T-tests

Decisions with Data | Inference for means

STAT5002

The University of Sydney

Apr 2025

THE UNIVERSITY OF SYDNEY

Decisions with Data

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

Topic 10: The t-test

Topic 11: The two-sample test

Topic 12: χ^2-test

Outline

The T-statistic

Student's t-distribution

P-values and confidences intervals via simulation

Review

The T-statistic

The Z-statistic

- The Z-statistic, which measures how many SEs the sample mean $\bar{X}$ is away from the expected value μ_0:

$$Z = \frac{\bar{X} - \mu_0}{SE_0(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}}$$

  can only be computed when $SE_0(\bar{X})$ (SE of $\bar{X}$ under H_0) is **known**.
- Due to the Central Limit Theorem, so long as n is large enough, Z will behave like a single draw from a standard normal box **if H_0 is true**.
- What should we do when $SE_0(\bar{X})$ is **unknown**?
  - ➠ **Estimate it** using sample SD.

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

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

is the sample SD.

- 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
    - ➠ How does it behave?

Simulations

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

- Lets compare the behaviour of the Z- and T-statistics via simulation (using the for loop)

For loop

```
samp.sums = 0  # initialise a list of sample sums
for (i in 1:10) {
    samp = sample(1:6, size = i, rep = T)
    samp.sums[i] = sum(samp)
}
```

- `for(i in 1:10)` iterates through the sequence from 1 to 10. Each iteration is indexed by `i`.
- Within the curly brackets `{` and `}`, it executes a set of statements:
  - `samp = sample(box, size=i, replace=T)`:
    - sampling with replacement to form a sample of size `i`
  - `samp.sums[i] = sum(samp)`:
    - computes the sample sum and assign it to the `i`th location of the list `samp.sums`

```
samp.sums
```

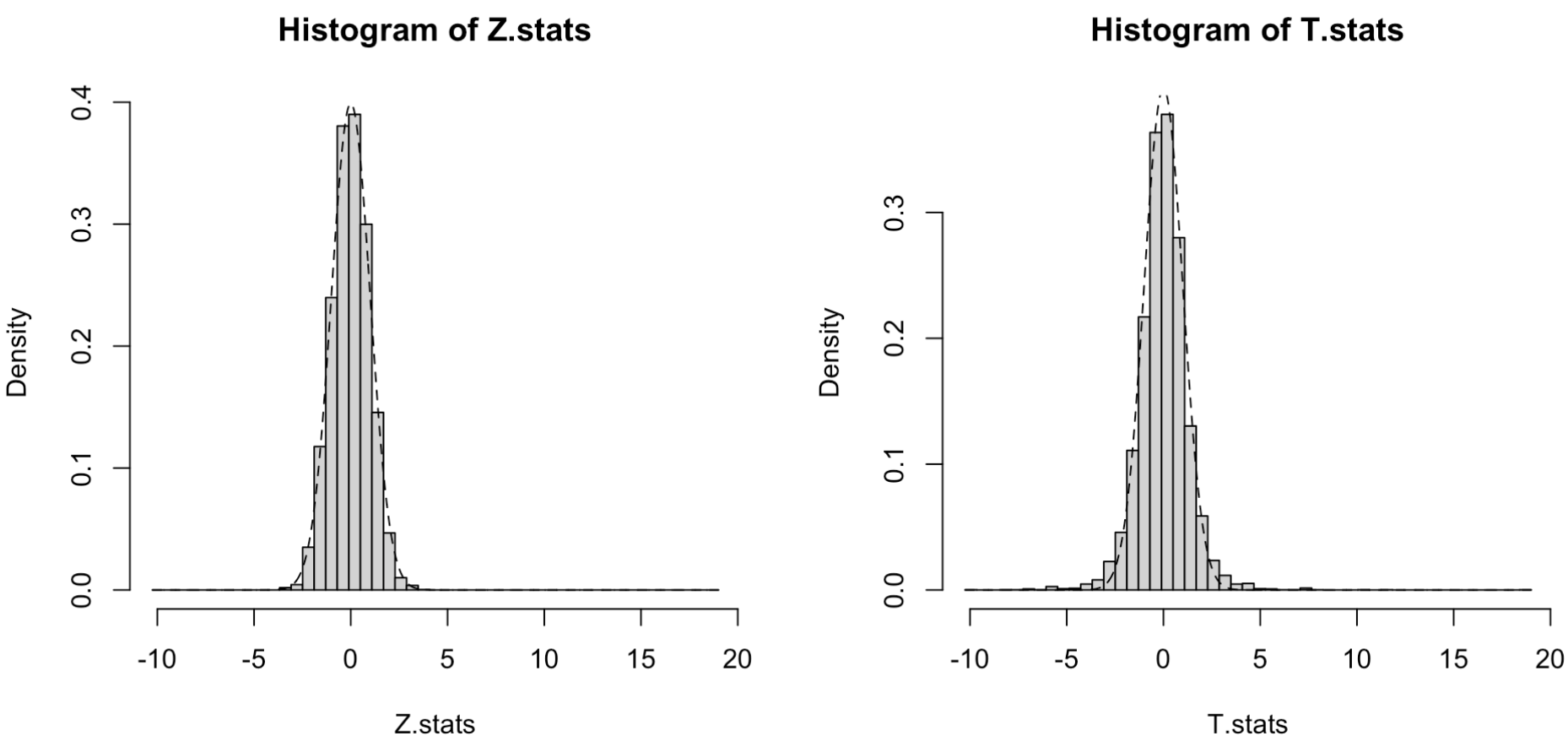
```
[1]  5  7  5  9 14 31 24 42 34 36
```

Simuilated Z- and T-statistics

```r
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
}
```

