

# Stat7350: Experimental Design 2 - Examples

*AC Gerstein*

*2019-03-07*

## Learning Objectives

- See how simple biological experiments can lead to important insights (with a statistical controversy thrown in)
- Walk-through of ‘typical’ experiments that test association between two categorical variables
- Walk-through of ‘typical’ experiments that seek to compare population means















## Genetics: Mendel's pea plants

**Genetics:** the branch of biology that deals with heredity and variation of organisms.

In eukaryotic organisms (animals, plants, fungi), chromosomes carry the hereditary information (genes) that are comprised solely of four base pairs, A, C, T, G that combine in different ways to code information.

Gregor Mendel, an Austrian Monk, performed experiments that demonstrated how the law of inheritance works. Before Mendel's experiments, it was thought that traits were passed on through a blending process, where offspring inherited a mix of both parental characteristics.

Mendel looked at seven different traits (phenotypes) from pea plants:

Seed		Flower	Pod		Stem	
Form	Cotyledon	Color	Form	Color	Place	Size
						
Round	Yellow	White	Full	Green	Axial pods	Tall
						
Wrinkled	Green	Violet	Constricted	Yellow	Terminal pods	Short
1	2	3	4	5	6	7

Credit: Rupali Raju Source: CK-12 Foundation

Mendel is credited as the first biologist to use mathematics to quantitatively explain his results. From his pea breeding experiments he predicted.

- the concept of genes as the unit of heredity
- that genes occur in pairs (i.e., there are two alleles that occupy at the same locus (the same position on a strand of DNA) on homologous chromosomes (matching chromosomes) that influence the same trait)
- that one gene of each pair is present in the gametes (i.e., you get one from each parent)

---

## Some terms:

**genotype** - the genetic makeup

**phenotype** - the physical appearance (genotype + environmental effects)

**homozygous** - having the same two alleles of a gene (of position within a gene)

**heterozygous** - having two different alleles

**dominant** - the allele of a gene that masks or suppresses the expression of an alternative allele

**recessive** - an allele that is masked by a dominant allele

---

## Monohybrid cross

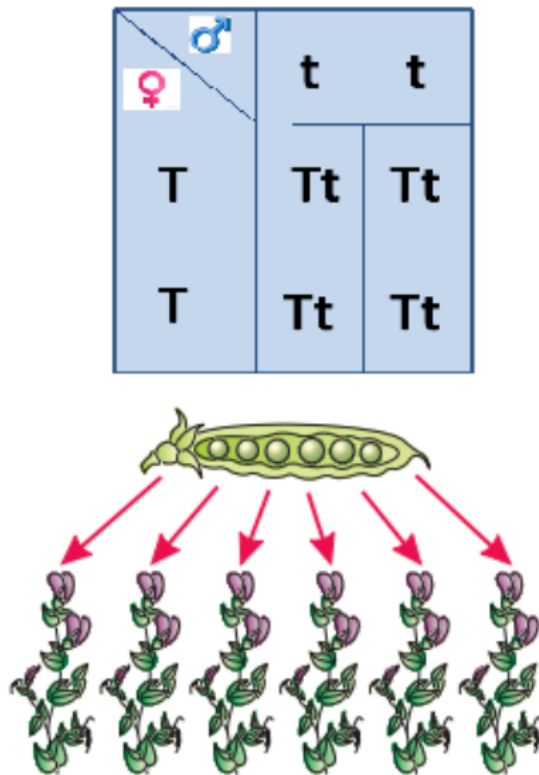
A *monohybrid cross* is a genetic cross involving parents that differ in only a single trait (single gene).

$P$  = Parental generation

$F_1$  = First filial generation; the first set of offspring from a genetic cross

$F_2$  = Second filial generation of a genetic cross

Mendel conducted a monohybrid cross between parents that were tall (genotype:  $TT$ ) and dwarf (Genotype:  $tt$ ). The genotype of all  $F_1$  generation plants is  $Tt$ . Phenotypically, all plants were tall.

