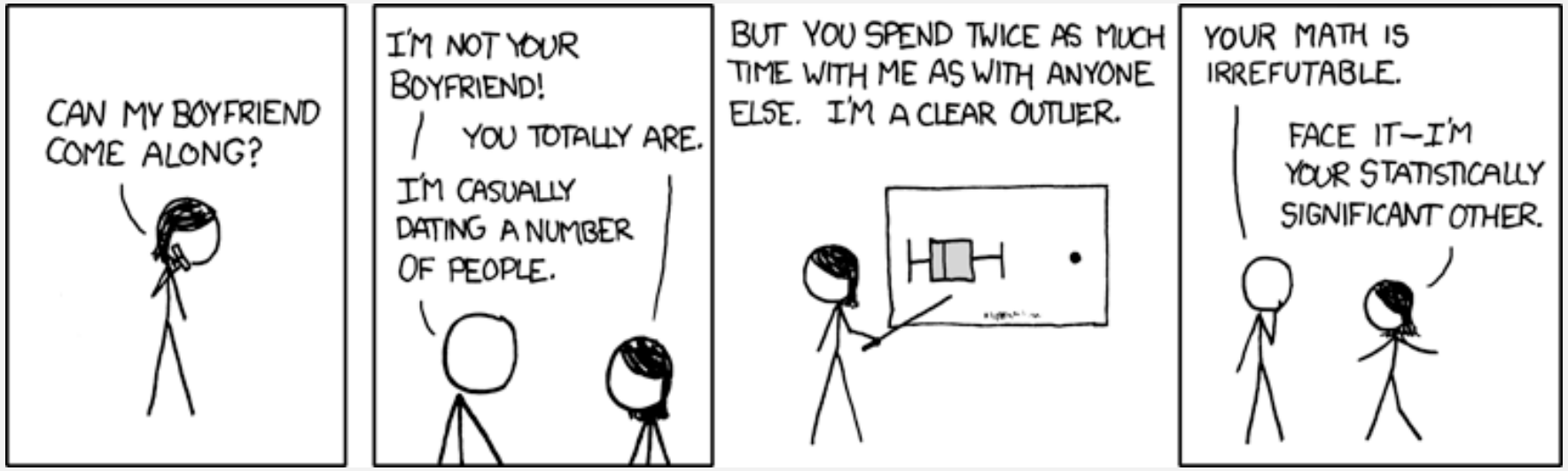
Statistical Principles

Biology 683

Heath Blackmon



Example test questions:

Name 3 causes of the reproducibility crisis

Name 3 possible solutions to the reproducibility crisis

This week

1. Terminology (a lot of it!)
2. P-values
3. Terminology
4. Central Limit Theorem

Data

Data terminology

Data

Variables

The characteristics that differ among individuals

Data

The measurements of variables taken for a sample of individuals

Data

Numerical Variables

Individuals vary on a quantitative scale

Figure 1. Joiner's living histogram of student height.

Categorical Variables

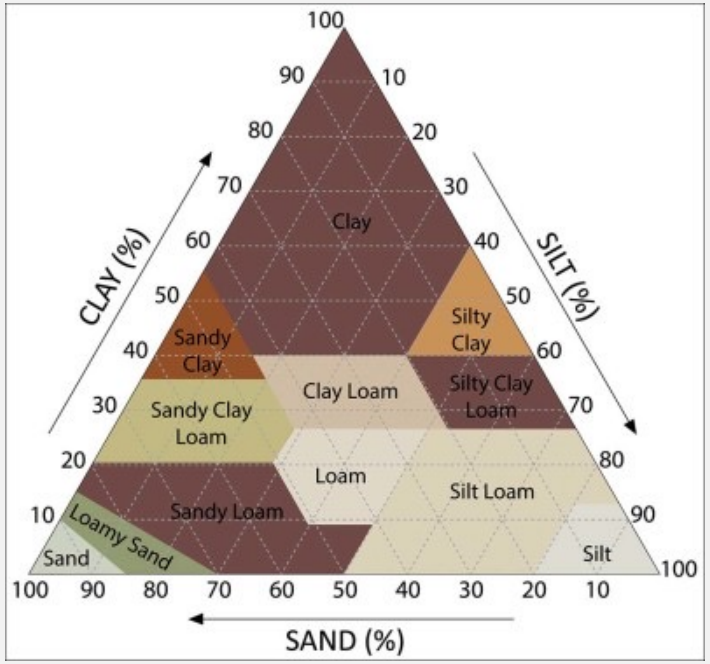
Individuals are in qualitative categories

