MATH10096

Applied Statistics

2020-21

Tim Cannings [1]
Email[2]: timothy.cannings@ed.ac.uk

THE UNIVERSITY
of EDINBURGH

SCHOOL OF MATHEMATICS

[1]Many thanks to Serveh Sharifi for providing material for the course
[2]Please point out any typos or corrections via Email with the subject "Applied Stats Correction"

Preface

Course outline

Many of the standard statistical tests, for example *t*-tests, depend for their theoretical justification on the assumption that the population being sampled is normally distributed. Real populations may differ significantly from normal, which suggests that there is need to seek procedures which are not dependent on the specific form of the distribution being sampled. For many problems such procedures can indeed be found, and are termed *distribution-free* or *nonparametric*.

In this course, we will first study some statistical tests that assess whether the data are sampled from a particular distribution (for example from a normal distribution). Then a variety of nonparametric tests for comparing two or more populations are discussed, including§ the Mann-Whitney test, the Wilcoxon signed-rank test, the sign test, randomisation tests, the run test, the Kruskal-Wallis test.

Nonparametric methods can be extended to generalised linear models, Bootstrap methods, and Bayesian nonparametric methods, but they will not be discussed here as some of them are covered in other courses like Statistical Computing and Nonparametric Regression Models.

Textbooks

There are several textbooks which cover the material and can be used for extra help:

- W. J. Conover, (1999), *Practical Nonparametric Statistics*, 3rd edition, Wiley. library link
- J. Kloke and J. W. McKean, (2015), *Nonparametric Statistical Methods Using R*, CRC Press.
- B. F. J. Manly, (1997), *Randomisation, Bootstrap and Monte Carlo Methods in Biology*, Chapman & Hall. library link
- I. P. Sprent and N.C. Smeeton, (2001), *Applied Nonparametric Statistical Methods*, 3rd edition, Chapman & Hall. library link

Contents

Chapter 1

Introduction

We begin by briefly recapping some useful topics in statistics and probability – much of the introduction should be revision, but do take some time to familiarise yourself with the material.

1.1 Types of data

The modelling and analysis that has been covered in previous courses (for example in Statistics 2) has typically concerned continuous data. In this course, we will look at the analysis of categorical data data as well. These data are often in the form of frequencies (or counts) in a number of different categories (or cells). For example this may be the number of students graduating in Mathematics in each of different degree classifications (for example, Mathematics and Statistics, Mathematics and Biology, Mathematics and Music, etc).

We distinguish between four types of data:

1. **Nominal data:** This is the weakest type of "measurement". We simply classify the members of a population into categories, so the data are simply labels and just qualitative. There is no particular relationship between the categories. For a well-defined nominal scale, the categories must be

   (a) exhaustive (each member belongs to some category)
   (b) mutually exclusive (no member belongs to more than one category).

   A number can be assigned to each class, but the number assists only as a label (e.g. convenient to use 0 or 1 if there are only two categories).

   Examples Colour, gender, nationality, blood group of individuals in a sample.

2. **Ordinal data:** Again each member of the population belongs to exactly one category, but now the categories are ordered. The differences between measurements on an ordinal scale are not informative (e.g. in athletics, the gap between 1st and 4th may be smaller than that between 5th and 6th). Many nonparametric tests require the data to be *ranked* in increasing order, and are easier to implement if the number of *tied* ranks (e.g. equal 3rd) is small.

   Example Height of tall/medium/short, degree classes, Mohs hardness scale (what scratches what), in a questionnaire respondents might be asked to rate something as 'bad', 'average' or 'good'.

3. **Interval data:** Data of this type satisfy the requirements of ordinal data, but data are measured in equal units and the difference between any two numbers on the scale must also be of known size. An interval scale involves both a unit distance and a zero distance, but both these quantities are arbitrary. e.g. Celsius and Fahrenheit scales have different zeros, and the Celsius unit is 1.8 times the Fahrenheit one.

   Examples Temperature in degrees Celsius: e.g. at 12.00 GMT today the temperature in Edinburgh was 8°C, while it was 24°C in Cape Town. Thus it was 16°C cooler in Edinburgh than in Cape Town.

4. **Ratio data:** Assume the temperature in Edinburgh was 8°C at 12.00 GMT today while it was 24°C in Cape Town. Do these temperatures indicate that it was three times as hot in Cape Town as in Edinburgh at 12.00 GMT today?

   No. Because the zero on the Celsius scale is arbitrary in contrast with degrees Kelvin.

   The strongest form of measurement, the ratio scale, possesses the properties of the interval scale, but the ratio between two measurements is also meaningful. Thus on a ratio scale, the zero is naturally defined but the unit is still arbitrary. Ratio data are real numbers and they can be subjected to standard mathematical procedures (e.g., addition, subtraction, multiplication, division).

   Example Height, income, weight (e.g. someone weighing 100kg is twice the weight of someone weighing 50kg).

In practice, observations are usually treated as either categorical (nominal, ordinal) or continuous (interval, ratio). Most parametric tests need interval or ratio measurements, but nonparametric ones often require only nominal or ordinal data.

1.2 Some revision

1.2.1 Revision: The χ^2 distribution

In this course, we will make use of the χ^2 (or chi-squared) distribution. Recall that the chi-squared distribution is defined as follows: Given $k \in \{1, 2, \ldots, \}$, let $Z_1, \ldots, Z_k$ be independent and identically distributed $N(0, 1)$ random variables, then $Y = Z_1^2 + \ldots + Z_k^2$ has a chi-squared distribution with k degrees of freedom, and we write

$$Y \sim \chi^2_k.$$

Some basic properties: This has probability density function (p.d.f.)

$$f_Y(y) = \frac{y^{k/2-1} \exp(-y/2)}{2^{k/2}\Gamma(k/2)}; \qquad y > 0.$$

where Γ is the gamma function: $\Gamma(t) := \int_0^\infty x^{t-1} \exp(-x)\, dx$. We have that $\mathbb{E}(Y) = k$ and $\mathrm{Var}(Y) = 2k$.

1.2.2 Revision: Hypothesis testing

The general framework for hypothesis testing that we will use is as follows.

- **Step 1**: State our *null hypothesis* H_0 and *alternative hypothesis* H_1.
- **Step 2**: Calculate an appropriate *test statistic* T, and its observed value based on the data t.
- **Step 3**: Determine the distribution of the test statistic T under H_0.
- **Step 4**: Calculate the *critical region* C for the test statistic for a given *significance level* (or size) $\alpha \in (0, 1)$ – this is the set of values of the test statistic for which we will reject H_0;
- **Step 5**: Conclude the test:
  - If $t \in C$, then we reject H_0 in favour of H_1.
  - If $t \notin C$, then we do not reject H_0 in favour of H_1

Some key definitions:

Type I error: reject H_0 when H_0 is true.

Type II error: do not reject H_0 when H_1 is true.

Size: $\alpha = \mathbb{P}(\text{Type I error}) = \mathbb{P}(\text{Reject } H_0 | H_0 \text{ True})$.

Power: $1 - \beta = 1 - \mathbb{P}(\text{Type II error}) = \mathbb{P}(\text{Reject } H_0 | H_1 \text{ True})$.

***p*-value**: For a real valued test statistic T and a critical region of the form $C = \{T > c\}$, the p-value given an observed value of the test statistic t is $\mathbb{P}(T \geq t | H_0 \text{ True})$. In this case we may carry out a test of size α by rejecting H_0 if the p-value is less than α.

1.2.3 Revision: The multinomial distribution

The multinomial distribution is a generalisation of the binomial distribution to the case where there are more than two possible outcomes. Recall that the Binomial distribution models the number X of successes in $n \geq 1$ (where n is known) independent Bernoulli trials, each of which has two possible outcomes ('success' and 'failure') with probabilities $p \in [0, 1]$ (typically unknown) of 'success' and $1 - p$ of 'failure' – in short, we write $X \sim \text{Bin}(n, p)$.

We are often interested in modelling n independent trials, each of which results in exactly one of $K \geq 2$ outcomes or categories, (for example, the responses 'yes', 'no' and 'don't know' to a question in a questionnaire). For $k = 1, \ldots, K$, let O_k denote the number of trials which result in outcome k. Then, we have

$$\sum_{k=1}^{K} O_k = n,$$

and $(O_1, \ldots, O_K)$ are said to have a *multinomial distribution*. We will write

$$(O_1, \ldots, O_K) \sim \text{MN}(n; p_1, \ldots, p_K)$$

where p_k is the probability that any given trial results in outcome k. Since each trial results in exactly one outcome, we have that $\sum_{k=1}^{K} p_k = 1$. We refer to $(O_1, \ldots, O_K)$ as the observed frequencies.

Additional properties: we have

$$\begin{aligned} E_k := \mathbb{E}(O_k) &= np_k \qquad k = 1, \ldots, K; \\ E_1 + \cdots + E_K &= n. \end{aligned}$$

The values $E_1, \ldots, E_K$ are called the expected frequencies. The probability mass function is

$$\mathbb{P}\{(O_1, \ldots, O_K) = (o_1, \ldots, o_K)\} = \frac{n!}{o_1! \ldots o_K!} p_1^{o_1} \ldots p_K^{o_K},$$

for $o_1, \ldots o_K \in \{0, 1, \ldots, n\}$ and $\sum_{k=1}^{K} o_k = n$.

Of course, the binomial distribution is a special case of the multinomial distribution with $K = 2$: we have

$$X \sim \text{Bin}(n, p) \Leftrightarrow (X, n - X) \sim \text{MN}(n; p, 1 - p).$$

Example 1 Suppose we roll a fair six-sided die n times and count the number of times O_k each number $k = 1, \ldots, 6$ is obtained. Then

$$(O_1, \ldots, O_6) \sim \text{MN}(n; 1/6, 1/6, 1/6, 1/6, 1/6, 1/6).$$

Exercise: Try an experiment by rolling a dice many times – what do you notice about the observed counts as n gets larger? Confirm your thinking by repeating the experiment in `R` with a very large value of n.

Example 2 The following table gives the results of an experiment counting number of sixes in $n = 216$ trials, where each trial consisted of rolling three (possibly unfair) 6-sided dice.

Count	0	1	2	3
Frequency	110	85	20	1

In this case we can model $(O_1, O_2, O_3, O_4) \sim \text{MN}(216; p_1, p_2, p_3, p_4)$ – since we don't know whether or not the dice are fair in this case, the probabilities p_1, p_2, p_3, p_4 are unknown. In the next chapter, we will see how we can formally test the hypothesis: *are the dice fair?*

Chapter 2

Goodness-of-fit tests

In this chapter we will learn about *goodness-of-fit* tests. In the simplest case, we aim to test whether the data arose from the distribution F_0, where F_0 is a prespecified cumulative distribution function (CDF) of interest. More generally, we may be in interested in whether the distribution of our data belongs to a particular *family* of distributions.

We will consider two goodness-of-fit tests; 1) the *chi-square-test* (or χ^2-test) for categorical data; and 2) the *Kolmogorov-Smirnov* test for continuous data. These tests allow a formal statistical assessment of whether a distribution adequately fits data.

2.1 The χ^2 goodness-of-fit test

In Example 2 above, we wish to test the hypothesis that the dice are fair – we can use the **chi-square test**! We model the counts $(110, 85, 20, 1)$ with a multinomial distribution $\text{MN}(216; p_1, p_2, p_3, p_4)$. In this example, $K = 4$ and $n = 216$, and, **if the dice is fair** we have

$$\begin{aligned}
\mathbb{P}(\text{No sixes in 3 throws}) &= (5/6)^3 = 125/216 \\
\mathbb{P}(\text{1 six in 3 throws}) &= 3(1/6)(5/6)^2 = 75/216 \\
\mathbb{P}(\text{2 sixes in 3 throws}) &= 3(1/6)^2(5/6) = 15/216 \\
\mathbb{P}(\text{3 sixes in 3 throws}) &= (1/6)^3 = 1/216,
\end{aligned}$$

where we have used the Binomial distribution to calculate the probabilities. Testing if the dice are fair then corresponds to testing whether $(p_1, \ldots, p_4) = (125/216, 75/216, 15/216, 1/216)$; in this Section we will see how this can be done by assessing (formally) whether the observed counts are close to the expected counts under the null hypothesis.

Let's return the to general case: Suppose that $(O_1, \ldots, O_K) \sim MN(n; p_1, \ldots, p_K)$, where n and K are fixed and known, but $(p_1, \ldots, p_K)$ is a vector of unknown probabilities satisfying $\sum_{k=1}^{K} p_k = 1$. We wish to test the hypothesis

$$H_0 : (p_1, \ldots, p_K) \in P_0 \quad vs. \quad H_1 : p_1, \ldots, p_K \quad \text{unrestricted}$$

where the set P_0 of probabilities can be either:

1. specified fully: i.e. $P_0 = \{(p_1^*, \ldots p_K^*)\}$; or

2. specified as a function of the parameter θ, denoted $p_1(\theta), \ldots, p_k(\theta)$, where θ is a vector of parameters with **unknown** values in some set $\Theta \subseteq \mathbb{R}^q$: i.e. $P_0 = \{(p_1(\theta), \ldots p_K(\theta)) : \theta \in \Theta\}$

We will deal with these two cases separately, but note that the first is a special case of the second with $q = 0$.

1. In the fully specified case, the expected values of $O_1, \ldots, O_K$ under the null hypothesis are given by $E_k = np_k^*$, for $k = 1, \ldots, K$. We compare the observed counts $O_1, \ldots, O_K$ with the corresponding expected values $E_1, \ldots, E_K$ using *Pearson's chi-squared statistic*
$$X^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}. \tag{2.1}$$
we will **reject** H_0 if the the test statistic X^2 is large. To find the critical value to give the appropriate size, under H_0, we have that[1]
$$X^2 \stackrel{\cdot}{\sim} \chi^2_{K-1}. \tag{2.2}$$

2. When P_0 specifies a set of possible probabilities indexed by an unknown parameter θ, there are a few (minor) changes: now the expected count for the kth outcome is $np_k(\theta)$, but θ is unknown! We will estimate it using the maximum likelihood estimator $\hat{\theta}$, and let $E_k = np_k(\hat{\theta})$, the the test statistic is X^2 the same as above. In this case, since there are q unknown parameters in θ, we have that
$$X^2 \stackrel{\cdot}{\sim} \chi^2_{K-q-1} \qquad \text{under } H_0. \tag{2.3}$$

Note (i) The usual rule of thumb for the approximations (2.2) and (2.3) is that they are reasonable if $E_k \geq 5$ for $k = 1, \ldots, K$.

(ii) Since X^2 measures the discrepancy between the observed counts $O_1, \ldots, O_k$ and the expected counts $E_1, \ldots, E_k$, H_0 is only rejected for 'large' values of X^2, so that this test is 1-tailed.

(iii) The way to remember the number of degrees of freedom needed is:

$d.f. = K - q - 1 = \#\{\text{potential outcomes}\} - \#\{\text{unknown parameters estimated}\} - 1.$

In general, this is $\#\{\text{unknown parameters under } H_1\} - \#\{\text{unknown parameters under } H_0\}$. Where $\#\{\text{unknown parameters under } H_1\} = K - 1$: we have $p_1, \ldots, p_K$, subject to $\sum_{k=1}^{K} p_k = 1$. This second formulation is important to remember when we test for homogeneity in a contingency table – see Section 2.3.2.

(iv) Some heuristic understanding of (2.2): suppose $K = 2$ the marginal distribution of the count O_1 for outcome 1 can be modelled as a Binomial random variable, i.e. $O_1 \sim \text{Bin}(n, p_1)$. Therefore, using the Normal approximation to the Binomial distribution, we have
$$\frac{O_1 - np_1}{\sqrt{np_1(1-p_1)}} \stackrel{\cdot}{\sim} N(0,1).$$
Now
$$\frac{(O_1 - np_1)^2}{np_1(1-p_1)} = \frac{(O_1 - np_1)^2}{np_1} + \frac{\{n - O_1 - n(1-p_1)\}^2}{n(1-p_1)} = \frac{(O_1 - np_1)^2}{np_1} + \frac{(O_2 - np_2)^2}{np_2}.$$

[1] the notation $\stackrel{\cdot}{\sim}$ means that X^2 is approximately distributed as χ^2_{K-1}. More formally, we have that $X^2 \to^d \chi^2_{K-1}$ as $n \to \infty$.

Therefore, since $E_k = np_k^*$ under H_0, we have

$$\sum_{k=1}^{2} \frac{(O_k - E_k)^2}{E_k}$$

is (approximately) the square of a standard normal random variables, suggesting that this has a χ_1^2 distribution (when $K = 2$). In general, we have K components in the sum, but $O_1, \ldots, O_K$ are not independent – they satisfy the constraints that $O_1 + \cdots + O_k = n$. Thus, if we know $K - 1$ of $O_1, \ldots, O_K$ then the remaining one is completely determined, which suggests that we only have $K - 1$ degrees of freedom, i.e. that $X^2 \stackrel{\cdot}{\sim} \chi^2_{K-1}$. For (2.3) things are a little more complicated – see (non-examinable) Section 2.1.1.

Example 2 (continued) Recall the results table from Example 2.

Count	0	1	2	3
Frequency	110	85	20	1

We wish to test the hypothesis that the dice are fair, i.e. test whether or not the data are consistent with a $Bin(3, 1/6)$ distribution at $\alpha = 0.05$. Let Y denote the number of 6's that are thrown with the 3 dice on a single trial so that we can specify

$$H_0 : Y \sim Bin(3, 1/6); \quad \text{vs.} \quad H_1 : Y \nsim Bin(3; 1/6).$$

or equivalently

$$H_0 : (p_1, p_2, p_3, p_4) = (125/216, 75/216, 15/216, 1/216) \quad \text{vs.} \quad H_1 : p_k \text{ unspecified.}$$

Answer: We begin by calculating the expected frequencies under H_0. If $Y \sim Bin(3, 1/6)$, the associated probabilities for each possible value is given by,

y	0	1	2	3
$\mathbb{P}(Y = y)$	125/216	75/216	15/216	1/216

Thus we have the expected values $E_k = np_k^* = 216p_k^*$:

	0	1	2	3
O_k	110	85	20	1
E_k	125	75	15	1

Here the expected cell probabilities are completely specified and we are in case 1 above and thus use (2.2). The test statistic is given by:

$$\sum_{i=1}^{4} \left(\frac{(O_i - E_i)^2}{E_i} \right) = 4.8.$$

Under H_0: we have that $X^2 \stackrel{\cdot}{\sim} \chi^2_{K-1} = \chi^2_3$. Then the approximate p-value is $\mathbb{P}(\chi^2_3 > 4.8) = 0.19$. Thus, there is no evidence to reject H_0 corresponding to the dice being fair. This can be obtained in R using: `1-pchisq(4.8,3)` which gives 0.1870417.

