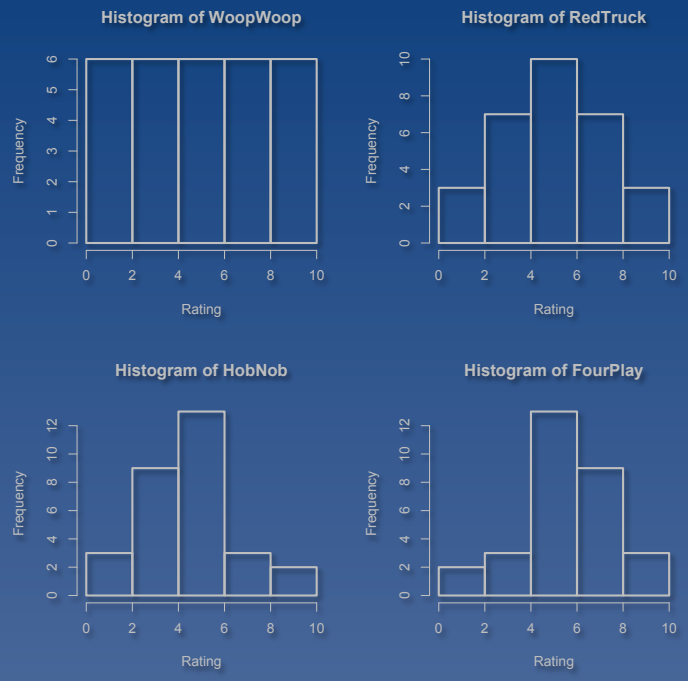
Lecture 5
Segment 2
Sampling

Sampling

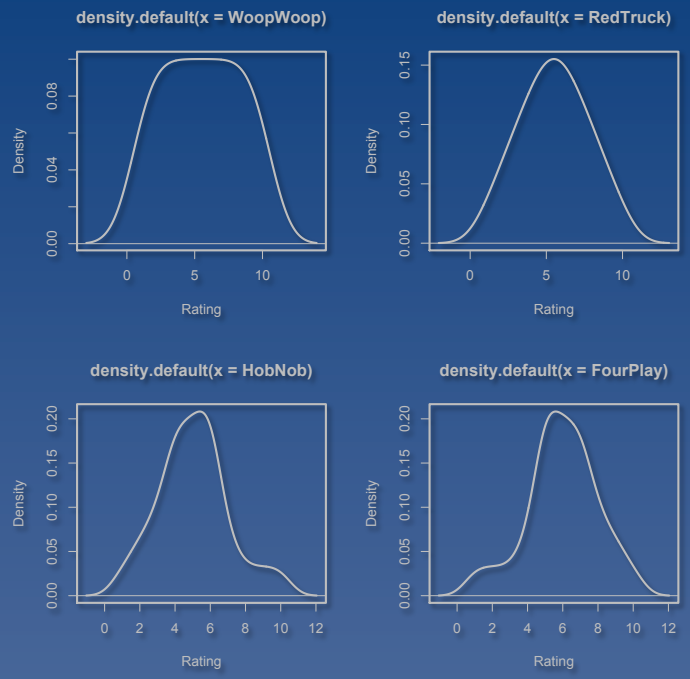
- Important topics
  - Sampling error
  - Standard error
  - Probability histograms
  - Central limit theorem

Wine tasting!

