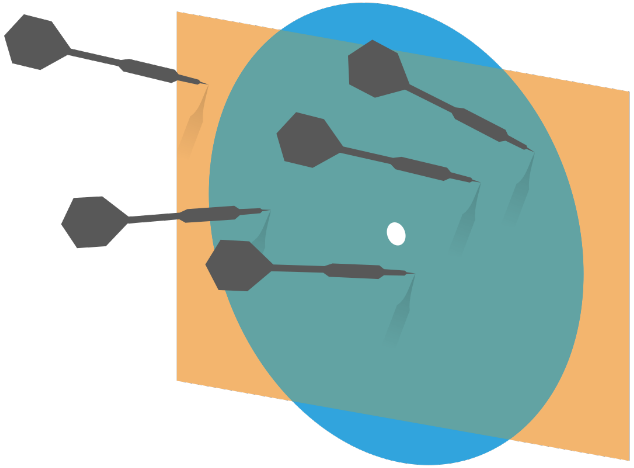
Probability and Statistics for Computer Science



Credit: wikipedia

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

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 10.06.2021

Last time

- Exponential distribution
- Normal (Gaussian) distribution

Objectives

- Sample mean
- confidence interval
- t-distribution

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.

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

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

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

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

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.

Standard deviation of the sample mean

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

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

- 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

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

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

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

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%

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

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

Standard error: election poll

- $stdunbiased(\{x\})$

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

- $stderr(\{x\})$

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

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

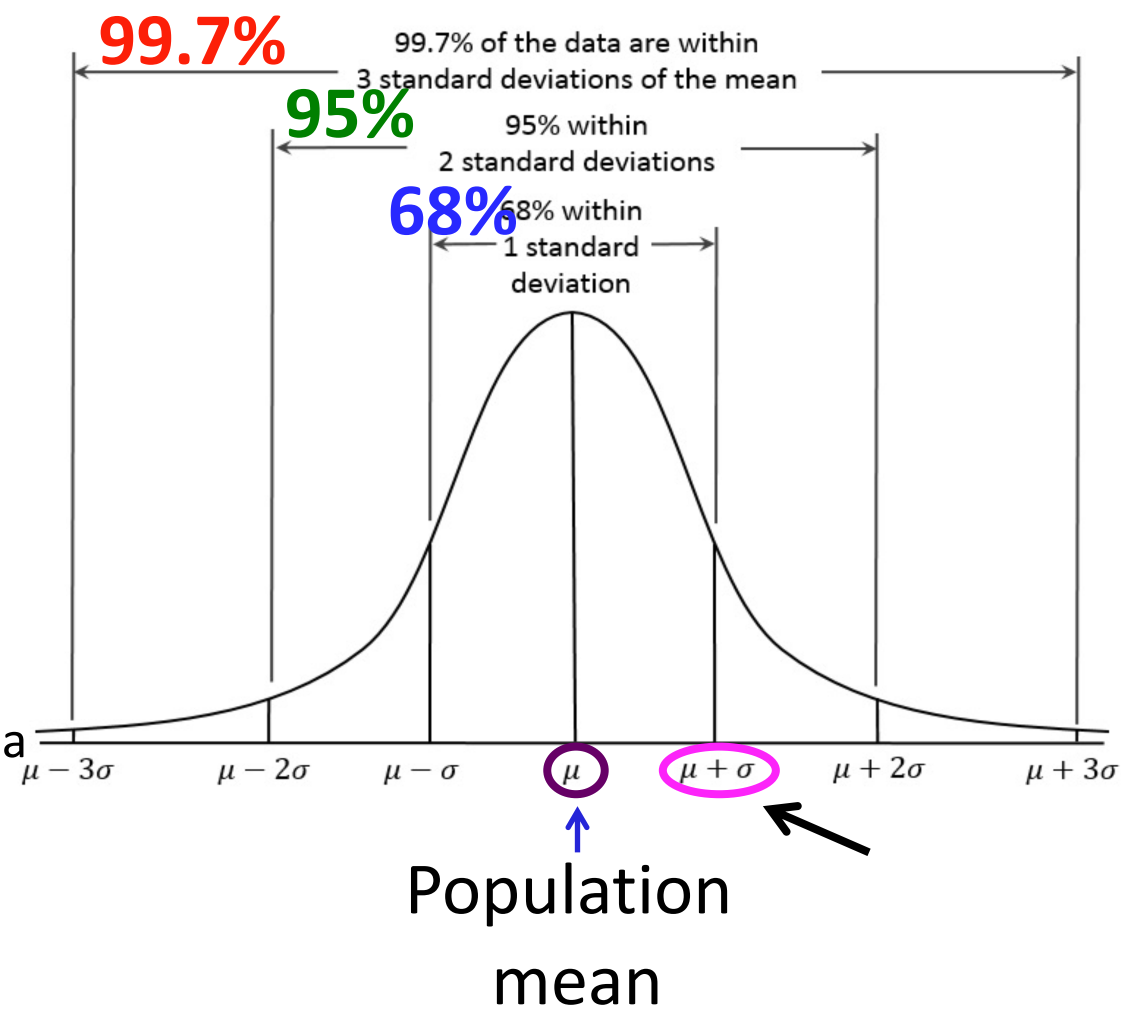
- When ***N*** is very large, according to the **Central Limit Theorem**, sample mean is approaching a normal distribution with

