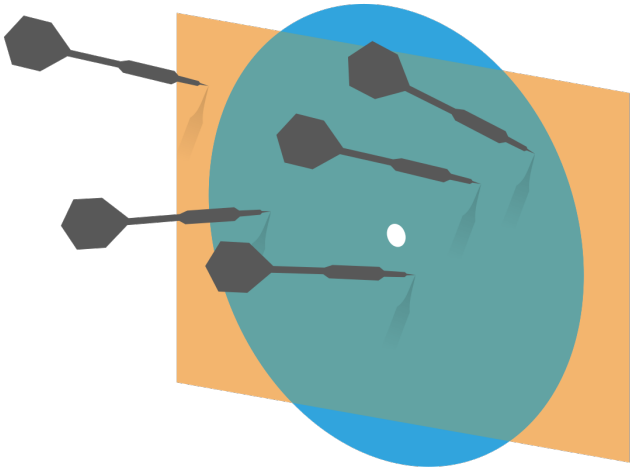
Probability and Statistics for Computer Science



Credit: wikipedia

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

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 3.03.2020

Last time

- Statistical inference overview
- Population and sample mean
- Confidence interval

Midterm 1 Mar. 5th Thurs.
in NHB 2079

A. Anxious
B. Not Anxious

→ Write about what you're stressed about to release the anxiety
→ Mock the exam with practice exam
→ Prepare the note sheet.

Contents

- Review of sample mean and more on confidence interval
- Bootstrap simulation of other sample statistic
- Hypothesis test

Q. Sample mean

- Among the students in CS361, I randomly sampled 10 students' grade for HW5 , and then averaged these numbers, the value I got is:

A. a population

B. a sample mean variable

C. a realized sample mean

D. a random sample

Q.

- Among the students in CS361, I randomly sampled 10 students' attendances, and then averaged these numbers, the value I got is:

A. a population

B. a sample mean random variable

C. a realized sample mean

D. a random sample

Calculate the standard error for sample mean

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

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

delta

- So the standard error is estimated as

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

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

Interpreting the standard error

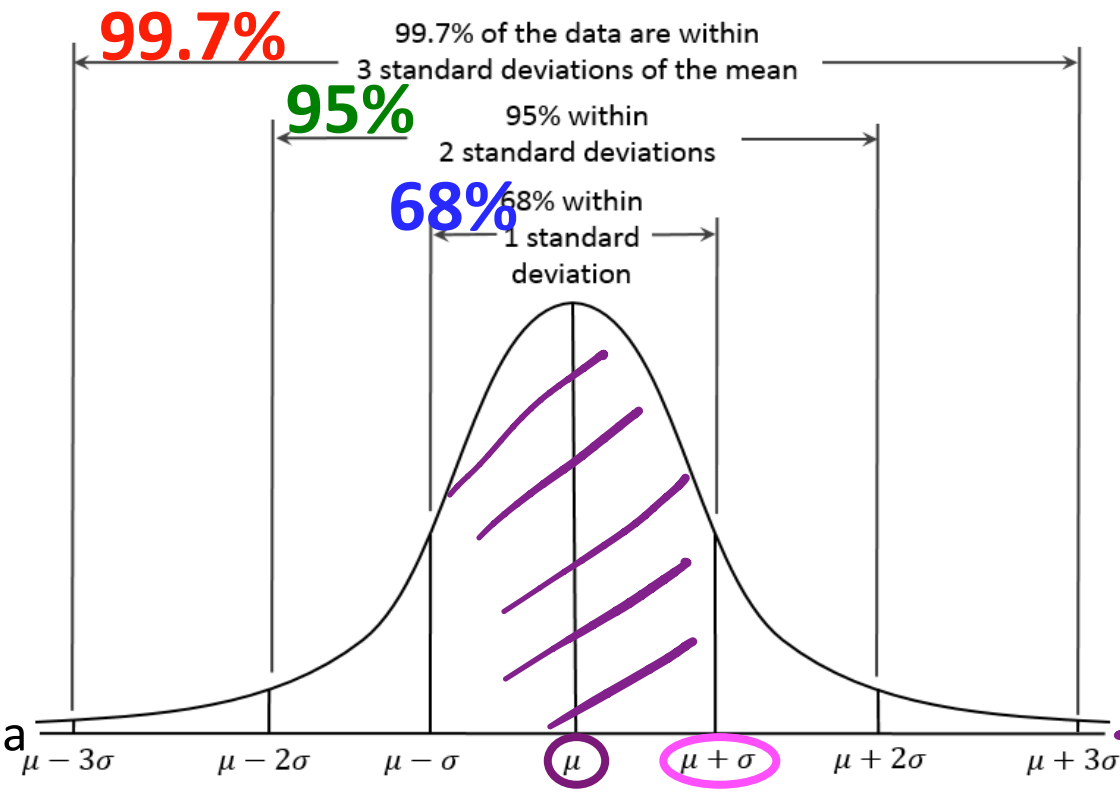
- **Sample mean** is a random variable and has its own probability distribution, stderr is sample mean's standard deviation
- When ***N*** is very large, according to the **Central Limit Theorem**, sample mean is approaching a normal distribution with

$$\mu = popmean(\{x\}) \ \& \ \sigma = stderr(\{x\})$$

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

Interpreting the standard error

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



Credit: wikipedia

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

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

- For about 99.7% of realized sample means

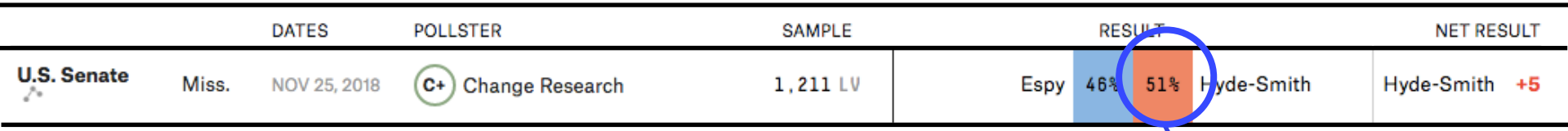
$$mean(\{x\}) - 3 \cdot stderr(\{x\}) \leq popmean(\{x\}) \leq mean(\{x\}) + 3 \cdot stderr(\{x\})$$

What does 95% Confidence interval mean?

- For 95% of the realized sample means, if you construct the confidence interval this way, the population mean will lie within such interval.

- It does **not** mean the population mean lies in the interval with probability 95%

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

$$stdunbiased(\{x\})$$

51%

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

$$stderr(\{x\}) \simeq \frac{0.5}{\sqrt{1211}} \simeq 0.0144$$

✺ So we estimate the population mean as 51% with stderr 1.44%

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%]

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

- If samples are taken from normal distributed population, the following statistic is from 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:

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

if we know N-1 data points, the last one will be fixed.

