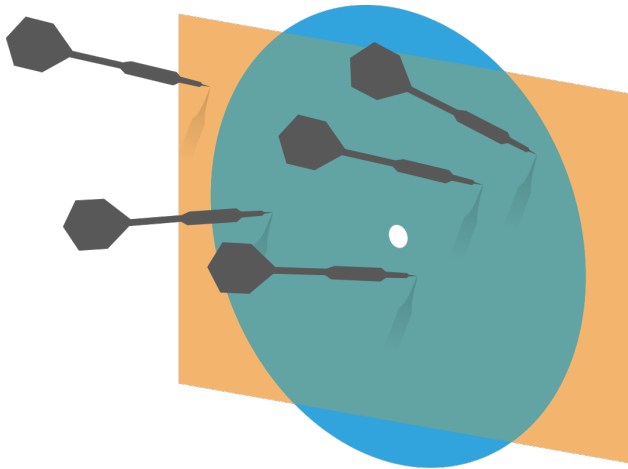


Probability and Statistics for Computer Science



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

Credit: wikipedia



Last time

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

Objectives

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

Motivation of sampling: the poll example

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

Source: [FiveThirtyEight.com](https://www.forty8eight.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?

Expected value of one random sample is the population mean

- ✱ Since each sample is drawn uniformly from the population

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

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

- ✱ 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

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

- ✱ The standard deviation of the sample mean

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

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

Unbiased estimate of population standard deviation & Stderr

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

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

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

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

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

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:
 $1211(0.51) \approx 618$

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

Standard error: election poll

✱ $stdunbiased(\{x\})$

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

✱ $stderr(\{x\})$

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

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

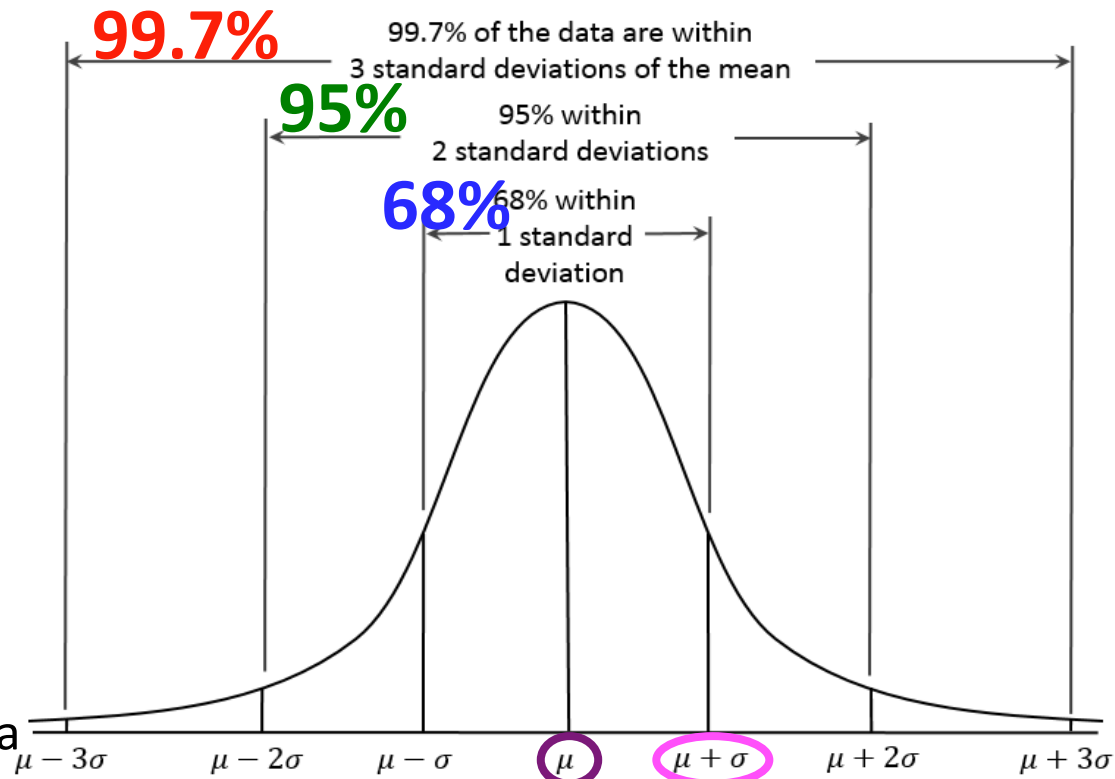
$$\mu = \text{popmean}(\{X\}) ; \sigma = \frac{\text{popstd}(\{X\})}{\sqrt{N}} \doteq \text{stderr}(\{x\})$$

