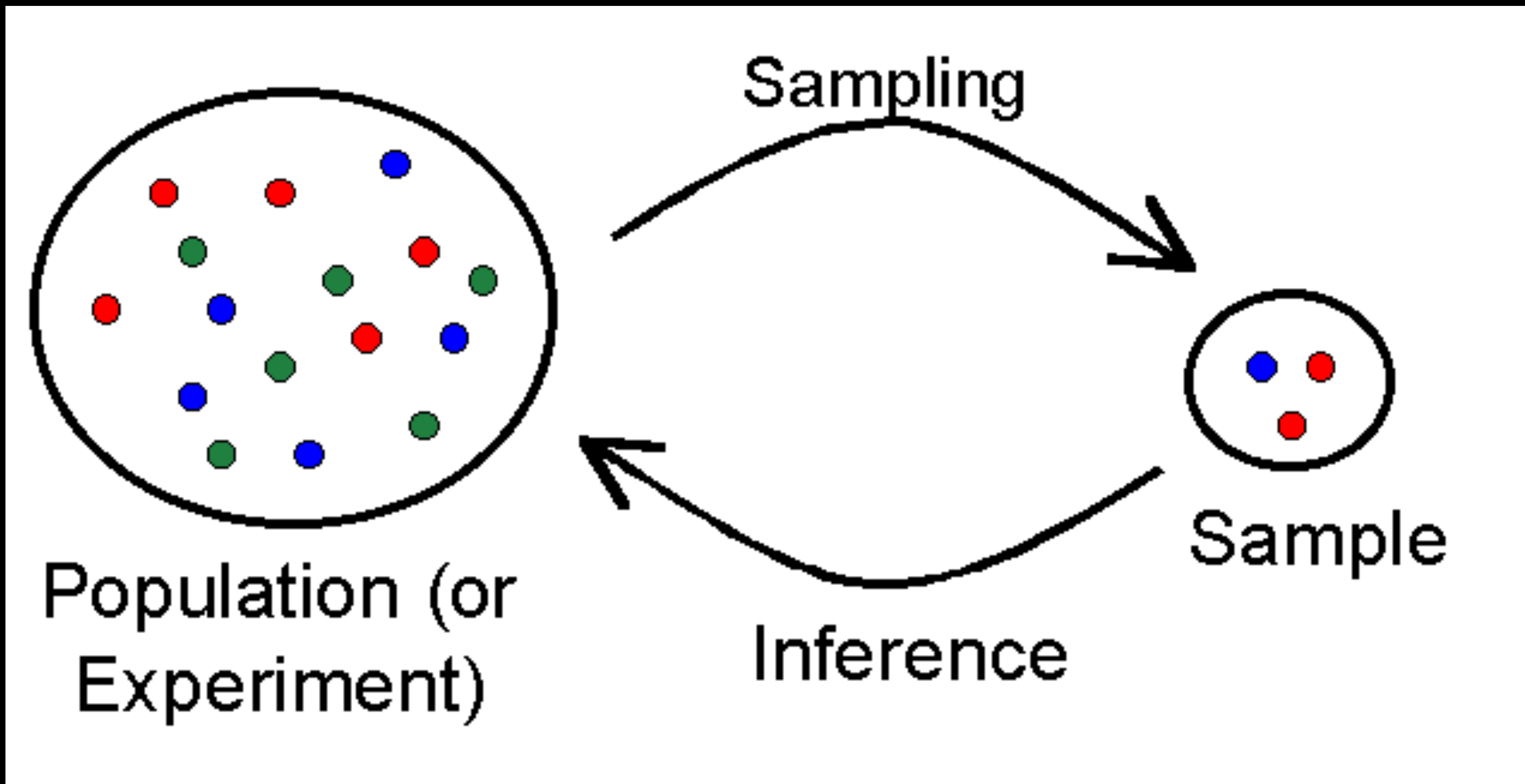
Probability and Statistics Refresher

Biological Statistics Course

Week 2

basic statistics terminology


Sampling
Population (or Experiment)
Sample
Inference

What is statistics?

describing and measuring aspects of nature from samples

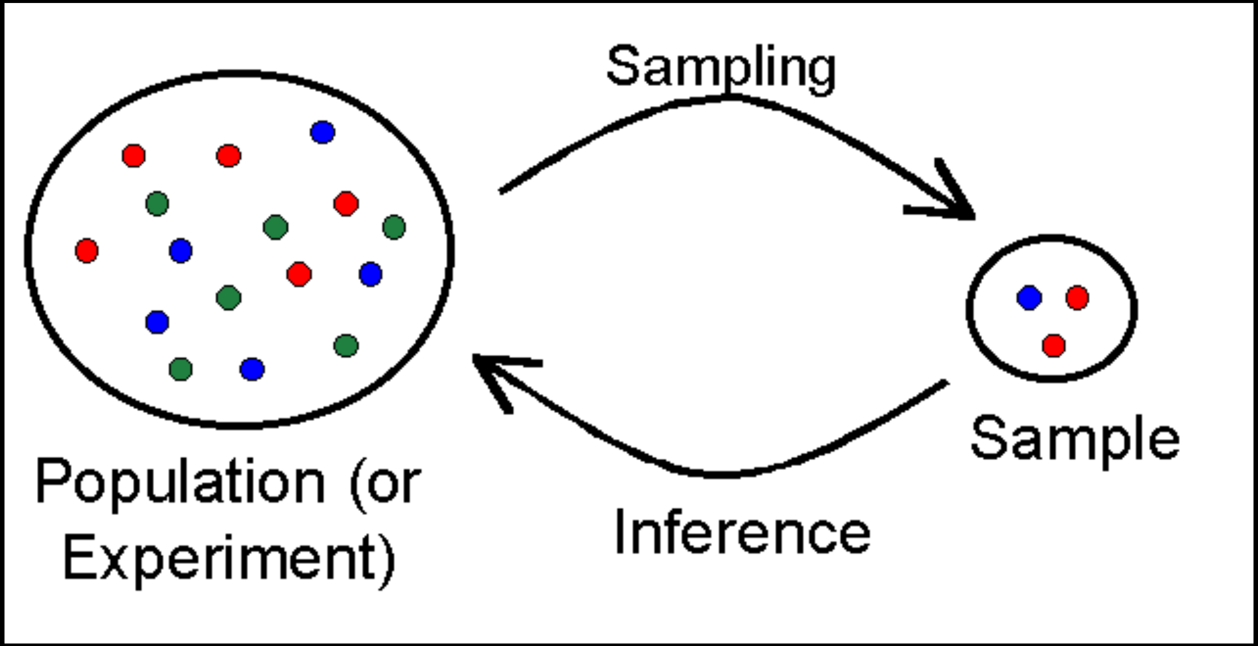
lets us *quantify the uncertainty* of these measurements estimation: process of inferring an unknown quantity of a population using sample data

parameter: quantity describing a population

estimate: quantity calculated from a sample

population: entire collection of individuals or units that a researcher is interested in

