Chapter 6: Basic Inference Methods

Alex Chi

Exercises

```
marathoners = read.csv("http://personal.bgsu.edu/~mrizzo/Rx/Rx-data/nyc-marathon.csv")
```

6.1 (Gender of marathoners).

In 2000, the proportion of females who competed in marathons in the United States was 0.375. One wonders if the proportion of female marathoners has changed in the ten-year period from 2000 to 2010. One collects the genders of 276 people who competed in the 2010 New York City Marathon – in this sample, 120 were women.

a. If p denotes the proportion of 2010 marathoners who are female, use the prop.test function to test the hypothesis that p = 0.375. Store the calcu- lations of the test in the variable Test.

```
Test = prop.test(120, 276, p=0.375)
Test
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  120 out of 276, null probability 0.375
## X-squared = 3.9575, df = 1, p-value = 0.04666
## alternative hypothesis: true p is not equal to 0.375
## 95 percent confidence interval:
##  0.3758309 0.4955799
## sample estimates:
##         p
## 0.4347826
```

b. From the components of Test, construct a 95% interval estimate for p.

```
Test$conf.int
```

```
## [1] 0.3758309 0.4955799
## attr(,"conf.level")
## [1] 0.95
```

c. Using the function binom.test, construct an exact-test of the hypothesis. Compare this test with the large-sample test used in part (a).

```
binom.test(120, 276, p=0.375)
```

```
##
##  Exact binomial test
##
## data:  120 and 276
## number of successes = 120, number of trials = 276, p-value =
## 0.04644
## alternative hypothesis: true probability of success is not equal to 0.375
## 95 percent confidence interval:
##  0.3754670 0.4955137
## sample estimates:
## probability of success
##              0.4347826
```

6.2 (Ages of marathoners)

The datafile "nyc.marathon.txt" contains the gender, age, and completion time (in minutes) for 276 people who completed the 2010 New York City Marathon. It was reported that the mean ages of men and women marathoners in 2005 were respectively 40.5 and 36.1.

a. Create a new dataframe "women.marathon" that contains the ages and completion times for the women marathoners.

```r
women.marathon = marathoners[marathoners$Gender == "female", ]
women.marathon
```

```
##       Minutes Gender Age
## 2    268.4667 female  30
## 3    463.2833 female  43
## 4    286.5500 female  54
## 5    408.1000 female  37
## 9    220.6667 female  44
## 15   454.5667 female  32
## 17   281.7500 female  41
## 19   389.2500 female  58
## 20   217.3000 female  40
## 21   401.4500 female  38
## 26   299.4000 female  51
## 27   241.6333 female  47
## 31   274.5500 female  59
## 32   365.5333 female  27
## 33   301.7500 female  43
## 34   294.1333 female  38
## 35   334.6667 female  55
## 37   308.1500 female  31
## 38   405.1333 female  51
## 40   240.3167 female  42
```

```
## 42  378.0167 female  45
## 44  357.3333 female  35
## 52  322.9333 female  46
## 53  326.2500 female  31
## 54  358.9500 female  43
## 55  301.2333 female  52
## 57  359.7000 female  37
## 62  427.4333 female  58
## 66  332.9500 female  41
## 67  317.8167 female  35
## 74  312.8333 female  34
## 75  314.7667 female  26
## 79  224.3167 female  45
## 83  365.3833 female  40
## 87  328.0667 female  35
## 89  236.1500 female  42
## 90  206.8833 female  43
## 93  238.4000 female  33
## 94  276.0000 female  28
## 98  396.4333 female  48
## 105 272.1667 female  40
## 106 435.1500 female  53
## 108 348.7833 female  39
## 109 292.5667 female  43
## 113 295.5167 female  23
## 115 298.2667 female  51
## 119 226.8667 female  36
## 120 201.8333 female  26
## 122 210.8167 female  28
## 123 302.4667 female  29
## 131 195.2833 female  22
## 132 223.0000 female  45
## 133 302.4167 female  30
## 134 305.9000 female  48
## 135 235.4167 female  43
## 139 298.5333 female  44
## 144 359.6667 female  55
## 145 408.4667 female  49
## 146 330.2000 female  43
## 148 354.8667 female  50
## 149 313.2333 female  45
## 150 212.3500 female  44
## 152 285.5167 female  28
## 153 218.8333 female  46
## 157 299.9167 female  56
```

```
## 160 194.9333 female  46
## 161 239.3167 female  26
## 162 223.0500 female  41
## 163 450.7333 female  60
## 164 259.1167 female  29
## 167 275.3500 female  44
## 168 378.3333 female  57
## 169 334.6500 female  59
## 170 285.8500 female  51
## 174 358.0167 female  31
## 176 366.1000 female  52
## 179 316.9500 female  49
## 183 287.8667 female  49
## 187 223.5833 female  35
## 193 379.3000 female  41
## 202 252.3333 female  40
## 204 349.7667 female  42
## 205 294.1667 female  50
## 206 286.4833 female  52
## 207 329.4667 female  39
## 208 208.8833 female  43
## 209 289.7333 female  34
## 215 378.6333 female  36
## 216 206.2333 female  36
## 217 344.8833 female  33
## 219 379.3333 female  43
## 221 286.8833 female  47
## 222 207.9333 female  28
## 223 287.8500 female  36
## 231 325.7667 female  44
## 232 224.3167 female  40
## 235 234.4167 female  47
## 237 373.3667 female  48
## 241 238.7000 female  46
## 242 194.5500 female  41
## 244 362.6333 female  52
## 247 391.4167 female  27
## 251 220.2333 female  51
## 255 210.8167 female  29
## 256 326.4500 female  27
## 257 224.1833 female  42
## 275 231.4333 female  50
```