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

Interpreting the standard error

Probability distribution of sample mean tends normal when N is large

Credit: wikipedia

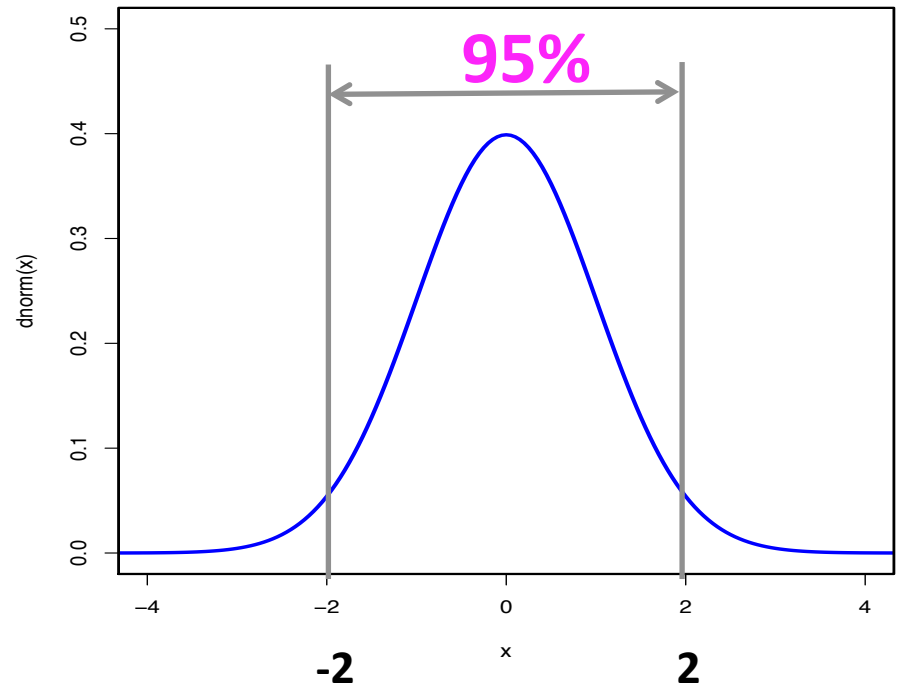


Population mean

$\mu + \text{Standard error}$

Confidence intervals

- ✱ Confidence interval for a population mean is defined by fraction
- ✱ Given a percentage, find how many units of stderr 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

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

- ✱ For about 95% of realized sample means

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



- ✱ For about 99.7% of realized sample means

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

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	 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: $\text{strerr} > 0.05$)

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.

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

Degree of freedom is **N-1** due

to this constraint:
$$\sum_i (x_i - \text{mean}(\{x\})) = 0$$

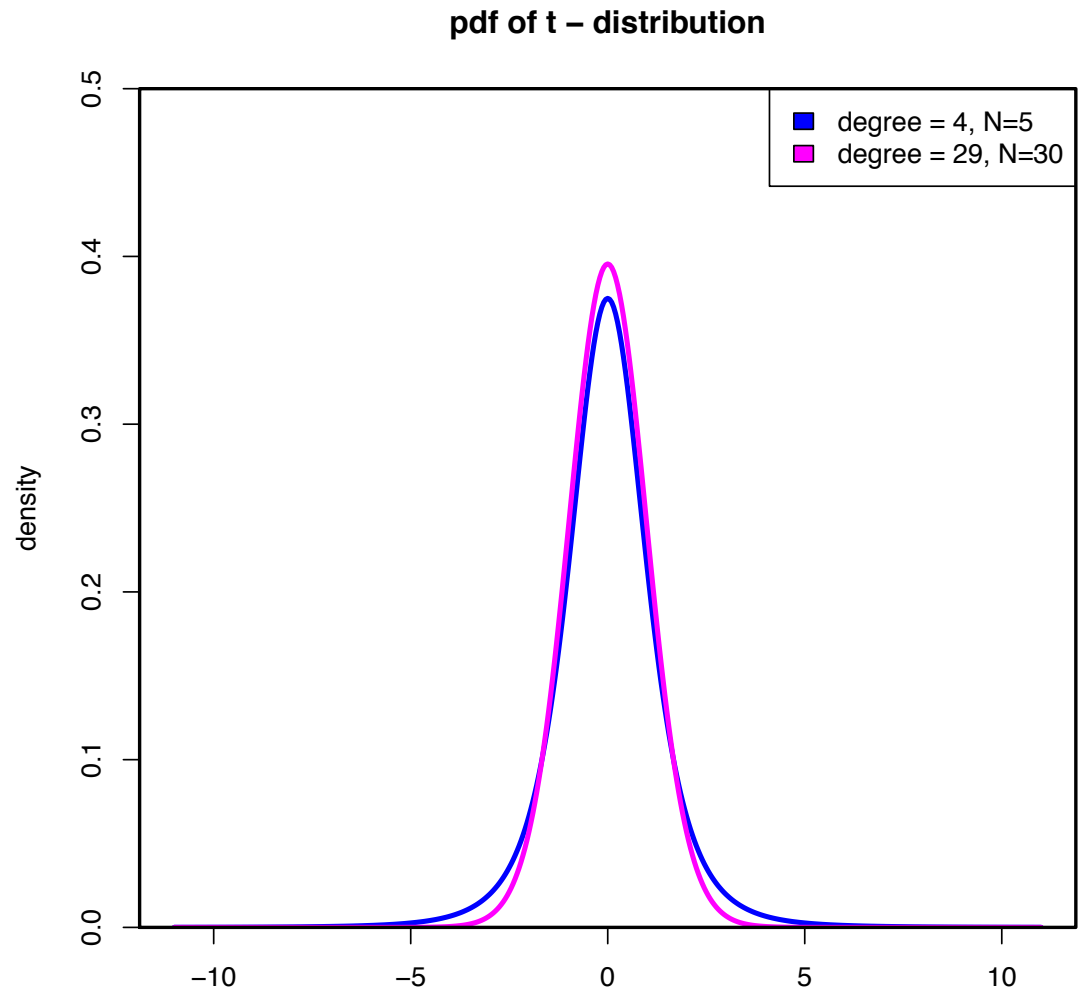
t-distribution is a family of distri. with different degrees of freedom