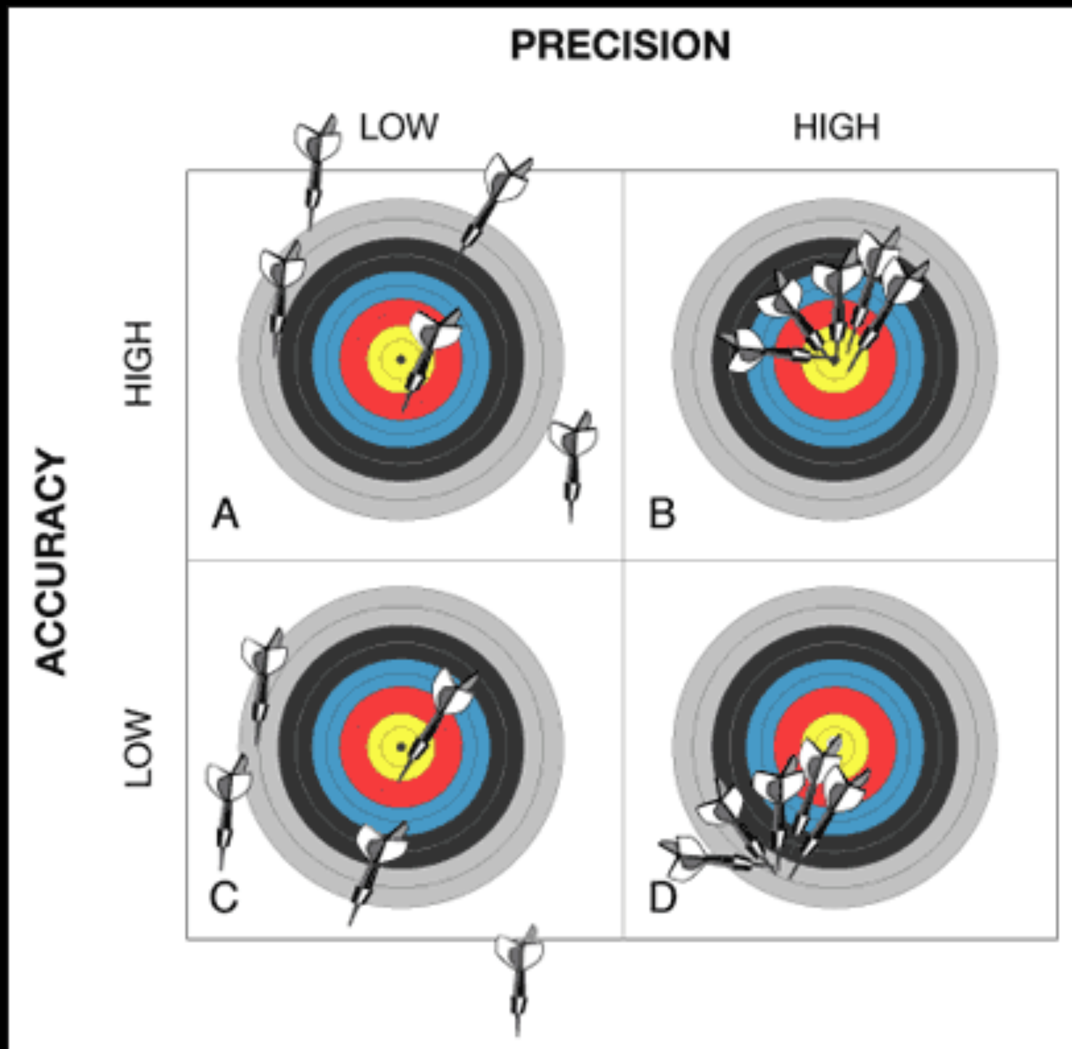
sample: much smaller set of individuals selected from the population



sampling error: chance difference between estimate and population parameter being estimated

bias: systematic discrepancy between estimates and the true population characteristic

properties of good samples



random sampling

every unit in the population has an equal chance of being included in the sample

selection of units must be independent

variable: characteristics that differ from individual to individual

data: raw measurements of one or more variables made on a sample of individuals

types of variables

categorical: characteristics of individuals do not have magnitude on a numerical scale

