Continuous Variables
Biology 683

Heath Blackmon

Today

1) Comparing means
   a) One sample t-test
   b) Two sample t-test
   c) Paired value t-test
2) Comparing Variances
   a) Levene's test
3) Assumptions
4) Transformations
5) Non-parametric tests
   a) Sign test
   b) Mann-Whitney / Wilcoxon Rank
   c) Permutation test

Continuous Data (more or less)

Many datasets are technically categorical but can be treated as continuous.
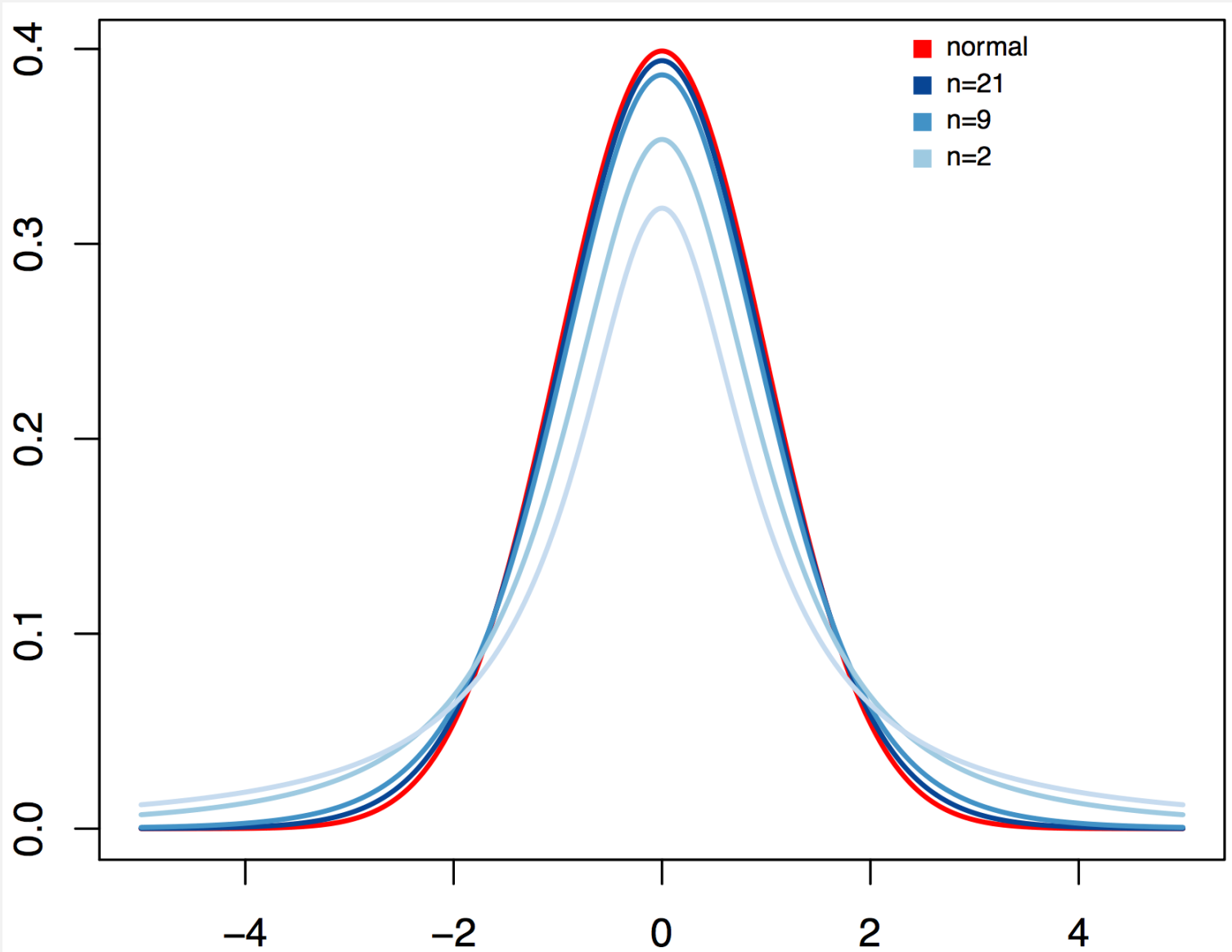
 - count data

 chromosomes

 offspring

 species

 gene copy number

T distribution

Power increases dramatically for the first few samples

One-sample t-test

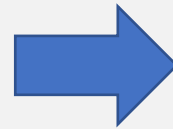The one sample T-test compares a sample to a hypothesis

Assumptions:
- The variable is normally distributed in the population

$$t = \frac{(\bar{Y} - \mu)}{SEM}$$

$$\frac{\sigma}{\sqrt{n}}$$

Implementation

```
x <- c(.45, .43, .31,
       .54, .47, .39)

t.test(x, mu=.5)
```

```
        One Sample t-test

data:  x
t = -2.1579, df = 5, p-value = 0.08341
alternative hypothesis: true mean is not equal to 0.5
95 percent confidence interval:
 0.3502649 0.5130684
sample estimates:
mean of x
0.4316667
```

One-sample t-test

The one sample T-test compares a sample to a hypothesis

$$t = \frac{(\bar{Y} - \mu)}{SEM}$$

$$\frac{\sigma}{\sqrt{n}}$$

Assumptions:
- The variable is normally distributed in the population

Implementation
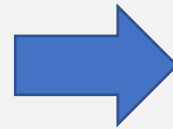
```
x <- c(.45, .43, .31,
       .54, .47, .39)


t.test(x, mu=.5)
```

```
        One Sample t-test

data:  x
t = -2.1579, df = 5, p-value = 0.08341
alternative hypothesis: true mean is not equal to 0.5
95 percent confidence interval:
 0.3502649 0.5130684
sample estimates:
mean of x
0.4316667
```
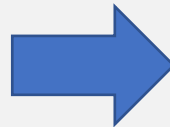
One-sample t-test

Arcsine is the standard transformation for proportion data. Lets see how our answer changes.

$$t = \frac{(\bar{Y} - \mu)}{SEM}$$

```
x <- c(.45, .43, .31,
       .54, .47, .39)
tx <- asin(sqrt(x))
t.test(tx, mu=asin(sqrt(.5)))
```

```
        One Sample t-test

data:  tx
t = -2.1444, df = 5, p-value = 0.08484
alternative hypothesis: true mean is not equal to 0.7853982
95 percent confidence interval:
 0.6329179 0.7991801
sample estimates:
mean of x
 0.716049
```

If you transform don't forget to back transform

```
      One Sample t-test

data:  tx
t = -2.1444, df = 5, p-value = 0.08484
alternative hypothesis: true mean is not equal to 0.7853982
95 percent confidence interval:
 0.6329179 0.7991801
sample estimates:
mean of x
 0.716049
```

```
x <- c(.45, .43, .31,
       .54, .47, .39)
tx <- asin(sqrt(x))
res <- t.test(tx, mu=asin(sqrt(.5)))
sin(res$conf.int)^2

[1] 0.3498723 0.5137801
```
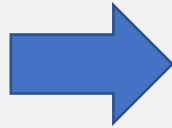
Two-sample t-test

Many times we will want to compare two populations strains or treatments and see if their means differ.

