ASSIGNMENT 7

Brandon Lampe
STAT 527
Advanced Data Analysis I

November 19, 2014

1 ET:

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

```r
n <- 1014          # total number of adults surveyed
p.sup <- 0.16      # proportion of adults surveyed willing to support tax hike
alpha <- 1 - 0.95
z.crit <- qnorm(1 - alpha/2)
SE <- sqrt(p.sup*(1 - p.sup)/n)
L.CI <- p.sup - z.crit * SE
U.CI <- p.sup + z.crit * SE
L.CI # lower CI limit
## [1] 0.1374353
U.CI # upper CI limit
## [1] 0.1825647
```

The population paramter (p) is the proportion of adults willing to support tax hikes to find extra-terrestrials. The sample proportion ($\hat{p}$), of the adults surveyed ($n = 1,014$), was 16%. With 95% confidence, the proportion of adults willing to support tax hikes to find extra-terrestrials (p) is between 13.7% and 18.3%. This estimate of the CI uses the normal approximation.

1(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 $\alpha = 0.05$.

```r
prop.test(p.sup * n, n, p = 0.2, correct = FALSE) # test for equal proportions
##
##  1-sample proportions test without continuity correction
##
## data:  p.sup * n out of n, null probability 0.2
## X-squared = 10.14, df = 1, p-value = 0.001451
## alternative hypothesis: true p is not equal to 0.2
## 95 percent confidence interval:
##  0.1387246 0.1838418
## sample estimates:
##    p
## 0.16
```

The null hypothesis is that the proportion of adults willing to support tax hikes to find extral-terresetrials in 1997 (p) is equal to 0.2 (p_0), which is the proportion of adults supporting this tax in 1990. That is $H_0 : p = p_0$ and the alternative is $H_A : p \neq p_0$.

The test for equal proportions resulted in a p-value of 0.001, which is less than 0.05. Therefore, I reject the null hypothesis in favor of the alternative. The proportion of adults willing to support tax hikes to find extra-terrestrials in 1990 is not equal to the proportion of adults willing to support this tax hike in 1997. Additionally, the confidence interval does for is: $0.138 < p < 0.184$, which does not include p_0.

2 Side effects

2(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.

```
# confidence interval from a normal distribution
n <- 15       # total number of adults surveyed
p <- 0     # proportion of adults surveyed willing to support tax hike
alpha <- 1 - 0.95
z.crit <- qnorm(1 - alpha/2)
SE <- sqrt(p.sup*(1 - p.sup)/n)
U.CI <- p.sup + z.crit * SE
U.CI # upper CI limit
## [1] 0.3455249
# Exact binomial test
binom.test(0, 15, p = 0.5)
##
##  Exact binomial test
##
## data:  0 and 15
## number of successes = 0, number of trials = 15, p-value =
## 6.104e-05
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.0000000 0.2180194
## sample estimates:
## probability of success
##                      0
```

The population parameter (p) is the probability of major side effects from using mephetamines to treat children with traumatic brain injuries. The sample probability ($\hat{p}$) of major side effects was 0%. With 95% confidence, the maximum plausible value for the probability of major side effects from using amphetamines to treat children with traumatic brain injuries is 21.8 %.

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

Calculation the bound based on the normal distribution is not appropriate. The bound based on the normal distribution assumes a large sample size and as a rule of thumb is appropriate when (the expected number of successes) $n * p_0 \geq 5$ and (the expected number of failures) $n(1 - p_0) \geq 5$. Additionally, the normal approximation is less reliable for extreme values of $\hat{p}$ (e.g., as the sample proportion approaches zero or unity).

