Stat 430 Homework 3

due: November 6, 2009 at 5pm

1. Access the Shakespeare word usage data. Some researchers have claimed they can identify gender of modern writers from as few as 300 words. Rather than focus on the obvious words (e.g. "princess" or "fight"), which can depend on the topic of the article, the researchers focus on seemingly unsubstantial words. In particular, men are supposed to prefer words communicating quantities, like "more" and "some," while women are supposed to prefer personal pronouns and possessives, like "you" and "your." Shakespeare was a man, by all accounts, but he wrote the words of many female characters. You will test if he was able to recreate these subtle differences in word usage of males vs. females.

   (a) Use the data to test if there is a difference in Shakespeare's gender word usage for the more active words [love (v), hate (v), love (n), kill (v)].

   Solution: The null hypothesis is independence between gender and words. We use a chi-square test of independence, finding no evidence to reject the independence assumption. The R code and output are below.

   R Code:

```
d <- read.table("http://thirteen-01.stat.iastate.edu/wiki/stat430/files?filename=shakespeare.txt",
   header=T, row.names=1)
d.a <- d[1:4,]
r.sums <- rowSums(d.a)
c.sums <- colSums(d.a)
n <- sum(c.sums)
e <- r.sums%*%t(c.sums)/n
cat("\nCell Contributions:\n")
print((d.a-e)^2/e)
X.2 <- sum((d.a - e)^2/e)
cat("chi-square statistic:", X.2, "\n")
df <- (length(r.sums)-1)*(length(c.sums)-1)
cat("p-value (df = ", df, "): ", 1-pchisq(X.2, df=df), "\n", sep="")
```

   R Output:

```
Cell Contributions:
              Male     Female
love(v) 0.27155205 0.81283366
hate(v) 0.01320927 0.03953914
love(n) 0.39582624 1.18482218
kill(v) 0.17327782 0.51867053
chi-square statistic: 3.409731
p-value (df = 3): 0.3326598
```

   (b) Test if there is a gender difference in the usage of the words "more" and "some." Repeat for "you" and "yours."

   Solution: This question caused confusion for good reason. It doesn't say what I meant. Regardless, here is an answer to what it asked.

The null hypothesis is, again, independence of gender and word. Both these tests can be done with a Fisher's exact test because the table is 2-by-2, but the counts are so high that the asymptotic chi-square distribution of the chi-square test will be satisfied. We find no significant difference in the usage of words more and some, but there is significant difference in the usage of you and your ("yours" with the "s" was a typo).

R Code:

```
d.ms <- d[c("more","some"),]
d.yy <- d[c("you","your"),]
for(d in list(d.ms, d.yy)) {
   print(d)
   r.sums <- rowSums(d)
   c.sums <- colSums(d)
   n <- sum(c.sums)
   e <- r.sums%*%t(c.sums)/n
   cat("\nCell Contributions:\n")
   print((d-e)^2/e)
   X.2 <- sum((d - e)^2/e)
   cat("chi-square statistic:", X.2, "\n")
   df <- (length(r.sums)-1)*(length(c.sums)-1)
   cat("p-value (df = ", df, "): ", 1-pchisq(X.2, df=df), "\n", sep="")
}
```

R Output:

```
Cell Contributions:
          Male    Female
more 0.1503590 0.7095215
some 0.2682143 1.2656628
chi-square statistic: 2.393758
p-value (df = 1): 0.1218205

Cell Contributions:
          Male   Female
you  0.7540464 2.926934
your 1.6417569 6.372703
chi-square statistic: 11.69544
p-value (df = 1): 0.0006265344
```

(c) You may combine words that do not have usage differences from part b. Now choose one or a group of words to "represent" each gender. Is there a difference in Shakespeare's gender-based use of these words (or sets of words)?

Solution: This question continues with the confusion of the second part, but ignoring the reference to part b, here is an answer.

There is no need to merge words, let's just use all the subtle words to test for independence of usage and gender. We find significant differences in usage between the genders. Considering that 18.31 is the critical value for the chi-square with 10 degrees of freedom, we can look at the cell contributions (signed in this output) and identify words that clearly contribute significant amounts. (Notice, other words may have significant differences, because 18.31 is not a cell-by-cell critical value. The words "you," "the," and "not" are the most variable. "You" and "not" are overused by females, while "the" is overused by males.

What does it mean? No clue, but according to The Gender Genie, "the" is a male word, while "you" and "not" are indeed female words. One explanation given in the linked cites is that men prefer to discuss things, while women are comfortable discussing personal relationships.

