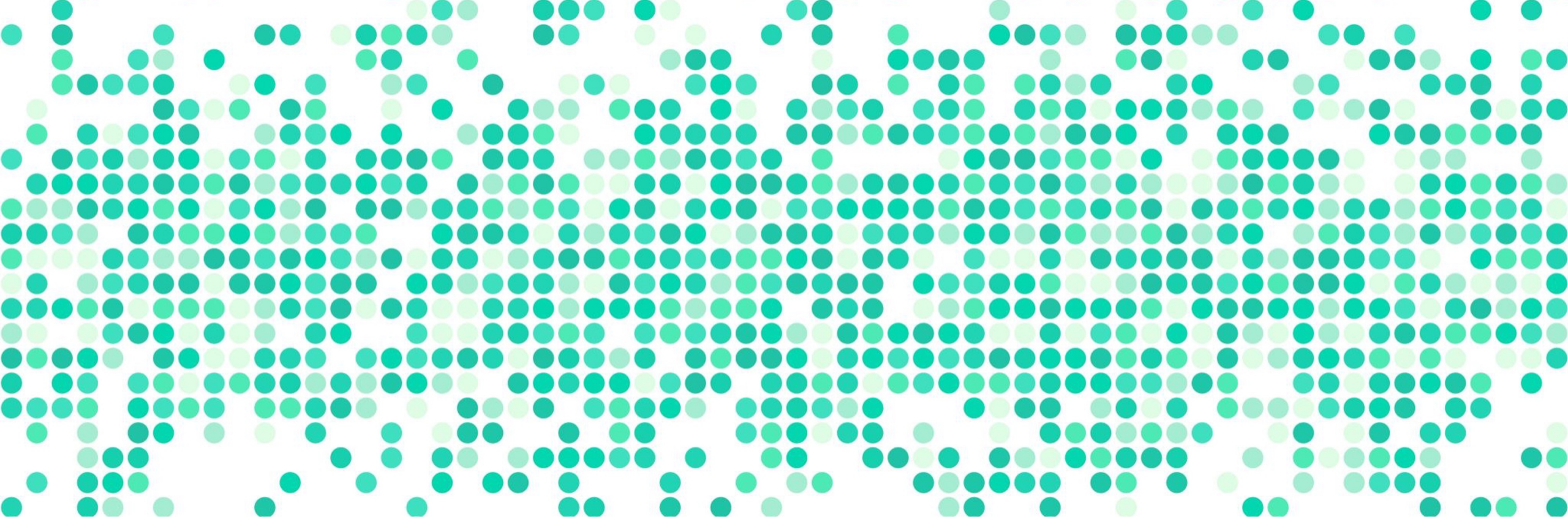


DATA ANALYSIS

Week 10: Modeling Relationships I continued...

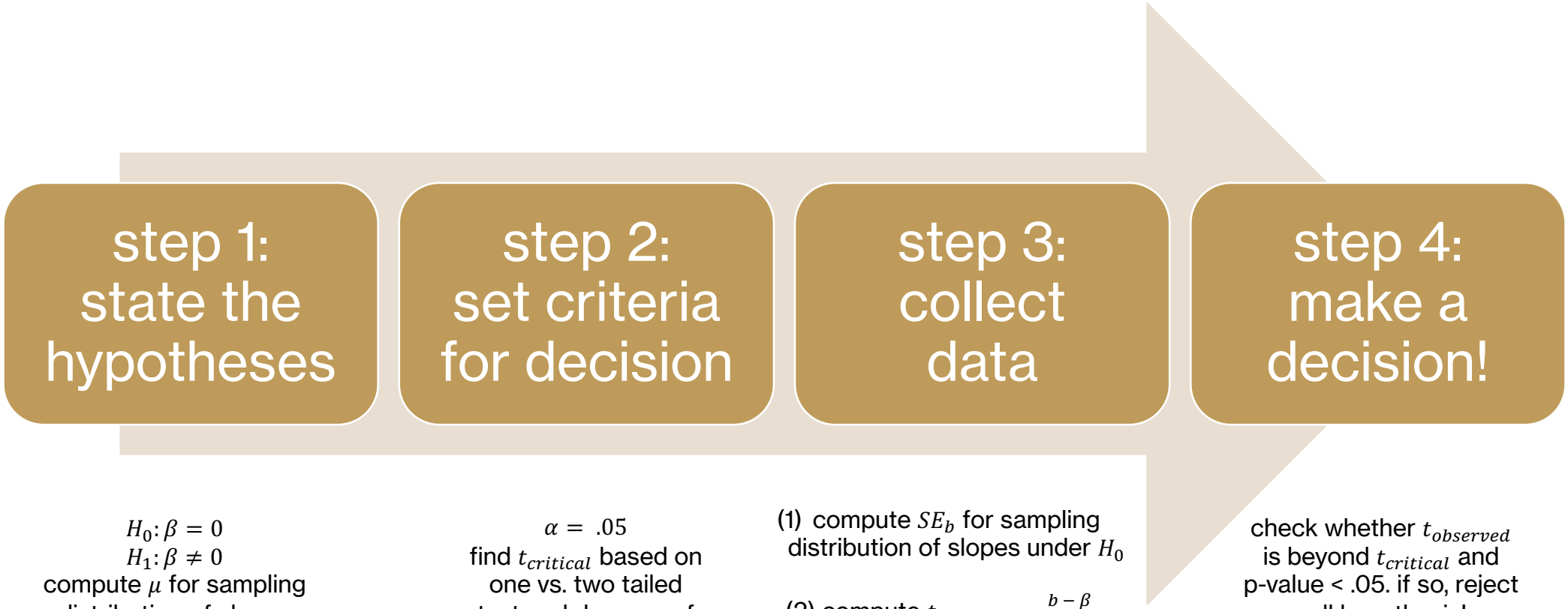
logistics

- problem set #5 (due April 2)
  - might change/remove some problems, will send email by tonight
- extra credit March survey (due April 9, link on Canvas)
  - make sure to submit completion code on Canvas!

today's agenda

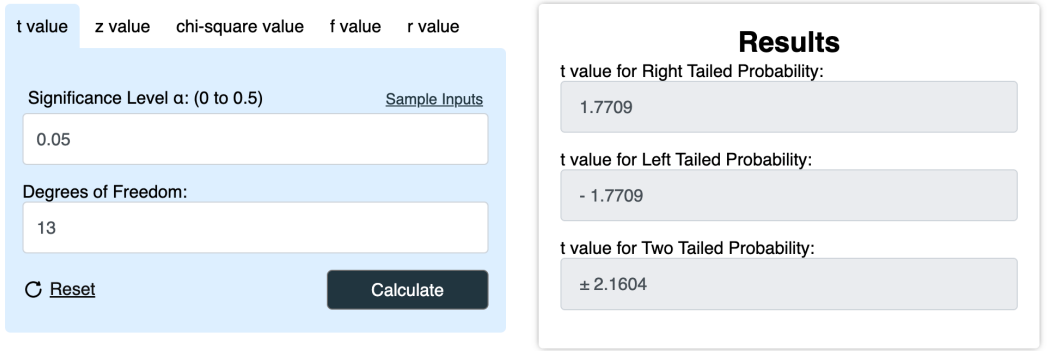
hypothesis testing with one independent variable continued...

