Chi-squared tests

Decisions with Data | Inference for Frequencies

STAT5002

The University of Sydney

May 2025

CRICOS 00026A TEQSA PRV12057

Decisions with Data

Topics 8 and 9: Confidence intervals and the z-test

Topic 10: The t-test

Topic 11: The two-sample test

Topic 12: χ^2-test

Chi-squared tests

Suspicious dice

- A gambler is accused of using a loaded (6-sided) die, but he pleads innocent.
- A record has been kept of the last 60 throws.

```r
die <- c(4,3,3,1,2,3,4,6,5,6,
         2,4,1,3,3,5,3,4,3,4,
         3,3,4,5,4,5,6,4,5,1,
         6,4,4,2,3,3,2,4,4,5,
         6,3,6,2,4,6,4,6,3,2,
         5,4,6,3,3,3,5,3,1,4)
```

- Let's summarise these:

```r
table(die)
```

```
die
 1  2  3  4  5  6 
 4  6 17 16  8  9 
```

- These counts should be "roughly equal" for a fair die, but these look a bit **too** unequal.
- How can we test if the die is fair?

Box model for (possibly loaded) die

- We are very familiar with our box model for a **fair** die:

$$\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}\,\boxed{4}\,\boxed{5}\,\boxed{6}}$$

- A single random draw X from this box has the distribution

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- A box for a **loaded** die might be

$$\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}\,\boxed{3}\,\boxed{4}\,\boxed{4}\,\boxed{5}\,\boxed{6}}$$

giving

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

Goodness of fit test

- We can define the distribution by a probability vector $\boldsymbol{p} = (p_1, \ldots, p_6)$ of (rational) probabilities
  - so each $p_j \geq 0$ and $p_1 + \cdots + p_6 = 1$); and
  - we can imagine a box with a certain number of each ticket, so the proportion of tickets with integer j is p_j.
- We would like to test the hypothesis $H_0\colon p_1 = \cdots = p_6 = \frac{1}{6}$.
- We are interested in **any alternative that is not H_0**.
  - That is, $p_j \neq \frac{1}{6}$ for at least one j in $1, \ldots, 6$.
  - In brief the alternative is H_1: not H_0.
- This is an example of a **goodness of fit test**:

Expected frequencies after 60 “draws”

- Suppose H_0 is true. we have a fair die.
- Since each value 1,2,…,6 is equally likely, after 60 draws we would **expect** to get 10 of each:

Outcome	1	2	3	4	5	6
Prob.	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
Expected frequency	10	10	10	10	10	10

Comparison with observed frequencies

- The table below compares observed and expected frequencies:

Outcome	1	2	3	4	5	6
H_0 Prob.	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
Expected frequency	10	10	10	10	10	10
Observed frequency	4	6	17	16	8	9

- Due to random sampling the observed are not **exactly** equal to the expected, we anticipate some “small” discrepancies.
- We want to know *How different* do these have to be before it gets suspicious??

General formulation

- Suppose we have data $X_1, \ldots, X_n$ only taking k distinct values (categories), modelled as a random sample taken with replacement from a box.
    - ➟ The tickets of the box take k distinct values (categories).
    - ➟ We use integers $j = 1, 2, \ldots, k$ (or any other distinct values/labels) to **label the categories**.
    - ➟ The testing procedure we use can deal with general categorical data

$$\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}\,\boxed{4}\,\boxed{5}\,\boxed{6}}$$

$$\boxed{\boxed{\text{a}}\,\boxed{\text{b}}\,\boxed{\text{c}}\,\boxed{\text{d}}\,\boxed{\text{e}}\,\boxed{\text{f}}}$$

- Write $p_j = P(X_1 = j) =$ the proportion of tickets in box labelled j (for $j = 1, \ldots, k$).
- Write also $\boldsymbol{p} = (p_1, \ldots, p_k)$.
- We wish to test $H_0\colon \boldsymbol{p} = \boldsymbol{p_0}$ for some hypothesised $\boldsymbol{p_0} = (p_{01}, \ldots, p_{0k})$.
- The alternative we are interested in is H_1**: not** H_0.

Observed and Expected frequencies

- We summarise the data to **observed frequencies**: O_j = number of data points labelled j.
- We compare these to the corresponding **expected frequencies**: $E_j = np_{0j}$, i.e. the number of data points labelled j we would expect **under** H_0.

Outcome	1	2	$\cdots$	k
H_0 Prob.	p_{01}	p_{02}	$\cdots$	p_{0k}
Expected frequency	$E_1 = np_{01}$	$E_2 = np_{02}$	$\cdots$	$E_k = np_{0k}$
Observed frequency	O_1	O_2	$\cdots$	O_k

