ECo 602 - Analysis of Environmental Data

Interactions, Dummy Variables, and Model Interpretation

Michael France Nelson

Eco 602 – University of Massachusetts, Amherst – Fall 2021

Michael France Nelson

Dummy Variables and Interactions

What's in This Section?

Slide Show	Take-Home Concepts

• Dummy variables
• Categorical predictors (factors)
• Dummy variables
• Design matrix
• Factor levels and model coefficients
• Interactions
	• How to represent categorical data in a regression equation.
	• Interpreting factor coefficients as slopes.
	• What does the base case represent?
	  • For categorical data
	  • For numerical data
	• Representing and interpreting an interaction

Analysis of Variance and Linear Models

I claim that all of the Group 1 methods are really linear regressions.

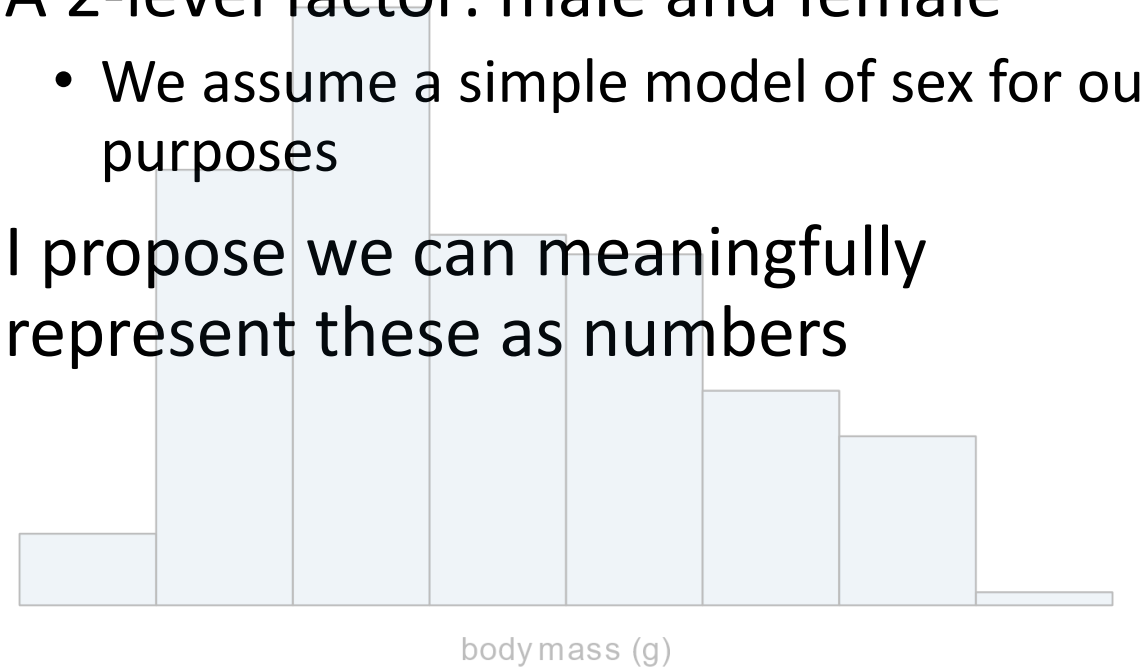
This includes the models we've used for one- and multi-way ANoVA

Regression Equation

- But… how can we draw a line on an x-y plane when the x-axis is a category?
  - What would the slope represent?
- A linear model: $y = \alpha + \beta x + \epsilon$
  - Recall what the components mean?

Penguin Sex

- A 2-level factor: male and female
  - We assume a simple model of sex for our purposes
- I propose we can meaningfully represent these as numbers



Dummy Variables

Our Sex Model

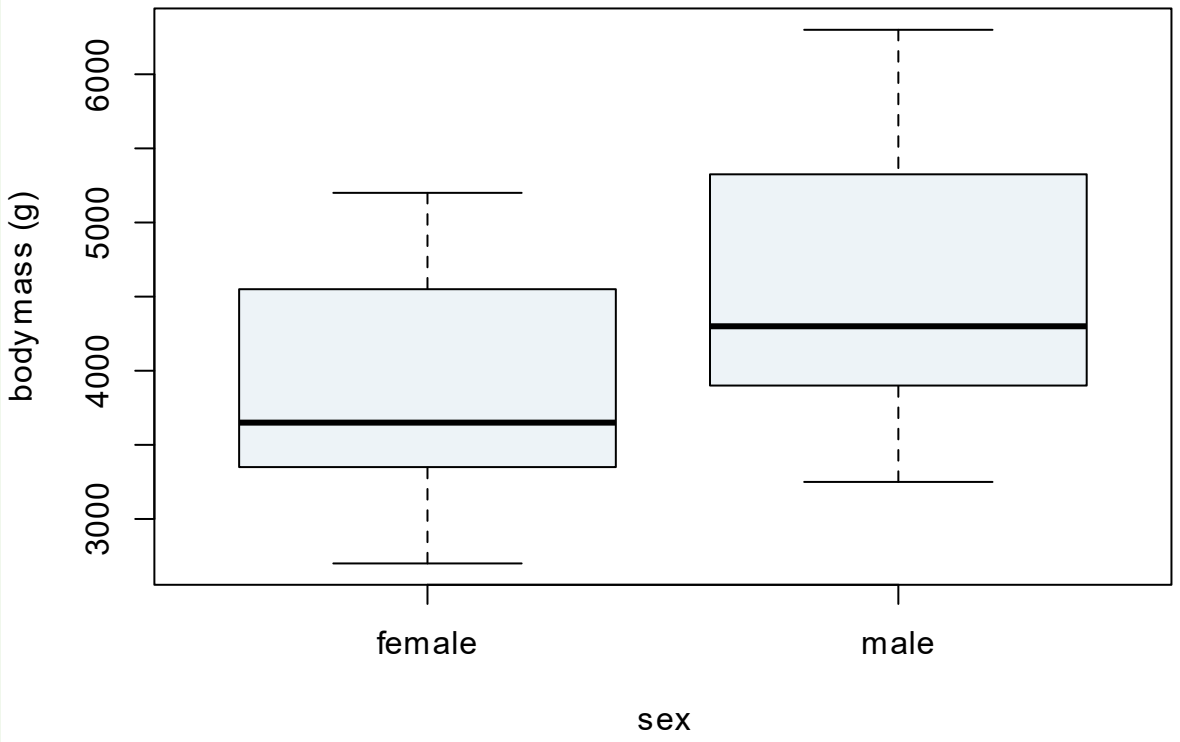
A deterministic function is a *model of the means*. With a model of penguin body mass as a function of sex:

$$y = [\text{intercept}] + \text{sex} + \epsilon$$

We propose to explain body mass by sex only. Well, sex and an intercept, that is!

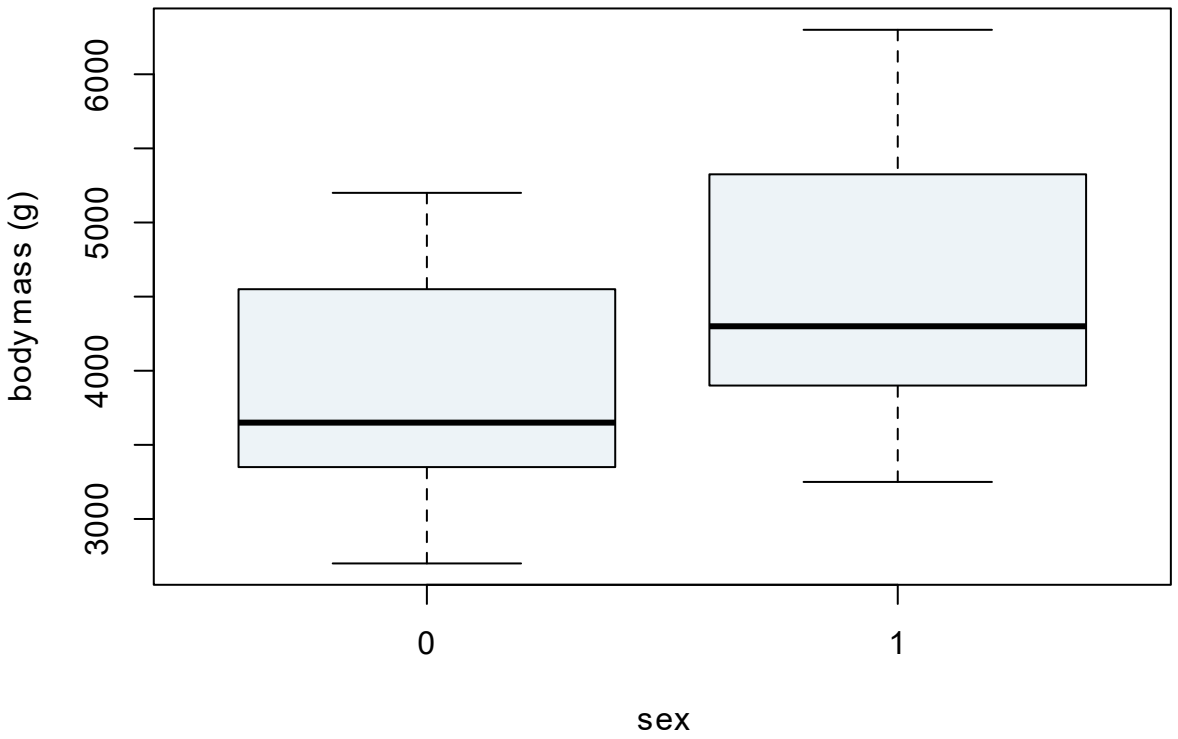
… but wait, the intercept turns out to represent our base-case sex, but we'll get to that.

