Matlab exercise:

- Generate a sample of 100,000 variables with "Harry Potter" Gamma distribution with r =0.1 and k=9 ¾ (9.75)
- Calculate mean and compare it to k/r (Gamma)
- Calculate standard deviation and compare it to sqrt(k)/r (Gamma)
- Plot semilog-y plots of PDFs **and CCDFs.**
- Hint: read the help page (better yet documentation webpage) for random('Gamma'…): one of their parameters is different than r

Matlab exercise: Gamma

```
Stats=100000; r=0.1; k=9.75;
r2=random('Gamma', k,1./r, Stats,1);
disp([mean(r2),k./r]);
disp([std(r2),sqrt(k)./r]);
step=0.1; [a,b]=hist(r2,0:step:max(r2));
pdf_g=a./sum(a)./step;
figure;
subplot(1,2,1); semilogy(b,pdf_g,'ko-'); hold on;
x=0:0.01:max(r2); clear cdf_g;
for m=1:length(x);
    cdf_g(m)=sum(r2>x(m))./Stats;
end;
subplot(1,2,2); semilogy(x,cdf_g,'rd-');
```

Continuous Probability Distributions

Normal or Gaussian Distribution

PAY
ATTENTION

Normal or Gaussian Distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$-\infty < x < \infty$

is a normal random variable

with mean μ,

and standard dewviation σ

sometimes denoted as

$N(\mu, \sigma)$

Carl Friedrich Gauss (1777 –1855)
German mathematician

Normal Distribution

- The location and spread of the normal are independently determined by mean (μ) and standard deviation (σ)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

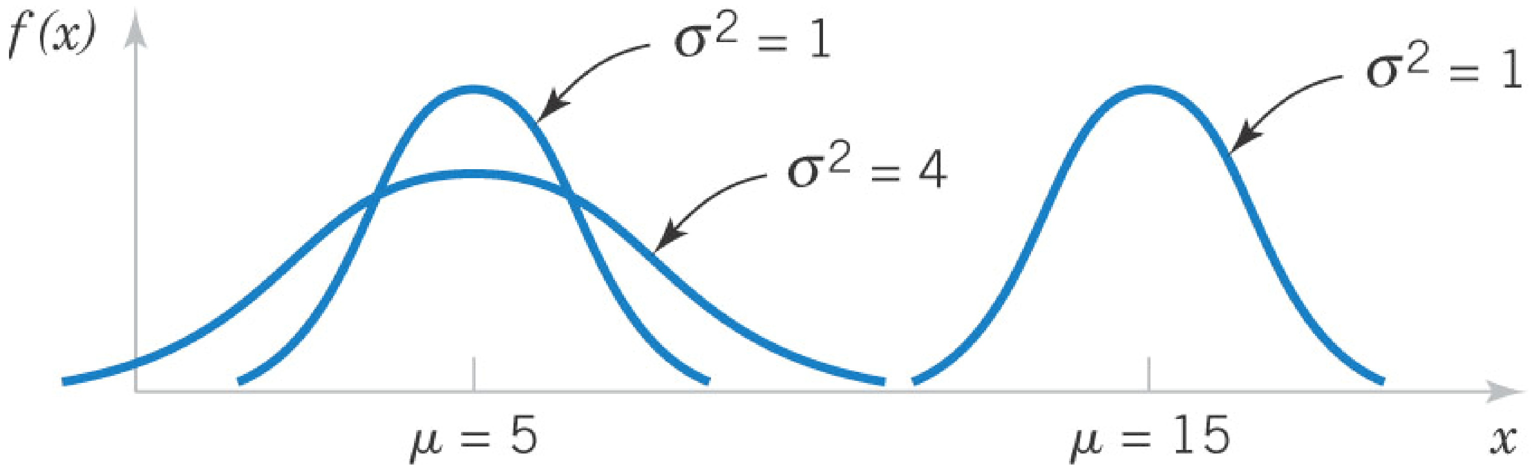


Figure 4-10 Normal probability density functions

Matlab exercise:

plot PDF of the Gaussian distribution

with **mu=3; sigma=2**

calculate mean, standard deviation and variance,

Linear-y and Semilog-y plots of PDF

Hint:

Generate Standard normal distribution using

randn(Stats,1) then

multiply and add using sigma, mu

Matlab exercise solution

- **Stats=100000;**
- **mu=3; sigma=2;**
- **r1=sigma.*randn(Stats,1)+mu;**
- **step=0.1;**
- **[a,b]=hist(r1,(mu-10.*sigma):step:(mu+10.*sigma));**
- **pdf_n=a./sum(a)./step;**
- **figure; subplot(1,2,1); plot(b,pdf_n,'ko-');**
- **subplot(1,2,2); semilogy(b,pdf_n,'ko-');**

Gaussian (Normal) distribution is very important because any sum of many independent random variables can be approximated with a Gaussian

Standard Normal Distribution

- A normal (Gaussian) random variable with

  $$\mu = 0 \text{ and } \sigma^2 = 1$$

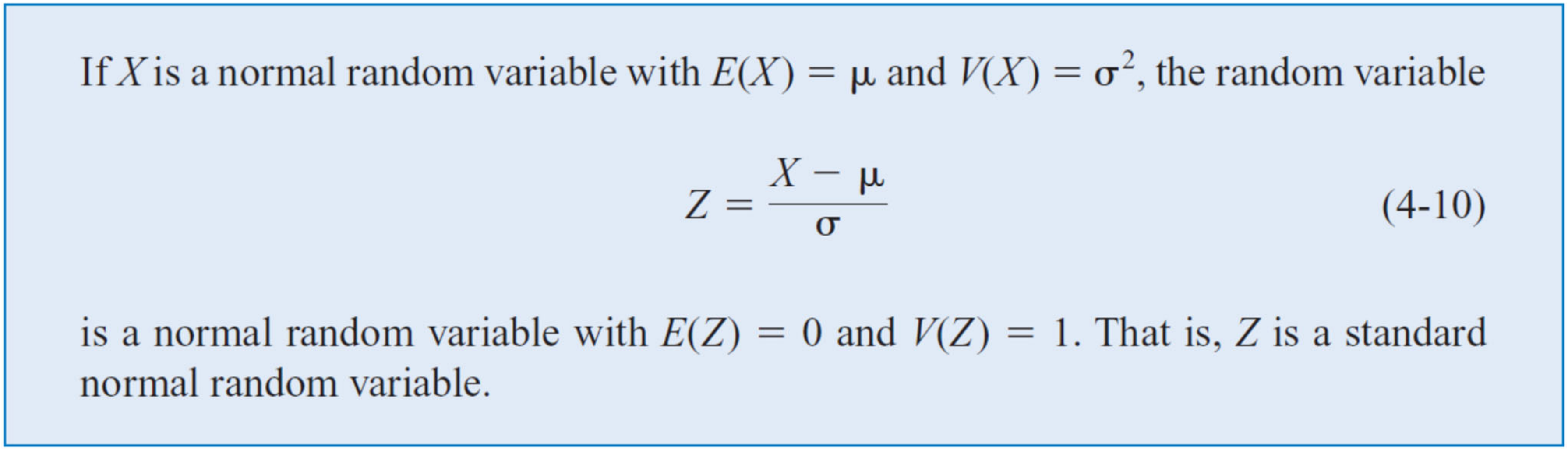
  is called a standard normal random variable and is denoted as Z.

- Thed cumulative distribution function of a standard normal random variable is denoted as:

  $$\Phi(z) = P(Z \leq z)$$

- Values are found in Appendix A Table III to Montgomery and Runger textbook

Standardizing



If X is a normal random variable with $E(X) = \mu$ and $V(X) = \sigma^2$, the random variable

$$Z = \frac{X - \mu}{\sigma} \qquad (4\text{-}10)$$

is a normal random variable with $E(Z) = 0$ and $V(Z) = 1$. That is, Z is a standard normal random variable.

Suppose X is a normal random variable with mean μ and variance σ^2.

Then, $P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P(Z \le z)$ (4-11)

where Z is a standard normal random variable, and

$z = \frac{(x-\mu)}{\sigma}$ is the z-value obtainedby standardizing x.

The probability is obtained by using Appendix Table III

$P(X < \mu - \sigma) = P(X > \mu + \sigma) = (1\text{-}0.68)/2=0.16=16\%$

$P(X < \mu - 2\sigma) = P(X > \mu + 2\sigma) =(1\text{-}\ 0.95)/2=0.023=2.3\%$

$P(X < \mu - 3\sigma) = P(X > \mu + 3\sigma) =(1\text{-}\ 0.997)/2=0.0013=0.13\%$

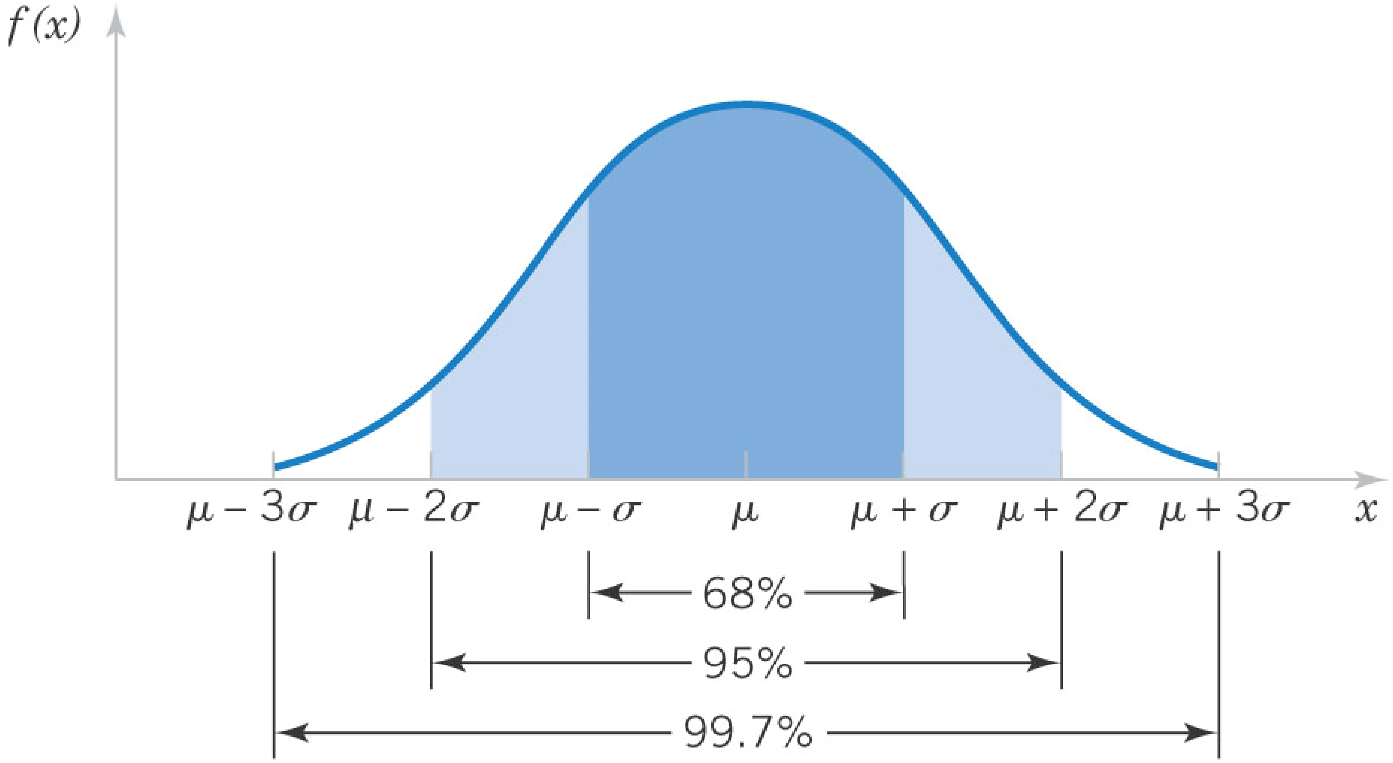


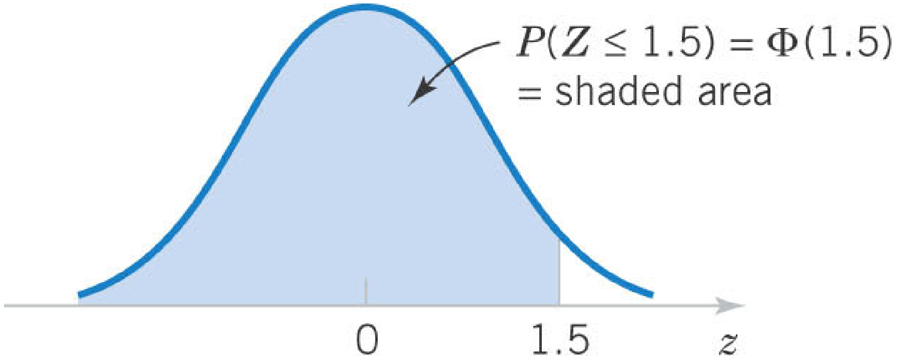
Figure 4-12 Probabilities associated with a normal distribution – well worth remembering to quickly estimate probabilities.

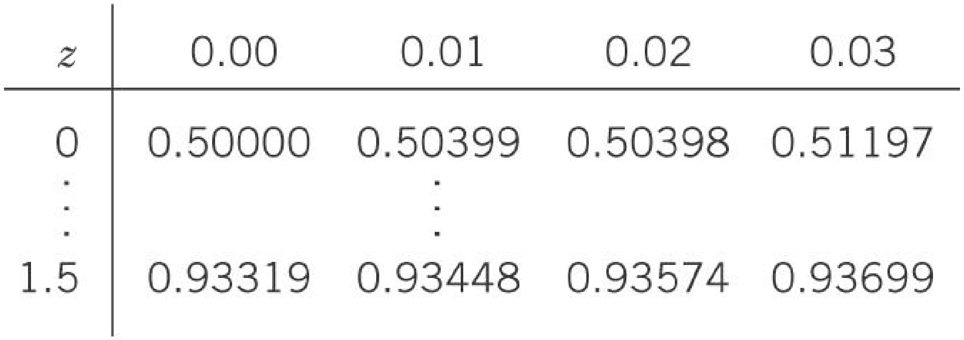
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

Standard Normal Distribution Tables

Assume *Z* is a standard normal random variable.

Find $P(Z \leq 1.50)$. Answer: 0.93319





z	0.00	0.01	0.02	0.03
0	0.50000	0.50399	0.50398	0.51197
⋮		⋮		
1.5	0.93319	0.93448	0.93574	0.93699

Figure 4-13 Standard normal PDF

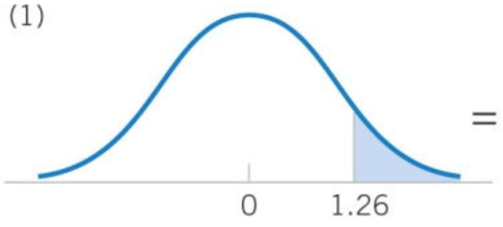
Table III from, Appendix A in Montgomery & Runger

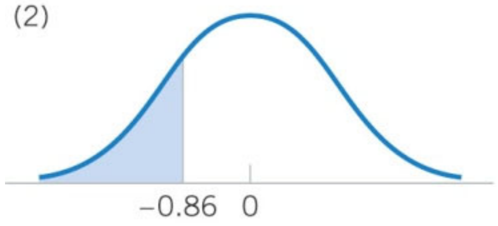
Find $P(Z \leq 1.53)$. Answer: 0.93699

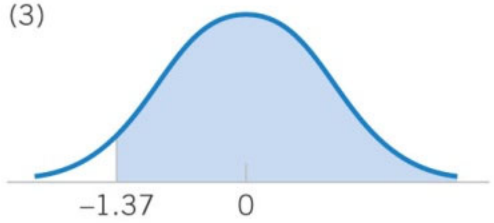
Find $P(Z \leq 0.02)$. Answer: 0.50398

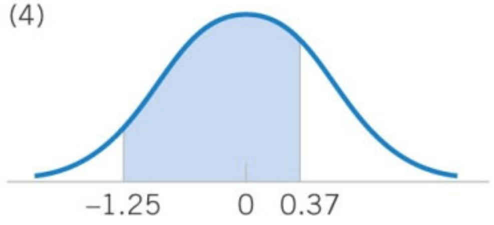
Standard Normal Exercises

1. P(Z > 1.26) =1- P(Z < 1.26) =1-0.8962= =0.1038
2. P(Z < -0.86) = P(Z >0.86)=1- P(Z <0.86)= 1-0.815=0.195
3. P(Z > -1.37) =P(Z<1.37)= 0.915
4. P(-1.25 < Z < 0.37) = P(Z < 0.37)- P(Z<-1.25) =P(Z < 0.37)- P(Z>1.25) = P(Z < 0.37)- (1-P(Z<1.25))= 0.6443-(1-0.8944)=0.5387

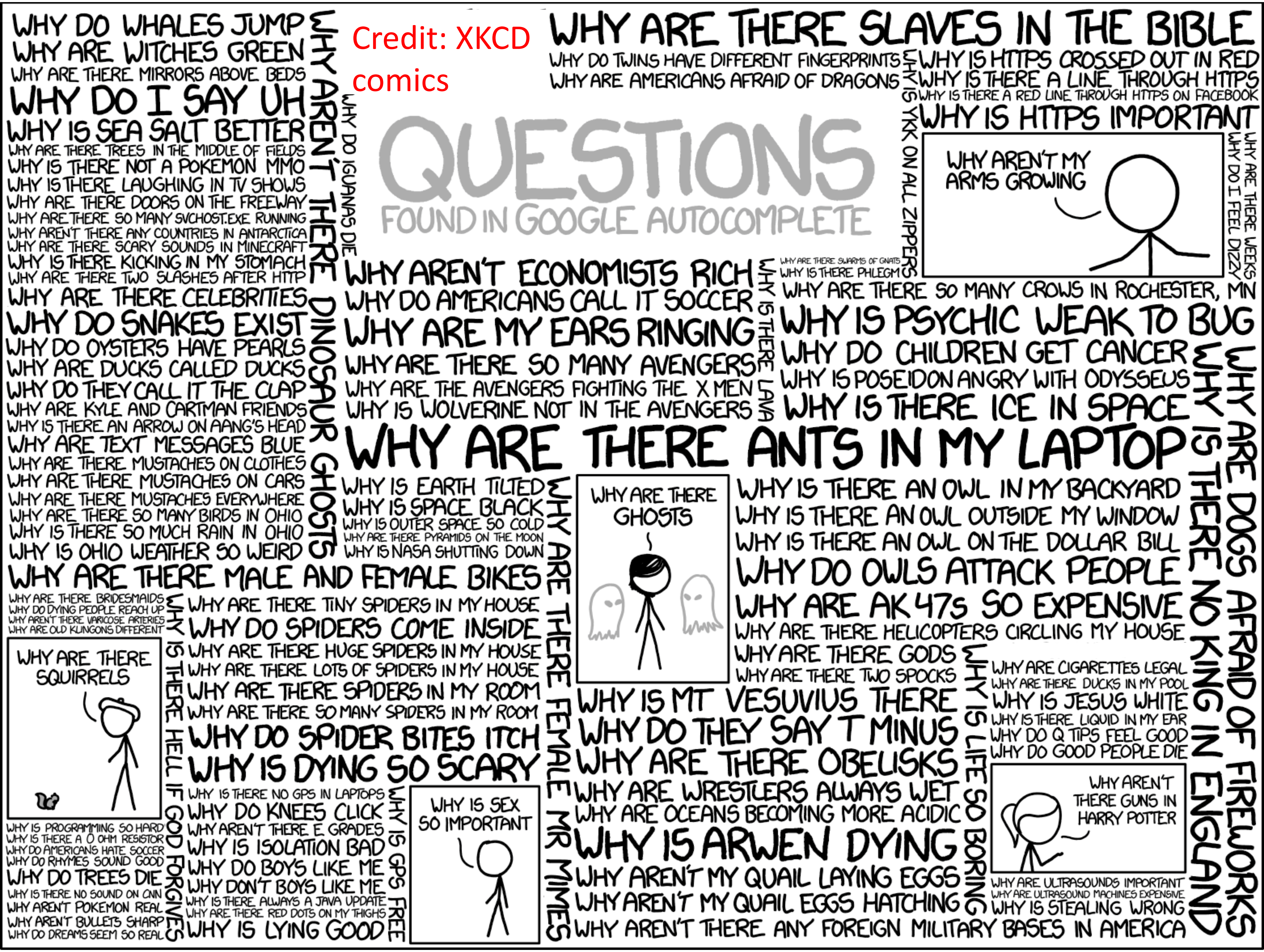








z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

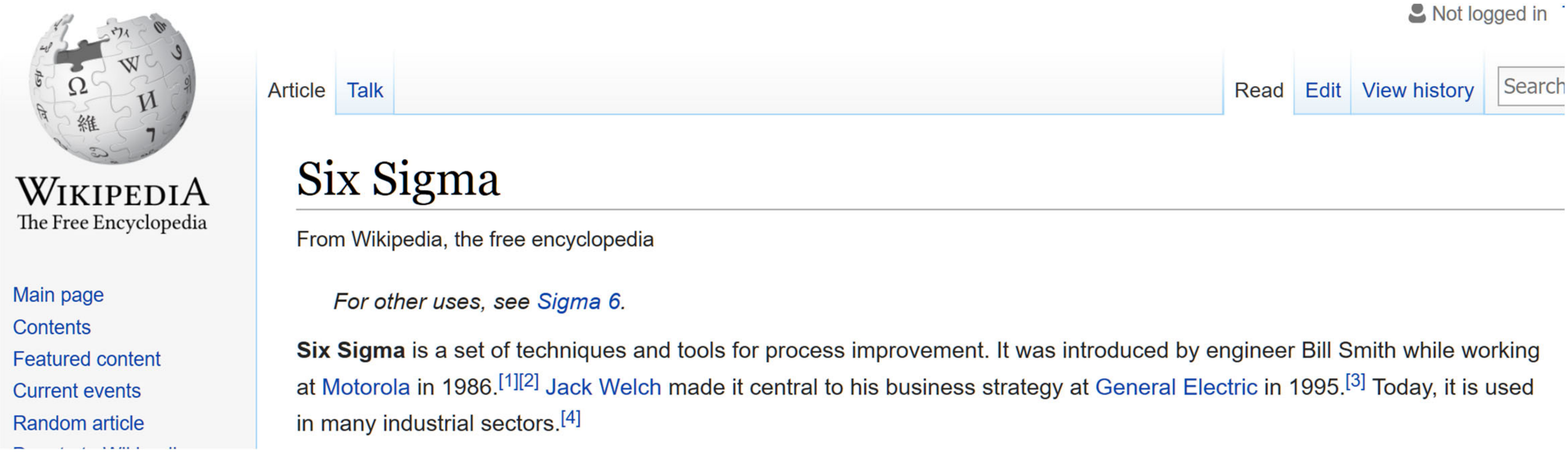

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Credit: XKCD comics

Range	The expected fraction of population inside the range	Approximate expected frequency outside the range	The approximate frequency for daily event
μ ± 0.5σ	0.382924922548026	2 in 3	Four or five times a week
μ ± 1σ	0.682689492137086	1 in 3	Twice a week
μ ± 1.5σ	0.866385597462284	1 in 7	Weekly
μ ± 2σ	0.954499736103642	1 in 22	Every three weeks
μ ± 2.5σ	0.987580669348448	1 in 81	Quarterly
μ ± 3σ	0.997300203936740	1 in 370	Yearly
μ ± 3.5σ	0.999534741841929	1 in 2149	Every six years
μ ± 4σ	0.999936657516334	1 in 15787	Every 43 years (twice in a lifetime)
μ ± 4.5σ	0.999993204653751	1 in 147160	Every 403 years (once in the modern era)
μ ± 5σ	0.999999426696856	1 in 1744278	Every 4776 years (once in recorded history)
μ ± 5.5σ	0.999999962020875	1 in 26330254	Every 72090 years (thrice in history of modern humankind)
μ ± 6σ	0.999999998026825	1 in 506797346	Every 1.38 million years (twice in history of humankind)
μ ± 6.5σ	0.999999999919680	1 in 12450197393	Every 34 million years (twice since the extinction of dinosaurs)
μ ± 7σ	0.999999999997440	1 in 390682215445	Every 1.07 billion years (four times in history of Earth)

Source: Wikipedia

Business buzzword: Six Sigma



Six Sigma

From Wikipedia, the free encyclopedia

For other uses, see Sigma 6.

Six Sigma is a set of techniques and tools for process improvement. It was introduced by engineer Bill Smith while working at Motorola in 1986.[1][2] Jack Welch made it central to his business strategy at General Electric in 1995.[3] Today, it is used in many industrial sectors.[4]

Business literature defined six sigma as no more than 3.4 defective products per million

Matlab group exercise 3

- $P(X-\mu>z\cdot\sigma)=P(Z>z)=$(1-erf(z./sqrt(2)))/2
- You can also use 1-normcdf(z)
- Calculate Prob($X-\mu>6\sigma$) and compare with expected 3.4 errors per million
- Find z such that Prob($X-\mu>z\cdot\sigma$)=3.4 errors per million

What Six Sigma should be really called if *$P(X-\mu > z\cdot \sigma) = 3.4e-6$*

A. 6 sigma
B. 7 sigma
C. 3 sigma
D. 4.5 sigma
E. I could not figure it out

Get your i-clickers

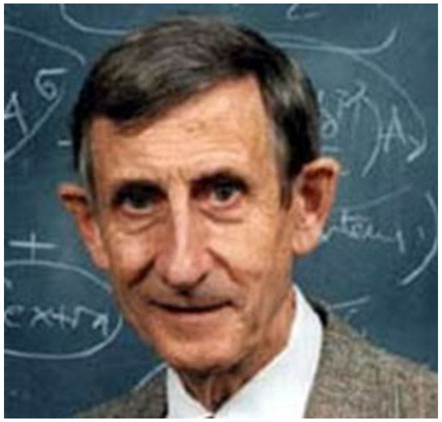
Appendix Table III is no good for 6-sigma How to calculate in Matlab?

- Matlab has a built-in function normcdf
- 1-normcdf(z) is the Prob[X-μ>z·σ]
- I expected: P(Z>6)= 3.4e-6
- Matlab says 1-normcdf(6)~ 1e-9
- Six sigma is not 6σ at all !!!
- Let’s find out how many simas are in six sigma
- Matlab says: invnorm(3.4e-6)=4.5
- Six sigma should be called 4.5σ
- Does not have the same buzz

What’s wrong with Six Sigma?

- Motorola has determined, through years of process and data collection, that processes vary and drift over time – what they call the Long-Term Dynamic Mean Variation. This variation typically falls between 1.4 and 1.6. They shifted their
sigma down by 1.5.
- The statistician Donald J. Wheeler has dismissed the 1.5 sigma shift as "goofy" because of its arbitrary nature.
- A *Fortune* article stated that "of 58 large companies that have announced Six Sigma programs, 91 percent have trailed (performed below) the S&P 500 index since"

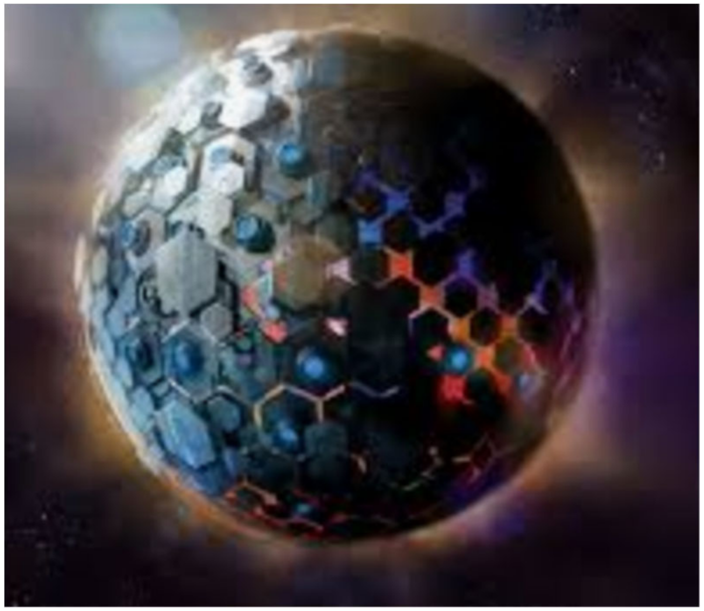
- Freeman Dyson (a famous theoretical physicist) once sat on a committee reviewing Department of Energy Joint Genomics Institute (DOE JGI)
- Motorola sent their six-sigma preacher Freeman Dyson asked him:
  - D: Can you explain me what is six–sigma?
    - P: Mumbling something about it being the gold standard of reliability
  - D: Can you at least define one-sigma?
    - P: Silence
- Six-sigma was never implemented at JGI

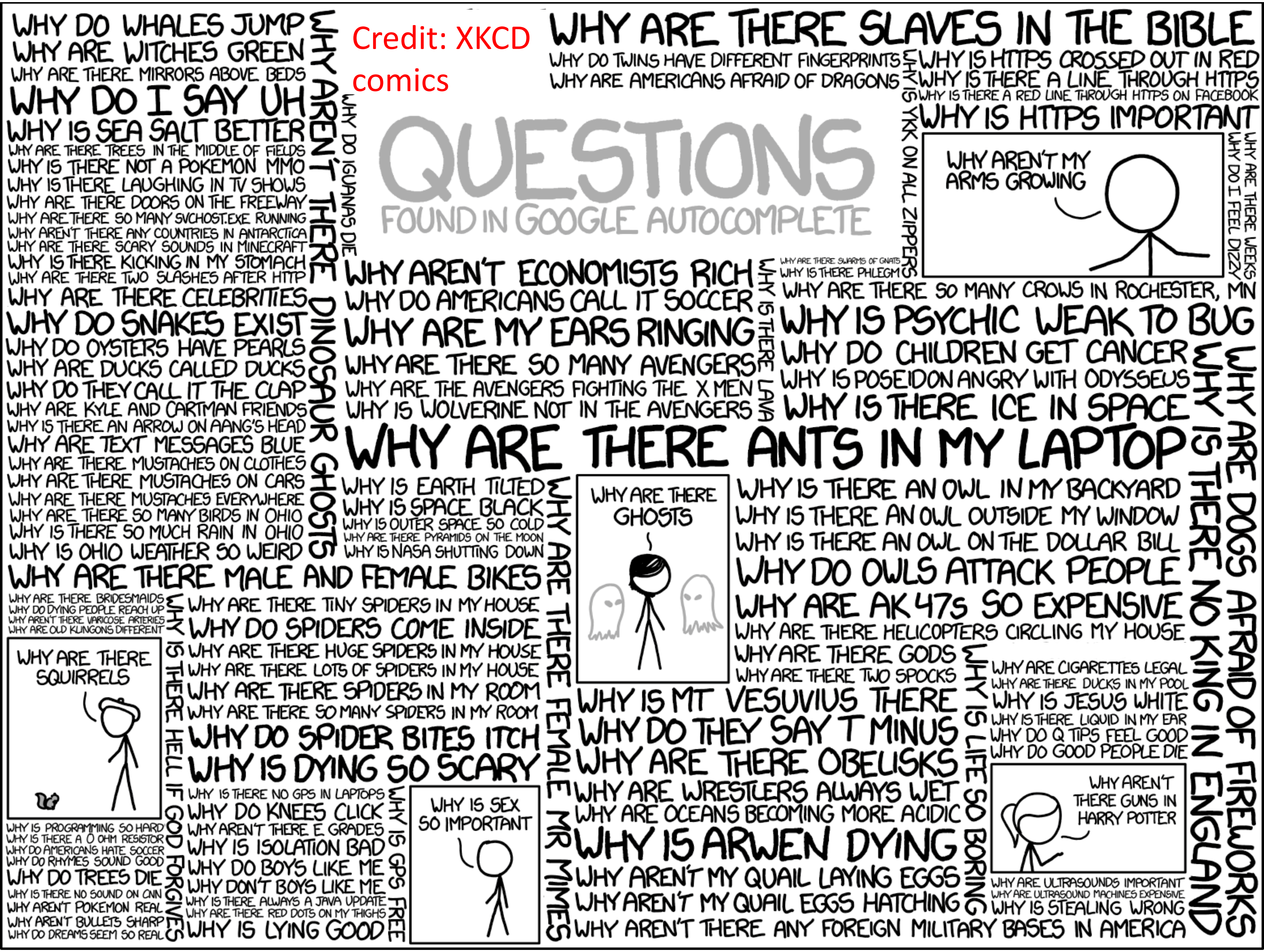


Born:
December 15, 1923,
Crowthorne, UK
Died:
February 28, 2020
Princeton, NJ USA

Dyson’s legacy

- **Seminal contributions to quantum mechanics**
- The Origin of Life:
  Cells → Enzymes → DNA/RNA later
  First proposed by Alexander Oparin in 1922
- Dyson sphere:
  Completely
  captures light from a star
- Dyson tree:
  genetically engineered
  tree growing inside a
  comet



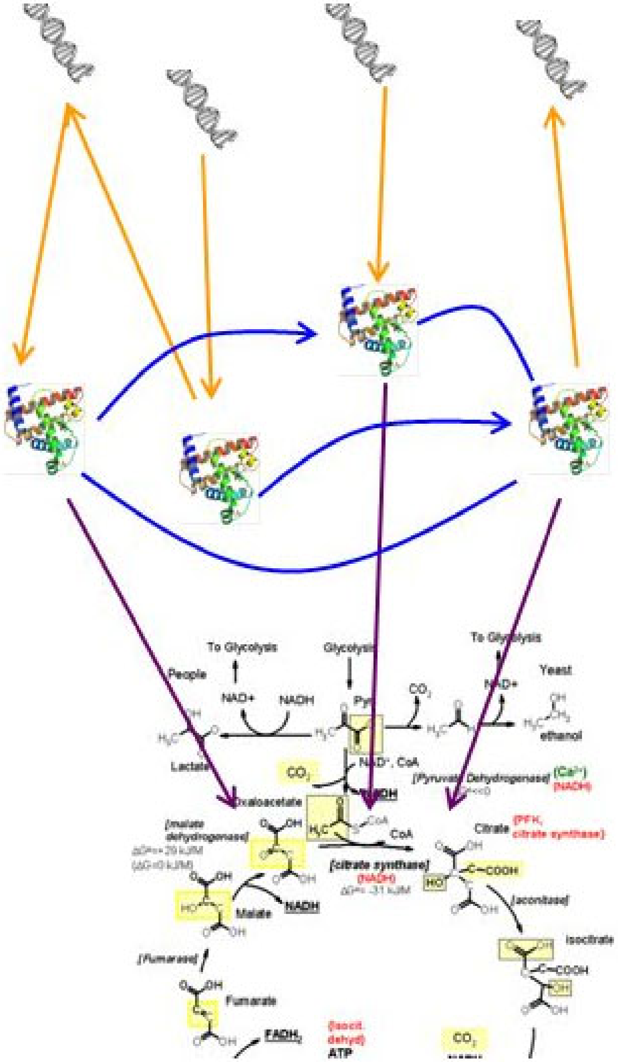

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Credit: XKCD comics

Fitting a Gaussian distribution: a biological example

Molecular binding is used at multiple levels

Each level has its own molecular interaction network



Regulatory network:

RNA-level regulation

By DNA-binding Proteins

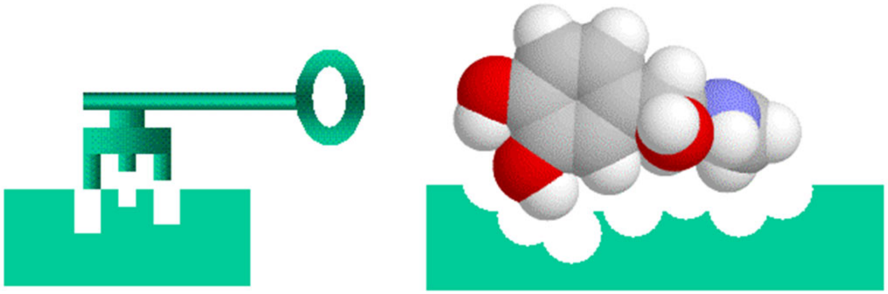
Protein-Protein (binding) Interaction Network

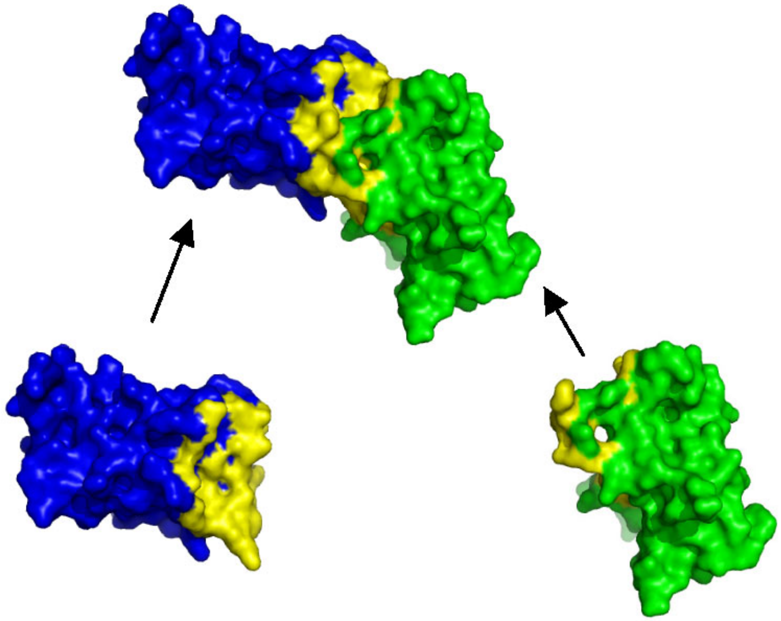
Protein-Metabolite Interactions:

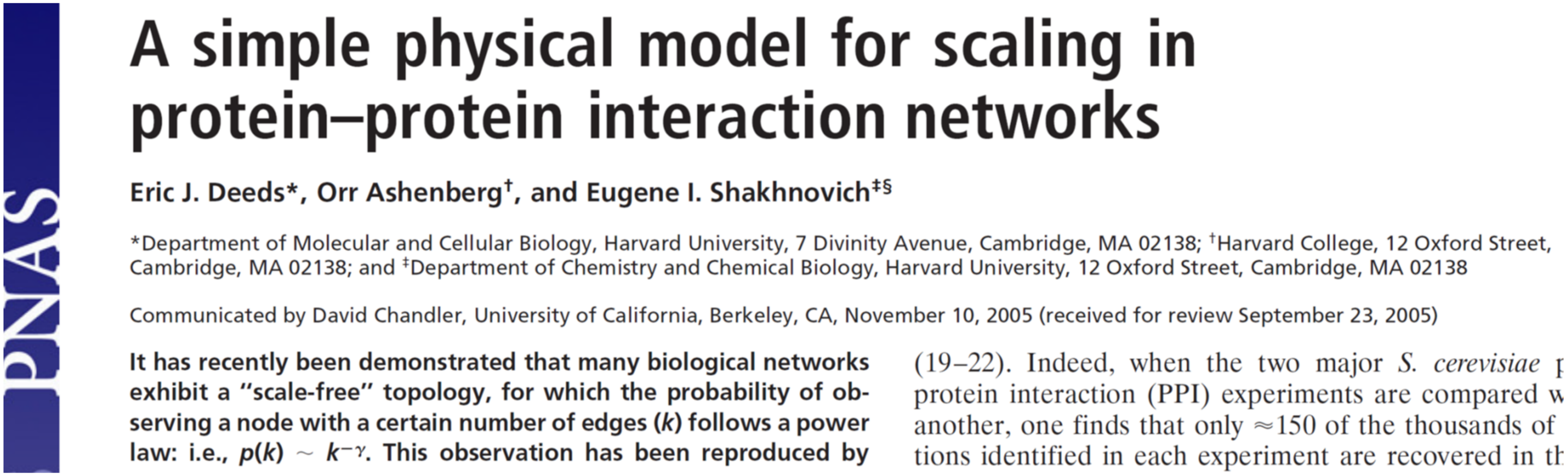
Metabolic network

Biological example of a Gaussian: Energy of Protein-Protein Binding Interactions

- Proteins and other biomolecules (metabolites, drugs, DNA) specifically (and non-specifically) bind each other
- For specific bindings: Lock-and-Key theory
- For non-specific bindings: random contacts







PNAS

A simple physical model for scaling in protein–protein interaction networks

Eric J. Deeds*, Orr Ashenberg[†], and Eugene I. Shakhnovich[‡§]

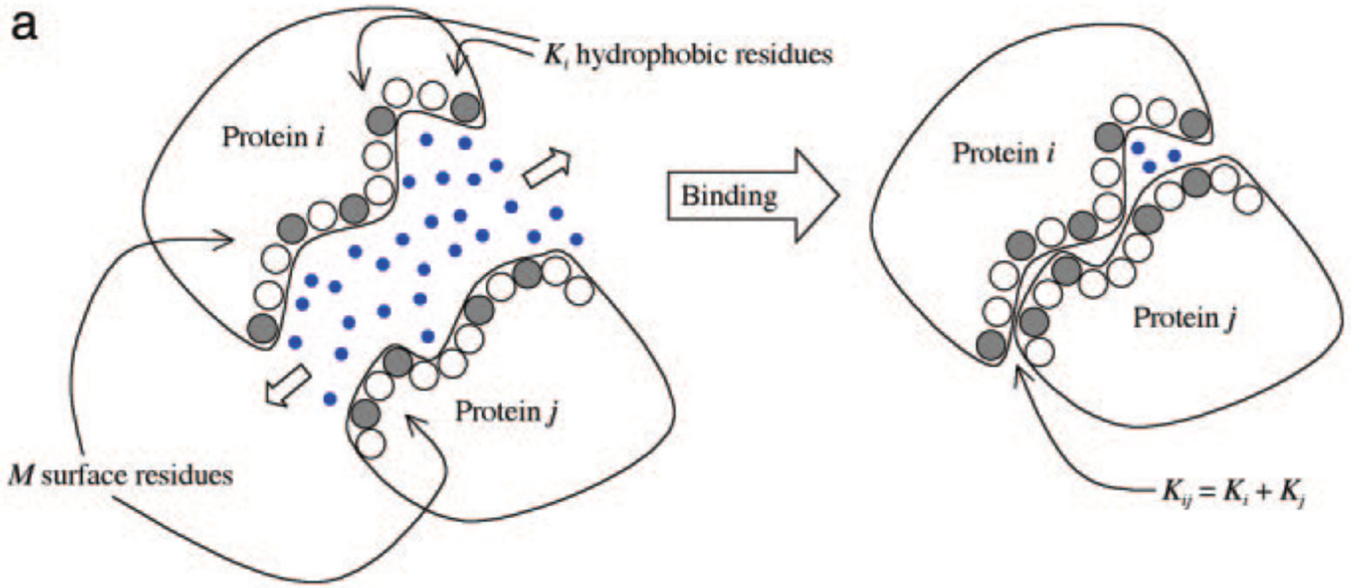
*Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138; [†]Harvard College, 12 Oxford Street, Cambridge, MA 02138; and [‡]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138

Communicated by David Chandler, University of California, Berkeley, CA, November 10, 2005 (received for review September 23, 2005)

It has recently been demonstrated that many biological networks exhibit a "scale-free" topology, for which the probability of observing a node with a certain number of edges (k) follows a power law: i.e., $p(k) \sim k^{-\gamma}$. This observation has been reproduced by

(19–22). Indeed, when the two major *S. cerevisiae* p
protein interaction (PPI) experiments are compared w
another, one finds that only ≈150 of the thousands of
tions identified in each experiment are recovered in th

Most Binding energy is due to hydrophobic amino-acid residues being screened from water

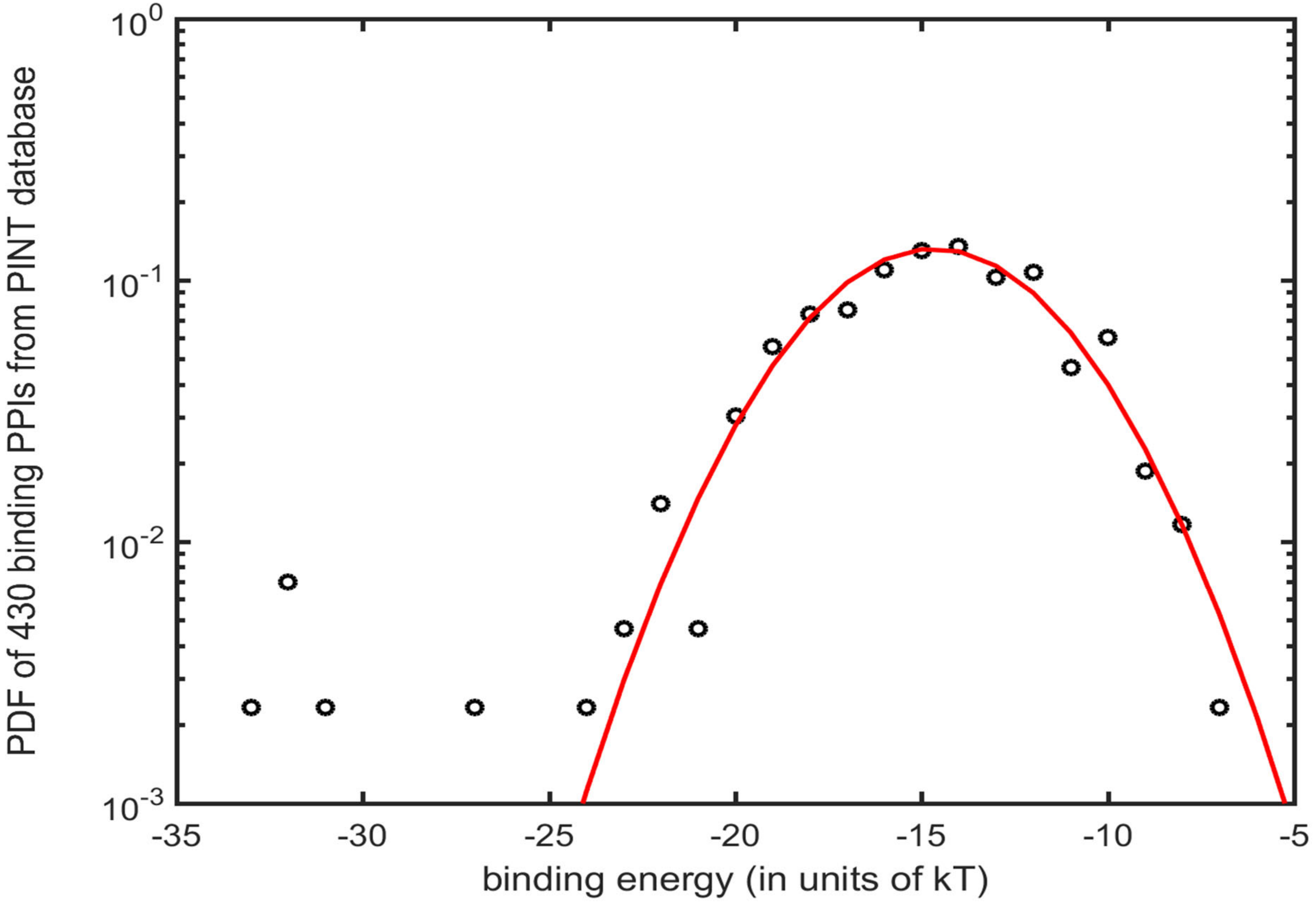


Predicted Gaussian distribution: PDF(E_{ij}=E)– because E_{ij} – sum of hydrophobicities of many independent residues

Matlab exercise

- In Matlab load PINT_binding_energy.mat with binding energy E_{ij} (in units of kT at room temperature) for 430 pairs of interacting proteins from human, yeast, etc.
- Data collected in 2007 from the PINT database http://www.bioinfodatabase.com/pint/ and analyzed in J. Zhang, **S. Maslov**, E. Shakhnovich, Molecular Systems Biology (2008)
- Fit Gaussian to the distribution of E_{ij} using dfittool
- Use “Exclude” button to generate the new exclusion rule to drop all points with X<-23 from the fit
- Use "New Fit" button to generate the new “Normal” fit with the exclusion rule you just created
- Find mean (mu) and standard deviation (sigma)
- Select “probability plot” from “Display type” dropdown menu to evaluate the quality of the plot. Where does the probability plot deviate from a straight line?

How does it compare with the experimental data ?



J. Zhang, **S. Maslov**, E. Shakhnovich,
Nature/EMBO Molecular Systems Biology (2008)

Data on binding interactions
from PINT database

Dissociation constant

- Interaction between two molecules (say, proteins) is usually described in terms of dissociation constant
  K_{ij}=1M exp($-E_{ij}/kT$)
- Law of Mass Action: the concentration D_{ij} of a heterodimer formed out of two proteins with free (monomer) concentrations C_i and C_j : $D_{ij}=C_iC_j/K_{ij}$
- What is the distribution of K_{ij}?
- Answer: it is called log-normal since the logarithm of K_{ij} is the binding energy $-E_{ij}/kT$ which is normally distributed

Lognormal Distribution

- Let *W* denote a normal random variable with mean of θ and variance of ω^2, i.e., $E(W) = \theta$ and $V(W) = \omega^2$
- As a change of variable, let $X = e^W = \exp(W)$ and $W = \ln(X)$
- Now X is a lognormal random variable.

$$F(x) = P[X \le x] = P[\exp(W) \le x] = P[W \le \ln(x)]$$

$$= \mathrm{P}\left[\mathrm{Z} \le \frac{\ln(\mathrm{x})\text{-}\theta}{\omega}\right] = \Phi\left[\frac{\ln(\mathrm{x})\text{-}\theta}{\omega}\right] = \quad \text{for} \quad x > 0$$

$$= 0 \quad \text{for} \quad x \le 0$$

$$f(x) = \frac{dF(x)}{dx} = \frac{1}{x\omega\sqrt{2\pi}} e^{-\left[\frac{\ln(x)-\theta}{2\omega}\right]^2} \qquad \text{for } 0 < x < \infty$$

$$E(X) = e^{\theta+\omega^2/2} \qquad \text{and} \qquad V(X) = e^{2\theta+\omega^2}\left(e^{\omega^2} - 1\right) \qquad \text{(4-22)}$$

Lognormal Graphs

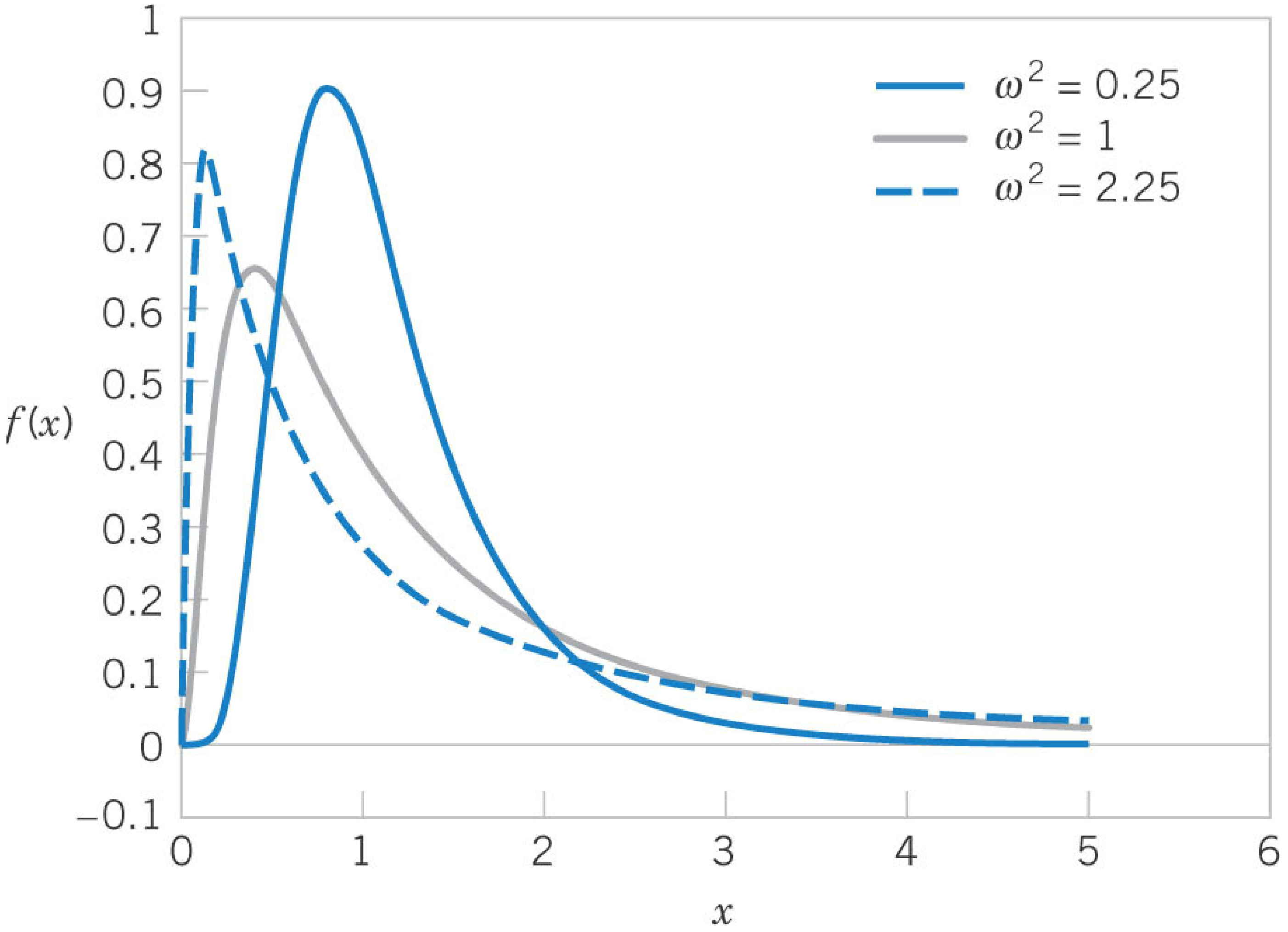
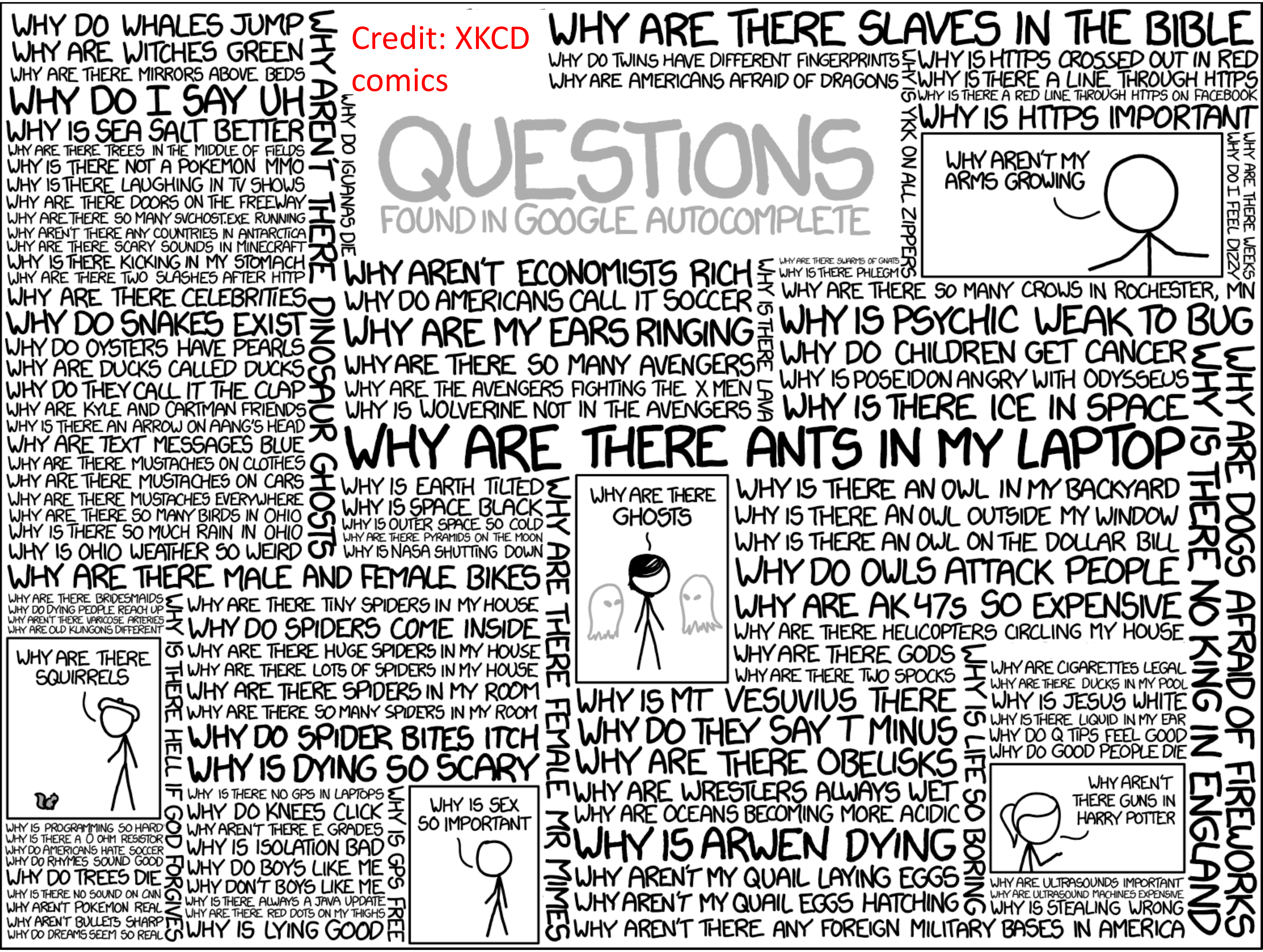


Figure 4-27 Lognormal probability density functions with θ = 0 for selected values of ω^2.


WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Credit: XKCD comics

Multiple random variables, Correlations

What we learned so far…

- Random Events:
  - Working with events as sets: union, intersection, etc.
    - Some events are simple: Head vs Tails, Cancer vs Healthy
    - Some are more complex: 10<Gene expression<100
    - Some are even more complex: Series of dice rolls: 1,3,5,3,2
  - Conditional probability: *P(A|B)=P(A ∩ B)/P(B)*
  - Independent events: *P(A|B)=P(A)* or *P(A ∩ B)= P(A)*P(B)*
  - Bayes theorem: relates *P(A|B) to P(B|A)*
- Random variables:
  - Mean, Variance, Standard deviation. How to work with E(g(X))
  - Discrete (Uniform, Bernoulli, Binomial, Poisson, Geometric, Negative binomial, Power law);
    PMF: f(x)=Prob(X=x); CDF: F(x)=Prob(X≤x);
  - Continuous (Uniform, Exponential, Erlang, Gamma, Normal, Log-normal);
    PDF: f(x) such that Prob(X inside A)= $\int_A$ f(x)dx; CDF: F(x)=Prob(X≤x)
- Next step: work with **multiple random variables** measured together in the same series of random experiments

Concept of Joint Probabilities

- Biological systems are usually described not by a single random variable but by many random variables
- Example: The expression state of a human cell: 20,000 random variables X_i for each of its genes
- A joint probability distribution describes the behavior of several random variables
- We will start with just two random variables X and Y and generalize when necessary

Joint Probability Mass Function Defined

The joint probability mass function of the discrete random variables X and Y,
denoted as $f_{XY}(x, y)$, satifies:
(1) $f_{XY}(x, y) = P$
(2) $f_{XY}(x, y) \geq 0$ All probabilities are non−negative
(3) $\sum_x \sum_y f_{XY}(x, y) = 1$ The sum of all probabilities is 1

Montgomery Runger 5th edition Equation (5−1)

Example 5-1: # Repeats vs. Signal Bars

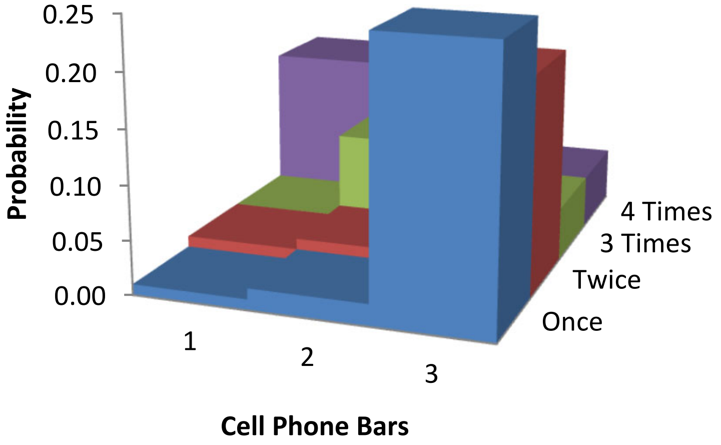
You use your cell phone to check your airline reservation. It asks you to speak the name of your departure city to the voice recognition system.

- Let Y denote the number of times you have to state your departure city.
- Let X denote the number of bars of signal strength on you cell phone.

y = number of times city name is stated	x = number of bars of signal strength		
	1	2	3
1	0.01	0.02	0.25
2	0.02	0.03	0.20
3	0.02	0.10	0.05
4	0.15	0.10	0.05

Figure 5-1 Joint probability distribution of X and Y. The table cells are the probabilities. Observe that more bars relate to less repeating.

Bar Chart of Number of Repeats vs. Cell Phone Bars



Marginal Probability Distributions (discrete)

For a discrete joint PDF, there are marginal distributions for each random variable, formed by summing the joint PMF over the other variable.

$$f_X(x) = \sum_y f_{XY}(x,y)$$

$$f_Y(y) = \sum_x f_{XY}(x,y)$$

y = number of times city name is stated	x = number of bars of signal strength			
	1	2	3	$f_Y(y)$ =
1	0.01	0.02	0.25	0.28
2	0.02	0.03	0.20	0.25
3	0.02	0.10	0.05	0.17
4	0.15	0.10	0.05	0.30
$f_X(x)$ =	0.20	0.25	0.55	1.00

Called marginal because they are written in the margins

Figure 5-6 From the prior example, the joint PMF is shown in green while the two marginal PMFs are shown in purple.

Mean & Variance of X and Y are calculated using marginal distributions

y = number of times city name is stated	x = number of bars of signal strength					
	1	2	3	$f(y) =$	$y*f(y) =$	$y^2*f(y) =$
1	0.01	0.02	0.25	0.28	0.28	0.28
2	0.02	0.03	0.20	0.25	0.50	1.00
3	0.02	0.10	0.05	0.17	0.51	1.53
4	0.15	0.10	0.05	0.30	1.20	4.80
$f(x) =$	0.20	0.25	0.55	1.00	2.49	7.61
$x*f(x) =$	0.20	0.50	1.65	2.35		
$x^2*f(x) =$	0.20	1.00	4.95	6.15		

$\mu_X = E(X) = 2.35; \quad \sigma_X^2 = V(X) = 6.15 - 2.35^2 = 6.15 - 5.52 = 0.6275$

$\mu_Y = E(Y) = 2.49; \quad \sigma_Y^2 = V(Y) = 7.61 - 2.49^2 = 7.61 - 16.20 = 1.4099$

Conditional Probability Distributions

Recall that $P(B|A) = \dfrac{P(A \cap B)}{P(A)}$

$P(Y{=}y\,|\,X{=}x){=}P(X{=}x{,}Y{=}y)/P(X{=}x){=}$
${=}f(x{,}y)/f_X(x)$

From Example 5-1

$P(Y{=}1\,|\,X{=}3) = 0.25/0.55 = 0.455$

$P(Y{=}2\,|\,X{=}3) = 0.20/0.55 = 0.364$

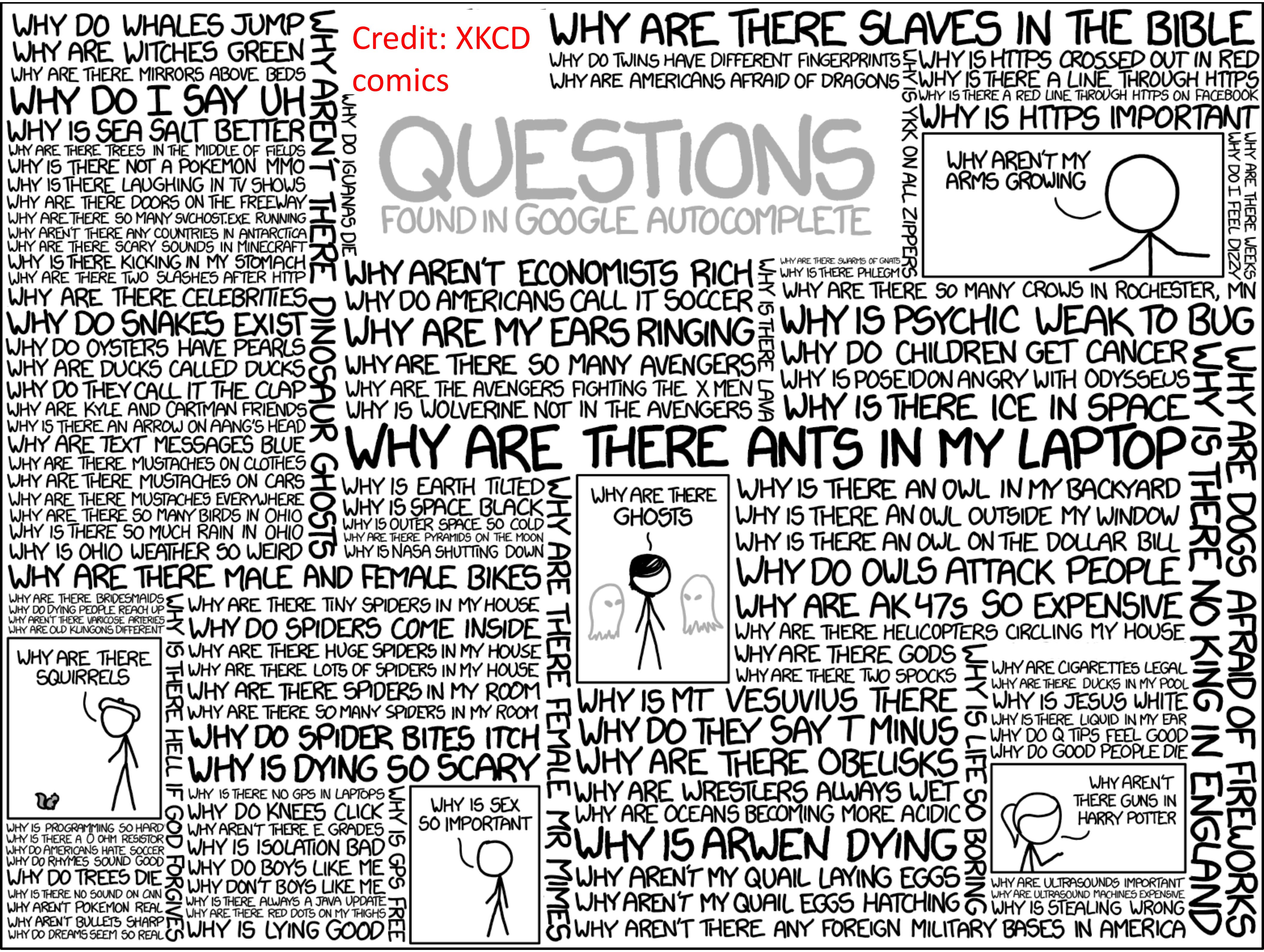
$P(Y{=}3\,|\,X{=}3) = 0.05/0.55 = 0.091$

$P(Y{=}4\,|\,X{=}3) = 0.05/0.55 = 0.091$

Sum = *1.00*

y = number of times city name is stated	x = number of bars of signal strength			
	1	2	3	$f_Y(y)$ =
1	0.01	0.02	0.25	0.28
2	0.02	0.03	0.20	0.25
3	0.02	0.10	0.05	0.17
4	0.15	0.10	0.05	0.30
$f_X(x)$ =	0.20	0.25	0.55	1.00

Note that there are 12 probabilities conditional on *X*, and 12 more probabilities conditional upon *Y*.


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS

Credit: XKCD comics

Joint Probability Density Function Defined

The joint probability density function for the continuous random variables *X* and *Y*, denotes as $f_{XY}(x,y)$, satisfies the following properties:

$$(1)\ \ f_{XY}(x,y) \geq 0 \text{ for all } x, y$$

$$(2)\ \ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1$$

$$(3)\ \ P\big((X,Y) \subset R\big) = \iint_{R} f_{XY}(x,y)dxdy \qquad (5\!-\!2)$$

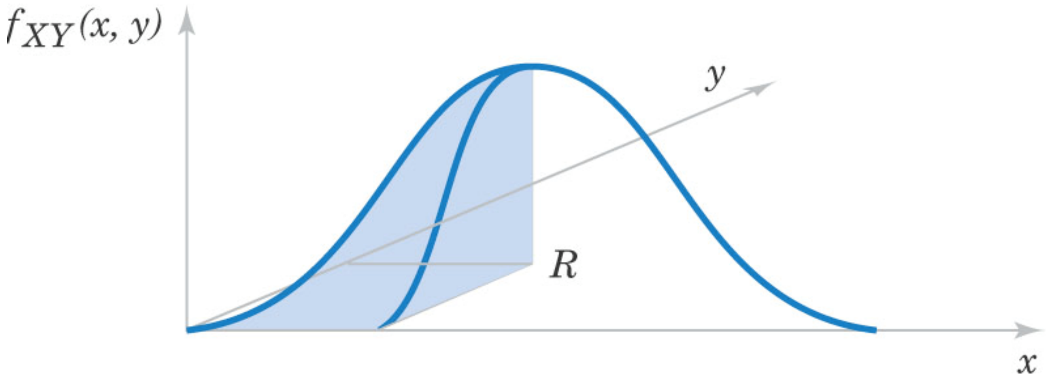


Figure 5-2 Joint probability density function for the random variables *X* and *Y*. Probability that (*X*, *Y*) is in the region *R* is determined by the volume of $f_{XY}(x,y)$ over the region *R*.

Joint Probability Density Function Graph

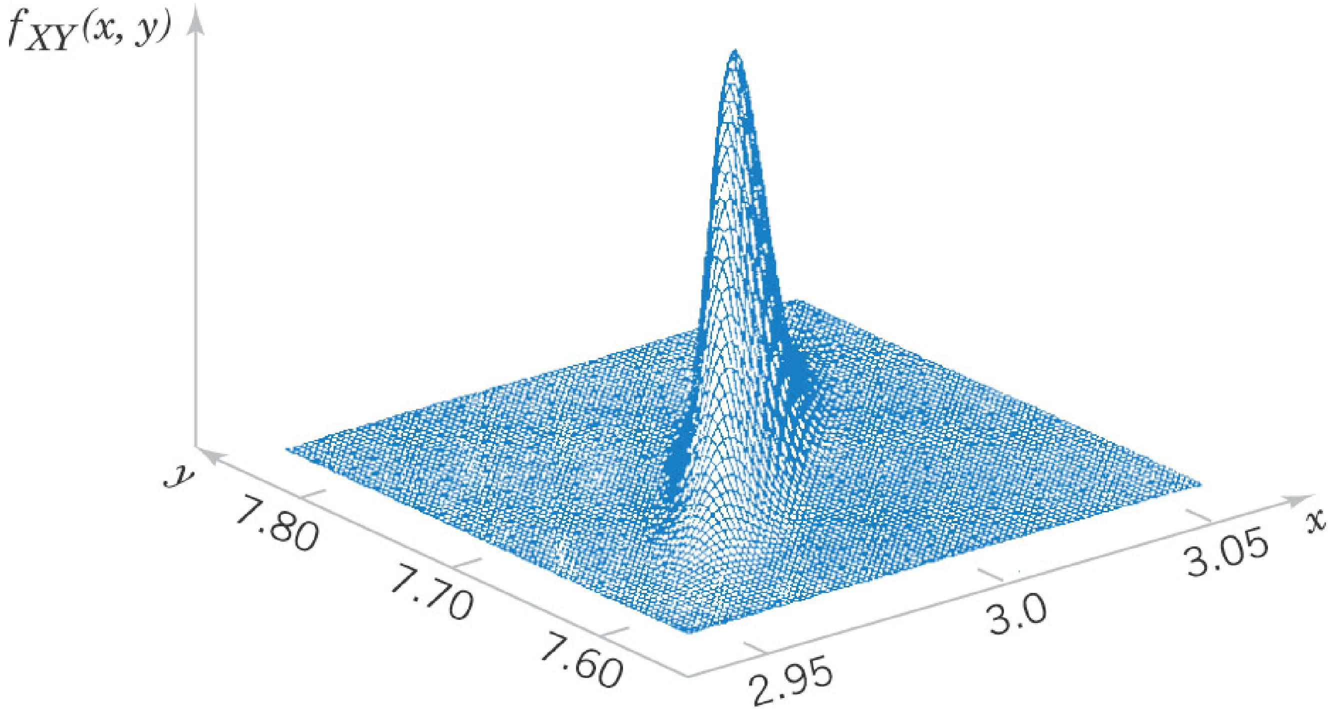


Figure 5-3 Joint probability density function for the continuous random variables *X* and *Y* of expression levels of two different genes. Note the asymmetric, narrow ridge shape of the PDF – indicating that small values in the *X* dimension are more likely to occur when small values in the *Y* dimension occur.

Marginal Probability Distributions (continuous)

- Rather than summing a discrete joint PMF, we integrate a continuous joint PDF.
- The marginal PDFs are used to make probability statements about one variable.
- If the joint probability density function of random variables *X* and *Y* is $f_{XY}(x,y)$, the marginal probability density functions of *X* and *Y* are:

$$f_X(x) = \int_y f_{XY}(x,y)\,dy$$

$$f_Y(y) = \int_x f_{XY}(x,y)\,dx \qquad (5\!-\!3)$$

$$f_X(x) = \sum_y f_{XY}(x,y)$$

$$f_Y(y) = \sum_x f_{XY}(x,y)$$

Conditional Probability Density Function Defined

Given continuous random variables X and Y with joint probability density function $f_{XY}(x, y)$, the conditional probability densiy function of Y given X=x is

$$f_{Y|x}(y) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{XY}(x,y)}{\int_y f_{XY}(x,y)\,dy} \text{ if } \; f_X(x) > 0 \qquad (5\text{−}4)$$

which satifies the following properties:

(1) $f_{Y|x}(y) \geq 0$

(2) $\int f_{Y|x}(y)dy = 1$

(3) $P(Y \subset B|X = x) = \int_B f_{Y|x}(y)dy$ for any set B in the range of Y

Compare to discrete: *$P(Y{=}y|X{=}x){=}f_{XY}(x,y)/f_X(x)$*

Conditional Probability Distributions

- Conditional probability distributions can be developed for multiple random variables by extension of the ideas used for two random variables.
- Suppose *p* = 5 and we wish to find the distribution of X_1, X_2 and X_3 conditional on X_4=x_4 and X_5=x_5.

$$f_{X_1X_2X_3|x_4x_5}(x_1,x_2,x_3)=\frac{f_{X_1X_2X_3X_4X_5}(x_1,x_2,x_3,x_4,x_5)}{f_{X_4X_5}(x_4,x_5)}$$

for $f_{X_4X_5}(x_4,x_5)>0$.

Independence for Continuous Random Variables

For random variables *X* and *Y*, if any one of the following properties is true, the others are also true. Then *X* and *Y* are independent.

(1) $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$

(2) $f_{Y|x}(y) = f_Y(y)$ for all x and y with $f_X(x) > 0$

(3) $f_{X|y}(y) = f_X(x)$ for all x and y with $f_Y(y) > 0$

(4) $\mathrm{P}(X \subset A, Y \subset B) = P(X \subset A) \cdot P(Y \subset B)$ for any sets A and B in the range of X and Y, respectively. (5−7)

P(Y=y|X=x)=P(Y=y) **for any *x*** or

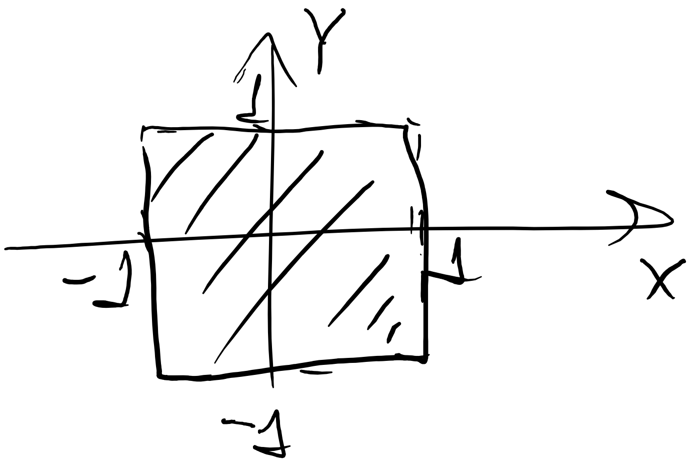
P(X=x|Y=y)=P(X=x) **for any *y*** or

P(X=x, Y=y)=P(X=x)·P(Y=y) **for any *x* and *y***

Example 1:

Uniform distribution in the square

$-1 < X < 1, \quad -1 < Y < 1$



$$f_{XY}(x,y) = \begin{cases} c & \text{if } -1 < x < 1 \text{ and } -1 < y < 1 \\ 0 & \text{- otherwise} \end{cases}$$

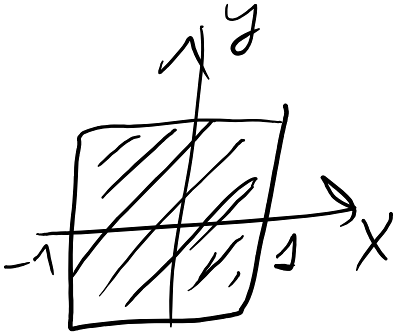
$$1 = \int_{\text{square}} dx\,dy\, f_{XY}(x,y) = c \cdot \text{Area} = c \cdot 4 \Rightarrow c = \frac{1}{4}$$

Are X and Y independent? Yes they are

Let's test if $f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dy =$$

$$= \int_{-1}^{1} \frac{1}{4}\,dy = \frac{1}{2} \text{ if } -1 < x < 1$$



Same for $f_Y(y) = \frac{1}{2}$ if $-1 < y < 1$

$$\frac{1}{4} = f_{XY}(x,y) = \frac{1}{2} \cdot \frac{1}{2} = f_X(x) \cdot f_Y(y)$$

if both x & y are in $[-1,1]$

0 otherwise

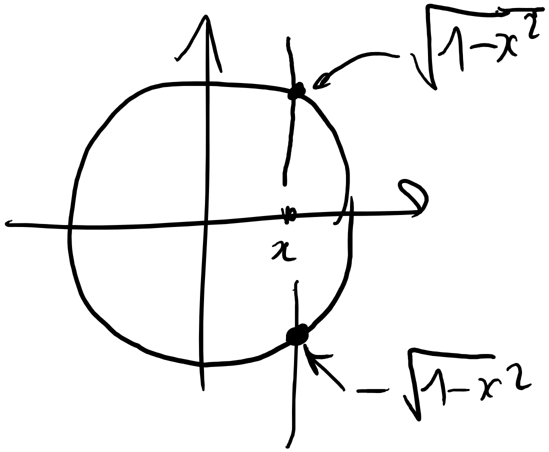
Joint PDF $f_{XY}(x,y) = \frac{1}{\text{area}} = \frac{1}{\pi}$ if x, y in the disc

0 – otherwise

Marginal distributions:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{dy}{\pi} = \frac{2\sqrt{1-x^2}}{\pi}$$

Same for $f_Y(y) = \frac{2\sqrt{1-y^2}}{\pi}$



$$\frac{1}{\pi} = f_{XY}(x,y) \neq \frac{2}{\pi}\sqrt{1-x^2}\cdot\frac{2}{\pi}\sqrt{1-y^2} = f_X(x)\cdot f_Y(y)$$

Variables are NOT independent

Covariation, Correlations

Quick and dirty check for **linear** (in)dependence between variables

Covariance Defined

Covariance is a number quantifying the average *linear* dependence between two random variables.

The covariance between the random variables X and Y, denoted as $\mathrm{cov}(X, Y)$ or σ_{XY} is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X\mu_Y$$

Montgomery, Runger 5th edition Eq. (5–14)

The units of σ_{XY} are the units of X times the units of Y.

Unlike the range of the variance, covariance can be negative: $-\infty < \sigma_{XY} < \infty$.

Covariance – 1 number to measure dependance between random variables

$\mathrm{Cov}(X,Y)$ or σ_{XY}

$$\sigma_{XY} = E\left[(X-\mu_X)\cdot(Y-\mu_Y)\right] =$$

$$= E(X \cdot Y) - \mu_X \cdot \mu_Y$$

- $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$
- If X & Y are independent

$$\mathrm{Cov}(X,Y) = E[X-\mu_X]\cdot E[Y-\mu_Y] = 0$$

- $\underline{\underline{\underline{-\infty}}} < \mathrm{Cov}(X,Y) < +\infty$ Can be negative!

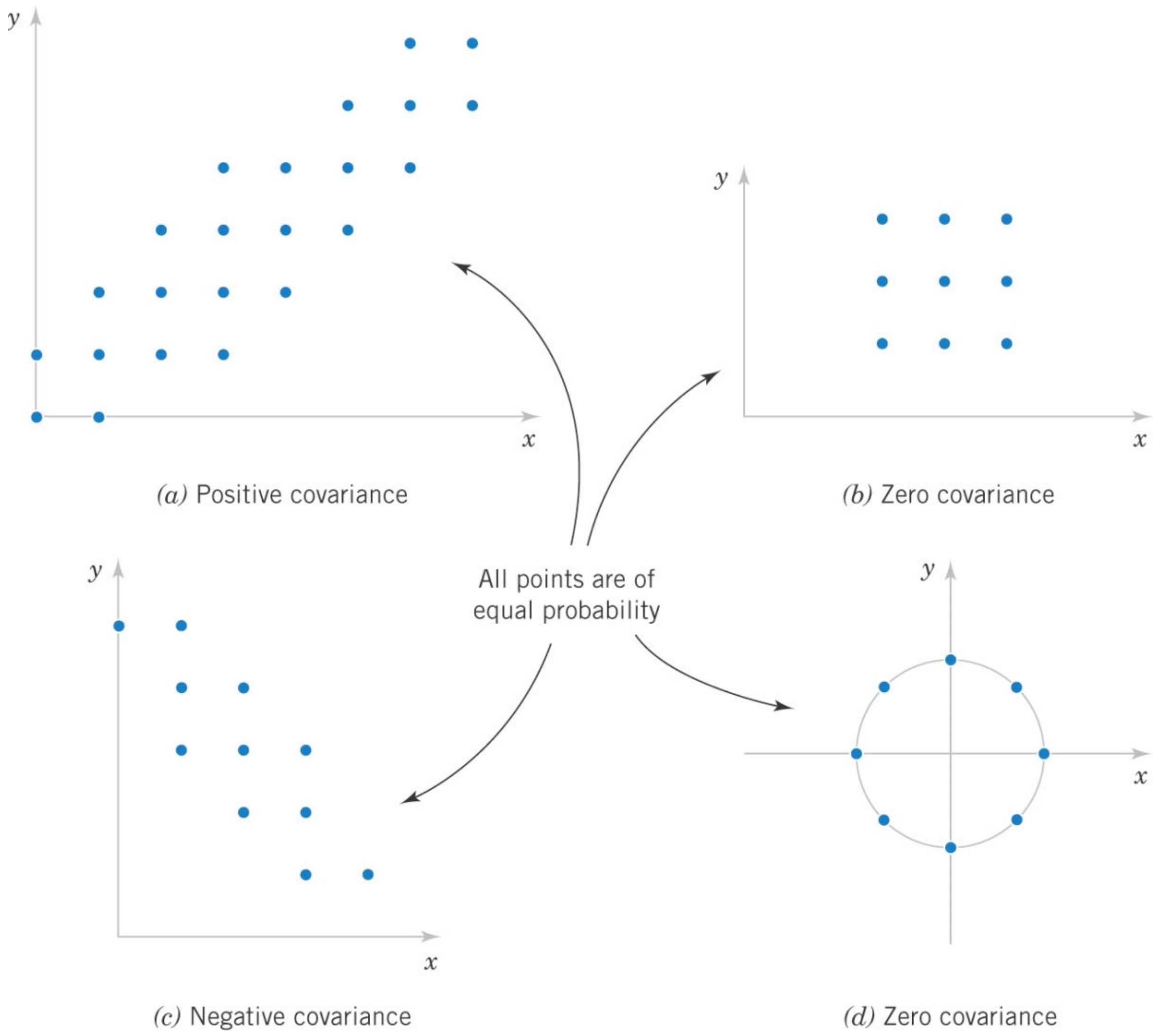
Covariance and PMF tables

y = number of times city name is stated	x = number of bars of signal strength		
	1	2	3
1	0.01	0.02	0.25
2	0.02	0.03	0.20
3	0.02	0.10	0.05
4	0.15	0.10	0.05

The probability distribution of Example 5-1 is shown.

By inspection, note that the larger probabilities occur as *X* and *Y* move in opposite directions. This indicates a negative covariance.

Covariance and Scatter Patterns



(a) Positive covariance

(b) Zero covariance

(c) Negative covariance

(d) Zero covariance

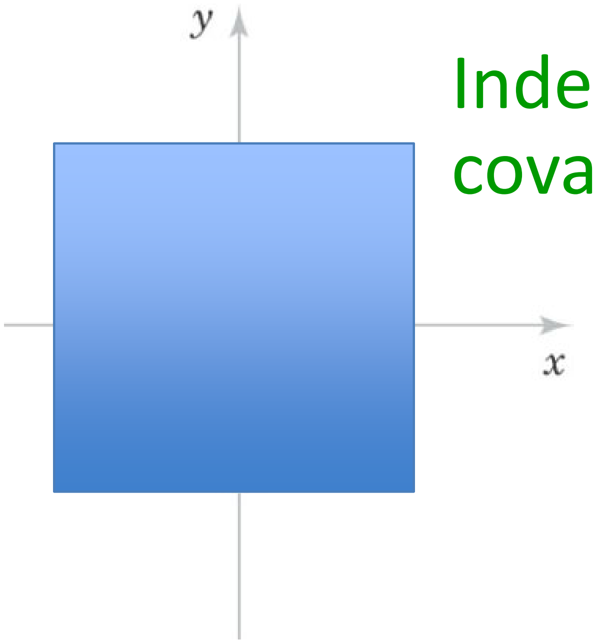
Figure 5-13 Joint probability distributions and the sign of cov(*X*, *Y*). Note that covariance is a measure of linear relationship. Variables with non-zero covariance are correlated.

