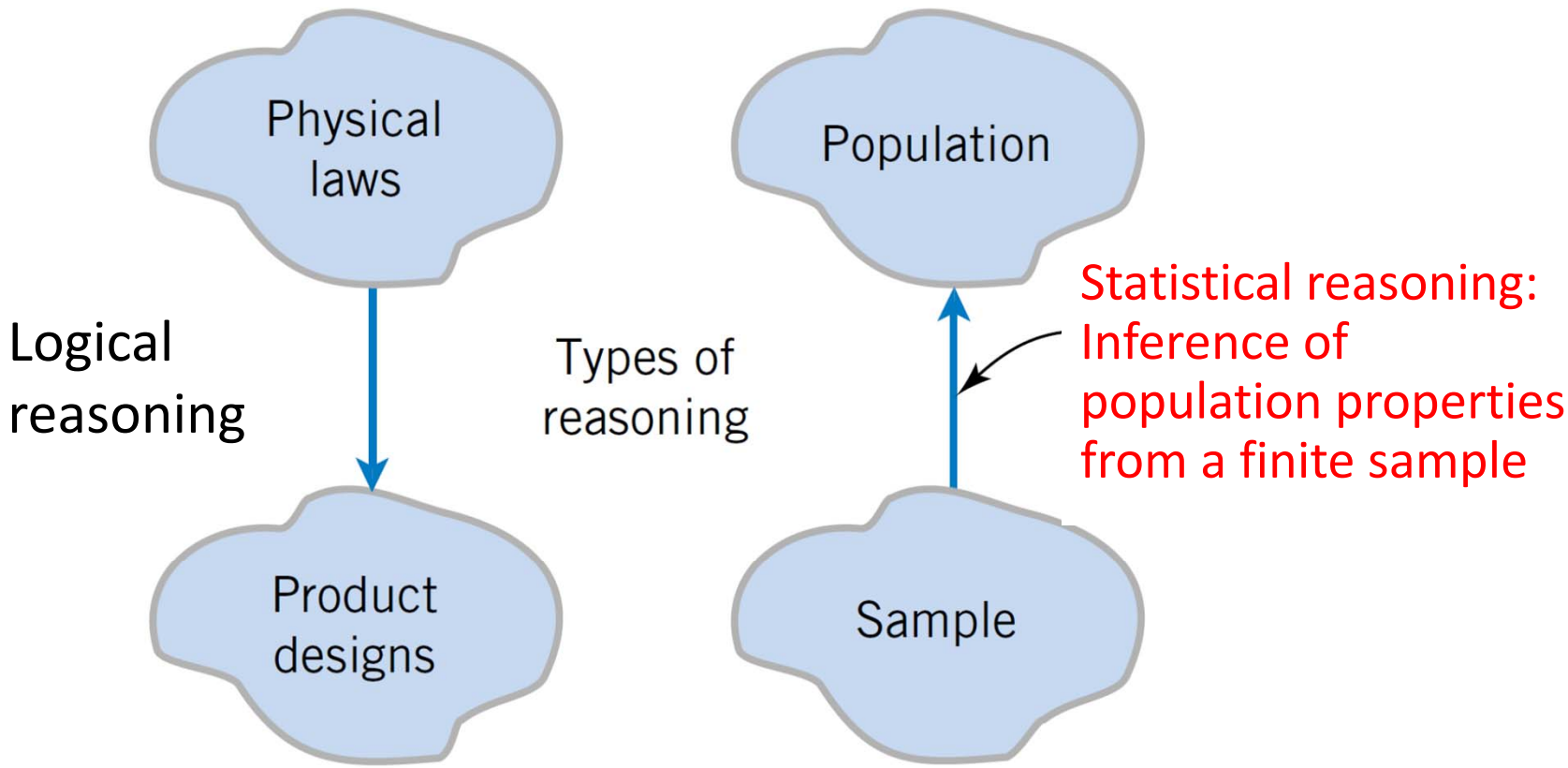
# Descriptive statistics:

Populations, Samples

Histograms, Quartiles

# Two types of reasoning



# Numerical Summaries of Data

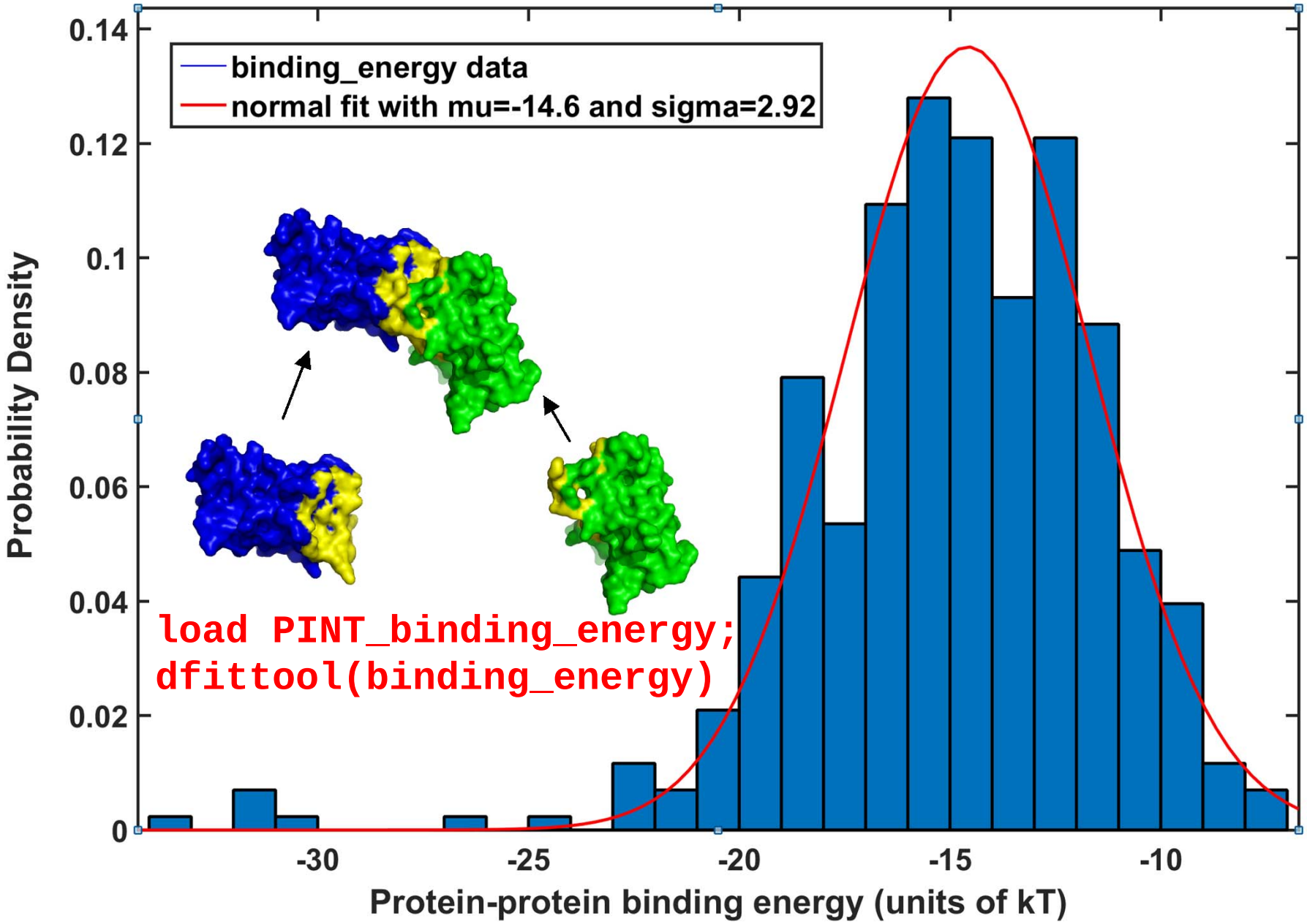
- Data are the numerical observations of a phenomenon of interest.
- The totality of all observations is a population.
  - Population can be infinite (e.g. random variable)
  - It can be very large (e.g. 7 billion humans)
- A (usually small) portion of the population collected for analysis is a random 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 $X_1, X_2, \ldots, X_n$ 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

