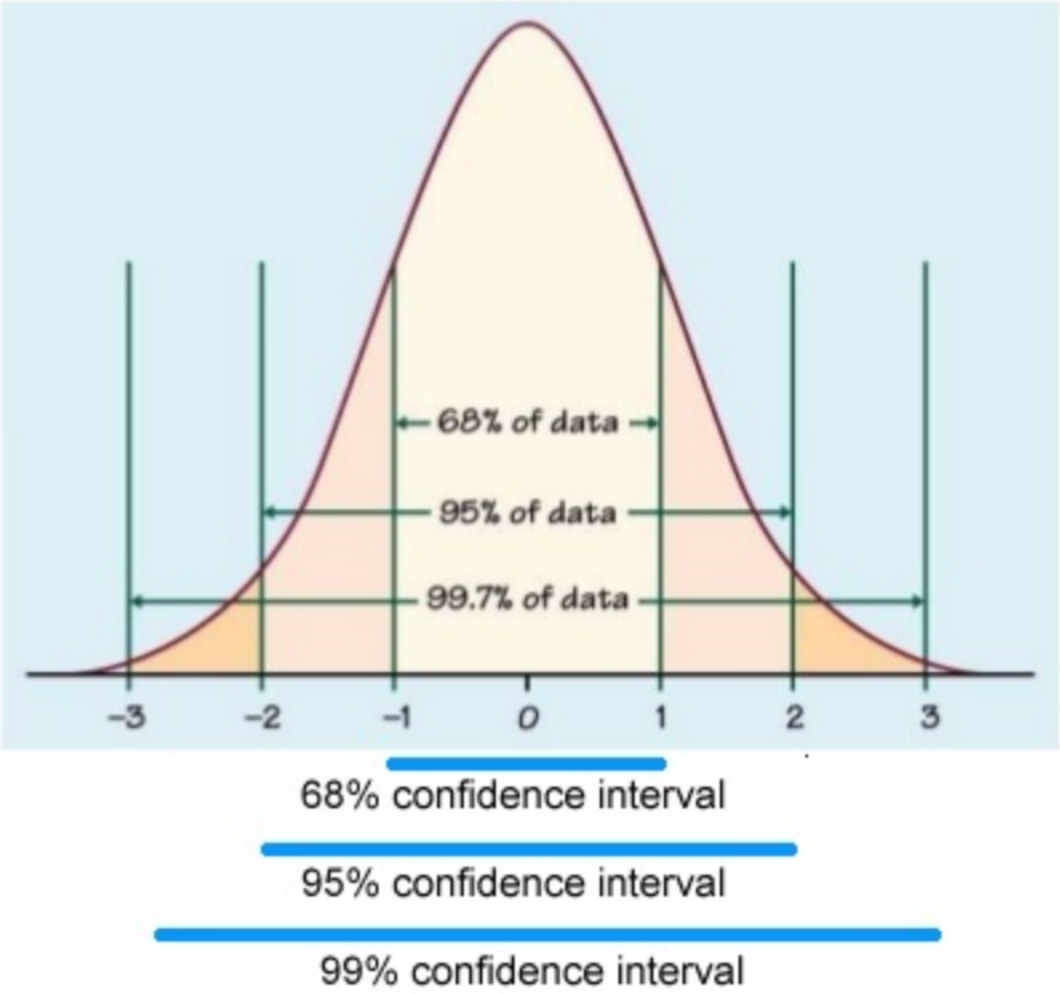
CSCI 3022: Intro to Data Science

Lecture 14: Introduction to Statistical Inference and Confidence Intervals

Rachel Cox

Department of Computer Science



Announcements & Reminders

Homework 3 - Due Friday March 13

Quizlet 05 - Due Wednesday March 11

What will we learn today?