nominal: different categories have no inherent order

ordinal: categories can be ordered

types of variables

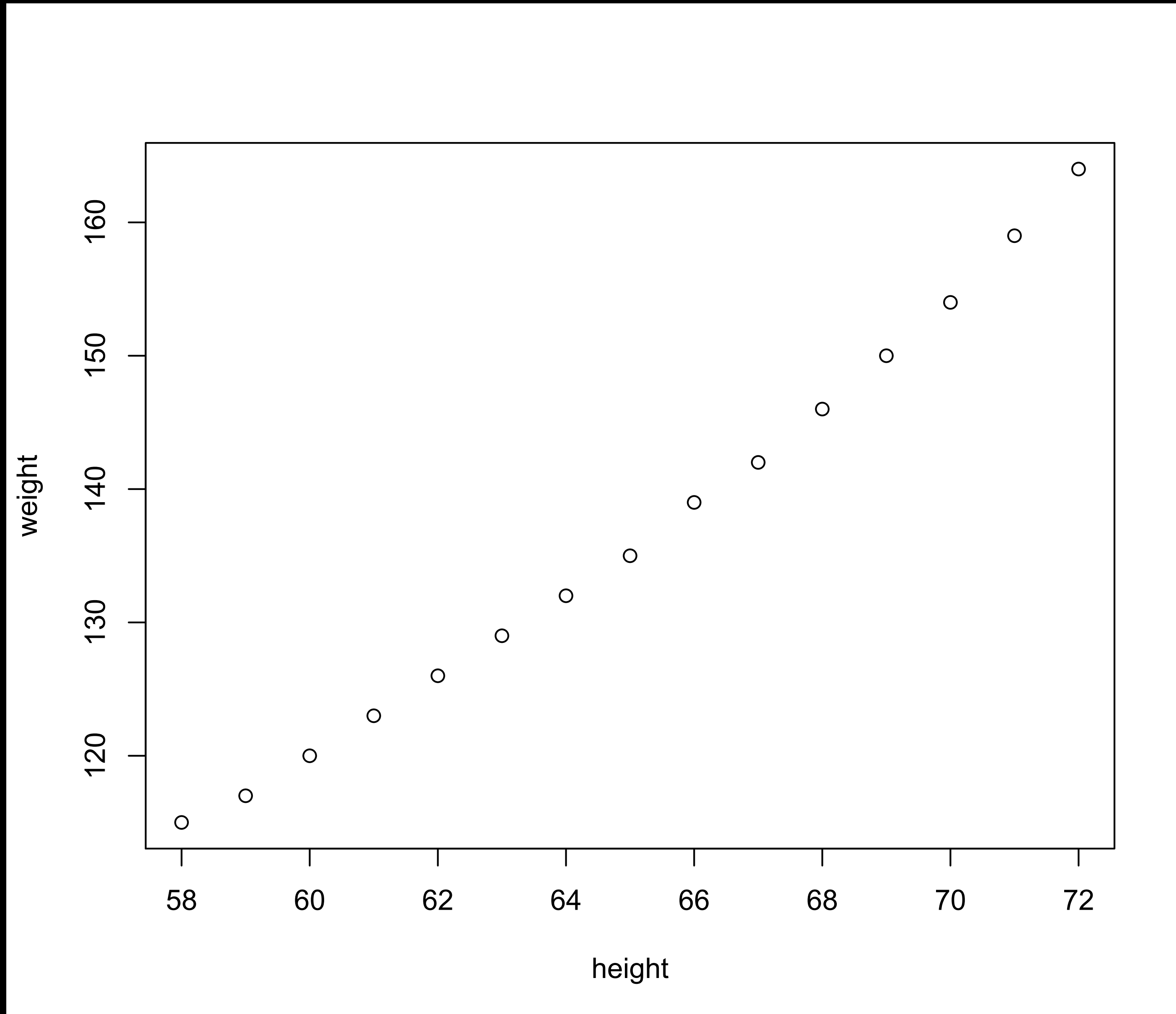
numerical: measurements are quantitative and have magnitude on a numerical scale

continuous: take on any real number value within some range

discrete: take on indivisible units

response (dependent) variable: variable we are trying to predict from the explanatory variable

explanatory (independent) variable: variable we measure to try to determine its relationship with the response variable


weight
120
130
140
150
160
58
60
62
64
66
68
70
72
height

types of studies

experimental: researcher assigns treatments randomly to individuals

observational: assignment of treatments not made by the researcher

experimental studies can determine cause-and-effect relationships between variables whereas observational studies can only point to associations

distributions

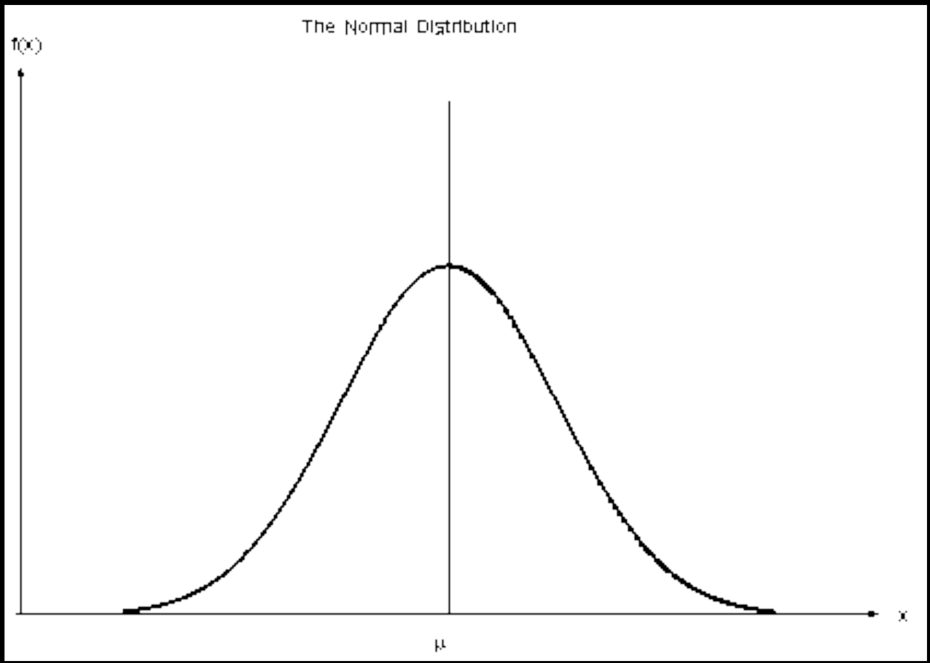
frequency distribution: shows how often each value of the variable occurs in a sample

probability distribution: shows probability of each value of the variable

normal distribution

"bell curve"

