Statistical Hypothesis Testing Basics

Gladstone Institutes

Reuben Thomas & Michela Traglia
Associate Core Director @ Bioinformatics Core @ GIDB
01/18/2023

Leading the discussion today…

- Reuben Thomas – Associate Core Director
- Michela Traglia– Biostatistician

Days 2 and 3

- Very basic introduction to the concepts and terminology of hypothesis testing
- Some guidance on choosing tests in relatively simple situations
- Hands-on training on implementing statistical tests in R, requires some basic familiarity in working with R/Rstudio
- Two days: 1/18-1/19 @1PM for 2 hours
- Today: Mostly concepts and some practical implementation
- Tomorrow: Mostly hands-on plus some concepts
- Both days: Your specific problems

Poll: Why do we perform statistical hypothesis testing?

- It allows us to make claims claims that are reproducible and generalizable with limited resources

Poll: What hypothesis tests have you used?

Terms one commonly encounters in hypothesis testing

- Null hypothesis versus Alternative hypothesis
- P-values
- Two-sided test *versus* One-sided test
- Test statistic
- Sampling distribution
- Type I and Type II errors (Power)
- Multiple testing
- Assumptions of different tests
- Linear models
- ANOVA

Typical scenario

- Setting: I have generated data from very cool experiment that I hope would resolve a long standing question
- Problem: I don't know how to use my data to conclude in a convincing manner one way or other
- Possible solution: Pose the problem as a statistical association problem
  - Changing something has a consequence on something else of biological relevance
  - E.g.: Change dose of drug treatment and phenotype changes

Outline

- **Introduction to hypothesis testing**
- Define variables
- Choosing the right test
- Basic concepts in hypothesis testing
- Demo

Introduction to Hypothesis Testing

- We would like to make **generalizable claims about an entire target population** with data **from only a random subset** of this population.
- **Random sampling**, **appropriate experimental design** and **Central Limit Theorem** allows us to make generalizable claims
- Hypothesis testing rests on assuming the **skeptical point of view** and testing for deviations from this assumption – **Null** versus **Alternative Assumption**
- Equally relevant to Hypothesis Testing is the idea of measuring **the strength of association/effect size**

Outline

- Introduction to hypothesis testing
- **Define variables**
- Choosing the right test
- Basic concepts in hypothesis testing
- Hands-on

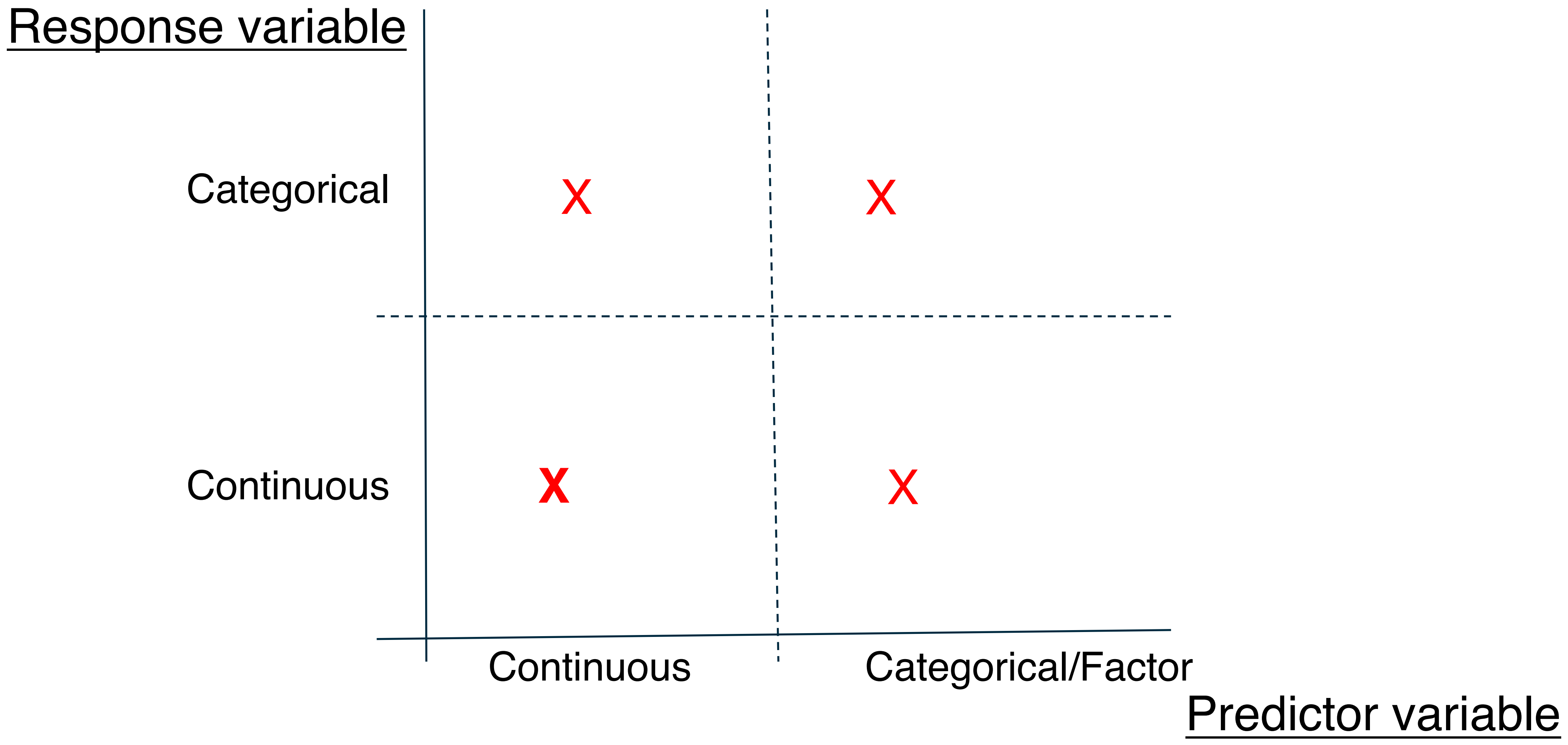
Variables

- Response: Gene expression, Chicken weight
- Predictor: Genotype, treatment, chicken feed
- Types: Categorical or Continuous
  - Categorical – genotype (mutant versus wild-type), disease vs normal
  - Continuous – age, dose of drug treatment

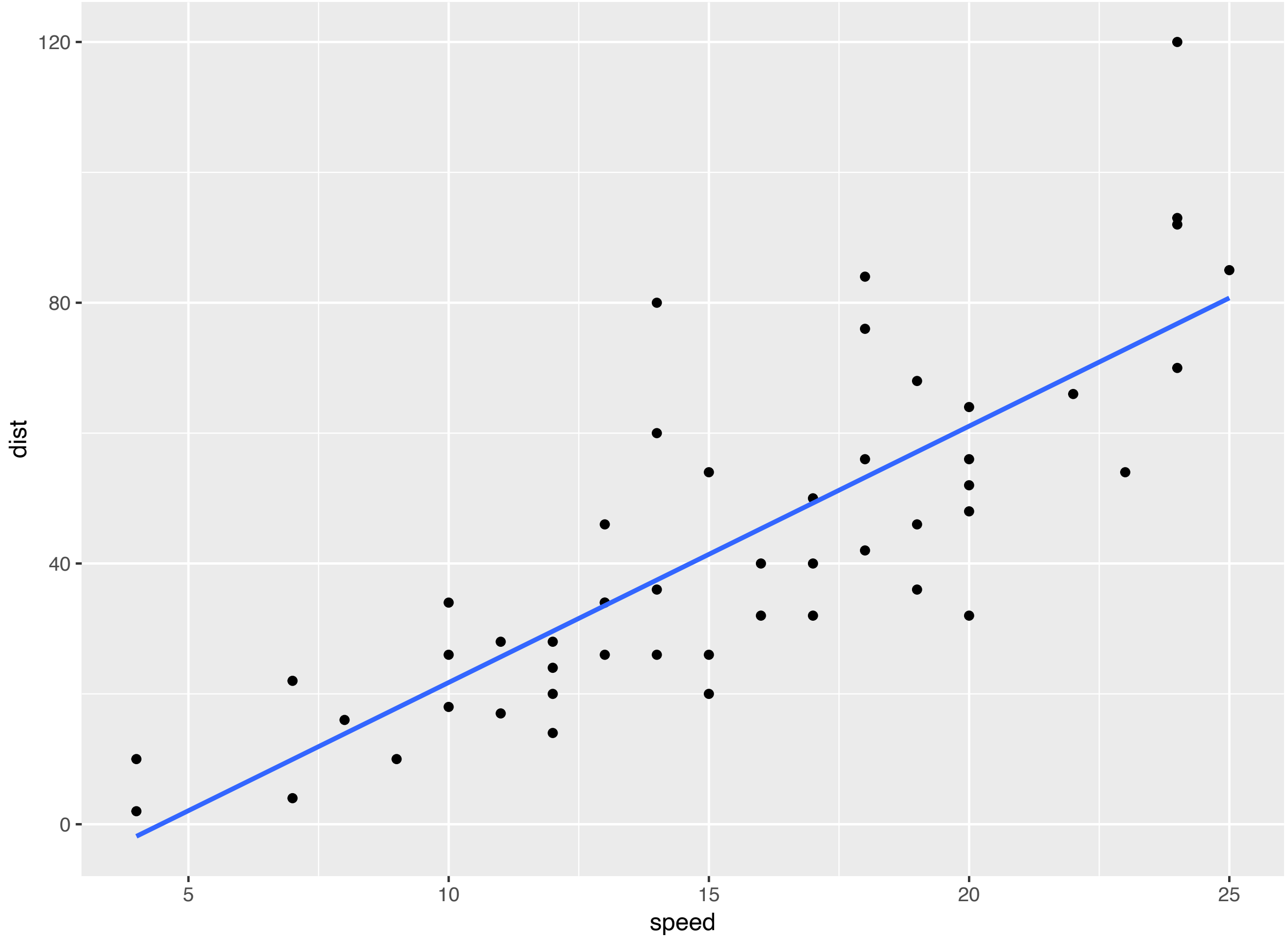
Outline

- Introduction to hypothesis testing
- Define variables
- **Choosing the right test**
- Basic concepts in hypothesis testing
- Hands-on

How do I choose which statistical test to use?



Response: Continuous
Predictor: Continuous



Linear regression

Parameter/effect size: slope

How do I choose which statistical test to use?

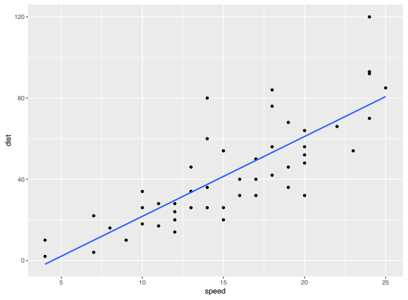
Response variable

Categorical

X

X

Continuous



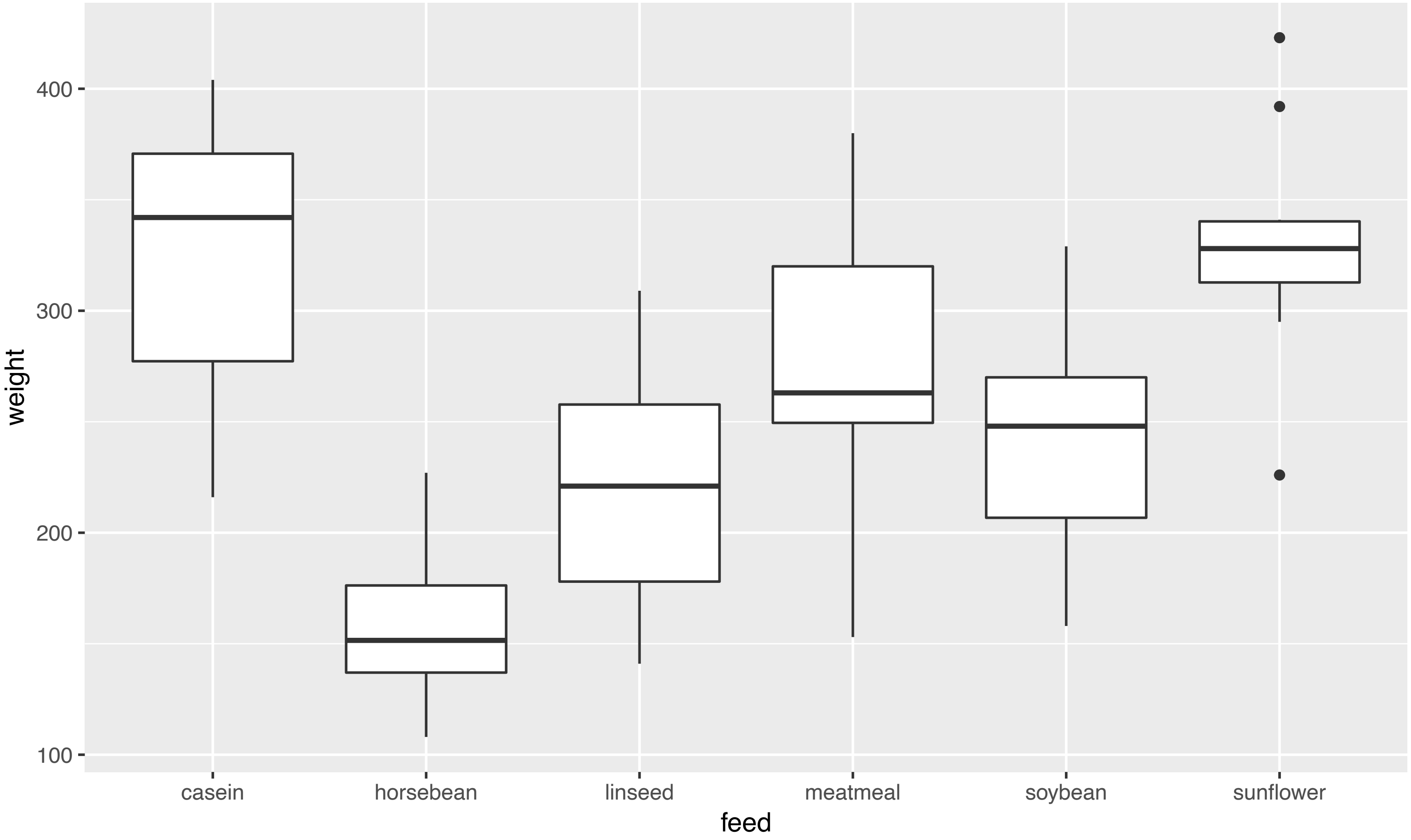
X

Continuous

Categorical

Predictor variable

Response: Continuous
Predictor: Categorical



T-tests, ANOVA

Parameter/effect size: difference of means

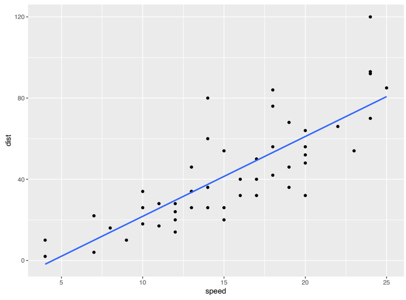
How do I choose which statistical test to use?

