Chapter 22

Alan T. Arnholt - Modified notes from the Second Edition of *Probability and Statistics with R*

Last compiled: May 10, 2022 at 09:27:36 AM

Contents

1 Comparing Counts

Objectives:

I. Goodness of fit test

II. Independence test

III. Homogeneity test

IV. What can go wrong?

Many statistical procedures require knowledge of the population from which the sample is taken. For example, using Student's t-distribution for testing a hypothesis or constructing a confidence interval for μ assumes that the parent population is normal. In this section, **goodness-of-fit** (GOF) procedures are presented that will help to identify the distribution of the population from which the sample is drawn. The null hypothesis in a goodness-of-fit test is a statement about the form of the cumulative distribution. When all the parameters in the null hypothesis are specified, the hypothesis is called simple. Recall that in the event the null hypothesis does not completely specify all of the parameters of the distribution, the hypothesis is said to be composite. Goodness-of-fit tests are typically used when the form of the population is in question. In contrast to most of the statistical procedures discussed so far, where the goal has been to reject the null hypothesis, in a GOF test one hopes to retain the null hypothesis.

1.1 The Chi-Square Goodness-of-Fit Test

Given a single random sample of size n from an unknown population F_X, one may wish to test the hypothesis that F_X has some known distribution $F_0(x)$ for all x. For example, using the data frame `SOCCER` from the `PASWR2` package, is it reasonable to assume the number of goals scored during regulation time for the 232 soccer matches has a Poisson distribution with $\lambda = 2.5$?

The chi-square goodness-of-fit test is based on a normalized statistic that examines the vertical deviations between what is observed and what is expected when H_0 is true in k mutually exclusive categories. At times, such as in surveys of brand preferences, where the categories/groups would be the brand names, the sample

will lend itself to being divided into k mutually exclusive categories. Other times, the categories/groupings will be more arbitrary. Before applying the chi-square goodness-of-fit test, the data must be grouped according to some scheme to form k mutually exclusive categories. When the null hypothesis completely specifies the population, the probability that a random observation will fall into each of the chosen or fixed categories can be computed. Once the probabilities for a data point to fall into each of the chosen or fixed categories is computed, multiplying the probabilities by n produces the expected counts for each category under the null distribution. If the null hypothesis is true, the differences between the counts observed in the k categories and the counts expected in the k categories should be small. The test criterion for testing $H_0 : F_X(x) = F_0(x)$ for all x against the alternative $H_1 : F_X(x) \neq F_0(x)$ for some x when the null hypothesis is completely specified is

$$\chi^2_{\text{obs}} = \sum_{i=1}^{k} \frac{(O_k - E_k)^2}{E_k}, \tag{1}$$

where χ^2_{obs} is the sum of the squared deviations between what is observed (O_k) and what is expected (E_k) in each of the k categories divided by what is expected in each of the k categories. Large values of χ^2_{obs} occur when the observed data are inconsistent with the null hypothesis and thus lead to rejection of the null hypothesis. The exact distribution of χ^2_{obs} is very complicated; however, for large n, provided all expected categories are at least 5, χ^2_{obs} is distributed approximately χ^2 with $k - 1$ degrees of freedom. When the null hypothesis is composite, that is, not all of the parameters are specified, the degrees of freedom for the random variable χ^2_{obs} are reduced by one for each parameter that must be estimated.

Example

Test the hypothesis that the number of goals scored during regulation time for the 232 soccer matches stored in the data frame `SOCCER` has a Poisson **cdf** with $\lambda = 2.5$ with the chi-square goodness-of-fit test and an α level of 0.05. Produce a histogram showing the number of observed goals scored during regulation time and superimpose on the histogram the number of goals that are expected to be made when the distribution of goals follows a Poisson distribution with $\lambda = 2.5$.

Solution

Since the number of categories for a Poisson distribution is theoretically infinite, a table is first constructed of the observed number of goals to get an idea of reasonable categories.

```
library(PASWR2)
xtabs(~goals, data = SOCCER)
```

```
goals
 0  1  2  3  4  5  6  7  8
19 49 60 47 32 18  3  3  1
```

Based on the table, a decision is made to create categories for 0, 1, 2, 3, 4, 5, and 6 or more goals. Under the null hypothesis that $F_0(x)$ is a Poisson distribution with $\lambda = 2.5$, the probabilities of scoring 0, 1, 2, 3, 4, 5, and 6 or more goals are computed with `R` as follows:

```
PX <- c(dpois(0:5, 2.5), ppois(5, 2.5, lower = FALSE))
PX
```

```
[1] 0.08208500 0.20521250 0.25651562 0.21376302 0.13360189 0.06680094 0.04202104
```

