3Y03 - 3J04 PROBABILITY & STATISTICS FOR (CIVIL) ENGINEERING

C01 - Lecture 34

Last time CORRELATION ρ

estimated by

$$R = \frac{\overbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}^{S_{xy}}}{\sqrt{\underbrace{\sum (x_i - \bar{x})^2}_{S_{xx}} \underbrace{\sum (Y_i - \bar{Y})^2}_{SS_T}}}$$

COEFFICIENT OF DETERMINATION

$$R^2 = \frac{SS_R}{SS_T}$$

$\in [0,1]$

Big = linear relationship

$$R^2 = \frac{S_{xy}^2}{S_{xx} SS_T} = \frac{S_{xy}}{S_{xx}} \cdot \frac{S_{xy}}{SS_T} = \hat{\beta}_1 \cdot \frac{S_{xy}}{SS_T} = \frac{SS_R}{SS_T}.$$

coeff. of determination

Example For our running example, find R, the correlation coefficient.

Solution

$$R = \frac{S_{xy}}{\sqrt{S_{xx} SS_T}} = \frac{-16}{\sqrt{(32.8)(10)}} = \underline{\underline{-0.88}}$$

We see that X, Y are strongly negatively correlated.

(**Note**: $R = \pm\sqrt{R^2} = \pm\sqrt{0.784} = \pm 0.88$ but need extra info. e.g. $-$ve slope to know $R = -0.88$ from R^2.)

we found this last lecture.

13.2 Single Factor Experiments and Analysis of Variance (ANOVA)

The influence of 1 factor on a response is being investigated & tested at different 'levels' called "treatments".

e.g.

Response	Treatments
• Compression strength of concrete	Different mixing techniques
• Midterm grades	Types of studying (e.g. last-minute cramming v. distributed studying)
• Ability to stay awake	Different amounts of caffeine consumed

- Visual performance of LCD screens — Different screen luminescence levels
- Batting performance in baseball — Different BMI (body mass indices)

<u>Factor</u> : the single variable in each case

Observations grouped by "treatment" ...

a = # treatments

n_i = # observations at treatment level i

$i = 1, \ldots, a$.

For ith treatment get a random sample

↳ the random variables are $Y_{i1}, \ldots, Y_{in_i}$

↳ Y_{ij} is the <u>response</u> from observation j in treatment i

We model $Y_{ij} = \mu_i + \varepsilon_{ij}$

μ_i: mean at ith treatment

ε_{ij}: random error dependent on treatment

↳ usually think of this as $\mu + \tau_i$

μ: overall mean

τ_i: "ith treatment effect"

We assume $\varepsilon_{ij} \sim N(0, \sigma_i^2)$; then

$$Y_{ij} \sim N(\mu_i, \sigma_i^2)$$ for $j = 1, \dots, i$
$i = 1, \dots, a$

We want to compare means μ_i.

We've already looked at this with 2 treatments.

10-2 : Tests & C.I.s for difference of 2 means of Normal pop. variances unknown. (& not equal ?)

↳ the point is, nothing I've said here about the setup so far lets conclude they are equal, so for now we'll not assume it.

We had $H_0 : \mu_1 = \mu_2$
H_1 : say $\mu_1 \neq \mu_2$

We used t-test (whether assuming $\sigma_1^2 = \sigma_2^2$ or not).

Now we'll have $H_0 : \mu_1 = \mu_2 = \dots = \mu_a$
H_1 : at least one pair differs (some $\mu_i \neq \mu_k$)

One option : pair up. Do a t-test for each pair:

$H_0 : \mu_i = \mu_k$
H_1 : say $\mu_i \neq \mu_k$

for each pair (i, k) with $i \neq k$.

We'll have test stat $t_{ik} = \dfrac{\bar{x}_i - \bar{x}_k}{\sqrt{\dfrac{s_i^2}{n_i} + \dfrac{s_k^2}{n_k}}}$.

(in case we <u>don't</u> assume $\sigma_i^2 = \sigma_k^2$)
(as noted in blue above)

We'll have $\binom{a}{2}$ tests.

At significance level α, prob. of incorrectly concluding a difference between some fixed pair of means μ_i and μ_k is α (= P(Type I error))

So EER = Experiment-wise Error Rate

= P(at least 1 Type I error across all $\binom{a}{2}$ tests)

= 1 − P(no type I errors)

$= 1 - (1-\alpha)^{\binom{a}{2}}$

If $\alpha = 0.05$, and we have 6 treatments to test, we have $\binom{6}{2}$ pairs of means to test, so

$= \dfrac{6!}{4!\,2!} = 15$ tests!!!

$$EER = P(\text{at least 1 type I error})$$
$$= 1 - (0.95)^{15} = \underline{\underline{0.54}}$$

$>50\%$ chance of a Type I error somewhere!! Oh dear!

This is NOT a good approach !!!

We need one test that will compare <u>all</u> means $\mu_1, \ldots, \mu_a$ simultaneously.

This is where ANOVA comes into its own.
We will use an F - test.

For simplicity, for now we assume $n_i = n$ for all i
(i.e. same # observations for each treatment)

We also assume <u>completely</u> <u>randomized</u> <u>experimental</u> <u>design</u>. i.e. the samples are tested across all treatments in random order & with uniform environment i.e. the $(n \times a)$-many Y_{ij} s are tested in random order

This simplifies our model to

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \text{where}$$

$i = 1, \ldots, a$ $\quad$ $j = 1, \ldots, n$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

i.e. assume errors have same underlying distribution.

(which, from above, we are assuming is normal).