SML 201 – Week 8

John D. Storey

Spring 2016

Contents

CLT Revisited

Standardized RVs

Note that in general for a rv Y it is the case that

$$\frac{Y - \mathrm{E}[Y]}{\sqrt{\mathrm{Var}(Y)}}$$

has population mean 0 and variance 1.

CLT for Standardized RVs

Suppose $X_1, X_2, \ldots, X_n$ are iid rv's with population mean $\mathrm{E}[X_i] = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$.

Then for "large n",

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

approximately follows the Normal$(0, 1)$ distribution.

As $n \to \infty$, this approximation becomes exact.

Example: Standardized Poisson

Let $X_1, X_2, \ldots, X_{40}$ be iid Poisson(λ) with $\lambda = 6$.

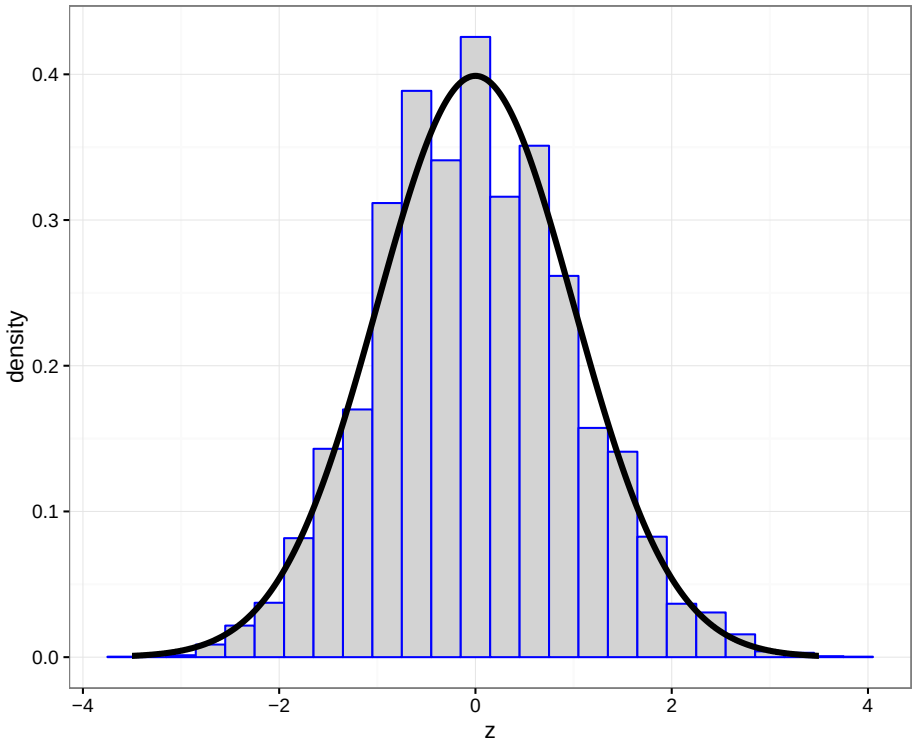
We will form

$$\frac{\overline{X} - 6}{\sqrt{6}/\sqrt{40}}$$

over 10,000 realizations and compare their distribution to a Normal(0, 1) distribution.

```
> x <- replicate(n=1e4, expr=rpois(n=40, lambda=6),
+                simplify="matrix")
> x_bar <- apply(x, 2, mean)
> clt_std <- (x_bar - 6)/(sqrt(6)/sqrt(40))
>
> df <- data.frame(z=clt_std, x = seq(-3.5,3.5,length.out=1e4),
+                  y = dnorm(seq(-3.5,3.5,length.out=1e4)))
> # note that df$y are Normal(0,1) pdf values
```

```
> ggplot(data=df) +
+   geom_histogram(aes(x=z, y=..density..), color="blue",
+                  fill="lightgray", binwidth=0.3) +
+   geom_line(aes(x=x, y=y), size=1.5)
```



Approximate Pivotal Statistics

Normal Distribution, Known Variance

Last week we considered data modeled by $X_1, X_2, \ldots, X_n$ iid Normal(μ, σ^2) where we assumed that σ^2 is known.

We derived $(1-\alpha)$-level confidence intervals and also hypothesis tests based on the pivotal statistic:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1).$$

Wider Application

As it turns out, we can use these results for a wider range of distributions. Those we earlier introduced have approximately pivotal Normal$(0, 1)$ statistics.

They have the form:

$$Z = \frac{\text{estimator} - \text{parameter}}{\text{standard error}} \sim \text{Normal}(0, 1),$$

where "standard error" is what we call an estimator of the standard deviation of the estimator.

Justification

The CLT from the previous section provides a justification for why these Z statistics are approximately Normal$(0, 1)$.

Some additional mathematics and assumptions must be detailed, but the basic justification is through the CLT.

Summary of Statistics

Distribution	Estimator	Std Err	Z Statistic
Binomial(n, p)	$\hat{p} = X/n$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$
Normal(μ, σ^2)	$\hat{\mu} = \overline{X}$	$\frac{S}{\sqrt{n}}$	$\frac{\hat{\mu}-\mu}{S/\sqrt{n}}$
Poisson(λ)	$\hat{\lambda} = \overline{X}$	$\sqrt{\frac{\hat{\lambda}}{n}}$	$\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}/n}}$

In all of these scenarios, Z is approximately Normal$(0, 1)$ for large n.

