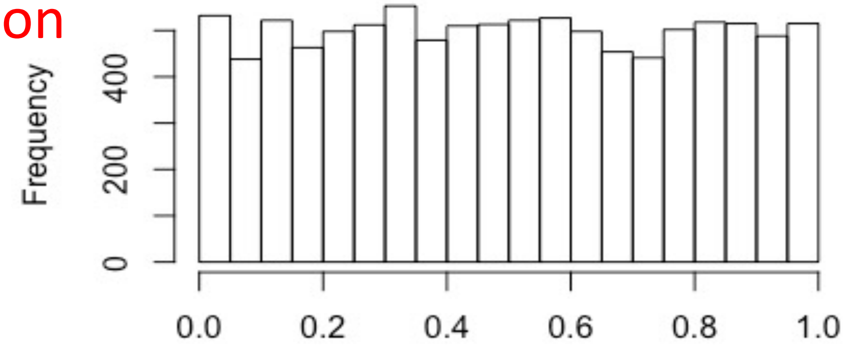
All handouts for this class on Blackboard

IST772 (IST772) Probabilities Over the Long Run (Week 3)

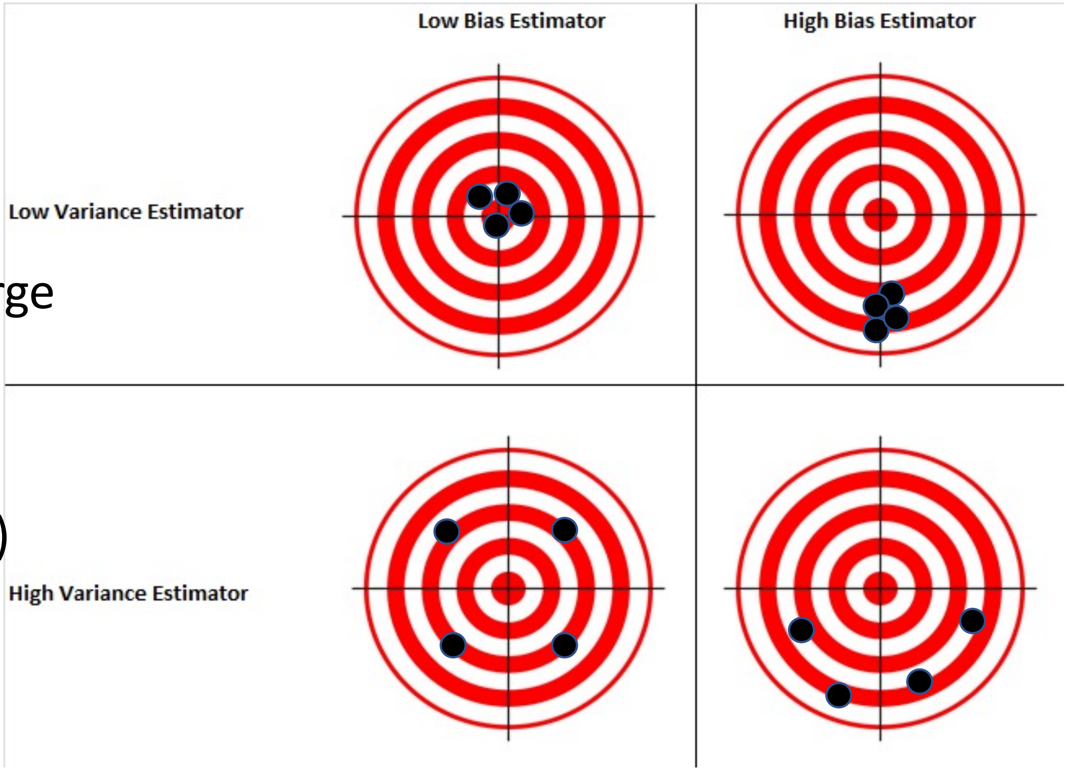
Pre-class activity: Is this a distribution of raw data or a sampling distribution? How do you know?



