Advanced Data Analysis 1, Stat 427/527 Homework 7 Name: Answer Key
Fall 2014 Prof. Erik B. Erhardt

Part I. (105 points) Do all calculations in R. All R code for the assignment should be included with the part of the problem it addresses (for code and output use a fixed-width font, such as Courier). Code is used to calculate result. Text is used to report and interpret results. Do not report or interpret results in the code. Also:

1. Clearly define population parameters in each problem. That is, give a verbal description of what the population mean is in the context of the problem.
2. Clearly specify hypotheses when appropriate (not every problem involves a test of hypothesis).
3. Write a coherent conclusion based on each CI or test.

(10 pts) **1. ET:** A 1997 survey conducted by M.I.T. reported that 16% of the 1,014 adults surveyed would be willing to support tax hikes to find extra-terrestrials.

(a) (5 pts) Find a 95% CI for the proportion p of all adults that favor such a tax hike.

Solution: $1014 * 0.16 = 162$ adults of 1014 are willing. Both "successes" and "failures" exceed 5, so the normal approximation is appropriate here.
We are 95% confident that between 13.9% and 18.4% of adults support tax hikes to phone E.T.

```
# Approximate normal test for proportion, without Yates' continuity correction
p.summary <- prop.test(162, 1014, p = 0.2, correct = FALSE)
p.summary
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  162 out of 1014, null probability 0.2
## X-squared = 10.26, df = 1, p-value = 0.001359
## alternative hypothesis: true p is not equal to 0.2
## 95 percent confidence interval:
##  0.1385 0.1836
## sample estimates:
##      p
## 0.1598
```

(b) (5 pts) Suppose it was known that in 1990 that the proportion of all adults willing to support tax hikes to find extra-terrestrials was 0.2. Is there evidence that the proportion of adults in 1997 willing to spring for tax hikes for this purpose has changed since 1990? Carry out a test to answer this question. Use $\alpha = 0.05$.

Solution: We know from above that the CI does not include 0.2 (the p-value=0.001 also reflects this specific test). Therefore we reject the null hypothesis in favor of the alternative concluding that the proportion of adults is different in 1997 from 1990.

(15 pts) **2. Side effects:** Prof. Ed Bedrick was involved in a study that examined the extent of side effects from using amphetamines to treat children with traumatic brain injuries. Prior information suggests that the probability of major side effects is very small, so it is of interest to estimate the maximum plausible value that the probability of major side effects might be. In the study none of the 15 children had major side effects.

(a) (10 pts) Compute an exact upper 95% confidence bound for the probability of major side effects. Write a short conclusion to your analysis, interpreting the results of the exact bound in the context of the problem.

Solution:

```
# Exact binomial test for proportion
b.summary <- binom.test(0, 15, alternative="less")
b.summary
```

```
##
##  Exact binomial test
##
## data:  0 and 15
## number of successes = 0, number of trials = 15, p-value =
## 3.052e-05
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.000 0.181
## sample estimates:
## probability of success
##                      0
```

The exact CI is $(0, 0.181)$, thus we are 95% confident that the true probability of major side effect from using amphetamines is no greater than 0.181.

(b) (5 pts) What would your reponse be to someone asking you to compute the bound based on the normal distribution, and why?

Solution: The assumptions regarding the minimum number of successes and failures (at least 5 each) has not been met. Do not use the normal approximation using `prop.test()`.

(25 pts) **3. Suicides:** The National Center for Health Statistics (NCHS) gave the following data on the distribution of suicides in the U.S. by month in 1990. Is there any evidence that the suicide rate varies monthly, or are the data consistent with the hypothesis that the rate is constant?
To simplify your analysis, assume the months have the same numbers of days. Compare the observed proportions across months qualitatively and through a formal goodness-of-fit test. Write a short and coherent summary to this problem. Make sure to include relevant graphical summaries with your presentation. The code below will create a data.frame for you to get started.

```
# read data from space delimited text
suicide <- read.table(text="
  Month  Suicides
  01Jan      1867
  02Feb      1789
  03Mar      1944
  04Apr      2094
  05May      2097
  06Jun      1981
  07Jul      1887
  08Aug      2024
  09Sep      1928
  10Oct      2032
  11Nov      1978
  12Dec      1859
", header=TRUE)

# show the structure of the data.frame
str(suicide)
```

```
## 'data.frame': 12 obs. of  2 variables:
##  $ Month   : Factor w/ 12 levels "01Jan","02Feb",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Suicides: int  1867 1789 1944 2094 2097 1981 1887 2024 1928 2032 ...
```