3 Suicides

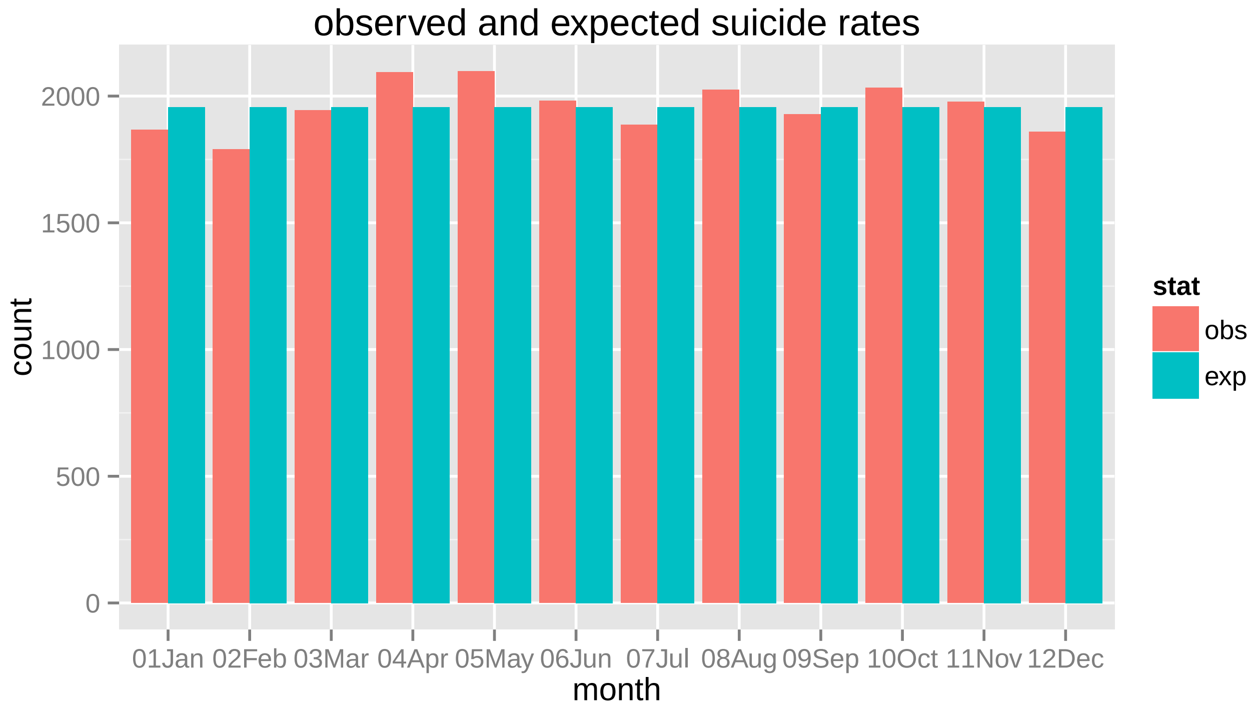
```
# read data from space delimited text
suicide <- read.table(text="
  Month  Suicides
  01Jan      1867
  02Feb      1789
  03Mar      1944
  04Apr      2094
  05May      2097
  06Jun      1981
  07Jul      1887
  08Aug      2024
  09Sep      1928
  10Oct      2032
  11Nov      1978
  12Dec      1859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

suicide$prop <- suicide$Suicides / sum(suicide$Suicides) # calculate proportions
suicide$prop.eq <- (sum(suicide$Suicides)/12) / sum(suicide$Suicides) # calculate equal proportions
sum(suicide$prop) # check that sum is equal to 1
```

```
# calculate chi-square goodness-of-fit
x.summ <- chisq.test(suicide$Suicides, correct = FALSE, p = suicide$prop.eq)
x.summ
##
##  Chi-squared test for given probabilities
##
## data:  suicide$Suicides
## X-squared = 51.7905, df = 11, p-value = 2.975e-07
x.table <- data.frame(month = suicide$Month,
                      obs = suicide$Suicides,
                      exp = x.summ$expected,
                      res = x.summ$residuals,
                      chisq = x.summ$residuals^2,
                      stdres = x.summ$stdres)

x.table.long <- melt(x.table, id.vars = c("month"), measure.vars = c("obs", "exp"),
                     variable.name = "stat", value.name = "value")
```