Since there were a total of $n = 232$ soccer games, the expected number of goals for the six categories is simply $232 \times$ `PX`.

```
EX <- 232*PX
OB <- c(as.vector(xtabs(~goals, data = SOCCER)[1:6]),
        sum(xtabs(~goals, data = SOCCER)[7:9]))
OB
```

```
[1] 19 49 60 47 32 18  7
```

```
ans <- cbind(PX, EX, OB)
row.names(ans) <- c(" X=0"," X=1"," X=2"," X=3"," X=4"," X=5","X>=6")
ans
```

```
             PX        EX OB
 X=0 0.08208500 19.043720 19
 X=1 0.20521250 47.609299 49
 X=2 0.25651562 59.511624 60
 X=3 0.21376302 49.593020 47
 X=4 0.13360189 30.995638 32
 X=5 0.06680094 15.497819 18
X>=6 0.04202104  9.748881  7
```

Hypotheses— The null and alternative hypotheses for using the chi-square goodness-of-fit test to test the hypothesis that the number of goals scored during regulation time for the 232 soccer matches stored in the data frame `SOCCER` has a Poisson **cdf** with $\lambda = 2.5$ are

$$H_0 : F_X(x) = F_0(x) \sim \text{Pois}(\lambda = 2.5) \text{ for all } x \text{ versus}$$
$$H_A : F_X(x) \neq F_0(x) \text{ for some } x.$$

Test Statistic:— The test statistic chose is χ^2_{obs}.

Rejection Region Calculations—Reject if $\chi^2_{\text{obs}} > \chi^2_{1-\alpha;k-1}$. The χ^2_{obs} is computed with (1) in `R` below.

```
chi_obs <- sum((OB - EX)^2/EX)
chi_obs
```

```
[1] 1.39194
```

$1.3919402 = \chi^2_{\text{obs}} \overset{?}{>} \chi^2_{0.95;6} = 12.5915872$.

Statistical Conclusion—The $p-$value is 0.9663469.

```
p_val <- pchisq(chi_obs, 7-1, lower = FALSE)
p_val
```

```
[1] 0.9663469
```

I. Since $\chi^2_{\text{obs}} = 1.3919402$ is not greater than $\chi^2_{0.95;6} = 12.5915872$, fail to reject H_0.

II. Since the $p-$value $= 0.9663469$ is greater than 0.05, fail to reject H_0.

Fail to reject H_0.

English Conclusion—There is no evidence to suggest that the true **cdf** does not equal the Poisson distribution with $\lambda = 2.5$ for at least one x.

To perform a goodness-of-fit test with the function `chisq.test()`, one may specify a vector of observed values for the argument `x=`, and a vector of probabilities of the same length as the vector passed to `x=` to the argument `p=`.

```
chisq.test(x = OB, p = PX)
```

```
    Chi-squared test for given probabilities

data:  OB
X-squared = 1.3919, df = 6, p-value = 0.9663
```

The code below uses base graphics to create a histogram with superimposed expected goals and the result is shown in Figure 1.

```
hist(SOCCER$goals, breaks = c((-0.5 + 0):(8 + 0.5)), col = "lightblue",
     xlab = "Goals scored", ylab = "", freq = TRUE, main = "")
x <- 0:8
fx <- (dpois(0:8, lambda = 2.5))*232
lines(x, fx, type = "h")
lines(x, fx, type = "p", pch = 16)
```