review: NHST for linear regression (t-test)



review: NHST for linear regression

- **step 1: state the hypotheses**
  - H_0: $\beta = 0$
  - H_1: $\beta \neq 0$
- **step 2: set criteria for decision**
  - $t_{n-2} = t_{13} = t_{critical} = 2.16 \; at \; \alpha = .05$
- **step 3a: fit the line**
  - compute the slope
    - $b = r \frac{s_y}{s_x} = 3.45$
  - compute the intercept
    - $a = M_y - bM_x$ = -87.51667



r	s_y	s_x	b = r s_y/s_x	a = My - bMx
0.9954947678	15.49869426	4.472135955	3.45	-87.51666667

review: computing SE_b (option 1)

- **step 3b: compute standard error SE_b for slope**
  - $SE_b = \frac{SE_{model}}{\sqrt{\sum(X-M_x)^2}}$
  - compute $SS_{error} = \sum(Y - \hat{Y})^2$ = 30.23
  - compute $SE_{model} = \sqrt{\frac{SS_{error}}{n-2}}$ = 1.525
  - compute $SS_X = \sum(X - M_x)^2$ = 280
  - compute $SE_b = \frac{SE_{model}}{\sqrt{\sum(X-M_x)^2}}$= .0911

Yhat=a + bX	Y-Yhat	Y-Yhat sq
112.5833333	2.416666667	5.840277778
116.0333333	0.9666666667	0.9344444444
119.4833333	0.5166666667	0.2669444444
122.9333333	0.06666666667	0.004444444444
126.3833333	-0.3833333333	0.1469444444
129.8333333	-0.8333333333	0.6944444444
133.2833333	-1.283333333	1.646944444
136.7333333	-1.733333333	3.004444444
140.1833333	-1.183333333	1.400277778
143.6333333	-1.633333333	2.667777778
147.0833333	-1.083333333	1.173611111
150.5333333	-0.5333333333	0.2844444444
153.9833333	0.01666666667	0.000277777777
157.4333333	1.566666667	2.454444444
160.8833333	3.116666667	9.713611111
		SSerror
		30.23333333

SEmodel
1.525005254

(X-Mx)^2
49
36
25
16
9
4
1
0
1
4
9
16
25
36
49
Sum of X-Mx sq
280

SSerror	Sum of X-Mx sq
30.23333333	280
SEmodel	
1.525005254	
SE_b	
=N20/SQRT(O18)	

review: computing SE_b (option 2)

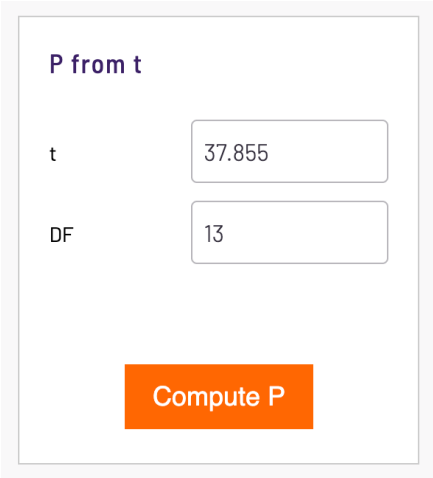
- **step 3b: compute standard error SE_b for slope**
  - $SE_b = \frac{SE_{model}}{\sqrt{\sum(X-M_x)^2}} = SE_r \frac{s_y}{s_x}$ (proof)
  - compute $SE_r = \sqrt{\frac{1-r^2}{n-2}}$ = 0.026
  - compute $SE_b = SE_r \frac{s_y}{s_x}$ = .0911

r	s_y	s_x
0.9954947678	15.49869426	4.472135955
SE_r		
0.02629736359		
SE_b		
0.09113649547		

review: NHST for linear regression

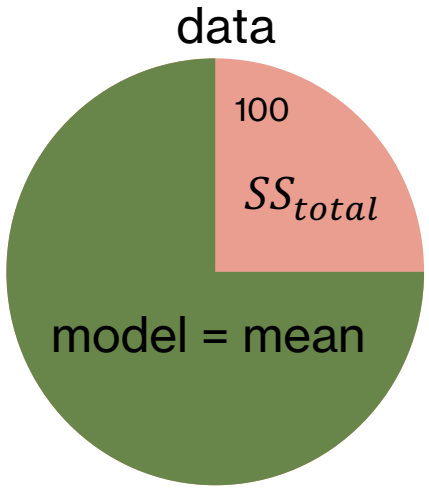
- **step 3c: compute $t_{observed}$ and p-value**
  - compute $t_{observed} = \frac{b-0}{SE_b} = \frac{3.45}{.0911} = 37.855$
  - obtain p-value: $p < .0001$
  - note that $t_{observed}$ is the SAME as the t-value obtained from the correlation t-test!!
- **step 4: decide!**
  - height significantly predicts weight, $b = 3.45$, $t(13) = 37.86$, $p < .001$

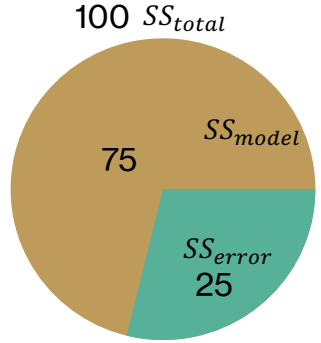
t_observed	p_value
37.85530684	< .0001
check	
TRUE	

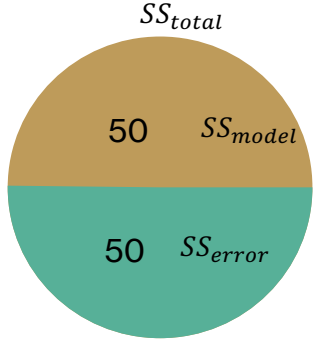


overall test of model: ANOVA

- sometimes, you may want to conduct a test for the overall model instead of testing for significance of individual coefficients (multiple b's)
- in such cases, we resort to an **an**alysis **o**f **va**riance (ANOVA)
  - $SS_{total} = SS_{model} + SS_{error}$
- we can calculate the ratio between the variance explained by the model and the natural variance expected/left over in the dependent variable
  - if $\frac{SS_{model}}{SS_{error}}$ is high, the model explains **more** variance than expected
  - if $\frac{SS_{model}}{SS_{error}}$ is low, the model explains **less** variance than expected
- typically, we want the "average" variance explained, so we also divide this by *df*







F ratio

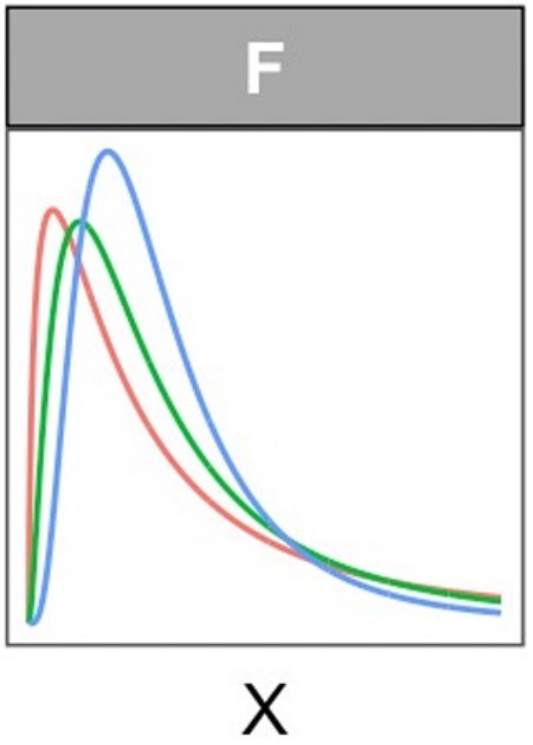
- The F ratio compares the "average" squared error between model (explained variance) and the natural (unexplained) variance (data = model + error)

$$F = \frac{explained\ variance}{unexplained\ variance} = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}}$$

