

# Distribution-free methods

(Module 7)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2019

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Testing for a difference in location</b>	<b>2</b>
2.1	Sign test . . . . .	2
2.2	Wilcoxon signed-rank test (one-sample) . . . . .	4
2.3	Wilcoxon rank-sum test (two-sample) . . . . .	6
<b>3</b>	<b>Goodness-of-fit tests (<math>\chi^2</math>)</b>	<b>7</b>
3.1	Introduction . . . . .	7
3.2	Two classes . . . . .	8
3.3	More than two classes . . . . .	8
3.4	Estimating parameters . . . . .	10
<b>4</b>	<b>Tests of independence (contingency tables)</b>	<b>12</b>

## Aims of this module

- Introduce inference methods that do not make strong distributional assumptions
- Explain the highly used Pearson's chi-squared test

## 1 Introduction

### Distribution-free methods

- So far, have only considered tests that assume a specified form for the population distribution.
- We don't always want to make such assumptions.
- Instead, we can use *distribution-free* methods.
- Here, we will learn about various distribution-free hypothesis tests.

### An aside: *distribution-free* versus *non-parametric*

- The term *non-parametric* is also often used to describe methods that do not assume a specific distributional form.
- It is usually a misnomer: the methods typically **do** make use of parameters, but there are usually a large number of them and they adapt to the data.
- Thus, a better term might be **super-parameteric**.
- (Note: we won't be covering any advanced methods of this form in this subject.)
- In any case, the convention has stuck, so you will see either of the labels 'distribution-free' or 'non-parameteric' being used.

## Distribution-free tests

- Even without making distributional assumptions, it is possible to obtain exact or asymptotic sampling distributions for various statistics.
- Can use these as a basis for hypothesis tests.
- Often the distribution-free test statistic is approximately normally distributed
- ... the Central Limit Theorem strikes again!

## 2 Testing for a difference in location

### Extracting information with fewer assumptions

- How can we assess the information in a sample without assuming a distribution?
- Specifying a distribution is somewhat analogous to specifying a scale of measurement, so...
- How do we compare numbers without a scale?
- Two strategies:
  1. **(Sign)** Only record whether a number is smaller or greater than a reference number, i.e. replace them by binary indicator variables.
  2. **(Rank)** Only retain information about the order of the numbers, i.e. replace them by their rank order.
- Each of these throws away some information, but hopefully retains enough to be useful.
- We now look at a few methods that use these strategies.

### Aim: test for the median

- Let  $X$  have median  $m$
- We have an iid sample of size  $n$  from  $X$
- Can we test  $H_0: m = m_0$  with very few assumptions?
- (Want to find distribution-free alternatives to tests about the mean, such as the t-test)
- (Typically consider medians rather than means when distribution-free)

### 2.1 Sign test

#### Sign test

- We assume  $X$  is continuous
- (No further assumptions!)
- Compute,  $Y$ , the number of positive numbers amongst  $X_1 - m_0, \dots, X_n - m_0$
- In other words, replace  $X_i$  with  $\text{sgn}(X_i - m_0)$
- Under  $H_0$ , we have  $Y \sim \text{Bi}(n, 0.5)$
- Tests proceed as usual...

#### Example (sign test)

The time between calls to a switchboard is represented by  $X$ .

$$H_0: m = 6.2 \quad \text{versus} \quad H_1: m < 6.2$$

$i$	$x_i$	$x_i - 6.2$	Sign	$i$	$x_i$	$x_i - 6.2$	Sign
1	6.80	0.60	+1	11	18.90	12.70	+1
2	5.70	-0.50	-1	12	16.90	10.70	+1
3	6.90	0.70	+1	13	10.40	4.20	+1
4	5.30	-0.90	-1	14	44.10	37.90	+1
5	4.10	-2.10	-1	15	2.90	-3.30	-1
6	9.80	3.60	+1	16	2.40	-3.80	-1
7	1.70	-4.50	-1	17	4.80	-1.40	-1
8	7.00	0.80	+1	18	18.90	12.70	+1
9	2.10	-4.10	-1	19	4.80	-1.40	-1
10	19.00	12.80	+1	20	7.90	1.70	+1