- ❑ Statistical Inference
- ❑ Confidence Intervals

- ❑ *A Modern Introduction to Probability and Statistics, Chapter 23 & 24*

Review from Last Time

Proposition: If X is a normally distributed random variable with mean μ and standard deviation σ, then Z follows a standard normal distribution if we define:

$$Z = \frac{X-\mu}{\sigma} \quad \text{and} \quad X = \sigma Z + \mu$$

If Z is a standard normal random variable, then we can compute probabilities using the standard normal cdf

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(x)\, dx$$

The Central Limit Theorem: Let $X_1, X_2, \ldots, X_n$ be iid draws from some distribution. Then, as n becomes large…

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Statistical Inference

$\bar{x}, s, \hat{p}$ sample statistics μ, σ, p population statistics

Goal: We want to extract properties of an underlying population by analyzing sampled data.

- Is the sample mean $\bar{x}$ a good approximation of the population mean μ?
- Is the sample proportion $\hat{p}$ a good approximation of population proportion p?
- Is there a statistically significant difference between the mean of two samples?
- If the answer is Yes, how sure are we?
- How much data do we need in order to be confident in our conclusion?

Confidence Intervals

The Central Limit Theorem tells us that as the sample size n increases, the sample mean of X is close to normally distributed with expected value μ and standard deviation $\sigma/\sqrt{n}$

Standardizing the sample mean by first subtracting the expected value and dividing by the standard deviation yields a standard normal random variable.

How large does the sample need to be if (in order to use the central Limit Theorem)

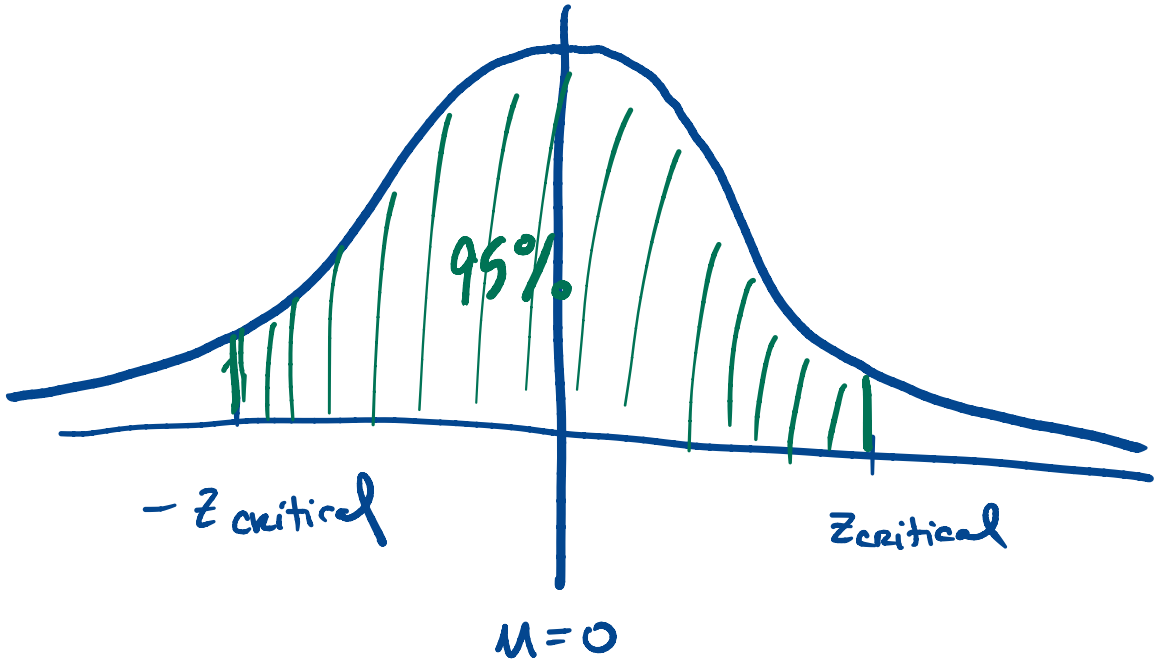
1) The population is normally distributed? $n \geq 1$
2) The population is *not* normally distributed? $n \geq 30$

