Introduction to Statistics Part 2

Natasha Stanbridge, Tiangang Cui, Jaslene Lin

Table of contents

8 Unknown Proportions **4**
8.1 A review of box models and the Central Limit Theorem . . . 4
8.1.1 Random samples from box models . . . 4
8.1.2 Important special case: 0-1 Box . . . 4
8.2 Prediction Interval for the 0-1 box . . . 5
8.2.1 Example . . . 5
8.2.2 Alternative exampe: $p = 0.2$ . . . 7
8.2.3 Simulations . . . 8
8.2.4 Main take-aways . . . 10
8.2.5 Prediction Interval for the sample proportion $\bar{X}$ . . . 10
8.2.6 General prediction intervals under the Central Limit Theorem . . . 11
8.3 Inference for an unknown proportion: Confidence Interval . . . 11
8.3.1 What if the box proportion p is unknown? . . . 11
8.3.2 Example . . . 12
8.3.3 Confidence interval for p . . . 12
8.3.4 Properties of the Wilson confidence interval . . . 14
8.3.5 Different confidence levels . . . 14
8.3.6 Case study . . . 16
8.4 Inference for an unknown proportion: Hypothesis Test . . . 18
8.4.1 Elements of a hypothesis test . . . 18
8.5 A review of inference for unknown proportions . . . 22
8.5.1 HATPC framework . . . 22
8.5.2 Summary of Z-test procedure using HATPC . . . 22
8.5.3 Not rejecting H_0, but not accepting it either . . . 25
8.5.4 Different confidence/significance levels . . . 25
8.6 One-sided tests . . . 26
8.6.1 Case study: production lines . . . 26
8.6.2 Apply the HATPC framework . . . 26
8.6.3 Practical note: specify conclusions before seeing the data . . . 28

8.6.4 Summary . . . 29

9 Unknown Means **30**
9.1 General Z-test . . . 30
9.1.1 Case study: standardised school exams . . . 30
9.1.2 Formal hypothesis test . . . 31
9.1.3 Confidence interval . . . 32
9.2 Estimating the standard error . . . 33
9.2.1 Sample SD in depth (not for assessment) . . . 34
9.3 The T-statistic . . . 36
9.3.1 Simulations . . . 36
9.3.2 Fatter tails . . . 37

10 Unknown Means: T-test **39**
10.1 The T-statistic . . . 39
10.2 Student's t-distribution . . . 39
6-sided die example . . . 40
10.3 One-sample T-test . . . 42
10.3.1 HATPC . . . 43
10.3.2 The function `t.test()` . . . 45
10.3.3 Confidence Interval . . . 46
10.4 Bootstrap simulation . . . 47
10.4.1 Skewed example . . . 47
10.4.2 Bootstrap principle . . . 48
10.4.3 Skewed box example . . . 48
10.4.4 "Equal-tailed" confidence interval . . . 51
10.5 A review of tests for unknown mean with an example . . . 51
10.5.1 Z-test . . . 52
10.5.2 T-test using Student's t . . . 52
10.5.3 T-test using bootstrap simulation . . . 53
10.5.4 Simulation-based confidence interval . . . 54
10.6 Equal-tailed $(1-\alpha)$ confidence interval for unknown mean . . . 55

11 Unknown Means: Two-Sample T-test **57**
11.1 Comparing two (sample) means . . . 57
11.1.1 Case study: red bull . . . 57
11.1.2 Two-box model . . . 58
11.1.3 Two-sample Test Statistics . . . 58
11.2 The Classical Two-Sample T-test . . . 59
11.2.1 The pooled estimate $\hat{\sigma}$ . . . 59
11.2.2 Red Bull example . . . 60
11.3 The Welch Test . . . 63
11.3.1 Relaxing the equal variance assumption . . . 63

11.3.2 Default two-sample `t.test()` . . . 63
11.4 Bootstrap simulation . . . 64
11.4.1 Simulate the Welch statistic . . . 64
11.4.2 Two-sided P-value by simulation . . . 65
11.4.3 Confidence interval by simulation . . . 66
11.5 Paired (two-sample) T-test . . . 66
11.5.1 Student's sleep data . . . 66
11.5.2 Paired T-test . . . 69

12 Chi-squared tests **70**
12.1 Chi-squared test of goodness of fit . . . 70
12.1.1 Suspicious dice . . . 70
12.1.2 Box model for (possibly loaded) die . . . 71
12.1.3 Goodness of fit test . . . 71
12.1.4 General formulation of Pearson's χ^2 statistic . . . 72
12.1.5 The χ^2_d distribution . . . 73
12.1.6 Our dice example . . . 73
12.1.7 Assumptions required . . . 75
12.1.8 Equivalence with z-test . . . 76
12.2 Simulation (test of goodness of fit) . . . 77
12.2.1 Using simulation: the dice example . . . 77
12.2.2 Small expected frequencies . . . 79
12.2.3 Using simulation . . . 80
12.3 Chi-squared tests with estimated parameters . . . 82
12.4 Chi-squared test of independence . . . 83
12.4.1 Pearson's statistic . . . 83
12.4.2 Full model . . . 84
12.4.3 Null hypothesis . . . 84
12.4.4 Estimate marginal probabilities $p_{i\bullet}$s **and** $p_{\bullet j}$s . . . 84
12.4.5 Expected frequencies . . . 85
12.4.6 Degrees of freedom . . . 85
12.4.7 Handedness example . . . 85
12.4.8 Using simulation . . . 88
12.4.9 Standardised residuals (not for assessment) . . . 88

8 Unknown Proportions

8.1 A review of box models and the Central Limit Theorem

8.1.1 Random samples from box models

We have seen that if $X_1, \ldots, X_n$ represent the numbers obtained in a random sample taken *with replacement* from a box, we have a good understanding of the **random behaviour** of both

- The sum $S = X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i$ *and*
- The average $\bar{X} = \frac{X_1 + \cdots + X_n}{n} = S/n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

As long as n is "large enough", if we know *both* the mean μ and the SD σ of the numbers in the box, then

- The box of all possible **sample sums** has mean $E(S) = n\mu$, SD $SE(S) = \sigma\sqrt{n}$ and has a **normal shape**;
- The box of all possible **sample means** has mean $E(\bar{X}) = \mu$, SD $SE(\bar{X}) = \sigma/\sqrt{n}$ and has a **normal shape**.

8.1.2 Important special case: 0-1 Box

An important example is where the box only contains $\boxed{0}$ and $\boxed{1}$. Suppose that for some $0 \le p \le 1$, and integer N, the box contains

- $(1-p)N$ $\boxed{0}$s and
- pN $\boxed{1}$s:

$$\underbrace{\boxed{0} \cdots \boxed{0}}_{(1-p)N \text{ of these}} \quad \underbrace{\boxed{1} \cdots \boxed{1}}_{pN \text{ of these}}$$

The parameter p is thus the "proportion of $\boxed{1}$s". Then

- The mean of the box $\mu = \frac{pN}{N} = p$;
- The mean square of the box is also p, and so the SD of the box is

$$\sigma = \sqrt{\text{mn.sq.} - (\text{mean})^2} = \sqrt{p - p^2} = \sqrt{p(1-p)}\,,$$

  only depending on p.

Behaviour of sample sum and sample average (which is also the sample proportion of $\boxed{1}$s) only depends on p and n.

8.2 Prediction Interval for the 0-1 box

8.2.1 Example

Suppose we draw $n = 49$ times randomly from a box with $p = 0.4$. How would $\bar{X}$, the sample mean/proportion of $\boxed{1}$s behave? Can we find an interval *centred* at 0.4 so that the chance of landing in this interval is about 95%?

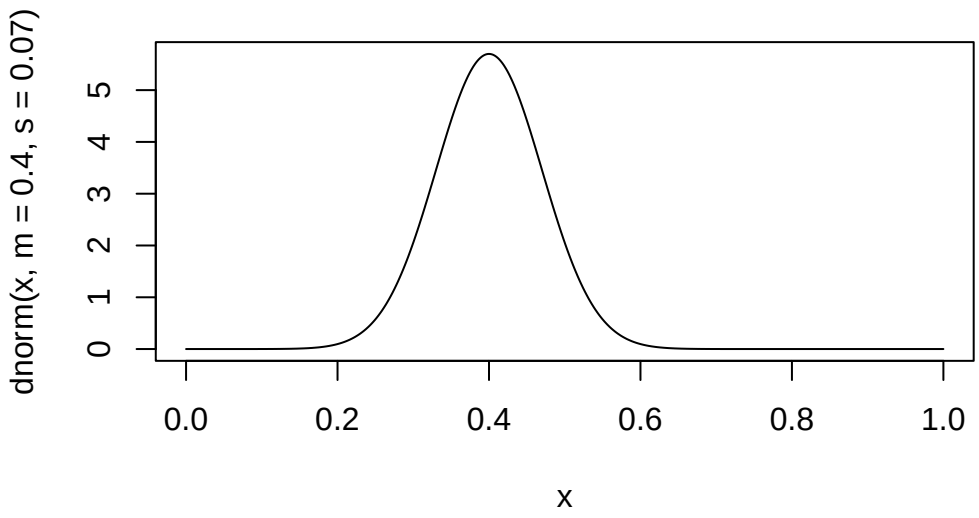
Solution: We know that the expected value is $E(\bar{X}) = \mu = p = 0.4$; the standard error is

$$SD(\bar{X}) = \sigma/\sqrt{n} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{1}{49} \times \frac{2}{5} \times \frac{3}{5}} = \frac{\sqrt{6}}{35} \approx 0.07\,;$$

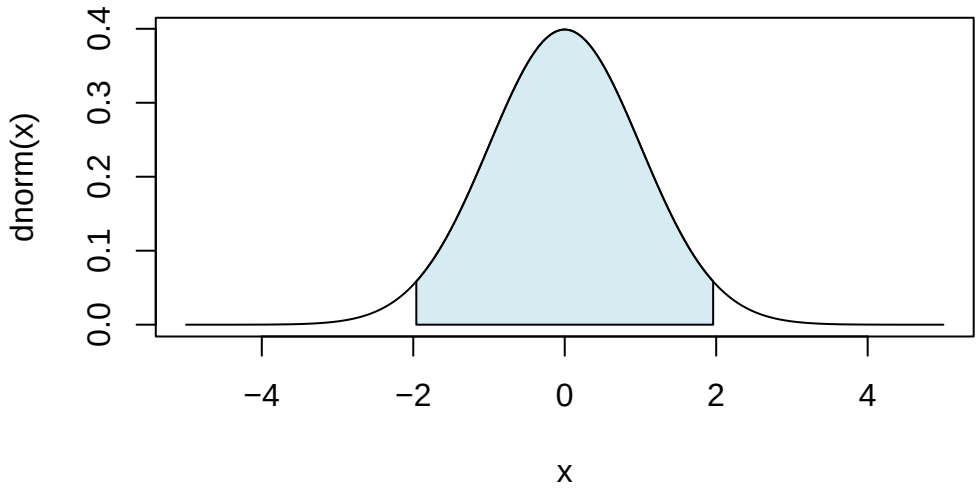
and the distribution of $\bar{X}$ has a (roughly) normal shape.

The random behaviour of the sample proportion is precisely like a **single draw** from the *bigger* box of all possible sample proportions. This bigger box has mean equal to $E(\bar{X}) = 0.4$ and SD equal to $SE(\bar{X}) \approx 0.07$. This bigger box also has an (approx.) normal shape (by the Central Limit Theorem). We may visualise the distribution of the sample proportion $\bar{X}$ in this case as follows:

```
## n=1000 joins 1000 points
curve(dnorm(x, m = 0.4, s = 0.07), from = 0, to = 1, n = 1000)
```



Consider the standard normal curve



Note the following:

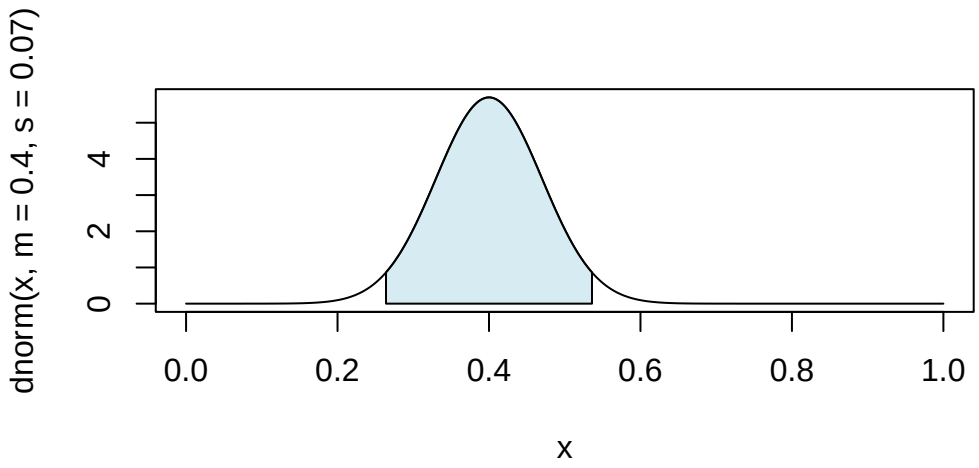
```
round(qnorm(0.025), 2)
```

```
[1] -1.96
```

This means that for any normal-shaped histogram, about 2.5% of the values are more than 1.96 SDs below the mean. By symmetry, about 2.5% of the values are more than 1.96 SDs **above** the mean too. The remaining 95% of values are thus within 1.96 SDs of the mean.

Back on the original scale

Converting back to "original units", the box of all possible sample proportions looks like a normal curve centred at 0.4, but scaled down by a factor of 0.07:



Our prediction interval is thus $0.4 \pm (1.96 \times 0.07)$, i.e. roughly $(0.26, 0.54)$:

```
0.4 + c(-1, 1) * 1.96 * 0.07
```

```
[1] 0.2628 0.5372
```

8.2.2 Alternative exampe: $p = 0.2$

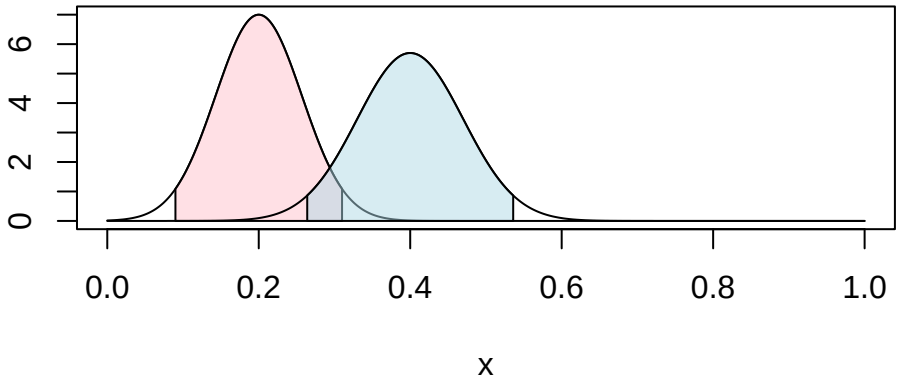
It is interesting to see how this changes if the proportion in the box is 0.2 instead of 0.4. We then get

- $E(\bar{X}) = \mu = p = 0.2$
- $SE(\bar{X}) = \sigma/\sqrt{n} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{1}{49} \times \frac{1}{5} \times \frac{4}{5}} = \frac{2}{35} = 0.057$

So the box of all possible $\bar{X}$ values has

- Mean 0.2;
- SD 0.057;
- A normal shape.

Interval now a bit narrower



Interval is now roughly $(0.09, 0.31)$, i.e. 0.22 units wide (compare with 0.28 when $p = 0.4$).

```
0.2 + c(-1, 1) * 1.96 * 0.057
```

```
[1] 0.08828 0.31172
```

8.2.3 Simulations

We can use a **for loop** to simulate multiple realisations of samples drawn from the 0-1 box, and then check the percentage of sample averages (which is a sample proportion) falling into the prediction interval. The following is an example of a for loop that iterates through a list `x.list`.

```
x.list = 3:7  # a list
y.list = 0  # initialise a second list for saving the output
for (i in 1:length(x.list)) {
    print(x.list[i])  # print the i-th element of x.list
    y.list[i] = x.list[i] * i
}
```

```
[1] 3
[1] 4
[1] 5
[1] 6
[1] 7
```

The loop `for(i in 1:length(x.list))` iterates through the sequence from 1 to `length(x.list)`=5. Each iteration is indexed by `i`. Within the curly brackets`{` and `}`, it executes a set of statements:

- `print(x.list[i])`:
  - print the i-th element of `x.list`
- `y.list[i] = x.list[i]*i`:
  - computes the product `x.list[i]*i` and assign it to the `i`th location of the list `y.list`

```
y.list
```

```
[1]  3  8 15 24 35
```

- Now let's apply the for loop to simulate samples for the case $p = 0.4$

```
too.big = 0
too.small = 0
for (i in 1:1000) {
    samp = sample(c(0, 1), prob = c(0.6, 0.4), repl = T, size = 49)
    prop = mean(samp)
    too.big[i] = prop > (0.4 + 1.96 * 0.07)
    too.small[i] = prop < (0.4 - 1.96 * 0.07)
}
num.too.small = sum(too.small)
num.too.big = sum(too.big)
num.just.right = 1000 - num.too.small - num.too.big
cbind(num.too.small, num.just.right, num.too.big)
```

```
     num.too.small num.just.right num.too.big
[1,]            13            967          20
```

We have nearly 95% of sample proportions falling into the prediction interval.

- Apply the for loop to simulate samples for the case $p = 0.2$

```
too.big = 0
too.small = 0
for (i in 1:1000) {
    samp = sample(c(0, 1), prob = c(0.8, 0.2), repl = T, size = 49)
    prop = mean(samp)
    too.big[i] = prop > (0.2 + 1.96 * 0.057)
    too.small[i] = prop < (0.2 - 1.96 * 0.057)
}
num.too.small = sum(too.small)
num.too.big = sum(too.big)
num.just.right = 1000 - num.too.small - num.too.big
cbind(num.too.small, num.just.right, num.too.big)
```

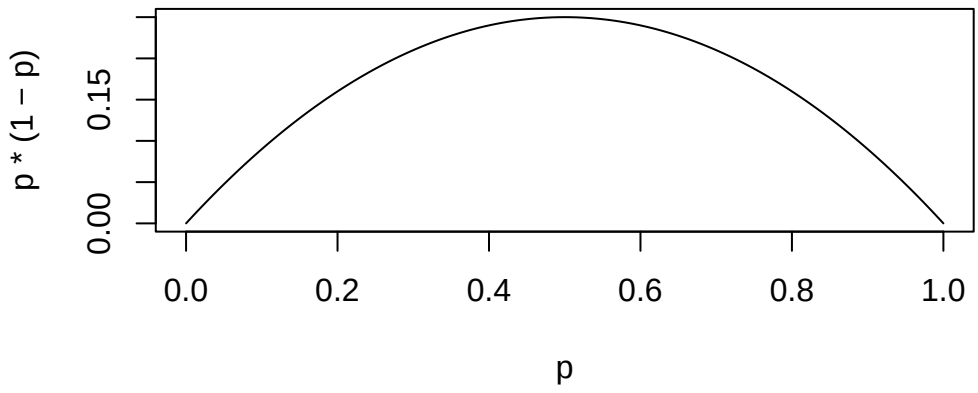
```
     num.too.small num.just.right num.too.big
[1,]            26            948          26
```

Again, we have nearly 95% of sample proportions falling into the prediction interval.

