L10

Monday, October 22, 2018
8:32 PM

<<Lecture10_SS.pdf>>

Stat 324 –
Introduction to Statistics for Engineers

LECTURE 10: BOOTSTRAP CI AND HYPOTHESIS TEST (REFERENCE SEC 5.8 IN OTT AND LONGDECKER * - USE LECTURE CI CONSTRUCTION)

Hypothesis testing and CI with a Bootstrap

What if data cannot be assumed to come from a normal distribution and our sample size is borderline small we're not confident the CLT will do its thing?

One option, is to do a bootstrap hypothesis test or confidence interval.

Ex: Secondhand smoke is of great health concern, especially for children. An SRS of 15 children is collected and the amount of cotanine (a metabolite of nicotine) in the urine was measured. Cotanine in unexposed children should be below 75 units. The data were as follows:

29, 30, 53, 75, 34, 21, 12, 58, 117, 119, 115, 134, 253, 289, 287

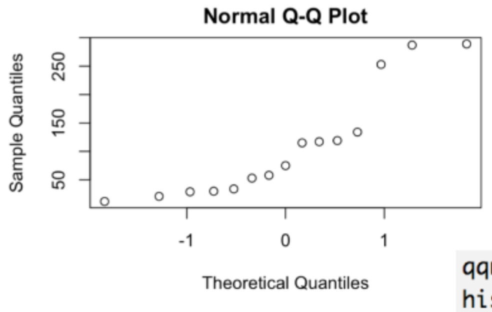
Make a 95% confidence interval for the mean amount of cotanine for the population these children were chosen from.

What first steps should we take to construct the interval?

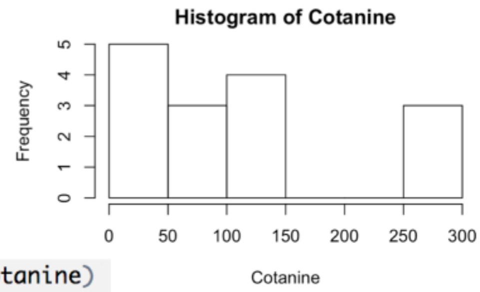
1. Look at QQ plot to see if normal population. distribution assumption is reasonably met.

check for independence

Bootstrap Confidence Interval




```
qqnorm(Cotanine)
hist(Cotanine)
```



It doesn't look great, and with the sample size on the edge of 'large' we might worry about the normal assumption (___CLT___ may not have kicked in yet).

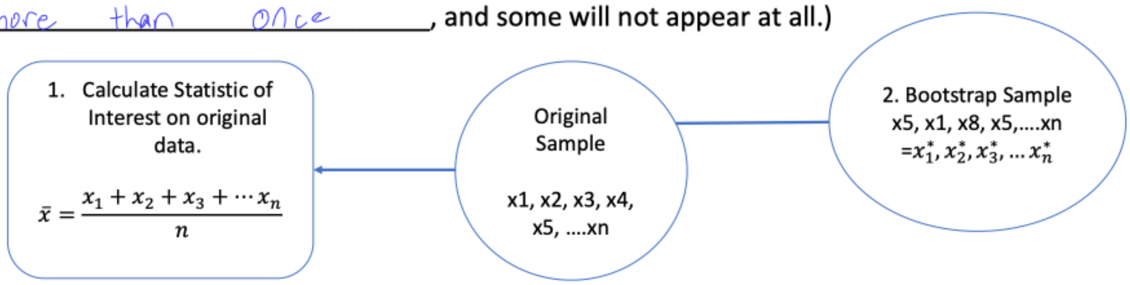
The problem with not assuming normality is that the quantity: $T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ will not have a t-distribution. So, a t-confidence interval will not have the coverage rate that it is theoretically supposed to.