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

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

Interpreting the standard error

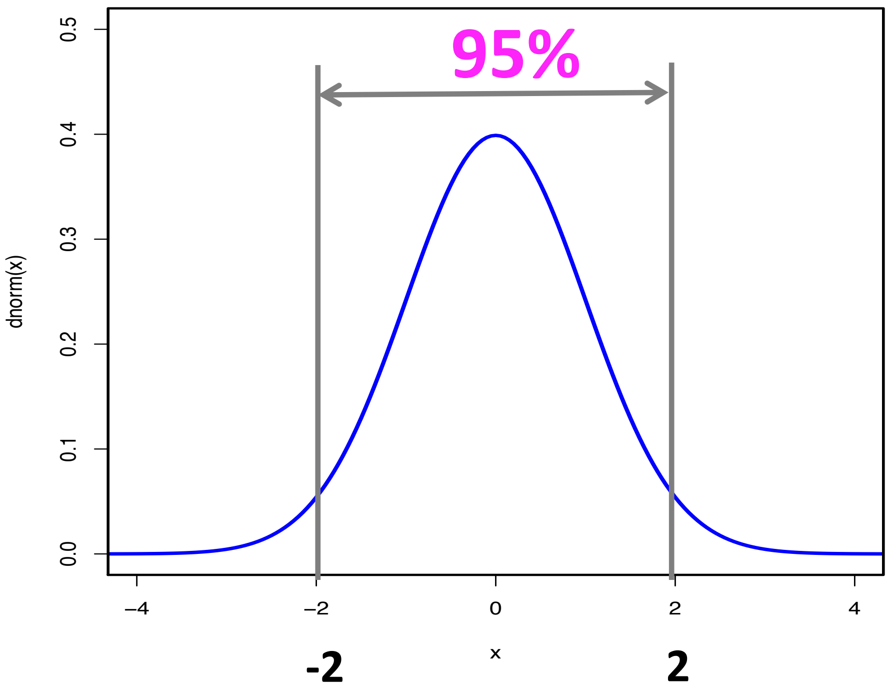
Probability distribution of sample mean tends **normal when N is large**



Credit: wikipedia

Confidence intervals

- Confidence interval for a population mean is defined by fraction
- Given a percentage, find how many units of strerr it covers.



For **95%** of the **realized sample means**,
the population mean lies in
[sample mean-2 stderr, sample mean+2 stderr]

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

51%

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

- ✺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.