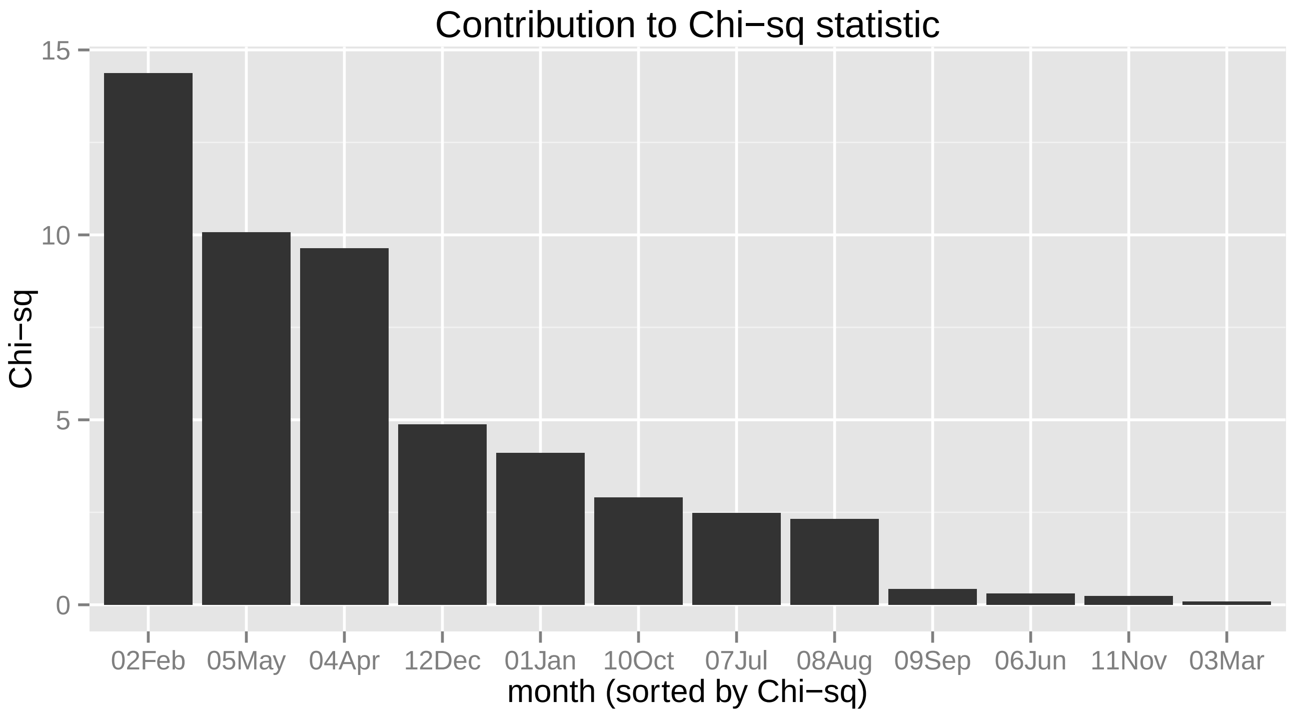
This analysis was formulated as a goodnes-of-fit test. The plausibility of a constant suicide rate per month was tested by comparing the proportions of the estimated (observed) suicides each month to the hypothesized proportion (sum of suicides divided by 12 months). The null hypothesis was that each month had a proportion of suicides equal to the hypothesized proportion; that is $H_0 : p_1 = p_{01}, p_2 = p_{02}, ..., p_{12} = p_{012}$; and H_A : not H_0.

The p-value from the Chi-squared test was $2.98e^{-7}$; therefore, at the 5% level I reject the null hypothesis in favor of the alternative. That is, it is not plausible that the suicide rate is constant. Comparisons between the expected and observed suicide counts are shown below along with the respective Chi-squared values for each month. From these data, it is apparent that February, May, and April have high Chi-squared values that do not agree with the null hypothesis.

```
#create bar plot comparing recorded to expected suicides by month
s.bar <- ggplot(x.table.long, aes(x = month, fill = stat, weight = value))
s.bar <- s.bar + geom_bar(position = "dodge")
s.bar <- s.bar + labs(xlab = "Month", title = "observed and expected suicide rates")
s.bar
```



```r
x.table.chisq <- x.table[,c("month", "chisq")]
x.table.chisq$month <- with(x.table, reorder(month, -chisq))

# create bar plot showing contribution of chi-sq
p <- ggplot(x.table.chisq, aes(x = month, weight = chisq))
p <- p + geom_bar()
p <- p + labs(title = "Contribution to Chi-sq statistic")
p <- p + xlab("month (sorted by Chi-sq)")
p <- p + ylab("Chi-sq")
p
```



```
b.sum1 <- binom.test(suicide$Suicides[1], sum(suicide$Suicides), p = suicide$prop.eq[1], alternative = "two.si
b.sum2 <- binom.test(suicide$Suicides[2], sum(suicide$Suicides), p = suicide$prop.eq[2], alternative = "two.si
b.sum3 <- binom.test(suicide$Suicides[3], sum(suicide$Suicides), p = suicide$prop.eq[3], alternative = "two.si
b.sum4 <- binom.test(suicide$Suicides[4], sum(suicide$Suicides), p = suicide$prop.eq[4], alternative = "two.si
b.sum5 <- binom.test(suicide$Suicides[5], sum(suicide$Suicides), p = suicide$prop.eq[5], alternative = "two.si
b.sum6 <- binom.test(suicide$Suicides[6], sum(suicide$Suicides), p = suicide$prop.eq[6], alternative = "two.si
b.sum7 <- binom.test(suicide$Suicides[7], sum(suicide$Suicides), p = suicide$prop.eq[7], alternative = "two.si
b.sum8 <- binom.test(suicide$Suicides[8], sum(suicide$Suicides), p = suicide$prop.eq[8], alternative = "two.si
b.sum9 <- binom.test(suicide$Suicides[9], sum(suicide$Suicides), p = suicide$prop.eq[9], alternative = "two.si
b.sum10 <- binom.test(suicide$Suicides[10], sum(suicide$Suicides), p = suicide$prop.eq[10], alternative = "two
b.sum11 <- binom.test(suicide$Suicides[11], sum(suicide$Suicides), p = suicide$prop.eq[11], alternative = "two
b.sum12 <- binom.test(suicide$Suicides[12], sum(suicide$Suicides), p = suicide$prop.eq[12], alternative = "two

b.sum  <- data.frame(
            rbind( c(b.sum1$p.value, b.sum1$conf.int)
                 , c(b.sum2$p.value, b.sum2$conf.int)
                 , c(b.sum3$p.value, b.sum3$conf.int)
                 , c(b.sum4$p.value, b.sum4$conf.int)
                 , c(b.sum5$p.value, b.sum5$conf.int)
                 , c(b.sum6$p.value, b.sum6$conf.int)
                 , c(b.sum7$p.value, b.sum7$conf.int)
                 , c(b.sum8$p.value, b.sum8$conf.int)
                 , c(b.sum9$p.value, b.sum9$conf.int)
                 , c(b.sum10$p.value, b.sum10$conf.int)
                 , c(b.sum11$p.value, b.sum11$conf.int)
                 , c(b.sum12$p.value, b.sum12$conf.int)
) )

names(b.sum) <- c("p.value", "CI.lower", "CI.upper")
b.sum$Month <- suicide$Month
b.sum$Observed <- x.table$obs/sum(x.table$obs)
b.sum$EqualProp <- suicide$prop.eq
b.sum <- b.sum[,c(4,1:3,5,6)]
```