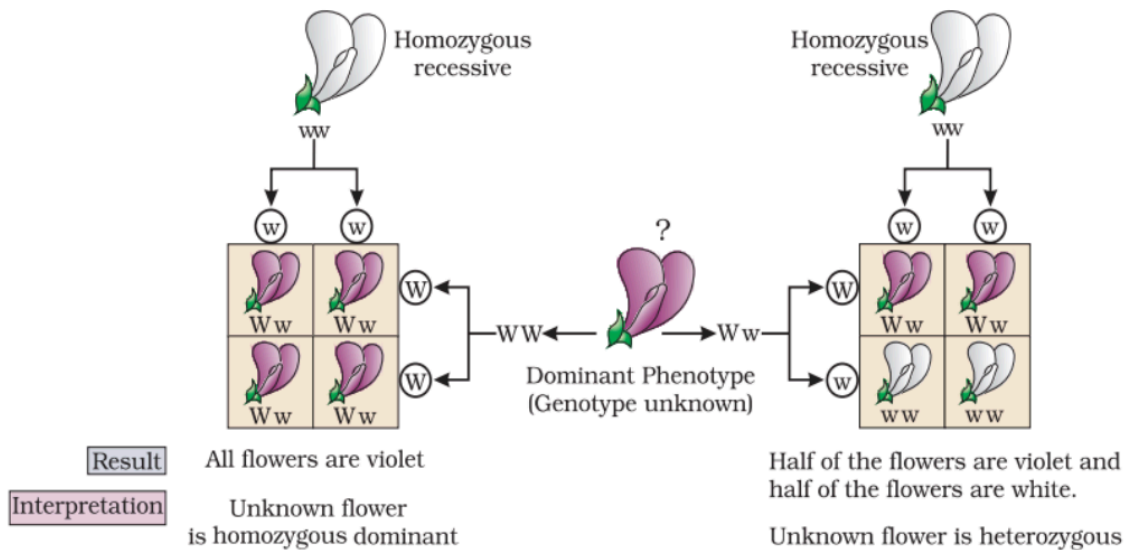


If you let the  $F_1$  generation self-fertilize, the next monohybrid cross is  $Tt * Tt$ .

♀ \ ♂	T	t
	T	t
T	TT	Tt
t	Tt	tt



You can also use this information to perform a ‘test cross’ to elucidate an unknown genotype of one parent.



These experiments led to:

- (1) **the principle of dominance**— one allele is dominant over another  
(2) **the principle of segregation**— that when gametes are formed, each sex cell (e.g., sperm/egg) receives only one copy of each gene.

Traits that follow strict dominance/recessive relationships (one gene, 0/1 phenotype) are referred to as Mendelian. Many traits are not Mendelian, though some diseases (such as cystic fibrosis) are.

## Dihybrid cross

The second type of cross was a **dihybrid cross** involving two traits.

P cross: One parent has a phenotype of round yellow seeds (genotype: RRYy) and the other parent has a phenotype of wrinkled green seeds (genotype: rryy).

F<sub>1</sub> cross: All have a phenotype of round yellow seeds (genotype: RrYy).

---

## CHALLENGE

What is the ratio of genotypes and phenotypes in the offspring of the dihybrid cross?

---

These experiments led Mendel to postulated:

(3) **The principle of independent assortment**— members of one gene pair segregate independently from other gene pairs during gamete formations.

That genes get “shuffled” and thus many combinations are formed is one of the major advantages of sexual reproduction.

The original paper (published in 1866) was initially poorly received (3 citations in the first 35 years) before it was simultaneously rediscovered by multiple researchers. Shortly thereafter Mendel was accused of falsifying his data, by Oxford biologist W. F. R. Weldon and then by R.A. Fisher. as the results were too close to expectation when analyzed statistically.

Many words (an entire book) has been written about this. Here’s a few suggestions for additional reading:

Gregory Radick. Beyond the Mendel-Fisher controversy. Science Vol. 350, Issue 6257, pp. 159-160 (2015) <http://science.sciencemag.org.uml.idm.oclc.org/content/350/6257/159>

Ana M. Pires and João A. Branco. A Statistical Model to Explain the Mendel–Fisher Controversy. Statist. Sci. Volume 25, Number 4, 545-565 (2010). <https://projecteuclid.org/euclid.ss/1300108237>

Additional source:

<https://hemantmore.org.in/science/biology/monohybrid-cross/10676/>

---

## Association tests between categorical variables

### Cancer and Aspirin: 2x2 Contingency Table

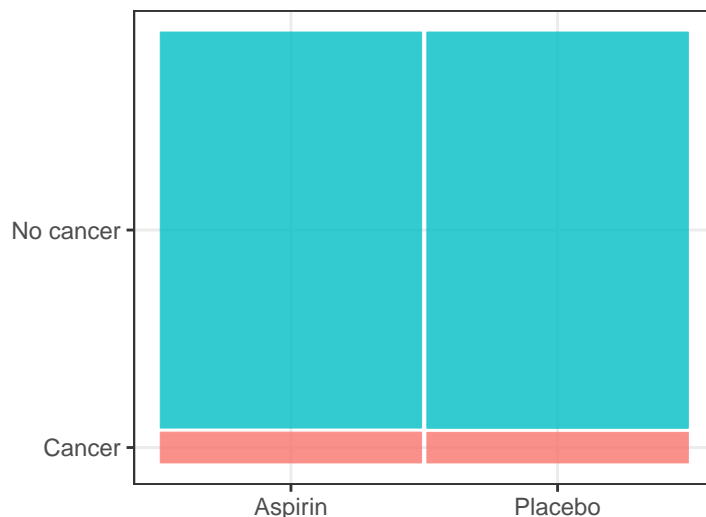
Whitlock & Schluter Example 9.2:

Aspirin has been thought to reduce the risk of stroke and heart attack in susceptible people. An experimental study was designed to test this. 39,876 women were randomly assigned to two treatments: 19,934 received 100 mg of aspirin every other day while 19,942 women received a placebo. The experiment was single-blind: the women did not know which treatment group they were in. During the study, 1438 women on aspirin and 1427 of those on the placebo were diagnosed with cancer.

