Advanced decision modelling in the context of Health Technology Assessment

Juan Pablo Diaz Martinez

2021-11-19

Contents

Introduction

Bringing a new health technology to market and into the hands of a patient is a long process. Most of the times patients, who have a medical need, ask themselves why does it take so long to make the health technology available to everyone. When a health technology is in the market, it usually took between 5 to 10 years to make it available.

Depending on the country, governments usually are involved in the reimbursement process. They usually ask the next questions when a new health technology is available:

- How much does it cost?
- Will it save lives and/or improve quality of life?
- Do we have enough budget to fund it?
- If we have a pool of interventions for a specific disease, which one/ones should we reimburse?

Moreover, physicians, patients, insurance plans, and advocacy groups play an important role when new technologies are available in the market (why?). Even though a new technology see the light (i.e. it has proved to be safe and effective), insurance providers or the government will not necessarily cover it. Usually they argue that the new technology is "Not cost-effective" or "Not have good value for money". *These notes aim to provide all the necessary tools to decide if a new intervention has a good value-for-money.* It is important to stress that value-for-money decision is only one of many questions that are asked by one of the users of a **health technology assessment (HTA)**: patients, healthcare workers, government, and others.

Why reimbursment submissions fail?

According to Goeree (2015), the reasons for rejection are:

1. Inappropriate comparator. Lack of proper statistical analysis.

2. Inappropriate outcome. Use of surrogates.
3. Inappropriate analysis. Lack of robust evidence for costs and quality of life.
4. High cost to the government.

Topics of the course

1. What is HTA?
2. Introduction to decision-analytic models
3. Good practices in decision modelling
4. Evidence-based medicine
5. Decision tree-models
6. State-transition models with the Markov assumption
7. Partitioned survival models
8. Microsimulation
9. Discrete-event simulation
10. Uncertainty and decision-making
11. Presentation of results

Statistical computing

The use of open-source programming languages, such as `R`, in health decision sciences is growing and has the potential to facilitate model transparency, reproducibility, and shareability. However, realizing this potential can be challenging. Models are complex and primarily built to answer a research question, with model sharing and transparency relegated to being secondary goals. Moreover, many decision modelers are not formally trained in computer programming and may lack good coding practices, further compounding the problem of model transparency. **Therefore, throughout this course, the programming language `R` will be used to show its potential for advanced modelling in the context of HTA.**

For this course, we will be using the book "R for Data Science". To install `R` and `Rstudio`, instructions are provided in Chapter 1 of this book. We will also use Excel throughout this course.

Evaluation

Item	Percentage	Due date
Assignment 1	15%	Nov 27, 2021

Item	Percentage	Due date
Assignment 2	15%	Dec 23, 2021
Take-home exam	30%	Jan 7, 2022
Project proposal	5%	Nov 22, 2021
Project presentation	5%	Jan 14, 2022
Final project	30%	Jan 17, 2022

The intent is to allow the students to demonstrate their mastery of this class through the following way. **Project proposal, presentation and final project will be done in pairs**.

Asssignments

The assignments are handed out approximately two weeks prior to the due date. Late work will not be marked, with the exception of an advance permission from the instructor.

Project proposal

(1 page)

The final deliverable for this course is a mini-HTA on a medical technology (preferably something topical), with a focus on the quantitative aspect of it. Given that the translation of a health policy question into a relevant research question is an essential first step in the conduct of HTA, students are required to formulate a research question and submit for grading purposes. This should include at leas some of the following: an overview of the technology being assessed; a clear specification of the policy problem; and the research question(s) (including PICO) with objectives.

Project presentation

(20 minutes with extra 5 minutes for questions)

Students will be expected to present their final course paper and answer questions. Student will be graded on their presentations.

Final project

(20 pages double-spaced)

The main assignment will require students to produce a scaled down HTA, with a focus on the quantitative aspect of it. The objective of the final project is for the

student to show that they have obtained a clear understanding of the advanced methods in decision modelling in the context of HTA. More information will be provided throughout the course, but the paper should contain the following:

a) Background and technology overview
b) Formulation of the question you are trying to answer through your mini-HTA
c) Review of the clinical literature
d) Description of the structure of the model
e) Description of the function of the model
f) Results
g) Conclusions

Bibliography

Briggs, A., Sculpher, M., & Claxton, K. (2006). Decision modelling for health economic evaluation. Oxford University Press.

Gray, A. M., Clarke, P. M., Wolstenholme, J. L., & Wordsworth, S. (2011). Applied methods of cost-effectiveness analysis in healthcare (Vol. 3). Oxford University Press.

Edlin, R., McCabe, C., Hulme, C., Hall, P., & Wright, J. (2015). Cost effectiveness modelling for health technology assessment: a practical course. Springer.

Chapter 1

What is HTA?

1.1 Pre-session readings

Goodman, C. S. (2004). Introduction to health technology assessment. The Lewin Group. virginia, USA. link. Chapters 1, 2, and 5.

Briggs, A., Sculpher, M., & Claxton, K. (2006). Decision modelling for health economic evaluation. Oxford University Press. Chapter 1.

Chapters 1 and 2 of *R for data science*.

1.2 Definition and rationale

The first thing that we need to know is the definition of a **health technology**. A health technology is any intervention that may be used to promote health, to prevent, diagnose or treat disease or for rehabilitation or long-term care.

Questions

1. List some examples of health technologies.

Depending on the agency, health technology assessment has a broad spectrum of definitions:

"HTA is a multidisciplinary process that uses explicit methods to determine the value of a health technology at different points in its lifecycle. The purpose is to inform decision-making in order to promote an equitable, efficient, and high-quality health system." INAHTA

"Health technology assessment is a multidisciplinary process that uses explicit methods to determine the value of a health technology at different points in its

lifecycle. The purpose is to inform decision-making in order to promote an equitable, efficient, and high-quality health system." EUnetHTA

"A comprehensive, objective, evidence-based analysis of the clinical effectiveness, cost-effectiveness and broader impact of drugs, medical technologies and health systems. HTA examines technologies at all stages of their life cycle, from development through to maturity and obsolescence." CADTH

The purpose of HTA is to support/help decision makers by identifying technologies that will improve health outcomes and deliver value for every dollar invested.

- Does a new health technology offer a clinical advantage over the alternatives/standard approaches?
- Is it worth the investment?
- Can I pay for it?
- Who would benefit from it?
- Any ethical, social or legal issues

But, what are the reasons for conducting HTAs?

- Increased demand for healthcare (why?)
- Soaring healthcare costs
- Increased rate of diffusion of new technologies and associated evidence

Once we have seen the definition and rationale for conducting HTAs, it is important to talk about the potential users.

