# Exercise Solutions - Statistical methods for clinical trials

Alireza Ghorbani

2025-03-08

## Contents

# 1 Exercise 1 (Bias)

(a) Explain briefly (1-2 sentences) the terms “selection bias”, “allocation bias” and “assessment bias”.

(b) Which design techniques and/or statistical methods can be used to obtain unbiased estimates of the treatment effect in spite of these potential biases?

**Answer:**

a)

- **Selection bias**: Bias occurring when the decision to enter a patient to an RCT is influenced by knowledge of which treatment the patient will receive when entered and when this decision is related to the outcome, for example when selection affected by a risk factor. The patient must enter the study before the treatment is chosen.
- **Allocation bias**: Patients have many factors that can affect the outcome of their therapy. In simple randomization like toss of a coin in some cases especially when sample size is small, balance may not be achieved. This makes formation of comparable treatment groups to fail and the comparison will be biased. Arises from flaws in the randomization process, causing systematic differences between treatment groups. This is often due to inadequate allocation concealment (e.g., unsealed envelopes) or improper randomization, particularly in small trials.
- **Assessment bias**: At the end of or during the trial, various outcome variables are observed, some of them objective like death and some other are subjective like quality of life. If the observer knows the treatment being given to the patient and if the measurement of an outcome variable contains an element of subjectivity, then it is possible that the value of an observation might be influenced by the knowledge of the treatment. The assessment bias can be avoided by blinding. Occurs when outcome evaluators’ knowledge of treatment assignments influences measurements, especially for subjective outcomes (e.g., pain, quality of life).

**Note:** selection bias is a systematic error while allocation bias happens at random.

b)
- *Selection bias*, we can use blinding and randomization,
- *Allocation bias*, we can use randomization after patients entered the study. also adjustment for important factors.
- *Assessment bias*, we can use blinding.

**EXTRA:**

A randomized concurrently controlled clinical trial is simply an experiment performed on human subjects to assess the efficacy of a new treatment (or, in a broader sense, a medical intervention) for some condition. It has two key features, which in the simplest case are as follows:

1. The new treatment is given to a group of patients (called the treated group) and another treatment, often the one most widely used, is given to another group of patients at the same time; this is usually called the control group.
2. Patients are allocated to one group or another by randomization. This can be thought of as deciding on the treatment to be given by the toss of a coin, although more sophisticated methods are usually employed.

Suppose that the outcome variable in group 1 resp. 2 of an RCT is represented by a random variable $X_1$ resp. $X_2$ and that the treatment effect is additive, giving $E(X_1) = \mu + \tau_1$ and $E(X_2) = \mu + \tau_2$, where $\mu$ is the expected value of either $X$ at randomization. $\tau = \tau_2 - \tau_1$ is referred to as the treatment effect and that is what we usually want to estimate. If all goes well with the trial, $X_2 - X_1$ is an unbiased estimator of $\tau$. In practice, however, many things can happen which lead to the possibility that $X_2 - X_1$ is not an unbiased estimator of $\tau$. Types of bias: 1. Selection bias 2. Allocation bias 3. Assessment bias 4. Publication bias 5. Early stopping (see B8)

## 2 Exercise 2 (Randomization—RPB design)

The Random Permuted Block (RPB) design with fixed block length is used in a clinical trial to randomize $n = 48$ patients into two treatment groups. The beginning of the sequence is

$$BBBAAAABABBA...$$

(a) Which block length was used to generate this sequence: 4 or 6?

(b) Considering this sequence, can a doctor predict the treatment assignment of the 6th patient given that he/she knows (i) which treatment was assigned to the 5 first patients and (ii) that an RPB design with fixed block length 6 is used. Which type of bias can this generate? Explain using an example why this type of bias may affect the estimated treatment effect.

(c) Suppose now that the trialists chose the RPB design with random block length (chosen from $\{4, 6\}$) instead of the RPB design with fixed block length. Can the problem outlined in b) be completely ruled out and why?

**Answer:**

a) Block length is 6 since the pattern of BBBAAA fits into a fixed length RPB design with block of six that ensures equal treatments allocation while BBBA in fixed length RBP design with 4 block does not provide equal allocation of treatments.

b) If the doctor knows which treatment was assigned to the 5 first patients

   i) Yes they can predict the assignment of the 6th patient with certainty, since they are using a RBP with fixed block length. [**Note:** In this case even the forth and the fifth patient can be predicted.]

   ii) Since the 6th patient will not receive the treatment randomly this could potentially lead to selection bias also if the doctor involved in the study this could also lead to assessment bias. [The doctor could not let a specific patient take the 6th treatment.]

c) If the sample size is small or with small sequences we are still prone to same bias but as the sample or the sequence is getting bigger the bias is negligible. However, the main reason of randomization is to make sure that the doctor has no information regarding the assignment of the treatment. For example if a doctor guess the pattern of assigning a treatment to a patients are of blocks 6 and 4, then after observing a pattern like BBB… he can for sure know the assignment of the treatment, and the same problem is back again.