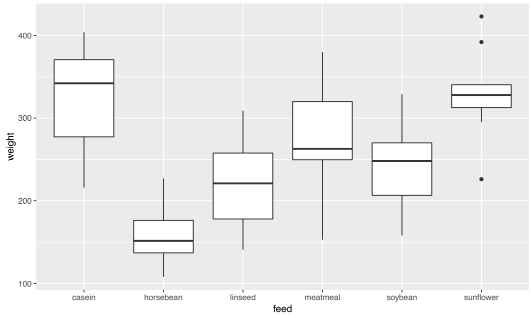
Response variable

Categorical

X X

Continuous

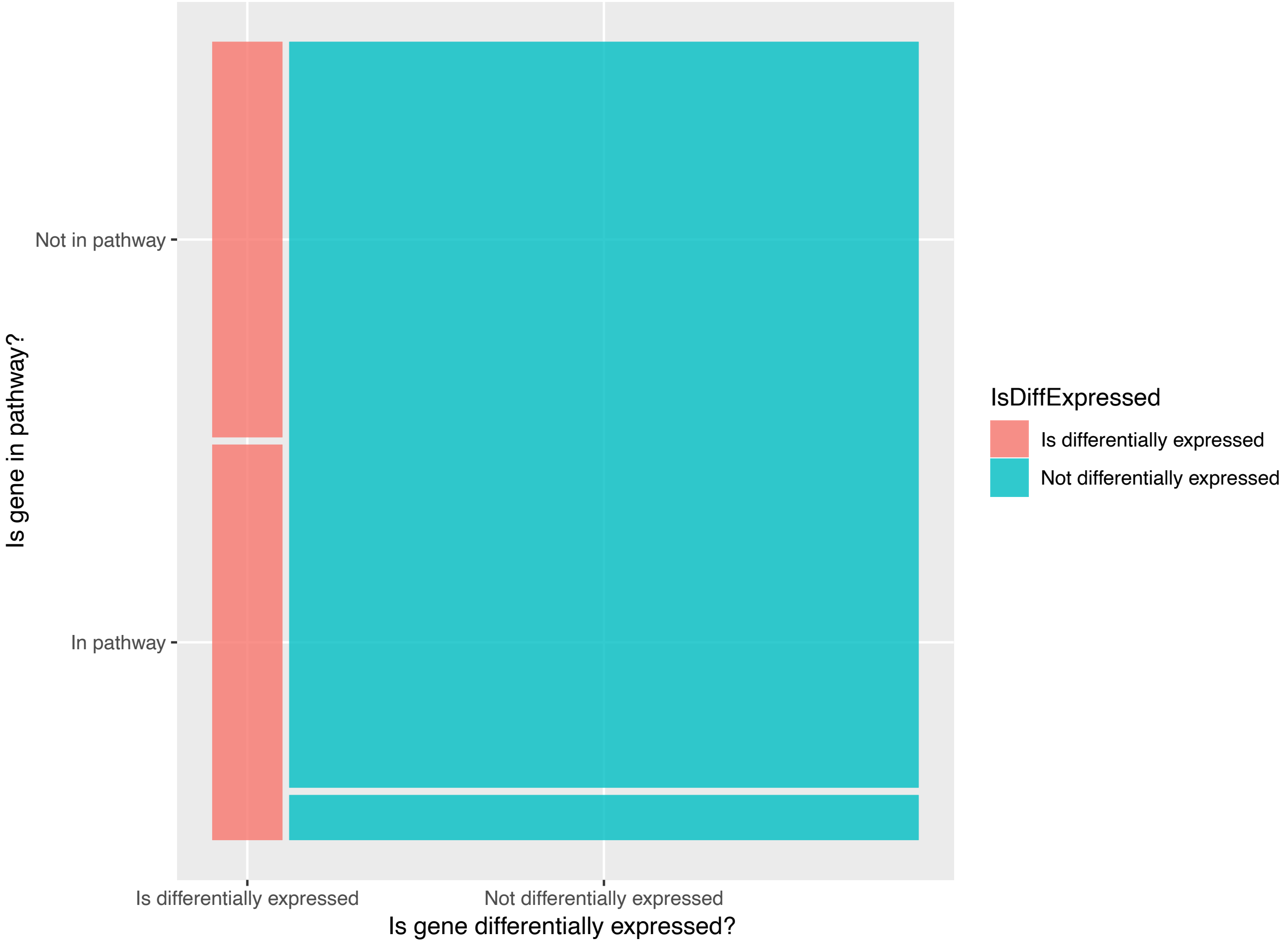




Continuous Categorical

Predictor variable

Response: Categorical
Predictor: Categorical



Fisher's test, Chiq-square test, 2x2 tables

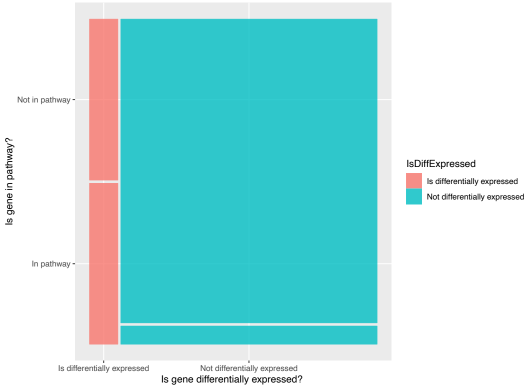
Parameter/effect size: odds ratio

How do I choose which statistical test to use?

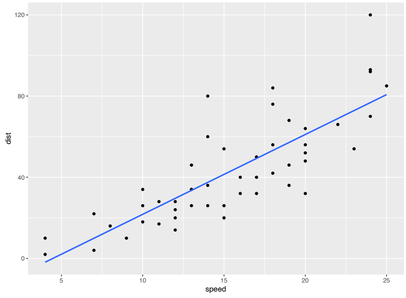
Response variable

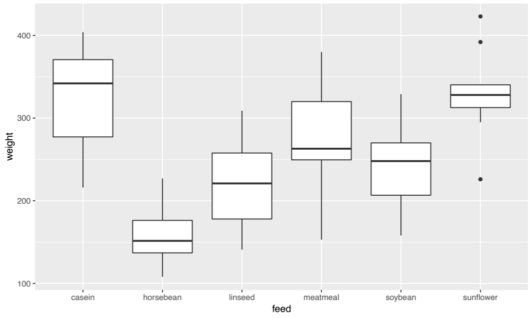
Categorical

X



Continuous



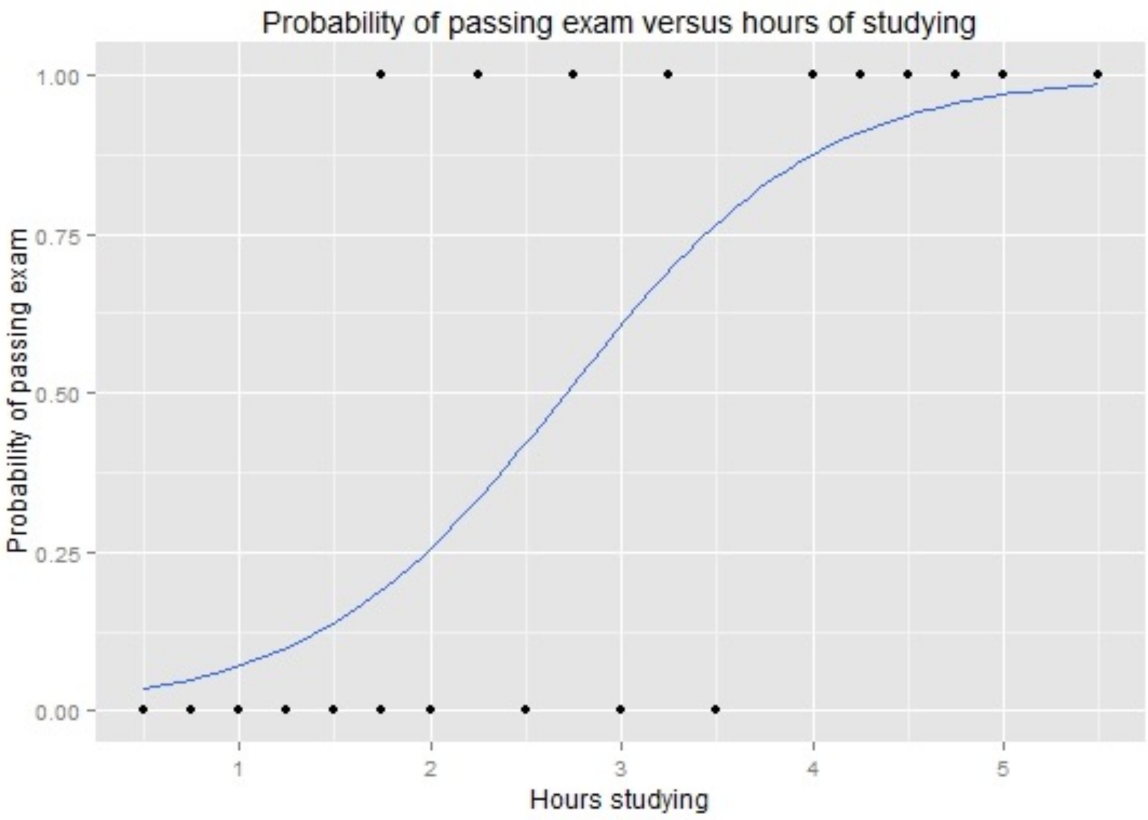


Continuous

Categorical

Predictor variable

Response:Categorical
Predictor: Continuous



Logistic regression

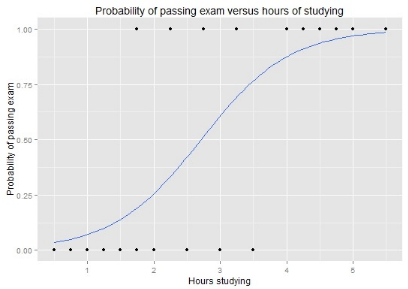
Parameter/effect size: odds ratio

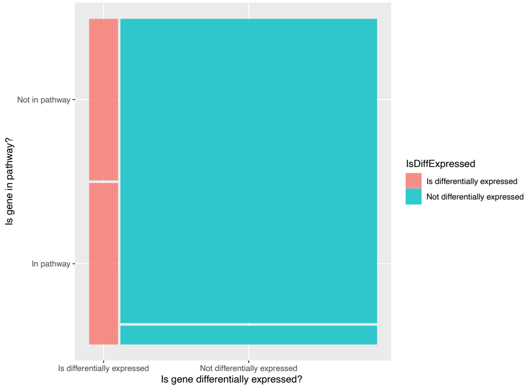
https://upload.wikimedia.org/wikipedia/commons/6/6d/Exam_pass_logistic_curve.jpeg

How do I choose which statistical test to use?