t-distribution with $N=5$
and $N=30$



Credit :
wikipedia

William Sealy Gosset 1876-1937

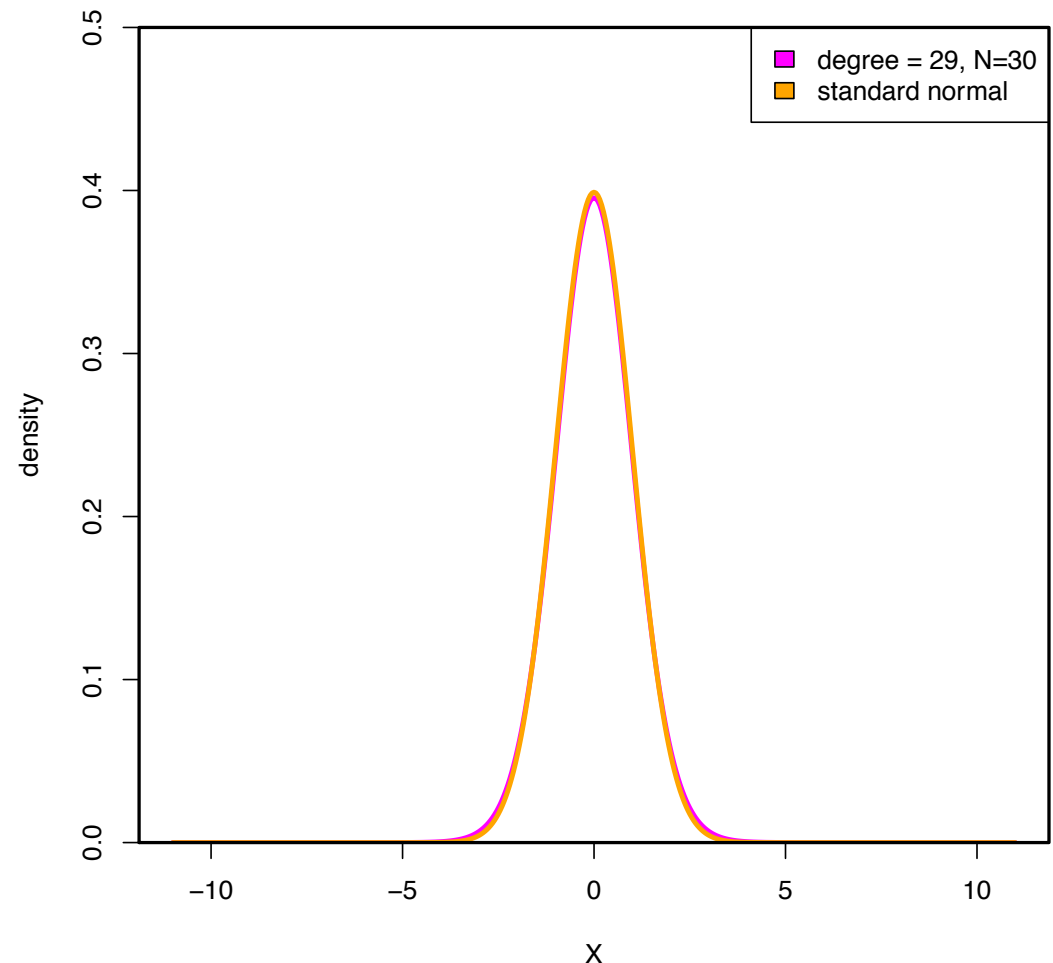


When $N=30$, t-distribution is almost Normal

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

pdf of t (n=30) and normal distribution



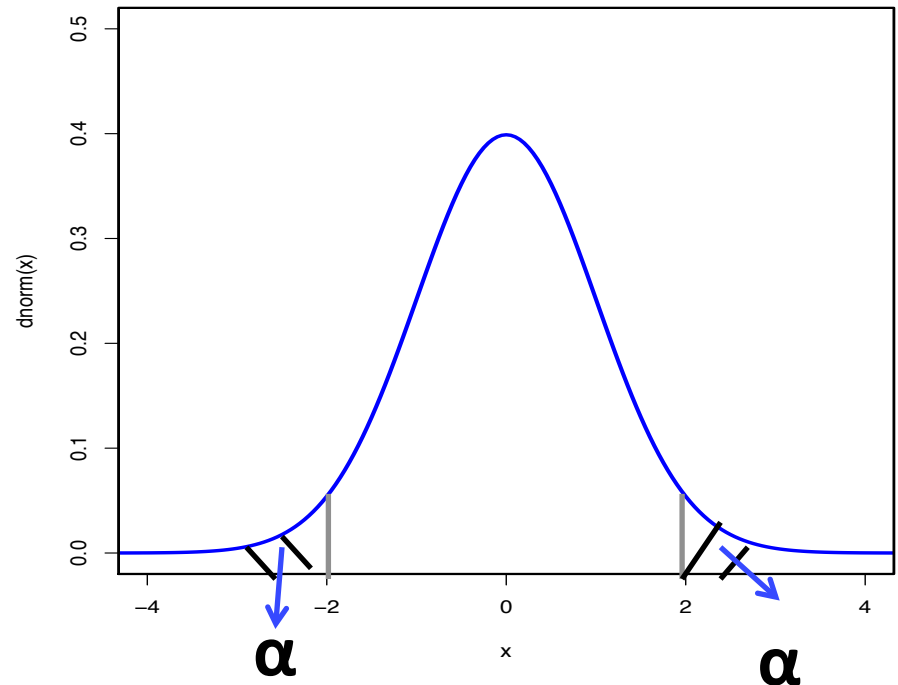
Confidence intervals when $N < 30$

- ✱ 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

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

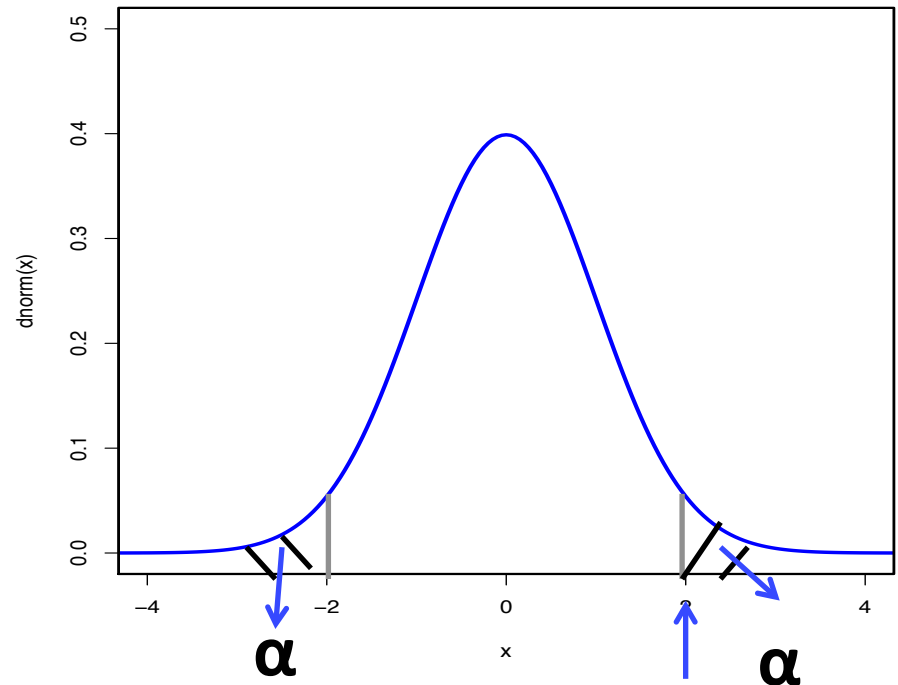


For $1-2\alpha$ of the realized sample means,
the population mean lies in
