Variation in Estimates of Sample Properties

Samples So Far

Here's what we've talked about with respect to samples of a population:

- Estimating Mean value of a sample property
- Sample design to get correct *unbiased* estimate of that mean
- Standard deviation to describe dispersion $\sqrt{Variance}$
- 67% of the values within a population fall within 1 SD of the Mean
- 95% of the values within a population fall within 2 SD of the Mean

Sample Properties: Variance

How variable was that population?

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$$

- **Sums of Squares** over n-1
- n-1 corrects for both sample size and sample bias
- σ^2 if describing the population
- Units in square of measurement...

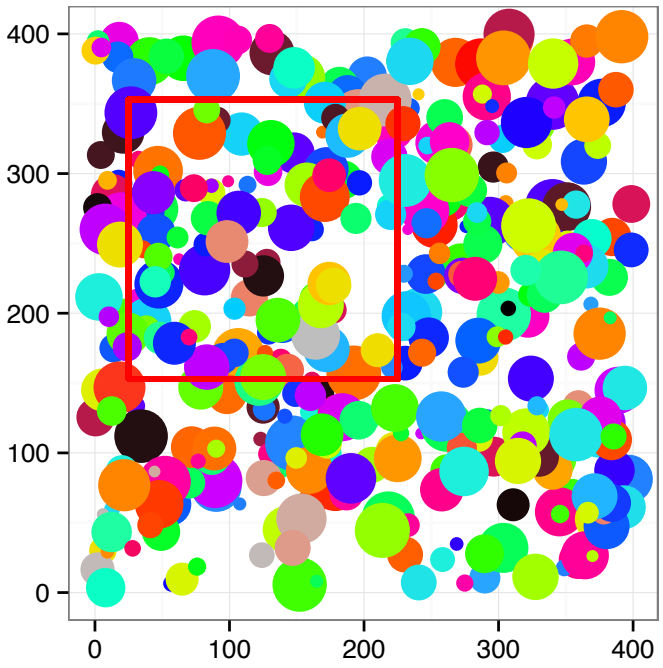
Sample Properties: Standard Deviation

$$s = \sqrt{s^2}$$

- Units the same as the measurement
- If distribution is normal, 67% of data within 1 SD
- 95% within 2 SD
- σ if describing the population

Populations, Samples, and Repeatability

How good is our estimate of a population parameter?



Populations, Samples, and Repeatability

We've seen that we get variation in point estimates at any sample size

What does that variation look like?

Exercise: Variation in Estimation

- Consider a population with some distribution (rnorm, runif, rgamma)
- Think of the mean of one sample as one individual replicate
- Take many (50) 'replicate' means from this population of means
- What does the distribution of means look like? Use the *hist* function
- How does it depend on sample size (within replicates) or distribution type?

Extra: Show the change in distributions with sample size in one figure.

Central Limit Theorem Simulation

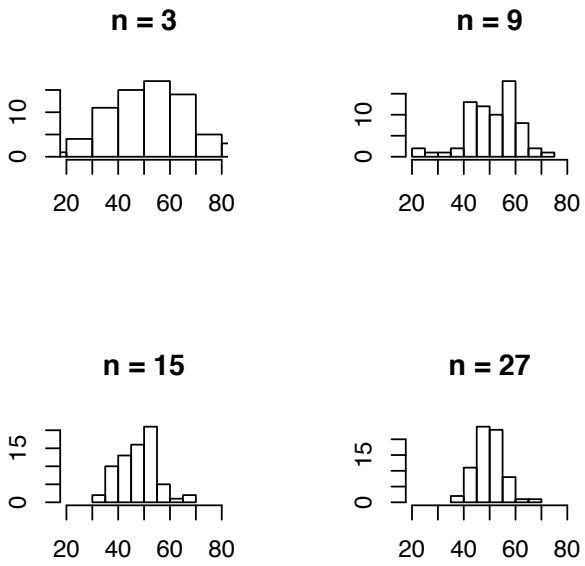
```
set.seed(607)
n<-3
mvec<-rep(NA, times=100)

#simulate sampling events!
for(i in 1:length(mvec)){
  mvec[i]<-mean(runif(n, 0,100))
}

hist(mvec, main="n=3")
```

Central Limit Theorem

The distribution of means converges on normality



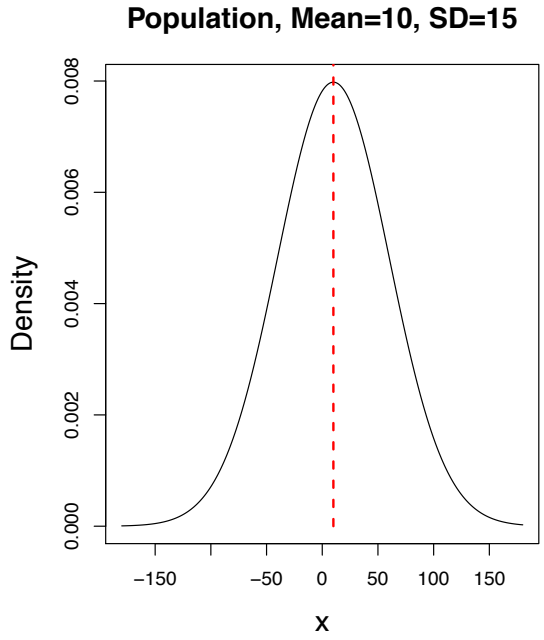
Central Limit Theorem: The distribution of means of a sufficiently large sample size will be approximately normal

