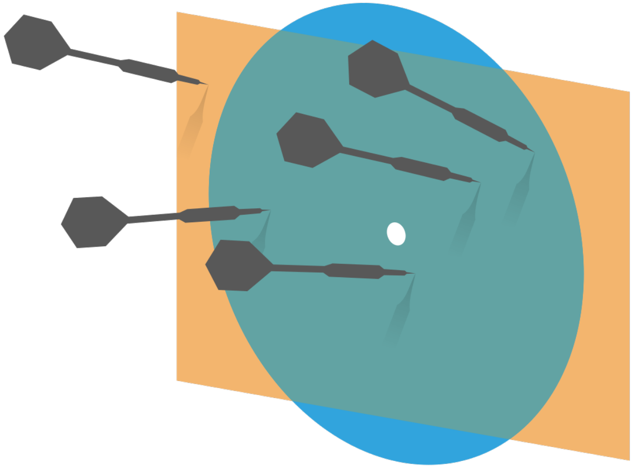
Probability and Statistics for Computer Science



Credit: wikipedia

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

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 10.13.2020

Which of the following do you feel like to be?

A. Theorist

B. Experimentalist

C. Both

D. Others

Last time

- Exponential Distribution
- Sample mean and confidence interval

Objectives

Recap of Sample mean, confidence interval

Bootstrap Simulation

Hypothesis test

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

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

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

A tale of two statisticians

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

The task: use only a subset of the population with $N = 5$ to estimate the popmean with some confidence report.

RV $X^{(N)} \longrightarrow$ simple mean

A tale of two statisticians

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

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

One random sample $\boxed{\{x\}} = \{1, 4, 5, 7, 11\}$ $N=5$

iid

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

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

A tale of two statisticians

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

One random sample $\rightarrow \boxed{\{x\}} = \{1, 4, 5, 7, 11\}$ $\quad N=5$

if $N \to \infty$

$$X^{(N)} \stackrel{\cdot}{\approx} \mathcal{N}\left(\mu(X^{(N)}), \sigma(X^{(N)})\right)$$

$$\mu(X^{(N)}) \doteq \text{mean}(\{x\})$$

$$\sigma(X^{(N)}) \doteq \text{stderr}(\{x\})$$

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

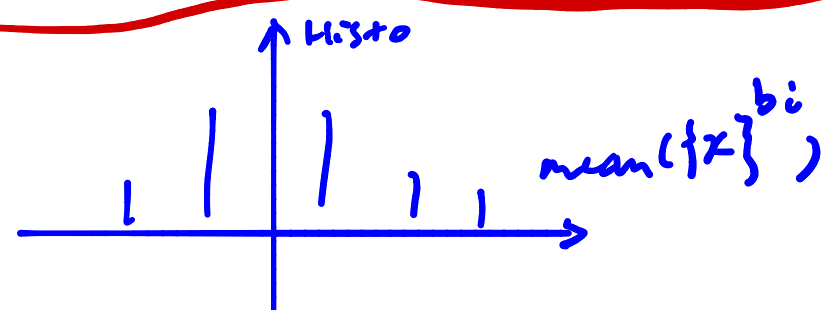
$$\{x\}^{b1} = \{1, 1, 4, 5, 7\}$$

$$\{x\}^{b2} = \{4, 5, 7, 7, 11\}$$

$$\vdots$$

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

$$5^5 = 3125$$



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

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

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

- 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

$\{x\} = \{x_i\} = \{x_1, x_2, x_3, \cdots x_N\}$

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

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

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

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

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

- $stdunbiased(\{x\})$ $= \sqrt{\frac{1}{N-1}\left(\sum_{i=1}^{N}(x_i - mean(\{x\}))^2\right)}$

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

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

$\ldots 0, 1\}$

618 "1"

593 "0"

- $stderr(\{x\})$

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

$N = 1211$

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

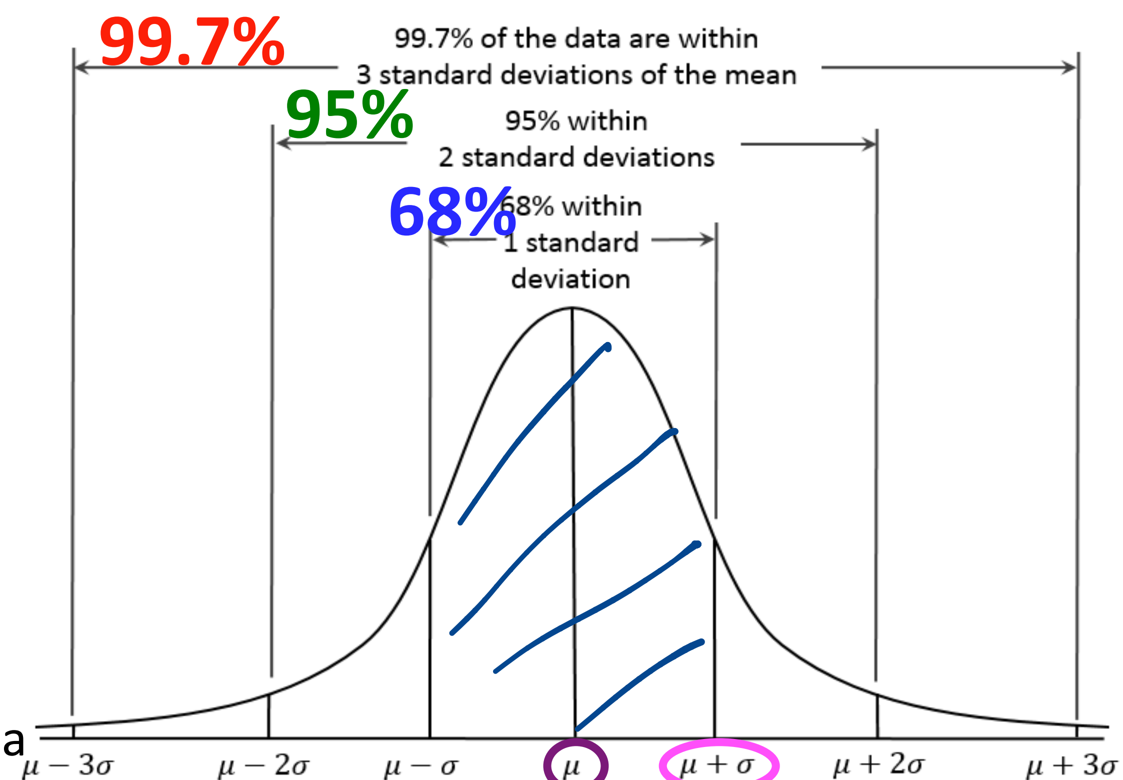
- When ***N*** is very large, according to the **Central Limit Theorem**, sample mean is approaching a normal distribution with