- Government
- Managers in hospitals
- Healthcare workers
- Researchers

1.3 HTA process

1. Identification and prioritization of technologies
2. Clear specification of the problem
3. Technology assessment and review
    - Evidence and systematic literature review
    - Aggregation and appraisal of evidence
    - Synthesize and consolidate
    - Collect primary data (if necessary)
    - Economic evaluation, budget and health system impact

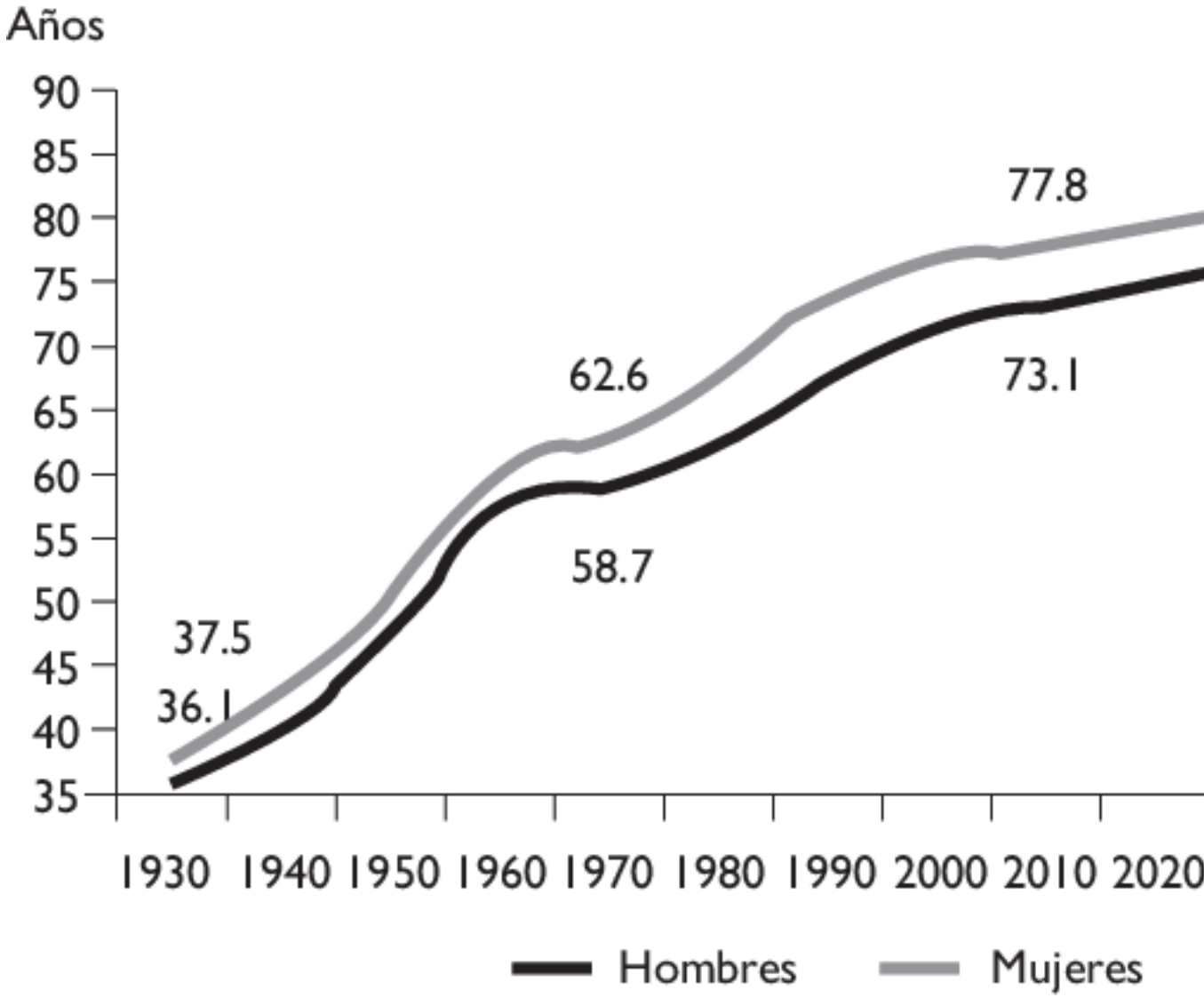

Figure 1.1: Life expectancy in Mexico. Source: CONAPO

    - Assessment of social, legal, and ethical consideration
    - Formulation of finding

4. Dissemination and implementation of recommendations
5. Monitor the impact of assessment reports

1.3.1 Identification and prioritization of technologies

- Drugs seeking public or private reimbursement
- Variable for non-drug technologies. However candidates:
    - High potential to improve health outcomes, reduce harm or decrease costs with similar efficacy
    - Large numbers of individuals affected
    - Political pressure
    - Unmet needs—no current treatment

1.3.2 Clear specification of the problem

- Problem statements need to consider:
    - Patient population affected (indication; epidemiology)
    - Intervention being considered (drug, device, new/old)
    - Comparators
    - Outcome(s) or interest
    - Setting (e.g. hospital, community)
- Well formulated question

1.3.3 Technology assessment and review

1.3.3.1 Evidence and systematic literature review

- A comprehensive search of the literature based on systematic methods is essential
- 2 main types of resources relevant to HTA:
    - Published literature
    - Grey literature

1.3.3.2 Identification, aggregation & appraisal of evidence

- Objective, systematic process for screening and determine studies to be included in the synthesis
- Classify the studies

  - Randomised, non-randomised and economic
- Critical appraisal of the quality of the evidence

1.3.3.3 Synthesize & consolidate

- Findings from multiple studies often combined to respond to the HTA question
- Techniques commonly used to synthesize data in HTA are:
  - Meta-analysis, meta-regression
  - Network meta-analysis

1.3.3.4 Economic evaluation

- Measures the incremental costs and benefits of the technology under review compared to one or more relevant technologies
- CEA, CUA and CBA
- Budget impact

1.3.3.5 Assessment of social, legal & ethical considerations

- Example: genetic information (why?)
- Any access or equity issues following the dissemination and implementation of technologies?

1.3.3.6 Formulation of findings

- Explicitly link quality of the evidence to the strength of findings and recommendations as well as any limitations
- Recommendations based on the findings that reflect the original question(s)

1.3.4 Dissemination of recommendations

- Findings translated into relevant and understandable information
- Knowledge translation

1.3.5 Monitoring the impact of reports

- Some recommendations are translated into policies with clear and quantifiable impacts (e.g. adoption of new technology, change in reimbursement)
- Others go ignored and are not readily adopted into general practice

1.4 Excercises

Read the following HTA published by NICE in the UK. Do the following:

- What is the population?
- What is the intervention and comparators?
- Is there a reproducible search strategy for the clinical evidence in the HTA?
- Was the clinical evidence critically appraised? How?
- Describe the evidence synthesis process
- What type of economic evaluation they used?
- What type of model was used in the economic evaluation?
- How was the uncertainty handled in the economic evaluation?
- Is there a budget impact in the HTA?
- What is the recommendation?

Chapter 2

Introduction to decision-analytic models

2.1 Pre-session readings

Chapter 3 of *R for data science*.

Economic evaluation

Gray, A. M., Clarke, P. M., Wolstenholme, J. L., & Wordsworth, S. (2011). Applied methods of cost-effectiveness analysis in healthcare (Vol. 3). Oxford University Press. Chapter 2.

Birch, S., & Gafni, A. (1992). Cost effectiveness/utility analyses: do current decision rules lead us to where we want to be?. Journal of health economics, 11(3), 279-296. Doubilet, P., Weinstein, M. C., & McNeil, B. J. (1986). Use and misuse of the term "cost effective" in medicine.

