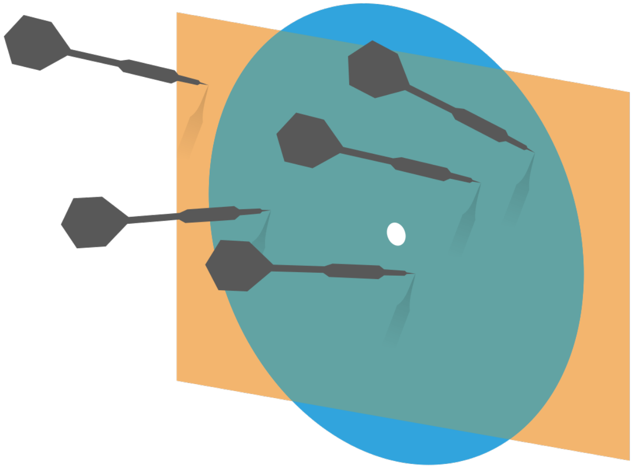
Probability and Statistics for Computer Science



Credit: wikipedia

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

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 10.13.2021

Last time

- Sample mean
- Confidence interval
- t-distribution (I)

Objectives

- Review Sample mean, CI
- t-distribution (II)
- Bootstrap simulation

Review with Questions

1) Why is sample mean $X^{(N)}$ a random variable?

2) Is $E[X^{(N)}] = \text{mean}(\{x\})$? $\{x\}$ is some realized data of size N, drawn from the population $\{X\}$ with replacement.

$\downarrow$ popmean

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

3) What is the distribution of $X^{(N)}$?

4) What are $E[X^{(N)}]$, $\text{var}[X^{(N)}]$?

About the distribution of $X^{(N)}$

if $N \to \infty$ $\quad X^{(N)} \sim \text{Normal}(\mu, \sigma)$ $\quad \text{mean}(\{x\}) \approx$

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

$\sigma = \text{std}[X^{(N)}] = \dfrac{\text{popstd}}{\sqrt{N}}$

---

if $X^{(1)}$ is from a Normal like population,

$$T = \frac{\text{mean}(\{x\}) - \text{popmean}}{\text{stderr}(\{x\})} \sim t \text{ distribution with DOF } N-1$$

$X^{(N)}$

i.i.d. $X^{(1)}$

A tale of two statisticians

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

The task: use only a subset of $\{X\}$: $\{x\}$ with $N = 5$ to estimate the popmean($\{X\}$) with some confidence report.

A tale of two statisticians

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

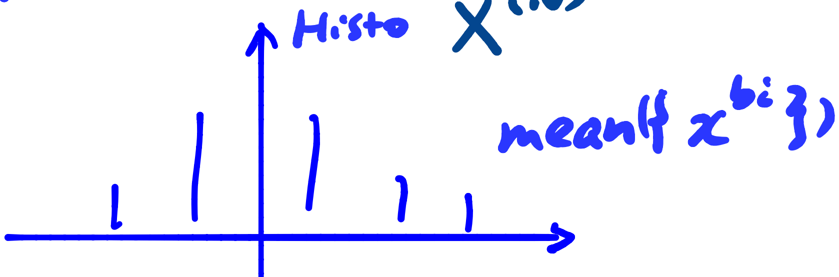
$\{x^{b}\} = \{1, 4, 5, 7, 11\}$

$\{x^{b1}\} = \{1, 1, 4, 5, 7\}$ $\quad x^{b1(n)} = \frac{18}{5}$

$\{x^{b2}\} = \{4, 5, 7, 7, 1\}$

$\vdots$

$\{x^{bn}\} = \{5, 5, 5, 5, 5\}$



i.i.d $\quad X^{(1)}$

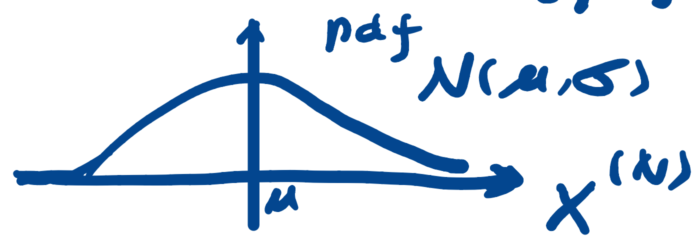
$\{x\} = \{1, 4, 5, 7, 11\}$ $\quad N = 5$