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

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

Interpreting the standard error

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



Credit: wikipedia

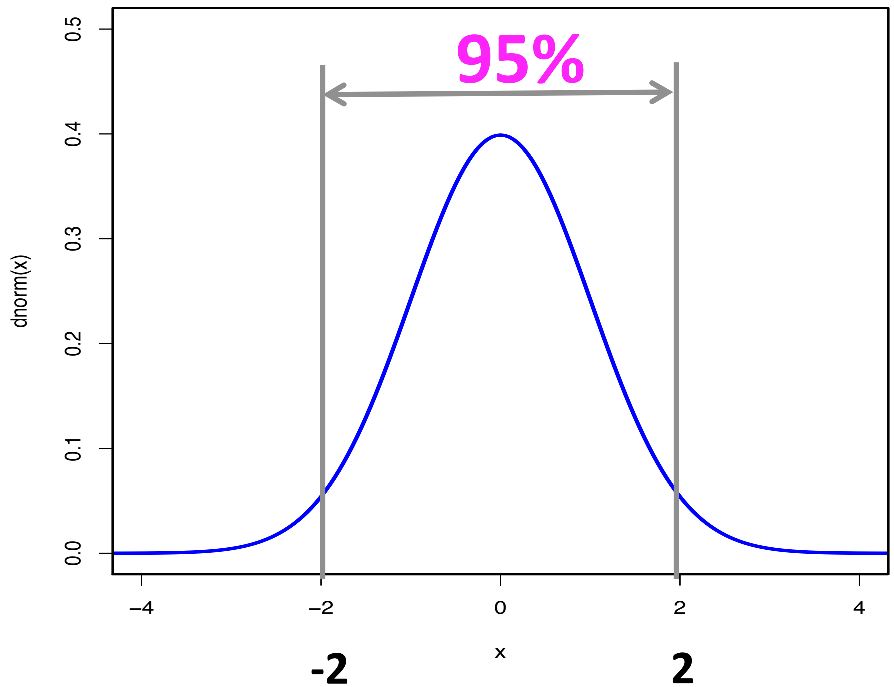
$X^{(N)}$

Population mean

μ+Standard error

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

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

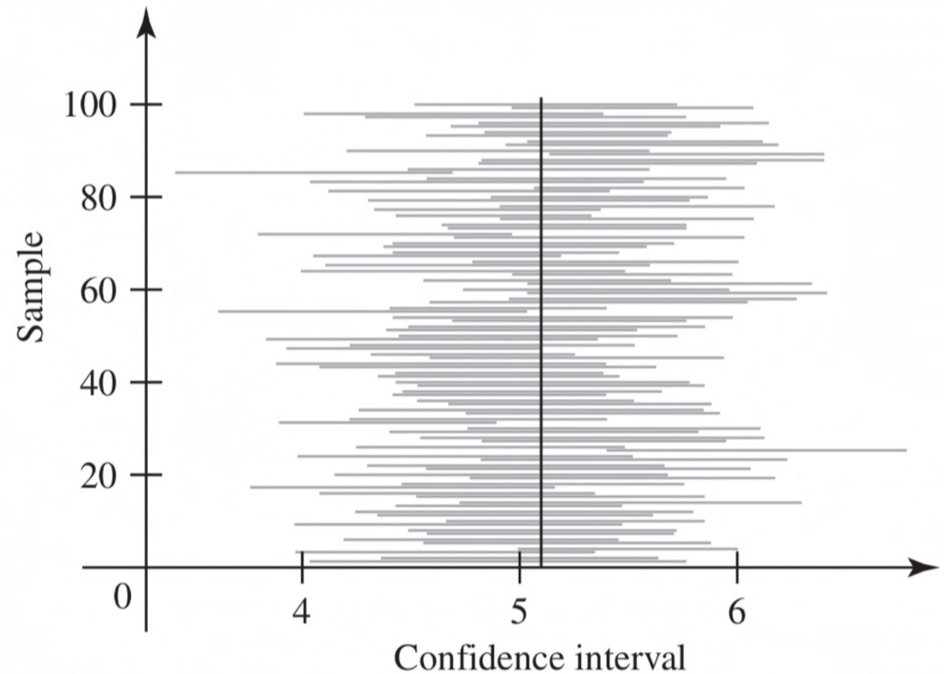
Q. Confidence intervals

- 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]

Interpreting the confidence intervals

Figure 8.5 A sample of one hundred observed 95% confidence intervals based on samples of size 26 from the normal distribution with mean $\mu = 5.1$ and standard deviation $\sigma = 1.6$. In this figure, 94% of the intervals contain the value of μ.



Degroot Pg 487

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.