Implementation

$$t = \frac{(\bar{Y}_a - \bar{Y}_b)}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

```
# population 1 measures
x <- c(.45, .43, .31, .37,
       .47, .25, .39, .22)


# population 2 mesures
y <- c(.42, .49, .45, .49,
       .59, .37, .57, .53)


t.test(x, y)
```

```
        Welch Two Sample t-test

data:  x and y
t = -3.0297, df = 13.377, p-value = 0.009405
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.21815501 -0.03684499
sample estimates:
mean of x mean of y
  0.36125   0.48875
```
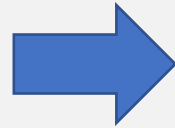
Paired-sample t-test

In some experiments you measure individuals before and after a manipulation. These values can be analyzed with the paired-sample t-test.

$$t = \frac{\bar{d}}{SE_d} \qquad \bar{d} = \frac{\sum_{i=1}^{n} y_{ai} - y_{bi}}{n}$$

Implementation

```
# first measure
x <- c(.24, .16, .14, .33, .04)
# second measure
y <- c(.52, .41, .10, .47, .39)

t.test(x, y, paired = T)
```
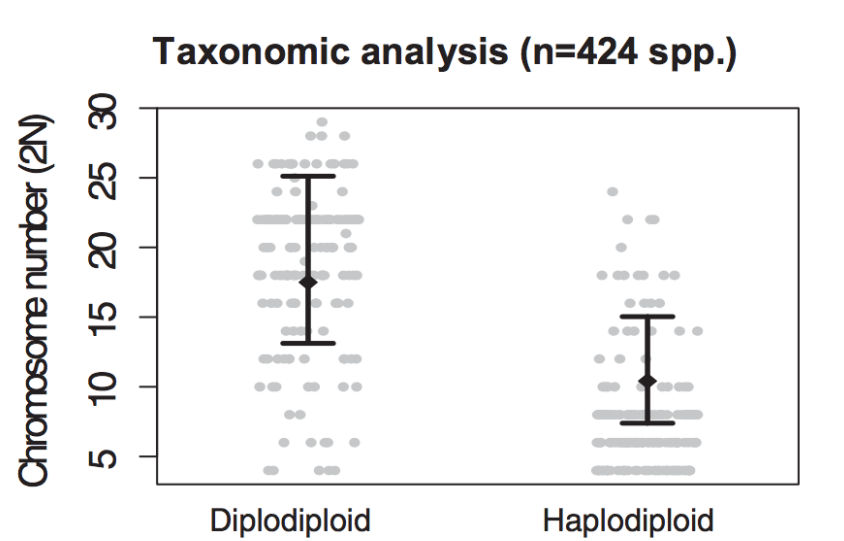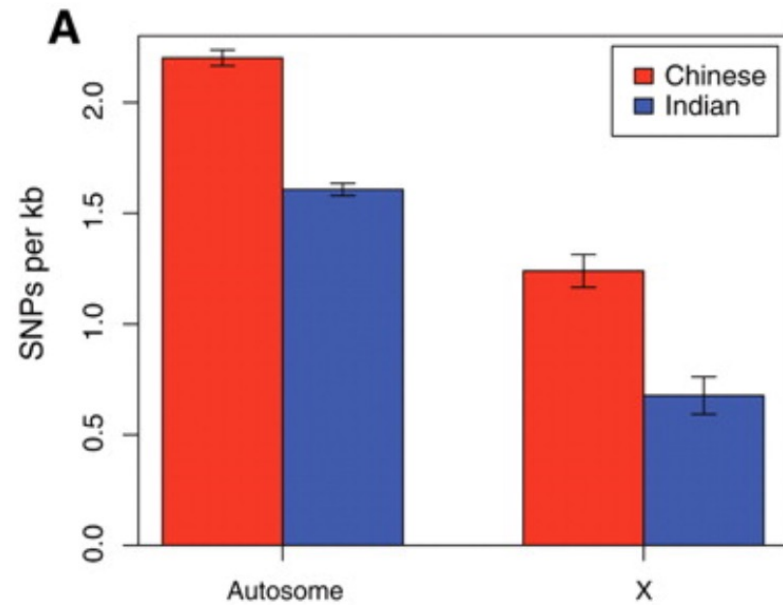
```
        Paired t-test

data:  x and y
t = -2.8817, df = 4, p-value = 0.04493
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.384839093 -0.007160907
sample estimates:
mean of the differences
                 -0.196
```
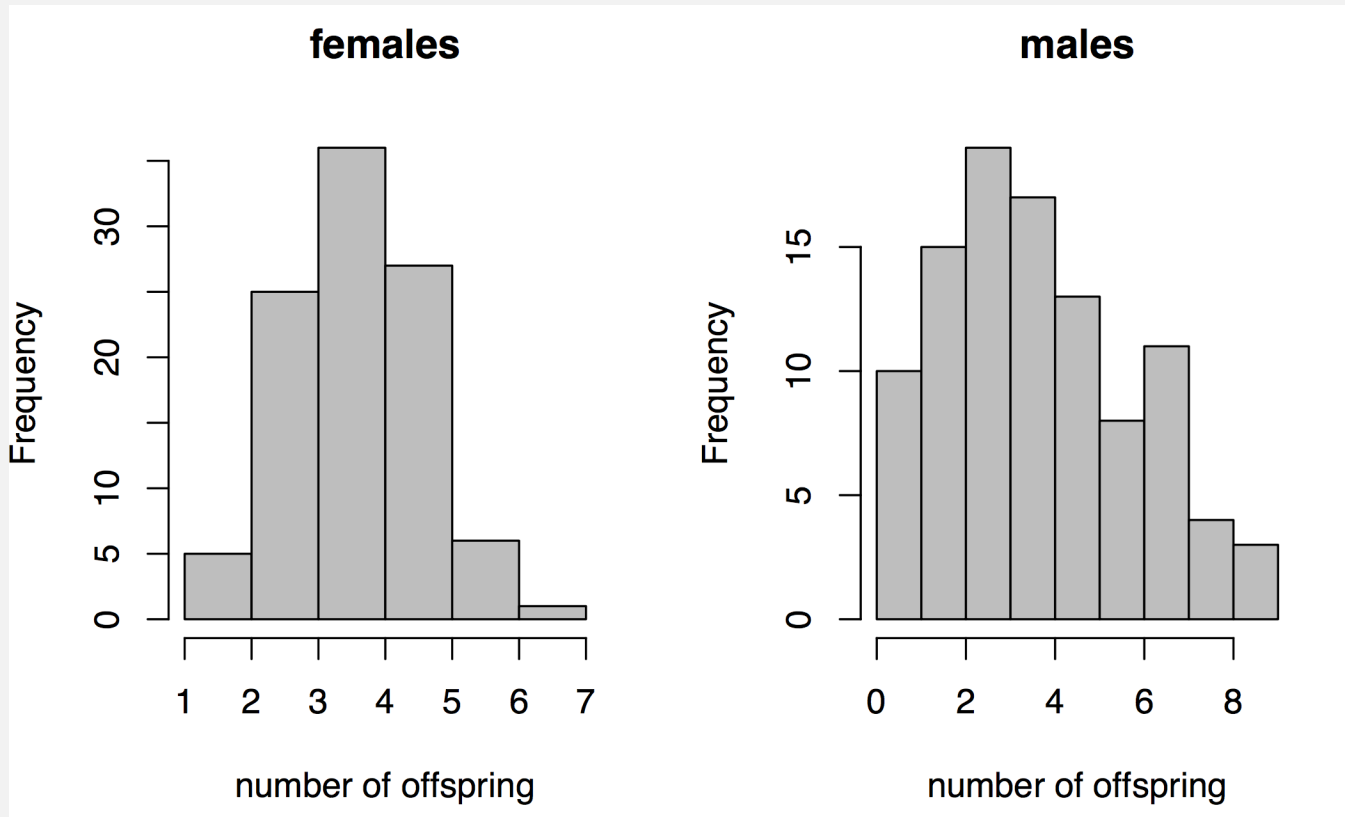
Taxonomic analysis (n=424 spp.)