Independence Implies σ=ρ = 0 but not vice versa

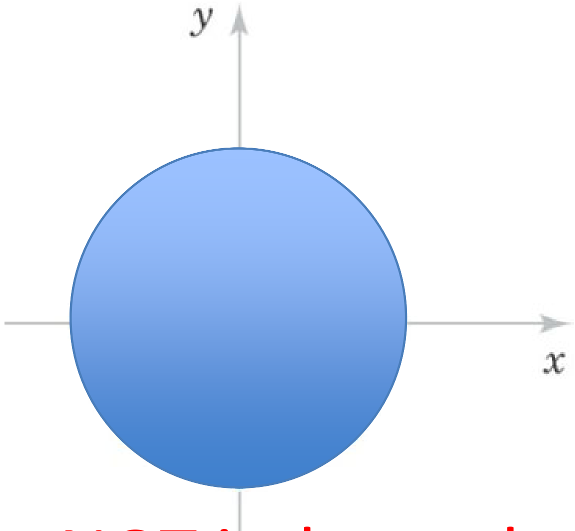
- If *X* and *Y* are independent random variables,

  $\sigma_{XY} = \rho_{XY} = 0$ (5-17)

- $\rho_{XY} = 0$ is necessary, but not a sufficient condition for independence.



Independent
covariance=0



NOT independent
covariance=0

Correlation is "normalized covariance"

- Also called:
  Pearson correlation coefficient

  $\rho_{XY}=\sigma_{XY}/\sigma_X\sigma_Y$
  is the covariance normalized to be $-1 \leq \rho_{XY} \leq 1$

Karl Pearson (1852– 1936)
English mathematician and biostatistician

Prove that ρ_{xy} is in $[-1, 1]$

$$Z_x = \frac{X-\mu_x}{\sigma_x} \quad ; \quad Z_y = \frac{Y-\mu_y}{\sigma_y}$$

$$0 \le E((Z_x - Z_y)^2) = E(Z_x^2) + E(Z_y^2) -$$

$$- 2E(Z_x \cdot Z_y) = 2 - 2\frac{1}{\sigma_x \sigma_y} E((X-\mu_x)(Y-\mu_y)) =$$

$$2 - 2\rho_{xy} \Longrightarrow \boxed{\rho_{xy} \le 1}$$

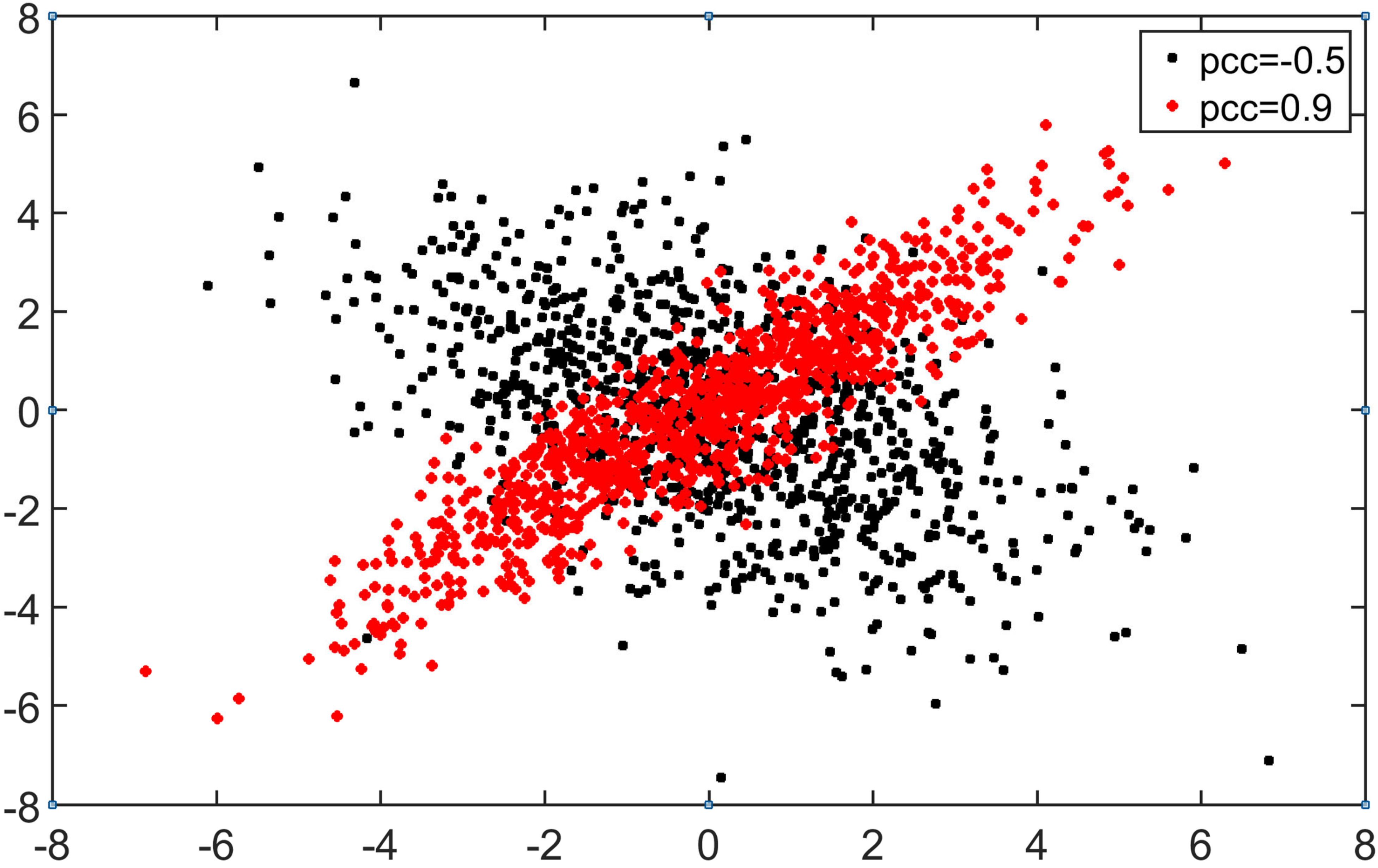
$$0 \le E((Z_x + Z_y)^2) = E(Z_x^2) + E(Z_y^2) +$$

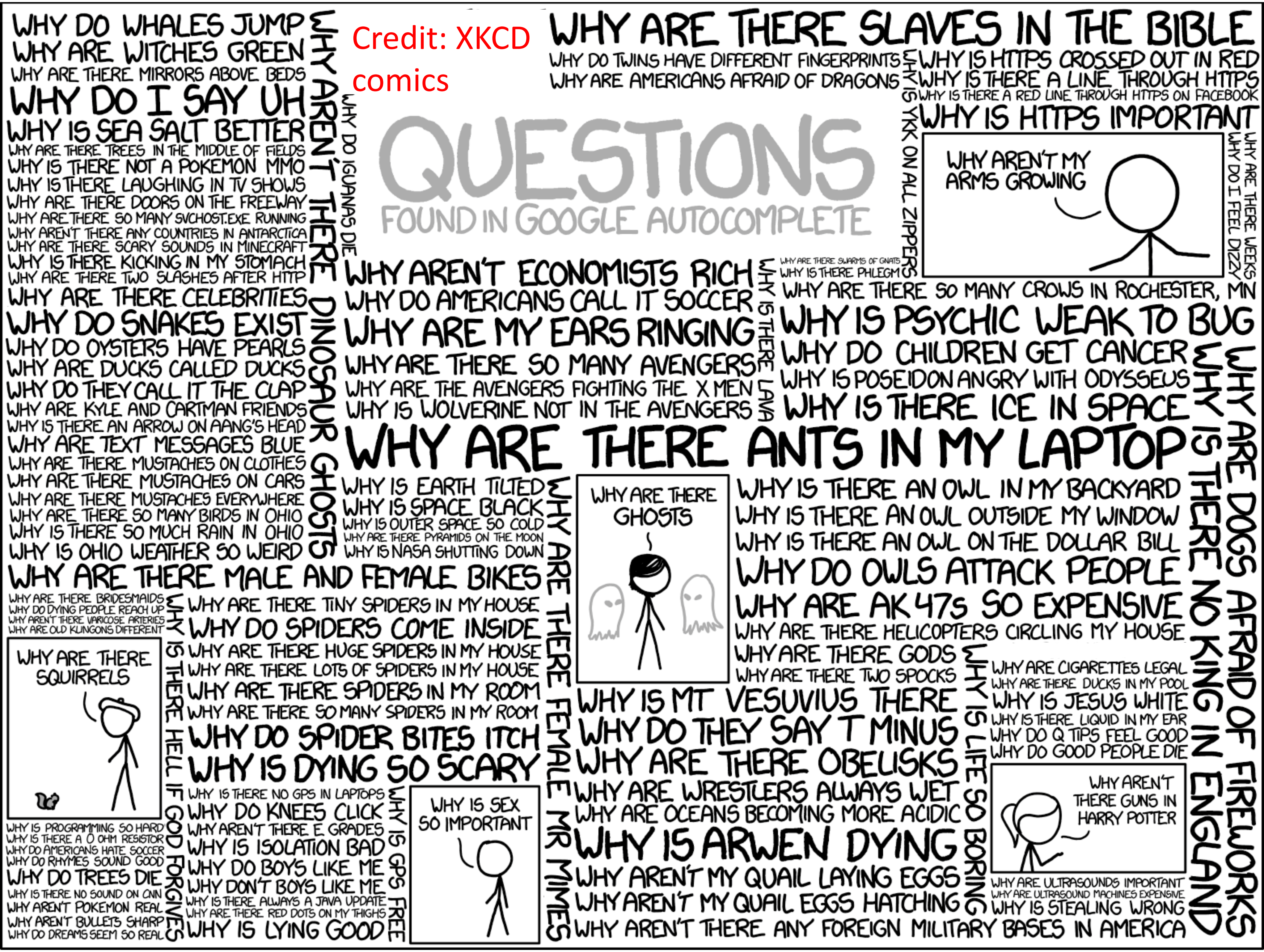
$$+ 2E(Z_x \cdot Z_y) = 2 + 2\rho_{xy} \Longrightarrow$$

$$\Longrightarrow \boxed{\rho_{xy} \ge -1}$$

Spearman rank correlation

- Pearson correlation tests for linear relationship between X and Y
- Unlikely for variables with broad distributions → non-linear effects dominate
- Spearman correlation tests for any monotonic relationship between X and Y
- Calculate ranks (1 to n), $r_X(i)$ and $r_Y(i)$ of variables in both samples. Calculate Pearson correlation between ranks: Spearman(X,Y) = Pearson(r_X, r_Y)
- Ties: convert to fractions, e.g. tie for 6s and 7s place both get 6.5. This can lead to artefacts.
- If lots of ties: use Kendall rank correlation (Kendall tau)

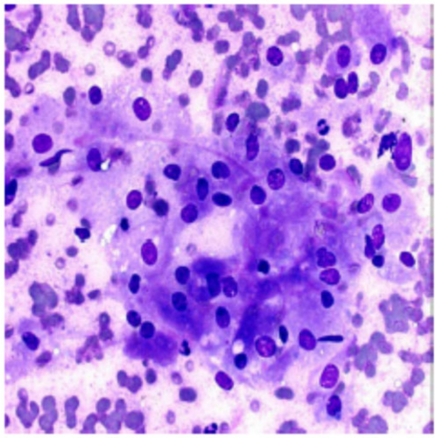

8
6
4
2
0
-2
-4
-6
-8
-8
-6
-4
-2
0
2
4
6
8
pcc=-0.5
pcc=0.9


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

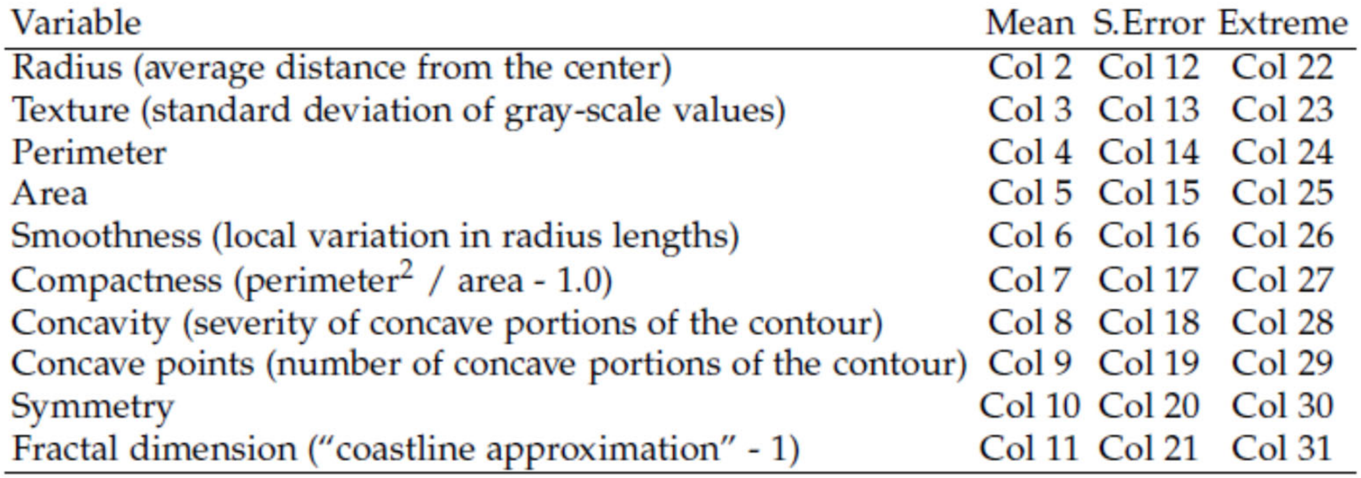
Credit: XKCD comics

Let’s work with real cancer data!

- Data from Wolberg, Street, and Mangasarian (1994)
- Fine-needle aspirates = biopsy for breast cancer
- Black dots – cell nuclei. Irregular shapes/sizes may mean cancer



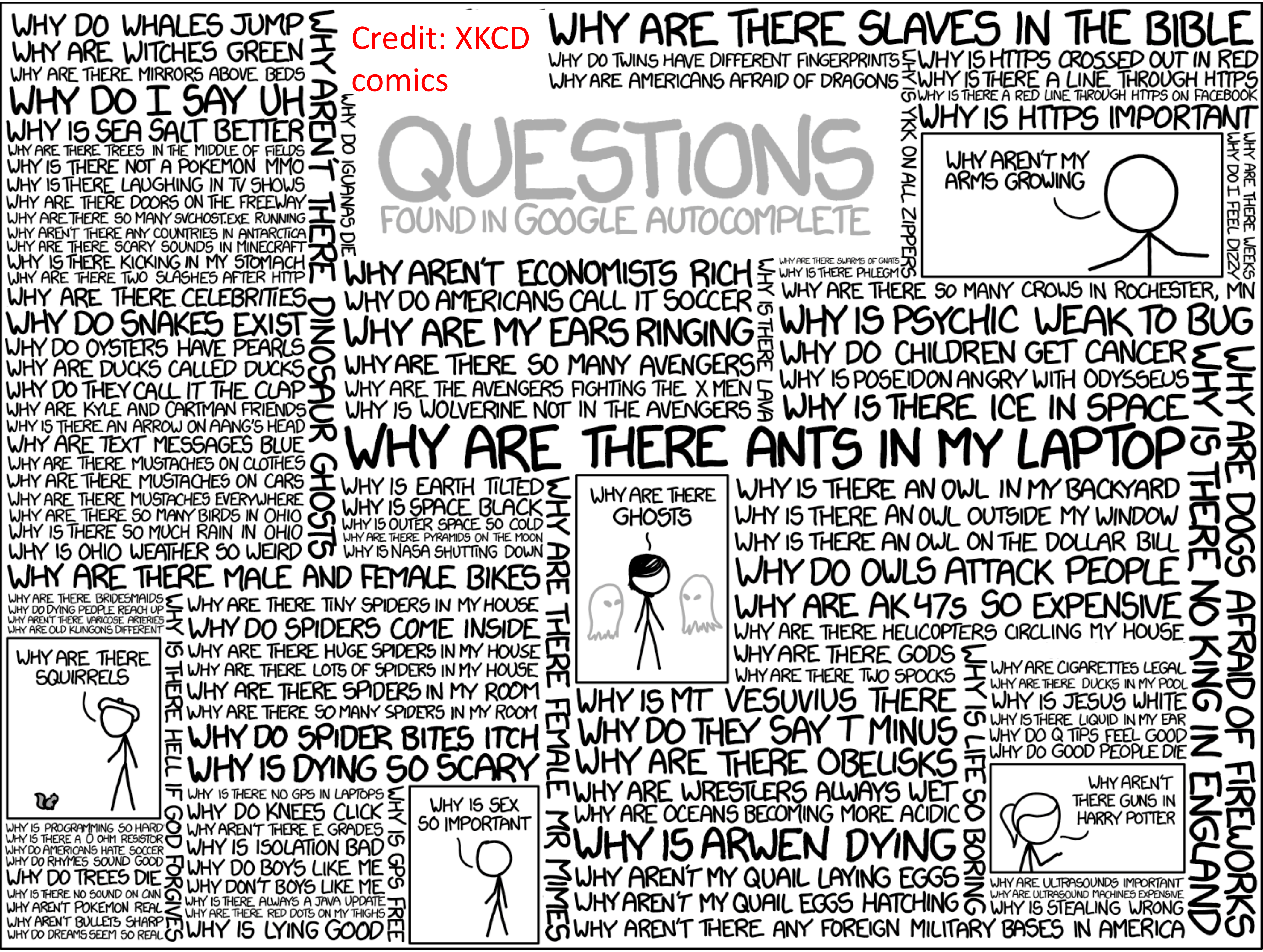
- Statistics of all cells in the image
- 212 cancer patients and 357 healthy individuals (column 1)
- 30 other properties (see table)



Variable	Mean	S.Error	Extreme
Radius (average distance from the center)	Col 2	Col 12	Col 22
Texture (standard deviation of gray-scale values)	Col 3	Col 13	Col 23
Perimeter	Col 4	Col 14	Col 24
Area	Col 5	Col 15	Col 25
Smoothness (local variation in radius lengths)	Col 6	Col 16	Col 26
Compactness (perimeter^2 / area - 1.0)	Col 7	Col 17	Col 27
Concavity (severity of concave portions of the contour)	Col 8	Col 18	Col 28
Concave points (number of concave portions of the contour)	Col 9	Col 19	Col 29
Symmetry	Col 10	Col 20	Col 30
Fractal dimension (“coastline approximation” - 1)	Col 11	Col 21	Col 31

Matlab exercise #2

- Download cancer data in cancer_wdbc.mat
- Data in the table cancerwdbc (569x30). First 357 patients are healthy. The remaining 569-357=212 patients have cancer.
- Make scatter plots of area vs perimeter and texture vs radius.
- Calculate Pearson and Spearman correlations
- Calculate the correlation matrix of all-against-all variables: there are 30*29/2=435 correlations. Hint: corr_mat=corr(cancerwdbc);
- Plot the histogram of these 435 correlation coefficients. Hint: use [i,j,v]=find(corr_mat); then find all i>j and analyze v evaluated on this subset of 435 matrix elements

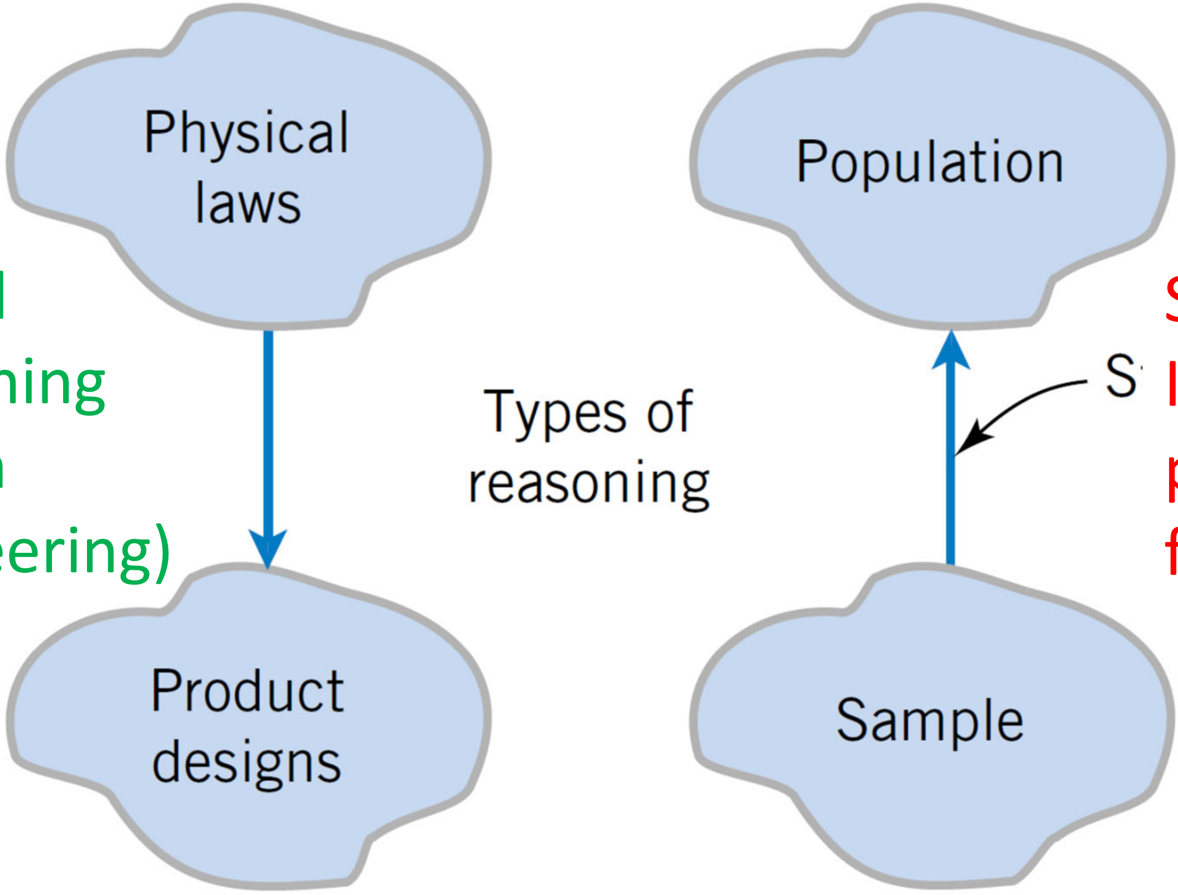

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE SQUIRRELS
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Credit: XKCD comics

Descriptive statistics:

Populations, Samples
Histograms, Quartiles
Sample mean and variance

Two types of reasoning



Logical Reasoning (e.g. in Engineering)

Statistical reasoning: Inference of population properties from a finite sample

Numerical Summaries of Data

- Data are the numerical observations of a phenomenon of interest.
- The totality of all observations is a population.
  - Population can be infinite (e.g. abstract random variables)
  - It can be very large (e.g. 7 billion humans or all patients who have cancer of a given type)
- A (usually small) portion of the population collected for analysis is a random sample.
- We want to use sample to infer facts about populations
- The inference is not perfect but gets better and better as sample size increases.

Some Definitions

- The random variables X_1, X_2,…,X_n are a random sample of size *n* if:
  a) The X_i are independent random variables.
  b) Every X_i has the same probability distribution.

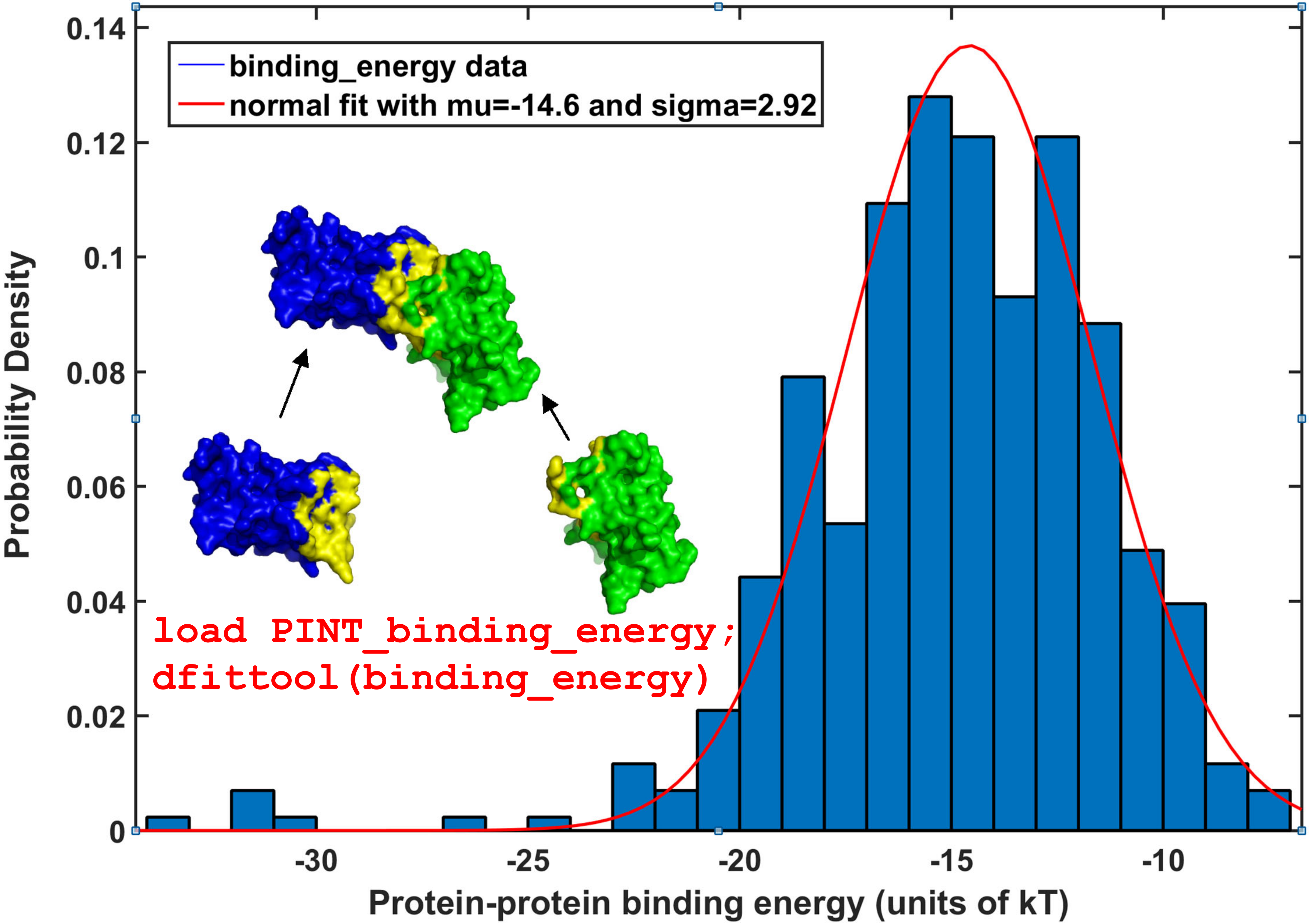
- Such X_1, X_2,…,X_n are also called independent and identically distributed (or i. i. d.) random variables

Ways to describe a sample:

Histogram

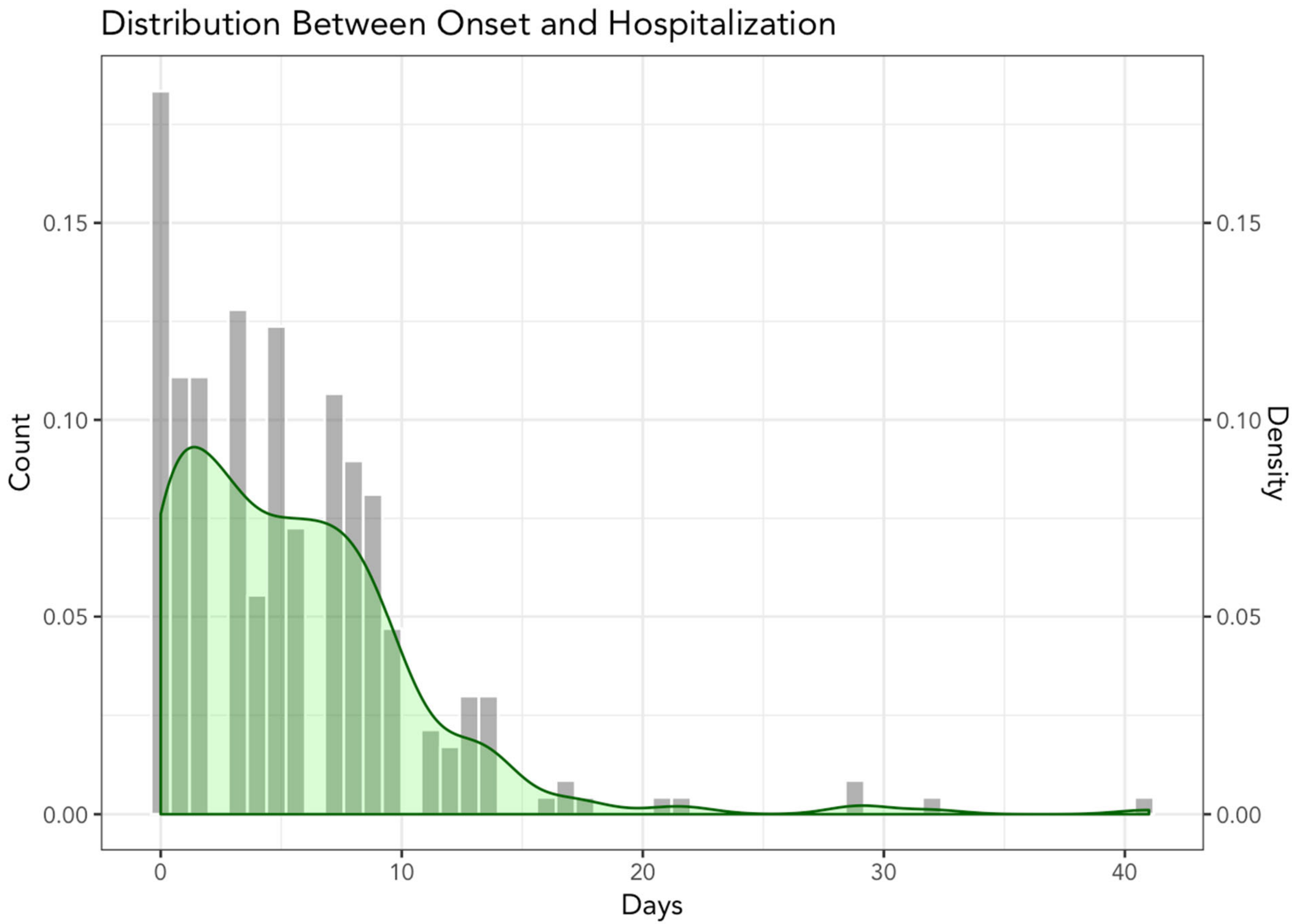
approximates PDF

(or PMF)



```
load PINT_binding_energy;
dfittool(binding_energy)
```

PDF of time between COVID-19 symptoms onset and hospitalization in IL, April 2020

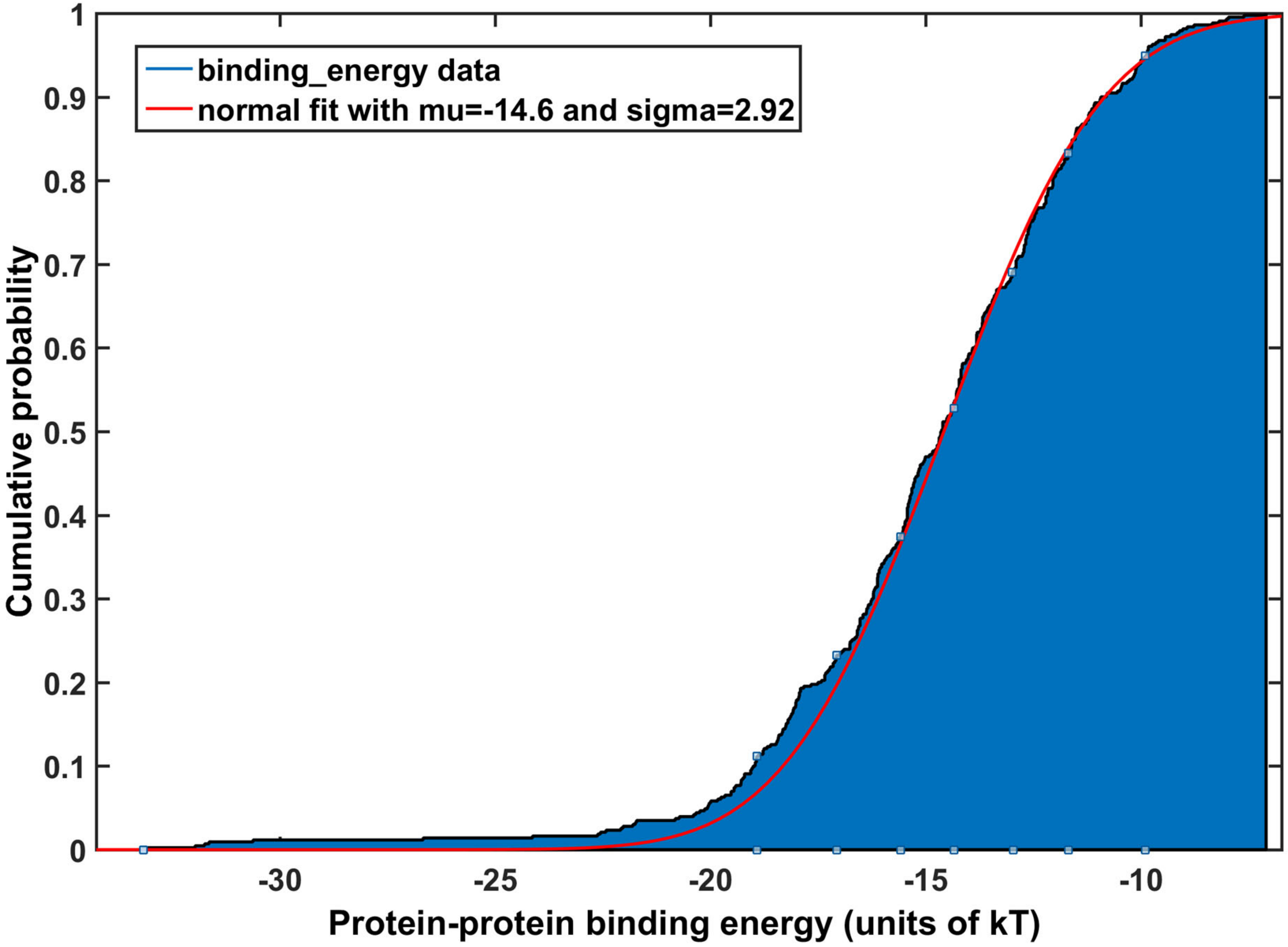


Histograms with Unequal Bin Widths

- If the data is tightly clustered in some regions and scattered in others, it is visually helpful to use narrow bin widths in the clustered region and wide bin widths in the scattered areas.
- To approximate the PDF, the rectangle area, not the height, must be proportional to the bin relative frequency.

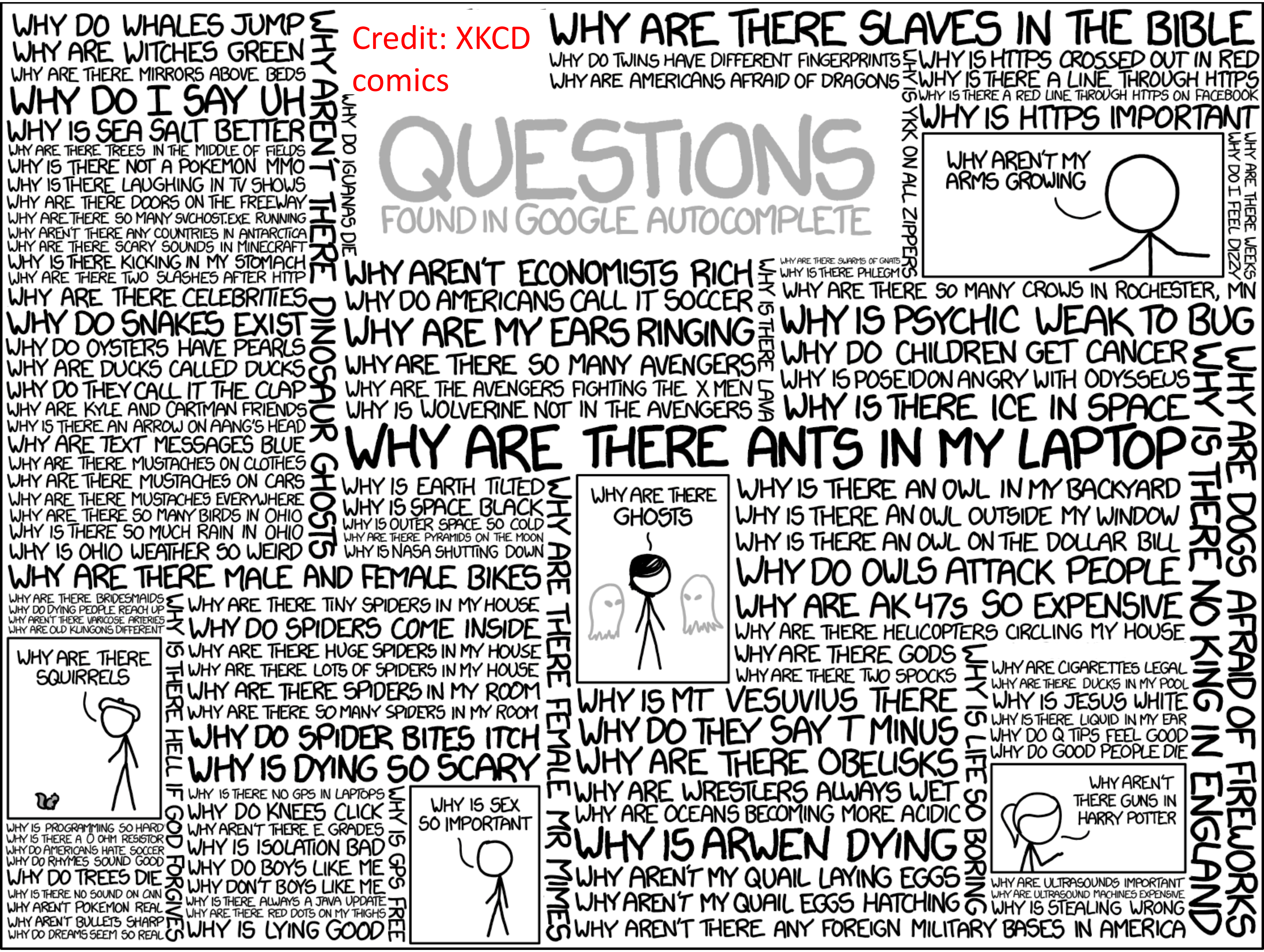
$$\text{Rectangle height} = \frac{\text{bin relative frequency}}{\text{bin width}}$$

Cumulative Frequency Plot



Median, Quartiles, Percentiles

- The median q_2 divides the sample into two equal parts: 50% (n/2) of sample points below q_2 and 50% (n/2) points above q_2
- The three quartiles partition the data into four equally sized counts or segments.
  - 25% of the data is less than q_1.
  - 50% of the data is less than q_2, the median.
  - 75% of the data is less than q_3.
- There are 100 percentiles. n-th percentile p_n is defined so that n% of the data is less than p_n


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Credit: XKCD comics

Box-and-Whisker Plot (or better use Cat-and-Whiskers plots)

- A box plot is a graphical display showing Spread, Outliers, Center, and Shape (SOCS).
- It displays the 5-number summary: *min*, q_1, *median*, q_3, and *max*.

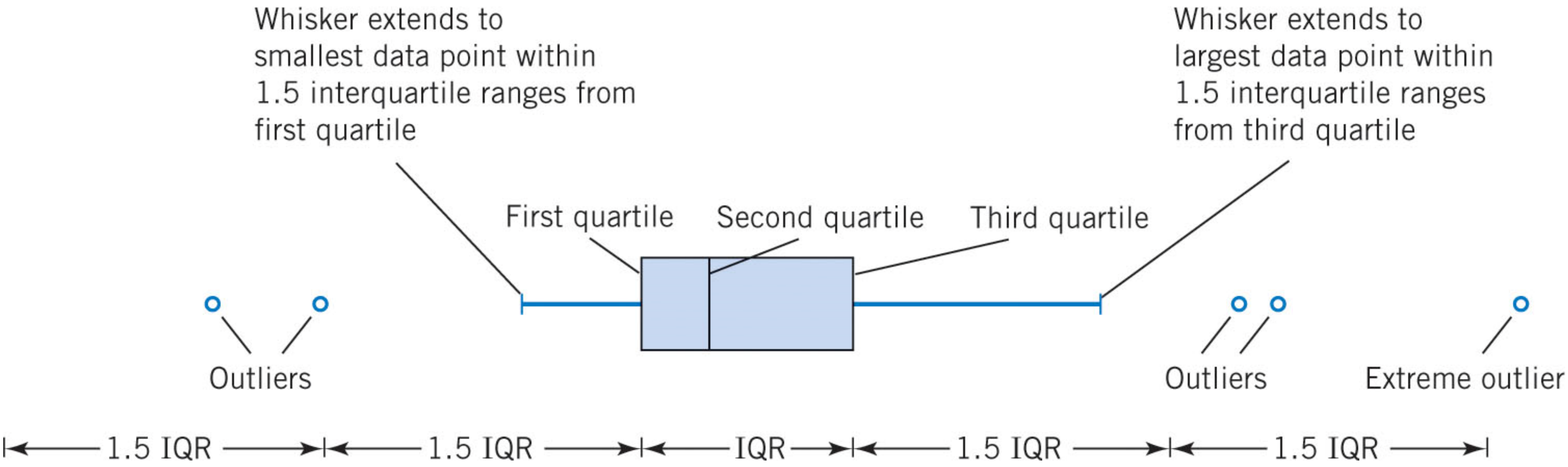
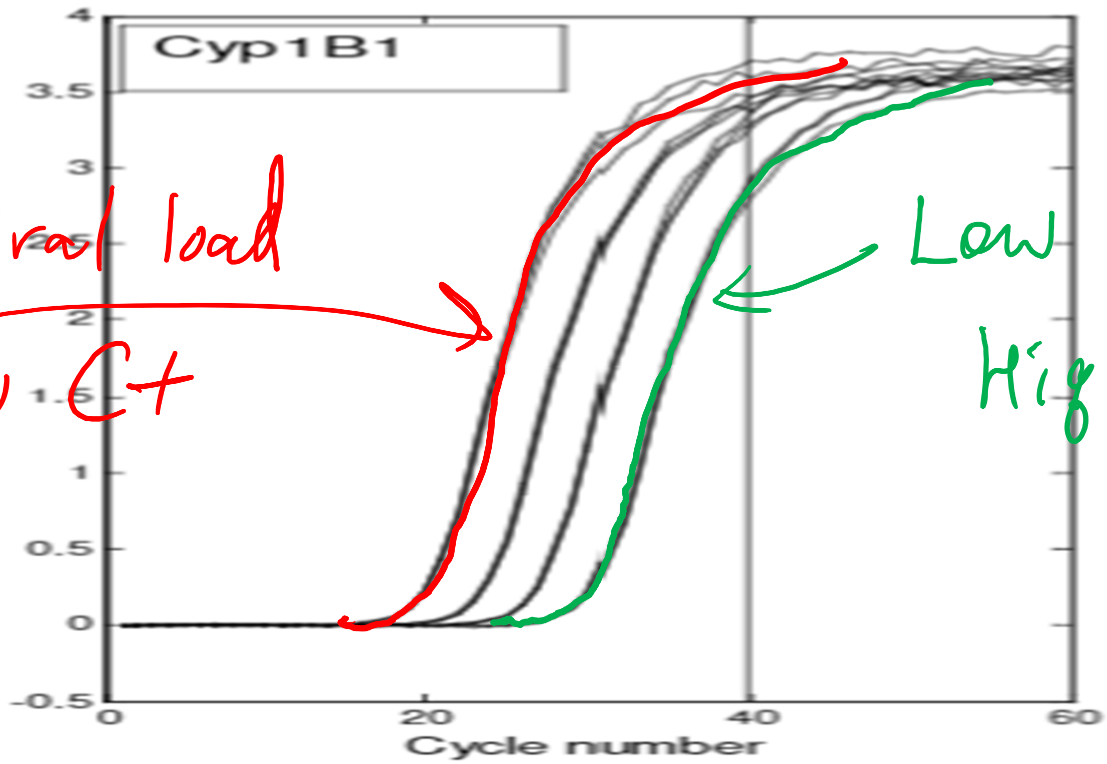

Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

Figure 6-13 Description of a box plot.

Reminder

What is the Cycle threshold (Ct) value of a PCR test?

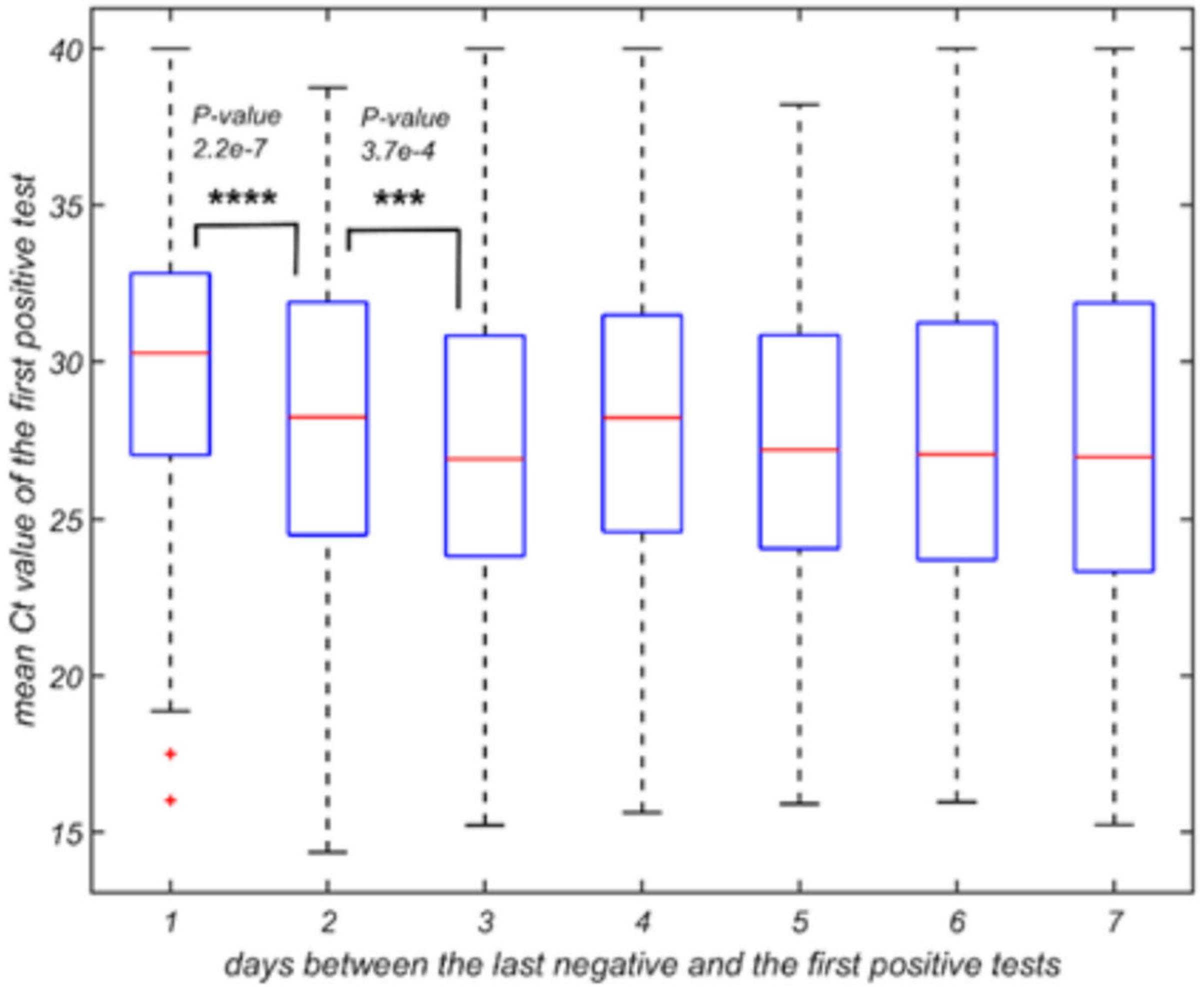
Ct = const – log2(viral DNA concentration)



High viral load = Low Ct

Low viral load = High Ct

Bar plot based on COVID-19 tests at UIUC



Mitigation of SARS-CoV-2 Transmission at a Large Public University
Diana Rose E. Ranoa, et al. , medRxiv 2021 https://doi.org/10.1101/2021.08.03.21261548

Midterm will be held
here in class
this Tuesday 11/07
during regular class hours
12pm-1:50pm

Midterm Info

- Closed book exam; no books, notes, laptops, phones…

- Calculators (not on smartphones) can be used

- You can prepare one 2-sided cheat sheet

- The following two printouts will be provided

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

	Probability			Section
Name	Distribution	Mean	Variance	in Book
Discrete				
Uniform	$\frac{1}{n}, a \le b$	$\frac{(b+a)}{2}$	$\frac{(b-a+1)^2-1}{12}$	3-5
Binomial	$\binom{n}{x} p^x (1-p)^{n-x},$ $x = 0, 1, \ldots, n, 0 \le p \le 1$	np	$np(1-p)$	3-6
Geometric	$(1-p)^{x-1}p,$ $x = 1, 2, \ldots, 0 \le p \le 1$	$1/p$	$(1-p)/p^2$	3-7.1
Negative binomial	$\binom{x-1}{r-1}(1-p)^{x-r}p^r$ $x = r, r+1, r+2, \ldots, 0 \le p \le 1$	r/p	$r(1-p)/p^2$	3-7.2
Hypergeometric	$\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$ $x = \max(0, n-N+K), 1, \ldots$ $\min(K, n), K \le N, n \le N$	$np,$ where $p = \frac{K}{N}$	$np(1-p)\left(\frac{N-n}{N-1}\right)$	3-8
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, 0 < \lambda$	λ	λ	3-9
Continuous				
Uniform	$\frac{1}{b-a}, a \le x \le b$	$\frac{(b+a)}{2}$	$\frac{(b-a)^2}{12}$	4-5
Normal	$\frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$ $-\infty < x < \infty, -\infty < \mu < \infty, 0 < \sigma$	μ	σ^2	4-6
Exponential	$\lambda e^{-\lambda x}, 0 \le x, 0 < \lambda$	$1/\lambda$	$1/\lambda^2$	4-8
Erlang	$\frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!}, 0 < x, r = 1, 2, \ldots$	r/λ	r/λ^2	4-9.1
Gamma	$\frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}, 0 < x, 0 < r, 0 < \lambda$	r/λ	r/λ^2	4-9.2

What is included in the midterm?

- Probability of events (set operations), Multiplication rules. Combinatorics
- Bayes Theorem
- Discrete Random Variables
- Continuous Random Variables
- Other topics covered
  (see HW1-HW2 for inspiration)
- No joint probabilities, correlation and covariation
- No Matlab exercises (since no computers)

Probability Multiplication Rules
Combinatorics

Mr. Jones has 6 different books that he is going to put on his bookshelf. Of these, 3 are chemistry books, 2 are physics books, and 1 is a mathematics book. Jones wants to arrange his books so that two conditions are met:

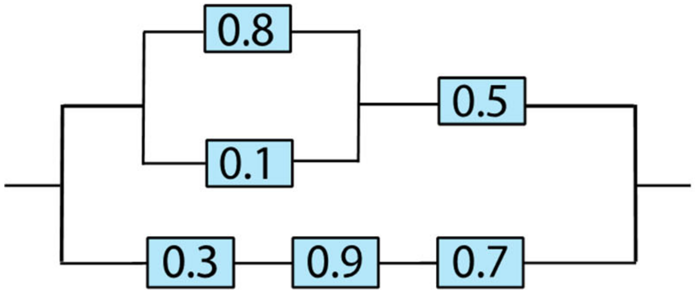
(1) all the books dealing with the same subject are together on the shelf

AND

(2) all chemistry books are on the leftmost side.

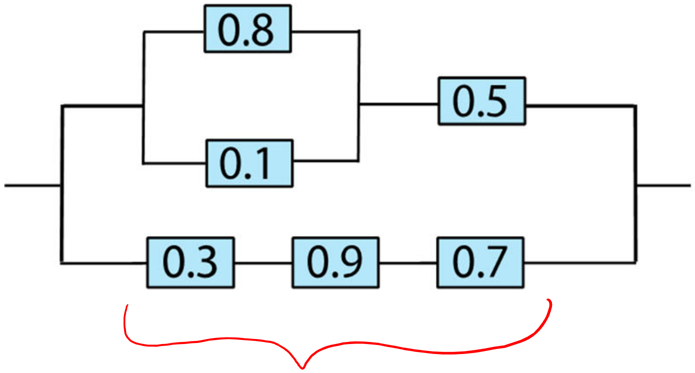
How many such different arrangements are possible?

4. **(4 points)** The following circuit operates if and only if there is a path of functional devices from left to right. The probability that each device functions is as shown. Assume that the probability that a device is functional does not depend on whether or not other devices are functional. What is the probability that the circuit operates?



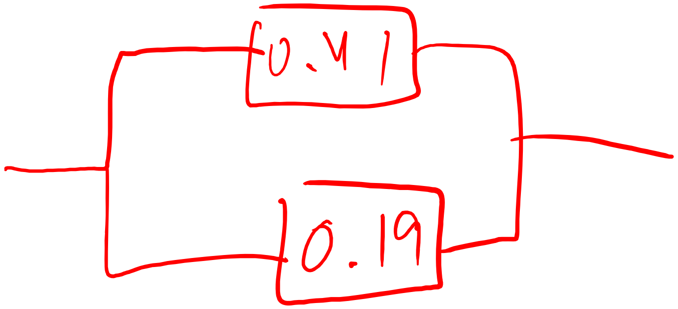
4. (4 points) The following circuit operates if and only if there is a path of functional devices from left to right. The probability that each device functions is as shown. Assume that the probability that a device is functional does not depend on whether or not other devices are functional. What is the probability that the circuit operates?

$1-(1-0.8)\cdot(1-0.1)=0.82$



$0.3\cdot 0.9\cdot 0.7=0.19$

$0.82\times 0.5=0.41$



$1-(1-0.41)\cdot(1-0.19)=0.52$

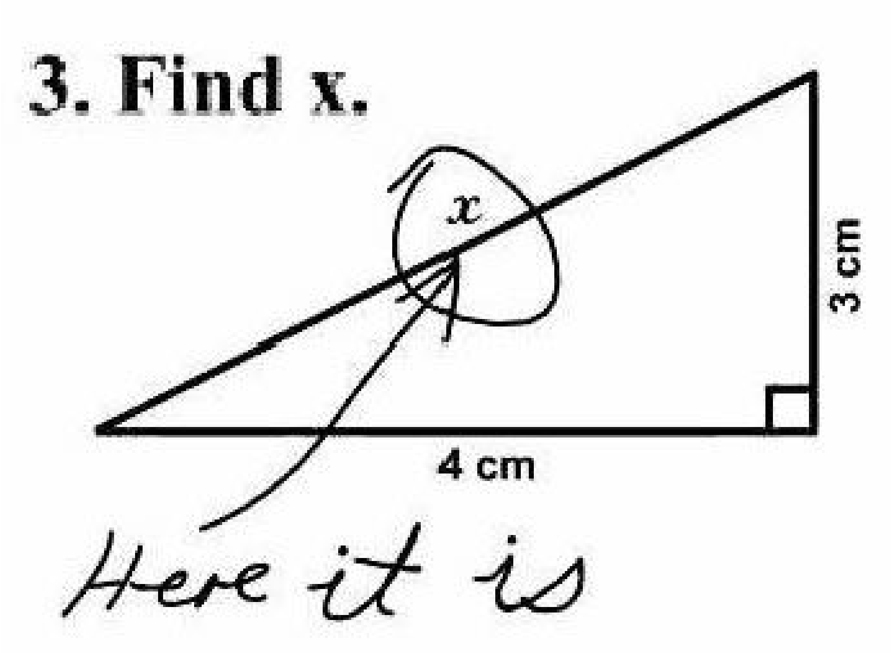
Bayes theorem

(10 points) Suppose that a bag contains ten coins, three of which are fair, while the remaining seven are biased: they have probability of 0.6 of heads when flipped. A coin was taken at random from the bag and flipped five times. All five flips gave heads. What's the probability that this coin is fair?

Discrete Probability Distributions

What is X in this problem?

- What is the random variable: Look for keywords:
  - – Find the probability that….
  - – What is the mean (or variance) of…
- What are parameters? Look for keywords:
  - – Given that…
  - – Assuming that…



Guide to probability distributions

- Binomial: # of samples, n, is fixed, # of successes, x, is variable

$$P(X=x)=\frac{n!}{x!(n-x)!}p^{x}(1-p)^{n-x}$$

- Geometric: # of samples, x is variable. # of successes 1 is fixed. Success comes in the end

$$P(X=x)=(1-p)^{x-1}\cdot p$$

- Negative binomial: # of samples, x is variable. # of successes, r, is fixed rth success in ~~the~~ end

$$P(X=x)=\frac{(x-1)!}{(r-1)!(x-r)!}p^{r}(1-p)^{x-r}$$

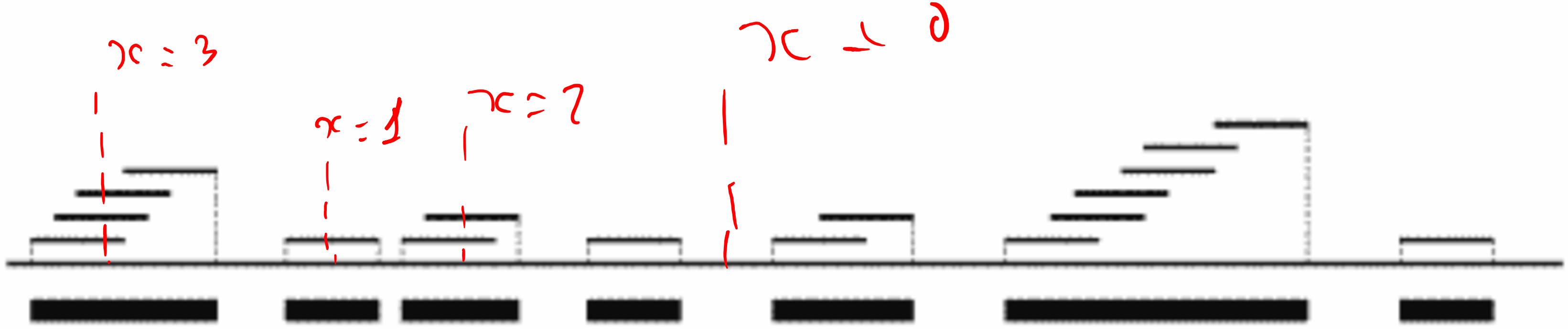
Poisson distribution in genomics

- *G - genome length (in bp)*
- *L - short read average length*
- *N – number of short read sequenced*
- *λ – sequencing redundancy = LN/G*
- *x- number of short reads covering a given site on the genome*

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Ewens, Grant, Chapter 5.1

Poisson as a limit of Binomial. For a given site on the genome for each short read Prob(site covered): p=L/G is very small. Number of attempts (short reads): N is very large. Their product (sequencing redundancy): λ = NL/G is O(1).

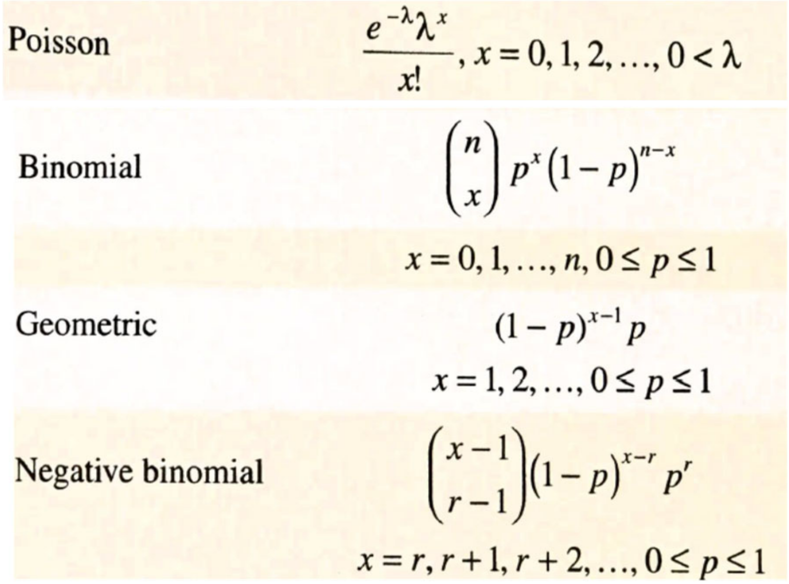


Probability that a base pair in the genome is not covered by any short reads is 0.1

One randomly selects base pairs until exactly 5 uncovered base pairs are found.

Which discrete probability distribution describes the number of attempts?

A. Poisson
B. Binomial
C. Geometric
D. Negative Binomial
E. I have no idea



Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, 0 < \lambda$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \ldots, n, 0 \le p \le 1$
Geometric	$(1-p)^{x-1} p$ $x = 1, 2, \ldots, 0 \le p \le 1$
Negative binomial	$\binom{x-1}{r-1} (1-p)^{x-r} p^r$ $x = r, r+1, r+2, \ldots, 0 \le p \le 1$

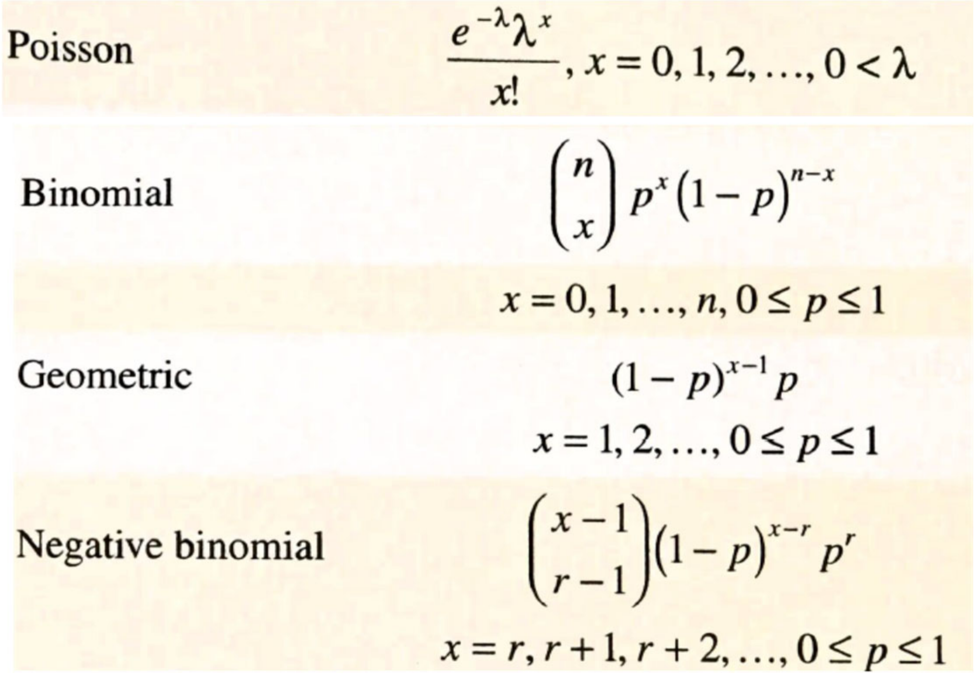
Get your i-clickers

Probability that a base pair in the genome is not covered by any short reads is 0.1

One randomly selects base pairs until exactly 5 uncovered base pairs are found.

What are the values of p, r?

A. p=0.5, r=5
B. p=0.1, r=0.5
C. p=0.1, r=5
D. p=0.5, r=0.1
E. I have no idea



Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, 0 < \lambda$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \ldots, n, 0 \le p \le 1$
Geometric	$(1-p)^{x-1} p$ $x = 1, 2, \ldots, 0 \le p \le 1$
Negative binomial	$\binom{x-1}{r-1} (1-p)^{x-r} p^r$ $x = r, r+1, r+2, \ldots, 0 \le p \le 1$

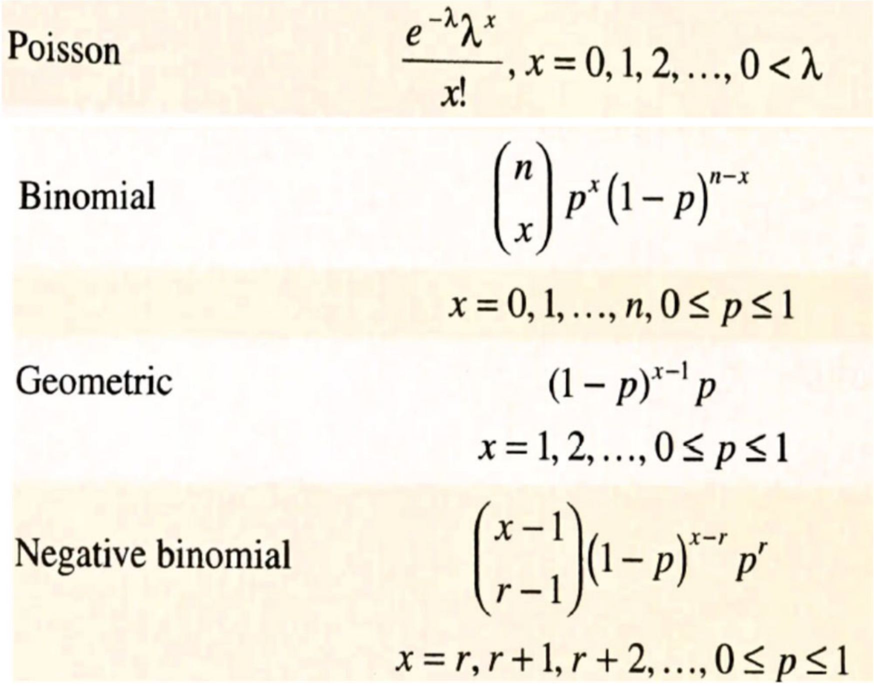
Get your i-clickers

Cancer happens when the gene p53 mutates.
Probability of p53 to mutate per year is 5%.

How many years before a patient gets disease?

Which discrete probability distribution would you use to answer?

A. Poisson
B. Binomial
C. Geometric
D. Negative Binomial
E. I have no idea



Poisson	$\frac{e^{-\lambda}\lambda^{x}}{x!}, x=0,1,2,\ldots,0<\lambda$
Binomial	$\binom{n}{x}p^{x}(1-p)^{n-x}$ $x=0,1,\ldots,n,0\le p\le 1$
Geometric	$(1-p)^{x-1}p$ $x=1,2,\ldots,0\le p\le 1$
Negative binomial	$\binom{x-1}{r-1}(1-p)^{x-r}p^{r}$ $x=r,r+1,r+2,\ldots,0\le p\le 1$

Get your i-clickers

Continuous Probability Distributions

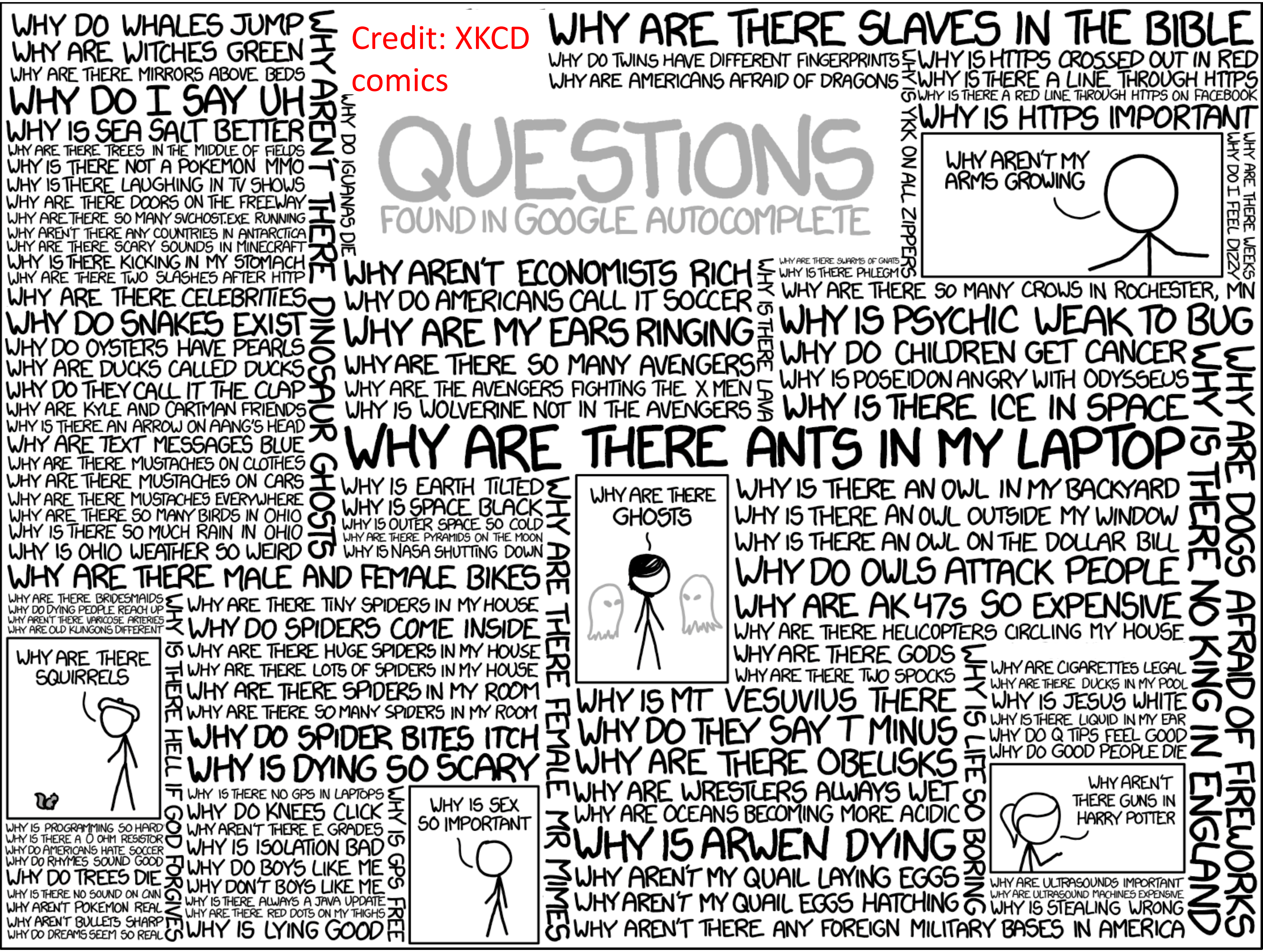
1. **(8 points)** The expression level of a *TP53* tumor suppressor gene in a randomly selected cell is normally distributed with mean μ= 20, and standard deviation σ = 8.

   (A)(4 points) What is the probability that the expression level in a given cell will be between 24 and 16?

   (B)(4 points) How many cells does one have to sample (on average) until there will be exactly 2 cells with such "close to average" *TP53* expression?

I can show you how to solve any HW1-HW2 problem.

Which one do you choose?


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS

Credit: XKCD comics

Box-and-Whisker Plot

- A box plot is a graphical display showing Spread, Outliers, Center, and Shape (SOCS).
- It displays the 5-number summary: *min*, q_1, *median*, q_3, and *max*.

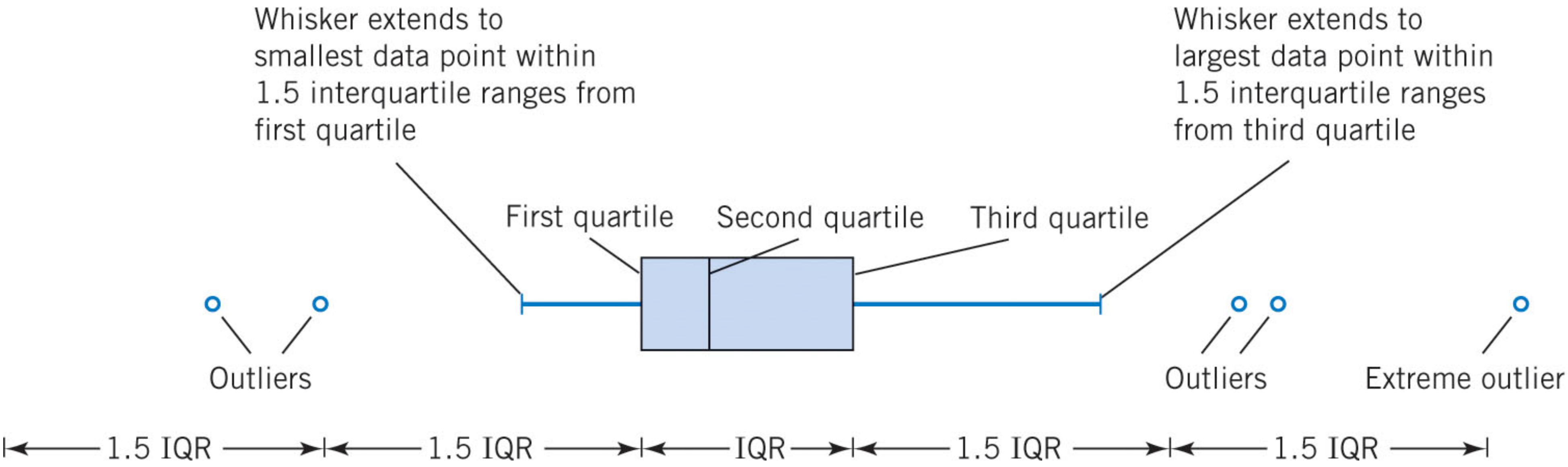
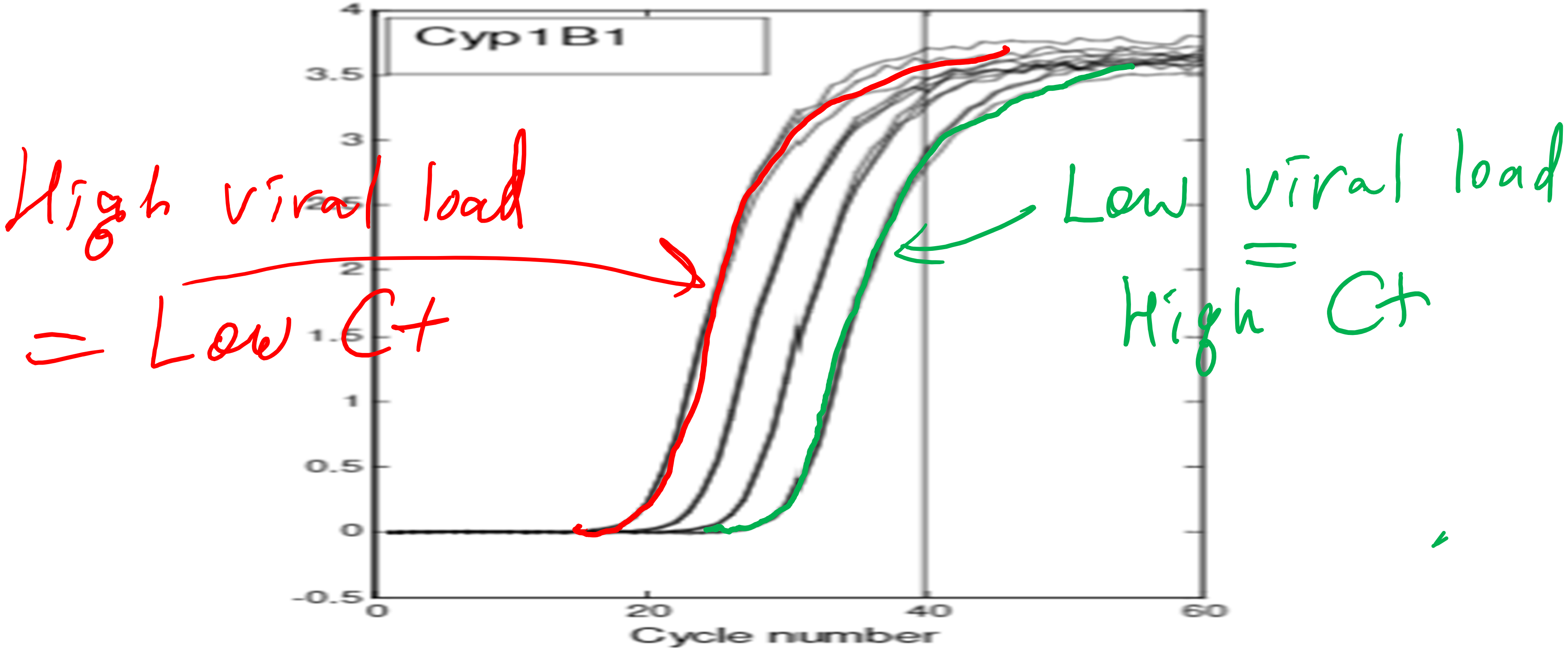

Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

Figure 6-13 Description of a box plot.

Reminder

What is the Cycle threshold (Ct) value of a PCR test?

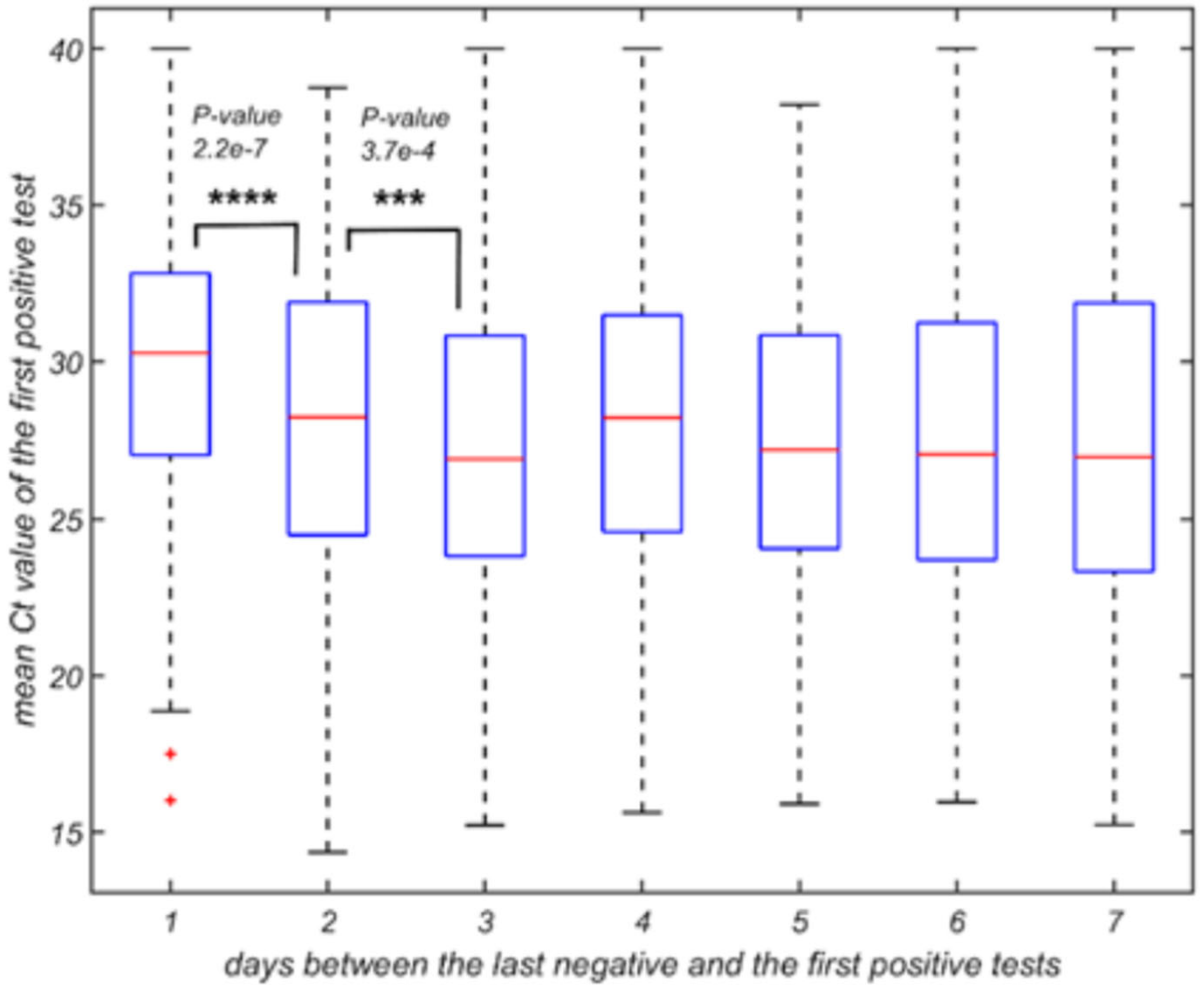
Ct = const – log2(viral DNA concentration)



High viral load = Low Ct

Low viral load = High Ct

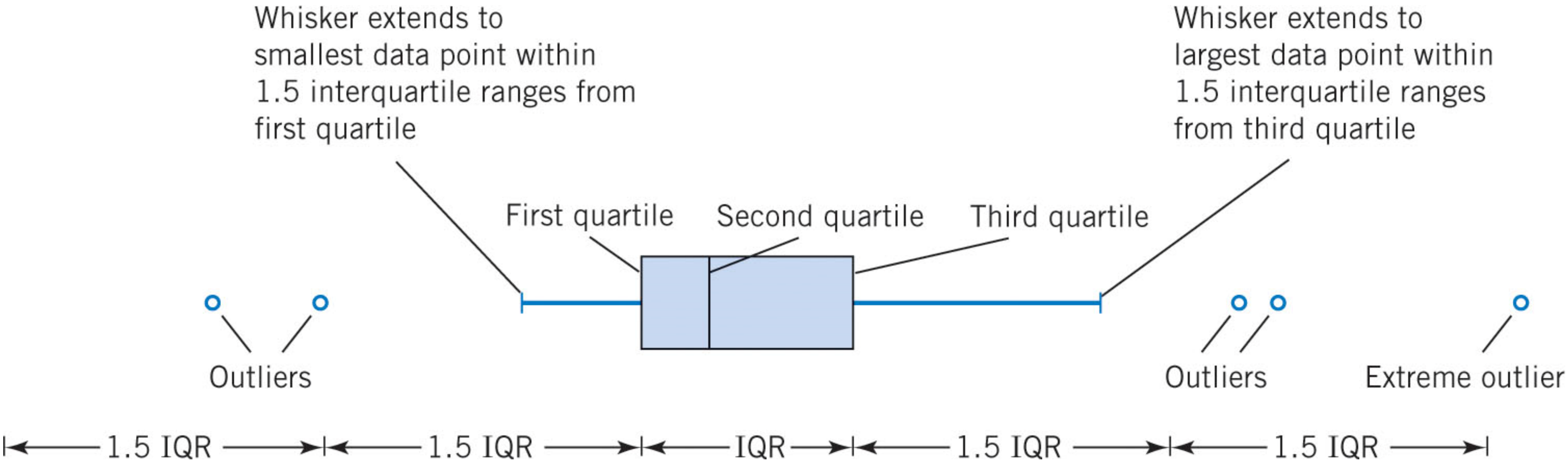
Bar plot based on COVID-19 tests at UIUC

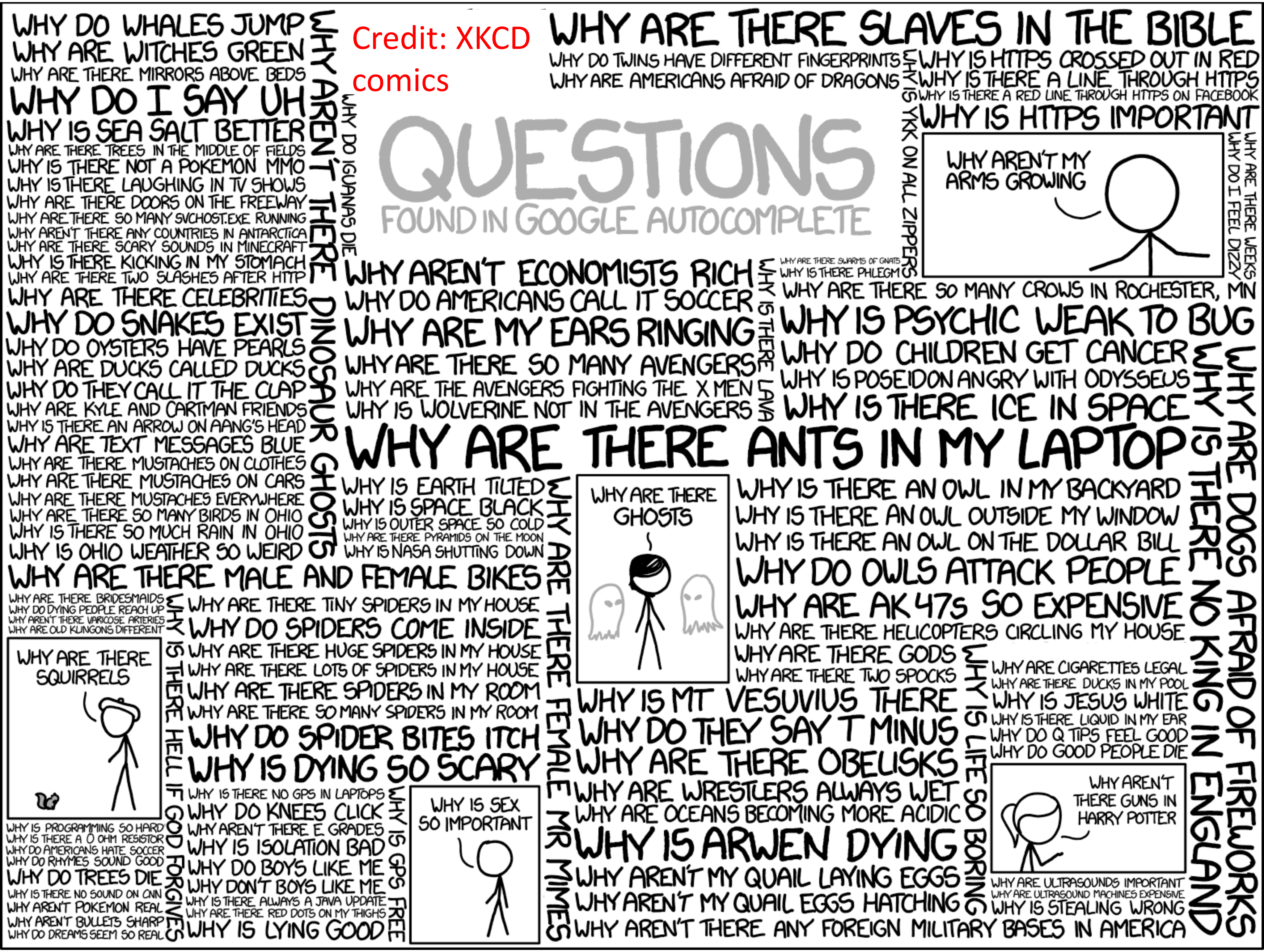


Ranoa, D. R. E. et al. Mitigation of SARS-CoV-2 transmission at a large public university. Nat Commun 13, 3207 (2022)

Matlab exercise #2:

- Generate a sample with n= 1000 following standard normal distribution
- Calculate median, first, and third quartiles
- Calculate IQR and find ranges shown below
- Find and count left and right outliers
- Do not use built-in Matlab functions for this!
- Make box and whisker plot: use boxplot


Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Credit: XKCD comics

Descriptive statistics:

Sample mean and its variance

Standard error vs Standard deviation

Some Definitions

- The random variables $X_1, X_2, \ldots, X_n$ are a random sample of size n if:

  a) The X_i are independent random variables.

  b) Every X_i has the same probability distribution.