However, we did not satisfy the rule of thumb that $E_i > 5$. We could have remedied this by pooling (i.e. combining) the last two groups. This would give:

	0	1	≥ 2
O_k	110	85	21
E_k	125	75	16

Now, the test statistic is 4.6958 and under H_0, $X^2 \dot{\sim} \chi^2_2$. Then, $\mathbb{P}(\chi^2_2 > 4.6958) = 0.096$, so the conclusion remains the same.

Using R

Here is R code to do all this.

```
> n <- 216
> y <- 0:3
  # 3 rolls of the die; p(success)=1/6 on each trial
> p <- dbinom(y,3,1/6);  p
[1] 0.57870370 0.34722222 0.06944444 0.00462963
  # pool last two cells (count = 2 and 3)
> E <- n*c(p[1],p[2],(p[3]+p[4])); E
[1] 125  75  16
  # pool last two cells (count = 2 and 3)
> O <- c(110,85,21); O
[1] 110  85  21
> X2 <- sum((O-E)^2/E); X2
[1] 4.695833
> df <- length(O)-1; df
[1] 2
  # p-value
> pchisq(X2,df,lower.tail=FALSE)
[1] 0.09556806
```

The inbuilt R function to conduct this type of hypothesis test is `chisq.test`:

```
> O <- c(110 , 85, 20, 1)
> probs <- dbinom(0:3, 3, 1/6) # Calculate the probabilities for
                                 each of the Multinomial cells
> chisq.test(O,p=probs)

Chi-squared test for given probabilities
data:  O
X-squared = 4.8, df = 3, p-value = 0.187

Warning message:
In chisq.test(O, p=probs): Chi-squared approximation may be incorrect
```

Thus we obtain the same results as above. However, note the warning message. This is because the expected cell counts are not all > 5. We can again conduct this in R using:

```
> O1 <- c(110 , 85 , 21)
> p <- c(probs[1:2],probs[3]+probs[4])
> chisq.test(O1,p=p)
```

```
Chi-squared test for given probabilities
data:  O1
X-squared = 4.6958, df = 2, p-value = 0.09557
```

Example 3: hypothesised distribution with one unknown parameter Suppose we observe the following data

	0	1	2	3	4	5	6
O_k (frequency)	121	85	19	1	0	0	1

where the counts arise from unobserved independent and identically distributed discrete random variables $X_1, \ldots, X_n$. Note that we do not observe $X_1, \ldots, X_n$ themselves, only the counts $O_k = \#\{X_i = k-1\}$ for $k = 1, \ldots, 7$. Let X denote the a generic random variable with the same distribution as X_i.

We wish to test the following hypothesis at $\alpha = 0.05$ level:

$$H_0 : X \sim Poisson(\lambda) \quad \text{for some } \lambda > 0$$

against

$$H_1 : X \nsim Poisson(\lambda) \quad \text{for any } \lambda > 0.$$

Answer: We note that the null hypothesis does not specify the value of λ – therefore we are in case 2 above and have one unknown parameter. We will estimate λ using the MLE. Recall that (e.g. from Statistics2 and Statistical Methodology), if $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$, then the MLE for λ is

$$\hat{\lambda} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We do not observe $X_1, \ldots, X_n$, but in this case we have $\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{k=1}^{K}(k-1)O_k$. For the above data we have $n = 227$ and $\hat{\lambda} = \frac{1}{227}\sum_{k=1}^{K}(k-1)O_k = 0.5815$. Now to calculate E_i we first need to calculate $p_k(\hat{\lambda})$. Under H_0, we know $X \sim \text{Poisson}(\lambda)$, so that

$$p_k(\lambda) = \mathbb{P}_\lambda(X = k-1) = \frac{\exp(-\lambda)\lambda^{k-1}}{(k-1)!}$$

for $k = 0, 1, 2, 3 \ldots,$. Then plugging in $\lambda = \hat{\lambda} = 0.5815$ and calculating $np_k(\hat{\lambda})$ we obtain:

	0	1	2	3	4	5	6
O_k	121	85	19	1	0	0	1
E_k	126.91	73.80	21.46	4.16	0.60	0.07	0.007

Notes

- The calculation of E_k is most easily done in R, for example, using the following command:

```
> dpois(0:6,0.5815)*227
[1] 1.269064e+02 7.379609e+01 2.145621e+01 4.158929e+00 6.046043e-01
    7.031548e-02 6.814742e-03
```

- We have only tabulated E_k for values ≤ 6 but of course for the Poisson random variable there is no upper limit (we have that for $Y \sim Poisson(\hat{\lambda})$, $\mathbb{P}(Y \geq 7) = 2.68 \times 10^{-6}$ so that $E_{\geq 7} = 0.0006$).
- Given the small E_k we pool for values ≥ 3, so that we consider:

	0	1	2	≥ 3
O_k	121	85	19	2
E_k	126.91	73.80	21.46	4.84

The test statistic is given by:

$$X^2 = \sum_{k=1}^{4} \frac{(O_k - E_k)^2}{E_k} = 3.9246.$$

In this case we use (2.3) – we have $K = 4$ classes and $q = 1$ unknown parameters (λ) – therefore we have $K - q - 1 = 2$ degrees of freedom. The critical point for a χ^2_2 distribution at level 0.05 is 5.991. Thus we do not reject the null hypothesis.

Note that we cannot use the `chisq.test` function in `R` to conduct the hypothesis test since we are unable to tell it the correct number of degrees of freedom. However, we could use the command to calculate the observed test statistic and use this with the correct χ^2 distribution to conduct the hypothesis test.

Using R

```
> y <- c(0,1,2,3,4,5,6)
> O <- c(121,85,19,1,0,0,1)
> mean <- sum(y*O)/227
> E <- dpois(0:6,0.5815)*227
> prob <- dpois(0:6,0.5815)
> O1 <- c(121,85,19,2)
> prhs <- 1-ppois(6, mean)
> prob1 <- c(prob[1:3],prob[4]+prob[5]+prob[6]+prob[7]+prhs)
> chisq.test(O1,p=prob1)

Chi-squared test for given probabilities

data:  O1
X-squared = 3.9246, df = 3, p-value = 0.2697

Warning message:
In chisq.test(O1, p=prob1):Chi-squared approximation may be incorrect

> p.value <- 1-pchisq(3.9246, 2);
> p.value
[1] 0.1405348
```

2.1.1 Non-examinable: Generalised likelihood ratio tests

Where does the test statistic X^2 come from? Intuitively it's measuring the difference between the observe counts and the expected counts under the null hypothesis. The statistical derivation comes from the *generalised likelihood ratio test*: Suppose that

$$H_0 : (p_1, \ldots, p_K) \in \{(p_1(\theta), \ldots, p_K(\theta)) : \theta \in \Theta\} \quad vs. \quad H_1 : p_1, \ldots, p_K \quad \text{unrestricted}$$

where $\theta = (\theta_1, \ldots, \theta_q)$ has q unknown components.

Then, under H_0, the generalised likelihood ratio test statistic

$$2\log\{L(H_0, H_1)\} = 2\log\{L(H_1)\} - 2\log\{L(H_0)\} \stackrel{\cdot}{\sim} \chi^2_{K-q-1}$$

where $L(H)$ denotes the maximum likelihood under the hypothesis H. In this case, we have

$$2\log\{L(H_1)\} - 2\log\{L(H_0)\} = 2\sum_{k=1}^{K} O_k \log(O_k/n) - 2\sum_{k=1}^{K} O_k \log(p_k(\hat{\theta})) = 2\sum_{k=1}^{K} O_k \log\Big(\frac{O_k}{np_k(\hat{\theta})}\Big),$$

where $\hat{\theta}$ is the MLE for θ. Finally, writing $\delta_k = O_k - E_k$, observe that

$$\begin{aligned} 2\sum_{k=1}^{K} O_k \log\Big(\frac{O_k}{np_k(\hat{\theta})}\Big) = 2\sum_{k=1}^{K} O_k \log\Big(\frac{O_k}{E_k}\Big) &= 2\sum_{k=1}^{K} (\delta_k + E_k)\log\Big(1 + \frac{\delta_k}{E_k}\Big) \\ &= 2\sum_{k=1}^{K} (\delta_k + E_k)\Big(\frac{\delta_k}{E_k} - \frac{\delta_k^2}{2E_k^2} + \ldots\Big) \\ &\approx \sum_{k=1}^{K} \frac{\delta_k^2}{E_k} = X^2. \end{aligned}$$

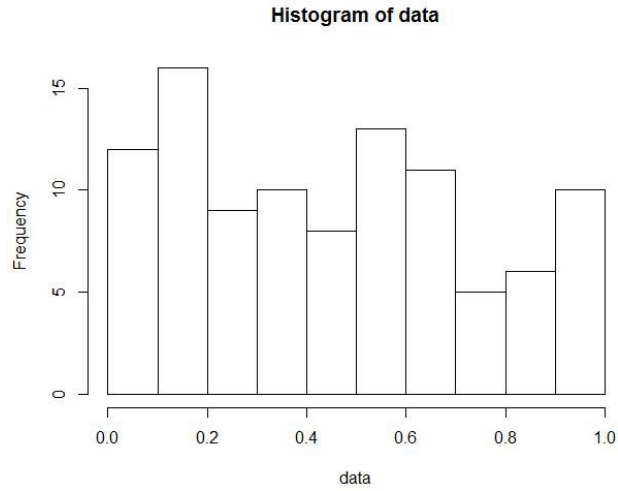
2.2 The χ^2 goodness of fit test for continuous data

The above χ^2 goodness of fit test considered multinomial distributions but the test can be applied to continuous distributions. This is done by dividing the real line into K intervals such that a trial results in outcome k if the observation falls in the kth interval. The intervals can be arbitrary, but it is sensible to choose the intervals such that the expected number of observations in each interval is at least 5.

Example 4: Testing uniformity We observe independent variables $X_1, \ldots, X_{100}$ taking values in $[0, 1]$ and wish to test whether they are uniformly distributed. We conduct a goodness of fit test as follows. We divide the interval $[0, 1]$ into equal spaced "bins" and count the number of observations that falls into each bin. We use 10 bins of width 0.1. This summary data are:

Interval k	1	2	3	4	5	6	7	8	9	10
O_k	12	16	9	10	8	13	11	5	6	10

(Interval $k = [(k-1)/10, k/10)$). The data can be visualised via a histogram:



If the X_is are uniformly distributed, they will fall into each (equal width) bin with equal probability. Thus we specify the hypotheses:

$$H_0 : p_k = 0.1 \text{ for } k = 1, \ldots, 10$$

against

$$H_1 : p_k \quad \text{unspecified.}$$

Answer: We have that $E_i = 100 \times 0.1 = 10$ for all $i = 1, \ldots, 10$. The test statistic is given by,

$$X^2 = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{10} \frac{(O_i - 10)^2}{10} = 9.6.$$

If H_0 is true $X^2 \sim \chi^2_{10-1} = \chi^2_9$. The corresponding p-value is given by $\mathbb{P}(\chi^2_9 > 9.6) = 0.38$. Thus there is no evidence to reject H_0 at $\alpha = 0.05$ level.

Using R

The analysis can be conducted in R.

We only need the frequencies and conduct the hypothesis test using:

```
> freq <- c(12, 16,  9, 10,  8, 13, 11,  5,  6, 10)
> chisq.test(freq) #or chisq.test(freq, p=rep(0.1,10))

Chi-squared test for given probabilities
data:  freq
X-squared = 9.6, df = 9, p-value = 0.3838
```

The hypothesis of randomness of the data is not rejected at 5% level.

Note This is not the only option to test for uniformity – we'll see some other options later in the course.

2.3 Contingency tables

It is often the case that count data consist of counts of outcomes of events which can be classified in (at least) 2 ways. In this course, we'll focus on the binary case – this gives a 2-way contingency table. Each observation is classified according to two characteristics: it belongs to exactly one class r from $\{1, \ldots, R\}$ **and** exactly one class c from $\{1, \ldots, C\}$. We let O_{rc} denote the number of observations in cell (r, c). Further let $O_{r\cdot} = \sum_{c=1}^{C} O_{rc}$ denote the (marginal) number of observations in the rth row and $O_{\cdot c} = \sum_{r=1}^{R} O_{rc}$ the (marginal) number of observations the cth column.

The data can then be displayed in the form:

	1	2	...	C	row total
1	O_{11}	O_{12}	...	O_{1C}	$O_{1\cdot}$
2	O_{21}	O_{22}	...	O_{2C}	$O_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
R	O_{R1}	O_{R2}	...	O_{RC}	$O_{R\cdot}$
column total	$O_{\cdot 1}$	$O_{\cdot 2}$	...	$O_{\cdot C}$	n

In this case we let p_{rc} denote the probability that a trial results in row r and column c, $p_{r\cdot}$ denotes the marginal probability of the rth row and $p_{\cdot c}$ for the cth column. We have the constraints that $\sum_{r=1}^{R}\sum_{c=1}^{C} p_{rc} = 1$, $\sum_{r=1}^{R} p_{r\cdot} = 1$ and $\sum_{c=1}^{C} p_{\cdot c} = 1$. (Note that in the special case that $C = 1$, we have a 1-way contingency table with $K = R$ possible outcomes.)

It is useful to introduce a random pair (X, Y), where X and Y take values in $\{1, \ldots, R\}$ and $\{1 \ldots, C\}$ respectively. The data in a contingency table can arise in two ways:

(i) We have n independent and identically distributed pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, where n is fixed and known (i.e. a fixed total).

(ii) for $c = 1, \ldots, C$, we have $O_{\cdot c}$ (a fixed number) observations with $Y_i = c$, with associated X values $X_1, \ldots X_n$ (i.e. fixed marginal, or column totals).

In both cases, (as in the previous chapters) the pairs themselves are unobserved, and we only observe the counts $O_{rc} = \#\{(X_i, Y_i) = (r, c)\}$. We will introduce two tests, one for each case: in case (i) we will test for **independence** between the rows and the columns; in case (ii) we will test for **homogeneity** – is the distribution across the rows the same in each column?.

Although the two cases arise in different contexts: in case (i) n is fixed, whereas in case (ii) $O_{\cdot 1}, \ldots, O_{\cdot C}$ are fixed, (a) it is not possible to tell the context in which a table arose just from looking at the table; (b) it turns out that the method of analysis is the same in the two cases - this is seen by considering each test in turn.

2.3.1 Case (i): Testing for independence

In case (i) we wish to test

$$H_0 : X \text{ and } Y \text{ are independent} \quad \text{vs.} \quad H_1 : X \text{ and } Y \text{ are dependent.}$$

If X and Y are independent, then

$$p_{rc} = \mathbb{P}\{(X, Y) = (r, c)\} = \mathbb{P}(X = r)\mathbb{P}(Y = c) = p_{r\cdot}p_{\cdot c}$$

Alternatively, therefore, we could write:

$$H_0 : p_{rc} = p_{r\cdot}p_{\cdot c} \quad \text{for } r = 1, \ldots, R \text{ and } c = 1, \ldots, C, \text{ where } \sum_{r=1}^{R} p_{r\cdot} = \sum_{c=1}^{C} p_{\cdot c} = 1$$

against

$$H_1 : p_{rc} \qquad \text{unrestricted.}$$

To carry out the test in this case, we follow the same steps as in Case 2 in Section 2.1. We first need to calculate the MLEs for p_{rc} under the null hypothesis. We have that

$$\hat{p}_{rc} = \hat{p}_{r\cdot}\hat{p}_{\cdot c},$$

where

$$\hat{p}_{r\cdot} = \frac{O_{r\cdot}}{n} \quad \text{and} \quad \hat{p}_{\cdot c} = \frac{O_{\cdot c}}{n}$$

are the (marginal) proportions of observations in categories r and c, respectively. Therefore, the expected count in the cell (r, c) under the null hypothesis is

$$E_{rc} = n\hat{p}_{rc} = \frac{O_{r\cdot}O_{\cdot c}}{n}.$$

The test statistic in this case is:

$$X^2 = \sum_{r=1}^{R}\sum_{c=1}^{C} \frac{(O_{rc} - E_{rc})^2}{E_{rc}}.$$

Now, from (2.3), X^2 (approximately) has a χ^2 distribution under the null hypothesis with $K-q-1$ degrees of freedom, where K denotes the total number of categories and q the number of parameters to be estimated.

- K: There are R possible X values and C possible Y values, so in total there are $K = RC$ possible outcomes;

- q: It might seem that we have had to estimate $R + C$ parameters: $p_{1\cdot}, \ldots, p_{R\cdot}$ and $p_{\cdot 1}, \ldots, p_{\cdot C}$. However, since $\sum_{r=1}^{R} p_{r\cdot} = 1$ and $\sum_{c=1}^{C} p_{\cdot c} = 1$, we in fact have only $R - 1$ and $C - 1$ parameters, respectively. Thus the total is number of parameters is $q = (R - 1) + (C - 1) = R + C - 2$.

Putting this together, we have $K - q - 1 = RC - (R + C - 2) - 1 = (R - 1)(C - 1)$. Thus

$$X^2 \mathrel{\dot{\sim}} \chi^2_{(R-1)(C-1)} \qquad \text{under } H_0.$$

2.3.2 Case (ii): Test of homogeneity

In case (ii) we wish to test for homogeneity across columns,
H_0 : the distribution is the same in each column vs. H_1 the distribution is not the same in each column.
Let $p_{r|c} = \mathbb{P}(X = r|Y = c)$, then we can reformulate the hypothesis as

$$H_0 : p_{r|c} = p_{r\cdot} \qquad \text{for all } c = 1, \ldots, C \text{ and } r = 1, \ldots, R, \quad \sum_{r=1}^{R} p_{r\cdot} = 1$$

against

$$H_1 : p_{r|c} \quad \text{unrestricted.}$$

Recall that the column totals are fixed by design to be $O_{\cdot 1}, \ldots, O_{\cdot C}$. Then, writing $p_{rc} = p_{r|c}p_{\cdot c}$, we see that under the null hypothesis the MLE for $p_{r|c}$ is $\hat{p}_{r\cdot} = \frac{O_{r\cdot}}{n}$, and thus the expected

count E_{rc} in cell (r, c) is

$$E_{rc} = O_{\cdot c}\hat{p}_{r\cdot} = \frac{O_{r\cdot}O_{\cdot c}}{n}.$$

Perhaps surprisingly, even though the data we generated in a different way, and we are testing a different hypothesis, the expected values E_{rc} (under the null hypothesis) are the same as for the test of independence.