**EXTRA:**

**Naive randomization:** is when a doctor flips a coin. this is a natural procedure, but…

1. coin might be biased;
2. doctor knows what treatment is allocated, so double blindness is impossible;
3. overemphasizes the aspect of uncertainty in front of the patient;
4. creates groups of different sizes, problematic in small trials.

Problem 1 can be avoided through the use of a reliable random number generator; problems 2 and 3 if this is done by another person (so that the doctor remains blind).

**Random permuted blocks:** The problem of unbalanced group sizes can be solved by a form of restricted randomization known as random permuted blocks (RPB):

- RPBs of fixed block length
- RPBs with random block length

### Random Permuted Blocks of Fixed Length

Random permuted blocks (RPB) can help solve the problem of unbalanced group sizes in clinical trials by using restricted randomization. Consider sequences of length 4 that comprise two 1s and two 2s:

$$\{1.1122, 2.1221, 3.1212, 4.2211, 5.2112, 6.2121\}.$$

A list of independent identically distributed random numbers is then generated, each element being chosen from 1, 2, 3, 4, 5, and 6 with equal probability. This process results in a sequence in which each patient is equally likely to receive treatment 1 or treatment 2, but the randomization is restricted to ensure that the number of 1s and 2s do not differ by more than 2 at any point.

However, If the trial is organized such that doctors know the treatments patients have received, after 3 patients (modulo 4) are admitted, the next treatment can be predicted with certainty, leading to: **Selection Bias** (Knowledge of treatment allocation) and **Assessment Bias** (If the doctor is not blinded).

### RPBs with Random Block Length

To counteract the drawback, random permuted blocks with random block lengths can be used, e.g., blocks of sizes 4 (6 possible blocks) and 6 (20 possible blocks): 1. **Generate a random number** from the set {4, 6} where Pr(4) = 1/2. 2. **If 4** is chosen, generate a random number from {1, 2, 3, 4, 5, 6} (each equally likely) and set the block accordingly. 3. **If 6** is chosen, generate a random number from {1, 2, . . . , 20} (each equally likely) and set the block accordingly.

In this approach, each patient is equally likely to receive either treatment, The number of patients allocated to two groups can never differ by more than 3, and The possibility of selection bias is negligible.This happens at the beginning of the sequence. This approach ensures better balance and reduces the chance of selection bias while maintaining randomness in treatment allocation.

# 3 Exercise 3 (Randomization—Minimization)

We consider a randomized clinical trial comparing two treatments A and B that should include $n = 20$ patients in total. The trialists assume that the covariates age ( $< 40$ vs. $\geq 40$), sex and smoking status (smoker vs. non smoker) may have a non-negligible effect on the considered outcome. They decide to randomize the patients using the minimization procedure, where the probability to assign a patient to treatment group A is either $p = 0.2, 0.5$ or $1 - p = 0.8$. The 14 first recruited patients are assigned as follows.

| | A | B | Total |
|---|---|---|---|
| | ($n_A = 7$) | ($n_B = 7$) | |
| *Age* <= 40 | 6 | 4 | 10 |
| Male | 2 | 5 | 7 |
| Smoker | 1 | 2 | 3 |

(a) What is the goal of the minimization procedure? Why did the trialist prefer it to RPB with stratification in this case?

(b) Which is the probability that the next patient entering the study, a 38-year old female smoking patient, is assigned to treatment A? Justify your answer.

**Answer:**

a) They want to control the imbalance between the groups not with respect to their size but with respect to their composition. In principle, randomization will produce balance but in practice treatment groups that are not alike with respect to important prognostic factors.

b) If we calculate the summation below for the patients already entered the study with the same prognostic factors as 38-year old female smoking patient,

$$(n^A_{i++} - n^B_{i++}) + (n^A_{+j+} - n^B_{+j+}) + (n^A_{++k} - n^B_{++k}) = (6 - 4) + (5 - 2) + (1 - 2) = 2 + 3 - 1 = 4$$

We can see that more patients with these prognostic features have been assigned more to treatment A rather than B. Therefore, to have a equal assignment of patients to both treatments the next patient with the same features should be less likely to be assignmed to the treatment A.

$$P(\text{Treatment} = A \mid \text{Age} = 38,\ \text{Gender} = \text{Female},\ \text{Smoking} = \text{Yes}) < 0.5$$

And since there are aselction of probabilities $(0.2, 0.5, 0, 8)$, then we can say that $P = 0.2$.

**EXTRA:**

**Stratification**: RPBs are often used in practice in combination with stratification. Stratification is used to control the imbalance between the groups not with respect to their size but with respect to their composition. Although randomization will, in principle, produce groups that are balanced with respect to any prognostic factor, in practice, treatment groups that are not alike with respect to important prognostic factors can and do occur.

**Principle of Minimization:** 1. The 1st patient is allocated by simple randomization. 2. Denote as $n^A_{ijkl}$ and $n^B_{ijkl}$ the number of patients with prognostic factors $i, j, k, l$ allocated to treatment A and B at some stage of the trial. 3. A new patient is entered into the trial who has prognostic factors $w, x, y, z$.

