Analysis of Variance (ANOVA)

Grinnell College

Review + Outline

Hypothesis Tests

- comparison of proportions
- comparison of means
  - $H_0 : \mu_1 = \mu_2$

Multiple Regression

- Linear relationship between response and predictors
- Can use quant. or cat. predictors
- Quality of fit $\rightarrow$ multiple or adj. R^2
  - can help us compare models
- Hypothesis tests related to slopes/intercept
  - F-test, p-value for overall performance
  - t-tests for slope/intercept being 0

Today we are going to see what else we can do with these F- and t-tests, and talk more about what the F-test is actually doing. Some parts may be a bit 'mathy'

ANOVA

The **Analysis of variance (ANOVA)** is a collection of statistical *models* used to analyze difference among many means

The null hypothesis is testing the difference of means between k groups

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_A : \text{at least one } \mu_i \neq \mu_j$$

But what does this have to do with variance?

Review – Standardizing

Think back to what we were doing when calculating test-statistics before:

$$T = \frac{\overline{x} - \mu_0}{(\hat{\sigma}/\sqrt{n})}$$

The 'formula' for a test-statistic has always been to do the following

$$\text{Test-stat} = \frac{\text{statistic} - \text{hypothesized value}}{\text{standard error}}$$

The test-statistic

- measures how 'far' the statistic is from the hypothesized value
- uses the standard error as a 'ruler'
- we have been using variance to compare distance this whole time

Dog Speed

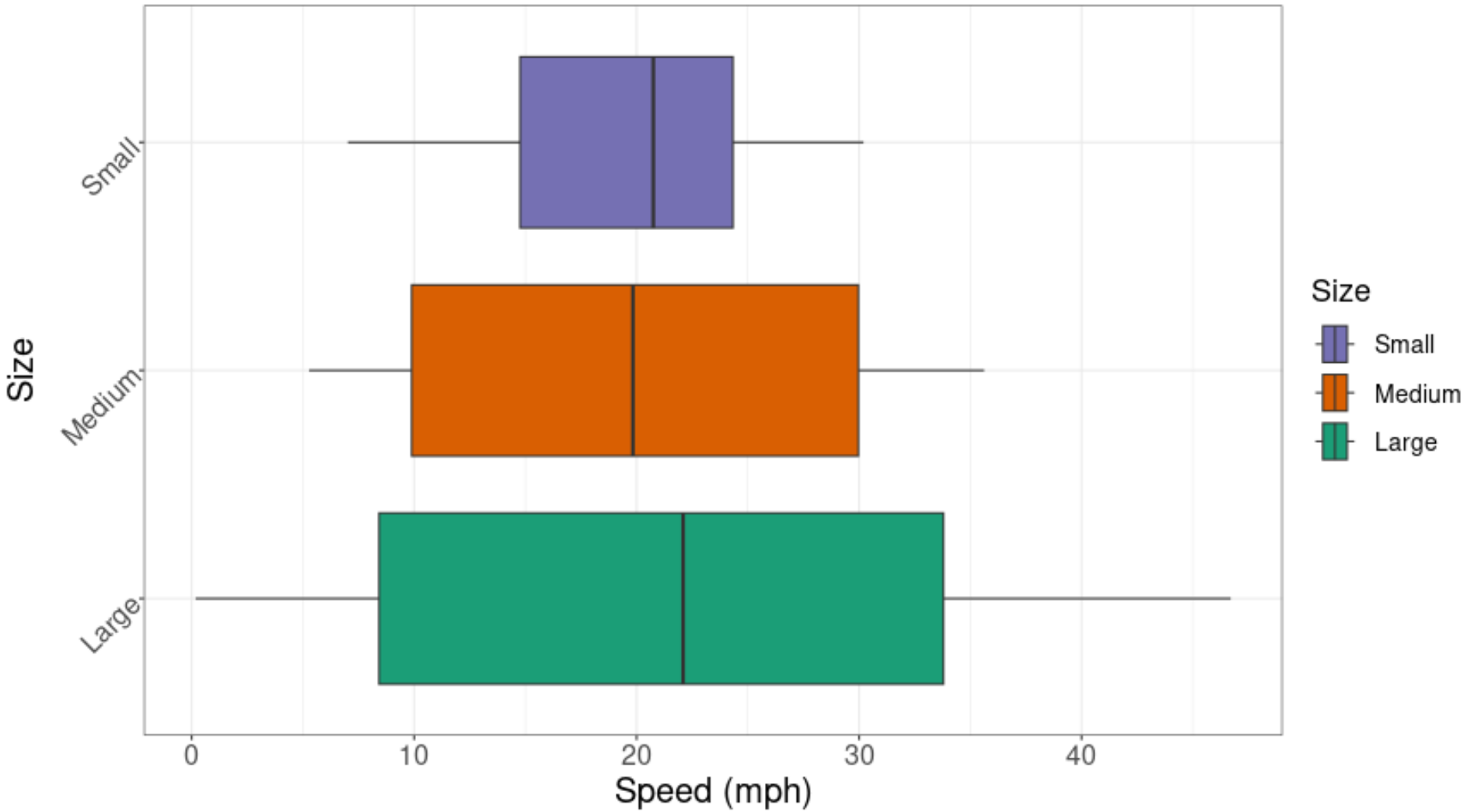
Data collected by Professor Nolte:
400 dogs from 8 different breeds (random sample of size 50 for each breed)

Each set has 25 black dogs and 25 dogs of one other color. For each dog, we have recorded land speeds in miles per hour (mph)