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

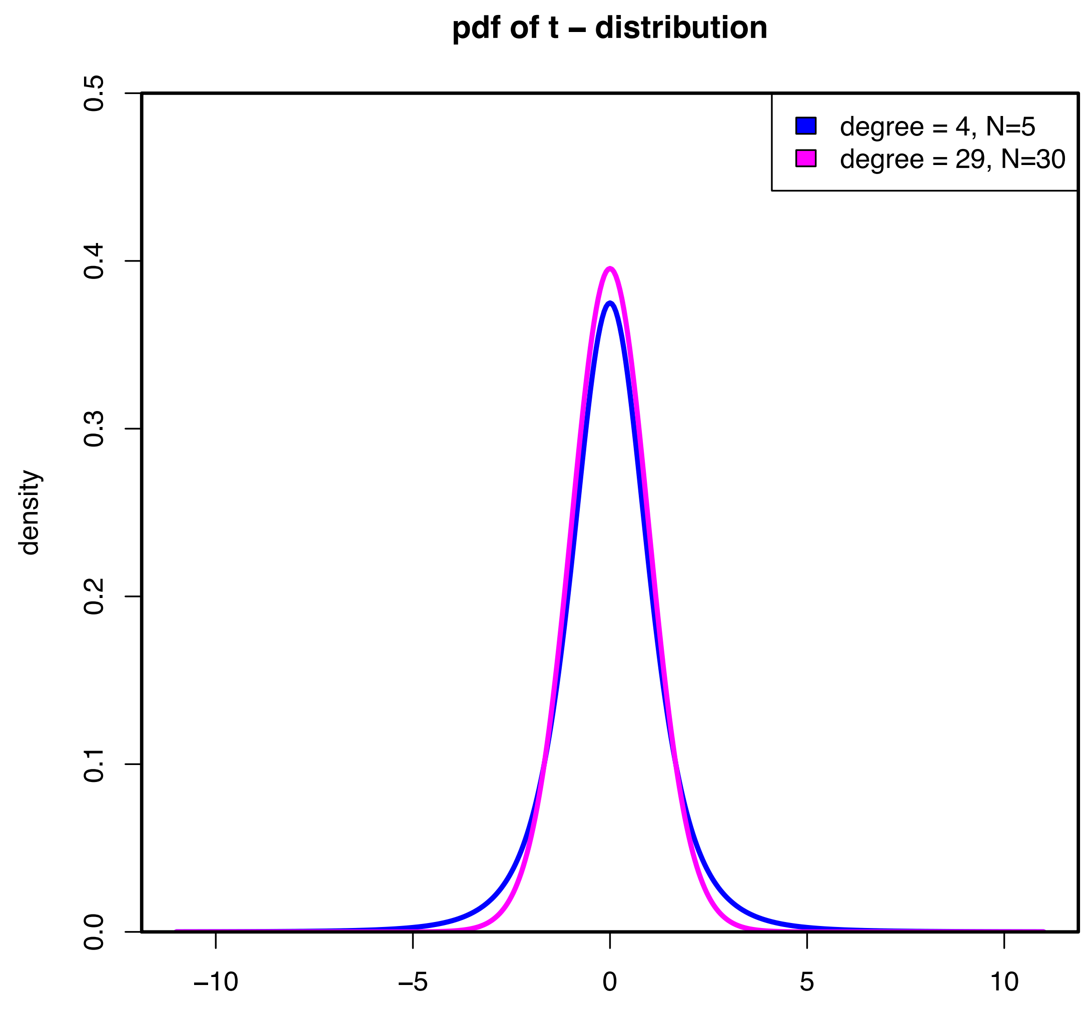
Degree of freedom is ***N***-1 due to this constraint: $$\sum_i (x_i - mean(\{x\})) = 0$$

