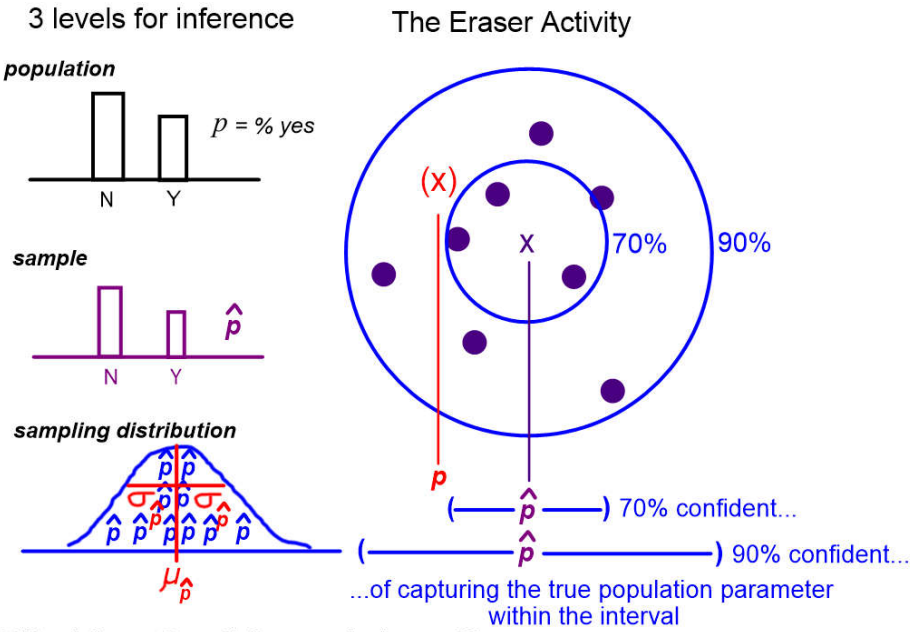
AP Statistics – Lesson Notes - Chapter 19: Confidence Intervals for Proportions



What is a Confidence Interval?

An example from our textbook:
One type of coral, the sea fan, is threatened by a particular disease called aspergillosis. In June 2000, a team examined a sample of 104 sea fans off the coast of Mexico and found that 54 of the sea fans had this disease. *Is it possible to say anything about the prevalence of this disease among sea fans in general from the data about this sample?*

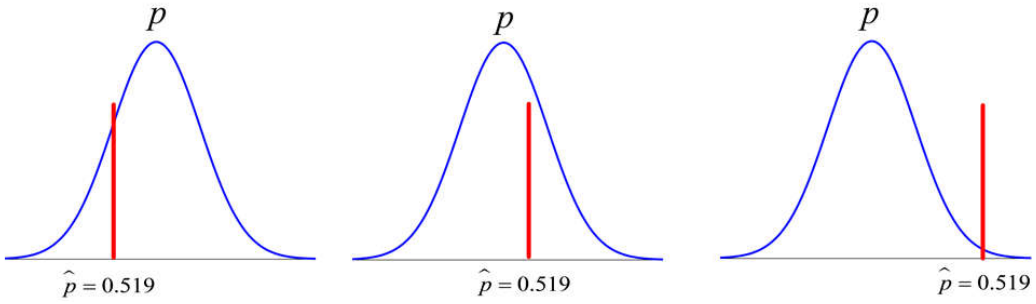
This sample had a sample proportion $\hat{p} = \frac{54}{104} = 0.519$ of sea fans which have the disease.

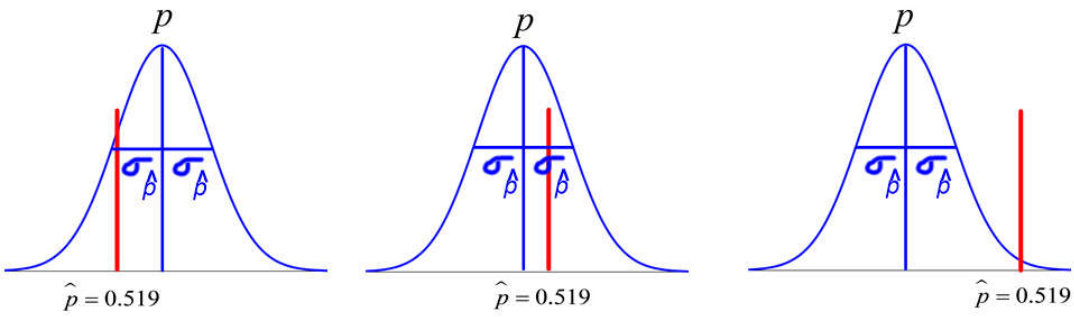
We would like to know the proportion, p , of the entire sea fan population which has the disease.