# Ways to describe a sample: Histogram

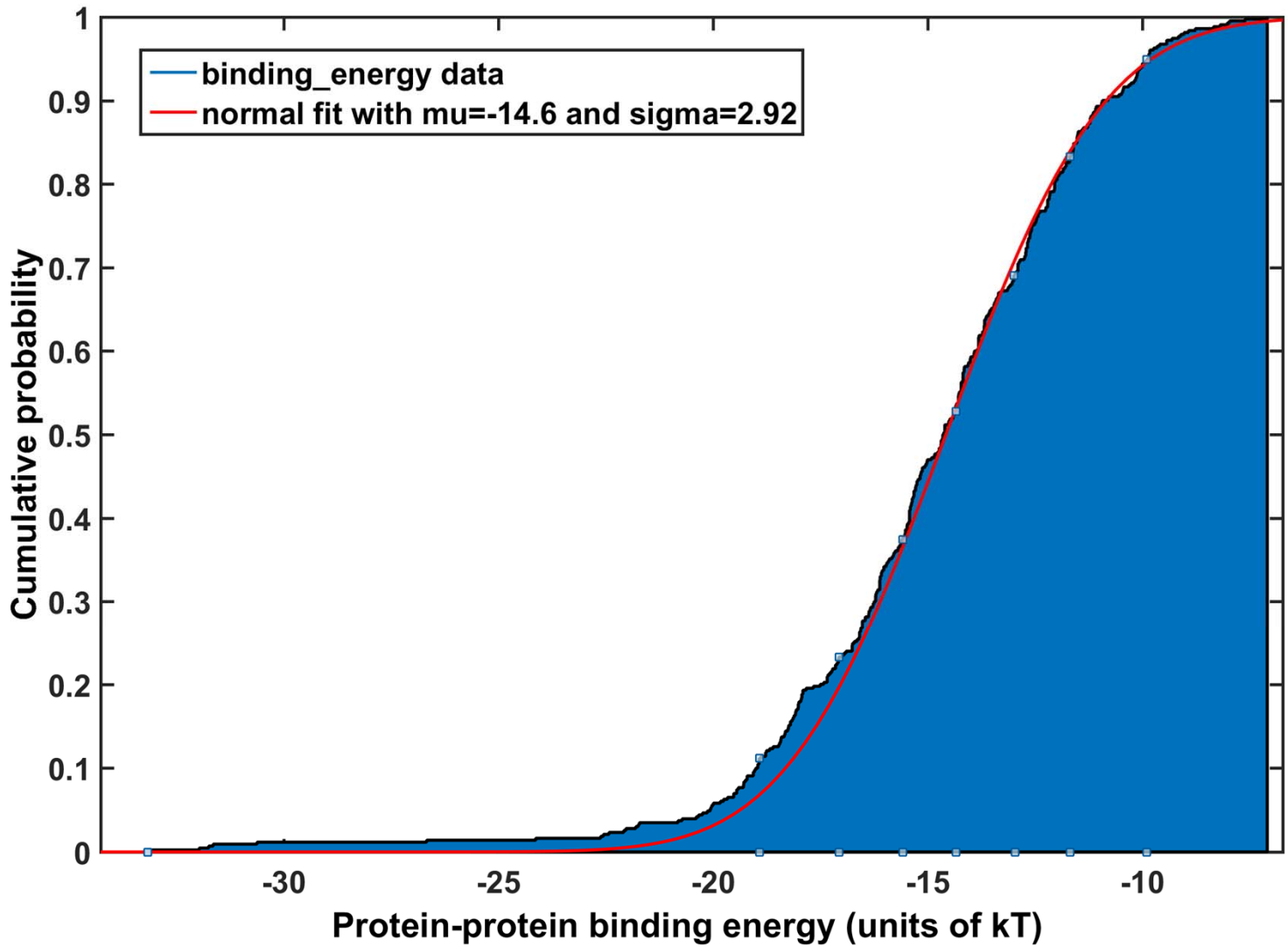

binding_energy data
normal fit with mu=-14.6 and sigma=2.92
Probability Density
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-30
-25
-20
-15
-10
Protein-protein binding energy (units of kT)
load PINT_binding_energy;
dfittool(binding_energy)

# 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 $p_n$

# Matlab exercise

- Find the median and lower & upper quartiles of n=100 sample drawn from a continuous uniform distribution in [0,1]
- Do not use built-in Matlab functions for this exercise!
- Hint: use [a,b]=sort(r1); to rank order your sample. The variable a returns r1 sorted in the order of increasing values.
- How to find quartiles from a?