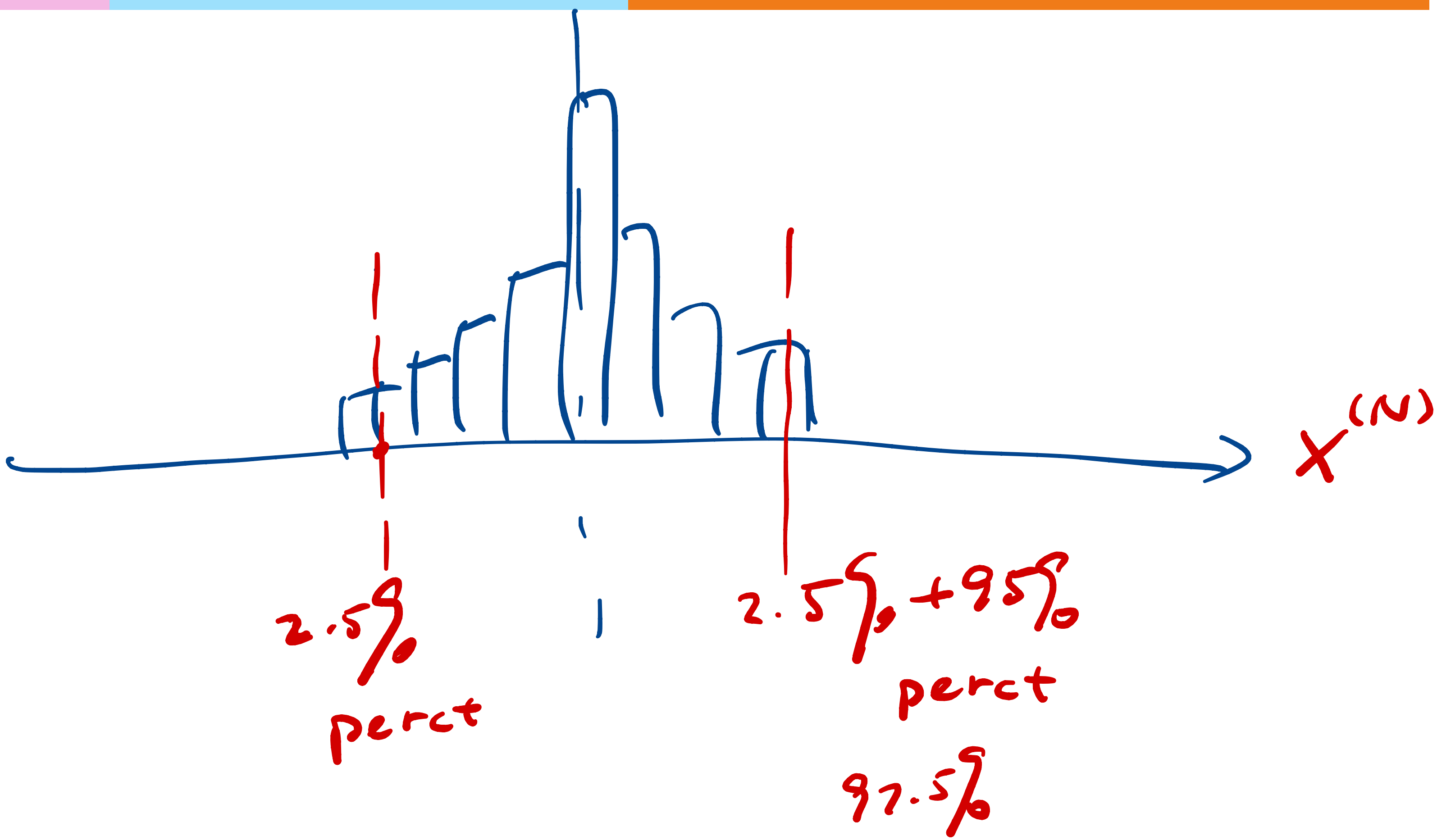
if $N \to \infty$

$X^{(N)} \sim N(\mu, \sigma)$

$\mu = E[X^{(N)}] \doteq \mathrm{mean}(\{x\})$

$\sigma = \mathrm{std}[X^{(N)}] \doteq \mathrm{stderr}(\{x\})$





$X^{(N)}$
2.5%
perct
2.5% + 95%
perct
97.5%

review on your own

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%

$X^{(1211)}$ takes one value of mean $(\{x\})$ from $\{x\}$ of size 1211

$\{1, 0, 1, 0, 0, \ldots\}$

- What is the estimate of the percentage of votes for Hyde-smith?
- How confident is that estimate?

review on your own

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.

$E[X^{(N)}] \approx mean(\{x\})$

realized sample

review on your own

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

review on your own

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

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

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

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

review on your own

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%

$X = \begin{cases} 1 & \text{Smith} \\ 0 & \text{Not Smith} \end{cases}$

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

review on your own

Standard error: election poll

- $stdunbiased(\{x\})$

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

- $stderr(\{x\})$

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

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

$N = 1211$

review on your own

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

$X^{(N)} \sim \mathcal{N}(\mu, \sigma)$

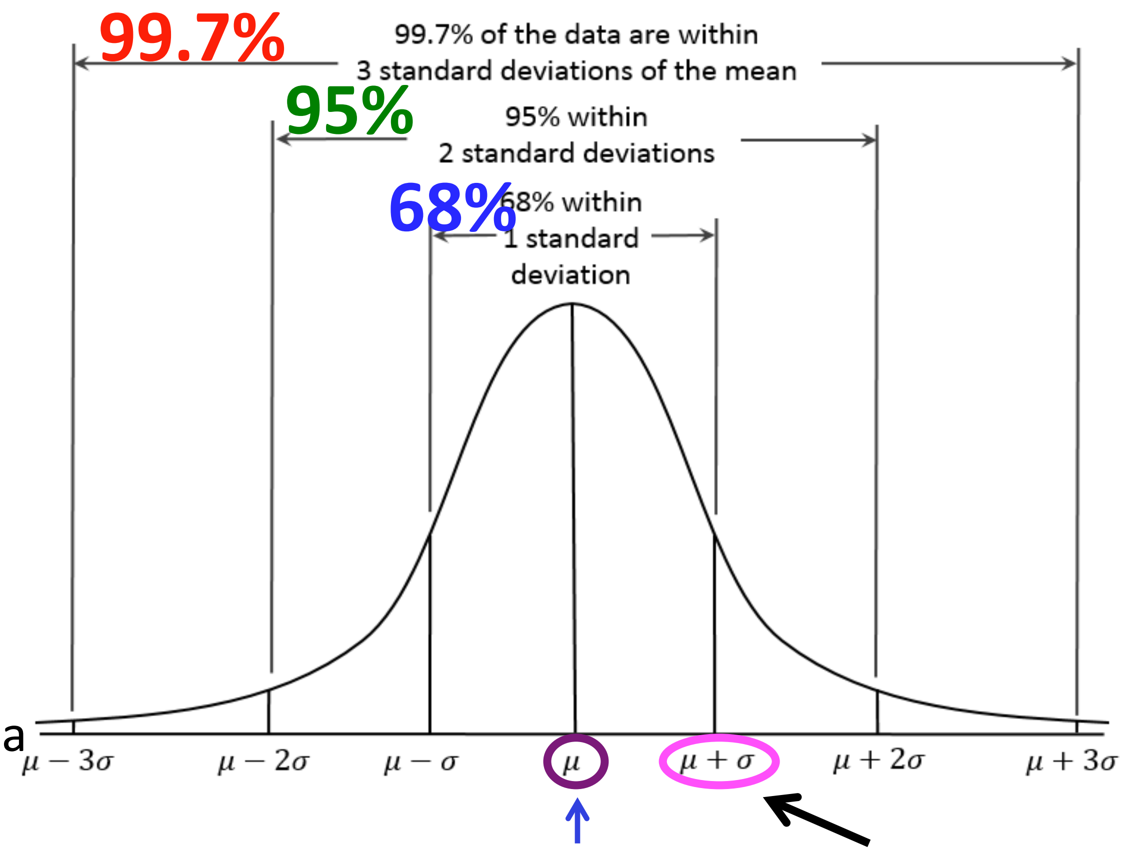
$$\boxed{\mu = \underbrace{popmean(\{X\})}_{\approx mean(\{x\})} \; ; \; \sigma = \frac{popsd(\{X\})}{\sqrt{N}} \overset{\approx}{\doteq} stderr(\{x\})}$$

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

review on your own

Interpreting the standard error

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



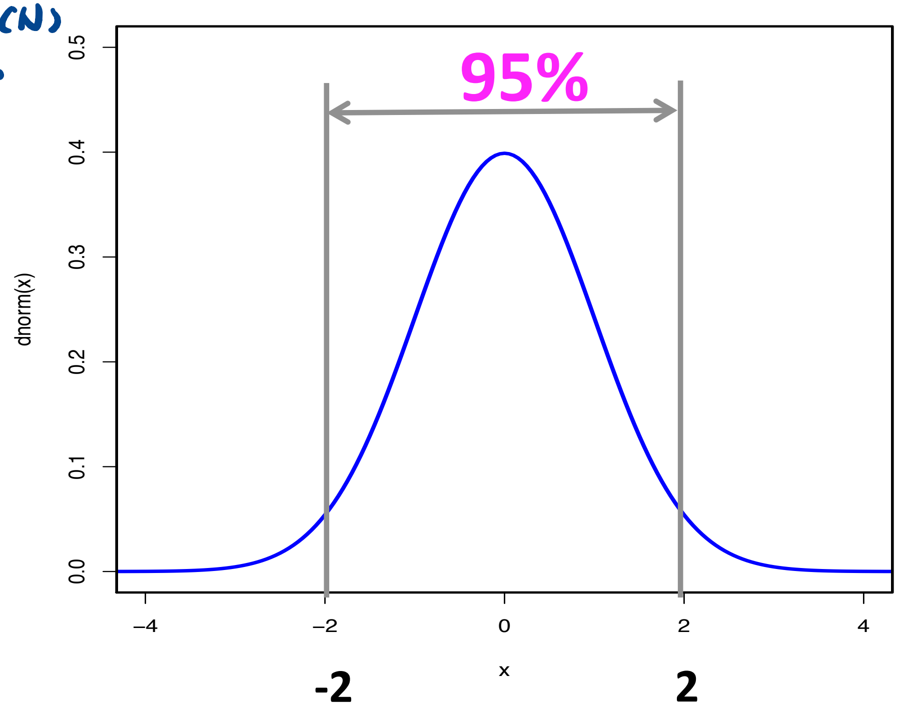
Credit: wikipedia

Population mean

μ+Standard error

Confidence intervals

- Confidence interval pdf of $\hat{X}(N)$ 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]
↳ one value of the specific sample

review on your own

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

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

review on your own

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]

review on your own

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%]

review on your own

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]

review on your own

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.