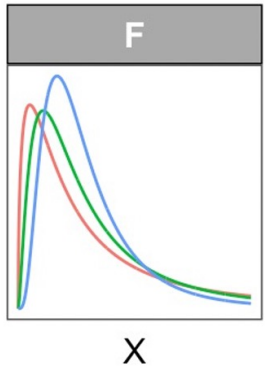
- obtaining SS_{model} and SS_{error}
  - $SS_{error} = \sum(Y - \hat{Y})^2$ and $SS_{total} = \sum(Y - M_y)^2$
  - $SS_{model} = SS_{total} - SS_{error}$
- obtaining df_{model} and df_{error}
  - k denotes the number of levels of the independent variable OR number of estimated parameters
  - $df_{model} = k - 1$
  - $df_{error} = n - k$

interpreting F values

- The F-distribution is a positively skewed distribution
- defined by two parameters (df_1 and df_2) that determine the exact form/shape
- F-values are typically **non-negative**: why??
  - $F = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}}$
  - $F = 1$: $MS_{model} = MS_{error}$ i.e., the model does not do any better than random chance
  - $F > 1$: more variance explained by model than random chance
- F ratios are another way to assess model fit (in addition to R^2 and standard error!)



NHST for linear regression (F-test)



step 1: state the hypotheses

$H_0: \beta = 0$
$H_1: \beta \neq 0$

step 2: set criteria for decision

$\alpha = .05$
find $F_{critical}$ based on **right** tailed test and degrees of freedom
$df_1 = k - 1$
$df_2 = n - k$

step 3: collect data

(1) compute SS_{model} and SS_{error}

(2) compute $F_{observed} = \frac{MS_{model}}{MS_{error}}$

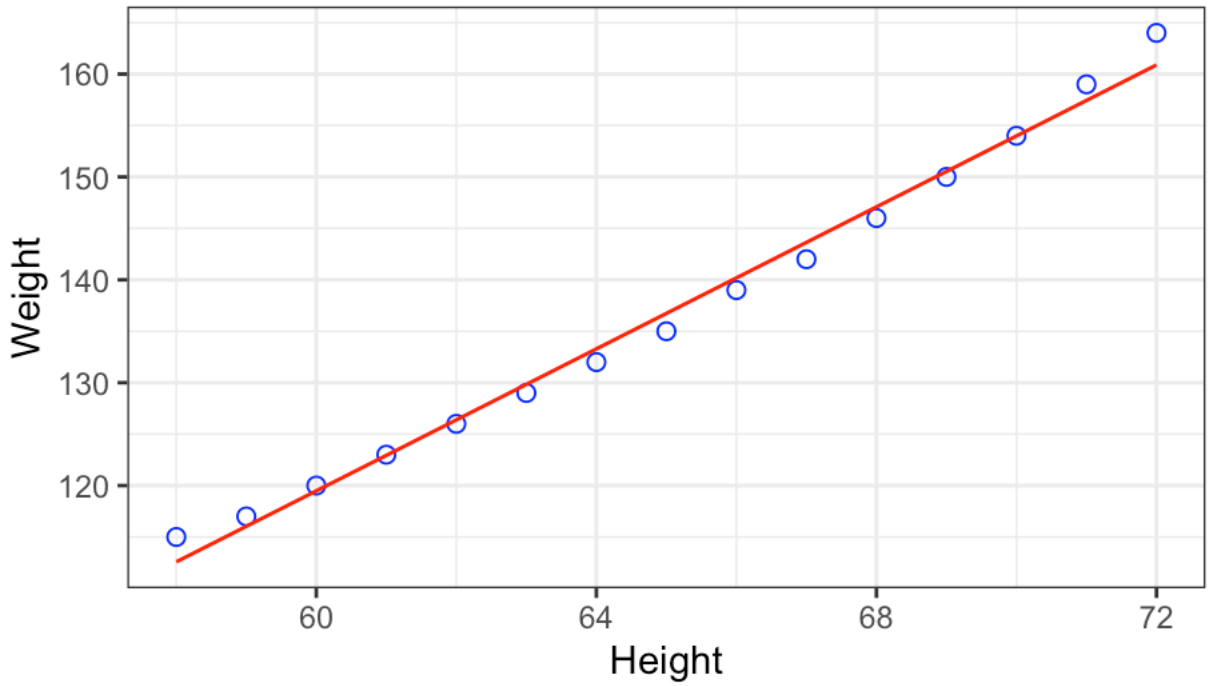
(3) find p-value for F-score

step 4: make a decision!

check whether $F_{observed}$ is beyond $F_{critical}$ and p-value < .05. if so, reject null hypothesis!

F-test for women dataset

- $F_{critical} = F\ (k-1, n-k) =\ F\ (1, 13) = 4.667$
- SS_{error} = 30.23 and SS_{total} = 3362.93
- thus, SS_{model} = $SS_{total}\ - SS_{error}$ = 3332.7
- $F = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}} = \frac{3332.7/1}{30.23/13} = 1433$
- $F\ (1, 13) = 1433, p <\ .0001$
- again, the model explains significantly more variance in the data than chance
- in fact, $t^2 =\ F$!!



F-tables

- F-tests are typically represented in tables

		SS	df	MS	F	p-value
SS_{model}	IV	3332.7	1	3332.7	1433.02	<.0001
SS_{error}	residual	30.23	13	2.33		

- knowing parts of the F table are sufficient for completing it!

questions

hypothesis tests in R

```
data("women")
View(women)

weight_model = lm(data = women, weight ~ height)
summary(weight_model)
car::Anova(weight_model)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height        3.45000    0.09114   37.85 1.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

		SS	df	MS	F	p-value
SS_{model}	IV	3332.7	1	3332.7	1433.02	<.0001
SS_{error}	residual	30.23	13	2.33		

```
Anova Table (Type II tests)

Response: weight
           Sum Sq Df F value    Pr(>F)
height    3332.7  1    1433 1.091e-14 ***
Residuals   30.2 13
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

data come in all forms

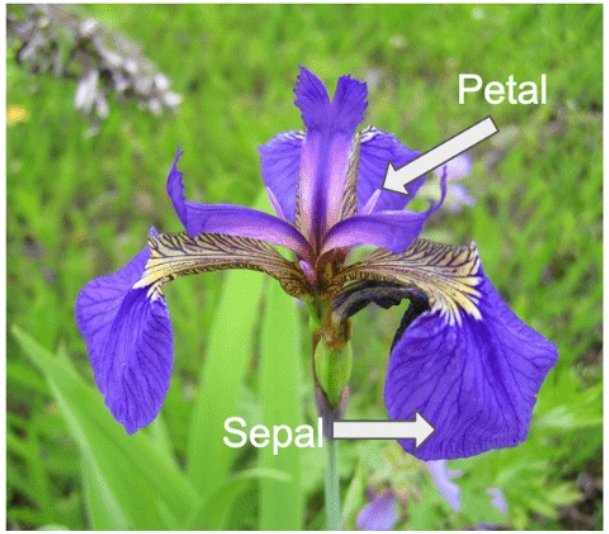
- think back to NOIR: what kinds of data have we worked with so far?
- independent variables can be nominal/ordinal
  - examples?
- dependent variables can be nominal/ordinal
  - examples?
- when data are **not interval/ratio**, the same general framework of linear models can be applied, with a few modifications

	independent variable		
dependent variable	nominal	ordinal	interval/ ratio
nominal			
ordinal			
interval/ratio			r or b

