Team Project: Great Barrier Reef Tourism Operators

Statistical Decision Making, Autumn 2019

Sample Solution

0 Setting up

We load the data and define the number of bootstrap or randomisation samples for use in the report.

```
T13 = read.csv("TourismOperators2013.csv")
T17 = read.csv("TourismOperators2017.csv")
head(T13)
```

```
##   Years Optimistic QualityOfLife WantToProtect ClimateChangeView
## 1    20          6             8             9                 1
## 2    44          5            10             5                 1
## 3    15          8            10            10                 1
## 4    38          6             6             9                 1
## 5    14         10            10            10                 1
## 6    20          8            10             9                 3
```

```
head(T17)
```

```
##   Years Optimistic QualityOfLife WantToProtect ClimateChangeView
## 1     4          7             6            10                 5
## 2     3          1            10            10                 1
## 3    20         10            10             1                 3
## 4     8          8             7             6                 1
## 5    17          8            10            10                 2
## 6    19          1            10            10                 1
```

```
B = 100000  # number of bootstrap or randomisation samples; the more, the better
```

1 Number of years in GBR tourism

(a) Distribution

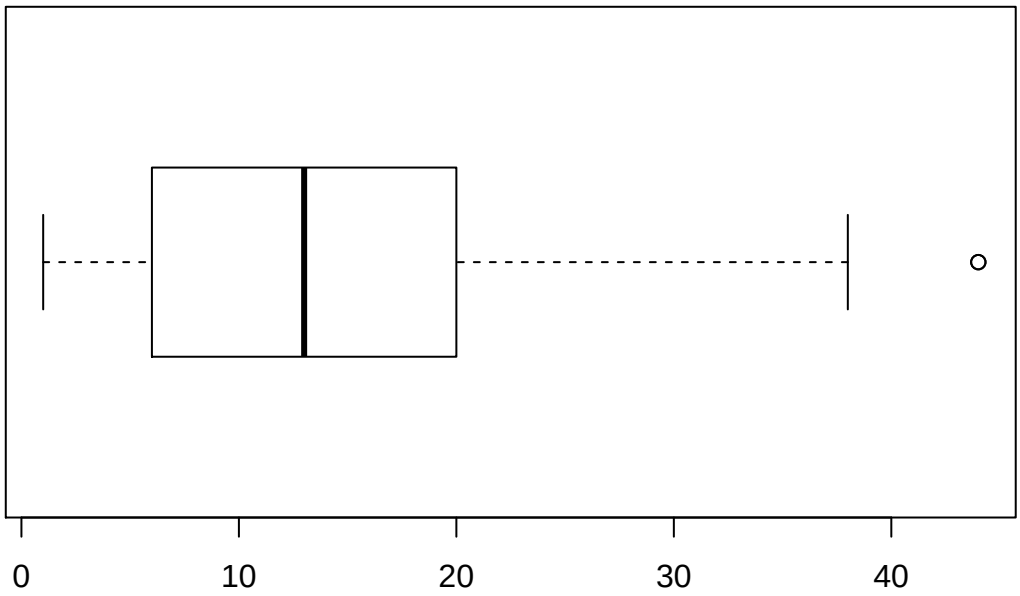
Useful plots for getting an idea of the distribution of the variable `T13$Years` are a histogram and a boxplot; we can also look at the five number summary.

```
summary(T13$Years)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    6.00   13.00   14.43   20.00   44.00
```

```
hist(T13$Years)
```


```
boxplot(T13$Years, horizontal=TRUE)
```



Both plots show a right-skewed distribution. The fact that the mean of the data is larger than the median also indicates a right-skewed distribution.

[**Note for markers:** Any reasonable justification is fine.]

(b) 99% confidence interval from bootstrapping

Before constructing the bootstrap distribution, it is wise to check whether we need to worry about missing values.

```
mean(T13$Years)  # this returns NA if there are missing values
```

```
## [1] 14.42857
```

So there are no missing values. We construct a bootstrap distribution by repeatedly sampling with replacement from the given data, preserving the sample size, and check the histogram of the constructed bootstrap distribution.

```
boot.dist = rep(NA, B)
for (i in 1:B) {
  boot.dist[i] = mean(sample(T13$Years, replace=TRUE))  # no missing values
}
hist(boot.dist, breaks=30)
```

The bootstrap distribution is smooth and close to bell-shaped, and it is centred at the sample mean of the data as it should be. A 99% confidence interval for the population mean contains the middle 99% of the bootstrap sample means, so its boundaries are the 0.005-quantile and the 0.995-quantile of the bootstrap distribution.

```
quantile(boot.dist, c(0.005,0.995))
```

```
##     0.5%    99.5%
## 12.21849 16.75630
```

So the obtained 99% confidence interval for the mean number of years of experience tourism operators on the GBR had in 2013 is [12.22,16.76].