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

T is data from

$$S = \frac{X^{(N)} - \text{Popmean}(\{x\})}{\text{stderr}(\{x\})}$$

→ Random variable $X^{(N)}$ standardized

if x_i is from $\mathcal{N}(\mu, \sigma)$

μ → popmean

σ → popsd

$$S \sim t(N-1)$$

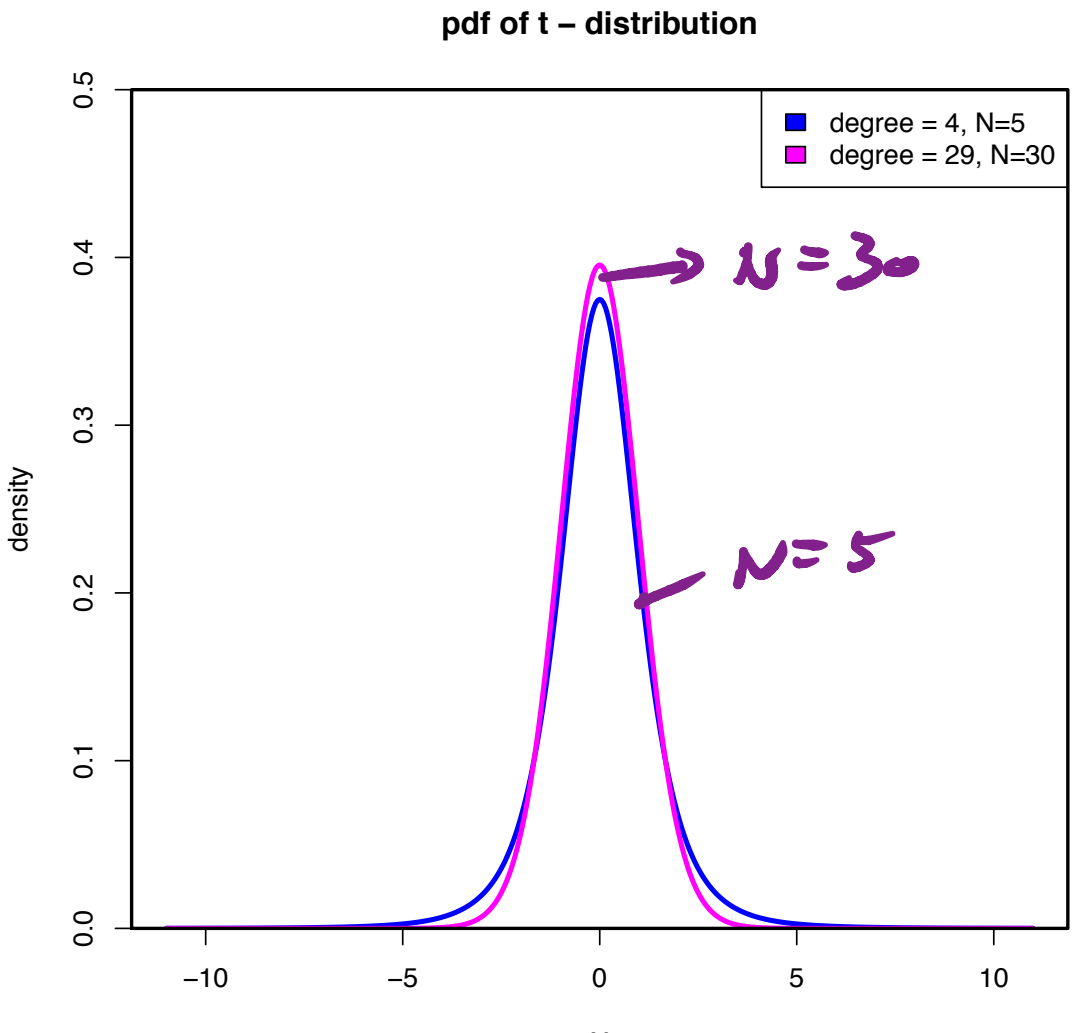
↑ t

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

t-distribution with N=5 and N=30

Credit : wikipedia

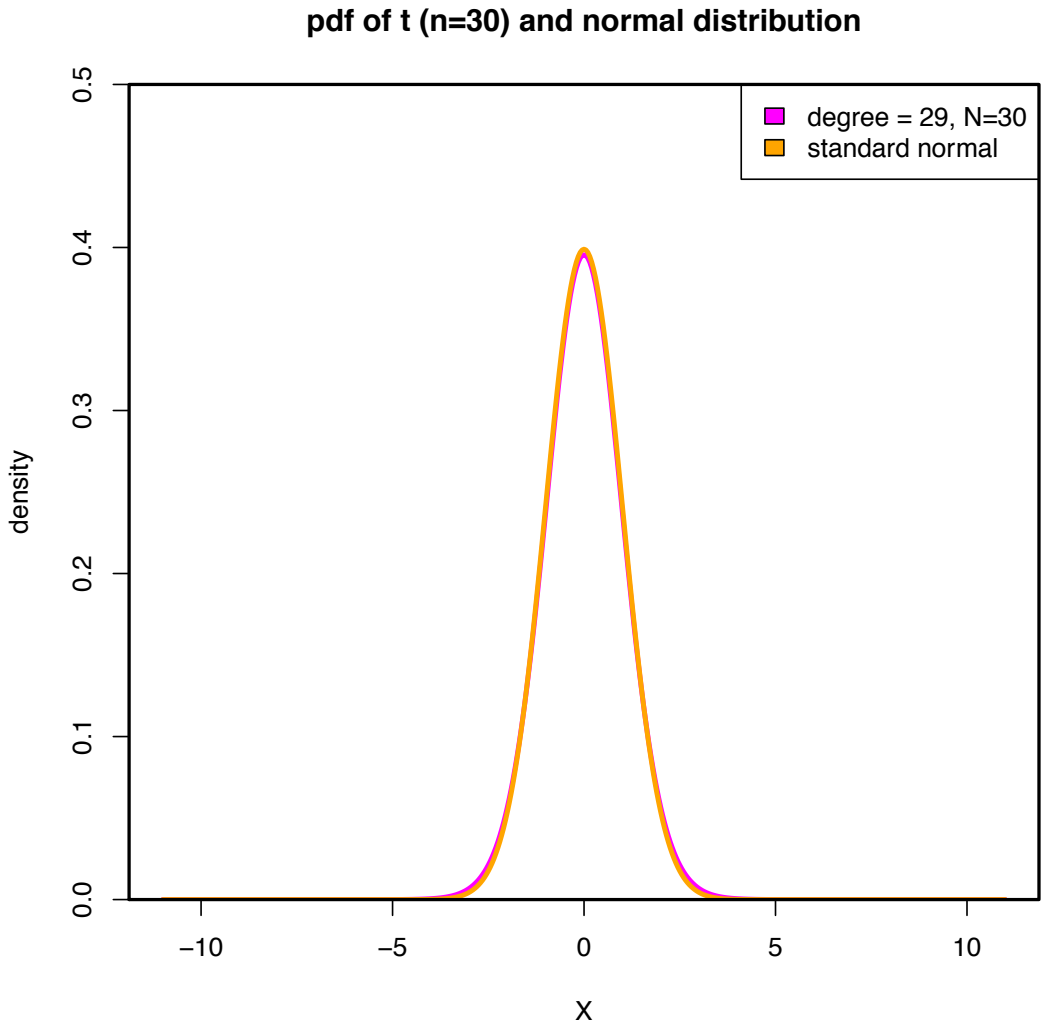
William Sealy Gosset 1876-1937



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

Contents

- Review of sample mean and more on confidence interval
- **Bootstrap simulation of any sample statistic**
- Hypothesis test

