7 Chi-squared Tests for Proportions (and Independence)

The chi-squared test (chisq.test() in R) can be used in the following 3 situations:

1. Testing whether k proportions in one population are equal to k specific (NULL) values
2. Testing **homogeneity** of r populations with respect to k categories
3. Testing whether two categorical variables are independent

7.1 Chi-squared test of k proportions in 1 population

Example: Tornados and El Nino

The data are as follows:
Number of tornadic days during El Nino years: 14
Number of tornadic days during La Nina years: 28
Number of tornadic days during Normal years: 44
Total number of days: 86

Number of years classified as El Nino: 12
Number of years classified as La Nina: 17
Number of years classified as Normal: 25
Total number of years: 54

We will first assume the following:

1. The proportion of tornadoes occurring in El Nino years is equal to the proportion of El Nino years.
2. The proportion of tornadoes occurring in La Nina years is equal to the proportion of La Nina years.
3. The proportion of tornadoes occurring in Normal years is equal to the proportion of Normal years.

If the above conditions are NOT supported by data, then we can say that "Data suggests that climate has an effect on tornadic activity." To answer the question, we set up the following hypothesis:

$$H_0\text{: } p_1 = \tfrac{12}{54},\ p_2 = \tfrac{17}{54},\ p_3 = \tfrac{25}{54}$$
$$H_1\text{: At least one of the three specifications in } H_0 \text{ is false}$$

```
obs.counts <- c(14, 28, 44)  # Note the data are entered as *counts*.
p0 <- c(12/54, 17/54, 25/54)  # But the null values are given as proportions.
chisq.test(obs.counts, p = p0 )  # Make sure to always specify p = p0.


Chi-squared test for given probabilities

data:  obs.counts
X-squared = 1.8, df = 2, p-value = 0.4
```

The exact p-value is 0.3988. At $\alpha = 0.05$, since p-value $> \alpha$, we cannot reject the null hypothesis (that climate has no effect on tornadic activity) in favor of the alternative (that it does). In short, there is no evidence that climate effects tornadic activity, at $\alpha = 0.05$.

```
# To see the p.value alone or the expected counts, use the following command:
chisq.test(obs.counts, p = p0)$p.value

[1] 0.3988

chisq.test(obs.counts, p = p0)$expected

[1] 19.11 27.07 39.81

# Diagnosis: check the individual terms in the observed X^2. The way
# to do that in R is to look for residuals. However, R defines the residual
# as (observed - expected)/sqrt(expected), and so to get the terms in X^2,
# you need to square these residuals:

chisq.test(obs.counts, p = p0)$residuals^2

[1] 1.36693 0.03167 0.43993
```

In this example, we can see that the biggest residual is from El Nino. I.e., the biggest difference between observed tornadic counts and the expected counts (if there were no effect between tornadoes and climate) is in El Nino years.

7.2 Testing homogeneity

The chisq.test() function can also be used to test homogeneity of r populations with respect to k categories in each. What that means is whether the k proportions in population 1 are equal to the k proportions in population 2, are equal to the k proportions in population 3, etc. A mathematically equivalent test is whether two categorical variables are independent. In other words, let us consider the contingency table. The question can be translated to whether the column and the row variable are independent. In the following example, the two variables are education level and religiosity. The first one is measured by the highest degree earned: Jr. College, College, and Grad School, and the second one is whether the subject declares him/herself as Fundamentalist or Moderate. Note that the first variable has 3 levels, and the second variable has 2 levels. That's a 2×3 (or 3×2) contingency table.