(c) 99% confidence interval using the Central Limit Theorem

In order to compute a 99% confidence interval using the Central Limit Theorem, we need the sample mean $\bar{x}$, the sample standard deviation s, and quantiles of the t-distribution with $n-1$ degrees of freedom, where n is the sample size.

```
xbar = mean(T13$Years)
s = sd(T13$Years)
n = length(T13$Years)
```

For a 99% confidence interval, the quantiles to use are again the 0.005-quantile and the 0.995-quantile; as t-distributions are symmetric, the former quantile is just the negative of the latter.

```
tstar = qt(0.995, df=n-1)
tstar
```

```
## [1] 2.618137
```

```
qt(0.005, df=n-1)
```

```
## [1] -2.618137
```

We can now compute the confidence interval:

```
xbar + c(-1,1)*tstar*s/sqrt(n)
```

```
## [1] 12.10762 16.74953
```

So the obtained 99% confidence interval for the mean number of years experience tourism operators on the GBR had in 2013 is [12.11,16.75]. This is roughly consistent with the confidence interval computed using bootstrapping, although the lower boundaries are slightly different.

2 Proportion of GBR optimists

(a) Proportion of GBR optimists in 2013

(i) Hypothesis test and p-value using `prop.test`

Denoting by p_{13} the (population) proportion of tourism operators on the GBR who were optimistic about the future of the GBR in 2013, the hypothesis test is

$$\begin{aligned} H_0: &\quad p_{13} = 0.7 \\ H_a: &\quad p_{13} < 0.7 \end{aligned}$$

We first should check whether we need to worry about missing values.

```
mean(T13$Optimistic >= 6)  # this returns NA if there are missing values
```

```
## [1] NA
```

We do! The best course of action is to ignore all respondents who didn't answer that question and count, separately, those that gave an optimistic view (level of agreement ≥ 6) and those that gave a pessimistic view (level of agreement < 6).

```
n_opt13 = sum(T13$Optimistic >= 6, na.rm=TRUE)
n_pes13 = sum(T13$Optimistic < 6, na.rm=TRUE)
n_tot13 = n_opt13 + n_pes13
n_opt13
```

```
## [1] 72
```

```
n_pes13
```

```
## [1] 45
```

```
n_tot13
```

```
## [1] 117
```

Now we just have to feed this information into `prop.test`.

```
prop.test(n_opt13, n_tot13, p=0.7, alternative="less")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  n_opt13 out of n_tot13, null probability 0.7
## X-squared = 3.5963, df = 1, p-value = 0.02895
## alternative hypothesis: true p is less than 0.7
## 95 percent confidence interval:
##  0.0000000 0.6899887
## sample estimates:
##         p
## 0.6153846
```

So the computed p-value of the data is 0.029.

(ii) p-value from randomisation

We first calculate the sample proportion of optimistic responses in the 2013 survey:

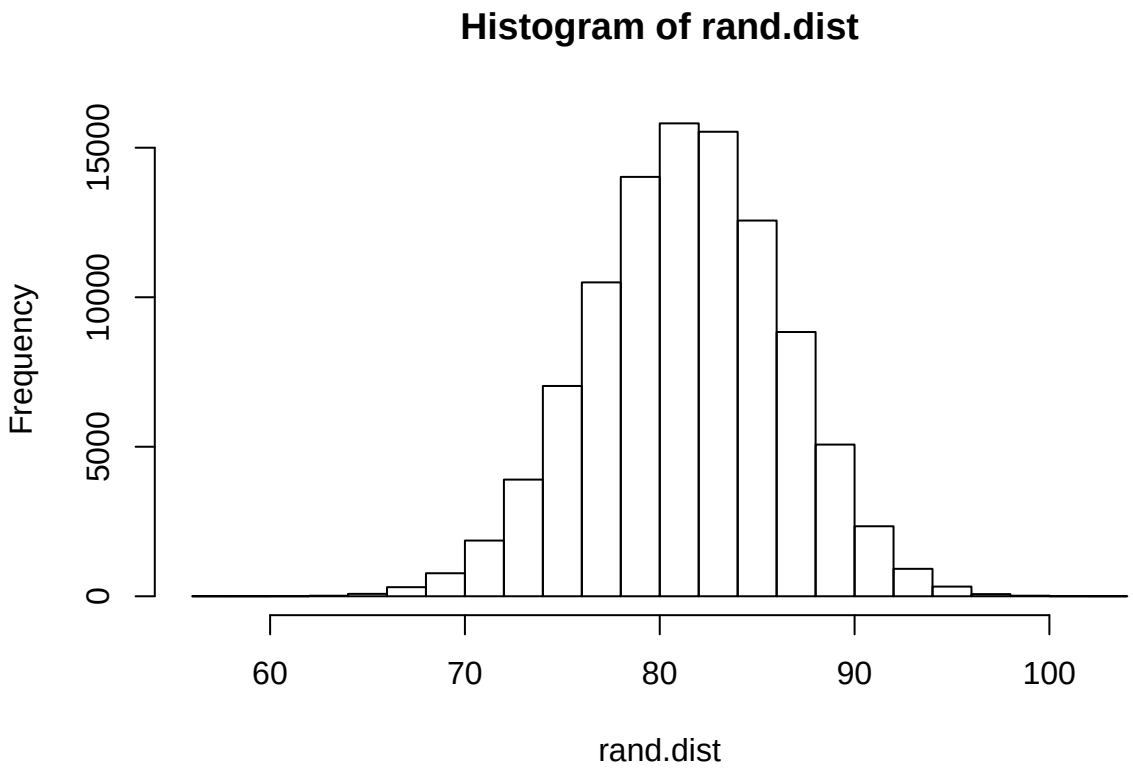
```
phat_13 = n_opt13 / n_tot13
phat_13
```

```
## [1] 0.6153846
```

Under the assumption of the null hypothesis $p_{13} = 0.7$, the number of optimistic responses in the 2013 survey is given by the binomial distribution with sample size `n_tot13` = 117 and success probability 0.7, so we can use this distribution to construct a randomisation distribution.

We should look at a histogram of the randomisation distribution to check that it is centred at the value `0.7*n_tot13` = 81.9 as it should be.

```
rand.dist = rbinom(B, n_tot13, p=0.7)
hist(rand.dist, breaks=30)
```



This looks good. As the observed sample proportion is less than 0.7 and the alternative hypothesis is one-sided, the p-value of the data is the proportion of randomisation samples that have at most `n_opt13` successes.

```
mean(rand.dist <= n_opt13)
```

```
## [1] 0.0304
```

So the computed p-value of the data is 0.030.

(iii) Comparison and conclusion of test

The p-value computed using `prop.test` is slightly smaller than the p-value obtained by randomisation. The binomial distribution describes the sampling distribution under the assumption of the null hypothesis *exactly*, hence the latter is more reliable, provided that the number of randomisation samples is sufficiently large.

In any case, both p-values are quite a bit smaller than the significance level of 5%, so the data provide enough evidence to reject the null hypothesis at this significance level: At a significance level of 5%, there is evidence that the proportion of tourism operators who were generally optimistic about the future of the GBR in 2013 was less than 70%.

(b) Proportions of GBR optimists in 2013 vs. 2017

(i) Hypothesis test and p-value using `prop.test`

We keep the notation from part (a). Denoting by p_{17} the (population) proportion of tourism operators on the GBR who were optimistic about the future of the GBR in 2017, the hypothesis test is

$$\begin{aligned} H_0: &\quad p_{13} = p_{17} \\ H_a: &\quad p_{13} \neq p_{17} \end{aligned}$$

We again ignore all respondents who didn't answer that question and count, separately, those that gave an optimistic view (level of agreement ≥ 6) and those that gave a pessimistic view (level of agreement < 6).

```
n_opt17 = sum(T17$Optimistic >= 6, na.rm=TRUE)
n_pes17 = sum(T17$Optimistic < 6, na.rm=TRUE)
n_tot17 = n_opt17 + n_pes17
n_opt17
```

```
## [1] 55
```

```
n_pes17
```

```
## [1] 38
```

```
n_tot17
```

```
## [1] 93
```

Now we just have to feed the information into `prop.test`.

```
prop.test(c(n_opt13,n_opt17), c(n_tot13,n_tot17))
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(n_opt13, n_opt17) out of c(n_tot13, n_tot17)
## X-squared = 0.044557, df = 1, p-value = 0.8328
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1189019  0.1668754
## sample estimates:
##    prop 1    prop 2
## 0.6153846 0.5913978
```

So the computed p-value of the data is 0.833.

