Introduction to Statistical Inference and Confidence Intervals

Administrivia

- **Homework 4** due Friday.
- **Midterm** returned in class on Wednesday (PROBABLY)

Previously on CSCI 3022

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

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

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

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

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

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

Statistical Inference

Goal: Want to extract properties of an underlying population by analyzing sampled data

Questions:

- Is sample mean $\bar{x}$ a good approximation of the population mean μ ?
- Is sample proportion $\bar{p}$ a good approximation of the 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 $\sigma/\sqrt{n}$

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

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

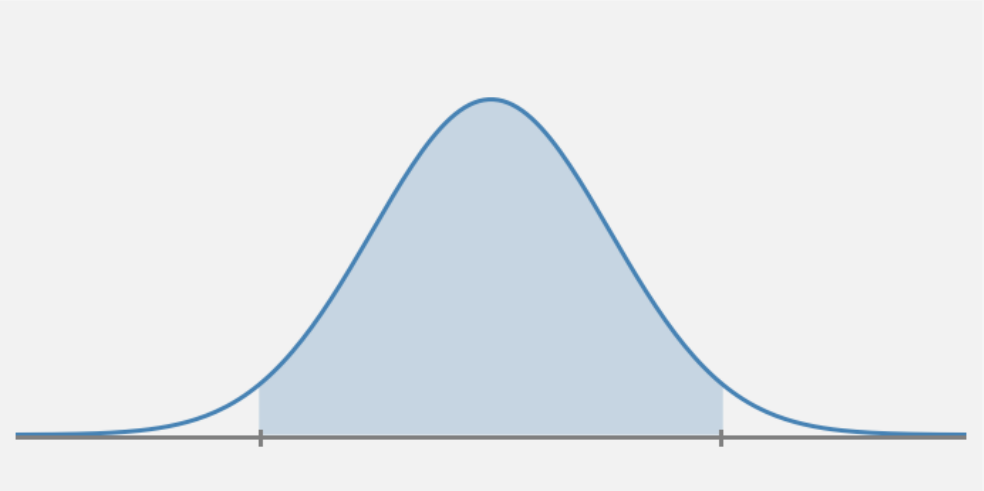
$$Z = \frac{X-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Question: How big does our sample need to be if

- The population is normally distributed $n \geq 1$
- The population is not normally distributed $n \geq 30$

Confidence Intervals

We saw a while ago that the 95% of the area under the standard normal curve falls between -1.96 and +1.96, so we know that



$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

WHERE $Z \sim N(0,1)$

This is equivalent to:

$$P\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

Confidence Intervals

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

$$P\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95 \Rightarrow$$

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \Rightarrow$$

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

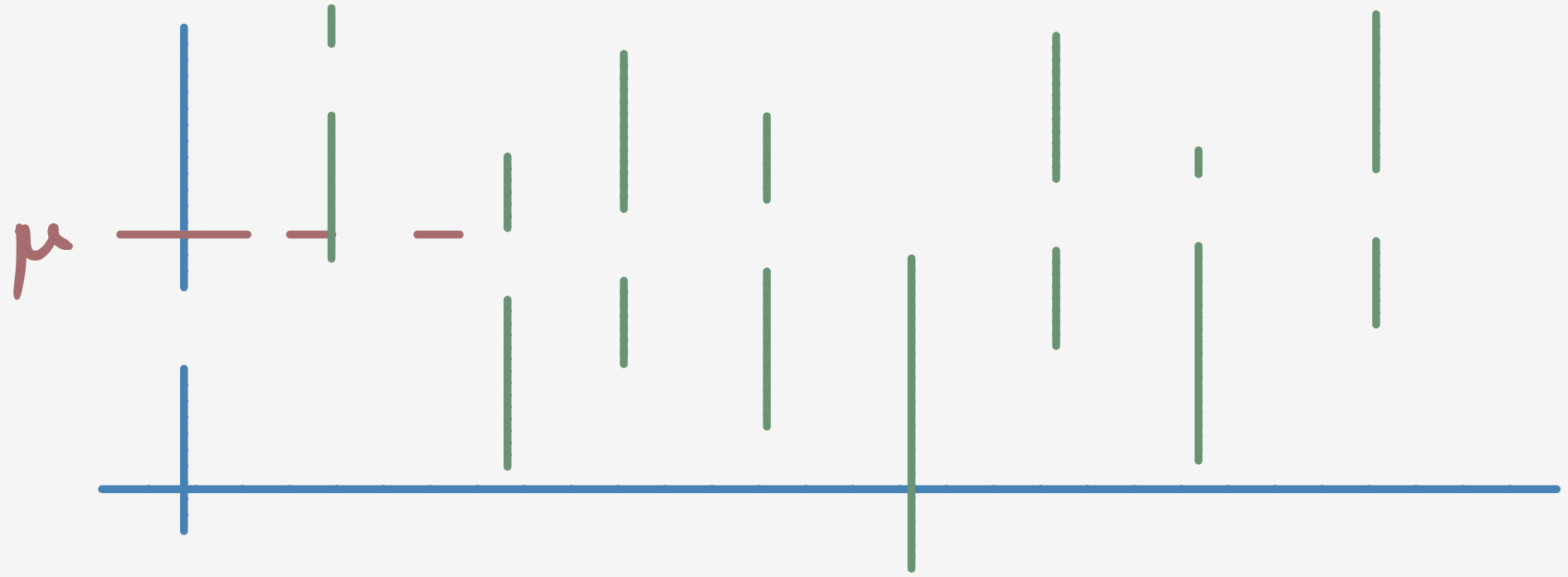
RANDOM — FIXED — RANDOM

Question: Which things in this expression are random and which things are fixed?

Confidence Intervals

The CI varies from sample to sample, so the CI is really a random interval

Question: Suppose 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 ?



Confidence Intervals

The CI is centered at _____ and extends _____ to each side of $\bar{X}$

The CIs width is _____ which is **not** random; only the location of the interval's midpoint $\bar{X}$ is random.

We often write the CI _____ as _____

Interpreting the Confidence Level

Statement: We are 95% confident that the true population mean is in this interval

WHAT DOES THIS EVEN MEAN ???

Interpreting the Confidence Level

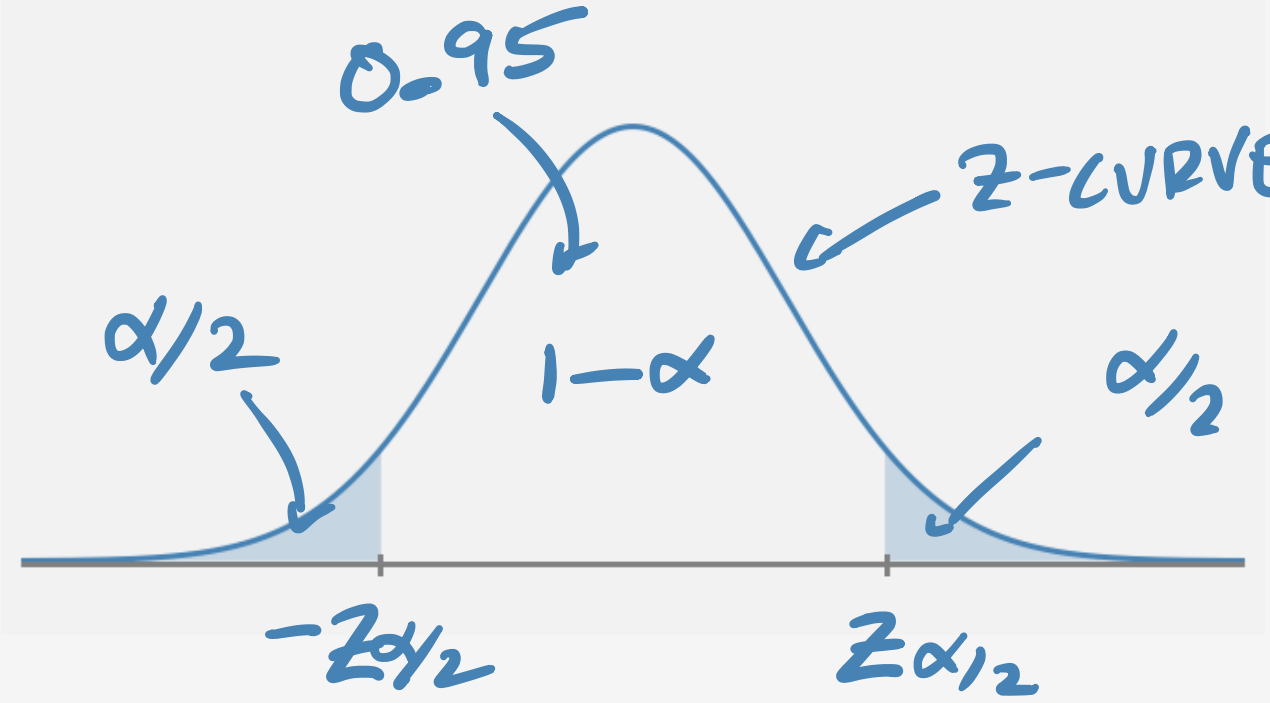
Statement: We are 95% confident that the true population mean is in this interval

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.

