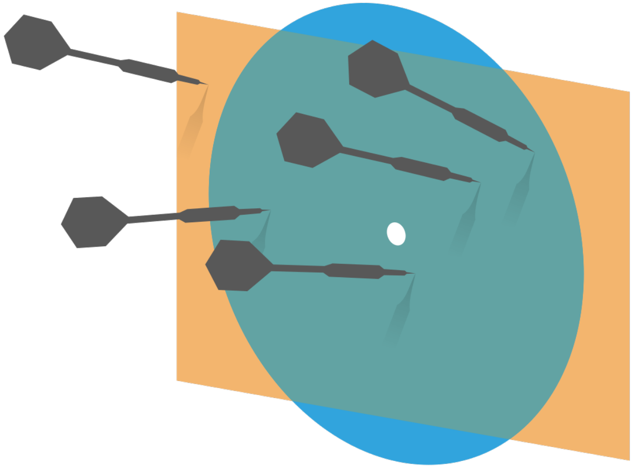
Probability and Statistics for Computer Science



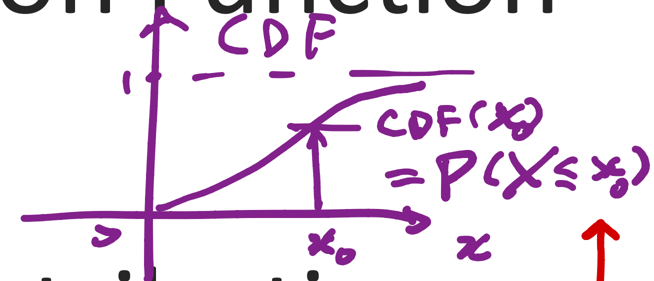
Credit: wikipedia

"In statistics we apply probability to draw conclusions from data."
---Prof. J. Orloff