(ii) p-value from randomisation

We first calculate the sample proportion of optimistic responses in the 2017 survey and the difference of the proportions of optimistic responses in the 2013 survey and in the 2017 survey.

```
phat_17 = n_opt17 / n_tot17
phat_17
```

```
## [1] 0.5913978
```

```
dphat = phat_17 - phat_13
dphat
```

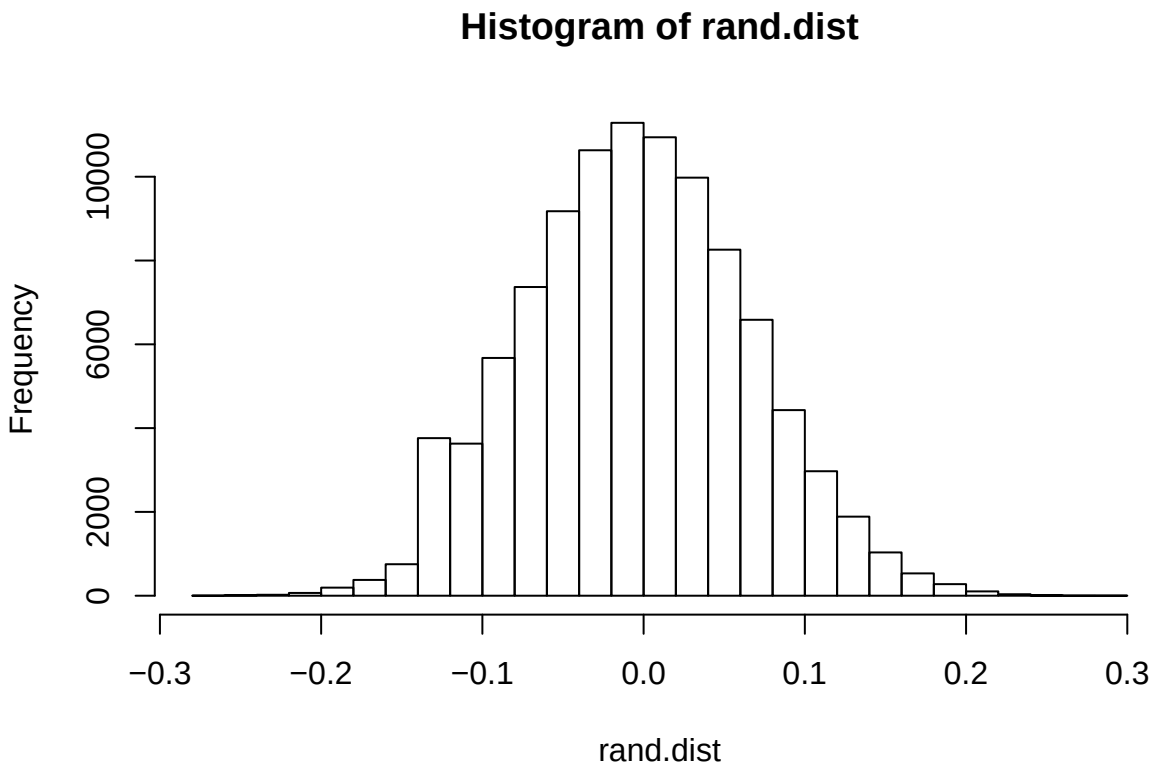
```
## [1] -0.02398677
```

The null hypothesis $p_{13} = p_{17}$ says that there is no difference between 2013 and 2017 as far as optimism about the future of the GBR is concerned. Therefore, we can enforce the null hypothesis by constructing a randomisation distribution as follows:

- Merge the 2013 responses and the 2017 responses to this question.
- Repeatedly
  - allocate all responses to fake-2013 and fake-2017 groups at random, preserving the group sizes; and
  - compute the difference of the proportions of optimists for the fake-2013 and the fake-2017 groups.

Once we have computed the randomisation distribution, we should look at a histogram to check that it is smooth and centred at the value 0 as it should be.

```
n = n_tot13 + n_tot17  # total number of responses
responses = c(rep(TRUE,n_opt13+n_opt17), rep(FALSE,n_pes13+n_pes17))  # all responses
rand.dist = rep(NA, B)
for (i in 1:B) {
  idx_13 = sample(1:n, n_tot13)  # select responses for a fake 2013 group at random...
  fake_13 = responses[idx_13]    # ...put them in the fake 2013 group...
  fake_17 = responses[-idx_13]   # ...and all others in the fake 2017 group
  rand.dist[i] = sum(fake_17)/n_tot17 - sum(fake_13)/n_tot13
}
hist(rand.dist, breaks=30)
```



This looks good. As the observed difference in the sample proportion of optimistic responses is negative and the alternative hypothesis is two-sided, the p-value of the data is twice the proportion of randomisation samples that have a difference in sample proportions of optimistic responses less than or equal to `dphat = -0.0239868`.

```
2*mean(rand.dist <= dphat)
```

```
## [1] 0.83334
```

So the computed p-value of the data is 0.833.

