Inference for two Population Means

Bret Hanlon and Bret Larget

Department of Statistics
University of Wisconsin—Madison

October 27–November 1, 2011

Case Study

Case Study

Example 19.2 on page 545 of the text describes an experiment with pseudoscorpions of the species *Cordylochernes scorpiodes* which live in the tropics. The females typically mate multiple times, even though mating once provides sufficient sperm for fertilization. Researchers conducted an experiment to examine a possible evolutionary explanation for this behavior. If there is genetic incompatibility between some pairs, a female may be more fertile by having multiple partners. In the experiment, females were randomly assigned to one of two treatment groups. In one group, a female was mated with the same male twice; in the other group, two different males were mated to the same female. The response variable is the number of successful broods. This variable is a small integer, ranging from 0 in some cases to a maximum of observed value of 7.

Data

- The data set is small enough that we can display it.
- The mean of the 20 values in the *Same* treatment group is 2.2.
- The mean of the 16 values in the *Different* treatment group is 3.6.

Group	Number of Successful Broods
Same	0 0 0 0 1 1 2 2 2 2 2 2 3 3 3 3 4 4 4 6
Different	0 1 2 2 2 3 3 4 4 4 4 4 6 6 6 7

Questions

- Do female pseudoscorpions have more successful broods on average, when they have multiple partners?
- The real biological question of interest involves the pseudoscorpions in their natural environment.
- The experimental setting seeks to explore the question in a controlled setting.
- In the experiment, the sample means are 2.2 and 3.6; what does this imply about the populations?
- Here, there is a single biological population of pseudoscorpions that can be thought of as two statistical populations on the basis of assignment to experimental treatment groups.

A statistical model

- A statistical model for the experiment is that there are two probability distributions for the number of successful broods, one for each treatment group.
- The specific distributions are not specified, but each is summarized my a mean or expected value, say μ_1 for the *Same* treatment group and μ_2 for the *Different* treatment group.
- The null hypothesis is $\mu_1 = \mu_2$; the biologically interesting alternative here is $\mu_1 < \mu_2$.
- How can we test this hypothesis?
- For the sample estimates, $2.2 < 3.6$, but what can we infer about the populations?

The Big Picture

- We have two populations with means μ_1 and μ_2.
- We have independent samples from each sample means $\bar{x}_1$ and $\bar{x}_2$ of sizes n_1 and n_2 respectively.
- We want to test $H_0\colon \mu_1 = \mu_2$ versus the alternative $H_A\colon \mu_1 < \mu_2$ with few assumptions about the distributions in the populations.
- The key idea of a *randomization test* is to consider the null distribution of the difference in sample means for all possible random samples assuming that the randomization is independent of the observed data.

Example

- In the example, there are 20 females in the *Same* treatment group and 16 in the *Different* treatment group.
- The observed difference in sample means is $2.2 - 3.625 = -1.425$ (without roundoff).
- What if the number of successful broods each female pseudoscorpion had was independent of the assignment to treatment group?
- If this were the case, we could compare the observed difference in sample means to the null distribution of differences in sample means for all other possible results of the randomization.
- There are $\binom{36}{20} = 7307872110$ possible ways to randomly separate 36 individuals into groups of size 20 and 16.
- Instead of finding exactly how many of these 7 billion+ possible randomization results have differences in sample means at least as extreme as the observed -1.425, we can use the computer to simulate the randomization process and estimate the p-value.
- Before beginning, we will examine methods to graph the data.

Types of Graphs

- When comparing two independent samples, we can use extensions of the types of graphs for single samples:
  - density plots;
  - histograms;
  - box-and-whisker plots;
  - dot plots.
- As this example data set is small, a dot plot is best because there is no compelling reason to summarize the data.
- Points should be *jittered* so equal values are not directly on top of one another.

Dot plot of the data
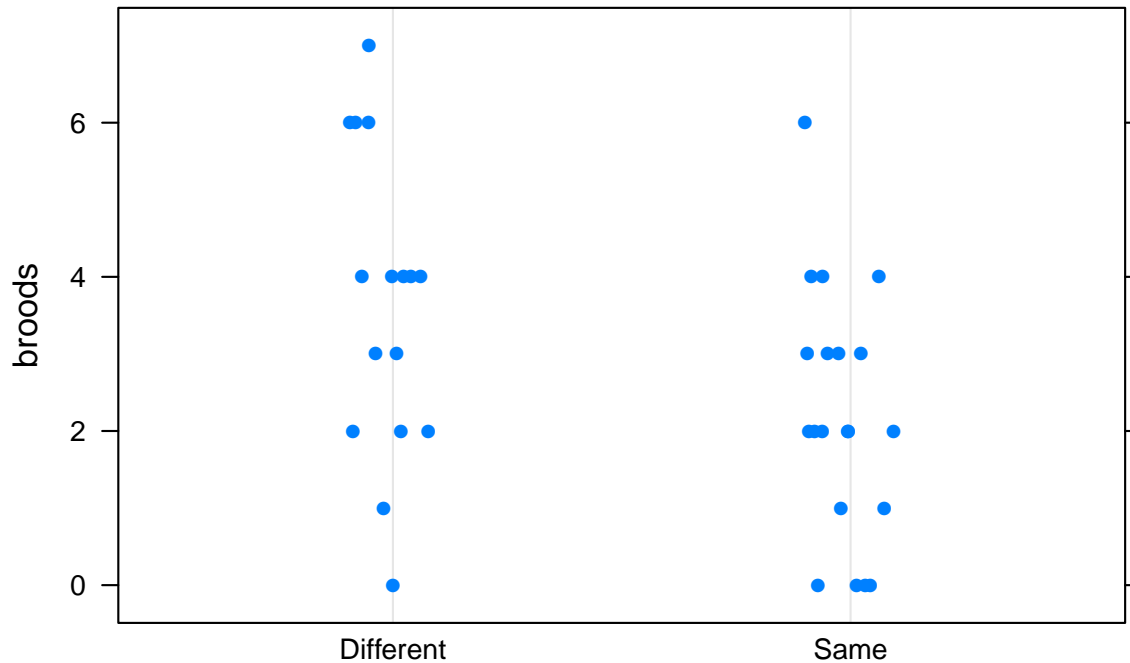
Comments on the Graphics

- There is a lot of overlap between the samples.
- It would be difficult to place an individual in one group or the other on the basis of the number of successful broods.
- But the centers of the distributions appear to be a bit different with generally larger values for the *Different* group on average.

Components of a Randomization Test

Randomization Tests

1. State hypotheses;
2. Select and calculate a test statistic;
3. Use simulation to find the null distribution of the test statistic;
4. Compare the value of the actual test statistic to its null distribution to compute a p-value;
5. Summarize the results in the context of the problem.

State Hypotheses

- Hypotheses are statements about populations;
- Here we are assuming that the pseudoscorpions in the sample may be treated as if they were randomly sampled from the population of these pseudoscorpions in the wild;
- In words, the hypotheses are:

  H_0: There would be no difference in the mean number of successful broods for each experimental condition among all female pseudoscorpions in the population.

  H_A: The experimental condition with different partners produces a larger mean number of successful broods than the experimental condition with the same partner mating twice.

