Ma 322: Biostatistics
Homework Assignment 8

Prof. Wickerhauser

Read Chapter 14, “Hypothesis Testing,” pages 240–262 of our text.

1. Following are 14 samples from a normal population with unknown mean and unknown standard deviation:

```
2.68 5.11 3.66 0.87 4.34 4.12 3.79 2.05 2.59 2.56 4.10 1.99 3.46 1.97
```

(a) Estimate the mean μ, the standard deviation σ, and the variance σ^2 from this sample.

(b) Test the hypothesis $H_0 : \mu = 3.0$, using the significance level $\alpha = 0.05$.

(c) Test the hypothesis $H_0 : \mu \leq 2.5$, using the significance level $\alpha = 0.05$.

Solution: Read the data by copy-pasting the line of numbers after this command:

```
x<-scan()
```

(a) Use the following R code:

```
mean(x); sd(x); var(x);
```

That yields the values $\mu \approx 3.092143$, $\sigma \approx 1.171752$, and $\sigma^2 \approx 1.373003$.

(b) Use the following R code:

```
t.test(x, mu=2.0)
```

That yields the following output:

```
data:  x
t = 0.29423, df = 13, p-value = 0.7732
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.415593 3.768692
sample estimates:
mean of x
 3.092143
```

Since the p-value is greater than 0.05, **Do not reject** H_0.

(c) Use the following `R` code:

```
t.test(x,mu=2.5,alternative="greater")
```

That performs the one-sided test with $H_A : \mu > 2.5$ and yields the output

```
data:  x
t = 1.8908, df = 13, p-value = 0.04057
alternative hypothesis: true mean is greater than 2.5
95 percent confidence interval:
 2.53755     Inf
sample estimates:
mean of x
 3.092143
```

Since the p-value is less than 0.05, **Reject** H_0. □

2. Using the sample standard deviation from Exercise 1 and a significance level of $\alpha = 0.05$, determine:

   (a) The power $1 - \beta$ of the t-test to reject the two-sided null hypothesis on the mean in Exercise 1b when there is a true difference $\delta = 0.5$.

   (b) The power $1 - \beta$ of the t-test to reject the one-sided null hypothesis on the mean in Exercise 1c when there is a true difference $\delta = 0.5$.

   (c) The number of samples needed to get a power $1 - \beta = 99\%$ in the t-test of the two-sided null hypothesis on the mean in Exercise 1b when there is a true difference $\delta = 0.5$.

   (d) The number of samples needed to get a power $1 - \beta = 99\%$ in the t-test of the one-sided null hypothesis on the mean in Exercise 1c when there is a true difference $\delta = 0.5$.

Solution: First enter the data as in Q.1 so that `sd(x)` gives the standard deviation and `length(x)` gives n. Then use `power.t.test()`:

(a)

```
power.t.test(n=length(x), sd=sd(x), delta = 0.5, sig.level = 0.05,
  power = NULL, strict=TRUE, type="one.sample", alternative="two.sided");
```

We get the result `power = 0.3157387`.

(b)

```
power.t.test(n=length(x), sd=sd(x), delta = 0.5, sig.level = 0.05,
  power = NULL, strict=TRUE, type="one.sample", alternative="one.sided");
```

We get the result `power = 0.4475606`.

(c)

```
power.t.test(n=NULL, sd=sd(x), delta = 0.5, sig.level = 0.05, power = 0.99,
  strict=TRUE, type="one.sample", alternative="two.sided");
```

We get the result `n = 102.8482`, so we would take 103 samples.

(d)

```
power.t.test(n=NULL, sd=sd(x), delta = 0.5, sig.level = 0.05, power = 0.99,
   strict=TRUE, type="one.sample", alternative="one.sided");
```

We get the result `n = 87.98652`, so we would take 88 samples. □

3. (a) Using the following data, and assuming that both populations are normal with equal variance, test the null hypothesis that male and female turtles have the same mean serum cholesterol concentrations.

Serum cholesterol (mg/100 ml) of turtles.

Male	248,329,223,313,271,324,255,255,423,332,311,264
Female	341,311,362,371,419,366,246,273,312,331

(b) The following data were found in Table 1 of C. M. Holcomb, C. G. Jackson, Jr., and M. M. Jackson, "Serum Cholesterol Values in Three Species of Turtles," J. Wildlife Diseases 8(1972), pp.181–182. `<www.jwildlifedis.org/cgi/reprint/8/2/181.pdf>`

Serum cholesterol (mg/100 ml) in turtles.