If two 95% confidence intervals don't overlap then the values are significantly different at the α = 0.05 level.

If there is overlap in the CI then you can't conclude anything.

Comparing variances

- Some experiments might focus on differences in variance. Think about how sexual selection works with regard to number offspring.

- Two most common tests are the F-test and Levine's test. F-test assumes that the variable is normally distributed in the population and is sensitive to violations of this assumption.

- For this reason we will focus on Levine's test which assumes the variables have roughly symmetrical distributions and is furthermore quite robust to violations of this assumption.

Levene's test

- Levene's test works by calculating a central value for each group and then comparing the deviations of individuals in each group from these to calculate a W statistic.

- Compared to a F distribution

R insect spray data
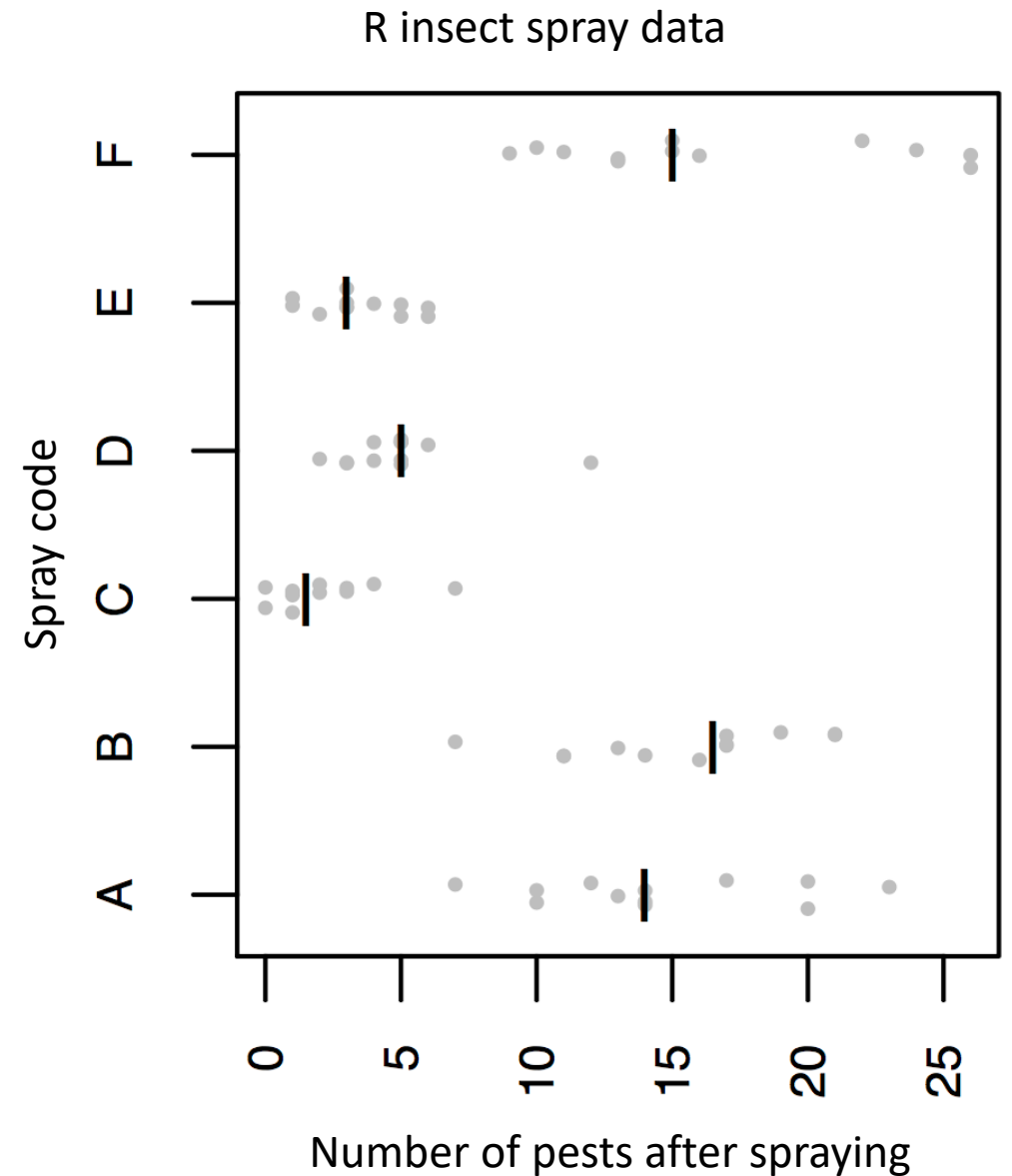
Levene's test

```r
library(car)
data("InsectSprays")
leveneTest(y = InsectSprays$count,
           group = InsectSprays$spray,
           center = "median")
```

```
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value    Pr(>F)
group  5  3.8214 0.004223 **
      66
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```
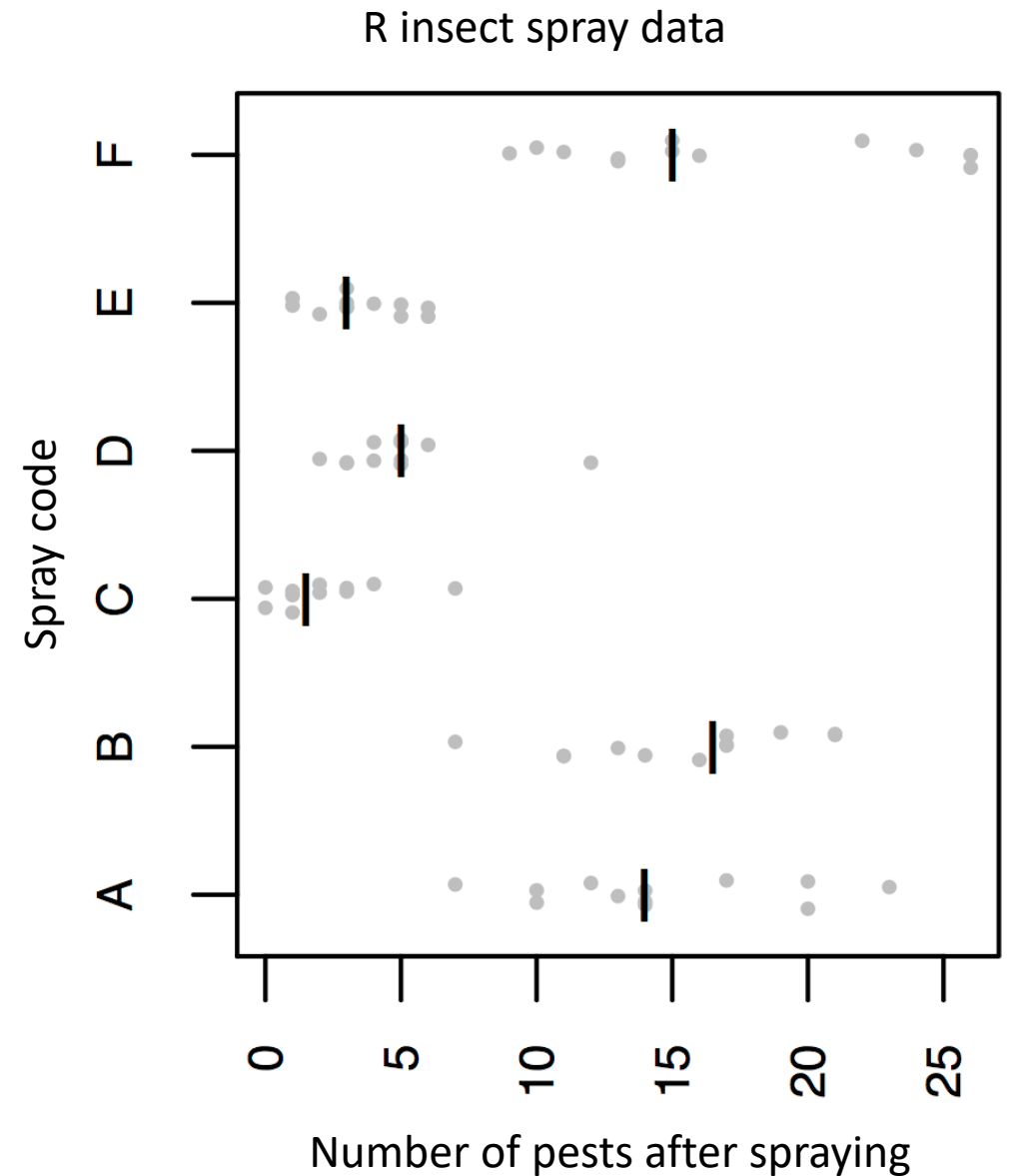

R insect spray data

Assumptions

- **One-sample T-test**
  - The variable is normally distributed in the population
- **Two-sample T-test**
  - The variable is normally distributed in the population
  - unequal variance ok
- **Paired-sample T-test**
  - Differences are normally distributed
- **Levine's test**
 - Distributions are roughly symmetrical

Assumptions

- Many tests are quite robust to minor violations of assumptions.

- Tests of normality (Shapiro-Wilks) are not terribly useful. `shapiro.test(x)`
  - Small sample size have little power
  - Large sample size reject with even minor deviations from normality

- Often most important to simply visualize distributions
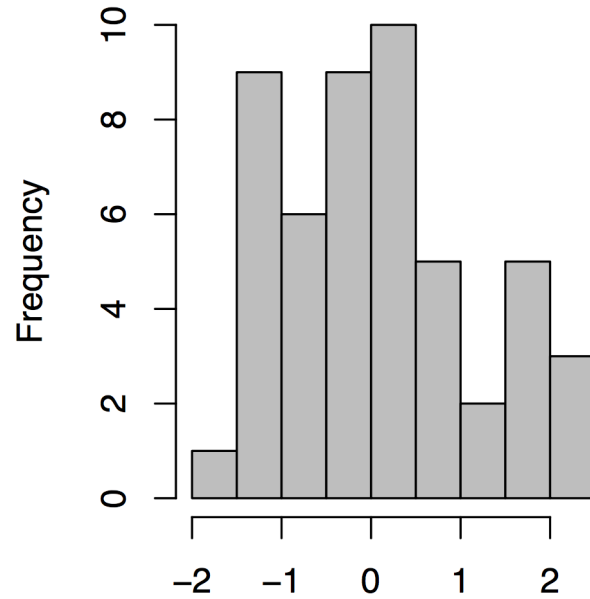 - histograms
  - qqplots

histograms
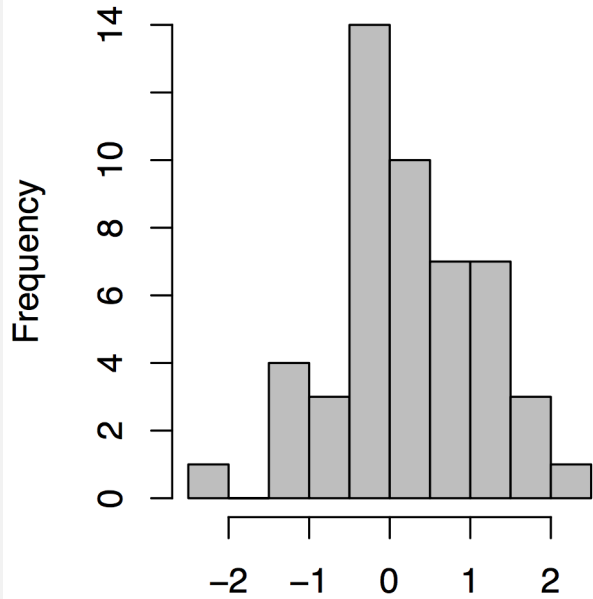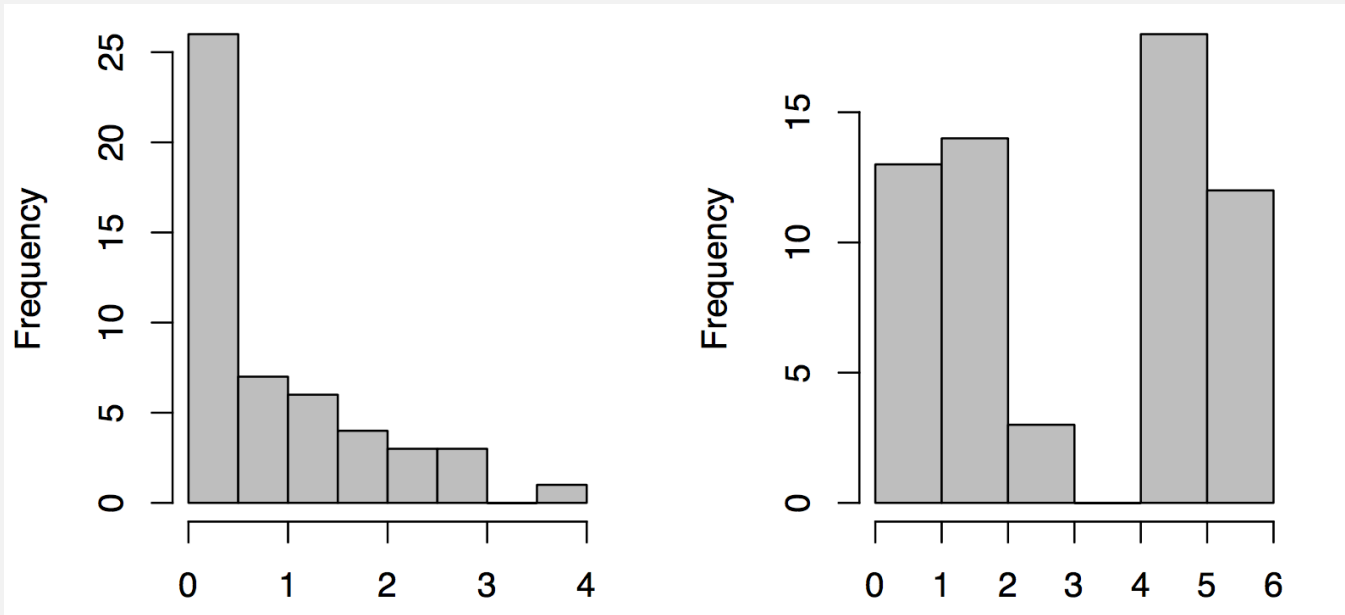

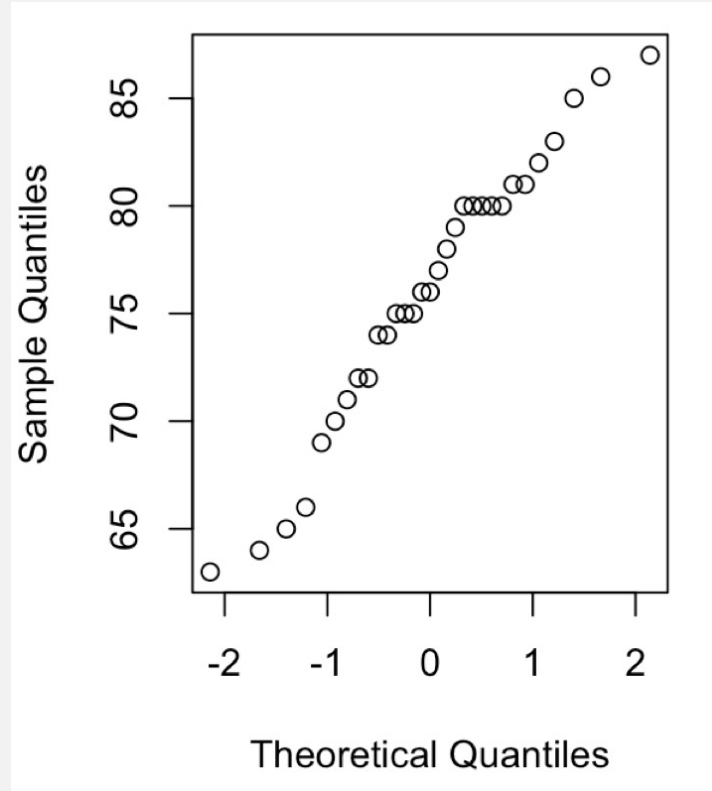You don't need or even expect a perfect bell curve from a finite sample of a population.

histograms

You should be worried if the histogram suggests a different distribution

qqplots

qqnorm(x)

normal
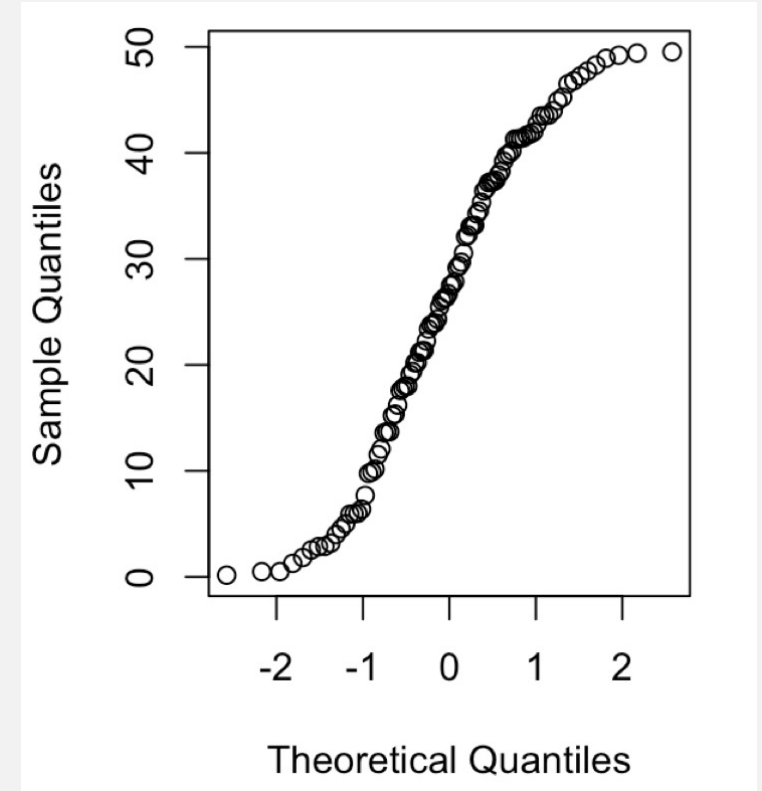
Too many extreme values

The time when you should be most hesitant is when your biological knowledge tells you that you don't expect a variable to have a normal distribution!

Data really isn't close to normal

First, be sure that the statistical test actually requires normality as an assumption – this constraint isn't universal (e.g., Levene's test merely requires symmetry)

Second, try to **transform** the variable to make it roughly normal

Third, as a last resort, use a non-parametric test

Transformations

Log transformation
- Take the natural log of every observation
- Especially for right-skewed data.