[sample mean- $b \times$ stderr, sample mean+ $b \times$ stderr]

Centered Confidence intervals

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

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



For $1-2\alpha$ of the realized sample means,
the population mean lies in
[sample mean- $b \times \text{stderr}$, sample mean+ $b \times \text{stderr}$]

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. $\alpha = 0.1$

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

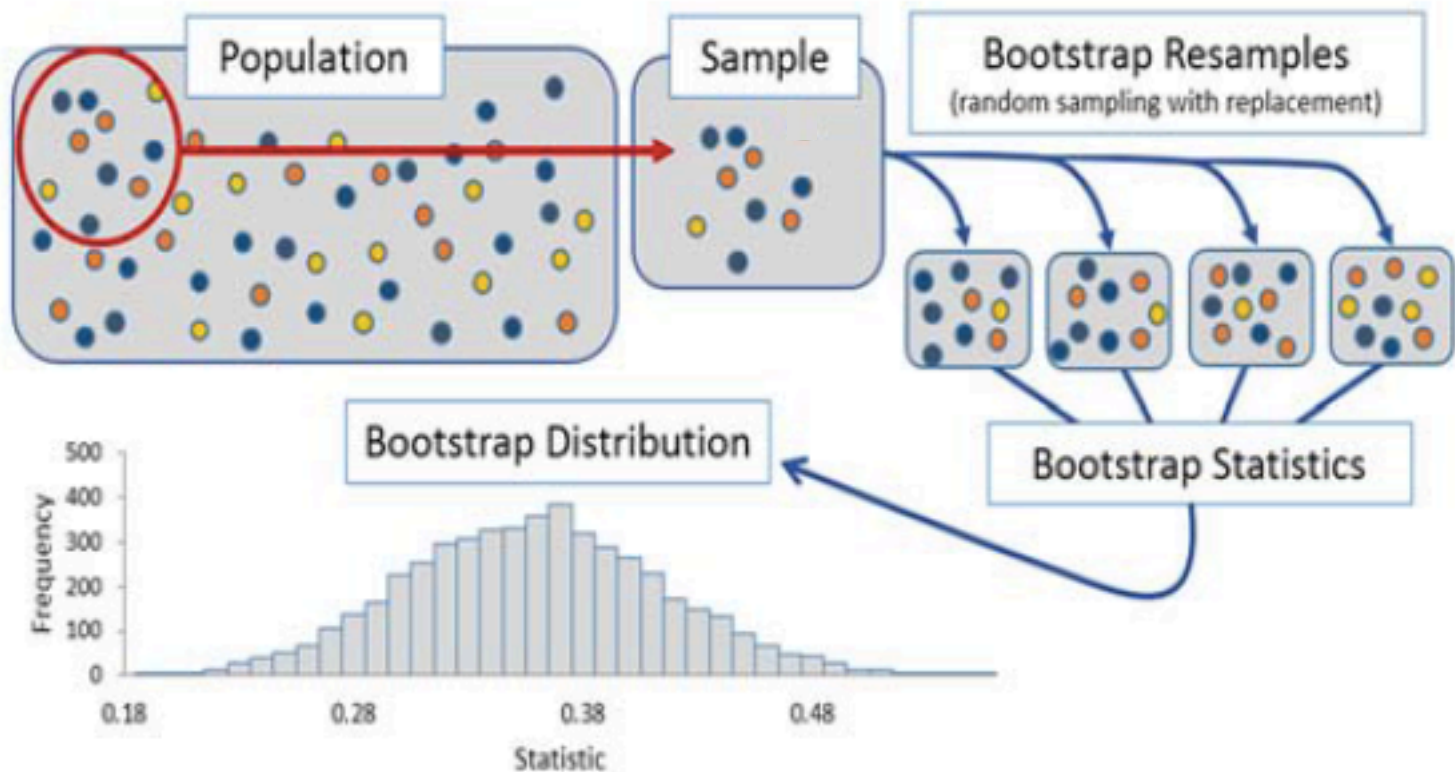


Figure 1. Summary of Bootstrapping Process

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
...	...

