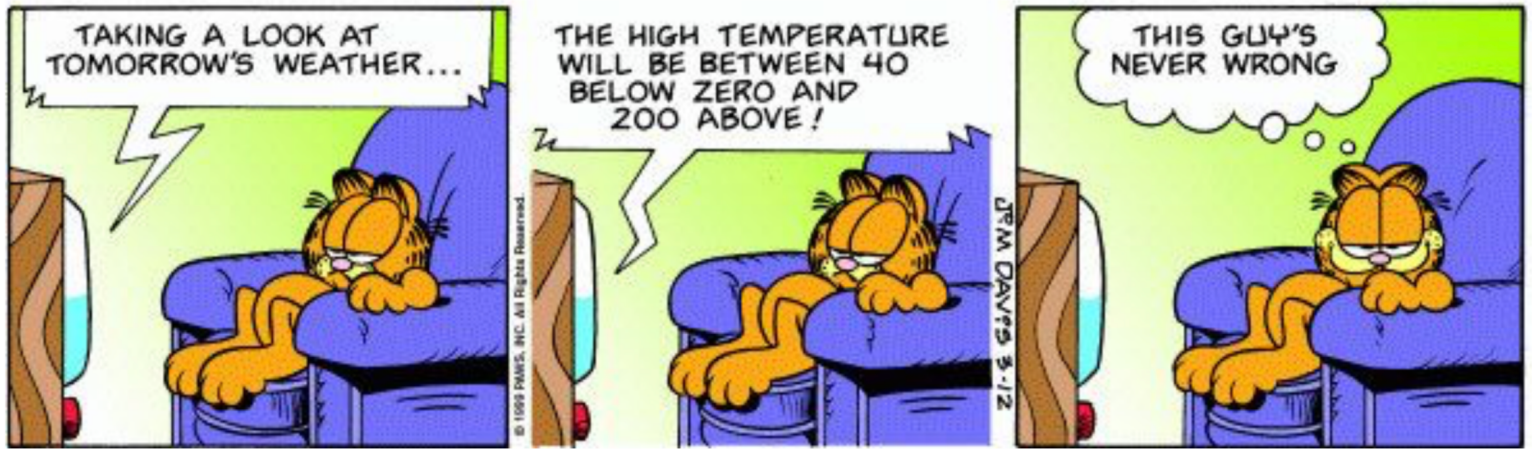
CSCI 3022: Intro to Data Science

Lecture 15: Two-Sample Confidence Intervals

Rachel Cox

Department of Computer Science



Announcements & Reminders

- ☐ Quizlet 05 Due tonight at 11:59pm
- ☐ HW03 Due Friday at 11:59pm
- ☐ We begin Remote lectures + office hours on Monday March 16th.

What will we learn today?

- ❑ Two-sample 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:

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

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

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

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

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{for propositions: } \hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Statistical Inference

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

Last Time:

- How to calculate a CI for population mean μ
- How to calculate a CI for population proportion p

Next:

- How to get a CI for the difference between the mean of two populations.
- How to get a CI for the difference between the proportions of two populations.

Difference between population means

Question: How do two sub-populations compare? Are their means the same?

Classic motivating examples:

- Is a drug's effectiveness the same in children and adults?
- Does cigarette brand X contain more nicotine than brand Y?
- Does a class perform better when taught using method One or method Two?
- Does organizing a website give better user experience using format A or format B?

→ A/B testing

Difference between population means

How do two sub-populations compare? 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:

$X_1, X_2, \dots, X_m$ is a random sample from distribution 1 with mean μ_1 and SD σ_1
$Y_1, Y_2, \dots, Y_n$ is a random sample from distribution 2 with mean μ_2 and SD σ_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 in sample means: $\bar{x} - \bar{y}$

But is it a good estimator?

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

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

The SD of $\bar{X} - \bar{Y}$ is given by:

$$= \mu_1 - \mu_2$$

$$SD = \sqrt{Var(\bar{X} - \bar{Y})} = \sqrt{Var(\bar{x}) + Var(-\bar{Y})} = \sqrt{Var(\bar{x}) + (-1)^2 Var(\bar{Y})}$$

$$= \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

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 also independent.

- The difference in sample means is normally distributed, for any sample sizes, with:

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

Normal Distribution
for the difference of sample means.