  Such $X_1, X_2, \ldots, X_n$ are also called independent and identically distributed (or i. i. d.) random variables

- A **statistic** is any function of the observations in a random sample.
- The probability distribution of a statistic is called a **sampling distribution**.

Statistic #1: Sample Mean

If the values of n observations in a random sample are denoted by $x_1, x_2, \ldots, x_n,$ the sample mean is

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (6\text{–}1)$$

New random variable $\overline{X}$ is a linear combination of *n* independent identically distributed variables $X_1, X_2, \ldots, X_n$

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Mean & Variance of a Linear Function

$Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p$

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \ldots + c_pE(X_p) \qquad (5\text{-}25)$$

$$\mathrm{V}(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) + 2\sum_{i<j}\sum c_ic_j \operatorname{cov}(X_iX_j) \qquad (5\text{-}26)$$

If $X_1, X_2, \ldots, X_p$ are independent, then $\operatorname{cov}(X_iX_j) = 0$,

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) \qquad (5\text{-}27)$$

IMPORTANT:

Sample mean $\overline{x}$ is drawn from a random variable

$$\overline{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$$E(\overline{X}) = \frac{n \cdot E(X_i)}{n} = \frac{n \cdot \mu}{n} = \mu$$

$$V(\overline{X}) = \frac{n \cdot V(X_i)}{n^2} = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\underline{\text{Stand. dev.}\,(\overline{X})} = \underline{\frac{\sigma}{\sqrt{n}}}$$

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample of size n is taken from a population with mean μ and finite variance σ^2, and **any distribution**. If $\bar{X}$ is the sample mean, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (7\text{–}1)$$

for large n, is the standard normal distribution.

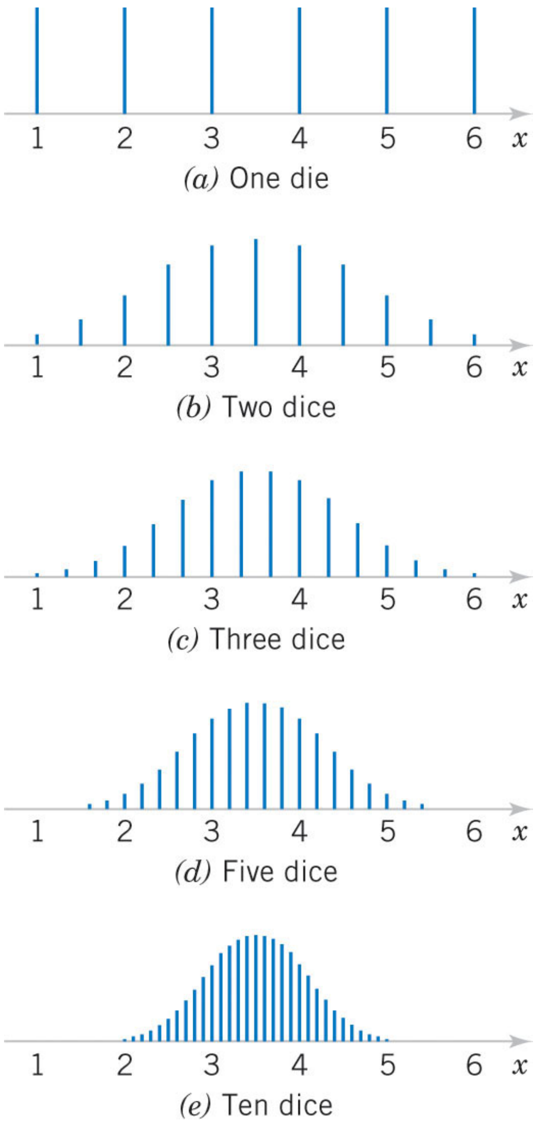
If $X_1, X_2, \ldots, X_n$ are themselves normally distributed – for any n

Test CLT for your own random variable

- Go to: https://onlinestatbook.com/stat_sim/sampling_dist/
- Select "Custom" at the top and use mouse to sketch the PMF of your own random variable
- Select "mean" and n=5 in the third panel
- Choose "Animated" in the second panel and use number_of_experiments=5 to see one sample being generated
- Repeat with number_of _experiments =10,000
- Now select "mean" and n=25 in the fourth panel
- Skewness and Curtosis are measures of how good is the normal (Gaussian) fit (choose "fit normal")

Sampling Distributions of Sample Means

Figure 7-1 Distributions of average scores from throwing dice.
Mean = (6+1)/2=3.5
Sigma^2 = [(6-1+1)^2-1]/12=2.92
Sigma=1.71



(a) One die

(b) Two dice

(c) Three dice

(d) Five dice

(e) Ten dice

Formulas

$$\mu = \frac{b+a}{2} = 3.5$$

$$\sigma_X^2 = \frac{(b-a+1)^2-1}{12} = 35/12$$

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$$

show Matlab

Matlab exercise

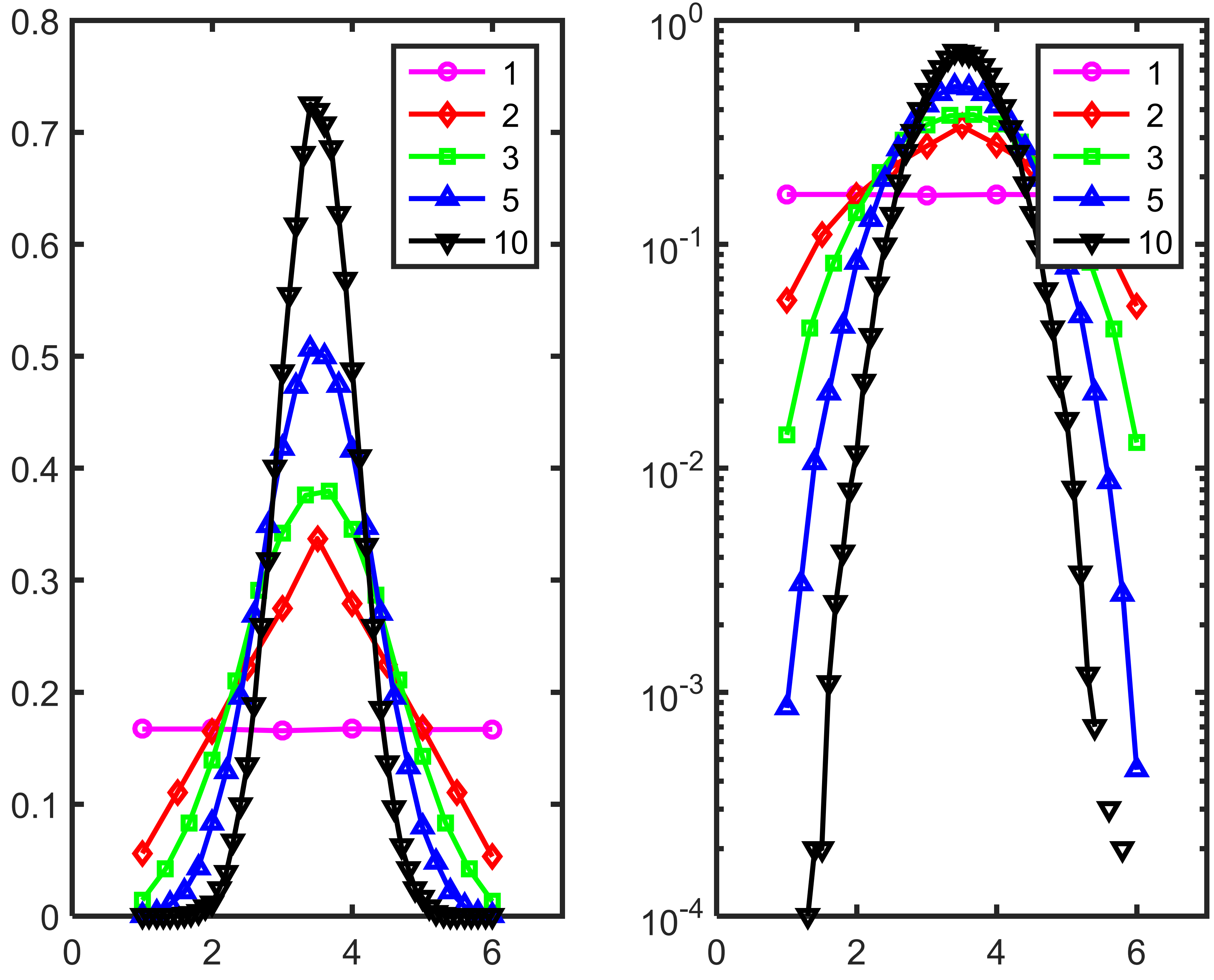
- Do a numerical experiment: generate a sample of size n by rolling n fair dice
- Calculate the sample mean $$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$
- Repeat Stats=100,000 times
- Generate PDFs of sample means for different samples sizes: n=1, n=2, n=3, n=5, and n=10
- Plot them in the same (semi-logarithmic) figure
- **What do you see?**
- Template is at the website: central_limit_theorem_template.m

How did I do it?

- **Stats=100000;**
- **figure;**
- **for n=[1,2,3,5,10];**
- **r_sample=floor(6.*rand(Stats,n))+1;**
- **sample_mean=sum(r_sample,2)./n;**
- **step=1./n;**
- **[a,b1]=hist(sample_mean,1:step:6);**
- **pdf_r1=a./sum(a)./step;**
- **semilogy(b1,pdf_r1,'o-'); hold on;**
- **end;**
- **legend('1','2','3','5','10');**

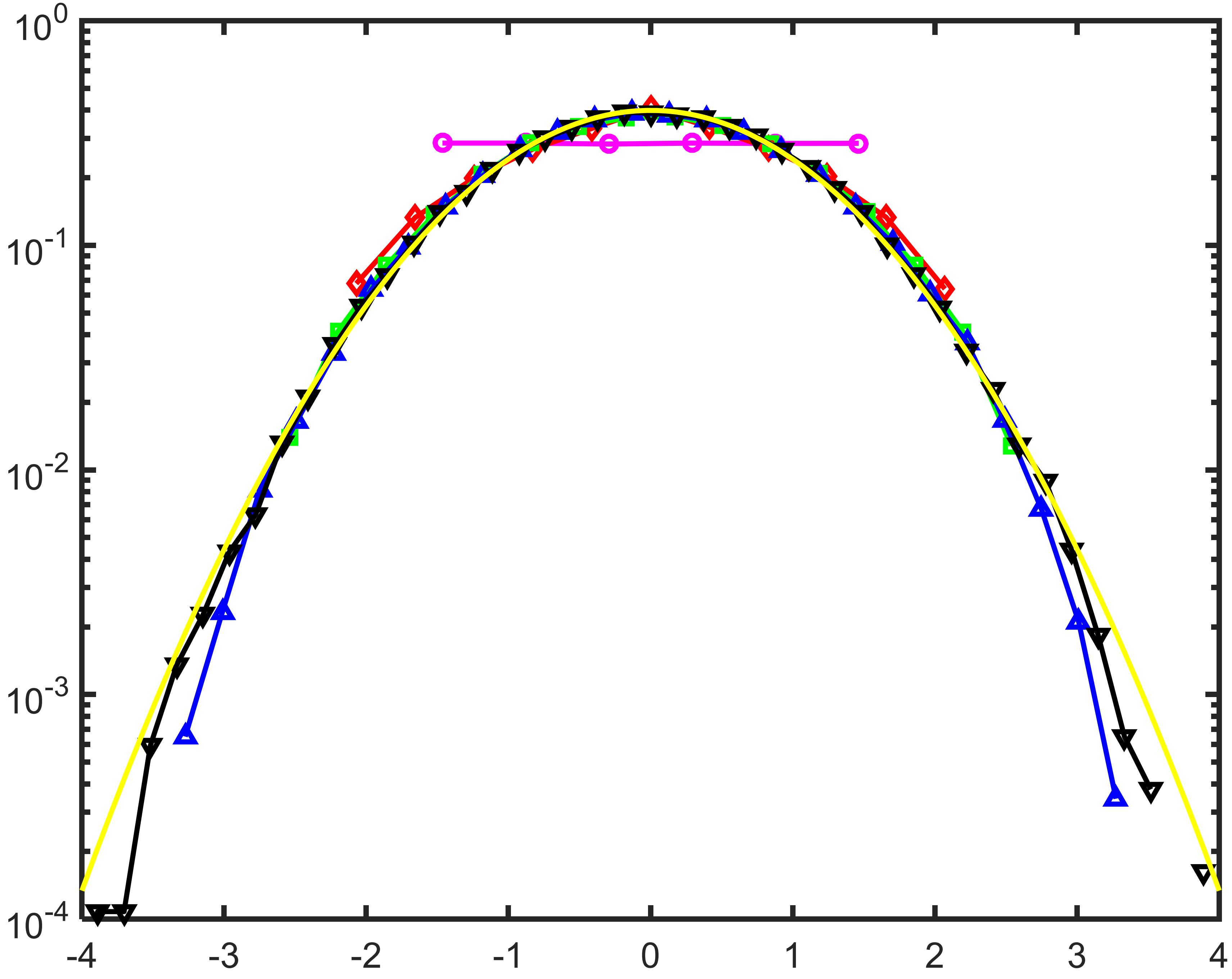
Matlab demonstration

```
Stats=100000; N=10;
r_table=floor(6.*rand(Stats,N))+1;
%%
r1=r_table(:,1);
step=1; [a,b1]=hist(r1,1:step:6);
pdf_r1=a./sum(a)./step;
figure; hold on; subplot(1,2,1); plot(b1,pdf_r1,'mo-'); hold on; axis([0 7 0 0.2]); subplot(1,2,2); semilogy(b1,pdf_r1,'mo-'); hold on; axis([0 7 1e-3 1]);
%%
r2=(r_table(:,1)+r_table(:,2))./2;
step=0.5; [a,b2]=hist(r2,1:step:6); pdf_r2=a./sum(a)./step;
subplot(1,2,1); plot(b2,pdf_r2,'rd-'); axis([0 7 0 0.4]); subplot(1,2,2); semilogy(b2,pdf_r2,'rd-');
%%
r3=(r_table(:,1)+r_table(:,2)+r_table(:,3))./3;
step=1./3; [a,b3]=hist(r3,1:step:6); pdf_r3=a./sum(a)./step;
subplot(1,2,1); plot(b3,pdf_r3,'gs-');  axis([0 7 0 0.4]); subplot(1,2,2); semilogy(b3,pdf_r3,'gs-');
%%
r5=sum(r_table(:,1:5),2)./5;
step=1./5; [a,b5]=hist(r5,1:step:6); pdf_r5=a./sum(a)./step;
subplot(1,2,1); plot(b5,pdf_r5,'b^-');  axis([0 7 0 0.6]); subplot(1,2,2); semilogy(b5,pdf_r5,'b^-'); axis([0 7 1e-4 1]);
%%
r10=sum(r_table(:,1:10),2)./10;
step=1./10; [a,b10]=hist(r10,1:step:6); pdf_r10=a./sum(a)./step;
subplot(1,2,1); plot(b10,pdf_r10,'kv-'); axis([0 7 0 0.8]); legend(num2str([1,2,3,5,10]'));
subplot(1,2,2); semilogy(b10,pdf_r10,'kv-'); legend(num2str([1,2,3,5,10]'));
```

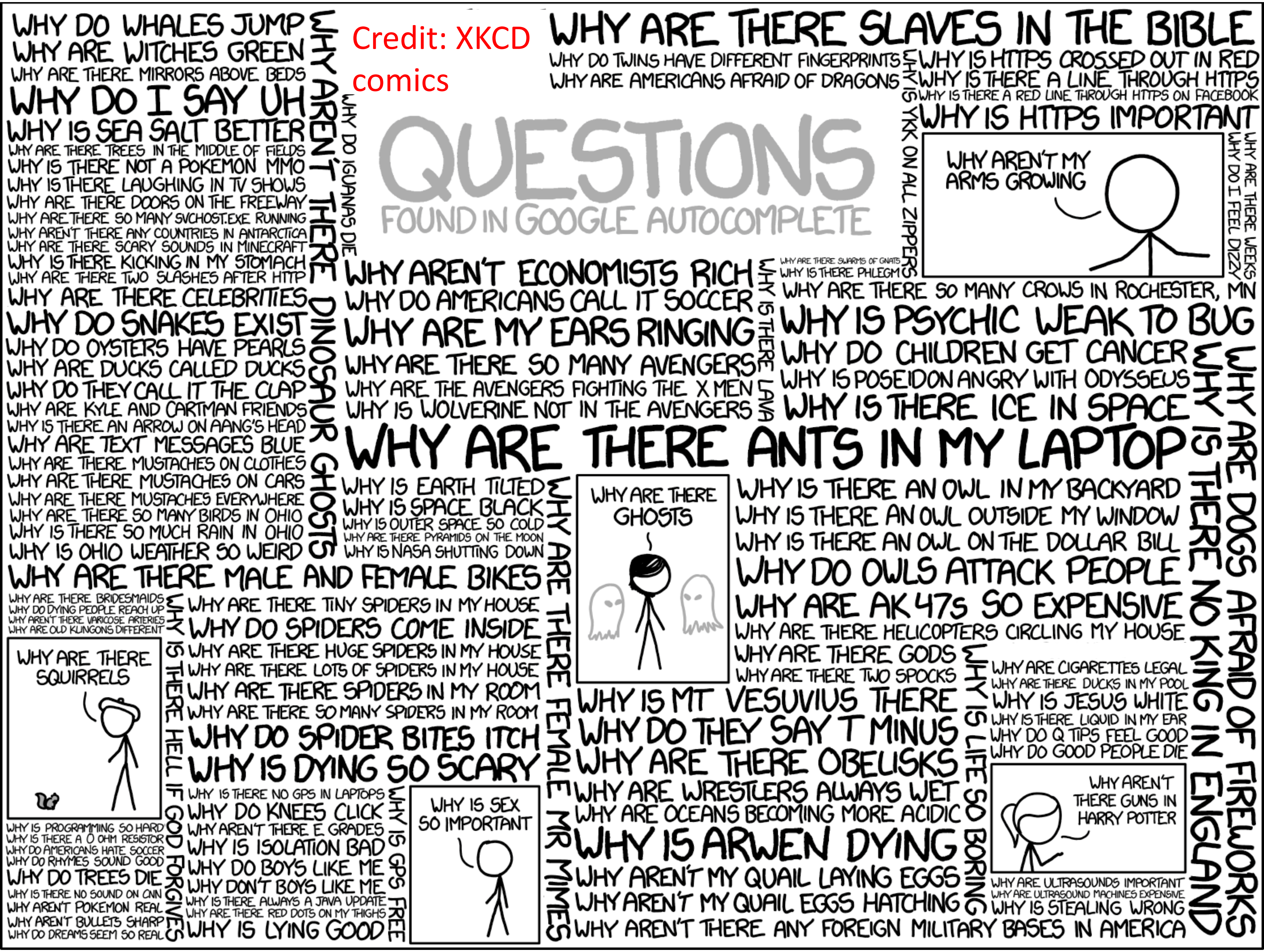

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
1
2
3
5
10
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
2
4
6
1
2
3
5
10

Matlab demonstration; part 2

```
 %%Now plot all of them normalized to 0 and std 1
sigma=sqrt(35/12);
mu=3.5;
figure;
sigma1=sigma;
semilogy((b1-mu)./sigma1,pdf_r1.*sigma1,'mo-');
axis([-4 4 1e-3 1]);
hold on;
%%
sigma2=sigma./sqrt(2);
semilogy((b2-mu)./sigma2,pdf_r2.*sigma2,'rd-');
%%
sigma3=sigma./sqrt(3);
semilogy((b3-mu)./sigma3,pdf_r3.*sigma3,'gs-');
%%
sigma5=sigma./sqrt(5);
semilogy((b5-mu)./sigma5,pdf_r5.*sigma5,'b^-');
axis([-4 4 1e-4 1]);
%%
sigma10=sigma./sqrt(10);
semilogy((b10-mu)./sigma10,pdf_r10.*sigma10,'kv-');
axis([-4 4 1e-4 1]);
%%
%Let's see how well does the Gaussian fits it
x=-4:0.1:4;
semilogy(x,1./sqrt(2*pi)*exp(-x.^2./2),'y-');
```



10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
-4
-3
-2
-1
0
1
2
3
4


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT

Credit: XKCD comics

Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of $n = 25$ resistors will have an average resistance of less than 95 ohms?

Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of *n* = 25 resistors will have an average resistance of less than 95 ohms?

$\mu = 100 \text{ ohms}, \quad \sigma = 10 \text{ ohms}, \quad n = 25$

$$\mu_{\bar{X}} = \mu \,; \; \sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2 \text{ ohms}$$

$$Z_{\bar{X}} = \frac{95 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{95-100}{2} = -2.5$$

$$\text{Prob}\left(\bar{X} < 95\right) = \Phi\left(Z_{\bar{X}}\right) = \Phi(-2.5) =$$

$$= 0.0062$$

Example 7-1: Resistors

An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. What is the approximate probability that a random sample of *n* = 25 resistors will have an average resistance of less than 95 ohms?

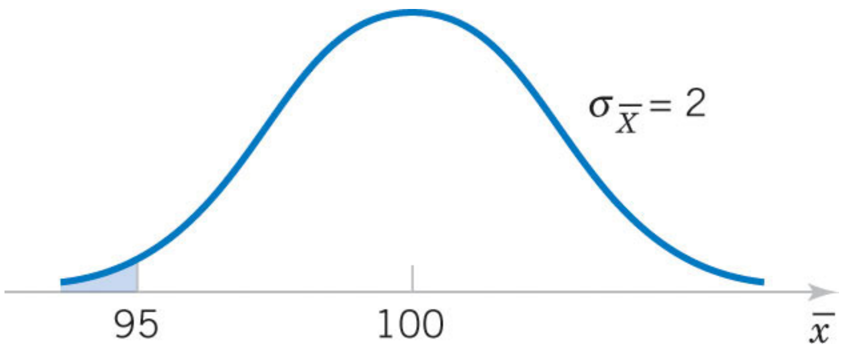


Figure 7-2 Desired probability is shaded

Answer:

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2.0$$

$$\Phi\left(\frac{\overline{X}-\mu}{\sigma_{\overline{X}}}\right) = \Phi\left(\frac{95-100}{2}\right)$$

$$= \Phi(-2.5) = 0.0062$$

Two Populations

We have two independent populations. What is the distribution of the difference of their sample means?

The sampling distribution of $\bar{X}_1 - \bar{X}_2$

has the following mean anad variance:

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Sampling Distribution of a Difference in Sample Means

- If we have two independent populations with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2,
- And if *X*-bar$_1$ and *X*-bar$_2$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations:
- Then the sampling distribution of:

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (7\text{-}4)$$

  is approximately standard normal, if the conditions of the central limit theorem apply.

- If the two populations are normal, then the sampling distribution is exactly standard normal.

Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a random variable with μ_{old}=5000 hours and σ_{old}=40 hours (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to μ_{new}=5050 hours and σ_{new}=30 hours (new).

Random samples of 16 components manufactured using “old” process and 25 components using “new” process are chosen.

What is the probability new sample mean is at least 25 hours longer than old?

Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a random variable with μ_{old}=5000 hours and σ_{old}=40 hours (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to μ_{new}=5050 hours and σ_{new}=30 hours (new).

Random samples of 16 components manufactured using "old" process and 25 components using "new" process are chosen.

What is the probability new sample mean is at least 25 hours longer than old?

$$\sigma_{\bar{X}_{old}} = \frac{\sigma_{old}}{\sqrt{16}} = 10 \text{ hrs}$$

$$\sigma_{\bar{X}_{new}} = \frac{\sigma_{new}}{\sqrt{25}} = 6 \text{ hrs}$$

$$\sigma_{TOT} = \sqrt{\sigma_{\bar{X}_{old}}^2 + \sigma_{\bar{X}_{new}}^2} =$$

$$= \sqrt{100+36} \approx 11.7 \text{ hrs}$$

$$\mu_{new} - \mu_{old} = 50 \text{ hrs}$$

$$z = \frac{25-(50)}{11.7} = -2.14$$

$$\text{Prob}(z > -2.14) = 0.9840$$

Example 7-3: Aircraft Engine Life

The effective life of a component used in jet-turbine aircraft engines is a normal-distributed random variable with parameters shown (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters mu and sigma as shown (new).

Random samples are selected from the "old" process and "new" process as shown.

What is the probability new sample mean is at least 25 hours longer than old?

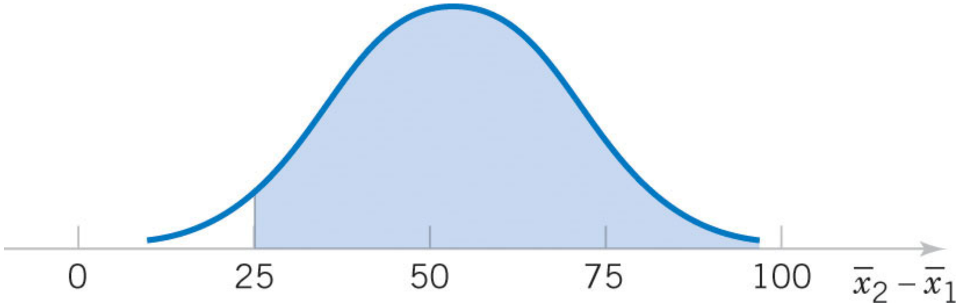
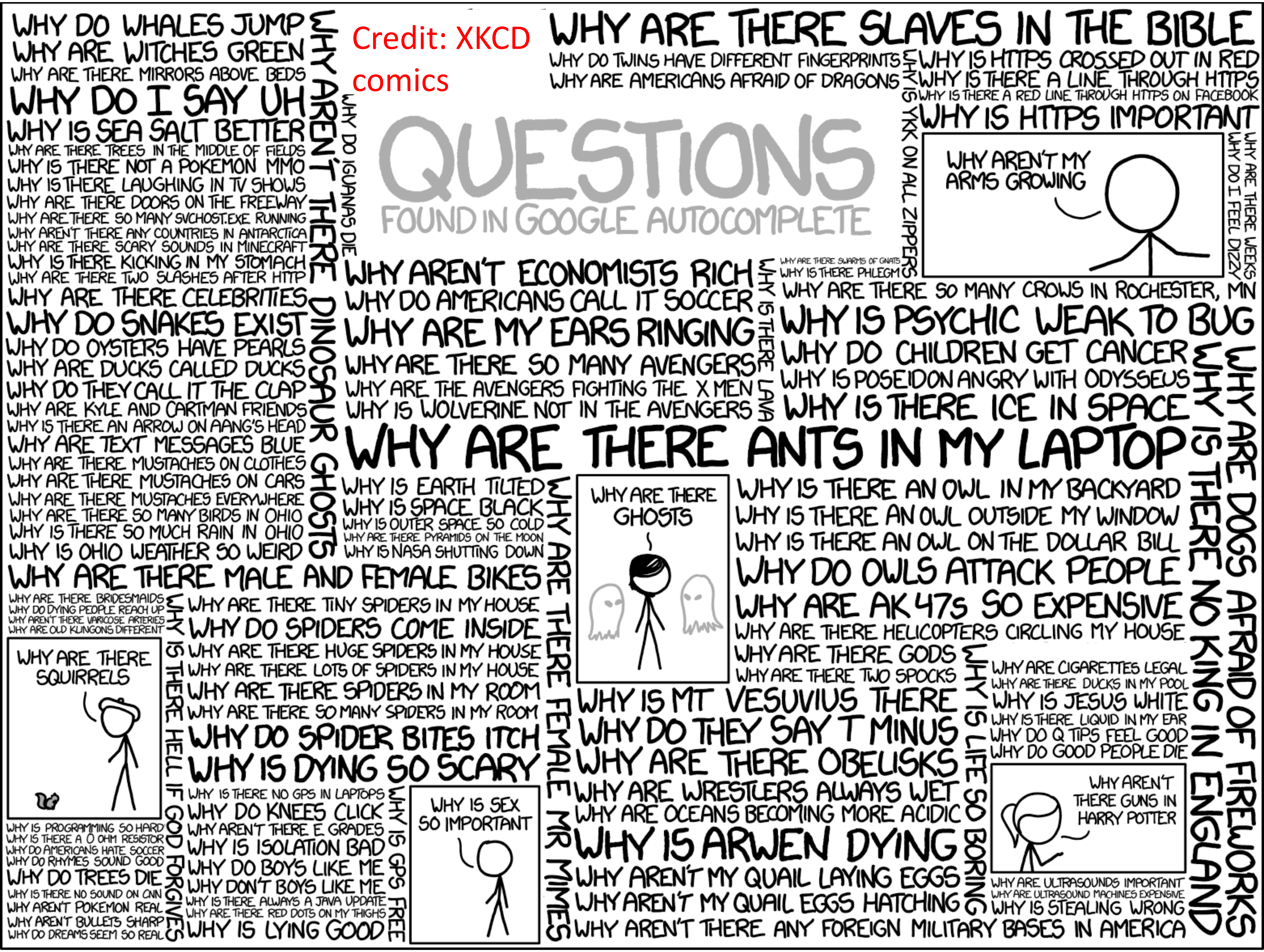


Figure 7-4 Sampling distribution of the sample mean difference.

	Process		
	Old (1)	New (2)	Diff (2-1)
mu =	5,000	5,050	50
sigma =	40	30	50
n =	16	25	
Calculations			
s / √*n* =	10	6	11.7
		z =	-2.14
P(xbar$_2$-xbar$_1$ > 25) = P(Z>z) =			0.9840

Credit: XKCD comics


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY ARE THERE MALE AND FEMALE BIKES
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY ARE THERE SQUIRRELS
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT

Descriptive statistics:

Point estimation:

Some Definitions

- The random variables X_1, X_2,…,X_n are a random sample of size *n* if:
  a) The X_i are independent random variables.
  b) Every X_i has the same probability distribution.

  Such X_1, X_2,…,X_n are also called independent and identically distributed (or i. i. d.) random variables
- A **statistic** is any function of the observations in a random sample.
- The probability distribution of a statistic is called a **sampling distribution**.

Point Estimation

- A sample was collected: X_1, X_2,…, X_n
- We suspect that sample was drawn from a random variable distribution *f(x)*
- *f(x)* has k parameters that we do not know
- Point estimates are estimates of the parameters of the *f(x)* describing the population based on the sample
  - For exponential PDF: $f(x)=\lambda\exp(-\lambda x)$ one wants to estimate λ
  - For Bernoulli PDF: $p^x(1-p)^{1-x}$ one wants to estimate p
  - For normal PDF one wants to estimates both μ and σ
- Point estimates are uncertain: therefore we can talk of averages and standard deviations of point estimates

Point Estimator

A point estimate of some parameter θ describing population random variable is a single numerical value $\hat{\theta}$ depending on all values $x_1, x_2, \ldots x_n$ in the sample.
The sample statistic (whis a random variable $\widehat{\Theta}$ defined by a function $\widehat{\Theta}(\mathrm{X}_1, \mathrm{X}_2, \ldots X_n)$) is called the
point estimator.

- There could be multiple choices for the point estimator of a parameter.
- To estimate the mean of a population, we could choose the:
  - Sample mean
  - Sample median
  - Peak of the histogram
  - ½ of (largest + smallest) observations of the sample.
- We need to develop criteria to compare estimates using statistical properties.

Unbiased Estimators Defined

The point estimator $\widehat{\Theta}$ is an unbiased estimator

for the parameter θ if:

$$E(\widehat{\Theta}) = \theta \qquad (7-5)$$

If the estimator is not unbiased, then the difference:

$$E(\widehat{\Theta}) - \theta \qquad (7-6)$$

is called the bias of the estimator $\widehat{\Theta}$.

Mean Squared Error

The **mean squared error** of an estimator $\widehat{\Theta}$ of the parameter θ is defined as:

$$\mathrm{MSE}\left(\widehat{\Theta}\right) = E\left(\widehat{\Theta} - \theta\right)^2 \qquad (7\text{—}7)$$

Can be rewritten as

$$= E\left[\widehat{\Theta} - E\left(\widehat{\Theta}\right)\right]^2 + \left[\theta - E\left(\widehat{\Theta}\right)\right]^2$$

$$= V\left(\widehat{\Theta}\right) + (\mathrm{bias})^2$$

Methods of Point Estimation

- We will cover two popular methodologies to create point estimates of a population parameter.
  - Method of moments
  - Method of maximum likelihood
- Each approach can be used to create estimators with varying degrees of biasedness and relative MSE efficiencies.

Method of moments for point estimation

What are moments?

- A k-th moment of a random variable is the expected value *E(X^k^)*
  - First moment: $\mu = \int_{\infty}^{+\infty} xf(x)\,dx$
  - Second moment: $\mu^2 + \sigma^2 = \int_{\infty}^{+\infty} x^2 f(x)\,dx$
- A population moment relates to the entire population
- A sample moment is calculated like its population moments but for a finite sample
  - Sample first moment = sample mean = $\frac{1}{n}\sum_{i=1}^{n} x_i$
  - Sample k-th moment $\frac{1}{n}\sum_{i=1}^{n} x_i^k$

Moment Estimators

Let $X_1, X_2, ..., X_n$ be a random sample from either a probability mass function or a probability density function with m unknown parameters $\theta_1, \theta_2, ..., \theta_m$.

The moment estimators $\hat{\Theta}_1$ $\hat{\Theta}_2$ $..., \hat{\Theta}_m$ are found by equating the first m population moments to the first m sample moments and solving the resulting simultaneous equations for the unknown parameters.

Exponential Distribution: Moment Estimator-1st moment

- Suppose that x_1, x_2, …, x_n is a random sample from an exponential distribution $f(x)=\lambda\exp(-\lambda x)$ with parameter λ.
- There is only one parameter to estimate, so equating population and sample first moments, we have one equation: $E(X) = \bar{x}$.
- $E(X) = 1/\lambda$ thus
  $\lambda = 1/\bar{x}$ is the 1st moment estimator.

Descriptive statistics:

Point estimation:

Point Estimation

- A sample was collected: $X_1, X_2, \ldots, X_n$
- We suspect that sample was drawn from a random variable distribution *f(x)*
- *f(x)* has k parameters that we do not know
- Point estimates are estimates of the parameters of the *f(x)* describing the population based on the sample
  - For exponential PDF: $f(x)=\lambda \exp(-\lambda x)$ one wants to estimate λ
  - For Bernoulli PDF: $p^x(1-p)^{1-x}$ one wants to estimate p
  - For normal PDF one wants to estimates both μ and σ
- Point estimates are uncertain: therefore, we can talk of averages and standard deviations of point estimators

Point Estimator

A point estimate of some parameter θ describing population random variable is a single numerical value $\hat{\theta}$ depending on all values $x_1, x_2, \ldots x_n$ in the sample.
The sample statistic (whis a random variable $\widehat{\Theta}$ defined by a function $\widehat{\Theta}(\mathrm{X}_1, \mathrm{X}_2, \ldots X_n)$) is called the
point estimator.

- There could be multiple choices for the point estimator of a parameter.
- To estimate the mean of a population, we could choose the:
  - Sample mean
  - Sample median
  - Peak of the histogram
  - ½ of (largest + smallest) observations of the sample.
- We need to develop criteria to compare estimates using statistical properties.

Unbiased Estimators Defined

The point estimator $\widehat{\Theta}$ is an unbiased estimator

for the parameter θ if:

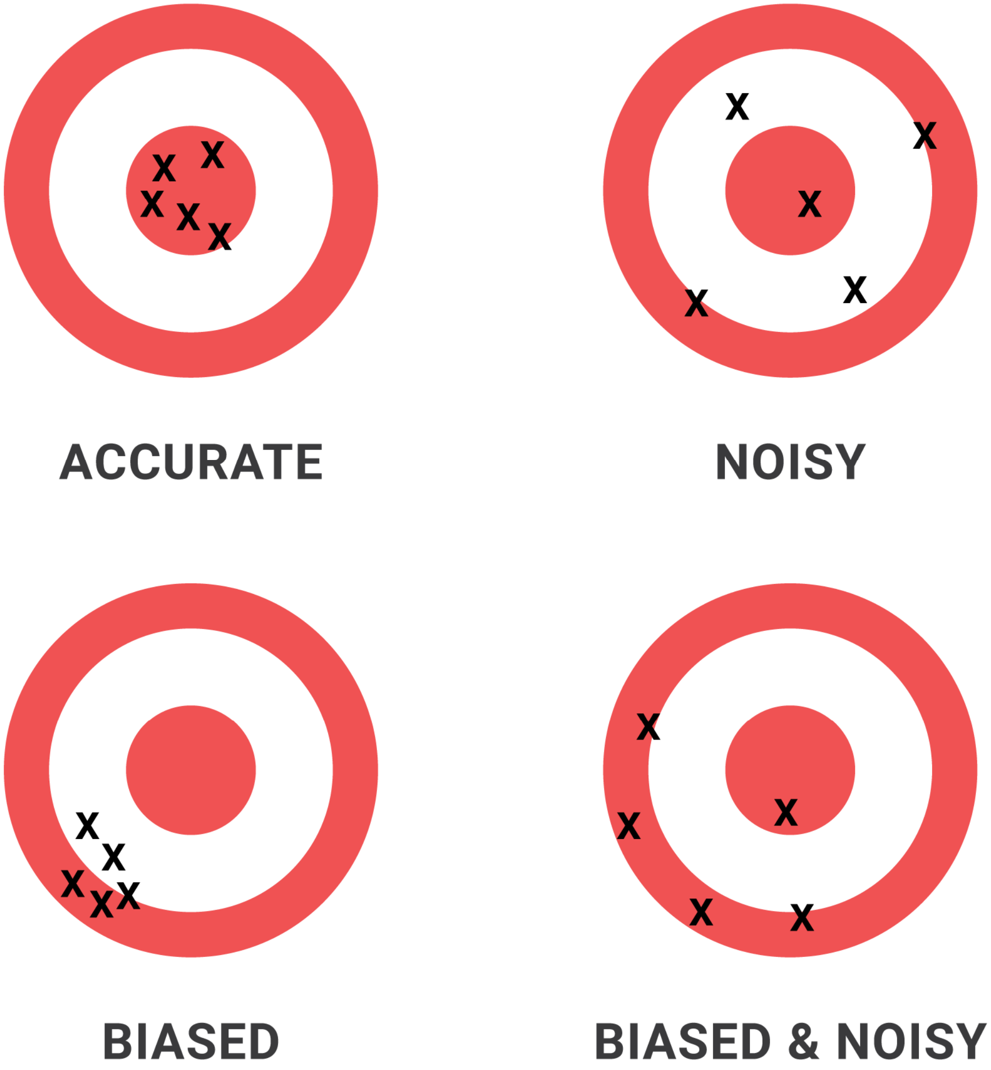
$$E\left(\widehat{\Theta}\right) = \theta \qquad (7-5)$$

If the estimator is not unbiased, then the difference:

$$E\left(\widehat{\Theta}\right) - \theta \qquad (7-6)$$

is called the bias of the estimator $\widehat{\Theta}$.

Bias vs Noise



ACCURATE

NOISY

BIASED

BIASED & NOISY

NOISE: Daniel Kahneman, Olivier Sabony, and Cass R. Sunstein

Mean Squared Error

The **mean squared error** of an estimator $\widehat{\Theta}$ of the parameter θ is defined as:

$$\mathrm{MSE}(\widehat{\Theta}) = E(\widehat{\Theta} - \theta)^2 \qquad (7\text{–}7)$$

Can be rewritten as

$$= E[\widehat{\Theta} - E(\widehat{\Theta})]^2 + [\theta - E(\widehat{\Theta})]^2$$

$$= V(\widehat{\Theta}) + (\text{bias})^2$$

Statistic #1: Sample Mean

If the values of n observations in a random sample are denoted by $x_1, x_2, \ldots, x_n,$ the sample mean is

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (6-1)$$

New random variable $\overline{X}$ is a linear combination of *n* independent identically distributed variables $X_1, X_2, \ldots, X_n$

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Sample mean $\bar{x}$ is drawn from a random variable

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$E(\bar{X}) = \frac{n \cdot E(X_i)}{n} = \frac{n \cdot \mu}{n} = \mu$$

Sample mean, $\bar{X}$, is an <u>unbiased</u> estimator of the population mean, μ

Sample variance S^2 –
is an estimator of
the population variance σ^2

Sample Variance

If n observations in a sample are denoted by $x_1, x_2, \ldots, x_n,$ the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \boldsymbol{\bar{x}})^2}{n-1} \qquad (6\text{–}3)$$

If one knows the population average, μ, one divides by n to estimate the variance

$$s(\mu)^2 = \frac{\sum_{i=1}^{n}(x_i - \boldsymbol{\mu})^2}{\boldsymbol{n}}$$

Why divide by n-1 instead of n?

- The sample mean $\bar{x}$ is on average closer to points $x_1, x_{2,} \ldots x_n$ than the true mean μ $\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq \sum_{i=1}^{n}(x_i - \mu)^2$
- Consider a sample of size n=1. Then $\bar{x}$= x_1 while μ≠ x_1 . Dividing by n gives s^2 =0, while dividing by n-1 leaves s^2 undefined (0/0)
- For n=2, $\bar{x}$ is exactly halfway between x_1 and x_2 making its sum of squares smaller than that of μ
- Dividing by n-1 on average corrects for a smaller sum of squares: S^2 is an unbiased estimator of σ^2

Show that S^2 is unbiased estimate of $\sigma 2$

$$E(S^2) = E\left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}\right) = \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i^2+\bar{X}^2-2\bar{X}X_i\right)\right] =$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}X_i^2 + n\bar{X}^2 - 2\bar{X}n\bar{X}\right] =$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right) = \frac{1}{n-1}\left(nE(X_i^2) - nE(\bar{X}^2)\right)$$

$$= \frac{1}{n-1}\left(n(\mu^2+\sigma^2) - n\left(\mu^2+\frac{\sigma^2}{n}\right)\right) = \frac{n-1}{n-1}\sigma^2 = \underline{\underline{\sigma^2}}$$

Example 7-4: Sample Variance S^2 is Unbiased

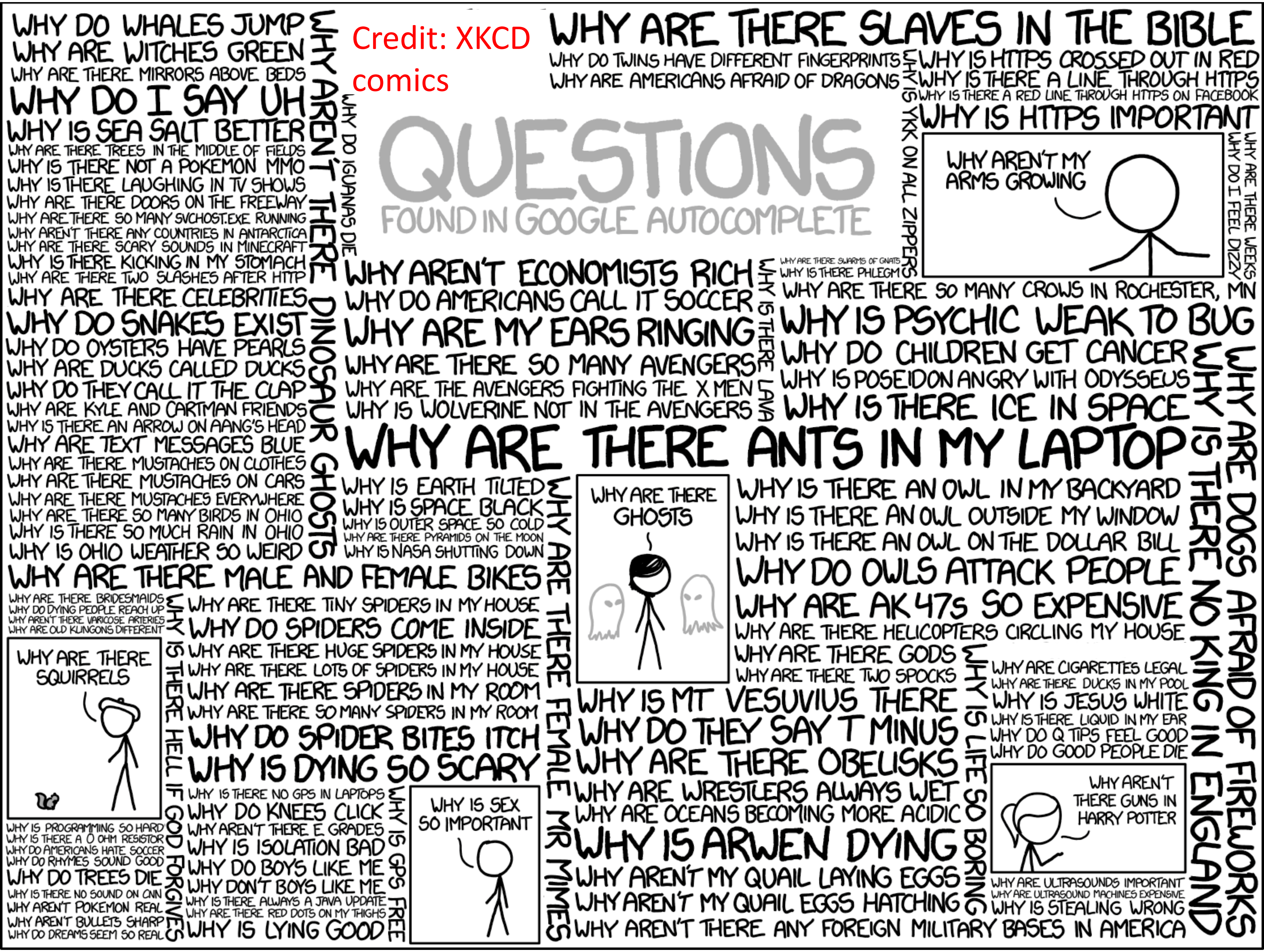
$$E\left(S^2\right) = E\left(\frac{\sum_{i=1}^{n}\left(X-\bar{X}\right)^2}{n-1}\right)$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i^2+\bar{X}^2-2\bar{X}X_i\right)\right]$$

$$= \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\mu^2+\sigma^2\right)-n\left(\mu^2+\frac{\sigma^2}{n}\right)\right]$$

$$= \frac{1}{n-1}\left[n\mu^2+n\sigma^2-n\mu^2-\sigma^2\right] = \frac{1}{n-1}\left[(n-1)\sigma^2\right]$$


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT

Credit: XKCD comics

Methods of Point Estimation

- We will cover two popular methodologies to create point estimates of a population parameter.
  - Method of moments
  - Method of maximum likelihood
- Each approach can be used to create estimators with varying degrees of biasedness and relative MSE efficiencies.

Method of moments for point estimation

What are moments?

- The p-th population moment of a random variable is the expected value of *X^p*
  - First moment: $\mu = \int_{\infty}^{+\infty} xf(x)\,dx$
  - Second moment: $\mu^2 + \sigma^2 = \int_{\infty}^{+\infty} x^2 f(x)\,dx$
  - p-th moment: $\int_{\infty}^{+\infty} x^p f(x)\,dx$
  - The population moment relates to the entire population

- A sample moment is calculated like its population moments but for a finite sample
  - Sample first moment = sample mean = $\frac{1}{n}\sum_{i=1}^{n} x_i$
  - Sample p-th moment $\frac{1}{n}\sum_{i=1}^{n} x_i^p$

Moment Estimators

Let $X_1, X_2, \ldots, X_n$ be a random sample from either a probability mass function or a probability density function with p unknown parameters $\theta_1, \theta_2, \ldots, \theta_p$.

The moment estimators $\widehat{\Theta}_1$ $\widehat{\Theta}_2$ $\ldots, \widehat{\Theta}_p$ are found by equating the first p population moments to the first p sample moments and solving the resulting simultaneous equations for the unknown parameters.

Exponential Distribution: Moment Estimator-1st moment

- Suppose that $x_1, x_2, \ldots, x_n$ is a random sample from an exponential distribution $f(x)=\lambda\exp(-\lambda x)$ with parameter λ.
- There is only one parameter to estimate, so equating population and sample first moments, we have one equation: $E(X) = \bar{x}$.
- $E(X) = 1/\lambda$ thus
  $\lambda = 1/\bar{x}$ is the 1st moment estimator.

How I solved it

- Stats=100000;
- Y=random('Exponential', 1/3, Stats, 1);

%parametrization in MATLAB is 1/lambda

- 1/mean(Y) %matching the first moment

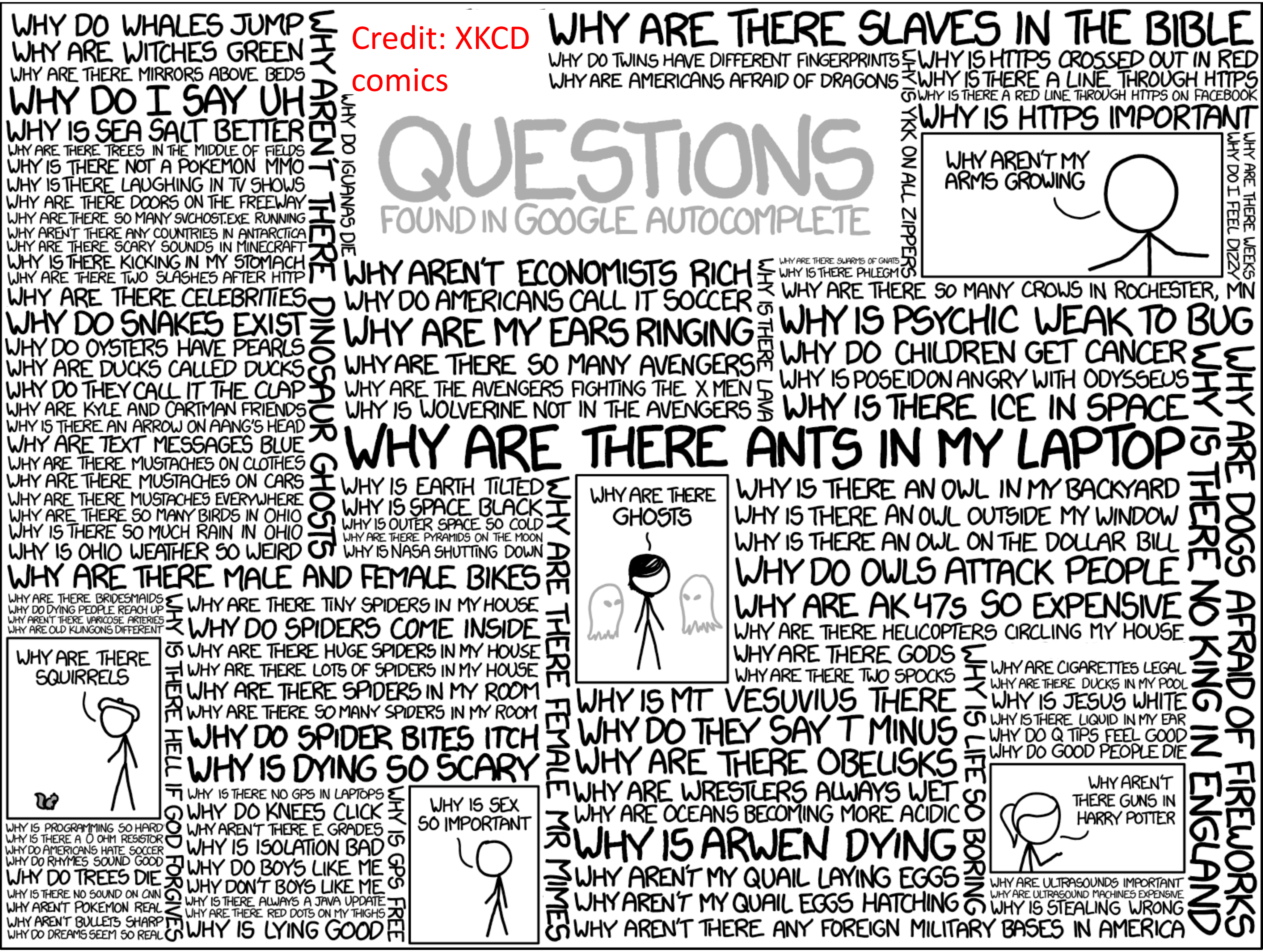
% ans = 3.0086

- sqrt(2/mean(Y.^2)) %matching the second moment

% ans = 3.0081

- (factorial(20)/mean(Y.^20))^(1./20) %matching the 20th moment

Credit: XKCD comics


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Method of Maximum Likelihood for point estimation

Maximum Likelihood Estimators

- Suppose that X is a random variable with probability distribution $f(x, \theta)$, where θ is a single unknown parameter. Let $x_1, x_2, \ldots, x_n$ be the observed values in a random sample of size n. Then the likelihood function of the sample is the probability to get it in a random variable with PDF $f(x, \theta)$:

$$L(\theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n , \theta) \qquad (7\text{-}9)$$

- Note that the likelihood function is now a function of only the unknown parameter θ. The maximum likelihood estimator (MLE) of θ is the value of θ that maximizes the likelihood function $L(\theta)$.
- Usually, it is easier to work with logarithms: $l(\theta) = \ln L(\theta)$

Exponential MLE:

$$f(x_i) = \lambda e^{-\lambda x_i}$$

$$L(\lambda) = P(x_1, x_2 \ldots x_n \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} =$$

$$= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$\ln L(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum x_i = 0$$

$$\hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{X}}$$

Same as 1st moment estimator

Example 7-11: Exponential MLE

Let *X* be a exponential random variable with parameter λ. The likelihood function of a random sample of size n is:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$\ln L\,(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{d \ln L\,(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{X}} \quad \text{(same as moment estimator)}$$

Bernoulli MLE

$$f(x, p) = p^{x}(1-p)^{1-x}$$

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} =$$

$$= p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$$\ln L(p) = \left(\sum x_i\right)\ln p + \left(n - \sum x_i\right)\ln(1-p)$$

$$\frac{d\ln L(p)}{dp} = \frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p} = 0 \text{ at } \hat{p}$$

$$0 = \frac{(1-\cancel{\hat{p}})\sum x_i - \hat{p}\,(n-\cancel{\sum} x_i)}{\hat{p}(1-\hat{p})}$$

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example 7-9: Bernoulli MLE

Let *X* be a Bernoulli random variable. The probability mass function is $f(x;p) = p^x(1-p)^{1-x}$, $x = 0, 1$ where P is the parameter to be estimated. The likelihood function of a random sample of size *n* is:

$$L(p) = p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \cdot \ldots \cdot p^{x_n}(1-p)^{1-x_n}$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

$$\ln L\,(p) = \left(\sum_{i=1}^{n} x_i\right)\ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p)$$

$$\frac{d \ln L\,(p)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{(n - \sum_{i=1}^{n} x_i)}{(1-p)} = 0$$

$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$ (same as moment estimator)

Normal MLE for μ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$L(\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(-\frac{\sum(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\ln L(\mu,\sigma) = -n\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{d\ln L(\mu,\sigma)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \text{ at } \hat{\mu}$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example 7-10: Normal MLE for μ

Let *X* be a normal random variable with unknown mean μ and variance σ^2. The likelihood function of a random sample of size n is:

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{(2\sigma^2)}}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\ln L\,(\mu) = \frac{-n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{d\ln L\,(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X} \text{ (same as moment estimator)}$$

Example 7-11: Normal MLE for σ²

Let *X* be a normal random variable with the estimate of mean μ determined by MLE (see the previous slide) and an unknown variance σ². The likelihood function of a random sample of size n is:

$$L(\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{(2\sigma^2)}}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\ln L(\sigma) = \frac{-n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{d\ln L(\sigma)}{d\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} \quad \text{(biased estimator)}$$

MLE for Poisson distribution

$$f(x_1, \ldots, x_n|\lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$

$$= \frac{e^{-n\lambda}\lambda^{\Sigma_1^n x_i}}{x_1! \ldots x_n!}$$

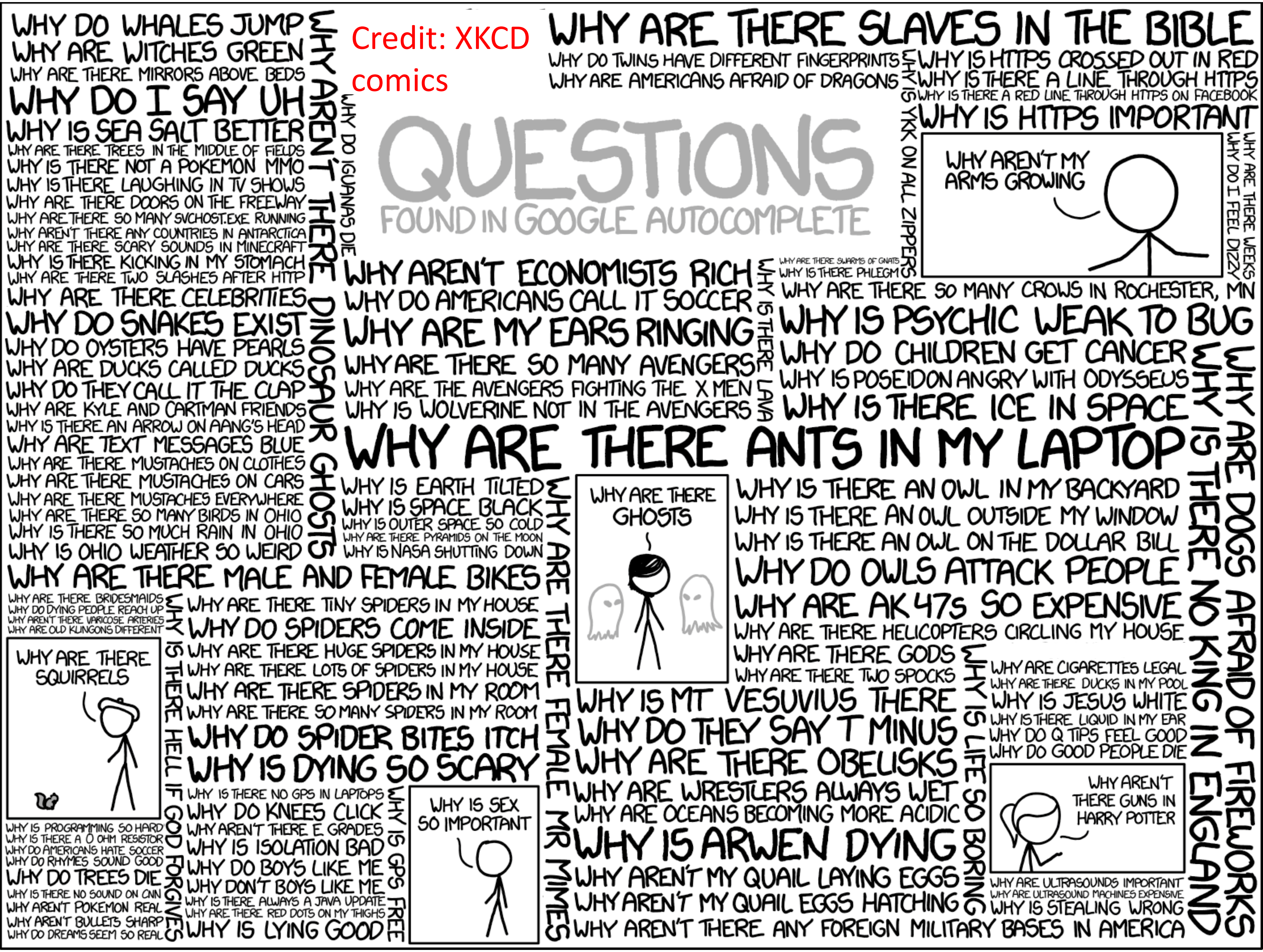
$$\log f(x_1, \ldots, x_n|\lambda) = -n\lambda + \sum_1^n x_i \log \lambda - \log c$$

where $c = \prod_{i=1}^n x_i!$ does not depend on λ, and

$$\frac{d}{d\lambda} \log f(x_1, \ldots, x_n|\lambda) = -n + \frac{\sum\limits_1^n x_i}{\lambda}$$

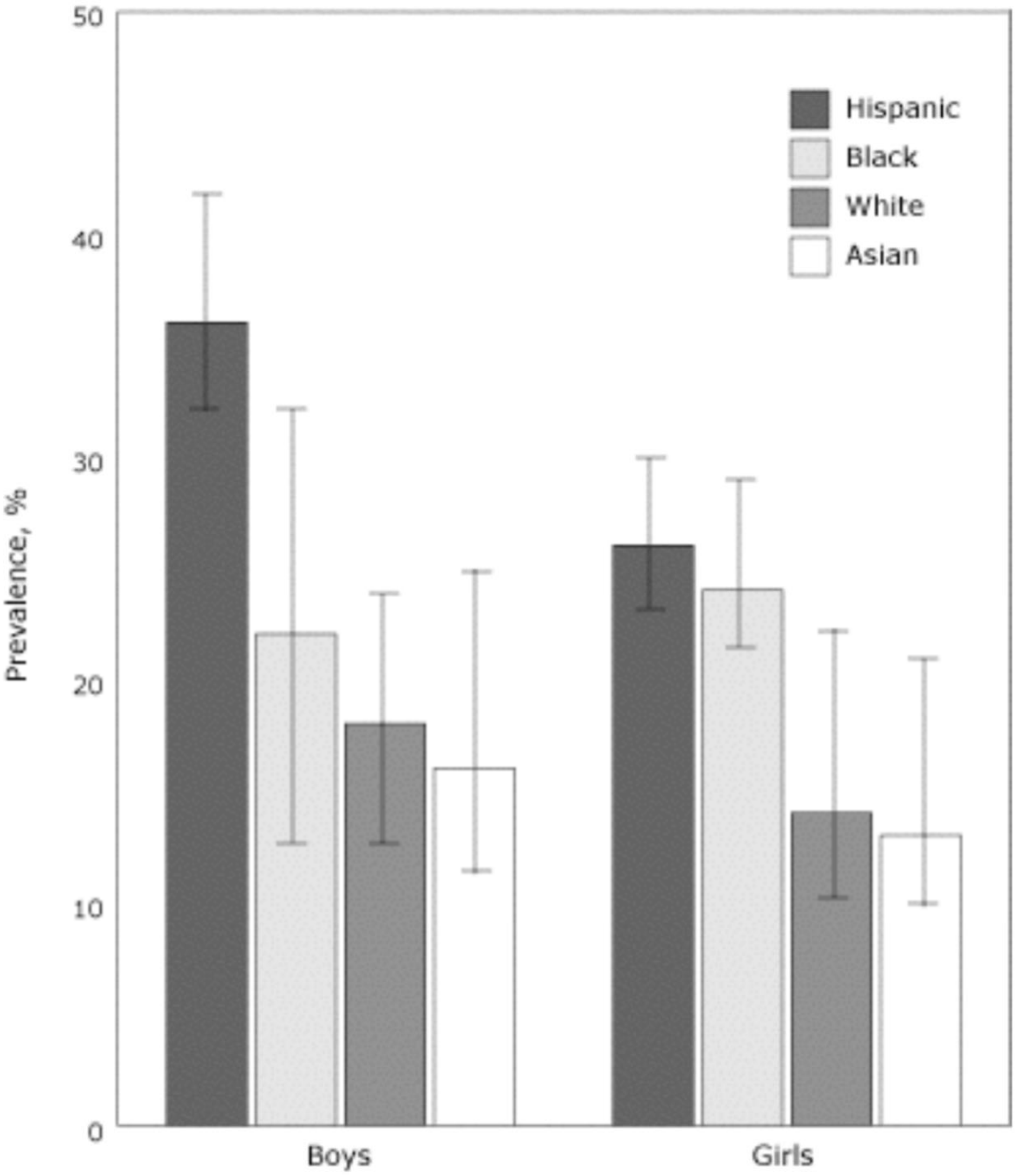
By equating to zero, we obtain that the maximum likelihood estimate $\hat{\lambda}$ equals

$$\hat{\lambda} = \frac{\sum\limits_1^n x_i}{n}$$


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT

Credit: XKCD comics

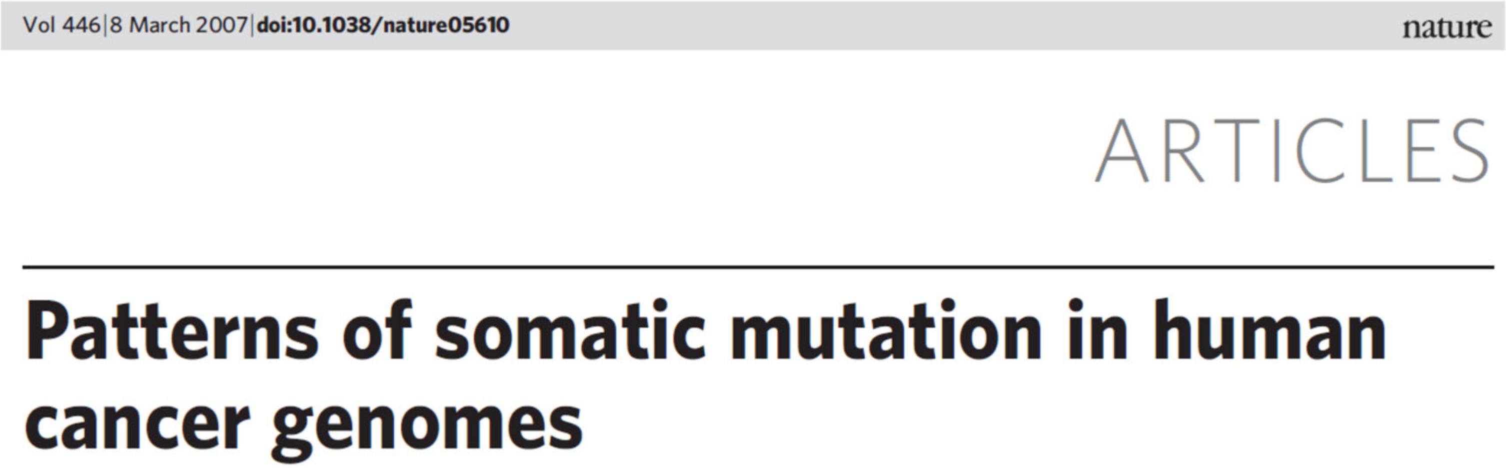
Confidence Intervals



Prevalence (with 95% CI bars) of obesity among New York City public elementary schoolchildren, by sex and race/ethnicity, 2003.

(source: CDC.GOV)

What do those bars actually mean?



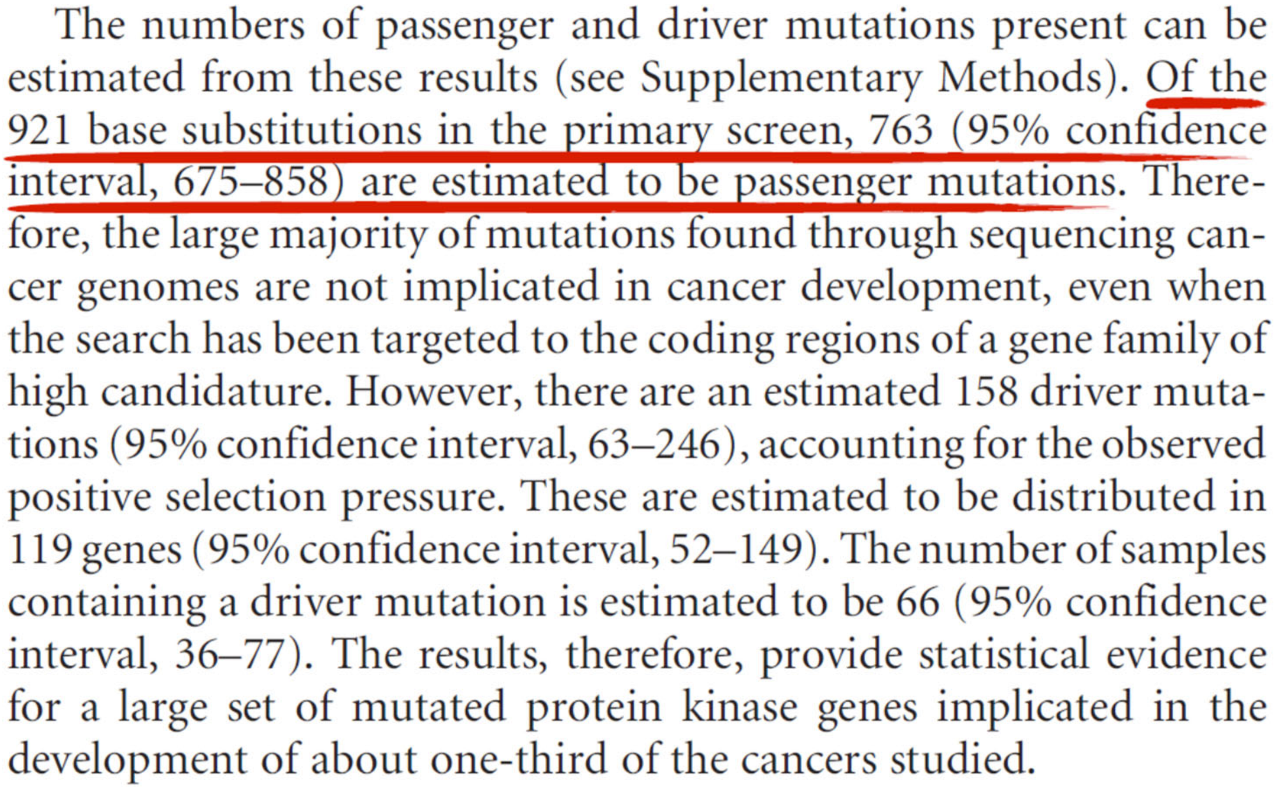
Vol 446|8 March 2007|doi:10.1038/nature05610

nature

ARTICLES

Patterns of somatic mutation in human cancer genomes

What does confidence interval mean?



The numbers of passenger and driver mutations present can be estimated from these results (see Supplementary Methods). Of the 921 base substitutions in the primary screen, 763 (95% confidence interval, 675–858) are estimated to be passenger mutations. Therefore, the large majority of mutations found through sequencing cancer genomes are not implicated in cancer development, even when the search has been targeted to the coding regions of a gene family of high candidature. However, there are an estimated 158 driver mutations (95% confidence interval, 63–246), accounting for the observed positive selection pressure. These are estimated to be distributed in 119 genes (95% confidence interval, 52–149). The number of samples containing a driver mutation is estimated to be 66 (95% confidence interval, 36–77). The results, therefore, provide statistical evidence for a large set of mutated protein kinase genes implicated in the development of about one-third of the cancers studied.

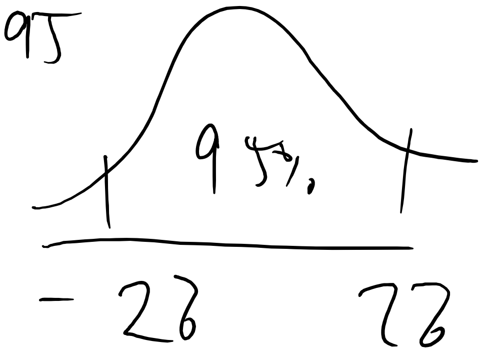
- We have talked about how a parameter can be estimated from sample data. However, it is important to understand how good is the estimate obtained.

- Bounds that represent an interval of plausible values for a parameter are an example of an interval estimate.

Two-sided confidence intervals

- Calculated based on the sample $X_1, X_2,\ldots,X_n$
- Characterized by:
  - lower- and upper- confidence limits *L* and *R*
  - the confidence coefficient $1\text{-}\alpha$
- Objective: for two-sided confidence interval, find L and R such that
  - Prob$(\mu>R)=\alpha/2$
  - Prob$(\mu<L)=\alpha/2$
  - Therefore, Prob$(L<\mu<R)=1-\alpha$
- For one-sided confidence interval, say, upper bound of μ *,* find R that
  - Prob$(\mu>R)=\alpha$
- **Assume standard deviation sigma is known**

Consider $1-\alpha = 95\% = 0.95$



$\alpha = 0.05$; $\frac{\alpha}{2} = 0.025$

$z_{\alpha/2} = 1.96 \Rightarrow \text{Prob}(Z > z_{\alpha/2}) = \frac{\alpha}{2}$

$$\text{Prob}\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\text{Prob}\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

For one sided lower bound on μ

$$\text{Prob}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\underline{\underline{\underline{\alpha}}}}\right) \rightarrow$$

$$\mu > \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$$

$z_\alpha = 1.65 <$

$z_{\alpha/2} = 1.96$

Exercise

Ishikawa et al. (Journal of Bioscience and Bioengineering 2012) studied the force with which bacterial biofilms adhere to a solid surface.

Five measurements for a bacterial strain of Acinetobacter gave readings 2.69, 5.76, 2.67, 1.62, and 4.12 dyne-cm2.

Assume that the standard deviation is known to be 0.66 dyne-cm2

(a) Find 95% confidence interval for the mean adhesion force

(b) If scientists want the width of the confidence interval to be below 0.55 dyne-cm2 what number of samples should be?

Ishikawa et al. (Journal of Bioscience and Bioengineering 2012) studied the force with which bacterial biofilms adhere to a solid surface. Five measurements for a bacterial strain of Acinetobacter gave readings 2.69, 5.76, 2.67, 1.62, and 4.12 dyne-cm2. Assume that the standard deviation is known to be 0.66 dyne-cm2

(a) Find 95% confidence interval for the mean adhesion force

(b) If scientists want the width of the confidence interval to be below 0.55 dyne-cm2 what number of samples should be?

a) 95% CI for μ, $n = 5$ $\sigma = 0.66$ $\bar{x} =$ 3.372, $z = 1.96$

$$\bar{x} - z\sigma/\sqrt{n} \le \mu \le \bar{x} + z\sigma/\sqrt{n}$$

$$3.372 - 1.96(0.66/\sqrt{5}) \le \mu \le 3.372 + 1.96(0.66/\sqrt{5})$$

$$2.79 \le \mu \le 3.95$$

b) Width is $2z\sigma/\sqrt{n} = 0.55$, therefore n = [2zσ/0.55]2 = [2(1.96)(0.66)/0.55]2 = 22.13

Round up to $n = 23$.

Confidence Intervals

- We have talked about how a parameter can be estimated from sample data. However, it is important to understand how good is the estimate obtained.

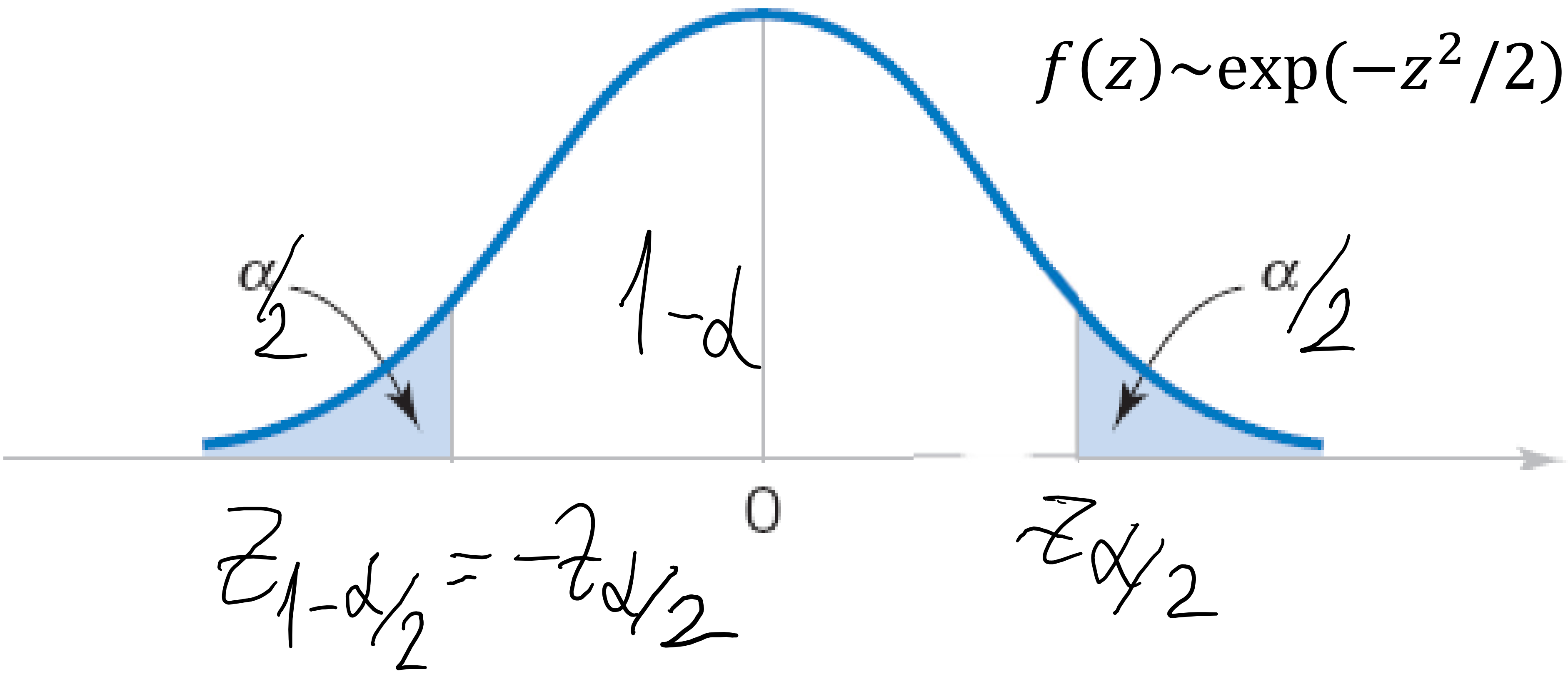
- Bounds that represent an interval of plausible values for a parameter are an example of an interval estimate.

Two-sided confidence intervals

- Calculated based on the sample $X_1, X_2,\ldots,X_n$
- Characterized by:
  - lower- and upper- confidence limits *L* and *U*
  - the confidence coefficient *1-α*
- Objective: for two-sided confidence interval, find L and R such that
  - Prob$(\mu > U) = \alpha/2$
  - Prob$(\mu < L) = \alpha/2$
  - Therefore, Prob$(L < \mu < U) = 1-\alpha$
- For one-sided confidence interval, say, upper bound of *μ* , find R that
  - Prob$(\mu > U) = \alpha$
- **Assume standard deviation σ is known**

Confidence Interval on the Population Mean, Variance Known

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$



Matlab exercise

- 1000 labs measured average P53 gene expression using n=20 samples drawn from the Gaussian distribution with mu=3; sigma=2;
- Each lab found 95% confidence estimates of the population mean mu **based on its sample only**
- Count the number of labs, where the population mean lies **outside their bounds**
- You should get ~50 labs out of 1000 labs

8-2 Confidence Interval on the Mean of a Normal Distribution, Variance Known

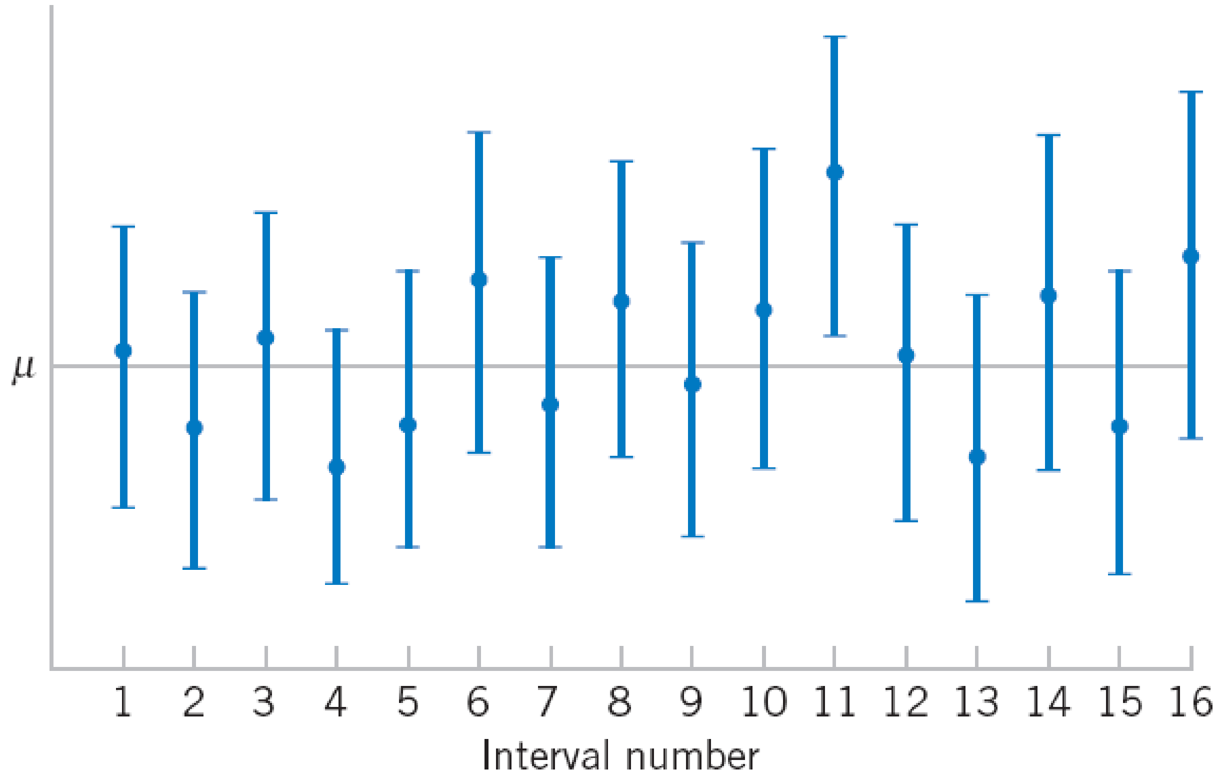


Figure 8-1 Repeated construction of a confidence interval for μ.

Figure 8-1 Repeated construction of a confidence interval for μ.

So far in estimating
confidence intervals for population mean μ
we assumed that the population variance σ^2
is known

Then (or when n>>1, say 20 and above)
one can use the Normal Distribution
to calculate confidence intervals

Q: What to do if the sample is small
& the population variance is **not known**?

A: Use the sample variance

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

but carefully:

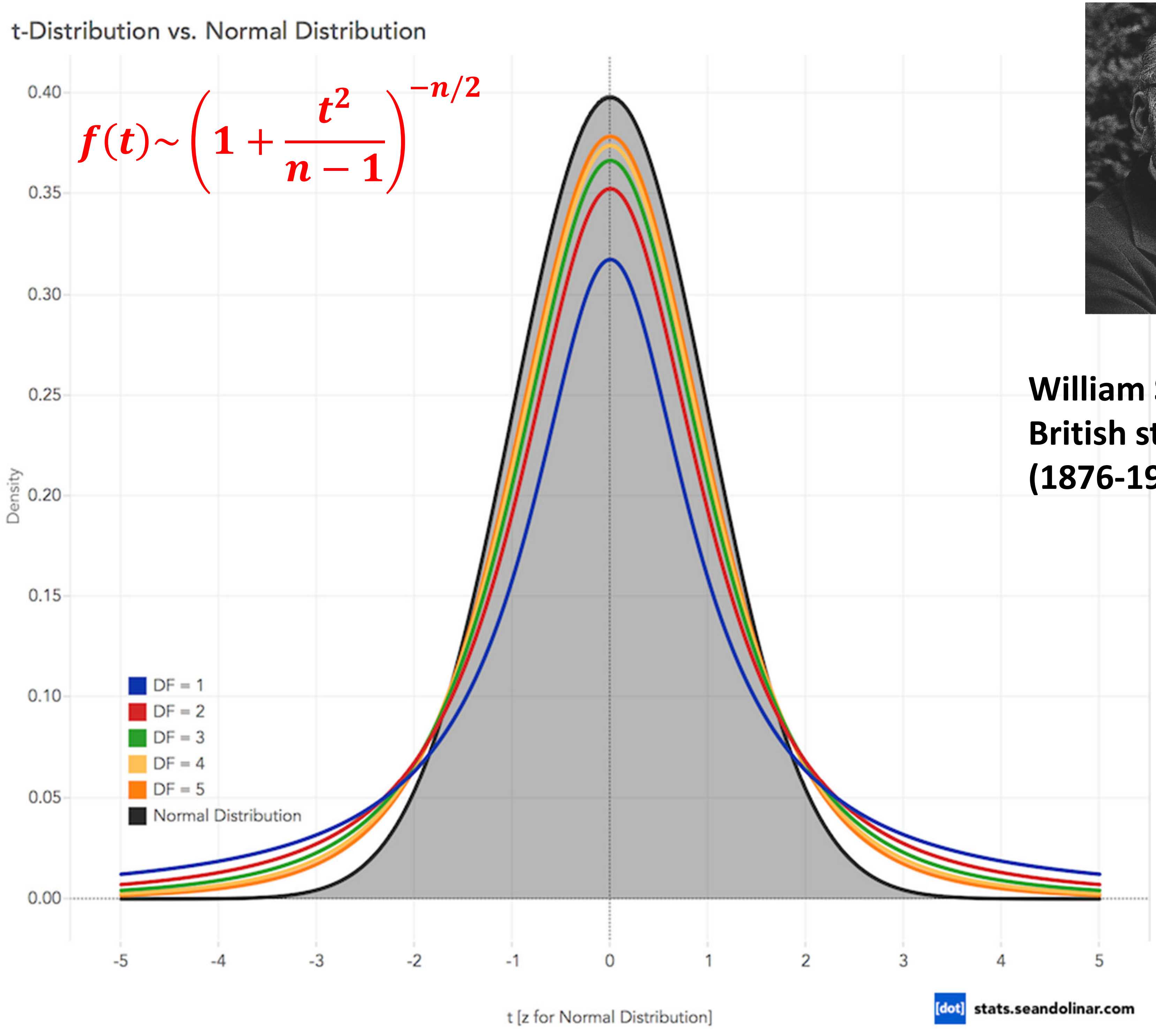
- Variable X has to be **normally distributed**
- **Student t-distribution** has to be used instead of
  the normal distribution (z-distribution).