Confidence Intervals

$-a \leq 3 \rightarrow a \geq -3$

95% of the area under the standard normal curve falls between -1.96 and +1.96, so we know that: $P(-1.96 \leq z \leq 1.96) = 0.95$

[Note: z = stats.norm.ppf(0.025) = -1.96]



$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$P(-1.96 \leq z \leq 1.96)$$

$$= P\left(-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right)$$

$$= P\left(-1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

$$\left(-\bar{x} \quad -\bar{x} \quad -\bar{x}\right)$$

$$= P\left(-\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = P\left(\bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

$$= 0.95$$

Confidence Intervals

The 95% confidence interval for the mean μ is given by:

$z_{crit} = 1.96$ for the 97.5th percentile

$$\left[\bar{x} - \boxed{1.96} \cdot \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + \boxed{1.96} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Which things in this expression are random, and which things are fixed?

Random: $\bar{X}$

fixed: n ← can be changed

1.96 ← can be changed depending on n% CI

σ

Confidence Intervals

Example: Suppose that you perform 20 random samples of the population and compute 95% CIs for each sample. How many of the intervals do you expect to contain the true population mean μ?

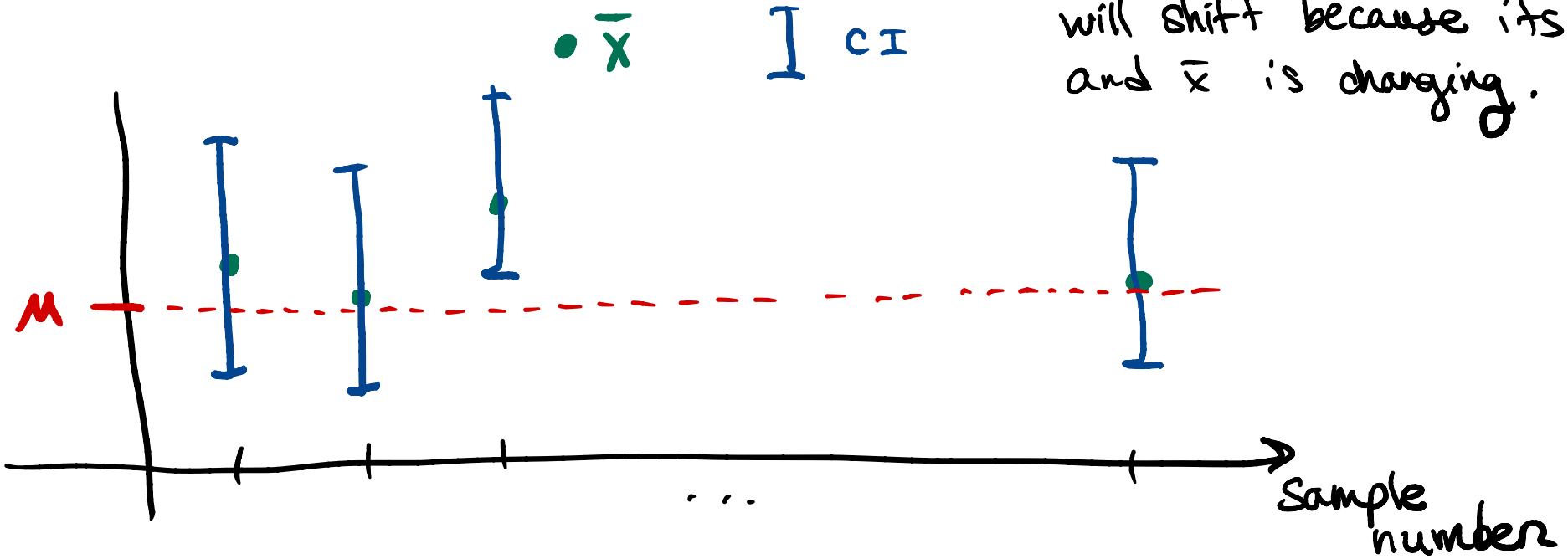
$.95 * 20 = 19$

we expect 19 of them to contain the true mean μ

$\bar{X}$ is changing

CI: $\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$

width of CI does not change, but it will shift because its centered on $\bar{x}$ and $\bar{x}$ is changing.