State Hypotheses (cont.)

- In symbols, letting μ_1 and μ_2 represent the mean number of successful broods in the population for the *Same* and *Different* groups, respectively, the hypotheses are:

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 < \mu_2$$

- One could also test the alternative hypotheses $H_A: \mu_1 \neq \mu_2$ or $H_A: \mu_1 > \mu_2$ if appropriate for the setting.

Select a Test Statistic

- The difference in sample means is the natural test statistic for a hypothesis that compares population means.
- As we are determining the null distribution by simulation, there is no need to standardize the test statistic so it can be compared to some well-known benchmark distribution (such as standard normal, chi-square, or t).
- For the observed data, $\bar{x}_1 = 44/20 = 2.2$ and $\bar{x}_2 = 58/16 = 3.625$ and the difference is $\bar{x}_1 - \bar{x}_2 = -1.425$.

Compute the Null Distribution

- Conceptually, we take a random sample of 20 without replacement from the 36, compute its mean and the mean of the 16 remaining values, and take the difference.
- Repeat this process very many times and see how may differences are -1.425 or smaller.
- The proportion of such values is the p-value.

Graph of Null Distribution
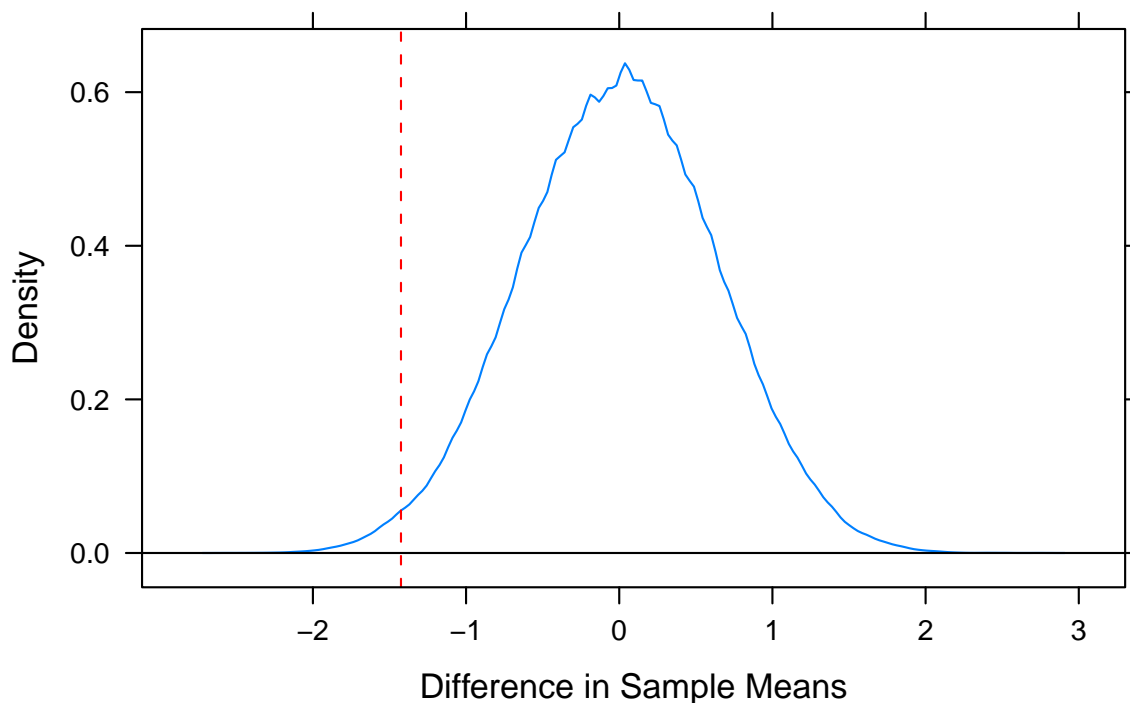
P-value

- It is evident from the graph that the observed difference is fairly unusual relative to the sampling distribution.
- The p-value is the actual proportion of sampled randomizations with a difference at least as extreme as that observed.
- For the 100,000 randomly selected randomizations, the p-value is estmated to be 0.0093.
- A different simulation would estimate this differently, but not by too much.
- The jaggedness of the preceding graph is caused by lots of ties in the randomization distribution of the difference between sample means.

Conclusions

- The p-value is fairly small.
- In the context of the problem, we can say this.

> *There is strong evidence that female pseudoscorpions have fewer successful broods when they mate with only one partner than when they mate with two partners under the given experimental conditions (two independent sample randomization test, $p = 0.009$, $n_1 = 20, n_2 = 16$). This result is consistent with the evolutionary explanation that the behavior of having multiple partners as seen in nature may overcome the possibility of genetic incompatibility among some partners.*

Summary

- Randomization tests are useful for comparing population means (or other population characteristics).
- The method simply considers the test statistic under the null hypothesis of independence between the randomization and the response variable of interest.
- As simulation determines the null distribution, there is no need to scale the test statistic so that it is comparable to a standard benchmark null distribution.
- Randomization tests are only practical with a computer.
- We see that the shape of the null distribution is symmetric and bell-shaped, which suggests that an approximation with a standard distribution may be accurate.
- Later in these notes we will reexamine this data using a *t-test*.

Two Different Designs

- There are two standard designs to compare two treatment groups;
  1. In a *paired design*, there is a single sample of pairs of observations and each treatment is applied to each sampled unit.

Example

In a sample of people, scores from vision tests are recorded separately for each eye. There is interest in comparing scores between right and left eyes.

  2. In a *two-independent-sample* design, there are two separate samples, and all elements of one sample get one treatment, all elements of the other sample get the other treatment.

Example

In a dairy cattle study with two different feed supplements, cattle are randomly separated into two groups and each group is given one of two dietary supplements. Average daily milk yield is compared between the groups.

Butterfat Study

This study and the accompaning data is from a former Statistics 571 student.

Case Study

The butterfat content in milk is an important factor in determining its economic value and in how it is processed to form dairy products such as cheese, ice cream, and butter. In an experiment, a company is interested in comparing the performances of two different labs which measure the butterfat content of milk. Two separate samples were collected from 107 loads of milk, and one sample from each load was sent to one of two labs. Butterfat content changes based on the identity of the cows, the time of milking, the time since the last milking, and other factors, so the percentage butterfat can be expected to vary from load to load, but should be consistent for samples taken from the same load as each load is properly agitated before sampling to promote mixing throughout the load.

How should this data be examined to compare the performances of the labs?

Horned Lizards

Case Study

The horned lizard *Phrynosoma mcalli* has horns it uses for protection. Researchers tested a hypothesis that longer horns are more protective then shorter horns. A predator of these lizards is the loggerhead shrike, a bird that impales the lizards on thorns or barbed wire. Researchers compared the horn lengths of 30 skewered lizards with 154 horned lizards that were living. The average length of the skewered lizards was 21.99 mm and the average length of the living lizards was 24.28 mm. Is this evidence that longer horns are more protective?

