Week 7: Confidence Intervals
Part 1

Jessie Yeung

STA 220

Fall 2023

Overview

- Over the next 2 weeks, we will be learning about confidence intervals
- Topics for this week
  - Inference and Sampling Distributions
  - Confidence Intervals for Proportions
  - Sample size for proportions
- This content corresponds with Module 6

Second half of the course

- We will be working with data!
- In most of the applications in the rest of the course, we will NOT know the true population parameters
- We will be using the data to estimate the true populations parameters
- But how do we know if our estimates are good or not?
  - We have to make sure the data is gathered well (Week 5)
- Let's assume the data is gathered perfectly (simple random sampling with no issues, no missing data, can access whole population…)
  - Is our estimate useful or not? Can use our estimate to say anything about the population parameter?

Review: Sampling Distributions

Sampling Distributions

- Let's take a quick refresher on Week 4 and 5 content regarding sampling distributions.
- When we study things, we are often dealing with a population that we want to know something about:
    - e.g. suppose we want to know the proportion of Americans that drink more than one soda per day.
    - this proportion is the population parameter p
    - we can't see this because we can't measure all Americans
    - but we know that the number of Americans that drink more than one soda per day is Binomial, with some probability of success p
    - we want to try to estimate this probability of success with some data

Sampling Distributions

- We have our population of Americans, and the number of them that drink more than 1 soda per day can be represented as a Binomial distribution.
  - we want to figure out the overall proportion of Americans who drink >1 soda/ day
- So we take a sample of say 50 of them. From the sample, we can find the number of them that drink >1 soda/day
  - suppose we find that 17 of them do… that's a sample proportion of $\hat{p} = 0.34 = \frac{17}{50}$
  - This sample proportion is a sample statistic
  - we now want to see if this is a *good guess* (or estimate) for this unknown population proportion of all Americans that drink >1 soda/day.

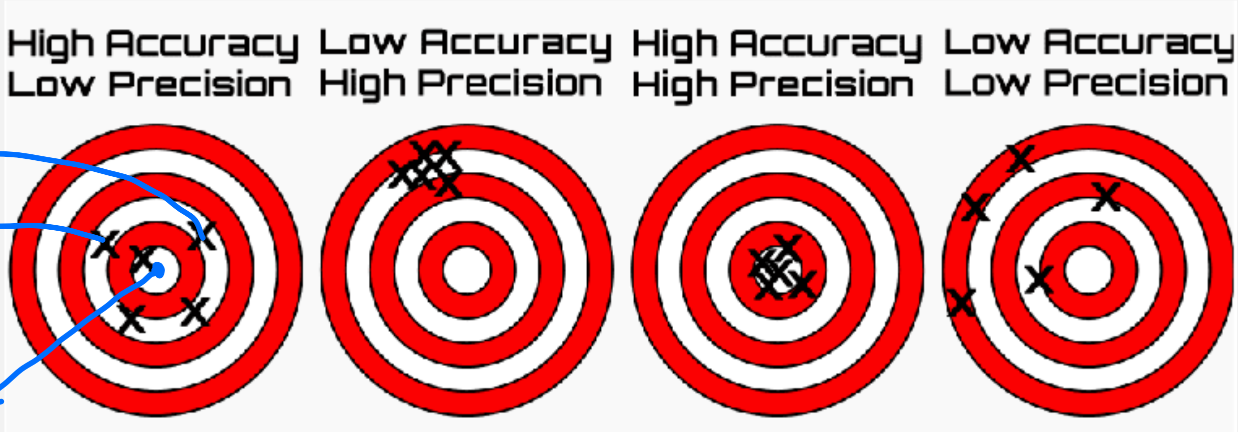
What do we mean by a good estimate?

- High accuracy is desirable:
  - High accuracy means that it is measuring what we want it to measure and there is no/low bias
- High precision is desirable:
  - High precision means that the method of estimating has low variability

$\hat{p}$ from another sample

$\hat{p}$ from a sample

p



Sampling Distributions

- For our particular sample of 50 people, we have $\hat{p} = 0.34$ which is now our estimate for the unknown population parameter p
- We know that the value of $\hat{p}$ will vary from sample to sample.
- But if $\hat{p}$ will vary from sample to sample, how do I know whether the value of $\hat{p}$ from my sample will be a good/bad estimate for p?
- The sampling distribution of $\hat{p}$ can tell us about the value of $\hat{p}$ on average and how much it's value varies
- Recall that when the CLT applies,