Species	n	Mean	S.E.	Range	Coef. of Var.
C. scripta	8	290.0	±42.3	174–512	41.2%
T. carolina	31	339.7	±15.6	178–511	25.6%

Assuming that both populations are normal with equal variance, test the alternative hypothesis that *T. carolina* has higher mean serum cholesterol concentrations than *C. scripta*.

Solution: (a) Denote the male and female turtle serum cholesterol population means by μ_m and μ_f. The hypotheses are: $H_0 : \mu_m = \mu_f$ versus $H_A : \mu_m \neq \mu_f$. Test these with a two-tailed two-sample t-test. The R commands are:

```
male <-c(248,329,223,313,271,324,255,255,423,332,311,264);
female<-c(341,311,362,371,419,366,246,273,312,331);
t.test(male,female, var.equal=TRUE);
```

The default specification for H_A is `alternative="two.sided"` and need not be invoked in this case. However, we must specify the non-default assumption of equal population variances. The p value is 0.1113, so **do not reject** the null hypothesis. NOTE: for this experiment, assuming unequal variances and thus using Welch's approximation gives a p value of 0.1092 from `t.test(male,female)`, so again we would not reject H_0.

(b) Denote the two species by subscripts of 1 (for *C. scripta*) and 2 (for *T. carolina*). The hypotheses are: $H_0 : \mu_1 \geq \mu_2$ versus $H_A : \mu_1 < \mu_2$.

This is a one-tailed, two-sample t-test from reduced data. Prepare the test statistic from the given data as follows. Compute the two sample standard deviations from the published standard errors:

$$s_1 = SE_1\sqrt{n_1} = 119.6; \qquad s_2 = SE_2\sqrt{n_2} = 86.86.$$

The homoscedasticity assumption allows us to compute the pooled variance using these sample standard deviations:

$$s_p^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2} = 8825,$$

where $\nu_1 = n_1 - 1 = 7$ and $\nu_2 = n_2 - 1 = 30$. We now use Welch's approximation to compute the variance of the difference of the means:

$$s^2_{\bar{X}_1 - \bar{X}_2} = s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 1388.$$

Thus $s_{\bar{X}_1-\bar{X}_2} = \sqrt{s^2_{\bar{X}_1-\bar{X}_2}} = 37.25$. From this and the two means, we form the difference statistic:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1-\bar{X}_2}} = -1.33.$$

This negative t-statistic allows us to use the area under the lower tail as the one-tailed probability of that value or worse, given H_0. There are $\nu = \nu_1 + \nu_2 = n_1 - 1 + n_2 - 1 = 37$ total degrees of freedom, so the p-value for a one-tailed test of these hypotheses is given by the cdf `pt(t, df=37)`≈ 0.095. Using a significance level $\alpha = 0.05$, we **do not reject** the null hypothesis.

The R commands are:

```
n1<-8; n2<-31; nu1<-n1-1; nu2<-n2-1; nu<-nu1+nu2;
SE1 <- 42.3; SE2<-15.6; s1 <- SE1*sqrt(n1); s2<-SE2*sqrt(n2);
s2p <- (nu1*s1**2+nu2*s2**2)/nu;  s1; s2; s2p;
s2xbar<-s2p*(1/n1+1/n2);   s2xbar; sxbar<-sqrt(s2xbar);  sxbar;
m1<-290.0; m2<-339.7; t<-(m1-m2)/sxbar; t; p<- pt( t, df=nu );  p;
```

Alternatively, generate fake data with the same means and variances with the program `faker()` on the class website, then use `t.test()`:

```
source("faker.R")     # or cut/paste the function into this console
SE1 <- 42.3; n1 <-  8; s1 <- SE1*sqrt(n1); m1<-290.0;
SE2 <- 15.6; n2 <- 31; s2 <- SE2*sqrt(n2); m2<-339.7;
x1 <- faker(n1, mu=m1, sd=s1);  # fake C.scripta data
x2 <- faker(n2, mu=m2, sd=s2);  # fake T.carolina data
t.test(x1, x2, var.equal=TRUE, alt="less")
```

□

4. For a fair coin, expect a binomial distribution with "heads" probability $p = 1/2$. A certain guilder coin is tossed 2000 times and comes up heads just 962 times.

   (a) Rosencrantz does not believe that this guilder is a fair coin. Use the experimental data and a significance threshold of $\alpha = 0.05$ to test Rosencrantz's one-sided hypothesis H_A: heads are less likely than tails in a toss of that coin.

