ANOVA

Ryan Miller

Outline

1. Introduction to statistical modeling
2. The one-way ANOVA model
3. Post-hoc testing

Statistical modeling - introduction

- A **model** is a simplified representation of some phenomenon intended to aide in *explanation* or *prediction*
  - A **statistical model** is one that involves a *probability distribution*

Statistical modeling - introduction

- A **model** is a simplified representation of some phenomenon intended to aide in *explanation* or *prediction*
  - A **statistical model** is one that involves a *probability distribution*
- All statistical models include a *systematic component* and a *random component*:

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

Arguably the simplest statistical model uses $f(X) = \mu$ and $\epsilon \sim N(0, \sigma)$, which suggest data-points are centered at the population's mean (μ) with random variability following a Normal curve

Statistical modeling - fitted models

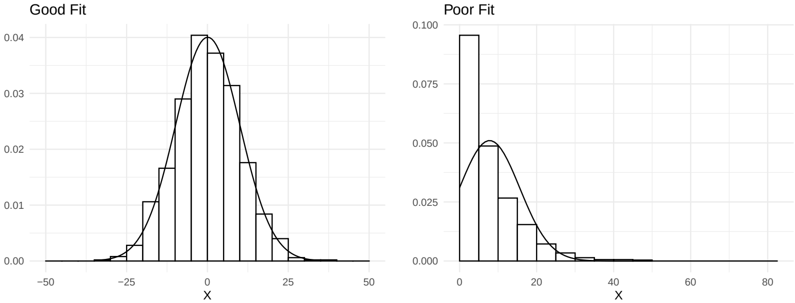
Applying a statistical model has two steps:

1) Specifying the model's systematic and random components (done at the population-level)
2) Estimating the model parameters (done using the sample data)

Our simple model (from the last slide) would require us to estimate two parameters: μ and σ

Statistical modeling - fitted models

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



Clearly some model fits are better than others, we'll need a way of quantifying this.

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 predicts $\bar{y}$ (the sample mean) for each data-point

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 predicts $\bar{y}$ (the sample mean) for each data-point
- The accuracy of an individual prediction is expressed as a **residual**. In general:

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

- For our simple model, residuals look like:

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

Statistical Modeling - residuals and sums of squares

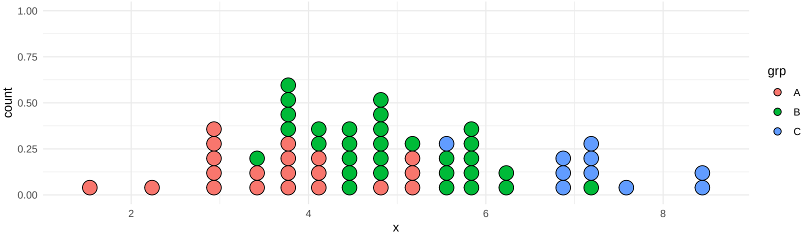
We can summarize a model's overall fit by considering *all* of its residuals:

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

- 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

The one-way ANOVA model

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

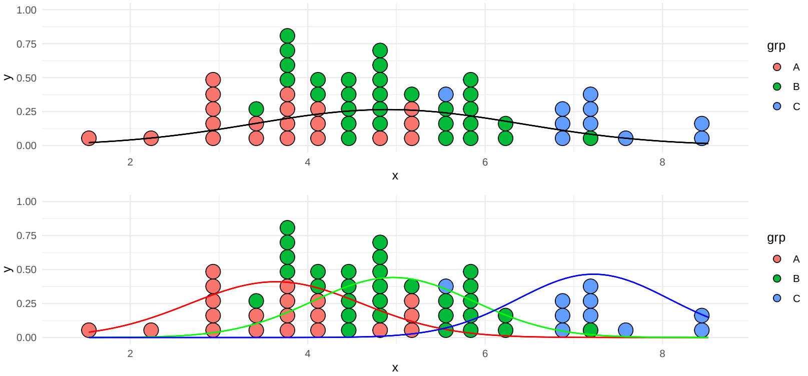


grp	n	Mean	StdDev
A	20	3.64	0.97
B	30	4.96	0.90
C	10	7.22	0.86

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

The one-way ANOVA model

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?

The F-test

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

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

The F-test

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

1) H_0 involves the simpler model, in our case $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$, 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")

The F-test

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

1) H_0 involves the simpler model, in our case $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$, 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)

The F-test

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

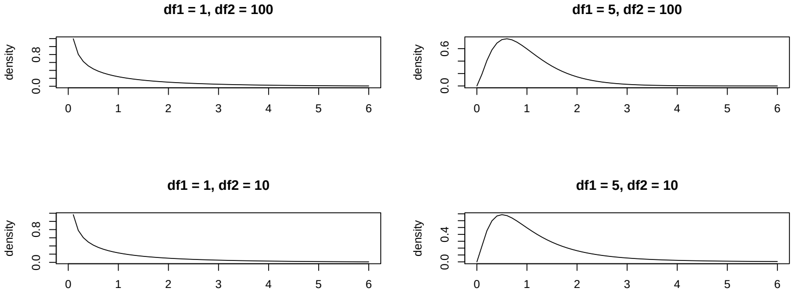
$$\text{Std. Error} = \frac{SSE}{n - d_1}$$