The question we want to answer is the following: Do the data provide evidence that religiosity and education are independent? (Equivalently, do the data provide evidence that religion is not homogeneous with respect to education?) To answer the question, we will set up the following hypothesis:

H_0: Religiosity is independent of education. (Religion is homogeneous with respect to education.)
H_1: Religiosity is dependent on education. (Religion is not homogeneous with respect to education.)

```
obs.counts = matrix(c(728, 1304, 495, 1072, 2800, 1193), ncol = 3, byrow = T)
chisq.test(obs.counts)


Pearson's Chi-squared test

data:  obs.counts
X-squared = 58, df = 2, p-value = 3e-13
```

```
# The columns are highest degree earned: Jr. College, College, and Grad School.
# The rows are religious belief: Fundamentalist, and Moderate.
```

Conclusion: Given that p-value $< \alpha$, we can reject the null hypothesis in favor of the alternative. I.e., there is evidence from data that religiosity is dependent on education. (Equivalently, religion is not homogeneous with respect to education.)

```
# To see the expected counts and the individual terms in X^2:
chisq.test(obs.counts)$expected
```

```
       [,1] [,2]   [,3]
[1,]  599.1 1366  561.9
[2,] 1200.9 2738 1126.1
```

```
chisq.test(obs.counts)$residuals^2
```

```
      [,1]  [,2]  [,3]
[1,] 27.72 2.816 7.954
[2,] 13.83 1.405 3.968
```

We can see that the biggest discrepancy between expected and observed is in the first category, i.e., in Jr, College. The next biggest discrepancy is in Graduate school. In "English:" There is a relationship between religiosity and education level, and the relationship is strongest in Jr. College and Graduate school. But what is that relationship? Nothing in chi.sq answers that question. For that we need to look at the data itself. For example, we can look at the proportion of Moderates within each of the education levels:

```
obs.counts[2, ] / apply(obs.counts, 2, sum)
```

```
[1] 0.5956 0.6823 0.7068
```

One can describe the relationship by saying that Moderateness increases with education level. There may be many different explanations for this pattern (e.g., parental factor, income level, etc.), but it certainly says that there is a positive relationship between moderateness and education.

7.3 Chi-squared using basic formulas

```
obs.counts <- matrix(c(435, 58, 89, 375, 50, 84), ncol = 3, byrow = T)
total <- sum(obs.counts)
rowsum <- apply(obs.counts, 1, sum) # If unfamiliar with apply(), look-up help. Here,
colsum <- apply(obs.counts, 2, sum) # it simply finds row and column marginals.
expected <- (matrix(rowsum) %*% t(matrix(colsum))) / total
expected
```

```
      [,1]  [,2]  [,3]
[1,] 432.1 57.61 92.29
[2,] 377.9 50.39 80.71
```

```
# This is the matrix of expected counts if the populations are homogeneous
# with respect to the categories (or if the row and column variables were
# independent).
residuals <- (obs.counts - expected) / sqrt(expected)
```

```r
df <- prod(dim(obs.counts) - 1)  # This df is just the product (nrow-1)*(ncol-1).
X2 <- sum(residuals^2)  # = observed X-squared.
1-pchisq(X2, df)
```

```
[1] 0.8614
```

```r
# p-value = area under the chi-squared distribution to the right of the
# observed X-squared.
```

8 F-test for 1-way ANOVA

Recall that the main question that can be addressed by 1-way ANOVA is whether the means of k samples are equal. Thus, we have the following hypothesis:

$$H_0\text{: } \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$
$$H_1\text{: At least two of the } \mu\text{'s are different.}$$

Example: Table 9.1 Vibration (in microns) in five groups of electric motors with each group using a different brand of bearing