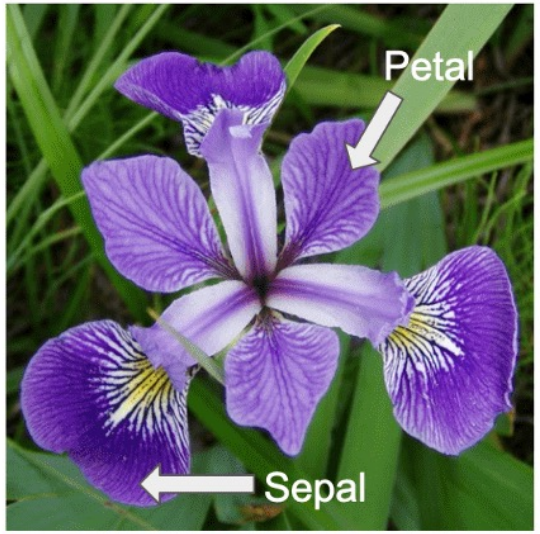
example: iris dataset

- the iris dataset contains petal and sepal dimensions for three species (setosa, virginica, and versicolor)

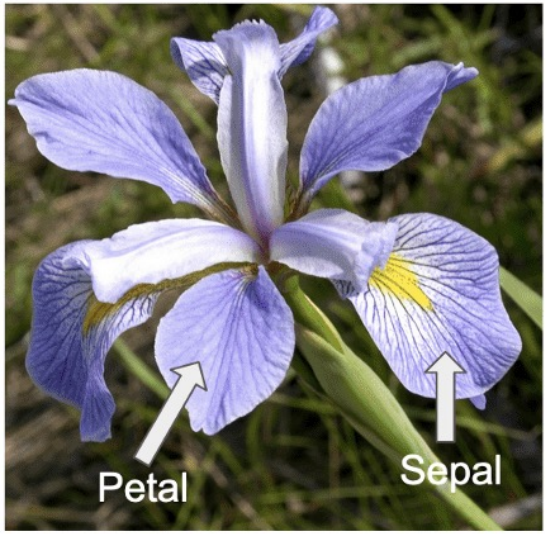
Iris setosa



Iris versicolor

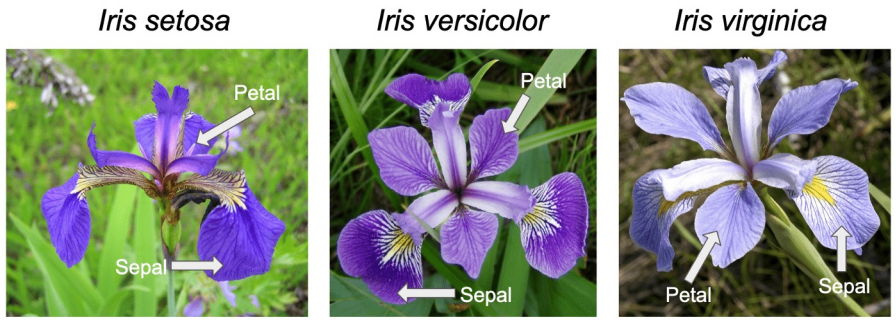


Iris virginica



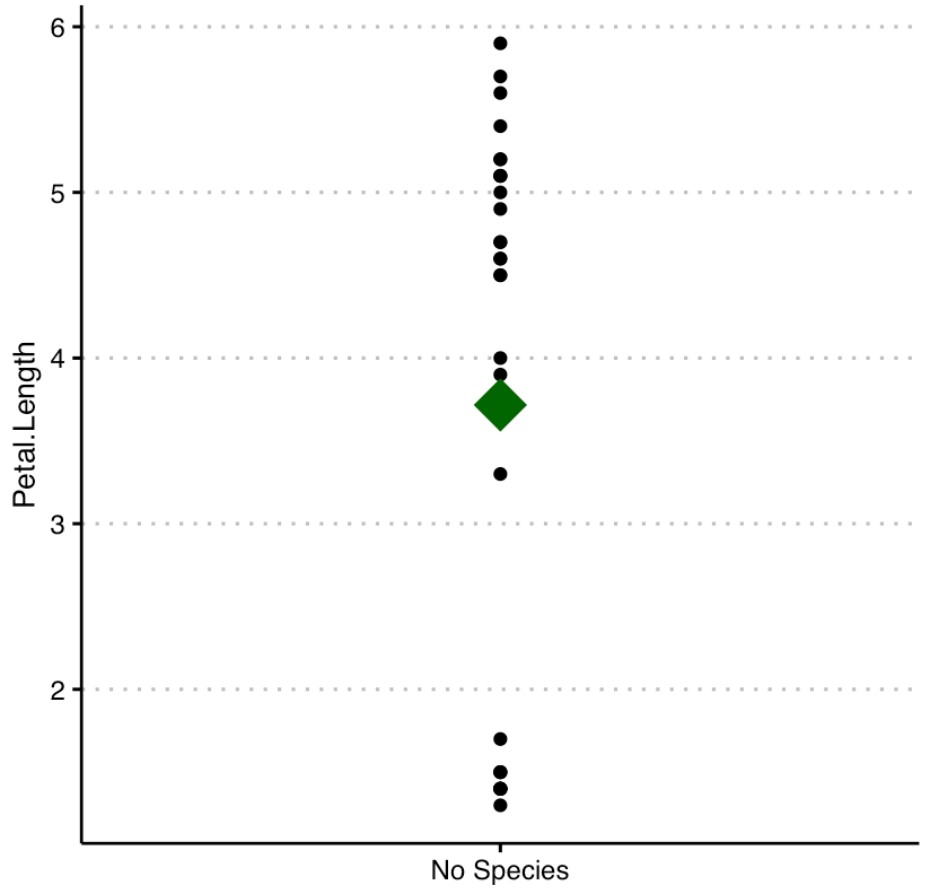
example: iris dataset

- our goal is to build the best model for petal lengths
- if there were no species labels in this dataset, what would be the best model of petal lengths?



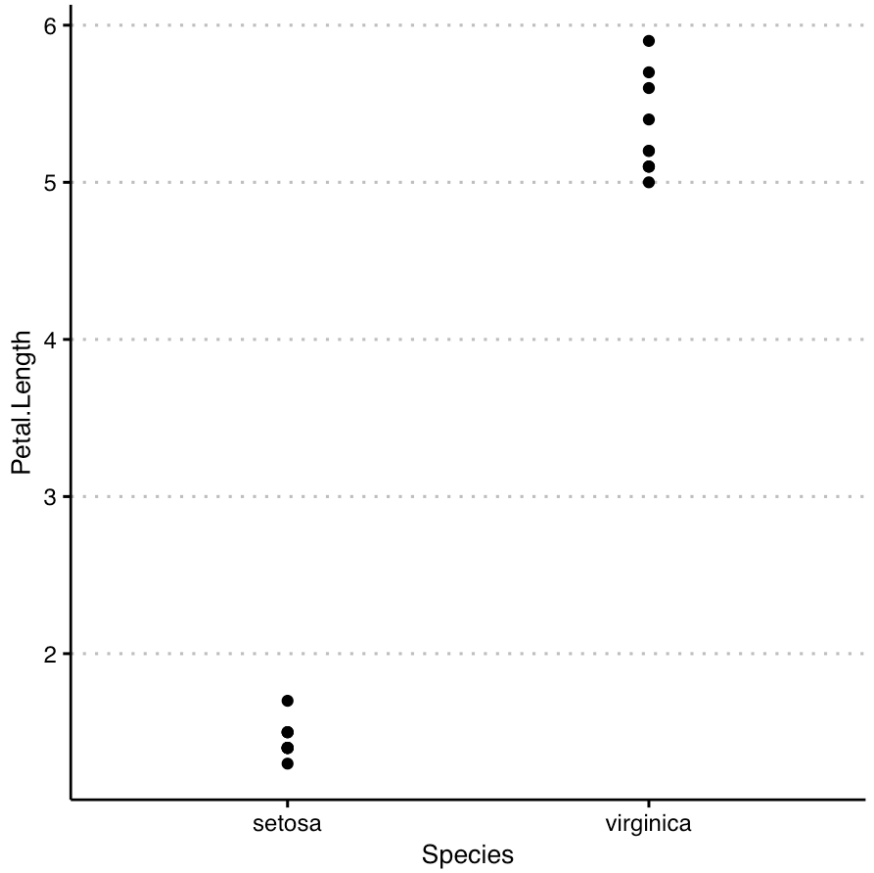
example: iris dataset

- if there were no species labels in this dataset, the overall or “grand mean” of all petal lengths would be the best model for the data
- this “grand mean” will provide our baseline, i.e., how much better can we do than the grand mean in fitting a model to the data?