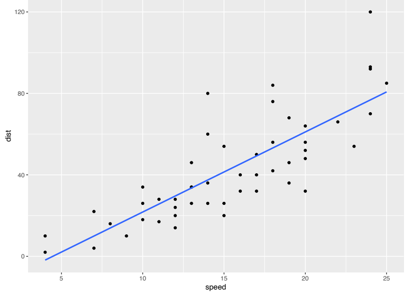
Response variable

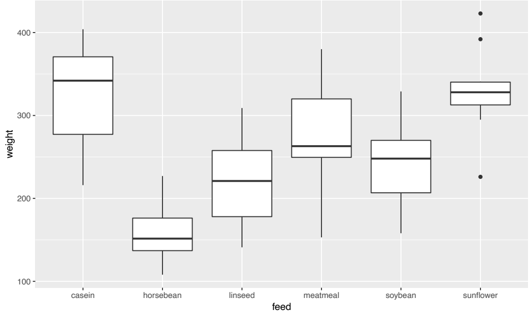
Categorical





Continuous





Continuous

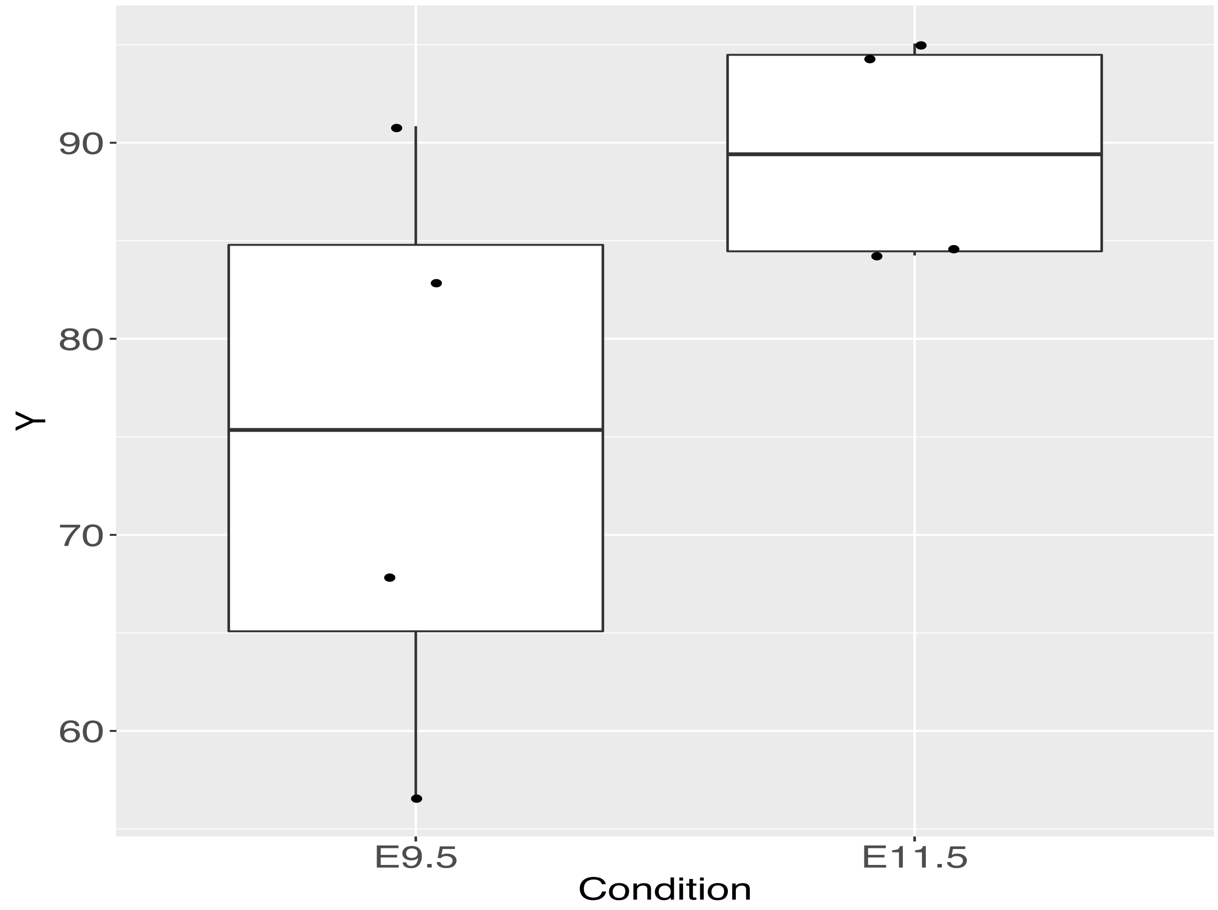
Categorical

Predictor variable

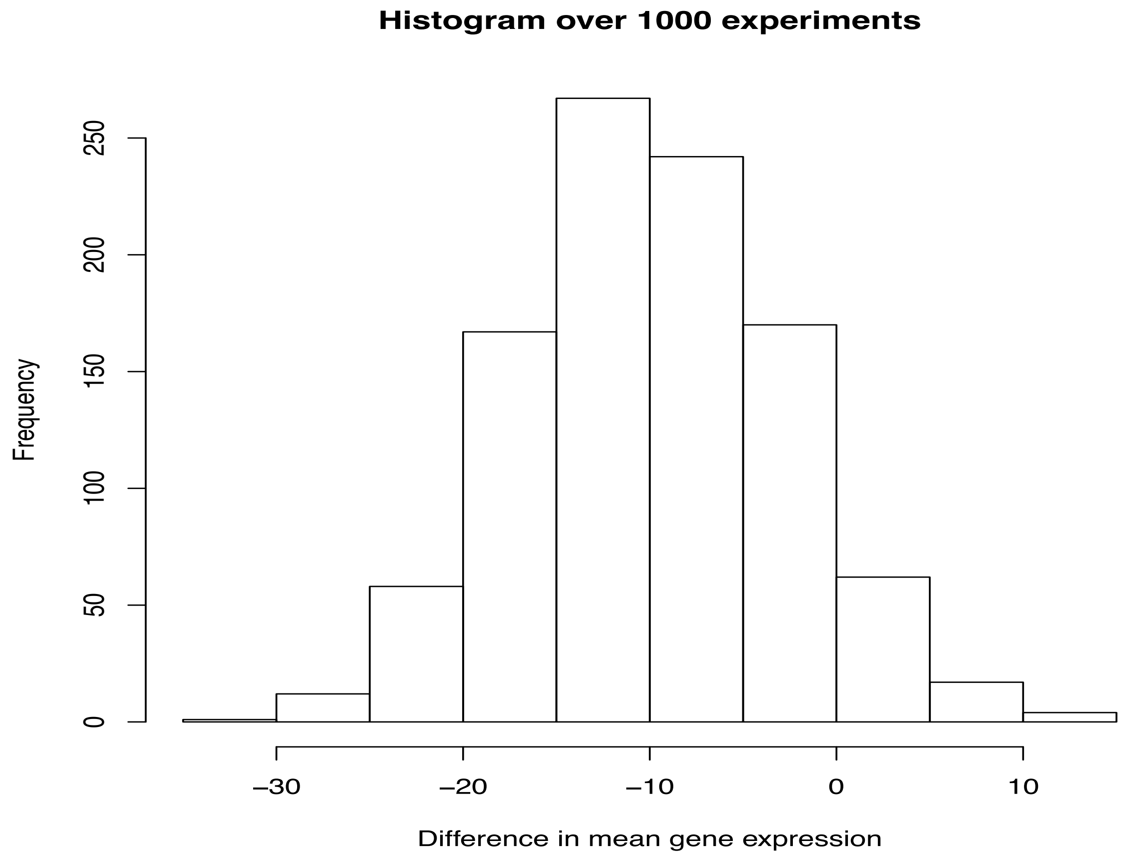
Outline

- Introduction to hypothesis testing
- Define variables
- Choosing the right test
- **Basic concepts in hypothesis testing**
- Hands-on

Is gene differentially expressed between the two developmental time-points?

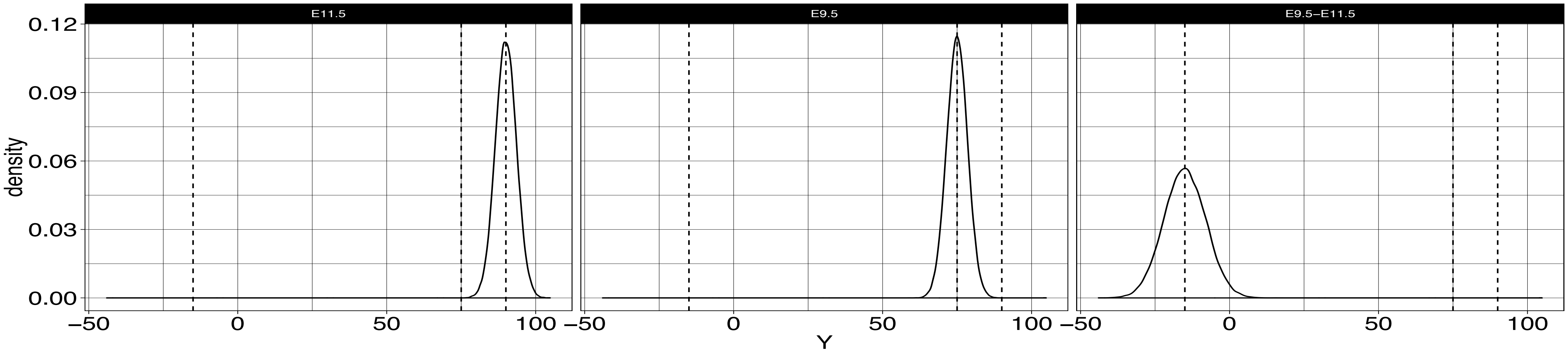


Convince a skeptic: Repeat this experiment 1000 times

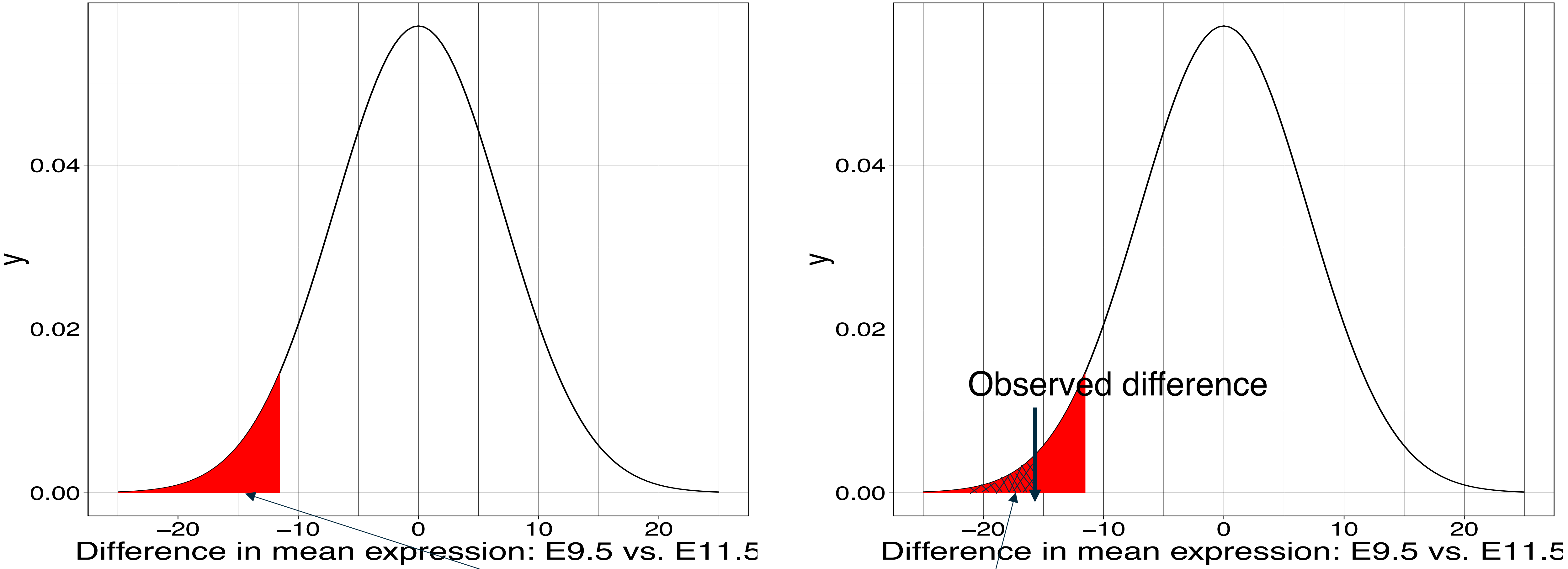


Central limit theorem allows us to estimate the variation of the location of the distribution

$E11.5 : \mathrm{N}ormal\left(90, \frac{7}{\sqrt{4}}\right)$ $E9.5 : \mathrm{N}ormal\left(75, \frac{7}{\sqrt{4}}\right)$ $E9.5 - E11.5 : \mathrm{N}ormal\left(75-90, \frac{7+7}{\sqrt{4}}\right)$



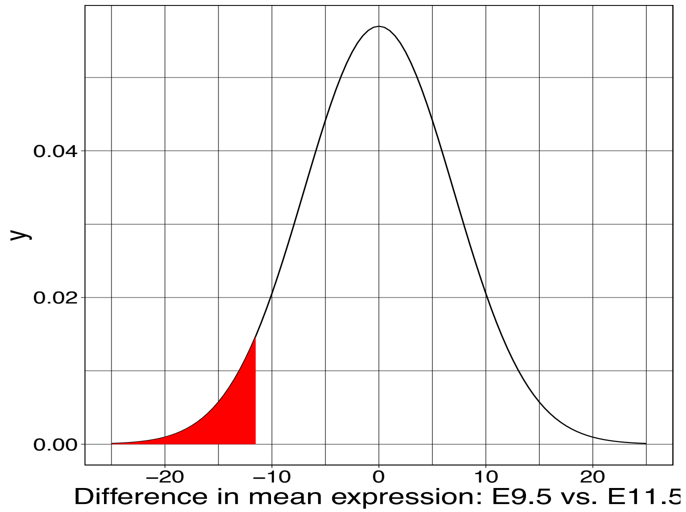
Theoretical distribution of difference in means under Null Hypothesis