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 10.06.2020

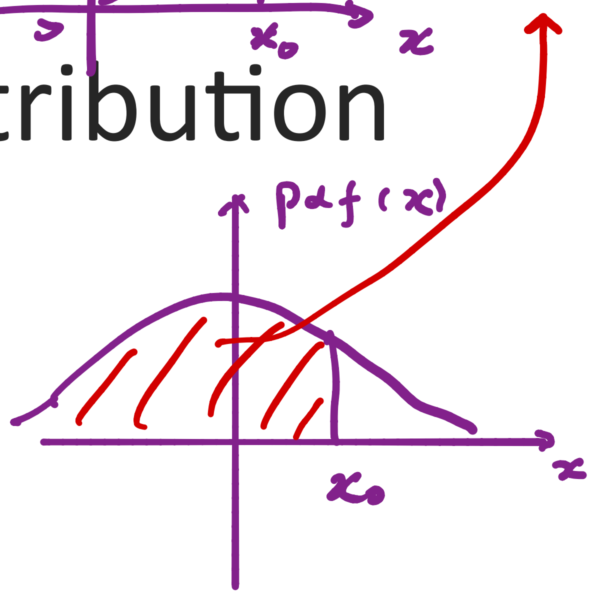
Last time

- Cumulative Distribution Function of a continuous RV



- Normal (Gaussian) distribution

CLT

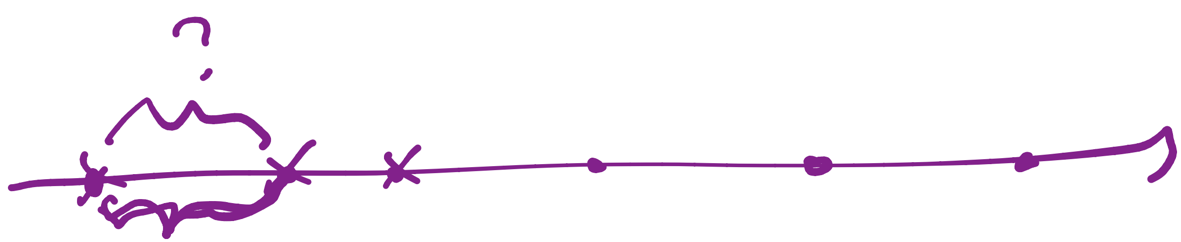


$$P(X \le x_0) = \int_{-\infty}^{x_0} p(x)\,dx$$

Objectives

- Exponential Distribution
- Sample mean and confidence interval

1 hr

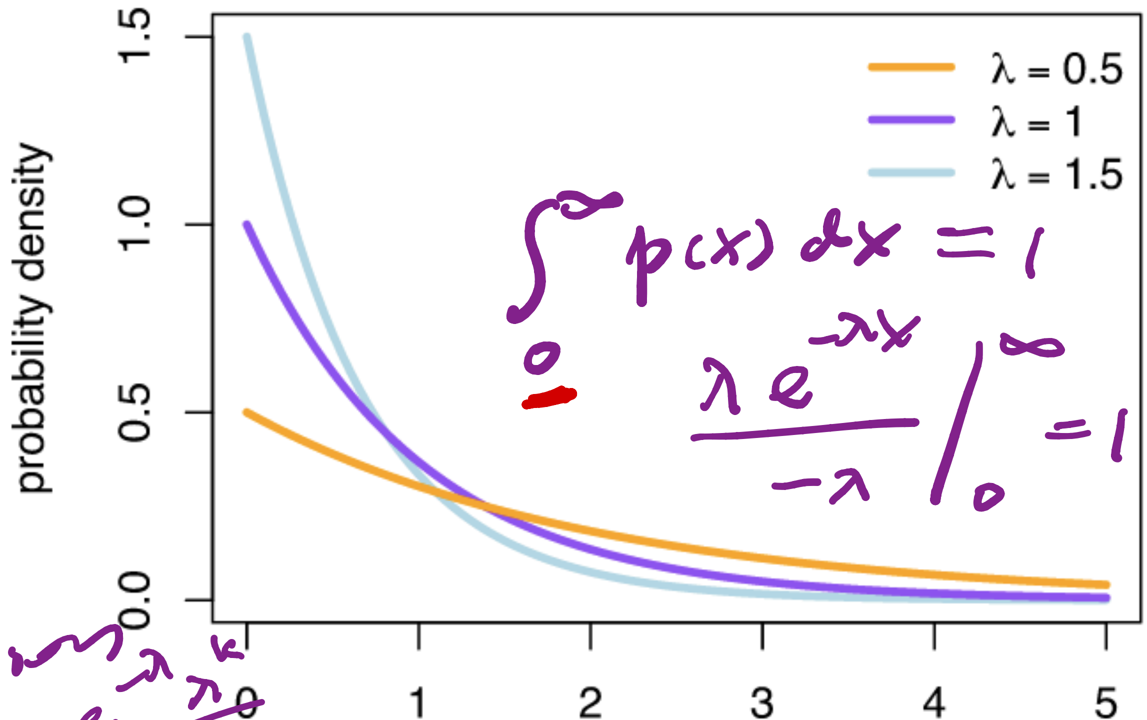


$\lambda =$

Exponential distribution

- Common Model for waiting time
- Associated with the Poisson distribution with the same **λ**

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & for\ x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



$$\int_0^\infty p(x)\,dx = 1$$

$$\frac{\lambda e^{-\lambda x}}{-\lambda}\Big|_0^\infty = 1$$

Poisson $\frac{e^{-\lambda}\lambda^k}{k!}$

Credit: wikipedia

Exponential distribution

- A continuous random variable X is exponential if it represent the “time” until next incident in a Poisson distribution with intensity **λ**. Proof See Degroot et al Pg 324.

$$p(x) = \lambda e^{-\lambda x} \quad for\ x \geq 0$$

- It’s **similar** to **Geometric distribution** – the discrete version of waiting in queue

Expectations of Exponential distribution

- A continuous random variable X is exponential if it represent the "time" until next incident in a Poisson distribution with intensity **λ**.

$$\int_0^\infty x p(x) dx = \frac{1}{\lambda}$$

$$p(x) = \lambda e^{-\lambda x} \quad for \ x \geq 0$$

$$\int_0^\infty (x-\bar{x})^2 p(x) dx = \frac{1}{\lambda^2}$$

$$\boxed{E[X] = \frac{1}{\lambda} \quad \& \quad var[X] = \frac{1}{\lambda^2}}$$

Example of exponential distribution

- How long will it take until the next call to be received by a call center? Suppose it's a random variable **T**. If the number of incoming call is a Poisson distribution with intensity **λ = 20 in an hour**. What is the expected time for T?

$$T = \frac{1}{\lambda} = \frac{1}{20} = 0.05 \text{ (hr)}$$

Exponential has the same λ!

Motivation for drawing conclusion from samples

- In a study of new-born babies' health, random samples from different time, places and different groups of people will be collected to see how the overall health of the babies is like.

Weights of babies at 1 month?

Motivation of sampling: the poll example

		DATES	POLLSTER	SAMPLE	RESULT				NET RESULT
U.S. Senate	Miss.	NOV 25, 2018	C+ Change Research	1,211 LV	Espy	46%	51%	Hyde-Smith	Hyde-Smith +5

Source: FiveThirtyEight.com

- This senate election poll tells us:
  - The sample has 1211 likely voters
  - Ms. Hyde-Smith has realized sample mean equal to 51%
- What is the estimate of the percentage of votes for Hyde-smith?
- How confident is that estimate?

