CSCI 3022

intro to data science with probability & statistics

Lecture 15
March 7, 2018

Introduction to Statistical Inference & Confidence Intervals

NOTE: It was L'Hospital who happily stole Johann Bernoulli's work & published it as his own. The next time you take an indeterminant limit, remember that & call it "Bernoulli's Rule" instead!

TONY WUZ HERE

Department of Computer Science
UNIVERSITY OF COLORADO BOULDER

Da~~n L~~arremore

Last time 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

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

- **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

non-normal
or
normal

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

2

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

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

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
  - How to determine a confidence interval for the population proportion
- **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 hris teaches it or Professor D? an
  - 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**:

m X's indep. $\bar{X}$ S_x

OLD NEWS

  - $X_1, X_2, \ldots, X_m$ is a random sample from a distribution with mean μ_1 and sd σ_1

n Y's

  - $Y_1, Y_2, \ldots, Y_n$ is a random sample from a distribution with mean μ_2 and sd σ_2
  - The X and Y samples are independent of each other.

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 $\mu_1 - \mu_2$?

$\bar{X}$ & $\bar{Y}$ r.v.

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

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

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

$$SD[\bar{X}-\bar{Y}] = \sqrt{Var[\bar{X}-\bar{Y}]} = \sqrt{Var[\bar{X}] + Var[\bar{Y}]} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

$(-1)^2$

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 sample means are independent.
- Therefore, the difference between the means is normally distributed, for any sample sizes, with:

$$\underbrace{\bar{X} - \bar{Y}}_{\text{estimator}} \sim N\Big(\underbrace{\mu_1 - \mu_2}_{\text{expected value of est.}}, \underbrace{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}_{\text{variance}}\Big)$$

Confidence Interval for the difference

messy?

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

But oh so nice! Z

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

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

$$\underbrace{\bar{X} - \bar{Y}} \pm Z_{\alpha/2} \underbrace{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \quad *$$

$$\boxed{\text{central est.} \pm Z\text{-score} * \underline{SD}}$$

Large sample CIs for the difference

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

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

$$\sigma_1^2 \mapsto s_1^2 = \frac{1}{m-1}\sum_i (x_i - \bar{x})^2$$

$$\sigma_2^2 \mapsto s_2^2 = \frac{1}{n-1}\sum_j (y_j - \bar{y})^2$$

Confidence Interval for the Difference

$[-0.508,\ 0.068]$

- **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).

$\bar{X} = 2$

$m = 50$

$s_1 = 1$

$\bar{Y} = 2.25$

$n = 40$

$s_2 = 0.5$

$z_{\frac{\alpha}{2}} = z_{\frac{.05}{2}} = 1.96$

$$95\%\ CI = (\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

$$= (2 - 2.25) \pm 1.96 \cdot \sqrt{\frac{1^2}{50} + \frac{0.5^2}{40}} = [-0.508,\ 0.068]$$

Confidence Interval 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 Interval for the Difference

HERE

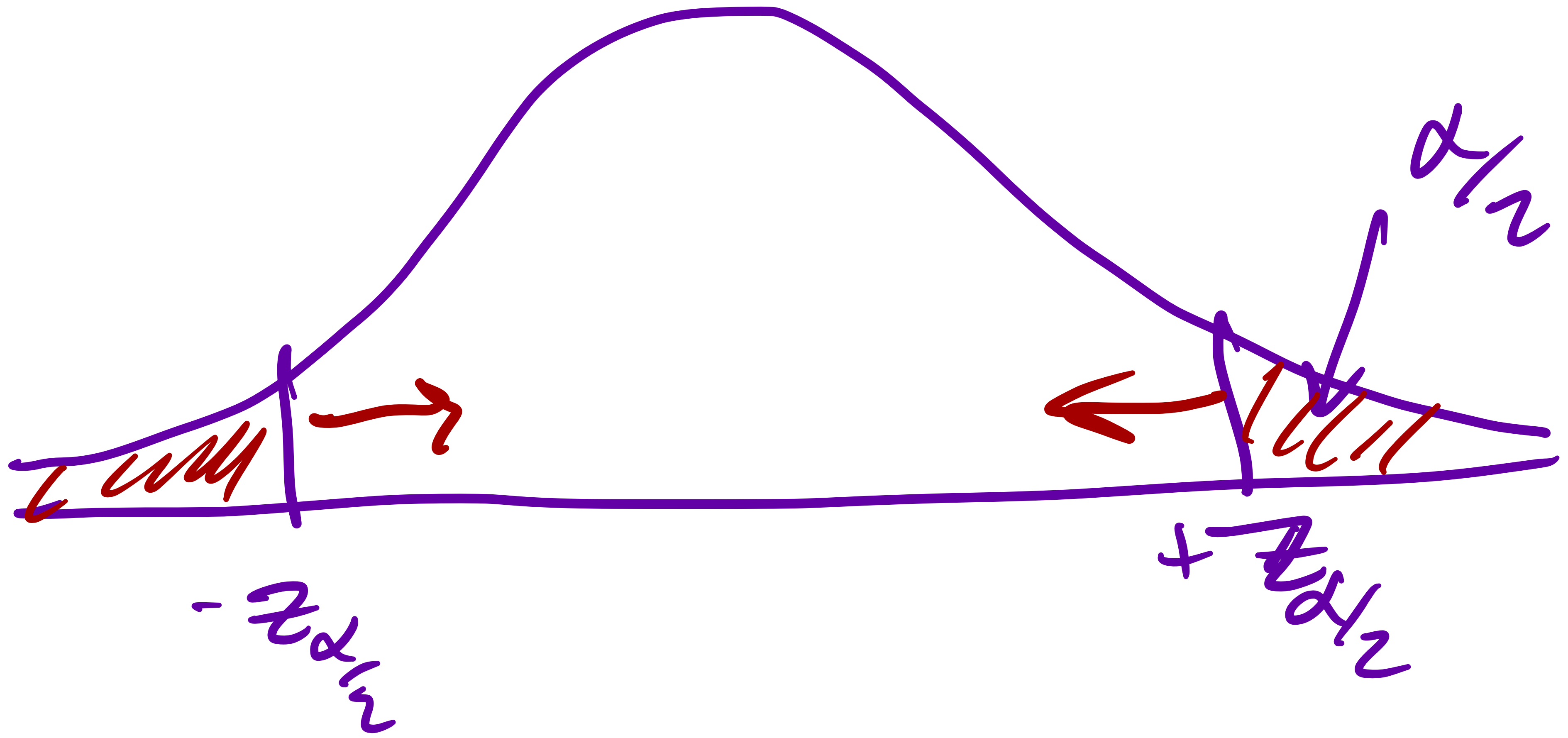
CI width: $2 \cdot z_{\alpha/2} \cdot SD$

