Unit6: Inference

Correction to Unit 5

So far when we talked about the distribution of the sample mean, $\overline{X}$, we assumed it was normally (or t) distributed about the true mean μ, however, there are two conditions

1. The samples $X_1, X_2, ..., X_n$ are all independent random samples from population
2. When n is small ($\lessapprox 30$), we also requre that the sample observations come from a normally distributed population

If the sample size is small (e.g. $n = 8$) you can still get pretty "good" results as long as the samples don't deviate too strongly from normality - which is why the examples from last week worked!

Now, back to Unit 6: Inference...

What is Inference?

What do YOU think inference is?

Examples: jurors are performing inference in court cases. They are trying to **infer** from data (evidence) whether the plaintiff is not guilty or guilty.

What other examples can you think of?

inference is about estimating stuff when we don't know the whole story... making guesses from sample data in a **controlled** manner. For that we need both a guess (point estimate) as well as a sense of how good is the guess (confidence interval)

Point Estimation

- Use of **sample data** from the population to calculate a **single value** which is the best guess of a **population parameter**
- **Sampling error** describes how much the estimate will vary from one sample to the next. This is the standard deviation in the point estimator often denoted as standard error, SE
- **Bias** a systematic tendency to under, or overestimate the population parameter

Example

The sample mean $\overline{X}$ is an unbiased point estimator of the (true) population mean μ. The sampling error is just the standard error of the mean $SEM = \frac{\sigma}{\sqrt{n}}$.

Point estimator biases

- There are many different ways to estimate population parameters given some sample data
- For example, Suppose that I want to estimate the number of tickets in a box of N tickets, each ticket is labelled $1, 2, ..., N$
- I don't know how many tickets are in the box
- Suppose I take one ticket out of the box and read the number, call it X

- What are some ways that I could estimate the total number of tickets N, from this single ticket value X?

$\hat{N} = X$ is the biased *maximum likelihood* esimator for N

$\hat{N} = 2X$ would be an unbiased estimator for N*

Let's say $N = 100$, repeat the experiment $m = 1000$ times

```
box_of_tickets <- 1:100
mean(replicate(1000,sample(box_of_tickets,size=1)))   # biased
```

```
## [1] 51.026
```

```
mean(2*replicate(1000,sample(box_of_tickets,size=1)))  # unbiased
```

```
## [1] 101.488
```

*You can use your knowledge that X is discrete uniform with expectation $E(X) = N/2$ to form $\hat{N} = 2X$

Example with proportions

We will use the following example to demonstrate confidence intervals for a proportion

1. Suppose that of 300,000 US citizens in a certain district, a proportion, $p = 0.40$, of them "approve" of President Trump
2. In order to know this you would have to ask 300,000 people their opinion. What if instead we took a sample of $n = 1000$ observations and estimated the true proportion using the sample proportion, $\hat{p}$ as a point estimator *for* p
3. Since there will be variability in the point estimate that will depend on n, we would like to give both the point estimator as well as **a range of plausible values**

What is a proportion, really?

A proportion is a ratio of successes to total. Let's estimate the population proportion ($p = 0.4$) from $n = 1000$ observations

```
pop_size <- 300000
p <- 0.40
opinions <- c(rep("approve",p*pop_size), rep("not",(1-p)*pop_size))
sampled <- sample(opinions,size=1000) # WITHOUT REPLACEMENT
sum(sampled == "approve")/1000  # proportion estimator
```

```
## [1] 0.371
```

Notice that the proportion is the number of approvals divided by total asked, where the *probability of each person approving is p, independently over n trials*

This should ring some bells!

The estimator for a proportion is just a binomial random variable, X, divided by n, that is, $\hat{p} = X/n$!!