```
log(x)
# if zeros are present
log(x + 1)
```

Arcsine transformation
- Take the arcsine of the square root of each observation
- Used mainly for proportions

```
asin(sqrt(x))
```

Square-root transformation
- Add 0.5 and take the square root
- Solves similar problems to the log transformation

```
sqrt(x + .5)
```

For other transformations, see p. 381 in Whitlock and Schluter

Don't P-hack

- Try different transformations to see which improve normality or homoscedasticity not to see which one gives the lowest p-value.

- If you use a transformation that is not standard in your field or for the data type you will often need to justify it in the text of your paper or to the reviewers of your work

- Best case scenario is when you can take two routes and show the same result.
    - *The number of genes in the P53 network that showed signs of selection were positively correlated with the maximum life span of taxa (B = 2.5, R^2 = .26, p = .024). Analysis of square root transformed gene count data produced qualitatively similar results.*

    - *A Welch's two sample t-test showed that the two populations had significantly different mean body size. However despite log transformation of body size both populations failed a Shapiro-Wilks test of normality. Therefor we also performed a Mann-Whitney U-test which also showed that population A was significantly larger than population B.*

Non-parametric tests

1. Non-parametric tests still have assumptions but fewer.
2. You lose a lot of power when you switch from parametric to non-parametric.

Mann-Whitney / Wilcoxon Rank

Replaces two sample t-test

- Tests the hypothesis based upon a ranked list

- The tests are done on the ranks rather than the actual values

- Implementation: `wilcox.test(x, y)`

- Assumptions:
  - The distribution of measurements is symmetrical
  - Distributions of the two variables must have the same shape
  - A **significant result implies that the two variables have different distributions** (possibly due to different means but also possibly due to variances or skewness)

Doing a Mann-Whitney / Wilcoxon Rank in R

```
x <- c(45, 43, 46, 47,
       47, 25, 39, 22)
y <- c(42, 49, 45, 49,
       59, 37, 57, 53)
wilcox.test(x, y)
```

        Wilcoxon rank sum test with continuity
        correction

data: x and y
W = 14.5, p-value = 0.07357

```
> t.test(x,y)
```

        Welch Two Sample t-test

data: x and y
t = -2.1694, df = 12.883, p-value = 0.04937
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -19.21901046 -0.03098954
sample estimates:
mean of x mean of y
   39.250   48.875

Sign test

Replaces one-sample or paired-sample t-test

- Convert the data to a score (+ or -)
  - One-sample: whether the value is above or below the hypothesized value.
  - Paired-sample: whether the second value is higher or lower

- Under the null hypothesis, the number of +'s should equal the number of –'s, so this test boils down to a binomial test

- Obviously, it's throwing away a lot of information

- Implementation: `binom.test(x=4, n=14, p=.5)`

Doing sign test in R

- Implementation:

```
x <- c(45, 43, 31, 37,
       47, 25, 39, 22)
y <- c(42, 49, 45, 49,
       59, 37, 57, 53)
z <- x>y
[1]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
binom.test(x=sum(z), n=length(z), p=.5)
```

