Central Limit Theorem Applications Solutions

Lt Col Ken Horton Professor Bradley Warner

18 July, 2020

Exercises

1. Suppose we roll a fair six-sided die and let X be the resulting number. The distribution of X is discrete uniform. (Each of the six discrete outcomes is equally likely.)

   a. Suppose we roll the fair die 5 times and record the value of $\bar{X}$, the *mean* of the resulting rolls. Under the central limit theorem, what should be the distribution of $\bar{X}$?

The mean of X is 3.5 and the variance of $X = \frac{(b-a+1)^2-1}{12} = \frac{35}{12}$ is 2.9167. So,

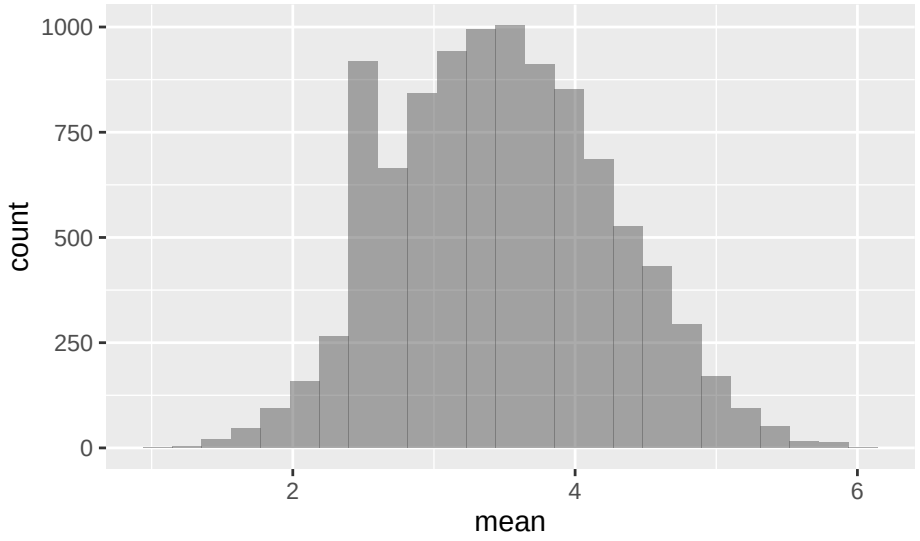
$$\bar{X} \overset{approx}{\sim} \text{Norm}(3.5, 0.764)$$

   b. Simulate this process in `R`. Plot the resulting empirical distribution of $\bar{X}$ and report the mean and standard deviation of $\bar{X}$. Was it what you expected?

(HINT: You can simulate a die roll using the `sample` function. Be careful and make sure you use it properly.)

```
set.seed(2003)
results<-do(10000)*mean(sample(6,5,replace=T))
```

```
results %>%
   gf_histogram(~mean)
```



```
favstats(~mean,data=results)
```

```
##  min Q1 median Q3 max    mean       sd     n missing
##    1  3    3.6  4   6 3.51278 0.772254 10000       0
```

It appears to be roughly normally distributed with the mean and standard deviation we expected.

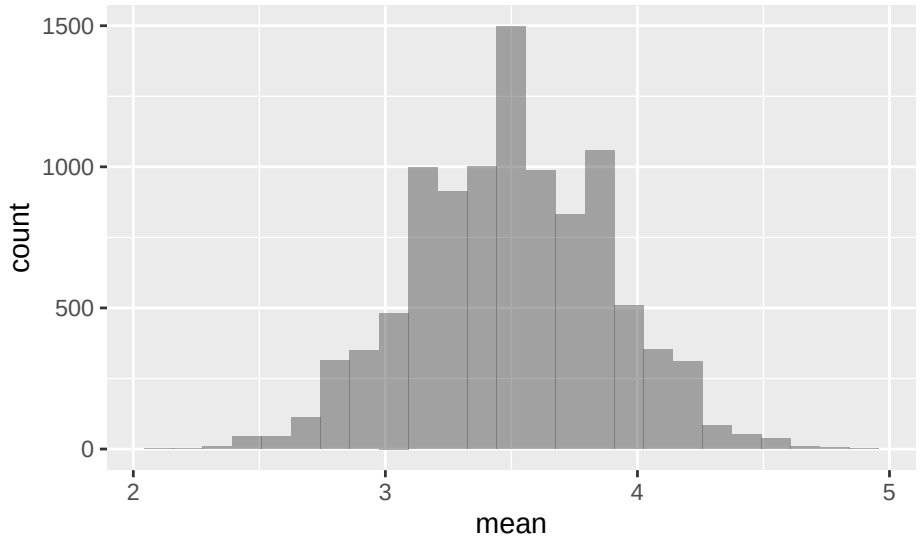
c. Repeat parts a) and b) for $n = 20$ and $n = 50$. Describe what you notice. Make sure all three plots are plotted on the same x-axis scale. You can use facets if you combine your data into one `tibble`.