Recall, if X is binomial with parameters n, p then

$E(X) = np$

$V(X) = np(1-p)$

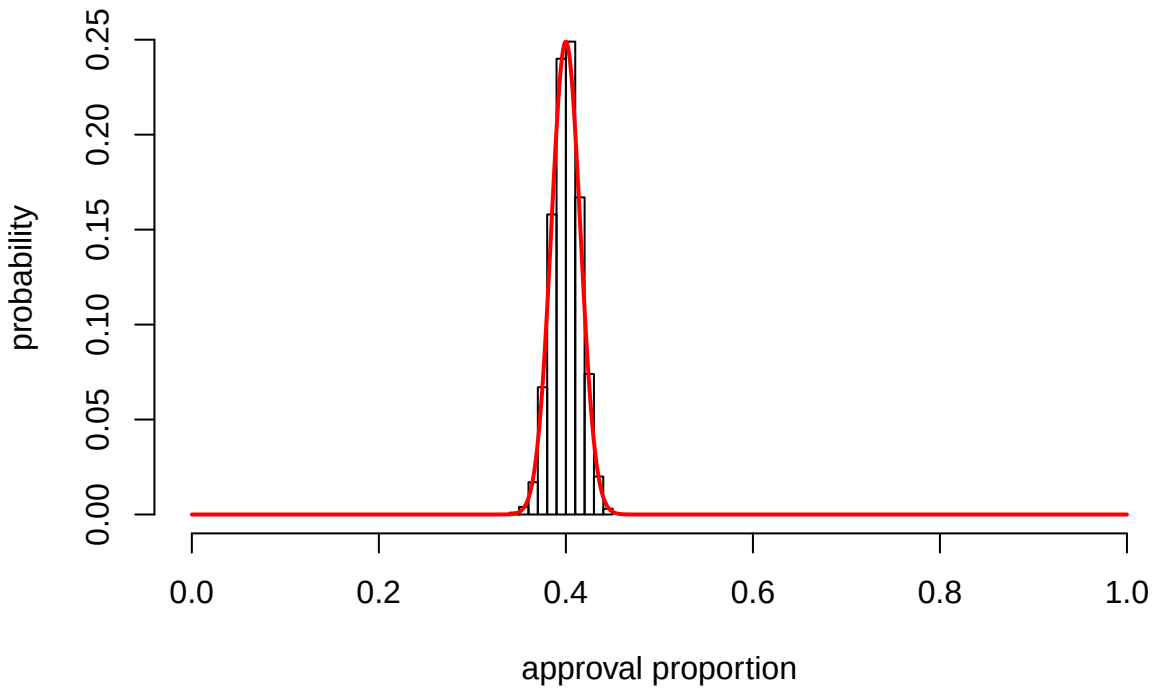
therefore

$E(\hat{p}) = \frac{1}{n}E(X) = p$

$SE = \sqrt{V(\hat{p})} = \sqrt{V(\frac{X}{n})} = \frac{1}{n}\sqrt{V(X)} = \sqrt{\frac{p(1-p)}{n}}$

remember when n is "large" the binomial turns into a normal

Below is a histogram of the sample proportions along with a binomial $\frac{X}{n}$ where $X \sim \text{binom}(1000, 0.4)$



Confidence Interval for a proportion

When

1. Observations are independent and
2. Sample size is sufficiently large, that is $np \gtrapprox 10$ and $n(1-p) \gtrapprox 10$*

then the sample proportion, $\hat{p}$ is **normally** distributed with mean p and $SE = \sqrt{\frac{p(1-p)}{n}}$ so that we can construct the

confidence interval as:

point estimate $\pm\, z^* \times SE = \hat{p} \pm z^* \sqrt{\frac{p(1-p)}{n}}$

*this is called the *success-failure* condition

Derivation

Well, we know that the distribution of $\hat{p}$ is

$\hat{p} \sim N(p, SE)$

therefore

$\frac{\hat{p}-p}{SE} \sim N(0,1)$

Now, you can flip this equation around

$p \sim \hat{p} \pm z^* \times SE, \qquad SE = \sqrt{\frac{p(1-p)}{n}}$

z^* is a value corresponding to the chosen confidence level

Example: 95% confidence level

What is the interval, centered around $z = 0$ corresponding to a 95% probability that z takes a value on that interval?

