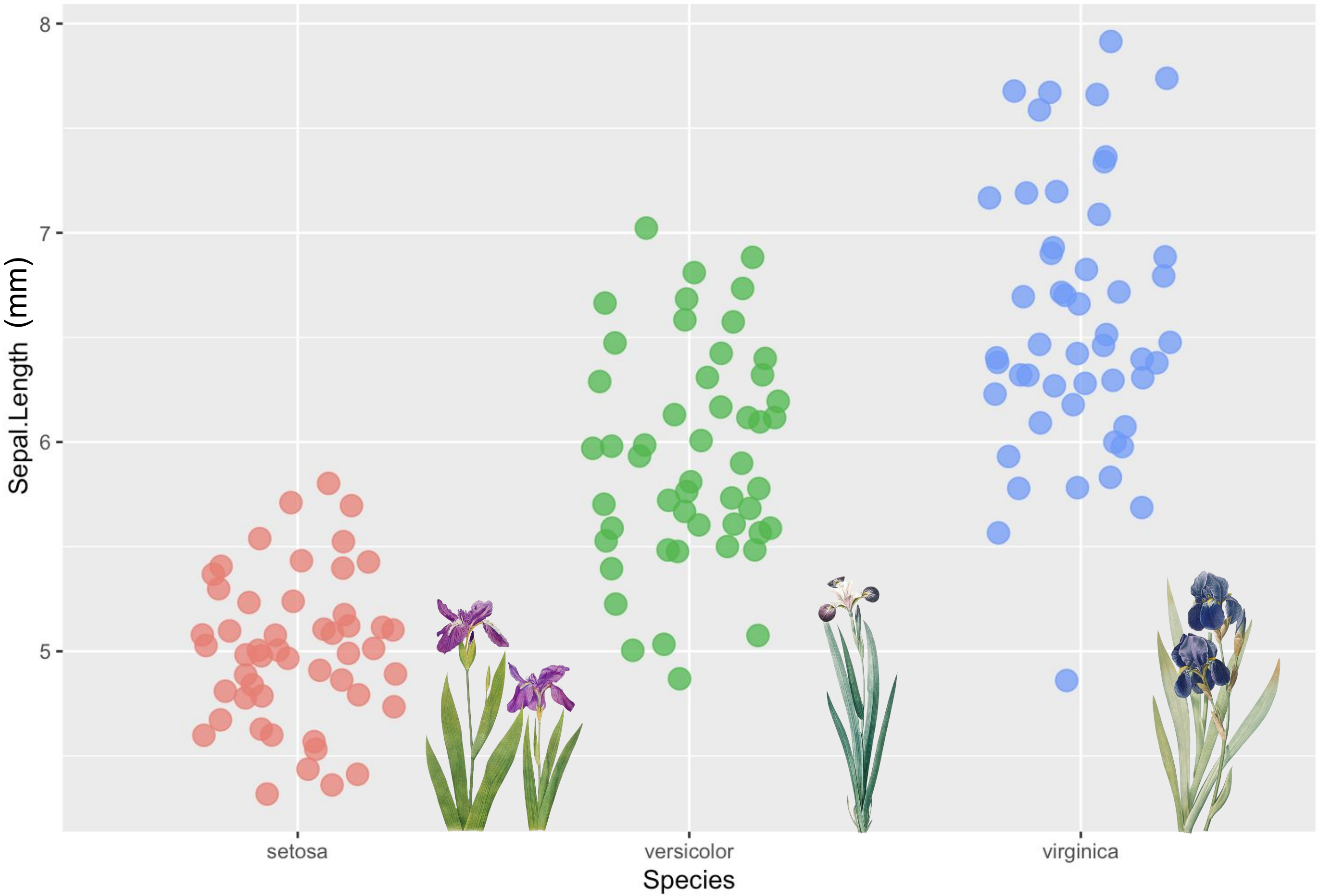
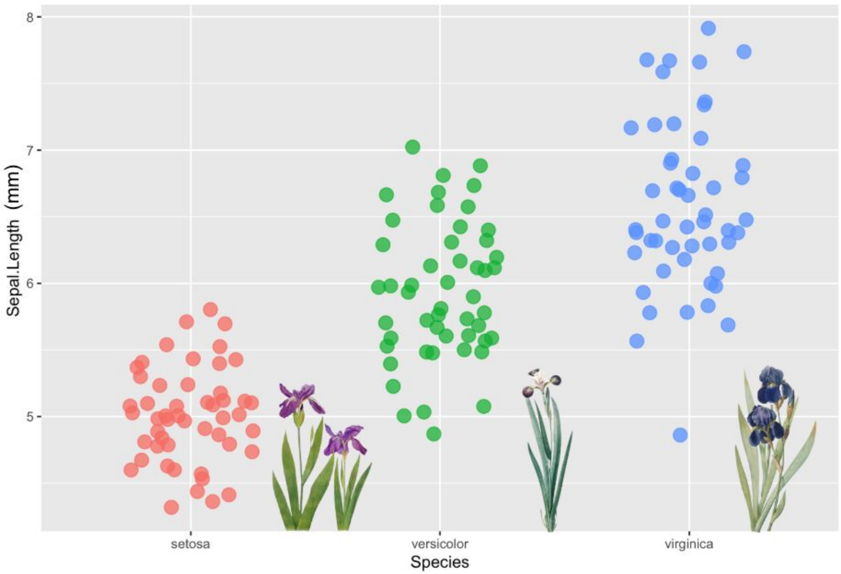
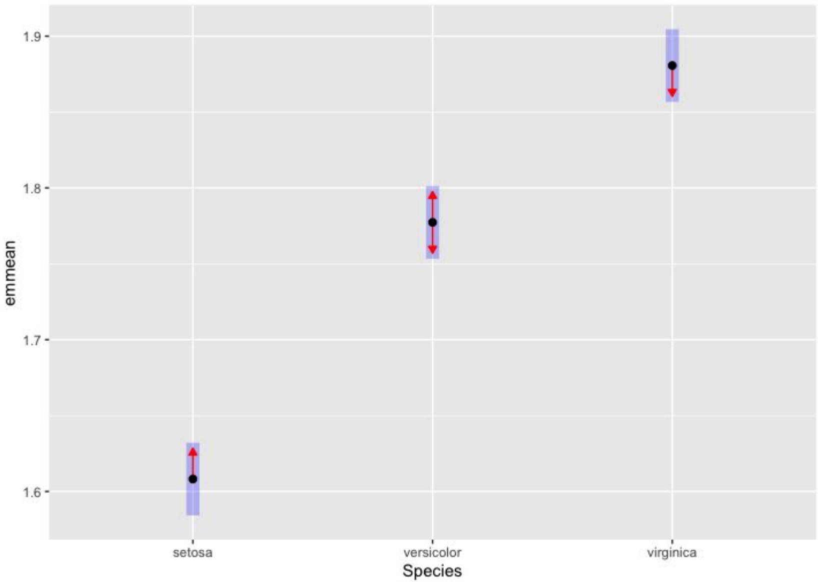
Are the sepal lengths of these three Iris species different?



