3Y03 - Probability and Statistics for Engineering

WS19 Lecture 29

Last Time C.I.s/Hypothesis Tests on the Difference of 2 Means of a Normal Population, Variance Unknown

$H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$. H_1: (I) $\mu_1 - \mu_2 \neq 0$; (II) $\mu_1 - \mu_2 > 0$; (III) $\mu_1 - \mu_2 < 0$.

2 CASES: 1/ $\sigma_1^2 = \sigma_2^2$ 2/ $\sigma_1^2 \neq \sigma_2^2$

TEST STAT.

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ - distribution}$$

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim t_\nu \text{ - distr.}$$

100(1-α)% C.I.

$$\bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$\bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}, \nu}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$ ← Pooled Sample Variance

$$\nu = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}$$ (Round DOWN ↓.)

Example Two extrusion machines produce steel rods, diameter Normally distributed.

Sample of 15 ($=n_1$) rods from machine 1 has $\bar{x}_1 = 8.73$, $s_1^2 = 0.35$

" " 17 ($=n_2$) " " " 2 " $\bar{x}_2 = 8.68$, $s_2^2 = 0.40$

Test the claim that the machines produce with different mean diameters.

Use $\alpha = 0.05$ [& do NOT assume variances are equal].

2-sided $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$

Solution

$$t_0^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = \frac{8.73 - 8.68}{\sqrt{\frac{0.35}{15}+\frac{0.4}{17}}} = 0.231$$ ← test statistic value

Now find

$$\nu = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}} = \frac{\left(\frac{0.35}{15}+\frac{0.4}{17}\right)^2}{\frac{(0.35/15)^2}{14}+\frac{(0.4/17)^2}{16}} = 29.88 \rightarrow \nu = 29$$

Look up $t_{\frac{\alpha}{2},\nu} = t_{0.025,29} = 2.045$. ← critical value

Compare test stat. value $0.231 < 2.045$ critical value, so do not reject H_0.

Example Find a 95% C.I. for $\mu_1 - \mu_2$ in previous setup — but now you can assume variances equal.

Solution Given by $\bar{x}_1 - \bar{x}_2 \pm t_{\frac{\alpha}{2}, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{14(0.35) + 16(0.4)}{15+17-2}$$

$$= 0.37666\ldots$$

So $s_p = 0.613\ldots$

So our C.I. is $(8.73-8.68) \pm \underbrace{t_{0.025,30}}_{2.042}(0.613\ldots)\sqrt{\frac{1}{15}+\frac{1}{17}}$

i.e. 0.05 ± 0.444

i.e. $(-0.394, 0.494)$.

The null value under H_0 of $\mu_1-\mu_2$. ↑

Notice 0 is in this C.I. So could ~~also~~ test $H_0 : \mu_1 = \mu_2$ when variances equal by checking $H_1 : \mu_1 \neq \mu_2$ if 0 is in C.I. → here yes, so do not reject H_0.

In general, to test $H_0 : \theta = \theta_0$ v. $H_1 : \theta \neq \theta_0$ at level α, you can use a $100(1-\alpha)\%$ C.I. for θ & check if θ_0 is in the C.I. (if yes, accept H_0).

11 Linear Regression

X, Y — assume a linear relationship

— want to understand this using data

We model it with $Y = \beta_0 + \beta_1 X + \varepsilon$

- β_0: intercept
- β_1: slope
- β_0, β_1: regression coefficients
- ε: error term "random noise" with $E(\varepsilon) = 0$