The test statistic is the same:

$$X^2 = \sum_{r=1}^{R}\sum_{c=1}^{C}\frac{(O_{rc} - E_{rc})^2}{E_{rc}}.$$

The degrees of freedom calculation is slightly different in this case – here the number of unknown parameters under H_1 and H_0 is $C(R-1)$ and $(R-1)$, respectively. Therefore the degrees of freedom of the χ^2 statistic X^2 is $C(R-1) - (R-1) = (R-1)(C-1)$. So that, as before, we have

$$X^2 \mathrel{\dot\sim} \chi^2_{(R-1)(C-1)} \qquad \text{under } H_0.$$

Yates' continuity correction For the special case where $C = 2$ and $R = 2$ we have a 2×2 table. In this case, Yates' continuity correction is often used. This involves subtracting $1/2$ from the absolute difference between the observed and expected values in the numerator of the test statistic:

$$\tilde{X}^2 = \sum_{r=1}^{R}\sum_{c=1}^{C}\frac{(|O_{rc} - E_{rc}| - 1/2)^2}{E_{rc}}.$$

Example 5: Test of independence Consider data from the US General Social Survey on party support and gender in 1991:

	Democrat	Independent	Republican	Total
Female	279	73	225	577
Male	165	47	191	403
Total	444	120	416	980

We wish to test whether party support is the same for each gender. In other words we wish to test:

H_0 : party support and gender are independent

against

H_1 : party support is dependent on gender.

Answer: Thus we wish to apply a test of independence. Under H_0 the expected cell counts are given by:

	Democrat	Independent	Republican
Female	261.41	70.7	244.93
Male	182.60	49.34	171.07

Then,

$$X^2 = \sum_{r=1}^{2}\sum_{c=1}^{3}\frac{(O_{rc} - E_{rc})^2}{E_{rc}} = 7.01.$$

Using R

We use R to conduct the test:

```
> party <- matrix(c(279,165,73,47,225,191), ncol = 3)
> party
     [,1] [,2] [,3]
[1,]  279   73  225
[2,]  165   47  191

> chisq.test(party)

Pearson's Chi-squared test
data:  party
X-squared = 7.0095, df = 2, p-value = 0.03005
```

The small p-value < 0.05 indicates that we reject the null hypothesis at the 5% level and conclude that there is evidence that political orientation differs between (American) men and women. Note that we can also extract the expected cell counts using the function `chisq.test` in R. For example:

```
> chisq.test(party)$expected
         [,1]     [,2]     [,3]
[1,] 261.4163 70.65306 244.9306
[2,] 182.5837 49.34694 171.0694
```

Example 6: Test of homogeneity 100 school leavers in the North of England and 50 school leavers in Scotland were classified according to whether or not they had found work 6 months after leaving school. The number of Scottish and English are fixed by the survey design.

	Scotland	N. England	Total
Unemployed	16	41	57
Employed	34	59	93
Total	50	100	150

We wish to apply a test of homogeneity, corresponding to whether there is a difference in the unemployment rates between Scottish and North English school leavers.

Formally we wish to test

$$H_0 : p_S = p_E \quad \text{vs} \quad H_1 : p_S \neq p_E,$$

where p_S and p_E are the probabilities of a school leaver being unemployed in Scotland and N. England, respectively.

Answer: Under H_0, let $p_S = p_E = p$. We can estimate p from the data as $\hat{p} = 57/150 = 0.38$. Using this, calculate the expected number in each cell of the table under H_0.

	Scotland	N. England
Unemployed	19	38
Employed	31	62

Then, using Yates' continuity correction, we have

$$\tilde{X}^2 = \sum_{r=1}^{2}\sum_{c=1}^{2} \frac{(|O_{rc} - E_{rc}| - 1/2)^2}{E_{rc}} = 0.79.$$

Using R
We calculate the test statistic using R. For a 2×2 table, Yates' correction is automatically applied in R:

```
> school <- matrix(c(16,34,41,59),ncol=2)
> school
     [,1] [,2]
[1,]   16   41
[2,]   34   59
> chisq.test(school)

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

data:  school
X-squared = 0.79584, df = 1, p-value = 0.3723
```

Thus we do not reject the null hypothesis and conclude that there is no evidence that unemployment rates of school leavers differ between the two sides of the border.

2.4 The Kolmogorov-Smirnov test

2.4.1 Q-Q plots

In this section, instead of observing counts of outcomes, we have access to n continuous observations, and wish to test whether they arose from a particular distribution. Before carrying out a formal hypothesis test, we may seek to informally assess the distribution using a quantile-quantile (Q-Q) plot. Q-Q plots are a useful visual representation of how well data fit a distribution. The main idea is to plot the sample quantiles against the true quantiles of the hypothesised distribution – if the hypothesis is correct we expect to see (approximately) a straight line on the diagonal.

In the case of a normal distribution the R function `qqnorm` produces a Q-Q plot, but R does not seem to have a general-purpose Q-Q plot function. Below is one that should work for the three distributions listed (and you can easily add others). Note that this produces a slightly different kind of plot to `qqnorm`, but it is still a useful graphical representation of the fit.

Using R

```
QQplot <- function(x,dbn,par, ...) {
  sx <- sort(x)
  nx <- length(sx)
  edf=order(sx)/nx # calculate EDF
  cdf=switch(dbn,
```

```
            pnorm = pnorm(sx,mean=par$mean,sd=par$sd),
            pbinom = pbinom(sx,size=par$size,prob=par$prob),
            punif = punif(sx,min=par$min,max=par$max)
  )
  # find max difference (should consider current and next CDF values)
  D=abs(cdf-edf)
  d=abs(cdf[-1]-edf[-nx])
  Dmax=which(D==max(D))
  dmax=which(d==max(d))
  if(D[Dmax]>=d[dmax]) imax=Dmax
  if(d[dmax]>D[Dmax]) imax=dmax
  # do the plot
  plot(cdf, edf, ylab="EDF",xlab="CDF",xlim=c(0,1),ylim=c(0,1),...)
  points(cdf[imax],edf[imax],col="red",lwd=2)
  abline(0,1)
  # return EDF and CDF
  list(edf=edf,cdf=cdf)
}
```

As with `ks.test`, you pass it the data and the name of the R CDF function to use (as a character), but you must also pass it the required parameters of the R CDF function in a list, with elements named according to the names of the function arguments. For example, below is code to produce a Q-Q plot for the above data, and Figure 2.4.1 shows the plot it produces.

```
> y <- c(0.70, 0.29, 0.88, 0.22, 0.74)
> QQplot(y,"punif",par=list(min=0,max=1))
```

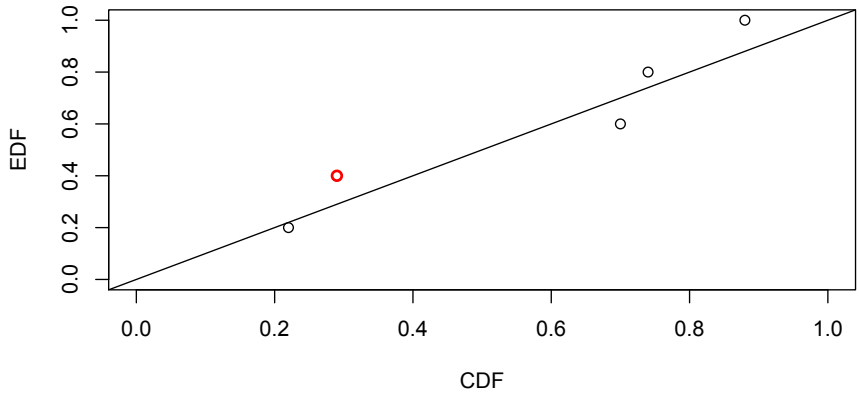


Figure 2.1: Q-Q plot for a sample of purportedly uniformly distributed data. The red dot is the point associated with the K-S test. "EDF" is for "empirical distribution function" and "CDF" for the hypothesised cumulative distribution function.

2.4.2 The Kolmogorov-Smirnov (KS) one-sample test

A Q-Q plot provides a heuristic tool for us to assess to the goodness of fit of a distribution – in this section we wish to formally test for a particular distribution function. More precisely, suppose we have independent and identically distributed random variables $X_1, \ldots, X_n \sim F$, where F is an unknown cumulative distribution function (c.d.f.) on the real line. Given a known cumulative distribution function F_0 of particular interest, we wish to test the hypothesis

$$H_0 : F = F_0 \quad versus \quad H_1 : F \neq F_0.$$

The Kolomogorov–Smirnov (KS) test procedure uses the empirical distribution function: Given data $x_1, \ldots, x_n$, define the *empirical* c.d.f. at $x \in \mathbb{R}$, by

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}} = \frac{1}{n} \#\{x_i \leq x\}.$$

In other words, the empirical c.d.f. $\hat{F}_n(x)$ gives the proportion of observations in the sample which are less than or equal to x. It is an empirical estimate of the (population) cumulative distribution function $F(x)$.

The main idea behind the KS test is that, if H_0 is true, then the empirical c.d.f. at x should be a close approximation to $F_0(x)$. We need a measure of the distance between the two functions: the KS test takes as its test statistic the maximum absolute difference between the two functions, given by

$$D = \max_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_0(x) \right|. \tag{2.4}$$

Since $\hat{F}_n$ and F_0 are both increasing functions, and that $\hat{F}_n$ is piecewise constant, the maximum deviation between the two functions must occur at or 'immediately before' one of the observations x_i. More precisely, let $x_{(1)} \leq \cdots \leq x_{(n)}$ be a reordering of the observations in increasing order, then

$$D = \max_{i=1,\ldots,n} \left\{ \left| \hat{F}_n(x_{(i)}) - F_0(x_{(i)}) \right|, \left| \hat{F}_n(x_{(i-1)}) - F_0(x_{(i)}) \right| \right\}.$$

This is most easily seen from an example.

Example 7 In a sample of size $n = 5$, the observations $x_1, \cdots, x_5$ equal 0.70, 0.29, 0.88, 0.22 and 0.74 respectively. So we have:

i	1	2	3	4	5
Order statistic $x_{(i)}$	0.22	0.29	0.70	0.74	0.88
e.d.f. $\hat{F}_5(x_{(i)})$	0.20	0.40	0.60	0.80	1.00

Suppose that we wish to test whether the data are uniformly distributed,

$$H_0 : F(x) = x, \quad \text{for } x \in (0, 1) \quad versus \quad H_1 : F(x) \neq x.$$

Answer: The distribution function of a $U(0, 1)$ random variable is

$$F_0(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 < x \leq 1 \\ 1, & x > 1 \end{cases}$$

The two functions we need to compare are shown in Figure 2.2. The maximum deviation can be calculated using a table:

i	0	1	2	3	4	5
Order statistic $x_{(i)}$	N/A	0.22	0.29	0.70	0.74	0.88
$F_0(x_{(i)})$	N/A	0.22	0.29	0.70	0.74	0.88
e.d.f. $\hat{F}_5(x_{(i)})$	0	0.2	0.4	0.6	0.8	1.0
Max. deviation near $x_{(i)}$	N/A	0.22	0.11	0.30	0.14	0.12

We have that $D = 0.30$. If H_0 is true, the value of D should be small, so the **critical region** lies in the upper tail. i.e. reject H_0 when D is large.

The exact distribution of D under H_0 is known, and is the same for all continuous cumulative distribution functions $F_0(x)$ – this is due to the fact that, if $X \sim F_0$ for a continuous c.d.f. F_0, then $F_0(X) \sim U(0,1)$. In this case, we say that the test is **distribution-free**.

The full derivation of the distribution of D under H_0 is complicated – in practice we only require the relevant critical value, which can be obtained from tables or using R for small sample sizes.

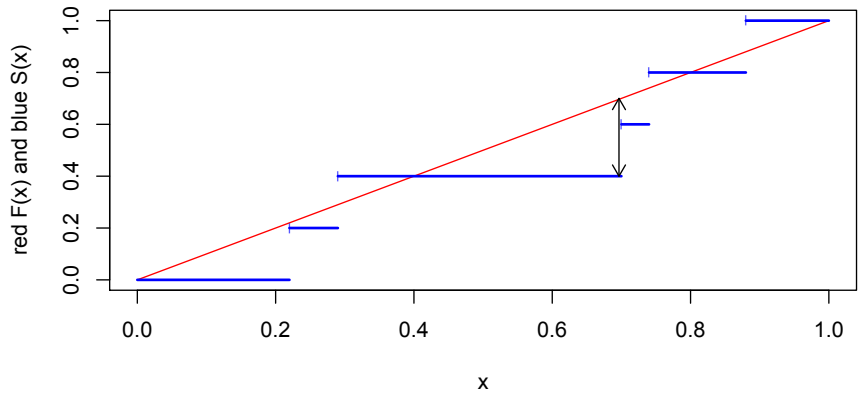


Figure 2.2: $F_0(x)$ (red) and $\hat{F}_n(x)$ (blue) for the sample of 5 observations, under H_0: the observations are from a $U(0,1)$ distribution. Observations are at the small blue vertical bars. The arrow shows the maximum difference between $F(x)$ and $S_n(x)$.

Using R

The Kolmogorov-Smirnov test is available under the name `ks.test` (within the `stats` package which does not need to be explicitly called).

For our example, the test of whether the sample of size 5 could be treated as a random sample from a uniform distribution can be carried out as follows. (The "`two.sided`" is a little misleading here, but you can see from the `ks.test` help documentation that it refers to whether the true distribution function is either greater than or less than the hypothesised one, not whether we are considering in the left and right tails of the test statistic distribution).

```
> x <- c(0.70,0.29,0.88,0.22,0.74)
```

```
> ks.test(x,"punif",0,1,alternative="two.sided")

One-sample Kolmogorov-Smirnov test
data:  x
D = 0.3, p-value = 0.664
alternative hypothesis: two-sided
```

The first argument of `ks.test` indicates the vector of data values which is to be tested. The second argument gives the name of the postulated distribution under the null hypothesis, whilst the third and fourth give its parameters. If there are no ties and the sample size is less than 100, an exact $p-$value is given. Otherwise the following asymptotic distribution is used.

Procedure for large samples For large n, the critical values can be obtained from the asymptotic distribution of D. It can be shown that the critical value is $1.358/\sqrt{n}$ for a 5% test and $1.628/\sqrt{n}$ for a 1% test.

2.4.3 Comparison of the Kolmogorov-Smirnov with the χ^2 goodness-of-fit test

1. Applicability.

   (a) The Kolmogorov-Smirnov (KS) test cannot be used for nominal data.

   (b) However, although the theory of Kolmogorov-Smirnov assumes that $F_0(x)$ is continuous (and I do not present it here), the test can also be used on discrete data. Use of the standard tables results in a conservative test (i.e. it sometimes accepts H_0 when it should reject it), but exact calculations are also feasible (see W.J. Conover: Practical Nonparametric Statistics, 3rd edn.).

2. Sample size. If you follow the rule of thumb that cell expectations should not be less than 5, the χ^2 test can only be used if sample size is at least 10, whereas KS can be used even for $n = 1$, although its power would be tiny!

3. Confidence band. Another advantage of the Kolmogorov-Smirnov test is that a confidence band for the distribution function can be obtained, but we will not pursue that here.

2.4.4 The Kolmogorov-Smirnov two-sample test

Finally in this chapter we consider the case that we have two independent samples and wish to test whether the data arose from the same distribution in each case. Suppose we observe $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_X$ and $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} F_Y$, where F_X and F_Y are the true (unknown) distribution functions of the X and Y samples, respectively. We assume that F_X and F_Y are both continuous. Note that this means that ties should only arise due to imprecise measurements. [2]

[2] In fact, as in the case of the Kolmogorov-Smirnov one-sample test, this test can be used with discrete data, but use of the standard tables will then give a conservative result.

Formally, we wish to test the following:

Null hypothesis $H_0 : F_X = F_Y$ (i.e. the X and Y observations are from the same distribution).

Alternative hypotheses: we consider three possible alternatives:

(A) $H_1 : F_X(t) \geq F_Y(t)$ for all t, with strict inequality holding over some interval.

(B) $H_1 : F_X(t) \leq F_Y(t)$ for all t, with strict inequality holding over some interval.

(C) $H_1 : F_X(t) \neq F_Y(t)$ over some interval.

Here (A) and (B) give one-sided tests, while (C) is two-sided.

The tests use the empirical distribution function (e.d.f) of each sample: if H_0 is true, the e.d.f. $\hat{F}_X(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq t\}}$ of the X sample should closely approximate the e.d.f. $\hat{F}_Y(t) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{y_i \leq t\}}$ of the Y sample. The **test statistics** here are

Test (A): $D^+ = \max_t \left[\hat{F}_X(t) - \hat{F}_Y(t)\right]$

Test (B): $D^- = \max_t \left[\hat{F}_Y(t) - \hat{F}_X(t)\right]$

Test (C): $D = \max_t \left|\hat{F}_X(t) - \hat{F}_Y(t)\right| = \max[D^+, D^-]$

The evaluation of the statistics is easier than for the corresponding one-sample test as the maximum deviation between the two functions must occur at an observed value in either the X or the Y samples. There is no difficulty in evaluating the statistics if there are repeated readings within a sample or if x and y observations are tied: the e.d.f. of the X sample simply takes a step of the appropriate size at each x observation, and the same is true of the e.d.f. of the Y sample.

Example 8 A motorist, who stays in St Andrews, makes frequent journeys to visit relatives in Perth. On five occasions he travels via Newburgh, and his outward journey times $x_1, x_2, \cdots, x_5$ are 51, 55, 58, 50 and 53 minutes. On another five occasions he tries going via Dundee. His outward journey times $y_1, y_2, \cdots, y_5$ are 57, 60, 54, 63 and 56 minutes. He wishes to carry out a test to see if the journeys have the same distribution of time – we will consider the alternative hypothesis given by option (C).

To evaluate the test statistic, we first pool the samples and list the distinct values in increasing order:

Answer:

i	1	2	3	4	5	6	7	8	9	10
Order statistics	50	51	53	54	55	56	57	58	60	63
$\hat{F}_X(t)$	0.2	0.4	0.6	0.6	0.8	0.8	0.8	1.0	1.0	1.0
$\hat{F}_Y(t)$	0.0	0.0	0.0	0.2	0.2	0.4	0.6	0.6	0.8	1.0
$\lvert\hat{F}_X(t) - \hat{F}_Y(t)\rvert$	0.2	0.4	0.6	0.4	0.6	0.4	0.2	0.4	0.2	0.0

Hence the test statistic takes the value $D = D^+ = 0.6$.