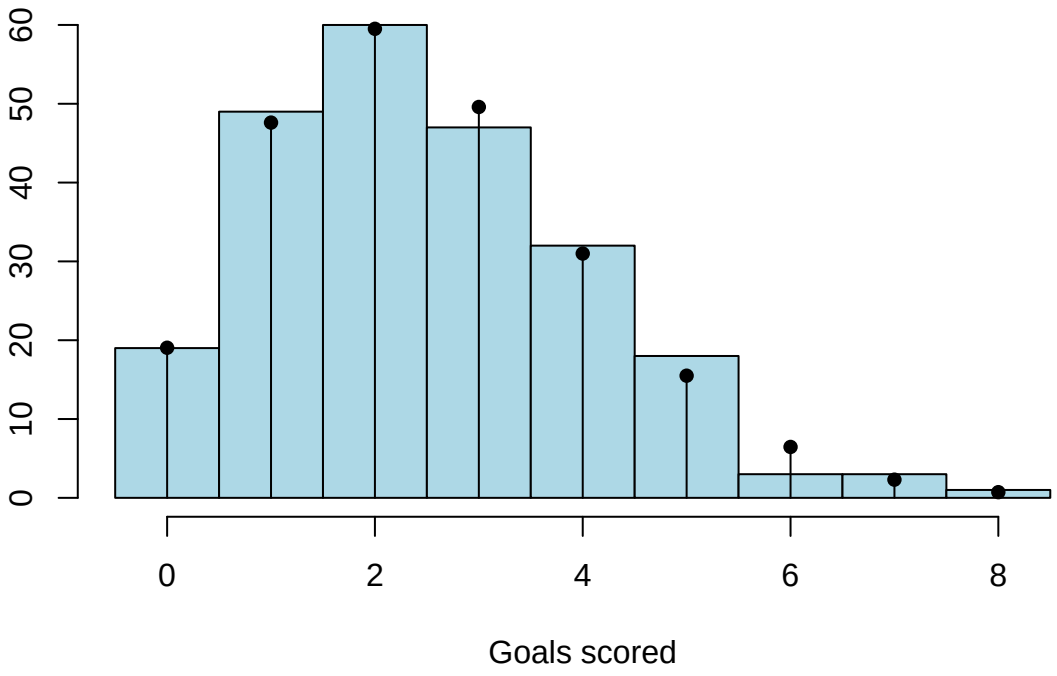

Figure 1: Histogram of observed goals for 'SOCCER' with a superimposed Poisson distribution with lambda = 2.5 (vertical lines)

Although the chi-square goodness-of-fit test is primarily designed for discrete distributions, it can also be used with a continuous distribution if appropriate categories are defined.

Example

Use the chi-square goodness-of-fit test with $\alpha = 0.05$ to test the hypothesis that the SAT scores stored in the data frame `GRADES` have a normal **cdf**. Use categories $(-\infty, \mu - 2\sigma]$, $(\mu - 2\sigma, \mu - \sigma]$, $(\mu - \sigma, \mu]$, $(\mu, \mu + \sigma]$, $(\mu + \sigma, \mu + 2\sigma]$, and $(\mu + 2\sigma, \infty]$. Produce a histogram using the categories specified and superimpose on the histogram the expected number of SAT scores in each category when $F_0(x) \sim N(\mu = \bar{x}, \sigma = s)$.

Solution

Hypotheses—The null and alternative hypotheses for using the chi-square goodness-of-fit test to test the hypothesis that the SAT scores stored in the data frame `GRADES` have a Normal **cdf** are

$$H_0 : F_X(x) = F_0(x) \sim N(\mu = \bar{x}, \sigma = s) \text{ for all } x \text{ versus}$$
$$H_A : F_X(x) \neq F_0(x) \text{ for some } x.$$

Test Statistic—Since the mean and standard deviation are unknown, the first step is to estimate the unknown parameters μ and σ using $\bar{x} = 1134.65$ and $s = 145.6086774$.

```
mu <- mean(GRADES$sat)
sig <- sd(GRADES$sat)
c(mu, sig)
```

```
[1] 1134.6500  145.6087
```

Because a normal distribution is continuous, it is necessary to create categories that include all the data. The categories $\mu - 3\sigma$ to $\mu - 2\sigma, \ldots, \mu + 2\sigma$ to $\mu + 3\sigma$ are 697.8239678 to 843.4326452, 843.4326452 to 989.0413226, 989.0413226 to 1134.65, 1134.65 to 1280.2586774, 1280.2586774 to 1425.8673548, and 1425.8673548 to 1571.4760322. These particular categories include all of the observed SAT scores; however, the probabilities actually computed for the largest and smallest categories will be all of the area to the right and left, respectively, of $\bar{x} \pm 2s$. This is done so that the total area under the distribution in the null hypothesis is one.

```
bin <- seq(from = mu - 3*sig, to = mu + 3*sig, by = sig)
round(bin, 0)                       # vector of bin cut points
```

```
[1]  698  843  989 1135 1280 1426 1571
```

```
T1 <- table(cut(GRADES$sat, breaks = bin))
T1                                  # count of observations in bins
```

```
          (698,843]           (843,989]      (989,1.13e+03] (1.13e+03,1.28e+03]
                  4                  27                  65                  80
(1.28e+03,1.43e+03] (1.43e+03,1.57e+03]
                 21                   3
```

```
OB <- as.vector(T1)
OB                                  # vector of observations
```

```
[1]  4 27 65 80 21  3
```

```
PR <- c(pnorm(-2), pnorm(-1:2) - pnorm(-2:1),
        pnorm(2, lower = FALSE))    # area under curve
EX <- 200*PR                        # Expected count in bins
ans <- cbind(PR, EX, OB)            # column bind values in ans
ans
```

```
            PR        EX OB
[1,] 0.02275013  4.550026  4
[2,] 0.13590512 27.181024 27
[3,] 0.34134475 68.268949 65
[4,] 0.34134475 68.268949 80
[5,] 0.13590512 27.181024 21
```

```
[6,] 0.02275013  4.550026  3
```

