

Analysis of Variance (ANOVA)

Ryan Miller

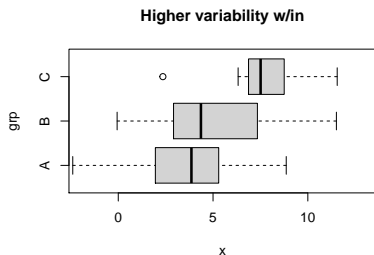
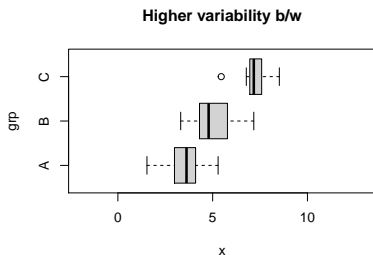
1. Statistical models
2. Comparing models
3. Post-hoc testing and ANOVA assumptions

This semester we've learned that the choice of statistical test depends upon the types of variable(s) we are considering:

- ▶ categorical outcome
 - ▶ one-sample Z -test if binary, or Chi-squared goodness of fit test if nominal
 - ▶ comparing two groups - two-sample Z -test
 - ▶ comparing many groups - Chi-squared test of association
- ▶ quantitative outcome
 - ▶ one-sample T -test
 - ▶ comparing two groups - two-sample T -test
 - ▶ comparing many groups - ?

ANOVA (big picture)

- ▶ One-way ANOVA (analysis of variance) is a statistical test for comparing a *quantitative outcome across many groups*
- ▶ The basic idea of ANOVA is to use a *statistical model* to split the *total variability* in the outcome variable into two parts, variability *between groups* and *variability within groups*:



A statistical model is a representation of an observed phenomena that involves a *systematic component* and a *random component*:

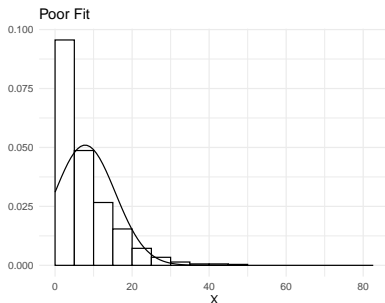
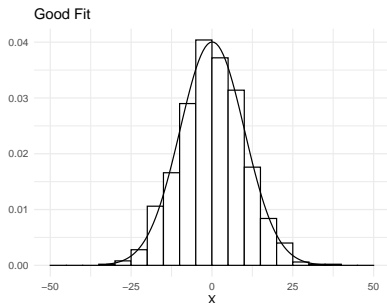
$$y = f(X) + \epsilon$$

A simple model uses:

- ▶ $f(X) = \mu$, which implies the data are centered at the population's mean (μ)
- ▶ $\epsilon \sim N(0, \sigma)$, which implies random variability around μ following a Normal curve

Statistical Modeling - Introduction

Below are two examples of the model $f(X) = \mu$ and $\epsilon \sim N(0, \sigma)$:



We'll want to *mathematically* describe how well such a model fits the observed data

Statistical Modeling - Residuals and sums of squares

- ▶ A good model produces *predictions* that closely resemble the observed data
 - ▶ Predictions only use the model's *systematic component*, so our simple model would predict \bar{y} (the sample mean) for each data-point

- ▶ A good model produces *predictions* that closely resemble the observed data
 - ▶ Predictions only use the model's *systematic component*, so our simple model would predict \bar{y} (the sample mean) for each data-point
- ▶ The accuracy of each prediction can be expressed as a **residual**. In general:

$$r_i = y_i - \hat{y}_i$$

- ▶ For our simple model, this is:

$$r_i = y_i - \bar{y}$$

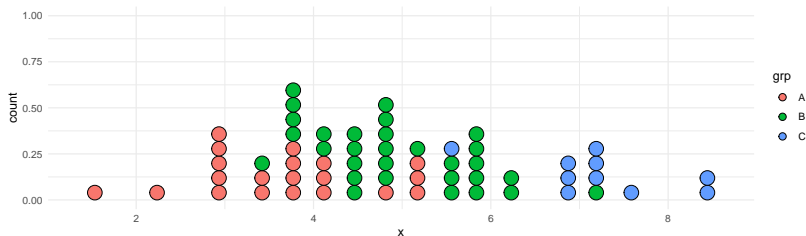
We can summarize a model's overall fit using *all* of the residuals:

$$SS = \sum_{i=1}^n r_i^2$$

- ▶ A smaller *sum of squares* indicates a better fit between the model and the observed data
- ▶ **Analysis of variance** (ANOVA) is a statistical test used to determine whether a more complex model fits the data better than a less complex model by an amount that is more than would be expected by random chance

