# Two-Sample Confidence Intervals

# Administrivia

- **Homework 4** due Friday
- **Extra Office Hours** Thursday from 12:30-2pm in ECOT 731

# Previously on CSCI 3022

**Proposition**: If X is a normally distributed random variable with mean $\mu$ and standard deviation $\sigma$ , then Z is a standard normal distribution if

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

**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

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

A $100(1-\alpha)\%$ **confidence interval** for the mean $\mu$ with known sd. $\sigma$ is given by

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

# Statistical Inference

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

**Last time we saw**:

- How to determine a confidence interval for the population mean $\mu$
- How to determine a confidence interval for the population proportion $p$

**This time we'll see:**

- How to put a confidence interval on the **difference** between means of two populations
- How to put a confidence interval on the **difference** between proportions of two populations
- How we can get a good numerical estimate of a CI using something called the **Bootstrap**

# Difference Between Population Means

How do two sub-populations compare? In particular, are their means the same?

**Classic Motivating Examples:**

- Is a drug's effectiveness the same in children and adults?
- Does cigarette brand A contain more nicotine that cigarette brand B?
- Does a class perform better when Professor C teaches it or Professor D?
- Does email Ad E generate more customers than email Ad F?

# Difference Between Population Means

How do two sub-populations compare? In particular, are their means the same?

**Solution Process**: Collect samples from both sub-populations, and perform inference on both samples to make conclusions about $\mu_1 - \mu_2$

# Difference Between Population Means

How do two sub-populations compare? In particular, are their means the same?

**Solution Process**: Collect samples from both sub-populations, and perform inference on both samples to make conclusions about $\mu_1 - \mu_2$

**Basic Assumptions**:

○ $X_1, X_2, \ldots, X_m$ is a random sample from a distribution with mean $\mu_1$ and sd $\sigma_1$

○ $Y_1, Y_2, \ldots, Y_n$ is a random sample from a distribution with mean $\mu_2$ and sd $\sigma_2$

○ The $X$ and $Y$ samples are independent of one another.

# Difference Between Population Means

The natural estimator of $\mu_1 - \mu_2$ is the difference of the sample means, $\bar{x} - \bar{y}$

Is $\bar{x} - \bar{y}$ a good estimator for $\mu_1 - \mu_2$ ?

UNBIASED ESTIMATOR

The expected value of $\bar{X} - \bar{Y}$ is given by

$$E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = \mu_1 - \mu_2$$

The standard deviation of $\bar{X} - \bar{Y}$ is given by

$$\sqrt{VAR(\bar{X} - \bar{Y})} = \sqrt{VAR(\bar{X}) + VAR(\bar{Y})} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

# Normal Populations with Known SDs

If both populations are normal, then both $\bar{X}$ and $\bar{Y}$ are normally distributed

Independence of the two samples implies that the samples means are independent

Thus, the difference between the means is normally distributed, for any sample sizes, with:

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

# Confidence Intervals for the Difference

Standardizing $\bar{X} - \bar{Y}$ gives a standard normal random variable

$$Z = \frac{(\bar{X}-\bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$

And so we can compute a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ as

$$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

# Large Sample CIs for the Difference

Not surprisingly, if both m and n are large, then the CLT kicks in, and our confidence interval for the difference of means is valid, even when the populations are not normally distributed

$$(\bar{x}-\bar{y}) \sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}\right)$$

As $m, n \to \infty$

Furthermore, if m and n are large, and we don't know the standard deviations, we can replace them with the sample standard deviations

$$\sigma_1 \to S_1 \qquad \sigma_2 \to S_2 \quad \Rightarrow \quad \bar{x}-\bar{y} \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}}$$

# Confidence Intervals for the Difference

**Example**: Suppose you run two different email Ad campaigns over many days and record the amount of traffic driven to your website on days that each Ad is sent. In particular, suppose that Ad 1 is sent on 50 different days and generates an average of 2 million page views per day with an sd of 1 million views, and Ad 2 is sent on 40 different days and generates an average of 2.25 million page views per day with an sd of a half million views. Find a 95% confidence interval for the difference in average page views per day (in units of millions of views).

