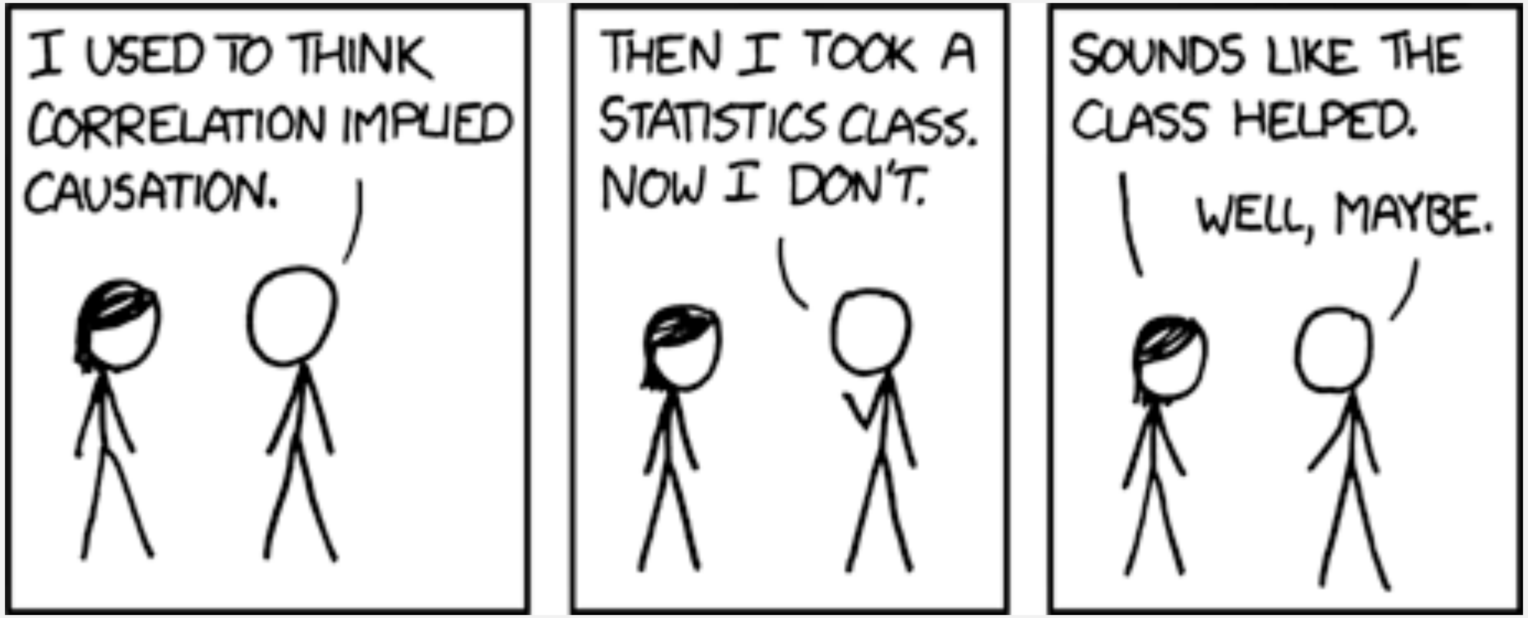
Sampling and Summary Statistics

Biology 683

Lecture 2

Heath Blackmon


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Last week

- Rules and suggestions for plotting.
- What was hardest on the homework?