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 largest value 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

realized

  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

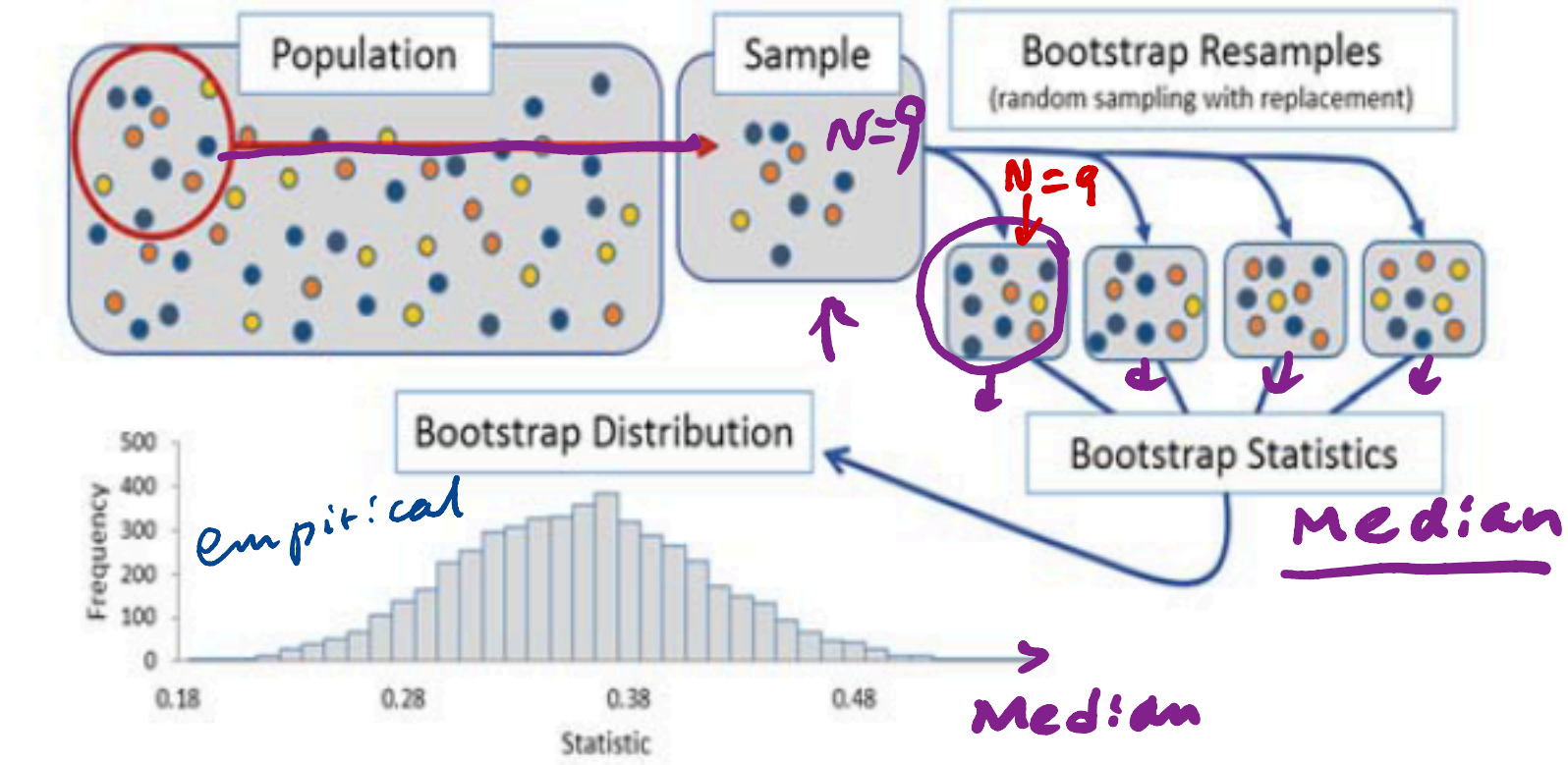


Figure 1. Summary of Bootstrapping Process

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 {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~~ value 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
…	…

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)

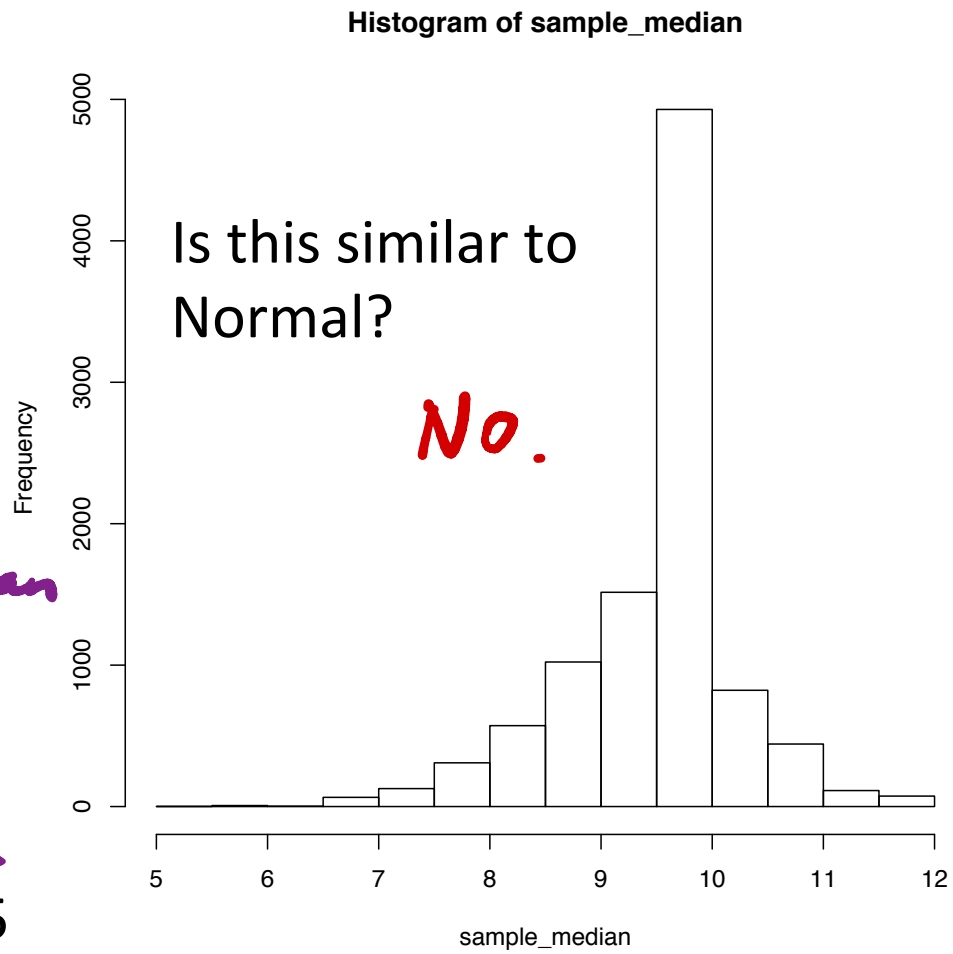
Bootstrap replicate	**Sample median**
{11, 11, 12, 10, 10, 10, 12, 10, 7, 10}	10
{7, 10, 10, 10, 9, 7, 9, 10, 12, 10}	10
{9, 7, 10, 8, 5, 10, 7, 10, 12, 8}	8.5
…	…

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.

stderr of (sample median)

