Outlier

Section Summary | Introduction to Statistics

Hypothesis Testing With Two Samples

2-Sample Tests: Independent or Dependent

1. 2 Sample Hypothesis Testing
    a. Chapter 9 is about 2 sample hypothesis tests
    b. These tests consist of the same mechanics of a 1 sample test: null hypothesis, alternative hypothesis, test statistic, and critical region.
    c. The main difference will come in the form of independent or dependent samples
2. Two Sample Independent Test
    a. There are two groups, but no matching pair or natural pairing between samples.
    b. An example is “two different people each with their own scores”.
3. Two Sample Dependent Test
    a. The two samples are related to each other through a natural pairing.
    b. An example is “same person with two different scores” or “a before treatment value and after treatment value”.
4. Conclusion
    a. Before proceeding with a hypothesis test, always ask if the samples are independent or dependent.
    b. Dependent and Independent Samples have different procedures when it comes to testing.

2-Sample Hypothesis Test: Matched Samples Test

1. Procedure
    d. Convert the 2 samples into 1 sample by taking the difference between values in each pair. This creates a new list of values which will be treated as the sample.

e. Find the mean and standard deviation of those differences.

f. State the null and alternative hypothesis

- H_0: $\mu_d = 0$ H_a:$\mu_d \neq 0$
- H_0: $\mu_d = 0$ H_a:$\mu_d > 0$
- H_0: $\mu_d = 0$ H_a:$\mu_d < 0$
- Note: the subscript d stands for differences and the assumed difference doesn't have to be zero either (0 is most commonly used).

d. Assumptions:

- Data are independent and identically distributed
- Sample size of differences is large, or the underlying distribution of differences is normally distributed

e. If the assumptions hold, then the test statistic $t = \frac{\bar{X}_d - \mu_d}{\frac{s_d}{\sqrt{n}}}$ will be used.

This test statistic is associated with the t distribution with $n - 1$ degrees of freedom. $\bar{X}_d$ refers to the sample mean of the differences. The symbol μ_d refers to the assumed population mean of the differences found in the null hypothesis. The symbol s_d represents the standard deviation for the sample of differences. As usual, n refers to the sample size.

f. The rejection region will be determined by the t distribution with n-1 degrees of freedom, the level of significance, and the alternative hypothesis (in terms of which tails being used).

g. The p-value is computed using the t distribution with n-1 degrees of freedom in the same way it would be for a 1-sample hypothesis test.

2. Example

a. The example is testing to see if the number of runs scored per baseball team has increased from 1968 to 1969. The question being examined is did the average number of runs per team actually increase.

b. This means the null hypothesis is $H_0\colon \mu_d = 0$ and the alternative hypothesis is $H_a\colon \mu_d > 0$

c. The level of significance is $\alpha = 0.05$.

d. Now taking the pairwise difference of the runs scored by each team in 1968 and1969, a new sample of differences is created. This sample has a mean $\bar{X}_d = 85.85$ and standard deviation $s_d = 87.775$, and a sample size of $n = 20$ (there were only 20 teams that existed in 1968)

e. With the sample size of 20, the normality assumption must be checked because the central limit theorem does not apply. Using a histogram of the differences, we see a somewhat unimodal, symmetric, and bell-shaped graph. This means we have approximate normality and can use a t statistic here.

f. The t-test statistic is $t = \dfrac{85.85 - 0}{\frac{87.775}{\sqrt{20}}} \approx 4.3741$

g. Since the alternative hypothesis is $H_a\colon \mu_d > 0$, $\alpha = 0.05$, and the t-test statistic was used, the critical value that defines the rejection region is going to be the value from a t distribution with 19 degrees of freedom (sample size minus 1) that has 5% of the area to the right (right-tailed test) and 95% of the area to the left. This means the critical value is the 95th percentile of the distribution and is about 1.73.

h. Given that the test statistic is 4.3741; the value is larger than 1.73 and falls into the rejection region. Therefore, the null hypothesis is rejected in favor of the alternative hypothesis.

i. The p-value is the probability of observing a value larger than 4.3741 in a t-distribution with 19 degrees of freedom. The p-value is 0.00016 approximately.

j. There is statistical evidence that on average, teams scored more runs in 1969 then in 1968.

