ISE 315: Engineering Statistics

Lecture 23: After the F-Test — Multiple Comparisons, Residuals, Sample Size

Instructor: Mansur M. Arief, PhD
Industrial and Systems Engineering, KFUPM

Office: 22-219 — Email: mansur.arief@kfupm.edu.sa

Based on Montgomery & Runger, Applied Statistics and Probability for Engineers, 6th Ed.

Lecture 23

After the F-Test: Multiple Comparisons, Residual Analysis, Sample Size
(§8.7–§8.9)

Lecture 23 Outline

- Recap: what the F-test does and does not say

Lecture 23 Outline

- Recap: what the F-test does and does not say
- Running example: a startup's onboarding experiment

Lecture 23 Outline

- Recap: what the F-test does and does not say
- Running example: a startup's onboarding experiment
- Fisher's Least Significant Difference (8.14), Procedure 8.2

Lecture 23 Outline

- Recap: what the F-test does and does not say
- Running example: a startup's onboarding experiment
- Fisher's Least Significant Difference (8.14), Procedure 8.2
- Tukey's Honestly Significant Difference (8.15), Procedure 8.3

Lecture 23 Outline

- Recap: what the F-test does and does not say
- Running example: a startup's onboarding experiment
- Fisher's Least Significant Difference (8.14), Procedure 8.2
- Tukey's Honestly Significant Difference (8.15), Procedure 8.3
- LSD vs. HSD on the same data — where they disagree

Lecture 23 Outline

- Recap: what the F-test does and does not say
- Running example: a startup's onboarding experiment
- Fisher's Least Significant Difference (8.14), Procedure 8.2
- Tukey's Honestly Significant Difference (8.15), Procedure 8.3
- LSD vs. HSD on the same data — where they disagree
- Residual analysis: checking the assumptions (8.16)

Lecture 23 Outline

- Recap: what the F-test does and does not say
- Running example: a startup's onboarding experiment
- Fisher's Least Significant Difference (8.14), Procedure 8.2
- Tukey's Honestly Significant Difference (8.15), Procedure 8.3
- LSD vs. HSD on the same data — where they disagree
- Residual analysis: checking the assumptions (8.16)
- Sample size via the non-centrality parameter (8.17)

Recap: What ANOVA Tells Us (and Doesn't)

Last lecture, the red-team strategy experiment gave

$$F_0 = 8.00 > f_{0.05,\,2,\,12} = 3.89 \quad \Rightarrow \quad \text{reject } H_0 : \mu_1 = \mu_2 = \mu_3$$

Recap: What ANOVA Tells Us (and Doesn't)

Last lecture, the red-team strategy experiment gave

$$F_0 = 8.00 > f_{0.05,\,2,\,12} = 3.89 \quad \Rightarrow \quad \text{reject } H_0 : \mu_1 = \mu_2 = \mu_3$$

What this tells us: *at least one* pair of treatment means differs.
What it does not tell us: *which* pairs differ.

Recap: What ANOVA Tells Us (and Doesn't)

Last lecture, the red-team strategy experiment gave

$$F_0 = 8.00 > f_{0.05,\,2,\,12} = 3.89 \quad \Rightarrow \quad \text{reject } H_0 : \mu_1 = \mu_2 = \mu_3$$

What this tells us: *at least one* pair of treatment means differs.
What it does not tell us: *which* pairs differ.

Multiple comparisons answer the follow-up question — which specific means differ — while watching the family-wise error rate (8.1).

Recap: What ANOVA Tells Us (and Doesn't)

Last lecture, the red-team strategy experiment gave

$$F_0 = 8.00 > f_{0.05,\,2,\,12} = 3.89 \quad \Rightarrow \quad \text{reject } H_0 : \mu_1 = \mu_2 = \mu_3$$

What this tells us: *at least one* pair of treatment means differs.
What it does not tell us: *which* pairs differ.

Multiple comparisons answer the follow-up question — which specific means differ — while watching the family-wise error rate (8.1).

Two procedures today: Fisher's LSD (simple, powerful, no family-wise control) and Tukey's HSD (controls the family-wise rate exactly).

Textbook: Chapter 8, §8.7 ((8.14) onward)

Running Example: Onboarding Flows

Table 8.6

A startup tests three onboarding flows: **A** (current design), **B** (checklist on first screen), **C** (guided interactive tour). New signups are randomly assigned. Response: activation rate — % of a weekly cohort completing a key action within 7 days. Each flow gets $n = 6$ cohorts.

Running Example: Onboarding Flows

Table 8.6

A startup tests three onboarding flows: **A** (current design), **B** (checklist on first screen), **C** (guided interactive tour). New signups are randomly assigned. Response: activation rate — % of a weekly cohort completing a key action within 7 days. Each flow gets $n = 6$ cohorts.

Flow	**Activation rate (%) by cohort**						$y_{i\cdot}$	$\bar{y}_{i\cdot}$
A. Current	48	55	51	56	50	52	312	52.0
B. Checklist	53	59	54	58	52	57	333	55.5
C. Tour	57	63	59	62	58	61	360	60.0
						Grand total $y_{\cdot\cdot}$	1005	
						Grand mean $\bar{y}_{\cdot\cdot}$		55.83

Running Example: Onboarding Flows

Table 8.6

A startup tests three onboarding flows: **A** (current design), **B** (checklist on first screen), **C** (guided interactive tour). New signups are randomly assigned. Response: activation rate — % of a weekly cohort completing a key action within 7 days. Each flow gets $n = 6$ cohorts.

Flow	**Activation rate (%) by cohort**						$y_{i\cdot}$	$\bar{y}_{i\cdot}$
A. Current	48	55	51	56	50	52	312	52.0
B. Checklist	53	59	54	58	52	57	333	55.5
C. Tour	57	63	59	62	58	61	360	60.0
						Grand total $y_{\cdot\cdot}$	1005	
						Grand mean $\bar{y}_{\cdot\cdot}$		55.83

$a = 3$, $n = 6$, $N = 18$. First step: the overall F-test (Procedure 8.1).

Step 0: The Overall F-Test

Prerequisite for any pairwise comparison

Sums of squares (shortcut forms (8.6), (8.7)):

$$\sum\sum y_{ij}^2 = 56421, \qquad \frac{y_{\cdot\cdot}^2}{N} = \frac{1005^2}{18} = 56112.5 \;\Rightarrow\; SS_T = 308.5$$

Step 0: The Overall F-Test

Prerequisite for any pairwise comparison

Sums of squares (shortcut forms (8.6), (8.7)):

$$\sum\sum y_{ij}^2 = 56421, \qquad \frac{y_{\cdot\cdot}^2}{N} = \frac{1005^2}{18} = 56112.5 \;\Rightarrow\; SS_T = 308.5$$

$$SS_{\text{Flows}} = \frac{312^2 + 333^2 + 360^2}{6} - 56112.5 = 56305.5 - 56112.5 = 193.0$$

$$SS_E = 308.5 - 193.0 = 115.5$$

Step 0: The Overall F-Test

Prerequisite for any pairwise comparison

Sums of squares (shortcut forms (8.6), (8.7)):

$$\sum\sum y_{ij}^2 = 56421, \qquad \frac{y_{..}^2}{N} = \frac{1005^2}{18} = 56112.5 \;\Rightarrow\; SS_T = 308.5$$

$$SS_{\text{Flows}} = \frac{312^2 + 333^2 + 360^2}{6} - 56112.5 = 56305.5 - 56112.5 = 193.0$$

$$SS_E = 308.5 - 193.0 = 115.5$$

Source	SS	df	MS	F_0
Flows	193.0	2	96.50	12.53
Error	115.5	15	7.700	

Step 0: The Overall F-Test

Prerequisite for any pairwise comparison

Sums of squares (shortcut forms (8.6), (8.7)):

$$\sum\sum y_{ij}^2 = 56421, \qquad \frac{y_{..}^2}{N} = \frac{1005^2}{18} = 56112.5 \;\Rightarrow\; SS_T = 308.5$$

$$SS_{\text{Flows}} = \frac{312^2 + 333^2 + 360^2}{6} - 56112.5 = 56305.5 - 56112.5 = 193.0$$

$$SS_E = 308.5 - 193.0 = 115.5$$

Source	SS	df	MS	F_0
Flows	193.0	2	96.50	12.53
Error	115.5	15	7.700	

Fisher's Least Significant Difference (LSD)

§8.7.1

Idea: a two-sample t-test for each pair (i, k), but with the pooled MS_E from the *full* ANOVA (all $N - a$ error df):

$$|t_0| = \frac{|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}|}{\sqrt{2\, MS_E / n}} > t_{\alpha/2,\, N-a}$$

Fisher's Least Significant Difference (LSD)

§8.7.1

Idea: a two-sample t-test for each pair (i, k), but with the pooled MS_E from the *full* ANOVA (all $N - a$ error df):

$$|t_0| = \frac{|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}|}{\sqrt{2\,MS_E/n}} > t_{\alpha/2,\,N-a}$$

Multiply through by the denominator: one yardstick for every pair,

$$\mathrm{LSD} = t_{\alpha/2,\,N-a}\sqrt{\frac{2\,MS_E}{n}} \tag{8.14}$$

declare $\bar{y}_{i\cdot}$, $\bar{y}_{k\cdot}$ different when $|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}| > \mathrm{LSD}$.

Fisher's Least Significant Difference (LSD)

§8.7.1

Idea: a two-sample t-test for each pair (i, k), but with the pooled MS_E from the *full* ANOVA (all $N - a$ error df):

$$|t_0| = \frac{|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}|}{\sqrt{2\, MS_E / n}} > t_{\alpha/2,\, N-a}$$

Multiply through by the denominator: one yardstick for every pair,

$$\text{LSD} = t_{\alpha/2,\, N-a} \sqrt{\frac{2\, MS_E}{n}} \tag{8.14}$$

declare $\bar{y}_{i\cdot}$, $\bar{y}_{k\cdot}$ different when $|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}| > \text{LSD}$.
Note: the LSD is exactly the half-width of the CI for a difference (8.13) — a pair is LSD-significant iff that interval excludes zero.

Fisher's Least Significant Difference (LSD)

§8.7.1

Idea: a two-sample t-test for each pair (i, k), but with the pooled MS_E from the *full* ANOVA (all $N - a$ error df):

$$|t_0| = \frac{|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}|}{\sqrt{2\, MS_E / n}} > t_{\alpha/2,\, N-a}$$

Multiply through by the denominator: one yardstick for every pair,

$$\text{LSD} = t_{\alpha/2,\, N-a} \sqrt{\frac{2\, MS_E}{n}} \tag{8.14}$$

declare $\bar{y}_{i\cdot}$, $\bar{y}_{k\cdot}$ different when $|\bar{y}_{i\cdot} - \bar{y}_{k\cdot}| > \text{LSD}$.
Note: the LSD is exactly the half-width of the CI for a difference (8.13) — a pair is LSD-significant iff that interval excludes zero.

Caution: LSD controls α *per pair*, not family-wise. Procedure 8.2 therefore

LSD Applied to the Onboarding Flows

Following Procedure 8.2

Step 1 — prerequisite: $F_0 = 12.53 > 3.68$ ✓. **Step 2 — ingredients:** $MS_E = 7.700$, $N - a = 15$ df, and $t_{0.025,\,15} = 2.131$ from the t table (Table 8.8).

LSD Applied to the Onboarding Flows

Following Procedure 8.2

Step 1 — prerequisite: $F_0 = 12.53 > 3.68$ ✓. **Step 2 — ingredients:** $MS_E = 7.700$, $N - a = 15$ df, and $t_{0.025,\,15} = 2.131$ from the t table (Table 8.8).

Step 3 — the yardstick (8.14):

$$\text{LSD} = 2.131\sqrt{\frac{2(7.700)}{6}} = 2.131\sqrt{2.567} = 2.131 \times 1.602 = 3.414$$

LSD Applied to the Onboarding Flows

Following Procedure 8.2

Step 1 — prerequisite: $F_0 = 12.53 > 3.68$ ✓. **Step 2 — ingredients:** $MS_E = 7.700$, $N - a = 15$ df, and $t_{0.025,\,15} = 2.131$ from the t table (Table 8.8).

Step 3 — the yardstick (8.14):

$$\text{LSD} = 2.131\sqrt{\frac{2(7.700)}{6}} = 2.131\sqrt{2.567} = 2.131 \times 1.602 = 3.414$$

Step 4 — compare every pair:

Pair	$\lvert\bar{y}_{i\cdot} - \bar{y}_{k\cdot}\rvert$	$> \text{LSD} = 3.414$?	LSD verdict
A vs. B	$\lvert 52.0 - 55.5\rvert = 3.5$	Yes	differ
A vs. C	$\lvert 52.0 - 60.0\rvert = 8.0$	Yes	differ
B vs. C	$\lvert 55.5 - 60.0\rvert = 4.5$	Yes	differ

LSD Applied to the Onboarding Flows

Following Procedure 8.2

Step 1 — prerequisite: $F_0 = 12.53 > 3.68$ ✓. **Step 2 — ingredients:** $MS_E = 7.700$, $N - a = 15$ df, and $t_{0.025,\,15} = 2.131$ from the t table (Table 8.8).

Step 3 — the yardstick (8.14):

$$\text{LSD} = 2.131\sqrt{\frac{2(7.700)}{6}} = 2.131\sqrt{2.567} = 2.131 \times 1.602 = 3.414$$

Step 4 — compare every pair:

Pair	$\lvert\bar{y}_{i\cdot} - \bar{y}_{k\cdot}\rvert$	$> \text{LSD} = 3.414$?	LSD verdict
A vs. B	$\lvert 52.0 - 55.5\rvert = 3.5$	Yes	differ
A vs. C	$\lvert 52.0 - 60.0\rvert = 8.0$	Yes	differ
B vs. C	$\lvert 55.5 - 60.0\rvert = 4.5$	Yes	differ

Quick Check: Fisher LSD

Supply chain

A freight broker's ANOVA of $a = 5$ carriers gives $F_0 = 8.4$ ($p < 0.001$), $MS_E = 2.0$, $n = 6$ deliveries per carrier, $df_E = 25$; $t_{0.025,\,25} = 2.060$.

Q. Two carrier means differ by 1.5 days. Which conclusion is correct?

(A) They differ significantly because F_0 is large.

(B) They do not differ significantly because $1.5 < \text{LSD} = 1.68$.

(C) They differ significantly because $1.5 > 0.05 = \alpha$.

(D) We cannot tell without all five carrier means.

Quick Check: Fisher LSD

Supply chain

A freight broker's ANOVA of $a = 5$ carriers gives $F_0 = 8.4$ ($p < 0.001$), $MS_E = 2.0$, $n = 6$ deliveries per carrier, $df_E = 25$; $t_{0.025,\,25} = 2.060$.

Q. Two carrier means differ by 1.5 days. Which conclusion is correct?

(A) They differ significantly because F_0 is large.

(B) They do not differ significantly because $1.5 < \text{LSD} = 1.68$.

(C) They differ significantly because $1.5 > 0.05 = \alpha$.

(D) We cannot tell without all five carrier means.

Answer: (B). By (8.14), $\text{LSD} = 2.060\sqrt{2(2.0)/6} = 2.060(0.8165) = 1.68$. Since $1.5 < 1.68$, this pair is *not* separated. (A) confuses the omnibus test with a pairwise one; (C) compares quantities with different units; (D) is wrong — LSD needs only the two means and MS_E.

Tukey's Honestly Significant Difference (HSD)

§8.7.2

The sharper question: under H_0, how far apart should the *largest* and *smallest* of a sample means fall just by chance? The **Studentized range**

$$q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{\sqrt{MS_E/n}}$$

has upper-α point $q_\alpha(a, f)$, with f = error df.

Tukey's Honestly Significant Difference (HSD)

§8.7.2

The sharper question: under H_0, how far apart should the *largest* and *smallest* of a sample means fall just by chance? The **Studentized range**

$$q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{\sqrt{MS_E/n}}$$

has upper-α point $q_\alpha(a, f)$, with f = error df.

Calibrate the yardstick to the most extreme pair, and *every* pair is simultaneously tested at family-wise α:

$$T_\alpha = q_\alpha(a, f)\sqrt{\frac{MS_E}{n}}, \qquad f = N - a \tag{8.15}$$

Tukey's Honestly Significant Difference (HSD)

§8.7.2

The sharper question: under H_0, how far apart should the *largest* and *smallest* of a sample means fall just by chance? The **Studentized range**

$$q = \frac{\bar{y}_{\text{max}} - \bar{y}_{\text{min}}}{\sqrt{MS_E/n}}$$

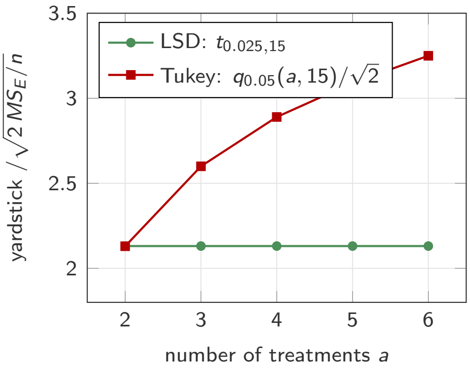
has upper-α point $q_\alpha(a, f)$, with f = error df.
Calibrate the yardstick to the most extreme pair, and *every* pair is simultaneously tested at family-wise α:

$$T_\alpha = q_\alpha(a, f)\sqrt{\frac{MS_E}{n}}, \qquad f = N - a \tag{8.15}$$

Two properties worth memorizing:

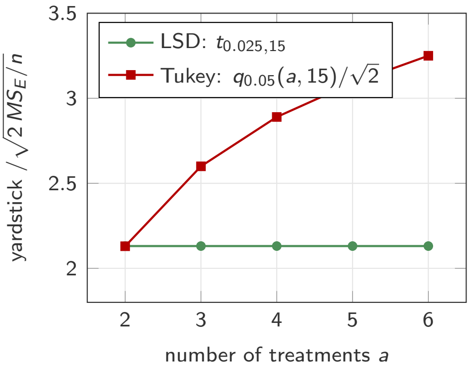
The Two Yardsticks as a Grows

Figure 8.4



The Two Yardsticks as a Grows

Figure 8.4



Both yardsticks on a common scale ($f = 15$, $\alpha = 0.05$):

- They coincide at $a = 2$ — one comparison, nothing to correct.
- For $a \geq 3$, Tukey's grows with the number of treatments — the price of family-wise control.
- LSD's stays flat — which is exactly why its family-wise rate drifts up.

Tukey HSD Applied to the Onboarding Flows

Following Procedure 8.3

Step 1 — ingredients: $MS_E = 7.700$, $f = N - a = 15$.

Tukey HSD Applied to the Onboarding Flows

Following Procedure 8.3

Step 1 — ingredients: $MS_E = 7.700$, $f = N - a = 15$.

Step 2 — table lookup (Table 8.9); first argument is a, *not* the number of pairs:

$f \backslash a$	2	3	4
14	3.03	3.70	4.11
15	3.01	3.67	4.08
16	3.00	3.65	4.05

$q_{0.05}(3, 15) = 3.67$

Tukey HSD Applied to the Onboarding Flows

Following Procedure 8.3

Step 1 — ingredients: $MS_E = 7.700$, $f = N - a = 15$.

Step 2 — table lookup (Table 8.9); first argument is a, *not* the number of pairs:

$f \backslash a$	2	3	4
14	3.03	3.70	4.11
15	3.01	3.67	4.08
16	3.00	3.65	4.05

$q_{0.05}(3, 15) = 3.67$

Step 3 — the yardstick (8.15):

$$T_{0.05} = 3.67\sqrt{\frac{7.700}{6}} = 3.67\sqrt{1.283} = 3.67 \times 1.133 = 4.157$$

Tukey HSD Applied to the Onboarding Flows

Following Procedure 8.3

Step 1 — ingredients: $MS_E = 7.700$, $f = N - a = 15$.

Step 2 — table lookup (Table 8.9); first argument is a, *not* the number of pairs:

$f \backslash a$	2	3	4
14	3.03	3.70	4.11
15	3.01	3.67	4.08
16	3.00	3.65	4.05

$q_{0.05}(3, 15) = 3.67$

Step 3 — the yardstick (8.15):

$$T_{0.05} = 3.67\sqrt{\frac{7.700}{6}} = 3.67\sqrt{1.283} = 3.67 \times 1.133 = 4.157$$

LSD vs. HSD on the Same Data

Table 8.10

Pair	$\lvert \bar{y}_{i\cdot} - \bar{y}_{k\cdot} \rvert$	$> \text{LSD} = 3.414?$	$> T_{0.05} = 4.157?$	Verdict
A vs. B	3.5	Yes	**No**	methods disagree
A vs. C	8.0	Yes	Yes	differ
B vs. C	4.5	Yes	Yes	differ

LSD vs. HSD on the Same Data

Table 8.10

Pair	$\lvert \bar{y}_{i\cdot} - \bar{y}_{k\cdot} \rvert$	$>$ LSD $= 3.414$?	$> T_{0.05} = 4.157$?	Verdict
A vs. B	3.5	Yes	**No**	methods disagree
A vs. C	8.0	Yes	Yes	differ
B vs. C	4.5	Yes	Yes	differ

Underline diagram (Figure 8.5) — underline sets Tukey cannot separate:

LSD vs. HSD on the Same Data

Table 8.10

Pair	$\lvert \bar{y}_{i\cdot} - \bar{y}_{k\cdot} \rvert$	$>$ LSD $= 3.414$?	$> T_{0.05} = 4.157$?	Verdict
A vs. B	3.5	Yes	**No**	methods disagree
A vs. C	8.0	Yes	Yes	differ
B vs. C	4.5	Yes	Yes	differ

Underline diagram (Figure 8.5) — underline sets Tukey cannot separate:

Honest summary for the product team: the tour (C) is better than both alternatives; the checklist-vs-current gap (3.5 points) is suggestive but not con-

Quick Check: Tukey HSD vs. Fisher LSD

AI safety

Q. An AI safety team compares $a = 6$ guardrail configurations with one-way ANOVA. Which statement about Fisher LSD vs. Tukey HSD is *correct*?

(A) LSD and Tukey always declare the same pairs significant.

(B) Tukey is less conservative than LSD when a is large.

(C) For $a \geq 3$, Tukey's critical value is larger, so it rejects fewer pairs at the same α.

(D) LSD is preferred because it controls the family-wise error rate.

Quick Check: Tukey HSD vs. Fisher LSD

AI safety

Q. An AI safety team compares $a = 6$ guardrail configurations with one-way ANOVA. Which statement about Fisher LSD vs. Tukey HSD is *correct*?

(A) LSD and Tukey always declare the same pairs significant.

(B) Tukey is less conservative than LSD when a is large.

(C) For $a \geq 3$, Tukey's critical value is larger, so it rejects fewer pairs at the same α.

(D) LSD is preferred because it controls the family-wise error rate.

Answer: (C). The Studentized range satisfies $q_\alpha(a, f) > \sqrt{2}\, t_{\alpha/2,f}$ for $a \geq 3$, so $T_\alpha >$ LSD (8.15), (8.14). (A) fails whenever a gap lands between the two yardsticks — as A–B just did; (B) reverses the relationship; (D) is backwards — Tukey controls the family-wise rate, LSD does not.

Residual Analysis: Checking the Assumptions

§8.8

The F-test, the intervals, LSD, and HSD all lean on: normal errors, equal variances, independence. The raw material for checking is the residuals:

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot} \tag{8.16}$$

The fitted value in one-way ANOVA is simply the treatment mean.

Residual Analysis: Checking the Assumptions

§8.8

The F-test, the intervals, LSD, and HSD all lean on: normal errors, equal variances, independence. The raw material for checking is the residuals:

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot} \tag{8.16}$$

The fitted value in one-way ANOVA is simply the treatment mean.

Three plots, one assumption each:

Plot	**Healthy pattern**	**Warning sign**
Normal probability plot	straight line	heavy tails, S-shape
Residuals vs. fitted	even vertical spread	funnel (variance grows)
Residuals vs. run order	no trend	drift, oscillation

Residual Analysis: Checking the Assumptions

§8.8

The F-test, the intervals, LSD, and HSD all lean on: normal errors, equal variances, independence. The raw material for checking is the residuals:

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot} \tag{8.16}$$

The fitted value in one-way ANOVA is simply the treatment mean.

Three plots, one assumption each:

Plot	Healthy pattern	Warning sign
Normal probability plot	straight line	heavy tails, S-shape
Residuals vs. fitted	even vertical spread	funnel (variance grows)
Residuals vs. run order	no trend	drift, oscillation

Remedies: variance grows with mean $\to$ $\sqrt{y}$ or $\log y$; severe non-normality $\to$ Kruskal–

Residuals in Practice

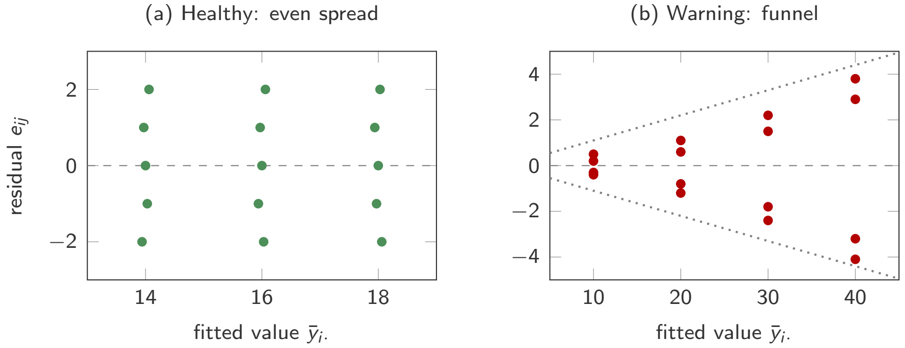
Red-team data (healthy) vs. a funnel (warning)

For the red-team data: group 1 residuals $(-2, 1, 0, -1, 2)$, group 2 $(1, -1, 0, 2, -2)$, group 3 $(-1, 1, 0, -2, 2)$ — each set sums to zero by construction.

Residuals in Practice

Red-team data (healthy) vs. a funnel (warning)

For the red-team data: group 1 residuals $(-2, 1, 0, -1, 2)$, group 2 $(1, -1, 0, 2, -2)$, group 3 $(-1, 1, 0, -2, 2)$ — each set sums to zero by construction.

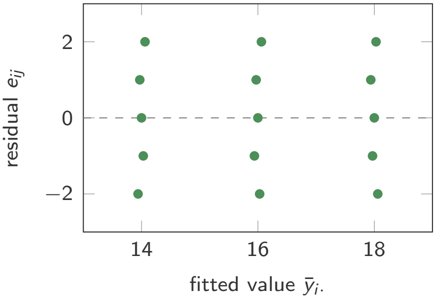


Residuals in Practice

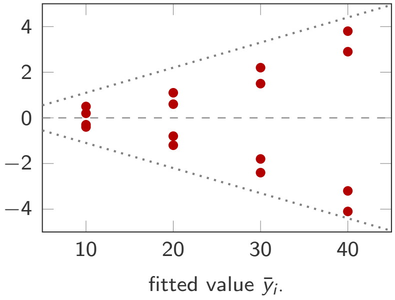
Red-team data (healthy) vs. a funnel (warning)

For the red-team data: group 1 residuals $(-2, 1, 0, -1, 2)$, group 2 $(1, -1, 0, 2, -2)$, group 3 $(-1, 1, 0, -2, 2)$ — each set sums to zero by construction.

(a) Healthy: even spread



(b) Warning: funnel



Choosing the Sample Size

§8.9

Question: how many replicates n per treatment to detect a meaningful effect with adequate power $1-\beta$?

Choosing the Sample Size

§8.9

Question: how many replicates n per treatment to detect a meaningful effect with adequate power $1-\beta$?

The difficulty of the detection problem is summarized by the **non-centrality parameter**:

$$\Phi^2 = \frac{n\sum_{i=1}^{a}\tau_i^2}{a\,\sigma^2} \tag{8.17}$$

Large when effects are big, replication generous, noise small.

Choosing the Sample Size

§8.9

Question: how many replicates n per treatment to detect a meaningful effect with adequate power $1-\beta$?

The difficulty of the detection problem is summarized by the **non-centrality parameter**:

$$\Phi^2 = \frac{n \sum_{i=1}^{a} \tau_i^2}{a\,\sigma^2} \tag{8.17}$$

Large when effects are big, replication generous, noise small.
Planning shortcut — if some two means are at least D apart, the worst case gives

$$\Phi^2 \;\geq\; \frac{nD^2}{2a\,\sigma^2}$$

then read power from operating characteristic (OC) curves as a function of Φ, $a-1$, and df_E.

Choosing the Sample Size

§8.9

Question: how many replicates n per treatment to detect a meaningful effect with adequate power $1-\beta$?

The difficulty of the detection problem is summarized by the **non-centrality parameter**:

$$\Phi^2 = \frac{n\sum_{i=1}^{a} \tau_i^2}{a\,\sigma^2} \tag{8.17}$$

Large when effects are big, replication generous, noise small.

Planning shortcut — if some two means are at least D apart, the worst case gives

$$\Phi^2 \;\geq\; \frac{nD^2}{2a\,\sigma^2}$$

then read power from operating characteristic (OC) curves as a function of Φ, $a-1$, and df_E.

Inputs: a, α, target power, smallest meaningful D, and a pilot estimate of σ. Because D

Sample Size Worked: Planning a Guardrail Comparison

Table 8.11

Setup. An AI safety team will compare $a = 4$ guardrail configurations by false-positive rate on benign traffic. Pilot data: $\sigma \approx 2$ points. Goal: detect any gap of $D = 3$ points with power ≈ 0.90 at $\alpha = 0.05$.

Sample Size Worked: Planning a Guardrail Comparison

Table 8.11

Setup. An AI safety team will compare $a = 4$ guardrail configurations by false-positive rate on benign traffic. Pilot data: $\sigma \approx 2$ points. Goal: detect any gap of $D = 3$ points with power ≈ 0.90 at $\alpha = 0.05$.

Step 1 — the bound from (8.17):

$$\Phi^2 = \frac{nD^2}{2a\sigma^2} = \frac{n(3)^2}{2(4)(2)^2} = \frac{9n}{32} = 0.2813\, n$$

Sample Size Worked: Planning a Guardrail Comparison

Table 8.11

Setup. An AI safety team will compare $a = 4$ guardrail configurations by false-positive rate on benign traffic. Pilot data: $\sigma \approx 2$ points. Goal: detect any gap of $D = 3$ points with power ≈ 0.90 at $\alpha = 0.05$.

Step 1 — the bound from (8.17):

$$\Phi^2 = \frac{nD^2}{2a\sigma^2} = \frac{n(3)^2}{2(4)(2)^2} = \frac{9n}{32} = 0.2813\,n$$

Step 2 — try candidate n, read power from the OC chart:

n	$\Phi^2 = 9n/32$	Φ	$df_E = a(n-1)$	Power (OC chart)
4	1.13	1.06	12	≈ 0.55
5	1.41	1.19	16	≈ 0.70
6	1.69	1.30	20	≈ 0.82

Sample Size Worked: Planning a Guardrail Comparison

Table 8.11

Setup. An AI safety team will compare $a = 4$ guardrail configurations by false-positive rate on benign traffic. Pilot data: $\sigma \approx 2$ points. Goal: detect any gap of $D = 3$ points with power ≈ 0.90 at $\alpha = 0.05$.

Step 1 — the bound from (8.17):

$$\Phi^2 = \frac{nD^2}{2a\sigma^2} = \frac{n(3)^2}{2(4)(2)^2} = \frac{9n}{32} = 0.2813\, n$$

Step 2 — try candidate n, read power from the OC chart:

n	$\Phi^2 = 9n/32$	Φ	$df_E = a(n-1)$	Power (OC chart)
4	1.13	1.06	12	≈ 0.55
5	1.41	1.19	16	≈ 0.70
6	1.69	1.30	20	≈ 0.82

Quick Check: Sample Size

Q. Planning a one-way ANOVA, an engineer doubles the smallest detectable difference D from 2 to 4, holding a, σ, and n fixed. What happens to Φ^2?

(A) It doubles.

(B) It quadruples.

(C) It is halved.

(D) It is unchanged, because D does not appear in the formula.

Quick Check: Sample Size

Q. Planning a one-way ANOVA, an engineer doubles the smallest detectable difference D from 2 to 4, holding a, σ, and n fixed. What happens to Φ^2?

(A) It doubles.

(B) It quadruples.

(C) It is halved.

(D) It is unchanged, because D does not appear in the formula.

Answer: (B). In the planning bound $\Phi^2 \geq nD^2/(2a\sigma^2)$ built on (8.17), D enters squared: doubling D multiplies Φ^2 by four. Detecting larger effects needs far fewer observations — the relationship is strongly nonlinear. (D) is wrong because D does appear; (A) misses the square.

Putting It Together: The Complete Workflow

Step 0 — plan. Choose n from power requirements (8.17) *before* collecting data.

Putting It Together: The Complete Workflow

Step 0 — plan. Choose n from power requirements (8.17) *before* collecting data.

Step 1 — overall test. Build the ANOVA table (Procedure 8.1); if F_0 is not significant, stop and report.

Putting It Together: The Complete Workflow

Step 0 — plan. Choose n from power requirements (8.17) *before* collecting data.

Step 1 — overall test. Build the ANOVA table (Procedure 8.1); if F_0 is not significant, stop and report.

Step 2 — pairwise comparisons. If significant: Procedure 8.2 as an exploratory screen, Procedure 8.3 for the report.

Putting It Together: The Complete Workflow

Step 0 — plan. Choose n from power requirements (8.17) *before* collecting data.

Step 1 — overall test. Build the ANOVA table (Procedure 8.1); if F_0 is not significant, stop and report.