Informal investigation vs. statistics





Informal investigation: Rely on intuition and observation

Statistics: Collect, analyze, interpret data

Week 10: Statistical analysis in R

Today's Agenda:

1. Introduction to stats in R and example
2. We do an example together
3. If time: You analyze your response data

Due Next:

- Read Chapters 22 & 23
- Research Days Extra Credit Opportunity (Mon 4/7)
- Early Plant Development, Cells & Tissues Lab Quiz (Friday 4/4)

Lecture
Lab
Greenhouse

UNM Biology Research Days Extra Credit

- This Friday, 4/4!!!
- Optional, but fun to see research here at UNM
- Extra credit - 2 pts (due next Mon 4/7)
- Morning talks (session 1) and afternoon poster sessions
- Link to schedule

UNM Biology Research Days (Optional Extra Credit)

2 Points Possible

Due: Mon Apr 7, 2025 11:59pm

Attempt 1

In Progress
NEXT UP: Submit Assignment

Add Comment

Unlimited Attempts Allowed

Details

Instructions:

** This is an optional assignment worth 2 points of extra credit on a future assignment **

On **April 4, 2025**, UNM Biology is hosting Research Days. Please find 3+ posters (or talks) that pique your interest. Here is a link to the schedule. Take notes on the **research question**, the **hypothesis/hypotheses**, the **results**, and **future directions** (if they don't mention one or more of these, ask them!). Take a picture of each poster for reference.

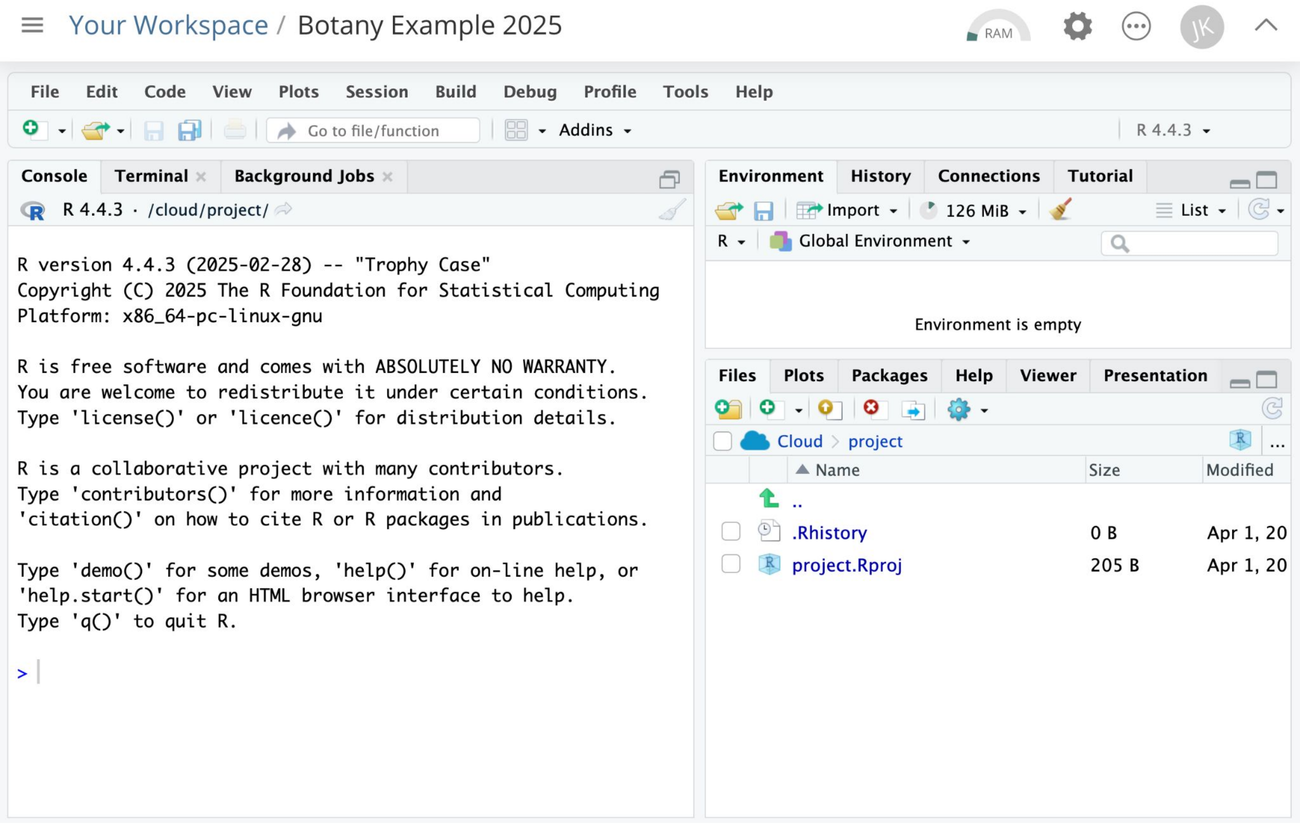
In the text box below, answer the following questions:

1. What 3 posters or talks (title and name of author) did you visit and what interested you about them?
2. Which poster/talk was your favorite and why? (Optionally, include a picture of that poster/talk)
3. For your favorite poster/talk, what were the...
   1. research questions?
   2. hypotheses?
   3. results?
   4. future directions?
4. What is one question you asked (or would ask) the author about their work?

Edit View Insert Format Tools Table

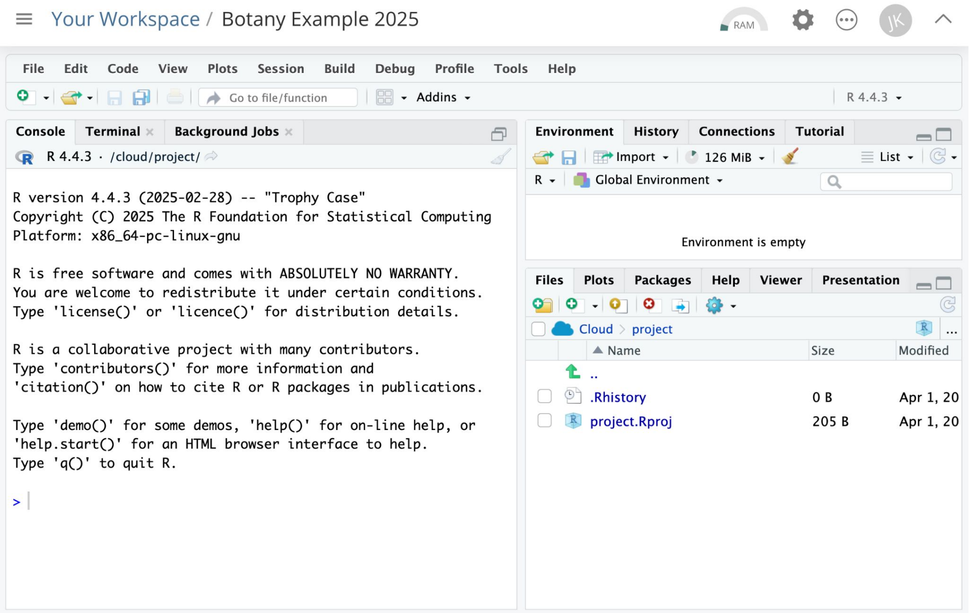
12pt Paragraph

R Studio Cloud



Parts to analyzing our data

1. Cleaning and loading our data
2. Informal investigation
3. Hypothesis testing
4. Interpreting results
5. Testing for differences among groups
6. Graphing our results



Part 1: Cleaning and loading data

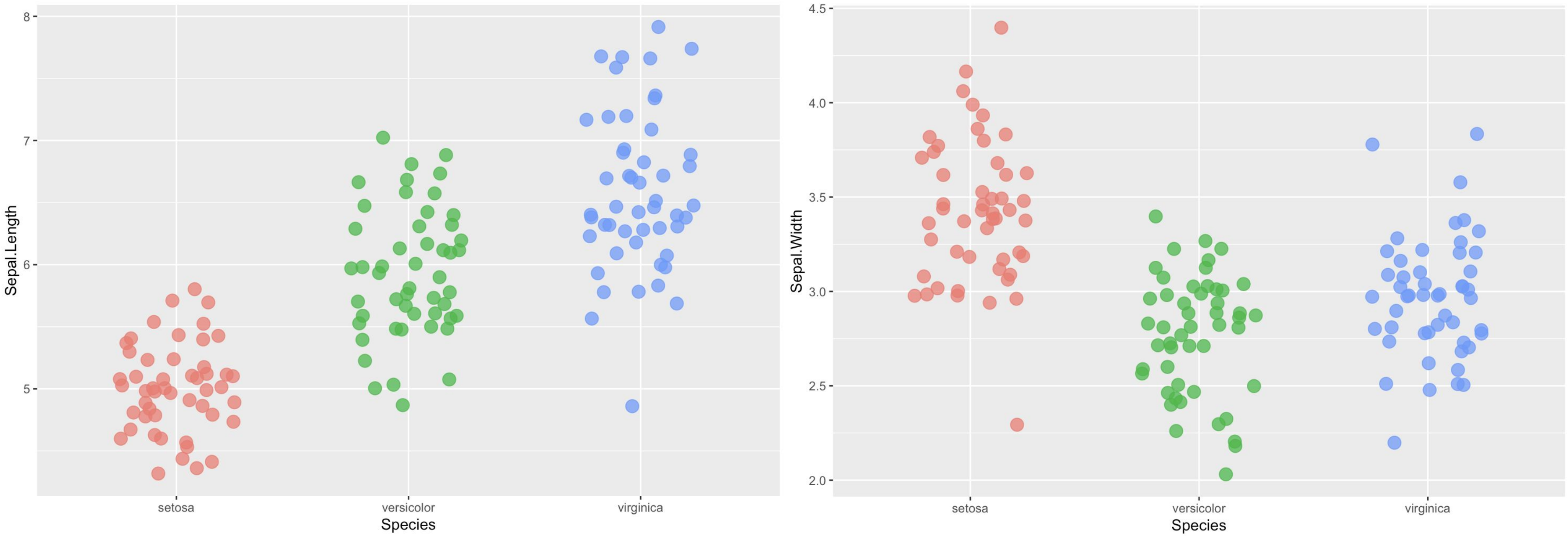
- Column headers (no spaces)
- Quantitative variables are numbers, categorical variables are categories
- No missing data, but missing data is OK (just leave cells blank)

```
mydata <- read.csv("iris.csv")
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3	1.4	0.1	setosa
4.3	3	1.1	0.1	setosa
5.8	4	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa
5.4	3.4	1.7	0.2	setosa
5.1	3.7	1.5	0.4	setosa
4.6	3.6	1	0.2	setosa

Part 2: Informal investigation