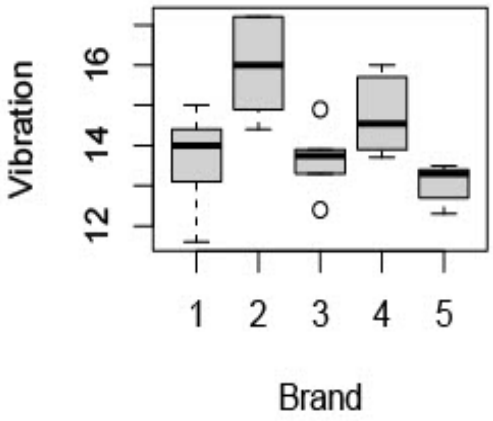
	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5
	13.1	16.3	13.7	15.7	13.5
	15.0	15.7	13.9	13.7	13.4
	14.0	17.2	12.4	14.4	13.2
	14.4	14.9	13.8	16.0	12.7
	14.0	14.4	14.9	13.9	13.4
	11.6	17.2	13.3	14.7	12.3
Mean:	13.68	15.95	13.67	14.73	13.08
St. dev:	1.194	1.167	.816	.940	.479
ANOVA Table					
Source	df	SS	MS	F	
Factor	4	30.88	7.72	8.45	
Error	25	22.83	.913		
Total	29	53.71			

Note that the data provided by the following link are entered in a form that makes ANOVA look like regression: i.e., the 1st column is x and the 2nd column is y. Although lm() is capable of handling situations where x is discrete/categorical (in which case that x is referred to as a dummy variable), generally when one speaks of regression it is assumed that x is continuous. Regardless, in regression the reponse y is continuous, but in ANOVA it's discrete/categorical.

```
dat <- read.table("9_1_dat.txt",  header = TRUE)

aov.1 = aov(Vibration ~ as.factor(Brand), data = dat)
summary(aov.1)

                Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(Brand)  4   30.9    7.71    8.44 0.00019 ***
Residuals        25   22.8    0.91
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Note that similar results can be obtained from general linear models (glm),
# which is a generalization of linear regression using the following command.

# glm.1 = glm(Vibration ~ as.factor(Brand), data = dat)
# anova(glm.1)
```

Since the p-value (.00018 < most α's) is really small, we reject the null in favor of the alternative. I.e., the data suggest that at least 2 of the means are different. One way to identify which two means are different is to at the following boxplots. This plot shows the 5-number summary at each level of x, i.e., something about the **spread** of the data.

```
boxplot(Vibration ~ Brand, data = dat)
```



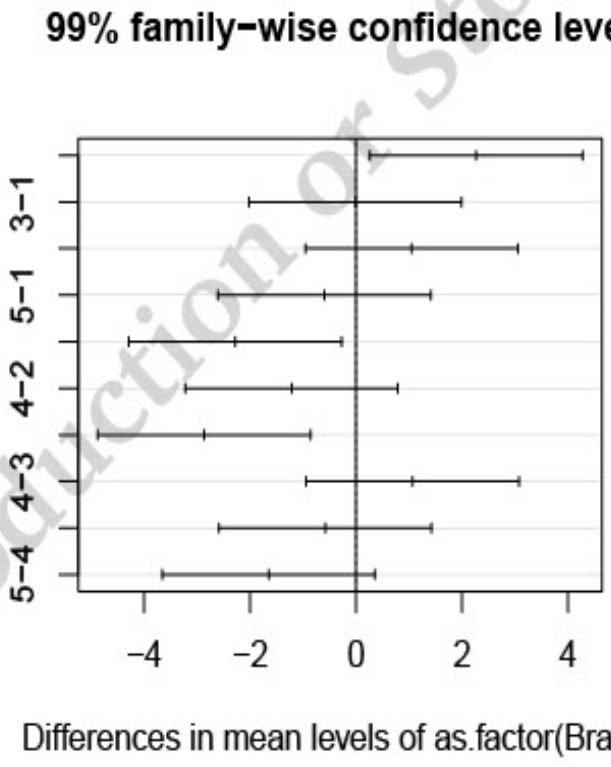
This allows for a visual comparison of the distribution of the 5 populations. The p-value suggests that at least 2 of the means are different. It's evident, for example, that the population means of brand 2 and 5 are probably different. But to quantify this observation, we need to do a "post hoc" analysis, an example of which is Tukey's.

