A warmup

The following data show the number of job offers that a sample of EC students had at the time of graduation, together with their major field of study

Data science	Biology	Economics
2	3	2
5	0	2
1	1	0
	2	
$n = 3$	$n = 4$	$n = 3$
$\bar{y} = 2.67$	$\bar{y} = 1.5$	$\bar{y} = 1.33$

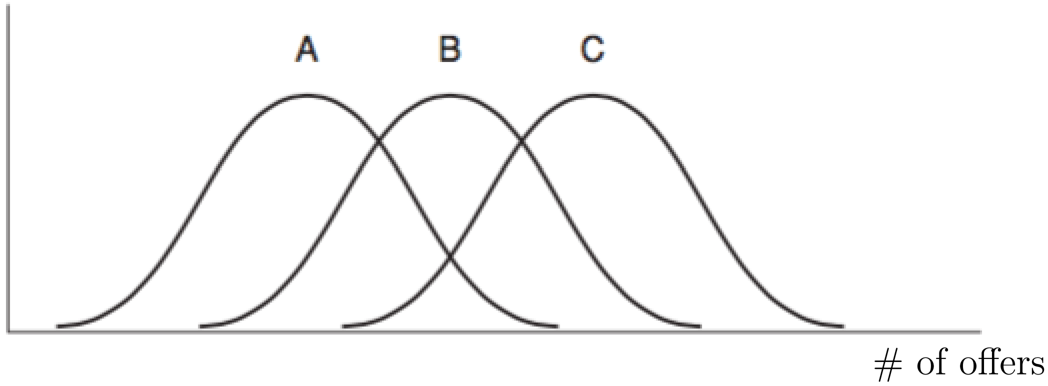
Grand mean = 1.8

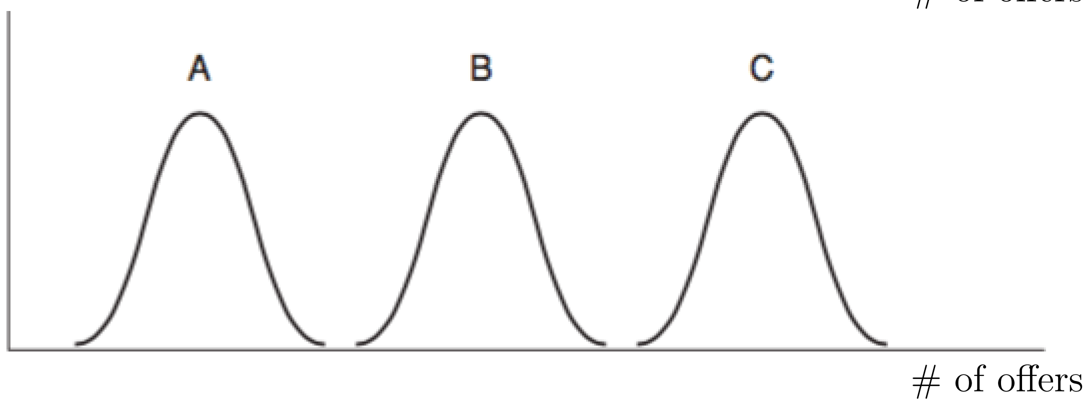
(here "grand mean" refers to the mean of the entire sample)

- **Q**: Is there a statistically significant difference between the group means?
- ANOVA (ANalysis Of VAriance) is the branch of inferential stats that deals with such questions.
- Before getting into details, let's develop some intuition for ANOVA.

Intuition for ANOVA

The following graphs show two different hypothetical distributions of number of job offers for the student groups





1. Which graph makes the difference seem more significant?
2. Why? Think in terms of variability (or variances).

ANOVA Intuition:

- How much variability is seen **within** each group?
- How much variability is seen **across** different groups?
- Which will be higher if the groups have significantly different means?
- Consider the ratio: $F = \dfrac{var_{\text{across}}}{var_{\text{within}}}$

  What would $F > 1$ suggest? What about $F \approx 0$?

Warmup exercise (continued)

Here is the data again (number of job offers at the time of graduation)

Data science	Biology	Economics
2	3	2
5	0	2
1	1	0
	2	
$n = 3$	$n = 4$	$n = 3$
$\bar{y} = 2.67$	$\bar{y} = 1.5$	$\bar{y} = 1.33$

Grand mean $= 1.8$

1. For each group, compute the **within groups** sum of squares: $\sum(y_i - \bar{y})^2$.
2. Add those 3 numbers to get: SS_W
   This is the net sum of squares within the groups.
3. To find the sum of squares **between groups**, compute:
   $$n(\bar{y} - \text{grand mean})^2 \quad \text{for each group.}$$
   Add those 3 numbers to get SS_B.
4. Next, we want to turn those sums of squares into variances, by dividing by an appropriate number. Those variances are often called "mean squares" or "MS" in ANOVA.
5. For the **within groups** mean squares, compute: $MS_W = \frac{SS_w}{10-3}$
6. For the **between groups** mean squares, compute: $MS_B = \frac{SS_B}{3-1}$
7. Finally, compute the F ratio: $F = \frac{MS_B}{MS_W}$

Remarks:

ANOVA is, essentially, a hypothesis test with null hypothesis saying there is no difference between the group means.

We look up the P-value corresponding to the computed F statistic.