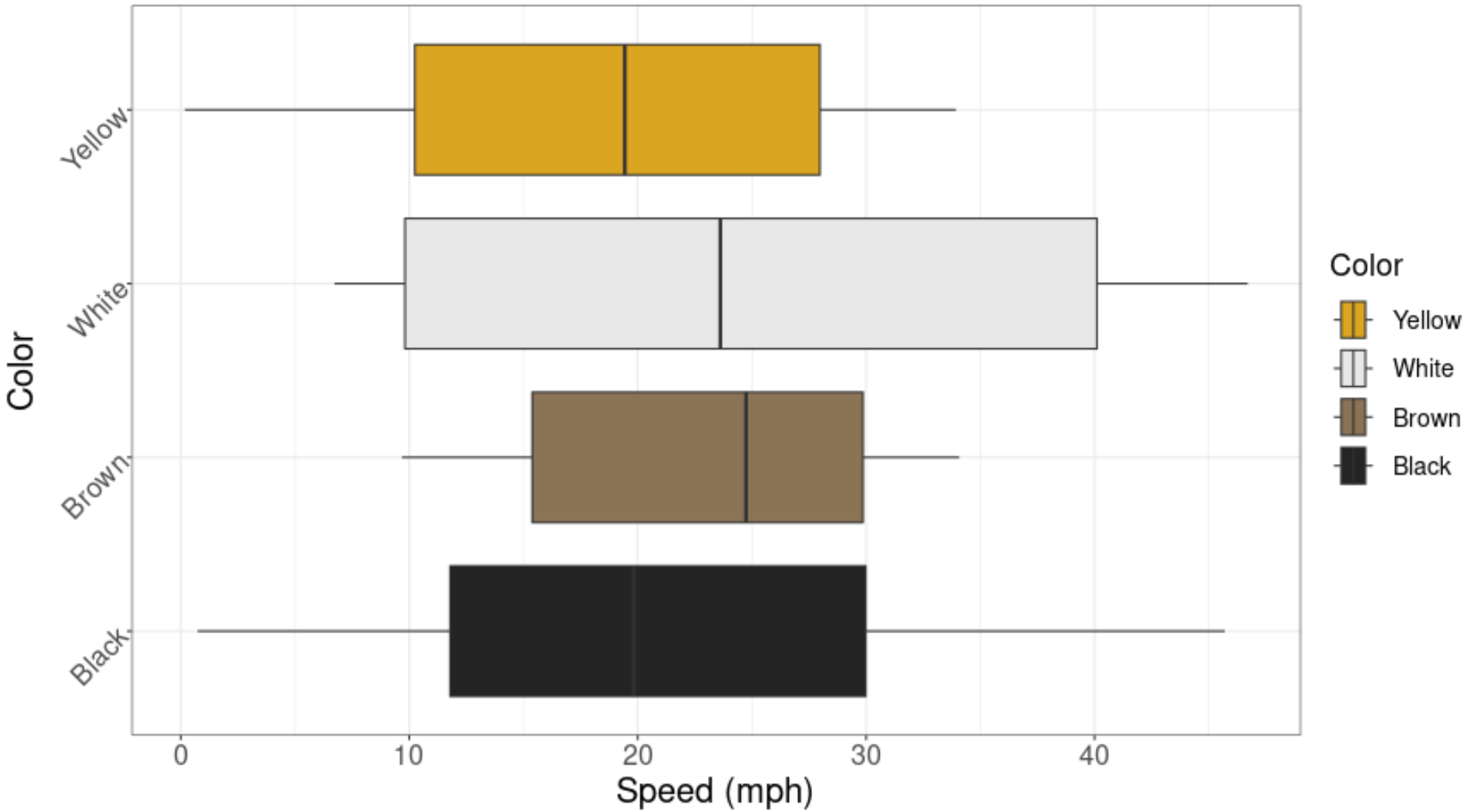
Breed	Size	Other Color	N
Chihuahua	Small	Brown	50
Corgie	Small	Yellow	50
Poodle	Medium	Brown	50
Bulldog	Medium	White	50
Saint Bernard	Large	Yellow	50
German Shepard	Large	Yellow	50
Mastiff	Large	Yellow	50
Greyhound	Large	White	50

What variables will do best in helping me predict speed? Let's find out

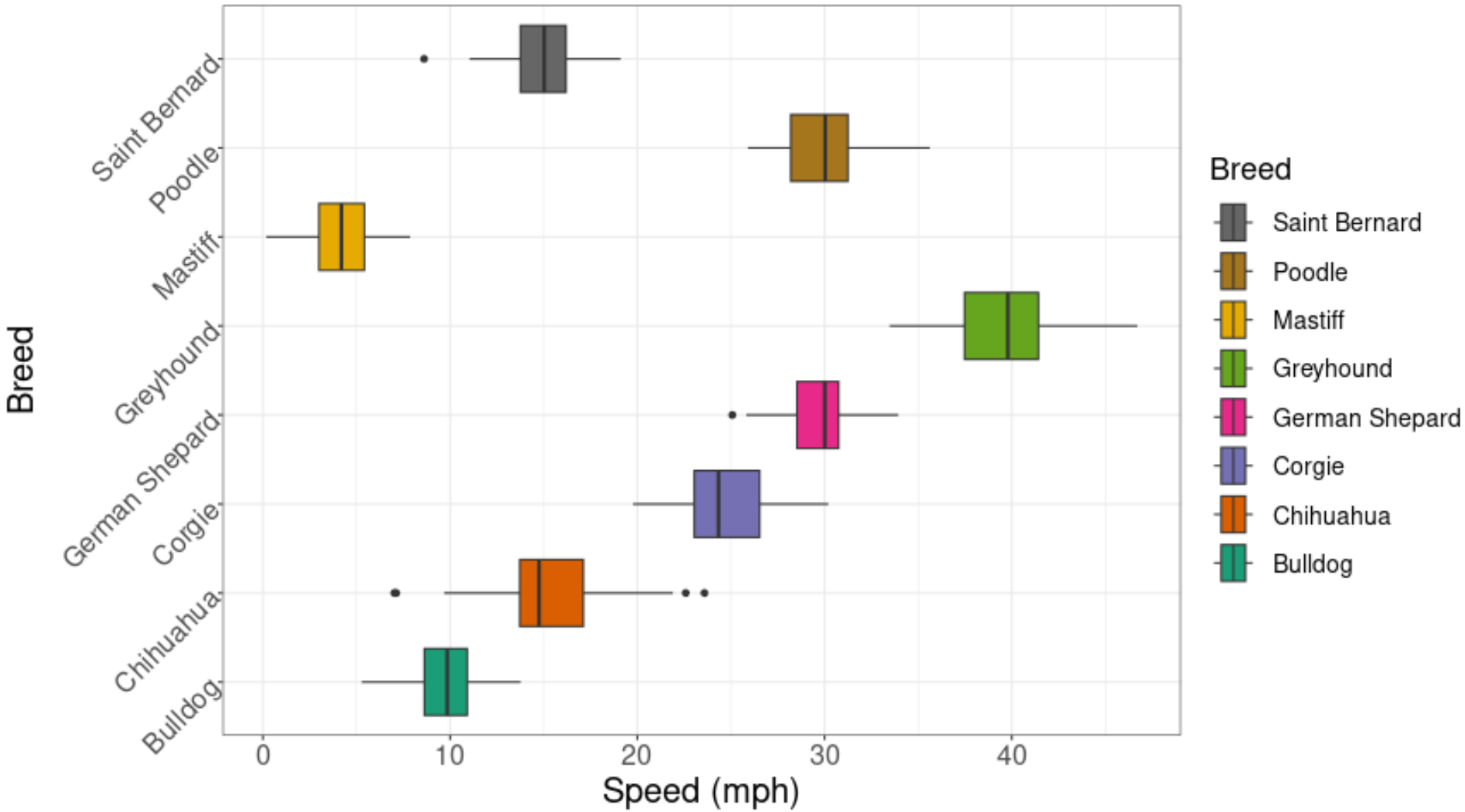
Dog Size


Size
Small
Medium
Large
0
10
20
30
40
Speed (mph)
Size
Small
Medium
Large

Dog Color


Color
Yellow
White
Brown
Black
0
10
20
30
40
Speed (mph)
Color
Yellow
White
Brown
Black

Dog Breed


Breed
Saint Bernard
Poodle
Mastiff
Greyhound
German Shepard
Corgie
Chihuahua
Bulldog
0
10
20
30
40
Speed (mph)
Breed
Saint Bernard
Poodle
Mastiff
Greyhound
German Shepard
Corgie
Chihuahua
Bulldog

The General Idea

The total variability of a sample can be broken into two parts:

- Variability within each group
- Variability between groups
  - how different are the group means/medians

How did variability *between groups* and *within groups* compare when we looked at dogs grouped by size versus by color or by breed?

