DATA2002

Testing in small samples

Garth Tarr

THE UNIVERSITY OF SYDNEY

Fisher's exact test

Yate's corrected χ^2 test

Monte Carlo p-values

Fisher's exact test

Lady tasting tea

Given a cup of tea with milk, a lady claims she can discriminate as to whether milk or tea was first added to the cup.

How could we test this claim?

What information would we need?

Lady tasting tea

Fisher proposed a preparing 8 cups of tea

- 4 cups where tea was added before milk
- 4 cups where milk was added before tea

The lady would then be randomly given the cups of tea and asked to identify the 4 where tea was added before milk.

We would then need to record:

- Which cups had tea or milk added first (**truth**).
- Which cups the lady claimed had tea or milk added first (**predicted**).

Ronald Fisher (1913)

Lady tasting tea - hypothesis

For Fisher's experiment we were left with two categorical variables.

$$\text{Truth} = \{\text{Milk}, \text{Tea}, \text{Tea}, \text{Milk}, \text{Tea}, \text{Tea}, \text{Milk}, \text{Milk}\}$$

$$\text{Prediction} = \{\text{Milk}, \text{Tea}, \text{Tea}, \text{Milk}, \text{Tea}, \text{Tea}, \text{Milk}, \text{Milk}\}$$

- And the hypothesis

$$\text{Are the predictions independent of the truth?}$$

```r
truth = c("milk", "tea", "tea", "milk", "tea", "tea", "milk", "milk")
predicted = c("milk", "tea", "tea", "milk", "tea", "tea", "milk", "milk")
y.mat = table(truth, predicted)
y.mat
```

```
##       predicted
## truth  milk tea
##   milk    4   0
##   tea     0   4
```

```r
chisq.test(y.mat, correct = FALSE)
```

```
## Warning in chisq.test(y.mat, correct = FALSE): Chi-squared
## approximation may be incorrect
```

```
##
##     Pearson's Chi-squared test
##
## data:  y.mat
## X-squared = 8, df = 1, p-value = 0.004678
```

Why do we get a warning message?

Fisher's exact test

The χ^2 approximation for the test statistic is only reasonable when n is sufficiently large. I.e. we need the expected cell frequencies to all be 5 or more. However, if this is not the case, then we need to take care and maybe consider **exact** tests, i.e. calculating the exact p-value for the test statistic.

In R the function `fisher.test()` is available to carry out these calculations both for 2×2 tables and general contingency tables.

Fisher's exact test and the hypergeometric distrubution

The simplest exact test for contingency tables is Fisher's test for 2×2 tables. Consider the table:

	A_1	A_2	**Total**
B_1	y_{11}	y_{12}	$y_{1\bullet}$
B_2	y_{21}	y_{22}	$y_{2\bullet}$
Total	$y_{\bullet 1}$	$y_{\bullet 2}$	n

For a 2×2 table, if we know the row and column and y_{11} then the table is completely specified.

Let θ be the odds ratio. A test of

$$H_0\colon\ \theta = 1 \quad \text{vs} \quad H_1\colon\ \theta > 1$$

(or $H_1\colon\ \theta < 1$) can be based on the observed value of y_{11} given the marginal totals.

If H_0 is true and we know the $y_{1\bullet}, y_{\bullet 1}$ and n values we expect $y_{\bullet 1} \times \frac{y_{1\bullet}}{n}$ in the $(1,1)$th cell.

To obtain the distribution of y_{11} given the marginal values note the situation is like selecting $y_{\bullet 1}$ values from n where $y_{1\bullet}$ are type B_1 and $y_{2\bullet}$ are type B_2. Then

$$P(Y_{11} = y_{11}) = \frac{\binom{y_{1\bullet}}{y_{11}}\binom{y_{2\bullet}}{y_{\bullet 1} - y_{11}}}{\binom{n}{y_{\bullet 1}}},$$

which is the hypergeometric distribution.

The hypergeometric distribution

The hypergeometric distribution relates to sampling without replacement from a finite population.

The following conditions characterise the **hypergeometric distribution**:

- The result of each draw (the elements of the population being sampled) can be classified into one of two mutually exclusive categories (e.g. Pass/Fail or Employed/Unemployed).
- The probability of a success changes on each draw, as each draw decreases the population (sampling without replacement from a finite population).

A random variable X follows the hypergeometric distribution if its probability mass function (pmf) is given by:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where

- N is the population size,
- K is the number of success states in the population,
- n is the number of draws (i.e. quantity drawn in each trial), and
- k is the number of observed successes.

p-values

To calculate the p-value for a particular table we need to:

- enumerate all tables, as extreme, or more extreme than the observed table **with the same marginal totals**; and
- sum up the probability of each of these tables.