What is more interesting is that Shakespeare seemed to be able to capture these differences. On the other hand, Shakespeares men did not use "more" more (strike that!) often than women, which they are supposed to, and they use "with" more often than females, when they are not supposed to. So, who knows if Shakespeare was really capturing subtle word differences. In addition, of course, he lived during another time when language usage was different.

R **Code:**

```
d.a <- d[-c(1:4),]
print(d.a)
r.sums <- rowSums(d.a)
c.sums <- colSums(d.a)
n <- sum(c.sums)
e <- r.sums%*%t(c.sums)/n
cat("\nCell Contributions:\n")
sign <- sign(d.a-e)
print(sign*(d.a-e)^2/e)
X.2 <- sum((d.a - e)^2/e)
cat("chi-square statistic:", X.2, "\n")
df <- (length(r.sums)-1)*(length(c.sums)-1)
cat("p-value (df = ", df, "): ", 1-pchisq(X.2, df=df), "\n", sep="")
```

R **Output:**

```
Cell Contributions:
             Male       Female
more   -0.09235172    0.4310164
you   -21.41712278   99.9562577
she    -0.24243725    1.1314835
that   -1.21346297    5.6633759
```

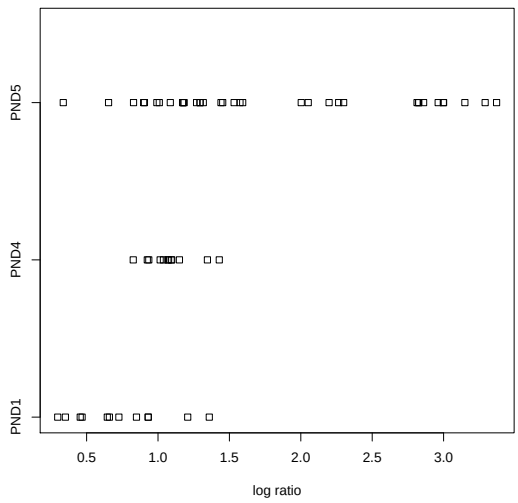
```
some    0.33810812   -1.5779908
these   0.85228315   -3.9777068
with    0.29270768   -1.3661016
your   -1.68756395    7.8760615
the    29.98405151 -139.9391324
for     0.16112583   -0.7519934
not    -5.67933610   26.5061366
chi-square statistic: 351.1378
p-value (df = 10): 0
```

2. Access the virus data. To test virus fitness (as in, Darwin's survival of the fittest), live virus are added to cell cultures and the amount of virus is measured after a fixed period of time. Since growth conditions may vary tremendously, the virus is added with a standard virus and the log ratio of the numbers of the viruses is reported (a kind of paired sample, where the number reported is a measure of the difference between the members of pair). In this question, you will use ANOVA to detect if there is a fitness difference in three different EIA viruses, HIV-like viruses that infect horses.

   (a) There are three named viruses in the file, PND1, PND4, and PND5, that differ in their envelop protein (the protein that is responsible for binding and infecting target cells in the host and the part of the virus most susceptible to attack by the immune system). Do they differ in fitness as measured by a significant different log ratio relative to the standard virus?

Solution: First, we plot the data using

`stripchart(V3 ~ V1, data=d, method="stack", ylab="", xlab="log ratio").`



It seems there may be differences across the three conditions, and some things that worry us, but we'll discuss that later.

We will use the `aov()` function in `R` to perform the ANOVA. The ANOVA table is show below in the `R` output. The between sum-of-squares is $SS_B = 12.2858$, with mean-square, and estimate of σ^2 under H_0, 6.14. The within sum-of-squares is $SS_W = 28.0283$, with mean-square, and estimate of σ^2, 0.49. Under an alternative hypothesis with different means, the first estimate *overestimates* σ^2, and we can see there must be some difference in means. The F statistic is $\frac{6.14}{0.49} = 12.493$, which `R` indicates is highly significant evidence of a difference in means.

The coefficients indicate that PND1 has the lowest mean 0.74, PND4 has intermediate fitness with mean 1.08, and PND5 most fit at 1.81. Here (Intercept) is $\hat{\mu} + \alpha_1$, PND4 is α_2, and PND5 is α_3. `R`, by default, constrains $\alpha_1 = 0$, which is different from the constraint $\sum \alpha_i = 0$ we used in class.