$X^{(N)}$

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

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

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

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

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

Degroot Pg. 480

PDF of t-distri. of m degree

$$\frac{\Gamma\left(\frac{m+1}{2}\right)}{(m\pi)^{1/2}\,\Gamma\left(\frac{m}{2}\right)}\left(1+\frac{x^2}{m}\right)^{-(m+1)/2} \qquad -\infty < x < \infty.$$

mean doesn't exist for $m \le 1$

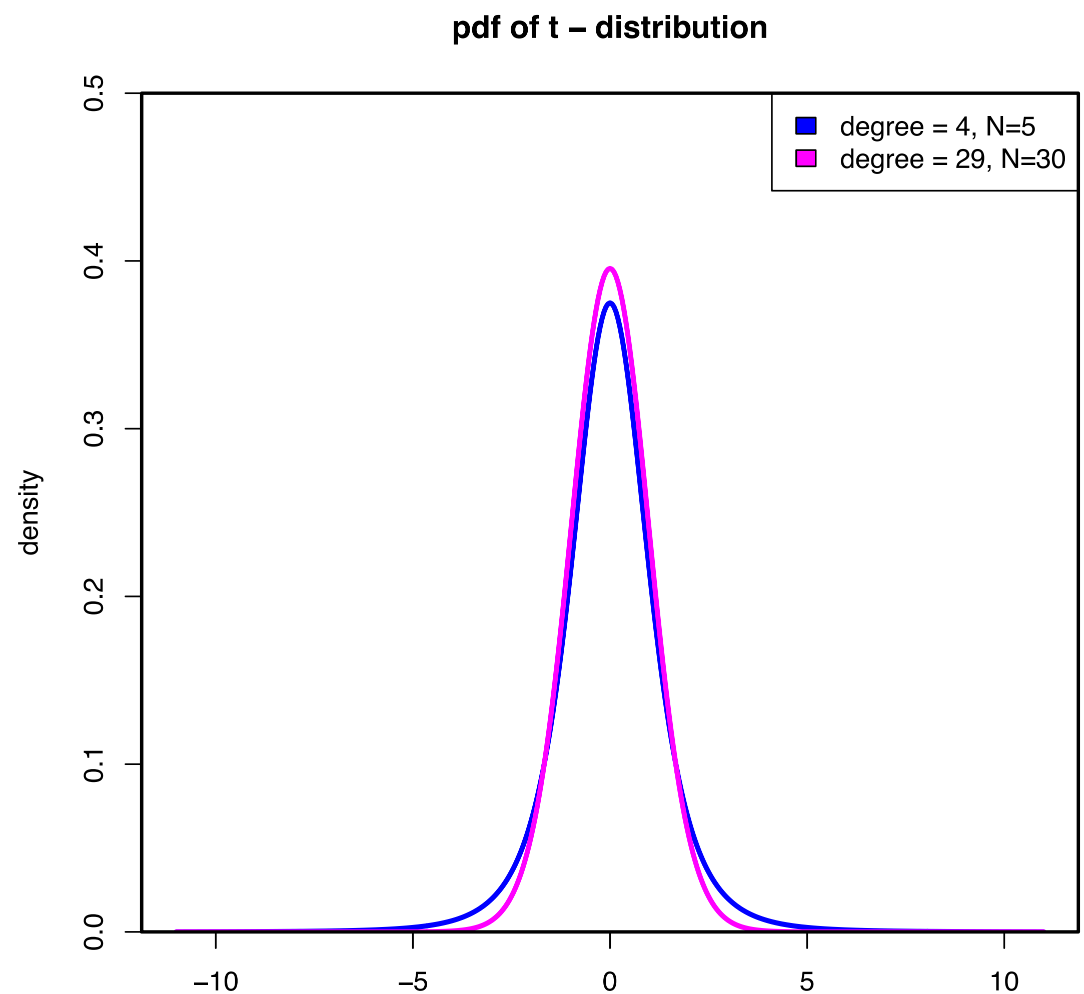
review on your own

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

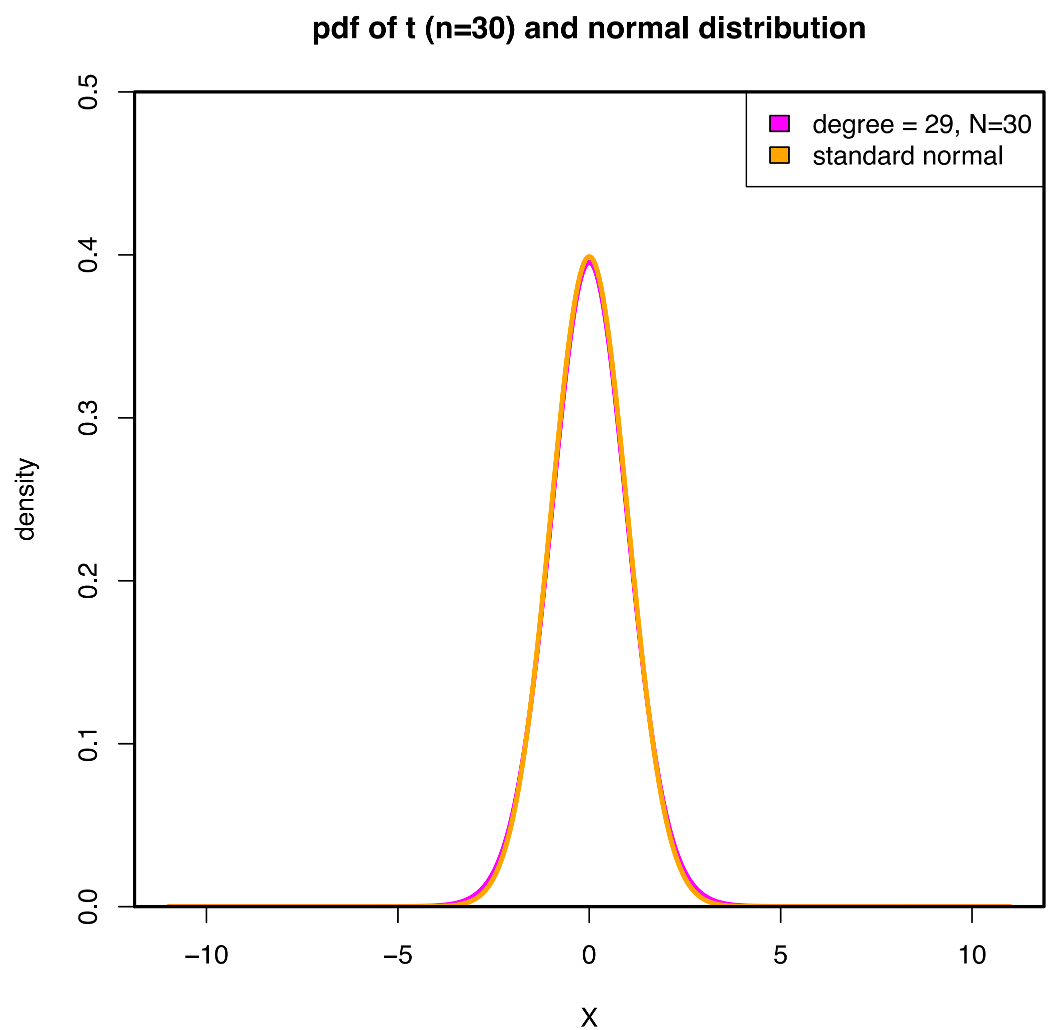


review on your own

When N=30, t-distribution is almost Normal

t-distribution looks very similar to standard 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