b. Use the t.test function to construct a test of the hypothesis that the mean age of women marathoners is equal to 36.1.

```r
Test = t.test(women.marathon$Age, mu=36.1)
Test
```

```
##
##  One Sample t-test
##
## data:  women.marathon$Age
## t = 6.1735, df = 106, p-value = 1.249e-08
## alternative hypothesis: true mean is not equal to 36.1
## 95 percent confidence interval:
##  39.80704 43.31446
## sample estimates:
## mean of x
##  41.56075
```

c. As an alternative method, use the wilcox.test function to test the hy- pothesis that the median age of women marathoners is equal to 36.1. Com- pare this test with the t-test used in part (b).

```r
Test = wilcox.test(women.marathon$Age, mu=36.1)
Test
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  women.marathon$Age
## V = 4522, p-value = 3.879e-07
## alternative hypothesis: true location is not equal to 36.1
```

d. Construct a 90% interval estimate for the mean age of women marathoners

```r
Test = t.test(women.marathon$Age, conf.level=0.9)
Test
```

```
##
##  One Sample t-test
##
## data:  women.marathon$Age
## t = 46.985, df = 106, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  40.09296 43.02854
## sample estimates:
## mean of x
##  41.56075
```

```r
Test$conf.int
```

```
## [1] 40.09296 43.02854
```

```
## attr(,"conf.level")
## [1] 0.9
```

6.3 (Ages of marathoners, continued).

From the information in the 2005 report, one may believe that men marathoners tend to be older than women marathons.

a. Use the t.test function to construct a test of the hypothesis that the mean ages of women and men marathoners are equal against the alternative hypothesis that the mean age of men is larger.

```r
t.test(Age ~ Gender, marathoners, alternative="less")
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by Gender
## t = -2.4519, df = 252.66, p-value = 0.007443
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -0.974724
## sample estimates:
## mean in group female   mean in group male
##             41.56075             44.54438
```