Today

1. Terminology
2. Summarizing Data
3. Central Limit Theorem

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

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

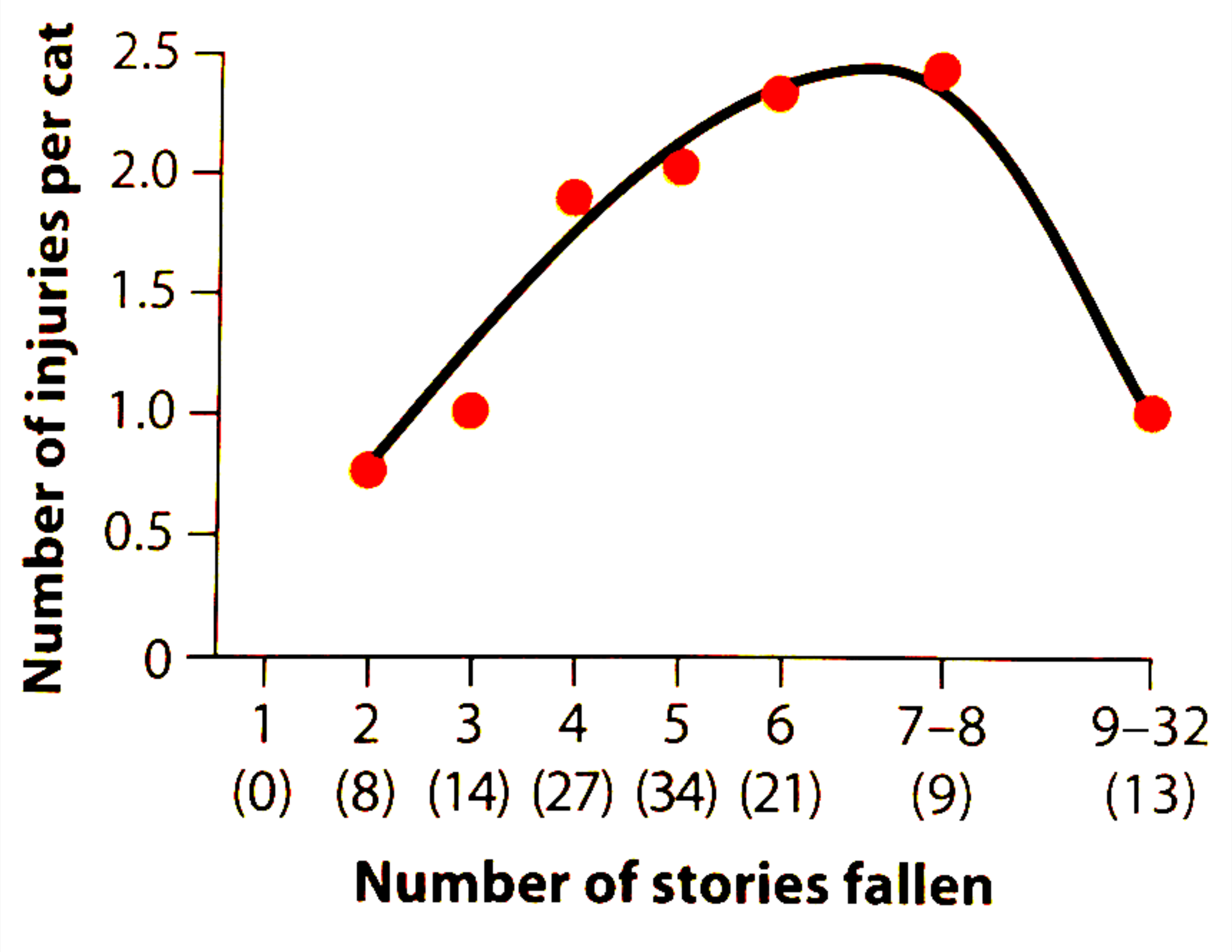
What is the population?

We wanted to examine any association between the severity of injuries, and the height from which cats fall in high-rise buildings.

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?



Sampling Considerations

Target population

- Need to sample a representative population
- A sample of people from College Station, for instance, would probably not be representative of New Yorkers