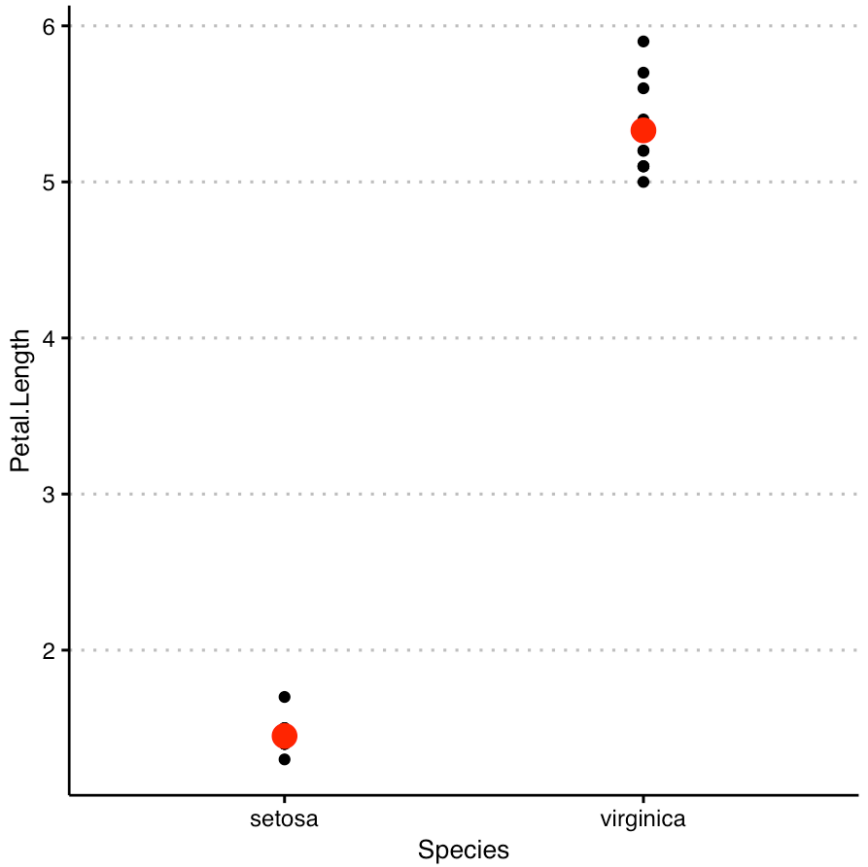
comparing two groups

- our goal is to fit a different model to the data that includes species as additional information and evaluate how much better we can do than the grand mean
- Y (petal lengths) = X (species) + error
- instead of a continuous scale of values, X (species) can only take two values: how can we fit a line?