4. Form the sum

$$(n^A_{w+++} - n^B_{w+++}) + (n^A_{+x++} - n^B_{+x++}) + (n^A_{++y+} - n^B_{++y+}) + (n^A_{+++z} - n^B_{+++z})$$

5. If the sum is negative (respectively, positive), then the new patient is allocated to A (respectively, B) with probability $P > 0.5$.

$P < 1$ protects against selection bias, but selection bias is very unlikely in this setting anyway.

# 4 Exercise 4 (Sample size)

(a) Give two reasons why sample size calculation is needed when planning an RCT.

(b) For a standard two-arm clinical trial, write the (approximate) formula giving the needed number of patients per group as a function of the significance level, the power, the clinically relevant difference, and the within-group variance.

(c) All other things remaining equal, does the needed sample size increase or decrease when

1) *the clinically relevant difference increases;*
2) *the significance level increases;*
3) *the power increases;*
4) *the within-group variance increases;*

(d) The size of the control group and of the intervention group has to be determined for a clinical trial on the effect of a special diet on the cholesterin level. The sample size should be (approximately) the same in both groups and such that a power of 80% is achieved. The significance level is set to $\alpha = 0.05$. A reduction of $15mg/dl$ from a basis value of $250mg/dl$ is considered clinically relevant and should be identified by the trial with a power of 80%.The standard deviation with in groups is assumed to be approximately $40mg/dl$. Compute the needed number of patients per group assuming the use of a two-sided two-sample t-test. Take into account the dropout rate of 10% (for both groups).

**Answer:**

a) Calculating the necessary sample size needed to achieve a certain power, given other assumptions. It is also not ethical to expose patients to a treatment if we can already know that it is inferior. Also, it is not ethical to perform experiments on humans if the study is not likely to produce valuable evidence. Furthermore, we can know how fast we can obtain the results and this way the RCT will be also cheaper.

b) We consider the general case of the test statistic $Z = \frac{(\theta-\theta_0)}{\text{se}(n_1 n_2)}$ which (approximately) follows a standard normal distribution under the null-hypothesis $H_0 : \theta = \theta_0$. Let $n_1 = \gamma N$ and $n_2 = (1-\gamma)N$, where $N$ is the total sample size. $N \approx \frac{(Z_{1-\alpha/2}+Z_{1-\beta})^2\sigma^2}{(\theta-\theta_0)^2\gamma(1-\gamma)}$

c) Sample size (Intuitive and mathematical explanation)

1) Needed samples size will **decrease** when the relevant difference increases. It is because now it is much more easier to detect any difference. Also using the sample size formula the denominator increases then therefore the sample size decreases.
2) Needed samples size will **decrease** when the significance level ($\alpha$) increases. Since we would be accepting more type I error. Also When $\alpha$ increases its corresponding $Z$ value decreases and therefore the numerator of the formula.
3) Needed samples size will **increase** when the power increases. It is because we would like to decrease the type II error and since again the numerator increases.
4) Needed samples size will **increase** when the within-group variance increases. It is because the data has more variation and it is harder to detect any difference. Also in the formula, the numerator increases.

d)

$$n_1 = n_2 \approx \frac{2\sigma^2(Z_{1-\alpha/2}+Z_{1-\beta})^2}{\tau_c^2}/(1-rate_{dropout}) =$$

```r
(2*40^2*(qnorm(0.975)+qnorm(0.8))^2)/(15^2)/(1-0.1)
```

```
## [1] 124.0317
```

Therefore $N = 2 \times 124 = 248$

---

# 5 Exercise 5 (Baseline values)

The following data come from a study by Giardiello et al. (1993) in which patients with familiar adenomatous polyposis (FAP) were treated either with Sulindac or Placebo. FAP is an autosomal dominant inherited condition in which numerous adenomatous polyps form mainly in the epithelium of the large intestine.The number of polyps in the colon was determined before randomization as well as 12 months after treatment with Sulindac or Placebo. The following table contains the $log_{10}$ values at randomization and 12 months after treatment.

| Patient ID | Baseline value | 12 months after treatment | treatment (1: Sulindac, 0: Placebo) |
|---|---|---|---|
| 1 | 0.84510 | 0.60206 | 1 |
| 2 | 0.69897 | 1.41497 | 0 |
| 3 | 1.36173 | 1.20412 | 1 |
| 4 | 1.54407 | 1.60206 | 0 |
| 5 | 1.04139 | 1.14613 | 1 |
| 6 | 1.07918 | 1.20412 | 0 |
| 7 | 0.84510 | 1.04139 | 0 |
| 8 | 2.50243 | 2.63749 | 0 |
| 9 | 2.20412 | 1.41497 | 1 |
| 10 | 0.90309 | 0.84510 | 1 |
| 11 | 1.30103 | 1.65321 | 0 |
| 12 | 1.04139 | 1.50515 | 0 |
| 13 | 1.38021 | 1.90309 | 0 |
| 14 | 1.53148 | 1.53148 | 1 |
| 15 | 1.73239 | 1.57978 | 0 |
| 16 | 1.47712 | 1.75587 | 0 |
| 17 | 1.00000 | 0.84510 | 1 |
| 18 | 1.30103 | 0.00000 | 1 |
| 19 | 1.07918 | 0.90309 | 1 |

(a) A medical doctor says: "According to the paired sample t-test, the number of polyps significantly increased within 12 months after treatment in the placebo group, but decreased in the Sudinlac group, so Sudinlac helps significantly". Do you agree with this statement? If no, what is wrong in it?