We know that our particular sample of 104 is just one of many possible samples of 104 sea fans that could have been taken, and if a different sample had been taken, the proportion would likely be different. If we imagine taking many such samples, the proportions of each would form a sampling distribution of sample proportions.

This sample had a sample proportion $\hat{p} = \frac{54}{104} = 0.519$ of sea fans which have the disease.

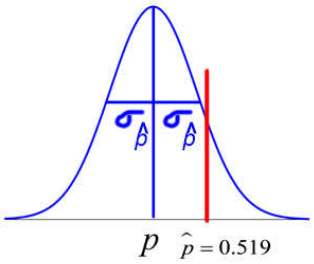
But we don't know the true population proportion, so is our proportion might be close to the true proportion or far away and might be above or below the true proportion:





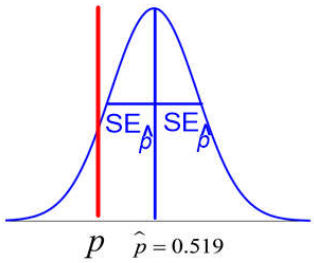
We also know the shape and SD for how the $\hat{p}$ would vary because this would be a sampling distribution of sample proportions.

So what can we do?



We can't center our sampling distribution at the population parameter, *p*, because we don't know it.

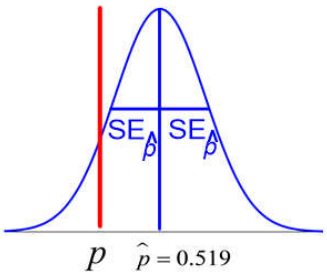
So instead, we...



- **Center things at our sample's statistic.**
- **Use the standard deviation in the sample as an estimate for the population and find the standard error:**

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Confidence Interval



Then we add some space above and below the statistic for our sample to form the *Confidence Interval*.

(—— $\hat{p}$ ——) 70% confident...

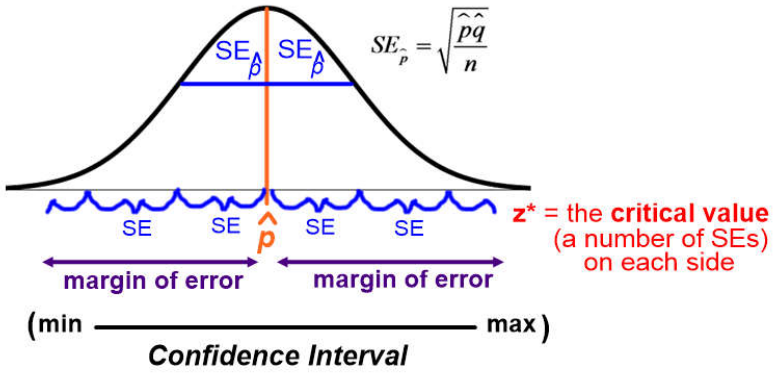
(———— $\hat{p}$ ————) 90% confident...

Confidence Interval

We can adjust the width of the confidence interval depending upon how confident we want to be that this interval will capture the true population portion *(here, the percentage of <u>all</u> sea fans with this disease).*

The amount we go above and below the statistic is called the 'margin of error'.

A number of things affect the margin of error, and these combine to form an equation for computing the minimum and maximum values for the confidence interval.

