CSCI 556 Data Analysis & Visualization

Data and Sampling Distributions

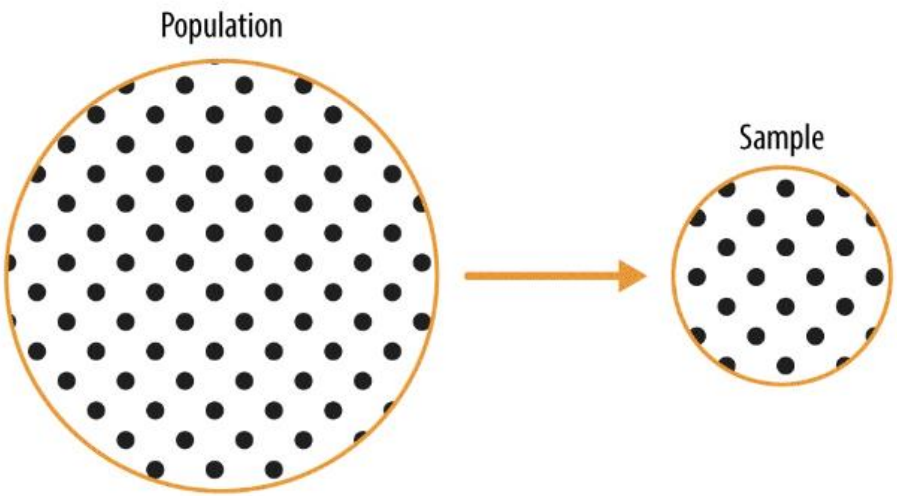
Instructor: Dr. Jinoh Kim

Data and sampling distributions

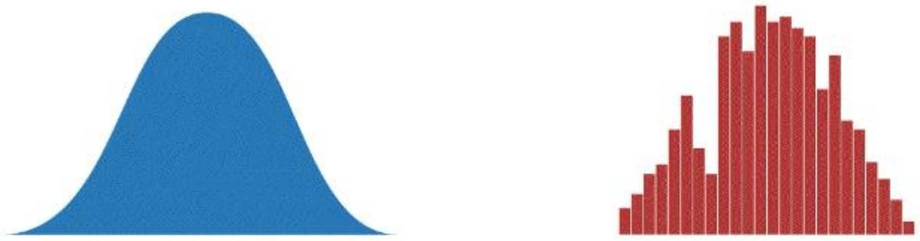
- Population vs. sample
- Random sampling and stratified sampling
- Sampling distribution of a statistic
- Confidence interval
- Distributions – normal, long-tailed, t-, binomial distributions

Population vs. sample

- Population
  - The larger data set or idea of a data set
  - Underlying but unknown distribution
  - Focus of traditional statistics



- Sample
  - A subset from a larger data set
  - Empirical distribution
  - Focus of modern statistics