Lady tasting tea

```
truth = c("milk", "tea", "tea", "milk", "tea", "tea", "milk", "milk")
predicted = c("milk", "tea", "tea", "milk", "tea", "tea", "milk", "milk")
y.mat = table(truth, predicted)
y.mat
```

```
##       predicted
## truth  milk tea
##   milk    4   0
##   tea     0   4
```

For Fisher's exact test we:

1. Consider all possible permutations of the 2×2 contingency table with the same *marginal totals* (in this case $y_{i\bullet} = y_{\bullet j} = 4$).
2. Calculate how many of these were equal to or "more extreme" than what we observed.

Truth \ Predicted	Milk	Tea	
Milk	4	0	$y_{1\bullet} = 4$
Tea	0	4	$y_{2\bullet} = 4$
	$y_{\bullet 1} = 4$	$y_{\bullet 2} = 4$	$y_{\bullet\bullet} = n = 8$

How do we define more extreme?

Let us define a test statistic

$$T = \text{number of cups of tea before milk that she got correct.}$$

This test statistic has 5 outcomes: $\{0, 1, 2, 3, 4\}$.

Given that there are 8 cups of tea, there are $\binom{8}{4} = 70$ ways that we could predict which cups had tea added before milk.

We can look at all 70 permutations of prediction vs truth and calculate how often we see a test statistic of $0, 1, 2, 3$ or 4.

t_i	0	1	2	3	4	
f_i	$\binom{4}{0}\binom{4}{4} = 1$	$\binom{4}{1}\binom{4}{3} = 16$	$\binom{4}{2}\binom{4}{2} = 36$	$\binom{4}{3}\binom{4}{1} = 16$	$\binom{4}{4}\binom{4}{0} = 1$	70
p_i	$\frac{1}{70}$	$\frac{16}{70}$	$\frac{36}{70}$	$\frac{16}{70}$	$\frac{1}{70}$	1

$P(T = 4) = \frac{1}{70} = 0.014$

```
fisher.test(y.mat)
```

```
##
##     Fisher's Exact Test for Count Data
##
## data:  y.mat
## p-value = 0.02857
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.339059      Inf
## sample estimates:
## odds ratio
##        Inf
```

```
fisher.test(y.mat, alternative = "greater")
```

```
##
##     Fisher's Exact Test for Count Data
##
## data:  y.mat
## p-value = 0.01429
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  2.003768      Inf
## sample estimates:
## odds ratio
##        Inf
```

Cancer of the larynx

Mendenhall, Million, Sharkey, et al. (1984) report the results of a study comparing radiation therapy with surgery in treating cancer of the larynx.

	Cancer controlled	Cancer not controlled	Total
Surgery	21	2	23
Radiation therapy	15	3	18
Total	36	5	41

Suppose that we wish to test $H_0\colon\ \theta = 1$ (both treatments equally effective) against $H_1\colon\ \theta > 1$ (surgery more effective).

Cancer of the larynx

First we need to enumerate all tables which are as extreme or more extreme than the observed table. These are:

The original table:

	Cancer controlled	Cancer not controlled	Total
Surgery	21	2	**23**
Radiation therapy	15	3	**18**
Total	**36**	**5**	**41**

And the tables where surgery controlled more than 21 (i.e. 22 or 23) patients, holding the **margins** constant.

	Cancer controlled	Cancer not controlled	Total
Surgery	22	1	**23**
Radiation therapy	14	4	**18**
Total	**36**	**5**	**41**

	Cancer controlled	Cancer not controlled	Total
Surgery	23	0	**23**
Radiation therapy	13	5	**18**
Total	**36**	**5**	**41**

Cancer of the larynx

Let X be the number of surgery cases where cancer is controlled. Applying Fisher's approach

$$
\begin{aligned}
\text{p-value} &= P(X \geq 21 \mid \text{marginal totals}) \\
&= P(X = 21, 22, 23 \mid \text{marginal totals}) \\
&= P(X = 21 \mid \text{marginal totals}) + P(X = 22 \mid \text{marginal totals}) + P(X = 23 \mid \text{marginal totals}) \\
&= \frac{\binom{23}{21}\binom{18}{15}}{\binom{41}{36}} + \frac{\binom{23}{22}\binom{18}{14}}{\binom{41}{36}} + \frac{\binom{23}{23}\binom{18}{13}}{\binom{41}{36}} \\
&= 0.3808.
\end{aligned}
$$

```
y_mat = matrix(c(21, 15, 2, 3), ncol = 2)
colnames(y_mat) = c("Controlled", "Not controlled")
rownames(y_mat) = c("Surgery", "Radiation therapy")
y_mat
```

```
##                    Controlled Not controlled
## Surgery                    21              2
## Radiation therapy          15              3
```

```
fisher.test(y_mat, alternative = "greater")
```

```
##
##     Fisher's Exact Test for Count Data
##
## data:  y_mat
## p-value = 0.3808
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.2864828       Inf
```

Drawbacks

Why don't we use Fisher's exact test all the time?

