STAT115: Introduction to Biostatistics

University of Otago
Ōtākou Whakaihu Waka

Lecture 16: Errors and Power in Tests

Outline

- Previous:
  - Confidence interval for μ
  - Hypothesis test
- Today:
  - Explore more of the properties around the hypothesis test
  - Type I and Type II errors
  - Power of a test
  - Trade-offs between errors and power

Height of 100-level STAT students

- In previous years there was a questionnaire (optional) for STAT110 students
  - Questions about age, height, sex, ...
- Exploratory study
  - Explore the height of females in STAT110 relative to national average
    - Average height for NZ female aged 15-24 is 164.7 cm (`figure.nz`)[1]
    - Restrict ourselves to female STAT110 students aged 15-24

```
STAT110 = read.csv('../data/STAT110_height_f.csv')
head(STAT110$height)
```

```
## [1] 167 153 171 177 161 173
```

- Heights from $n = 451$ female students aged 15-24

[1] Data from New Zealand Health Survey, 2023

Hypothesis test

- Write down the null and alternate hypothesis
  - $H_0 : \mu = 164.7$
  - $H_A : \mu \neq 164.7$
- Use $\alpha = 0.05$
- We can conduct the test in R

```
h_test = t.test(STAT110$height, mu = 164.7)
h_test
##
##  One Sample t-test
##
## data:  STAT110$height
## t = 8.073, df = 450, p-value = 6.32e-15
## alternative hypothesis: true mean is not equal to 164.7
## 95 percent confidence interval:
##  166.891 168.301
## sample estimates:
## mean of x
##   167.596
```

Interpretation

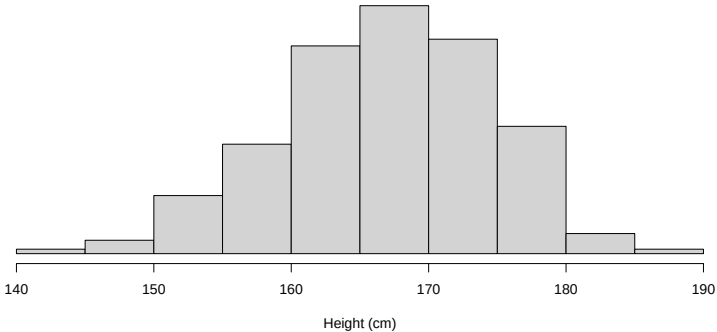
- Exploratory study: interpret p-values (no formal test)
- There is evidence that the data are incompatible with null hypothesis
  - p-value is approximately 1 in a quadrillion[2]
- Evidence that the (mean) height of female STAT110 students is incompatible with national average
- Do we trust it? It pays to be cautious
  - Students in STAT110 are not a random selection of 15 – 24 year olds in NZ
  - STAT110 data are voluntary and heights are self-reported
    - There are also very different rates of left-handedness from national averages
- If this question were of interest...
  - There is 'enough' to look into designing a (confirmatory) study

[2] The progression is million (10^6), billion (10^9), trillion (10^{12}), quadrillion (10^{15}), ...

Assumptions

- We have made an assumption that our data are normally distributed
  - Just as we did with confidence intervals
- To check this assumption: looking for serious departures from normality
  - We check visually (histogram)
- As with confidence intervals: if the sample size is large enough
  - p-values are reasonable for non-normal data
  - Discuss more in a few weeks

Histogram



- No obvious departures from normality
- Large sample (~ 450)

Setup

- We want to better understand how hypothesis testing works
- We do this in the context of formal hypothesis test
  - If p-value $< \alpha$ we reject H_0
  - If p-value $> \alpha$ we fail to reject H_0
- There are four possibilities:

	Decision	
	Do not reject H_0	Reject H_0
H_0 true	✓	Type I error
H_0 not true	Type II error	✓

Setup

- Consider a specific gene: GENE-X
  - Reference expression value of 5.0 TPM (transcripts per million) in healthy individuals
- Design a confirmatory study to test if GENE-X is expressed differently in a sample of people with a specific disease
  - $H_0 : \mu = 5$ (the mean expression for the diseased group is the same as the reference)
  - $H_A : \mu \neq 5$
- In this study:
  - We want to find evidence against the null
  - We want to find evidence that gene expression differs in the diseased group
- In the rest of the lecture an effect is defined as:
  - Effect: difference between the mean for the disease group and $\mu_0 = 5$

A tale of two errors

- Type I Error (α): Rejecting H_0 when it is true.
  - Concluding the expression of GENE-X is different for the diseased group, when it isn't
- Type II Error (β): Failing to reject H_0 when H_A is true.
  - Concluding that there is no evidence that expression of GENE-X differs for diseased group, when there is a non-zero effect