To answer, notice we want **2.5% of the excluded probability to be in each tail**

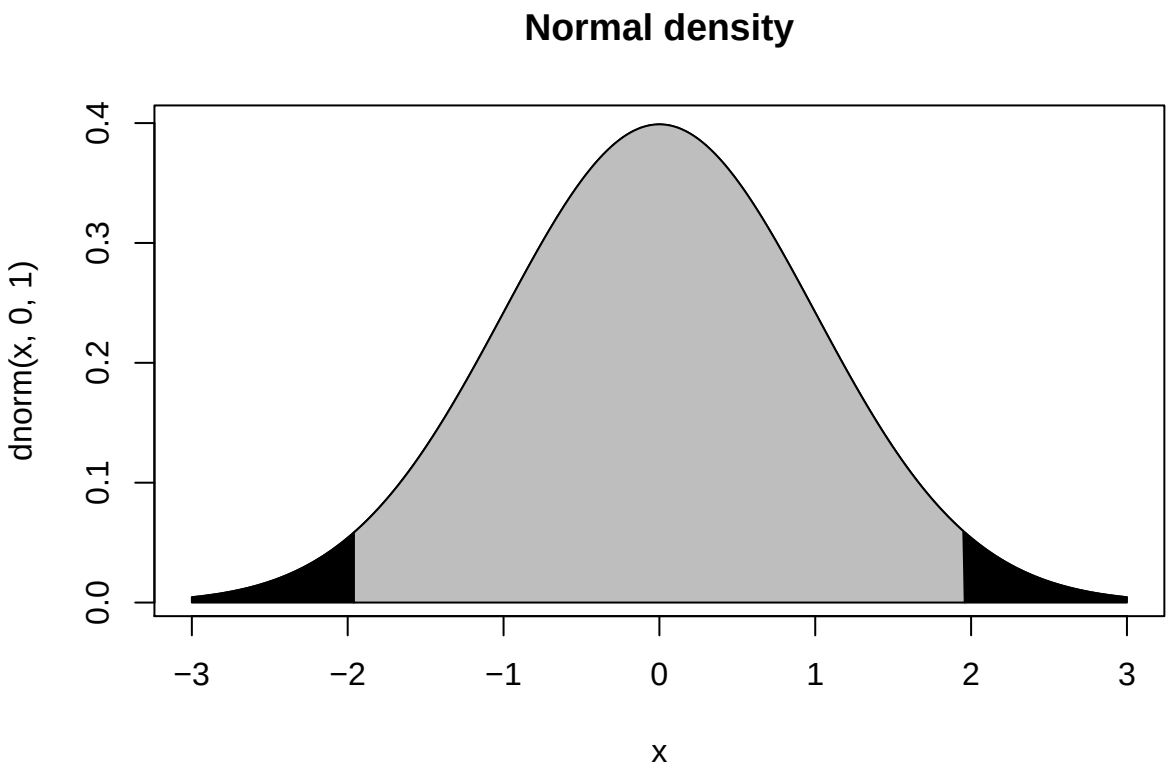
That means we need to find the 0.975 and the 0.025 quantiles of the standard normal which can be obtained by `qnorm(0.975)` = 1.96 = $+z^*$ and `qnorm(0.025)` = -1.96 = $-z^*$

The 95% confidence interval for z is the interval (-1.96, 1.96)

Therefore, the 95% confidence interval for p is the interval $(\hat{p} - 1.96 \times SE, \hat{p} + 1.96 \times SE)$

Three areas perspective

Thus z^* is the cutoff (quantiles) that divides $N(0,1)$ into a chunk in the middle with area (probability) =0.95 in the middle chunk and area (probability) = 0.025 in each of the tail chunks