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

t-distribution with N=5 and N=30

Credit : wikipedia

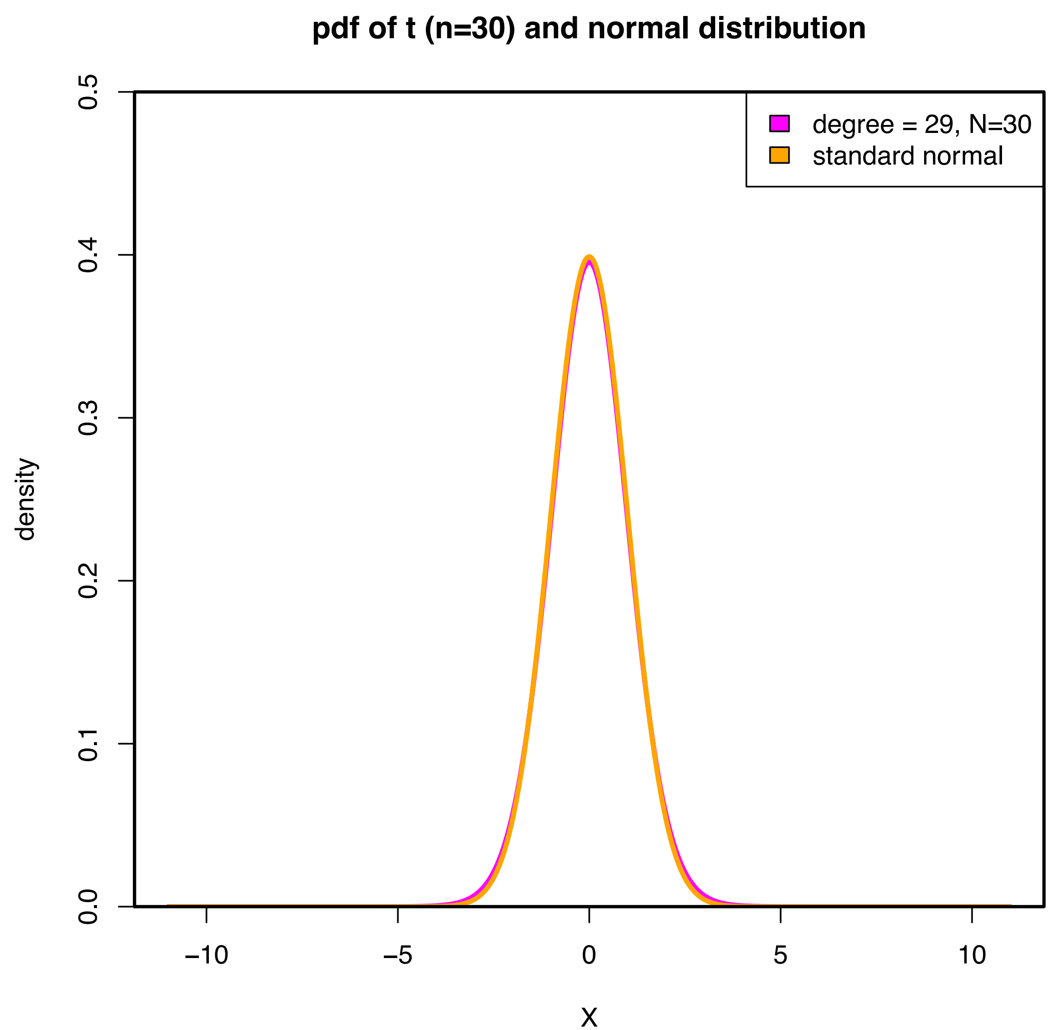
William Sealy Gosset 1876-1937



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**



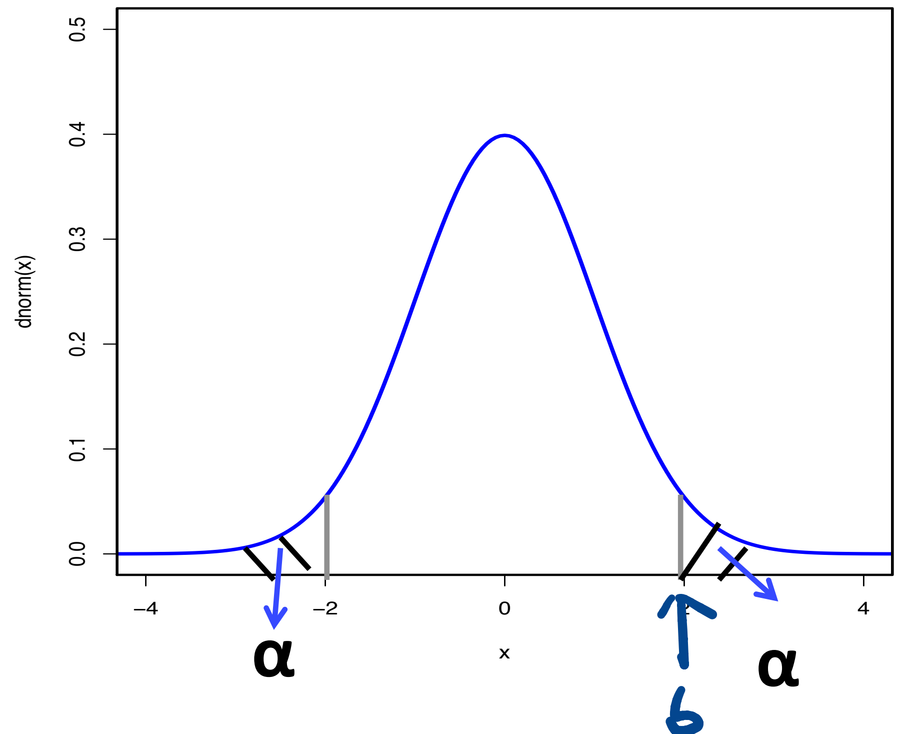
Confidence intervals when N< 30

- 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

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



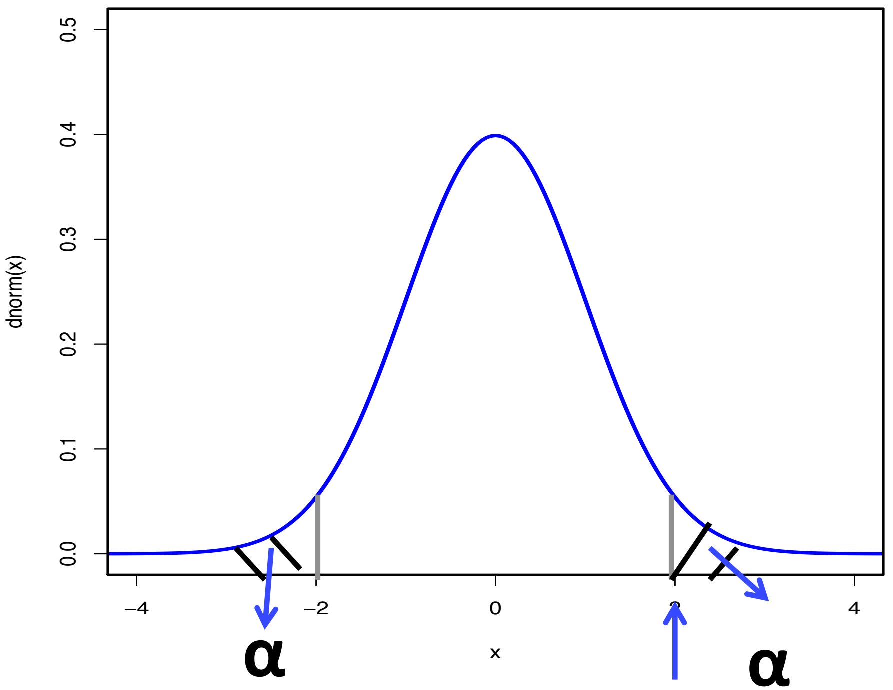
$1-2\alpha$

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. $\alpha = 0.05$

B. $\alpha = 0.025$

C. $\alpha = 0.1$

$$\frac{1 - 95\%}{2} = \alpha$$

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.

