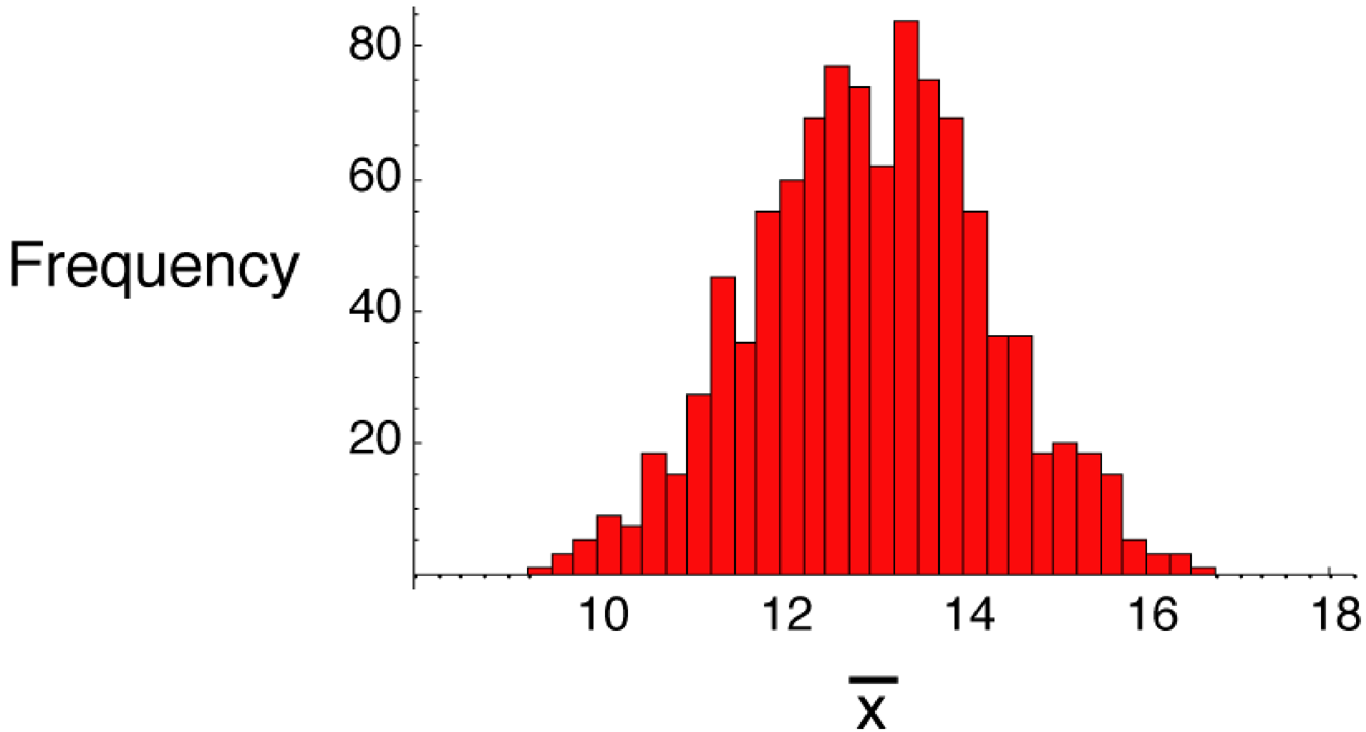
Estimating with uncertainty

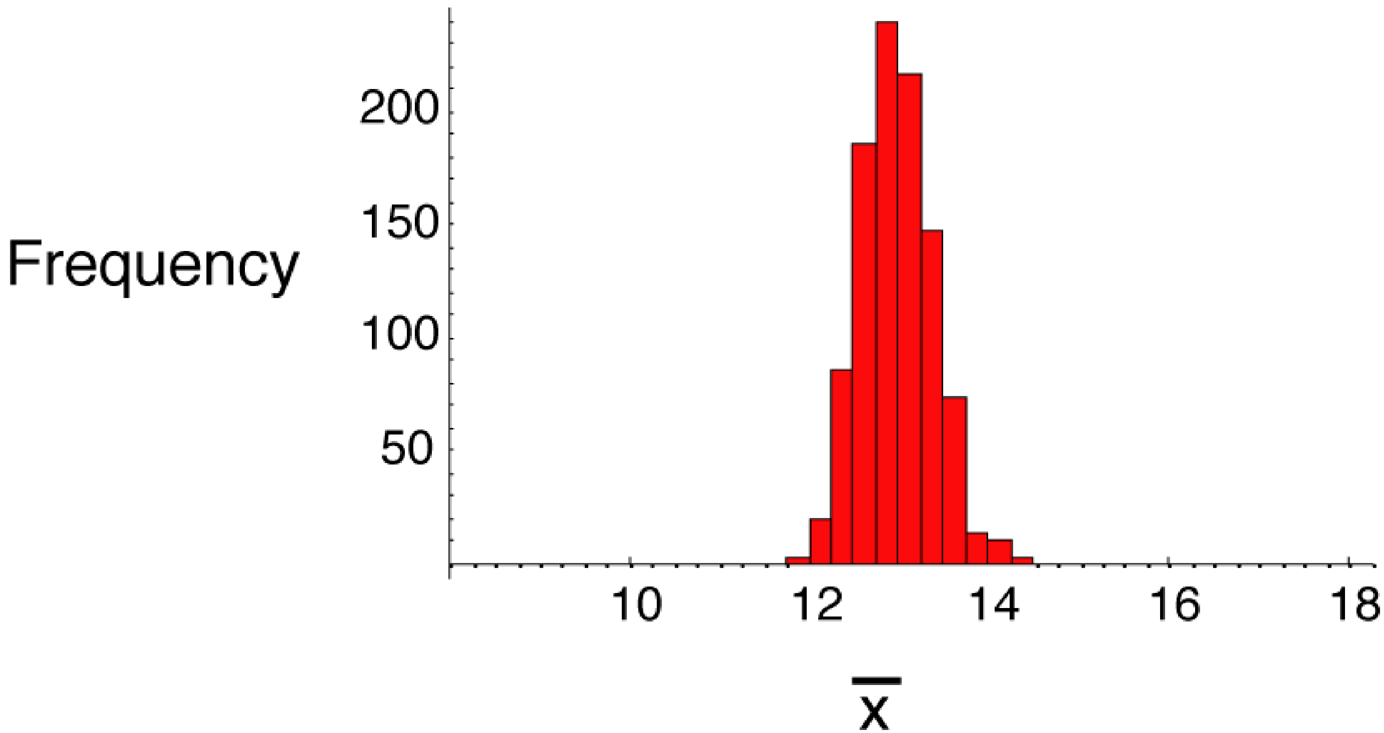
"We demand rigidly defined areas of doubt and uncertainty!"

Douglas Adams

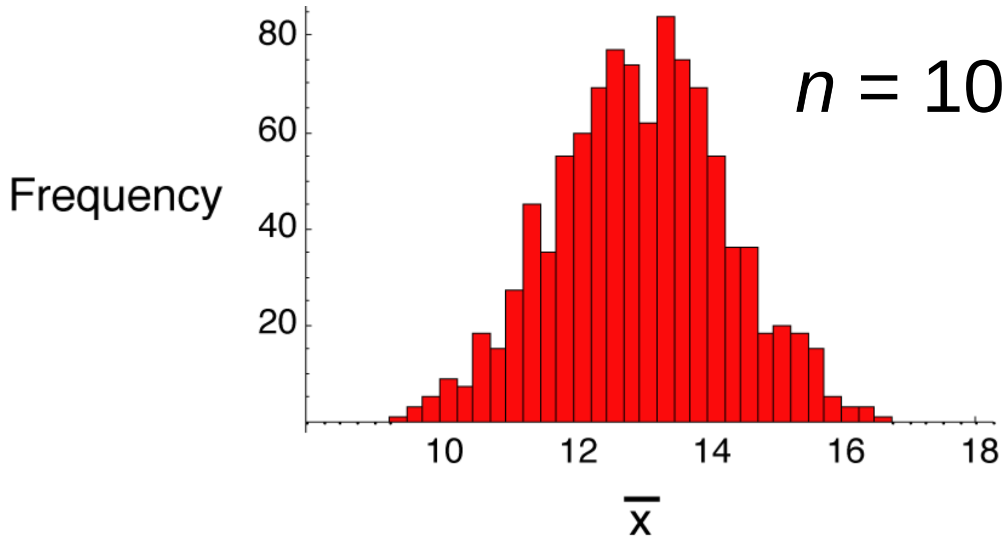
Distribution of the means of many samples, each of sample size 10

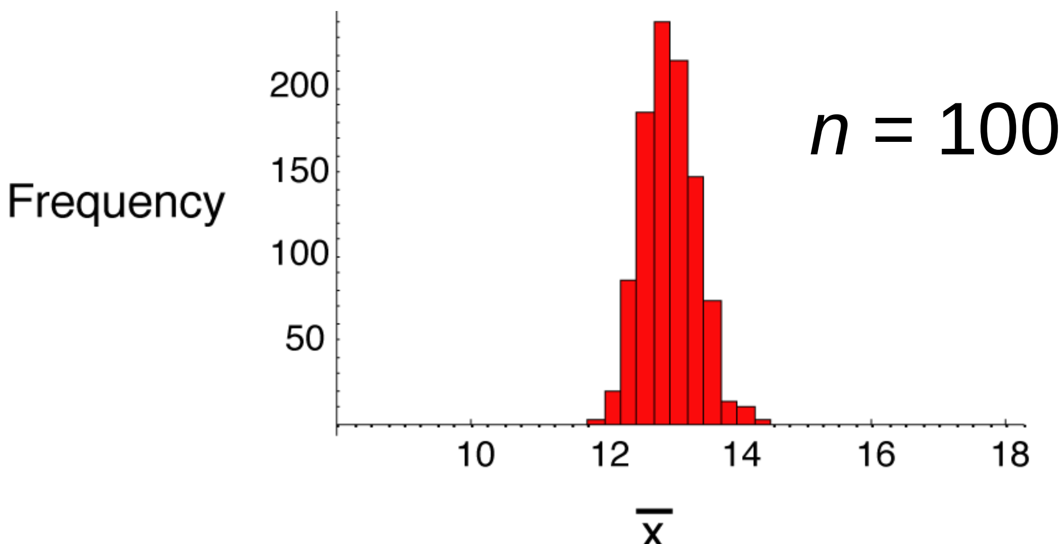


Distribution of the means of many samples, each of sample size 100



Variation in sample means decreases with sample size





1000 samples each

The *standard error* of an estimate is the standard deviation of its sampling distribution. The standard error predicts the sampling error of the estimate.

Confidence interval

The 95% confidence interval provides a plausible range for a parameter. All values for the parameter lying within the interval are plausible, given the data, whereas those outside are unlikely.