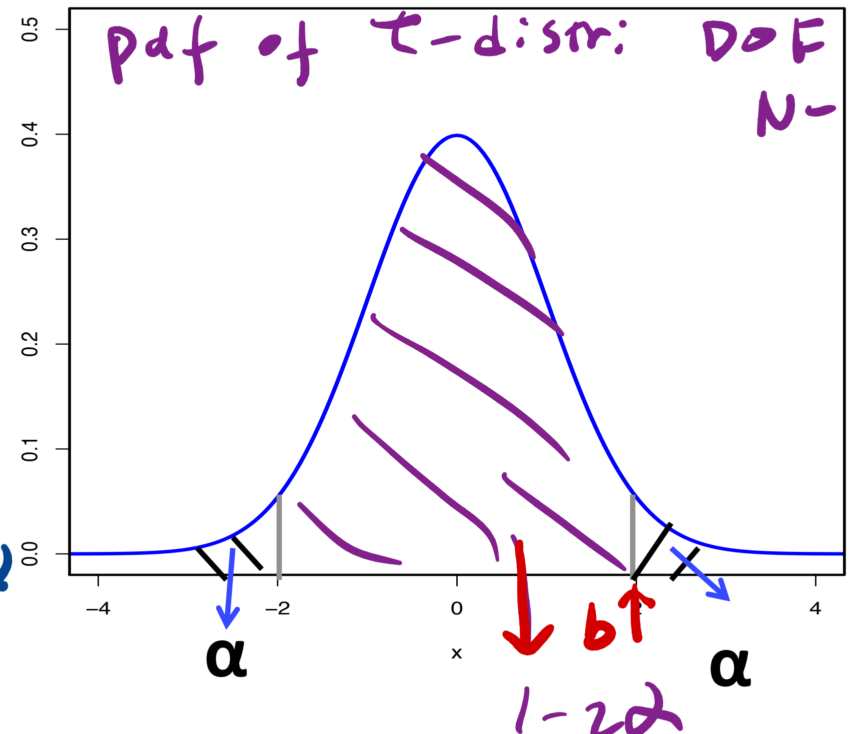
t- distr: is also symmetric.

Centered Confidence intervals

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

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

$$P\left(\left|\frac{\text{mean}\{x\} - \text{popmean}}{\text{stderr}}\right| \geq b\right) = \alpha$$



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

2α Confidence Interval

$$P\left(\frac{\text{mean}(\{x\}) - \text{popmean}}{\text{stderr}(\{x\})} \geq b\right)$$

$$= P(\text{popmean} \leq \text{mean}(\{x\}) - b \cdot \text{stderr}(\{x\}))$$
$$= \alpha$$

$$P\left(\frac{\text{mean}(\{x\}) - \text{popmean}}{\text{stderr}(\{x\})} \leq -b\right)$$

$$= P(\text{popmean} \geq \text{mean}(\{x\}) + b \cdot \text{stderr}(\{x\}))$$
$$= \alpha$$

$\alpha = 5\% \rightarrow 1 - 2\alpha = 90\%$

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

$1 - 2\alpha = 95\%$

Sample statistic

- A **statistic** is a function of a dataset
  - For example, the mean or median of a dataset is a statistic
- **Sample statistic**
  - Is a statistic of the data set that is formed by the realized sample
  - For example, the realized sample mean

Q. Is this a sample statistic?

- The largest integer that is smaller than or equal to the mean of a sample

  A. Yes

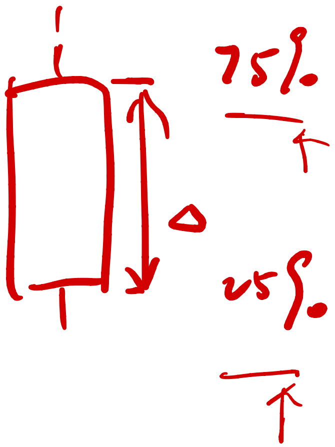
  B. No.

Q. Is this a sample statistic?

- The interquartile range of a sample

  A. Yes

  B. No.



Confidence intervals for other sample statistics

- **Sample statistic** such as ***median*** and others are also interesting for drawing conclusion about the population
- It's often difficult to derive the analytical expression in terms of stderr for the corresponding random variable
- So we can use simulation…

Bootstrap for confidence interval of other sample statistics

- Bootstrap is a method to construct confidence interval for *any** sample **statistics** using resampling of the sample data set

- Bootstrapping is essentially uniform random sampling with replacement on the sample of size ***N***