The Frequentist Philosophy

- There is an unknown population parameter that has a fixed value
- We want to estimate that parameter from sample data
- The goal is to create estimators (statistical procedures) that will converge on the fixed, but unknown parameter over many replications
- Good estimators have the smallest possible bias (distance between the estimate and the fixed unknown value) and the smallest possible variance (variation among replications)

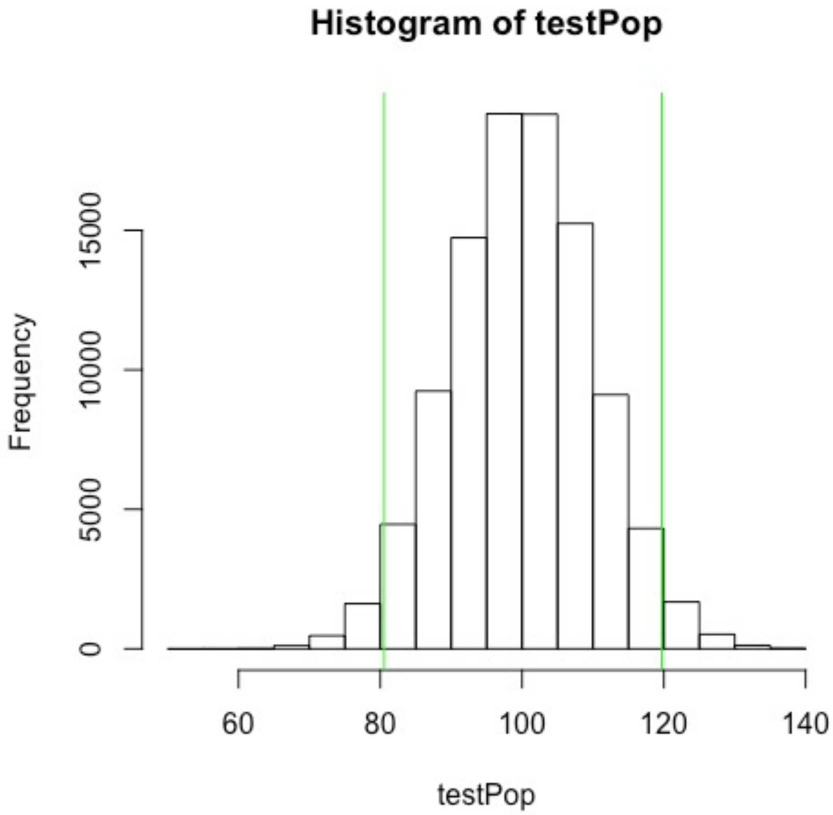
(Simulations provide a method of understanding the properties of such estimators)



Breakout 1 – Marking a Population Distribution

- Open `1. week3breakout1.Rmd`
- Create a large population of test scores with a random number generator
- Display a histogram with the result
- Mark various points on the X-axis with abline( )
- Share your code on https://codeshare.io/aJDyRX