Note In cases where there are ties, enter only the distinct values in increasing order in the first row of the table. i.e. avoid repetitions and ensure $\hat{F}_X(t)$ and $\hat{F}_Y(t)$ are uniquely defined for each value of t.

For each of the tests, it is clear that, if H_0 is true, the test statistic is likely to be small, whilst if the alternative hypothesis is true, the test statistic will tend to be larger. Thus the critical region lies in the upper tail. i.e. reject H_0 when the test statistic is large. Critical values can be found in tables (e.g. see W.J. Conover: Practical Nonparametric Statistics, 3rd edn.), but we will use R.

The Kolmogorov-Smirnov two-sample test can also be carried out using the function `ks.test`. The default for a two-sided test with no ties is to calculate exact $p-$values, provided the product of the sample sizes is less than 10000. In other cases, the asymptotic distribution is used.

Using R
The above test can be carried out within R as follows:

```
> x <- c(51,55,58,50,53)
> y <- c(57,60,54,63,56)
> ks.test(x,y,alternative="tw")
```

The alternative "tw" is short for "two.sided". The other options are "less" or "greater". The following output is generated:

```
Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.6, p-value = 0.3571
alternative hypothesis: two-sided
```

2.4.5 Non-examinable: Other goodness-of-fit tests

The Cramer von Mises (CvM) test and the Andreson-Darling (AD) test are similar to the Kolmogarov-Smirnov one-sample test, and can be used wherever the KS test can be. Unlike the KS test, these use the total discrepancy between the empirical distribution function (EDF), $\hat{F}_n(x)$, and hypothesised cumulative distribution function (CDF), $F_0(x)$, not just the maximum discrepancy. For the CvM and AD tests, the statistics are

$$\begin{aligned} C_n &= \int_{-\infty}^{\infty} (S_n(x) - F_0(x))^2 dx \quad \text{and} \\ A_n &= \int_{-\infty}^{\infty} \frac{(S_n(x) - F_0(x))^2}{F_0(x)[1 - F_0(x)]} dx. \end{aligned} \tag{2.5}$$

respectively.

Using R
Both tests are implemented in the R library `goftest`, in functions `cvm.test` and `ad.test`. Using it with the example data above:

```
> library(goftest)
> y <- c(0.70, 0.29, 0.88, 0.22, 0.74)
> cvm.test(y,"punif",0,1)

Cramer-von Mises test of goodness-of-fit
Null hypothesis: uniform distribution
data:  y
omega2 = 0.073167, p-value = 0.7545

> ad.test(y,"punif",0,1)

Anderson-Darling test of goodness-of-fit
Null hypothesis: uniform distribution
data:  y
An = 0.41686, p-value = 0.8254
```

Chapter 3

Permutation and randomisation tests

3.1 Introduction

In this chapter, we introduce *nonparametric* approaches to hypothesis testing, which utilise *permutations* of the data. The idea is to generate all possible equally-likely permutations (reorderings) of the data, and then examine whether the actual permutation that we observed is likely or unlikely under the null hypothesis when compared with the other possible permutations that might have arisen. By permuting the observed data, we emulate the distribution of the test statistic under the null hypothesis, which enables us to calculate a p-value without relying on asymptotic approximations for the distribution of the test statistic.

In contrast to the previous section, where we were typically testing for a fixed distribution or family of distributions indexed by a finite-dimensional parameter, we will focus on particular aspects of the distribution (e.g. the mean or median). Furthermore, some of the new methods introduced in this chapter will be used to test for particular values of the mean without the assumption of an underlying parametric (eg. Normal) model; the tests introduced thus provide alternatives to t-tests and Z-tests in settings where we are unhappy making a normality assumption. We may wish to avoid a parametric assumption since such a model may be too restrictive. We will introduce the following permutation tests: the sign test, Wilcoxon's signed ranks test, the Mann-Whitney test, and randomisation tests (in which we generate a random subset of all possible permutations, as opposed to all possible permutations).

Permutation and randomisation tests are nonparametric in that we do not assume an underlying distribution for the process that generated the observed data. There are some advantages in adopting such an approach:

- Some of the tests are valid for continuous measurements which do not follow the Normal distribution (or other parametric distributions);
- The methods are robust to outliers: even if the vast majority of the observations are in accordance with the Normal distribution, outliers may influence statistical inference when using classical t-or Z-tests;
- Some of the tests are valid (i.e have the correct size) even for small samples, since we do not rely on asymptotic approximations
- Some of the tests are are valid when the underlying distribution is not continuous – in some cases we only need ordinal data.

On the other hand, if we are satisfied that a normality assumption holds, then a classical test (e.g. a t-test) is likely to be a better (more powerful) option.

3.2 Two-Sample Permutation Tests

In this section we have two samples, an X sample and a Y sample, and we wish to test:

$$H_0\text{: } X \text{ and } Y \text{ have the same distribution}$$
$$\text{vs.}$$
$$H_1\text{: } X \text{ and } Y \text{ do not have the same distribution}$$

We first consider a simpler test to demonstrate the main idea.

3.2.1 Permutation test using the sample means to construct the test statistic

The concept of a permutation test is perhaps best illustrated by an example – we'll use the sample means to construct the test statistic. The test here does not add much beyond what we can already do with the KS test (though ties are easier to deal with here), but will act as a straightforward introduction to permutation based techniques.

Example 9: Suppose we have two samples, one from an X distribution and one from a Y distribution,

x	8	6	3	9		
y	7	10	10	12	18	15

To simplify our aim, we initially consider testing whether the two distributions have different underlying means. Let μ_X and μ_Y denote the means for X and Y populations, respectively. Then, we under the null hypothesis above we have

$$H_0 : \mu_X = \mu_Y$$

and we test against the simpler alternative $H_1 : \mu_X \leq \mu_Y$. Note that the hypotheses on the means are not equivalent to the initial hypothesis here – we may have two different distributions under the alternative in the initial case with the same means.

We can estimate μ_X and μ_Y by the corresponding sample means $\bar{x}$ and $\bar{y}$ and consider the test statistic: $d = \bar{y} - \bar{x}$. Under H_0, we expect d to be close to 0; whereas if H_0 is not true then d should be nonzero. Now the idea underlying the permutation test is if we *permute* the data randomly between the two groups (keeping each group size fixed), then, under H_0, the observed dataset should be a typical member of these permuted datasets. And, since the permutation merges the samples, for any one of the permuted datasets, the two-samples will have the same distribution. We then calculate a p-value by evaluating how many of the permutations result in a value of d that is at least as extreme as the observed difference from the original data.

Returning to our example, we begin by calculating the test statistic for the observed data, which gives an observed test statistic of $d = 5.5$. We now need to consider all permutations of the data which give a value for d of 5.5 or more. This is equivalent to finding all samples of size 4 that give a sample total that is equal to or smaller than the total of the xs in the real sample. That sample comprised the values 3, 6, 8, 9 with a total of 26.

Only two samples will give a smaller total: 3, 6, 7, 8 and 3, 6, 7, 9. There are three that give the same total: the original sample, together with two samples comprising the values 3, 6, 7, 10. There are two 10s in the combined sample, and so there are two 10s to choose from. Thus, there are five possible combinations with a test statistic at least as large as the observed test statistic.

x				Total x	y						Total y	$\bar{x}$	$\bar{y}$	$d = \bar{y} - \bar{x}$
3	6	7	8	24	9	10	10	12	15	18	74	6	$12\frac{1}{3}$	$6\frac{1}{3}$
3	6	7	9	25	8	10	10	12	15	18	73	6.25	$12\frac{1}{6}$	$5\frac{11}{12}$
3	6	7	10	26	8	9	10	12	15	18	72	6.5	12	5.5
3	6	7	10	26	8	9	10	12	15	18	72	6.5	12	5.5
3	6	8	9	26	7	10	10	12	15	18	72	6.5	12	5.5

In total, there are $\binom{10}{4} = 210$ possible permutations, all of which are equally likely under the null hypothesis. Hence the one-tailed probability of obtaining a test statistic at least as large as the observed one is $\frac{5}{210}$; this is the p-value! Thus we reject H_0 at the 5% significance level.

We can also consider a two-tailed alternative hypothesis:

$$H_0 : \mu_X = \mu_Y \qquad \text{vs} \qquad H_1 : \mu_X \neq \mu_Y.$$

In this case, we also need to worry about the left hand tail of the distribution. Let p denote the one-sided p-value, then $2\min\{p, 1-p\}$ gives a valid p-value for the two-sided test.

Example 10 Consider two samples of observations

$$(x_1, \ldots, x_4) = (13, 14, 10, 13), \quad (y_1, \ldots, y_6) = (19, 17, 18, 13, 20, 15).$$

A permutation test was conducted on the null hypothesis that both samples come from the same distribution, against the alternative that the sample y comes from a distribution with a larger mean than the distribution yielding sample x. Using a permutation test, which of the following p-values is correct?

(a) 8/210

(b) 8/252

(c) 4/210

(d) 4/252

Answer: We have in total 10 observations, 4 of which make up the x sample and 6 the y sample, hence there are $\binom{10}{4} = 210$ possible permutations.

We could calculate the difference in sample means as our test statistic, alternatively we could simply calculate the total of the x sample. We observe $\sum x = 50$.

The permutations which will result in an x total less than or equal to the one observed are $\{10, 13, 13, 13\}$ and $\{10, 13, 13, 14\}$, for the latter there are three possible ways to achieve this permutation (there are three 13s in the observed samples and we want to choose 2). Therefore, there are 4 permutations that are at least as extreme as the observed permutation and there are 210 possible permutations.

The test is a one-sided test and so the correct answer is (c).

Example 11 A physiotherapy trial measures the increase in muscle strength of patients undergoing a particular standard form of therapy. A new form of therapy is developed which is thought to be more efficient than the standard therapy. A physiotherapist conducts a trial to assess the performance of the two therapies by measuring the increase in muscle strength over a six month period produced by patients receiving either therapy. The data consist of an index of relative increase in muscle strength for each patient classified by therapy.

Standard	15	12	18	13	14	10
New	16	15	19	20		

The physiotherapist is interested in testing whether the new therapy produces a greater increase in muscle strength than the standard therapy.

(a) State the hypotheses that the physiotherapist should test.

(b) Conduct a permutation test at the 5% significance level to test these hypotheses.

(c) What are your conclusions? Given this analysis what advice would you give to the physiotherapist?

Answer: The hypotheses we should test are, $H_0 : \mu_X = \mu_Y$, vs, $H_1 : \mu_X > \mu_Y$. (Note you could justify using a two-sided alternative if you wanted to take a very conservative approach.) We could work with the difference in sample means, however, it is equivalent to consider the total of the new treatment results. The permutations that are at least as extreme as the one observed (which has a total of 70) are

x				frequency	total
16	18	19	20	1	73
15	18	19	20	2	72
14	18	19	20	1	71
13	18	19	20	1	70
15	16	19	20	2	70

In total there are 210 possible permutations of the data. Hence the p-value of the test is $\frac{7}{210} = 0.033$. Since this p-value is smaller than 0.05 (because we are conducting the test at 5% significance), we reject the null hypothesis. Note, that had we been using the two-sided alternative hypothesis our p-value would have been $0.066 = 2\min\{0.033, 1 - 0.033\}$, and so we would have failed to reject the null hypothesis at the 5% level.

Given the results of our permutation test we would advise the physiotherapist that there is moderate support that the new treatment does perform better than the standard treatment. However, we note that the evidence against the null hypothesis of no difference is not particularly strong (does 5% significance seem appropriate for a medical trial) and would strongly suggest the trial is repeated for a much larger sample.

3.2.2 The Mann–Whitney U-test

We now turn to a more sophisticated two-sample test called the Mann-Whitney U-test (AKA, among other names, the Wilcoxon rank sum test). Recall that we have two samples, an X sample $(x_1, \ldots, x_n)$ and a Y sample $(y_1, \ldots, y_m)$, and we wish to test:

H_0: X and Y have the same distribution

vs.

H_1: X and Y do not have the same distribution

The Mann–Whitney U-test will have power against more alternatives than the simple test based on the mean in the previous section.

We will assume the the X and Y samples are independent from each other and that the data is at least ordinal (we will need to rank the observations when calculating the test statistic). The idea is calculate the number of times an X sample ranks above a Y sample – if the null hypothesis is true, this should happen around "half" the time. (I put half in quotes here since I haven't been too explicit on how the calculation is carried out, we'll formalise this below).

The test statistic in this case is

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(x_i, y_j),$$

where

$$S(x_i, y_j) = \begin{cases} 1 & x_i > y_j \\ 1/2 & x_i = y_j \\ 0 & x_i < y_j. \end{cases}$$

One way to calculate U is to pool the two samples together and list the observations in increasing order. Then replace each observation by an x or a y, according to which sample it came from. For each x_i, evaluate v_i the number of y observations which are smaller than x_i (adding a half for each tie). Then $U = \sum_{i=1}^{n} v_i$. The test is a permutation test, where the permutations are all possible orderings of n of xs and m of ys.

If there are no ties, then the distribution of U under H_0 is known and we can calculate an exact p-value based on the permutations (there are also tables available, but we will use `R`) . In examples where there are ties, `R` will provide an approximate p-value based on a normal approximation.

Example 12 Consider the same data used to illustrate the two-sample permutation test based on sample means,

x	8	6	3	9		
y	7	10	10	12	18	15

We wish to test

H_0: X and Y have the same distribution

vs.

H_1: X and Y do not have the same distribution

at the 5% significance level.

Answer: If we place the observations in increasing order, we have

3 6 7 8 9 10 10 12 15 18

Replacing observations by an x if they are from the first sample, or a y otherwise, gives

x x y x x y y y y y

We have four xs. No ys are less than two of these, and a single y is less than the other two. Hence we have $U = 0 + 0 + 1 + 1 = 2$. We can list every possible ordering of 4 xs and 6 ys, and determine the proportion of these that give a test statistic equal to 2 or less. The following permutations meet this condition

										Test statistic
x	x	x	x	y	y	y	y	y	y	0
x	x	x	**y**	x	y	y	y	y	y	1
x	x	**y**	x	x	y	y	y	y	y	2
x	x	x	**y**	**y**	x	y	y	y	y	2

The total number of permutations is $\binom{10}{4} = 210$, giving us a $p-$value of 4/210 for a one-tailed test of whether the x distribution is *stochastically* less than that of the y distribution. We have a two-tailed alternative hypothesis, so our $p-$value is $8/210 = 0.038$ or 3.8% (similarly to the previous section we are really calculating $2 * min(p, 1 - p)$, where p is the one-sided p-value here). This compares with 4.76% that we obtained from our earlier two-sample permutation test.

In practice, we do not need to enumerate the permutations that are at least as extreme as the observed one, as we can look up the Mann-Whitney U-statistic in published tables, or use R (see below).

Notes

- The value that U takes will lie between 0 (when the smallest n observations all come from the X sample) and $n \times m$ (when the smallest m observations all come from the Y sample).
- The distribution is symmetric about $\frac{nm}{2}$ if the null hypothesis holds. Hence $\mathbb{E}(U) = \frac{nm}{2}$ if H_0 is true.
- It can be shown that the variance under H_0 is $\frac{nm(n+m+1)}{12}$.
- For large n and m, we can use these results together with the Central Limit Theorem to obtain a test based on a normal approximation. A continuity correction improves this approximation.
- If the two samples are normally distributed with the same variance, then this permutation test is slightly less powerful than the two-sample $t-$test. In general, it can also be expected to be less powerful than the two-sample permutation test based on the sample means. It loses power by throwing away the actual values of the observations, but the power loss is modest. The advantage is that outliers do not affect statistical inferences disproportionately.

Using R
The test can be carried out using the `wilcox.test` function. The function uses the

statistic U but denotes it by W. Hypotheses are expressed for the case that the two distributions differ only in location (and the default is to test for the difference to be zero under the null hypothesis). The default, if there are no ties and the sample sizes are less than 50, is to calculate exact p-values. Otherwise, a normal approximation is used. For our example, we have

```
> x <- c(8,6,3,9)
> y <- c(7,10,10,12,18,15)
> wilcox.test(x, y, paired=FALSE, alternative="two.sided")
```

The `alternative` argument of the function can also be specified as `"greater"` or `"less"`. The output from the function is as follows,

```
        Wilcoxon rank sum test with continuity correction
data:  x and y
W = 2, p-value = 0.0422
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x,y, paired = FALSE, alternative="two.sided"):
  cannot compute exact p-value with ties
```

Why does this give a different result than we obtained previously? The clue is in the warning. R has detected a tie, and has therefore switched to a normal approximation. But this doesn't make sense, as the tie is within one of the samples, and therefore does not affect the test! To verify, we can replace one of the 10s by 11, which should not alter our result (it would still just be labelled as coming from sample y):

```
> x <- c(8,6,3,9)
> y <- c(7,10,11,12,18,15)
> wilcox.test(x, y, paired=FALSE, alternative="two.sided")

        Wilcoxon rank sum test
data:  x and y
W = 2, p-value = 0.03810
alternative hypothesis: true location shift is not equal to 0
```

We now get the same result as previously when we performed the test by hand (the programmers of R are clearly not infallible!).

3.3 Permutation Tests for Matched Pairs

Suppose we have n pairs of data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. For example, these might be before and after readings on a set of n individuals, or two different hand creams might be tested by applying one to the left hand and the other to the right hand of an individual.

Typically, analysis of matched pairs is conducted by first calculating the differences,

$$d_i = x_i - y_i \qquad \text{for} \qquad i = 1, \ldots, n.$$

If these differences can reasonably be assumed to be normally distributed with constant but unknown variance, then we can use a paired t-test of $H_0 : \mu_X = \mu_Y$ against a one- or two-sided alternative. In this section, consider options for paired data when the differences are not normally distributed.

In contrast to the previous section, we cannot generate permutations of the data by reallocating the pooled set of observations to the two samples – this would fail to preserve the pairs. Here we will introduce tests that do not assume that the two samples are independent. To do this, we work with the differences $d_1, \ldots, d_n$ (which we do assume to be independent). The idea is that if the (marginal) X and Y distributions are the same, then the d_i should be symmetric about 0 (and in particular should be positive with probability 1/2.)