Bootstrap for confidence interval of other sample statistics

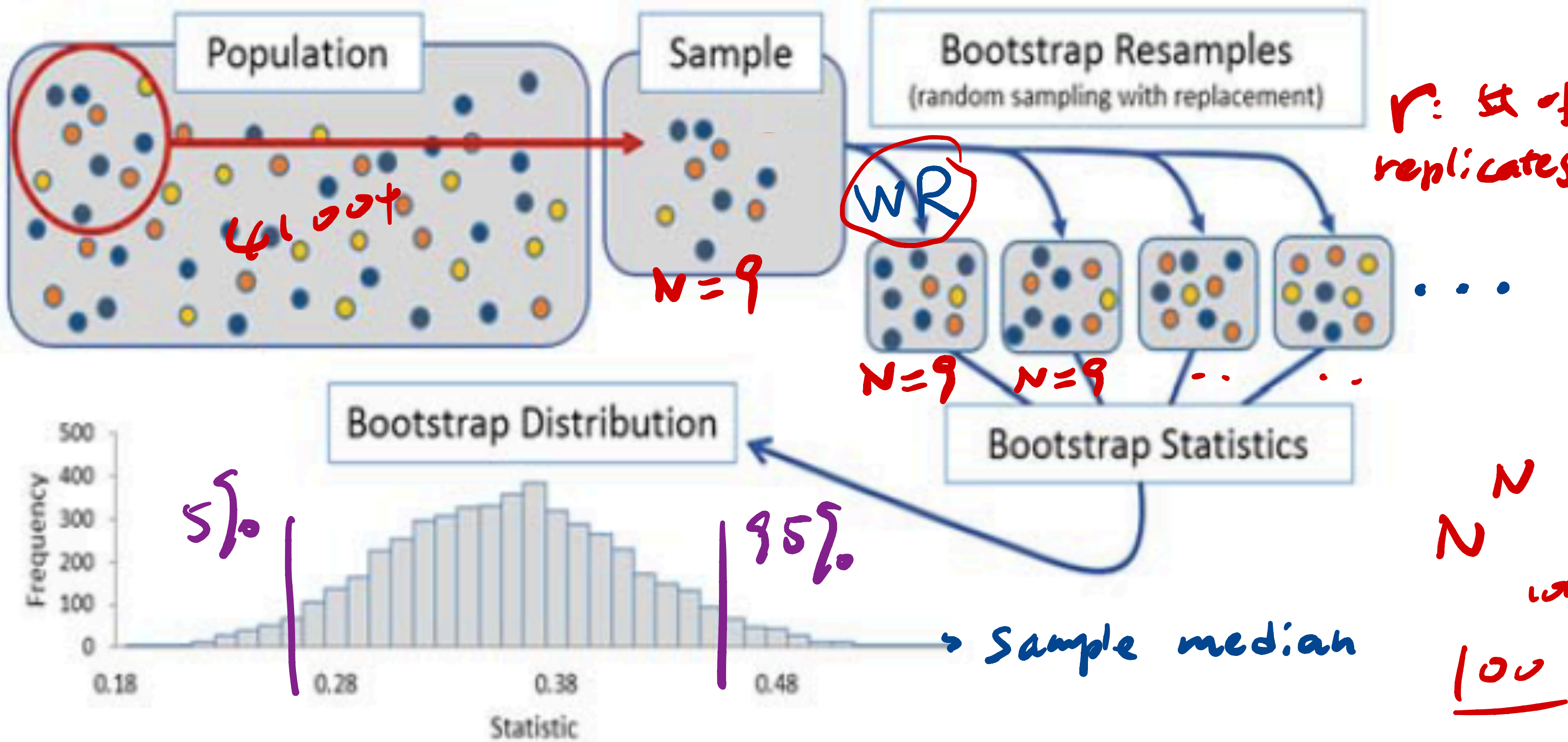


Figure 1. Summary of Bootstrapping Process

Credit: E S. Banjanovic and J. W. Osborne, 2016, PAREonline

Example of Bootstrap for confidence interval of sample median

- The realized sample of student attendance
$\{x^b\}$ {12,10,9,8,10,11,12,7,5,10}, N=10, median=10

- Generate a random index uniformly from [1,10] that correspond to the 10 numbers in the sample, ie. if index=6, the bootstrap sample's number will be 11.
$N = 10$

- Repeat the process 10 times to get one bootstrap sample

Bootstrap replicate	Sample median
{11, 11, 12, 10, 10, 10, 12, 10, 7, 10}	10

$\{x^{b1}\}$

Example of Bootstrap for confidence interval of sample median

- The realized sample of student attendance {12,10,9,8,10,11,12,7,5,10}, *N*=10, median=10

Bootstrap replicate	**Sample median**	
{11, 11, 12, 10, 10, 10, 12, 10, 7, 10}	10	$\{x^{b1}\}$
{7, 10, 10, 10, 9, 7, 9, 10, 12, 10}	10	$\{x^{b2}\}$
{9, 7, 10, 8, 5, 10, 7, 10, 12, 8}	8.5	$\vdots$
…	…	

Q. How many possible bootstrap replicates?

- A. $\frac{10^{10}}{3!\,2!}$ B. 10! C. e^{10}

Bootstrap replicate	Sample median
{11, 11, 12, 10, 10, 10, 12, 10, 7, 10}	10
{7, 10, 10, 10, 9, 7, 9, 10, 12, 10}	10
{9, 7, 10, 8, 5, 10, 7, 10, 12, 8}	8.5
...	...

Example of Bootstrap for confidence interval of sample median

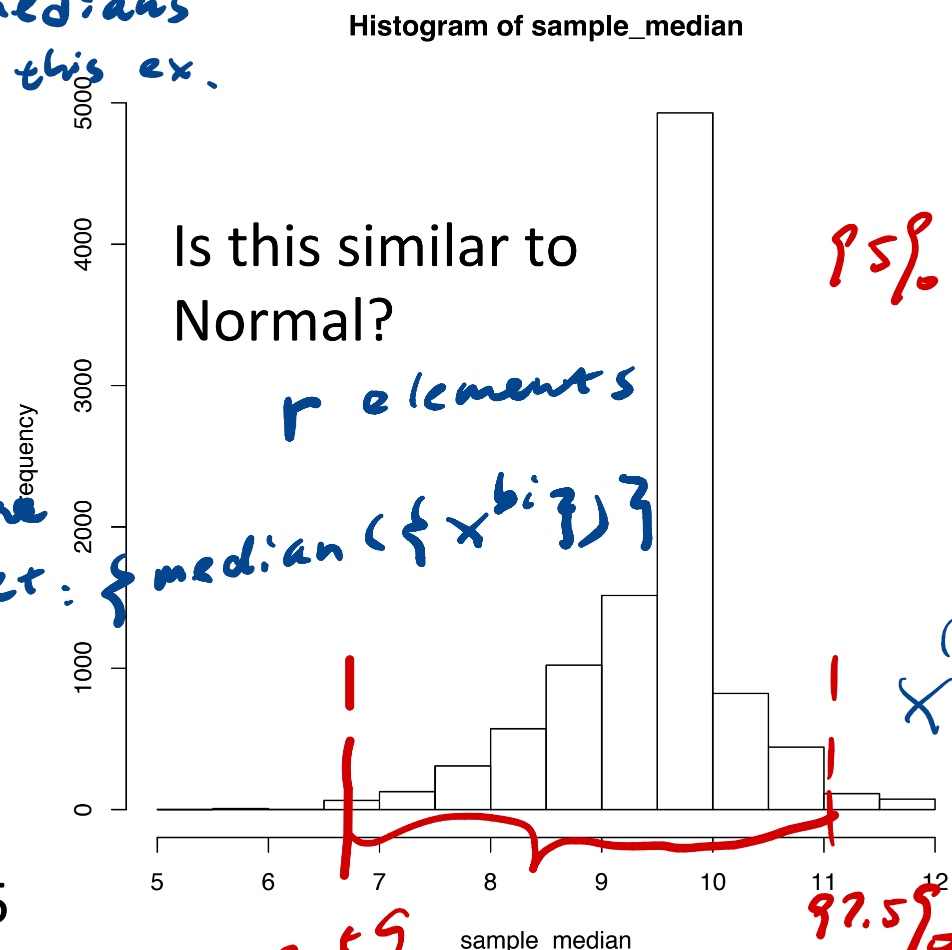
- Do the bootstrapping for **r = 10000** times, then draw the histogram and also find the stderr of sample median)

Bootstrap replicate	**Sample median**
{11, 11, 12, 10, 10, 10, 12, 10, 7, 10}	10
{7, 10, 10, 10, 9, 7, 9, 10, 12, 10}	10
{9, 7, 10, 8, 5, 10, 7, 10, 12, 8}	8.5
...	...

Example of Bootstrap for confidence interval of sample median

- Bootstrapping for **r = 10000** times, then draw the histogram and also find the stderr of sample median.

r medians for this ex.



Is this similar to Normal?

95%

r elements

stdevs of the data set: $\{median(\{x^{bi}\})\}$

$X^{(N)}$

2.5%

97.5%

$$stderr(\{S\}) = \sqrt{\frac{\sum_i [S(\{x\}_i) - \overline{S}]^2}{r-1}}$$

mean(Sample Median) = 9.73625

stderr(Sample Median) = 0.7724446

Errors in Bootstrapping

- The distribution simulated from bootstrapping is called empirical distribution. It is not the true population distribution. **There is a statistical error**.