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



mean(Sample Median) = 9.73625

stderr(Sample Median) = 0.7724446

Errors in Bootstrapping

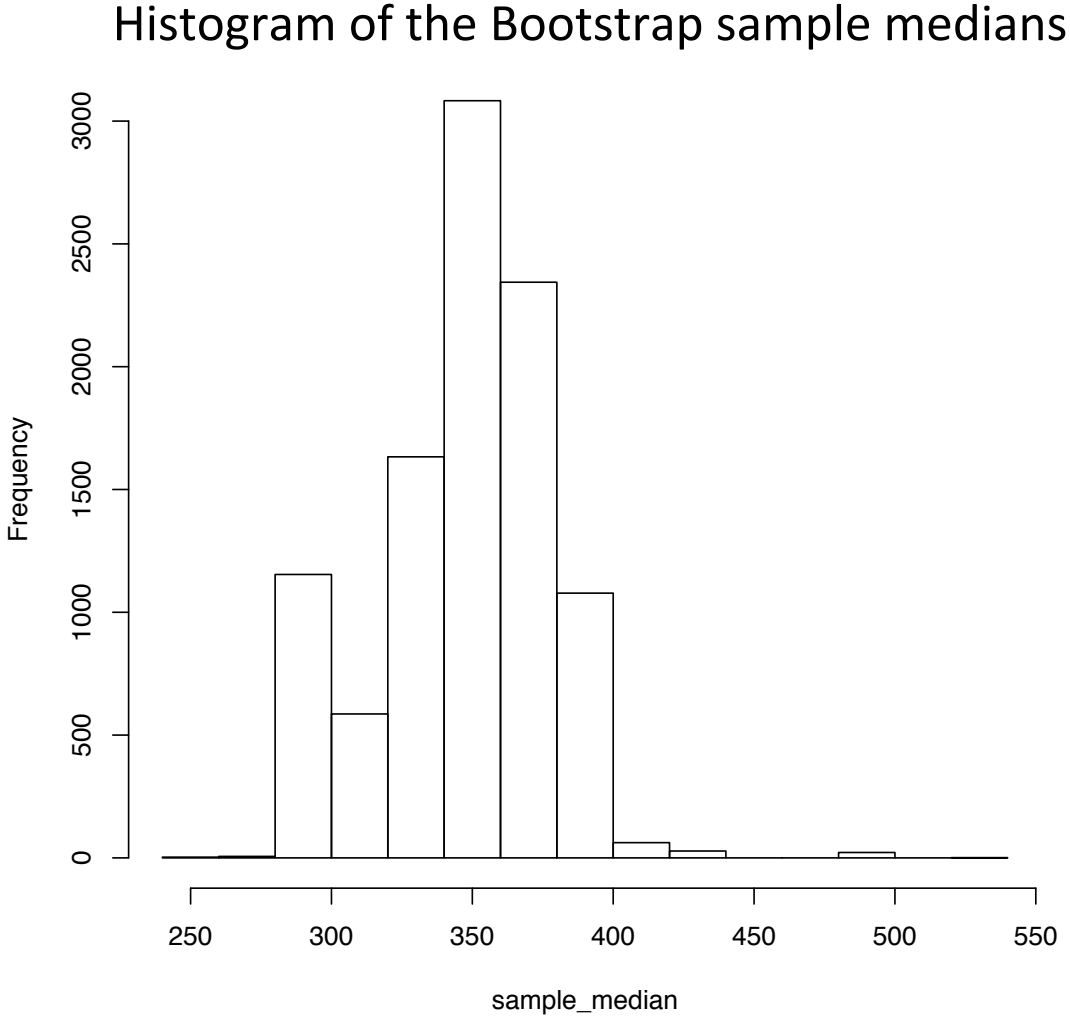
- The distribution simulated from bootstrapping is called empirical distribution. It is not the true population distribution. **There is a statistical error**.
- The number of bootstrapping replicates may not be enough. **There is a numerical error**.
- When the statistic is not a well behaving one,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

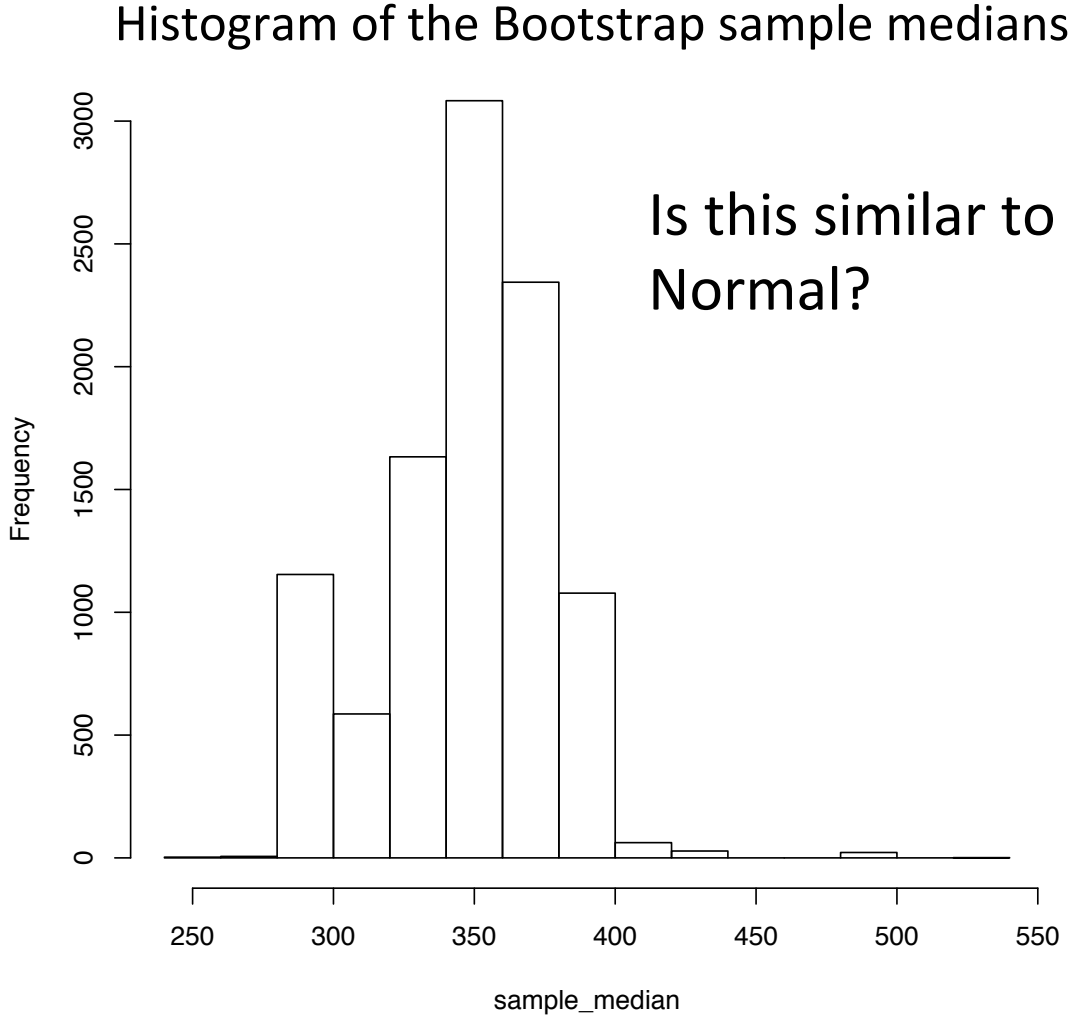
stderr(Sample Median) = **27.30539**



CEO salary example with larger N = 59

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

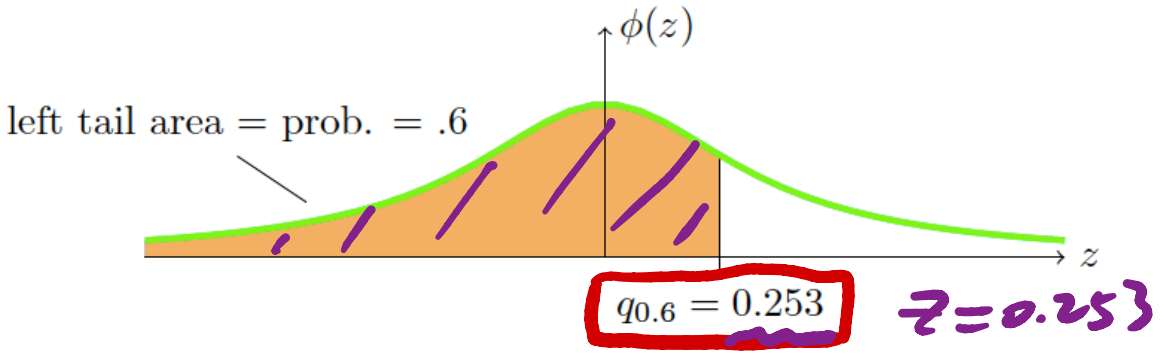
mean(Sample Median) = **348.0378**
stderr(Sample Median) = **27.30539**