Use correct language when talking about confidence intervals

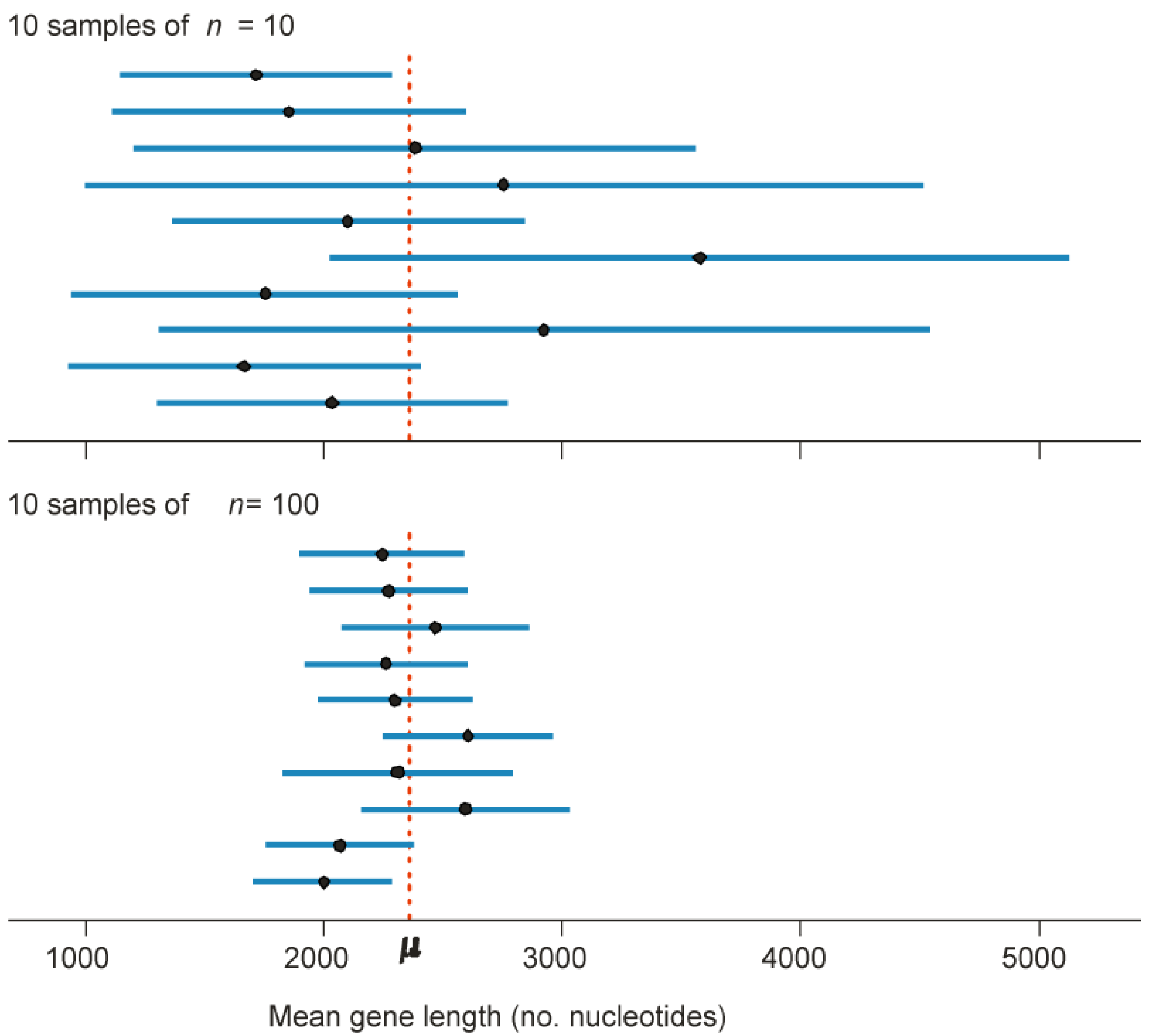
Not correct:

"There is a 95% probability that the population mean is within a particular 95% confidence interval"

Correct:

"We are 95% confident that the population mean lies within the 95% confidence interval."

Confidence interval



Hypothesis testing

Hypothesis testing asks how unusual it is to get data that differ from the null hypothesis.

If the data would be quite unlikely under H_0, we reject H_0.

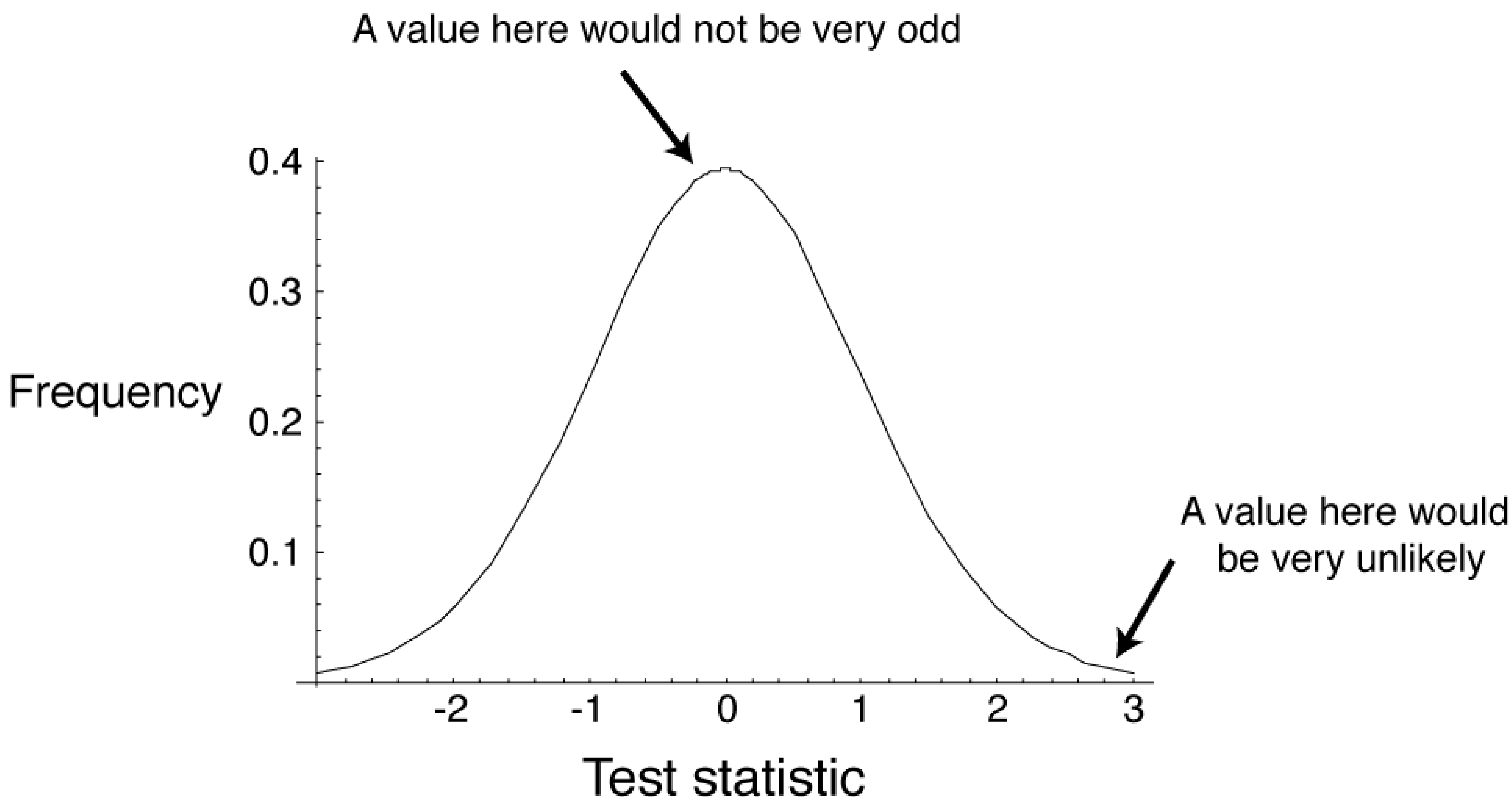
Null hypothesis: a specific statement about a population parameter made for the purposes of argument.

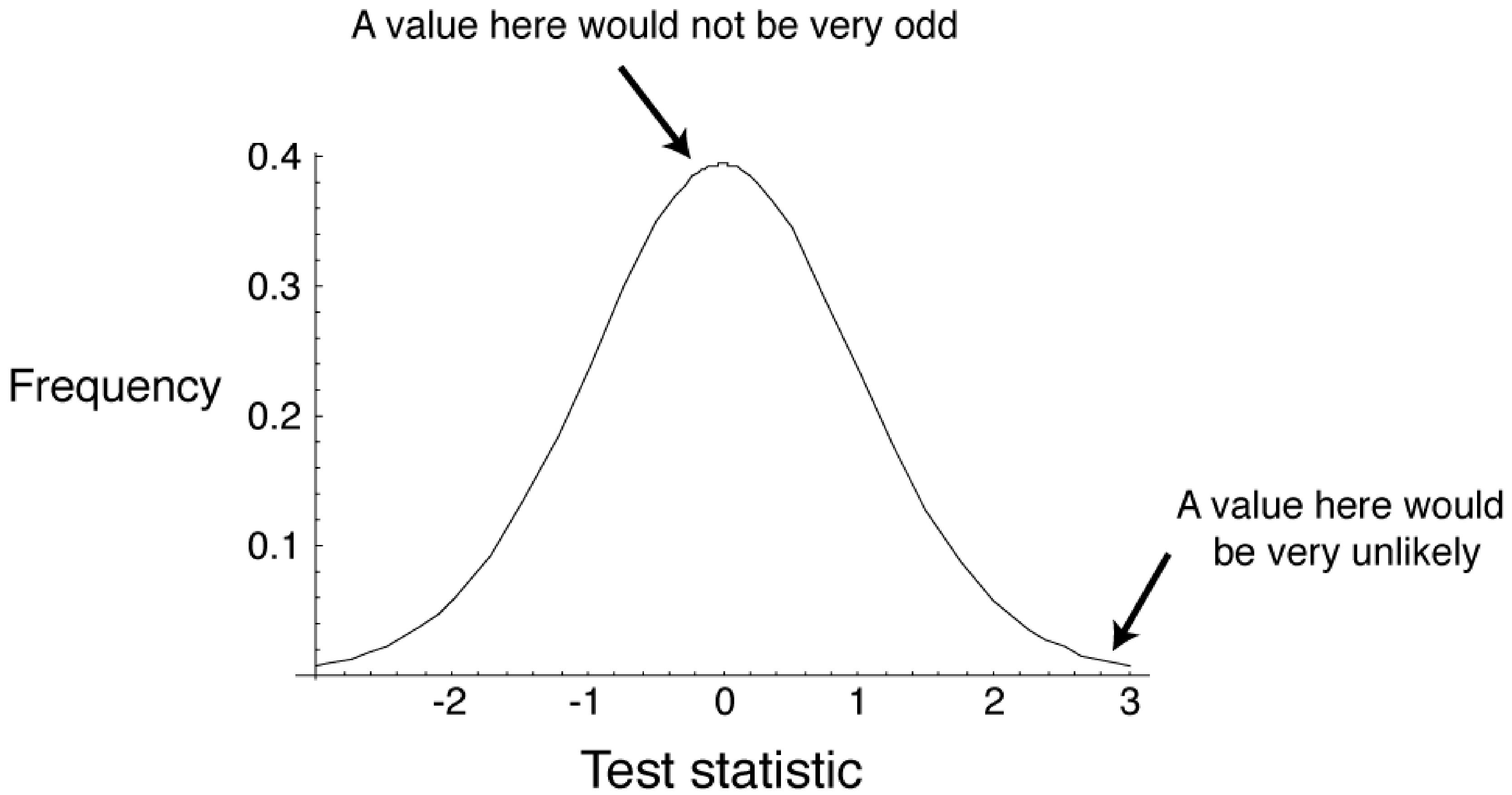
Alternate hypothesis: represents all other possible parameter values except that stated in the null hypothesis.

The *null hypothesis* is usually the simplest statement, whereas the *alternative hypothesis* is usually the statement of greatest interest.

A good null hypothesis would be interesting if proven wrong.