- $Y$  is the number of positive signs. Reject  $H_0$  if  $Y$  too small. (If median  $< 6.2$  then expect fewer than 1/2 of the observations to be greater than 6.2.)
- Since  $\Pr(Y \leq 6) = 0.0577 \approx 0.05$ , an appropriate rejection rule is to reject  $H_0$  if  $Y \leq 6$ . (In R: `pbinom(6, 20, 0.5)`)
- We observed  $y = 11$ , so cannot reject  $H_0$ .
- The p-value is  $\Pr(Y \leq 11) = 0.75 > 0.05$  so cannot reject  $H_0$ . (In R: `pbinom(11, 20, 0.5)`)

### R code

```
> binom.test(11, 20, alternative = "less")

Exact binomial test

data: 11 and 20
number of successes = 11, number of trials = 20,
p-value = 0.7483
alternative hypothesis: true probability of
          success is less than 0.5
95 percent confidence interval:
 0.0000000 0.7413494
sample estimates:
probability of success
          0.55
```

### Sign test for paired samples

Can also use the sign test for paired samples: simply replace  $(x_i, y_i)$  with  $\text{sgn}(x_i - y_i)$ .

For example:

$i$	$x_i$	$y_i$	Sign
1	8.9	10.3	-1
2	26.7	11.7	+1
3	12.4	5.2	+1
4	34.3	36.9	-1

### Use of the sign test

- The sign test requires few assumptions
- But it doesn't use information on the size of the differences, so it can be insensitive to departures from  $H_0$
- In other words, large type II error or small power
- Tends to only be used when the data are not numerical but for which comparisons between values are meaningful (e.g. ordinal data)

## 2.2 Wilcoxon signed-rank test (one-sample)

### Wilcoxon one-sample test

- Now, assume the underlying distribution is also symmetrical (as well as continuous)
- Same null hypothesis ( $H_0: m = m_0$ ) against a one-sided or two-sided alternative
- Determine the ranks of:  $|X_1 - m_0|, \dots, |X_n - m_0|$
- Replace the data by signed ranks,  $X_i$  becomes  $\text{sgn}(X_i - m_0) \cdot \text{rank}(|X_i - m_0|)$
- The *Wilcoxon signed-rank statistic*,  $W$ , is the sum of these signed ranks
- Using this as a basis for a test gives the *Wilcoxon signed-rank test*, also known as the *Wilcoxon one-sample test*.

### Alternative definitions

- Textbooks and software packages vary in the statistic they use
- We just defined:  $W$  is the sum of the signed ranks
- A popular alternative:  $V$  is the sum of the positive ranks only
- $V$  is a bit easier to calculate, esp. by hand
- R uses  $V$
- $V$  and  $W$  are deterministically related (can you derive the formula?)
- $V$  and  $W$  have different (but related) sampling distributions
- Using either statistic leads to equivalent test procedures

### Example (Wilcoxon one-sample test)

- The lengths of 10 fish are:  
5.0, 3.9, 5.2, 5.5, 2.8, 6.1, 6.4, 2.6, 1.7, 4.3
- Interested in testing:  $H_0: m = 3.7$  versus  $H_1: m > 3.7$

$i$	$x_i$	$x_i - 3.7$	$ x_i - 3.7 $	Rank	Signed rank
1	5.0	1.3	1.3	5	5
2	3.9	0.2	0.2	1	1
3	5.2	1.5	1.5	6	6
4	5.5	1.8	1.8	7	7
5	2.8	-0.9	0.9	3	-3
6	6.1	2.4	2.4	9	9
7	6.4	2.7	2.7	10	10
8	2.6	-1.1	1.1	4	-4
9	1.7	-2.0	2.0	8	-8
10	4.3	0.6	0.6	2	2

- The sum of signed ranks is:

$$W = 5 + 1 + 6 + 7 - 3 + 9 + 10 - 4 - 8 + 2 = 25$$

- Alternatively, the sum of positive ranks is:

$$V = 5 + 1 + 6 + 7 + 9 + 10 + 2 = 40$$

### Decision rule

- What is an appropriate critical region?
- If  $H_1: m > 3.7$  is true, we expect more positive signs. Then  $W$  should be large, so the critical region should be  $W \geq c$  for a suitable  $c$ .
- (For other alternative hypotheses, e.g. two-sided, need to modify this accordingly.)

