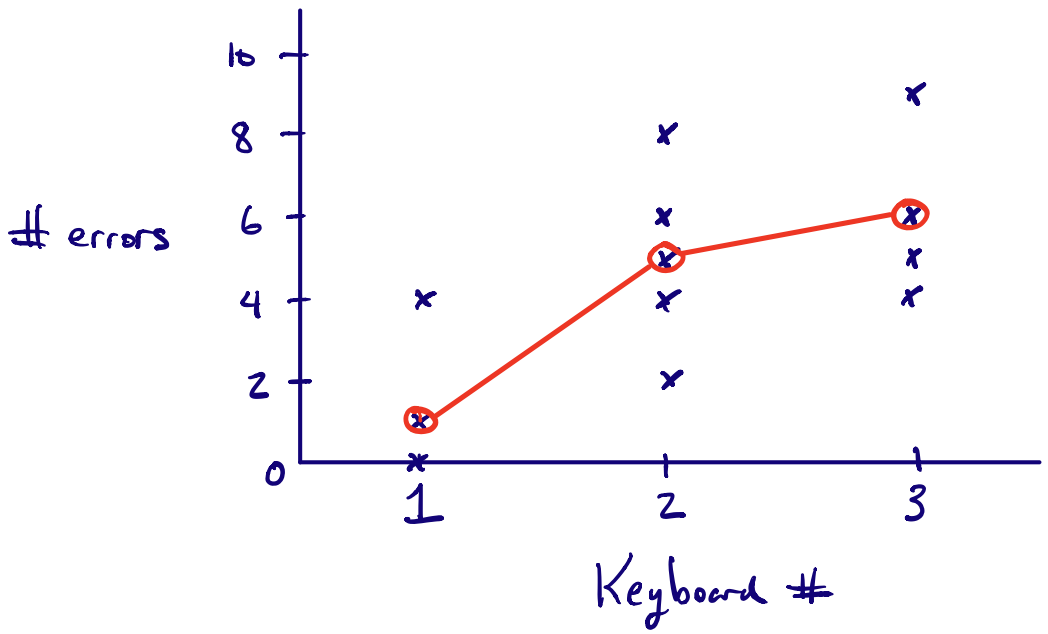
Lecture 3 – Inference in Single Factor Designs

Guiding example: An I/O psychologist studied three computer keyboard designs. Subjects were randomly assigned to type on one of the three designs, and the # of errors was recorded.

Keyboard 1	Keyboard 2	Keyboard 3
0	6	6
4	8	5
0	5	9
1	4	4
0	2	6
$\bar{X}_1 = 1$	$\bar{X}_2 = 5$	$\bar{X}_3 = 6$

To answer questions about these data, we need to put some **structure** on the data



Structure = statistical model

* write a formula that generates each observed data point
* "data story"

Linear model: $$X_{ij} = \mu_j + \varepsilon_{ij}$$

where: X_{ij} = score of subject # i $(i = 1, \cdots, 5)$ in condition #j $(j = 1, 2, 3)$

μ_j = population mean for condition #j $(j = 1, 2, 3)$

ε_{ij} = deviation of subject i in condition j from μ_j (i.e., model error)

In words, this means that each observed score can be written as the sum of

(1) the condition mean

(2) individual variation around the condition mean.

Inference goals:

(1) test hypotheses about the condition means μ_j

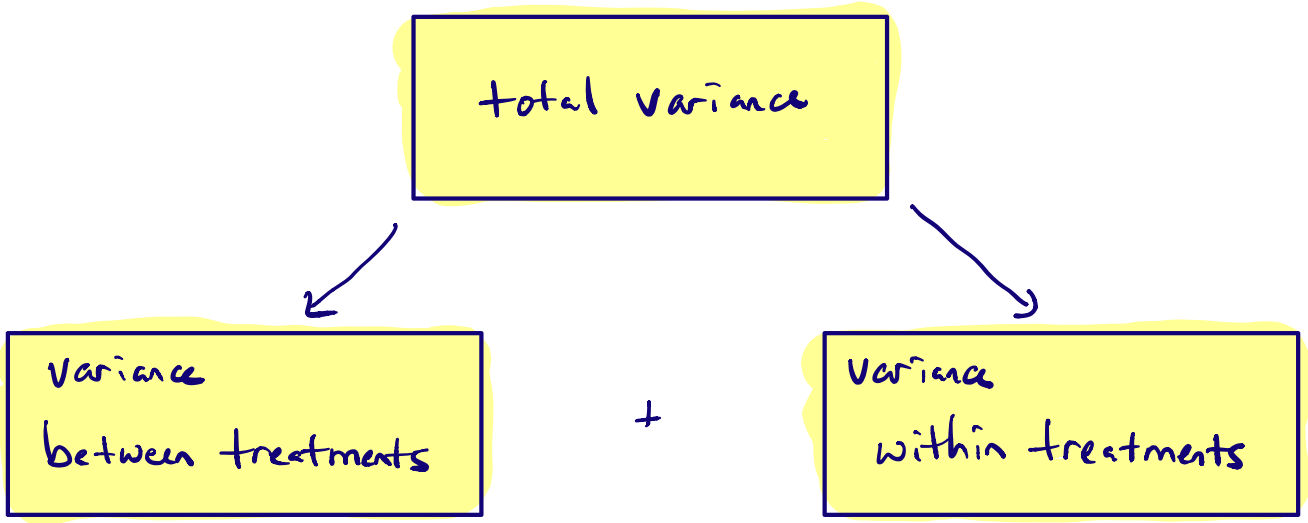
(2) estimate the condition means μ_j

Hypothesis testing:

Consider $\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \text{not all } \mu_i\text{'s equal.} \end{cases}$ Which better predicts the observed data?

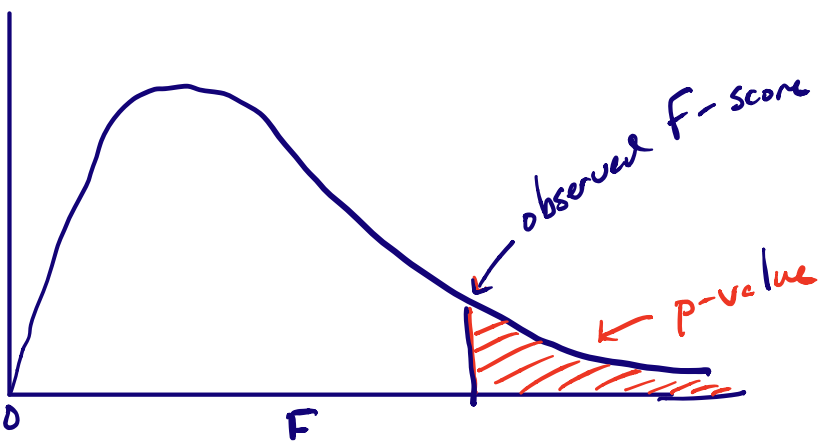
Recall: frequentist approach → compute $p(\text{data} \mid H_0)$.

R.A. Fisher (1925) developed a creative method for computing this likelihood – partitioning the variance in the observed data.



$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

Under H_0, this statistic forms an F- distribution



So how do we compute these two different variances?

Recall: variance = $\frac{SS}{df}$ ← need to compute

(1) SS between group means

(2) SS within groups

(1) To compute SS between group means, we sum the squared deviations of each group mean $(\bar{X}_1, \bar{X}_2, \bar{X}_3)$ from the grand mean $(\bar{\bar{X}})$

$$SS = (\bar{X}_1 - \bar{\bar{X}})^2 + (\bar{X}_2 - \bar{\bar{X}})^2 + (\bar{X}_3 - \bar{\bar{X}})^2$$

$$= \sum_{j=1}^{3} (\bar{X}_j - \bar{\bar{X}})^2$$

Technical note: this SS, when divided by the appropriate df, will estimate the variance of the distribution of sample means, σ^2_{means}.