```
        Exact binomial test

data:  sum(z) and length(z)
number of successes = 1, number of trials = 8,
p-value = 0.07031
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.003159724 0.526509671
sample estimates:
probability of success
               0.125
```

Test	P-value
T-test	0.049
Wilcox rank test	0.074
Sign test	0.125

Permutation test

- Makes almost no assumptions about the underlying data and allow very specific questions to be asked.

- Under the null hypothesis, the two groups are drawn from identical distributions (they are from identical populations)

- By randomizing the observations between groups, an expected distribution under the null hypothesis can be generated

- If we are interested in differences between the mean, then we might choose mean 1-mean 2 as our test statistic

Permutation test

Step 1: Create a permuted set of data in which the values of the response variable are randomly reordered

Step 2: Calculate the test statistic from the randomly reordered groups

Step 3: Repeat Steps 1 and 2 thousands of times

Step 4: Compare the actual value of the test statistic from the original data to the distribution of values from the permutation steps

Permutation test

```r
crickets <- read.csv("crickets.csv")
obs.vals <- tapply(X = crickets$timeToMating,
                   INDEX = crickets$feedingStatus,
                   FUN = mean)
obs.d <- obs.vals[2]-obs.vals[1]
perm <- 10000
perm.d <- vector()
for(i in 1:perm){
  newX <- sample(crickets$timeToMating, replace = F)
  perm.vals <- tapply(X = newX,
                      INDEX = crickets$feedingStatus,
                      FUN = mean)
  perm.d[i] <- perm.vals[2] - perm.vals[1]
}
plot(density(perm.d))
abline(v = obs.d, col = "red")
sum(perm.d < obs.d) / perm * 2