# How to find median & quartiles

```
% Example: find median and lower quartile of
% a sample with n=100 drawn from uniform
r1=rand(100,1);
[a,b]=sort(r1);
med=(a(50)+a(51))./2
sum(r1<med) % verify
q1=(a(25)+a(26))./2
sum(r1<q1) % verify
```

# 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*.

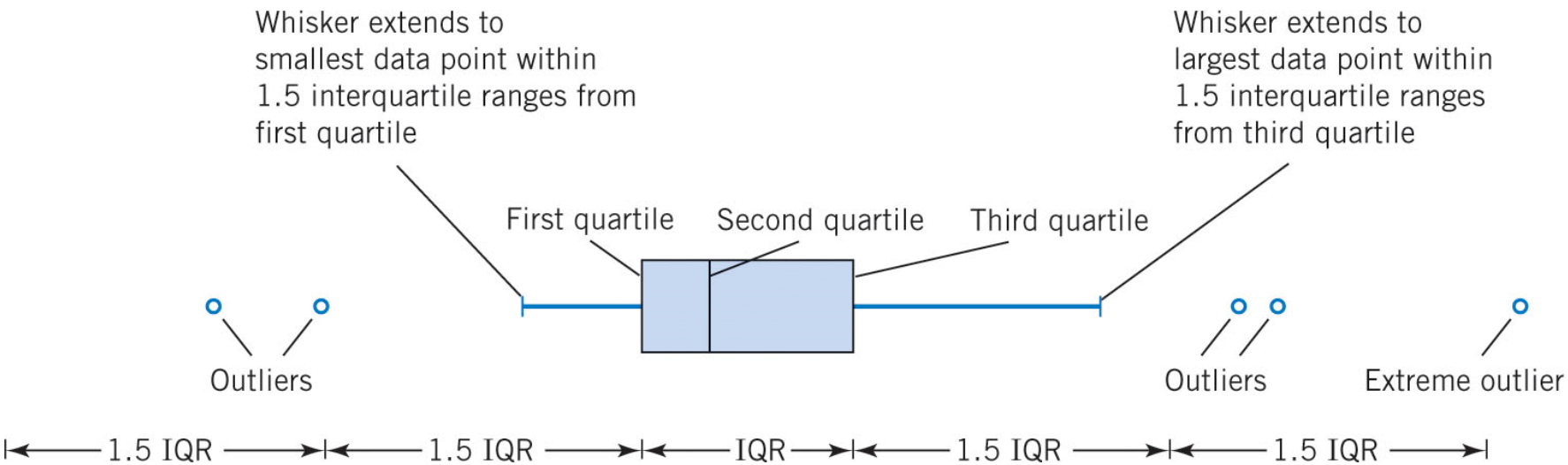
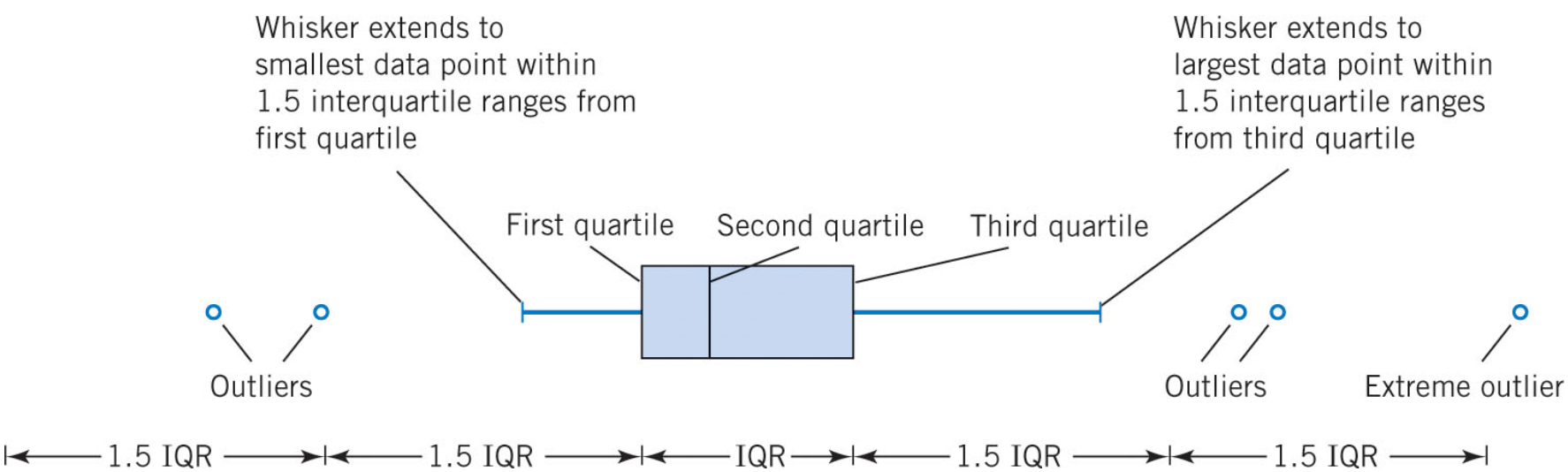