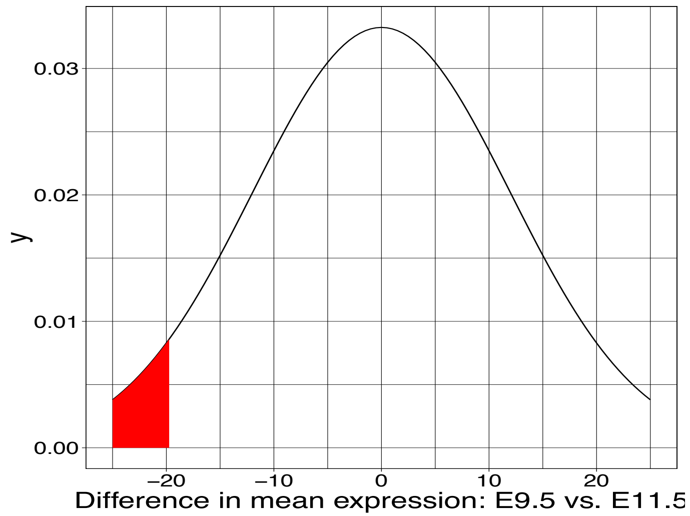
Type I error and **p-value**

Alter underlying variation

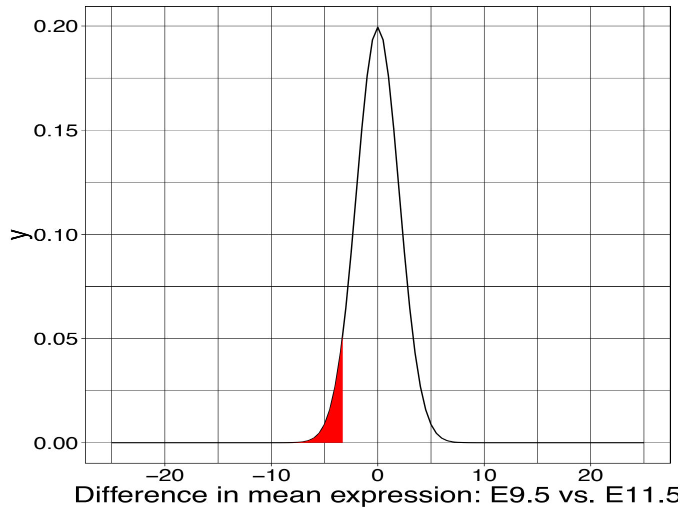
sd=14



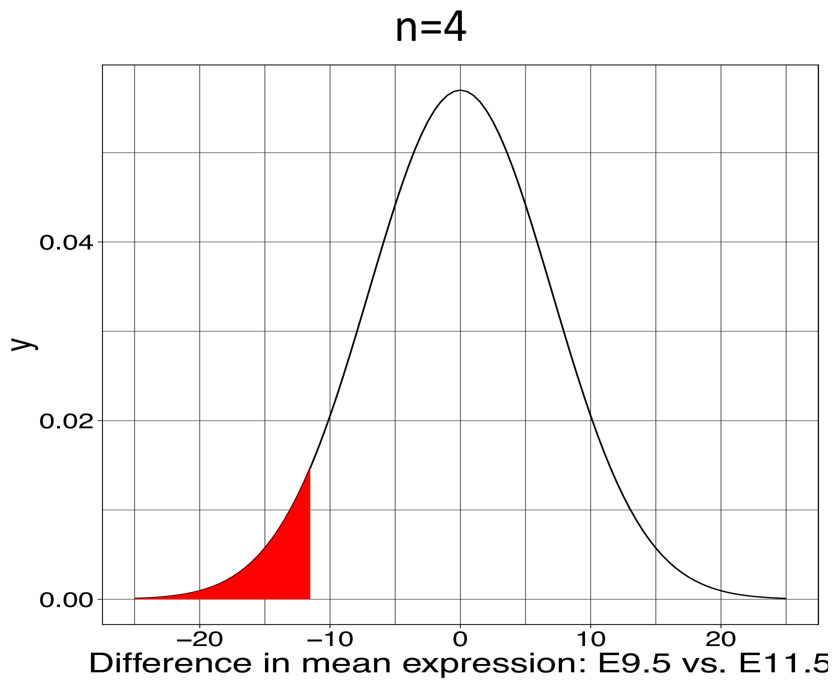
sd=24

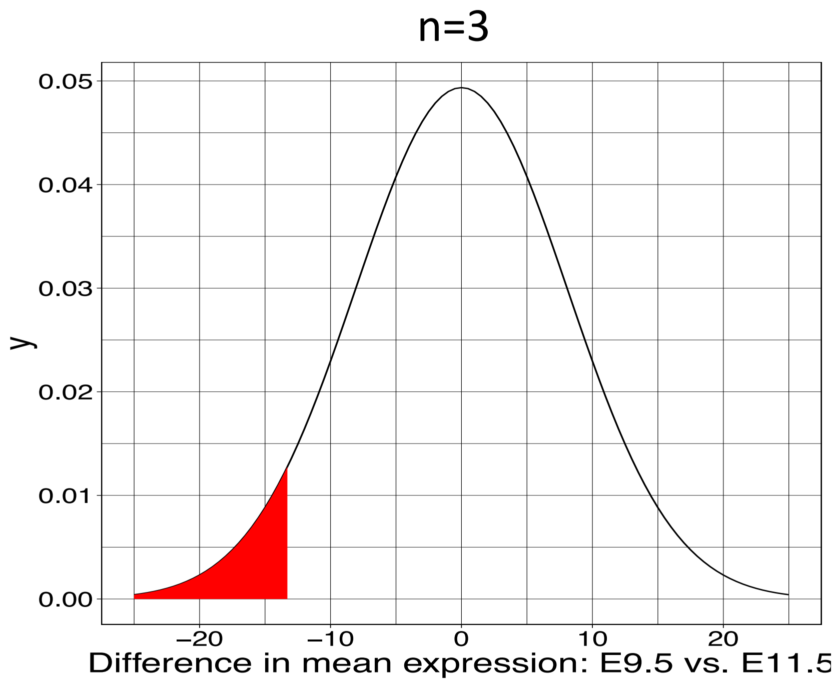


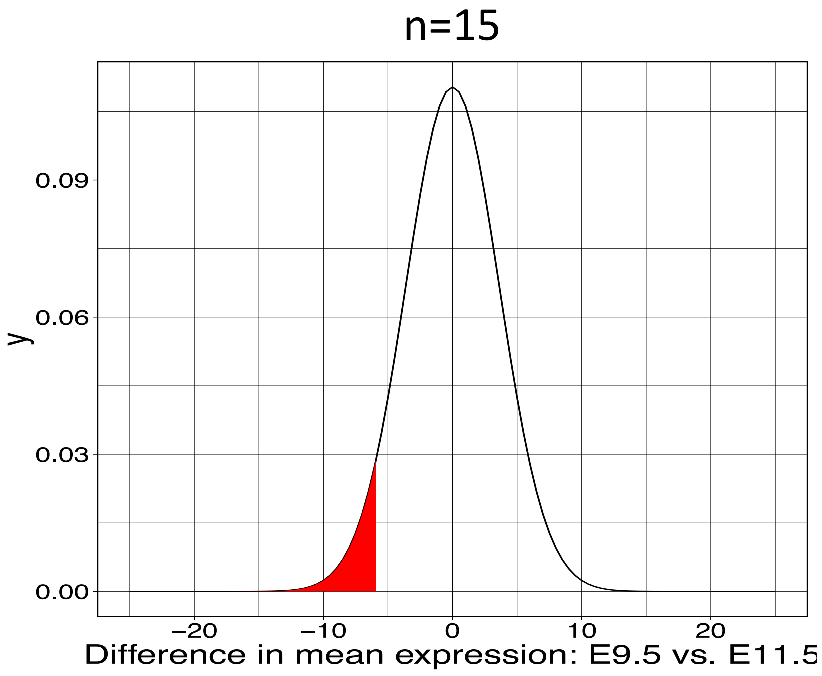
sd=4



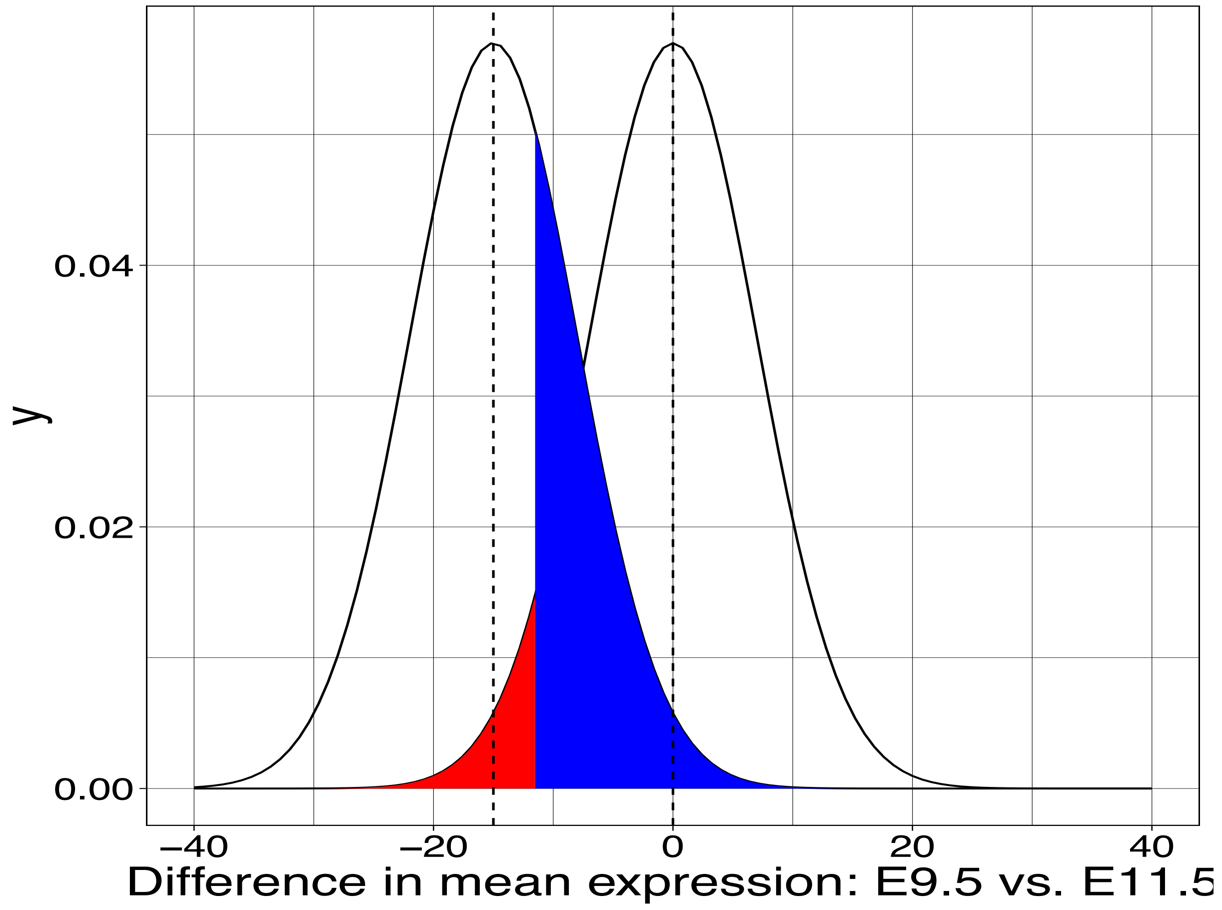
Alter the number of replicates







Power to detect a difference of means of -15



Type I and **Type II** error

You are willing to be mistaken that there is a true difference **Type I error** fraction of time you repeat this experiment

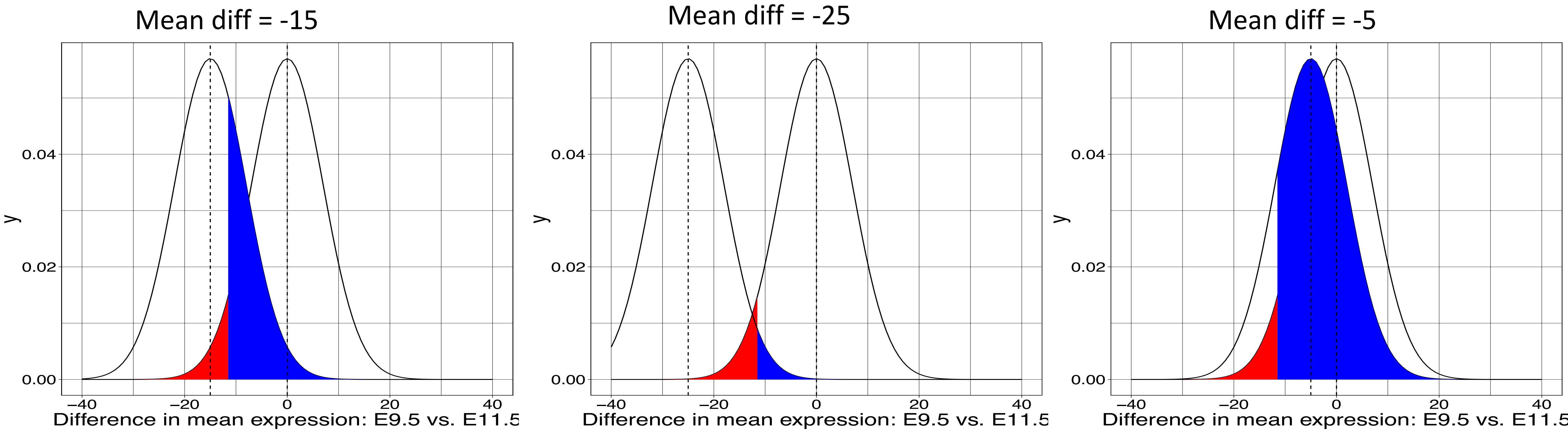
You are mistaken that there is no difference **Type II error** fraction of time you repeat this experiment

Power = 1 – **Type II error**

You correctly say that there is a difference **Power** fraction of time you repeat this experiment

Poll: What are the factors that affect Power or the fraction of time you claim that there is a real difference when there is actually a difference?

