Homework set 9 - Due 4/19/2013

Math 3200 – Renato Feres

Preliminaries

The purpose of this homework assignment is to introduce some of the R-functions that can be used to perform the kind of data analysis we studied in chapters 7, 8, and 9. Not all the tests listed below will be needed for this problems set. I've added them for reference; the main tests for inference on proportions you are going to use are `prop.test` and `chisq.test`.

Since the main point of this assignment is to illustrate R's capabilities, I recommend that you do all the problems using it (rather than doing them "by hand.") This is, in any event, the easiest way. A few references for this material (from which I'm taking freely here) are: *Introductory Statistics with R*, by Peter Dalgaard, Springer 2002; *R Cookbook*, by Paul Teetor, O'Reilly 2011; and *Simple R — Using R for introductory statistics*, by John Verzani `http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf`.

- **One-sample t test.** We have already used R's `t.test` in previous assignments. Recall that t-tests are based on the assumption that data come from a normal distribution. In the one-sample case we have data $x_1, x_2, \ldots, x_n$ assumed independent realizations of random variables with distribution $N(\mu, \sigma^2)$, and we wish to test the null hypothesis that $\mu = \mu_0$.

  Here is an example (from Dalgaard's book) of using `t.test`. Consider the daily energy intake in kJ for 11 women:

  ```
  daily.intake=c(5260,5470,5640,6180,6390,6515,6805,7515,7515,8230,8770)
  ```

  The recommended value of women's energy intake is 7725 kJ. We wish to know whether the energy intake in the sample deviates systematically from the recommended value. Thus we wish to test whether this distribution might have a mean $\mu = 7725$. For this we use `t.test` as follows:

  ```
  > t.test(daily.intake,mu=7725)

  One Sample t-test

  data:  daily.intake
  t = -2.8208, df = 10, p-value = 0.01814
  alternative hypothesis: true mean is not equal to 7725
  95 percent confidence interval:
   5986.348 7520.925
  sample estimates:
  mean of x
   6753.636
  ```

  Most of this output should be self-explanatory. Note that the level of the test and the type of alternative hypothesis were defaulted to 95% and "two-sided." These values can be changed. If we want a confidence level 0.80

and the alternative H_1 that the recommended mean μ is greater than the actual mean of the distribution from which the data was sampled:

```
> t.test(daily.intake,mu=7725,conf.level=0.80,alternative="less")

One Sample t-test

data:  daily.intake
t = -2.8208, df = 10, p-value = 0.009069
alternative hypothesis: true mean is less than 7725
80 percent confidence interval:
     -Inf 7056.351
sample estimates:
mean of x
 6753.636
```

The values for `alternative` are `two.sided, less, greater`.

- **Two-sample t-test.** The two-sample t-test is used to test the hypothesis that tow samples may be assumed to come from distributions with the same mean. Suppose the data vectors are $x = (x_1, \ldots, x_{n_1})$ and $y = (y_1, \ldots, y_{n_2})$, sampled from normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. We illustrate the test with simulated data:

```
> x=rnorm(10,0,1)
> y=rnorm(15,1.5,1)
> t.test(x,y)

Welch Two Sample t-test

data:  x and y
t = -3.5348, df = 17.538, p-value = 0.002445
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.7924390 -0.7080206
sample estimates:
 mean of x  mean of y
-0.1737888  1.5764410
```

- **Comparison of variances.** Although we can perform the t-test as above without assuming that the two sampled distributions have the same variance, it may still be useful to know whether the variances are the same or not. When they are the same, it is not necessary to estimate the d.f. from the data as we have seen in class. We can compare variances using an F-test. (This relates to the theory of section 8.4, which we didn't cover in class, but is useful to know.) The F-test is implemented in R by the function `var.test`:

```
> var.test(x,y)

F test to compare two variances
```

```
data:  x and y
F = 1.3154, num df = 9, denom df = 14, p-value = 0.6229
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4098717 4.9958324
sample estimates:
ratio of variances
          1.315402
```

- **T-test for matched pairs data.** Paired tests for means are used when there are two measurements on the same experimental unit. Their theory is essentially based on taking differences and thus reducing the problem that that of a one-sample test. The function `t.test` can be used for this by specifying that the data are paired:

```
> x=1+rnorm(10,0,1)
> y=2+rnorm(10,0,1)
> t.test(x,y,paired=TRUE)

Paired t-test

data:  x and y
t = -2.2646, df = 9, p-value = 0.0498
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.8020868104 -0.0009904147
sample estimates:
mean of the differences
             -0.9015386
```