Decision modelling

Briggs, A., Sculpher, M., & Claxton, K. (2006). Decision modelling for health economic evaluation. Oxford University Press. Chapter 2. Sections 2.1 and 2.2

Gray, A. M., Clarke, P. M., Wolstenholme, J. L., & Wordsworth, S. (2011). Applied methods of cost-effectiveness analysis in healthcare (Vol. 3). Oxford University Press. Chapter 8. Sections 8.1 to 8.4

Buxton, M. J., Drummond, M. F., Van Hout, B. A., Prince, R. L., Sheldon, T. A., Szucs, T., & Vray, M. (1997). Modelling in ecomomic evaluation: an unavoidable fact of life. Health economics, 6(3), 217-227.

2.2 Economic evaluation

Cost-effectiveness analysis (CEA) and cost-utility analysis (CUA) have been proposed as methods for comparing alternative uses of scarce health-care resources. The difference between CEA and CUA lies in the way outputs are measured.

We can find different objectives of CEA across the literature:

"The underlying premise of cost-effectiveness analysis in health problems is that for any given level of resources available, society (or the decision making jurisdiction involved) wishes to maximize the total aggregate health benefits conferred" Weinstein and Stason (1977).

"For any given rate of output [the combination of inputs] . that costs the decision maker least" Culyer (1980)

"A method of determining the most efficient way of dealing with a specified health problem" Green and Barker (1988)

The goal of CEA is to maximize health benefits produced from a given level of resources. Therefore, it is consistent with welfare economics concept of Pareto efficiency (Birch and Gafni, 1992) (Figure 2.1).

In practice, CEA **relaxes the constraint on available resources**. Because the focus of the evaluation is not a fixed resource pool, but a specific programme making demands on resources, the comparison of the programme under evaluation with an existing programme (e.g., services aimed at improving FS) must consider not only the inter-programme differences in outputs (incremental benefits), but also the inter-programme differences in resources used (incremental costs).

Weinstein and Stason (1977) state that: *"the criterion for cost-effectiveness is the ratio of the net increase of health-care costs to the net effectiveness. The lower the value of this ratio, the higher the priority in terms of maximizing benefits derived from a given health expenditure"*. Issue is that applying a ratio does nod lead to the maximization of benefits from a fixed resource pool (movement from A to C in Figure). **The real problem of the application of CEA is the failure to adhere strictly to the notion of opportunity cost in the measurement of the (incremental) cost of a programme.** The existing programme represents a true opportunity cost for the entire resource requirements of the new programme, even though it does not absorb this level of resources.

2.2.1 Cost-effectiveness in practice

As mentioned before, CEA include both costs and effects, which are represented graphically in a plane:

Questions

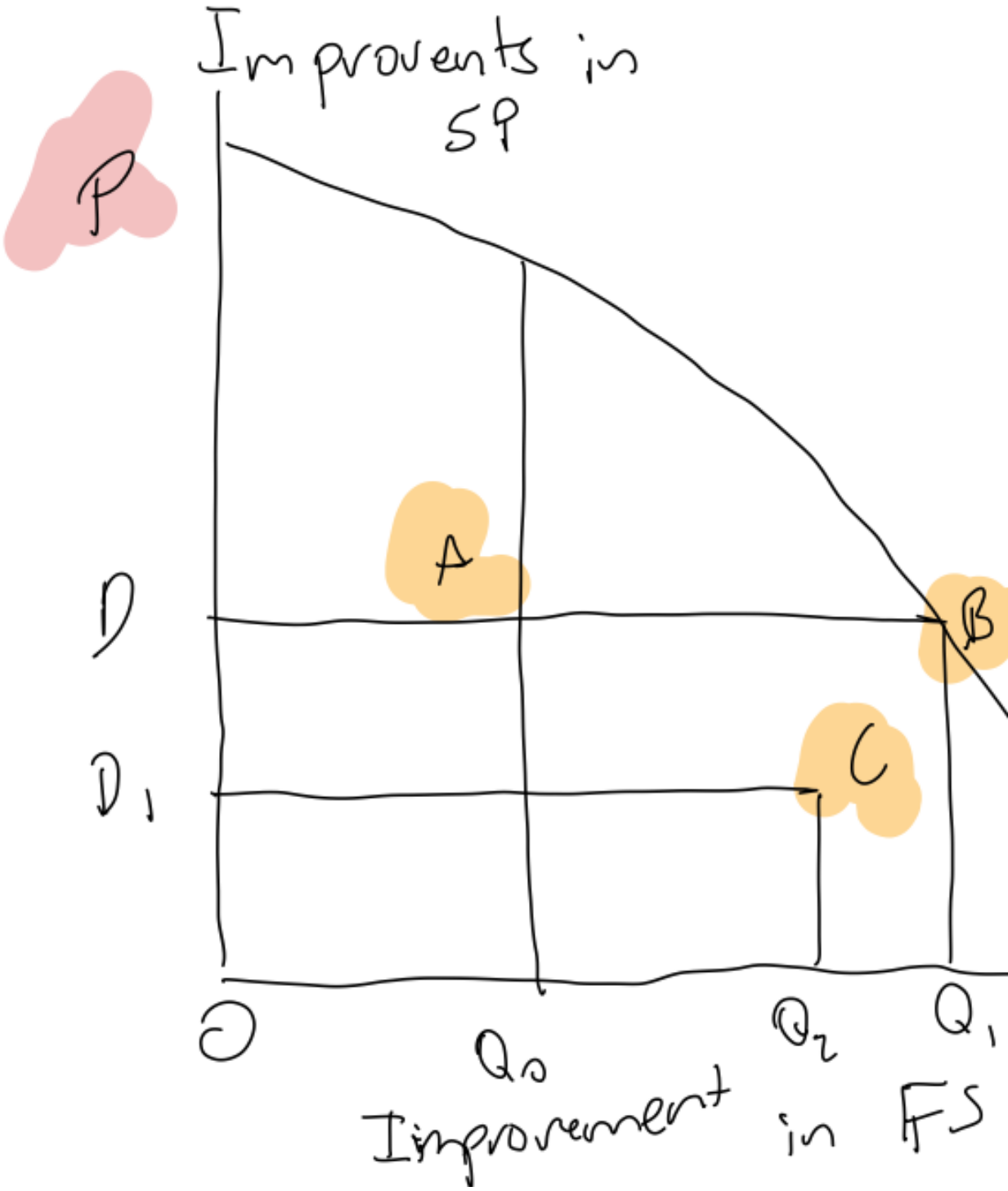

Figure 2.1: Production possibilities frontier

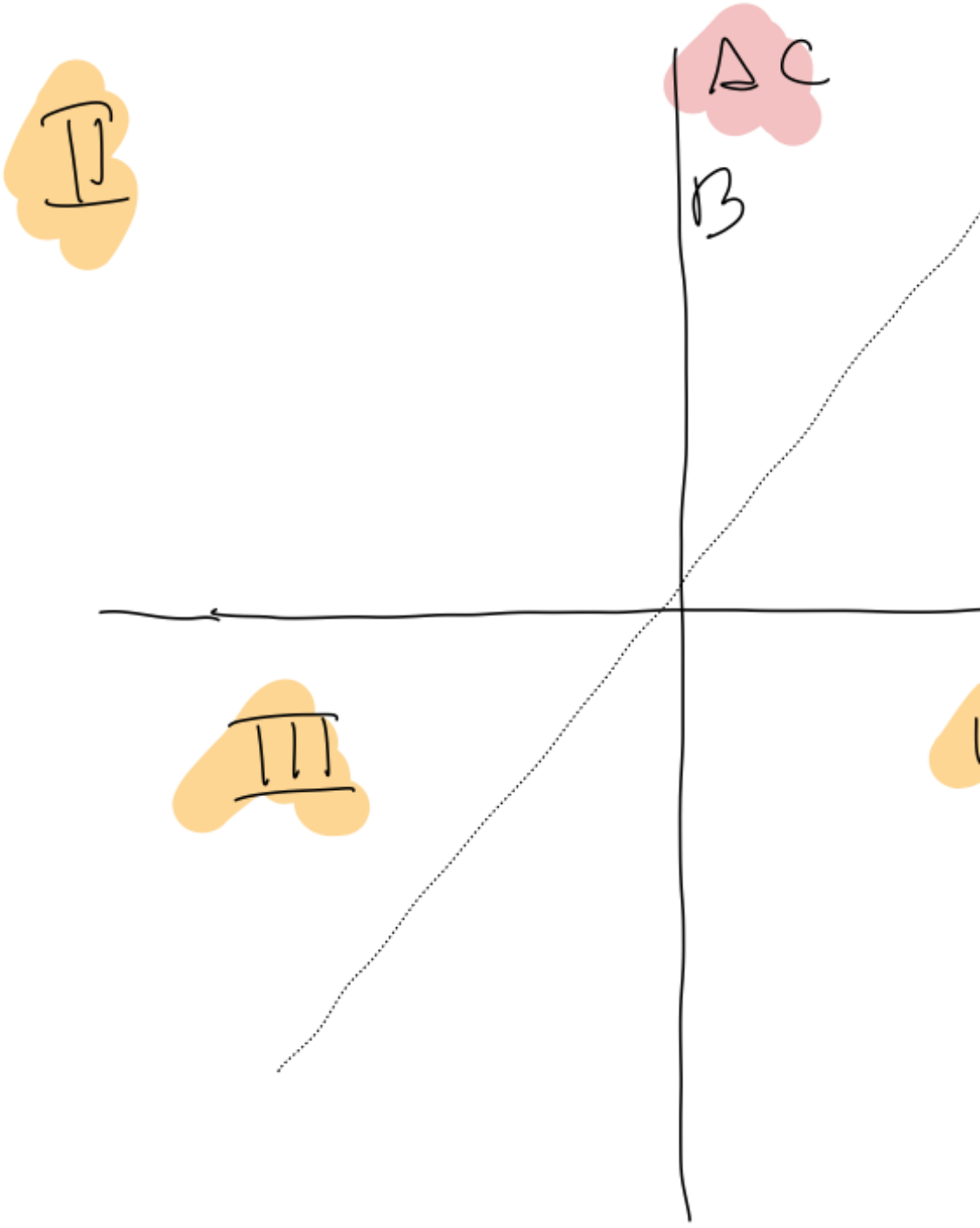

Figure 2.2: Incremental CE plane

Table 2.1: Calculating incremental cost-effectiveness

Option	Cost	QALYs	Incremental cost	Incremental QALY	ICER
A	0				
B	10,000	0.40	10,000	0.40	25,000
C	22,000	0.55	12,000	0.15	80,000
D	25,000	0.50	3,000	-0.05	-60,000
E	40,000	1.00	15,000	0.50	30,000

1. Describe the plane.
2. What is the willingness-to-pay in this plane?
3. What type of uncertainties are encountered in this plane?

Why incremental? We can see it using the example from Gray et al. (2011) applied to mutually exclusive options (see Figure 2.3):

- Diet and excercise (C): Reference
- Metformin (A): \$500k/250 life-years
- New drug (B): \$2500k/300 life-years

Clearly, as this example shows, it is quite misleading to calculate average cost-effectiveness ratios, as they ignore the alternatives available.

Questions

1. What is the difference between average cost-effectiveness ratios vs incremental?

The idea of CEA is to maximize health benefits with the available resources, which in terms of the CE plane represents pushing as far to the right as possible while moving up the vertical axis as little as possible. The next example from Gray et al. (2011) shows the ideas behind **cost-effectivness frontier, dominance, extended dominance** (Table 2.1).

Once we have the ICERs for different independent programmes. How can we maximize health gains with this information? Note that now we are comparing different programmes as opposed to mutually exclusive options. Let's work in the next example:

Finally, we can ask ourselves. What is the maximum value of the incremental cost-effectivess ratio (λ)?

- Rule-based approaches - Adopting arbitrary thresholds. For example, Laupacis et al. (1992) adopted identical cut-off points of CAN\$20,000 per