- `for(i in 1:10000)`: iterates through the sequence from 1 to 10000.
  - `samp = sample(box, size=n, replace=T)`: draw a sample of size n
  - `m = mean(samp)`: sample mean
  - `sig.hat = sd(samp)`: 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
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)
```



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**.

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

```
[1] 0.0472
```

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

```
[1] 0.0958
```

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

Expression(s) for the P-value

- Suppose we have a box with unknown mean μ and unknown SD and wish to test the null hypothesis $H_0: \mu = \mu_0$.
- We consider the T-statistic

$$T = \frac{\bar{X} - E_0(\bar{X})}{\widehat{SE}(\bar{X})} = \frac{\bar{X} - \mu_0}{\frac{1}{\sqrt{n}} SD(X)} = \frac{\bar{X} - \mu_0}{\frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}}.$$

- Suppose that t is the observed value of T.
- If we knew the distribution of T when $H_0: \mu = \mu_0$, then...
  - If the alternative was $H_1: \mu > \mu_0$, the P-value would be $P_{H_0}\{T \geq t\}$;
  - If the alternative was $H_1: \mu < \mu_0$, the P-value would be $P_{H_0}\{T \leq t\}$;
  - If the alternative was $H_1: \mu \neq \mu_0$, the P-value would be

$$P_{H_0}\{|T| \geq |t|\} = P_{H_0}\{T \geq |t|\} + P_{H_0}\{T \leq -|t|\}.$$

- How do we determine/approximate $P_{H_0}\{T \geq t\}$, etc.?

Two options

There are two ways to get a P-value based on the T-statistic:

- 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;
  - ➟ If it is “reasonable” to assume the box is of this special type, we can use this option.
- 2. Use simulation:
  - ➟ **Note**: no extra assumptions!

Student's t-distribution

Sampling from a normal population

- W.S. Gossett was a statistician working for Guinness (the brewer!).
- He theoretically derived the distribution of T under H_0 when sampling from a normal population.
  - ➠ A **normal population** is an infinite, idealisation of a “box with a normal shape”.
- He wanted to publish his result in a Statistics journal, but his employer did not want him using his real name.
  - ➠ The brewing industry was obviously pretty competitive at the time.
- He thus published his result under the pseudonym **Student** in the journal *Biometrika* (in 1908).

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

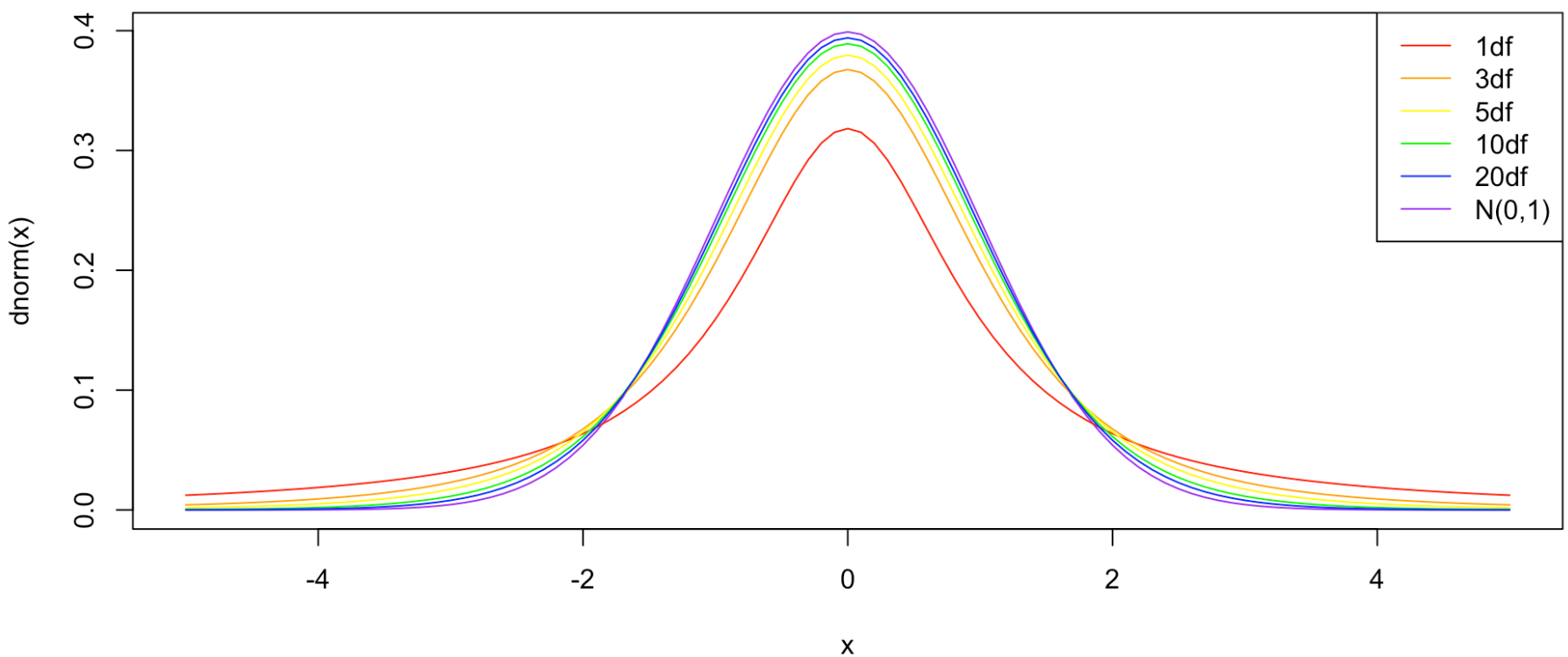
Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.

Student's t-distribution

- The mathematical form is not important, but the t-distribution's curve may be computed using R.
- The "density" is computed using `dt(x, df=`$n-1$`)`.
  - The distribution depends on a "degrees of freedom" (d.f.) 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, ...)`.
- Quantiles may be obtained using e.g. `qt(0.975, df=`$n-1$`)`.
  - This is the analogue of `qnorm(0.975)`.