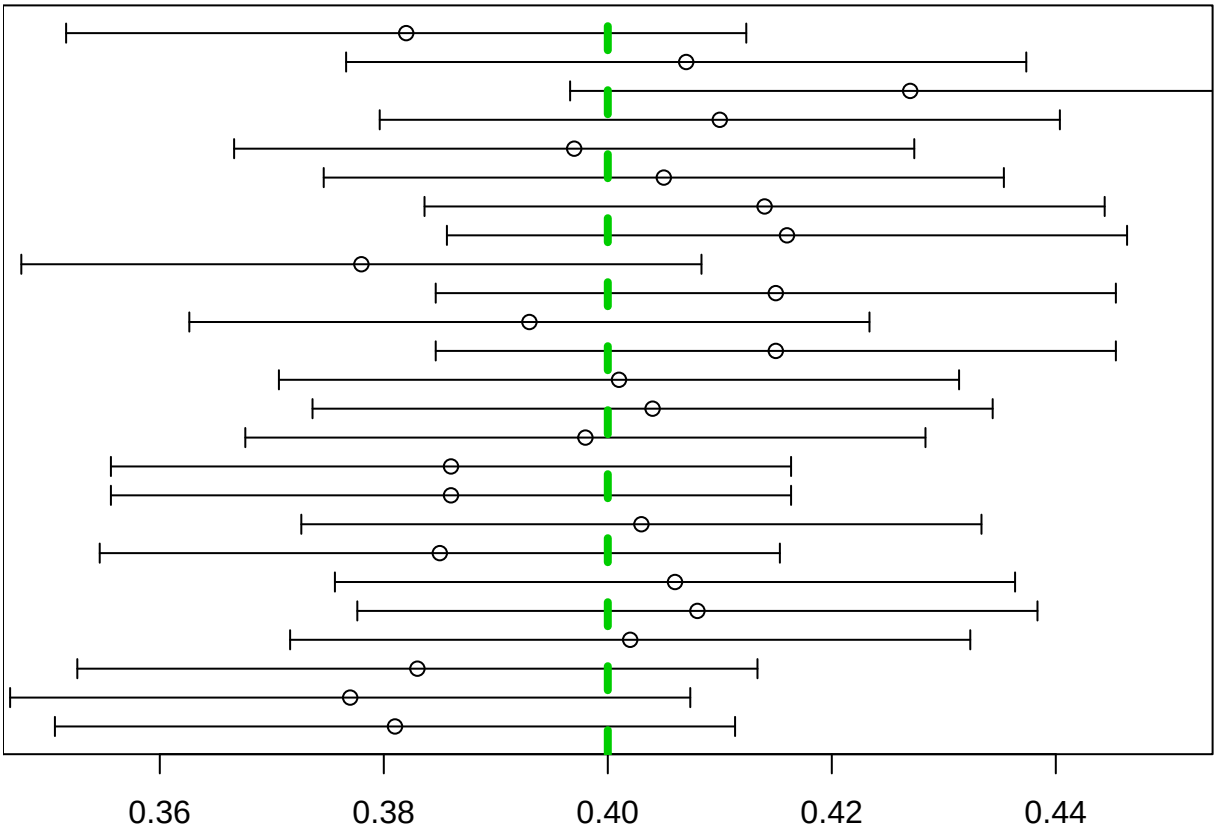


Constructing confidence intervals

m=25, 95% CIs from Trump approval samples with $n = 1000$

```
vb <- vector()
for (m in 1:25) {
pop_size <- 300000
opinions <- c(rep("approve",0.4*pop_size), rep("not",0.6*pop_size))
sampled <- sample(opinions,size=1000) # WITHOUT REPLACEMENT
vb[m] <- sum(sampled == "approve")/1000
}  # proportion estimator
sdev = rep(1.96*sqrt(0.4*0.6/1000),25)

y<-1:25
par(mar=c(1,0,1,1))
plot(vb,y,ylab = "", yaxt='n',xlim = c(0.35,0.45),xlab=" ")
arrows(vb-sdev, y,vb+sdev, y, length=0.05, angle=90, code=3)
abline(v=0.4,col=3,lty=2,lwd=4)
```



Common misinterpretation of CIs

The true population parameter is a fixed unknown value that is either inside or outside the CI with 100% certainty.