- If  $H_0$  is true then  $\Pr(X_i < m_0) = \Pr(X_i > m_0) = \frac{1}{2}$ .
- Assignment of the  $n$  signs to the ranks are mutually independent (due to symmetry assumption)
- $W$  is the sum of the integers  $1, \dots, n$ , each with a positive or negative sign
- Under  $H_0$ ,  $W = \sum_{i=1}^n W_i$  where

$$\Pr(W_i = i) = \Pr(W_i = -i) = \frac{1}{2}, \quad i = 1, \dots, n$$

- The mean under  $H_0$  is  $\mathbb{E}(W_i) = -i \cdot \frac{1}{2} + i \cdot \frac{1}{2} = 0$ , so  $\mathbb{E}(W) = 0$
- Similarly,  $\text{var}(W_i) = \mathbb{E}(W_i^2) = i^2$  and

$$\text{var}(W) = \sum_{i=1}^n \text{var}(W_i) = \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$$

- A more advanced argument shows that for large  $n$  this statistic approximately follows a normal distribution when  $H_0$  is true. In other words,

$$Z = \frac{W - 0}{\sqrt{n(n+1)(2n+1)/6}} \approx N(0, 1)$$

- $\Pr(W \geq c \mid H_0) \approx \Pr(Z \geq z \mid H_0)$ , which allows us to determine  $c$ .
- In this case, for  $n = 10$  and  $\alpha = 0.05$ , we reject  $H_0$  if

$$Z = \frac{W}{\sqrt{10 \cdot 11 \cdot 21/6}} \geq 1.645$$

(because  $\Phi^{-1}(0.95) = 1.645$ ) which is equivalent to

$$W \geq 1.645 \times \sqrt{\frac{10 \cdot 11 \cdot 21}{6}} = 32.27$$

- For the example data we have  $w = 25$ , so we do not reject  $H_0$

## Using R

- R uses  $V$  rather than  $W$
- For small sample sizes R will use the exact sampling distribution (which we haven't explored) rather than the normal approximation.
- To carry out the test, use: `wilcox.test`
- To work with the sampling distribution of  $V$ , use: `psignrank`
- Note:  $\mathbb{E}(V) = n(n+1)/4$  and  $\text{var}(V) = n(n+1)(2n+1)/24$ . You can derive these in a similar way to  $W$ .

```
> wilcox.test(x, mu = 3.7, alternative = "greater",
  exact = TRUE)
```

Wilcoxon signed rank test

```
data: x
V = 40, p-value = 0.1162
alternative hypothesis: true location is greater than 3.7
```

```
# Calculate exact p-value manually.
> 1 - psignrank(39, 10)
[1] 0.1162109
```

```
# Calculate approximate p-value, based on W.
> z <- 25 / sqrt(10 * 11 * 21 / 6)
> 1 - pnorm(z)
[1] 0.1013108
```

⇒ Close agreement between exact and approximate p-values

## Paired samples

- Like other tests, we can use the Wilcoxon signed-rank test for paired samples by first taking differences and treating these as a sample from a single distribution.
- The assumption of symmetry is quite reasonable in this setting, since under  $H_0$  we would typically assume  $X$  and  $Y$  have the same distribution and therefore  $X - Y \sim Y - X$ .
- Indeed, this test is most often used in such a setting, due to the plausibility of this assumption.