Test statistic: Pearson's χ^2 statistic

- A "foundational" paper in modern statistics was by Karl Pearson in 1900.
- He considered the statistic

$$T = \frac{(O_1 - E_1)^2}{E_1} + \cdots + \frac{(O_k - E_k)^2}{E_k}\,.$$

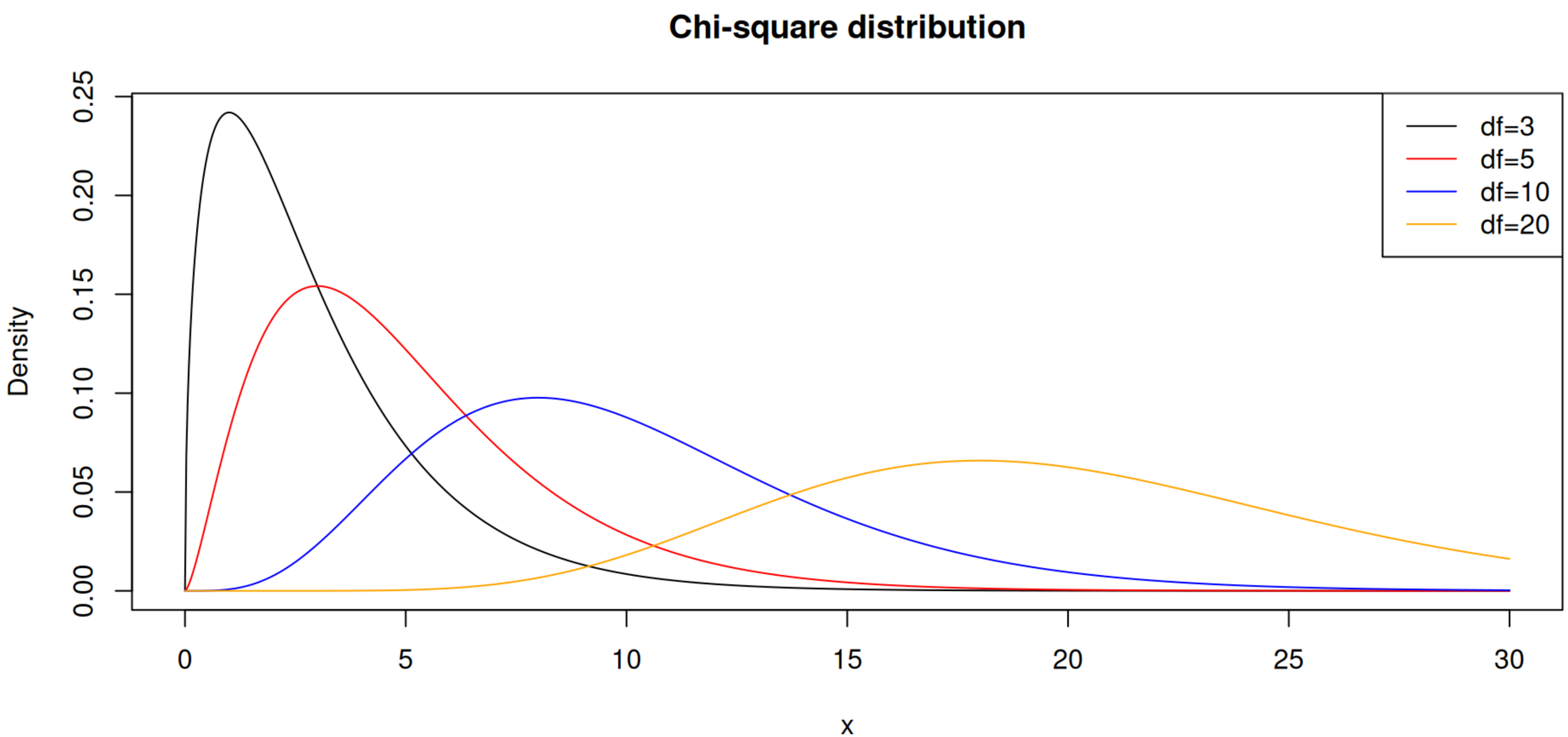
- For categories with larger E_i, the "error" $O_i - E_i$ tends to be bigger;
  - ➟ Dividing $(O_i - E_i)^2$ by E_i, this "normalised squared error" makes each term "comparable".
- He argued that under H_0, for "large n",

$$T \overset{\text{approx.}}{\sim} \chi^2_{k-1}\,,$$

the chi-squared distribution with $k - 1$ degrees of freedom.

The χ^2_d distribution

- Suppose we take d independent (i.e. with replacement) random draws from a $N(0, 1)$ box: $Z_1, Z_2, \ldots, Z_d$.
- Then the sum of squares $Z_1^2 + Z_2^2 + \cdots + Z_d^2$ has a χ^2_d distribution.
- It is a skewed (to the right) distribution, but gets more symmetric as d increases.



P-value

- Suppose we have k categories, and the observed value of Pearson's statistic is t_{obs}.
- The **larger** t_{obs}, the more evidence against H_0.
  - ➠ One-sided test.
  - ➠ The P-value is given by the area under the χ^2_{k-1} curve to the **right** of t_{obs}.
- This is the chance of
  - ➠ observing something more extreme than t_{obs}, assuming H_0 true.
- Why the degrees of freedom is $k-1$ in χ^2_{k-1}?
  - ➠ The test statistic T for k categories behaves like the summation of $Z_1^2 + \ldots + Z_{k-1}^2$ – the actual derivation of this is beyond the scope of this unit.
  - ➠ Quick way to remember (more later): there are k elements in the probability vector, but $\sum_{j=1}^{k} p_j = 1$, so we only need $k-1$ of "free" probability parameters to define the entire vector.

Our dice example

$\boxed{\text{H}}$

- Null hypothesis ($H_0 : p_0 = (\frac{1}{6}, \ldots, \frac{1}{6})$): the die is fair.
- Alternative hypothesis (H_1 :) at least one of $p_{0j} \neq \frac{1}{6}, j = 1, \ldots, 6$, indicating the die is loaded.

$\boxed{\text{A}}$ We need a sufficiently large n, what else? We will discuss this later.