Confidence intervals for the difference

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

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

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

transformation is analogous to what we've been doing

So we can compute a $100 \cdot (1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$

Recall $\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{n}$

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

Large–sample CIs

If both m and n are large ($n \geq 30$), then the CLT kicks in and our confidence interval for the difference of means is valid, even if the populations are not normally distributed.

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

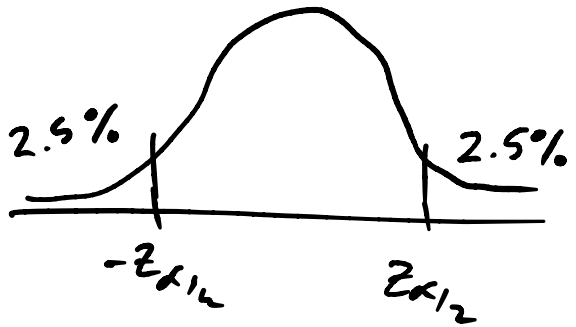
Confidence Intervals for the Difference

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 was sent. Ad 1 was sent on 50 different days and generates an average of 2 million page views per day, with a SD of 1 million page views. Ad 2 was sent on 40 different days and generates an average of 2.25 million page views per day, with SD of half a million views. Find a 95% confidence interval for the average page views for each ad (in units of millions of views).

Ad 1	Ad 2
$n_1 = 50$	$n_2 = 40$
$\bar{x}_1 = 2$	$\bar{x}_2 = 2.25$
$SD_1 = 1$	$SD_2 = 0.5$

CI for Ad 1:

$$\bar{x}_1 \pm z_{\alpha/2} \cdot \frac{SD_1}{\sqrt{n_1}}$$



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

$= z_{.025}$

$= \text{stats.norm.ppf}(1 - .025)$

$= 1.96$

$$2 \pm 1.96 \cdot \frac{1}{\sqrt{50}}$$

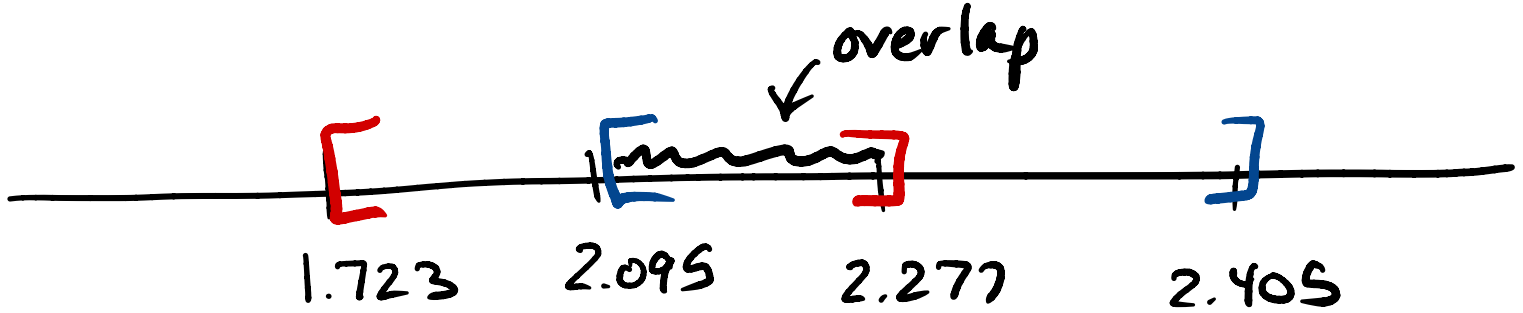
$\Rightarrow CI_1 : [1.723, 2.277]$

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 was sent. Ad 1 was sent on 50 different days and generates an average of 2 million page views per day, with a SD of 1 million page views. Ad 2 was sent on 40 different days and generates an average of 2.25 million page views per day, with SD of half a million views. Find a 95% confidence interval for the average page views for each ad (in units of millions of views).

CI_2: $2.25 \pm 1.96 * \frac{0.5}{\sqrt{40}}$ $[2.095, 2.405]$

from previous page: CI_1: $[1.723, 2.277]$



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 was sent. Ad 1 was sent on 50 different days and generates an average of 2 million page views per day, with a SD of 1 million page views. Ad 2 was sent on 40 different days and generates an average of 2.25 million page views per day, with SD of half a million views. Find a 95% confidence interval for the **difference** in average page views for each ad (in units of millions of views).