Graphs

- Density plots, histograms, dot plots, and box-and-whisker plots are all useful for graphical comparisons between samples.
- For paired samples, graphs of differences are also useful.
- It is very important to graph data and look for patterns or *outliers* before carrying out statistical inferences.

Butterfat Study

- The butterfat percentage data is from a *paired data design*.
- There is a single sample of size $n = 107$.
- Each sample unit (a load) is measured twice (the two butterfat measurements from the two labs).
- Paired data design data is analyzed by:
  - ▶ taking differences within each pair (same order);
  - ▶ analyzing the single sample of differences using one-sample methods.
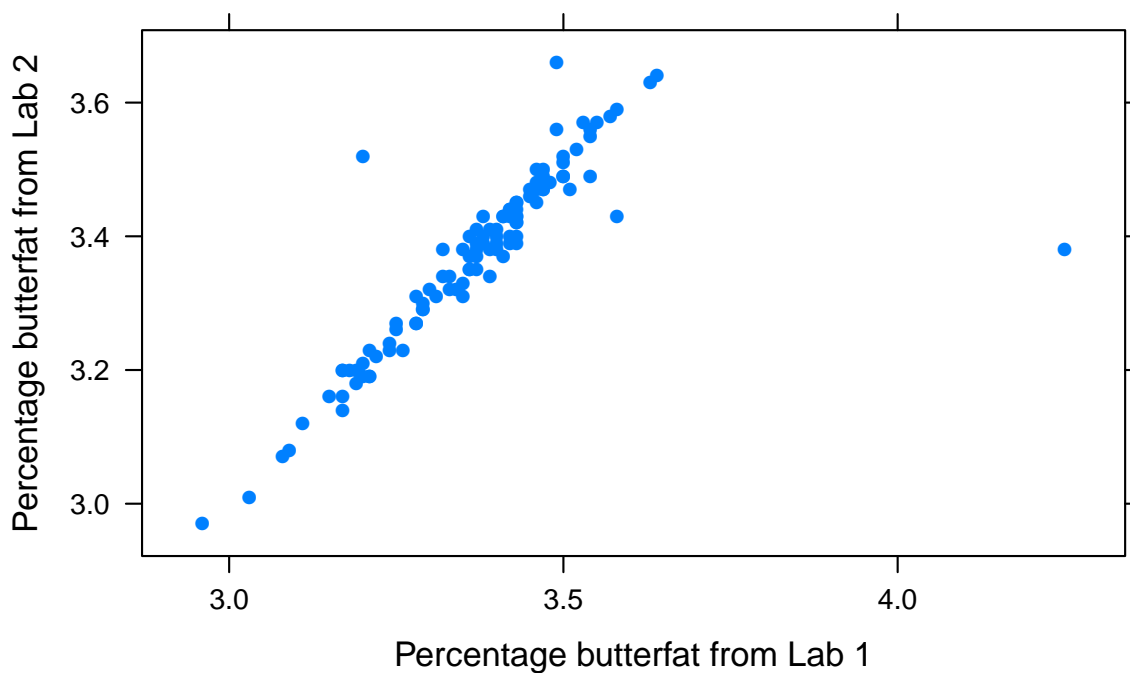
Scatter plot

- For paired data, a *scatter plot* of one measurement versus the other helps show the relationship and identify *potential outliers*.
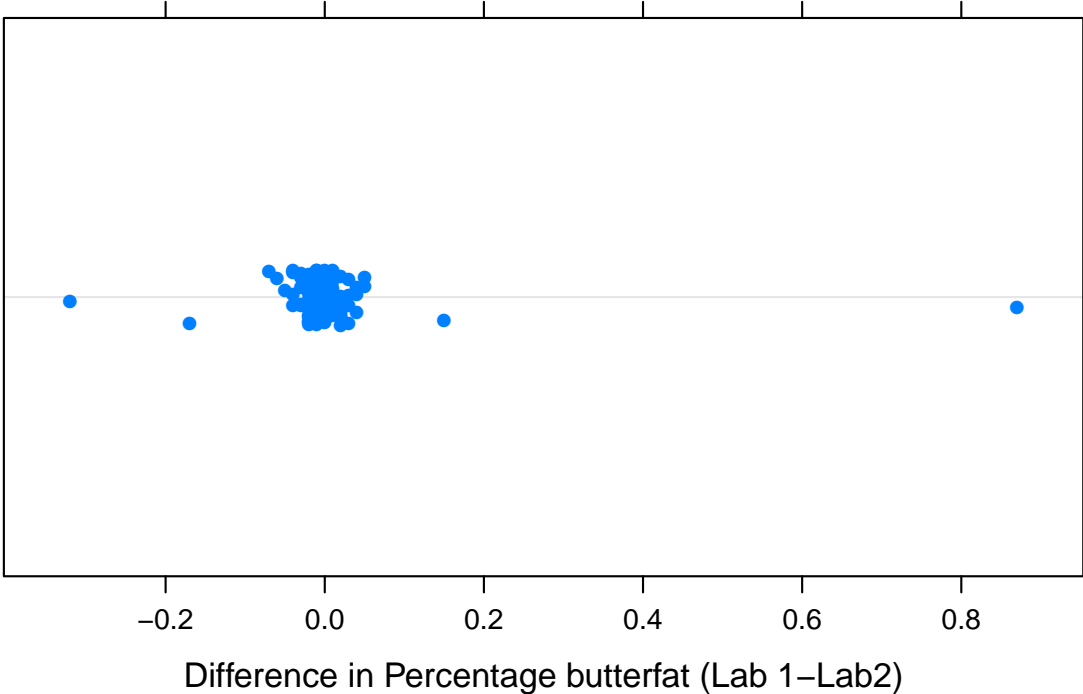
> **Definition**
>
> An *outlier* is a single observation that sticks out in a plot as unusual and not part of the general trend.