Rejection Region Calculations—Reject if $\chi^2_{\text{obs}} > \chi^2_{1-\alpha;k-p-1}$. Now that the expected and observed counts for each of the categories are computed, the χ^2_{obs} value can be computed according to (1) and is 4.1736536.

```
chi_obs <- sum((OB - EX)^2/EX)
chi_obs
```

```
[1] 4.173654
```

Statistical Conclusion—In this problem, two parameters were estimated, and as a consequence, the degrees of freedom are computed as $6 - 2 - 1 = 3$. The $p-$value is 0.2433129.

```
p_val <- pchisq(chi_obs, 6 - 2 - 1, lower = FALSE)
p_val
```

```
[1] 0.2433129
```

I. Since $\chi^2_{\text{obs}} = 4.1736536$ is not greater than $\chi^2_{0.95;3} = 7.8147279$, fail to reject H_0.

II. Since the $p-$value $= 0.2433129$ is greater than 0.05, fail to reject H_0.

Fail to reject H_0

English Conclusion—There is no evidence to suggest that the true **cdf** of SAT scores is not a normal distribution.

Caution: If one uses the `R` function `chisq.test()`, the degrees of freedom and the subsequent $p-$value will be incorrect, as illustrated below.

```
chisq.test(x = OB, p = PR)  # returns incorrect dof and p-value
```

```
    Chi-squared test for given probabilities

data:  OB
X-squared = 4.1737, df = 5, p-value = 0.5247
```

Since it is not feasible to produce a histogram that extends from $-\infty$ to ∞, a histogram is created where the categories will simply cover the range of observed values. In this problem, the range of the SAT scores is 720 to 1550. The histogram with categories $(\mu - 3\sigma, \mu - 2\sigma]$, $(\mu - 2\sigma, \mu - \sigma]$, $(\mu - \sigma, \mu]$, $(\mu+, \mu + \sigma]$, $(\mu + \sigma, \mu + 2\sigma]$, and $(\mu + 2\sigma, \mu + 3\sigma]$, superimposed with the expected number of SAT scores for the categories $(-\infty, \mu - 2\sigma]$, $(\mu - 2\sigma, \mu - \sigma]$, $(\mu - \sigma, \mu]$, $(\mu, \mu + \sigma]$, $(\mu + \sigma, \mu + 2\sigma]$, and $(\mu + 2\sigma, \infty]$ is computed below and depicted in Figure 2.

```
hist(GRADES$sat, breaks = bin, col = "lightblue",
     xlab = "SAT scores", ylab="", freq = TRUE, main = "")
x <- bin[2:7] - sig/2
fx <- PR*200
lines(x, fx, type = "h")
lines(x, fx, type = "p", pch = 16)
```

Your Turn

A psychology professor reports that historically grades in her intro class have been distributed 15% A, 30% B, 40% C, 10% D, and 5% F. Grades this year were distributed:

```
  A   B   C   D   E
 89 121  78  25  12
```

