Central Limit Theorem Notes

Lt Col Ken Horton   Professor Bradley Warner

18 July, 2020

Objectives

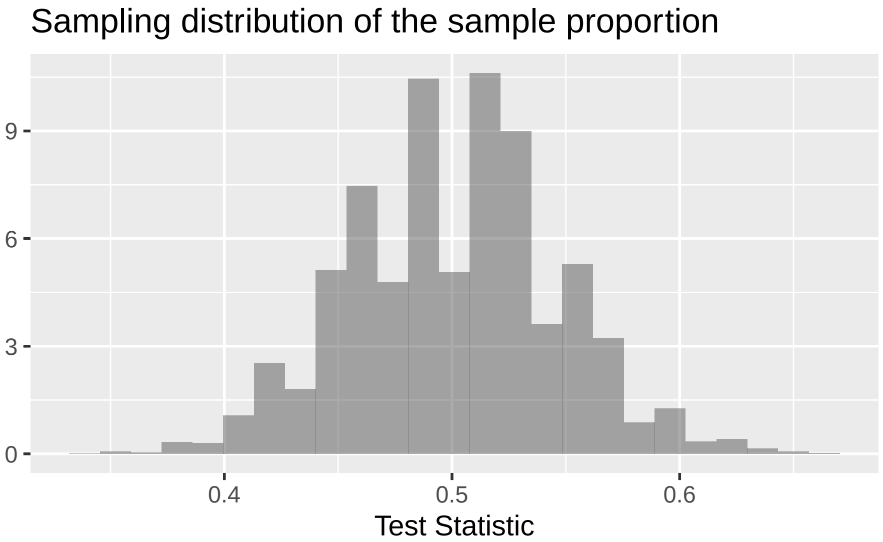
1) Explain the central limit theorem and describe, in general, how it can be used to make inferences.
2) Conduct hypotheses test of a single mean and proportion using the CLT and `R`.
3) Understand the chi-squared and t distributions and describe their parameters.

Central Limit Theorem

We've encountered several problems so far this chapter. While they differ in the settings, in their outcomes, and also in the technique we've used to analyze the data, they all have something in common: for a certain class of test statistics, the general shape of the null distribution.

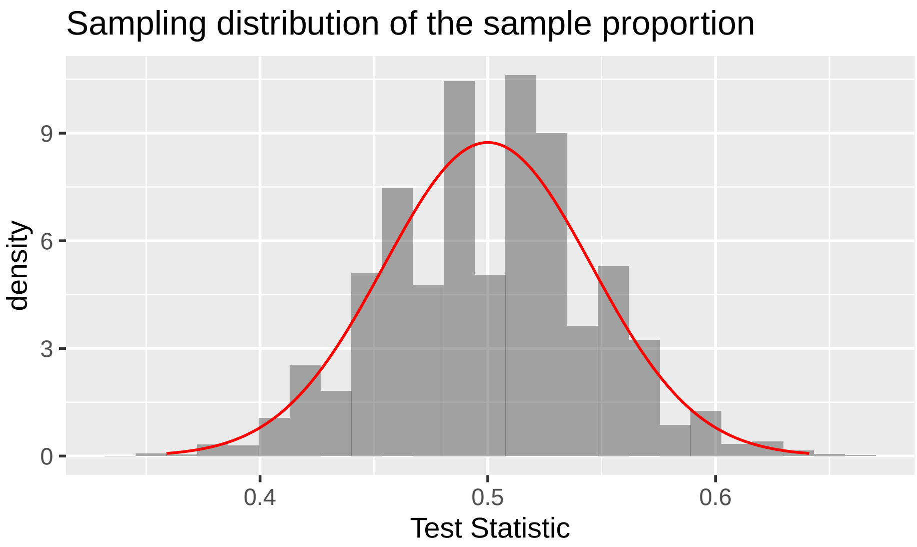
Null distribution

As a reminder, if the tapping and listening problem, we used the proportion of correct answers as our test statistic. Under the null hypothesis we assumed the probability of success was 0.5. The sampling distribution of our test statistic is shown in the figure below.