(iii) Comparison and conclusion of test

Both p-values are very similar and much larger than the significance level of 5%, so the data do not provide enough evidence to reject the null hypothesis at this (or indeed at any reasonable) significance level: There is no evidence that the proportions of tourism operators who were generally optimistic about the future of the GBR in 2013 and 2017 were different.

3 Contribution to quality of life vs. willingness to protect the GBR

(a) Hypothesis test

Denoting by ρ the (population) correlation between the 2013 levels of tourism operators' agreement with the statements "The GBR contributes to my quality of life." and "I would like to do more to help protect the GBR.", the hypothesis test is

$$H_0: \quad \rho = 0$$
$$H_a: \quad \rho > 0$$

(b) p-value of the data

We first calculate the sample correlation.

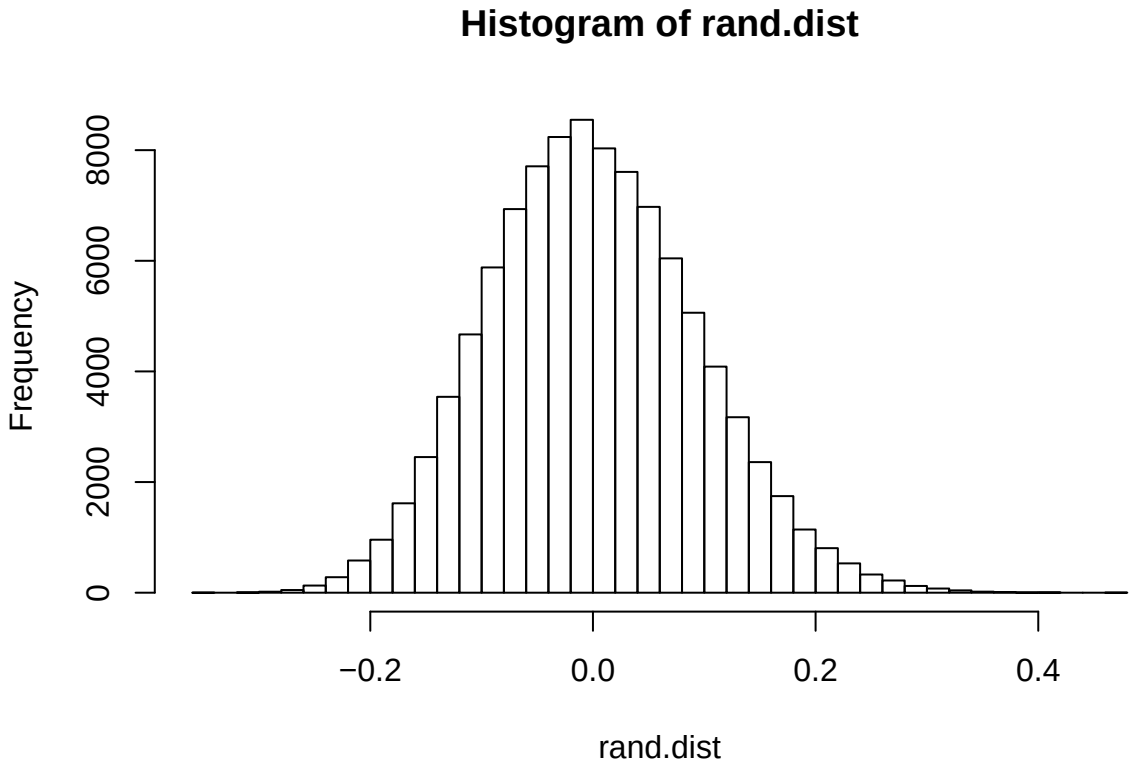
```
r = cor(T13$QualityOfLife, T13$WantToProtect, use="complete.obs")  # missing values!
r
```

```
## [1] 0.08340365
```

The null hypothesis $\rho = 0$ says that there is no linear association between the variables `T13$QualityOfLife` and `T13$WantToProtect`. We can enforce the null hypothesis by repeatedly

- permuting the values of one of the variables at random; and
- computing the sample correlation of the resulting pairs.

```
rand.dist = rep(NA, B)
for (i in 1:B) {
  rand.dist[i] = cor(sample(T13$QualityOfLife), T13$WantToProtect, use="complete.obs")
}
hist(rand.dist, breaks=30)  # always check for possible trouble
```



The randomisation distribution is smooth and centred at the value 0, as it should be.