By the central limit theorem, $\sigma^2_{means} = \frac{\sigma^2_{subjects}}{N}$

Equivalently, $\sigma^2_{subjects} = N \cdot \sigma^2_{means}$

So we need to multiply by N to "scale up" to the level of variation among subjects:

$$SS_{between} = N \sum_{j=1}^{3} (\bar{X}_j - \bar{\bar{X}})^2$$

For our data:

$\bar{X}_1 = 1$

$\bar{X}_2 = 5$ $\quad \bar{X} = 4 \quad N = 5$

$\bar{X}_3 = 6$

$$SS_{between} = N \sum_{j=1}^{3} (\bar{X}_j - \bar{X})^2$$

$$= 5 \left[(1-4)^2 + (5-4)^2 + (6-4)^2 \right]$$

$$= 5 (9 + 1 + 4)$$

$$= 70$$

Convert to variance (which we denote MS for "mean squared deviation")

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{70}{2} = 35$$

(2) to compute SS within groups, we pool together the SS for each group:

$$SS_{within} = \sum_{j=1}^{3} \sum_{i=1}^{5} \left(X_{ij} - \overline{X}_j\right)^2$$

$$SS_{within} = (0-1)^2 + (4-1)^2 + (0-1)^2 + (1-1)^2 + (0-1)^2 \quad \text{group 1}$$
$$+ (6-5)^2 + (8-5)^2 + (5-5)^2 + (4-5)^2 + (2-5)^2 \quad \text{group 2}$$
$$+ (6-6)^2 + (5-6)^2 + (9-6)^2 + (4-6)^2 + (6-6)^2 \quad \text{group 3}$$

$$= 1 + 9 + 1 + 0 + 1$$
$$+ 1 + 9 + 0 + 1 + 9$$
$$+ 0 + 1 + 9 + 4 + 0$$

$$= 46$$

Convert to variance: since there are $5 - 1 = 4$ degrees of freedom in each of the three groups, we have a total of $4 \times 3 = 12$ degrees of freedom within groups

$$\longrightarrow MS_{within} = \frac{46}{12} = 3.833$$

Finally, we can compute F:

$$F = \frac{\text{variance between groups}}{\text{variance within groups}} = \frac{MS_{between}}{MS_{within}} = \frac{35}{3.833} = 9.14$$

How surprising is this if $\mathcal{H}_0$ true?

$$p = p\left(F > 9.14 \mid \mathcal{H}_0\right) = 0.00387$$

So, data are rare under $\mathcal{H}_0$ $\longrightarrow$ reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$.

Bayes factor?

* a quick and dirty approximation comes from Faulkenberg (2018)

$\longrightarrow$ convert observed F to BF_{01}

$$BF_{01} \approx \sqrt{n^{df_1}\left(1 + \frac{F\,df_1}{df_2}\right)^{-n}}$$

OR: use online app, described in Faulkenberg (2019)

https://tomfaulkenberg.shinyapps.io/anovaBFcalc

Estimation

Let's get a confidence interval for μ_2

$$95\%\ CI = \bar{X}_2 \pm t^*_{df} \cdot \frac{\hat{\sigma}_2}{\sqrt{N_2}}$$

$$\bar{X}_2 = 5$$

$$t^*_{df} = 2.78 \quad (\text{use } df = 5-1 = 4)$$

$$\hat{\sigma}_2 = \sqrt{MS_{within}} = \sqrt{3.833} = 1.96$$

$$\text{so } 95\%\ CI = 5 \pm 2.78 \cdot \frac{1.96}{\sqrt{5}}$$

$$= 5 \pm 2.44$$

$$= (2.56, 7.44)$$