- **Looking forward to interpretation**: What does our confidence interval tell us about the effectiveness of the two advertisements?

[-0.5 , +0.1] (ish)

$\bar{X} < \bar{Y} \rightarrow \bar{Y}$ is better?

contains 0 → so no statistically significant difference at $\alpha = 0.05$ confidence level



What happens if we increase α?

$\alpha \uparrow \Rightarrow z_{\alpha/2} \downarrow \Rightarrow$ CI width $\downarrow \Rightarrow$ 0 gets kicked out

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 population 1 (number of "successes") and Y denote the number of units with the characteristic in population 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

$\hat{p}_1 - \hat{p}_2$ ← estimate of $p_1 - p_2$

↑ real thing wanted

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

$$E[\hat{p}_1 - \hat{p}_2] = E[\hat{p}_1] - E[\hat{p}_2] = E\left[\frac{X}{m}\right] - E\left[\frac{Y}{n}\right] = \frac{1}{\cancel{m}}\cancel{m}p_1 - \frac{1}{\cancel{n}}\cancel{n}p_2 = p_1 - p_2$$

- The standard deviation is approximated well by

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

I know $var(\hat{p}_1 - \hat{p}_2) = var(\hat{p}_1) + var(\hat{p}_2)$

next

Difference Between Population Proportions

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

$$= var\left(\frac{X}{m}\right) + var\left(\frac{Y}{n}\right)$$

$$= \frac{1}{m^2} var(X) + \frac{1}{n^2} var(Y)$$

$$= \frac{1}{m\cancel{^2}} \cancel{m}\, p_1(1-p_1) + \frac{1}{n\cancel{^2}} \cancel{n}\, p_2(1-p_2)$$

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

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

CIs for the Difference of Proportions

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

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}$$

CIs for the Difference of Proportions

$Z_{0.005} = 2.576$

$\frac{76}{154} \approx 0.494$

$98/164 \approx 0.598$

- **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?**

1 2

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

$\alpha = 0.01 \qquad Z_{0.005} = 2.576$

chemo only: $\frac{76}{154} = \hat{p}_1 \approx 0.494 \qquad m = 154$

hybrid: $\frac{98}{164} = \hat{p}_2 \approx 0.598 \qquad n = 164$

$$0.494 - 0.598 \pm 2.576\sqrt{\frac{0.494(1-0.494)}{154} + \frac{0.598(1-0.598)}{164}}$$

Writing an Autograder

- 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.

1. We know true mean of Chuck-a-Luck winnings → we calculated it!
2. Run the student's code n times
3. Compute a CI for the student's code's mean.
4. Is the true mean in the CI?

Writing an Autograder

- 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?

① This is about proportions.

② We don't have true proportion.

→ but we have a correct simulation.

③ compute $\hat{p}_1$ (student) via n simulations

$\hat{p}_2$ (correct) via n simulations

④ compute CI for diff in proportions.

⑤ does it contain 0?

⑥ If not, run codes again.