```
ggplot(mydata) +
  aes(x = Species, y = Sepal.Length) +
  geom_jitter(aes(colour = Species), width = 0.25, size = 4, alpha = 0.75) +
  theme(legend.position = "none")
```



Part 3: Hypothesis testing

Our question: Does (simulated) herbivory affect plant growth and/or reproduction?

Our hypotheses (for each response):

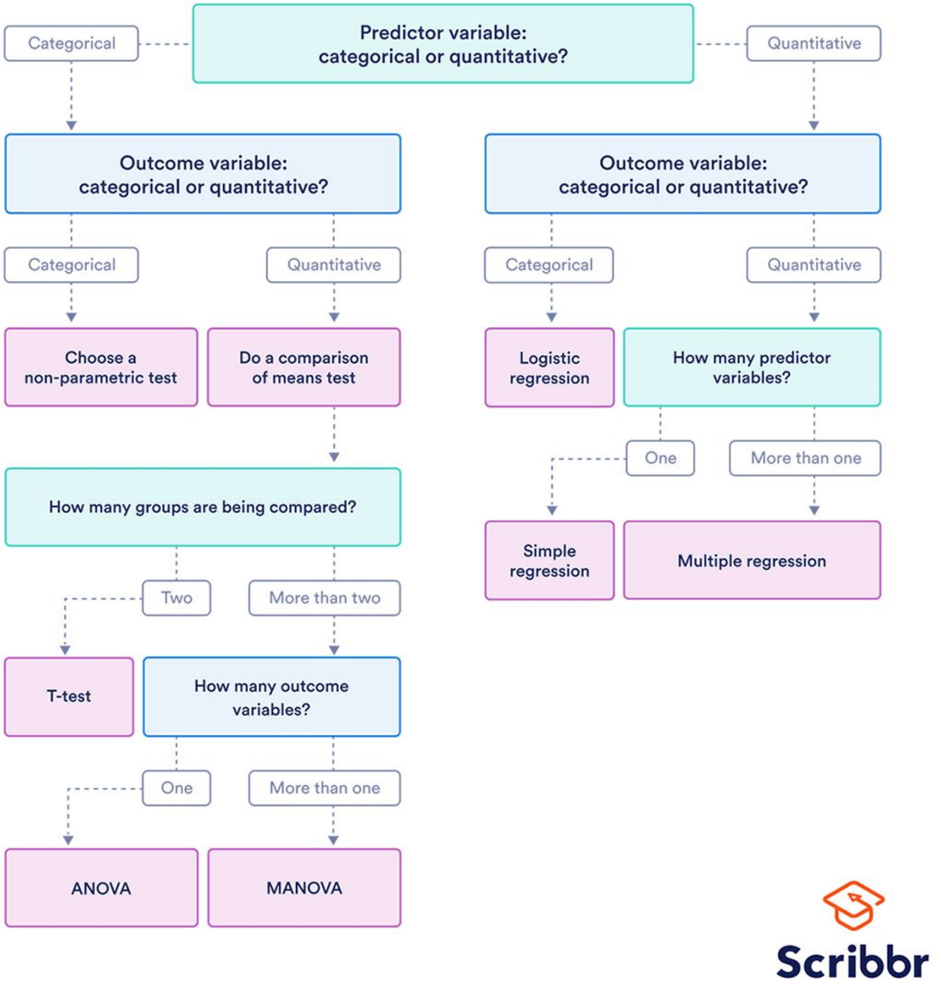
H_0 (Null): Response means for each treatment group are not significantly different from each other.

H_A(Alt): The response mean for at least one treatment group is significantly different!

How do we choose a statistical test to use?

We have to look at the predictor and response variables:

- Are they categorical or quantitative?
- How many?
- Think about our question!

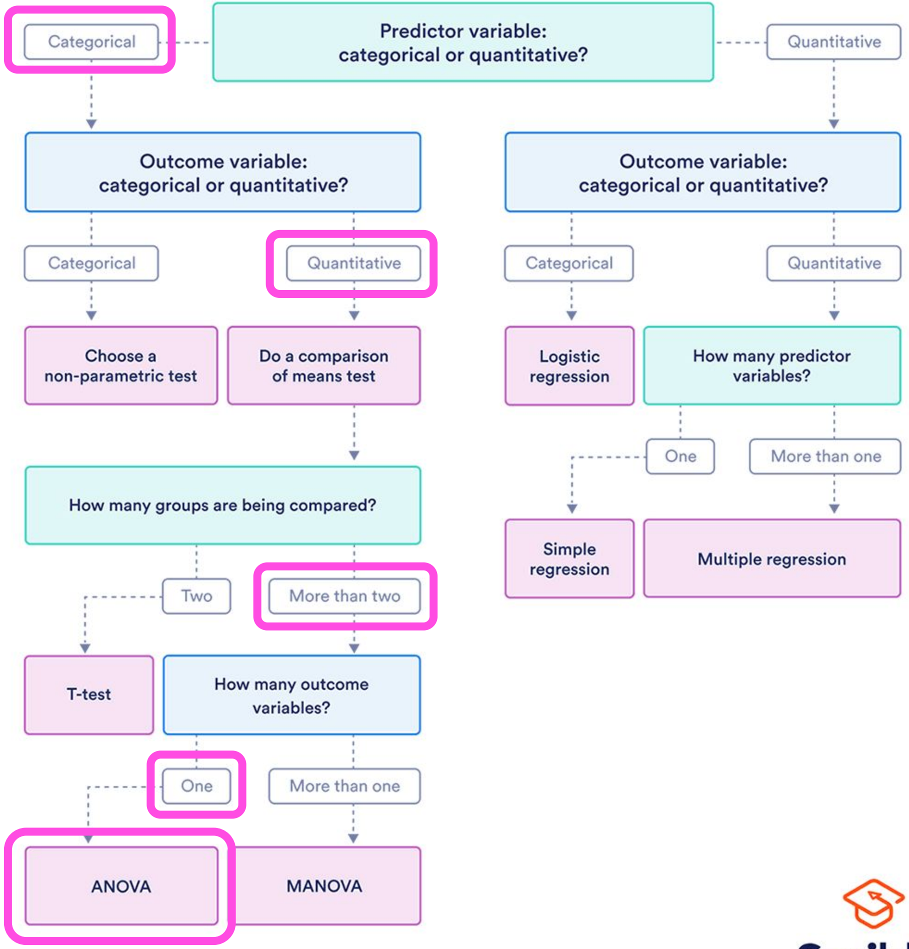


Independent (Predictor) Variable

- Treatment type
- Categorical variable

Dependent (Response) Variable

- Plant height, biomass, your choice
- Quantitative variables

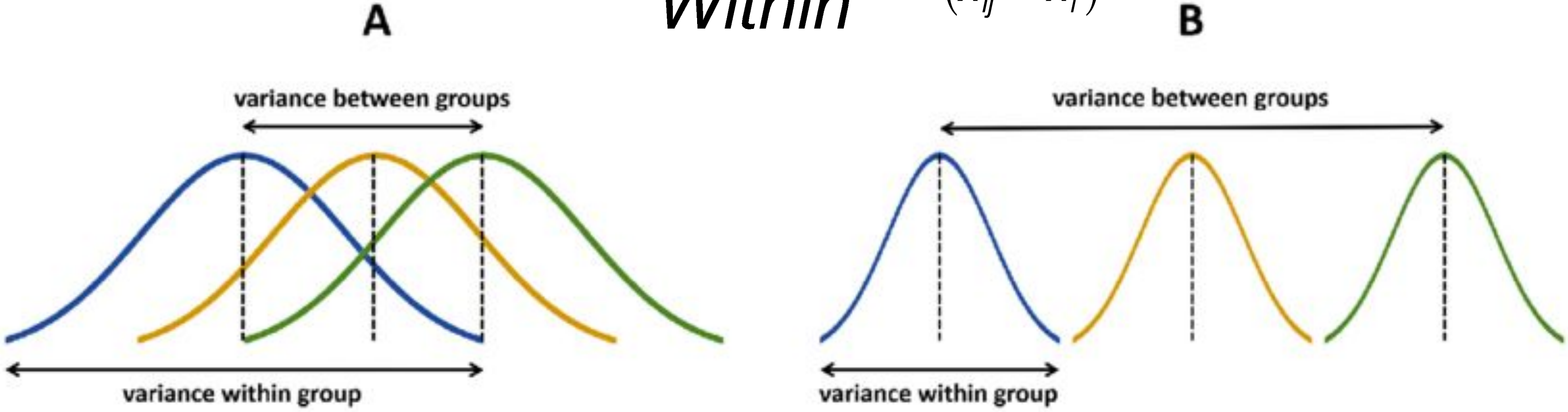


Scribbr

ANOVA compares variances within & between groups

(Analysis of Variance)