Q. How many possible bootstrap replicates?

- ✱ A. 10^{10} 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)

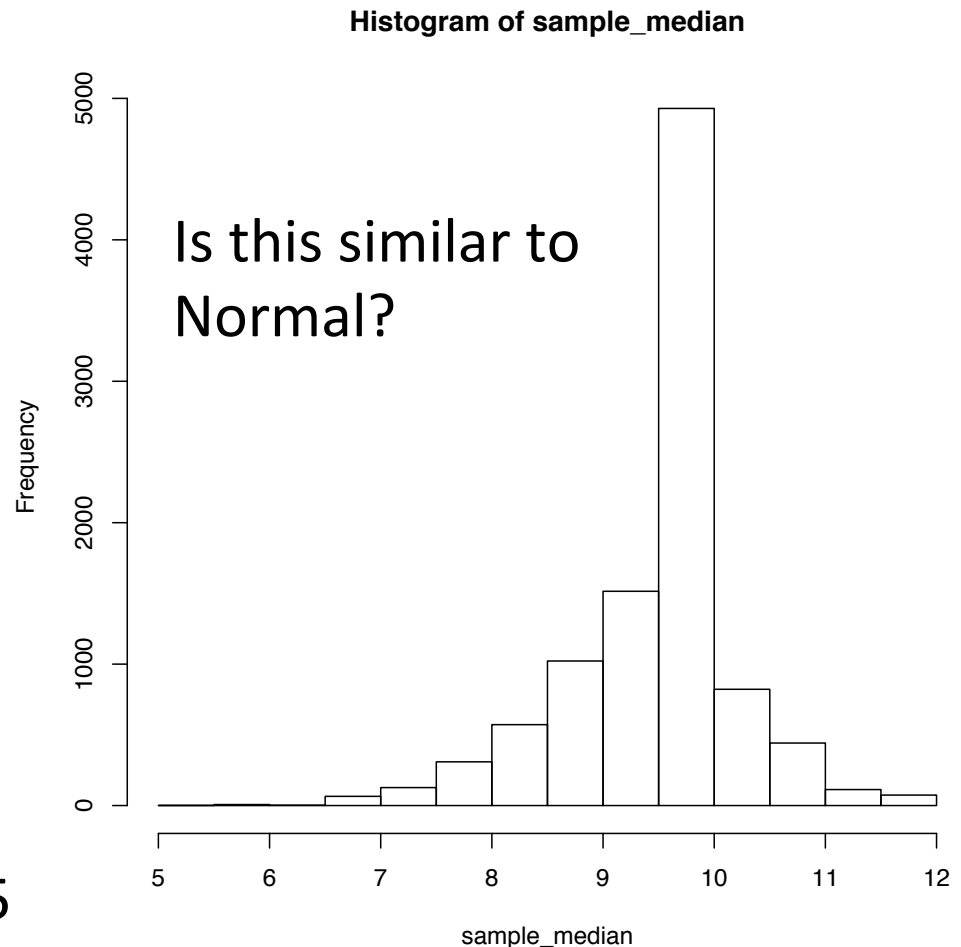
Bootstrap replicate	Sample median
{11, 11, 12, 10, 10, 10, 12, 10, 7, 10}	10
{7, 10, 10, 10, 9, 7, 9, 10, 12, 10}	10
{9, 7, 10, 8, 5, 10, 7, 10, 12, 8}	8.5
...	...

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.

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

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.**
- ✱ 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