Figure 6-13 Description of a box plot.

# Matlab exercise:

- 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



# How many right outliers one expects in a sample of n=1000 following normal distribution?

- % find the third quartile of a standard distribution
- norminv(0.75)  %ans = 0.6745
- % Calculate IQR - Inter Quartale Range
- IQR=2.*norminv(0.75) % 1.3490
- % Calculate 0.5*IGR+1.5*IQR - the right whisker position
- whisker=0.5.*IQR+1.5*IQR %ans =  2.6980
- % Find the probability to be above the right whisker
- 1-normcdf(whisker) %ans =  0.00349
- % Find number of right outliers in a sample of 1000 points
- 1000.*(1-normcdf(whisker)) %ans = 3.49

# Descriptive statistics:

# Sample mean and variance

# Linear Functions of Random Variables

- A function of random variables is itself a random variable.
- A function of random variables can be formed by either linear or nonlinear relationships. We start with linear functions.
- Given random variables $X_1, X_2, \ldots, X_p$ and constants $c_1, c_2, \ldots, c_p$

  $$Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p \qquad (5\text{-}24)$$

  is a linear combination of $X_1, X_2, \ldots, X_p$.

# 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 $\operatorname{cov}(X_iX_j) = 0$,

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

# Example 5-31: Error Propagation

A semiconductor product consists of three layers. The variances of the thickness of each layer is 25, 40 and 30 $nm^2$. What is the variance of the finished product?

Answer:

IMPORTANT

$$X = X_1 + X_2 + X_3$$

Add variations not standard deviations

$$V(X) = \sum_{i=1}^{3} V(X_i) = 25 + 40 + 30 = 95 \text{ nm}^2$$

$$SD(X) = \sqrt{95} = 9.7 \text{ nm}$$

If adding SDs one would get $\sqrt{25}$ nm + $\sqrt{40}$ nm + $\sqrt{30}$ nm = 16.08 nm

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

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

IMPORTANT:

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

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

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

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

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

# Central Limit Theorem

If $X_1, X_2, ..., X_n$ is a random sample of size $n$ is taken from a population (either finite or infinite) with mean $\mu$ and finite variance $\sigma^2$, and if $\bar{X}$ is the sample mean,

then the limiting form of the distribution of

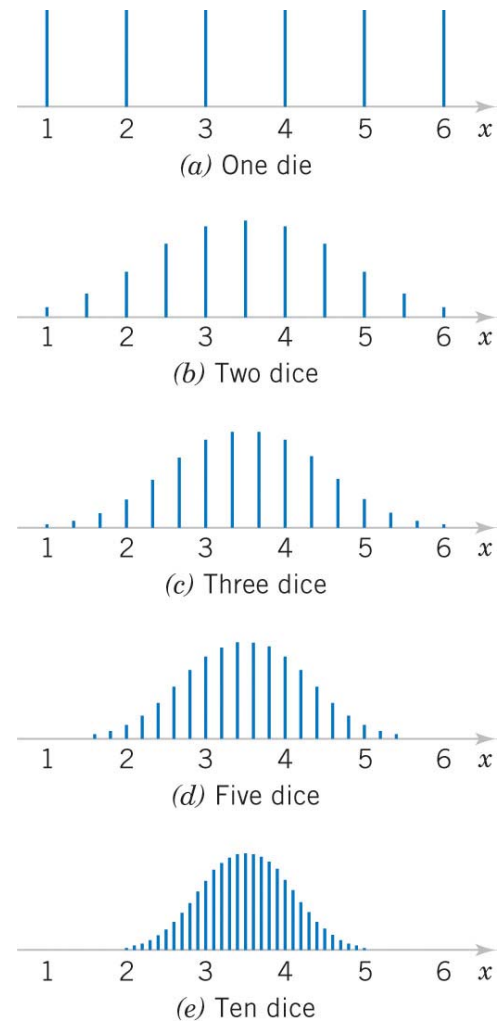
$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \qquad (7\text{-}1)$$

for large $n$, is the standard normal distribution.