8.1 Tukey's Test

The following performs Tukey's method (section 9.3 of the textbook) for identifying the different means. It gives confidence intervals and p-values for pairwise tests of population means. Recall that if the confidence interval does NOT include zero, then we conclude that the two means being tested are different.

```
library(stats)
tuk.1 <- TukeyHSD(aov.1, conf.level = 0.99)
tuk.1

  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = Vibration ~ as.factor(Brand), data = dat)

$`as.factor(Brand)`
        diff     lwr     upr  p adj
2-1  2.26667  0.2595  4.2738 0.0032
3-1 -0.01667 -2.0238  1.9905 1.0000
4-1  1.05000 -0.9571  3.0571 0.3418
5-1 -0.60000 -2.6071  1.4071 0.8113
3-2 -2.28333 -4.2905 -0.2762 0.0029
4-2 -1.21667 -3.2238  0.7905 0.2107
5-2 -2.86667 -4.8738 -0.8595 0.0002
4-3  1.06667 -0.9405  3.0738 0.3268
```

```
5-3 -0.58333 -2.5905  1.4238 0.8262
5-4 -1.65000 -3.6571  0.3571 0.0445

# The lower-bound (lwr) and upper-bound (upr) are given for the difference
# in the mean of two pops. These values are affected by the confidence level in
# TukeyHSD(). Then, look at the p-values in the last column; they test
# H0 vs. H1: two means are different. At alpha = 0.01, it's evident that the
# means of brand 1 and 2 are different. Other different pairs are 3-2, and
# 5-2.  Compare these results with those on page 419; the difference is
# in the alpha level.

plot(tuk.1)  # Shows the CIs as a figure.
abline(v = 0)

# This makes it visually clear that 1-2, 3-2, and 5-2 are three different-mean
# pairs. Note that Tukey's method is done after ANOVA, and its calculations
# depend on the results of ANOVA. As such, it cannot be done before
# ANOVA. There is no point in testing the means pairwise, unless there
# is evidence that at least two of the means are different - and that's
# what the F-test does.
```



8.2 1-Way ANOVA using basic formulas

If only means and standard deviations are given from data, then ANOVA must be done using basic formulas.

For a 1-way ANOVA test, we will first compute the mean and standard deviation for the data in Table 9.1, then use the basic ANOVA equations to show that we get the same answers as above.

```
dat <- read.table("9_1_dat.txt", header = TRUE)

attach(dat)  # The attach function loads in all variables in the data set.
k <- 5  # Number of categories.
n <- m <- s <- numeric(k)  # Space for mean and sd in each category.
for(i in 1:k) {
    n[i] <- 6  # Sample size in each category.
    m[i] <- mean(dat[Brand == i, 2])  # Mean in each category.
    s[i] <- sd(dat[Brand == i, 2])  # Standard deviation in each category.
}  # Ignore R warnings, if any.

# To do ANOVA by hand, we need n, m and s:
n

[1] 6 6 6 6 6

m

[1] 13.68 15.95 13.67 14.73 13.08

s

[1] 1.1940 1.1675 0.8165 0.9395 0.4792

# ANOVA using basic anova equations:
df.1 <- k - 1  # Numerator degrees of freedom.
df.2 <- k * 6 - k  # Denominator df.
SSB <- sum(n * (m - mean(m)) ^ 2)  # Sum of squares between groups.
SSW <- sum((n - 1) * s ^ 2 )  # Sum of squares within groups.
MSB <- SSB / df.1  # Mean-squared between groups.
MSW <- SSW / df.2  # Mean-squared within groups.
FF <- MSB/MSW  # F-ratio.
p.value <- 1-pf(FF, df.1, df.2)  # p-value.

df.1; df.2; SSB; SSW; MSB; MSW; FF; p.value

[1] 4
[1] 25
[1] 30.86
[1] 22.84
[1] 7.714
[1] 0.9135
[1] 8.444
[1] 0.0001871

# Note that results are the same as the ANOVA table above.
```