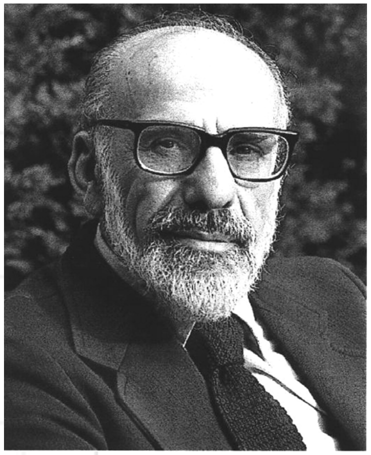
Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery. To prevent further disclosure of confidential information, Guinness prohibited its employees from publishing any papers regardless of the contained information. However, after pleading with the brewery and explaining that his mathematical and philosophical conclusions were of no possible practical use to competing brewers, he was allowed to publish them, but under a pseudonym ("Student"), to avoid difficulties with the rest of the staff. Thus, his most noteworthy achievement is now called Student's, rather than Gosset's, t-distribution.

Gosset had almost all his papers including “The probable error of a mean” (1908) published in Pearson's journal Biometrika under the pseudonym Student

Student's t-distribution

$$f(t) \sim \left(1 + \frac{t^2}{n-1}\right)^{-n/2}$$





William Sealy Gosset
British statistician
(1876-1937)

Play with Mathematica notebook

http://demonstrations.wolfram.com/ComparingNormalAndStudentsTDistributions/

By Gary McClelland

8-3 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu$$
$$< \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

Student's *t* distribution

$$f(t) \sim \left(1 + \frac{t^2}{n-1}\right)^{-n/2}$$

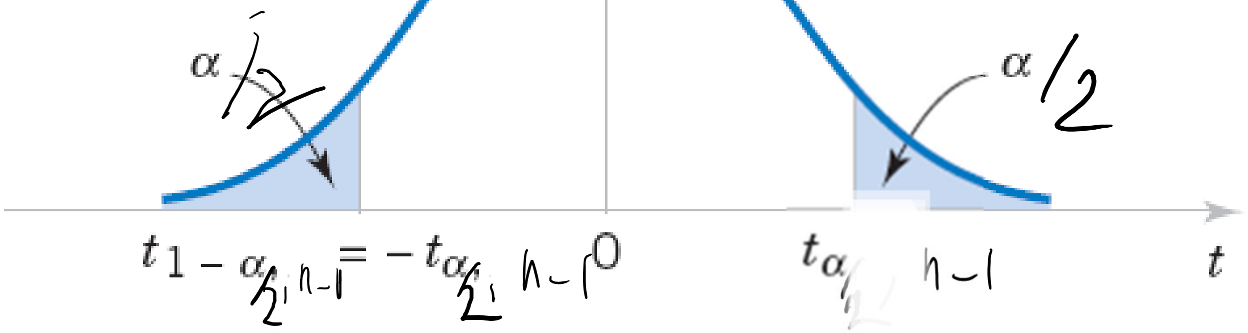
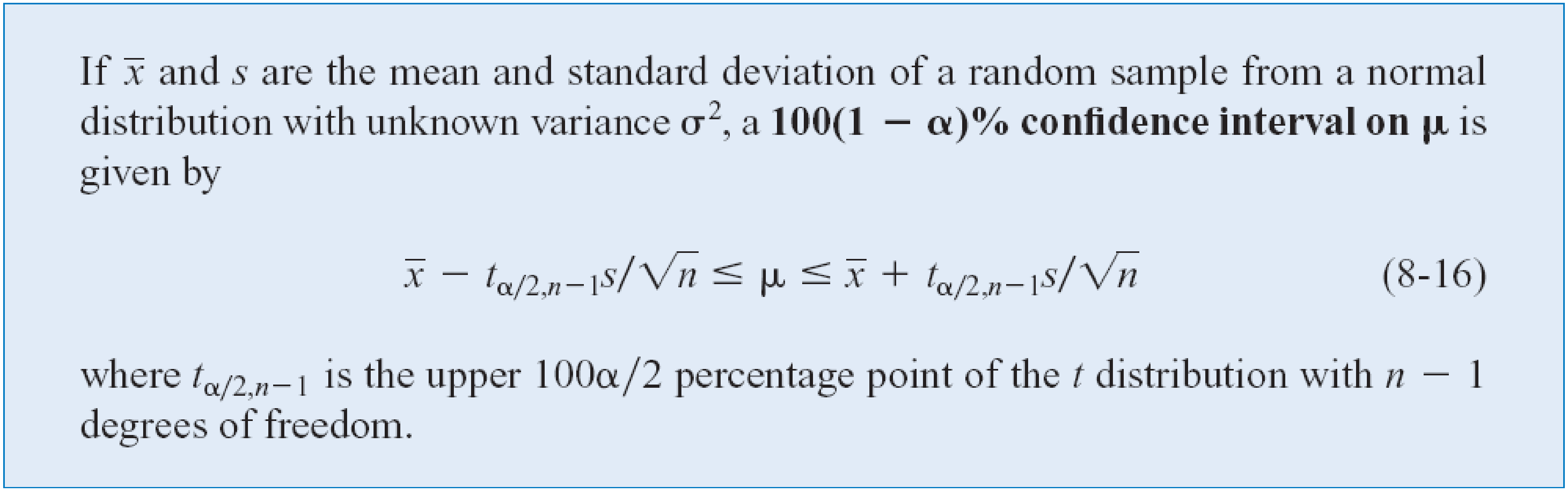


Figure 8-5 Percentage points of the *t* distribution.

8-3 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

8-3.2 The *t* Confidence Interval on μ

(Eq. 8-16)



If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a **100(1 − α)% confidence interval on μ** is given by

$$\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n} \qquad (8\text{-}16)$$

where $t_{\alpha/2,n-1}$ is the upper $100\alpha/2$ percentage point of the t distribution with $n - 1$ degrees of freedom.

One-sided confidence bounds on the mean are found by replacing $t_{\alpha/2,n-1}$ in Equation 8-16 with $t_{\alpha,n-1}$.

Confidence intervals for the population variance σ^2 based on the sample variance s^2

Confidence interval for the population variance σ^2

- Up until now we were calculating the confidence interval on the **population average μ**
- What if one wants to put **confidence interval on the population variance $\boldsymbol{\sigma}^2$**?
- We know an unbiased estimator of σ^2:

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

- How to determine the confidence interval?

Assume $\sigma=1, \mu=0$

$$\vec{X} = (x_1, \dots x_n)$$

$$y = |\vec{X}|^2 = \sum_{i=1}^{n} x_i^2 = (n-1)S^2$$

$$P(x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right)$$

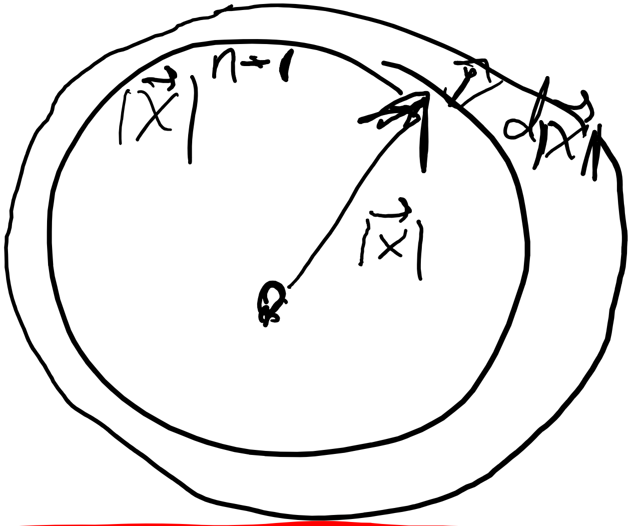
$$P(|\vec{X}|) \sim \exp\left(-\frac{|\vec{X}|^2}{2}\right) \cdot$$ surface of the sphere

$$|\vec{x}| = \sqrt{y}$$

$$d|\vec{X}| = \frac{1}{2}\frac{dy}{\sqrt{y}}$$

$$|\vec{X}|^{n-1} d|\vec{X}| = y^{\frac{n-1}{2}} y^{-\frac{1}{2}} dy =$$

$$= y^{\frac{n}{2}-1} dy$$

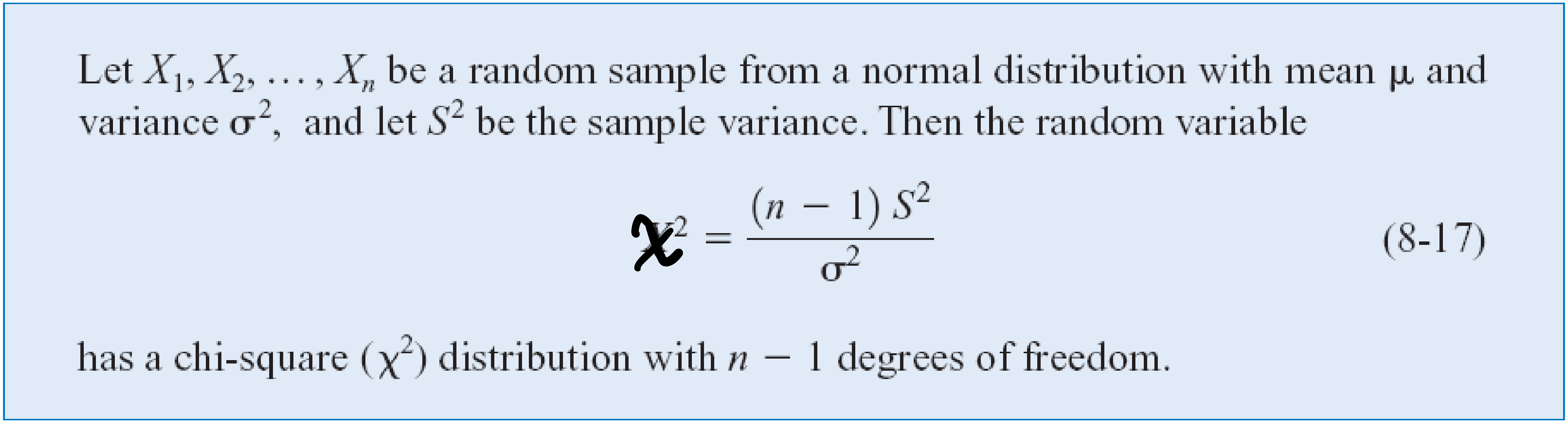


$$\boxed{P(y)dy = y^{\frac{n}{2}-1} \exp\left(-\frac{y}{2}\right) dy}$$

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

Definition

(Eq. 8-17)



Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2, and let S^2 be the sample variance. Then the random variable

$$\chi^2 = \frac{(n-1)\,S^2}{\sigma^2} \qquad (8\text{-}17)$$

has a chi-square (χ^2) distribution with $n - 1$ degrees of freedom.

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

$X=(n-1)S^2/\sigma^2$

We know n, S^2

want to estimate σ^2

$f(x,n) \sim x^{(n-1)/2-1} exp(-x/2)$

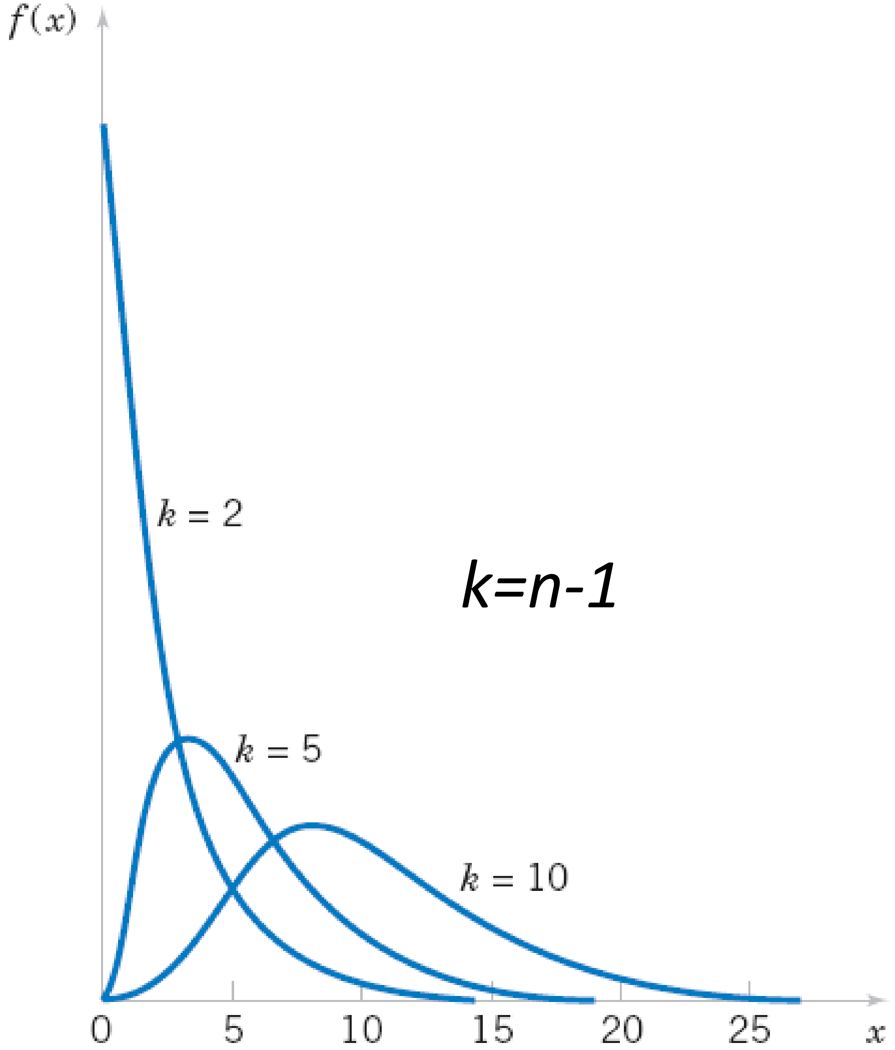
It is just Gamma PDF with $r=(n-1)/2$, and $\lambda=1/2$

Mean value:

$n-1$

Standard deviation:

$$\sqrt{2(n-1)}$$

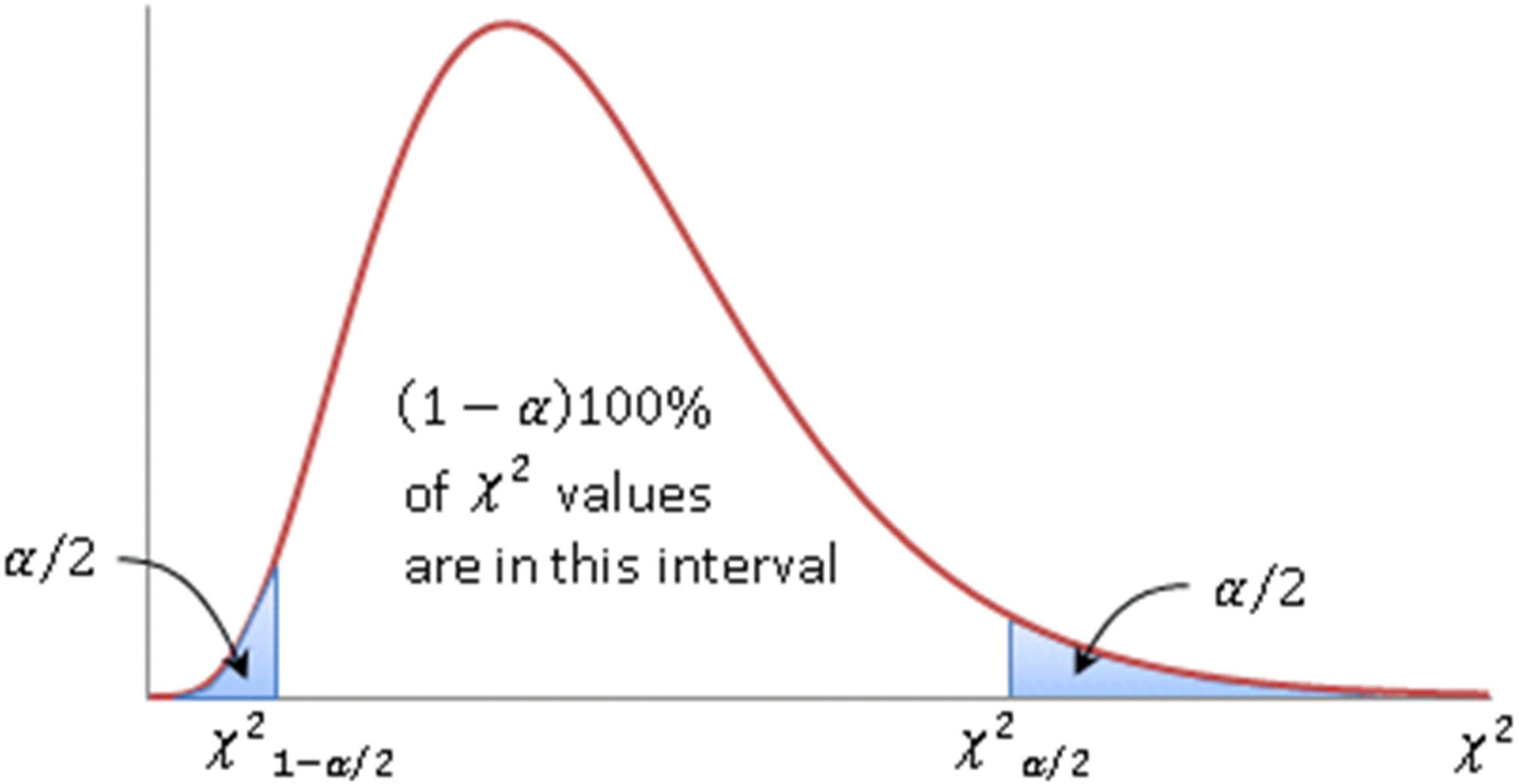


k=n-1

Play with Mathematica notebook

http://demonstrations.wolfram.com/ChiSquaredDistributionAndTheCentralLimitTheorem/

By Peter Falloon



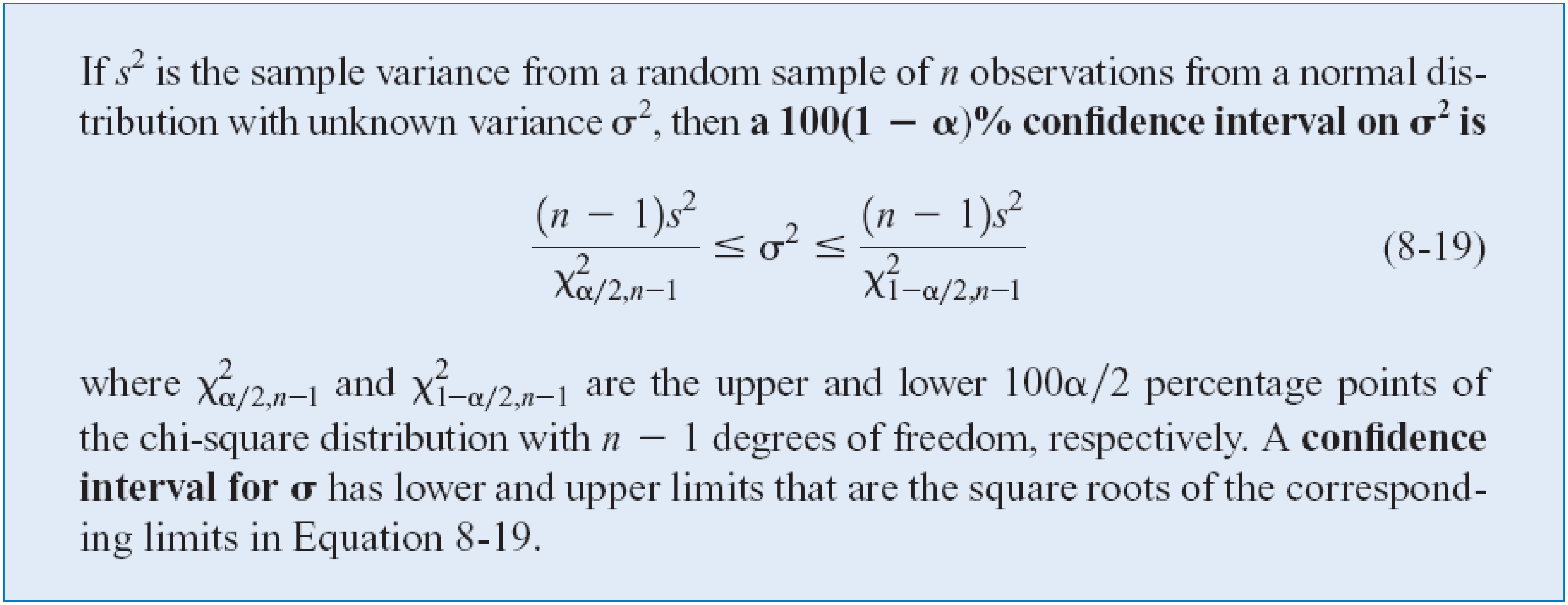
$$\chi^2_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

Definition

(Eq. 8-19)

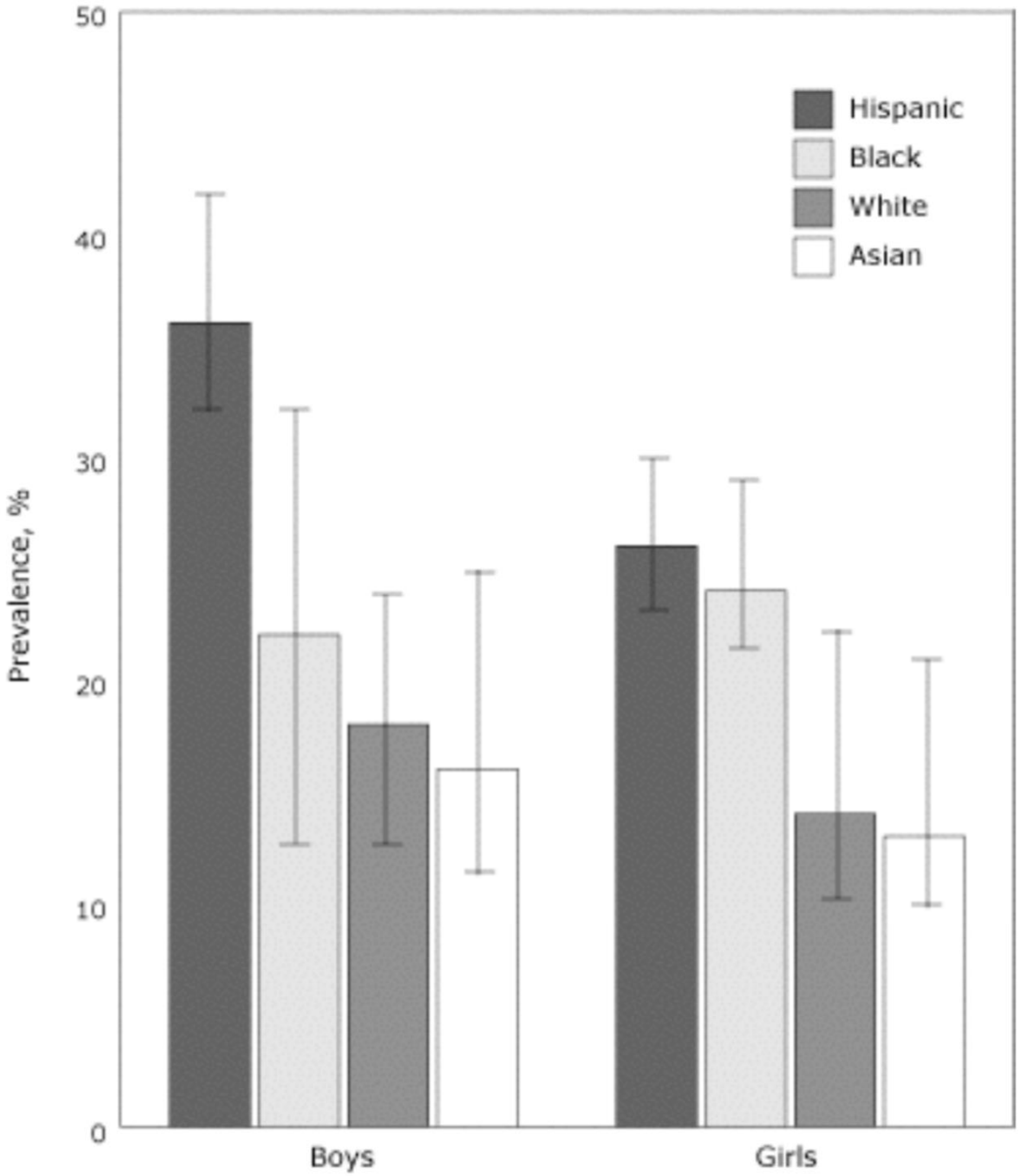


If s^2 is the sample variance from a random sample of n observations from a normal distribution with unknown variance σ^2, then **a 100(1 − α)% confidence interval on σ^2 is**

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \qquad (8\text{-}19)$$

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the upper and lower $100\alpha/2$ percentage points of the chi-square distribution with $n-1$ degrees of freedom, respectively. A **confidence interval for σ** has lower and upper limits that are the square roots of the corresponding limits in Equation 8-19.

Confidence estimates of the population proportion



Prevalence (with 95% CI bars) of obesity among New York City public elementary schoolchildren, by sex and race/ethnicity, 2003.

(source: CDC.GOV)

Collect a sample of BMI values

Obese means BMI>30

What do those bars actually mean?

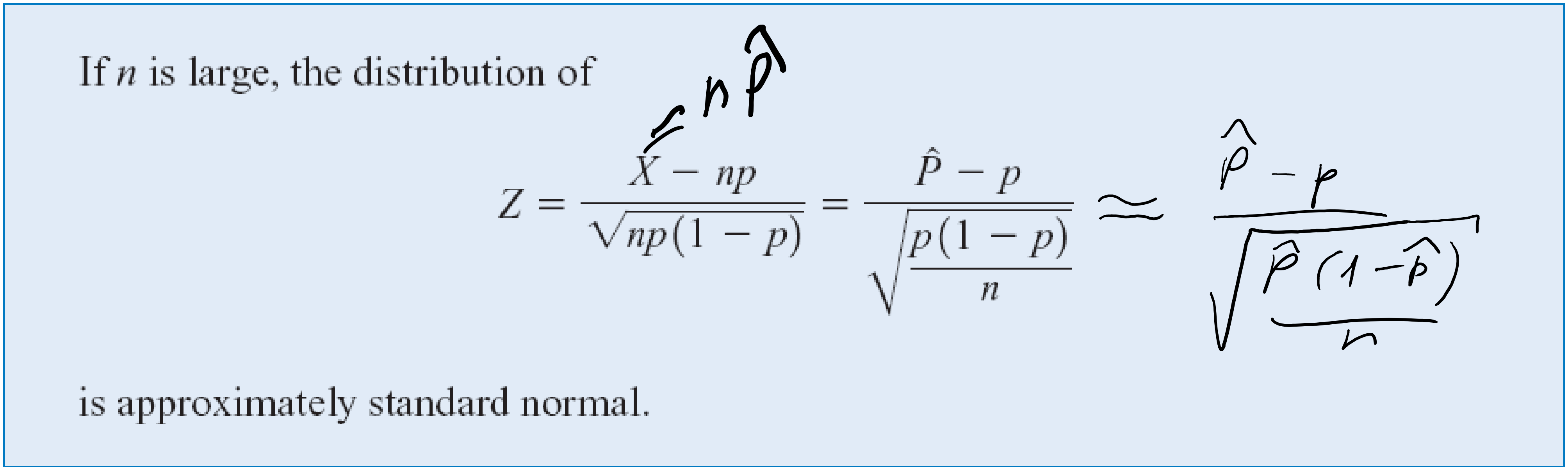
Large sample confidence estimate of population proportion

- Want to know the fraction *p* of the population that belongs to a class, e.g., the class “obese” kids defined by BMI>30.
- Each variable is a Bernoulli trial with one parameter p.
  We can use moments or MLE estimator to estimate p
- Both give the same estimate: sample fraction $\hat{p}$*=(# of obese kids in the sample)/(sample size n)*
- How to put confidence bounds on *p* based on $\hat{p}$
- # of obese kids in the sample follows the binomial distribution:
  “success” = sampled kid is obese : -(
  p – probability of success, *1-p* – failure
- Expected # of successes is *np* → Expected fraction of successes is *p*
- Standard deviation of # of successes is $\sqrt{np(1-p)}$→

Standard deviation of fraction of successes is $\sqrt{p(1-p)/n}$

8-5 A Large-Sample Confidence Interval For a Population Proportion

Normal Approximation for Binomial Proportion



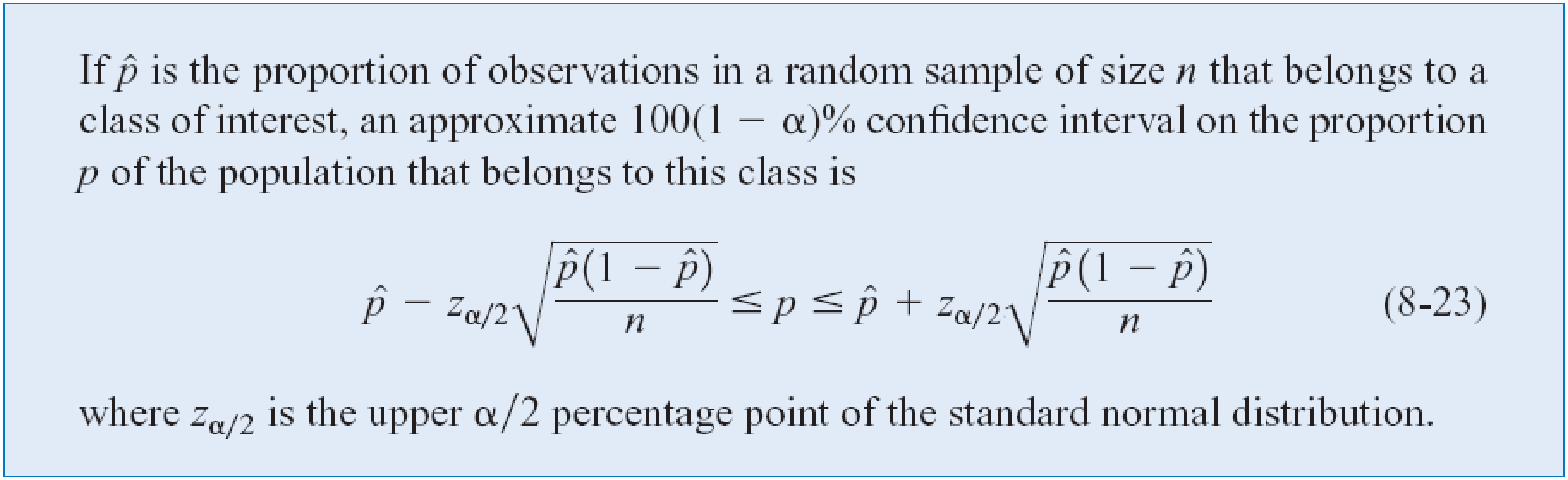
If n is large, the distribution of

$$Z = \frac{\overset{= n\hat{P}}{X} - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \simeq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

is approximately standard normal.

The quantity $\sqrt{\hat{p}(1-\hat{p})/n}$ is the standard error of the point estimator . $\hat{P}$

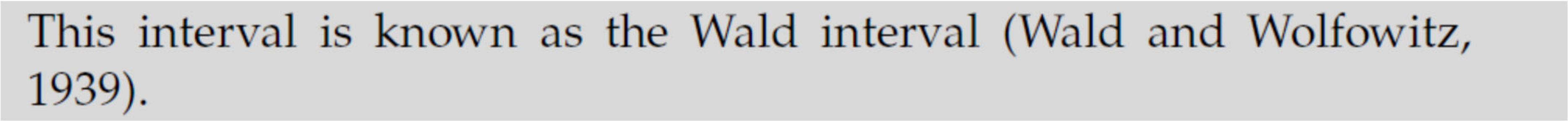
8-5 A Large-Sample Confidence Interval For a Population Proportion (Eq. 8-23)



If $\hat{p}$ is the proportion of observations in a random sample of size n that belongs to a class of interest, an approximate $100(1 - \alpha)\%$ confidence interval on the proportion p of the population that belongs to this class is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (8\text{-}23)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.



This interval is known as the Wald interval (Wald and Wolfowitz, 1939).

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

http://www.scientificameriken.com/candy5.asp

“To our surprise M&Ms met our demand to review their procedures in determining candy ratios. It is, however, noted that the figures presented in their email differ from the information provided from their website (http://us.mms.com/us/about/products/milkchocolate/). An email was sent back informing them of this fact. To which M&Ms corrected themselves with one last email:

In response to your email regarding M&M'S CHOCOLATE CANDIES

Thank you for your email.
On average, our new mix of colors for M&M'S® Chocolate Candies is:

M&M'S® Milk Chocolate: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown.

M&M'S® Peanut: 23% blue, 23% orange, 15% green, 15% yellow, 12% red, 12% brown.

M&M'S® Kids MINIS®: 25% blue, 25% orange, 12% green, 13% yellow, 12% red, 13% brown.

M&M'S® Crispy: 17% blue, 16% orange, 16% green, 17% yellow, 17% red, 17% brown.

M&M'S® Peanut Butter and Almond: 20% blue, 20% orange, 20% green, 20% yellow, 10% red, 10% brown.

Have a great day!

Your Friends at Masterfoods USA
A Division of Mars, Incorporated

How to estimate these probabilities from a finite sample and how to set confidence interval on these estimates?

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

How large is a sample needed for 95% CI
on the percentage of blue M&Ms to be less than +/- 4%

Same question for red M&Ms?

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

How large is a sample needed for 95% CI
on the percentage of blue M&Ms to be less than +/- 4%

Same question for red M&Ms?

For blue M&Ms $p = 0.24$

$$1.96\sqrt{\frac{0.24(1-0.24)}{n}} < 0.04$$

$$n > \left(\frac{1.96}{0.04}\right)^2 0.24 \times (1-0.24) = 438 \text{ M\&Ms or}$$

~ 2 × 7oz bags with 210 candies each

For red M&Ms $p = 0.13$

$$n > \left(\frac{1.96}{0.04}\right)^2 \times 0.13 \times (1-0.13) \simeq 271 \text{ M\&Ms or}$$

~ 1 × 7oz bag

Hypothesis testing:
one sample

Is P53 gene expressed at a lower level in cancer patients than in healthy people?

- We are interested if a P53 gene expression is **lowered** in population of cancer patients compared to the healthy population.

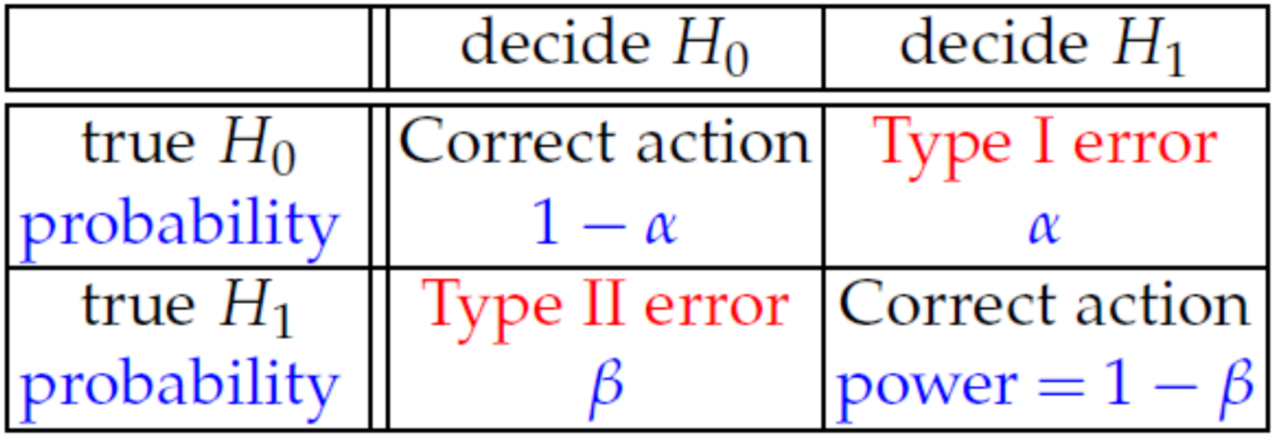
- We know that mean gene expression in the healthy population is μ_h=50 mRNAs/cell We are interested in deciding whether or not the mean expression in cancer population is **lower than** in healthy population. Let's call hypothesis H_1. Here H_1 is **one-sided**

- If we asked: cancer is not equal to healthy H_1 would be a **two-sided hypothesis**

- Assume we have a sample of 100 cancer patients with sample mean $\bar{x}$ =48 mRNAs/cell and standard deviation σ=10 mRNA/cell

- Can we use our sample to reject the "business as usual" or **null hypothesis H_0**: cancer = healthy and select **one-sided hypothesis H_1**: cancer < healthy

Two types of errors



	decide H_0	decide H_1
true H_0 probability	Correct action $1-\alpha$	Type I error α
true H_1 probability	Type II error β	Correct action power $= 1-\beta$

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

Sometimes the **type I error probability α** is called the **significance level,** or the **α-error**

Instructions: get α from your boss or PI (e.g., 5% or 1%)

Prob(H_0 is true given the sample data) < α
→reject H_0 and accept H_1

Prob(H_0 is true given the sample data) > α
→ accept H_0 and reject H_1

Type II error is much harder to estimate. Will deal with it later

P-Values of Hypothesis Tests

- P-value: what is the probability to get the observed value of sample mean of $\bar{x}$ =48 mRNAs/cell (or even smaller) and σ=10 mRNAs/cell in a healthy population with μ_h=50 mRNAs/cell

- If P-value is small – the null hypothesis is likely wrong and thus, the probability of making a type I error (incorrectly rejecting the null hypothesis) is small

- P-value answers the question: if I reject the null hypothesis H_0 based on the sample, what is the probability that I am making a type I error?

P-Value vs α in Hypothesis Testing

- Problem with using a predefined α: you don't know by how much you exceeded it
- Another approach is to calculate Prob(H_0 is true given the sample data) referred to as P-value.
  It the smallest α that would lead to rejection of null hypothesis
- You give your boss the P-value and let him/her decide if it is good enough
- Routinely with big datasets in genomics and systems biology P-values can be $10^{\text{-large number~10-100}}$. This number is used to judge the quality of the hypothesis

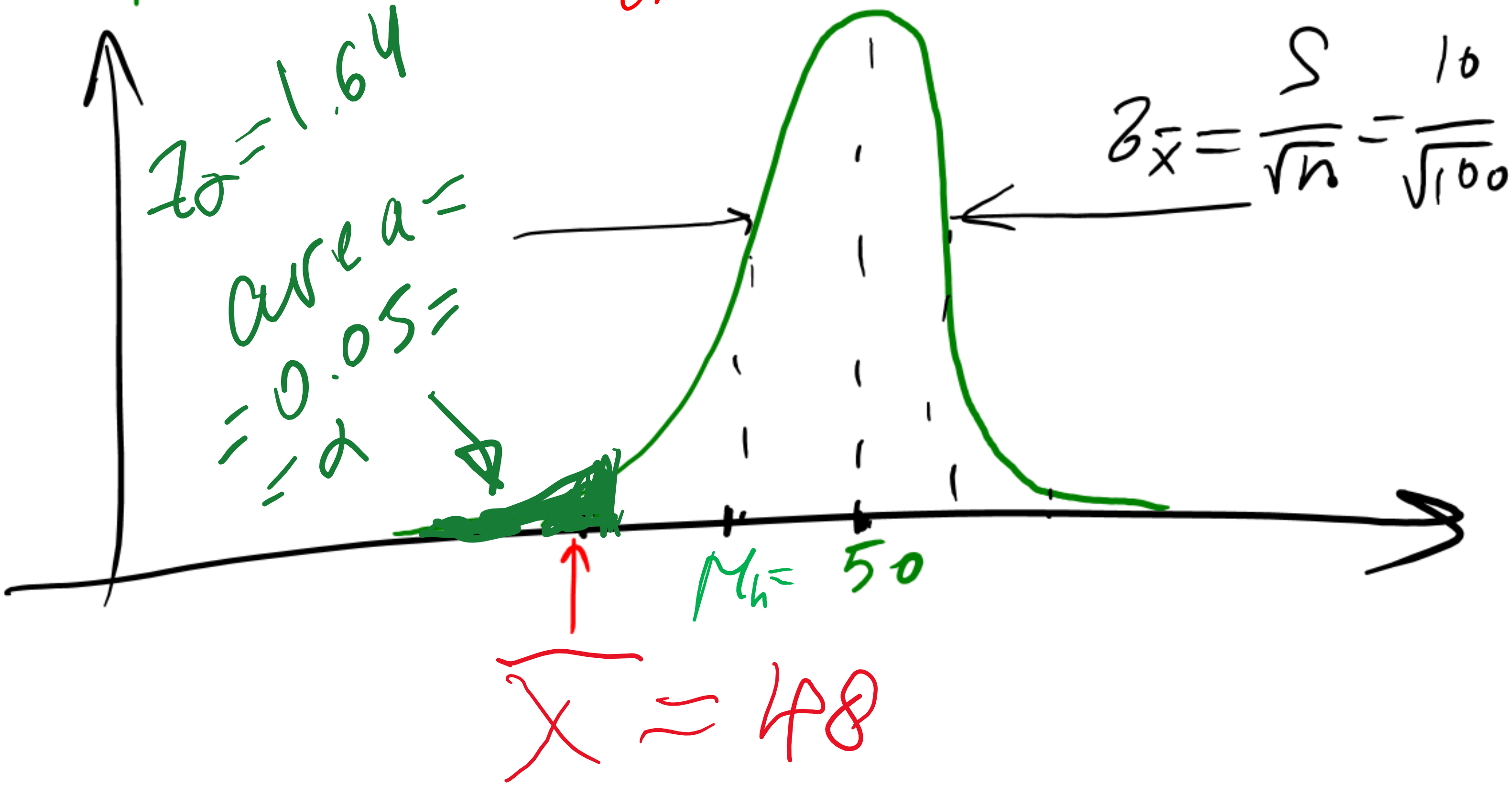
$\mu_h = 50$

$H_0: \mu_c = \mu_h$

$n=100, \ \bar{X}=48, \ s=10$

One-sided hypothesis $\quad H_1: \mu_c < \mu_h$

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$$



$$\text{P-value} = \text{Prob}(\bar{X}_h < 48 \mid H_0) = \cong 2.5\%$$

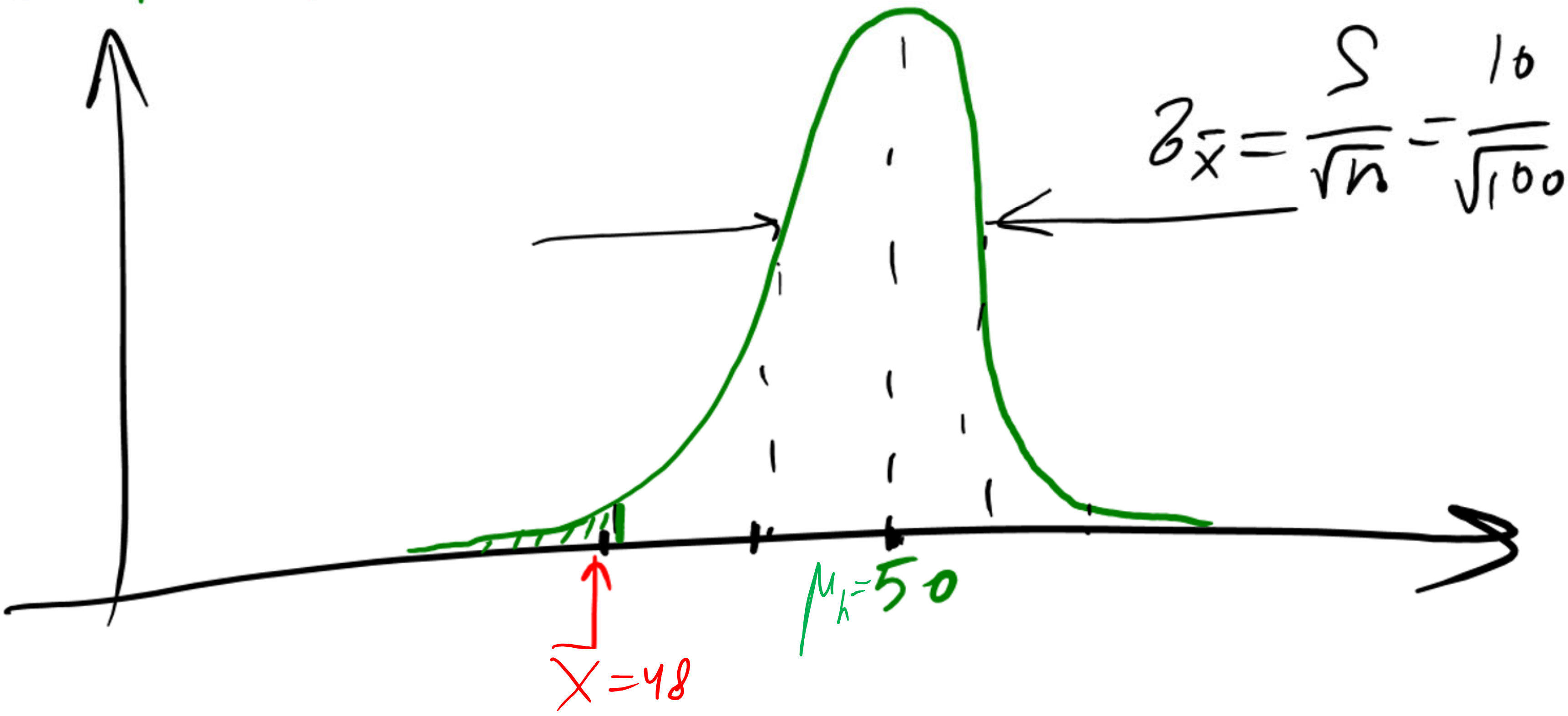
$\mu_h = 50$

$H_0: \mu_c = \mu_h$

$n = 100, \quad \bar{X} = 48, \quad s = 10$

$H_1: \mu_c < \mu_h$

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$$

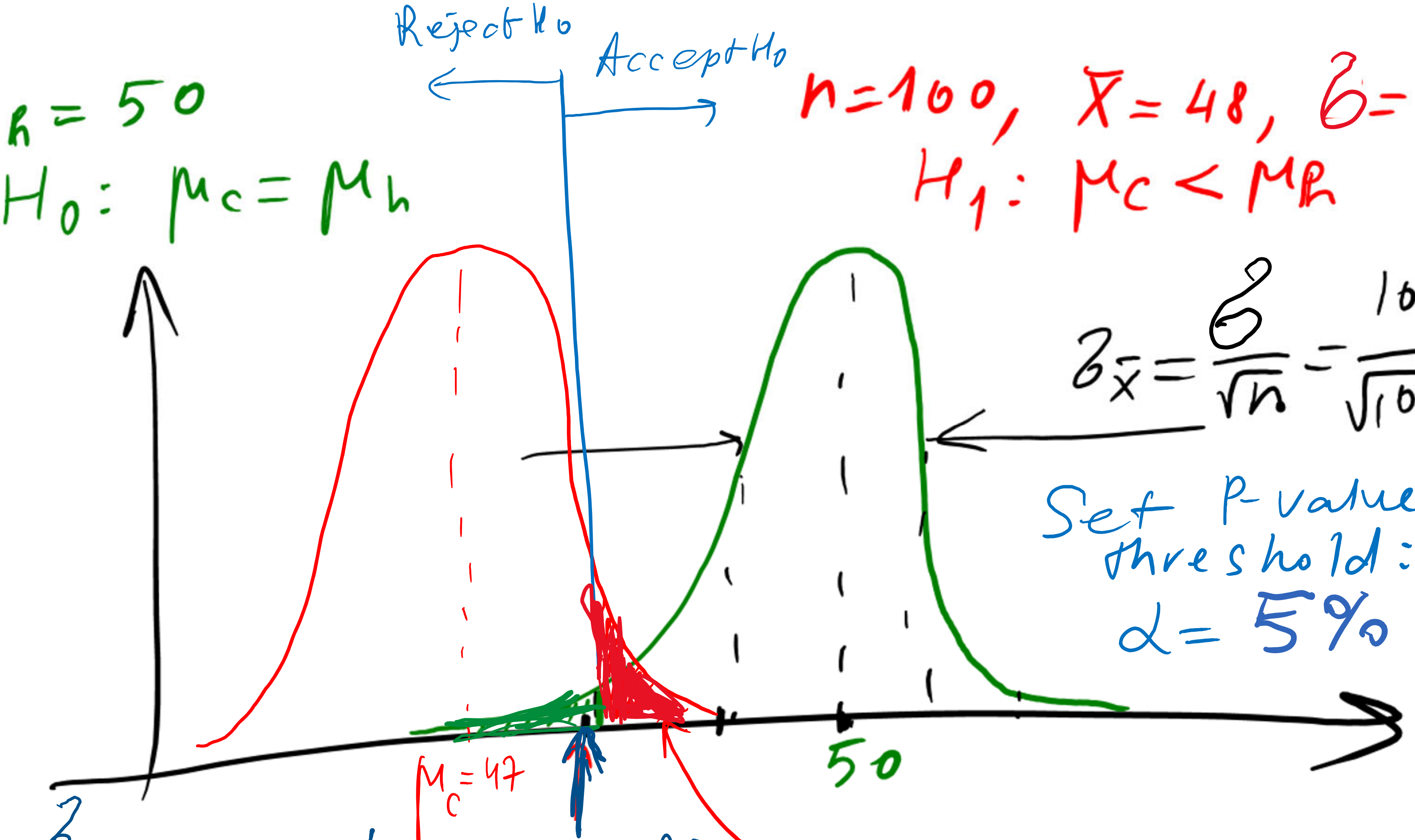


$\mu_h = 50$

$H_0: \mu_c = \mu_h$

$n = 100, \quad \bar{X} = 48, \quad \sigma = 10$

$H_1: \mu_c < \mu_h$



Type II error

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$$

Set P-value threshold:

$\alpha = 5\%$

$$\mu_h - Z_\alpha \frac{\sigma}{\sqrt{n}} = 50 - 1 \cdot 1.64 = 48.36$$

$$\beta = P(\text{Accept } H_0 \mid H_1 \text{ is true}) = \int_{48.36}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-47)^2}{2}\right) dx =$$

$$\alpha = 1 - \Phi(1.64) = 5\%$$

$$= 1 - \Phi(1.36) = 8.8\%$$

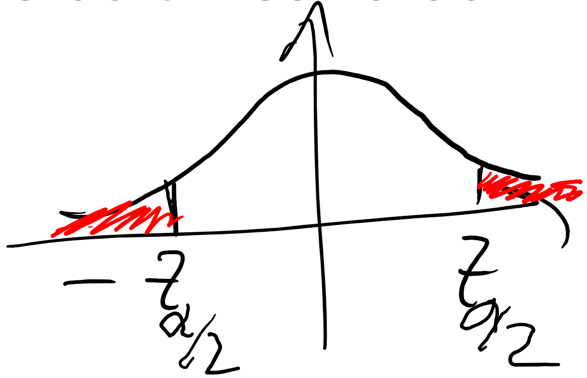
Generalizations

- What if H_1 is a two-sided hypothesis?
- A: P-value is *2(1-Φ(|Z|))*, where $Z=(\bar{X}-\mu_0)/[S/\sqrt{n}]$

  Compare it to: For one sized $\mu_1 > \mu_0$ it is *1-Φ(Z)*

  For one sized $\mu_1 < \mu_0$ it is *Φ(Z)*
- If α is given, use $\mu_0 +/- z_{\alpha/2}*S$ as thresholds

to reject the null hypothesis



- What if the sample size n is small (say n<10):
- A: Use t-distribution with n-1 degrees of freedom for 2-sided *P-value=2(1-CDF_Tdist(|T|))* where $T=(\bar{X}-\mu_0)/[S/\sqrt{n}]$.
- For a given α use $\mu_0 +/- t_{\alpha/2,n-1}\, T$ to reject the null hypothesis

Type II Error and Choice of Sample Size

Assume you know the minimum $\delta = |\mu_1 - \mu_0|$ that you care about.
What is the minimal sample you should use to separate H0 and H1 hypotheses if your tolerance to type I and type II errors is α and β ?

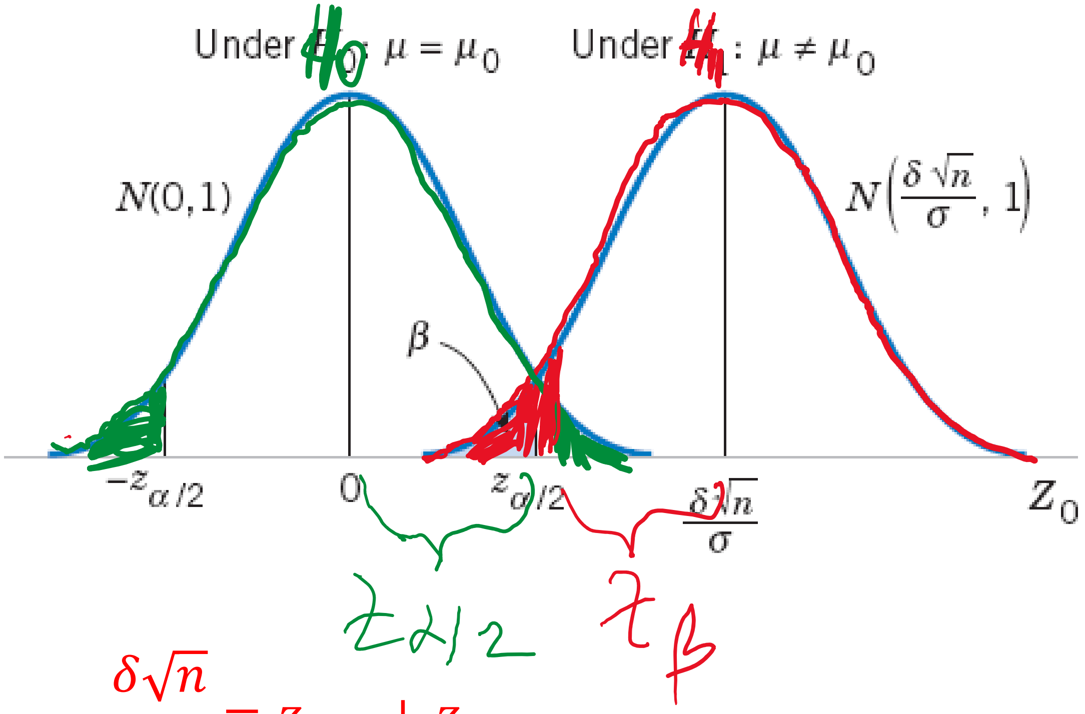


Figure 9-9 The distribution of Z_0 under H_0 and H_1.

$$\frac{\delta\sqrt{n}}{\sigma} = z_{\alpha/2} + z_\beta$$

$$n \simeq \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2} \qquad \text{where} \qquad \delta = \mu - \mu_0 \qquad (9\text{-}22)$$

Standard notation to indicate P-value with *, **, ***

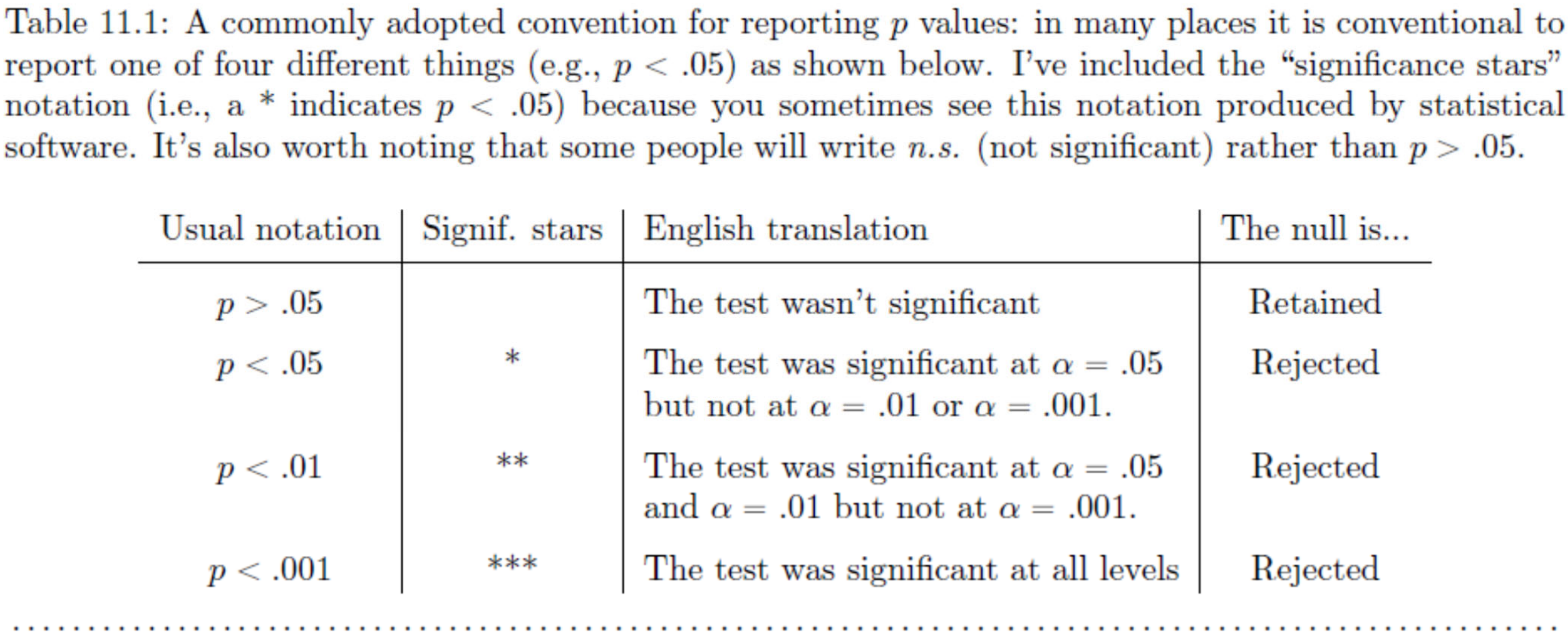
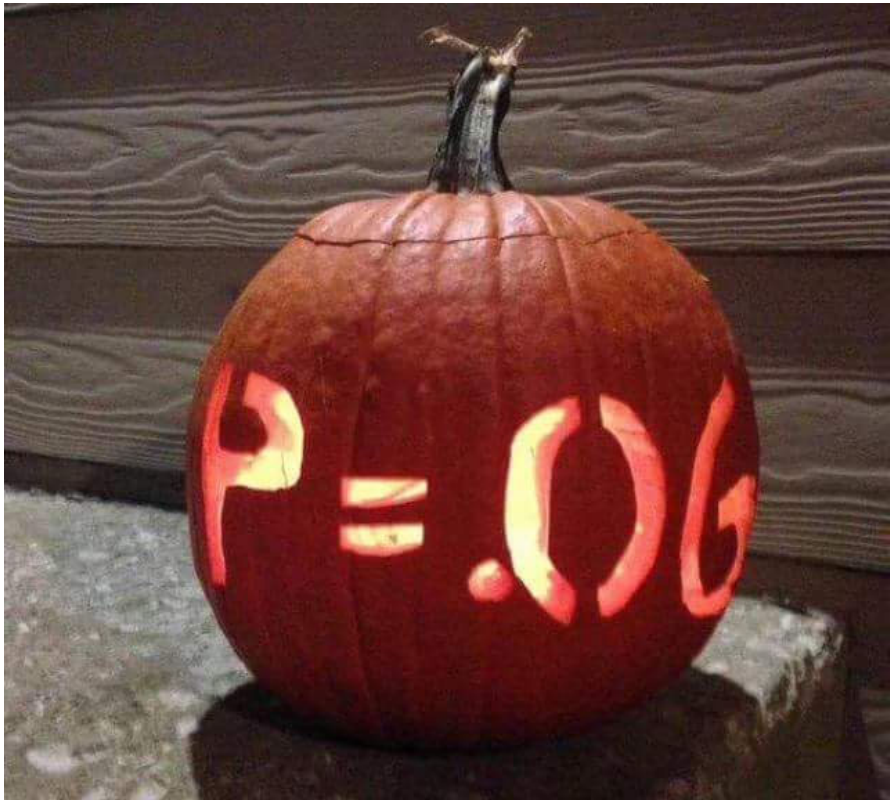


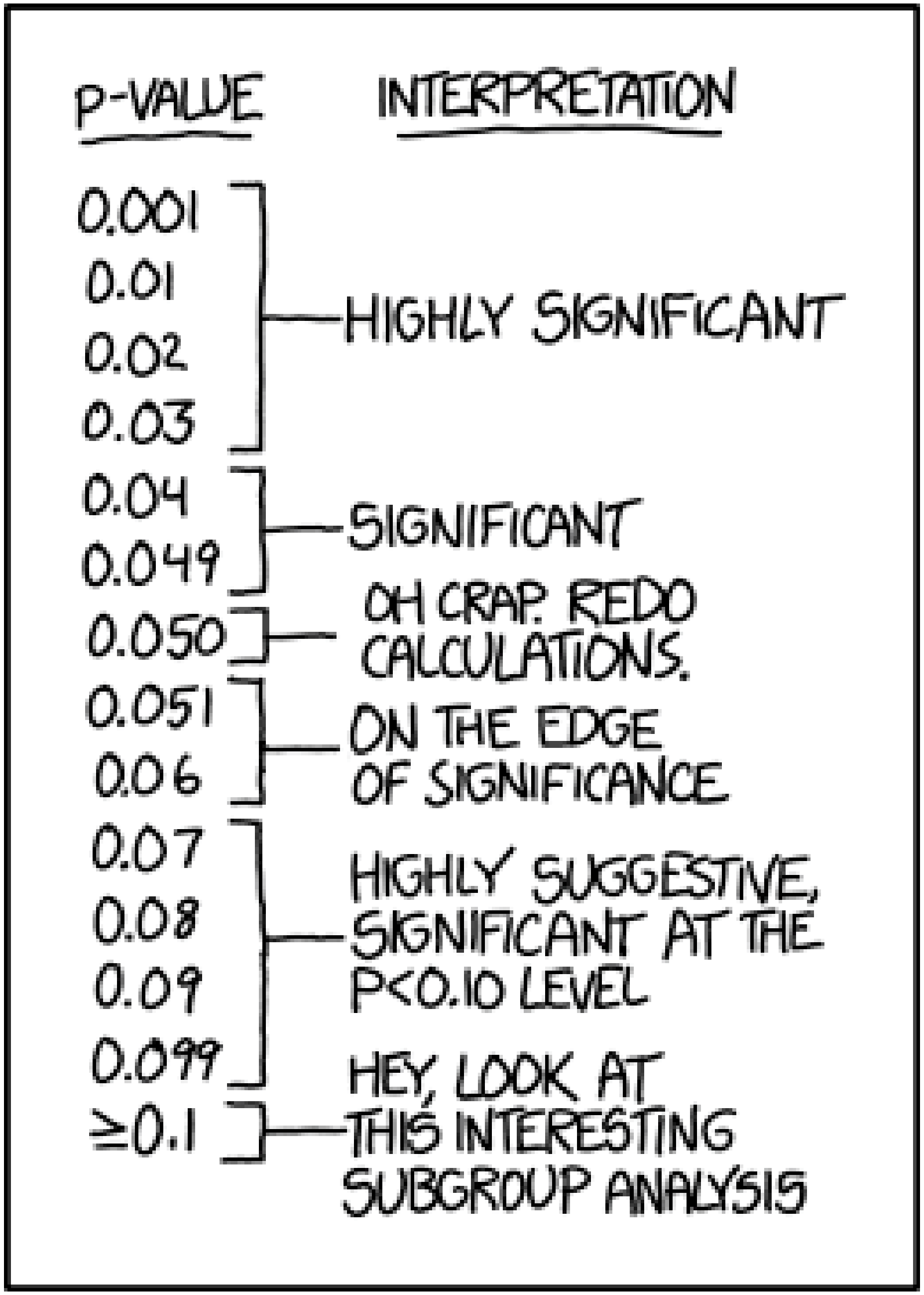
Table 11.1: A commonly adopted convention for reporting p values: in many places it is conventional to report one of four different things (e.g., $p < .05$) as shown below. I've included the "significance stars" notation (i.e., a * indicates $p < .05$) because you sometimes see this notation produced by statistical software. It's also worth noting that some people will write $n.s.$ (not significant) rather than $p > .05$.

Usual notation	Signif. stars	English translation	The null is...
$p > .05$		The test wasn't significant	Retained
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.	Rejected
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$.	Rejected
$p < .001$	***	The test was significant at all levels	Rejected



Happy
Halloween!
(belated)

Credit: Trust me,
I’m a “Biologist”
Facebook community



P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Credit: XKCD
comics

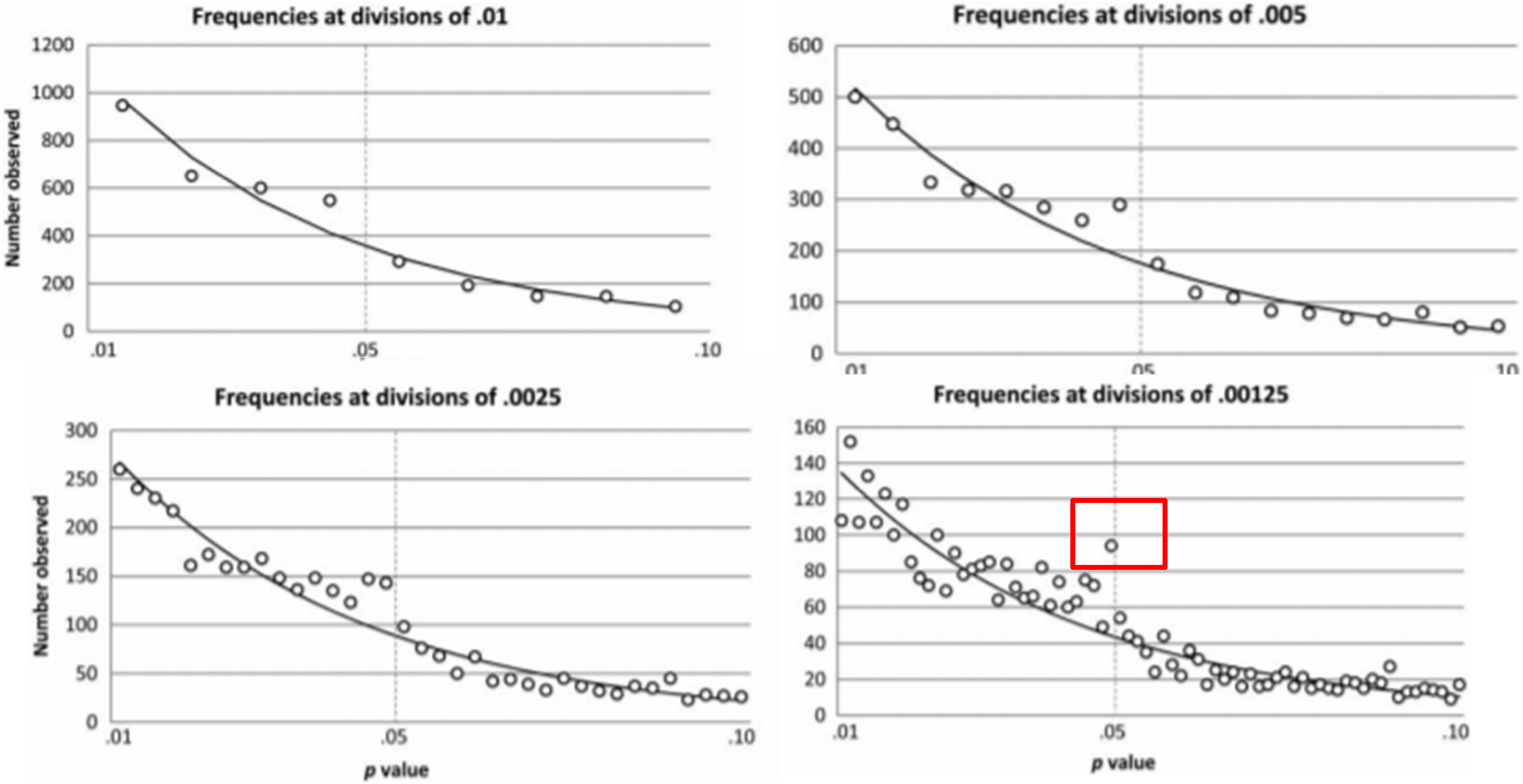
THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
2012, 65 (11), 2271–2279

Psychology Press
Taylor & Francis Group

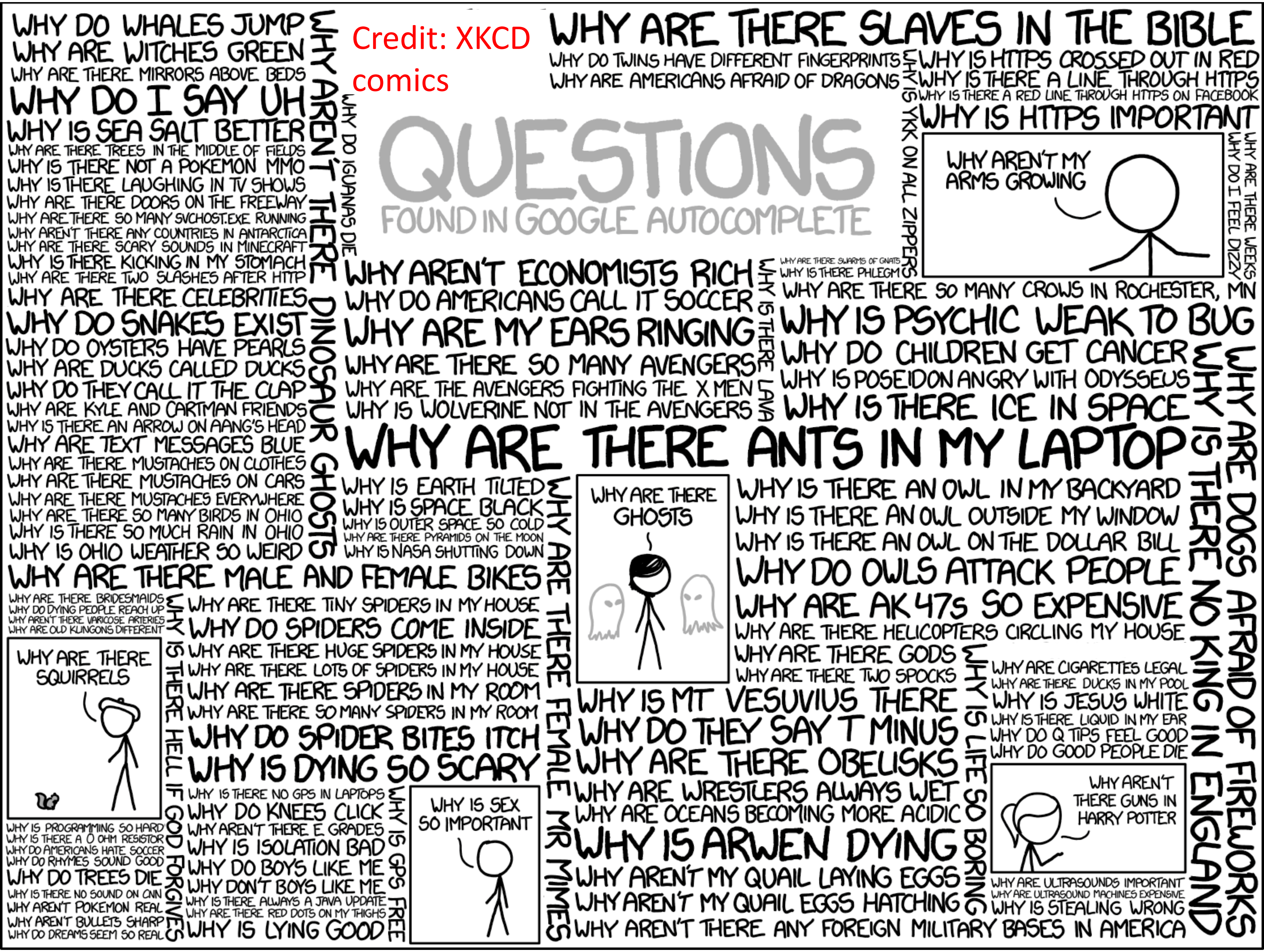
A peculiar prevalence of *p* values just below .05

E. J. Masicampo[1], and Daniel R. Lalande[2]

[1]Department of Psychology, Wake Forest University, Winston-Salem, NC, USA
[2]Department of Health Sciences, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada



Credit: XKCD comics


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT

Hypothesis testing:
two samples

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

Figure 10-1 Two independent populations.

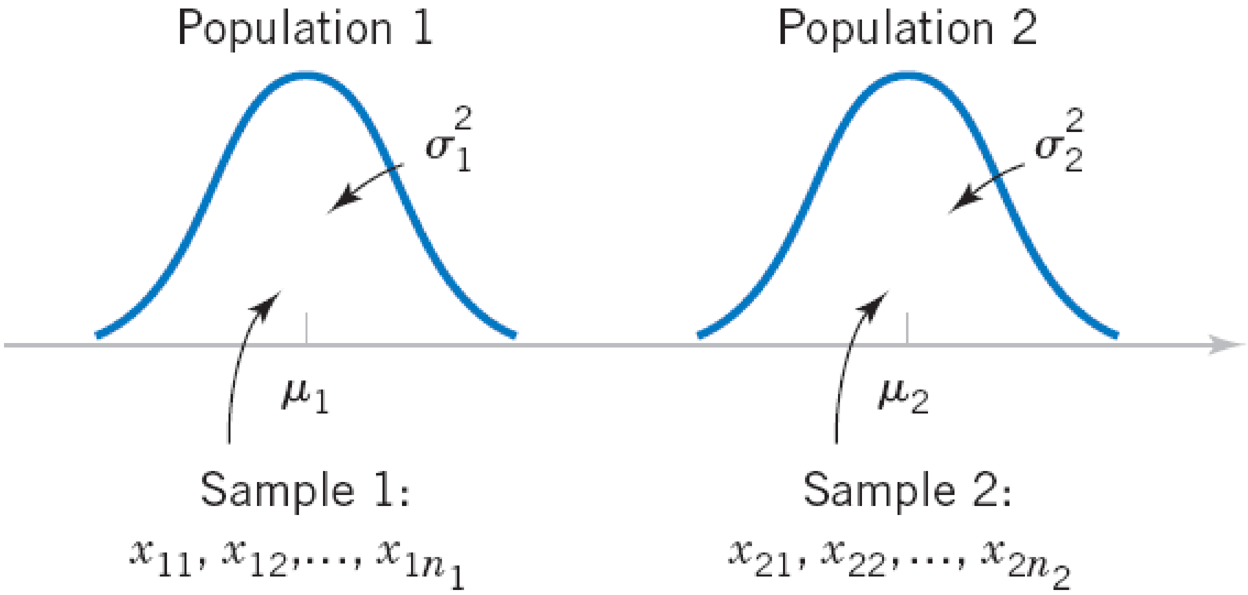


Figure 10-1 Two independent populations.

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

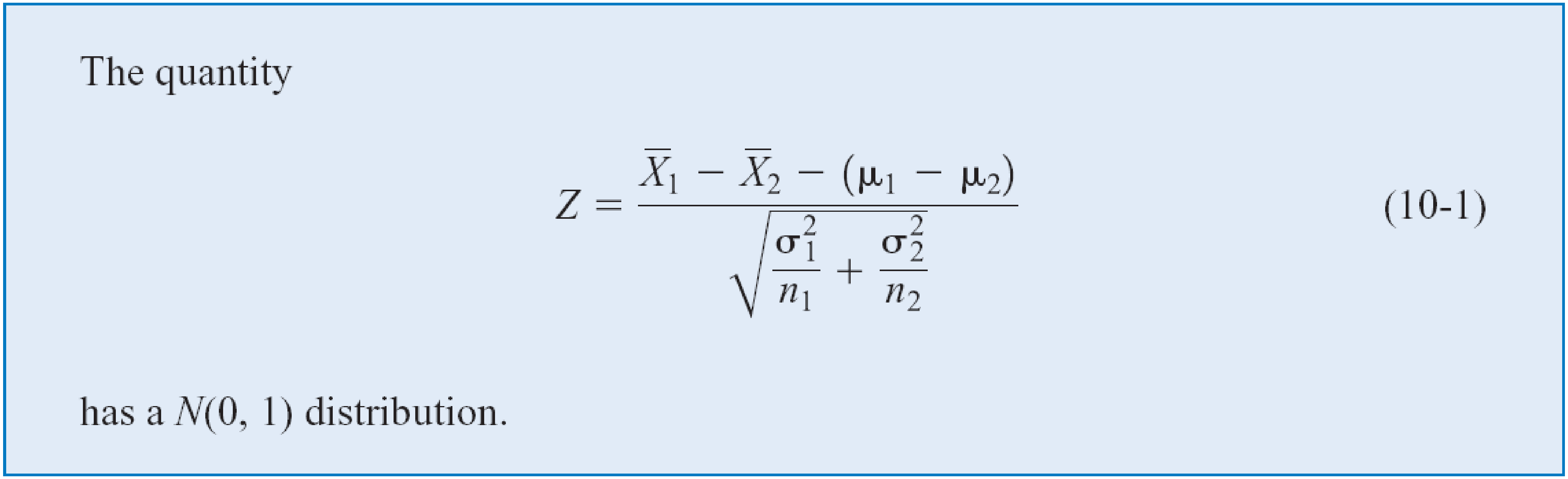
Assumptions

1. $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1.
2. $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample from population 2.
3. The two populations represented by X_1 and X_2 are independent.
4. Both populations are normal.

$$E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2$$

$$V(\overline{X}_1 - \overline{X}_2) = V(\overline{X}_1) + V(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known



The quantity

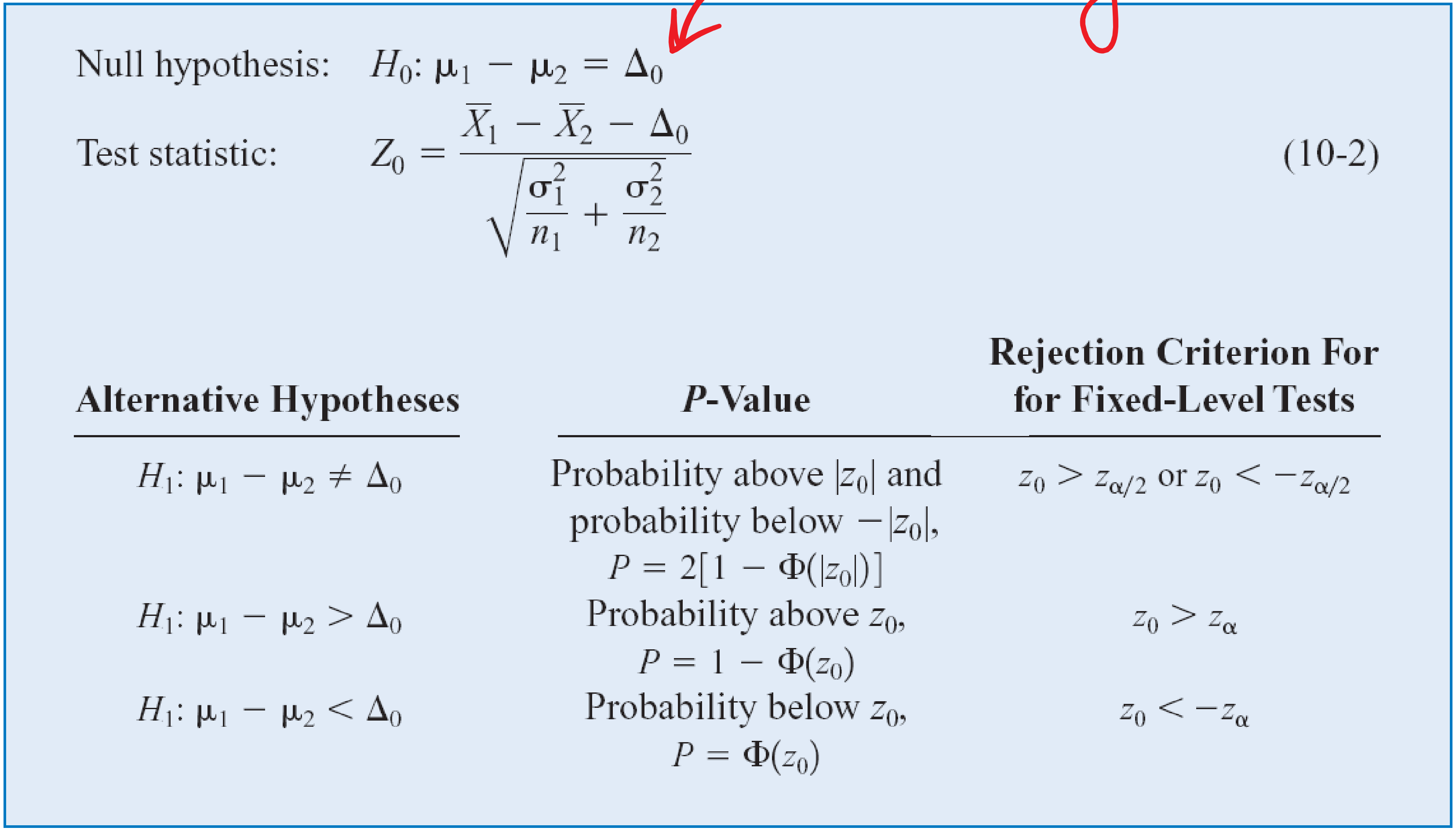
$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{10-1}$$

has a $N(0, 1)$ distribution.

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

10-2.1 Hypothesis Tests for a Difference in Means, Variances Known

usually $\Delta_0 = 0$



Null hypothesis: $H_0\colon \mu_1 - \mu_2 = \Delta_0$

Test statistic:
$$Z_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad (10\text{-}2)$$

Alternative Hypotheses	*P*-Value	Rejection Criterion For for Fixed-Level Tests
$H_1\colon \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert z_0\rvert$ and probability below $-\lvert z_0\rvert$, $P = 2[1 - \Phi(\lvert z_0\rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1\colon \mu_1 - \mu_2 > \Delta_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1\colon \mu_1 - \mu_2 < \Delta_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

10-2.1 Hypotheses Tests on the Difference in Means, Variances Unknown

Case 2: $\sigma_1^2 \neq \sigma_2^2$

If H_0: $\mu_1 - \mu_2 = \Delta_0$ is true, the statistic

$$T_0^* = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad (10\text{-}15)$$

is distributed as t-distribution with degrees of freedom given by

$$v = n_1 + n_2 - 2,$$

or more generally

Multiple null hypotheses: Bonferroni correction

- What if you have **m** independent null hypotheses? Say you have **m=25,000 genes** in a genome?
- What is the probability that at least one of the null-hypotheses will be shown to
  be false at significance threshold α_1?
- Answer:
  Family-Wise Error Rate
  or $FWER=1-(1- \alpha_1)^m \approx m\alpha_1$
- If $m=20$ and $\alpha_1=0.05$,
  $FWER= 0.6415$
- If you want to get $FWER< \alpha$, use
  $\alpha_1= \alpha/m$

Carlo Emilio Bonferroni
(1892 –1960)
Italian mathematician
who worked on
probability theory.

Example 10-7 **Chocolate and Cardiovascular Health** An article in *Nature* (2003, Vol. 424 ~~48,~~ p. 1013) described an

chocolate c
In the expe
late per day
consisted o
average bo

Is there ev
plasma ant

Vol. 424

Plasma antioxidants from chocolate

Dark chocolate may offer its consumers health benefits the milk variety cannot match.

There is some speculation that dietary flavonoids from chocolate, in particular (−)epicatechin, may promote cardiovascular health as a result of direct antioxidant effects or through antithrombotic mechanisms[1–3]. Here we show that consumption of plain, dark chocolate (Fig. 1) results in an increase in both the total antioxidant capacity and the (−)epicatechin content of blood plasma, but that these effects are markedly reduced when the chocolate is consumed with milk or if milk is incorporated as milk chocolate. Our findings indicate that milk may interfere with the absorption of antioxidants from chocolate *in vivo* and may therefore negate the potential health benefits that can be derived from eating moderate amounts of dark chocolate.

To determine the antioxidant content of different chocolate varieties, we took dark chocolate and milk chocolate prepared from the same batch of cocoa beans and defatted them twice with *n*-hexane before extracting them with a mixture of water, acetone and acetic acid (70.0:29.8:0.2 by volume). We measured their *in vitro* total antioxidant capacities using the ferric-reducing antioxidant potential (FRAP) assay[4]; FRAP reduced iron per 100 g for dark and milk chocolate, respectively. Volunteers must therefore consume twice as much milk chocolate as dark chocolate to receive a similar intake of antioxidants.

We recruited 12 healthy volunteers (7 women and 5 men with an average age of 32.2 ± 1.0 years (range, 25–35 years). Subjects were non-smokers, had normal blood lipid levels, were taking no drugs or vitamin supplements, and had an average weight of 65.8 ± 3.1 kg (range, 46.0–86.0 kg) and body-mass index of 21.9 ± 0.4 kg m^{-2} (range, 18.6–23.6 kg m^{-2}). On different days, following a crossover experimental design, subjects consumed 100 g dark chocolate, 100 g dark chocolate with 200 ml full-fat milk, or 200 g milk chocolate (containing the equivalent of up to 40 ml milk).

One hour after subjects had ingested the chocolate, or chocolate and milk, we measured the total antioxidant capacity of their plasma by FRAP assay. Plasma antioxidant levels increased significantly after consumption of dark chocolate alone, from 100 ± 3.5% to 118.4 ± 3.5% (*t*-test, $P < 0.001$), returning to baseline values (95.4 ± 3.6%) after 4 h (Fig. 2a). There was

TION.COM/ALAMY

Mauro Serafini*, Rossana Bugianesi*, Giuseppe Maiani*, Silvia Valtuena*, Simone De Santis*, Alan Crozier†
**Antioxidant Research Laboratory, Unit of Human Nutrition, National Institute for Food and Nutrition Research, Via Ardeatina 546, 00178 Rome, Italy*
e-mail: serafini@inran.it
†*Plant Products and Human Nutrition Group, Graham Kerr Building, Division of Biochemistry and Molecular Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK*

Figure 1 Stack of benefits? Unlike its milky counterpart, dark chocolate may provide more than just a treat for the tastebuds.

could be due to the formation of secondary bonds between chocolate flavonoids and milk proteins[6,7], which would reduce the biological accessibility of the flavonoids and therefore the chocolate's potential antioxidant properties *in vivo*.

