Lecture 18: Analysis of Variance

STAT 515 — Comparing More Than Two Means

Dr. Yen-Yi Ho

Department of Statistics, University of South Carolina

Outline

1. Why not just run all the two-sample t tests?
2. The completely randomized design and the ANOVA model
3. Partitioning the variation: SST, SSE, and the ANOVA table
4. The F distribution and the ANOVA F test
5. Checking the assumptions
6. After a significant F: Tukey's comparisons
7. Summary: one-way ANOVA from start to finish

Outline

1. Why not just run all the two-sample t tests?
2. The completely randomized design and the ANOVA model
3. Partitioning the variation: SST, SSE, and the ANOVA table
4. The F distribution and the ANOVA F test
5. Checking the assumptions
6. After a significant F: Tukey's comparisons
7. Summary: one-way ANOVA from start to finish

Where We Are

- **Lectures 14–15:** one sample against a number — $\mu = \mu_0$, $p = p_0$, $\sigma^2 = \sigma_0^2$.
- **Lecture 16:** two samples — is $\mu_1 - \mu_2 = 0$?

Today

Many studies compare k **groups at once**: four packaging methods, five fertilizers, three teaching formats. The question is no longer about a single difference but about **a whole set of means**: are they all the same?

- The tool is the **analysis of variance (ANOVA)** — and despite the name, it is a test about **means**.
- The six-step recipe is unchanged. Only the test statistic is new.

A Motivating Study: Packaging and Bacterial Growth

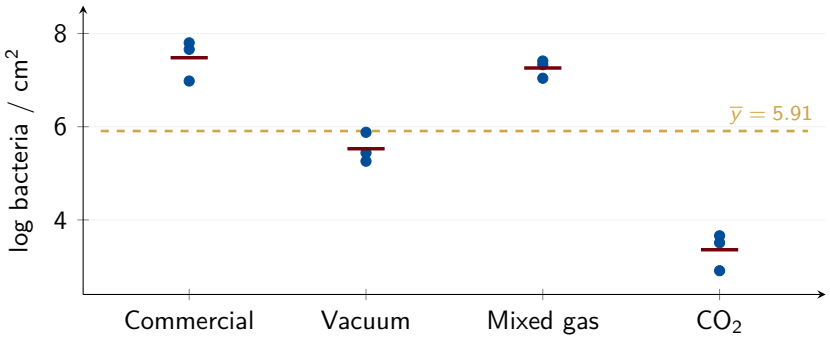
Nichols (1980), M.S. thesis, University of Arizona

How much do bacteria grow on steaks under different packaging? Steaks of about 75 grams each were **randomly assigned** to one of four packaging methods: commercial plastic wrap; vacuum packed; a mixed gas atmosphere of 1% CO, 40% CO_2, 59% N; or 100% CO_2. After nine days at 4°C, the **logarithm** of the number of psychrotrophic bacteria per cm^2 was recorded.

Packaging	**log bacteria per cm^2**		
Commercial	7.66	6.98	7.80
Vacuum	5.26	5.44	5.88
Mixed gas	7.41	7.33	7.04
CO_2	3.51	2.91	3.66

Response: log bacterial count. **Factor:** packaging, at $k = 4$ levels. $n_i = 3$, $n = 12$.

Look at the Data First



Garnet dashes are the four **group means**; the gold line is the **grand mean**. The groups clearly sit at different heights relative to how tightly the three points within each group cluster — **that comparison is exactly what ANOVA formalizes**.

The Tempting Shortcut: Test Every Pair

We know how to compare two means. Why not just do it six times?

$$\mu_1 = \mu_2, \quad \mu_1 = \mu_3, \quad \mu_1 = \mu_4, \quad \mu_2 = \mu_3, \quad \mu_2 = \mu_4, \quad \mu_3 = \mu_4$$

With k groups there are $\binom{k}{2} = \dfrac{k(k-1)}{2}$ pairs; with $k = 4$ that is **6** separate t tests.

The problem

Each test is run at $\alpha = .05$, so each carries a 5% chance of a **false positive**. Run six of them and the chance that **at least one** comes out significant by pure luck is far larger than 5%.

How Fast the Error Rate Inflates

If all k means really are equal and the tests were independent, then

$$P(\text{at least one false rejection}) = 1 - (1 - \alpha)^m, \qquad m = \binom{k}{2}$$

Groups k	**Comparisons** m	P(**at least one Type I error**)
2	1	0.05
3	3	0.14
4	6	**0.26**
5	10	0.40
10	45	0.90

With four groups, a procedure we believe is running at 5% is really running near 26%. This is the **multiple comparisons problem**, and the inflated rate is called the **familywise error rate**.

The Fix: One Test First

The rule

Ask one global question first. Test

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_k$$

with a **single** procedure at level α. Only if that test rejects do we go looking for *which* means differ — and then with a method that controls the familywise error rate.

- The single global test is the **ANOVA F test** (Sections 3–4).
- The follow-up is **Tukey's multiple comparisons** (Section 6).
- This two-stage structure — **omnibus test, then comparisons** — is the shape of almost every ANOVA you will ever see.

Outline

1. Why not just run all the two-sample t tests?
2. The completely randomized design and the ANOVA model
3. Partitioning the variation: SST, SSE, and the ANOVA table
4. The F distribution and the ANOVA F test
5. Checking the assumptions
6. After a significant F: Tukey's comparisons
7. Summary: one-way ANOVA from start to finish

The Completely Randomized Design

Completely randomized design

A design with a **single factor** at k levels (the **treatments**), where the treatments are **randomly assigned** to the experimental units.

- Equivalently: k **independent random samples**, one from each of k populations.
- In the steak study the factor is **type of packaging** and it has $k = 4$ treatments; the experimental units are the steaks.
- **Balanced** means all n_i are equal — as here, $n_i = 3$. ANOVA does not require balance, but balance makes it more robust.

A Motivating Study: Packaging and Bacterial Growth

Nichols (1980), M.S. thesis, University of Arizona

How much do bacteria grow on steaks under different packaging? Steaks of about 75 grams each were **randomly assigned** to one of four packaging methods: commercial plastic wrap; vacuum packed; a mixed gas atmosphere of 1% CO, 40% CO_2, 59% N; or 100% CO_2. After nine days at 4°C, the **logarithm** of the number of psychrotrophic bacteria per cm^2 was recorded.

Packaging	**log bacteria per cm^2**		
Commercial	7.66	6.98	7.80
Vacuum	5.26	5.44	5.88
Mixed gas	7.41	7.33	7.04
CO_2	3.51	2.91	3.66

Response: log bacterial count. **Factor:** packaging, at $k = 4$ levels. $n_i = 3$, $n = 12$.

Notation

Symbol	Meaning
y_{ij}	the jth observation under treatment i (sample i)
$\overline{y_i}$	the sample mean of sample i
s_i^2	the sample variance of sample i
n_i	the sample size of sample i
$\overline{y}$	the **grand mean** — the mean of all observations
$n = \sum_i n_i$	the total sample size
k	the number of treatments

The subscripts always read **group first, observation second**: y_{23} is the third steak in the vacuum group, $y_{23} = 5.88$.

The One-Way ANOVA Model

Model

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad i = 1, \ldots, k, \quad j = 1, \ldots, n_i, \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

- Every observation is its **group mean** plus **noise**.
- The noise has the **same variance σ^2 in every group** — one spread, shared by all k populations.
- ANOVA asks whether the μ_i are all the same. The noise term is what we must measure in order to answer that.
- Writing $\mu_i = \mu + \tau_i$ gives the equivalent "**grand mean plus treatment effect**" form; then H_0 says every $\tau_i = 0$.

The Assumptions Required for ANOVA

Three assumptions

- The k samples are **independent** of each other, and are random samples from their populations.
- Each sample comes from a **normal population**.
- The population variances are **equal**: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$.

- These are exactly the **pooled t** assumptions from Lecture 16, extended from $k = 2$ to general k.
- **Independence is the one that cannot be repaired after the fact** — it is a property of how the data were collected.
- Normality and equal variance are checked with **residual plots** (Section 5).

The Hypotheses

One-way ANOVA hypotheses

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_k \qquad \text{(all } k \text{ treatment means are equal)}$$

$$H_a: \ \text{at least two of the means are different}$$

- In plain words: H_0 says the **factor has no effect** on the response; H_a says the factor **has some effect**.
- H_a **is not "all the means are different."** One group out of line is enough to make H_a true.
- **There is no "$<$" or "$>$" version.** With more than two means there is no single direction to point in, so ANOVA has **one alternative only**.

Outline

1. Why not just run all the two-sample t tests?
2. The completely randomized design and the ANOVA model
3. Partitioning the variation: SST, SSE, and the ANOVA table
4. The F distribution and the ANOVA F test
5. Checking the assumptions
6. After a significant F: Tukey's comparisons
7. Summary: one-way ANOVA from start to finish

The Central Idea

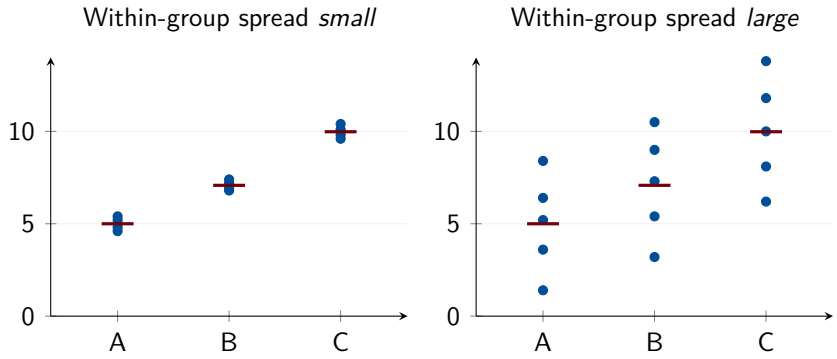
Analysis of variance

Compare the variation **between** the treatment groups with the variation **within** the treatment groups.

- **Within-group** variation is pure noise: steaks given the *same* treatment still differ. It estimates σ^2.
- **Between-group** variation is noise *plus* any real differences among the μ_i.
- If the between-group variation is **much larger**, then it favors H_a.

We test a hypothesis about means by analyzing variances — which is where the name comes from.

The Same Means, Two Different Verdicts



The three **group means are identical** in both panels, so the **between**-group variation is identical. On the left the groups barely overlap and the separation is convincing; on the right it could easily be noise. **Group means alone cannot settle the question — the within-group spread is the yardstick they must be measured against.**

SST — Sum of Squares for Treatment

Between-group variation

$$SST = \sum_{i=1}^{k} n_i \left(\overline{y}_i - \overline{y}\right)^2, \qquad df_T = k - 1$$

- Each group contributes the squared distance from **its own mean** to the **grand mean**, weighted by its sample size.
- $SST = 0$ exactly when all the group means are equal. The larger SST, the more spread out the group means are.
- $df_T = k - 1$**:** once you know the grand mean and $k - 1$ of the group means, the last one is determined.

SSE — Sum of Squares for Error

Within-group variation

$$SSE = \sum_{i=1}^{k} (n_i - 1)\, s_i^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y_i} \right)^2, \qquad df_E = n - k$$

- Each observation is compared to **its own group's mean**, so group differences are swept out entirely — what is left is random error.
- $df_E = (n_1 - 1) + \cdots + (n_k - 1) = n - k$**:** each group donates $n_i - 1$ degrees of freedom, exactly as in the pooled t.

Total sum of squares splits in two

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}\right)^2}_{TSS} = \underbrace{\sum_{i=1}^{k} n_i\left(\overline{y_i}-\overline{y}\right)^2}_{SST} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y_i}\right)^2}_{SSE}$$

And the degrees of freedom split the same way:

$$\underbrace{n-1}_{df_{\text{total}}} = \underbrace{(k-1)}_{df_T} + \underbrace{(n-k)}_{df_E}$$

Every observation's distance from the grand mean is $(y_{ij}-\overline{y}) = (\overline{y_i}-\overline{y}) + (y_{ij}-\overline{y_i})$: **how far my group is from the center**, plus **how far I am from my group**. Squaring and summing, the cross-product term vanishes and the two pieces separate cleanly.

Mean Squares

A sum of squares divided by its degrees of freedom is a **mean square** — that is, a variance.

MST and MSE

$$MST = \frac{SST}{df_T} = \frac{SST}{k-1} \qquad MSE = \frac{SSE}{df_E} = \frac{SSE}{n-k}$$

- MSE **always** estimates σ^2, whether H_0 is true or not.
- MST estimates σ^2 **only if H_0 is true**. If the μ_i differ, MST estimates something larger.
- **Mean squares do not add:** $MST + MSE \neq TSS/(n-1)$. Only the **sums of squares** and the **degrees of freedom** add.

MSE Is the Pooled Variance from Lecture 16

$$MSE = \frac{SSE}{n-k} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_k-1)s_k^2}{n_1 + n_2 + \cdots + n_k - k}$$

A familiar face

Set $k = 2$ and this is exactly

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2},$$

the **pooled sample variance** from the two-sample t test. ANOVA's MSE is the same weighted average of the group variances, extended to k groups.

This is why the equal-variance assumption appears in both places — in both procedures we pool all the groups into **one** estimate of a **single** σ^2.

The Test Statistic

The ANOVA F statistic

$$F = \frac{MST}{MSE} = \frac{\text{variation } \textit{between} \text{ treatments}}{\text{variation } \textit{within} \text{ treatments}}$$

- If H_0 is true, numerator and denominator both estimate σ^2, so F should land **near 1**.
- If H_0 is false, the numerator is inflated by the real differences and F grows.
- **The larger the F, the stronger the evidence against H_0.** Only **large** values of F are evidence against H_0 — which is why the test is **upper-tailed**.

The One-Way ANOVA Table

Source	SS	DF	MS	F
Between (Treatment)	$SST = \sum_{i=1}^{k} n_i(\overline{y_i} - \overline{y})^2$	$k-1$	$MST = \frac{SST}{k-1}$	$\frac{MST}{MSE}$
Within (Error)	$SSE = \sum_{i=1}^{k}(n_i - 1)s_i^2$	$n-k$	$MSE = \frac{SSE}{n-k}$	
Total	$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$	$n-1$		

Reading the table

The first two rows add to the third — in the SS column and in the DF column. That is the arithmetic check: if your rows do not add, you have made a mistake.

Outline

1. Why not just run all the two-sample t tests?
2. The completely randomized design and the ANOVA model
3. Partitioning the variation: SST, SSE, and the ANOVA table
4. The F distribution and the ANOVA F test
5. Checking the assumptions
6. After a significant F: Tukey's comparisons
7. Summary: one-way ANOVA from start to finish

The F Distribution

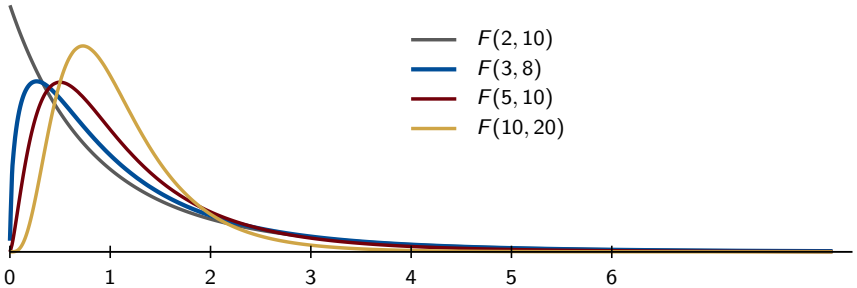
Fact

Under the ANOVA assumptions and when H_0 is true,

$$F = \frac{MST}{MSE} \sim F\Big(df_n = k - 1, \quad df_d = n - k\Big)$$

- The F distribution has **two** degrees of freedom: a **numerator** df and a **denominator** df, in that order.
- It is a distribution of a **ratio of two variances**, so it lives on $[0, \infty)$ — **never negative**.
- It is **right-skewed**, and centered near 1 (its mean is $df_d/(df_d - 2)$).

What the F Curves Look Like

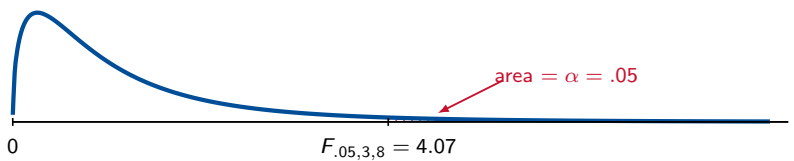


All F curves start at 0, are skewed to the right, and have a long upper tail. As **both** degrees of freedom grow the curve becomes more symmetric and piles up tightly around 1 — so a given F value is more surprising with large samples than with small ones.

Critical Values: $F_{\alpha,\, df_n,\, df_d}$

Notation

$F_{\alpha,\, df_n,\, df_d}$ is the value with **area α to its right** under the $F(df_n, df_d)$ curve.



Order matters: $F_{.05,3,8} = 4.07$ but $F_{.05,8,3} = 8.85$. In R: `qf(.95, 3, 8)` — note that `qf` counts area to the **left**, so the α in our notation becomes $1 - \alpha$ there.

The ANOVA F Test

Test of H_0 : $\mu_1 = \mu_2 = \cdots = \mu_k$

$$F = \frac{MST}{MSE}, \qquad RR = \left\{ F > F_{\alpha,\ df_n = k-1,\ df_d = n-k} \right\}$$

$$p\text{-value} = P_{H_0}\Big(F > F_{calc}\Big)$$

- **Always upper-tailed**, with the whole of α in the right tail — there is no $\alpha/2$ here, even though H_a is "the means differ."
- Reject H_0 when $F > F_{\alpha,k-1,n-k}$, equivalently when $p < \alpha$.
- A small F (near or below 1) means the groups are no further apart than the noise would predict: **fail to reject**.

Example 1: Computing the Pieces

$$n = 12, \qquad k = 4, \qquad n_i = 3, \qquad \overline{y} = 5.907$$

Group i	n_i	$\overline{y_i}$	s_i	s_i^2
1 Commercial	3	7.480	0.4386	0.1924
2 Vacuum	3	5.527	0.3190	0.1017
3 Mixed gas	3	7.260	0.1947	0.0379
4 CO_2	3	3.360	0.3969	0.1575

$$SST = 3\Big[(7.480 - 5.907)^2 + (5.527 - 5.907)^2 + (7.260 - 5.907)^2 + (3.360 - 5.907)^2\Big] = \mathbf{32.810}$$

$$SSE = 2\Big(0.1924 + 0.1017 + 0.0379 + 0.1575\Big) = \mathbf{0.979}$$

Example 1: The ANOVA Table

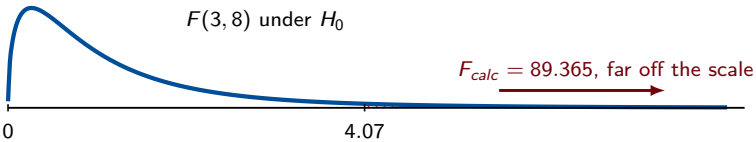
Source	SS	DF	MS	F
Packaging (Treatment)	32.810	3	10.9368	**89.365**
Error	0.979	8	0.1224	
Total	33.789	11		

$$MST = \frac{32.810}{4-1} = 10.9368, \quad MSE = \frac{0.979}{12-4} = 0.1224, \quad F = \frac{10.9368}{0.1224} = 89.4$$

Check the additivity: $32.810 + 0.979 = 33.789$ and $3 + 8 = 11$. Also note $\sqrt{MSE} = 0.350$ — the pooled estimate of σ, on the log-count scale.

Example 1: Decision

$$RR = \{ F > F_{.05,\ 3,\ 8} \} = \{ F > 4.07 \}$$



Since $89.365 > 4.07$, the statistic falls in the rejection region: **reject** H_0 at $\alpha = .05$.

$$p\text{-value} = P\Big(F_{3,8} > 89.365\Big) = 1.714 \times 10^{-6} \ll .05$$

Conclusion

At the .05 level there is sufficient evidence that **at least two** of the four packaging methods lead to different mean log bacterial counts.

Example 1 in R

```
y    <- c(7.66,6.98,7.80, 5.26,5.44,5.88,
          7.41,7.33,7.04, 3.51,2.91,3.66)
grp <- factor(c(rep("Comm",3), rep("Vac",3),
                rep("Mixed",3), rep("CO2",3)))
anova(lm(y ~ grp))
```

```
Analysis of Variance Table

Response: y
          Df  Sum Sq  Mean Sq  F value     Pr(>F)
grp        3  32.810  10.9368   89.365  1.714e-06 ***
Residuals  8   0.979   0.1224
```

`grp` is the **treatment** row and `Residuals` is the **error** row. R omits the Total row — add the two rows yourself when you need it. Note R does *not* print the hypotheses: **you still have to state them.**

Outline

1. Why not just run all the two-sample t tests?
2. The completely randomized design and the ANOVA model
3. Partitioning the variation: SST, SSE, and the ANOVA table
4. The F distribution and the ANOVA F test
5. Checking the assumptions
6. After a significant F: Tukey's comparisons
7. Summary: one-way ANOVA from start to finish

Residuals: What We Actually Check

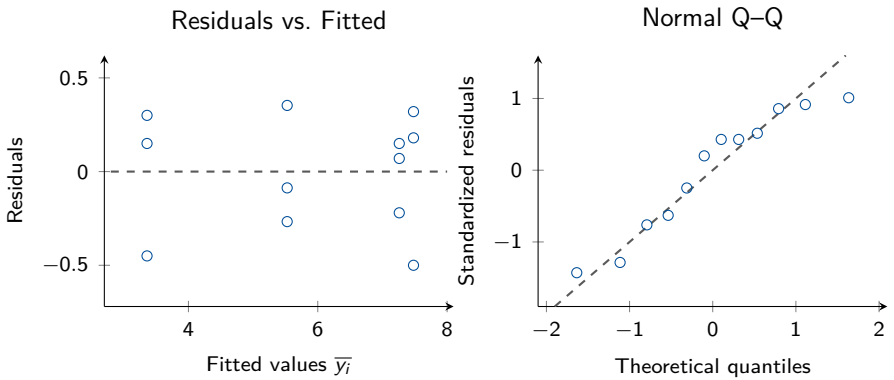
We cannot see ε_{ij}, but we can estimate it.

Fitted values and residuals

$$\widehat{y}_{ij} = \overline{y_i} \qquad e_{ij} = y_{ij} - \widehat{y}_{ij} = y_{ij} - \overline{y_i}$$

- The model's prediction for *every* observation in group i is simply that group's mean.
- The **residual** is what the model missed. The assumptions "normal, constant variance" are statements about the ε_{ij}, so we check them on the e_{ij}.
- Two standard plots: **residuals vs. fitted values** (equal variance) and a **Q–Q plot** of the residuals (normality).

Example 1: The Two Diagnostic Plots



Left: the vertical spread is about the same above each of the four fitted values — the **equal-variance assumption holds up**. **Right:** the points follow the reference line reasonably closely — **the normality assumption is acceptable**. With $n = 12$ we cannot ask for more than "no obvious violation."

Reading the Diagnostic Plots

What you see	What it means
Residuals vs. fitted: even band	equal variance is fine
Residuals vs. fitted: funnel / megaphone	variance grows with the mean
Residuals vs. fitted: curvature	the model is missing structure
Q–Q plot: points near the line	normality is fine
Q–Q plot: curved ends	skewness
Q–Q plot: S-shape	heavy or light tails

If an assumption fails

A transformation often fixes both problems at once — a funnel shape and right skewness usually disappear together under a log or square root. The steak response was *already* recorded on the **log** scale, for exactly this reason.

How Much Do the Assumptions Matter?

- **Normality:** the F test is **robust**. Mild to moderate non-normality barely affects it, especially with equal sample sizes.
- **Equal variance:** matters more, and matters most when the **sample sizes are unequal**. A common rule of thumb is to worry once the largest s_i exceeds about **twice** the smallest.
- **Independence: not robust at all.** If observations are paired, clustered, or repeated on the same subject, one-way ANOVA is simply the wrong procedure and no amount of sample size repairs it.

In Example 1 the four standard deviations are 0.44, 0.32, 0.19, 0.40 — the largest is about 2.3 times the smallest, which looks alarming until you remember each is based on only 3 observations. With n_i all equal, the F test tolerates this comfortably.

Outline

1. Why not just run all the two-sample t tests?
2. The completely randomized design and the ANOVA model
3. Partitioning the variation: SST, SSE, and the ANOVA table
4. The F distribution and the ANOVA F test
5. Checking the assumptions
6. After a significant F: Tukey's comparisons
7. Summary: one-way ANOVA from start to finish

A Significant F Does Not Say Which

We rejected H_0 in Example 1. All that tells us is that *somewhere* among the four means there is a difference.

- Is CO_2 different from vacuum? Is commercial different from mixed gas? **The F test is silent on all of it.**
- Going back to six separate t tests would reintroduce exactly the error-rate inflation that sent us to ANOVA in the first place.

Tukey's Honest Significant Difference (HSD)

A procedure that forms a confidence interval for **every** pairwise difference $\mu_i - \mu_{i'}$, with the intervals widened just enough that the **familywise** confidence level across the whole set is 95%.

Tukey's HSD

Simultaneous intervals, balanced design ($n_i = n_0$)

$$(\overline{y_i} - \overline{y_{i'}}) \pm \underbrace{q_{\alpha,\,k,\,n-k}\sqrt{\frac{MSE}{n_0}}}_{\text{HSD}}$$

- $q_{\alpha,k,n-k}$ comes from the **studentized range** distribution — it depends on how many groups are being compared, which is how the procedure pays for the multiplicity.
- *MSE* is taken from the ANOVA table: **all k groups contribute to the standard error**, not just the two being compared.
- Declare $\mu_i \neq \mu_{i'}$ when the interval **excludes** 0, equivalently when the adjusted p-value is below α.

Example 1: Tukey Output

```
fit <- aov(y ~ grp);  TukeyHSD(fit, conf.level = .95)
```

Comparison	diff	lwr	upr	p adj
Comm − CO_2	4.120	3.205	5.035	0.0000025
Mixed − CO_2	3.900	2.985	4.815	0.0000038
Vac − CO_2	2.167	1.252	3.081	0.0002935
Mixed − Comm	−0.220	−1.135	0.695	**0.8656902**
Vac − Comm	−1.953	−2.868	−1.039	0.0006036
Vac − Mixed	−1.733	−2.648	−0.819	0.0013498

Every HSD here is $q_{.05,4,8}\sqrt{MSE/3} = 4.529\sqrt{0.1224/3} = 0.915$ wide on each side — the design is balanced, so all six intervals have the same half-width. **Five intervals exclude 0; one does not.**

Example 1: The Underline Display

Sort the group means and **underline groups that are not significantly different**:

CO_2	Vacuum	Mixed gas	Commercial
3.36	5.53	7.26	7.48

Conclusion

CO_2 gives a significantly **lower** mean log bacterial count than every other method, and vacuum packing is significantly lower than both mixed gas and commercial wrap. **Mixed gas and commercial wrap cannot be distinguished** from each other.

Example 2: Golf Ball Brands

Example 2

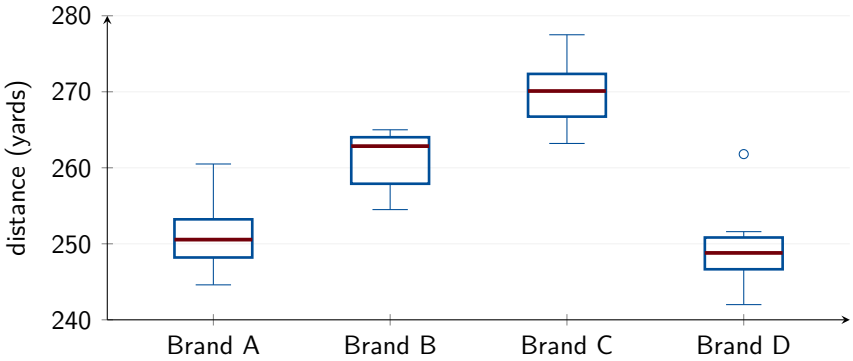
The USGA wants to compare the mean distances reached by four brands of golf ball struck with a driver. A completely randomized design is used: *Iron Byron*, the USGA's robotic golfer, hits a random sequence of 10 balls of each brand, and the distance (yards) is recorded for each hit.

Brand A	251.2	245.1	248.0	251.1	260.5	250.0	253.9	244.6	254.6	248.8
Brand B	263.2	262.9	265.0	254.5	264.3	257.0	262.8	264.4	260.6	255.9
Brand C	269.7	263.2	277.5	267.4	270.5	265.5	270.7	272.9	275.6	266.5
Brand D	251.6	248.6	249.4	242.0	246.5	251.3	261.8	249.0	247.1	245.9

Response: distance travelled. **Factor:** brand, at $k = 4$ levels. $n_i = 10$, $n = 40$.

$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ (brand has no effect on distance) vs. H_a : at least two means differ.

Example 2: Side-by-Side Boxplots



Brand C is clearly the longest and A and D look similar. The four boxes have **comparable heights**, which is the quick visual check on the equal-variance assumption.

Example 2: ANOVA Table and Decision

```
Analysis of Variance Table

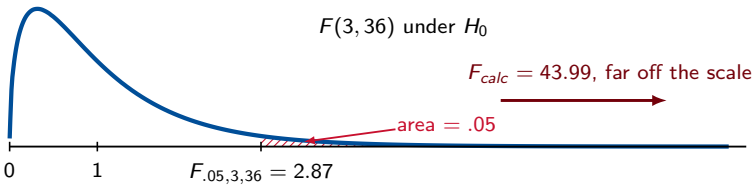
Response: distance
          Df  Sum Sq  Mean Sq  F value     Pr(>F)
brand      3  2794.4   931.47   43.989  3.973e-12 ***
Residuals 36   762.3    21.18
```

$$F = \frac{SST/(k-1)}{SSE/(n-k)} = \frac{2794.4/3}{762.3/36} = \frac{931.47}{21.18} = \mathbf{43.99}, \qquad df_n = 3, \quad df_d = 36$$

$$RR = \{\, F > F_{.05,3,36} \,\} = \{\, F > 2.87 \,\}$$

Since $43.99 > 2.87$ (and $p = 3.97 \times 10^{-12} < .05$) we **reject** H_0: at least two brands give significantly different mean distances.

Example 2: Where the F Statistic Falls



Compare this curve with the $F(3, 8)$ earlier: same numerator df, but 36 denominator degrees of freedom instead of 8 pull the critical value down from 4.07 to 2.87. **More data, less tolerance for large F values under H_0.**

Example 2: Tukey Comparisons

Comparison	diff	lwr	upr	p adj
B − A	10.28	4.738	15.822	0.0000869
C − A	19.17	13.628	24.712	0.0000000
D − A	−1.46	−7.002	4.082	**0.8926914**
C − B	8.89	3.348	14.432	0.0006522
D − B	−11.74	−17.282	−6.198	0.0000100
D − C	−20.63	−26.172	−15.088	0.0000000

Brand D	Brand A	Brand B	Brand C
249.3	250.8	261.1	270.0

Brands D and A are indistinguishable; every other pair differs significantly. Brand C travels farthest, by about 9 yards over brand B.

Outline

1. Why not just run all the two-sample t tests?
2. The completely randomized design and the ANOVA model
3. Partitioning the variation: SST, SSE, and the ANOVA table
4. The F distribution and the ANOVA F test
5. Checking the assumptions
6. After a significant F: Tukey's comparisons
7. Summary: one-way ANOVA from start to finish

The Whole Procedure, in Order

1. **Check the design.** One factor, k levels, independent groups? Then one-way ANOVA.
2. **State the hypotheses.** $H_0 : \mu_1 = \cdots = \mu_k$ versus H_a : at least two differ.
3. **Build the ANOVA table.** SST, SSE, df, MST, MSE.
4. **Compute** $F = MST/MSE$ and compare to $F_{\alpha,k-1,n-k}$, or compute the p-value.
5. **Check the residuals** — residuals vs. fitted, and Q–Q.
6. **If and only if** H_0 **is rejected**, run Tukey's HSD and report which pairs differ.
7. **State the conclusion in context**, in the units of the response.

Choosing the Right Procedure

Design	Groups	Statistic	Reference
Independent, large n	2	$\dfrac{(\overline{y_1} - \overline{y_2}) - D_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	$N(0,1)$
Independent, small n, $\sigma_1^2 = \sigma_2^2$	2	$\dfrac{(\overline{y_1} - \overline{y_2}) - D_0}{\sqrt{s_p^2(1/n_1 + 1/n_2)}}$	$t_{n_1+n_2-2}$
Independent, small n, $\sigma_1^2 \neq \sigma_2^2$	2	$\dfrac{(\overline{y_1} - \overline{y_2}) - D_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	$t_{\text{Satterthwaite}}$
Paired	2	$\dfrac{\overline{d} - \mu_{d_0}}{s_d/\sqrt{n}}$	t_{n-1}
Completely randomized	$k \geq 2$	$\dfrac{MST}{MSE}$	$F_{k-1,\,n-k}$

A reassuring fact

With $k = 2$ the ANOVA F test is the pooled t test: $F = t^2$, and the two p-values agree exactly — ANOVA extends Lecture 16 rather than replacing it.

Common Mistakes to Avoid

- **Running all the pairwise t tests** instead of one F test. With $k = 4$ that inflates the error rate from 5% to 26%.
- **Reading H_a as "all the means differ."** It says at least two do.
- **Putting $\alpha/2$ in each tail of the F curve.** The ANOVA F test is upper-tailed; all of α goes on the right.
- **Swapping the two degrees of freedom.** Numerator $df = k - 1$ comes first; $F_{.05,3,8} \neq F_{.05,8,3}$.
- **Running Tukey after a non-significant F**, or skipping Tukey and declaring which group is best from the F test alone.
- **Using one-way ANOVA on repeated measures.** Paired or clustered data break the independence assumption.
- **Forgetting that the SS and df columns must add.**

Summary

- With k groups we ask **one** global question first, because many pairwise tests inflate the familywise error rate.
- The total variation **partitions**: $TSS = SST + SSE$, with $n - 1 = (k - 1) + (n - k)$.
- MSE always estimates σ^2; MST estimates σ^2 *only* under H_0. Their ratio $F = MST/MSE$ is therefore near 1 under H_0 and **large** when the means differ.
- Reject when $F > F_{\alpha,\, k-1,\, n-k}$. **The test is always upper-tailed.**
- Assumptions — independent samples, normal populations, equal variances — are checked on the **residuals**.
- A significant F is a beginning, not an end: **Tukey's HSD** says which pairs actually differ, with the familywise confidence level protected.