Lesson 24. One-Way ANOVA – Conditions, Scope and Limits – Part 1

Note. In Part 2 of this lesson, you can run the R code that generates the plots and outputs in here Part 1.

1 Previous lesson...

- **The one-way ANOVA model**
  - We need: one quantitative response variable and one categorical explanatory variable with K values
  - The model:

Response	=	Grand Mean	+	Group Effect	+	Error Term	
Y	=	μ	+	α_k	+	ε	$\varepsilon \sim N(0, \sigma_\varepsilon^2)$

  - Parameter estimates:

$$\hat{\mu} = \bar{y} \qquad \hat{\alpha}_k = \bar{y}_k - \bar{y}$$

  - Inference: ANOVA table, one-way ANOVA F-test for K groups

2 Conditions on the error terms

- The theory behind the one-way ANOVA F-test assumes the following are true:

1. Each population (group) has the same standard deviations
   - Check:
2. Each population (group) is Normal
   - Check:
3. After accounting for group membership, responses are independent
   - Check:

Example 1. Continuing with the `FatRats` example from the previous lesson... Recall:

A study was designed to compare the effect of three different high-protein diets on weight gain in baby rats. The subjects for the study were 30 baby rats. Each was fed a high-protein diet from one of three sources: beef, cereal, or pork. Their weight gains were recorded in grams. We would like to test whether average weight gain differs from protein source.

Check whether each of the three conditions appears to be met.

1. We can compute the standard deviations for each group using `tapply`:

```
tapply(FatRats.HighProtein$Gain, FatRats.HighProtein$Source, sd)
```

Here is the output:

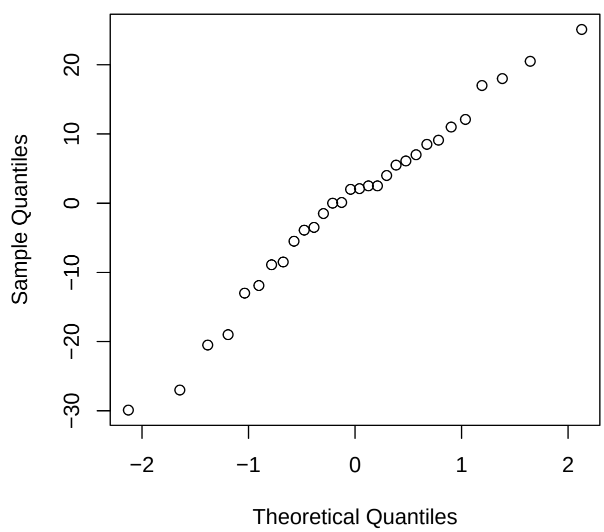
Beef: 15.136416719657 **Cereal:** 15.0218359582161 **Pork:** 10.9163485958752

Is the standard deviation condition met?

2. We can create a Normal Q-Q plot of the residuals values as follows:

```
qqnorm(residuals(test))
```

Here is the output:



Is the Normality condition met?

3. After accounting for group membership, are the responses independent?

3 Scope and limits of ANOVA inference

- In practice, almost all ANOVA datasets come from:
  (a) randomized, controlled experiments
  (b) independent random samples
  (c) observational data with no basis to claim a random sample

Example 2. Consider the `FatRats` example from the previous lesson and Example 1.

- The data for this example is

- In this case, a significant p-value from a one-way ANOVA F-test provides evidence that there are detectable differences among the means of the treatments studied

- Since the explanatory variable was assigned at random, the data

Example 3. Do average commute times in various U.S. cities differ? If so, by how much? To investigate these questions, we can use data from the 2007 American Housing Survey, which involved representative samples of American households in various metropolitan areas. Suppose we select independent random samples of 500 commuters from Boston, Houston, Minneapolis, and Washington, D.C. The data contains one-way daily commuting distance (in miles) and time (in minutes).

- Is this an experiment or an observational study?

- The data for this example is

- In this case, since we have random samples from each city, we can make inferences about the differences in mean commute times for the larger populations of commuters in those cities

- Since the explanatory variable was not assigned, the data

Example 4. The data in `TeenPregnancy` from the `Stat2Data` library give the teen pregnancy rate (number of pregnancies per 1000 teenage girls) for each of the 50 U.S. states in 2010. One of the other variables in that dataset records the role of the state in the U.S. Civil War in the table below, which also shows the mean teen pregnancy rate for each group of states.

Group	Number of states	Mean Teen Pregnancy
Confederate states	11	64.64
Border states	3	61.67
Union states	21	48.24
Other states (admitted later)	15	55.07

- Is this an experiment or an observational study?
- The data for this example is
- In this case, making inference (i.e., making statistically significant conclusions) is <u>not</u> appropriate
- A small or large p-value might still indicate whether the pattern of association is worth further investigation
- Since we explanatory variable was not assigned, the data