2-Sample Hypothesis Test: Two Independent Samples

1. Comparing population means
    h. There are now two different populations that whose means are being compared.

   i. A sample is drawn from both populations and the sample means are used to test.
   j. If the population variances are not known, then the sample variances will be used.
2. Null and Alternative Hypothesis
   a. Notation:
      - μ_1 represents the population mean for population 1
      - μ_2 represents the population mean for population 2
   b. The null hypothesis could be stated as $H_0\!:\mu_1 = \mu_2$ or as $H_0\!:\mu_1 - \mu_2 = 0$
   c. The alternative hypothesis can be stated in the following ways:
      - $H_a\!:\mu_1 > \mu_2$ or $H_a\!:\mu_1 - \mu_2 > 0$ (right-tailed test)
      - $H_a\!:\mu_1 < \mu_2$ or $H_a\!:\mu_1 - \mu_2 < 0$ (left-tailed test)
      - $H_a\!:\mu_1 \neq \mu_2$ or $H_a\!:\mu_1 - \mu_2 \neq 0$ (two-tailed test)
3. Assumptions
   a. The data from both samples are i.i.d.
   b. The populations are approximately normal, or the sample sizes are large enough for the central limit theorem to apply.
4. Test Statistics
   a. Notation:
      - Sample 1 mean: $\underline{X}_1$ and sample 2 mean: $\underline{X}_2$
      - Sample 1 variance: s_1^2 and sample 2 variance: s_2^2
      - Assumed population 1 mean: $:\mu_1$ and assumed population 2 mean: μ_2
      - Sample 1 size: n_1 and sample 2 size: n_2
      - Population 1 variance: σ_1^2 and population 1 variance: σ_2^2
   b. The test statistic for Welch's T-test is $t = \dfrac{\underline{X}_1 - \underline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$
      - This is used when population variances are unknown and sample variances cannot be assumed equal to each other.
      - This test statistic is associated with a t-distribution with degrees of freedom calculated from the following formula:

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

c. The test statistic for a Pooled T-test is $t = \dfrac{\underline{X}_1 - \underline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

- $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ is the pooled variance which is used in the pooled t-test.
- This test is used when population variances are unknown but can be assumed to be equal.
- This test statistic is associated with a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

d. The z test statistic is $z = \dfrac{\underline{X}_1 - \underline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

- This is used with both population variances are known and has the standard normal distribution.

5. Example: 2016 Major League Baseball Home Runs
   a. We are comparing the average number of home runs per team of the American League to the average number of home runs per team of the National League. Testing to see if there is a statistically significant difference between the two league averages.
   b. Data
      - American League: Sample mean is $\underline{X}_1 = 196.87$, sample standard deviation is $s_1 = 28.72$, and the sample size is $n_1 = 15$.
      - American League: Sample mean is $\underline{X}_2 = 177.15$, sample standard deviation is $s_2 = 32.85$, and the sample size is $n_2 = 15$.
   c. Null hypothesis, alternative hypothesis and level of significance are $H_0: \mu_1 = \mu_2$, $H_a: \mu_1 \neq \mu_2$, and $\alpha = 0.05$. With μ_1 being the population mean for the American League and μ_2 is the population mean for the National League
   d. Assumptions
      - The populations appear to be approximately normal using histograms (the sample sizes are too small to use the central limit theorem).
      - Data are i.i.d.
      - Box plots show that both samples have similar range, this means variances can be assumed equal.

e. Given the assumptions, a pooled t-test will be used. First, we should calculate the pooled variance.
$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(14)28.72^2 + (14)32.85^2}{15 + 15 - 2}$=951.98045

f. Test statistic:
$t = \frac{\underline{X}_1 - \underline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{196.87 - 177.15 - 0}{\sqrt{951.98045\left(\frac{1}{15} + \frac{1}{15}\right)}} \approx 1.75$

g. The degrees of freedom is $15 + 15 - 2 = 28$. This means the critical values (remember that this is a two-tailed) test will be the 2.5 percentile and 97.5 percentile of a t-distribution with 28 degrees of freedom. The critical values are approximately ± 2.05.

h. The p-value will be twice the probability of observing a value greater than 1.75. This is approximately .091

i. Since the p-value is larger than alpha and the test statistic does not fall in the rejection region; we fail to reject the null hypothesis.