Ordinal

The categories can be ordered

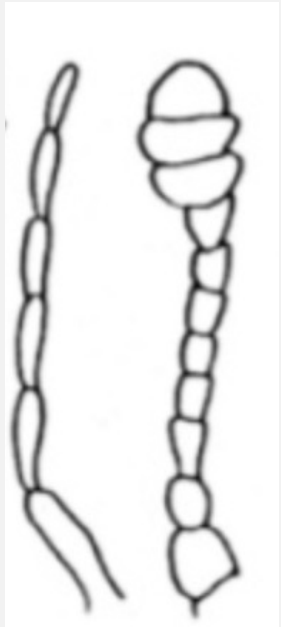
Nominal

The categories have no inherent order

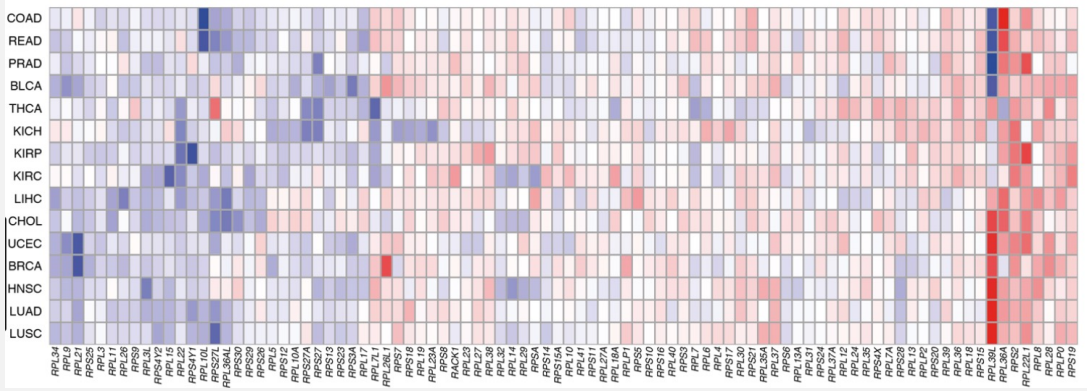


Continuous vs Discrete

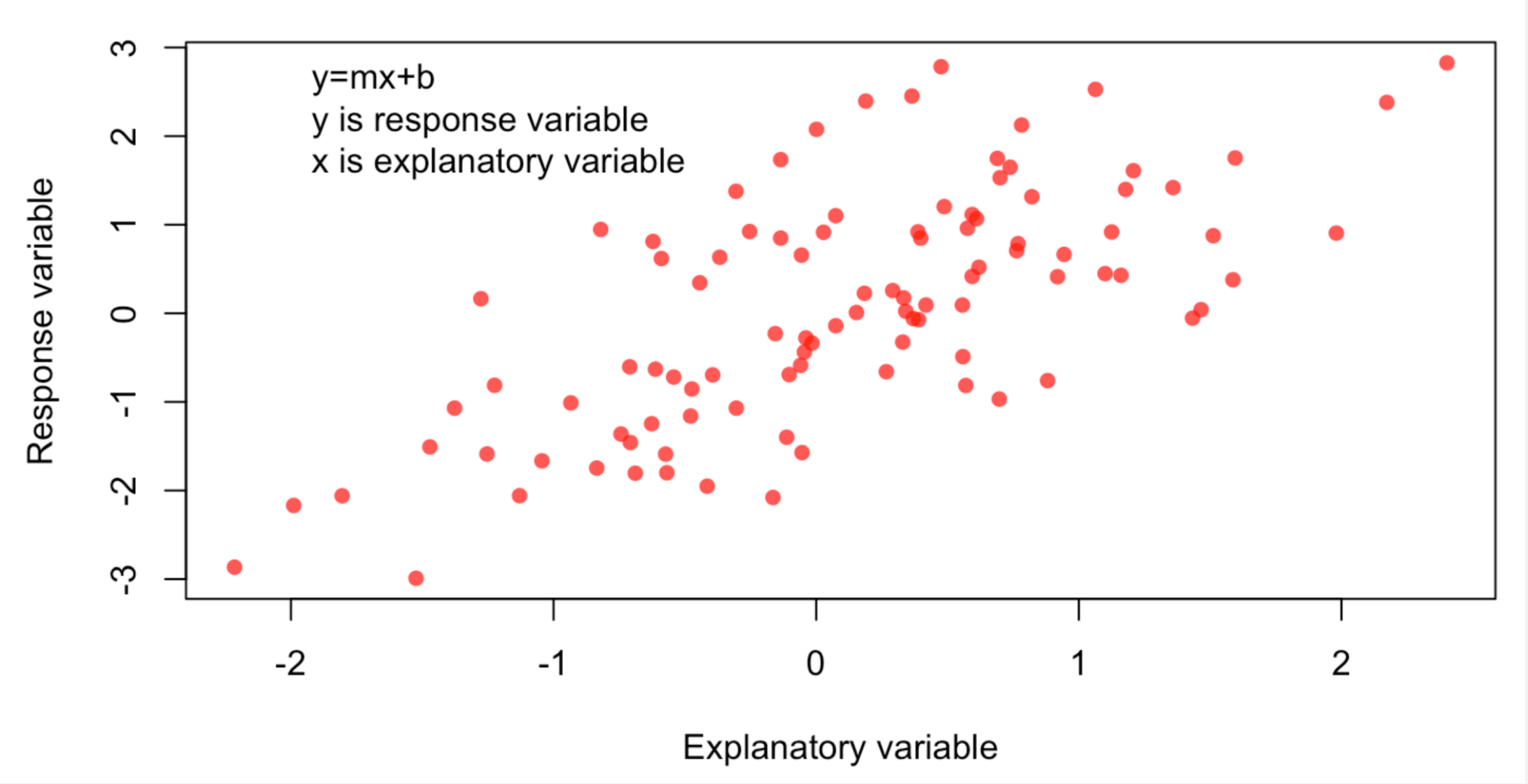
Continuous variables
a variable that has an infinite number of possible values



Discrete variables
a variable that has a finite number of possible values



Explanatory and Response Variables



Sampling

What is sample?

Populations and Samples

- **Populations**

  Some sort of group of something - could be anything
  - Undergraduates at Texas A&M
  - Jewel beetles in Arizona
  - Strain of flies in the lab
  - People on the titanic

- **Samples**
  - A subset of individuals drawn from a population

Going from samples to populations

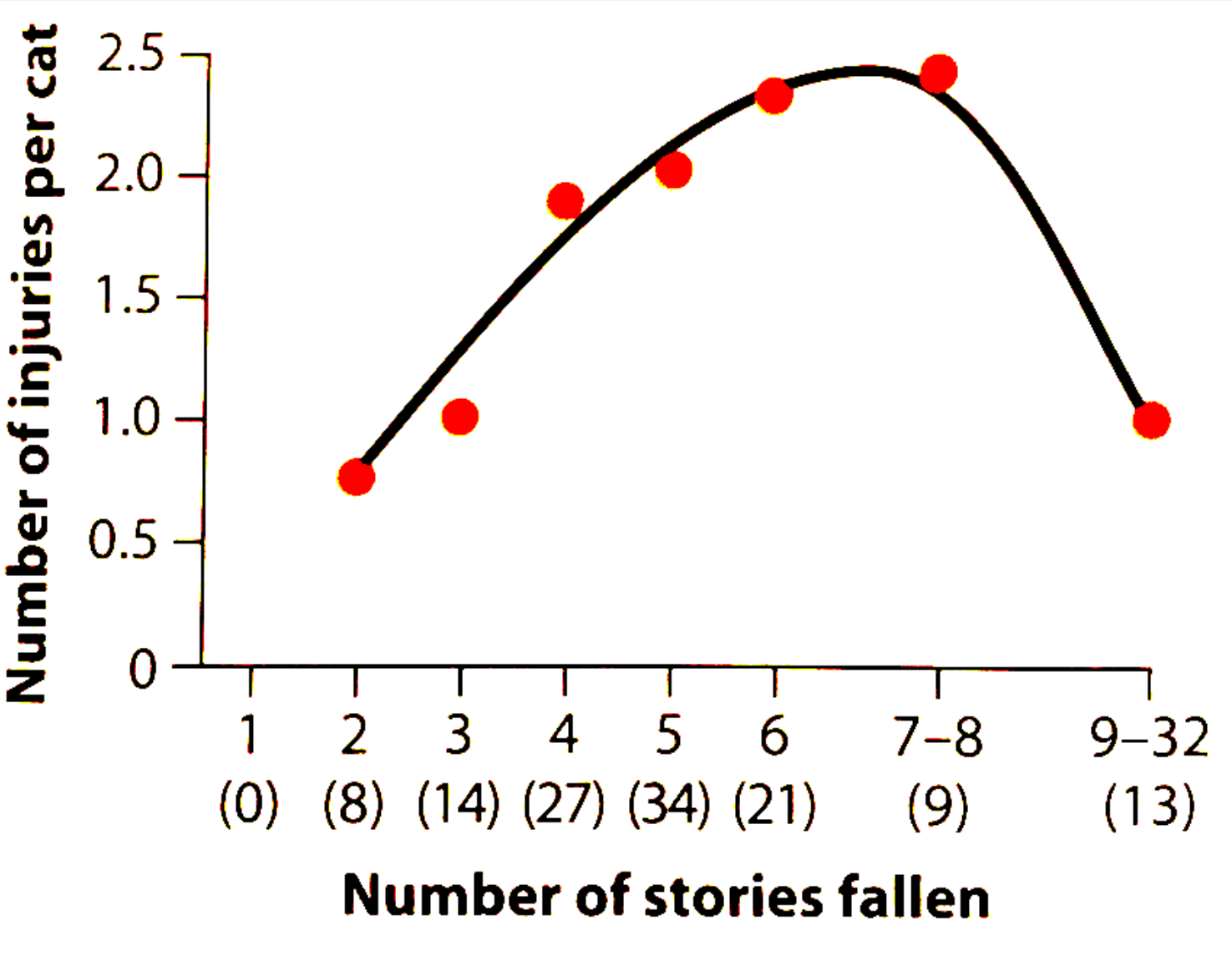
- The group C57BL/6 mice on a high fat diet in the BSBW vivarium
- C57BL/6 mice on a high fat diet
- All *Mus musculus*
- All rodents
- All mammals
- All animals
- All life