8.2.4 Main take-aways

The function $p \mapsto p(1-p) = p - p^2$ is a quadratic function of p:

```
p = 0:1000/1000
plot(p, p * (1 - p), type = "l")
```



So that the variability in the sample proportion gets **smaller** as the p in the box gets **further from 0.5**. This is precisely reflected in $SE(\bar{X}) = \sqrt{\frac{p(1-p)}{n}}$.

8.2.5 Prediction Interval for the sample proportion $\bar{X}$

If we know the value of p, we can "predict" the value $\bar{X}$ will take. In any normal-shaped box, (approx.) 95% of values are within 1.96 SDs of the mean. Since the box of all possible sample means is normal-shaped, with mean p and SD $\sqrt{\frac{p(1-p)}{n}}$, (approx.) 95% of all possible sample means are within $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$. Thus, since the process of obtaining $\bar{X}$ is like a *single* draw from this bigger box,

$$P\left(p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq \bar{X} \leq p + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95\,.$$

So our **95% prediction interval** for $\bar{X}$ is $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$.

8.2.6 General prediction intervals under the Central Limit Theorem

If the Central Limit Theorem holds, the z-score of the sample mean approximately follows the standard normal curve

$$Z = \frac{\bar{X} - E(\bar{X})}{SE(\bar{X})} \sim N(0,1)$$

With the quantile $q =$ `qnorm(1 - `$\frac{\alpha}{2}$` )` we have $P(Z < -q) = P(Z > q) = \frac{\alpha}{2}$; e.g., $q =$ `qnorm(0.975)` ≈ 1.96 for $\alpha = 5\%$. This way, we also have

$$P(a \leq \bar{X} \leq b) = P\left(\underbrace{\frac{a - E(\bar{X})}{SE(\bar{X})}}_{=-q} \leq Z = \frac{\bar{X} - E(\bar{X})}{SE(\bar{X})} \leq \underbrace{\frac{b - E(\bar{X})}{SE(\bar{X})}}_{=q} \right) = 1 - \alpha.$$

So $(1-\alpha)$ proportion of the sample means, $\bar{X}$, lands in the **two-sided prediction interval**

$$\Big[E(\bar{X}) - q \times SE(\bar{X}), E(\bar{X}) + q \times SE(\bar{X}) \Big]$$

which is symmetric about $E(\bar{X})$.

For example, when $\alpha = 0.05$, a $1 - \alpha = 95\%$ prediction interval is given by the multiplier $q \approx 1.96$:

$$\Big[E(\bar{X}) - 1.96 \times SE(\bar{X}), E(\bar{X}) + 1.96 \times SE(\bar{X}) \Big].$$

8.3 Inference for an unknown proportion: Confidence Interval

The previous section showed us how the sample mean/proportion $\bar{X}$ behaves for a **known** box proportion p. We saw that each value p has associated with it an *interval of values consistent with that* p, characterised as a 95% "prediction interval" for the sample proportion.

- The interval is centred at p
- Its width depends on n and p
- Interval is wider the closer p is to 0.5.

8.3.1 What if the box proportion p is unknown?

Prediction intervals are useful for predicting $\bar{X}$ when p is **known**. They are however not directly useful if p is **unknown**.

We are able to derive an alternative procedure (the confidence interval) for estimating the unknown box proportion p using an interval. - It is based on the idea that if observed value $\bar{x}$ lies in the prediction interval for some p, it is **consistent** with that value of p. - It is an interval based on the observed sample for estimating the box proportion p.

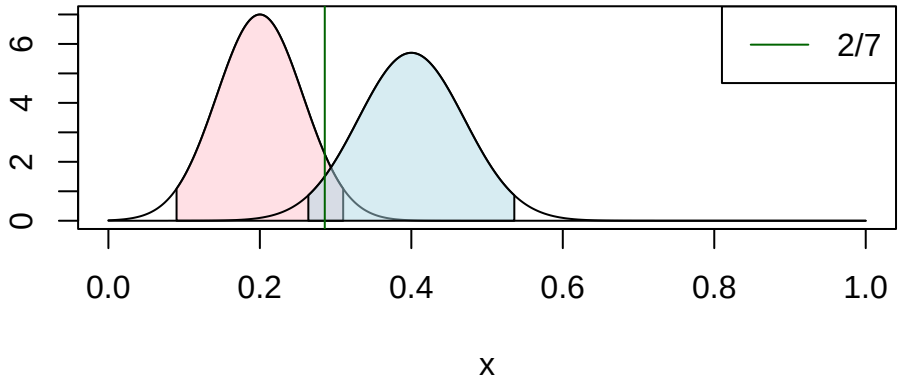
8.3.2 Example

Let's begin with an example. Suppose we have a sample of size $n = 49$ from a box with unknown p. The observed sample sum is $s = 14$, so that the observed sample proportion is $\bar{x} = \frac{s}{n} = \frac{14}{49} = \frac{2}{7} \approx 0.2857$.

We want to know which values of p this observation is consistent with (in the 95% prediction interval sense)?

How about both $p = 0.2$ and $p = 0.4$?

- We replicate our graph from before, showing intervals of values consistent with both $p = 0.2$ and $p = 0.4$, when $n = 49$.
- The vertical green line below shows our observed value $\bar{x} = \frac{2}{7}$.



- Note that $\bar{x} = \frac{2}{7}$ is consistent with both $p = 0.2$ and $p = 0.4$.

In fact, there is a range of p values that are consistent with $\frac{2}{7}$ in the sense their prediction intervals contain the observed $\frac{2}{7}$.

8.3.3 Confidence interval for p

Since an observation $\bar{x}$ has a range of p values such that $\bar{x}$ lies in the 95% prediction interval for that p, that is

$$p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq \bar{x} \leq p + 1.96\sqrt{\frac{p(1-p)}{n}}\,.$$

This range of p values define a **95% confidence interval** for p, which is thus given by the set

$$\left\{ p \,:\, -1.96 \leq \frac{\bar{x} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq 1.96 \right\}\,.$$

Explicit endpoints of this interval can be obtained by solving a quadratic equation, but its easier to use the R commands.

The R package `binom` computes these endpoints using the `binom.confint()` function. In our case, we compute the endpoints as follows:

- Note here the argument 'x' is the sample **sum s** $= n\bar{x}$.

```
require(binom)  # this makes sure the binom package is loaded
```

```
Loading required package: binom
```

```
binom.confint(x = 14, n = 49, method = "wilson")
```

```
  method  x  n      mean     lower     upper
1 wilson 14 49 0.2857143 0.1784959 0.4240888
```

- This shows us the "extreme" values of p for which $\bar{x} = \frac{2}{7}$ is in the 95% prediction interval are $p = 0.178$ and $p = 0.424$.
- As a "sanity check", we can easily check this

```
0.178 + 1.96 * sqrt(0.178 * 0.822/49)  # upper endpoint of values consistent
 ↪  with 0.178
```

```
[1] 0.2851036
```

```
0.424 - 1.96 * sqrt(0.424 * 0.576/49)  # lower endpoint of values consistent
 ↪  with 0.424
```

```
[1] 0.2856267
```

Let's see how the Wilson confidence interval works when repeatedly sampling from a box with a known p

```r
p = 0.3
n = 50
over.est = 0
under.est = 0
for (i in 1:1000) {
    samp = sample(c(0, 1), prob = c(1 - p, p), replace = T, size = n)
    s = sum(samp)
    w = binom.confint(s, n, method = "wilson")
    over.est[i] = w$lower > p
    under.est[i] = w$upper < p
}
num.over.est = sum(over.est)
num.under.est = sum(under.est)
num.covering = 1000 - num.over.est - num.under.est
cbind(num.under.est, num.covering, num.over.est)
```

```
     num.under.est num.covering num.over.est
[1,]            21          948           31
```

We see that close to 95% of the time, the interval covers the "true" value of $p = 0.3$.

8.3.4 Properties of the Wilson confidence interval

- Under repeated sampling from a 0-1 box, the 95% Wilson confidence interval covers the "true" proportion p in (approx.) 95% of samples.
- This is a long-run property of the procedure.
- For a *single* data set, you don't know if it has covered the true value or not.
  - You just know that the procedure you have used is 95% reliable in the long run.
- Note that the interval is not (in general) symmetric about the observed sample proportion $\bar{x}$.
  - The midpoint of the interval is somewhere between $\bar{x}$ and 0.5.

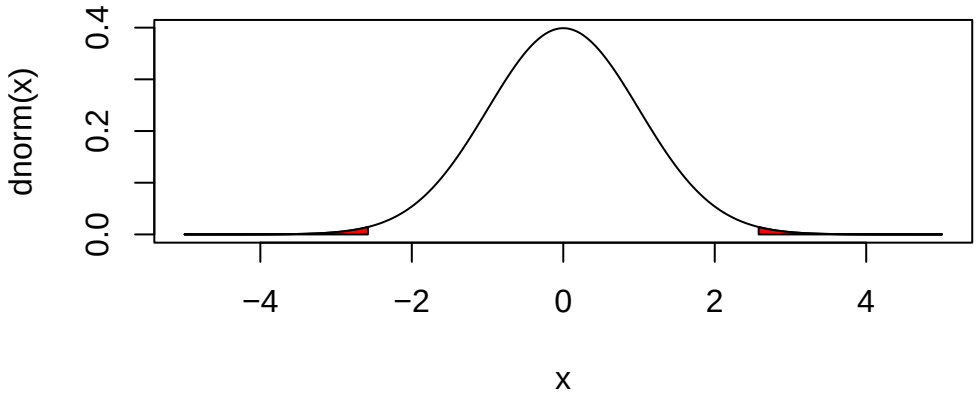
8.3.5 Different confidence levels

- We can change the confidence level by replacing 1.96 with another value.
- E.g., for 99% we should replace `1.96` with

```
qnorm(0.995)
```

```
[1] 2.575829
```

(which gives 0.5% in the upper tail under the standard normal curve).



Using `binom.confint()` we simply set the `conf.level=` argument to the desired level:

```
binom.confint(14, 49, conf.level = 0.99, method = "wilson")
```

```
  method  x  n      mean     lower     upper
1 wilson 14 49 0.2857143 0.1531828 0.4693562
```

As a *sanity check*, we can manually verify that the observed value $\frac{2}{7} \approx 0.2857\ldots$ is "right on the edge" for each of the endpoints 0.153 and 0.469, using the larger multiplier 2.576:

```
0.469 - 2.576 * sqrt(0.469 * 0.531/49)
```

```
[1] 0.285354
```

```
0.153 + 2.576 * sqrt(0.153 * 0.847/49)
```

```
[1] 0.2854754
```

8.3.6 Case study

The file `march2024.csv` has daily weather observations from the Canterbury Racecourse weather station for March 2024.

```
mar.2024 = read.csv("data/march2024.csv", skip = 5)
summary(mar.2024)
```

```
   X                Date           Minimum.temperature..degC.
Mode:logical   Length:31          Min.   :12.30
NA's:31        Class :character   1st Qu.:14.95
               Mode  :character   Median :16.90
                                  Mean   :16.72
                                  3rd Qu.:18.00
                                  Max.   :23.20

 Maximum.temperature..degC. Rainfall..mm.     Evaporation..mm. Sunshine..hours.
 Min.   :21.50              Min.   : 0.000   Mode:logical     Mode:logical
 1st Qu.:24.85              1st Qu.: 0.000   NA's:31          NA's:31
 Median :27.30              Median : 0.000
 Mean   :27.19              Mean   : 1.759
 3rd Qu.:29.00              3rd Qu.: 0.200
 Max.   :34.20              Max.   :35.600
                            NA's   :2
 Direction.of.maximum.wind.gust. Speed.of.maximum.wind.gust..km.h.
 Length:31                       Min.   :24.0
 Class :character                1st Qu.:28.5
 Mode  :character                Median :35.0
                                 Mean   :36.4
                                 3rd Qu.:42.5
                                 Max.   :56.0
                                 NA's   :1
 Time.of.maximum.wind.gust X9am.Temperature..degC. X9am.relative.humidity....
 Length:31                 Min.   :16.00           Min.   : 56.00
 Class :character          1st Qu.:19.30           1st Qu.: 76.00
 Mode  :character          Median :21.50           Median : 81.00
                           Mean   :21.04           Mean   : 81.03
                           3rd Qu.:23.05           3rd Qu.: 86.00
                           Max.   :24.80           Max.   :100.00

 X9am.cloud.amount..oktas. X9am.wind.direction X9am.wind.speed..km.h.
 Mode:logical              Length:31           Length:31
```

```
 NA's:31                   Class :character    Class :character
                           Mode  :character    Mode  :character



 X9am.MSL.pressure..hPa. X3pm.Temperature..degC. X3pm.relative.humidity....
 Mode:logical            Min.   :20.10           Min.   :18.00
 NA's:31                 1st Qu.:23.65           1st Qu.:49.50
                         Median :25.50           Median :57.00
                         Mean   :25.74           Mean   :56.58
                         3rd Qu.:27.35           3rd Qu.:63.50
                         Max.   :33.30           Max.   :91.00

 X3pm.cloud.amount..oktas. X3pm.wind.direction X3pm.wind.speed..km.h.
 Mode:logical              Length:31           Min.   : 2.00
 NA's:31                   Class :character    1st Qu.:13.00
                           Mode  :character    Median :17.00
                                               Mean   :17.29
                                               3rd Qu.:22.00
                                               Max.   :30.00

 X3pm.MSL.pressure..hPa.
 Mode:logical
 NA's:31
```

```
mar.2024$Rain
```

```
 [1]  0.0  0.0  1.0  0.2  0.0  0.0  0.0  0.0  0.0  0.0   NA  0.2   NA  0.0  4.2
[16]  1.0 35.6  1.2  6.2  0.2  0.6  0.0  0.2  0.2  0.2  0.0  0.0  0.0  0.0  0.0
[31]  0.0
```

Suppose we can model the presence or absence of rain as being like a random sample from a 0-1 box with an unknown proportion p of 1s. What is a 95% Wilson confidence interval for p?

```
rain = na.omit(mar.2024$Rain)
n = length(rain)
s = sum(rain > 0)
binom.confint(s, n, method = "wilson")
```

```
  method  x  n      mean     lower     upper
1 wilson 13 29 0.4482759 0.2841317 0.6245204
```

The data is thus consistent with the "true" p being anywhere in the range $(0.284, 0.625)$.

8.4 Inference for an unknown proportion: Hypothesis Test

Suppose we have data modelled as a random sample from a 0-1 box, with unknown proportion p of 1s. We have seen how to produce a confidence interval, which is a collection of p values that the data is consistent with. In some scentific scenarios, there is a special value of the parameter which might be of interest. For example: A company claims that 60% of customers prefer their product ($p = 0.6$). We want to collect data to assess this claim. In this case, instead of estimating p, we may want to test whether the data supports or contradicts this special value.

8.4.1 Elements of a hypothesis test

Let's continue work on the rainfall data set. Suppose that historical data indicates that the proportion of rainy days at Canterbury in March is 0.2. Is the March 2024 data (i.e. $\bar{x} = \frac{s}{n} = \frac{13}{29} \approx 0.448$) consistent with this?

Let us interpret "consistent" to mean "in the sense of 95% prediction". The answer is NO. We can see this two ways

1. The **prediction interval** way: We can explicitly construct a 95% prediction interval for $\bar{X}$ when the true $p = 0.2$. This would look like:

```
0.2 + c(-1, 1) * 1.96 * sqrt(0.2 * 0.8/29)
```

```
[1] 0.05441485 0.34558515
```

Since this does not include 0.45, we conclude that an observation of 0.448 is *not* consistent with $p = 0.2$ (in this sense).

2. The **confidence interval** way: We have already computed a 95% confidence interval based on this data: $(0.284, 0.625)$. These are the values of the p parameter that the data are consistent with. Since 0.2 is not included, the data is not consistent with 0.2, in this sense.

We want to use the prediction interval way to introduce elements of a hypothesis test.

8.4.1.1 False alarm rate / level of significance

In fact, before we see the data, we can indicate what our conclusion would be for any potential observation:

- If the observed $\bar{x}$ is within the range (0.054, 0.346) we would conclude "**data is consistent with the hypothesis** $p = 0.2$";
- If the observed $\bar{x}$ is *outside* the range (0.054, 0.346) we would **"reject" the hypothesis** $p = 0.2$

This "rejection" statement entails some risk: there is a chance we can reject the hypothesis incorrectly.

- When the hypothesis is true (i.e true $p = 0.2$) there is a 5% chance $\bar{X}$ lands outside (0.054, 0.346).
- This is the **false alarm rate** or **level of significance** of our procedure (the chance we reject the hypothesis when it is true).

The smaller the false alarm rate, the more "cautious" we are:

- We then only reject the hypothesis if there is overwhelming evidence in the data.

8.4.1.2 Measuring strength of "evidence against"

What if we instead start with a 99% prediction interval for $p = 0.2$? As we have seen, this would give $(0.009, 0.391)$:

```
round(qnorm(0.995), 3)
```

```
[1] 2.576
```

```
0.2 + c(-1, 1) * 2.576 * sqrt(0.2 * 0.8/29)
```

```
[1] 0.008659524 0.391340476
```

Our observed $\bar{x} \approx 0.448$ is also outside this, so we would *also* reject the hypothesis $p = 0.2$ at the 1% level of significance. How small do we have to make the false alarm rate before we do not reject?

A 99.9% prediction interval would use a multiplier which has only 0.05% in the upper tail of the standard normal curve:

```
qnorm(0.9995)
```

```
[1] 3.290527
```

The corresponding prediction interval is

```
0.2 + c(-1, 1) * 3.29 * sqrt(0.2 * 0.8/29)
```

```
[1] -0.04437507  0.44437507
```

This *still* does not include 0.448, so we *still* reject the hypothesis $p = 0.2$ even at the 0.1% level of significance.