As the observed sample correlation is positive and the alternative hypothesis is one-sided, the p-value of the data is the proportion of randomisation samples that have a sample correlation greater than or equal to `r` = 0.0834037.

```
mean(rand.dist >= r)
```

```
## [1] 0.18746
```

So the computed p-value of the data is 0.1875.

NOTE: The p-value reported by the function `lm` is **not** the correct p-value for the test from part (a): the p-value reported by `lm` is for a test for non-zero correlation, that is, for a two-sided test; it is approximately by a factor of 2 larger than the p-value for the test from part (a).

```
summary(lm(T13$WantToProtect~T13$QualityOfLife))
```

```
##
## Call:
## lm(formula = T13$WantToProtect ~ T13$QualityOfLife)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1644 -1.1425  0.0107  1.8356  2.6234
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.28901    0.87370   8.343 2.11e-13 ***
## T13$QualityOfLife  0.08754    0.09883   0.886    0.378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.056 on 112 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.006956,   Adjusted R-squared:  -0.00191
## F-statistic: 0.7845 on 1 and 112 DF,  p-value: 0.3777
```

```
2*mean(rand.dist >= r)
```

```
## [1] 0.37492
```

[**Note for markers:** At most 50% of the marks if the p-value reported by the function `lm` is used without comment; full marks if the p-value reported by the function `lm` is scaled by 0.5 with justification.]

(c) Conclusion of the test

The p-value is quite a bit larger than the significance level of 10%, so the data do not provide enough evidence to reject the null hypothesis at this significance level: There is no evidence that the 2013 levels of tourism operators' agreement with the statements "The GBR contributes to my quality of life." and "I would like to do more to help protect the GBR." are positively correlated.

4 Changes in views on climate

(a) Goodness of fit test

We're taking the 2013 proportions as fixed reference and test whether the 2017 counts differ significantly from these proportions.

```
count_13 = table(T13$ClimateChangeView)
prop_13 = count_13/sum(count_13)
prop_13
```

```
##
##          1          2          3          4          5
## 0.50000000 0.17796610 0.22881356 0.04237288 0.05084746
```

```
count_17 = table(T17$ClimateChangeView)
chisq.test(count_17, p=prop_13)
```

```
## Warning in chisq.test(count_17, p = prop_13): Chi-squared approximation may
## be incorrect
```

```
##
##  Chi-squared test for given probabilities
##
## data:  count_17
## X-squared = 12.624, df = 4, p-value = 0.01327
```

The warning indicates that some of the expected counts are too small to trust p-values computed by using a χ^2-distribution, so we should use randomisation; this can be done either by calling `chisq.test` with the appropriate parameters, or by using a `for`-loop.