$\{x^b\}$

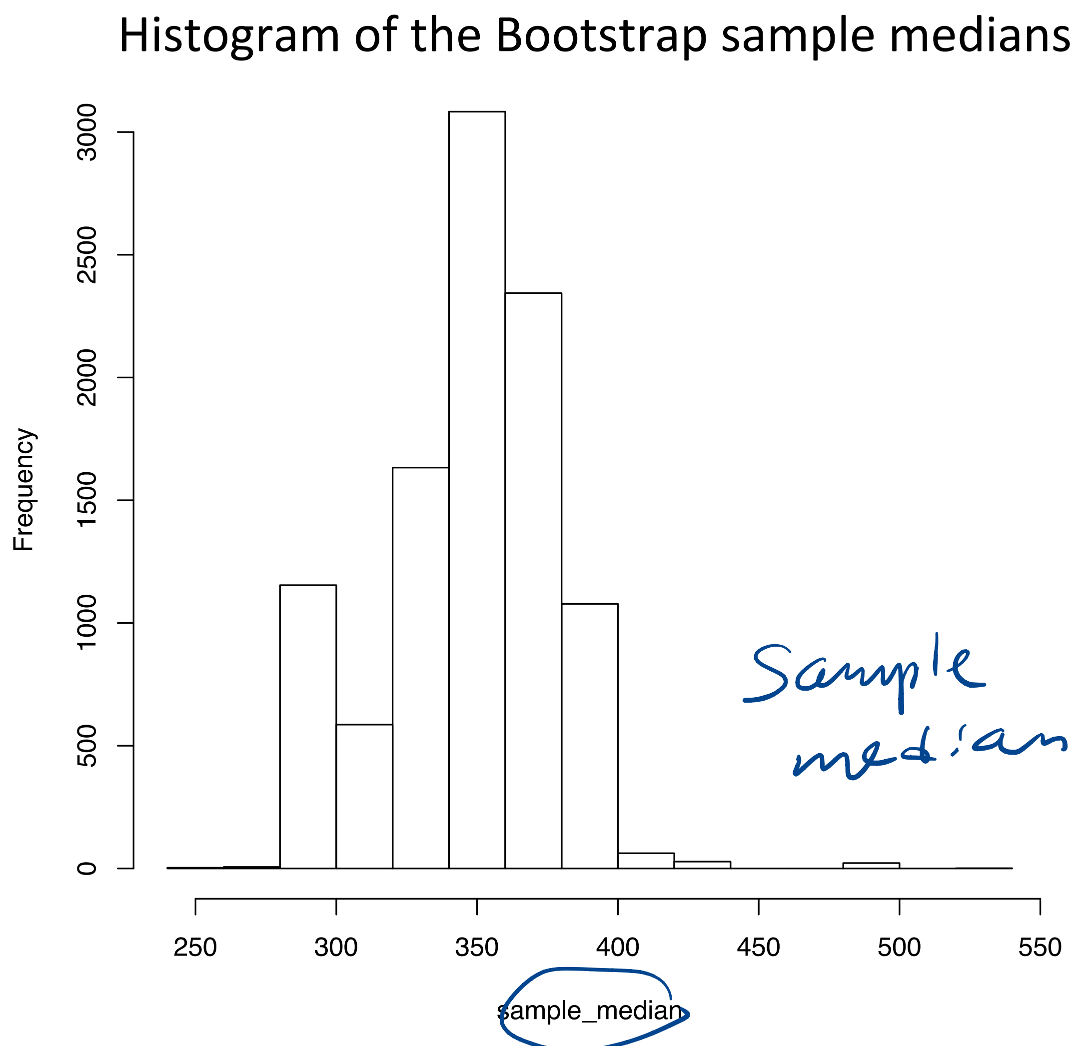
- The number of bootstrapping replicates may not be enough. **There is a numerical error**. 1000

- When the statistic is not a well behaving one, such as maximum or minimum of a data set, the bootstrap method may fail to simulate the true distribution.

