Analysis of Variance (ANOVA)

Grinnell College

Dec 2, 2024

Review

Identify the statistic and null hypothesis associated with the following:

- Single quantitative variable
- Two quantitative variables (two types)
- Single categorical variable
- Two categorical variables

Also be prepared to explain:

- What is the relationship between the null hypothesis and a sampling distribution?
- What is the relationship between a sampling distribution and an observed statistic?

ANOVA

Tests so far:

- Difference in means
- Difference in proportion
- Goodness of fit
- Independence of two categorical variables

Today we are extending our set of tests to include testing the difference in means in multiple groups

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?

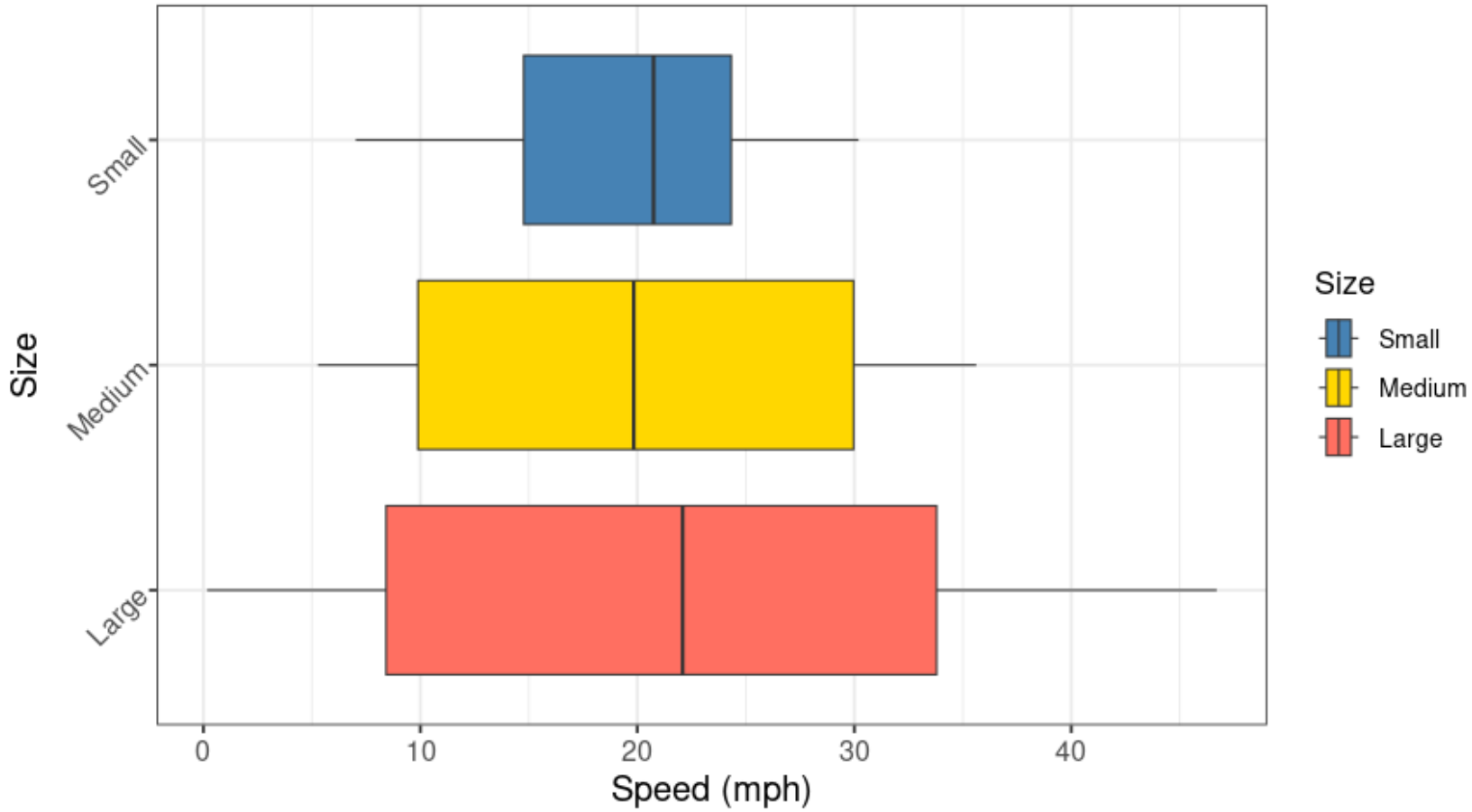
Dog Speed

Collected 400 dogs from 8 different breeds, each a sample of 50. Each set has 25 black dogs and 25 dogs of one other color. For each dog, I also recorded land speeds in miles per hour (mph)

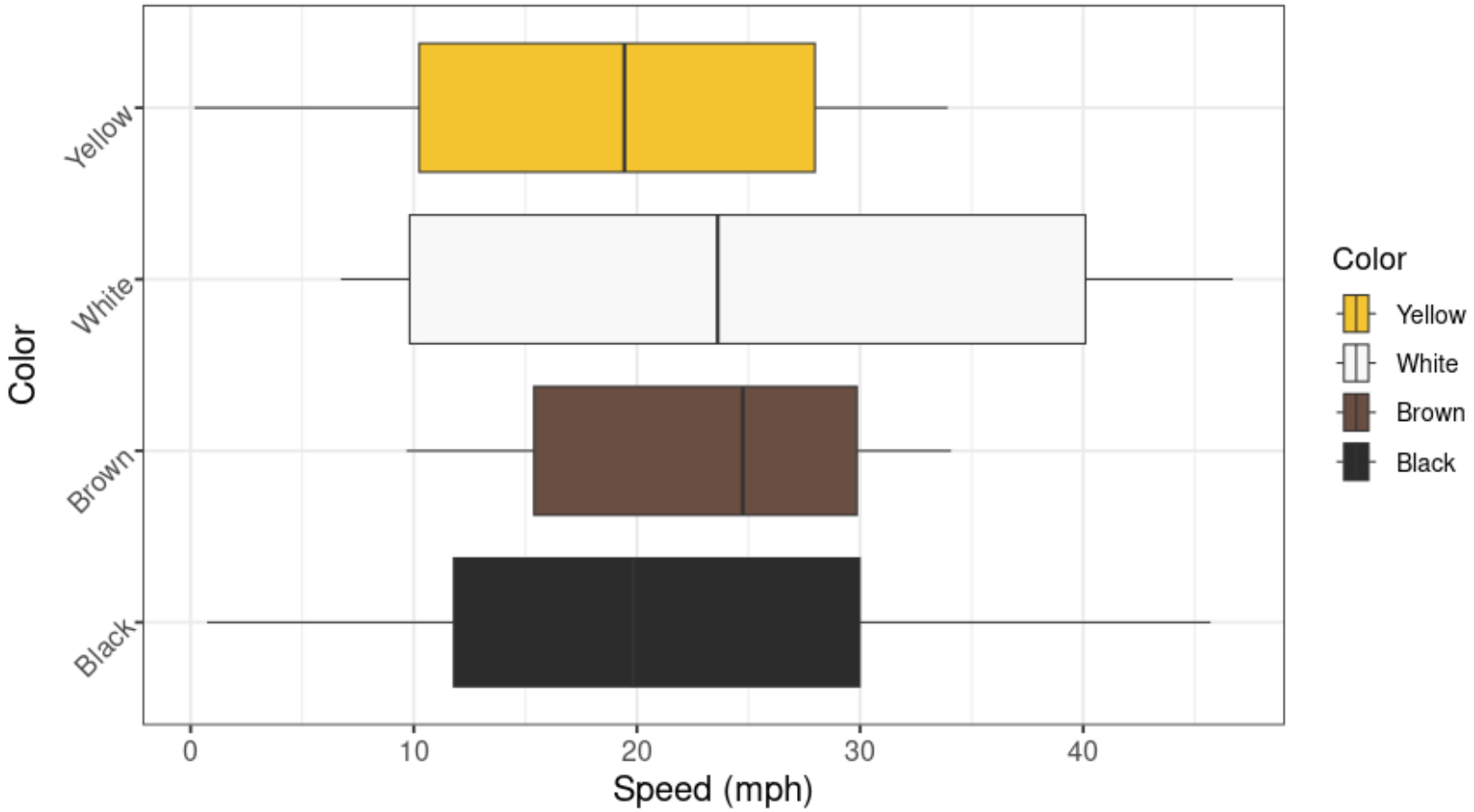
Breed	Size	Other Color	N
Chihuahua	Small	Brown	50
Corgie	Small	Yellow	50
Poodle	Medium	Brown	50
Bulldog	Medium	White	50
Saint Bernard	Large	Yellow	50
German Shepard	Large	Yellow	50
Mastiff	Large	Yellow	50
Greyhound	Large	White	50

What variables will do best in helping me predict speed?