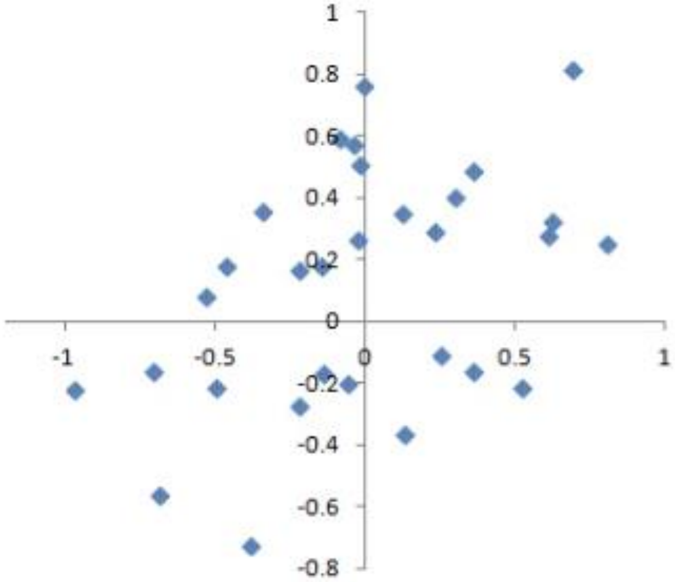


Random sampling

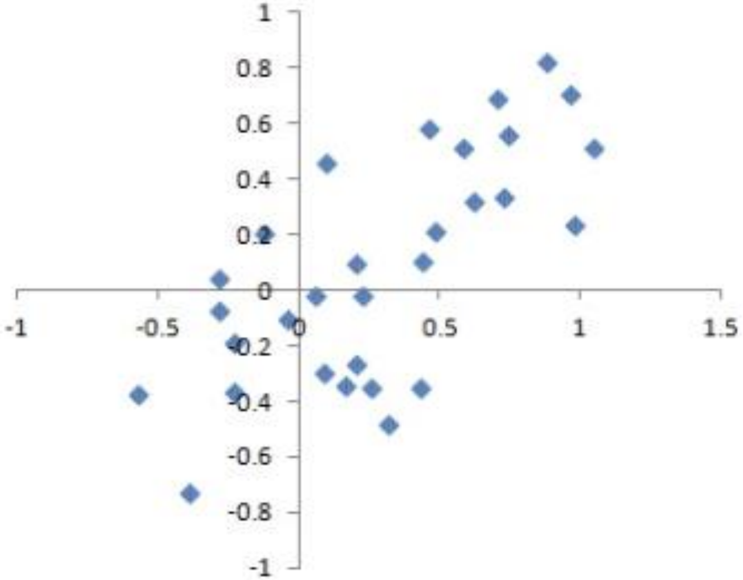
- Drawing elements into a sample at random
  - With replacement: observations are put back in the population after each draw for possible future reselection
  - Without replacement: selected observations are unavailable for future draws
- Simple random sample: the sample that results from random sampling without stratifying the population
- Sample bias: a sample that misrepresents the population.

Statistical bias

- Measurement/sampling errors produced by the measurement/sampling process
- Errors due to random chance vs. bias



Scatterplot of shots from a gun with true aim



Scatterplot of shots from a gun with biased aim

Representativeness

- Sample bias: a sample that misrepresents the population
- How to avoid the problem of sample bias?
    - Or how to get a sample more representative (of the population)?
- Stratified sampling: dividing the population into strata and randomly sampling from each strata
- Stratification: the strata are organized based on the shared characteristics (e.g., race) of the members in the group

