STATS 2107
Statistical Modelling and Inference II
Lecture notes
Chapter 2: Distributions from the normal

Jono Tuke

School of Mathematical Sciences, University of Adelaide

Semester 2 2017

Distributions from the normal

Lemma

Suppose that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$, and let

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}},$$

then

$$Z \sim N(0, 1).$$

Case where σ^2 is not known

We use the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

in its place.

The **key distributional result** used in inference is

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Theorem

Suppose $Y_1, Y_2, \ldots, Y_n$ are independent random variables with

$$E[Y_i] = \mu \text{ and } var(Y_i) = \sigma^2,$$

and let S^2 be defined as above, then

$$E[S^2] = \sigma^2.$$

Proof:

Lemma

Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$, then $\bar{Y}$ and S^2 are independent.

Proof

Definition

Suppose $Z_1, Z_2, \ldots, Z_p$ are i.i.d. $N(0,1)$ random variable, then the random variable

$$X = \sum_{i=1}^{p} Z_i^2$$

is said to have the χ^2 distribution with p degrees of freedom and we write

$$X \sim \chi_p^2.$$

Distributions from the normal part 2

Theorem

Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$, then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Proof

Definition

Suppose $Z \sim N(0,1)$ and $X \sim \chi_p^2$ independently, and let

$$T = \frac{Z}{\sqrt{X/p}},$$

then T is said to have a t distribution with p degrees of freedom and we write

$$T \sim t_p.$$

Theorem

Suppose that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$, then

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Definition

Let W_1 and W_2 be independent χ^2 distributed random variables with ν_1 and ν_2 degrees of freedom respectively, then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

is said to have an F distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom.

One-sample T-test

Setup

Suppose that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables with σ^2 are known.

- The BLUE for μ is $\bar{Y}$.
- The estimated standard error for $\bar{Y}$ is $S/\sqrt{n}$.

Hypothesis test

To test

$$H_0 : \mu = \mu_0,$$
$$H_a : \mu \neq \mu_0,$$

the test statistic is

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}.$$

We reject H_0 iff

$$|t_{obs}| \geq t_{n-1,\alpha/2}$$

This procedure has a significance level of α.

Proof:

Confidence interval for μ

The confidence interval

$$\left(\bar{Y} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{Y} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for μ.

Proof:

Inference for σ^2

Suppose that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$. Let c_1, c_2 be such that

$$P(c_1 < X < c_2) = 1 - \alpha,$$

where

$$X \sim \chi^2_{n-1},$$

then

$$\left(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1} \right)$$

is a $100(1-\alpha)\%$ confidence interval for σ^2.

Choices for c_1, c_2.

- Symmetric: $P(X < c_1) = \alpha/2$ and $P(X > c_2) = \alpha/2$.
- Lower bound: $c_1 = 0$ and $P(X > c_2) = \alpha$.
- Upper bound: $P(X < c_1) = \alpha$ and $c_2 = \infty$.

Definition

A random variable

$$H = H(Y_1, Y_2, \ldots, Y_n, \theta)$$

with a known distribution that does not depend on θ is called a **pivotal quantity**.

Two-sample t-test - pooled

Setup

Consider independent random variables

$$Y_{ij},\ \ i = 1, 2;\ j = 1, 2, \ldots, n_i,$$

such that

$$Y_{ij} \sim N(\mu_i, \sigma^2).$$

Estimation of $\mu_1 - \mu_2$

Let

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \text{for } i = 1, 2,$$

then

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right).$$

Estimation of σ^2

The pooled estimator

$$S_p = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

where

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

is an unbiased estimator of σ^2, and

$$\frac{(n_1 + n_2 - 2)S_P^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

Pooled two-sample t-test

As S_p^2 is independent of $\bar{Y}_i$ $i = 1, 2$, it follows that

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

From this we can get the hypothesis test and confidence interval.

Hypothesis test

To test

$$H_0 : \mu_1 - \mu_2 = 0,$$
$$H_0 : \mu_1 - \mu_2 \neq 0,$$

the test statistic is

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

We reject H_0 iff

$$|t_{obs}| \geq t_{n_1+n_2-2,\alpha/2}.$$

Confidence interval for $\mu_1 - \mu_2$

The interval

$$\left(\bar{Y}_1 - \bar{Y}_2 - t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{Y}_1 - \bar{Y}_2 + t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \cdots} \right.$$

is a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$.

Two-sample t-test - not pooled

Setup

Consider independent random variables

$$Y_{ij}, \ i = 1, 2; \ j = 1, 2, \ldots, n_i,$$

such that

$$Y_{ij} \sim N(\mu_i, \sigma_i^2),$$

where $\sigma_1^2 \neq \sigma_2^2$.

Estimation of $\mu_1 - \mu_2$

Let

$$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}, \text{for } i = 1, 2,$$

then

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Estimation of σ_i^2

We can use

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2,$$

which is an unbiased estimator of σ_i^2, but what about

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}?$$

Test statistic

We could use the test statistic

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_i^2}{n_1} + \frac{S_2^2}{n_2}}},$$

but T does not have a t_k-distribution for any value of k.

Approximate t-distribution

Instead, we choose a t_k distribution that approximates the true distribution of T.

Method 1

Choose

$$k = \min(n_1 - 1, n_2 - 1).$$

Method 2

Use

$$k = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Pooled versus not-pooled

'Rule of thumb'

Use a pooled two-sample t-test if

$$\frac{\max(s_1, s_2)}{\min(s_1, s_2)} < 2.$$