A 99.99% prediction interval for $\bar{X}$ when the true $p = 0.2$ uses the multiplier

```
qnorm(0.99995)
```

```
[1] 3.890592
```

```
0.2 + c(-1, 1) * 3.89 * sqrt(0.2 * 0.8/31)
```

```
[1] -0.07946585  0.47946585
```

Finally, this includes the observed value 0.448. So we would *not* reject the hypothesis $p = 0.2$ if we used the **super-cautious** 0.01% false alarm rate.

8.4.1.3 Observed level of significance / P-value.

We can work out the **exact** false alarm rate at which the observed $\bar{x} = 0.448$ is *right on the edge*. This is the level which uses as multiplier the value z such that

$$0.2 + z\sqrt{\frac{0.2 \times 0.8}{29}} = 0.448\,, \quad \text{that is} \quad z = \frac{0.448 - 0.2}{\sqrt{0.2 \times 0.8/29}} \approx 3.339\,.$$

The desired false alarm rate is simply **twice** the upper tail area beyond 3.339:

```
2 * pnorm(3.339, lower.tail = F)
```

```
[1] 0.0008408056
```

This (small) quantity is the **observed level of significance** or **P-value** based on the data.

The smaller the P-value, the stronger the evidence in the data against the hypothesis. The P-value may be interpreted as a probability: **The probability of getting as much evidence against the hypothesis as was observed, when the hypothesis is true.**

8.4.1.4 Z-statistic

Using the normal approximation of the box model, we know that the observed sample mean follows a normal curve. The "crucial value of the multiplier" is

$$z = \frac{0.45 - 0.2}{\sqrt{0.2 \times 0.8/31}} = \frac{\bar{x} - E_0(\bar{X})}{SE_0(\bar{X})}$$

where $p_0 = 0.2$ is the *hypothesised value*. This simply measure how many SEs away the observed value $\bar{x}$ is from the expected value, converting the observed $\bar{x}$ into *standard units*, assuming the hypothesis is true. The value z is in turn the observed value of the (random) **Z-statistic**:

$$Z = \frac{\bar{X} - E_0(\bar{X})}{SE_0(\bar{X})}$$

which is approximately distributed like a draw from a standard normal box **if the hypothesis is true**.

8.5 A review of inference for unknown proportions

8.5.1 HATPC framework

When conducting hypothesis tests, it's helpful to follow a structured approach. One such approach is the HATPC framework, which consists of five key steps:

H Hypotheses

- Set up the two hypotheses: the null H_0 and the alternative H_1.

A Assumptions

- State the assumption(s) of the test, and justify if they are valid based on the sample and the sampling process

T Test Statistic

- State the Test Statistic and it's distribution (the underlying model) **assuming H_0 is true**.
- State what value of the test statistic argue against H_0 for H_1.
- Find the observed value of the Test Statistic.

P P-value

- Calculate the P-value, the probability of observing the sample (or more extreme) under H_0.

C Conclusion

- Weigh up the conclusion, based on the P-value and the level of significance α.

8.5.2 Summary of Z-test procedure using HATPC

H The hypotheses are commonly articulated in terms of the unknown population parameter.

The **null hypothesis** H_0 is the default hypothesis: what we currently believe to be true about the population. In this case, $H_0 : p_0 = 0.2$.

The **alternate hypothesis** H_1 is a new claim about the population. It can take 2 forms:

- 1-sided ($H_1 : p_0 > 0.2$ or $H_1 : p_0 < 0.2$)
- 2-sided ($H_1 : p_0 \neq 0.2$).

We decide between a 1- or 2-sided test using the context of the problem. In this case, we take $H_1 : p_0 \neq 0.2$, as this is the alternative we want to detect. Note that, the decision must not be influenced by the data – we must specify the hypotheses before we do the actual test.

A The assumptions are necessary for the test to be valid. We justify whether they are valid based on the sample and the sampling process. In the rainfall data case, we have a sample size $n = 29$. It may be sufficient large for assuming the central limit theorem (CLT). Then the normal curve can be used to approximate the sample proportion.

T The test statistic can be viewed as a random draw from a box that depends on the unknown population parameter. We derive either the test statistic or its distribution (box) from H_0. The sample proportion $\bar{X}$ has $E_0(\bar{X}) = p_0$ and $SE_0(\bar{X}) = \sqrt{\frac{p_0(1-p_0)}{n}}$. The z-score of the sample proportion approximately follows the standard normal curve

$$Z = \frac{\bar{X} - E_0(\bar{X})}{SE_0(\bar{X})} = \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

In this case, Z is the test statistic, and hence **Z-statistic**.

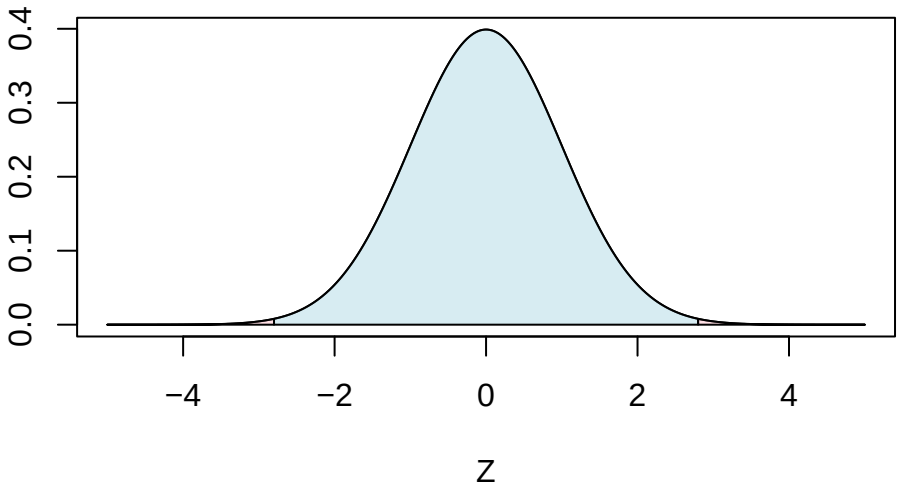
Depending on whether it is a one-sided test or a two-sided test, we need to determine what values of the Test Statistic will argue against H_0. In this case, both large and small values of Z-statistic will argue against H_0.

Calculate the **observed value of the Test Statistic** (see Page 11 for details)

$$z = \frac{\bar{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx 3.339$$

P The P-value is the probability of observing something more extreme than the observed sample (under H_0).

- "something more extreme" = test statistics that argue against H_0 more than the observed one
- two-sided test in this case: small and large values of Z-statistic will argue against H_0
- P-value $P(|Z| > |z|) = P(Z < -3.339) + P(Z > 3.339)$, by symmetry, which is
  - $2\times$ `pnorm(3.339, lower.tail=F)` ≈ 0.00084



A small P-value either means that either H_0 is true but the sample is highly rare, or H_0 is false. The smaller the P-value, the stronger the evidence against H_0 for H_1. A large P-value means that the sample is consistent with H_0.

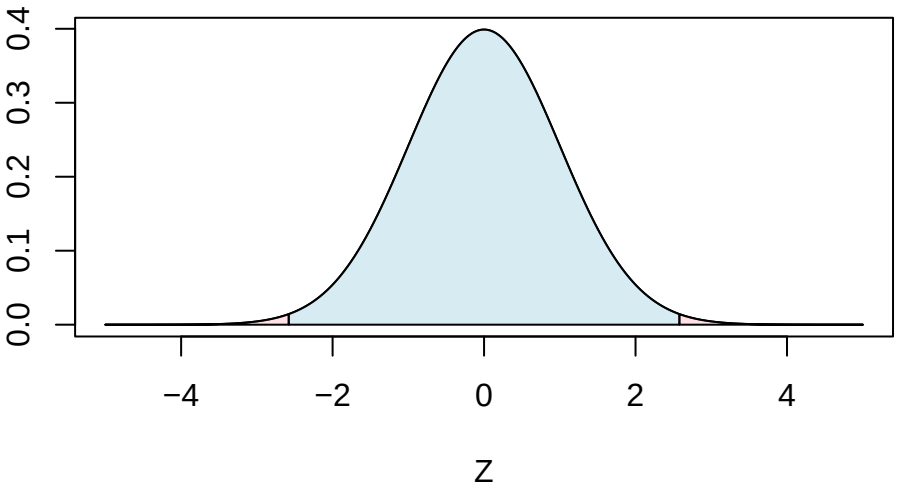
C We often make decision based on P-value of the **level of significance** α. For example, if we use $\alpha = 1\%$, which allows 1% of false alarm rate assuming H_0 is true. For the P-value $= 0.00084 < \alpha$, we reject H_0. This constitutes evidence against H_0, suggesting the historical proportion of rainy days in March is unlikely to be 0.2.

In addition, we don't need the P-value to make the decision. Based on the level of significance α, we can determine a **critical region** of test statistics such that their corresponding P-values are smaller than α.

In this case, given it's two-sided test (small and large values of Z-statistic arguing against H_0), we have the multiplier corresponding to the $1 - \alpha/2 = 99.5\%$ quantile

```
round(qnorm(0.995), 3)
```

```
[1] 2.576
```



The critical region is given by $|z| > 2.576$. The observed Z-statistic 3.339 is in the critical region, so the corresponding P-value is less than the level of significance, and thus we reject H_0 at the 1% level of significance.

8.5.3 Not rejecting H_0, but not accepting it either

If $\bar{x}$ lands within the prediction interval, i.e. if

$$|z| = \frac{|\bar{x} - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq 2.576\,,$$

we say the data is **consistent with** H_0 (at the 1% level of significance). However, we do not "accept" $H_0 : p = p_0$, since a single observation does not "prove a hypothesis true".

8.5.4 Different confidence/significance levels

The multiplier/critical value cuts off a certain area in the upper tail under the standard normal curve. In general, suppose we have a significance level $0 \leq \alpha \leq 1$. Then the multiplier/critical value cuts of an area of $\alpha/2$ in the upper tail of the standard normal curve. Multipliers/critical values for common confidence/significance levels are:

Sig. level	Conf. level	Upper tail area	Multiplier/critical value	R command
0.05	95%	0.025	1.960	`qnorm(0.975)`
0.02	98%	0.010	2.326	`qnorm(0.99)`
0.01	99%	0.005	2.576	`qnorm(0.995)`
0.001	99.9%	0.0005	3.291	`qnorm(0.9995)`

Sig. level	Conf. level	Upper tail area	Multiplier/critical value	R command
0.0001	99.99%	0.00005	3.891	`qnorm(0.99995)`

8.6 One-sided tests

In many practical hypothesis-testing scenarios, values both above and below the hypothesised value p_0 might be of interest. For examples:

- p = proportion of days with rain in March: is climate change increasing or decreasing rain in March? Here p_0 represents historical proportion of days with rain in March.
- p = proportion of patients showing improvement using a new drug: is the new drug better or worse than the current standard treatment? Here p_0 represents proportion of patients showing improvement with current standard treatment.

We will use the following example to study the details of the one-sided alternative.

8.6.1 Case study: production lines

Suppose a production line produces items at a rate of 5000 per day. The process occasionally produces faulty items. It is deemed "acceptable" if 3% of the items are faulty. As a quality control measure, once a week a random sample of $n = 200$ items is taken and the proportion of faulty items $\bar{x}$ determined. If there is evidence that the "failure rate" is higher than 3%, they plan to halt production and overhaul the machines using a 1% false alarm rate.

This is a costly process, so it should only be done if the evidence is "clear". How should such a test be performed so that the false alarm rate (the chance of needless shutdown) is no more than 1%?

8.6.2 Apply the HATPC framework

H Let the parameter p represents the actual proportion of faulty items being produced.

- Our **null hypothesis** is $H_0 : p = 0.03$, which represents "nothing interesting going on".
- We declare our **alternative hypothesis** to be $H_1 : p > 0.03$, which is the alternatives we are "trying to detect". Note that the direction of the alternative suggests what procedure to use.

A The total number of items produced a week is large $5,000 \times 7 = 35,000$. Selecting a sample with $n = 200$ items is a sample without replacement from a very large box, and thus taking a sample without replacement from a large box does not modify the box much. So a sample of 200 items can be viewed as "almost" independent. This way, we can justify that $n = 200$ is a

sufficiently large sample size so that **CLT** may hold, and hence the normal curve can be used to approximate the sample proportion.

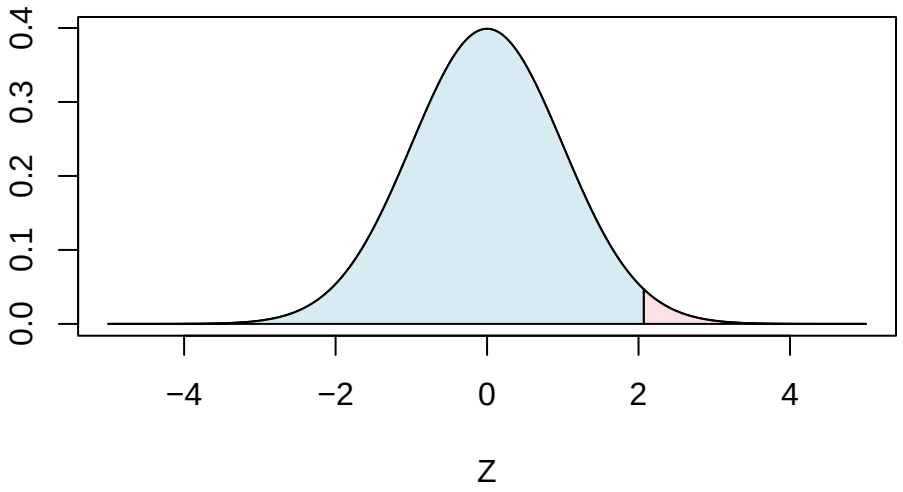
T Assuming H_0 is true, the z-score of the sample proportion (**Z-statistic**)

$$Z = \frac{\bar{X} - E_0(\bar{X})}{SE_0(\bar{X})} = \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

approximately follows the standard normal curve. In this case, large values of Z-statistic will argue against H_0. Calculate the **observed value of the Test Statistic** using the observed proportion $\bar{x} = \frac{11}{200} = 0.055$:

$$z = \frac{\bar{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.055 - 0.03}{\sqrt{\frac{0.03\times(1-0.03)}{200}}} \approx 2.07.$$

P The P-value is the probability of observing something more extreme than the observed sample (under H_0). We have a one-sided test in this case, and large values of Z-statistic will argue against H_0, so P-value $= P(Z > z) = P(Z > 2.07) =$ `pnorm(2.07, lower.tail=F)` $=$ 0.019.



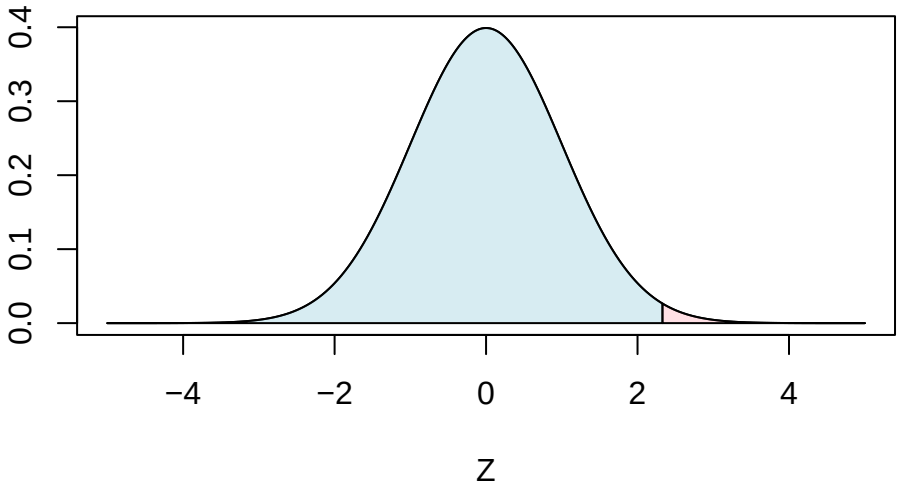
C In this example, the P-value is 0.019 (with observed Z-statistic 2.07). Since P-value $= 0.019 > \alpha$, so we can't reject H_0. We say "the data is consistent with the null hypothesis H_0".

We can also construct the region of rejection to make an equivalent decision. In this example, an $\alpha = 1\%$ **level of significance** is used. Given it's one-sided and large values of Z-statistic arguing against H_0. We have the multiplier:

```
round(qnorm(0.99), 2)
```

```
[1] 2.33
```

If observed statistic falls above this multiplier, the corresponding P-value is less than the level of significance. This way, the critical regions is given by $[2.33, \infty)$.



Since the obsered Z-statistic is below the multiplier 2.33, we consider the data is consistent with H_0 in this case.

8.6.3 Practical note: specify conclusions before seeing the data

We can indicate possible conclusions *before seeing the data* to prevent "data snooping" (letting the data suggest the procedure). In this example, we can provide an operation guide (in the following code) if one should shut the production line based on different levels of significance.

```
false.alarm.rate = c(0.01, 0.1, 1, 5)  # in percentage
critical.values = qnorm(1 - false.alarm.rate/100)  # multipliers/critical
 ↪  values
n = 200  # sample size
p0 = 0.03  # H_0
# expected sample proportion
E.Xbar = p0
## SE of the sample proportion
SE.Xbar = sqrt(p0 * (1 - p0)/n)
```

```r
## critical values in sample proportions
critical.values.props = (critical.values * SE.Xbar + E.Xbar)
## critical values in sample sums
critical.values.sums = critical.values.props * n
## rounding to the nearest interger larger than sums
observed.faulty.items = ceiling(critical.values.sums)
## table of operational guide
cbind(false.alarm.rate, observed.faulty.items)
```

```
     false.alarm.rate observed.faulty.items
[1,]             0.01                   15
[2,]             0.10                   14
[3,]             1.00                   12
[4,]             5.00                   10
```

Note that we use `ceiling()` to round to the nearest interger larger than the critical values of sample sum. This way, for each observed no. of faulty items, its P-value is less than the listed level of significance. If we "reject" based on such an observation, we shut the production line more cautiously than the specified false alarm rate.

