Statistics 301: Probability and Statistics

2-sample Methods

Module 10

2018

Comparing two groups

Comparisons:

(1) Two independent means
  (a) When $\sigma_1^2 \approx \sigma_2^2$: Pooled
  (b) When $\sigma_1^2 \neq \sigma_2^2$: Unpooled (also called a Welch or Satterthwaite test)
(2) Dependent means
(3) Two proportions (independent)

Use of t when comparing two means

While we could know the population standard deviations (σ_1, σ_2), that rarely happens, and usually in practice we most often just use t. So while the textbook covers using known σ_1 and σ^2, we will not cover that and *only use t*

Independent means

This compares the means of two distinct (separate) groups of units or subjects. The wording used is **the difference of two (2) means**

While there are two cases for this (when variances are equal or unequal), we will only use the unequal variances (unpooled) method. If the two variances are unequal or equal, the unpooled is appropriate in either case.

Formula: df for use of t

Degrees of freedom for independent means (unpooled – the one we are using) is calculated rather than using $n-1$ or something similar:

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

We will use:

$$df = min(n_1 - 1, n_2 - 1)$$

Formula: CI

CI for the difference of two (independent) means:

$$\overline{X}_1 - \overline{X}_2 \pm t^{\star}(se) \ where \ se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \ and \ t^{\star} = t_{\alpha/2,df}$$

$$df = min(n_1 - 1, n_2 - 1)$$

Hypotheses

For the difference of two (independent) means[1]:

$$H_0 : \mu_1 = \mu_2 \quad H_a : \mu_1 \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} \mu_2$$

$$Or$$

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} 0$$

Formula: Test Statistic

$$t = \frac{\overline{X}_1 - \overline{X}_2}{se} \quad where \ se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

With

$$df = min(n_1 - 1, n_2 - 1)$$

Assumptions

(1) Independence (if random met, this is met)
(2) Randomization
(3) Each group of observations have approximate normal distribution

Dependent means

This compares the mean of the difference between two measurements of the same unit or subject. The wording used is **the mean difference**. This analysis is for comparing measurements on the same subject/unit; usually once before a treatment and once again after the treatment, to detect if there is a difference due to the treatment.

Examples are weight loss programs, Coke vs. Pepsi, compare GDP of countries at 2 different dates (time is treatment)

[1] In practice, you could test the difference of means equal to a value other than zero

Formula: CI

d_i: individual differences between measurements

$\overline{X}_d = \frac{\sum d_i}{n}$ sample mean difference (mean of the differences)

$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$: sample standard deviation of the differences

CI for the mean difference:

$$\overline{X}_d \pm t^\star(se) \;\; where \; se = \frac{s_d}{\sqrt{n}} \;\; and \;\; t^\star = t_{\alpha/2,df}, \; df = n-1$$

Hypotheses

For the mean difference[2]:

$$H_0 : \mu_d = 0 \;\; H_a : \mu_d \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} 0$$

Formula: Test Statistic

$$t = \frac{\overline{X}_d - 0}{se} \text{ where } se = \frac{s_d}{\sqrt{n}}$$

Assumptions

(1) Dependence (two measurements per unit/subject)
(2) Randomization
(3) Differences have approximate normal distribution

Two Proportions

This compares the means of two distinct (separate) groups of units or subjects. The wording used is **the difference of two (2) proportions**

The *se* for the test is different from the *se* for the CI

Assumptions

(1) Independent groups (if random met, this is met)
(2) Randomization
(3) success/failure condition to have normality
    (a) either $n_1 \geq 60$ AND $n_2 \geq 60$ or
    (b) $n_1p_1 \geq 5$, $n_1q_1 \geq 5$, $n_2p_2 \geq 5$, **AND** $n_2q_2 \geq 5$

[2] In practice, you could test the mean difference equal to a value other than zero. This concept could be a supplemental module

