SML 201 – Week 9

John D. Storey

Spring 2016

Contents

Inference on Binomial Data in R

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?

Modeling Relationships Among Variables

Rationale

One of the most important goals when analyzing data is to understand how variables relate to one another. This may include:

- Characterizing how variables covary
- Measuring and identifying associations between variables
- Explaining the variation of one variable in terms of others
- Predicting the outcome of a variable in terms of others

Strategies

We will consider both categorical and quantitative variables to achieve these goals. Over the next few weeks we will study:

- Analyzing two categorical variables
- Analyzing two quantitative variables
- Least squares linear regression to characterize variation of a quantitative variable in terms of other variables
- Logistic regression to characterize the probability distribution of a dichotomous variable in terms of other variables
- Predicting future values of a given variable based on measured values of other variables

Two Categorical Variables

Survey Data

```
> library("MASS")
> data("survey", package="MASS")
> survey <- tbl_df(survey)
> head(survey)
Source: local data frame [6 x 12]

     Sex Wr.Hnd NW.Hnd  W.Hnd    Fold Pulse    Clap   Exer
  (fctr)  (dbl)  (dbl) (fctr)  (fctr) (int)  (fctr) (fctr)
1 Female   18.5   18.0  Right  R on L    92    Left   Some
2   Male   19.5   20.5   Left  R on L   104    Left   None
3   Male   18.0   13.3  Right  L on R    87 Neither   None
```

```
4   Male   18.8   18.9  Right  R on L    NA Neither   None
5   Male   20.0   20.0  Right Neither    35   Right   Some
6 Female   18.0   17.7  Right  L on R    64   Right   Some
Variables not shown: Smoke (fctr), Height (dbl), M.I (fctr),
  Age (dbl)
```

2 x 2 Table

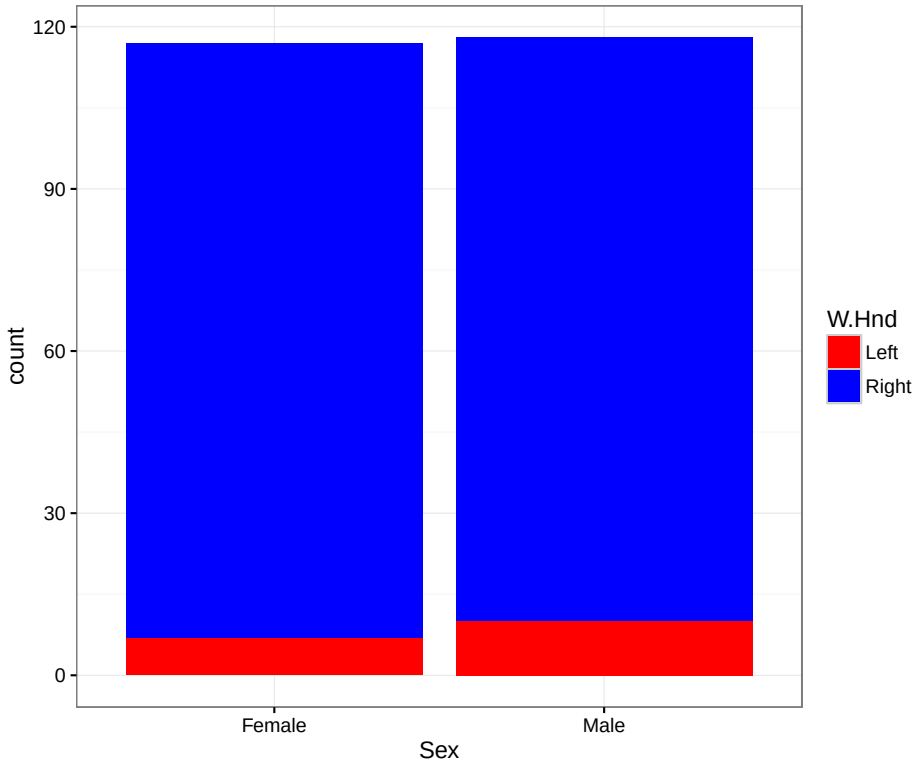
A contingency table:

```
> tbl = table(survey$Sex, survey$W.Hnd)
> tbl

        Left Right
  Female    7   110
  Male     10   108
```

Let's test the null hypothesis that sex and writing hand are independent vs. the alternative hypothesis that they are dependent.