8.6.4 Summary

If we are testing $H_0 : p = p_0$ based on an observed proportion $\bar{x}$, or equivalently observed Z-statisic

$$z = \frac{\bar{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

we can identify which procedure we should use by identifying which of these questions is appropriate:

- If we ask "*Is the population proportion significantly* ***greater*** *than* p_0*?*", we should use the (one-sided) alternative $H_1 : p > p_0$ and compute the P-value using `pnorm(`z`,lower.tail=F)`;
- If we ask "*Isthe population proportion significantly* ***less*** *than* p_0*?*", we should use the (one-sided) alternative $H_1 : p < p_0$ and compute the P-value using `pnorm(`z`)`;
- If we ask "*Is the population proportion significantly* ***different*** *to* p_0*?*", we should use the (two-sided) alternative $H_1 : p \neq p_0$ and compute the P-value using `2*pnorm(abs(`z`),lower.tail=F)`.

9 Unknown Means

9.1 General Z-test

So far we have focussed on the scenario of inference for a proportion. Since a hypothesis $H_0 : p = p_0$ fixes both the mean $\mu = p_0$ and SD $\sigma = \sqrt{p_0(1-p_0)}$ of the box, both $E(\bar{X})$ and $SE(\bar{X})$ are known when H_0 is true. This allows us to perform the Z-test, since we only need to know how the Z-statistic behaves **when H_0 is true**.

We can have more general settings where

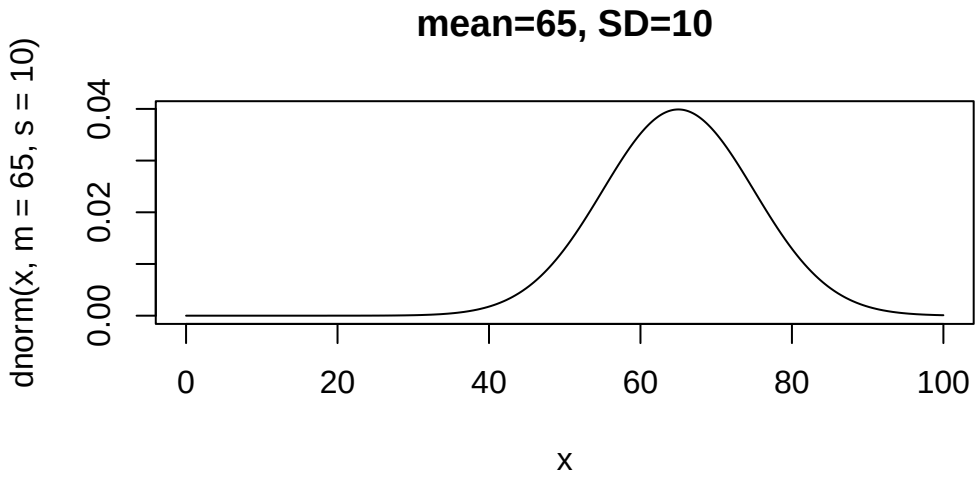
- the box mean μ is the unknown parameter of interest but
- The SD σ_0 of the box is **known**.

In this case, when a hypothesis $H_0 : \mu = \mu_0$ is true, we again have $E(\bar{X})$ and $SE(\bar{X})$ known, so a Z-test can still be performed.

9.1.1 Case study: standardised school exams

In many jurisdictions, students are assessed using standardised exams, where marks for several subjects are combined to give a single score. It is thus important that exam marks from different subjects are comparable. To achieve this, the scores for each exam should follow a "standard" distribution, e.g. a normal distribution with mean 65 and SD 10.

```
x = 0:1000/10
plot(x, dnorm(x, m = 65, s = 10), type = "l", main = "mean=65, SD=10")
```



Suppose that for a certain subject, ordinary exams are such that the spread of marks from year to year is much the same, with SD $= \sigma_0 = 10$ (known population SD) **but** the average mark μ tends to vary from year to year.

In order to produce a "standardised" exam, a draft can be tested on a small group of students. If there is evidence that the "population mean" mark μ is different to 65, the exam is moderated (i.e. made a bit easier or a bit harder).

Suppose a group of 100 students take the draft exam, and obtain the marks below

```
marks
```

```
  [1] 64 57 67 66 69 53 67 49 67 64 71 62 63 51 51 59 59 54 70 44 68 47 40 49 57 62 58 48 63
 [42] 67 53 41 72 85 52 54 84 57 81 79 58 45 69 59 68 64 57 70 64 55 66 45 73 68 78 54 65 49
 [83] 62 73 80 70 57 78 56 59 65 73 60 72 76 62 57 68 77 71
```

```
summary(marks)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  35.00   55.00   63.00   62.46   70.00   85.00
```

The average mark here is a bit **different** to 65, but it is **significantly different**? We want to apply the Z-test procedure to find this out.

9.1.2 Formal hypothesis test

A Students are like a random sample taken without replacement from a very large box (there are many students in the population), so we consider the resulting marks as independent draws. A sample size $n = 100$ is sufficiently large for apply the Central Limit Theorem. The underlying population has unknown mean μ but **known** SD $\sigma_0 = 10$.

H

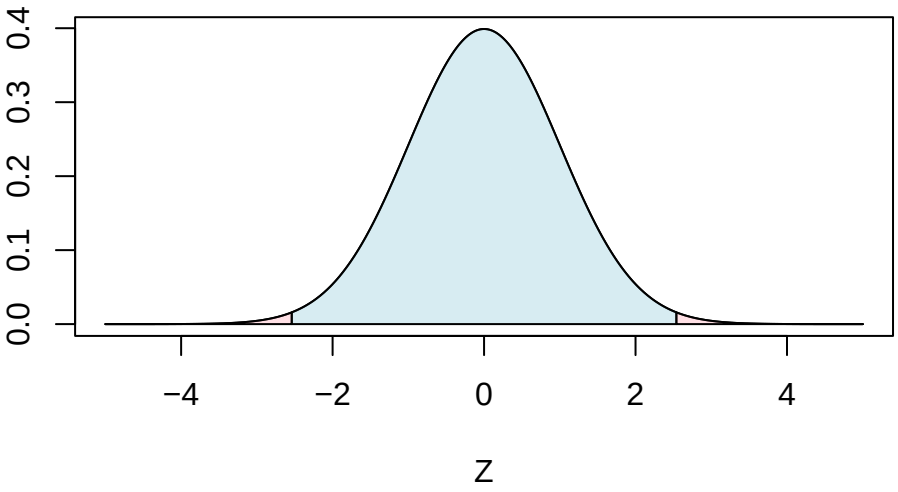
- **Null hypothesis**: $H_0 : \mu = 65$.
- **Alternative hypothesis**: $H_1 : \mu \neq 65$. A *two-sided* test is appropriate here, since the exam could be too easy or too hard.

T When H_0 is true the sample mean $\bar{X}$ is like a random draw from a normal box with mean equal to $E(\bar{X}) = \mu = 65$ and SD equal to $SE(\bar{X}) = \frac{\sigma_0}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$. Equivalently the Z-statistic

$$Z = \frac{\bar{X} - 65}{1} = \bar{X} - 65$$

is like a random draw from a standard normal box.

P Two-sided test here: large and small values of Z-statistic will argue against H_0, so P-value $= P(Z > |z|) + P(Z < -|z|) =$ `2 * pnorm(abs(-2.54), lower.tail=F)` $= 0.011$.



C P-value $= 0.011 < \alpha = 0.05$, so we reject H_0. The observed mean is significantly different to 65 at the 5% level of significance. This constitutes evidence against the null hypothesis, suggesting the exam needs moderation. Note that for the 5% level of significance, the critical region is given by $|z| \geq 1.96$.

9.1.3 Confidence interval

Since the alternative is two-sided, we can also consider the construction of a confidence interval for μ based on the observed data. This is because alternative values both above and below 65 are of equal interest.

Let's start with building a prediction interval for the sample mean $\bar{X}$. Under the assumptions, for any value μ, the (random) sample mean $\bar{X}$ is like a single random draw from a normal box with mean μ and SD $\frac{\sigma_0}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$. Note that the standard error of $\bar{X}$ does not depend on μ (*under the assumptions*). This way, a 95% prediction interval for μ is then given by

$$\mu \pm 1.96 \frac{\sigma_0}{\sqrt{n}}$$

that is, $\mu \pm 1.96$.

An observed sample mean $\bar{x}$ is **consistent** with a population mean μ if $\bar{x}$ lands in the 95% prediction interval of μ, that is, they are consistent if $\bar{x} \in (\mu - 1.96, \mu + 1.96)$.

We may define a 95% confidence interval based on the observed $\bar{x}$ as "all μ such that $\bar{x}$ landed in the corresponding 95% prediction interval". That is all μ such that $\mu - 1.96 \leq \bar{x} \leq \mu + 1.96$. But this is the same as $\bar{x} - 1.96 \leq \mu \leq \bar{x} + 1.96$. Because the standard error does not depend on the parameter μ, it is easier to explicitly solve for the interval endpoints, than in the case of an unknown proportion. This gives

```
62.5 + c(-1, 1) * 1.96
```

```
[1] 60.54 64.46
```

Note this does not include the value 65, so in this sense, the data is not consistent with μ being 65.

We can also get confidence intervals at different confidence levels:

Conf. level	Multiplier	Interval	Includes 65?
95%	1.960	(60.540, 64.460)	No
98%	2.326	(60.264, 64.736)	No
99%	2.576	(59.924, 65.076)	Yes

So we need to go to the "rather cautious" 99% confidence level before we agree the data is consistent with $\mu = 65$. This is in agreement with the hypothesis test:

- At the 5% level of significance, we reject H_0 (since P-value smaller than 0.05)
- At the 2% level of significance, we reject H_0 (since P-value smaller than 0.02)
- At the 1% level of significance, we do **not** reject H_0 (since P-value bigger than 0.01).

We will discuss the general formulation of the confidence interval for unknown population mean in next topic.

9.2 Estimating the standard error

In the previous example, we assumed the data was like a random sample from a box with **known** SD $\sigma = \sigma_0 = 10$ but unknown mean μ. Note the following R output:

```
popsd = function(x) sqrt(mean(x^2) - (mean(x))^2)
popsd(marks)
```

```
[1] 10.65685
```

```
sd(marks)
```

```
[1] 10.71053
```

Recall that if the observed marks are $x_1, \ldots, x_{100}$,

- `popsd(marks)` gives the population SD $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ while
- `sd(marks)` gives the sample SD $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Both may be viewed as "estimates" of the σ in the box. The population SD is a natural measure of spread of a given list of numbers which we are regarding as a **population** i.e. contents of a box, whereas the sample SD is better to use when the numbers we have are a **sample** from some population with SD σ unknown and we wish to **estimate** σ in the population.

9.2.1 Sample SD in depth (not for assessment)

Using the box model, we will explain why the sample SD is a better choice for **estimating** σ in the population for a given sample.

Suppose $X_1, \ldots, X_n$ is a random sample with replacement from a box

$$\boxed{\boxed{y_1}\ \boxed{y_2}\ \cdots\ \boxed{y_N}}$$

with mean $\mu = \frac{1}{N}\sum_{j=1}^{N} y_j$ and SD $\sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \mu)^2}$. We already know how the sample sum $S = \sum_{i=1}^{n} X_i$ and sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}$ behave. In particular

- $E(S) = n\mu$
- $E(\bar{X}) = \mu$

We want to know how the *sum of squared deviations (from μ)*, i.e. $SSD = \sum_{i=1}^{n}(X_i - \mu)^2$ behave.

We start with a different box

$$\boxed{\boxed{(y_1-\mu)^2}\ \boxed{(y_2-\mu)^2}\ \cdots\ \boxed{(y_N-\mu)^2}}$$

This box has mean

$$\frac{1}{n}\sum_{j=1}^{N}(y_j - \mu)^2 = \sigma^2\,.$$

So we have the identify

$$E[(X-\mu)^2] = \sigma^2 = SE(X)^2$$

the expected value of $(X-\mu)^2$ is the same as the standard error of X. The SSD is exactly like the *sum* of a random sample with replacement from this box. So

$$E(SSD) = n\sigma^2 .$$

We can use a trick. We can "add and subtract" the (random) sample mean inside the squareto get

$$SSD = \sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2$$

Expanding out the square we get

$$SSD = \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2(\bar{X}-\mu)\underbrace{\sum_{i=1}^{n}(X_i-\bar{X})}_{=0} + n(\bar{X}-\mu)^2$$

$$= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2 .$$

Taking expected values and rearranging we see that

$$n\sigma^2 = E(SSD) = E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] + E\left[n(\bar{X}-\mu)^2\right]$$

Dividing by the sample size n, we have

$$\sigma^2 = E\left(\frac{1}{n}SSD\right) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] + E\left[(\bar{X}-\mu)^2\right]$$

Rearranging, it leads to

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \sigma^2 - E\left[(\bar{X}-\mu)^2\right]$$

It turns out that $E\left[(\bar{X}-\mu)^2\right] = SE(\bar{X})^2 = \frac{\sigma^2}{n}$ so

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \sigma^2\left(1-\frac{1}{n}\right) = \left(\frac{n-1}{n}\sigma^2\right)$$

and thus the population SD $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ **underestimates** σ^2 in expectation. Replacing $\frac{1}{n}$ with $\frac{1}{n-1}$ corrects this:

$$E\left(\hat{\sigma}^2\right) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \sigma^2$$

We see that the expected squared sample SD (which is the sample variance) is the population variance (σ^2). This means the sample variance $\hat{\sigma}^2$ is the correct estimation to the population variance in average.

9.3 The T-statistic

For more realistic cases where both the population SD and the population mean are unknown, we can estimate the population SD σ by the sample SD

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

to get the estimated SE of the sample mean. This gives the T-statistic

$$T = \frac{\bar{X} - \mu_0}{\widehat{SE}_0(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}} = \sqrt{n} \times \frac{\bar{X} - \mu_0}{\hat{\sigma}}$$

Here the "hats" $\widehat{\cdot}$ over $SE_0(\cdot)$ and σ indicate "estimate of". **However**, due to the "extra randomness" in the denominator, this no longer behaves like a single draw from a standard normal box. Let's check its behaviour using simulations.

9.3.1 Simulations

Consider samples of size $n = 8$ from the "6-sided die" box

1	2	3	4	5	6

Lets compare the behaviour of the Z- and T-statistics via simulation.

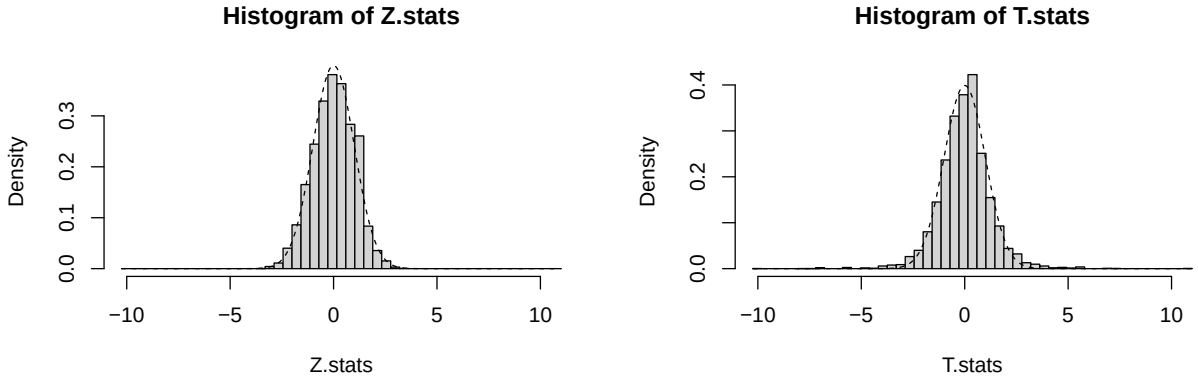
```
box = 1:6
mu = mean(box)
sig = sqrt(mean(box^2) - mean(box)^2)
n = 8
Z.stats = 0
T.stats = 0
for (i in 1:10000) {
    samp = sample(box, size = n, replace = T)
    m = mean(samp)  # sample mean
    sig.hat = sd(samp)  # sample SD
    Z.stats[i] = sqrt(n) * (m - mu)/sig
    T.stats[i] = sqrt(n) * (m - mu)/sig.hat
}
```

Here the loop `for(i in 1:10000)` iterates through the sequence from 1 to 10000.

- `samp = sample(box, size=n, replace=T)` draws a sample of size n

- `m = mean(samp)` is the sample mean
- `sig.hat = sd(samp)` is the sample SD
- `Z.stats[i] = sqrt(n)*(m-mu)/sig` computes the Z-statistic and assign it to the `i`th location of the list `Z.stats`
- `T.stats[i] = sqrt(n)*(m-mu)/sig.hat` computes the T-statistic and assign it to the `i`th location of the list `T.stats`

```
r = range(c(Z.stats, T.stats))
par(mfrow = c(1, 2))
br = seq(from = r[1], to = r[2], length = 50)
hist(Z.stats, breaks = br, pr = T, xlim = r)
curve(dnorm(x), n = 1001, lty = 2, add = T)
hist(T.stats, breaks = br, pr = T, xlim = r)
curve(dnorm(x), n = 1001, lty = 2, add = T)
```



9.3.2 Fatter tails

The general shape of the histograms are similar. The Z-statistics' one follows `dnorm(x)` pretty closely, **but** the T-statistics' one has **fatter tails**.

```
mean(abs(Z.stats) >= 1.96)
```

```
[1] 0.0476
```

```
mean(abs(T.stats) >= 1.96)
```

```
[1] 0.0942
```

Roughly 5% of the Z-statistics exceed 1.96 (in absolute value), *as we would expect*, whereas roughly 10% of the T-statistics exceed 1.96 (in absolute value). We cannot use `pnorm()` to get P-values any more.

10 Unknown Means: T-test

10.1 The T-statistic

The T-statistic simply replaces σ_0 with an estimate based on the sample:

$$T = \frac{\bar{X} - \mu_0}{\widehat{SE}_0(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

where

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}\,.$$

Due to the "extra randomness" in the denominator, this no longer behaves like a single draw from a standard normal box.

Suppose we have a box with unknown mean μ and wish to test the null hypothesis $H_0 \colon \mu = \mu_0$ and t is the observed value of T. We have to determine/approximate $P_{H_0}\{T \geq t\}$ in order to calculate the P-value. There are two ways to do this:

1. Impose extra assumptions and use theory. If the box is of a "special type", some theory tells us the distribution of T under H_0.
2. Use simulation with no extra assumptions imposed.

10.2 Student's t-distribution

W.S. Gossett was a statistician working for Guinness. He theoretically derived the distribution of T under H_0 when sampling from a normal population. Here we can consider that a **normal population** is an infinite, idealisation of a "box with a normal shape". Gossett wanted to publish his result in a Statistics journal. However, Guinness did not want him using his real name. So Gossett published his result under the pseudonym **Student** in the journal *Biometrika* (in 1908).

