Chi-Squared Goodness of Fit Tests

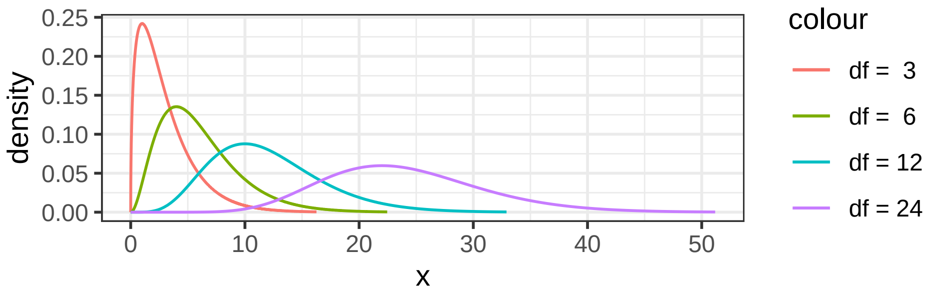
Stat 145

The Chi-Squared Distribution

When sample sizes are large enough, the shape of the null distribution for the chi-squared statistic is approximately a **chi-squared distribution**. Here are some important facts about this family of distributions.

- Here are a few examples of chi-squared distributions.

```
gf_dist("chisq", df = 3, color = ~ "df =  3") %>%
  gf_dist("chisq", df = 6, color = ~"df =  6") %>%
  gf_dist("chisq", df = 12, color = ~ "df = 12") %>%
  gf_dist("chisq", df = 24, color = ~ "df = 24")
```



- Unlike the normal distributions, which are determined by two parameters (mean and standard deviation), a chi-squared distribution is determined by just one, called **degrees of freedom** and abbreviated **df**.
- Chi-squared distributions are **skewed** and **never take on negative values**.
- They become more symmetric when the degrees of freedom is larger and for very large degrees of freedom, they are approximately normal.
- The mean value of a chi-squared distribution is the degrees of freedom.
- We can use `pchisq()` and `qchisq()` just like `pnorm()` and `qnorm()`.

Comparing the test statistic to a Chi-squred distribution

Golf balls, revisited In our golf balls example, the null distribution is approximately a chi-squared distribution with 3 (one less than the number of cells in our table) degrees of freedom.

```
set.seed(1234)
golfballs <- as.table(c(137, 138, 107, 104));
dimnames(golfballs) <- list(1:4)
golfballs
```

```
##   1   2   3   4
## 137 138 107 104
```

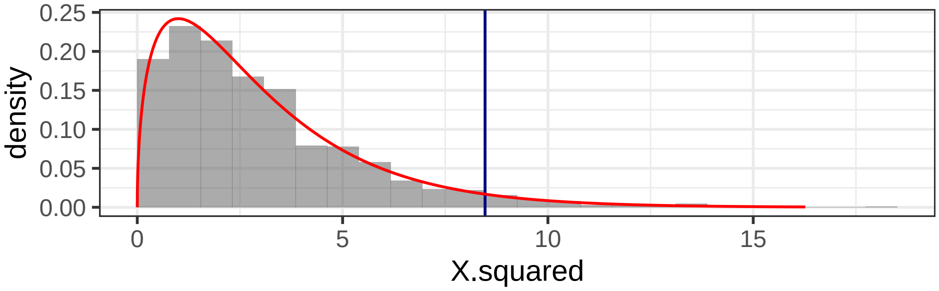
```
chisq(golfballs)
```

```
## X.squared
##  8.469136
```

```
Golf_rand <- do(2000) * chisq(tally( ~ resample(1:4, size = 486)))
head(Golf_rand, 3)
```

```
##   X.squared
## 1 8.7818930
## 2 2.3292181
## 3 0.3539095
```

```
gf_dhistogram( ~ X.squared, data = Golf_rand, boundary = 0) %>%
  gf_vline(xintercept = ~ 8.469, color = "navy") %>%
  gf_dist('chisq', df = 3, color = "red")
```



```
prop( ~(X.squared >= 8.469), data = Golf_rand)
```