Visualization



Pearson's Chi-Squared Test

```
> str(chisq.test)
function (x, y = NULL, correct = TRUE, p = rep(1/length(x),
    length(x)), rescale.p = FALSE, simulate.p.value = FALSE,
    B = 2000)
```

```
> chisq.test(tbl)

    Pearson's Chi-squared test with Yates' continuity
    correction

data:  tbl
X-squared = 0.23563, df = 1, p-value = 0.6274
```

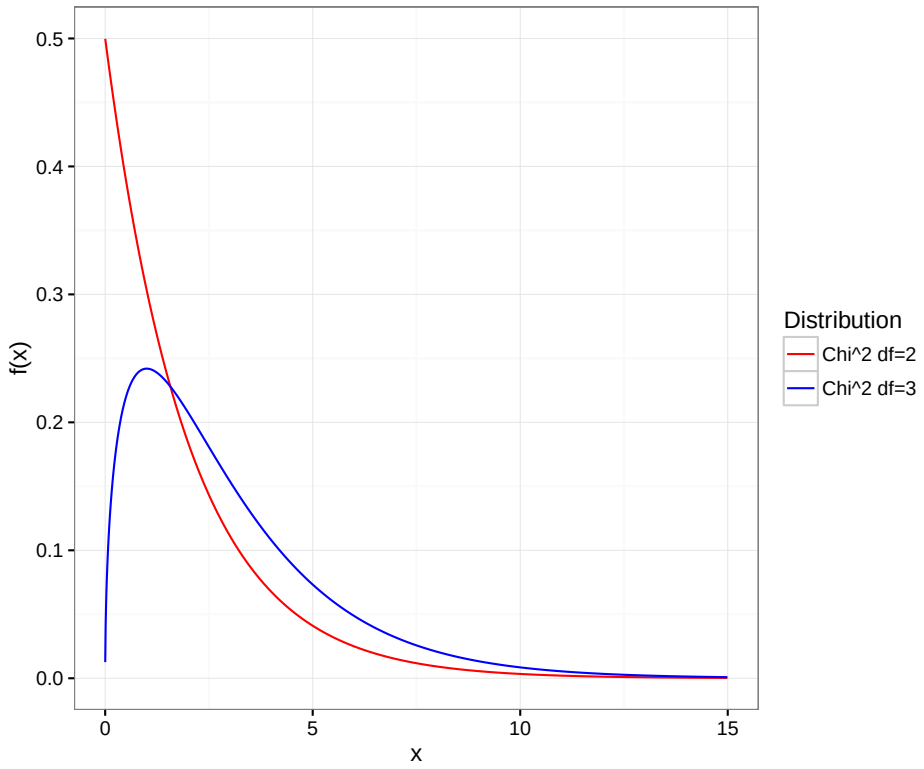
Chi-Squared Distribution

A χ^2 distribution with d degrees of freedom is equivalent to the sum of d independent Normal$(0, 1)$ random variables.

$$\chi^2_d \sim Z_1^2 + Z_2^2 + \cdots + Z_d^2$$

where $Z_1, Z_2, \ldots, Z_d$ are iid Normal$(0, 1)$.

Chi-Squared PDFs



Expected Counts

Observed counts:

```
> tbl

        Left Right
  Female    7   110
  Male     10   108
```

Expected (under H_0) counts:

```
> n <- sum(tbl)
> p <- sum(tbl[1,])/n # freq Female
> q <- sum(tbl[,1])/n # freq Left
> expected <- n * matrix(c(p*q, (1-p)*q, p*(1-q), (1-p)*(1-q)),
+                        nrow=2)
> expected
         [,1]     [,2]
[1,] 8.46383 108.5362
[2,] 8.53617 109.4638
```

Chi-Squared Statistic

The chi-squared statistic is calculated as

$$X^2 = \sum \frac{(O-E)^2}{E}$$

where O is the observed count, E is the expected count, and the sum is taken over all cells in the table.

Calculate the Statistic

```
> X2 <- sum((tbl - expected)^2 / expected)
> X2
[1] 0.5435149
>
> chisq.test(tbl, correct=FALSE)$statistic # equals X2
X-squared
0.5435149
> chisq.test(tbl)$statistic # with continuity correction
X-squared
0.2356302
```

Calculate the P-value

The null distribution of X^2 is a χ^2 distribution with d degrees of freedom. We calculate d by $d = (r-1)(c-1)$ where r is the number of rows and c is the number of columns.

```
> 1-pchisq(X2, df=1)
[1] 0.4609797
>
> chisq.test(tbl, correct=FALSE)$p.value
[1] 0.4609797
```

Derivation

The theoretical derivation of this test is beyond the scope of this course.