Population

- What is a population?
  - It's the entire possible data set $\{X\}$
  - It has a countable size N_p
  - The population mean $popmean(\{X\})$ is a number
  - The population standard deviation is $popsd(\{X\})$ and is also a number

- The population mean and standard deviation are the same as defined previously in chapter 1

Population

$\boxed{\{X\}} = \{1, 2, 3, \cdots 12\}$

$N_p = 12$

6.5

$\text{popmean}(\{X\}) = ?$

$\text{popstd}(\{X\}) = ? \quad \sqrt{\frac{\Sigma(X_i - \ \ )^2}{12}}$

Sample

- The sample is a random subset of the population and is denoted as $\{x\}$, where sampling is done with **replacement**
  - The sample size N is assumed to be much less than population size N_p

    $N << N_p$
  - The **sample mean of a population** is $X^{(N)}$ and is a **random variable**

Sample $\{x\}$ and Sample Mean $X^{(N)}$

1 1 3 4 5 6 . . .

$\{X\} = \{1, 2, 3, \ldots 12\}$

One random sample $\boxed{\{x\}} = \{1, 1, 2, 3, 3\}$ $N = 5$

$\boxed{X^{(N)}}$ RV takes value ?

$\frac{10}{5}$

$$X^{(N)} = \frac{x_1 + x_2 + \cdots + x_N}{N} = 2$$

Another random sample $\rightarrow \{1, 1, 1, 1, 1\} \Rightarrow X^{(N)} = 1$

Sample mean of a population

- The sample mean of a population is very similar to the sample mean of ***N*** random variables if the samples are **IID samples** -randomly & independently drawn with replacement.

- Therefore the expected value and the standard deviation of the sample mean can be derived similarly as we did in the proof of the weak law of large numbers.

Sample mean of a population

- The sample mean is the average of **IID** samples

$$X^{(N)} = \frac{1}{N}(X_1 + X_2 + ... + X_N)$$

- By linearity of the expectation and the fact the sample items are identically drawn from the same population with replacement

$$E[X^{(N)}] = \cancel{\frac{1}{N}}(\underbrace{E[X^{(1)}] + E[X^{(1)}].. + E[X^{(1)}]}_{\cancel{N} \cdot E[X^{(1)}]}) = E[X^{(1)}]$$

$N=1$

Expected value of one random sample is the population mean

- Since each sample is drawn uniformly from the population

$$E[X^{(1)}] = popmean(\{X\})$$

therefore $$E[X^{(N)}] = popmean(\{X\})$$

- We say that $X^{(N)}$ is an unbiased estimator of the population mean.

$\frac{1}{N} \cdot x_1 + \frac{1}{N} \cdot x_2 \cdots \frac{1}{N} \cdot x_{N_p}$

$= popmean(\{X\})$

Standard deviation of the sample mean

- We can also rewrite another result from the lecture on the weak law of large numbers

$$var[X^{(N)}] = \frac{popvar(\{X\})}{N}$$

$std[x^{(N)}]$

$= \sqrt{var[x^{(N)}]}$

- The standard deviation of the sample mean

$$std[X^{(N)}] = \frac{popsd(\{X\})}{\sqrt{N}}$$

- But we need the population standard deviation in order to calculate the $std[X^{(N)}]$!

Unbiased estimate of population standard deviation & Stderr

- The unbiased estimate of $popsd(\{X\})$ is defined as

$$stdunbiased(\{x\}) = \sqrt{\frac{1}{N-1} \sum_{x_i \in \ sample} (x_i - mean(\{x_i\}))^2}$$

- So the **standard error** is an estimate of

$$std[X^{(N)}] \quad \boxed{std[X^{(N)}]} \overset{?}{=} \frac{popsd(\{X\})}{\sqrt{N}} \ ?$$

approxi. of popsd

$$\frac{popsd(\{X\})}{\sqrt{N}} \doteq \frac{stdunbiased(\{x\})}{\sqrt{N}} = \boxed{stderr(\{x\})}$$

The reason to use the unbiased standard deviation(s) for popsd

Example 6.4-5 We have shown that when sampling from $N(\theta_1 = \mu, \theta_2 = \sigma^2)$, one finds that the maximum likelihood estimators of μ and σ^2 are

ch. 9

$$\widehat{\theta}_1 = \widehat{\mu} = \overline{X} \qquad \text{and} \qquad \widehat{\theta}_2 = \widehat{\sigma^2} = \frac{(n-1)S^2}{n}.$$

Recalling that the distribution of $\overline{X}$ is $N(\mu, \sigma^2/n)$, we see that $E(\overline{X}) = \mu$; thus, $\overline{X}$ is an unbiased estimator of μ.

In Theorem 5.5-2, we showed that the distribution of $(n-1)S^2/\sigma^2$ is $\boxed{\chi^2(n-1)}$. Hence,

$$E(S^2) = E\left[\frac{\sigma^2}{n-1}\,\frac{(n-1)S^2}{\sigma^2}\right] = \frac{\sigma^2}{n-1}(n-1) = \sigma^2.$$

That is, the sample variance

Hogg et. al.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an unbiased estimator of σ^2. Consequently, since

$$E(\widehat{\theta}_2) = \frac{n-1}{n}E(S^2) = \frac{n-1}{n}\sigma^2,$$

$\widehat{\theta}_2$ is a biased estimator of $\theta_2 = \sigma^2$.