## Tied ranks

- We assumed a continuous population distribution
- Thus, all observations will differ (with probability 1)
- In practice, the data are reported to finite precision (e.g. due to rounding), so we could have exactly equal values
- This will lead to ties when ranking our data
- If this happens, the ‘rank’ assigned for the tied values should be equal to the average of the ranks they span
- Example:

Value:	2.1	4.3	4.3	5.2	5.7	5.7	5.7	5.9
Rank:	1	2.5	2.5	4	6	6	6	8

- The presence of ties complicates the derivation of the sampling distribution, but R knows how to do the right thing

## 2.3 Wilcoxon rank-sum test (two-sample)

### Wilcoxon two-sample test

- We can create a two-sample version of the Wilcoxon test.
- Independent random samples  $X_1, \dots, X_{n_X}$  and  $Y_1, \dots, Y_{n_Y}$  from two different populations with medians  $m_X$  and  $m_Y$  respectively.
- Want to test  $H_0: m_X = m_Y$  against a one-sided or two-sided alternative
- Order the **combined** sample and let  $W$  be the sum of the ranks of  $Y_1, \dots, Y_{n_Y}$ . This is the *Wilcoxon rank-sum statistic*.
- Note: this captures information on  $X$  as well as  $Y$ ! (Why?)
- The test based on this statistic is called the *Wilcoxon rank-sum test*, also known as the *Wilcoxon two-sample test* and the *Mann-Whitney U test*.

### Rejection region

- Suppose our alternative hypothesis is  $H_1: m_X > m_Y$
- If  $m_X > m_Y$  then we expect  $W$  to be small, since the  $Y$  values will tend to be smaller than  $X$  and thus have smaller ranks
- Therefore, the critical region should be of the form  $W \leq c$  for a suitable  $c$ .
- Properties of  $W$  (derivation not shown):

$$\begin{aligned}\mathbb{E}(W) &= \frac{n_Y(n_X + n_Y + 1)}{2} \\ \text{var}(W) &= \frac{n_X n_Y (n_X + n_Y + 1)}{12}\end{aligned}$$

- $W$  is approximately normally distributed when  $n_X$  and  $n_Y$  are large

## Alternative definitions

- Like for the one-sample version, the definition of the statistic varies
- We just defined:  $W$  is the sum of the ranks in the  $Y$  sample
- A popular alternative:  $U$  is the number of all pairs  $(X_i, Y_j)$  such that  $Y_j \leq X_i$  (the number of ‘wins’ out of all possible pairwise ‘contests’)
- $U$  and  $W$  are deterministically related (can you derive the formula?)
- $U$  and  $W$  have different (but related) sampling distributions
- Using either statistic leads to equivalent test procedures
- Note:  $\mathbb{E}(U) = n_X n_Y / 2$  and  $\text{var}(U) = \text{var}(W)$

## Example (Wilcoxon two-sample test)

Two companies package cinnamon. Samples of size eight from each company yield the following weights:

$X$	117.1	121.3	127.8	121.9	117.4	124.5	119.5	115.1
$Y$	123.5	125.3	126.5	127.9	122.1	125.6	129.8	117.2

Want to test  $H_0: m_X = m_Y$  versus  $H_1: m_X \neq m_Y$

Use a significance level of 5%

## Using R

- R uses  $U$ ... but calls it  $W$ !
  - For small sample sizes R will use the exact sampling distribution, otherwise it will use a normal approximation
  - To carry out the test, use: `wilcox.test`
  - To work with the sampling distribution of  $U$ , use: `pwilcox`
- ```
> wilcox.test(x, y)
```

Wilcoxon rank sum test

```
data: x and y
W = 13, p-value = 0.04988
alternative hypothesis:
  true location shift is not equal to 0

# Calculate exact p-value manually.
> 2 * pwilcox(13, 8, 8)
[1] 0.04988345
```

We reject  $H_0$  and conclude that we have sufficient evidence to show that the median weights differ between the two companies.