Quantile

- Quantiles give a measure of location, the median is the 0.5 quantile

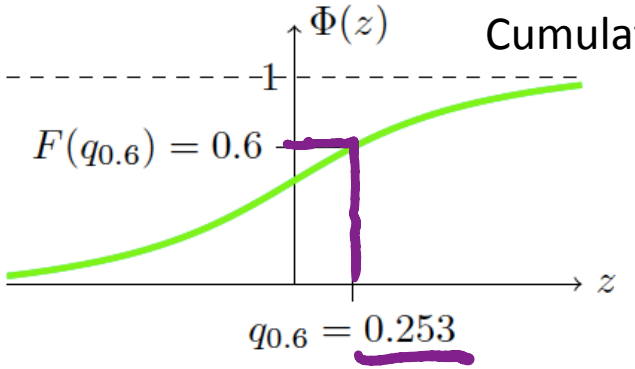
Probability density



$z = 0.253$

$q_{0.6} = 0.253$

Cumulative Probability (CDF)



Credit:
J. Orloff et al

$q_{0.6}$: left tail area $= 0.6 \Leftrightarrow F(q_{0.6}) = 0.6$

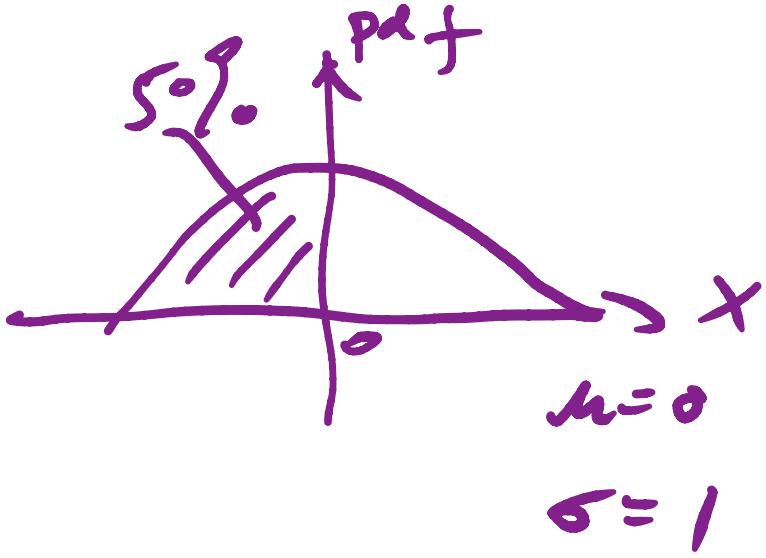
Q.

- What is the value of 50% quantile in a standard normal distribution?

A. -1

B. 0

C. 1



Q.

- What is the value of 50% quantile in a standard normal distribution?