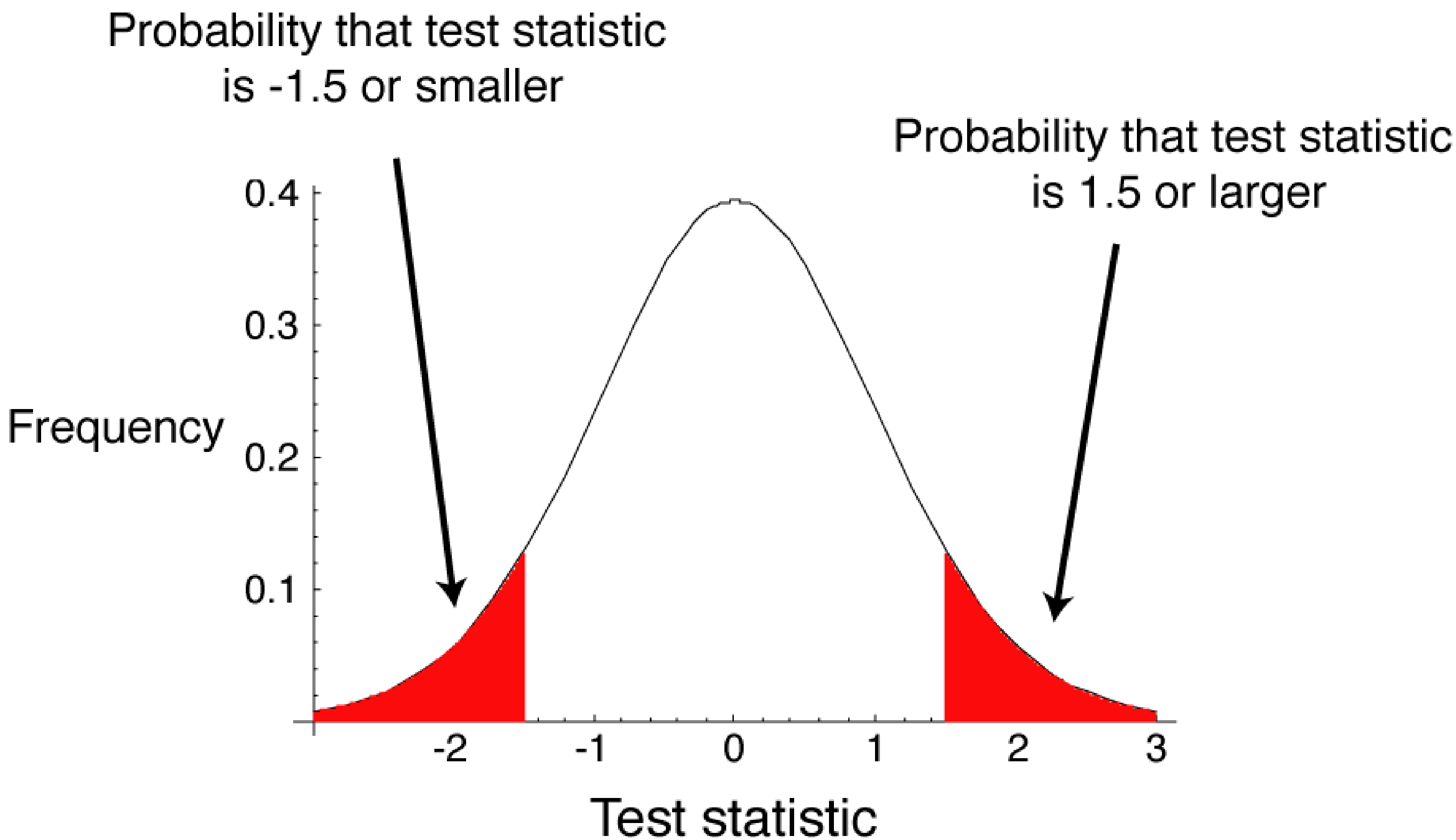
A null hypothesis is specific;
an alternate hypothesis is not.


A value here would not be very odd
A value here would be very unlikely
Frequency
0.4
0.3
0.2
0.1
-2
-1
0
1
2
3
Test statistic



A test statistic summarizes the match
between the data and the null hypothesis

P-value


Probability that test statistic is -1.5 or smaller
Probability that test statistic is 1.5 or larger
Frequency
0.4
0.3
0.2
0.1
-2
-1
0
1
2
3
Test statistic

A *P*-value is the probability of getting the data, or something as or more unusual, if the null hypothesis were true.

Hypothesis testing: an example

Does a red shirt help win wrestling?

The experiment and the results

- Animals use red as a sign of aggression

- Does red influence the outcome of wrestling, taekwondo, and boxing?

    - 16 of 20 rounds had more red-shirted than blue-shirted winners in these sports in the 2004 Olympics

    - Shirt color was randomly assigned

Hill, RA, and RA Burton 2005. Red enhances human performance in contests Nature 435:293.

Stating the hypotheses

H_0: Red- and blue-shirted athletes are equally likely to win (*proportion* = 0.5).

H_A: Red- and blue-shirted athletes are not equally likely to win (*proportion* ≠ 0.5).

Estimating the value

- 16 of 20 is a proportion of *proportion* = 0.8

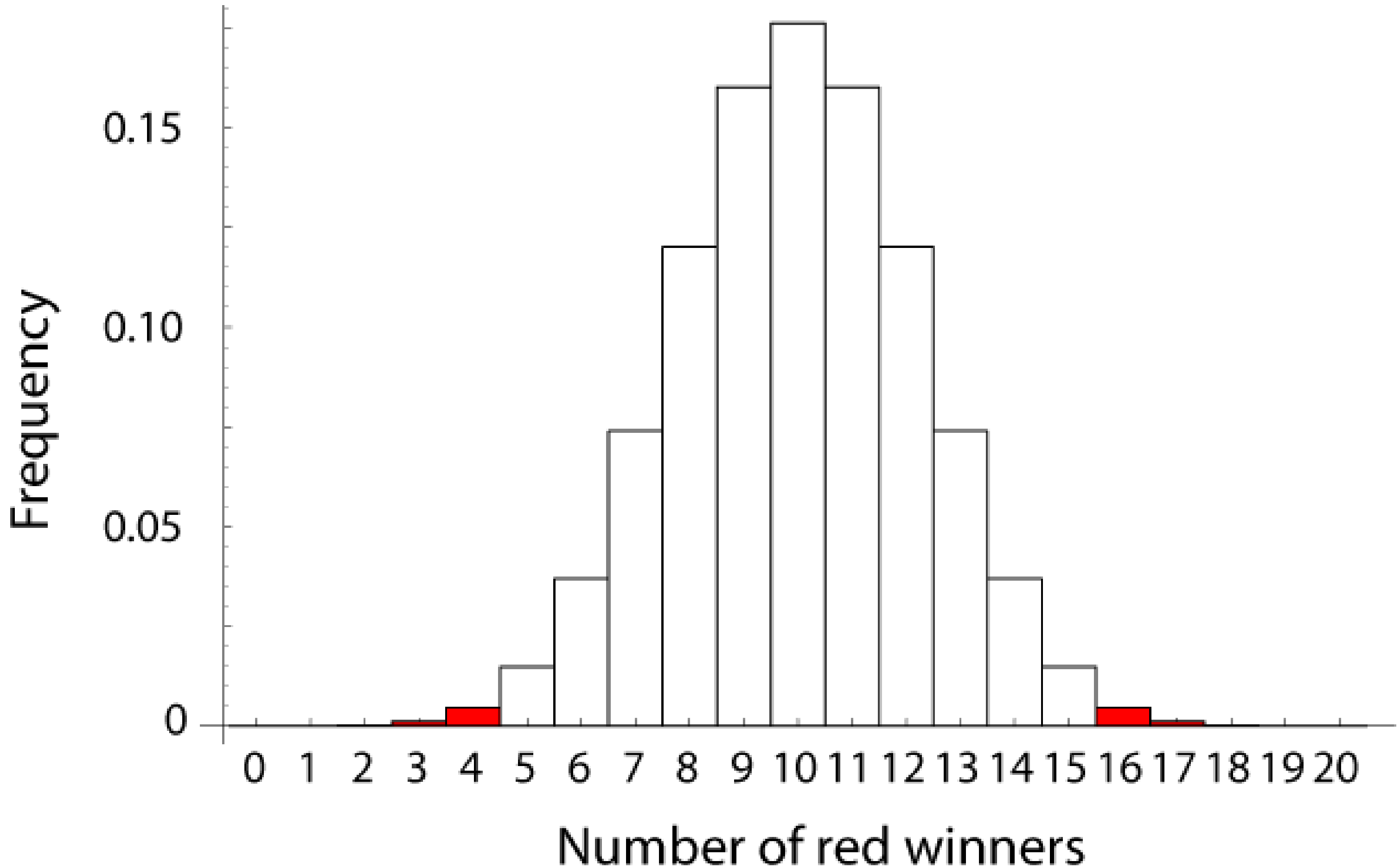
- This is a discrepancy of 0.3 from the proportion proposed by the null hypothesis, *proportion* = 0.5

Is this discrepancy by chance alone?:

Estimating the probability of such an extreme result

- The *null distribution* for a test statistic is the probability distribution of alternative outcomes when a random sample is taken from a population corresponding to the null expectation.

The null distribution of the *sample proportion*



Calculating the *P*-value from the null distribution

The *P*-value is calculated as

$$P = 2 \times [\Pr(16) + \Pr(17) + \Pr(18) + \Pr(19) + \Pr(20)] = 0.012.$$

Statistical significance

The *significance level*, α, is a 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 rejected.

α is often 0.05

Significance for the red shirt example

- $P = 0.012$

- $P < \alpha$, so we can reject the null hypothesis

- Athletes in red shirts were more likely to win.

Significance level

- The acceptable probability of rejecting a true null hypothesis
- Called α
- For many purposes, $\alpha = 0.05$ is acceptable

Type I error

- Rejecting a true null hypothesis

- Probability of Type I error is α (the significance level)

Type II error

- Not rejecting a false null hypothesis
- The probability of a Type II error is β*.*
- The smaller β, the more *power* a test has.

Power

- The ability of a test to reject a false null hypothesis

- Power = 1- β

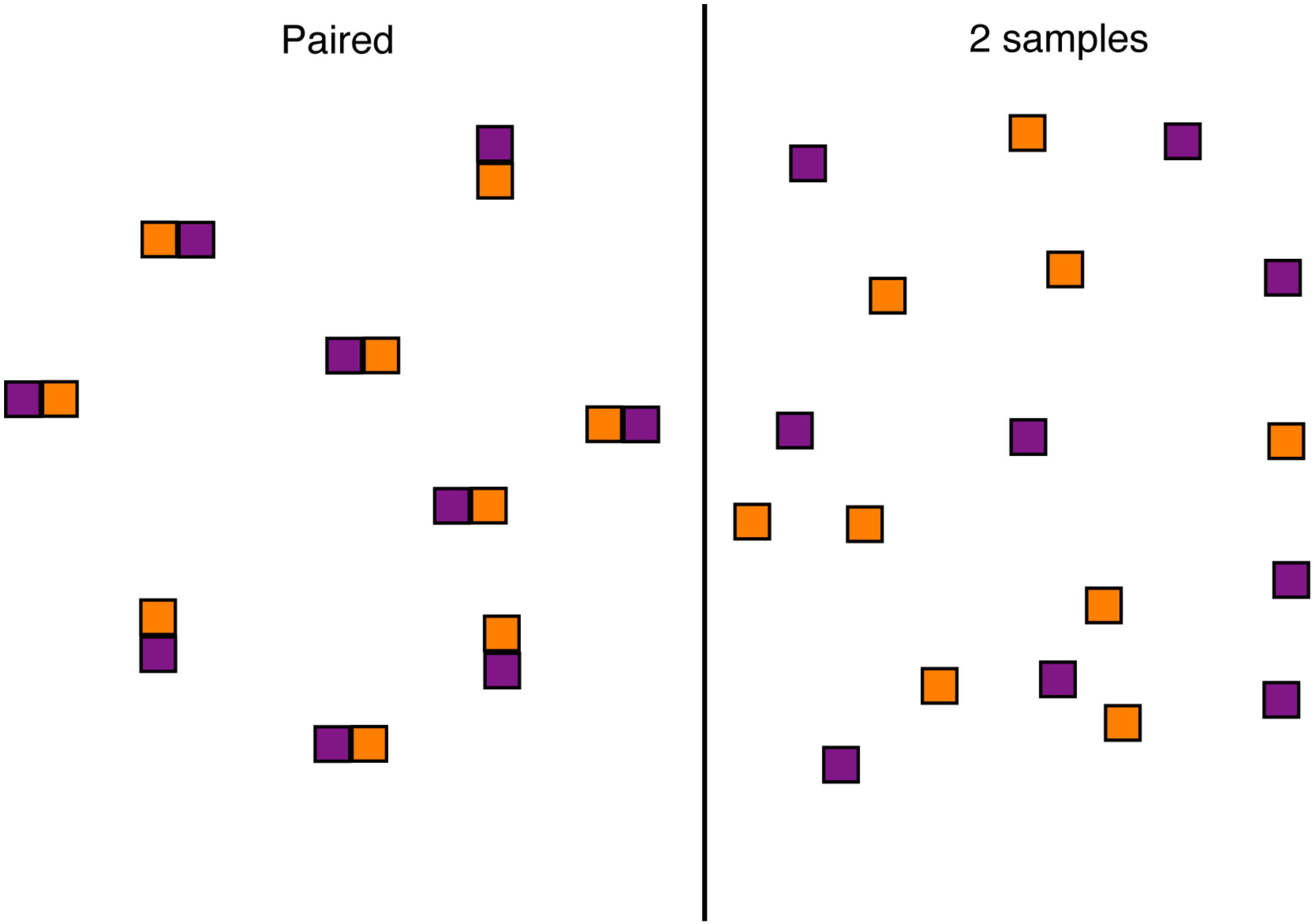
We never “accept the null hypothesis”

Comparing means

- Tests with one categorical and one numerical variable

- Goal: to compare the mean of a numerical variable for different groups.