$$\begin{bmatrix} 9 \\ 5 \end{bmatrix}$$

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

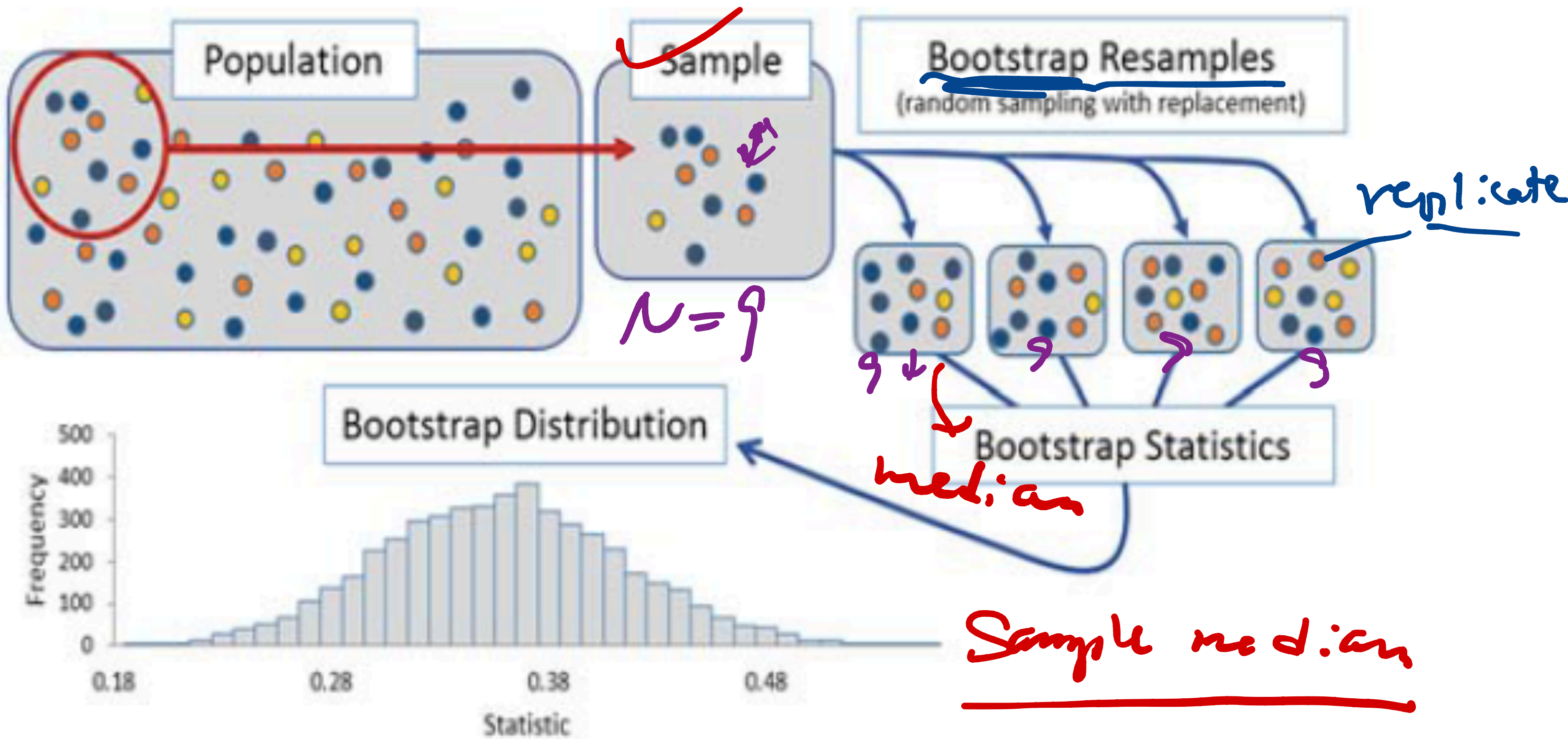


Figure 1. Summary of Bootstrapping Process

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 {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.
- 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

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
{7, 10, 10, 10, 9, 7, 9, 10, 12, 10}	10
{9, 7, 10, 8, 5, 10, 7, 10, 12, 8}	8.5
…	…

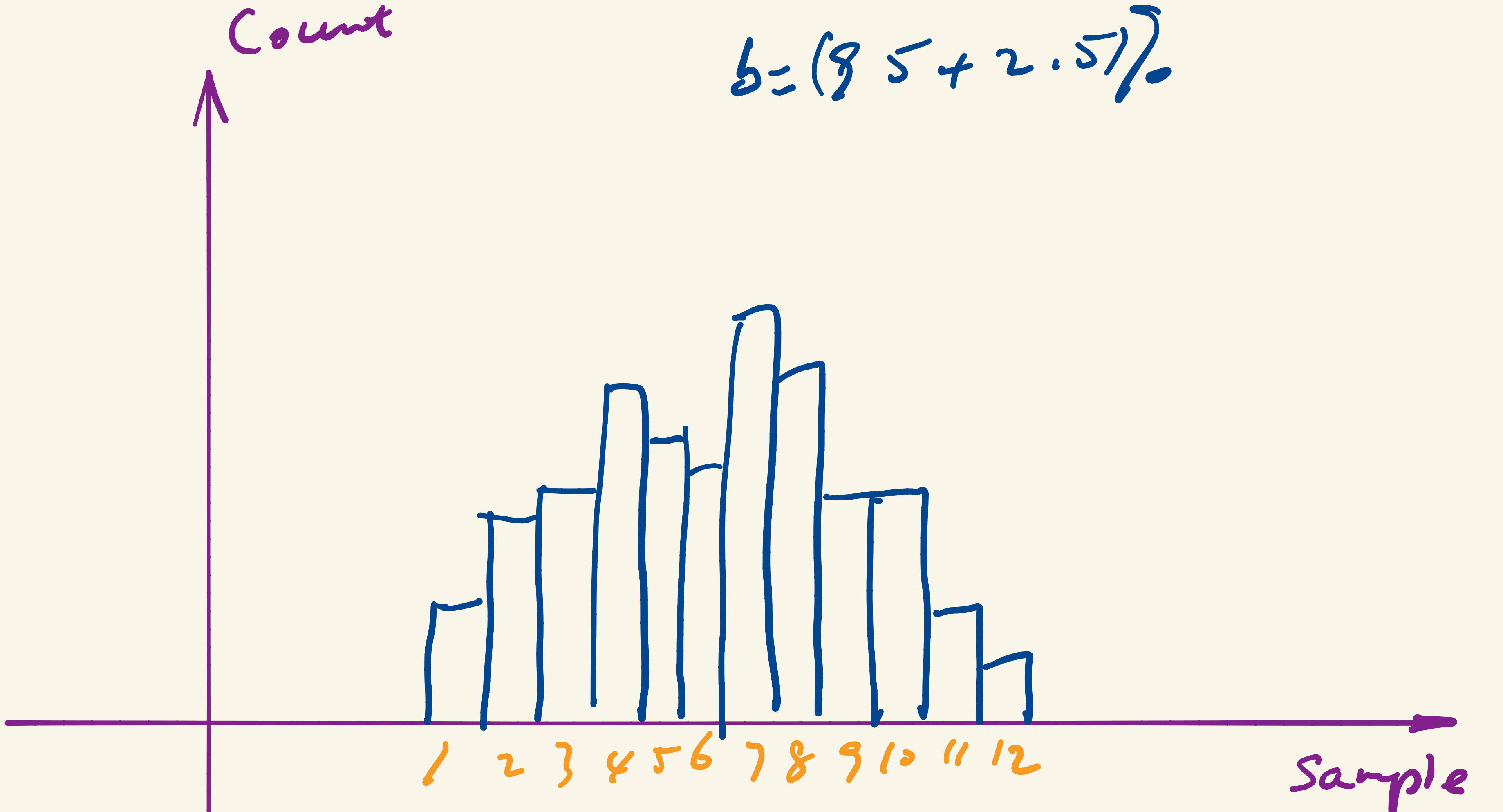
Example of Bootstrap for confidence interval of sample median