Sex and Body Mass

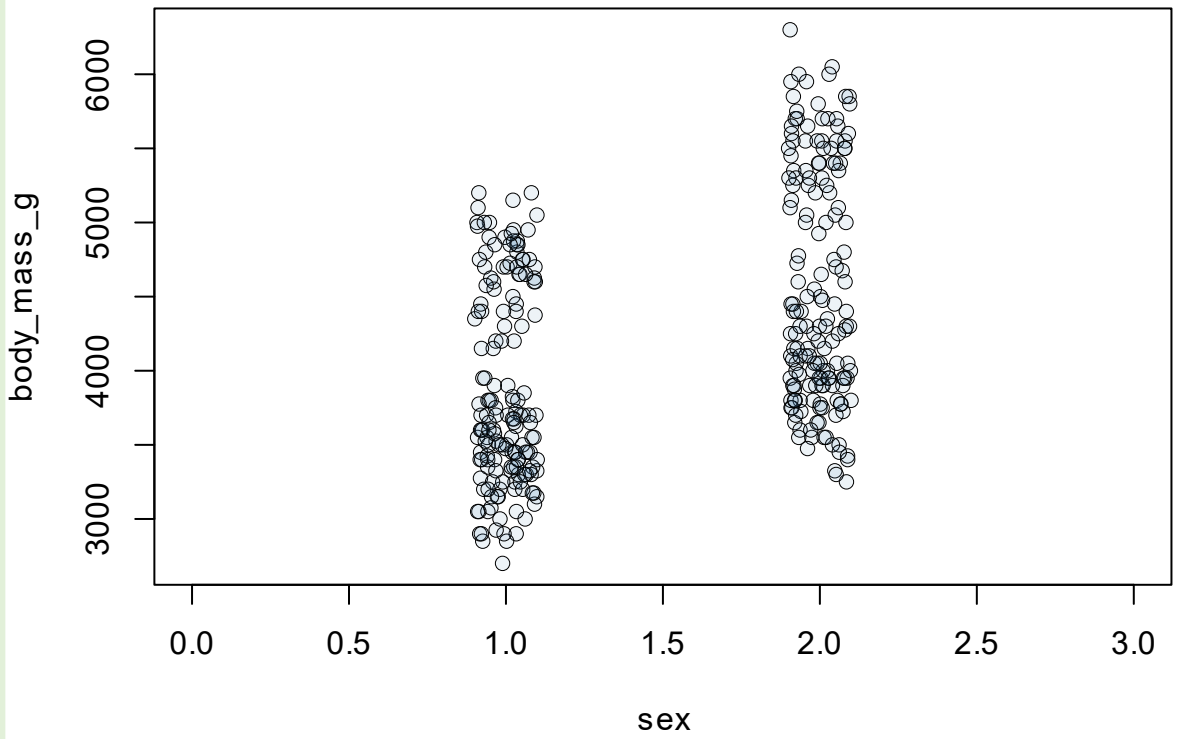


Dummy Variables: Sex as Numeric

We could relabel sex as a number.



Plot the points rather than a summary.



Dummy Variables: Sex as Numeric

Re-code sex as a number

```
# load the package
require(palmerpenguins)

# recode to 0 and 1
penguins$n_sex =
  as.numeric(penguins$sex) - 1

# take a look head(penguins)
```

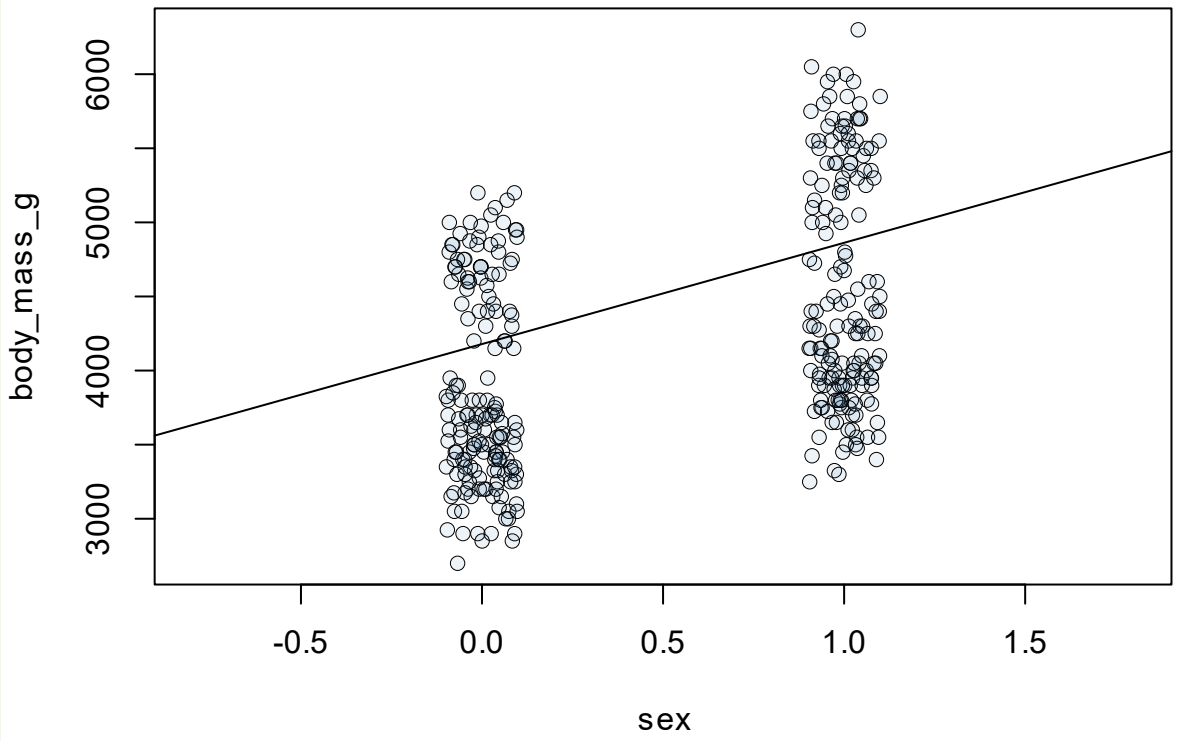
	species	sex	n_sex	body_mass_g
1	Adelie	male	1	3750
2	Adelie	female	0	3800
3	Adelie	female	0	3250
4	Adelie	NA	NA	NA
5	Adelie	female	0	3450
6	Adelie	male	1	3650

Dummy Variables: Sex as Numeric

We could fit a linear model to that!

Ignore the obvious normality issues!

```
fit_sex =
   lm(
      body_mass_g ~ sex_n,
      data = penguins)
```



Model Design Matrix

We recoded sex to numeric

It's a binary representation

- Every entry in n_sex is either 0 or 1
- It's categorical, but we can consider it numeric, but why?

	species	sex	n_sex	body_mass_g
1	Adelie	male	1	3750
2	Adelie	female	0	3800
3	Adelie	female	0	3250
4	Adelie	NA	NA	NA
5	Adelie	female	0	3450
6	Adelie	male	1	3650