Paired vs. 2 sample comparisons



Paired comparisons allow us to account for a lot of extraneous variation

2-sample methods are sometimes easier to collect data for

Paired *t* test

- Compares the mean of the differences to a value given in the null hypothesis

- For each pair, calculate the difference. The paired *t*-test is simply a one-sample *t*-test on the differences.

Assumptions of paired *t* test

- Pairs are chosen at random

- The differences have a normal distribution

It does *not* assume that the individual values are normally distributed, only the differences.

One-sample *t*-test

The *one-sample* t*-test* compares the mean of a random sample from a normal population with the population mean proposed in a null hypothesis.

Test statistic for one-sample *t*-test

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

μ_0 is the mean value proposed by H_0

Hypotheses for one-sample *t*-tests

H_0 : The mean of the population is $\mu_{0.}$

H_A: The mean of the population is not μ_0.

Comparing the means of two groups

Hypothesis test: 2-sample *t* test

Testing hypotheses about the difference in two means

2-sample *t*-test

The *two sample* t*-test* compares the means of a numerical variable between two populations.

2-sample t-test

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{SE_{\bar{Y}_1 - \bar{Y}_2}}$$

Hypotheses

H_0: There is no difference between the number of buds in the susceptible and resistant plants. ($\mu_1 = \mu_2$)

H_A: The resistant and the susceptible plants differ in their mean number of buds. ($\mu_1 \neq \mu_2$)

Comparing means when variances are not equal

Welch's *t* test

Welch's approximate t-test compares the means of two normally distributed populations that have unequal variances.

Welch's t

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

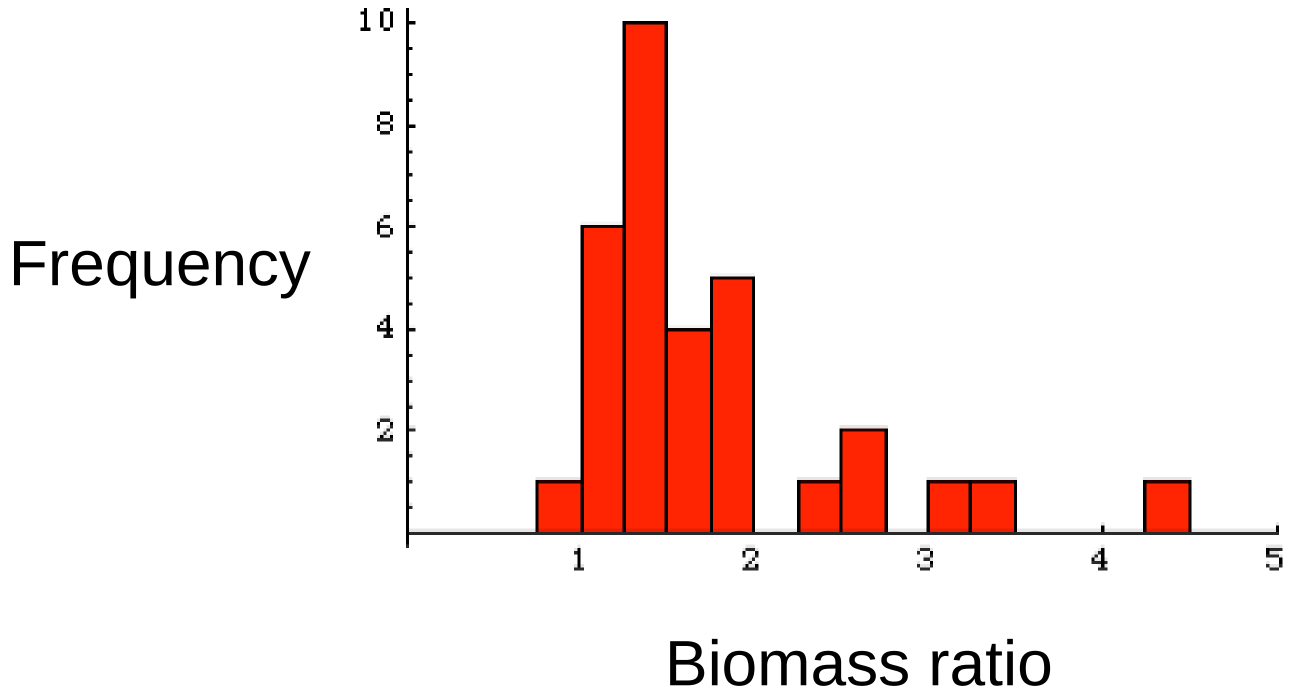
$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}\right)}$$

Round down *df* to nearest integer

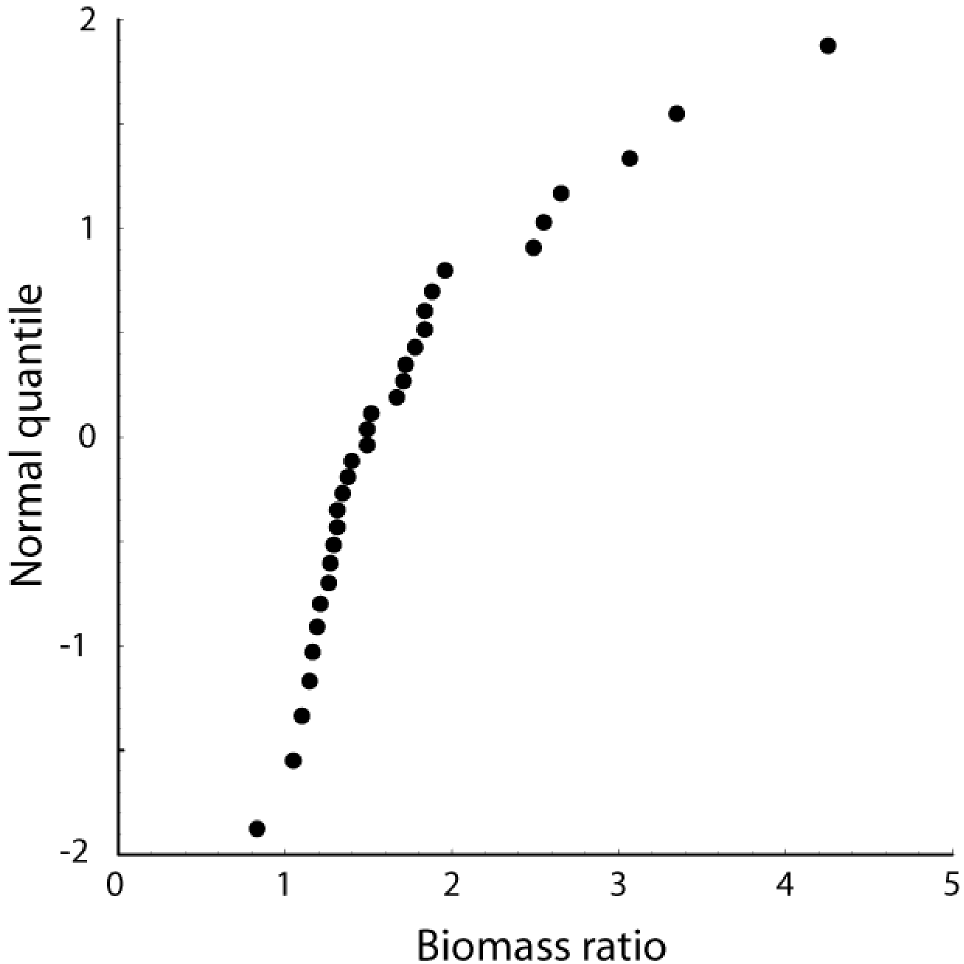
Detecting deviations from normality

- Previous data/ theory
- Histograms
- Quantile plots
- Shapiro-Wilk test

Detecting deviations from normality: by histogram

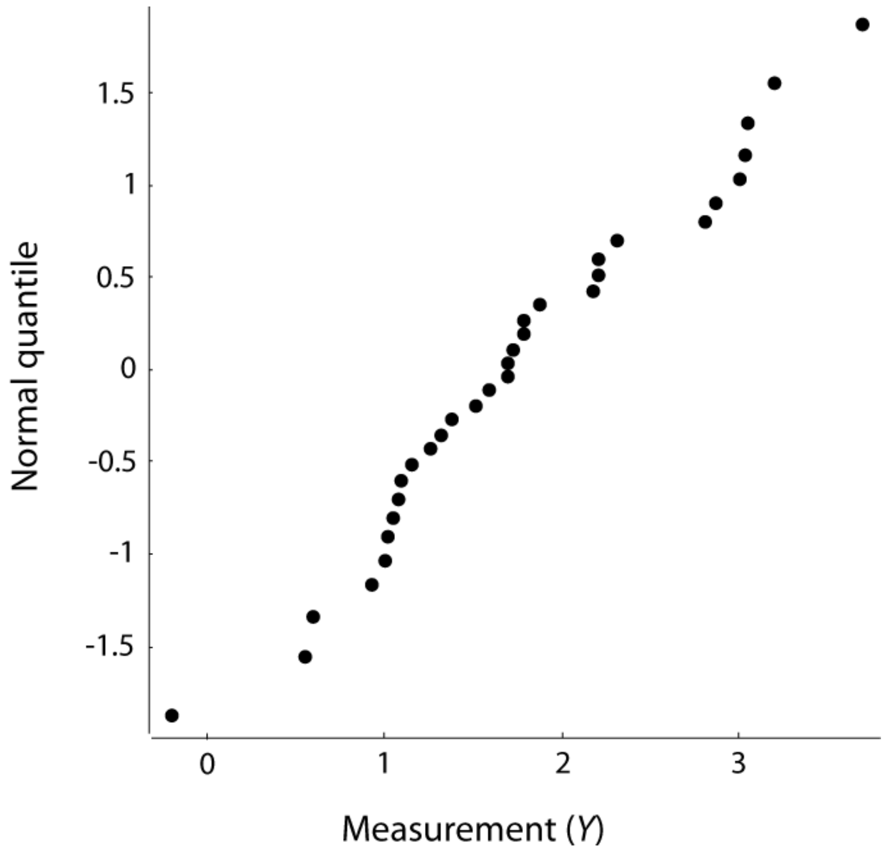


Detecting deviations from normality: by quantile plot



Detecting deviations from normality: by quantile plot

Normal data



What to do when the assumptions are not true

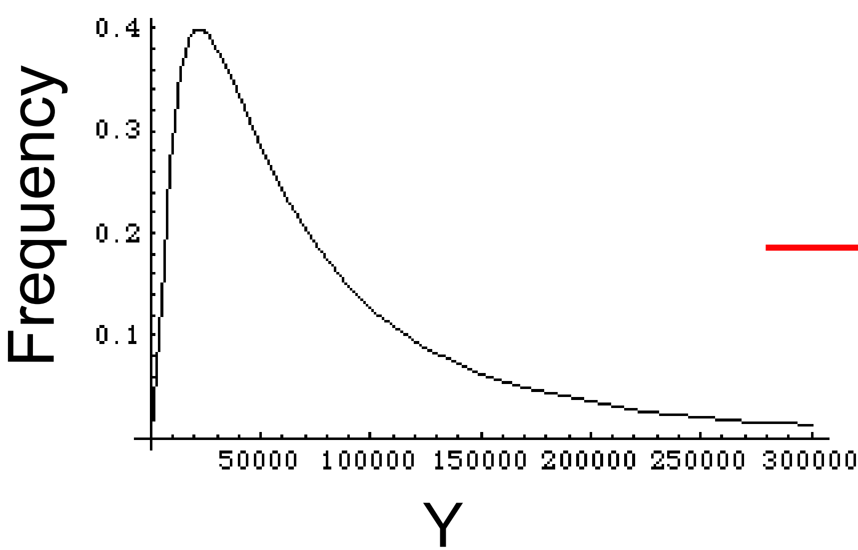
- If the sample sizes are large, sometimes the parametric tests work OK anyway
- Transformations
- Non-parametric tests
- Randomization and resampling