$\bar{x} = 2 \quad \sigma_1 = 1 \qquad m = 50$

$\bar{y} = 2.25 \quad \sigma_2 = 0.5 \qquad n = 40$

$\bar{x} - \bar{y} = 2 - 2.25 = -0.25$

$\sqrt{\frac{1^2}{50} + \frac{.5^2}{40}} = 0.162$

$\alpha = .05 \Rightarrow z_{.025} = 1.96$

$\Rightarrow \quad -0.25 \pm 1.96 \times 0.162 = -0.25 \pm 0.318$

$\Rightarrow [-0.568, 0.068]$

# Confidence Intervals for the Difference

**Example**: Suppose you run two different email Ad campaigns over many days and record the amount of traffic driven to your website on days that each Ad is sent. In particular, suppose that Ad 1 is sent on 50 different days and generates an average of 2 million page views per day with an sd of 1 million views, and Ad 2 is sent on 40 different days and generates an average of 2.25 million page views per day with an sd of a half million views. Find a 95% confidence interval for the difference in average page views per day (in units of millions of views).

# Confidence Intervals for the Difference

**Looking Forward:** What does our confidence interval tell us about the effectiveness of the two advertisements?

THE DATA suggests THAT AD 1 might BE BETTER, BUT THE 95% CONFIDENCE INTERVAL COVERS 0, SO THERES A VERY REASONABLE CHANGE THERE IS NO SIGNIFICANT DIFFERENCE.

# Difference Between Population Proportions

What if we want to compare population proportions?

Suppose that a sample of size m is selected from the first population and a sample of size n is selected from the second population.

Let X denote the number of units with the characteristic in pop 1 (number of "successes") and Y denote the number of units with the characteristic in pop 2

Reasonable estimators for the population proportions are: $\hat{p}_1 = \frac{X}{m}$ , $\hat{p}_2 = \frac{Y}{n}$

The natural estimator for the difference between population proportions $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2$$

# Difference Between Population Proportions

Now, let $\hat{p}_1 = \dfrac{X}{m}$ and $\hat{p}_2 = \dfrac{Y}{n}$ where $X \sim Bin(m, p_1)$ and $Y \sim Bin(n, p_2)$

Assuming that X and Y are independent, we can show that

$$E[\hat{p}_1 - \hat{p}_2] = E[\hat{p}_1] - E[\hat{p}_2] = \frac{1}{m}E[X] - \frac{1}{n}E[Y]$$

$$= \frac{1}{m}mp_1 - \frac{1}{n}np_2 = p_1 - p_2$$

where the standard deviation is approximated well by

# Difference Between Population Proportions

$$VAR(\hat{p}_1 - \hat{p}_2) = VAR(\hat{p}_1) + VAR(-\hat{p}_2)$$

$$= VAR(\hat{p}_1) + VAR(\hat{p}_2) = VAR\left(\frac{X}{m}\right) + VAR\left(\frac{Y}{n}\right)$$

$$= \frac{1}{m^2} VAR(X) + \frac{1}{n^2} VAR(Y) = \frac{m p_1(1-p_1)}{m^2} + \frac{n p_2(1-p_2)}{n^2}$$

$$= \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n} \Rightarrow \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}$$

$$\longrightarrow \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

# CIs for the Difference of Proportions

The $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is then given by

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

# CIs for the Difference of Proportions

The $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is then given by

# CIs for the Difference of Proportions

**Example:** A study was published in the New Engl. J. of Med. In 1997 describing an experiment designed to compare treating cancer patients with chemotherapy only and a course of treatment involving both chemo and radiation. Of 154 individuals who received the chemo-only treatment, 76 survived at least 15 years, whereas 98 of the 164 patients who received the hybrid treatment survived at least 15 years.

What is the 99% confidence interval for this difference of proportions?

$$\hat{p}_1 = \frac{76}{154} \approx .494 \quad , \quad \hat{p}_2 = \frac{98}{164} \approx 0.598$$

$$\alpha = .01 \Rightarrow \alpha/2 = .005 \Rightarrow z_{\alpha/2} = 2.576$$

$$\left( \frac{.494(1-.494)}{154} + \frac{.598(1-.598)}{198} \right)^{1/2} =$$

# CIs for the Difference of Proportions

**Example**: A study was published in the New Engl. J. of Med. In 1997 describing an experiment designed to compare treating cancer patients with chemotherapy only and a course of treatment involving both chemo and radiation. Of 154 individuals who received the chemo-only treatment, 76 survived at least 15 years, whereas 98 of the 164 patients who received the hybrid treatment survived at least 15 years.

What is the 99% confidence interval for this difference of proportions?

# CIs for the Difference of Proportions

**Example**: A study was published in the New Engl. J. of Med. In 1997 describing an experiment designed to compare treating cancer patients with chemotherapy only and a course of treatment involving both chemo and radiation. Of 154 individuals who received the chemo-only treatment, 76 survived at least 15 years, whereas 98 of the 164 patients who received the hybrid treatment survived at least 15 years.

What is the 99% confidence interval for this difference of proportions?