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Sampling Distributions

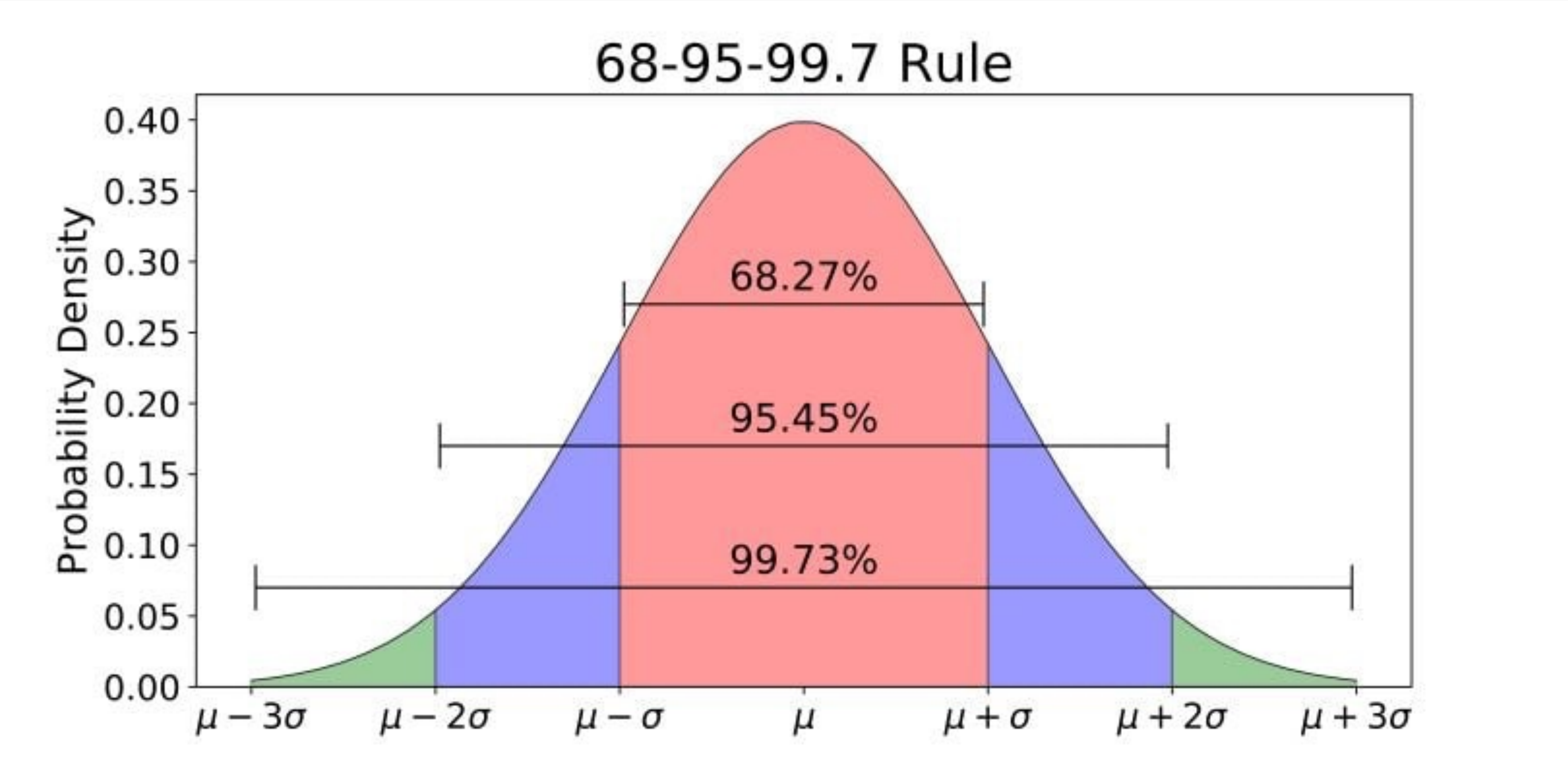
$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- The mean of the sampling distribution tells us that, even though $\hat{p}$ changes from sample to sample, its value on average is p
  - the mean here matches with the mean in the population, which we are trying to estimate

- The variance of the sampling distribution tells us how much $\hat{p}$ changes from sample to sample
  - This gives me some information on how far $\hat{p}$ might be from the unknown p