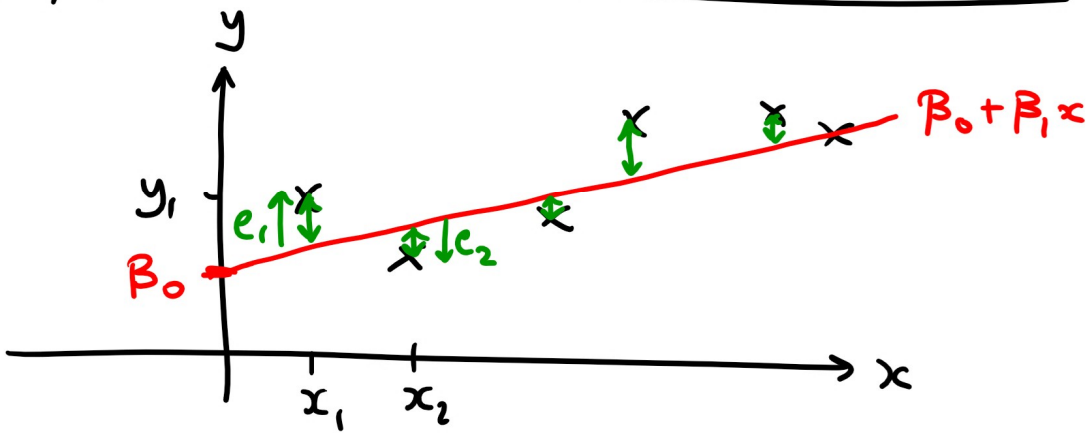
i.e. we expect on average a linear relationship, so we model that with a linear relationship + error, where average error is zero.

So $E(Y) = E(\beta_0 + \beta_1 X + \varepsilon)$

$= \beta_0 + \beta_1 E(X)$

Given data, how to find β_0, β_1?

Scatter plot:



n pairs of observations x (x_i, y_i)

$y_i = \beta_0 + \beta_1 x_i + e_i$ ← residual

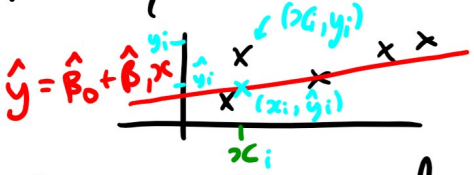
If we think of the model point by point as $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, then e_i is the realization of ε_i.

Want line $y = \beta_0 + \beta_1 x$ "best fit" i.e. minimize distances of error

So minimize $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

The resulting estimates for β_0, β_1 are called the **least squares estimates** $\hat{\beta}_0, \hat{\beta}_1$

The resulting best fit line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is the **least squares regression line**.



(Allows us to estimate y-value given an x-value.)

How do we find $\hat{\beta}_0, \hat{\beta}_1$?

"Different values of β_0, β_1" is another way of saying "Different choices of line"!

Different values of β_0, β_1 give different values of

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

← want to minimize

↖ "Different values of L" = "different amounts of (square) error".

Think of β_0, β_1 as variables for now.

At minimum of L, we must have $\frac{\partial L}{\partial \beta_0} = 0 = \frac{\partial L}{\partial \beta_1}$

Get 2 equations in 2 unknowns (β_0, β_1); solving:

*see below

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$\sum_{i=1}^{n} x_i = n\bar{x}$, $\sum_{i=1}^{n} y_i = n\bar{y}$, $\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = \bar{x} \cdot (n\bar{x}) = n\bar{x}^2$

& $\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{\hat{\beta}_1}{n}\sum_{i=1}^{n} x_i = \bar{y} - \hat{\beta}_1 \bar{x}$

$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$. The values $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values of β_0 and β_1 for which $\frac{\partial L}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = 0 = \frac{\partial L}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1)$

So $\frac{\partial L}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = 0$ becomes $\underbrace{-2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}_{\text{so this} = 0 \text{ (i.e. forget the } -2)} = 0$

(rearranging) $\Rightarrow$ $\boxed{\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i}$

and $\frac{\partial L}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = 0$ becomes $\underbrace{-2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}_{\text{i.e. this} = 0 \text{ (forget the } -2 \text{ again)}} = 0$

(rearranging) $\Rightarrow$ $\boxed{\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2}$

The two equations in boxes are called the "Least Squares Normal Equations." Solving these gives the formulas for $\hat{\beta}_1$ first, then $\hat{\beta}_0$ in terms of $\hat{\beta}_1$ as given in class.