$\boxed{\text{T}}$ The degrees of freedom is $6 - 1 = 5$, so χ^2_5 is the test distribution.

- One-sided test: large values of test statistics argue against H_0.
- For the record of results from the die

```
Oi = table(die)
Ei = rep(10, 6)
rbind( Ei, Oi)
```

```
    1  2  3  4  5  6
Ei 10 10 10 10 10 10
Oi  4  6 17 16  8  9
```

```
sum(((Oi-Ei)^2)/Ei)
```

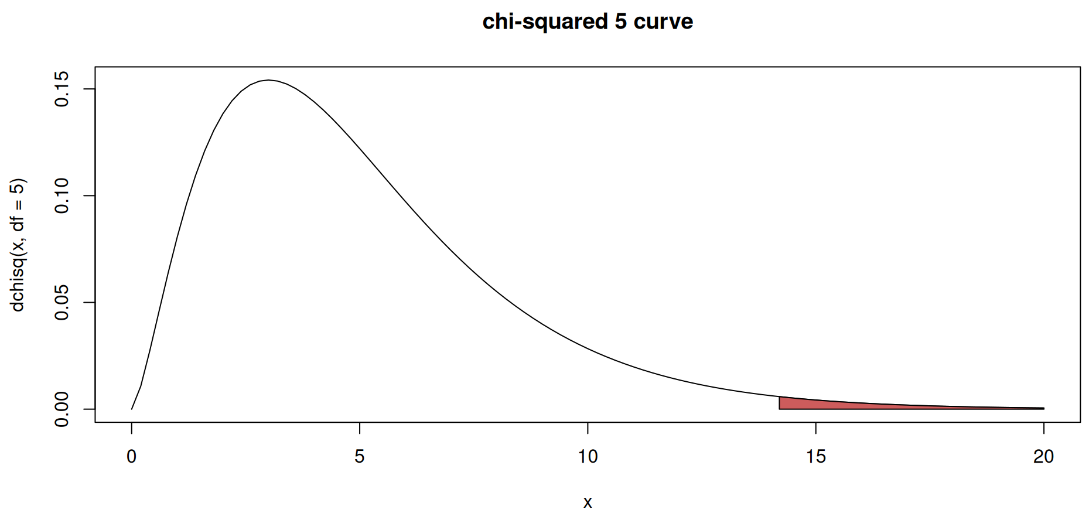
```
[1] 14.2
```

So the observed value is $t_{\text{obs}} = 14.2$.

$\boxed{\text{P}}$ Obtain P-value using `pchisq( ... , df= ... , lower.tail=F)`: we need the *upper tail* (large values of t_{obs} argue against H_0).

```
pchisq(14.2, df=5, lower.tail=F)
```

```
[1] 0.01438768
```



$\boxed{\text{C}}$ Is the value $t_{\text{obs}} = 14.2$ consistent with H_0?

- The P-value is a rather small.
- At a rather small false alarm rate (e.g., 2%), the data is significantly different from the claim of H_0 (all 6 sides equally likely).
  - Indirectly suggests the die may be loaded.

Using `chisq.test()`

- We can also use the built-in function `chisq.test()`.
- If we give it a vector of counts, it compares it to the vector of probabilities in `p`:

```r
chisq.test(Oi, p=c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6))
```

```

	Chi-squared test for given probabilities

data:  Oi
X-squared = 14.2, df = 5, p-value = 0.01439
```

- Note that by default it takes `p` as the same length as the vector containing observed frequencies, with equal probabilities:

```r
chisq.test(Oi)
```

```

	Chi-squared test for given probabilities

data:  Oi
X-squared = 14.2, df = 5, p-value = 0.01439
```

A Assumptions required

- The χ^2_{k-1} distribution is a "large-sample approximation" to the exact sampling distribution of Pearson's statistic when H_0 is true.
- It may not be a good approximation if
    - *either* the sample size n is not very large
    - *or* some categories have very small hypothesised probabilities.
- A "rule of thumb" is that if all expected frequencies E_j are at least 5, the χ^2_{k-1} approximation should be reasonably accurate.
    - The R function `chisq.test()` prints a warning if this condition is violated:

```r
Oi = c(5, 3, 4)
chisq.test(Oi, p=c(1/3, 1/3, 1/3))
```

```
Warning in chisq.test(Oi, p = c(1/3, 1/3, 1/3)): Chi-squared approximation may
be incorrect


    Chi-squared test for given probabilities

data:  Oi
X-squared = 0.5, df = 2, p-value = 0.7788
```

Special case: wquivalence with Z-test for 0-1 box

- We can draw a connection between the chi-squared test and a **two-sided Z-test for proportion**.
- Consider a box containing only $\boxed{0}$s and $\boxed{1}$s, let p denote the proportion of $\boxed{1}$s in the box.
- Suppose we have a random sample $X_1, \dots, X_n$ taken with replacement from the box.
- Consider testing the null hypothesis $H_0\colon p = p_0$ with the two-sided $H_1\colon p \neq p_0$
- We have already done this using a Z-test with the statistic

$$Z = \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{S - np_0}{\sqrt{np_0(1 - p_0)}} \, ,$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = S/n$ is the sample proportion of $\boxed{1}$s.