Notes

- For the Normal and Poisson distributions, our model is $X_1, X_2, \ldots, X_n$ iid from each respective distribution
- For the Binomial distribution, our model is $X \sim \text{Binomial}(n, p)$
- In the Normal model, $S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$ is the sample standard deviation

- The above formulas were given in terms of the random variable probability models; on observed data the same formulas are used except we observed data lower case letters, e.g., replace $\overline{X}$ with $\overline{x}$

Binomial

Approximate $(1-\alpha)$-level two-sided CI:

$$\left(\hat{p} - |z_{\alpha/2}|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + |z_{\alpha/2}|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Hypothesis test, $H_0 : p = p_0$ vs $H_1 : p \neq p_0$:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \text{ and p-value} = \Pr(|Z^*| \geq |z|)$$

where Z^* is a Normal$(0, 1)$ random variable.

Normal

Approximate $(1-\alpha)$-level two-sided CI:

$$\left(\hat{\mu} - |z_{\alpha/2}|\frac{s}{\sqrt{n}}, \hat{\mu} + |z_{\alpha/2}|\frac{s}{\sqrt{n}}\right)$$

Hypothesis test, $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$:

$$z = \frac{\hat{\mu} - \mu_0}{s/\sqrt{n}} \text{ and p-value} = \Pr(|Z^*| \geq |z|)$$

where Z^* is a Normal$(0, 1)$ random variable.

Poisson

Approximate $(1-\alpha)$-level two-sided CI:

$$\left(\hat{\lambda} - |z_{\alpha/2}|\sqrt{\frac{\hat{\lambda}}{n}}, \hat{\lambda} + |z_{\alpha/2}|\sqrt{\frac{\hat{\lambda}}{n}}\right)$$

Hypothesis test, $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda \neq \lambda_0$:

$$z = \frac{\hat{\lambda} - \lambda_0}{\sqrt{\frac{\hat{\lambda}}{n}}} \text{ and p-value} = \Pr(|Z^*| \geq |z|)$$

where Z^* is a Normal$(0, 1)$ random variable.

One-Sided CIs and HTs

The one-sided versions of these approximate confidence intervals and hypothesis tests work analogously.

The procedures shown for the Normal(μ, σ^2) case with known σ^2 from last week are utilzied with the appropriate subsitutions as in the above examples.

Comment

This gives you a framework to do many common inference tasks "by hand" (i.e., calculating each component directly in R).

However, R uses a much more comprehensive set of theory, methods, and computational approximations.

Therefore, this "large n, z-statistic" framework serves as a guide so that you know approximately what R does, but we will learn specific functions that are tailored for each data type.

Two-Sample Inference

Comparing Two Populations

So far we have concentrated on analyzing n observations from a single population.

However, suppose that we want to do inference to compare two populations?

The framework we have described so far is easily extended to accommodate this.

Two RVs

If X and Y are independent rv's then:

$$\mathrm{E}[X - Y] = \mathrm{E}[X] - \mathrm{E}[Y]$$

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

Two Sample Means

Let $X_1, X_2, \ldots, X_{n_1}$ be iid rv's with population mean μ_1 and population variance σ_1^2.

Let $Y_1, Y_2, \ldots, Y_{n_2}$ be iid rv's with population mean μ_2 and population variance σ_2^2.

Assume that the two sets of rv's are independent. Then when the CLT applies to each set of rv's, it approximately holds that:

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \text{Normal}(0, 1)$$

Same Rationale

Just as we formed Z-statistics earlier of the form

$$Z = \frac{\text{estimator} - \text{parameter}}{\text{standard error}} \dot{\sim} \text{Normal}(0, 1),$$

we can do the analogous thing in the two-sample case, except now we're considering differences.

Poisson

Let $X_1, X_2, \ldots, X_{n_1}$ be iid Poisson(λ_1) and $Y_1, Y_2, \ldots, Y_{n_2}$ be iid Poisson(λ_2).

We have $\hat{\lambda}_1 = \overline{X}$ and $\hat{\lambda}_2 = \overline{Y}$. For large n_1 and n_2, it approximately holds that:

$$\frac{\hat{\lambda}_1 - \hat{\lambda}_2 - (\lambda_1 - \lambda_2)}{\sqrt{\frac{\hat{\lambda}_1}{n_1} + \frac{\hat{\lambda}_2}{n_2}}} \sim \text{Normal}(0, 1).$$

Normal (Unequal Variances)

Let $X_1, X_2, \ldots, X_{n_1}$ be iid Normal(μ_1, σ_1^2) and $Y_1, Y_2, \ldots, Y_{n_2}$ be iid Normal(μ_2, σ_2^2).

We have $\hat{\mu}_1 = \overline{X}$ and $\hat{\mu}_2 = \overline{Y}$. For large n_1 and n_2, it approximately holds that:

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \text{Normal}(0, 1).$$

Normal (Equal Variances)

Let $X_1, X_2, \ldots, X_{n_1}$ be iid Normal(μ_1, σ^2) and $Y_1, Y_2, \ldots, Y_{n_2}$ be iid Normal(μ_2, σ^2).