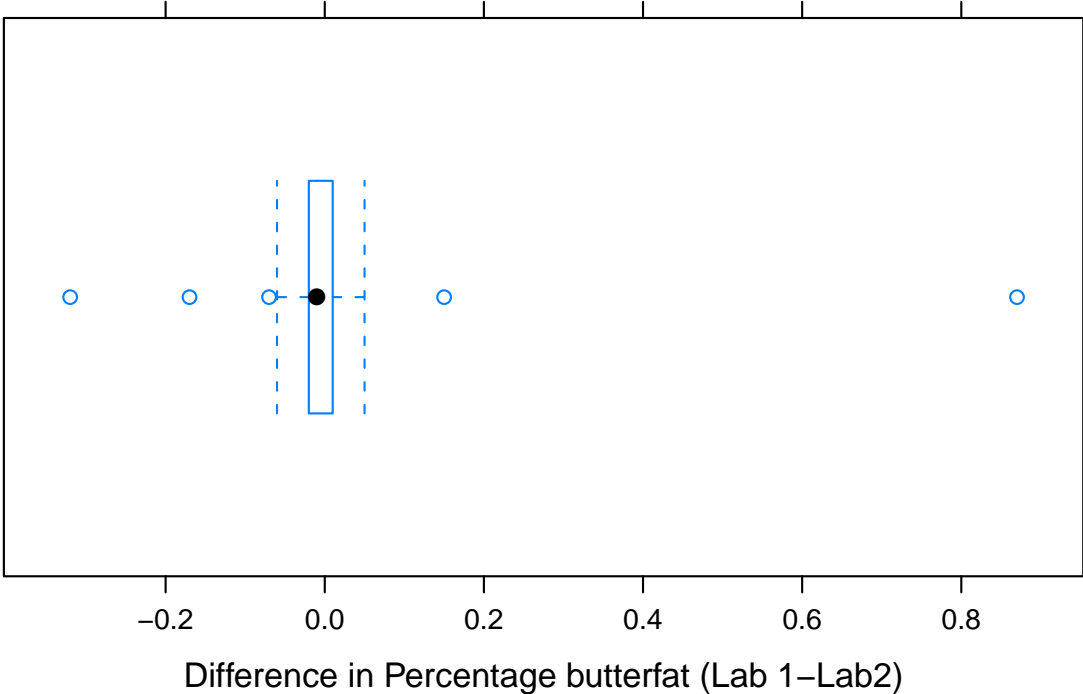There are multiple ways to deal with outliers in an analysis.
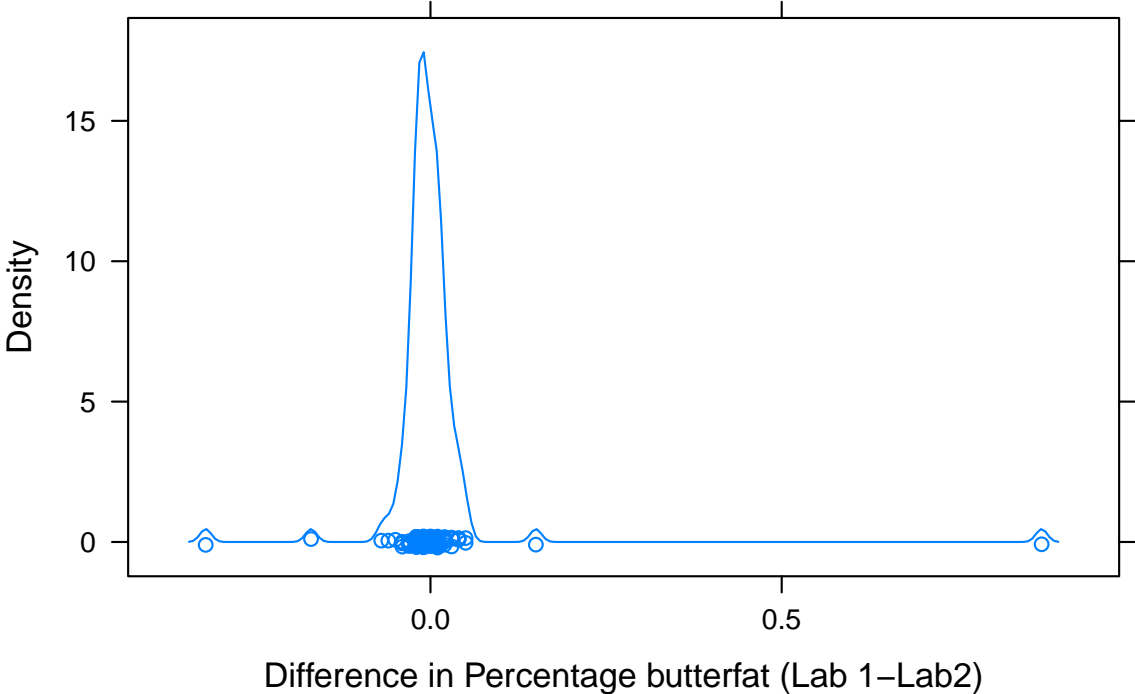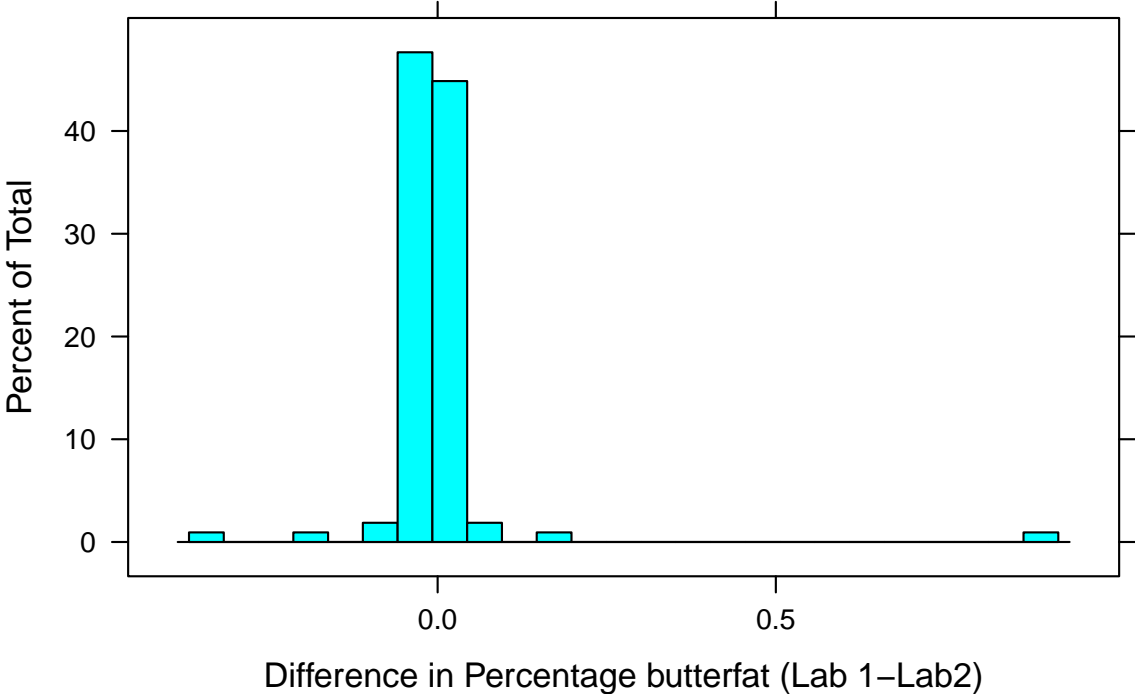
Scatter plot

Dot plot

Difference in Percentage butterfat (Lab 1–Lab2)

Box-and-Whisker Plot

Difference in Percentage butterfat (Lab 1–Lab2)

Density Plot

Histogram

Observations

- There are four individual measurements that are *outliers*.
- There are several possible explanations:
  1. data may be recorded incorrectly;
  2. the load may not have been agitated properly before sampling;
  3. one or both labs occasionally makes a poor measurement;
- With additional background information, we understand that the second explanation is most plausible, as some of the individuals that do the work of taking samples fail to properly agitate the milk, and as cream rises to the top, there is the possibility that two separate samples from the same load might differ considerably in percentage of butterfat.
- As these observations are likely not telling us about the performance of the laboratories, but about how a small part of the data is collected, and our desired inference is about the laboratories, it is reasonable to discard the outliers.