t.test(crickets$timeToMating~crickets$feedingStatus)
```


density.default(x = perm.d)

N = 10000   Bandwidth = 1.709
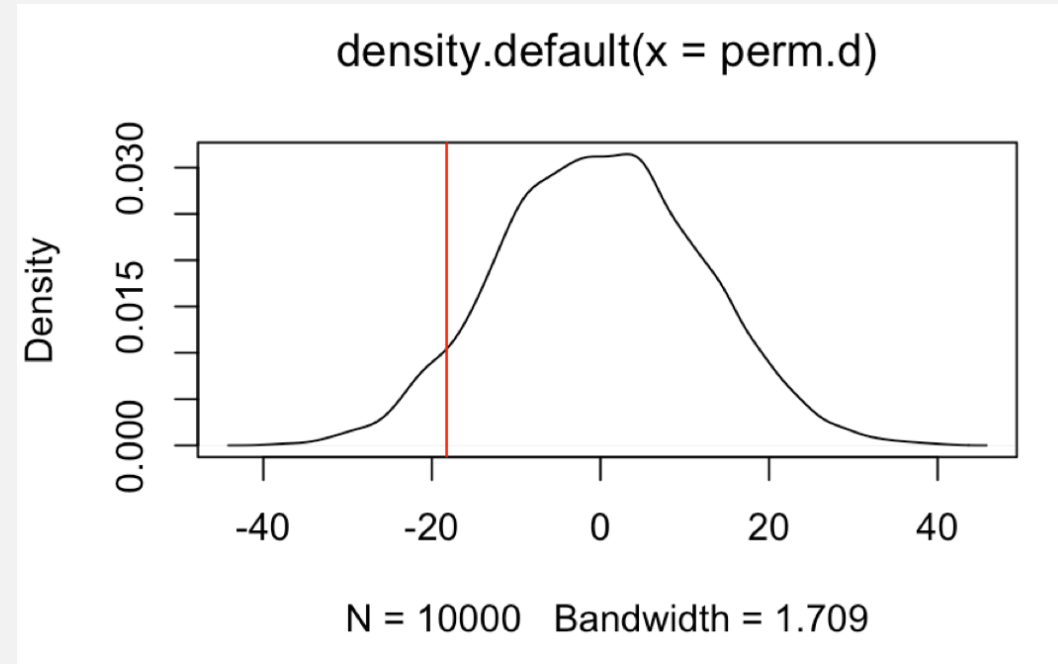
Correlation and Causation

The correlation is a foundational concept in the biological sciences

Pearson's correlation coefficient:

$$r(X,Y) = \frac{cov(X,Y)}{s_x s_y}$$

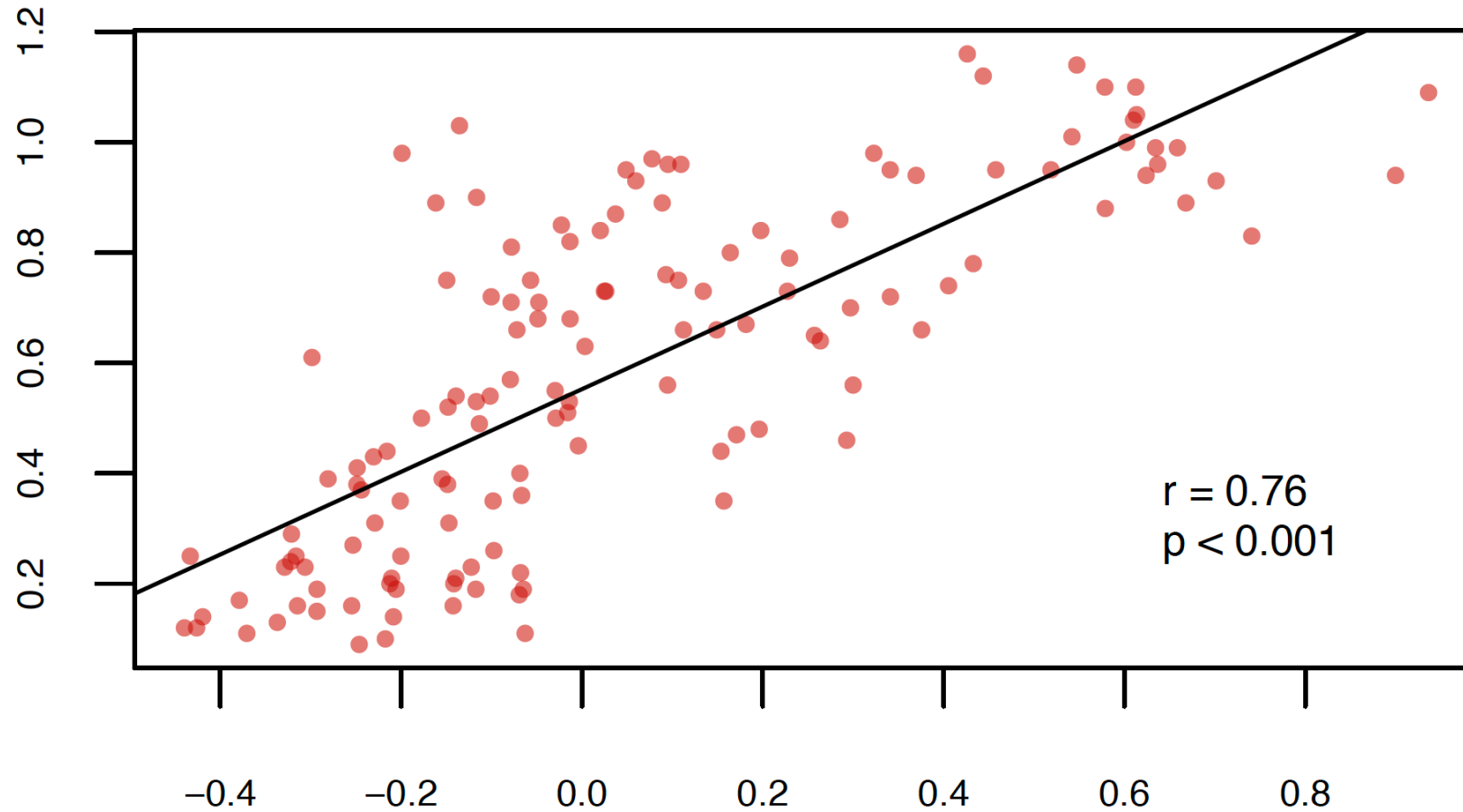$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Francis Galton
1822 - 1911

Karl Pearson
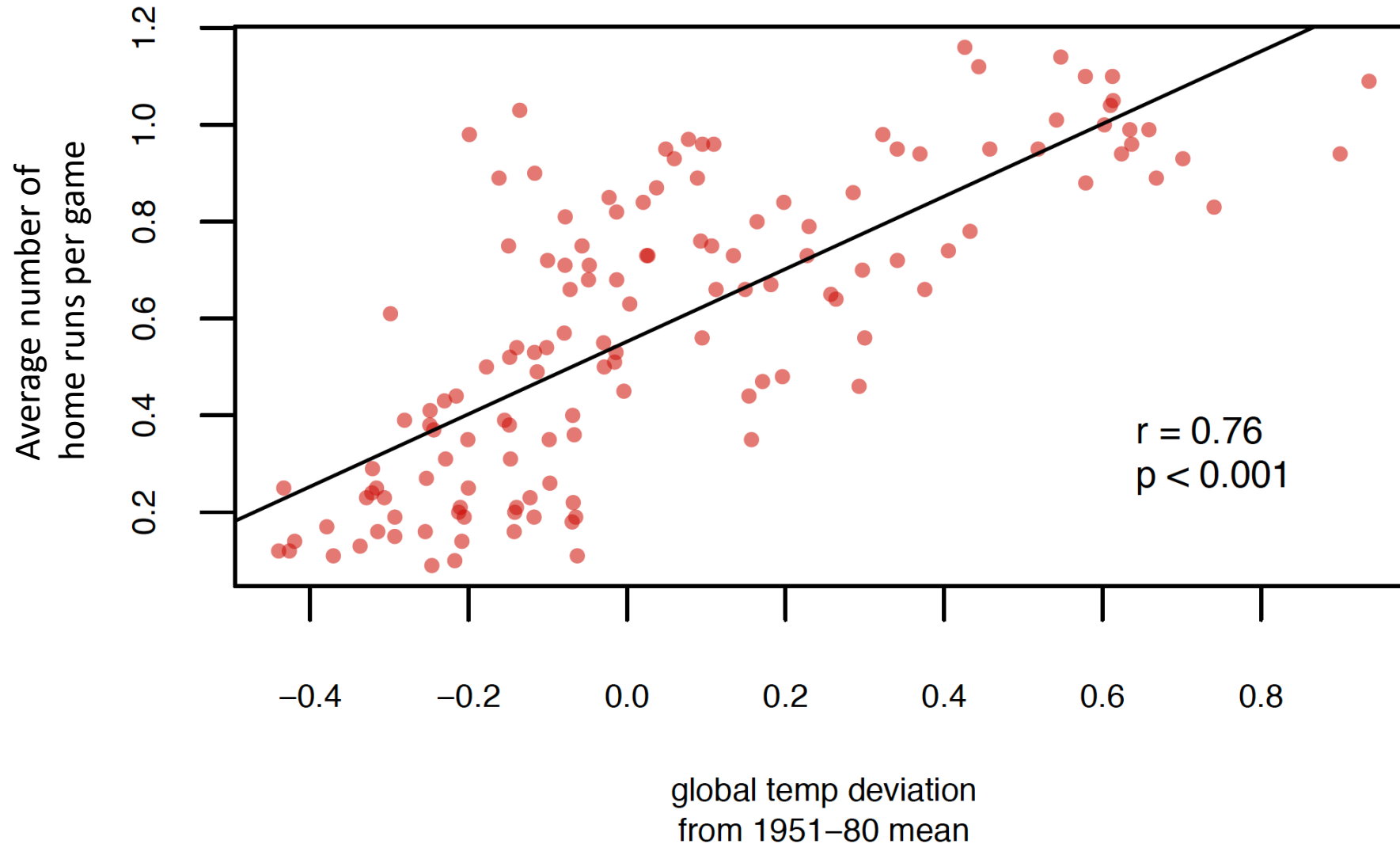1857 - 1936

Interpret with caution


r = 0.76
p < 0.001
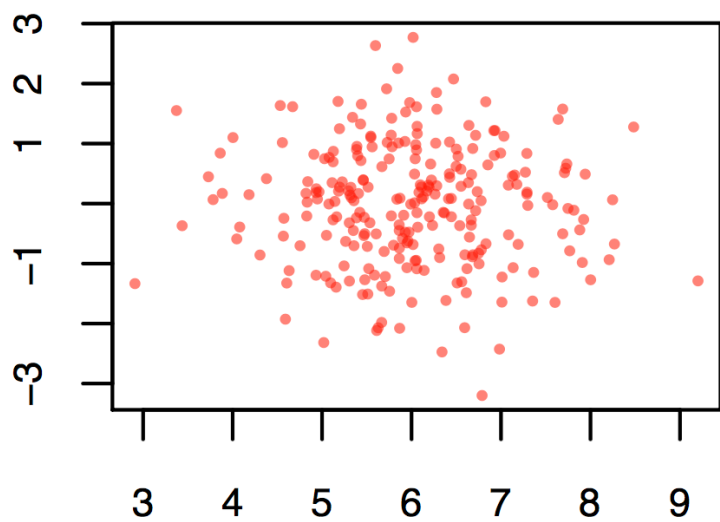
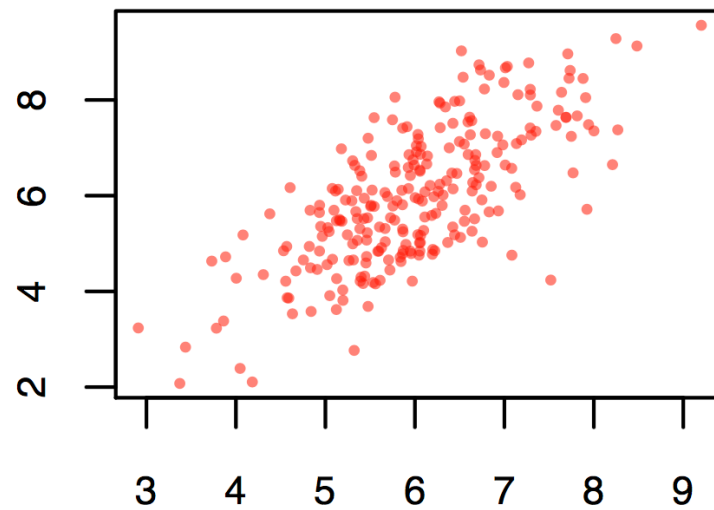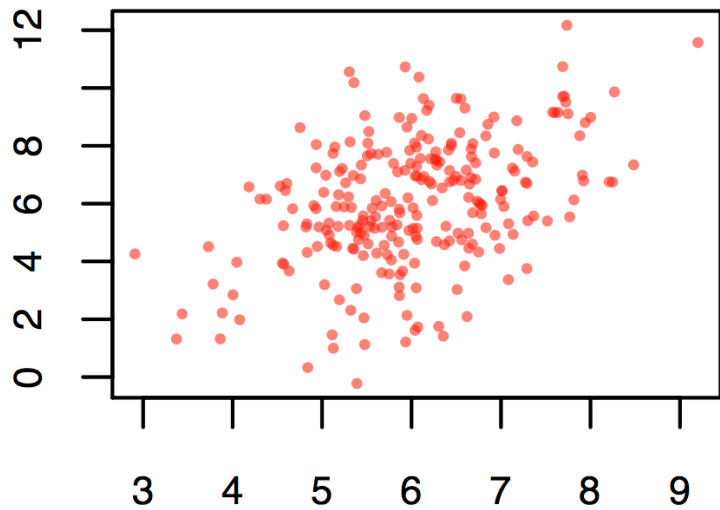# Obviously these covary with time