Selection bias

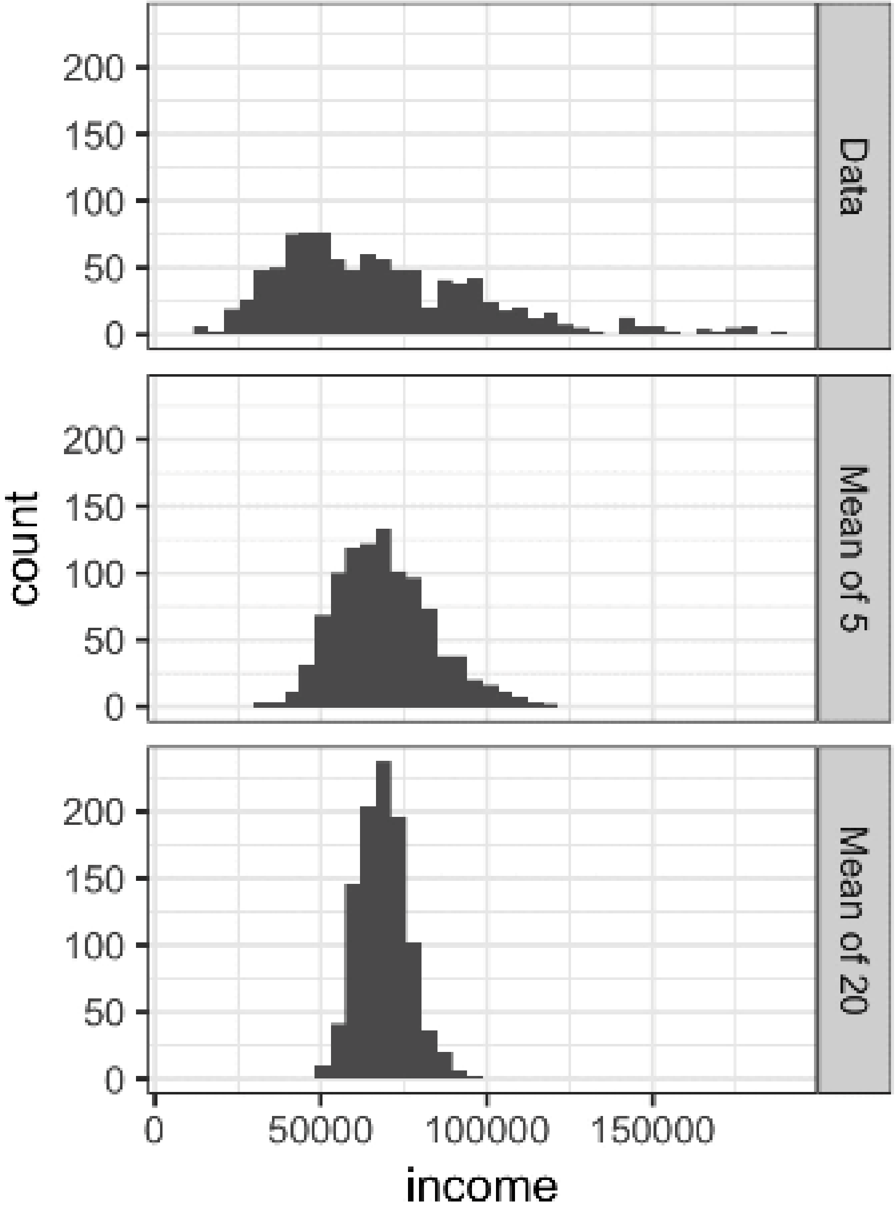
- Practice of selectively choosing data
  - Lead to a conclusion that is misleading or ephemeral
- Data snooping: extensive hunting through data in search of something interesting
- Vast search effect: bias or nonreproducibility resulting from repeated data modeling, or modeling data with large numbers of predictor variables
- Regression to the mean: if one sample of a random variable is extreme, the next sampling of the same random variable is likely to be closer to its mean

Sampling distribution of a statistic

- Sample statistic: A metric calculated for a sample of data drawn from a larger population
- Data distribution: The frequency distribution of individual values in a data set
- Sampling distribution: The frequency distribution of a sample statistic over many samples or resamples
- Central limit theorem: The tendency of the sampling distribution to take on a normal shape as sample size rises
- Standard error: The variability (standard deviation) of a sample statistic over many samples (not to be confused with standard deviation, which, by itself, refers to variability of individual data values)

Example

1) Data distribution: A sample of 1000 values
2) Sampling distribution of a statistic: A sample of 1,000 means of 5 values
3) Sampling distribution of a statistic: A sample of 1,000 means of 20 values