Exercise:
Describe the shape of the distribution and note anything that you find interesting.[1]

In the following figure we have overlayed a normal distribution on the histogram, this allows us to visual compare a normal probability density curve with the empirical distribution of the sampling distribution.



As we observed in the first and second blocks, many of the sampling distributions we've so far encountered have all looked somewhat similar and, for the most part, symmetric. They all resemble a bell-shaped curve. This is not a coincidence, but rather, is guaranteed by mathematical theory. This lesson will be a little more mathematical than the previous lessons. However, the goal is to develop a tool that will help us find sampling distributions for test statistics and thus find p-values. Remember that before the advances of modern computing, these mathematical solutions were all that was available.

Theorem

Theorem: Let $X_1, X_2, ..., X_n$ be a sequence of iid random variables from a distribution with mean μ and standard deviation $\sigma < \infty$. Then,

$$\bar{X} \overset{approx}{\sim} \textsf{Norm}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

There is a lot going on in this theorem. First is that we are drawing independent samples from the sample parent population. The central limit theorem (CLT) does not specify the form of this distribution, only that it has a finite variance. If we then form a new random variable, in this case the sample mean, the distribution is approximately normal with the same mean as the parent population and a variance that is the variance of the parent population divided by the sample size. Let's summarize these ideas.

1. The process of creating a new random variable from independent identically distributed random variable is approximately normal.
2. The approximation to a normal distribution improves with sample size n.
3. The mean and variance of the sampling distribution are a function of the mean and variance of the parent population and the sample size n.

[1] In general, the distribution is reasonably symmetric. It is unimodel and looks like a normal distribution.

If you go back and review examples, exercises, and homework problems from the previous lesson, you will see that we get symmetric normal “looking” distributions when we created test statistics that involved the process of summing. The example of a skewed distribution was the golf ball example where our test statistic was the difference of the max and min.

It is hard to overstate the historical importance of this theorem to the field of inferential statistics and science in general.

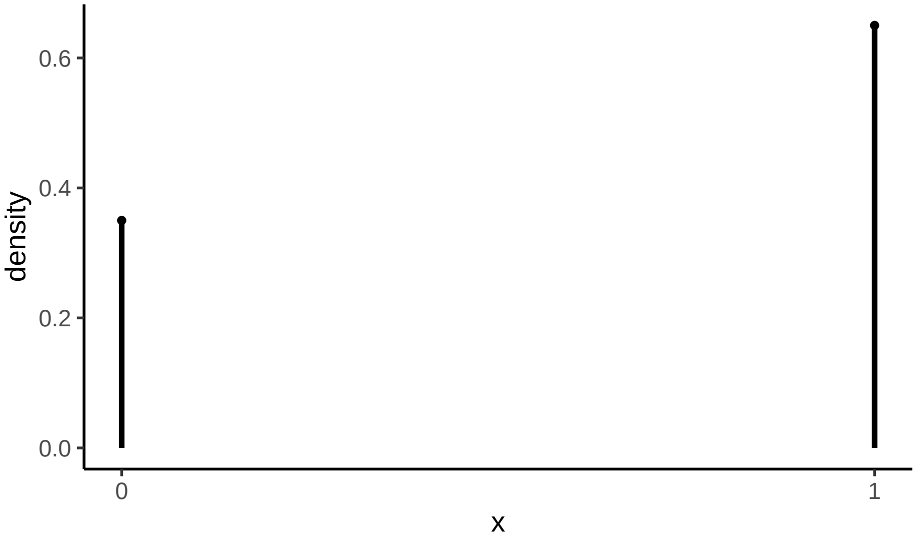
To get a sense of the central limit theorem, let’s simulate some data.

Simulating data for CLT

We are going to use an artificial example where we know the population distribution and parameters. We will repeat sampling from this many times and plot the distribution of the summary statistic of interest to demonstrate the CLT. This is purely an educational thought experiment to convince ourselves about the validity of the CLT.