Falling Cats

In the period between January 1, 1998 and December 12, 2001 at the Clinic of Surgery, Orthopedics and Ophthalmology of the Veterinary Faculty, 119 cats were treated after a fall or jump from a balcony or window, where the owners saw the fall, or where there was a reasonable suspicion that a fall had occurred. Only those cats that fell from the second or higher stories were included. The owners brought the cats for treatment within varying periods of time after the fall (from 30 min to over a month).

Vnuk, et al. "Feline high-rise syndrome: 119 cases (1998–2001).
Journal of Feline Medicine & Surgery 6.5 (2004): 305-312.

What is the population?



What is your interpretation of the results from this study?

Sampling Considerations

A researcher is interested in understanding how exercise impacts blood pressure. To collect data for their study they go door to door in their neighborhood and ask if any individuals in the house would be interested in participating in the study (a blood pressure reading and a quick questionnaire. To encourage people to participate the researcher is providing $10 Starbucks gift cards.

Random Sampling

1. Every unit in a population should have an equal chance of being sampled.

2. The selection of units must be independent.

3. What are some ways of being non-random?

Hypotheses and P-values

The Null Hypothesis

To analyze your data, you will need a statistical hypothesis to go with your scientific hypothesis

A statistical hypothesis is most easily constructed as a null hypothesis

A null hypothesis posits that the factor of interest has no effect

Frequentist test we will be looking at p-value

Bayesian approaches usually tells us if the posterior estimate of the parameter of interest overlap in our two treatments.

Examples of Null Hypotheses

Fertilizer has no effect on the growth rate of oak trees.

Blocking olfactory cues has no effect on mate choice in swordtail fishes.

Rates of genome evolution are the same in two populations.

Mutations in the 5’ UTR of msl-2 have no effect on translation.

Rejecting the Null

- You apply a statistical test to determine whether your data can reject the null hypothesis

- If you reject the null, then one of the alternative hypotheses must be true – though not necessarily the one you believe or even one you've ever imagined!

- You cannot **prove** a hypothesis, but
  - As frequentist you can find support for an alternative by rejecting the null. The more convincing the null and the more well designed the experiment the more evidence you provide for your alternative.
  - As a Bayesian you can compare support for two competing hypotheses.

What is a p-value?

Is the probability of finding the observed, or more extreme, statistic when the null hypothesis is true (generating the data).

```
> x
     [,1] [,2]
[1,]  140    4
[2,]   80   13
[3,]   76   89
[4,]   20    3
> chisq.test(x)

        Pearson's Chi-squared test

data:  x
X-squared = 117.31, df = 3, p-value < 2.2e-16
```

Number of women on titanic who survived (first column) or died (second column) in first, second, third, or crew classes (rows 1:4 respectively).

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Misconceptions about p-values

The p-value is not the probability that the observed statistic is due to random chance.

A p-value is not the probability that your alternative is false.

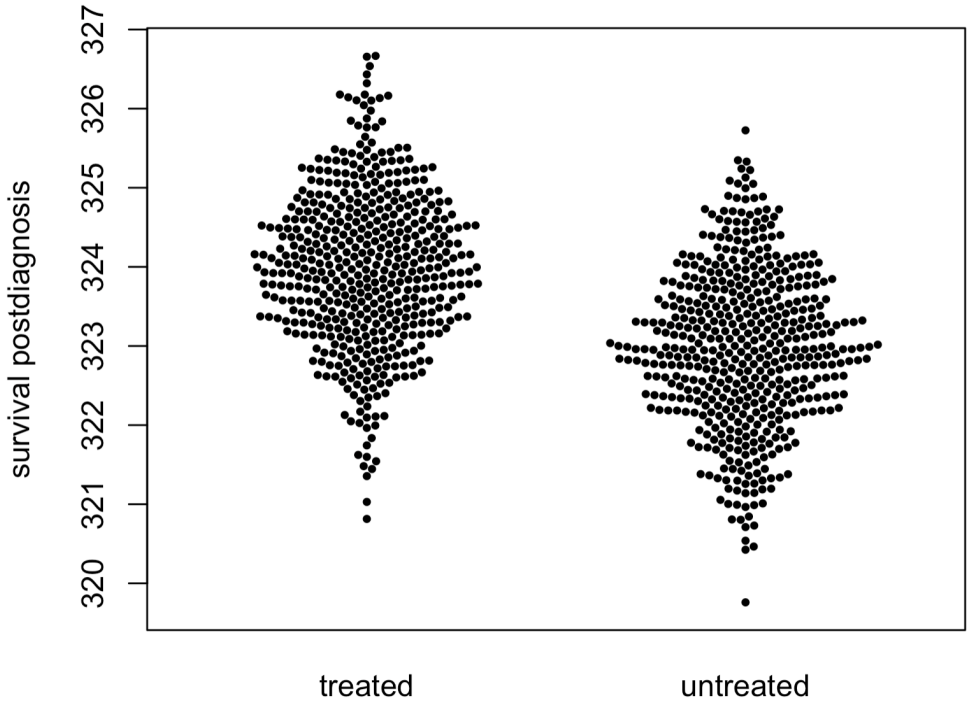
A p-value is not the probability that the null hypothesis is true.