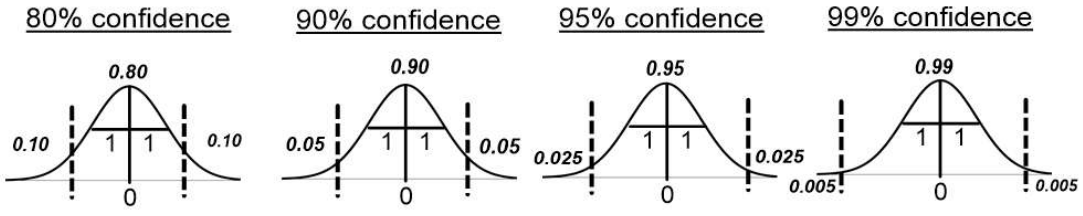


Confidence Interval (CI) = statistic $\pm$ (critical value) ($SE_{statistic}$)

Confidence Interval (CI) = $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$

z* is determined by the desired Confidence Level

The confidence level is a percentage value which roughly means what percent of the time the true population parameter will be within the confidence interval. Since everything is modeled on a sampling distribution with a normal shape, we can use invNorm to select a number of standard deviations (really standard errors) to go above and below the mean to enclose this percentage of the normal distribution:



These z-scores represent how many Standard Errors (SEs) we should include on either side of the mean for our confidence interval, and when used in this way, the z-score is called the ***critical value*** and is denoted by z^*.

For our sea fan problem, let's assume we want a confidence level of 95%, so we'll use z* = 1.96 (the confidence interval will extend 1.96 SEs above and below the statistic).

Standard Error (SE) is determined by n and $\hat{p}$

The Standard Error (SE) is the standard deviation of the sampling distribution for sample proportions, but calculated using our sample value for p instead of the (unknown) population value.

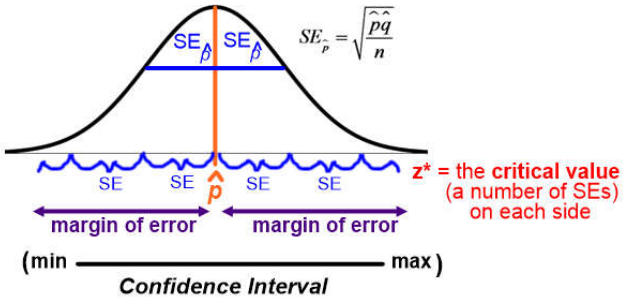
For our sea fan problem...

We know that the sample proportion is $\hat{p} = \frac{54}{104} = 0.519$

Given n=104, we can calculate the Standard Error:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.519)(0.481)}{104}} = 0.049$$

Calculating the confidence interval for the sea fan problem



$$\text{Confidence Interval (CI)} = \text{statistic} \pm (\text{critical value})(SE_{\text{statistic}})$$

$$= \hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$= .519 \pm (1.96)(.049)$$

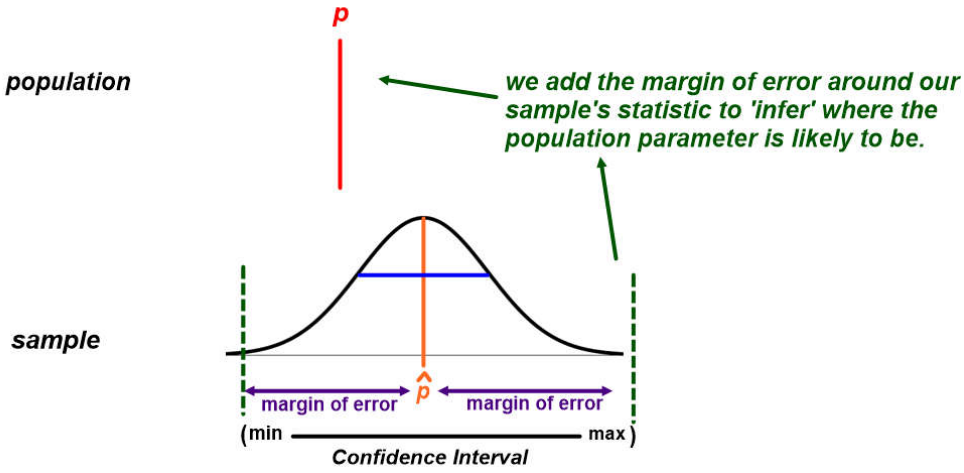
$$= .519 \pm (.096)$$

margin of error

$$= (.423,\ .615) \text{ in interval notation}$$

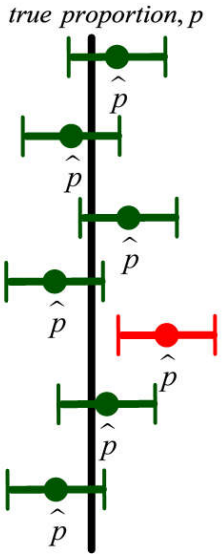
"We are 95% confident that between 42.3% and 61.5% of <u>all</u> sea fans have this disease."

