Unknown Proportions and Means

Decisions with Data | Inference for proportions and means

STAT5002

The University of Sydney

Apr 2025

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 general framework of hypothesis tests

Z-test for proportion (review)

Z-test for mean with known SD

The general framework of hypothesis tests

HATPC framework

It's helpful to follow the HATPC framework to conduct hypothesis tests:

- $\boxed{\text{H}}$ Hypotheses
  - Set up the two hypotheses: the null H_0 and the alternative H_1.
- $\boxed{\text{A}}$ Assumptions
  - State the assumption(s) of the test, and justify if they are valid based on the sample and the sampling process
- $\boxed{\text{T}}$ 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.
  - Find the observed value of the Test Statistic.
- $\boxed{\text{P}}$ P-value
  - Calculate the P-value, the probability of observing the sample (or more extreme) under H_0.
    - we summarise the sample by the observed test statistic
- $\boxed{\text{C}}$ Conclusion
  - Weigh up the conclusion, based on the P-value and the level of significance α.

Z-test for proportion under HATPC

Example (last week)

- A production line produces items at a rate of 5000 per day.
- It is deemed "acceptable" if **3%** of the items are faulty.
- Every week a random sample of $n = 200$ items is taken and the proportion of faulty items $\bar{x}$ is determined.
- If there is evidence that the "failure rate" is higher than 3%, they stop the production and repair the machines.
- How should such a test be performed so that is at $\alpha = 1\%$ level of significance?
    - ➟ This is the false alarm rate, which is the chance of needless shutdown.
- We observe $s = 11$ faulty items in one sample, what decision should we make?

Hypotheses

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

- The H_0 is the default hypothesis: what we currently believe to be true about the population.
    - In this case, $H_0 : p_0 = .03$.
- The H_1 is a new claim about the population.
    - It can take 2 forms:
        - 1-sided ($H_1 : p_0 > 0.03$ or $H_1 : p_0 < 0.03$)
        - 2-sided ($H_1 : p_0 \neq 0.03$).
- How to decide between a 1 or 2 sided test?
    - Depending on the context.
        - In this case, we take $H_1 : p_0 > 0.03$, as this is the alternative we want to detect.
    - The decision must not be influenced by the data – we must specify the hypotheses before we do the actual test.

A Assumptions

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 this case, 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
  - ➟ so a sample of 200 items can be viewed as "almost" independent
- We can justify that $n = 200$ is a sufficiently large sample size so that **CLT** may hold
- Then the normal curve can be used to approximate the sample proportion

T Test Statistic

- 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 $E(\bar{X}) = p_0$ and $SE(\bar{X}) = \sqrt{\frac{p_0(1-p_0)}{n}}$. The z-score of the sample proportion approximately follows the standard normal curve

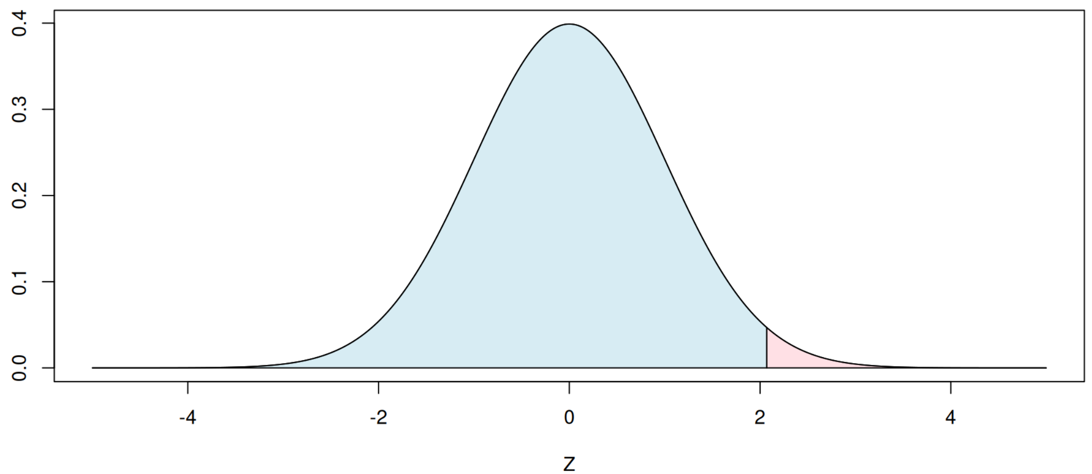
$$Z = \frac{\bar{X} - E_0(\bar{X})}{SE_0(\bar{X})} = \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

  - ➟ 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, large values of Z-statistic will argue against H_0.
- Calculate the **observed value of the Test Statistic** ($\bar{x} = \frac{11}{200} = 0.055$)

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

P-value

- 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
  - ➟ We have a one-sided test in this case, and large values of Z-statistic will argue against H_0
  - ➟ P-value $= P(Z > z) = P(Z > \mathbf{2.07}) =$ `pnorm(2.07, lower.tail=F)` = 0.019



- 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.