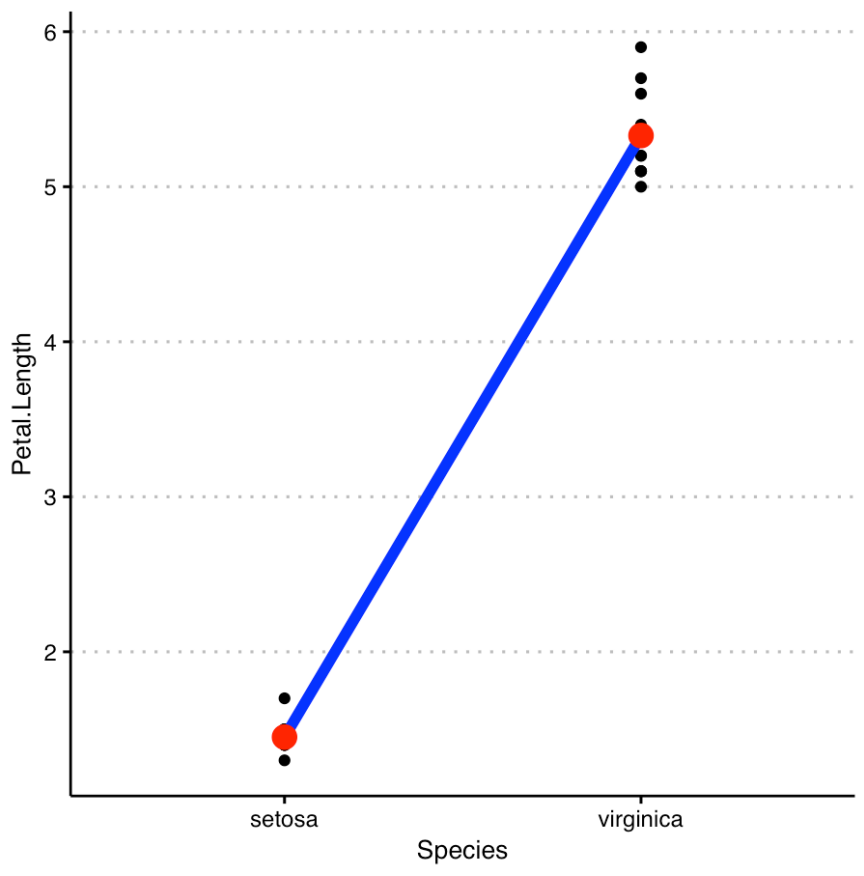
comparing two groups

- we could take the mean of each group



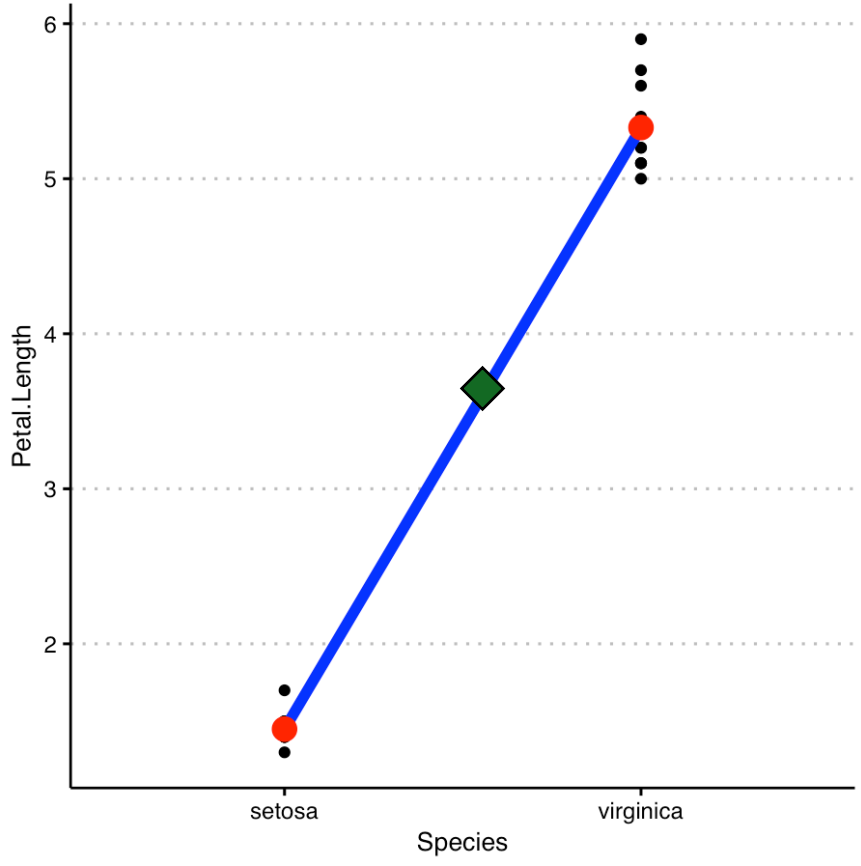
comparing two groups

- we could join the line that connects the two means?
- the slope of this line is the change in the dependent variable (petal length) when the level of the independent variable (species) changes
- the intercept is the mean of one level of the IV
- Y (petal lengths) = X (species) + error
- a “one-unit” change in the IV is simply going from one level of the IV to another (i.e., from one group to another)



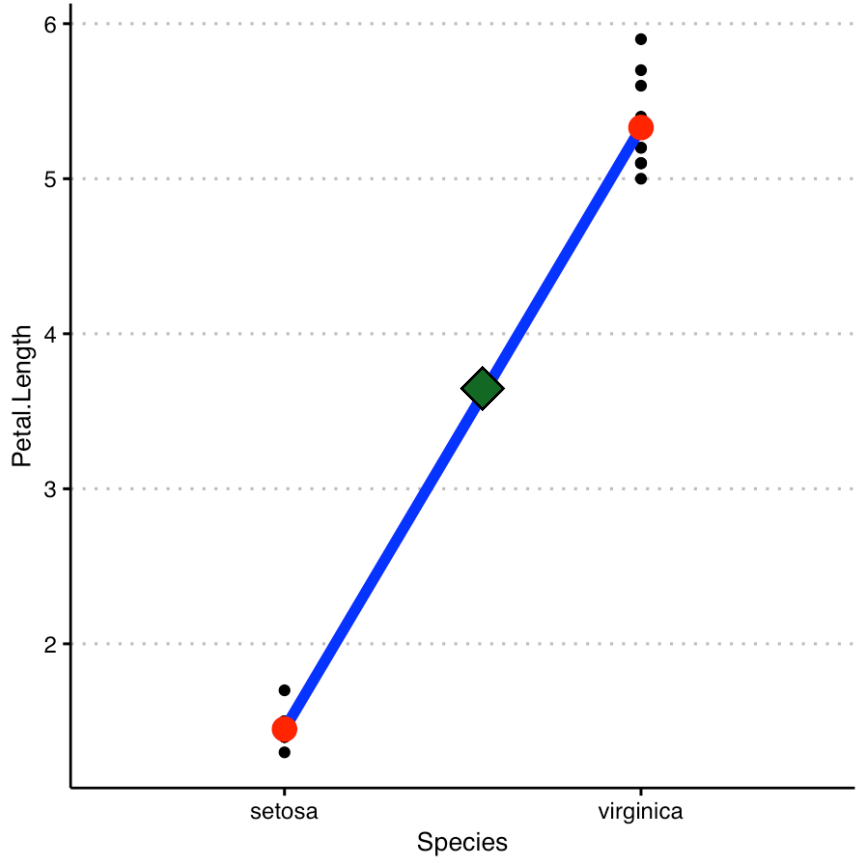
comparing two groups

- Y (petal lengths) = a + b X (species) + error
- X is now an indicator of *levels* of the independent variable
  - when X= 0 (first group), Y = a
  - when X = 1 (second group), Y = a + b
- the intercept is the mean for the first group (X = 0)
  - a = M_{group1}
- the slope is the change in means from first to second group (X = 1)
  - b= $M_{group2} - M_{group1}$
- essentially, we are “fitting” a model to the data that substitutes the mean of the individual species instead of the grand mean



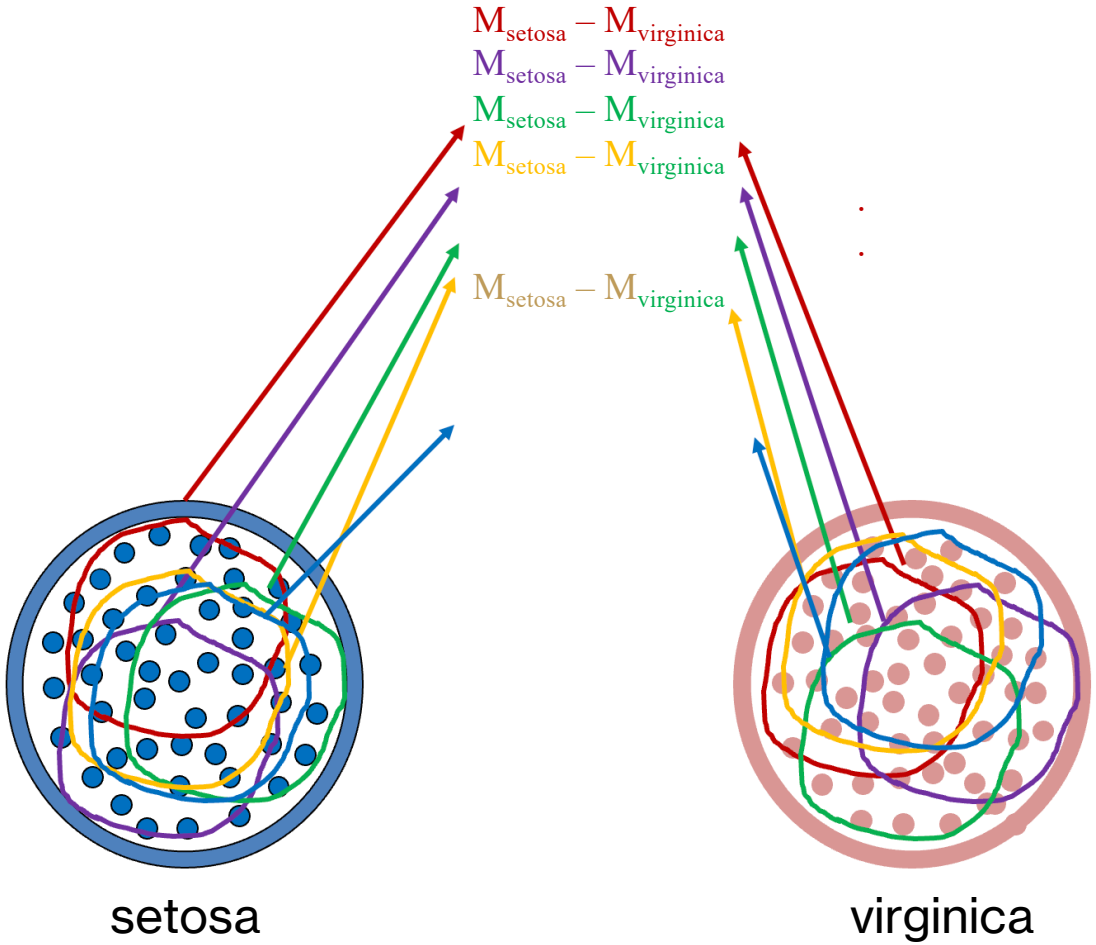
comparing two groups

- we have the slope and intercept formulation
  - a = M_{group1}
  - b= $M_{group2} - M_{group1}$
- we compute the slope and intercept for iris data
  - a = M_{setosa} = 1.45
  - b = $M_{virginica} - M_{setosa}$ = 5.33 – 1.45 = 3.88
- line's equation is:
  - Y = 1.45 + 3.88 (X) where X = {0,1}
- the model is the species means instead of the grand mean