When $n = 20$:

$$\bar{X} \overset{approx}{\sim} \text{Norm}(3.5, 0.382)$$

```
results2<-do(10000)*mean(sample(6,20,replace=T))
```

```
results2 %>%
   gf_histogram(~mean)
```




```
favstats(~mean,data=results2)
```

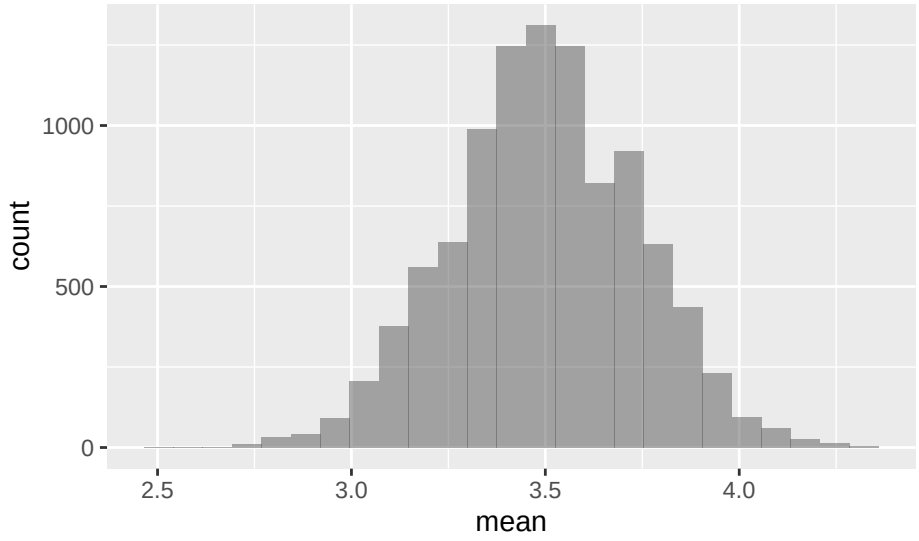
```
##   min   Q1 median   Q3  max    mean        sd     n missing
##  2.15 3.25    3.5 3.75 4.95 3.49896 0.3828754 10000       0
```

When $n = 50$:

$$\bar{X} \overset{approx}{\sim} \text{Norm}(3.5, 0.242)$$

```
results3<-do(10000)*mean(sample(6,50,replace=T))
```

```
results3 %>%
   gf_histogram(~mean)
```



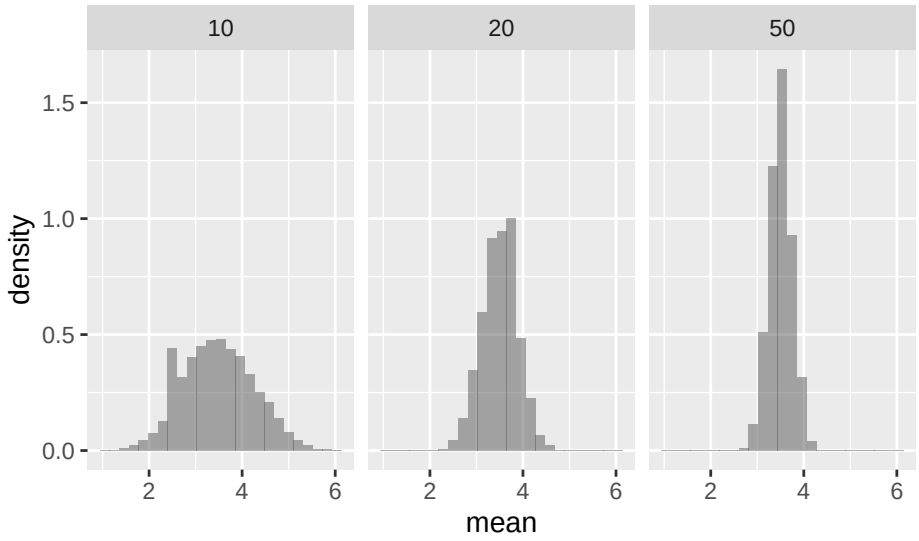

```
favstats(~mean,data=results3)
```

```
##   min   Q1 median   Q3  max    mean        sd     n missing
##  2.54 3.34    3.5 3.66 4.36 3.49852 0.2423665 10000       0
```

Now let's put them all together to make it easier to compare.

```
final_results<-rbind(cbind(results,n=10),cbind(results2,n=20),cbind(results3,n=50))
```

```
final_results %>%
   gf_dhistogram(~mean|n)
```




```
favstats(~mean|n,data=final_results) %>%
   select(mean,sd,n)
```

```
##      mean        sd  n
## 1 3.51278 0.7722540 10
## 2 3.49896 0.3828754 20
## 3 3.49852 0.2423665 50
```

All results were as expected. As n increased, the variance of the sample mean decreased.