(b) Test the null hypothesis that the treatment with Sulindac has no effect on the logarithmized number of polyps 12 months after treatment at the significance level 5%, while ignoring the baseline values. You may use R.

(c) Test the same null hypothesisas in (a), but this time considering as an endpoint the difference between logarithmized number 12 months after treatment and at randomization. You may again use R.

(d) Explain qualitatively in which case analysis (c) is preferred to (b) and why. Optional: you might base your explanation on the following formulas (see the course materials for a definition of the notations):

$$var(X_2 - X_1) = 2\sigma^2/N,$$

$$var((X_2 - B_2) - (X_1 - B_1)) = 4\sigma^2(1 - \rho)/N,$$

where $\sigma 2 = var(X_1) = var(X_2) = var(B_1) = var(B_2)$ and $\rho = cor(X_1, B_1) = cor(X_2, B_2)$.

(e) Explain why a regression analysis with the logarithmized number at baseline as covariate may be seen as a good compromise between (b) and (c).

(f) Write the corresponding model, defining all notations carefully.

(g) Explain how you would do this analysis in R.

**Answer:**

a) We can say that number of polyp increased in placebo group and there was a insignificant decrease in Sudinlac group. The treatment effect is not significant after 12 months. The statement "Sudinlac helped significantly" is not correct.

```
# Import data:
data <- data.frame(PatientID = 1:19,
                   Baseline = c(0.8451, 0.69897, 1.36173, 1.54407, 1.04139,
                                1.07918, 0.84510, 2.50243, 2.20412, 0.90309,
                                1.30103, 1.04139, 1.38021, 1.53148, 1.73239,
                                1.47712, 1.0000, 1.30103, 1.07918),
                   AfterTreat = c(0.60206, 1.41497, 1.20412, 1.60206, 1.14613,
                                  1.20412, 1.04139, 2.63749, 1.41497, 0.84510,
                                  1.65321, 1.50515, 1.90309, 1.53148, 1.57978,
                                  1.75587, 0.84510, 0.0000, 0.90309),
                   Rand = c(1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1,
                            1, 1))
head(data,2)
```

```
##   PatientID Baseline AfterTreat Rand
## 1         1  0.84510    0.60206    1
## 2         2  0.69897    1.41497    0
```

```
# (a) Test different before-after, two groups successively
t.test(x = data$Baseline[data$Rand == 1], y = data$AfterTreat[data$Rand == 1], paired=TRUE)
```

```
## 
##  Paired t-test
## 
## data:  data$Baseline[data$Rand == 1] and data$AfterTreat[data$Rand == 1]
## t = 2.0601, df = 8, p-value = 0.07335
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.03681277  0.65349499
## sample estimates:
## mean difference 
##       0.3083411
```

```
t.test(x = data$Baseline[data$Rand == 0], y = data$AfterTreat[data$Rand == 0], paired=TRUE)
```

```
## 
##  Paired t-test
## 
## data:  data$Baseline[data$Rand == 0] and data$AfterTreat[data$Rand == 0]
## t = -3.3706, df = 9, p-value = 0.008249
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.45041281 -0.08863519
## sample estimates:
## mean difference 
##       -0.269524
```

b) By ignoring the baseline values and comparing the logarithmized number of polyps 12 months after treatment for both groups we can see that there is a significant different between the two groups. The log number of polyps are higher in Sulindac group.

```
# (b) Analysis ignoring baseline values
# t-test:
t.test(x = data$AfterTreat[data$Rand == 1], y = data$AfterTreat[data$Rand == 0])
```

```
##
##  Welch Two Sample t-test
##
## data:  data$AfterTreat[data$Rand == 1] and data$AfterTreat[data$Rand == 0]
## t = -3.3262, df = 16.512, p-value = 0.004128
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.1223623 -0.2499415
## sample estimates:
## mean of x mean of y
## 0.9435611 1.6297130
```

```
t.test(x = data$AfterTreat[data$Rand == 1], y = data$AfterTreat[data$Rand == 0],
       var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  data$AfterTreat[data$Rand == 1] and data$AfterTreat[data$Rand == 0]
## t = -3.3374, df = 17, p-value = 0.003902
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.1199147 -0.2523891
## sample estimates:
## mean of x mean of y
## 0.9435611 1.6297130
```

```
qt(0.975, df = 17)
```

```
## [1] 2.109816
```

c) The effect of Sulindac on the log number of polyps is significantly higher, decreased more, than the effect of placebo.

```
# (c) Analysis considering difference between values 12 months after treatment and baseline values
t.test(x = (data$AfterTreat - data$Baseline)[data$Rand == 1],
       y = (data$AfterTreat - data$Baseline)[data$Rand == 0], var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  (data$AfterTreat - data$Baseline)[data$Rand == 1] and (data$AfterTreat - data$Baseline)[data$
## t = -3.5053, df = 17, p-value = 0.002713
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9256790 -0.2300513
## sample estimates:
##  mean of x  mean of y
## -0.3083411  0.2695240
```

d) Basically when $\rho$ is high and close to one, $var((X_2-B_2)-(X_1-B_1))$ could be smaller than $var(X_2-X_1)$ which can increase the power of the test otherwise, when the correlation between baseline value and after 12 months values is low then this could lead to noise and decrease in the power of the test.