What to do with outliers

- For the purposes of this example in lecture, we will *discard the four outliers from further analysis*, as explained on the previous slide.
- For 103 of the 107 data points, differences between measurements are no more than 0.07 percent.
- For the other loads, the differences are at least twice as large, ranging from 0.15 to 0.87.
- If this was my data, I would seek explanations in each case for the discrepancy: (was the data recorded improperly, is there a note to cast suspicion on a particular measurement)
- Generally, one should not discard data unless there is a good reason.
- If I had more information that indicated that the observed measurements were genuinely accurate and just unusual, I would not discard the data.
- This example highlights *the importance of graphing data before conducting an analysis for inference*.
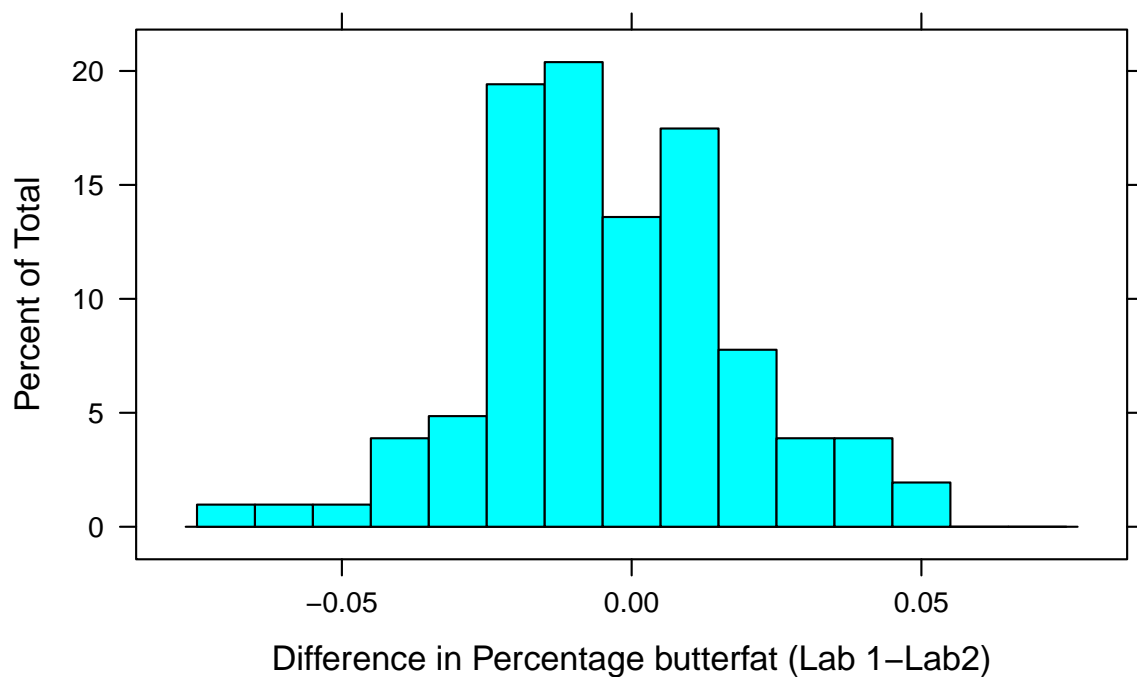
Dot plot (after outliers removed)

After Removal of Outliers

Difference in Percentage butterfat (Lab 1–Lab2)
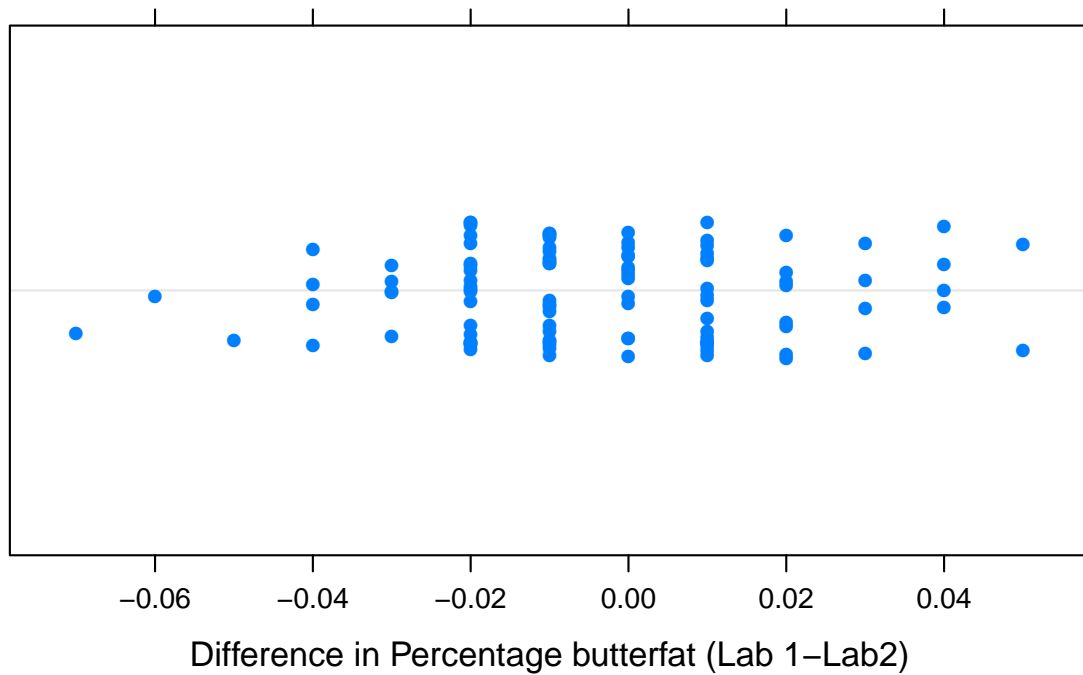
Histogram (after outliers removed)

After Removal of Outliers

Difference in Percentage butterfat (Lab 1–Lab2)

Estimation for Paired Designs

- For paired designs, inference is on the *differences in measurements*.
- The data is considered to be *a single sample of differences*.
- Confidence intervals use the single sample methods we have seen before;
    1. a t distribution method; or
    2. the bootstrap
- Here we will use the t distribution as the sample size is large enough to overcome even extreme nonnormality and the (remaining) data has a fairly symmetric distribution.

Confidence Intervals from Paired Designs

Confidence Interval for $\mu_D = \mu_1 - \mu_2$

A P% confidence interval for $\mu_D = \mu_1 - \mu_2$ has the form

$$\bar{D} - t^* \frac{s}{\sqrt{n}} < \mu < \bar{D} + t^* \frac{s}{\sqrt{n}}$$

where t^* is the critical value such that the area between $-t^*$ and t^* under a t-density with $n - 1$ degrees of freedom is $P/100$, where n is the sample size.

Chalkboard example

> **Example**
>
> - Here are summary statistics:
>   - ▸ The sample size is 103.
>   - ▸ The mean difference (Lab 1 - Lab 2) is -0.0037.
>   - ▸ The standard deviation of the differences (Lab 1 - Lab 2) is 0.0223.
>   - ▸ Find a 95% confidence interval using the formula on the board.
>
> $-0.008 < \mu_D < 7e - 04$

Compare to R Solution

```
> t.test(lab1, lab2, paired = T)

Paired t-test

data:  lab1 and lab2
t = -1.6811, df = 102, p-value = 0.0958
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0080422274  0.0006635867
sample estimates:
mean of the differences
          -0.003689320
```

Interpretation

After excluding observations with unexplained unusually large differences, we are 95% confident that the mean difference (Lab1 - Lab 2) in butterfat determination is between -0.008 and 0.001 in the common situation where there is an absence of large discrepancies in the measurements.