- It assumes that row and column margins are fixed.
- Computationally difficult for large samples.
- It can be generalized to $r \times c$ two-way contingency tables but is very difficult to compute. Generally requires use of Monte Carlo (i.e. random permutation).

$\mathrm{X} \setminus \mathrm{Y}$	y_1	y_2	y_3	Row total
x_1	a	b	c	$a+b+c$
x_2	d	e	f	$d+e+f$
Row total	$a+d$	$b+e$	$c+f$	$n=a+b+c+d+e+f$

Yates' chi-squared test

Yates' Corrected χ^2 Test

Yates (1934) modified the standard chi-squared test with a continuity correction. It is usually more accurate when counts in each cell are small. Yates' statistic for 2×2 tables is:

$$T = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|Y_{ij} - e_{ij}| - 0.5)^2}{e_{ij}}$$

which approximately follows a χ^2_1 distribution under H_0.

Logic behind continuity corrections

In general, if we have an *integer-valued* random variable X which we would like to approximate with a continuous random variable Y then

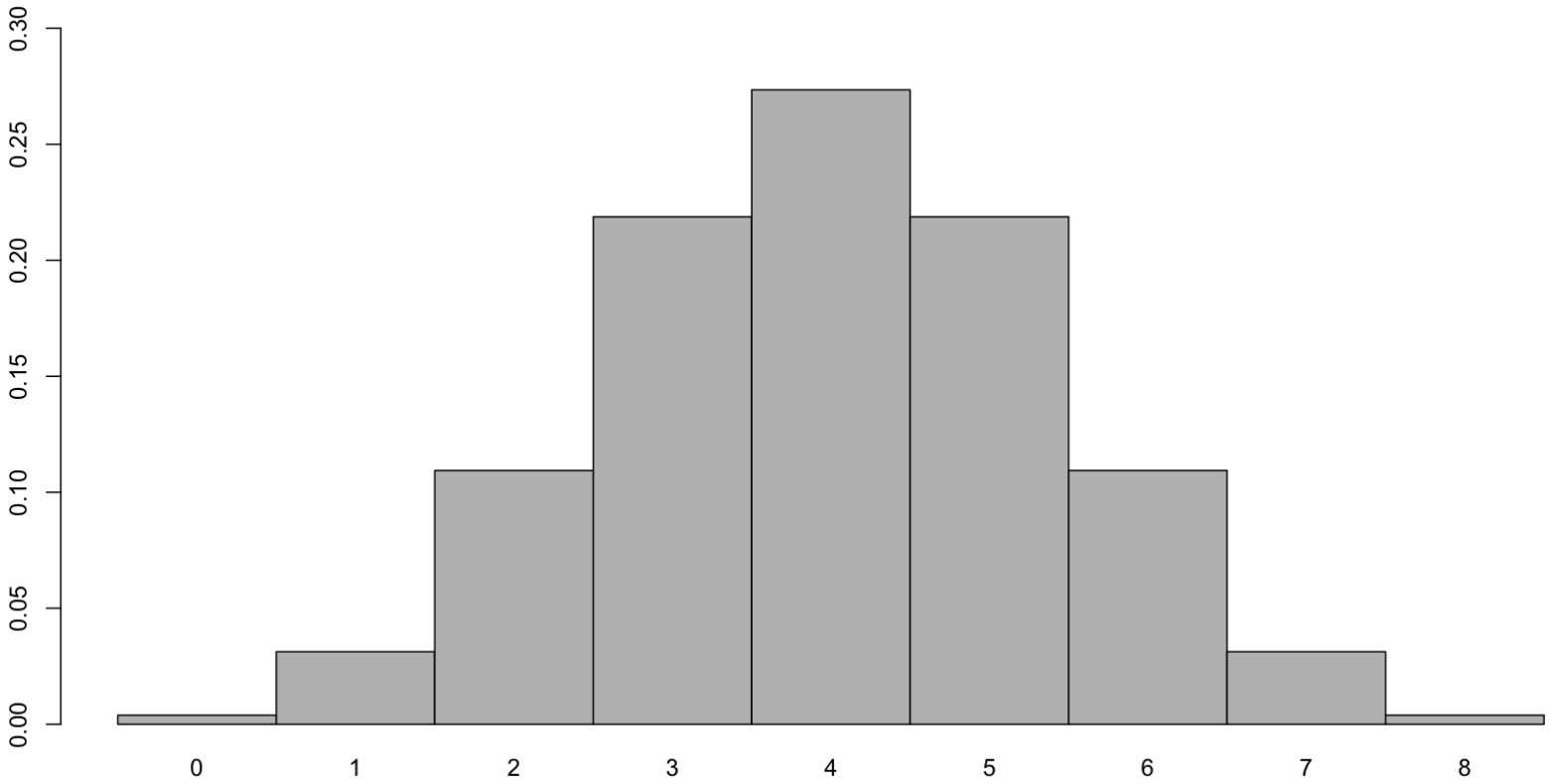
$$P(X \leq x) \approx P(Y \leq x + 0.5)$$