Aside: 68-95-99 Rule

- The standard deviation, in general, acts as a yard stick
  - it tells us how far things are from the mean/centre
- There is empirical rule to get a rough idea how far we are from the mean when values are normally distributed
  - about 68% of observations are within 1 SD of the mean
  - about 95% are within 2 SD of the mean
  - about 99% are within 3 SD of the mean
- Even if the data is not normal, this can be used as an approximation to tell you how far away you are from the mean
  - Eg. "1 SD away from the mean is not that far, but 4 SD is very far"
  - Use your own judgement on when this approximation is appropriate


68-95-99.7 Rule
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
68.27%
95.45%
99.73%
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Standard Deviation with Normals

- The sampling distribution of $\hat{p}$ tells us that values of $\hat{p}$ are normally distributed
- When we are working with a Normal distribution, we can get a precise idea of exactly how many standard deviations a value is from the mean
  - This is because we know how to use the z-score to get a standard Normal distribution
  - and, we know the quantiles of the standard Normal from the Normal table
- So we just need to turn the sampling distribution of a proportion into a standard Normal to find out how far away my sample value is from what we are trying to estimate.

Z-score of a Sample Proportion

- We can now take the z-score of our sample proportion, by using its sampling distribution
  - the mean of the sampling distribution is p
  - the standard deviation of the sampling distribution is $\sqrt{\frac{p(1-p)}{n}}$
- To find the z-score, we subtract off the mean, and divide by the standard deviation:
- $z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$
- We know that this process will give us a standard Normal

Soda Example

- So in our example, we have seen that we get a sample proportion of 0.34 $\hat{p} = 0.34$
- Further suppose the true population proportion is 0.3 $p = 0.3$
- So, $\hat{p} \sim N\left(0.3, \frac{0.3(1-0.3)}{50}\right)$
- Let's find the z-score for the sample proportion:

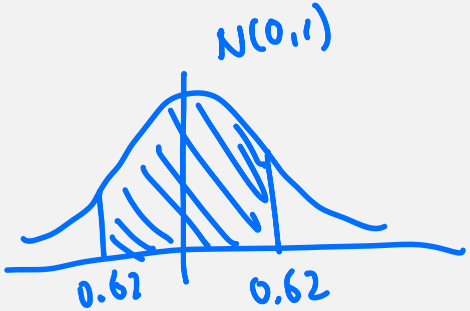
$$z\text{-score} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.34 - 0.30}{\sqrt{\frac{0.3 \times 0.7}{50}}} = 0.62$$

- This seems to suggest that we are reasonably close to the population parameter we want to estimate.
  - but this is still a bit vague… how close is close enough?

Normal Probabilities

- The closer my sample proportion $\hat{p}$ is to the population proportion p, the closer the z-score is to 0.
- My current sample proportion is 0.04 away from the population proportion. What is the probability that a random sample of size 50 will yield a sample proportion that is at most 0.04 away from p?

$$P(0.26 < \hat{p} < 0.34) = P\left(\frac{0.26-0.3}{\sqrt{\frac{0.3\times 0.7}{50}}} < Z < \frac{0.34-0.3}{\sqrt{\frac{0.3\times 0.7}{50}}}\right), \quad Z \sim N(0,1)$$



$$= P(-0.62 < Z < 0.62)$$

$$= 0.4648$$

But…Population parameters are unknown

- I can find all this information because we were given the value of $p = 0.3$
- This gave us $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \equiv N\left(0.3, \frac{0.3(1-0.3)}{50}\right)$ which tells us everything we need to know about the behaviour of $\hat{p}$
- Recall that in real life we don't know the value of population parameters
  - Or else why even collect data and do statistics???