Confidence Intervals

Example: Fill in the blank.

for a 95% CI

The CI is centered at $\bar{X}$ and extends $1.96 * \frac{\sigma}{\sqrt{n}}$ to each side of $\bar{x}$.

The CI's width is $2 * 1.96 * \frac{\sigma}{\sqrt{n}}$, which is not random; only the location of the interval's midpoint $\bar{X}$ is random.

Confidence Intervals

What does it mean to say that we are 95% confident that the true population mean is in this interval?

Correct interpretation: In repeated sampling, 95% of all CIs obtained from sampling will actually contain the true population mean. The other 5% of CIs will not.

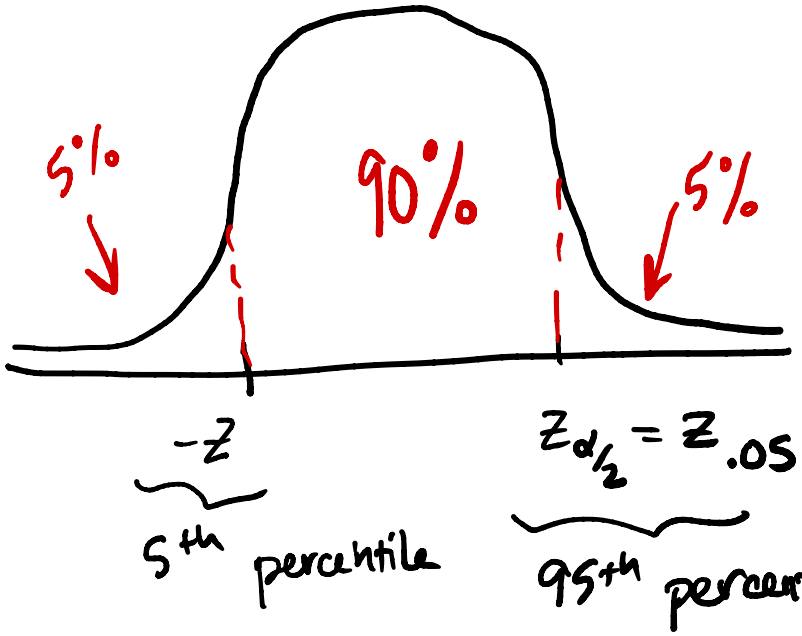
Wrong interpretation: The mean is in my 95% CI with probability 95%

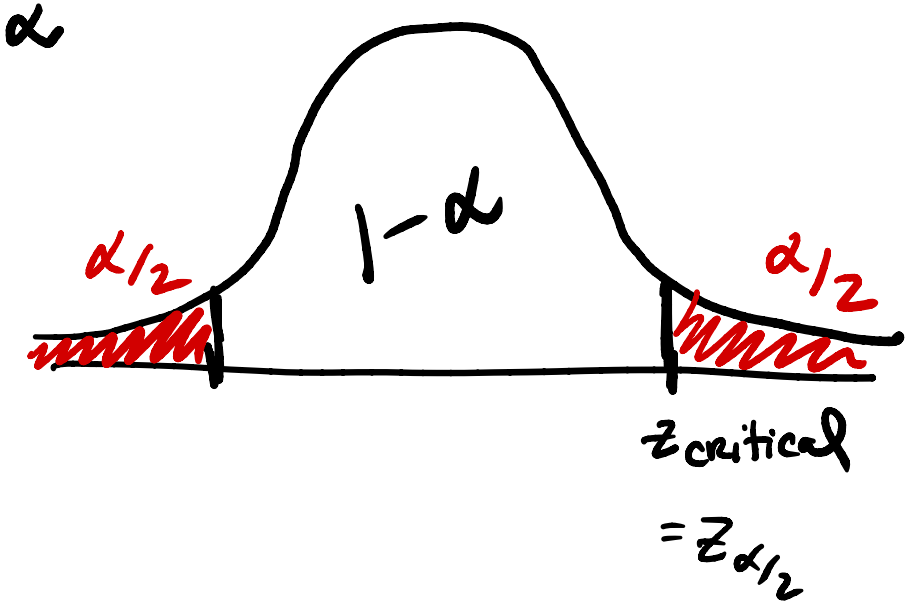
Confidence Intervals

A $100 \cdot (1-\alpha)\%$ confidence interval for the mean μ when the value of σ is known is given by