Four histograms



Four density plots

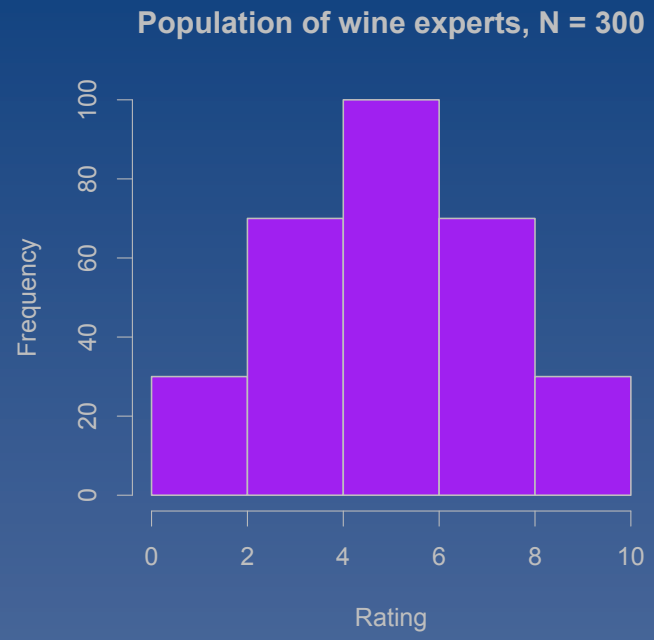


Red Truck: A closer look

Sampling

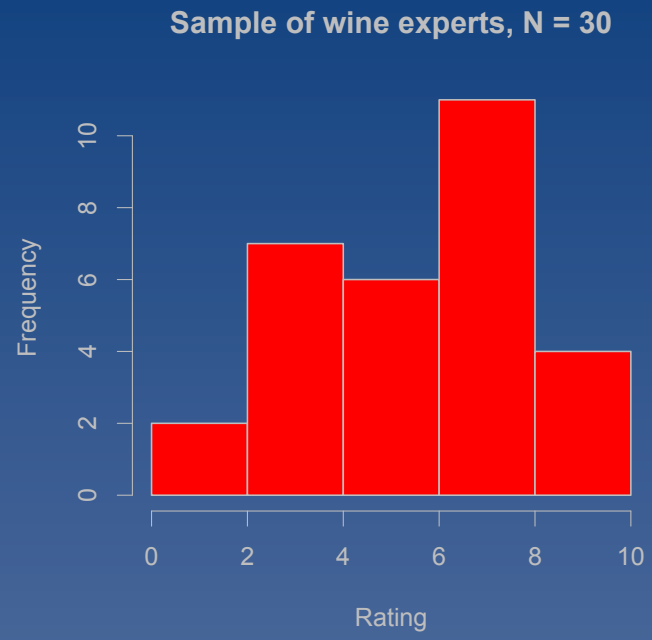
- Suppose there are only 300 certified "wine experts" in the entire world
  - In other words, the population N = 300
  - Let's also assume that the ratings for Red Truck are normally distributed in the population

Histogram for the population



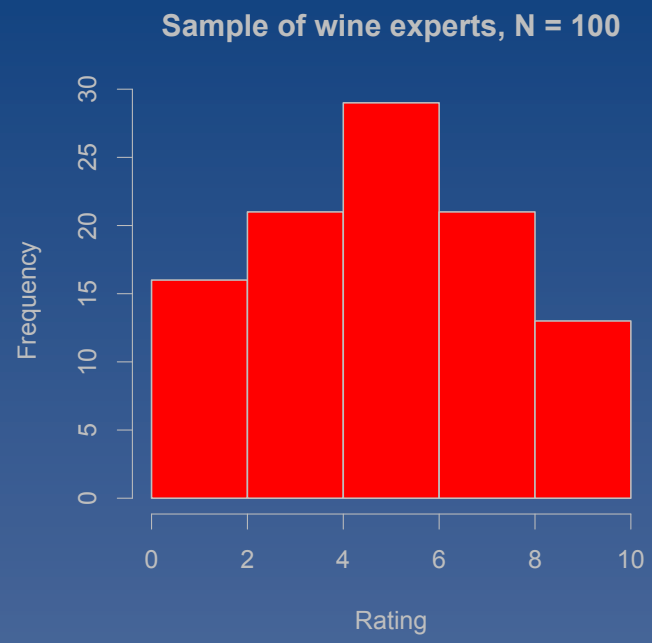
M = 5.5
SD = 2.22

Random sample, N = 30



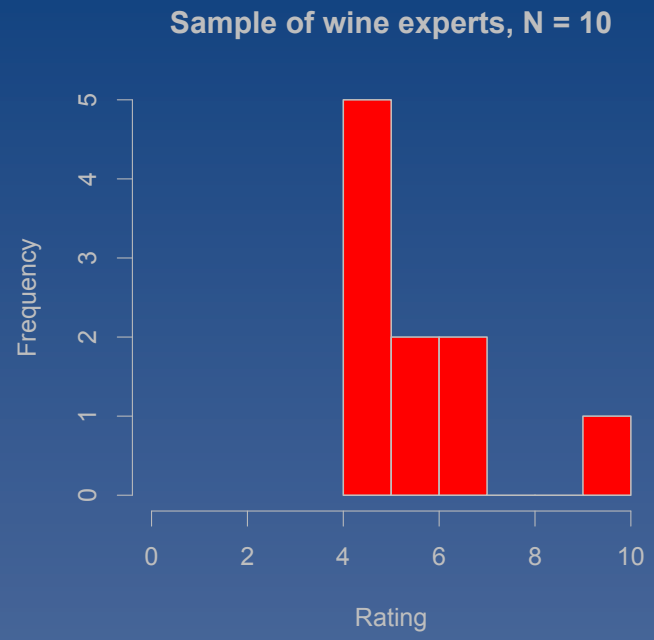
M = 5.93
SD = 2.45

Random sample, N = 100



M = 5.47
SD = 2.19

Random sample, N = 10



M = 6.00
SD = 1.70

Sampling error

- The difference between the population and the sample
- PROBLEM!
  - We typically don't know the population parameters
  - So, how do we estimate sampling error?