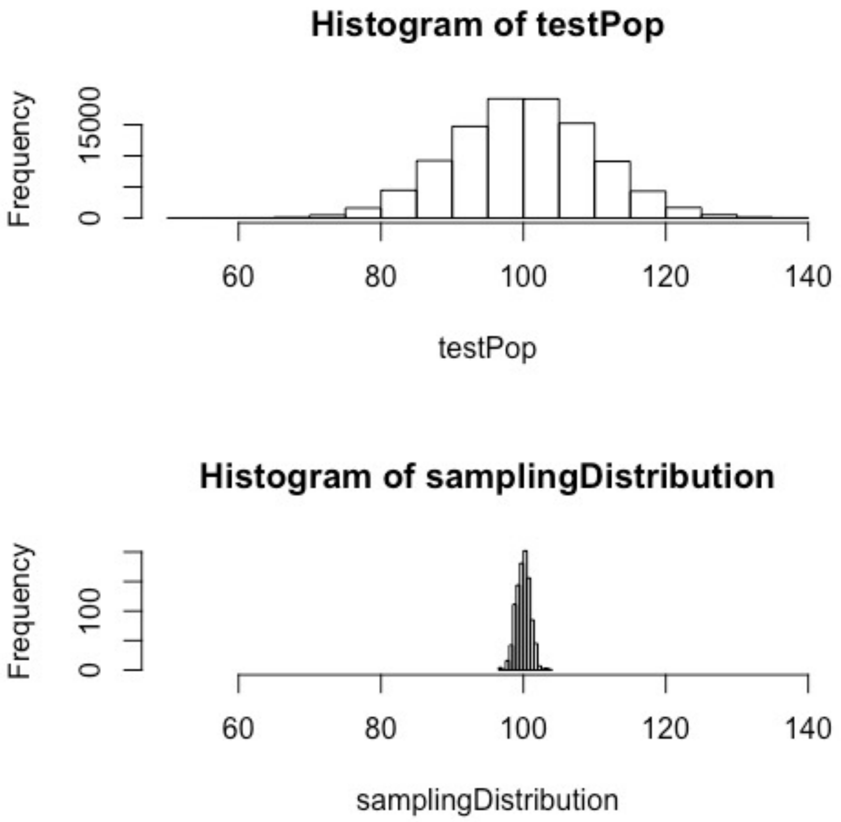
```
hist(testPop)
abline(v=quantile(testPop, probs=0.025),col="green") # Lower tail
abline(v=quantile(testPop, probs=0.975),col="green") # Upper tail
```



R-Studio Clinic #4

- Answer the questions in `2. week3rstudioclinic.Rmd`

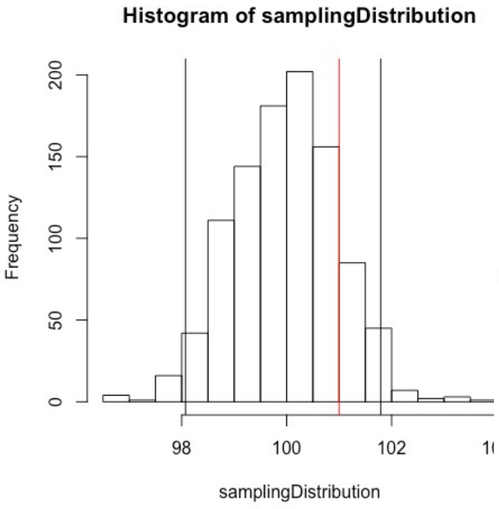
Original Population Versus Sampling Dist



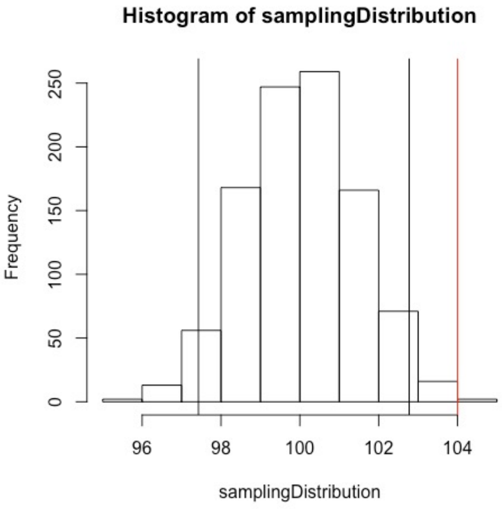
Breakout 2 – Sampling Distribution Case Studies

- Open `3. week3breakout2.Rmd`
- Start by re-generating the plot on the previous slide
- Then create additional plots of sampling distributions to reason about five case studies
- Each case study has a slightly different scenario for sample size: You will mark the boundaries of the central region and the position of the sample mean provided in the scenario
- Then reason about what the position of the sampling mean implies