Bootstrap T Distribution

We can instead use the data we have observed to make an estimate of what the statistic will look like. One method to construct bootstrap ci is to simulate a "T-like" distribution from your observed data:

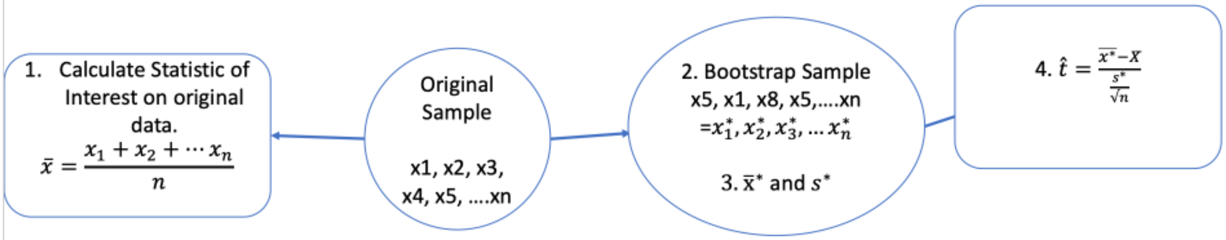
(1) Compute the estimate of the sample mean from the data sampled, ___$\bar{X}$___.

(2) Draw a simple random sample, ___with replacement___, of size n, from the sample data. Call these observations, $x_1^*, x_2^*, x_3^*, \dots x_n^*$. (Often this means that some of the data points will appear ___more than once___, and some will not appear at all.)



Bootstrap T Distribution

We can instead use the data we have observed to make an estimate of what the statistic will look like. One method to construct bootstrap ci is to simulate a "T-like" distribution from your observed data:

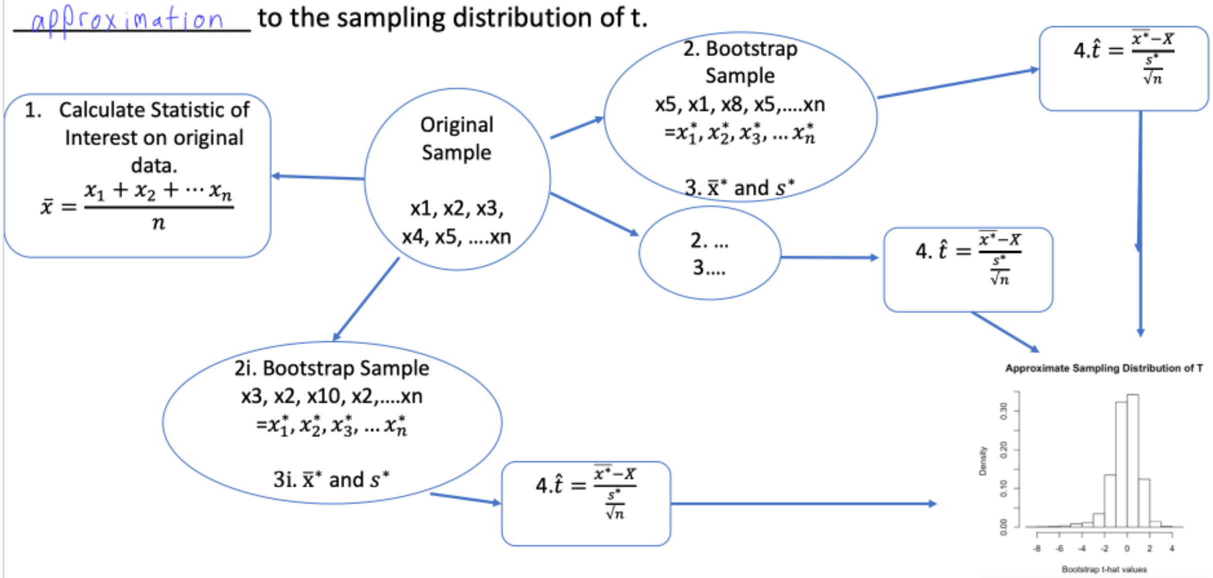


3. Compute the __mean__ and __sd__ of the resampled data. Call these things $\bar{x}^*$ and s^*.

4. Compute __statistic__ on the resampled data : $\hat{t} = \frac{\bar{X}^* - \bar{X}}{\frac{s^*}{\sqrt{n}}}$ (and retain this value).

Bootstrap T Distribution

5. Repeat steps 2-4 a large number of times, and compute $\hat{t}$ from each one. This is an __approximation__ to the sampling distribution of t.



