7 – The chi-squared test

The figure below shows the density function of the χ^2-distribution for several degrees of freedom.

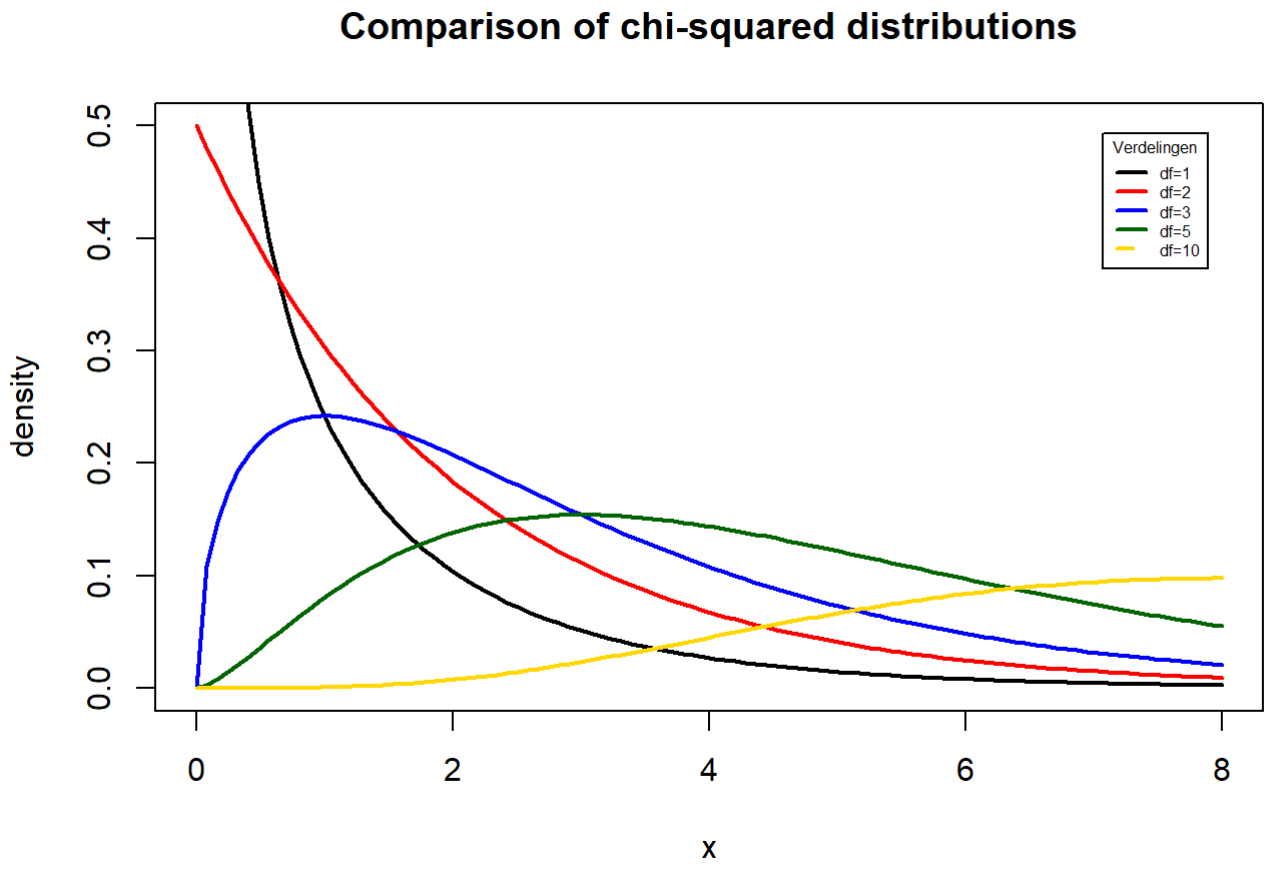
```
x <- seq(0, 8, length=100)  # x-values
degf <- c(1, 2, 3, 5, 10)   # degrees of freedom to be plotted

# line colors and legend labels
colors <- c("black", "red", "blue", "darkgreen", "gold")
labels <- c("df=1", "df=2", "df=3", "df=5", "df=10")

# plot of the distribution with df = 1
plot(x, dchisq(x, degf[1]),
     col = colors[1],
     type = "l", lwd = 2,
     ylim = c(0, .5),
     ylab = "density",
     main = "Comparison of chi-squared distributions")

# plots of the distribution with higher degrees of freedom
for (i in 2:5){
  lines(x, dchisq(x, degf[i]), lwd=2, col=colors[i])
}

legend("topright", inset=.05, title="Verdelingen",
       labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors,
       cex = .5)
```