Using `chisq.test` with randomisation

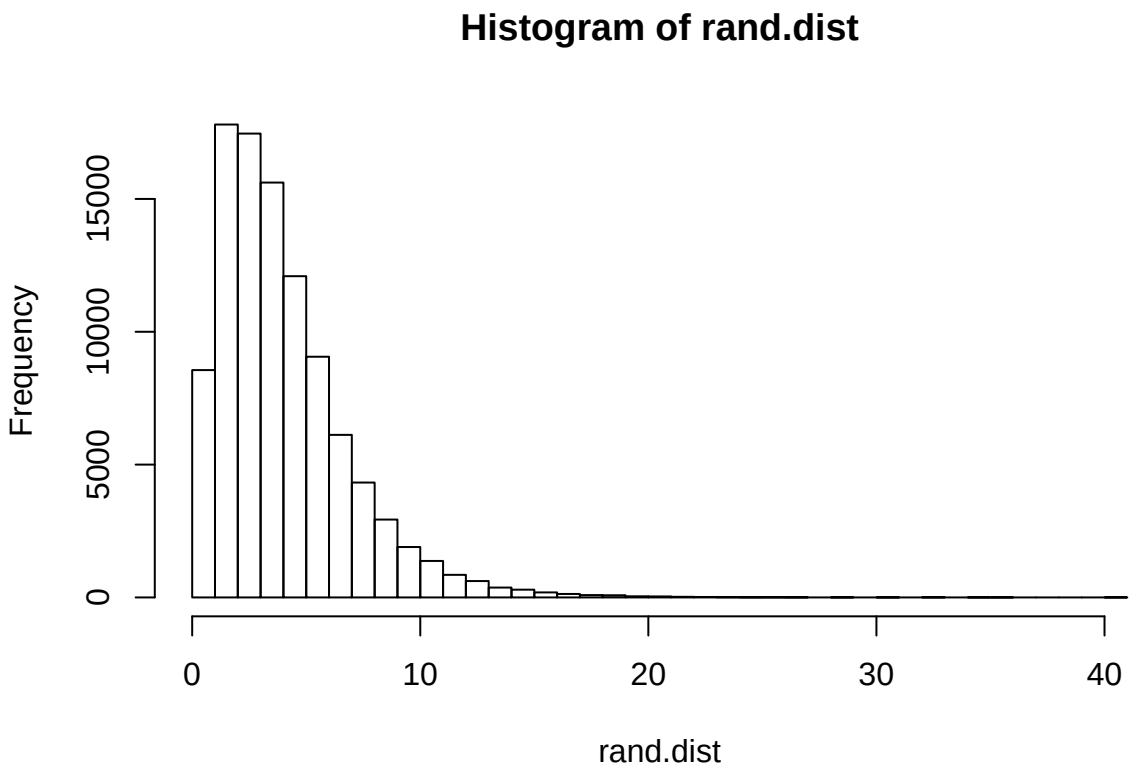
```
chisq.test(count_17, p=prop_13, simulate.p.value=TRUE, B=B)
```

```
##
##  Chi-squared test for given probabilities with simulated p-value
##  (based on 1e+05 replicates)
##
## data:  count_17
## X-squared = 12.624, df = NA, p-value = 0.01447
```

Using a `for`-loop

To construct the randomisation distribution, we repeatedly create a sample of responses at random, using the 2013 proportions and the 2017 sample size, and compute the test statistic of the constructed sample.

```
e = prop_13*sum(count_17)  # expected counts
rand.dist = rep(NA, B)
for (i in 1:B) {
  s = sample(1:5, sum(count_17), replace=TRUE, prob=prop_13)
  t = table(factor(s,1:5))  # use 'factor' to make sure that counts of 0 are not removed
  rand.dist[i] = sum((t - e)^2/e)
}
hist(rand.dist, breaks=30)  # always check for possible trouble
```



The randomisation distribution is sufficiently smooth to be trusted. The p-value is the proportion of randomisation samples that have a test statistic at least as large as the 2017 data.

```
pval = mean(rand.dist >= sum((count_17 - e)^2/e))
pval
```

```
## [1] 0.01514
```

Both estimates for the p-value obtained by simulation are around 0.015. (Note that the p-value computed by using a χ^2-distribution seems to be a bit too small.)

[**Note for markers:** At most 50% of the marks if the p-value reported by the function `chisq.test` without simulation is used.]

(b) Test for independence

We're analysing two categorical variables: the date, taking 2 values; and the views on climate change, taking 5 values. So we have to construct a 2-by-5 (or a 5-by-2) matrix and test for independence of the variables.

```
responses = na.omit(c(T13$ClimateChangeView, T17$ClimateChangeView))  # all responses...
years = c(rep(2013,sum(count_13)), rep(2017,sum(count_17)))  # ...and the respective years
M = table(years, responses)
M
```

```
##       responses
## years   1  2  3  4  5
##   2013 59 21 27  5  6
##   2017 58  4 22  4  4
```

```
chisq.test(M)
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  M
## X-squared = 9.5167, df = 4, p-value = 0.04941
```

Again, the warning indicates that some of the expected counts are too small to trust p-values computed by using a χ^2-distribution, so we should use randomisation; this can be done either by calling `chisq.test` with the appropriate parameters, or by using a `for`-loop.

Using `chisq.test` with randomisation

```
chisq.test(M, simulate.p.value=TRUE, B=B)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 1e+05
##  replicates)
##
## data:  M
## X-squared = 9.5167, df = NA, p-value = 0.04754
```