Source: <https://jamanetwork.com/journals/jama/fullarticle/10.1001/jama.294.1.47> —

```
cancer <- read_csv(url("http://whitlockschluter.zoology.ubc.ca/wp-content/data/chapter09"))

ggplot(data = cancer) +
  geom_mosaic(aes(x = product(cancer, aspirinTreatment), fill=cancer),
             na.rm=TRUE, show.legend = FALSE) +
  xlab("") +
  ylab("") +
  theme_bw()
```



```
chiTest <- chisq.test(cancer$cancer, cancer$aspirinTreatment, correct = FALSE)
chiTest
```

##

```
## Pearson's Chi-squared test
##
## data: cancer$cancer and cancer$aspirinTreatment
## X-squared = 0.050383, df = 1,
## p-value = 0.8224
```

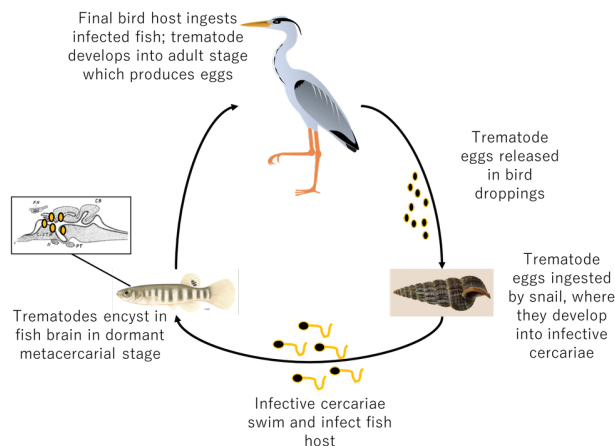
## Parasites

Whitlock & Schluter Example 9.3:

Many parasites have multiple species of hosts that are required for them to complete their life cycle. Trematodes of the species *Euhaplorchis californienis* use three hosts during their life cycle. The worms mature in birds and lay eggs that pass through the bird feces. The horn snail *Cerithidea californica* eats the eggs that hatch and castrate the snail. When an infected snail is eaten by the killifish *Fundulus parviinnis* the parasite further develops and encysts in the fish's brain. Finally, when the killifish is eaten by a bird, the worm becomes a mature adult and starts again.

Lafferty and Morris (1996) tested whether infected fish are more likely to be ingested by a bird than non-infected fish. They set up an experiment: a large, open, outdoor tank was stocked with three types of killifish: unparasitized, lightly parasitized, and heavily infected. Foraging birds were naturally able to eat fish directly from the tank.

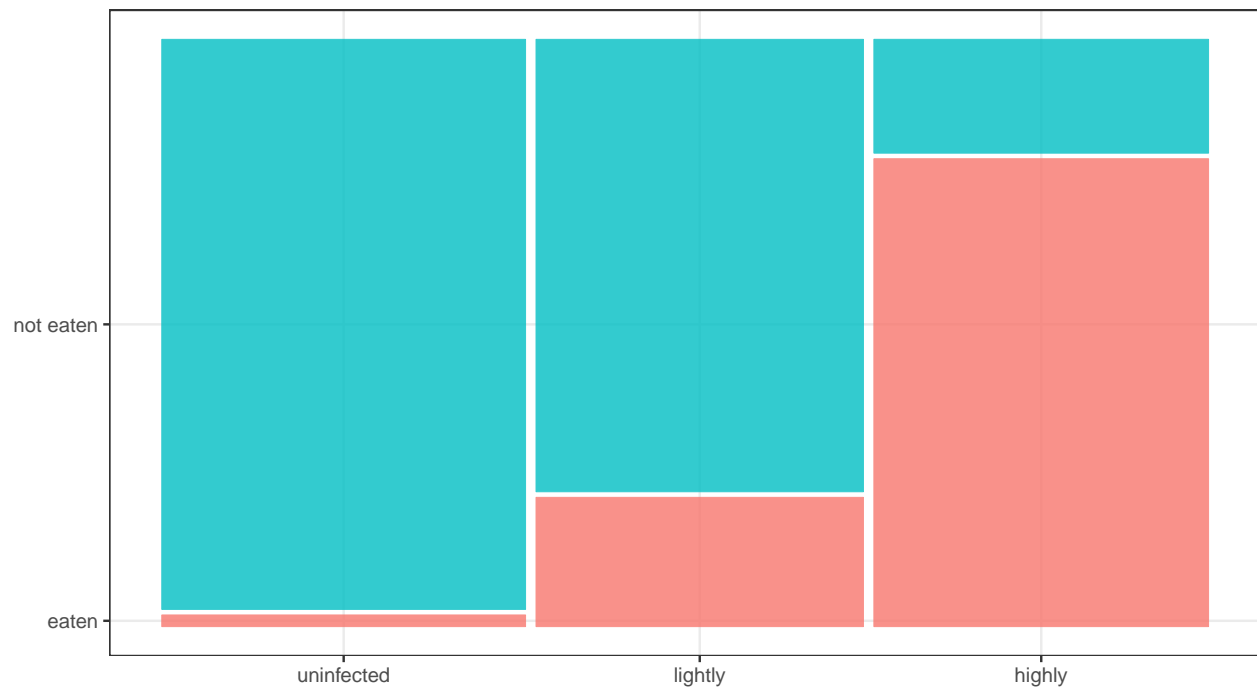
<https://esajournals-onlinelibrary-wiley-com.uml.idm.oclc.org/doi/abs/10.2307/2265536>



