Introduction to Statistical Analysis: a regression-from-the-outset approach

Sahir, Shirin and Jim

2020-03-17

Contents

Preface

0.1 Target

The target is graduate students in population health sciences in their first year. Concurrently, they take their first courses on epidemiologic methods. The department is known for its emphasis on quantitative methods, and students' ability to carry out their own quantitative work. Since most of the data they will deal with are non-experimental, there is a strong emphasis on multivariable regression. While some students will have had some statistical courses as undergraduates, the courses start at the beginning, and are pitched at the Master's level.

In the last decade, the incoming classes have become more diverse, both in their backgrounds, and in their career plans. Some of those in the recently begun MScPH program plan to me consumers rather than producers of research; previously, the majority of students pursued a thesis-based Masters that involved considerable statistical analyses to produce new statistical evidence.

0.2 Topics/textbooks

For the **first term** course 607, recent choices have been *The Practice of Statistics in the Life Sciences* by Baldi and Moore, and *Stats* by de Veaux, Velleman and Bock. Others that have been recommended are the older texts by Pagano and Gauvreau, and by Rosner. Some of us have also drawn on material in *Statistics* by Freedman, Pisani, Purves and Adkikari, and *Statistical Methods in Medical Research*, 4th Edition_ by Armitage,Berry, and Matthews.

The newer books have tried to teach the topic more engagingly, by starting with where data come from, and (descriptively) displaying single distributions, or relationships between variables. They and the many others then typically move on to Probability; Random Variables; Sampling Distributions; Confidence intervals and Tests of Hypotheses; Inference about/for a single Mean/Proportion/Rate and a difference of two Means/Proportions/Rates; Chi-square Tests for 2 way frequency tables; Simple Correlation and Regression. Most include a (or point

to an online) chapter on Non-Parametric Tests. They typically end with tables of probability tail areas, and critical values.

Bradford Hill's *Principles of Medical Statistics* followed the same sequence 80 years ago, but in black type in a book that measured 6 inches by 9 inches by 1 inch, and weighed less than a pound. Today's multi-colour texts are 50% longer, 50% wider, and twice as thick, and weigh 5 pounds or more.

The topics to be covered in the **second term** course include multiple regression involving Gaussian, Binomial, and Poisson variation, as well as (possibly censored) time-durations – or their reciprocals, event rates. Here is is more difficult to point to one modern comprehensive textbook. There is pressure to add even more topics, such as correlated data, missing data, measurement error etc. top the second statistics course.

0.3 Regression from the outset

It is important to balance the desire to cover more of these regression-based topics with having a good grounding, from the first term, in the basic concepts that underlie all statistical analyses.

The first term *epidemiology* course deals with proportions and rates (risks and hazards) and – at the core of epidemiology – comparisons involving these. Control for confounding is typically via odds/risk/rate differences/ratios obtained by standardization or Mantel-Haenszel-type summary measures. Teachers are reluctant to spend the time to teach the classical confidence intervals for these, as they are not that intuitive and – once students have covered multiple regression – superceded by model-based intervals.

One way to synchronize with epidemiology, is to teach the six separate topics Mean/Proportion/Rate and differences of two Means/Proportions/Rates in a more unified way by embedding all 6 in a regression format right from the outset, to use generalized linear models, and to focus on all-or-none contrasts, represented by binary 'X' values.

This would have other benefits. As of now, a lot of time in 607 is spent on 1-sample and 2-sample methods (and chi-square tests) that don't lead anywhere (generalize). Ironically, the first-term concerns with equal and unequal variance tests are no longer raised, or obsessed about, in the multiple regression framework in second term.

The teaching/learning of statistical concepts/techiques is greatly enriched by real-world applications from published reports of public health and epidemiology research. In 1980, a first course in statistics provided access to 80% of the articles in NEJM articles. This large dividend is no longer the case – and even less so for journals that report on non-experimental research. The 1-sample and 2-sample methods, and chi-square tests that have been the core of first statistics courses are no longer the techniques that underlie the reported summaries in

the abstracts and in the full text. The statistical analysis sections of many of these articles do still start off with descriptive statistics and a perfunctory list of parametric and non-parametric 1 and 2 sample tests, but most then describe the multivariable techniques used to produce the reported summaries. [Laboratory sciences can still get by with t-tests and 'anova's – and the occasional ancova'; studies involving intact human beings in free-living populations can not.] Thus, if the first statistical course is to to get the same 'understanding' dividend from research articles as the introductory epidemiology course does, that first statistical course needs to teach the techniques that produce the results in the abstracts. Even if it can only go so far, such an approach can promote a regression approach right from week one, and build on it each week, rather than introduce it for the first time at week 9 or 10, when the course is already beginning to wind down, and assignments from other courses are piling up.