- 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
…	…

$a = 2.5\%$

$b = (95 + 2.5)\%$

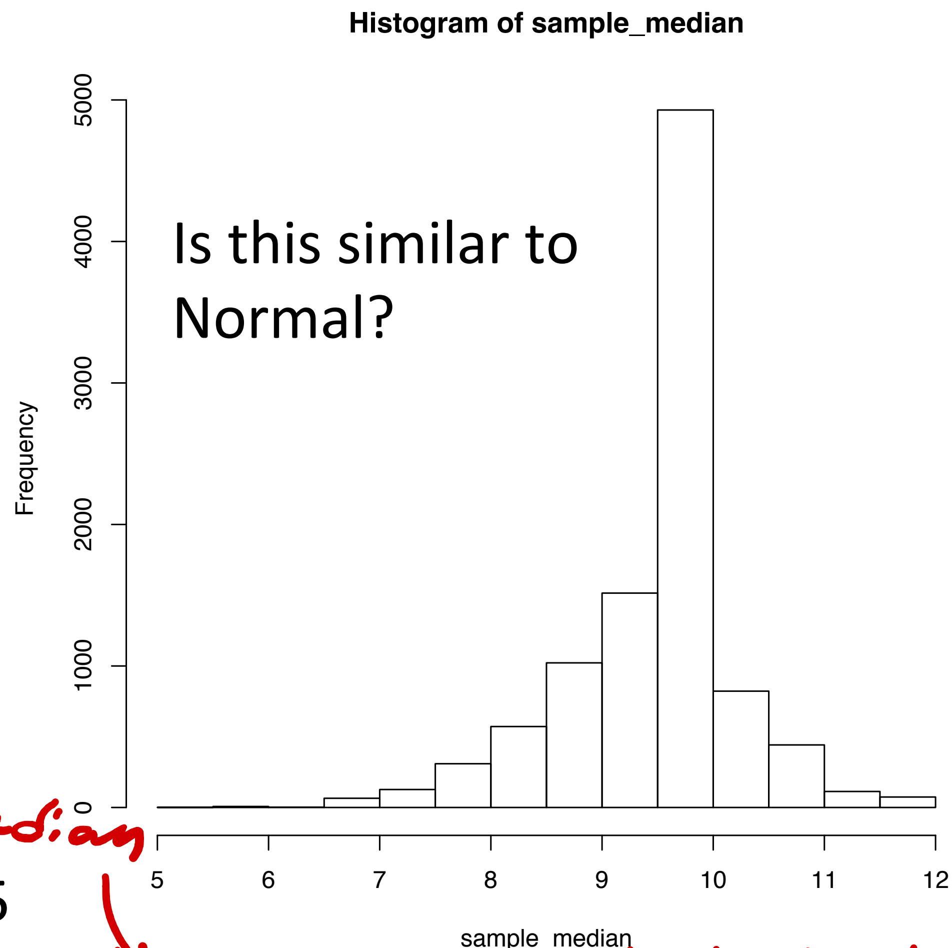


mean({sample medians})

std({sample medians})

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.



Is this similar to Normal?

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

std unbiased of sample median

mean(Sample Median) = 9.73625

stderr(Sample Median) = 0.7724446

notation was following the book.

Errors in Bootstrapping

- The distribution simulated from bootstrapping is called empirical distribution. It is not the true population distribution. **There is a statistical error**.
- The number of bootstrapping replicates may not be enough. **There is a numerical error**.
- 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

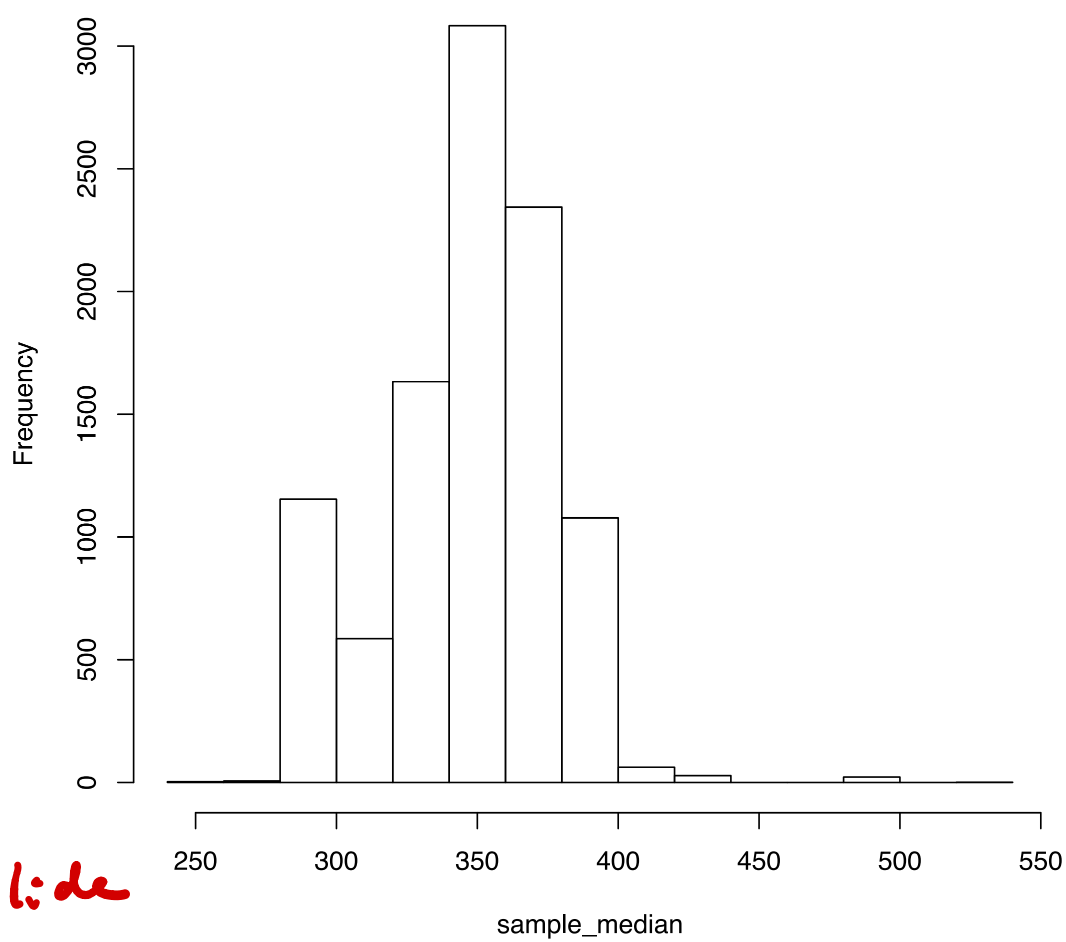
- The realized sample of CEO salary *N*=59, median=350 K
- r = 10000