We have $\hat{\mu}_1 = \overline{X}$ and $\hat{\mu}_2 = \overline{Y}$. For large n_1 and n_2, it approximately holds that:

$$\frac{\hat{\mu}_1 - \hat{\mu}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}} \sim \text{Normal}(0, 1)$$

where

$$S^2 = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}{n_1 + n_2 - 2}$$

Binomial

Let $X \sim \text{Binomial}(n_1, p_1)$ and $Y \sim \text{Binomial}(n_2, p_2)$.

We have $\hat{p}_1 = X/n_1$ and $\hat{p}_2 = Y/n_2$. For large n_1 and n_2, it approximately holds that:

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim \text{Normal}(0, 1).$$

Example: Binomial CI

A 95% CI for the difference $p_1 - p_2$ can be obtained by unfolding the above pivotal statistic:

$$\left((\hat{p}_1 - \hat{p}_2) - 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\ , \right.$$

$$\left. (\hat{p}_1 - \hat{p}_2) + 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right)$$

Example: Binomial HT

Suppose we wish to test $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$.

First form the z-statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

Now, calculate the p-value:

$$\Pr(|Z^*| \geq |z|)$$

where Z^* is a Normal(0,1) random variable.

Z Statistic Inference in R

`BSDA` Package

```
> install.packages("BSDA")
```

```
> library(BSDA)
> str(z.test)
function (x, y = NULL, alternative = "two.sided", mu = 0,
    sigma.x = NULL, sigma.y = NULL, conf.level = 0.95)
```

Example: Poisson

Apply `z.test()`:

```
> set.seed(210)
```

```
> n <- 40
> lam <- 14
> x <- rpois(n=n, lambda=lam)
> lam.hat <- mean(x)
> stddev <- sqrt(lam.hat)
> z.test(x=x, sigma.x=stddev, mu=lam)

    One-sample z-Test
```

```
data:  x
z = 0.41885, p-value = 0.6753
alternative hypothesis: true mean is not equal to 14
95 percent confidence interval:
 13.08016 15.41984
sample estimates:
mean of x
    14.25
```

By Hand Calculations

Confidence interval:

```
> lam.hat <- mean(x)
> lam.hat
[1] 14.25
> stderr <- sqrt(lam.hat)/sqrt(n)
> lam.hat - abs(qnorm(0.025)) * stderr # lower bound
[1] 13.08016
> lam.hat + abs(qnorm(0.025)) * stderr # upper bound
[1] 15.41984
```

Hypothesis test:

```
> z <- (lam.hat - lam)/stderr
> z # test statistic
[1] 0.4188539
> 2 * pnorm(-abs(z)) # two-sided p-value
[1] 0.6753229
```

Exercise

Figure out how to get the `z.test()` function to work on Binomial data.

Hint: Are n iid observations from the Binomial$(1, p)$ distribution equivalent to one observation from the Binomial(n, p)?

The t Distribution

Normal Distribution, Unknown Variance

Suppose data a sample of n data points is modeled by $X_1, X_2, \ldots, X_n$ iid Normal(μ, σ^2) where σ^2 is *unknown*.

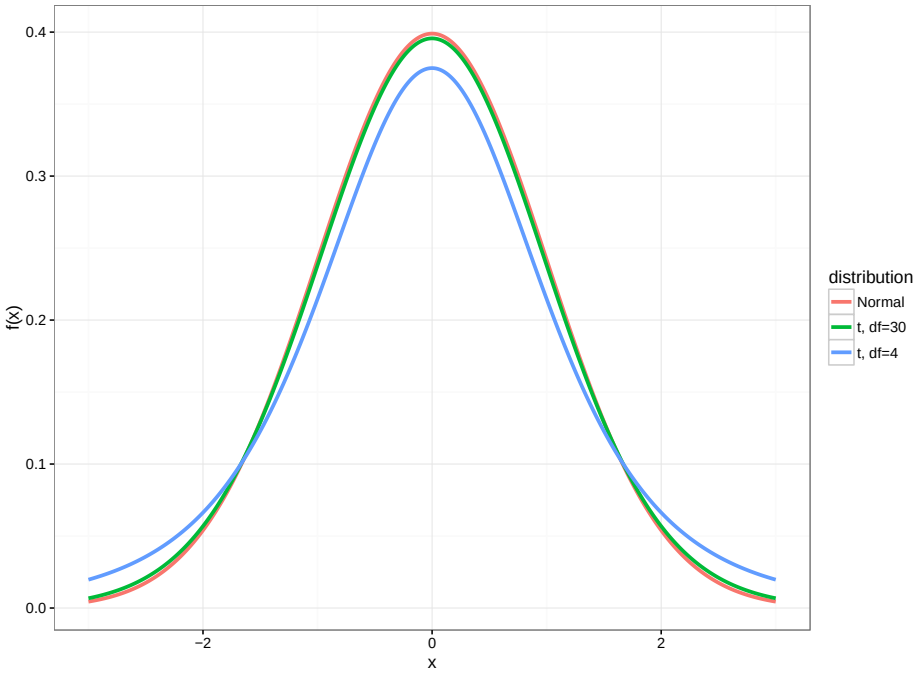
We still have a pivotal statistic. Recall that $S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$ is the sample standard deviation.

The statistic

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a t_{n-1} distribution, a t-distribution with $n-1$ degrees of freedom.

t vs Normal



t Percentiles

We calculated percentiles of the Normal(0,1) distribution (e.g., z_α). We can do the analogous calculation with the t distribution.