Chi-squared test for 0-1 box

- Note that the two-sided P-value $P(|Z| > |z|)$ is the same as

$$P(Z^2 > z^2) = P\left(Z^2 > \frac{(s - np_0)^2}{np_0(1-p_0)}\right)$$

  where $Z^2 \sim \chi^2_1$ and s is the observed sample sum (number of 1's in a sample).
- We may also view this as a χ^2-test.

Outcome	0	1
Prob.	$1 - p_0$	p_0
Expected frequency	$E_0 = n(1-p_0)$	$E_1 = np_0$
Observed frequency	$O_0 = n - S$	$O_1 = S$

Both tests are equivalent for 0-1 box

- Pearson's statistic is then

$$
\begin{aligned}
T &= \frac{(O_0 - E_0)^2}{E_0} + \frac{(O_1 - E_1)^2}{E_1} \\
&= \frac{[(n - S) - n(1 - p_0)]^2}{n(1 - p_0)} + \frac{(S - np_0)^2}{np_0} \\
&= \frac{(n - S - n + np_0)^2}{n(1 - p_0)} + \frac{(S - np_0)^2}{np_0} \\
&= \frac{(S - np_0)^2}{n} \left(\frac{1}{1 - p_0} + \frac{1}{p_0} \right) \\
&= \frac{(S - np_0)^2}{n} \left(\frac{p_0 + (1 - p_0)}{p_0(1 - p_0)} \right) \\
&= \frac{(S - np_0)^2}{np_0(1 - p_0)} \\
&= Z^2 \, .
\end{aligned}
$$

- The chi-squared test is **exactly** a two-sided Z-test for 0-1 box, as for $Z \sim N(0, 1)$, Z^2 follows χ^2_1.

Example: 5% level of significance

```r
round(qchisq(.95, df=1),2)
```

```
[1] 3.84
```

- An upper 5% percentage point for χ^2_1 is
  - ➠ The critical region of rejection is $T > 3.84$.

```r
round(qnorm(0.975),2)
```

```
[1] 1.96
```

```r
round(qnorm(0.975)^2,2)
```

```
[1] 3.84
```

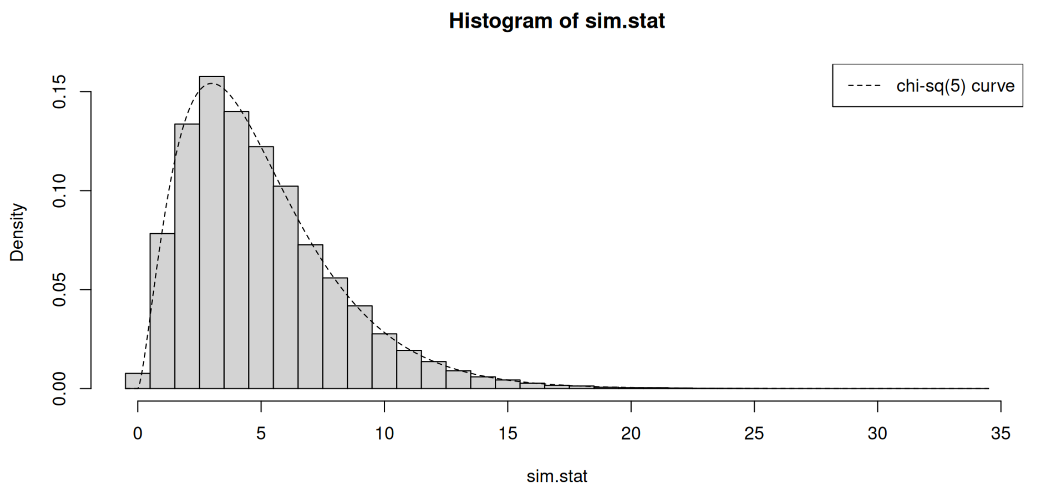
- An upper 2.5% percentage point (97.5% quantile) for $N(0,1)$ is approximately 1.96
  - ➠ $P(|Z| > 1.964)$ is the same as $P(|Z|^2 = Z^2 > 1.96^2 \approx 3.84)$
  - ➠ The region of rejection is $|Z| > 1.96$ or $Z^2 > 3.84$.
- The chi-squared test may be viewed as a **generalisation** of the **two-sided** Z-test for a proportion, to a box with more than 2 different values in it.

Simulation

Using simulation: the dice example

- We can approximate the sampling distribution of the test statistic by simulating an appropriate (approximate if necessary) box model.
- Straightforward for chi-sq tests – H_0 completely specifies the distribution of X_i, and hence the box.

```
sim.stat=0 # the dice example
for(i in 1:100000) {
   sim.rolls=sample(1:6, size=60, replace=T)
   freqs = tabulate(sim.rolls, nbins=6)  # works even with zero freqs, better than table()
   sim.stat[i] = chisq.test(freqs)$stat  # save the test statistics
}
```



- Nice agreement between the histogram of simulated Pearson statisics and the χ^2_5 curve.

Simulated P-value

- The observed Pearson statistic

```r
Oi = table(die)
Ei = rep(10, 6)
rbind( Ei, Oi)
```

```
    1  2  3  4  5  6
Ei 10 10 10 10 10 10
Oi  4  6 17 16  8  9
```

```r
stat=sum(((Oi-Ei)^2)/Ei)
stat
```

```
[1] 14.2
```

- P-value obtained using the simulated test distribution
  - Note that it’s a one-sided test.

```r
mean(sim.stat>=stat)
```

```
[1] 0.0139
```

- P-value obtained using the theoretical χ^2_5

```r
chisq.test(Oi)$p.value
```

```
[1] 0.01438768
```

- The simulation-based P-value is close to that obtained using the χ^2_5 approximation.