✱ $r = 10000$

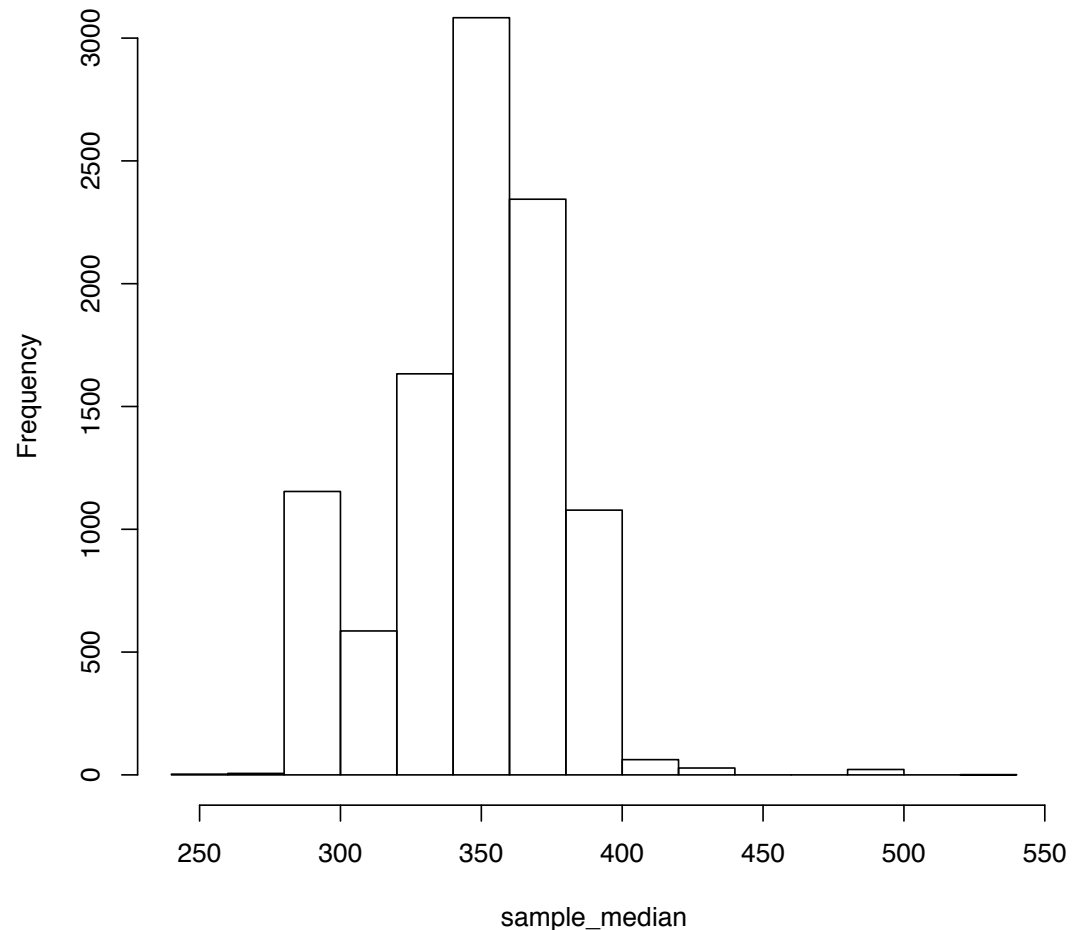
mean(Sample Median) =

348.0378

stderr(Sample Median) =

27.30539

Histogram of the Bootstrap sample medians



CEO salary example with larger $N = 59$

✱ The realized sample of CEO salary $N=59$, median=350 K

✱ $r = 10000$