Let t_α be the α percentile of the t distribution. Examples:

```
> qt(0.025, df=4) # alpha = 0.025
[1] -2.776445
> qt(0.05, df=4)
[1] -2.131847
```

```
> qt(0.95, df=4)
[1] 2.131847
> qt(0.975, df=4)
[1] 2.776445
```

Confidence Intervals

Here is a $(1-\alpha)$-level CI for μ using this distribution:

$$\left(\hat{\mu} - |t_{\alpha/2}|\frac{s}{\sqrt{n}}, \hat{\mu} + |t_{\alpha/2}|\frac{s}{\sqrt{n}}\right),$$

where as before $\hat{\mu} = \overline{x}$. This produces a wider CI than the z statistic analogue.

Hypothesis Tests

Suppose we want to test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ where μ_0 is a known, given number.

The t-statistic is

$$t = \frac{\hat{\mu} - \mu_0}{\frac{s}{\sqrt{n}}}$$

with p-value

$$\Pr(|T^*| \geq |t|)$$

where $T^* \sim t_{n-1}$.

Two-Sample Inference

In the **Two-Sample Inference** section we presented pivotal statistics for the two-sample case with unequal and equal variances.

When there are equal variances, the pivotal statistic follows a $t_{n_1+n_2-2}$ distribution.

When there are unequal variances, the pivotal statistic follows a t distribution where the degrees of freedom comes from a more complex formula, which R calculates for us.

When Is t Utilized?

- The t distribution and its corresponding CI's and HT's are utilized when the data are Normal (or approximately Normal) and n is small
- Small typically means that $n < 30$
- In this case the inference based on the t distribution will be more accurate
- When $n \geq 30$, there is very little difference between using t-statistics and z-statistics

Inference in R

Functions in R

R has the following functions for doing inference on the distributions we've considered.

- Normal: `t.test()`
- Binomial: `binomial.test()` or `prop.test()`
- Poisson: `poisson.test()`

These perform one-sample and two-sample hypothesis testing and confidence interval construction for both the one-sided and two-sided cases.

About These Functions

- We covered a convenient, unified framework that allows us to better understand how confidence intervals and hypothesis testing are performed
- However, this framework requires large sample sizes and is not necessarily the best method to apply in all circumstances

About These Functions (cont'd)

- The above R functions are versatile functions for analyzing Normal, Binomial, and Poisson distributed data (or approximations thereof) that use much broader theory and methods than we will cover in this course
- The arguments these functions take and the ouput of the functions are in line with the framework that we have covered

Inference on Normal Data in R

Setup

```
> library("dplyr")
> library("ggplot2")
> theme_set(theme_bw())
> library("broom")
```

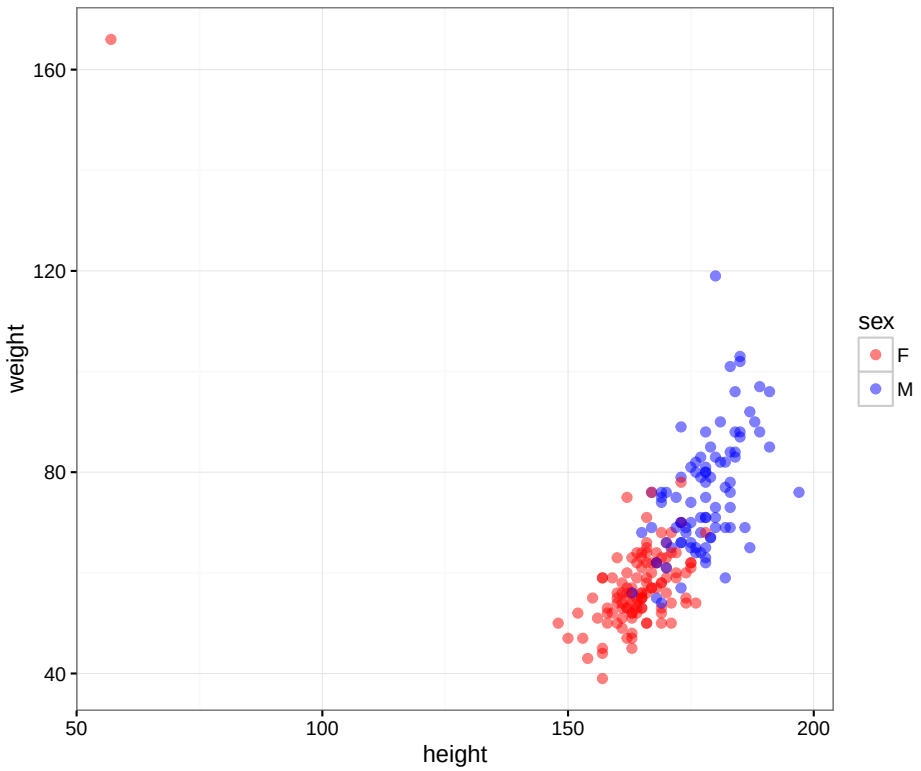
"Davis" Data Set

```
> library("car")

Attaching package: 'car'
The following objects are masked from 'package:BSDA':

    Vocab, Wool
> data("Davis")
```

```
> htwt <- tbl_df(Davis)
> htwt
Source: local data frame [200 x 5]

      sex weight height repwt repht
   (fctr)  (int)  (int) (int) (int)
1       M     77    182    77   180
2       F     58    161    51   159
3       F     53    161    54   158
4       M     68    177    70   175
5       F     59    157    59   155
6       M     76    170    76   165
7       M     76    167    77   165
8       M     69    186    73   180
9       M     71    178    71   175
10      M     65    171    64   170
..    ...    ...    ...   ...   ...
```