If $X_1, X_2, ..., X_n$ are themselves normally distributed - for any $n$

# 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



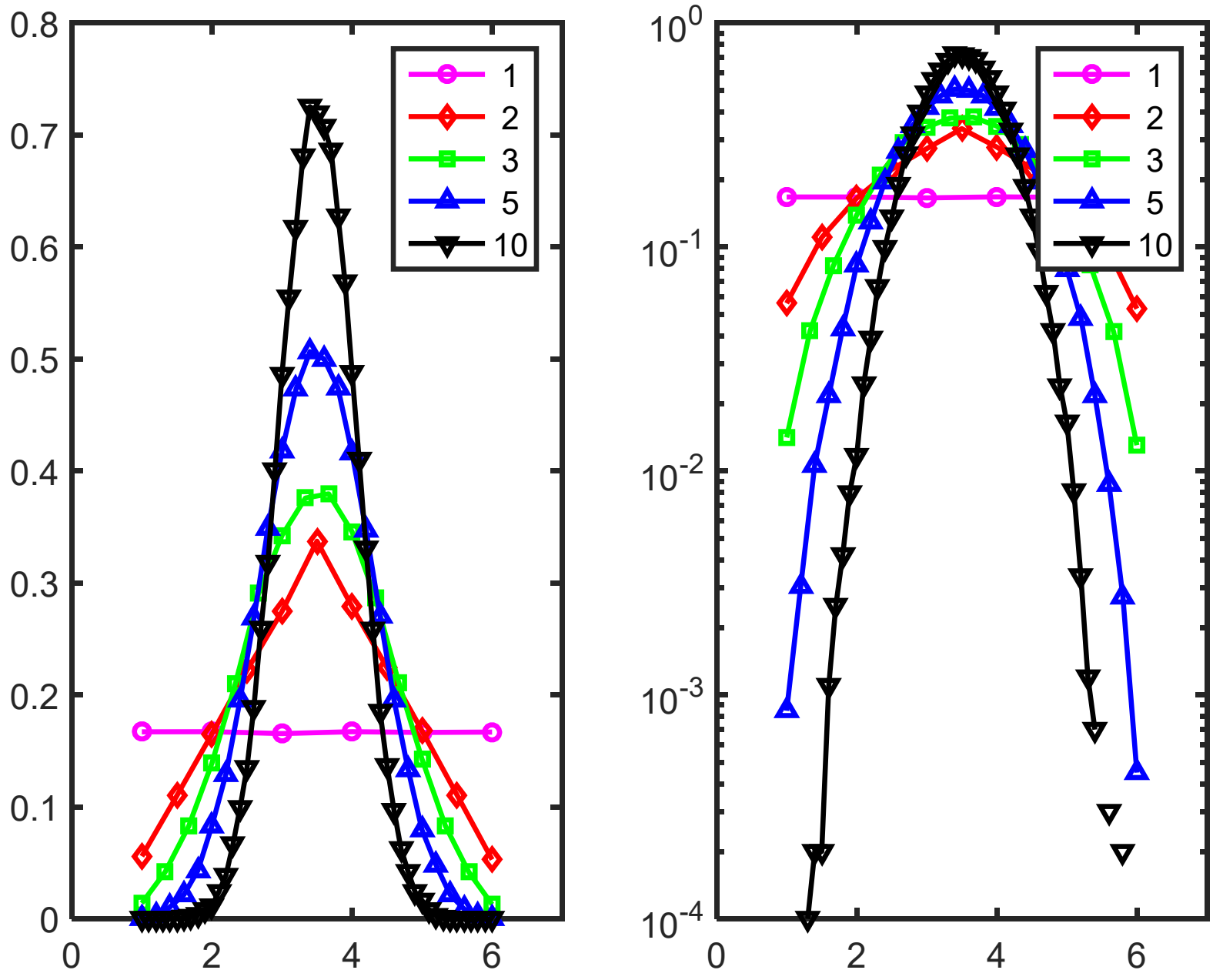

## Formulas

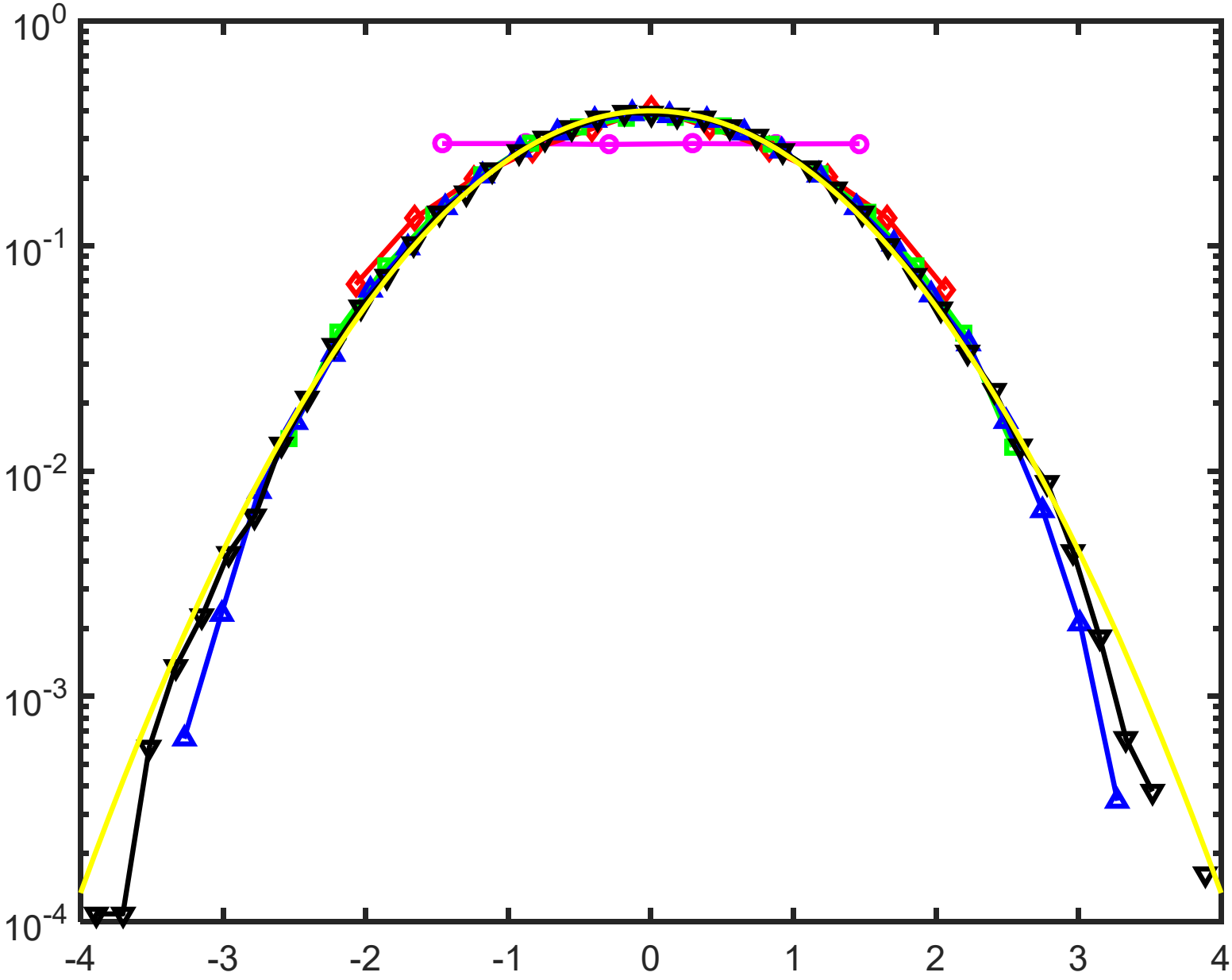
$$\mu = \frac{b+a}{2}$$

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

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

show Matlab


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


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

# 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?

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

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

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

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

$$= 0.0062$$

# 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?

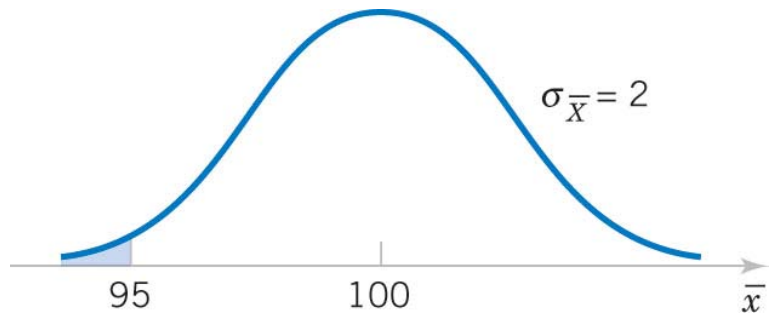



Figure 7-2 Desired probability is shaded

Answer:

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

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

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

# 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 $\mu_1$ and $\mu_2$, and variances $\sigma_1^2$ and $\sigma_2^2$,
- 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:

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