For any particular confidence interval, it is **NOT** correct to say that there is a 95% *probability* that the population proportion is in the interval.

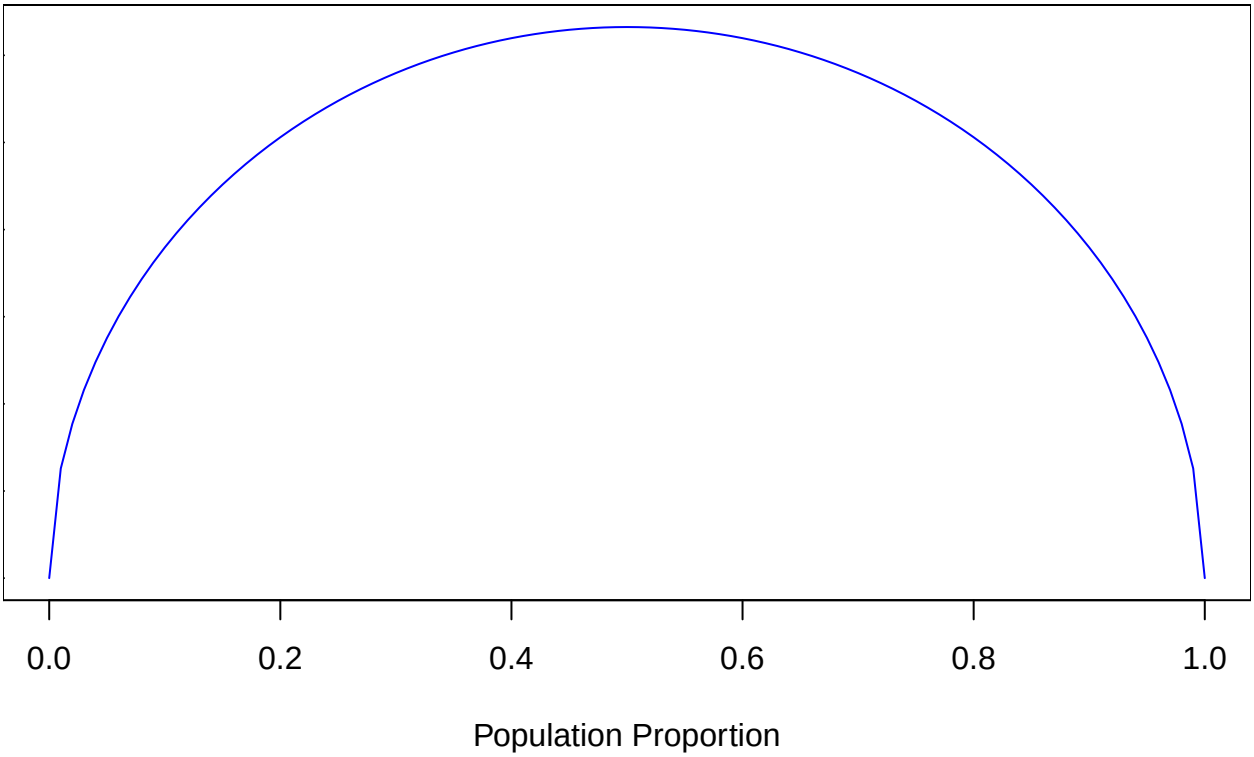
In reality, for a particular confidence interval, centered around an estimate $\hat{p}$, it will either contain p or not contain p, it is not correct to speak of the probability of p being inside the confidence interval.