$$F = \frac{Between}{Within} \quad \frac{(\bar{x}_i - \bar{x})^2}{(x_{ij} - \bar{x}_i)^2}$$



A small *F* is evidence that we must accept $\mathbf{H_0}$: equal means

A larger *F* is evidence against $\mathbf{H_0}$: means may be different

Doing an ANOVA (significance a = 0.05)

In the "old days" you would do this by hand...

Fortunately, we can do ANOVA in R pretty easily

F-table of Critical Values of α = 0.05 for F(df1, df2)

	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07

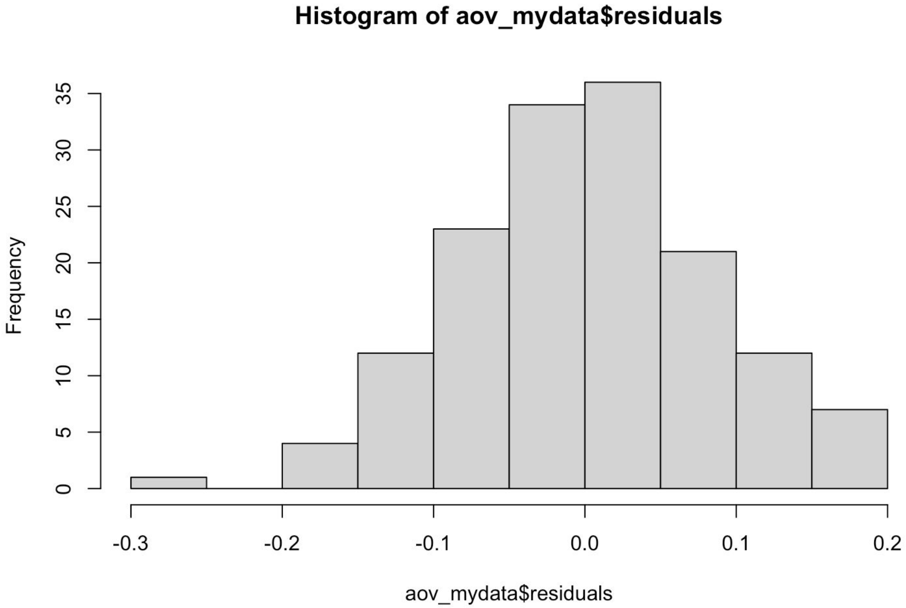
```
aov_mydata <- aov(Sepal.Length ~ Species, data = mydata)
```

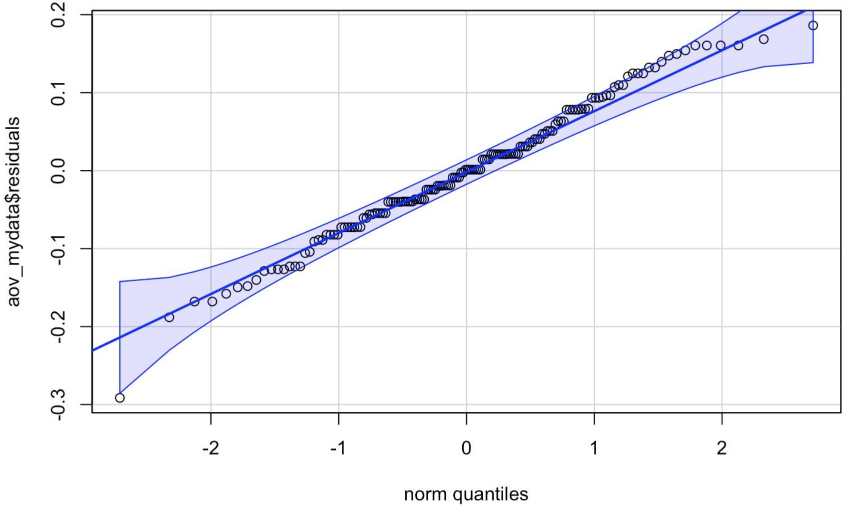
Assumptions of ANOVA

1. **Data points are all independent**
   - Verify during data collection
2. **Data (residuals) are normally distributed**
   - Plot (informal) and Shapiro-Wilk test (formal)
3. **Variances are equal among treatment groups**
   - Plot (informal) and Bartlett's Test (formal)

Assumption 2: Data are normally distributed

Plot (informal)





Assumption 2: Data are normally distributed

Shapiro-Wilk Test (formal)