6. Example: Square Footage of Old Houses vs. New Houses in Wake County, NC

a. Definitions
- Old houses are houses built before the year 1980 and shall be referred to as sample 1 with population mean μ_1.
- New houses are houses built in or after 1980 and shall be referred to as sample 2 with population mean μ_2.

b. The null hypothesis, alternative hypothesis, and level of significance are $H_0: \mu_1 = \mu_2$, $H_a: \mu_1 \neq \mu_2$, and $\alpha = 0.05$.

c. Data
- Old Houses: Sample mean is $\underline{X}_1 = 1538.87$, sample standard deviation is $s_1 = 510.382$, and the sample size is $n_1 = 45$.
- New Houses: Sample mean is $\underline{X}_2 = 1885.47$, sample standard deviation is $s_2 = 825.5401$, and the sample size is $n_2 = 55$.

d. Assumptions
- Data are i.i.d.
- Sample sizes are large enough to use the central limit theorem.
- The box plots indicate that we cannot assume the population variances are equal.

Outlier

e. Based on the previous assumptions, Welch's T-test will be used. The test statistic is
$$t = \frac{\underline{X}_1 - \underline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1538.87 - 1885.47 - 0}{\sqrt{\frac{510.382^2}{45} + \frac{825.5401^2}{55}}} \approx -2.57$$ and

the degrees of freedom for the distribution is
$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{510.382^2}{45} + \frac{825.5401^2}{55}\right)^2}{\left(\frac{1}{45 - 1}\right)\left(\frac{510.382^2}{45}\right)^2 + \left(\frac{1}{55 - 1}\right)\left(\frac{825.5401^2}{55}\right)^2} \approx 91.682$$

f. Using the degrees of freedom and that this is a two-tailed test, the critical values are ± 1.99 and the p-value is twice the probability of observing a value less than -2.57. The p-value is 0.012.

g. Conclusion: Reject the null hypothesis.

7. Comparing Population Proportions
    a. There are two populations and we are comparing the proportion of each population that meets a specific criterion.
    b. Notation that you will need to know; p_1: population 1 proportion, p_2: population 2 proportion, x_1: number of successes found in sample 1, x_2: number of successes found in sample 2, $\hat{p}_1$: sample 1 proportion, $\hat{p}_2$: sample 2 proportion, n_1: sample 1 size, n_2: sample 2 size, and $p_c = \frac{x_1 + x_2}{n_1 + n_2}$: proportion of successes across both samples.
    c. The null hypothesis can be expressed as $H_0: p_1 = p_2$ or $H_0: p_1 - p_2 = 0$.
    d. Alternative Hypothesis
        - $H_a: p_1 > p_2$ or $H_a: p_1 - p_2 > 0$ (right-tailed test)
        - $H_a: p_1 < p_2$ or $H_a: p_1 - p_2 < 0$ (left-tailed test)
        - $H_a: p_1 \neq p_2$ or $H_a: p_1 - p_2 \neq 0$ (two-tailed test)
    e. Assumptions
        - Data are i.i.d.
        - There must be at least 5 successes and 5 failures in both samples
        - The population must be at least 10 times the size of the sample
    f. The test statistic is $z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{p_c(1 - p_c)(\frac{1}{n_1} + \frac{1}{n_2})}}$.

g. The p-value and critical values will come from a standard normal distribution.

8. Example: Proportion of Old vs. New Houses with Fireplaces in Wake County, NC
   a. This example will use the same definitions for old and new houses as the previous examples.
   b. Notation:
      - Old houses will represent population and sample 1. The population proportion of old houses that have a fireplace will be called p_1, and x_1 represents the number of old houses with fireplaces.
      - New houses will represent population and sample 2. The population proportion of new houses that have a fireplace will be called p_2 and x_2 represents the number of new houses with fireplaces.
   c. The null hypothesis, alternative hypothesis, and level of significance are $H_0: p_1 = p_2$, $H_a: p_1 < p_2$, and $\alpha = 0.05$
   d. Data