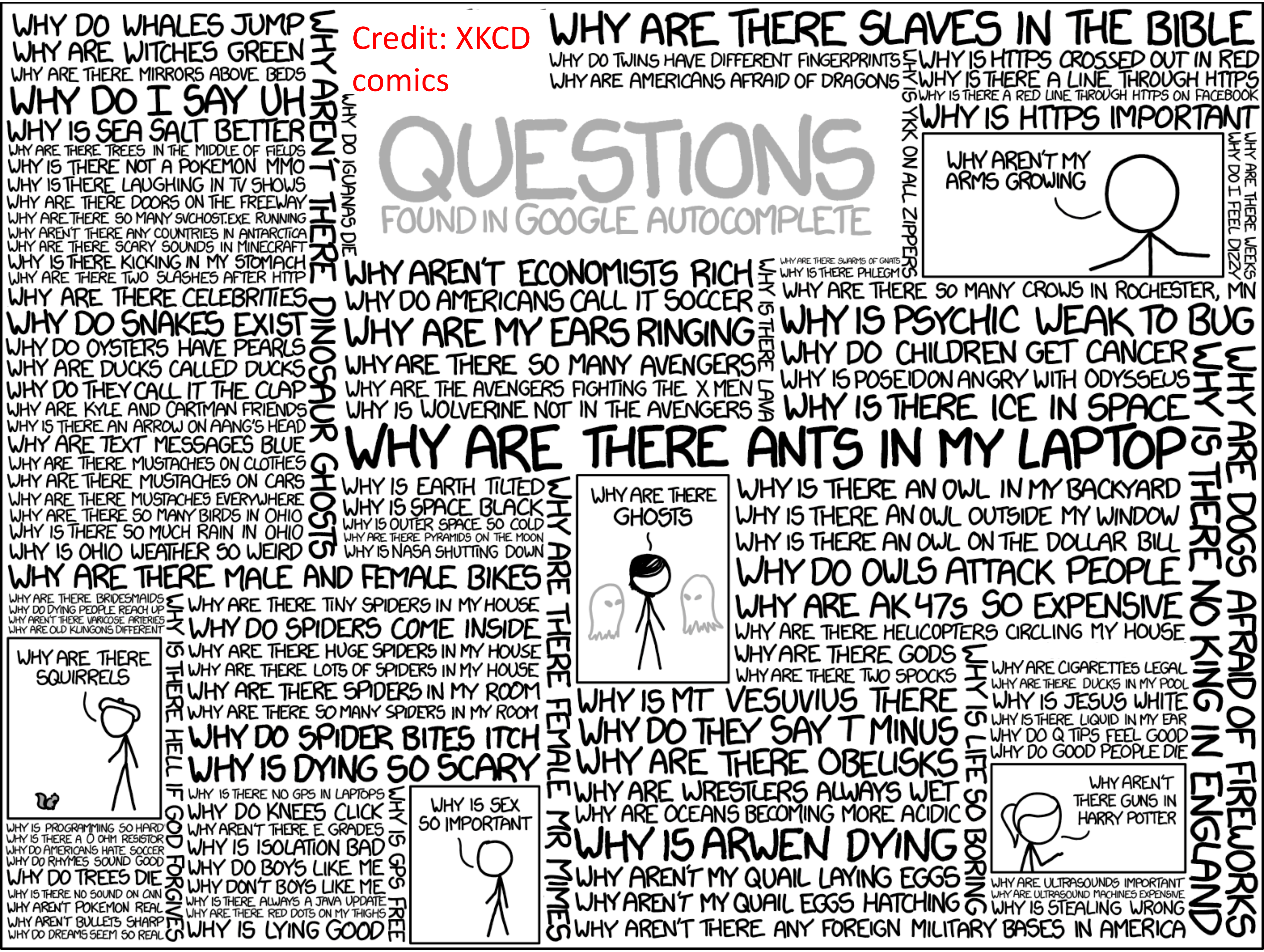
Our findings highlight the possibility

Sweet matlab exercise #1

- **Download dark_vs_milk_chocolate_analysis_template.m at the course website. Correct all ?? In the file**
- **dark=[118.8 122.6 115.6 113.6 119.5 115.9 115.8 115.1 116.9 115.4 115.6 107.9];**
- **milk=[102.1 105.8 99.6 102.7 98.8 100.9 102.8 98.7 94.7 97.8 99.7 98.6]**
- **Use Z-statistics to calculate P-value of the null hypothesis H_0 that milk = dark against H_1 that dark > milk. P_value_z=2*[1-normcdf(|Z|)]**
- **Repeat using T-statistics. # of degrees of freedom is dof=2*(n-1)**
  P_value_t=2*tcdf(|T|, dof)

Sweet matlab exercise #1

- **dark=[118.8 122.6 115.6 113.6 119.5 115.9 115.8 115.1 116.9 115.4 115.6 107.9];**
- **milk=[102.1 105.8 99.6 102.7 98.8 100.9 102.8 98.7 94.7 97.8 99.7 98.6]**
- **x_dark=mean(dark) % sample mean dark chocolate**
- **x_milk=mean(milk) % sample mean milk chocolate**
- **s_dark=std(dark) % sample std dark chocolate**
- **s_milk=std(milk) % sample std milk chocolate**
- **n=12 % sample size of both dark and milk**
- **std_xdiff=sqrt(s_dark.^2./2+s_milk.^2./n) % std diff x**
- **z_stat=(x_dark-x_milk)./std_xdiff % z-statistic**
- **P_value_z=erfc(z_stat./sqrt(2))./2 % P-value of null true**
- **% P_value_z=9.9629e-34**
- **dof=(n-1)+(n-1) % # of degrees of freedom**
- **P_value_t=tcdf(z_stat,dof,'upper') % P-value of null true**
- **%P_value_t= 1.8417e-11**


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Credit: XKCD comics

Regression analysis
Two variables

(Montgomery and Runger: ch 11
Brani Vidakovic: ch 14)

Reminder

Covariance Defined

Covariance is a number qunatifying average dependence between two random variables.

The covariance between the random variables X and Y, denoted as $\operatorname{cov}(X,Y)$ or σ_{XY} is

$$\sigma_{XY} = E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right] = E\left(XY\right)-\mu_X\mu_Y \qquad (5\text{-}14)$$

The units of σ_{XY} are units of X times units of Y.

Unlike the range of variance, $-\infty < \sigma_{XY} < \infty$.

Correlation is “normalized covariance”

- Also called:
  Pearson correlation coefficient

  $\rho_{XY}=\sigma_{XY}/\sigma_X\sigma_Y$
  is the covariance normalized to be $-1 \le \rho_{XY} \le 1$

Karl Pearson (1852– 1936)
English mathematician and biostatistician

Covariance and Scatter Patterns

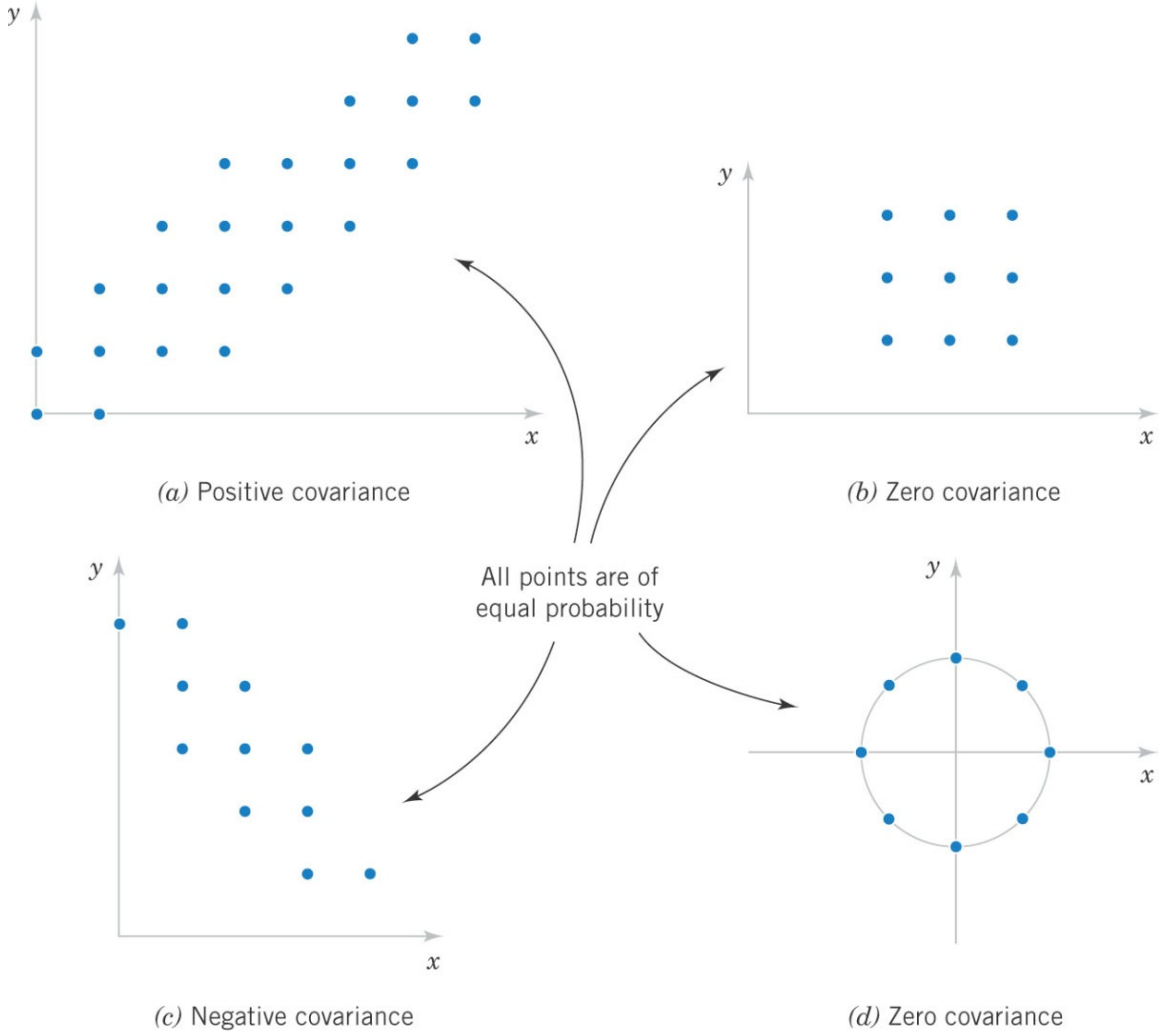


Figure 5-13 Joint probability distributions and the sign of cov(*X*, *Y*). Note that covariance is a measure of linear relationship. Variables with non-zero covariance are correlated.

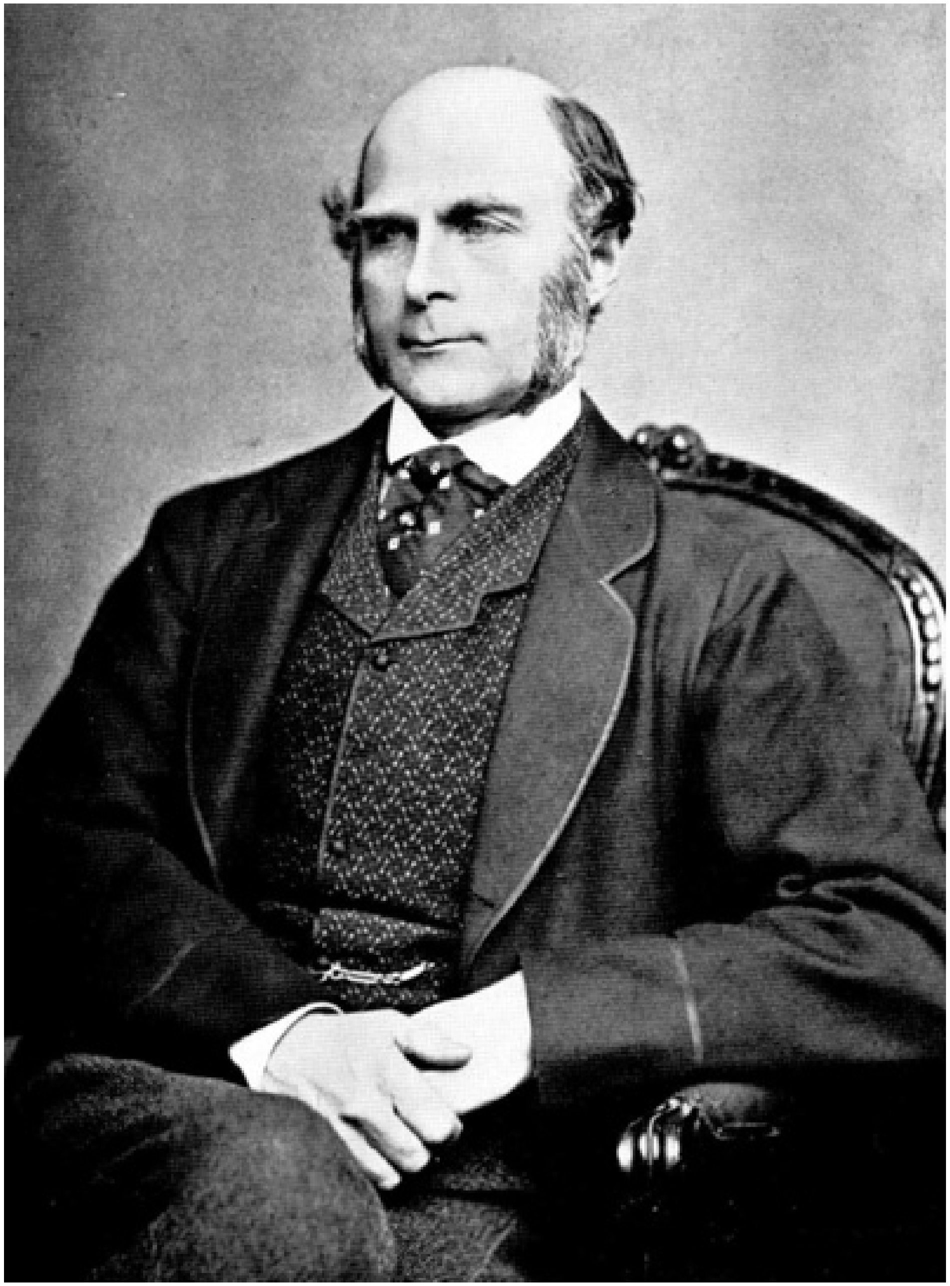
Regression analysis

- Many problems in engineering and science involve sample in which two or more variables were measured. They may not be independent from each other and one (or several) of them can be used to predict another

- Everyday example: in most samples height and weight of people are related to each other

- Biological example: in a cell sorting experiment the copy number of a protein may be measured alongside its volume

- Regression analysis uses a sample to build a model to predict protein copy number given a cell volume

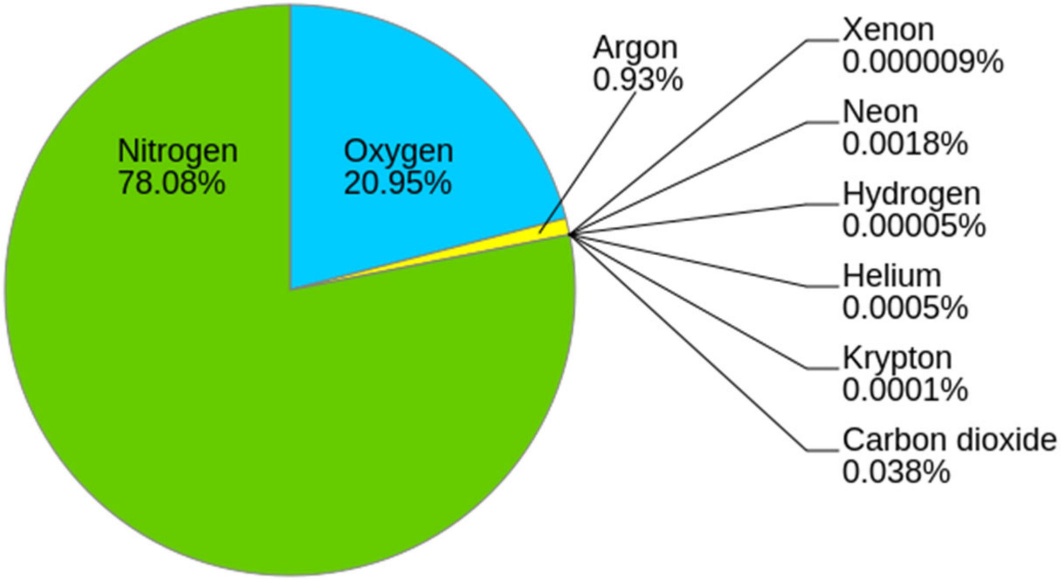


Sir Francis Galton,
(1822 -1911) was an English statistician, anthropologist, proto-geneticist, psychometrician, eugenicist, ("Nature vs Nurture", inheritance of intelligence), tropical explorer, geographer, inventor (Galton Whistle to test hearing), meteorologist (weather map, anticyclone).

Invented both correlation and regression analysis when studied heights of fathers and sons

Found that fathers with height above average tend to have sons with height also above average but closer to the average.
Hence "regression" to the mean

Two variable samples



- Oxygen can be distilled from the air
- Hydrocarbons need to be filtered out or the whole thing would go kaboom!!!
- When more hydrocarbons were removed, the remaining oxygen stays cleaner
- Except we don't know how dirty was the air to begin with

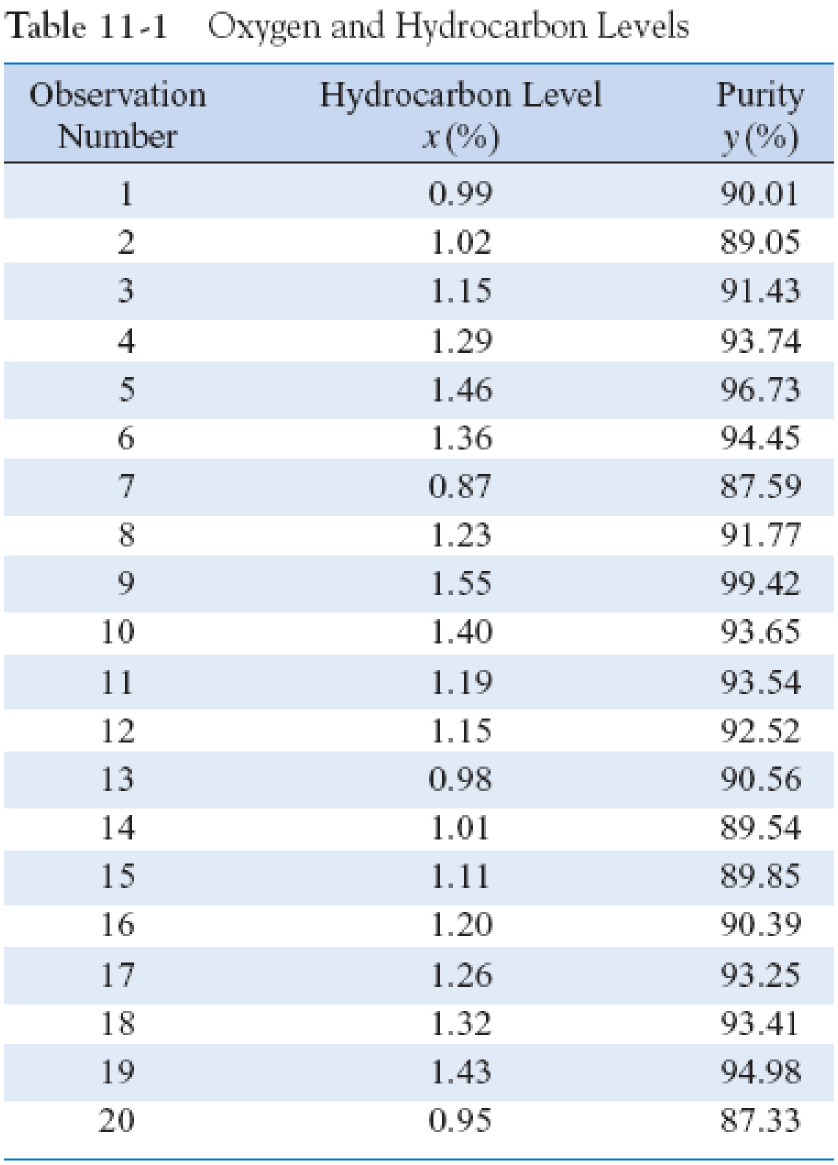


Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

$$Y = \beta_0 + \beta_1 X + \epsilon$$

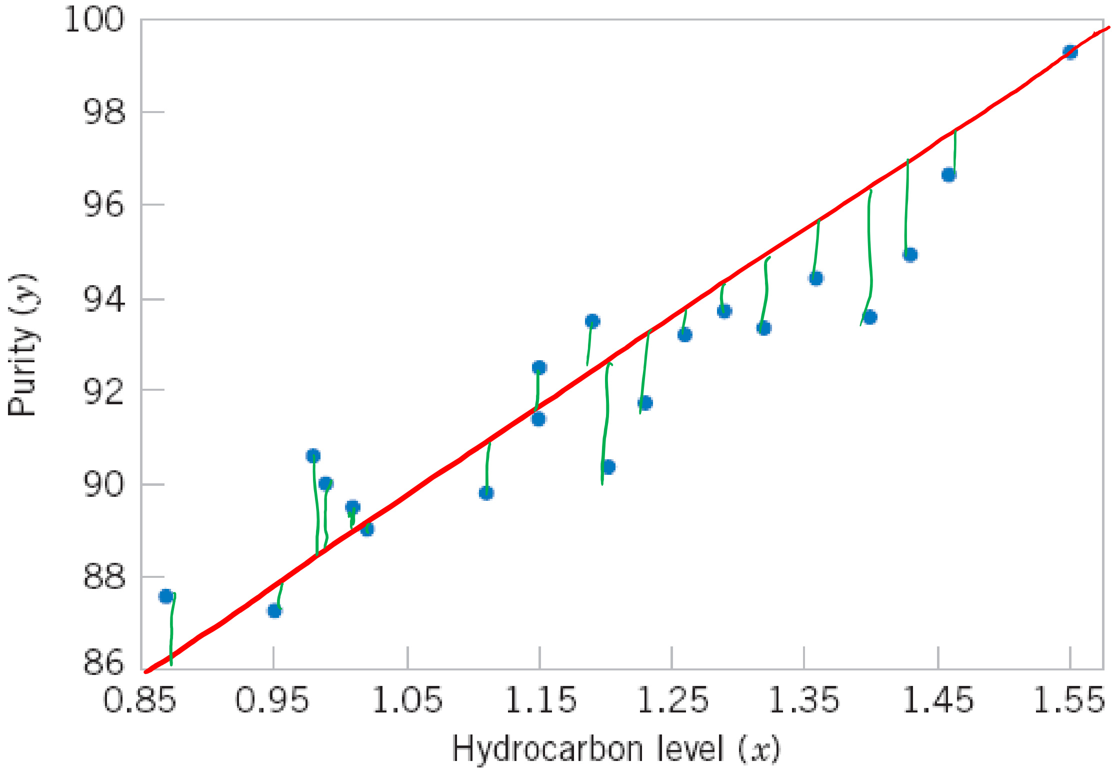


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

$$Y = 75 + 15 \cdot X + \epsilon$$

Linear regression

The **simple linear regression model** is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

ε is the **random error term**

slope β_1 and intercept β_0 of the line are called **regression coefficients**

Note: Y, X and ε are random variables
The minimal assumption: $E(\varepsilon \mid x)=0$ →
$E(Y \mid x) = \beta_0 + \beta_1 x + E(\varepsilon \mid x) = \beta_0 + \beta_1 x$

$$Y=\beta_0+\beta_1 X+\epsilon \;;\quad E(\epsilon|x)=0 \;\; \forall x$$

How does one find β_0 & β_1?

$$Cov(Y;X)=Cov((\beta_0+\beta_1 X+\epsilon);X)=$$

$$=\cancel{Cov(\beta_0;X)}+\beta_1 Cov(X,X)+\cancel{Cov(\epsilon;X)}$$

$Cov(\beta_0,X)=0$ since β_0 is constant

$$Cov(X,X)=E(X^2)-E(X)^2=Var(X)$$

$$Cov(\epsilon,X)=E(\epsilon\cdot X)-E\cancel{(\epsilon)}\cdot E(X)=$$

$$=E(\epsilon\cdot X)=\sum_{\text{all } x} x\cdot E(\cancel{\epsilon|}x)=0$$

Thus $\boxed{\beta_1=\frac{Cov(X,Y)}{Var(X)}}$; $\boxed{\beta_0=E(Y)-\beta_1 E(X)}$

Method of least squares

- The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

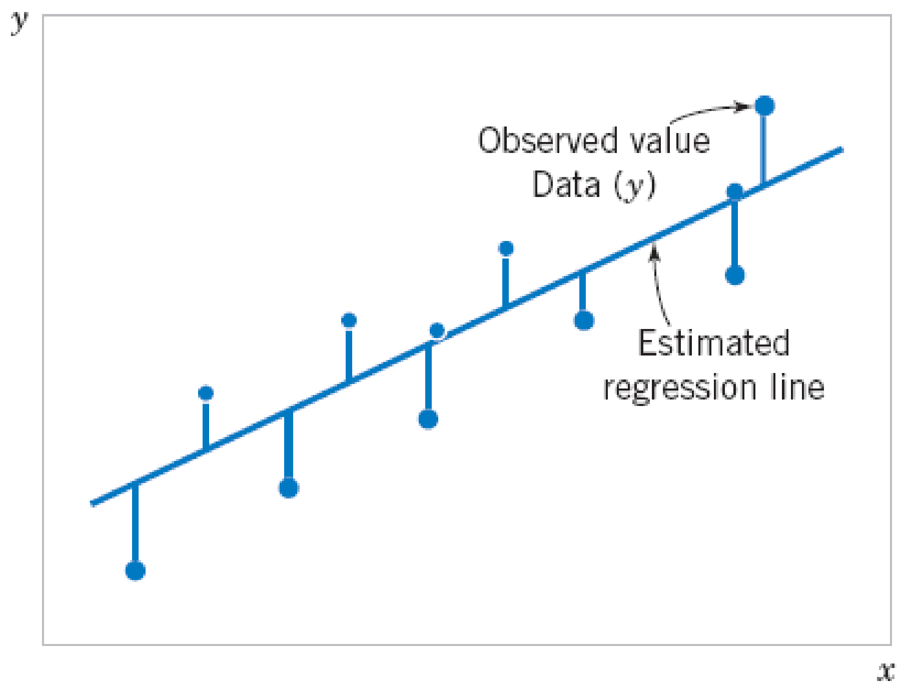
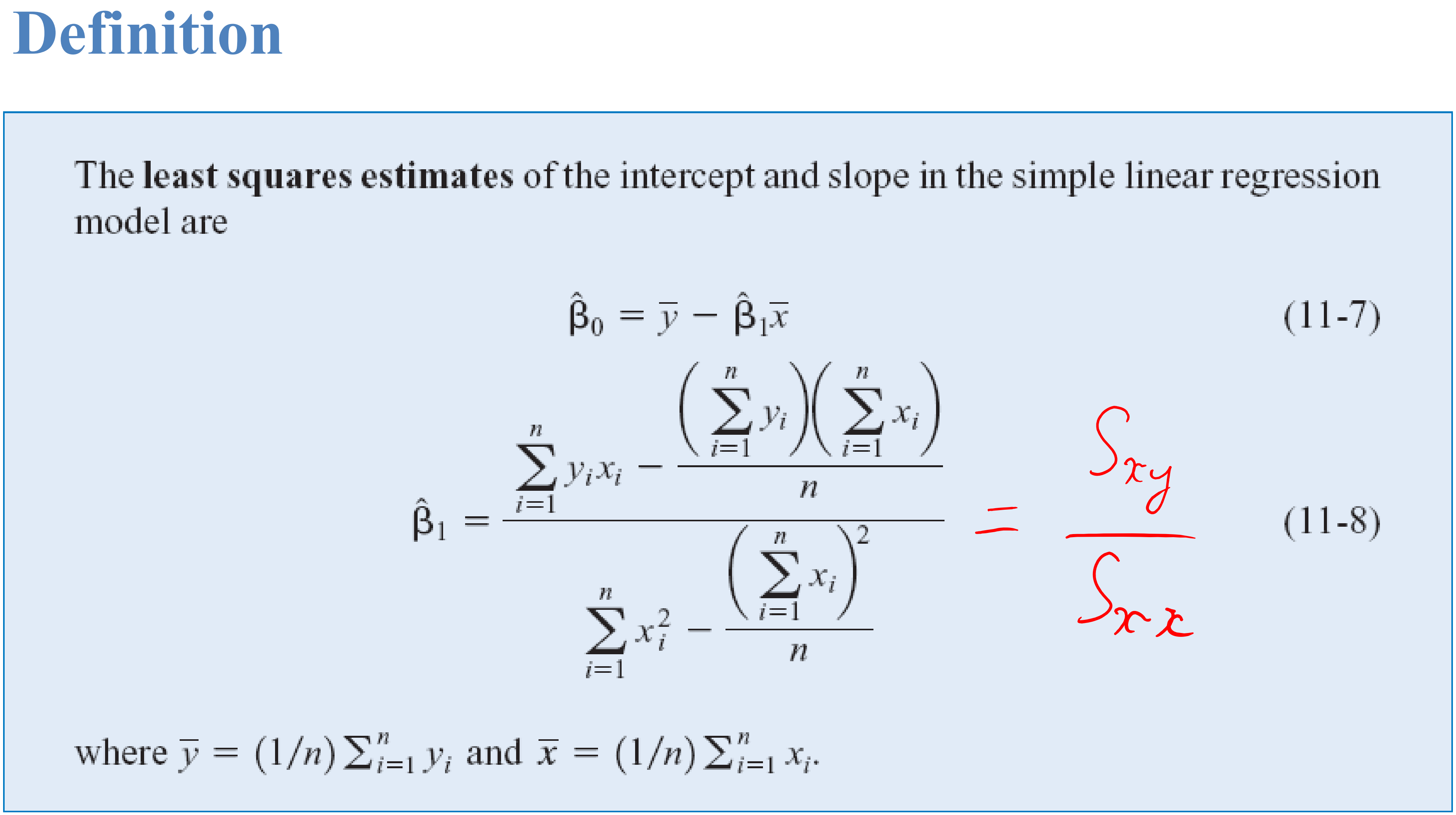


Figure 11-3 Deviations of the data from the estimated regression model.

Figure 11-3 Deviations of the data from the estimated regression model.

Traditional notation



Definition

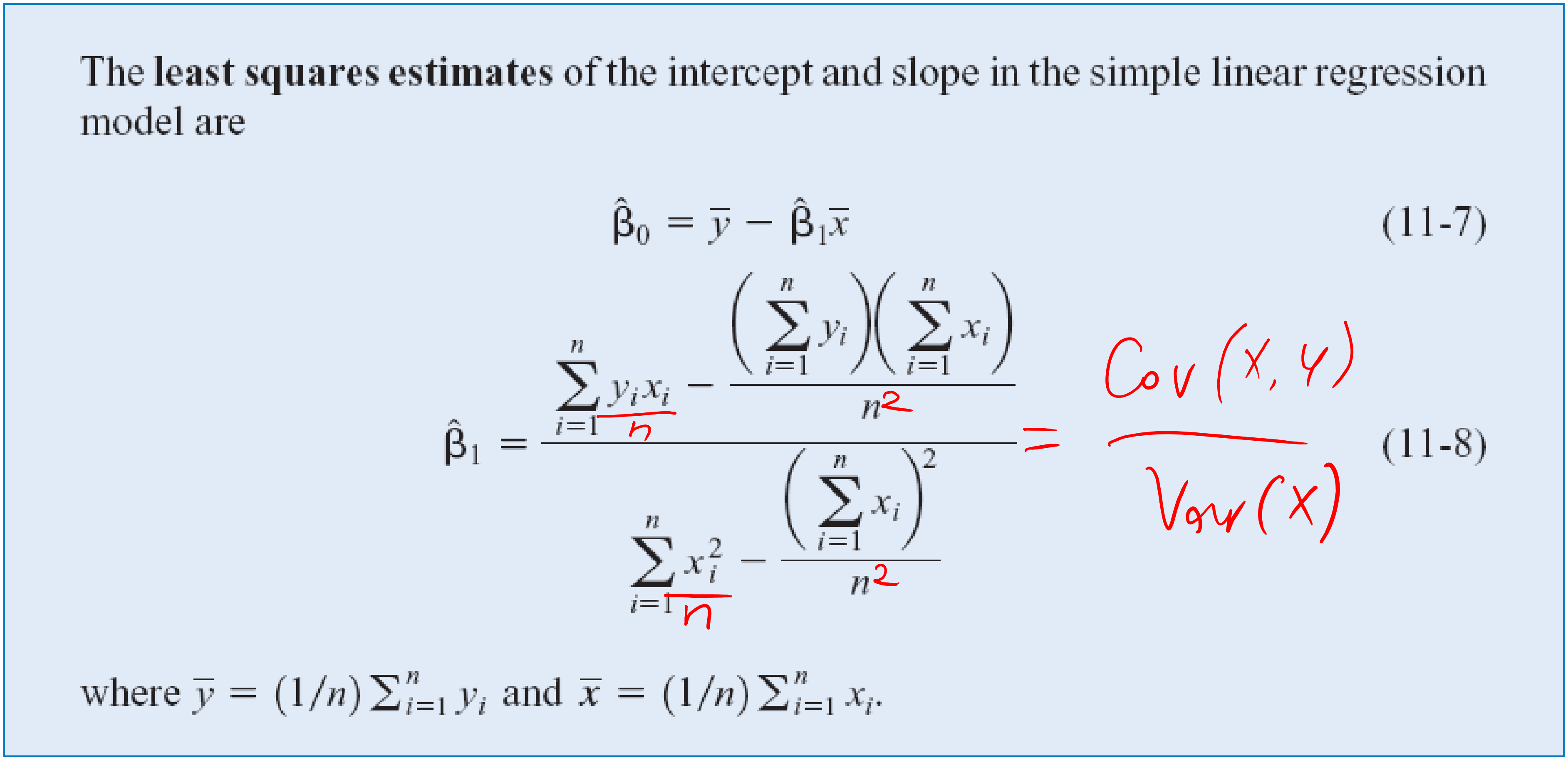
The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (11\text{-}7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}} \qquad (11\text{-}8)$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

11-2: Simple Linear Regression



Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (11\text{-}7)$$

$$\hat{\beta}_1 = \frac{\dfrac{\sum_{i=1}^{n} y_i x_i}{n} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n^2}}{\dfrac{\sum_{i=1}^{n} x_i^2}{n} - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2}} = \frac{Cov(X,Y)}{Var(X)} \qquad (11\text{-}8)$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

11-4: Hypothesis Tests in Simple Linear Regression

11-4.2 Analysis of Variance Approach to Test Significance of Regression

The analysis of variance identity is

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (11\text{-}24)$$

Symbolically,

$$SS_T = SS_R + SS_E \qquad (11\text{-}25)$$

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

VERY COMMONLY USED

- The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

- $0 \leq R^2 \leq 1$;
- We often refer (loosely) to R^2 as the amount of variability in the data explained or accounted for by the regression model.

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

- For the oxygen purity regression model,

$$\begin{aligned} R^2 &= SS_R/SS_T \\ &= 152.13/173.38 \\ &= 0.877 \end{aligned}$$

- Thus, the model accounts for 87.7% of the variability in the data.

11-2: Simple Linear Regression

Estimating σ_ε^2

An **unbiased estimator** of σ_ε^2 is

$$\boxed{\hat{\sigma}_{\varepsilon}^2} = \frac{SS_E}{n - 2} \tag{11-13}$$

where SS_E can be easily computed using

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \tag{11-14}$$

11-3: Properties of the Least Squares Estimators

- Slope Properties

$$E(\hat{\beta}_1) = \beta_1 \qquad V(\hat{\beta}_1) = \frac{\hat{\sigma}_\varepsilon^2}{S_{xx}} = \frac{\hat{\sigma}_\varepsilon^2}{n\,\sigma_x^2}$$

Large $n \rightarrow$ small variance of β_1

- Intercept Properties

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \hat{\sigma}_\varepsilon^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] =$$

$$= \hat{\sigma}_\varepsilon^2\left[1 + \frac{\mu_x^2}{\sigma_x^2}\right] \times \frac{1}{n}$$

11-4: Hypothesis Tests in Simple Linear Regression

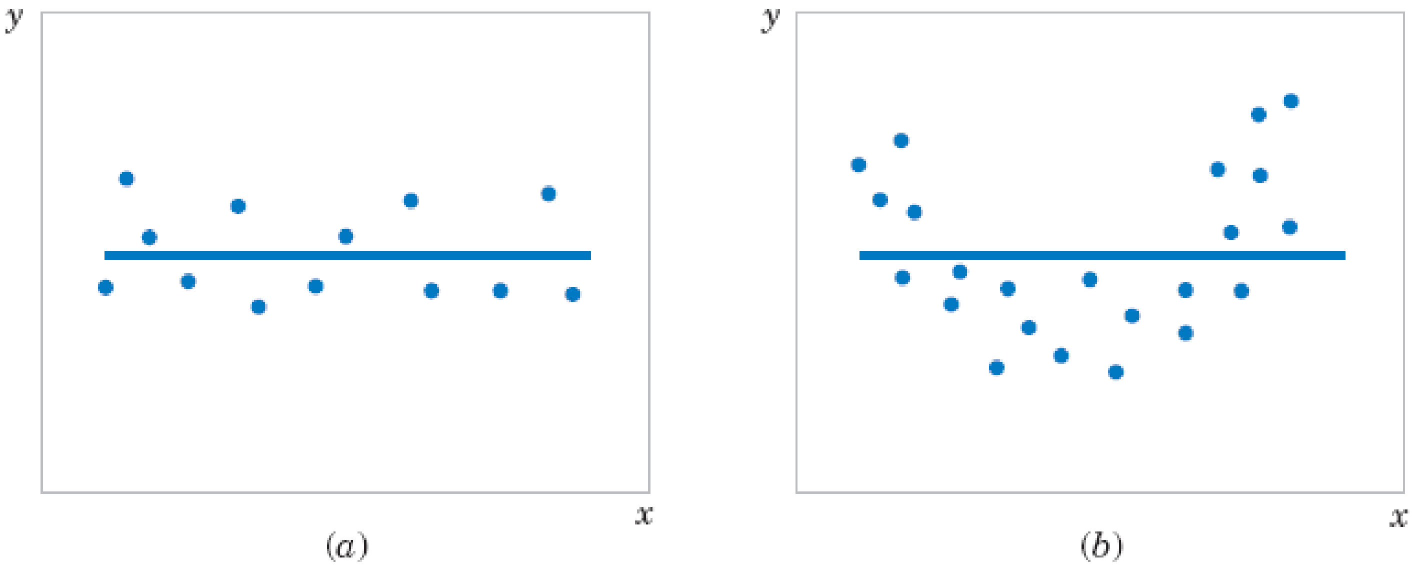


Figure 11-5 The hypothesis H_0: $\beta_1 = 0$ is not rejected.

Figure 11-5 The null hypothesis H_0: $\beta_1 = 0$ is accepted.

11-4: Hypothesis Tests in Simple Linear Regression

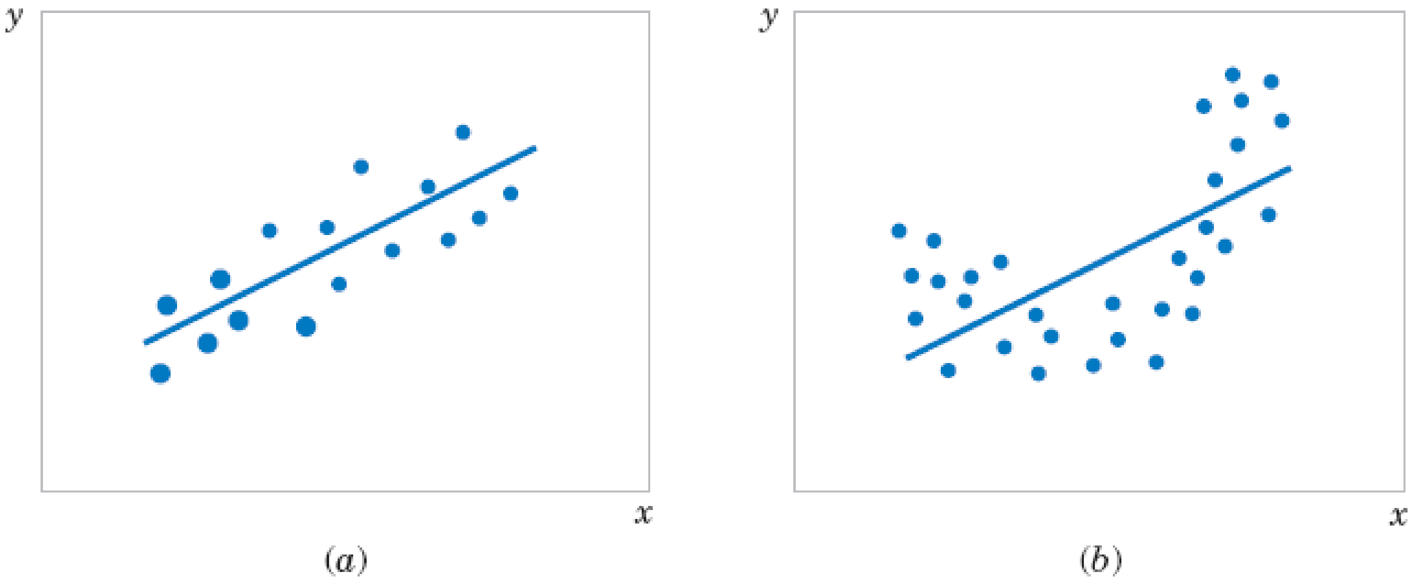


Figure 11-6 The hypothesis H_0: $\beta_1 = 0$ is rejected.

Figure 11-6 The **null hypothesis H_0: $\beta_1 = 0$ is rejected**.

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *Z*-tests for large n

An important special case of the hypotheses of Equation 11-18 is

$$H_0\colon \beta_1 = 0$$
$$H_1\colon \beta_1 \neq 0$$

These hypotheses relate to the **significance of regression**. *Failure* to reject H$_0$ is equivalent to concluding that there is no linear relationship between *X* and *Y*.

11-4: Hypothesis Tests in Simple Linear Regression

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Choose α

(e.g. $\alpha = 5\%$ for 95% confidence in rejecting H_0)

$$Z = \frac{\hat{\beta}_1 - 0}{\frac{\hat{\sigma}_\varepsilon}{\hat{\sigma}_x} \cdot \frac{1}{\sqrt{n}}}$$

$$\sqrt{Var(\hat{\beta}_1)} = \frac{\sigma_\varepsilon}{\sigma_x} \frac{1}{\sqrt{n}}$$

for $\alpha = 5\%$

Reject H_0 if $|Z| > Z_{\alpha/2} = 1.96$

11-4: Hypothesis Tests in Simple Linear Regression

11-4.1 Use of *t*-tests for smaller n.

The number of degrees of freedom in **n-2**

n-2

One can always fit a straight line through two points so one needs n>=3

11-4: Hypothesis Tests in Simple Linear Regression

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

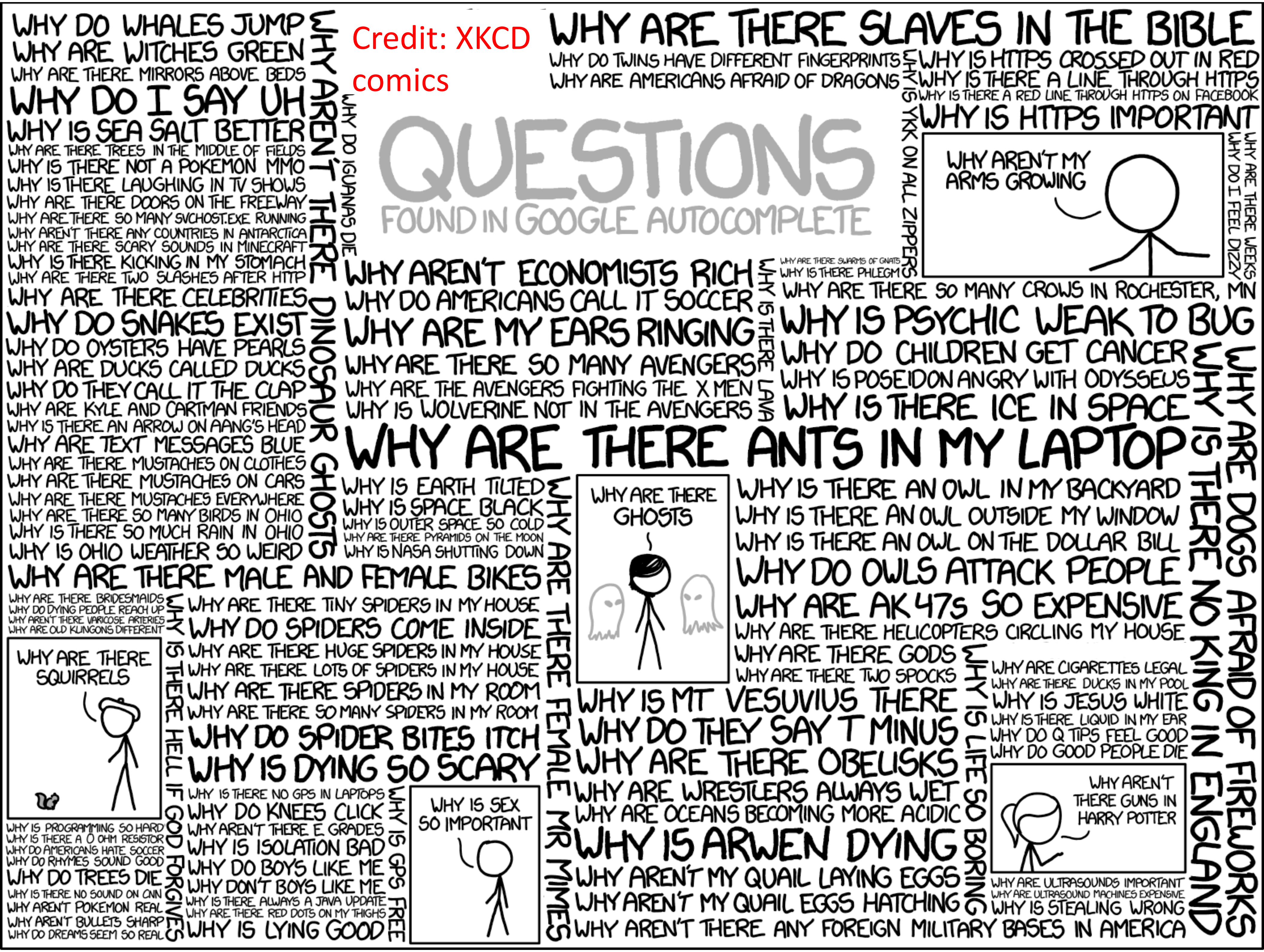
Choose α (e.g. $\alpha = 5\%$ for 95% confidence in rejecting H_0)

$$T = \frac{\hat{\beta}_1}{\frac{\hat{\sigma}_\epsilon}{\hat{\sigma}_x} \cdot \frac{1}{\sqrt{n}}}$$

$t_{\alpha/2, n-2}$ is such

$$1 - \frac{\alpha}{2} = \text{tcdf}(t_{\alpha/2, n-2}, n-2)$$

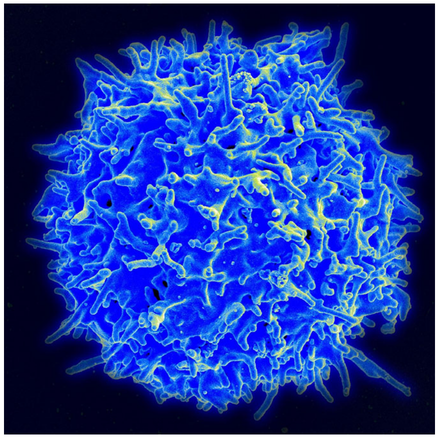
Reject H_0 if $|T| > t_{\alpha/2, n-2}$


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE GHOSTS
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT

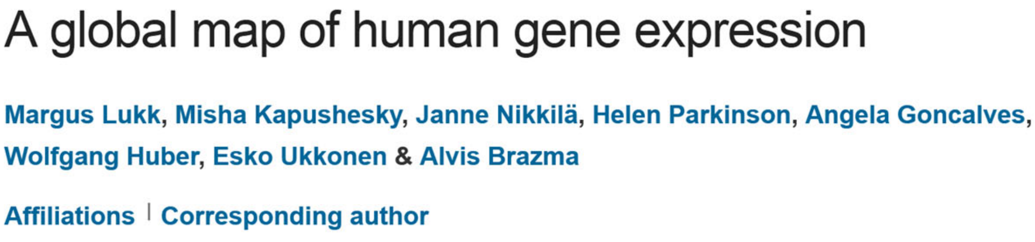
Credit: XKCD comics

Human T cell expression data

- The matrix contains 47 expression samples from Lukk et al, Nature Biotechnology 2010
- All samples are from T cells in different individuals
- Only the top 3000 genes with the largest variability were used
- The value is log2 of gene's expression level in a given sample as measured by the microarray technology



a T cell

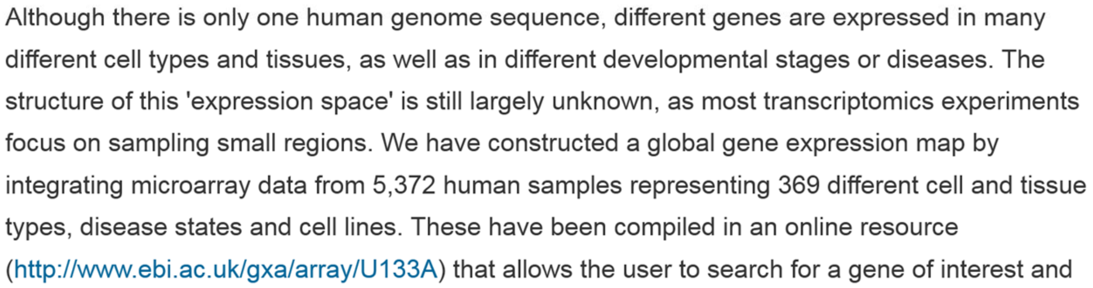


A global map of human gene expression

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

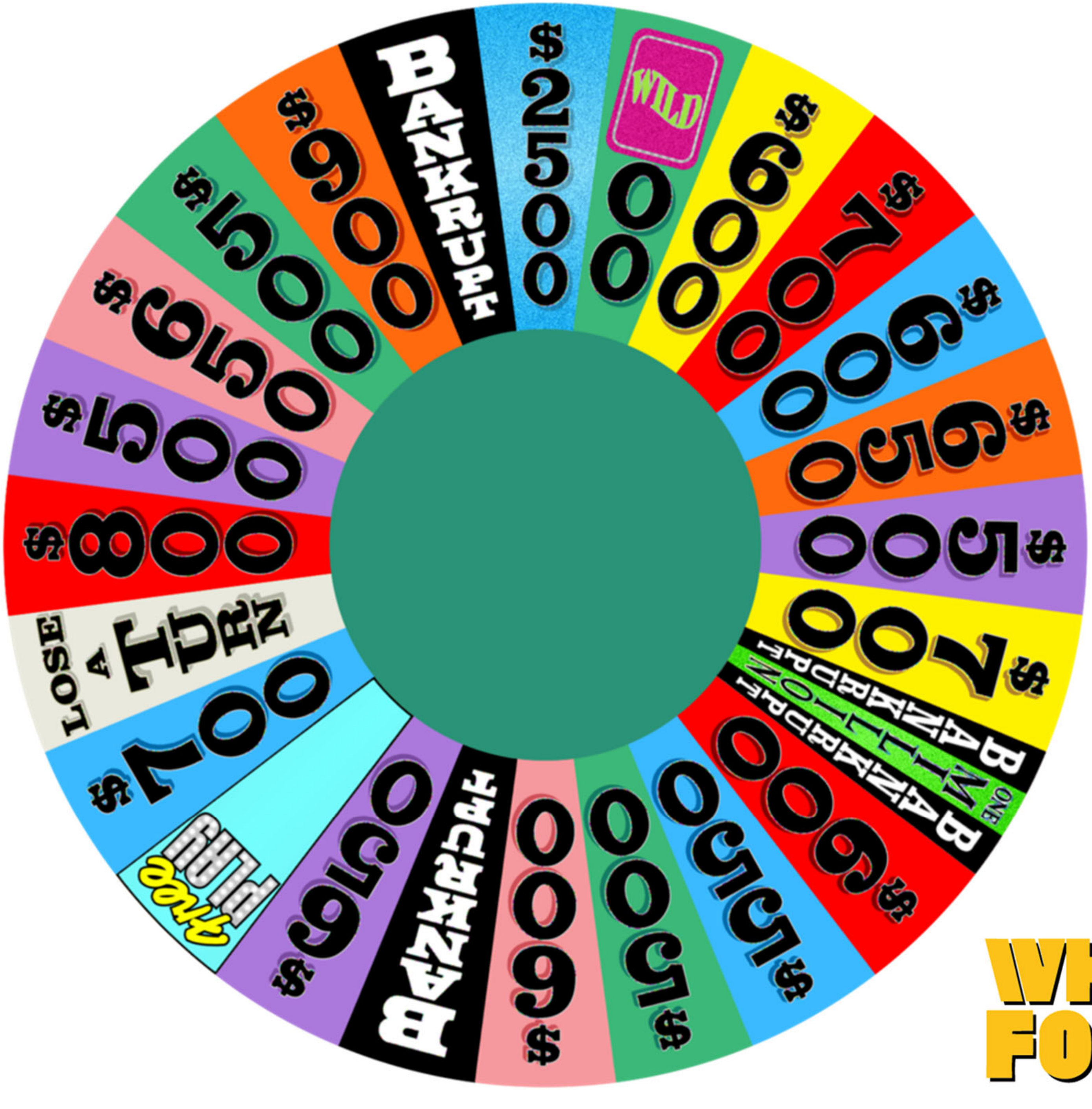
Affiliations | Corresponding author

Nature Biotechnology **28**, 322–324 (2010) | doi:10.1038/nbt0410-322



Although there is only one human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. The structure of this 'expression space' is still largely unknown, as most transcriptomics experiments focus on sampling small regions. We have constructed a global gene expression map by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. These have been compiled in an online resource (http://www.ebi.ac.uk/gxa/array/U133A) that allows the user to search for a gene of interest and

“Let’s Make a Deal” show with Monty Hall aired on NBC/ABC 1963-1986



WHEEL OF FORTUNE®

Gene Expression "Wheel of Fortune"

- Each group gets a pair of genes that are known to be correlated.
- Each group also gets a random pair of genes selected by the "Wheel of Fortune". They may or may not be correlated
- Download (log-transformed) expression_table.mat
- Run command fitlm(x,y) on assigned and random pairs
- Record β_0, β_1, R^2, P-value of the slope β_1 and write them on the blackboard
- Validate Matlab result for R^2 using your own calculations
- Look up gene names (see gene_description in your workspace) and write down a brief description of biological functions of genes. Does their correlation make biological sense?

Correlated pairs

plausible biological connection based on short description

g1=1994; g2=188; group 1

g1=2872; g2=1269; group 2

g1=1321; g2=10; group 3

g1= 886; g2=819; group 4

g1=2138; g2=1364; group 5

no obvious biological common function

```
g1=1+floor(rand.*3000); g2=1+floor(rand.*3000);
disp([g1, g2])
```

Confidence interval for population variance σ^2

- Up until now we were calculating the confidence interval on the **population average μ**
- What if one wants to put **confidence interval on population variance $\boldsymbol{\sigma^2}$**?
- We know an unbiased estimator of σ^2:

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

- How to determine confidence interval?

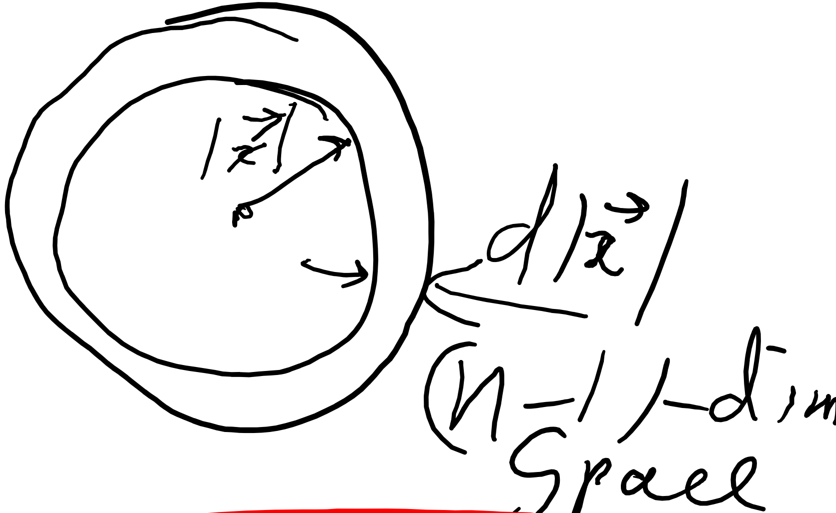
$\vec{X} = (x_1, x_2 \ldots x_n)$ $\qquad x_i \to x_i - \bar{X}$

$y = |\vec{x}|^2 = \sum x_i^2 = (n-1)s^2 \qquad \sum_{i=1}^{n} x_i = 0$

$$P(|\vec{x}|)\, d|\vec{x}| \sim \prod_{i=1}^{n-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) dx_i$$

(left the last one since $x_n = -\sum_{i=1}^{n-1} x_i$)

sphere area $\sim |\vec{x}|^{n-2}$



$|\vec{x}| = \sqrt{y}$

$d|\vec{x}| = \frac{1}{\sqrt{y}} dy$

$\prod dx_i \sim |\vec{x}|^{n-2} d|\vec{x}|$

$$\boxed{P(y)\,dy = y^{\frac{n-1}{2}-1} \exp\left(-\frac{y}{2}\right) dy}$$

8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

$X=(n-1)S^2/\sigma^2$

We know n, S^2

want to estimate σ^2

$f(x,n) \sim x^{(n-1)/2-1}\exp(-x/2)$

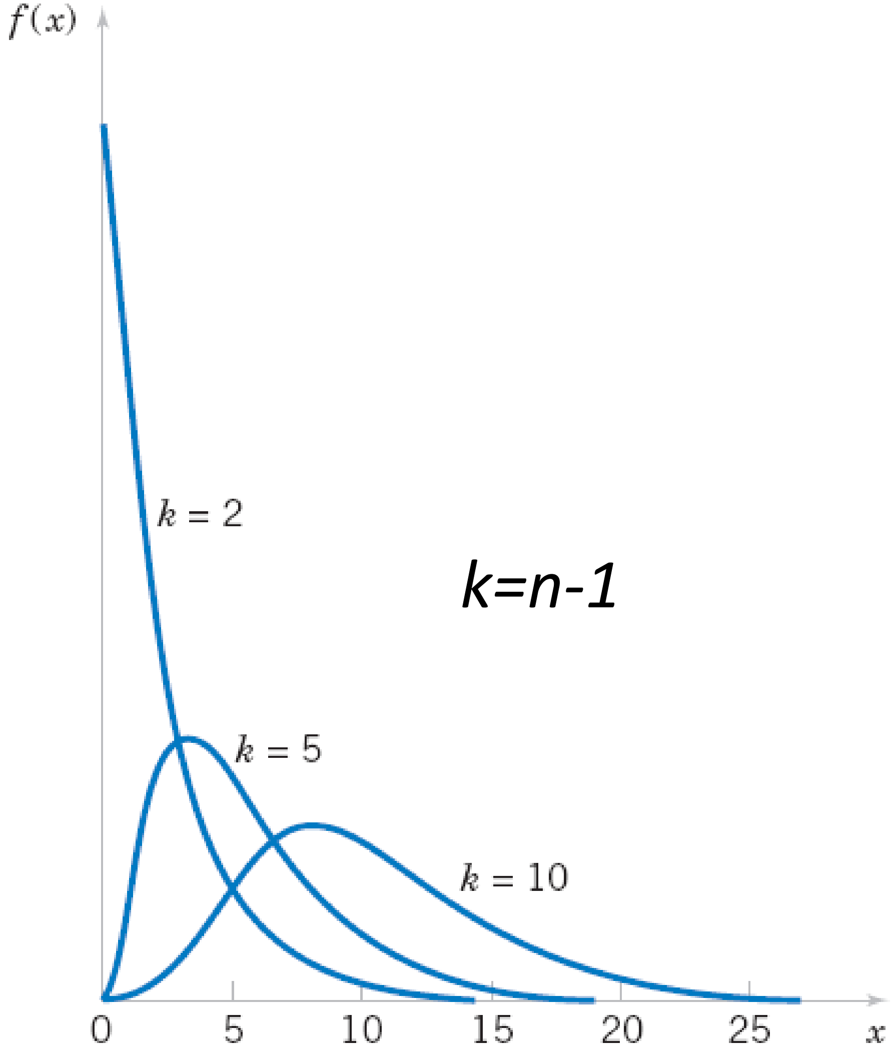
It is just Gamma PDF
with $r=(n-1)/2$, and $\lambda=1/2$

Mean value:

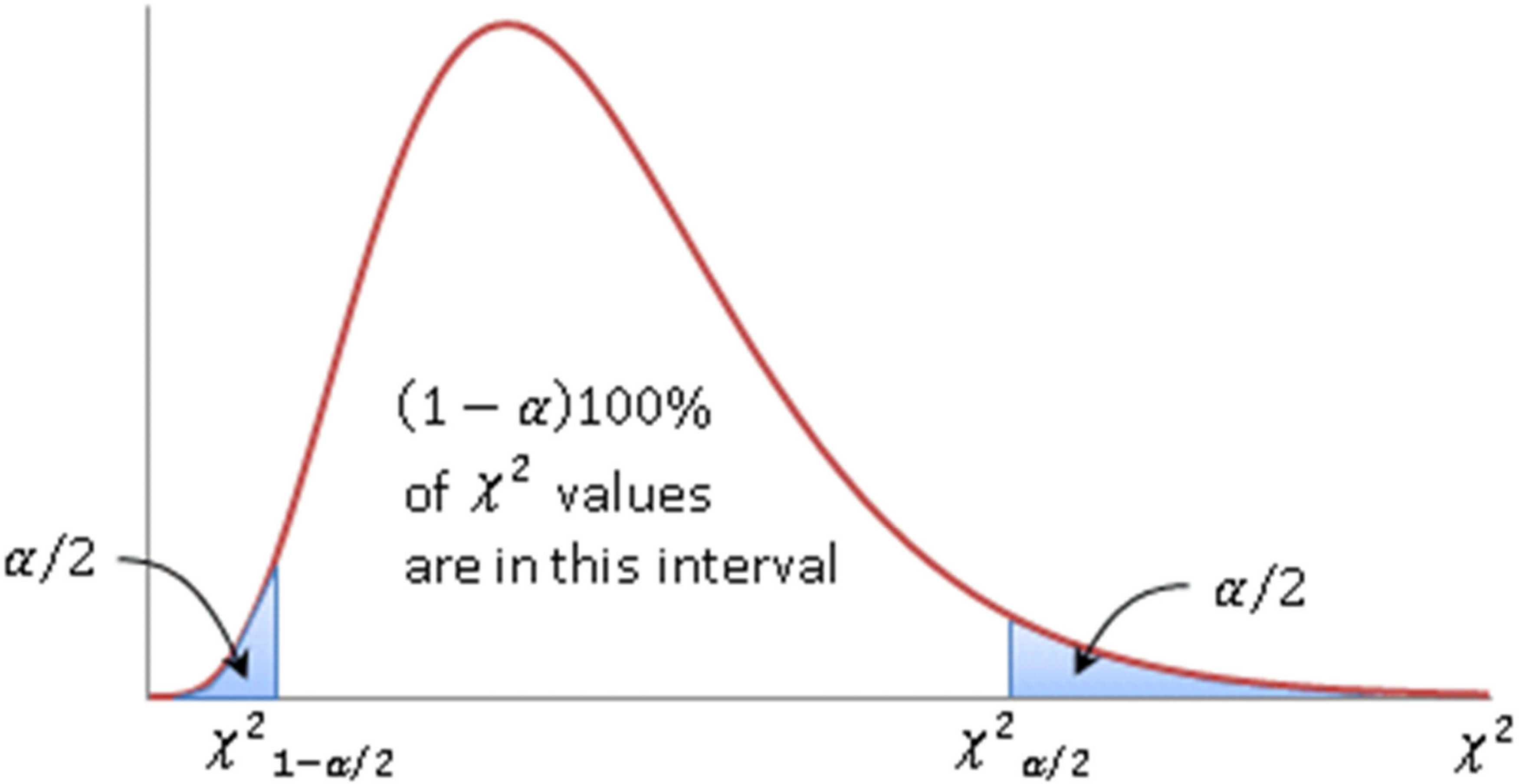
$$n-1$$

Standard deviation:

$$\sqrt{2(n-1)}$$



k=n-1



$$\chi^2_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Person’s chi-squared Goodness of fit test

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

http://www.scientificameriken.com/candy5.asp

“To our surprise M&Ms met our demand to review their procedures in determining candy ratios. It is, however, noted that the figures presented in their email differ from the information provided from their website (http://us.mms.com/us/about/products/milkchocolate/). An email was sent back informing them of this fact. To which M&Ms corrected themselves with one last email:

In response to your email regarding M&M'S CHOCOLATE CANDIES

Thank you for your email.
On average, our new mix of colors for M&M'S® Chocolate Candies is:

M&M'S® Milk Chocolate: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown.

M&M'S® Peanut: 23% blue, 23% orange, 15% green, 15% yellow, 12% red, 12% brown.

M&M'S® Kids MINIS®: 25% blue, 25% orange, 12% green, 13% yellow, 12% red, 13% brown.

M&M'S® Crispy: 17% blue, 16% orange, 16% green, 17% yellow, 17% red, 17% brown.

M&M'S® Peanut Butter and Almond: 20% blue, 20% orange, 20% green, 20% yellow, 10% red, 10% brown.

Have a great day!

Your Friends at Masterfoods USA
A Division of Mars, Incorporated

How to accept or reject the null hypothesis that these probabilities are correct from a finite sample?

Pearson chi^2 Goodness of Fit Test

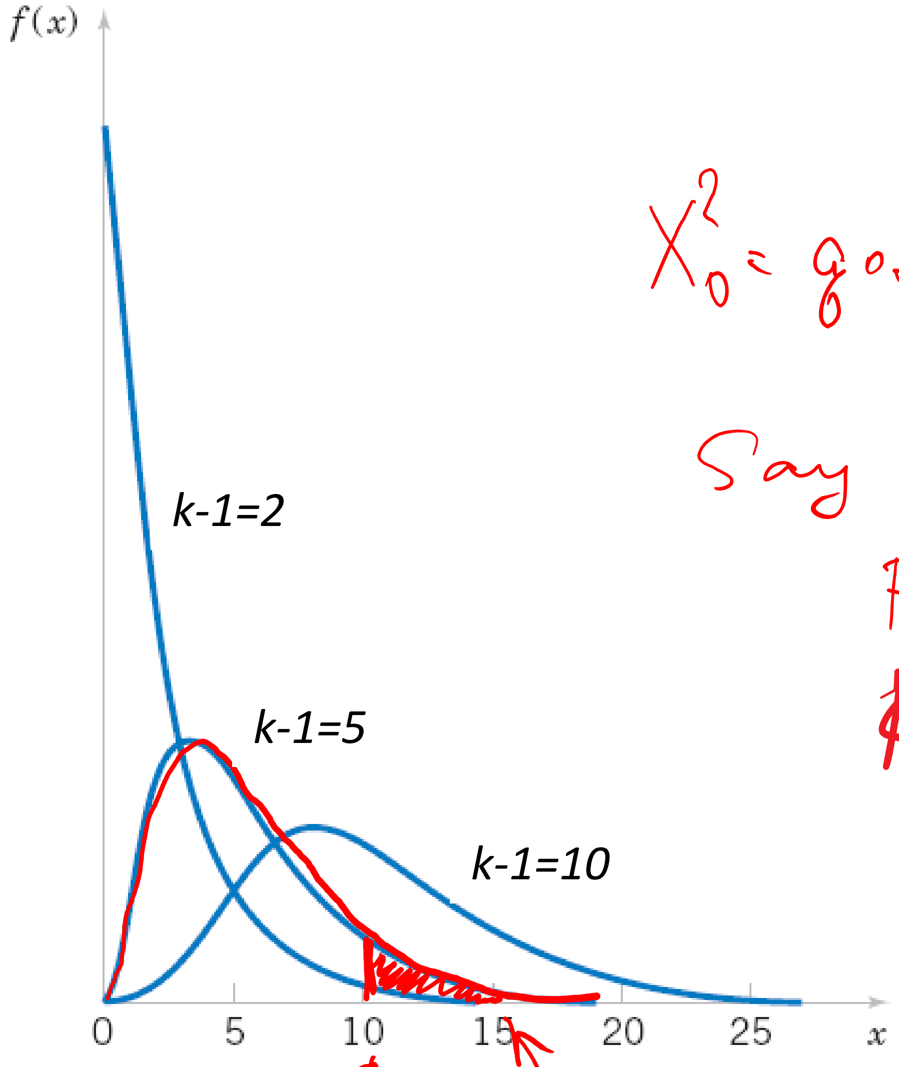
- Assume there is a sample of size *n* from a population with *k* classes (e.g. 6 M&M colors)
- Null hypothesis H_0: class *i* has frequency f_i in the population
- Alternative hypothesis H_1: some population frequencies are inconsistent with f_i
- Let O_i be the observed number of sample elements in the *i*th class and $E_i = n\, f_i$ be the expected number of sample elements in the *i*th class.
- Group any bin with E_i <3 with
  a) if numerical value of i is important, group it with its neighbor (k=i-1 or k=i+1) which has the smallest E_k until E_{group} >=3;
  b) If numerical value of i is irrelevant, group together all E_i <3 bins until E_{group} >=3
- The test statistic is

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (9\text{-}47)$$

P-value is calculated based on the chi-square distribution with k-1 degrees of freedom:

P-value = Prob(H_0 is correct) =1-CDF_chi-squared($X_0{}^2$, *k-1*)

chi^2 Goodness of Fit Test is a *one-sided* hypothesis



X_0^2

p-Value that null hypothesis is correct

$$X_0^2 = gof = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Say $X_0^2 = 10$

For M&M

$K=6 \rightarrow K-1=5$

M&M group exercise

- DO NOT EAT CANDY BEFORE COUNTING IS FINISHED! THEN, **PLEASE, DO**.
- We will be testing three null hypotheses one after another:
  - M&M official data: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown
  - Website (fan collected) data from http://joshmadison.com/2007/12/02/mms-color-distribution-analysis: 18.36% blue, 20.76% orange, 18.44% green, 14.08% yellow, 14.20% red, 14.16% brown
  - Uniform distribution: 1/6~16.67% of each candy color
- You will estimate P-values for each one of these null hypotheses
- Hints: O_i – is the observed # of candies of color i; calculate the expected # E_i=(# candies in your sample)*f_i

Use **1-chi2cdf(X0squared, 5)** for P-value

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Statistical tests of independence

- Did I mix M&M candy well?

	blue	orange	green	yellow	red	brown
group 1	55	33	39	61	69	32
group 2	59	34	31	84	52	28
group 3	27	15	46	6	40	4
group 4	33	28	57	22	34	20

How to **test the hypothesis** if multiple sample are drawn from the same population?

- Table: samples (Student groups) – rows, classes (M&M colors) – columns
- Test if color fractions are **independent** from group
- ***P(Group 1 and Color = green) = P(Group 1)*P(Color green)***
- Compute for all groups/colors 6*4=24 in our case

  $E_{green}(group\ 1) = n_{tot}*(group1/n_{tot})*(green/n_{tot})$

- $\chi^2 = \sum_{groups\ \&\ colors}^{n_{tot}} \frac{(O_{color}(group) - E_{color}(group))^2}{E_{color}(group)}$
- # degrees of freedom=(colors-1)*(groups-1)

- Was the M&M box from Costco well mixed? Let’s compare the first two groups' data

	blue	orange	green	yellow	red	brown
group 1	56	62	36	36	37	35
group 2	59	67	29	39	32	25
group 3	58	63	29	28	33	24
group 4	58	60	36	22	37	36

- Using $\chi^2 = \sum_{groups\,\&\,colors} \frac{\left(O_{color}\,(group) - E_{color}(group)\right)^2}{E_{color}(group)}$

with # degrees of freedom (colors-1)*(groups-1)

Find P-value of null hypothesis H_0 that samples are independent from each other

Batch effect

Does color composition vary between Costco and Schnucks

- Costco: 114 67 70 145 121 60
- Schnucks: 60 43 103 28 74 24
- Test if they are significantly different from each other:
- Same test expect ngroups=2; ncolors=6;
- Results:

  Goodness of Fit =73.4774

  P-value = 1.9318e-14

- Batch effect is highly statistically significant!

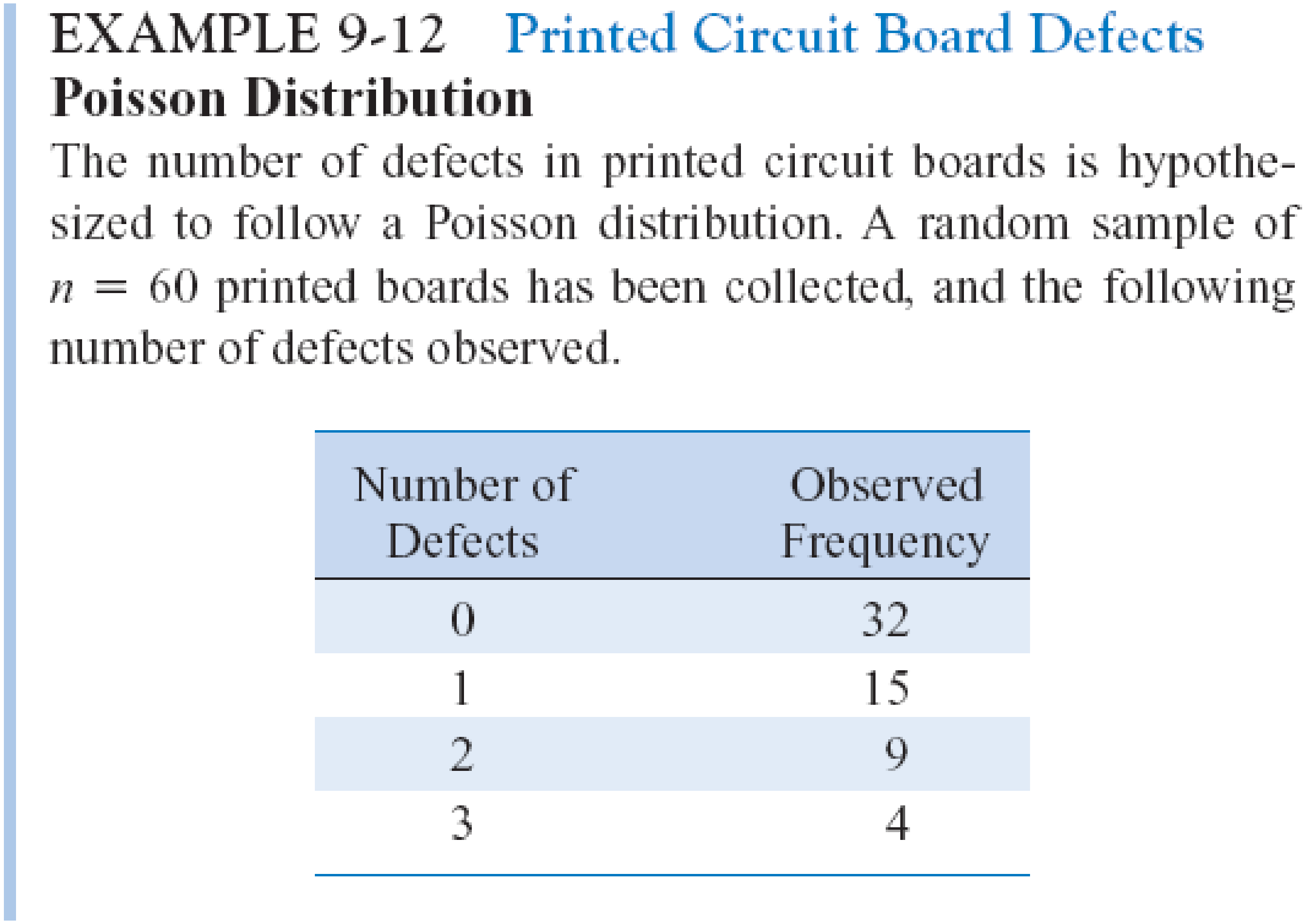
Goodness of fit with a PDF defined by m parameters

- As before: k classes (e.g. M&M colors)
- Use parameter estimators to find the best parameters for the fit
  - Method of moments
  - MLE: method of maximum likelihood
- Use chi-squared distribution with k-1-m degrees of freedom
- As before: if E_i <3, group it together with another group and reduce k by 1

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (9\text{-}47)$$

9-7 Testing for Goodness of Fit

Example 9-12



EXAMPLE 9-12 Printed Circuit Board Defects

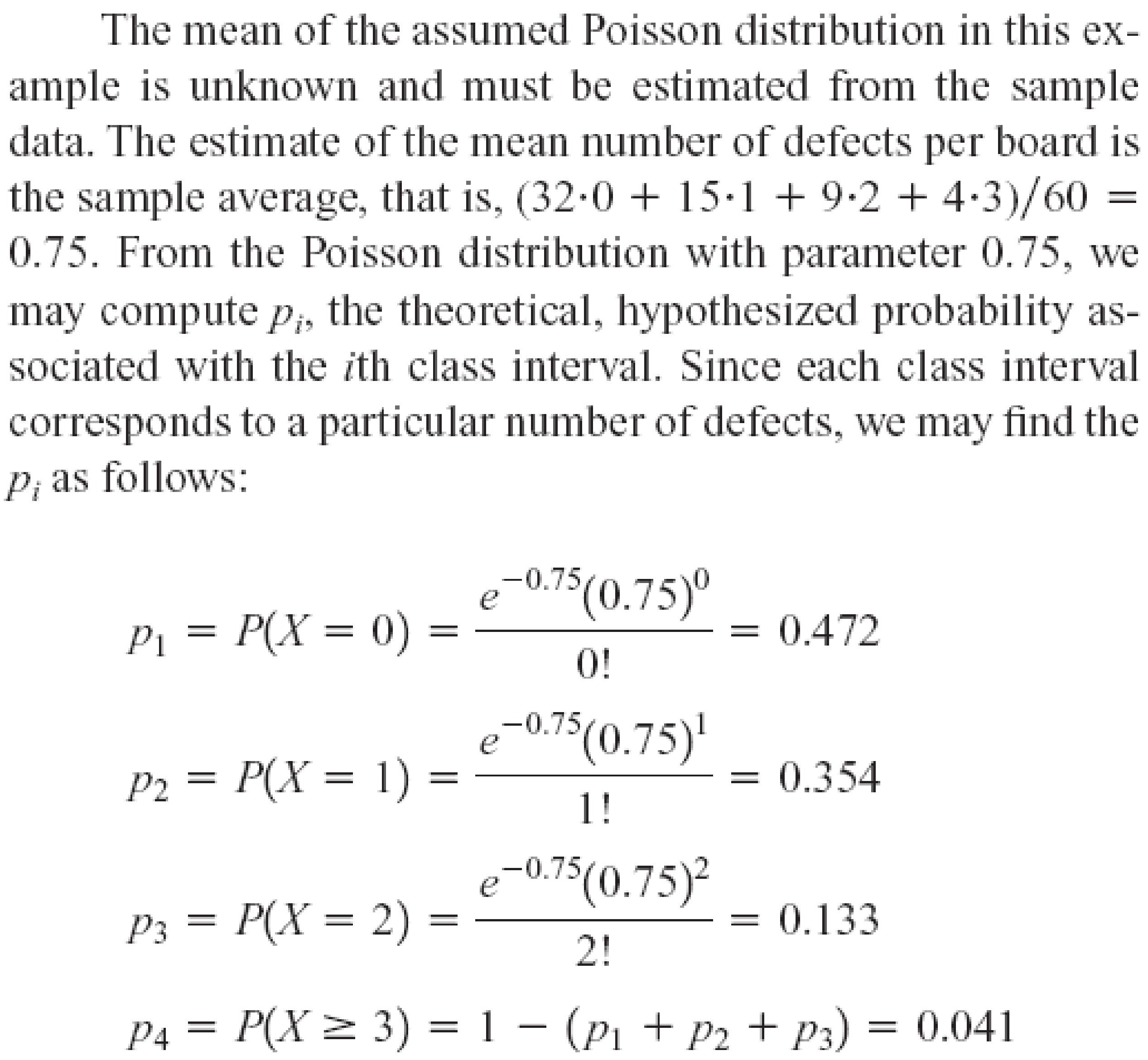
Poisson Distribution

The number of defects in printed circuit boards is hypothesized to follow a Poisson distribution. A random sample of $n = 60$ printed boards has been collected, and the following number of defects observed.

