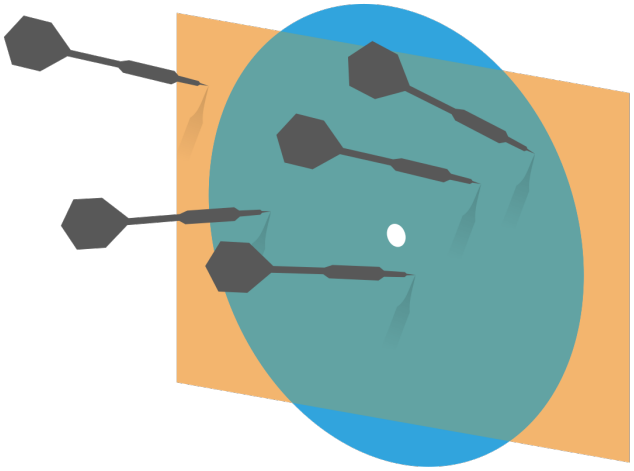
Probability and Statistics for Computer Science



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

Objectives

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

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

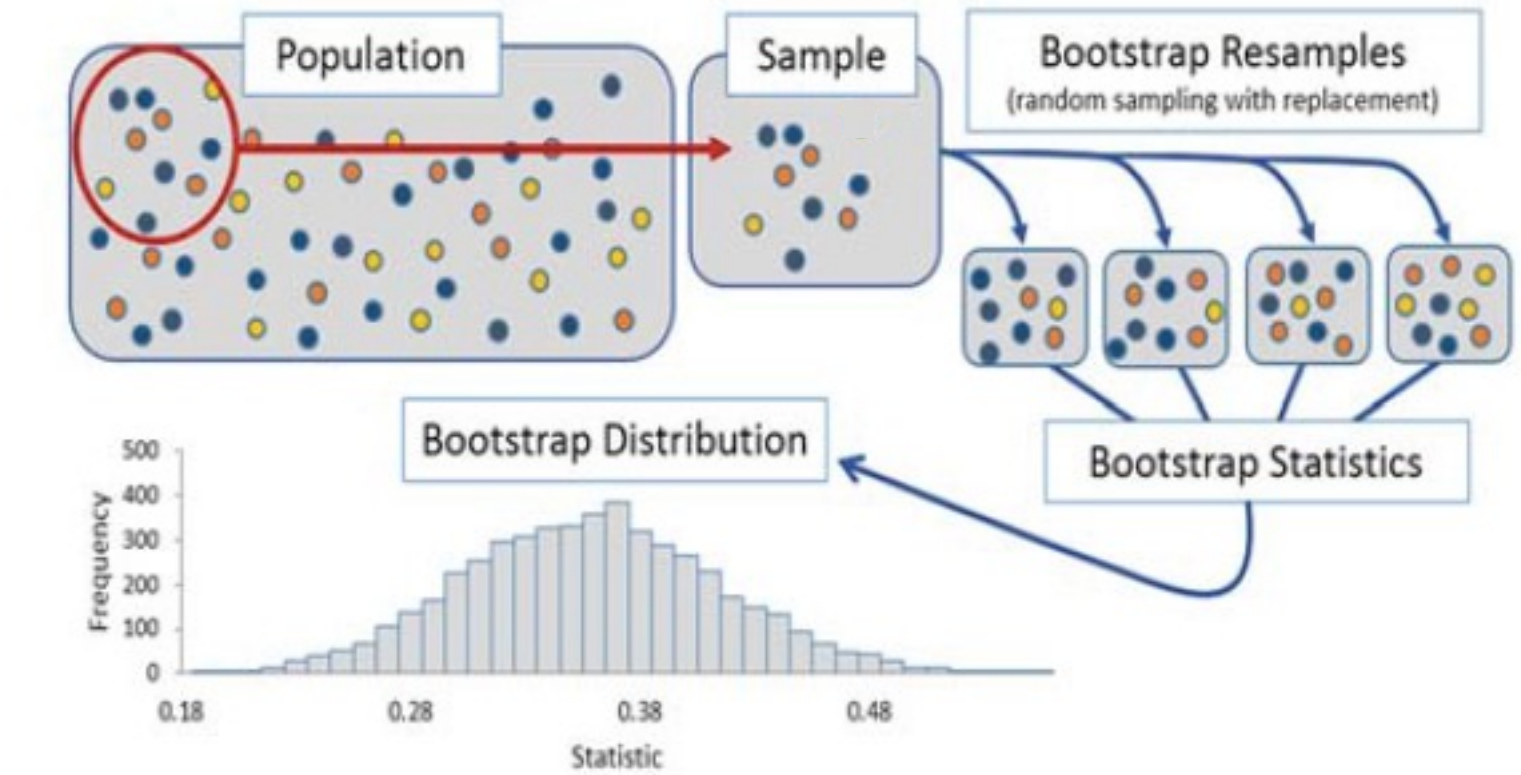


Figure 1. Summary of Bootstrapping Process