Solution: We assume each month has the same number of days, and therefore the number of suicides is directly proportional to the proportion of suicides for each month. We test the hypothesis that the suicide rate for each month is the same, $H_0 : p_1 = \cdots = p_{12} = 1/12$, against the hypothesis that at least one month is different from the others, H_A : not H_0.

```
# calculate chi-square goodness-of-fit test
x.summary <- chisq.test(suicide$Suicides, correct = FALSE, p = rep(1/12,12))
# print result of test
x.summary
```

```
##
```

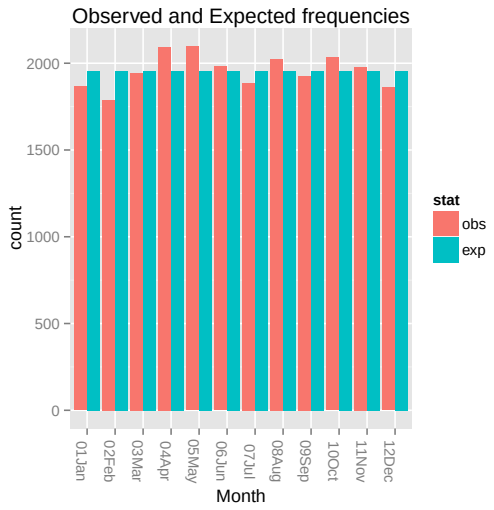
```
##  Chi-squared test for given probabilities
##
## data:  suicide$Suicides
## X-squared = 51.79, df = 11, p-value = 2.975e-07
```

```
# use output in x.summary and create table
x.table <- data.frame(Month = suicide$Month
                      , obs = x.summary$observed
                      , exp = x.summary$expected
                      , res = x.summary$residuals
                      , chisq = x.summary$residuals^2
                      , stdres = x.summary$stdres)
#x.table
```

	Month	obs	exp	res	chisq	stdres
1	01Jan	1867	1956.6666667	-2.0270880	4.1090857	-2.1172244
2	02Feb	1789	1956.6666667	-3.7904285	14.3673481	-3.9589734
3	03Mar	1944	1956.6666667	-0.2863544	0.0819989	-0.2990875
4	04Apr	2094	1956.6666667	3.1046850	9.6390687	3.2427377
5	05May	2097	1956.6666667	3.1725057	10.0647927	3.3135742
6	06Jun	1981	1956.6666667	0.5501019	0.3026122	0.5745627
7	07Jul	1887	1956.6666667	-1.5749494	2.4804656	-1.6449810
8	08Aug	2024	1956.6666667	1.5221999	2.3170926	1.5898860
9	09Sep	1928	1956.6666667	-0.6480653	0.4199886	-0.6768821
10	10Oct	2032	1956.6666667	1.7030553	2.9003975	1.7787833
11	11Nov	1978	1956.6666667	0.4822812	0.2325951	0.5037262
12	12Dec	1859	1956.6666667	-2.2079434	4.8750142	-2.3061217

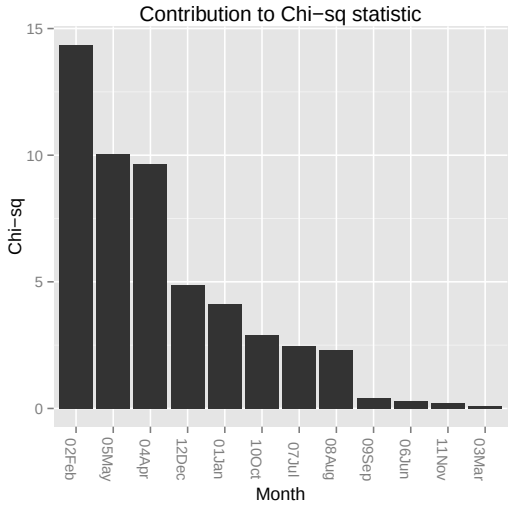
Plot observed vs expected values to help identify months that deviate the most.

```
library(reshape2)
x.table.obsexp <- melt(x.table,
              # id.vars: ID variables
              #   all variables to keep but not split apart on
              id.vars=c("Month"),
              # measure.vars: The source columns
              #   (if unspecified then all other variables are measure.vars)
              measure.vars = c("obs", "exp"),
              # variable.name: Name of the destination column identifying each
              #   original column that the measurement came from
              variable.name = "stat",
              # value.name: column name for values in table
              value.name = "value"
           )

# Observed vs Expected counts
library(ggplot2)
p <- ggplot(x.table.obsexp, aes(x = Month, fill = stat, weight=value))
p <- p + geom_bar(position="dodge")
p <- p + labs(title = "Observed and Expected frequencies")
p <- p + xlab("Month")
p <- p + theme(axis.text.x = element_text(angle=-90))
print(p)
```