Sampling error

- Clearly depends on the size of the sample, relative to the size of the population
- Also depends on the variance in the population

Sampling error

- We therefore estimate sampling error from the size of the sample and the variance in the sample
  - Under the assumption that the sample is random and representative of the population

Standard error

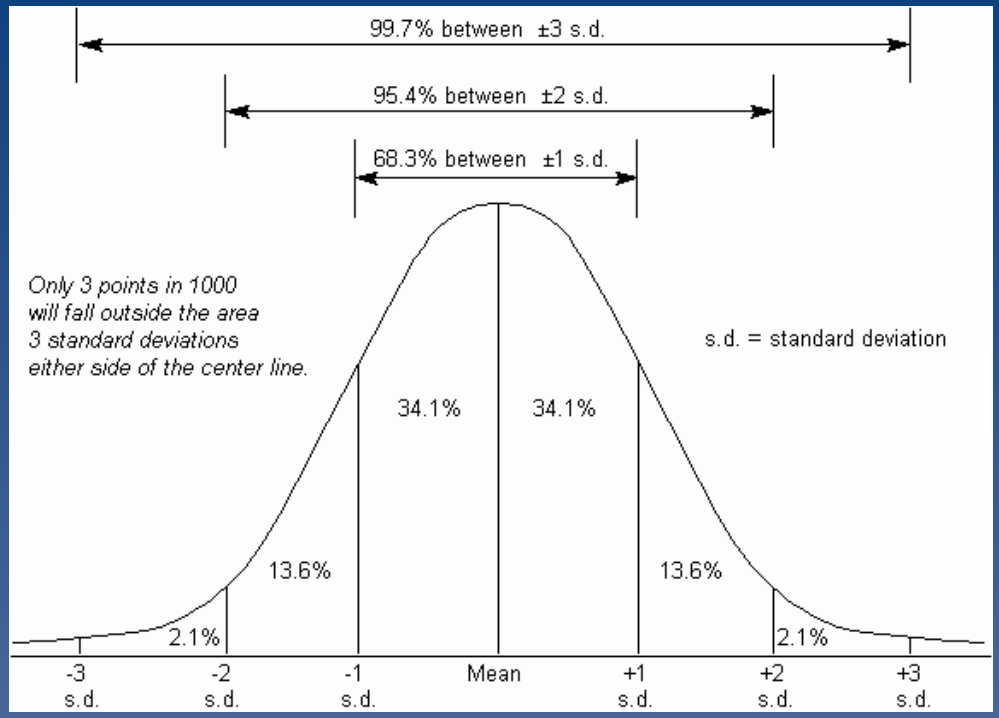
- Standard error is an estimate of amount of sampling error
  - SE = SD / SQRT(N)

  - SE: Standard error
  - SD: Standard deviation of the sample
  - N: Size of the sample

Back to histograms

- If a variable X is perfectly normal, then we know a lot about it's distribution

The normal distribution



Probability histogram

- Probability histogram
  - A distribution of sample means
  - Assume we took multiple samples of the same size and then plotted all the sample means
    - N = 10
    - N = 30
    - N = 100

Probability histogram

- Standard error is the standard deviation of the probability histogram

Distribution of sample means

- Characteristics
  - It is hypothetical, i.e., we don't *know* the dimensions of the distribution as we do with a distribution of individual scores (we estimate the dimensions)

Distribution of sample means

- Characteristics
  - The mean of the distribution of sample means should be the same as the mean of the population of individuals

Distribution of sample means

- Characteristics
  - The variance of the distribution of sample means is less than the variance in the population of individuals
    - $\sigma^2_M = \sigma^2 / N$

Distribution of sample means

- Characteristics
  - The shape of the distribution of sample means is approximately normal

Distribution of sample means

- $\sigma^2_M = \sigma^2 / N$
  - σ^2_M is the variance of the distribution of sample means
  - σ_M is the standard deviation of the distribution of sample means (standard error)
  - σ^2 is the variance of the population
  - σ is the standard deviation of the population
  - N is the sample size

Central Limit Theorem

- Three principles
  - The mean of the distribution of sample means is the same as the mean of the population
  - The standard deviation of the distribution of sample means is the square root of the variance of the distribution of sample means, which is $\sigma^2_M = \sigma^2 / N$
  - The shape of the distribution of sample means is approximately normal if either (a) N >= 30 or (b) the shape of the population distribution is normal

Sampling

- Important topics
  - Sampling error
  - Standard error
  - Probability histograms
  - Central limit theorem

Image in slide 3 was retrieved from
http://www.delawareonline.com/blogs/secondhelpings/
uploaded_images/redwinegl-758416.JPG

Image in slide 6 was retrieved from
http://www.winebusiness.com/content/Image/
brands-redTruck.jpg

Image in slide 17 was retrieved from
http://www.syque.com/quality_tools/toolbook/Variation/
Image375.gif