Analysis of Variance
(ANOVA)

November 28, 2022

Warming up

- Data
- Inference
- Hypothesis testing
- Distributions
- *p*-values?
- Null, alternative, and errors

ANOVA

The null hypothesis is testing the difference of means between 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?

Questions

Here are some questions we should be able to answer by the end of class

1. What is the relationship between in-group and within-group variability?
2. If we are interested in the difference of means, why consider variability at all?
3. What is the relationship between *p*-values, sample sizes, and variance?
4. Why use ANOVA instead of multiple *t*-tests?

Modeling Variance

One of the main goals in statistical inference is to keep track of uncertainty

In a t-test, for example, it's not enough to simply look at the difference of the sample means – we also want an estimate of how certain we are about those sample means (in other words, how much variability)

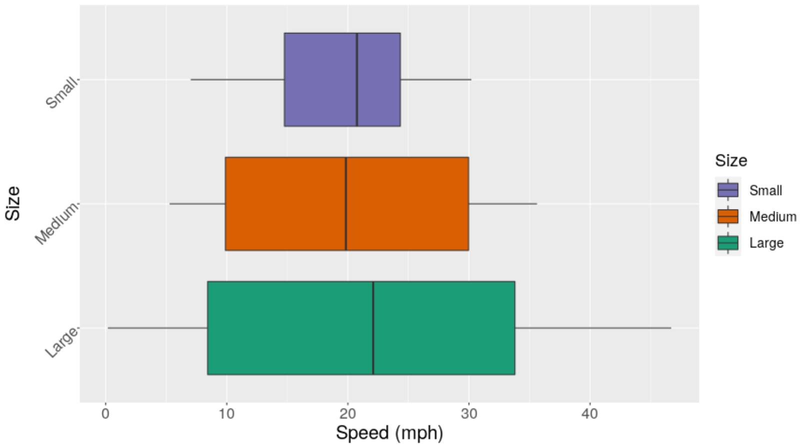
We can use groups to "bin" our observations to try and account for variability, where some groups arrangements better than others

Dog Speed

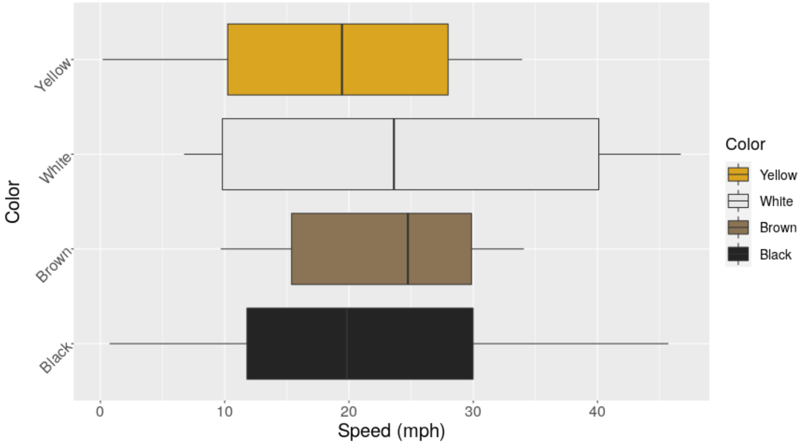
Collected 400 dogs from 8 different breeds, each a sample of 50. Each set has 25 black dogs and 25 dogs of one other color. We also took a collection of land speeds in miles per hour

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

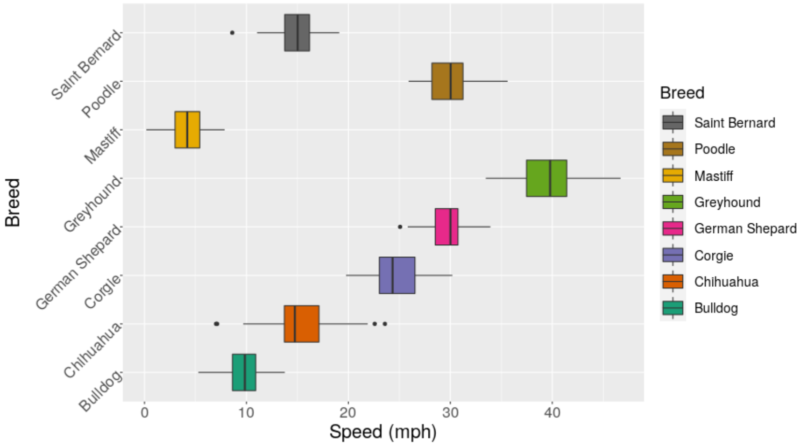
Dog Size


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 and within groups compare when we looked at dogs grouped by size versus by color or by breed?