The magnitude of the p-value does not indicate the importance of an effect. Statistical significance does not equate to biological significance.

Studies with p-values on opposite sides of 0.05 are equally “correct”.

Thoughts

A drug company has just developed a new drug Naquadah. This drug is tailored to treat a terminal illness the company reports that their drug increases life span in people who are diagnosed with diseases with a p-value of 1.9×10^{-9}. Unfortunately this drug is very expensive and the prescribed treatment regime would run approximately $40,000 per patient. Should your insurer pay for this drug if you are diagnosed with the diseases.



There are many ways of calculating p-values

Traditional statistical tests

For many questions/experiments there isn't a ready made statistical test.
- Randomization of datasets
- Comparison to simulated datasets

I will ask questions about p-values on tests!

Type I versus Type II Error

Type I error refers to rejecting a true null hypothesis

Type II error refers to failing to reject a false null hypothesis

Power is a description of our probability of rejecting a false null hypothesis

We usually set up statistical tests to avoid Type I errors, at the expense of possibly committing Type II errors

Type I error = FALSE POSITIVE
1 – Type 2 error = POWER

Parameter, estimates, sampling considerations

Parameter: Population-level variables we are trying to estimate

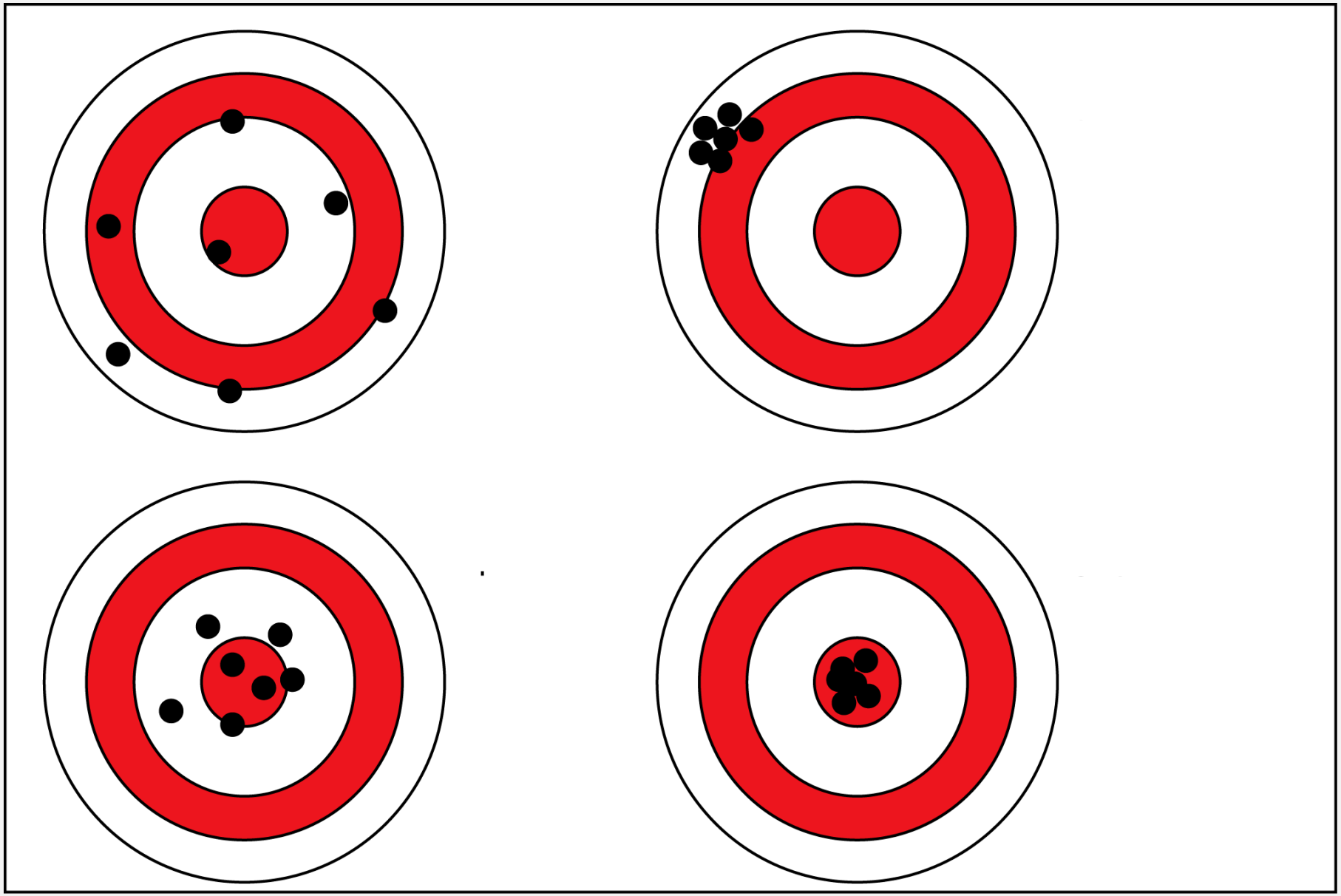
Estimate or Statistic: The value of the parameter inferred from the sample

Bias: If something about the sampling or measuring procedure causes the sample to systematically misrepresent the population.