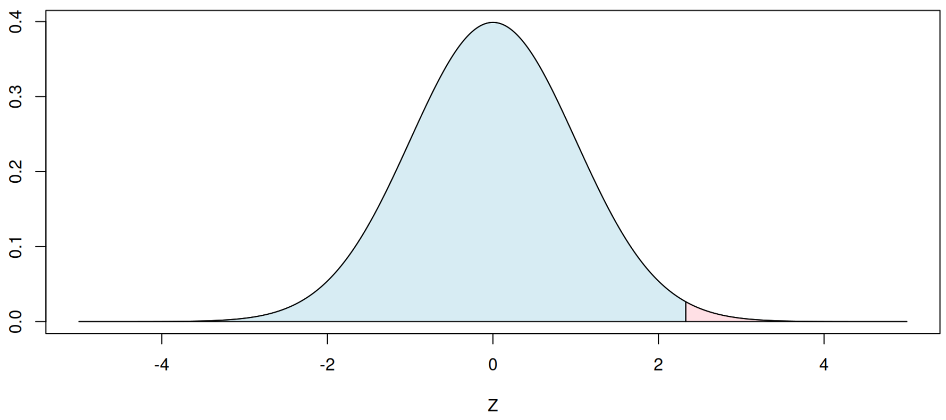
C Conclusion

We often make decision based on P-value of the **Level of significance α**. In this example, $\alpha = 1\%$.

- Allowing 1% of false alarm rate assuming H_0 is true.
- Given it's one-sided and large values of Z-statistic arguing against H_0. We have the multiplier:

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

```
[1] 2.33
```



- If the obsered Z-statistic is above the multiplier 2.33, we consider it's inconsistent with H_0.
  - ➟ If this happens, it means that the corresponding P-value is less than the level of significance.
  - ➟ Then, we reject H_0 at the 1% level of significance

C Conclusion

- In this example, the P-value is $\mathbf{0.019}$ (with observed Z-statistic $\mathbf{2.07}$).
    - P-value $= \mathbf{0.019} > \boldsymbol{\alpha}$, so we can't reject $\boldsymbol{H_0}$.
    - We say "the data is consistent with the null hypothesis $\boldsymbol{H_0}$" but **never accept $\boldsymbol{H_0}$**.
        - **A single observation does not prove a hypothesis true**.
- We don't need the P-value to make the decision. Based on the level of significance $\boldsymbol{\alpha}$, we can determine a **critical region** of test statistics such that their corresponding P-values are smaller than $\boldsymbol{\alpha}$.
    - In this example, $\mathbf{1\%}$ significance level corresponds to the multiplier 2.33 ($\mathbf{1} - \boldsymbol{\alpha} = \mathbf{99\%}$ quantile).
    - The critical regions is given by $[\mathbf{2.33}, \infty)$.
    - Quicker decison: Z-statsitic $= \mathbf{2.07} < \mathbf{2.33}$, outside the critical region, so data is consistent $\boldsymbol{H_0}$.

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).
- An operation guide if one should shut the production line based on different levels of significance.

```r
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
E.Xbar = p0  # expected sample proportion
SE.Xbar = sqrt(p0 * (1 - p0)/n)  # SE of the sample proportion
critical.values.props = (critical.values * SE.Xbar + E.Xbar)  # critical values in sample proportions
critical.values.sums = critical.values.props * n  # critical values in sample sums
observed.faulty.items = ceiling(critical.values.sums)  # rounding to the nearest interger larger than sums
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
```

- We use `ceiling()` to round to the nearest interger larger than the critical values of sample sum
  - ➟ 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.

2025 election

YouGov's latest Public Data survey

YouGov's latest Public Data survey published on 11 April reveals that Labor Party now leads the Liberal/Nationals coalition 52.5% to 47.5% in the two-party preferred vote.

- This survey was conducted between April 4th and April 10th with a sample of 1505.
- A TV commentator (I made this part up) claimed that we would have a hung parliament if the election were held on 11 April. Using a 10% level of significance, Let's test the claim.
    - Assume a hung parliament means 50% support for each side.