Power to detect varying levels of difference in mean differences



Type II error smaller for larger effect sizes

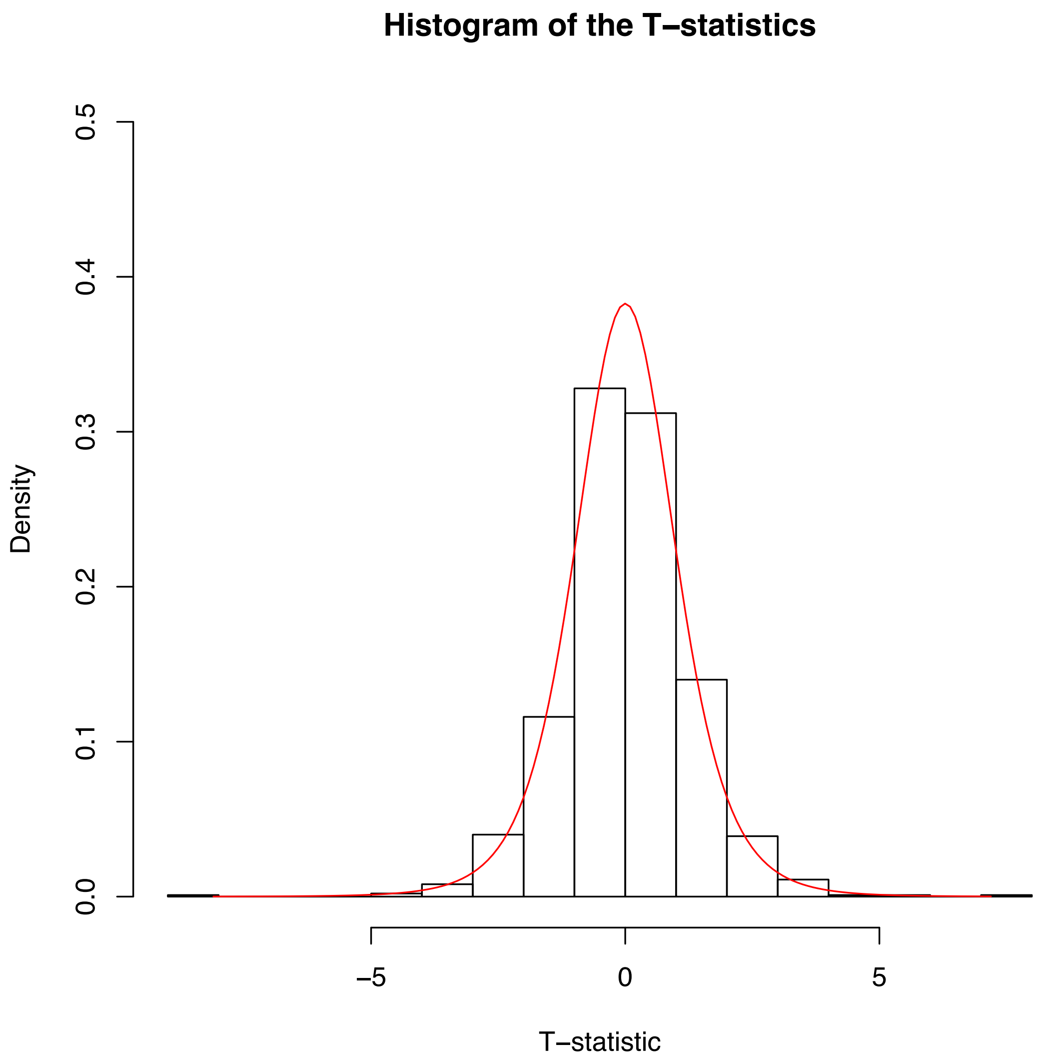
Larger effect sizes are easier to estimate compared to smaller effect sizes

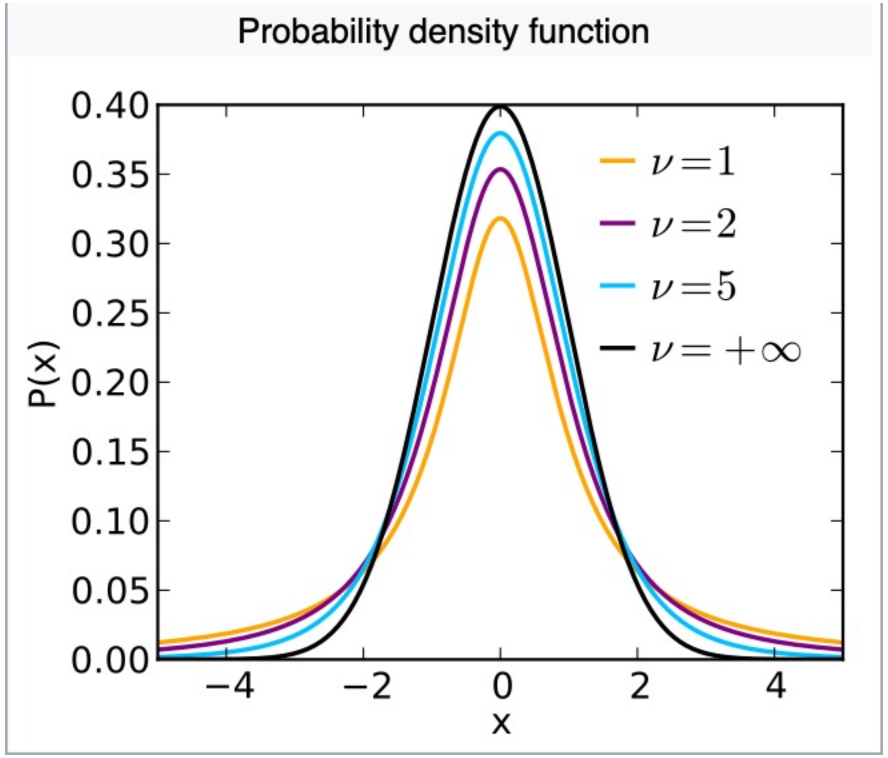
Poll: If Type II error for a given hypothesis test is zero then what is its statistical power?

Z/T-statistic (Two-sample t-test)

$$Z = \frac{mean\left(Y_{E9.5}\right) - mean\left(Y_{E11.5}\right)}{sd(Y)\sqrt{\frac{1}{n} + \frac{1}{n}}}$$

Sampling distribution of T-statistic under the Null hypothesis





T-tests requires assumptions of…

- Normality of the responses
- Equal variance of the two groups being compared

Parametric versus non-parametric tests

- Parametric tests make distributional assumptions about the response variables (Example: Normal probability distribution for the t-test)
- Non-parametric tests do not make such assumptions (Example: Mann-Whitney test (next))

U-statistic (Mann Whitney test, two sample test)

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

- Two groups
- Rank all observations across both groups, smallest observation given rank 1.
- The sum of ranks of observations within group 1 with n1 observations is R1

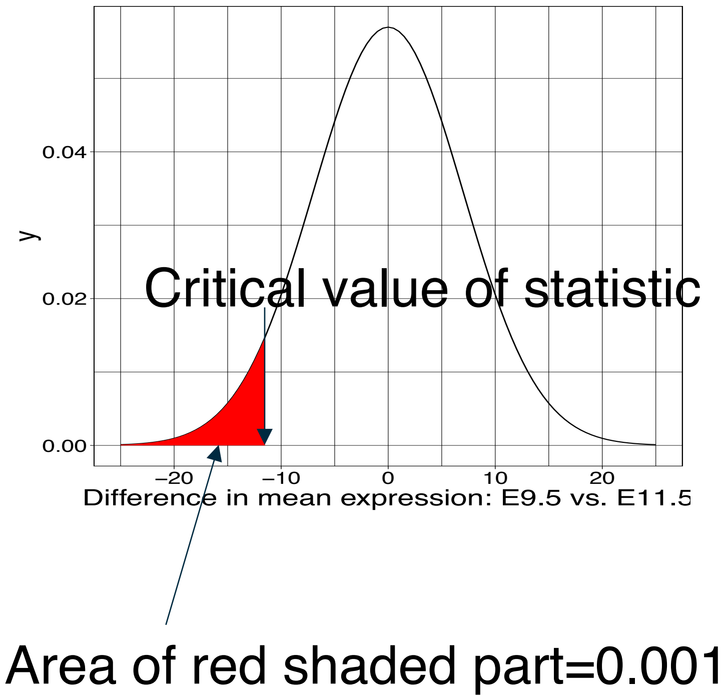
U-statistic sampling distribution in terms of tables

Mann-Whitney Table

The following tables provide the critical values of *U* for various values of alpha and the sizes of the two samples for the two-tailed test. For one-tail tests double the value of alpha and use the appropriate two-tailed table. See Mann-Whitney Test for details.

Alpha = .001 (two-tailed)

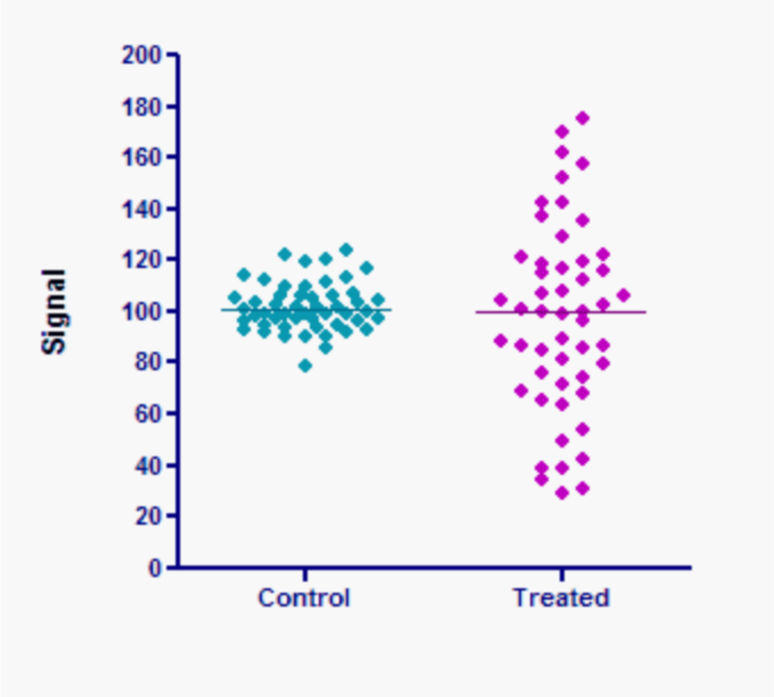
n1\n2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2																			
3																			
4												0	0	0	1	1	1	2	2
5								0	0	1	1	2	2	3	3	4	4	5	5
6							0	1	2	2	3	4	5	5	6	7	8	8	9
7						0	1	2	3	4	5	6	7	8	9	10	11	13	14
8					0	1	2	4	5	6	7	9	10	11	13	14	15	17	18
9				0	1	2	4	5	7	8	10	11	13	15	16	18	20	21	23
10				0	2	3	5	7	8	10	12	14	16	18	20	22	24	26	28
11				1	2	4	6	8	10	12	15	17	19	21	24	26	28	31	33
12				1	3	5	7	10	12	15	17	20	22	25	27	30	33	35	38
13			0	2	4	6	9	11	14	17	20	23	25	28	31	34	37	40	43
14			0	2	5	7	10	13	16	19	22	25	29	32	35	39	42	45	49
15			0	3	5	8	11	15	18	21	25	28	32	36	39	43	46	50	54
16			1	3	6	9	13	16	20	24	27	31	35	39	43	47	51	55	59
17			1	4	7	10	14	18	22	26	30	34	39	43	47	51	56	60	65
18			1	4	8	11	15	20	24	28	33	37	42	46	51	56	61	65	70
19			2	5	8	13	17	21	26	31	35	40	45	50	55	60	65	70	76
20			2	5	9	14	18	23	28	33	38	43	49	54	59	65	70	76	81