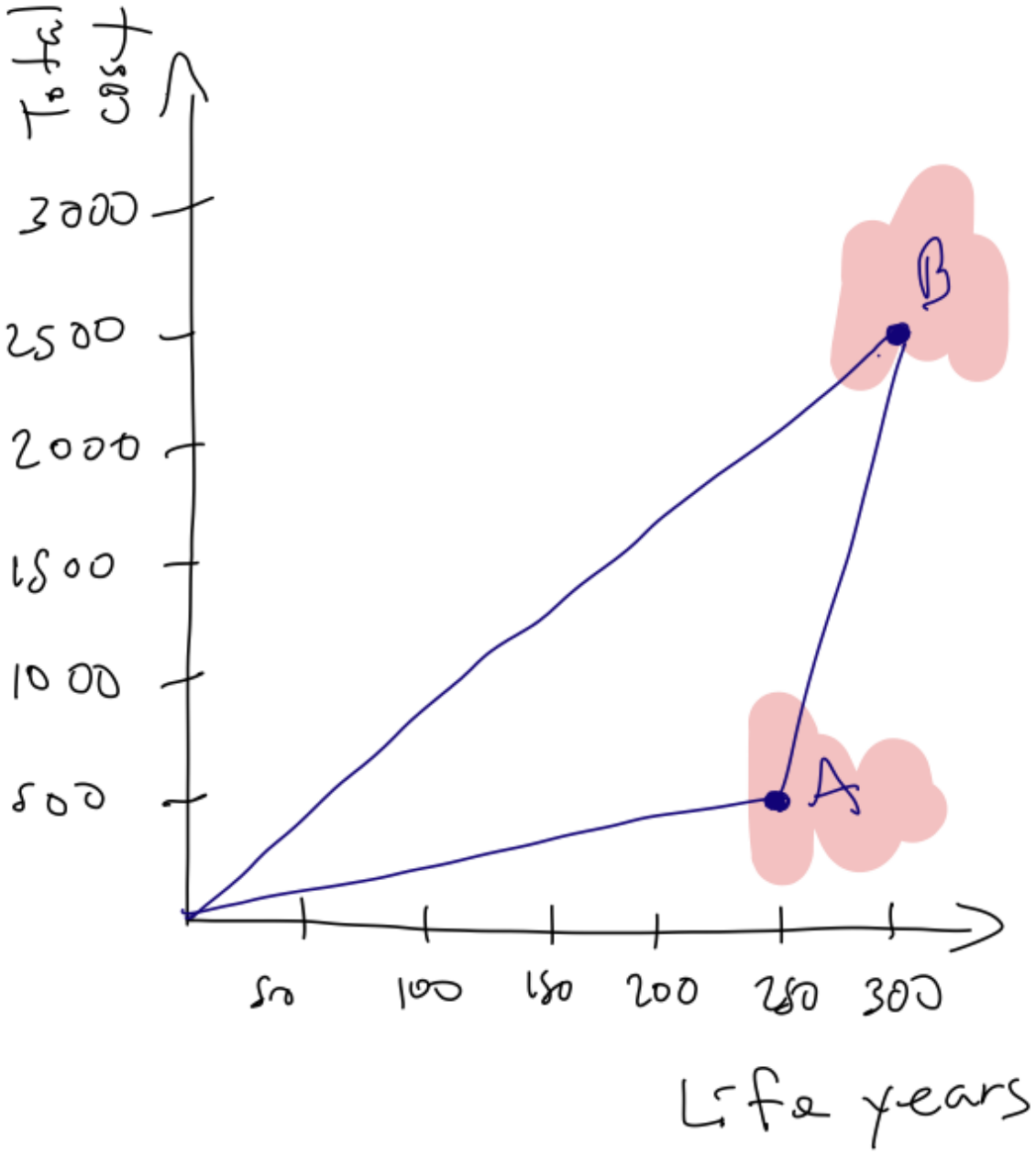

Figure 2.3: Average and incremental cost-effectivenessr

Table 2.2: Using cost-effectiveness to maximize health gain

Intervention	Incremental cost (per k)	Incremental QALYs	ICER
1	1,300	165	7,879
2	600	28	21,429
3	750	110	6,818
4	750	13	57,692
5	2,200	75	29,333
6	400	85	4,706
Total	6,000	476	

QALY gained, up to which level they considered that there would be strong grounds for adoption, and CAN$100,000 per QALY gained, above which they considered that evidence for adoption was weak. An alternative approach is to use the country's gross domestic product (GDP) per capita. Williams et al. (2004) advanced one line of reasoning in support of this, arguing that each person has an entitlement to a 'fair share' of the country's wealth.

- League table approach - If we had full information on the ICERs of all available interventions, it would be possible to rank them by ICER. See the example in Table 2.1.

What are the issues with these approaches?

1. λ is a function of the budget. Therefore it is constantly changing
2. The dynamic nature of λ. As new programs are funded and others replaced, the identification of the last program funded changes.

2.2.2 Conclusions

- Irrespective of whether the problem face by decision-makers is simple (maximizing health gains from available resources) or complex (subject to considerations of equity, accessibility, etc.), if it is not to be considered in the context of a resource constraint there is little use for economics in the way the problem is considered.
- Other approaches have been considered, such as methods in linear programming, which is consistent with the objective of CEA.

$$\max_{x_i,\ldots,x_n} \sum_{i=1}^{n} x_i E_i \text{s.t.} \qquad \sum_{i=1}^{n} x_i c_i \leq C$$

where x_i is the health intervention presented as a binary outcome (1 or 0), E_i is the present value of the health benefits (measured using QALYs) generated by programme over the planning period, and c_i is the present value of the cost of providing programme i over the planning period.

2.3 Decision modelling

HTA is undertaken in order to inform decision-making regarding the appropriate use of particular healthcare programs and interventions, and involves the synthesis of a range of evidence. Broadly speaking, this implies two important components of this process:

1. Gathering evidence for the disease and technology of concern from a range of primary studies. This includes effectiveness, costs, epidemiology, natural history of the disease, quality of life, etc.
2. Synthesizing the evidence found in the first component in order to inform policy and decision-making.

Because of the nature of these two components, decision-analytic modelling has played an important role in HTA. This is because it represents an explicit approach to synthesizing currently available evidence regarding the effectiveness and costs of alternative (mutually exclusive) healthcare strategies (Philips et al., 2006). Therefore one of the main objectives of decision-analytic modelling is to address the relationship between the effectiveness and costs of alternative healthcare strategies in order to assess relative cost-effectiveness (CE) and to determine which options should be adopted given existing information (Philips et al., 2006). Consequently, modelling in the context of HTA is a typical problem of decision-making under uncertainty.

Traditionally randomised controlled trials (RCTs) have been a key component of many HTA process. This is because randomsation protects against selection bias and confounding. Nevertheless, the information produced by RCTs can be limited with respect to evaluating health care as delivered in the real world. This can be because:

1. Not all interventions to be compared are included in the trial.
2. Not enough follow-up.
3. Not enough flexibility (controlled trial).
4. Small sample sizes.
5. Patients in trial are not representative to the target population.

Consequently, it is advised that HTA submissions incorporate information from as many sources as possible to address some of these problems (Sculpher et al.,

2006). As mentioned before, decision analytical models allow for the synthesis of information across multiple sources and for the comparison of multiple options that might not have been included as part of an RCT.

Briefly, Dahabreh et al. (2016) consider some potential goals of modelling in a health-care context:

- To structure investigators' thinking and to facilitate the communication of data, assumptions, and results.
- To synthesize data from disparate sources.
- To make predictions.
- To support causal explanations.
- To inform decision making.

Moreover, they also distinguish different stages for the development of a decision analytic model

1. Define a question
2. Decide on the type of decision model most appropriate
3. Conceptualize the model (and its mathematical structure).
4. Gather all the evidence required for the model and synthesize it.
5. Implement and run the model.
6. Assess the model.

The development of models, especially those trying to explain complex phenoma and informing difficult decisions, is a demanding task. Choosing between alternative modeling approaches can be difficult because the correct choice would not be obvious at early stages in developing a decision analytical model. In general, modelling is most useful when data have limitations (e.g. non-randomised evidence, sparse evidence, etc.), when the research question is complex and when choices are preference laden (Dahabreh et al., 2016). Another important aspect when choosing between modelling approaches is whether the model in question is likely to show results that the intended audience will consider credible and useful. Multiple iterations are typically needed between the key activities outlined previously because at each activity the need for changes at earlier stages may become evident.

2.3.1 Importance of decision-analytic models in HTA

When doing a cost-effectiveness analysis in the context of HTA, one usually starts conceptualizing the model that will help answer the research question. But what is a model? A model a simplified representation of reality (Roberts et al., 2012), where inputs from different sources inform it and its purpose, in the context of HTA, is to inform medical decisions and health-related resource allocation questions (Roberts et al., 2012).

Methods for the conduct of decision-analytic modelling have continued evolving to address the ever-increasing information needs of decision makers. The complexity and continued advances of the relevant methods have spurred the publication of recommendation statements on "best practices" for modeling in the context of HTA (Roberts et al., 2012; Briggs et al., 2012). Some of these modelling techniques (Caro et al., 2012) include:

- Decision-tree models.
- State-transition models.
- Micro-simulation models.
- Discrete event simulation (DES) models.
- Dynamic transmission models.

2.4 Excercises

1. Question 1 and 2 from Chapter 2 in *Applied methods of cost-effectiveness analysis in healthcare* .
2. Read the articles in our google classroom.

Chapter 3

Good practices in decision modelling and decion-tree models

3.1 Pre-session readings

Good practices

Roberts, M., Russell, L. B., Paltiel, A. D., Chambers, M., McEwan, P., & Krahn, M. (2012). Conceptualizing a model: a report of the ISPOR-SMDM modeling good research practices task force–2. Medical Decision Making, 32(5), 678-689.

Decision-tree models

Tarride, J. E., Blackhouse, G., Bischof, M., McCarron, E. C., Lim, M., Ferrusi, I. L., … & Goeree, R. (2009). Approaches for economic evaluations of health care technologies. Journal of the American College of Radiology, 6(5), 307-316.

