ANOVA test

One-Way ANOVA

- One-way analysis of variance (ANOVA) is a statistical method for testing for differences in the means of **three or more groups**.
- One-way ANOVA is typically used when you have a single independent variable or factor. The independent variable divides cases into two or more mutually exclusive levels, categories, or groups. Your goal is to investigate if variations or different levels of that factor have a measurable effect on a dependent variable.
- One-way ANOVA can only be used when investigating a single factor and a single dependent variable. When comparing the means of three or more groups**, it can tell us if at least one pair of means is significantly different**, **but it can't tell us which pair**.

One-Way ANOVA (cont..)

- One-way ANOVA is a test for differences in group means
- One-way ANOVA is a statistical method to test the null hypothesis (H_0) that three or more population means are equal vs. the alternative hypothesis (H_a) that at least one mean is different.

Examples

- 1# Your independent variable is social media use, and you assign groups to low, medium, and high levels of social media use to find out if there is a difference in hours of sleep per night.
- 2# Your independent variable is brand of soda, and you collect data on Coke, Pepsi, Sprite, and Fanta to find out if there is a difference in the price per 100ml.
- 3# You independent variable is type of fertilizer, and you treat crop fields with mixtures 1, 2 and 3 to find out if there is a difference in crop yield.

Example 2

- Imagine you work for a company that manufactures an adhesive gel that is sold in small jars. The viscosity of the gel is important: too thick and it becomes difficult to apply; too thin and its adhesiveness suffers. You've received some feedback from a few unhappy customers lately complaining that the viscosity of your adhesive is not as consistent as it used to be. You've been asked by your boss to investigate.
- You decide that a good first step would be to examine the average viscosity of the five most recent production lots. If you find differences between lots, that would seem to confirm the issue is real. It might also help you begin to form hypotheses about factors that could cause inconsistencies between lots.

You measure viscosity using an instrument that rotates a spindle immersed in the jar of adhesive. This test yields a measurement called torque resistance. You test five jars selected randomly from each of the most recent five lots

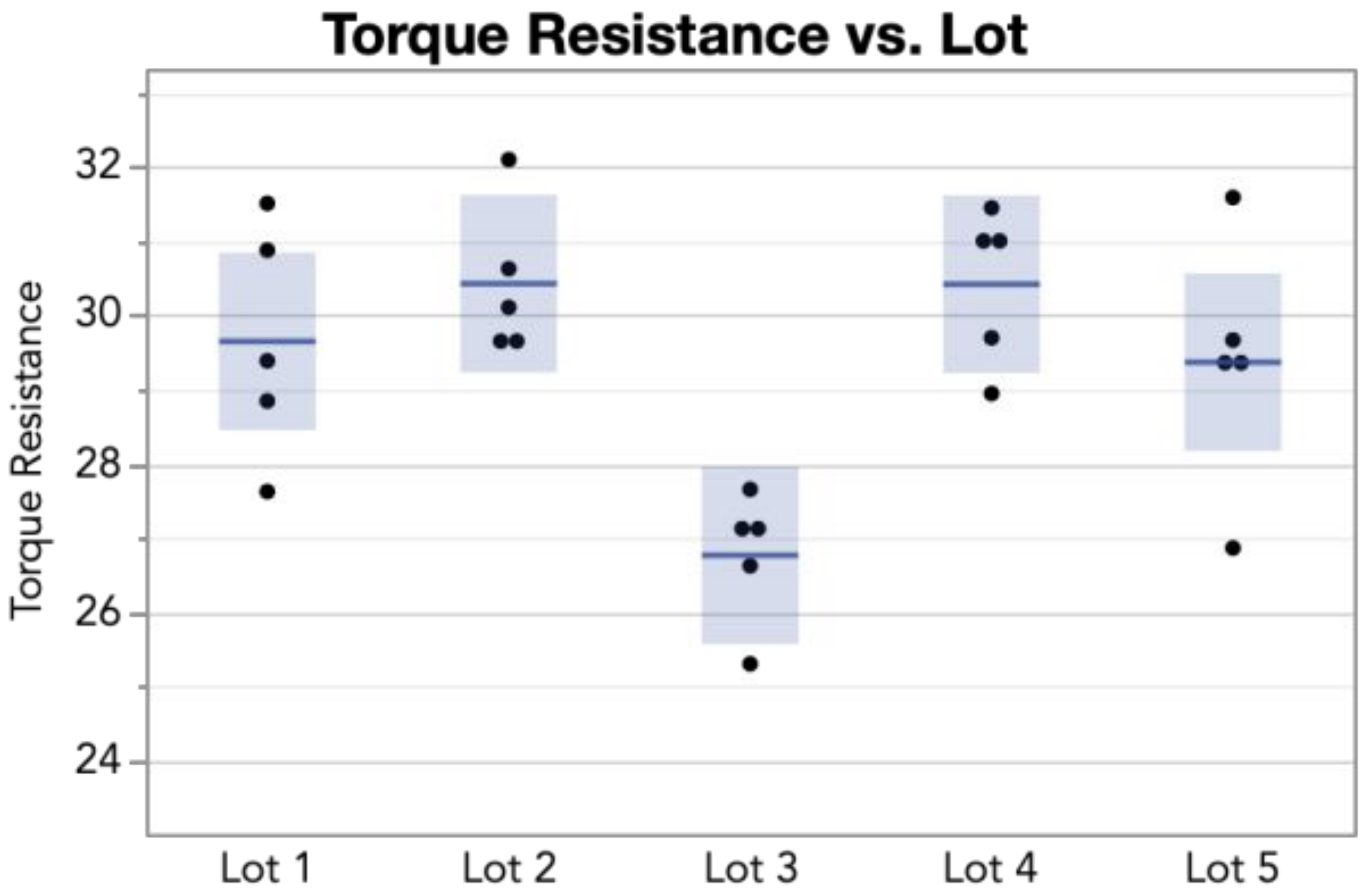

Table 1: Mean torque measurements from tests of five lots of adhesive

Lot #	N	Mean
1	5	29.65
2	5	30.43
3	5	26.77
4	5	30.42
5	5	29.37

Table 2: ANOVA table with results from our torque measurements

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob-> F
Lot	4	45.25	11.31	6.90	0.0012
Error	20	32.80	1.64		
Total	24	78.05			

- One key element in this table to focus on for now is the p-value. The p-value is used to evaluate the validity of the null hypothesis that all the means are the same. In our example, the p-value (Prob-> F) is 0.0012. This small p-value can be taken as evidence that the means are not all the same
- Remember, an ANOVA test will not tell you which mean or means differs from the others, and (unlike our example) this isn't always obvious from a plot of the data

- ***Hypothesis****: the viscosity of the adhesive is not consistent among different lots*
- ***Null Hypothesis:*** *the viscosity of the adhesive is consistent in all lots*

One-way ANOVA calculation

Table 3: Torque measurements by Lot

	Lot 1	Lot 2	Lot 3	Lot 4	Lot 5
Jar 1	29.39	30.63	27.16	31.03	29.67
Jar 2	31.51	32.10	26.63	30.98	29.32
Jar 3	30.88	30.11	25.31	28.95	26.87
Jar 4	27.63	29.63	27.66	31.45	31.59
Jar 5	28.85	29.68	27.10	29.70	29.41
Mean	**29.65**	**30.43**	**26.77**	**30.42**	**29.37**

n_i = Number of observations for treatment i (in our example, Lot i)

N = Total number of observations

Y_{ij} = The jth observation on the ith treatment

$\overline{Y}_i$ = The sample mean for the ith treatment

$\overline{\overline{Y}}$ = The mean of all observations (grand mean)

Sum of Squares

- The sum of squares gives us a way to quantify variability in a data set by focusing on the difference between each data point and the mean of all data points in that data set. The formula below partitions the overall variability into **two parts**: the variability due to the model or the factor levels, and the variability due to random error.