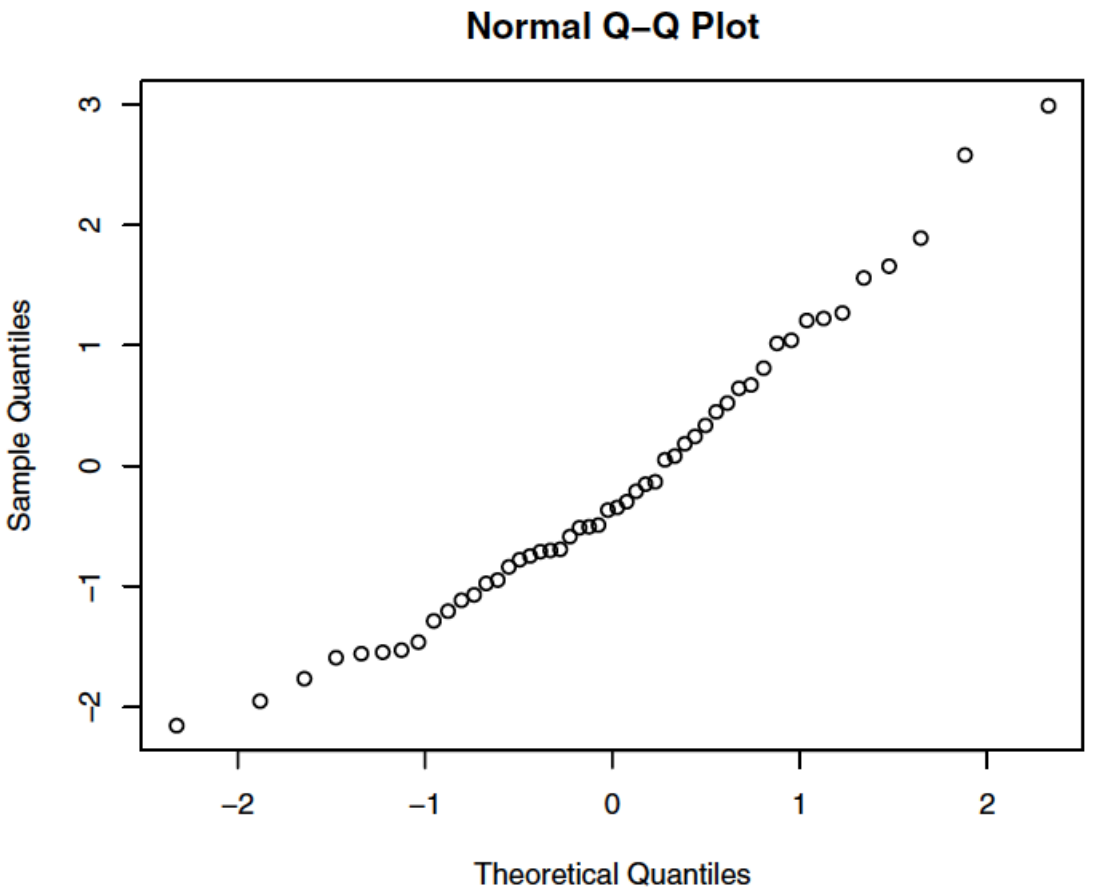
A. -1

B. 0

C. 1

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

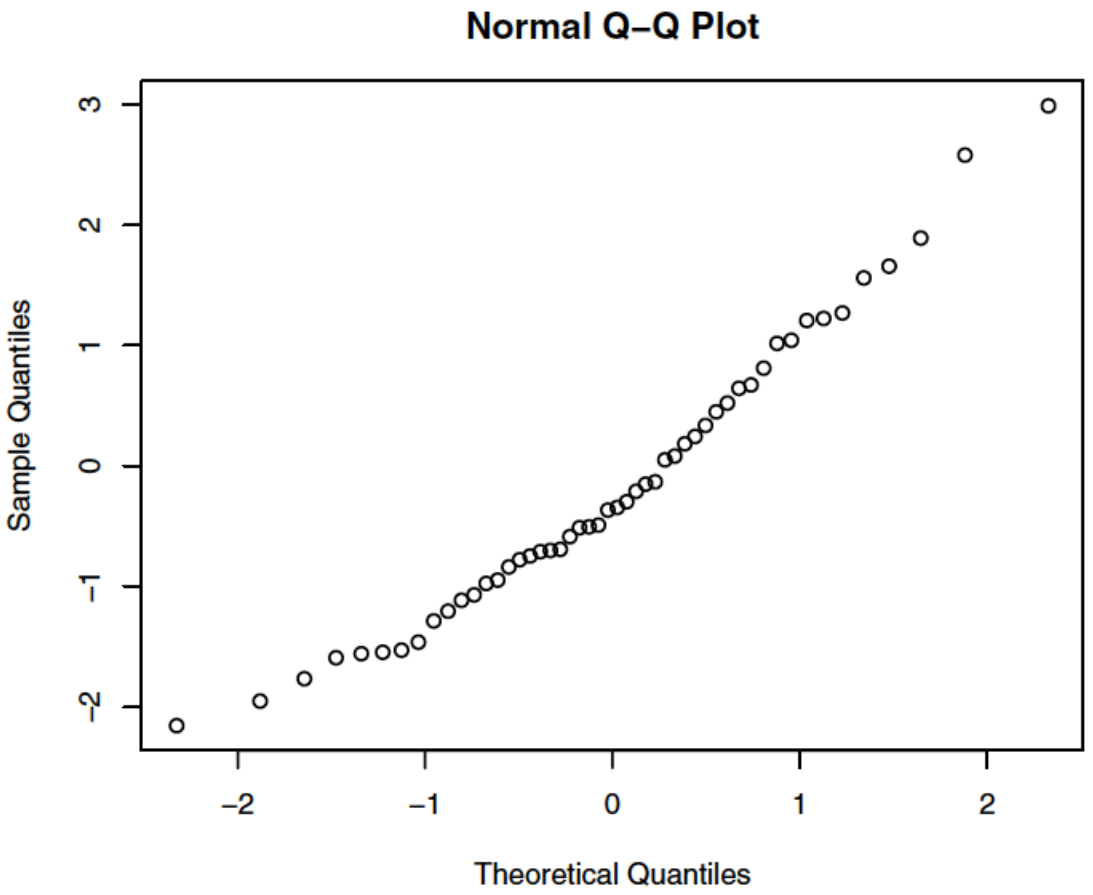
- Q-Q compares a distribution with normal by matching the kth smallest quantile value pairs and plot as a point in the graph
- **Linear means similar to normal!**