Example 13 We have 400m race times for 8 athletes competing at both sea level and altitude:

Runner	1	2	3	4	5	6	7	8
Time at sea level	48.3	47.6	49.2	50.3	48.8	51.1	49.0	48.1
Time at altitude	50.4	47.3	50.8	52.3	47.7	54.5	48.9	49.9
Difference	-2.1	0.3	-1.6	-2.0	1.1	-3.4	0.1	-1.8

The "permutations" in this case are obtained by considering all possible combinations of $(\pm d_1, \pm d_2, \ldots, \pm d_n)$. Eg:

```
  -0.1 -0.3 -1.1 -1.6 -1.8 -2.0 -2.1 -3.4
   0.1 -0.3 -1.1 -1.6 -1.8 -2.0 -2.1 -3.4
  -0.1  0.3 -1.1 -1.6 -1.8 -2.0 -2.1 -3.4
   0.1  0.3 -1.1 -1.6 -1.8 -2.0 -2.1 -3.4
...
   0.1  0.3  1.1 -1.6 -1.8 -2.0 -2.1 -3.4
...
   0.1  0.3  1.1  1.6  1.8  2.0  2.1  3.4
```

Similarly to the unmatched two-sample tests above, different choices for the test statistic lead to slightly different tests. We consider the some standard options below.

3.3.1 Using the sample difference mean as the test statistic

Assume that the measurements are continuous observations (this is needed so that calculating the differences makes sense). We like to test whether X and Y have the same (marginal) distributions. As in Section 3.2.1, we simplify our initial goal and test the following hypotheses instead

$$H_0 : \mu_X = \mu_Y$$

versus

$$H_1 : \mu_X \neq \mu_Y.$$

(we consider a two-tailed alternative here, but one tailed alternatives are possible too.) Note that if X and Y do have the same (marginal) distributions, then $\mu_X = \mu_Y$. Let $d_i = x_i - y_i$ and let $z_i = -d_i = y_i - x_i$ for $i = 1, \ldots, n$. If H_0 is true, then the distribution of d_i will have mean zero, and so will the distribution of z_i. Now we generate "permutations" by taking either d_1 or z_1 with either d_2 or z_2, and so on, giving 2^n possible combinations in all.

We take the test statistic to be the mean of the differences $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$, and evaluate the mean difference for every possible permutation. If H_0 is true, then the observed value should be a "typical" one of these, and the p-value is the number of permutations with a test statistic at least as extreme as that obtained from the original data – see the examples for explicit calculations.

Example 13 (cont.) Athletes' 400m times:

Runner	1	2	3	4	5	6	7	8
Time at sea level (x)	48.3	47.6	49.2	50.3	48.8	51.1	49.0	48.1
Time at altitude (y)	50.4	47.3	50.8	52.3	47.7	54.5	48.9	49.9
Difference	-2.1	0.3	-1.6	-2.0	1.1	-3.4	0.1	-1.8

We wish to test $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X < \mu_Y$, where μ_X and μ_Y is the population mean Time at sea level and Time at altitude, respectively.

Answer: We use the mean difference as our test statistic. We list all permutations, and evaluate our test statistic for each. We then count up the number of permutations with a mean less than or equal to the observed sample mean, $\bar{d} = -1.175$.

Permutations (sign flips)								Mean difference ($\bar{d}$)
-0.1	-0.3	-1.1	-1.6	-1.8	-2.0	-2.1	-3.4	-1.55
0.1	-0.3	-1.1	-1.6	-1.8	-2.0	-2.1	-3.4	-1.525
-0.1	0.3	-1.1	-1.6	-1.8	-2.0	-2.1	-3.4	-1.475
-0.1	-0.3	1.1	-1.6	-1.8	-2.0	-2.1	-3.4	-1.275
0.1	0.3	-1.1	-1.6	-1.8	-2.0	-2.1	-3.4	-1.45
0.1	-0.3	1.1	-1.6	-1.8	-2.0	-2.1	-3.4	-1.25
-0.1	0.3	1.1	-1.6	-1.8	-2.0	-2.1	-3.4	-1.2
0.1	0.3	1.1	-1.6	-1.8	-2.0	-2.1	-3.4	-1.175

This number of as extreme permutations divided by $2^8 = 256$ (the total number of permutations possible) gives a p-value of 0.031. There is moderate evidence that the race times at sea level are faster. If the alternative hypothesis was two-sided, we would double this p-value.

3.3.2 Wilcoxon's signed rank test Or: Using the ranks of the observations to construct the test statistic

Assume that the measurements are continuous observations (as above this is needed so that calculating the differences makes sense. In some circumstances, we may wish to adopt a test

statistic that is largely unaffected by outliers – a common strategy (which we adopt here) for nonparametric tests is to replace the observations by their ranks. Hypothesis tests based on ranks exploit the relative size of the observations, but do not use the actual values. In order to reflect this in our hypothesis, we use

$$H_0 : \text{the distribution of } X - Y \text{ is symmetric about 0}$$

versus

$$H_1 : \text{the distribution of } X - Y \text{ is not symmetric about 0}$$

(we can also consider one sided alternatives – see example below.)

The test works as follows: we first rank the absolute values of the differences ($|d_i|$) from smallest to largest, ignoring their sign, then multiply the rank by the corresponding sign of the difference; any pairs for which the difference is zero are discarded. We can then conduct a permutation test, using T^+ the sum of the positive (or T^- the sum of the negative) ranks as a test statistic. We generate all possible permutations by flipping the signs of the ranks. In a setting with n pairs, T^+ takes values between 0 and $\sum_{i=1}^{n} i = n(n+1)/2$ (e.g., with $n = 8$, T^+ can take any value between 0 and 36). If H_0 is true, then the signs of the ranks should be evenly distributed (ie. they have probability 1/2 of being positive or negative), which means the T^+ should be close to $n(n+1)/4$; we investigate this in more detail below, but first let's consider the Athlete's example.

Example 13(b): Conduct a Wilcoxon's signed rank test on the Athlete's 400m times.

Answer:

Runner	1	2	3	4	5	6	7	8
Time at sea level	48.3	47.6	49.2	50.3	48.8	51.1	49.0	48.1
Time at altitude	50.4	47.3	50.8	52.3	47.7	54.5	48.9	49.9
Difference	-2.1	0.3	-1.6	-2.0	1.1	-3.4	0.1	-1.8
Absolute difference	2.1	0.3	1.6	2.0	1.1	3.4	0.1	1.8
Rank	7	2	4	6	3	8	1	5
Signed rank	-7	2	-4	-6	3	-8	1	-5

Here we were interested in the alternative that the athletes are *slower* at altitude – therefore we wish to have a one-sided hypothesis and test, for the alternative hypothesis we can specify this using the median. We have H_0 : the distribution of $d = X - Y$ is symmetric about 0; versus H_1 : the median of $d = X - Y$ is negative.

The test statistic here is $T^+ = 2 + 3 + 1 = 6$. Given our observed test statistic is small, and we have a small sample size we can easily find the flips which will result in a test statistic as extreme as the one we have observed. These are those with the following positive signed ranks:

$$\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{2,3\}, \{2,4\}, \{1,2,3\}.$$

Hence our p-value is $14/2^8 = 0.0547$.

Alternatively, we can use published tables, or we can use R:

Using R
The test can be carried out using the `wilcox.test` function.

```
> x <- c(48.3, 47.6, 49.2, 50.3, 48.8, 51.1, 49.0, 48.1)
> y <- c(50.4, 47.3, 50.8, 52.3, 47.7, 54.5, 48.9, 49.9)
> wilcox.test(x, y, paired=TRUE, alternative="less")

    Wilcoxon signed rank test
data:  x and y
V = 6, p-value = 0.05469
alternative hypothesis:  true location shift is less than 0
```

Hence for our one-sided test, the p-value is just above 0.05. We have weak evidence (we would reject at 10% significance, but not 5% significance) that the distribtuion of the difference is not symmetric about zero. If we were to conduct a two-sided test, we would double the p-value.

We can find the expectation and variance of T^+ under the null hypothesis (ignoring any zero differences) by writing

$$T^+ = \sum_{i=1}^{n} (\delta_i \times i)$$

where $\delta_i = 1$ if the ith rank is positive and $\delta_i = 0$ otherwise. Then, under H_0, we have $\mathbb{P}(\delta_i = 1) = \mathbb{P}(\delta_i = 0) = 0.5$, and it follows that $\mathbb{E}(\delta_i) = 0.5$,, $\mathbb{E}(\delta_i^2) = 0.5$ and $\text{Var}(\delta_i) = \mathbb{E}(\delta_i^2) - \mathbb{E}(\delta_i)^2 = 0.25$.. We conclude that, under H_0,

$$\mathbb{E}(T^+) = \sum_{i=1}^{n} i\mathbb{E}(\delta_i) = \frac{1}{2}\sum_{i=1}^{n} i = \frac{n(n+1)}{4}$$

$$\text{Var}(T^+) = \sum_{i=1}^{n} i^2 \text{Var}(\delta_i) = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{24}.$$

For large sample sizes, this allows us to use a Normal approximation. Under H_0, we have approximately

$$\frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0,1)$$

Here, the approximation is improved using a continuity correction, in this case we reduce the magnitude of the numerator by 0.5.

Critical values in published tables assume that there are no *ties*. That is, we assume that no two differences are identical in magnitude. The effect of ties is minor when the normal approximation is used. In R, if there are no ties, the exact $p-$value is calculated for $n < 50$, but the default is to use the Normal approximation whatever the sample size if ties are present.

This Wilcoxon Signed Rank test is based only on the ranks of the differences. It uses less information compared to using the actual observed values to calculate the sample mean. Therefore, it is less powerful than some other tests we have seen when interval or ratio data are available without any outliers.

3.3.3 The Sign test (Using the sign of the differences to construct the test statistic)

Our third paired test, the Sign test, is based only which of the two samples is larger for each observation. In other words, for each pair, we only use whether x_i is smaller or larger than y_i (any ties are discarded). Therefore, the measurement scale of the data may be ordinal, interval or ratio.

It is useful to introduce a sign variable S_i for each pair, given by

$$S_i = \begin{cases} 1 & x_i > y_i \\ 0 & x_i = y_i \\ -1 & x_i < y_i. \end{cases}$$

[Note that this notation is in slight contrast to that used for the Mann-Whitney U-Test above]

Now, if the distribution of X and Y are the same, then the distribution of S_i will be symmetric about zero – thus, we can reformulate our hypotheses as

$$H_0 : \mathbb{P}(S = -1) = \mathbb{P}(S = 1)$$

versus

$$H_1 : \mathbb{P}(S = -1) \neq \mathbb{P}(S = 1).$$

We can also consider one-sided alternatives here (ie. $\mathbb{P}(S = -1) < \mathbb{P}(S = 1)$., or $\mathbb{P}(S = -1) > \mathbb{P}(S = 1)$.).

After discarding the pairs with $S_i = 0$, suppose we are left with r pairs, which we relabel as $s_1, \ldots, s_r$. Our test statistic is S^+, the number of differences that are positive. We evaluate all the possible flips $\pm s_1, \pm s_2, \ldots, \pm s_r$ (there will be 2^r permutations in total) and simply count the number of positive differences. This is a rather lengthy way of calculating the p-value – fortunately, there is a quicker method! We have r independent trials, and under H_0, the probability of a 'success' (i.e. a positive difference) is 0.5. Hence if H_0 is true, S^+ is simply an observation from a binomial distribution with r trials and probability of success 0.5. For large r, we can use a Normal approximation to the binomial distribution.

Example 13(b): Conduct a sign test on the Athlete's 400m times.

Answer: If we have $H_1 : \mathbb{P}(S = 1) < \mathbb{P}(S = -1)$, thus the p-value of our test is given by $\mathbb{P}(S^+ \leq 3)$, which we can evaluate in R:

```
> pbinom(3, 8, 0.5)
[1] 0.3632813
```

This test is easy to perform and has the advantage of not assuming symmetry of the differences between X and Y. On the other hand, since it is based only on the signs of the differences, it is typically less powerful when interval or ratio data are available, compared to using the sample mean or the ranks of the observations.

3.4 One-sample permutation tests

Suppose now that we only observe one sample of data, $(x_1, \ldots, x_n)$, say. We may be interested in testing properties of this distribution (without comparing it to a second distribution). For instance, a one-sample permutation test of $H_0 : \mu_X = \mu_0$ can be conducted in exactly the same way as in Section 3.3.1 above, except that the differences $d_1, \ldots, d_n$ are replaced by $x_1 - \mu_0, \ldots, x_n - \mu_0$. This is best demonstrated using an example:

Example 14: Pine martens As part of a study on pine martens in Kinlochewe, Scotland, radio tags were placed on 12 animals. Subsequent records of habitat usage allowed a habitat utilisation index to be evaluated for each animal in several habitat types. The index was constructed so that a value of zero would be expected if an animal used a habitat type in proportion to its occurrence within the animal's territory. In the case of deciduous woodland, the following values were obtained:

$x = (x_1, \ldots, x_{12}) = (0.13, -0.01, -0.01, 0.42, -0.02, 0.01, 0.09, 0.03, 0.04, 0.06, 0.12, 0.03)$

These values suggest that most animals use deciduous woodland roughly in proportion to its occurrence, but that one or two animals, especially animal four, may have a particular preference for the habitat. We use these data to show how to test the null hypothesis that the mean index value for the population represented by these 12 animals is zero.

The classical approach would be to conduct a one-sample t-test, assuming the data are normally distributed, but that looks dubious here due to the outliers. Instead, we could use a nonparametric test, such as Wilcoxon's signed ranks test, or a sign test.

We can avoid the assumption that the observations are normally distributed, without reducing them to ranks, by using the mean as the test statistic, one slight drawback is that we need to assume that the distribution is symmetric about it's mean μ_X. We will test $H_0 : \mu_X = 0$, versus $H_1 : \mu_X \neq 0$.

Note that if the distribution is symmetric about 0, then the mean must be zero. Under H_0, and a symmetry assumption, the sign of x_i given $|x_i|$ is equally likely to be positive or negative. Thus, each permutation of $\{\pm x_i\}$ is equally likely under H_0.

Consider all possible permutations:

```
 0.13  0.01  0.01  0.42  0.02  0.01  0.09  0.03  0.04  0.06  0.12  0.03
-0.13  0.01  0.01  0.42  0.02  0.01  0.09  0.03  0.04  0.06  0.12  0.03
 0.13 -0.01  0.01  0.42  0.02  0.01  0.09  0.03  0.04  0.06  0.12  0.03
 0.13  0.01 -0.01  0.42  0.02  0.01  0.09  0.03  0.04  0.06  0.12  0.03
:
:
-0.13 -0.01 -0.01 -0.42 -0.02 -0.01 -0.09 -0.03 -0.04 -0.06 -0.12 -0.03
```

In all, there are $2^{12} = 4096$ different permutations. To calculate the p-value, we need to determine how many of these are at least as extreme as the observed permutation of

```
0.13 -0.01 -0.01 0.42 -0.02 0.01 0.09 0.03 0.04 0.06 0.12 0.03
```

We'll use the sample mean as the test statistic: here $\bar{x} = 0.074$. Under H_0, all permutations are equally likely, so we simply need to evaluate the proportion of permutations that yield a sample mean $\geq \bar{x}$. Since the alternative hypothesis is two tailed, we double this proportion to obtain a valid p-value.

These are lengthly calculations to do by hand, we typically use R instead. For the above data, we find that 24 permutations out of 4096 yield a mean ≥ 0.074; thus our p-value is $48/4096 = 0.012$ and we reject H_0.

Although the above test does not assume normality, it does assume symmetry, and in reality, does not offer much more robustness than the t-test. However, the permutation test has the advantage that it can be applied using an assortment of test statistics. Equivalent to the sample mean would be the sum of the observations, as the number of observations n is the same for every permutation. Other options include the sample median, and the number of observations that are positive. The second option turns the permutation test into a sign test. If observations are replaced by signed ranks, the test becomes Wilcoxon's signed ranks test.

The best choice of statistic depends on the null and alternative hypotheses of interest, as well as on the data.

Some advantages of permutation tests:

1. The method is exact.
2. No specific distribution is assumed for the data.
3. The analytic distribution of the test statistic is not required.

Some disadvantages:

1. Each permutation of the data must be equally likely (above we assumed symmetry to ensure this). If they are not, but the probabilities of occurrence are known, permutations can be sampled with appropriate probability.
2. We might suffer from a loss of power is normality assumptions (or similar) are valid.

3.4.1 Non-examinable: Permutation Intervals

We may wish to find a nonparametric confidence interval for μ. To that end, consider testing $H_0 : \mu = \mu_0$. We form the differences $0.13 - \mu_0$, $-0.01 - \mu_0$, ..., $0.03 - \mu_0$, and proceed as before, enumerating all permutations $\{\pm(x_i - \mu_0)\}$.

We found above that for $\mu_0 = 0$ and a two-tailed alternative hypothesis, the p-value was 0.012. If we want a 95% permutation confidence interval, we can increase μ_0 from zero until the p-value (one-tailed test) just rises above 0.025. The corresponding negative value of μ_0 provides the lower 95% confidence limit. In other words, we increase μ_0 until a test of $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$ yields a p-value below 0.025; the upper 95% confidence limit is given by the last value for which the p-value is at least 0.025. Because the permutation distribution is symmetric in this example, this 95% permutation interval comprises all values of μ_0 that are not rejected at the 5% level when $H_0 : \mu = \mu_0$ is tested against the alternative, $H_1 : \mu \neq \mu_0$.

For our example above, we obtain the interval (0.013, 0.150). Note that this interval is not symmetric about the estimate $\bar{x} = 0.074$.

3.5 Randomisation

An important and appealing feature of permutation tests based on the ranks of the observations, is that the distribution of the test statistic under H_0 does not depend on the data, just the sample sizes. Therefore, for small sample sizes, the critical values and/or p-values are stored in R (tables are also available). We simply need to find the observed value of the test statistic and we can immediately find the p-value without actually doing any of the permutations! Different samples of size n all have exactly the same set of ranks, 1 to n, provided there are no ties. For larger sample sizes, we can rely on Normal approximations.

On the other hand, permutation tests based on the original observations (eg. using the mean as the test statistic) are not included in standard R packages. The distribution of the test statistic under the Null will be different for each different dataset, and it can be tricky to program the tests to generate every possible permutation by ourselves. Even for quite small sample sizes, the amount of computation required quickly becomes excessive.

For example, suppose that we observe two (not paired) samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$. If $n = 10$ and $m = 10$, how many permutations of the data between the two samples are there? Double the sample size to 20 in each group, now how many permutations are there?

Answer: For $n = m = 10$ there are $\binom{20}{10} = 184756$ possible combinations with 10 observations in each sample. Doubling our sample size for each group we then have a total of 40 observations and 1.38×10^{11} possible combinations for dividing the observations into two groups of 20. If we calculated one combination per second, it would take 4371 years to enumerate them. Even a computer working a million combinations per second would still take 38 hours to enumerate all possible combinations. Clearly it is not always feasible to consider all possible permutations.