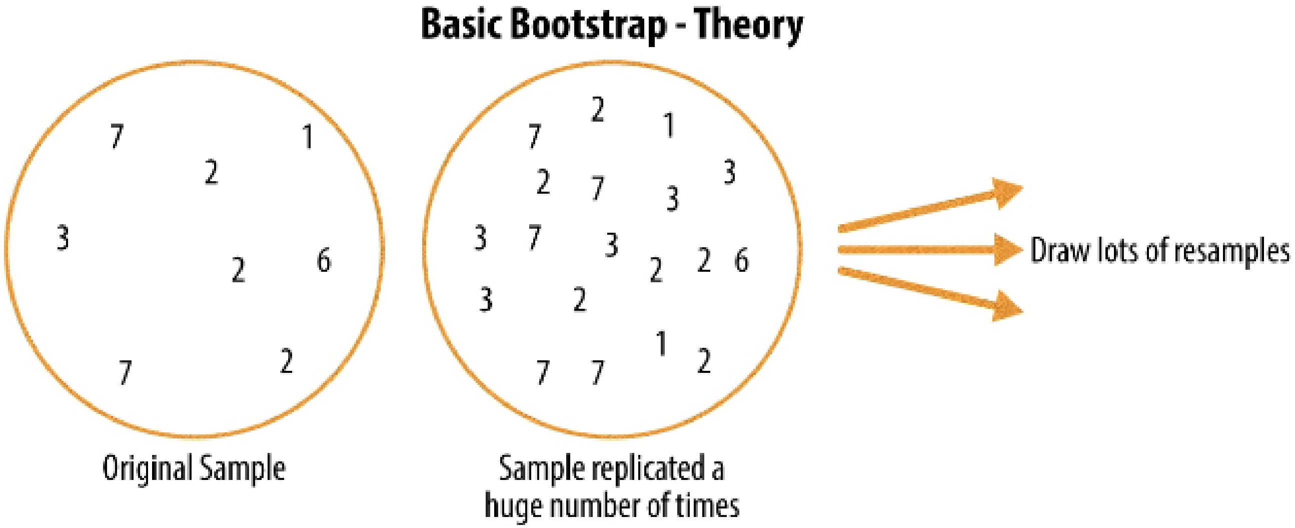
Central limit theorem

- The means drawn from multiple samples will resemble the familiar bell-shaped normal curve (even if the source population is not normally distributed)
- Standard error: A single metric that sums up the variability in the sampling distribution for a statistic
- For standard deviation *s* of the sample values, and the sample size *n*:

$$\text{Standard error} = SE = \frac{s}{\sqrt{n}}$$

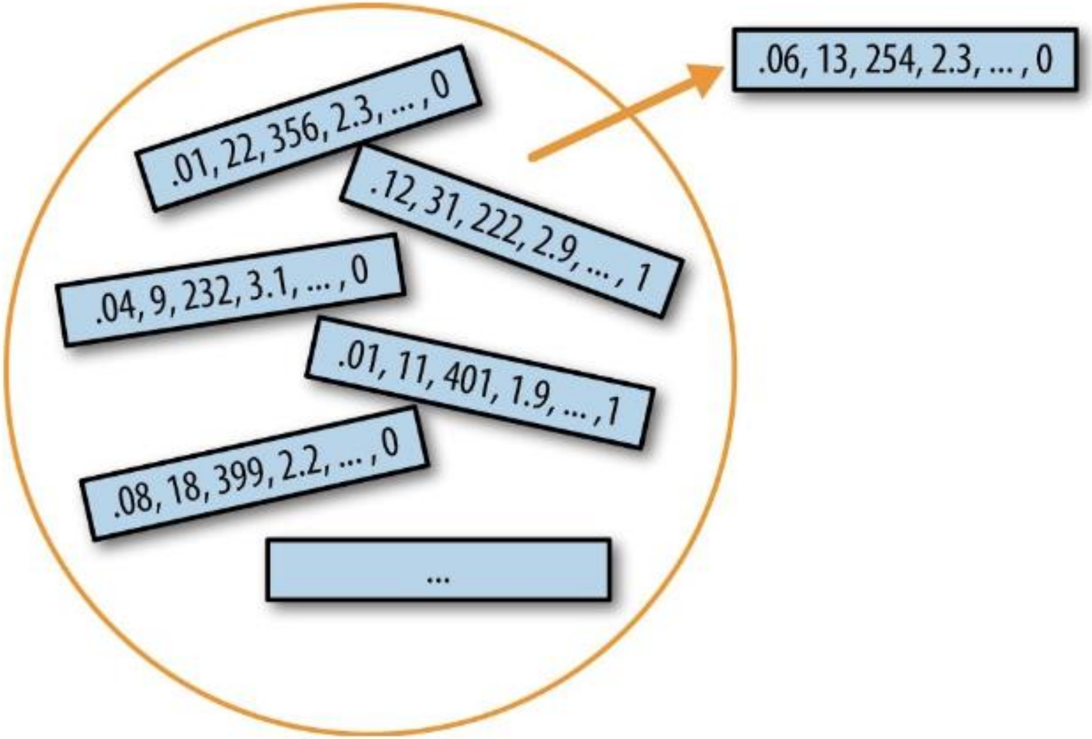
Bootstrap

- Approach of collecting new samples may be expensive
- Bootstrap: drawing additional samples, with replacement, from the sample itself and recalculate the statistic or model for each resample



