Stat 20: Discussion at section

B. M. Bolstad, bolstad@stat.berkeley.edu

Nov 3, 2003

Two sample normal model

For the two sample normal model we assume that we have a sample $Y_{11}, \ldots, Y_{n_1 1}$ of independent variables from a normal distribution with mean μ_1 and standard deviation σ_1 and a sample $Y_{12}, \ldots, Y_{n_2 2}$ of independent variables from a normal distribution with mean μ_2 and standard deviation σ_2. In addition we assume each of the two samples are independent. The homoskedastic two sample normal model further assumes that $\sigma_1 = \sigma_2 = \sigma$. Our sample estimates of μ_1 and σ_1 are given by sample mean and standard deviations

$$\bar{Y}_1 = \frac{\sum_{i=1}^{n_1} Y_{i1}}{n_1}$$

and

$$s_1 = \sqrt{\frac{\sum_{i=1}^{n_1} \left(Y_{i1} - \bar{Y}_1\right)}{n_1 - 1}} = \sqrt{\frac{\sum_{i=1}^{n_1} Y_{i1}^2 - n_1 \bar{Y}_1^2}{n_1 - 1}}$$

and the sample estimates of μ_2 and σ_2 are given by sample mean and standard deviations

$$\bar{Y}_2 = \frac{\sum_{i=1}^{n_2} Y_{i2}}{n_2}$$

and

$$s_2 = \sqrt{\frac{\sum_{i=1}^{n_2} \left(Y_{i2} - \bar{Y}_2\right)}{n_2 - 1}} = \sqrt{\frac{\sum_{i=1}^{n_2} Y_{i2}^2 - n_2 \bar{Y}_2^2}{n_2 - 1}}$$

For the homoskedastic model we use the pooled sample standard deviation

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Common tests

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_A : \mu_1 \neq \mu_2$$

One sided alternatives

$$H_0 : \mu_1 \leq \mu_2 \text{ vs } H_A : \mu_1 > \mu_2$$

or

$$H_0 : \mu_1 \geq \mu_2 \text{ vs } H_A : \mu_1 < \mu_2$$

Test statistic

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

under the null distribution this test statistic will have t distribution with $n_1 + n_2 - 2$ degrees of freedom. Note that $\text{SE}\left(\bar{Y}_1 - \bar{Y}_2\right) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

Confidence interval for the difference between two means

A 100C% confidence interval for the difference $\mu_1 - \mu_2$ is given by

$$\bar{Y}_1 - \bar{Y}_2 \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where t^* is given by the value satisfying

$$P\left(-t^* < T < t^*\right) = C$$

where T has t-distribution with $n - 1$ degrees of freedom.

Using the t distribution table

The t table gives the t^* values which satisfy $P(T > t^*) = p$ for fixed degrees of freedom and known upper tail probabilities p.

Confidence intervals

If you desire a 100C% confidence interval then the upper tail probability is given by $\frac{1-C}{2}$.

Examples: Suppose df = 8, then the t^* for a 95% confidence interval is given by 2.306. Suppose df=13 then the t^* for a 99% confidence interval is given by 3.012.

Hypothesis tests

There are fewer P-values in a t-table then in the normal distribution table we have used earlier. There are three options we can use

1. Pick fixed significance level for the test and then find the critical value to compare with your computed t-statistic.

2. Bracket (Bound) your P-value. In particular we find t_{left}, with upper tail probability p_{left}, and t_{right}, with upper tail probability p_{right}, from the table such that $t_{\text{left}} < t < t_{\text{right}}$. We may then conclude that $p_{\text{left}} >$ P-value $> p_{\text{right}}$

3. Use linear interpolation to approximate the P-value. This will be given by

$$p_{\text{left}} + \frac{t - t_{\text{left}}}{t_{\text{right}} - t_{\text{left}}} (p_{\text{right}} - p_{\text{left}})$$

It is recommended that you know how to do method 1 and 2.

Question 1

A study of iron deficiency among infants compared different feeding regimens. One group contained breast-fed infants, while the other group received only a standard baby formula without additional iron supplements. After six months the hemoglobin levels in the infants blood were measured. The following table gives us summary statistics for the study:

Group	n	$\bar{Y}$	s
Breast Fed	23	13.3	1.7
Formula	19	12.4	1.8

Is there evidence that the hemoglobin levels are higher in breast fed babies? Carry out the appropriate test. Then give the 95% confidence for the difference in haemoglobin levels between the two groups.