However, it is worth noting that it is related to the Z-statistic approximation from last week:

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

Guidelines for Practice

- The total number of observations should be "large" so that ...
- The expected number of counts per cell should be 10 or greater
- The observed number of counts per cell shoudl be 5 or greater

When these are violated, continuity corrections and simulation based p-values can be used... or other tests can be used such as Fisher's Exact Test — see `fisher.test()`.

Clapping and Writing Hand

```
> tbl = table(survey$Clap, survey$W.Hnd)
> tbl

         Left Right
  Left      9    29
  Neither   5    45
  Right     4   143
```

Note that now one of the categorical variables takes three values.

Also note the existence of low cell counts.

Chi-Squared Test Via Simulation

```
> chisq.test(tbl)
Warning in chisq.test(tbl): Chi-squared approximation may be
incorrect

    Pearson's Chi-squared test

data:  tbl
X-squared = 19.252, df = 2, p-value = 6.598e-05
```

We address this warning by simulating tables from the null hypothesis.

```
> chisq.test(tbl, simulate.p.value = TRUE, B=10000)

    Pearson's Chi-squared test with simulated p-value
    (based on 10000 replicates)

data:  tbl
X-squared = 19.252, df = NA, p-value = 9.999e-05
```

Exercise Vs. Writing Hand

```
> tbl = table(survey$Exer, survey$W.Hnd)
> tbl

       Left Right
  Freq    7   107
  None    3    21
  Some    8    90
>
> chisq.test(tbl, simulate.p.value = TRUE, B=10000)

    Pearson's Chi-squared test with simulated p-value
    (based on 10000 replicates)

data:  tbl
X-squared = 1.2065, df = NA, p-value = 0.5532
```

Smoking Vs. Exercise

```
> tbl = table(survey$Smoke, survey$Exer)
> tbl

        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
>
> chisq.test(tbl, simulate.p.value = TRUE, B=10000)

    Pearson's Chi-squared test with simulated p-value
    (based on 10000 replicates)

data:  tbl
X-squared = 5.4885, df = NA, p-value = 0.4794
```

What feature of the data is this test ignoring?

Goodness of Fit Tests

The `chisq.tes()` function also performs goodness of fit tests. These are goodness of fit tests of a set of probabilities, very related to our tests of proportions from last week.

For example, suppose we want to test whether a six-sided die is fair. We roll the die 100 times and record the frequency with which we observe each face.

```
> die
die
 1  2  3  4  5  6
17 14 21 28 14  6
> chisq.test(x=die, p=rep(1/6, 6))

    Chi-squared test for given probabilities

data:  die
X-squared = 16.52, df = 5, p-value = 0.005506
```

Two Quantitative Variables

Correlation

- It is often the case that two or more quantitative variables are measured on each unit of observation (such as an individual).
- We are then often interested in characterizing how pairs of variables are associated or how they vary together.
- A common measure that is used is called "correlation", which is most well suited for measuring linear associations

Sample Correlation

Suppose we observe n pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Their sample correlation is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (1)$$

$$= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} \quad (2)$$

where s_x and s_y are the sample standard deviations of each measured variable.

Ranked-Based Correlation

- There are other ways to measure correlation that are less reliant on linear trends in covariation and are also more robust to outliers.
- Specifically, one can convert each measured variable to ranks by size (1 for the smallest, n for the largest) and then use a formula for correlation designed for these ranks.
- One popular measure of rank-based correlation is the Spearman correlation.