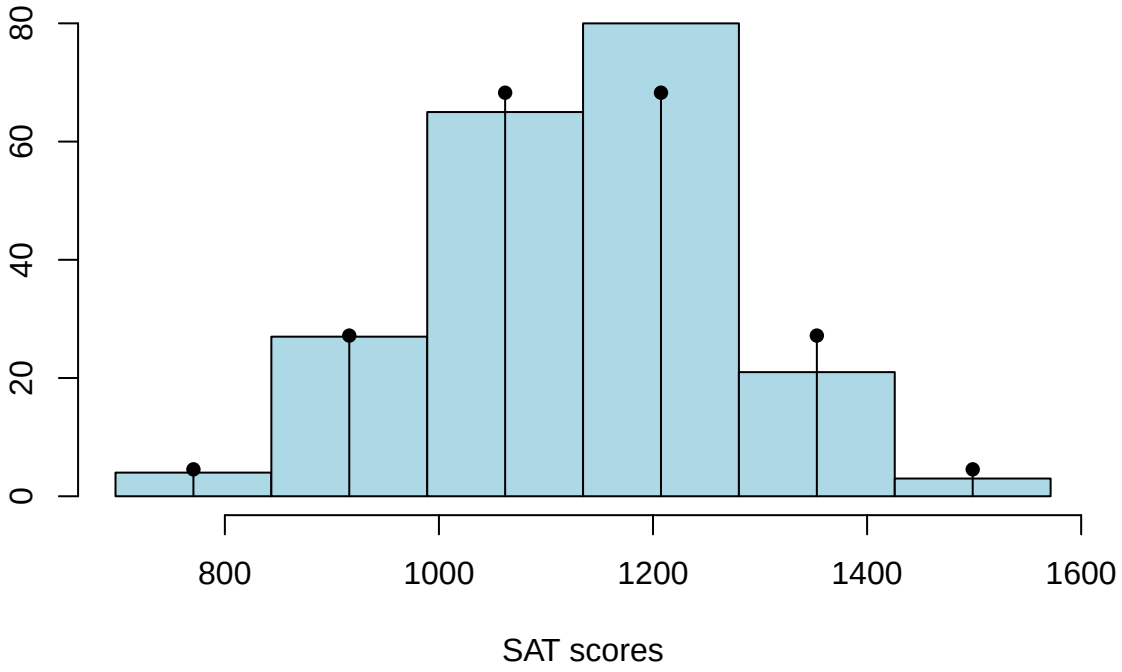

Figure 2: Histogram of SAT scores in 'Grades'

Is there evidence that this year's distribution is different from the historical distribution? If yes, which grade impacted that decision most?

```
# Your Code Here
OB <- c(89, 121, 78, 25, 12)
names(OB) <- LETTERS[1:5]
OB
```

```
  A   B   C   D   E
 89 121  78  25  12
```

```
EX <- sum(OB)*c(0.15, 0.3, 0.4, 0.1, 0.05)
SS <- (OB - EX)^2/EX
SS
```

```
        A         B         C         D         E
33.232051  5.664103 20.800000  1.730769  1.111538
```

```
# Largest component is A's
chi_obs <- sum(SS)
p_val <- pchisq(chi_obs, 4, lower = FALSE)
p_val
```

```
[1] 8.48661e-13
```

```
chisq.test(x = OB, p = c(0.15, 0.3, 0.4, 0.1, 0.05))
```

```
    Chi-squared test for given probabilities

data:  OB
X-squared = 62.538, df = 4, p-value = 8.487e-13
```

Given a contingency table (Question 5)

```
ad <- c(20, 45, 35, 25, 50, 25)
adm <- matrix(ad, byrow = TRUE, nrow = 2)
adm
```

```
     [,1] [,2] [,3]
[1,]   20   45   35
[2,]   25   50   25
```

```
dimnames(adm) <- list(Branch = c("In-Town", "Mall"), Age = c("Less Than 30", "30-55", "56 or older"))
adm
```

```
         Age
Branch    Less Than 30 30-55 56 or older
  In-Town           20    45          35
  Mall              25    50          25
```

```
chisq.test(adm)
```

```
    Pearson's Chi-squared test

data:  adm
X-squared = 2.4854, df = 2, p-value = 0.2886
```

```
# To get the expected counts use $exp
chisq.test(adm)$exp
```

```
         Age
Branch    Less Than 30 30-55 56 or older
  In-Town         22.5  47.5          30
  Mall            22.5  47.5          30
```

To get **Standardized Residuals** compute $\frac{(\text{Obs - Exp})}{\sqrt{\text{Exp}}}$. If using the `chisq.test()` function one may extract the residuals using `$residuals`

```
chisq.test(adm)$residuals
```

```
         Age
Branch    Less Than 30      30-55 56 or older
  In-Town   -0.5270463 -0.3627381   0.9128709
  Mall       0.5270463  0.3627381  -0.9128709
```

```
# Equivalent to
(chisq.test(adm)$obs - chisq.test(adm)$exp)/chisq.test(adm)$exp^.5
```

```
         Age
Branch    Less Than 30      30-55 56 or older
  In-Town   -0.5270463 -0.3627381   0.9128709
  Mall       0.5270463  0.3627381  -0.9128709
```

Questions 8

Given a ratio 9:3:3:1, the fraction for each category will be 9/16, 3/16, 3/16, 1/16.

```
obs <- c(54, 21, 13, 12)
obs2 <- obs*2
p <- c(9/16, 3/16, 3/16, 1/16)
chisq.test(x = obs, p = p)
```

```
    Chi-squared test for given probabilities
```

```
data:  obs
X-squared = 7.4133, df = 3, p-value = 0.05983
```

```
chisq.test(obs2, p = p)
```

```
    Chi-squared test for given probabilities

data:  obs2
X-squared = 14.827, df = 3, p-value = 0.001971
```