Sampling Error

- Chance alone will cause your sample to depart from the population

Parameter, estimates, sampling considerations

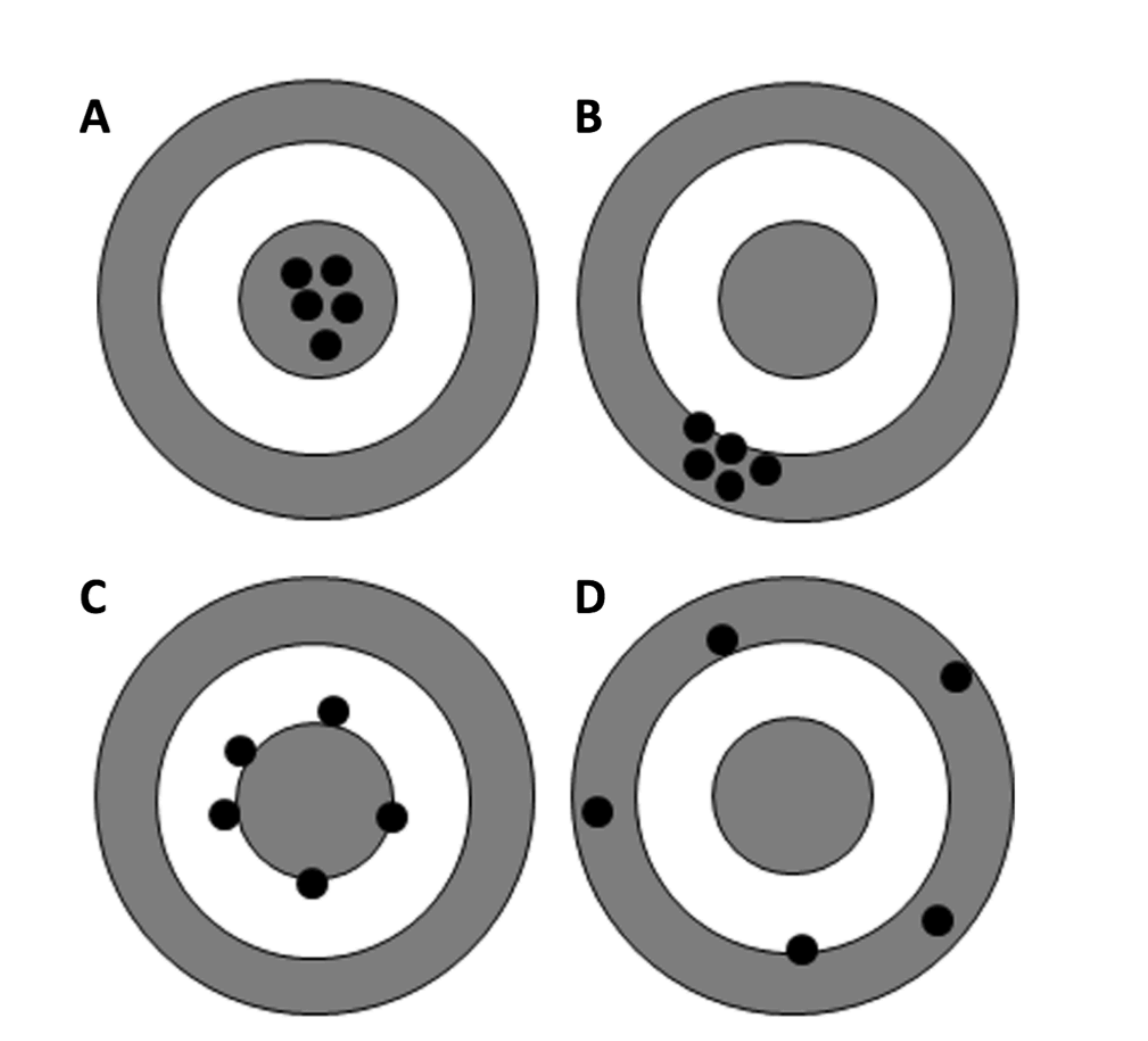
Parameter: Population-level variables we are trying to estimate

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

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

Precision: How tightly grouped are the estimates?

Accuracy vs Precision



- Precision is a measure of spread

- Accuracy is a measure of bias

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. Lots of ways of being non-random…