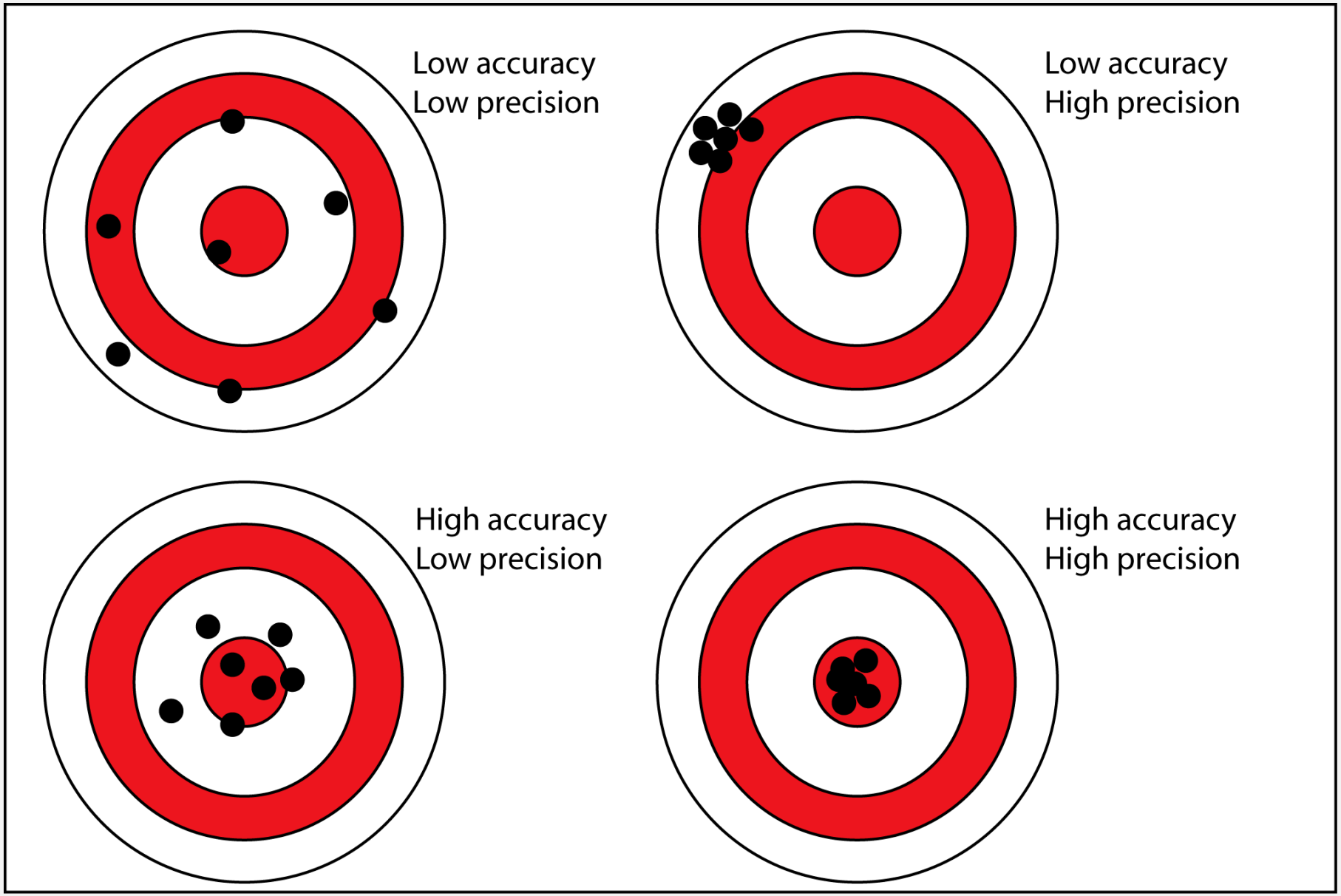
Precision: How tightly grouped are the estimates.

Accuracy: How close estimates are to the true value.

Accuracy vs Precision



Accuracy vs Precision



Experimental vs observational studies

- Does caloric restriction increase lifespan in mice?
- Is global warming caused by human activities?
- Does smoking cause lung cancer in humans?
- Does parasite infection reduce mating success of beetles?
- Does oxytocin affect sexual attraction in humans?
- Do sex chromosomes increase the rate of speciation?
- Do chromosome fusions reduce fitness?

Why should we summarize data?

- Many datasets are simply too big to look at all values and form an impression?

- Our impressions of small datasets are often misled by our tendency to look for patterns.

Typical summary statistics

- **Mean:** Sum of the observations divided by the number of observations
- **Median:** The middle observation in a set of data
- **Variance:** The average squared deviation from the mean
- **Standard Deviation:** The square root of the variance

Symbols for samples and populations

Samples versus Populations

The mean or standard deviation statistic you calculate from your sample is an estimate of the population parameter.

Parameter Symbols:

μ (mu): population mean

σ (sigma): population standard deviation

Statistic Symbols:

$\bar{Y}$ (Y bar): sample mean

s : samples standard deviation

For a sample of a population

The mean is just: $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$

The standard deviation is $s = \sqrt{s^2}$

Where s^2 or the variance is: $s^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$

Central limit theorem

- Imagine that we sample from the same population many times, so we have a bunch of different, independent samples.

- Each sample will have a mean, but the means will be different due to chance. In principle, we could draw a histogram of these means.

- In general, you only have one sample from a given population, however, so what can you infer about the distribution of the means from your sample?

- The Central Limit Theorem states that regardless of the underlying population distribution of the variable of interest, the distribution of the population of means will be roughly normal.

Central limit theorem

Your estimate of the sample mean is an estimate of the mean of this distribution of means (that is, it's your best estimate of the population mean).

The hypothetical distribution of sample means has a standard deviation equal to s divided by the square root of n.

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

We call this standard deviation the standard error of the mean (SEM). The true population mean should be within $\bar{Y} \pm 1.96SE_{\bar{Y}}$ 95% of the time