most important probability distribution in statistics



R Exercise:
Normal Distribution

R Exercise:
Frequency Distribution

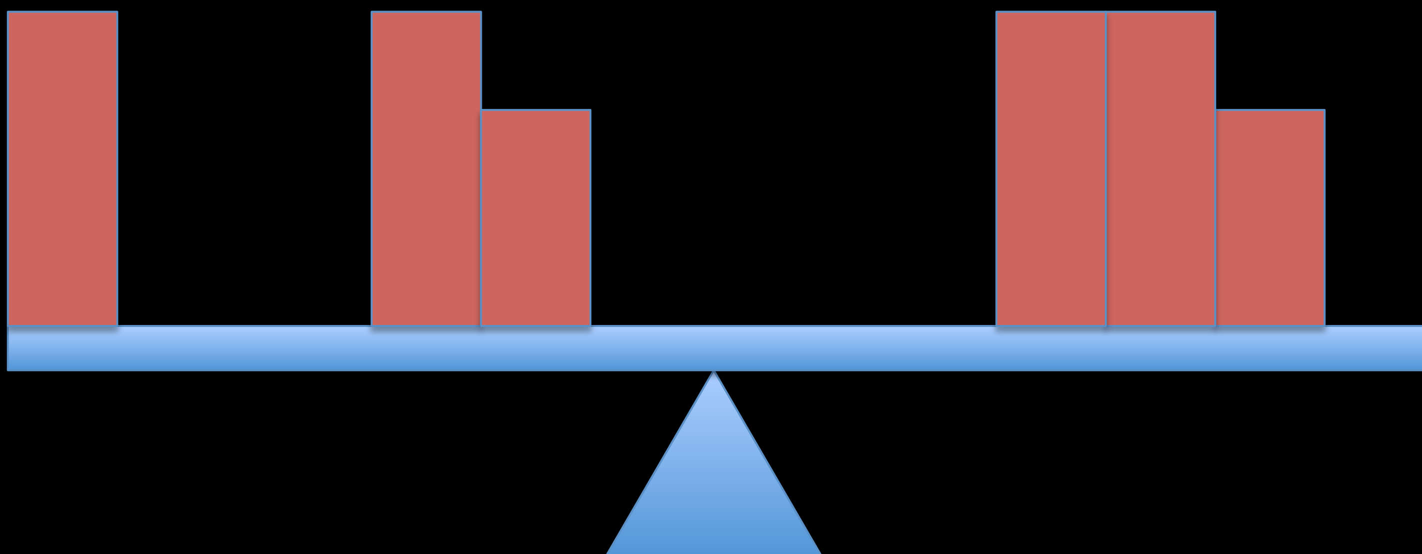
describing data

descriptive statistics

location: tells us something about the average or typical individual

spread: tells us how variable the measurements are from individual to individual

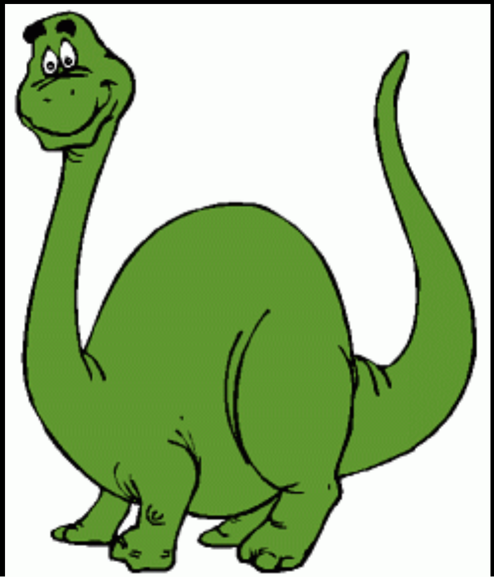
proportion: measures fraction of observations in a given category sample mean: sum of all the observations in a sample divided by the number of observations



standard deviation: common measure of the spread of a distribution, measures how different measurements typically are from the mean

the standard deviation is the square root of the variance coefficient of variation: standard deviation expressed as a percentage of the mean

good for when you want to look at relative variation

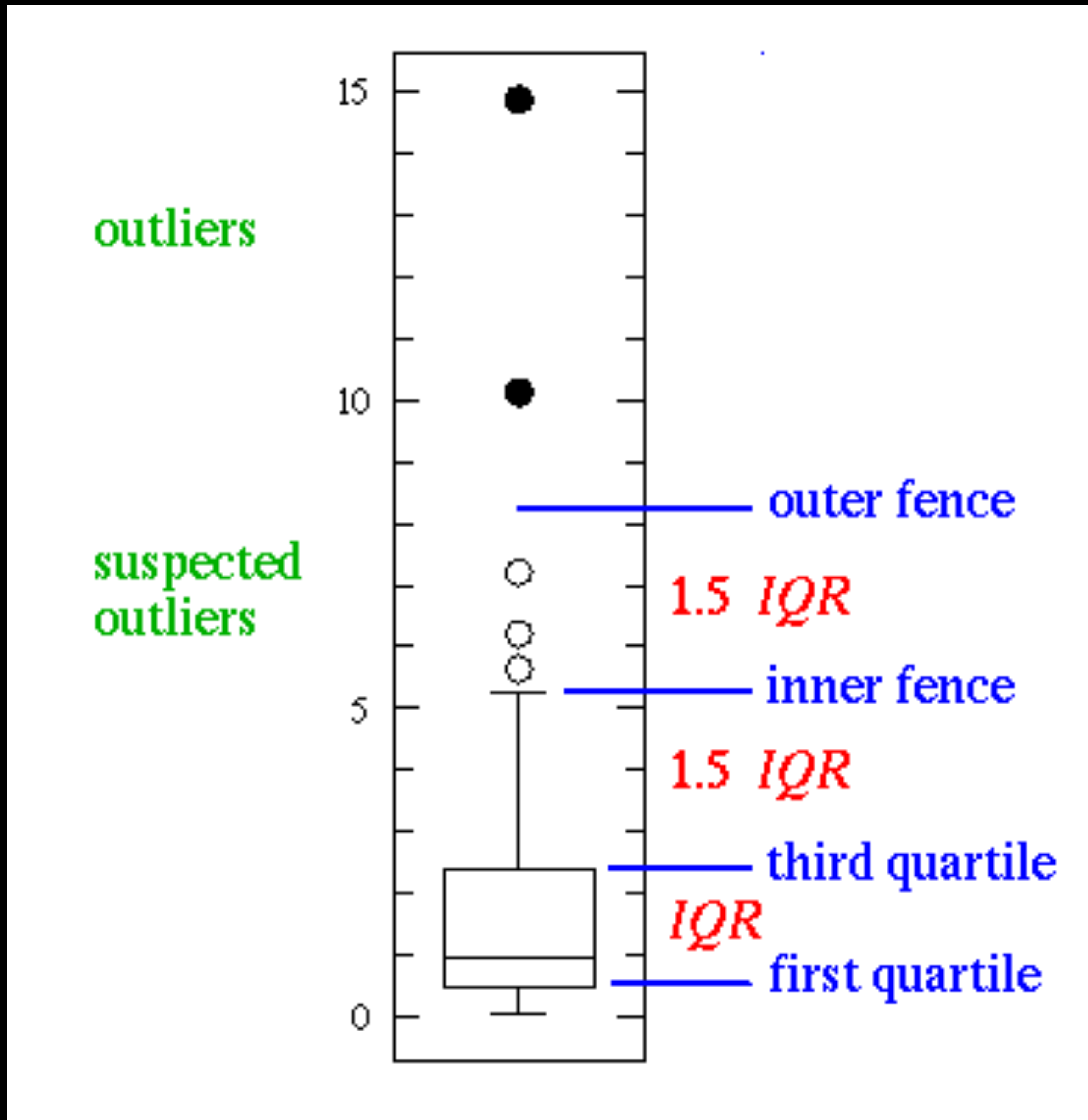


median: middle observation in a set of data

quantile: values that partition the data into quarters

interquantile range: difference between first and third quantiles of the data

R Exercise:
Summary Statistics



outliers
suspected
outliers
15
10
5
0
outer fence
1.5 IQR
inner fence
1.5 IQR
third quartile
IQR
first quartile

probability

random trial: process or experiment that has two or more possible outcomes whose occurrence cannot be predicted