Type I error

- Type I error rate is given by α, the significance level
  - Decreasing α from 0.05 to 0.01 will reduce the number of type I errors we make
    - Recall: α is the threshold for incompatibility with null
    - A lower α is applying a higher threshold for incompatibility

Type II error

- The type II error rate is representated as β
- We often refer to the power $= 1 - \beta$
- Power: the probability of rejecting the null hypothesis, given it is incorrect
  - i.e. it is the probability of detecting an effect, given there is one
- All else equal, we want a powerful test
  - More likely to correctly reject H_0
  - More likely to correctly conclude that gene expression differs in diseased group
- We will look at four factors that change the type II error / power

Type I error rate α

- Trade off between type I error rate and power
  - If we decrease α (lower type I error rate)
    - Increase type II error rate β
    - Decrease power
- If we increase α (higher type I error rate)
  - Decrease type II error rate β
  - Increase power

Effect size

- Recall: $\mu_0 = 5$ TPM (transcripts per million)
- Consider two scenarios:
  1. The true mean of the diseased population is $\mu_A = 5.1$ TPM
  2. The true mean of the diseased population is $\mu_A = 12$ TPM
- In which scenario will power be higher (all else equal)?

[3] $|x|$ is the absolute value of x

Effect size

- Recall: $\mu_0 = 5$ TPM (transcripts per million)
- Consider two scenarios:
  1. The true mean of the diseased population is $\mu_A = 5.1$ TPM
  2. The true mean of the diseased population is $\mu_A = 12$ TPM
- In which scenario will power be higher (all else equal)?
- The larger[3] the effect $|\mu_A - \mu_0|$
  - The more powerful the test, all else equal
- The size of the effect is not something we can typically control

[3] $|x|$ is the absolute value of x

Sample size

- For a fixed α and effect size, consider these two scenarios:
  1. The sample size (of diseased participants) is $n = 20$
  2. The sample size (of diseased participants) is $n = 200$
- In which scenario will power be higher?

Sample size

- For a fixed α and effect size, consider these two scenarios:
  1. The sample size (of diseased participants) is $n = 20$
  2. The sample size (of diseased participants) is $n = 200$
- In which scenario will power be higher?
- The larger the sample size
  - The more powerful the test, all else equal
- Scientific research (grant) funding in ecology, food science, global health, etc
  - Typically have to justify your research design
  - Power calculation: determining sample size needed to achieve a certain power

Population standard deviation

- For a fixed n, α, and effect size, consider these two scenarios:
  1. The population standard deviation (of gene expression in the disease group) is $\sigma = 0.1$
  2. The population standard deviation (of gene expression in the disease group) is $\sigma = 1$
- In which scenario will power be higher?

Population standard deviation

- For a fixed n, α, and effect size, consider these two scenarios:
  1. The population standard deviation (of gene expression in the disease group) is $\sigma = 0.1$
  2. The population standard deviation (of gene expression in the disease group) is $\sigma = 1$
- In which scenario will power be higher?
- The smaller the population standard deviation
  - The smaller the standard error
  - The more precise $\bar{y}$ is
  - The more powerful the test, all else equal
- The value of σ is not something we can typically control

p-value

- ASA principle: "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result"
- Suppose we have $p = 0.0000001$. This could be because:
  - This could be because the effect size is large
  - It could occur when the effect size is small (but non-zero) and sample size is large
- Care is needed that we don't confuse a small p-value, with an important result

Relationship with confidence intervals

- If we are testing the hypothesis:
  - $\mathrm{H}_0 : \mu = \mu_0$
  - $\mathrm{H}_A : \mu \neq \mu_0$
- There is an equivalence between p-value and confidence interval
  - p-value $< \alpha$ is equivalent to μ_0 outside the $(1-\alpha)100\%$ confidence interval
    - e.g. if p-value < 0.05, then μ_0 is outside 95% confidence interval
    - e.g. if p-value > 0.01, then μ_0 is inside 99% confidence interval

Quiz

- It's quiz time!
- Three possible answers for the questions below:
  - (1) increase; (2) decrease; (3) can't tell
- What is the effect on (i) type I error rate, and (ii) power if we:
  - Increase the sample size?
  - Decrease α?
  - Decrease the sample size and increased α?
  - Changed the research design so that the type II error rate β decreased?
  - Collected a sample twice the size for a different gene (GENE-Y) that has a smaller effect and larger σ?

Summary

- Checking assumptions
- Looked more at the properties of hypothesis testing
  - Type I error
  - Type II error
  - Power
- Looked at the effect of
  - Sample size
  - Effect size
  - α
  - σ