and

$$P(X \geq x) \approx P(Y \geq x - 0.5)$$

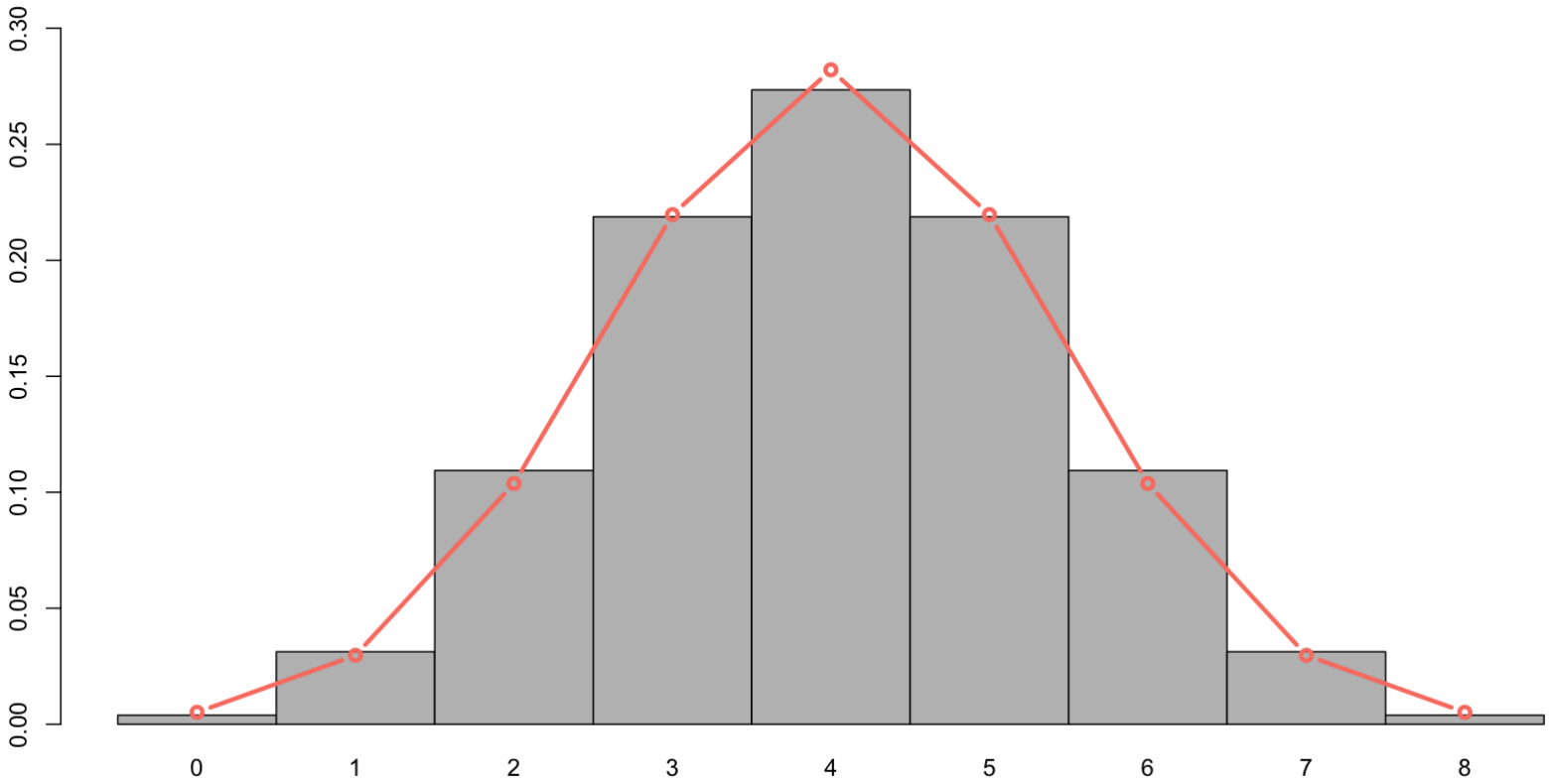
Continuity correction

Below is the distribution for a $B(n, p)$.