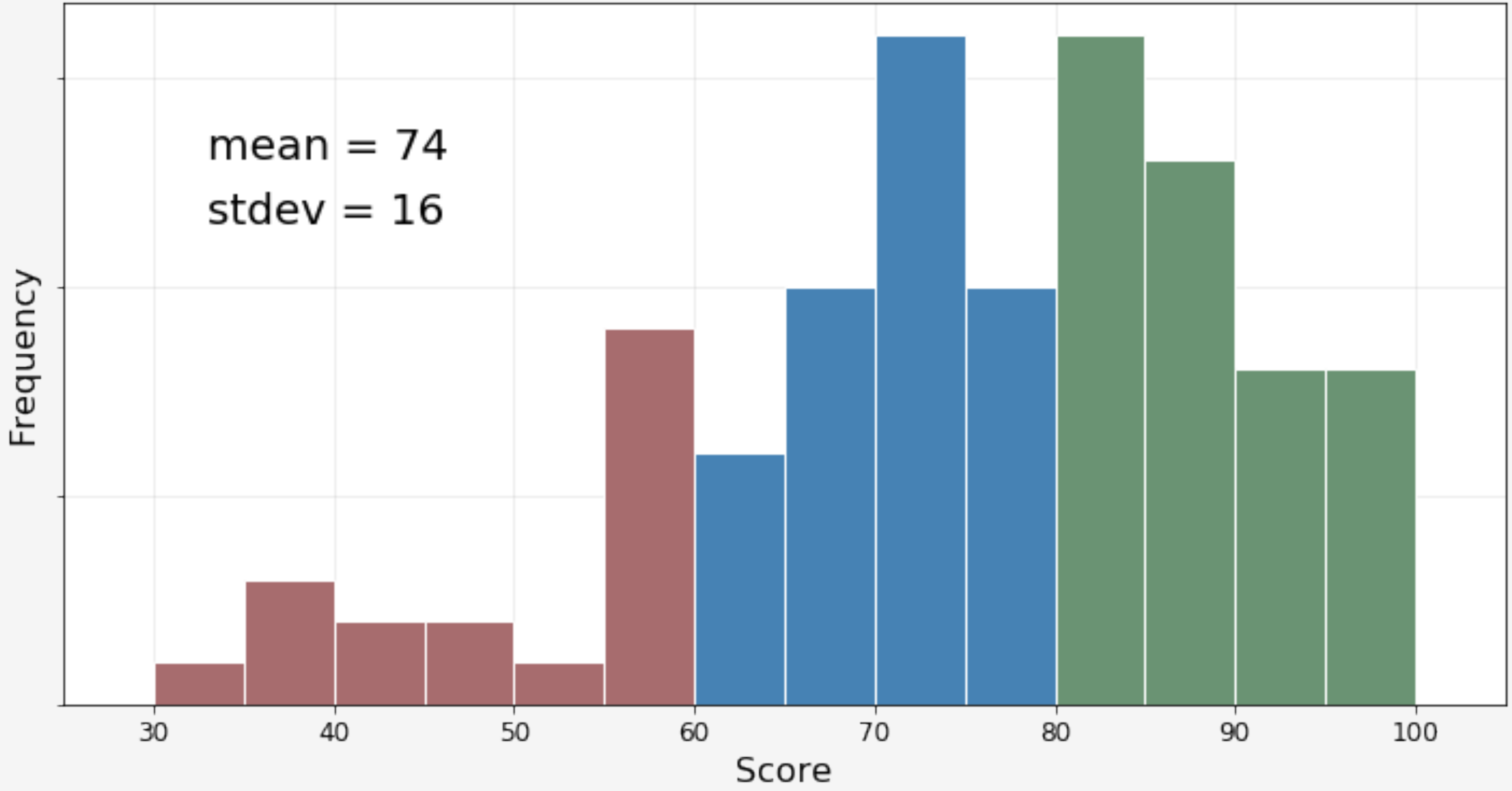
# Writing Autograders

Suppose you're a TA for Intro Data Science, and your professor-boss has tasked you with writing an autograder for a homework assignment which asks students to write a simulation to estimate the expected winnings in the game of Chuck-a-Luck.

# Writing Autograders

Now suppose your professor-boss asks you to write an autograder for a simulation of Miniopoly. Specifically, she asks you to check solutions to the function that estimates the probability that a player goes Bankrupt within the first 20 turns of the game. How is this problem different from the Chuck-a-Luck problem? How should you proceed?

# Midterm Results



# OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 15 In-Class Notebook

**Let's**:

- Get some more practice computing confidence intervals

# Acknowledgements

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