Step 2 — pairwise comparisons. If significant: Procedure 8.2 as an exploratory screen, Procedure 8.3 for the report.

Step 3 — diagnostics. Residual plots (8.16) to verify the assumptions the conclusions lean on.

Putting It Together: The Complete Workflow

Step 0 — plan. Choose n from power requirements (8.17) *before* collecting data.

Step 1 — overall test. Build the ANOVA table (Procedure 8.1); if F_0 is not significant, stop and report.

Step 2 — pairwise comparisons. If significant: Procedure 8.2 as an exploratory screen, Procedure 8.3 for the report.

Step 3 — diagnostics. Residual plots (8.16) to verify the assumptions the conclusions lean on.

Step 4 — practical interpretation. Translate into engineering language: which levels are best, by how much, within what uncertainty.

Lecture 23 Summary

- The F-test says *some* means differ; **multiple comparisons** say *which*.

Lecture 23 Summary

- The F-test says *some* means differ; **multiple comparisons** say *which*.
- **Fisher LSD** (8.14): $t_{\alpha/2,N-a}\sqrt{2MS_E/n}$ — per-pair α only; run only after a significant F (Procedure 8.2).

Lecture 23 Summary

- The F-test says *some* means differ; **multiple comparisons** say *which*.
- **Fisher LSD** (8.14): $t_{\alpha/2,N-a}\sqrt{2MS_E/n}$ — per-pair α only; run only after a significant F (Procedure 8.2).
- **Tukey HSD** (8.15): $q_\alpha(a,f)\sqrt{MS_E/n}$ — larger yardstick, family-wise α controlled (Procedure 8.3).

Lecture 23 Summary

- The F-test says *some* means differ; **multiple comparisons** say *which*.
- **Fisher LSD** (8.14): $t_{\alpha/2,N-a}\sqrt{2MS_E/n}$ — per-pair α only; run only after a significant F (Procedure 8.2).
- **Tukey HSD** (8.15): $q_\alpha(a, f)\sqrt{MS_E/n}$ — larger yardstick, family-wise α controlled (Procedure 8.3).
- On the onboarding data: LSD flagged all three pairs; Tukey declined A–B. When they disagree, report Tukey.

Lecture 23 Summary

- The F-test says *some* means differ; **multiple comparisons** say *which*.
- **Fisher LSD** (8.14): $t_{\alpha/2,N-a}\sqrt{2MS_E/n}$ — per-pair α only; run only after a significant F (Procedure 8.2).
- **Tukey HSD** (8.15): $q_\alpha(a,f)\sqrt{MS_E/n}$ — larger yardstick, family-wise α controlled (Procedure 8.3).
- On the onboarding data: LSD flagged all three pairs; Tukey declined A–B. When they disagree, report Tukey.
- **Residuals** (8.16) check normality, equal variance, independence — minutes well spent.

Lecture 23 Summary

- The F-test says *some* means differ; **multiple comparisons** say *which*.
- **Fisher LSD** (8.14): $t_{\alpha/2,N-a}\sqrt{2MS_E/n}$ — per-pair α only; run only after a significant F (Procedure 8.2).
- **Tukey HSD** (8.15): $q_\alpha(a,f)\sqrt{MS_E/n}$ — larger yardstick, family-wise α controlled (Procedure 8.3).
- On the onboarding data: LSD flagged all three pairs; Tukey declined A–B. When they disagree, report Tukey.
- **Residuals** (8.16) check normality, equal variance, independence — minutes well spent.
- **Sample size** comes from Φ^2 (8.17) and OC curves, decided before the experiment.

Lecture 23 Summary

- The F-test says *some* means differ; **multiple comparisons** say *which*.
- **Fisher LSD** (8.14): $t_{\alpha/2,N-a}\sqrt{2MS_E/n}$ — per-pair α only; run only after a significant F (Procedure 8.2).
- **Tukey HSD** (8.15): $q_\alpha(a,f)\sqrt{MS_E/n}$ — larger yardstick, family-wise α controlled (Procedure 8.3).
- On the onboarding data: LSD flagged all three pairs; Tukey declined A–B. When they disagree, report Tukey.
- **Residuals** (8.16) check normality, equal variance, independence — minutes well spent.
- **Sample size** comes from Φ^2 (8.17) and OC curves, decided before the experiment.
- **Next lecture:** random effects and the randomized complete block design.

Reminders

- **Quiz on multiple comparisons** is due before class.

Reminders

- **Quiz on multiple comparisons** is due before class.
- Homework 7 (single-factor experiments) is due this week on Blackboard.

Reminders

- **Quiz on multiple comparisons** is due before class.
- Homework 7 (single-factor experiments) is due this week on Blackboard.
- Office hours: Tuesdays 9–10 AM, Room 22-219 or Zoom.

Reminders

- **Quiz on multiple comparisons** is due before class.
- Homework 7 (single-factor experiments) is due this week on Blackboard.
- Office hours: Tuesdays 9–10 AM, Room 22-219 or Zoom.
- Read §8.10 and §8.11 to prepare for the next lecture.