```
shapiro.test(aov_mydata$residuals)
```

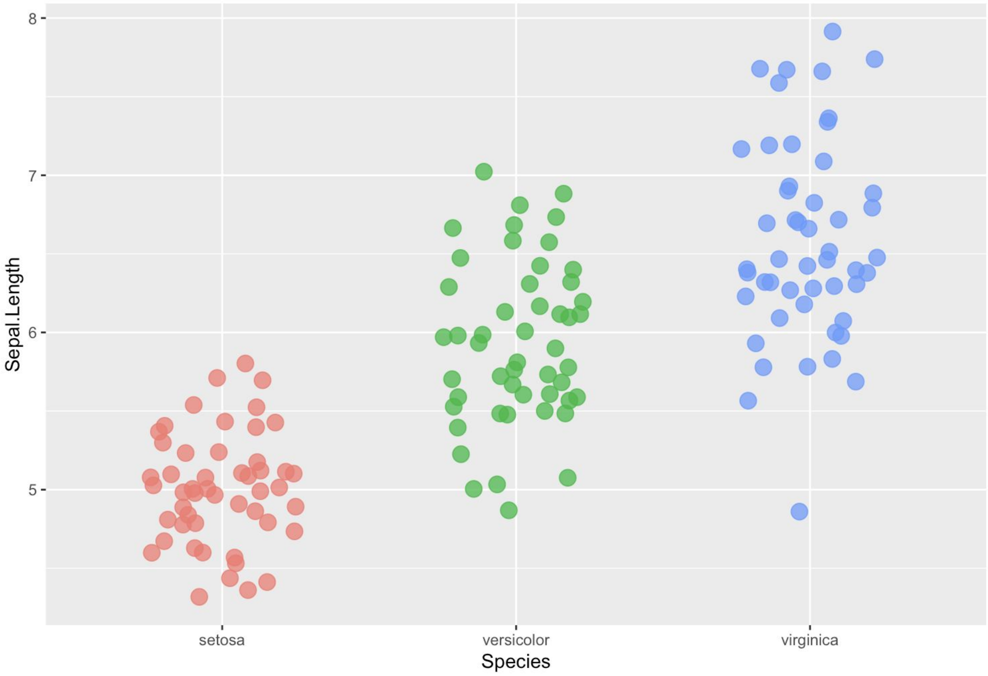
```
	Shapiro-Wilk normality test