* The notation might be different in this ref.

Standard error: election poll

		DATES	POLLSTER	SAMPLE	RESULT	NET RESULT
U.S. Senate	Miss.	NOV 25, 2018	C+ Change Research	1,211 LV	Espy 46% 51% Hyde-Smith	Hyde-Smith +5

51%

- What is the estimate of the percentage of votes for Hyde-smith? 51%

Sample mean value

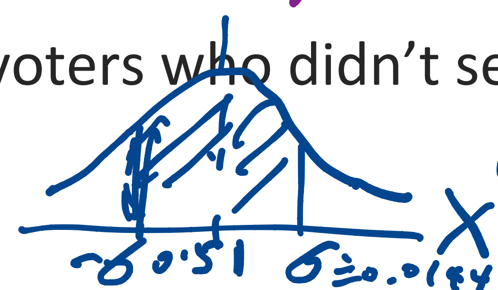
One sample value of $X^{(N)}$

$\sigma = std(X^{(N)})$

Number of sampled voters who selected Ms. Smith is:
1211(0.51) ≅ 618

$N = 1211$

Number of sampled voters who didn't selected Ms. Smith was
1211(0.49) ≅ 593



pop mean

Standard error: election poll

- $stdunbiased(\{x\})$

$$= \sqrt{\frac{1}{1211-1}(618(1-0.51)^2 + 593(0-0.51)^2)} = 0.5001001$$

- $stderr(\{x\})$

$$= \frac{0.5}{\sqrt{1211}} \simeq 0.0144$$

$VX_1 = \begin{cases} 1 & \text{vote for Hyde-Smith} \\ 0 & \text{No} \end{cases}$

$\vdots$

VX_{1211}

618 "yes"

593 "No"

$N = 1211$

Interpreting the standard error

- **Sample mean** is a random variable and has its own probability distribution, stderr is an estimate of the sample mean's standard deviation
- When ***N*** is very large, according to the **Central Limit Theorem**, sample mean is approaching a normal distribution with

$$\boxed{\mu \doteq \text{mean}(\{x\}); \quad \sigma \doteq \text{stderr} = \frac{\text{stdunbi}(\{x\})}{\sqrt{N}}}$$

$$E[X^{(N)}] = E[x^{(1)}] = \text{popmean}(\{X\})$$

Interpreting the standard error

- **Sample mean** is a random variable and has its own probability distribution, stderr is an estimate of sample mean's standard deviation

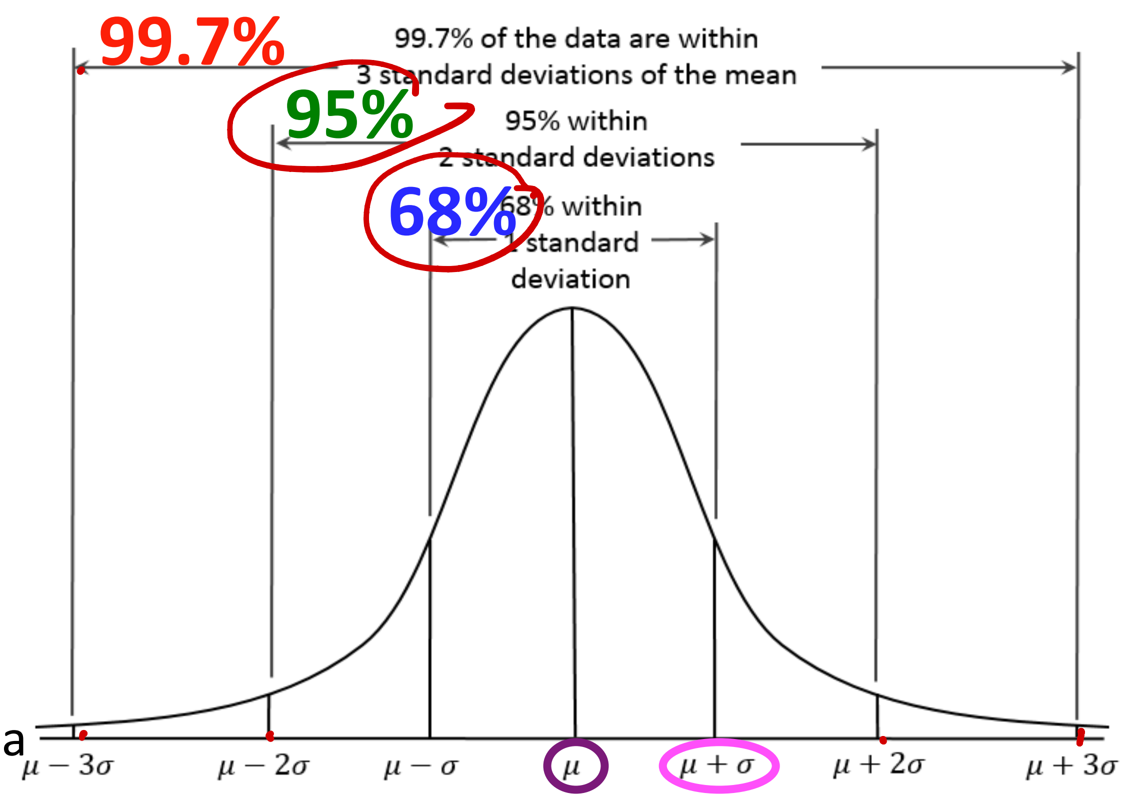
- When ***N*** is very large, according to the **Central Limit Theorem**, sample mean is approaching a normal distribution with