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Suppose $\alpha = .10$





$z_{\alpha/2}$ = stats.norm.ppf$(1 - \alpha/2)$

Confidence Intervals

Example: The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the US. In 2010, the survey collected responses from 1000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours/day. Suppose further that the known standard deviation of the characteristic is 2 hours/day.

Find a 90% confidence interval for the amount of relaxation hours/day.

CLT applies

$n = 1000$

$\bar{x} = 3.6$

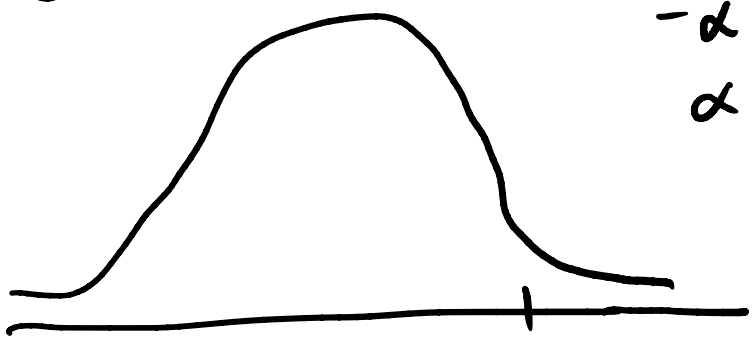
$\sigma = 2$

90% CI

$\Rightarrow 1 - \alpha = .9$

$-\alpha = -.1$

$\alpha = .1$



$z_{crit} = z_{\alpha/2} = z_{.05}$

stats.norm.ppf(1 − .05) ≈ 1.645

general CI: $\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right]$

$= \left[3.6 - 1.645 \cdot \frac{2}{\sqrt{1000}}, 3.6 + 1.645 \cdot \frac{2}{\sqrt{1000}}\right]$

$= [3.496, 3.704]$

Confidence Intervals

Example: The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the US. In 2010, the survey collected responses from 1000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours/day. Suppose further that the known standard deviation of the characteristic is 2 hours/day.

Find a 95% confidence interval for the amount of relaxation hours/day.

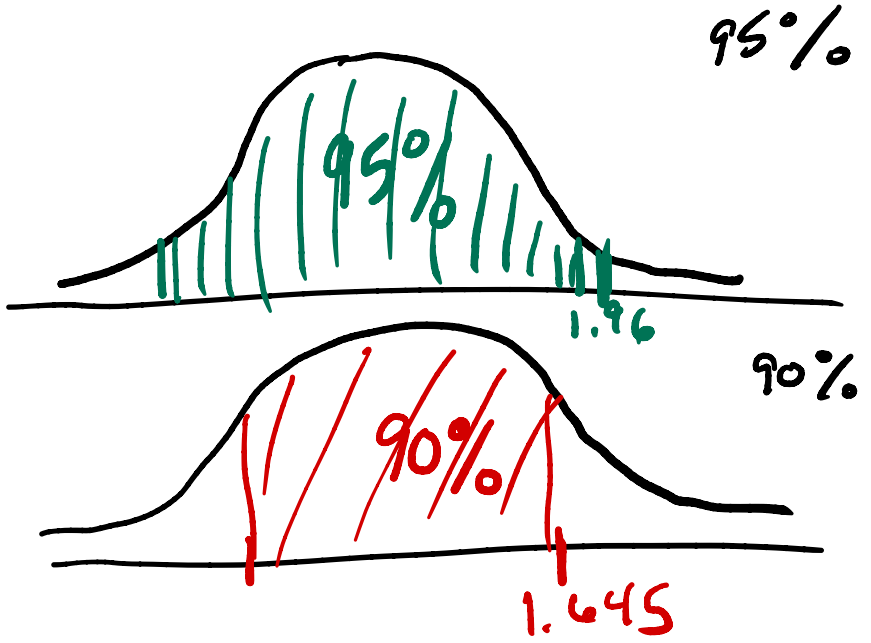
$$CI = 3.6 \pm 1.96 \cdot \frac{2}{\sqrt{1000}}$$

