Schedule

9 to 9.15 Recap
9.15 to 9.40 Basic statistics: mean, sd, median, quantiles

9.40 to 10 Hypothesis testing: Null hypothesis, type I and II error
10 to 10.10 pause
10.10 to 11 Basic tests: T-test, chi-square, hypergeometric
11 to 11.15 pause
11.15 to 12 Multiple hypothesis correction: Bonferroni, Benjamini Hochberg

12 to 13 lunch break
13 to 16 Exercise
Lesson goals

- Learn the basics of statistics
- Perform statistical tests with R
- Be able to understand the main pitfalls
Statistics

Population

Estimate population features from calculations on a sample population
Statistics

Estimate population features from calculations on a sample.
Estimate population features from calculations on a sample
Estimate population features from calculations on a sample.

In R: your sample will likely be stored as an array, vector, matrix or data.frame.
Estimators

Bias and imprecision affect our ability to study the population. They can depend on sample size.
Examples

How many of us have a first name that starts with a vowel?
Examples

How many of us have a first name that starts with a vowel?

Population -> Hypothesis -> Sample -> estimates
What statistics is not

- Study of causes and consequences
- Probability
- 100% Objective
- Easy
Statistics 101

Average or mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Numeric variable (e.g. height, foot size, letters in your name)

OBS: this definition is for both population and sample

in R: mean(x)
Statistics 101

Median: Value that divides the population in 2

Numeric variable (e.g. height, foot size, letters in your name)

in R: median(x)
Statistics 101

Standard deviation: \[ SD = \sqrt{\frac{\sum(x - \bar{x})^2}{n}} \]
Standard deviation:

$$SD = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

This definition is for the whole population, it does not work for samples

Biased estimator!!
Statistics 101

Standard deviation: \[SD = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}\]

Unbiased estimator

in R: sd(x)
X% quantile: value y for which x\geq y is true for X% of the samples

50% quantile = median

in R: quantile(x, c(.05,.10,.25,.50,.75,.90,.95))
Exercise

R : connect to the server or locally

file at
www.cbs.dtu.dk/~pmar/nba.txt

or

/home/people/pmar/public_html/nba.txt

Name, Role, Height, Weight and Age of NBA players
Questions:
1- read the file nba.txt with the function read.table
2- calculate the mean, median and sd of the height
3- calculate the mean height of guards only
Hypothesis testing

We want to check if sampled data supports an hypothesis on the overall population

IN GENERAL

Do some subgroups show a difference in some features?

Do guards (G) and forwards (F) have the same height?
Null hypothesis

In order to do this, we formulate a Null hypothesis $H_0$
(same height for guards and forwards)

and an alternate hypothesis $H_A$
(Guards and forwards do not have similar height)

Then:
supposing that $H_0$ is true,
how likely would it be to observe the data in my sample?
Caveat

- Often $H_0$ is the opposite of what you want to prove
- It is easier to check if the data support the "no difference" hypothesis
- There might be many $H_A$
- Find relevant statistical measures (e.g. average, sd, etc.)
- $H_0$ is about population statistics, not on the interpretation
- Be specific and precise when stating $H_0$
## Type I and II errors

<table>
<thead>
<tr>
<th>True State</th>
<th>Test Result –</th>
<th>H_0 True</th>
<th>H_0 False</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_0 True</td>
<td>Correct Decision</td>
<td>Type I Error</td>
<td></td>
</tr>
<tr>
<td>H_0 False</td>
<td>Type II Error</td>
<td>Correct Decision</td>
<td></td>
</tr>
</tbody>
</table>

\[ \alpha = P(\text{Type I Error}) \quad \beta = P(\text{Type II Error}) \]

- Goal: Keep \( \alpha, \beta \) reasonably small
Type I and II errors

$\alpha = P(\text{Type I Error})$
G and F have similar height, test says they don't

$\beta = P(\text{Type II Error})$
G and F don't have similar height, test says they do
Tests

if $H_0$, $\text{mean}(F) = \text{mean}(G)$

if $H_0$, $\text{sd}(F) = \text{sd}(G)$

if $H_0$, $|\text{mean}(F) - \text{mean}(G)| < \text{sd}(G)$

if $H_0$, $|\text{mean}(R_1) - \text{mean}(R_2)| = |\text{mean}(F) - \text{mean}(G)|$
Example - Efficacy Test for New drug