Variability within a group

This describes how much data within a group differs from its group mean

In anova, we collect the total of all of the within-group variation, which we associated with standard sum of squared errors

$$SSE = \sum (x - \overline{x}_i)^2$$

We can also write it as the weighted sum of group standard deviations

$$SSE = \sum (n_i - 1)s_i^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2$$

In either event, we can find the mean squared error by dividing by the number of observations, n, less the number of groups k

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

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 means, we can use the F distribution to find a p-value for assessing the null hypothesis for ANOVA

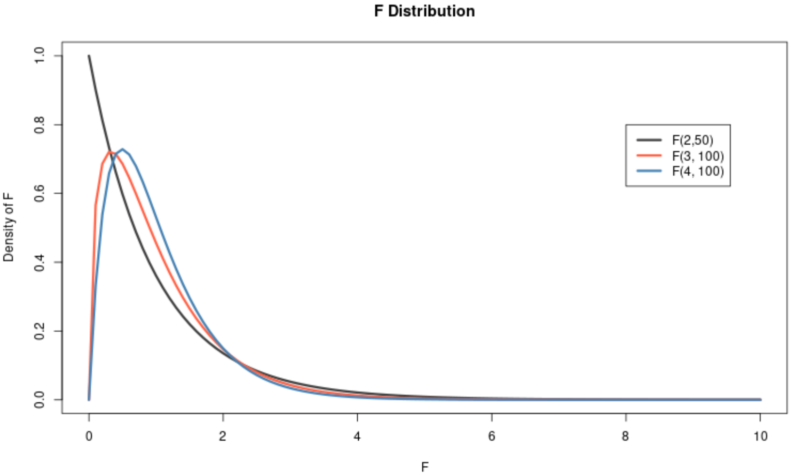
Generally speaking, we are in good shape if:

- 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

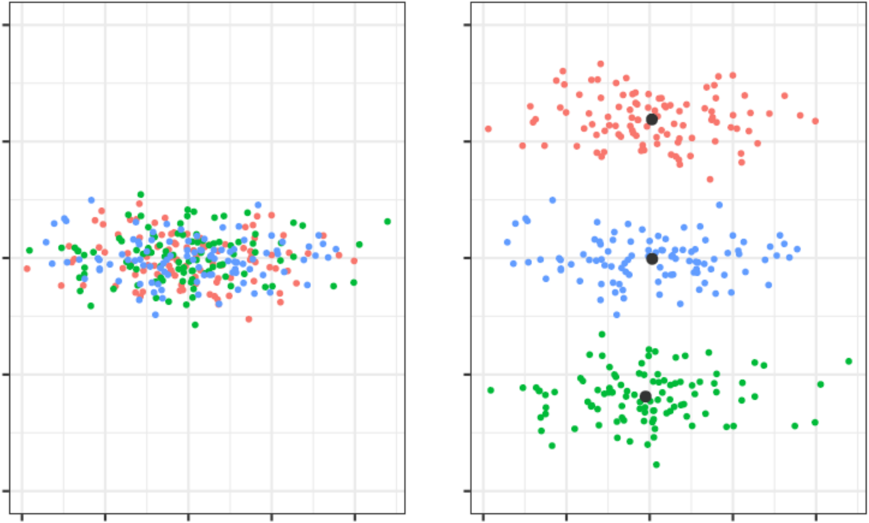
Again similar to the t-distribution, the F distribution is associated with degrees of freedom, in this case two, one for each of the mean squares in the ratio.