$$\frac{2\sigma^2}{N} > \frac{4\sigma^2(1-\rho)}{N} \rightarrow \frac{1}{2} > (1-\rho) \rightarrow p > \frac{1}{2}$$

e) A regression analysis that includes baseline values as a covariate adjusts for individual differences in initial polyp counts, addressing potential imbalances between treatment groups. This approach reduces residual variance in the outcome (post-treatment values) by accounting for the correlation ($\rho$) between baseline and post-treatment measurements, as implied by the variance formulas in part (d). When $\rho$ is high, including baseline improves precision similar to analyzing differences (part c), but without discarding information about baseline variability. Unlike part (b), which ignores baseline and risks confounding, and part (c), which assumes differences are the optimal metric, regression leverages baseline as a predictor. This achieves a balance: it controls for baseline variability while directly modeling the treatment effect, leading to more reliable and efficient estimates.

f)

$$x_i = \beta_0 + \beta_1 * G_i + \beta_2 b_i + \epsilon_i \rightarrow x_i - \beta_2 b_i = \beta_0 + \beta_1 * G_i + \epsilon_i$$

Where $x_i$ is the $log_{10}$ number of polyps after 12 months, $\beta_0$ mean base effect after 12 months, $\beta_1$ is the treatment group effect, and $b_i$ is the baseline $log_{10}$ number of polyps.

**Note:** The regression model automatically considers $\rho$.

g) The coefficient for Rand ($\beta_1$) estimates the treatment effect. A significant p-value (e.g., $< 0.05$) indicates Sulindac's effect after adjusting for baseline. The coefficient for Baseline ($\beta_2$) reflects how baseline values predict post-treatment outcomes.

```r
# (g) Baseline as covariate
model <- lm(AfterTreat ~ Baseline + Rand, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = AfterTreat ~ Baseline + Rand, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97598 -0.15293  0.06775  0.20482  0.40345
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7322     0.2532   2.892 0.010617 *
## Baseline      0.6598     0.1695   3.892 0.001296 **
## Rand         -0.6147     0.1530  -4.018 0.000994 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3306 on 16 degrees of freedom
## Multiple R-squared:  0.6896, Adjusted R-squared:  0.6508
## F-statistic: 17.78 on 2 and 16 DF,  p-value: 8.608e-05
```

# 6 Exercise 6 (Cross-over design)

The following data come from a cross-over trial with AB/BA-design including $n = 20$ patients comparing a verum with a placebo for the treatment of alcoholism with respect to anapproximately normally distributed clinical endpoint. Each patient received the verum in one of the periods and the placebo in the other.

| | Placebo/Verum | | | Placebo/Verum | |
|---|---|---|---|---|---|
| ID | Period 1 | Period 2 | ID | Period 1 | Period 2 |
| 1 | 326 | 228 | 13 | 243 | 618 |
| 2 | 808 | 353 | 14 | 179 | 177 |
| 3 | 355 | 188 | 15 | 215 | 38 |
| 4 | 295 | 220 | 16 | 388 | 341 |
| 5 | 121 | 156 | 17 | 188 | 159 |
| 6 | 301 | 286 | 18 | 249 | 246 |
| 7 | 270 | 211 | 19 | 292 | 465 |
| 8 | 549 | 298 | 20 | 190 | 229 |
| 9 | 340 | 138 | | | |
| 10 | 385 | 264 | | | |
| 11 | 186 | 207 | | | |
| 12 | 206 | 216 | | | |
| Average | 345.17 | 230.42 | | 243.00 | 284.13 |
| Empirical variance of the difference | 19640.93 | | | 27562.41 | |

(a) Explain the following model that can be used to analyse the data:

| | Period 1 | Period 2 |
|---|---|---|
| Group 1 | $x_{i1} = \mu + \pi_1 + \tau_A + \xi_i + \epsilon_{i1}$ | $x_{i2} = \mu + \pi_2 + \tau_B + \xi_i + \epsilon_{i2}$ |
| Group 2 | $x_{i1} = \mu + \pi_1 + \tau_B + \xi_i + \epsilon_{i1}$ | $x_{i2} = \mu + \pi_2 + \tau_A + \xi_i + \epsilon_{i2}$ |

(b) Explain how you would compute an estimate of the treatment effect $\tau$ using the data at hand.

(c) Explain the term "carry-over effect" from a medical perspective.

(d) Modify the model specified in (a) such that it takes carry-over effects into account.

(e) Explain based on (d) in which situation the estimate given in (b) is biased.