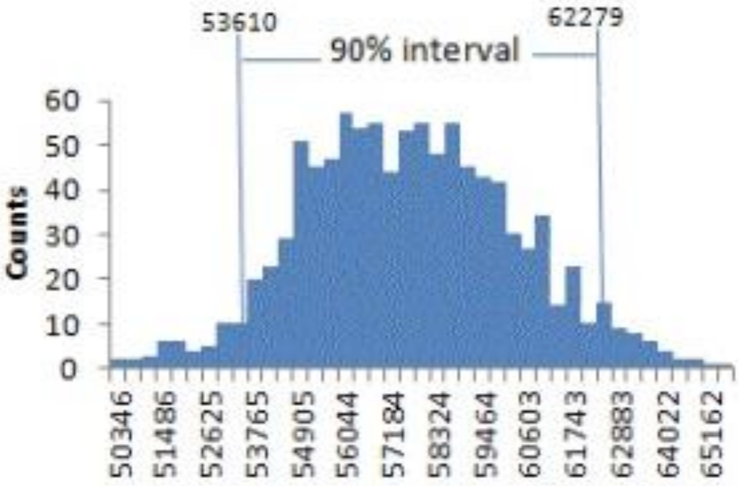
Bootstrap for multivariate data

- Rows are sampled as units
- Bagging: running multiple trees on bootstrap samples and then averaging their predictions