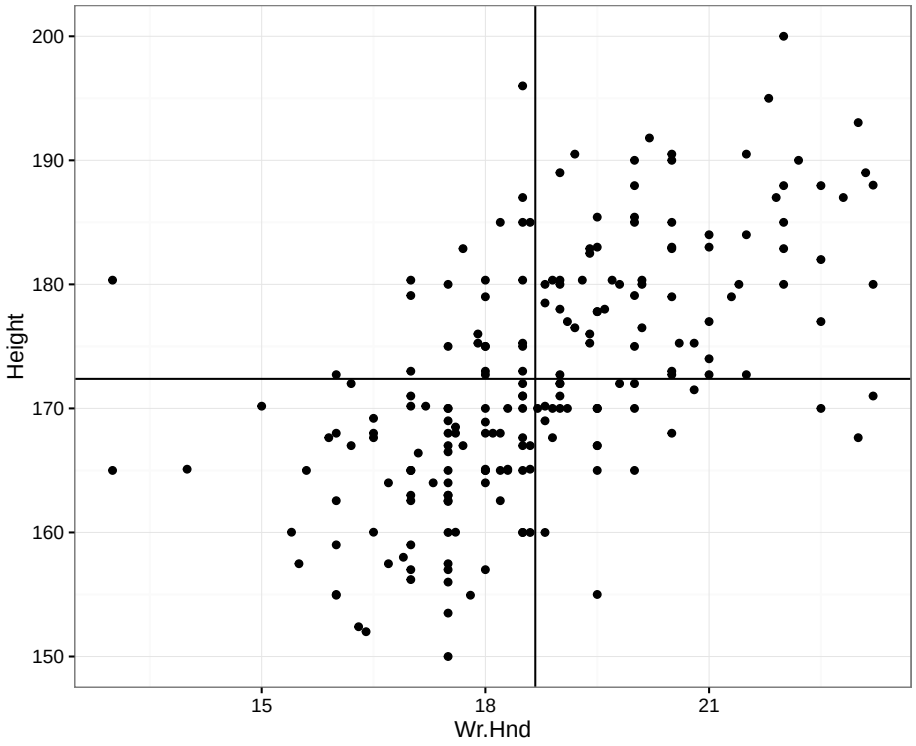
Population Correlation

Suppose there are two random variables \$X and Y. Their population correlation is

$$\rho_{XY} = \frac{\mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Hand Size Vs. Height

```
> ggplot(data = survey, mapping=aes(x=Wr.Hnd, y=Height)) +
+   geom_point() + geom_vline(xintercept=mean(survey$Wr.Hnd, na.rm=TRUE)) +
+   geom_hline(yintercept=mean(survey$Height, na.rm=TRUE))
```



Calculating Correlation

```
> str(cor)
function (x, y = NULL, use = "everything", method = c("pearson",
    "kendall", "spearman"))
>
> cor(survey$Wr.Hnd, survey$Height,
+     use="pairwise.complete.obs")
[1] 0.6009909
```

```
> df <- survey %>% dplyr::select(Wr.Hnd, Height) %>% na.omit()
> sum((df$Wr.Hnd - mean(df$Wr.Hnd)) *
+       (df$Height - mean(df$Height))) /
+   ((nrow(df)-1) * sd(df$Wr.Hnd) * sd(df$Height))
[1] 0.6009909
```

Example Correlations

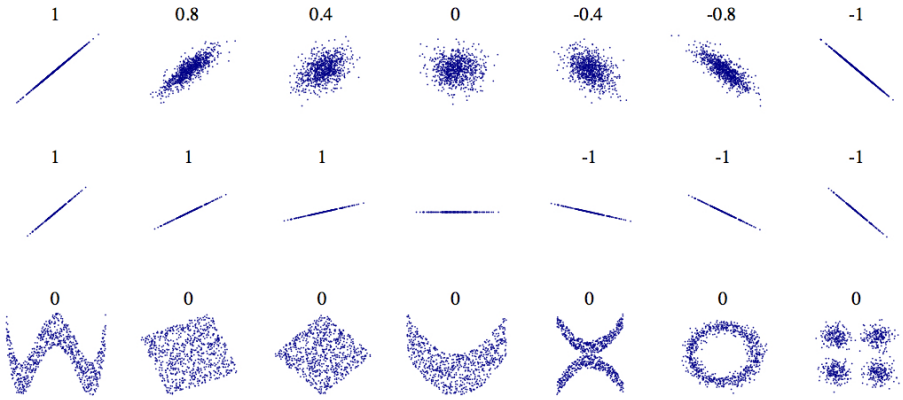


Image from Wikipedia.

HT of Correlation

```
> str(cor.test)
function (x, ...)
```

From the help file:

```
Usage

cor.test(x, ...)

## Default S3 method:
cor.test(x, y,
         alternative = c("two.sided", "less", "greater"),
         method = c("pearson", "kendall", "spearman"),
         exact = NULL, conf.level = 0.95, continuity = FALSE,
         ...)

## S3 method for class 'formula'
cor.test(formula, data, subset, na.action, ...)
```

HT of Correlation

```
> cor.test(x=survey$Wr.Hnd, y=survey$Height)

    Pearson's product-moment correlation

data:  survey$Wr.Hnd and survey$Height
t = 10.792, df = 206, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5063486 0.6813271
sample estimates:
      cor
0.6009909
```