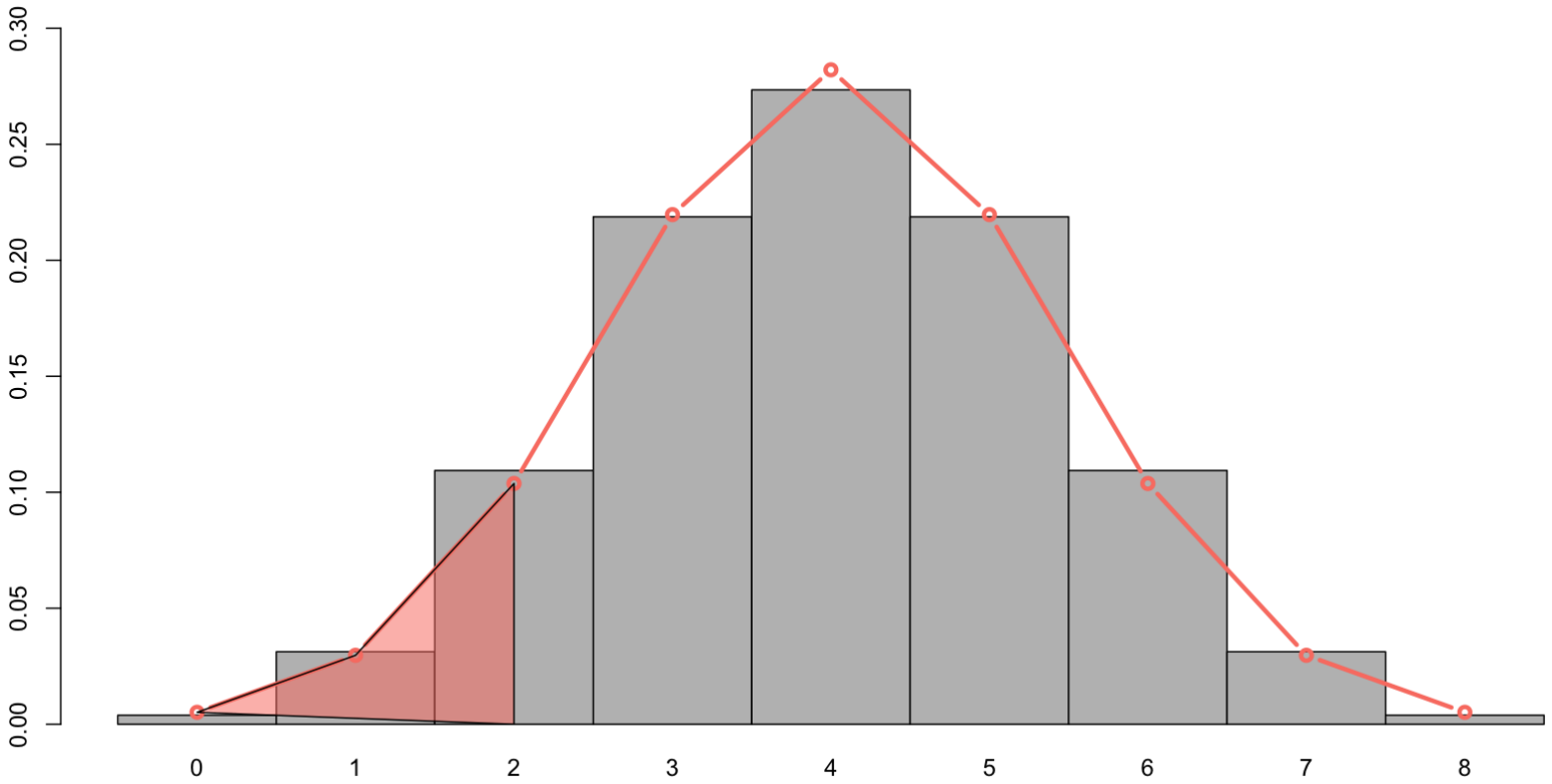
Continuity correction

We can approximate a $X \sim B(n, p)$ with a $Y \sim N(np, np(1-p))$.