Briggs, A., Sculpher, M., & Claxton, K. (2006). Decision modelling for health economic evaluation. Oxford University Press. Chapter 2. Sections 2.2 and 2.3.1.

Gray, A. M., Clarke, P. M., Wolstenholme, J. L., & Wordsworth, S. (2011). Applied methods of cost-effectiveness analysis in healthcare (Vol. 3). Oxford University Press. Chapter 8. Sections 8.5 and 8.6.

Introduction to R

The goal of this tutorial is to orient the learner to R Studio and the R programming language.

Please complete the following:

1. Create a folder somewhere you can easily find it (e.g., on your desktop) called 'R Course'.
2. Open R Studio
3. Session -> Set Working Directory -> to source file location

Outline

1. Read in the 'babies' data set
2. Basic data manipulation
3. Functions
4. Saving and viewing results

Babies Dataset

The `babies` dataset will be used throughout this session to illustrate basic R concepts.

The dataset is a collection of variables taken for each new mother in a Child and Health Development Study. It has 1,236 observations on male live births for the following 23 variables.

Variables in data file

```
id            identification number
date          birth date as character string (mon-dd-yyyy)
ddate         day of birth
mdate         month of birth
```

```
ydate          year of birth
gestation      length of gestation in days
wt             birth weight in ounces (999 unknown)
parity         total number of previous pregnancies
               including fetal deaths and still births
age            mother's age in years at end of pregnancy
                 999=unknown
ed             mother's education
                  0=less than 8th grade
                  1=8th-12th grade - did not graduate
                  2=HS graduate???no other schooling
                  3=HS+trade
                  4=HS+some college
                  5=College graduate,
                  6=Trade school HS unclear
                  9=unknown
ht             mother's height in inches to the last completed inch
                  999=unknown
wt1            mother prepregnancy wt in pounds
                  999=unknown
dage           father's age, coding same as mother's age.
ded            father's education, coding same as mother's education.
dht            father's height, coding same as for mother's height
dwt            father's weight coding same as for mother's weight
inc            family yearly income in $2500 increments
                  0=under 2500
                  1=2500-4999
                  ...,
                  8= 12,500-14,999
                  9=15000+
                  998=unknown
                  999=not asked
smoke          does mother smoke?
                  0=never
                  1=smokes now
                  2=until current pregnancy
                  3=once did, not now
                  9=unknown
time           If mother quit smoking, how long ago?
                  0=never smoked
                  1=still smokes
                  2=during current preg
                  3=within 1 yr
                  4=1 to 2 years ago
                  5=2 to 3 years ago
                  6=3 to 4 years ago
```

```
                  7=5 to 9 years ago
                  8=10+years ago
                  9=quit and don't know,
                  998=unknown
                  999=not asked
number        number of cigs smoked per day for past and current smokers
                  0=never
                  1=1-4
                  2=5-9
                  3=10-14
                  4=15-19
                  5=20-29
                  6=30-39
                  7=40-60
                  8=60+
                  9=smoke but don't know
                  998=unknown
                  999=not asked
race          mother's race
marital       marital status of mother
drace         father's race, coding same as mother's race
```

Loading babies dataset

We can can load the data from our working directory as shown below. It is a .csv file, so can be read in with `read.csv`. specifying that

- the file has a "header" row (`header=T`) with variable names
- values are separated by commas.

- values of '', '998' or '999' represent missing data

We need to be confident that 998 and 999 are not legitimate values for any variables, so that R does not interpret legitimate values as missing. We will not distinguish between 'unknown' and 'not asked' in the data, so the code below will read two consecutive commas (''), 998, and 999 as missing values. We will have to deal separately with the values '9' that mean a value is missing.

```
babies <- read.csv('babies.csv',header=T, sep=",",
                     na.strings=c("", "998", "999"))
```

Data manipulation

What are the dimensions of the data set?

```
dim(babies)
```

[1] 1236 23

What are the names of the variables?

```
names(babies)
```

```
##  [1] "id"        "date"      "gestation" "wt"        "parity"    "age"
##  [7] "ed"        "ht"        "wt1"       "dage"      "ded"       "dht"
## [13] "dwt"       "inc"       "smoke"     "time"      "number"    "race"
## [19] "marital"   "drace"     "ddate"     "mdate"     "ydate"
```

```
#head(babies)
#tail(babies)
```

Isolating variables

```
# babies$id
# babies[, 1]
# babies[, 'id']
# babies[1:10, c('id', 'date', 'gestation')]
```

Assignment and subsetting

```
under.30<- babies$age<30
b<- babies[under.30, ]
summary(b$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   15.00   22.00   24.00   24.03   27.00   29.00       2
```

```
b<- babies[which(under.30), ]
summary(b$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   22.00   24.00   24.03   27.00   29.00
```

Exercises (1)

1. Create a new data set consisting of only the rows where the father's age is 40 or over (use '>=')

2. Include only the mother's age, father's age and marital status in the new dataset
3. Display the dimensions of this data set in the console.
4. Locate the new dataset in th Global Environment and view it

NB: Be sure that you have handled missing father's ages properly

Exercises (1) solutions

Creating some new variables

A baby weight more than 4kg is defined as macrosomia. We can create this variable (converting ounces to kg) as:

```
babies$wtKg <- babies$wt/(16*2.2)
babies$macrosomia <- babies$wtKg > 4
```

A baby weighing less than 2.5kg is defined as "small". We can create this variable as:

```
babies$smallBaby <- babies$wtKg < 2.5
```

We can make a single character variable with three categories of weight:

```
babies$birthWeightCat <- ifelse(babies$wtKg < 2.5, "small",
                                ifelse(babies$wtKg > 4, "large","normal"))
```

Notice that when we tabulate these, they appear in alphabetical order.

```
table(babies$birthWeightCat)
```

```
##
##  large normal  small
##    145   1033     58
```

To impose an order, we can convert to a factor and specify the order:

```
babies$birthWeightCat <- factor(babies$birthWeightCat,
                                levels=c("small","normal","large"),
                                labels=c("Small baby","Normal weight baby","Large baby"))
table(babies$birthWeightCat)
```

```
##
##         Small baby Normal weight baby         Large baby
##                 58               1033                145
```

A baby born before 37 completed weeks of pregnancy is defined as preterm. We can create this logical variable as:

```
babies$preterm <- babies$gestation < 37*7
table(babies$preterm)
```

```
##
## FALSE  TRUE
##  1126    97
```

This is just a logical variable (`TRUE/FALSE` or `T/F`) so we need to make a factor to have it appear more user-friendly:

```
babies$whenBorn <- factor(babies$preterm,
                          levels=c(TRUE,FALSE),
                          labels=c("Preterm","Full-term"))
```

We can make a binary variable for the mother's and father's races:

```
babies$whiteRace <- ifelse(babies$race=="white","white","other")
babies$dwhiteRace <- ifelse(babies$drace=="white","white","other")
```

Exercises (2)

1. Starting with the numeric variable `smoke`, in the babies dataset, create a new variable `smokeCat` in the data set that has three levels: `never smoker`, `past smoker`,`current smoker`. Use the `factor` function and ensure that the levels appear in that order.
2. Tabulate the numerical variable against the new one to make sure you have not made an error.
3. Make a binary variable in the babies dataset `smokeNow` that is 'Yes' when the mother smokes and 'No' otherwise (i.e., it is not yes (1) or it is missing (9).)
4. [optional] Make a factor variable `eduCat` from the mother's education variable `ed` using the information at the top if this file about the meaning of 1,2,3,...