We expect population p to lie in 95% of confidence intervals

How margin of error depends on p

$z^* \times SE = z^* \times \sqrt{\frac{p(1-p)}{n}}$ is called the *margin of error* (ME)

Notice that when p is around 0.5 you get the biggest ME



How to choose sample size

A university newspaper is conducting a survey to determine what fraction (**proportion**) of students support a $200 per year increase in fees to pay for a new football stadium. How big of a sample is required to ensure the margin of error is smaller than 0.04 using a 95% confidence level?

$ME = 1.96 \times \sqrt{\frac{p(1-p)}{n}} = 0.04$

$n = \frac{1.96^2 p(1-p)}{0.04^2}$

We don't know p, BUT, remember, ME is ***worst-case*** when p = 0.5, so we can set an upper bound as

$n > \frac{1.96^2 0.5(1-0.5)}{0.04^2} = 600.25$ so n should be more than 600

Usually you don't know p

similar to how you replace the σ with an estimate $\hat{\sigma} = S$ to get the t-distribution, you replace p with $\hat{p}$ to estimate the SE when you don't know the population proportion p!

We introduce a subscript to SE to indicate:

$SE_p = \sqrt{\frac{p(1-p)}{n}} \approx SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$ we hope $\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$ is also $\sim N(0,1)$

The procedure is to replace p with $\hat{p}$ also, when checking the **success-failure** condition. Note that there is no equivalent to the t distribution for proportions.

Hypothesis testing framework

H_0 – the *null hypothesis*: often a skeptical perspective of a claim to be tested. Often, it is the statement of "no difference" between two things

H_A – the *alternative hypothesis*: an alternative claim, often represented by a range of parameter values

- We reject the null hypothesis when the data provides **statistically significant** evidence against it
- what means *significant* depends on a chosen level α,
- Even if we fail to reject the null hypothesis, we typically do not accept the null hypothesis as true.
- Failing to find strong evidence for the alternative hypothesis is not equivalent to accepting the null hypothesis.

Example: Trump approval ratings

Suppose that we want to test the claim that the population approval rating is $p = 0.40$, then,

$$H_0 : p = p_0 = 0.40$$
$$H_A : p \neq p_0 \qquad (p \neq 0.40)$$

The thing to do next would be to collect some sample data from the population and get a point estimator $\hat{p}$.

Naturally, we say that the evidence supports the alternative hypothesis if $p_0 = 0.40$ lies outside the $0.95 = 1-\alpha$, CI:

$$(\hat{p} - 1.96 \times SE_{\hat{p}}, \hat{p} + 1.96 \times SE_{\hat{p}})$$

notice that we are using $SE_{\hat{p}}$ here since we *don't know* p

Decision errors

When we choose $0.95 = 1 - \alpha$ or $\alpha = 0.05$ that means that we **acknowledge that we will be wrong 5% of the time!!** So you should hesitate when someone gives you a 95% CI.

Recall that there are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a statement about which one might be true, but we might choose incorrectly. There are four possible scenarios

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	**Type 1 Error**
	H_A true	**Type 2 Error**	okay

Choosing the signifcance level

- If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.
- If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we might choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the alternative hypothesis is actually true.

Example 1: When comparing effectiveness of two drugs, one cheap, one very expensive, choose small α because of high cost type 1 error

Example2 : Two drugs are known to be equally effective for a certain condition. They are also each equally affordable. However, there is some suspicion that Drug 2 causes a serious side-effect in some patients, whereas Drug 1 has been used for decades with no reports of the side effect. The null hypothesis is "the incidence of the side effect in both drugs is the same", and the alternate is "the incidence of the side effect in Drug 2 is greater than that in Drug 1." Falsely rejecting the null hypothesis when it is in fact true (Type I error) would have no great consequences for the consumer, but a Type II error (i.e., failing to reject the null hypothesis when in fact the alternate is true, which would result in deciding that Drug 2 is no more harmful than Drug 1 when it is in fact more harmful) could have serious consequences from a public health standpoint. So setting a large significance level is appropriate.

p-values (alternative to CIs)

The p-value is the probability of observing data at least as favorable to the alternative hypothesis as our current data, if the null hypothesis were true.