Data

Variables

The characteristics that differ among individuals

Data

The measurements of variables taken for a sample of individuals

Categorical Variables

Individuals are in qualitative categories

Data

Numerical Variables

Individuals vary on a quantitative scale

Ordinal

The categories can be ordered

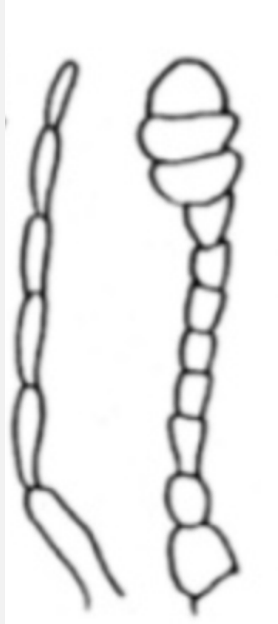
Nominal

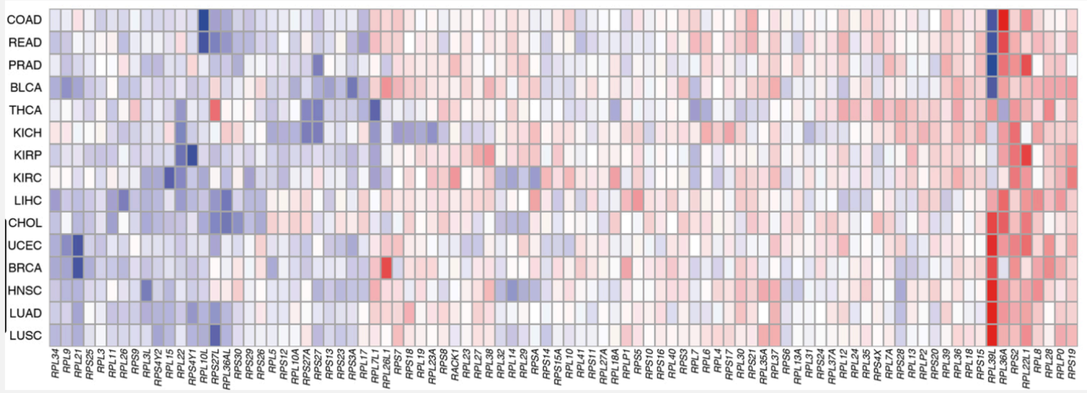
The categories have no inherent order

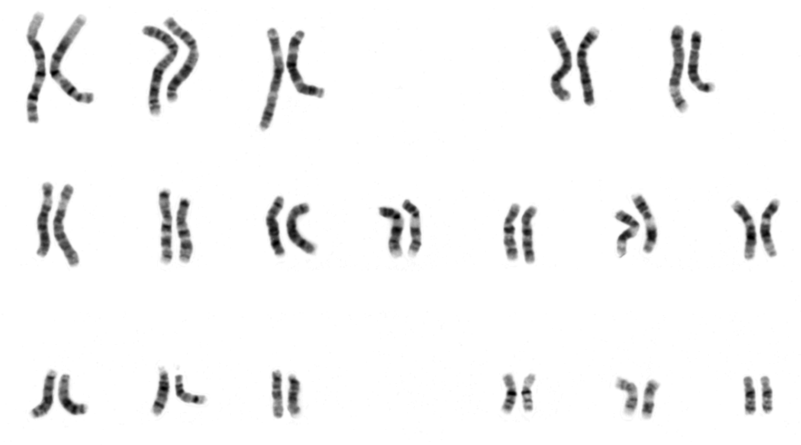
Continuous vs Discrete

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

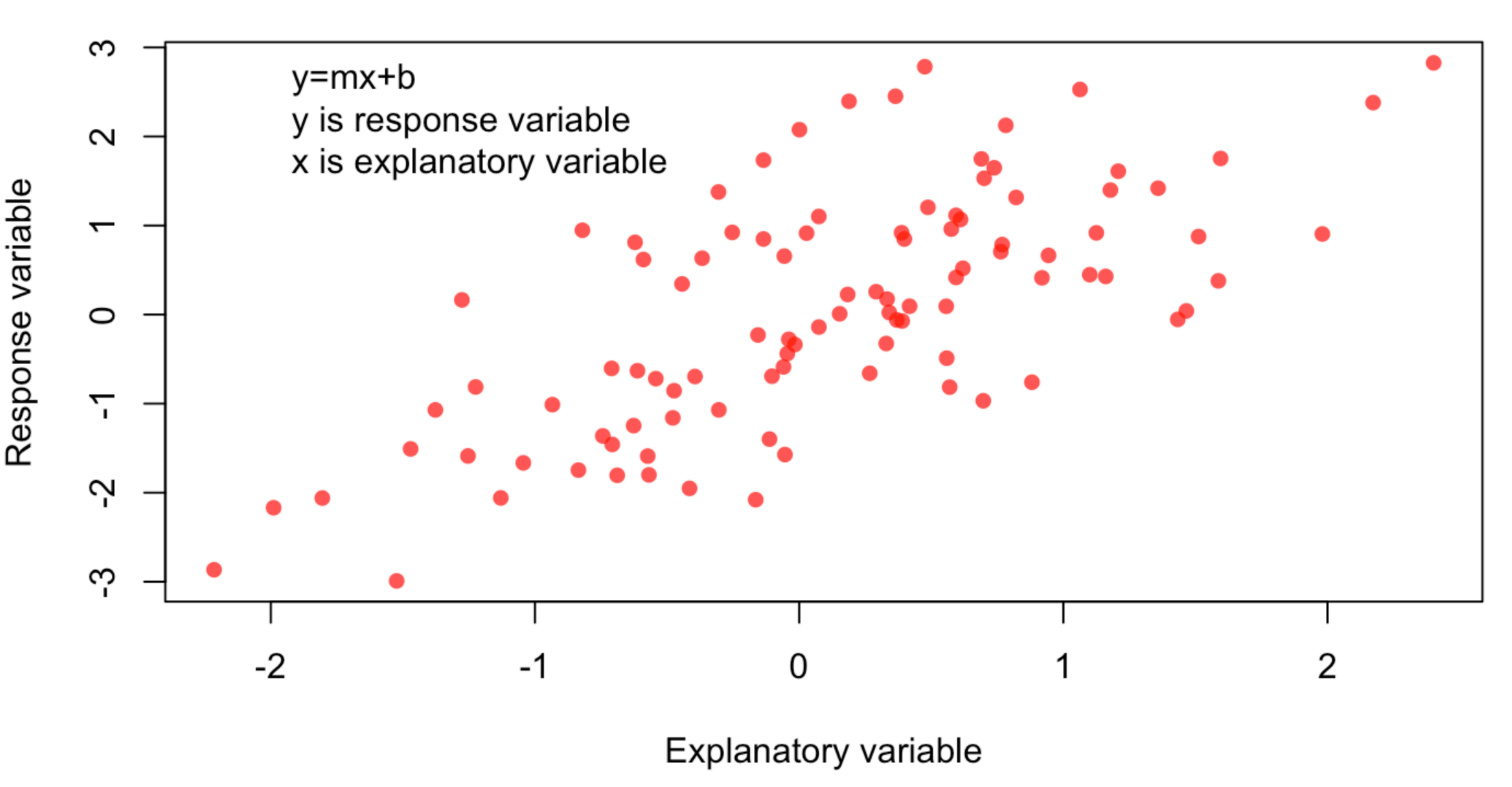
Discrete variables
a variable that has a finite number of possible values