Source: <https://theethogram.com/2018/06/12/creature-feature-euhaplorchis-californienis/>

```
worm <- read_csv(url("http://www.zoology.ubc.ca/~schluter/WhitlockSchluter/wp-content/d
worm$infection <- factor(worm$infection,
                        levels = c("uninfected", "lightly", "highly"))
```

```
ggplot(data = worm) +
  geom_mosaic(aes(x = product(fate, infection), fill=fate),
             na.rm=TRUE, show.legend = FALSE) +
  xlab("") +
  ylab("") +
  theme_bw()
```



```
chiTest2 <- chisq.test(worm$fate, worm$infection, correct = FALSE)
chiTest2
```

```
##
##  Pearson's Chi-squared test
##
## data:  worm$fate and worm$infection
## X-squared = 69.756, df = 2, p-value
## = 7.124e-16
```



# Comparing Population Means

## Human body temperature

Whitlock & Schluter example 11.3:

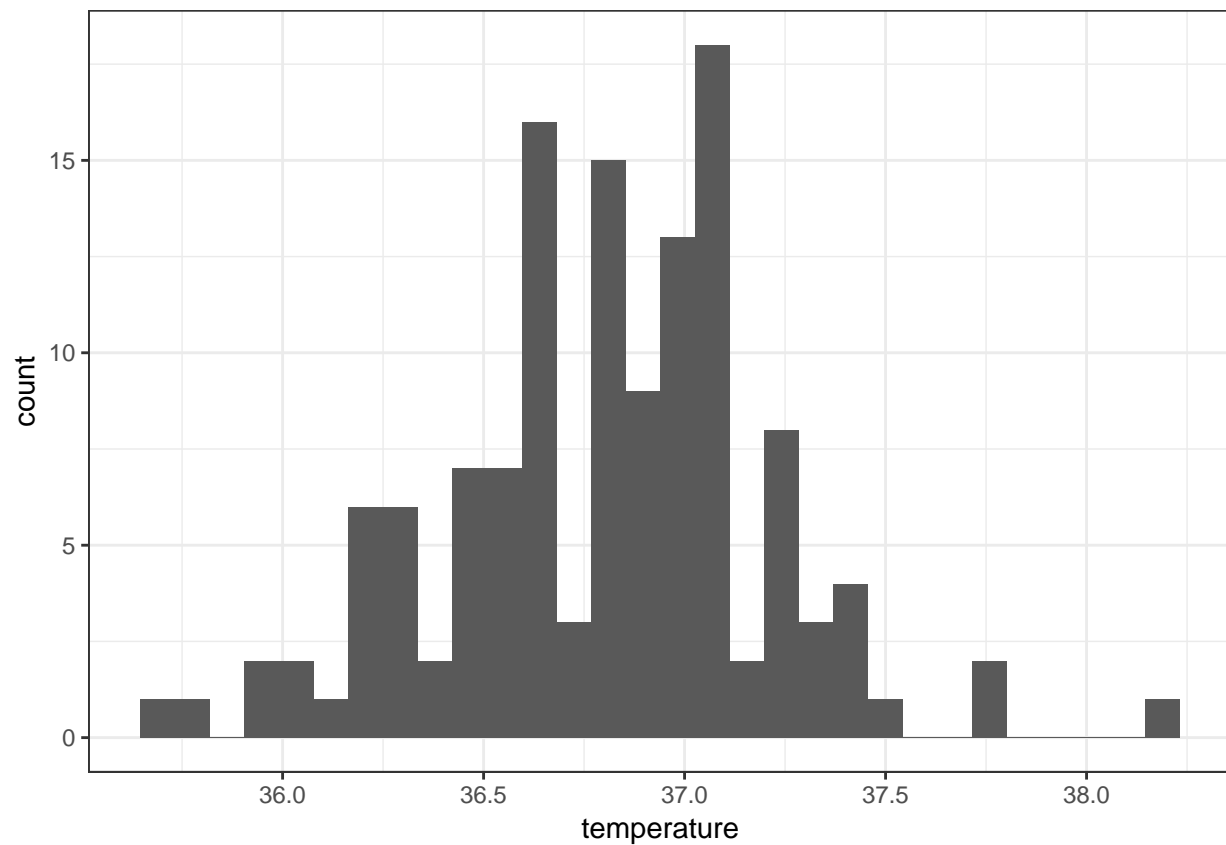
It has often been reported that the average human body temperature is 37 degrees Celsius. This stems from a book published in 1868 by German physician Carl Reinhold August Wunderlich who reported an analysis of over one million temperature readings from 24,000 patients.