Suppose there is an upcoming election in Colorado and Proposition A is on the ballot. Now suppose that 65% of Colorado voters support Proposition A. We poll a random sample of n Colorado voters. Prior to conducting the sample, I can think about the sample as a sequence of iid random variables from the binomial distribution with 1 trial and a probability of success (support for the measure) of 0.65. In other words, each random variable will take value 1 (support) or 0 (oppose). We can plot the pmf of the parent distribution ($\textsf{Binom}(1, 0.65)$):

```r
gf_dist("binom",size=1,prob=.65,plot_size=1) %>%
  gf_theme(theme_classic()) %>%
  gf_theme(scale_x_continuous(breaks = c(0,1)))
```



This is clearly not normal, it is in fact discrete. The mean of X is 0.65 and the standard deviation is $\sqrt{0.65(1-0.65)} = 0.477$.

In our first simulation $n = 10$. In the code box below, we will obtain a sample of size 10 from this distribution and record the mean $\bar{X}$, which is our method of moments estimate of the probability of success. We will repeat this 10,000 times to get an empirical distribution of $\bar{X}$. (Note that $\bar{X}$ is a mean of 1s and 0s and can be thought of as a proportion of voters in the sample that support the measure. Often, population proportion is denoted as π and the sample proportion is denoted as $\hat{\pi}$.)

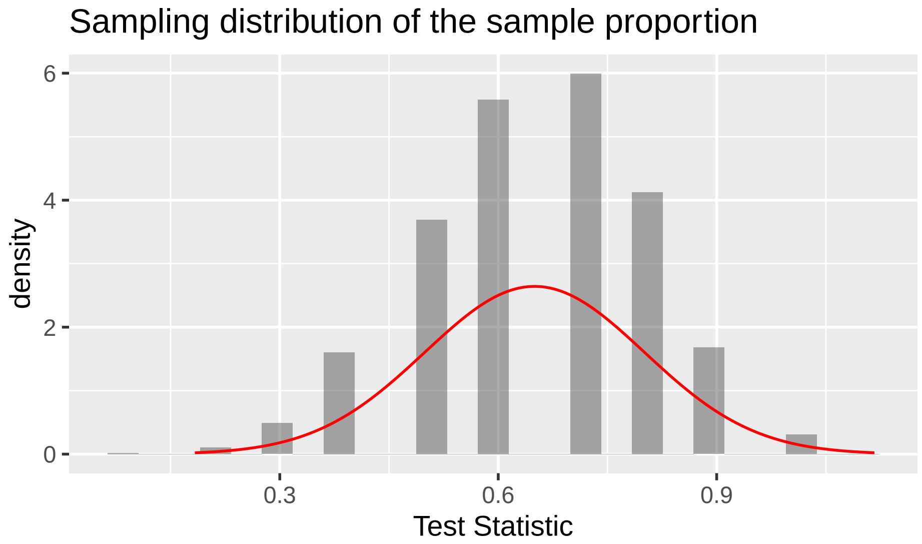
```r
set.seed(5501)
results<-do(10000)*mean(rbinom(10,1,0.65))
```

Since the central limit theorem applies, the distribution should appear of the sampling distribution of the mean should look normal. The mean should be close to 0.65, and the standard deviation $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.65(1-0.65)}{10}} = 0.151$

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

```
##  min  Q1 median  Q3 max    mean        sd     n missing
##  0.1 0.5    0.7 0.8   1 0.64932 0.1505716 10000       0
```

Remember from our lessons on probability, these results for the mean and standard deviation do not depend on the CLT, they are results from the properties of expectation on independent samples. The distribution of the sampling mean, approximately normal, does depend on CLT.



Note that the sampling distribution of the sample mean has a bell-curvish shape, but with some skew to the left for this particular sample size. That is why we state that the approximation improves with sample size. As a way to determine the impact of the sample, let's record how often did the sample fail to indicate support for the measure. (How often was the sample proportion less than or equal to 0.5?)