Height vs Weight

```
> ggplot(htwt) +
+ geom_point(aes(x=height, y=weight, color=sex), size=2, alpha=0.5) +
+ scale_colour_manual(values=c("red", "blue"))
```



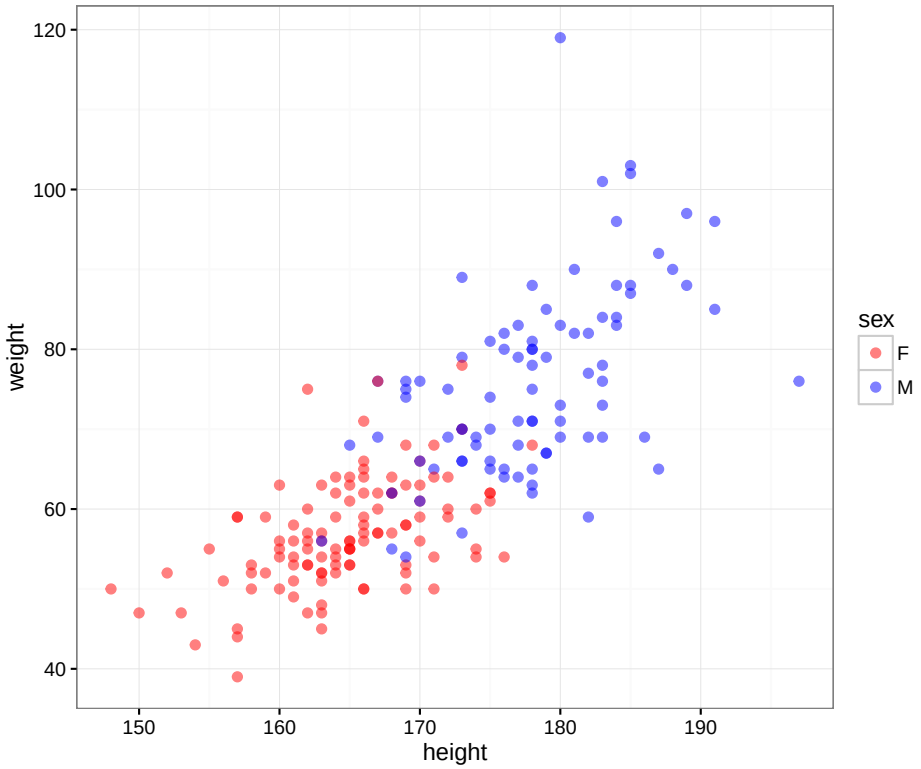
An Error?

```
> which(htwt$height < 100)
[1] 12
> htwt[12,]
Source: local data frame [1 x 5]

     sex weight height repwt repht
  (fctr)  (int)  (int) (int) (int)
1      F    166     57    56   163
```

```
> htwt[12,c(2,3)] <- htwt[12,c(3,2)]
```

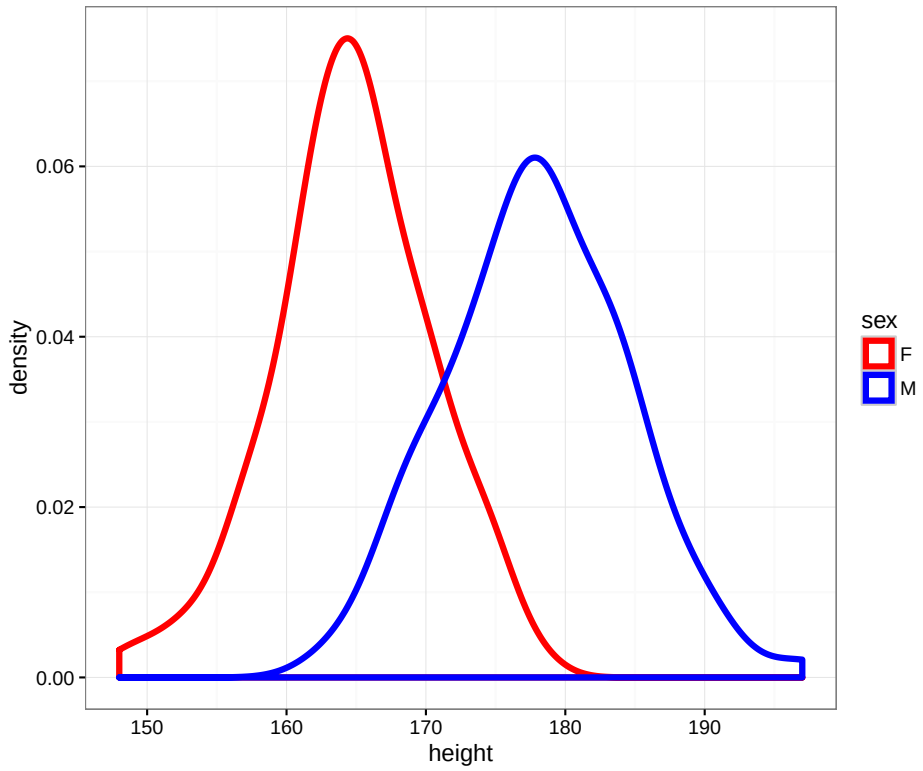
Updated Height vs Weight

```
> ggplot(htwt) +
+   geom_point(aes(x=height, y=weight, color=sex), size=2, alpha=0.5) +
+   scale_color_manual(values=c("red", "blue"))
```