Central limit theorem

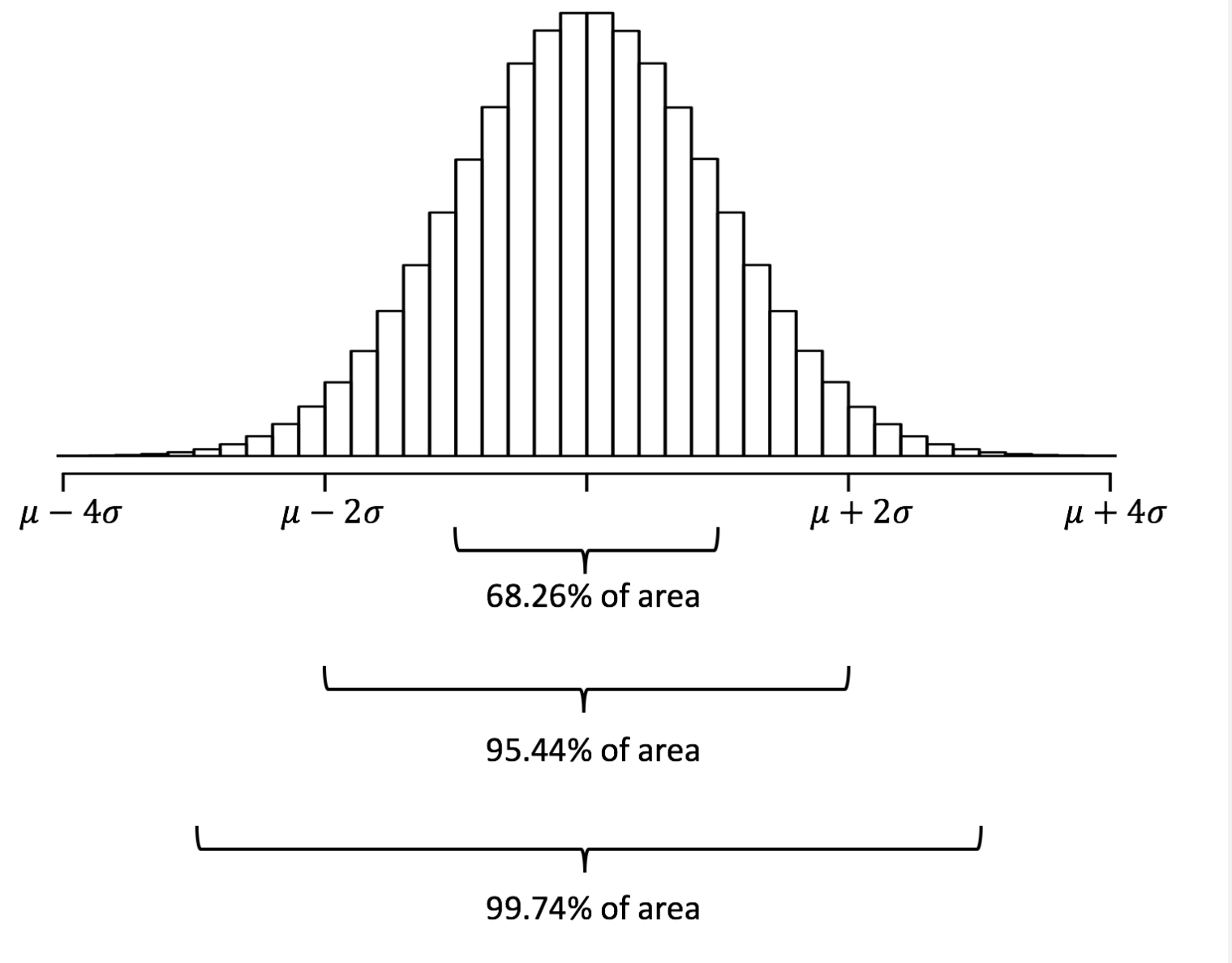
Lets try that

create a population with a known mean.

sample from it and calculate the mean and standard error and see if it includes the true mean.

tally results and see if it worked about 95% of the time

Estimating with uncertainty



Confidence Interval vs Credible Interval

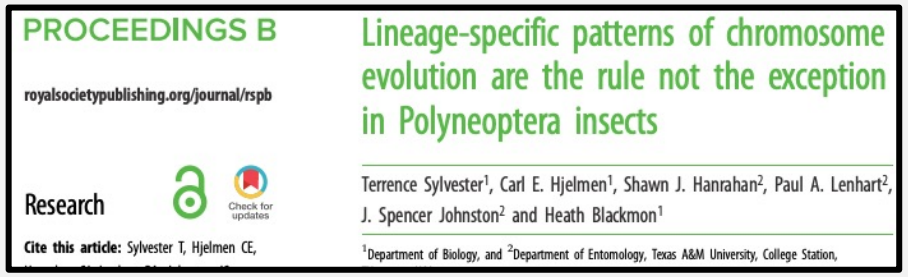
$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

z = 1.65 fo 90%
= 1.96 for 95%
= 2.58 for 99%

natural choice for things we go and measure in biological entities and we are interested in what the “true” mean value of the population is

Credible intervals are often used in Bayesian approaches. In these methods we often run an MCMC which yields an arbitrarily large number of estimates of our parameter of interest. It is not sensible to talk about the CI of a parameter estimate like this because it can always be narrowed to a point estimate with sufficient sample size.

Confidence Interval vs Credible Interval



PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Sylvester T, Hjelmen CE,

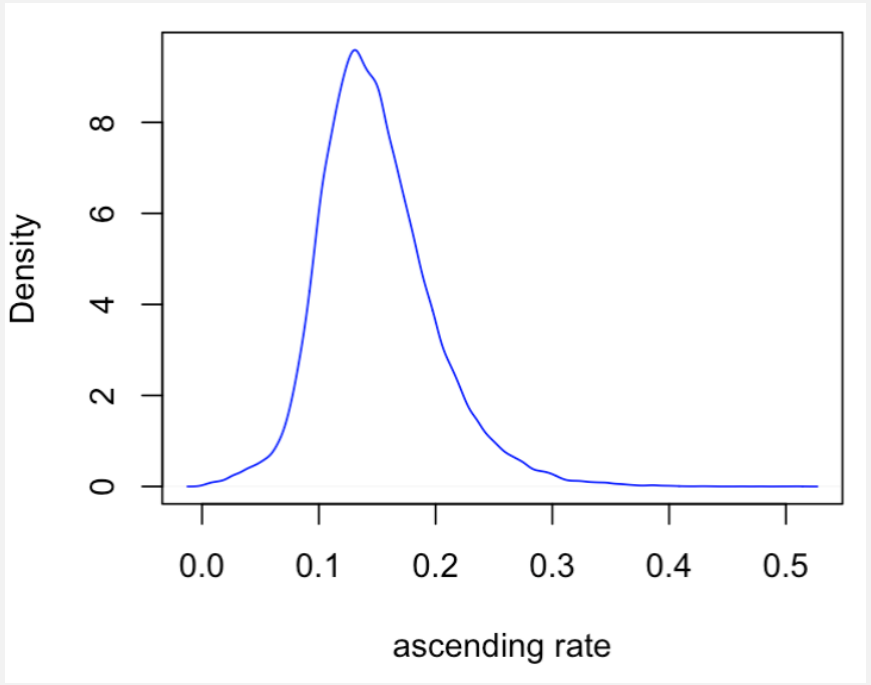
Lineage-specific patterns of chromosome evolution are the rule not the exception in Polyneoptera insects

Terrence Sylvester[1], Carl E. Hjelmen[1], Shawn J. Hanrahan[2], Paul A. Lenhart[2], J. Spencer Johnston[2] and Heath Blackmon[1]

[1]Department of Biology, and [2]Department of Entomology, Texas A&M University, College Station,

column 2: numeric with range 0 - 0.55	desc1	pol1	p
0.158975292	0.13567824	0.0012022928	-267.9407
0.141901449	0.17588564	0.0011763734	-269.2247
0.138279931	0.13940276	0.0021394110	-268.2814
0.123512205	0.11179867	0.0028047752	-268.0275

10,000 rows



Frequentist 95% CI
0.149-0.150

95% HPD (credible interval)
0.06-0.26

Some Experimental Design Considerations

Why do I need a control?

To interpret an experiment, we need to compare the experimental subjects to the correct reference group.

What about observational studies?

What is an appropriate control?

