Chapter 8: Hypothesis Testing with Two Samples

DSCC 462
Computational Introduction to Statistics

Anson Kahng
Fall 2022

Comparison of Two Means

- We used a hypothesis test to compare the unknown mean of a single population to some fixed, known value, μ_0
- Often, we want to compare the means of two separate populations, where *both* means are unknown
  - E.g., is the average height of Americans equal to the average height of Canadians?
- Must determine whether the two samples are paired or independent
  - Paired: Weight before and after surgery for a group of men
  - Independent: Height of Americans compared to height of Canadians

Paired Samples

- For each observation in the first group, there is a corresponding observation in the second group
- *Self-pairing*: Measurements are taken on a single subject at two distinct time points (before and after)
- *Matched pairing*: Match two individuals with similar demographics / characteristics and compare their differences in response
    - Make the pair as similar as possible with respect to important characteristics (e.g., age, gender, socioeconomic status, etc.) depending on the setting

Paired Samples

- Use pairing to control for extraneous sources of variation that may otherwise influence results
- By measuring on the same sample, we remove natural biological variability between people
- If we pair based on a certain characteristic (e.g., age), then we do not have to worry about that characteristic (age) influencing the results
- In general, pairing makes comparisons more *precise*

Paired Samples

- In the situation of paired samples, our data often is the difference between elements in each pair
  - Sample 1: $x_{11}, x_{21}, \ldots, x_{n1}$
  - Sample 2: $x_{12}, x_{22}, \ldots, x_{n2}$
  - Difference: $d_1 = x_{11} - x_{12}, d_2 = x_{21} - x_{22}, \ldots, d_n = x_{n1} - x_{n2}$

Paired Samples: Example

- We measure the sitting heart rate for 40 people after they have been sitting for 5 minutes
- These 40 people then run on a treadmill for 30 minutes, sit for 5 minutes, and then have their heart rate measured
- Sample 1: “before” heart rate
- Sample 2: “after” heart rate
- Goal: We are interested in how heart rate changed after running; we care about *before – after*

Paired Samples: Example

- The data is paired, so we can use the differences $d_i = b_i - a_i$ as the data
  - Reduces to the one-sample problem: compare differences to 0
- We need the mean of the d_i's and its sampling distribution
- Mean: $\bar{x}_d$, sample standard deviation s_d (unknown true σ_d)
- Standard error: $\dfrac{s_d}{\sqrt{n}}$
- Assumption: $\bar{x}_d \sim N\left(\mu_d, \sigma_d/\sqrt{n}\right)$

Paired Samples: Example

- Let the true difference in population mean be μ_d
- Then, our hypotheses are:
  - $H_0 : \mu_d = 0$ and $H_1 : \mu_d \neq 0$
- Test statistic: $t = \dfrac{\bar{x}_d - \mu_d}{s_d/\sqrt{n}}$
- Under the null hypothesis, $t \sim t_{n-1}$ ($n - 1$ degrees of freedom)
- Calculate p-value, p, by finding the probability of observing a mean difference at least as extreme as $\bar{x}_d$, given that $\mu_d = 0$
- Reject H_0 if $p \leq \alpha$

Paired Samples: Example

- Consider the heart rates before and after running on a treadmill example

Before	**After**	**Difference**
55	60	-5
62	75	-13
61	65	-4
72	89	-17
57	70	-13

- Does heart rate significantly change after running on a treadmill?
- Test at the $\alpha = 0.01$ significance level

Paired Samples: Example

Before	After	Difference
55	60	-5
62	75	-13
61	65	-4
72	89	-17
57	70	-13

- $\bar{x}_d = \dfrac{-5 - 13 - 4 - 17 - 13}{5} = -10.4$
- $s_d^2 = \dfrac{(-5 + 10.4)^2 + (-13 + 10.4)^2 + (-4 + 10.4)^2 + (-17 + 10.4)^2 + (-13 + 10.4)^2}{5 - 1} = 31.8$
- $s_d = \sqrt{31.8} = 5.64$

Paired Samples: Example

- $H_0 : \mu_d = 0$ vs. $H_1 : \mu_d \neq 0$
- Calculate t statistic: $t = \dfrac{\bar{x}_d - \mu_d}{s_d/\sqrt{n}} = \dfrac{-10.4 - 0}{5.64/\sqrt{5}} = -4.12$
- Degrees of freedom: df = 5 - 1 = 4
- In R: `2*pt(-4.12,4) = 0.0146`
- Since $0.0146 > \alpha = 0.01$, we fail to reject H_0
- There is insufficient evidence to conclude that the average heart rate after running on a treadmill for 30 minutes is significantly different than the average sitting heart rate at the $\alpha = 0.01$ significance level