Pearson's Correlation Coefficient

Assumes that the data are roughly from a bivariate normal distribution
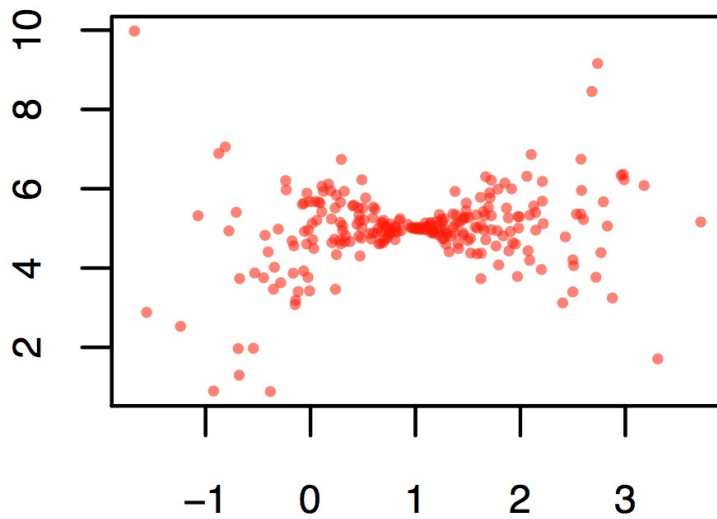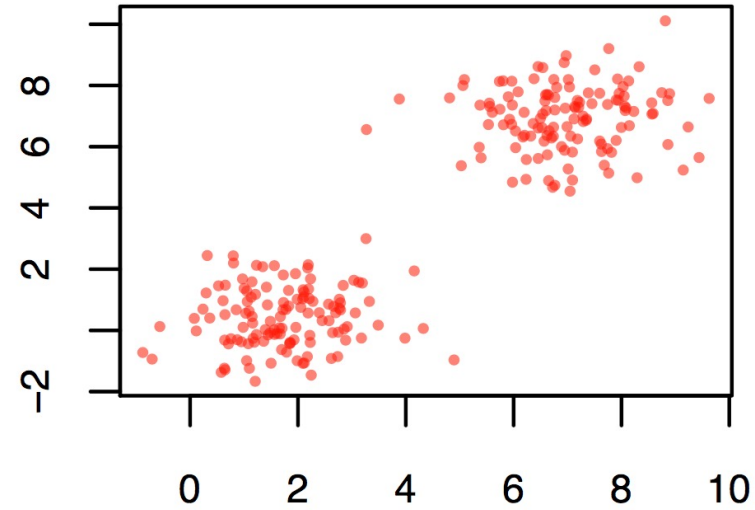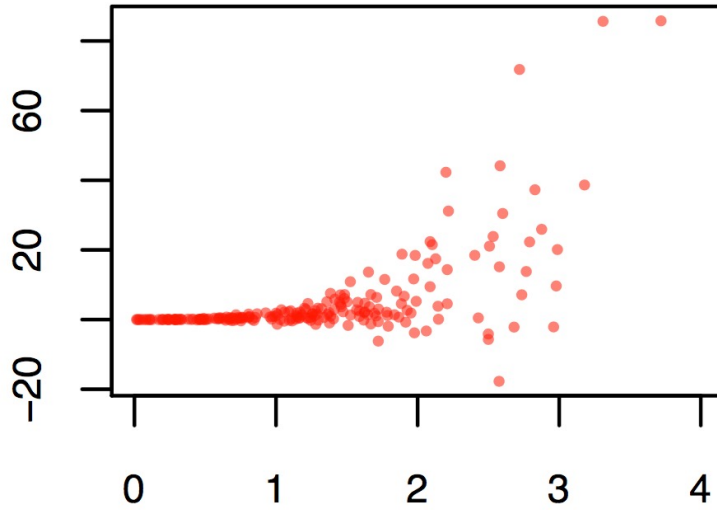
Would **not** be appropriate when:

- Relationship is non-linear (possible solutions?)

- Dataset contains extreme outliers (evaluate leverage)

- Data are in multiple distinct clouds of points (what could cause this?)

Good data
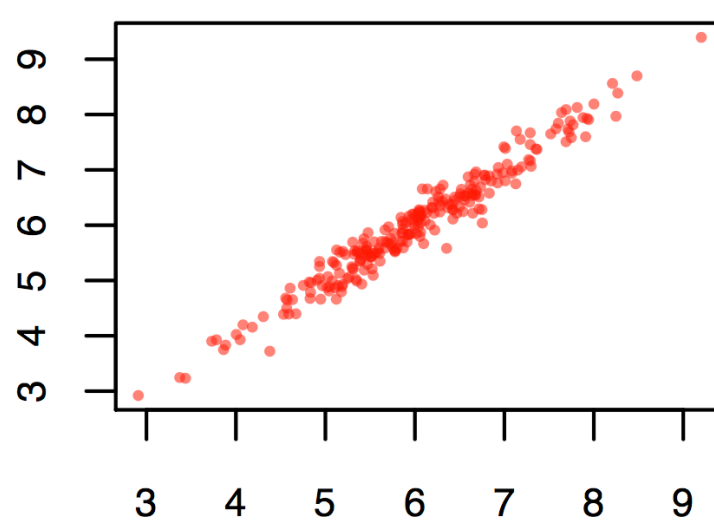
Bad data

Correlation simulator

Spearman's Rank Correlation

An alternative to Pearson's correlation coefficient for data that depart substantially from bivariate normality

Based on ranks