$$\mu = popmean(\{X\}) \;\; ; \;\; \sigma = \frac{popsd(\{X\})}{\sqrt{N}} \doteq stderr(\{x\})$$

$$stderr(\{x\}) = \frac{stdunbiased(\{x\})}{\sqrt{N}}$$

Interpreting the standard error

Probability distribution of sample mean tends **normal when N is large**



Credit: wikipedia

$\doteq \text{mean}(\{x\})$

Population mean

$\sigma \doteq \text{stderr} = \frac{\text{stdunb}}{\sqrt{N}}$

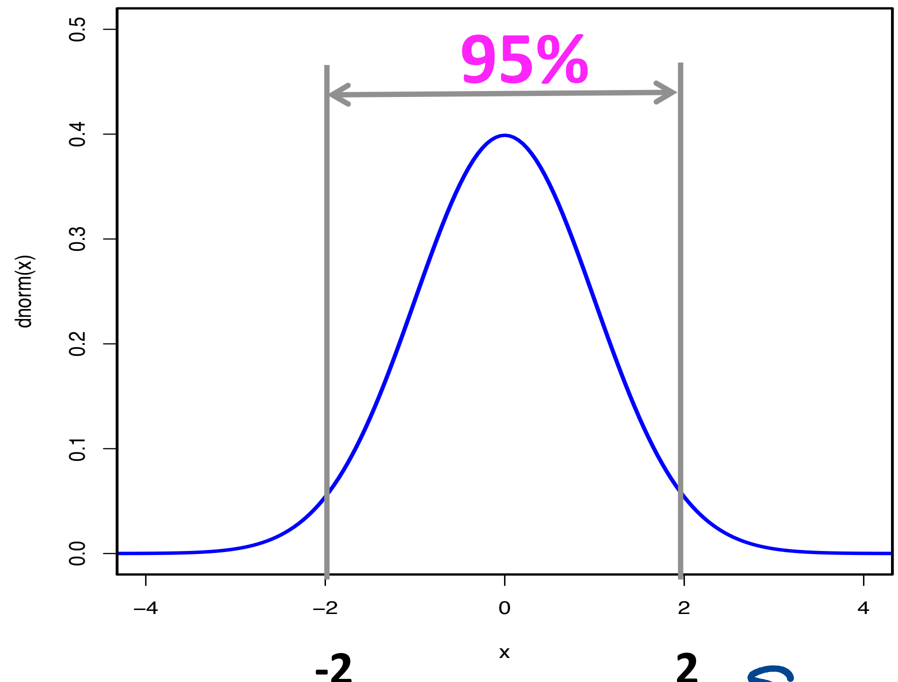
$\underline{x^{(N)}}$

values

Confidence intervals

- Confidence interval for a population mean is defined by fraction

- Given a percentage, find how many units of strerr it covers.



95% of $X^{(N)}$ values

For **95%** of the **realized sample means**,

the population mean lies in

realized value

[sample mean-2 stderr, sample mean+2 stderr]

the mean({x}) value

Confidence intervals when N is large

- For about 68% of realized sample means

$$mean(\{x\}) - stderr(\{x\}) \leq popmean(\{X\}) \leq mean(\{x\}) + stderr(\{x\})$$

- For about 95% of realized sample means

$$mean(\{x\}) - 2stderr(\{x\}) \leq popmean(\{X\}) \leq mean(\{x\}) + 2stderr(\{x\})$$

- For about 99.7% of realized sample means

$$mean(\{x\}) - 3stderr(\{x\}) \leq popmean(\{X\}) \leq mean(\{x\}) + 3stderr(\{x\})$$

Q. Confidence intervals

- What is the 68% confidence interval for a population mean?