Number of Defects	Observed Frequency
0	32
1	15
2	9
3	4

9-7 Testing for Goodness of Fit

Example 9-12



The mean of the assumed Poisson distribution in this example is unknown and must be estimated from the sample data. The estimate of the mean number of defects per board is the sample average, that is, $(32 \cdot 0 + 15 \cdot 1 + 9 \cdot 2 + 4 \cdot 3)/60 = 0.75$. From the Poisson distribution with parameter 0.75, we may compute p_i, the theoretical, hypothesized probability associated with the ith class interval. Since each class interval corresponds to a particular number of defects, we may find the p_i as follows:

$$p_1 = P(X = 0) = \frac{e^{-0.75}(0.75)^0}{0!} = 0.472$$

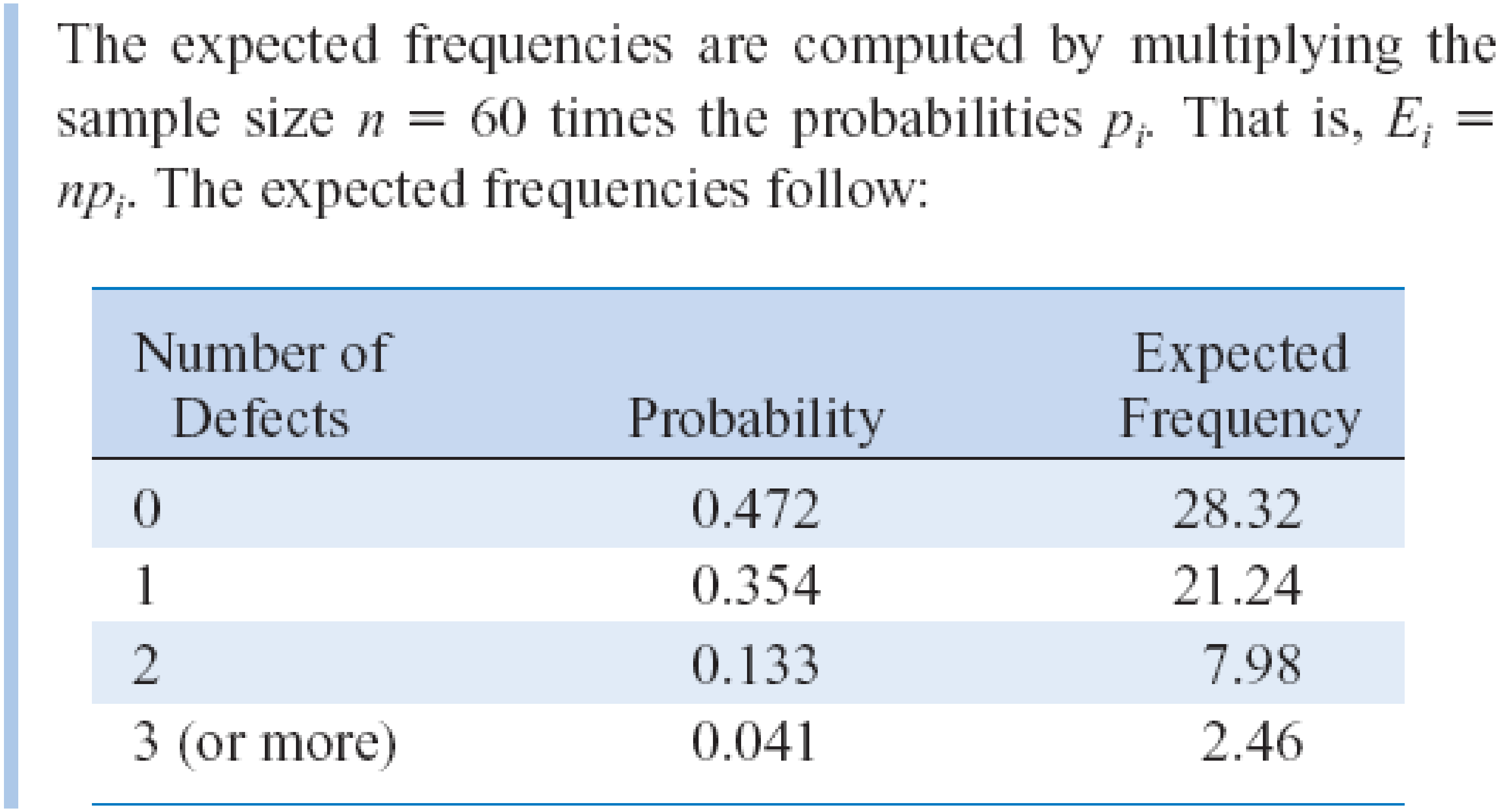
$$p_2 = P(X = 1) = \frac{e^{-0.75}(0.75)^1}{1!} = 0.354$$

$$p_3 = P(X = 2) = \frac{e^{-0.75}(0.75)^2}{2!} = 0.133$$

$$p_4 = P(X \geq 3) = 1 - (p_1 + p_2 + p_3) = 0.041$$

9-7 Testing for Goodness of Fit

Example 9-12

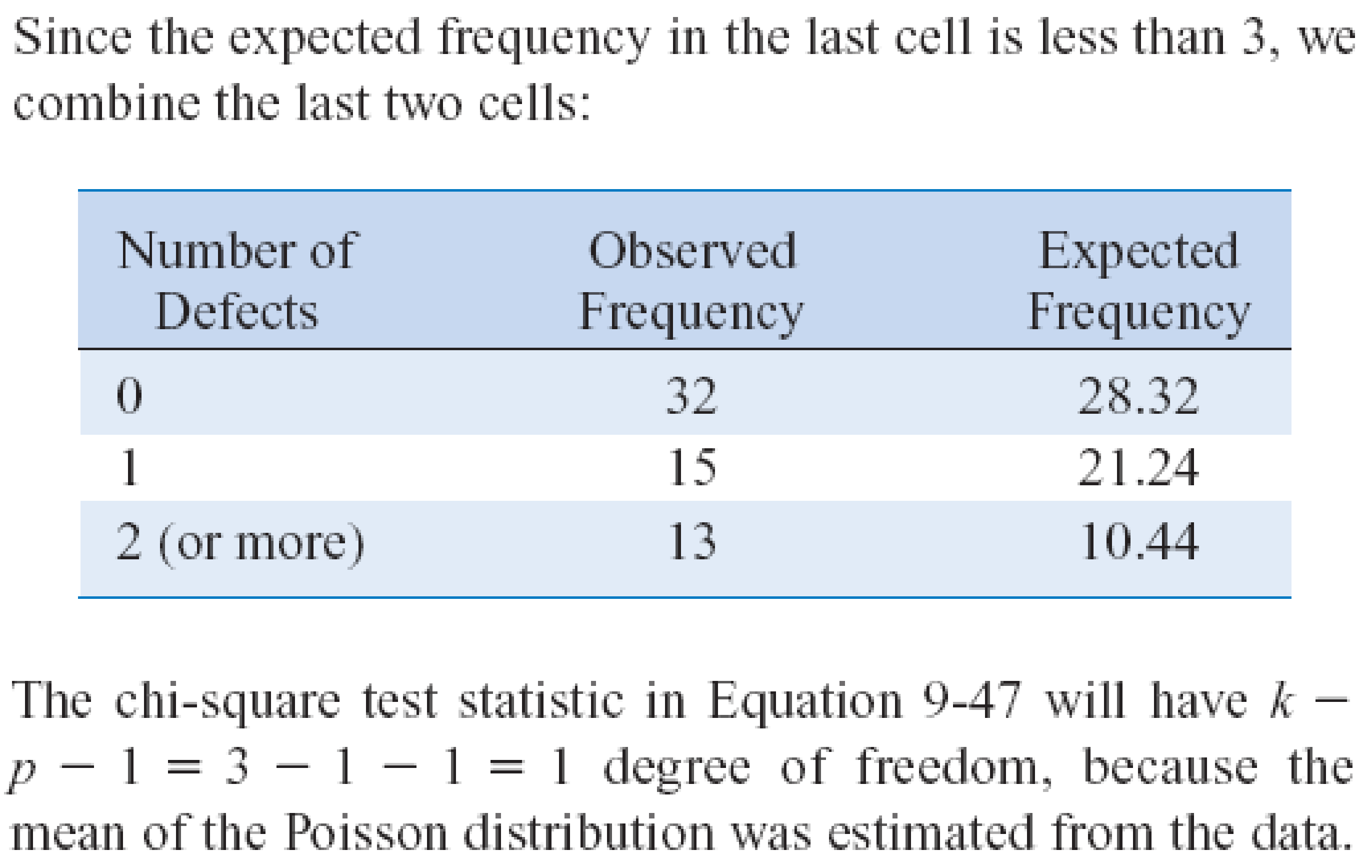


The expected frequencies are computed by multiplying the sample size $n = 60$ times the probabilities p_i. That is, $E_i = np_i$. The expected frequencies follow:

Number of Defects	Probability	Expected Frequency
0	0.472	28.32
1	0.354	21.24
2	0.133	7.98
3 (or more)	0.041	2.46

9-7 Testing for Goodness of Fit

Example 9-12



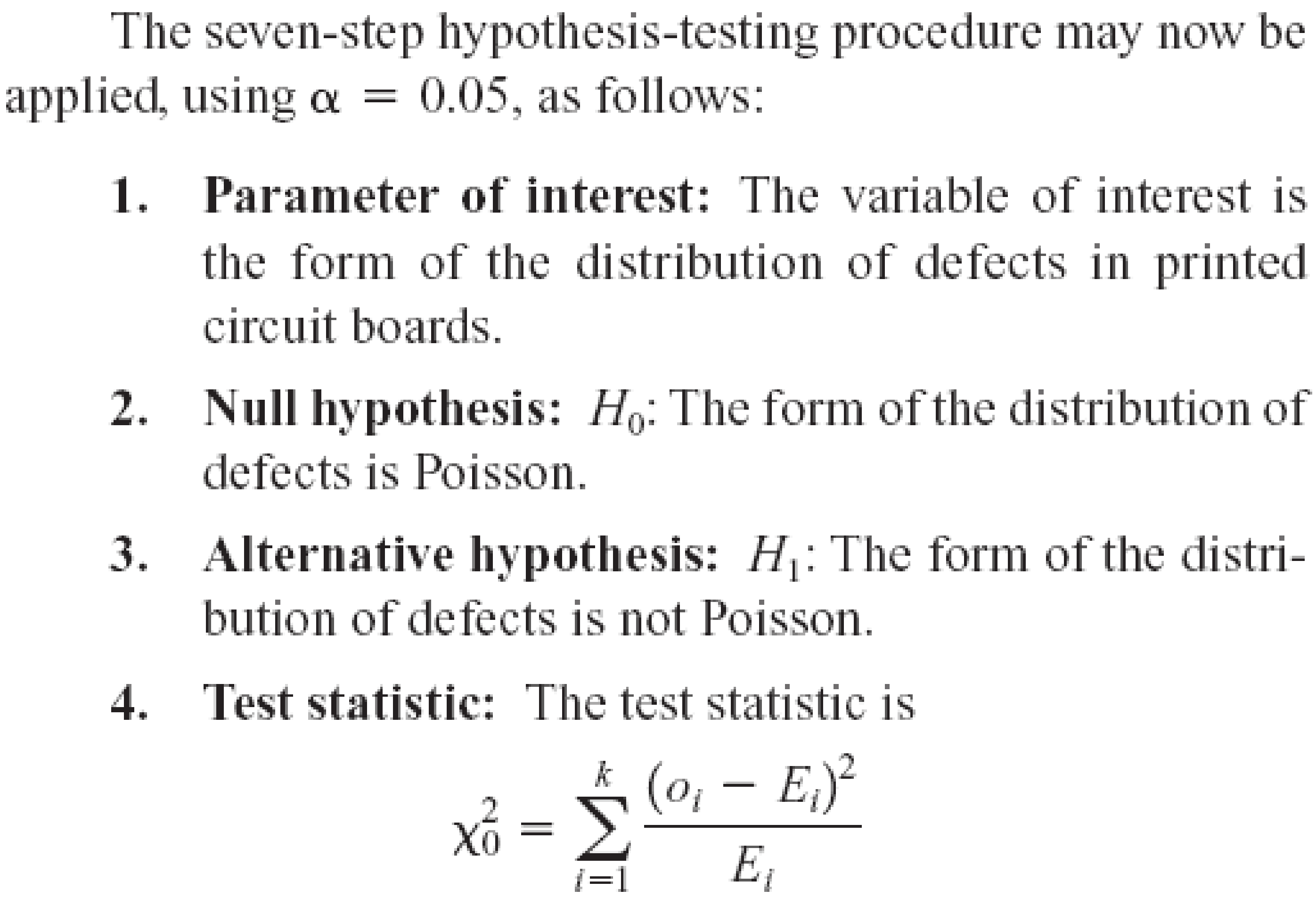
Since the expected frequency in the last cell is less than 3, we combine the last two cells:

Number of Defects	Observed Frequency	Expected Frequency
0	32	28.32
1	15	21.24
2 (or more)	13	10.44

The chi-square test statistic in Equation 9-47 will have $k - p - 1 = 3 - 1 - 1 = 1$ degree of freedom, because the mean of the Poisson distribution was estimated from the data.

9-7 Testing for Goodness of Fit

Example 9-12



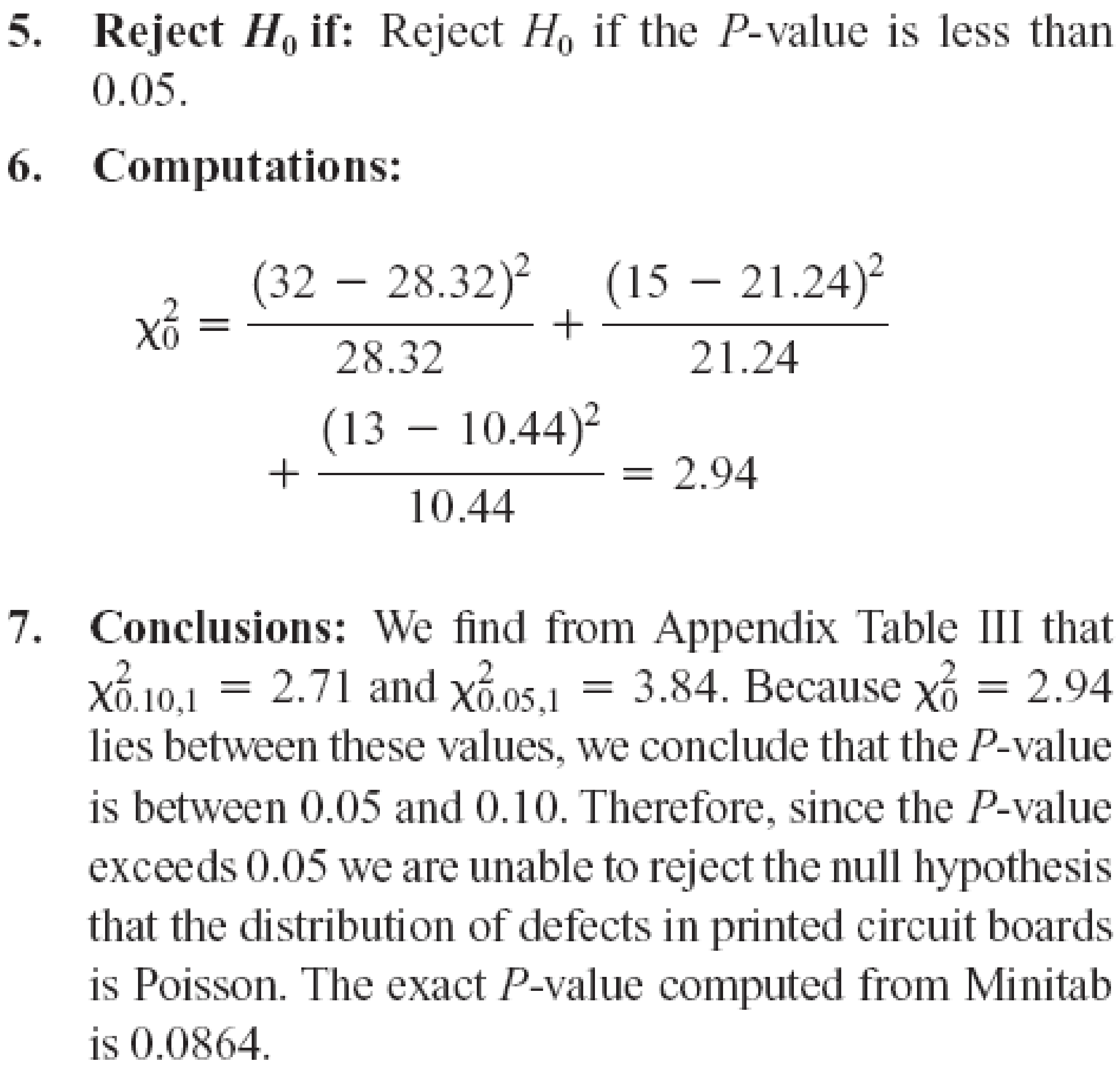
The seven-step hypothesis-testing procedure may now be applied, using $\alpha = 0.05$, as follows:

1. **Parameter of interest:** The variable of interest is the form of the distribution of defects in printed circuit boards.
2. **Null hypothesis:** H_0: The form of the distribution of defects is Poisson.
3. **Alternative hypothesis:** H_1: The form of the distribution of defects is not Poisson.
4. **Test statistic:** The test statistic is

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(o_i - E_i)^2}{E_i}$$

9-7 Testing for Goodness of Fit

Example 9-12



5. **Reject H_0 if:** Reject H_0 if the P-value is less than 0.05.

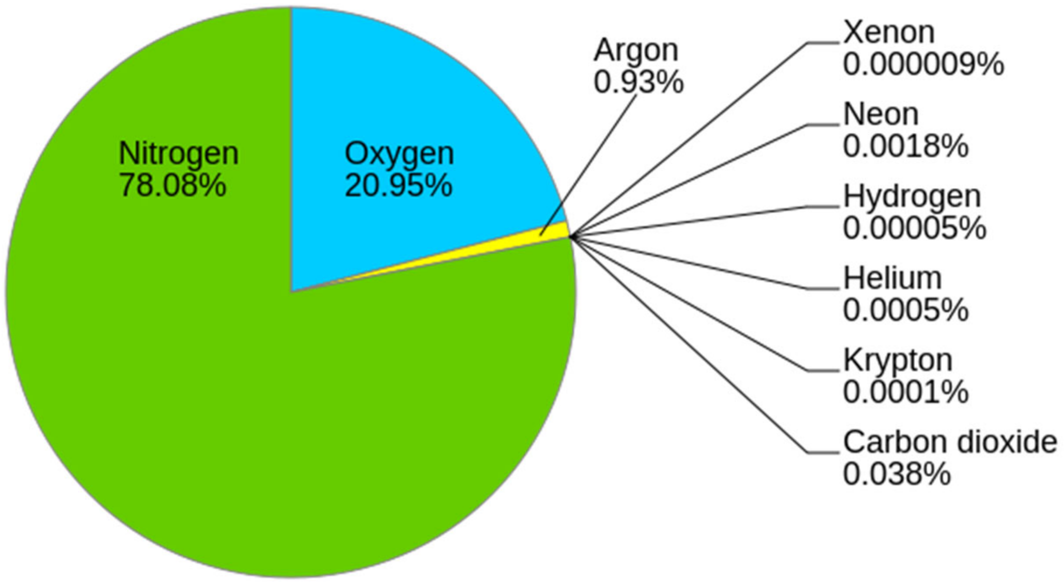
6. **Computations:**

$$\chi_0^2 = \frac{(32 - 28.32)^2}{28.32} + \frac{(15 - 21.24)^2}{21.24} + \frac{(13 - 10.44)^2}{10.44} = 2.94$$

7. **Conclusions:** We find from Appendix Table III that $\chi^2_{0.10,1} = 2.71$ and $\chi^2_{0.05,1} = 3.84$. Because $\chi_0^2 = 2.94$ lies between these values, we conclude that the P-value is between 0.05 and 0.10. Therefore, since the P-value exceeds 0.05 we are unable to reject the null hypothesis that the distribution of defects in printed circuit boards is Poisson. The exact P-value computed from Minitab is 0.0864.

Reminder

Two variable samples



- Oxygen can be distilled from the air

- Hydrocarbons need to be filtered out or the whole thing would go kaboom!!!

- When more hydrocarbons were removed, the remaining oxygen stays cleaner

- Except we don't know how dirty was the air to begin with

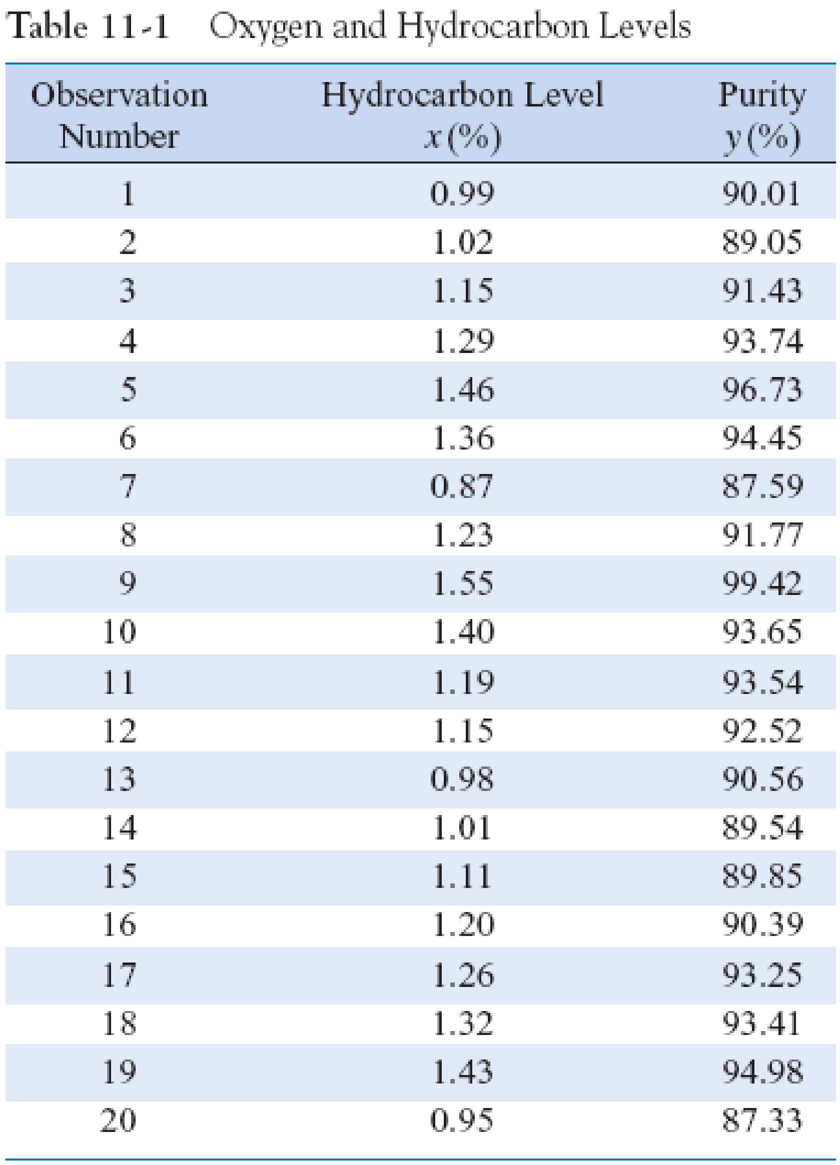


Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Linear regression

The **simple linear regression model** is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon = \hat{Y} + \varepsilon$$

ε is the **random error term**

slope β_1 and intercept β_0 of the line are called **regression coefficients**

Note: Y , $\hat{Y}$, X and ε are random variables
The minimal assumption: $E(\varepsilon \mid x)=0$ →
$E(Y \mid x) = \beta_0 + \beta_1 x + E(\varepsilon \mid x) = \beta_0 + \beta_1 x$

$$Y = \beta_0 + \beta_1 X + \epsilon$$

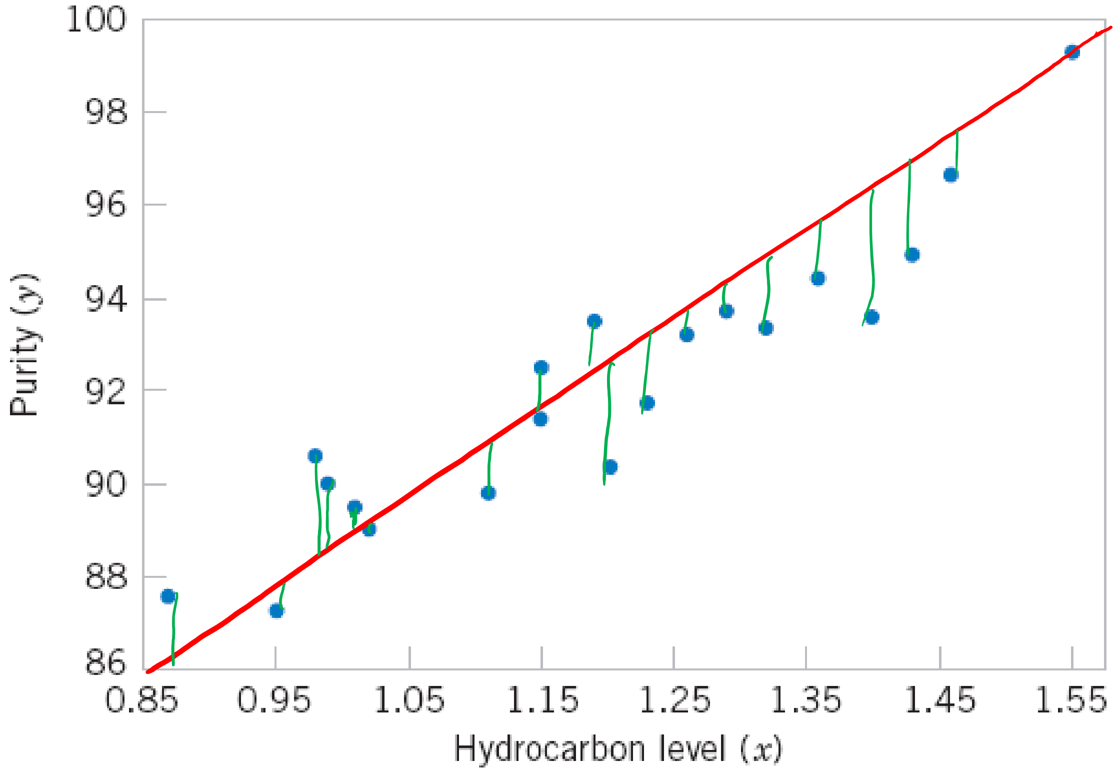


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

$$Y = 75 + 15 \cdot X + \epsilon$$

$$Y = \beta_0 + \beta_1 X + \epsilon \ ; \quad E(\epsilon | x) = 0 \ \forall x$$

How does one find β_0 & β_1?

$$Cov(Y;X) = Cov((\beta_0 + \beta_1 X + \epsilon); X) =$$

$$= \cancel{Cov(\beta_0; X)} + \beta_1 Cov(X, X) + \cancel{Cov(\epsilon; X)}$$

$Cov(\beta_0, X) = 0$ since β_0 is constant

$$Cov(X, X) = E(X^2) - E(X)^2 = Var(X)$$

$$Cov(\epsilon, X) = E(\epsilon \cdot X) - \cancel{E(\epsilon)} \cdot E(X) =$$

$$= E(\epsilon \cdot X) = \sum_{\text{all } x} x \cdot E(\cancel{\epsilon |} x) = 0$$

Thus $\boxed{\beta_1 = \frac{Cov(X, Y)}{Var(X)}}$; $\boxed{\beta_0 = E(Y) - \beta_1 E(X)}$

Method of least squares

- The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

Figure 11-3 Deviations of the data from the estimated regression model.

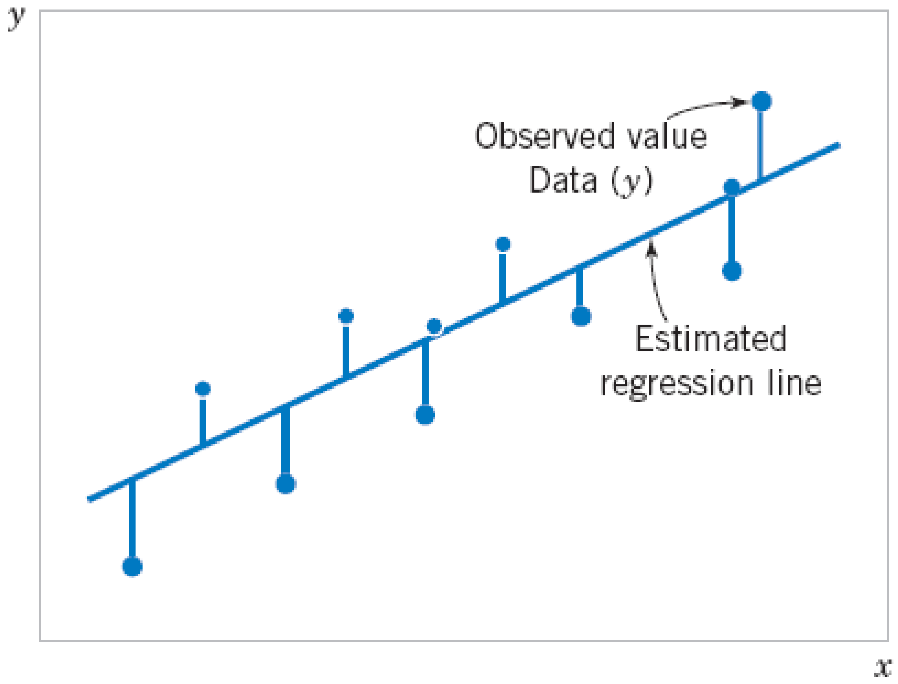
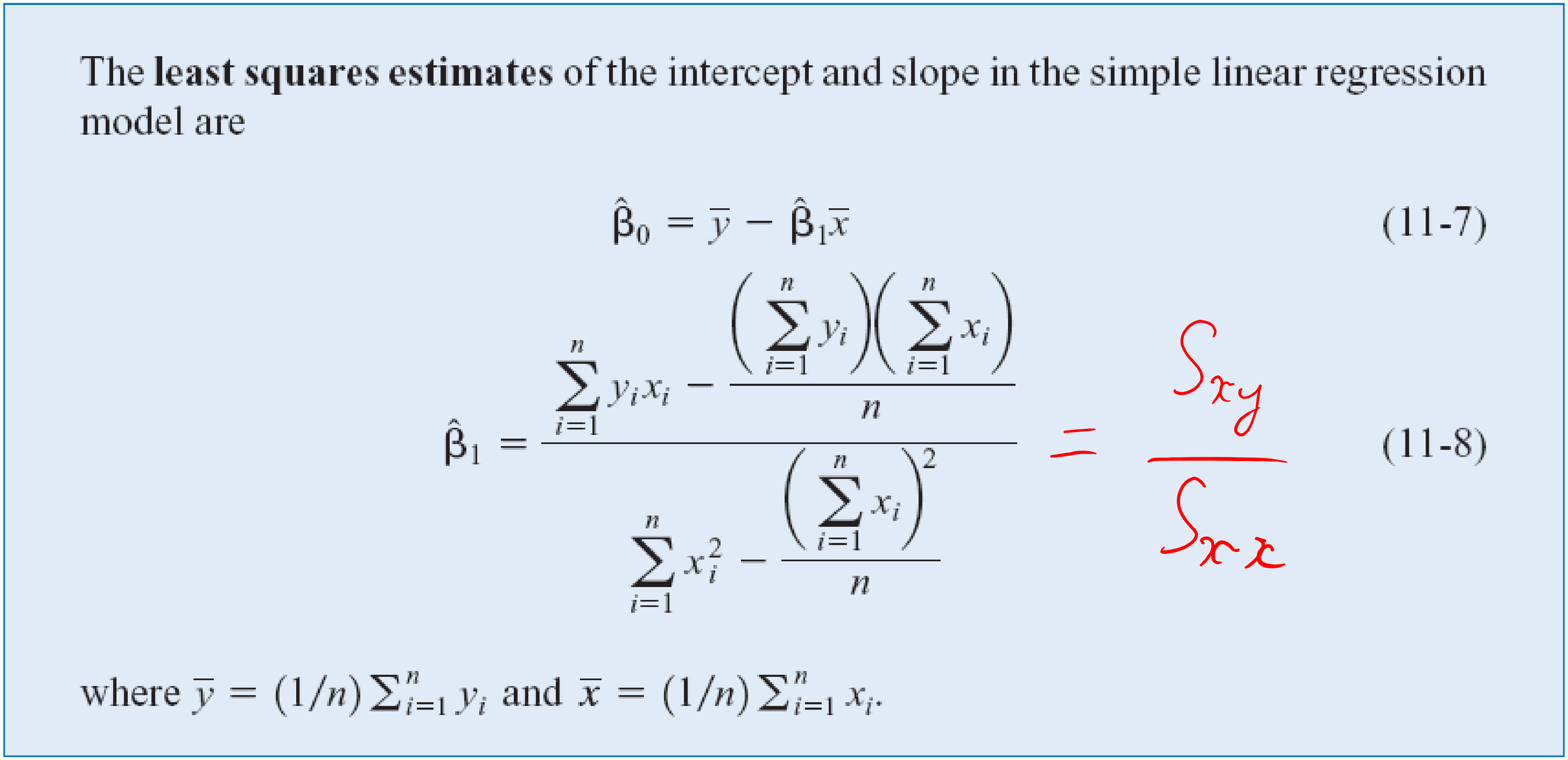


Figure 11-3 Deviations of the data from the estimated regression model.

Traditional notation

Definition



The **least squares estimates** of the intercept and slope in the simple linear regression model are

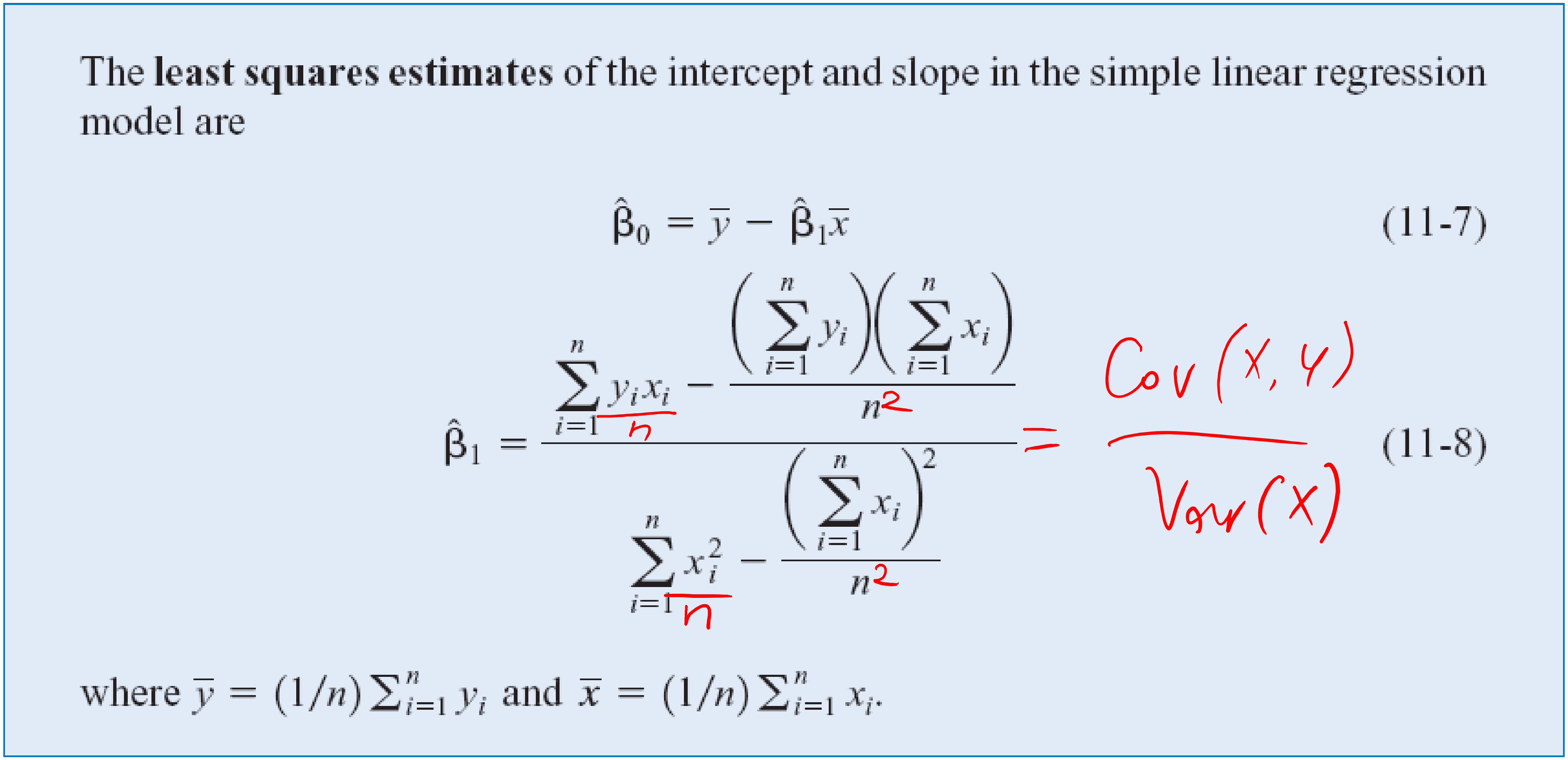
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (11\text{-}7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}} \qquad (11\text{-}8)$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

Connection to Cov(X,Y)/Var(X) result

Definition



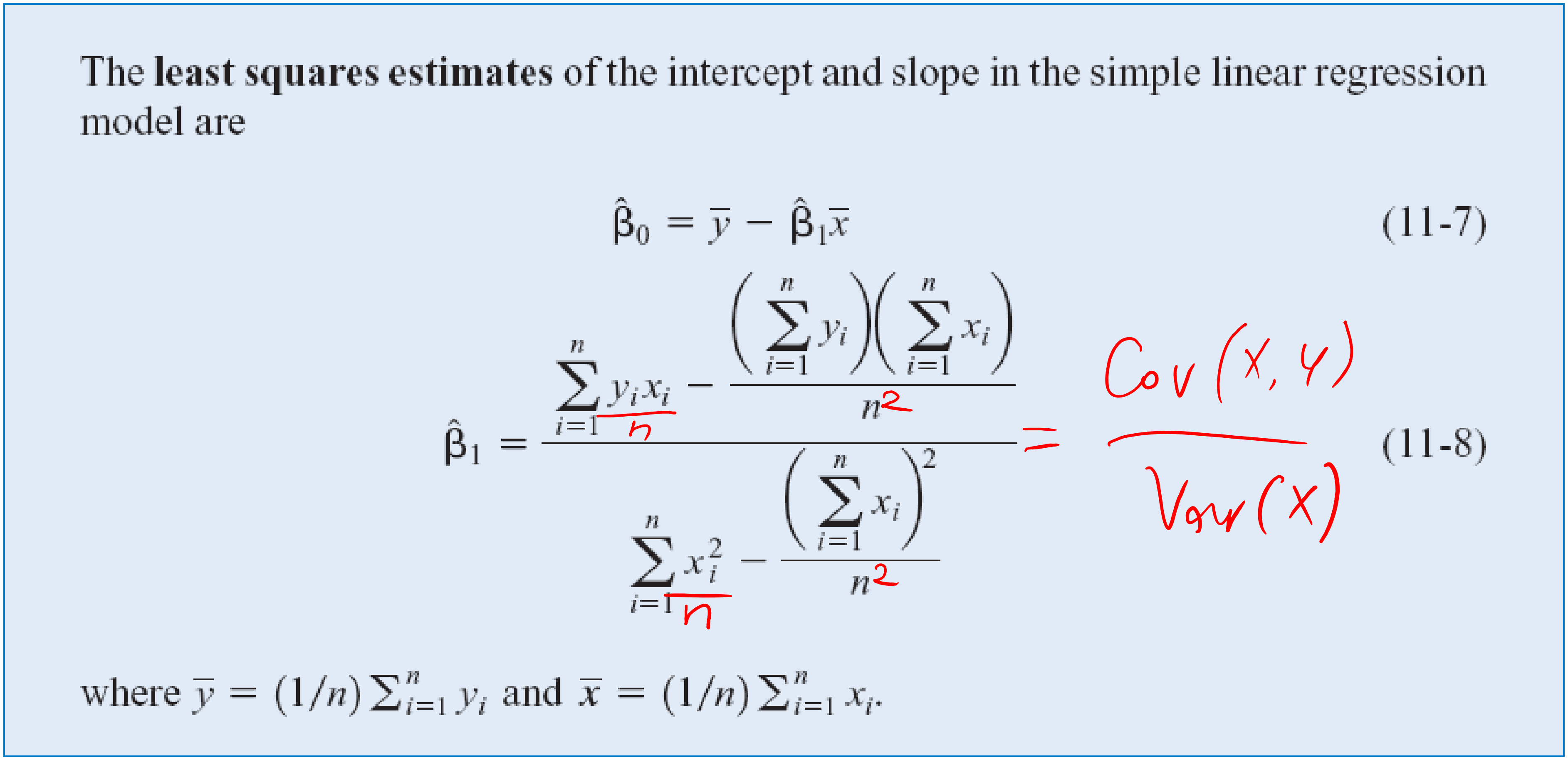
The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (11\text{-}7)$$

$$\hat{\beta}_1 = \frac{\dfrac{\sum_{i=1}^{n} y_i x_i}{n} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n^2}}{\dfrac{\sum_{i=1}^{n} x_i^2}{n} - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2}} = \frac{Cov(X,Y)}{Var(X)} \qquad (11\text{-}8)$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

Different types of y



The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (11\text{-}7)$$

$$\hat{\beta}_1 = \frac{\dfrac{\sum_{i=1}^{n} y_i x_i}{n} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n^2}}{\dfrac{\sum_{i=1}^{n} x_i^2}{n} - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2}} = \frac{Cov(X,Y)}{Var(X)} \qquad (11\text{-}8)$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

$$\bar{y} = \sum y_i / n$$

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$$

$$\varepsilon_i = y_i - \hat{y}_i$$

The analysis of variance identity is

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (11\text{-}24)$$

Symbolically,

$$SS_T = SS_R + SS_E \qquad (11\text{-}25)$$

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2)

VERY COMMONLY USED

- The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

- $0 \leq R^2 \leq 1$;
- We often refer (loosely) to R^2 as the amount of variability in the data explained or accounted for by the regression model.

11-2: Simple Linear Regression

Estimating σ_{ε}^2

An **unbiased estimator** of σ_{ε}^2 is

$$\hat{\sigma}_{\varepsilon}^2 = \frac{SS_E}{n-2} \qquad (11\text{-}13)$$

where SS_E can be easily computed using

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \qquad (11\text{-}14)$$

Multiple Linear Regression

(Chapters 12-13 in
Montgomery, Runger)

12-1: Multiple Linear Regression Model

12-1.1 Introduction

- Many applications of regression analysis involve situations in which there are more than one regressor variable X_k used to predict Y.

- A regression model then is called a **multiple regression model**.

Multiple Linear Regression Model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots \beta_k x_k + \varepsilon$$

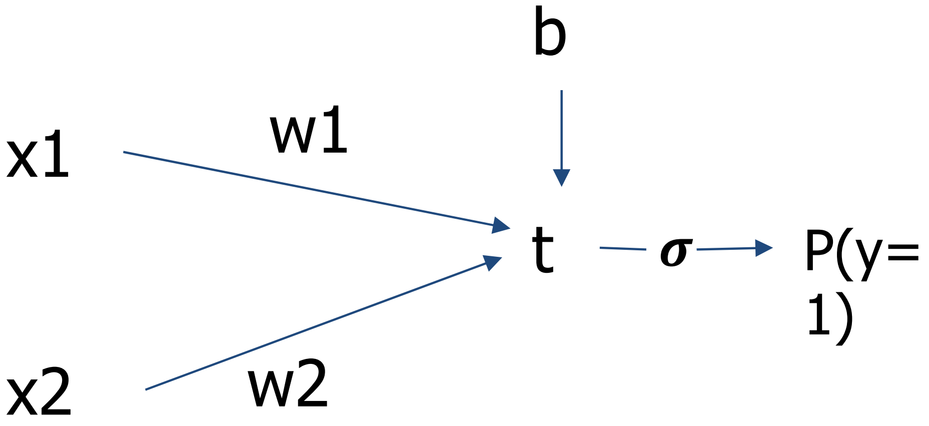
One can also use powers and products of other variables or even non-linear functions like $\exp(x_i)$ or $\log(x_i)$ instead of $x_3, \ldots x_k$.

Example: the general two-variable quadratic regression has 6 constants:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1)^2 + \beta_4 (x_2)^2 + \beta_5 (x_1 x_2) + \varepsilon$$

Logistic Regression

P(y=1) = 𝝈(x1*w1 + x2*w2 + b)



b

x1

w1

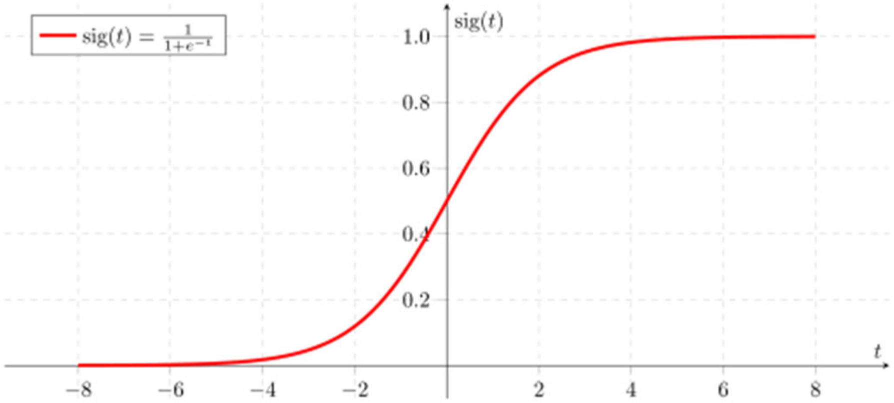
t

𝝈

P(y=1)

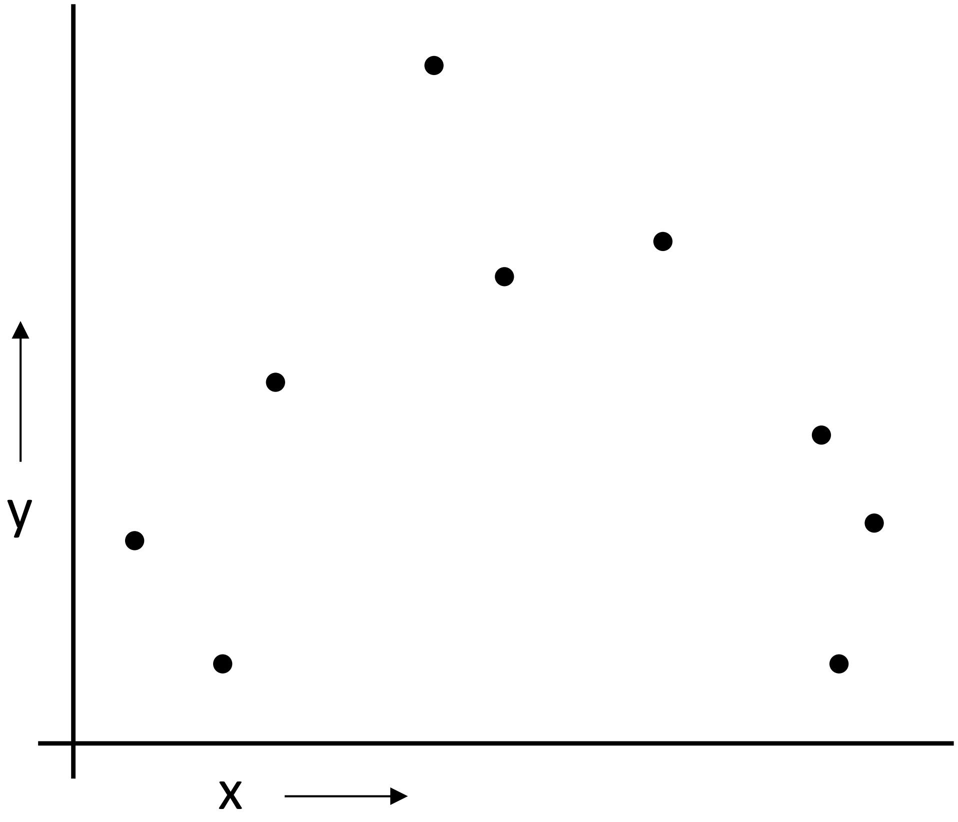
x2

w2



How to know where to stop adding new variables or powers of old variables?

A Regression Problem

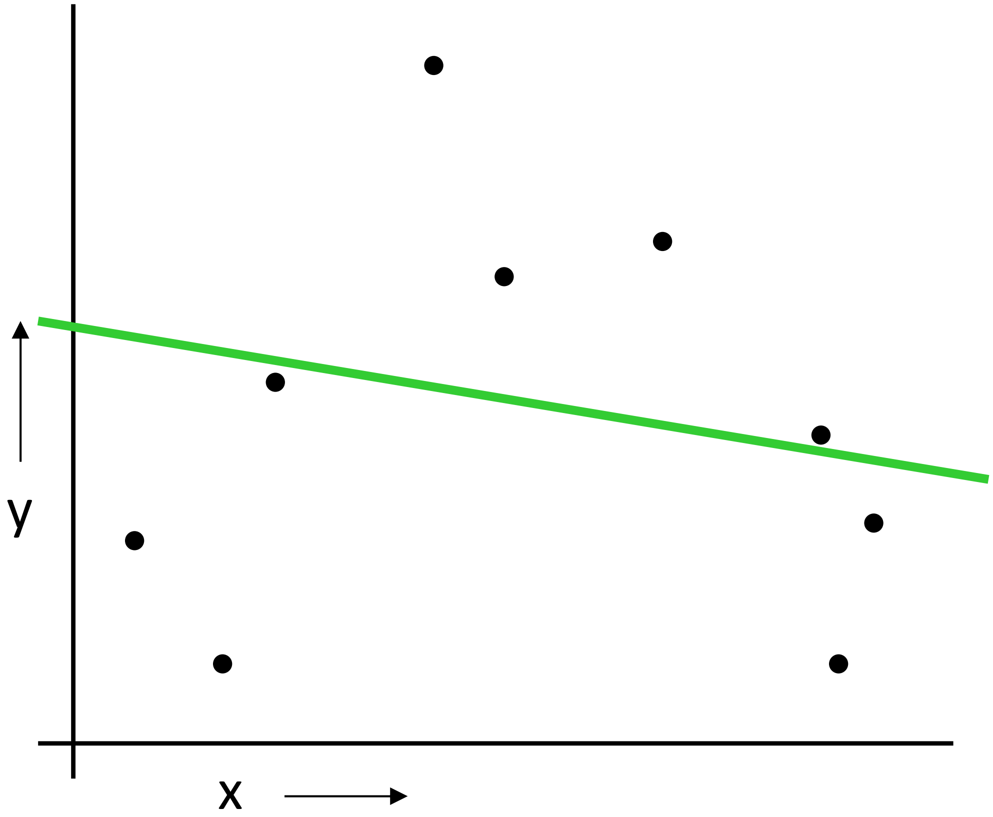


y = f(x) + noise
Can we learn f from this data?

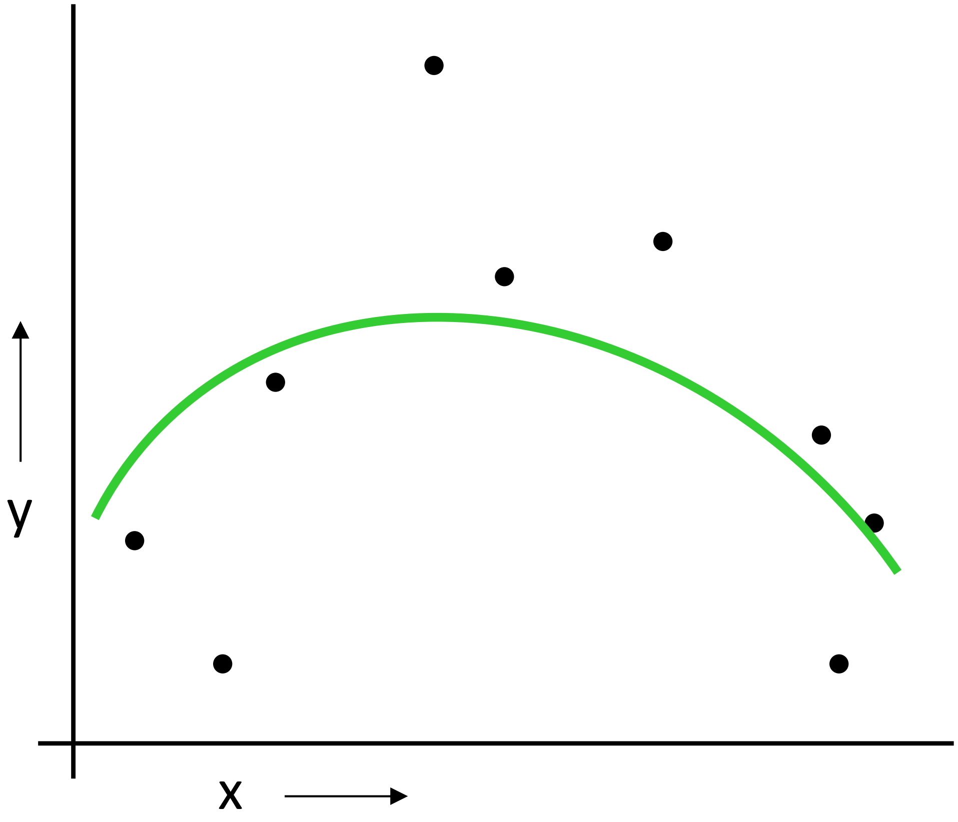
Let’s consider three methods…

Copyright © Andrew W. Moore

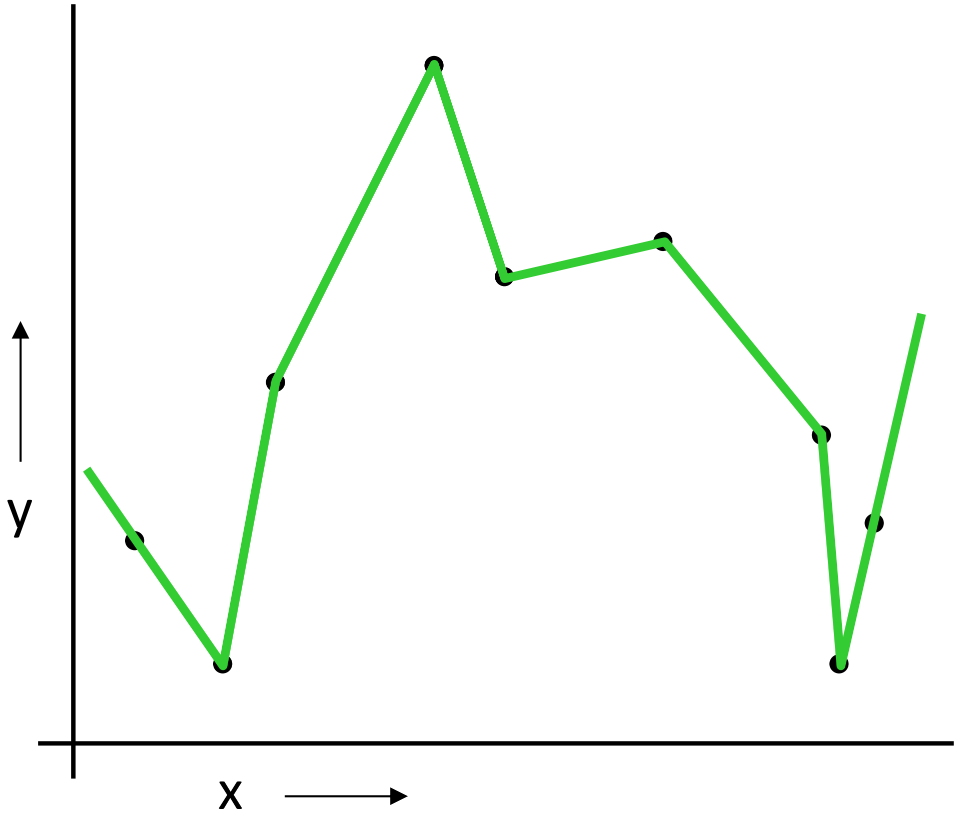
Linear Regression



Quadratic Regression

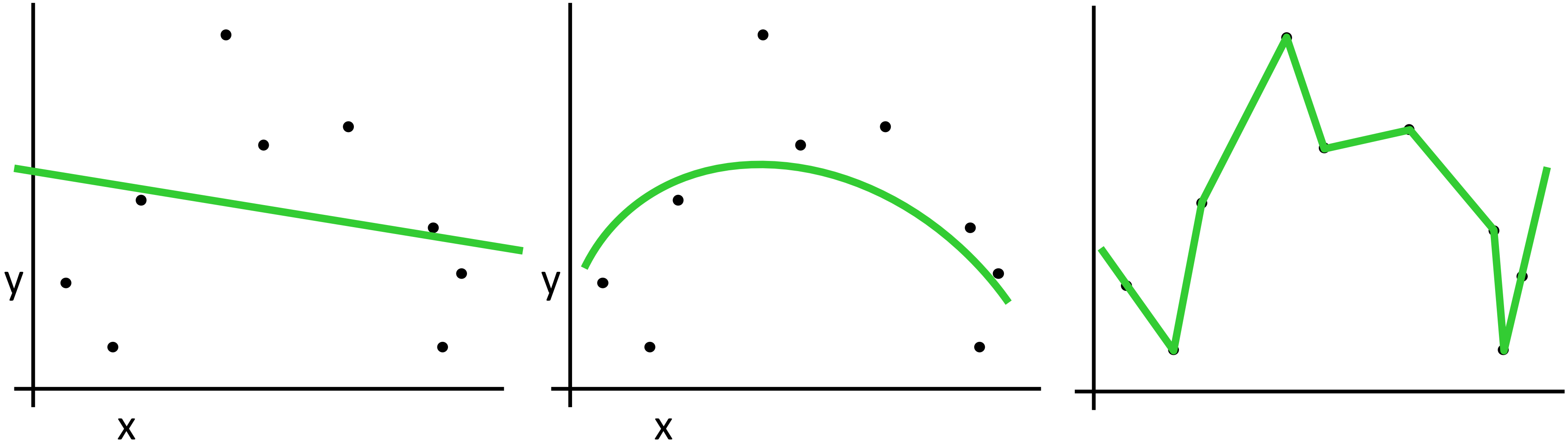


Join-the-dots



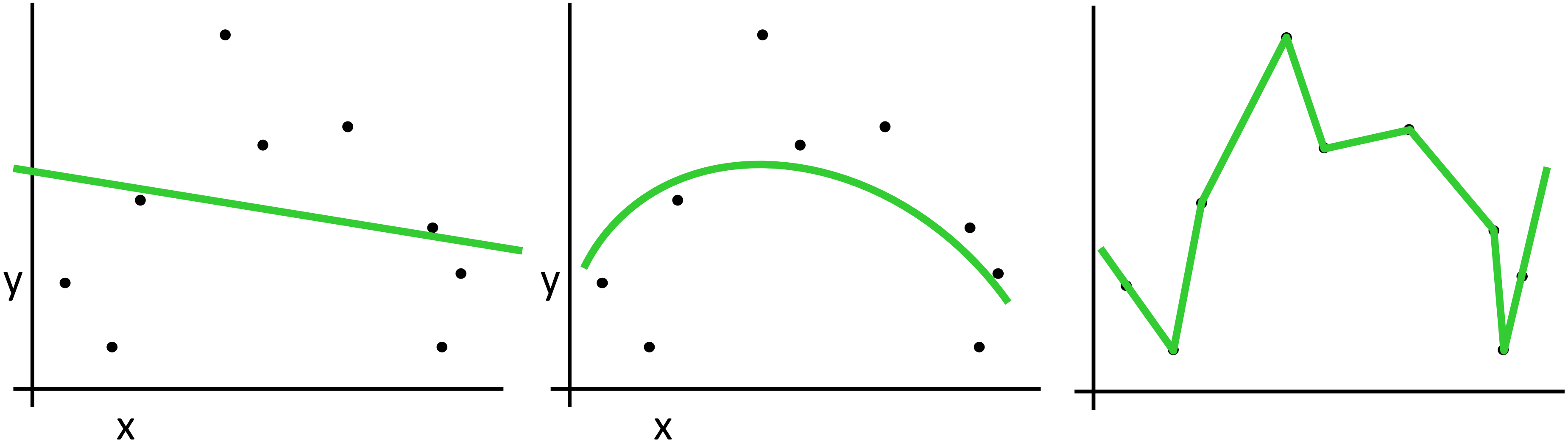
Also known as piecewise linear nonparametric regression if that makes you feel better

Which is best?



Why not choose the method with the best fit to the data?

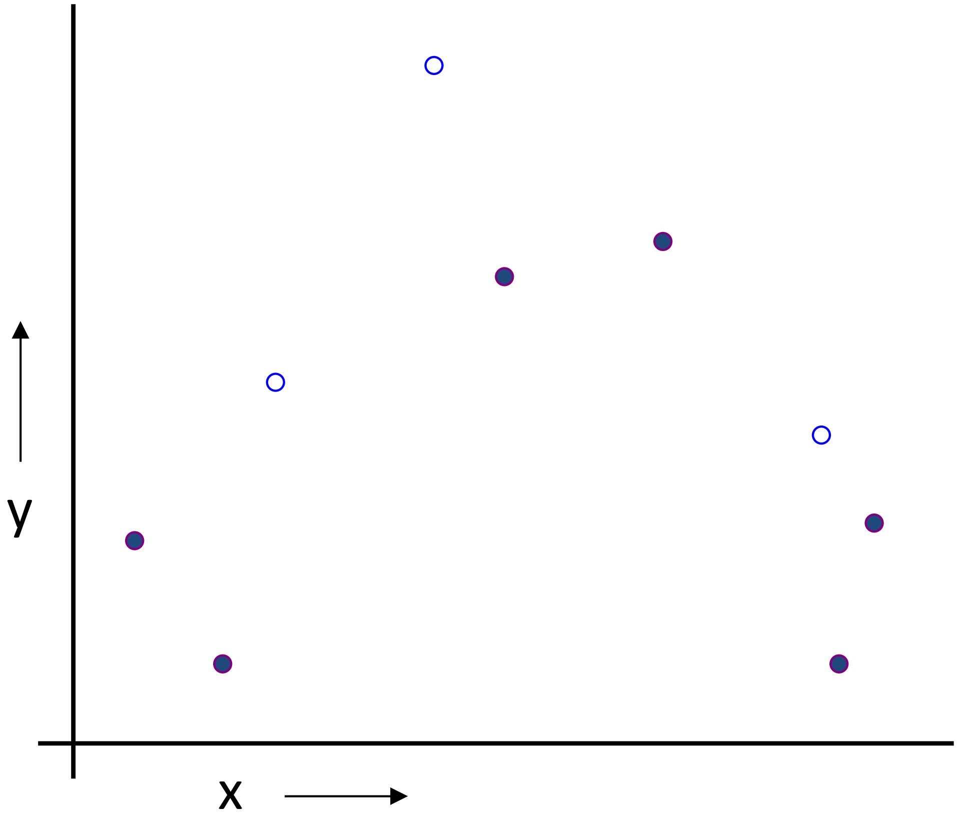
What do we really want?



Why not choose the method with the best fit to the data?

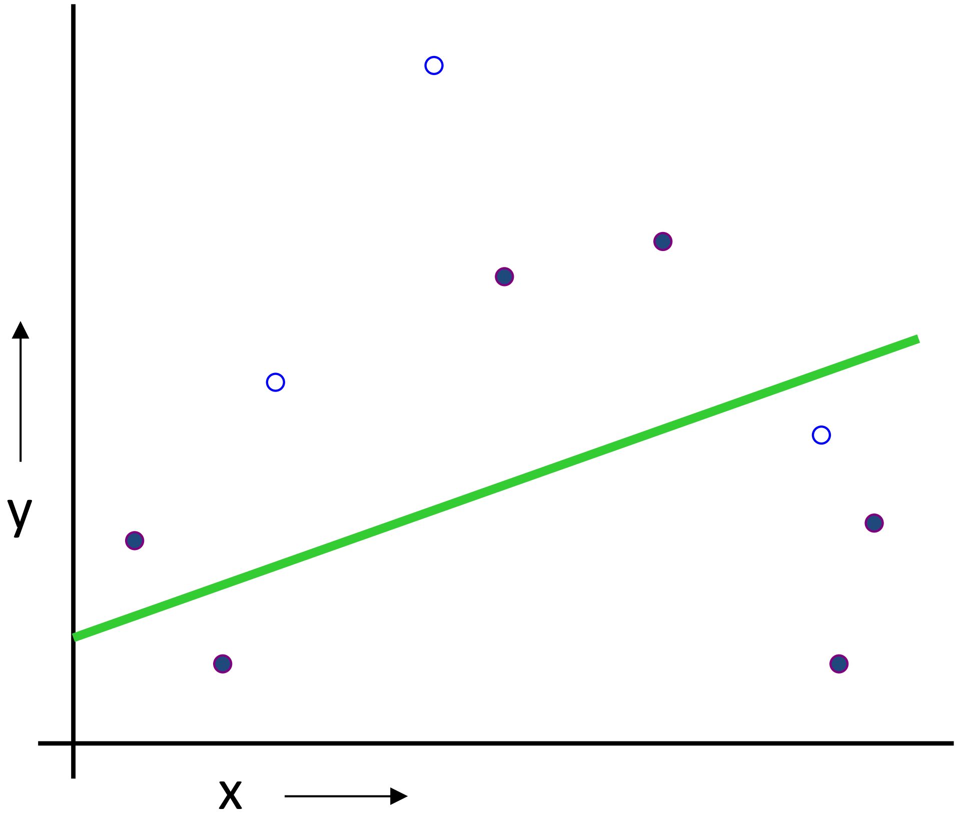
"How well are you going to predict future data drawn from the same distribution?"

The test set method



1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set

The test set method



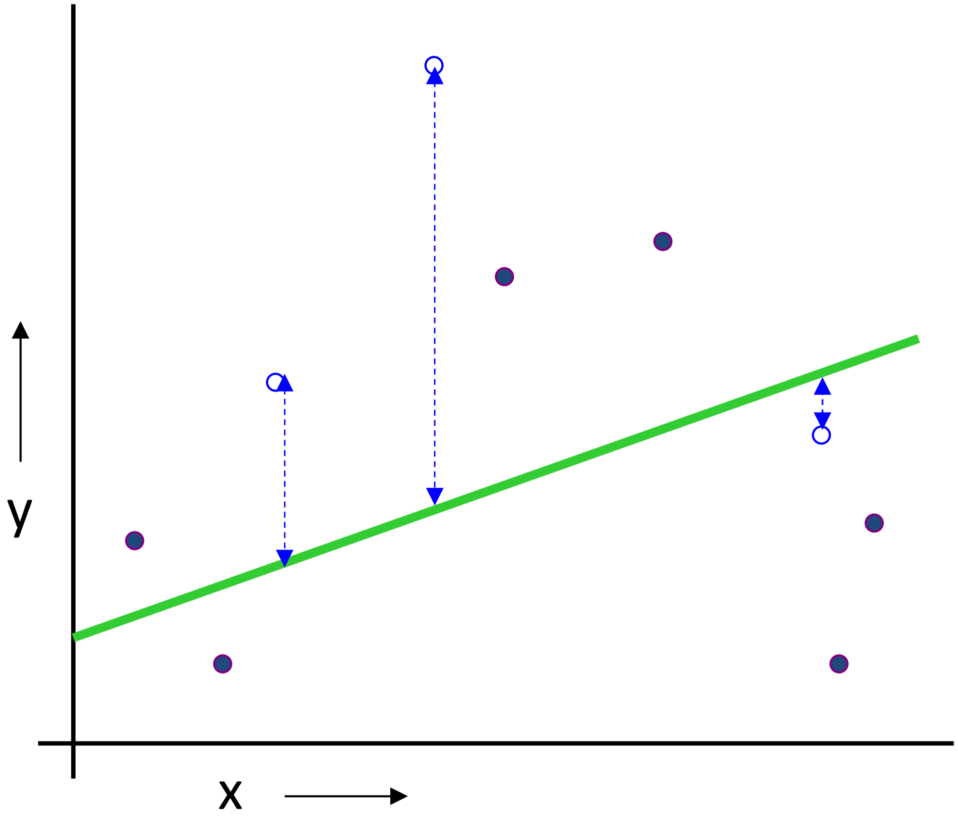
(Linear regression example)

1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

3. Perform your regression on the training set

The test set method



1. Randomly choose 30% of the data to be in a test set

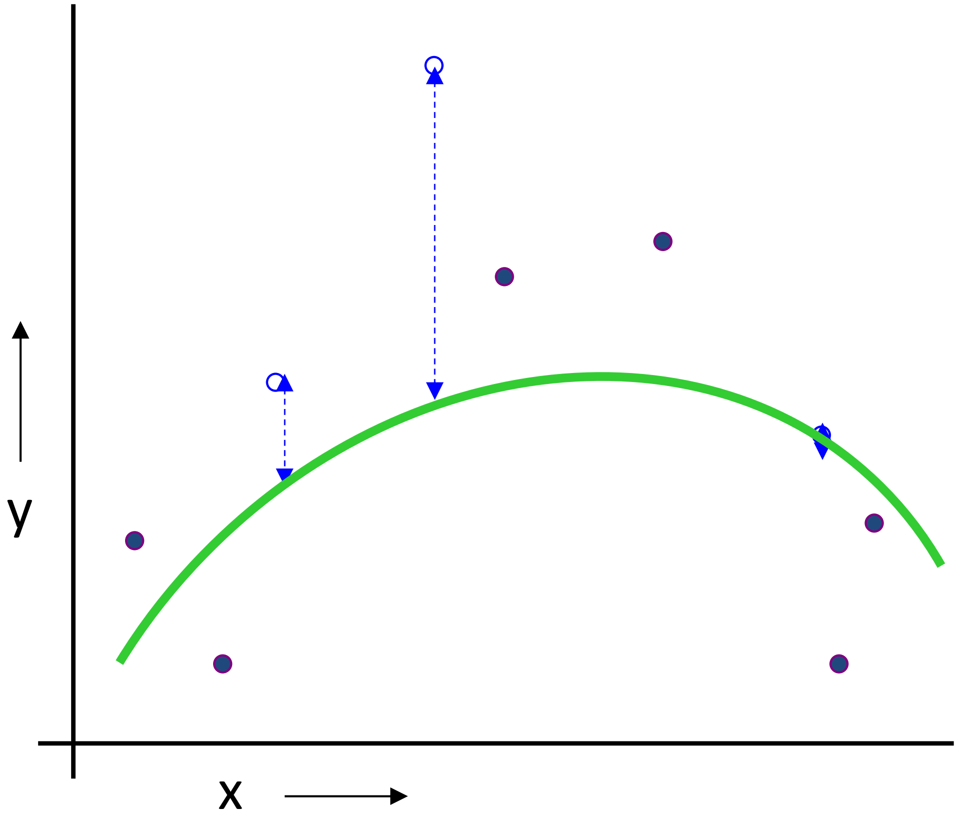
2. The remainder is a training set

3. Perform your regression on the training set

4. Estimate your future performance with the test set

(Linear regression example)
Mean Squared Error = 2.4

The test set method



(Quadratic regression example)
Mean Squared Error = 0.9

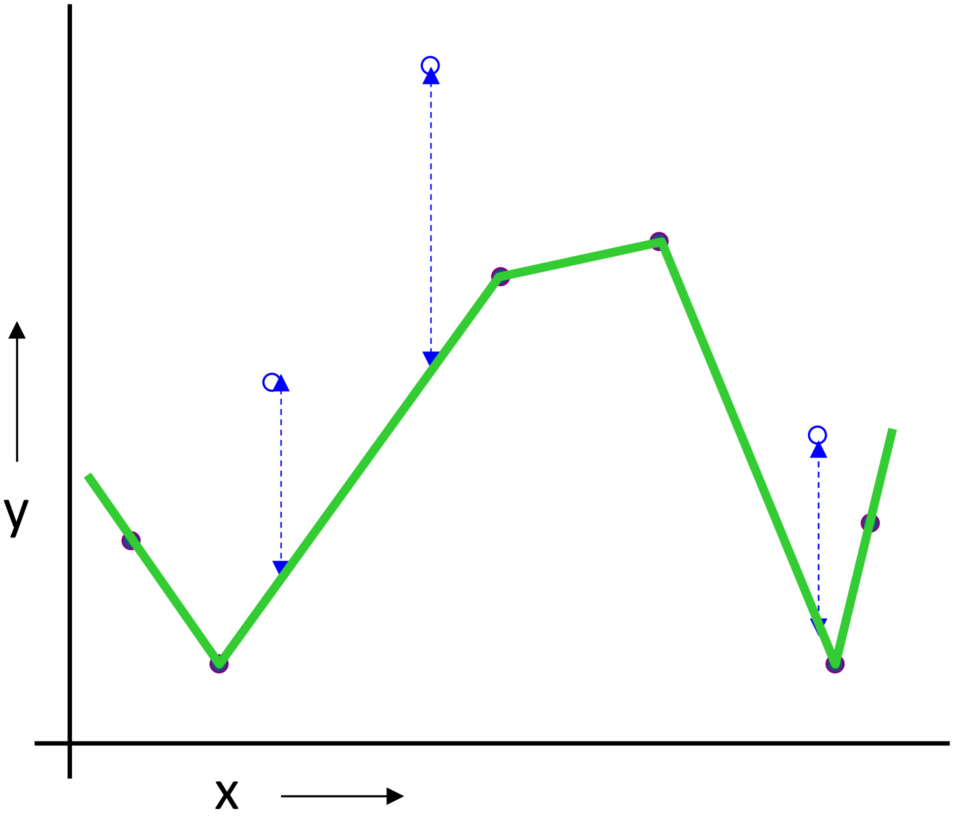
1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

3. Perform your regression on the training set

4. Estimate your future performance with the test set

The test set method



(Join the dots example)
Mean Squared Error = 2.2

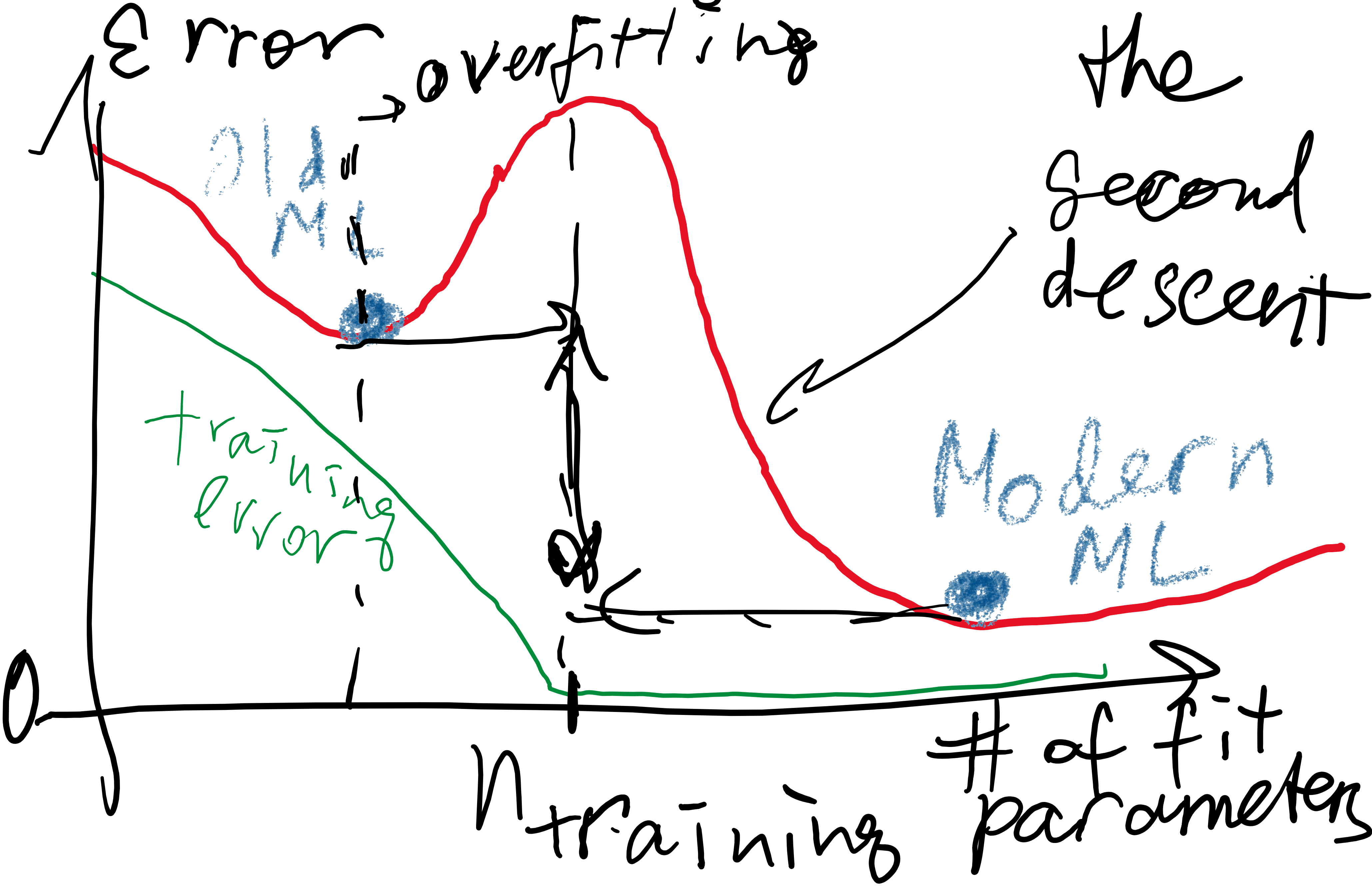
1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

3. Perform your regression on the training set

4. Estimate your future performance with the test set

Double descend- the main reason modern Machine Learning works so well



12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

Suppose the model relating the regressors to the response is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad i = 1, 2, \ldots, n$$

In matrix notation this model can be written as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \qquad (12\text{-}6)$$

12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

12-1.3 Matrix Approach to Multiple Linear Regression

We wish to find the vector $\hat{\beta}$ that minimizes the sum of squares of error terms:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

$$0 = \frac{\partial L}{2\partial \beta} = -X'(\mathbf{y} - \mathbf{X}\beta) = -X'y + (X'X)\boldsymbol{\beta}$$

The resulting least squares estimate is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y} \qquad (12\text{-}7)$$

Analog of $\frac{1}{Var(X)}$

Analog of $Cov(X,Y)$

Multiple Linear Regression Model

$$\hat{\boldsymbol\beta} = (\mathbf{X}'\mathbf{X})^{-1}\ \mathbf{X}'\mathbf{y}$$

H is an idempotent matrix

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y,$$

$$\hat{y} = Hy, \quad \text{and} \quad e = (I - H)y.$$

$$H = H' ; \quad H^2 = X\underbrace{(X'X)^{-1} X'X}_{I}(X'X)^{-1}X' = X(X'X)^{-1}X' = H$$

Vectors $\hat{y}$ & e are orthogonal since

$$\hat{y}' \cdot e = y'H(I-H)y = 0$$ since

$$\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{H} - \mathbf{H}^2 = \mathbf{H} - \mathbf{H} = 0.$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Unbiased estimators:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}] \\ &= \boldsymbol{\beta} \end{aligned}$$

Covariance Matrix of Estimators:

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix}$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Individual variances and covariances:

$$V(\hat{\beta}_j) = \sigma^2 C_{jj}, \qquad j = 0, 1, 2$$

$$\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}, \qquad i \neq j$$

In general,

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\mathbf{C}$$

12-1: Multiple Linear Regression Models

Estimating error variance σ_{ε}^2

An unbiased estimator of error variance σ_{ε}^2 is

$$\hat{\sigma}_{\varepsilon}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - p} = \frac{SS_E}{n - p} \qquad (12\text{-}16)$$

Here p=k+1 for k-variable multiple linear regression

R^2 and Adjusted R^2

The **coefficient of multiple determination R^2**

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

The **adjusted R^2** is

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \approx 1 - \frac{\hat{\sigma}^2_\varepsilon}{S^2_y} \qquad (12\text{-}23)$$

- The adjusted R^2 statistic penalizes adding terms to the MLR model.
- It can help guard against overfitting (including regressors that are not really useful)

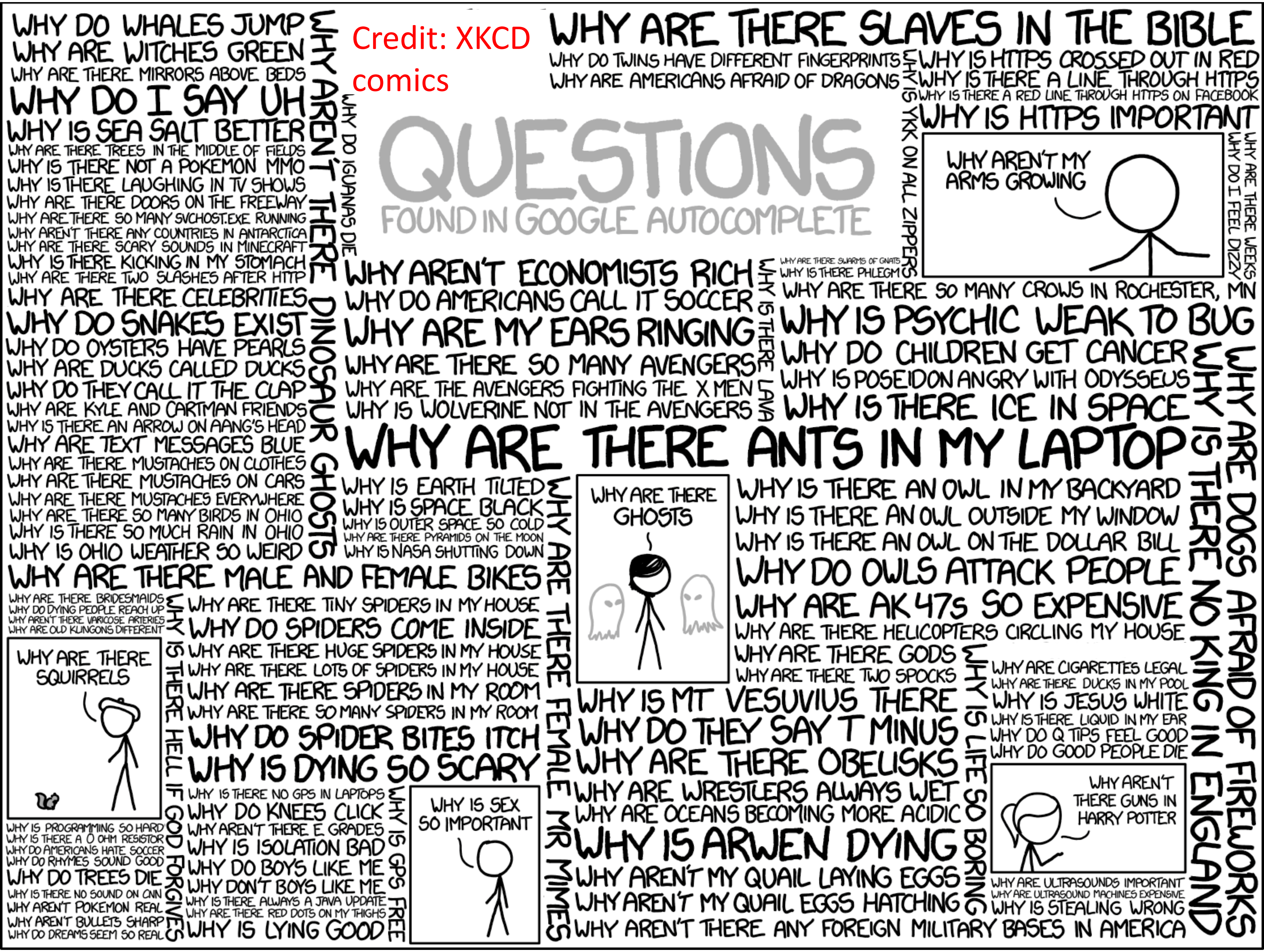
How to know where to stop adding variables?

- Adding new variables x_i to MLR
  watch the adjusted R^2

- Once the adjusted R^2
  no longer increases = stop.
  Now you did the best you can.

Matlab exercise

- Every group works with g0=2907; g1=1527; g2=2629; g3=2881; g4=1144; g5=1066;
- Compute Multiple Linear Regression (MLR): where y=exp_t (g0); x1= exp_t (g1); x2= exp_t (g2);
- How much better the MLR did compared to the Single Linear Regression (SLR)?
- Continue increasing the number of genes in x until R_adj starts to decrease


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

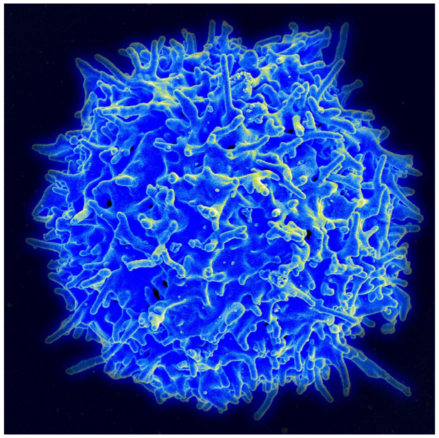
Credit: XKCD comics

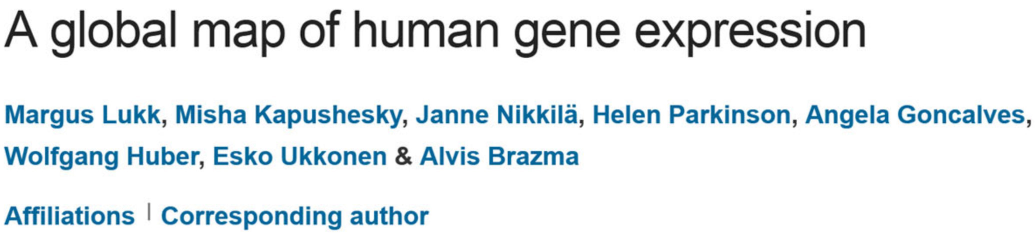
Clustering analysis of gene expression data

Chapter 11 in
Jonathan Pevsner,
Bioinformatics and Functional Genomics,
3rd edition
(Chapter 9 in 2nd edition)

Human T cell expression data

- The matrix contains 47 expression samples from Lukk et al, Nature Biotechnology 2010
- All samples are from T cells in different individuals
- Only the top 3000 genes with the largest variability were used
- The value is log2 of gene's expression level in a given sample as measured by the microarray technology

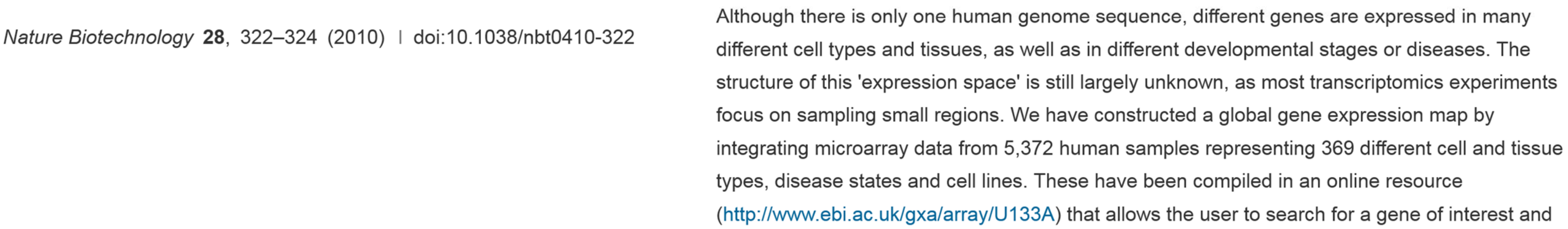




A global map of human gene expression

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

Affiliations | Corresponding author



Nature Biotechnology **28**, 322–324 (2010) | doi:10.1038/nbt0410-322

Although there is only one human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. The structure of this 'expression space' is still largely unknown, as most transcriptomics experiments focus on sampling small regions. We have constructed a global gene expression map by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. These have been compiled in an online resource (http://www.ebi.ac.uk/gxa/array/U133A) that allows the user to search for a gene of interest and

$2500
WILD
00
$900
$700
$600
$650
$500
$700
BANKRUPT
ONE MILLION
BANKRUPT
$900
$550
$500
$900
BANKRUPT
$650
Free PLAY
$700
LOSE A TURN
$800
$500
$650
$500
$900
BANKRUPT

WHEEL OF FORTUNE®

Correlated pairs

plausible biological connection based on short description

g1=1994; g2=188; group 1

g1=2872; g2=1269; group 2

g1=1321; g2=10; group 3

g1= 886; g2=819; group 4

g1=2138; g2=1364; group 5

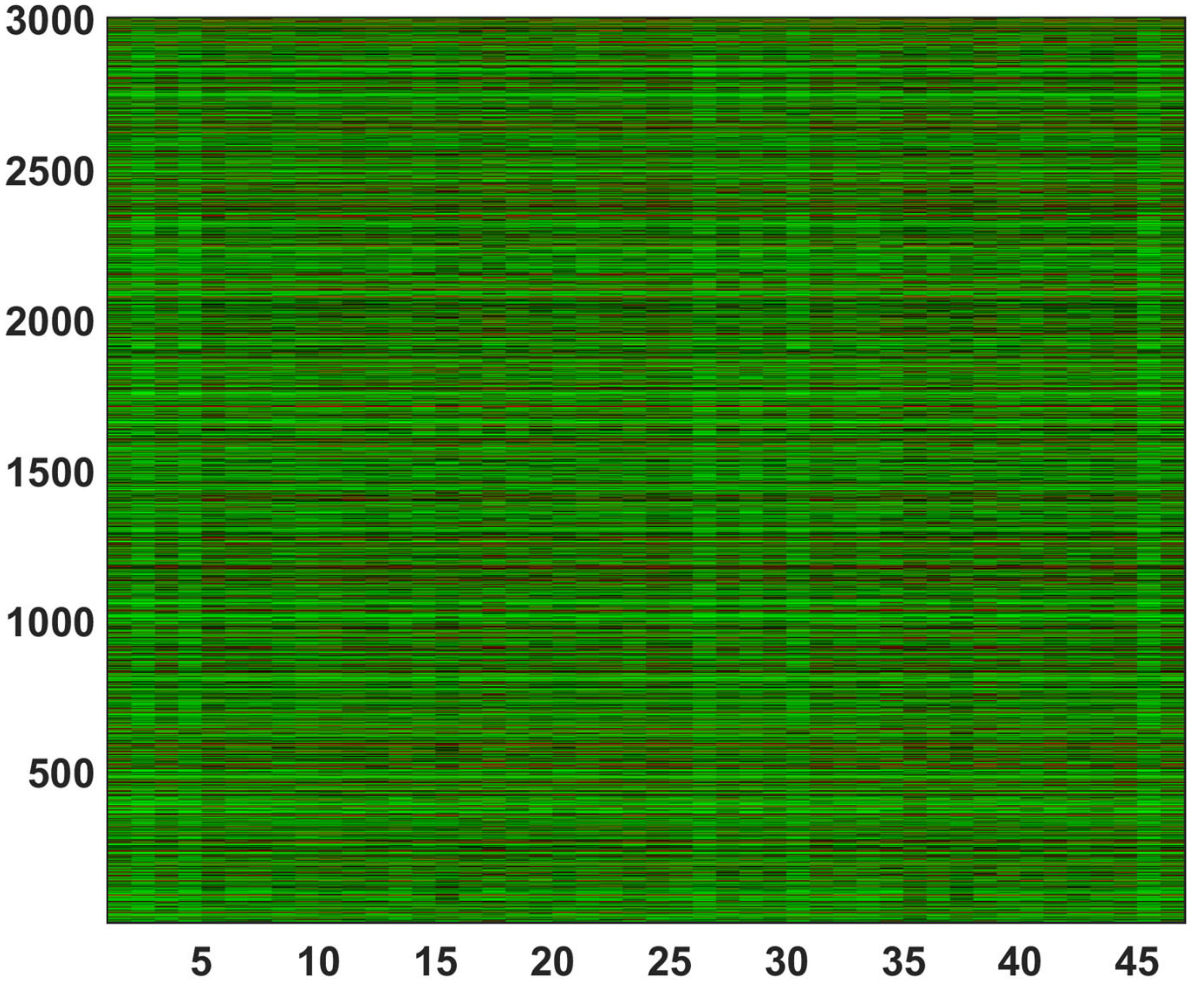
no obvious biological common function

```
g1=1+floor(rand.*3000); g2=1+floor(rand.*3000);
disp([g1, g2])
```

Matlab exercise

- Every group works with
  g0=2907; g1=1527; g2=2629; g3=2881;
  g4=1144; g5=1066;
- Compute Multiple Linear Regression (MLR),
  where y=exp_t (g0);
  x1= exp_t (g1); x2= exp_t (g2);
- How much better the MLR did compared to the Single Linear Regression (SLR)?
- Continue increasing the number of genes in x until R_adj starts to decrease

How to find the entire groups of mutually correlated genes if you have many genes and many samples?