```r
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
    "green", "blue", "purple"))
```



What if the box/population is not normal?

- Gosset/Student himself noted that the "normality assumption" was probably not critical:

of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here

- It seems then that an "nearly normal" box should also be OK to apply this theory.
- But how close to normal is "nearly normal"?

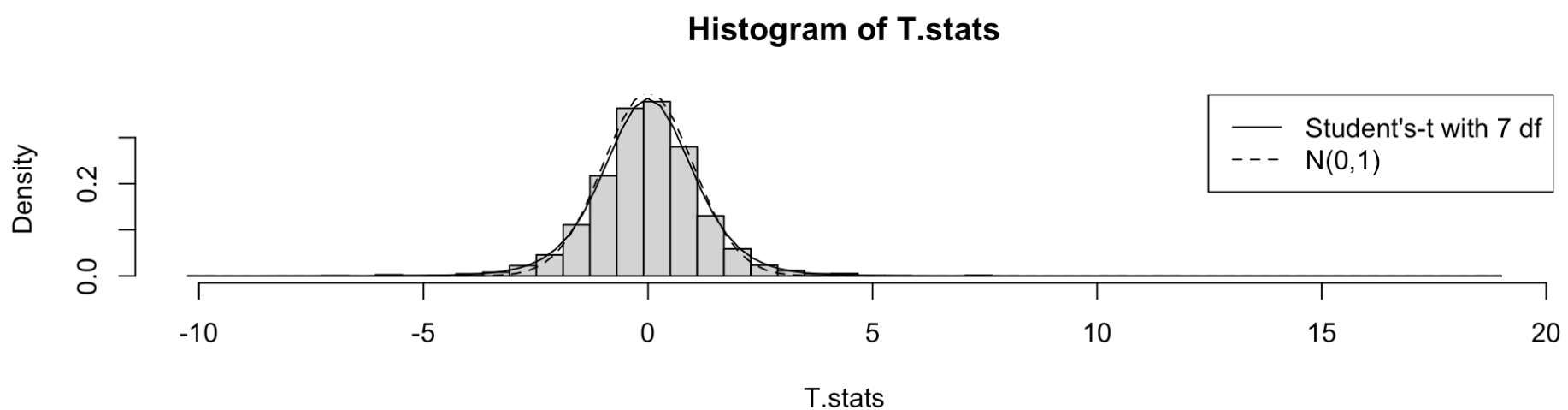
6-sided die example

- Let's see how closely this works for our $n = 8$ "6-sided die box" example.
  - ➠ This box is certainly not "normal", but let's see anyway.

```
1 n
```

```
[1] 8
```

```
1 r = range(T.stats)
2 br = seq(from = r[1], to = r[2], length = 50)
3 hist(T.stats, freq = F, breaks = br, n = 40)
4 curve(dt(x, df = n - 1), add = T)
5 curve(dnorm(x), add = T, lty = 2)
6 legend("topright", leg = c(paste("Student's-t with", n - 1, "df"), "N(0,1)"), lty = c(1, 2))
```



Looks better, but…

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

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

```
[1] 2.364624
```

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

```
[1] 0.0588
```

- It is still a bit over 5%, but this t-distribution is doing a *much better* job (although it is not perfect.)
- Compared to the quantiles of normal curve, the quantiles of Students-t is closer to those of the simulated T-statistics.

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

```
       1%        2%        5%       95%       98%       99%
-3.326337 -2.679033 -1.901612  1.888396  2.645751  3.415650
```

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

```
[1] -2.997952 -2.516752 -1.894579  1.894579  2.516752  2.997952
```

```
1 qnorm(c(0.01, 0.02, 0.05, 0.95, 0.98, 0.99))
```

```
[1] -2.326348 -2.053749 -1.644854  1.644854  2.053749  2.326348
```

One sample T-test

Exam marks

- Let us reanalyse the `marks` data via a formal hypothesis test
  - ➠ 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

```
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 52 64 42 78 60 57
61 47 75 58 51 35
 [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 76 52 77 65 75 80
73 70 61 55 69 66
 [83] 62 73 80 70 57 78 56 59 65 73 60 72 76 62 57 68 77 71
```