Comparison with outliers included

- An analysis with the outliers included would be appropriate if it were determined that the outliers were, in fact, genuinely correct.
- Notice that the interval still includes 0 (consistent with no differences in the labs), but that the confidence interval is much wider.

```
Paired t-test

data:  bfat[lab == "Lab1"] and bfat[lab == "Lab2"]
t = 0.1522, df = 106, p-value = 0.8793
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01686179  0.01966553
sample estimates:
mean of the differences
          0.001401869
```

Hypothesis Tests

- We can also formally test the null hypothesis that the mean difference in measurements in the two labs is zero.

$$H_0 \colon \mu_1 - \mu_2 = 0$$

$$H_A \colon \mu_1 - \mu_2 \neq 0$$

Paired t test

If differences $D_1, D_2, \ldots, D_n$ are normally distributed, then

$$T = \frac{\bar{D} - d_0}{s/\sqrt{n}} \sim t(n-1)$$

where d_0 (usually 0) is the mean difference in the null hypothesis, s is the sample standard deviation of differences, and n is the sample size.

Chalkboard example

Example

- Here are summary statistics:
  - ▹ The sample size is 103.
  - ▹ The mean difference (Lab 1 - Lab 2) is -0.0037.
  - ▹ The standard deviation of the differences (Lab 1 - Lab 2) is 0.0223.
  - ▹ Test the hypothesis of no difference between the labs.

$T = -1.68$, $p = 0.0958$, $df = 102$

Compare to R Solution

```
> t.test(lab1, lab2, paired = T)

Paired t-test

data:  lab1 and lab2
t = -1.6811, df = 102, p-value = 0.0958
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0080422274  0.0006635867
sample estimates:
mean of the differences
         -0.003689320
```

Two Independent Samples

- The lizard example is *not paired*.
- There is no reason to match a specific individual in one sample with an individual in the other.
- In fact, the two sample sizes are not even equal.
- Inference methods and graphs are different for independent samples.
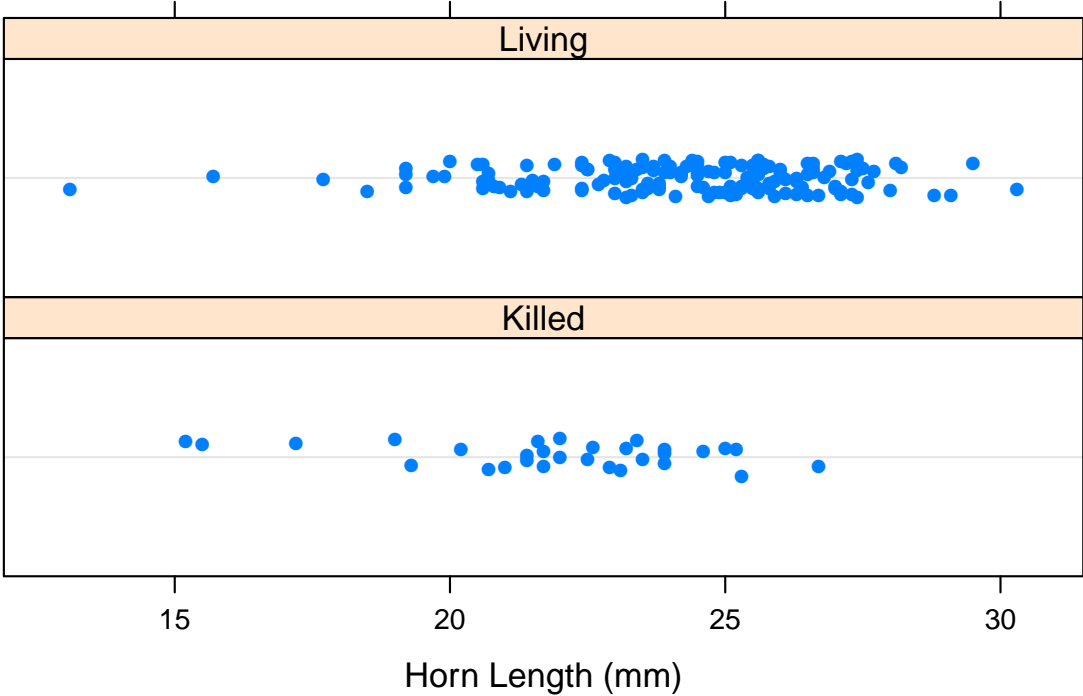
Graphics