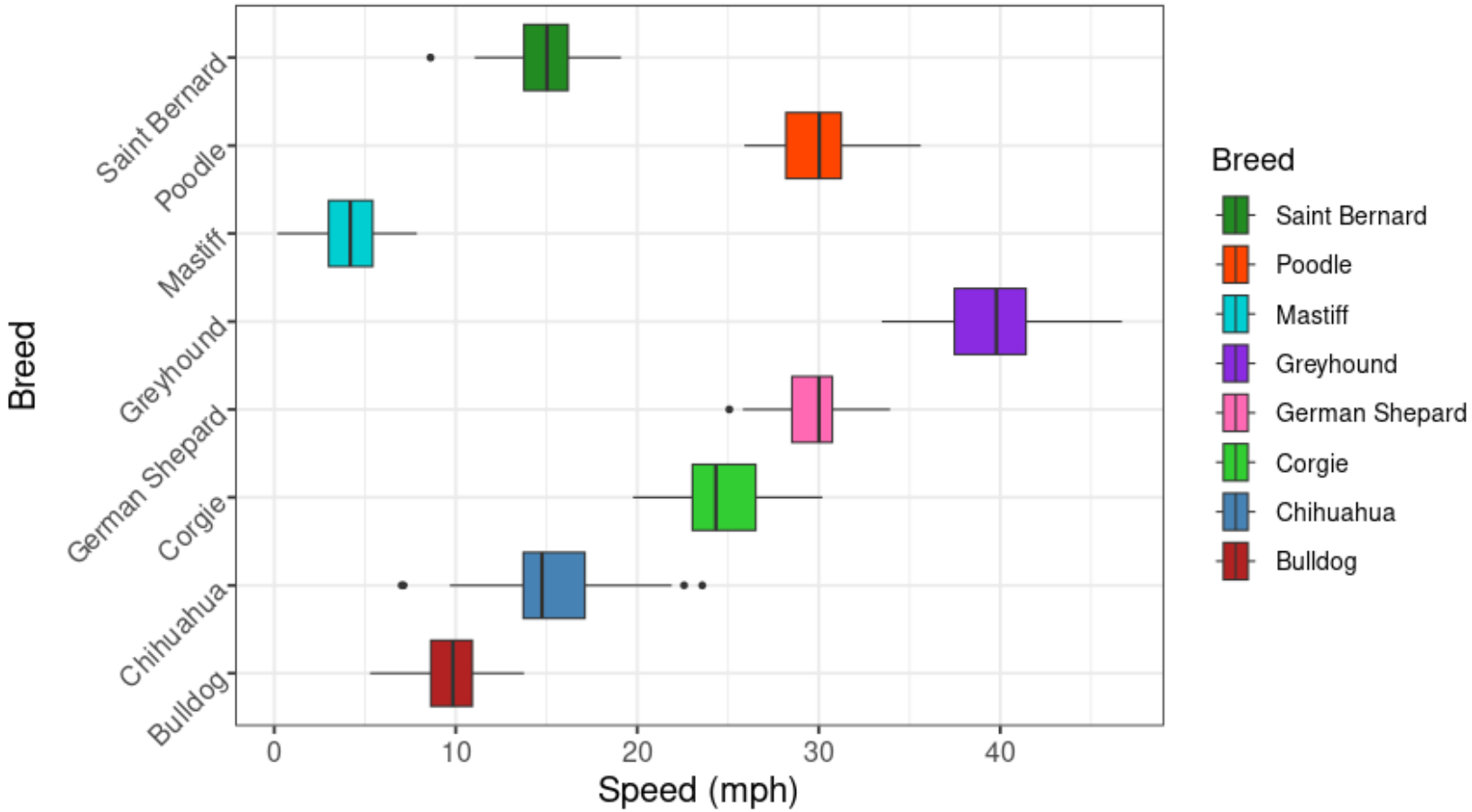
Dog Size


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 groups
- Variability between groups

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

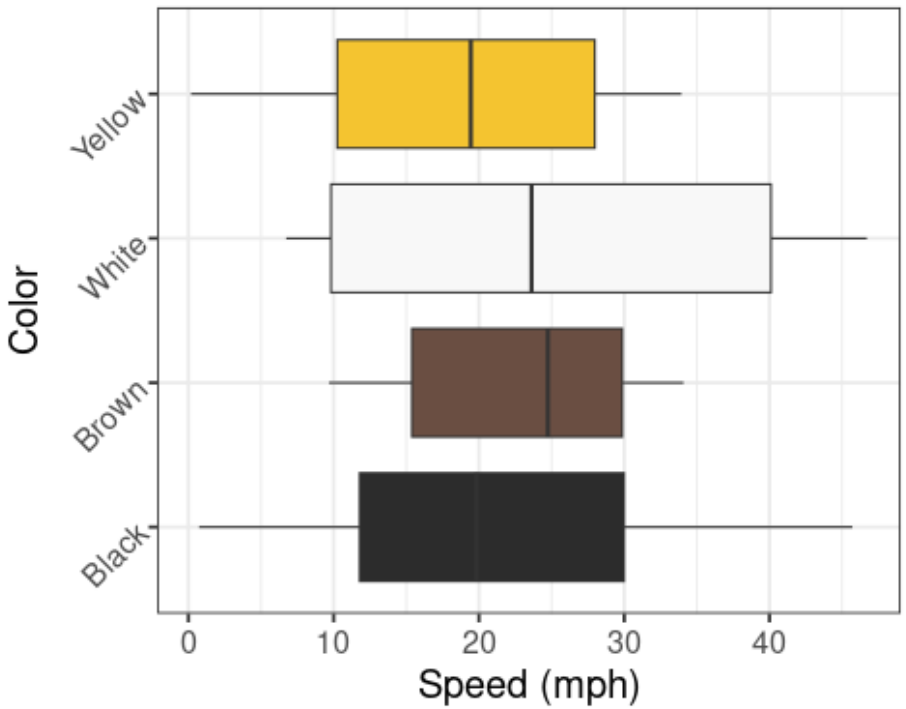
Variability and You

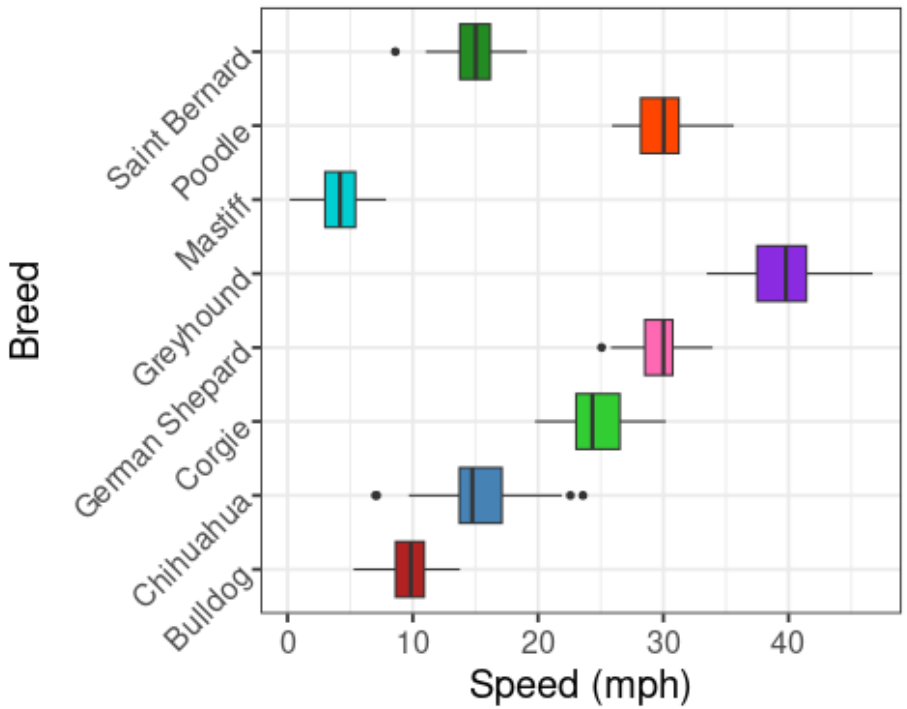
Recall our null hypothesis for ANOVA

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

- 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

Variability and You





Variability and You

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

$$\text{TSS} = \sum_{i}^{n} (x_{ij} - \overline{x})^2$$

where $i = 1, \ldots, n$ indicates the observation and $j = 1, \ldots, k$ indicates the group. We can always decompose this into a sum of two sums of squares

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

Within-group Variability

Within-group variability is associated with a standard *sum of squared errors (SSE)*

$$\text{SSE} = \sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2$$

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

$$\text{SSE} = \sum_{j=1}^{k}(n_j - 1)s_j^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2$$

As this sum involves estimating k different group means, $\overline{x}_j$, we have $n - k$ degrees of freedom, giving the mean of this metric to be

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

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:

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

As there are k groups, we can find the mean by dividing by k, less 1 degree of freedom from finding $\overline{x}$,

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

F-statistic Ratio

Finding the means of SSG and SSE help keep the metrics interpretable when the number of groups or samples increases

Ultimately, then, what determines value the outcome of our test is the ratio between group variations and variation from error

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

What makes the F statistic larger:

- MSG increases
- MSE decreases

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

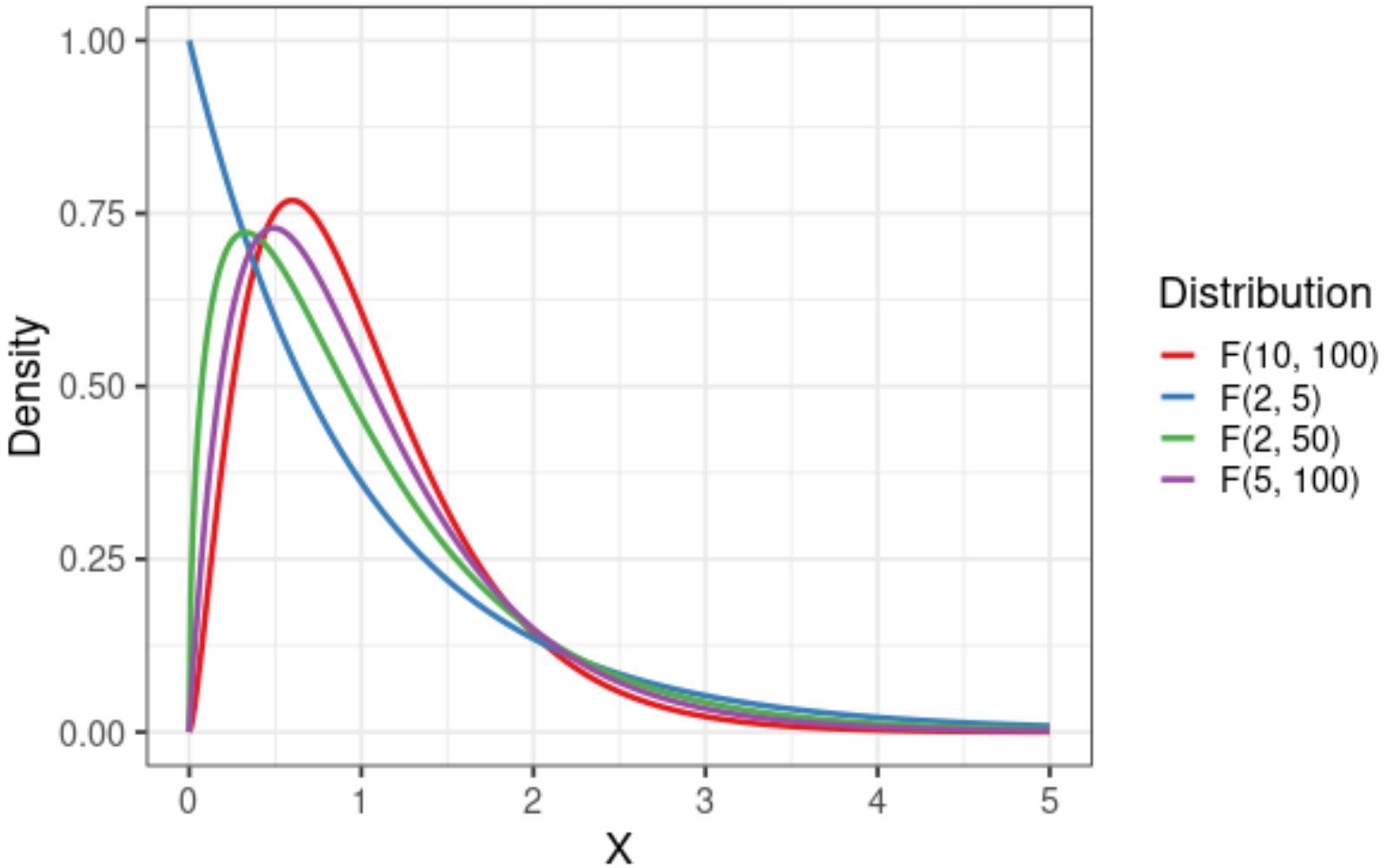
Generally speaking, we are in good shape if:

- 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

Again similar to the t-distribution, the F distribution is associated with degrees of freedom, in this case two, one for each of the mean squares in the ratio.

F distribution

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



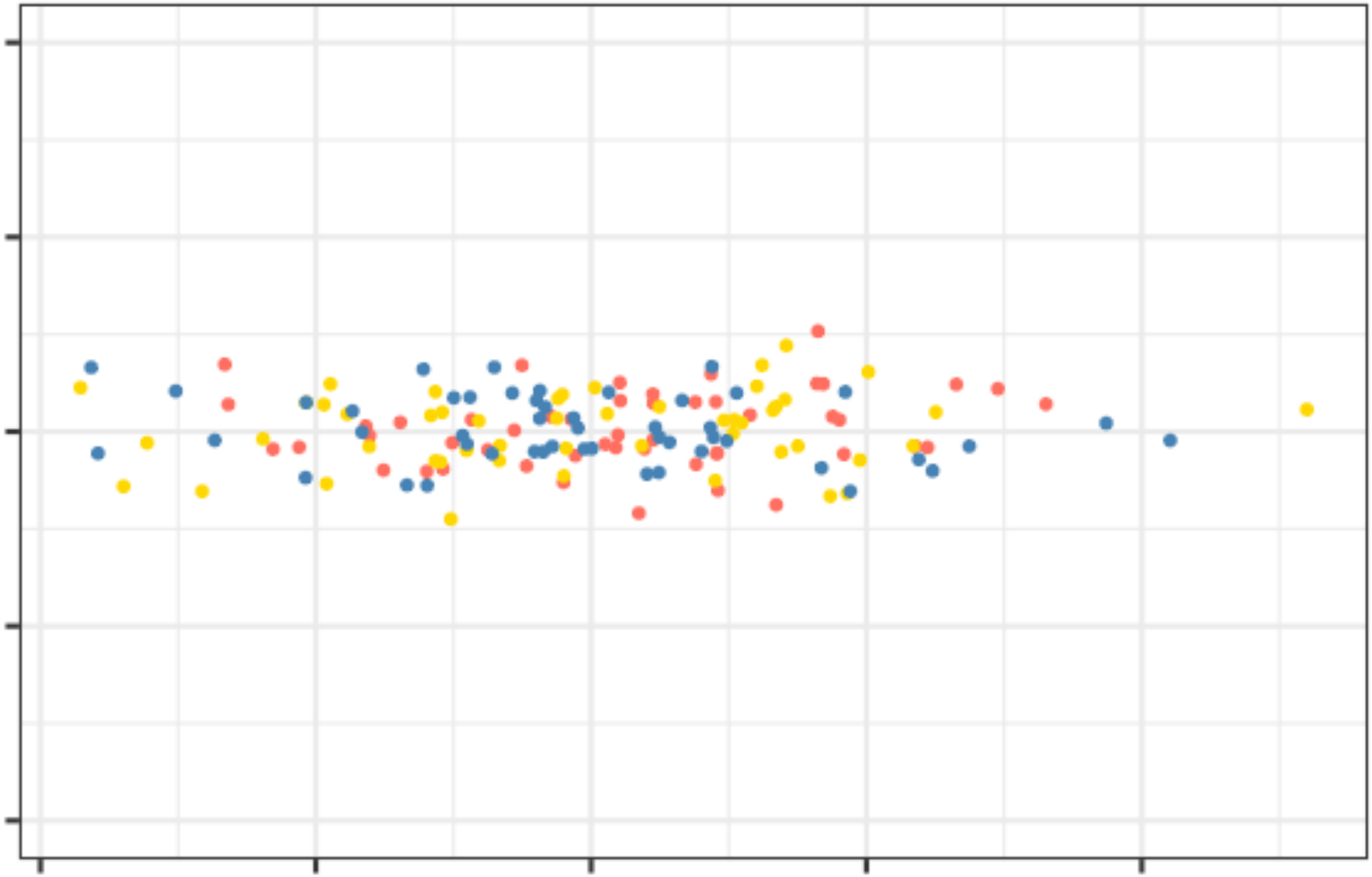
Formulas

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

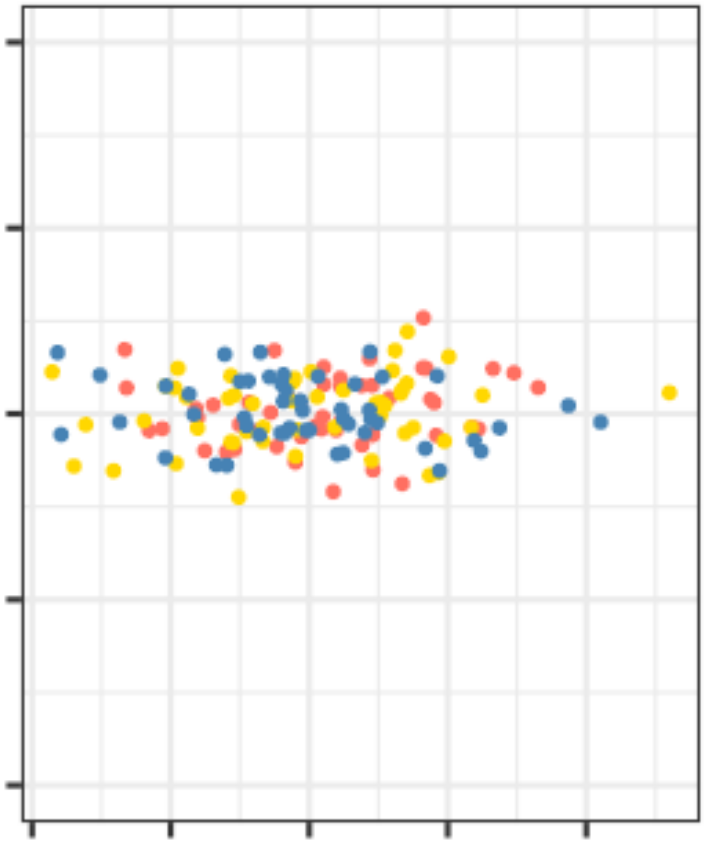
- SST = SSE + SSG
- SSE = sum of squares *within groups*
- SSG = sum of squares *between groups*
- $MSG = \frac{SSG}{k-1}$
- $MSE = \frac{SSE}{n-k}$
- $F = \frac{MSG}{MSE}$