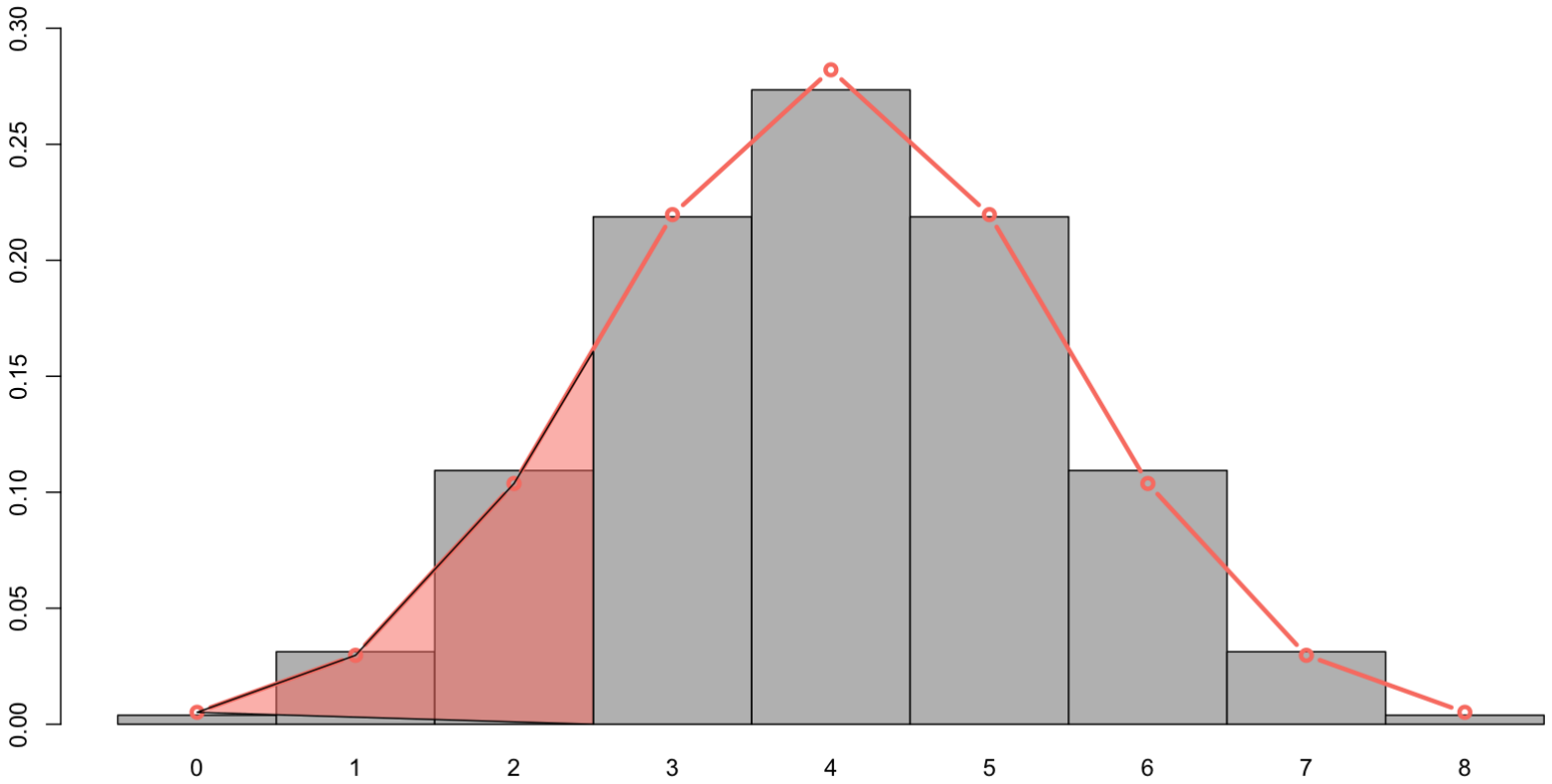
Continuity correction

We can estimate $P(X \leq 2)$ with $P(Y \leq 2)$.