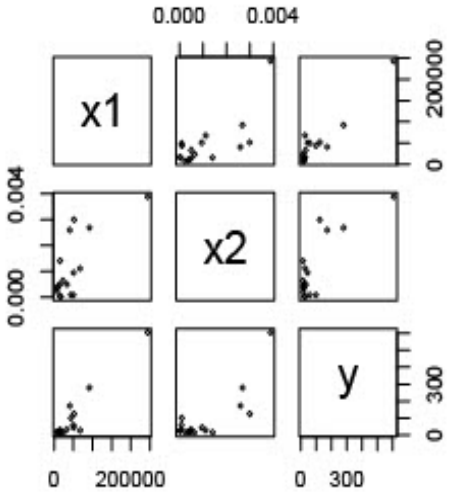
9 T-test and F-test for Regression Coefficients

In multiple regression the F-test is for testing if **at least one** of the regression coefficients is nonzero, because then there is evidence that at least one of the predictors is useful, i.e., the model has some utility. Also, if the model has some utility, then we can try to identify which regression coefficients are the nonzero ones, because then the corresponding predictors are the useful ones. The latter step is done with a sequence of t-tests, each on a different coefficient.

Example: Problem 11.39 (in 2nd edition)

Snowpacks contain a wide spectrum of pollutants which may represent environmental hazards. The article "Atmospheric PAH Deposition: Deposition Velocities and Washout Ratios" (J. of Environmental Engineering, 2002: 186???195) focused on the deposition of polyaromatic hydrocarbons. The authors proposed a multiple regression model for relating deposition over a specified time period y to two rather complicated predictors x_1 and x_2 defined in terms of PAH air concentrations for various species, total time and total amount of precipitation. The data is on the web at:

```
dat <- read.table("11_39_dat.txt", header = TRUE)

plot(dat, cex = 0.5)  # Look at the data, and note the collinearity.
model.1 <- lm(y ~ x1 + x2, data = dat)  # Fit a linear model.
summary(model.1)


Call:
lm(formula = y ~ x1 + x2, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-111.6  -19.7   18.2   27.4   44.9

Coefficients:
               Estimate   Std. Error t value  Pr(>|t|)
(Intercept)   -33.463810    14.896258   -2.25     0.041 *
x1              0.002055     0.000295    6.98 0.0000065 ***
x2          29835.665532 13653.728296    2.19     0.046 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44.3 on 14 degrees of freedom
Multiple R-squared:  0.923,Adjusted R-squared:  0.912
F-statistic: 84.4 on 2 and 14 DF,  p-value: 0.0000000155
```



Note that the F-ratio tests for "model utility", i.e., if at least one variable is a significant predictor of y. The way it's done in practice is to compare the so-called "full model" $(y = intercept + x_1 + x_2)$, against the so-called "null model" $(y = intercept)$. The appropriate statistic is the F-ratio, which is returned in summary():

```
summary(model.1)$fstatistic  # Returns the F-ratio and degrees of freedom

value numdf dendf
84.39  2.00 14.00

summary(model.1)$fstatistic[1]  # Selects only F-ratio.

value
84.39
```

The 3 p-values appearing in the table test the full model against a model without one variable. These are based on the t-tests for testing whether the respective regression coefficient is nonzero. In this case, at $\alpha = 0.01$, it looks like x1 is the useful predictor and at $\alpha = 0.05$, both x1 and x2 are useful. We can also compute a confidence interval for each of the regression coefficients:

```
confint(model.1, level = 0.99)

                    0.5 %       99.5 %
(Intercept)    -77.807627    10.880008
x1               0.001178     0.002932
x2          -10809.336342 70480.667406

# The CI for beta1 does not include zero, but the CI for beta2 does. So the
# conclusions from the CI's are consistent with the conclusions from the p-values.
```

9.1 Confidence Interval vs. Prediction Interval

Recall that the confidence interval is a confidence interval for the population mean of y given x and the prediction interval is not a CI at all because it is not referring to a population parameter. Instead, it