We test all pairwise differences using the Tukey method (`TukeyHSD()`), finding that PND5 is significantly different from the other two, but PND1 and PND4 are not significantly different.

`R` Code:

```
d <- read.table("virus.txt", header=F)
m <- aov(V3 ~ V1, data=d)
summary(m)
coefficients(m)
TukeyHSD(m)
```

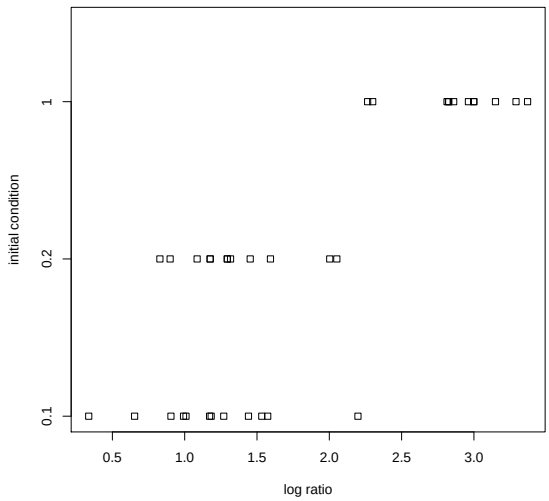
`R` Output:

```
             Df  Sum Sq Mean Sq F value    Pr(>F)
V1            2 12.2858  6.1429  12.493 3.169e-05 ***
Residuals    57 28.0283  0.4917
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
      (Intercept)           PND4           PND5
        0.7401250      0.3428583      1.0682028

$V1
               diff        lwr      upr     p adj
PND4-PND1 0.3428583 -0.3460415 1.031758 0.4594486
PND5-PND1 1.0682028  0.5057184 1.630687 0.0000782
PND5-PND4 0.7253444  0.1628601 1.287829 0.0082490
```

(b) You will notice there is substantially more data for virus PND5. It is because that virus was tested with three different initial conditions, indicated in the second column. Do these initial conditions affect the measurement of viral fitness?

Solution: As for part one, we examine the data first.



This time again, the F test is highly significant. Tukey's method reveals that initial condition 1 is significantly different from the other two. We conclude the initial conditions, particularly 1, *do* impact the results.

R Code:

```
d.5 <- subset(d, V1=="PND5")
stripchart(V3 ~ V2, data=d.5, method="stack",
   ylab="initial condition", xlab="log ratio")

m.5 <- aov(V3 ~ as.factor(V2), data=d.5)
summary(m.5)
coefficients(m.5)
TukeyHSD(m.5)
```

R Output:

```
              Df  Sum Sq Mean Sq F value    Pr(>F)
as.factor(V2)  2 21.1468 10.5734  65.052 3.547e-12 ***
Residuals     33  5.3638  0.1625
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
     (Intercept) as.factor(V2)0.2   as.factor(V2)1
       1.1889083        0.1588333        1.6994250

$`as.factor(V2)`
             diff       lwr       upr     p adj
0.2-0.1 0.1588333 -0.245036 0.5627027 0.6037368
1-0.1   1.6994250  1.295556 2.1032943 0.0000000
```

```
1-0.2   1.5405917  1.136722 1.9444610 0.0000000
```

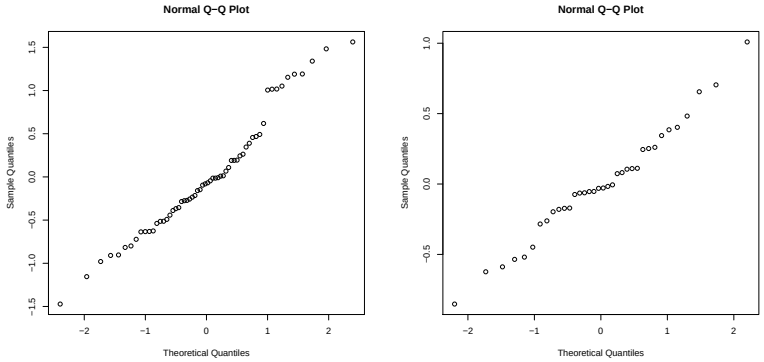
(c) Evaluate the ANOVA assumptions and whether they apply to this data.