Small expected frequencies

- Consider another example where we the assumptions are not reasonable:
  - ➟ suppose we draw a sample of size $n = 10$ from the box (with 11 tickets)

$$\boxed{\boxed{1}\ \boxed{1}\ \boxed{1}\ \boxed{1}\ \boxed{2}\ \boxed{2}\ \boxed{2}\ \boxed{2}\ \boxed{3}\ \boxed{4}\ \boxed{5}}$$

- How does Pearson's statistic behave when we test $H_0\colon \boldsymbol{p}_0 = \left(\frac{4}{11}, \frac{4}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}\right)$?
- Note that H_0 is ture in this example.
- The expected frequencies are then **all** < 5:

```r
n = 10
p0=c(4,4,1,1,1)/11
n*p0
```

```
[1] 3.6363636 3.6363636 0.9090909 0.9090909 0.9090909
```

- So we suspect the χ^2_4 approximation may not be so good.

Using `chisq.test()`

- Sure enough, `chisq.test()` tells us this: suppose we draw the sample

```
samp
```

```
[1] 1 3 3 2 3 2 2 2 2 2
```

```
table(samp) # skips categories with zero frequency, can't be used here
```

```
samp
1 2 3
1 6 3
```

```
Obs.freq = tabulate(samp, nbins=5) # works even if some values don't appear
Obs.freq
```

```
[1] 1 6 3 0 0
```

```
chisq.test(Obs.freq, p=p0)
```

```
Warning in chisq.test(Obs.freq, p = p0): Chi-squared approximation may be
incorrect


    Chi-squared test for given probabilities

data:  Obs.freq
X-squared = 10.075, df = 4, p-value = 0.03918
```

- the function `tabulate(samp, nbins=5)` counts the frequencies of categories from 1 to 5 in this case, without skipping labels.

Using simulation

- Simulate the box under H_0

```r
box = c(1, 1, 1, 1, 2, 2, 2, 2, 3, 4, 5)
sim.stat=0
for(i in 1:100000) {
    sim.obs = sample(box, size=n, replace=T)
    freqs = tabulate(sim.obs, nbins=5)
    sim.stat[i] = suppressWarnings(chisq.test(freqs, p=p0)$stat)
    # without supressWarnings() we get
    # 10000 "approximation may be incorrect"
    # warnings
}
```

- Compare the quantiles of simulated Pearson's statistics with the theoretical ones

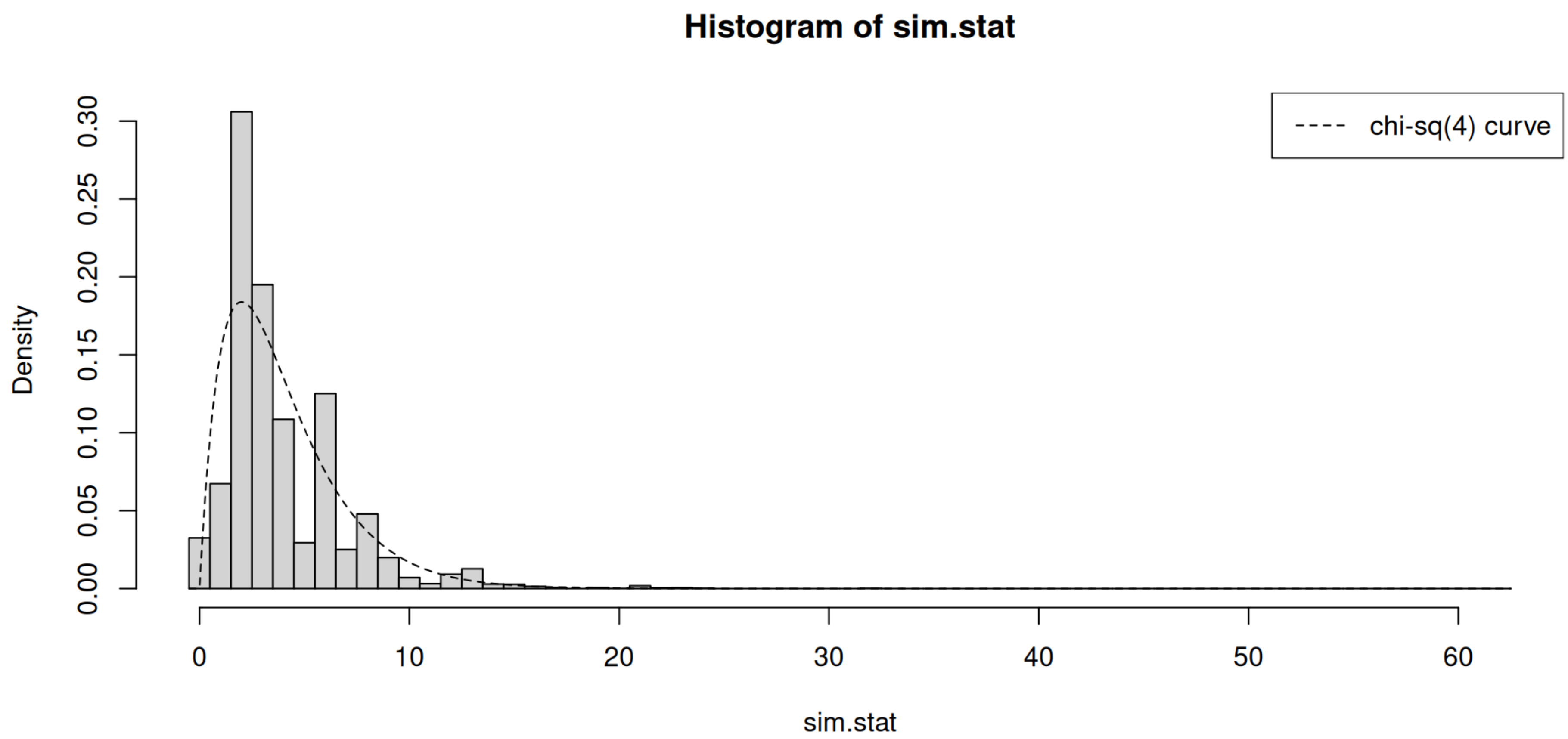
```r
quantile(sim.stat, probs=c(0.95, 0.98, 0.99))
```

```
   95%    98%    99%
 8.975 12.550 14.200
```

```r
qchisq(c(.95, .98, .99), df=4)
```

```
[1]  9.487729 11.667843 13.276704
```

