STT 3850 : Chi-Square Tests

Fall 2023

Appalachian State University

Section 1

Chi-Square Goodness-of-Fit Tests

Background

- 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.
- **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**.

Background

- 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.
- 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.

Background

- 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 $\lambda = 2.5$?
- 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.

Background

- 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.

Background

- 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.

Background

- NOTE: 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.

Soccer 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 $\lambda = 2.5$.

Soccer 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
```

Soccer Solution

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[1:4] # Probabilities for categories 0, 1, 2, 3
```

```
[1] 0.0820850 0.2052125 0.2565156 0.2137630
```

```
PX[4:6] # Probabilities for categories 4, 5, and 6 or more
```

```
[1] 0.21376302 0.13360189 0.06680094
```

Soccer Solution

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

```
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")
```

Soccer Solution

```
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
```

Soccer Solution

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

$$H_0 : F_X(x) = F_0(x) \sim Pois(\lambda = 2.5) \text{ for all } x \text{ versus}$$
$$H_1 : F_X(x) \neq F_0(x) \text{ for some } x.$$

Soccer Solution

- The test statistic chosen is χ^2_{obs}.
- Reject if $\chi^2_{\text{obs}} > \chi^2_{1-\alpha;k-1}$.

```
chi.obs <- sum((OB-EX)^2/EX)
chi.obs
```

```
[1] 1.39194
```

Soccer Solution

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

```

	Chi-squared test for given probabilities

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

Soccer Solution

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

The p-value is 0.9663469.

```
p.val <- pchisq(chi.obs, 7-1, lower = FALSE)
p.val
```

```
[1] 0.9663469
```

Soccer Solution

- Since $\chi^2_{\textsf{obs}} = 1.3919402$ is not greater than $\chi^2_{0.95;6} = 12.5915872$, fail to reject H_0.
- Since the p-value $=$ 0.9663469 is greater than 0.05, fail to reject H_0.

Soccer Solution

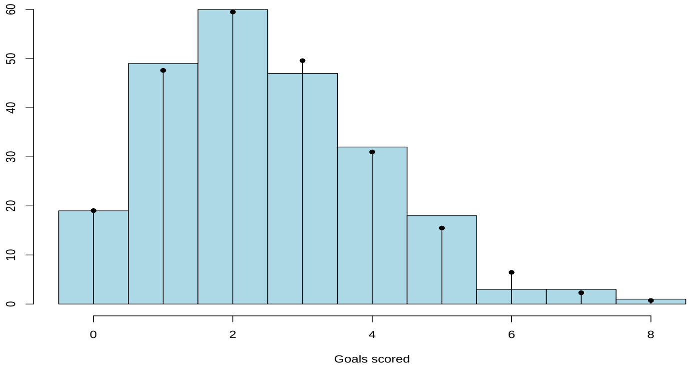
English Conclusion: There is no evidence to suggest that the true **cdf** does not equal the Poisson distribution with $\lambda = 2.5$ for at least one x.

Soccer Solution

The following code can be used to create a histogram with superimposed expected goals.

```
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)
```

Soccer Solution



All Parameters Known

- Bansley et al. (1992) investigated the relationship between month of birth and achievement in sport. Birth dates were collected for players in teams competing in the 1990 World Cup soccer games.

```
Observed <- c(150, 138, 140, 100)
names(Observed) <- c("Aug-Oct", "Nov-Jan",
                     "Feb-April", "May-July")
Observed
```

```
  Aug-Oct   Nov-Jan Feb-April  May-July 
      150       138       140       100 
```

All Parameters Known

We wish to test whether these data are consistent with the hypothesis that birthdays of soccer players are uniformly distributed across the four quarters of the year. Let P_i denote the probability of a birth occurring in the i^{th} quarter; the hypotheses are as follows:

$H_0 : p_1 = \frac{1}{4}, p_2 = \frac{1}{4}, p_3 = \frac{1}{4}, p_4 = \frac{1}{4}$ versus $H_A : p_i \neq \frac{1}{4}$ for at least one i.

There were a total of $n = 528$ players considered for this study, so the expected count for each quarter is $528/4 = 132$.

All Parameters Known

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(150-132)^2}{132} + \frac{(138-132)^2}{132} + \frac{(140-132)^2}{132} + \frac{(100-132)^2}{132} = 10.97$$

```
(chi_obs <- sum((Observed - 132)^2/132))
```

```
[1] 10.9697
```

```
# Or
chisq.test(Observed, p = c(1/4, 1/4, 1/4, 1/4))$stat
```

```
X-squared
  10.9697
```

All Parameters Known

```
chisq.test(Observed, p = c(1/4, 1/4, 1/4, 1/4)) -> CST
CST
```

```

    Chi-squared test for given probabilities

data:  Observed
X-squared = 10.97, df = 3, p-value = 0.01189
```

```
CST$observed
```

```
  Aug-Oct   Nov-Jan Feb-April  May-July
      150       138       140       100
```