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

Degree of freedom is ***N***-1 due to this constraint:

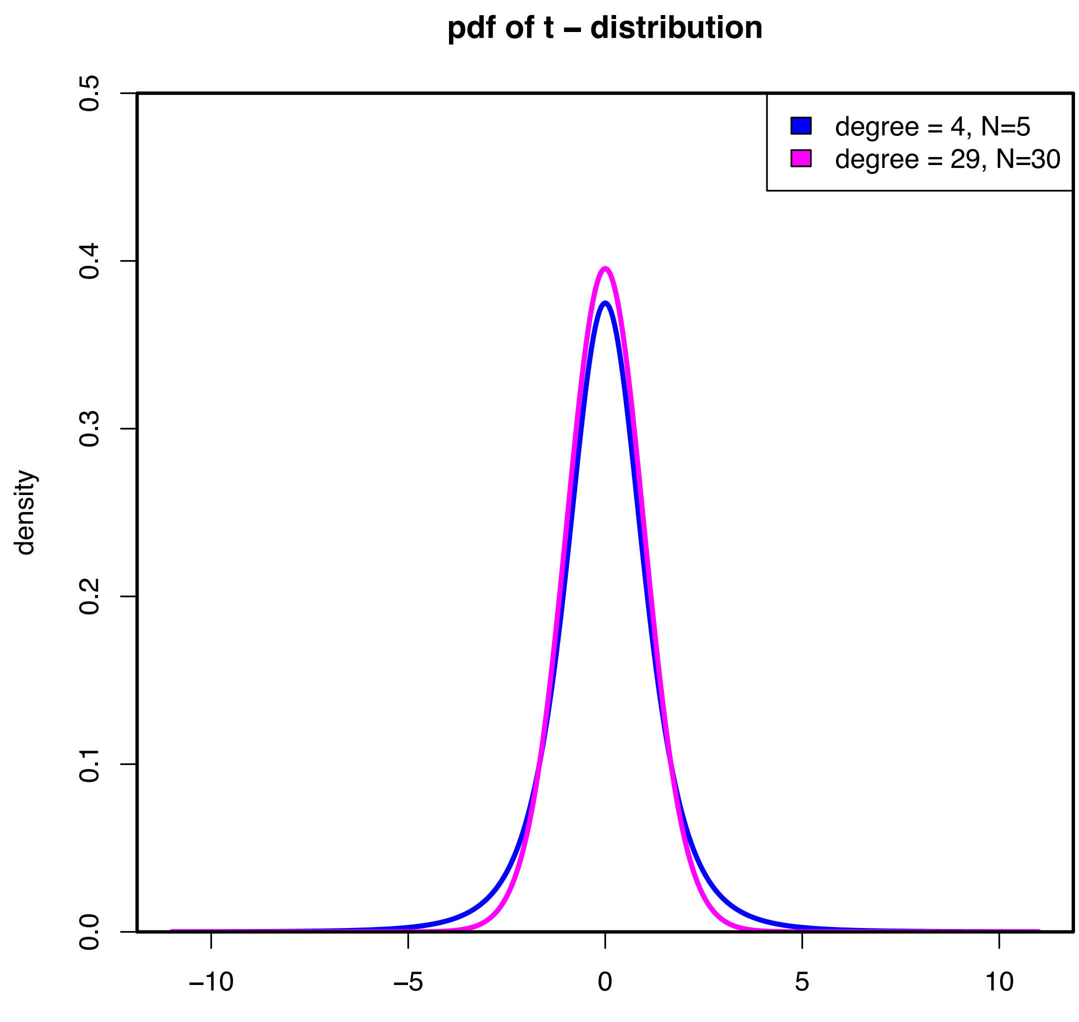
$$\sum_i (x_i - mean(\{x\})) = 0$$

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

t-distribution with N=5 and N=30

Credit : wikipedia

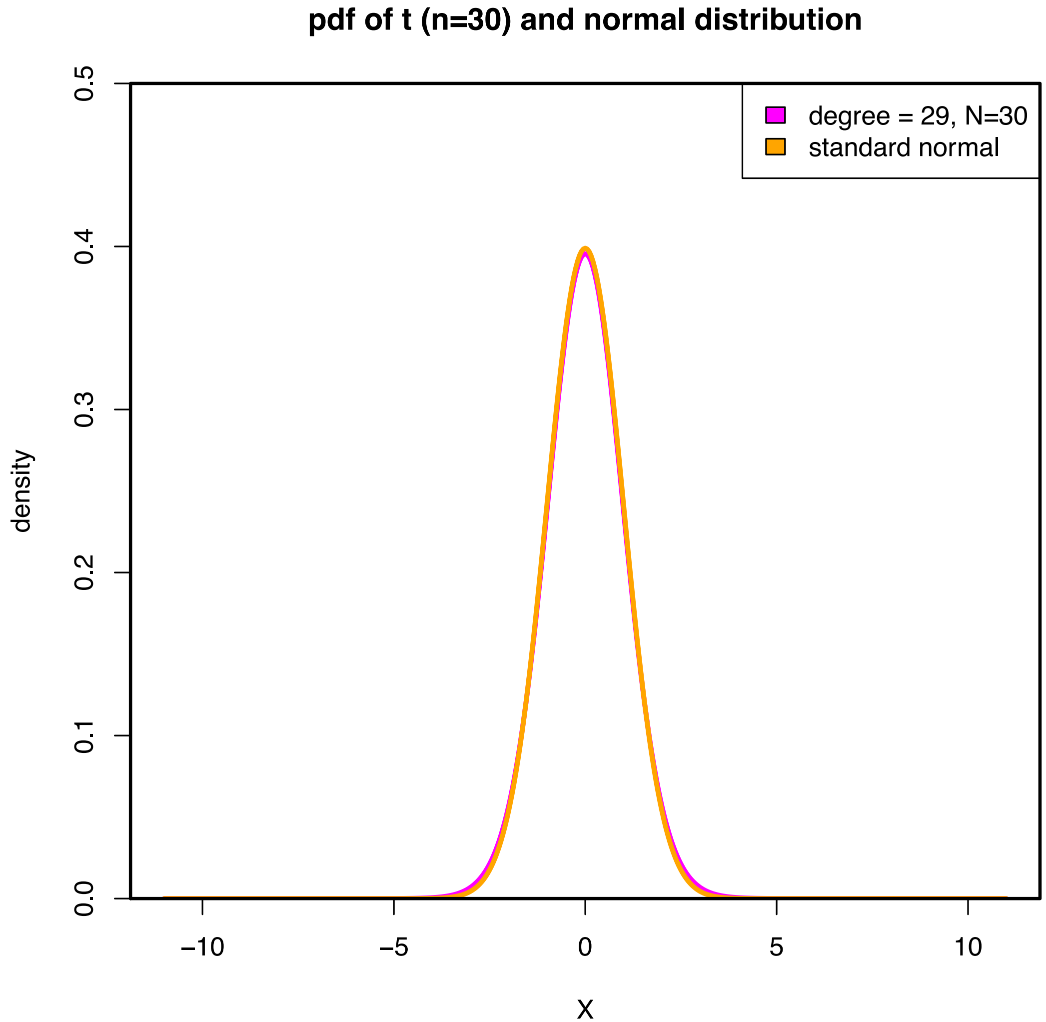
William Sealy Gosset 1876-1937