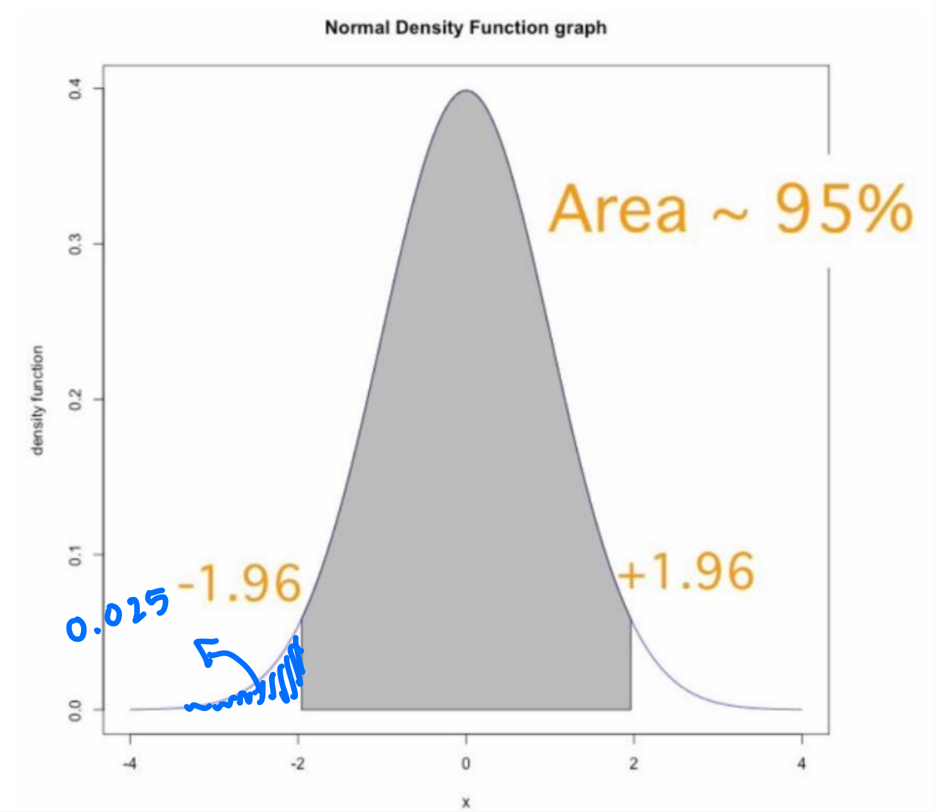
Idea of Confidence Intervals

- Even though I won't be able to fill in the value of p into the sampling distribution, all is not lost
- We can still use the sampling distribution in order to create a **confidence interval**
- Confidence intervals are an interval that describes:
  - It describes a range of plausible values for the population parameter
  - It describes the confidence we have in our estimate
- Statements like "*We are 95% confident that between 63% and 69% of all adult Canadians support the government's response to the refugee crisis*" is example of a confidence interval

Confidence Intervals for Proportions

Standard Normal Distribution

- The standard Normal distribution is very useful because we can graph it and we can determine the probabilities in different regions.
- For example, the area under the curve between –1.96 and 1.96 is equal to 95% of the total area.
- If a random variable follows a standard Normal distribution only 5% of the time will an observation be less than –1.96 or greater than 1.96.



Standard Normal Distribution

- In other words,

$$P(-1.96 < Z < 1.96) = 0.95$$

- We also know that

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

- Let's substitute in the standardized sampling distribution for $\hat{p}$

Deriving the confidence interval

$$P(-1.96 < Z < 1.96) = 0.95$$

sub in $\hat{p}$ after standardizing

$$P\left(-1.96 < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < 1.96\right) = 0.95$$

multiply by denominator

$$P\left(-1.96\sqrt{\frac{p(1-p)}{n}} < \hat{p} - p < 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

subtract $\hat{p}$

$$P\left(-\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < -p < -\hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

$2 < 3$

$-2 > -3$

multiply by -1

$$P\left(\hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}} > p > \hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

Rearrange

$$P\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

Therefore, the 95% confidence interval for p is

$$\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}, \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right)$$

Standard Error

- The standard error refers to the standard deviation of an estimator (sampling statistic)
- It represents how precise an estimate is
- In the case of the sample proportion, $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$
- So, another way to state the confidence interval is:

$$\hat{p} \pm 1.96 \times SE(\hat{p})$$

- Or

$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

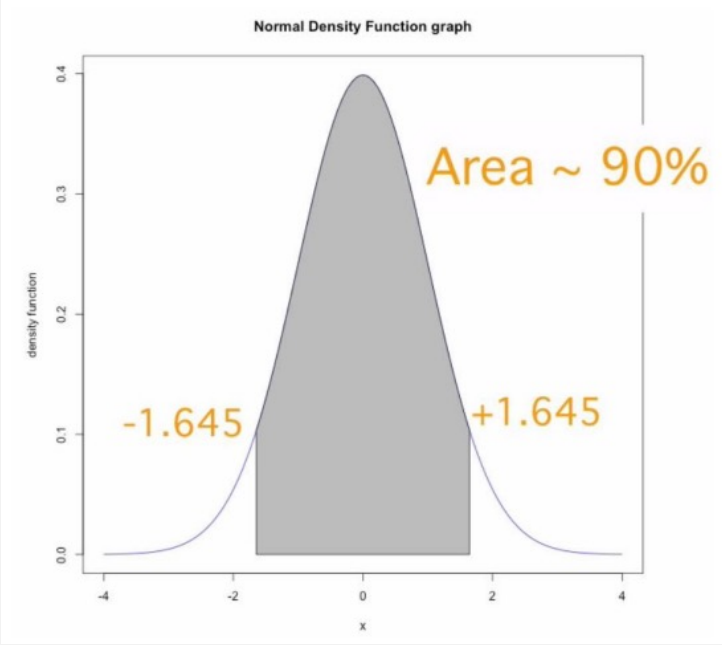
Other levels of confidence

- Not all confidence intervals need to be at the 95% confidence interval. They can be at any level of confidence.
- Suppose you were interested in a 90% confidence interval? How would the formula change?