	Month	p.value	CI.lower	CI.upper	Observed	EqualProp
1	01Jan	0.035	0.076	0.083	0.080	0.083
2	02Feb	0.000	0.073	0.080	0.076	0.083
3	03Mar	0.777	0.079	0.087	0.083	0.083
4	04Apr	0.001	0.085	0.093	0.089	0.083
5	05May	0.001	0.086	0.093	0.089	0.083
6	06Jun	0.563	0.081	0.088	0.084	0.083
7	07Jul	0.101	0.077	0.084	0.080	0.083
8	08Aug	0.114	0.083	0.090	0.086	0.083
9	09Sep	0.509	0.078	0.086	0.082	0.083
10	10Oct	0.077	0.083	0.090	0.087	0.083
11	11Nov	0.612	0.081	0.088	0.084	0.083
12	12Dec	0.021	0.076	0.083	0.079	0.083

Results from the exact binomial tests show that months Feb., May, Apr., Dec, and Jan. all have p-values less than 0.05. Again, bases an these analysis I reject the null hypothesis in favor of the alternative; that is the suicide rate is not constant.

4 Welsh and Breton

```
prop.test(c(76, 57), c(86, 77), correct = FALSE)
##
##  2-sample test for equality of proportions without continuity
##  correction
##
```

```
## data:  c(76, 57) out of c(86, 77)
## X-squared = 5.5677, df = 1, p-value = 0.0183
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.02437108 0.26255130
## sample estimates:
##    prop 1    prop 2
## 0.8837209 0.7402597
```

Let p_1 and p_2 be the proportion of bilingual adults in Welsh who speak Welsh fluently and the proportion of bilingual adults in Brittany who speak Brenton fluently, respectively. Then the null hypothesis is that both population proportions are equal, $H_0 : p_1 = p_2$, agains the alternative $H_A : p_1 \neq p_2$.

The p-value for the test of equal proportions was 0.018; therefore, at the 5% level I reject the null hypothesis in favor of the alternative. That is the proportion of Welsh bilingual adults who speak Welsh fluently is not equal to the proportion of Brittany bilingual adults who speak Brenton fluiently. With 95% confidence, the proportion of Welsh bilingual adults who speak Welsh fluetnly is greater than the proportion of Brittany bilingual adults who speak Breton fluently by at least 0.024 but less than 0.263.

