9 Inference Based On Two Samples

Introduction

In chapters 7–8, we looked at using hypothesis tests and confidence intervals to complete statistical inference on one population parameter. In this chapter, we'll extend these methods to do statistical inference on two population parameters.

9.1: z-tests and Confidence Intervals for a Difference Between Two Population Parameters

<u>Example Question:</u> Suppose you want to buy a dog and are considering Beagles and Pugs. You want to select the breed that lives longer. A sample of 12 Beagles lived on average 11.2 years with a known population standard deviation of 3 years. A sample of 12 Pugs lived on average 10.8 years with a known population standard deviation of 3.4 years. Is there sufficient evidence to suggest that the two breeds have a difference in average lifespans?

(a) One way to test this claim is create a confidence interval for the difference in means $(\mu_1 - \mu_2)$.
(b) Another approach to testing this claim is to see how consistent that sample is with initial hypothesis of $\mu_1 - \mu_2 = 0$.

<u>Note:</u> We must make an assumption of whether the samples from the two populations are independent or dependent. In this section (9.1) and the next section (9.2), we will assume the samples are independent.

<u>Assumptions for Comparison of Independent Samples:</u>

(1) $X_1, X_2, \ldots, X_{n_1}$ are a random sample from a distribution with mean μ_1 and variance σ_1^2.
(2) $Y_1, Y_2, \ldots, Y_{n_2}$ are a random sample from a distribution with mean μ_2 and variance σ_2^2.
(3) The X and Y samples are independent.

Notation:

	True Population Parameters			Sample Statistics		
Group	Mean	Variance	SD	Mean	Variance	SD
Group #1	μ_1	σ_1^2	σ_1	$\bar{x}_1$	s_1^2	s_1
Group #2	μ_2	σ_2^2	σ_2	$\bar{x}_2$	s_2^2	s_2

Theorem: If X and Y are independent, then
(a) $E[\bar{X} - \bar{Y}] = \mu_1 - \mu_2$
(b) $Var(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ (σ_1^2 and σ_2^2 may be unknown)

Proof:

Writing Hypotheses for Difference in Means

(a) Testing $\mu_1 \neq \mu_2$

(b) Testing $\mu_1 > \mu_2$

(c) Testing $\mu_1 < \mu_2$

Example: Write the hypotheses to test $\mu_1 > \mu_2 + 4$

We can complete a hypothesis test for the difference in means (assuming the variances or standard deviations of the two populations are known) by using the following test statistic.

$$\boxed{z = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$

Example: A sample of 12 Brand A LED bulbs lived on average 11.2 years with a population standard deviation of 3 years. A sample of 12 Brand B LED bulbs lived on average 10.8 years with a population standard deviation of 3.4 years. Is there sufficient evidence to suggest that the bulbs from the two companies have a difference in average lifespans?

A two-sided confidence interval for the difference in means can also be completed with the following formula, assuming again that the population standard deviations are known.

$$\boxed{(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

9.2: The Two-Sample t-test and Confidence Interval

If the population standard deviations of the two populations are unknown, we can estimate them and use similar methods as in section 9.1.

A confidence interval for the difference in means can be completed with the following formula if the population standard deviations are unknown.

$$\boxed{(\bar{x_1} - \bar{x_2}) \pm t_{\alpha/2,\nu} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

For the above confidence interval, you must calculate the degrees of freedom (ν). Round the value of ν to the next lowest integer if it is not a whole number.

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

Example: Suppose it is claimed that people in Okinawa live longer on average than people on the Japanese mainland. A sample of 25 people from Okinawa found an average lifespan of 84.3 years with a sample variance of 123. A sample of 22 people from Japanese mainland found an average life expectancy of 83.1 years with a sample variance of 140. Verify the correct degrees of freedom is 43 and create a 95% confidence interval for the difference in means of the two populations.

We can complete a hypothesis test for the difference in means (assuming the variances or standard deviations of the two populations are unknown) by using the following test statistic. This is called a two-sample t-test.

$$\boxed{t = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, df = \nu}$$

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

Example: For the Okinawa life expectancy example, test the claim that $\mu_1 > \mu_2$ at $\alpha = 0.05$.

9.3: Analysis of Paired Data

In the previous two sections, we assumed the two samples were independent. This may not always be the case.

Example: Before/after studies (same individuals measured twice)

Def: A sample is paired if the sample subjects are measured twice.