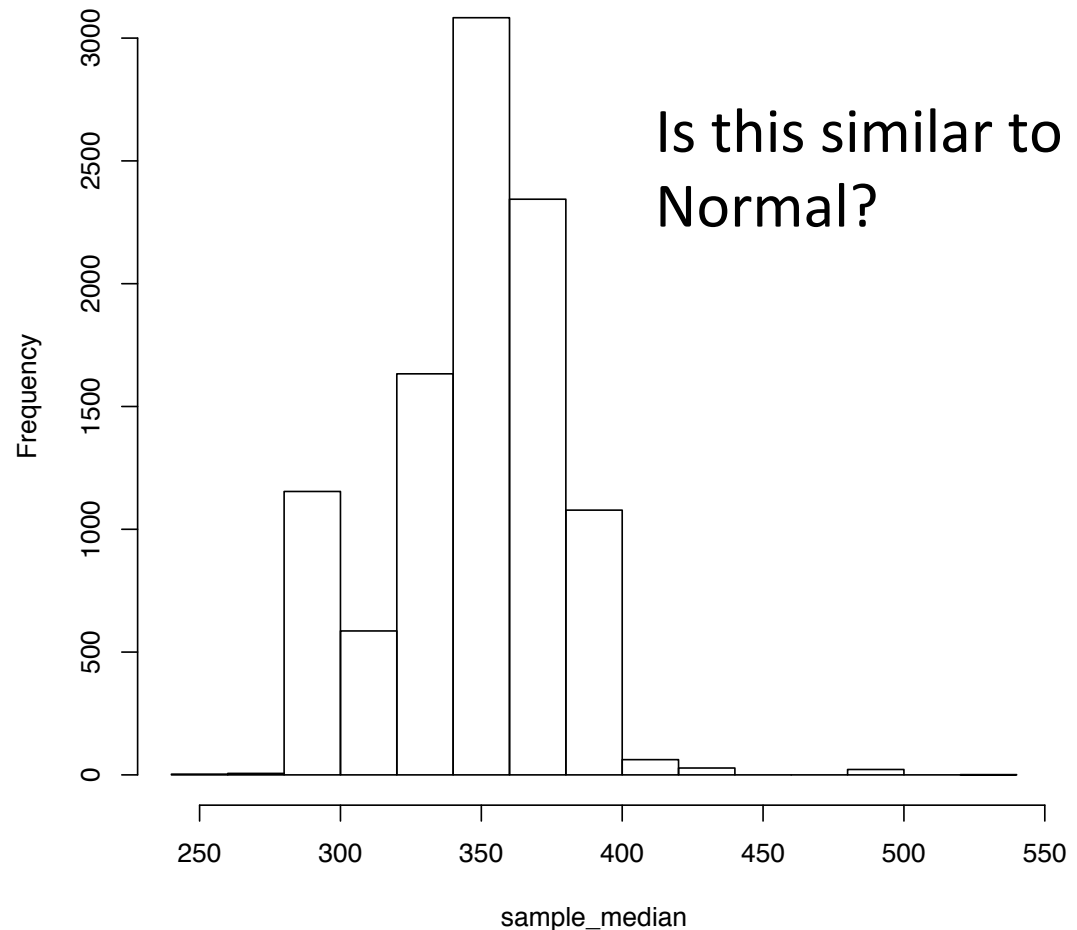
mean(Sample Median) =

348.0378

stderr(Sample Median) =

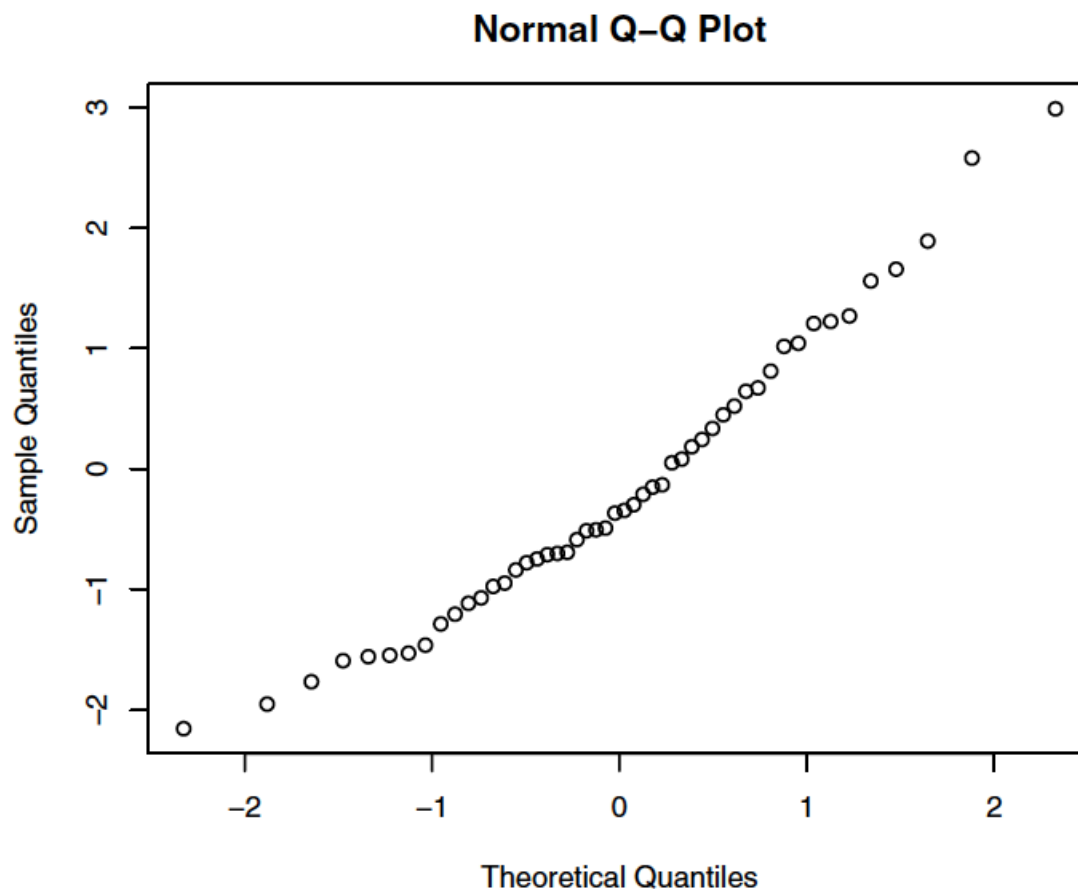
27.30539

Histogram of the Bootstrap sample medians



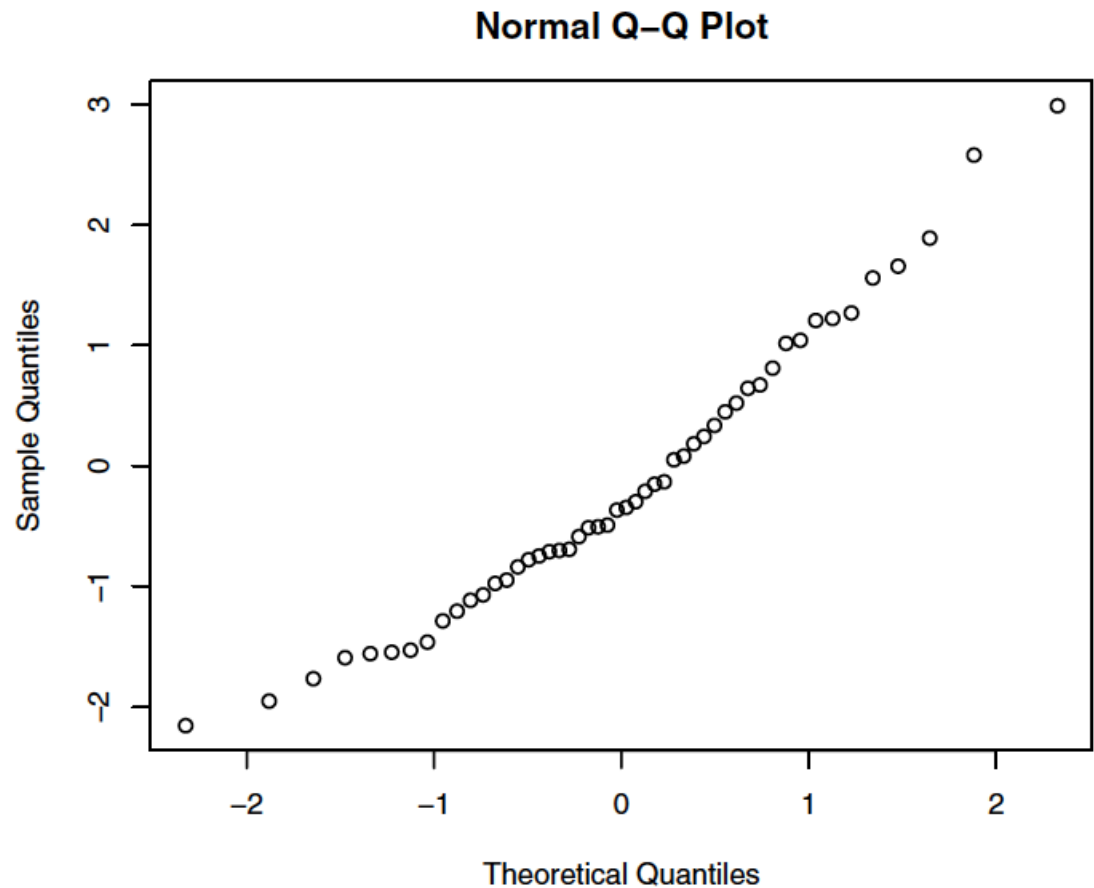
Checking whether it's normal by Normal Q-Q plot