5 Hawaiian blood

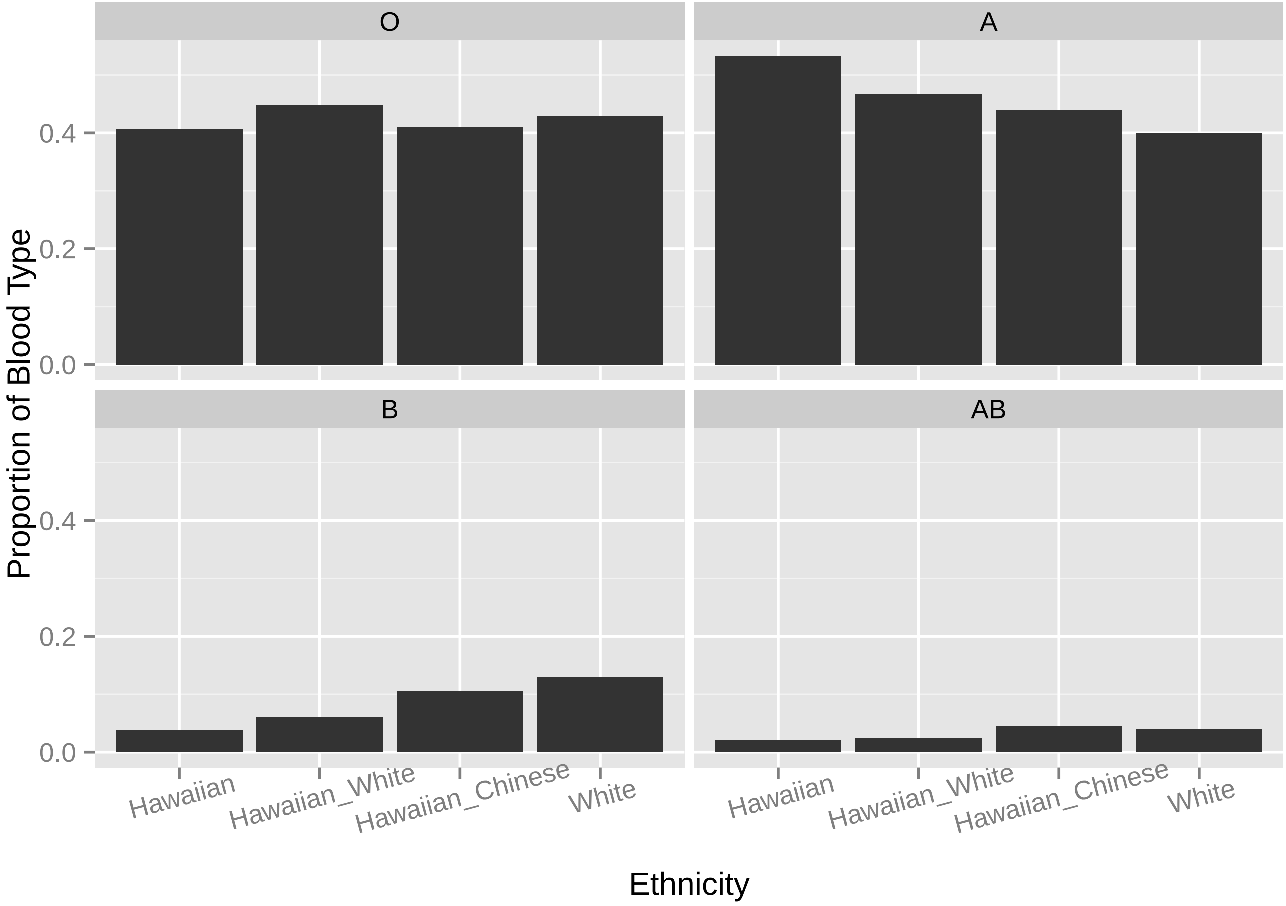
```
blood <- read.table(text="
Blood_Type Hawaiian   Hawaiian_White   Hawaiian_Chinese White
O 1903          4469             2206 53759
A 2490          4671             2368 50008
B 178           606              568 16252
AB 99           236              243 5001
", header=TRUE, skip=1)

# 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")))

eth.sum <- colSums(blood.matrix)
blood.matrix.sum <- rbind(blood.matrix, eth.sum)
rownames(blood.matrix.sum) <- c(rownames(blood.matrix),"Total")
blood.table <- t(apply(blood.matrix.sum, 1, function(x) x/eth.sum))
```

5(a) (15 pts) Summarize these data, focusing on comparing the proportions or percents in the 4 blood categories across the 4 ethnic groups.