b. Construct a 90% interval estimate for the difference in mean ages of men and women marathoners.

```r
Test = t.test(Age ~ Gender, marathoners, conf.level=0.9)
Test
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by Gender
## t = -2.4519, df = 252.66, p-value = 0.01489
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -4.992538 -0.974724
## sample estimates:
## mean in group female   mean in group male
##             41.56075             44.54438
```

```r
diff(Test$estimate)
```

```
## mean in group male
##           2.983631
```

c. Use the alternative Mann-Whitney-Wilcoxon test (function wilcox.test) to test the hypothesis that the ages of the men and ages of the women come from populations with the same location parameter against the alternative that the population of ages of the men have a larger location parameter. Compare the result of this test with the t-test performed in part (a).

```
wilcox.test(Age ~ Gender, marathoners, alternative="less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Age by Gender
## W = 7694.5, p-value = 0.01852
## alternative hypothesis: true location shift is less than 0
```

6.4 (Measuring the length of a string).

An experiment was performed in an introductory statistics class to illustrate the concept of measurement bias. The instructor held up a string in front of the class and each student guessed at the string's length. The following are the measurements from the 24 students (in inches).

22 18 27 23 24 15 26 22 24 25 24 18 18 26 20 24 27 16 30 22 17 18 22 26

a. Use the scan function to enter these measurements into R.

```
x = c(22, 18, 27, 23, 24, 15, 26, 22, 24, 25, 24, 18, 18, 26, 20, 24, 27, 16, 30, 22, 17, 18
```

b. The true length of the string was 26 inches. Assuming that this sample of measurements represents a random sample from a population of student measurements, use the t.test function to test the hypothesis that the mean measurement μ is different from 26 inches.

```
t.test(x, mu=26)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -4.6148, df = 23, p-value = 0.0001216
## alternative hypothesis: true mean is not equal to 26
## 95 percent confidence interval:
##  20.569 23.931
## sample estimates:
## mean of x
##     22.25
```

c. Use the t.test function to find a 90% confidence interval for the popula- tion mean μ.

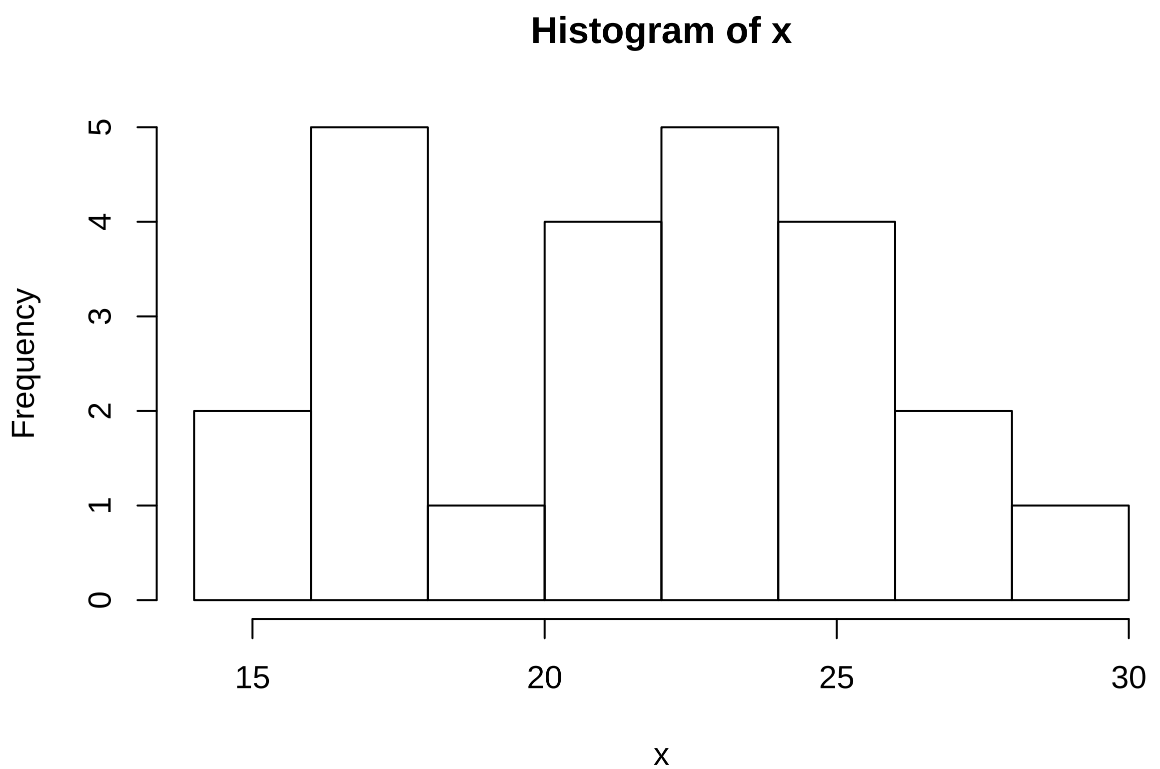
```
t.test(x, mu=26, conf.level=0.9)
```

```
##
##  One Sample t-test
```

```
##
## data:  x
## t = -4.6148, df = 23, p-value = 0.0001216
## alternative hypothesis: true mean is not equal to 26
## 90 percent confidence interval:
##  20.8573 23.6427
## sample estimates:
## mean of x
##     22.25
```