**Answer:**

a) This model is used to analyze AB/BA-design in which it ensures that pairs of measurements from a single patient be kept together and considers the systematic differences between the treatment periods as well as treatment effects.

b) If $\bar{d_i}$ is mean sample difference in group $i$, we can use $\frac{(\bar{d_1}-\bar{d_2})}{2}$ to estimate $\tau$.

$$E(\bar{d_i}) = \sum_{i=1}^{2} E(x_{i1} - x_{i2}) = (\pi_1 + \tau_A - \pi_2 - \tau_B) + (\pi_1 + \tau_B - \pi_2 - \tau_A) = 2(\pi_1 - \pi_2)$$

c) It is potential problem with a crossover trial in which the effects of the treatment given in period 1 may still persist during period 2.

d) We can add a term $\gamma_A$ to the previous model for period 2 of group 1 and a term $\gamma_B$ for period 2 of group 2.

| | Period 1 | Period 2 |
|---|---|---|
| Group 1 | $x_{i1} = \mu + \pi_1 + \tau_A + \xi_i + \epsilon_{i1}$ | $x_{i2} = \mu + \pi_2 + \tau_B + \gamma_A + \xi_i + \epsilon_{i2}$ |
| Group 2 | $x_{i1} = \mu + \pi_1 + \tau_B + \xi_i + \epsilon_{i1}$ | $x_{i2} = \mu + \pi_2 + \tau_A + \gamma_B + \xi_i + \epsilon_{i2}$ |

e) If we ignore carryover and estimate $\tau$ as $\frac{(\bar{d_1}-\bar{d_2})}{2}$, what would we actually be estimating is $E(\frac{(\bar{d_1}-\bar{d_2})}{2}) = \tau - \frac{1}{2}\gamma$, where $\gamma = \gamma_A - \gamma_B$. If there is a carryover effect ($\gamma \neq 0$), then the estimator is biased.

**EXTRA**

### Crossover Trials

In all trials considered so far, each patient has received just one of the treatments being compared. This is natural for a majority of diseases and conditions, for example the investigation of a new material for use in the construction of plasters for fractures, new approaches to removing the appendix, and antithrombolytic treatment following heart attacks. However, for some conditions such as asthma or diabetes which cannot be cured or for renal disease necessitating kidney dialysis twice a week, each patient could be given several treatments successively. One may then consider the within-patient difference $d_i = x_{i2} - x_{i1}$. Such trials are known as crossover trials and have the advantage that they allow more precise treatment comparisons.

### The AB/BA Design

For two treatments, the simplest form of crossover design would be to give each patient treatment A and then follow it with B. This may introduce a bias for different reasons:

- Bias Type 1 (Evolution of the Disease): Disease may tend to get worse or better independently of treatment.
- Bias Type 2 (Patients and Doctors' Behavior, Organization of the Trial): Patients may acclimate to trial procedures (e.g., lower blood pressure due to reduced anxiety in later visits. Staff may improve at administering treatments or measurements over time.
- Bias Type 3 (Confounding with Temporal Effects): Time-dependent factors like seasonal effects (e.g., pollen allergies in spring) or weekday-specific lab variations (e.g., equipment calibration differences) may skew results. Example: All readings from the laboratory were higher on Monday than on Tuesday. Respiratory diseases may depend on the season due to allergies, viral infections, etc.

Solution: The simplest form of crossover trial consists of randomly allocating patients to two groups. Patients in group 1 receive the treatments in the order AB, whereas those in group 2 receive them in the opposite order. The times when treatments are given are referred to as periods.

| | **Period 1** | **Period 2** |
|---|---|---|
| **Group 1** | A | B |
| **Group 2** | B | A |

# 7 Exercise 7 (Cluster-randomized design)

(a) Outline situations in which a cluster-randomized design might be appropriate (1-2 sentences).

(b) Why is it wrong from a statistical point of view to analyse the data using, say, a t-test for two independent samples while ignoring the dependence between patients from the same cluster?

(c) Explain the following formula commonly used for sample size calculations for cluster-randomized designs

$$N \approx \frac{2\sigma^2(1+\rho(n_a-1))(Z_{1-\alpha/2}+Z_{1-\beta})^2}{\tau_c^2}$$

(d) The total number of patients in the two treatment arms being equal, is the power higher in a cluster randomized trial or in a standard trial with independent patients who are individually randomized?

**Answer:**

a) A cluster-randomized design is appropriate when interventions are applied to groups or clusters (e.g., communities, schools, or clinics) rather than individuals, to account for group-level effects and reduce contamination across participants within the same cluster. This design is used when individual randomization is impractical or when the intervention's impact is inherently linked to the group's collective behavior or environment.

b) Analyzing cluster-randomized data with a t-test for independent samples ignores the intra-cluster correlation, leading to an underestimation of variance and incorrect p-values. This can result in unreliable conclusions and compromised statistical validity.

c) In this formula, $n_a$ is the expected average number of individuals in a cluster, $\rho$ denotes the correlation of the outcomes within a cluster and with the rest of the notation as before for sample size calculation. Here the the variance of mean is $\frac{\sigma^2(1+\rho(n_a-1))}{N}$.