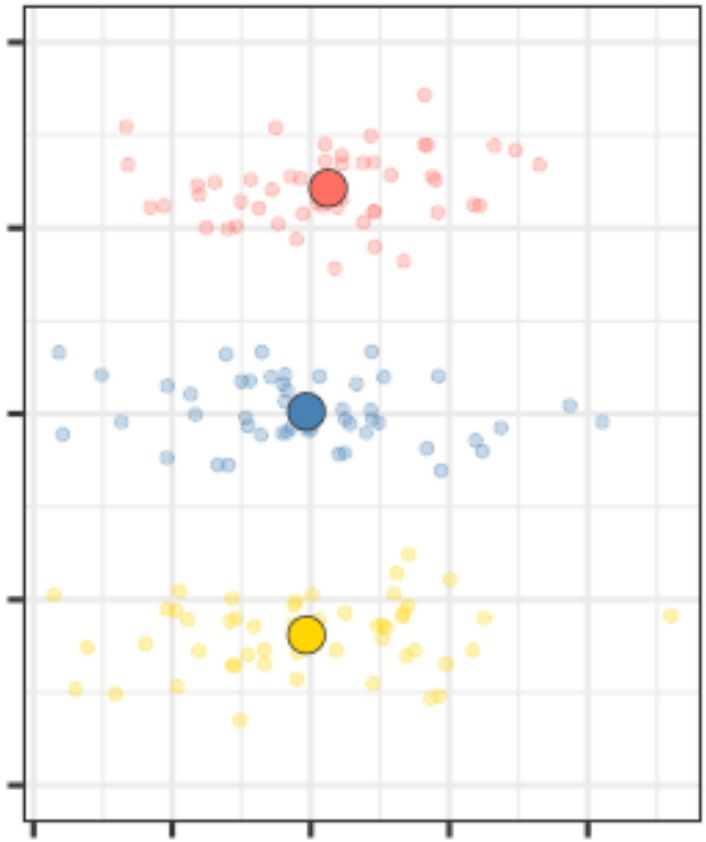
Source	df	Sum Sq	Mean Sq	F value	Pr(>F) / *p*-value
Group	k-1	SSG	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSE}$	Upper tail
Error	n-k	SSE	$MSE = \frac{SSE}{n-k}$		
Total	n - 1	SSTotal			