   (b) Guildenstern does not share Rosencrantz's suspicions about the coin. Use the experimental data and a significance threshold of $\alpha = 0.05$ to test Guildenstern's two-sided hypothesis H_0: heads and tails are equally likely in a toss of that coin.

   Solution: (a) Rosencrantz's hypothesis test is one-tailed: $H_0 : p \geq 1/2$ versus $H_A : p < 1/2$. The `R` command to perform this test is:

```
binom.test(x=962,n=2000,p=1/2,alternative="less");
```

This yields `p-value = 0.04675`. With significance threshold of $\alpha = 0.01$, we therefore **reject** the null hypothesis in favor of Rosencrantz's alternative that heads are less likely than tails in a toss of the guilder coin.

(b) Guildenstern's hypothesis test is two-tailed: $H_0 : p = 1/2$ versus $H_A : p \neq 1/2$. Use this `R` command to perform this test (the input `alternative="two.sided"` may be omitted as it is the default):

```
binom.test(x=962,n=2000,p=1/2,alternative="two.sided")
```

This yields `p-value = 0.09351`. With significance threshold of $\alpha = 0.05$, we therefore **do not reject** the null hypothesis that heads and tails are equally likely in a toss of the guilder coin. □

5. (a) Using the data for Problem 3, part a, test the null hypothesis that male and female turtles have the same serum cholesterol variance.

(b) Using the data for Problem 3, part b, test the alternative hypothesis that *C. scripta* has a higher serum cholesterol variance than *T. carolina*.

Solution: (a) This is a two-tailed test of the hypotheses H_0: male and female turtles have the same serum cholesterol variance, versus H_A: male and female turtles have different serum cholesterol variance. It is performed with an F-test of the variance ratio, all done by the following `R` commands:

```
male <-c(248,329,223,313,271,324,255,255,423,332,311,264);
female<-c(341,311,362,371,419,366,246,273,312,331);
var.test(male,female)
```

This returns `p-value = 0.8392`, so with a significance threshold of $\alpha = 0.05$ we certainly **do not reject** H_0.

NOTE: the default in `var.test()` is to test the two-sided alternative hypothesis $H_A : \sigma_1 \neq \sigma_2$. For a one-sided H_A such as $H_A : \sigma_1 > \sigma_2$, we would call `var.test( male, female, alternative="greater")`.

(b) This is a one-tailed test of of the hypotheses H_0: $\sigma_1 \leq \sigma_2$ (*C. scripta* has no higher serum cholesterol variance than *T. carolina*), versus H_A: $\sigma_1 > \sigma_2$ (*C. scripta* has a higher serum cholesterol variance than *T. carolina*).

To perform the test, we must construct the F statistic from the reduced data. So, first recover the two sample standard deviations from the published standard errors:

$$s_1 = SE_1\sqrt{n_1} = 119.6; \qquad s_2 = SE_2\sqrt{n_2} = 86.86.$$

Note that $s_1 > s_2$, which is the ordering consistent with H_A, so compute the F statistic with the species 1 quantities in the numerator:

$$F = \frac{s_1^2}{s_2^2} = 1.897.$$

The numerator degrees of freedom are $\nu_1 = n_1 - 1 = 7$, and the denominator degrees of freedom are $\nu_2 = n_2 - 1 = 30$. The p value for the variance ratio is computed by `pf(F,nu1,nu2)`≈ 0.105. Since this is greater than the significance threshold $\alpha = 0.05$, **do not reject** the null hypothesis.

The R commands are:

```
SE1 <- 42.3; SE2<-15.6; n1<-8; n2<-31; nu1<-n1-1; nu2<-n2-1;
s1s <- n1*SE1**2; s2s<-n2*SE2**2; F <- s1s/s2s;   s1s; s2s; F;
p<-pf(F, df1=nu1, df2=nu2, lower.tail=FALSE); p
```

NOTE: We must specify the non-default `lower.tail=FALSE` when placing the larger variance in the numerator of the F ratio to perform the one-tailed test H_A: numerator $>$ denominator.

NOTE: Perform the two-tailed test, H_A: numerator $\neq$ denominator, by using two calls to `pf()`, one with F and one with $1/F$ as the first argument.