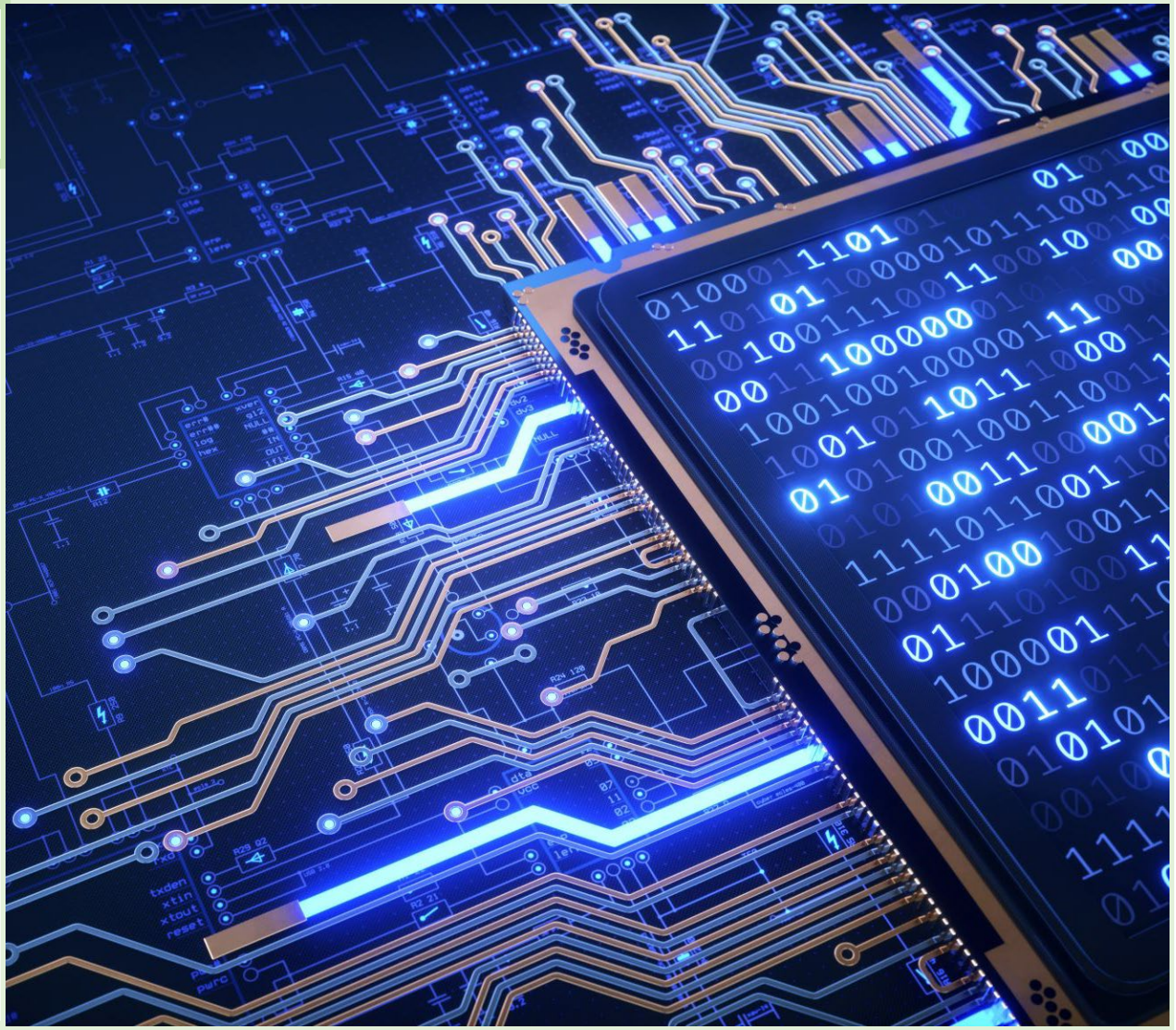
Model Design Matrix

Recall the *matrix/vector* form of the regression equation:

$$Y = \alpha + \beta X + \epsilon$$

Using linear algebra, i.e. working with matrices and vectors, we can use the matrix/vector form to calculate all of the predicted values at once.

The key is converting sex to *binary*.



Model Design Matrix

Keep only the numbers and add an 'intercept' column.

	intercept	n_sex	body_mass_g
1	1	1	3750
2	1	0	3800
3	1	0	3250
5	1	0	3450
6	1	1	3650
7	1	0	3625

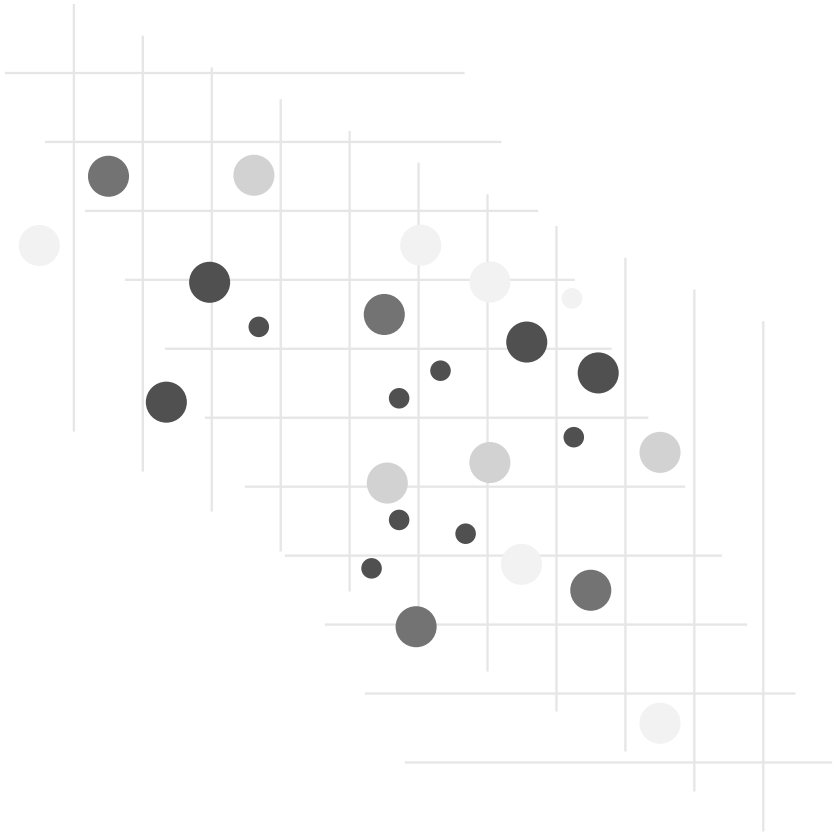
And remember the regression equation:
$\mathbf{Y} = \alpha + \beta\mathbf{X} + \epsilon$

	alpha	beta	y
1	1	1	3750
2	1	0	3800
3	1	0	3250
5	1	0	3450
6	1	1	3650
7	1	0	3625

Model Design Matrix and Dummy Variables

Finally…..

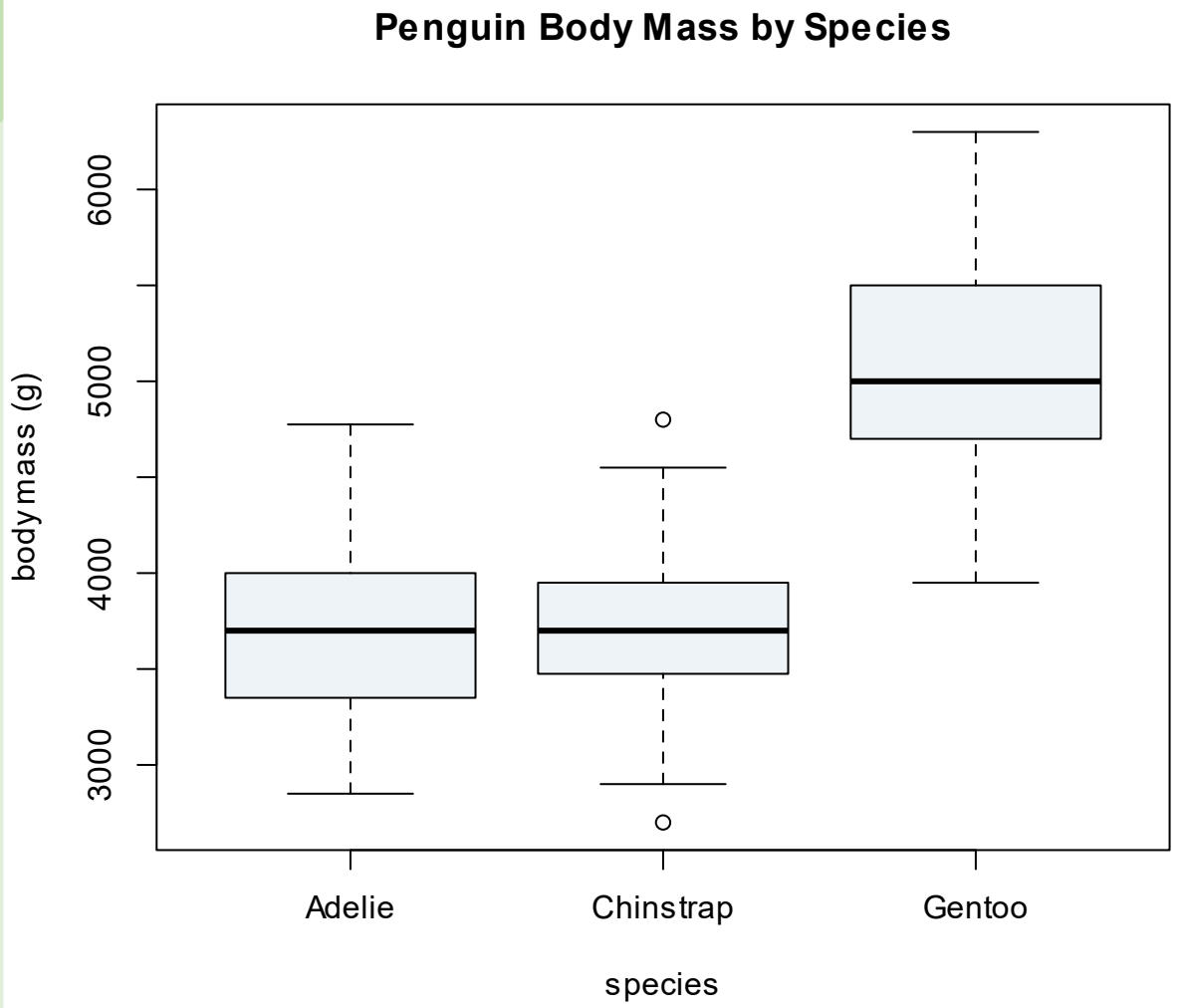
Now we can do matrix-vector multiplication using a vector of the model coefficients and the matrix/vector form to multiply the model matrix directly!



Factor Levels

What if a factor variable has more than 2 levels?

- We could use 0 and 1 to represent a two-level factor.
- Could we use 0, 1, 2?
- Categorical scale: “interval” between levels is not consistent.
- Is the “distance” between Adelie and Gentoo twich the the “distance” between Adelie and Chinstrap?



Factor Levels: Numeric

For n-level factors, we have to create n-1 dummy variables.