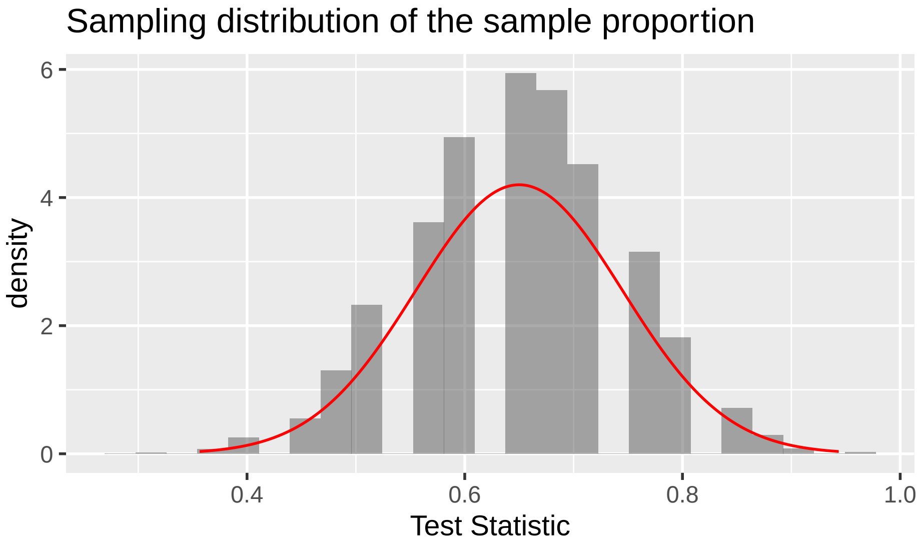
```r
results %>%
 summarise(low_result=mean(~mean<=0.5))
```

```
##   low_result
## 1     0.2505
```

Even though we know that 65% of Colorado voters support the measure, a sample of size 10 failed to indicate support 25.05% of the time.

Let's take a larger sample. In the box below, I will repeat the above but with a sample of size 25:

```
set.seed(5501)
results<-do(10000)*mean(rbinom(25,1,0.65))
```

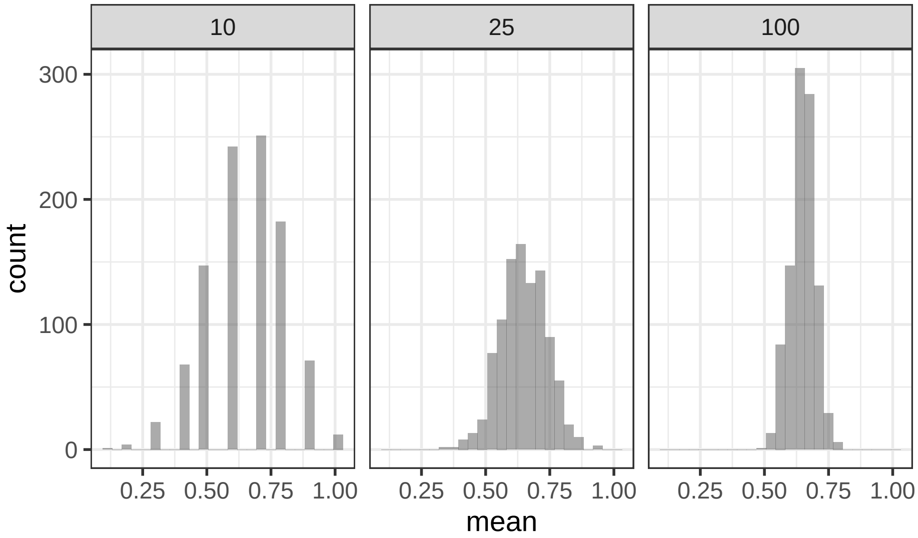


```
results %>%
 summarise(low_result=mean(~mean<=0.5))
```

```
##   low_result
## 1     0.0623
```