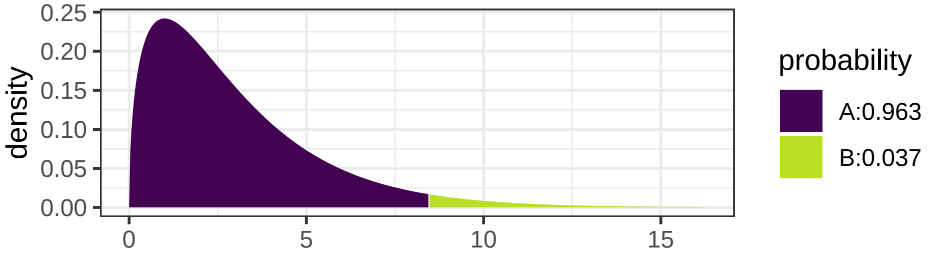
```
## prop_TRUE
##     0.036
```

```
1 - pchisq(8.469, df = 3)
```

```
## [1] 0.03725096
```

```
# illustrated version
1 - xpchisq(8.469, df = 3)
```

```
## [1] 0.03725096
```



As we see, we get essentially the same p-value either way.

Rule of Thumb The results should be quite similar to the randomization results as long as the sample is large enough. The usual rule of thumb is that the expected counts should be at least 5 in each cell.[1] Our expected counts are much larger than this.

[1] Two notes: First, if you use `chisq.test()`, it will warn you if it thinks the p-value might not be very accurate. In those situations you should use `simulate.p.value = TRUE`. Second, the rule of thumb isn't magic. As always, bigger is better: The larger the expeted counts are, the more accurate the approximation is. But there is no hard cut-off between large enough and not large enough. If there five or ore cells and most of the expected counts are above 5, it is probably OK if one of then is a bit smaller.

Using chisq.test()

The `chisq.test()` function can do all this work for us.

```r
# randomization method
chisq.test(c(137, 138, 107, 104), p = c(0.25, 0.25, 0.25, 0.25),
           simulate.p.value = TRUE)
```

```
##
##  Chi-squared test for given probabilities with simulated p-value (based
##  on 2000 replicates)
##
## data:  c(137, 138, 107, 104)
## X-squared = 8.4691, df = NA, p-value = 0.03848
```

```r
# theory method
chisq.test(c(137, 138, 107, 104), p = c(0.25, 0.25, 0.25, 0.25))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(137, 138, 107, 104)
## X-squared = 8.4691, df = 3, p-value = 0.03725
```

You can get even more information if you use `xchisq.test()`

```r
xchisq.test(c(137, 138, 107, 104), p = c(0.25, 0.25, 0.25, 0.25))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  x
## X-squared = 8.4691, df = 3, p-value = 0.03725
##
##  137.00   138.00   107.00   104.00
## (121.50) (121.50) (121.50) (121.50)
##  [1.98]   [2.24]   [1.73]   [2.52]
## < 1.41>  < 1.50>  <-1.32>  <-1.59>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

Why 3 degrees of freedom? If we know the sample size (the number of golf balls we collect) and we know any 3 of the cell counts in the table, then the last cell count is completely determined, it is not free. We have only 3 **degrees of freedom**.

Example:

Suppose we collect a new data set. This time there are only 100 golf balls. In each table below, determine the missing observed count.

number on golf ball	1	2	3	4
observed count	20	30	25	?

number on golf ball	1	2	3	4
observed count	31	28	?	20

number on golf ball	1	2	3	4
observed count	?	28	24	20

Goodness of Fit Practice

Skill building

1. In each example below, use the test statistic and degrees of freedom to compute a p-value using the chi-squared distribution.

   a. $X^2 = 3.5$; df $= 3$.
   b. $X^2 = 7.2$; df $= 3$.
   c. $X^2 = 7.2$; df $= 2$.
   d. $X^2 = 15.52$; df $= 5$.
   e. $X^2 = 15.52$; df $= 10$.

2. How large is the chi-squared statistic if df $= 4$ and the p-value is 0.05?

3. How large is the chi-squared statistic if df $= 2$ and the p-value is 0.05?

4. Answer the questions below about this use of `xchisq.test()`.