The solution here is to use randomisation! This is a standard trick in permutation based tests – instead of calculating the test statistic for all observations, we can generate a (relatively large) subset at random. Instead of enumerating the full distribution of the test statistic under the null hypothesis by considering all permutations, we generate a random sample of the permutations to get an approximate distribution. Thus, to perform a randomisation test start by generating r independent random permutations and calculating the test statistic for each one. Add the original observed test statistic to the list. Now calculate the proportion of the $r + 1$ values that are at least as extreme as the observed test statistic; this gives the estimated p-value of the test. The larger r is, the better the approximation – the limiting factor is the computational cost.

Example 15: Consider the following data on the production of nitrogen-bound serum albumen in diabetic and non-diabetic mice:

Diabetic (x)					Non-diabetic (y)				
391	46	469	86	174	156	282	197	297	116
133	13	499	168	62	127	119	29	253	122
127	276	176	146	108	249	110	143	64	26
276	50	73			86	122	455	655	14

Here, $n = 18$ and $m = 20$, so there are $\binom{38}{18} = 3.4 \times 10^{10}$ possible permutations, and many will give a test statistic at least as extreme as the observed test statistic.

Suppose, we wish to test the hypothesis:

$$H_0 : \mu_X = \mu_Y \qquad \text{vs} \qquad H_1 : \mu_X \neq \mu_Y,$$

where μ_X and μ_Y are the underlying means for diabetic and non-diabetic mice, respectively. We use the test statistic $d = \bar{x} - \bar{y}$. Since we're using randomisation, it's best done in R:

Answer: Under the null hypothesis the two groups have the same mean and so randomising the group that the observations come from should have little effect on the group means. The following code performs the desired randomisation test.

Using R

```
# Read data in form of a dataframe
> Level <- c(391,46,469,86,174,133,13,499,168,62,127,276,176,146,108,
      276,50,73,156,282,197,297,116,127,119,29,253,122,249,110,143,
      64,26,86,122,455,655,14)
> Group <- rep(c(1,2),c(18,20))
> serum <- data.frame(Level,Group)
# Take a look to make sure the data were read in correctly
> head(serum)
  Level Group
1   391     1
2    46     1
3   469     1
4    86     1
5   174     1
6   133     1
# Attach the dataframe; it saves having to type 'serum$Level' etc.
> attach(serum)

# Calculate the observed test statistic
obstest <- mean(Level[Group==1]) - mean(Level[Group==2])

# Define our own randomising function
> rand.function <- function(nrand, data) {
    # Define a vector r of length nrand+1, this will store
    # the test statistics
    r <- vector(length=nrand+1)
    # For each randomisation do the following
    for (i in 1:nrand) {
        # Randomly reorder the n observations in Level
        randlevel <- sample(data[,1],size=length(data[,1]),replace=F)
        # Calculate the test statistic for each randomisation
        randmean <- mean(randlevel[data[,2]==1])-
                    mean(randlevel[data[,2]==2])
        # Put the test stat for i into element i of vector r
        r[i] <- randmean
    }
    # Add 'obstest' to the vector
```

```
    obstest <- mean(data[,1][data[,2]==1])-mean(data[,1][data[,2]==2])
    r[nrand+1] <- obstest
    # Calculate the p-value of the test as the proportion at least as
    # big as obstest
    p <- length(r[r >= obstest])/length(r)
    # Check we are looking in the correct tail
    if (p > 0.5) {
       # if wrong tail then calculate the proportion at least as
       # small as obstest
       p <- length(r[r <= obstest])/length(r)
    }
    # double it for two-tailed test
    p <- 2*p
    # double check we have a valid probability
    if (p > 1) {
       p <- 1
    }
    # We want the function to return the p-value
    p
}

> rand.function(9999, serum)
[1] 0.984
```

When we used nrand = 99 and ran the code a number of times, I obtained p-values ranging from 0.86 to 0.98. Increasing nrand to 999, we obtain p-values in the range (0.912, 0.988); for nrand = 9999, approximate p-values were in the range (0.972, 0.987); for nrand = 99999, approximate p-values were in the range (0.981, 0.988). By increasing the number of randomisations of the data we obtain more consistent estimates of the p-value; this is because we are evaluating the test statistic for more of the possible permutations.

Example 16: Pine martens example In the one-sample case, random permutations can be generated by listing the set of values $|x_i - \mu|$ (or $|x_i - M|$ if we wish to draw inference about the median M), and randomly assigning a sign to each value ($Pr(+) = Pr(-) = 0.5$). The preferred test statistic is then evaluated for each randomization, and the proportion of these that is at least as extreme as the test statistic for the real data gives the p-value.

For the pine marten example, and using 1000 randomizations, we obtain a p-value of 0.015. This compares well with the exact p-value of 0.012 when all 4096 permutations are enumerated. The following R code performs this test.

Using R

```
> rand.func.one <- function (nrand, data) {
    r <- vector(length=nrand+1)
    n <- length(data)
```

```
    pm <- c(-1,1)
    for (i in 1:nrand){
       rdata <- sample(pm, size=n, replace=T)*data
       r[i] <- mean(rdata)
      }
    teststat <- mean(data)
    r[nrand+1] <- teststat
    # find the p-value
    p <- length(r[r >= teststat])/length(r)
    if (p > 0.5) {
       p <- length(r[r <= teststat])/length(r)
    }
    p <- 2*p
    p
  }

# pine martens example
> data <- c(0.13,-0.01,-0.01,0.42,-0.02,0.01,0.09,0.03,0.04,
            0.06,0.12,0.03)
> rand.func.one(999, data)
```

3.5.1 Matched-Pairs Randomisation Test

We can also use randomisation tests for matched pairs. Recall that for a matched pair, we flip the sign of each of the n observations to generate the permutations – this gives 2^n possibilities. If n is around 20 or less, then we likely can try all possibilities, but for $n = 100$ or more (which is very modest by modern standards) we have 1×10^{30} possibilities, and we will likely need to resort to using a randomised method.

Example 17::For a manageable example, consider the Athletes' 400m times again (here $n = 8$, so we don't need randomisation, but it's useful to demonstrate how it works):

Runner	1	2	3	4	5	6	7	8
Time at sea level	48.3	47.6	49.2	50.3	48.8	51.1	49.0	48.1
Time at altitude	50.4	47.3	50.8	52.3	47.7	54.5	48.9	49.9
Difference	-2.1	0.3	-1.6	-2.0	1.1	-3.4	0.1	-1.8

Suppose we wish to test $H_0 : \mu_x = \mu_y$ against $H_1 : \mu_x < \mu_y$, and we use the mean difference as our test statistic. Use R to run an appropriate randomisation test.

Answer: If the differences are in a variable called `diff`, we can generate a single randomisation as follows:

```
> n <- length(diff)
> pm <- c(-1,1)
> rdiff <- sample(pm, n, replace=T)*diff
```

The proportion of randomly-generated permutations that have a mean less than or equal to the mean of the original sample (-1.175) is the approximate p-value. For a two-sided alternative, we would double this p-value.

Using R

```
> rand.func.pairs <- function(nrand, data) {
    r <- vector(length=nrand+1)
     n <- length(data)
     pm <- c(-1,1)
    for (i in 1:nrand){
       rdiff <- sample(pm, n, replace=T)*data
       r[i] <- mean(rdiff)
     }
     teststat <- mean(data)
     r[nrand+1] <- teststat
     # find the p-value
     p <- length(r[r <= teststat])/length(r)
     p
  }

> rand.func.pairs(999,diff)
[1] 0.025
```

3.5.2 A more general way of looking at randomisation

In the permutation and randomisation tests we have seen so far, we have typically calculated p-values by reassigning the data randomly to groups. An alternative approach is to reassign the groups randomly to the data. This is particularly easy to see with reference to a randomisation part of the functions seen so far:

```
> randlevel <- sample(data, n, replace=F)
```

This generates a single random permutation of the data in `data`; the variable `index` indicates the sample to which the randomised (randomly reordered) observations are assigned. We could instead randomise the index:

```
> randindex <- sample(index, n, replace=F)
```

We now assign the observations to the two samples according to the index values in `randindex`. Both methods are equivalent: *it does not matter whether we permute the data or the labels identifying groups*. This observation shows that randomisation methods are widely applicable, we demonsrate this with an example.

Randomisation tests for linear models: Suppose that we have pairs of observations, $(x_1, y_1), \ldots, (x_n, y_n)$, that satisfy a linear model

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

where ϵ_i is a mean zero (not necessarily Normal) error term. Then we can use randomisation to test $H_0 : Y_i = \alpha + \epsilon_i$, versus $H_1 : Y_i = \alpha + \beta x_i + \epsilon_i$.

A standard approach to fitting such a model is to use least squares: Let

$$S(\alpha,\beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

Then we estimate α and β by

$$(\hat{\alpha},\hat{\beta}) = \text{argmin}_{\alpha,\beta}\{S(\alpha,\beta)\}.$$

and $\hat{S} = S(\hat{\alpha},\hat{\beta})$ is a measure of how well the model fits.

Now, under H_0, the distribution of Y_i does not depend on x_i, thus the observed data should appear to be a typical member of the population of data sets obtained by reassigning the $x_1,\ldots,x_n$ randomly to $y_1,\ldots,y_n$. As above, 'appearance' will be judged by an appropriate test statistic. We can use $\hat{S}$.

Under the null hypothesis, the x_i term does not offer any improvement to the model fit, so that the observed $\hat{S}$ should be a typical element of the distribution of $\hat{S}$ values obtained by reassigning x_is randomly to the y_is and recalculating $\hat{S}$ (by re-fitting the model). Our p-value will be the proportion of times we obtain a smaller value on $\hat{S}$ than the one we observed Performing such a hypothesis test by randomisation gives a means for deciding whether or not there is sufficient evidence to retain a given explanatory variable in the model. This is best seen by the following simple example:

Example 18: Using R

```
x <- c(1,2,3,4,5,6,7,8)
y <- c(27,32,39,45,34,48,39,41)
#test for beta = 0
n <- length(x)
teststat <- cor(x,y)
nr <- 9999
rcor <- vector(length=nr+1)
for(i in 1:nr){
     rx <- sample(x, n, replace=FALSE)
     rcor[i+1] <- cor(rx,y) }
rcor[1] <- teststat
p <- 2*sum(rcor>=teststat)/length(rcor)
p
[1] 0.0966

#test for  y = a + err  vs y = a + bx + err,

#centre the xs to simplify working
x <- x - mean(x)

plot(x,y)

#define parameter
n <- length(x)

#least squares function
S <- function(a,b)
```

```
{
  return(sum((y - a - b*x)^2))
}

##fitted parameters
ahat <- mean(y)
bhat <- sum(x*y)/sum(x^2)
abline(a = ahat, b = bhat)

#test stat
Shat <- S(ahat, bhat)


#Number of permutations
R <- 1000

#Null variable to store answers
Sout <- NULL

#loop over 1:R
for(r in 1:R)
  {
  xr <- sample(x, n, replace=FALSE)
  ahatr <- mean(y)
  bhatr <- sum(xr*y)/sum(xr^2)
  Sout <- c(Sout, sum((y - ahatr - bhatr*xr)^2))
  }

#add obs value
Sout <- c(Sout,Shat)

#check dist
plot(hist(Sout))

#calculate p value
pvalue <- mean(Sout <= Shat)
```

For more information on the full spectrum of randomisation methods, see B. F. J. Manly, *Randomisation, Bootstrap and Monte Carlo Methods in Biology*, Chapman & Hall, 1997.

3.5.3 Non-examinable: Randomisation confidence intervals

Suppose we seek a 95% confidence interval for the mean index value in our example. By assuming that index values are normally distributed and using the t-statistic, we can obtain a confidence interval of (-0.002, 0.150). We will use these limits as the start points for searches

for better confidence limits, using randomization tests and Robbins-Monro search.

Consider the upper limit μ_U, and set up $H_0 : \mu = \mu_U$ against $H_1 : \mu < \mu_U$. At step j, denote the current estimate of μ_U by U_j. Subtract U_j from each observation, then generate a permutation at random from the resulting values. If $\bar{x}$ is the mean of the original data, and $\bar{y}_j = \bar{x}_j - U_j$ is the mean of values in the permutation generated at step j, then update the estimate of μ_U as follows:

$$U_{j+1} = \begin{cases} U_j - c\alpha/j, & \text{if } \bar{x}_j > \bar{x} \\ U_j + c(1-\alpha)/j, & \text{if } \bar{x}_j \le \bar{x} \end{cases} \tag{3.1}$$

where $\alpha = 0.025$ and c is a steplength constant.

This search oscillates widely for small j. Hence we start the search from a reasonable approximation to the upper confidence limit, 0.150 in the example, and arbtrarily set say $U_{40} = 0.150$. The number of steps might vary from a few hundred to many thousand. The optimal value of the steplength constant depends on the true distribution of the test statistic, but performance of the method is generally good if it is set to a generic value, equal to double the optimal value for the normal distribution. It is calculated at step j as $c = k(U_j - \bar{x})$ where

$$k = \frac{2\sqrt{2\pi}}{z_\alpha \exp(-z_\alpha^2/2)} \tag{3.2}$$

and $z_\alpha = 1.96$ for $\alpha = 0.025$. Under very general conditions, U_j converges to μ_U as $j \to \infty$. The lower limit is found similarly.

3.6 Comparison of several samples

3.6.1 Kruskal–Wallis Test

In this section we introduce the Kruskal-Wallis test, which is an extension to $K > 2$ samples of the two-sample Wilcoxon test. To motivate it, we first review some standard parametric theory (ANOVA).

ANOVA: Comparison of K samples, assuming normal distributions. Suppose we have K samples of sizes $n_1, \cdots, n_K$, the kth sample being $X_{k,1}, \cdots X_{k,n_k}$, the samples being from normal distributions with the same (unknown) variance σ^2 but possibly different means, denoted by μ_k. The standard test of the null hypothesis that the means are the same uses the variance-ratio statistic:

$$F = \frac{S_B/(K-1)}{S_W/(n-K)} = \frac{(K-1)^{-1}\sum_{k=1}^K n_k(\bar{X}_k - \bar{X})^2}{(n-K)^{-1}\sum_{k=1}^K \sum_{i=1}^{n_k}(X_{k,i} - \bar{X}_k)^2}.$$

If the null hypothesis is true (ie. the means are all the same), then this has a $F_{K-1,n-K}$ distribution. Here $n = \sum_{k=1}^K n_k$ and $\bar{X}_k$ is the mean of the kth sample, and $\bar{X}$ is the overall mean. The denominator in the expression for F functions is an estimate of σ^2; if σ^2 was known we could instead use the statistic

$$S = \frac{\sum_{k=1}^K n_k(\bar{X}_k - \bar{X})^2}{\sigma^2}$$

whose null distribution is χ^2_{K-1}. This is the standard (parametric) approach to ANOVA (which many of you will have seen before).

The Kruskal-Wallis test: To obtain a distribution-free procedure we replace the observations in the K samples by their ranks in the combined sample of size n. That is we replace $X_{k,i}$ by its rank $R_{k,i}$ and $\bar{X}_k$ by the mean rank $\bar{R}_k = n_k^{-1} \sum_{i=1}^{n_k} R_{k,i}$. Then $\bar{X}$ is replaced by the overall mean rank $\frac{1}{2}(n+1)$ and σ^2 by the variance of the ranks, which is $\frac{1}{12}(n^2-1)$. This leads to the statistic

$$S = \frac{12}{n^2-1} \sum_{k=1}^{K} n_k \left\{ \bar{R}_k - \frac{1}{2}(n+1) \right\}^2.$$

In fact the test introduced by Kruskal and Wallis used the statistic $H = (n-1)S/n$. Writing $R_k = \sum_{i=1}^{n_k} R_{k,i} = n_k \bar{R}_k$, the sum of the ranks of the kth sample, we have the alternative formula

$$H = \frac{12}{n(n+1)} \sum_{k=1}^{K} \frac{R_k^2}{n_k} - 3(n+1).$$

Note that this is just another permutation test – for small sample sizes the null distribution of H is tabulated (or we can use R of course). For large samples the null distribution is approximately χ^2_{K-1}.

If ties are present, the effect is to reduce the variance of the ranks, from $\frac{1}{12}(n^2-1)$ to $\frac{1}{12}(n^2-1) - \frac{1}{12n}\sum_{ties}(t^3-t)$, where sum is over all ties, t being the number of tied members in each tie. Taking the ties into account one gets the modified definition of H:

$$H = \frac{12(n-1)}{n^3 - n - \sum(t^3-t)} \left\{ \sum_{k=1}^{K} \frac{R_k^2}{n_k} - \frac{n(n+1)^2}{4} \right\}.$$

The effect is usually small, unless there are lots of ties.

Example 19: The following data are on measurements of maximum head width in units of 0.01 mm for three species of *Chaetocnema*. Determine if there is a species difference in head widths.

Species 1	53	50	52	50	49	47	54	51	52	57	
Species 2	49	49	47	54	43	51	49	51	50	46	49
Species 3	58	51	51	45	53	49	51	50	51		

Answer: The ranks in the three samples are as follows:

Species 1	53	50	52	50	49	47	54	51	52	57
Rank	25.5	13.5	23.5	13.5	8.5	4.5	27.5	19	23.5	29
Species 2	49	49	47	54	43	51	49	51	50	46	49
Rank	8.5	8.5	4.5	27.5	1	19	8.5	19	13.5	3	8.5
Species 3	58	51	51	45	53	49	51	50	51		
Rank	30	19	19	2	25.5	8.5	19	13.5	19		

We have $K = 3$, $n_1 = 10$, $n_2 = 11$, $n_3 = 9$, $n = 30$ and summing the ranks gives $R_1 = 188$, $R_2 = 121.5$, $R_3 = 155.5$. Then $\sum_{k=1}^{3} \frac{R_k^2}{n_k} = \frac{188^2}{10} + \frac{121.5^2}{11} + \frac{155.5^2}{9} = 7563.12$. Then without correction for ties we get $H = \frac{12 \times 7563.12}{30 \times 31} - 3 \times 31 = 4.59$.

To correct for ties, we have four 2-fold ties and one 4-fold, one 6-fold and one 7-fold tie. So $\sum(t^3 - t) = 4(2^3 - 2) + 4^3 - 4 + 6^3 - 6 + 7^3 - 7 = 630$. Then $H = \frac{12\times29}{30^3-30-630}(7563.12 - \frac{30\times31^2}{4}) = 4.70$. The sample sizes are too large to be covered by the specified table, so we use the χ^2_2 distribution. For the 5% level the critical value of χ^2_2 is 5.991 so we do not reject the hypothesis of no difference.