Statistical Modeling - Null and alternative models

Summarized below are quantitative data for three different groups (A, B, and C):

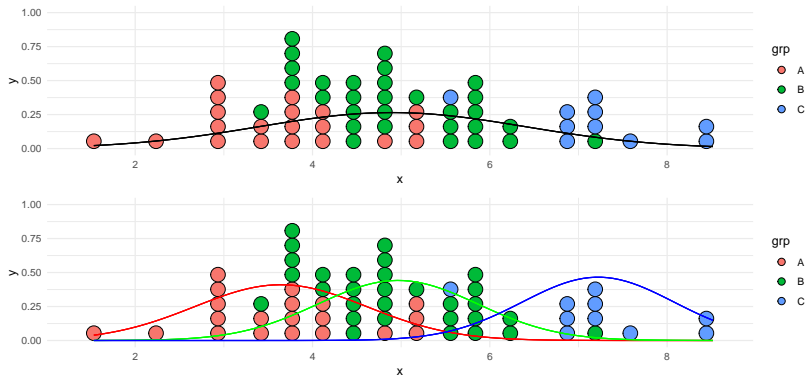


grp	n	Mean	StdDev
A	20	3.64	0.97
B	30	4.96	0.90
C	10	7.22	0.86

Can you think of two different models for these data? (Hint: think about one that uses the “group” and one that doesn’t)

Statistical Modeling - Null and alternative models

One model might use a *single mean* to represent all of the data, while another might use *group-specific means*:



Is there enough of a difference for us to *reject* the simpler model in favor of the more complex model?

ANOVA uses an F -test to compare models using the following steps:

- 1) H_0 involves the simpler model, in our case $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$, while H_a describes the more complex model, in our case “at least one mean is different”

ANOVA uses an F -test to compare models using the following steps:

- 1) H_0 involves the simpler model, in our case
 $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$, while H_a describes the more complex model, in our case “at least one mean is different”
- 2) Each model is summarized using a *sum of squares* (SS), we'll use SST for the null model and SSE for the alternative model
- 3) We then calculate an F -value:

$$F = \frac{(SST - SSE)/(d_1 - d_0)}{\text{Std. Error}}$$

- ▶ d_1 and d_0 describe the number of parameters in each model
 - ▶ In our example, $d_0 = 1$ (the single overall mean) and $d_1 = 3$ (the means of groups “A”, “B”, and “C”)

Statistical Modeling - ANOVA

ANOVA uses an F -test to compare models using the following steps:

- 1) H_0 involves the simpler model, in our case
 $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$, while H_a describes the more complex model, in our case “at least one mean is different”
- 2) Each model is summarized using a *sum of squares* (SS), we'll use SST for the null model and SSE for the alternative model
- 3) We then calculate an F -value:

$$F = \frac{(SST - SSE)/(d_1 - d_0)}{\text{Std. Error}}$$

- ▶ d_1 and d_0 describe the number of parameters in each model
 - ▶ In our example, $d_0 = 1$ (the single overall mean) and $d_1 = 3$ (the means of groups “A”, “B”, and “C”)

So, the F -value is a standardized measure of improvement in model fit (via the per parameter drop in SS)