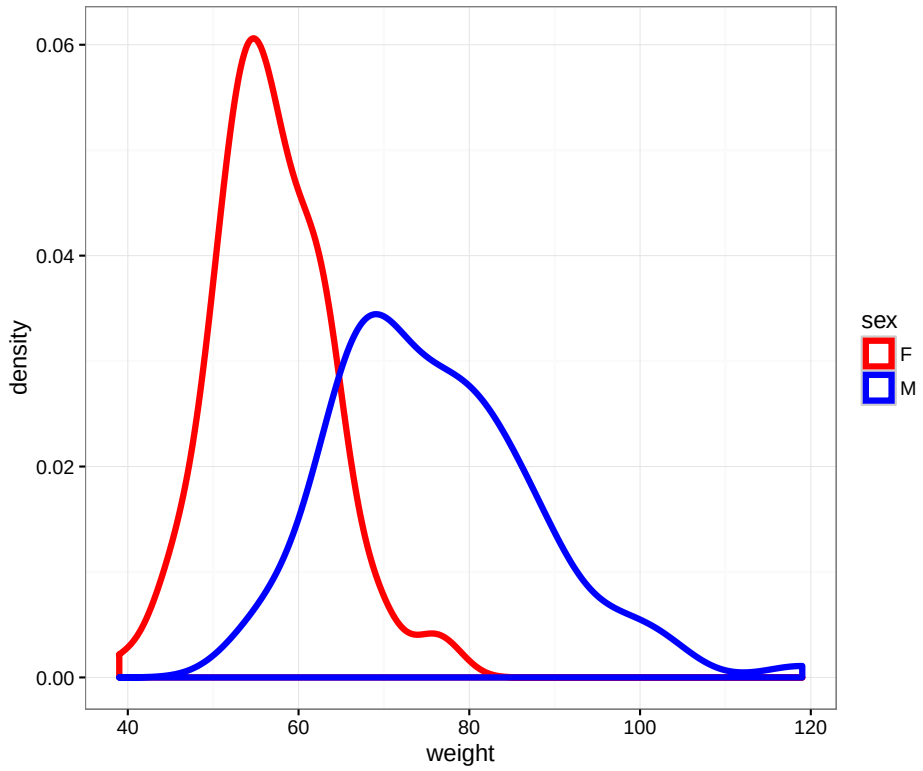
Density Plots of Height

```
> ggplot(htwt) +
+   geom_density(aes(x=height, color=sex), size=1.5) +
+   scale_color_manual(values=c("red", "blue"))
```



Density Plots of Weight

```
> ggplot(htwt) +
+   geom_density(aes(x=weight, color=sex), size=1.5) +
+   scale_color_manual(values=c("red", "blue"))
```



`t.test()` Function

From the help file...

```
Usage

t.test(x, ...)

## Default S3 method:
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)

## S3 method for class 'formula'
t.test(formula, data, subset, na.action, ...)
```

Two-Sided Test of Male Height

```
> m_ht <- htwt %>% filter(sex=="M") %>% select(height)
> testresult <- t.test(x = m_ht$height, mu=177)
```

```
> class(testresult)
[1] "htest"
> is.list(testresult)
[1] TRUE
```

Output of `t.test()`

```
> names(testresult)
[1] "statistic"   "parameter"   "p.value"     "conf.int"
[5] "estimate"    "null.value"  "alternative" "method"
[9] "data.name"
> testresult

    One Sample t-test

data:  m_ht$height
t = 1.473, df = 87, p-value = 0.1443
alternative hypothesis: true mean is not equal to 177
95 percent confidence interval:
 176.6467 179.3760
sample estimates:
mean of x
 178.0114
```

Tidying the Output

```
> library(broom)
> tidy(testresult)
  estimate statistic   p.value parameter conf.low conf.high
1 178.0114  1.473043 0.1443482        87 176.6467   179.376
```

Two-Sided Test of Female Height

```
> f_ht <- htwt %>% filter(sex=="F") %>% select(height)
> t.test(x = f_ht$height, mu = 164)

    One Sample t-test

data:  f_ht$height
t = 1.3358, df = 111, p-value = 0.1844
alternative hypothesis: true mean is not equal to 164
95 percent confidence interval:
 163.6547 165.7739
sample estimates:
mean of x
 164.7143
```

Difference of Two Means

```
> t.test(x = m_ht$height, y = f_ht$height)

    Welch Two Sample t-test

data:  m_ht$height and f_ht$height
t = 15.28, df = 174.29, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 11.57949 15.01467
sample estimates:
mean of x mean of y
 178.0114  164.7143
```

Test with Equal Variances

```
> htwt %>% group_by(sex) %>% summarize(sd(height))
Source: local data frame [2 x 2]

     sex sd(height)
  (fctr)      (dbl)
1      F   5.659129
2      M   6.440701
> t.test(x = m_ht$height, y = f_ht$height, var.equal = TRUE)

    Two Sample t-test
```

```
data:  m_ht$height and f_ht$height
t = 15.519, df = 198, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 11.60735 14.98680
sample estimates:
mean of x mean of y
 178.0114  164.7143
```