Types of Variability

A common metric for describing variability is the 'sum of squares' giving total squared distance between observations and the overall mean

$$\text{SS}_{\text{total}} = \text{SST} = \sum_{\text{all } i,j} (x_{ij} - \overline{x})^2$$

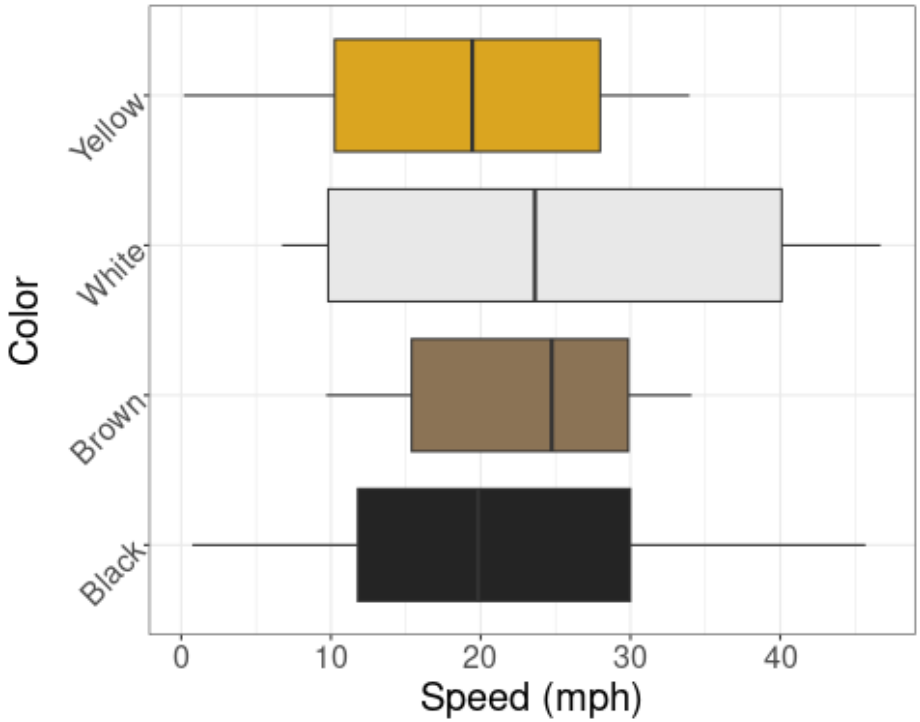
where $i = 1, \ldots, k$ indicates the group and $j = 1, \ldots, n$ indicates the observations within the group.

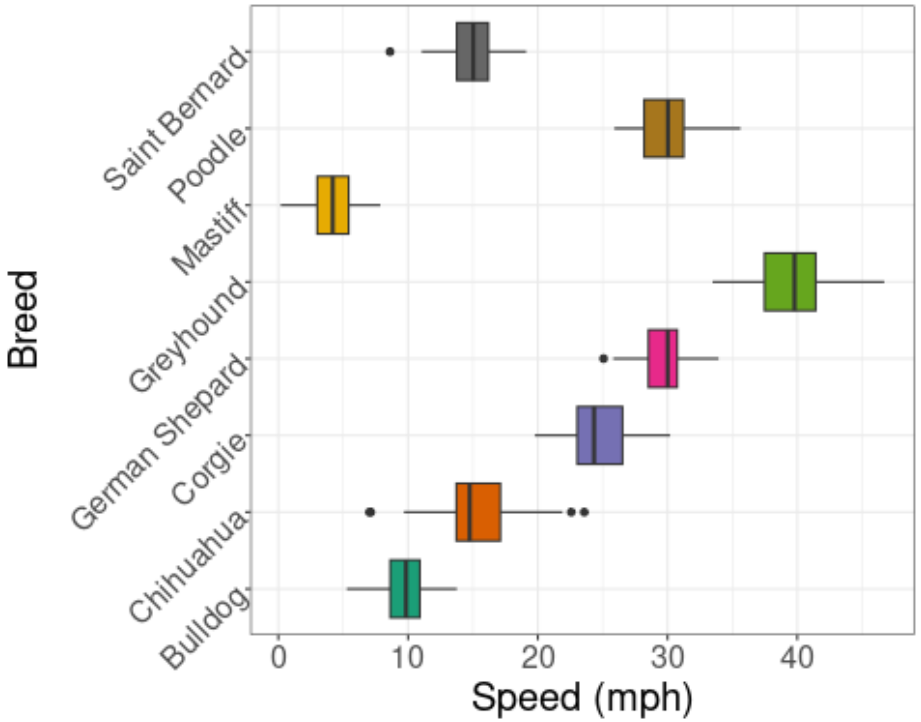
We can always decompose this into two other sums of squares

$$\underbrace{\sum (x_{ij} - \overline{x})^2}_{\substack{\text{Total} \\ \text{variability}}} = \underbrace{\sum_{j=1}^{n} (x_{ij} - \overline{x}_i)^2}_{\substack{\text{Variability} \\ \text{within groups}}} + \underbrace{\sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})^2}_{\substack{\text{Variability} \\ \text{between groups}}}$$

Types of Variability

$$\underbrace{\sum (x_{ij} - \overline{x})^2}_{\text{Total variability}} = \underbrace{\sum_{j=1}^{n} (x_{ij} - \overline{x}_i)^2}_{\text{Variability within groups}} + \underbrace{\sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})^2}_{\text{Variability between groups}}$$





Types of Variability

Recall our null hypothesis for ANOVA

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_A : \text{at least one } \mu_i \neq \mu_j$$

- Low within group variabilty $\Leftrightarrow$ tight knit clearly defined group around a mean
- High between group variability $\Leftrightarrow$ the groups are clearly distinct from one another

Within-group Variability

This measures how much variability there is within each group

The summation corresponding to within-group variability is given a special name: *sum of squared errors (SSE)* or SS_{error}

$$\text{SSE} = \sum_{j=1}^{k} (x_{ij} - \overline{x}_i)^2$$

This can also be written as the *weighted* sum of group standard deviations

$$\text{SSE} = \sum_{i=1}^{k} (n_i - 1)s_i^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2$$

Intuition: SSE is finding how far each observation is from it's group mean

Variability between groups

This describes how different each of the groups are from one another

Also known as the sum of squares between groups (SSG), we can compute it by finding the weighted mean of group deviations:

$$\begin{aligned} SSG &= \sum n_i(\overline{x}_i - \overline{x})^2 \\ &= n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \cdots + n_k(\overline{x}_k - \overline{x})^2 \end{aligned}$$

Intuition: SSG is finding how different each group mean is from the overall mean (and adjusting for # of observations within that group)

Degrees of Freedom

When we calculated test-statistics before, we had a degree of freedom corresponding to how much information we used to estimate things

$$\text{SSE} = \sum_{j=1}^{n} (x_{ij} - \overline{x}_i)^2, \text{ for each of group i} = 1,\ldots, \text{k}$$

- using k group means, each has 50 observations $\rightarrow n - k$ df

$$\begin{aligned} SSG &= \sum n_i(\overline{x}_i - \overline{x})^2 \\ &= n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \cdots + n_k(\overline{x}_k - \overline{x})^2 \end{aligned}$$

- using k groups to estimate this $\rightarrow$ df $= k - 1$

Degrees of Freedom

Frequently instead of using SSE and SSG directly, we will divide them by their respective degrees of freedom

Mean-Square-Error (Mean Square of Errors):

$$\text{MSE} = \frac{\text{SSE}}{n-k}$$

Mean-Square-Groups (Mean Square of Groups):

$$\text{MSG} = \frac{\text{SSG}}{k-1}$$

Dividing the Sums of Squares like this adjusts them to account for the number of observations and numbers of groups, and makes comparing them easier

F-statistic

Ultimately what we will use to determine the outcome of our test is the ratio of 'between group variations' and 'variation from error'

$$F = \frac{MSG}{MSE}$$

What makes the F statistic large?