A. [sample mean-2stderr, sample mean+2stderr]
B. [sample mean-stderr, sample mean+stderr]
C. [sample mean-std, sample mean+std]

Standard error: election poll

		DATES	POLLSTER	SAMPLE	RESULT	NET RESULT
U.S. Senate	Miss.	NOV 25, 2018	C+ Change Research	1,211 LV	Espy 46% 51% Hyde-Smith	Hyde-Smith +5

$X^{(N)}$ here is $X^{(1211)}$

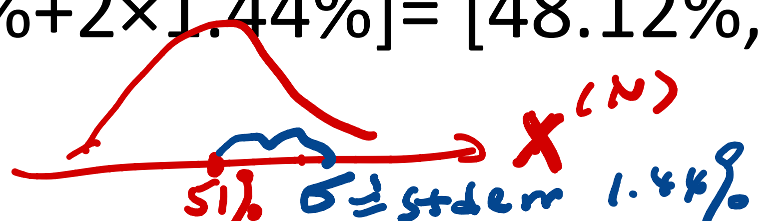
51%

- We estimate the population mean as 51% with stderr 1.44%

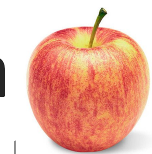
→ approx. of $std(X^{(N)})$

$mean\{x\}$, $\{x\}$ has $N=1211$

- The 95% confidence interval is [51%-2×1.44%, 51%+2×1.44%]= [48.12%, 53.88%]



Q.

- A store staff mixed their fuji  and gala  apples and they were individually wrapped, so they are indistinguishable. if I pick 30 apples and found 21 fuji , what is my 95% confidence interval to estimate the popmean is 70% for fuji? (hint: strerr > 0.05)

A. [0.7-0.17, 0.7+0.17]

B. [0.7-0.056, 0.7+0.056]

What if N is small? When is N large enough?

- If samples are taken from normal distributed population, the following variable is a random variable whose distribution is Student's **t**-distribution with ***N*-1** degree of freedom.

N = sample size M → random sample {x} from

$$T = \frac{mean(\{x\}) - popmean(\{X\})}{stderr(\{x\})}$$

eg. 12.1.1 E[mean({x})] R

$= \frac{1}{N} \cdot N \cdot E[x^{(i)}]$

$= popmean$

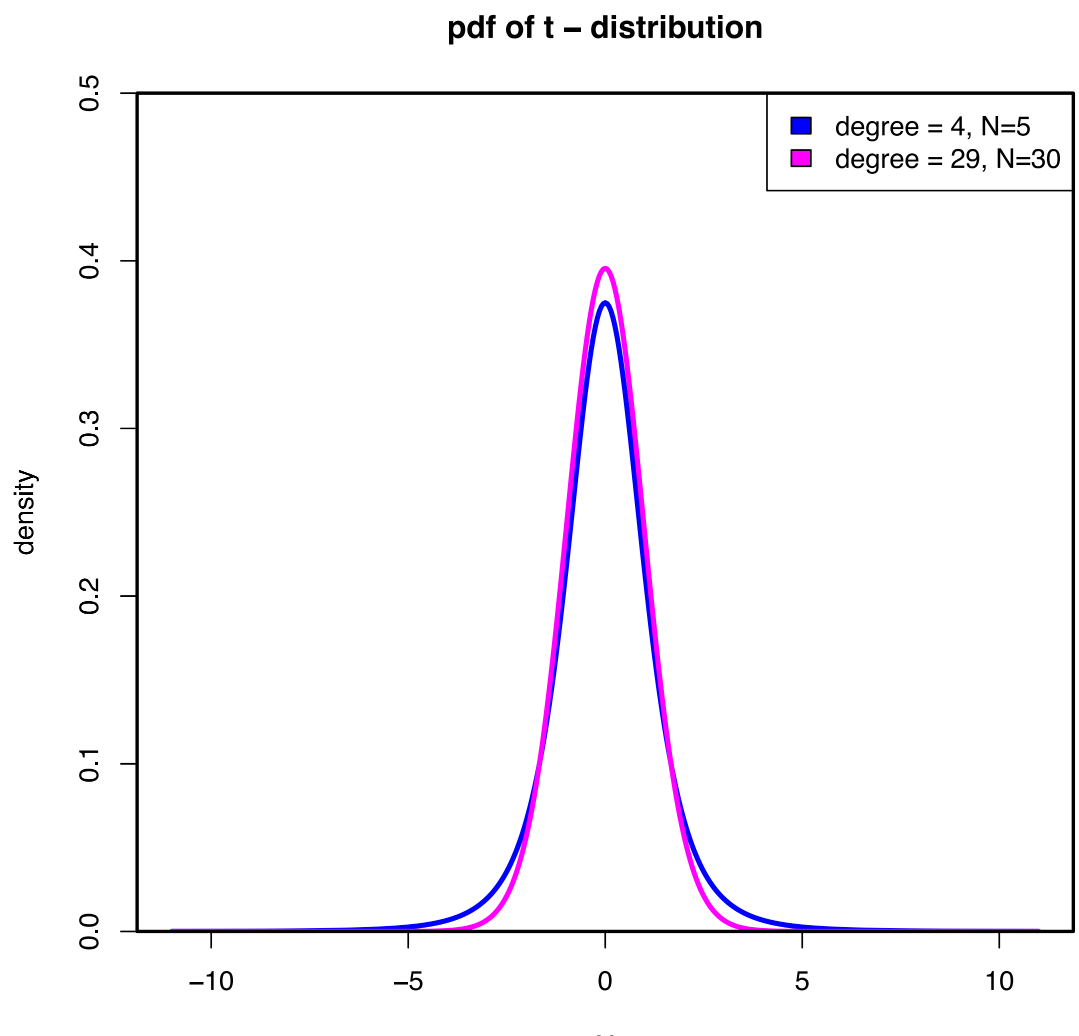
Degree of freedom is ***N***-1 due to this constraint: $\sum_i (x_i - mean(\{x\})) = 0$