Explanatory and Response Variables



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 : population mean

σ : population standard deviation

Statistic Symbols:

$\bar{Y}$: 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

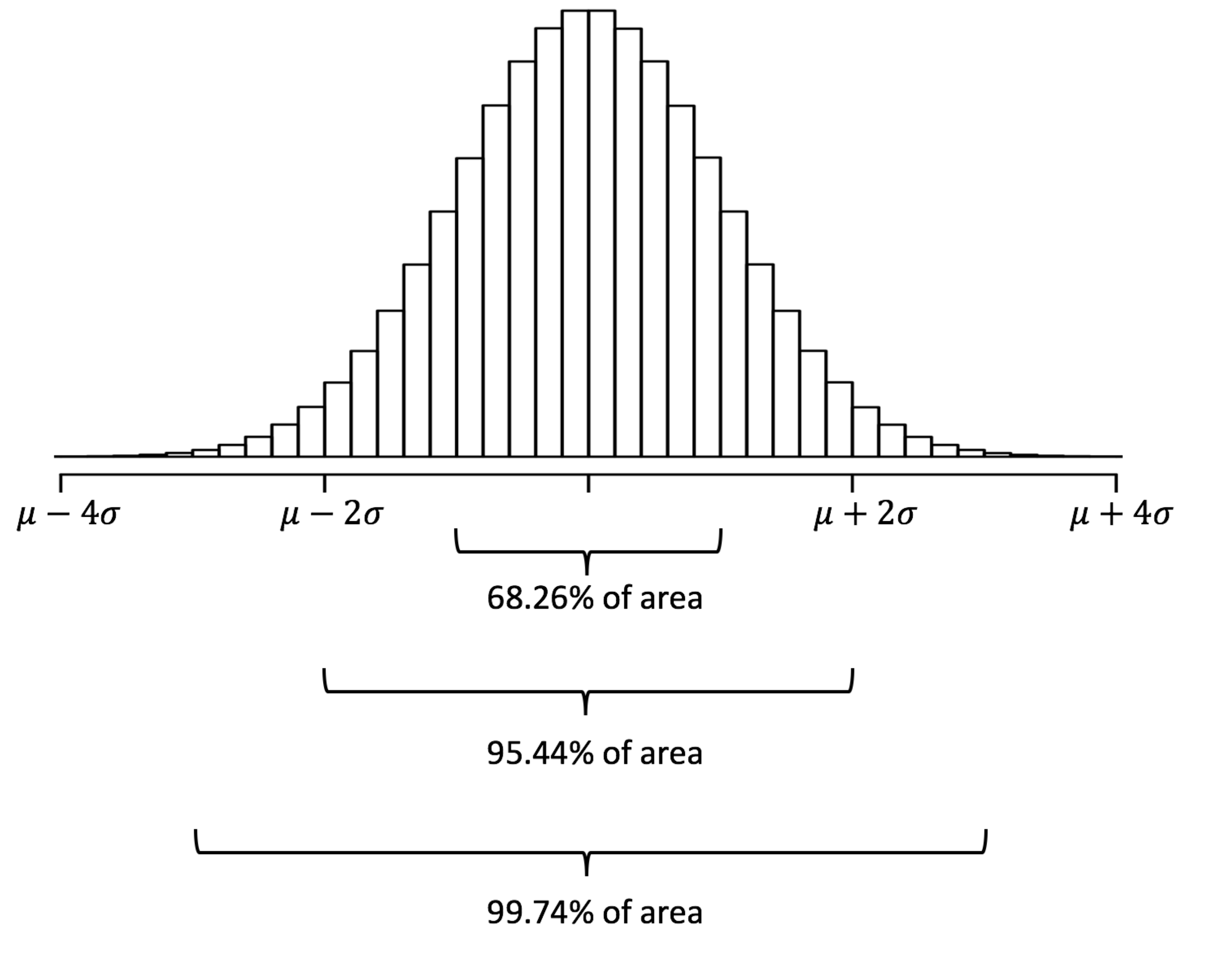
Lets try that

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

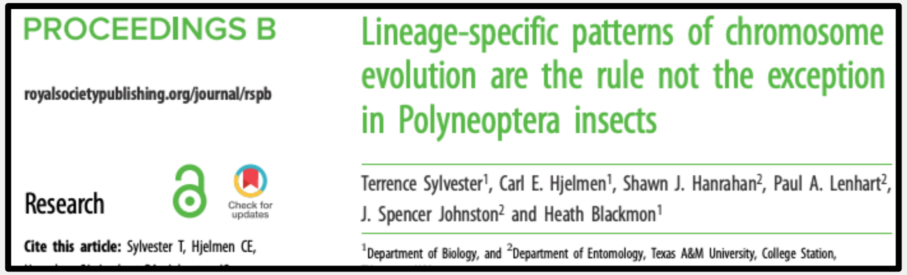
$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