CEO salary example with larger N = 59

- The realized sample of CEO salary N=59, median=350 K
- r = 10000

mean(Sample Median) = **348.0378**
stderr(Sample Median) = **27.30539**



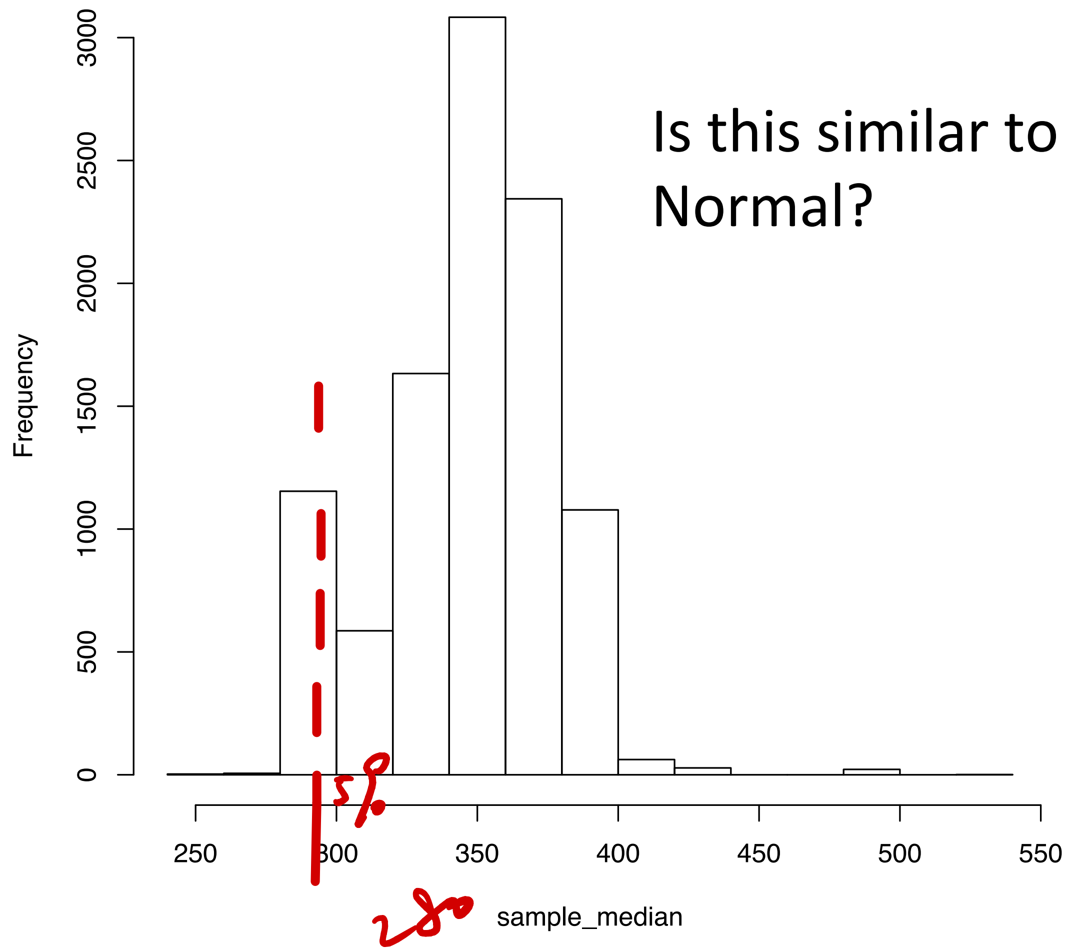
CEO salary example with larger N = 59

- The realized sample of CEO salary *N*=59, median=350 K
- r = 10000

mean(Sample Median) = **348.0378**

stderr(Sample Median) = **27.30539**

Histogram of the Bootstrap sample medians

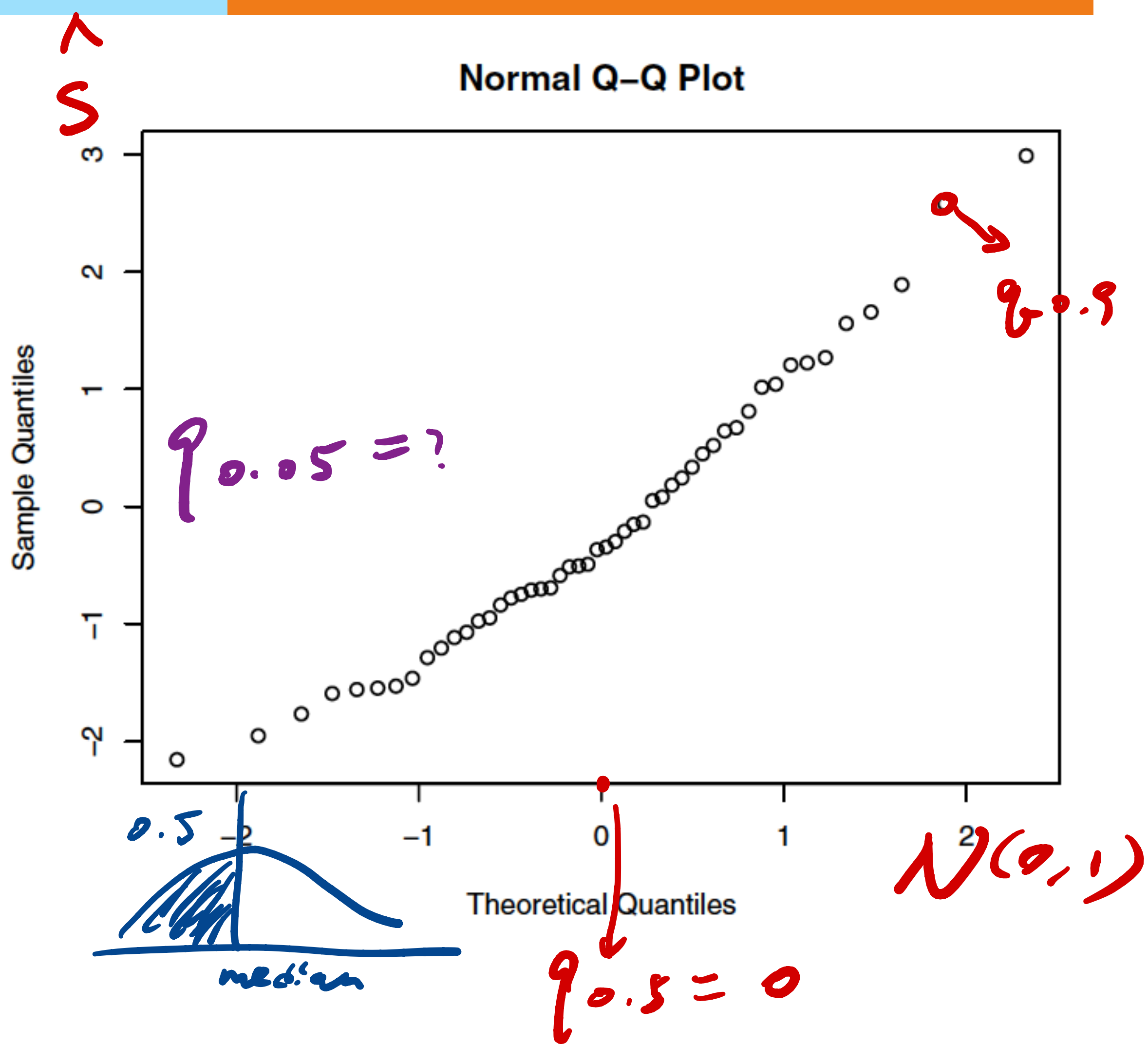


Is this similar to Normal?

Checking whether it's normal by Normal Q-Q plot

- Q-Q compares a distribution with normal by matching the kth smallest quantile value pairs and plot as a point in the graph

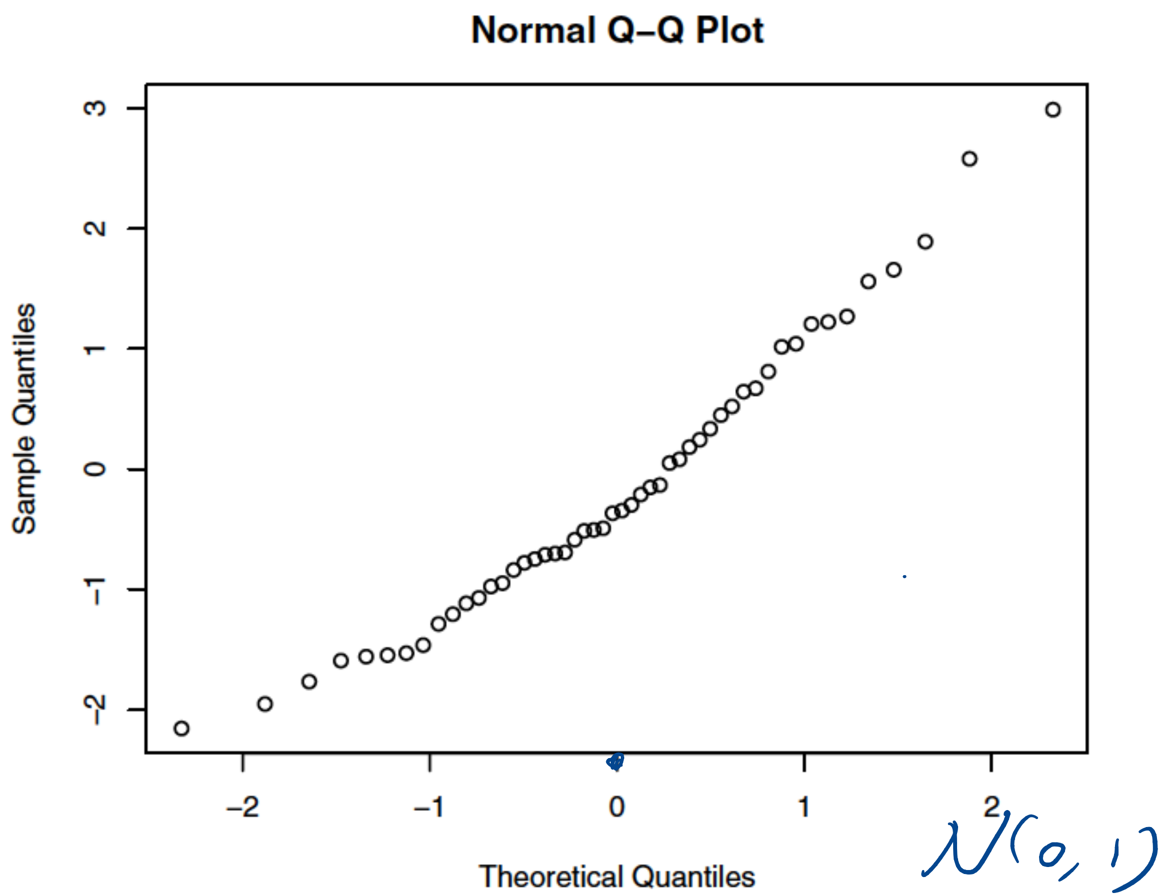
- **Linear means similar to normal!**



Read Pg 64, 3.2.3, "Introductory statistics with R"

Checking whether it's normal by Normal Q-Q plot

- Q-Q compares a distribution with normal by matching the kth smallest quantile value pairs and plot as a point in the graph
- **Linear means similar to normal!**