mean(Sample Median) = **348.0378**

$\boxed{\text{stderr}}$*(Sample Median) = **27.30539**

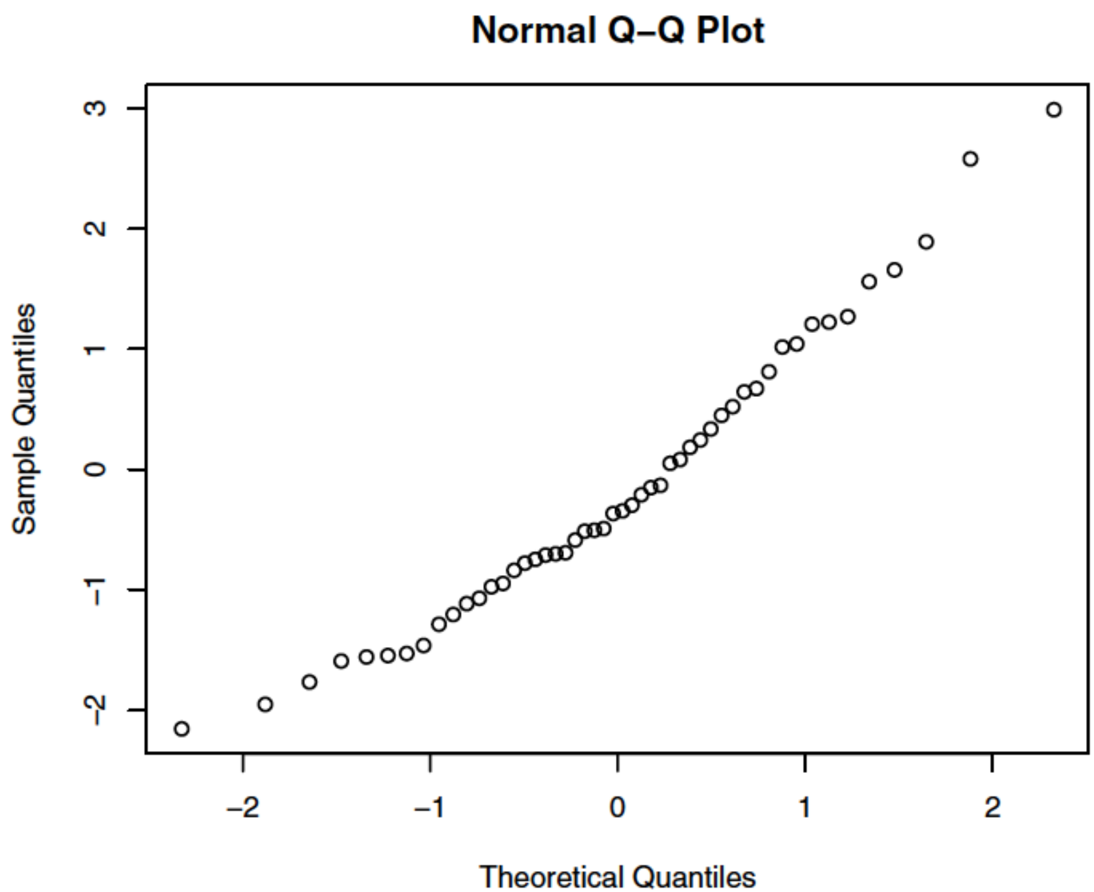
note as in previous slide

Histogram of the Bootstrap sample medians



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

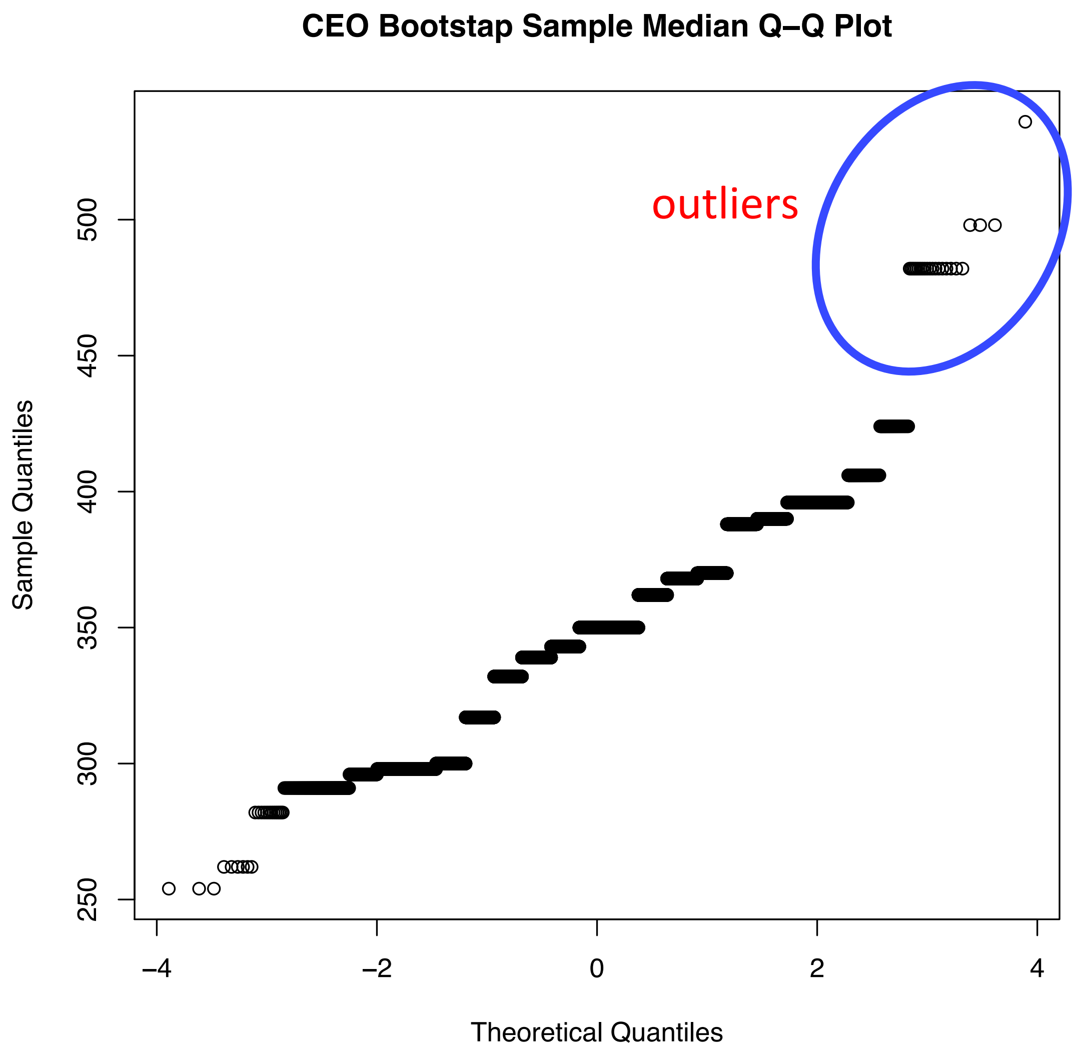
- Q-Q compares a distribution with normal by matching the kth smallest quantile value pairs and plot as a point in the graph
- **Linear means similar to normal!**