Formula: CI

CI for the difference of two (independent) means:

$$\hat{p}_1 - \hat{p}_2 \pm z^\star(se) \text{ where } se = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \text{ and } z^\star = z_{\alpha/2}$$

Hypotheses

For the difference of two (independent) proportions[3]:

$$H_0 : p_1 = p_2 \quad H_a : p_1 \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} p_2$$

Or

$$H_0 : p_1 - p_2 = 0 \quad H_a : p_1 - p_2 \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} 0$$

Formula: Test Statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{se} \text{ where } se = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Where $\hat{p}\hat{q}$ without subscripts is the pooled proportion used when assuming the difference of proportions is equal to 0.

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

X_1, X_2 are the successes from each group. If you are given percents, then you will have to calculate the successes by:

$$X_1 = n_1\hat{p}_1 \quad X_2 = n_2\hat{p}_2$$

Rejection regions

Thankfully, they are the same as for 1-sample methods. Make sure to still follow the 4 steps to hypothesis testing:

(1) Hypotheses
  - assumptions (if asked for)

(2) Test statistic
(3) Rejection region (use either critical value approach or *pvalue* approach)
(4) Results and conclusion in context

[3] In practice, you could test the difference of proportions equal to a value other than zero. This concept could be a supplemental module

2 independent means

Some archaeologists theorize that ancient Egyptians interbred with several different immigrant populations over thousands of years. To see if there is any indication of changes in body structure that might have resulted, they measured 30 skulls of male Egyptians dated from 4000 BCE and 30 others dated from 200 BCE. Using the provided summary statistics, is there sufficient evidence that the mean breadth of males' skulls changed over this period? Estimate the true difference of means with 95% confidence and interpret; conduct hypothesis test.

```
        meanse    sdse
200BCE  135.633 4.03846
4000BCE 131.367 5.12925
```

2 independent proportions

Sludge is a dried product remaining from processed sewage and is often used as a fertilizer on crops, but the nickel may be at a dangerous concentration in the crops. A new method of processing sewage has been developed and an experiment conducted to evaluate its effectiveness in removing heavy metals. Sewage of a known concentration of nickel is treated using both old and new methods. One hundred tomato plants were randomly assigned to pots containing sewage sludge processed by one of the two methods and the nickel was measured in the tomatoes. Is there sufficient evidence that the new treatment has a lower concentration of nickel than the old treatment?

2 proportions con't

Estimate the true difference of proportions with 95% confidence and interpret; conduct hypothesis test.

```
      Toxic Non-toxic Total
New       5        45    50
Old       9        41    50
Total    14        86   100
```

Dependent means

Trace metals in drinking water afftect the flavor; high concentrations can be a health hazard. A randomized study looked at six river locations along the South Indian River (6 units) and the zinc concentration in mg/L was measured for both surface and bottom water at each location. Is there sufficient evidence the true mean concentration in bottom water exceeds that of surface water? Estimate the true mean difference with 90% confidence and interpret; conduct hypothesis test, let $\alpha = 0.10$.

Dependent means con't

```
               1     2     3     4     5     6
Bottom     0.430 0.266 0.567 0.531 0.707 0.716
Surface    0.415 0.238 0.390 0.410 0.605 0.609
Difference 0.015 0.028 0.177 0.121 0.102 0.107
```

```
               means       sds
Bottom     0.5361667 0.1713259
Surface    0.4445000 0.1417699
Difference 0.0916667 0.0606883
```

2 independent means: test

$$H_0 : \mu_{200BCE} = \mu_{4000BCE} \text{ vs. } H_a : \mu_{200BCE} \neq \mu_{4000BCE}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{se} \text{ with } se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

2 independent means: test con't

$$t = \frac{135.63 - 131.37}{1.192} = 3.574$$

Reject H_0 if $|t_{calc}| \geq |t_{\alpha/2,df}|$ with $df = \frac{\left(\frac{4.038^2}{30} + \frac{5.129^2}{30}\right)^2}{\frac{(4.038^2/30)^2}{30-1} + \frac{(5.129^2/30)^2}{30-1}} = 54.973 \approx 55$, $t_{\alpha/2,df} = t_{.05/2,55} = 2.004$ Since $|3.574| \geq 2.004$, H_0 is rejected. There is a significant difference in skull breadths of the two time periods.