Paired Sample Test (v. 1)

First take the difference between the paired observations. Then apply the one-sample t-test.

```
> htwt <- htwt %>% mutate(diffwt = (weight - repwt),
+                         diffht = (height - repht))
> t.test(x = htwt$diffwt) %>% tidy()
     estimate statistic   p.value parameter   conf.low
1 0.005464481 0.0319381 0.9745564       182 -0.3321223
  conf.high
1 0.3430513
> t.test(x = htwt$diffht) %>% tidy()
  estimate statistic      p.value parameter conf.low conf.high
1 2.076503  13.52629 2.636736e-29       182 1.773603  2.379403
```

Paired Sample Test (v. 2)

Enter each sample into the `t.test()` function, but use the `paired=TRUE` argument. This is operationally equivalent to the previous version.

```
> t.test(x=htwt$weight, y=htwt$repwt, paired=TRUE) %>% tidy()
     estimate statistic   p.value parameter   conf.low
1 0.005464481 0.0319381 0.9745564       182 -0.3321223
  conf.high
1 0.3430513
> t.test(x=htwt$height, y=htwt$repht, paired=TRUE) %>% tidy()
  estimate statistic      p.value parameter conf.low conf.high
1 2.076503  13.52629 2.636736e-29       182 1.773603  2.379403
> htwt %>% select(height, repht) %>% na.omit() %>%
+   summarize(mean(height), mean(repht))
Source: local data frame [1 x 2]

  mean(height) mean(repht)
```

```
         (dbl)       (dbl)
1     170.5738    168.4973
```

Inference on Binomial Data in R

The Coin Flip Example

I flip it 20 times and it lands on heads 16 times.

1. My data is $x = 16$ heads out of $n = 20$ flips.
2. My data generation model is $X \sim \text{Binomial}(20, p)$.
3. I form the statistic $\hat{p} = 16/20$ as an estimate of p.

Let's do hypothesis testing and confidence interval construction on these data.

`binom.test()`

```
> str(binom.test)
function (x, n, p = 0.5, alternative = c("two.sided",
    "less", "greater"), conf.level = 0.95)
> binom.test(x=16, n=20, p = 0.5)

    Exact binomial test

data:  16 and 20
number of successes = 16, number of trials = 20,
p-value = 0.01182
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.563386 0.942666
sample estimates:
probability of success
                   0.8
```

`alternative = "greater"`

Tests $H_0 : p \leq 0.5$ vs. $H_1 : p > 0.5$.

```
> binom.test(x=16, n=20, p = 0.5, alternative="greater")

    Exact binomial test

data:  16 and 20
number of successes = 16, number of trials = 20,
p-value = 0.005909
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.5989719 1.0000000
sample estimates:
probability of success
                   0.8
```

`alternative = "less"`

Tests $H_0 : p \geq 0.5$ vs. $H_1 : p < 0.5$.

```
> binom.test(x=16, n=20, p = 0.5, alternative="less")

    Exact binomial test

data:  16 and 20
number of successes = 16, number of trials = 20,
p-value = 0.9987
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.9286461
sample estimates:
probability of success
                   0.8
```

`prop.test()`

This is a "large n" inference method that is very similar to our z-statistic approach.

```
> str(prop.test)
function (x, n, p = NULL, alternative = c("two.sided",
    "less", "greater"), conf.level = 0.95, correct = TRUE)
> prop.test(x=16, n=20, p=0.5)

    1-sample proportions test with continuity correction
```

```
data:  16 out of 20, null probability 0.5
X-squared = 6.05, df = 1, p-value = 0.01391
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5573138 0.9338938
sample estimates:
  p
0.8
```

An Observation

```
> p <- binom.test(x=16, n=20, p = 0.5)$p.value
> binom.test(x=16, n=20, p = 0.5, conf.level=(1-p))

    Exact binomial test

data:  16 and 20
number of successes = 16, number of trials = 20,
p-value = 0.01182
alternative hypothesis: true probability of success is not equal to 0.5
98.81821 percent confidence interval:
 0.5000000 0.9625097
sample estimates:
probability of success
                   0.8
```

Exercise: Figure out what happened here.

OIS Exercise 6.10

The way a question is phrased can influence a person's response. For example, Pew Research Center conducted a survey with the following question:

"As you may know, by 2014 nearly all Americans will be required to have health insurance. [People who do not buy insurance will pay a penalty] while [People who cannot afford it will receive financial help from the government]. Do you approve or disapprove of this policy?"

For each randomly sampled respondent, the statements in brackets were randomized: either they were kept in the order given above, or the two statements were reversed.

The Data

Table 6.2 shows the results of this experiment, reproduced below.