The confidence level is not a statement about any one particular interval. Instead it describes what would happen if a very large number of CIs were computed using the same CI formula.

Other Levels of Confidence

A probability of $1-\alpha$ is achieved by using $z_{\alpha/2}$ in place of $z_{.05/2} = z_{.025} = 1.96$



FOR 99% CI : $\alpha = .01 \Rightarrow z_{.01/2} = z_{.005}$

SCIPY.STATS.PPF(.995) = 2.57

Other Levels of Confidence

A probability of $1-\alpha$ is achieved by using $z_{\alpha/2}$ in place of $z_{.05/2} = z_{.025} = 1.96$

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

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \text{ or } \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Confidence Intervals

Example: The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of the residents of the US. In 2010, the survey collected responses from 1,000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours per day. Suppose further that the known standard deviation of the characteristic is 2 hours per day.

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

$90\% = \alpha = 0.1 \Rightarrow z_{\alpha/2} = z_{.05}$

$= \text{scipy.stats.norm.ppf}(.95) = 1.645$

$\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{1000}} = .0632 \Rightarrow 3.6 \pm 1.645 \cdot .0632$

$\Rightarrow 3.6 \pm 0.104$

Confidence Intervals

Example: The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of the residents of the US. In 2010, the survey collected responses from 1,000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours per day. Suppose further that the known standard deviation of the characteristic is 2 hours per day.

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

$95\% \Rightarrow \alpha = .05 \Rightarrow z_{.025} = \text{norm.ppf}(.975) = 1.96$

$\frac{2}{\sqrt{1000}} = 0.0632 \Rightarrow 3.6 \pm 1.96 \times .0632$

$3.6 \pm .124 \Rightarrow [3.48, 3.72]$

Question: What are the advantages/disadvantages to a wider confidence interval?

Confidence Intervals

Concept Check: 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 per day.

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

Computing Required Sample Size

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

$$\text{WIDTH} = 2 * 1.96 \frac{2}{\sqrt{n}} \underbrace{=}_{\text{WANT}} 0.1$$

SOLVE FOR n ... $\sqrt{n} = \dfrac{2 \times 1.96 \times 2}{0.1} = 78.4$

$\Rightarrow \quad n = (78.4)^2 = 6147$

Unknown Variance

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

Question: How often does this happen in real life?

Unknown Variance

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

Question: How often does this happen in real life? **NEVER!**

Solution: If n is large we use the sample variance instead

$$CI = \bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Solution: If n is small we have to do something else (more on this later)

Confidence Intervals for Proportions

Let p denote the proportion of "successes" in a population (e.g. individuals who graduated from college, compute nodes that did 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:

$$E[X] = np \quad , \quad VAR(X) = np(1-p)$$

Confidence Intervals for Proportions

The Estimator for p is given by $\hat{p} = \frac{X}{n}$

The estimator is approximately normally distributed with:

$$E[\hat{p}] = p \quad , \quad VAR(\hat{p}) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

Standardizing the estimate yields:

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

This gives us a confidence interval of

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

DON'T KNOW p!

← SUB SAMPLE PROPORTION!

Confidence Intervals for Proportions

Example: The EPA considers indoor radon levels above 4 picocuries per 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 above 4 pCi/L. Calculate the 99% confidence interval for the proportion of homes with indoor radon levels above 4 pCi/L.

$\hat{p} = \frac{127}{200} \approx 0.635$ $\quad$ 99% $\rightarrow \alpha = .01$

$\Rightarrow z_{\alpha/2} = z_{.005} = \text{norm.ppf}(.995) = 2.57$

$\hat{p} \pm z_{.005} \sqrt{\frac{.635(1-.635)}{200}} = .635 \pm 2.57 * .034$

$\Rightarrow [0.548, 0.722]$

OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 14 In-Class Notebook

Let's:

- Get some more practice computing confidence intervals
- See how we write autograders for your simulation homework

Acknowledgements

- Some of the slides in this lecture were adopted from Brian Zaharatos