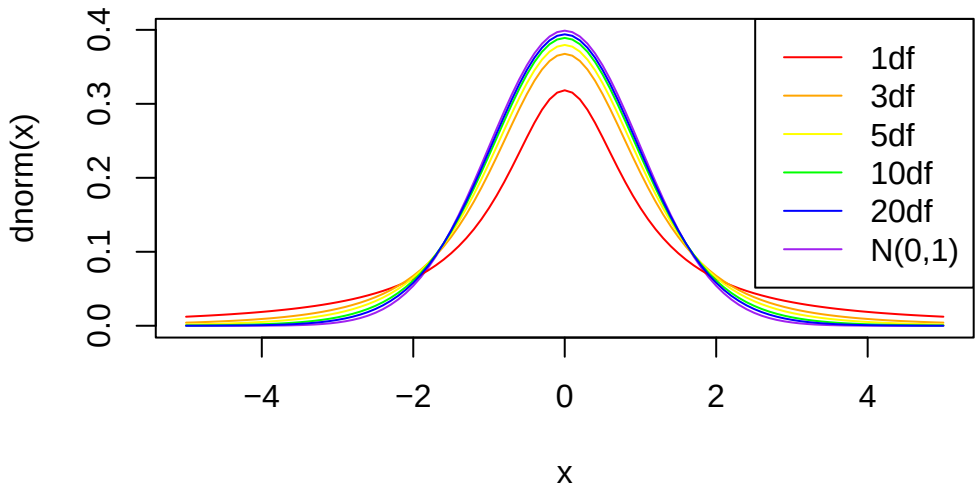
Here we are not interested in the actual form of Student's t-distribution. We only need to know how to use it in R.

- Its "density" is computed using `dt(x, df=`$n-1$`)`.
  - The distribution depends on a "degrees of freedom" parameter, which is set equal to $n-1$.
    - The tails get shorter for larger n.
  - This is the analogue of `dnorm()` for the standard normal distribution.
- Tail areas are computed using `pt(x, df=`$n-1$`,...)`.

  - This is the analogue of `pnorm(x,...)`.

- Percentage points may be obtained using e.g. `qt(0.975, df=`$n-1$`)`.
  - This is the analogue of `qnorm(0.975)`.

```
curve(dnorm(x), from = -5, to = 5, col = "purple")
curve(dt(x, df = 1), add = T, col = "red")
curve(dt(x, df = 3), add = T, col = "orange")
curve(dt(x, df = 5), add = T, col = "yellow")
curve(dt(x, df = 10), add = T, col = "green")
curve(dt(x, df = 20), add = T, col = "blue")
legend("topright", leg = c("1df", "3df", "5df", "10df", "20df", "N(0,1)"),
 ↪  lty = c(1, 1, 1, 1, 1, 1), col = c("red", "orange", "yellow",
    "green", "blue", "purple"))
```



One may ask what if the box/population is not normal?

Gosset/Student himself noted that the "normality assumption" was probably not critical. It seems then that an "nearly normal" box should also be OK to apply this theory. However, it is rather hard to define how close to normal is "nearly normal". Let's use some simulation to demonstate this.

6-sided die example

Consider samples of size $n = 8$ from the "6-sided die" box

1	2	3	4	5	6

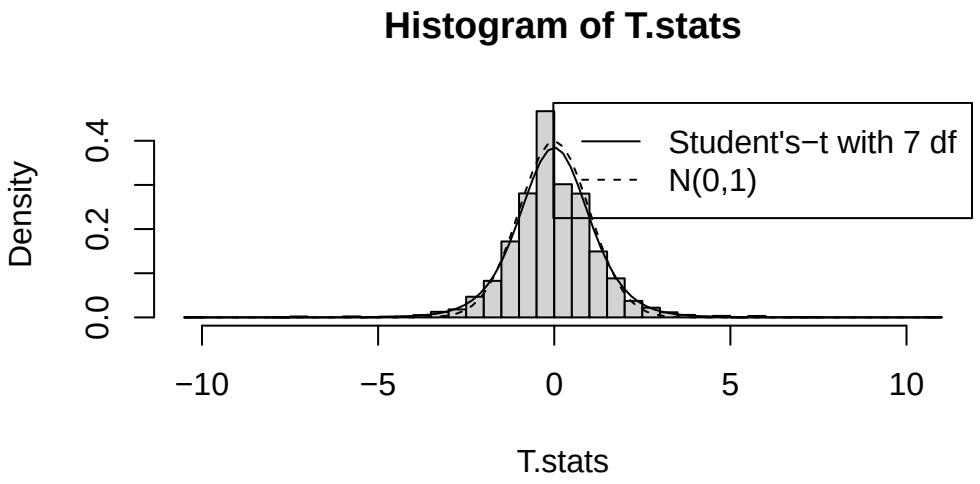
we have seen before. This box is certainly not "normal", but let's see anyway.

```
box = 1:6
mu = mean(box)
sig = sqrt(mean(box^2) - mean(box)^2)
n = 8
T.stats = 0
for (i in 1:10000) {
    samp = sample(box, size = n, replace = T)
    m = mean(samp)  # sample mean
    sig.hat = sd(samp)  # sample SD
    T.stats[i] = sqrt(n) * (m - mu)/sig.hat
}
```

```
n
```

```
[1] 8
```

```
hist(T.stats, pr = T, n = 40)
curve(dt(x, df = n - 1), add = T)
curve(dnorm(x), add = T, lty = 2)
legend("topright", leg = c(paste("Student's-t with", n - 1, "df"), "N(0,1)"),
 ↪  lty = c(1, 2))
```



The solid (Students-t with 7 d.f.) curve follows the histogram much better than the dashed standard normal curve. In particular, the upper 2.5% point for Students-t with 7 d.f. is given by

```r
qt(0.975, df = n - 1)
```

```
[1] 2.364624
```

```r
mean(abs(T.stats) >= 2.364)
```

```
[1] 0.0586
```

It is still a bit over 5%, but this t-distribution is doing a *much better* job (although it is not perfect.)

10.3 One-sample T-test

Let us reanalyse the `marks` data via a formal hypothesis test using the T-statistic. We want to know if there is evidence that the "population mean" mark μ of this exam will be different to 65. Suppose a group of 100 students take the draft exam, and obtain the marks below

```r
marks
```

```
  [1] 64 57 67 66 69 53 67 49 67 64 71 62 63 51 51 59 59 54 70 44 68 47 40 49 57 62 58 48 63
 [42] 67 53 41 72 85 52 54 84 57 81 79 58 45 69 59 68 64 57 70 64 55 66 45 73 68 78 54 65 49
 [83] 62 73 80 70 57 78 56 59 65 73 60 72 76 62 57 68 77 71
```

```r
mean(marks)
```

```
[1] 62.46
```

```r
sd(marks)
```

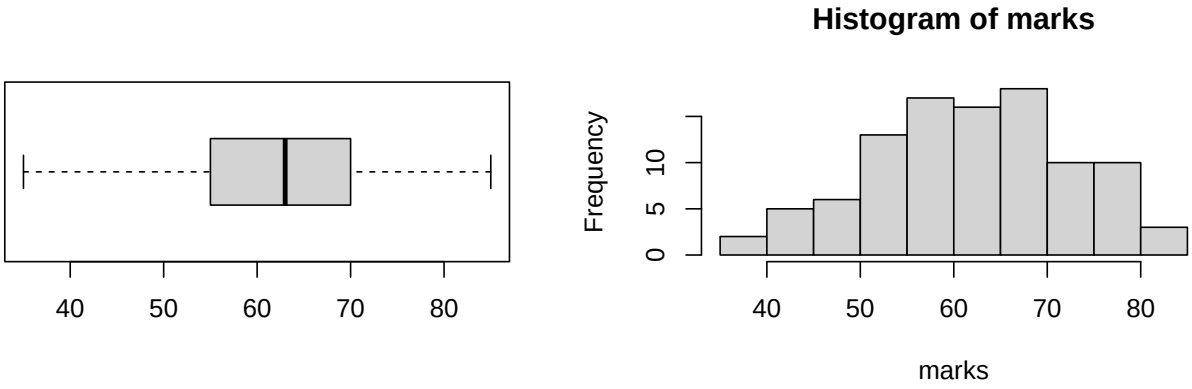
```
[1] 10.71053
```

10.3.1 HATPC

H **Null hypothesis**: $H_0 : \mu_0 = 65$ and **alternative hypothesis**: $H_1 : \mu_0 \neq 65$. A *two-sided* test is appropriate here, as the exam could be too easy or too hard.

A We should look at the sample to see if the "approximately normal box" assumption is reasonable.

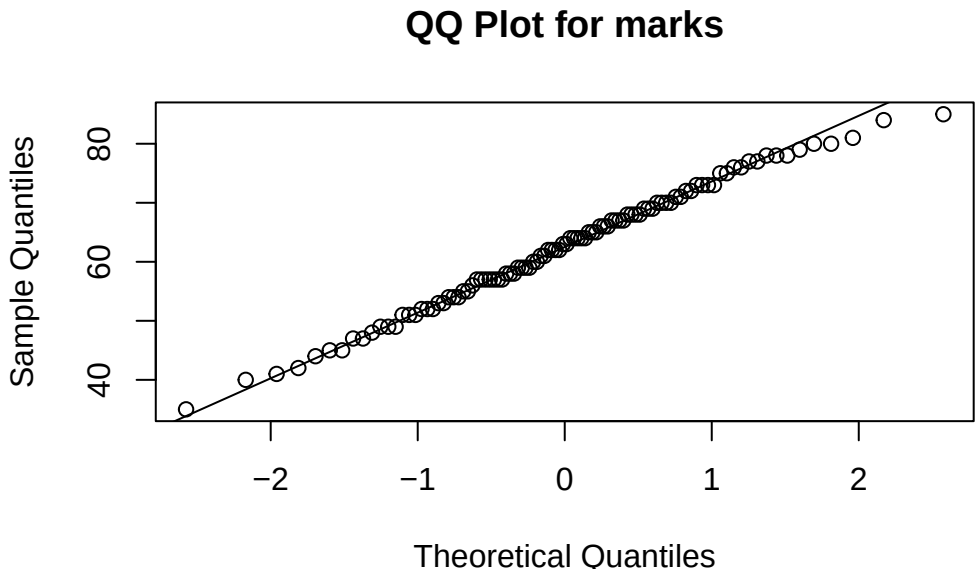
```
par(mfrow = c(1, 2))
boxplot(marks, horizontal = T)
hist(marks)
```



This boxplot is reasonably symmetric, with no outliers and the histogram looks quite bell-shaped, so it appears the "approximately normal box" assumption is reasonable.

We can also use the **quantile-quantile (QQ) plot** to determine if a data set is drawn from a distribution, e.g., a normal curve. A QQ plot is given by the function `qqnorm()`, we often pair it with the `qqline()`.

```
qqnorm(marks, main = "QQ Plot for marks")
qqline(marks)
```



A circle (x, y) on the QQ plot corresponds to the quantile of the data (y) plotted against the same quantile of the normal curve (x), while the QQ-line indicate what the QQ plot should look like if the data perfectly follow a normal curve. This way,

- Normally distributed data will have their points close to the line.
- The linearity of the points suggests that the marks are approximately normally distributed.

Note that even for data drawn from a normal population, we often don't have the points of QQ plot follow the QQ line exactly, as data are randomly drawn and subject to chance errors.

T We use the T-statistic

$$T = \frac{\bar{X} - 65}{\hat{\sigma}/\sqrt{n}}\,.$$

If H_0 is true this is (approximately) follows Student's t-distribution with $n-1 = 99$ degrees of freedom. This is a two-sided test, and hence small and large values of T-statistic argue against H_0. The observed value of T-statistic is

```
n = length(marks)
sig.hat = sd(marks)
t.stat = sqrt(n) * (mean(marks) - 65)/sig.hat
t.stat
```

```
[1] -2.371497
```

Note that the degrees of freedom of Student's t-distribution is $n-1$ which is exactly the effective dimension of sample deviations $X_j - \bar{X}$ (since $\sum_{j=1}^{n}(X_j - \bar{X}) = 0$, we lose one degree of freedom in the sample deviations).

P If T takes the value t, the P-value will be given by `2*pt(abs(t), df=99, lower.tail=F)` for a two-sided test.

```
2 * pt(abs(t.stat), df = 99, lower.tail = F)
```

```
[1] 0.01965109
```

C We conclude that the observed sample mean is significantly different from 65 at the $\alpha = 2\%$ level of significance, but *not* at the $\alpha = 1\%$ level. If we use the default 5% level of significance, we would reject H_0 and suggest that the exam should be moderated.

We can also obtain the critical regions of rejection:

- At $\alpha = 5\%$, $|T| > 1.984$
- At $\alpha = 2\%$, $|T| > 2.365$
- At $\alpha = 1\%$, $|T| > 2.626$

```
round(qt(1 - c(0.05, 0.02, 0.01)/2, df = 99), 3)
```

```
[1] 1.984 2.365 2.626
```

10.3.2 The function `t.test()`

Since performing a T-test is a common task, R has a "built-in" function which does all the necessary calculations in one step.

```
t.test(marks, mu = 65)
```

```
    One Sample t-test

data:  marks
t = -2.3715, df = 99, p-value = 0.01965
alternative hypothesis: true mean is not equal to 65
95 percent confidence interval:
 60.3348 64.5852
```

```
sample estimates:
mean of x
    62.46
```

Note that a two-sided test is performed by default. One-sided tests can be performed by adding arguments: `alternative="greater"` or `alternative="less"`.

10.3.3 Confidence Interval

Note the 95% confidence interval given in the `t.test()` output. Under Student's t-distribution, given the (unknown) true population mean μ,

$$\frac{\bar{X}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{99}$$

and so we can find multipliers $-\ell = u$ (by the symmetry of Student's t) such that

$$P\left\{\ell \leq \frac{\bar{X}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \leq u\right\} = 0.95\,.$$

where u is the upper 2.5% percentage point (97.5% quantile) taking the value below

```
round(qt(0.975, df = 99), 3)
```

```
[1] 1.984
```

Rearranging gives the probability statement

$$P\left\{\bar{X} - 1.984\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.984\frac{\hat{\sigma}}{\sqrt{n}}\right\} \approx 0.95\,.$$

This justifies the form of the interval where the interval is random whereas the population mean μ is fixed.

For the exam marks data, we have the 95% confidence interval

```
mean(marks) + c(-1, 1) * 1.984 * sig.hat/sqrt(100)
```

```
[1] 60.33503 64.58497
```

which is identical to that obtained by `t.test()`.

10.4 Bootstrap simulation

10.4.1 Skewed example

Let's modify the symmetric box

1	2	3	4	5	6

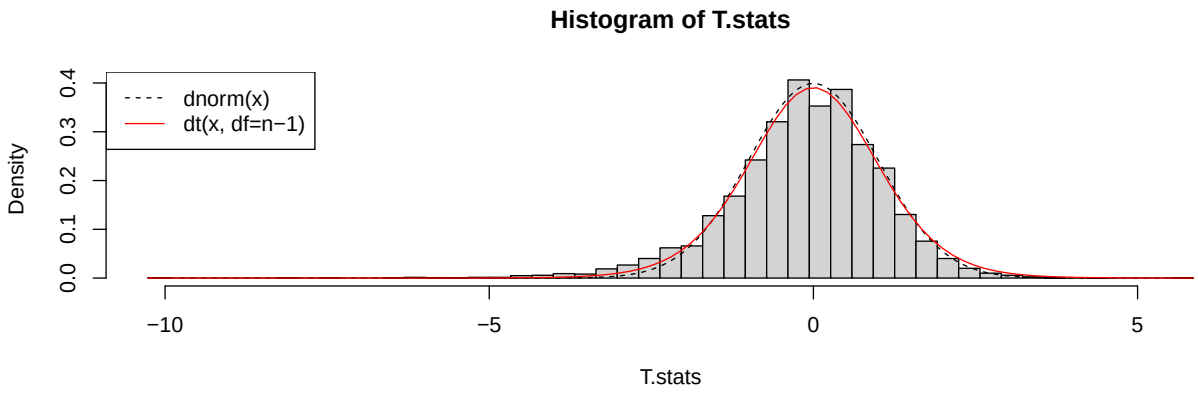
by changing the value of the last ticket to 8. The new box looks like:

1	2	3	4	5	8

.

We want to use a simulation to check if the behaviour of the T-statistic can be modelled by the Student's t-distribution in this case. Let's simulate 10,000 T-statistics from this box with sample size $n = 13$

```
box = c(1, 2, 3, 4, 5, 8)
mu = mean(box)
sig = popsd(box)
Z.stats = 0
T.stats = 0
n = 13
for (i in 1:10000) {
    samp = sample(box, size = n, replace = T)
    m = mean(samp)
    Z.stats[i] = sqrt(n) * (m - mu)/sig
    sig.hat = sd(samp)
    T.stats[i] = sqrt(n) * (m - mu)/sig.hat
}
```



```r
quantile(T.stats, prob = c(0.01, 0.02, 0.05, 0.95, 0.98, 0.99))
```

```
       1%        2%        5%       95%       98%       99%
-3.483140 -2.923581 -2.179174  1.589899  2.030729  2.362975
```

```r
qt(c(0.01, 0.02, 0.05, 0.95, 0.98, 0.99), df = 12)
```

```
[1] -2.680998 -2.302722 -1.782288  1.782288  2.302722  2.680998
```

The distribution of the T-statistics in samples from this box is clearly asymmetric. The quantiles of the simulated T-statistics and the quantiles of Student's t-distribution have clear discrepancies.

- We certainly cannot use even `pt()` to get P-values for a scenario like this.
- We need an alternative approach. We can try to approximate the distribution of the T-statistic by **approximating the box**.

10.4.2 Bootstrap principle

The theory behind the T-test suggests that if the box is "normal" (or "nearly normal") we can use e.g. `pt()` to get P-values using the T-statistic. But if the box is not "nearly normal", we want to use simulation to approximate the bahaviour of the T-statistic.

However, We may not know the *exact distribution* of T when H_0 is true, as the population (the true box model) is unknown. We can try simulating from a box model that is "reasonably close" to the "real" population. In this regard, we can **use the observed sample can be used as a "surrogate box"**, as this is the best we know about the actual population.

The idea of approximating a statistic's behaviour by using a "best guess" to the underlying population is known **the bootstrap principle**. It works because the T-statistic is "centred" and "normalised" using the population mean and sample SD (which is close to the population SD), so its behaviour does not change much if the underlying population changes a bit.