- Drug company has new drug, wishes to compare it with current standard treatment
- Federal regulators tell company that they must demonstrate that new drug is better than current treatment to receive approval
- Firm runs clinical trial where some patients receive new drug, and others receive standard treatment
- Numeric response of therapeutic effect is obtained (higher scores are better).
- Parameter of interest: $\mu_{\text{New}} - \mu_{\text{Std}}$
Example - Efficacy Test for New drug

- **Null hypothesis** - New drug is no better than standard trt

\[ H_0 : \mu_{New} - \mu_{Std} \leq 0 \quad (\mu_{New} - \mu_{Std} = 0) \]

- **Alternative hypothesis** - New drug is better than standard trt

\[ H_A : \mu_{New} - \mu_{Std} > 0 \]

- **Experimental (Sample) data:**

<table>
<thead>
<tr>
<th>_y_{New}</th>
<th>_y_{Std}</th>
</tr>
</thead>
<tbody>
<tr>
<td>s_{New}</td>
<td>s_{Std}</td>
</tr>
<tr>
<td>n_{New}</td>
<td>n_{Std}</td>
</tr>
</tbody>
</table>
Difference in Means

- In large samples, the difference in two sample means is approximately normally distributed:

\[ \bar{Y}_1 - \bar{Y}_2 \sim N \left( \mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \]

• Under the null hypothesis, \( \mu_1 - \mu_2 = 0 \) and:

\[ Z = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \]

• \( \sigma_1^2 \) and \( \sigma_2^2 \) are unknown and estimated by \( s_1^2 \) and \( s_2^2 \)
Example - Efficacy Test for New drug

- **Type I error** - Concluding that the new drug is better than the standard ($H_A$) when in fact it is no better ($H_0$). Ineffective drug is deemed better.
  - Traditionally $\alpha = P(\text{Type I error}) = 0.05$

- **Type II error** - Failing to conclude that the new drug is better ($H_A$) when in fact it is. Effective drug is deemed to be no better.
  - Traditionally a clinically important difference ($\Delta$) is assigned and sample sizes chosen so that:
    $$\beta = P(\text{Type II error} \mid \mu_1 - \mu_2 = \Delta) \leq .20$$
Elements of a Hypothesis Test

- **Test Statistic** - Difference between the **Sample means**, scaled to number of standard deviations (standard errors) from the null difference of 0 for the **Population means**:

  \[ T.S.: z_{obs} = \frac{- \bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \]

- **Rejection Region** - Set of values of the test statistic that are consistent with \( H_A \), such that the probability it falls in this region when \( H_0 \) is true is \( \alpha \) (we will always set \( \alpha = 0.05 \))

  \[ R.R.: z_{obs} \geq z_\alpha \quad \alpha = 0.05 \Rightarrow z_\alpha = 1.645 \]
**P-value (aka Observed Significance Level)**

- **P-value** - Measure of the strength of evidence the sample data provides against the null hypothesis:

\[
P(\text{Evidence This strong or stronger against } H_0 \mid H_0 \text{ is true})
\]

\[
P-val : p = P(Z \geq z_{obs})
\]
Find $H_0$

Check if a antipsychotic drug works properly

Check if different ethnic groups have different social behaviours

Check if female workers are underpaid wrt male colleagues

Check if a given mutation is over-represented in some cancer patients
Statistical tests