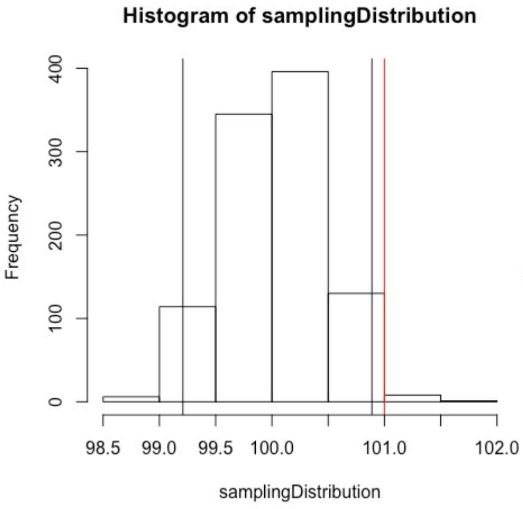
Case Studies A through D



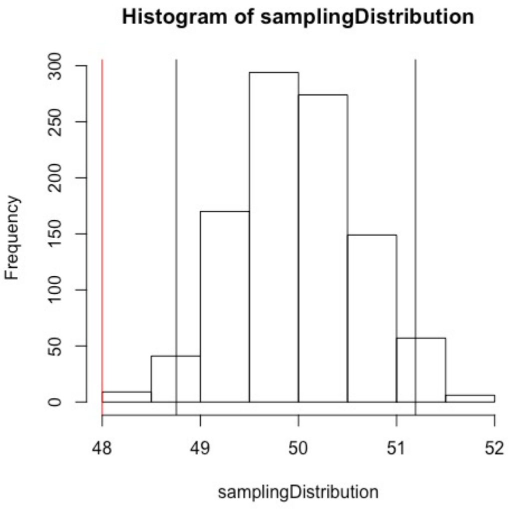
A: n=100, Xbar=101



B: n=49, Xbar=104



C: n=500, Xbar=101



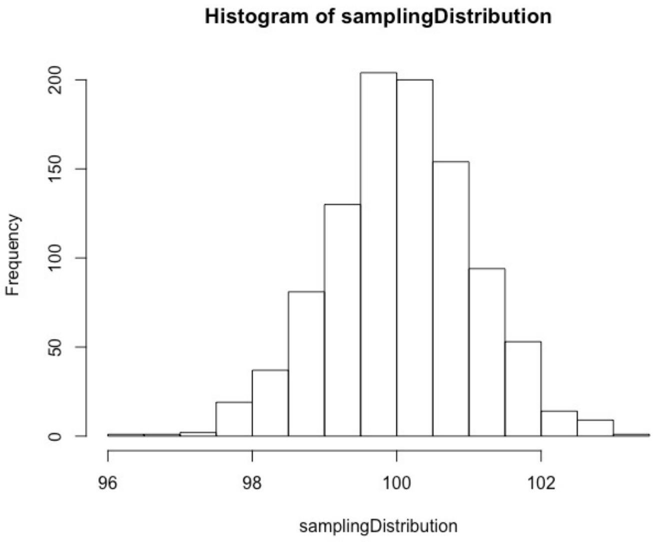
D: mu=50, n=64, Xbar=48

Dispersion of the Sampling Distribution

The Central Limit Theorem states that over the long run the mean of the sampling distribution will match the mean of the underlying population.

The **dispersion** of the sampling distribution depends upon the dispersion of the underlying population.

But there is also another influence that affects the dispersion of the sampling distribution. Consider the code on the following slide.