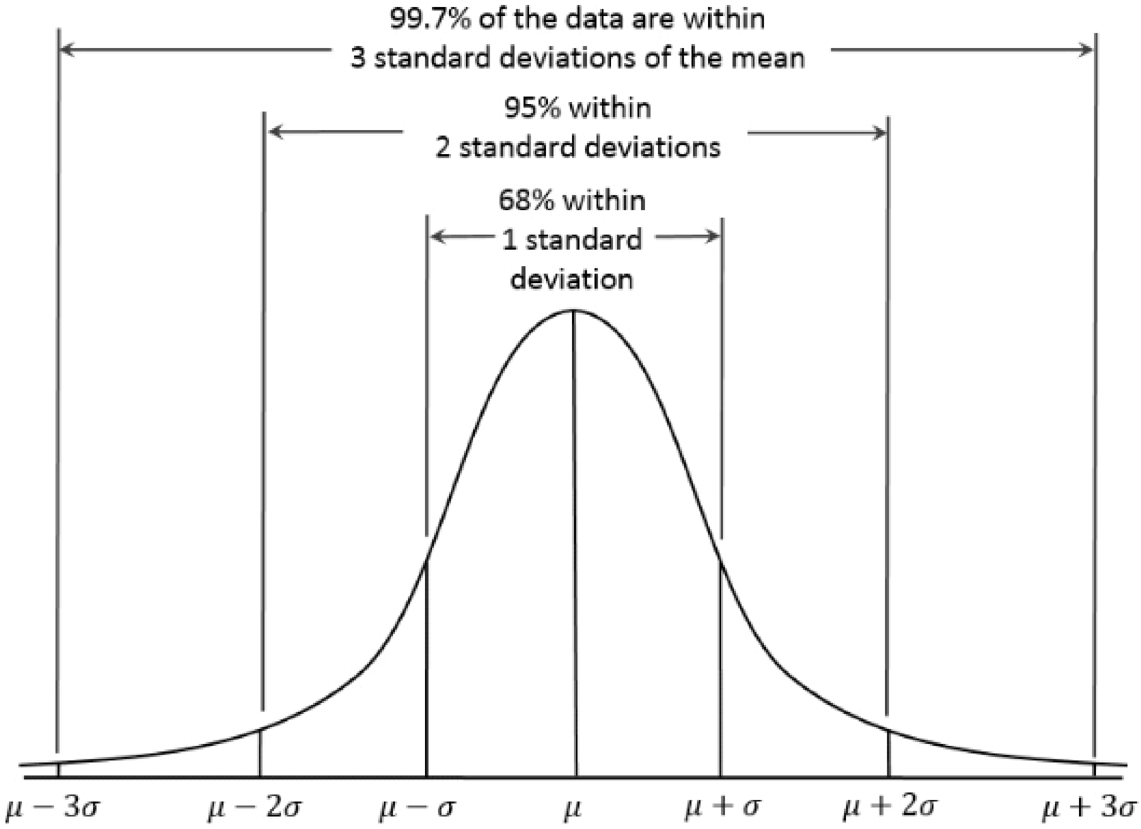
Confidence interval

- Presenting an estimate not as a single number but as a range
  - E,g., 90% confidence interval = the interval that encloses the central 90% of the bootstrap sampling distribution of a sample statistic
- Level of confidence: The higher the level of confidence, the wider the interval, and vice versa



Normal distribution

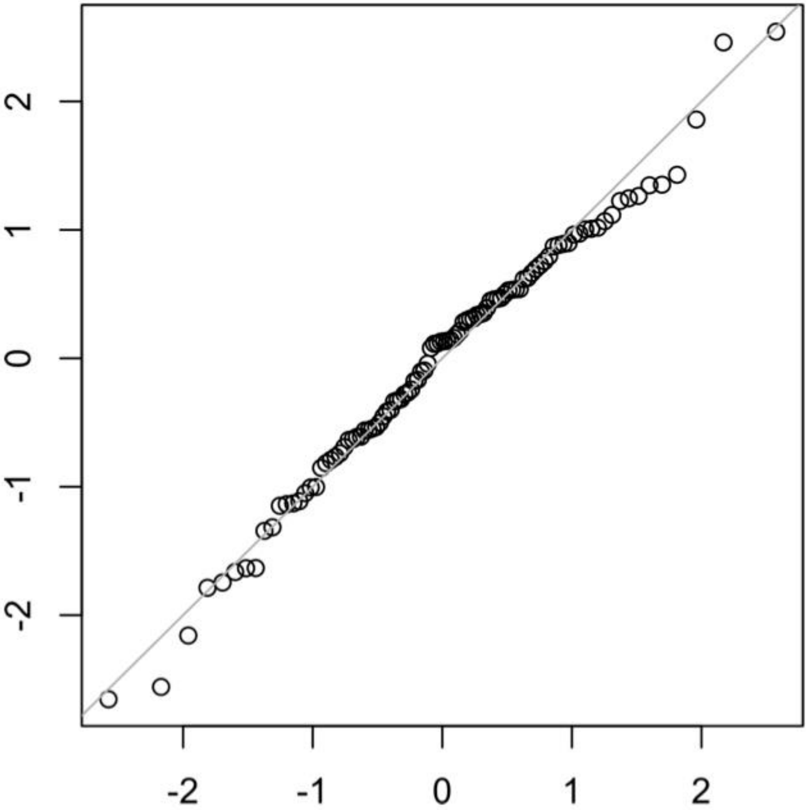
- Bell-shaped distribution with mean (μ) and standard deviation (σ)