With the information from the sample, we infer something about the population



What does "95% confidence" really mean?

If we had taken different samples of 104 sea fans, we would have different sample proportions and different confidence intervals:



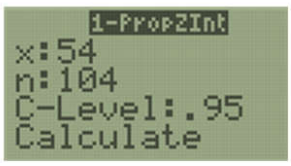
If we were to take many samples of this sample size, and compute confidence intervals for each, 95% of these confidence intervals would contain the true proportion of the population of sea fans with the disease.

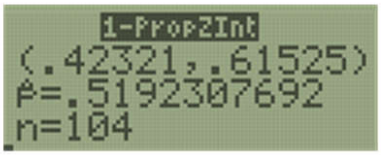
EXPLORE: rossmanchance.com/applets/

Calculator Functions

Calculators provide statistical 'tests' to do this analyses very quickly. In our class (at least for now) we are doing thing by hand, but you can use the calculator tests to verify your answer.

Stat, Tests, A: 1-PropZInt





Default value for $\hat{p}$

Usually, we will be given a value for $\hat{p}$, which will be the proportion we actually had in our sample. But sometimes we don't have that value, or we want to find the *worst case margin of error* which would be valid regardless of what the $\hat{p}$ value was for any particular sample.

In these situations, we must use 0.5 for $\hat{p}$, because this will give us the widest confidence interval that would be valid for any sample's $\hat{p}$.

Assumptions and Conditions

In order to use the Normal distribution, we need to make some assumptions. We can never be certain that an assumption is true, but we can often decide whether an assumption is plausible by checking related conditions.

Independence Assumption: The data values need to be independent from each other.
- **Plausible independence condition:** Is there any reason to believe data values somehow affect each other? (this would mean *not* independent)
- **Randomization condition:** The data samples need to represent a random sampling of the population. Were these data samples taken at random or generated from a properly randomized experiment?
- **10% condition:** Sample w/o replacement...can we still assume that probability isn't changing for later samples?

Sample Size Assumption: Based on the Normal approximation for a Binomial distribution, number of samples must be high enough to assume that the sampling distribution is Normal.
- **Success/Failure condition:** $np \geq 10 \quad and \quad nq \geq 10$

Try this one: *In May 2002, the Gallup Poll asked 537 randomly selected U.S. adults the question, "Generally speaking, do you believe the death penalty is applied fairly or unfairly in this country today?" Of these, 53% answered 'fairly'. What can we conclude about the proportion of all U.S. who would answer 'fairly' to this question?*

1) Identify the population and parameter you wish to estimate about this population, and select a confidence level.

2) Verify that the conditions are met for a Normal distribution.

3) Determine $n,\ \hat{p},\ SE_{\hat{p}}$

4) Determine the critical level z^* for your chosen confidence level.

5) Calculate the margin of error (ME) and the confidence interval.

6) State your result in the context of the problem.

You have two ways to change the confidence interval:

1) Select the confidence level:

confidence level ↘, z* ↘ , ME ↘ , width of CI ↘

2) Select the sample size:

n ↗, σ ↘, ME ↘ , width of CI ↘

Choosing a sample size

If you have a desired confidence level and precision of margin of error, you could choose sample size to match:

$ME = (z^*)\left(SE_{\hat{p}}\right)$ and $SE_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$ so you could solve for n.

Example: *Suppose a candidate is planning a poll and wants to estimate voter support within 3% with a 95% confidence level. How large a sample is required?*

But...be aware that larger sample sizes cost money and effort. Because the standard error declines only with the square root of the sample size, to cut the standard error (and the ME) in half, we must quadruple the sample size.

Explanations

Confidence Interval

We are 95% confident that between 57% and 68% of all U.S. adults would say the death penalty is applied fairly.

Confidence Level

If we took many samples of size 537 and computed confidence intervals for each, 95% of these confidence intervals would contain the true proportion of all U.S. Adults saying the death penalty is applied fairly.

Things to be careful about...

- Be careful to state the conclusion correctly (don't overstate the result).
- Inappropriate balance of margin of error and confidence (avoid huge margin of error to be able to claim 99% confidence, avoid confidence levels lower than 80% to get small margin of error).
- Choose appropriate sample size balancing precision with cost.
- Check conditions/assumptions. Especially watch for lack of independence or bias in sampling.