T-test
Fisher exact test
Chi-square
Randomization
Tests characteristics

Data: categorical, discrete, continuous

Samples: 1, 2 or more

multiple samples: paired – unpaired

Population distribution: Normal, not Normal

Test: one- or two-tailed
Example

100 consumers, wine A vs B scoring

Basketball player BMI

Is racism more common in wealthy countries

Do girls have better marks than boys in high school?
<table>
<thead>
<tr>
<th>How many samples?</th>
<th>Paired/Unpaired</th>
<th>All sample(s) are drawn from a normal distribution?/Parametric (P) or non-parametric test (NP)?</th>
<th>Name of the statistical test</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 sample only</td>
<td>Yes/P</td>
<td>One sample t-test</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No/NP</td>
<td>Wilcoxon rank sum test, One Sample Chi-Square test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paired</td>
<td>Yes/P</td>
<td>Paired t-test</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No/NP</td>
<td>Wilcoxon matched pairs test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 samples</td>
<td>Unpaired</td>
<td>Yes/P</td>
<td>Independent samples t-test</td>
<td>It assumes that the two samples have equal variance (in other words that the difference between the variance of the two samples has not statistical significance). The F test may be used to prove this assumption.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No/NP</td>
<td>Mann-Whitney U test</td>
<td>We may observe that is only one nonparametric test for unpaired data, instead of 2 tests for parametric data. This is happened because a nonparametric tests will not rely on assumptions that the data are drawn from a normal distribution, thus the use of variance become meaningless</td>
</tr>
<tr>
<td>Paired</td>
<td>Yes/P</td>
<td>Repeated-measures one-way ANOVA</td>
<td></td>
<td>Some post hoc tests are available, able to make comparison between each and every pair of samples from the experiment.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Friedman's test</td>
<td>Post hoc tests are available, able to make comparison between each and every pair of samples from the experiment.</td>
<td></td>
</tr>
<tr>
<td>3 or more samples</td>
<td>Paired</td>
<td>Yes/P</td>
<td>One-way ANOVA</td>
<td>Post hoc tests are available</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No/NP</td>
<td>Kruskal-Wallis test</td>
<td>Post hoc tests are available</td>
</tr>
</tbody>
</table>
T-test

Parametric

1 or 2 classes

paired or unpaired

continuous variable

in R:

t.test

William Sealy Gosset
Student distribution

t-test

The standard error for means is:

$$SE = \frac{\sigma_0}{\sqrt{n}} = \sigma_{\bar{x}}$$
t-test

Hence for one mean compared to a hypothesis:

\[ t = \frac{\bar{x} - H_0}{\frac{\sigma_0}{\sqrt{n}}} \]

Each t value comes with a certain degree of freedom  \( \text{df} = n - 1 \)
The test statistic looks like this:

\[ t = \frac{98 - 100}{15} = \frac{-2}{0.75} = 2.67 \]

There are \( n - 1 = 399 \) degrees of freedom.

The results are printed out by a computer or looked up on a t-test table.
The critical value for 399 degrees of freedom is about 1.65
Exercise

The average height of the American population is 73''

Are NBA players significantly taller?

Are Guards taller than Forwards?

Are older (>25) players taller than younger ones?

Are older (>25) players heavier than younger ones?
in R

One sample T test:

t.test(x=V1, mu=73)

Two samples (unpaired)

t.test(x=V1,y=V2)

alternative : "two.sided", "greater" or "less"
Fisher exact test

Used for categorical variables

Small sample size

Contingency table
Fisher exact test

Which packaging do you like the best

Box

Bottle
```r
milk <- matrix(c(10, 5, 3, 8), nrow = 2,
               dimnames = list(Container = c("Bottle", "Box"),
                                Sex = c("Male", "Female")))

fisher.test(milk)
```
Exercise

Calculate if Forwards and Centers have the same probability of being taller than 80''
Chi Square

Similar to Fisher exact Test

non-parametric

test whether the observed proportions for a categorical variable differ from hypothesized proportions.

Less accurate for small samples (<20), good for large samples
Exercise

Use the chi square test on the milk preference data

let's see if the 2 predicted distributions were accurate
Randomization tests

Non parametric, brute force

Low accuracy

Computationally intensive

Complex hypotheses (e.g. not involving means)
Multiple hypothesis correction: Bonferroni, Benjamini Hochberg
Problem of multiple testing

Controlling the FWER:
  • Bonferroni
  • Bonferroni-Holm

Controlling the FDR:
  • Benjamini-Hochberg

Case study
Claim: Wonder pill has an effect!