R has several functions for calculations related to the χ^2 density function, just like those for the normal and t distributions.

- `dchisq(x, df)` : the density function
- `pchisq(x, df)` : the left-tail probability $P(\chi^2 < x)$
- `qchisq(p, df)` : the inverse of `pchisq` , "find a number x so that $P(\chi^2 < x) = p$"
- `rchisq(n, df)` : generate n random numbers for a χ^2 distribution

The parameter `df` denotes the degrees of freedom, and generally equals the sample size minus one: $n - 1$.

Goodness of fit test

In a sample of super heroes, the following types were observed:

```
types      <- c("mutant", "human", "alien", "god", "demon")
observed   <- c(   127,      75,      98,     27,     73)
expected_p <- c(   .35,     .17,     .23,    .08,    .17)
```

The question now is, is this sample representative for the population w.r.t. the different types? I.e. does each type of super hero occur in the sample proportional to the expected percentages in the population as a whole?

Test procedure

The *goodness of fit test* is suitable to answer this type of question. The procedure is as follows:

1. Formulate the hypotheses:
   - H_0: the sample is representative for the population (i.e. the proportions of each class in the sample closely matches those of the population)
   - H_0: the sample is **not** representative for the population (i.e. the proportions diverge *significantly*)
2. Determine a significance level, e.g. $\alpha = 0.05$ and the sample size

   ```
   alpha <- 0.05
   n <- sum(observed)
   expected <- n * expected_p
   expected
   ```

   ```
   ## [1] 140  68  92  32  68
   ```

3. Calculate the test statistic, in this case χ^2:

   ```
   chisq <- sum((observed - expected)^2 / expected)
   chisq
   ```

   ```
   ## [1] 3.467932
   ```

4. Determine the p-value or the critical value g. Remark that in practice, you only need to calculate **one** of the two. Both methods are equivalent.

   a. In a χ^2-test, the critical value is a number g with property $P(\chi^2 > g) = \alpha$ (where α is our chosen significance level). *Left* of g is the acceptance region, *right* of g the critical region (see the plot below). This number can be calculated with:

```
l <- length(types)
g <- qchisq(p = 1 - alpha, df = l - 1)
g
```

```
## [1] 9.487729
```

b. The p-value is given by:

```
p <- 1 - pchisq(chisq, df = l - 1)
p
```

```
## [1] 0.482771
```

5. Draw a conclusion:

   a. In the case of the critical value g:
      - if $\chi^2 < g$, accept the null hypothesis,
      - if $\chi^2 > g$, reject the null hypothesis
   b. In the case of the p-value:
      - if $p > \alpha$, accept the null hypothesis,
      - if $p < \alpha$, reject the null hypothesis

```
# Critical value $g$
paste(ifelse(chisq < g, "Accept", "Reject"), "the null hypothesis")
```

```
## [1] "Accept the null hypothesis"
```

```
# Probability value $p$
paste(ifelse(p > alpha, "Accept", "Reject"), "the null hypothesis")
```