t-distribution is a family of distri. with different degrees of freedom

t-distribution with N=5 and N=30

Credit : wikipedia

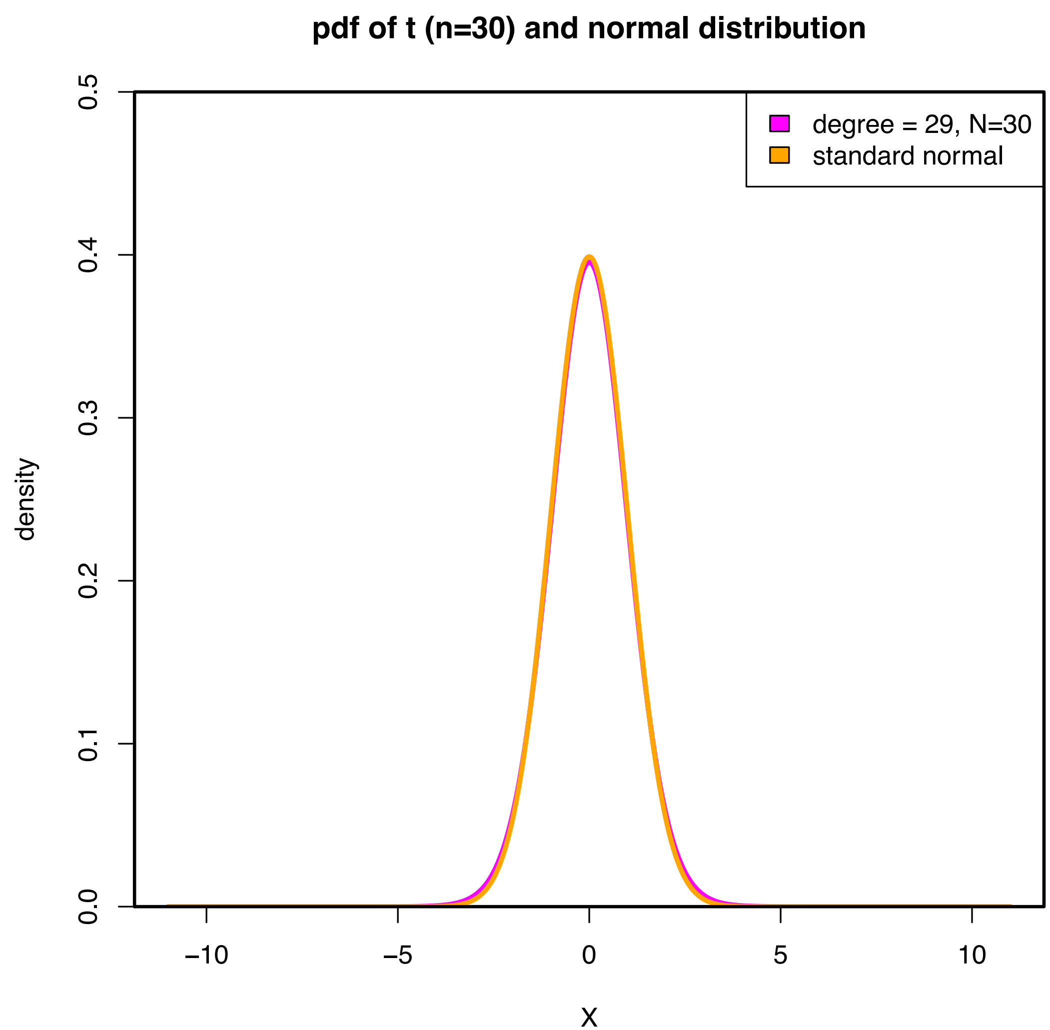
William Sealy Gosset 1876-1937



When N=30, t-distribution is almost Normal

t-distribution looks very similar to normal when N=30.