Continuity correction

But $P(Y \leq 2.5)$ is better.



```
(y.mat = table(truth, predicted))
```

```
##       predicted
## truth  milk tea
##   milk    4   0
##   tea     0   4
```

```
r = c = 2
yr = apply(y.mat, 1, sum)  # Or try rowSums()
yc = apply(y.mat, 2, sum)  # Or try colSums()
yr.mat = matrix(yr, r, c, byrow = FALSE)
yc.mat = matrix(yc, r, c, byrow = TRUE)
# ey.mat = yr%*%t(yc)/n
(ey.mat = yr.mat * yc.mat/sum(y.mat))
```

```
##      [,1] [,2]
## [1,]    2    2
## [2,]    2    2
```

```
(res.yates = (abs(y.mat-ey.mat)-0.5)^2/ey.mat)
```

```
##       predicted
## truth   milk   tea
##   milk 1.125 1.125
##   tea  1.125 1.125
```

```
(t0 = sum(res.yates))
```

```
## [1] 4.5
```

```
#Calculate p-values
pchisq(t0, 1, lower.tail = FALSE)
```

```
## [1] 0.03389485
```

Traditional chi-squared test

```
chisq.test(y.mat, correct = FALSE)
```

```
## Warning in chisq.test(y.mat, correct = FALSE): Chi-squared
## approximation may be incorrect

##
##     Pearson's Chi-squared test
##
## data:  y.mat
## X-squared = 8, df = 1, p-value = 0.004678
```

Yates' continuity correction

```
chisq.test(y.mat, correct = TRUE)
```

```
## Warning in chisq.test(y.mat, correct = TRUE): Chi-squared
## approximation may be incorrect

##
##     Pearson's Chi-squared test with Yates' continuity
##     correction
##
## data:  y.mat
## X-squared = 4.5, df = 1, p-value = 0.03389
```

Fisher's exact test

```
fisher.test(y.mat)
```

```
##
##     Fisher's Exact Test for Count Data
##
## data:  y.mat
## p-value = 0.02857
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.339059      Inf
## sample estimates:
## odds ratio
##        Inf
```

Permutation testing

Fingerprints

The study of Galton (1892) marked one of the first formal statistical examinations of association for contingency tables. His work involved determining the association of fingerprint characteristics of 105 fraternal (or dizygotic) male twins. One male twin was "earmarked" as twin A and his brother was "earmarked" as twin B. For each twin, the number of Arches, Loops and Whorls was counted and summarised in the 3×3 contingency table of Galton (1892, Table XXII, pg. 175).

TABLE XXII.

Observed Fraternal Couplets.

B children.	A children.			Totals in B children.
	Arches.	Loops.	Whorls.	
Arches . . .	5	12	2	*19*
Loops . . .	4	42	15	*61*
Whorls . . .	1	14	10	*25*
Totals in A children	*10*	*68*	*27*	105

```r
galton.dat <- matrix(c(5, 4, 1, 12, 42, 14, 2, 15, 10), 3, 3)
rownames(galton.dat) = c("Arches-B", "Loops-B", "Whorls-B")
colnames(galton.dat) = c("Arches-A", "Loops-A", "Whorls-A")
galton.dat
```

```
##           Arches-A Loops-A Whorls-A
## Arches-B         5      12         2
## Loops-B          4      42        15
## Whorls-B         1      14        10
```

```r
chisq.test(galton.dat)
```

```
## Warning in chisq.test(galton.dat): Chi-squared approximation
## may be incorrect

##
##     Pearson's Chi-squared test
##
## data:  galton.dat
## X-squared = 11.17, df = 4, p-value = 0.02472
```

Monte Carlo simulation

The **Monte Carlo simulation procedure** is as follows:

- Analyse the sample as one would normally do in a hypothesis test (up to, and including, the calculation of the test statistic)
- From the original sample being analysed, resample it LOTS of times
- The test statistic of interest is calculated for each of the resamples (so that we have the sampling distribution of the test statistic)
- This leads to LOTS of test statistics that will be used to calculate p-values for the observed statistic.

Monte Carlo p-values are calculated by determining the proportion of the resampled test statistics as or more extrame than the observed test statistic.

No assumptions are made about the underlying distribution of the population.

Fingerprints

Monte Carlo p-values may be obtained by randomly generating contingency tables given that the margins are assumed fixed.

To randomly generate a contingency table with the same margins as the original table we use the `r2dtable` function in R.

```
rcounts = rowSums(galton.dat)
ccounts = colSums(galton.dat)
B = 10000
set.seed(123)
x_list = r2dtable(B, rcounts, ccounts)
```

`r2dtable()` generates a list of random 2-way tables given marginals. If we consider the first element in the list we can perform a chi-square test for independence on the first random table.

```
x_list[[1]]
```

```
##      [,1] [,2] [,3]
## [1,]    2   10    7
## [2,]    7   43   11
## [3,]    1   15    9
```

```
chisq.test(x_list[[1]])
```

```
## Warning in chisq.test(x_list[[1]]): Chi-squared
## approximation may be incorrect

##
##     Pearson's Chi-squared test
##
## data:  x_list[[1]]
## X-squared = 5.2367, df = 4, p-value = 0.2639
```

Here a test statistic for the first random sample is 5.24.