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

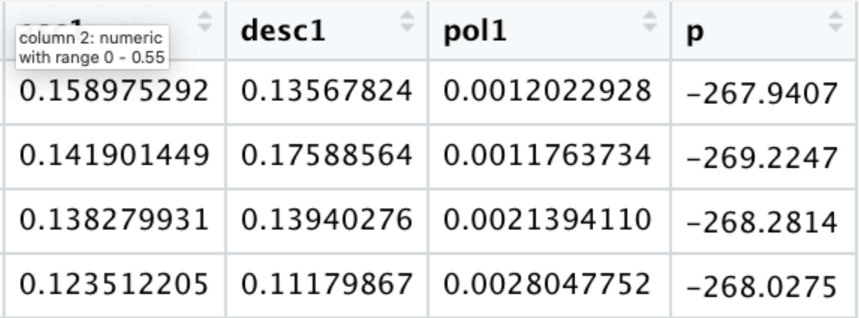
Check for updates

Cite this article: Sylvester T, Hjelmen CE,

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

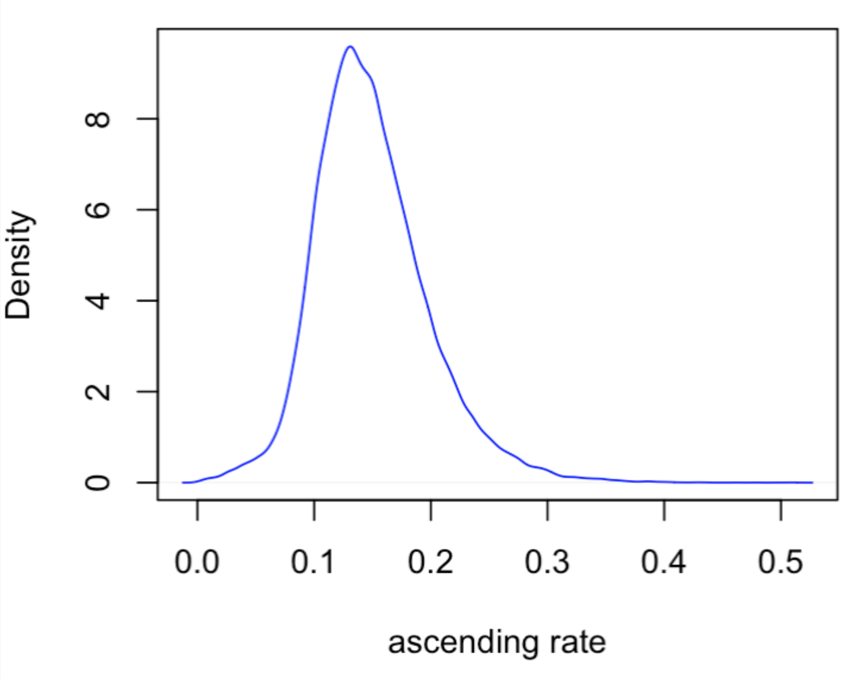
Terrence Sylvester[1], Carl E. Hjelmen[1], Shawn J. Hanrahan[2], Paul A. Lenhart[2], J. Spencer Johnston[2] and Heath Blackmon[1]

[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

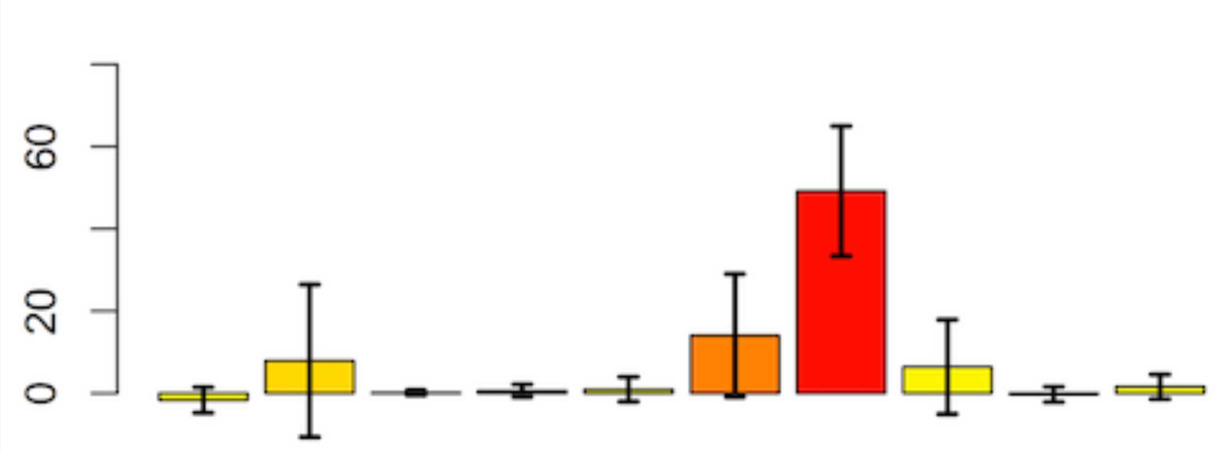
10,000 rows



Frequentist 95% CI
0.149-0.150

95% HPD (credible interval)
0.06-0.26

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!**

Thursday

1. Simulate a populations of 10,000 individuals each with exponential distributions.
2. Record the true population means
3. Take a sample of 100 individuals from population 1.
4. Estimate mean and 95% confidence interval of this estimate
5. Record whether the CI includes true mean
6. Repeat steps 3-5 1000 times
7. Calculate the proportion of samples that included true mean… 95%?

Homework 3

Create a vector of 1000 normally distributed values with a mean of 5.7 and a sd of .2
Q1 true mean

Calculate the mean and confidence interval of a sample of 50 individuals from the population.
Q2 lower end of confidence interval on first try

Calculate the mean and confidence interval of a sample of 50 individuals from the population.
Q3 lower end of confidence interval on second try

Create a new vector of 1000 normally distributed values with a mean of 5.7 and a sd of 2.0
Q4 true mean

Calculate the mean and confidence interval of a sample of 50 individuals from the population.
Q5 upper end of the confidence interval on the first try

Calculate the mean and confidence interval of a sample of 50 individuals from the population.
Q6 upper end of the confidence interval on the second try

Enter answers on blackboard