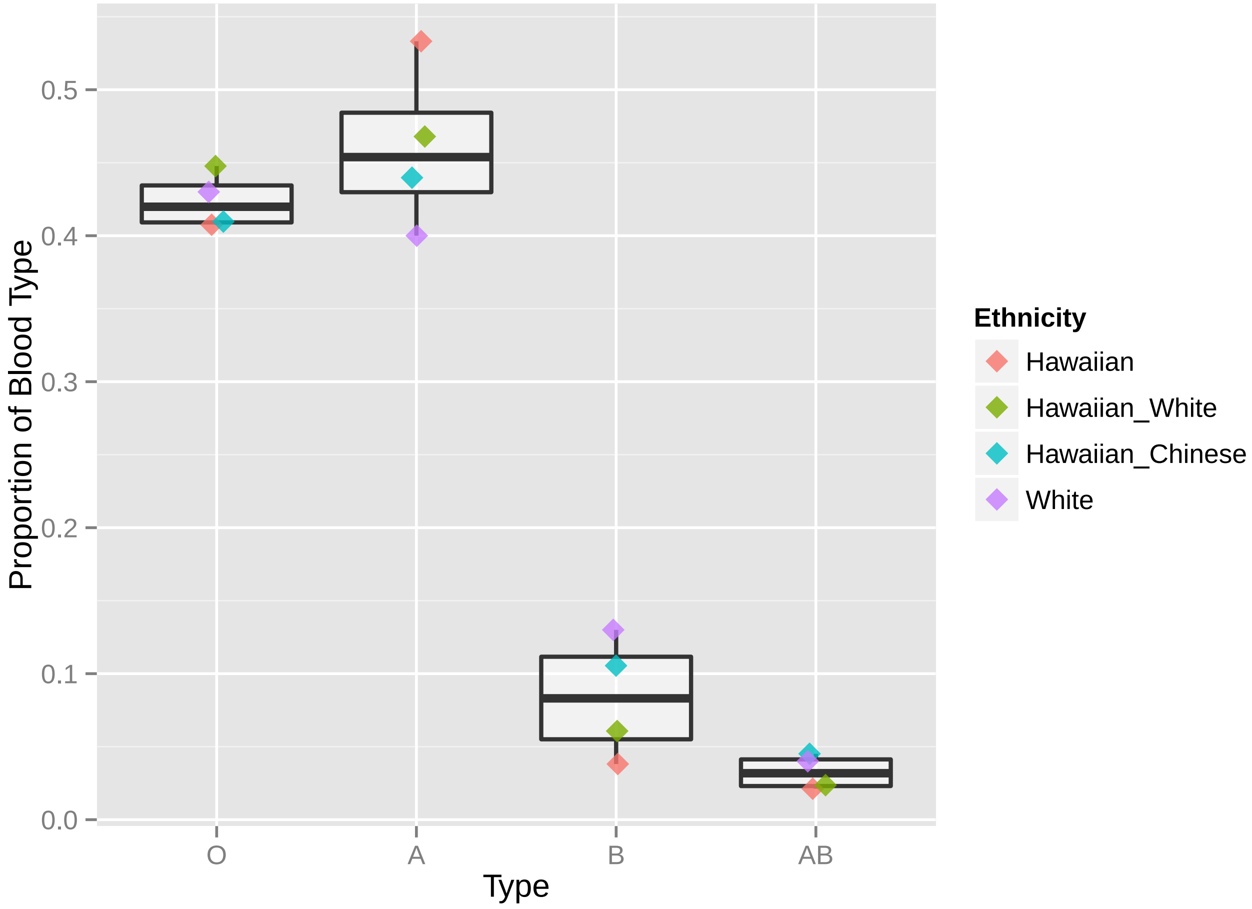
	Hawaiian	Hawaiian_White	Hawaiian_Chinese	White
O	0.407	0.448	0.410	0.430
A	0.533	0.468	0.440	0.400
B	0.038	0.061	0.105	0.130
AB	0.021	0.024	0.045	0.040
Total	1.000	1.000	1.000	1.000

```
blood.stat <- blood.table[1:4,]
blood.stat.long <- melt(blood.stat,
                        id.vars = colnames(blood.stat),
                        measure.vars = rownames(blood.stat),
                        variable.name = "Prop",
                        value.name = "Proportion")
colnames(blood.stat.long)[1:2] <- c("Type", "Ethnicity")
```

```
library(ggplot2)
p.b <- ggplot(blood.stat.long, aes(x = Ethnicity, weight = Proportion))
p.b <- p.b + geom_bar()
p.b <- p.b + facet_wrap(~Type)
p.b <- p.b + labs(y = "Proportion of Blood Type")
p.b <- p.b + theme(axis.text.x = element_text(angle = 15))
p.b
```



```
p.5 <- ggplot(blood.stat.long, aes(x = Type, y = Proportion))
p.5 <- p.5 + geom_boxplot(size = 0.75, alpha = 0.5)
p.5 <- p.5 + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 4,
                          aes(color = Ethnicity), alpha = 0.8,
                     position = position_jitter(w = 0.05, h = 0))
p.5 <- p.5 + labs(y = "Proportion of Blood Type")
p.5
```



The proportions of type O and AB blood have little varation across all ethnicities and have proportions ranging from 0.407 to 0.448 and 0.021 to 0.045, respectively. Proportions of Type A and B blood have much larger of variation across all ethnicities, with ranges from 0.400 to 0.533 and 0.038 to 0.130, respectively. The two most common blood types for all four ethnicities were Type O and A, while types B and AB were significantly less prevalent in all ethnicities. Hawaiians appear to have disproportionally high Type A blood while Whites appear to be disproportionally low. Conversely, Whites appear to have a disproportionally high rate of Type B blood while Hawaiians are disproportionally low in this blood type.