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!**

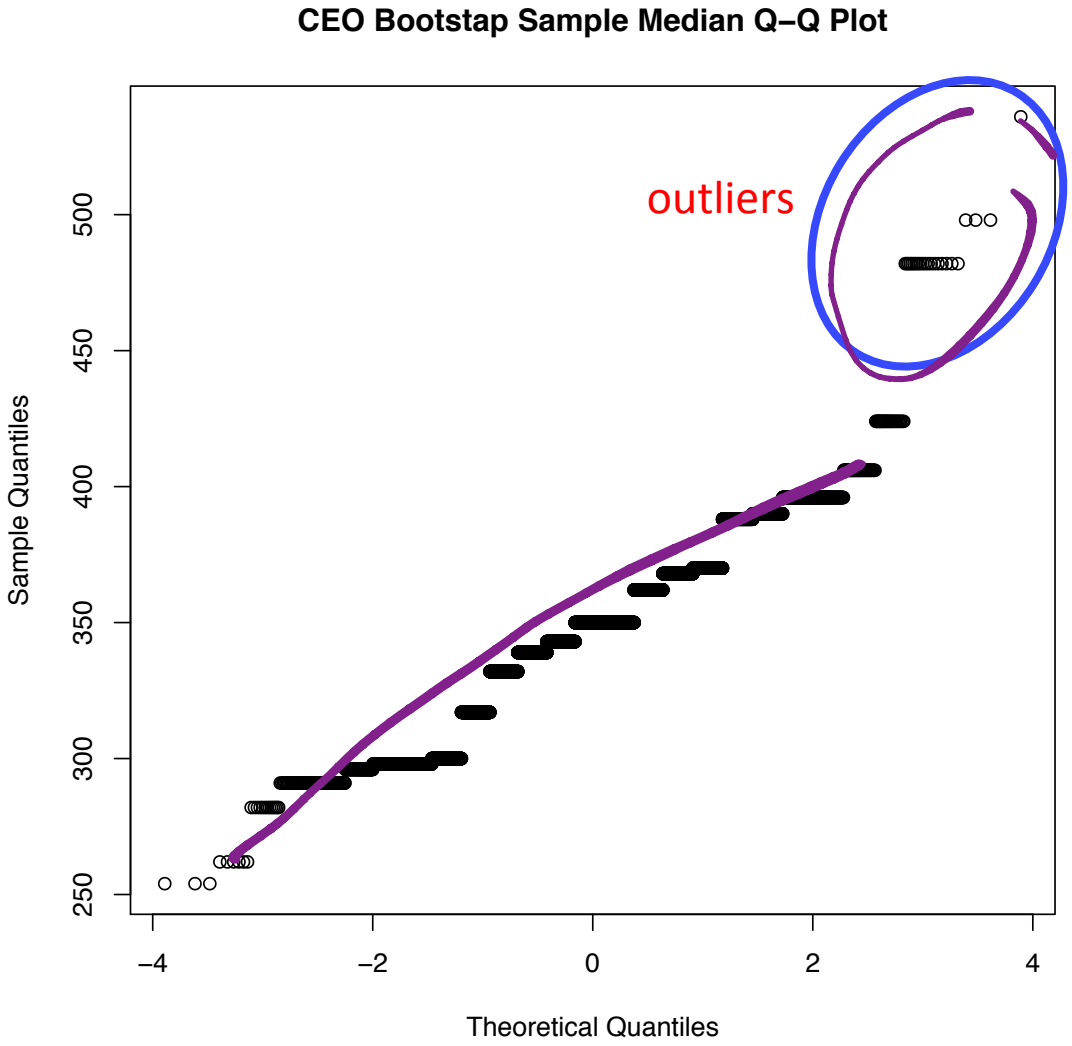


Normal Distribution's Quantile

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

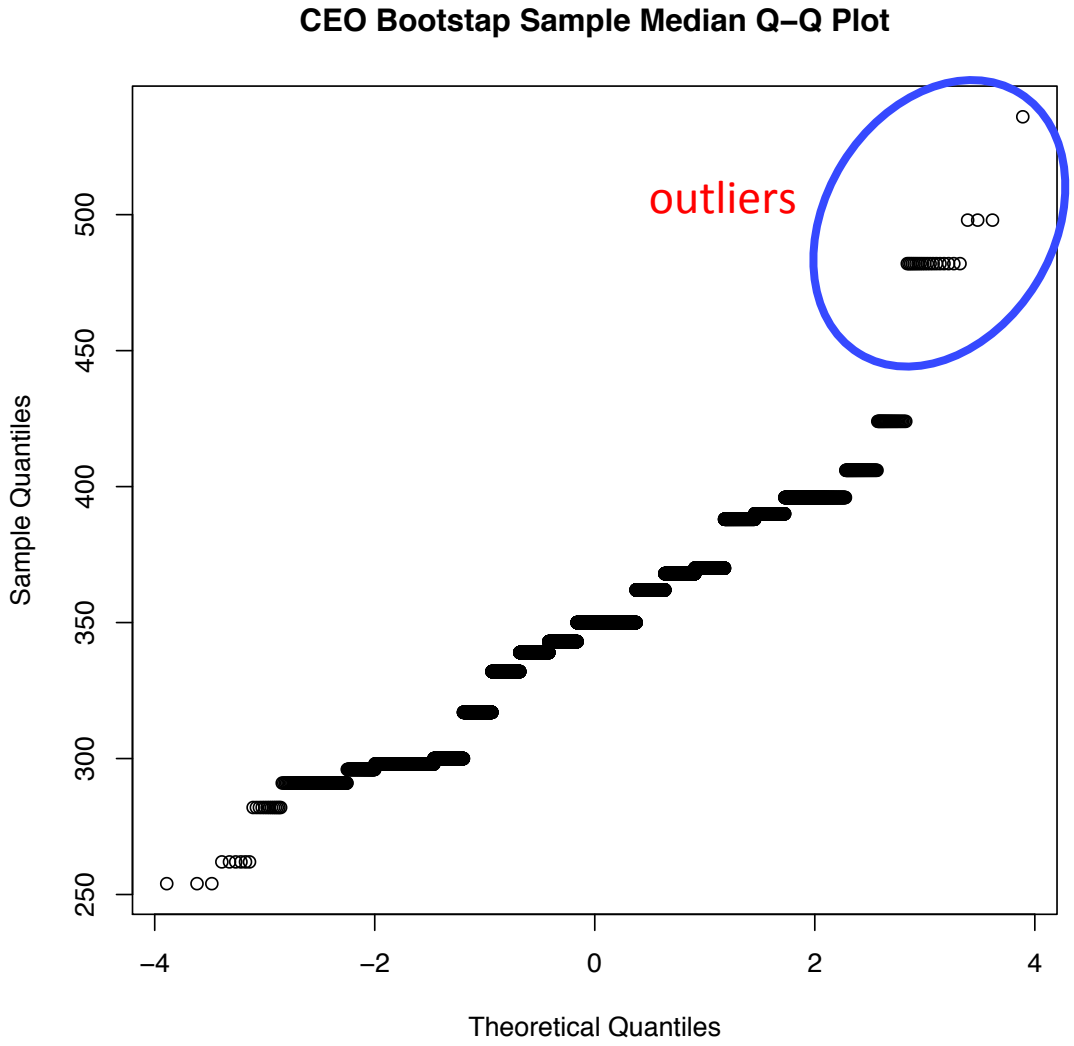


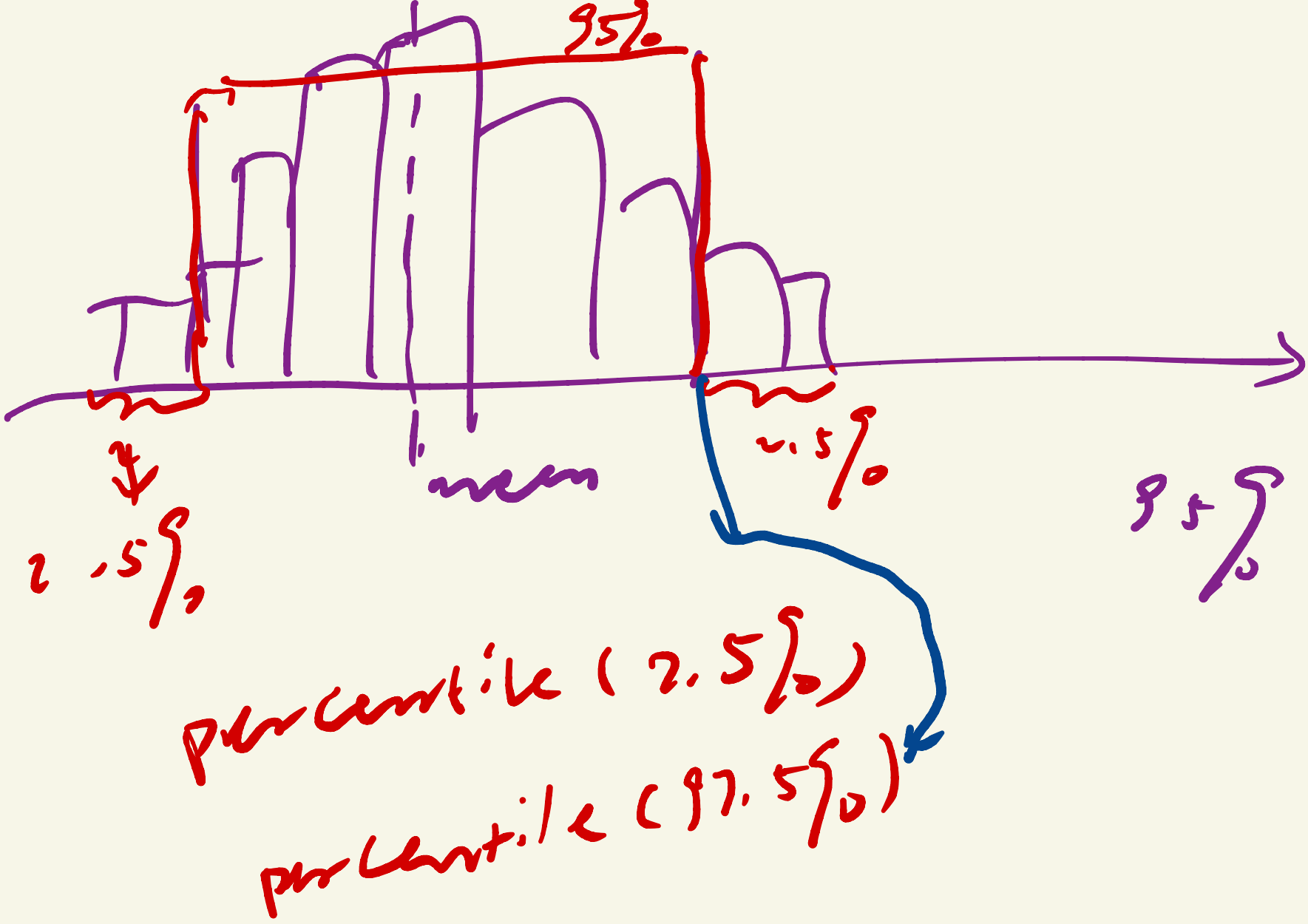
CEO salary sample median's Q-Q plot

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

- **348.0378± 2×27.30539**

= [293.427, 402.6486]




95%
2.5%
mean
2.5%
95%
percentile (2.5%)
percentile (97.5%)

Contents

- Review of sample mean and confidence interval
- Bootstrap simulation of other sample statistic
- **Hypothesis test**

Hypothesis in Science

- When doing science we form hypothesis first
- Then we design experiment and collect/ generate data
- Then we test whether the hypothesis is true given the data, **reject** it if it contradicts the data significantly

A hypothesis

- Ms. Smith's vote percentage is 55%

  This is what we want to test, often called null hypothesis H_0

		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%

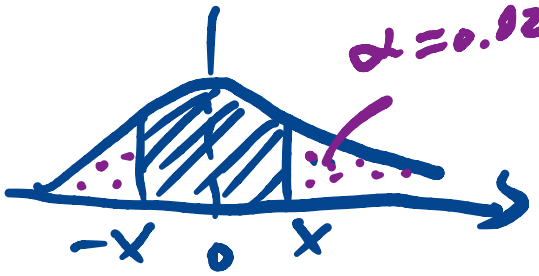
- Should we reject this hypothesis given the poll data?