sample space: list of all possible outcomes of a random trial

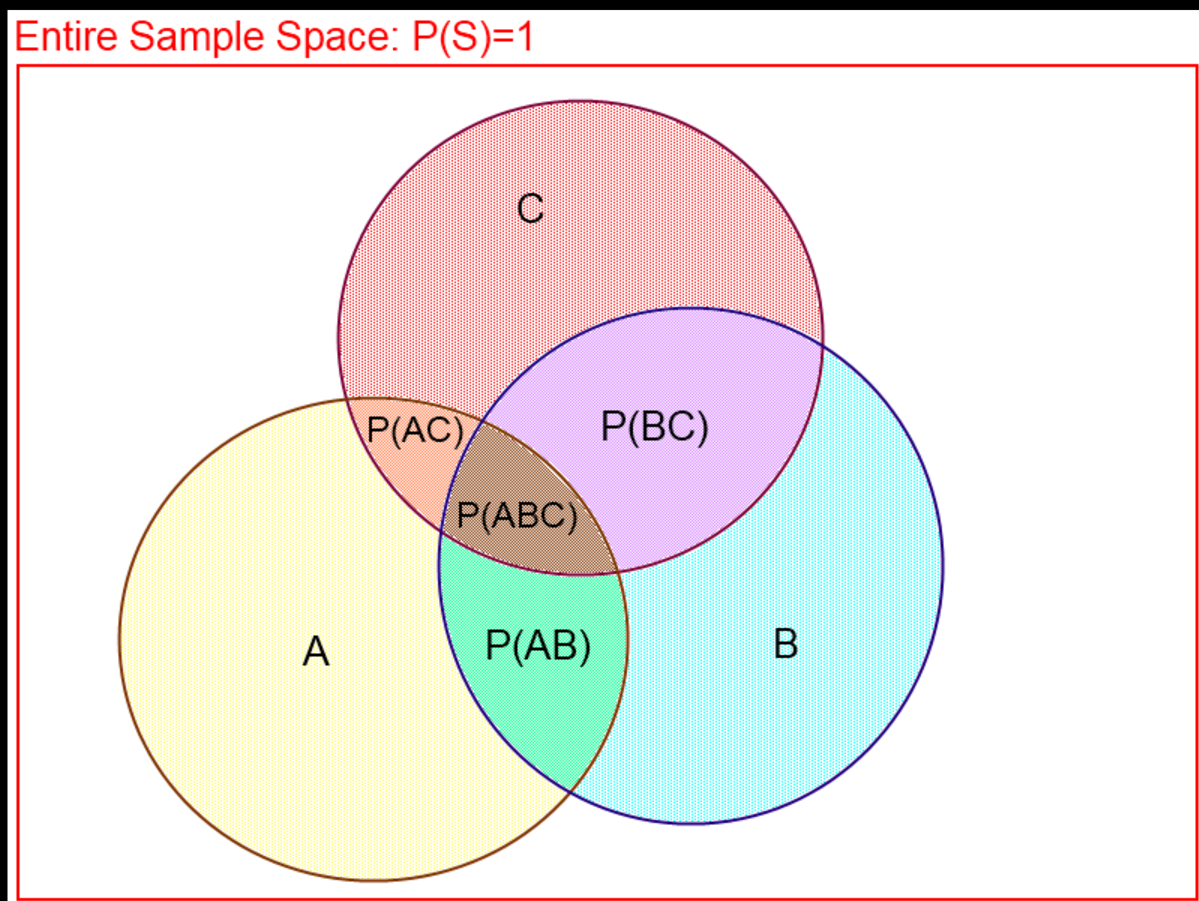
event: any potential subset of the sample space the probability of an event is the proportion of times the event would occur if we repeated a random trial over and over again under the same conditions

Pr[A] means “the probability of event A”

Some people do P[A] or P(A)