So **N=30 is a rule of thumb to decide N is large or not**



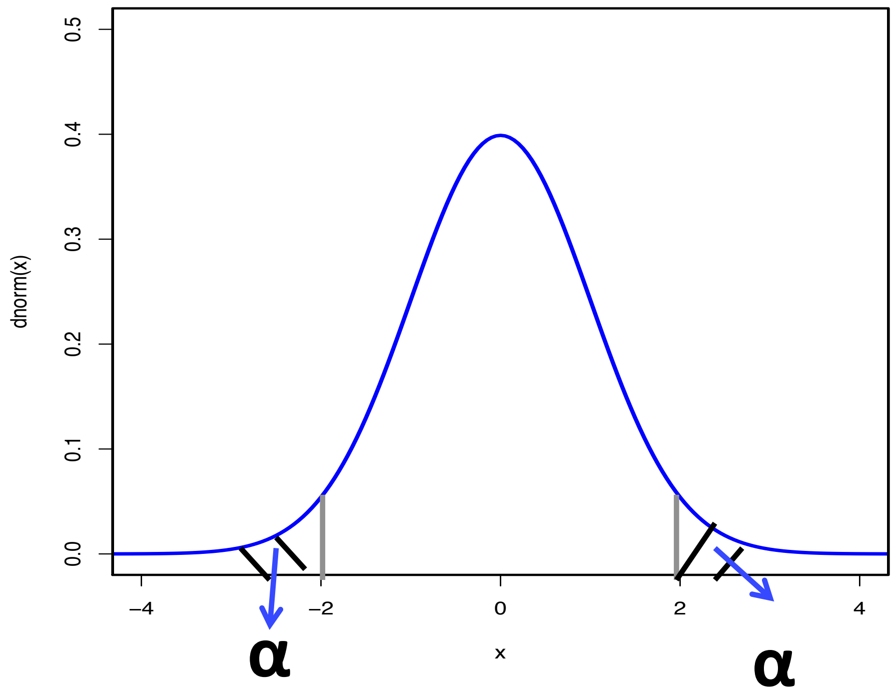
Confidence intervals when N< 30

- If the sample size N< 30, we should use t-distribution with its parameter (the degrees of freedom) set to N-1

Centered Confidence intervals

- Centered Confidence interval for a population mean by **α** value, where

$$P(T \geq b) = \alpha$$

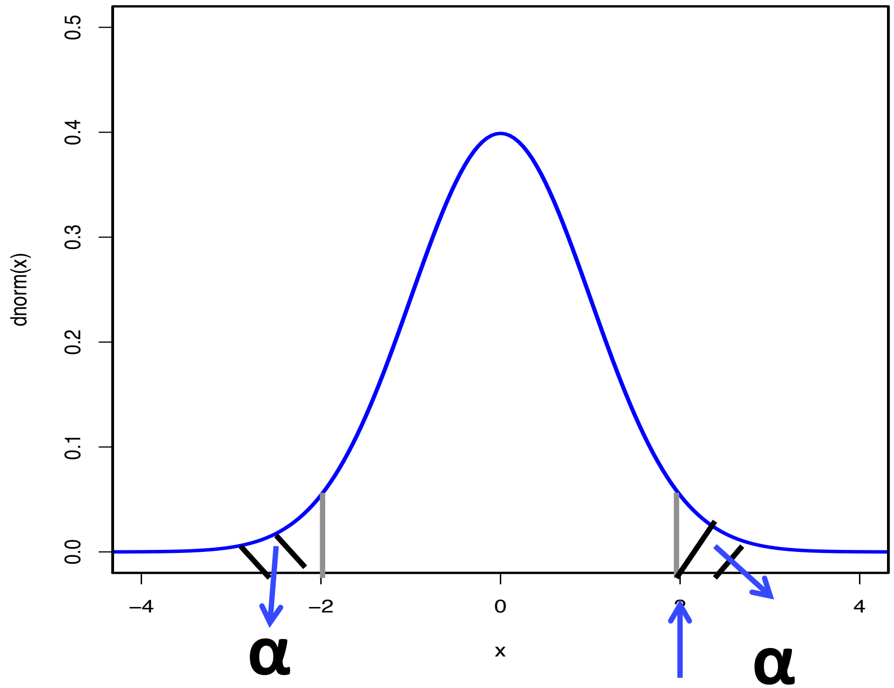


For **1-2α** of the realized sample means,
the population mean lies in
[sample mean-**b**×stderr, sample mean+**b**×stderr]

Centered Confidence intervals

- Centered Confidence interval for a population mean by **α** value, where

$$P(T \geq b) = \alpha$$



For 1-2α of the realized sample means,
the population mean lies in
[sample mean-**b**×stderr, sample mean+**b**×stderr]

Q.

- The 95% confidence interval for a population mean is equivalent to what 1-2α interval?

A. α= 0.05

B. α= 0.025

C. α= 0.1

Assignments

- Read Chapter 7 of the textbook
- Next time: Bootstrap, Hypothesis tests

Additional References

- Charles M. Grinstead and J. Laurie Snell "Introduction to Probability"
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics"

* Hogg et al. "Probability and Statistical Inference"

See you next time

See you!