Mann-Whitney test valid as a comparison of location only if…

- The two distributions have the same underlying shape, variance

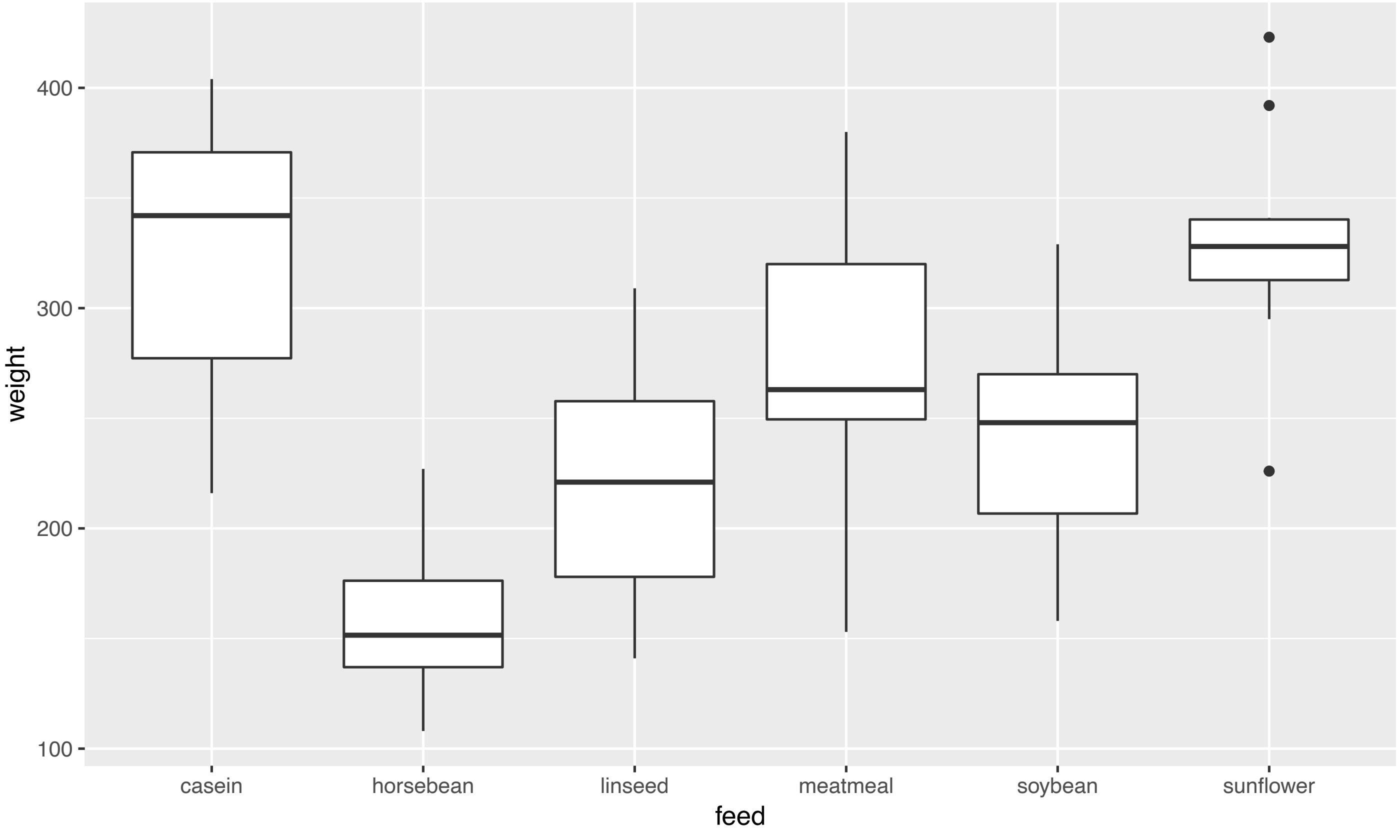
Same location, significant p-value



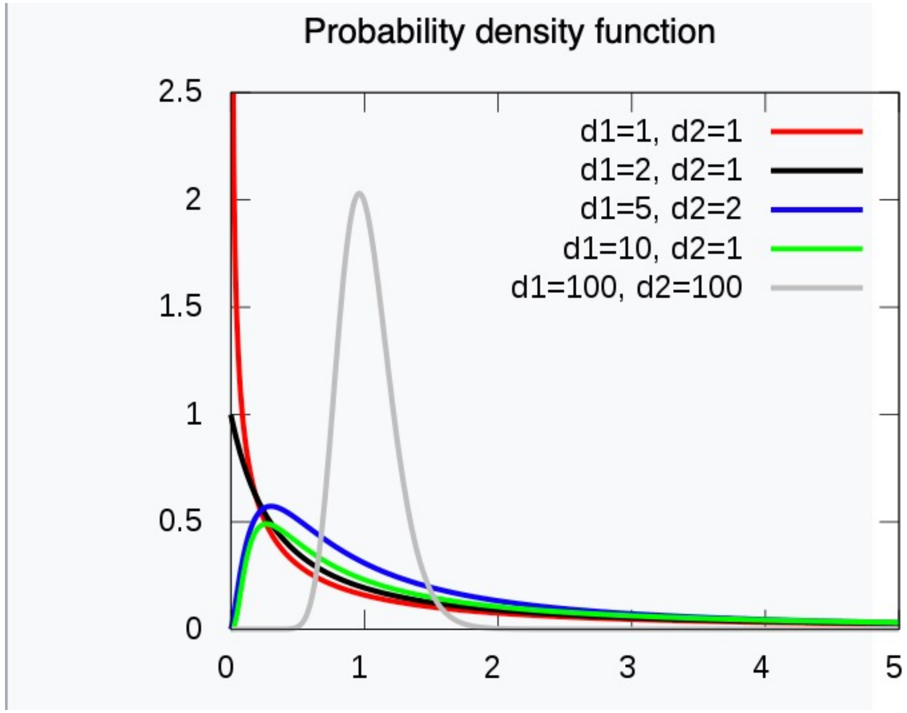
Same location, non-significant p-value

F-statistic (ANOVA)

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$



Sampling distribution of the F-statistic



1-way ANOVA requires assumptions of…

- Normality of the responses
- Equal variance of the responses with each of the groups being compared

Poll: Are you aware of the difference between the t-test, Welch t-test, Mann-Whitney test?

Why do we have so many different tests?

- Sampling distribution derived via Central Limit Theorem only valid only if certain **<u>assumptions</u>** met with underlying data
- E.g. of assumptions could be Normality, Equality of variances etc.

Every hypothesis test requires…

- Test statistic
- Sampling distribution of test statistic under the null hypothesis
- A Type I error that will be allowable – fraction of times you are willing to accept a false-positive as a real result
- Note: Use of test statistic and associated sampling distribution depends on your data meeting certain assumptions
- A Type II error given the effect size of the association you are expect to estimate

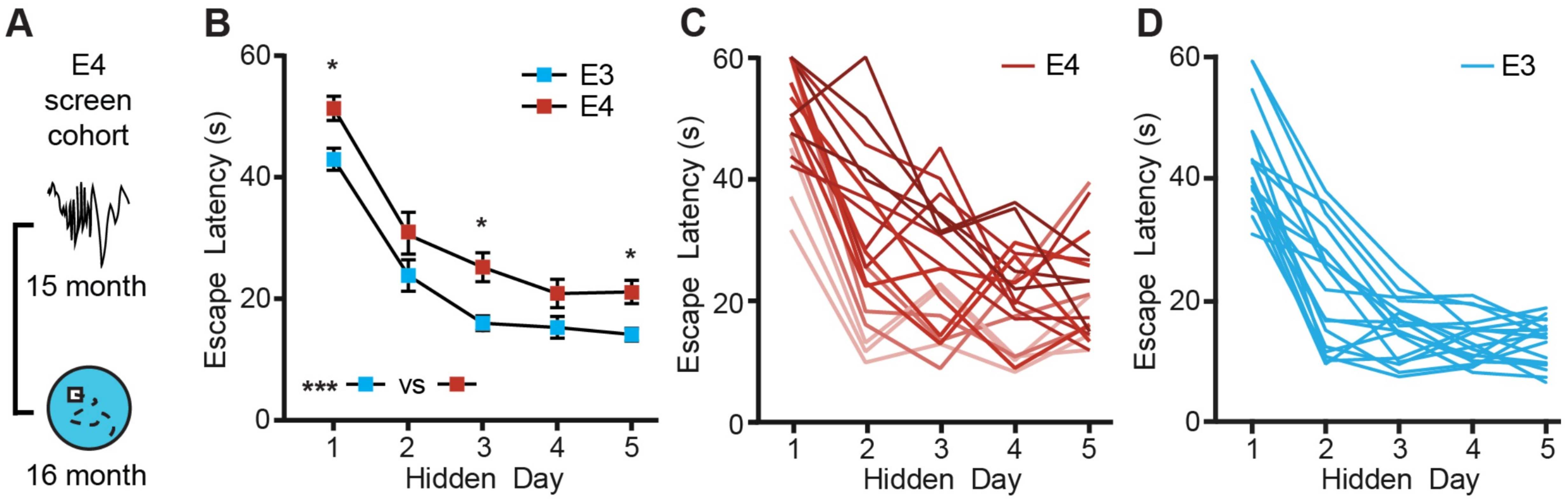
Outline

- Introduction to hypothesis testing
- Define variables
- Choosing the right test
- Basic concepts in hypothesis testing
- **Hands-on**

Repeated measures experimental design

- Designs where multiple responses from the same biological unit are assessed
  - Examples include measuring changes in biomarker levels (e.g. CD4 counts) in subjects over time

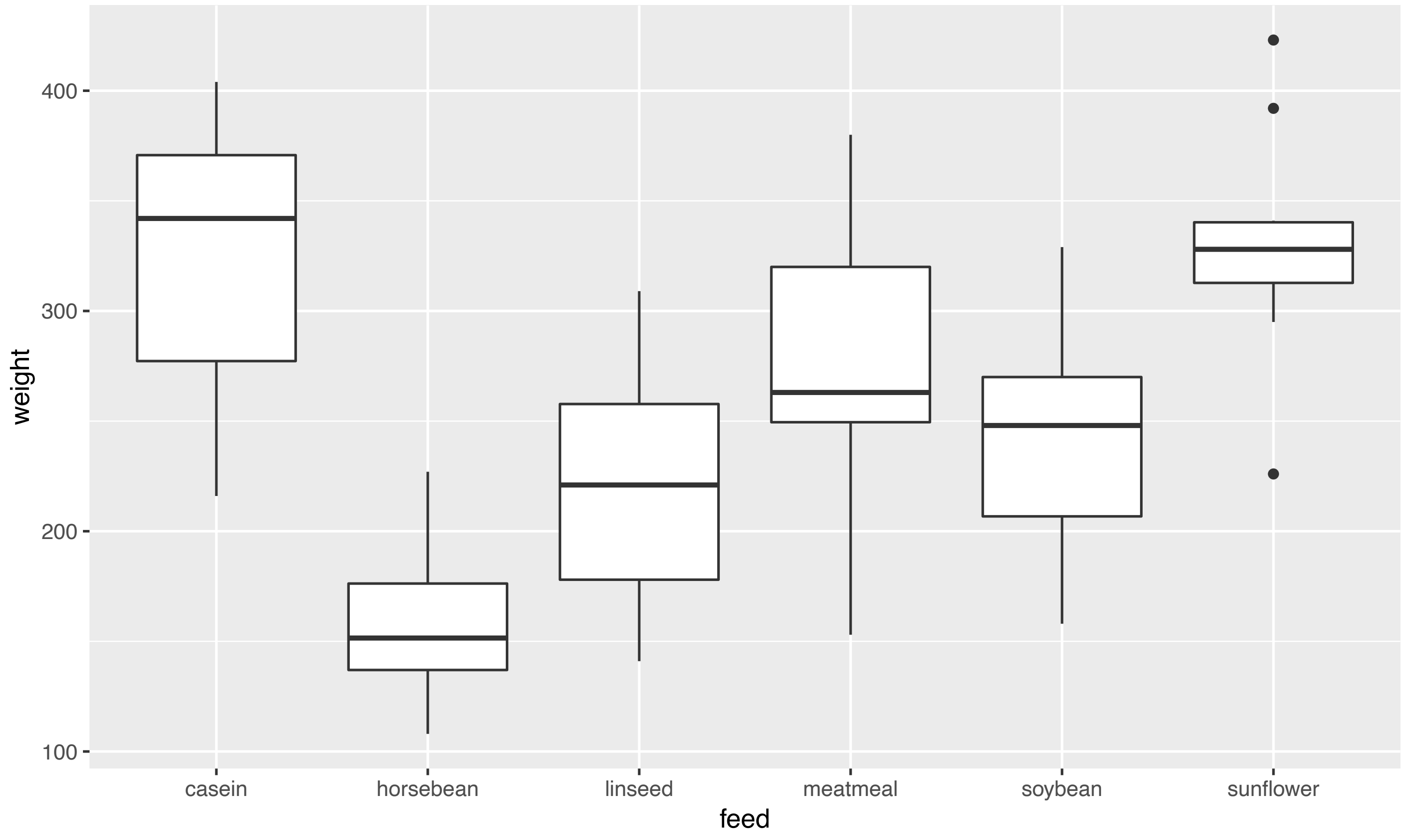
Learning in Alzheimer's Disease mice assayed in the Morris-Water Maze