Venn diagrams

useful way to think about probabilities



two events are mutually exclusive if they cannot both occur simultaneously

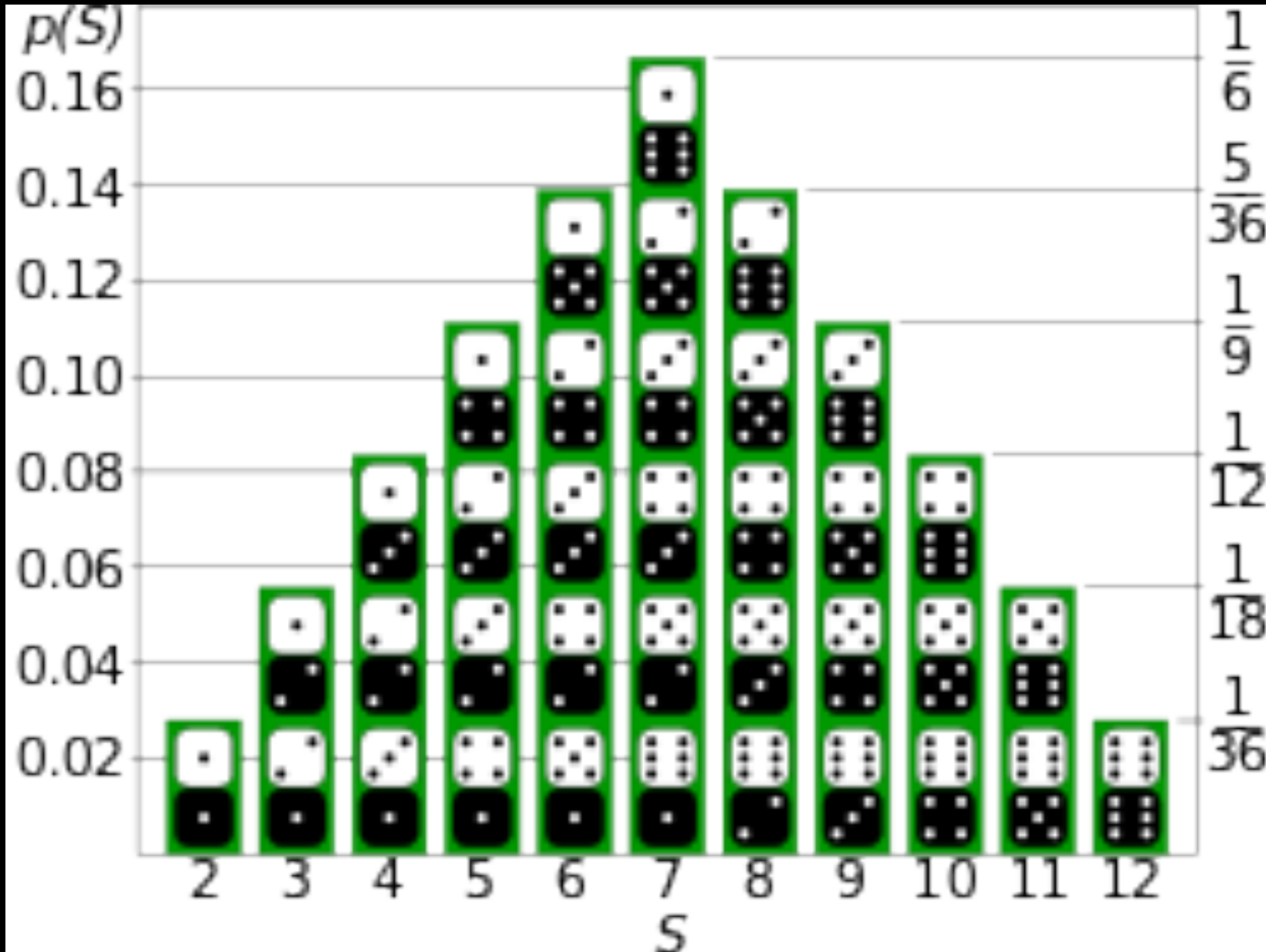
Pr[A and B]=0

probability distribution: list of the probabilities of all mutually exclusive outcomes of a random trial

discrete probability distributions

height at that point is equal to probability of that outcome

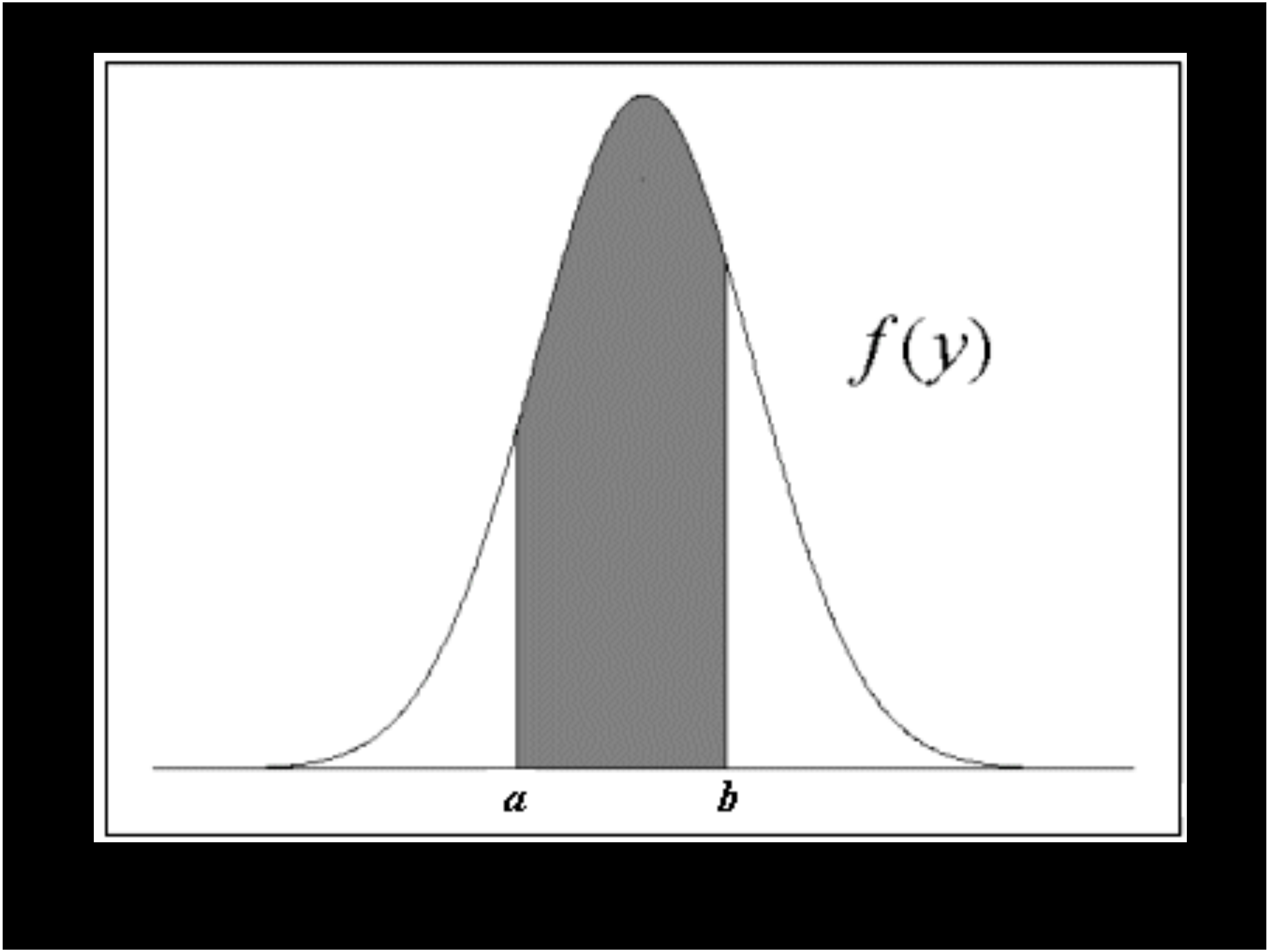
by definition, the sum of all of the probabilities in the distribution should be 1


p(S)
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
1/6
5/36
1/9
1/12
1/18
1/36
2 3 4 5 6 7 8 9 10 11 12
S

continuous probability distribution

height at that point is not the probability of that occurring

makes more sense to talk about probabilities of ranges


f(y)
a
b

we'll talk more next time about special types of distributions

adding mutually exclusive probabilities

if two events A and B are mutually exclusive, then

Pr[A or B]=Pr[A]+Pr[B]

extends to more than two events as long as they are all mutually exclusive the probabilities of all possible mutually exclusive events add to one

Pr[not A]=1-Pr[A]

general addition property

for not mutually exclusive events

Pr[A or B]=Pr[A]+Pr[B]-Pr[A and B]

independence

two events are independent if the occurrence of one does not change the probability that the second will occur

two events are dependent if the probability of one event depends on the result of another event

multiplication rule

If two events A and B are independent

Pr[A and B]=Pr[A] x Pr[B]

applies to more than two events as well

and versus *or*

if you would use *or* in the sentence, add

Pr[A or B]=Pr[A]+Pr[B]
(if A and B are mutually exclusive)

if you would use *and* in the sentence, multiply

Pr[A and B]=Pr[A]xPr[B]
(if A and B are independent)

conditional probability: probability of that event occurring given that a condition is met

$Pr[X|Y]$

law of total probability: if we want to know the overall probability of an event, we sum its probability across every possible condition, weighted by the probability of that condition

$Pr[X]=\Sigma Pr[Y]Pr[X|Y]$

general probability rule

probability that both of two events occur, even if the two are dependent

$Pr[A \text{ and } B]=Pr[A]Pr[A|B]$

probability trees

The jewel wasp, *Nasonia vitripennis*,, is a parasite that lays its eggs on the pupae of flies. Larvae emerge, feed on the fly pupae, and emerge as adults. Emerging males and females mate on the spot.