10.4.3 Skewed box example

Suppose we only see a sample `x` from the "unknown" box

1	2	3	4	5	8

```
box
```

```
[1] 1 2 3 4 5 8
```

```
n
```

```
[1] 13
```

```
x = sample(box, size = n, replace = T)
x
```

```
 [1] 3 8 3 2 2 8 3 5 4 8 8 1 2
```

We want to check how the T-statistic behave when we simulate from a surrogate box defined by this sample.

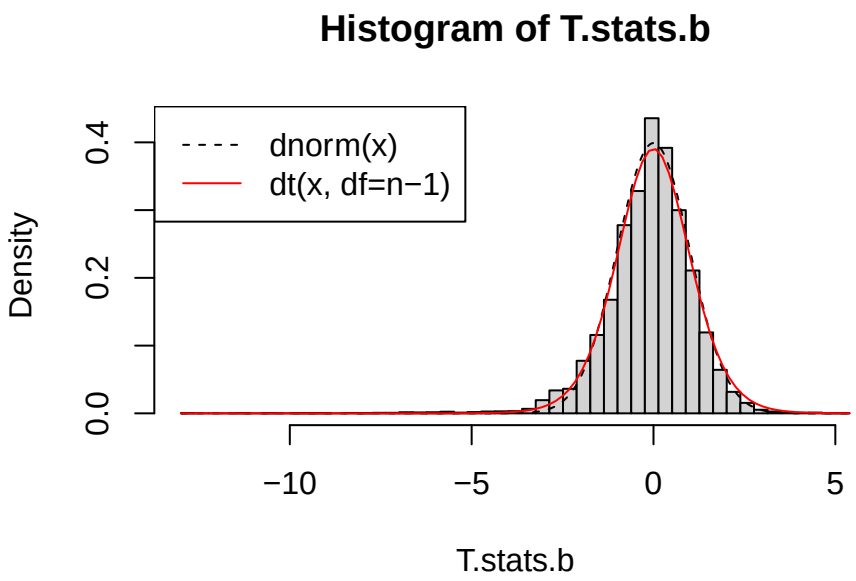
```
box.g = sort(x)  # g for 'guess'
box.g
```

```
 [1] 1 2 2 2 3 3 3 4 5 8 8 8 8
```

```
table(box.g)
```

```
box.g
1 2 3 4 5 8
1 3 3 1 1 4
```

```
mu.g = mean(box.g)
sig.g = popsd(box.g)
T.stats.b = 0  # b for 'bootstrap'
n = 13
for (i in 1:10000) {
    samp = sample(box.g, size = n, replace = T)
    m = mean(samp)
    sig.hat = sd(samp)  # for this 'best guess' population
    T.stats.b[i] = sqrt(n) * (m - mu.g)/sig.hat
}
```



The histogram of `T.stats.b` is not *quite* as asymmetric as that of `T.stats`, but it is a lot closer than the Student's-t curve.

```
quantile(T.stats, probs = c(0.01, 0.02, 0.05, 0.95, 0.98, 0.99))
```

```
       1%        2%        5%       95%       98%       99%
-3.483140 -2.923581 -2.179174  1.589899  2.030729  2.362975
```

```
quantile(T.stats.b, probs = c(0.01, 0.02, 0.05, 0.95, 0.98, 0.99))
```

```
       1%        2%        5%       95%       98%       99%
-3.310742 -2.805431 -2.038099  1.593375  2.044864  2.379248
```

```
qt(c(0.01, 0.02, 0.05, 0.95, 0.98, 0.99), df = n - 1)
```

```
[1] -2.680998 -2.302722 -1.782288  1.782288  2.302722  2.680998
```

Comparing the percentage points of `T.stats.b` with those of `T.stats`, they agree quite well, so we could try to use the percentage points from `T.stats.b` to approximate the distribution of T when H_0 is true (i.e. the percentage points of `T.stats`).

10.4.4 "Equal-tailed" confidence interval

We can also use the bootstrap principle to construct confidence intervals via simulation. Suppose we wish to construct an "equal-tailed" 95% confidence interval for μ. What we are really after a two values ℓ and u so that *whatever be* the value of μ,

$$P\left\{\ell \leq \frac{\bar{X}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \leq u\right\} \approx 0.95\,.$$

Since we no longer necessarily have symmetry, we may not have $\ell = -u$. Rearranging, we get

$$P\left\{\bar{X} - u\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X} - \ell\frac{\hat{\sigma}}{\sqrt{n}}\right\} \approx 0.95\,.$$

For observed $\bar{x}$ the interval is $\left[\bar{x} - u\frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} - \ell\frac{\hat{\sigma}}{\sqrt{n}}\right]$. Remember, ℓ is typically negative, u is typically positive.

We can use the lower and upper 2.5% percentage points of the *simulated* versions of T. For our skewed example, we would thus get

```
## this puts the values l and u in the right order!
u.l = quantile(T.stats.b, prob = c(0.975, 0.025))
u.l
```

```
    97.5%      2.5%
 1.949420 -2.617968
```

```
mean(x) - u.l * sig.hat/sqrt(n)
```

```
  97.5%    2.5%
3.09465 6.11697
```

Note that (like the Wilson interval) this is not necessarily symmetric about the "point estimate" $\bar{x}$.

10.5 A review of tests for unknown mean with an example

Using the exam mark example, we want to see how to apply three testing procedures: Z-test, T-test using Student's t, and T-test using bootstrap simulation.

We wanted to see if the average mark was **significantly different to 65**:

```
marks
```

```
  [1] 64 57 67 66 69 53 67 49 67 64 71 62 63 51 51 59 59 54 70 44 68 47 40 49 57 62 58 48 63
 [42] 67 53 41 72 85 52 54 84 57 81 79 58 45 69 59 68 64 57 70 64 55 66 45 73 68 78 54 65 49
 [83] 62 73 80 70 57 78 56 59 65 73 60 72 76 62 57 68 77 71
```

```
mean(marks)
```

```
[1] 62.46
```

10.5.1 Z-test

We first used a Z-test, assuming the SD of the "box" was known = 10. We can obtain the P-value:

```
sig0 = 10
n = length(marks)
SE = sig0/sqrt(n)
z = (mean(marks) - 65)/SE
z
```

```
[1] -2.54
```

```
2 * pnorm(abs(z), lower.tail = F)
```

```
[1] 0.01108525
```

10.5.2 T-test using Student's t

We then perform a T-test, where instead of **assuming** the box SD σ is known, we estimate it using the *sample SD* of the data and using the Student's t-distribution to calculate the P-value

```
sig.hat = sd(marks)
sig.hat
```

```
[1] 10.71053
```

```
est.SE = sig.hat/sqrt(n)
t = (mean(marks) - 65)/est.SE
t
```

```
[1] -2.371497
```

```
2 * pt(abs(t), df = n - 1, lower.tail = F)
```

```
[1] 0.01965109
```

Note the P-value is *slightly* larger than that of the Z-test.

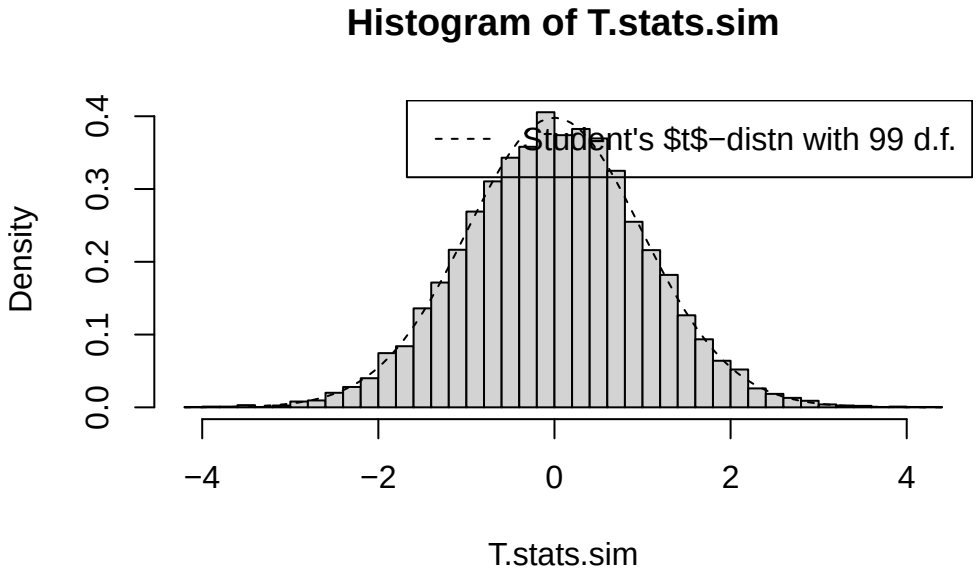
10.5.3 T-test using bootstrap simulation

Now we consider the bootstrap simulation. We repeatedly sample from the data (the surrogate box), and compute the value taken by the T-statistic

```
T.stats.sim = 0
for (i in 1:10000) {
    samp = sample(marks, size = n, replace = T)
    T.stats.sim[i] = (mean(samp) - mean(marks))/(sd(samp)/sqrt(n))
}
```

NOTE: `mean(marks)` is the "population mean" of the surrogate box, as we are trying to simulate the distribution of T-statistic here.

```
hist(T.stats.sim, n = 50, pr = T)  # n=50 here gives (approx.) no. bins
curve(dt(x, df = n - 1), add = T, lty = 2)
legend("topright", legend = c("Student's $t$-distn with 99 d.f."), lty = 2)
```



We can calculate the proportion of simulated T-statistic exceeding our observed T-statistic value of `-2.371` in absolute value:

```
mean(abs(T.stats.sim) > abs(t))
```

```
[1] 0.0209
```

The resulting P-value `0.0209` summarises how significant our observed T-statistic value is based on the simulation. Since this agrees very closely with the T-test result using Student's t, we feel comfortable in making the "approximately normal box" assumption that underlies the validity of the classical T-test.

Note that we are using the same T-statistic here, but a different testing distribution based on simulated T-statistic.

10.5.4 Simulation-based confidence interval

We firstly get the upper and lower 2.5% points from `T.stats.sim`:

```
u.l = quantile(T.stats.sim, prob = c(0.975, 0.025))
u.l
```

```
    97.5%      2.5%
 2.022866 -1.966097
```

These are then used to construct the interval $\left[\bar{X} - u\frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} - \ell\frac{\hat{\sigma}}{\sqrt{n}}\right]$:

```
mean(marks) - u.l * sd(marks)/sqrt(n)
```

```
   97.5%     2.5% 
60.29340 64.56579 
```

This is also very close to the confidence interval obtained using Student's t before.

10.6 Equal-tailed $(1-\alpha)$ confidence interval for unknown mean

We show here how to construct a confidence interval for unknown mean with a confidence level $(1-\alpha)$, e.g., $\alpha = 0.05$ for a 95% confidence interval. Consider the standard unit of the sample mean

$$T = \frac{\bar{X} - E(\bar{X})}{\widehat{SE}(\bar{X})} = \frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

where we have

- $\hat{\sigma} = \sigma$ if the population SD is known (Z-statistic); and
- $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ if the population SD is unknown (T-statistic).

We want to first find lower and upper multipliers ℓ and u such that $P\{T < \ell\} = P\{T > u\} = \frac{\alpha}{2}$. Their values depend on α and the distribution of the statistic T

- $-\ell = u =$`qnorm(`$1 - \frac{\alpha}{2}$`)` for standard normal (known population SD and assuming CLT);
- $-\ell = u =$`qt(`$1 - \frac{\alpha}{2}$`,df=n-1)` for Student's t (unknown population SD and assuming nearly normal box);
- $\ell =$`quantile(sim.stats, prob=`$\frac{\alpha}{2}$`)` and $u =$`quantile(sim.stats, prob=`$1 - \frac{\alpha}{2}$`)` for bootstrap simulation.

Given ℓ and u such that $P\{T < \ell\} = P\{T > u\} = \frac{\alpha}{2}$, we have

$$P\left\{\ell \le T = \frac{\bar{X} - E(\bar{X})}{\widehat{SE}(\bar{X})} \le u\right\} = 1 - \alpha.$$

Multiplying through by $-\widehat{SE}(\bar{X})$, we have

$$P\left\{-\ell\widehat{SE}(\bar{X}) \ge \mu - \bar{X} \ge -u\widehat{SE}(\bar{X})\right\} = 1 - \alpha.$$

Reversing inequalities (also reversing the positions of u and ℓ), this leads to

$$P\left\{-u\widehat{SE}(\bar{X}) \leq \mu - \bar{X} \leq -\ell\widehat{SE}(\bar{X})\right\} = 1-\alpha.$$

Adding $\bar{X}$, we obtain the $(1-\alpha)$-confidence interval

$$P\left\{\bar{X} - u\widehat{SE}(\bar{X}) \leq \mu \leq \bar{X} - \ell\widehat{SE}(\bar{X})\right\} = 1-\alpha.$$

This way, $(1-\alpha)$ proportions of the intervals

$$\left(\bar{X} - u\widehat{SE}(\bar{X}), \bar{X} - \ell\widehat{SE}(\bar{X})\right),$$

which depend on the sample $\bar{X}$, covers the true population mean μ. We can easily apply it to different confidence levels $1-\alpha = 95\%, 98\%, 99\%, ...$ Note that this doesn't apply to Wilson's confidence interval, as $SE(\bar{X})$ depends on μ for the 0-1 box.

11 Unknown Means: Two-Sample T-test

11.1 Comparing two (sample) means

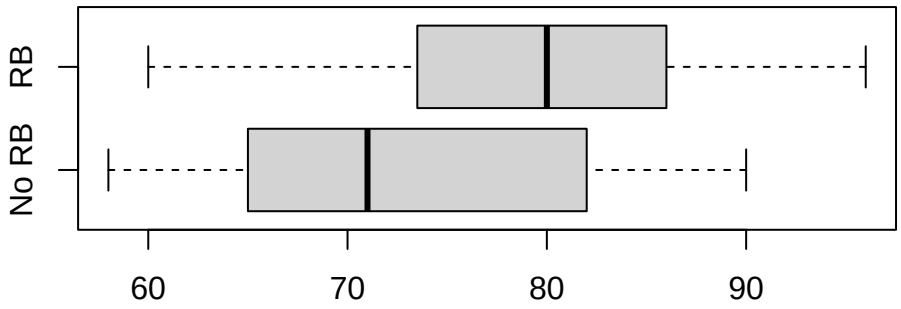
Previously, we learned how to test hypotheses about the mean of a single population using the t-test. The learning objective of this chapter is to extend that idea to testing the difference between population means

11.1.1 Case study: red bull

Red Bull is an energy drink advertised to "give you wings". We are interested in testing the medical effects of drinking a Red Bull, particularly in terms of heart rates. Researchers collected the following data on heart rates (beats per minute), for 2 independent groups of Sydney students, collected 20 minutes after the 'RedBull' group had drunk a 250ml cold can of Red Bull.

No Red Bull	84	76	68	80	62	58	74	84	68	90	82	64	65	66
Red Bull	72	88	72	88	76	75	84	80	60	96	80	84	-	-

```
No_RB <- c(84, 76, 68, 80, 62, 58, 74, 84, 68, 90, 82, 64, 65, 66)
RB <- c(72, 88, 72, 88, 76, 75, 84, 80, 60, 96, 80, 84)
boxplot(No_RB, RB, names = c("No RB", "RB"), horizontal = T)
```



By comparing the boxplots, the Red Bull group seems to have a higher heart rate. We want to apply the T-test to figure out if the apparent difference significant.

11.1.2 Two-box model

As a starting point, we want to first understand the behaviour of the difference between sample means, so later we can derive the test statistic.

We can model the two groups as samples taken from two separate boxes (independently of each other). That is, we model the "No Red Bull" group as a random sample $X_1, \dots, X_m$ taken with replacement from a box with mean μ_X and SD σ_X. Similarly, we model the "Red Bull" group as a random sample $Y_1, \dots, Y_n$ taken (with replacement) from a box with mean μ_Y and SD σ_Y.

We *really* wish to make a statement about the **population** mean difference μ_X and μ_Y, based on the **sample** mean difference $\bar{X} - \bar{Y}$. So we need to work out the expected value and the standard error of the **sample** mean difference $\bar{X} - \bar{Y}$ based on

- $E(\bar{X}) = \mu_X$, $SE(\bar{X}) = \frac{\sigma_X}{\sqrt{m}}$; and
- $E(\bar{Y}) = \mu_Y$, $SE(\bar{Y}) = \frac{\sigma_Y}{\sqrt{n}}$.

Note that the sample mean difference difference can be expressed as $\bar{X} - \bar{Y} = \bar{X} + (-\bar{Y})$, which is the summation of random draws $\bar{X}$ and $-\bar{Y}$. Since we have

- $E(-\bar{Y}) = -E(\bar{Y})$ and $SE(-\bar{Y}) = SE(\bar{Y})$;

we can use results of Topic 6 to conclude

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) + E(-\bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$$

and most importantly

$$SE(\bar{X} - \bar{Y})^2 = SE(\bar{X})^2 + SE(-\bar{Y})^2 = \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\,.$$

That is

$$SE(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}\,.$$

11.1.3 Two-sample Test Statistics

We wish to test the null hypothesis $H_0 \colon \mu_X = \mu_Y$. If the two box SDs σ_X and σ_Y were known, we could test H_0 using the Z-statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim N(0, 1)$$

if H_0 true. In general, σ_X and σ_Y are both unknown. In this case we have three options:

- Assuming $\sigma_X = \sigma_Y = \sigma$ is the same in both boxes and normality of both boxes, we can apply the **Classical Two-Sample T-test**
- Without the equal SD assumption $\sigma_X = \sigma_Y = \sigma$ but assuming normality of both boxes: we can apply the **Welch Test**.
- Without the normality assumption, we can apply the bootstrap simulation.

11.2 The Classical Two-Sample T-test

In some cases it is reasonable to assume $\sigma_X = \sigma_Y = \sigma$, which is often called an **equal variances** assumption, i.e. $\sigma_X^2 = \sigma_Y^2$ since for "normal populations" (idealised, infinite boxes whose histograms follow a normal curve exactly) it is more common to refer to variances, i.e. squared SDs. Then the SE of the sample mean difference may be written as

$$SE(\bar{X} - \bar{Y}) = \sigma\sqrt{\frac{1}{m} + \frac{1}{n}}\,.$$

In this case, *if* it is also assumed the boxes are (approx.) normal-shaped, a special "combined" or "pooled" estimate $\hat{\sigma}_p$ of the common σ is used. Then Student's theory can be applied to show the statistic

$$T = \frac{\bar{X} - \bar{Y}}{\hat{\sigma}_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

i.e. has Student's-t distribution with $m + n - 2$ degrees of freedom.