Using R

We can do the Kruskal-Wallis test in R as bellow:

```
> Species1 <- c(53, 50, 52, 50, 49, 47, 54, 51, 52, 57)
> Species2 <- c(49, 49, 47, 54, 43, 51, 49, 51, 50, 46, 49)
> Species3 <- c(58, 51, 51, 45, 53, 49, 51, 50, 51)
> width <- c(Species1, Species2, Species3)
> group <- rep(1:3, c(10,11,9))
> kruskal.test(width, group)

        Kruskal-Wallis rank sum test
data:  width and group
Kruskal-Wallis chi-squared = 4.6984, df = 2, p-value = 0.09545
```

A boxplot can visualise the difference among the groups:

```
> boxplot(width ~ group, xlab="group", ylab="width")
```

3.6.2 Multiple-sample Randomisation Tests

An alternative option to using the ranks in the Kruskal-Wallis test in multi-sample problems, is to use the sample means, similar to the way we did at the beginning of this chapter with $K = 2$.

$$H_0\text{: all of the distributions are the same}$$
$$\text{vs}$$
$$H_1\text{: at least one distribution is different.}$$

Suppose that we are primarily interested in detecting differences in the means of the populations. Then the underlying idea remains the same as in the two-sample case. Under H_0, we can reshuffle the observations between groups and the observed data will appear similar to those simulated if H_0 is true. We just need to define a test statistic t that summarises the data in a useful way.

Suppose again we observe $(x_{k,i} : i = 1, \ldots, n_k, k = 1, \ldots, K)$. Here, k denotes the group of the data point and i is indexing the observations within each group, so that we have n_k observations in the kth group.

Let $\bar{x} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k} x_{k,i}}{\sum_{k=1}^{K} n_k}$ be the overall sample mean and let $\bar{x}_k = \frac{\sum_{i=1}^{n_k} x_{k,i}}{n_k}$ the sample mean of group j. Then let

$$T := \sum_{k=1}^{K} n_k(\bar{x}_k - \bar{x})^2.$$

be our test statistic. Note that $\text{Var}(\bar{x}_k) \propto \frac{1}{n_k}$. Now, small values of t would support H_0, whereas large values would be evidence against H_0 in favour of H_1. In order to calculate a p-value we can use randomisation – we will reassign the observations to the K groups (keeping the group sample sizes fixed) uniformly at random. This is best seen using an example.

Example 20: We observe data relating to the treatment of anorexia using three different treatments. We are interested in whether the different treatments produce different results in terms of weight gain of the individuals. The treatments are CBT (cognitive behavioural therapy, $j = 1$), a standard treatment ($j = 2$) and family therapy ($j = 3$). The data are the weight gains (in lb) that resulted:

C.B.T.:

$$(x_{1,1}, \ldots, x_{n_1,1}) = (1.7, 0.7, -0.1, -0.7, -3.5, 14.9, 3.9, 17.1, -7.6, 1.6, 11.7, 6.1, 1.1, -4.0, 20.9, -9.1, 2.1, -1.4, 1.4, -0.3, -3.7, -0.8, 2.4, 12.6, 1.9, 3.9, 0.1, 15.4, -0.7)$$

Standard:

$$(x_{1,2}, \ldots, x_{n_2,2}) = (-0.5, -9.3, -5.4, 12.3, -2.0, -10.2, -12.2, 11.6, -7.1, 6.2, -0.2, -9.2, 8.3, 3.3, 11.3, 0.0, -1.0, 11.6, -4.6, -6.7, 2.8, 0.3, 2.0, 3.7, 5.9, 10.2)$$

and **Family therapy**:

$$(x_{1,3}, \ldots, x_{n_3,3}) = (11.4, 11.0, 5.5, 9.5, 13.6, -2.9, -0.1, 7.4, 21.5, -5.3, -3.8, 13.4, 13.1, 9.0, 3.9, 5.7, 10.7)$$

Answer: The following R code could be used to implement a randomisation test.

Using R

```
# Read in, look at and attach data
> anorexia <- read.table("anorexia.txt", header=T)
> head(anorexia)
   Diff Group
1   1.7     1
2   0.7     1
3  -0.1     1
4  -0.7     1
5  -3.5     1
6  14.9     1
> attach(anorexia)

# find sample sizes
> n1 <- length(Diff[Group==1])
> n2 <- length(Diff[Group==2])
> n3 <- length(Diff[Group==3])

# calculate the observed test statistic
> teststat <- n1*(mean(Diff[Group==1])-mean(Diff))^2 +
              n2*(mean(Diff[Group==2])-mean(Diff))^2 +
              n3*(mean(Diff[Group==3])-mean(Diff))^2
```

```
# create a function to perform the randomisation test
> multrand.function <- function(nrand, data) {
   r <- vector(length=nrand+1)
   for (i in 1:nrand) {
     randdiff <- sample(data[,1], length(data[,1]), replace=F)
     randstat <-
     length(data[,1][data[,2]==1])*(mean(randdiff[data[,2]==1])
     -mean(data[,1]))^2+
     length(data[,1][data[,2]==2])*(mean(randdiff[data[,2]==2])
     -mean(data[,1]))^2 +
     length(data[,1][data[,2]==3])*(mean(randdiff[data[,2]==3])
     -mean(data[,1]))^2
     r[i] <- randstat
     }
  # add the observed test statistic to the list
  teststat <-
  length(data[,1][data[,2]==1])*(mean(data[,1][data[,2]==1])
  -mean(data[,1]))^2+
  length(data[,1][data[,2]==2])*(mean(data[,1][data[,2]==2])
  -mean(data[,1]))^2+
  length(data[,1][data[,2]==3])*(mean(data[,1][data[,2]==3])
  -mean(data[,1]))^2
  r[i+1] <- teststat
  # find the p-value
  p <- length(r[r >= teststat])/length(r)
  p
}

> multrand.function(9999, anorexia)
[1] 0.024
```

Running the R code with nrand=9999, I obtained a p-value of 0.024. Thus, there is evidence to reject H_0 in favour of H_1, we reject H_0 at the 5% level. There is moderate evidence that at least one treatment has a different mean.

We could have tried to calculate the test statistic for all possible permutations here, but even with small sample sizes the number of possible permutations is prohibitively large.

Chapter 4

Testing for independence Or: More nonparametric tests

4.1 Introduction

In this chapter, we discuss (nonparametric) tests for independence. The first, *the Runs Test*, assesses whether our data sample is independent and identically distributed. We will then introduce the nonparametric equivalent of Pearson correlation coefficient between two variables in order to test for (in)dependence. Our main tool will again be the use of permutations of the data.

4.2 The runs test

This test examines whether the order of a set of observations is consistent with the observations forming a random sample (i.e. being independent and identically distributed). The ordering is usually determined by the order in which observations are recorded.

The data need only be nominal. In fact, we just need each observation in the data sequence to fall into one of two distinct types (e.g. Heads or Tails). Data that are not initially of this type must be put into the required form. For nominal data with more than two categories, this can be achieved by pooling categories. For data that are at least ordinal, observations can be classified by their position relative to the median (if any observations actually equal the median, either assign all of them to the upper category or all to the lower category).

We begin with an example:

Example 21 A psychologist recorded the order in which 26 adults arrived at a social function. The guests were later asked to complete a social skills test. Their scores, when sorted according to their arrival order at the function, were as follows (read row by row):

```
52  42  26  39  41  54  36  55  56  53  42  59  36
54  21  24  49  57  29  53  59  47  52  36  43  48
```

These data are plotted in Figure 4.2. The psychologist wishes to know if the test score is related to arrival order, i.e., do people with low social skills scores arrive earlier or later than those with high scores?

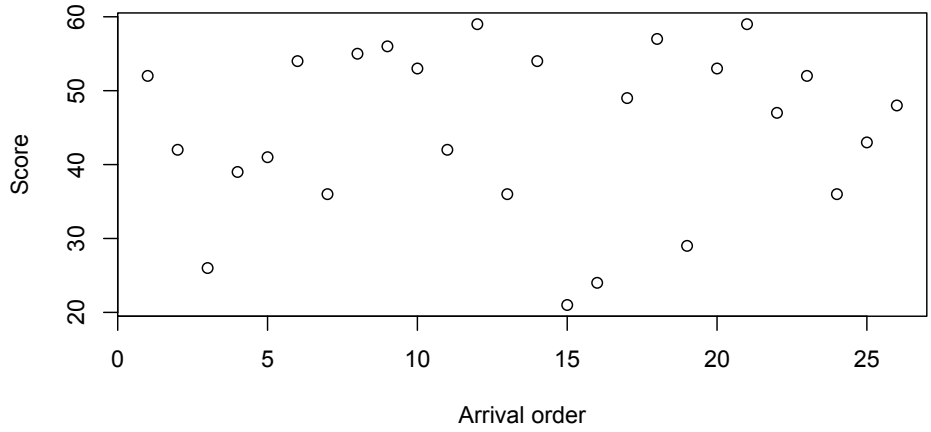


Figure 4.1: Social skills scores against arrival order.

We could try to answer this question using linear regression (with arrival order as the explanatory variable and score as the response), and testing the null hypothesis that the slope of the regression is zero. But to do this we need to make some assumption about the distribution of scores (e.g. that they are normal random variables), and we are not confident that the assumption is reasonable.

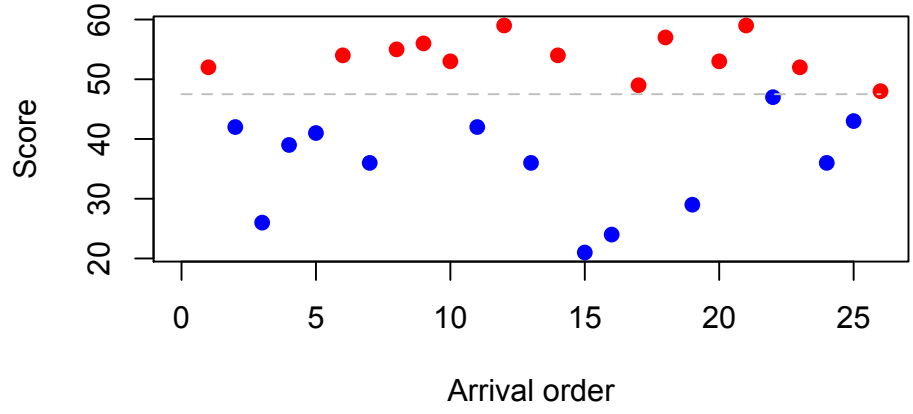


Figure 4.2: Social skills scores against arrival order. Red dots are people with scores greater than the median score, blue dots are people with scores less than the median. The grey dashed line is the median score.

We'd like an assumption-free test. If there is no relationship between arrival order and score, then these scores are in random order. We can assess whether or not this is the case by testing H_0 : $\{x_1, x_2, \cdots x_n\}$ is a random sequence, i.e. the observed values of a sequence of i.i.d. (independent, identically distributed) random variables (without specifying anything

about the distribution). To test the hypothesis we need a statistic with a known distribution under H_0, and whose distribution departs from this under the alternative hypothesis. We will make use of the notion of a *run*; a maximal subsequence of somehow "similar" observations. The general method will involve transforming the data into a binary sequence (0s and 1s, or heads and tails), by splitting at the median, each outcome should be equally likely.

Example 22 Suppose a coin is tossed 12 times and the outcomes in order of occurrence are:-

H H H T H H T T T T H H

Here we have 7 heads and 5 tails, which occur in $Q = 5$ runs (3 runs of heads and 2 runs of tails). We will take as our test statistic $Q = \{\text{total number of runs}\}$. If the null hypothesis is true, then the sequence should look like a typical sequence of independent coin flips.

On the other hand, if there are a small number of runs, i.e. all heads then all tails ($Q = 2$), or we alternate between heads and tails (Q large), this provides evidence against H_0. Notice that this means we have a two-sided test.

4.2.1 The exact test

We need to determine the distribution of Q under H_0. This could be calculated exactly (although this is not easy to do except for very small sample sizes. For example, if $n = 5$, and we have 3 heads and 2 tails, we can enumerate all possibilities: we need to choose which 2 positions out of 5 in which to put the tails, thus we have $\binom{5}{2} = 10$ possibilities. We can calculate Q for each possible sequence as follows

$$\begin{cases} Q = 2 & \text{is 2 (HHHTT; TTHHH)} \\ Q = 3 & \text{is 3 (HHTTH; HTTHH; THHHT)} \\ Q = 4 & \text{is 4 (HHTHT; HTHHT; THHTH; THTHH)} \\ Q = 5 & \text{is 1 (HTHTH)} \end{cases}$$

Hence the distribution of Q under H_0 is:

q	2	3	4	5
$P(Q = q)$	0.2	0.3	0.4	0.1

In general, this distribution depends on n_1 and n_2, the number of observations of the first and second type, respectively. In practice, we can find p-values by calculating Q all possible permutations; for small values of n_1 and n_2 we can use R (or tables) to find the p-value and critical values. For larger samples, we can use a normal approximation.

4.2.2 Normal approximation

When $n = n_1 + n_2$ is large, under H_0 we have that

$$Q \dot{\sim} N\left(\mu_Q, \sigma_Q^2\right)$$

where,

$$\mu_Q = 1 + \frac{2n_1n_2}{n_1 + n_2}, \qquad \sigma_Q^2 = \frac{2n_1n_2\left(2n_1n_2 - n_1 - n_2\right)}{(n_1 + n_2)^2\left(n_1 + n_2 - 1\right)} \tag{4.1}$$

Thus, for a two-sided test of size $\alpha = 0.05$, we reject H_0 if

$$\frac{|Q - \mu_Q|}{\sigma_Q} \geq z_{0.975} = 1.96.$$

Example 21 (part 2): Social Skills We return to the social skills example above. The median score in the sample is 47.5. We wish to test whether the arrivals in random order?

```
52 42 26 39 41 54 36 55 56 53 42 59 36
54 21 24 49 57 29 53 59 47 52 36 43 48
```

Answer: Denoting A for above the median and B for below, we have:

ABBBBABAAABABABBAABAABABBA

Hence the observed number of runs is $q = 17$. Our sample size is on the low side for the normal approximation, but we will use it nonetheless and then compare our result with what we get using R.

We have $n_1 = 13$ observations above the median and $n_2 = 13$ observations below the median. From equation (4.1), the mean and variance of Q under H_0 are 14 and 6.24. By normal approximation, we have

$$P(Q \geq 17) = P\left(Z > \frac{17 - 14}{\sqrt{6.24}}\right) = P(Z > 1.201) = 0.115$$

Thus, for a two-tailed test, the (approximate) p-value is 0.230. Hence, we have no evidence against the null hypothesis that the arrivals are in random order.

Using R

As our sample sizes were on the small side for the normal approximation to be good, we check our result using R. Although the base package does not contain the runs test, it is available in the TSA (Time Series Analysis) package, which can be installed and called up by

```
> install.packages("TSA")
> library(TSA)
```

The runs test is then implemented as follows:

```
> x <- c(52,42,26,39,41,54,36,55,56,53,42,59,36,54,21,24,49,57,29,
         53,59,47,52,36,43,48)
> runs(x,median(x))
$pvalue
[1] 0.313
$observed.runs
[1] 17
$expected.runs
[1] 14
$n1
[1] 13
$n2
[1] 13
$k
[1] 47.5
```

There is also a `runs.test` option in R, which requires the data vector to be a factor:

```
> install.packages("tseries")
```

```
> library("tseries")
> x <- c(52,42,26,39,41,54,36,55,56,53,42,59,36,54,21,24,49,57,
         29,53,59,47,52,36,43,48)
> y <- rep(0,length(x))
> y[x<median(x)] <- 1
> runs.test(factor(y),alternative="two.sided")

Runs Test
data:  factor(y)
Standard Normal = 1.201, p-value = 0.2298
alternative hypothesis: two.sided
```

4.2.3 Using the runs test for equality of two distributions

In addition to testing whether a series of observations of two kinds occur in random order, we can use the runs test to test whether two **independent** samples are drawn from the same distribution. This is called the Wald–Wolfowitz two-sample test. It tests the equality of two cumulative distribution functions F_X and F_Y. We sort the $n_1 + n_2$ combined observations from F_X and F_Y in increasing order then count the number of runs of observations from each distribution in this sorted list (where we have n_1 observations from F_X and n_2 from F_Y). Note that we do not need to observe the xs and ys to do this test - we just have to be able to order them.

Consider the arrival times of people at a party in the previous example. Let the random variable X be the arrival times of people with high social skills scores and Y be the arrival times of those with low scores. If $F_X = F_Y$, then the observed xs and ys should occur in random order. We can find the distribution of the number of runs, Q, under the null hypothesis that $F_X = F_Y$, in which case the xs and ys occur in random order. Now if F_X and F_Y are not the same, we should observe too few runs – see Figures 4.3 to 4.5. Because we are looking for deviations in the direction of Q being too small here only, we would use a one-sided test.

Why one-sided? We are testing whether the distributions of two *independent* random variables are the same. Figures 4.6 has the red and blue distributions identical and the number of runs higher than would be expected from two identical independent distributions. The high number of runs results from x and y not being independent (if reds and blues arrived in pairs, for example). If we used a two-sided test we would likely reject the hypothesis that the two distributions are identical if we got this many runs, even though the two distributions are identical. In doing this, we would reach the wrong conclusion. Whenever the blue and red distributions are different, we get too few runs. So in looking for departures from what we would expect if the two *independent* sets of random variables came from two different distributions, we should only consider departures in the direction of too few runs.

Example 23: (Example 8 revisited) A motorist, who stays in St Andrews, makes frequent journeys to visit relatives in Perth. On five occasions he travels via Newburgh, and his outward journey times $x_1, x_2, \cdots, x_5$ are 51, 55, 58, 50 and 53 minutes. On another five occasions he tries going via Dundee. His outward journey times $y_1, y_2, \cdots, y_5$ are 57, 60, 54, 63 and 56 minutes. We can use the runs test to assess whether the distributions are the same.

We first list the observations in increasing order, with the corresponding sample label:

Figure 4.3: Example of runs from two identical distributions. The plot shows two cumulative distribution functions and a sample of 7 random variables from each.

