resampling of observerad date
of a known data set

THE BOOTSTRAP

each sample has sam # of obs as original data set

I. The bootstrap is a computer-intensive method that is used in applied statistics. In its simplest form, the bootstrap is resampling of observed data. There are several purposes of the bootstrap. It can be used to obtain estimates of variance, standard errors, central tendency, and confidence intervals. In most instances the bootstrap is used when there is no sampling distribution or to apply a conventional sampling distribution is unwise.

II. In its simplest form, the bootstrap, to me, is an algorithm. You are going to use the observed data to generate samples (sometimes called pseudo-samples) and then construct a sampling distribution from the bootstrap samples. The number of samples to draw is usually 1000 to 10,000. In general, the more samples the better.

a. The algorithm
   i. What is the statistic (mean, variance, etc.).
   ii. What measure of variability, precision, central tendency, etc. of the statistic will be calculated.
   iii. Select the number of bootstrap samples to obtain. * Each bootstrap sample is the same size as the original data set.
   iv. Third, using sampling with replacement randomly select the number of bootstrap samples. For example, if the original data set has 20 observations and I am going to generate 1000 bootstrap samples; then I will use sampling with replacement to select 20,000 observations from the original data set. The first 20 are bootstrap sample 1, the next 20 values are bootstrap sample 2, so on.
   v. Calculate the statistic of each sample.
   vi. If needed, calculate the measure of variability, precision, central tendency, etc. of the statistic.

III. Example

a. Calculate the mean of pseudo-sample means. Also, express variability among pseudo-sample means by calculating the 2.5th and 97.5th percentiles of all pseudo-sample meals.
b. Generate 10,000 bootstrap samples.
c. For each sample calculate the mean.
d. Report the 2.5th and 97.5th percentiles of the 10,000 bootstrapped sample means.

IV. Caveat

a. The robustness of inferences using the bootstrap is founded in the original data. You have to assume it is representative of the population. Consequently, the bootstrap is not recommended with small sample sizes. I treat 'small' (when bootstrapping) as $n \leq 15$. Bootstrap - need 16 values

V. R example

a. The data set: $n = 20$ (weights of mice in grams in a museum collection). I will obtain 10,000 bootstrap pseudo-samples, each pseudo-sample $n = 20$, and report the mean of pseudo-sample means as well as the 2.5th and 97.5th percentiles of the 10,000 pseudo-sample means.

mean = 12.1

b. R

```
> orig.data=c(13,11,13,11,11,10,11,14,13,13,14,11,14,12,14,12,11,12,11,11)
> boot.dump=matrix(sample(orig.data,200000,TRUE),nrow=20,ncol=10000)
> dim(boot.dump)#Dimensions of matrix boot.dump, 20 rows and 10,000
[1]    20 10000 columns#
> boot.mn=apply(boot.dump,2,mean)
 [1] 12.10098
> mean(boot.mn)#Mean of pseudo-sample means#
 [1] 12.10098
> quantile(boot.mn,c(0.025,0.975))
 2.5% 97.5%
11.55 12.65 # 2.5th percentile = 11.55, 97%th percentile = 12.65#
> hist(boot.mn)#What is the shape of the dist'n of sample means?#
> sd(boot.mn) #Standard deviation of pseudo-sample means#
[1] 0.2831644
> sd(orig.data)/sqrt(20) #FYI, Standard error of mean using conventional formula#
[1] 0.2892822
```

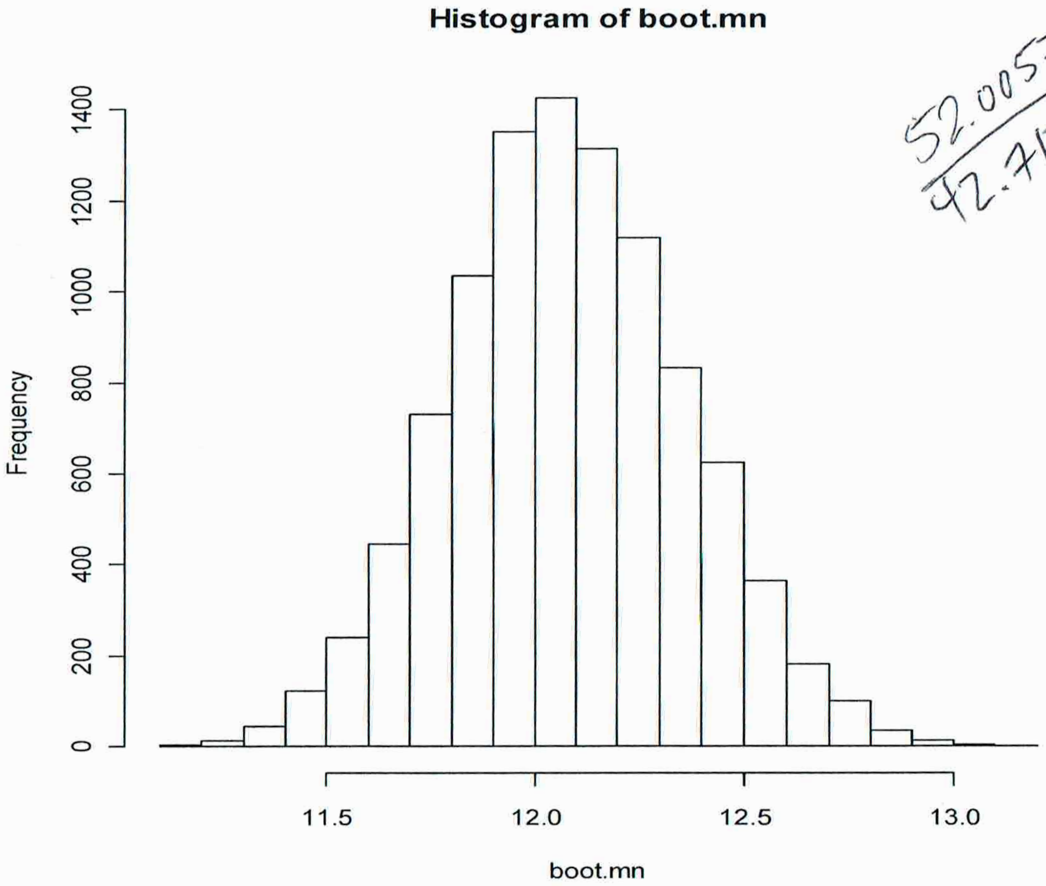
(.025, .05, .975) includes median

95% confidence interval #s should be lower + higher than mean

→ ~~Std dev of sample variability around mean~~

Use of orig data set

♣sd(boot.mn) and sd(orig.data)/sqrt(20) are similar, what does this tell you?



52.0053
42.7136

mean of sample & mean of means should be = to pop mean

.025 .975

par(mfrow = c(R,C))