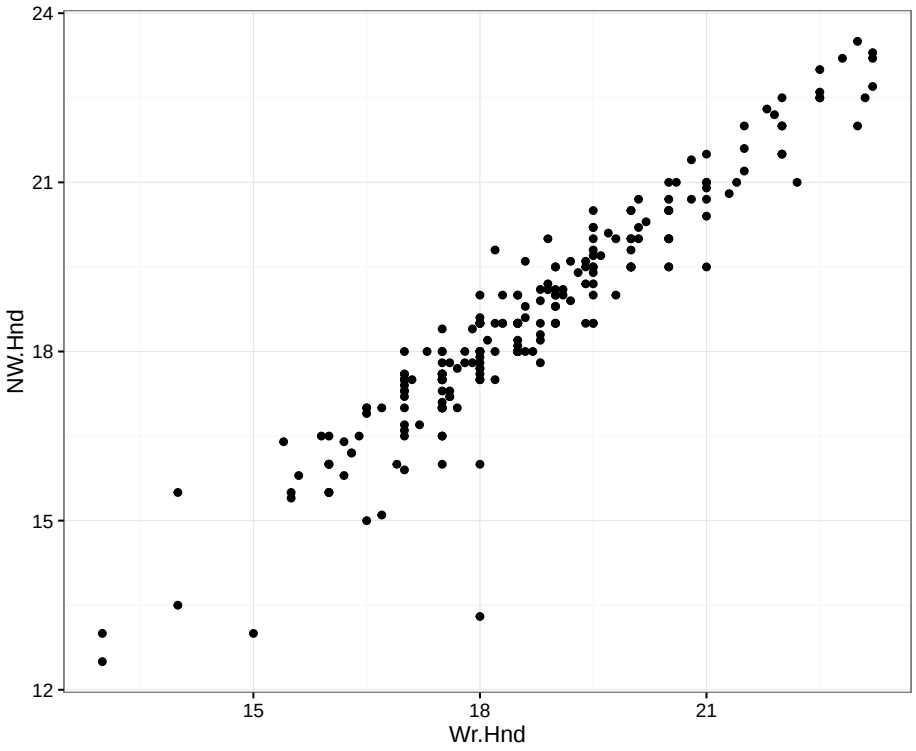
HT By Hand

Compare the following to the above output of `cor.test()`.

```
> r <- cor(survey$Wr.Hnd, survey$Height,
+     use="pairwise.complete.obs")
> df <- sum(complete.cases(survey[,c("Wr.Hnd", "Height")]))-2
> # dplyr way to get df:
> # df <- (survey %>% select(Wr.Hnd, Height) %>%
> #         na.omit() %>% nrow())-2
>
> tstat <- r/sqrt((1 - r^2)*df)
> tstat
[1] 0.05239001
>
> pvalue <- 2*pt(q=-abs(tstat), df=df)
> pvalue
[1] 0.9582687
```

Hand Sizes

```
> ggplot(data = survey) +
+   geom_point(aes(x=Wr.Hnd, y=NW.Hnd))
```



Correlation of Hand Sizes

```
> cor.test(x=survey$Wr.Hnd, y=survey$NW.Hnd)

    Pearson's product-moment correlation

data:  survey$Wr.Hnd and survey$NW.Hnd
t = 45.712, df = 234, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9336780 0.9597816
sample estimates:
      cor
0.9483103
```