- Separate graphics are produced for each separate sample.
- The graphs use a layout to make comparisons between them easy.
- Often, such graphs are side by side or on top of one another.
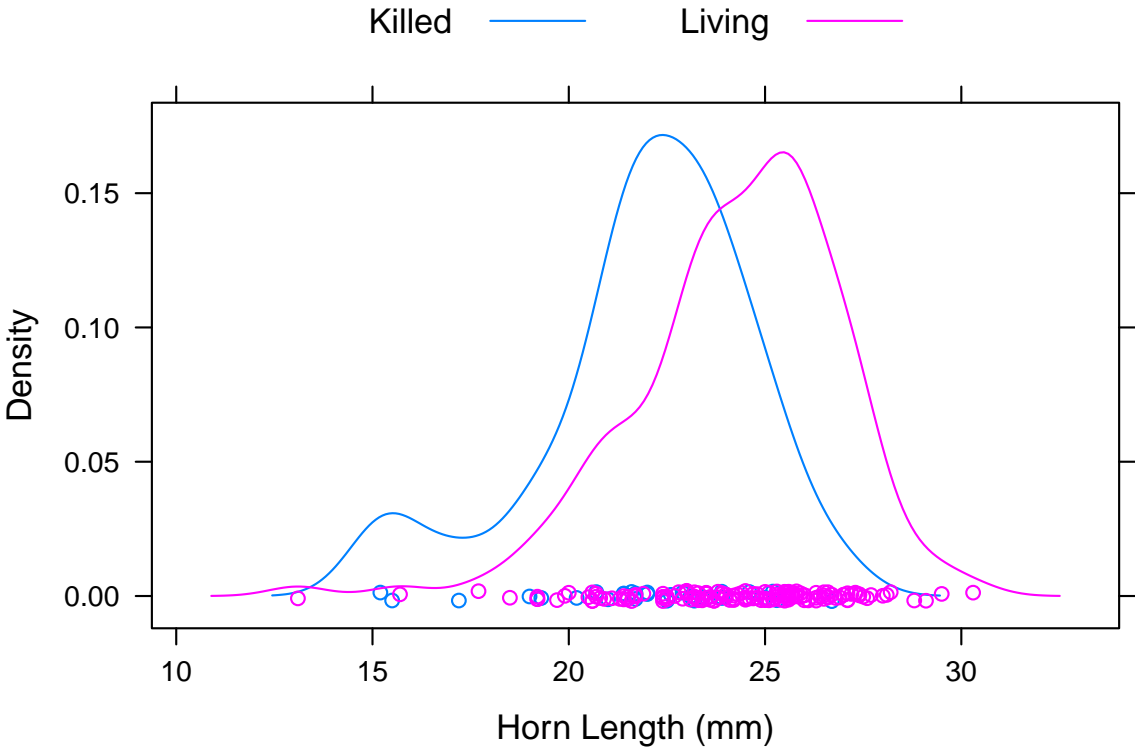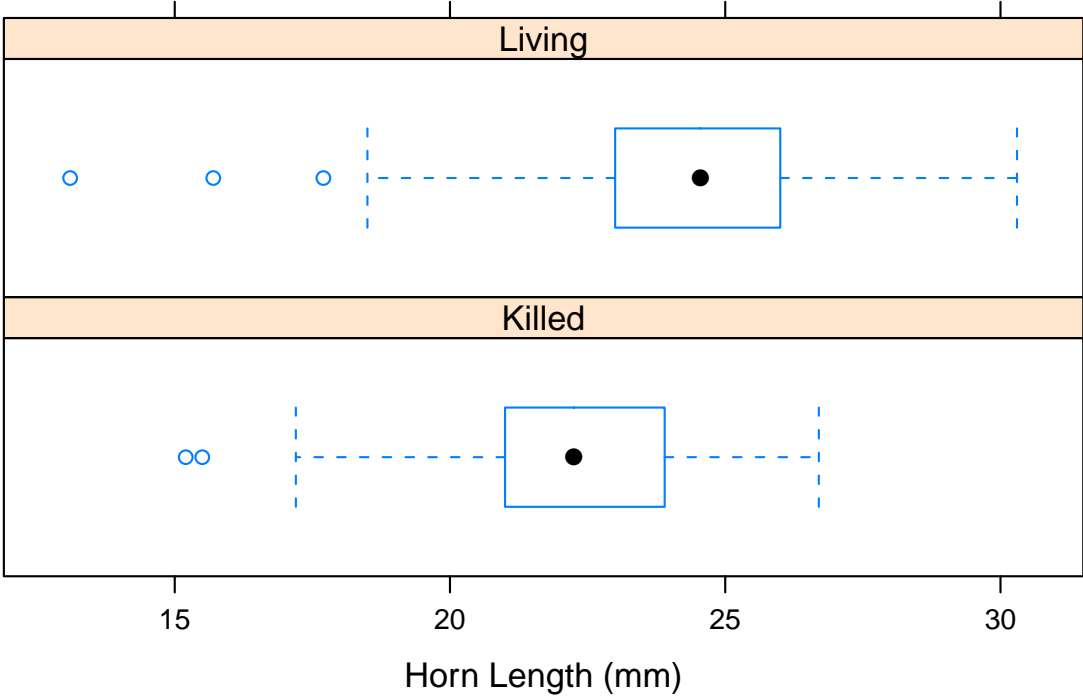- It is best to use common scales on axes.
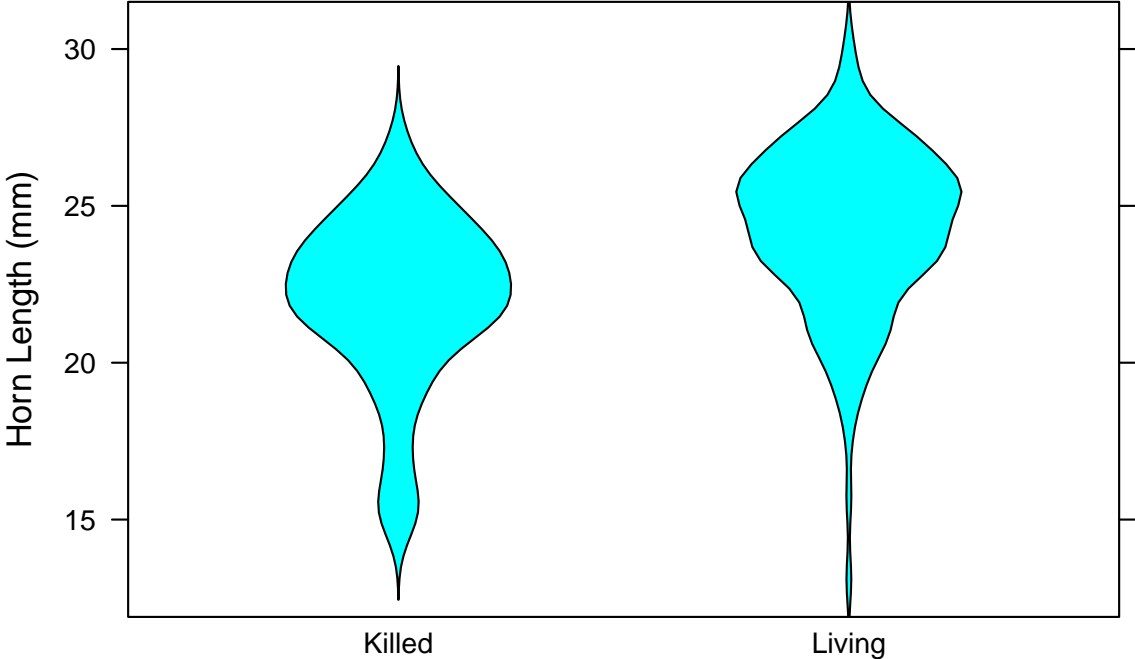
Histograms
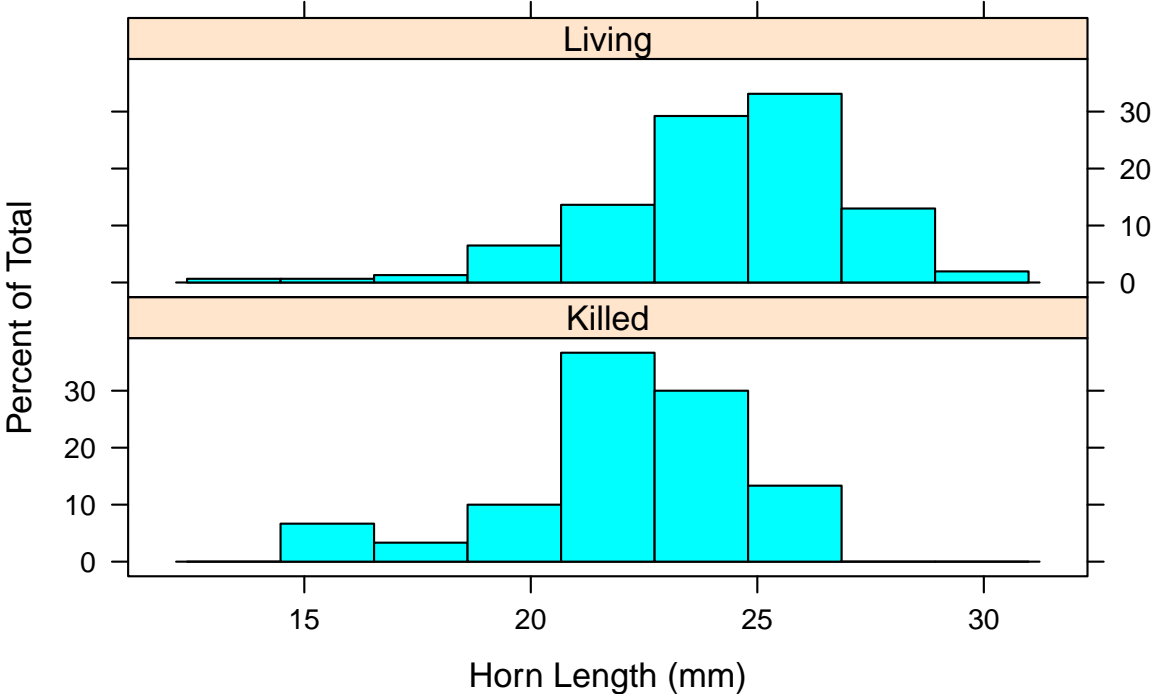
Dot Plots