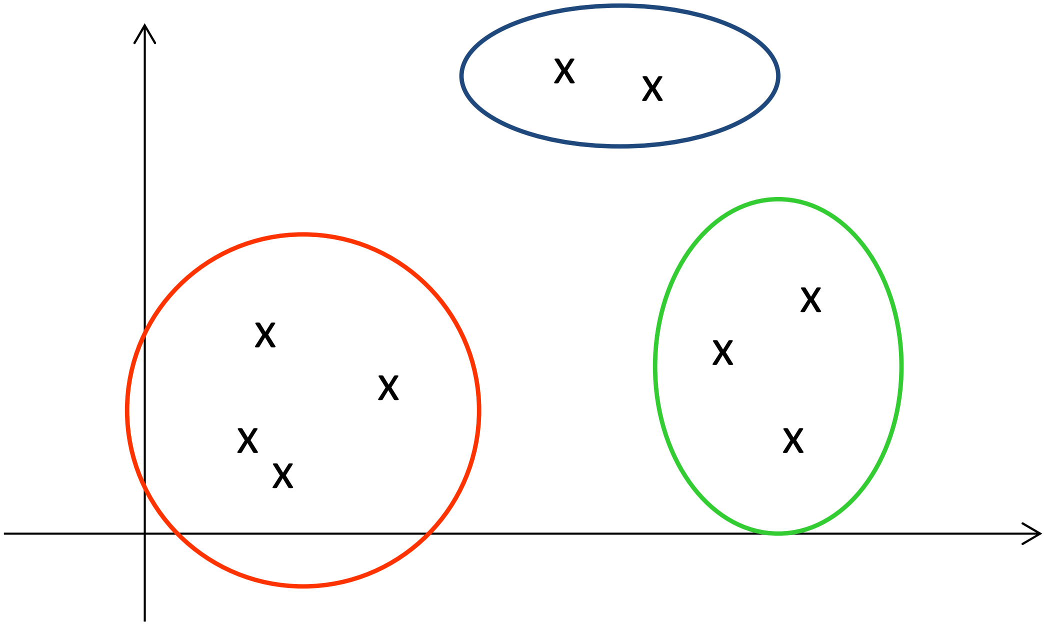
Clustering to the rescue!

Clustering is a part of Machine Learning

- ***Supervised Learning:***
  A machine learning technique whereby a system uses a set of human-labelled training examples to learn how to correctly perform a task
  Example: a sample of cancer expression profiles each **annotated** with cancer type
  Goal: predict cancer type based on expression pattern
- ***Unsupervised Learning (including clustering):***
  In machine learning, unsupervised learning is a class of problems in which one seeks to determine how the data are organized. One only has unlabeled examples.
  Example: a sample of breast cancer expression profiles.
  Goal: Identify several different (yet unknown) subtypes with potentially different treatments

What is clustering?

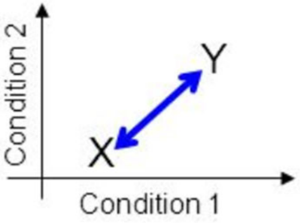
- The goal of clustering is to
  - group data points that are close (or **similar**) to each other
  - Usually, one needs to identify such groups (or clusters) in an **unsupervised** manner
  - Sometimes one takes into account **prior information** (Bayesian methods)
- Need to define some distance d_{ij} between objects i and j
- Clustering is easy in 2 dimensions but hard in 3000 dimensions -> need to somehow reduce dimensionality



How to define the distance?

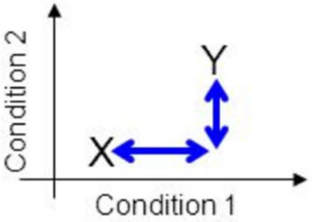
- Euclidean distance:
  - Most commonly used distance
  - Sphere shaped cluster
  - Corresponds to the geometric distance into the multidimensional space

$$d(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}$$



- City Block (Manhattan) distance:
  - Sum of differences across dimensions
  - Less sensitive to outliers
  - Diamond shaped clusters

$$d(X,Y) = \sum_i |x_i - y_i|$$



The Canberra distance metric is calculated in R by

$$\sum \left(\frac{|x_i - y_i|}{|x_i + y_i|} \right).$$

Correlation coefficient distance

$$d(X,Y) = 1 - \rho(X,Y) = 1 - \frac{Cov(X,Y)}{\sqrt{(Var(X) \cdot Var(Y)}}$$

Common types of clustering algorithms

- Hierarchical if one doesn’t know in advance the # of clusters
  - Agglomerative: start with N clusters and gradually merge them into 1 cluster
  - Divisive: start with 1 cluster and gradually break it up into N clusters
- Non-hierarchical algorithms
  - K-means clustering:
    - Iteratively apply the following two steps:
    - Calculate the centroid (center of mass) of each cluster
    - Assign each to the cluster to the nearest centroid
  - Principal Component Analysis (PCA)
    - plot pairs of top eigenvectors of the covariance matrix Cov(X_i, X_j) and uses visual information to group

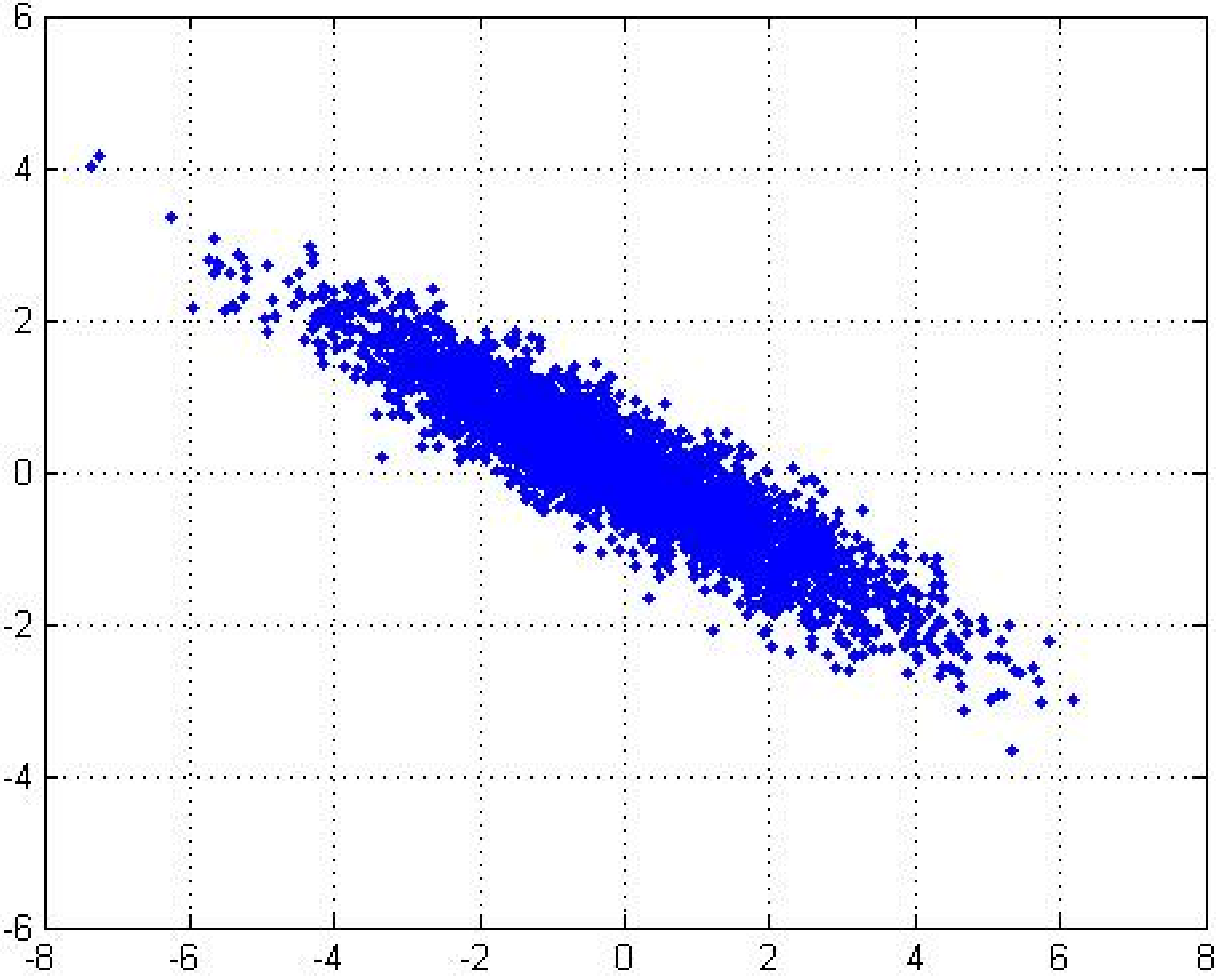
The Principal Components

- **Vectors** originating from the center of mass

- Principal component #1 points
  in the direction of the **largest variance**.

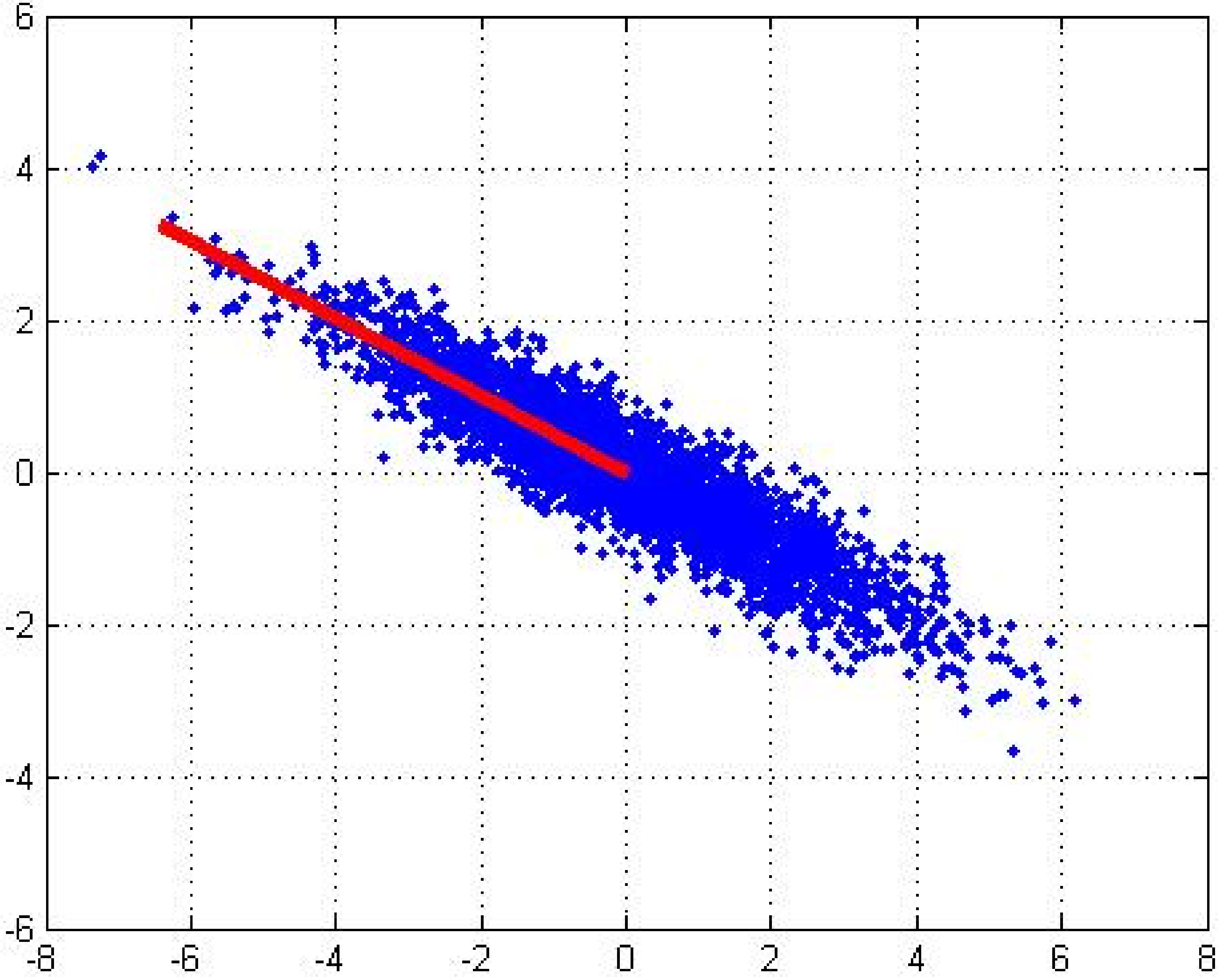
- Each subsequent principal component…
  - is **orthogonal** to the previous ones, and
  - points in the directions of the **largest variance of the residual subspace**

2D Gaussian dataset



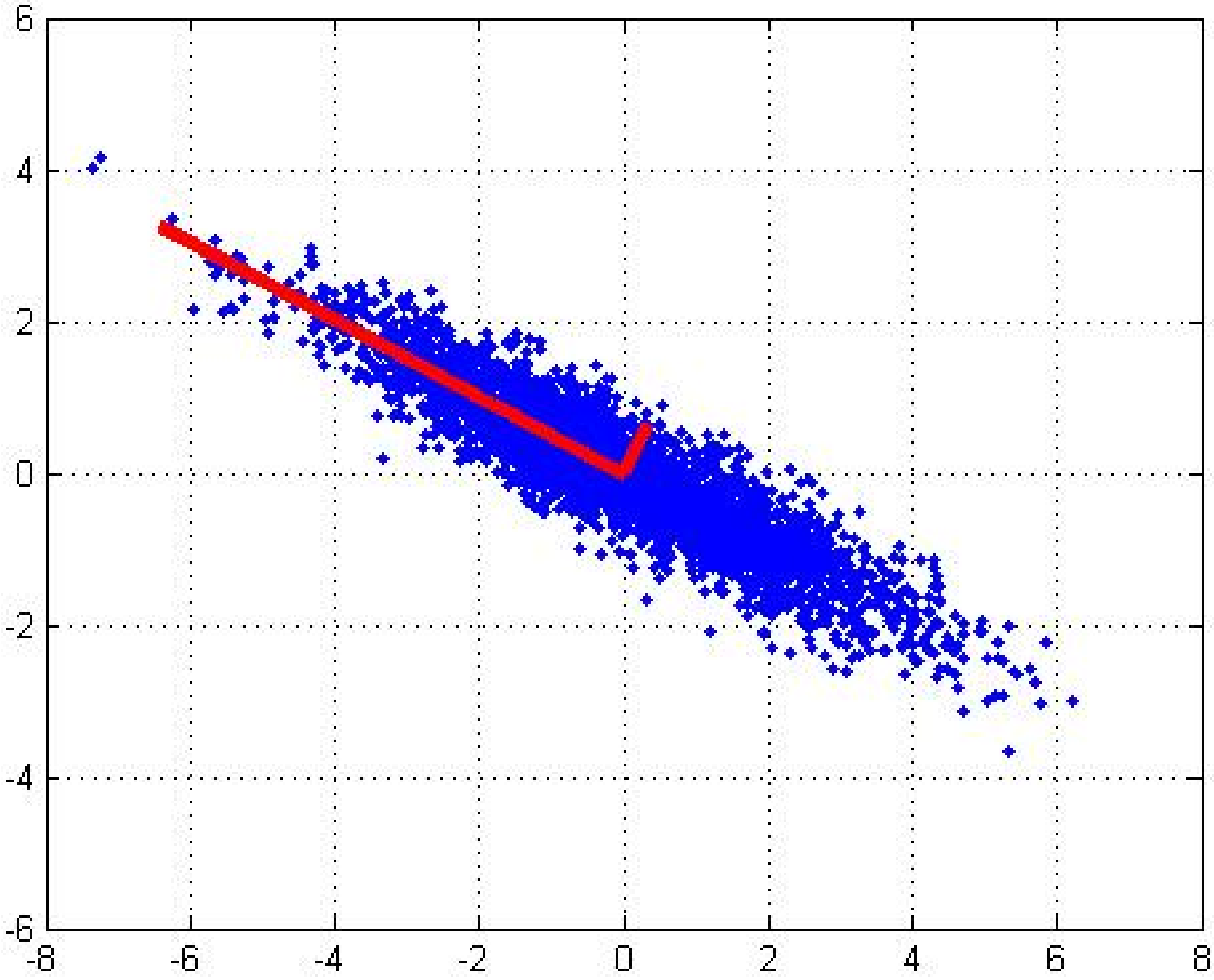
Adapted from lectures Prof. Pat Virtue at CMU based on original slide from Barnabas Poczos

1st PCA axis



Adapted from lectures Prof. Pat Virtue at CMU based on original slide from Barnabas Poczos

2nd PCA axis



Adapted from lectures Prof. Pat Virtue at CMU based on original slide from Barnabas Poczos

Data for PCA

$$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$$

$$\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \vdots \\ (\mathbf{x}^{(N)})^T \end{bmatrix}$$

We assume the data is **centered**

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)} = \mathbf{0}$$

Q: What if your data is **not** centered?

A: Subtract off the sample mean

Adapted from lectures Prof. Pat Virtue at CMU based on the original slides from Matt Gormley

Sample Covariance Matrix

The sample covariance matrix is given by:

$$\Sigma_{jk} = \frac{1}{N} \sum_{i=1}^{N} (x_j^{(i)} - \mu_j)(x_k^{(i)} - \mu_k)$$

Since the data matrix is centered, we rewrite as:

$$\mathbf{\Sigma} = \frac{1}{N} \mathbf{X}^T \mathbf{X}$$

$$\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \vdots \\ (\mathbf{x}^{(N)})^T \end{bmatrix}$$

Adapted from lectures Prof. Pat Virtue at CMU based on the original slides from Matt Gormley

PCA algorithm

PCA algorithm($\mathbf{X}$, k): top k eigenvalues/eigenvectors

- { λ_i, $\mathbf{u}_i$ }$_{i=1:m}$ = eigenvectors/eigenvalues of Σ … $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$
- **PCA** basis vectors = the eigenvectors of Σ
- Larger eigenvalue ⇒ more important eigenvectors

PCA and units

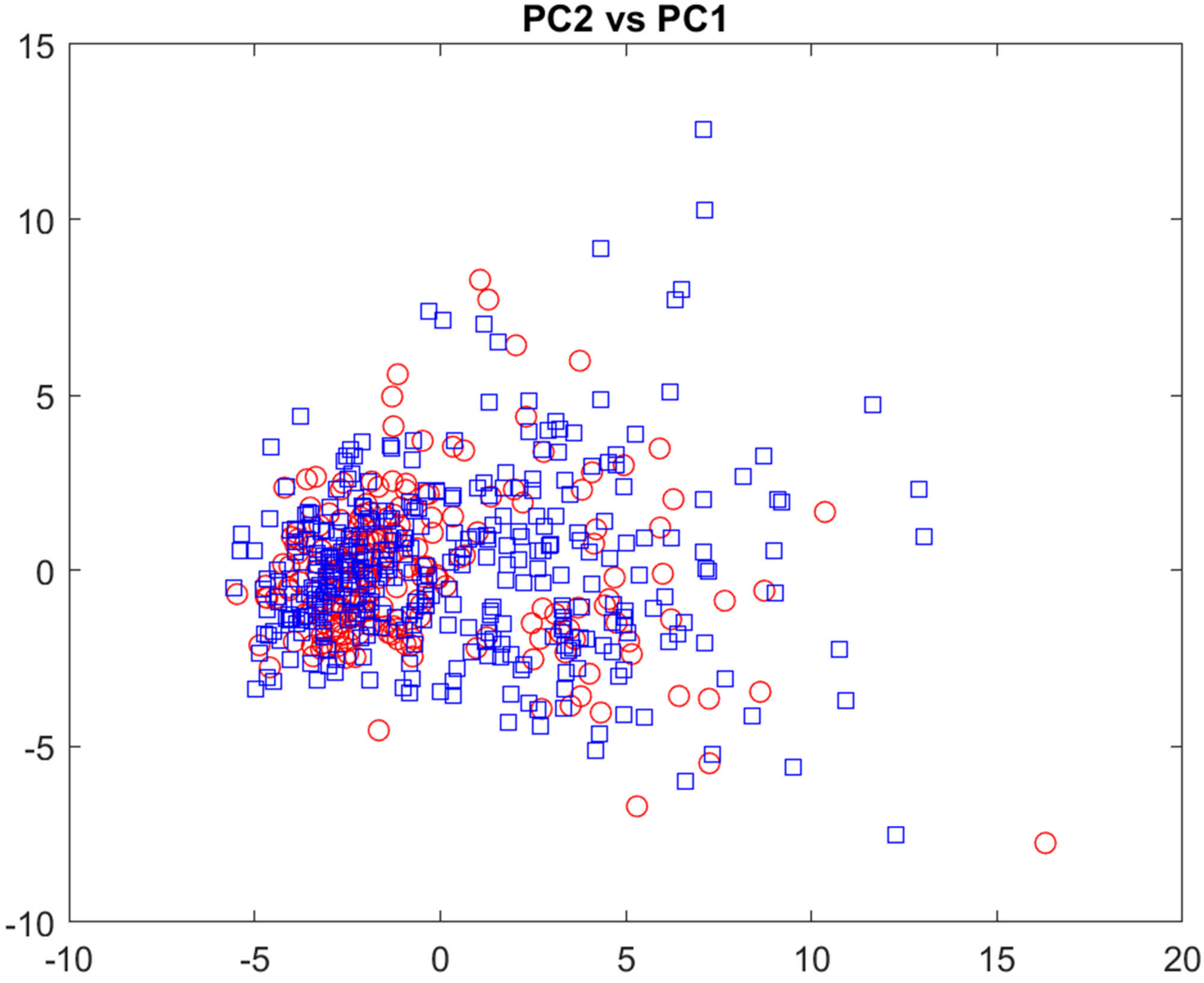
- When different variables have different units (like temperature and mass), the meaning of principal components is a somewhat arbitrary

- One way of making the PCA less arbitrary is to use variables scaled so as to have unit variance, by standardizing the data

- Before making PCA of X transform it using Z=zscore(X);

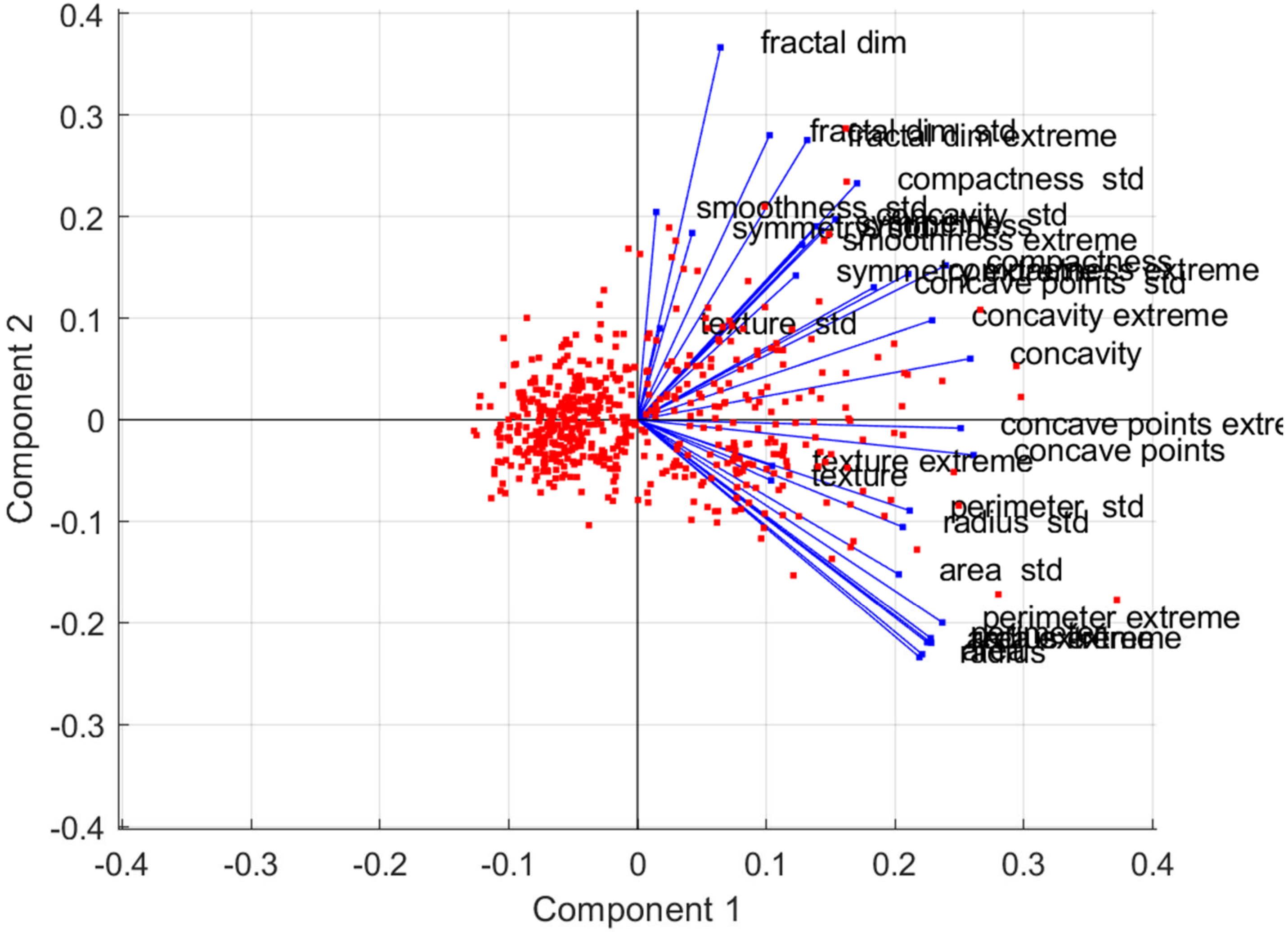
Group project 4

- load cancer_wdbc.mat
- Z=zscore(cancerwdbc);
- [coeff_z, score_z, latent_z] = pca(Z);
- ic=find(cancer_yn==1); whos ic; inc=find(cancer_yn==0); whos inc;
- figure; plot(score_z(ic,1), score_z(ic,2),'ro'); hold on; plot(score_z(inc,1), score_z(inc,2),'bs’); title('PC2 vs PC1');
- Plot pairs of score_z components
  - 1st principal component vs 2nd principal component.
  - 1st principal component vs 3rd principal component
  - 3rd principal component vs 2nd principal component

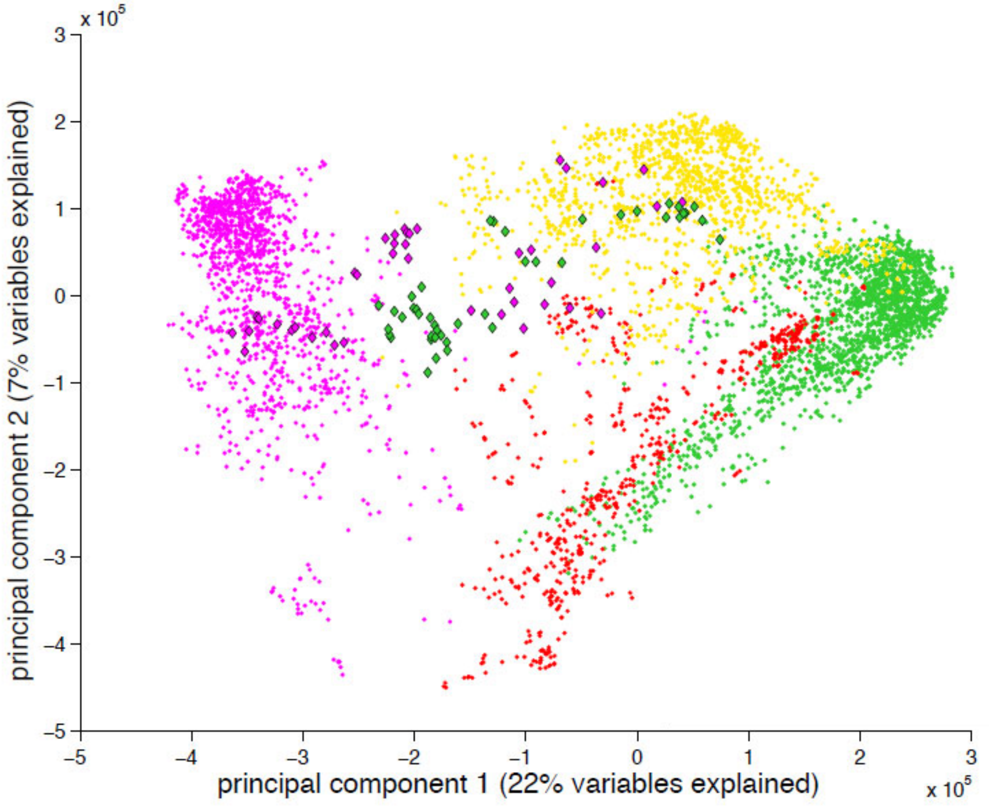

PC2 vs PC1
15
10
5
0
-5
-10
-10
-5
0
5
10
15
20

Which variables contribute to which PC? Add loadings (coeff eigenvectors)

- figure; biplot(coeff_z(:,1:2),'scores',score_z(:,1:2), 'VarLabels', feature_names);



Example of Principal Component Analysis (PCA) clustering



7000 gene expression samples of model plant *Arabidopsis thaliana*

●	root
●	seedling
●	vegetative(leaf+shoot+stem)
●	reproductive(flower+seed+silique)
◆	cultured root
◆	cultured aerial

Plant J. 2016 Mar 25. doi: 10.1111/tpj.13175. [Epub ahead of print]

Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis.

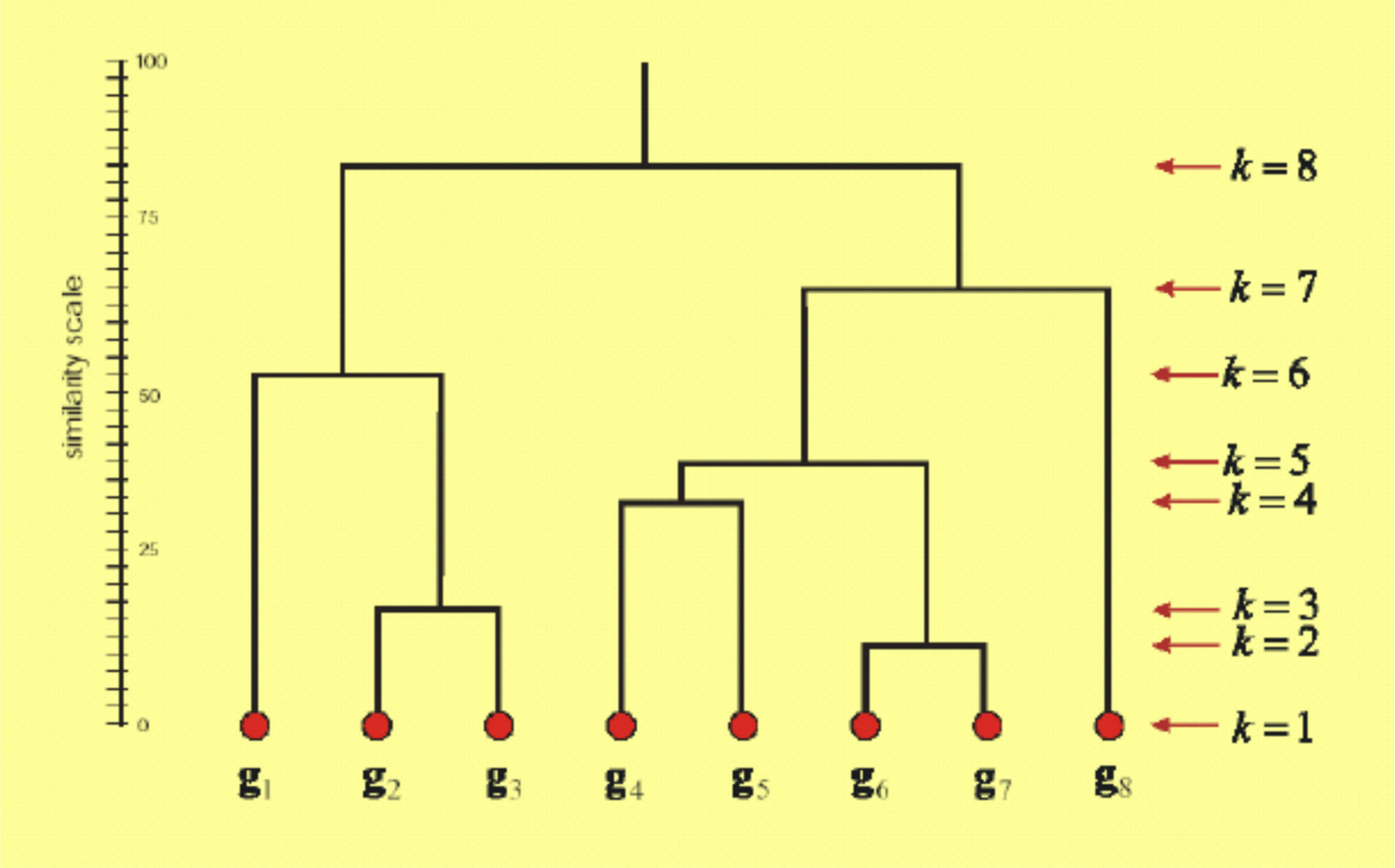
He F[1], Yoo S[2,3], Wang D[4], Kumari S[5], Gerstein M[4], Ware D[5,6], Maslov S[1,7].

Hierarchical clustering

UPGMA algorithm

- Hierarchical agglomerative clustering algorithm
- **UPGMA** = Unweighted Pair Group Method with Arithmetic mean
- Iterative algorithm:
- Start with a pair with the smallest d(X,Y)
- Cluster these two together and replace it with their arithmetic mean (X+Y)/2
- Recalculate all distances to this new “cluster node”
- Repeat until all nodes are merged

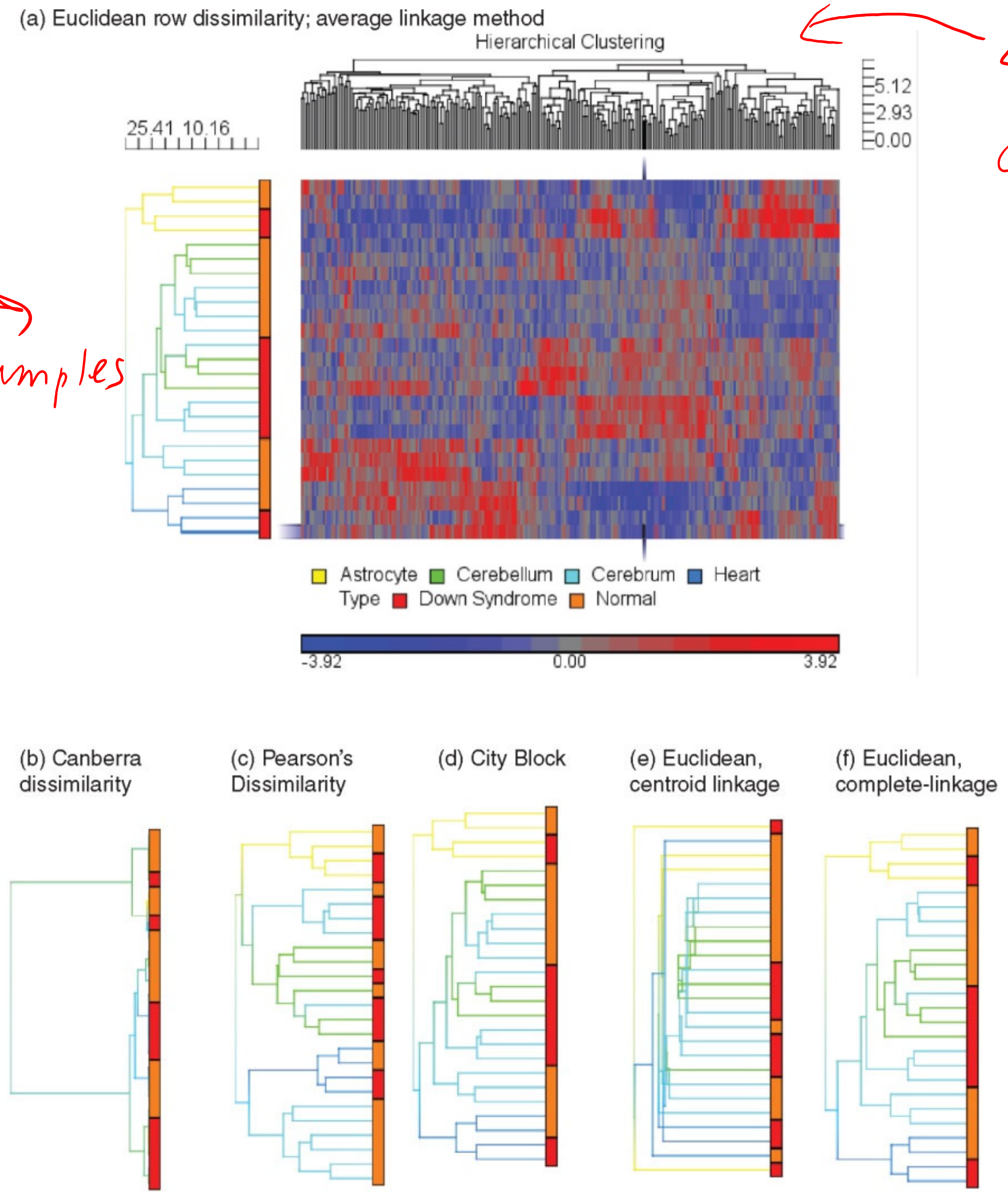
Output of UPGMA algorithm



UPGMA algorithm

250 genes on chromosome 21

25 samples


(a) Euclidean row dissimilarity; average linkage method

(b) Canberra dissimilarity

(c) Pearson's Dissimilarity

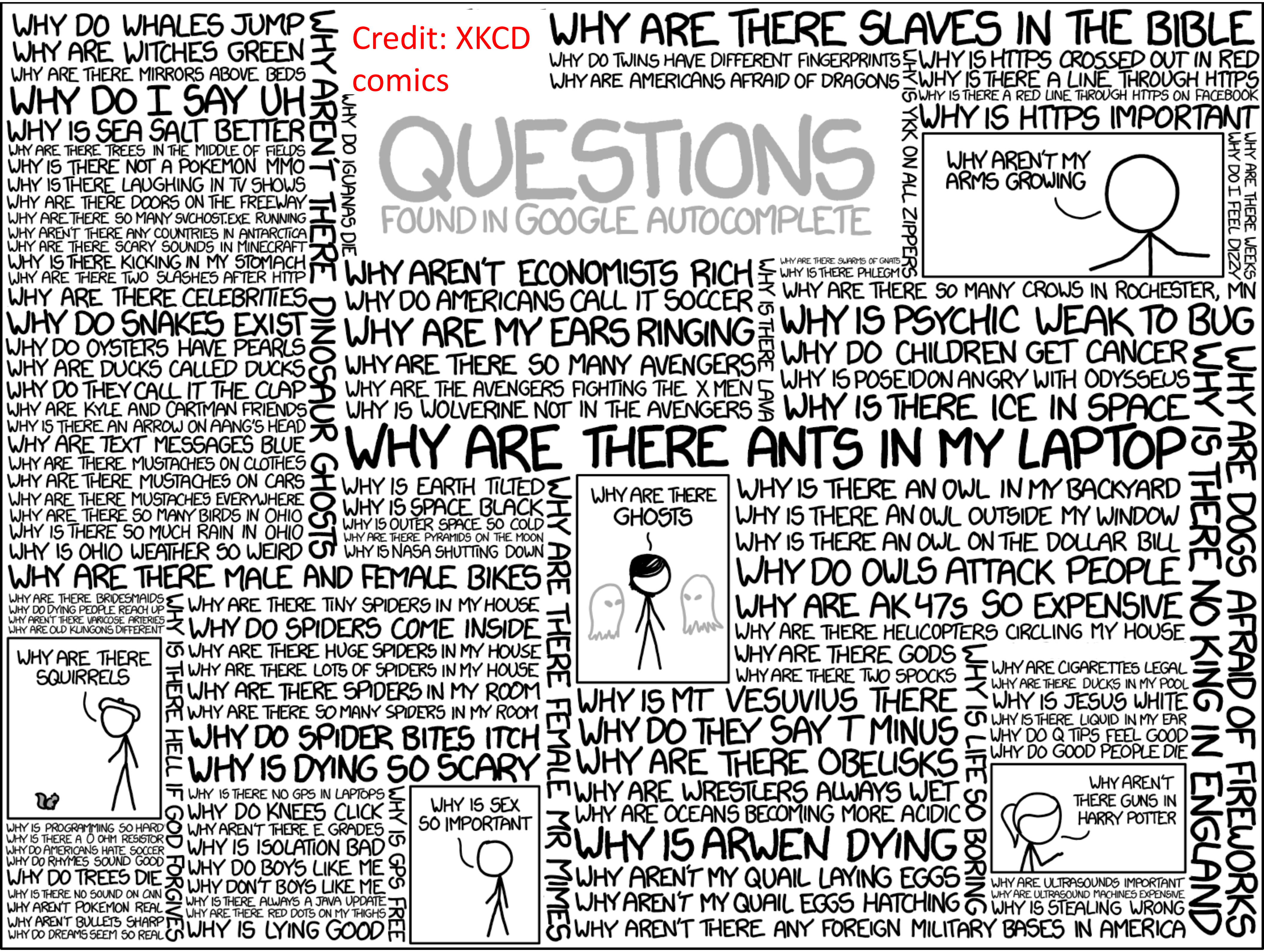
(d) City Block

(e) Euclidean, centroid linkage

(f) Euclidean, complete-linkage

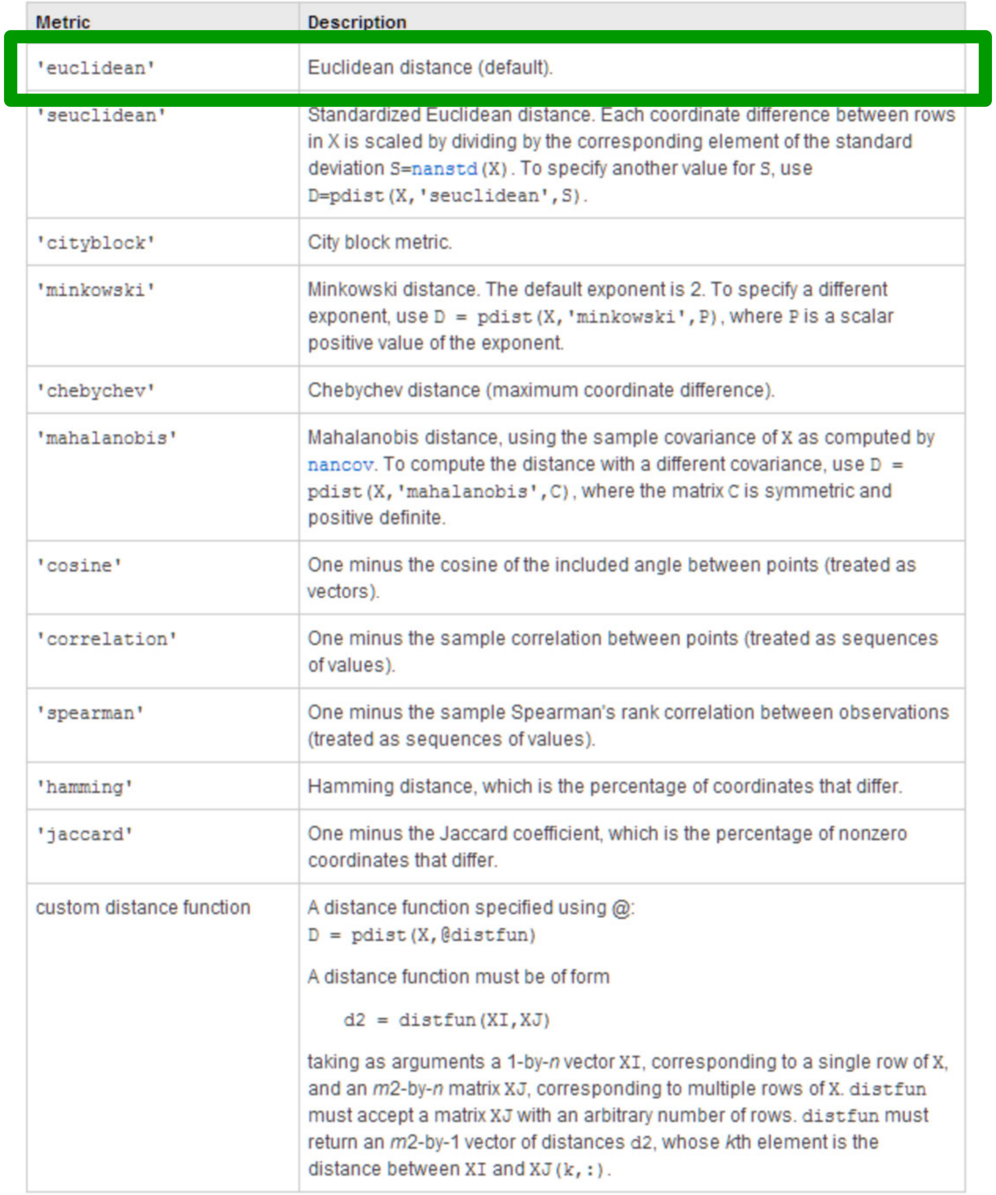
FIGURE 11.16 Hierarchical clustering of 250 chromosome 21 transcripts in 25 samples using Partek software. (a) Hierarchical clustering of microarray data using the default settings of Euclidean dissimilarity for rows (samples) and columns (transcripts). Colors correspond to expression intensity values.

Credit: XKCD comics


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Clustering
Matlab demo

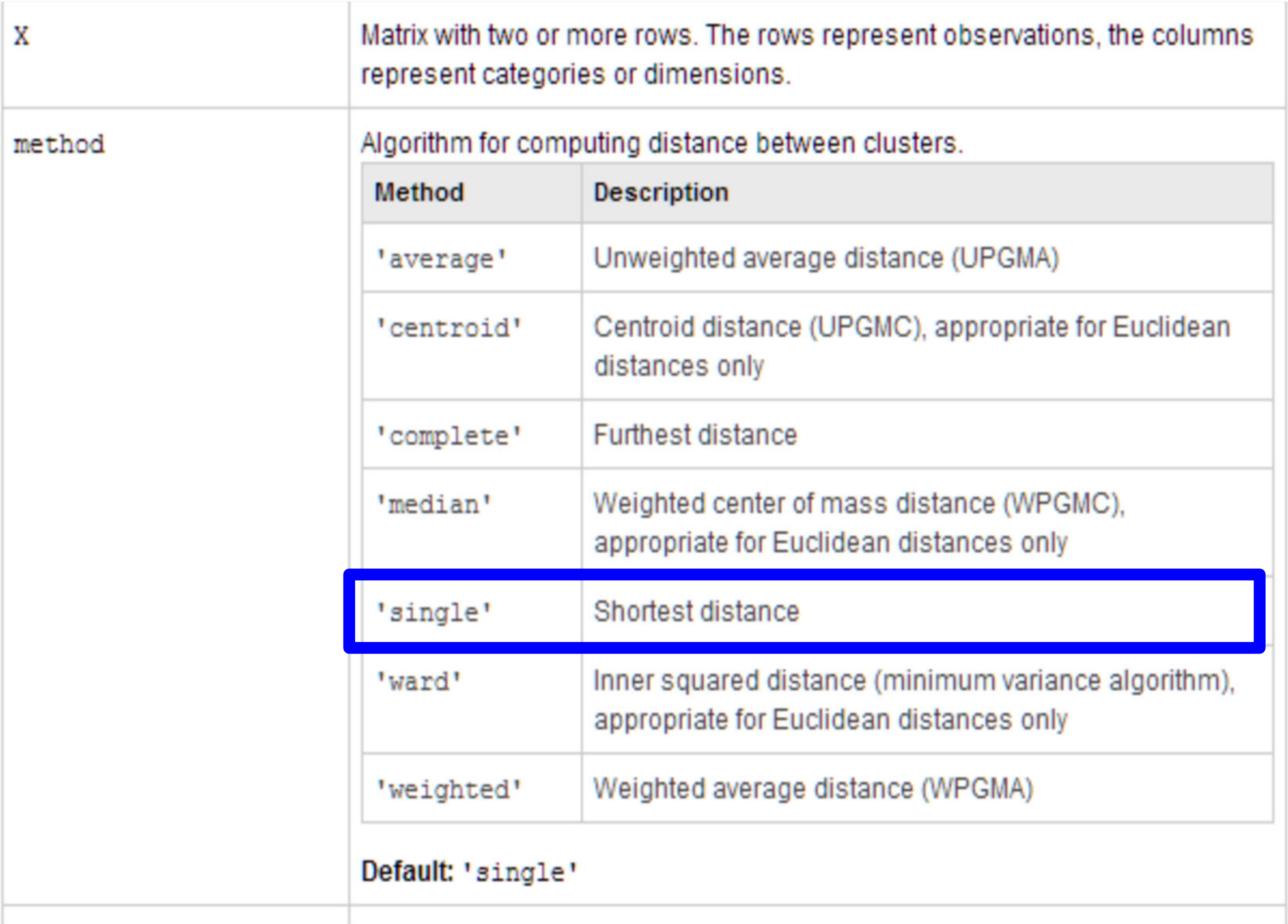
Choices of distance metrics in clustergram(… 'RowPDistValue' …, 'ColumnPDistValue' …,)



Metric	Description
`'euclidean'`	Euclidean distance (default).
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between rows in X is scaled by dividing by the corresponding element of the standard deviation `S=nanstd(X)`. To specify another value for `S`, use `D=pdist(X,'seuclidean',S)`.
`'cityblock'`	City block metric.
`'minkowski'`	Minkowski distance. The default exponent is 2. To specify a different exponent, use `D = pdist(X,'minkowski',P)`, where `P` is a scalar positive value of the exponent.
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'mahalanobis'`	Mahalanobis distance, using the sample covariance of `X` as computed by `nancov`. To compute the distance with a different covariance, use `D = pdist(X,'mahalanobis',C)`, where the matrix `C` is symmetric and positive definite.
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors).
`'correlation'`	One minus the sample correlation between points (treated as sequences of values).
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ.
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.

custom distance function	A distance function specified using @:
	`D = pdist(X,@distfun)`
	A distance function must be of form
	`d2 = distfun(XI,XJ)`
	taking as arguments a 1-by-*n* vector `XI`, corresponding to a single row of `X`, and an *m*2-by-*n* matrix `XJ`, corresponding to multiple rows of `X`. `distfun` must accept a matrix `XJ` with an arbitrary number of rows. `distfun` must return an *m*2-by-1 vector of distances `d2`, whose *k*th element is the distance between `XI` and `XJ(k,:)`.

Choices of hierarchical clustering algorithm in clustergram(…’linkage’,…)

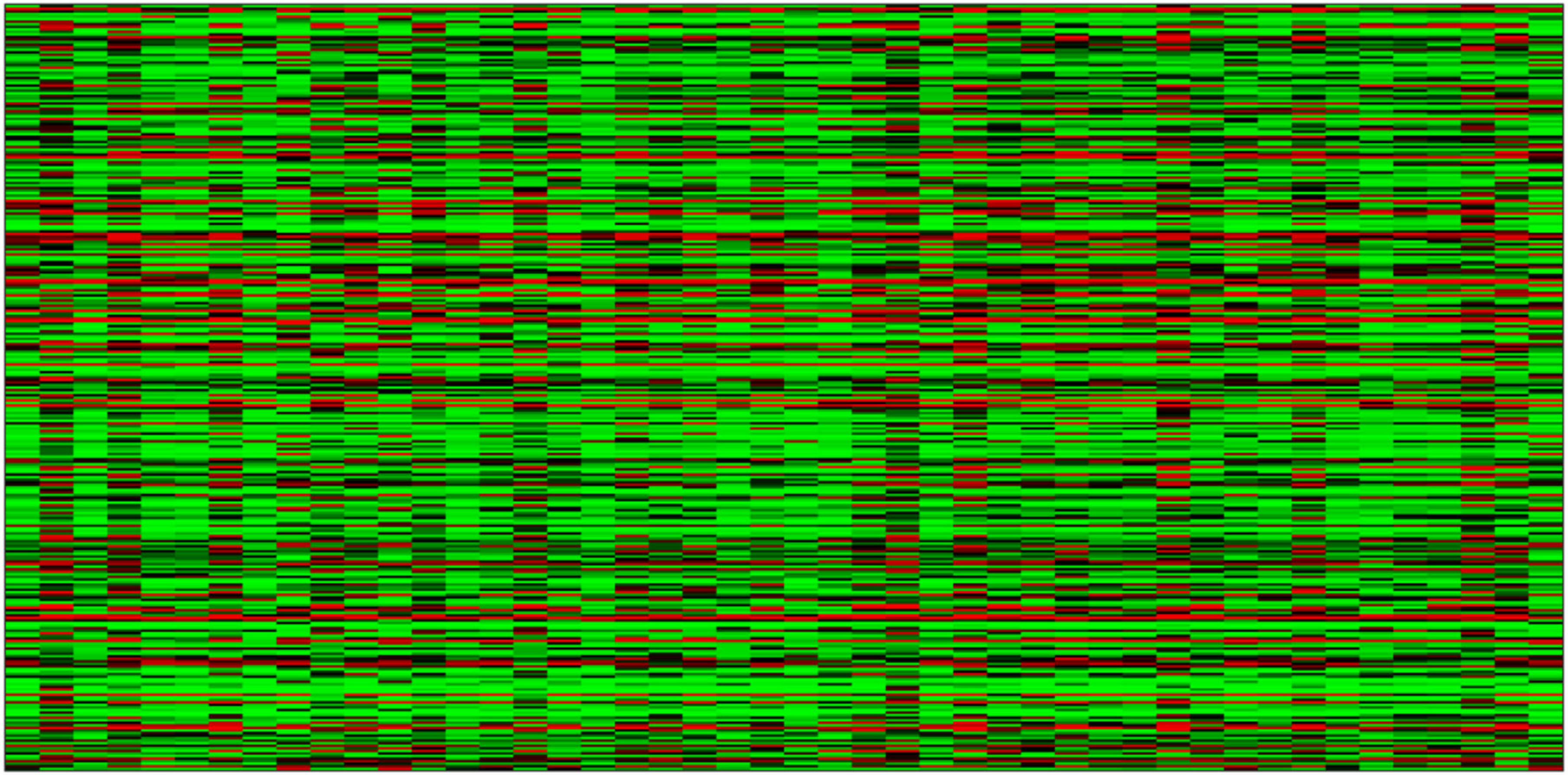


<table>
<tr><td><code>X</code></td><td>Matrix with two or more rows. The rows represent observations, the columns represent categories or dimensions.</td></tr>
<tr><td><code>method</code></td><td>Algorithm for computing distance between clusters.
<table>
<tr><th>Method</th><th>Description</th></tr>
<tr><td><code>'average'</code></td><td>Unweighted average distance (UPGMA)</td></tr>
<tr><td><code>'centroid'</code></td><td>Centroid distance (UPGMC), appropriate for Euclidean distances only</td></tr>
<tr><td><code>'complete'</code></td><td>Furthest distance</td></tr>
<tr><td><code>'median'</code></td><td>Weighted center of mass distance (WPGMC), appropriate for Euclidean distances only</td></tr>
<tr><td><code>'single'</code></td><td>Shortest distance</td></tr>
<tr><td><code>'ward'</code></td><td>Inner squared distance (minimum variance algorithm), appropriate for Euclidean distances only</td></tr>
<tr><td><code>'weighted'</code></td><td>Weighted average distance (WPGMA)</td></tr>
</table>
<b>Default:</b> <code>'single'</code></td></tr>
</table>

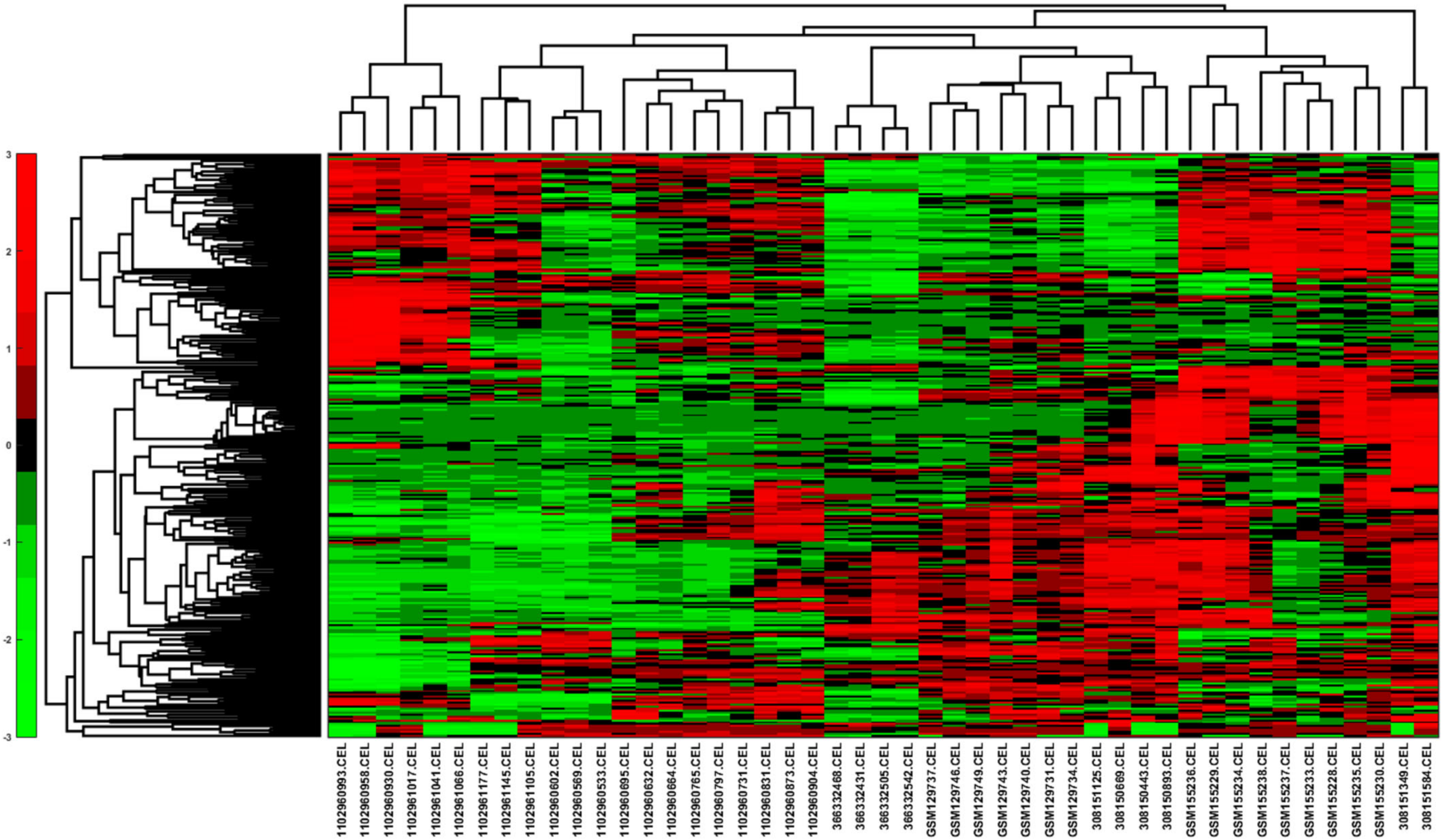
Clustering group exercise

- Each group will analyze a cluster of genes identified in the T cell expression table
- Analyze the table of top 100 genes **by variance** in 47 samples
- Cluster them using:
  - Group 1: UPGMA = ‘linkage’, ‘average’, ‘RowPDistValue’, ’euclidean’,
  - Group 2: ‘linkage’, ‘single’, ‘RowPDistValue’, ’cityblock’,
  - Group 3: ‘linkage’, ‘average’, ‘RowPDistValue’, ’correlation’,
  - Group 4: UPGMA = ‘linkage’, ‘single’, ‘RowPDistValue’, ’euclidean’,
  - Group 5: UPGMA = ‘linkage’, ‘weighted’, ‘RowPDistValue’, ’correlation’,
- Use clustergram(…, 'Standardize','Row',
  ‘linkage’, ***as specified for your group,***
  ‘RowPDistValue’ ***as specified for your group,***
  'RowLabels',gene_names1,'ColumnLabels‘, array_names)

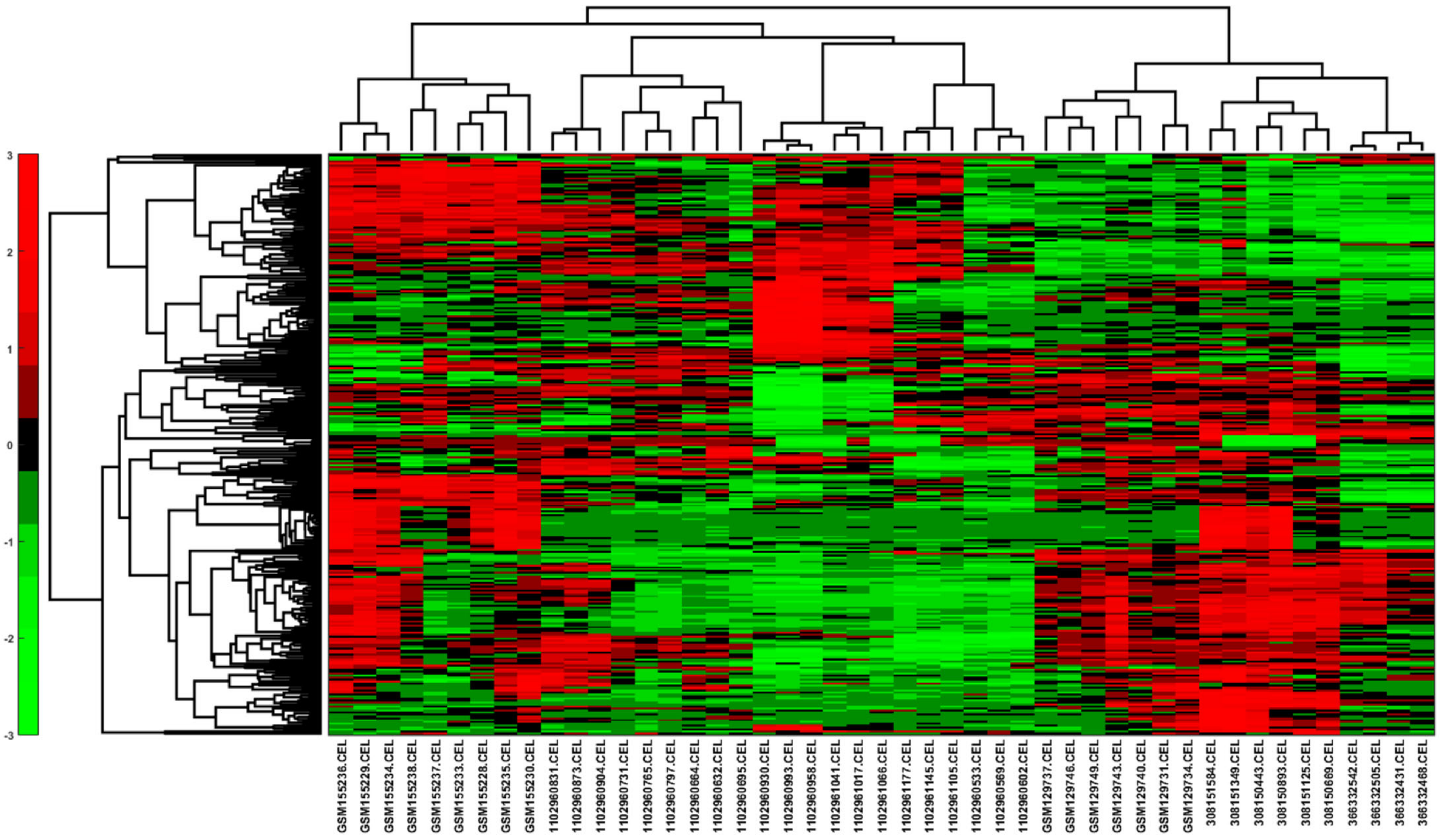
Before clustering



UPGMA hierarchical clustering, Euclidian distance



UPGMA hierarchical clustering, correlation distance



Clustering group exercise

- Each group will analyze a cluster of genes identified in the T cell expression table
- Analyze the table of top 100 genes **by variance** in 47 samples
- Cluster them using:
  - Group 1: UPGMA = ‘linkage’, ‘average’, ‘RowPDistValue’, ’euclidean’,
  - Group 2: ‘linkage’, ‘single’, ‘RowPDistValue’, ’cityblock’,
  - Group 3: ‘linkage’, ‘average’, ‘RowPDistValue’, ’correlation’,
  - Group 4: UPGMA = ‘linkage’, ‘single’, ‘RowPDistValue’, ’euclidean’,
  - Group 5: UPGMA = ‘linkage’, ‘weighted’, ‘RowPDistValue’, ’correlation’,
- Use clustergram(…, 'Standardize','Row', ‘linkage’, ***as specified for your group,*** ‘RowPDistValue’ ***as specified for your group,*** 'RowLabels',gene_names1,'ColumnLabels‘, array_names)

Cluster analysis group exercise

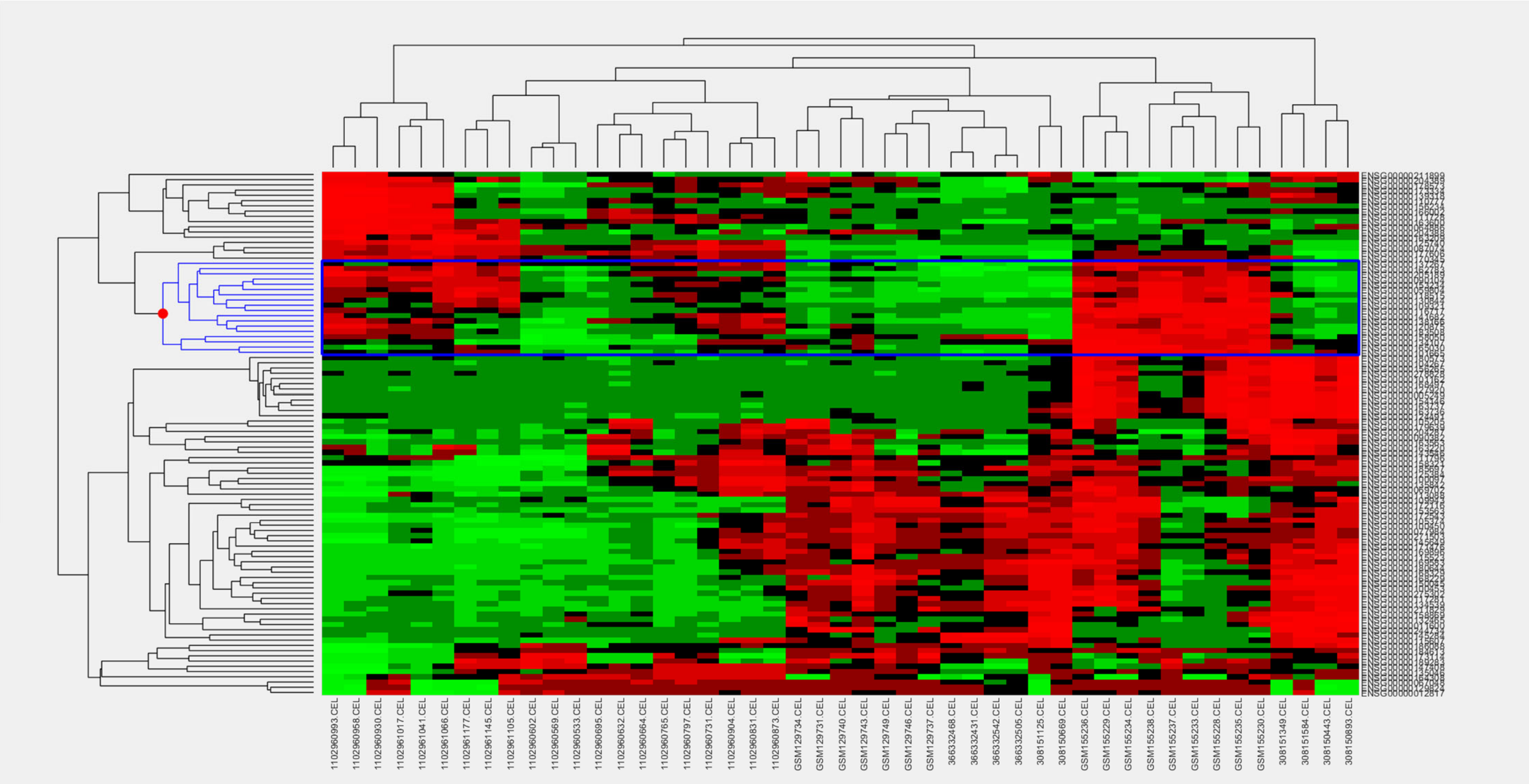
- Which biological functions are overrepresented in different clusters?
- Pick a cluster:
  - Select a node on the tree of rows,
  - Right click
  - Choose “export group info” into the workspace
  - Name it gene_list
- Run the following two Matlab commands to display genes
  - g1=gene_list.RowNodeNames;
  - for m=1:length(g1); disp(g1{m}); end;

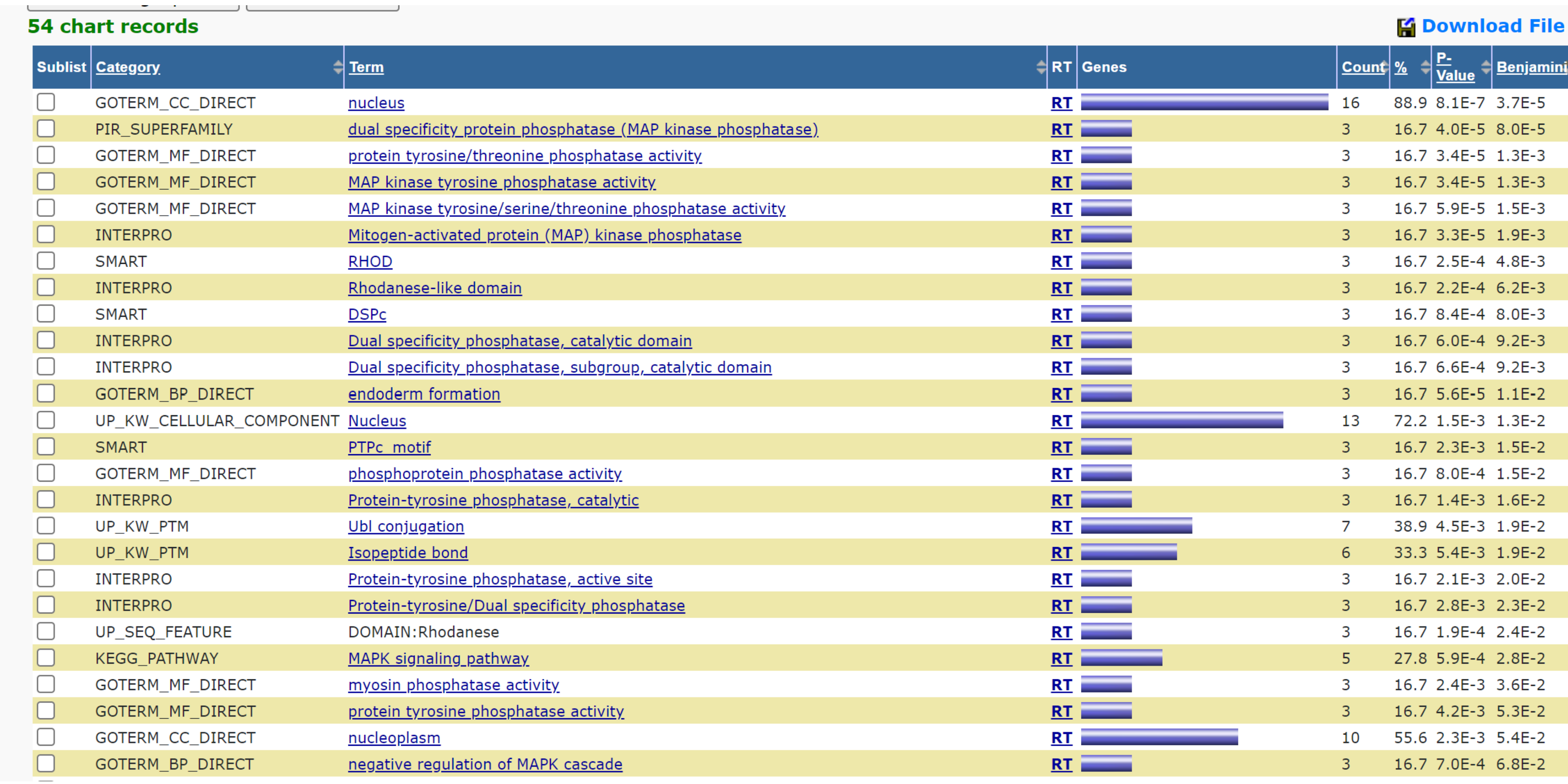
Search for shared biological functions

- copy the list of displayed genes
- go to "Start Analysis" on https://david.ncifcrf.gov/tools.jsp
- Paste genes from gene list displayed by Matlab into the box in the left panel of the website
- select ENSEMBL_GENE_ID and “gene list” radio button
- Click "Functional Annotation Clustering“
- Select groups in “Annotation Summary Results” which have many genes from your list. Definitely select “PUBMED_ID” and interaction databases like “Biogrid”
- First look at "Functional Annotation Chart" rectangular button below to display all overrepresented terms. Sort by “Benjamini” correction for multiple hypotheses tetsting
- Select "Functional Annotation Clustering" rectangular button below to display annotation results for gene list broken into multiple groups (clusters) each with related biological functions
- Write down the # of genes in the cluster and the top functions in two most interesting clusters

Using Group 1 options:

‘linkage’, ‘average’, ‘RowPDistValue’, ’euclidean’,





54 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	GOTERM_CC_DIRECT	nucleus	RT		16	88.9	8.1E-7	3.7E-5
☐	PIR_SUPERFAMILY	dual specificity protein phosphatase (MAP kinase phosphatase)	RT		3	16.7	4.0E-5	8.0E-5
☐	GOTERM_MF_DIRECT	protein tyrosine/threonine phosphatase activity	RT		3	16.7	3.4E-5	1.3E-3
☐	GOTERM_MF_DIRECT	MAP kinase tyrosine phosphatase activity	RT		3	16.7	3.4E-5	1.3E-3
☐	GOTERM_MF_DIRECT	MAP kinase tyrosine/serine/threonine phosphatase activity	RT		3	16.7	5.9E-5	1.5E-3
☐	INTERPRO	Mitogen-activated protein (MAP) kinase phosphatase	RT		3	16.7	3.3E-5	1.9E-3
☐	SMART	RHOD	RT		3	16.7	2.5E-4	4.8E-3
☐	INTERPRO	Rhodanese-like domain	RT		3	16.7	2.2E-4	6.2E-3
☐	SMART	DSPc	RT		3	16.7	8.4E-4	8.0E-3
☐	INTERPRO	Dual specificity phosphatase, catalytic domain	RT		3	16.7	6.0E-4	9.2E-3
☐	INTERPRO	Dual specificity phosphatase, subgroup, catalytic domain	RT		3	16.7	6.6E-4	9.2E-3
☐	GOTERM_BP_DIRECT	endoderm formation	RT		3	16.7	5.6E-5	1.1E-2
☐	UP_KW_CELLULAR_COMPONENT	Nucleus	RT		13	72.2	1.5E-3	1.3E-2
☐	SMART	PTPc_motif	RT		3	16.7	2.3E-3	1.5E-2
☐	GOTERM_MF_DIRECT	phosphoprotein phosphatase activity	RT		3	16.7	8.0E-4	1.5E-2
☐	INTERPRO	Protein-tyrosine phosphatase, catalytic	RT		3	16.7	1.4E-3	1.6E-2
☐	UP_KW_PTM	Ubl conjugation	RT		7	38.9	4.5E-3	1.9E-2
☐	UP_KW_PTM	Isopeptide bond	RT		6	33.3	5.4E-3	1.9E-2
☐	INTERPRO	Protein-tyrosine phosphatase, active site	RT		3	16.7	2.1E-3	2.0E-2
☐	INTERPRO	Protein-tyrosine/Dual specificity phosphatase	RT		3	16.7	2.8E-3	2.3E-2
☐	UP_SEQ_FEATURE	DOMAIN:Rhodanese	RT		3	16.7	1.9E-4	2.4E-2
☐	KEGG_PATHWAY	MAPK signaling pathway	RT		5	27.8	5.9E-4	2.8E-2
☐	GOTERM_MF_DIRECT	myosin phosphatase activity	RT		3	16.7	2.4E-3	3.6E-2
☐	GOTERM_MF_DIRECT	protein tyrosine phosphatase activity	RT		3	16.7	4.2E-3	5.3E-2
☐	GOTERM_CC_DIRECT	nucleoplasm	RT		10	55.6	2.3E-3	5.4E-2
☐	GOTERM_BP_DIRECT	negative regulation of MAPK cascade	RT		3	16.7	7.0E-4	6.8E-2

Gene list being analyzed

Clustering options and stringency

score for the group based on the EASE scores of each term members. The higher, the more enriched.

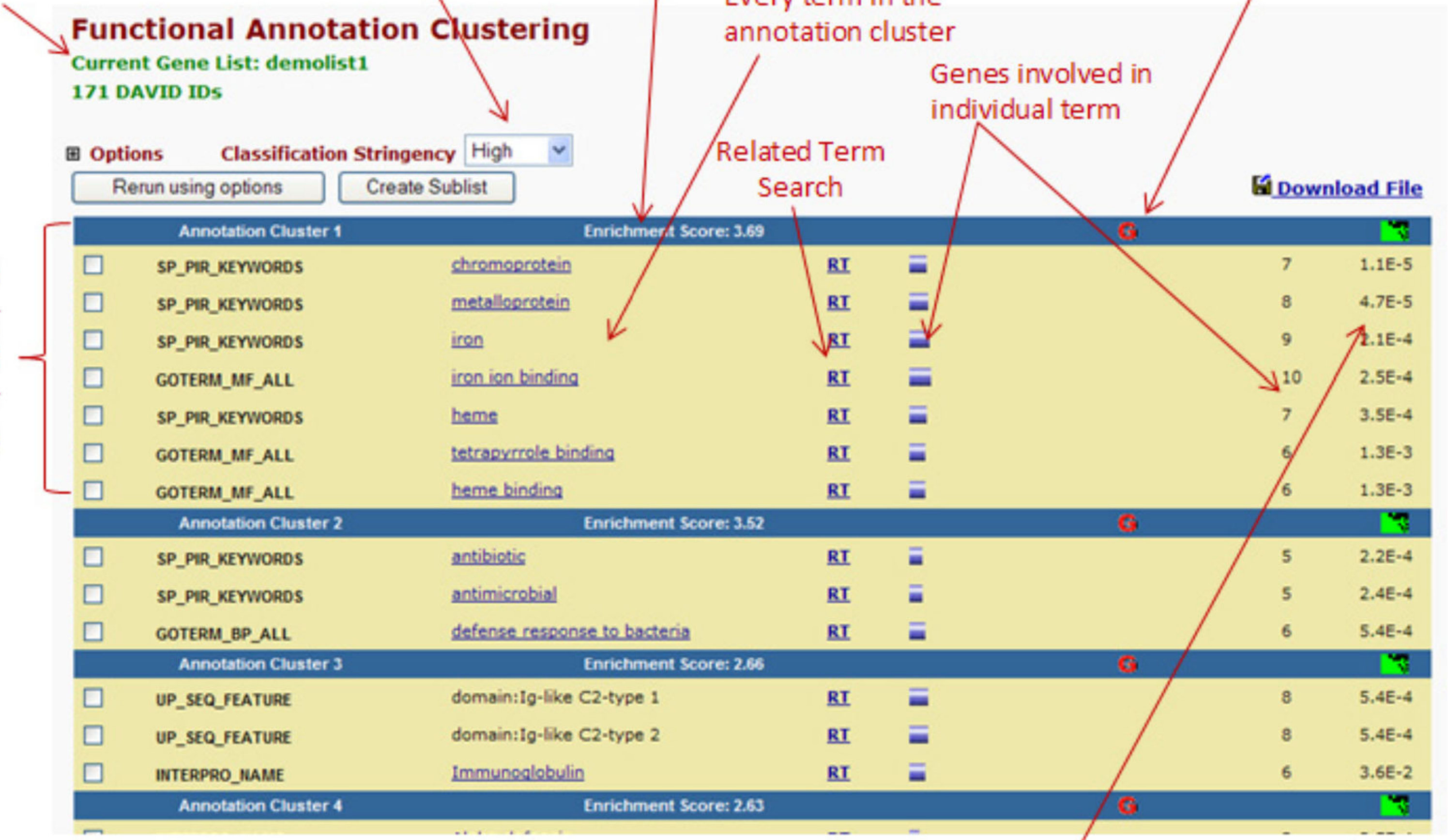
Every term in the annotation cluster

ALL genes involved in this annotation cluster

Genes involved in individual term

Related Term Search

A group of terms having similar biological meaning due to sharing similar gene members



Functional Annotation Clustering

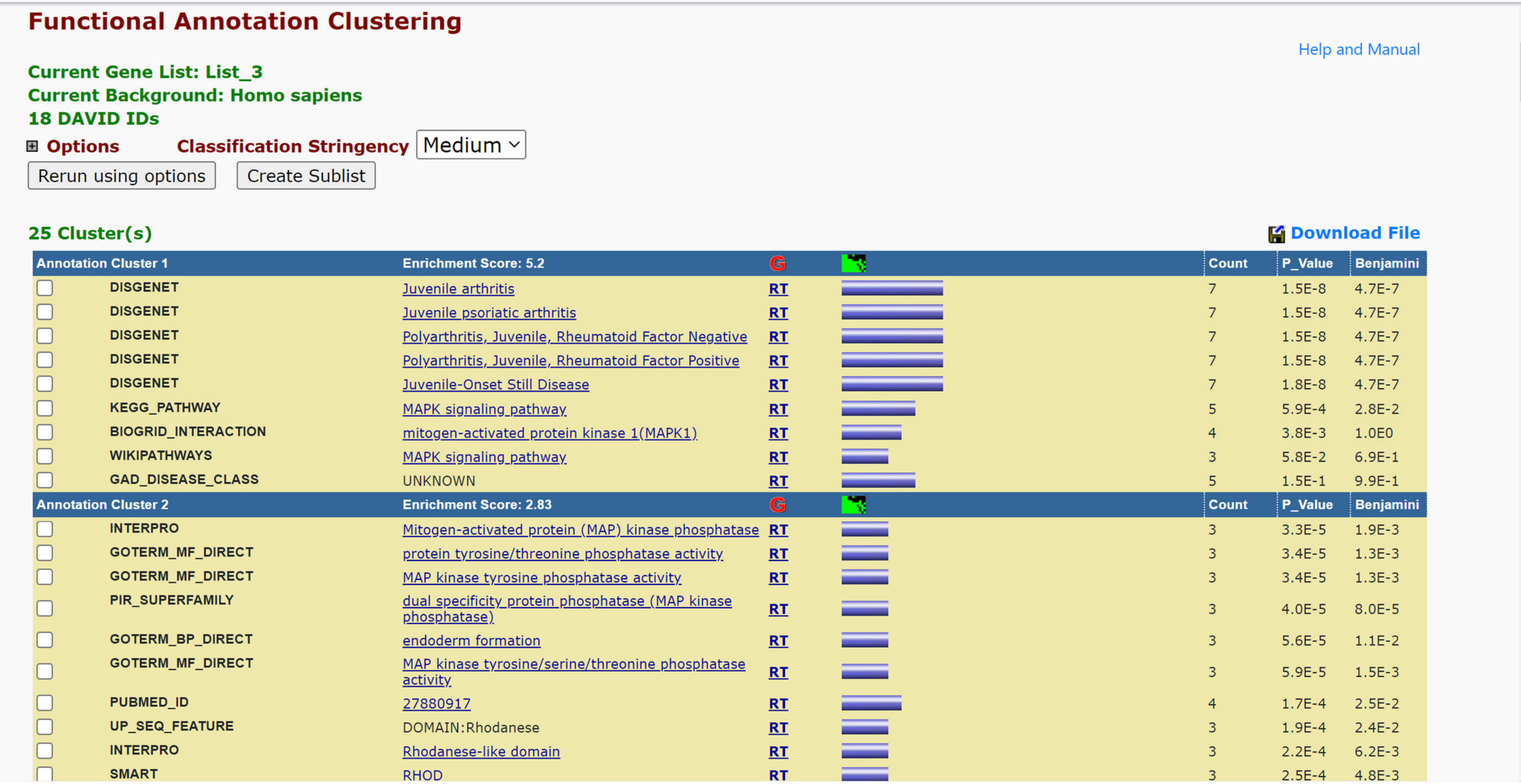
Current Gene List: demolist1

171 DAVID IDs

⊞ **Options** **Classification Stringency** High

Annotation Cluster 1		**Enrichment Score: 3.69**			●	🟩
□	SP_PIR_KEYWORDS	chromoprotein	RT	▤	7	1.1E-5
□	SP_PIR_KEYWORDS	metalloprotein	RT	▤	8	4.7E-5
□	SP_PIR_KEYWORDS	iron	RT	▤	9	2.1E-4
□	GOTERM_MF_ALL	iron ion binding	RT	▤	10	2.5E-4
□	SP_PIR_KEYWORDS	heme	RT	▤	7	3.5E-4
□	GOTERM_MF_ALL	tetrapyrrole binding	RT	▤	6	1.3E-3
□	GOTERM_MF_ALL	heme binding	RT	▤	6	1.3E-3
Annotation Cluster 2		**Enrichment Score: 3.52**			●	🟩
□	SP_PIR_KEYWORDS	antibiotic	RT	▤	5	2.2E-4
□	SP_PIR_KEYWORDS	antimicrobial	RT	▤	5	2.4E-4
□	GOTERM_BP_ALL	defense response to bacteria	RT	▤	6	5.4E-4
Annotation Cluster 3		**Enrichment Score: 2.66**			●	🟩
□	UP_SEQ_FEATURE	domain:Ig-like C2-type 1	RT	▤	8	5.4E-4
□	UP_SEQ_FEATURE	domain:Ig-like C2-type 2	RT	▤	8	5.4E-4
□	INTERPRO_NAME	Immunoglobulin	RT	▤	6	3.6E-2
Annotation Cluster 4		**Enrichment Score: 2.63**			●	🟩

EASE Score, the modified Fisher Exact P-Value. They are identical to that in the Chart Report. The smaller, the more enriched.