Nasonia females can manipulate the sex of the eggs they lay. On a fresh host, she lays mainly female eggs. If the host has already been parasitized, she lays mainly males.

If the host is not already parasitized, 95% chance the egg is female, 5% chance the egg is male.

If the host is already parasitized, 10% chance the egg is female, 90% chance the egg is male.

Let's look at the Venn diagram and the probability tree.

Bayes' theorem

powerful mathematical relationship about conditional probability

$$Pr[A|B]=Pr[B|A]Pr[A]/Pr[B]$$

hypothesis testing

hypothesis testing compares data to the expectations of a specific null hypothesis. If the data are too unusual, assuming the null hypothesis is true, then the null hypothesis is rejected

null hypothesis

abbreviated H_o

specific statement about a population parameter made for the purposes of argument

a good null hypothesis is one that would be interesting to reject

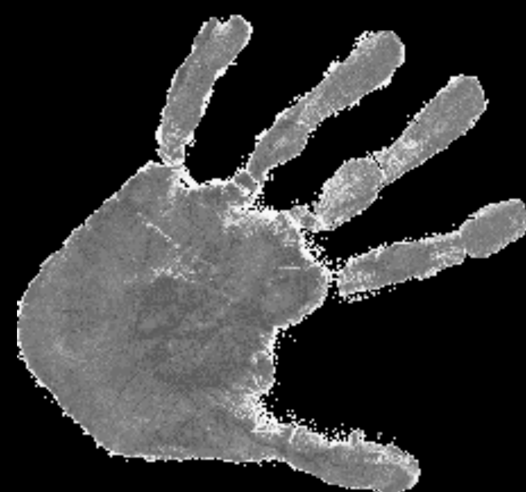
alternative hypothesis

abbreviated H_A

includes all other possible values for the population parameter besides the value states in the null hypothesis an example

Bisazza et al. (1996) tested the possibility of handedness in European toads, *Bufo bufo*, by sampling and measuring 18 toads from the wild.

Of the 18 toads tested, 14 were right-handed and 4 were left-handed. Are these results evidence of a predominance of one type of handedness in toads?



stating the hypothesis

number of interest in population is the proportion that are right-handed

our null hypothesis should be that the two handedness types are equally frequent in the population

H_o: p=0.5

stating the hypothesis

our alternative hypothesis should be that left- and right-handed toads are not equally frequent in the population

H_A: p is not equal to 0.5

this is a two-sided hypothesis because the alternative hypothesis includes values on both sides of the value specified by the null hypothesis

test statistic

quantity calculated from the data that is used to evaluate how compatible the data are with the result expected under the null hypothesis

test statistic

on average, if the null hypothesis were correct, we would expect to observe nine right-handed toads (out of the 18)

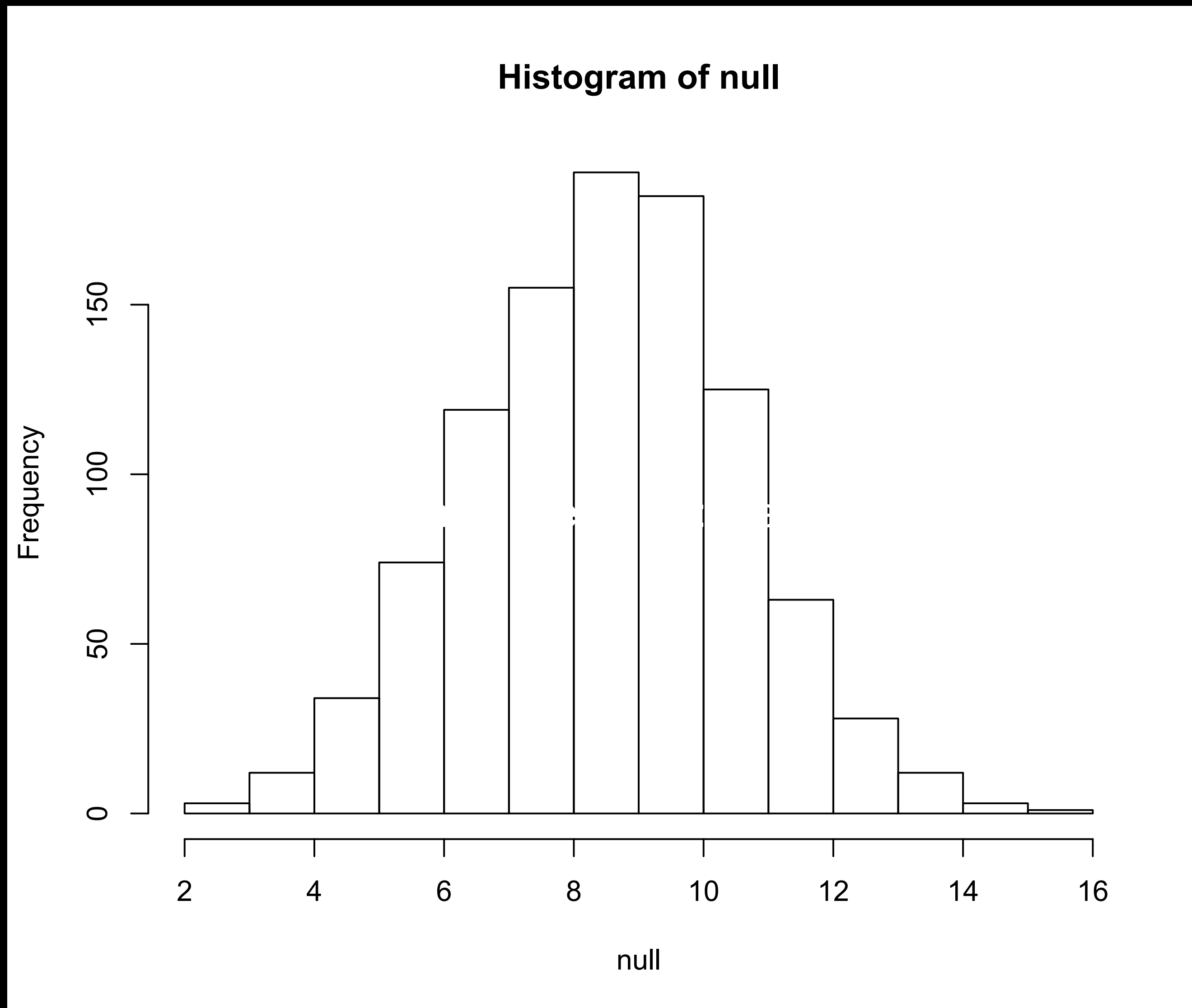
instead, we observed 14 right-handed toads out of the 18 sampled