Jones et al. 2019

Comparing every feed to every other one

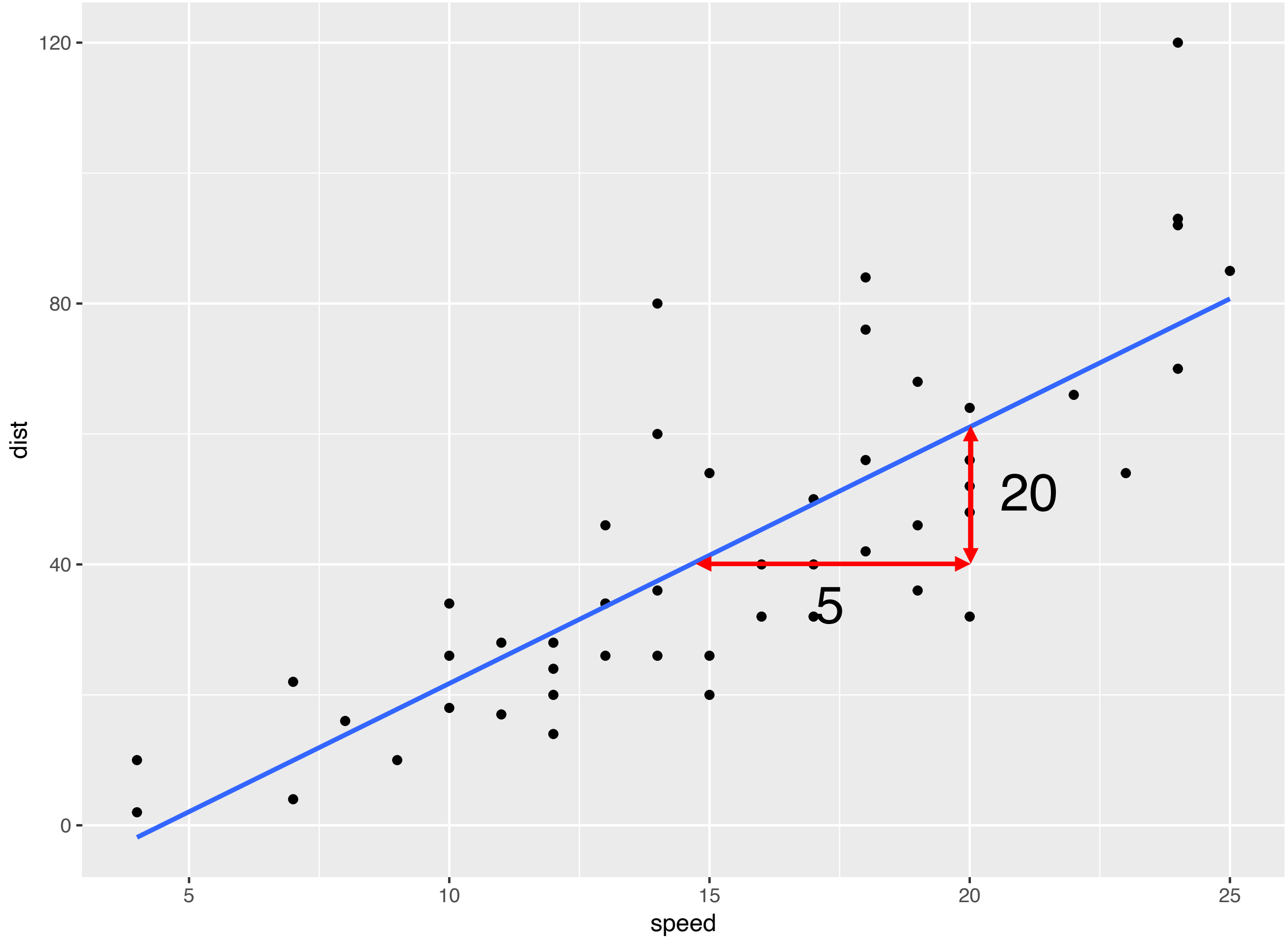
There are 15 possible comparisons



Why do we need multiple testing?

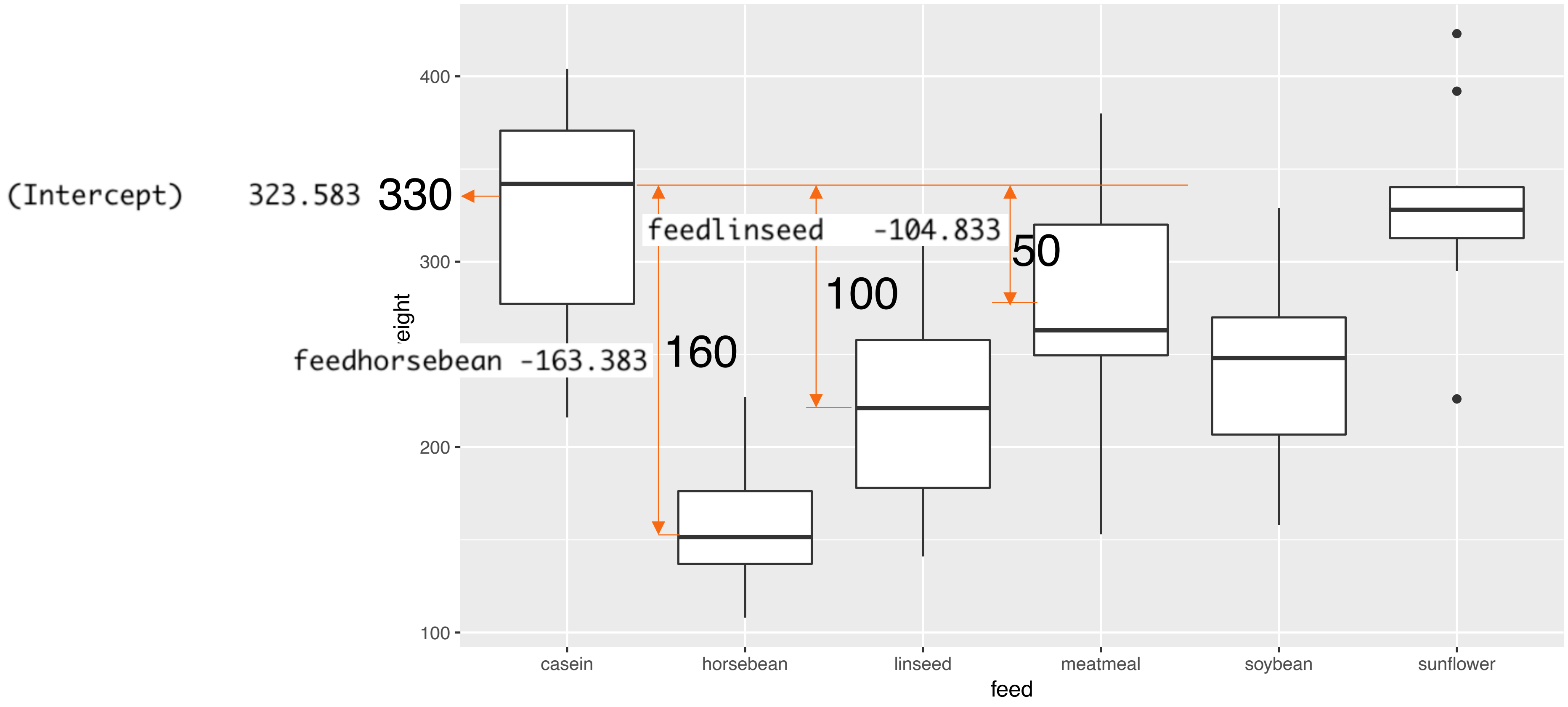
- We have 15 possible comparisons between feeds
- Assume no. of true associations = 8
- We set Type I error = 0.05
- Assume statistical power to detect differences = 0.8
- We will detect 8x0.8 ~ 6 true differences
- #false positives = 15x0.05~1
- False Discovery Rate = #false positives/(# false positives + #true positives) = 1/(1+6) ~ 14% - pretty high!

Interpret parameters from linear model to estimate slope

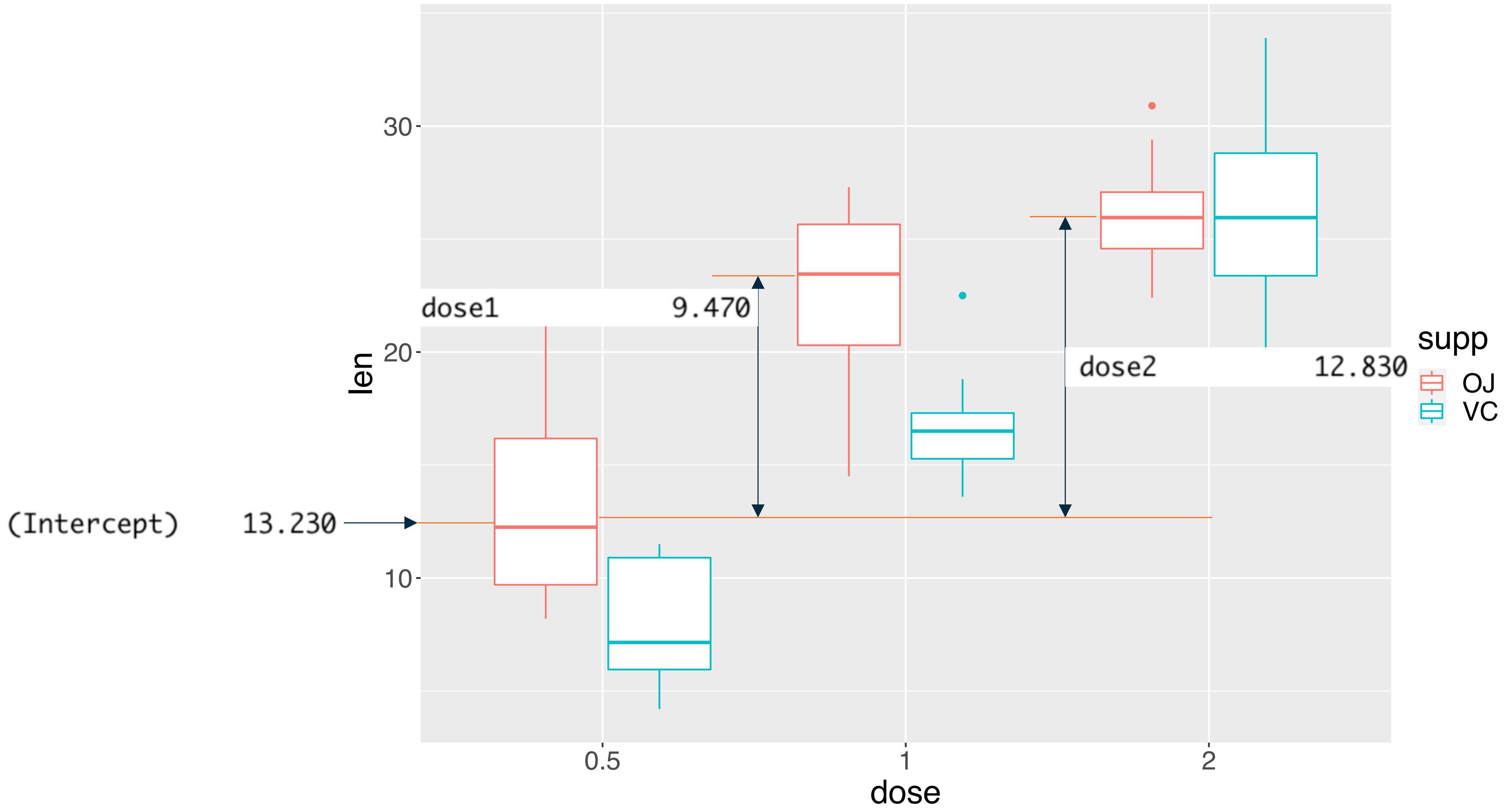


Slope ~ 20/5 = 4

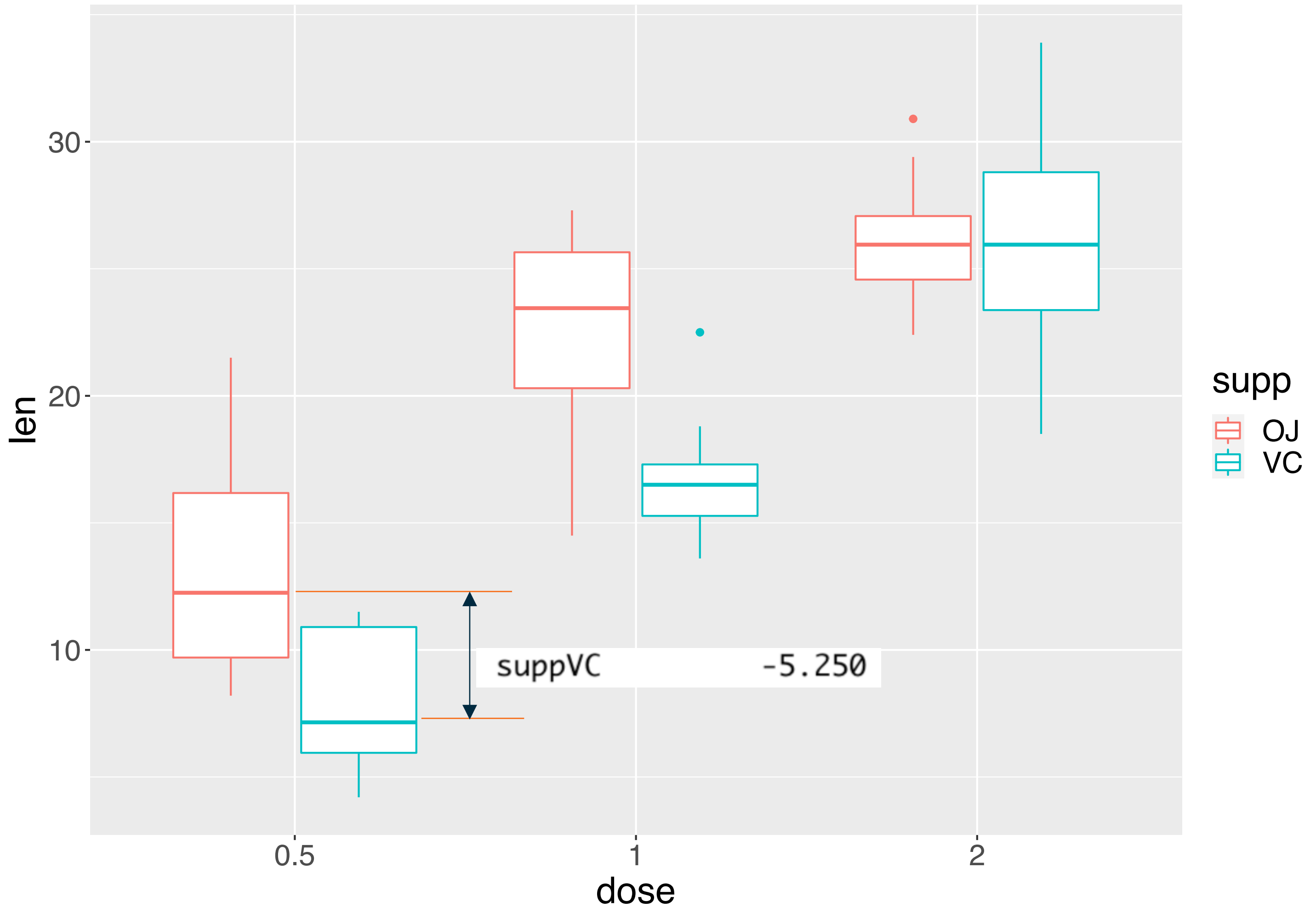
Interpret parameters from linear model implementation of one-way ANOVA