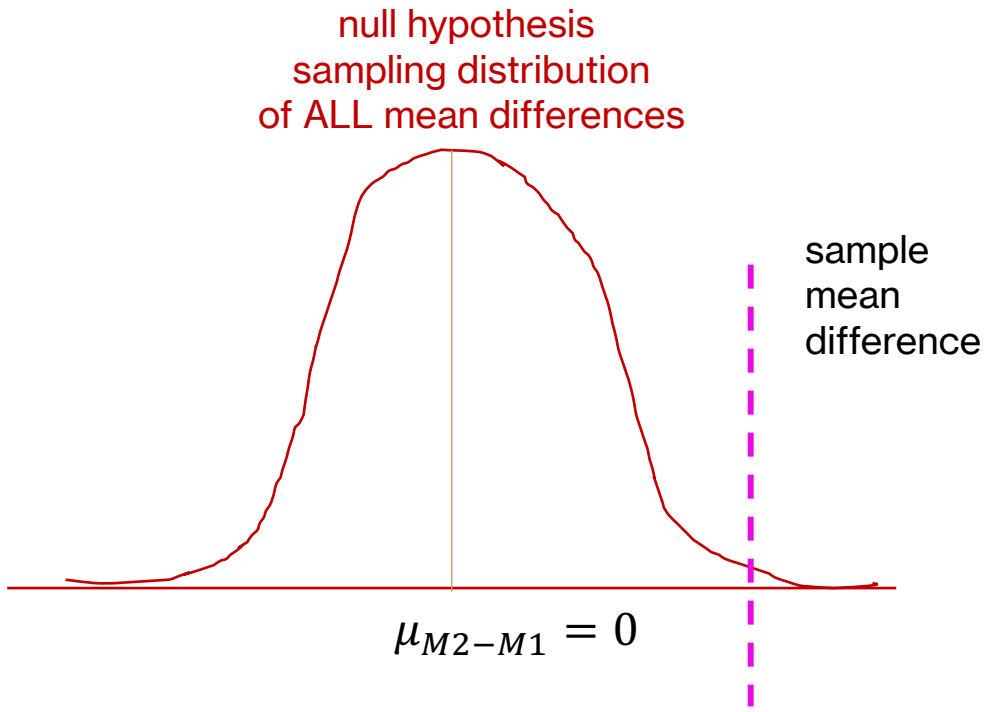
hypothesis testing for two groups

- our hypothesis is that species is informative about petal lengths, OR that the group means are different OR come from different populations
- how do we test this hypothesis?
- we could take ALL samples of size n from the two groups and compute the difference between those group means
- these mean differences will form the sampling distribution
- we can then compare where our actual sample mean difference lies relative to this sampling distribution of ALL mean differences



hypothesis testing for two groups

- under the null hypothesis, this sampling distribution of mean differences for all pairs of possible samples of size n follows a t-distribution with a mean $\mu_{M2-M1} = 0$
- next, we need to figure out degrees of freedom for this t-distribution and standard error (or standard deviation of this sampling distribution)



independent measures t-test

- step 1: state the hypotheses
  - H_0: β = 0: mean petal lengths for both species are equal, i.e.,
    - $\mu_{setosa} = \mu_{virginica}$ OR $\mu_{virginica} - \mu_{setosa} = 0$
  - H_1: β ≠ 0: mean petal lengths for both species are not equal, i.e.,
    - $\mu_{setosa} \neq \mu_{virginica}$ OR $\mu_{virginica} - \mu_{setosa} \neq 0$
- step 2: set criteria for decision

$$t_{df} = \frac{sample\ statistic\ (b) - population\ parameter\ (\beta)}{standard\ error} = t_{critical}$$

- step 3: collect data

$$t_{df} = \frac{b - 0}{SE} = \frac{(M_2 - M_1) - 0}{SE}$$

- step 4: make a decision!

step 2a: setting criteria for t-test

- to set criteria for decision, we need to define which t-distribution we will use
- for each sample, we take away 1 degree of freedom

$$df = df_1 + df_2 = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$$

$$t_{n_1+n_2-2} = t_{critical}$$

step 2b: standard error for t-test

- how do we calculate standard error? typically, when we had data from a single sample of scores, we looked at the “average” deviation from the mean

$$t = \frac{M - \mu}{s_M} \text{ and } s_M = \frac{s}{\sqrt{n}}$$

- we now have two different samples, so the standard error needs to take into account the variation from both samples (it is the average error to be expected between any two sample means under the null hypothesis)

$$s_{M_1} = \frac{s_1}{\sqrt{n_1}} \text{ and } s_{M_2} = \frac{s_2}{\sqrt{n_2}}$$

$$s_{M_2 - M_1} = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}} = \sqrt{{s_{M1}}^2 + {s_{M2}}^2}$$

step 2b: pooled variance

- when $n_1 = n_2$, $s_{M2-M1} = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$ makes sense as it equally weights both samples
- when $n_1 \neq n_2$, bigger samples should yield more accurate estimates of population variance than smaller samples and this formula may not be appropriate
- an estimate of pooled variance can be computed in such cases

$$ {s_{pooled}}^2 = {s_p}^2 = \frac{SS_1 + SS_2}{df_1 + df_2} $$

- then we can compute standard error as follows:

$$ s_{M2-M1} = \sqrt{\frac{{s_p}^2}{n_1} + \frac{{s_p}^2}{n_2}} $$

step 3: collect data

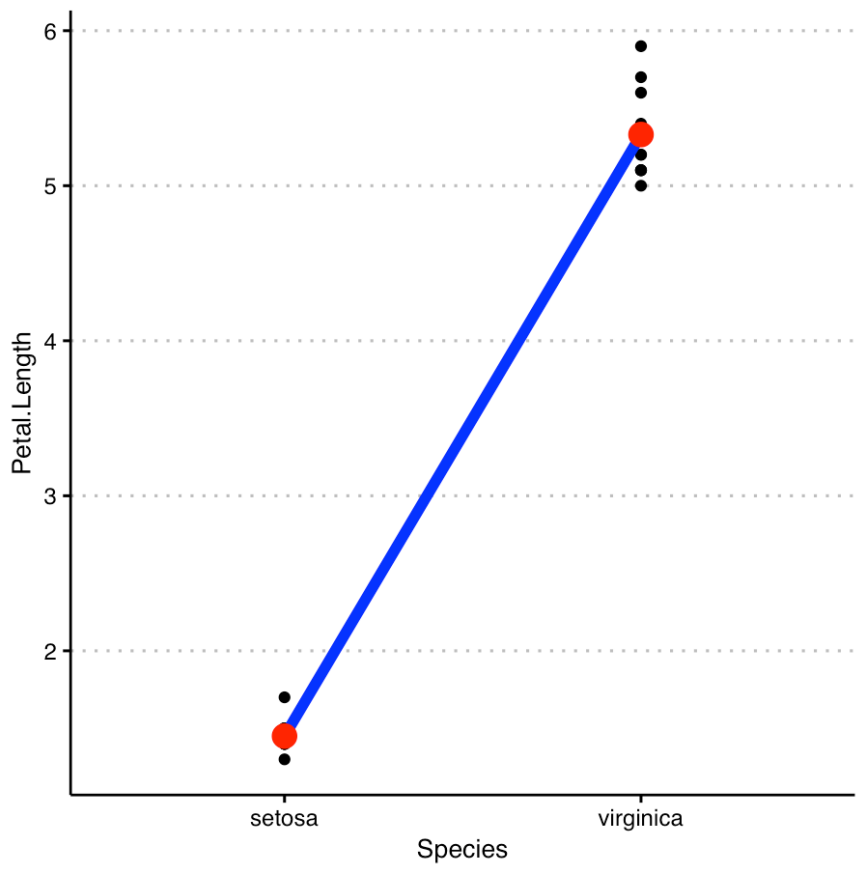
- we compute the standard error for our sample