```
CST$expected
```

```
  Aug-Oct   Nov-Jan Feb-April  May-July
      132       132       132       132
```

All Parameters Known

```r
(pvalue <- pchisq(CST$stat, 3, lower = FALSE))
```

```
 X-squared
0.01189087
```

```r
# Or
CST$p.value
```

```
[1] 0.01189087
```

All Parameters Known - Conclusion

Given the $p - value$ of 0.012 evidence suggests birthdays for World Cup soccer players are not uniformly distributed.

All Parameters Known - Example 2

Suppose you draw 100 numbers at random from an unknown distribution. Thirty values fall in the interval $(0, 0.25]$, 30 fall in $(0.25, 0.75]$, 22 fall in $(0.75, 1.25]$, and the rest fall in $(1.25, \infty]$. Your friend claims that the distribution is exponential with parameter $\lambda = 1$. Do you believe her?

- A random variable X has the exponential distribution with parameter $\lambda > 0$ if its **pdf** is

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

All Parameters Known - Example 2

We wish to test the following:

H_0 : The data are from an exponential distribution with $\lambda = 1$.
H_A : The data are not from an exponential distribution with $\lambda = 1$.

All Parameters Known - Example 2

Given $X \sim \text{Exp}(\lambda = 1)$. The probabilities for each interval are as follows:

$p_1 = P(0 \leq X \leq 0.25) = \int_0^{0.25} e^{-x}\, dx = 0.2211992$

$p_2 = P(0.25 \leq X \leq 0.75) = \int_{0.25}^{0.75} e^{-x}\, dx = 0.3064342$

$p_3 = P(0.75 \leq X \leq 1.25) = \int_{0.75}^{1.25} e^{-x}\, dx = 0.1858618$

$p_4 = P(1.25 \leq X \leq \infty) = \int_{1.25}^{\infty} e^{-x}\, dx = 0.2865048$

All Parameters Known - Example 2

```r
p1 <- pexp(0.25, 1)
p2 <- pexp(0.75, 1) - pexp(0.25, 1)
p3 <- pexp(1.25, 1) - pexp(0.75, 1)
p4 <- pexp(1.25, 1, lower = FALSE)
ps <- c(p1, p2, p3, p4)
ps
```

```
[1] 0.2211992 0.3064342 0.1858618 0.2865048
```

All Parameters Known - Example 2

```r
EXP <- ps*100
EXP
```

```
[1] 22.11992 30.64342 18.58618 28.65048
```

```r
OBS <- c(30, 30, 22, 18)
test_stat <- sum((OBS - EXP)^2/EXP)
test_stat
```

```
[1] 7.406963
```

All Parameters Known - Example 2

```r
# Another approach
chisq.test(OBS, p = ps)
```

```
    Chi-squared test for given probabilities

data:  OBS
X-squared = 7.407, df = 3, p-value = 0.06
```

```r
pvalue <- chisq.test(OBS, p = ps)$p.value
pvalue
```

```
[1] 0.05999777
```

All Parameters Known - Example 2 - Conclusion

If you test using $\alpha = 0.05$, you will fail to reject the null hypothesis since the $p - value = 0.0599 > \alpha = 0.05$. There is not convincing evidence that the data do not come from an Exp($\lambda = 1$).

Section 2

Categorical Data

Different Scenarios

The 2×2 contingency table can be generalized for I rows and J columns and is referred to as an $I \times J$ contingency table. The sampling scheme employed to acquire the information in the table will determine the type of hypothesis that can be tested. Consider the following two scenarios:

Scenario One:

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 in the next slide shows how a subset of respondents (26-year-olds) were classified with respect to the variables HAPPY and SEX.

Scenario One:

```r
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 - Expected Values

```
E <- outer(rowSums(HAT), colSums(HAT), "*")/sum(HAT)
E
```

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

```
# OR
chisq.test(HAT)$expected
```

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

Scenario Two

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 on the next slide shows how the pschiatrists classified the patients. Are the proportions in the three status categories the same for the two treatments?

Scenario Two

```r
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
```

Scenario Two - Expected Values

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

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

Categorical Data

The two scenarios illustrate two different sampling schemes that both result in $I \times J$ contingency tables. In the first scenario, there is a single population (Americans) and individuals are sampled from this single population and classified into one of the IJ cells of the $I \times J$ contingency table based on the $I = 2$ SEX categories and the $J = 3$ HAPPY categories. The format of an $I \times J$ contingency table when sampling from a single population is shown in Table 1. The number of observations from the i^{th} row classified into the j^{th} column is denoted by n_{ij}. It follows that the number of observations in the j^{th} column $(1 \leq j \leq J)$ is $n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{Ij}$, while the number of observations in the i^{th} row $(1 \leq i \leq I)$ is $n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{iJ}$.