2 independent means: CI

$$\overline{X}_1 - \overline{X}_2 \pm t^{\star}(se)$$

$t^{\star} = t_{\alpha/2,df} = t_{.05/2,55} = 2.004$ and $se = 1.192$

$$135.63 - 131.37 \pm (2.004)(1.192) = 4.26 \pm 2.39 = 1.87, 6.65$$

We are 95% confident the true difference of mean skull breadths of Egyptians from 200 BCE and 4000 BCE is between 1.89 and 6.66 mm.

2 proportions: test

$$H_0 : p_{new} = p_{old} \text{ vs. } H_a : p_{new} < p_{old}$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{se} \text{ with } se = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ and } \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{5 + 9}{50 + 50} = 0.14 \text{ , } \hat{q} = 1 - 0.14 = 0.86$$

2 proportions: test con't

$$\hat{p}_1 = \frac{5}{50} = 0.1 \text{ ; } \hat{p}_2 = \frac{9}{50} = 0.18$$

$$se = \sqrt{(0.14)(0.86)\left(\frac{1}{50} + \frac{1}{50}\right)} = 0.0694$$

$$z = \frac{0.1 - 0.18}{0.0694} = -1.15$$

Reject H_0 if $pvalue \leq \alpha$. $H_a :<$, $pvalue = P(Z < z_{calc}) = P(Z < -1.15) = 0.1251$.

2 proportions: test con't (pt 3)

$0.1251 \nleq 0.05$, H_0 is not rejected. The new method is not significantly different; the proportions of toxic plants treated with both new and old methods is not significantly different.

2 proportions: CI

$$\hat{p}_1 - \hat{p}_2 \pm z^\star(se)$$

$z^\star = z_{\alpha/2} = z_{.05/2} = 1.96$ and $se = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} = \sqrt{\frac{(0.1)(0.9)}{50} + \frac{(0.18)(0.82)}{50}} = 0.0689$

$$0.1 - 0.18 \pm (1.96)(0.0689) = -0.08 \pm 0.1350$$

$$= -0.215, 0.055$$

2 proportions: CI interpretation

We are 95% confident the true difference of proportions of toxic plants treated with new and old methods is between -21.5% and 5.5%. Since the interval contains 0 (which is the hypothesized difference $p_1 = p_2$ implies that $p_1 - p_2 = 0$), there is no difference between the methods.

Dependent means: test part 1

$$H_o : \mu_d = 0 \text{ vs. } H_a : \mu_d > 0$$

Since we are doing an upper tail test, and we want to see if bottom water is more than surface water, then we shoulld make sure the difference is calculated as: $d_i = bottom - surface$ where d_i are the individual differences between measurements.

$$se = \frac{s_d}{\sqrt{n}} = \frac{0.0607}{\sqrt{6}} = 0.0248$$

$$t = \frac{\overline{X}_d - 0}{se} = \frac{0.0917}{0.0248} = 3.699$$

Dependent means: test con't

Reject H_0 iff $t_{calc} \geq t_{\alpha,df} t_{0.1,5} = 1.476$.

Since $3.699 \geq 1.476$, we reject H_0. The zinc concentration is significantly higher in the bottom water than in the surface water.

Dependent means: CI

$$\overline{X}_d \pm t^{\star}(se) = 0.0917 \pm (2.015)(0.0248)$$

$$= 0.0917 \pm 0.05 = 0.0417, 0.1417$$

We are 90% confident the true mean difference in zinc concentrations of bottom water vs. surface water is between 0.0417 and 0.1417 mg/L.