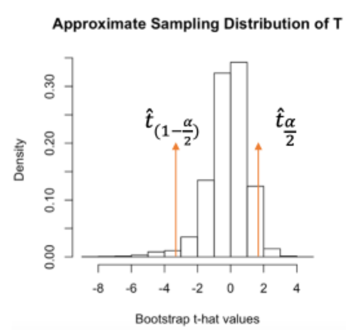
How can we do this re-sampling to see distribution of statistic?

```
```{r}
bootstrap = function(x, n.boot) {
1.  n = length(x)
2.  x.bar <- mean(x)
3.  t.hat <- numeric(n.boot) # create vector of length n.boot zeros
4.  for(i in 1:n.boot) {
5.    x.star <- sample(x, size=n, replace=TRUE)
6.    x.bar.star <- mean(x.star)
7.    s.star <- sd(x.star)
8.    t.hat[i] <- (x.bar.star - x.bar) / (s.star / sqrt(n))
  }
9.  return(t.hat)
}
```
```

Bootstrap Confidence Interval

6. Find the $1-\alpha/2$ and $\alpha/2$ upper critical values of $\hat{t}$ so the desired $100\left(1-\frac{\alpha}{2}\right)\%$ middle of $\hat{t}$ are between $\hat{t}_{(1-\frac{\alpha}{2})}$ and $\hat{t}_{\frac{\alpha}{2}}$.

7. Then make a $100\left(1-\frac{\alpha}{2}\right)\%$ CI :



lower $L = \bar{x} - \hat{t}_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ upper $U = \bar{x} - t_{(1-\frac{\alpha}{2})} \frac{s}{\sqrt{n}}$

*Just use your t relationship to back solve for the boundaries
*Notice the "upper" t-hat value is subtracted to make the lower boundary

(This equation is different from that given in your book – but this equation matches how we found t confidence intervals originally.)

Bootstrap Confidence Interval Smoke Data

E.g. For the secondhand smoke data, we find $\bar{x} = 108.4$ and $s = 95.6$. Bootstrapping 5000 times yields the following approximate distribution of t:

```
t.lower <- quantile(smoke.boot, probs=.025)
t.upper <- quantile(smoke.boot, probs=.975)
```

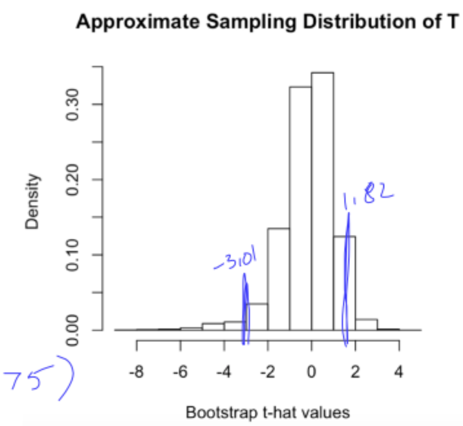
```
> t.lower
     2.5%
-3.012355
```

```
> t.upper
  97.5%
1.82445
```

$L = 108.4 - 1.82\left(\frac{95.6}{\sqrt{15}}\right) = 63.37$

$U = 108.4 - -3.01\left(\frac{95.6}{\sqrt{15}}\right) = 182.75$

$(63.37, 182.75)$



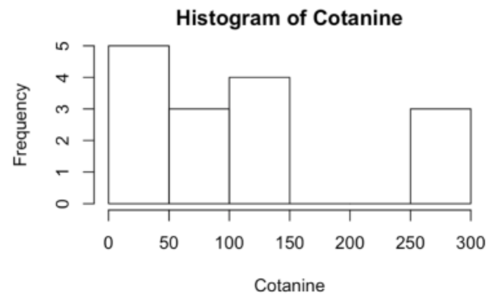

```
ci.low  = x.bar - t.upper * s / sqrt(n) # This is our lower interval endpoint.
ci.high = x.bar - t.lower * s / sqrt(n) # This is our upper interval endpoint.
interval = c(ci.low, ci.high)
names(interval)=NULL
print(interval)
```

Bootstrap Confidence Interval Smoke Data

E.g. For the secondhand smoke data, we find $\bar{x} = 108.4$ and $s = 95.6$.

Bootstrap: (63.37, 182.76)

T: (55.5, 161.3)



Similarities:
Both contain sample mean
Both utilize sample sd

Differences:
T has Normality of population (or CLT) assumption
T has symmetric interval around sample mean

Bootstrap requires computer resampling
Bootstrap has larger ME for values above[/below] sample mean

Bootstrap Hypothesis Test for $H_0: \mu = \mu_o$

1. Collect one simple random sample from population of size n.
   Compute sample mean $\bar{x}$, sample sd, s and $t_{obs} = \frac{\bar{x}-\mu_o}{s/\sqrt{n}}$.