- **Testing a sample proportion.** The R-function `prop.test(x,n,p)` can be used to do inference on proportion. Here `x` is the number of "successes" in a binary classification of the data, `n` is the number of observations (so that `x/n` is the observed proportion) and `p` is the probability of successes as specified by the null-hypothesis. For example,

```
> prop.test(47,259,0.2)

1-sample proportions test with continuity correction

data:  47 out of 259, null probability 0.2
X-squared = 0.4462, df = 1, p-value = 0.5042
alternative hypothesis: true p is not equal to 0.2
95 percent confidence interval:
 0.1375856 0.2350378
sample estimates:
        p
0.1814672
```

A similar but more precise test for proportion uses the binomial distribution rather than the approximate normal:

```
> binom.test(47,259,0.2)

Exact binomial test

data:  47 and 259
number of successes = 47, number of trials = 259, p-value = 0.4853
alternative hypothesis: true probability of success is not equal to 0.2
95 percent confidence interval:
 0.1364881 0.2339019
sample estimates:
probability of success
             0.1814672
```

- **Testing two independent proportions.** The function `prop.test` can also be used to compare two or more proportions. For this purpose, the arguments should be given as two vectors. For example, suppose that we wish to compare two proportions p_1, p_2 (say, of successes associated to two methods for baking a cake. Success means that the cake raises, and failure that it does not). A test is set to decide whether they are equal or not. We may also wish to obtain a confidence interval for the difference $p_1 - p_2$. Suppose that estimators for p_1 and p_2 computed from samples of sizes $n_1 = 12$ and $n_2 = 13$, and number of successes 9 and 4, respectively, give the following values:

$$\hat{p}_1 = 9/12, \;\; \hat{p}_2 = 4/13.$$

To apply `prop.test` we may do as follows:

```
> totals = c(12,13)
> successes = c(9,4)
> prop.test(successes,totals)

2-sample test for equality of proportions with
continuity correction

data:  successes out of totals
X-squared = 3.2793, df = 1, p-value = 0.07016
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01151032 0.87310506
sample estimates:
   prop 1    prop 2
0.7500000 0.3076923
```

This gives the P-value 0.07; so at the level 0.05 we cannot discard the hypothesis that the true proportions are equal. The 95% confidence interval given for the difference $p_1 - p_2$ is [0.0120.87]. Note how, in this case, the conclusion to derive from the P-value (that we cannot reject H_0) is contradicted by the confidence interval,

which excludes 0. Somehow the approximation used for the P-value gave a more conservative (bigger) value. It is possible to obtain a more precise approximation for the P-value using `fisher.test`. This can be done as follows. We first need to present the data as a contingency table:

	successes	failures
method 1	9	3
method 2	4	9

To apply `fisher.test` we first define the data matrix:

```
> data=matrix(c(9,4,3,9),2)
> data
     [,1] [,2]
[1,]    9    3
[2,]    4    9
```

Then

```
> fisher.test(data)

Fisher's Exact Test for Count Data

data:  data
p-value = 0.04718
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.9006803 57.2549701
sample estimates:
odds ratio
  6.180528
```

Using this test, a more precise P-value is now 0.047, which is in agreement with the conclusion derived from the confidence interval using `prop.test`. Note, though, that the confidence interval now is given for the *odds ratio* rather than difference of the proportions. The odd ratio is defined as $(p_1/(1-p_1))/(p_2/(1-p_2))$.

The `prop.test` assumes by default that you would like to use *Yates' continuity correction.* Without describing what it is, let me simply say that the test for proportion studied in section 9.2 of the textbook does not involve such a correction. The test without the correction is done as follows.

```
> totals = c(12,13)
> successes = c(9,4)
> prop.test(successes,totals,correct=FALSE)

2-sample test for equality of proportions without
continuity correction

data:  successes out of totals
```

```
X-squared = 4.8909, df = 1, p-value = 0.027
alternative hypothesis: two.sided
95 percent confidence interval:
 0.09163853 0.79297686
sample estimates:
   prop 1    prop 2
0.7500000 0.3076923
```

The P-value is now smaller; according to this new value we can reject H_0 at the level 0.05. When doing problem number 3, use `prop.test` without the correction.