- ▶ We've seen that standard errors tend to look like a measure of variability divided by the sample size, for ANOVA:

$$\text{Std. Error} = \frac{SSE}{n - d_1}$$

- ▶ This is the sum of squares of the alternative model divided by its *degrees of freedom*, $df = n - d_1$
- ▶ Using this standard error, the F -value can be expressed:

$$F = \frac{(SST - SSE)/(d_1 - d_0)}{SSE/(n - d_1)}$$

Example - Finding the F-value using StatKey

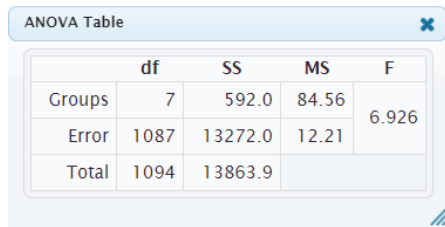
- ▶ Early in the semester we introduced the Colleges 2019 dataset, which contains data from The College Scorecard for all institutions that primarily bachelors degrees and enroll at least 400 students
 - ▶ In this example we'll see if "Region" is associated with "ACT_median"

Example - Finding the F-value using StatKey

- ▶ Early in the semester we introduced the Colleges 2019 dataset, which contains data from The College Scorecard for all institutions that primarily bachelors degrees and enroll at least 400 students
 - ▶ In this example we'll see if "Region" is associated with "ACT_median"
- ▶ To begin, download the data from our course website, then upload it into StatKey in the "ANOVA for Difference in Means" menu
 - ▶ You should see a table displaying the means and standard deviations for each of the 8 regions

Example - Finding the F-value using StatKey

An **ANOVA table** is a table used to organize the information needed to perform an F -test to compare two statistical models:



The screenshot shows a window titled "ANOVA Table" with a close button (X) in the top right corner. The window contains a table with the following data:

	df	SS	MS	F
Groups	7	592.0	84.56	6.926
Error	1087	13272.0	12.21	
Total	1094	13863.9		

- ▶ On StatKey, you can click “ANOVA Table” next to the Original Sample to open this table in a pop-up window
- ▶ So far, we’ve only discussed the sum of squares (SS) in the Error and Total rows
 - ▶ However, these are the only two things we need a computer to calculate in order to perform an ANOVA test

Example - Finding the F-value using StatKey

The F -value describes whether the difference in fit of the null and alternative models exceeds what we'd expect by random chance:

$$F = \frac{(SST - SSE)/(d_1 - d_0)}{SSE/(n - d_1)}$$

- ▶ $SST = 13863.9$ (sum of squares for the null model)
- ▶ $d_0 = 1$ (number of groups under the null model)
- ▶ $SSE = 13272.0$ (sum of squares for the alternative model)
- ▶ $d_1 = 8$ (number of groups under the alternative model)
- ▶ $n = 1095$ (total sample size)

Example - Using the F-distribution

In our colleges example, we found the following F -value:

$$F = \frac{(13863.9 - 13272.0)/(8 - 1)}{13272.0/(1095 - 8)} = 6.925$$

- ▶ To translate this test statistic into a p -value, we must compare it against an F -distribution
 - ▶ The **numerator** df are $d_1 - d_0$, or 7 in our example
 - ▶ The **denominator** df are $n - d_1$, or 1087 in our example

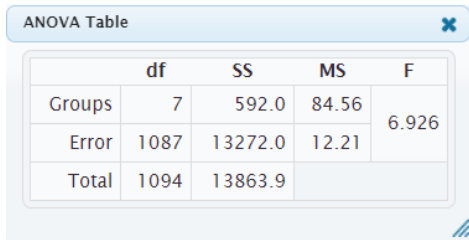
Example - Using the F-distribution

In our colleges example, we found the following F -value:

$$F = \frac{(13863.9 - 13272.0)/(8 - 1)}{13272.0/(1095 - 8)} = 6.925$$

- ▶ To translate this test statistic into a p -value, we must compare it against an F -distribution
 - ▶ The **numerator** df are $d_1 - d_0$, or 7 in our example
 - ▶ The **denominator** df are $n - d_1$, or 1087 in our example
- ▶ Therefore, the p -value is less than 0.0001, so there is overwhelming statistical evidence of a difference in ACT scores by region

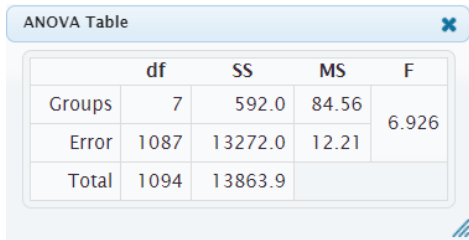
Example - Unpacking the ANOVA table



	df	SS	MS	F
Groups	7	592.0	84.56	6.926
Error	1087	13272.0	12.21	
Total	1094	13863.9		