In 1996 Allen Shoemaker published data from 130 body temperature readings for 65 males and 65 females. Is this data consistent with a mean body temperature of 37 degrees? <http://jse.amstat.org/v4n2/datasets.shoemaker.html>

```
temp <- read_csv("bodyTemp.txt", col_type = cols())

ggplot(temp, aes(temp)) +
  geom_histogram() +
  theme_bw() +
  xlab("temperature")
```

```
## `stat_bin()` using `bins = 30`. Pick
## better value with `binwidth`.
```



```
t.test(temp$temp, mu = 37)
```

```
##
##  One Sample t-test
##
## data:  temp$temp
## t = -5.4548, df = 129, p-value =
## 2.411e-07
## alternative hypothesis: true mean is not equal to 37
## 95 percent confidence interval:
##  36.73445 36.87581
## sample estimates:
## mean of x
##  36.80513
```

What other information might you like about this dataset?

## Horned lizard spikes

Whitlock & Schluter example 12.3:

The horned lizard *Phrynosoma mcalli* is named for the fringe of spikes surrounding its head. A group of herpetologists recently tested whether the long spikes help protect horned lizards from being eaten. They took advantage of the behaviour of one of the main natural predators, the loggerhead shrike *Lanius ludovicianus*. The loggerhead shrike skewers its victims on thorns or barbed wire to save for future eating.

Young et al (2004) wanted to test whether horn length influenced the likelihood of successful predation. To do this they measured the horn length from 30 horned lizards they found that had been killed by shrikes. For comparison, they measured the horn length on 154 horned lizards that were still alive in the same area.