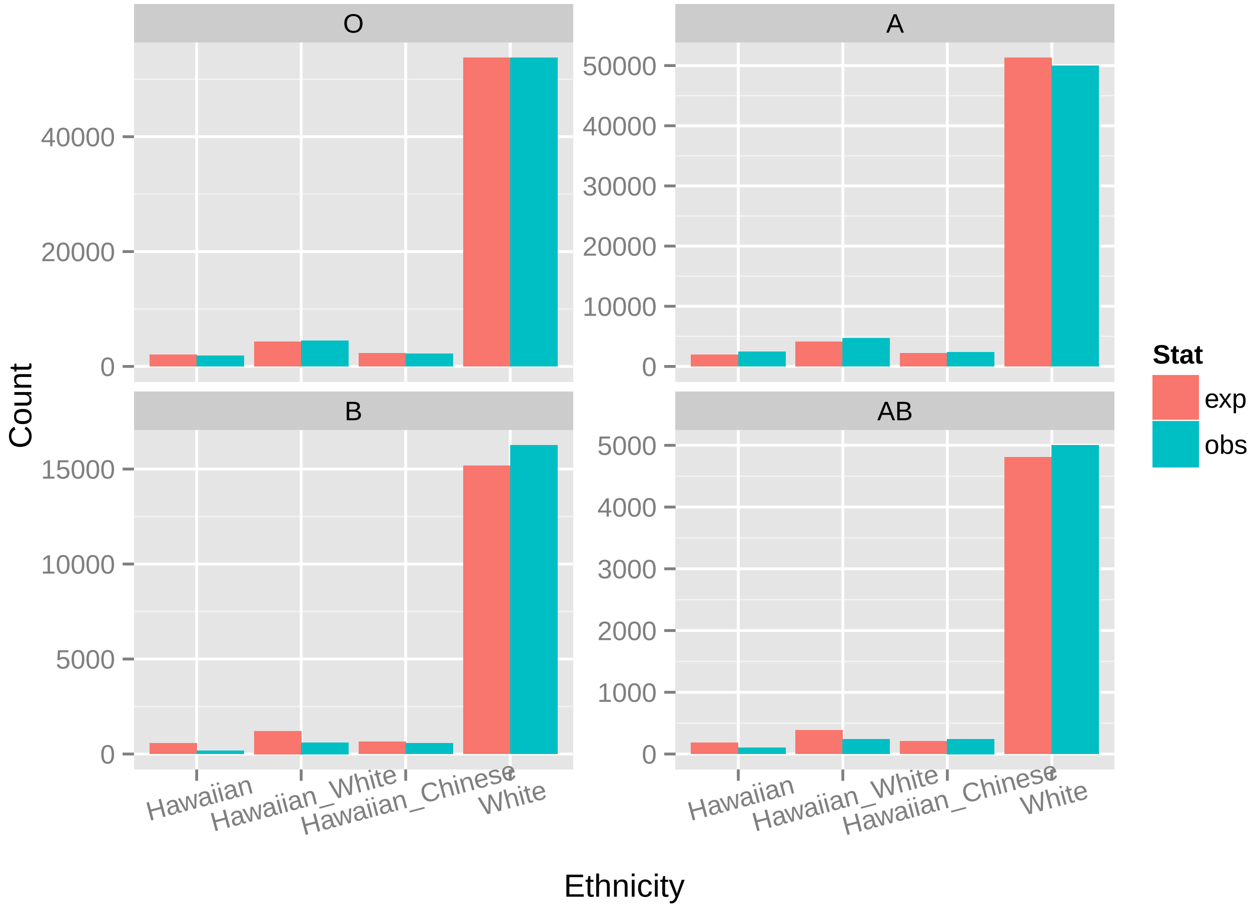
5(b) (15 pts) Is there evidence that blood type and ethnicity are associated in Hawaii? Explain.

A test for independence of blood type and ethnicity will be implemented using the chi-squared test. The null hypothesis is that the proportion of each blood type is consistent across ethnic groups in Hawaii $H_A : p_H = p_{HW} = p_{HC} = p_W$ against the alternative $H_A \neq H_0$ for blood types O, A, B, and AB. The subscripts H, HW, HC, and W represent Hawaiian, Hawaiian-White, Hawaiian-Chinese, and White, respectively.

```
chisq.5b <- chisq.test(blood.matrix, correct = FALSE)
chisq.5b
##
##  Pearson's Chi-squared test
##
## data:  blood.matrix
## X-squared = 1078.604, df = 9, p-value < 2.2e-16
```

The Pearson chi-squared test of independence resulted in a p-value of 2.2e-16, which is much less that 0.05. This result indicates an association between blood type and ethincity exists at the 5% level. The figure belows shows the observed versus expected values for each ethnicity and blood type.