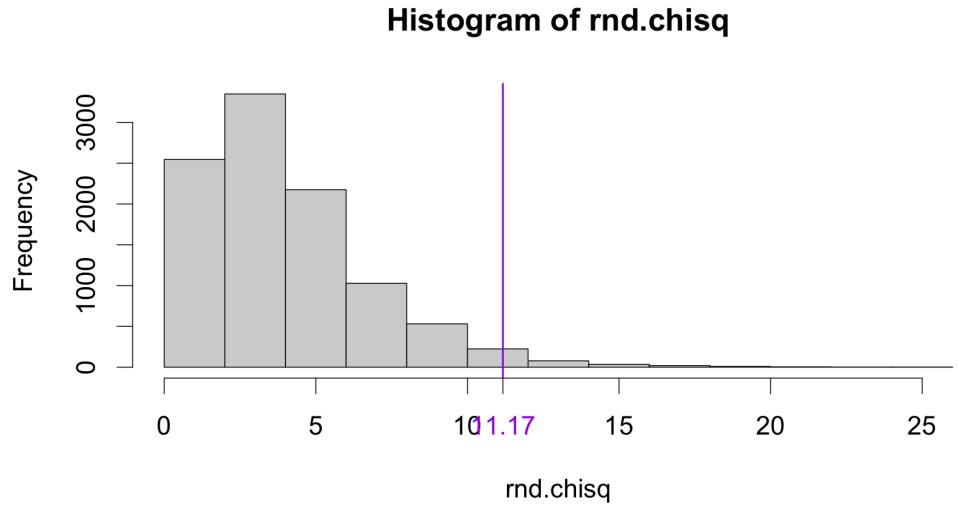
For each of the 10,000 randomly generated contingency tables, we can record their test statistic then determine what proportion of them are equal to (or exceed) the observed test statistic.

```
rnd.chisq = numeric(B)
for (i in 1:B){
  rnd.chisq[i] = chisq.test(x_list[[i]])$statis
}
sum(rnd.chisq > 11.1699)/B
```

```
## [1] 0.022
```

Here, the Monte Carlo p-value is 0.022 (comparable to theoretical p-value of 0.0247).

```
par(cex = 1.8)
hist(rnd.chisq)
abline(v = 11.17, col = "purple", lwd = 2)
axis(1, 11.17, col.axis = "purple")
```



`chisq.test()` can calculate Monte Carlo p-values and does so using `r2dtable` internally. You can do this by specifying the parameter `simulate.p.value = TRUE`

```
chisq.test(galton.dat, simulate.p.value = TRUE)
```

```
##
##     Pearson's Chi-squared test with simulated p-value
##     (based on 2000 replicates)
##
## data:  galton.dat
## X-squared = 11.17, df = NA, p-value = 0.01999
```

By default, R randomly generates 2000 contingency tables and from each calculates the test statistic, thereby producing a Monte Carlo p-value. We can use 10000 resamples by specifying

```
chisq.test(galton.dat, simulate.p.value = TRUE, B = 10000)
```

```
##
##     Pearson's Chi-squared test with simulated p-value
##     (based on 10000 replicates)
##
## data:  galton.dat
## X-squared = 11.17, df = NA, p-value = 0.0232
```

Notice that the degrees of freedom are `NA` because the test statistic is not being compared against any theoretical distribution rather, the reported p-value is calculated by comparing the test statistics from the simulated samples to the original observed test statistic.

References

MacMahon, B., P. Cole, T. M. Lin, C. R. Lowe, A. P. Mirra, B. Ravnihar, E. J. Salber, V. G. Valaoras, and S. Yuasa (1970). "Age at first birth and breast cancer risk". In: *Bulletin of the World Health Organization* 43.2, pp. 209-221. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2427645/.

Mendenhall, W. M., R. R. Million, D. E. Sharkey, and N. J. Cassisi (1984). "Stage T3 squamous cell carcinoma of the glottic larynx treated with surgery and/or radiation therapy". In: *International Journal of Radiation Oncology Biology Physics* 10.3, pp. 357-363. DOI: 10.1016/0360-3016(84)90054-3.

Yates, F. (1934). "Contingency tables involving small numbers and the χ^2 test". In: *Supplement to the Journal of the Royal Statistical Society* 1.2, pp. 217-235. DOI: 10.2307/2983604.