Davis Data

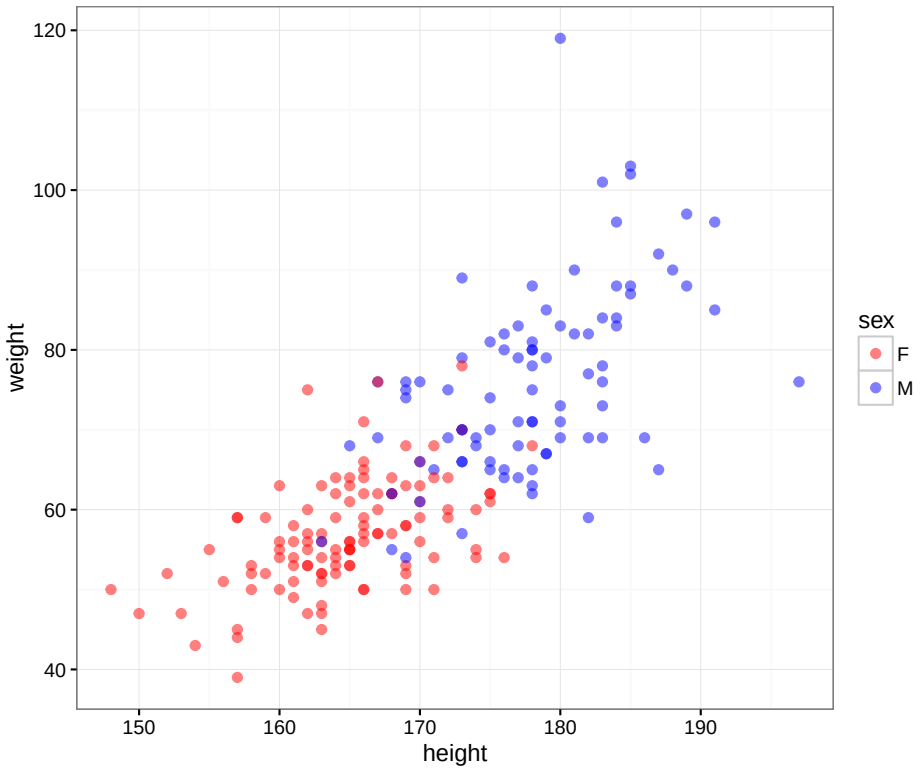
```
> library("car")
> data("Davis", package="car")
```

```
> htwt <- tbl_df(Davis)
> htwt[12,c(2,3)] <- htwt[12,c(3,2)]
> head(htwt)
Source: local data frame [6 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
```

Height and Weight

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



Correlation Test

```
> cor.test(x=htwt$height, y=htwt$weight)

    Pearson's product-moment correlation

data:  htwt$height and htwt$weight
t = 17.04, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7080838 0.8218898
sample estimates:
      cor
0.7710743
```

Correlation Test with Outlier

Recall we had to fix an error in the data, which we noticed as an outlier in the scatterplot. Here is the effect of the outlier:

```
> cor.test(x=Davis$height, y=Davis$weight)

    Pearson's product-moment correlation

data:  Davis$height and Davis$weight
t = 2.7179, df = 198, p-value = 0.007152
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.05228435 0.31997151
sample estimates:
      cor
0.1896496
```

Correlation Test with Outlier

Let's use the Spearman rank-based correlation:

```
> cor.test(x=Davis$height, y=Davis$weight, method="spearman")
Warning in cor.test.default(x = Davis$height, y = Davis$weight,
method = "spearman"): Cannot compute exact p-value with ties

    Spearman's rank correlation rho

data:  Davis$height and Davis$weight
S = 308750, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7684305
```

Correlation Among Females

```
> htwt %>% filter(sex=="F") %>%
+   cor.test(~ height + weight, data = .)

    Pearson's product-moment correlation

data:  height and weight
t = 6.2801, df = 110, p-value = 6.922e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3627531 0.6384268
sample estimates:
```

```
      cor
0.5137293
```

Correlation Among Males

```
> htwt %>% filter(sex=="M") %>%
+   cor.test(~ height + weight, data = .)

    Pearson's product-moment correlation

data:  height and weight
t = 5.9388, df = 86, p-value = 5.922e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3718488 0.6727460
sample estimates:
      cor
0.5392906
```

Why are the stratified correlations lower?

Extras

License

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

Source Code

https://github.com/SML201/lectures/tree/master/week9

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       MASS_7.3-45     broom_0.4.0
[4] dplyr_0.4.3     ggplot2_2.1.0   knitr_1.12.3
[7] magrittr_1.5    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          tools_3.2.3         digest_0.6.9
 [7] lme4_1.1-11         evaluate_0.8        memoise_1.0.0
[10] nlme_3.1-125        gtable_0.2.0        lattice_0.20-33
[13] mgcv_1.8-11         Matrix_1.2-3        psych_1.5.8
[16] DBI_0.3.1           yaml_2.1.13         parallel_3.2.3
[19] SparseM_1.7         stringr_1.0.0       MatrixModels_0.4-1
[22] grid_3.2.3          nnet_7.3-12         R6_2.1.2
[25] rmarkdown_0.9.5     minqa_1.2.4         reshape2_1.4.1
[28] tidyr_0.4.1         scales_0.4.0        htmltools_0.3
[31] splines_3.2.3       assertthat_0.1      pbkrtest_0.4-6
[34] mnormt_1.5-3        colorspace_1.2-6    quantreg_5.21
[37] labeling_0.3        stringi_1.0-1       lazyeval_0.1.10
[40] munsell_0.4.3
```