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

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

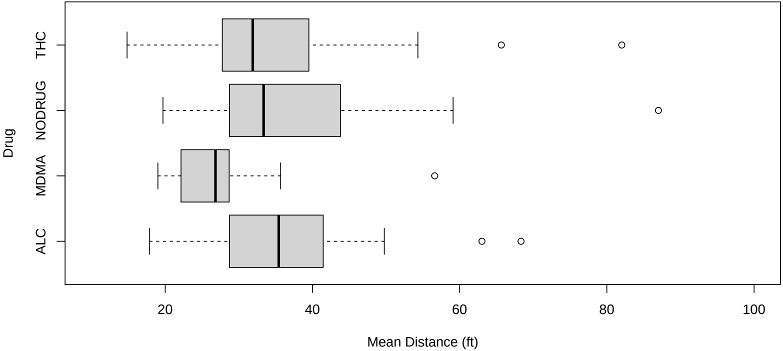
The F-distribution

- The observed F-value must be compared against the proper F-distribution to find the p-value
- Mathematically, F-distribution is the ratio of two Chi-squared distributions divided by their respective degrees of freedom
  - In practical terms, this means we need to specify numerator and denominator *df*



Example - introduction

We previously discussed a study exploring the driving of different categories of drug users:



Rather than individually comparing each group, we can instead begin by testing for an overall association.

Example - null and alternative models

- The null model is akin to modeling everyone's mean following distance using a single, overall mean
  - Statistical model: $y_i = \mu + \epsilon_i$, predictions: $\hat{y}_i = \bar{y}$
  - Corresponding hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
- The one-way ANOVA model is akin to using group-specific means
  - Statistical model: $y_i = \mu_i + \epsilon_i$, predictions: $\hat{y}_i = \bar{y}_i$
  - Corresponding hypothesis: "at least one group-specific mean differs from the others"

Example - ANOVA tables

Shown below is an **ANOVA table**, a common summary table used to describe a model:

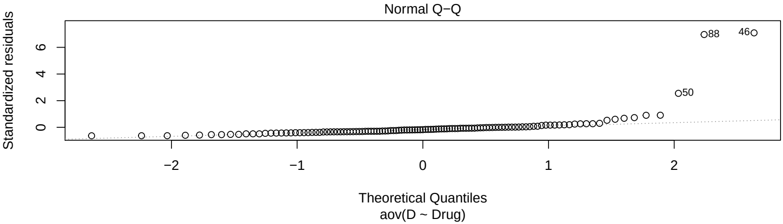
```r
tail <- read.csv("https://remiller1450.github.io/data/Tailgating.csv")
mod <- aov(D ~ Drug, data = tail)
summary(mod)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Drug          3   4989    1663    0.85   0.47
## Residuals   115 225127    1958
```

- The "residuals" row describes the fit of the alternative model (ie: SSE)
- The "Drug" row describes the improvement in fit that can be attributed to the variable "Drug" (ie: $SST - SSE$).

Example - diagnostics

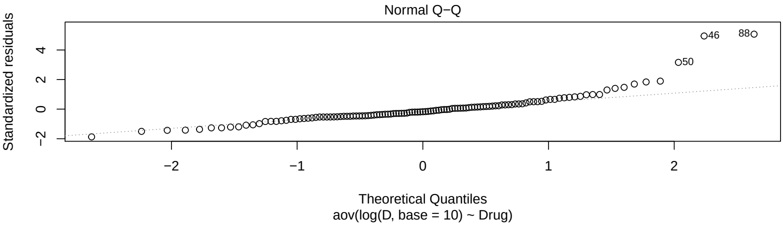
- ANOVA relies upon a probability model (the random component) that might not reasonably reflect the data
- A QQ-plot of the residuals is a popular diagnostic tool
  - If the residuals do not reflect a Normal distribution, the model is improper (as it specifies Normally distributed errors)



Example - a better model

- In our example, the right-skewed nature of these data is incompatible with the specified model
  - This is relatively common, and a simple solution is to apply a *log-transformation* to the outcome variable
  - The revised model still isn't good, but it's certainly an improvement

```
##              Df Sum Sq Mean Sq F value Pr(>F)  
## Drug          3  0.267 0.08898    2.23 0.0884 .
## Residuals   115  4.588 0.03990                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Practice

We've previously introduced data collected by a restaurant server at a chain restaurant in the suburbs of NYC. The code below reads these data and converts table size to a categorical variable:

```r
tips <- read.csv("https://remiller1450.github.io/data/Tips.csv")
tips$Size = as.factor(tips$Size)  ## Convert table size to categorical
```

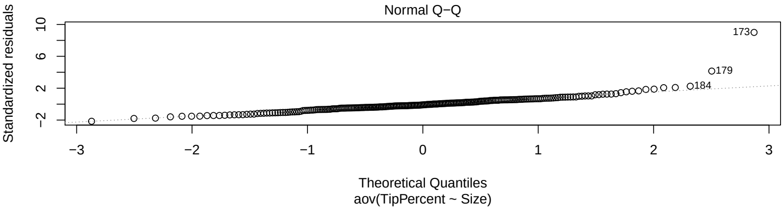
1) Use `R` to fit a one-way ANOVA model that uses table size to predict the percent tipped
2) Use the `summary()` function and an F-test to evaluate this model relative to the null model
3) Use a QQ-plot to evaluate whether this model one-way ANOVA model seems appropriate

Practice (solution)

```r
mod = aov(TipPercent ~ Size, data = tips)
summary(mod)
```

```
##              Df Sum Sq  Mean Sq F value Pr(>F)
## Size          5 0.0295 0.005897   1.601  0.161
## Residuals   238 0.8769 0.003684
```

```r
plot(mod, which = 2)
```



Equal variance assumption

- In addition to assuming Normally distributed errors, ANOVA also assumes the variance of the outcome is the same for each group (ie: a single value of σ in the population-level model)
- This can be checked by comparing sample standard deviations and assessing their similarity
  - Typically we are only concerned if there are very large differences (a ratio ≥ 3 for the largest/smallest)

```
library(dplyr)
tips %>% group_by(Size) %>% summarize(sd = sd(TipPercent))
```

```
## # A tibble: 6 x 2
##   Size      sd
##   <fct>  <dbl>
## 1 1     0.0803
## 2 2     0.0668
## 3 3     0.0455
## 4 4     0.0424
## 5 5     0.0677
## 6 6     0.0422
```

Post-hoc testing

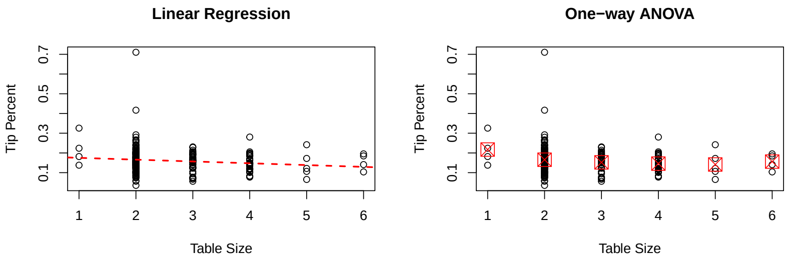
Tukey's Honest Significant Differences (HSD) is a post-hoc test that is designed to control the *family-wise Type I error rate*:

```r
tail <- read.csv("https://remiller1450.github.io/data/Tailgating.csv")
mod <- aov(LD ~ Drug, data = tail) ## Log-scale outcome
TukeyHSD(mod, conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = LD ~ Drug, data = tail)
##
## $Drug
##                    diff         lwr       upr     p adj
## MDMA-ALC    -0.27947379 -0.66645712 0.1075095 0.2411710
## NODRUG-ALC   0.07044162 -0.23914504 0.3800283 0.9339585
## THC-ALC     -0.01341974 -0.32449124 0.2976518 0.9994882
## NODRUG-MDMA  0.34991541 -0.00476067 0.7045915 0.0546053
## THC-MDMA     0.26605404 -0.08991885 0.6220269 0.2138699
## THC-NODRUG  -0.08386137 -0.35368446 0.1859617 0.8495067
```

ANOVA for other model comparisons

- ANOVA is a general statistical test that can be used to compare any two *nested* models
  - For example, we could also compare a *linear regression* model that treats table size as numeric (in the tipping example)



ANOVA for other model comparisons

Shown below are the ANOVA tables for each of these models (which cannot be directly compared since they are not nested):

```
tips <- read.csv("https://remiller1450.github.io/data/Tips.csv")
linmod = lm(TipPercent ~ Size, data = tips)
anova(linmod)
```

```
## Analysis of Variance Table
##
## Response: TipPercent
##            Df  Sum Sq   Mean Sq F value  Pr(>F)
## Size        1 0.01850 0.0184975  5.0418 0.02565 *
## Residuals 242 0.88785 0.0036688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tips$Size = as.factor(tips$Size)  ## Convert table size to categorical
aovmod = aov(TipPercent ~ Size, data = tips)
summary(aovmod)
```

```
##              Df Sum Sq  Mean Sq F value Pr(>F)
## Size          5 0.0295 0.005897   1.601  0.161
## Residuals   238 0.8769 0.003684
```

Conclusion

This presentation introduced ANOVA as a hypothesis test for comparing a statistical model against a simpler null model, I expect you to know the following:

- Situations where one-way ANOVA is used (ie: comparing the means of multiple groups)
- How to perform one-way ANOVA and post-hoc testing in `R` (ie: `aov()` and `TukeyHSD()`)
- How to interpret ANOVA output (ie: sums of squares, the F-statistic, etc.)
- Model assumptions made by the one-way ANOVA model (ie: Normality and equal variance)