data:  aov_mydata$residuals
W = 0.9883, p-value = 0.2424
```

If p-value < 0.05, your data is not normal!

Assumption 3: Equal variances among groups

Plot (informal)



Assumption 3: Equal variances among groups

Bartlett's Test (formal)

```
bartlett.test(Sepal.Length ~ Species, data = mydata)
```

```
	Bartlett test of homogeneity of variances


data:  Sepal.Length by Species
Bartlett's K-squared = 16.006, df = 2, p-value = 0.0003345
```

If p-value < 0.05, your variances are not equal!

What to do if assumptions are not met?!

1. **Identify and remove outliers** (one at a time!)
2. **Transform our data** (log, square root, inverse)

   Re-run your ANOVA and check assumptions

3. If you can't get your data to meet the assumptions, **use a non-parametric test** (Kruskal-Wallis Test)

If you do any of this, make a note of it somewhere

Part 4: Interpreting results

Now, we can look at the results of our ANOVA:

```
summary(aov_mydata)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)    
Species       2  2.724  1.3620   124.5 <2e-16 ***
Residuals   147  1.608  0.0109                   
```

If p-value < 0.05, we reject H_0!

Save the data using a descriptive name:

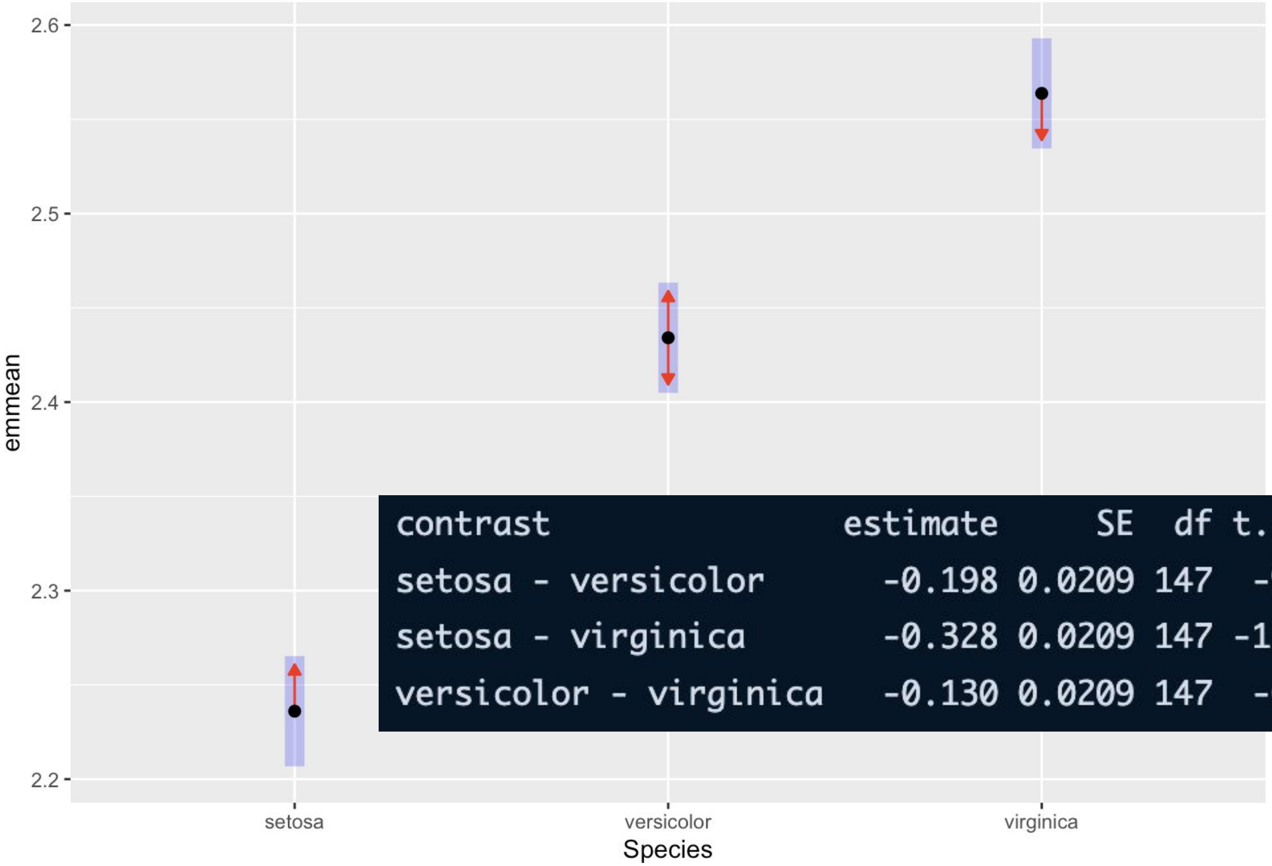
```
capture.output(summary(aov_mydata), file =
“Sepal.Length_anova.txt”)
```

Part 5: Test differences among groups

- ANOVA can tell us if we should reject the null hypothesis, but won't tell us which groups differ :(
- Use the Tukey Honestly Significant Difference (HSD) test :)

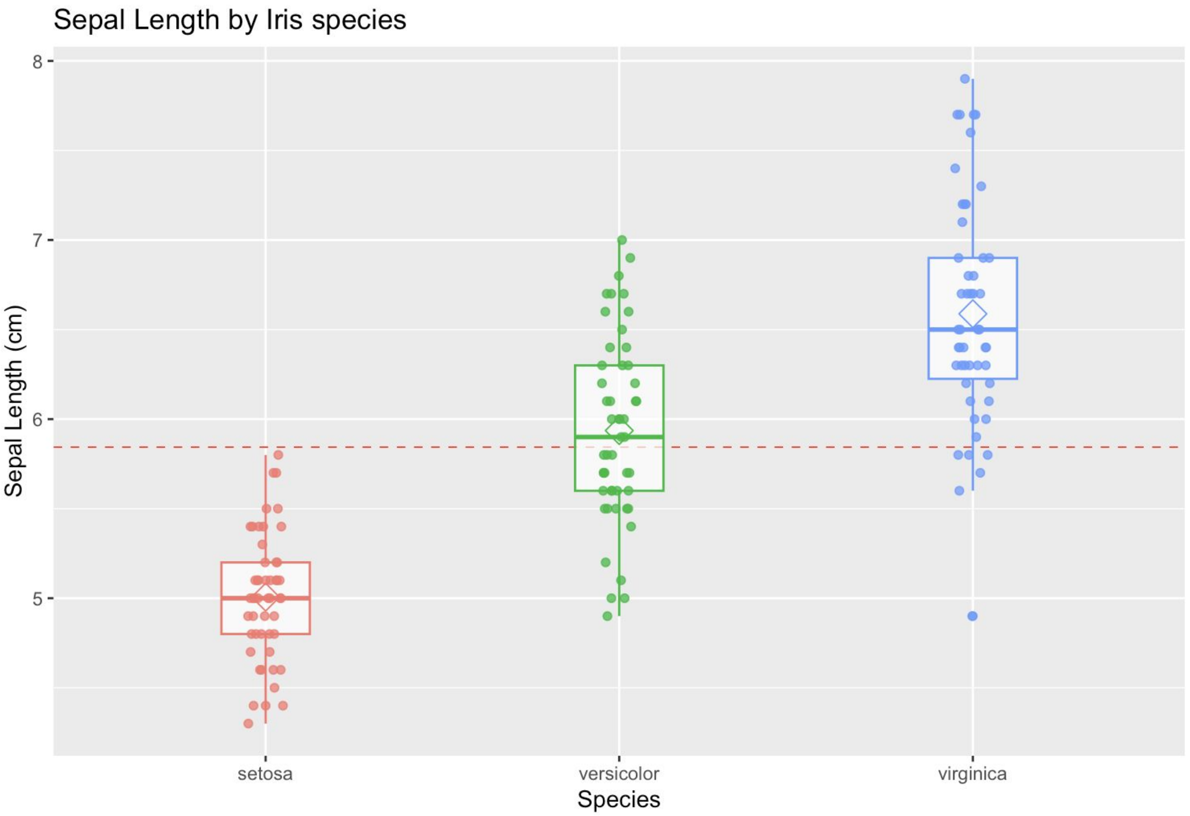
```
tukey_mydata <- TukeyHSD(aov_mydata)
emm_cont <- emmeans(aov_mydata, specs = "Species")
emm_cont %>% pairs(adjust = "tukey") %>% print()
plot(emm_cont,
     comparisons = TRUE,
     adjust = "tukey",
     horizontal = FALSE)
```

Part 5: Test differences among groups



contrast	estimate	SE	df	t.ratio	p.value
setosa - versicolor	-0.198	0.0209	147	-9.470	<.0001
setosa - virginica	-0.328	0.0209	147	-15.667	<.0001
versicolor - virginica	-0.130	0.0209	147	-6.198	<.0001