1.2 Independence test

SCENARIO ONE: Is there an association between gender and a person's happiness? To investigate whether happiness depends on gender, one might use information collected from the General Social Survey (GSS) http://sda.berkeley.edu/GSS. In each survey, the GSS asks, "Taken all together, how would you say things are these days — would you say that you are very happy, pretty happy, or not too happy?'' Respondents to each survey are coded as either male or female. The information below shows how a subset of respondents (26-year-olds) were classified with respect to the variables HAPPY and SEX.

```
HA <- c(110, 277, 50, 163, 302, 63)
HAT <- matrix(data = HA, nrow = 2, byrow = TRUE)
dimnames(HAT) <- list(SEX = c("Male", "Female"),
    Category = c("Very Happy", "Pretty Happy", "Not To Happy"))
HAT
```

```
        Category
SEX      Very Happy Pretty Happy Not To Happy
  Male          110          277           50
  Female        163          302           63
```

Scenario one asks if there is an association between gender and a person's happiness. Two events, A and B, were defined as independent when $P(A \cap B) = P(A) \times P(B)$ or, equivalently, when $P(A|B) = P(A)$. If, instead of having a random sample from a single population, an $I \times J$ contingency table consisted of entries from the population, association could be mathematically verified by showing that $P(n_{ij}) \neq P(n_{i\bullet}) \times P(n_{\bullet j})$ for some i and j. If by chance $P(n_{ij}) = P(n_{i\bullet}) \times P(n_{\bullet j})$ for all i and j, then one would conclude there is no association between gender and a person's happiness. That is, the variables gender and happiness would be considered mathematically independent. Since the entire population is not given but rather a sample from a population, the values in the $I \times J$ contingency table can be expected to change from sample to sample. The question is, "By how much can the variables deviate from the mathematical definition of independence and still be considered statistically independent?''

The null and alternative hypotheses to test for independence between row and column variables is written $H_0 : \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ versus $H_A : \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j}$. The test statistic is

$$\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2}$$

It compares the observed frequencies in the table with the expected frequencies when H_0 is true. Under the assumption of independence, and when the observations in the cells are sufficiently large (usually greater than 5), $\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \overset{\bullet}{\sim} \chi^2_{(I-1)(J-1)}$, where $\hat{\mu}_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n} = E_{ij}$ and $n_{ij} = O_{ij}$. The null hypothesis of independence is rejected when $\chi^2_{\text{obs}} > \chi^2_{1-\alpha;(I-1)(J-1)}$.

The chi-squared approximation is generally satisfactory if the E_{ij}s ($\hat{\mu}_{ij}$s) in the test statistic are not too small. Various rules of thumb exist for what might be considered too small. A very conservative rule is to require all E_{ij}s to be 5 or more. This can be accomplished by combining cells with small E_{ij}s and reducing the overall degrees of freedom. At times, it may be permissible to let the E_{ij} of a cell be as low as 0.5.

Test For SCENARIO ONE:

Hypotheses — $H_0 : \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ (Row and column variables are independent.) versus $H_A : \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j}$ for at least one i, j (Row and column variables are dependent.)

Test Statistic — The test statistic is

$$\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \overset{\bullet}{\sim} \chi^2_{(I-1)(J-1)} = \chi^2_{(2-1)(3-1)} = \chi^2_2$$

under the assumption of independence. The χ^2_{obs} value is 4.3214818.

```
chisq.test(HAT)$stat
```

```
X-squared
 4.321482
```

Rejection Region Calculations — The rejection region is

$$\chi^2_{\text{obs}} > \chi^2_{1-\alpha;(I-1)(J-1)} = \chi^2_{0.95;2} = 5.9914645.$$

Before the statistic $\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ can be computed, the expected counts for each of the ij cells must be calculated. Note that $O_{ij} = n_{ij}$ and $E_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n}$.

```
E <- chisq.test(HAT)$exp
E
```

```
        Category
SEX      Very Happy Pretty Happy Not To Happy
  Male      123.628        262.2     51.17202
  Female    149.372        316.8     61.82798
```

$$\chi^2_{\text{obs}} = \frac{(110-123.6280)^2}{123.6280} + \frac{(277-262.2)^2}{262.2} + \cdots + \frac{(63-61.828)^2}{61.828} = 4.3215.$$

The value of the test statistic is $\chi^2_{obs} = 4.3214818$. This can be done with code by entering

```
chisq.test(HAT)$stat
```

```
X-squared
 4.321482
```

```
# Or
chi_obs <- sum((HAT - E)^2/E)
chi_obs
```

```
[1] 4.321482
```

$4.3214818 = \chi^2_{\text{obs}} \overset{?}{>} \chi^2_{0.95,2} = 5.9914645.$

Statistical Conclusion — The $p-$value is 0.1152397

```
chisq.test(HAT)
```

```
    Pearson's Chi-squared test

data:  HAT
X-squared = 4.3215, df = 2, p-value = 0.1152
```

```
chisq.test(HAT)$p.value
```

```
[1] 0.1152397
```

```
# Or
p_val <- pchisq(chi_obs, 2, lower = FALSE)
p_val
```

```
[1] 0.1152397
```

I. From the rejection region, since $\chi^2_{\text{obs}} = 4.3214818 < \chi^2_{0.95;2} = 5.9914645$. fail to reject the null hypothesis of independence.

