QS101: Introduction to Quantitative Methods in Social Science

Week 17: Comparing Groups - Analysis of Variance (ANOVA)

Dr. Florian Reiche
Teaching Fellow in Quantitative Methods
Course Director BA Politics and Sociology
Deputy Director of Student Experience and Progression, PAIS

February 19, 2015

Recap

Queries

- What is the range of a correlation coefficient?

Queries

- What is the range of a correlation coefficient?
- Which is the most commonly employed correlation coefficient?

Queries

- What is the range of a correlation coefficient?
- Which is the most commonly employed correlation coefficient?
- Which distribution does the test of significance for a correlation use?

ANOVA

Introduction

- Objective: We want to establish whether there is an association between a quantitative dependent variable and a categorical independent variables

Introduction

- Objective: We want to establish whether there is an association between a quantitative dependent variable and a categorical independent variables
- Comparison of mean annual income between whites, blacks, and Hispanics

Introduction

- Objective: We want to establish whether there is an association between a quantitative dependent variable and a categorical independent variables
- Comparison of mean annual income between whites, blacks, and Hispanics
- What is the dependent variable here?

Notation

- Let g denote the number of groups we want to compare

Notation

- Let g denote the number of groups we want to compare
- The means of the dependent variable for the corresponding populations are denoted as μ_1, μ_2, μ_3, $\ldots$, μ_g

Notation

- Let g denote the number of groups we want to compare
- The means of the dependent variable for the corresponding populations are denoted as $\mu_1, \mu_2, \mu_3, \ldots, \mu_g$
- The sample means are denoted as $\bar{y}_1, \bar{y}_2, \bar{y}_3, \ldots, \bar{y}_g$

Hypotheses

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_g$

Hypotheses

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_g$
- H_a: at least two of the population means are unequal

Assumptions

- For each group, the population distribution of the dependent variable y is normal

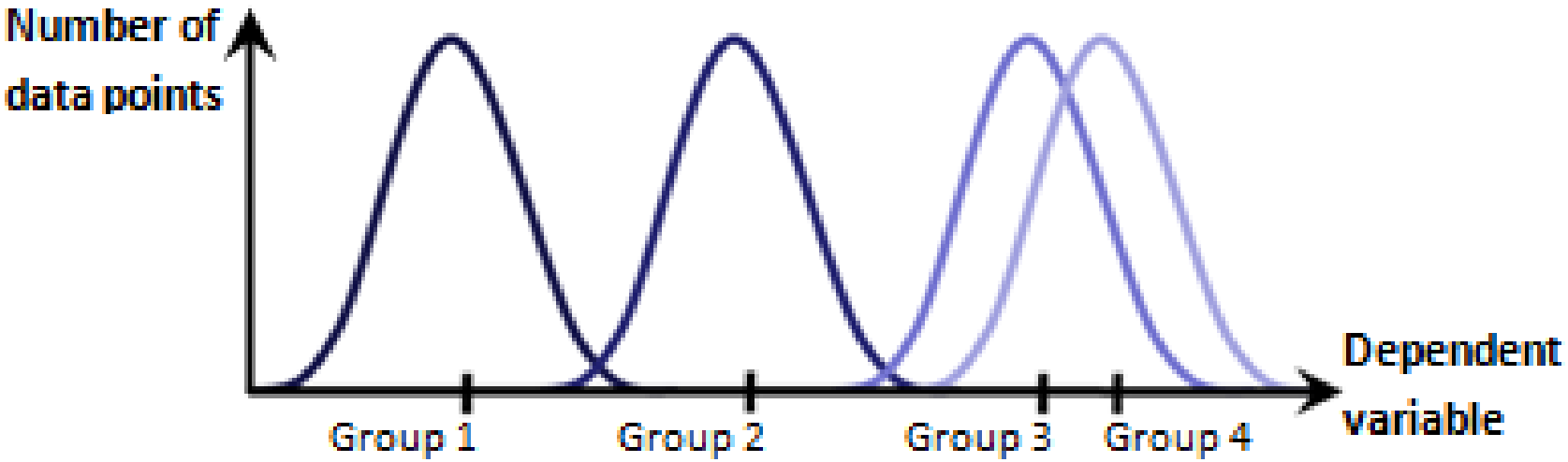
Assumptions

- For each group, the population distribution of the dependent variable y is normal
- The standard deviation of the population distribution is the same for each group. The common value is denoted by σ

Assumptions

- For each group, the population distribution of the dependent variable y is normal
- The standard deviation of the population distribution is the same for each group. The common value is denoted by σ
- The samples from the population are *independent* random samples

Assumptions (contd.)