- **Numeric coding can take on values 0 to n-1**
- Each dummy variable can only take on values of 0 or 1.
- When the factor level is the **base case**, all dummy variables have value 0.

```
penguins$n_sex =
  as.numeric(penguins$sex) - 1

penguins$n_species =
  as.numeric(penguins$species) - 1
```

Numeric Coding

	sex	n_sex	species	n_species	body_mass_g
1	female	0	Adelie	0	3150
2	female	0	Gentoo	2	4750
3	male	1	Adelie	0	3800
4	female	0	Gentoo	2	4625
5	male	1	Adelie	0	4475
6	male	1	Chinstrap	1	3900
7	male	1	Adelie	0	4600
8	female	0	Gentoo	2	4600

Factor Levels: Dummy Variables

For n-level factors, we have to create n-1 dummy variables.

- Numeric coding can take on values 0 to n-1
- **Each dummy variable can only take on values of 0 or 1.**
- When the factor level is the **base case**, all dummy variables have value 0.

```
penguins$n_sex =
  as.numeric(penguins$sex) - 1

penguins$n_species =
  as.numeric(penguins$species) - 1
```

Dummy Variables

	sex	n_sex	species	Sp_gen	Sp_chin
1	female	0	Adelie	0	0
2	female	0	Gentoo	1	0
3	male	1	Adelie	0	0
4	female	0	Gentoo	1	0
5	male	1	Adelie	0	0
6	male	1	Chinstrap	0	1
7	male	1	Adelie	0	0
8	female	0	Gentoo	1	0

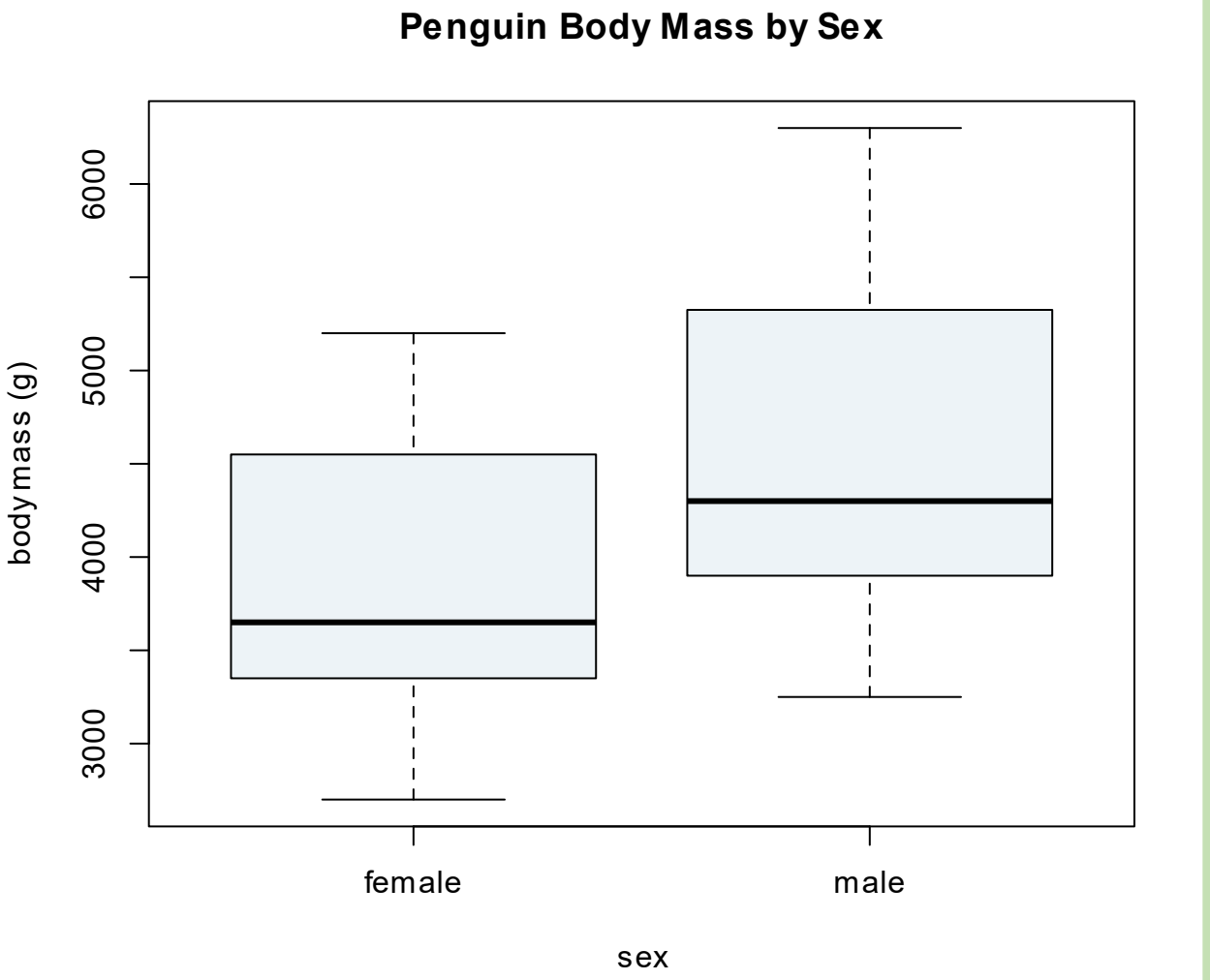
Factor Levels

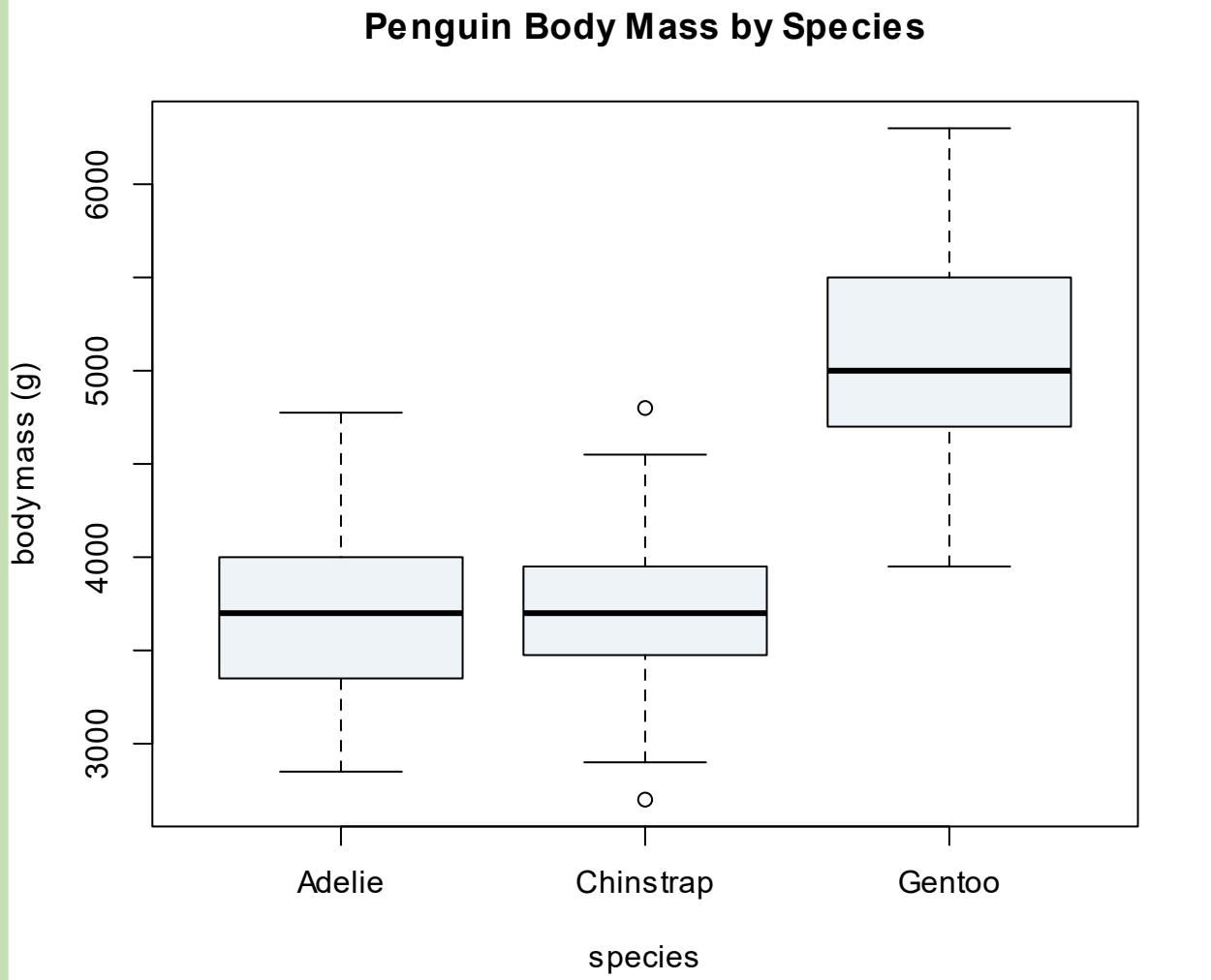
What are the *base cases*?

- Base species =?
- Base sex = ?

	sex	n_sex	species	sp_species	body_mass_g
1	female	0	Adelie	0	3150
2	female	0	Gentoo	2	4750
3	male	1	Adelie	0	3800
4	female	0	Gentoo	2	4625
5	male	1	Adelie	0	4475
6	male	1	Adelie	0	3900
7	male	1	Adelie	0	4600
8	female	0	Gentoo	2	4600

Factor Levels





Dummy Variables and Model Coefficients

How should we interpret the model coefficients for dummy variables?