2. The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a sample mean of 134 calories with a standard deviation of 17 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips? The conditions necessary for applying the normal model have been checked and are satisfied.

The question has been framed in terms of two possibilities: the nutrition label accurately lists the correct average calories per bag of chips or it does not, which may be framed as a hypothesis test.

a. Write the null and alternative hypothesis.

H_0: The average is listed correctly. $\mu = 130$
H_A: The nutrition label is incorrect. $\mu \neq 130$

b. What level of significance are you going to use?

I am going to use $\alpha = 0.05$.

c. What is the distribution of the test statistic $\frac{\bar{X}-\mu}{S/\sqrt{n}}$? Calculate the observed value.

The distribution of the test statistic is t with 34 degrees of freedom.

The observed average is $\bar{x} = 134$ and the standard error may be calculated as $SE = \frac{17}{\sqrt{35}} = 2.87$.

We can compute a test statistic as the t score:

$$t = \frac{134 - 130}{2.87} = 1.39$$

d. Calculate a p-value.

The upper-tail area is 0.0823,

```
pt(1.39,34,lower.tail = F)
```

```
## [1] 0.08678153
```

or

```
1-pt(1.39,34)
```

```
## [1] 0.08678153
```

so the p-value is $2 \times 0.0823 = 0.1646$.

e. Draw a conclusion.

Since the p-value is larger than 0.05, we do not reject the null hypothesis. That is, there is not enough evidence to show the nutrition label has incorrect information.

Extra material

If we had used a normal model from the CLT our p-value would have been close to the value from the t because our sample size is large.

```
pnorm(1.39,lower.tail = F)
```

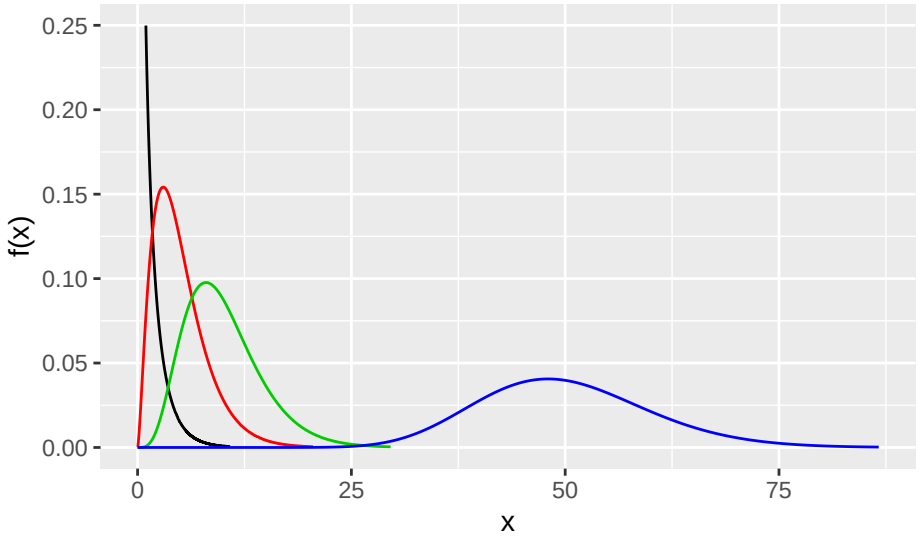
```
## [1] 0.08226444
```

3. Exploration of the chi-squared and t distributions.

   a) In R, plot the pdf of a random variable with the chi-squared distribution with 1 degree of freedom. On the same plot, include the pdfs with degrees of freedom of 5, 10 and 50. Describe how the behavior of the pdf changes with increasing degrees of freedom.

scale_color_manual(name = 'data source', values =c('b'='blue','a'='orange'), labels = c('df','empirical_data'))