Ideal controls are identical to the experimental population, except for the one parameter being manipulated

The control population should be similar in all other respects to the experimental population

The control population should experience sham manipulations that simulate any manipulations applied to the experimental population

Sometimes you might need multiple different controls.

Avoiding Experimenter Bias

Experimenter bias is real

The results of your study can be influenced by your expectations

Some precautions

Randomize assignment of subjects to controls and treatments (**use R or random.org**).

Humans are bad at recognizing and creating randomness.

Avoiding Experimenter Bias

Use a blind or double-blind experimental design

Blind: the subject doesn't know whether it's an experimental or control subject

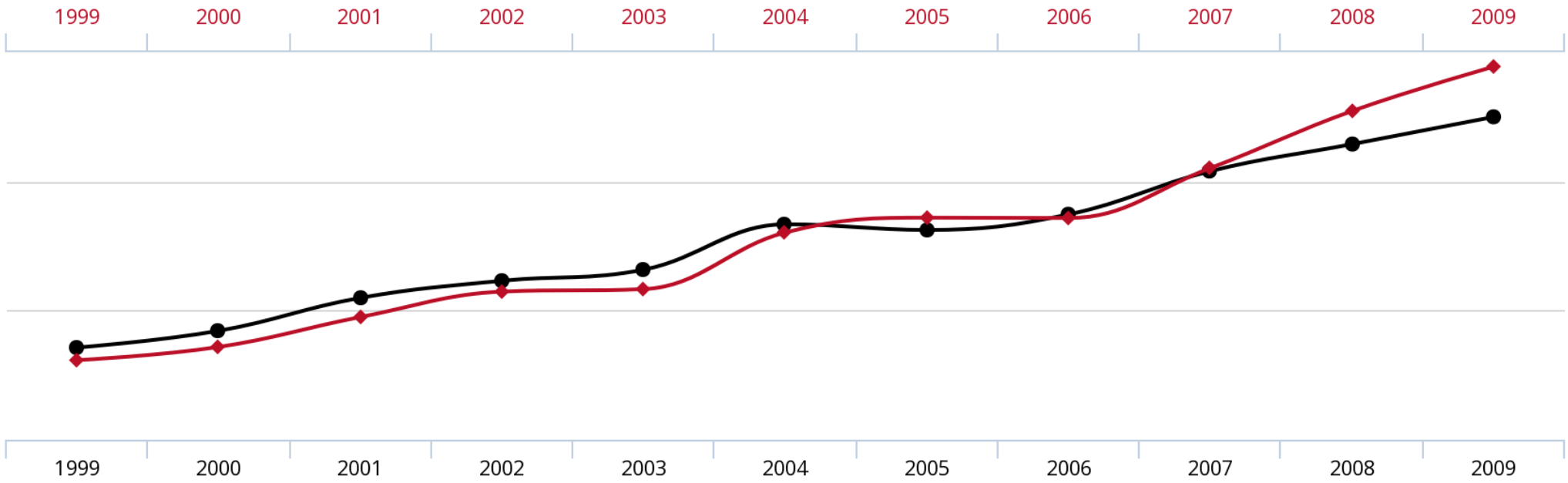
Double-blind: neither the researcher nor subject know which subjects are experimental versus control

How can you apply this to your research?

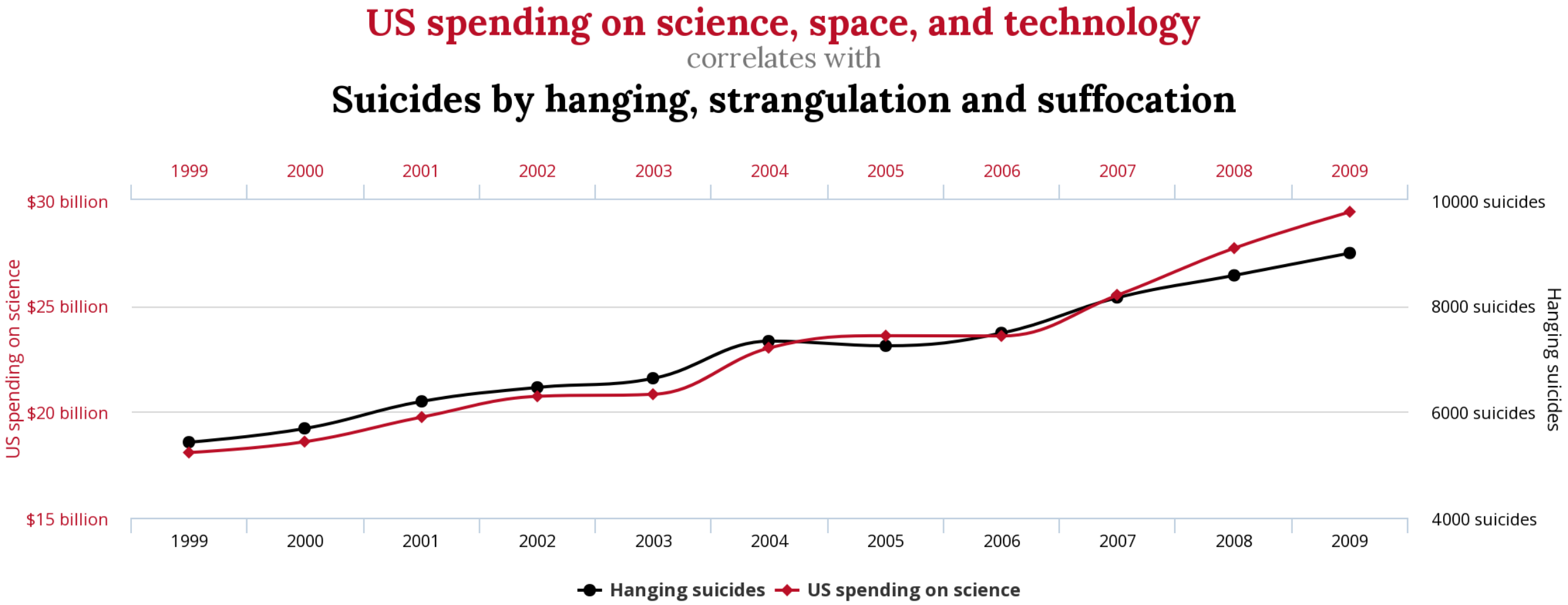
Confounding Variables

1. A difference between groups that the experimenter fails to account for
2. A hidden variable that creates an apparent causal relationship that isn't real
3. **An experiment with confounded variables can be impossible to interpret and impossible to fix**