Plot contribution to chi-square values to help identify age groups that deviate the most. The term "Contribution to Chi-Square" (`chisq`) refers to the values of $\frac{(O-E)^2}{E}$ for each category. χ^2_s is the sum of those contributions.

```
# Contribution to chi-sq
# pull out only the age and chisq columns
x.table.chisq <- x.table[, c("Month","chisq")]
# reorder the age categories to be descending relative to the chisq statistic
x.table.chisq$age <- with(x.table, reorder(Month, -chisq))

p <- ggplot(x.table.chisq, aes(x = age, weight = chisq))
p <- p + geom_bar()
p <- p + labs(title = "Contribution to Chi-sq statistic")
p <- p + xlab("Month")
p <- p + ylab("Chi-sq")
p <- p + theme(axis.text.x = element_text(angle=-90))
print(p)
```



The results of the goodness-of-fit test is below. The $\chi^2_s = 51.79$ has a p-value$= 2.975 \times 10^{-7}$, so we reject H_0 in favor of H_A concluding that not all months have the same suicide rate.
The larger "contribution to chi-squared" values indicate months that deviate most from the expected average month. February has fewer suicides than expected (with Dec and Jan having a little less than expected), while April and May each have more than expected.

(15pts) **4. Welsh and Breton:** The rising tide of national and regional loyalties around the world has bearing on the survival of minority or secondary languages. The article "Language Maintenance and Shift in a Breton and Welsh Sample," (Word, 1983; p. 67–88) describes one of the first comparative studies in this area. A random sample of 86 Welsh bilingual adults yielded 76 who spoke Welsh fluently, while another random sample of 77 bilingual adults from Brittany resulted in 57 who spoke Breton fluently. Both Welsh and Breton are southern Celtic in origin. Do these data support the conclusion that the true proportion of fluent speakers among Welsh bilingual adults differs from the corresponding proportion for Breton bilingual adults?
To answer this question, carry out an appropriate test at the 5% level, and quantify the difference in population proportions using a 95% CI.

Solution: The data are summarized as follows.

```
Bilinguals
Welsh     76 of 86
Brittany  57 or 77
```

We are interested in testing the hypothesis that the proportion of bilinguals from the Welsh and Breton populations is the same, $H_0 : p_W - p_B = 0$, against the alternative that the proportions are different, $H_A : p_W - p_B \neq 0$.

```
# Approximate normal test for two-proportions, without Yates' continuity correction
p.summary <- prop.test(c(76, 57), c(86, 77), correct = FALSE)
p.summary
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(76, 57) out of c(86, 77)
## X-squared = 5.568, df = 1, p-value = 0.0183
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.02437 0.26255
## sample estimates:
## prop 1 prop 2
## 0.8837 0.7403
```

The p-value= $0.0183 < 0.05$, therefore we reject H_0 in favor of H_A concluding that the proportions are different between these two populations. It appears that a greater proportion of Welsh are bilingual by between 0.0244 and 0.2626, with 95% confidence.

(40pts) **5. Hawaiian blood:** There are four major blood groups in humans: O, A, B, and AB. A sample of individuals with records at the Blood Bank of Hawaii was selected. Each individual was classified according to blood type and ethnicity. The following two-way table of counts was obtained.

```
# read data from space delimited text
    # skip=3 skips the blank line, "Ethnicity" line, and "-----" line
blood <- read.table(text="
                                Ethnicity
                ------------------------------------------------
 Blood_Type     Hawaiian   Hawaiian_White   Hawaiian_Chinese   White
     O            1903          4469             2206          53759
     A            2490          4671             2368          50008
     B             178           606              568          16252
    AB              99           236              243           5001
", header=TRUE, skip=3)

# reshape into matrix for chisq.test()
blood.matrix <- matrix(c(blood[,2], blood[,3], blood[,4], blood[,5]),
                  ncol = 4, byrow = FALSE,
                  dimnames = list("Blood_type" = c("O", "A", "B", "AB"),
   "Ethnicity" = c("Hawaiian", "Hawaiian_White", "Hawaiian_Chinese", "White")))
```

(a) (15 pts) Summarize these data, focusing on comparing the proportions or percents in the 4 blood categories across the 4 ethnic groups.
Create a plot of each population's proportions (y) over blood type (x) to help determine how the populations differ. (You can calculate proportions, reshape the data, then plot the proportions on the same axes or using facets.)