- big MSG
- small MSE

F distribution

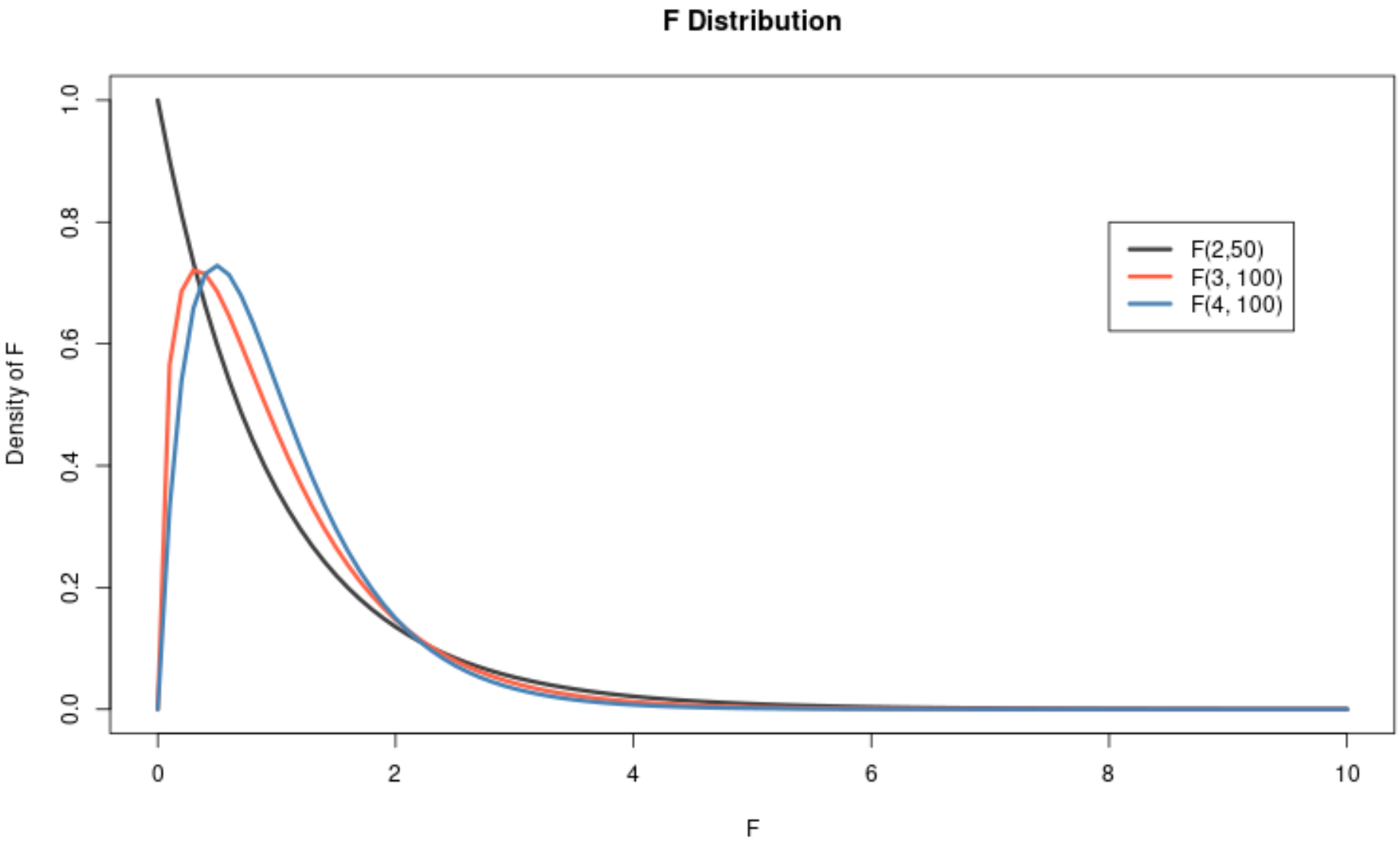
Just as we are able to use to t-distribution in finding p-values for the difference of two means, we can use the F distribution to find a p-value for assessing the null hypothesis for ANOVA

Generally speaking, we are in good shape if:

- The distributions of the groups are roughly normal
- The variances between the groups are roughly similar. Generally so long as the standard deviation of one group doesn't exceed twice that of another

Similar to the t-distribution, the F distribution is associated with degrees of freedom, in this case two different df's for each of MSG and MSE

F distribution



Formulas

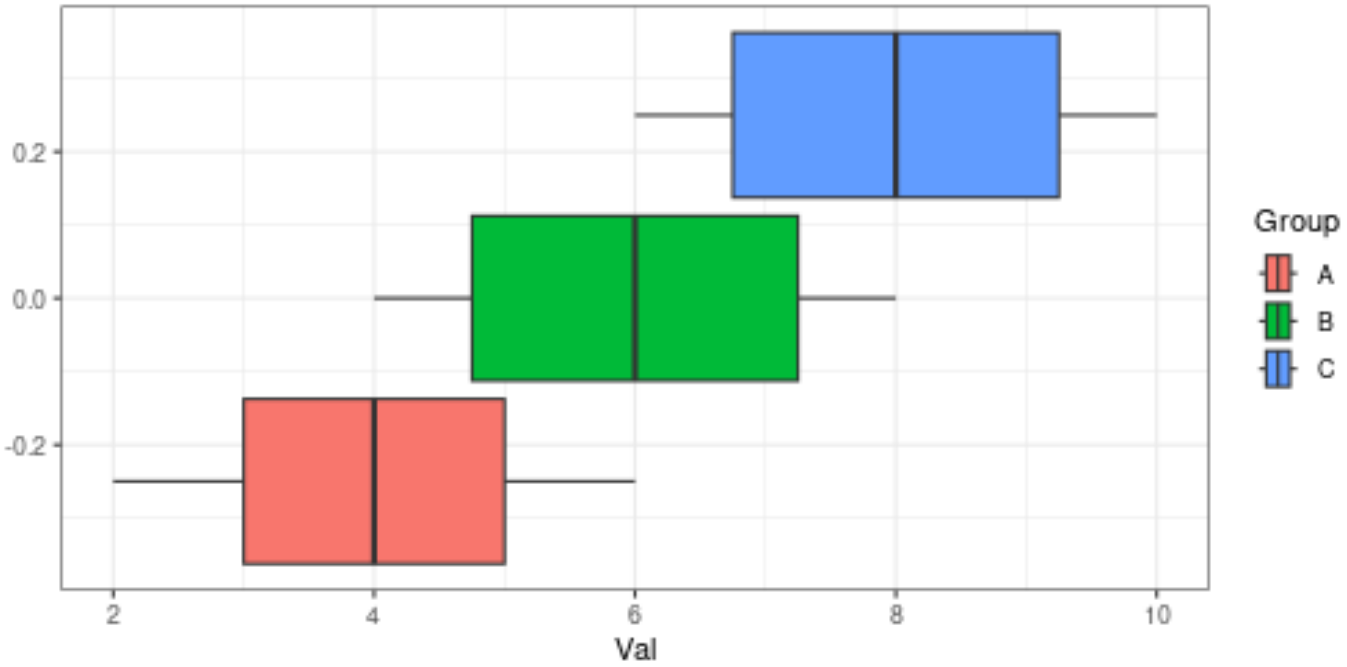
$$\underbrace{\sum (x_{ij} - \overline{x})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})^2}_{\text{SSG}} + \underbrace{\sum_{j=1}^{n} (x_{ij} - \overline{x}_i)^2}_{\text{SSE}}$$