Standard normal distribution

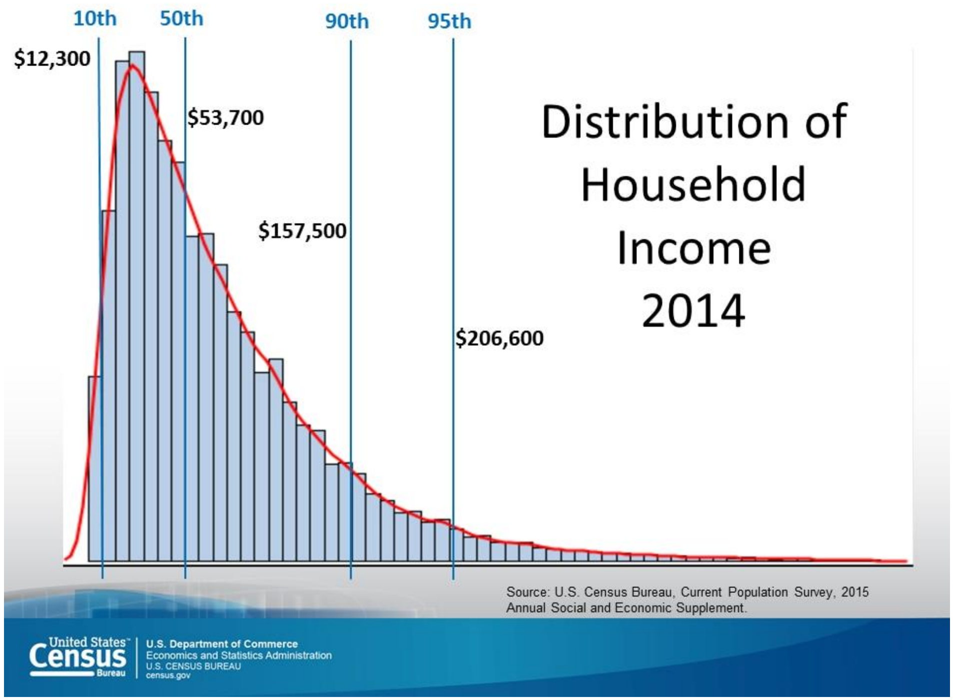
- Normal distribution with μ=0 and σ=1 (also known as z-distribution)
- Standardization: $z = \frac{x-\mu}{\sigma}$

QQ-plot: tests how close a sample is to the normal distribution;
If the points roughly fall on the diagonal line, then the sample distribution can be considered close to normal.



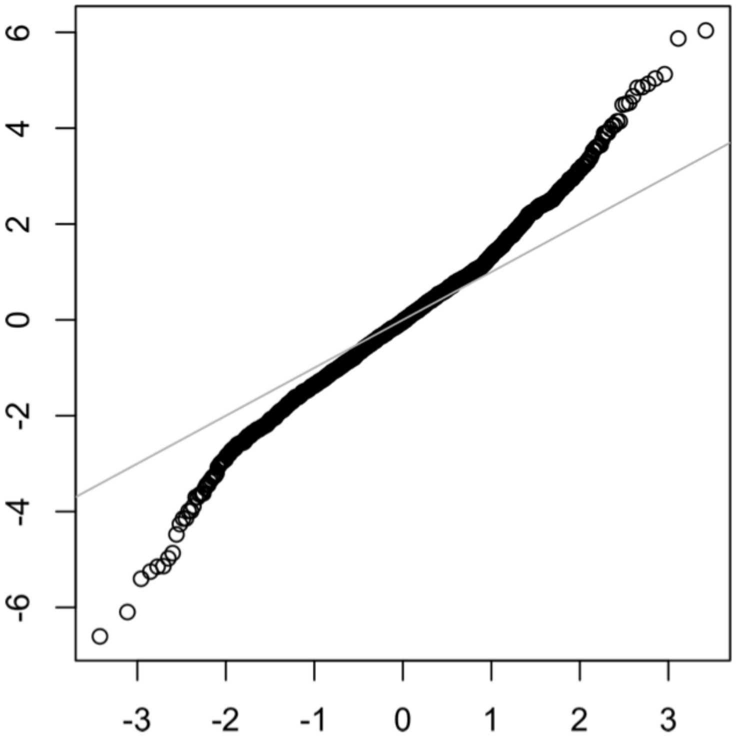
Long-tailed distribution

- In reality, the distribution is highly skewed (asymmetric)
- The tails of a distribution correspond to the extreme values (small and large)



Source: U.S. Census Bureau, Current Population Survey, 2015 Annual Social and Economic Supplement.

Example QQ-plot



Student t-distribution

- t-distribution is a normally shaped distribution, but a bit thicker and longer on the tails.
- Used extensively in depicting distributions of sample statistic
- Distributions of sample means are typically shaped like a t-distribution
- The larger the sample, the more normally shaped the t-distribution becomes
- Confidence interval calculation based on t-distribution is widely applied

Binomial distribution

- Binomial: having two outcomes (yes or no)
- Binomial distribution: The frequency distribution of the number of successes (*x*) in a given number of trials (*n*) with specified probability (*p*) of success in each trial
  - E.g., flipping a coin 10 times
- Mean = $n*p$
- Variance = $n*p(1-p)$