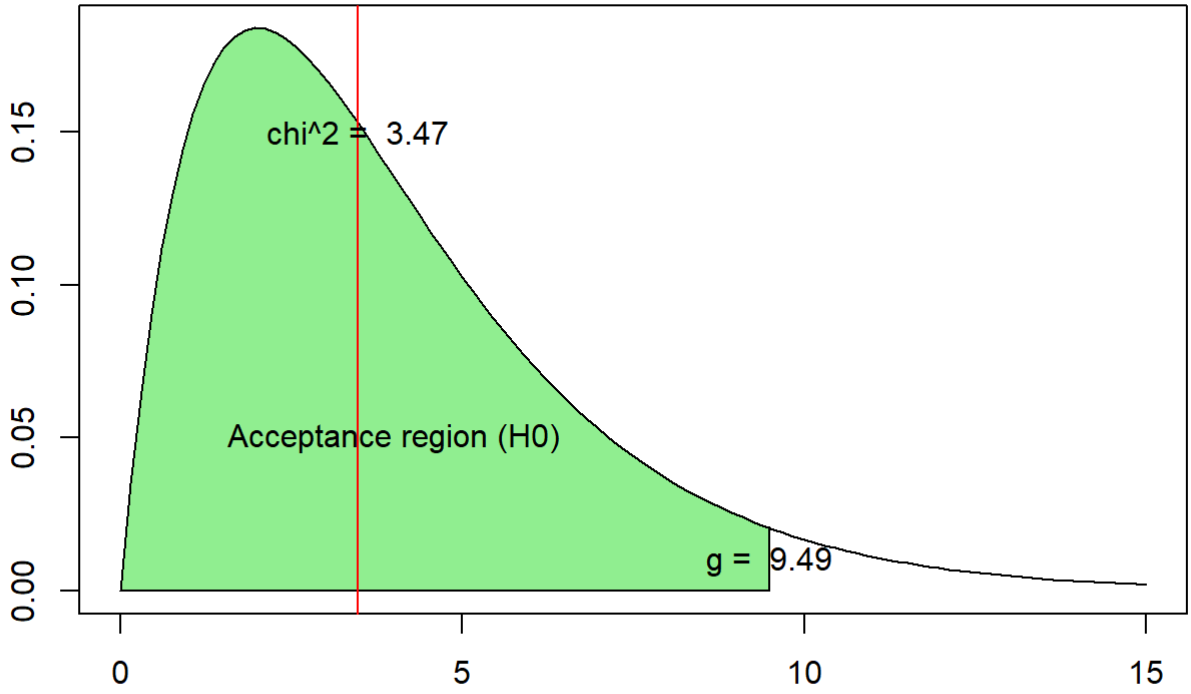
```
## [1] "Accept the null hypothesis"
```

Plot of this case

```
# Plot the chi-squared density function
x <- seq(0, 15, length = 100)
dist <- dchisq(x, df = l - 1)
plot (x, dist, type = 'l', xlab = '', ylab = '')

# The acceptance region (where H_0 is accepted)
i <- x <= g
polygon(c(x[i],    g,                    g),
        c(dist[i], dchisq(g, df = l - 1), 0),
        col = 'lightgreen')
text(x = 4, y = 0.05, 'Acceptance region (H0)')
text(x = g, y = 0.01, paste('g = ', round(g, digits=2)))

# The test statistic (chi squared)
abline(v = chisq, col = 'red')
text(x = chisq, y = 0.15, paste('chi^2 = ', round(chisq, digits=2)))
```



Using the R function `chisq.test`

The `chisq.test` function automates the entire process. You can provide it with the observed values in the sample and the expected proportions in the population, and it will perform all calculations. The most important part of the output is the p-value.

```
chisq_test_result <- chisq.test(observed, p = expected_p)
chisq_test_result
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 3.4679, df = 4, p-value = 0.4828
```