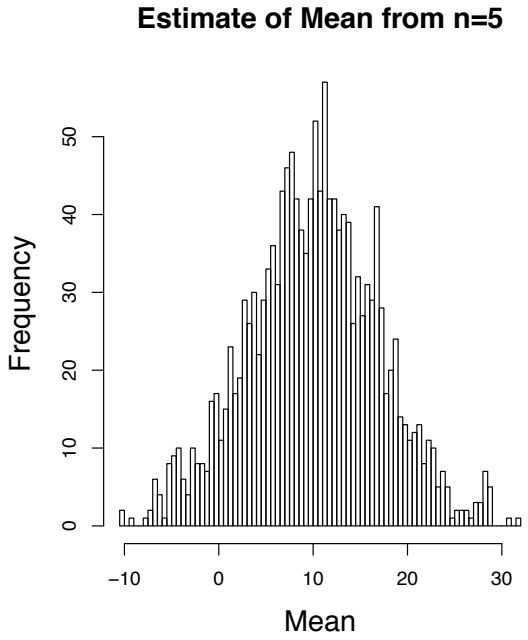
Estimating Variation Around a Mean

Great, so, if we can draw many replicated means from a larger population, we can the standard deviation of an estimate!

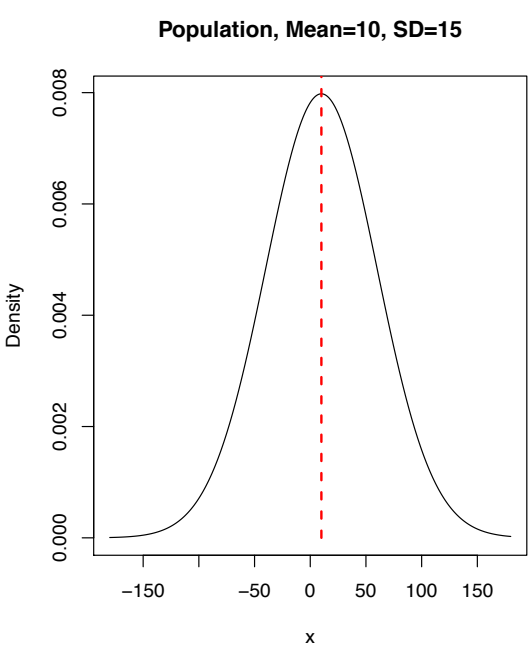
This standard deviation of the estimate of the mean is the **Standard Error.**

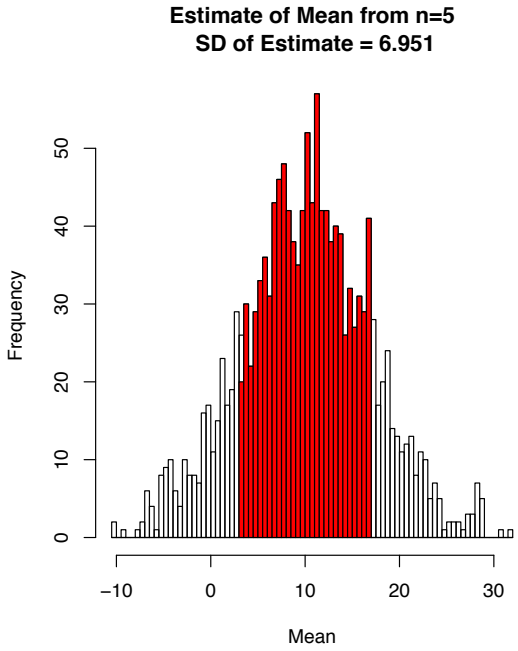
Variation in a Population v. Variation in Estimates of a Parameter