A paired t-test will control for differences between individuals.

Advantage: A paired t-test allows us to use a smaller sample size. This is because the variance of the difference of the two samples will be smaller than if the samples were independent.

Disadvantage: It's not always feasible. If this is the case, you can use a two-sample t-test and assume the samples are independent.

CI and Hypothesis Tests for Paired Samples (μ_d)

When doing inference for the true average change (μ_d), we compare "treatments" (e.g. before/after) on an individual level.

Create a "new", single sample based on the sample differences:

1. For each individual, define d_i = difference between treatments for subject i.
2. Calculate $\bar{d} = \frac{1}{n} \cdot \sum d_i$
3. Calculate $s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$

Using the "new" sample, use the same procedures for inference as for a single population.

$$\boxed{t = \frac{\bar{d} - \mu_{d_0}}{s_d/\sqrt{n}}, df = n - 1}$$

$$\boxed{\bar{d} \pm t_{\alpha/2,n-1} \cdot \frac{s_d}{\sqrt{n}}}$$

Example: A baseball coach randomly samples six players and has them watch an instructional video on fielding. The coach hopes that the video will help the players reduce the number of fielding errors they commit. Assume that the difference in the number of errors before and after watching the video is approximately normally distributed. Compute a six-step hypothesis test to test the claim the average number of errors is less after watching the video at $\alpha = 0.05$.

Player	1	2	3	4	5	6
Errors Before	12	9	0	5	4	3
Errors After	9	6	1	3	2	3
d_i						
$d_i - \bar{d}$						
$(d_i - \bar{d})^2$						

9.4: Inferences Concerning a Difference Between Population Proportions

We can complete a hypothesis test for the difference in proportions.

$$\boxed{z = \frac{(\bar{p_1} - \bar{p_2}) - D_0}{\sqrt{(\hat{p} \cdot (1 - \hat{p}) \cdot (\frac{1}{n_1} + \frac{1}{n_2})}}}$$

$$\boxed{\hat{p} = \frac{n_1}{n_1 + n_2}\bar{p_1} + \frac{n_2}{n_1 + n_2}\bar{p_2}}$$

We can also create a confidence interval for the difference in proportions.

$$\boxed{(\bar{p_1} - \bar{p_2}) \pm z_{\alpha/2} \cdot \sqrt{\frac{\bar{p_1}(1 - \bar{p_1})}{n_1} + \frac{\bar{p_2}(1 - \bar{p_2})}{n_2}}}$$

<u>Example:</u> Suppose a sample of 83 college students at UCD found 34 of them had binge drank in the last year and a sample of 80 college students at UCB found 45 of them had binge drank in the last year. Create a 90% confidence interval for the difference in proportions between the campuses. Is there evidence that one campus had a significantly higher proportion of binge drinkers?

Statistical Power

We know that $\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 \mid H_0 \text{ is true})$, so if $\alpha = 0.05$, then there is a 5% chance we reject the null hypothesis given that the null hypothesis is true. What if the null hypothesis is not true?

$\beta = P(\text{Fail to reject } H_0 \mid H_0 \text{ is false})$

$Power = P(\text{Reject } H_0 \mid H_0 \text{ is false})$

Consider the hypothesis test $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 = 3$. Given that H_a is true, what is the probability that we correctly reject H_0?

Illustration:

Def: Cohen's d is defined as the difference between two means divided by the standard deviation.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$$

If we are dealing with a two-sample test, then $s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$.

More about statistical power:

- Some rough guidelines on power are:
  1. $d < 0.2$ "small effect size"
  2. $0.2 < d < 0.8$ "moderate effect size"
  3. $d > 0.8$ "large effect size"
- Larger sample sizes result in higher power
- Larger differences in population means result in higher power
- A well-designed experimental design requires power analysis to determine the necessary sample size for a particular level of power.

Example: A new drug has been developed for lowering blood pressure. Participants are chosen to have a systolic blood pressure between 140 and 180 mmHg. Suppose that previous studies suggest that the standard deviation of patients' blood pressure is 12 mmHg. Suppose 100 participants are chosen for the treatment group and 100 participants are chosen for the control group. Additionally, suppose the new treatment is hypothesized to reduced blood pressure by 3 mmHg relative to the control group. What is the probability we detect a drop in blood pressure?