Paired Samples: Confidence Intervals

- We can also calculate confidence intervals for paired differences
- Our (two-sided) interval for unknown σ is $\bar{x}_d \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$
- Continuing with the heart rates example, we can get the following 99% confidence interval:

$$\bar{x}_d \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}} = -10.4 \pm 4.60 \cdot \frac{5.65}{\sqrt{5}}$$
$$= (-21.62, 1.62)$$

- I am 99% confident that the interval (-21.62, 1.62) captures the true mean before-after difference in heart rate for subjects running on a treadmill

Paired Samples: R Code

```
> before <- c(55, 62, 61, 72, 57)
>  after <- c(60, 75, 65, 89, 70)
>  d <- before-after
> t.test(d)

	One Sample t-test

data:  d
t = -4.1239, df = 4, p-value = 0.01457
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -17.401928  -3.398072
sample estimates:
mean of x
    -10.4

> t.test(before, after, paired=T)

	Paired t-test

data:  before and after
t = -4.1239, df = 4, p-value = 0.01457
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.401928  -3.398072
sample estimates:
mean of the differences
                  -10.4
```

One-sided Paired t-test Example

- Setup: Trace metals in drinking water affect the flavor, and an unusually high concentration can pose a health hazard

- Ten pairs of data were taken measuring zinc concentration in bottom water and surface water (two layers of water from same sources)

- Based on this sample, the mean difference in zinc concentrations between the bottom water and surface water was $\bar{x}_d = 0.0804$ and the standard deviation was $s_d = 0.9023$

- Q: Does the true average zinc concentration in the bottom water exceed that of surface water?

One-sided Paired t-test Example

- Let $\alpha = 0.05$
- $H_0 : \mu_d \leq 0$ vs. $H_1 : \mu_d > 0$
- Calculate t: $t = \dfrac{\bar{x}_d - \mu_d}{s_d/\sqrt{n}} = \dfrac{0.0804 - 0}{0.0923/\sqrt{10}} = 2.88$
- Degrees of freedom: df = 10 - 1 = 9
- R: `1-pt(2.88,9) = 0.0091`
- Since $0.0091 < \alpha = 0.05$, we reject H_0
- There is sufficient evidence to conclude that, on average, the bottom zinc concentration is higher than the surface zinc concentration

Independent Samples

- Suppose we have measurements on two samples of subjects
  - Heart rates for a group of people who have been sitting
  - Heart rates for a group of people who have been running
- The two underlying populations are independent and normally distributed
- The first population has mean μ_1 and the second population has mean μ_2
- Test whether the two populations are identical
  - $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$
  - Or, equivalently, $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$

Independent Samples

- Population 1: mean μ_1 and variance σ_1^2
  - Draw sample of size n_1 with sample mean $\bar{x}_1$ and sample variance s_1^2
- Population 2: mean μ_2 and variance σ_2^2
  - Draw sample of size n_2 with sample mean $\bar{x}_2$ and sample variance s_2^2
- We can compare the means of these two populations in two different ways
  - Equal population variances: known ($\sigma_1^2 = \sigma_2^2$) or unknown (t-test)
  - Unequal population variances: $\sigma_1^2 \neq \sigma_2^2$

Equal, Known Variances

- Consider situations where it is either known or reasonable to assume that the two population variances are the same, and we know $\sigma^2 = \sigma_1^2 = \sigma_2^2$
- Suppose we are interested in the difference between the two population means
- By extension of the CLT, $\overline{X}_1 - \overline{X}_2$ is approximately normal with mean $\mu_1 - \mu_2$ and standard error $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- Because we assume $\sigma^2 = \sigma_1^2 = \sigma_2^2$, the standard error is $\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$
- Then, we can use a z-test with $z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ as our test statistic

Equal, Unknown Variances

- When σ^2 is unknown, we use a sample estimate (t-test)
- We assume that both populations have the same variance, so we want to consider both samples in our calculation of the sample variance, s_p^2
- We can think of s_p^2 as the *pooled* estimate of common variance σ^2:
  - $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$
- In other words, s_p^2 is the weighted average of the two sample variances, s_1^2 and s_2^2, where the weights are the degrees of freedom for each sample