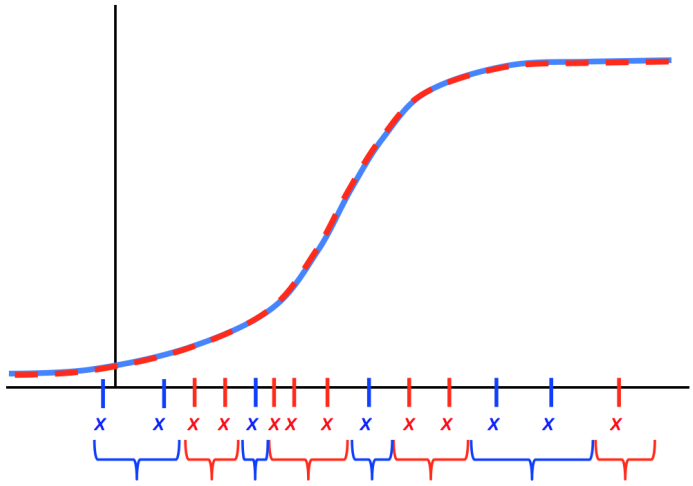
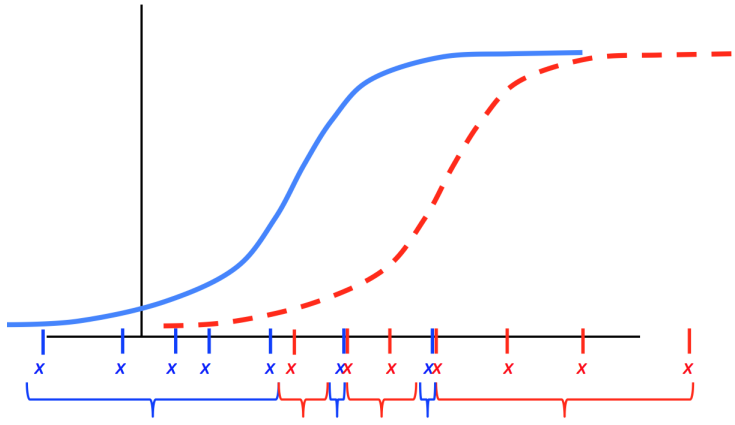


Figure 4.4: Example of runs from two distributions with different means and the same variance. The plot shows two cumulative distribution functions and a sample of 7 random variables from each.



	1	2	3	4	5	6	7	8	9	10
Order statistics	50	51	53	54	55	56	57	58	60	63
Sample	x	x	x	y	x	y	y	x	y	y

We see that there are 6 runs. This is exactly what we expect under the null here, so we will not reject the null; indeed the (approximate) one-sided p value in this case is (from the R code below) 0.5.

Figure 4.5: Example of runs from two distributions with different variances and the same mean. The plot shows two cumulative distribution functions and a sample of 7 random variables from each.

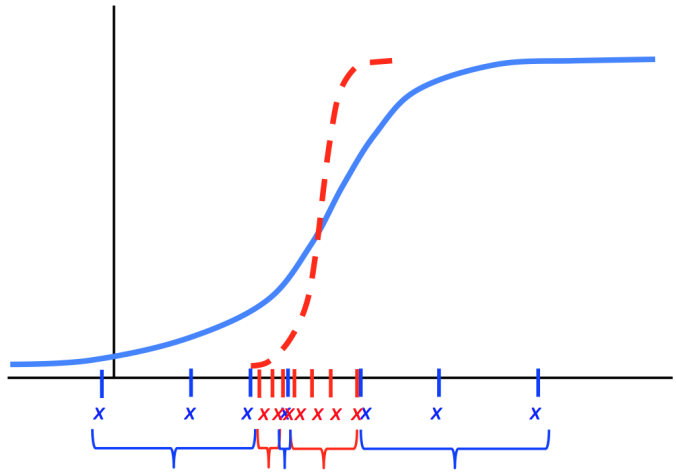
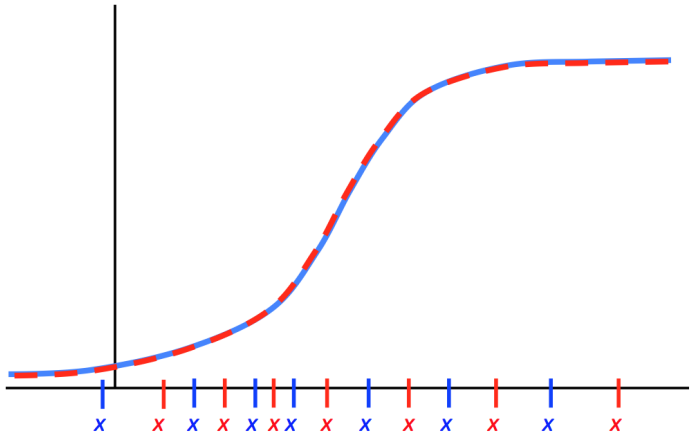


Figure 4.6: Example of runs from two identical distributions. The plot shows two cumulative distribution functions and a sample of 7 random variables from each. (Why might there be so many runs compared to Figure 4.3?)



```
> #Example 23
> x <- c(51, 55, 58, 50, 53)
> y <- c(57, 60, 54, 63, 56)
> library(tseries)
> sample <- c(rep("x",5),rep("y",5))
> runseq <- factor(sample[order(c(x,y))])
> runs.test(runseq, alternative = "less")

Runs Test

data:  runseq
Standard Normal = 0, p-value = 0.5
alternative hypothesis: greater
```

4.2.4 Notes on the runs test

- Instead of using the total number of runs, it is also possible to test for randomness using the length of the longest run as the test statistic.
- There are more powerful nonparametric tests than the Wald-Wolfowitz two sample test for testing whether two independent samples are drawn from the same distribution.

4.3 Testing for independence between two samples

We now move on to testing for dependence between two-samples. We saw a test for independence between the rows and columns in a two-way contingency table in Section 2.3.1. Here will will consider tests applicable to two continuous samples.

Our main tool will be the correlation between the samples. Suppose we have a pairs of observations $(X_1, Y_1), \cdots, (X_n, Y_n)$ of a pair of variables X, Y and are interested in whether X and Y are independent. If the pairs are jointly normal distribution, then the sample correlation coefficient

$$r := \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\{(\sum_{i=1}^n (X_i - \overline{X})^2)(\sum_{i=1}^n (Y_i - \overline{Y})^2)\}^{1/2}}$$

can be used as an estimate of the correlation between X and Y, and if X and Y are jointly normal, then the correlation is zero if and only if they are independent.

Here we wish to avoid a normality assumption and will therefore use other (more robust) measures of correlation.

4.3.1 Spearman's rank correlation (sometimes called Spearman's ρ)

We can obtain a distribution-free test of independence by replacing the observations by their ranks. Let Q_i be the rank of X_i amongst $\{X_1, \cdots, X_n\}$ and let R_i be the rank of Y_i amongst $\{Y_1, \cdots, Y_n\}$. Then Spearman's rank correlation coefficient r_s is

$$r_s = \frac{\sum_{i=1}^n (Q_i - \overline{Q})(R_i - \overline{R})}{\{(\sum_{i=1}^n (Q_i - \overline{Q})^2)(\sum_{i=1}^n (R_i - \overline{R})^2)\}^{1/2}}.$$

This expression can be simplified by noting that, in the absence of ties (we assume throughout this section that there are no ties within the samples) the Q_i are the numbers $\{1, \cdots, n\}$ in some order, likewise the R_i, and then we have $\overline{Q} = \overline{R} = \frac{n+1}{2}$ and $\sum_{i=1}^n (R_i - \overline{R})^2 = \frac{n(n^2-1)}{12}$ with the same for Q. Then we have (equivalently)

$$r_s = \frac{12}{n(n^2-1)} \sum_{i=1}^n \left(Q_i - \frac{n+1}{2}\right)\left(R_i - \frac{n+1}{2}\right) = \frac{12}{n(n^2-1)} \left(\sum_{i=1}^n Q_i R_i - \frac{n(n+1)^2}{4}\right).$$

Another alternative (and also equivalent) expression is

$$r_s = 1 - \frac{6S}{n(n^2-1)}$$

where $S := \sum_{i=1}^n (Q_i - R_i)^2$. Observe that $-1 \le r_s \le 1$; and $r_s = 1$ only when the two rankings coincide and $r_s = -1$ only when they are exactly opposite.

We wish to test

$$H_0 : X, Y \quad \text{independent; versus} \quad H_1 : X, Y \quad \text{not independent.}$$

We will use r_s as our test statistic, our first task is to find the distribution of r_s under the null. Suppose the X sample is arranged in order of increasing X_i then we have $Q_i = i$ and $r_s = \frac{12}{n(n^2-1)}(V - \frac{n(n+1)^2}{4})$ where $V = \sum_{i=1}^n iR_i$. Under H_0, the sequence $R_1, \cdots, R_n$ is equally likely to be any permutation of $\{1, \cdots, n\}$. Hence V is a linear rank statistic, and $\mathbb{E}(V) = \frac{n(n+1)^2}{4}$ and $\text{Var}(V) = \frac{1}{n-1}(\frac{n(n^2-1)}{12})^2$. It follows that, $\mathbb{E}(r_s) = 0$ and

$$\text{Var}(r_s) = \frac{1}{n-1}$$

In other words, the null distribution of r_s is symmetric about 0, with variance $1/(n-1)$. If n is large, we can use a normal approximation to this distribution. For small n, we could find the distribution using permutations. In practice, we can use the cor.test function in R.

Example 24(a) Eleven children are given an arithmetic test; after three weeks special tuition they are given a further test of equal difficulty. Their marks in each test are given below. Are the results on the two tests correlated?

First test	45	61	33	29	21	47	53	32	37	25	81
Second test	53	67	47	34	31	49	62	51	48	29	86

Answer: Let X_i be the score of the ith child on the first test and Y_i the score on the second. The scores and ranks are tabulated below:

X_i	45	61	33	29	21	47	53	32	37	25	81
Q_i	7	10	5	3	1	8	9	4	6	2	11
Y_i	53	67	47	34	31	49	62	51	48	29	86
R_i	8	10	4	3	2	6	9	7	5	1	11
$(Q_i - R_i)^2$	1	0	1	0	1	4	0	9	1	1	0

Summing the last row we get $S = 18$ and $r_s = 1 - \frac{6\times 18}{11^3-11} = \frac{101}{110} = 0.918$. We can do a test to check whether this correlation is significant. We have that $R_s \overset{\cdot}{\sim} N(0, \frac{1}{n-1})$ under H_0 and our (approximate) p-value (probability of observing a correlation at least as extreme as the observed one) is

$$2\mathbb{P}(R_s > 0.92) = 2\mathbb{P}(Z > \frac{0.92}{\sqrt{1/10}}) = 0.004.$$

The p-value is very small, so there is strong evidence against H_0.

4.3.2 Kendall's rank correlation (r_K) (sometimes called Kendall's τ)

A similar correlation statistic was proposed by Kendall. This is defined as the number of pairs $i < j$ such that $X_i < X_j$ and $Y_i < Y_j$. If the sample is ordered so that $X_1 < X_2 < \cdots < X_n$, then

$$K := \sum_{i=1}^{n-1} m_i$$

where m_i is the number of $j > i$ such that $Y_j > Y_i$ (again assuming there are no ties in each sample), which may be more convenient for calculation of K.

Null distribution: The range of K is from 0 to $\frac{n(n-1)}{2}$ and its null distribution is symmetric about $\frac{n(n-1)}{4}$; one can show that the variance is $\frac{n(n-1)(2n+5)}{72}$.

We can also define an associated correlation coefficient

$$r_K = \frac{4K}{n(n-1)} - 1$$

Then the range of r_K is from -1 to 1, and null distribution is symmetric about 0 and has variance $\frac{2(2n+5)}{9n(n-1)}$.

Example 24 (b) Consider the previous data set again, and find the Kendall's correlation coefficient.

Answer: If we reorder the sample so that $X_1 < X_2 < \cdots$ and let R_i be the rank of Y_i and let m_i be the number of $j > i$ such that $Y_j > Y_i$ (or $R_j > R_i$) as above, then we get the following table:

i	1	2	3	4	5	6	7	8	9	10	11
$X_{(i)}$	21	25	29	32	33	37	45	47	53	61	81
$Y_{(i)}$	31	29	34	51	47	48	53	49	62	67	86
$R_{(i)}$	2	1	3	7	4	5	8	6	9	10	11
m_i	9	9	8	4	6	5	3	3	2	1	

Then $K = \sum m_i = 9+9+8+4+6+5+3+3+2+1 = 50$ and $r_K = \frac{4K}{n(n-1)} - 1 = \frac{200}{11\times 10} - 1 = 0.82$. We can do a test to check whether this correlation is significant. We have $R_K \overset{\cdot}{\sim} N(0, \frac{2(2N+5)}{9N(N-1)})$ under H_0, and thus we can find the (approximate) p-value:

$$2 \times P(R_K > 0.82) = 2 \times P(Z > \frac{0.82}{\sqrt{0.054}}) = 2 \times (1 - P(Z < 3.53)) = 0.0004$$

Again the p-value is very small, so there is strong evidence against H_0.

Using R

To find these correlation coefficients and test whether the two variables are correlated, the function `cor.test` can be used.

```
> test1 <- c(45,61,33,29,21,47,53,32,37,25,81)
> test2 <- c(53,67,47,34,31,49,62,51,48,29,86)

> cor.test(test1, test2, method = "spearman")

Spearman's rank correlation rho
data:  test1 and test2
S = 18, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9181818

> cor.test(test1, test2, method = "kendall")

   Kendall's rank correlation tau
data:  test1 and test2
T = 50, p-value = 0.0001323
alternative hypothesis: true tau is not equal to 0
sample estimates:
     tau
0.8181818
```

4.4 U-statistic permutation (USP) tests for independence

In this last section, we return to testing for independence in a two-way contingency table. The method presented here, which was very recently proposed (Berrett et al.,2020[1]; Berrett and Samworth 2021[2]), is applicable in a wide range of settings, but we will specialise to two-way contingency tables. Recall the chi-squared test for independence introduced in Section 2.3: that test was based on a chi-squared approximation to the distribution of the test statistic – this will be fine in some cases, but it might happen that the approximation is not good and we may even end up with a test that does not have valid size. Perhaps more importantly, the chi-squared test is not appropriate when if the expected cell counts are small (< 5), or if there are no observations in a particular row or column (we can get around this by ignoring any such row, but as a result the test will not be sufficient to verify independence).

The USP test overcomes both of these drawbacks by using permutations of the data to find a critical value or p-value. The main idea is that after permuting the data, we "break" the dependence between the two samples. Suppose we observe $(X_1, Y_1), \ldots, (X_n, Y_n)$, where X_i takes values in $\{1, \ldots, R\}$ and Y_i takes values in $\{1, \ldots, C\}$ (it is sufficient for us to observe the contingency table here, but let's suppose we observe the pairs for simplicity.). The chi-squared test statistic used in Chapter 2 was

$$X^2 = \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{(O_{rc} - E_{rc})^2}{E_{rc}},$$

where the expected values $E_{rc} = \frac{O_{r\cdot} O_{\cdot c}}{n}$ were calculated by assuming independence.

The USP test uses a similar test statistic (the motivation for which is based on calculating the χ^2-divergence between two probability distributions), given by

$$U := \frac{1}{n(n-3)} \sum_{r=1}^{R} \sum_{c=1}^{C} (O_{rc} - E_{rc})^2 - \frac{4}{n(n-2)(n-3)} \sum_{r=1}^{R} \sum_{c=1}^{C} O_{rc} E_{rc}.$$

Large values of U provide evidence against H_0 (that X and Y are independent). We calculate a p-value or critical value using permutations – specifically, we keep the x order fixed and permute the y observations – there are $n!$ possible permutations, if n is small we can calculate them all, and if n is large, we can use a randomisation approach.

The key benefit to this approach is that the test will have exact size α (we don't need to rely on a chi-squared approximation). There are also scenarios in which the USP test has much higher power than the chi-squared test!

Example 25 The following table contains data on the education level and marital status of 300 people.

Observed counts	Middle School	High School	Bachelor's	Master's	PhD
Never Married	18	36	21	9	6
Married	12	36	45	36	21
Divorced	6	9	9	3	3
Widowed	3	9	9	6	3

We can calculate the corresponding expected counts under H_0 (rows and columns are independent).

[1] https://arxiv.org/abs/2001.05513
[2] https://arxiv.org/abs/2101.10880

Expected counts	Middle School	High School	Bachelor's	Master's	PhD
Never Married	11.7	27	25.2	16.2	9.9
Married	19.5	45	42	27	16.5
Divorced	3.9	9	8.4	5.4	3.3
Widowed	3.9	9	8.4	5.4	3.3

Here we see that many of the expected counts under H_0 are less than 5, and the chi-squared approximation to the distribution of X^2 will be inaccurate! Here the observed test statistic is

$$U = \frac{1}{n(n-3)} \sum_{r=1}^{4} \sum_{c=1}^{5} (O_{rc} - E_{rc})^2 - \frac{4}{n(n-2)(n-3)} \sum_{r=1}^{4} \sum_{c=1}^{5} O_{rc} E_{rc} = 0.0041.$$

and using a randomisation approach in R we get a p-value of 0.002 – there is strong evidence against H_0.

Using R

```
Data = matrix(c(18,12,6,3,36,36,9,9,21,45,9,9,9,36,3,6,6,21,3,3),4,5)

Data_y <- rep(1:ncol(Data),colSums(Data))
Data_x <- c(rep(1:nrow(Data),Data[,1]),rep(1:nrow(Data),Data[,2])
            ,rep(1:nrow(Data),Data[,3]),rep(1:nrow(Data),Data[,4])
            ,rep(1:nrow(Data),Data[,5]))
#sample size
n <- sum(Data)
#column and row sums
OC <- colSums(Data)
OR <- rowSums(Data)
#expected counts under H0
E <- outer(OR, OC)/n

#observed test statistic
U <- 1/(n*(n-3))*sum((Data - E)^2) - 4/(n*(n-2)*(n-3))*sum(Data*E)

B <- 10000
Ub <- NULL
for (b in 1:B)
{
  s <- sample(1:n, n, replace = F)

  Ob <- table(Data_x, Data_y[s])
  OCb <- colSums(Datab)
  ORb <- rowSums(Datab)
  Eb <- outer(ORb, OCb)/n

  Ub <- c(Ub,
   1/(n*(n-3))*sum((Ob - Eb)^2) - 4/(n*(n-2)*(n-3))*sum(Ob*Eb))
}
```

```
#add observed value
Ub <- c(Ub, U)

plot(hist(Ub))

#calculate p-value
pvalue <- mean(Ub >= U)

#we can try to do Pearson's chi-squared test
chisq.test(Data)
```