II. Since the $p-$value $= 0.1152397$ is greater than 0.05, fail to reject the null hypothesis of independence.

Fail to reject H_0.

English Conclusion — There is not sufficient evidence to suggest the variables gender and happiness are statistically dependent.

EXAMPLE

Is there an association between eye color and hearing loss in patients suffering from meningitis? British researcher Helen Cullingham recorded the eye color of 130 deaf patients and noted whether the deafness had followed treatment for meningitis.

```
HC <- c(30, 72, 2, 26)
HCM <- matrix(HC, byrow = TRUE, nrow = 2)
dimnames(HCM) <- list(Eye_Color = c("Light", "Dark"), Deafness_related_to = c("Meningitis", "Other"))
HCM
```

```
         Deafness_related_to
Eye_Color Meningitis Other
    Light         30    72
    Dark           2    26
```

```
#Your code to test the appropriate hypothesis

chisq.test(HCM, correct = FALSE)
```

```
    Pearson's Chi-squared test

data:  HCM
X-squared = 5.8712, df = 1, p-value = 0.01539
```

SCENARIO TWO: In a double blind randomized drug trial (neither the patient nor the physician evaluating the patient knows the treatment, drug or placebo, the patient receives), 400 male patients with mild dementia were randomly divided into two groups of 200. One group was given a placebo over three months while the second group received an experimental drug for three months. At the end of the three months, the physicians

(all psychiatrists) classified the 400 patients into one of three categories: improved, no change, or worse. The information below shows how the psychiatrists classified the patients. Are the proportions in the three status categories the same for the two treatments?

```
DT <- c(67, 76, 57, 48, 73, 79)
DTT <- matrix(data = DT, nrow = 2, byrow = TRUE)
dimnames(DTT) <- list(Treatment = c("Drug", "Placebo"),
   Category = c("Improve", "No Change", "Worse"))
DTT
```

```
         Category
Treatment Improve No Change Worse
  Drug         67        76    57
  Placebo      48        73    79
```

In the **second scenario**, there are two distinct populations from which samples are taken. The first population is the group of all patients receiving the experimental drug while the second population is the group of all patients receiving a placebo. In this scenario, there are $I = 2$ separate populations and $J = 3$ categories for the $I = 2$ populations. Individuals sampled from the $I = 2$ distinct populations are classified into one of the $J = 3$ status categories. This scenario has fixed row totals whereas the **first scenario** does not. In the first scenario, only the total sample size, n, is fixed. That is, neither the row nor the column totals are fixed. This is in contrast to **scenario two**, where the number of patients in each treatment group (row) was fixed.

1.3 Test of Homogeneity

The question of interest in **scenario two** is whether the proportions in each of the $j = 3$ categories for the $i = 2$ populations are equivalent. Specifically, is $\pi_{1j} = \pi_{2j}$ for all j? This question is answered with a test of homogeneity. In general, the null hypothesis for a test of homogeneity with $i = I$ populations is written

$$H_0 : \pi_{1j} = \pi_{2j} = \cdots = \pi_{Ij} \text{ for all } j \text{ versus } H_1 : \pi_{ij} \neq \pi_{i+1,j} \text{ for some } (i, j). \tag{3}$$

Expressed in words, the null hypothesis is that the I populations are homogeneous with respect to the J categories versus the I populations are not homogeneous with respect to the J categories.