HATPC

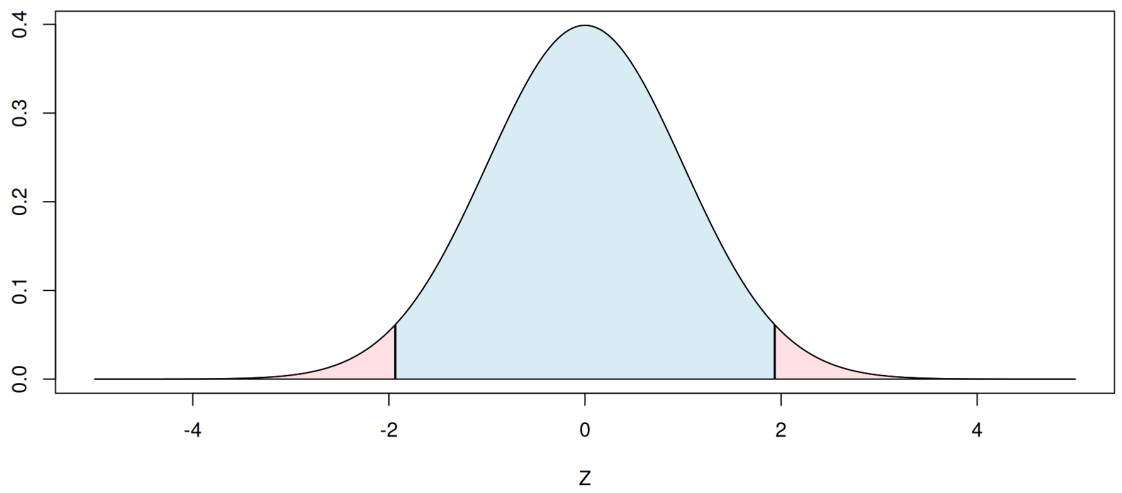
- $\boxed{\text{H}}$ **Null hypothesis**: $H_0\colon p_0 = .5$, and **alternative hypothesis**: $H_1\colon p_0 \neq .5$.
  - A *two-sided* test is appropriate here, as we are interested in not having a hung parliament.
- $\boxed{\text{A}}$
  - Sample without replacement again, but very large population (box) comapred to the sample size.
  - So we assume CLT holds.
- $\boxed{\text{T}}$ When H_0 is true the sample proportion $\bar{X}$ is like a random draw from a normal box with mean equal to $E(\bar{X}) = p_0 = .5$ and SD equal to $SE(\bar{X}) = \sqrt{\frac{p_0(1-p_0)}{n}} = \frac{0.5}{\sqrt{1505}} \approx \mathbf{0.0129}$. Equivalently the Z-statistic

$$Z = \frac{\bar{X} - E(\bar{X})}{SE(\bar{X})} = \frac{\bar{X} - .5}{0.0129}$$

approximately follows the standard normal curve.
  - Two-sided test, and hence small and large values of Z-statistic argue against H_0.
  - The observed value of Z-statistic is $z = \mathbf{1.937}$ (using the Labour's proportion of support 0.525).

HATPC

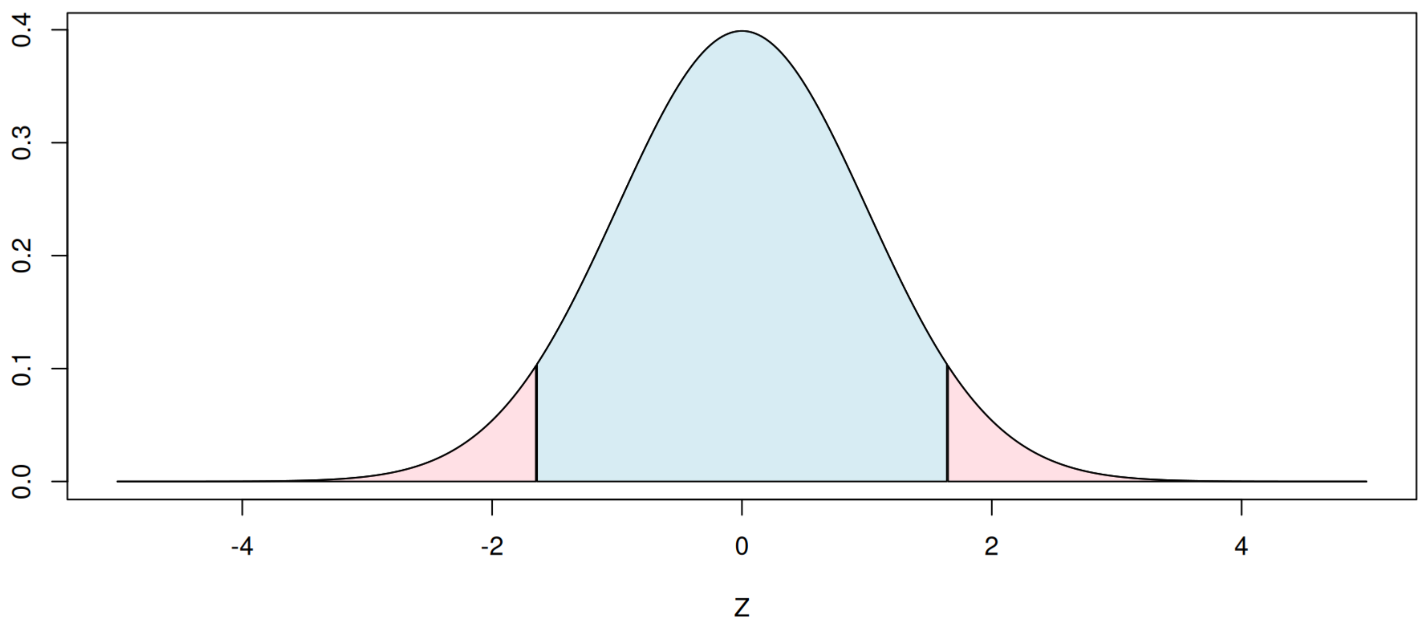
- $\boxed{\text{P}}$ The P-value is the probability of observing something more extreme than the observed sample.
    - "something more extreme" = test statistics that argue against H_0 more than the observed one.
    - Two-sided test here: large and small values of Z-statistic will argue against H_0
        - which is defined as $|Z| > |z|$
    - P-value $= P(Z > |z|) + P(Z < -|z|) =$ `2 * pnorm(1.937, lower.tail=F)` = 0.053