- What does the intercept mean?
- What is the base case?
- What does the slope coefficient mean?

Build a Model!

```
fit_sex = lm(
  body_mass_g ~ sex,
  data = penguins)

summary(fit_sex)
```

```
## Call:
## lm(formula = body_mass_g ~ sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1295.7  -595.7  -237.3   737.7  1754.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3862.27      56.83  67.963  < 2e-16 ***
## sexmale       683.41      80.01   8.542  4.9e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 730 on 331 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.1806, Adjusted R-squared:  0.1781
## F-statistic: 72.96 on 1 and 331 DF,  p-value: 4.897e-16
```

Dummy Variables and Model Coefficients

The coefficients for dummy variables are showin in the model coefficient table:

```
fit_species = lm(body_mass_g ~ species, data =
penguins)
summary(fit_species)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3700.662	37.619	98.371	0.000
speciesChinstrap	32.426	67.512	0.480	0.631
speciesGentoo	1375.354	56.148	24.495	0.000

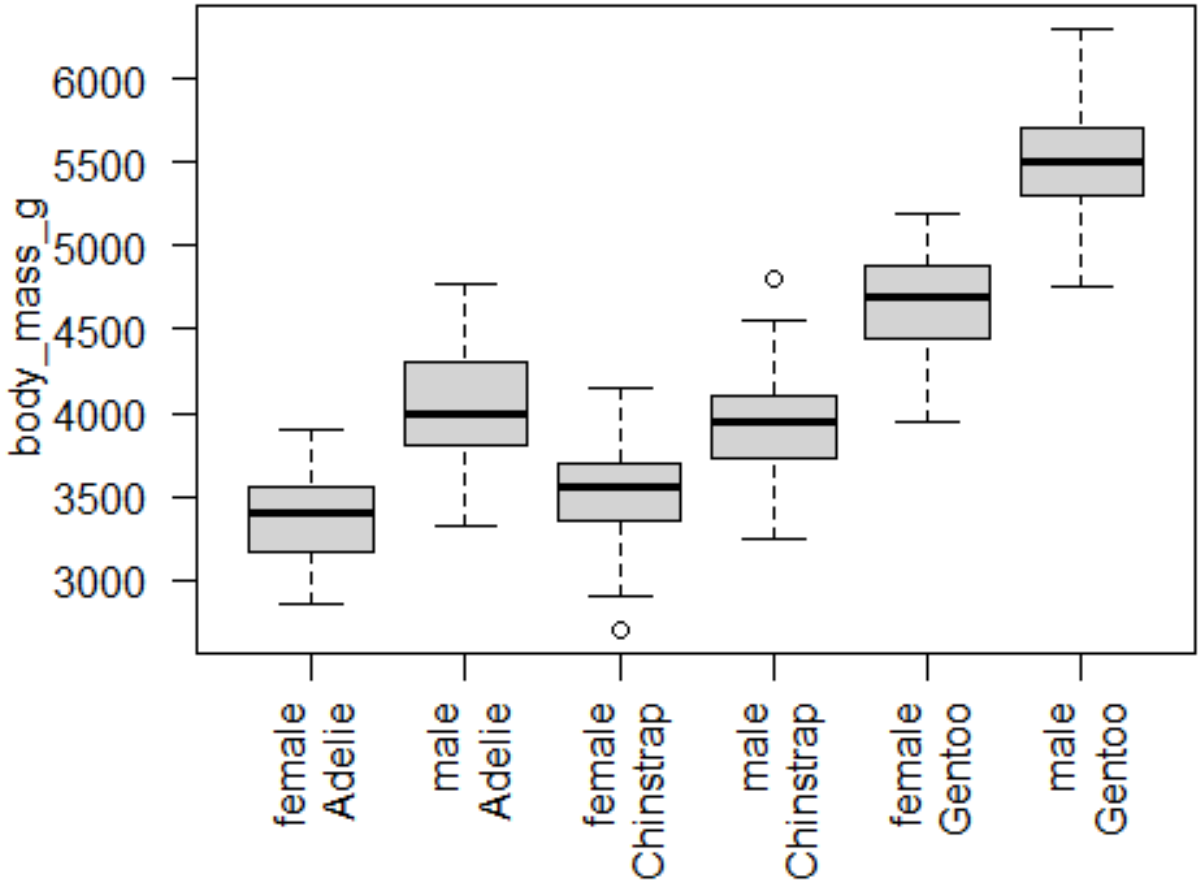
Dummy Variables and ANOVA

Since the dummy variables all *belong* to a single predictor variable, they collapse to a single line in the ANOVA table

```
anova(fit_species)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
species	2	146864214	73432107.1	343.6263	0
Residuals	339	72443483	213697.6	NA	NA

Interactions



Interactions

Consider two models:

Model 1 - body mass predicted by sex and species Model 2 - body mass predicted by the *interaction* between sex and species

In R:

```
fit_1 = lm(body_mass_g ~ sex + species, data =
penguins)
fit_2 = lm(body_mass_g ~ sex * species, data =
penguins)
```

Interactions: Model 1

What does model 1 propose?

1. A *species* effect: each species has a has a slope that defines the difference between the base case and the species.
2. A *sex* effect: There is a difference between the base case (female) and the male sex.

Interactions: Model 1

How are species and sex effects related?

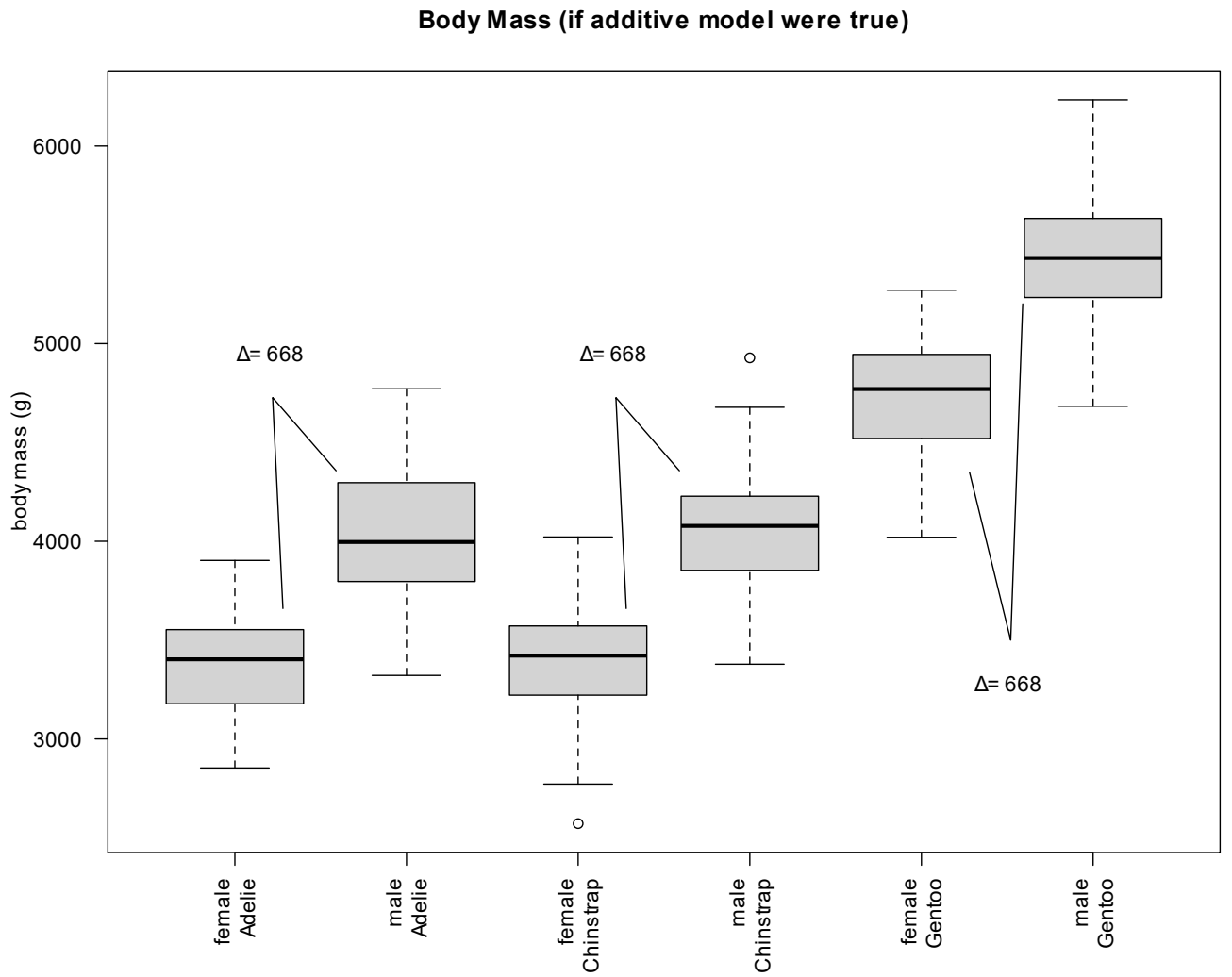
- The *species* effect is the same within a sex.
- The *sex* effect is the same within a species.
- Male penguins are always 668 grams heavier than females, regardless of species.
- Gentoo penguins are always 1378 grams heavier than Adelie penguins, regardless of sex. (male Gentoo weigh 1378 more than male Adelie)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3372.387	31.427	107.308	0.000
sexmale	667.555	34.704	19.236	0.000
speciesChinstrap	26.924	46.483	0.579	0.563
speciesGentoo	1377.858	39.104	35.236	0.000

Interactions: Model 1

Does the model 1 structure make sense? We can assess graphically, grouped by species:

- If males are always 668 grams heavier the boxplots would look like this:



Interactions: Model 2

What does model 2 propose?