<http://science.sciencemag.org.uml.idm.oclc.org/content/304/5667/65>

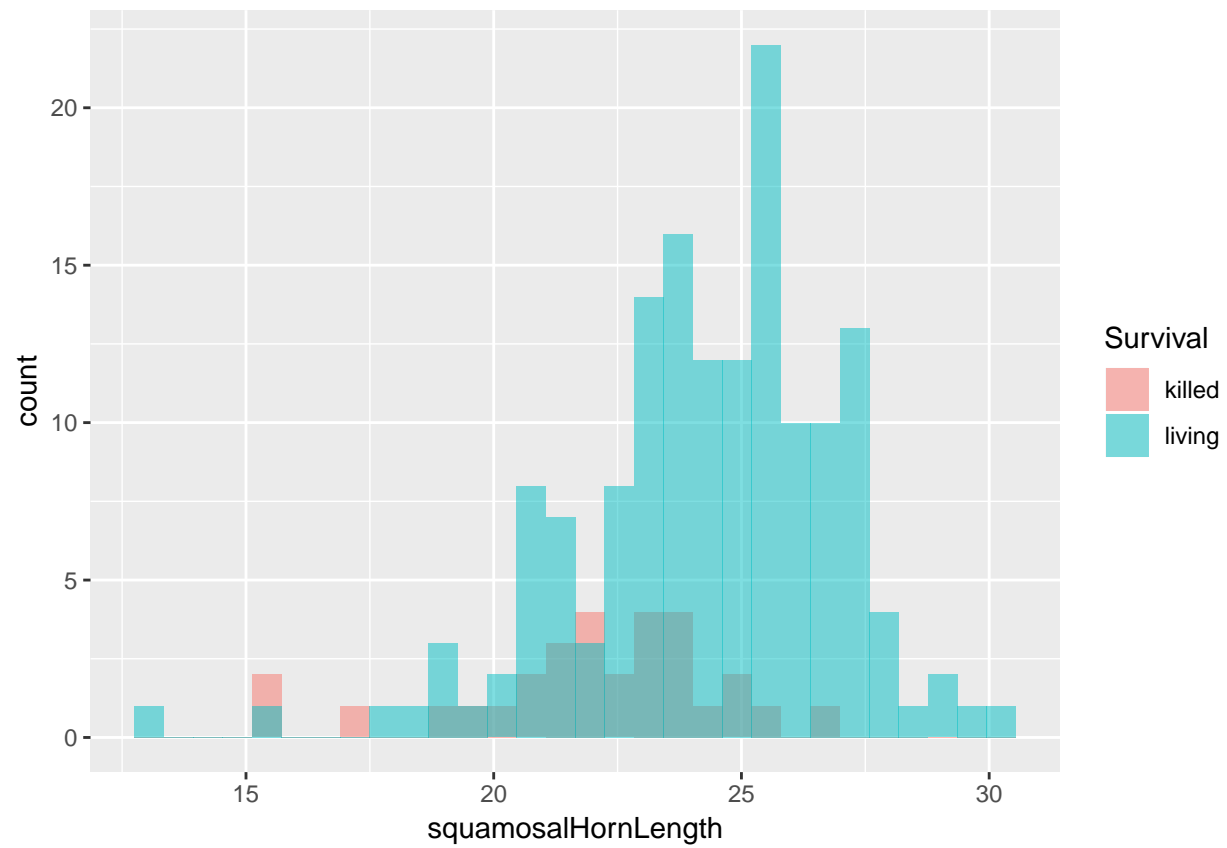


```
lizard <- read_csv(url("http://www.zoology.ubc.ca/~schluter/WhitlockSchluter/wp-content/
lizard2 <- lizard %>%
  na.omit()

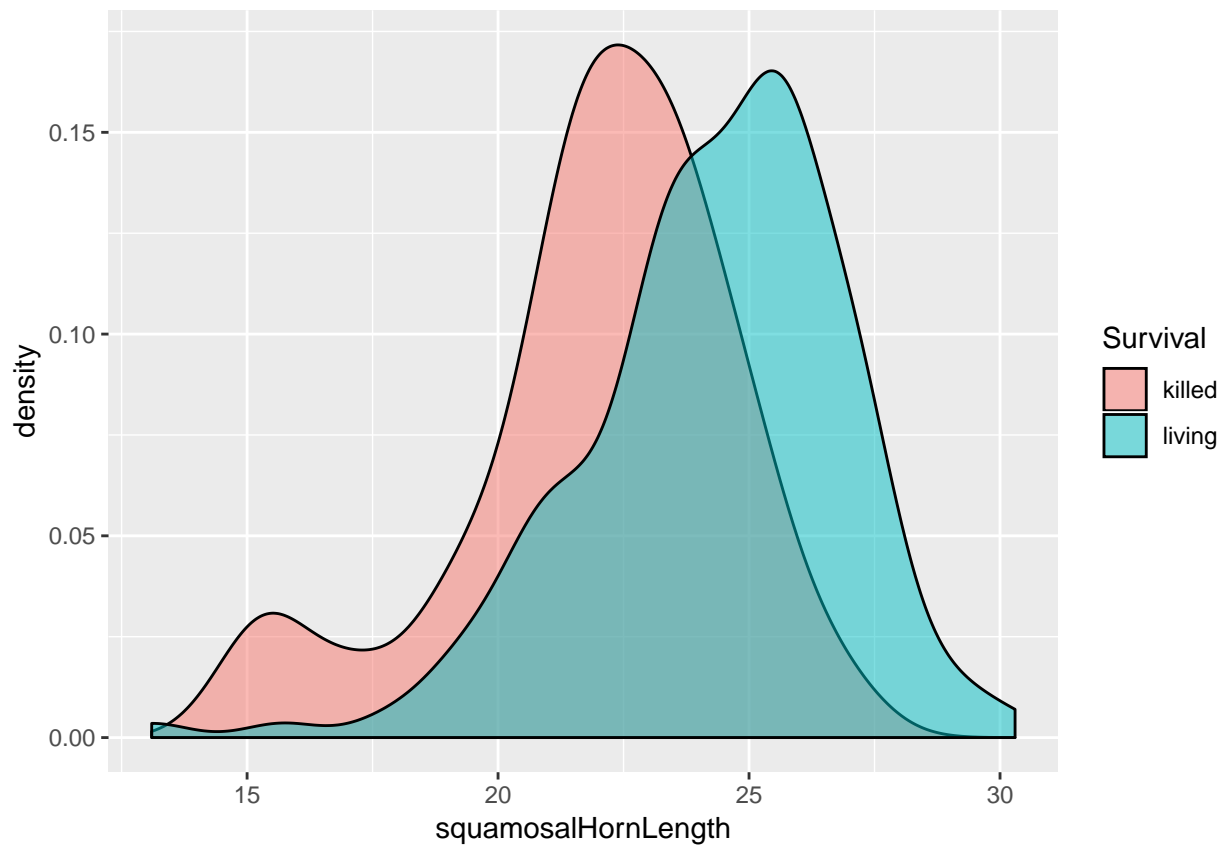
ggplot(lizard2, aes(squamosalHornLength, fill = Survival)) +
  geom_histogram(alpha=0.5, position="identity")
```

```
## `stat_bin()` using `bins = 30`. Pick
```

```
## better value with `binwidth`.
```



```
ggplot(lizard2, aes(squamosalHornLength, fill = Survival)) +  
  geom_density(alpha=0.5, position="identity")
```



```
t.test(squamosalHornLength ~ Survival, data = lizard, var.equal = TRUE)
```