Of means and variance ...

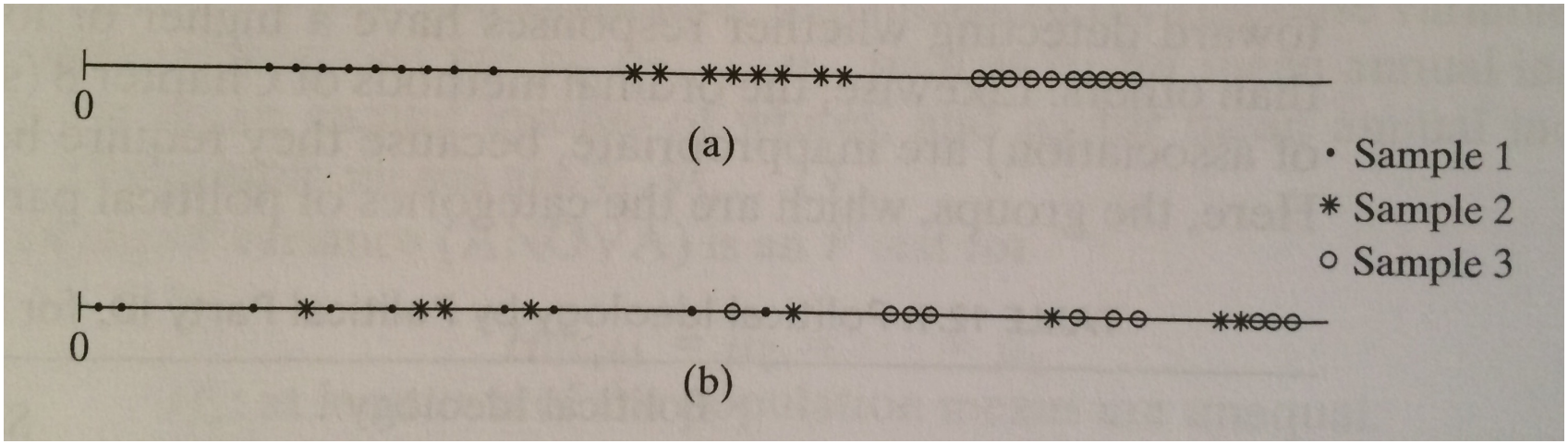
- Comparison of means is done by using two estimates of the variance, σ^2

Of means and variance ...

- Comparison of means is done by using two estimates of the variance, σ^2
- One estimate uses the variability between each sample mean $\bar{y}_i$ and the overall sample mean $\bar{y}$

Of means and variance ...

- Comparison of means is done by using two estimates of the variance, σ^2
- One estimate uses the variability between each sample mean $\bar{y}_i$ and the overall sample mean $\bar{y}$
- The other estimate uses the variability within each group of the sample observations about their separate means


0
(a)
• Sample 1
* Sample 2
o Sample 3
0
(b)

F Test Statistic

The F Test Statistic is a ratio of two variance estimates

$$F = \frac{\text{Between-groups estimate of variance}}{\text{Within-groups estimate of variance}} \tag{1}$$

This is called the ANOVA F statistic

Results

- If H_0 is true, then the F test statistic is equal to 1

Results

- If H_0 is true, then the F test statistic is equal to 1
- If H_0 is false, then the between-groups estimate will be larger, as it tends to overestimate σ^2

Results

- If H_0 is true, then the F test statistic is equal to 1
- If H_0 is false, then the between-groups estimate will be larger, as it tends to overestimate σ^2
- F test statistic has an F sampling distribution, with right hand probability for the p-value
- We will return to the F distribution next week when we tackle regression

How are the Variances Calculated?

- If you want to know how the within-group and between-group variance is calculated, turn to pp. 373-375 in Agresti and Finlay
- In Stata tables the results of the two variance estimates are presented as **Mean Square** (sum of squares, divided by df)

Example

```
. oneway a_fimngrs_dv a_sex, bonferroni tabulate
```

sex	Summary of total monthly personal income - gross Mean	Std. Dev.	Freq.
male	1820.5414	1895.08	16461
female	1242.8595	1245.0425	19308
Total	1508.7104	1603.8469	35769

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	2.9653e+09	1	2.9653e+09	1191.11	0.0000
Within groups	8.9042e+10	35767	2489491.5		
Total	9.2007e+10	35768	2572324.75		

Interpretation of Example

- What is the p-value?

Interpretation of Example

- What is the p-value?
- What does this tell us?

Interpretation of Example

- What is the p-value?
- What does this tell us?
- There is a difference amongst the population mean income for the two genders.