```
mean(marks)
```

```
[1] 62.46
```

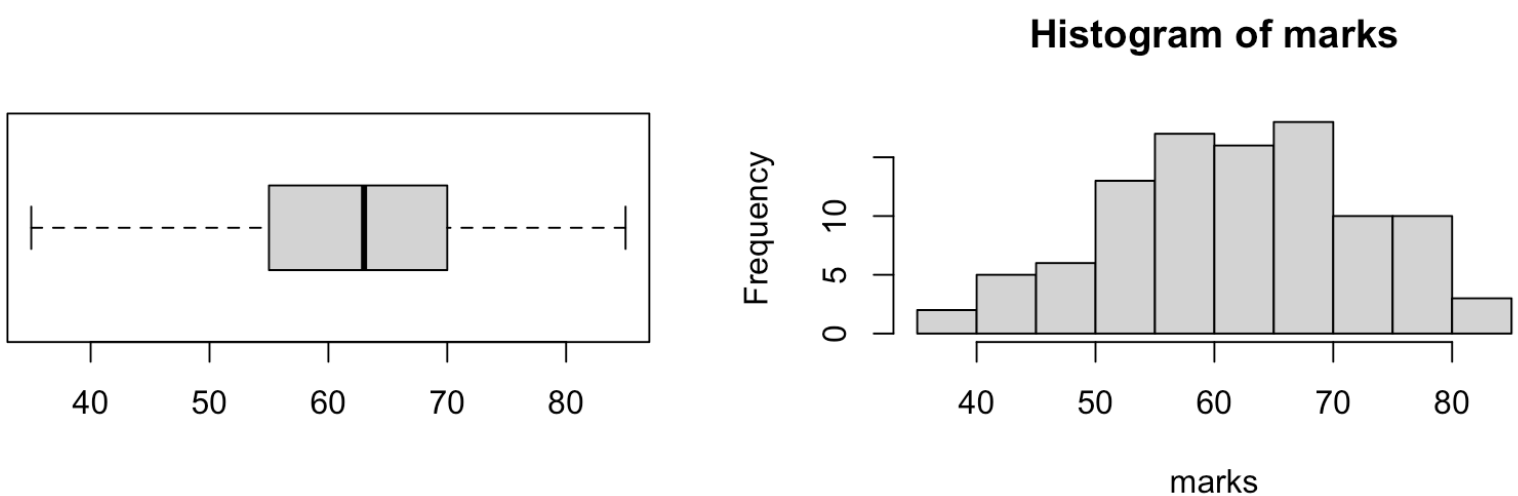
```
sd(marks)
```

```
[1] 10.71053
```

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.

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

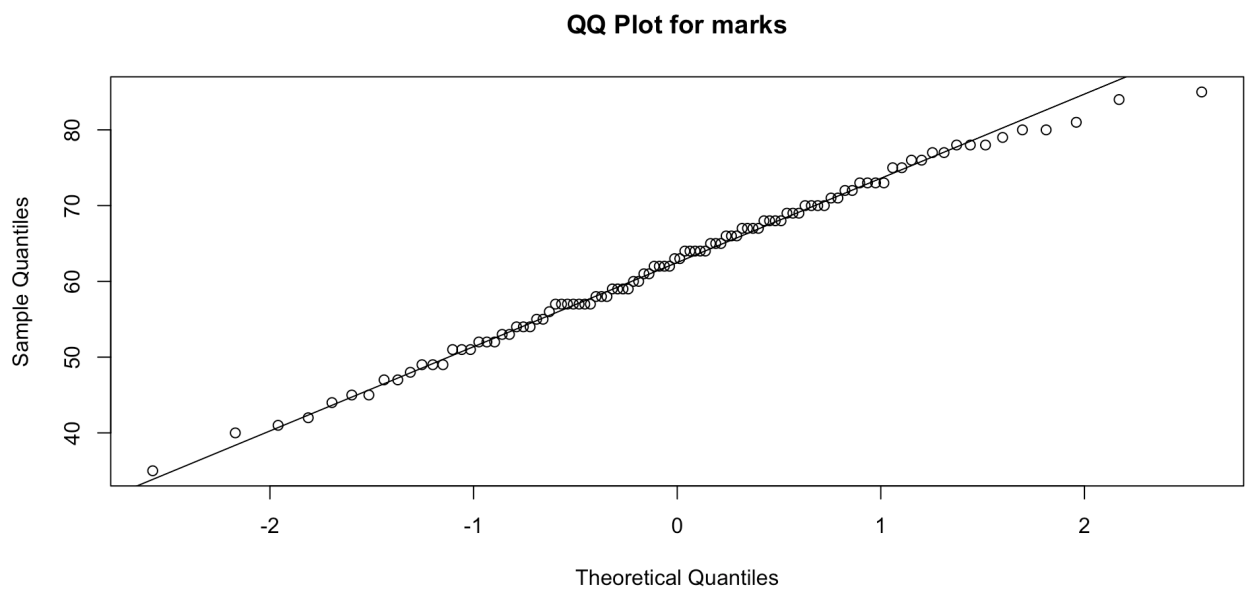


This boxplot is reasonably symmetric, with no outliers and the histogram looks quite bell-shaped, so it appears the "approximately normal box" assumption is reasonable.

QQ plot for checking normality

A better option: the **quantile-quantile (QQ) plot** is a graphical summary for determining if a data set is drawn from a distribution, e.g., a normal curve.

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



- A circle (x, y) corresponds to the quantile of the data (y) plotted against the same quantile of the normal curve (x).
  - ➠ Normally distributed data will have their points close to the line.
  - ➠ The linearity of the points suggests that the marks are approximately normally distributed.

- **T Test statistic:** 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 99 degrees of freedom.

  - 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
```

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

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

```
[1] 0.01965109
```

- This is not very different from the Z-test.

- [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.
  - Critical region of rejection:
    - At $\alpha = 5\%$, $|T| > 1.984$
    - At $\alpha = 2\%$, $|T| > 2.365$
    - At $\alpha = 1\%$, $|T| > 2.626$

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

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

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.

```
1  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: `alt="greater"` or `alt="less"`.

Confidence Interval

- Note the confidence interval given in the `t.test()` output.
- This takes the form $\bar{x} \pm q\frac{\hat{\sigma}}{\sqrt{n}}$, where q is the appropriate multiplier obtained using `qt()`.
- For a 95% confidence interval we obtain the upper 2.5% percentage point:

```
1  qt(0.975, df = 99)
```

```
[1] 1.984217
```

- Note this is not that different from the $N(0,1)$ upper 2.5% point 1.96;
- This reflects the fact that for larger n, the t-distribution gets closer to the $N(0,1)$ distribution.