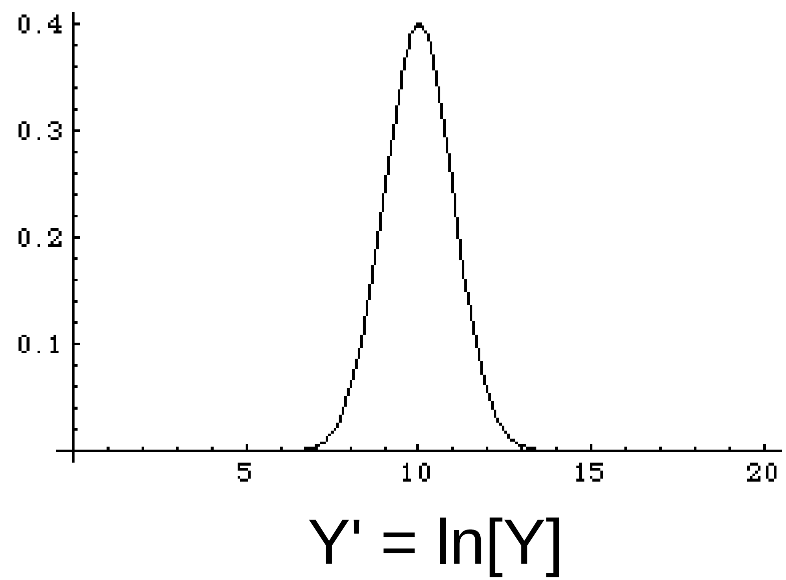
Data transformations

A *data transformation* changes each data point by some simple mathematical formula.

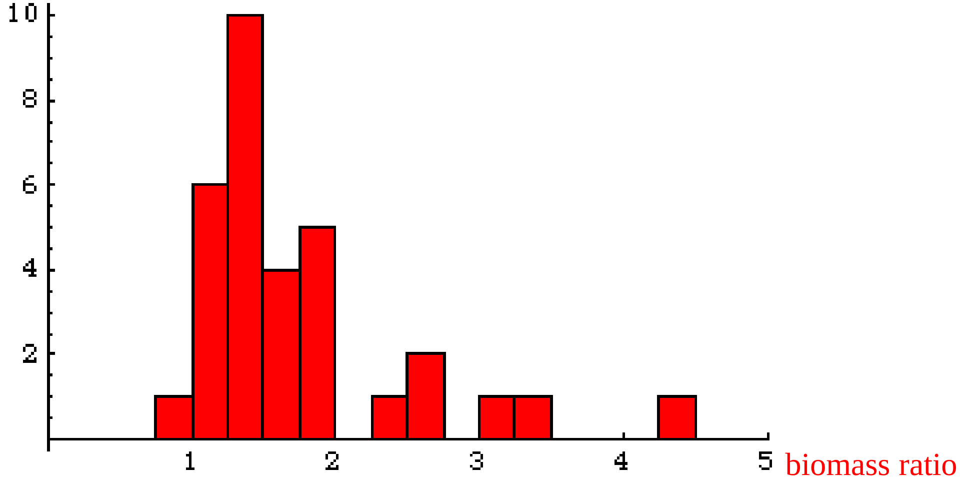
Log-transformation

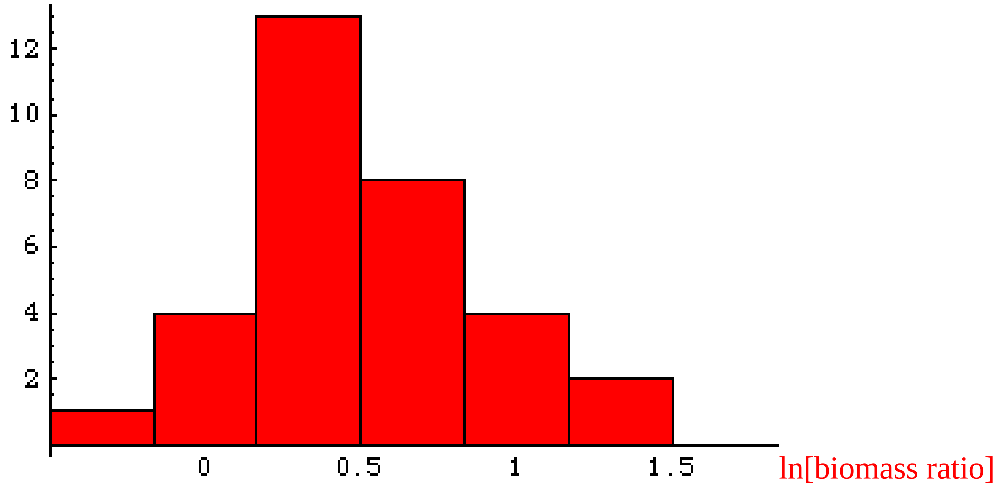
$$Y' = \ln[Y]$$





Biomass ratio	ln[Biomass Ratio]
1.34	0.30
1.96	0.67
2.49	0.91
1.27	0.24
1.19	0.18
1.15	0.14
1.29	0.26





Carry out the test on the transformed data!

Other transformations

Arcsine	$p' = \arcsin\left[\sqrt{p}\right]$
Square-root	$Y' = \sqrt{Y + 1/2}$
Square	$Y' = Y^2$
Reciprocal	$Y' = \dfrac{1}{Y}$
Antilog	$Y' = e^Y$

Valid transformations…

- Require the same transformation be applied to each individual
- Have one-to-one correspondence to original values
- Have a monotonic relationship with the original values (e.g., larger values stay larger)

Non-parametric methods

- Assume less about the underlying distributions
- Also called "distribution-free"
- "Parametric" methods assume a distribution or a parameter

Sign test

- Non-parametric test
- Compares data from one sample to a constant
- Simple: for each data point, record whether individual is above (+) or below (-) the hypothesized constant.
- Use a binomial test to compare result to 1/2.

Hypotheses

H_0: The median difference in number of species between singly-mating and multiply-mating insect groups is 0.

H_A: The median difference in number of species between these groups is not 0.

The sign test has very low power

So it is quite likely to *not* reject a *false* null hypothesis.

Most non-parametric methods use RANKS

- Rank each data point in all samples from lowest to highest

- Lowest data point gets rank 1, next lowest gets rank 2, ...

Permutation tests

- Also known as “randomization tests”
- Used for hypothesis testing on measures of association
- Mixes the real data randomly
- Variable 1 from an individual is paired with variable 2 data from a randomly chosen individual. This is done for all individuals.
- The estimate is made on the randomized data.
- The whole process is repeated numerous times. The distribution of the randomized estimates is the null distribution.

Without replacement

- Permutation tests are done without replacement.

- In other words, all data points are used exactly once in each permuted data set.

Permutation can be done for any test of association between two variables

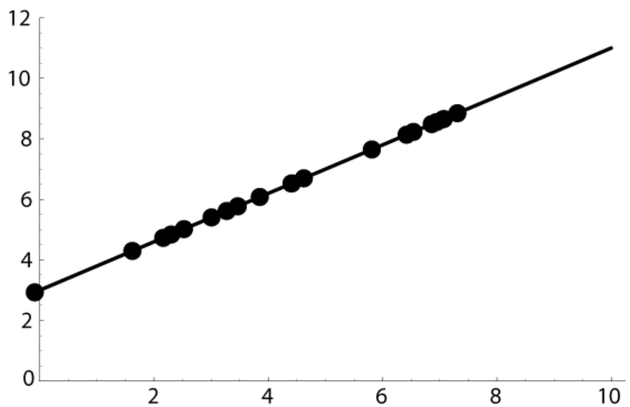
Linear Regression

Regression

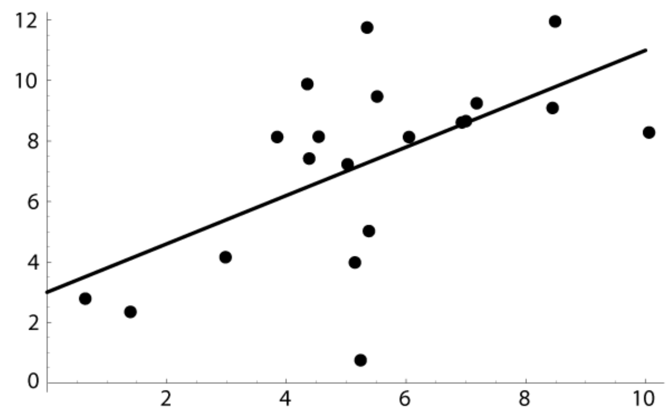
- Predicts Y from X

- Linear regression assumes that the relationship between X and Y can be described by a line