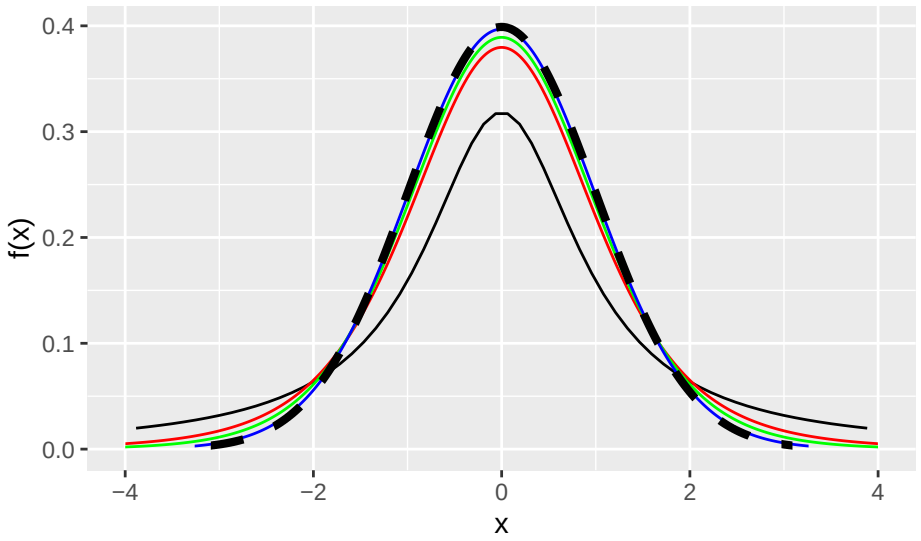
```
gf_dist("chisq",df=1,col=1) %>%
   gf_dist("chisq",df=5,col=2) %>%
   gf_dist("chisq",df=10,col=3) %>%
   gf_dist("chisq",df=50,col=4) %>%
   gf_lims(y=c(0,.25)) %>%
   gf_labs(y="f(x)")
```



The "bump" moves to the rights as the degrees of freedom increase.

   b) Repeat part (a) with the t distribution. Add the pdf of a standard normal random variable as well. What do you notice?

```
gf_dist("t",df=1,col="black") %>%
   gf_dist("t",df=5,col="red") %>%
   gf_dist("t",df=10,col="green") %>%
   gf_dist("t",df=50,col="blue") %>%
   gf_dist("norm",lty=2,lwd=1.5) %>%
   gf_lims(x=c(-4,4)) %>%
   gf_labs(y="f(x)")
```



As degrees of freedom increases, the t-distribution approaches the standard normal distribution.

4. In this lesson, we have used the expression *degrees of freedom* a lot. What does this expression mean? When we have sample of size n, why are there $n-1$ degrees of freedom? Give a short concise answer (about one paragraph). You will likely have to do a little research on your own.

Answers will vary. One possible explanation is that the degrees of freedom represents the number of indpendent pieces of information. For example, you'll notice that in order to get an unbiased estimate of σ^2, we have to divide by $n-1$. This is because in order to estimate σ^2, we need to first estimate μ, which is done by obtaining the sample mean. Once we know the sample mean, we only have $n-1$ pieces of independent information. For example, suppose we have a sample of size 10, and we know the sample mean. Once we are given the first 9 observations, we know exactly what the 10th observation must be.

5. Deborah Toohey is running for Congress, and her campaign manager claims she has more than 50% support from the district's electorate. Ms. Toohey's opponent claimed that Ms. Toohey has **less** than 50%. Set up a hypothesis test to evaluate who is right.

a. Should we run a one-sided or two-sided hypothesis test?

We should run a two-sided. She could be greater than 50% regardless of what the opponent claims.

b. Write the null and alternative hypothesis.

H_0: Ms. Toohey's support is 50%. $p = 0.50$.

H_A: Ms. Toohey's support is either above or below 50%. $p \neq 0.50$.

c. What level of significance are you going to use?

$\alpha = 0.05$

d. What are the assumptions of this test?

e. The observations are independent.

ii. There are at least 5 votes for her and 5 against.

Because this is a simple random sample that includes fewer than 10% of the population, the observations are independent. In a single proportion hypothesis test, the success-failure condition is checked using the null proportion, $p_0 = 0.5$: $np_0 = n(1 - p_0) = 500 \times 0.5 = 250 > 5$. With these conditions verified, the normal model may be applied to $\hat{p}$.

e. Calculate the test statistic.

A newspaper collects a simple random sample of 500 likely voters in the district and estimates Toohey's support to be 52%.

The test statistic is $\bar{x} = 0.52$

f. Calculate a p-value.

Based on the normal model, we can compute a one-sided p-value and then double to get the correct p-vlaue.

The standard error can be computed. The null value is used again here, because this is a hypothesis test for a single proportion with the specified value for the probability of success.

$$SE = \sqrt{\frac{p_0 \times (1 - p_0)}{n}} = \sqrt{\frac{0.5 \times (1 - 0.5)}{500}} = 0.022$$

```
2*pnorm(.52,mean=.5,sd=0.022,lower.tail = FALSE)
```

```
## [1] 0.3633021
```

g. Draw a conclusion.

Because the p-value is larger than 0.05, we do not reject the null hypothesis, and we do not find convincing evidence to support the campaign manager's claim.