null distribution

sampling distribution of outcomes for a test statistic under the assumption that the null hypothesis is true

sampling 18 toads under the null hypothesis is like tossing a coin in the air 18 times and counting the number of "heads" that come up (heads=right-handed)

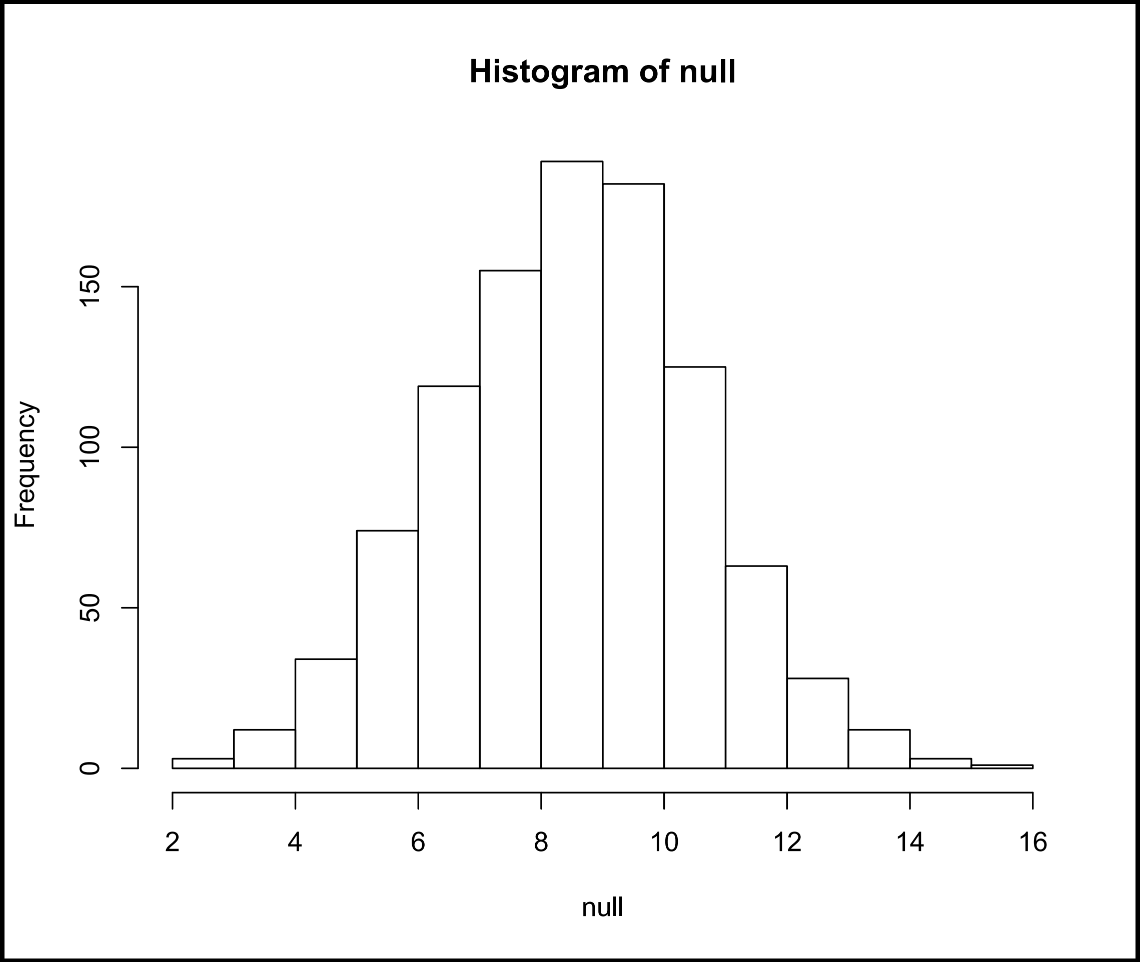
R Exercise:
Generate Null Distribution


Histogram of null
Frequency
0
50
100
150
2
4
6
8
10
12
14
16
null

quantifying uncertainty: the *P*-value

the *P*-value is the probability of obtaining the data (or data showing as great or greater difference from the null hypothesis) if the null hypothesis were true

quantifying uncertainty: the *P*-value



quantifying uncertainty: the *P*-value

don't worry about the calculation of the *P*-value at the moment—we'll get to that next week

our *P*-value is around 0.031

statistical significance

significance level (α): probability used as a criterion for rejecting the null hypothesis. If the *P*-value for a test is less than or equal to α, then the null hypothesis is not rejected

a widely used significance level is $\alpha=0.05$

interpreting non-significant results

can never conclude that the null hypothesis is true

always possible

- true value differs from the null hypothesis by a small amount
- null was not rejected because of chance
- power of the test was limited by sample size

We interpret our results as the data are *compatible* or *consistent* with the null hypothesis.

reporting the results

include the following information in the summery of the results of a statistical test

- value of the test statistic
- the sample size
- the *P*-value

It's also useful to provide confidence intervals, or at least the standard errors, for the parameters of interest

errors in hypothesis testing

Reality		
Decision	**H_0 True**	**H_0 False**
Reject H_0	Type I error	Correct
Do Not Reject H_0	Correct	Type II Error

Type I error is rejecting a true null hypothesis. The significance level α sets the probability of committing a Type I error.

Type II error is failing to reject a false null hypothesis.

The power of a test is the probability that a random sample will lead to rejection of a false null hypothesis

one-sided tests

alternative hypothesis includes parameter values on only one side of the value specified by the null hypothesis

H_o is rejected only if the data depart from it in the direction stated by the H_A

next time:

how to input data in R

proportions and contingency analysis

one sample tests

two sample tests

ANOVA