Alternatively, generate fake data with the same variances with the program `faker()` on the class website, then use `var.test()`:

```
source("faker.R")       # or cut/paste the function into this console
SE1 <- 42.3; n1 <-  8; s1 <- SE1*sqrt(n1); s1 # C.scripta, smaller
SE2 <- 15.6; n2 <- 31; s2 <- SE2*sqrt(n2); s2 # T.carolina, bigger
var.test(faker(n1,sd=s1),faker(n2,sd=s2), alt="greater")
```

NOTE: the population means are not used by `var.test()`; they are set to 0 by default in the fake data when `faker(n,mu,sd)` is called without specifying `mu`.

□

6. (a) Test the hypothesis that nucleotides a,c,g,t are equally likely in the GenBank sequence `NM_005369`, using the χ^2 goodness-of-fit method. Use significance level $\alpha = 0.01$.

   (b) Test the hypothesis that nucleotides a,c,g,t are equally likely in the GenBank sequence `NM_005367`, using the χ^2 goodness-of-fit method. Use significance level $\alpha = 0.01$.

   Solution: Use the following `R` code to install the APE package:

```
install.packages("ape");    require(ape);
```

The command `install.packages()` and subsequent commands `read.GenBank()` must be executed on an internet-connected computer.

(a) Obtain sequence `"NM_005369"` and perform the test:

```
ref<-c("NM_005369"); data<-read.GenBank(ref);
counts<-base.freq(data,freq=TRUE); counts; chisq.test(counts);
```

This finds counts of 1255 As, 660 Cs, 822 Gs, and 1127 Ts, yielding `X-squared = 231.69, df = 3, p-value < 2.2e-16`, indicating that we should definitely **reject** the null hypothesis H_0: all 4 nucleotides are equally likely, in favor of H_A: some of the nucleotides are more likely than others.

(b) Obtain sequence `"NM_005367"` and perform the test with slightly different function calls:

```
ref<-c("NM_005367"); data<-read.GenBank(ref, as.character=TRUE);
table(data); chisq.test(table(data));
```

This finds counts of 419 As, 403 Cs, 451 Gs, and 413 Ts, yielding `X-squared = 3.0629, df = 3, p-value = 0.382`, so we **do not reject** the null hypothesis H_0: all 4 nucleotides are equally likely. □

7. (a) How many 2×2 contingency tables are there with row sums $(2, 5)$ and column sums $(3, 4)$? (Hint: Write down all the solutions.)

(b) Assuming that the rows and columns are independent, compute the exact hypergeometric probability of each 2×2 contingency table in part a.

Solution: (a) This is easily done by hand, parametrizing by the value in the 1,1 location which is the intersection of the smallest-sum row and smallest-sum column. It can only take the values 0,1, or 2:

$$\begin{pmatrix} [0] & 2 \\ 3 & 2 \end{pmatrix}; \quad \begin{pmatrix} [1] & 1 \\ 2 & 3 \end{pmatrix}; \quad \begin{pmatrix} [2] & 0 \\ 1 & 4 \end{pmatrix}.$$

Thus, there are exactly 3 such tables.

(b) With row sums r_1, r_2, column sums c_1, c_2, and total sum $n = r_1 + r_2 = c_1 + c_2$, under the null hypothesis that the row and column variables are independent, a table containing frequencies $f_{11}, f_{12}, f_{21}, f_{22}$ satisfying

$$r_1 = f_{11} + f_{12}; r_2 = f_{21} + f_{22}; c_1 = f_{11} + f_{21}; c_2 = f_{12} + f_{22}; n = f_{11} + f_{12} + f_{21} + f_{22},$$

will have $p = p(f|r,c)$ given by the hypergeometric probability

$$\frac{r_1!r_2!c_1!c_1!}{n!f_{11}!f_{12}!f_{21}!f_{22}!} = \frac{\binom{n}{f_{11}, f_{12}, f_{21}, f_{22}}}{\binom{n}{r_1, r_2}\binom{n}{c_1, c_2}} = \frac{\binom{r_1}{f_{11}}\binom{r_2}{f_{21}}}{\binom{n}{c_1}} = \frac{\binom{c_1}{f_{11}}\binom{c_2}{f_{12}}}{\binom{n}{r_1}}$$

Hence the three tables in part a have probabilities

$$p[0] = \frac{2!5!3!4!}{7!0!2!3!2!} = \frac{2}{7}; \quad p[1] = \frac{2!5!3!4!}{7!1!1!2!3!} = \frac{4}{7}; \quad p[2] = \frac{2!5!3!4!}{7!2!0!1!4!} = \frac{1}{7}.$$