Density Plots

Box-and-Whisker Plots

Violin Plots

Remarks

- While there are some extreme values, they do not stick out far from the overall pattern of the data.
- Both distributions look to be not strongly skewed.
- Inferences based on t distributions will be accurate, even with moderate nonnormality in the underlying populations.

t Distribution for Independent Samples

t Distribution for Independent Samples

If two independent samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are each normally distributed with respective means μ_1 and μ_2 and if they *share a common variance* σ^2, then the test statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

is the *pooled sample standard deviation* and s_1 and s_2 are the respective single sample standard deviations.

- Note that s_p^2, the pooled variance, is a weighted average of the sample variances, weighted by the degrees of freedom.

Derivation

- Using linearity properties of expectation and variances of independent random variables, it is straightforward to show that if $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are independent samples with $E(X_i) = \mu_1$, $\text{Var}(X_i) = \sigma_1^2$, $E(Y_i) = \mu_2$, and $\text{Var}(Y_i) = \sigma_2^2$, then

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$$
$$\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- Furthermore, if both distributions are normal, the distribution of the difference in sample means is also normal.

- Under the additional assumption that $\sigma_1 = \sigma_2$,

$$\text{Var}(\bar{X} - \bar{Y}) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Standard Error

- The standard error for the difference in sample means is

$$SE(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Under the assumption that $\sigma_1 = \sigma_2$, this is estimated as

$$\widehat{SE}(\bar{X} - \bar{Y}) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Without this assumption, the estimate is

$$\widetilde{SE}(\bar{X} - \bar{Y}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- *Welch's t-test* uses $\widetilde{SE}$ for the T statistic; in this case the distribution is only approximate and the degrees of freedom is approximated with a messier formula (see page 304 in the textbook for details).

- Welch's approach is the default in the R function `t.test()`.

Chalkboard example

Example

- Here are summary statistics for the lizard data:
  - ▷ The sample sizes are 154 for the living lizards and 30 for the unfortunate skewered ones.
  - ▷ The sample means are 24.28 for the living lizards and 21.99 for the killed ones.
  - ▷ The standard deviations are respectively 2.63 and 2.71.
  - ▷ Find a 95% confidence interval for the difference in mean horn length for the two groups.
  - ▷ Test the hypothesis test of no difference in mean horn size.

Comparison with R

- Use `t.test()` with argument `var.equal=T` for the conventional two-sample test.

```
> t.test(hornLiving, hornKilled, var.equal = T)

Two Sample t-test

data:  hornLiving and hornKilled
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:
 1.253602 3.335402
sample estimates:
mean of x mean of y
 24.28117  21.98667
```

Comparison with R (Welch's Test)

- Use `t.test()` with default arguments for Welch's two-sample test which does not assume equal variances.
- Note that the approximate degrees of freedom is always at least as big as the smaller of $n_1 - 1$ and $n_2 - 1$ and no larger than $n_1 + n_2 - 2$.

```
> t.test(hornLiving, hornKilled)

Welch Two Sample t-test

data:  hornLiving and hornKilled
t = 4.2634, df = 40.372, p-value = 0.0001178
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.207092 3.381912
sample estimates:
mean of x mean of y
 24.28117  21.98667
```

Interpretation of Confidence Interval

We are 95% confident that the mean length of the horns is between 1.21 and 3.38 mm longer in surviving lizards than in lizards killed by the loggerhead shrike in the study population where the data was collected. This is consistent with the biological hypothesis that larger horns offer more protection. There are other possible interpretations, as other variables may be confounded with horn length. Perhaps lizards with longer horns are also older, heavier, faster, or something else that is the real cause of the greater protection.

A one-sided *t*-test

```
> t.test(hornLiving, hornKilled, alternative = "greater")

Welch Two Sample t-test

data:  hornLiving and hornKilled
t = 4.2634, df = 40.372, p-value = 5.889e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.388469       Inf
sample estimates:
mean of x mean of y
 24.28117  21.98667
```

Interpretation of Hypothesis Test

There is strong evidence that the mean horn length in living lizards is greater than that in lizards killed by the loggerhead shrike ($p < 0.0001$, $t = 4.26$, Welch's two-sample independent t-test, $n_1 = 154$, $n_2 = 30$). This is consistent with the biological hypothesis that longer horns offer greater protection, but is likewise consistent with explanations of protection offered by other characteristics associated with lizards with longer horns.

Case Study

> **Example**
>
> Recall the pseudoscorpions example we used for the permutation test. In one group, a female was mated with the same male twice; in the other group, two different males were mated to the same female. The response variable is the number of successful broods. This variable is a small integer, ranging from 0 in some cases to a maximum of observed value of 7.

- For the 100,000 randomly selected randomizations, the p-value is estmated to be 0.0093.

Comparison to Welch's t test

- Compare the corresponding p-values.
- There is also evidence that the mean number of successful broods is smaller in the *Same* treatment group.
- The p-values differ.

```
> t.test(same, different, alternative = "less")

	Welch Two Sample t-test

data:  same and different
t = -2.3424, df = 28.883, p-value = 0.01313
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.3911856
sample estimates:
mean of x mean of y
    2.200     3.625
```

Cautions and Concerns

- Always plot data before doing inferences.

- Handle potential outliers with care.

- Use the paired or two-sample t methods when appropriate; paired when the design is a single sample with each unit measured twice, independent when the samples are independent of one another.

- t methods are robust to nonnormality when sample sizes are large enough.

- Inferences to populations depends on random sampling; be prepared to argue when a sample can be thought of as representative despite nonrandom sampling.

- The two-sample independent t methods assume equal variances in the samples, but are robust to minor differences.

- Randomization or permutation methods are alternatives to t tests.

What You Should Know

You should know:

- how to use randomization/permutation methods or t methods for paired and independent sample problems;

- how to determine of paired or independent sample methods are appropriate;

- how to graph data in various ways to display differences in distributions;

- how to examine graphs of data to informally assess method assumptions;

- how to interpret inferences in context.

Extensions

- We have moved from one to two populations: what about three or more?
- ANOVA (Analysis of Variance) is on the way.