Fraction of "less extreme" samples

- Assuming the hypothesis H_0 is true
- Define a test statistic

$$x = \frac{(sample\ mean) - (hypothesized\ value)}{standard\ error}$$

- Since *N*>30, x should come from a standard normal
- So, the fraction of "less extreme" samples is:



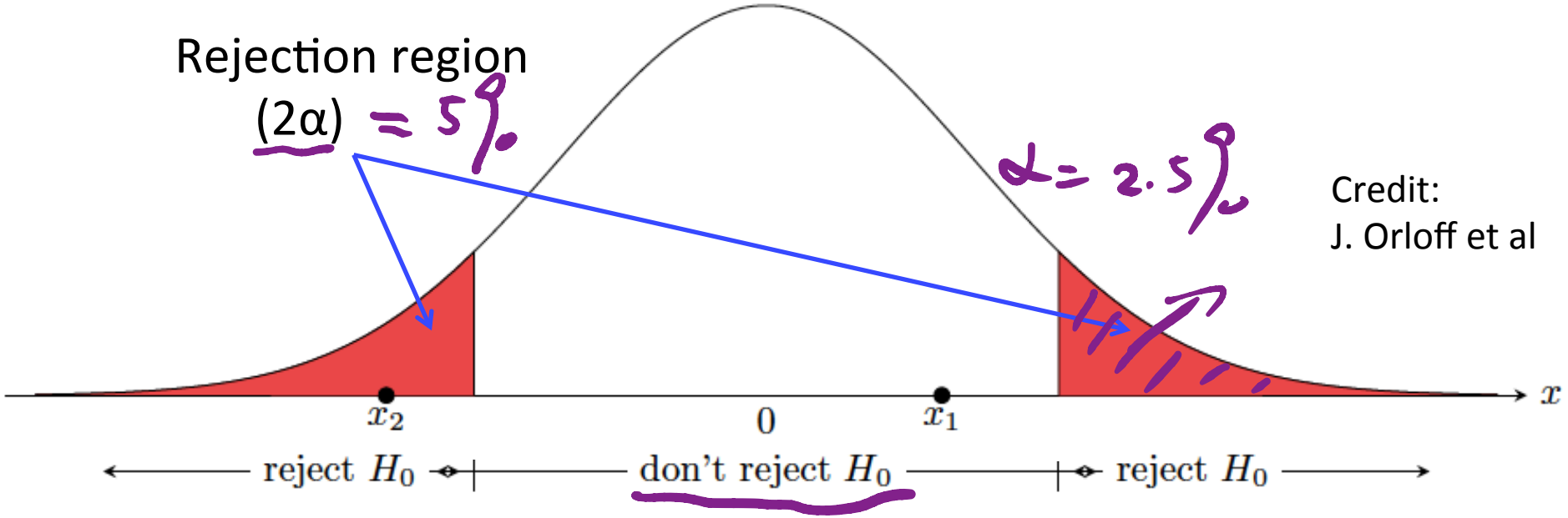
$$f = \frac{1}{\sqrt{2\pi}} \int_{-|x|}^{|x|} exp(-\frac{u^2}{2}) du$$

Rejection region of null hypothesis H_o

- Assuming the hypothesis H_0 is true
- Define a test statistic

$$x = \frac{(sample\ mean) - (hypothesized\ value)}{standard\ error}$$

- Since *N*>30, x should come from a standard normal



Pivoting

- If x is in the interval $\mu \pm c$, then μ is in the interval $x \pm c$

$$-c \leq \mu - x \leq c \quad \Leftrightarrow \quad -c \leq x - \mu \leq c$$

P-value: Rejection region- "The extreme fraction"

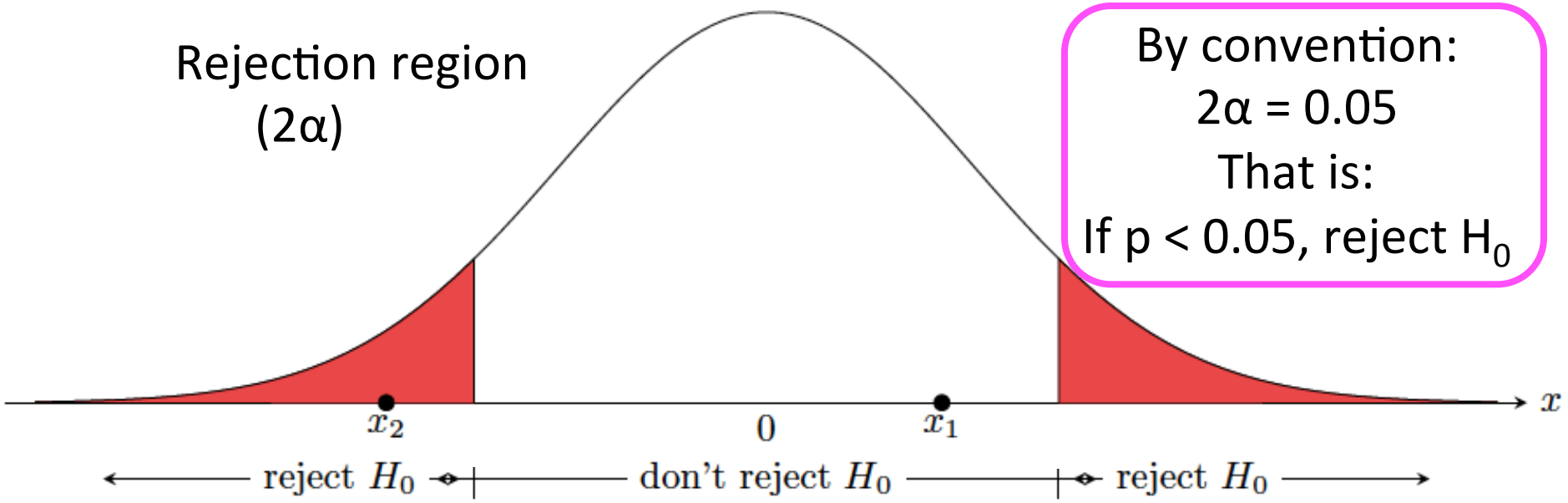
- It is conventional to report the p-value of a hypothesis test

$$p = 1 - f = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|x|}^{|x|} exp(-\frac{u^2}{2}) du$$

- Since *N*>30, x should come from a standard normal

Rejection region
(2α)

By convention:
2α = 0.05
That is:
If p < 0.05, reject H_0



p-value: election polling

- H_0: Ms. Smith's vote percentage is 55%
- The sample mean is 51% and stderr is 1.44%
- The test statistic $x = \frac{51 - 55}{1.44} = -2.7778$
- And the p-value for the test is:

$$p = 1 - \frac{1}{\sqrt{2\pi}} \int_{-2.7778}^{2.7778} exp(-\frac{u^2}{2}) du = 0.00547 < 0.05$$

- So we reject the hypothesis

Hypothesis test if N < 30

- Q: what distribution should we use to test the hypothesis of sample mean if N<30?
  A. Normal distribution
  B. t-distribution with degree =30
  C. t-distribution with degree = N
  D. t-distribution with degree = N-1

Hypothesis test if N < 30

- Q: what distribution should we use to test the hypothesis of sample mean if N<30?
  A. Normal distribution
  B. t-distribution with degree =30
  C. t-distribution with degree = N
  D. t-distribution with degree = N-1

The use and misuse of p-value

- p-value use in scientific practice
  - Usually used to reject the null hypothesis that the data is random noise
  - Common practice is $p < 0.05$ is considered significant evidence for something interesting
- Caution about p-value hacking
  - Rejecting the null hypothesis doesn't mean the alternative is true
  - $P < 0.05$ is arbitrary and often is not enough for controlling false positive phenomenon

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!