```
1  mean(marks) + c(-1, 1) * 1.984217 * sig.hat/sqrt(100)
```

```
[1] 60.3348 64.5852
```

Confidence Interval - probability statement

- Under our assumptions, since *whatever be* the true value of μ,

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

and so

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

- 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.

➠ Note that the interval is random whereas the population mean μ is fixed.

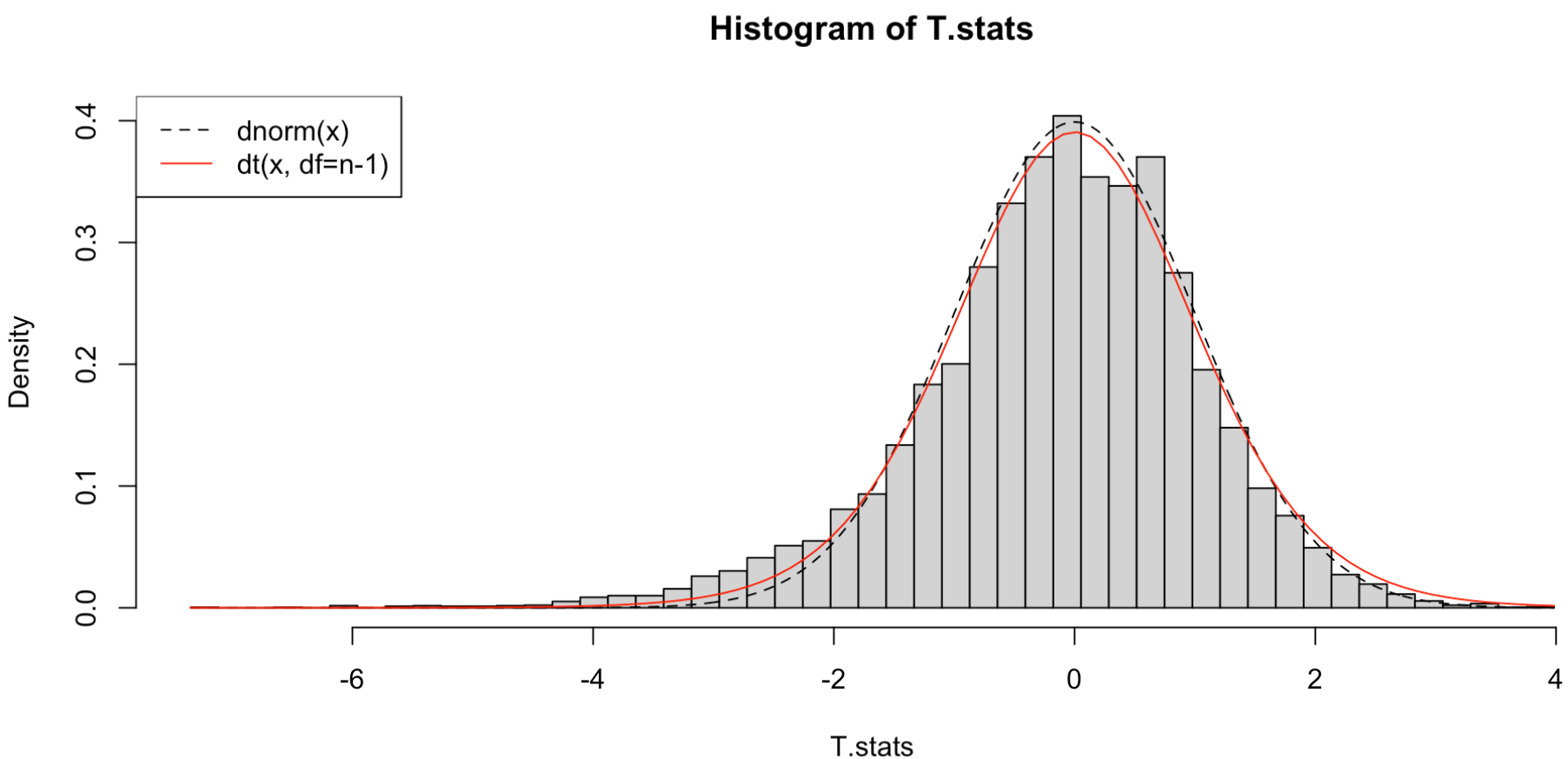
Skewed data

Skewed example

- What if instead we have a box like


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

```r
r = range(T.stats)
br = seq(from = r[1], to = r[2], length = 50)
hist(T.stats, breaks = br, pr = T, xlim = r)
curve(dnorm(x), n = 1001, lty = 2, add = T)
curve(dt(x, df = n - 1), add = T, col = "red")
legend("topleft", legend = c("dnorm(x)", "dt(x, df=n-1)"), lty = c(2, 1), col = c("black", "red"))
```



Clearly no good

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

```
      1%       2%       5%       95%      98%      99%
-3.480202 -2.986644 -2.216724  1.616247  2.034922  2.363002
```

```r
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
```

- The distribution of the T-statistics in samples from this box are clearly asymmetric.
  - ➠ 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 **simulation**.

P-values and confidence intervals via simulation

Bootstrap principle

- The theory behind the T-test suggests that if the box is "normal" (or "nearly normal") we can use `pt()` to get P-values using the T-statistic.
- But if the box is not "nearly normal"
  - ➠ We may not know the *exact distribution* of T when H_0 is true...
  - ➠ what should we do?
- We can try to **approximate** the distribution of T by **simulating** from a box which is "reasonably close" to the "real" box.
  - ➠ The observed sample can be used as a "surrogate box".
- This idea of approximating the behaviour of a statistic by using a "best guess" to the underlying population is known as **the bootstrap principle**.

Example

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