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 {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

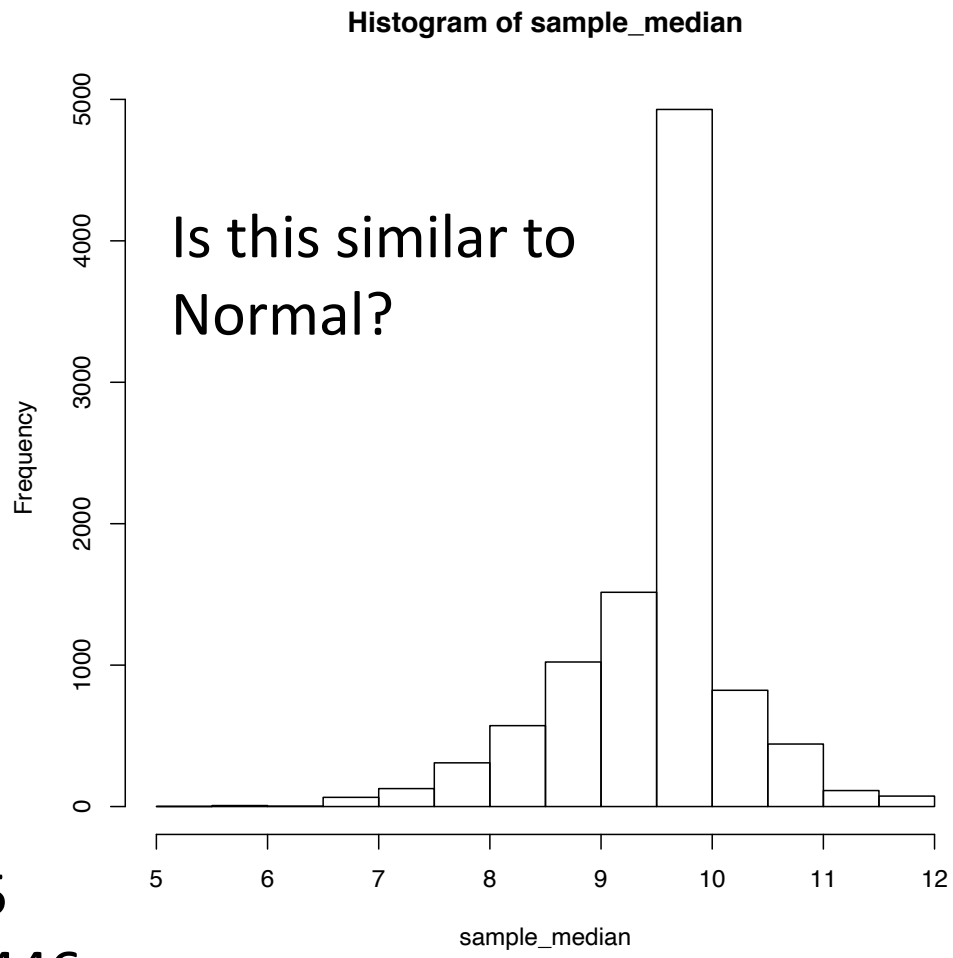
- Do the bootstrapping for **r = 10000** times, then draw the histogram and also find the stderr of sample median)

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.