An equivalent interpretation is that for each population $i = 1, 2, \ldots, I$, the proportion of people in the jth category is the same. When H_0 is true, $\pi_{1j} = \pi_{2j} = \cdots = \pi_{Ij}$ for all j. Under the null hypothesis, $\mu_{ij} = n_{i\bullet}\pi_{ij}$, $\hat{\pi}_{ij} = p_{ij} = \frac{n_{\bullet j}}{n_{\bullet\bullet}}$, and $\hat{\mu}_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}} = E_{ij}$. When H_0 is true, all the probabilities in the jth column are equal, and a pooled estimate of π_{ij} is obtained by adding all the frequencies in the jth column ($n_{\bullet j}$) and dividing the total by $n_{\bullet\bullet}$. The statistic used in this type of problem has the same form as the one used for the test of independence in (2). Substituting the homogeneity expressions for O_{ij} and E_{ij}, the statistic is expressed as

$$\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i\bullet}n_{\bullet j}/n_{\bullet\bullet})^2}{n_{i\bullet}n_{\bullet j}/n_{\bullet\bullet}} \overset{\cdot}{\sim} \chi^2_{(I-1)(J-1)}.$$

The null hypothesis of homogeneity is rejected when $\chi^2_{\text{obs}} > \chi^2_{1-\alpha;(I-1)(J-1)}$.

Test for SCENARIO TWO:

Hypotheses— $H_0 : \pi_{1j} = \pi_{2j}$ for all j versus $H_A : \pi_{i,\,j} \neq \pi_{i+1,\,j}$ for some (i, j). That is, all the probabilities in the same column are equal to each other versus at least two of the probabilities in the same column are not equal to each other.

Test Statistic—The test statistic is

$$\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-1)} = \chi^2_{(2-1)(3-1)} = \chi^2_2$$

under the null hypothesis. The χ^2_{obs} value is 6.7583566.

Rejection Region Calculations—The rejection region is

$$\chi^2_{\text{obs}} > \chi^2_{1-\alpha;(I-1)\cdot(J-1)} = \chi^2_{0.95;2} = 5.9914645.$$

Before the statistic $\chi^2_{\text{obs}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ can be computed, the expected counts for each of the ij cells must be determined. Recall that $O_{ij} = n_{ij}$ and $E_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}$.

```
DTT
```

```
         Category
Treatment Improve No Change Worse
  Drug         67        76    57
  Placebo      48        73    79
```

```
E <- chisq.test(DTT)$expected
E
```

```
         Category
Treatment Improve No Change Worse
  Drug       57.5      74.5    68
  Placebo    57.5      74.5    68
```

$$\chi^2_{\text{obs}} = \frac{(67-57.5)^2}{57.5} + \frac{(76-74.5)^2}{74.5} + \cdots + \frac{(79-68)^2}{68} = 6.7583566.$$

The value of the test statistic is $\chi^2_{obs} = 6.7583566$. This can be done with code by entering

```
chisq.test(DTT)
```

```
    Pearson's Chi-squared test

data:  DTT
X-squared = 6.7584, df = 2, p-value = 0.03408
```

```
chisq.test(DTT)$stat
```

```
X-squared
 6.758357
```

```
# Or
chi_obs <- sum((DTT - E)^2/E)
chi_obs
```

```
[1] 6.758357
```

$6.7583566 = \chi^2_{\text{obs}} \overset{?}{>} \chi^2_{.95,2} = 5.9914645$.

Statistical Conclusion—The $p-$value is 0.0340754.

```
p_val <- pchisq(chi_obs, 2, lower = FALSE)
p_val
```

```
[1] 0.03407544
```

I. From the rejection region, since $\chi^2_{\text{obs}} = 6.7583566 > \chi_{0.95;2} = 5.9914645$, reject the null hypothesis of homogeneity.

II. Since the $p-$value $= 0.0340754$ is less than 0.05, reject the null hypothesis of homogeneity.

Reject H_0.

English Conclusion—There is sufficient evidence to suggest that not all of the probabilities for the $i = 2$ populations with respect to each of the J categories are equal.

Example

In a study of the television viewing havits of children, a developmental psychologist selects two random samples of first graders: one with 100 boys and one with 200 girls. Each child is asked which of the following TV programs the like best: The Lone Ranger, Sesame Street, or The Simpsons. Results are shown below.

```
TV <- c(50, 30, 20, 50, 80, 70)
TVM <- matrix(data = TV, nrow = 2, byrow = TRUE)
dimnames(TVM) <- list(FirstGraders = c("Boys", "Girls"), ViewingPreference = c("Lone Ranger", "Sesame S
TVM
```

```
            ViewingPreference
FirstGraders Lone Ranger Sesame Street The Simpsons
       Boys           50            30           20
       Girls          50            80           70
```

Do the boys' preferences for these TV programs differ significantly from the girls' preferences? Use $\alpha = 0.05$ to test your hypotheses.

```
chisq.test(TVM)
```

```
    Pearson's Chi-squared test

data:  TVM
X-squared = 19.318, df = 2, p-value = 6.384e-05
```

Your English Conclusion

What can go wrong?

1. χ^2 tests are **only** for counts
2. Samples reasonably sized
3. Note that "not independent" $\neq$ dependent, only association exists

---