Solution: Our first concern is evident in the plot of all the virus data. There seems to be a difference in the variance for the three levels. The variance estimates for each level are

```
Std. Dev. of PND1 : 0.3309178
Std. Dev. of PND4 : 0.1687181
Std. Dev. of PND5 : 0.8703129
```

and particularly PND5 seems to have higher variance (perhaps because of the multiple initial conditions). Indeed, the gaps in the PND5 data disappear when considering initial conditions.

The probability plots of the residuals do not show any obvious problems with the assumption of normality.



We test the common variance assumption using Bartlett's test.

```
bartlett.test(V3 ~ V1, data=d)
bartlett.test(V3 ~ V2, data=d.5)
```

with output

```
Bartlett test of homogeneity of variances

data:  V3 by V1
Bartlett's K-squared = 31.7219, df = 2, p-value = 1.293e-07


Bartlett test of homogeneity of variances

data:  V3 by V2
Bartlett's K-squared = 1.3869, df = 2, p-value = 0.4998
```

finding as suspected that the combined data does not have constant variance. There seems to be no problem within the PND5 data. Since the assumption

of common variance is important for unbalanced samples, the significant result raises a concern.

Bartlett's test is sensitive to normality assumptions. There is an alternative called Levene's test. We did not discuss it in class, but it is easily performed with another ANOVA on the absolute residuals. It continues to show the same problem: non-constant variance across the three viruses, PND1, PND4, and PND5. Results for the part a data:

```
summary(aov(abs(m$res) ~ d$V1))
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
d$V1         2 4.9121  2.4561  23.679 3.27e-08 ***
Residuals   57 5.9123  0.1037
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The last assumption is independence. This is an important assumption, but we have no idea how the data were collected (we don't even know if the numbers are ordered in any meaningful way, for example by experimental day), so we cannot evaluate this assumption.

3. [For practice; not to be turned in.] Repeat question 2 using non-parametric tests. Do you reach the same conclusions?

Solution: We can use the Kruskal-Wallis test for testing equivalent means without assuming normality. The output below shows we can reject the null hypothesis of no virus effects and no initial condition effects.

We can perform rank sum tests for individual pairs of means. The Bonferroni adjusted type I error rate is $\frac{0.05}{2\binom{3}{2}} \approx 0.008$. We find significant differences between all viruses, and between initial condition 1 and the other two. The difference between PND1 and PND4 is borderline.

```
kruskal.test(x=d$V3, g=d$V1)
kruskal.test(x=d$V3[d$V1=="PND5"], g=d$V2[d$V1=="PND5"])

Kruskal-Wallis rank sum test

data:  d$V3 and d$V1
Kruskal-Wallis chi-squared = 21.3938, df = 2, p-value = 2.261e-05
```

```
Kruskal-Wallis rank sum test

data:  d$V3[d$V1 == "PND5"] and d$V2[d$V1 == "PND5"]
Kruskal-Wallis chi-squared = 23.6892, df = 2, p-value = 7.177e-06

# Pairwise tests
wilcox.test(V3 ~ V1, data=d, subset=(V1=="PND1"|V1=="PND5"))
wilcox.test(V3 ~ V1, data=d, subset=(V1=="PND1"|V1=="PND4"))
wilcox.test(V3 ~ V1, data=d, subset=(V1=="PND4"|V1=="PND5"))
wilcox.test(V3 ~ V2, data=d, subset=(V1=="PND5"&V2=="1"|V2=="0.1"))
wilcox.test(V3 ~ V2, data=d, subset=(V1=="PND5"&V2=="1"|V2=="0.2"))
wilcox.test(V3 ~ V2, data=d, subset=(V1=="PND5"&V2=="0.1"|V2=="0.2"))

Wilcoxon rank sum test

data:  V3 by V1
W = 49, p-value = 1.768e-05
alternative hypothesis: true location shift is not equal to 0


Wilcoxon rank sum test

data:  V3 by V1
W = 26, p-value = 0.006812
alternative hypothesis: true location shift is not equal to 0


Wilcoxon rank sum test

data:  V3 by V1
W = 96, p-value = 0.003475
alternative hypothesis: true location shift is not equal to 0


Wilcoxon rank sum test

data:  V3 by V2
W = 0, p-value = 7.396e-07
alternative hypothesis: true location shift is not equal to 0


Wilcoxon rank sum test
```

```
data:  V3 by V2
W = 0, p-value = 7.396e-07
alternative hypothesis: true location shift is not equal to 0


Wilcoxon rank sum test

data:  V3 by V2
W = 57, p-value = 0.4095
alternative hypothesis: true location shift is not equal to 0
```