F distribution

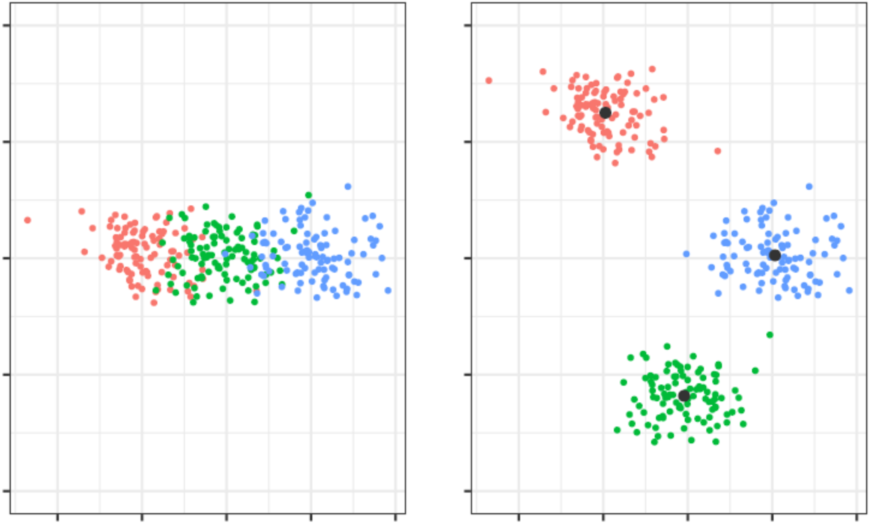
$$F = \frac{MSG}{MSE}$$



No Groups



Yes Groups

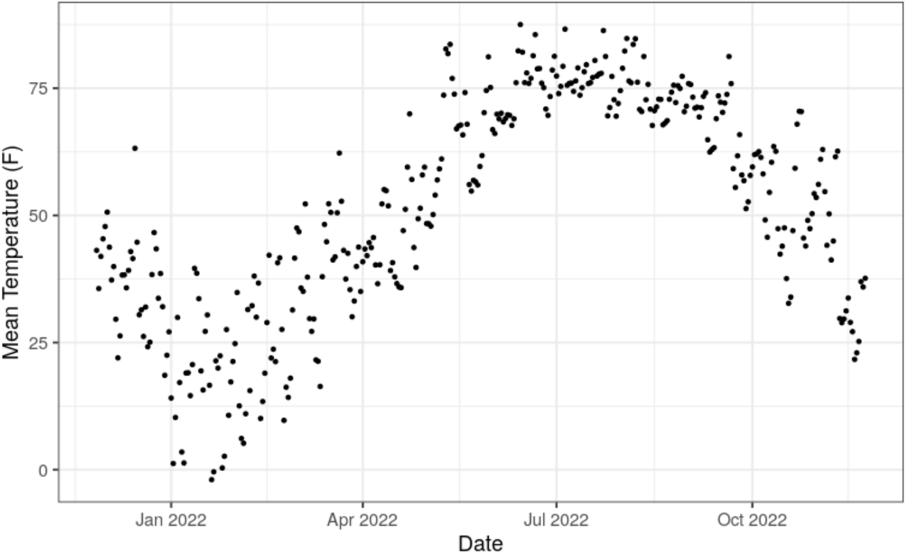


Formulas

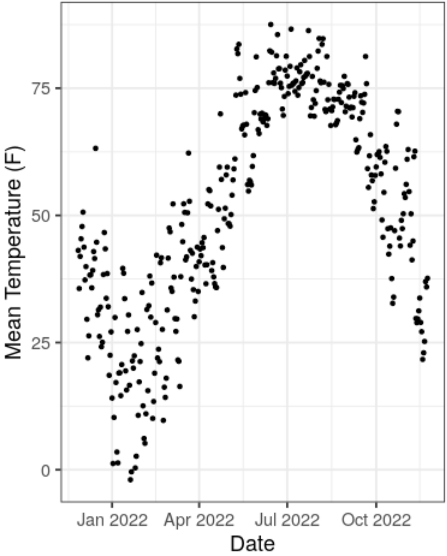
- SSTotal $=$ SSG $+$ SSE
- SSG $=$ sum of squares *between groups*
- SSE $=$ sum of squares *within groups*
- $MSG = \frac{SSG}{k-1}$
- $MSE = \frac{SSE}{n-k}$
- $F = \frac{MSG}{MSE}$