- SST = SSG + SSE
- SSE = sum of squares *within groups*
- SSG = sum of squares *between groups*
- $MSG = \frac{SSG}{k-1}$
- $MSE = \frac{SSE}{n-k}$
- $F = \frac{MSG}{MSE}$

Source	df	Sum Sq	Mean Sq	F value	Pr(>F) / *p*-value
Group	k-1	SSG	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSE}$	Upper tail
Error	n-k	SSE	$MSE = \frac{SSE}{n-k}$		
Total	n - 1	SSTotal			

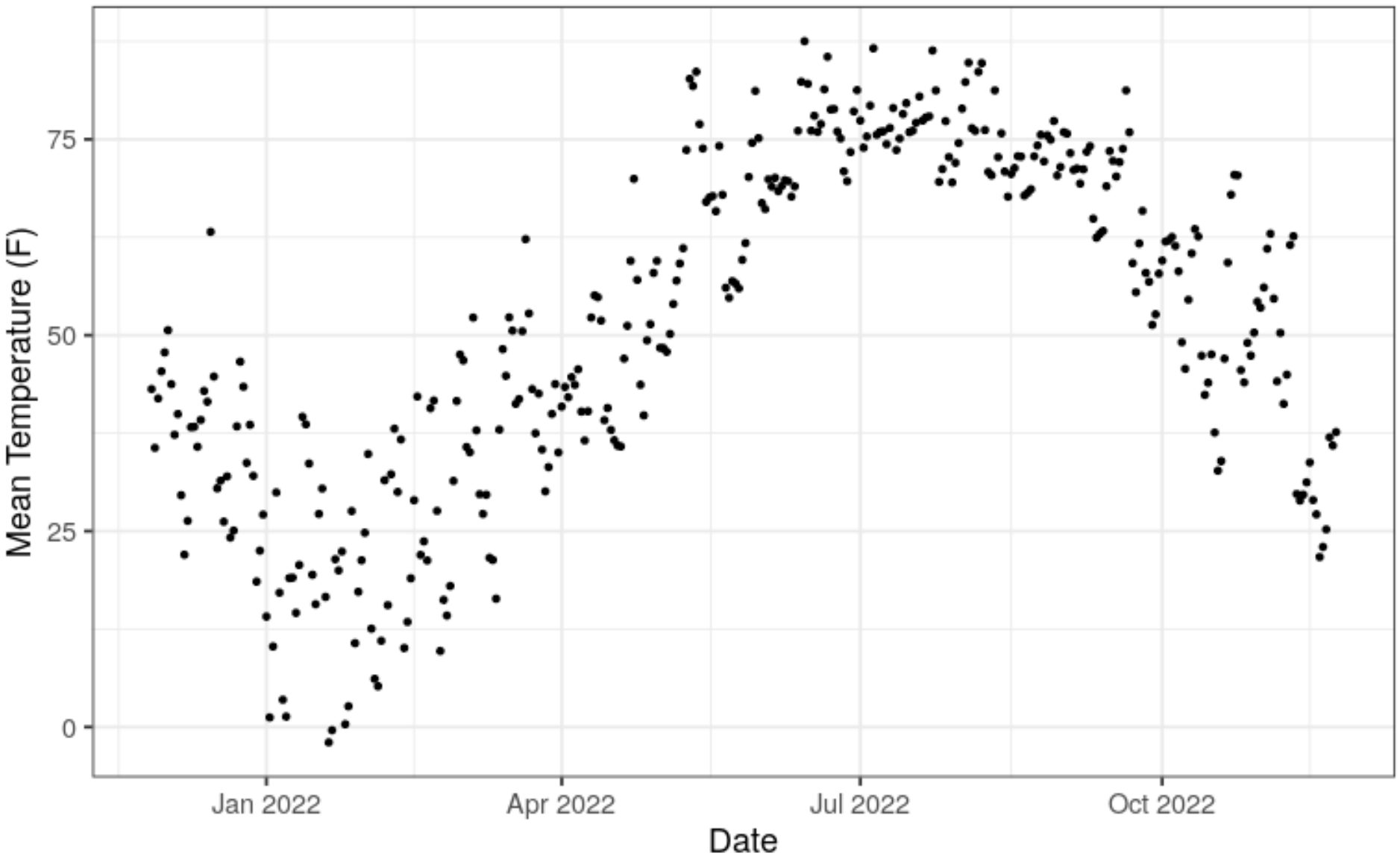
ANOVA in R

We can perform ANOVA in R using the `aov()` function

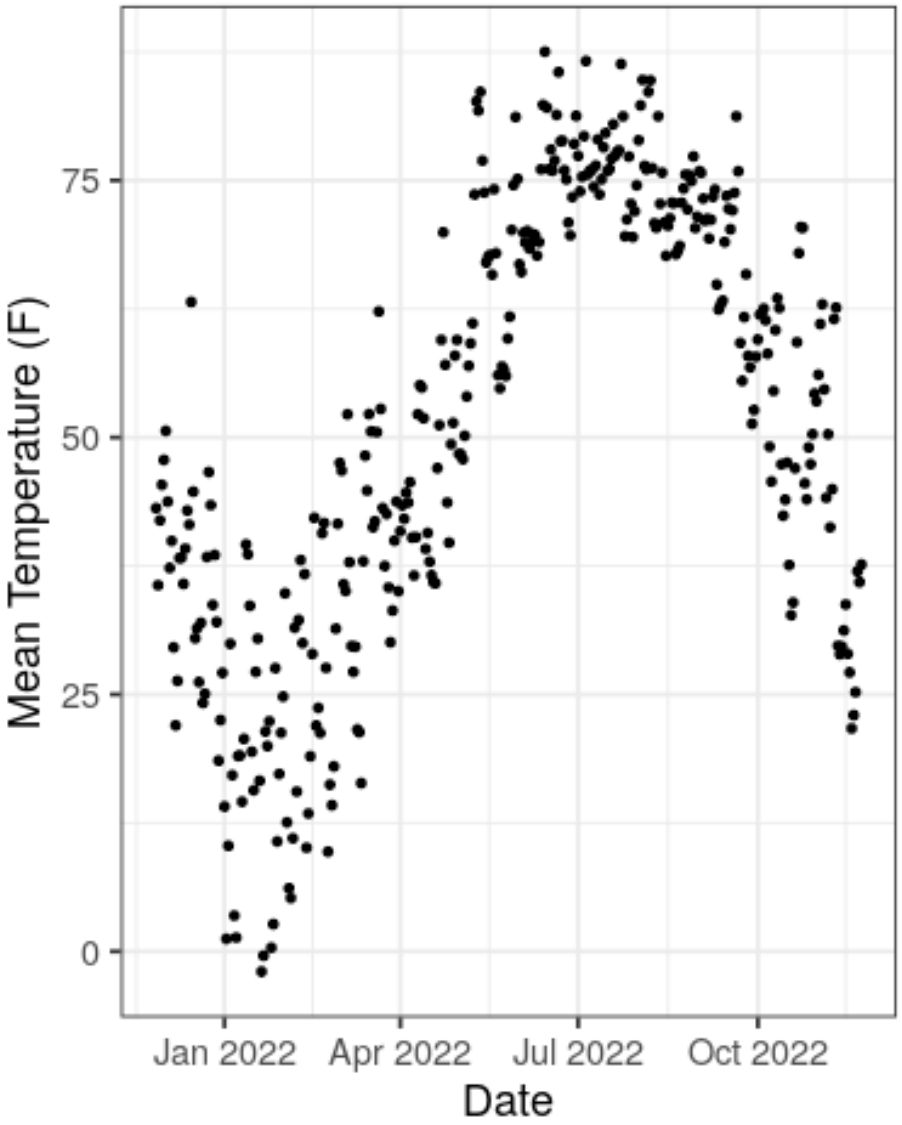


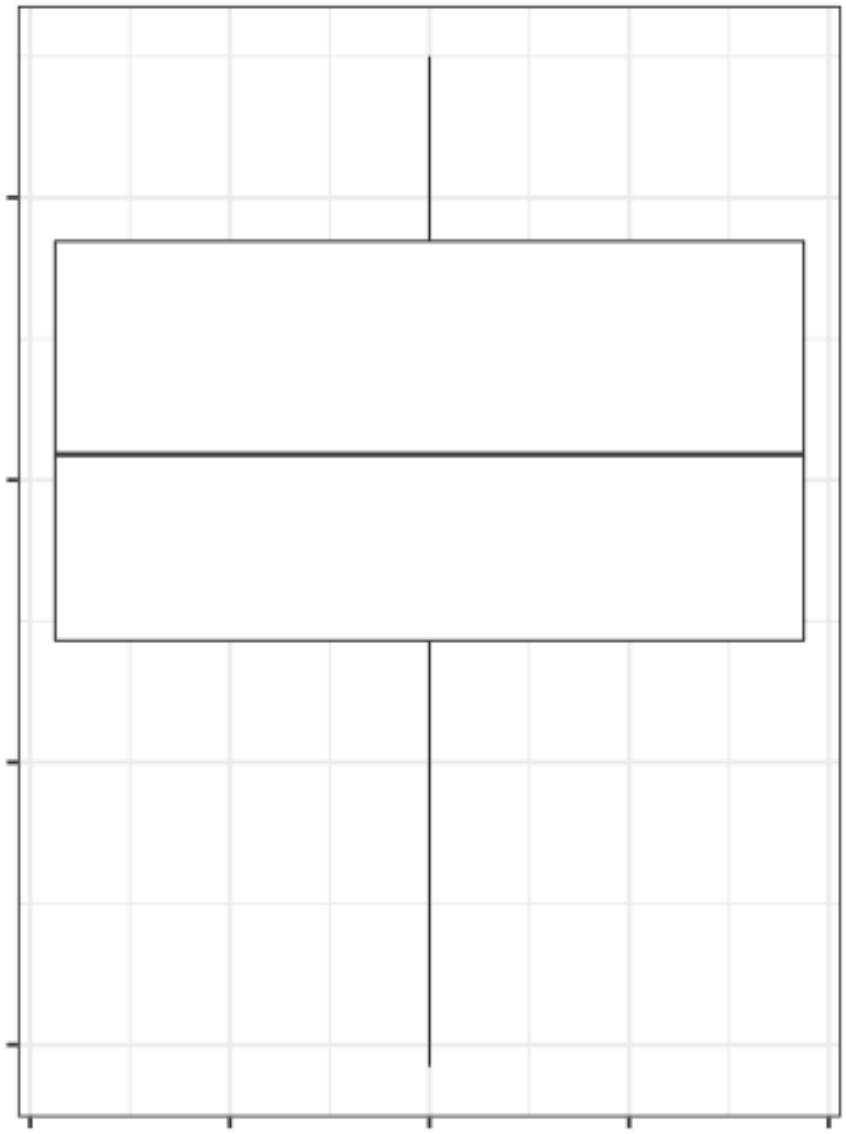

```
> aov(Val ~ Group, df) %>% summary()
             Df Sum Sq Mean Sq F value Pr(>F)
Group         2   35.7    17.9    5.95   0.02 *
Residuals    10   30.0     3.0
```