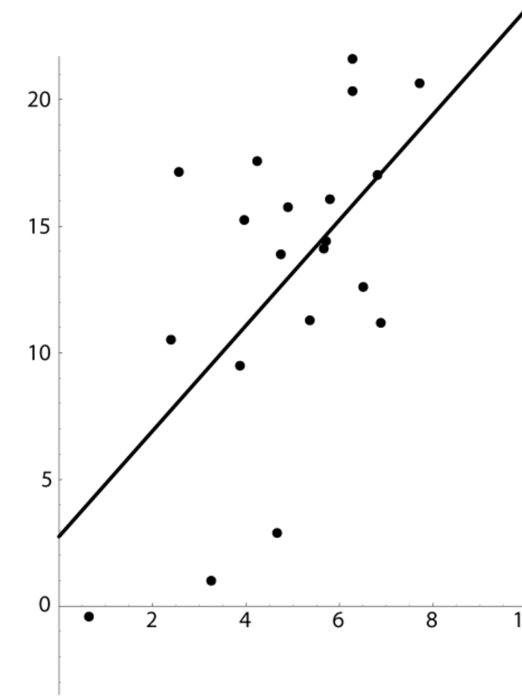
Correlation vs. regression



$r = 1$; $Y = 3 + 0.8X$



$r = 0.6$; $Y = 3 + 0.8X$



$r = 0.6$; $Y = 3 + 2\,X$

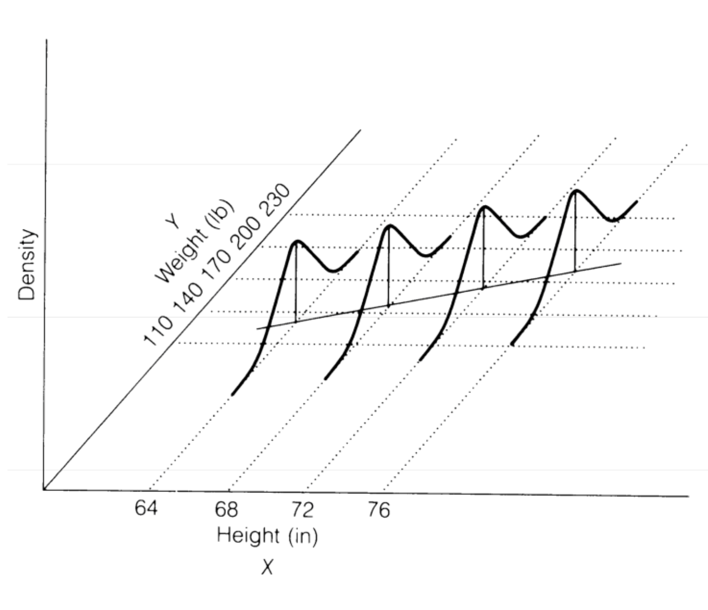
← Different correlation; same slope

Same correlation; different slope →

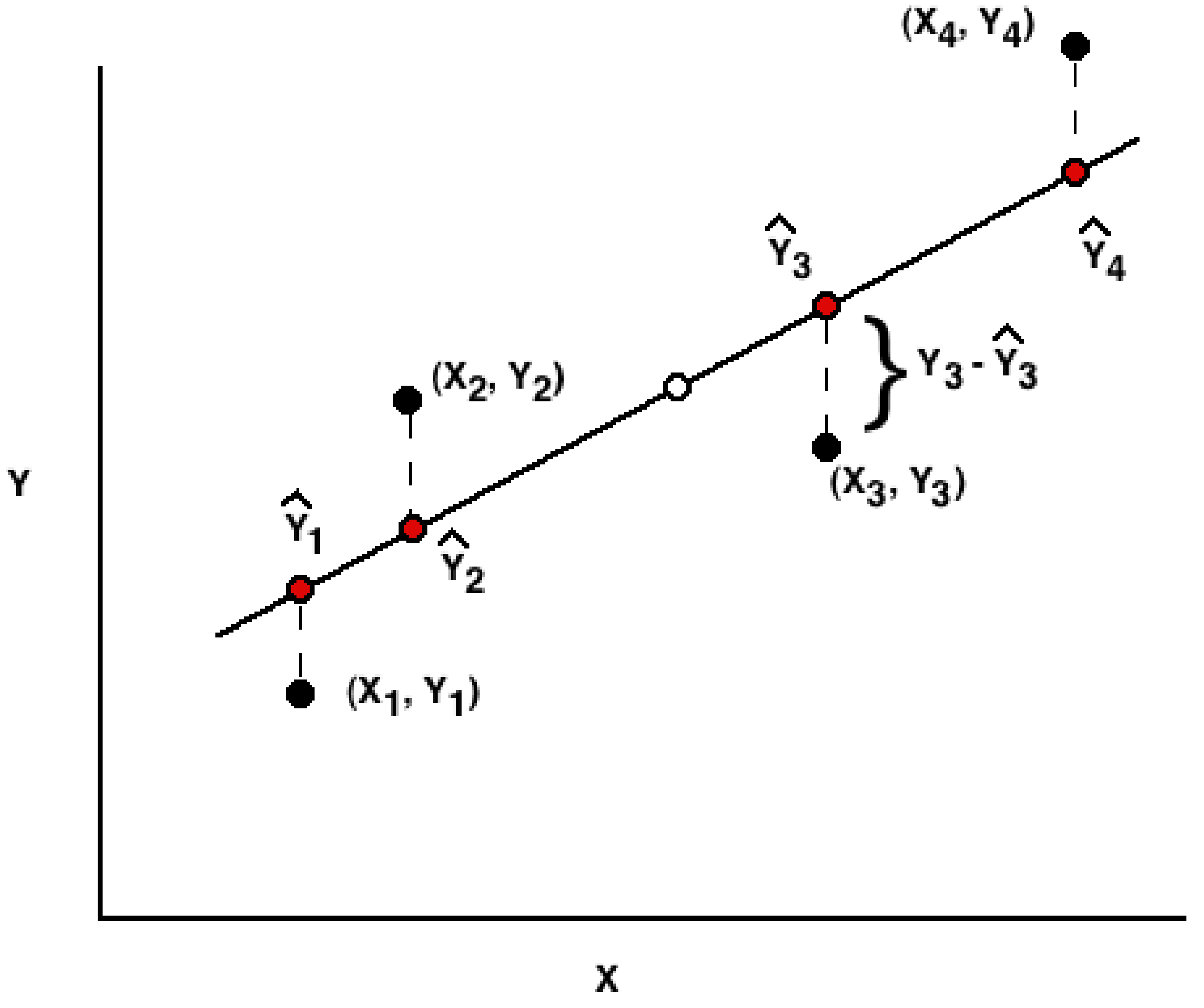
Regression assumes...

- Random sample

- *Y* is normally distributed with equal variance for all values of *X*



Nomenclature

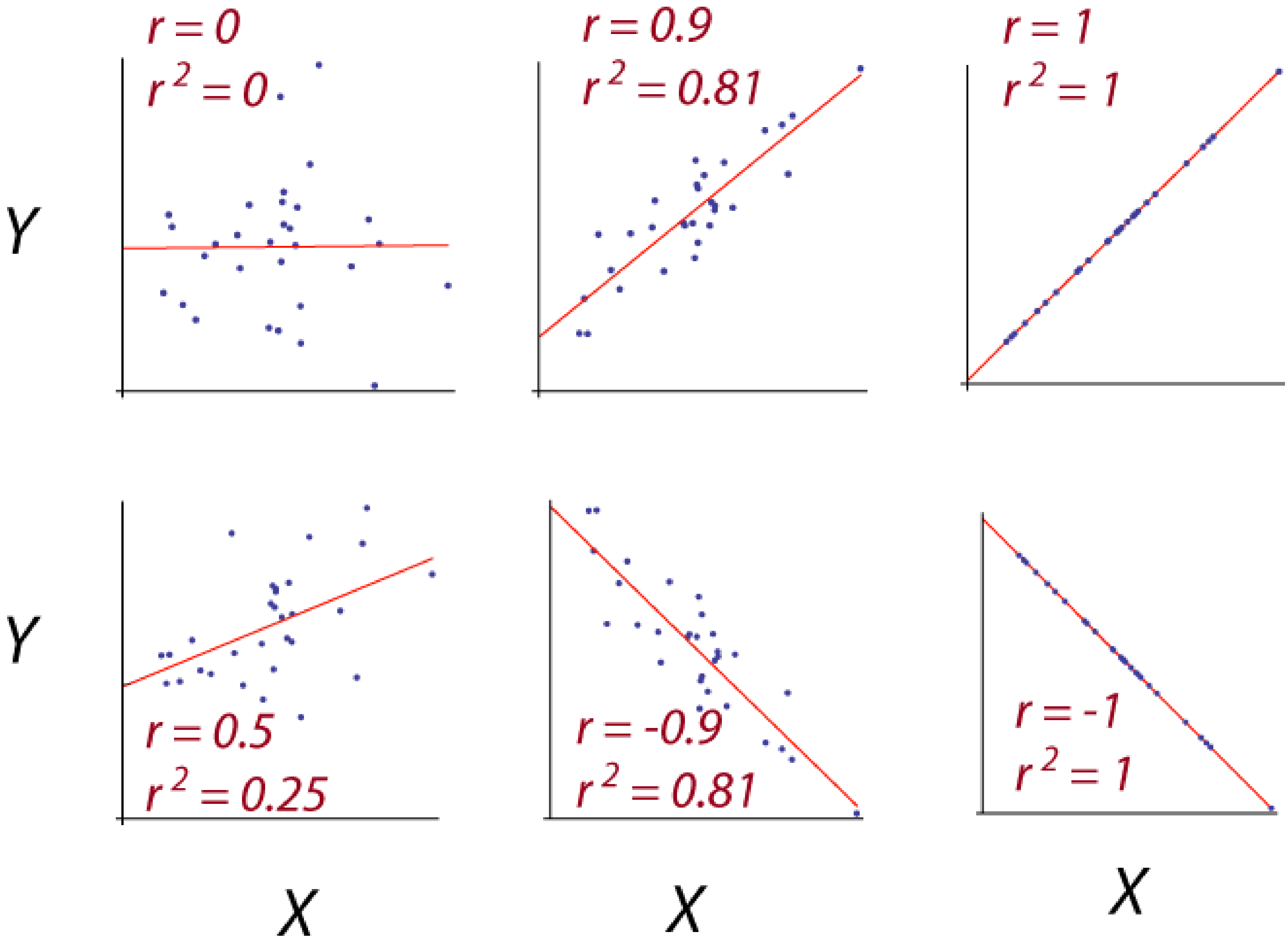


Residual:

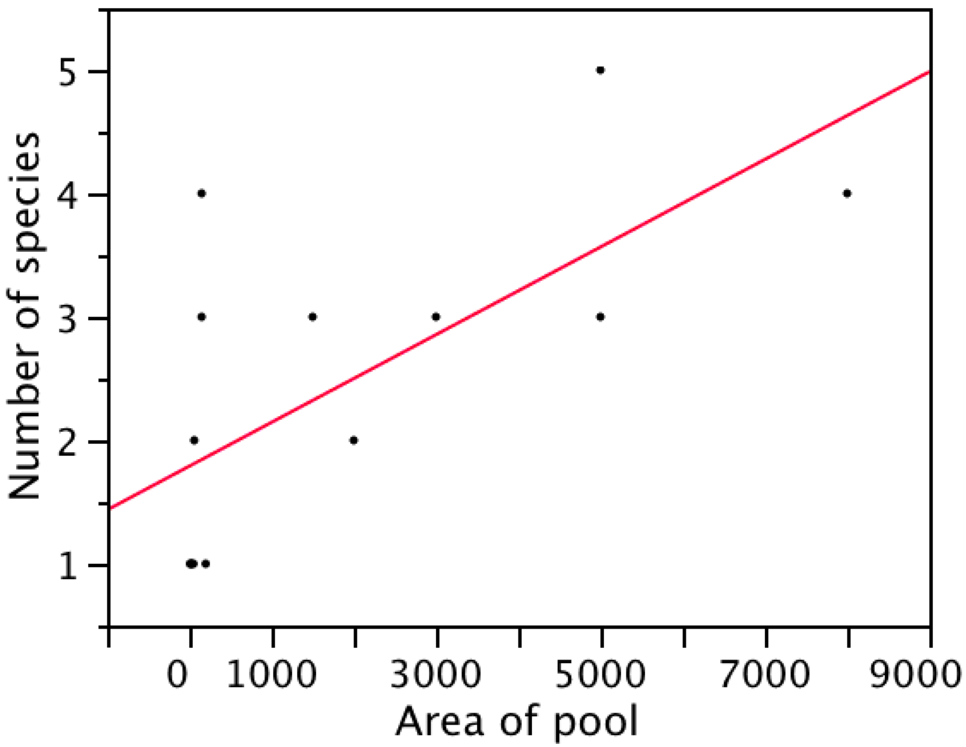
$$Y_i - \hat{Y}_i$$

Finding the "least squares" regression line

Minimize: $$SS_{residual} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$


$r = 0$
$r^2 = 0$
$r = 0.9$
$r^2 = 0.81$
$r = 1$
$r^2 = 1$
Y
$r = 0.5$
$r^2 = 0.25$
$r = -0.9$
$r^2 = 0.81$
$r = -1$
$r^2 = 1$
Y
X
X
X