Main Effects

1. A *species* effect: each species has a has a slope that defines the difference between the base case and the species.
2. A *sex* effect: There is a difference between the base case (female) and the male sex.

Interaction Effects

- The *species* and *sex* effects might not be independent:
  - The difference between sexes can be different for each species.
  - The differences among species can be different for each sex.

Interactions: Model 2

The model now has *interaction* slope coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3368.836	36.212	93.030	0.000
sexmale	674.658	51.212	13.174	0.000
speciesChinstrap	158.370	64.240	2.465	0.014
speciesGentoo	1310.906	54.422	24.088	0.000
sexmale:speciesChinstrap	-262.893	90.849	-2.894	0.004
sexmale:speciesGentoo	130.437	76.436	1.706	0.089

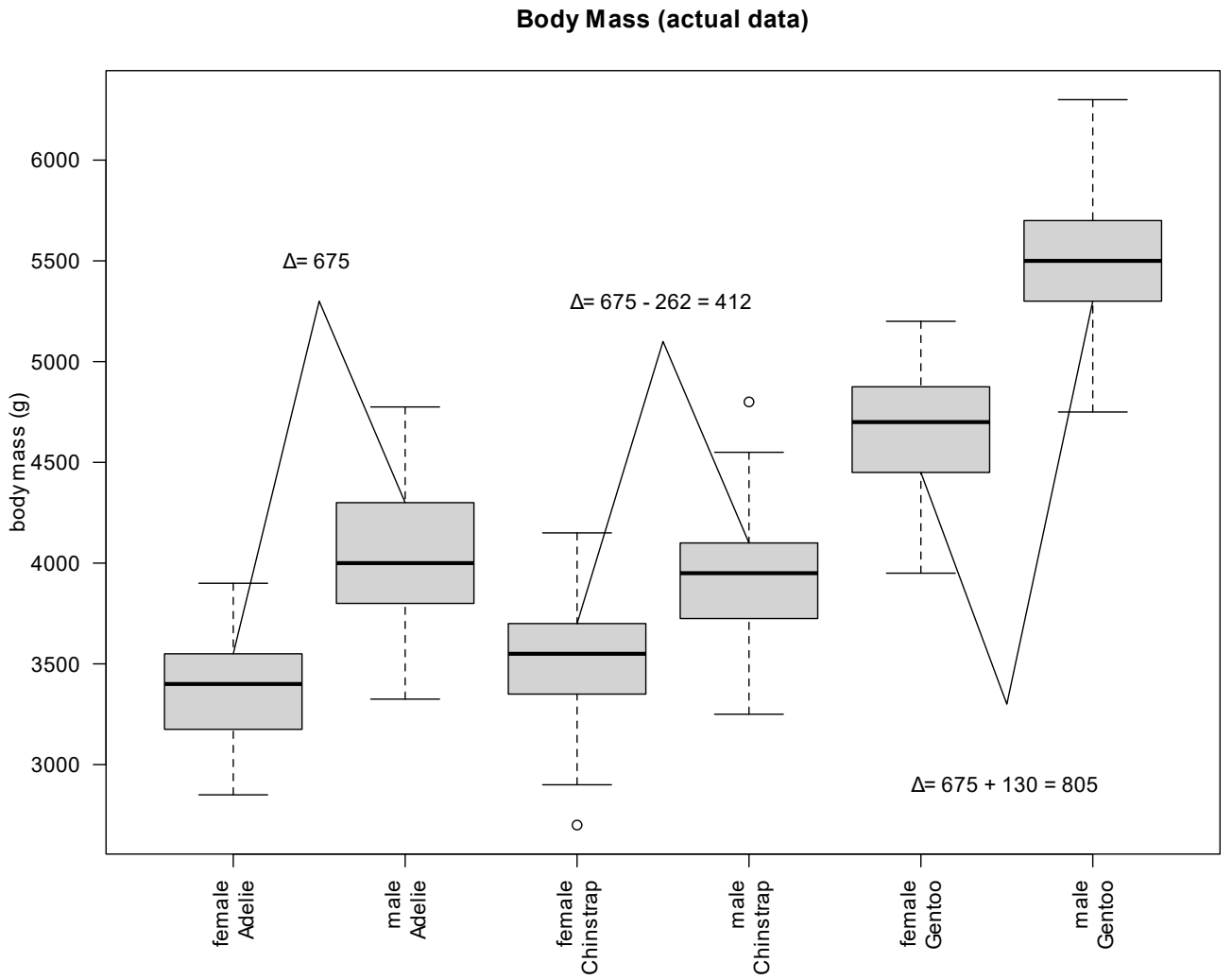
Interpreting the interaction coefficients

- The difference betweenmale and female Adelie penguins is 675 grams
- male/Gentoo interaction is positive: The difference between sexes is larger for Gentoo penguins
- male/Chinstrap interaction is negative: The difference between sexes is smaller for Chinstrap penguins

Model 2

We can see the interactions graphically:

Adelie male/female difference is 675 g



Model 2

We can verify our estimates numerically:

Adelie male/female difference is 675 g

Chinstrap difference is 412

Gentoo difference is 805

sex	species	body_mass_g
female	Adelie	3368.836
male	Adelie	4043.493
female	Chinstrap	3527.206
male	Chinstrap	3938.971
female	Gentoo	4679.741
male	Gentoo	5484.836

Interactions: ANOVA Tables

Compare the ANOVA tables:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	38878897	38878896.9	387.8555	0
species	2	143401584	71700792.0	715.2863	0
Residuals	329	32979185	100240.7	NA	NA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	38878897	38878896.91	406.14	0
species	2	143401584	71700791.99	749.02	0
sex:species	2	1676557	838278.37	8.76	0
Residuals	327	31302628	95726.69	NA	NA

Ginkgo Data Exploration

Instructions on GitHub

Interactive Model Matrix

```
head(model.matrix(fit_2))
##   (Intercept) sexmale speciesChinstrap speciesGentoo sexmale:speciesChinstrap
## 1           1       1                0             0                         0
## 2           1       0                0             0                         0
## 3           1       0                0             0                         0
## 5           1       0                0             0                         0
## 6           1       1                0             0                         0
## 7           1       0                0             0                         0
##   sexmale:speciesGentoo
## 1                     0
## 2                     0
## 3                     0
## 5                     0
## 6                     0
## 7                     0
```

Interactions

You can think of interactions in may ways, inclulding:

- Inhibiting
- Facilitating
- Synergistic
- Adjusting

Interactions are easiest to understand with factors, but they also work with continuous predictors.

Statistical Power

What's in This Section?

Slide Show	Take-Home Concepts
• Alpha: significance level, specified in advance • Beta: false negative rate, estimated after data collection • Critical value: test statistic must be more extreme than this value to reject null.	• Errors: false negatives and false positives. • Alpha and Beta • How to control the false negative rate

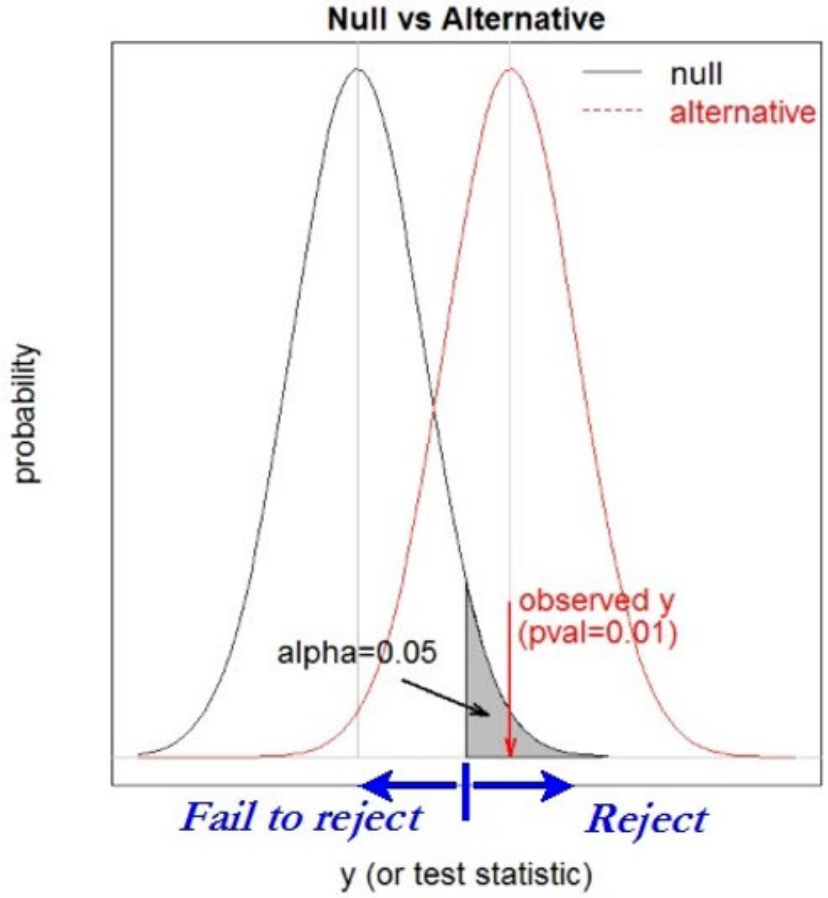
Hypothesis Testing

Hypothesis Testing Concepts

Neyman-Pearson decision framework

- *Reject* the null hypothesis if the *p*-value is less than a critical value (*alpha*), by convention usually ≤ 0.05
- *Fail to reject* the null hypothesis if the *p*-value is greater than *alpha* (i.e., there is insufficient evidence to disprove the null)