- **Testing multiple proportions.** The following example illustrates the use of `prop.test` when there are several proportions to compare.

```
## Data from Fleiss (1981), p. 139.
## H0: The null hypothesis is that the four populations from which
##     the patients were drawn have the same true proportion of smokers.
## H1: The alternative is that this proportion is different in at
##     least one of the populations.
> smokers  = c( 83, 90, 129, 70 )
> patients = c( 86, 93, 136, 82 )
> prop.test(smokers, patients)

4-sample test for equality of proportions without continuity
correction

data:  smokers out of patients
X-squared = 12.6004, df = 3, p-value = 0.005585
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4
0.9651163 0.9677419 0.9485294 0.8536585
```

Instead of entering into `prop.test` two data vectors as above, we could enter one data vector and one probability vector p as in the situation described in section 9.3.2 of the textbook.

- **Chi-squared goodness of fit tests.** A goodness of fit test checks to see if the data came from some specified probability distribution. The chi-squared goodness of fit test (discussed in section 9.3.2 of the textbook) can be used to test the hypothesis that the categories of some categorical random variable have the probabilities specified by the null-hypothesis.

  For example, say that we would like to investigate whether a die is fair. We toss the die 150 times and find the following counts:

face	1	2	3	4	5	6
number of rolls	22	21	22	27	22	36

  It is reasonable to take for the null-hypothesis that the die is fair, so $H_0 : p_1 = p_2 = \cdots = p_6 = 1/6$ versus the alternative that not all the p_i are the same. We can test these hypotheses using `chisq.test` as follows.

```
> counts=c(22,21,22,27,22,36)
> probabilities=c(1,1,1,1,1,1)/6
> chisq.test(counts,p=probabilities)

Chi-squared test for given probabilities

data:  counts
X-squared = 6.72, df = 5, p-value = 0.2423
```

Note that the P-value is relatively large. At the confidence level 0.80, for example, we do not reject the null-hypothesis that the die is fair.

Here is a nice example taken from *Simple R*, by J. Verzani. The letter distribution of the 5 most frequent letters in the English language can be found (by counting from large bodies of texts) to be approximately

letter	E	T	N	R	O
percentages	29	21	17	17	16

Suppose a text written in the latin alphabet is analyzed and the number of occurrences of those 5 letters are counted, giving the following frequencies:

letter	E	T	N	R	O
frequencies	100	110	80	55	14

We wish to test whether the text was written in English or not. We do a chi-squared test:

```
> x=c(100,110,80,55,14)
> probabilities=c(29,21,17,17,16)/100
> chisq.test(x,p=probabilities)

Chi-squared test for given probabilities

data:  x
X-squared = 55.3955, df = 4, p-value = 2.685e-11
```

The P-value is now extremely small. It is very unlikely (at any reasonable significance level) that the text from which the data were collected was written in English.

There are several other uses for the Chi-squared test. I refer to the above mentioned on-line text by Verzani for more examples.

Problems

Solve the following problems using the R-function `prop.test` or `chisq.test`. If the significance level of a test is not indicated, use $\alpha = 0.05$.

1. **Problem 9.2, page 331.** While imprisoned by the Germans during World War II, the English mathematician John Kerrich tossed a coin 10,000 times and obtained 5067 heads. Let p be the probability of a head on a single toss. We wish to check if the data are consistent with the hypothesis that the coin is fair.

   (a) Set up the hypotheses. Why should the alternative be two-sided?

   (b) Calculate the P-value. Can you reject H_0 at the 0.05 level?

   (c) Find a 95% CI for the proportion of heads for Kerrich's coin.

   Solution: (a) The null-hypothesis is $H_0 : p = 1/2$, and the alternative hypothesis is $H_1 : p \neq 1/2$. The alternative is two-sided since a coin can be biased to favor either heads or tails.

   For parts (b) and (c) we use R's `prop.test`:

```
  > prop.test(5067,10000,0.5)

1-sample proportions test with continuity
correction

data:  5067 out of 10000, null probability 0.5
X-squared = 1.7689, df = 1, p-value = 0.1835
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4968504 0.5165445
sample estimates:
     p
0.5067
```

   (b) From the above, the P-value of the test is 0.1835. This is greater than $\alpha = 0.05$, so we cannot reject the hypothesis that the coin is unbiased at this level α.