```r
xchisq.test(c(53, 108, 39), p = c(0.25, 0.5, 0.25))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  x
## X-squared = 3.24, df = 2, p-value = 0.1979
##
##    53.00   108.00    39.00
## ( 50.00) (100.00) ( 50.00)
##  [0.18]   [0.64]   [2.42]
## < 0.42>  < 0.80>  <-1.56>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

   a. What is the null hypothesis for this test?
   b. Why is df $= 2$?
   c. Explain how to compute the number 0.64 that appears in the output.
   d. Show how to calculate the p-value using `pchisq()` or `xpchisq()`.
   e. Square each of the Pearson residuals. What do you observe? (Note: residuals here are defined differently from regression residuals.)

More Examples

We now know 4 ways to do each of the examples below.

- Randomization
  - Using `resample()` to create a randomization distribution and using that to get your p-value.
  - Using `chisq.test()` or `xchisq.test()` with `simulate.p.value = TRUE`.
- Theoretical
  - "By hand" using the chi-squared distribution – computing the expected counts and chi-squared statistic and using `pchisq()` or `xpchisq()` to get the p-value.
  - Using `chisq.test()` or `xchisq.test()` to do all the work for you. (This is the default method for `chisq.test()` and `xchisq.test()`.)

Use the examples below to practice each method. If you use the theoretical method, make sure that our rule of thumb is satisfied. You don't necessarily need to do each method on each problem, but be sure you know how to do each method. In practice, we will prefer to use `chisq.test()` or `xchisq.test()` because they are so much easier to use and provide us all the information we need.

5. Plant Genetics. According to a genetic model, the ratio of colors of seed coats should be 12:3:1. The observed colors are given in the table below. Is this data compatible with this particular genetic theory?

Seed coat color	black	brown	pale
Count	321	77	31

6. More Plant Genetics. Below are some data from one of Mendel's original studies of pea plants.

Phenotype	yellow, round	yellow, wrinkled	green, round	green, wrinkled
Count	315	101	108	32

According to his theory of independent assortment (Mendel's second law), the phenotype ratios should be 9:3:3:1. Is this data compatible with this particular genetic theory?

7. Birth Days. A sample of 700 random births revealed the following data regarding the day of the week on which the birth occurred. Is there evidence that some days are more common than others or might the differences in our table simply be due to sampling variability?

Day	Sun	Mon	Tue	Wed	Thr	Fri	Sat
Births	84	110	124	104	94	112	72

8. One proportion or Chi-squared? Fifty subjects were given two unmarked cups of coffee and asked which one they preferred. 31 preferred the fresh brewed coffee and 19 preferred the instant coffee.

Is this evidence that a majority of people prefer fresh brewed coffee? Conduct the hypothesis test two ways.

- Using a test for one proportion
- Using a chi-squared goodness of fit test

How do the results compare?

9. Frizzled Feathers The Frizzle fowl is a variety of chicken with curled feathers. In a 1930 experiment, Launder and Dunn crossed Frizzle fowls with with the Leghorn variety which has straight feathers. The first generation (F1) produced all slightly frizzled chicks. This made the researchers suspect a co-dominant genetic model. To test this, they interbred F1 to get F2 chicks and recorded the feather type for each chick.

frizzled	slightly frizzled	straight
23	50	20

The codominant model predicts a 1:2:1 ratio. Why?

Use these data to assess the genetic model.