Solution:

```
# calculate the column sums to calculate the proportion blood type by ethnicity
blood.colsums <- matrix(rep(colSums(blood[, 2:5]), 4), nrow=4, byrow=TRUE)
blood.colsums
```

```
##      [,1] [,2] [,3]   [,4]
## [1,] 4670 9982 5385 125020
## [2,] 4670 9982 5385 125020
## [3,] 4670 9982 5385 125020
## [4,] 4670 9982 5385 125020
```

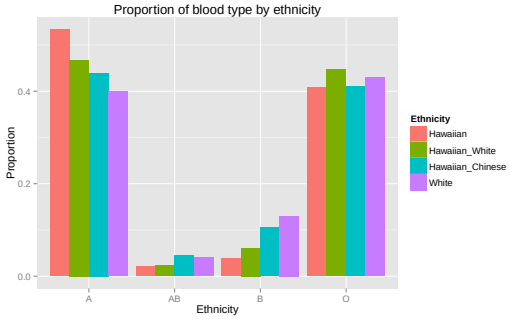
```
# proportion blood type
blood.prop <- blood[,2:5] / blood.colsums
blood.prop$Blood_Type <- blood$Blood_Type
blood.prop
```

```
##   Hawaiian Hawaiian_White Hawaiian_Chinese White Blood_Type
## 1  0.40749        0.44771          0.40966  0.43          O
## 2  0.53319        0.46794          0.43974  0.40          A
## 3  0.03812        0.06071          0.10548  0.13          B
## 4  0.02120        0.02364          0.04513  0.04         AB
```

```
library(reshape2)
blood.prop.long <- melt(blood.prop,
             id.vars=c("Blood_Type"),
             variable.name = "Ethnicity",
             value.name = "freq"
           )
blood.prop.long
```

```
##    Blood_Type        Ethnicity    freq
## 1           O         Hawaiian 0.40749
## 2           A         Hawaiian 0.53319
## 3           B         Hawaiian 0.03812
## 4          AB         Hawaiian 0.02120
## 5           O   Hawaiian_White 0.44771
## 6           A   Hawaiian_White 0.46794
## 7           B   Hawaiian_White 0.06071
## 8          AB   Hawaiian_White 0.02364
## 9           O Hawaiian_Chinese 0.40966
## 10          A Hawaiian_Chinese 0.43974
## 11          B Hawaiian_Chinese 0.10548
## 12         AB Hawaiian_Chinese 0.04513
## 13          O           White 0.43000
## 14          A           White 0.40000
## 15          B           White 0.13000
## 16         AB           White 0.04000
```

```
# Observed vs Expected counts
library(ggplot2)
p <- ggplot(blood.prop.long, aes(x = Blood_Type, fill = Ethnicity, weight = freq))
p <- p + geom_bar(position="dodge")
p <- p + labs(title = "Proportion of blood type by ethnicity")
p <- p + xlab("Ethnicity")
p <- p + ylab("Proportion")
print(p)
```



The most striking difference is that it appears that the Hawaiian population (the three groups with Hawaiian-[any]) has a higher proportion of A blood and a lower proportion of B than the white population. Secondly, AB blood is less common for Hawaiian and Hawaiian-White populations than for Hawaiian-Chinese and White.

(b) (15 pts) Is there evidence that blood type and ethnicity are associated in Hawaii? Explain.

Solution: We test the null hypothesis of no association between ethnicity and blood type against the alternative that there is an association (though we don't specify what that association is). Construct a chi-squared test.

```
chisq.summary <- chisq.test(blood.matrix, correct=FALSE)
chisq.summary
```

```
##
##  Pearson's Chi-squared test
##
## data:  blood.matrix
## X-squared = 1079, df = 9, p-value < 2.2e-16
```

```
# The sum of the squared residuals is the chi-squared statistic:
chisq.summary$residuals^2
```

```
##           Ethnicity
## Blood_type Hawaiian Hawaiian_White Hawaiian_Chinese    White
##          O     5.378          7.496            5.055  0.01994
##          A   171.445         80.419           11.265 33.19100
##          B   266.652        302.556           11.191 76.83042
##          AB   36.179         56.989            6.219  7.71767
```

Above, the p-value= $1.918 \times 10^{-226} < 0.05$, therefore we reject the null hypothesis in favor of the alternative that some association exists.

(c) (10 pts) Carry out any additional analyses that you deem relevant, and summarize your findings. For example, there are a number of possible additional analyses that could be done here.

- All six pairwise comparisons of blood type distributions between each pair of ethnicities.
- Comparisons between pairs suggested by ethnicities, for example: (White, Hawaiian), (White, Hawaiian-White), (Hawaiian, Hawaiian-White), and (Hawaiian, Hawaiian-Chinese).
- There are other sensible comparisons.

Then, in each of these, if you find significant differences between blood type proportions for a pair of ethnicities, you can continue to look at individual proportion comparisons (such as a two-sample proportion test of White vs Hawaiian blood type A).

You don't need to do tons of this. Do some amount of additional analysis to show that you understand the comparison principles. Then indicate (describe) some addition analyses that *could* be done.

Solution: Many solutions are possible. The primary conclusions are given above in part (b).