90% CI for p:

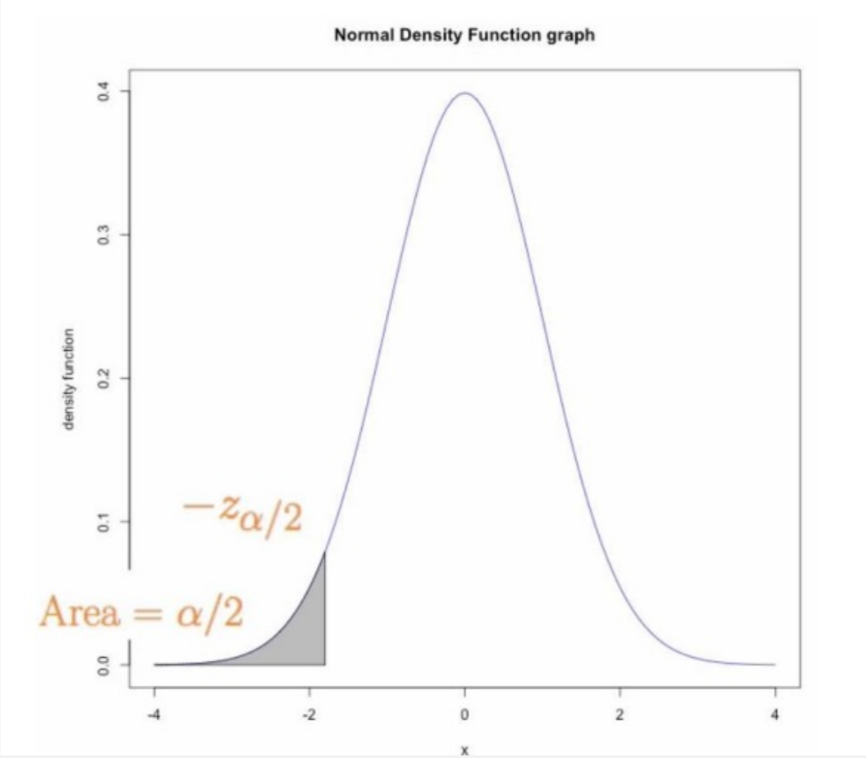
$$\hat{p} \pm 1.645 \times SE(\hat{p})$$

$$\left(\hat{p} - 1.645\sqrt{\frac{p(1-p)}{n}}\ ,\ \hat{p} + 1.645\sqrt{\frac{p(1-p)}{n}}\right)$$



Any level of confidence

- In general, you can find a $(1-\alpha)\%$ confidence interval for any value α
  - A 95% confidence level corresponds to $\alpha = 0.05$
- In general, for any value α, you can find a value $z_{\alpha/2}$ such that the area under the standard Normal curve which is less than $-z_{\alpha/2}$ is equal to $\alpha/2$.
  - Eg. For $\alpha = 0.05$, $z_{\alpha/2} = 1.96$ since $P(Z < -1.96) = \frac{0.05}{2} = 0.025$
- The value $z_{\alpha/2}$ is called the **critical value**



Don’t know parameter…

- But we don’t actually know what the population proportion is
- So we can’t use the interval as it is, because the standard deviation of the sampling distribution uses p.
- We have two options for dealing with this:
  1. Use $\hat{p}$ in place of p
     - if our sample is good, then it’s reasonable to think that $\hat{p}$ will be close to p
  2. Use p = 0.5
     - this is a conservative choice - it will give you the largest possible interval
     - this happens because you are basically saying you have no information to say that the true proportion is anything other than the result of a random coin flip.