Confounding Variables



Confounding Variables

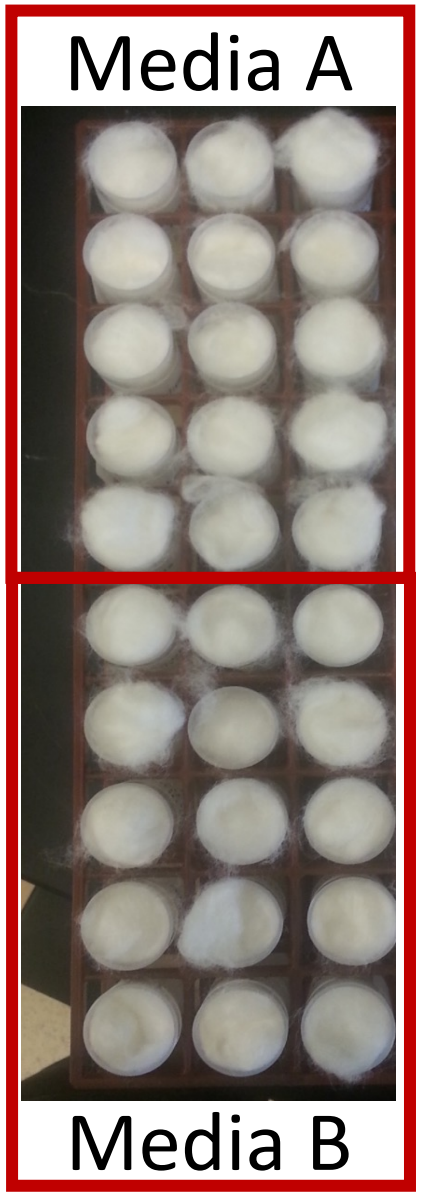


Confounding Example

Study type	Confounding variable
Gene expression level	Tissue used
Diversification	unobserved traits
Lung cancer and coffee	coffee smoking correlation
Behavior	maternal effects
Effective population size	breeding system

Redesign the procedure

- Collect 750 beetles from a population cage.
- Create 30 new vials with 25 beetles each.
- Make the first 15 of these control vials and use food media A.
- Make the next 15 of these experiment vials and use food media B.
- Place in a rack as shown and place in the incubator.
- Measure growth at day 15.



Pseudoreplication

Occurs when the apparent sample size is larger than true sample size

1. 10 rats are studied and tested on three consecutive days, resulting in 15 observations for the control group and 15 observations for the treatment groups
2. The experiment is conducted in two tanks: tank 1 has hormone added, tank 2 is the control tank. 10 fish are tested per tank.
3. We are testing for the effects of mating system on genome size. We use 5 outbreeding insects and 5 inbreeding species of beetles.
4. Beetles are segregated by sex into two vials, with 10 individuals per vial. I draw a male and female at random and test them, returning them to the vials at the end. I perform a total of 40 such tests.

Biological and Technical Replicates

- A biological replicate involves a new, independent test subject
- A technical replicate involves repeating the same procedure on a new sample from the same subject
- Technical replicates do not contribute to your estimates of population-level parameters, but they can increase the precision of measurements on individuals

Which kind of replication

- In general, biological replicates are superior to technical replicates, because biological replicates increase power.

- Technical replicates are useful when the technique in question sometimes produces extremely inaccurate results, which must be pruned from the dataset. An example is qPCR, where occasional extreme outliers are common.

Best Practices

- Ensure as much as possible that controls and experimental individuals are from identical populations (except for the factor of interest)
- Treat your controls as similarly as possible to the experimental subjects (sham injections, placebos, etc.)
- Conduct your control manipulations in parallel with your experimental manipulations
- Think about all possible confounding variables and establish a plan to eliminate or correct for them before you start!

Everything I do is an Experiment

You should approach everything you do in the lab from the perspective of an experiment

Always do the appropriate controls for PCR, transformations, etc.

Troubleshooting is experimenting

Think about how you will describe the experiment before you embark on it

You will see that simplicity is extremely valuable

Think about the analysis you will do before you get started

Terms to know for probability

Sample space: All the potential outcomes of a random trial.

Probability: The proportion of events with a given outcome if the random trial was repeated many times.

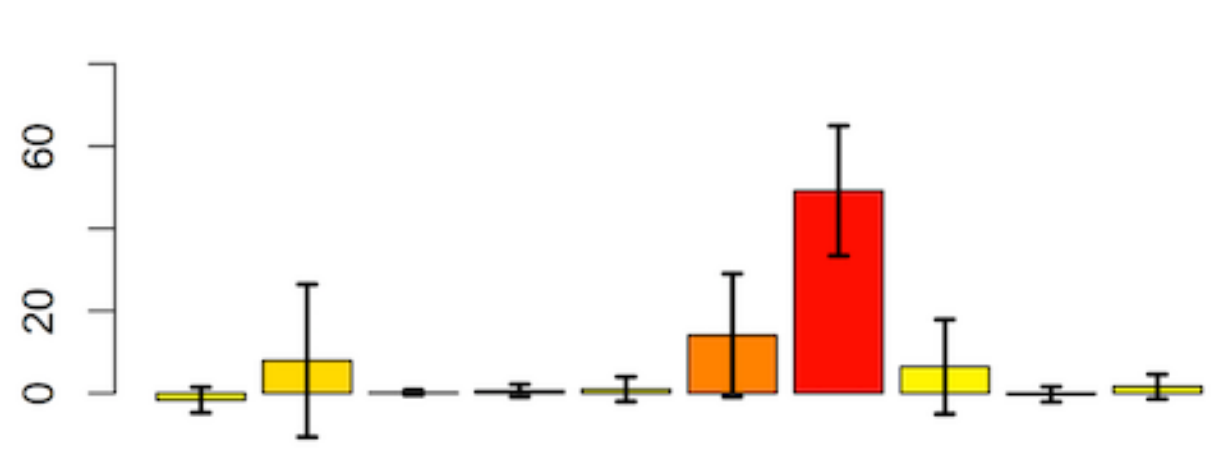
Mutually exclusive: If one outcome excludes the others, they are mutually exclusive.

Conditional probability: The probability of one outcome, if we know that some other outcome occurred.

Independent: When one outcome provides no information about another, they are independent.

Non-Independent: When knowing one outcome changes the probability of another, they are non-independent.

Error bars



- Error bars can be a useful way to show uncertainty when it's not possible to show the actual data points.
- Usually, they represent 1 SE or the 95% CI, but not always.
- **THE FIGURE LEGEND SHOULD INDICATE WHAT THE ERROR BARS REPRESENT!**

For Next Week!

- Have R up and running on a computer you can use!
- Learn any terminology from this lecture that wasn’t a review for you!
- Have a good weekend and watch House of Dragons