CI for Difference: $(\bar{X} - \bar{Y}) \pm z\alpha_{/2} \cdot \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$

$= (2 - 2.25) \pm 1.96 \sqrt{\frac{1^2}{50} + \frac{0.5^2}{40}}$

$= -0.25 \pm 1.96 * 0.162$

$= [-0.568, 0.068]$

Confidence Intervals for the Difference

What does our CI tell us about the effectiveness of the two advertisements.

Difference between population proportions

What if we want to compare population proportions instead of means?

Example: Suppose a sample of size *m* is selected from population 1, and a sample of size *n* from population 2.
Let X denote the number of units with the characteristic of interest in population 1 (# successes), and let Y denote the number of units with the characteristic of interest in population 2.

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

Natural estimator for the difference in population proportions $p_1 - p_2$: $\hat{p}_1 - \hat{p}_2$

Difference between population proportions

Example: (continued) Suppose a sample of size *m* is selected from population 1, and a sample of size *n* from population 2.
Let X denote the number of units with the characteristic of interest in population 1 (# successes), and let Y denote the number of units with the characteristic of interest in population 2.

Now let $\widehat{p_1} = \frac{X}{m}$ and $\widehat{p_2} = \frac{Y}{n}$, where X ~ Bin(m, p_1) and Y ~ Bin(n, p_2)

Assuming that X and Y are independent, we can show that the expected value and SD are estimated by:

$$E[\widehat{p_1} - \widehat{p_2}] = p_1 - p_2$$

$$Var(\widehat{p_1} - \widehat{p_2}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Difference between population proportions

$$E[\widehat{p_1} - \widehat{p_2}] = p_1 - p_2$$

$$Var(\widehat{p_1} - \widehat{p_2}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Derivation:

$$E[\hat{p}_1 - \hat{p}_2] = E[\hat{p}_1] - E[\hat{p}_2]$$

$$= E\left[\frac{x}{m}\right] - E\left[\frac{Y}{n}\right] = \frac{1}{m}E[x] - \frac{1}{n}E[Y]$$

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

$$= p_1 - p_2$$

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

$$= Var\left(\frac{x}{m}\right) + Var\left(-\frac{Y}{n}\right) = \frac{1}{m^2} Var(x) + \left(-\frac{1}{n}\right)^2 Var(Y)$$

$$= \frac{1}{m^2}(mp_1(1-p_1)) + \frac{1}{n^2}(np_2(1-p_2))$$

CIs for the Difference of Proportions

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

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

we often don't know p_1 or p_2
→ estimate with $\hat{p}_1$ and $\hat{p}_2$

CIs for the Difference of Proportions

Example: A study was published in the New England Journal of Medicine in 1997 describing an experiment designed to compare treating cancer patients with chemotherapy only and a course of treatment involving both chemotherapy and radiation. Of the154 individuals who received the chemo-only treatment, 76 survived at least 15 years, whereas 98 of the 165 patients who received the hybrid treatment survived at least 15 years. What is the 99% CI for this difference in proportions?

chemo-only $n_1 = 154$, $X = 76$ $\hat{p}_1 = \frac{76}{154}$

$z_{\alpha/2} = z_{.01/2} = z_{.005}$
$= \text{stats.norm.ppf}(.995)$
$= 2.576$

hybrid treatment $n_1 = 165$, $Y = 98$ $\hat{p}_2 = \frac{98}{165}$

$$\text{CI}: \frac{76}{154} - \frac{98}{165} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= 76/154 - 98/165 \pm 2.576 \cdot \sqrt{76/154} \longrightarrow \text{new slide}$$

CIs for the Difference of Proportions

Example: A study was published in the New England Journal of Medicine in 1997 describing an experiment designed to compare treating cancer patients with chemotherapy only and a course of treatment involving both chemotherapy and radiation. Of the154 individuals who received the chemo-only treatment, 76 survived at least 15 years, whereas 98 of the 165 patients who received the hybrid treatment survived at least 15 years. What is the 99% CI for this difference in proportions?

$$\text{CI: } \frac{76}{154} - \frac{98}{165} \pm 2.576 \cdot \sqrt{\frac{\frac{76}{154}\left(1-\frac{76}{154}\right)}{154} + \frac{\frac{98}{165}\left(1-\frac{98}{165}\right)}{165}}$$

$$= 0.494 - 0.598 \pm 2.576 * 0.0555$$

$$= [-0.247, 0.039]$$

Next Time:

- Two Sample Confidence Intervals