A Special Standard Deviation

```
sdVector <- NULL # Start a list with nothing
sampSizes <- (2:10)^2 # A list of 9 sample sizes ranging from 4 to 100

for (i in sampSizes)
{
  # Do a sampling distribution for each sample size
  samplingDistribution <- replicate(1000, mean(sampleTestScores(i)))

  # Add the standard deviation of that sampling distribution to the list
  sdVector <- c(sdVector, sd(samplingDistribution))
}
```

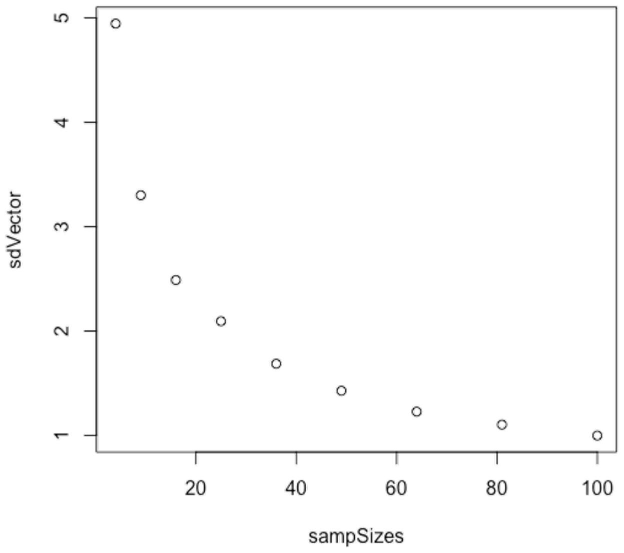
A Special Standard Deviation

This plot shows the standard deviation of sampling distributions that were created using different sample sizes (ranging from 4 to 100).

Remember that the original testPop population had sd = 10.

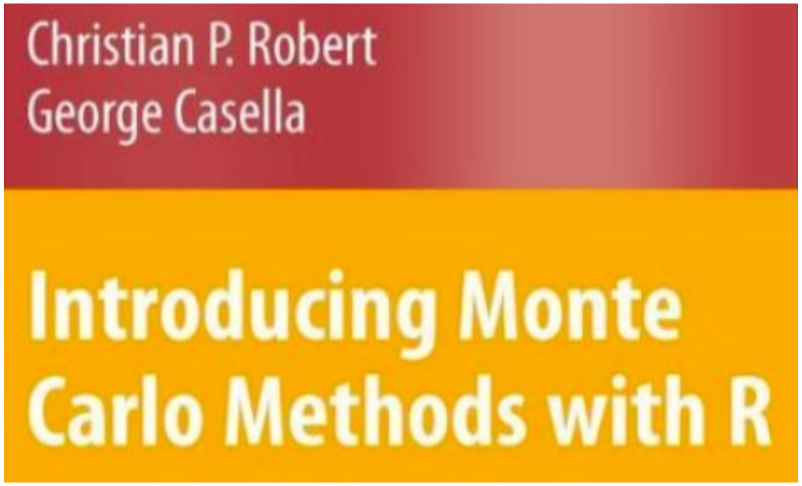
The standard deviation of a sampling distribution has a special name: the **Standard Error of the Mean**.

What is the relationship between population SD, sample size, and standard error of the mean?



Paper of the Week – Excerpt (Chapter 2) from Robert & Casella, 2010

- Generating random distributions to test our statistical reasoning requires accurate generation of “pseudo-random” numbers
- Today, we sampled hundreds of thousands of random samples that were each independent of one another
- Our capability to conduct Bayesian estimation depends upon being able to sample from a distribution where each sample is dependent solely on the previous sample
- Robert & Casella do a deep dive into sampling and Monte Carlo methods in their book (email me if you are interested in the whole book)



Christian P. Robert
George Casella

Introducing Monte Carlo Methods with R

Homework

- Make sure you are using the updated syllabus that I distributed at the beginning of the semester (on the wall and in the handouts folder).
- Make sure you are thoroughly commenting your code for future use!
- The homework for week three is based on exercises 2 through 7 on pages 50 and 51 (but be sure to answer the questions in the supplied notebook).