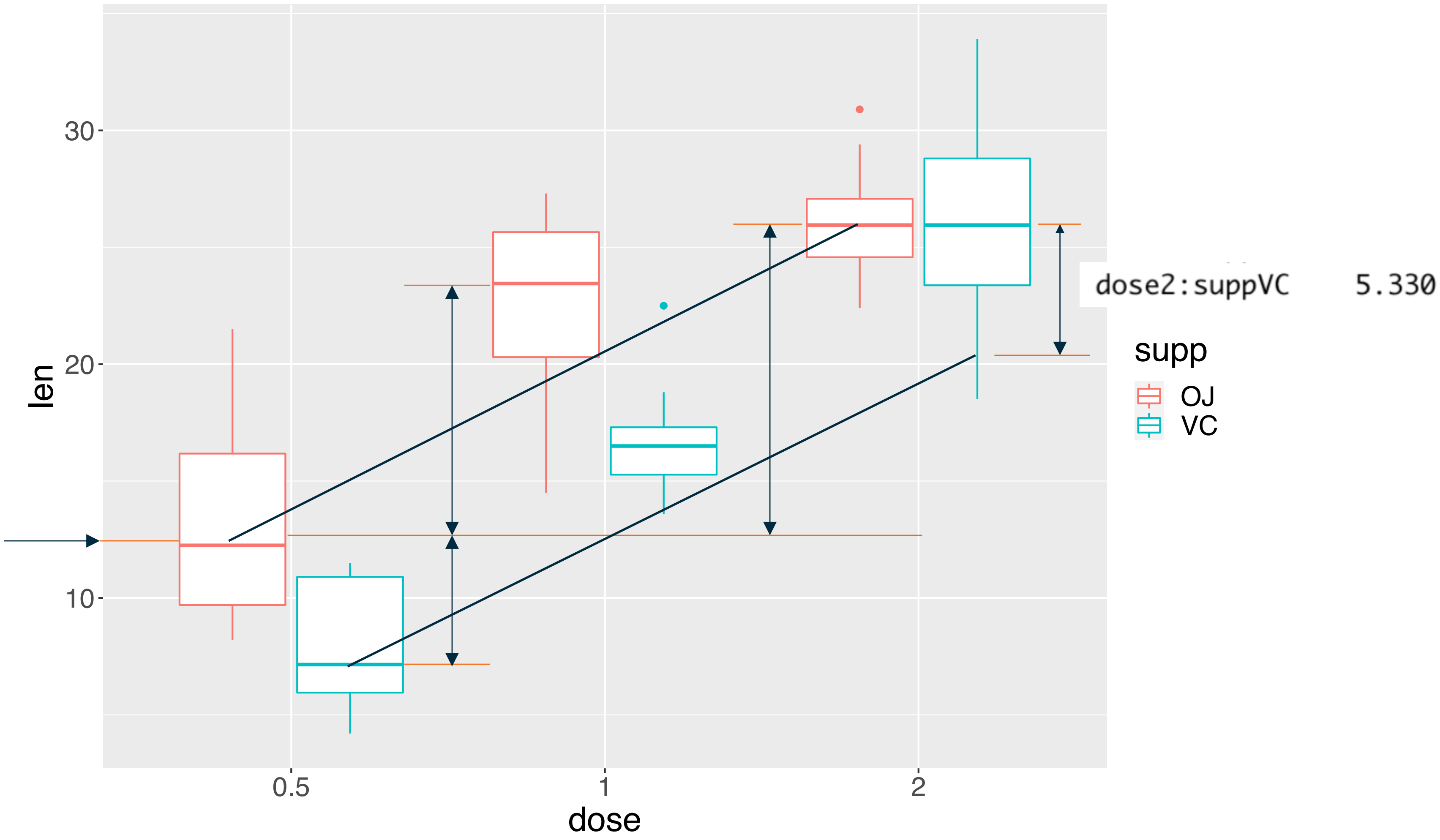
Interpret parameters from linear model implementation of two-way ANOVA

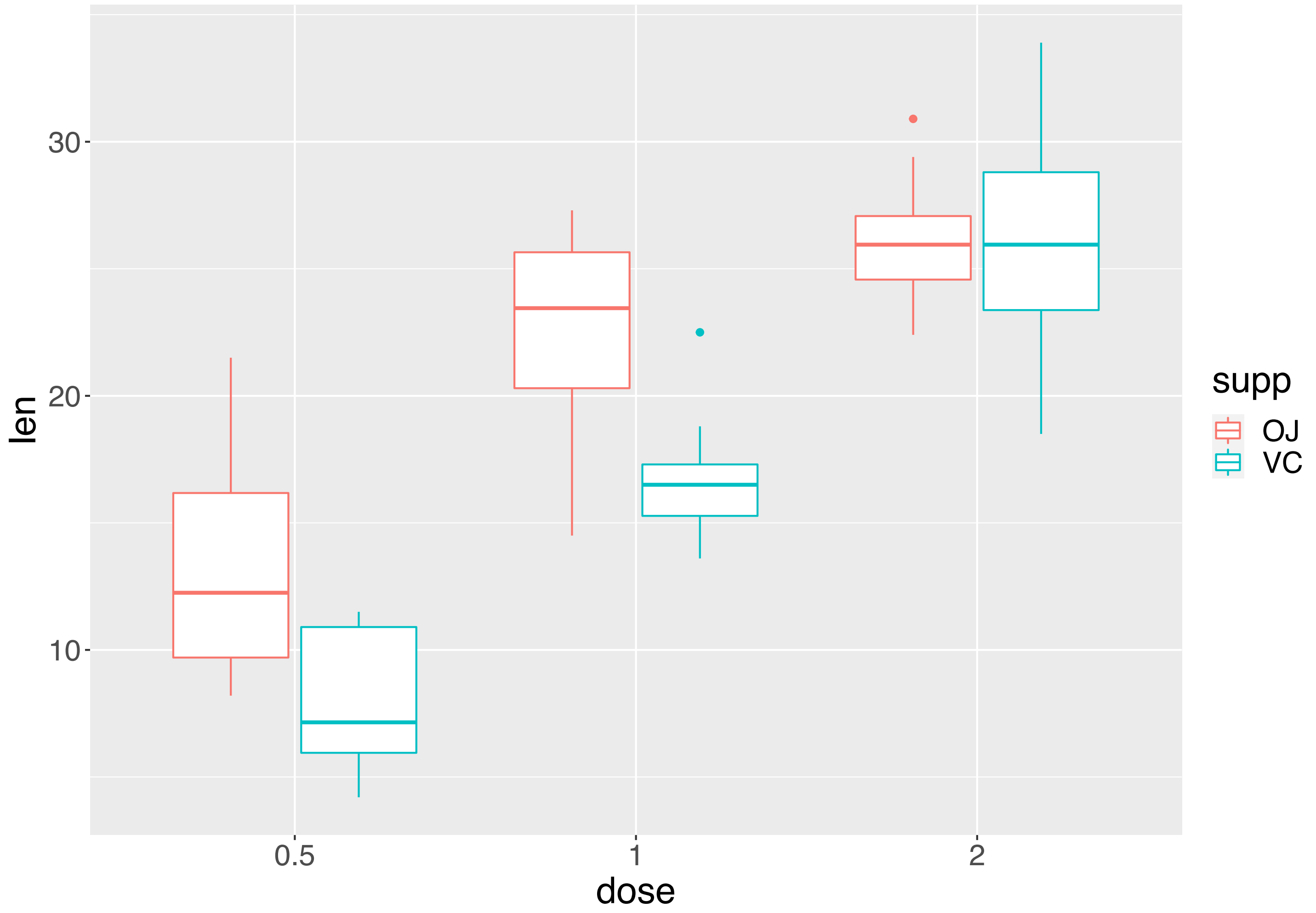


Interpret the main effect



Interpret the interaction term

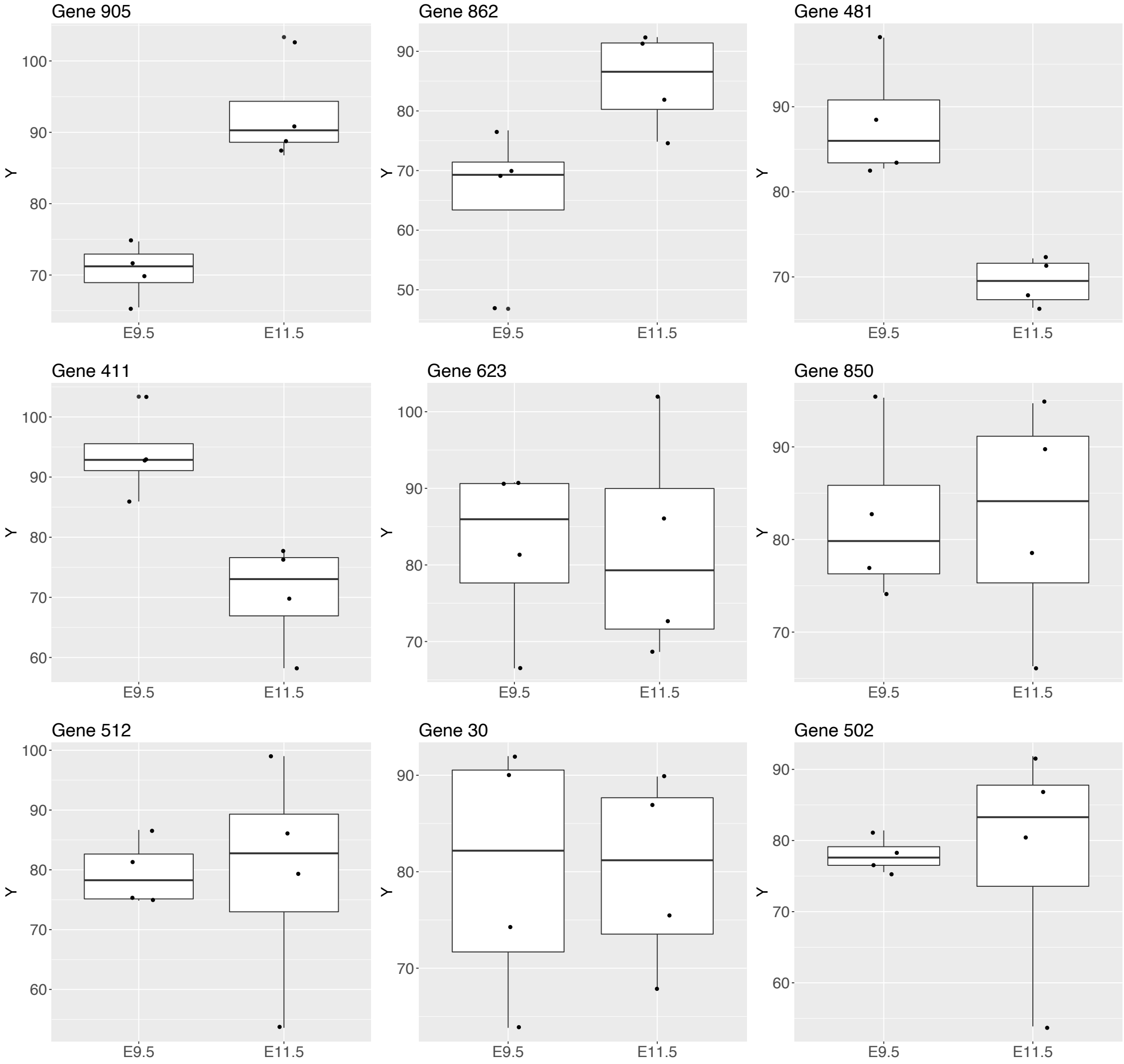



30
20
10
len
0.5
1
2
dose
supp
OJ
VC

Please fill-out survey

- https://www.surveymonkey.com/r/F75J6VZ
- ~ 3min

Multiple tests



Outline for this workshop