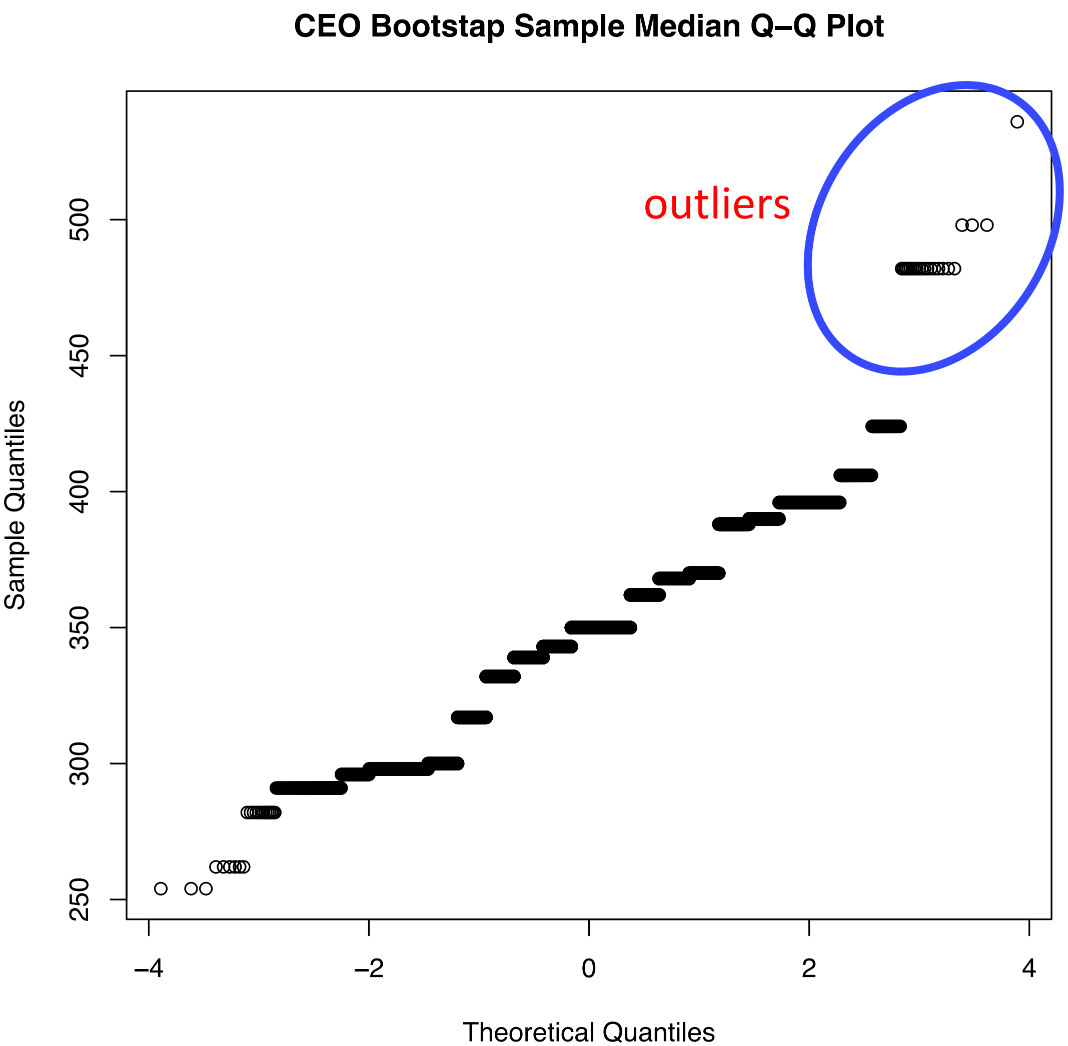


Normal Distribution's Quantile

Read Pg 64, 3.2.3, "Introductory statistics with R"

CEO salary sample median's Q-Q plot

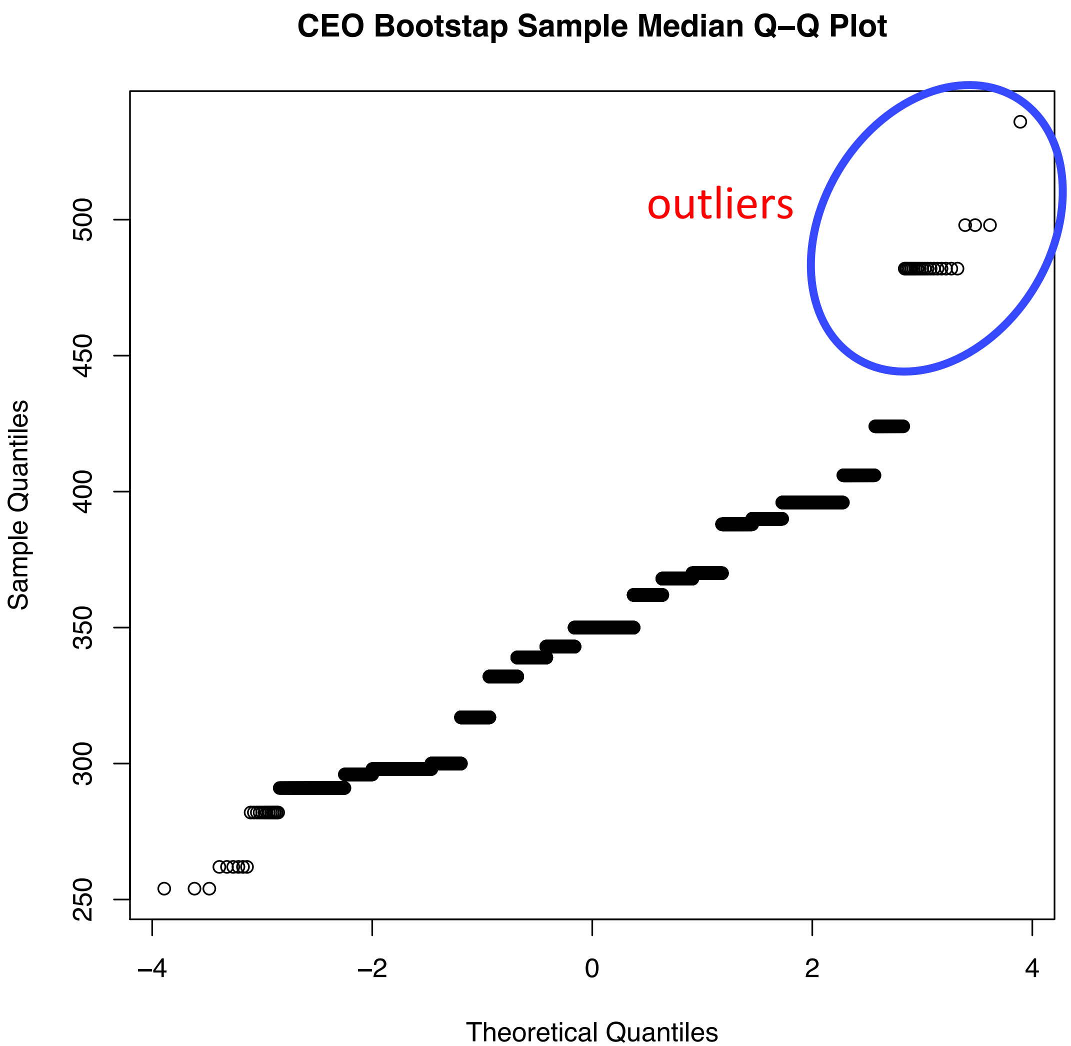
- Q-Q plot of CEO salary's bootstrap sample medians
- It's roughly linear so it's close to normal.
- We can use the normal distribution to construct the confidence intervals



CEO salary sample median's Q-Q plot

- 95% confidence interval for the median CEO salary from the bootstrap simulation
- **348.0378± 2×27.30539**

= [293.427, 402.6486]



Assignments

- Read Chapter 7 of the textbook
- Week 8 module on Canvas
- Next time: hypothesis testing

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!