Answer:

Let μ_1 be the mean hemoglobin level in breast fed infants and let μ_2 be the mean hemoglobin level in formula fed infants. The appropriate hypothesis test is

$$H_0 : \mu_1 \leq \mu_2 \text{ vs } H_A : \mu_1 > \mu_2$$

The pooled sample standard deviation is given by

$$s_p = \sqrt{\frac{(23-1)1.7^2 + (19-1)1.8^2}{23+19-2}} = 1.746$$

and so our t-statistic will be

$$t = \frac{13.3 - 12.4}{1.746\sqrt{\frac{1}{23} + \frac{1}{19}}} = 1.6627$$

note that the degrees of freedom are $23 + 19 - 2 = 40$

Method 1 - Fixed significance level

Note that

$$P(T > 1.684) = .05$$

and

$$P(T > 2.423) = .01$$

At the 5% level of significance: We would reject the null hypothesis if t was greater than 1.684 and could not reject the null hypothesis if $t < 1.684$.

At the 1% level of significance: We would reject the null hypothesis if t was greater than 2.423 and could not reject the null hypothesis if $t < 2.423$.

So we cannot reject the null hypothesis at the 5% level of significance andwe can not reject the null if we carry out the test at the 1% level of significance.

Method 2 - Bounding P-value

First we note that

$$1.303 < 1.6627 < 1.684$$

converting to P-values

$$0.1 > P(T > 1.6627) > 0.05$$

and so we may conclude that $0.05 <$ P-value < 0.1 which is not evidence to reject the null hypothesis.

Method 3 - Linear interpolation to approximate P-value

$$P(T > 1.6627) \approx .10 + (1.6627 - 1.303)/(1.684 - 1.303) * (.05 - .10) = 0.0527$$

Confidence interval

The 95% confidence interval is given by

$$13.3 - 12.4 \pm 2.021(1.746)\sqrt{\frac{1}{23} + \frac{1}{19}}$$

and so the interval is $(-0.19, 1.99)$ which does contain 0.

Question 2

Nitrites are often added to meat products as preservatives. A study of the effect of these chemicals on bacteria was carried out. The rate of uptake of a radio labeled amino acid was measured for a number of cultures of bacteria, some growing in a medium to which nitrites were then added. Here are some summary statistics for the the study:

Group	n	$\bar{Y}$	s
Nitrite	30	7480	1115
Control	30	8118	1250

Do we have evidence that nitrites decrease amino acid uptake?

Answer:

Let μ_1 be the mean amino acid uptake for nitrite treated cultures and let μ_2 be the mean amino acid uptake for control cultures. The appropriate hypothesis test is

$$H_0 : \mu_1 \geq \mu_2 \text{ vs } H_A : \mu_1 < \mu_2$$

The pooled sample standard deviation is given by

$$s_p = \sqrt{\frac{(30-1)1115^2 + (30-1)1250^2}{30+30-2}} = 1184.425$$

and so our t-statistic will be

$$t = \frac{7480 - 8118}{1184.425\sqrt{\frac{1}{30} + \frac{1}{30}}} = -2.0886$$

note that the degrees of freedom are $30 + 30 - 2 = 58$

Method 1 - Fixed significance level

Note that

$$P(T < -1.676) = P(T > 1.676) = .05$$

and

$$P(T < -2.403) = P(T > 2.403) = .01$$

At the 5% level of significance: We would reject the null hypothesis if t was less than -1.676 and could not reject the null hypothesis if $t > -1.676$.

At the 1% level of significance: We would reject the null hypothesis if t was less than -2.403 and could not reject the null hypothesis if $t > -2.403$.

So we can reject the Null hypothesis at the 5% level of significance, but we can not reject the null if we carry out the test at the 1% level of significance.

Method 2 - Bounding P-value

First we note that

$$-2.109 < -2.0886 < -2.009$$

and so

$$2.009 < 2.0886 < 2.109$$

converting to P-values

$$0.025 > P(T > 2.0886) > 0.02$$

and so we may conclude that $0.02 <$ P-value < 0.025 which is evidence to reject the null hypothesis.

Method 3 - Linear interpolation to approximate P-value

$P(T < -2.0886) = P(T > 2.0886) \approx .025 + (2.0886 - 2.009)/(2.109 - 2.009) * (.02 - .025) = 0.0210$