Example 1

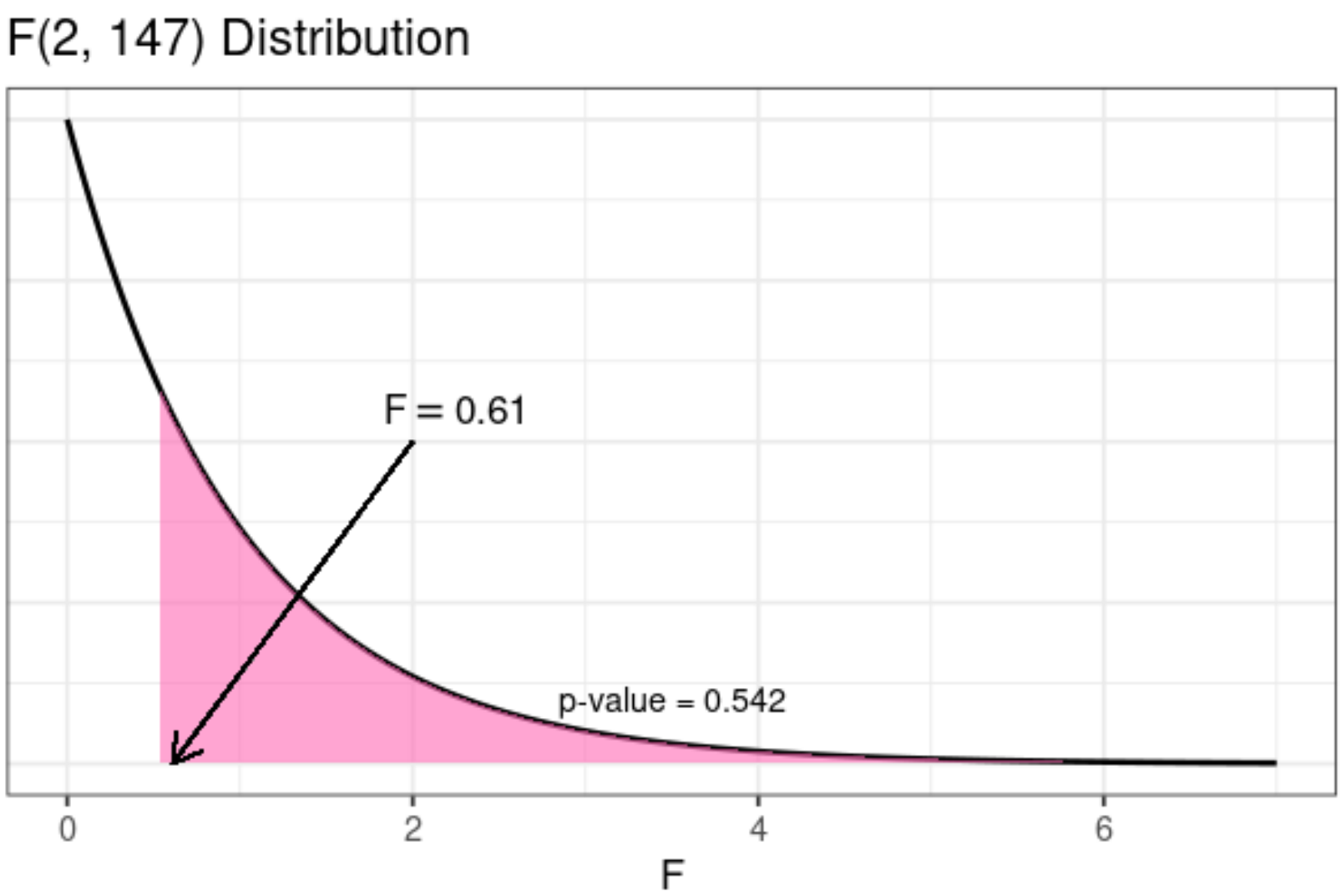


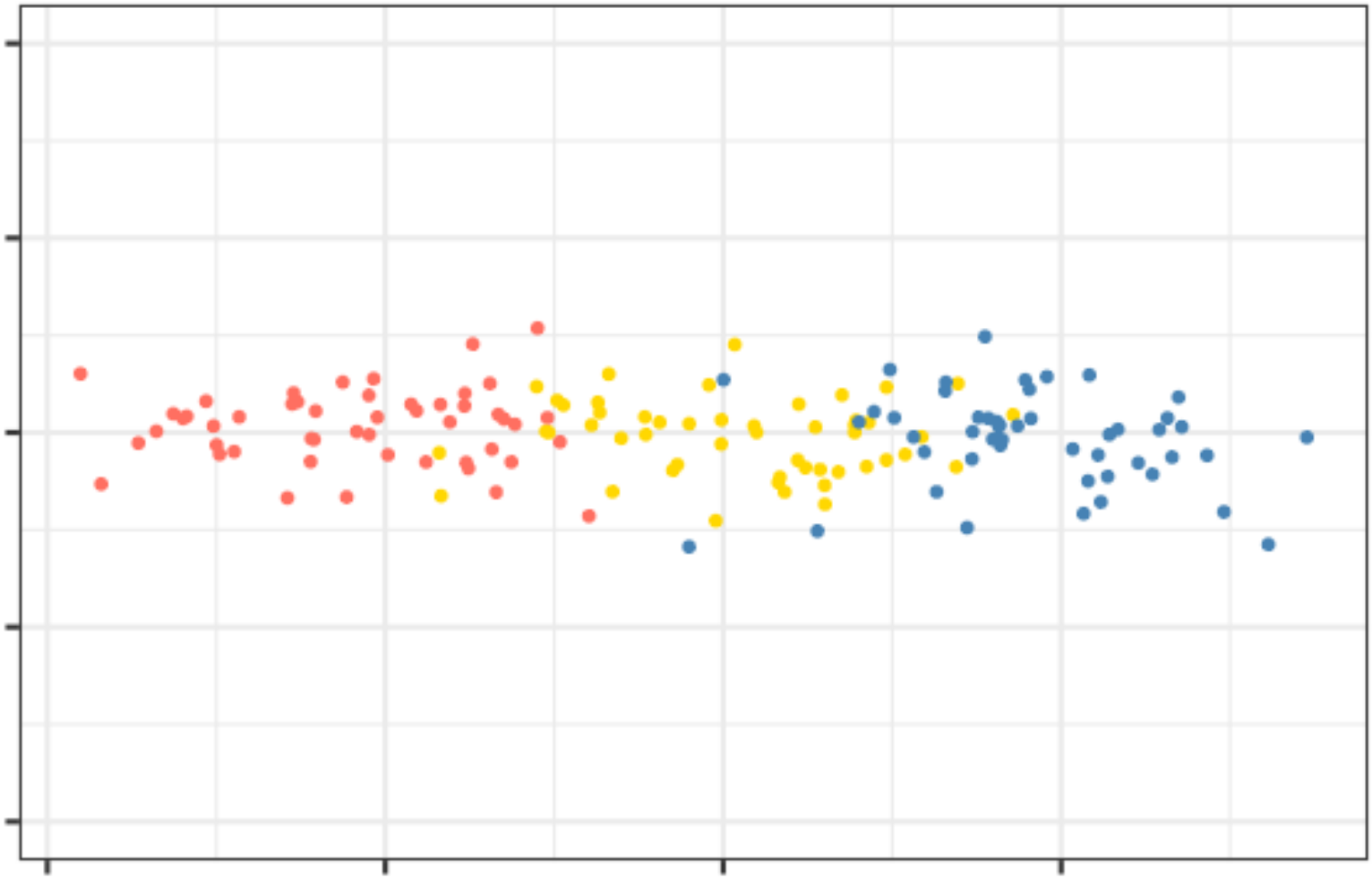
Example 1



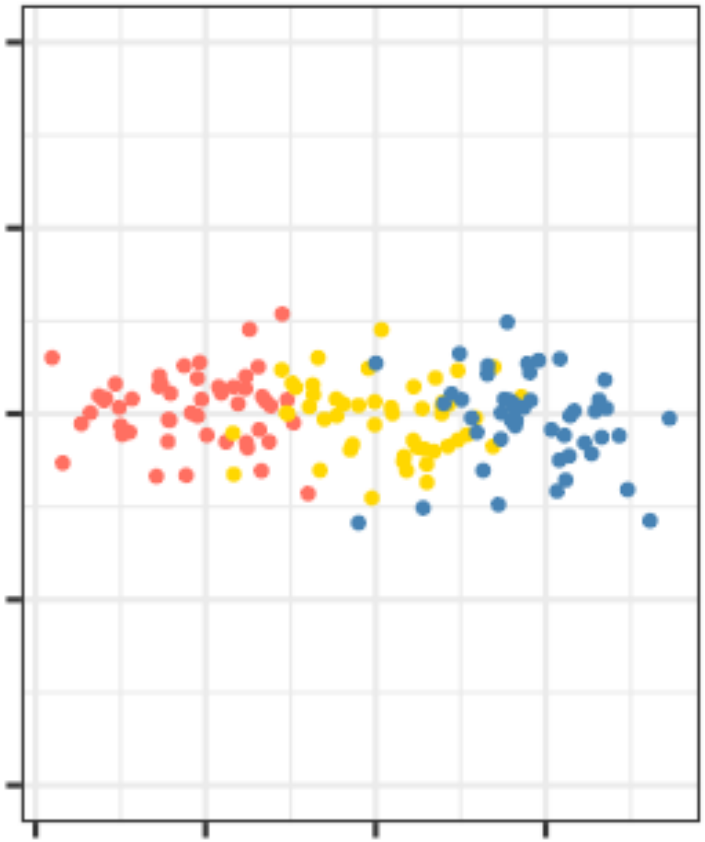


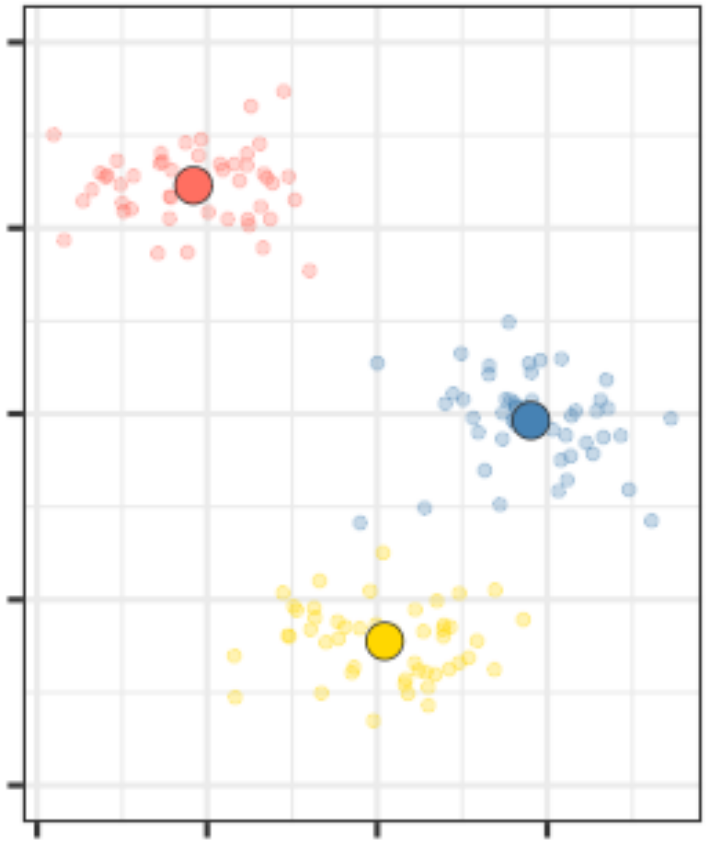
	Df	Sum Sq	Mean Sq	F value	p-value
Group	2	5.09	2.54	0.61	0.5427
Residuals	147	609.07	4.14		