11.2.1 The pooled estimate $\hat{\sigma}$

The form of the pooled estimate of σ is given by

$$\hat{\sigma}_p = \sqrt{\frac{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2}{m + n - 2}} = \sqrt{\frac{(m-1)\hat{\sigma}_X^2 + (n-1)\hat{\sigma}_Y^2}{m + n - 2}}\,.$$

Written this way, we see $\hat{\sigma}_p^2$ is a **weighted average** of σ_X^2 and σ_Y^2. The bigger sample gets more weight, so that the estimate from the larger sample is somehow "more trustworthy".

A quick way to remember the denominator $m + n - 2$ (which is the same as the degrees of freedom in Student's t) is that we lose one degree of freedom in each esimation of the sample SD. In the remaining of this section, we will show why this "pooled estimate" is the correct choice in expectation (not for assessment).

Recall that each sample variance (squared sample SD) estimates σ^2 "on average", in that

$$E(\hat{\sigma}_X^2) = E\left(\frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2\right) = \sigma^2$$

and so

$$E\left((m-1)\hat{\sigma}_X^2\right) = E\left(\sum_{i=1}^{m}(X_i-\bar{X})^2\right) = (m-1)\sigma^2$$

Similarly we have

$$E\left((n-1)\hat{\sigma}_Y^2\right) = E\left(\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right) = (n-1)\sigma^2$$

Then the numerator inside the $\sqrt{\cdot}$ of the pooled estimate $\hat{\sigma}_p$ has

$$\begin{aligned} E\left(\sum_{i=1}^{m}(X_i-\bar{X})^2+\sum_{j=1}^{n}(Y_j-\bar{Y})^2\right) &= E\left((m-1)\hat{\sigma}_X^2\right)+E\left((n-1)\hat{\sigma}_Y^2\right) \\ &= (m-1)\sigma^2+(n-1)\sigma^2 \\ &= (m+n-2)\sigma^2\,. \end{aligned}$$

Dividing through by $m+n-2$ we get

$$E\left(\hat{\sigma}_p^2\right) = \sigma^2\,,$$

so $\hat{\sigma}_p^2$ shares the "on-target on average" property that $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$ have. As in the one-sample T-test, the denominator in the estimate of σ^2 is also the degrees of freedom. We "lose one degree of freedom" in each set of sample deviations (for each sample mean we estimate). This way, degrees of freedom is then total sample size, minus 2.

11.2.2 Red Bull example

Based on the boxplots, we see that each sample looks reasonably symmetric and the spreads are similar. It may therefore be reasonable to assume we have approximately normal boxes with a common SD. We can further check the sample SDs.

```
sd(No_RB)
```

```
[1] 9.848579
```

```
sd(RB)
```

```
[1] 9.452833
```

The sample SDs are similar. The following code calculates the pooled estimate of the common SD.

```
m = length(No_RB)
n = length(RB)
m
```

```
[1] 14
```

```
n
```

```
[1] 12
```

```
numer = (m - 1) * (sd(No_RB)^2) + (n - 1) * (sd(RB)^2)
denom = m + n - 2
sig.hat.p = sqrt(numer/denom)
sig.hat.p
```

```
[1] 9.669206
```

We therefore compute the observed value taken by the (Classical) Two-Sample T-statistic: ::: {.cell}

```
est.SE = sig.hat.p * sqrt((1/m) + (1/n))
est.SE
```

```
[1] 3.803845
```

```
mean.diff = mean(No_RB) - mean(RB)
mean.diff
```

```
[1] -6.654762
```

```
stat = mean.diff/est.SE
stat
```

```
[1] -1.749483
```

:::

As originally phrased, i.e. "is the apparent difference significant?", it is (strictly speaking) two-sided test, so the two-sided p-value is thus give by

```
2 * pt(abs(stat), df = m + n - 2, lower.tail = F)
```

```
[1] 0.09298616
```

This is not small, so the apparent difference is **not** significant at the default 5% level of significance.

Of course, the `t.test()` function can do all of this in one line, in which we must supply the `var.equal=T` parameter:

```
t.test(No_RB, RB, var.equal = T)
```

```

    Two Sample t-test

data:  No_RB and RB
t = -1.7495, df = 24, p-value = 0.09299
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.505513   1.195989
sample estimates:
mean of x mean of y
 72.92857  79.58333
```

Note that the **confidence interval** given here is obtained in the familar way. Once an appropriate percentage point from the t_{m+n-2} distribution is obtained using `qt()`, we can obtain the confidence interval for the population mean difference:

```
qt(0.975, df = m + n - 2)
```

```
[1] 2.063899
```

```
mean.diff + c(-1, 1) * qt(0.975, df = m + n - 2) * est.SE
```

```
[1] -14.505513   1.195989
```

11.3 The Welch Test

11.3.1 Relaxing the equal variance assumption

If we want to apply Student's theory directly, we have to assume $\sigma_X = \sigma_Y$. However, this assumption is somewhat restrictive, as the two samples may have very different spreads. In a more general setting, an "obvious" approach would be to instead consider the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_X^2}{m} + \frac{\hat{\sigma}_Y^2}{n}}},$$

which just plugs in the two sample SD estimates in for σ_X and σ_Y. However, with this new test statistic, we cannot simply apply Student's t-distribution with $m + n - 2$ degrees of freedom to model its behaviour.

In 1947 (some time after Student's paper) B. L. Welch found that the statistic behaved **approximately** like a Students-t distribution whose degrees of freedom was a complicated function of m, n, σ_X and σ_Y, where the population SDs σ_X and σ_Y can be estimated using sample SDs $\hat{\sigma}_X$ and $\hat{\sigma}_Y$.

This gives the name **Welch Test**, where the test statistic T has a Student's-t distribution with a **data-dependent degrees of freedom**. We skip the detail of the data-dependent degrees of freedom, and rely on R to calculate it.

11.3.2 Default two-sample `t.test()`

It turns out Welch's procedure works very well, i.e. The "approximate" p-values returned have nice properties and rejection rates are in line with the desired false-alarm rate when simulating from normal boxes. So that R uses the Welch test as the default two-sample T-test:

```
## note: data-dependent d.f. close to Classical (which was 24 d.f.)
t.test(No_RB, RB)
```

```
    Welch Two Sample t-test

data:  No_RB and RB
t = -1.7552, df = 23.66, p-value = 0.09216
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.485720   1.176197
sample estimates:
```

```
mean of x mean of y
 72.92857  79.58333
```

The main difference between the Welch Test and the default two-sample t-test is how the degrees of freedom is calculated. For the Welch test, it is calculated with a complicated formula using the standard deviation of the two samples. For the two-sample t-test the degrees of freedom is calculated using the total sample size, $df = m + n - 2$.

11.4 Bootstrap simulation

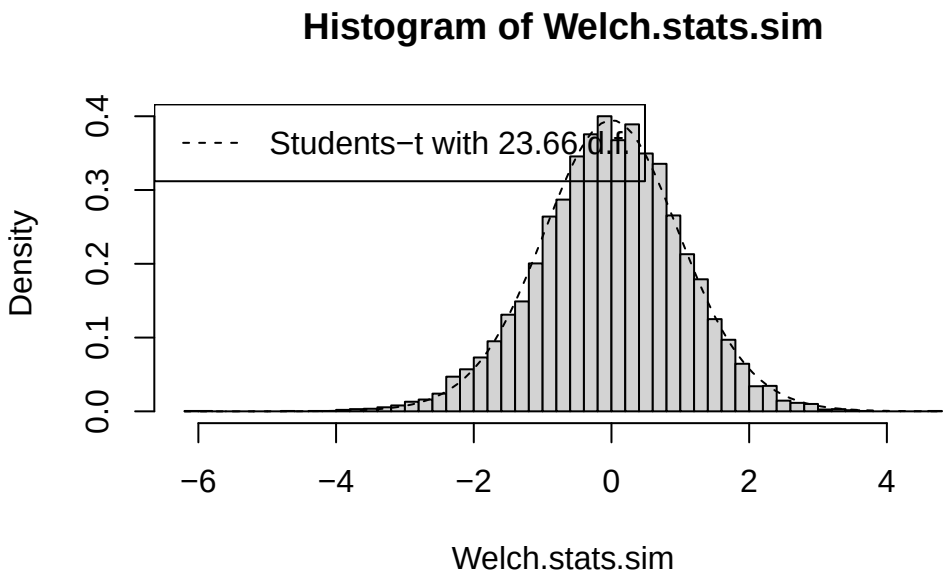
The Welch test does not assume $\sigma_X = \sigma_Y$, but it still assumes the two boxes are "approximately normal". We can carry out bootstrap simulation to model the behaviour of the Welch statistic if we are uncomfortable making this assumption.

11.4.1 Simulate the Welch statistic

Similar to the one-sample case, we can use two samples as surrogate boxes to represent the true but unknown populations. However, to simulate the test statistic under the null hypothesis (equal population means), we need to ensure that the surrogate boxes have the same mean. One way to achieve this is by centering each sample by subtracting its own mean, so that both surrogate boxes have a mean of zero.

```
No_RB.g = No_RB - mean(No_RB)  # both sorrogate boxes have
RB.g = RB - mean(RB)  # mean zero
Welch.stats.sim = 0
for (i in 1:10000) {
    samp.x = sample(No_RB.g, size = m, replace = T)
    samp.y = sample(RB.g, size = n, replace = T)
    est.SE = sqrt((sd(samp.x)^2)/m + (sd(samp.y)^2)/n)
    Welch.stats.sim[i] = (mean(samp.x) - mean(samp.y))/est.SE
}
```

```
hist(Welch.stats.sim, n = 50, freq = F)
curve(dt(x, df = 23.66), add = T, lty = 2)  # data-dependent d.f. from
 ↪  original sample
legend("topleft", legend = c("Students-t with 23.66 d.f."), lty = 2)
```



For the red bull example, the histogram of the simulated Welch statistic is a little left-skewed. This is maybe due to the slight departure from normal curve in the `No_RB` sample.

11.4.2 Two-sided P-value by simulation

Using the observed Welch statistic

```
mean.diff = mean(No_RB) - mean(RB)
est.SE = sqrt((sd(No_RB)^2)/m + (sd(RB)^2)/n)
stat = mean.diff/est.SE
stat
```

```
[1] -1.755237
```

and the simulated Welch statistic above, we can calculate the simulation-based P-value

```
mean(abs(Welch.stats.sim) >= abs(stat))
```

```
[1] 0.0931
```

which is very close to the earlier P-values (both Classical and Welch).

11.4.3 Confidence interval by simulation

We use the simulated values in `Welch.stats.sim` to approximate the "true distribution" of the Welch statistic when $\mu_X = \mu_Y$ (under H_0):

```
u.l = quantile(Welch.stats.sim, prob = c(0.975, 0.025))
u.l
```

```
    97.5%      2.5%
 1.957070 -2.197924
```

Note that these are not the same magnitude indicates the slight lack of symmetry. Similar to the one sample case, this in turn defines the simulation-based confidence interval for the population mean difference

```
mean.diff - u.l * est.SE
```

```
     97.5%       2.5%
-14.074746   1.678392
```

The interval is quite close to those obtained by (both versions of) `t.test()`.

11.5 Paired (two-sample) T-test

Another common scenario is where we have two samples of data **but** we have n individuals or "sampling units" and there is a reading in each sample associated with each individual. Thus, corresponding values in the two samples are "paired". In this case, the two samples are **not independent** and we cannot compare the two sample means using the methods we have already seen.

11.5.1 Student's sleep data

In his original paper W.S. Gossett ("Student") demonstrated his new testing technique on various examples. One was where 10 patients tried two different drugs designed to increase sleep time. An image of the table is linked to below.

Additional hours' sleep gained by the use of hyoscyamine hydrobromide.

Patient	1 (Dextro-)	2 (Laevo-)	Difference (2–1)
1.	+ ·7	+ 1·9	+ 1·2
2.	− 1·6	+ ·8	+ 2·4
3.	− ·2	+ 1·1	+ 1·3
4.	− 1·2	+ ·1	+ 1·3
5.	− 1	− ·1	0
6.	+ 3·4	+ 4·4	+ 1·0
7.	+ 3·7	+ 5·5	+ 1·8
8.	+ ·8	+ 1·6	+ ·8
9.	0	+ 4·6	+ 4·6
10.	+ 2·0	+ 3·4	+ 1·4
	Mean + ·75	Mean + 2·33	Mean + 1·58
	S. D. 1·70	S. D. 1·90	S. D. 1·17

from "The Probable Error of a Mean", Student(1908), Biometrika

Note that there is a typographical error: the −1 for patient 5 under "Dextro-" should be −0.1. Also, the SDs presented under the table are actually computed using the population SD, not the sample SD. Below is the corrected version of the first column of the data (labelled "Dextro-") and the the second column labelled "Laevo-" in R.

```
dextro = c(0.7, -1.6, -0.2, -1.2, -0.1, 3.4, 3.7, 0.8, 0, 2)
laevo = c(1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4)
diff = laevo - dextro
sleep = data.frame(dextro, laevo, diff)
```

We can use the R comand `apply(..., 2, mean)` to apply the function `mean()` to each column; and similarly `apply(..., 2, sd)` gives column `sd()`s:

```
apply(sleep, 2, mean)
```

```
dextro  laevo   diff
  0.75   2.33   1.58
```

```
apply(sleep, 2, sd)
```

```
  dextro    laevo     diff
1.789010 2.002249 1.229995
```

Note here that `sd(diff)` is much smaller than it would be if the two samples were independent. Let's verify this by randomly scrambling the pairing.

The command `sample(dextro)` randomly shuffles `dextro`:

```
dextro
```

```
 [1]  0.7 -1.6 -0.2 -1.2 -0.1  3.4  3.7  0.8  0.0  2.0
```

```
sample(dextro)
```

```
 [1]  0.7  0.8 -1.6 -0.2  3.7  0.0 -1.2  3.4 -0.1  2.0
```

After scrambling `dextro`, we eliminate any potential pairing information between the two variables. As a result, if `dextro` and `laevo` were originally paired, the standard deviation (SD) of their differences after scrambling will behave as if the variables are unpaired (i.e., independent). In the code below, the sample SD of the differences after random scrambling is:

```
sd(laevo - sample(dextro))
```

```
[1] 2.530613
```

which is what we would expect `sd(diff)` to be if the two sample were independent. Obviously, the SD of the differnece of the original data is much smaller than that after scrambling, suggesting that the data are not independent.

11.5.2 Paired T-test

To assess whether the sample mean difference is significantly different to zero *or* greater than zero *or* less than zero, we simply perform the appropriate **one-sample T-test** on the **sample differences**.

There are two ways to perform a paired T-test using `t.test()`. We can either manually taking differences

```
t.test(laevo - dextro)
```

```
    One Sample t-test

data:  laevo - dextro
t = 4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.7001142 2.4598858
sample estimates:
mean of x
     1.58
```

or using `t.test(..., paired=T)`:

```
t.test(laevo, dextro, paired = T)
```

```
    Paired t-test

data:  laevo and dextro
t = 4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.7001142 2.4598858
sample estimates:
mean difference
           1.58
```

One-sided tests can be done using `alternative=greater` or `alternative=less` as usual, but be careful of the order.

12 Chi-squared tests

12.1 Chi-squared test of goodness of fit

12.1.1 Suspicious dice

In games of chance, fairness is paramount. But what happens when a gambler is suspected of using a loaded die? In this example, we examine whether a six-sided die rolls each number with equal probability, or if the gambler's die is unfairly biased.

A record of the last 60 rolls has been kept, and we begin by summarising the data. If the die is fair, we would expect each face to appear roughly the same number of times. However, if certain numbers appear significantly more or less often than expected, this could suggest the die is loaded.

To formally test this, we introduce the goodness of fit test, which compares the observed frequencies of outcomes to what we would expect under the assumption of fairness. This statistical approach allows us to determine whether the differences we observe are due to random variation or if they provide evidence of an unfair die.

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

- Let's summarise these:

```
table(die)
```

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

- These counts should be "roughly equal", but these look a bit **too** unequal.
- Or do they?

12.1.2 Box model for (possibly loaded) die

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

1	2	3	4	5	6

A single random draw X from this box has the distribution

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

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

1	2	3	3	4	4	5	6

giving the distribution

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

12.1.3 Goodness of fit test

For any vector $p = (p_1, \dots, p_6)$ of (rational) probabilities (so each $p_j \geq 0$ and $p_1 + \cdots + p_6 = 1$) we can imagine a box with a certain number of each ticket, so the proportion of tickets with integer j is p_j.

- We would like to test the hypothesis $H_0 \colon p_1 = \cdots = p_6 = \frac{1}{6}$.
- We are interested in **any alternative that is not** H_0.
  - That is, $p_j \neq \frac{1}{6}$ for at least one $j = 1, \dots, 6$.
  - In brief the alternative is $H_1 \colon$ **not** H_0.

This is an example of a **goodness of fit test**. Since each value 1,2,…,6 is equally likely, after 60 draws we would **expect** to get 10 of each:

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

The table below compares observed and expected frequencies:

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

Due to random sampling the observed are not **exactly** equal to the expected, we anticipate some "small" discrepancies. We aim to figure out *how different* these have to be before we get suspicious.

12.1.4 General formulation of Pearson's χ^2 statistic

The testing problem can be generalised as follows. Suppose we have data $X_1, \ldots, X_n$ only taking k distinct categories, modelled as a random sample taken with replacement from a box. We may assume the possible values are the positive integers $1, 2, \ldots, k$, each integer labels a category.

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

We summarise the data to **observed frequencies**: $O_j =$ number of X_is equalling j (for $j = 1, \ldots, k$). We compare these to the corresponding **expected frequencies**: $E_j = np_{0j}$, i.e. the number of X_is equalling j we would expect **under** H_0.