For example, suppose our point estimate for Trump's approval was $\hat{p} = 0.33$ from a sample of $n = 1000$ opinions. Assume the null hypothesis $p_0 = 0.40$ is true, then $\hat{p}$ is normally distributed with mean 0.40 and $SE_{p_0} = \sqrt{\frac{0.4(1-0.4)}{1000}} = 0.0155$

next, compute the z-score

$z = \frac{\hat{p}-p_0}{SE_{p_0}} = \frac{0.4-0.33}{0.0155} = -4.52$

the left tail probability is `pnorm(-4.52)` $= 0.00000309$

We usually double this to account for values more than $\hat{p} = 0.47$ which are equally extreme in the right tail. The p-value 0.00000618 represents the probability of the observed $\hat{p} = 0.33$, or a $\hat{p}$ that is more extreme in either tail, if the null hypothesis were true.

To find the p-value, we generally

1. Find the null distribution
2. Then we find a tail area in that distribution corresponding to our point estimate

Notice that we used p_0 not $\hat{p}$ in SE_p since we assume H_0 is true - this is called the *null distribution*. The standard error is computed using $\hat{p}$ for the confidence interval approach. This is the difference between calculating p-values and confidence intervals!!

Evaluating p-values

If the null hypothesis were true, there's only a very small chance of observing such an extreme deviation of $\hat{p}$ from 0.4, so either

1. The null hypothesis is true, we just happened to observe something that only happens about once-in-a-million times
2. The alternative hypothesis is true, which would be consistent with observing a sample proportion far from 0.4

The first scenario is quite improbable, while the second scenario seems much more plausible.

Formally, when we evaluate a hypothesis test, we compare the p-value to the significance level, which in this case is $\alpha = 0.05$. Since the p-value is less than α , we reject the null hypothesis. That is, the data provide strong evidence against H_0.

When the p-value is greater than α, do not reject H_0, and report that we do not have suficient evidence to reject the null hypothesis.

In either case, it is important to describe the conclusion in the context of the data.

One sided hypothesis test

In this case, everything is the same, except that you would change the hypothesis framework to:

$$H_0 : p = p_0 = 0.40$$
$$H_A : p > 0.40 \qquad \text{OR} \qquad H_A : p < 0.40$$

Then you check the tail area in only **ONE TAIL** (don't double it)

You should be careful about this because you may miss out on some important significant stuff going on in the neglected tail.

proportion test in R with `prop.test()`

Suppose $\hat{p} = 0.33$ approved from $n = 1000$ surveyed

```
prop.test(330,1000,p=0.4,correct=F) # two-sided
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  330 out of 1000, null probability 0.4
## X-squared = 20.4167, df = 1, p-value = 0.0000062285
## alternative hypothesis: true p is not equal to 0.4
## 95 percent confidence interval:
##  0.30155554 0.35974556
## sample estimates:
##    p
## 0.33
```

R does a fancier version, which gives slightly different p-value

Chi squared goodness of fit test

Suppose we want to know if we have a **representative sample**

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

The idea is we want to test several proportions all at once!!

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

We can choose our test statistic to be

$$\left(\frac{205-198}{\sqrt{198}}\right)^2+\left(\frac{26-19.25}{\sqrt{19.25}}\right)^2+\left(\frac{25-33}{\sqrt{33}}\right)^2+\left(\frac{19-24.75}{\sqrt{24.75}}\right)^2$$

The expected counts $E_1, E_2, ..., E_k$ are based on the null hypothesis which is that the data is a representative random sample from the population. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a χ^2 distribution with $k-1$ degrees of freedom:

$$X^2=\frac{(O_1-E_1)^2}{E_1}+\frac{(O_2-E_2)^2}{E_2}+...+\frac{(O_k-E_k)^2}{E_k}$$

The p-value for this test statistic is found by looking at the upper tail of this chi-square distribution. We consider the upper tail because larger values of X^2 would provide greater evidence against the null hypothesis.