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

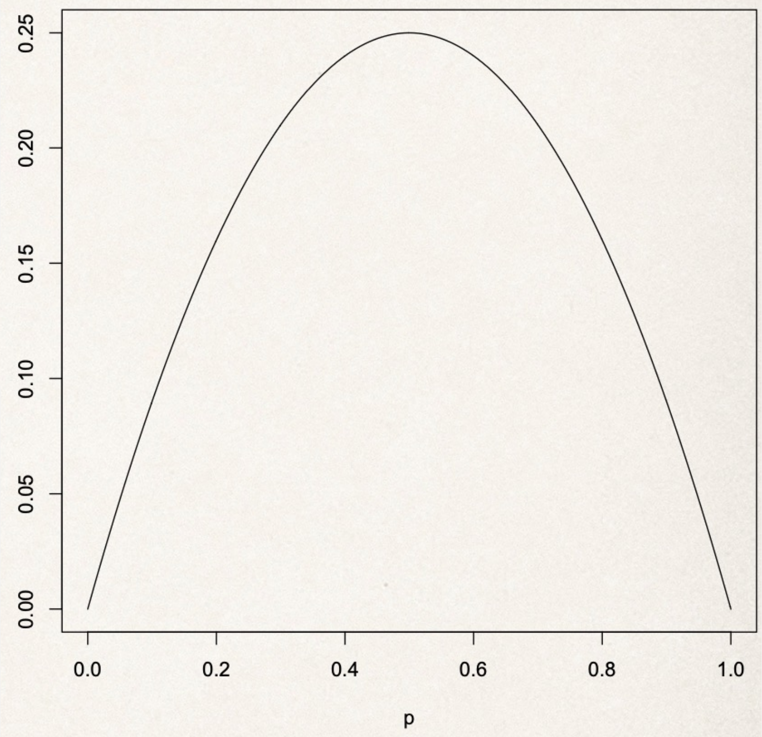
How is $p = 0.5$ the most conservative choice?

- To understand this, we need to understand what the standard deviation of the sampling distribution means in a confidence interval
- We can write our interval as

$$\hat{p} \pm z_{\alpha/2} \times SE(\hat{p})$$

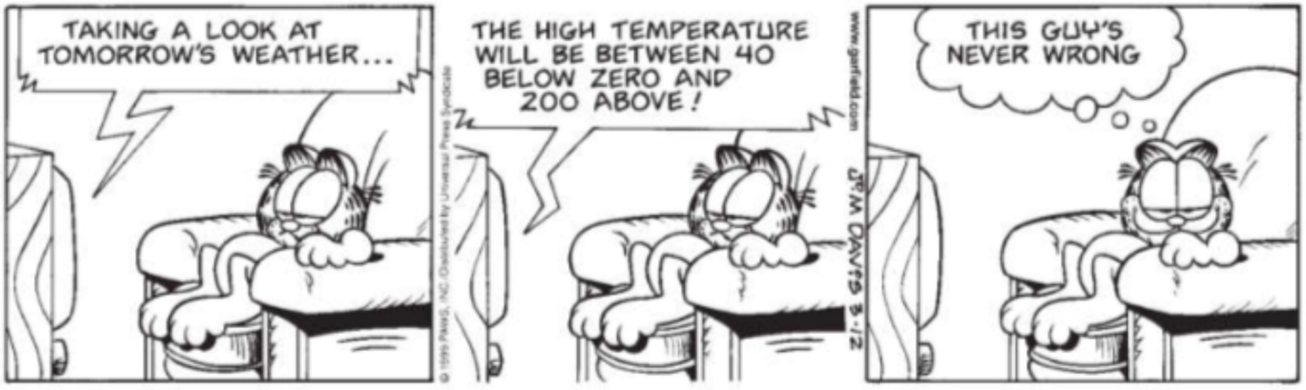
- Because $SE(\hat{p})$ is a function of p, as p changes, the error changes
- We see that when p = 0.5, the standard error is the largest it can be.

$p(1-p)$



Balance between error and confidence

- So what happens when the confidence is higher?
  - It turns out that this makes my confidence interval bigger too.
- This means we have to balance between the error of the estimate (i.e. precision) and the confidence that we could be close to the population value.



Example: Soda

We found from a sample of 50 Americans that 34% of them have more than 1 soda per day. Build a 95% confidence interval, using $p \approx \hat{p}$

$$\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}} \Rightarrow 0.34 \pm 1.96\sqrt{\frac{0.34(1-0.34)}{50}}$$

$$\Rightarrow (0.209,\ 0.471)$$

Example: Soda

Build a 95% confidence interval, using $p = 0.5$.