Categorical Data

Table 1: Contingency table when sampling from a single population

	Col 1	Col 2	$\cdots$	Col J	*Totals*
Row 1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Row I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I\bullet}$
Totals	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet J}$	n

Categorial Data

The true population proportion of individuals in cell (i, j) will be denoted π_{ij}. Under the assumption of independence between row and column variables (SEX and HAPPY in this example), $\pi_{ij} = \pi_{i\bullet} \times \pi_{\bullet j}$, where $\pi_{i\bullet} = \sum_{j=1}^{J} \pi_{ij}$ and $\pi_{\bullet j} = \sum_{i=1}^{I} \pi_{ij}$. That is, $\pi_{i\bullet}$ is the proportion of observations in the population classified in category i of the row variable and $\pi_{\bullet j}$ is the proportion of observations in the population classified in category j of the column variable. Since $\pi_{i\bullet}$ and $\pi_{\bullet j}$ are marginal population proportions, it follows that $\hat{\pi}_{i\bullet} = p_{i\bullet} = \frac{n_{i\bullet}}{n}$ and $\hat{\pi}_{\bullet j} = p_{\bullet j} = \frac{n_{\bullet j}}{n}$, where n is the sample size. Under the assumption of independence the expected count for cell (i, j) is $\mu_{ij} = n\pi_{ij} = n\pi_{i\bullet}\pi_{\bullet j}$ and $\hat{\mu}_{ij} = n\hat{\pi}_{ij} = n\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j} = n\frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}n_{\bullet j}}{n}$.

Categorical Data

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. The notation used for an $I \times J$ contingency table when I samples from I distinct populations differs slightly from the notation used in Table 1 with a contingency table from a single sample.

Categorical Data

Since the sample sizes of the I distinct populations are denoted $n_{i\bullet}$, the total for all I samples is denoted by $n_{\bullet\bullet}$ rather than the notation n used for a single sample in Table 1. Table 2 shows the general form and notation used for an $I \times J$ contingency table when sampling from I distinct populations. Each observation in each sample is classified into one of J categories. If $n_{i\bullet}$ denotes the number of observations in the i^{th} sample $(1 \leq i \leq I)$ and n_{ij} denotes the number of observations from the i^{th} sample classified into the j^{th} category $(1 \leq j \leq J)$, it follows that the number of observations in the j^{th} column is $n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{Ij}$, while the number of observations in the i^{th} row is $n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{iJ}$.

Categorical Data

Table 2: General form and notation used for an $I \times J$ contingency table when sampling from I distinct populations

	Category 1	Category 2	$\cdots$	Category J	Totals
Population 1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\bullet}$
Population 2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Population I	n_{I1}	n_{I2}	$\ldots$	n_{IJ}	$n_{I\bullet}$
Totals	$n_{\bullet 1}$	$n_{\bullet 2}$	$\ldots$	$n_{\bullet J}$	$n_{\bullet\bullet}$

Section 3

Chi-Square Tests of Independence

Example

```r
library(PASWR2)
(xtabs(~sex + survived, data = TITANIC3) -> T1)
```

```
        survived
sex        0   1
  female 127 339
  male   682 161
```

```r
chisq.test(T1, correct = FALSE) -> CST
CST
```

```

    Pearson's Chi-squared test

data:  T1
X-squared = 365.89, df = 1, p-value < 2.2e-16
```

Example

```
(EXP <- CST$expected)
```

```
        survived
sex             0        1
  female 288.0015 177.9985
  male   520.9985 322.0015
```

```
(OBS <- CST$observed)
```

```
        survived
sex        0   1
  female 127 339
  male   682 161
```

```
(chi_obs <- sum((OBS - EXP)^2/EXP))
```

```
[1] 365.8869
```

Section 4

Chi-Square Tests of Homogeneity

Example

- Data will often come summarized in contingency tables.

```
DP <- c(67, 76, 57, 48, 73, 79)
MDP <- matrix(data = DP, nrow = 2, byrow = TRUE)
dimnames(MDP) <- list(Pop = c("Drug", "Placebo"),
    Status = c("Improve", "No Change", "Worse"))
TDP <- as.table(MDP)
TDP
```

```
         Status
Pop        Improve No Change Worse
  Drug          67        76    57
  Placebo       48        73    79
```

Putting the data back in a tidy format

```
library(tidyverse)
NT <- TDP %>%
  tibble::as_tibble() %>%
  uncount(n)
head(NT, 3)
```

```
# A tibble: 3 x 2
  Pop   Status
  <chr> <chr>
1 Drug  Improve
2 Drug  Improve
3 Drug  Improve
```