- Note: if we use the coalition's proportion of support 0.475, we will have $z = -\mathbf{1.937}$, which leads to the same P-value.

HATPC

- $\boxed{\text{C}}$ P-value $= \mathbf{0.053} < \boldsymbol{\alpha} = \mathbf{0.1}$, so we reject $\boldsymbol{H_0}$.
  - The observed proportion is significantly different to .5 at the 10% level of significance.
  - This constitutes evidence against $\boldsymbol{H_0}$, suggesting we wouldn't have a hung parliament.
  - Note that for the $\mathbf{10\%}$ level of significance, the critical region is given by $|z| \geq \mathbf{1.645}$.
  - For two-sided test, the multiplier is given as `qnorm(1 -`α`/2)` – see below.



Z-test for unknown mean with known SD

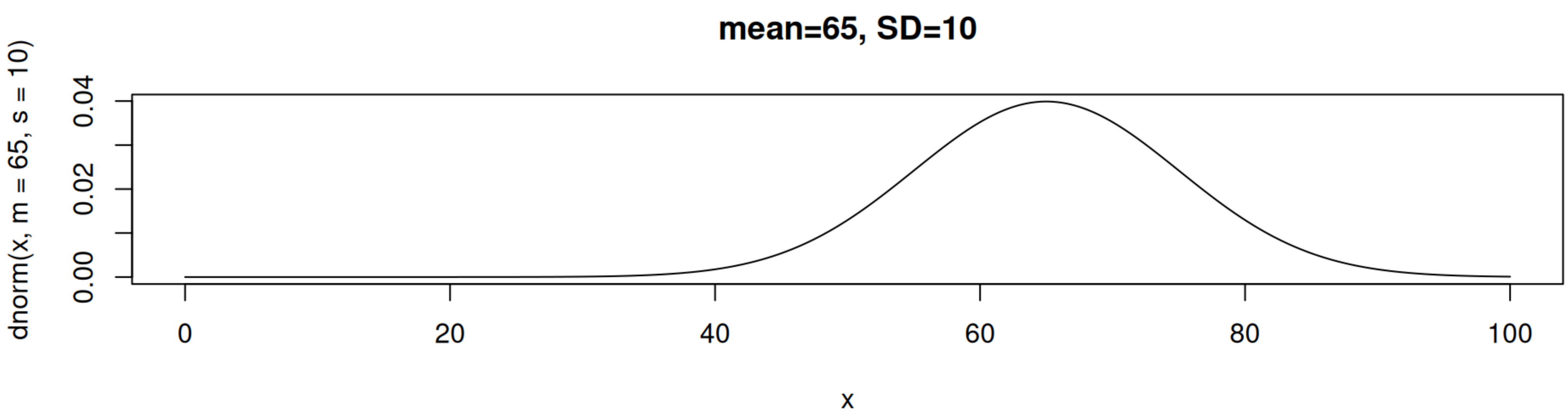
Wider scope of Z-test

- We have focussed on the scenario of inference for a proportion.
- Since the null 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 under H_0.
    - ➟ 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.