Using a `for`-loop

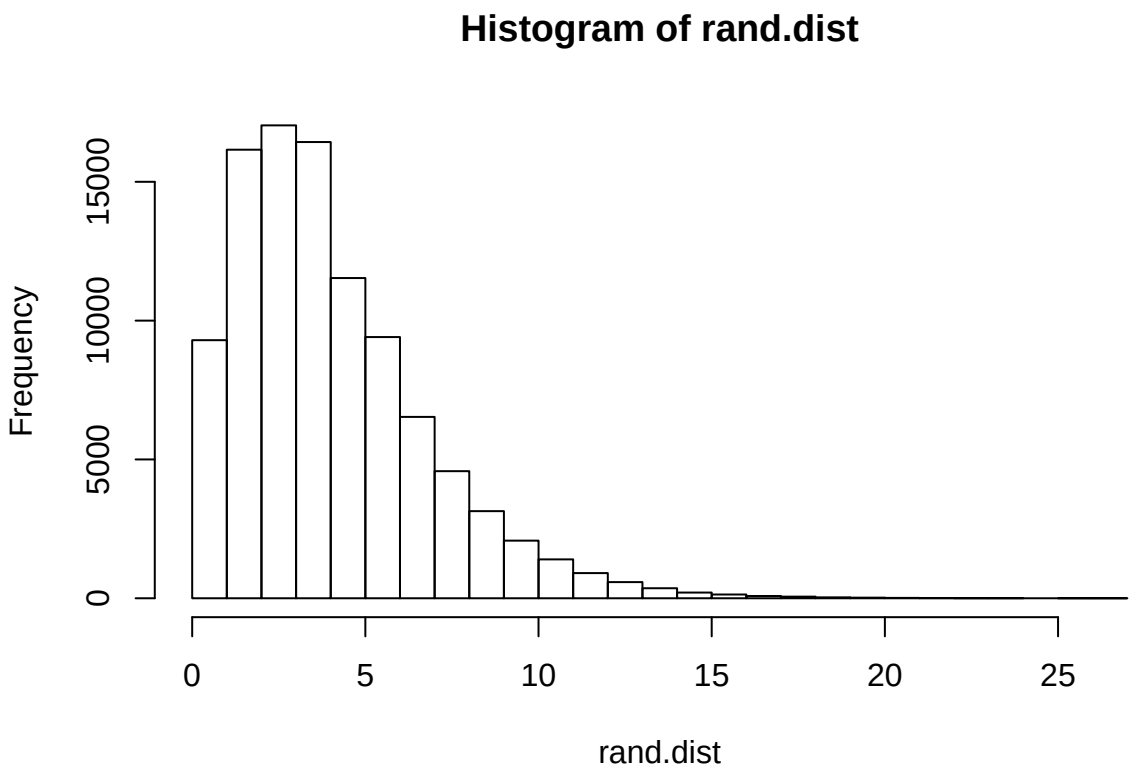
We compute the expected frequencies under the assumption of independence by multiplying the row and column totals and dividing by the number of cases:

```
n = sum(M)
E = rowSums(M) %o% colSums(M) / n
print(E)
```

```
##             1        2        3        4        5
## 2013 65.74286 14.04762 27.53333 5.057143 5.619048
## 2017 51.25714 10.95238 21.46667 3.942857 4.380952
```

To construct the randomisation distribution, we repeatedly assign all responses to 2013 or 2017 at random, preserving the row and column totals, and compute the test statistic of the constructed sample.

```
rand.dist = rep(NA, B)
for (i in 1:B) {
  T = table(sample(years), responses)
  rand.dist[i] = sum((T - E)^2/E)
}
hist(rand.dist, breaks=30)  # always check for possible trouble
```



The randomisation distribution is sufficiently smooth to be trusted. The p-value is the proportion of randomisation samples that have a test statistic at least as large as the actual data.

```
pval = mean(rand.dist >= sum((M - E)^2/E))
pval
```

```
## [1] 0.0472
```

Both estimates for the p-value obtained by simulation are around 0.047. (Note that the p-value computed by using a χ^2-distribution seems to be a bit too large.)

[**Note for markers:** At most 50% of the marks if the p-value reported by the function `chisq.test` without simulation is used.]

(c) Comparison

The p-values from parts (a) and (b) are quite different: The p-value from part (b) is about three times the p-value from part (a). – There is clearly a substantial difference between these tests.

For the analysis in part (a), we assume that we know the proportions in 2013 exactly, and only the data collected in 2017 are subject to statistical uncertainty; for the analysis in part (b), we treat both the data collected in 2013 and the data collected in 2017 as subject to statistical uncertainty. – It is clear that the analysis in part (b) will give a larger p-value (as there is more uncertainty) and that this is what should be used here: The 2013 data were collected in the same way as the 2017 data, so they add an equal amount of uncertainty.

[**Note for markers:** noting that p-values are substantially different: 0.5 marks; sensible explanation: 0.5 marks]