- **Random** group of people
- Measure 100 variables before and after taking the pill: Weight, blood pressure, heart rate, blood parameters, etc.
- Compare before and after using a paired t-test for each variable on the 5% significance level
- **Breaking news:** 5 out of 100 variables indeed showed a significant effect!
Type I and II errors – a quick brush up

Type I and II Errors

<table>
<thead>
<tr>
<th>Decision</th>
<th>$H_0$ True</th>
<th>$H_0$ False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do Not Reject $H_0$</td>
<td>Correct Decision $1 - \alpha$</td>
<td>Incorrect Decision Type II Error $\beta$</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>Incorrect Decision Type I Error $\alpha$</td>
<td>Correct Decision $1 - \beta$</td>
</tr>
</tbody>
</table>

$\alpha = P(\text{Type I Error}) \quad \beta = P(\text{Type II Error})$
The problem of Multiple Testing

Type 1 error: Rejects $H_0$ if $H_0$ is actually true

In example: Declare that wonder-pill \textit{changes variable}, if in reality there is \textit{no change}

Let's assume, that wonder-pill has no effect at all.

Then:
Every variable has a 5\% chance of being “significantly changed by the drug”
Why Multiple Testing Matters

• In general, if we perform $m$ hypothesis tests, what is the probability of at least 1 false positive?

\[
P(\text{Making an error}) = \alpha
\]

\[
P(\text{Not making an error}) = 1 - \alpha
\]

\[
P(\text{Not making an error in m tests}) = (1 - \alpha)^m
\]

\[
P(\text{Making at least 1 error in m tests}) = 1 - (1 - \alpha)^m
\]
So, what’s the solution?

What Correcting for Multiple Testing Mean?

When people say “adjusting p-values for the number of hypothesis tests performed” what they mean is controlling the Type I error rate.

- Very active area of statistics - many different methods have been described.
- Although these varied approaches have the same goal, they go about it in fundamentally different ways.
Different Approaches To Control Type I Errors

- **Per comparison error rate** (PCER): the expected value of the number of Type I errors over the number of hypotheses,
  \[ PCER = \frac{E(V)}{m} \]

- **Per-family error rate** (PFER): the expected number of Type I errors,
  \[ PFE = E(V) \]

- **Family-wise error rate**: the probability of at least one type I error
  \[ FEWR = P(V \geq 1) \]

- **False discovery rate** (FDR) is the expected proportion of Type I errors among the rejected hypotheses
  \[ FDR = E(V/R \mid R > 0)P(R > 0) \]

- **Positive false discovery rate** (pFDR): the rate that discoveries are false
  \[ pFDR = E(V/R \mid R > 0) \]
Family Wise Error Rate

Many procedures have been developed to control the FWER (the probability of at least one type I error):

- Two general types of FWER corrections:

1. Single step: equivalent adjustments made to each p-value
   - called Bonferroni

2. Sequential: adaptive adjustment made to each p-value
   - called Bonferroni-Holm
Single Step Approach: Bonferroni

Very simple method for ensuring that the overall Type I error rate of $\alpha$ is maintained when performing $m$ independent hypothesis tests

Rejects any hypothesis with p-value $\leq \alpha/m$:

$$\hat{p}_j = \min[mp_j, 1]$$

• For example, if we want to have an experiment wide Type I error rate of 0.05 when we perform 10,000 hypothesis tests, we’d need a p-value of $0.05/10000 = 5 \times 10^{-6}$ to declare significance!
Bonferroni method creates more problems than it solves
(Thomas Perneger, 1998):

Bonferroni correction leads to very high probability to miss proper association!
and...to high probability of type 2 errors

“Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference…”
Improving Bonferroni: Holm-Bonferroni Method

• Simplest sequential method is Holm’s Method
  ➢ Order the unadjusted $p$-values such that $p_1 \leq p_2 \leq \ldots \leq p_m$
  ➢ For control of the FWER at level $\alpha$, the step-down Holm adjusted $p$-values are
    \[ \tilde{p}_j = \min[(m - j + 1) \cdot p_j, 1] \]
  ➢ The point here is that we don’t multiply every $p_i$ by the same factor $m$

• For example, when $m = 10000$:
  \[ \tilde{p}_1 = 10000 \cdot p_1, \tilde{p}_2 = 9999 \cdot p_2, \ldots, \tilde{p}_m = 1 \cdot p_m \]
Example: Holm-Bonferroni