It's useful to assign the result of the `chisq.test` function to a variable, since it contains information that isn't printed:

```
chisq_test_result$statistic # The value of chi squared
```

```
## X-squared
##  3.467932
```

```
chisq_test_result$p.value   # The p-value
```

```
## [1] 0.482771
```

```
chisq_test_result$parameter # The degrees of freedom
```

```
## df
##  4
```

```
chisq_test_result$residuals # Residuals (o - e) / sqrt(e)
```

```
## [1] -1.0987005  0.8488747  0.6255432 -0.8838835  0.6063391
```

```
chisq_test_result$stdres    # Standardised residuals
```

```
## [1] -1.3627703  0.9317610  0.7128727 -0.9215122  0.6655436
```

The standardised residuals are noteworthy: they indicate the categories that are over- or underrepresented in the sample. If the value is between $[-2, 2]$, the category is considered to be well represented in the sample. Below -2, it's underrepresented, above 2, it is overrepresented.

Another example

In some research project, 1022 families with (exactly) five children are selected in a sample. The families are categorised according to the number of boys. Below, the frequencies for each category are given:

```
num_boys <- c( 0,   1,   2,   3,   4,  5)
observed <- c(58, 149, 305, 303, 162, 45)
n <- sum(observed)
```

The expected number of boys (assuming the probability of conceiving either a boy or a girl is 50%) can be calculated as follows:

```
prob_boy <- 0.5

expected_p <- choose(n = 5, k = num_boys) *
  prob_boy^num_boys *
  prob_boy^(5-num_boys)
expected_p
```

```
## [1] 0.03125 0.15625 0.31250 0.31250 0.15625 0.03125
```

```
expected <- expected_p * n
expected
```

```
## [1]  31.9375 159.6875 319.3750 319.3750 159.6875  31.9375
```

The test procedure:

1. H_0: the sample is representative; H_1: it is *not* representative
2. Choose significance level:

```
alpha <- 0.01
```

3. Calculate the test statistic

```
chisq <- sum((observed - expected)^2 / expected)
chisq
```

```
## [1] 28.84618
```

4. Calculate g or p

```
l <- length(num_boys)
g <- qchisq(p = 1 - alpha, df = l - 1)
g
```

```
## [1] 15.08627
```

```
p <- 1 - pchisq(chisq, df = l - 1)
p
```

```
## [1] 2.48555e-05
```

5. Draw a conclusion:

```
# Critical value $g$
paste(ifelse(chisq < g, "Accept", "Reject"), "the null hypothesis")
```

```
## [1] "Reject the null hypothesis"
```

```
# Probability value $p$
paste(ifelse(p > alpha, "Accept", "Reject"), "the null hypothesis")
```

```
## [1] "Reject the null hypothesis"
```

In order to find out which categories of families are under- or overrepresented, take a look at the standardized residuals:

```
stres <- (observed - n * expected_p) /
  sqrt(n * expected_p * (1 - expected_p))
stres
```

```
## [1]  4.6855411 -0.9207331 -0.9701101 -1.1050820  0.1992230  2.3483887
```