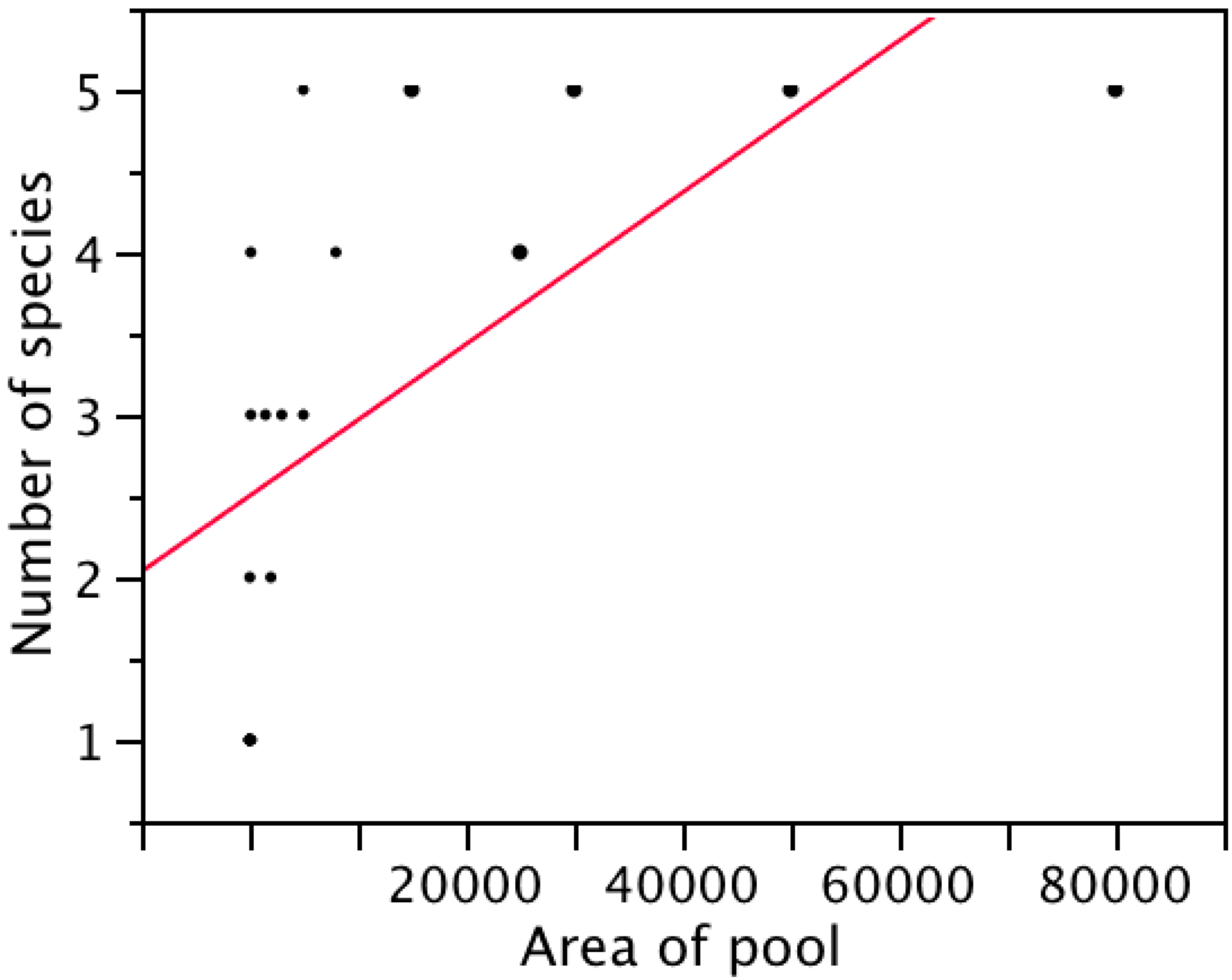
Caution: It is unwise to extrapolate beyond the range of the data.



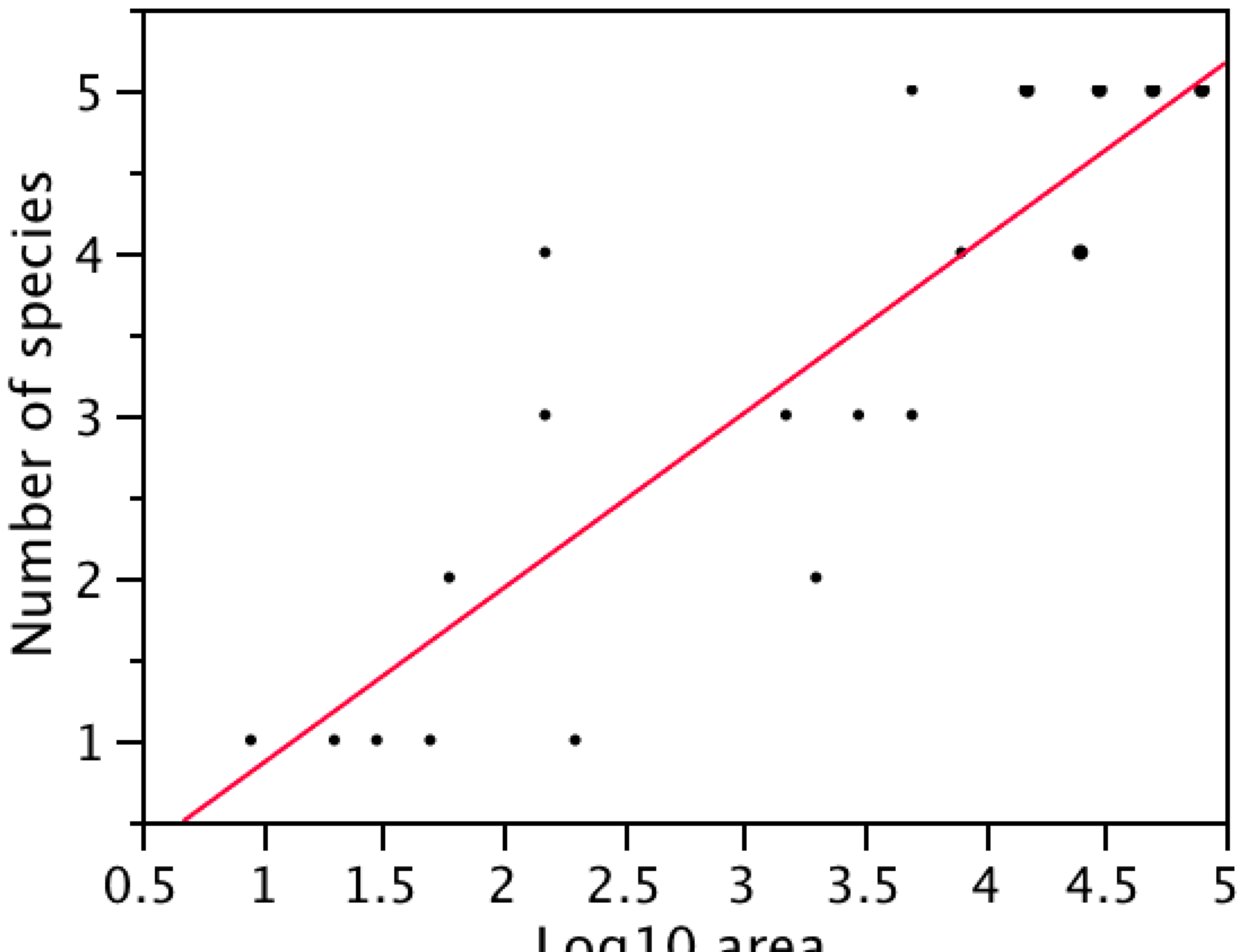
Number of species of fish as predicted by the area of a desert pool

If we were to extrapolate to ask how many species might be in a pool of 50000m^2, we would guess about 20.