Remember, this applies to any probability distribution



False Positives: alpha

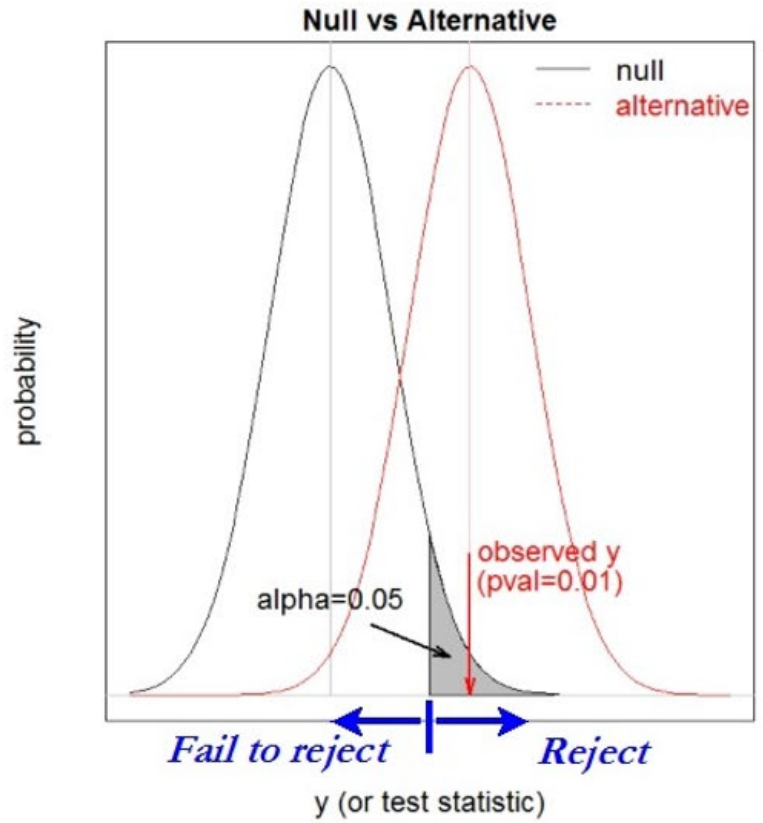
- alpha is the likelihood that we falsely reject a true null hypothesis.
- this is our p-value cutoff that we specify ahead of time.

Hypothesis Testing Concepts

Neyman-Pearson decision framework

- *Reject* the null hypothesis if the *p*-value is less than a critical value (*alpha*), by convention usually ≤ 0.05
- *Fail to reject* the null hypothesis if the *p*-value is greater than *alpha* (i.e., there is insufficient evidence to disprove the null)

Remember, this applies to any probability distribution



False Negatives: beta

Beta is the type II error rate: failing to reject a false null hypothesis.

- We select a p-value cutoff ahead of time: alpha
- The false negative rate depends on our choice of alpha and the data.
- We cannot know beta until after we have collected data :(

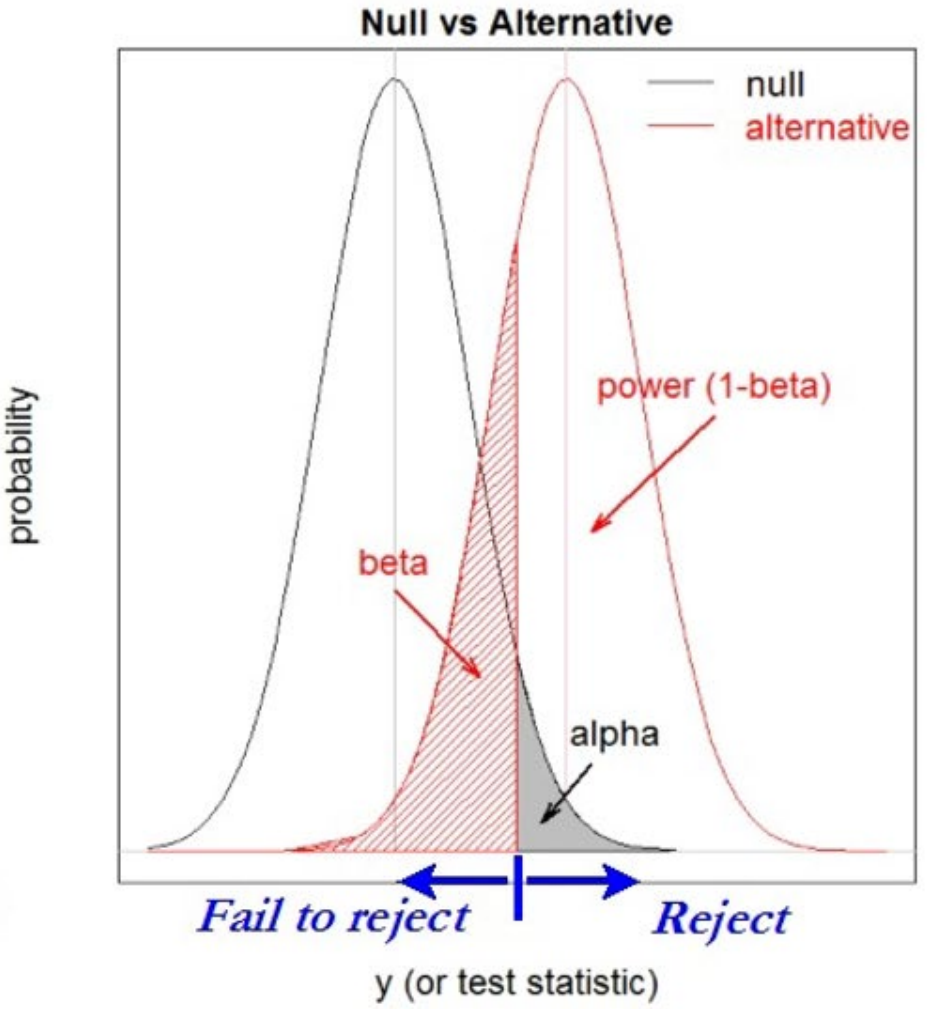
We can think of a false negative as when an observation *belongs* to the alternative hypothesis distribution, but falls outside of the *rejection region*.

- It belongs to the alternative distribution because the null hypothesis is false.
- But… it looks like it should belong to the null hypothesis because it is outside of the rejection region.

Alpha and Beta

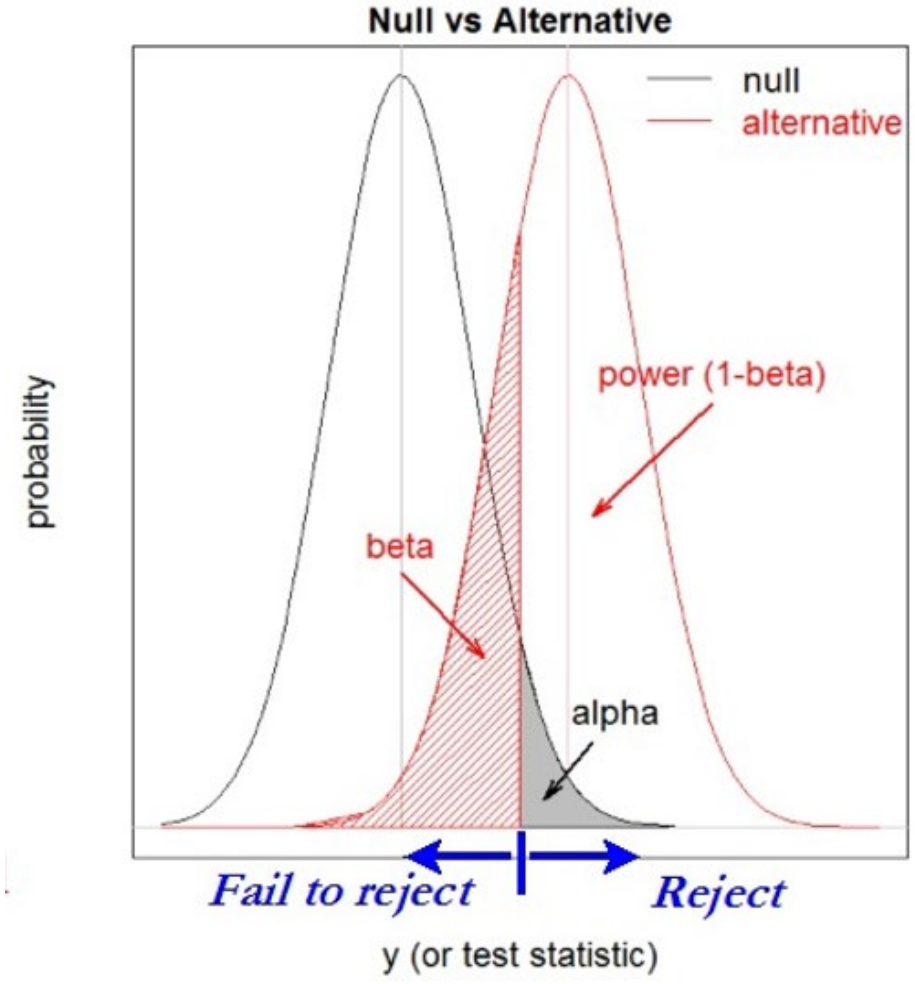
- *alpha* = probability of wrongly rejecting the null hypothesis (Type I error)
- *beta* = probability of wrongly accepting the null hypothesis (Type II error)
- *power* = probability of correctly rejecting the null hypothesis

alpha is under the null; *beta* and *power* are under the alternative



Power Analysis

- Statistical Power: the probability that we correctly reject a false null hypothesis.
- Statistical power is 1 – beta
- We can't know our statistical power until after we collect data.