  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 $\mu_{old}$=5000 hours and $\sigma_{old}$=40 hours (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to $\mu_{new}$=5050 hours and $\sigma_{new}$=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 $\mu_{old}$=5000 hours and $\sigma_{old}$=40 hours (old). The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to $\mu_{new}$=5050 hours and $\sigma_{new}$=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?

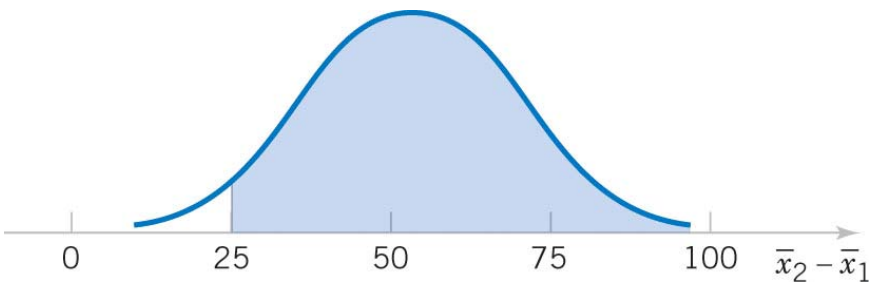



Figure 7-4 Sampling distribution of the sample mean difference.

| | 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 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 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 SEX SO IMPORTANT
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
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 SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
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 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 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 WEEKS
WHY DO I FEEL DIZZY
WHY IS YKK ON ALL ZIPPERS
WHY DO IGUANAS DIE
WHY ARE THERE DINOSAUR GHOSTS
WHY IS THERE LAVA
WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE THERE FEMALE MR MIMES
WHY IS GPS FREE
WHY IS THERE HELL IF GOD FORGIVES
WHY IS LIFE SO BORING