   (c) The 95% confidence interval for p is $[0.4969, 0.5166]$. Clearly, this interval contains 0.5, so we do not reject H_0 at the level 0.05.

2. **Problem 9.7, page 332.** People at high risk of sudden cardiac death can be identified using the change in a signal averaged electrocardiogram before and after prescribed activities. The current method is about 80% accurate. The method was modified, hoping to improve its accuracy. The new method is tested on 50 people and gave correct results on 46 patients. Is this convincing evidence that the new method is more accurate?

   (a) Set up the hypotheses to test that the accuracy of the new method is better than that of the current method.

   (b) Perform a test of the hypotheses at $\alpha = 0.05$. What do you conclude about the accuracy of the new method?

   Solution: (a) Let p represent the actual accuracy of the new test. The null-hypothesis is $H_0 : p = 0.8$, and the alternative hypothesis is $H_1 : p > 0.8$.

   (b) We again use R's `prop.test`.

```
> prop.test(46,50,0.8,alternative='greater')

1-sample proportions test with continuity correction

data:  46 out of 50, null probability 0.8
X-squared = 3.7812, df = 1, p-value = 0.02591
alternative hypothesis: true p is greater than 0.8
95 percent confidence interval:
 0.8207834 1.0000000
sample estimates:
   p
0.92
```

The P-value of the test is 0.026, which is less than 0.05. So we are justified in concluding, at the significance level 0.05, that the new test is more accurate. In fact, with 95% confidence, the true accuracy is greater than 0.82.

3. **Problem 9.11, page 332.** A high school had 17 students receive National Merit recognition (semifinalist or commendation) out of 482 seniors in 1992 and 29 students out of 503 seniors in 1995. Does this represent a significant change in the proportion recognized at this school? Answer by doing a two-sided test for the significance of the difference in two proportions at $\alpha = 0.10$. Why is a two-sided alternative appropriate here?

   Solution: Let p_1 and p_2 denote the true proportions in 1992 and 1995, respectively. We test the hypotheses: $H_0 : p_1 = p_2$ vs. $H_1 = p_1 \neq p_2$. A two-sided alternative is appropriate here since all we want is to detect change, in either direction.

   The observed proportions are

$$\hat{p}_1 = 17/482, \;\; \hat{p}_2 = 29/503.$$

   We apply the `prop.test`:

```
> totals=c(482,503)
> successes=c(17,29)
> prop.test(successes,totals,conf.level=0.9,correct=FALSE)

2-sample test for equality of proportions without
continuity correction

data:  successes out of totals
X-squared = 2.7702, df = 1, p-value = 0.09603
alternative hypothesis: two.sided
90 percent confidence interval:
 -0.0443667102 -0.0004020218
sample estimates:
    prop 1     prop 2
0.03526971 0.05765408
```

Note that both according to the P-value ($0.096 < 0.10$ and according to the confidence interval ($[-0.0444 - 0.0004]$, which does not contain 0), we are justified in rejecting H_0.

4. **Problem 9.17, page 334.** Use the following data to test the hypothesis that a horse's chances of winning are unaffected by its position on the starting lineup. The data give the starting position of each of 144 winners, where position 1 is closest to the inside rail of the race track.

Starting Position	1	2	3	4	5	6	7	8
Number of Wins	29	19	18	25	17	10	15	11

State the hypotheses and perform a test at $\alpha = 0.05$.

Solutions: We use the Chi-squared test for this problem. The null-hypothesis is that all the starting positions have the same winning probabilities: $H_0 : p_1 = \cdots = p_8 = 1/8$ and the alternative is that they are not all equal. Using R:

```
> wins=c(29,19,18,25,17,10,15,11)
> probabilities=c(1,1,1,1,1,1,1,1)/8
> chisq.test(wins,p=probabilities)

Chi-squared test for given probabilities

data:  wins
X-squared = 16.3333, df = 7, p-value = 0.02224
```

The P-value obtained is 0.022. Since this is smaller than 0.05 we conclude that we may reject, at this significance level, the assumption that the starting position does not affect the outcome of the horse-race.