Factors that influence statistical power

- Sampling error, sample size
- Population variability
- Effect size
- Our choice of alpha
- You cannot simultaneously decrease the false positive rate and increase statistical power!

Effect of Alpha

The choice of alpha affects our statistical power:

- Small alpha makes the *rejection region* smaller:
  - We have to observe a more extreme value to be in the *rejection region*.
  - Less overlap between the rejection region and the alternative distribution.
  - More overlap between the alternative distribution and the fail-to-reject region
- Large alpha moves the *rejection region* closer to the center of the null distribution.
  - We're more likely to observe a value within the rejection region by chance.
  - More overlap between the rejection region and the alternative distribution.
  - Less overlap between the alternative distribution and the fail-to-reject region

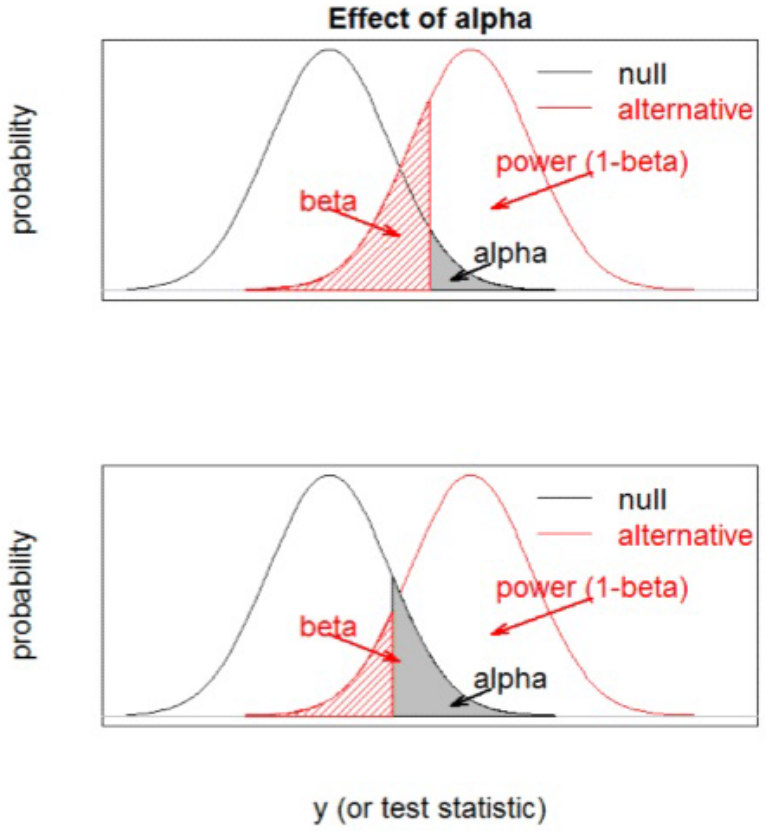
Effect of Alpha

There's a tradeoff between power and false positive rate.

If we're willing to accept more false positives, we have more power.

Effect of alpha?

- Increasing alpha, increases power, all other things being equal



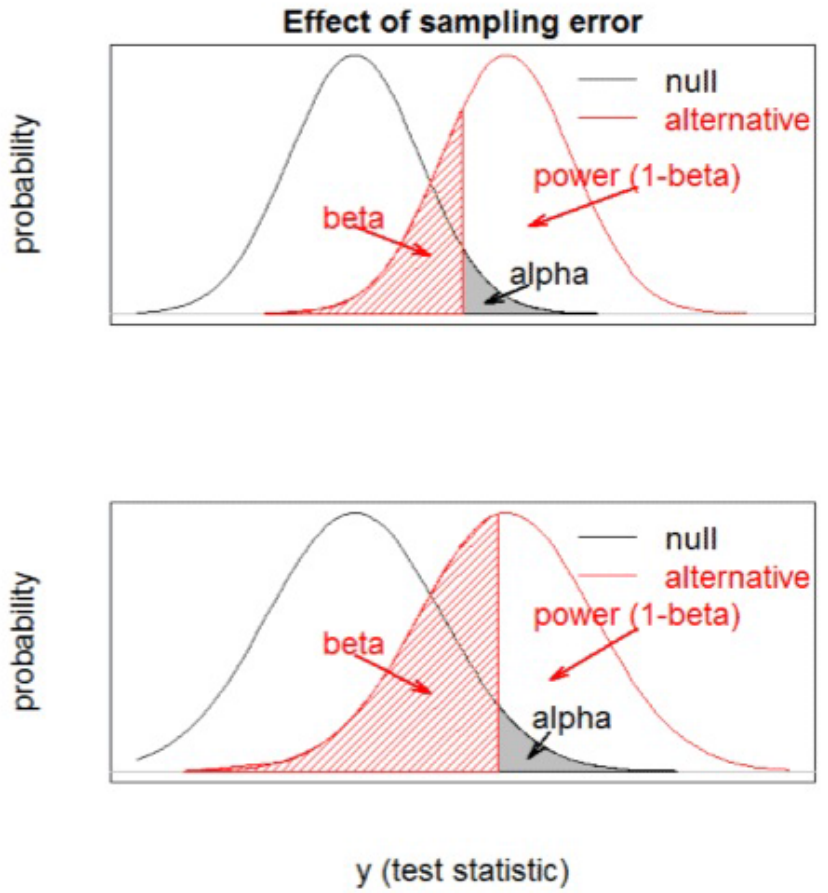
Effect of Population Standard Deviation

Smaller population standard deviation makes the sampling distribution narrower.

- Smaller overlap between null and alternative distributions.

Effect of sampling variability (standard error)?

- Increasing sampling variability, either by increasing the variance in the underlying distribution or decreasing sample size (both effect sampling precision), decreases power, all other things being equal

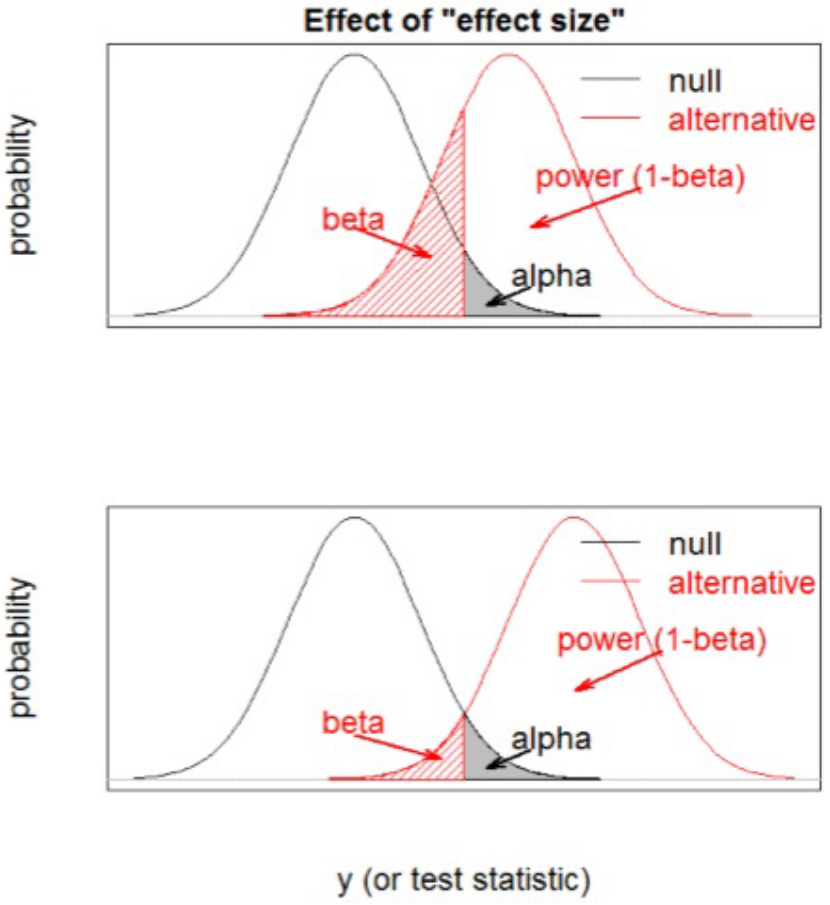


Effect of Effect Size

If the effect is larger, the null and alternative distributions are more separated.

Effect of effect size?

- Increasing the effect size, increases power, all other things being equal



Key Concepts

- Errors: false negatives and false positives.
- Alpha and Beta
- Tradeoff between false positive rate and statistical power.
- How to control the false negative rate