Intra-Cluster Correlation (ICC): The Intra-Cluster Correlation Coefficient ($\rho$) measures the similarity of outcomes within clusters compared to between clusters:$\rho = \frac{\sigma_b^2}{\sigma_b^2+\sigma_w^2}$ Where:

- $\sigma_b^2$ = Between-cluster variance
- $\sigma_w^2$ = Within-cluster variance

Interpretation:

- $\rho = 0$: No clustering effect (perfect independence)
- $\rho = 1$: Perfect similarity within clusters
- Typical values: 0.01-0.05 in healthcare studies

d) For equal total sample sizes ($N$ total patients): $\text{Power}_{\text{cluster}} \leq \text{Power}_{\text{individual}}$

1. Variance in Individual Randomization: $\text{Var}_{\text{ind}} = \frac{\sigma^2}{N}$
2. Variance in Cluster Randomization: $\text{Var}_{\text{cluster}} = \frac{\sigma^2[1+\rho(n_a-1)]}{N}$
3. Design Effect: $DE = 1+\rho(n_a-1) \geq 1$

**Key Implications**:

- When $\rho > 0$: $\text{Var}_{\text{cluster}} > \text{Var}_{\text{ind}}$ → Wider confidence intervals → **Lower power**
- When $\rho = 0$: $DE = 1$ → Equal power (special case)
- Typical scenario ($\rho > 0$): Cluster trials need $2-5\times$ more subjects for equivalent power

# 8 Exercise 8 (Subgroup analyses)

A clinical trial is conducted to compare two treatments of psoriasis. After 16 weeks a doctor evaluates whether an improvement can be observed. The data are given in the following table for patients with white skin and patients with dark skin separately

| | white skin | | dark skin | |
|---|---|---|---|---|
| treatment group | 1 | 2 | 1 | 2 |
| improved | 9 | 5 | 10 | 3 |
| not improved | 17 | 21 | 15 | 20 |

(a) Which test could be performed to test the treatment effect within a skin color subgroup? Explain it briefly (formal tested null-hypothesis, test statistic, null distribution).

(b) Explain how you would test the null hypothesis that the treatment effect is the same for the two considered skin types.

(c) Suppose the statistician performs the same analysis for a total of 25 partitions into two subgroups ((1) male and female, (2) age $< 50$ and age $\geq 50$, ..., (25) BMI $> 30$ and BMI $\geq 30$ ) and finds that the treatment effect is significantly better in young patients (age $< 50$) than in older patients (age $\geq 50$). Do you tend to trust this result? Justify your answer based on the concept of multiple testing.

(d) Sketch how the Bonferroni procedure works using the above example.

**Answer:**

a) The null hypothesis for treatment effect within a skin color should be that there are no difference between the two treatments, basically $\hat{\theta}_2^W - \hat{\theta}_1^W = 0$ and $\hat{\theta}_2^B - \hat{\theta}_1^B = 0$. The statistic should be 2 Z statistic for proportions, one for the white skin and the other for the dark skin. Under the null hypothesis both statistics distribution should follow a standard normal distribution. [Also $\chi^2$ and Fisher Exat Test]

b) To test the null-hypothesis $\tau_W = \tau_B$, we can use the following statistics

$$Z = \frac{(\hat{\pi}_2^W - \hat{\pi}_1^W) - (\hat{\pi}_2^B - \hat{\pi}_1^B)}{\sqrt{(\nu_1^W + \nu_2^w) + (\nu_1^B + \nu_2^B)}}$$

where $V_1^W = \frac{\pi_1^W \times (1-\pi_1^W)}{n_1^W}$ and under the null hypothesis $Z$ distribution is $N(0,1)$. It can also be solved by a logistic regression, $\ln(P(Y=1)) = \beta_1 + \beta_2 \times \text{Skin Color}_i + \beta_3 \times \text{group}_i + \epsilon_i$.

c) Subgroup analyses are generally not taken into account in sample size calculations. Too many subgroup analyses lead to a serious multiple testing problem. When we perform 25 multiple testing each with $\alpha = 0.05$, we get an overall type I error of much higher than 0.05. It is important to notice that the tests here are not independent.

d) Bonfroni correction considers the $\alpha_{multi} = \frac{\alpha}{n}$. In this case, $\alpha_{multi} = \frac{0.05}{25} = 0.002$. That means Bonfroni is more conservative meaning it has less power and the probability of early stopping is pretty low.

# 9 Exercise 9 (Interim analyses)

For a clinical trial comparing two treatment strategies for lung tumor, the trialist decides to use a group sequential design. Let K denote the maximum number of groups to be included in the trial.

(a) What is the maximal number of interim analyses?

(b) Explain why early stopping based on interim analyses may be recommended from an ethical point of view.

(c) The trialist sets $K$ to $K = 5$ with a group size of $2m = 40$ and chooses to apply the O'Brien-Fleming procedure. Suppose the four corresponding p-values are $p_1 = 0.149$, $p_2 = 0.014$, $p_3 = 0.067$, $p_4 = 0.018$ (where $p_k$ denotes the p-value obtained based on the first $2m * k$ patients). What happens with the O'Brien Flemming procedure? What would have happened if the trialist had chosen Pocock's procedure?