## 3 Goodness-of-fit tests ( $\chi^2$ )

### 3.1 Introduction

#### Goodness-of-fit tests

- How well does a given model fit a set of data?
- E.g. if we assume a Poisson model for a set of data, is it reasonable?
- We can assess this with a ‘goodness-of-fit’ test
- The most commonly used is *Pearson’s chi-squared test*

- Unlike most of the other tests we've seen, this operates on categorical (discrete) data
- Can also apply it on continuous data by first partitioning the data into separate classes

## 3.2 Two classes

### Binomial model

- Start with a binomial model  $Y_1 \sim \text{Bi}(n, p_1)$
- Our usual test statistic for this is

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}} \approx N(0, 1)$$

- Therefore,

$$Q_1 = Z^2 \approx \chi_1^2$$

- To test  $H_0: p = p_1$  versus  $H_1: p \neq p_1$ , we would reject  $H_0$  if  $|Z|$  (and, hence,  $Q_1$ ) is too large.
- Next, notice that

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1-p_1)}$$

- and

$$(Y_1 - np_1)^2 = (n - Y_1 - n(1-p_1))^2 = (Y_2 - np_2)^2$$

where  $Y_2 = n - Y_1$  and  $p_2 = 1 - p_1$ .

- Therefore,

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}$$

- $Y_1$  is the observed number of successes,  $np_1$  is the expected number of successes
- $Y_2$  is the observed number of failures,  $np_2$  is the expected number of failures
- So

$$Q_1 = \sum_{i=1}^2 \frac{(Y_i - np_i)^2}{np_i} = \sum_{i=1}^2 \frac{(O_i - E_i)^2}{E_i} \approx \chi_1^2$$

where  $O_i$  is the observed number and  $E_i$  is the expected number.

- Even though there are two classes, we have only **one** degree of freedom. This is due to the constraint  $Y_1 + Y_2 = n$ .

## 3.3 More than two classes

### Multinomial model

- Generalize to  $k$  possible outcomes (a multinomial model)
- $p_i$  = probability of the  $i$ th class ( $\sum_{i=1}^k p_i = 1$ )
- Suppose we have  $n$  trials, with  $Y_i$  being the number of outcomes in class  $i$
- $\mathbb{E}(Y_i) = np_i$
- Now we get,

$$Q_{k-1} = \sum_{i=1}^k \frac{(Y_i - np_i)^2}{np_i} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \approx \chi_{k-1}^2$$

- $k - 1$  degrees of freedom because  $Y_1 + \dots + Y_k = n$



### Setting up the test

- Specify a categorical distribution:  $p_1, p_2, \dots, p_k$
- We use the  $Q_{k-1}$  statistic to test whether data are consistent with this distribution
- The null hypothesis is that they do (i.e. the  $p_i$  define the distribution)
- The alternative is that they do not (i.e. a different set of probabilities define the distribution)
- Under the null, the test statistic will tend to be small (it measures ‘badness-of-fit’)
- Therefore, reject the null if  $Q_{k-1} > c$  where  $c$  is the  $1 - \alpha$  quantile from  $\chi^2_{k-1}$ .

### Remarks

- We are approximating a binomial with a normal
- Good approximation if  $n$  is large and the  $p_i$  are not too small
- Rule of thumb: need to have all  $E_i = np_i \geq 5$
- The larger the  $k$  (i.e. more classes), the more powerful the test. However, we need the classes to be large enough
- If any of the  $E_i$  are too small, can combine some of the classes until they are large enough
- If  $Q_{k-1}$  is very small, this indicates that the fit is ‘too good’. This can be used as a test for rigging of experiments / fake data. Typically need very large  $n$  to do this.
- Often refer to the test statistic as  $\chi^2$

### Example (completely specified distribution)

- Proportions of commuters using various modes of transport, based on past records:

| Bus  | Train | Car  | Other |
|------|-------|------|-------|
| 0.25 | 0.15  | 0.50 | 0.1   |

- After a 3-month campaign, a random sample ( $n = 80$ ) found:

| Bus | Train | Car | Other |
|-----|-------|-----|-------|
| 26  | 15    | 32  | 7     |

- Did the campaign alter commuters behaviour?
- The expected frequencies are:

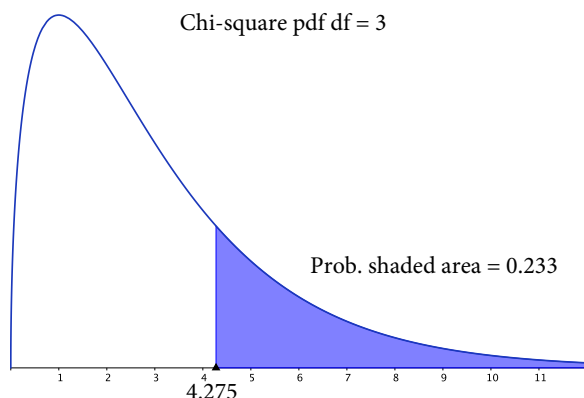
| Bus | Train | Car | Other |
|-----|-------|-----|-------|
| 20  | 12    | 40  | 8     |