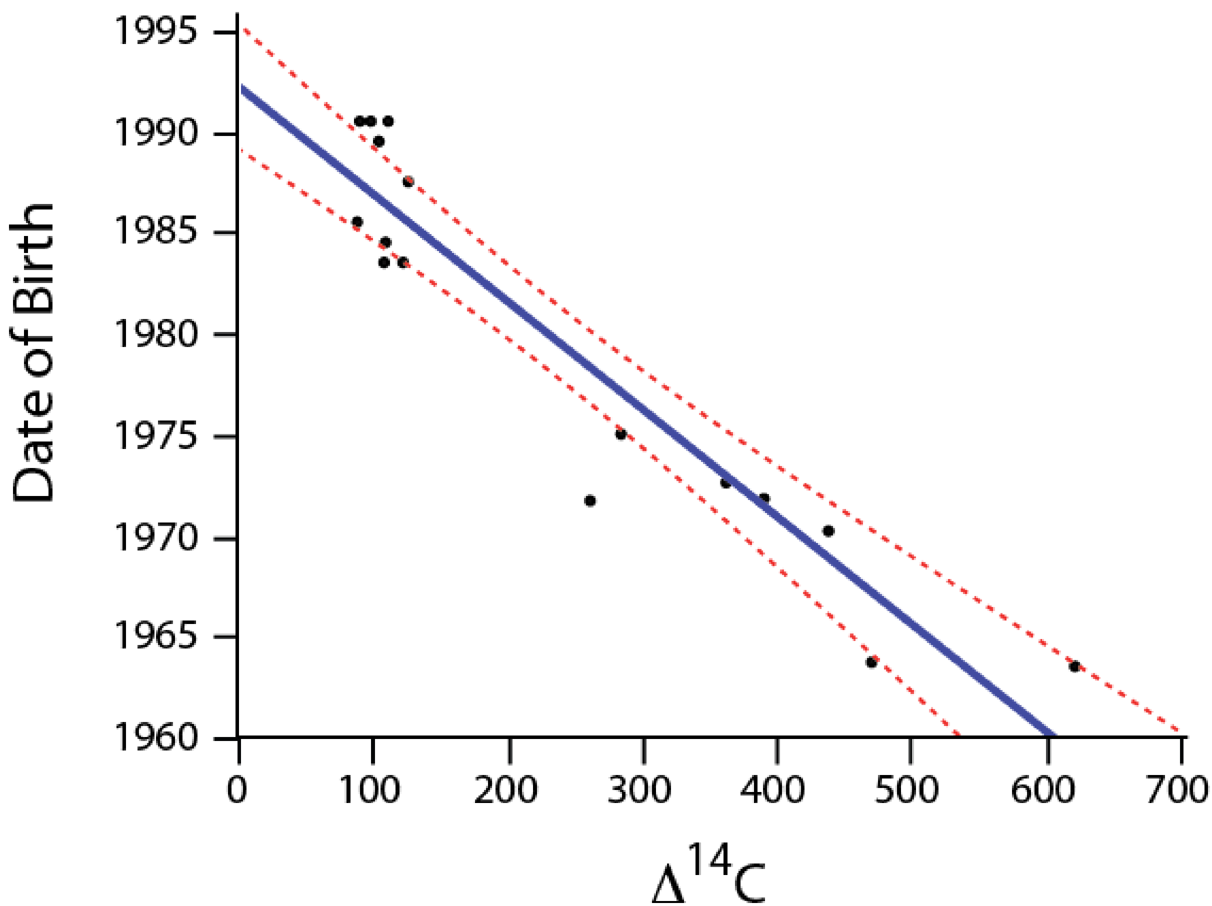
More data on fish in desert pools



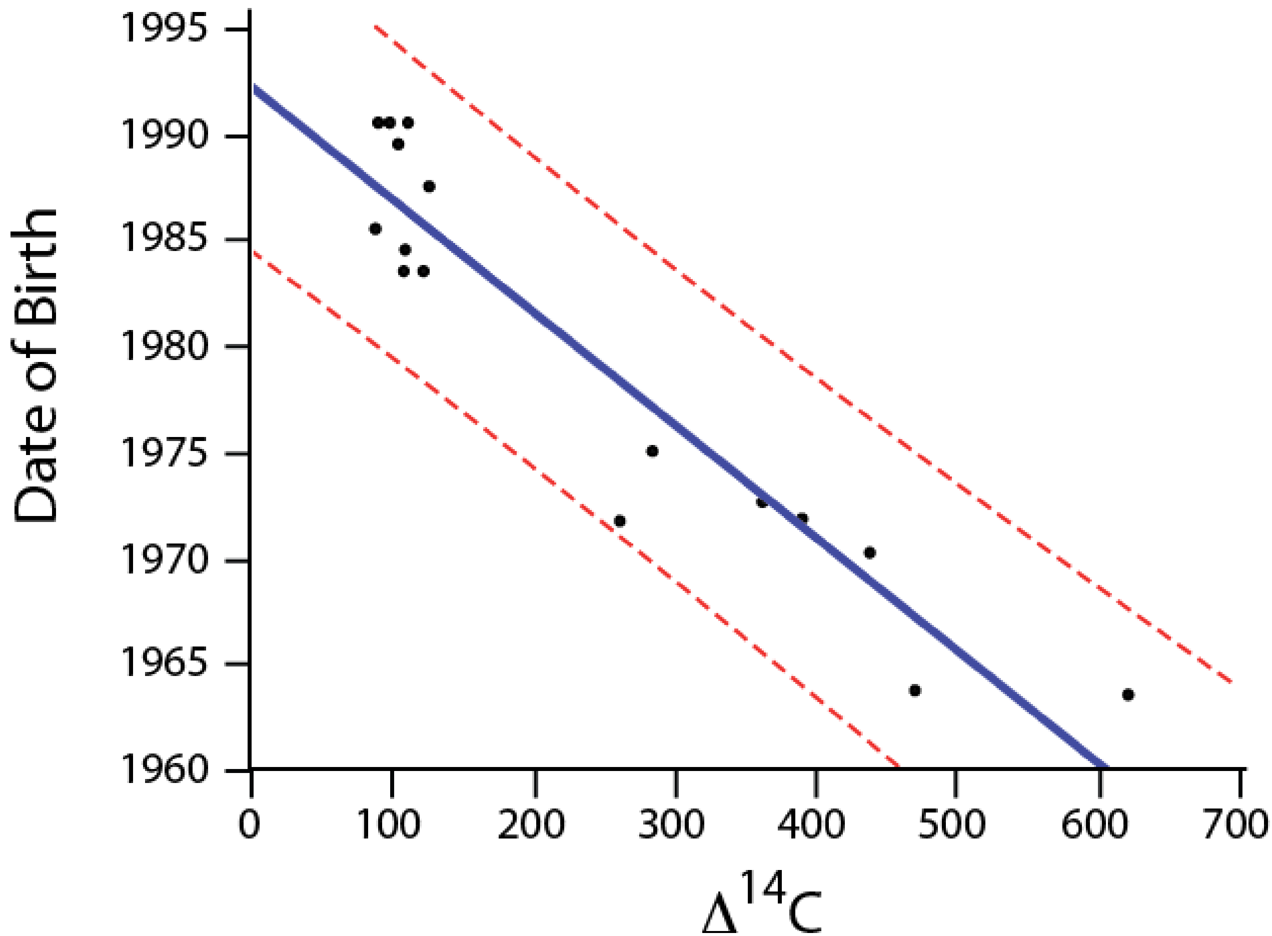
Log transformed data:



Confidence bands: confidence intervals for predictions of mean Y

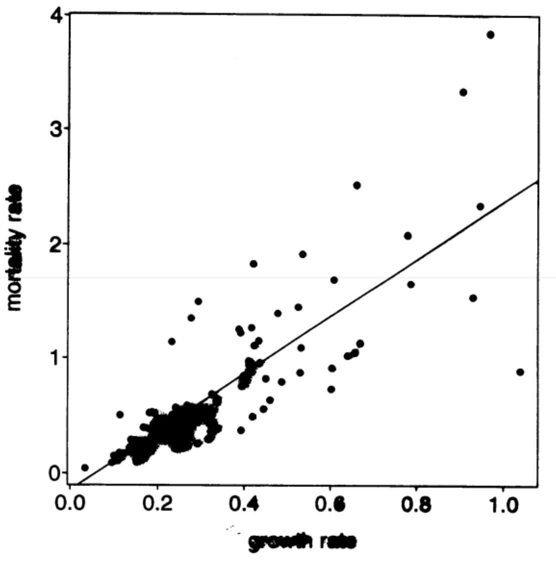


Prediction intervals: confidence intervals for predictions of individual *Y*

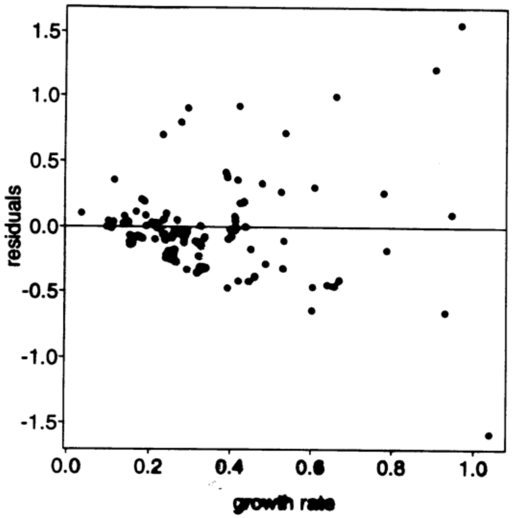


Residual plots help assess assumptions

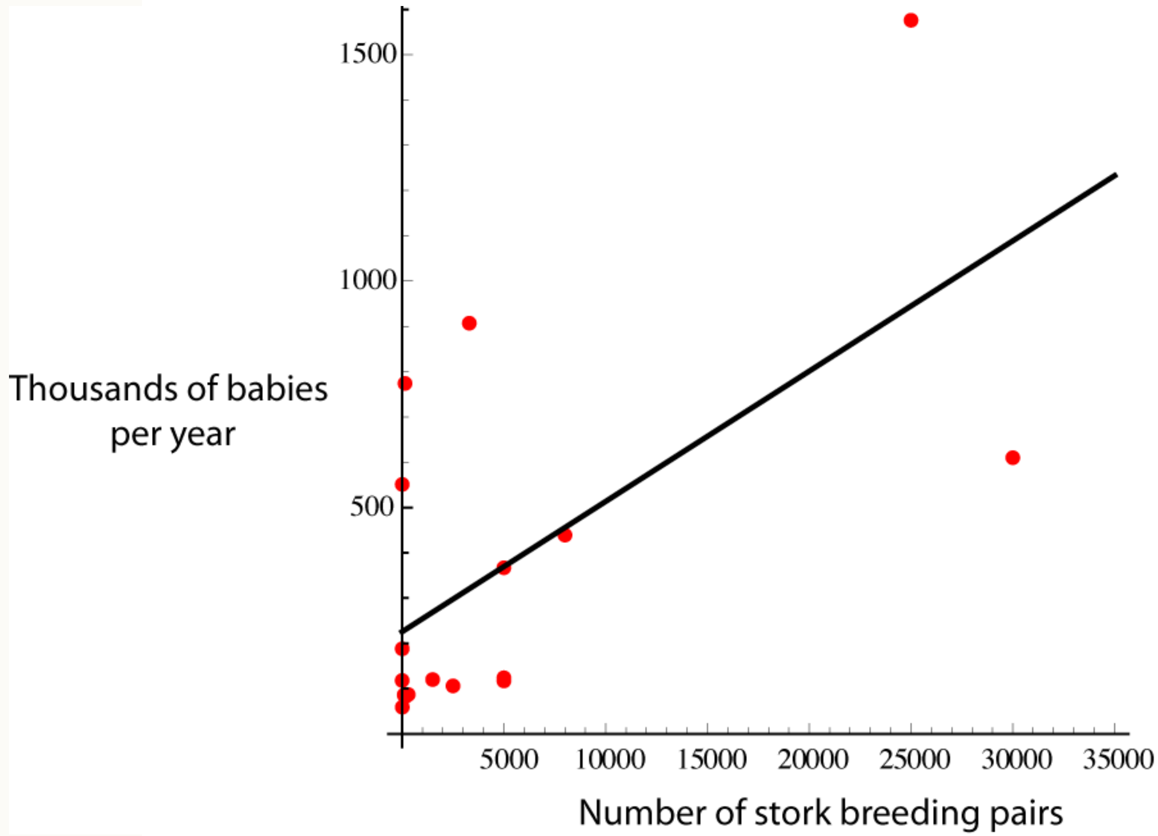
Original:

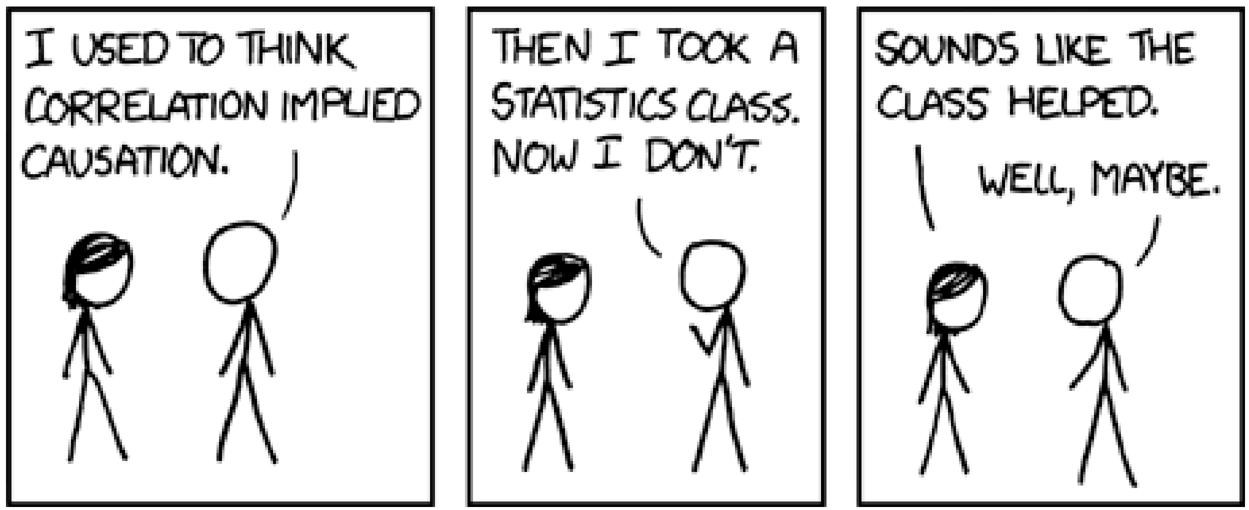


Residual plot



Correlation does not automatically imply causation




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

Statistical significance ≠ Biological importance

<table>
<tr><th></th><th>Important</th><th>Unimportant</th></tr>
<tr><td>Significant</td><td>Polio vaccine reduces incidence of polio</td><td>Things you don’t care about, or already well known things:

BRIEFS
Study Shows Frequent Sex Enhances Pregnancy Chances
By The Associated Press
BOSTON — A study that researchers say gives the best estimate ever of nature's window of female fertility comes to a not-so-startling conclusion: The best way to make babies is to have lots of sex.
Every other day is better than once a week. Daily is best.
The discovery, however, is not
ples don't want to use other forms of birth control.
Researchers say there are six days in every menstrual month when a woman can get pregnant.
Until now, estimates of women's fertility ranged from two days in a menstrual cycle to 10 or more. The new study found that conception is possible if a woman has intercourse on the five days before ovulation as well as on the day</td></tr>
<tr><td>Insignificant</td><td>Small study shows a possible effect, leading to larger study which finds significance.
or
Large study showing no effect of drug that was thought to be beneficial.</td><td>Studies with small sample size and high P-value
or
Things you don’t care about</td></tr>
</table>