When increasing the sample size to 25, the standard deviation of our sample proportion decreased. According to the central limit theorem, it should have decreased to $\sigma/\sqrt{25} = 0.095$. Also, the skew became less severe. Further, the sample of size 25 failed to indicate support only 6.12% of the time. It reasonably follows that an even larger sample would continue these trends. The following plot demonstrates these trends.

```
clt %>%
  gf_histogram(~mean) %>%
  gf_facet_grid(~samp) %>%
  gf_theme(theme_bw())
```



Summary of example

A quick note on the previous example. In this example, we knew the true proportion of voters who supported the proposition. Based on that knowledge, we simulated the behavior of the sample proportion. We did this by taking a sample of size n, recording the sample proportion and repeating that process thousands of times. In reality, we will not know the true underlying level of support; further, we will not take a sample thousands of times. Sampling can be expensive and time-consuming. Thus, we would take one random sample of size n, and acknowledge that the resulting sample proportion is but one observation from an underlying normal distribution. We would then figure out what values of π (the true unknown population proportion) could reasonably have resulted in the observed sample proportion.

Other Estimates

There are some useful results that come from the central limit theorem. A large part of theoretical statistics has been mathematically deriving the distribution of sample statistics. In these methods we obtain a sample statistic, determine the distribution of that statistic under certain conditions, and then use that information to make a statement about the population parameter.

Estimating Variance

Recall that the central limit theorem tells us that for reasonably large sample sizes, $\bar{X} \overset{approx}{\sim} \mathsf{Norm}(\mu, \sigma/\sqrt{n})$. However, this expression involves two unknowns: μ and σ. In the case of binary data (as in Example 23.2), population variance is a function of population proportion ($\text{Var}(X) = \pi(1-\pi)$), so there is really just one unknown. In the case of continuous data (as in Example 23.1), standard deviation would need to be estimated.

Let S^2 be defined as:

$$S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

Recall that this is an unbiased estimate for σ^2.

Lemma: Let $X_1, X_2, ..., X_n$ be an iid sequence of random variables from a normal population with mean μ and standard deviation σ. Then,

$$\frac{(n-1)S^2}{\sigma^2} \sim \mathsf{Chisq}(n-1)$$

The $\mathsf{Chisq}(n-1)$ distribution is read as the "chi-squared" distribution ("chi" is pronounced "kye"). The chi-squared distribution has one parameter: degrees of freedom. The chi-squared distribution is used in other contexts as well such as the golf ball example, we will discuss this in later lessons.

The proof of this lemma is outside the scope of this class, but it is not terribly complicated. It follows from the fact that the sum of n squared random variables, each with the standard normal distribution, follows the chi-squared distribution with n degrees of freedom.

This lemma can be used to draw inferences about σ^2. For a particular value of σ^2, we know how S^2 should behave. So, for a particular value of S^2, we can figure out reasonable values of σ^2.

In practice, one rarely estimates σ for the purpose of inferring on σ. Typically, we are interested in estimating μ and we need to account for the added uncertainty in estimating σ as well. That is what we will discuss in the following section.

Estimating Mean

Let $X_1, X_2, ..., X_n$ be an iid sequence of random variables, each with mean μ and standard deviation σ. Recall that the central limit theorem tells us that

$$\bar{X} \overset{approx}{\sim} \mathsf{Norm}(\mu, \sigma/\sqrt{n})$$

Rearranging:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{approx}{\sim} \mathsf{Norm}(0,1)$$

Again, σ is unknown. Thus, I have to estimate it. I can estimate it with S, but now I need to know the distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$. This *does not* follow the normal distribution.

Lemma: Let $X_1, X_2, ..., X_n$ be an iid sequence of random variables from a normal population with mean μ and standard deviation σ. Then,