Equal, Unknown Variances

- We can use a t-test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
- The test statistic t has a t distribution with $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ degrees of freedom
- Determine the p-value, p, based on this t distribution
- Reject H_0 if $p \leq \alpha$

Equal, Unknown Variances: Example

- Setup: A study examines whether targeted political ads makes potential voters more likely to donate money to candidates
  - A sample of 54 potential voters who saw targeted political ads had mean donation amount $\bar{x}_1 = 240$ USD with standard deviation $s_1 = 30$ USD
  - A sample of 49 potential voters who did not see targeted political ads had mean donation amount $\bar{x}_2 = 220$ USD with standard deviation $s_2 = 48$ USD
  - Is there a significant difference (at the $\alpha = 0.05$ significance level) in average donations between potential voters who saw targeted political ads and those who did not, assuming equal variances?

Equal, Unknown Variances: Example

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$, $H_1 : \mu_1 - \mu_2 \neq 0$, significance $\alpha = 0.05$
- Calculate the pooled (sample) variance:
- Next, calculate the t-statistic:

Equal, Unknown Variances: Example

- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$, $H_1 : \mu_1 - \mu_2 \neq 0$, significance $\alpha = 0.05$
- Calculate the pooled (sample) variance:

$$
\begin{aligned}
s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 2)} \\
&= \frac{(54 - 1)(30)^2 + (49 - 1)(48)^2}{54 + 49 - 2} \\
&= 1567.248
\end{aligned}
$$

- Next, calculate the t-statistic:

$$
\begin{aligned}
t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\
&= \frac{(240 - 220) - 0}{\sqrt{1567.248 \left(\frac{1}{54} + \frac{1}{49} \right)}} \\
&= 2.561
\end{aligned}
$$

Equal, Unknown Variances: Example

- We compare this t-value with a t distribution with (54 - 1) + (49 - 1) = 101 df
- In R: `2*(1-pt(2.561,101)) = 0.012`
- Since the p-value is 0.012 < 0.05, we reject the null hypothesis
- There is significant evidence to conclude that the average donation amount from potential voters who saw targeted political ads is significantly different from the average donation amount from potential voters who did not see targeted political ads

Equal, Unknown Variances: Example

- What if we want a $(1 - \alpha)\,\%$ confidence interval for $\mu_1 - \mu_2$?
- We can write this as $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$
- We are $(1 - \alpha)\,\%$ confident that this interval contains the true mean difference $\mu_1 - \mu_2$
- In our voter example, a 95% confidence interval for the mean difference in donations is as follows:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = (240 - 220) \pm 1.984\sqrt{1567.248\left(\frac{1}{54} + \frac{1}{49}\right)}$$
$$= (4.50, 35.50)$$

- We are 95% confident that the interval (4.50, 35.50) captures the true difference in average donation amounts

Unequal Variance (Welch t-test)

- Let's now consider the situation where the variances of the two populations are not assumed to be equal

- Need to modify the two-sample t-test

- Instead of used a pooled estimate of the variance, we substitute s_1^2 for σ_1^2 and s_2^2 for σ_2^2

- Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

- The exact distribution for this t-statistic is difficult to derive
  - Use an approximation

Unequal Variance (Welch t-test)

- Need to come up with a value for the degrees of freedom (Welch-Satterthwaite equation)

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}\right)}$$

- Under $H_0 : \mu_1 = \mu_2$, we can approximate the distribution of t by a t distribution with ν degrees of freedom

- Calculate the p-value, p, based on this t distribution

- Reject H_0 if $p \leq \alpha$

Unequal Variance (Welch t-test): Example

- Setup: Are average blood calcium levels different for people over 60 years old compared to 10-30 year olds?
- Take a sample of 15 people who are over 60 years old
  - Sample mean blood calcium level: 9.3
  - Sample standard deviation: 1.86
- Take a sample of 7 people between 10 and 30 years old
  - Sample mean blood calcium level: 10.6
  - Sample standard deviation: 0.92

Unequal Variance (Welch t-test): Example

- Setup: Are average blood calcium levels different for people over 60 years old compared to 10-30 year olds?
- Take a sample of 15 people who are over 60 years old
  - Sample mean blood calcium level: 9.3
  - Sample standard deviation: 1.86
- Take a sample of 7 people between 10 and 30 years old
  - Sample mean blood calcium level: 10.6
  - Sample standard deviation: 0.92
- Test at $\alpha = 0.05$ significance level