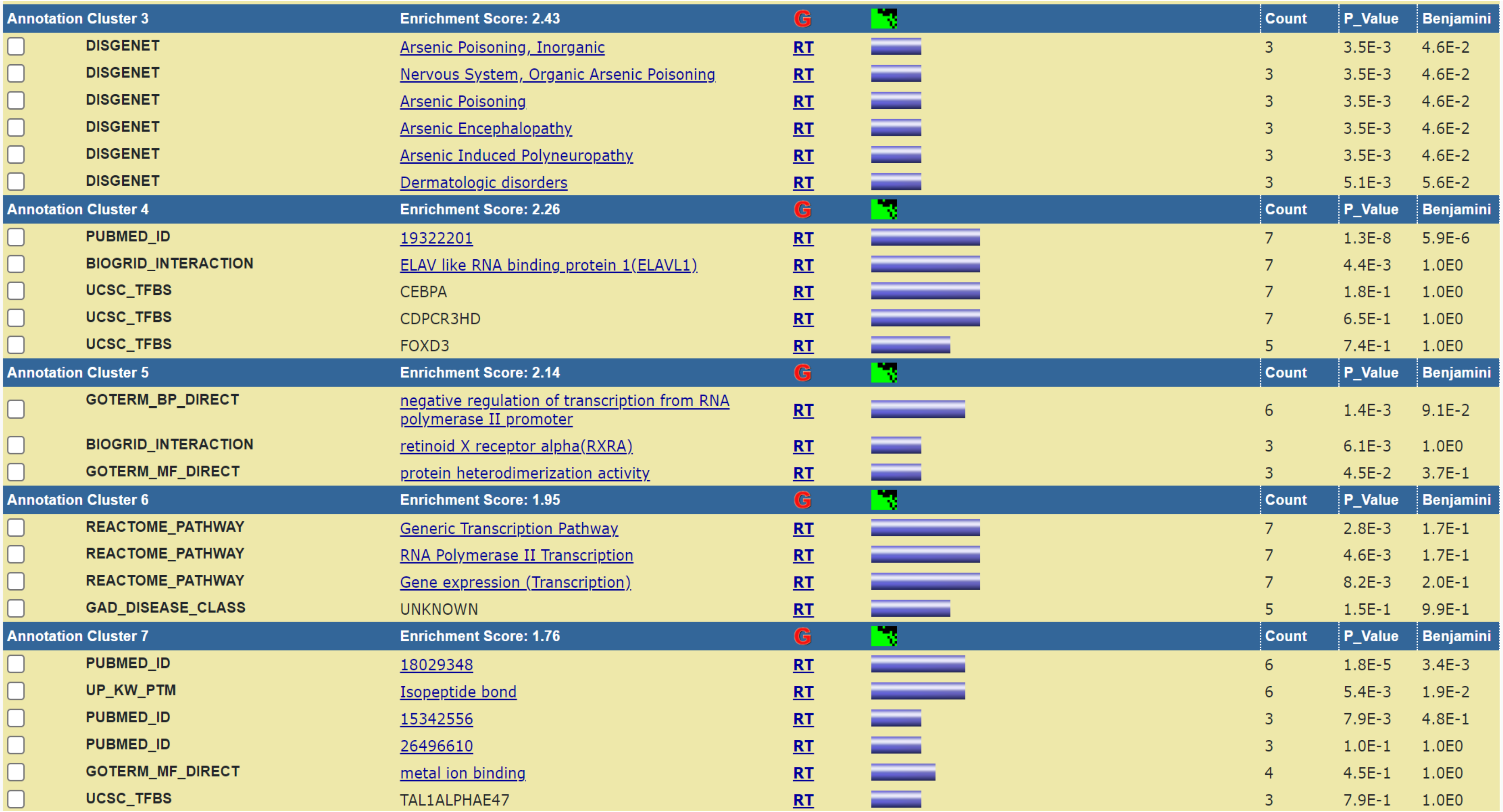
Functional Annotation Clustering

Help and Manual

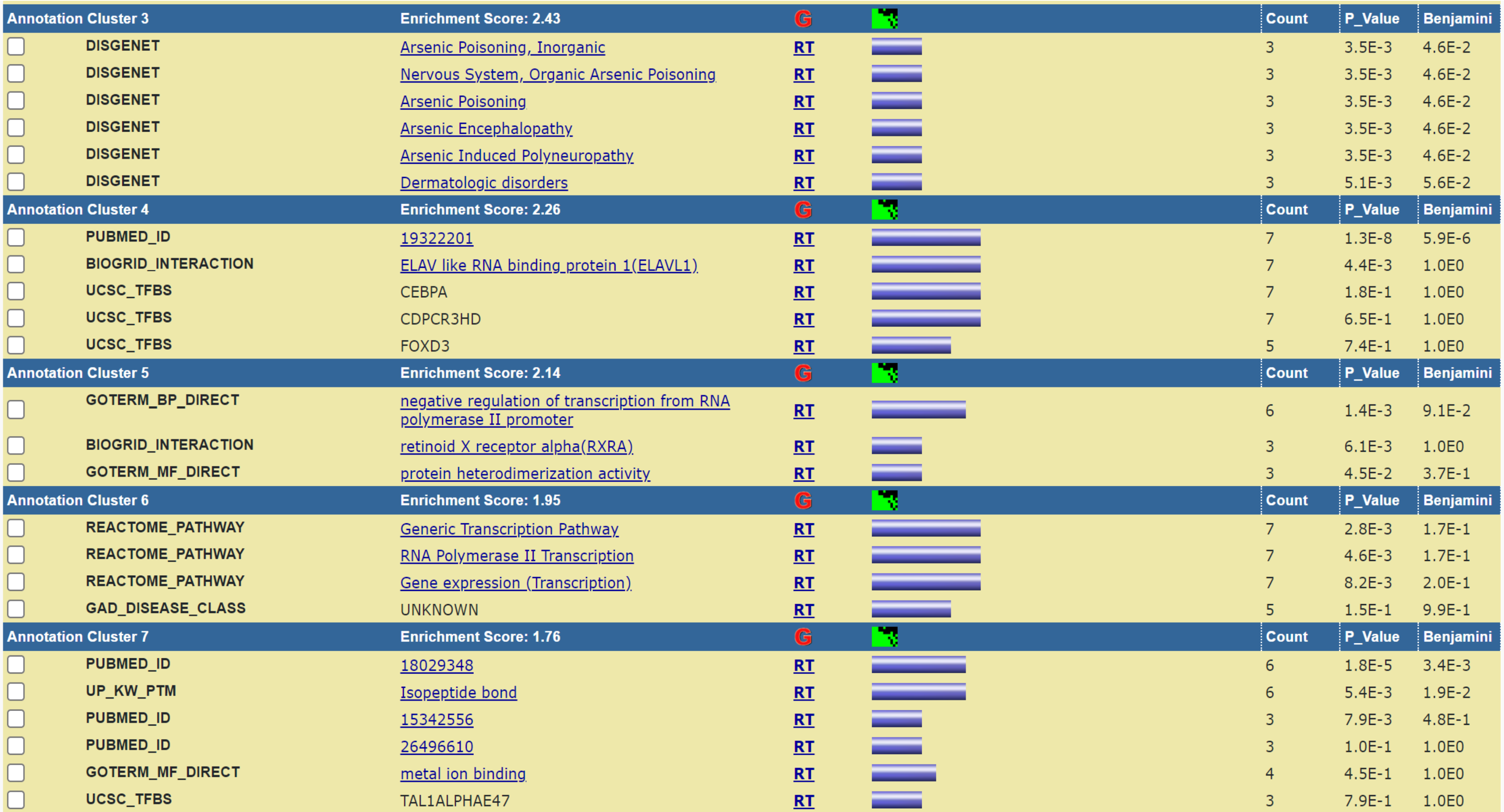
Current Gene List: List_3
Current Background: Homo sapiens
18 DAVID IDs

25 Cluster(s)

Annotation Cluster 1		Enrichment Score: 5.2	G	Count	P_Value	Benjamini
☐	DISGENET	Juvenile arthritis	RT	7	1.5E-8	4.7E-7
☐	DISGENET	Juvenile psoriatic arthritis	RT	7	1.5E-8	4.7E-7
☐	DISGENET	Polyarthritis, Juvenile, Rheumatoid Factor Negative	RT	7	1.5E-8	4.7E-7
☐	DISGENET	Polyarthritis, Juvenile, Rheumatoid Factor Positive	RT	7	1.5E-8	4.7E-7
☐	DISGENET	Juvenile-Onset Still Disease	RT	7	1.8E-8	4.7E-7
☐	KEGG_PATHWAY	MAPK signaling pathway	RT	5	5.9E-4	2.8E-2
☐	BIOGRID_INTERACTION	mitogen-activated protein kinase 1(MAPK1)	RT	4	3.8E-3	1.0E0
☐	WIKIPATHWAYS	MAPK signaling pathway	RT	3	5.8E-2	6.9E-1
☐	GAD_DISEASE_CLASS	UNKNOWN	RT	5	1.5E-1	9.9E-1
Annotation Cluster 2		Enrichment Score: 2.83	G	Count	P_Value	Benjamini
☐	INTERPRO	Mitogen-activated protein (MAP) kinase phosphatase	RT	3	3.3E-5	1.9E-3
☐	GOTERM_MF_DIRECT	protein tyrosine/threonine phosphatase activity	RT	3	3.4E-5	1.3E-3
☐	GOTERM_MF_DIRECT	MAP kinase tyrosine phosphatase activity	RT	3	3.4E-5	1.3E-3
☐	PIR_SUPERFAMILY	dual specificity protein phosphatase (MAP kinase phosphatase)	RT	3	4.0E-5	8.0E-5
☐	GOTERM_BP_DIRECT	endoderm formation	RT	3	5.6E-5	1.1E-2
☐	GOTERM_MF_DIRECT	MAP kinase tyrosine/serine/threonine phosphatase activity	RT	3	5.9E-5	1.5E-3
☐	PUBMED_ID	27880917	RT	4	1.7E-4	2.5E-2
☐	UP_SEQ_FEATURE	DOMAIN:Rhodanese	RT	3	1.9E-4	2.4E-2
☐	INTERPRO	Rhodanese-like domain	RT	3	2.2E-4	6.2E-3
☐	SMART	RHOD	RT	3	2.5E-4	4.8E-3

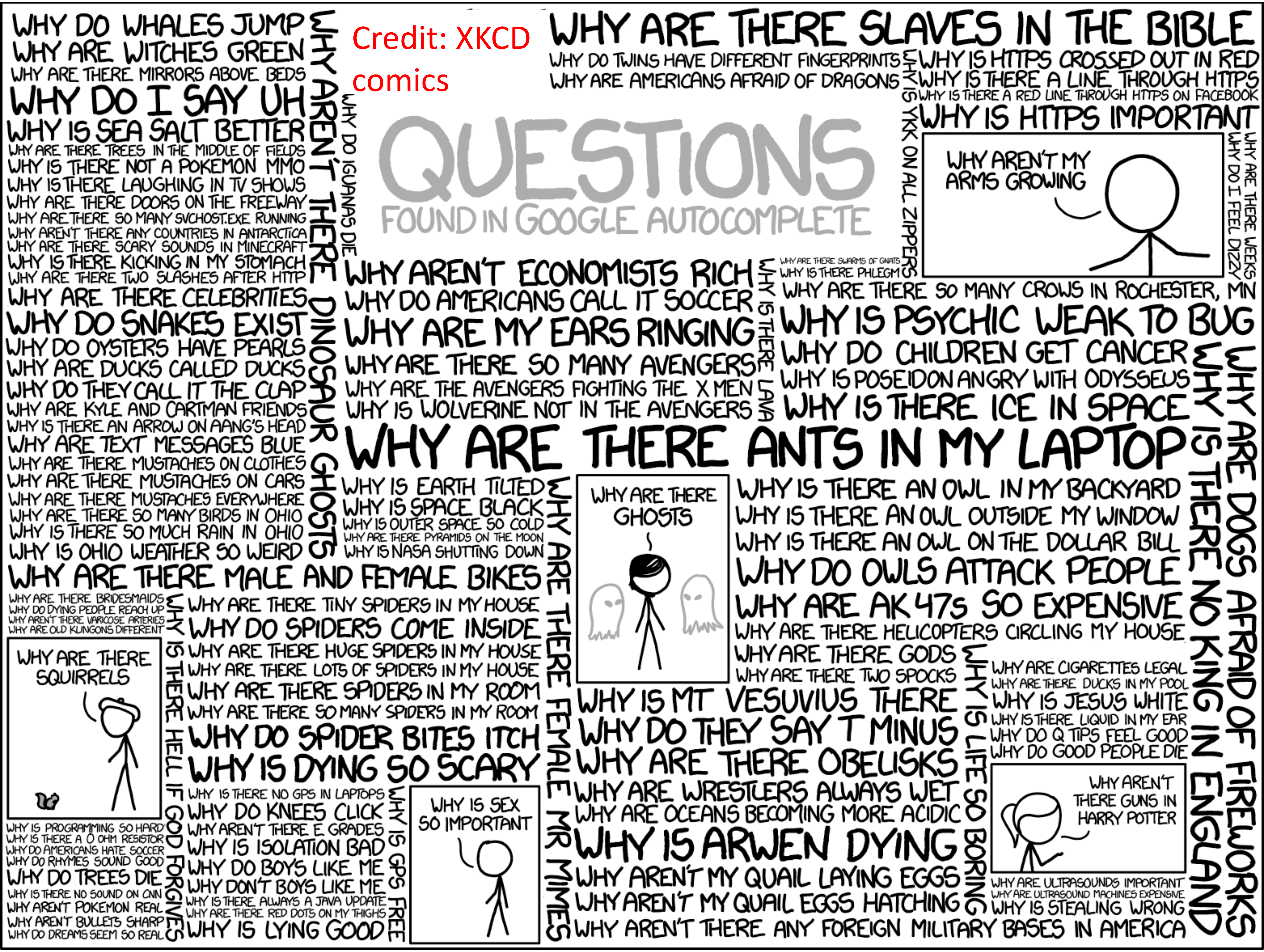


Annotation Cluster 3		Enrichment Score: 2.43	G		Count	P_Value	Benjamini
☐	DISGENET	Arsenic Poisoning, Inorganic	RT		3	3.5E-3	4.6E-2
☐	DISGENET	Nervous System, Organic Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
☐	DISGENET	Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
☐	DISGENET	Arsenic Encephalopathy	RT		3	3.5E-3	4.6E-2
☐	DISGENET	Arsenic Induced Polyneuropathy	RT		3	3.5E-3	4.6E-2
☐	DISGENET	Dermatologic disorders	RT		3	5.1E-3	5.6E-2
Annotation Cluster 4		Enrichment Score: 2.26	G		Count	P_Value	Benjamini
☐	PUBMED_ID	19322201	RT		7	1.3E-8	5.9E-6
☐	BIOGRID_INTERACTION	ELAV like RNA binding protein 1(ELAVL1)	RT		7	4.4E-3	1.0E0
☐	UCSC_TFBS	CEBPA	RT		7	1.8E-1	1.0E0
☐	UCSC_TFBS	CDPCR3HD	RT		7	6.5E-1	1.0E0
☐	UCSC_TFBS	FOXD3	RT		5	7.4E-1	1.0E0
Annotation Cluster 5		Enrichment Score: 2.14	G		Count	P_Value	Benjamini
☐	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		6	1.4E-3	9.1E-2
☐	BIOGRID_INTERACTION	retinoid X receptor alpha(RXRA)	RT		3	6.1E-3	1.0E0
☐	GOTERM_MF_DIRECT	protein heterodimerization activity	RT		3	4.5E-2	3.7E-1
Annotation Cluster 6		Enrichment Score: 1.95	G		Count	P_Value	Benjamini
☐	REACTOME_PATHWAY	Generic Transcription Pathway	RT		7	2.8E-3	1.7E-1
☐	REACTOME_PATHWAY	RNA Polymerase II Transcription	RT		7	4.6E-3	1.7E-1
☐	REACTOME_PATHWAY	Gene expression (Transcription)	RT		7	8.2E-3	2.0E-1
☐	GAD_DISEASE_CLASS	UNKNOWN	RT		5	1.5E-1	9.9E-1
Annotation Cluster 7		Enrichment Score: 1.76	G		Count	P_Value	Benjamini
☐	PUBMED_ID	18029348	RT		6	1.8E-5	3.4E-3
☐	UP_KW_PTM	Isopeptide bond	RT		6	5.4E-3	1.9E-2
☐	PUBMED_ID	15342556	RT		3	7.9E-3	4.8E-1
☐	PUBMED_ID	26496610	RT		3	1.0E-1	1.0E0
☐	GOTERM_MF_DIRECT	metal ion binding	RT		4	4.5E-1	1.0E0
☐	UCSC_TFBS	TAL1ALPHAE47	RT		3	7.9E-1	1.0E0



Annotation Cluster 3		**Enrichment Score: 2.43**	G		**Count**	**P_Value**	**Benjamini**
□	DISGENET	Arsenic Poisoning, Inorganic	RT		3	3.5E-3	4.6E-2
□	DISGENET	Nervous System, Organic Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
□	DISGENET	Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
□	DISGENET	Arsenic Encephalopathy	RT		3	3.5E-3	4.6E-2
□	DISGENET	Arsenic Induced Polyneuropathy	RT		3	3.5E-3	4.6E-2
□	DISGENET	Dermatologic disorders	RT		3	5.1E-3	5.6E-2
Annotation Cluster 4		**Enrichment Score: 2.26**	G		**Count**	**P_Value**	**Benjamini**
□	PUBMED_ID	19322201	RT		7	1.3E-8	5.9E-6
□	BIOGRID_INTERACTION	ELAV like RNA binding protein 1(ELAVL1)	RT		7	4.4E-3	1.0E0
□	UCSC_TFBS	CEBPA	RT		7	1.8E-1	1.0E0
□	UCSC_TFBS	CDPCR3HD	RT		7	6.5E-1	1.0E0
□	UCSC_TFBS	FOXD3	RT		5	7.4E-1	1.0E0
Annotation Cluster 5		**Enrichment Score: 2.14**	G		**Count**	**P_Value**	**Benjamini**
□	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		6	1.4E-3	9.1E-2
□	BIOGRID_INTERACTION	retinoid X receptor alpha(RXRA)	RT		3	6.1E-3	1.0E0
□	GOTERM_MF_DIRECT	protein heterodimerization activity	RT		3	4.5E-2	3.7E-1
Annotation Cluster 6		**Enrichment Score: 1.95**	G		**Count**	**P_Value**	**Benjamini**
□	REACTOME_PATHWAY	Generic Transcription Pathway	RT		7	2.8E-3	1.7E-1
□	REACTOME_PATHWAY	RNA Polymerase II Transcription	RT		7	4.6E-3	1.7E-1
□	REACTOME_PATHWAY	Gene expression (Transcription)	RT		7	8.2E-3	2.0E-1
□	GAD_DISEASE_CLASS	UNKNOWN	RT		5	1.5E-1	9.9E-1
Annotation Cluster 7		**Enrichment Score: 1.76**	G		**Count**	**P_Value**	**Benjamini**
□	PUBMED_ID	18029348	RT		6	1.8E-5	3.4E-3
□	UP_KW_PTM	Isopeptide bond	RT		6	5.4E-3	1.9E-2
□	PUBMED_ID	15342556	RT		3	7.9E-3	4.8E-1
□	PUBMED_ID	26496610	RT		3	1.0E-1	1.0E0
□	GOTERM_MF_DIRECT	metal ion binding	RT		4	4.5E-1	1.0E0
□	UCSC_TFBS	TAL1ALPHAE47	RT		3	7.9E-1	1.0E0

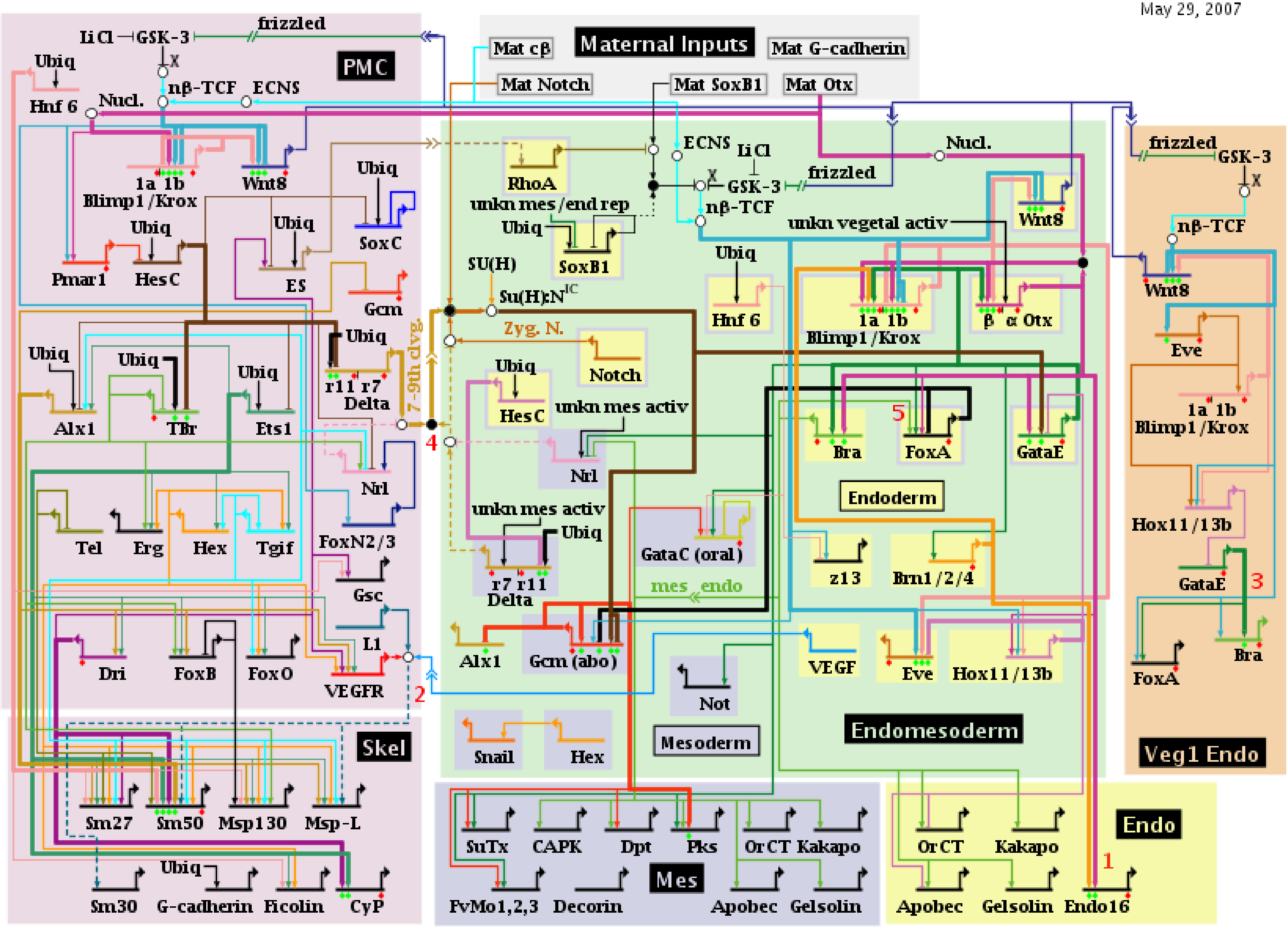
Credit: XKCD comics


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Reminder from the first lecture

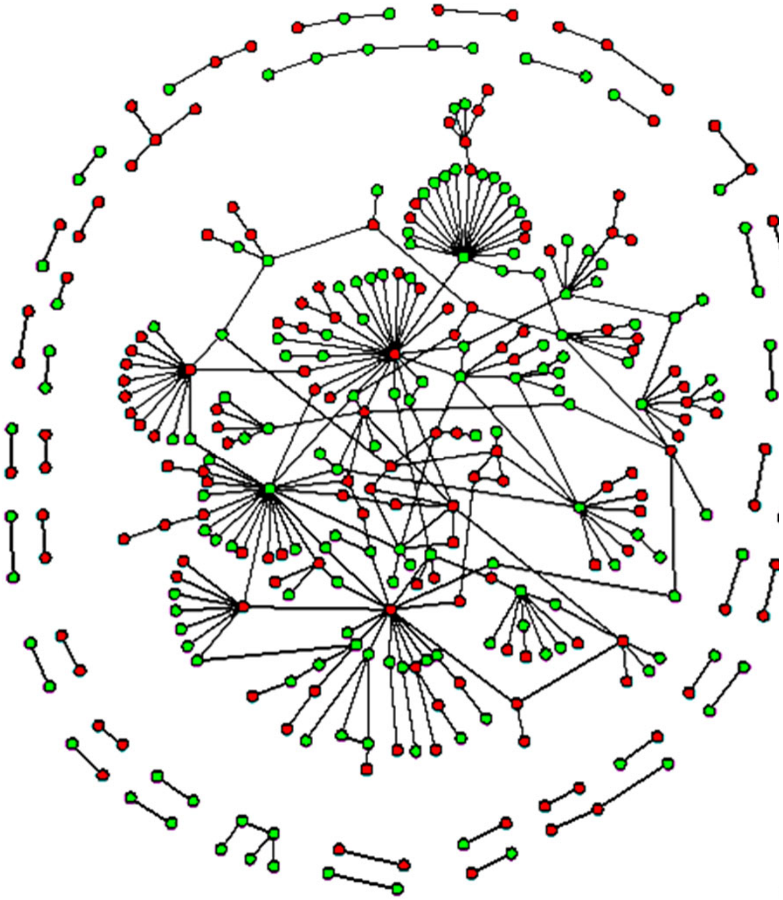
Sea urchin embryonic development (from endomesoderm up to 30 hours) by Davidson’s lab

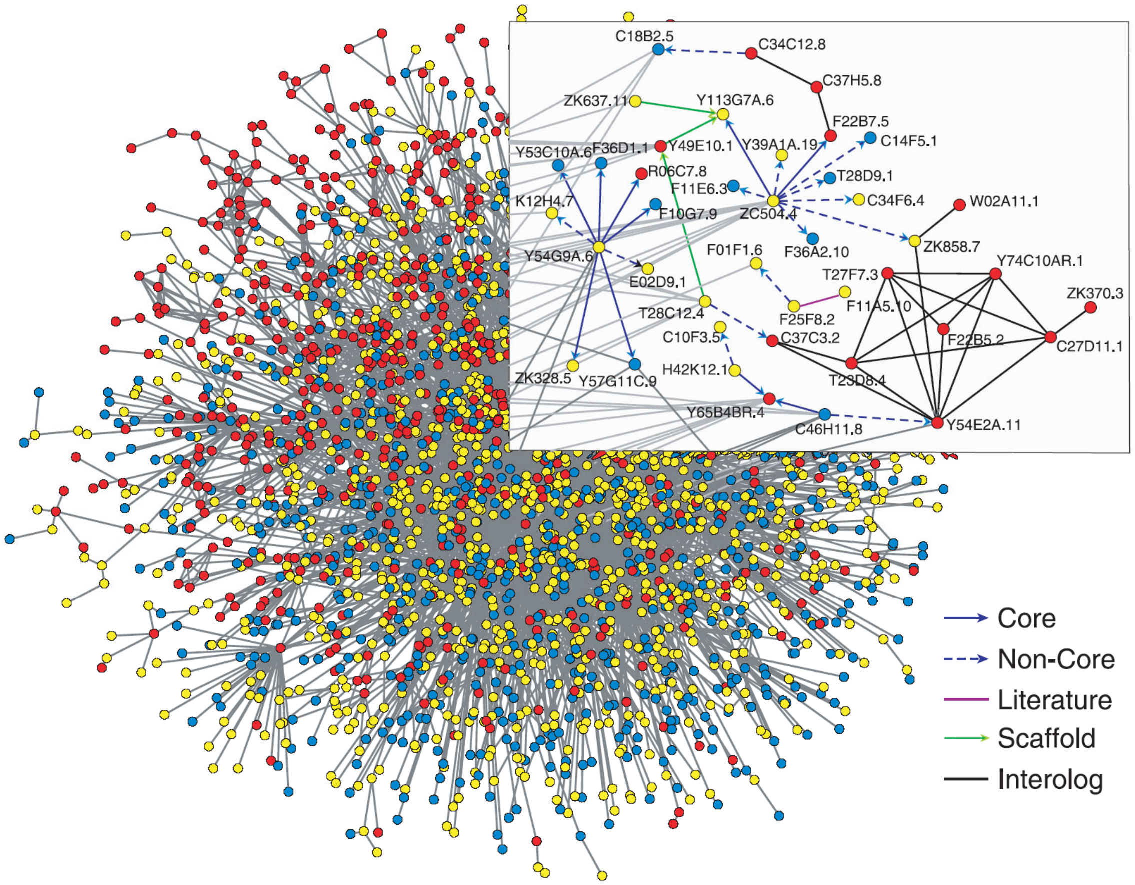
May 29, 2007



Ubiq=ubiquitous; Mat = maternal; activ = activator; rep = repressor;
unkn = unknown; Nucl. = nuclearization; χ = β-catenin source;
nβ-TCF = nuclearized b-β-catenin-Tcf1; ES = early signal;
ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

Protein-Protein binding
IntAct Database (Dec 2015)
Interactions: 577,297   Proteins: 89,716

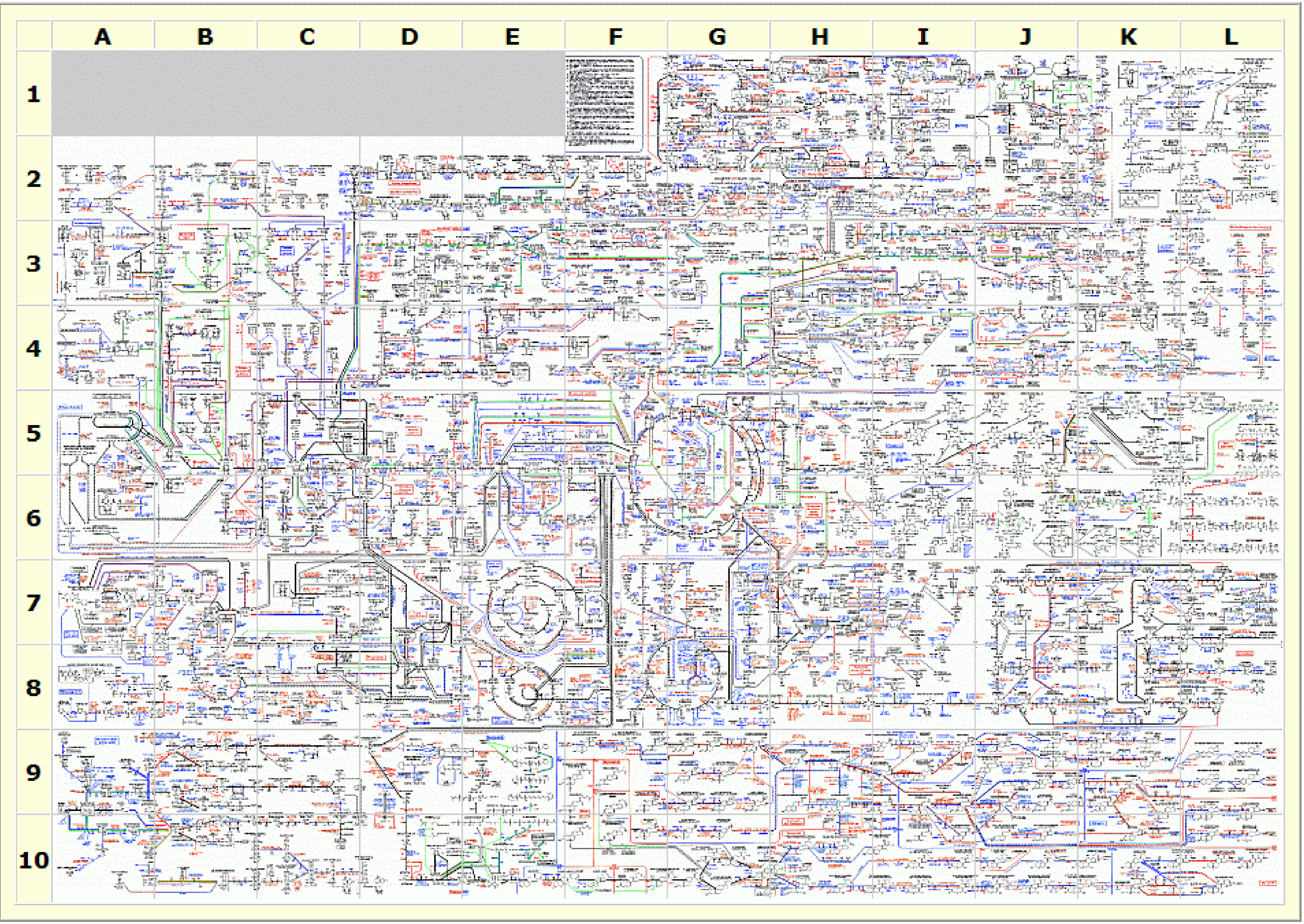




Baker's yeast *S. cerevisiae* (only nuclear proteins shown)
From S. Maslov, K. Sneppen, Science 2002

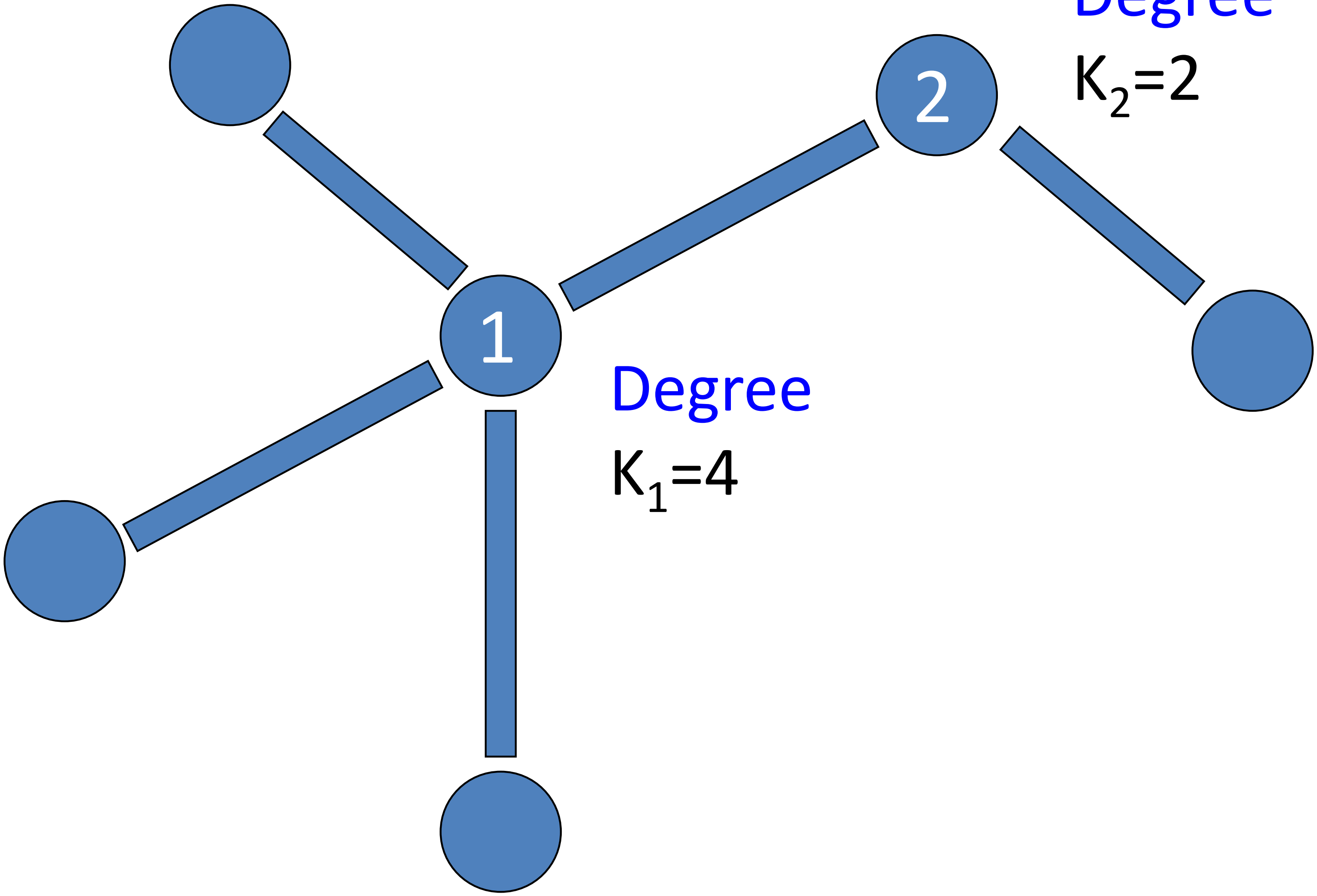
Worm *C. elegans*
From S. Lee et al , Science 2004

Metabolic pathway chart by ExPASy: **5702 reactions as of December 2015**



Basic concepts of network analysis

Degree of a node – its # of neighbors



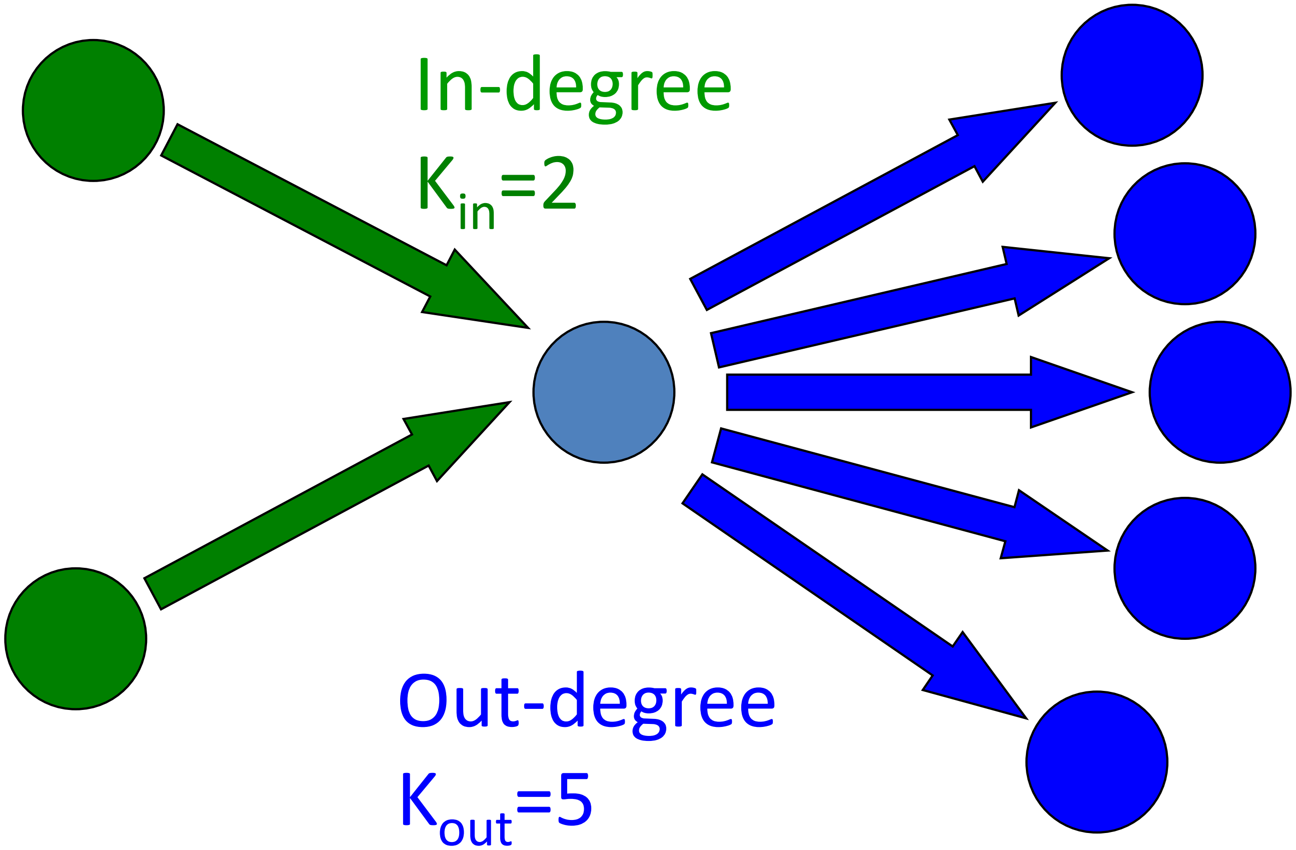
Degree

$K_2=2$

Degree

$K_1=4$

Directed networks have in- and out- degrees



In-degree

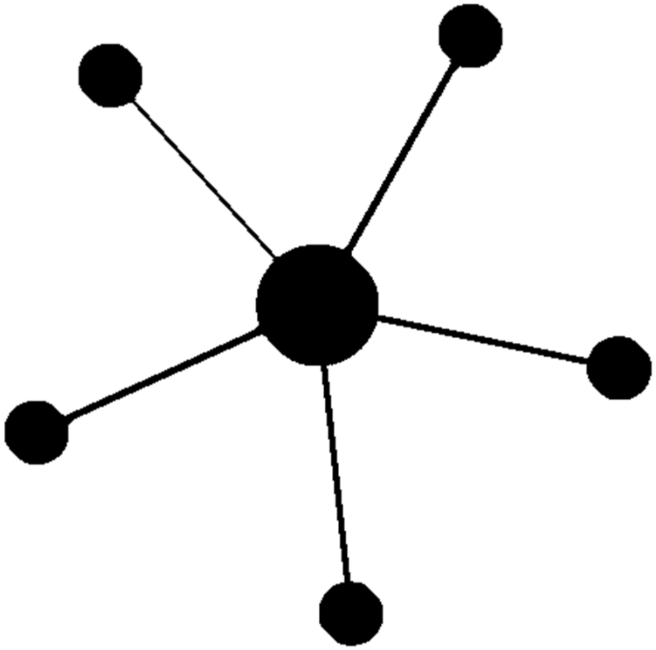
$K_{in}=2$

Out-degree

$K_{out}=5$

How to find “important” nodes?

- By their degree
- Hubs = important
- Example: Google’s PageRank



How Google PageRank algorithm works?

- Google was solving the following problem in mid-1990s: too many websites match a typical search query: need to rank websites.
- Other popular search engines (e.g. Altavista) count the # of times a query word appears in website's text. Websites respond by putting lots of invisible words
- One could rank the importance of webpages by number of hyperlinks pointing to it (in-degree K_{in}) but:
  - Too democratic: It doesn't take into account the importance of webpages sending hyperlinks
  - it's easy to trick and artificially boost the rank
- Google's solution: simulate the behavior of many "random surfers" and then count the number of times they visited each webpage = it's PageRank
  - Popular pages send more surfers your way → the PageRank weight is proportional to K_{in} but weighted by popularity

PageRank algorithm is Google's $2.8T idea

- PageRank assigns to every webpage an importance score G_i
- The meaning of G_i – how often random surfers visit this website
- To determine solves a self-consistent Eq.:

  $G_i \sim \sum_j T_{ij} G_{j.}$ Here $T_{ij}= A_{ij}/K_{out}$ (j) is the normalized adjacency matrix
- It finds the principal eigenvector (the one with the largest eigenvalue).

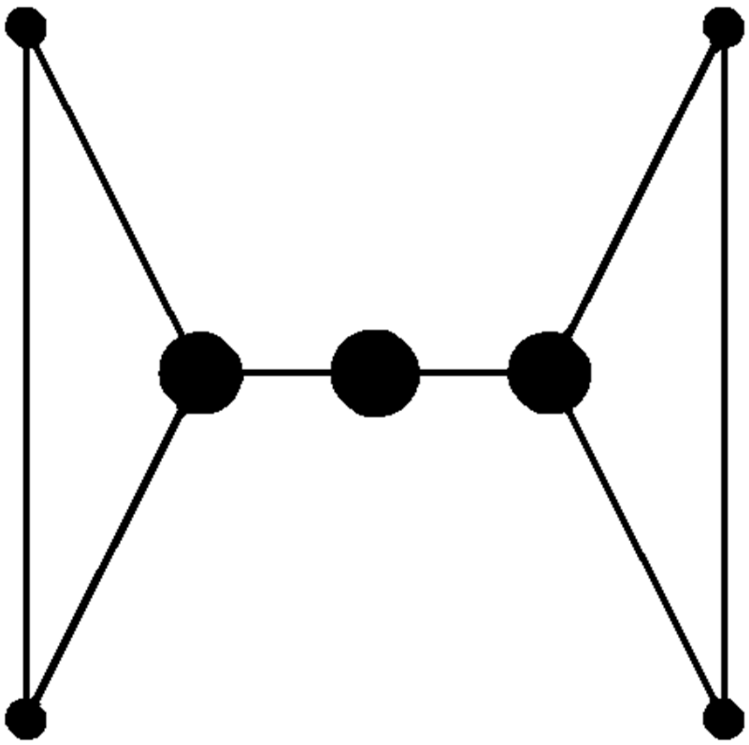
Problem with PageRank algorithm and how Google solved it

- Problem: surfers can be trapped in infinite loops with one or more entrances and no exits
- Model with random jumps mimicking surfers getting bored when following a chain of links

  $G_i \sim (1-\alpha)\sum_j T_{ij}\, G_j + \alpha \sum_j G_j$
- α=0.15 meaning that an average web surfer (circa 1995) on average jumped around 1/$\alpha\approx$6 webpages before going somewhere else

How to find “important” nodes?

- By their connectivity
- Connectors = important
- Betweenness-centrality



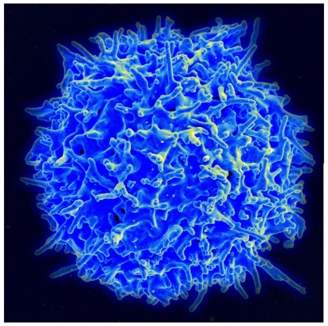
Betweenness centrality: definition

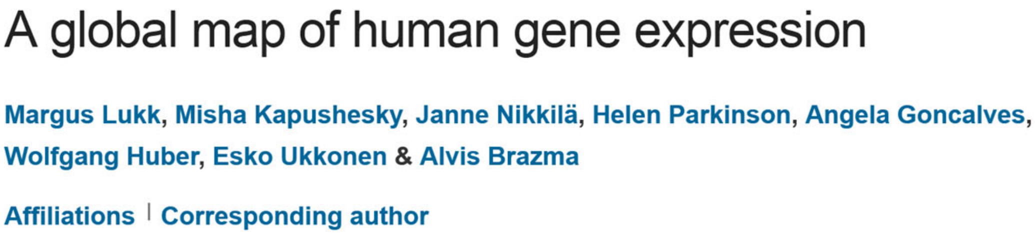
- Take a node i
- There are (N-1)*(N-2)/2 pairs of other nodes
- For each pair find the shortest path on the network
- If more than one shortest path, sample them equally
- Betweenness-centrality C(i) ~ the number of shortest paths going through node i

How is it connected to expression data analysis?

T-cell expression data

- The matrix contains 47 expression samples from Lukk et al, Nature Biotechnology 2011
- All samples are normal T-cells from different individuals
- Only the top 3000 genes with the largest variability were used
- The value is log2 of gene's expression level in a given sample as measured by microarray technology



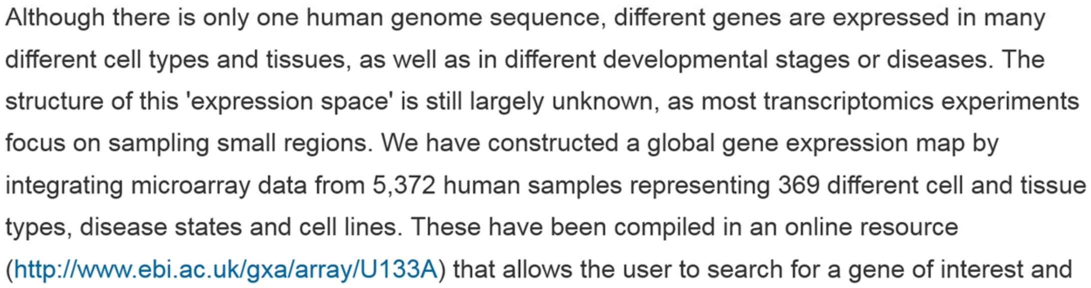


A global map of human gene expression

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

Affiliations | Corresponding author

Nature Biotechnology **28**, 322–324 (2010) | doi:10.1038/nbt0410-322



Although there is only one human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. The structure of this 'expression space' is still largely unknown, as most transcriptomics experiments focus on sampling small regions. We have constructed a global gene expression map by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. These have been compiled in an online resource (http://www.ebi.ac.uk/gxa/array/U133A) that allows the user to search for a gene of interest and

Correlated pairs

plausible biological connection based

on short description

g1=1994; g2=188; group 1

g1=2872; g2=1269; group 2

g1=1321; g2=10; group 3

g1= 886; g2=819; group 4

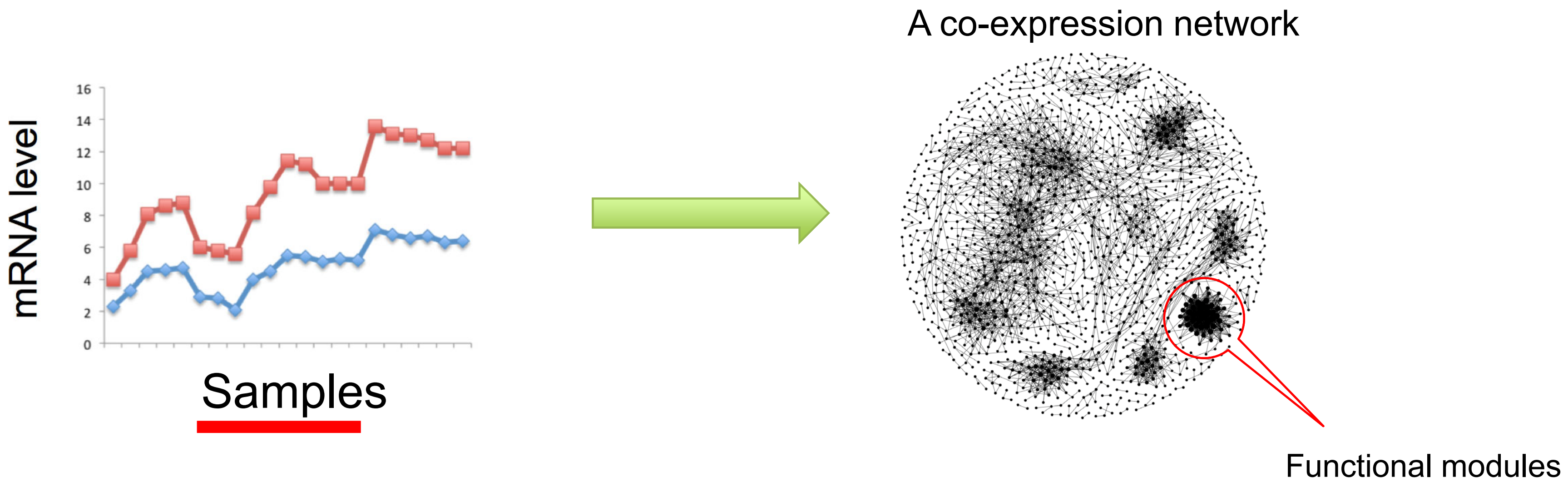
g1=2138; g2=1364; group 5

no obvious biological common function

```
g1=1+floor(rand.*3000); g2=1+floor(rand.*3000);
disp([g1, g2])
```

To analyze correlations in expression for all pairs of genes: Co-expression networks

How to construct a co-expression network?



- Start with a matrix of log2 of expression levels of N genes in K samples (conditions): for our T-cell data N=3000, K=47

- For each of N(N-1)/2 pairs of genes i and j calculate the correlation coefficient $\rho_{ij}=\sigma_{ij}/\sigma_i\sigma_j$ of gene levels across K samples

- Put a threshold, e.g. $\rho_{ij}>0.85$, or otherwise select the most correlated pairs of genes (~4500 in our case). Now you have a weighted network.

- Identify densely interconnected functional modules in this network.

- Modules can be used to infer unknown functions of genes via "Guilt by Association" principle.

How to install Gephi software for network analysis?

- Install Gephi from: https://gephi.org/users/download/
- One of the common problems with installation is the version of Java on your computer. One possible solution is here: https://github.com/gephi/gephi/issues/1787.

Sometimes after installation Gephi may complain that it cannot find java version 1.8 or higher. In this case you need to go to C:\Program Files\Gephi-0.9.2\etc
Open file gephi.conf using notepad.exe (MS Word does not work!).
Add a line jdkhome="C:\Program Files (x86)\Java\jre1.8.0_231"
(the numbers in …jre1.8.0_231 may be changed to reflect the actual directory where Java is installed on your computer).
If JDK is not installed on your computer, you need to install itfirst from https://www.java.com/en/download/win10.jsp"

Co-expression network analysis exercise

- Start Gephi and open coexpression_network_random_start.gephi
- Run "Layout" → Fruchterman Reingold → Speed 10.0
- Run "Average degree", "Network diameter", "Modularity" in the Statistics tab in the right panel.
- Color nodes by "modularity class":
  Appearance → Nodes → Partition → Palette Icon → Modularity class
- Size nodes first by "degree".
  Appearance → Nodes → Ranking → Multiple Circles Icon → Degree
  - If the nodes are too small, select "Min size": 10 and "Max size":80
  - Nodes in large tightly connected clusters have large degree
- Then size nodes by "betweenness-centrality"
  Appearance → Nodes → Ranking → Multiple Circles Icon → Betweenness-centrality
  - Large circles are "coordinator" genes connecting different co-expressed clusters to each other. Potentially biologically interesting

Disease-disease similarity network

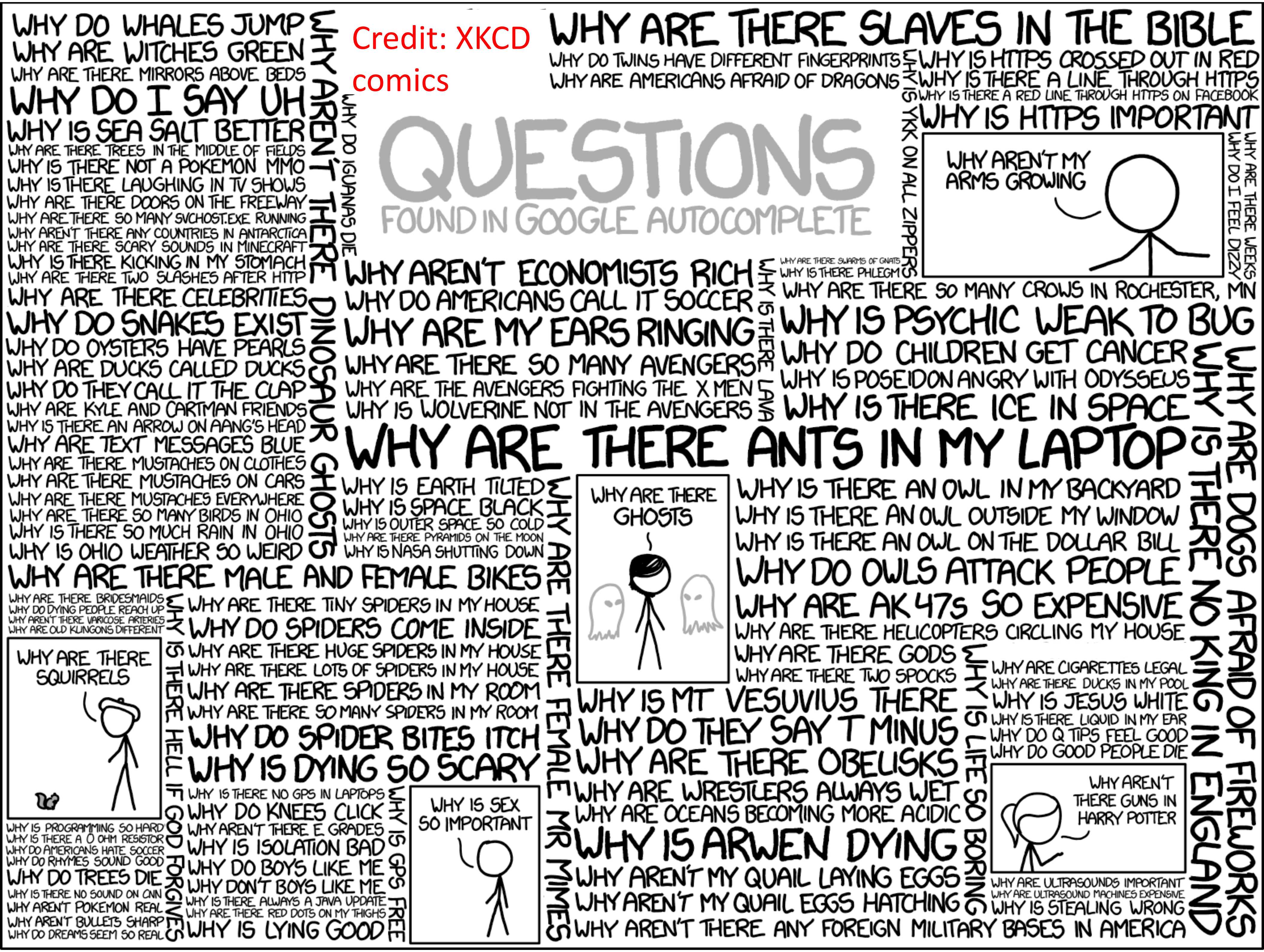
- Based on the table summarizing **all current medical knowledge** of genes implicated in diseases:
  - Rows: 516 common human diseases
  - Columns: 25,000 human genes
  - Matrix element $D_{i\alpha}$ =1 if the gene α is known to be involved in the disease i . 0 – otherwise
- Constructed disease-disease similarity network:
  - Weight of the edge - # of shared genes between two diseases
  - Easy to construct: the adjacency matrix A of the network is simply $A=D \bullet D^{+}$

Disease network analysis exercise

- Start Gephi and open disease_disease_random_start.gephi
- Run “Layout” → Fruchterman Reingold → Speed 10.0
  Observe how clusters emerge.
- Run “Average degree”, “Network diameter”, “Modularity” analysis tools in the right panel.
- Color nodes with medical term: “disorder class”
  Appearance → Nodes → Partition → Palette Icon → Disorder class
- Then color nodes by “modularity class”. See how well it agrees with the previous color.
  Appearance → Nodes → Partition → Palette Icon → Modularity class
- Size nodes first by “degree”.
  Appearance → Nodes → Ranking → Multiple Circles Icon → Degree
  - Which disease has the largest degree?
- Size nodes by “betweenness centrality”
  Appearance → Nodes → Ranking → Multiple Circles Icon → Degree
  - Which diseases have the largest betweenness-centrality?

  These “connector” diseases linking different diseases clusters to each other. They highlight potentially interesting connections between diseases

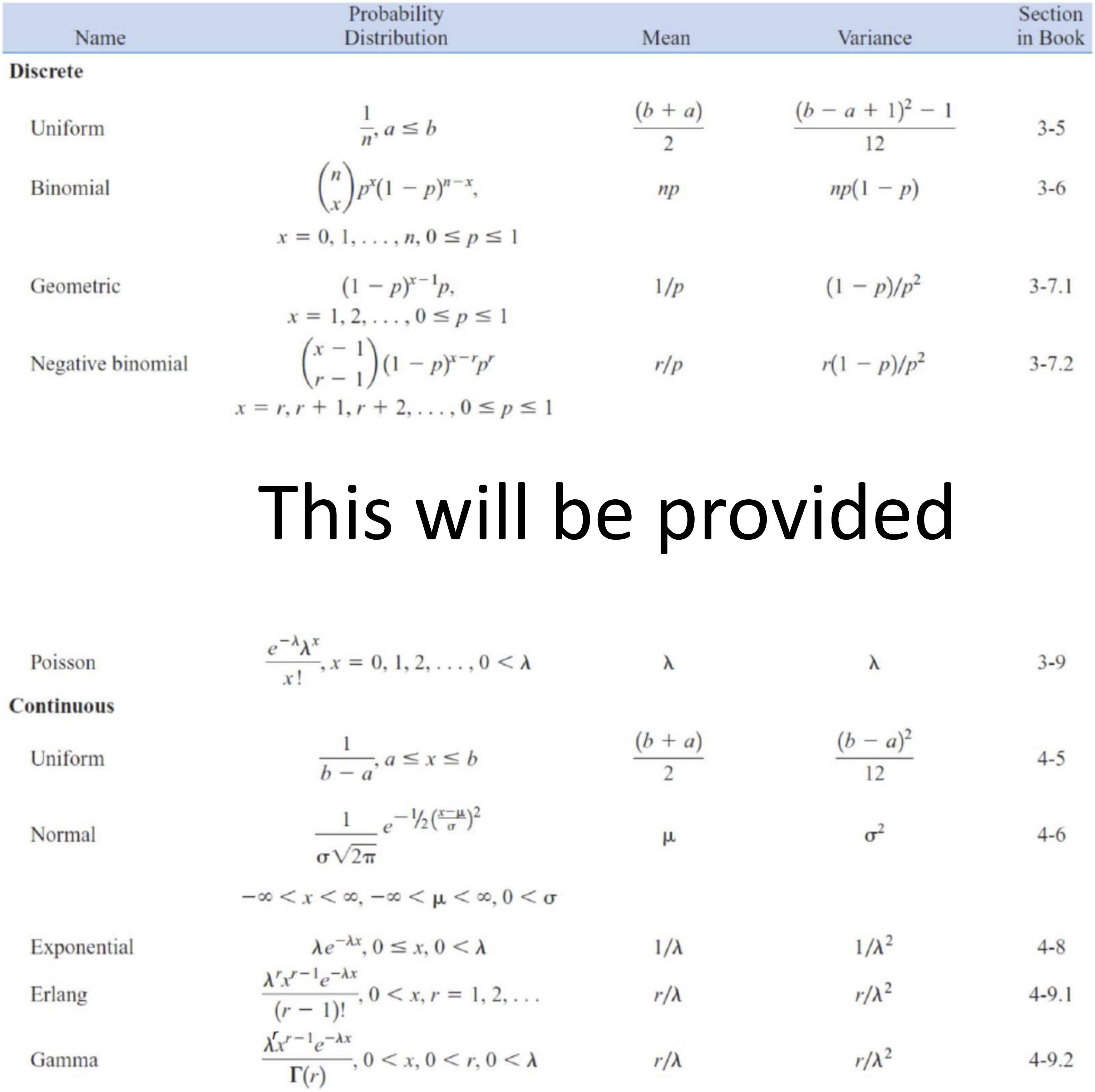
Credit: XKCD comics


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY IS LIFE SO BORING
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Review for the Final Exam

Rules

- Closed book exam; no books, notes, laptops, smartphones, etc.
- However, calculators (not on a smartphone) can be used.
- You can prepare one cheat sheet (letter size, two-sided if needed)
- Printouts provided:
  - Distributions means/variances/pdfs
  - Standard normal distribution CDF table



Name	Probability Distribution	Mean	Variance	Section in Book
Discrete				
Uniform	$\frac{1}{n}, a \le b$	$\frac{(b+a)}{2}$	$\frac{(b-a+1)^2-1}{12}$	3-5
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$,	np	$np(1-p)$	3-6
	$x = 0, 1, \ldots, n, 0 \le p \le 1$			
Geometric	$(1-p)^{x-1}p$,	$1/p$	$(1-p)/p^2$	3-7.1
	$x = 1, 2, \ldots, 0 \le p \le 1$			
Negative binomial	$\binom{x-1}{r-1}(1-p)^{x-r}p^r$	r/p	$r(1-p)/p^2$	3-7.2
	$x = r, r+1, r+2, \ldots, 0 \le p \le 1$			
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, 0 < \lambda$	λ	λ	3-9
Continuous				
Uniform	$\frac{1}{b-a}, a \le x \le b$	$\frac{(b+a)}{2}$	$\frac{(b-a)^2}{12}$	4-5
Normal	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2	4-6
	$-\infty < x < \infty, -\infty < \mu < \infty, 0 < \sigma$			
Exponential	$\lambda e^{-\lambda x}, 0 \le x, 0 < \lambda$	$1/\lambda$	$1/\lambda^2$	4-8
Erlang	$\frac{\lambda^r x^{r-1} e^{-\lambda x}}{(r-1)!}, 0 < x, r = 1, 2, \ldots$	r/λ	r/λ^2	4-9.1
Gamma	$\frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}, 0 < x, 0 < r, 0 < \lambda$	r/λ	r/λ^2	4-9.2

This will be provided

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

What may be on the final exam?

- Probability Multiplication, Combinatorics
- Bayes Theorem
- Discrete & Continuous Random Variables
- Joint Probability Distributions, Covariation/Correlations
- Sampling distributions and parameter point estimation
- Confidence Intervals
- Hypothesis testing for one and two samples
- Other topics
- Look at Homework 1-5 for examples of problems

One-sample hypothesis testing

3. **(8 points)** The college bookstore tells prospective students that the average cost of its textbooks is \$52 with a standard deviation of \$4.50. A group of statistics students think that the average cost is **actually higher**. In order to test bookstore's claim against this alternative hypothesis, the students bought a random sample of 100 books. The mean price of this sample was \$52.80. Perform the hypothesis test at the 5% level of significance and state your decision.

Two-sample hypothesis

Mating Calls. In a study of mating calls in the gray treefrogs *Hyla hrysoscelis* and *Hyla versicolor*, Gerhart (1994) reports that in a location in Lousiana the following data on the length of male advertisement calls have been collected:

	Sample size	Average duration	SD of duration	Duration range
Hyla chrysoscelis	43	0.65	0.18	0.36–1.27
Hyla versicolor	12	0.54	0.14	0.36–0.75

The two species cannot be distinguished by external morphology, but *H. chrysoscelis* are diploids while *H. versicolor* are tetraploids. The triploid crosses exhibit high mortality in larval stages, and if they attain sexual maturity, they are sterile. Females responding to the mating calls try to avoid mismatches.

Based on the data summaries provided, test whether the length of call is a discriminatory characteristic? Use $\alpha = 0.05$.

	Sample size	Average duration	SD of duration
Hyla chrysoscelis	43	0.65	0.18
Hyla versicolor	12	0.54	0.14

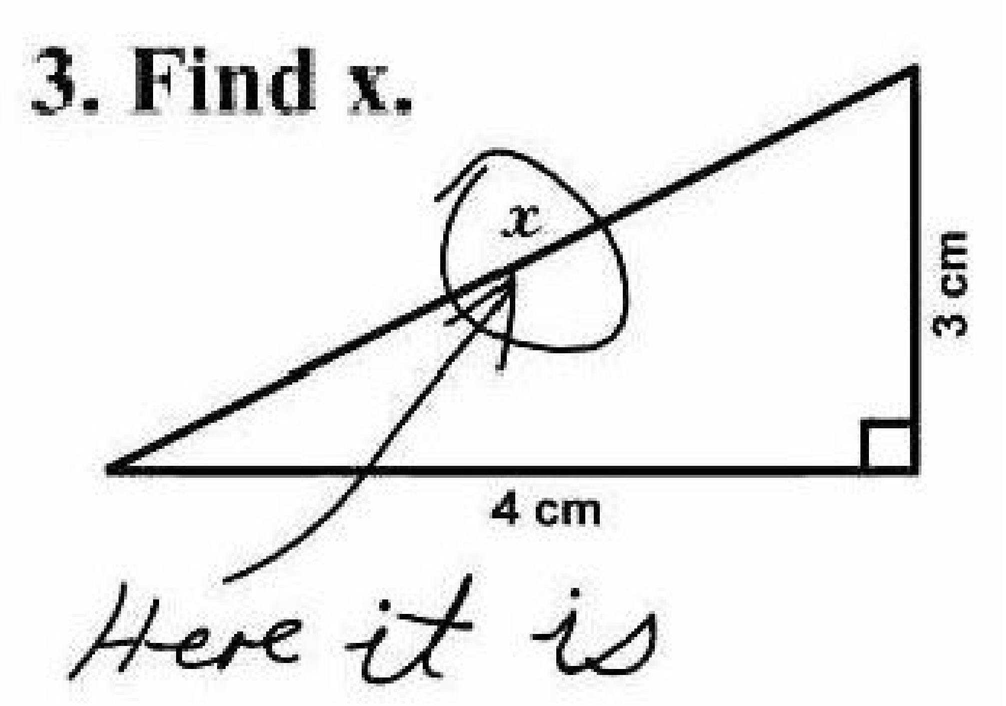
Based on the data summaries provided, test whether the length of call is a discriminatory characteristic? Use $\alpha = 0.05$.

Confidence intervals

2. **(6 points)** The operations manager of a large production plant would like to estimate the mean amount of time a worker takes to assemble a new electronic component. Assume that the standard deviation of this assembly time is 3.6 minutes. After observing a sample of 100 workers assembling similar devices, the manager noticed that their average time was 16.2 minutes. Construct a **90% confidence interval** for the population mean of the assembly time.

What is X in this problem?

- What is X? Look for keywords:
  - Find the probability that….
  - What is the mean (or variance) of…
- What are the parameters? Look for keywords:
  - Given that…
  - Assuming that…
- <u>**Is X discrete or continuous?**</u>



Discrete Probability Distributions

(8 points) You are doing a long series of experiments. Assume that each of your experiments has a probability of 0.02 of succeeding. Assume that your experiments are independent.

(A) (2 points) What is the probability that you first succeed on tenth experiment?

(B) (2 points) What is the probability that it requires more than five experiments for you to succeed?

(C) (2 points) What is the mean number of experiments needed to succeed once?

(D) (2 points) What is the probability that the second experiment that worked is the tenth one since you started?

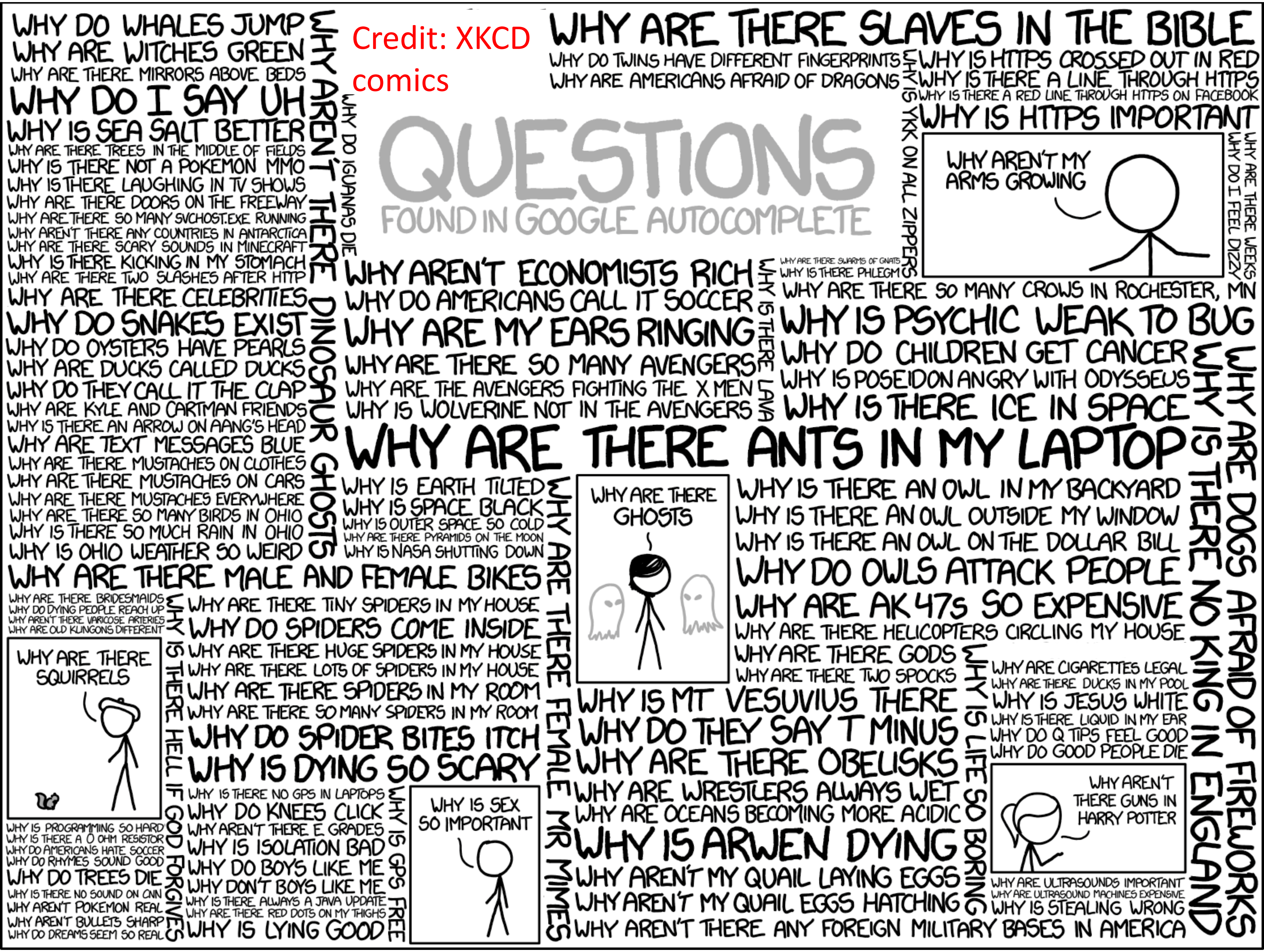
Continuous Probability Distributions

(12 points) Time interval separating subsequent bus arrivals at a stop is an exponential random variable with mean 20 minutes. Steve and Andrew work at the same place and each will be late to work unless they board a bus on or before 8:40am. Steve comes to the bus stop exactly at 8am. Andrew also comes to the same bus stop but at a random time, uniformly distributed between 8am and 8:30am. Both of them take the first bus that arrives.

(a) (4 points) What is the probability that Steve will be late for work tomorrow?

(b) **(4 points)** What is the probability that Andrew will be late for work tomorrow?

(c) **(4 points)** What is the probability that Steve and Andrew will ride the same bus

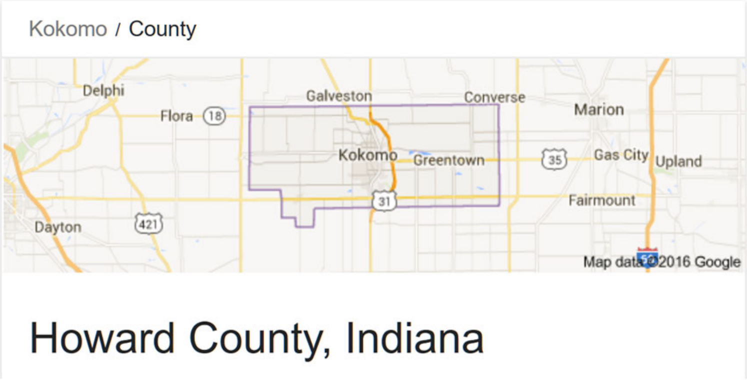


Credit: XKCD comics

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY ARE THERE FEMALE MR MIMES
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

Bayes theorem

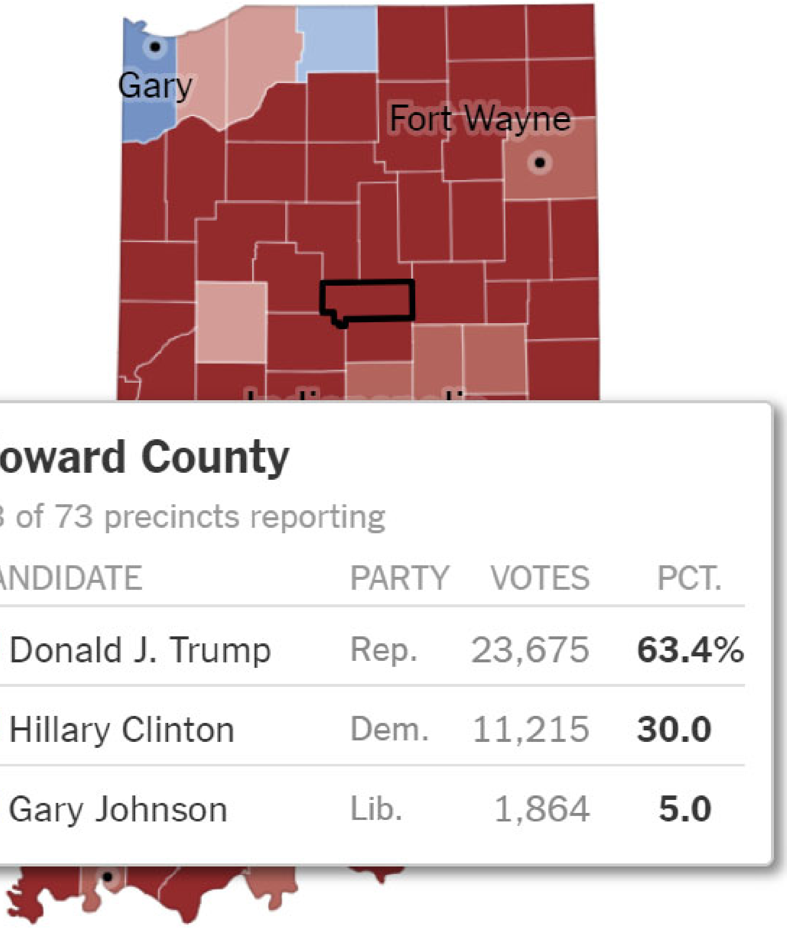
Kokomo, Indiana. In Kokomo, IN, 65% of the people are conservative, 20% are liberal, and 15% are independent. Records show that in a particular election, 82% of conservatives voted, 65% of liberals voted, and 50% of independents voted. If a person from the city is selected at random and it is learned that she did not vote, what is the probability that the person is liberal?



Kokomo / County

Howard County, Indiana

As of the 2010 census, the population was 82,752. The county seat is Kokomo, IN.



Howard County

73 of 73 precincts reporting

CANDIDATE	PARTY	VOTES	PCT.
🟥 Donald J. Trump	Rep.	23,675	**63.4%**
🟦 Hillary Clinton	Dem.	11,215	**30.0**
🟩 Gary Johnson	Lib.	1,864	**5.0**

Kokomo, Indiana. In Kokomo, IN, 65% of the people are conservative, 20% are liberal, and 15% are independent. Records show that in a particular election, 82% of conservatives voted, 65% of liberals voted, and 50% of independents voted. If a person from the city is selected at random and it is learned that she did not vote, what is the probability that the person is liberal?

Joint Probability Distributions

1. **(20 points)** The joint probability mass function of discrete random variables X and Y taking values x = 1, 2, 3 and y = 1, 2, 3, respectively, is given by $f_{XY}(x, y) = c^*(x + y)$. Determine the following:

a) **(2 points)** Find c
b) **(2 points)** Find probability of the event, where X = 1 and Y < 3
c) **(2 points)** Find marginal probability $P_Y(Y = 2)$
d) **(2 points)** Find marginal probability distribution of the random variable X
e) **(2 points)** Find E(X), E(Y), V(X), and V(Y)
f) **(2 points)** Find conditional probability distribution of Y given that X = 1
g) **(2 points)** Conditional probability distribution of X given that Y = 2
h) **(2 points)** Are X and Y independent?
i) **(2 points)** What is the covariance for X and Y?
j) **(2 points)** What is the correlation for X and Y?

1. **(20 points)** The joint probability mass function of discrete random variables X and Y taking values x = 1, 2, 3 and y = 1, 2, 3, respectively, is given by $f_{XY}(x, y) = c*(x + y)$. Determine the following:

a) **(2 points)** Find c

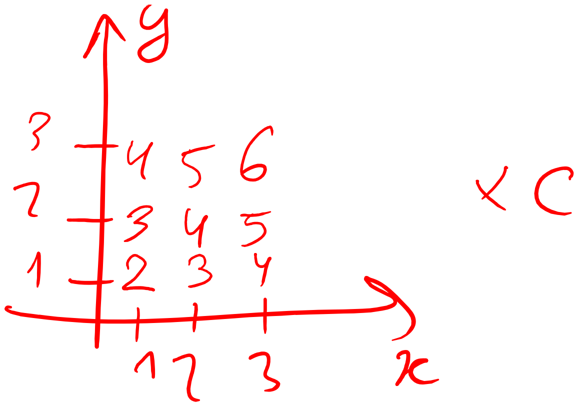
b) **(2 points)** Find probability of the event, where X = 1 and Y < 3

c) **(2 points)** Find marginal probability $P_Y(Y = 2)$

f) **(2 points)** Find conditional probability distribution of Y given that X = 1

(a) $1 = c\cdot(2+3+4+\ 3+4+5+4+5+6)$

$c = 1/36$

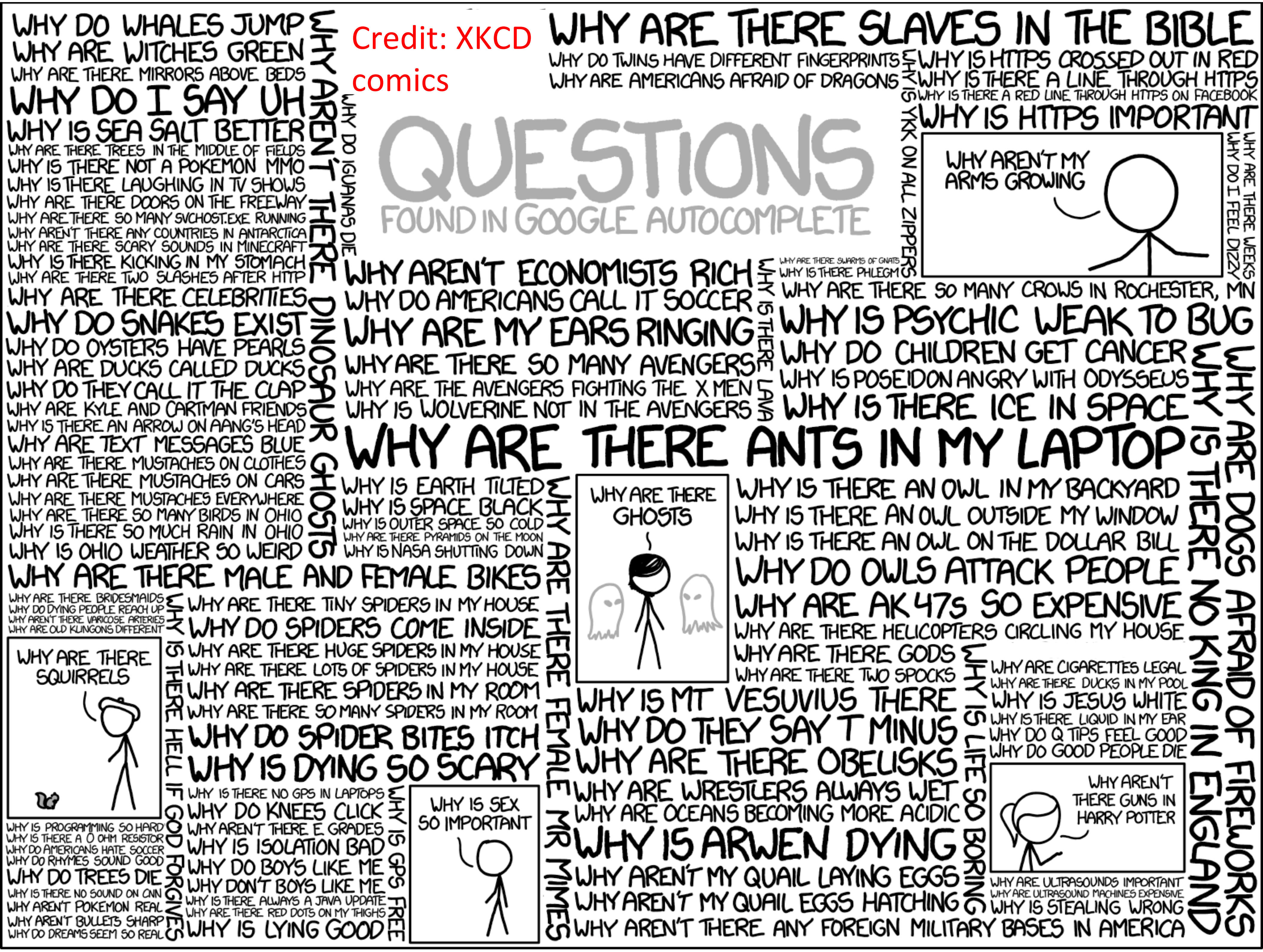


(b) $P(X=1, Y<3) = \dfrac{2+3}{36} = \dfrac{5}{\ }$

(c) $P_Y(Y=2) = \dfrac{3+4+5}{36} = \dfrac{12}{36} = \dfrac{1}{3}$

(f) $P(Y=2|X=1) = \dfrac{P(Y=2, X=1)}{P_X(X=1)} = \dfrac{3/36}{(2+3+4)/36} = \dfrac{1}{3}$

Credit: XKCD comics


QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE FEMALE MR MIMES
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG