ANOVA:
ANalysis Of VAriance

Mana Azizsoltani
HOA 730
Spring 2026

UNLV

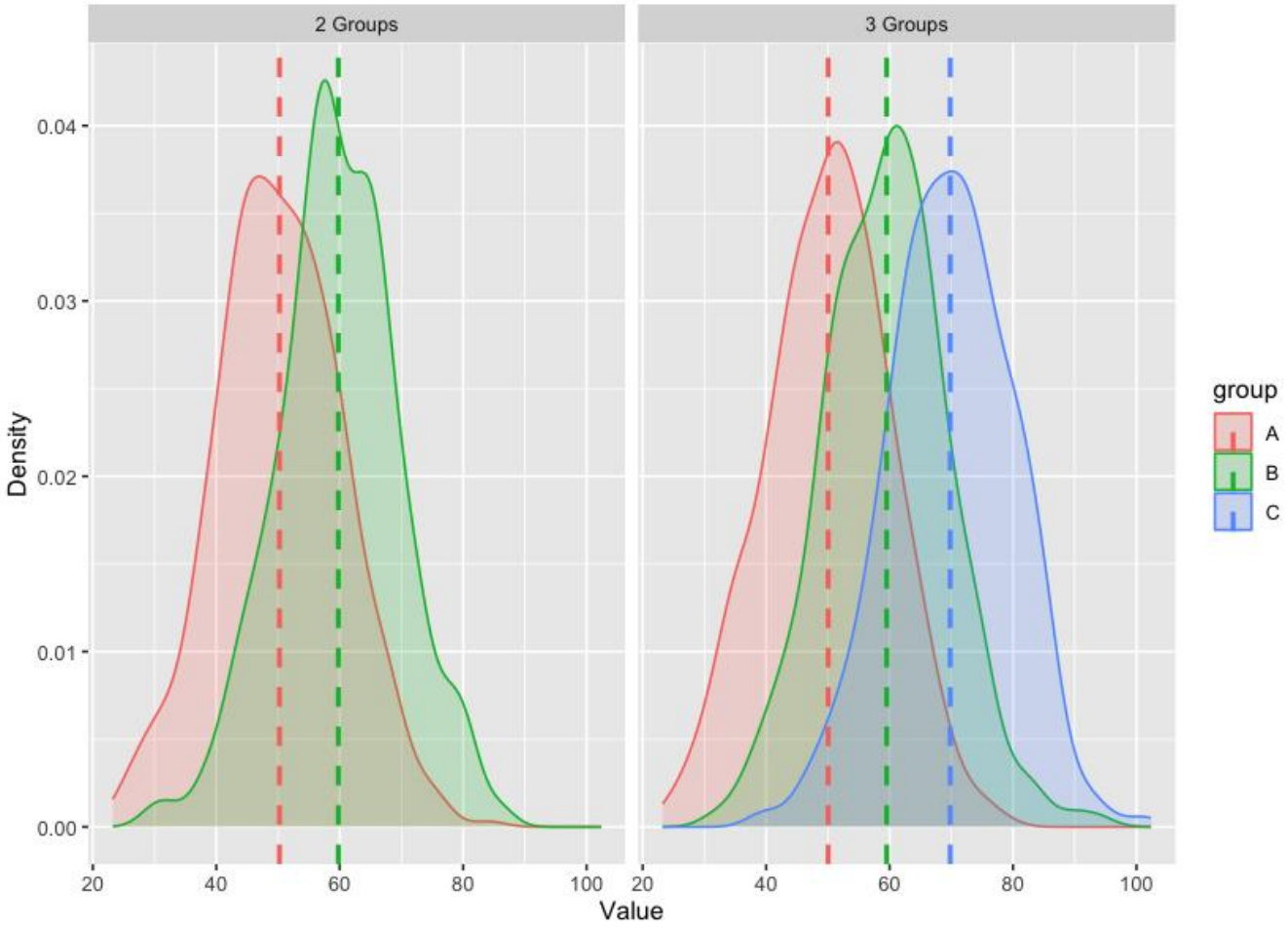
We have been building up to this

1. One sample t-test -> compare one group mean to a value
2. Two sample t-test -> compare two group means
3. ANOVA -> compare 2 or more group means

Examples:

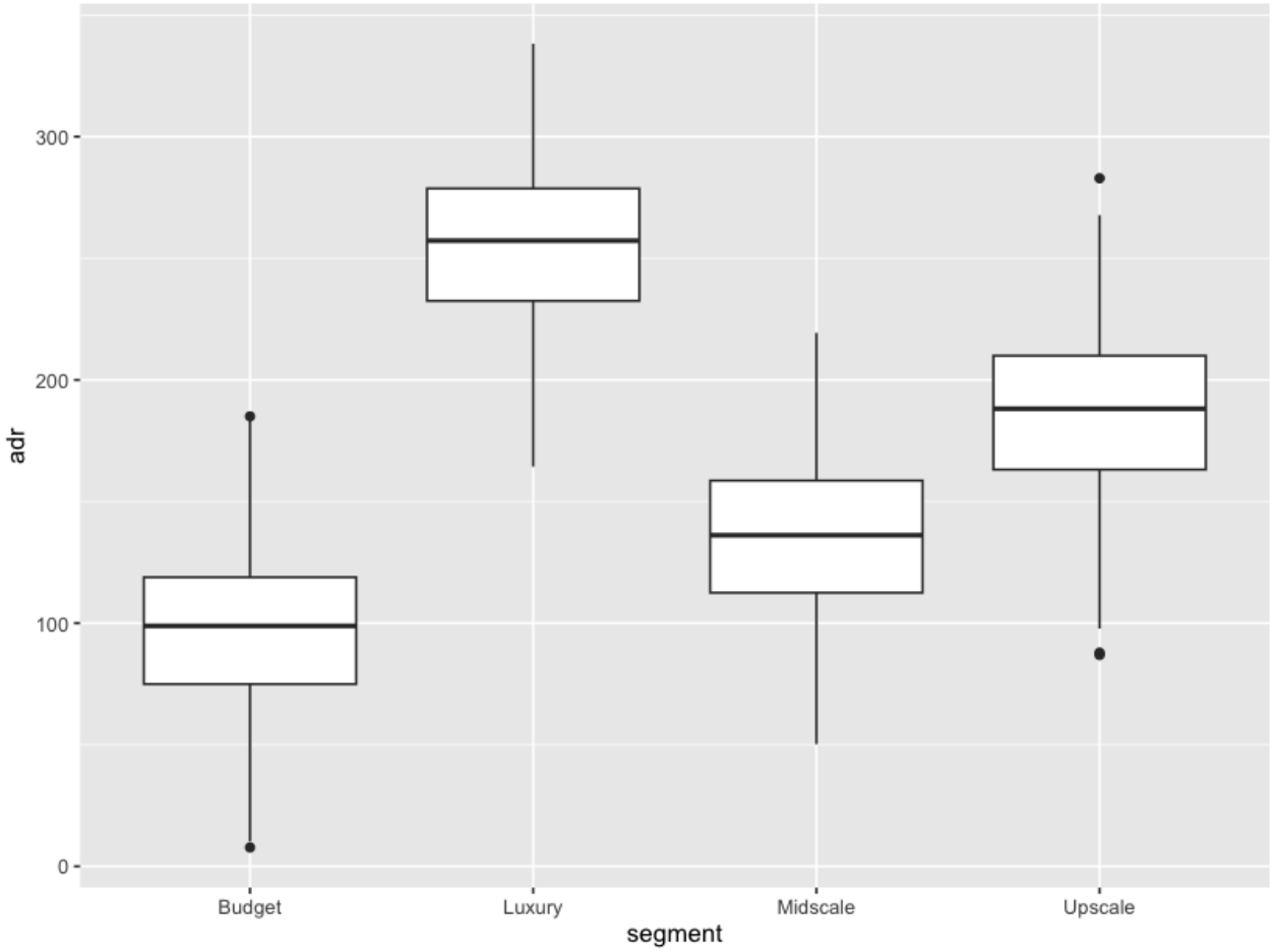
- Comparing ADR across hotel segments
- Comparing revenue across regions
- Comparing satisfaction across brands

Comparing group means



We want to generalize the t-test to have any number of group means…

What is ANOVA doing?



ANOVA compares within-group variance and between-group variance

- How does ADR vary within the luxury group?
- How does ADR vary between segments?

We extend our previous tests

	t-test	ANOVA
Number of groups	Two groups	Two or more groups
Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$	$H_0: \mu_1 = \mu_2 = \mu_3$ or $H_0: \mu_1 = \mu_2 = \ldots = \mu_n$
Alternative hypothesis	$H_a: \mu_1 - \mu_2 \neq 0$	$H_a: at\ least\ one\ mean\ is\ not\ equal$

In the case of exactly two groups, ANOVA is the same as the t-test

UNLV

What happens if the means aren't equal?

Step 1:
"Is there something different?"

$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_0 is false
At least one
mean is different

Step 2:
"what exactly is different?"

$$H_0: \mu_1 = \mu_2$$
$$H_0: \mu_1 = \mu_3$$
$$H_0: \mu_2 = \mu_3$$

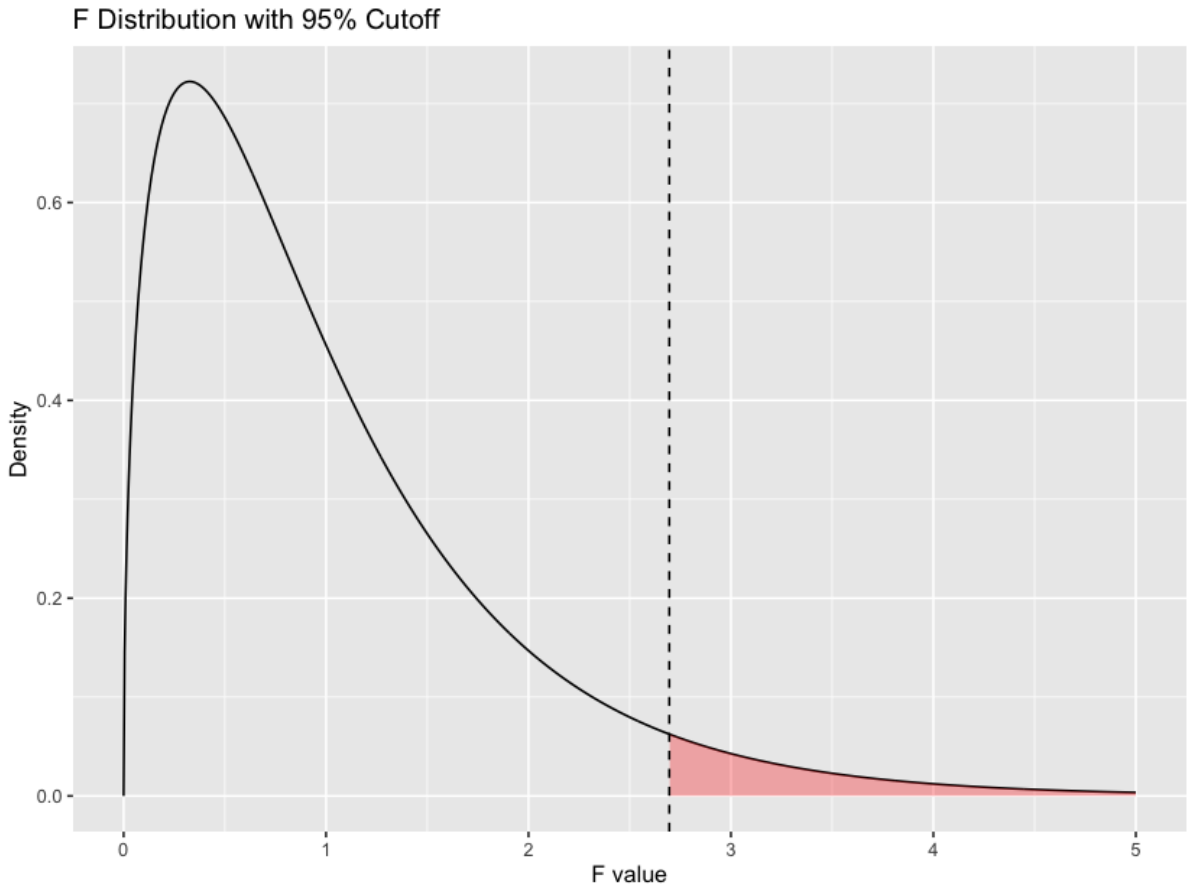
H_0 is true
All means equal

Step 1: testing for equal means

Instead of using a t-test, we use an F-test

$$F = \frac{bewteen\ -group\ variation}{within\ -group\ variation}$$

We reject that the means are equal when p-value < .05



Step 1: testing for equal means

- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- H_0: at least one mean unequal

Test comes back significant, meaning there is a difference, we just don't know where the difference(s) are…

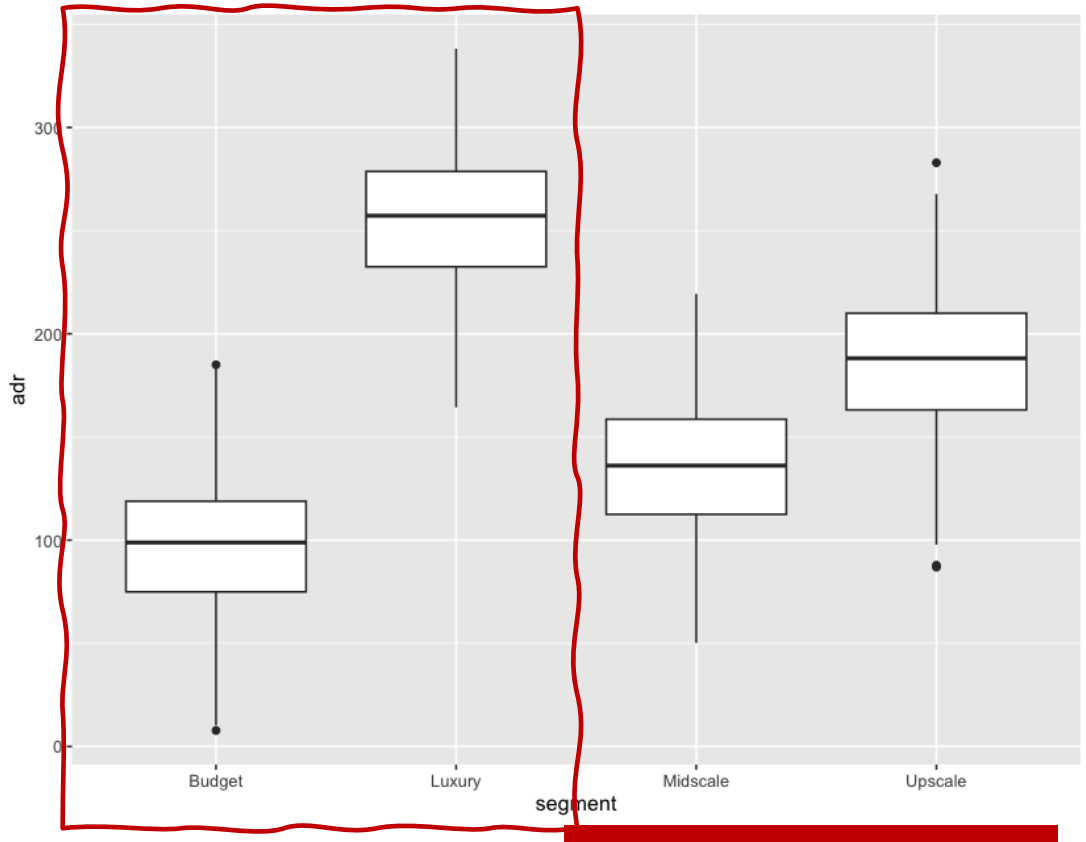
```
> anova_segment <- aov(adr ~ segment, data = hotels_anova)
> summary(anova_segment)
              Df   Sum Sq Mean Sq F value Pr(>F)
segment        3 10445396 3481799    3274 <2e-16 ***
Residuals   2996  3185850    1063
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

UNLV

Step 2: testing for difference between pairs

Tukey post-hoc analysis

- Compares every pair of group means
- Tells us which groups are different
- Used after ANOVA comes back with significant results (different means)



Step 2: testing for difference between pairs

We have four groups, so we have to check the pairwise differences:

1. Luxury vs budget
2. Midscale vs budget
3. Upscale vs budget
4. Midscale vs luxury
5. Midscale vs upscale
6. upscale vs luxury

```
> TukeyHSD(anova_segment)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = adr ~ segment, data = hotels_anova)

$segment
                         diff        lwr        upr p adj
Luxury-Budget      158.36320  154.02070  162.70570     0
Midscale-Budget     38.48559   34.14762   42.82355     0
Upscale-Budget      89.39882   85.12252   93.67513     0
Midscale-Luxury   -119.87762 -124.26054 -115.49469     0
Upscale-Luxury     -68.96438  -73.28628  -64.64248     0
Upscale-Midscale    50.91324   46.59589   55.23058     0
```

UNLV

Assumptions

1. Observations are independent
2. Groups have similar spread (variability)
3. Residuals are approximately normal

Same as regression assumptions

A final note on ANOVA and Regression

```
> anova_segment <- aov(adr ~ segment, data = hotels_anova)
> summary(anova_segment)
              Df   Sum Sq Mean Sq F value Pr(>F)
segment        3 10445396 3481799    3274 <2e-16 ***
Residuals   2996  3185850    1063
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

```
> TukeyHSD(anova_segment)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = adr ~ segment, data = hotels_anova)

$segment
                         diff        lwr        upr p adj
Luxury-Budget      158.36320  154.02070  162.70570     0
Midscale-Budget     38.48559   34.14762   42.82355     0
Upscale-Budget      89.39882   85.12252   93.67513     0
Midscale-Luxury   -119.87762 -124.26054 -115.49469     0
Upscale-Luxury     -68.96438  -73.28628  -64.64248     0
Upscale-Midscale    50.91324   46.59589   55.23058     0
```

```
> summary(lm_segment)

Call:
lm(formula = adr ~ segment, data = hotels_anova)

Residuals:
    Min      1Q  Median      3Q     Max
-99.521 -22.774   1.616  22.915  96.479

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       97.015      1.182   82.07   <2e-16 ***
segmentLuxury    158.363      1.689   93.74   <2e-16 ***
segmentMidscale   38.486      1.688   22.80   <2e-16 ***
segmentUpscale    89.399      1.664   53.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.61 on 2996 degrees of freedom
Multiple R-squared:  0.7663,    Adjusted R-squared:  0.766
F-statistic:  3274 on 3 and 2996 DF,  p-value: < 2.2e-16
```

ANOVA is just a regression, but instead of two numeric variables, you have one numeric and one categorical

UNLV