F(2, 147) Distribution
F = 0.61
p-value = 0.542
0
2
4
6
F

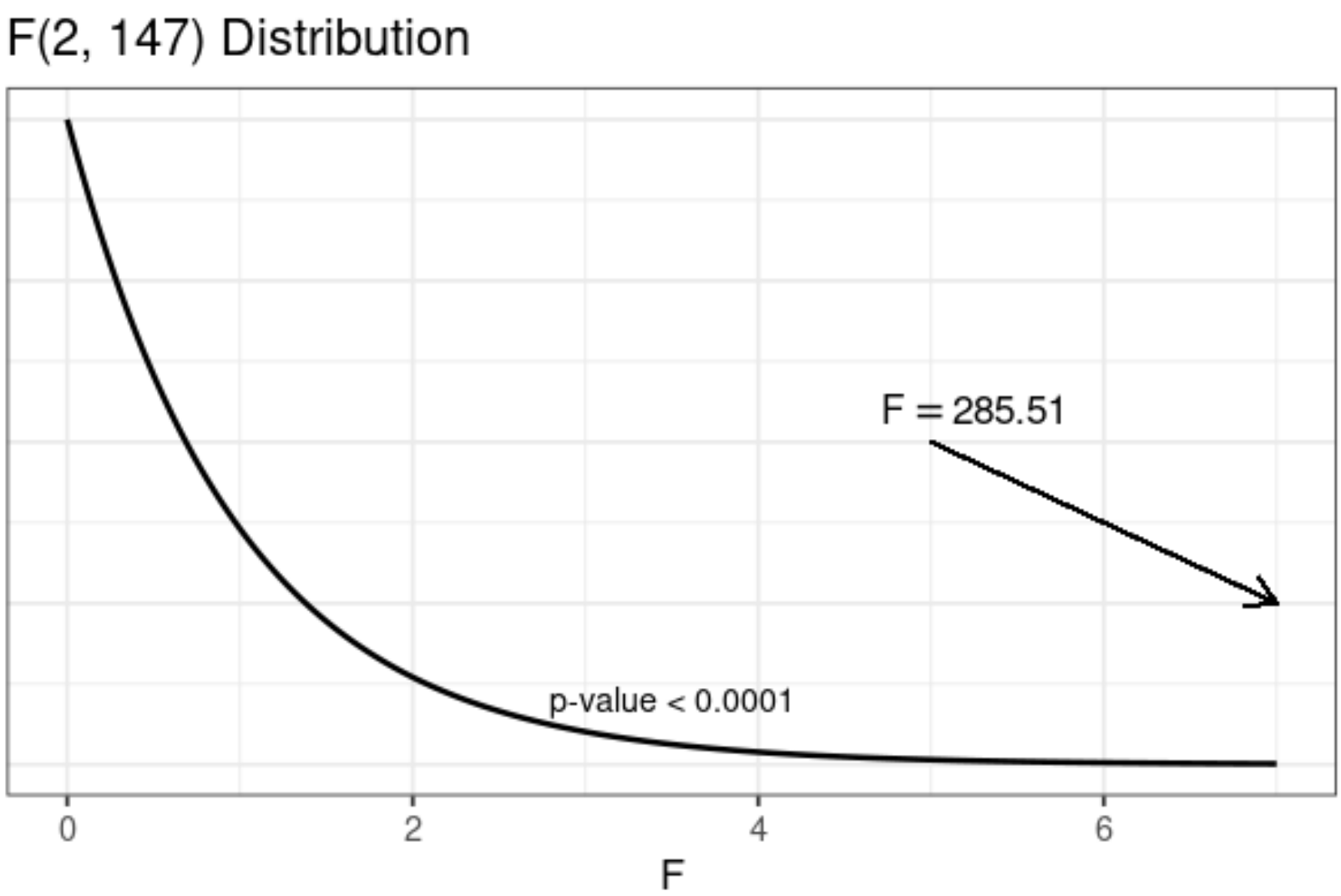


Example 2

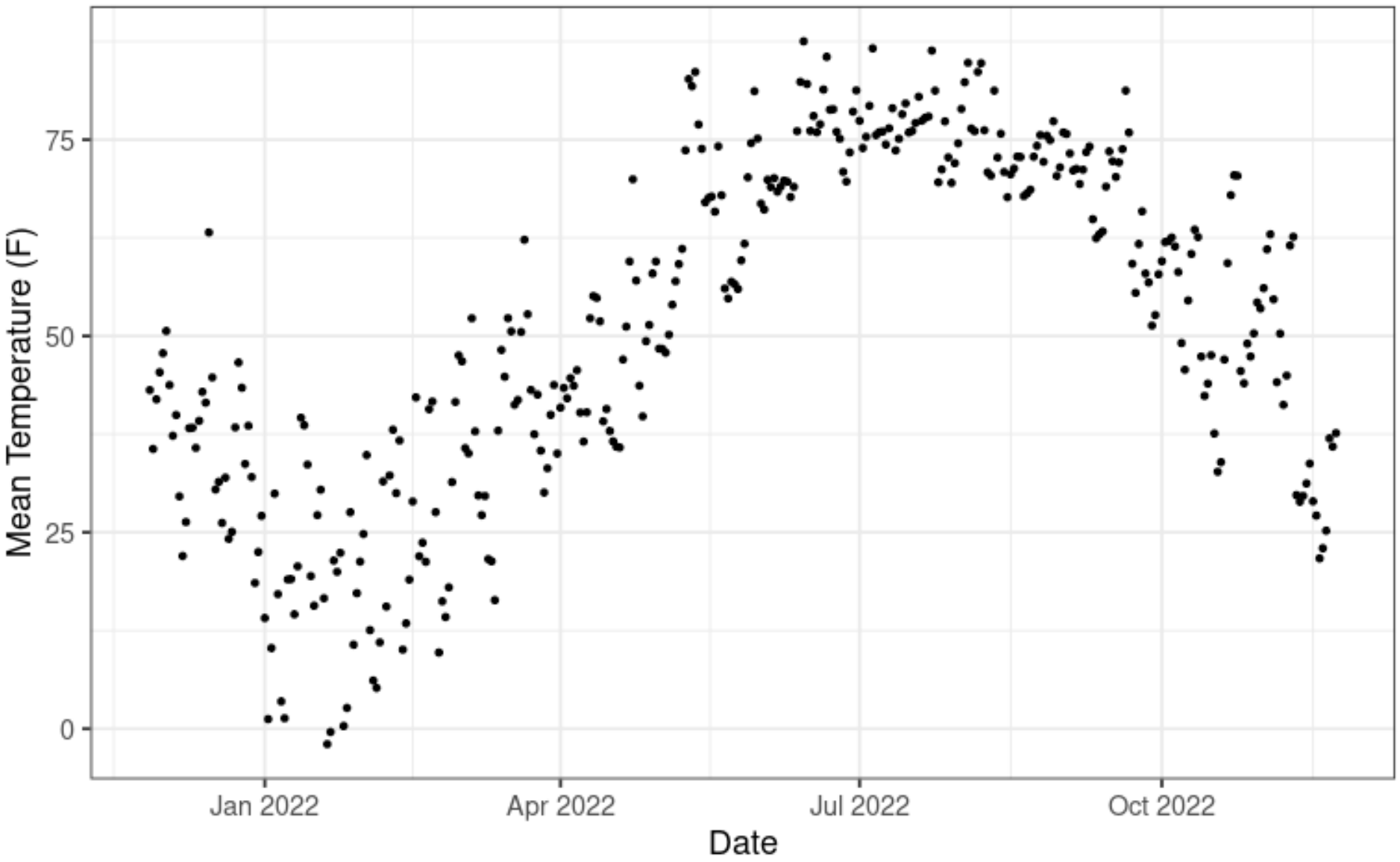




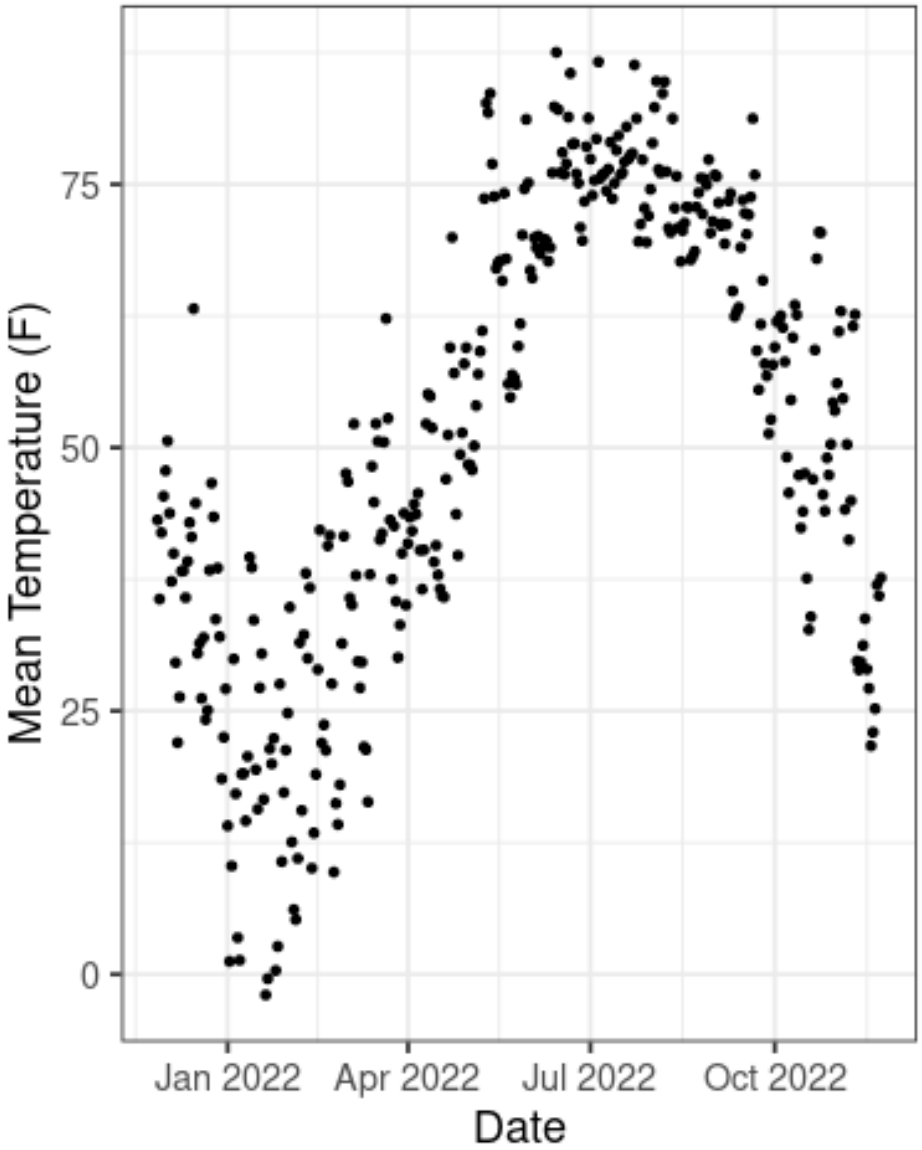
	Df	Sum Sq	Mean Sq	F value	p-value
Group	2	2587.33	1293.67	285.51	<0.00001
Residuals	147	666.07	4.53		