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



mean(Sample Median) = 9.73625

stderr(Sample Median) = 0.7724446

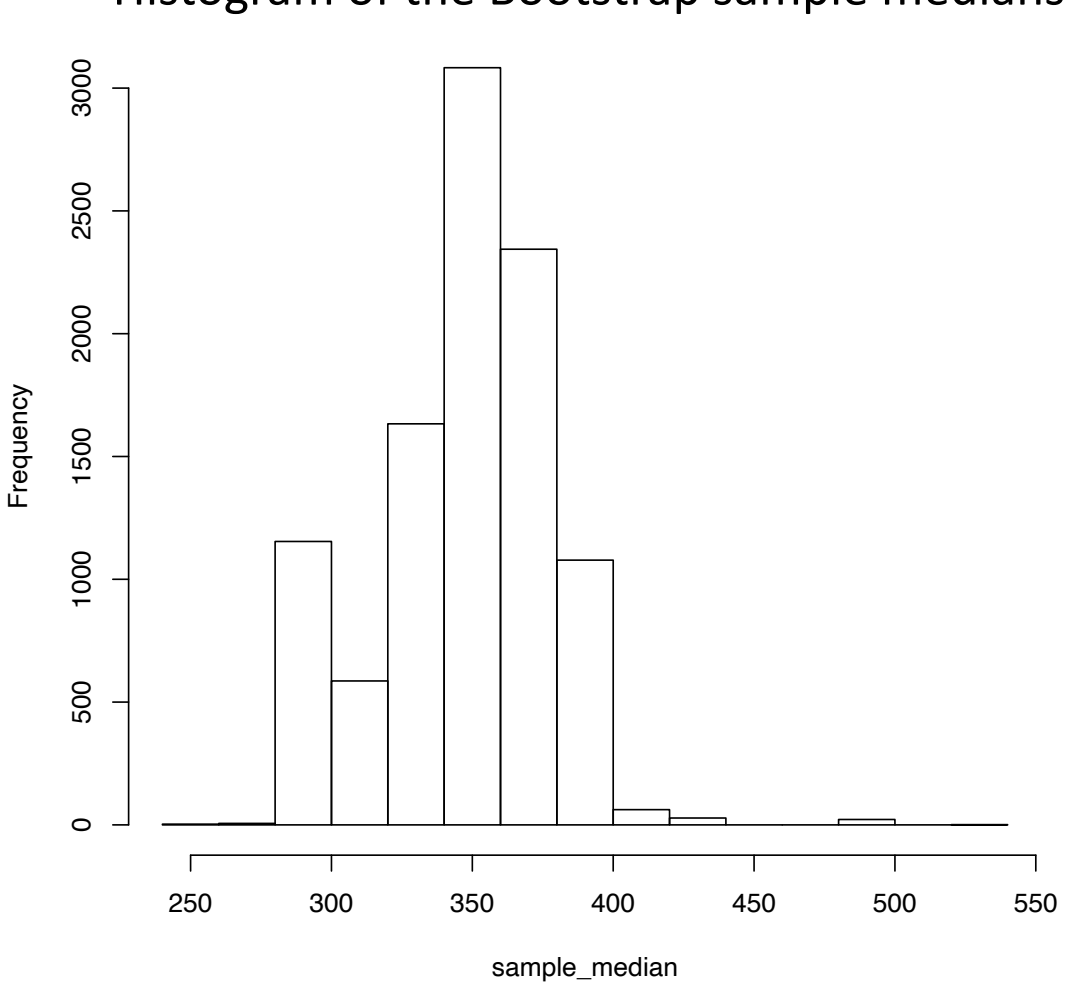
Errors in Bootstrapping