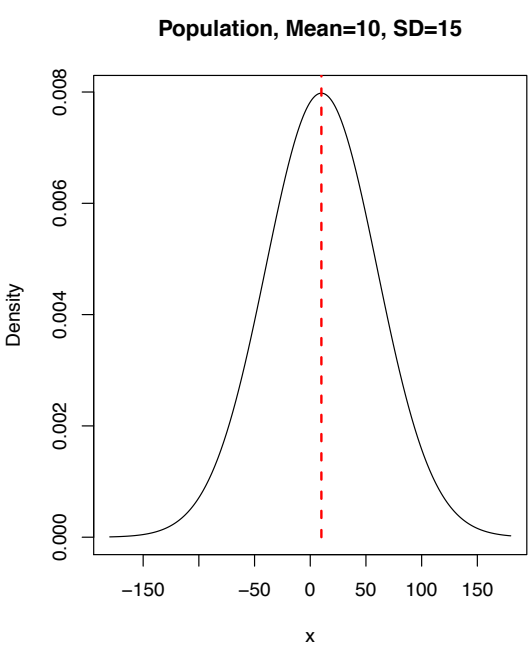


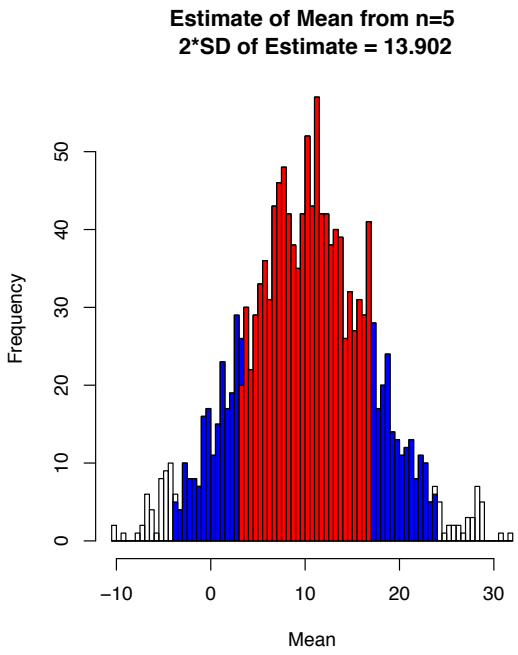
Variation in a Population v. Variation in Estimates of a Parameter: 66.7% of your Estimates





Variation in a Population v. Variation in Estimates of a Parameter: 95% of your Estimates





That's great, but for a single study, we only have one sample...

A Bootstrap Simulation Approach to Standard Error

- Our sample is representative of the entire population
- Therefore, we can resample it *with replacement* for 1 simulated sample
- We use our sample size as the new sample size as well

We set the replace=TRUE argument in the *sample* function
Try sampling from the bird count data with replacement.

A Bootstrap Simulation Approach to Standard Error

```
sample(bird$Count, replace=T, size=nrow(bird))

#  [1]   2   7  77 148  23   1   1 173   2   2  77   1   4   1
# [15] 300 297  67 173 148  23   4   1 135  18 173 135   3   2
# [29] 625 230 297 128  33  18  64  33   4  12  14  23   1  67
# [43]  16
```

```
sample(bird$Count, replace=T, size=nrow(bird))

#  [1]   1   5  13   3  28 173   7  13   2   2  10  16  16 625
# [15] 230   3   2 230 282   3 173 625 625   3  64   1   4  12
# [29]  64  14   1  13  77 135 135  25 173   5  14   1  33 230
# [43]  28
```

A Bootstrap Simulation Approach to Standard Error

```
n.sims<-100
birdMean <- rep(NA, n.sims)
for(i in 1:n.sims){
  birdMean[i] <- mean(sample(bird$Count, replace=T, size=nrow(bird)))
}

sd(birdMean)

# [1] 17.8
```

But what if we don't have simulation?

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

$\bar{Y}$ - sample mean

s - sample standard deviation

n - sample size

Bootstrap v. Formula for Standard Error

```
sd(birdMean)

# [1] 17.8

sd(bird$Count)/sqrt(nrow(bird))

# [1] 18.6
```

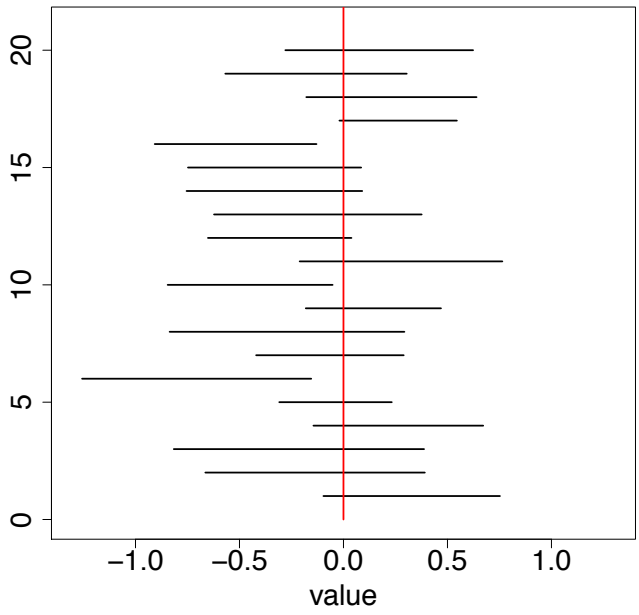
95% Confidence Interval and SE

- Recall that 95% of the data in a sample is within 2SD of its mean
- So, 95% of the times we sample a population, the *true* mean will lie within 2SE of our estimated mean
- This is the 95% **Confidence Interval**

$$\bar{Y} - 2SE \leq \mu \leq \bar{Y} + 2SE$$

How Often does the 95% Confidence Interval Contain the True Mean?

True Mean = 0 from Normal Distribution with SD=1



Variation in Other Estimates

- Many SEs and CIs of estimates have formulae and well understood properties
- For those that do not, we can bootstrap the SE of any estimate - e.g., the median
- Bootstrapped estimates (mean of simulated replicates) can be used to assess bias
- Bootstrapping is not a panacea - requires a good sample size to start