```
box
```

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

```
n
```

```
[1] 13
```

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

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

- How does the T-statistic behave when we simulate from a box that looks like this sample?

Simulating from a “best guess” to the box

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

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

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

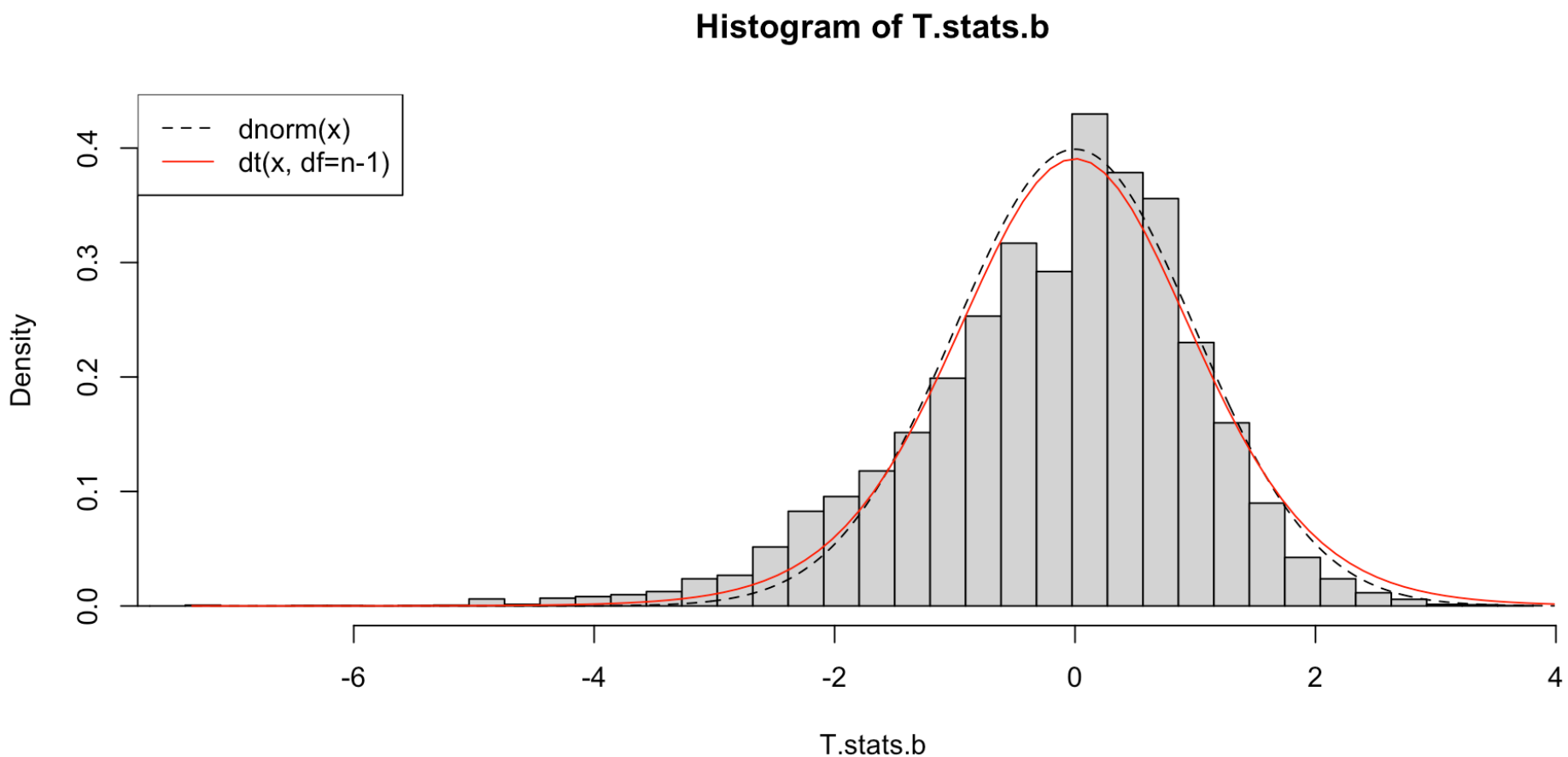
```
box.g
1 2 3 4 5 8
4 1 4 2 1 1
```

```r
mu.g = mean(box.g)
sig.g = sqrt(mean(box.g^2) - mean(box.g)^2)
Z.stats.b = 0  # b for 'bootstrap'
T.stats.b = 0
n = 13
for (i in 1:10000) {
    samp = sample(box.g, size = n, replace = T)
    m = mean(samp)
    sig.hat = sd(samp)
    T.stats.b[i] = sqrt(n) * (m - mu.g)/sig.hat
}
```

- Note: we calculate T-statistics of simulated samples drawn from the surrogate box

```
r.b = range(T.stats.b)
br = seq(from = r.b[1], to = r.b[2], length.out = 50)
```

```
hist(T.stats.b, breaks = br, pr = T, xlim = r)
curve(dnorm(x), n = 1001, lty = 2, add = T)
curve(dt(x, df = n - 1), add = T, col = "red")
legend("topleft", legend = c("dnorm(x)", "dt(x, df=n-1)"), lty = c(2, 1), col = c("black", "red"))
```



Not bad

- 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.
- Compare the quantiles of `T.stats.b` (from the surrogate box) with those of `T.stats` (from the true box):

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

```
       1%        2%        5%       95%       98%       99% 
-3.480202 -2.986644 -2.216724  1.616247  2.034922  2.363002 
```

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

```
       1%        2%        5%       95%       98%       99% 
-3.605551 -2.992528 -2.269127  1.469937  1.859962  2.132007 
```

```
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
```

- These agree quite well, so we could use the quantiles of `T.stats.b` to approximate the distribution of

$$T = \frac{\bar{x} - \mu_0}{\frac{\widehat{\sigma}}{n}}$$

when H_0 is true.

“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:

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

```
   97.5%      2.5% 
1.759765 -2.856099 
```

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