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**



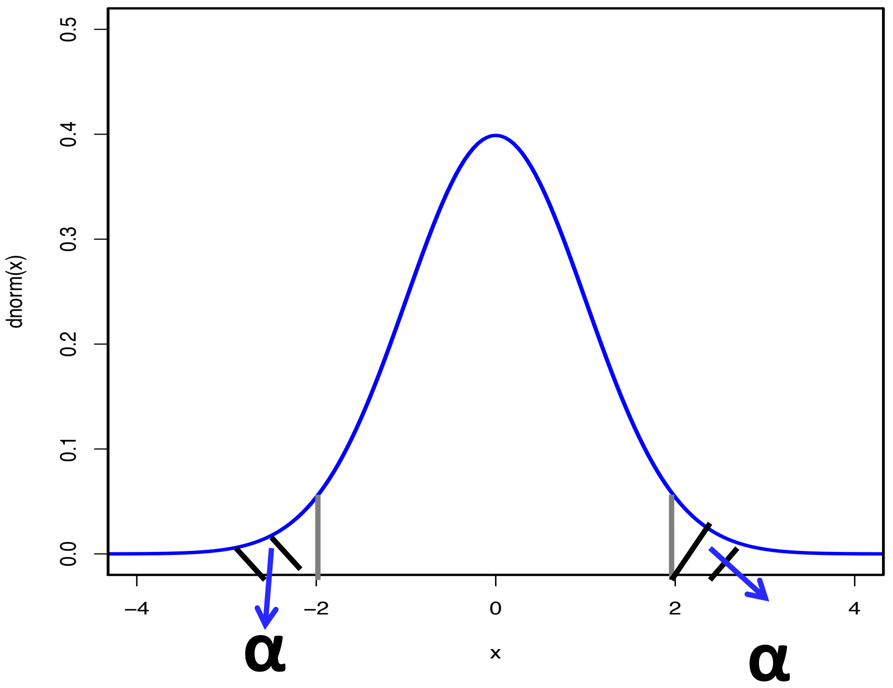
Confidence intervals when N< 30

- 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

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

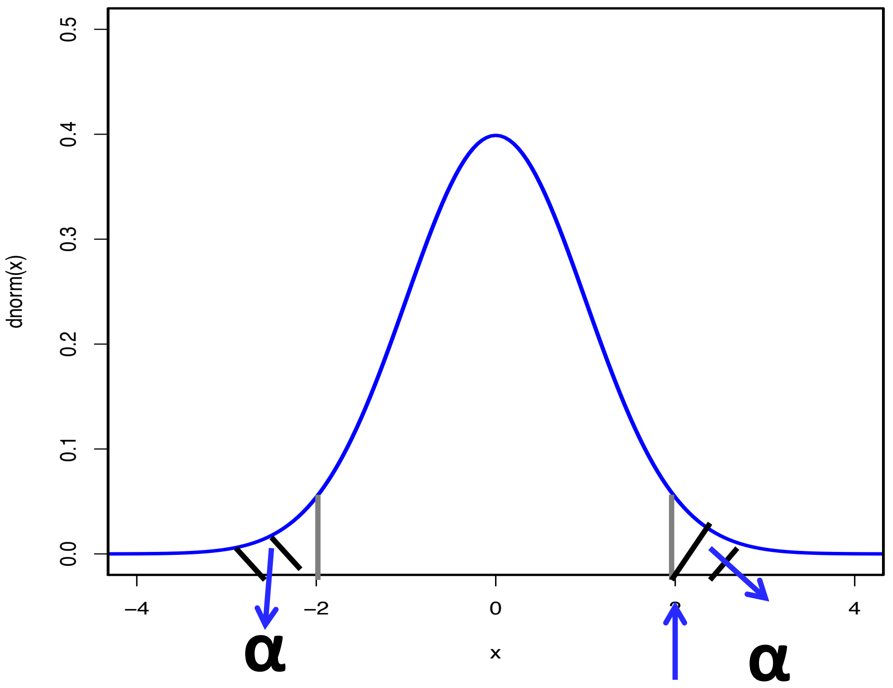


For **1-2α** of the realized sample means,
the population mean lies in
[sample mean-**b**×stderr, sample mean+**b**×stderr]

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 $1-2\alpha$ interval?

A. $\alpha = 0.05$

B. $\alpha = 0.025$

C. $\alpha = 0.1$

Assignments

- Read Chapter 7 of the textbook
- Next time: Bootstrap, Hypothesis tests
- Prepare for Midterm1

Additional References

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

See you next time

See you!