- 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

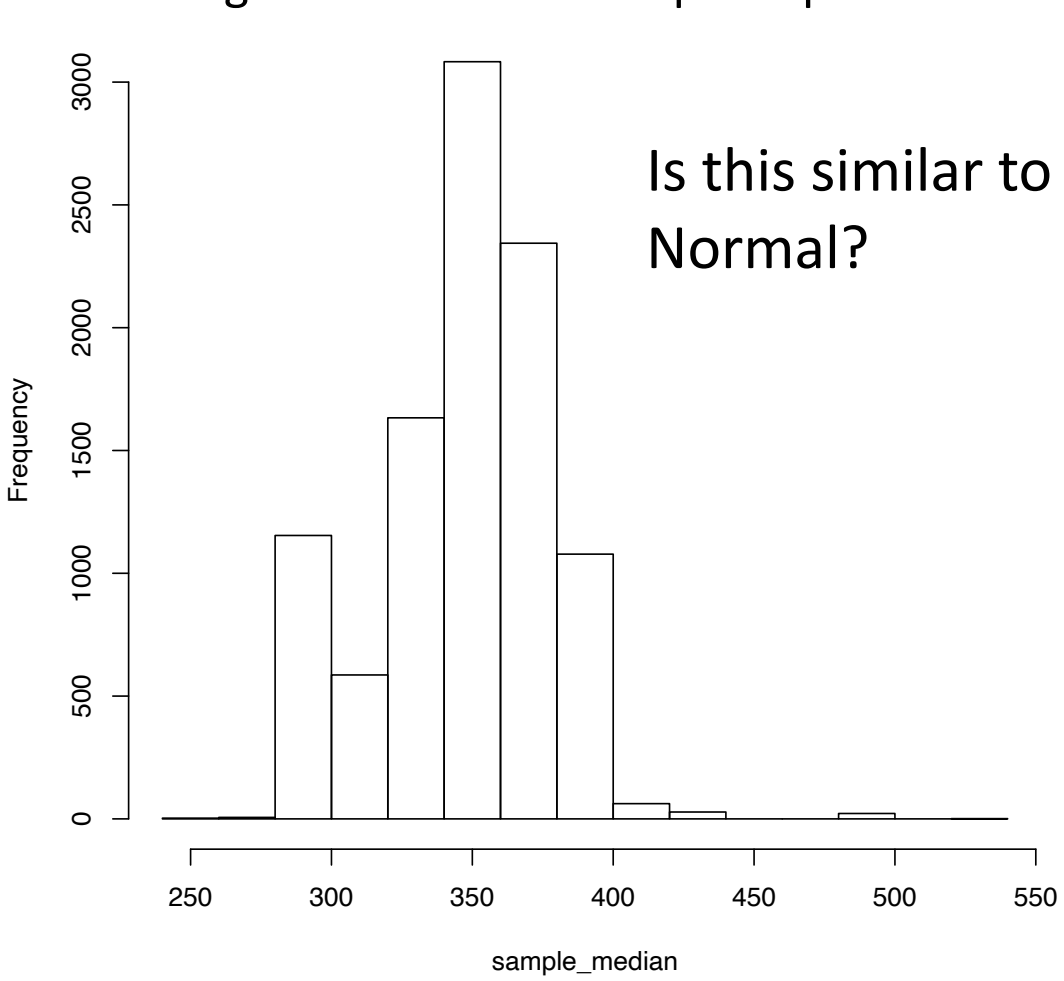
mean(Sample Median) = **348.0378**
stderr(Sample Median) = **27.30539**