Part 6: Graph and save results for each response



Expected outcomes and products:

1. **ANOVA results (in a table) for all three responses**
2. **Figures for all three response variables**
3. **Methods paragraph 4 (due Mon 4/14)**
    a. What statistical techniques did you use?
    b. Did you have to remove outliers or transform your data? ***Keep notes on this today!!!***
    c. I will give you a skeleton and sources for this, but you will be responsible for details!

Plant initial height example together

1. Cleaning and loading our data
2. Informal investigation
3. Hypothesis testing
4. Interpreting results
5. Testing for differences among groups
6. Graphing our results

Plant number	Treatment	Initial height	Final height	Root length	Above ground mass	notes
1	control	19	43.5	19.5	53.4	
2	control	15.1	44.7	23	49.91	
3	control	22.5	54.3	15	45.04	fun curve in stem
4	control	15	46.5	17	32.56	
5	control	14.6	39	21	42.35	
6	control	14.5	42	20	50	
7	control	8.3	35	29	28.3	
8	control	16.5	41.5	20	50.48	
9	control	16.5	48.1	21	46.65	
10	control	13	45.5	18	39.39	
11	control	13.5	36	31	38.19	
12	control	13.5	39.5	19	36.64	
13	control	12.4	36	30.5	29.39	very wilty
14	control	17.2	45	11.5	32.62	
15	control	14	31	17.5	35.73	
16	control					missing
17	control	9	40.5	20	42.19	
18	control	12	37	21	48.39	
19	control	16	34	23	27.86	
20	control	11.8	25	19	27.54	
21	control					missing
22	control	31	56	18	67.15	de3vion
23	control	16	36	19	41.35	
24	control	16.5	41.5	21	46.47	
25	control	13	37	18	36.12	
26	control	16	35	26	40.8	
27	control	17.5	41	19	35.91	
28	control	14.9	35	20	29.79	
29	control	17.3	41	18.5	40.27	
30	control	16.6	44.5	20.5	38.56	wilty
31	half leaf	14.6	36	27	31.54	a little crusty
32	half leaf	15.8	31	25	28.97	
33	half leaf	12.8	39	22	33.72	
34	half leaf	16.9	52	24	44.9	
35	half leaf	23	46	14	31.32	
36	half leaf					missing
37	half leaf	15.7	32	13	22.73	
38	half leaf	20.5	42	17	32.2	
39	half leaf	16.8	37	17	33.73	
40	half leaf	16	37	18.5	25.43	
41	half leaf	16.2	38.5	22	31.11	