- P-values:
  - $H_0(1)$: 0.005, $H_0(2)$: 0.011, $H_0(3)$: 0.02, $H_0(4)$: 0.04, $H_0(5)$: 0.13
  - $M = 5$ tests; Significance level: 0.05

- Corrected p-value: $0.005 \times 5 = 0.025 < 0.05$: Reject $H_0(1)$
- Corrected p-value: $0.011 \times 4 = 0.044$: Reject $H_0(2)$
- Corrected p-value: $0.02 \times 3 = 0.06$: Don’t reject $H0(3)$ and stop

- Conclusion:
  - Reject $H_0(1)$ and $H_0(2)$, don’t reject $H_0(3)$, $H_0(4)$, $H_0(5)$

Method “Holm” has never worse power than “Bonferroni” and is often better; still conservative
Who Cares About Not Making ANY Type I Errors?

- **FWER** is appropriate when you want to guard against ANY false positives.

- However, in many cases (particularly in *omics) we can live with a certain number of false positives.

- In these cases, the more relevant quantity to control is the false discovery rate (**FDR**).
False discovery rate (FDR) is designed to control the proportion of false positives among the set of rejected hypotheses (R)

- To control FDR at level $\delta$:

  1. Order the unadjusted $p$-values: $p_1 \leq p_2 \leq \ldots \leq p_m$

  2. Then find the test with the highest rank, $j$, for which the $p$ value, $p_j$, is less than or equal to $(j/m) \times \delta$

  3. Declare the tests of rank 1, 2, ..., $j$ as significant

$$p(j) \leq \delta \frac{j}{m}$$

Average fraction of false discoveries $< $ Significance level chosen
FDR example

Controlling the FDR at $\delta = 0.05$

<table>
<thead>
<tr>
<th>Rank (j)</th>
<th>P-value</th>
<th>$(j/m) \times \delta$</th>
<th>Reject $H_0$?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0008</td>
<td>0.005</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.009</td>
<td>0.010</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.165</td>
<td>0.015</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.205</td>
<td>0.020</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.396</td>
<td>0.025</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.450</td>
<td>0.030</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0.641</td>
<td>0.035</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0.781</td>
<td>0.040</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0.900</td>
<td>0.045</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0.993</td>
<td>0.050</td>
<td>0</td>
</tr>
</tbody>
</table>
When to correct for multiple testing?

- **Don’t correct:** Exploratory analysis; when generating hypothesis Report the number of tests you do.

- **Control FDR (typically FDR < 10%):** Exploratory analysis; Screening: Select some features for further, more expensive investigation Balance between high power and low number of false positives

- **Control FWER (typically FWER < 5%):** Confirmatory analysis; use if you really don’t want any false positives
```r
p=c(0.1,0.001,0.05,0.400,0.01,0.022,0.008)

p.adjust(p,method="bonferroni")

p.adjust(p,method="holm")

p.adjust(p,method="fdr")
```
Take Home

Calculate basic statistics in R (mean, median, etc.)
The null hypothesis $H_0$
Type I and II errors
p-value

How to perform in R:
T-test
Chi-square
Fisher exact test

Explain multiple testing corrections (FWER, FDR)
Perform basic corrections in R
data is at

www.cbs.dtu.dk/~pmar/data_epitope.csv

or

/home/people/pmar/public_html/data_epitope.csv
Exercise

1- Load the data and describe the variables
2- calculate the average, sd and median of Values
3- Calculate the mean number of "A" and "L" per sample
   hint: use apply or rowSums
4- Do samples with "A" at position 2
   have lower values than the others?

5- Do samples with "L" at position 2
   have lower values than the others?

6- Samples with "A" and "L" at position 2
   have similar values?
Exercise

Let's define samples with value < 100 as "Strong Binders" and the others as "Weak Binders"

7- Do Strong and Weak Binders have the same composition at position 4?

8- Calculate the same for all positions and store the p-values in the array p

9- correct p with bonferroni, holm and fdr

10 – redo 8 and 9 using a sub-sample of size 200
   hint: use the function sample()