Table 7: Nominal Significance level for Mth Interim Analysis for Various Stopping Rules with $\alpha = 0.05$

| M | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Pocock | 0.0158 | 0.0158 | 0.0158 | 0.0158 | 0.0158 |
| O'Brien Flemming | $5 \times 10^{-6}$ | 0.0013 | 0.0085 | 0.0228 | 0.0417 |
| Haybittle Pito | 0.001 | 0.001 | 0.001 | 0.001 | 0.05 |
| Flemming Harrington, & O'Brien | 0.0038 | 0.0048 | 0.0053 | 0.0044 | 0.0432 |

d) Discuss the advantages and inconveniences of the O'Brien Fleming procedure compared to Pocock's procedure (you may use the example from (c)). How does Bonferroni's procedure behave compared to Pocock's procedure?

**Answer:**

a) The analyses based on $1, ..., K-1$ groups are denoted as interim analyses; the analysis based on $K$ groups is the final analysis.

b) Early stopping focus on the importance of preventing patients from receiving inferior or harmful treatments and allowing more effective treatments to become available sooner. Early stopping also has financial and practical advantages.

c) In the O'Brien Flemming procedure, by comparing the p-values for each step by its nominal signifiacnce level, there would be a stoping at step 4 since $p_4 < M_4$. If the trialist have had chosen the Pocock's procedure, by the same comparision of nominal significance level for this procedure and the p-values the early stopping can happen at step 2.

d) i) O'Brien Flemming procedure is more conservative and less efficient with longer duration. However, it is more inforamative and more likely to show the the effect.

ii) In Bonfroni we assume that the tests are independent and control the overal type I error by deviding the overall alpha by the number of the tests $\alpha_{multi} = \frac{\alpha}{K}$. But, In the Pocock's we do not assume that tests are independent and using the formula below

$$\alpha_{multi} = \sum_{k=1}^{K} P(Z_k > c_k \mid Z_1 < c_1, Z_2 < c_2, \ldots, Z_{k-1} < c_{k-1})$$

to calculate type I error for each test as if they are equal. Here, $c_k$ are the critical values for each analysis. The significance levels $\alpha_k$ are derived to ensure the overall Type I error rate is controlled.

**EXTRA:**

**Accumulating Data in Clinical Trials**

Clinical trials generally run for many months or even years. As a consequence, for a trial with a fixed size, the results for the patients who enter at the start of the trial will become available before the patients entering the trial later are recruited.

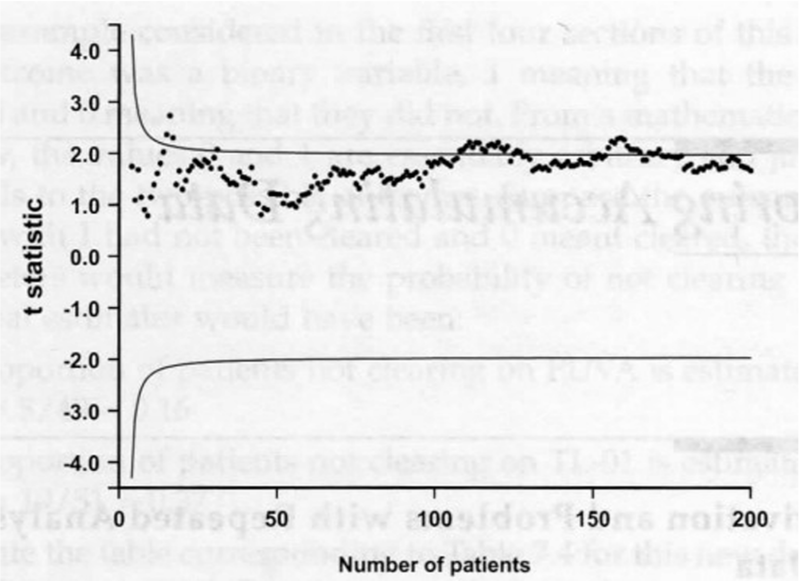
Trials are undertaken when the investigator does not have evidence whether one treatment is superior to another, but it is conceivable that this situation changes as the trial proceeds. If this is so, and the investigator does not take into account this new evidence, then the patients entered into the trial later, who receive the inferior treatment, will be receiving a treatment that could have been known to be inferior based on the evidence from the first part of the trial.

It is important to take this issue into account to maintain an ethically defensible trial. Moreover, early stopping also has financial and practical advantages. However, if a trial is stopped early on the basis of naive methods, serious statistical problems arise that can undermine the results of the study. In extreme cases, the medical community may discount the results from this trial and decide that a new trial is needed, resulting in more patients being exposed to the inferior treatment than if the trial had not been stopped.

**Simulation Study**

A simulation involving 200 patients was conducted to illustrate the impact of sequential analysis. Outcomes from a trial of 200 patients were simulated from a single normal distribution, and the patients were allocated to one of two treatments at random.

The two groups were compared with a t-test using just the first 4 patients, the first 5 patients, etc. The resulting t-statistics are plotted against the number of patients used in the comparison, together with the 95% confidence limits for the relevant t-statistics.


t statistic
Number of patients