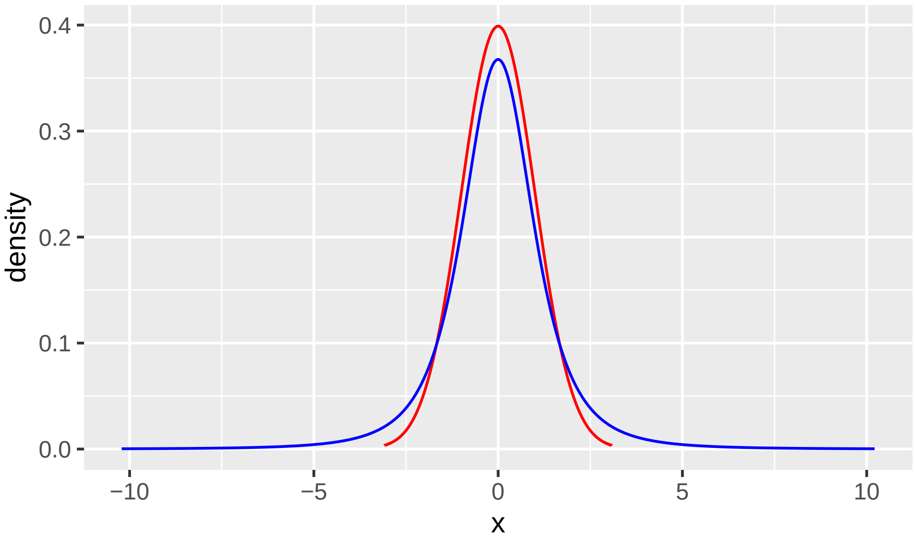
$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim \mathsf{t}(n-1)$$

The $\mathsf{t}(n-1)$ distribution is read as the "t" distribution. The t distribution has one parameter: degrees of freedom. The expression above $\left(\frac{\bar{X}-\mu}{S/\sqrt{n}}\right)$ is referred to as the t statistic.

Similar to the chi-squared distribution, we won't go over the proof, but it follows from some simple algebra and from the fact that the ratio between a standard normal random variable and the square root of a chi-squared random variable, divided by it's degrees of freedom follows a t distribution.

The t distribution is very similar to the standard normal distribution, but has longer tails. This seems to make sense in the context of estimating μ since substituting S for σ adds variability to the random variable. A t distribution, shown as a blue line, has a bell shape that looks very similar to a normal distribution, red line. However, its tails are thicker, which means observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution. When our sample is small, the value s used to compute the standard error isn't very reliable. The extra thick tails of the t distribution are exactly the correction we need to resolve this problem. When the degrees of freedom is about 30 or more, the t distribution is nearly indistinguishable from the normal distribution.

```
gf_dist("norm",color="red") %>%
  gf_dist("t",df=3,color="blue")
```



Important Note

You may have noticed an important condition in the two lemmas above. It was assumed that each X_i in the sequence of random variables was *normally* distributed. While the central limit theorem has no such normality assumption, the distribution of the *t*-statistic is subject to the distribution of the underlying population. With a large enough sample size, this assumption is not necessary. There is no magic number, but some resources state that as long as n is at least 30-40, the underlying distribution doesn't matter. For smaller sample sizes, the underlying distribution should be relatively symmetric and unimodal.

One advantage of simulation-based inference methods is that these methods do not rely on any such distributional assumptions.

Hypotheses tests using CLT

We are now ready to repeat some of our previous problems using the mathematically derived sampling distribution via the CLT.

Tappers and listeners

Step 1- State the null and alternative hypotheses Here are the two hypotheses:

H_0: The tappers are correct, and generally 50% of the time listeners are able to guess the tune. $p = 0.50$

H_A: The tappers are incorrect, and either more than or less than 50% of listeners will be able to guess the tune. $p \neq 0.50$

Step 2 - Compute a test statistic. The test statistic that we want to use is the sample mean $\bar{X}$, this is a method of moments estimate of the probability of success. Since these are independent samples from the same binomial distribution, by the CLT

$$\bar{X} \overset{approx}{\sim} \mathsf{Norm}\left(\pi, \frac{\pi(1-\pi)}{\sqrt{n}}\right)$$

As we learned, this approximation improves with sample size. As a rule of thumb, most analysts are comfortable with using the CLT for this problem if the number of success and failures are both 5 or greater.

In our study 42 out of 120 listeners ($\bar{x} = \hat{p} = 0.35$) were able to guess the tune. This is the test statistic.