Exercises (2) solutions

1. Make the two variables:
2. Check the coding was correct
3. Make a binary variable
4. [optional] make the education variable and check it

Functions

Two types:

1. Built in functions
2. User defined functions

Built in Functions

When you open R, there are many functions available to you: Here, we will review a few useful built in functions.

If you need help using a function, execute ? followed by the function name, with or without the parentheses.

```
? table
```

The following functions are widely used in descriptive statistics

table(), prop.table()

```
t<- with(babies, table(marital))
t
```

```
## marital
##          divorced legally separated           married     never married
##                 5                15              1208                 6
##           unknown
##                 2
```

```
p<- prop.table(t)
p
```

```
## marital
##          divorced legally separated           married     never married
##       0.004045307       0.012135922       0.977346278       0.004854369
##           unknown
##       0.001618123
```

```
p*100
```

```
## marital
##          divorced legally separated           married     never married
##         0.4045307         1.2135922        97.7346278         0.4854369
##           unknown
##         0.1618123
```

ifelse()

```
babies$first.preg<- with(babies, ifelse(parity==0, 'first','not first'))
table(babies$first.preg)
```

```
##
##     first not first
##       315       921
```

summary(), mean(), median(), sd(), quantile()

```
summary(babies)
```

```
##        id           date             gestation           wt
##  Min.   :  15   Length:1236        Min.   :148.0   Min.   : 55.0
##  1st Qu.:5286   Class :character   1st Qu.:272.0   1st Qu.:108.8
##  Median :6730   Mode  :character   Median :280.0   Median :120.0
##  Mean   :6001                      Mean   :279.3   Mean   :119.6
##  3rd Qu.:7583                      3rd Qu.:288.0   3rd Qu.:131.0
##  Max.   :9263                      Max.   :353.0   Max.   :176.0
##                                    NA's   :13
##      parity            age              ed              ht
```

```
##  Min.   : 0.000   Min.   :15.00   Min.   :0.000   Min.   :53.00
##  1st Qu.: 0.000   1st Qu.:23.00   1st Qu.:2.000   1st Qu.:62.00
##  Median : 1.000   Median :26.00   Median :2.000   Median :64.00
##  Mean   : 1.932   Mean   :27.26   Mean   :2.916   Mean   :64.05
##  3rd Qu.: 3.000   3rd Qu.:31.00   3rd Qu.:4.000   3rd Qu.:66.00
##  Max.   :13.000   Max.   :45.00   Max.   :9.000   Max.   :72.00
##                   NA's   :2                       NA's   :22
##       wt1             dage            ded             dht             dwt
##  Min.   : 87.0   Min.   :18.00   Min.   :0.000   Min.   :60.0   Min.   :110.0
##  1st Qu.:114.8   1st Qu.:25.00   1st Qu.:2.000   1st Qu.:68.0   1st Qu.:155.0
##  Median :125.0   Median :29.00   Median :4.000   Median :71.0   Median :170.0
##  Mean   :128.6   Mean   :30.35   Mean   :3.189   Mean   :70.2   Mean   :171.2
##  3rd Qu.:139.0   3rd Qu.:34.00   3rd Qu.:5.000   3rd Qu.:72.0   3rd Qu.:185.0
##  Max.   :250.0   Max.   :62.00   Max.   :9.000   Max.   :78.0   Max.   :260.0
##  NA's   :36      NA's   :7                       NA's   :492    NA's   :499
##       inc             smoke             time            number
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :3.000   Median :1.0000   Median :1.0000   Median :1.000
##  Mean   :3.701   Mean   :0.8681   Mean   :0.9625   Mean   :1.825
##  3rd Qu.:5.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :9.000   Max.   :9.0000   Max.   :9.0000   Max.   :9.000
##  NA's   :124                      NA's   :10       NA's   :10
##      race             marital             drace               ddate
##  Length:1236        Length:1236        Length:1236        Min.   : 1.00
##  Class :character   Class :character   Class :character   1st Qu.: 8.00
##  Mode  :character   Mode  :character   Mode  :character   Median :15.00
##                                                           Mean   :15.37
##                                                           3rd Qu.:23.00
##                                                           Max.   :31.00
##
##      mdate            ydate           wtKg       macrosomia
##  Min.   : 1.000   Min.   :1961   Min.   :1.562   Mode :logical
##  1st Qu.: 4.000   1st Qu.:1961   1st Qu.:3.089   FALSE:1091
##  Median : 7.000   Median :1962   Median :3.409   TRUE :145
##  Mean   : 6.617   Mean   :1962   Mean   :3.397
##  3rd Qu.: 9.000   3rd Qu.:1962   3rd Qu.:3.722
##  Max.   :12.000   Max.   :1962   Max.   :5.000
##
##  smallBaby                  birthWeightCat  preterm             whenBorn
##  Mode :logical   Small baby        :  58   Mode :logical   Preterm  :  97
##  FALSE:1178      Normal weight baby:1033   FALSE:1126      Full-term:1126
##  TRUE :58        Large baby        : 145   TRUE :97        NA's     :  13
##                                            NA's :13
##
##
```

```
##
##   whiteRace          dwhiteRace                  smokeCat     smokeNow
##  Length:1236        Length:1236        Never smoker  :544   Length:1236
##  Class :character   Class :character   Past smoker   :198   Class :character
##  Mode  :character   Mode  :character   Current smoker:484   Mode  :character
##                                        NA's          : 10
##
##
##
##               eduCat     first.preg
##  HS graduate     :444   Length:1236
##  HS+some college :298   Class :character
##  College graduate:219   Mode  :character
##  8th-12th grade  :183
##  HS+trade        : 65
##  (Other)         : 26
##  NA's            :  1
```

```
gestAge<- babies$gestation
mean(gestAge)
```

```
## [1] NA
```

```
mean(gestAge, na.rm=T)
```

```
## [1] 279.3385
```

```
gestAge<- gestAge[complete.cases(gestAge)]
summary(gestAge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   148.0   272.0   280.0   279.3   288.0   353.0
```

```
mean(gestAge)
```

```
## [1] 279.3385
```

```
median(gestAge)
```

```
## [1] 280
```

```
sd(gestAge)
```

```
## [1] 16.02769
```

```
range(gestAge)
```

```
## [1] 148 353
```

```
min(gestAge)
```

```
## [1] 148
```

```
max(gestAge)
```

```
## [1] 353
```

```
quantile(gestAge)
```

```
##   0%  25%  50%  75% 100%
##  148  272  280  288  353
```

```
quantile(gestAge, seq(0, 1, by = 0.2))
```

```
##   0%  20%  40%  60%  80% 100%
##  148  270  277  283  290  353
```

round()

```
x<- 1121.933384
round(x, 3)
```

```
## [1] 1121.933
```

```
round(x, -2)
```

```
## [1] 1100
```

paste()

We can build up complex quoted strings with `paste`:

```
paste("Mean (days) =", mean(gestAge))
```

```
## [1] "Mean (days) = 279.338511856092"
```

```
paste("Mean (days) =", round(mean(gestAge)))
```

```
## [1] "Mean (days) = 279"
```

```
paste('mean day (sd) = ',
      round(mean(gestAge)),
      ' (',
      round(sd(gestAge), 1),
      ')',
      sep='')
```

```
## [1] "mean day (sd) = 279 (16)"
```

apply()