d. The t-test procedure assumes the sample is from a population that is normally distributed. Construct a normal probability plot of the measure- ments and decide if the assumption of normality is reasonable.

```r
hist(x)
```




```r
ks.test(x, "pnorm")
```

```
## Warning in ks.test(x, "pnorm"): ties should not be present for the
## Kolmogorov-Smirnov test

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

6.5 (Comparing snowfall of Buffalo and Cleveland).

The datafile “buf- falo.cleveland.snowfall.txt” contains the total snowfall in inches for the cities Buffalo and Cleveland for the seasons 1968-69 through 2008-09.

```
snowfall = read.table("Rx-data/buffalo.cleveland.snowfall.txt", header=TRUE)
```

a. Compute the differences between the Buffalo snowfall and the Cleveland snowfall for all seasons.

```
snowfall$diff = snowfall$Cleveland - snowfall$Buffalo
snowfall
```

```
##       SEASON Cleveland Buffalo   diff
## 1  2008-2009      79.7   100.2  -20.5
## 2  2007-2008      77.2   103.8  -26.6
## 3  2006-2007      76.5    88.9  -12.4
## 4  2005-2006      50.6    78.2  -27.6
## 5  2004-2005     117.9   109.1    8.8
## 6  2003-2004      91.2   100.9   -9.7
## 7  2002-2003      95.7   111.3  -15.6
## 8  2001-2002      46.0   132.4  -86.4
## 9  2000-2001      78.1   158.7  -80.6
## 10 1999-2000      60.1    63.6   -3.5
## 11 1998-1999      62.4   100.5  -38.1
## 12 1997-1998      34.0    75.6  -41.6
## 13 1996-1997      55.9    97.6  -41.7
## 14 1995-1996     101.1   141.4  -40.3
## 15 1994-1995      43.6    74.6  -31.0
## 16 1993-1994      72.5   112.7  -40.2
## 17 1992-1993      88.5    93.2   -4.7
## 18 1991-1992      65.7    92.8  -27.1
## 19 1990-1991      47.1    57.5  -10.4
## 20 1989-1990      62.6    93.7  -31.1
## 21 1988-1989      54.8    67.4  -12.6
## 22 1987-1988      71.3    56.4   14.9
## 23 1986-1987      55.8    67.5  -11.7
## 24 1985-1986      58.3   114.7  -56.4
## 25 1984-1985      63.7   107.2  -43.5
## 26 1983-1984      79.4   132.5  -53.1
## 27 1982-1983      38.0    52.4  -14.4
## 28 1981-1982     100.5   112.4  -11.9
## 29 1980-1981      60.5    60.9   -0.4
## 30 1979-1980      38.7    68.4  -29.7
## 31 1978-1979      38.3    97.3  -59.0
## 32 1977-1978      90.1   154.3  -64.2
## 33 1976-1977      63.4   199.4 -136.0
## 34 1975-1976      54.4    82.5  -28.1
```