Step 3 - Determine the p-value. We now want to find the p-value from the one-sided probability P($\bar{X} \leq 0.35$) given the null hypothesis is true, that the probability of success is 0.50. We will use `R` to get the one-sided value and then double.

```
2*pnorm(0.35,mean=.5,sd=sqrt(.5*.5/120))
```

```
## [1] 0.001015001
```

That is a small p-value and consistent with what we would using the exact binomial and the simulation empirical p-value.

> **Important note**: In the calculation of the standard deviation of the sampling distribution, we used the null hypothesized value of the probability of success.

Step 4 - Draw a conclusion Based on our data, if the listeners were guessing correct 50% of the time, there is less than a 1.3×10^{-20} probability that only 42 or less or 78 or more listeners would get it right. This is much less than 0.05, so we reject that the listeners are guessing correctly half of the time.

Now `R` has built in functions to perform this test and the exact. If you explore these functions, use `?prop.test`, you will find options to improve the performance of the test. You are welcome and should read about these. Again, before computers, researchers spent time optimizing the performance of the asymptotic methods such as the CLT.

Here is the test of a single proportion using `R`.

```
prop.test(42,120)
```

```
## 
##  1-sample proportions test with continuity correction
## 
## data:  42 out of 120
## X-squared = 10.208, df = 1, p-value = 0.001398
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2667083 0.4430441
## sample estimates:
##    p 
## 0.35
```

The p-value is small, reported as 0.0014. We will study the confidence interval soon so don't worry about that part of the output. The alternative hypothesis is listed.

Exercise:
How do you conduct a one-sided test? What if the null value where 0.45?[2]

```
pval(prop.test(42,120,alternative="less",p=.45))
```

```
##   p.value
## 0.0174214
```

The exact test uses `binom.test()`.

Body Temperature

We will repeat the body temperature analysis using the CLT. We will use $\alpha = 0.05$

Step 1- State the null and alternative hypotheses

H_0: The average body temperature is 98.6 $\mu = 98.6$

H_A: The average body temperature is less than 98.6. $\mu < 98.6$

Step 2 - Compute a test statistic. We don't know the population variance, so we will use the t distribution. Remember that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathsf{t}(n-1)$$

thus our test statistic is

$$\frac{\bar{x} - 98.6}{S/\sqrt{n}}$$

```
favstats(~temperature,data=temperature)
```

```
##   min   Q1 median   Q3   max     mean        sd   n missing
##  96.3 97.8   98.3 98.7 100.8 98.24923 0.7331832 130       0
```

```
temperature %>%
  summarise(mean=mean(temperature),sd=sd(temperature),test_stat=(mean-98.6)/(sd/sqrt(130)))
```

```
## # A tibble: 1 x 3
##    mean    sd test_stat
##   <dbl> <dbl>     <dbl>
## 1  98.2 0.733     -5.45
```

We are over 5 standard deviation below the null hypothesis mean. We have some assumptions that we will discuss at the end of this problem.

[2] We will only exact the p-value in this exercise

Step 3 - Determine the p-value. We now want to find the p-value from $P(t \leq -5.45)$ on 129 degrees of freedom.given the null hypothesis is true, that the probability of success is 0.50. We will use R to get the one-sided value and then double.

```
pt(-5.45,129)
```

```
## [1] 1.232178e-07
```

We could also use the R function `t_test()`.

```
t_test(~temperature,data=temperature,mu=98.6,alternative="less")
```

```
##
##  One Sample t-test
##
## data:  temperature
## t = -5.4548, df = 129, p-value = 1.205e-07
## alternative hypothesis: true mean is less than 98.6
## 95 percent confidence interval:
##      -Inf 98.35577
## sample estimates:
## mean of x
##  98.24923
```

Step 4 - Draw a conclusion Based our data, if the true mean body temperature is 98.6, then the probability of observing a mean of 98.25 or less is 0.00000012. This is too unlikely so we reject the hypothesis that the average body temperature is 98.6.