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.

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



Moderating exams

- Suppose that for mathematics
  - The spread of marks from year to year is much the same, with SD $= \sigma_0 = 10$ but
  - The average mark μ tends to vary from year to year.
- To produce a "standardised" exam, a draft can be tested on a small group of students.
- 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 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
```

```r
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 **different** to 65, but what is our decision based on a 5% level of significance.

HATPC

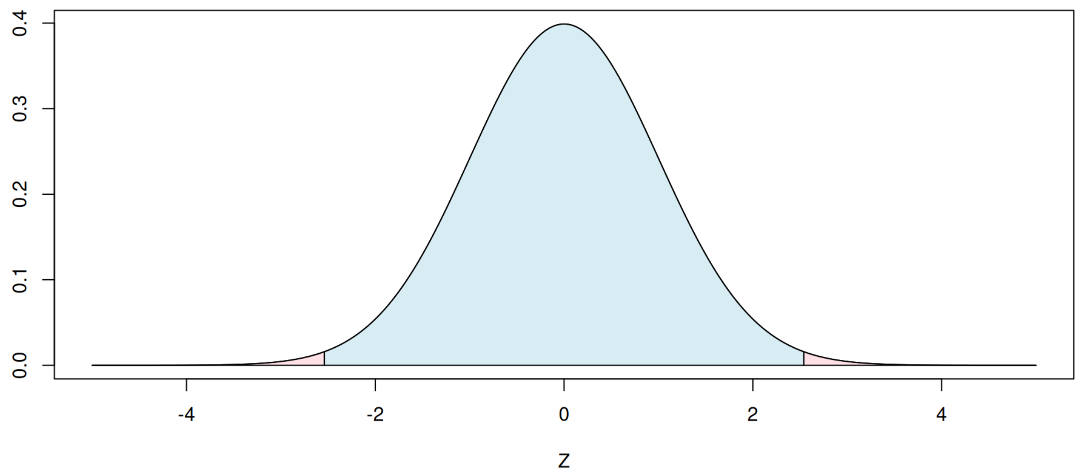
- $\boxed{\text{H}}$ **Null hypothesis**: H_0: $\mu_0 = 65$, and **alternative hypothesis**: H_1: $\mu_0 \neq 65$.
  - ➟ A *two-sided* test is appropriate here, as the exam could be too easy or too hard.
- $\boxed{\text{A}}$ Students are like a random sample taken without replacement from a very large box (there are many students in the population)
  - ➟ so we consider them as independent draws.
  - ➟ Their marks have unknown mean μ but **known** SD $\sigma_0 = 10$.
  - ➟ assuming CLT as $n = 100$ is reasonably large.
- $\boxed{\text{T}}$ 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

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

  approximately follows the standard normal curve.
  - ➟ Two-sided test, and hence small and large values of Z-statistic argue against H_0.
  - ➟ The observed value of Z-statistic is $z = \bar{x} - 65 = 62.5 - 65 = -2.54$.

HATPC

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



- $\boxed{\text{C}}$ P-value $= \mathbf{0.011} < \alpha = \mathbf{0.05}$, 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 $\mathbf{5\%}$ level of significance, the critical region is given by $|z| \geq \mathbf{1.96}$.

Decision based on the confidence interval

- Recall that for the unknown mean but known SD, the 95% confidence interval based on the observed $\bar{x}$ is

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

- For an observed $\bar{x} = 62.46$, this gives

```r
62.46 + c(-1, 1) * 1.96 * 1
```

```
[1] 60.50 64.42
```

- Note this does not include the value 65, so in this sense, the data is not consistent with the claimed μ_0 being 65.
- Both the confidence interval and the **two-sided** Z-test are defined using the consistency (based on the two-sided prediction interval), so $1 - \text{confidence level} = \text{significance level}$.

Different confidence levels

- We can get confidence intervals at different confidence levels:

Conf. level	Multiplier	Interval	Includes 65?
95%	1.960	(60.50, 64.42)	No
98%	2.326	(60.13, 64.79)	No
99%	2.576	(59.88, 65.04)	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).

Estimating the standard error

Assuming SD of the box is known

- The Z-statistic, which measures how many SEs $\bar{X}$ is away from 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.
- In the previous example, sample SD of the exam marks is

```
sd(marks)
```

```
[1] 10.71053
```

The T-statistic

- The T-statistic simply replaces $\boldsymbol{\sigma_0}$ 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" $\widehat{\quad\cdot\quad}$ over $SE_0(\cdot)$ and $\boldsymbol{\sigma}$ 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? – after the break.