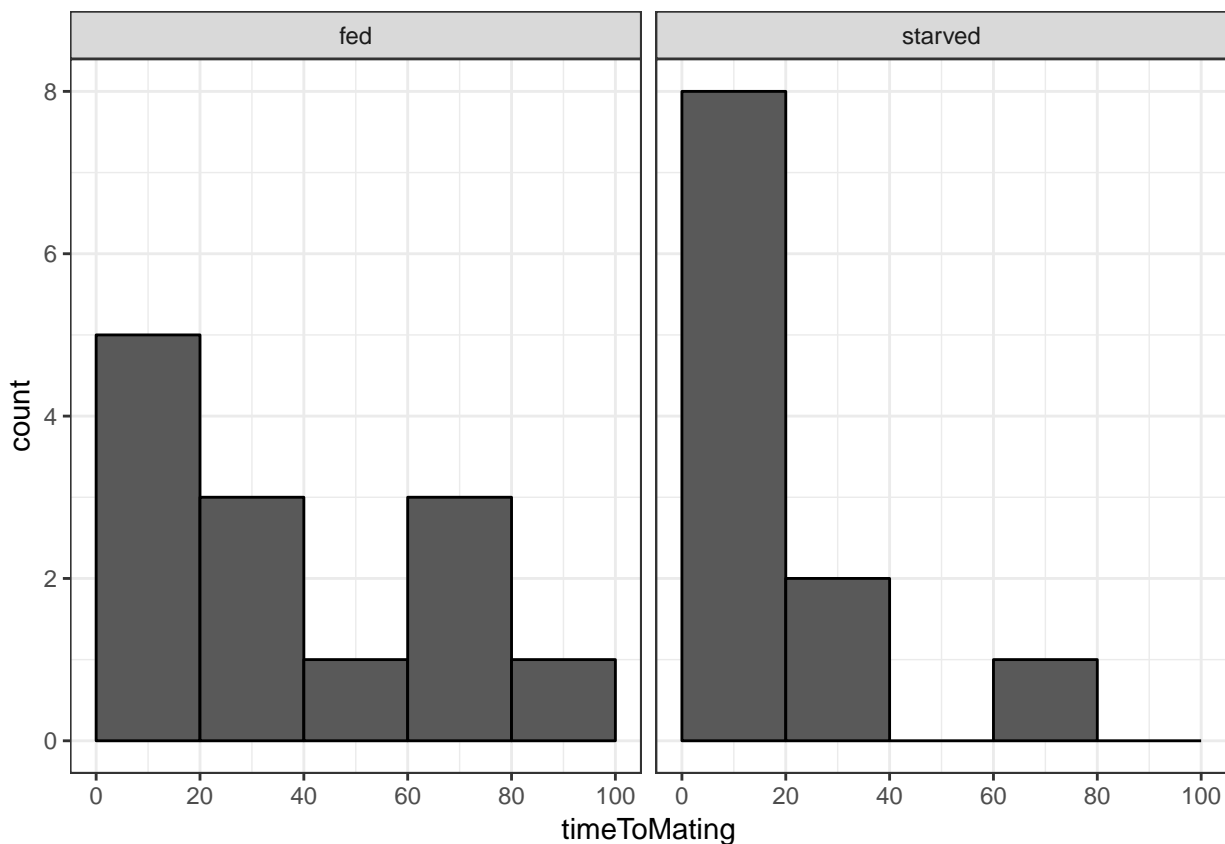
```
##
##  Two Sample t-test
##
## data:  squamosalHornLength by Survival
## t = -4.3494, df = 182, p-value =
## 2.27e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -3.335402 -1.253602
## sample estimates:
## mean in group killed
##          21.98667
## mean in group living
##          24.28117
```

## Sexual cannibalism in sagebrush crickets

During mating in the sage cricket, *Cyphoderris strepitans*, the male offers his fleshy hind wings to the female to eat. These wounds are not fatal, but a male that already has nibbled wings is less likely to be chosen by a female for subsequent mating. Since females get some nutrition from this process, Johnson et al. (1999) decided to test whether hungry females were more likely to mate than satiated females. <https://academic.oup.com/beheco/article/10/3/227/201476>