```
> cor.test(x, y, method = "pearson")

        Pearson's product-moment correlation

data:  x and y
t = 3.8345, df = 18, p-value = 0.001214
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3241974 0.8583527
sample estimates:
      cor
0.6705223


> cor.test(x, y, method = "spearman")

        Spearman's rank correlation rho

data:  x and y
S = 684, p-value = 0.03148
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4857143
```

Correlation Summary

- Correlations are very important

- We are interested in correlations between values of treatments and response variables

- We are interested in correlations between various factors that could affect the response variable

- We need to think about correlations between explanatory variables and potentially confounding variables

- **Many types of experiments call for the calculation and consideration of correlation coefficients**

t tests - history

- Based on the *t*-distribution, which is similar to a normal distribution, except it takes into account the fact that we don't know the population standard deviation with certainty. Published by Gosset in 1908 under the pseudonym Student because he was employed by Guinness brewery.

- Since we only know the sample standard deviation, instead of the population standard deviation, we have some additional uncertainty regarding the possible values of the mean

- Consequently, the *t*-distribution is wider than the normal distribution, and the *t*-distribution actually yields a more exact estimate of the confidence interval for the mean

BIOMETRIKA.

——————

THE PROBABLE ERROR OF A MEAN.

By STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

GUINNESS

The T-distribution

lets generate the t-distribution through simulation ourselves.