Summary When estimating the mean and standard error from a sample of numerical data, the t distribution is a little more accurate than the normal model. This is true for both small and large samples, though the benefits for larger samples are limited.

Use the t distribution for inference of the sample mean when observations are independent and nearly normal. You may relax the nearly normal condition as the sample size increases. For example, the data distribution may be moderately skewed when the sample size is at least 30, in a sense you are using the CLT and the fact that a t is close to a normal.

Before applying the t distribution for inference about a single mean, we check assumptions.

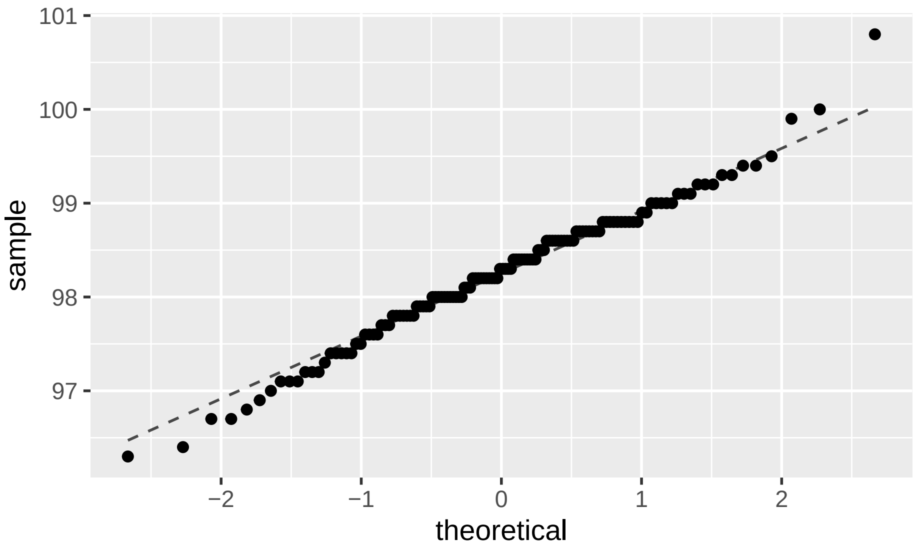
1. Independence of observations. This is a difficult assumption to verify. If we collect a simple random sample from less than 10% of the population, or if the data are from an experiment or random process, we feel better about this assumption. If the data comes from an experiment, we can plot the data versus time collected to see if there are any patterns that indicate a relationship. A design of experiment course discusses these ideas.

2. Observations come from a nearly normal distribution. This second condition is difficult to verify with small data sets. We often (i) take a look at a plot of the data for obvious departures from the normal model, usually in the form of prominent outliers, and (ii) consider whether any previous experiences alert us that the data may not be nearly normal. However, if the sample size is somewhat large, then we can relax this condition, e.g. moderate skew is acceptable when the sample size is 30 or more, and strong skew is acceptable when the size is about 60 or more.

A typical plot to use to evaluate is called the quantile-quantile plot. We form a scatterplot of the empirical quantiles from the data versus exact quantile values from the theoretical distribution. If the points fall along a line then the data match the distribution. An exact match is not realistic, so we look for major departures from the line.

Below is our normal-quantile plot for the data. The largest value may be an outlier, we may want to verify it was entered correctly. The fact that the point are above the line for the larger values and below the line for the smaller values indicates that our data may have longer tails than the normal. There are really only 3 values in the larger values so in fact the data may be slightly skewed to the left, this was also indicated my a comparison of the mean and median. However, since we have 130 data points these results should not impact our findings.

```
gf_qq(~temperature,data=temperature) %>%
  gf_qqline(~temperature,data=temperature)
```



We can also check the impacts of the assumptions by using other methods for the hypothesis test. If all methods lead to the same result, we can be confident in the results.

File Creation Information

- File creation date: 2020-07-18
- Windows version: Windows 10 x64 (build 17763)
- R version 3.6.3 (2020-02-29)
- `mosaic` package version: 1.6.0
- `tidyverse` package version: 1.3.0