```
   97.5%     2.5% 
1.691068 5.124397 
```

- Note that this is not necessarily symmetric about the "point estimate" $\bar{x}$, as it takes into account the underlying lack of symmetry.

A review of inference for unknown means

Inference for unknown mean μ of a box

- We model data $X_1, \ldots, X_n$ as a simple random sample taken with replacement from a box with unknown mean μ and SD σ (may be known or unknown).
- Inference is based on the behaviour of the random statistics

$$Z = \frac{\bar{X} - \mu}{SE(\bar{X})} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ or } T = \frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}},$$

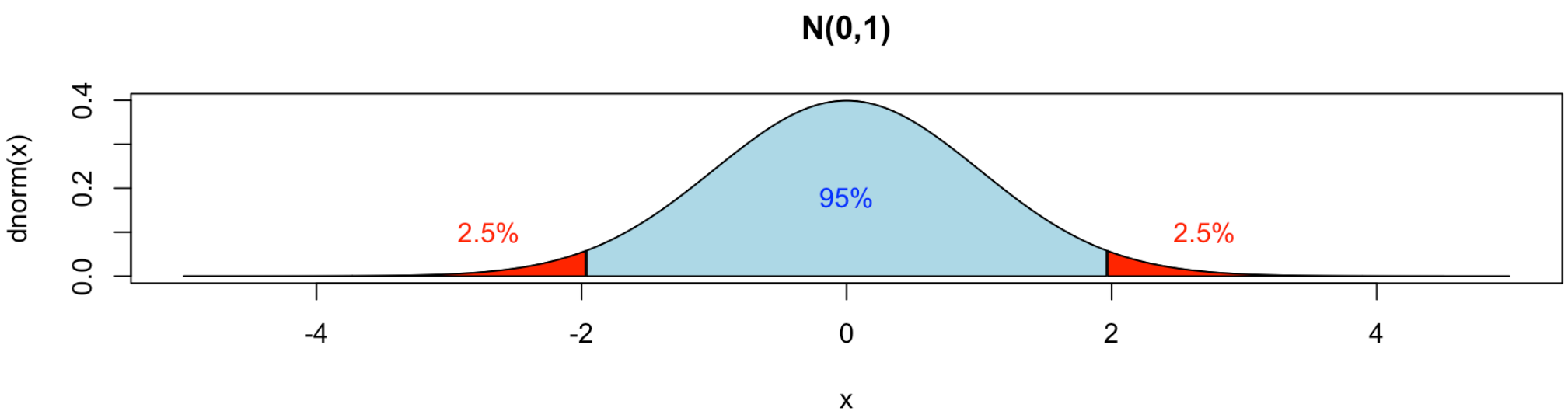
  where $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.
- Under certain assumptions, *theory* tells us how these behave:
  - **Central Limit Theorem**: if n "large enough" , Z (approx.) like $N(0, 1)$.
  - **Student's theory**: if box "approximately normal", T (approx.) like t_{n-1}.
- If we don't want to rely on theory, we can (maybe) use simulation instead.

Equal-tailed confidence interval

- If we know lower and upper values ℓ and u such that $P\{Z < \ell\} = 2.5\%$ and $P\{Z > u\} = 2.5\%$, then

$$P\left\{\ell \leq Z = \frac{\bar{X} - \mu}{SE(\bar{X})} \leq u\right\} = 95\%$$

When Z is like $N(0,1)$ we have $\ell = -1.96, u = +1.96$:



- Multiply through by $SE(\bar{X})$:

$$P\left\{\ell SE(\bar{X}) \leq \bar{X} - \mu \leq uSE(\bar{X})\right\} = 95\%$$

- Multiply through by -1:

$$P\left\{-\ell SE(\bar{X}) \geq \mu - \bar{X} \geq -uSE(\bar{X})\right\} = 95\%$$

- Add $\bar{X}$:

$$P\left\{\bar{X} - \ell SE(\bar{X}) \geq \mu \geq \bar{X} - uSE(\bar{X})\right\} = 95\%$$

- Reverse inequalities:

$$P\{\bar{X} - uSE(\bar{X}) \leq \mu \leq \bar{X} - \ell SE(\bar{X})\} = 95\%$$

- This probability statement justifies the form of a confidence interval.
  - ➠ Can easily change to 98%, 99%, whatever; same approach used for T.
  - ➠ Doesn't apply to Wilson's confidence interval, as $SE(\bar{X})$ depends on μ for the 0-1 box.

Hypothesis tests

- Again, **if we know the distribution of Z (or T)**, we can test H_0**:** $\mu = \mu_0$ (for some "special", "benchmark", etc. hypothesised value μ_0).
- For a given false-alarm rate (level of significance), we can construct a "critical region of rejection" for Z (or T) assuming $\mu = \mu_0$:
  - Could be one-sided or two-sided, depending on the alternative;
  - If Z (or T) takes a value in the critical region, we reject H_0.
- **Or**, we can convert Z (or T) into a P-value, e.g. if Z (assuming μ_0) takes value z, this is either
  - $P(Z > z)$,
  - $P(Z < z)$ or
  - $2P(Z > |z|)$,

  depending on the alternative (similarly for T taking value t).

Using simulation instead

- If we are not sure about the theory, we can replace the theoretical test distribution with computer simulation (cheap, accessible).
- The idea is, we don't know what the box is exactly, but we can use the data to construct a box that is "similar enough" to the "real" box.
- We can then draw random samples with replacement from this "approximate box":
  - ➠ This is just sampling with replacement from the data!
  - ➠ We then see how the T-statistic behaves for that box and use that as an approximation.

Exam marks example: Z-test

- 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 52 64 42 78 60 57
61 47 75 58 51 35
 [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 76 52 77 65 75 80
73 70 61 55 69 66
 [83] 62 73 80 70 57 78 56 59 65 73 60 72 76 62 57 68 77 71
```

```
mean(marks)
```

```
[1] 62.46
```

- We first used a Z-test, assuming the SD of the "box" was known $= 10$.

```
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
```

Exam marks example: T-test

- We then tried performing a T-test, where instead of **assuming** the box SD σ is known, we estimate it using the *sample SD* of the data.

```
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
```