F-test versus several t-tests

- For $g = 2$ the tests are identical

F-test versus several t-tests

- For $g = 2$ the tests are identical
- For $g > 2$, only the F test allows us to control the probability of the type I error

Two-Way Analysis of Variance

Let's Make Things more Realistic

- Usually we want to control for other influences

Let's Make Things more Realistic

- Usually we want to control for other influences
- This creates sub-groups, such as white male, white female, black male, black female, and so forth

Let's Make Things more Realistic

- Usually we want to control for other influences
- This creates sub-groups, such as white male, white female, black male, black female, and so forth
- In order to compare population means across categories of two independent variables, we can perform a two-way ANOVA

Example

Let's compare the mean income of students from different courses, controlling for gender:

	Course of Study		
Gender	Sociology	Politics	Q-Step
Male	25,000	28,000	30,000
Female	24,000	29,000	29,000

We can also use this to compare income between gender, controlling for course of study (compare within columns).

If these had no effect, the Table would look as follows: No effect of course of study (a), no effect of gender (b)

		Course of Study		
	Gender	Sociology	Politics	Q-Step
(a)	Male	25,000	25,000	25,000
	Female	24,000	24,000	24,000
(b)	Male	25,000	28,000	30,000
	Female	25,000	28,000	30,000

F Tests about Main Effects

- The effects of individual predictors tested in these two null hypotheses are called *main effects*

F Tests about Main Effects

- The effects of individual predictors tested in these two null hypotheses are called *main effects*
- Assumptions are the same as for one-way ANOVA

F Tests about Main Effects

- The effects of individual predictors tested in these two null hypotheses are called *main effects*
- Assumptions are the same as for one-way ANOVA
- This time, you really don't want to see the maths behind it ...

What's the Setup?

The F Test Statistic is the ratio of mean squares

$$F = \frac{\text{MS for the predictor}}{\text{MS error (MSE)}} \tag{2}$$

Remember: Mean Squares = Sum of Squares divided by degrees of freedom.

Interaction in Two-Way ANOVA

- An absence of interaction between two independent variables means that the effects of of either variable on the dependent variable (in the population) does not change for different levels of the other.

Interaction in Two-Way ANOVA

- An absence of interaction between two independent variables means that the effects of of either variable on the dependent variable (in the population) does not change for different levels of the other.
- It is not meaningful to test for the main effects hypotheses if interaction exists

Interaction in Two-Way ANOVA

- An absence of interaction between two independent variables means that the effects of of either variable on the dependent variable (in the population) does not change for different levels of the other.
- It is not meaningful to test for the main effects hypotheses if interaction exists
- Here, we would conclude that each variable has an effect, but that the nature of that effect changes according to the category of the other variable.

First Things First

- We therefore test for interaction first

First Things First

- We therefore test for interaction first
- If interaction does not exist, we can do the main effects

First Things First

- We therefore test for interaction first
- If interaction does not exist, we can do the main effects
- If interaction does exist, it is better to compare the means for one predictor separately within categories of the other

Effects of Violations of ANOVA Assumptions

Recap: Assumptions

- For each group, the population distribution of the dependent variable y is normal
- The standard deviation of the population distribution is the same for each group. The common value is denoted by σ
- The samples from the population are *independent* random samples

Recap: Assumptions

- For each group, the population distribution of the dependent variable y is normal
- The standard deviation of the population distribution is the same for each group. The common value is denoted by σ
- The samples from the population are *independent* random samples

These are never exactly met in practice.

Robustness of F tests

- Moderate departures from normality of the population distribution can be tolerated

Robustness of F tests

- Moderate departures from normality of the population distribution can be tolerated
- Moderate departures from equal standard deviations can also be tolerated (esp. if sample sizes are identical)

Robustness of F tests

- Moderate departures from normality of the population distribution can be tolerated
- Moderate departures from equal standard deviations can also be tolerated (esp. if sample sizes are identical)
- Check histograms to make sure these assumptions are satisfied

Robustness of F tests

- Moderate departures from normality of the population distribution can be tolerated
- Moderate departures from equal standard deviations can also be tolerated (esp. if sample sizes are identical)
- Check histograms to make sure these assumptions are satisfied
- ANOVA procedures are NOT robust to violations of sampling assumptions

Misleading Results MAY Occur, if . . .

- . . . the population distributions are highly skewed and the sample size is small

Misleading Results MAY Occur, if . . .

- . . . the population distributions are highly skewed and the sample size is small
- . . . large differences amongst the population standard deviations and the sample sizes are unequal

What to do then?

- Non-parametric approaches exist, such as the Kruskal-Wallis Test