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

Dan ~~Larremore~~

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

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

- 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 (his) 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**:
  - $X_1, X_2, \ldots, X_m$ is a random sample from a distribution with mean μ_1 and sd σ_1
  - $Y_1, Y_2, \ldots, Y_n$ 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 $\mu_1 - \mu_2$?
- The expected value of $\bar{X} - \bar{Y}$ is given by

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

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

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

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

$(-1)^2$

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:

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

estimator

expected value of est.

variance

Confidence Interval for the difference

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

messy!

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

But oh so nice! :)

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

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

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

central est. $\pm$ Z-score * 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).

$\bar{x} = 2$, $m = 50$, $s_1 = 1$

$\bar{y} = 2.25$, $n = 40$, $s_2 = 0.5$

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

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

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

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

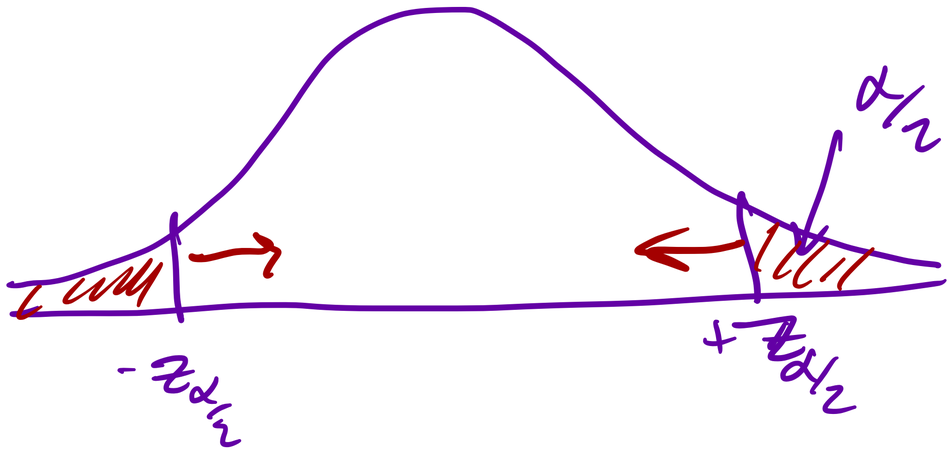
CI width: $2 \cdot z_{\alpha/2} \cdot SD$

- **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