```
## 35 1974-1975      67.0    95.6  -28.6
## 36 1973-1974      58.5    88.7  -30.2
## 37 1972-1973      68.5    78.8  -10.3
## 38 1971-1972      45.6   109.9  -64.3
## 39 1970-1971      51.4    97.0  -45.6
## 40 1969-1970      53.4   120.5  -67.1
## 41 1968-1969      37.0    78.4  -41.4
```

b. Using the t.test function with the difference data, test the hypothesis that Buffalo and Cleveland get, on average, the same total snowfall in a season.

```r
t.test(snowfall$diff, mu = 0)
```

```
## 
##  One Sample t-test
## 
## data:  snowfall$diff
## t = -7.5692, df = 40, p-value = 3.061e-09
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -42.45731 -24.56221
## sample estimates:
## mean of x 
## -33.50976
```

c. Use the t.test function to construct a 95% confidence interval of the mean difference in seasonal snowfall.

```r
t.test(snowfall$diff, mu = 0, conf.level = 0.95)
```

```
## 
##  One Sample t-test
## 
## data:  snowfall$diff
## t = -7.5692, df = 40, p-value = 3.061e-09
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -42.45731 -24.56221
## sample estimates:
## mean of x 
## -33.50976
```

6.6 (Comparing Etruscan and modern Italian skulls).

Researchers were interested if ancient Etruscans were native to Italy. The dataset"Etruscan- Italian.txt" contains the skull measurements from a group of Etruscans and modern Italians. There are two relevant variables in the dataset: x is the skull measurement and group is the type of skull.

```r
italian = read.table("Rx-data/Etruscan-Italian.txt")
```

a. Assuming that the data represent independent samples from normal dis- tributions, use the t.test function to test the hypothesis that the mean Etruscan skull measurement μE is equal to the mean Italian skull mea- surement μI.

```r
t.test(x ~ group, italian)
```

```
##
##  Welch Two Sample t-test
##
## data:  x by group
## t = 11.966, df = 148.82, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.459782 13.202123
## sample estimates:
## mean in group Etruscan  mean in group Italian
##               143.7738               132.4429
```

b. Use the t.test function to construct a 95% interval estimate for the dif- ference in means μE−μI.

```r
t.test(x ~ group, italian, conf.level=0.95)$conf.int
```

```
## [1]  9.459782 13.202123
## attr(,"conf.level")
## [1] 0.95
```

c. Use the two-sample Wilcoxon procedure implemented in the function wilcox.test to find an alter- native 95% interval estimate for the difference μE−μI.

```r
wilcox.test(x ~ group, italian, conf.int=TRUE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x by group
## W = 5401, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##   9.999978 13.000070
## sample estimates:
## difference in location
##               11.00004
```

6.7 (President's heights).

In Example 1.2, the height of the election winner and loser were collected for the U.S. Presidential elections of 1948 through 2008. Suppose you are interested in testing the hypothesis that the mean height of the election winner is equal to the mean height of the election loser. Assuming that this data represent paired data from a hypothetical population of elections, use the t.test function to test this hypothesis. Interpret the results of this test.

```r
winner = c(185, 182, 182, 188, 188, 188, 185, 185, 177,
           182, 182, 193, 183, 179, 179, 175)
opponent = c(175, 193, 185, 187, 188, 173, 180, 177, 183,
             185, 180, 180, 182, 178, 178, 173)
t.test(winner, opponent)
```

```
##
##  Welch Two Sample t-test
##
## data:  winner and opponent
## t = 1.2414, df = 29.012, p-value = 0.2244
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.456766  5.956766
## sample estimates:
## mean of x mean of y
##  183.3125  181.0625
```

Therefore, they're not equal.