Annual Temperature – Grinnell

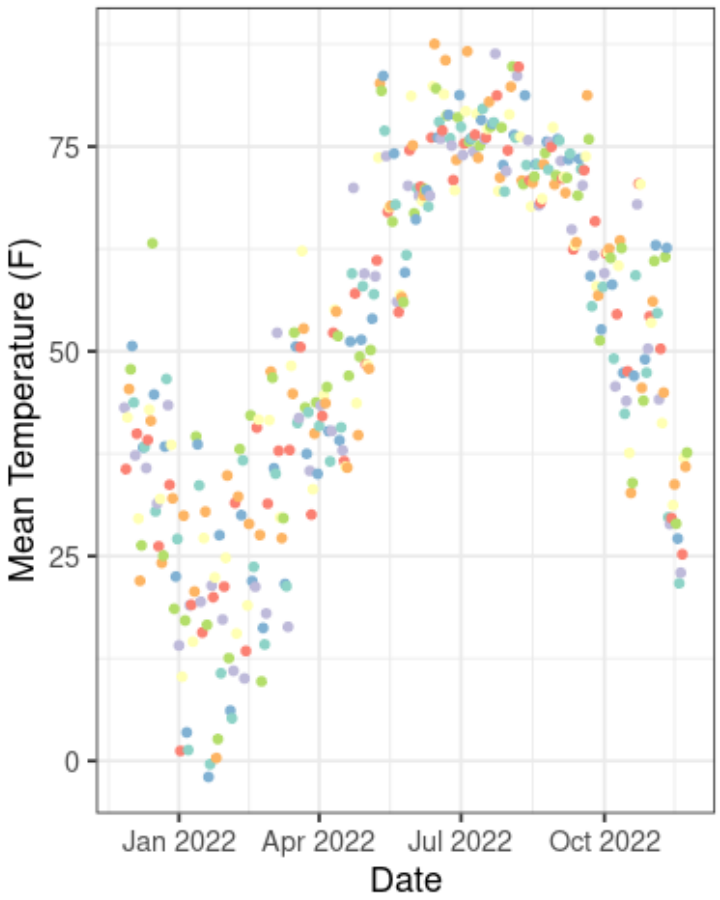


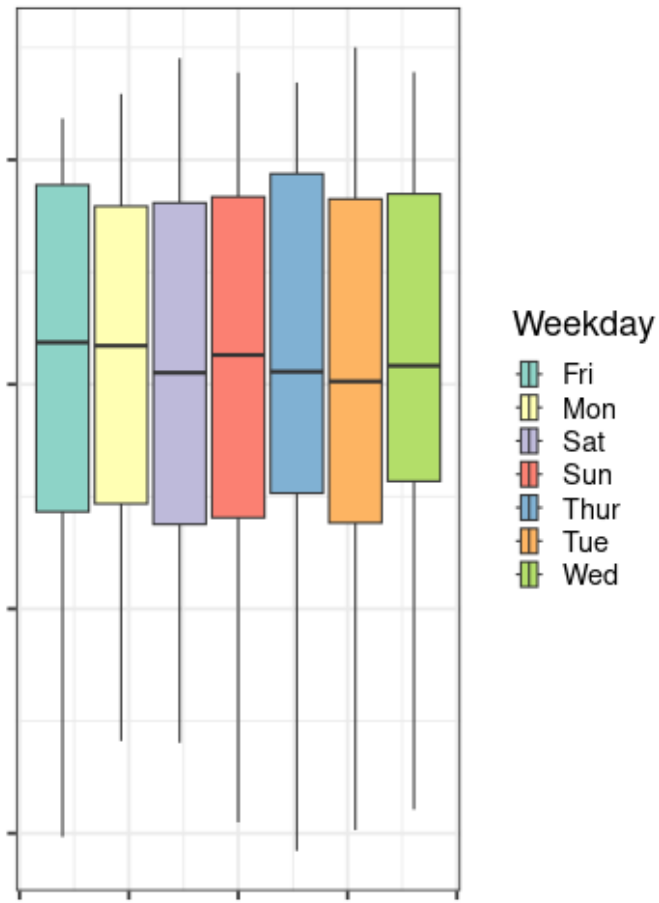
Annual Temperature





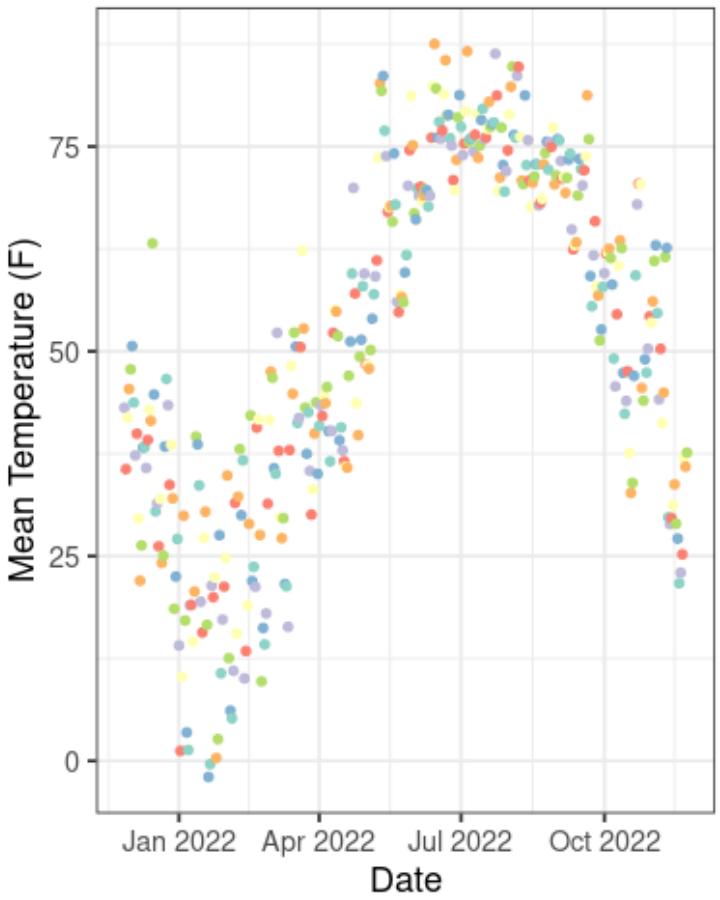
Temperature by Day

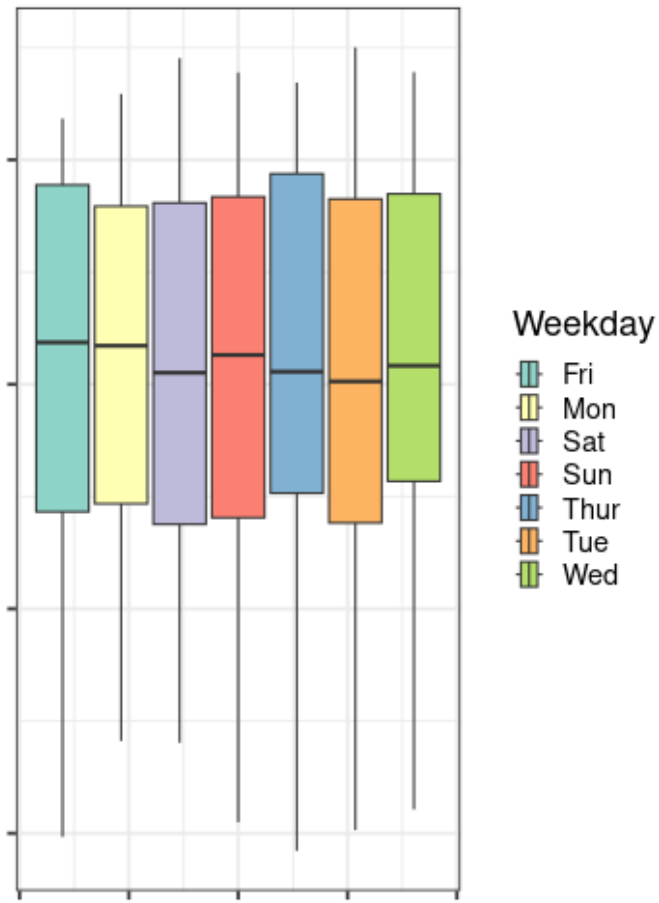




	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Weekday					
Residuals					

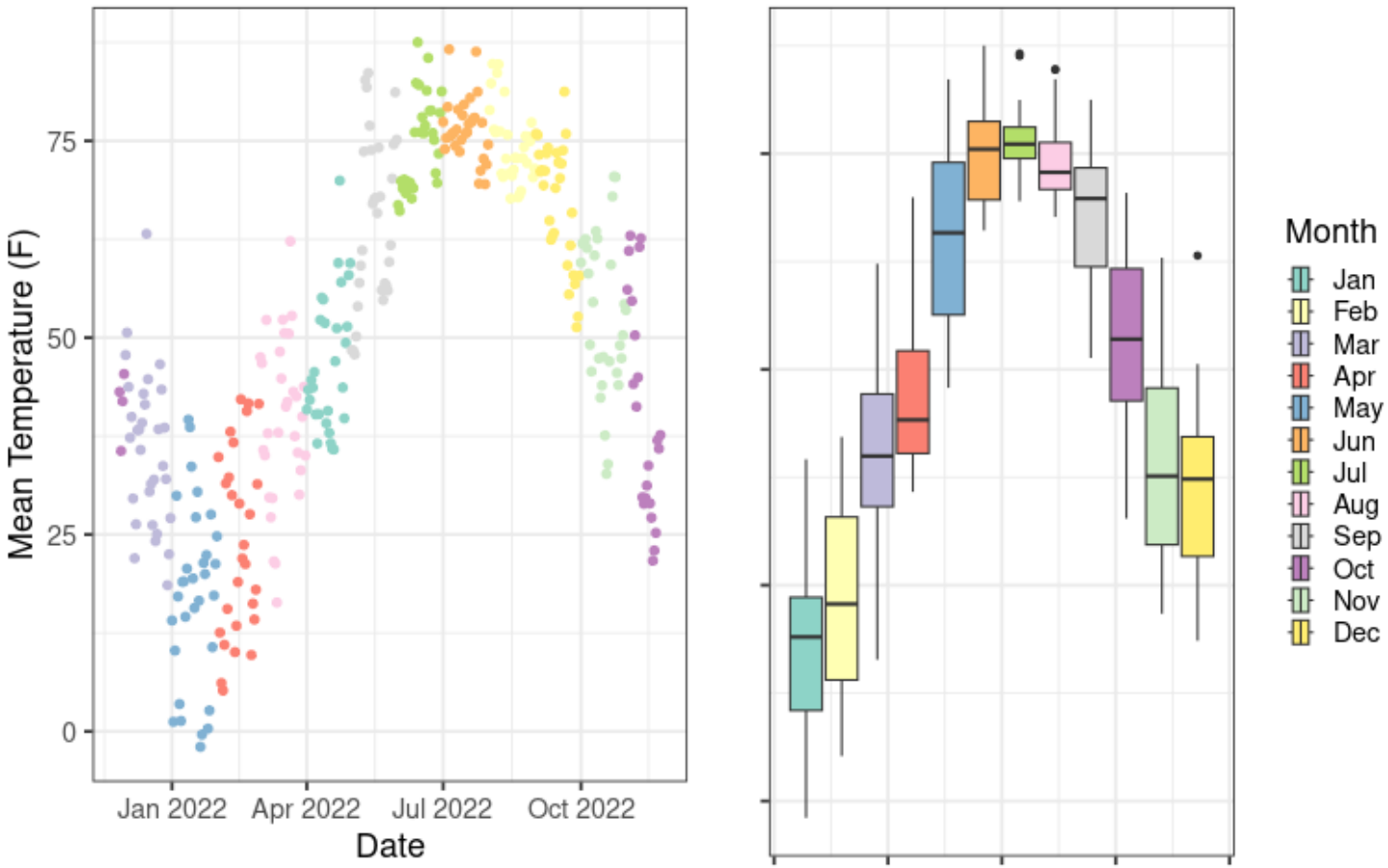
Temperature by Day





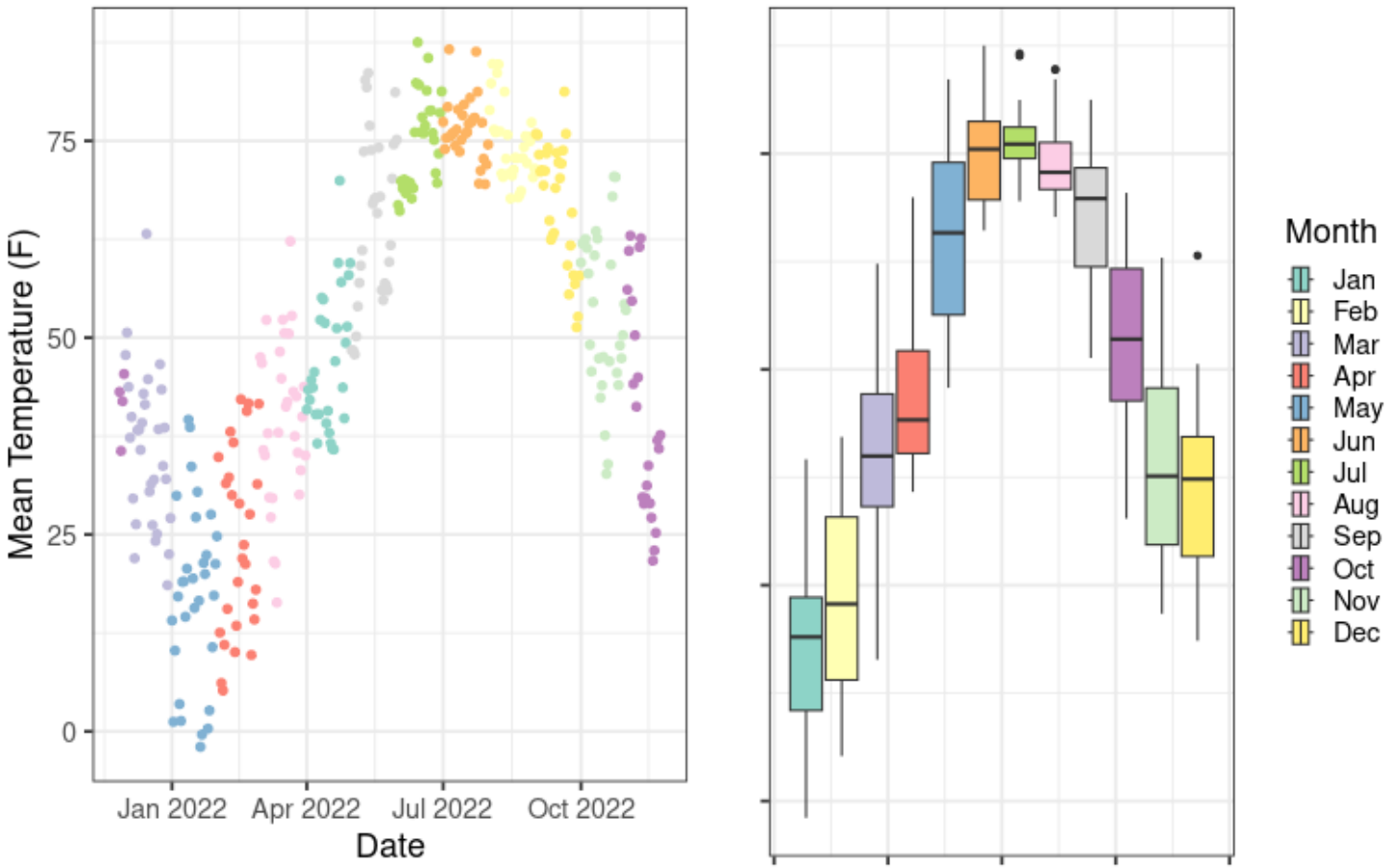
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Weekday	6	342.71	57.12	0.12	0.9939
Residuals	355	168524.83	474.72		

Temperature by Month



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Weekday					
Residuals					

Temperature by Month



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Month	11	138048.06	12549.82	142.52	<0.0001
Residuals	350	30819.48	88.06		

Review

ANOVA is a method for finding out if group means are different

- goal is to break down variability into different parts to help answer the question

- Total variability made up of 'within' and 'between group' variation
- F-statistic can be used to quantify how different these variabilities are from each other
  - F-distribution gives us our p-values

Most of the time ANOVA info is arranged in a table so we can keep track of it

You can follow the ANOVA up with individual t-tests for group differences if we find evidence of differences overall.