Using `t.test()`

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

Exam marks example: simulated T-test

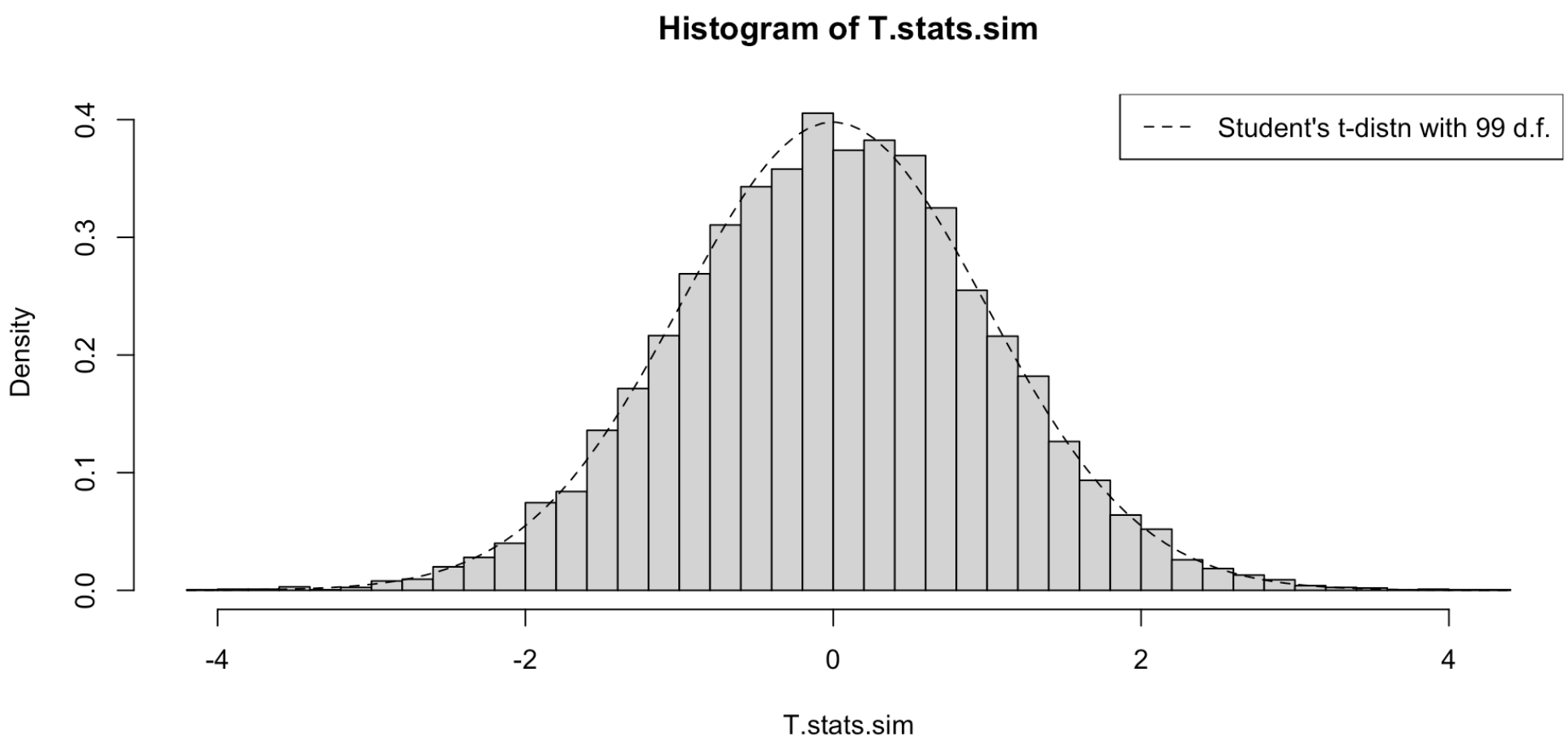
- What about a simulation? We repeatedly sample from the data, and compute the value taken by the T-statistic

```
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.

Histogram of simulated T-statstics

```r
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)
```



Simulation-based P-value

- How significant is our observed T-statistic value of `-2.371`, based on the simulation?
  - ➠ it's the same T-statistic, but a different testing distribution
- What proportion of the values in `T.stats.sim` exceed `abs(t)=` `2.371` (in absolute value)?

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

```
[1] 0.0209
```

- About 2%. Since this agrees very closely with the `t.test()` output, we feel comfortable in making the "approximately normal box" assumption that underlies the validity of the `t.test()` P-value.

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 given in the `t.test()` output.
  - ➠ Again, suggests the assumptions underlying `t.test()` are reasonable.

A review of sample SD using the box model

Sample SD

- The sample SD

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

  uses $\frac{1}{n-1}$ in the average. Previously (Topic 2), we justified this choice using the fact that the sum of deviations in a sample is zero, i.e.,

$$\sum_{i=1}^{n} X_i - \bar{X} = 0\,,$$

  so that there is only $\boldsymbol{n-1}$ pieces of effective information in the deviations.

- We want to understand this choice using **the box model and expectations**.
  - ➠ Not for assessment.

Sum of squared deviations

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

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

  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
  - $E(S) = n\mu$
  - $E(\bar{X}) = \mu$
- How does the **sum of squared deviations** (from the population mean μ)

$$SSD = \sum_{i=1}^{n}(X_i - \mu)^2$$

  behave?

Like the sum of a sample from a different box

- $(X-\mu)^2$ is a random draw from a different box

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

- This box has mean

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

  which is the squared population SD.

- 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.

- $SSD=\sum_{i=1}^{n}(X_i-\mu)^2$ is the *sum* of a random sample (with size n) drawn with replacement from this box.

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

Trick

- We can use a trick. We can "add and subtract" the (random) sample mean inside the square to 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 $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ **underestimates** σ^2 in expectation. Replacing $\frac{1}{n}$ with $\frac{1}{n-1}$ corrects this:

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