CEO salary example with larger N = 59

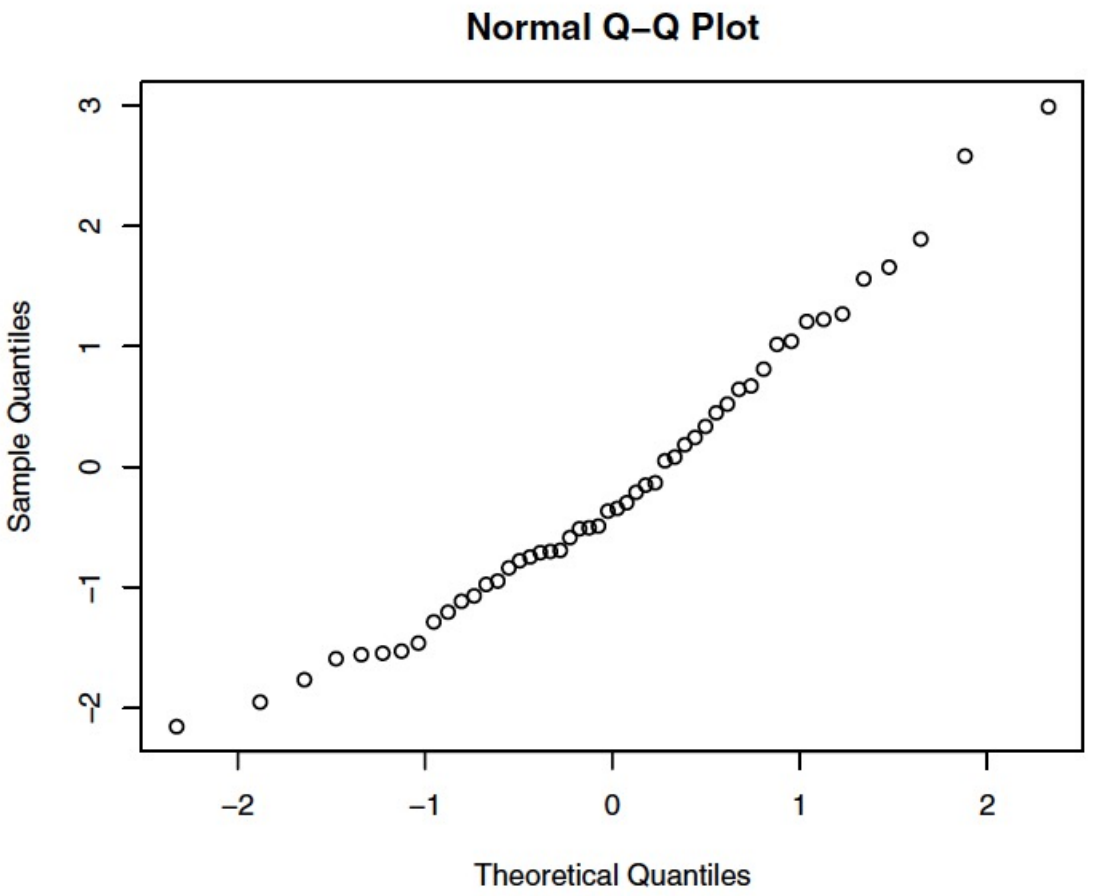
- The realized sample of CEO salary *N*=59, median=350 K
- r = 10000

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



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

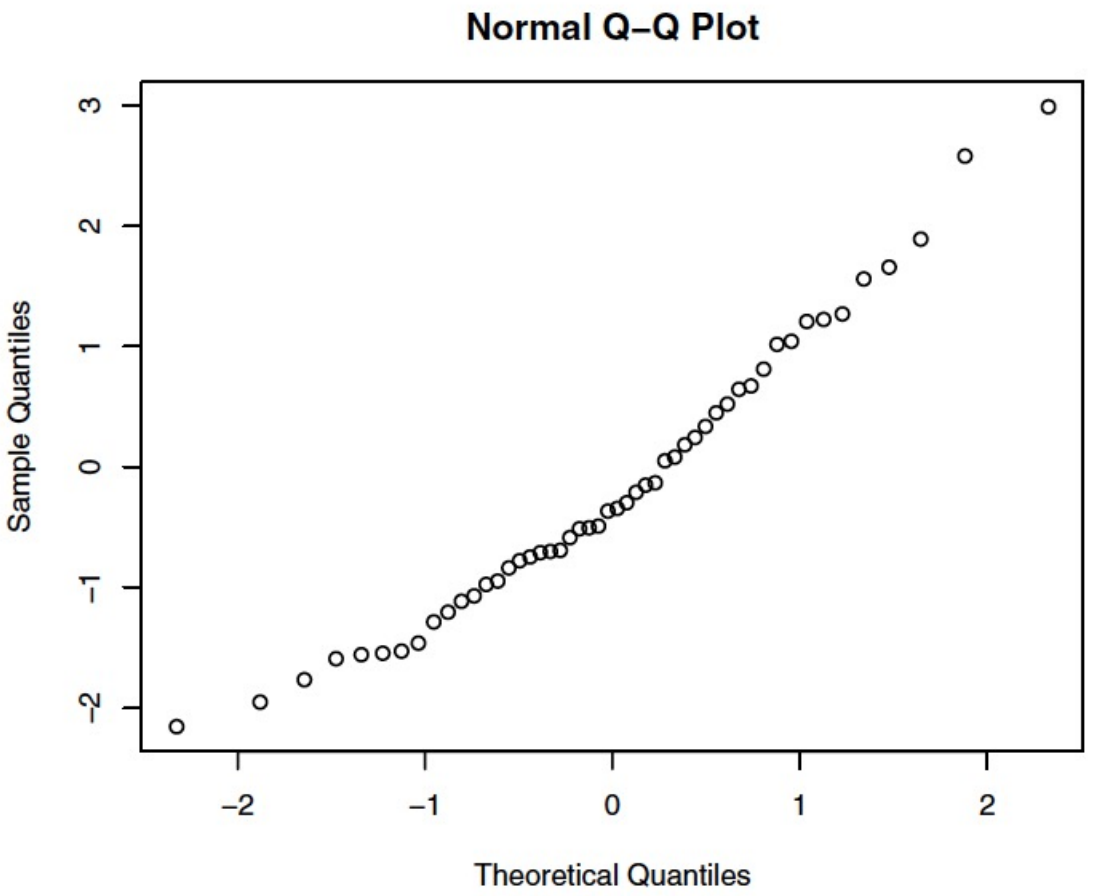
- Q-Q compares a distribution with normal by matching the kth smallest quantile value pairs and plot as a point in the graph
- **Linear means similar to normal!**