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



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

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

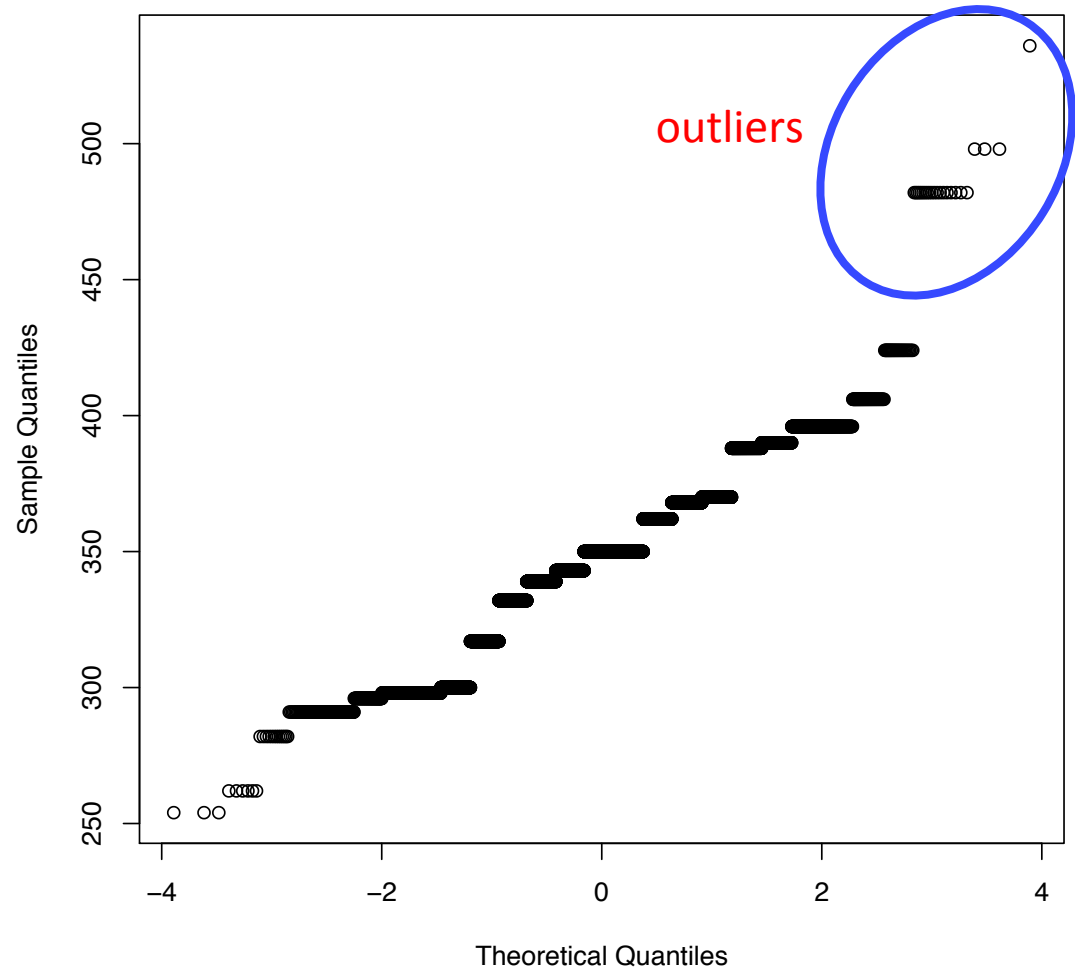


Normal Distribution's Quantile

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 Bootstrap Sample Median Q-Q Plot



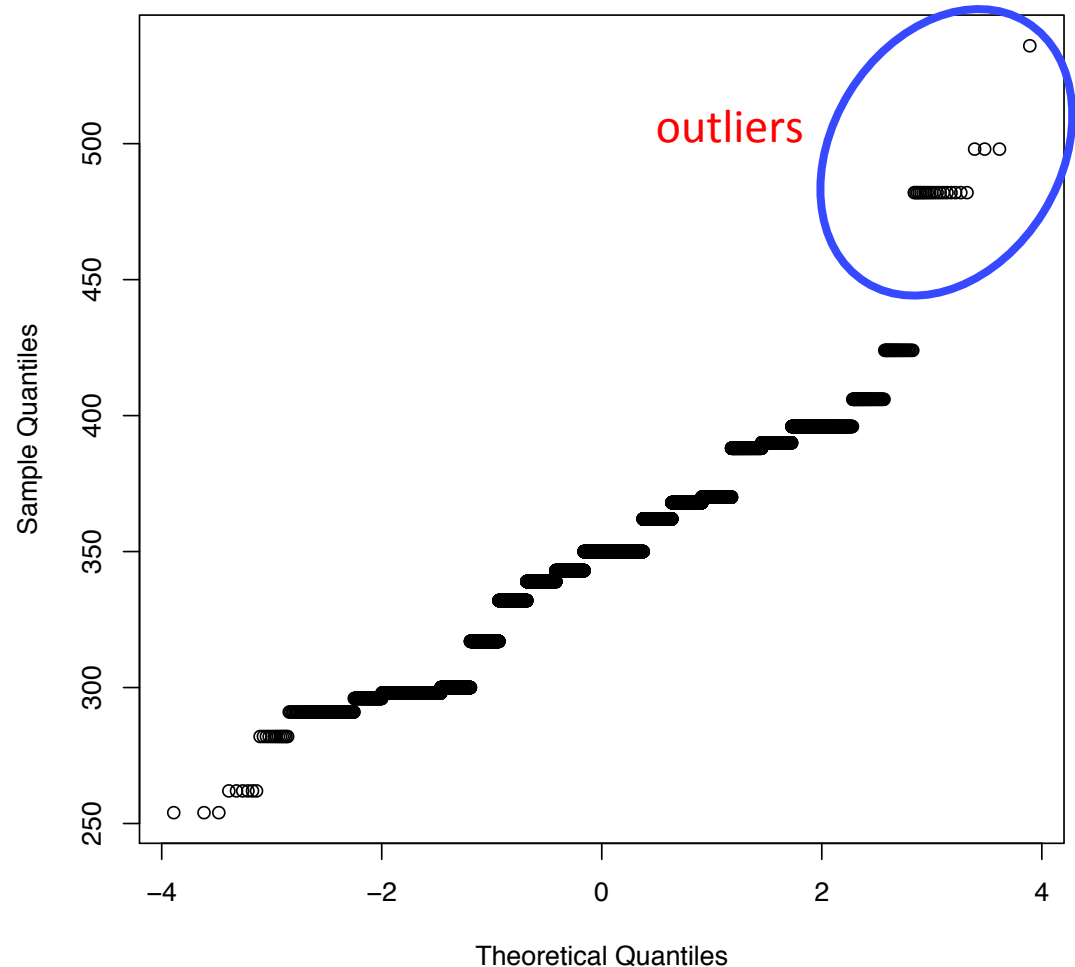
CEO salary sample median's Q-Q plot

✱ 95% confidence interval for the median CEO salary from the bootstrap simulation

✱ $348.0378 \pm 2 \times 27.30539$

$= [293.427, 402.6486]$

CEO Bootstrap Sample Median Q-Q Plot



Assignments

- ✱ Read Chapter 7 of the textbook
- ✱ Week 8 module on Compass
- ✱ 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!*