- ▶ The sum of squares in the “Group” row, or SSG , is part of the numerator of the F -value
 - ▶ The other part of numerator is the “Group” df
 - ▶ Notice SSG divided by it's degrees of freedom gives us the *mean square* (MS) value for “Group”

Example - Unpacking the ANOVA table



	df	SS	MS	F
Groups	7	592.0	84.56	6.926
Error	1087	13272.0	12.21	
Total	1094	13863.9		

- ▶ The sum of squares in the “Group” row, or SSG , is part of the numerator of the F -value
 - ▶ The other part of numerator is the “Group” df
 - ▶ Notice SSG divided by it's degrees of freedom gives us the *mean square* (MS) value for “Group”
- ▶ Similarly, the “Error” row describes denominator of the F -value
 - ▶ Then, the ratio of the two mean squares, MSG/MSE , is the F -value

Practice #1

The Breast Cancer Survival dataset documents the survival times from their initial diagnosis of a group of women who died from breast cancer. These data are available on our course website. For this question, we'll see if survival time is associated with cancer grade (I, II, or III)

- 1) State the null hypothesis in statistical terms
- 2) State the null and alternative models (in words)
- 3) Use StatKey to find the ANOVA table
- 4) Find a p -value and make a conclusion

Practice #1 (solution)

- 1) $H_0 : \mu_1 = \mu_2 = \mu_3$
- 2) The null model is that survival time can be represented by a single mean, the alternative model is the group-specific means (for each cancer grade) should be used to predict survival time.
- 3) Performed on StatKey
- 4) For an F -value of 5.307 with $df_1 = 2$ and $df_2 = 196$, the p -value is 0.0057, so we conclude there is a significant different in survival time by cancer grade.

Practice #2

Below are the partial results of a generic ANOVA test (multiple means) where $d_0 = 1$:

Source	df	Sum Sq.	Mean Sq.	F -statistic	p -value
"Group"	4	200	?	?	?
Error	?	440	?		
Total	59	?			

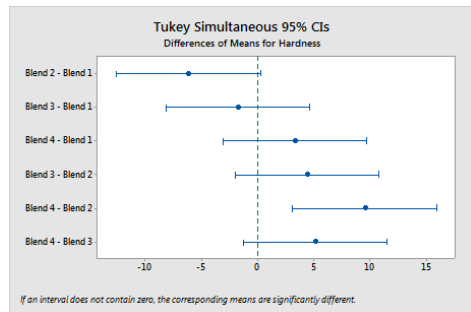
- 1) Fill in the remainder of the ANOVA table and calculate the F -value
- 2) Find the p -value and use it to make a generic conclusion
- 3) Sketch an example of what the side-by-side box plots might have looked like for these data

Practice #2 (solution)

- 1) Notice $d_1 = 5$ and $n = 60$, thus the F -value is 6.25
- 2) For $df_1 = 4$ and $df_2 = 55$, this leads to a p -value of 0.0003
- 3) Since the p -value is small, the distributions of at least two of the groups in the side-by-side boxplots should be far from each other.

Loose Ends - Post-hoc Testing

After a statistically significant ANOVA test, the natural next step is to identify which groups are different:



There are many approaches to this, but they're all similar to two-sample t -tests or confidence intervals for each difference in means

Loose Ends - Assumptions behind ANOVA

One-way ANOVA, like the other statistical tests we've covered this semester, relies upon certain assumptions:

- ▶ Normally distributed errors (residuals)
- ▶ Equal variance - all groups should have roughly the same standard deviation

Strategies commonly used to address violated assumptions include: transforming the outcome variable (converting it to the log-scale), randomization testing, reporting results with caution

This presentation introduced one-way ANOVA as a statistical test for comparing the means of many groups:

- ▶ ANOVA calculates an F-statistic that expresses whether using group-specific means can improve the sum of squares of a single-group model by more than would be expected by random chance.
- ▶ After a statistically significant ANOVA test, follow-up tests should be conducted to compare individual pairings of groups

As always, be cautious of our common hypothesis testing mistakes when interpreting the results of an ANOVA test.