$$s_{M2-M1} = \sqrt{\frac{{s_p}^2}{n_1} + \frac{{s_p}^2}{n_2}} = .1025$$

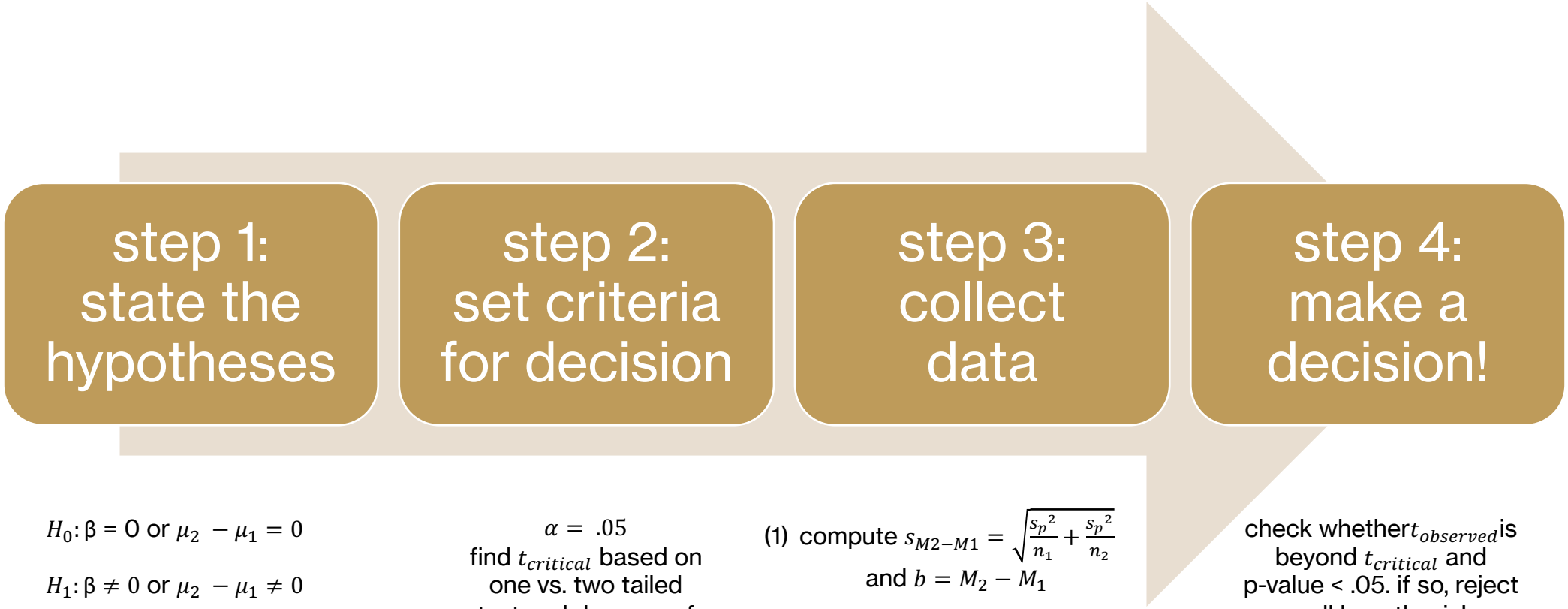
- we compute $t_{observed} = \frac{M_2 - M_1}{s_{M_2 - M_1}}$

$$t_{18} = \frac{b - \beta}{standard\ error} = \frac{b - 0}{s_{M2-M1}} = \frac{(M_2 - M_1) - 0}{s_{M_2 - M_1}} = \frac{3.88}{.1025} = 37.845$$

- $t_{18} = 37.845, p < .0001$
- therefore, we reject the null hypothesis that the petal lengths do not differ by species, i.e., species have different petal lengths!



NHST for two independent groups (t-test)



assumptions: t-test

- interval/ratio dependent variable
- independent observations
  - when observations are not independent, then we conduct a paired/dependent-samples t-test (later!)
- normality
  - when data are not normal, the t-test is not appropriate
  - BUT: t-tests are fairly robust to minor violations for large n
- homogeneity of variances
  - we assume that the populations from which samples are drawn have equal variances
  - Welch's test is done for unequal variances

independent t-test in R

```r
iris = read_csv("datasets/iris_subset.csv")
setosa = iris %>% filter(Species == "setosa")
virginica = iris %>% filter(Species == "virginica")
t.test(virginica$Petal.Length, setosa$Petal.Length, var.equal = TRUE)
```

```
        Two Sample t-test

data:  virginica$Petal.Length and setosa$Petal.Length
t = 37.845, df = 18, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.664606 4.095394
sample estimates:
mean of x mean of y
     5.33      1.45
```

comparing our tests so far: means

z-test: 1 DV, no IV	one sample t-test: 1 DV, no IV	two sample t-test: 1 DV, one nominal IV
• sample data: M • population parameter: μ • standard error: $\sigma_M = \frac{\sigma}{\sqrt{n}}$ • $z = \frac{M - \mu}{\sigma_M}$	• sample data: M • population parameter: μ • standard error: $s_M = \frac{s}{\sqrt{n}}$ • $t = \frac{M - \mu}{s_M}$	• sample data: $M_2 - M_1$ • population parameter: $\mu_2 - \mu_1$ • standard error: • $s_{M2-M1} = \sqrt{\frac{{s_p}^2}{n_1} + \frac{{s_p}^2}{n_2}}$ • $t = \frac{M_2 - M_1}{s_{M2-M1}}$

comparing our tests so far: correlation/slope

correlation t-test:
1 DV, 1 interval/ratio IV

- sample data: r
- population parameter: ρ
- standard error:
- $SE_r = \sqrt{\frac{1-r^2}{n-2}}$
- $z = \frac{r-\rho}{SE_r}$

slope t-test:
1 DV, 1 interval/ratio IV

- sample data: b
- population parameter: β
- standard error:
- $SE_b = \frac{SE_{model}}{\sqrt{\sum(X-M_x)^2}}$
- $t = \frac{b-\beta}{SE_b}$

slope F-test:
1 DV, 1 interval/ratio IV

- sample data: a, b
- population parameter: β
- $SS_{error} = \sum(Y - \hat{Y})^2$
- $SS_{total} = \sum(Y - M_y)^2$
- $SS_{model} = SS_{total} - SS_{error}$
- $F = \frac{MS_{model}}{MS_{error}}$

next time

- **before** class
  - *watch*: Hypothesis Testing (Linear regression: F-test) [13.5 min]
  - *read (optional)*: Chapter 10 from the Gravetter & Wallnau (2017) textbook.
  - *watch*: Hypothesis Testing (two-groups: t-test) [11 min]
  - *work on*: Problem Set 5!
- **during** class
  - F-tests/ANOVAs on nominal data

optional: why add the standard errors?

- range of scores in population I and II?
  - population I: 70-50 = 20
  - population II: 30-20 = 10
- when we take differences between means from population I and II, what is the largest difference possible?
  - largest: 70-20= 50
  - smallest: 50-30 = 20
- standard error of mean differences = range of population I + range of population 2! = 20+10 = 30!