Level *C* confidence interval

- Interval computed from sample data with probability *C* of producing an interval containing the true value of the parameter

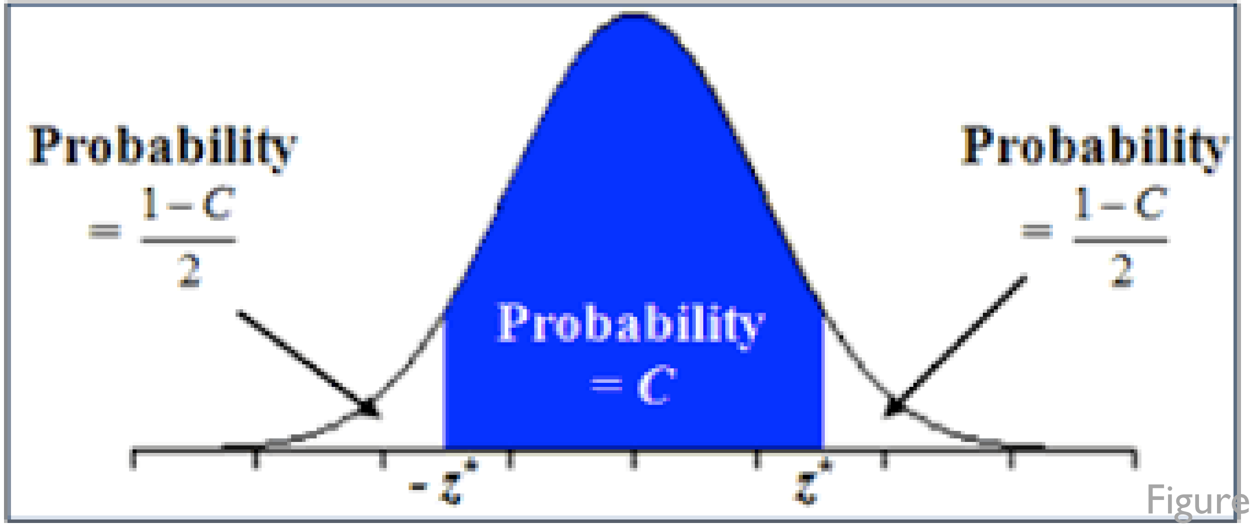


Figure from Wisconsin.edu

z*	1.645	1.960	2.576
C	90%	95%	99%

Confidence interval (for population mean)

- Suppose sample mean $\bar{x}$ has the normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, then confidence interval (for μ) is:

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} \text{ and } \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

- Margin of error $m = z^* \frac{\sigma}{\sqrt{n}}$
- Confidence interval = $\bar{x} \pm m$

Confidence interval example

- Question: The National student Loan Survey collects data to examine questions related to the amount of money that borrowers owe. The survey selected a sample of 1280 borrowers who began repayment of their loans between four and six months prior to the study. The mean of the debt for undergraduate study was $18,900 and the standard deviation was about $49,000. This distribution is clearly skewed but because our sample size is quite large, we can rely on the central limit theorem to assure us that the confidence interval based on the normal distribution will be a good approximation. Let's compute a 95% confidence interval for the true mean debt for all borrowers. Although the standard deviation is estimated from the data collected, we will treat it as a known quantity for our calculations here.

From Introduction to the practice of statistics, 4th edition

Confidence interval example

- Example: student loan data
  - (Sample) mean of the debt = $18,900
  - Standard deviation = $49,000
  - Number of students = 1,280

z*	1.645	1.960	2.576
C	90%	95%	99%

- 95% confidence interval?

Margin of error $m = z^* \frac{\sigma}{\sqrt{n}} = 1.960 \frac{49{,}000}{\sqrt{1280}} = 2684$

Confidence interval = $\bar{x} \pm m = 18900 \pm$2684
= (16216, 21584)

Summary

- Sampling strategies: random sampling and stratified sampling
- Distribution of sample statistic
- Central limit theorem, bootstrapping, confidence interval
- Distributions – normal, long-tailed, t-, binomial distributions
- Calculating confidence interval