STT 3850 : Chi-Square Tests

Fall 2023

Appalachian State University

Section 1

Chi-Square Goodness-of-Fit Tests

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.

```r
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

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

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

```
[1] 10.9697
```

```r
# 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

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

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

```
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

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

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

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

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

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

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

```
[1] 365.8869
```

Section 3

Chi-Square Tests of Homogeneity

Example

- Data will often come summarized in contingency tables.

```r
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
```