- The value of the test statistic is:

$$\chi^2 = \frac{(26 - 20)^2}{20} + \frac{(15 - 12)^2}{12} + \frac{(32 - 40)^2}{40} + \frac{(7 - 8)^2}{8} = 4.275$$

- $H_0$ : proportions have not changed,  $H_1$ : proportions have changed
- We have 4 classes, so the test statistic here has a  $\chi^2_3$  distribution.
- The 0.95 quantile is 7.81, which is greater than  $\chi^2 = 4.275$
- Therefore, there is **insufficient** evidence that the proportions have changed
- The p-value is

$$p = \Pr(\chi^2_3 > 4.275) = 0.233 > 0.05$$



### Using R

```
> x <- c( 26, 15, 32, 7)
> p <- c(0.25, 0.15, 0.5, 0.1)
> t1 <- chisq.test(x, p = p)
> t1
```

Chi-squared test for given probabilities

```
data: x
X-squared = 4.275, df = 3, p-value = 0.2333
> rbind(t1$observed, t1$expected)
      [,1] [,2] [,3] [,4]
[1,]  26  15  32   7
[2,]  20  12  40   8

> t1$residuals
[1]  1.3416408  0.8660254 -1.2649111 -0.3535534

> sum(t1$residuals^2)
[1] 4.275

> 1 - pchisq(4.275, 3)
[1] 0.2332594
```

## 3.4 Estimating parameters

### Fitting distributions

- We don't always have an exact model to compare against
- We might specify a family of distributions but still need to estimate some of the parameters
- For example,  $P_n(\lambda)$  or  $N(\mu, \sigma^2)$
- We would need to estimate the parameters using the sample, and use these to specify  $H_0$
- We need to adjust the test to take into account that we've used the data to define  $H_0$  (by design, it will be 'closer' to the data than if it we didn't need to do this)
- The 'cost' of this estimation is 1 degree of freedom for each parameter that is estimated
- The final degrees of freedom is  $k - p - 1$ , where  $p$  is the number of estimated parameters

### Example (Poisson distribution)

- $X$  is number of alpha particles emitted in 0.1 sec by a radioactive source
- Fifty observations:  
7, 4, 3, 6, 4, 4, 5, 3, 5, 3, 5, 5, 3, 2, 5, 4, 3, 3, 7, 6, 6, 4, 3, 9, 11, 6, 7, 4, 5, 4, 7, 3, 2, 8, 6, 7, 4, 1, 9, 8, 4, 8, 9, 3, 9, 7, 7, 9, 3, 10
- Is a Poisson distribution an adequate model for the data?
- $H_0$ : Poisson,  $H_1$ : something else
- We have only specified the family of the distribution, not the parameters
- Estimate the Poisson rate parameter  $\lambda$  by the MLE,  $\hat{\lambda} = \bar{x} = 5.4$
- Now we ask: does the  $Pn(5.4)$  model give a good fit?

First, find an appropriate partition of the value (collapse the data):

```
> X1 <- cut(X, breaks = c(0, 3.5, 4.5, 5.5, 6.5, 7.5, Inf))
> T1 <- table(X1)
> T1
X1
(0,3.5] (3.5,4.5] (4.5,5.5] (5.5,6.5] (6.5,7.5] (7.5,Inf]
      13         9         6         5         7        10
```

Then, prepare the data for the test:

```
> x <- as.numeric(T1)
> x
[1] 13  9  6  5  7 10

> n <- sum(x)
> p1 <- sum(dpois(0:3, 5.4));
> p2 <- dpois(4, 5.4)
> p3 <- dpois(5, 5.4)
> p4 <- dpois(6, 5.4)
> p5 <- dpois(7, 5.4)
> p6 <- 1 - (p1 + p2 + p3 + p4 + p5)
> p <- c(p1, p2, p3, p4, p5, p6)
```