- The upper 2% and 1% points are bigger than χ^2_4 would suggest.



- The distribution is multi-modal and there are more large values than χ^2_4 would suggest.
- Our earlier observed Pearson's statistic gives a simulation-based P-value of

```
stat = chisq.test(Obs.freq, p=p0)$stat
mean(sim.stat ≥ stat)
```

```
[1] 0.04037
```

Using `chisq.test(..., simulate=T)`

- The `simulate=T` argument gives a similar result.

```r
stat = chisq.test(Obs.freq, p=p0)$stat
mean(sim.stat >= stat)
```

```
[1] 0.04037
```

```r
chisq.test(Obs.freq, p=p0, simulate=T, B = 100000)
```

```

	Chi-squared test for given probabilities with simulated p-value (based
	on 1e+05 replicates)

data:  Obs.freq
X-squared = 10.075, df = NA, p-value = 0.04034
```

- `B = ...` specify the number of samples used in the simulation.

Chi-squared tests with estimated parameters

Parameters

- In Pearson's test
  - Observed frequency O of each category is compared with expected frequency $E = np$, where p is the probability of "landing" in that category.
  - We test **goodness of fit**, i.e. a null hypothesis H_0 specifies probabilities for each category - next we will see they possibly depend on some parameters.
    - alternative hypothesis is then H_1: not H_0.
- Pearsons statistic T is the sum of $\frac{(O-E)^2}{E}$ over all categories.
- When H_0 is true, T has an approximate χ^2_d distribution, where the degrees of freedom parameter d is given by

$$(\text{no. free parameters under full model}) - (\text{no. free parameters under } \ H_0)\ .$$

Completely specified probability vector

- In the previous examples, we had $\boldsymbol{k}$ categories and a vector of probabilities $\boldsymbol{p} = (p_1, \ldots, p_k)$ for each category.
- Then under the **full model** (where any probability vector is allowed), we have
  - ➠ $\boldsymbol{k}$ parameters **but**
  - ➠ only $\boldsymbol{k-1}$ of these are **free** since they add to 1, if we know $p_1, \ldots, p_{k-1}$,

$$p_k = 1 - (p_1 + \cdots + p_{k-1})$$

  is automatically determined.
- In H_0: $\boldsymbol{p} = \boldsymbol{p_0} = (p_{01}, \ldots, p_{0k})$, we had a completely specified probability vector $\boldsymbol{p_0}$.
  - ➠ Then there are zero free parameters under H_0.
- Therefore, T is approx. χ^2_d with

$$\begin{aligned} d &= (\text{no. free parameters under full model}) - (\text{no. free parameters under } \ H_0) \\ &= (k-1) - 0 = k-1\,. \end{aligned}$$

Two-way tables: test of independence

- Consider the following data giving biological sex (row categories) and handedness (column categories) for 2,237 people:

	Right-handed	Left-handed	Ambidextrous	Total
Men	934	113	20	1067
Women	1070	92	8	1170
Total	2004	205	28	2237

- Do the data suggest any evidence against that the handedness and the gender are independent?
    - ➠ Note that, if they are independent, there is no difference in handedness between men and women.

Pearson's statistic

- The statistic takes the same basic form: we add terms like $\frac{(O-E)^2}{E}$, but over all cells in the table:

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

  and so is now a "double sum".
- Here,
  - O_{ij} is the observed frequency in row i, column j
  - E_{ij} is the expected frequency in row i, column j **under the null hypothesis**.
- How do we formulate the null hypothesis exactly?
- How do we determine the expected frequencies, the E_{ij}s?

Full model

- Full model: rc different categories, unconstrained probabilities

	Col 1	Col 2	$\cdots$	Col c	Total
Row 1	p_{11}	p_{12}	$\cdots$	p_{1c}	$p_{1\bullet}$
Row 2	p_{21}	p_{22}	$\cdots$	p_{2c}	$p_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Row r	p_{r1}	p_{r2}	$\cdots$	p_{rc}	$p_{r\bullet}$
Total	$p_{\bullet 1}$	$p_{\bullet 2}$	$\cdots$	$p_{\bullet c}$	1

- We thus have $rc - 1$ free parameters under the full model.
- Here we use “dot” notation for sums. E.g.,
  - $p_{\bullet 1} = \sum_{i=1}^{r} p_{i1}$ (sum over a rows for a specified column)
  - $p_{1\bullet} = \sum_{j=1}^{c} p_{1j}$ (sum over columns for a specified row)
- The row sums $p_{\bullet j}$ gives the marginal probabilities for every column. That is,
  - the chance of landing in j-th column category of the table. E.g., handedness in this example.
- The column sums $p_{i\bullet}$ gives the marginal probabilities for for every rom. That is,
  - the chance of landing in i-th row category of the table. E.g., biological sex in this example.