To test this, they randomly divided 24 females into two groups: one group (n = 11) was starved for at least two days and the second group (n = 13) was fed during the same period. Each female was then separately put in a cage with a single (neww) male and the waiting time to mating was recorded.

```
cannibalism <- read_csv(url("http://www.zoology.ubc.ca/~schluter/WhitlockSchluter/wp-con  
ggplot(cannibalism, aes(timeToMating)) +  
  geom_histogram(position="identity", colour = "black", binwidth = 20, boundary= 0) +  
  facet_wrap(~feedingStatus) +  
  scale_x_continuous(breaks = c(0, 20, 40, 60, 80, 100)) +  
  theme_bw()
```



Neither group is normally distributed and we have a small sample size. So test using a Wilcoxon rank-sum test.

```
wilcox.test(timeToMating ~ feedingStatus, data = cannibalism)

##
##  Wilcoxon rank sum test
##
## data:  timeToMating by feedingStatus
## W = 88, p-value = 0.3607
## alternative hypothesis: true location shift is not equal to 0
```

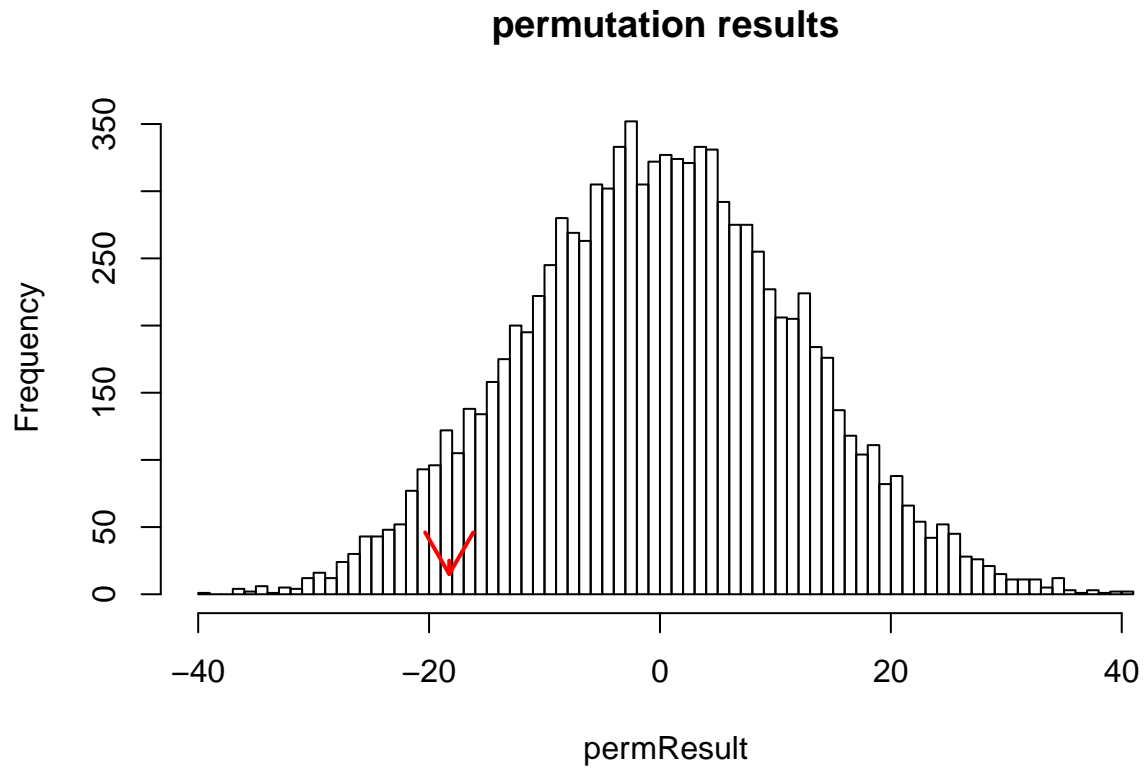
This is also the type of data we could run a permutation test on.

```
cricketMeans <-
  cannibalism %>%
  group_by(feedingStatus) %>%
  summarise(mean_timeToMating = mean(timeToMating))

diffMeans <- cricketMeans$mean_timeToMating[2] - cricketMeans$mean_timeToMating[1]

nPerm <- 10000
permResult <- vector() # initializes
for(i in 1:nPerm){
  # step 1: permute the times to mating
  permSample <- sample(cannibalism$timeToMating, replace = FALSE)
  # step 2: calculate difference between means
  permMeans <- tapply(permSample, cannibalism$feedingStatus, mean)
  permResult[i] <- permMeans[2] - permMeans[1]
}

hist(permResult, right = FALSE, breaks = 100, main= "permutation results")
arrows(diffMeans, 25, diffMeans, 15, col="red", lwd=2)
```



Use the null distribution to calculate an approximate p-value. Calculate the number of permuted means that fall below `diffMeans`

```
#two-tailed p-value  
2* (sum(as.numeric(permResult <= diffMeans))/nPerm)
```

```
## [1] 0.1322
```