Chisq test in R with `chisq.test(x,p)`

You can run a chisq test in R with `chisq(x,p)` where x, is your data, p are the expected proportions for each bin

```
observed <- c(205,26,25,19)
expected_probabilities <- c(198,19.25,33,24.75)/275
chisq.test(observed,p=expected_probabilities)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 5.88961, df = 3, p-value = 0.11711
```

The p-value 0.117 is the right tail area under the chi squared distribution with 3 degrees of freedom. If $p < \alpha$ we should reject the null hypothesis, H_0 (explain this statement)

Uses of chisquared test

1. You can use the chi squared test to see if your data fit a theoretical probability distribution (e.g.) exponentail. This will require some care if the theoretical probability distribution is continuous.
2. You can use the chi squared test as a test for independence in two way tables (See OpenIntro Ch. 6.4)

Finding a t-CI for the pop mean

Based on a sample of n independent and nearly normal observations, a confidence interval for the **population** mean is

point estimate $\pm t^*_{df} \times SE = \overline{X} \pm t^*_{df} \frac{S}{\sqrt{n}}$

where

$\overline{x}$ is the sample mean

t^*_{df} determined by confidence level and degrees of freedom

$SE = \frac{S}{\sqrt{n}}$ is the standard error as estimated by the sample

---

the data below has sample mean 5.25 we will find a 95% t-CI

```
x <- c(1,5,9,3,2,5,2,6,2,9,15,4)
sample_mean <- mean(x)
n <- length(x)
sample_SE <- sd(x)/sqrt(n)
round(sample_SE*qt(0.975,df=n-1),2)
```

```
## [1] 2.57
```

Therefore the 95% t-confidence interval for the sample mean is

$(5.25 - 2.57, 5.25 + 2.57) = (2.68, 7.82)$

If your null hypothesis was that $\mu_0 = 9$ then you would reject the hypothesis

R's `t.test`

An easier way to do the same, thing which gives you both the CI and the p-value is

```
x <- c(1,5,9,3,2,5,2,6,2,9,15,4)
t.test(x,mu=9)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -3.20908, df = 11, p-value = 0.0083191
## alternative hypothesis: true mean is not equal to 9
## 95 percent confidence interval:
##  2.6780188 7.8219812
## sample estimates:
## mean of x
##      5.25
```

Lab-Unit 6

Exercise 1

Suppose you want to know if a coin is a fair, that is if the probability of heads = probability of tails = 0.5

Suppose you flip the coin 30 times and it lands heads up 22 times

1a) Phrase this problem in terms of the hypothesis testing framework. State the null hypothesis and the alternative hypothesis

1b) Determine from the mathematical formulas, (you can check the jupyter notebook or search online), what is the probability of getting 22 or more heads from $n = 30$ trials, if the coin is fair (e.g. if the null hypothesis is true)

1c) state your result from part b) as a two tailed p-value and conclude whether to reject or not reject the null hypothesis

1d) Write a code in python or in R that flips a fair coin 30 times, and then replicates that experiment 10000 times. Each time you get 22 or more heads update a counter, then record the **simulated p-value** as this 2*counter value/10000

1e) which way of determining the p-value do you prefer, simulation approach or looking up the formula approach?

Exercise 2

You are given the following hypotheses:

$$H_0 : \mu = 60$$

$$H_0 : \mu \neq 60$$

Suppose that the sample standard deviation is S =8 and the sample size is $n = 20$. For what sample mean would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

Exercise 3

For a given confidence level,t^*_{df} is larger than z^*. Explain how t^*_{df} being slightly larger than z^* affects the width of the confidence interval.

Exercise 4

A 95% confidence interval for a population mean, μ, is given as (18.985, 21.015). This confidence interval is based on a simple random sample of 36 observations. Calculate the sample mean and standard deviation. Assume that all conditions necessary for inference are satisfied. Use the t-distribution in any calculations.