5(c) (10pts) 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.



```
blood.table.pair <- data.frame(Interval = rep(NA,24) ,
                           CI.lower = rep(NA,24) ,
                           CI.upper = rep(NA,24) ,
                           Z = rep(NA,24) ,
                           p.value = rep(NA,24))

blood.table.pair[,1] <- c("O:p_H - p_HW",
                      "O:p_H - p_HC",
                      "O:p_H - p_W",
                      "O:p_HW - p_HC",
                      "O:p_HW - p_W",
                      "O:p_HC - p_W",

                      "A:p_H - p_HW",
                      "A:p_H - p_HC",
                      "A:p_H - p_W",
                      "A:p_HW - p_HC",
                      "A:p_HW - p_W",
                      "A:p_HC - p_W",

                      "B:p_H - p_HW",
                      "B:p_H - p_HC",
                      "B:p_H - p_W",
                      "B:p_HW - p_HC",
                      "B:p_HW - p_W",
                      "B:p_HC - p_W",

                      "AB:p_H - p_HW",
                      "AB:p_H - p_HC",
                      "AB:p_H - p_W",
                      "AB:p_HW - p_HC",
                      "AB:p_HW - p_W",
                      "AB:p_HC - p_W")

 i.tab <- 0
 for(i in 1:4){ #loops over blood type
   for (j in 1:3){ #moves first test stats
     for(k in (j + 1):4) { #moves second test stat
       i.tab <- i.tab + 1
       blood.summary <- prop.test(rbind(c(blood.matrix.sum[i,j], blood.matrix.sum[5,j] - blood.matrix.sum[i,j])
                            c(blood.matrix.sum[i,k], blood.matrix.sum[5,k] - blood.matrix.sum[i,k])),
                            correct = FALSE, conf.level = 1 - 0.05/6)

       blood.table.pair[i.tab, 2:5] <- c(blood.summary$conf.int[1],
                                    blood.summary$conf.int[2],
                                    sign(-diff(blood.summary$estimate)) * blood.summary$statistic^0.5,
                                    blood.summary$p.value)
     }
   }
 }
```

	Interval	CI.lower	CI.upper	Z	p.value
1	O:p_H - p_HW	-0.063	-0.017	-4.575	0.000
2	O:p_H - p_HC	-0.028	0.024	-0.220	0.826
3	O:p_H - p_W	-0.042	-0.003	-3.051	0.002
4	O:p_HW - p_HC	0.016	0.060	4.540	0.000
5	O:p_HW - p_W	0.004	0.031	3.437	0.001
6	O:p_HC - p_W	-0.038	-0.002	-2.954	0.003
7	A:p_H - p_HW	0.042	0.089	7.363	0.000
8	A:p_H - p_HC	0.067	0.120	9.352	0.000
9	A:p_H - p_W	0.114	0.153	18.206	0.000
10	A:p_HW - p_HC	0.006	0.050	3.348	0.001
11	A:p_HW - p_W	0.054	0.082	13.307	0.000
12	A:p_HC - p_W	0.022	0.058	5.825	0.000
13	B:p_H - p_HW	-0.032	-0.013	-5.663	0.000
14	B:p_H - p_HC	-0.081	-0.054	-12.854	0.000
15	B:p_H - p_W	-0.100	-0.084	-18.534	0.000
16	B:p_HW - p_HC	-0.057	-0.032	-9.968	0.000
17	B:p_HW - p_W	-0.076	-0.062	-20.151	0.000
18	B:p_HC - p_W	-0.036	-0.013	-5.256	0.000
19	AB:p_H - p_HW	-0.009	0.004	-0.922	0.356
20	AB:p_H - p_HC	-0.033	-0.015	-6.601	0.000
21	AB:p_H - p_W	-0.025	-0.013	-6.491	0.000
22	AB:p_HW - p_HC	-0.030	-0.013	-7.311	0.000
23	AB:p_HW - p_W	-0.021	-0.012	-8.145	0.000
24	AB:p_HC - p_W	-0.002	0.013	1.874	0.061

Comparisons for each blood type across ethnicities with the Bonferroni adjustment to account for multiple comparisons was performed. Within each blood type there were 6 possible comparisons; therefore, with an overall Family Error Rate of 0.05, the individual hypothese tests were performed at the 0.008 level. This analysis resulted in identifying which pairs have significant differences and the 99.2% confidence interval of those differences. All pairs showed a significant difference at the 0.8% level except for Type O:$p_H - p_{HC}$, Type AB:$p_H - p_{HW}$, and Type AB:$p_{HC} - p_W$.

Some additional analyses that could be done include an analysis of variance and pairwise analysis of blood types within an ethnicity.