These may be computed with the following R codes:

```
r<-c(2,5); c<-c(3,4); n<-sum(r);
prod(factorial(c(r,c)))/prod(factorial(c(n, 0,2,3,2)));
prod(factorial(c(r,c)))/prod(factorial(c(n, 1,1,2,3)));
prod(factorial(c(r,c)))/prod(factorial(c(n, 2,0,1,4)));
```

Multiply the output decimals by 7 to get the exact rational number results. Equivalently, use the dhyper() function with all three possible values of f_{11} as the first argument:

```
r<-c(2,5); c<-c(3,4); dhyper(c(0,1,2), r[1],r[2],c[1]);
```

That procedure, however, does not generalize beyond 2 × 2 contingency tables. □

8. The following data are frequencies of bats found with and without rabies in two different geographic areas:

Area	With rabies	Without rabies
E	11	112
W	18	139

(a) Using the Yates-corrected χ^2 test at the $\alpha = 0.05$ significance level, test H_0: the incidence of rabies is the same in both areas.

(b) Use the Fisher exact test at the 0.05 level to test if the E population bats are less likely to have rabies than those in the W population.

Solution: (a) Use the following R commands to compute the results.

```
data<-c(11,112,18,139);
table<-matrix(data,nrow=2,byrow=TRUE); table
chisq.test(table);
```

That yields Pearson's Chi-squared test with Yates' continuity correction, with $\chi^2 = 0.23985$, df=1, and `p-value = 0.6243`. Hence **do not reject** H_0.

(b) Enter the same table as for part a, then issue the `R` command

```
fisher.test(table, alternative="less")
```

That performs a one-tailed Fisher Exact Test of H_A: E with-rabies percentage is less than the W with-rabies percentage. It yields `p-value = 0.3142`, which fails to meet the $\alpha = 0.05$ significance level. Hence **do not reject** H_0. □

9. A follow-on study was performed on the same bats data, similar to that of Problem 8 but with the additional tabulation of gender:

	With rabies		Without rabies	
Area	Male	Female	Male	Female
E	6	5	49	63
W	14	4	84	55

(a) Test for mutual independence at the $\alpha = 0.05$ significance level.

(b) Test for partial independence at the $\alpha = 0.05$ significance level.

Solution: Let area E/W be the rows (rA) index, gender Male/Female be the columns (cG) index, and Rabies/No rabies be the tiers (tR) index. Then $rA = cG = tR = 2$ and we may fill out the three-dimensional contingency table with the following `R` commands:

```
count<-c( 6,5,49,63,14,4,84,55);
# Create factor tags
area <- factor(gl(2,4,labels=c("E","W"))); area
rabies <- factor(gl(2,2,8,labels=c("With","Without")));  rabies
gender <- factor(gl(2,1,8,labels=c("Male","Female")));  gender
```

Make sure the variables are correctly labeled by examining the various cross-tabulations of the counts as follows:

```
xtabs(count  ~  area+gender+rabies)  # all three factors
xtabs(count  ~  gender+rabies)       # sum over ‘area’
xtabs(count  ~  area+rabies)         # sum over ‘gender’
xtabs(count  ~  area+gender)         # sum over ‘rabies’
xtabs(count  ~  area)          # sum over ‘gender’ and ‘rabies’
xtabs(count  ~  gender)        # sum over ‘area’ and ‘rabies’
xtabs(count  ~  rabies)        # sum over ‘area’ and ‘gender’
```

(a) Perform a 3-way test of factor independence as follows:

```
summary(xtabs(count  ~  area+rabies+gender))
```

That yields `Chisq=12.078, df=4, p-value=0.01678`, so we **reject** H_0 at the $\alpha = 0.05$ level in favor of H_A: the three factors are not independent.

(b) There are three tests to perform, one for each pair of factors chosen from the three: Partial independence: rA versus cG,tR:

```
summary(xtabs(count  ~  gender+rabies))
```

That yields `Chisq = 2.6776, df = 1, p-value = 0.1018`, so we **do not reject** H_0: factors Gender and Rabies are independent at the $\alpha = 0.05$ level.

Partial independence: tR versus rA,cG

```
summary(xtabs(count  ~  area+gender))
```

That yields `Chisq=8.723, df=1, p-value=0.003143`, so we **reject** H_0 at the $\alpha = 0.05$ level in favor of H_A: factors Area and Gender are not independent.

Partial independence: cG versus rA,tR

```
summary(xtabs(count  ~  area+rabies))
```

That yields `Chisq=0.4724, df=1, p-value=0.4919`, so we **do not reject** H_0: factors Area and Rabies are independent at the $\alpha = 0.05$ level. □