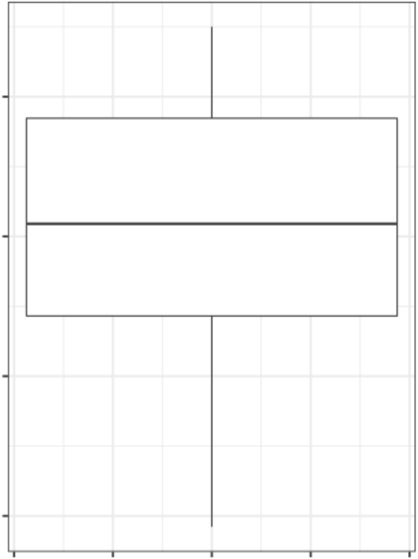
Source	df	Sum Sq	Mean Sq	F value	Pr($>$F) / p-value
Group	k-1	SSG	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSE}$	Upper tail
Error	n-k	SSE	$MSE = \frac{SSE}{n-k}$		
Total	n - 1	SSTotal			

Annual Temperature – Grinnell

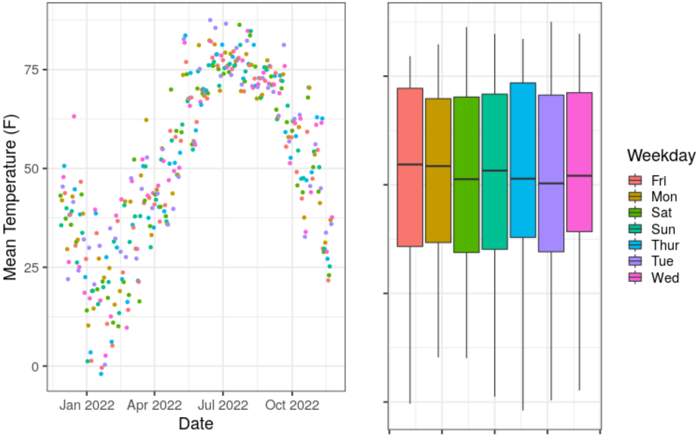


Annual Temperature



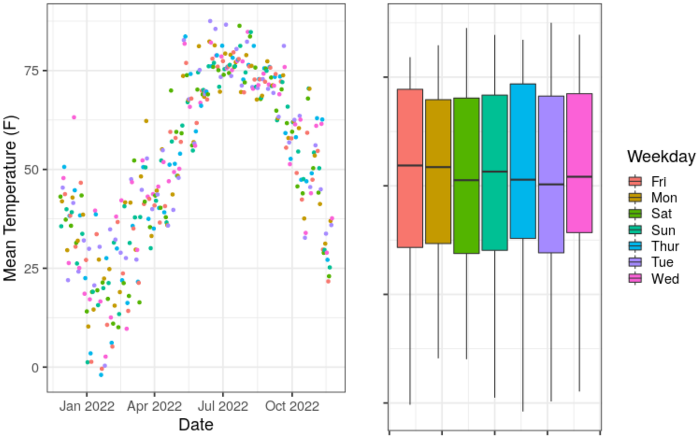


Temperature by Day



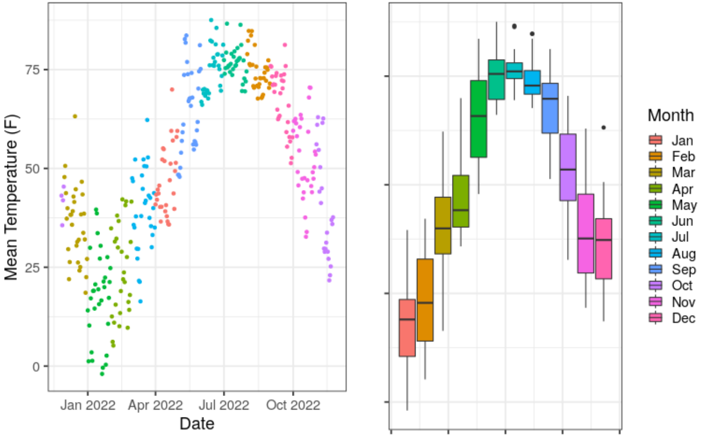
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Weekday					
Residuals					