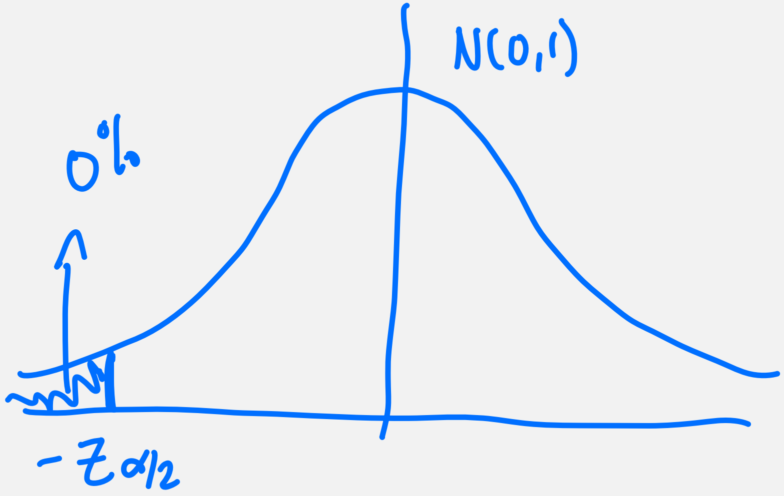
$$\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}} \Rightarrow 0.34 \pm 1.96\sqrt{\frac{0.5 \times 0.5}{50}}$$

$$\Rightarrow (0.201,\ 0.479)$$

Example: Soda

Build a 100% confidence interval, using $p = 0.5$. What happens?

Let's try to find critical value $z_{\alpha/2}$



We need a value such that

$P(Z < -z_{\alpha/2}) = 0$

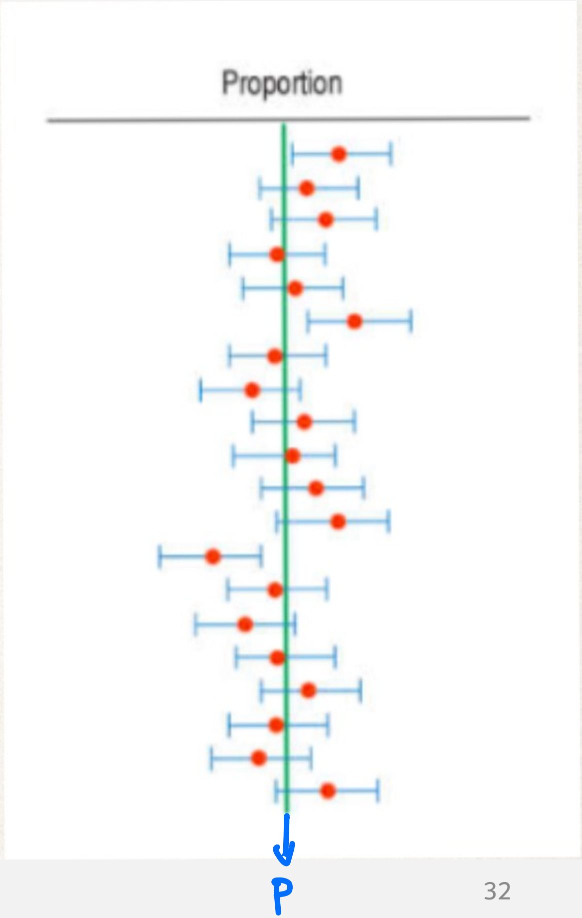
The critical value is infinity

Therefore, the CI is infinitely wide.

You can use any confidence level you want, but must pick one that is reasonable.

Meaning of Confidence Intervals

- What do these intervals mean?
- They are a statement about our sample.
- When we are talking about a 95% confidence interval of (a, b), we are saying that, for all samples of the same size as mine, 95% of confidence intervals on those samples will overlap with the population parameter I am trying to estimate.
- It is a statement about the variability in the sample that I took, and it reflects the idea of the sampling distribution.



How to interpret confidence intervals?

- “We are 95% confident that population proportion of ______ is between a and b”
- Avoid talking about probabilities that p is in the interval
  - Is it correct to say that there is a 95% probability that p is between a and b?
  - No, p is a fixed value that is unknown to us. Either p is in the interval or not.
  - Due to this, $P(a < p < b)$ is either 0 or 1
  - Instead, the correct way to think of this is that 95% of the confidence intervals created in this way across all possible samples will contain the true p

$p(0 < 1) = 1$

Sample Sizes

Sample Size

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

- If we look at the expression for the confidence interval, we basically are working with 3 pieces:
  - the sample proportion $\hat{p}$
  - the critical value $z_{\alpha/2}$ based on confidence level $(1-\alpha)\%$
  - the sample size n
- We have already discussed how the value of $\hat{p}$ and α can change the width of the confidence interval. But what about the sample size?
- Well we know that this changes the standard deviation of the sampling distribution... as n goes up, SD goes down!
  - if SD goes down, then width of confidence interval gets smaller
- So we can use the idea of confidence intervals to figure out how large my sample should be.