Null hypothesis

- The null hypothesis says: the events $\{\textbf{being in Row } i\}$ and $\{\textbf{being in Col } j\}$ are independent. That is

$$p_{ij} = P\{\text{in Row } i \text{ and Col } j\} = P\{\text{in Row } i\} \times P\{\text{in Col } j\} = p_{i\bullet} p_{\bullet j}$$

- Under H_0, the probability of each cell is

	Col 1	Col 2	$\cdots$	Col c	Total
Row 1	$p_{1\bullet}p_{\bullet 1}$	$p_{1\bullet}p_{\bullet 2}$	$\cdots$	$p_{1\bullet}p_{\bullet c}$	$p_{1\bullet}$
Row 2	$p_{2\bullet}p_{\bullet 1}$	$p_{2\bullet}p_{\bullet 2}$	$\cdots$	$p_{2\bullet}p_{\bullet c}$	$p_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Row r	$p_{r\bullet}p_{\bullet 1}$	$p_{r\bullet}p_{\bullet 2}$	$\cdots$	$p_{r\bullet}p_{\bullet c}$	$p_{r\bullet}$
Total	$p_{\bullet 1}$	$p_{\bullet 2}$	$\cdots$	$p_{\bullet c}$	1

Observed and expected frequencies

- Observed frequencies:

	Col 1	Col 2	$\cdots$	Col c	Total
Row 1	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1\bullet}$
Row 2	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Row r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r\bullet}$
Total	$O_{\bullet 1}$	$O_{\bullet 2}$	$\cdots$	$O_{\bullet c}$	n

- Expected frequencies under null hypothesis: $E_{ij} = np_{i\bullet}p_{\bullet j}$

	Col 1	Col 2	$\cdots$	Col c	Total
Row 1	$np_{1\bullet}p_{\bullet 1}$	$np_{1\bullet}p_{\bullet 2}$	$\cdots$	$np_{1\bullet}p_{\bullet c}$	$np_{1\bullet}$
Row 2	$np_{2\bullet}p_{\bullet 1}$	$np_{2\bullet}p_{\bullet 2}$	$\cdots$	$np_{2\bullet}p_{\bullet c}$	$np_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Row r	$np_{r\bullet}p_{\bullet 1}$	$np_{r\bullet}p_{\bullet 2}$	$\cdots$	$np_{r\bullet}p_{\bullet c}$	$np_{r\bullet}$
Total	$np_{\bullet 1}$	$np_{\bullet 2}$	$\cdots$	$np_{\bullet c}$	n

- We need to estimate the marginal probabilities $p_{i\bullet}$s **and** the $p_{\bullet j}$s.

Estimate marginal probabilities $p_{i\bullet}$s and $p_{\bullet j}$s

- Under H_0, we can collapse all the rows into a single row (last row of the observed table) to form a single sample from the "column" box. We can then estimate the column probability $P\{\text{in Col } j\}$ using

$$\hat{p}_{\bullet j} = \frac{O_{\bullet j}}{n},$$

- Similarly, we can collapse all the columns into a single column (last column of the observed table) to form a single sample from the "row" box. We can then estimate the row probability $P\{\text{in Row } i\}$ using

$$\hat{p}_{i\bullet} = \frac{O_{i\bullet}}{n}$$

- This gives expected frequencies

$$E_{ij} = n\hat{p}_{i\bullet}\hat{p}_{\bullet j} = n\,\frac{O_{i\bullet}}{n}\,\frac{O_{\bullet j}}{n} = \frac{(\text{Row } i \text{ total}) \times (\text{Col } j \text{ total})}{\text{Grand total}},$$

.

Degrees of freedom

- Pearson's statistic approximately follows a χ^2 distribution under H_0 with degrees of freedom given by

$$(\textbf{no. free parameters under full model}) - (\textbf{no. free parameters under } H_0)\,.$$

- There are $rc - 1$ free parameters under the full model.
- Under the null hypothesis there are
  - r row probabilities, giving $r - 1$ free parameters
  - c column probabilities, giving $c - 1$ free parameters
  - there are thus $(r-1) + (c-1)$ free parameters under H_0.
- The difference is

$$(rc-1) - (r-1) - (c-1) = rc - r - c + 1 = (r-1)(c-1)\,.$$

Handedness example

- Observed frequencies:

```
men = c(934, 113, 20)
women  = c(1070, 92, 8)
Oij = rbind(men, women) # define a matrix row by row
Oij
```

```
      [,1] [,2] [,3]
men    934  113   20
women 1070   92    8
```

```
colnames(Oij)=c("RH", "LH", "Ambi")
Oij
```

```
        RH  LH Ambi
men    934 113   20
women 1070  92    8
```

- Row and column sums:

```
R = rowSums(Oij)
R
```

```
  men women
 1067  1170
```

```
C = colSums(Oij)
C
```

```
  RH   LH Ambi
2004  205   28
```