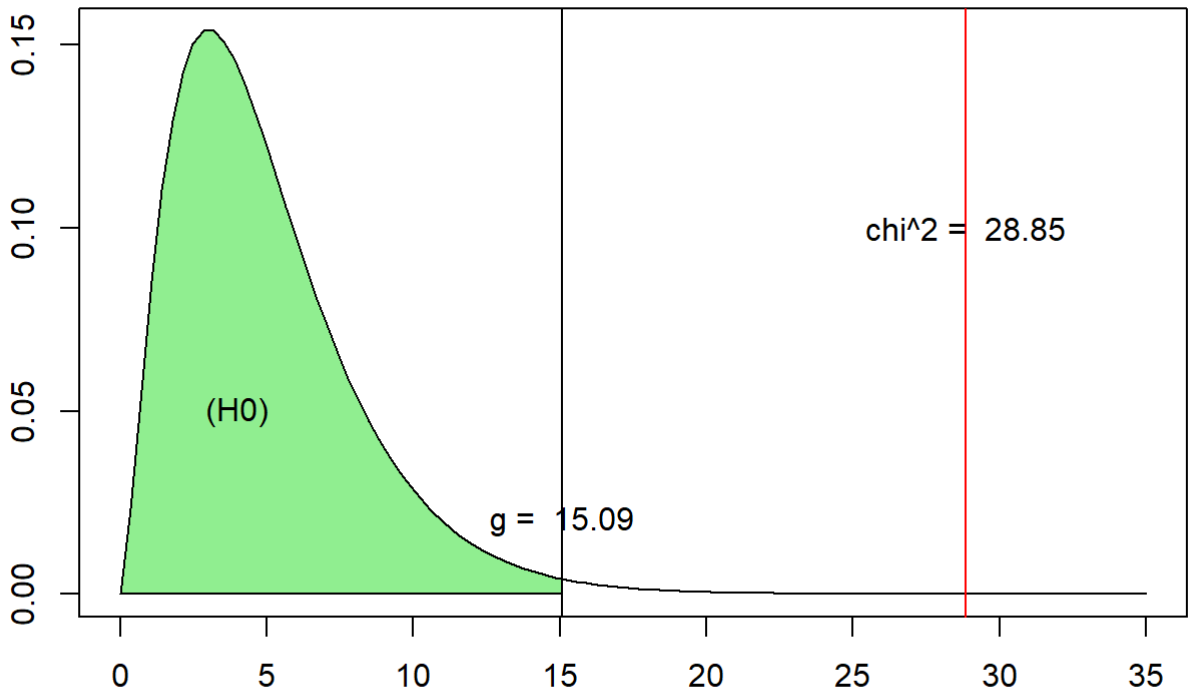
The standardised residuals for the families with either only girls (~4.69) or only boys (~2.35) are outside of the interval [-2, 2]. Consequently, these families are overrepresented in the sample.

Plot of this case

```
# Plot the chi-squared density function
x <- seq(0, 35, length = 100)
dist <- dchisq(x, df = l - 1)
plot (x, dist, type = 'l', xlab = '', ylab = '')

# The acceptance region (where H_0 is accepted)
i <- x <= g
polygon(c(x[i],    g,                      g),
        c(dist[i], dchisq(g, df = l - 1), 0),
        col = 'lightgreen')
abline(v = g)
text(x = 4, y = 0.05, '(H0)')
text(x = g, y = 0.02, paste('g = ', round(g, digits=2)))

# The test statistic (chi squared)
abline(v = chisq, col = 'red')
text(x = chisq, y = 0.1, paste('chi^2 = ', round(chisq, digits=2)))
```



Chi-squared test for categorical data

With this variant of the Chi-squared test, we can determine whether two categorical (i.e. qualitative) variables are associated.

As an example, we take a study by Doll and Hill, who in 1951 conducted a survey among British general practicioners with the request for data about their age and whether they smoked or not. They then followed up on the respondents for years and recorded whether they died of lung cancer or not.

The results of the survey are given in the following table:

```
doll_hill <- matrix(data = c(21178, 83, 3092, 1),
                    ncol = 2,
                    byrow = TRUE,
                    dimnames = list(c("Smoker", "Non-smoker"),
                                c("No Cancer", "Cancer")))
doll_hill
```

```
##            No Cancer Cancer
## Smoker         21178     83
## Non-smoker      3092      1
```

The research question now is: “is there an association with smoking and dying of lung cancer?”

```
doll_hill_result <- chisq.test(doll_hill)
doll_hill_result
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  doll_hill
## X-squared = 9.0565, df = 1, p-value = 0.002618
```

```
doll_hill_result$expected
```

```
##            No Cancer   Cancer
## Smoker     21187.668 73.33186
## Non-smoker  3082.332 10.66814
```

```
doll_hill_result$stdres
```

```
##            No Cancer    Cancer
## Smoker     -3.173528  3.173528
## Non-smoker  3.173528 -3.173528
```