Margin of Error

- The **margin of error** represents the distance from $\hat{p}$ to one end of the interval.

- Recall we have $\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$

- so the length of the interval is determined by the margin of error

- but the sample size plays a roll in deciding how big the margin of error will be.
  - as does the confidence level, and the choice for p in the standard error.

- If we know ahead of time that we want an interval with a certain width and at a specific confidence level, I can figure out how big of a sample I need.

Margin of Error

- I can write the margin of error as $ME = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
- If I decide how wide my interval should be (ME) and the confidence level (α) and thus $z_{\alpha/2}$, I can find a value for n, by rearranging the above expression.
- This would give me the expression $n = \left(\frac{z_{\alpha/2}\sqrt{p(1-p)}}{ME}\right)^2$
- So if I want to create a 95% confidence interval, where the total width of the interval is 0.2 (i.e. 0.1 on either side of $\hat{p}$), then I have $z_{\alpha/2} = z_{0.05/2}$= 1.96 and ME = 0.1.
- What to do with p?

Conservative choice of p

- This is where the idea of the conservative choice of p comes in.
- The reason why we want to use $p = 0.5$ to calculate a sample size is because we know this will give us the biggest possible standard error of our sampling distribution
  - and thus the widest interval.
- Using $p = 0.5$ means we should always get a sample size that will ensure we have more data than we should need to make a confidence interval with a particular width.
- And it is never a bad thing to have too much data!

Sample size formula

- So we should always use p = 0.5 to calculate a sample size because it will give us more data than we need.
- We can now write the final formula to use to find the sample size for proportion:
  - $n = \left(\frac{z_{\alpha/2} \times 0.5}{ME}\right)^2$
  - Always round up to the nearest integer
  - you are collecting units/objects so you can only collect whole units (not fractions)
  - you also want to have enough, so that's why we round up and never down.

Example: Health Care

In a 2009 Canada Day poll of 1000 Canadians, 58% said they were proud of Canadian health care.

A. Find the margin of error for the poll if we want 90% confidence in our estimate. Use a conservative choice for p.

B. What does this margin of error mean?

A. $\hat{p} = 0.58 \quad n = 1000$

90% confidence $\Rightarrow \alpha = 0.1 \Rightarrow z_{\alpha/2} = 1.645$

$$ME = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 1.645\sqrt{\frac{0.5 \times 0.5}{1000}} = 0.026$$

B. In a 90% CI for p, the distance from the centre ($\hat{p}$) to the end of the interval is 0.026

The total width of a CI would be 2×0.026

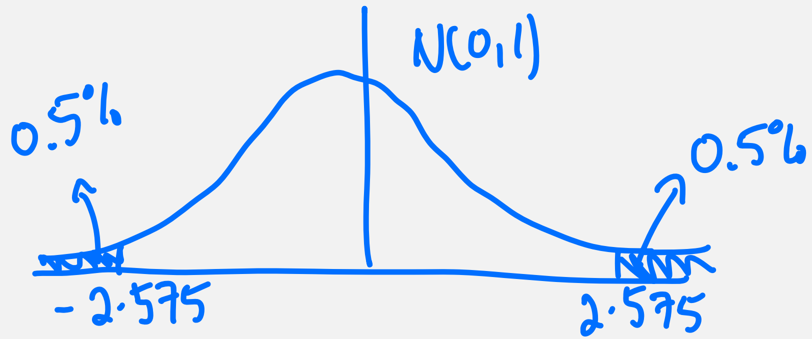
Example: Health Care

C. If we want to be 99% confident, would the margin of error be larger or smaller?

D. If we want to be 99% confident but have a margin of error of 0.02, how many people should we survey?

C. If we want to be more confident, we need a wider confidence interval.

D. 99% confidence $\Rightarrow$ $\alpha = 0.01$ $\Rightarrow$ $z_{\alpha/2} = 2.575$



$$n = \left(\frac{z_{\alpha/2} \times 0.5}{ME} \right)^2 = \left(\frac{2.575 \times 0.5}{0.02} \right)^2$$

$$= 4144.14 \xrightarrow[\text{UP}]{\text{ROUND}} 4145$$

Videos and Practice Problems

- In this lecture, we covered all of Module 6: Confidence Intervals Part 1
- Module 6 Practice Problems are posted on Quercus

Next Week

- Weekly quizzes resume this week
- Next week we will see confidence intervals for means