Pearson's statistic and P-value

- Pearson's statistic

```r
n = sum(Oij)
n
```

```
[1] 2237
```

```r
Eij = outer(R, C, FUN="*")/n
Eij
```

```
             RH        LH     Ambi
men    955.8641  97.78051 13.35539
women 1048.1359 107.21949 14.64461
```

```r
stat=sum( ((Oij-Eij)^2)/Eij )
stat
```

```
[1] 11.80613
```

- P-value，Pearson's statistic approximately follows $\chi^2_{(r-1)(c-1)}$

```r
r=length(R)
c=length(C)
d=(r-1)*(c-1)
d
```

```
[1] 2
```

```r
pchisq(stat, df=d, lower.tail=F)
```

```
[1] 0.002731055
```

Using `chisq.test()`

- The R function `chisq.test()` can also be used to test for relationships between rows and columns of two-way tables.
- The observed frequencies need to be in a matrix.

```r
Oij
```

```
        RH  LH Ambi
men    934 113   20
women 1070  92    8
```

```r
chisq.test(Oij)
```

```

	Pearson's Chi-squared test

data:  Oij
X-squared = 11.806, df = 2, p-value = 0.002731
```

Using simulation

- As with other tests, the chi-squared approximation may not be reasonable in some circumstances,
  - if either the overall sample size is small; or
  - we have too many small expected frequencies.
- In such a case, it is possible to use the simulation-based P-value (setting `simulate=T`).

```r
chisq.test(Oij, simulate=T)
```

```

	Pearson's Chi-squared test with simulated p-value (based on 2000
	replicates)

data:  Oij
X-squared = 11.806, df = NA, p-value = 0.003498
```

- The simulation for two way table is rather complicated (we skip the details here).
- Note that the chi-squared approximation gives a P-value that is about half the size it should be here:
  - trusting approximations blindly can lead to false significance in some cases.

Summary

- Pearson's statistic adds $\frac{(O-E)^2}{E}$ over each category, where O is the observed frequency and E is the expected frequency under the null hypothesis.
    - ➠ We may need to estimate parameters to compute the Es.
- For large enough sample sizes, the statistic has an approximate χ^2 distribution under the null hypothesis, with degrees of freedom given by

$$(\textbf{no. free parameters under full model}) - (\textbf{no. free parameters under } H_0)\,.$$

    - ➠ If we estimate parameters, we need to make sure they are estimated "properly" for this to be true.
- We can always use `chisq.test(..., simulate=T)` if unsure.
    - ➠ It is good practice to always compare the two.

Example: test of independence

Example

- The table below shows the results of a random sample of 100 males being classified according to amount of smoking (row categories) and age (column categories).

	Under 40 years	Over 40 years
< 20 cigarettes/day	50	15
≥ 20 cigarettes/day	10	25

```r
under.40 = c(50, 10)
over.40=c(15, 25)
Of = cbind(under.40, over.40)
rownames(Of)=c("less.20", "more.20")
Of
```

```
        under.40 over.40
less.20       50      15
more.20       10      25
```

Manual calculation: use `rowSums()`, `colSums()` and `outer()`.

- Row and column sums may be obtained using `apply()`:

```
rsums = rowSums(Of)
csums = colSums(Of)
rsums
```

```
less.20 more.20
     65      35
```

```
csums
```

```
under.40  over.40
      60       40
```

- Expected frequencies may be obtained using `outer()`:

```
n=sum(Of)
n
```

```
[1] 100
```

```
Ef = outer(rsums, csums, FUN="*")/n
Ef
```

```
        under.40 over.40
less.20       39      26
more.20       21      14
```

Pearson's statistic and (theoretical) P-value

```
stat = sum(((Of-Ef)^2)/Ef)
stat
```

```
[1] 22.16117
```

```
pchisq(stat, df=1, lower.tail=F)
```

```
[1] 2.506928e-06
```

- This is a very small P-value, providing very strong evidence against the hypothesis that smoking level and age are independent.

Using `chisq.test()` in the 2-by-2 case

- When we have a 2-by-2 table, the R function `chisq.test()` applies a (Yates') "continuity correction" by subtracting 0.5 from each $|O - E|$ before squaring.
  - ➟ This is designed to improve the chi-squared approximation.
  - ➟ We **do not want this** though: it confuses the issue (we prefer to use simulation if the chi-squared approximation is not reliable).
- We must thus use `chisq.test(... , correct=F)`.
- With the correction we get a *slightly smaller* statistic:

```r
chisq.test(Of)
```

```

	Pearson's Chi-squared test with Yates' continuity correction

data:  Of
X-squared = 20.192, df = 1, p-value = 7.003e-06
```

- Without the correction we get results that agree with our manual calculation:

```r
chisq.test(Of, correct=F)
```

```
	Pearson's Chi-squared test

data:  Of
X-squared = 22.161, df = 1, p-value = 2.507e-06
```

- We can also use `chisq.test()`'s built-in `simulate=T` option to obtain the simulated P-value:

```r
chisq.test(Of, simulate=T)
```

```
	Pearson's Chi-squared test with simulated p-value (based on 2000
	replicates)

data:  Of
X-squared = 22.161, df = NA, p-value = 0.0004998
```

- For all methods, we get a very small P-value: this data provides very strong evidence of a relationship between age and smoking level.