2nd Statement	Sample Size	Approve Law	Disapprove Law	Other
"people who cannot afford it will receive financial help from the government"	771	47	49	3
"people who do not buy it will pay a penalty"	732	34	63	3

Inference on the Difference

Create and interpret a 90% confidence interval of the difference in approval. Also perform a hyppthesis test that the approval rates are equal.

```
> x <- round(c(0.47*771, 0.34*732))
> n <- round(c(771*0.97, 732*0.97))
> prop.test(x=x, n=n, conf.level=0.90)

    2-sample test for equality of proportions with
    continuity correction

data:  x out of n
X-squared = 26.023, df = 1, p-value = 3.374e-07
alternative hypothesis: two.sided
90 percent confidence interval:
 0.08979649 0.17670950
sample estimates:
   prop 1    prop 2
0.4839572 0.3507042
```

OIS 90% CI

The book *OIS* does a "by hand" calculation using the z-statistics and comes up with a similar answer (but not identical).

```
> p1.hat <- 0.47
> n1 <- 771
> p2.hat <- 0.34
> n2 <- 732
> stderr <- sqrt(p1.hat*(1-p1.hat)/n1 + p2.hat*(1-p2.hat)/n2)
>
> # the 90% CI
> (p1.hat - p2.hat) + c(-1,1)*abs(qnorm(0.05))*stderr
[1] 0.08872616 0.17127384
```

Inference on Poisson Data in R

poisson.test()

```
> str(poisson.test)
function (x, T = 1, r = 1, alternative = c("two.sided",
    "less", "greater"), conf.level = 0.95)
```

From the help:

```
Arguments

x    number of events. A vector of length one or two.

T    time base for event count. A vector of length one or two.

r    hypothesized rate or rate ratio

alternative  indicates the alternative hypothesis and must be one of
"two.sided", "greater" or "less". You can specify just the initial letter.

conf.level  confidence level for the returned confidence interval.
```

Example: RNA-Seq

RNA-Seq gene expression was measured for p53 lung tissue in 12 healthy individuals and 14 individuals with lung cancer.

The counts were given as follows.

Healthy: 82 64 66 88 65 81 85 87 60 79 80 72

Cancer: 59 50 60 60 78 69 70 67 72 66 66 68 54 62

It is hypothesized that p53 expression is higher in healthy individuals. Test this hypothesis, and form a 99% CI.

$H_1 : \lambda_1 \neq \lambda_2$

```
> healthy <- c(82, 64, 66, 88, 65, 81, 85, 87, 60, 79, 80, 72)
> cancer <- c(59, 50, 60, 60, 78, 69, 70, 67, 72, 66, 66, 68,
+             54, 62)
```

```
> poisson.test(x=c(sum(healthy), sum(cancer)), T=c(12,14),
+              conf.level=0.99)

    Comparison of Poisson rates

data:  c(sum(healthy), sum(cancer)) time base: c(12, 14)
count1 = 909, expected count1 = 835.38, p-value =
0.0005739
alternative hypothesis: true rate ratio is not equal to 1
99 percent confidence interval:
 1.041626 1.330051
sample estimates:
rate ratio
  1.177026
```

$H_1 : \lambda_1 < \lambda_2$

```
> poisson.test(x=c(sum(healthy), sum(cancer)), T=c(12,14),
+              alternative="less", conf.level=0.99)

    Comparison of Poisson rates

data:  c(sum(healthy), sum(cancer)) time base: c(12, 14)
count1 = 909, expected count1 = 835.38, p-value =
0.9998
alternative hypothesis: true rate ratio is less than 1
99 percent confidence interval:
 0.000000 1.314529
sample estimates:
rate ratio
  1.177026
```

$H_1 : \lambda_1 > \lambda_2$

```
> poisson.test(x=c(sum(healthy), sum(cancer)), T=c(12,14),
+              alternative="greater", conf.level=0.99)

    Comparison of Poisson rates

data:  c(sum(healthy), sum(cancer)) time base: c(12, 14)
count1 = 909, expected count1 = 835.38, p-value =
0.0002881
alternative hypothesis: true rate ratio is greater than 1
99 percent confidence interval:
 1.053921      Inf
sample estimates:
rate ratio
  1.177026
```

Question

Which analysis is the more informative and scientifically correct one, and why?

Extras

License

https://github.com/SML201/lectures/blob/master/LICENSE.md

Source Code

https://github.com/SML201/lectures/tree/master/week8

Session Information

```
> sessionInfo()
R version 3.2.3 (2015-12-10)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X 10.11.3 (El Capitan)

locale:
```

```
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

other attached packages:
 [1] car_2.1-1       broom_0.4.0     dplyr_0.4.3
 [4] BSDA_1.01       lattice_0.20-33 e1071_1.6-7
 [7] ggplot2_2.1.0   knitr_1.12.3    magrittr_1.5
[10] devtools_1.10.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.3         nloptr_1.0.4        formatR_1.2.1
 [4] plyr_1.8.3          highr_0.5.1         class_7.3-14
 [7] tools_3.2.3         lme4_1.1-11         digest_0.6.9
[10] evaluate_0.8        memoise_1.0.0       gtable_0.2.0
[13] nlme_3.1-125        mgcv_1.8-11         Matrix_1.2-3
[16] psych_1.5.8         DBI_0.3.1           yaml_2.1.13
[19] parallel_3.2.3      SparseM_1.7         stringr_1.0.0
[22] MatrixModels_0.4-1 grid_3.2.3          nnet_7.3-12
[25] R6_2.1.2            rmarkdown_0.9.5     minqa_1.2.4
[28] reshape2_1.4.1      tidyr_0.4.1         splines_3.2.3
[31] scales_0.4.0        codetools_0.2-14    htmltools_0.3
[34] MASS_7.3-45         assertthat_0.1      pbkrtest_0.4-6
[37] mnormt_1.5-3        colorspace_1.2-6    quantreg_5.21
[40] labeling_0.3        stringi_1.0-1       lazyeval_0.1.10
[43] munsell_0.4.3
```