Carries out an operation on the rows or columns of a dataset. We pick the `MARGIN` to specify whether we want this to be rows or columns

- 1 = rows
- 2 = columns

```
apply(babies[, c('age', 'dage', 'wtKg')], MARGIN = 2, FUN = mean, na.rm=T)
```

```
##      age     dage     wtKg
## 27.25527 30.34825  3.39707
```

tapply()

Carries out an operation on one variable, split by a second variable (or group of variables) - e.g., means by group:

```
wt.by.smoke.term<- with(babies,
                        tapply(wt, list(smokeCat, whenBorn), mean, na.rm=T))
rw<- round(wt.by.smoke.term, 1)
rw
```

```
##                Preterm Full-term
## Never smoker     107.0     124.1
## Past smoker      103.4     125.9
## Current smoker    91.7     116.2
```

User Defined Functions

You can write your own functions, text(arguments){ body }

```
custom.summary<- function(x){
  out<- paste(round(mean(x, na.rm=T), 2),
              ' (SD=',
              round(sd(x, na.rm=T), 2),
              ')',
              sep='')
  return(out)
}
custom.summary(babies$age)
```

```
## [1] "27.26 (SD=5.78)"
```

Saving output and opening in word

```
write.csv(rw, 'test.csv', quote= F)
```

Now open 'test.csv' with word. Highlight and click Table -> convert text to table -> ok.

Exercises (3)

1. Create a function to summarize a binary (0/1 or FALSE/TRUE)) variable and return a character string that looks like, for example "18/54 (33.3%)"

Framework

```
bin.sum <- function(x){
t <- # tabulate x
n <- # how many nonmissing observations?
x <- # how many 1's?
pct <- round()
paste()

}
```

2. Use `tapply` to apply your function to the macrosomia using smokeCat as the grouping variable

Hint

```
result <- tapply(X =
                 INDEX=
                 FUN=bin.sum)
```

3. View your results as a Word table using by using write.csv to save to a CSV file.
4. [Advanced] Create a function to summarize a continuous variable in the following format: "Median (IQR.low, IQR.high), n". Use `apply` to apply this function to the variables `age`, `dage`, 'ht', 'dht'. The result should be a table that you can save in the same way as in question 3.
5. [More advanced] Use the function from question 4. to summarize a the variables `age`, `dage`, 'ht', 'dht' in groups formed by 'preterm'. Hint: Look at the help for the function `aggregate`.

Hint:

```
aggregate(x=DATASET,
          by=list(BY VARIABLES HERE),
          FUN=FUNCTION.FROM.4)
```

Exercises (3) solutions

1.
2.
3.
4.
5.

Working with more than one dataset

Suppose you want to know how far from the average for the ethnic group each baby's birth weight is as a Z-score:

```
Z = (wt - average)/SD
```

You need to

1. find the average and SD for each group
2. merge this data with the full babies dataset
3. calculate each baby's Z score

The R code for these steps is shown below

1. find the average and SD for each group and save it in a data.frame

```
average.by.group <- tapply(X = babies$wtKg,
                           INDEX=babies$race,
                           mean,na.rm=T)

sd.by.group <- tapply(X = babies$wtKg,
                      INDEX=babies$race,
                      sd,na.rm=T)
stats <- data.frame(race=names(average.by.group),
                    average=average.by.group,
                    SD=sd.by.group)
print(stats)
```

```
##             race  average        SD
## asian     asian 3.137268 0.4543554
## black     black 3.216980 0.5422872
## mexican mexican 3.526989 0.4017935
## mixed     mixed 3.360719 0.5363923
## unknown unknown 3.892045        NA
## white     white 3.455721 0.5027514
```

2. merge this data with the full babies dataset

```
babies <- merge(babies,
                stats,
                by="race")

head(babies[ , c("wtKg","average", "SD","race")])
```

```
##       wtKg  average        SD  race
## 1 3.409091 3.137268 0.4543554 asian
## 2 3.125000 3.137268 0.4543554 asian
## 3 2.642045 3.137268 0.4543554 asian
## 4 3.352273 3.137268 0.4543554 asian
## 5 2.982955 3.137268 0.4543554 asian
## 6 2.840909 3.137268 0.4543554 asian
```

3. calculate each baby's Z score

```
babies$Z.wt <- (babies$wtKg - babies$average)/babies$SD
```

Finally, please save the alterations to the babies data set for the next session:

```
write.csv(babies, 'babiesAugmented.csv', quote=F, row.names = F)
```

Bibliography

Birch, S. and Gafni, A. (1992). Cost effectiveness/utility analyses: do current decision rules lead us to where we want to be? *Journal of health economics*, 11(3):279–296.

Briggs, A. H., Weinstein, M. C., Fenwick, E. A., Karnon, J., Sculpher, M. J., and Paltiel, A. D. (2012). Model parameter estimation and uncertainty analysis: a report of the ispor-smdm modeling good research practices task force working group–6. *Medical decision making*, 32(5):722–732.

Caro, J. J., Briggs, A. H., Siebert, U., and Kuntz, K. M. (2012). Modeling good research practices—overview: a report of the ispor-smdm modeling good research practices task force–1. *Medical Decision Making*, 32(5):667–677.

Culyer, A. J. (1980). *The political economy of social policy*. St. Martin's Press.

Dahabreh, I. J., Trikalinos, T. A., Balk, E. M., and Wong, J. B. (2016). Guidance for the conduct and reporting of modeling and simulation studies in the context of health technology assessment. *Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet]*.

Goeree, R. (2015). *Health technology assessment: using biostatistics to break the barriers of adopting new medicines*. CRC Press.

Gray, A. M., Clarke, P. M., Wolstenholme, J. L., and Wordsworth, S. (2011). *Applied methods of cost-effectiveness analysis in healthcare*, volume 3. Oxford University Press.

Green, A. and Barker, C. (1988). Priority setting and economic appraisal: whose priorities—the community or the economist? *Social science & medicine*, 26(9):919–929.

Laupacis, A., Feeny, D., Detsky, A. S., and Tugwell, P. X. (1992). How attractive does a new technology have to be to warrant adoption and utilization? tentative guidelines for using clinical and economic evaluations. *CMAJ: Canadian Medical Association Journal*, 146(4):473.

Philips, Z., Bojke, L., Sculpher, M., Claxton, K., and Golder, S. (2006). Good practice guidelines for decision-analytic modelling in health technology assessment. *Pharmacoeconomics*, 24(4):355–371.

Roberts, M., Russell, L. B., Paltiel, A. D., Chambers, M., McEwan, P., and Krahn, M. (2012). Conceptualizing a model: a report of the ispor-smdm modeling good research practices task force–2. *Medical Decision Making*, 32(5):678–689.

Sculpher, M. J., Claxton, K., Drummond, M., and McCabe, C. (2006). Whither trial-based economic evaluation for health care decision making? *Health economics*, 15(7):677–687.

Weinstein, M. C. and Stason, W. B. (1977). Foundations of cost-effectiveness analysis for health and medical practices. *New England journal of medicine*, 296(13):716–721.

Williams, A. et al. (2004). What could be nicer than nice? *Monographs*.