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

CEO salary sample median's Q-Q plot

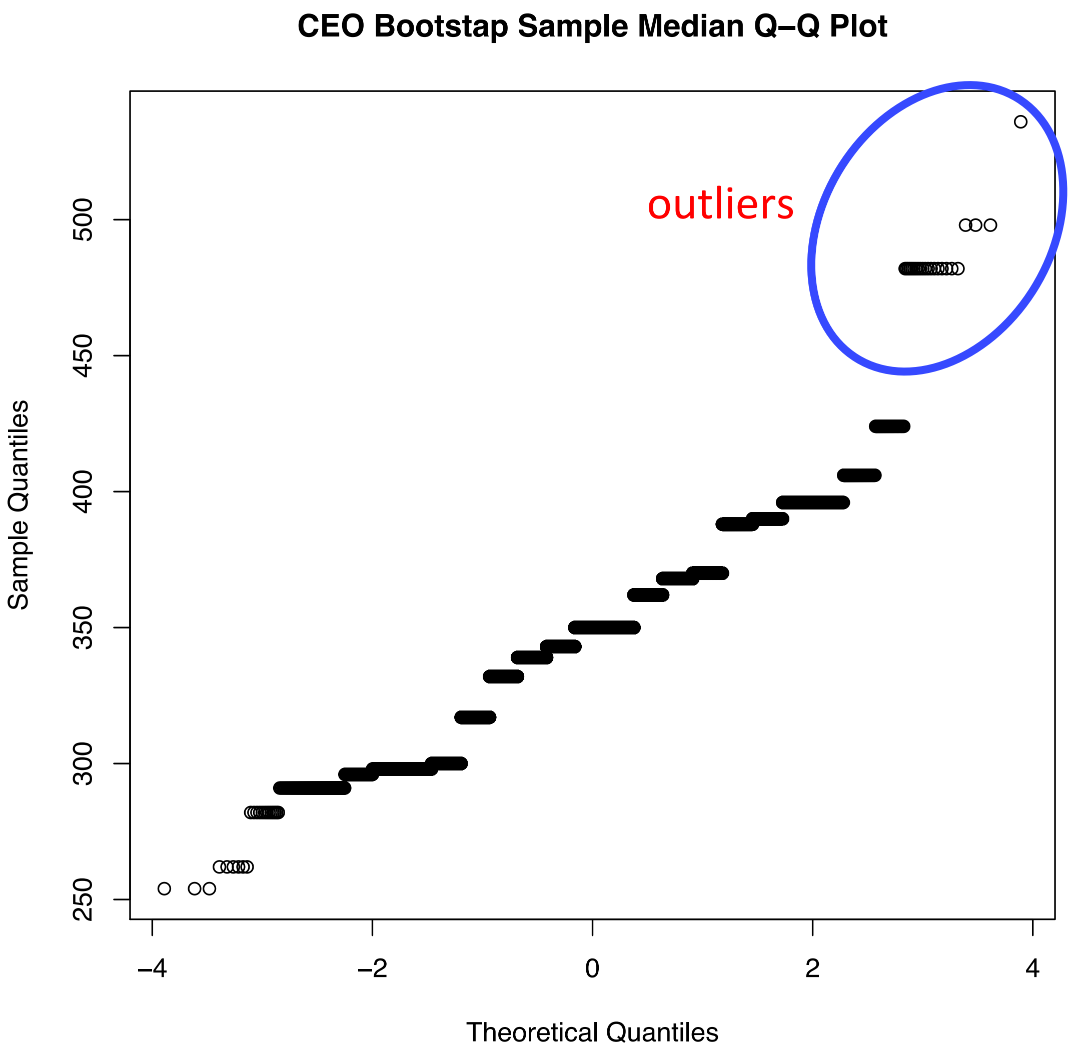
- Q-Q plot of CEO salary's bootstrap sample medians
- It's roughly linear so it's close to normal.
- We can use the normal distribution to construct the confidence intervals



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]



$$H_0 : \text{popmean}(\{X\}) = v_0$$

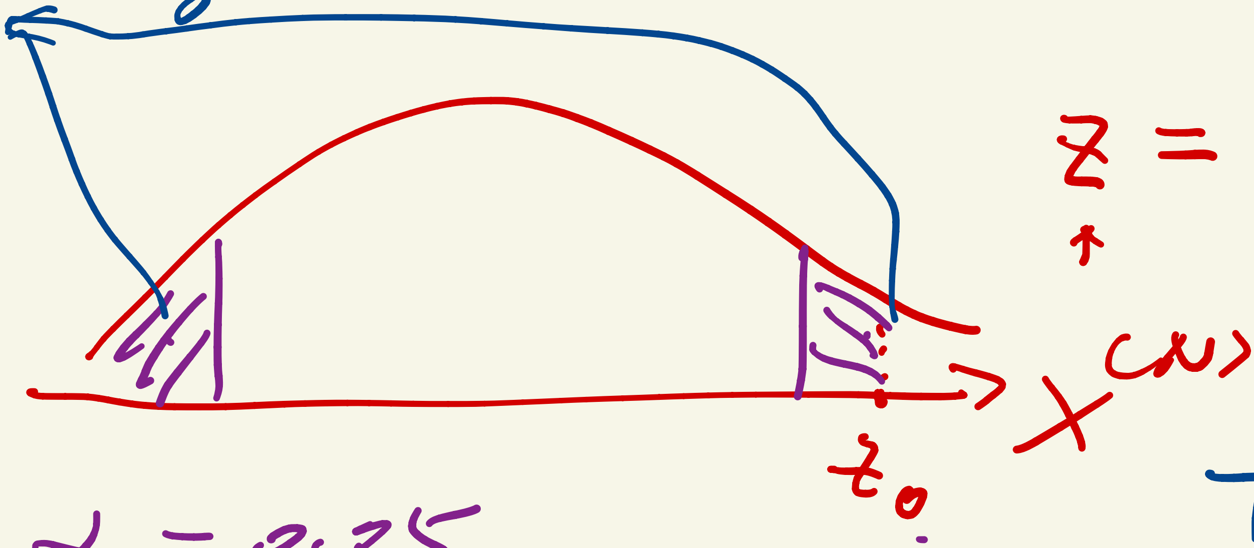
$$H_1 : \text{popmean}(\{X\}) \neq v_0$$

$\{x\}$ sample mean value

mean($\{x\}$)

$N > 30$

rejection region



$$Z = \frac{\text{mean}(\{x\}) - \underbrace{\text{popmean}}_{v_0}}{\text{stderr}}$$

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

$2\alpha = 0.05$

Assignments

- Read Chapter 7 of the textbook
- Next time: more on 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!