Temperature by Day



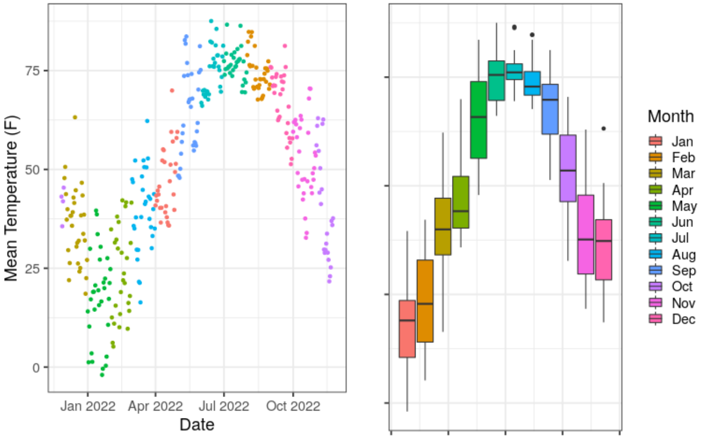
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Weekday	6	342.71	57.12	0.12	0.9939
Residuals	355	168524.83	474.72		

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	0.0000
Residuals	350	30819.48	88.06		

Temperature

Which of the two groupings ended up being better? What did that have to do with the variance within and between the groups?

Can you think of 1-2 other groups we could have used for this data?

Using the two groups presented (weekdays and months) and the 1-2 groups you came up with, order them by increasing F statistic. That is, which would have the largest F statistic? Why?

p-values and *F* statistics

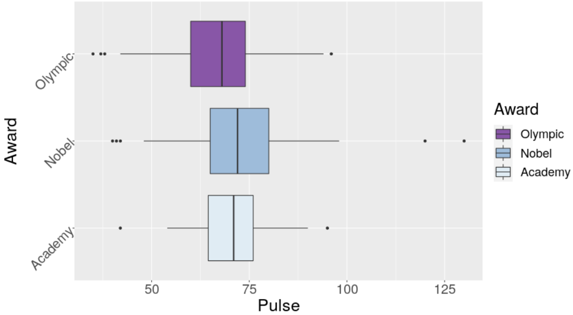
p-values don't exist in isolation – they have context

Reconsider our null hypothesis

What effect does sample size have on:

- confidence in our sample mean
- amount of variability *in the sample*

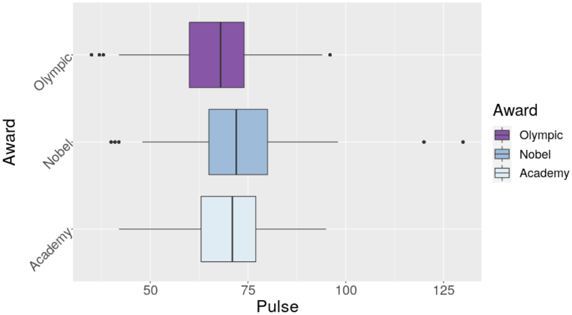
Student Cohort (8.5)



Variable	Award	Count	Mean	StDev
Pulse	Academy	31	70.52	12.36
	Nobel	149	72.21	13.09
	Olympic	182	67.25	10.97
Total		362	69.57	12.21

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Award	2	2047.24	1023.62	7.10	0.0009
Residuals	359	51729.24	144.09		

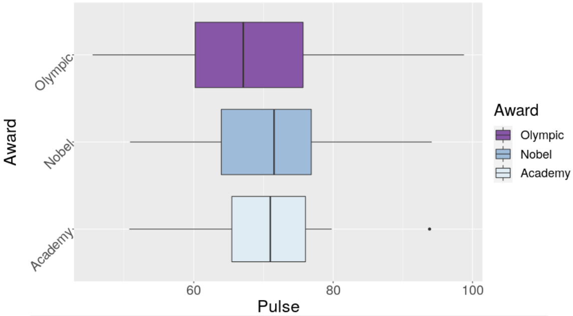
Student Cohort × 10



Variable	Award	Count	Mean	StDev
Pulse	Academy	310	70.52	12.36
	Nobel	1490	72.21	13.09
	Olympic	1820	67.25	10.97
Total		3620	69.57	12.21

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Award	2	20472.43	10236.22	71.57	< 0.0000000000000002
Residuals	3617	517292.43	143.02		

Student Cohort, a third the size



Variable	Award	Count	Mean	StDev
Pulse	Academy	10	70.52	12.48
	Nobel	50	71.61	10.44
	Olympic	60	67.86	10.94
Total		120	69.64	10.92

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Award	2	393.15	196.57	1.67	0.1935
Residuals	117	13807.10	118.01		

To t or not to t

ANOVA is fine and good, but why do we need so many test?