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

A "foundational" paper in modern statistics was by Karl Pearson in 1900. He considered the **Pearson's χ^2 statistic**

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

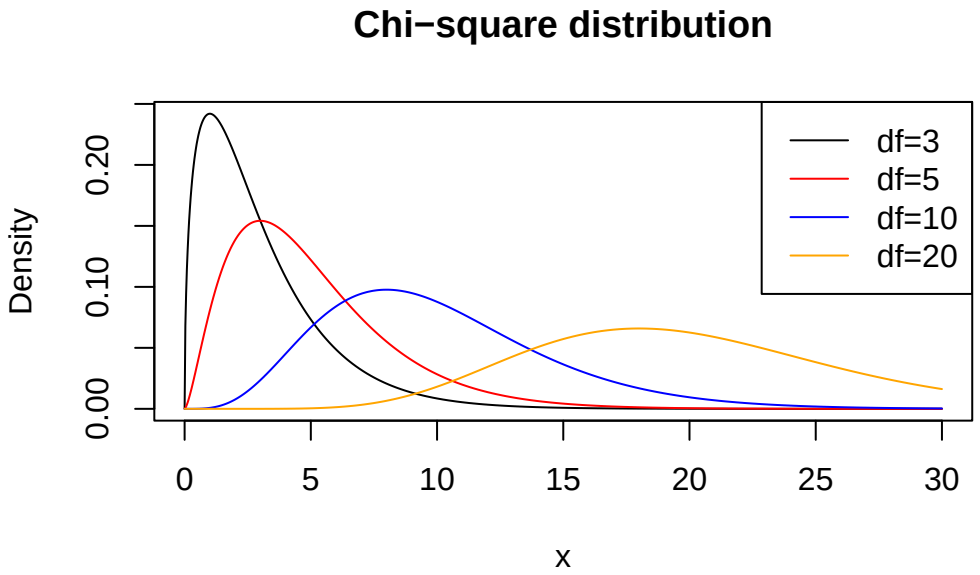
where for categories with larger E_i, the "error" $O_i - E_i$ tends to be bigger, so we divide $(O_i - E_i)^2$ by E_i makes each term "comparable". Pearson argued that under H_0, for "large n",

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

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

12.1.5 The χ^2_d distribution

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



Suppose we have k categories, and the observed value of Pearson's statisic is t_{obs}. The **larger** t_{obs}, the more evidence against H_0, and thus this is a one-sided test. Then, the (approximate) p-value is given by the area under the χ^2_{k-1} curve to the **right** of t_{obs}. This is (approximately) the chance of at least as much evidence against H_0 as was observed, assuming H_0 true.

12.1.6 Our dice example

H We have the null hypothesis ($H_0 : p_0 = (\frac{1}{6}, \ldots, \frac{1}{6})$) (indicating the die is fair); and the alternative hypothesis ($H_1 :$) at least one of $p_{0j} \neq \frac{1}{6}, j = 1, \ldots, 6$, indicating the die is loaded.

A We will discuss this later.

T The degrees of freedom is $6 - 1 = 5$, so χ^2_5 is the test distribution. For the record of results from the die

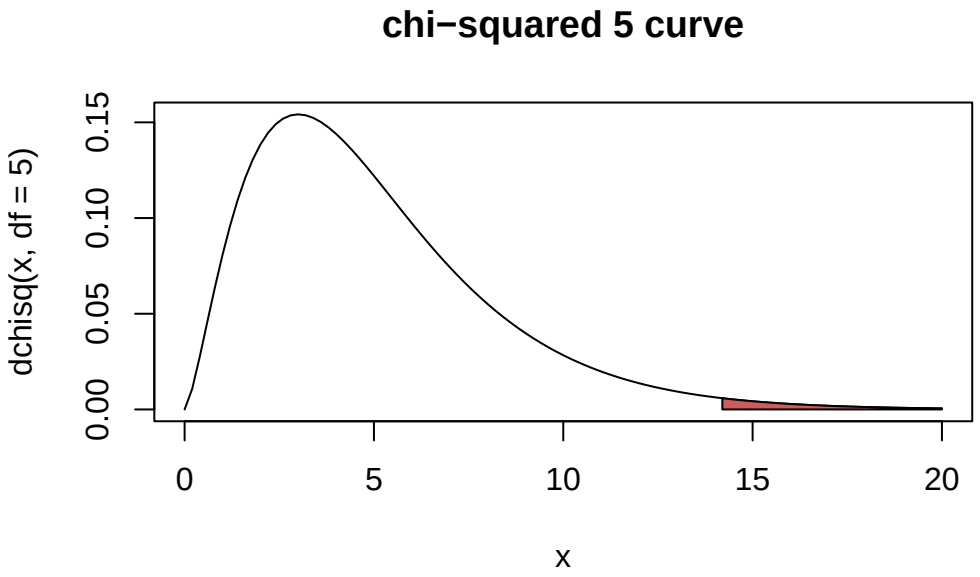
```
Oi = table(die)
Ei = rep(10, 6)
rbind(Ei, Oi)
```

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

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

```
[1] 14.2
```

So the observed value is $t_{\text{obs}} = 14.2$. Is the value $\chi^2 = 14.2$ plausible under H_0?



P Remember, we need the *upper tail* (because larger values of t_{obs} is more evidence against H_0).

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

```
[1] 0.01438768
```

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

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

```
    Chi-squared test for given probabilities

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

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

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

```
    Chi-squared test for given probabilities

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

C This is a rather small P-value. This provides evidence against the null hypothesis of "all 6 sides equally likely" using a false alarm rate of, for example, 2%, which indirectly suggests die is loaded.

12.1.7 Assumptions required

The χ^2_{k-1} distribution is a "large-sample approximation" to the exact sampling distribution of Pearson's statistic when H_0 is true. It may not be a good approximation if - *either* the sample size n is not very large - *or* some categories have very small hypothesised probabilities.

A "rule of thumb" for checking both assumptions is that if all expected frequencies E_j are at least 5, the χ^2_{k-1} approximation should be reasonably accurate.

The R function `chisq.test()` prints a warning if this condition is violated: ::: {.cell}

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

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

```
    Chi-squared test for given probabilities

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

:::

12.1.8 Equivalence with z-test

When there are only two categories, we can model the population by a 0-1 box. In this case, we can draw a connection between the chi-squared test and a **two-sided Z-test for a proportion**. Let's consider a box containing only $\boxed{0}$s and $\boxed{1}$s, let p denote the proportion of $\boxed{1}$s in the box. Suppose we have a random sample $X_1, \dots, X_n$ taken with replacement from the box. We aim to test the hypothesis $H_0 : p = p_0$ against the two-sided $H_1 : p \neq p_0$.

Using a Z-test, we have the statistic

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

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

We can also view this as a χ^2-test.

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

Then, we have Pearson's statistic

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

So we can see that in this case the chi-squared test is **exactly** a two-sided Z-test, because a squared $N(0,1)$ is exactly a χ^2_1.

For example, an upper 5% percentage point for χ^2_1 is

```
qchisq(0.95, df = 1)
```

```
[1] 3.841459
```

This is exactly the square of the upper 2.5% percentage point for $N(0,1)$

```
qnorm(0.975)
```

```
[1] 1.959964
```

```
qnorm(0.975)^2
```

```
[1] 3.841459
```

So we would reject (at the 5% level) if T exceeds 1.96^2, or equivalently, if Z^2 exceeds 1.96^2. This is exactly the same as $|Z|$ exceeds 1.96 (i.e. if $Z > 1.96$ **or** $Z < -1.96$). This shows that the chi-squared test may be viewed as a **generalisation** of the **two-sided** Z-test for a proportion, to a box with more than 2 different values in it.

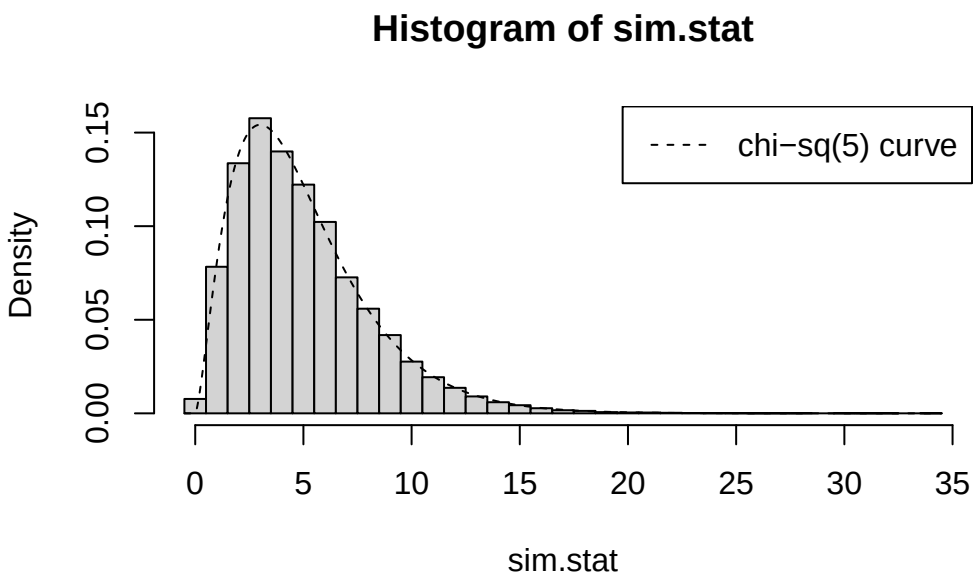
12.2 Simulation (test of goodness of fit)

12.2.1 Using simulation: the dice example

We can approximate the sampling distribution of the test statistic by simulating an appropriate (approximate if necessary) box model. This is straightforward for chi-squared tests, as H_0 completely specifies the distribution of X_i, and hence the box.

Using the loaded die example, the histogram of simulated Pearsion's statisitc under H_0 has nice agreement with the χ^2_5 curve (see below).

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



Recall the observed Pearson's statistic

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

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

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

```
[1] 14.2
```

We can obtain the P-value using the simulated test distribution (note that it's a one-sided test):

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

```
[1] 0.0139
```

Compare this to the P-value obtained using the theoretical χ^2_5

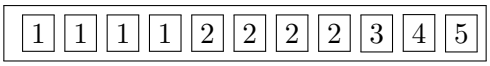
```
chisq.test(Oi)$p.value
```

```
[1] 0.01438768
```

We can see the simulation-based P-value is close to that obtained using the χ^2_5 approximation, which is not suprising as we have the neccessary assumption satisfied for using the χ^2_5 approximation.

12.2.2 Small expected frequencies

Let's consider another example where we the assumptions are not reasonable. Suppose we draw a sample of size $n = 10$ from the box (with 11 tickets)



We want to check how Pearson's statistic behaves when we test $H_0 : p_0 = (\frac{4}{11}, \frac{4}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11})$. Note that H_0 is true in this example.

The expected frequencies are then **all** < 5:

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

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

So we suspect the χ^2_4 approximation may not be so good. Sure enough, `chisq.test()` tells us this: suppose we draw the sample

```
samp
```

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

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

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

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

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

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

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

```

    Chi-squared test for given probabilities

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

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

12.2.3 Using simulation

Let's now apply the bootstrap simulation. We first simulate the box under H_0

```
box = c(1, 1, 1, 1, 2, 2, 2, 2, 3, 4, 5)
sim.stat = 0
for (i in 1:100000) {
    sim.obs = sample(box, size = n, replace = T)
    freqs = tabulate(sim.obs, nbins = 5)
    sim.stat[i] = suppressWarnings(chisq.test(freqs, p = p0)$stat)
    # without supressWarnings() we get 10000 'approximation may be incorrect'
     ↪  warnings
}
```

And then compare the quantiles of simulated Pearson's statistics with the theoretical ones

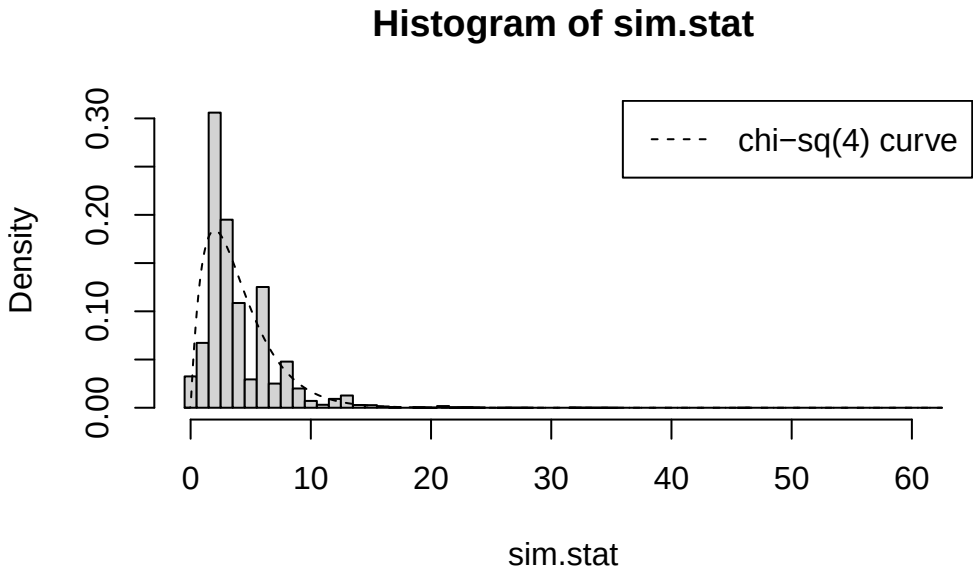
```
quantile(sim.stat, probs = c(0.95, 0.98, 0.99))
```

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

```
qchisq(c(0.95, 0.98, 0.99), df = 4)
```

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

We observe that the upper 2% and 1% points are bigger than χ^2_4 would suggest.



In addition, the distribution is multi-modal and there are more large values than χ^2_4 would suggest. This clearly indicates that the χ^2_4 is not approporiate here.

Using our earlier observed Pearson's statistic, we can use the bootstrap simulation to get a simulation-based P-value of

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

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

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

```
[1] 0.04037
```

This procedure works so well, so it's a built-in option in R. We can carry out bootstrap simulation by using `chisq.test(..., simulate=T)`, which gives a similar result:

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

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

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

```
[1] 0.04037
```

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

```
    Chi-squared test for given probabilities with simulated p-value (based on 100000 replicat

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

Here, `B = ...` specifues the number of samples used in the simulation.

12.3 Chi-squared tests with estimated parameters

In Pearson's test, the observed frequency O of each category is compared with expected frequency $E = np$, where p is the probability of "landing" in that category. We test **goodness of fit**, i.e. a null hypothesis H_0 specifies probabilities for each category, whereas the alternative hypothesis is H_1: **not** H_0.

Pearson's statistic T is the sum of $\frac{(O-E)^2}{E}$ over all categories. When H_0 is true, T has an approximate χ^2_d distribution. Next, we will see the degrees of freedom takes the form of

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

which possibly depends on some parameters.

In the test of good of fit, we had k categories and a vector of probabilities $p = (p_1, \dots, p_k)$ for each category. Then under the **full model** (where any probability vector is allowed), we

have k parameters **but** only $k-1$ of these are **free** since they add to 1. That is if we know $p_1, \ldots, p_{k-1}$, then

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

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

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

12.4 Chi-squared test of independence

The Chi-squared test can also be applied to **two way tables**, where one of the possible aims is to test if two categorical variables are independent.

Let's consider the following example, where data give biological sex (the row categorical variable) and the handedness (the column categorical variable) for 2,237 people:

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

We want to find out if the data suggest any evidence against that the handedness and the gender are independent. Note that, if they are independent, there is no difference in handedness between men and women.

12.4.1 Pearson's statistic

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

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

and so is now a "double sum", where r is the number of row categories and c is the number of column categories. Here,

- O_{ij} is the observed frequency in row i, column j
- E_{ij} is the expected frequency in row i, column j **under the null hypothesis**.

We needt to formulate the full model, the null hypothesis, and the expected frequencies, the E_{ij}s.

12.4.2 Full model

Given a two way table for two categorical variables, the full model contains $r \times c$ different categories with unconstrained probabilities in the table

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

We thus have $rc-1$ free parameters under the full model. Here we introduce the "dot" notation for sums. For example,

- $p_{\bullet 1} = \sum_{i=1}^{r} p_{i1}$ (sum over a rows for a specified column)
- $p_{1\bullet} = \sum_{j=1}^{c} p_{1j}$ (sum over columns for a specified row)

The row sums $p_{\bullet j}$ gives the marginal probabilities for every column. That is, the chance of landing in j-th column category of the table. E.g., handedness in this example. The column sums $p_{i\bullet}$ gives the marginal probabilities for for every row. That is, the chance of landing in i-th row category of the table. E.g., biological sex in this example.

12.4.3 Null hypothesis

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

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

Under H_0, the probability of each cell is

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

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

To figure out those probabilities p_{ij}'s under H_0, we need to **estimate** those marginal probabilities $P\{\text{in Row } i\}$ and $P\{\text{in Col } j\}$ using observed frequencies:

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

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

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

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

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

12.4.5 Expected frequencies

Since the expected frequencies under null hypothesis is $E_{ij} = np_{i\bullet}p_{\bullet j}$

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

Using the estimated marginal probabilities, $p_{i\bullet}$s **and** the $p_{\bullet j}$s, the expected frequencies are given as

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

12.4.6 Degrees of freedom

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

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

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

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

12.4.7 Handedness example

- Observed frequencies:

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

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

```
colnames(Oij) = c("RH", "LH", "Ambi")  # assign column names
Oij
```

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

- Row and column sums:

```
R = rowSums(Oij)  # sum over all the columns for each row
R
```

```
  men women
 1067  1170
```

```
C = colSums(Oij)  # sum over all the rows for each column
C
```

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

- Pearson's statistic, note that the assumption $E_{ij} \geq 5$ holds here.

```
n = sum(Oij)  # sample size
n
```

```
[1] 2237
```

```
Eij = outer(R, C, FUN = "*")/n  # expected frequencies
Eij
```

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

```
stat = sum(((Oij - Eij)^2)/Eij)
round(stat, 3)
```

```
[1] 11.806
```

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

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

```
[1] 2
```

```
round(pchisq(stat, df = d, lower.tail = F), 6)  # one-sided test
```

```
[1] 0.002731
```

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

```
Oij
```

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

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

```
    Pearson's Chi-squared test

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

12.4.8 Using simulation

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

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

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

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

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

12.4.9 Standardised residuals (not for assessment)

If Pearson's statistic is large, we might ask: "which cells are making it large"? Recall that dividing by E_{ij} makes each squared discrepancy "comparable". Thus we can compare the terms to see which are contributing the most to the large statistic. Better still, we can preserve the sign to see which are "more" or "less" than expected.

```
SR = (Oij - Eij)/sqrt(Eij)
SR
```

```
              RH        LH      Ambi
men   -0.7071859  1.539125  1.818199
women  0.6753406 -1.469817 -1.736324
```

Note that the sum of squares of the terms in `SR` is Pearson's statistic. There seem to be more `LH` and `Ambi` than expected for men, and less for women.