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!**

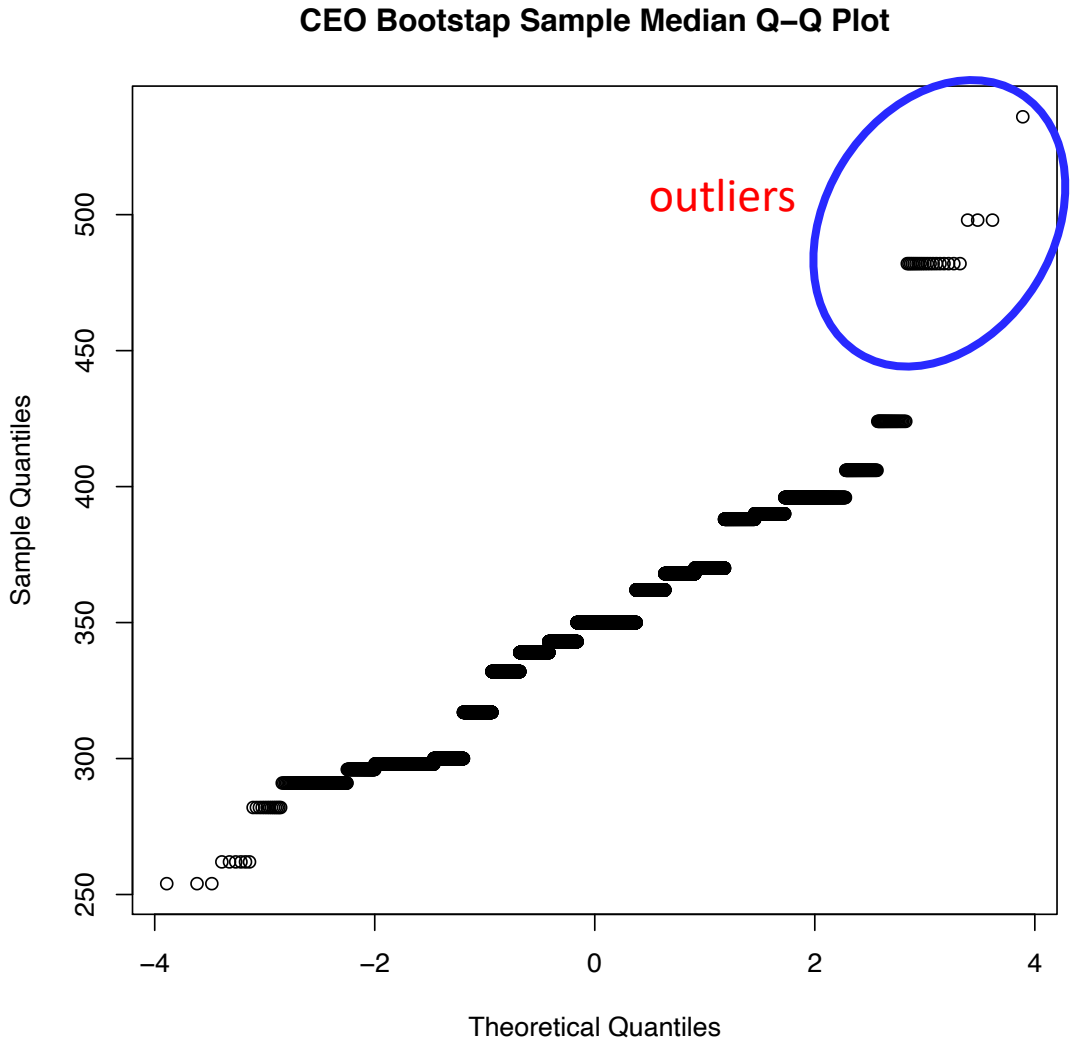


Normal Distribution's Quantile

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

CEO salary sample median's Q-Q plot

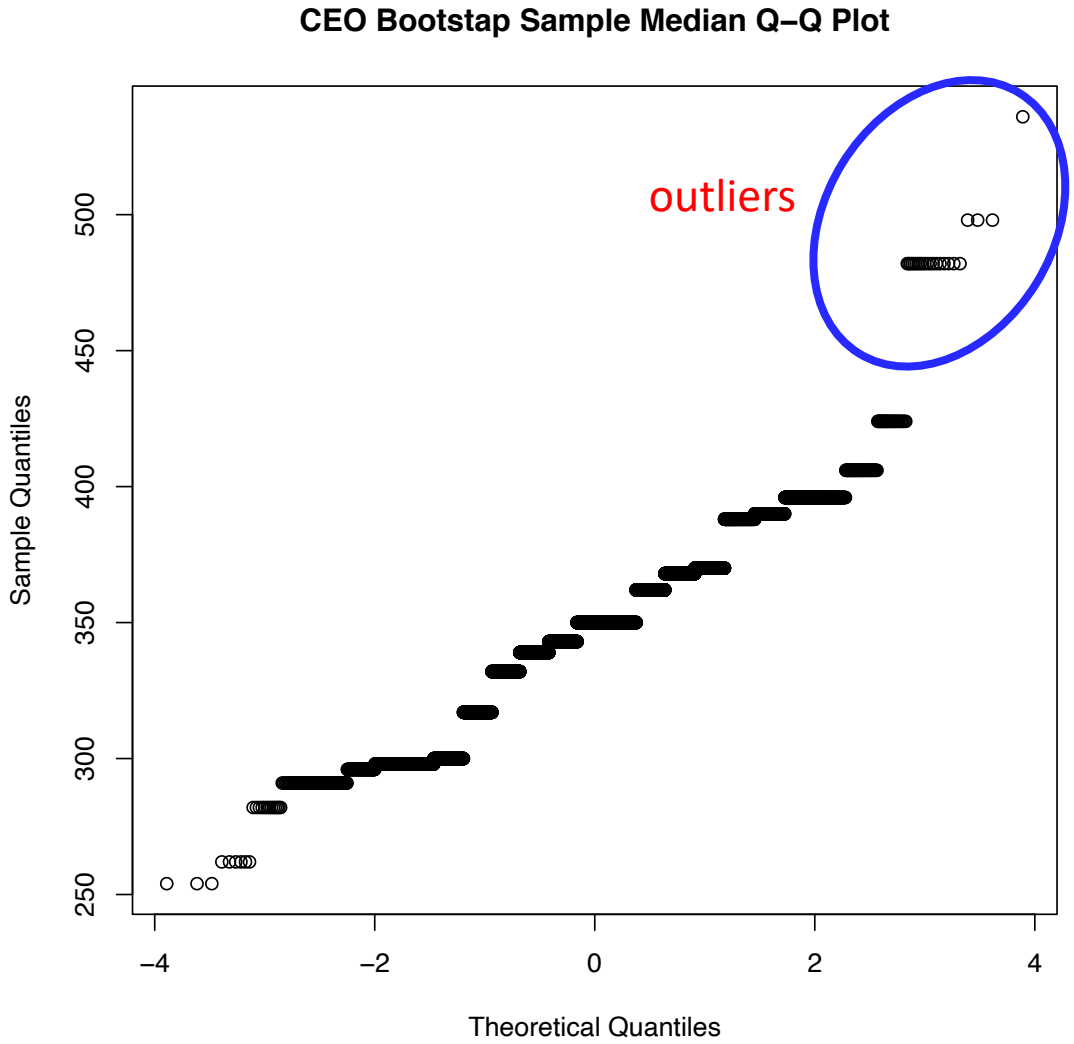
- Q-Q plot of CEO salary's bootstrap sample medians
- It's roughly linear so it's close to normal.
- We can use the normal distribution to construct the confidence intervals



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!