Is not $H_0 : \mu_A = \mu_B = \mu_C$ the same thing as $H_0 : \mu_A = \mu_B$, $H_0 : \mu_A = \mu_C$, and $H_0 : \mu_C = \mu_B$?

And in fact, if we do ANOVA with only two groups, we get the exact same p-value as if we had just done a t-test

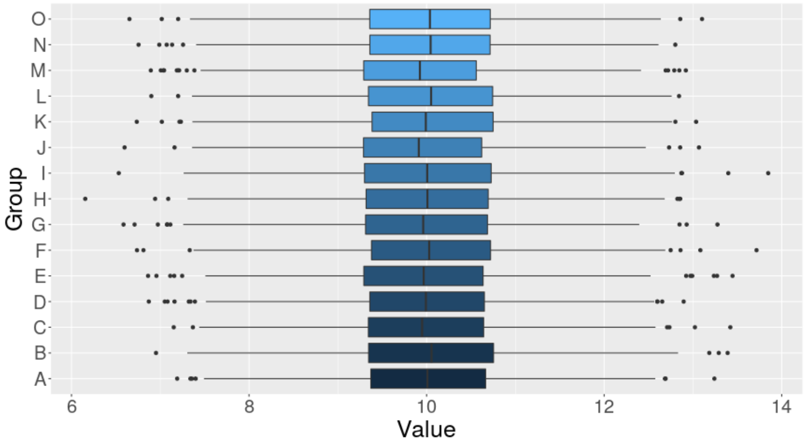
Well, but wait

Testing $\mu_A = \mu_B = \mu_C$ is really only one test, while $\mu_A = \mu_B$, $\mu_A = \mu_C$ and $\mu_B = \mu_C$ is three

What if instead of three groups I had five? Or ten? What effect does this have on the number of t-tests I am doing?

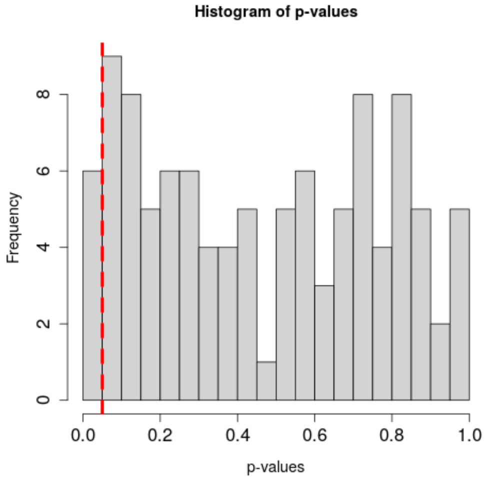
Especially if I am concerned with my Type I error rate at $\alpha = 0.05$.....

t-test mania!



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Groups	14	15.40	1.10	1.10	0.3504
Residuals	14985	14964.85	1.00		

p-values



Histogram of p-values

Of the 105 tests run, 6 were false positives, which is 5.7% as expected

Obligatory XKCD

Jelly Beans

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
- We should always take a step back and consider the questions we are trying to answer

Questions from the beginning

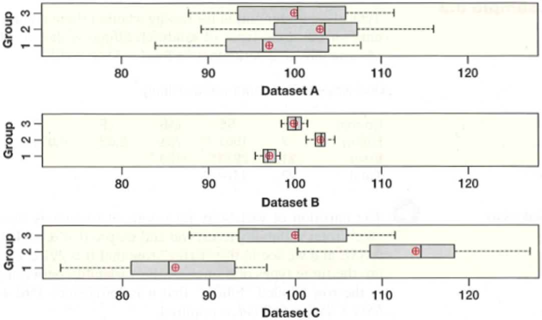
1. What is the relationship between in-group and within-group variability?
2. If we are interested in the difference of means, why consider variability at all?
3. What is the relationship between *p*-values, sample sizes, and variance?
4. Why use ANOVA instead of multiple *t*-tests?

60 second survey

- What is one takeaway from this class today?
- What is one thing that is still not very clear?

Figure 8.2 (bonus if time)

- How would you describe these datasets? Are the means different?
- Do A and B both have three groups in the same sense of the word?
- What about B and C? Which of these sets of groups would you say is more different?


Group
Dataset A
Dataset B
Dataset C