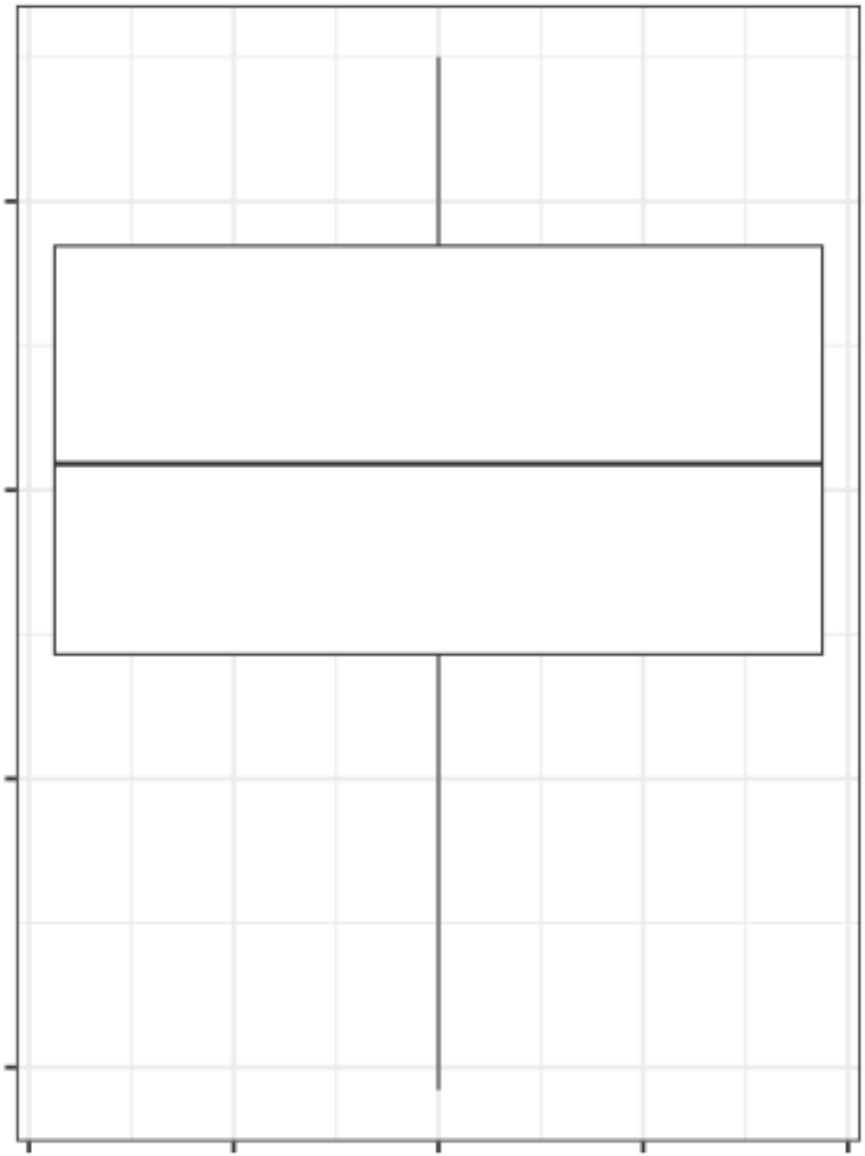

F(2, 147) Distribution
F = 285.51
p-value < 0.0001
0
2
4
6
F

Annual Temperature – Grinnell

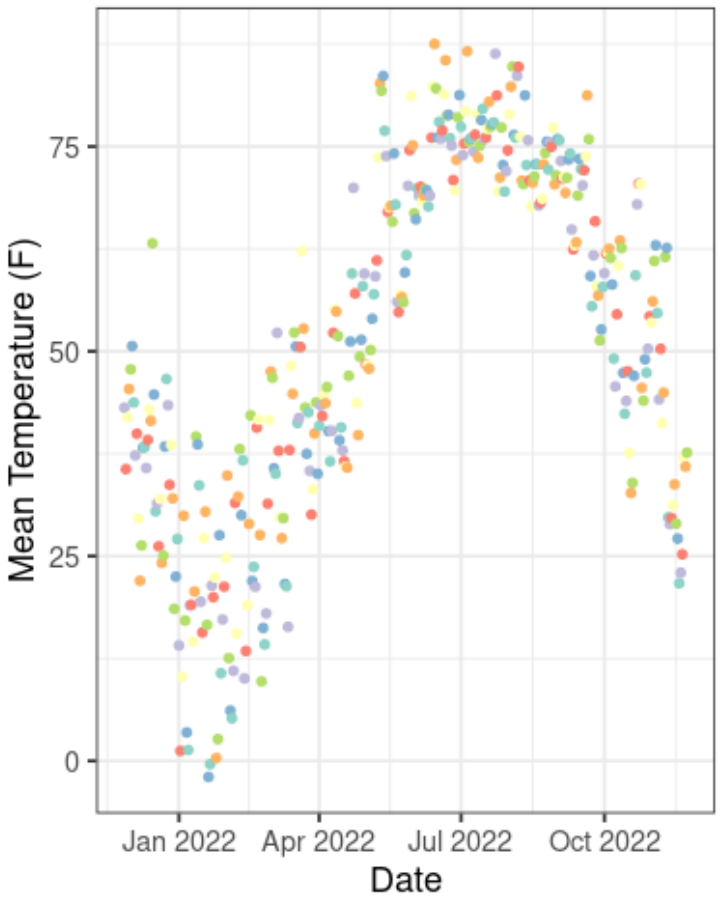


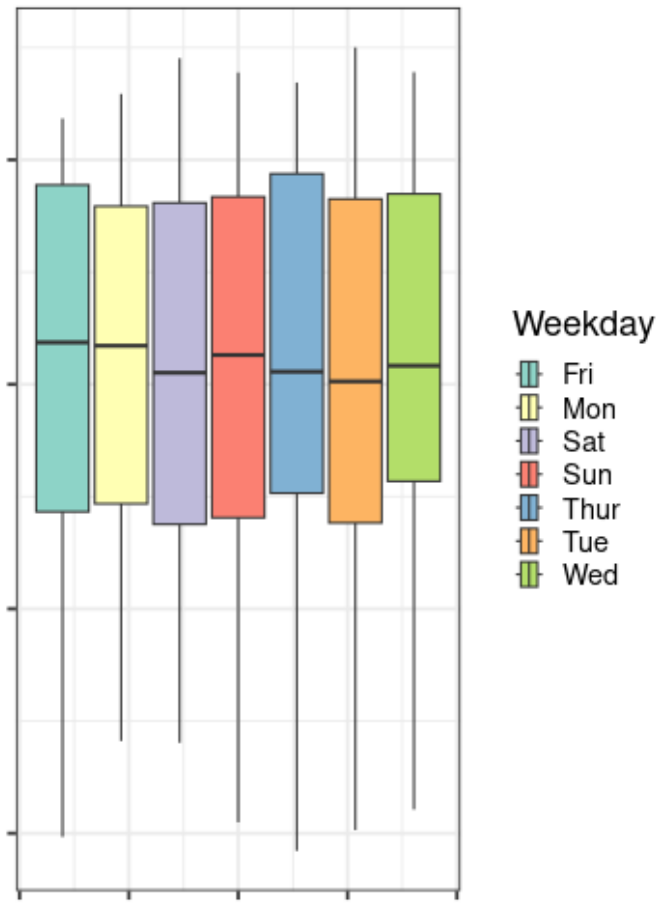
Annual Temperature





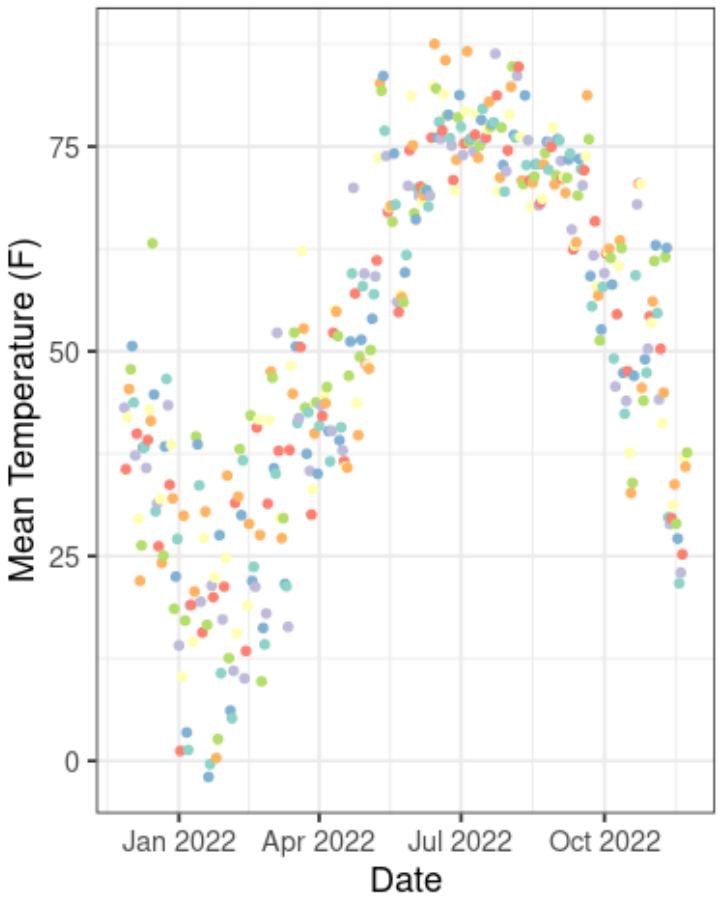
Temperature by Day

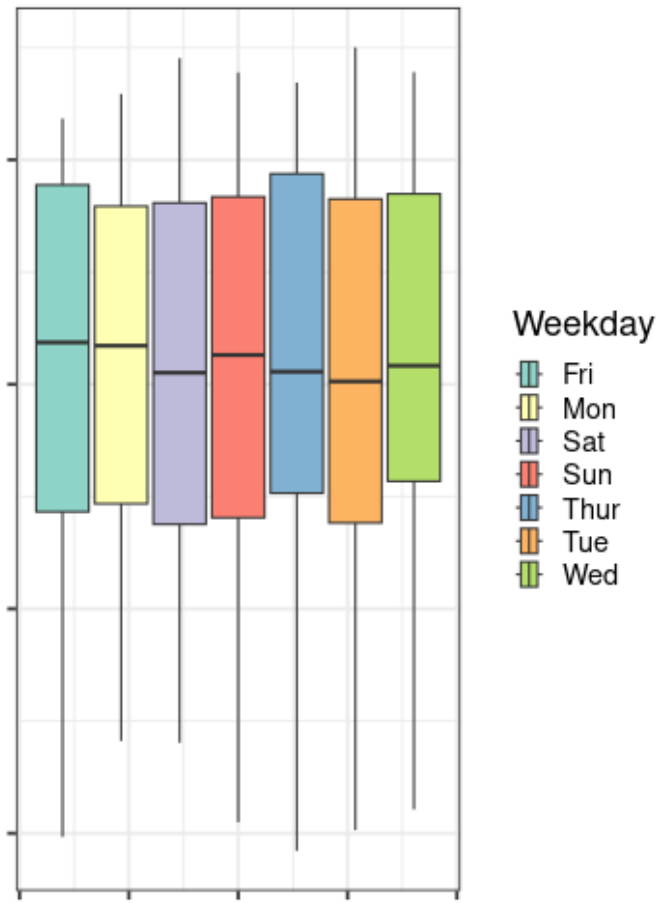




	Df	Sum Sq	Mean Sq	F value	*p*-value
Weekday					
Residuals					

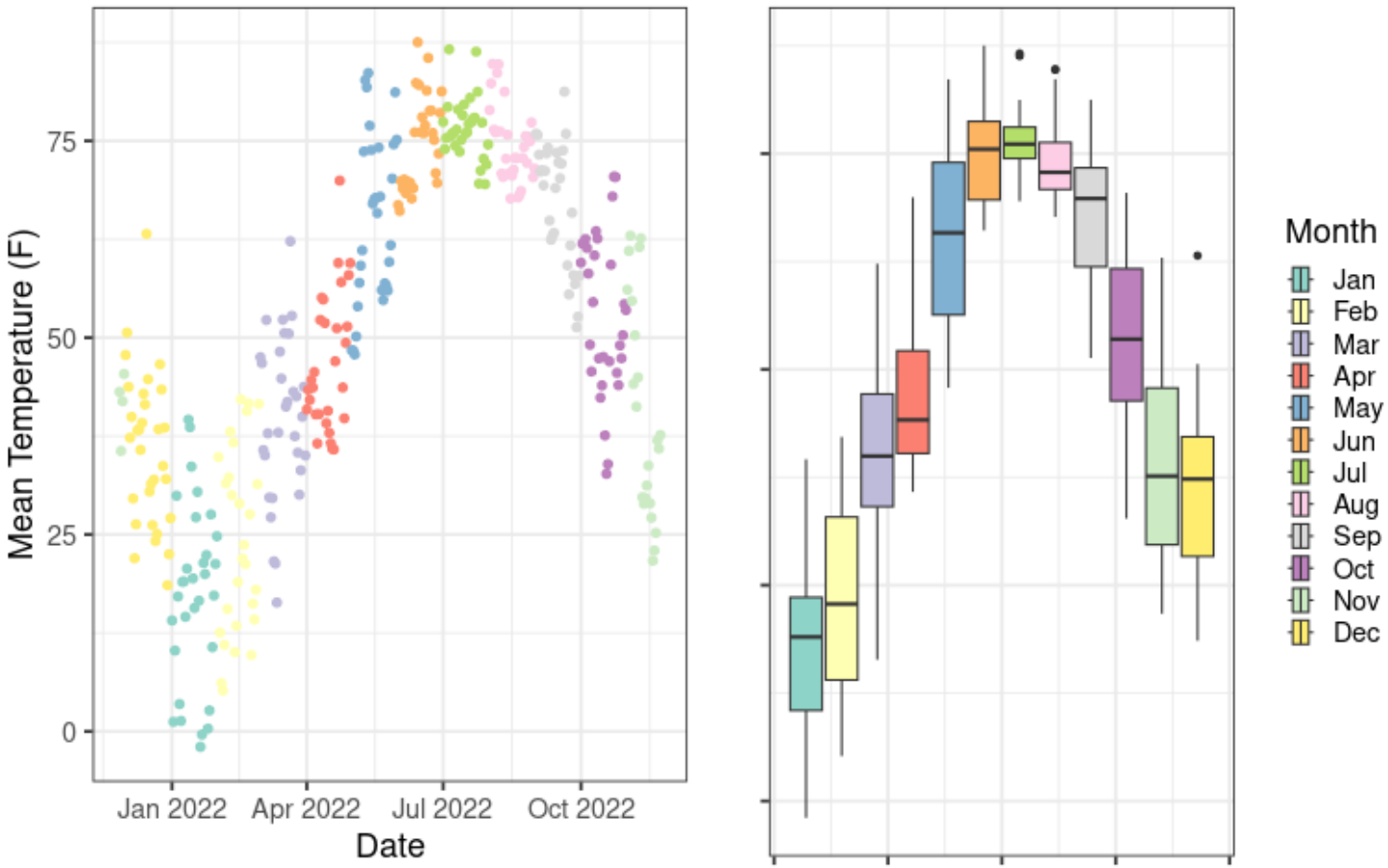
Temperature by Day





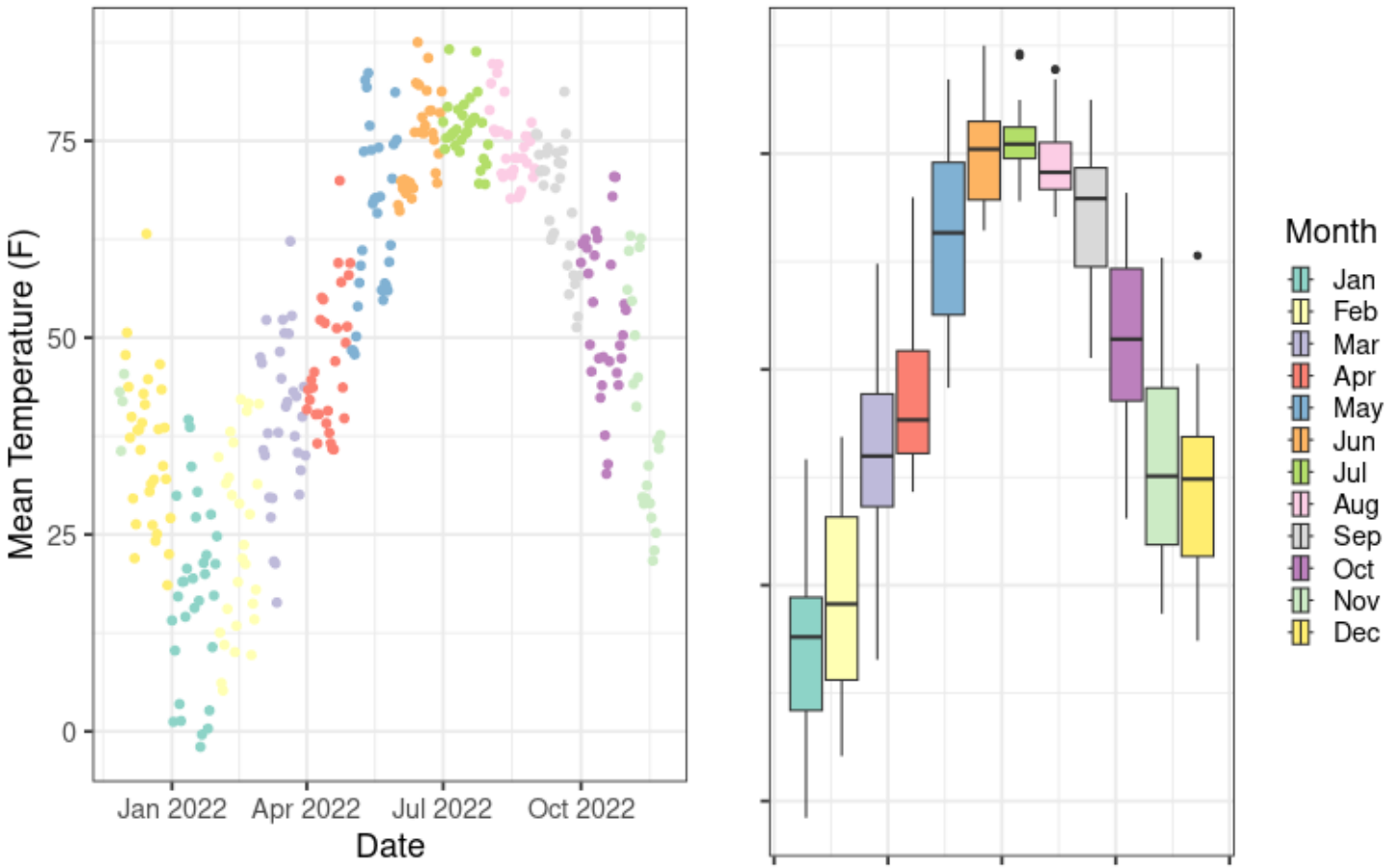
	Df	Sum Sq	Mean Sq	F value	*p*-value
Weekday	6	342.71	57.12	0.12	0.9939
Residuals	355	168524.83	474.72		

Temperature by Month



	Df	Sum Sq	Mean Sq	F value	*p*-value
Weekday					
Residuals					

Temperature by Month



	Df	Sum Sq	Mean Sq	F value	*p*-value
Month	11	138048.06	12549.82	142.52	<0.0001
Residuals	350	30819.48	88.06		

Review

- Total variability made up of within and between group variation
- ANOVA gives us a method for determining the relative sizes of these types of variation
- F statistic can be used as a measure of certainty