Then, run the test:

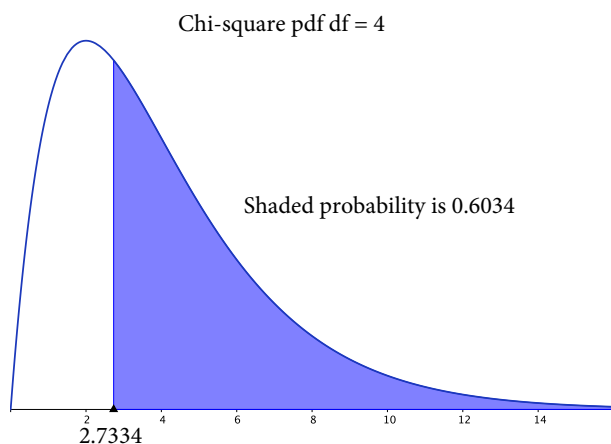
```
> chisq.test(x, p = p)
```

Chi-squared test for given probabilities

```
data:  x
X-squared = 2.7334, df = 5, p-value = 0.741
```

But this is the wrong df! Need to adjust manually:

```
> 1 - pchisq(2.7334, 4)
[1] 0.6033828
```



- Needed to adjust p-values as we have estimated the mean
- The critical value is the 0.95 quantile from  $\chi_4^2$ , which is 9.488, so we cannot reject  $H_0$
- Not enough evidence against the Poisson model
- Therefore, this is an adequate fit (at least, until further data proves otherwise)

|          | 0-3  | 4   | 5   | 6   | 7   | 8+   |
|----------|------|-----|-----|-----|-----|------|
| Observed | 13.0 | 9.0 | 6.0 | 5.0 | 7.0 | 10.0 |
| Expected | 10.7 | 8.0 | 8.6 | 7.8 | 6.0 | 8.9  |

## 4 Tests of independence (contingency tables)

### Contingency tables

- Suppose we have multiple categorical variables (which could be continuous variables partitioned into classes)
- A *contingency table* records the number of observations for each possible cross-classification of these variables
- We are often interested in whether two categorical variables are related to each other
- For example, height and weight
- Define height classes  $A_1, \dots, A_r$ , and weight classes  $B_1, \dots, B_c$
- Each person is assigned to a single combination  $(A_i, B_j)$
- A sample of people can be summarised with a  $r \times c$  table of counts (a contingency table)

### Independence model

- A general model for these data is:

$$p_{ij} = \Pr(A_i \cap B_j), \quad i = 1, \dots, r, \quad j = 1, \dots, c$$

- Are the two variables independent?
- We can set this up as a hypothesis test:

$$H_0: p_{ij} = \Pr(A_i) \Pr(B_j) \quad \text{versus} \quad H_1: p_{ij} \neq \Pr(A_i) \Pr(B_j)$$

- This has the same structure as a goodness-of-fit test, can use Pearson's chi-squared statistic
- Show how this works through an example...

### Example (contingency table)

150 executives were classified by sex,  $A$ , and whether or not they were firstborn,  $B$ :

|        | Firstborn | Not firstborn | Total |
|--------|-----------|---------------|-------|
| Male   | 34        | 74            | 108   |
| Female | 20        | 22            | 42    |
| Total  | 54        | 96            | 150   |

Let's test whether these two variables are independent.