2. Draw a random sample of size n from the data, with replacement. Call these observations: $x_1^*, x_2^*, \dots, x_n^*$. (same)

3. Compute the mean and standard deviation of the bootstrap sample. Call these: $\bar{x}^*$ and s^*. (same)

4. Compute the bootstrap statistic: $\widehat{t_B} = \frac{\overline{x^*} - \bar{x}}{s^*/\sqrt{n}}$. (same)

5. Repeat steps 2-4 a large number, B, times, accumulating B $\hat{t}$'s. These values approximate the sampling distribution of T. (same)

6. Find the p-value (the area under the approximate sampling distribution of T),given by: $\frac{m}{B}$, where m depends on H_A:
   $H_A: \mu > u_o \Rightarrow m_U$ is the number of values of $\hat{t}$ such that $\hat{t} > t_{obs}$
   $H_A: \mu < u_o \Rightarrow m_l$ is the number of values of $\hat{t}$ such that $\hat{t} < t_{obs}$
   $H_A: \mu \neq u_o \Rightarrow m = 2 * \min(m_l\, m_u)$

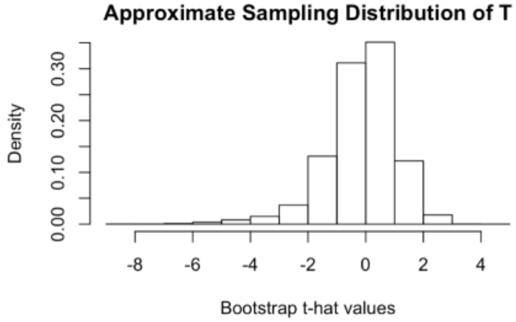
7. Draw conclusion: pvalue $< \alpha$ reject H_0; otherwise fail to reject H_0.

Hypothesis testing with a Bootstrap cont. (reran bootstrap for 5000 times)

29, 30, 53, 75, 34, 21, 12, 58, 117, 119, 115, 134, 253, 289, 287

$H_0: \mu = 75$

$\boldsymbol{H_0}: \mu = 75$; $\boldsymbol{H_a > 75}$ $\bar{X} = 108.4$ Sd(X)=95.60, so $t_{obs} = \frac{108.4-75}{95.6/\sqrt{15}} = 1.3551$



Range of $\hat{t}$ Values	Number of $\hat{t}$ Values
$\hat{t} \leq -1.35$	648
$-1.35 < \hat{t} < 1.35$	4352
$\hat{t} \geq 1.35$	360

```
> quantile(smoke.boot,probs=c(0.025,0.05,0.10,0.90,.95,0.975))
     2.5%        5%       10%       90%       95%     97.5%
-3.012355 -2.150563 -1.575449  1.184751  1.522036  1.824450
```

1.35 so approximate p value is between 5% - 10%

$0.05 < pvalue < 0.10$

$\boxed{H_A: \mu > 75, \quad \text{pvalue}: \frac{360}{5000} = 0.073}$

$H_A: \mu < 75$, pvalue: $\frac{648+4352}{5000} = 0.928$

$H_A: \mu \neq 75$, pvalue: $\frac{2 * \min(360, 648)}{5000} = \frac{720}{5000} = 0.146$

Hypothesis testing with a Bootstrap Compared with T

29, 30, 53, 75, 34, 21, 12, 58, 117, 119, 115, 134, 253, 289, 287

$\boldsymbol{H_0}: \mu = 75$; $\boldsymbol{H_a > 75}$ $\bar{X} = 108.4$ Sd(X)=95.60, so $t_{obs} = \frac{108.4-75}{95.60/\sqrt{15}} = 1.3531$

Bootstrap: Pvalue= $\frac{360}{5000} = 0.073$

T Test: $pvalue = P(T_{14} > 1.3531) = 0.05 < p - value < 0.10$

___lack___ of evidence against the null in this case. Bootstrap tends to be a bit more ____________________ (give higher p value and therefore less evidence against null).

In practice, I will often do both a t test and bootstrap and go with most conservative (unless Journal prefers one over the other or there is strong scientific knowledge about distribution of population).

For Next Time

- Continue working on Homework 4
- Review Returned Exam
- Quiz 3 Posted Soon….