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij}-\overline{\overline{Y}})^2 = \sum_{i=1}^{a} n_i(\overline{Y}_i-\overline{\overline{Y}})^2+\sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2$$

$$SS(Total) = SS(Factor) + SS(Error)$$

Lot	Y_{ij}	$\bar{Y}_i$	$\bar{\bar{Y}}$	$\bar{Y}_i - \bar{\bar{Y}}$	$Y_{ij} - \bar{\bar{Y}}$	$Y_{ij} - \bar{Y}_i$	$(\bar{Y}_i - \bar{\bar{Y}})^2$	$(Y_{ij} - \bar{Y}_i)^2$	$(Y_{ij} - \bar{\bar{Y}})^2$
1	29.39	29.65	29.33	0.32	0.06	-0.26	0.10	0.07	0.00
1	31.51	29.65	29.33	0.32	2.18	1.86	0.10	3.46	4.75
1	30.88	29.65	29.33	0.32	1.55	1.23	0.10	1.51	2.40
1	27.63	29.65	29.33	0.32	-1.70	-2.02	0.10	4.08	2.89
1	28.85	29.65	29.33	0.32	-0.48	-0.80	0.10	0.64	0.23
2	30.63	30.43	29.33	1.10	1.30	0.20	1.21	0.04	1.69
2	32.10	30.43	29.33	1.10	2.77	1.67	1.21	2.79	7.68
2	30.11	30.43	29.33	1.10	0.78	-0.32	1.21	0.10	0.61

⋮

4	31.45	30.42	29.33	1.09	2.12	1.03	1.19	1.06	4.49
4	29.70	30.42	29.33	1.09	0.37	-0.72	1.19	0.52	0.14
5	29.67	29.37	29.33	0.04	0.34	0.30	0.00	0.09	0.12
5	29.32	29.37	29.33	0.04	-0.01	-0.05	0.00	0.00	0.00
5	26.87	29.37	29.33	0.04	-2.46	-2.50	0.00	6.26	6.05
5	31.59	29.37	29.33	0.04	2.26	2.22	0.00	4.93	5.11
5	29.41	29.37	29.33	0.04	0.08	0.04	0.00	0.00	0.01
Sum of Squares							SS (Factor) = 45.25	SS (Error) = 32.80	SS (Total) = 78.05

Degrees of Freedom (DF)

- The number of values that are free to vary in a data set
- The degrees of freedom indicates the number of independent pieces of information used to calculate each sum of squares.
- For a one-factor design with a factor at k levels (five lots in our example) and a total of N observations (five jars per lot for a total of 25), the degrees of freedom are as follows:

	Degrees of Freedom (DF) Formula	Calculated Degrees of Freedom
SS (Factor)	k - 1	5 - 1 = 4
SS (Error)	N - k	25 - 5 = 20
SS (Total)	N - 1	25 - 1 = 24

Mean Squares (MS) and F Ratio

- We divide each sum of squares by the corresponding degrees of freedom to obtain mean squares (**MS)**. When the null hypothesis is true (i.e. the means are equal), **MS (Factor) and MS (Error) are both would be about the same size**. **Their ratio, or the F ratio, would be close to one.** When the null hypothesis is not true then the MS (Factor) will be larger than MS (Error) and **their ratio greater than 1**. In our adhesive testing example, the computed F ratio, 6.90, presents significant evidence **against** the null hypothesis that the means are equal.

Table 6: Calculating mean squares and F ratio

	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Squares	F Ratio
SS (Factor)	45.25	4	45.25/4 = 11.31	11.31/1.64 = 6.90
SS (Error)	32.80	20	32.80/20 = 1.64	

- The ratio of MS(factor) to MS(error)—the F ratio—has an F distribution. The F distribution is the distribution of F values that we'd expect to observe when the null hypothesis is true (i.e. the means are equal). F distributions have different shapes based on two parameters, called the numerator and denominator degrees of freedom. For an ANOVA test, the numerator is the MS(factor), so the degrees of freedom are those associated with the MS(factor). The denominator is the MS(error), so the denominator degrees of freedom are those associated with the MS(error).
- If your computed F ratio exceeds the expected value from the corresponding F distribution, then, assuming a sufficiently small p-value, you would reject the null hypothesis that the means are equal. **The p-value in this case is the probability of observing a value greater than the F ratio from the F distribution when in fact the null hypothesis is true.**