Unequal Variance (Welch t-test): Example

- Test at $\alpha = 0.05$ significance level
- Let μ_1 be the average blood calcium level for people over 60 years old
- Let μ_2 be the average blood calcium level for people between 10-30 years old
- Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$

Unequal Variance (Welch t-test): Example

- Calculate t-statistic:

- Calculate degrees of freedom (ν):

- R:

- Since the p-value is less than $\alpha = 0.05$, we reject the null hypothesis and conclude a difference in average blood calcium level between the two age groups

Unequal Variance (Welch t-test): Example

- Calculate t-statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{9.3 - 10.6}{\sqrt{\frac{1.86^2}{15} + \frac{0.92^2}{7}}} = -2.19$$

- Calculate degrees of freedom (ν):

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}\right)} = \frac{\left(\frac{1.86^2}{15} + \frac{0.92^2}{7}\right)^2}{\left(\frac{\left(\frac{1.86^2}{15}\right)^2}{15 - 1} + \frac{\left(\frac{0.92^2}{7}\right)^2}{7 - 1}\right)} = 19.82$$

- R: `2*pt(-2.19,19.82)=0.0407`

- Since the p-value is less than $\alpha = 0.05$, we reject the null hypothesis and conclude a difference in average blood calcium level between the two age groups

Unequal Variance (Welch t-test): Example

- Similarly, we can calculate a $(1-\alpha)\,\%$ confidence interval for $\mu_1 - \mu_2$ as

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

- We are $(1-\alpha)\,\%$ confident that this interval contains the true mean difference $\mu_1 - \mu_2$

- For our blood calcium example, a 95% confidence interval for the difference in average blood calcium levels for 60+ year olds and 10-30 year olds is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)} = (9.3 - 10.6) \pm 2.09\sqrt{\left(\frac{1.86^2}{15} + \frac{0.92^2}{7}\right)} = (-2.54, -0.06)$$

- I am 95% confident that the interval (-2.54, -0.06) captures the true difference in average blood calcium levels between 60+ year olds and 10-30 year olds

Unequal Variance (Welch t-test): Example

- Setup: Do babies born in Rochester have larger birthweights than babies born in Buffalo?
- Let μ_1 be the mean birthweight of babies born in Rochester
- Let μ_2 be the mean birthweight of babies born in Buffalo
- Assume unequal variances
- Let $\alpha = 0.05$ be our significance level

Unequal Variance (Welch t-test): Example

- Hypotheses: $H_0 : \mu_1 \leq \mu_2$ vs. $H_1 : \mu_1 > \mu_2$, samples of sizes $n_1 = 140$ and $n_2 = 172$
- $\bar{x}_1 = 8.2$ lbs, $\bar{x}_2 = 7.9$ lbs, $s_1^2 = 1.4$ lbs², and $s_2^2 = 1.1$ lbs²
- Calculating $t =$
- Calculating $\nu =$
- In R:
- Conclusion:

Unequal Variance (Welch t-test): Example

- Hypotheses: $H_0 : \mu_1 \leq \mu_2$ vs. $H_1 : \mu_1 > \mu_2$, samples of sizes $n_1 = 140$ and $n_2 = 172$
- $\bar{x}_1 = 8.2$ lbs, $\bar{x}_2 = 7.9$ lbs, $s_1^2 = 1.4$ lbs², and $s_2^2 = 1.1$ lbs²
- Calculating $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{8.2 - 7.9}{\sqrt{\dfrac{1.4^2}{140} + \dfrac{1.1^2}{172}}} = 2.101$
- Calculating $\nu = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\left(\dfrac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}\right)} = \dfrac{\left(\dfrac{1.4^2}{140} + \dfrac{1.1^2}{172}\right)^2}{\left(\dfrac{\left(\frac{1.4^2}{140}\right)^2}{140 - 1} + \dfrac{\left(\frac{1.1^2}{172}\right)^2}{172 - 1}\right)} = 260.353$
- In R: `1-pt(2.101,260.353)=0.018`
- Since the p-value 0.018 < 0.05, we reject H_0 and conclude that there is sufficient evidence that babies born in Rochester are heavier than babies born in Buffalo