### Estimating the marginals

- Recall discrete bivariate distributions:

|        | Firstborn     | Not firstborn | Total        |
|--------|---------------|---------------|--------------|
| Male   | $p_{11}$      | $p_{12}$      | $p_{1\cdot}$ |
| Female | $p_{21}$      | $p_{22}$      | $p_{2\cdot}$ |
| Total  | $p_{\cdot 1}$ | $p_{\cdot 2}$ | 1            |

- The marginals are:

$$p_{i\cdot} = \sum_{j=1}^c p_{ij} = \Pr(A_i)$$
$$p_{\cdot j} = \sum_{i=1}^r p_{ij} = \Pr(B_j)$$

- The null hypothesis of independence is just,  $H_0: p_{ij} = p_{i\cdot}p_{\cdot j}$
- Data:

|        | Firstborn     | Not firstborn | Total        |
|--------|---------------|---------------|--------------|
| Male   | $y_{11}$      | $y_{12}$      | $y_{1\cdot}$ |
| Female | $y_{21}$      | $y_{22}$      | $y_{2\cdot}$ |
| Total  | $y_{\cdot 1}$ | $y_{\cdot 2}$ | $n$          |

- Estimates:

$$\hat{p}_{i\cdot} = \frac{y_{i\cdot}}{n}$$
$$\hat{p}_{\cdot j} = \frac{y_{\cdot j}}{n}$$

where

$$y_{i\cdot} = \sum_{j=1}^c y_{ij}$$
$$y_{\cdot j} = \sum_{i=1}^r y_{ij}$$

- Pearson's  $\chi^2$  statistic for given  $p_{ij}$  is

$$Q = \sum_i \sum_j \frac{(Y_{ij} - np_{ij})^2}{np_{ij}}$$

- Under  $H_0$ , an estimator of  $p_{ij}$  is

$$\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{Y_{i\cdot}Y_{\cdot j}}{n^2}$$

- This gives the following,

$$Q = \sum_i \sum_j \frac{(Y_{ij} - Y_{i\cdot}Y_{\cdot j}/n)^2}{Y_{i\cdot}Y_{\cdot j}/n} \approx \chi_{(r-1)(c-1)}^2$$

## Explanation for degrees of freedom

- Recall that we should have  $k - p - 1$  degrees of freedom
- Here,  $k = rc$ , the total number of cells in the table
- We estimated  $r - 1$  marginal probabilities for the rows and  $c - 1$  for the columns, which makes  $p = (r - 1) + (c - 1)$
- Therefore, the number of degrees of freedom remaining is:

$$df = rc - (r - 1) - (c - 1) - 1 = (r - 1)(c - 1)$$

## Using R: set up the data

```
> x <- rbind( male = c(first = 34, later = 74),
+           female = c(first = 20, later = 22))
> x
```

|        | first | later |
|--------|-------|-------|
| male   | 34    | 74    |
| female | 20    | 22    |

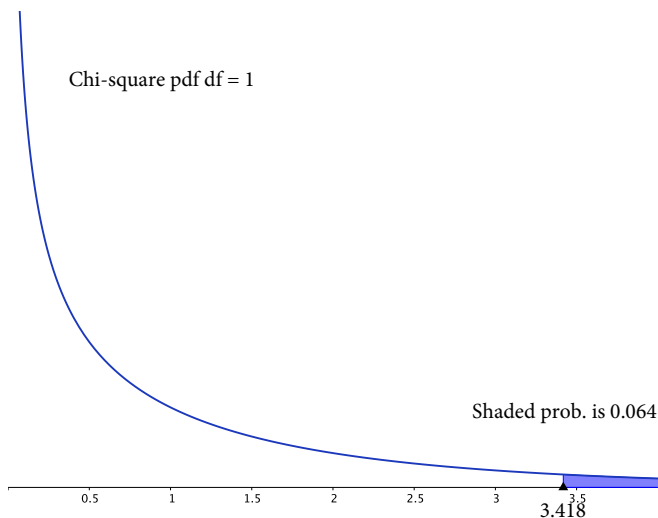
## Using R: run the test

```
> c1 <- chisq.test(x, correct = FALSE)
> c1
```

### Pearson's Chi-squared test

```
data:  x
X-squared = 3.418, df = 1, p-value = 0.06449
```

We do not have enough evidence to reject  $H_0$  at a 5% significance level.



## Using R: more output

```
> c1$observed
```

|        | first | later |
|--------|-------|-------|
| male   | 34    | 74    |
| female | 20    | 22    |

```
> c1$expected
```

|        | first | later |
|--------|-------|-------|
| male   | 38.88 | 69.12 |
| female | 15.12 | 26.88 |