$$[3.476, 3.724]$$

95% CI

90% CI

$$[3.496, 3.704]$$



Confidence Intervals

Example: In the previous example, we found a 95% CI for relaxation time to be [3.48, 3.72]. Which of the following statements are true?

A. 95% of Americans spend 3.48 to 3.72 hours per day relaxing after work.

B. 95% of random samples of 1000 residents will yield CIs that contain the true average number of hours that Americans spend relaxing after work each day.

C. 95% of the time the true average number of hours an American spends relaxing after work is between 3.48 and 3.72 hours/day.

D. We are 95% sure that Americans in this sample spend 3.48 to 3.72 hours/day relaxing after work.

Confidence Intervals

Example: For the GSS data, how large would n have to be to get a 95% CI with width at most 0.1?

$$\bar{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \qquad \sigma = 2$$

$$2 * 1.96 * \frac{\sigma}{\sqrt{n}} \leq 0.1$$

$$2 * 1.96 * \frac{2}{\sqrt{n}} \leq 0.1$$

$$\frac{4 * 1.96}{0.1} \leq \sqrt{n}$$

$$\left(\frac{4 * 1.96}{0.1}\right)^2 \leq n$$

$$n \geq 6147$$

Confidence Intervals

In the previous example, we assumed that we knew the population standard deviation σ

Question: How often does this happen in real life?

almost never

→ we often estimate σ with S - the sample standard deviation.

$$CI: \quad \bar{x} \pm \frac{s}{\sqrt{n}} * z_{\alpha/2}$$

Confidence Intervals – for proportions

Let p denote the proportion of "successes" in a population. (e.g. individuals who graduated from college, computer nodes that do not fail on a given day, etc)

A random sample of n individuals is selected, and X is the number of successes in the sample.

Then X can be modeled as a binomial random variable with E[X] = np and

Var(X) = np(1-p) .

Confidence Intervals – for proportions

X = # successes

n = sample size

The estimator for p is given by $\frac{X}{n}$, aka $\hat{p}$.

The estimator is approximately normally distributed with $E[\hat{p}] = p$ and

"unbiased estimator"

$E[\hat{p}] = E\left[\frac{X}{n}\right] = \frac{1}{n}E[X] = \frac{1}{n}(np) = p$

* $Var(\hat{p}) = \frac{p(1-p)}{n}$

* $Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}np(1-p)$

Standardizing the estimate yields: $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$

(Recall: $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$)

This gives us a $100 \cdot (1-\alpha)\%$ confidence interval of:

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$$

This is our CI

Confidence Intervals – for proportions

Example: The EPA considers indoor radon levels of about 4 picocuries/liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 homes found 127 of the sampled homes to have indoor radon levels of about 4 pCi/L. Calculate the 99% confidence interval for the proportion of homes with indoor radon levels above 4 pCi/L.

$n = 200$

$X = 127$

$$\hat{p} = \frac{X}{n} = \frac{127}{200} \approx 0.635$$

$$\text{CI: } \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$$

estimate p with $\hat{p}$

$z_{\alpha/2} = z_{.01/2} = z_{.005}$

$z_{crit} = \text{stats.norm.ppf}(1 - .005) \approx 2.57$

$$\text{CI: } 0.635 \pm 2.57 * \sqrt{\frac{.635(1-.635)}{200}} \Rightarrow [0.548, 0.722]$$

Next Time:

- Two Sample Confidence Intervals