0.4 Parameters first, data later

When many teachers and students think of regression, they imagine a cloud of points in x-y space, and the least squares fitting of a regression line. They start with thinking about the data.

A few teachers, when they introduce regression, do so by describing/teaching it as **an equation that connects parameters**, constructed in such a way that the parameter-constrast of interest is easily and directly visible. Three such teachers are Clayton and Hills 1995, Miettinen1985, and Rothman 2012. In each case, their first chapter on regression is limited to the parameters and to undersatnding what they mean; data only appear in the next chapter.

There is a lot to commend this approach. It reminds epidemiologists – and even statisticians – that statistical inference is about parameters. Before addressing data and data-summaries, we need to specify what the estimands are – i.e, what parameter(s) is(are) we pursuing.

It is easy and tempting to start with data, since the form of the summary statistic is usually easy to write down directly. It can also be used to motivate a definition: for example, we could define an odds ratio by its empirical computational form ad/bc. However, this 'give me the answer first, and the question later' approach comes up short as soon as one asks how statistically stable this estimate is. To derive a standard error or confidence interval, one has to appeal to a sampling distribution. To do this, one needs to identify the random variables involvced, and the parameters that determine/modulate their statistical distributions.

Once students master the big picture (the parameter(s) being pursued), the task of estimating them by fitting these equations to data is considerably simplified, and becomes more generic. In this approach more upfront thinking is devoted to the parameters – to what Miettinen calls the design of the study object –

with the focus on a pre-specified 'deliverable.'

0.5 Let's switch to "y-bar", and drop "x-bar".

The prevailing practice, when introducing descriptive statistics, and even to 1 and two sample procedures, is to use the term x-bar ($\bar{x}$) for an arithmetic mean (one notable execption is de Veaux at al.) This misses the chance to prepare students for regression, where E[Y|X] is the object of interest, and the X-conditional Y's are the random variables. Technically speaking, the X's are not even considered random variables. Elevating the status of the Y's and explaining the role of the X's, and the impact of the X distributions on precision might also cut down on the practice of checking the normality of the X's, even though the X's are not random variables. They are merely the X locations/profiles at which Y was measured/recorded. When possible, the X distribution should be determined by the investigators, so as to give more precise and less correlated estimates of the parameters being pursued. Switching from $\bar{x}$ to $\bar{y}$ is a simple yet meaningful step in this direction. JH made this switch about 10 years ago.

0.6 Computing from the outset

In 1980, most calculations in the first part of 607 were by hand calculator. Computing summary statistics by hand was seen as a way to help students understand the concepts involved, and the absence of automated rapid computation was not considered a drawback. However, doing so did not always help students understand the concept of a standard deviation or a regression slope, since these formulae were designed to minimize the number of keystrokes, rather than to illuminate the construct involved. For example, it was common to rescale and relocate data to cut down on the numbers of digits entered, to group values into bins, and use midpoints and frequencies. It was also common to use the computationally-economical 1-pass-through-the-data formula for the sample variance

$$s^2 = \frac{\sum y^2 - \frac{(\sum y)^2}{n}}{n-1},$$

even though the definitional formula is

$$s^2 = \frac{\sum (y-\bar{y})^2}{n-1}.$$

The latter (definitional) one was considered too long, even though having to first have to compute $\bar{y}$ and then go back and compute (and square) each $y - \bar{y}$ would have helped students to internalize what a sample variance is.

When spreadsheets arrived in the early 1980s, students could use the built-in mean formula to compute and display $\bar{y}$, another formula to compute and

display a new column of the deviations from $\bar{y}$, another to compute and display a new column of the squares of these deviations, another to count the number of deviations, and a final formula to arrives at s^2. The understanding comes from coding the definitional formula, and the spreadsheet simply and speedily carries them out, allowing to user to see all of the relevant components, and from noticing if each one looks reasonable. Ultimately, once students master the concept, they could move on to built-in formulae that hide (or themselves avoid) the intermediate quantities.

Few teachers actually encouraged the use of spreadsheets, and instead promoted commercial statistical packages such as SAS, SPSS and Stata. Thus, the opportunity to learn to 'walk first, run later' afforded by spreadsheets was not fully exploited.

RStudio is an integrated environment for R, a free software environment for statistical computing and graphics that runs on a wide variety of platforms. Just like spreadsheet software, one can use R not just as a calculator, but as a *programmable* calculator, and by programming them, learn the concepts before moving on to the built-in functions. There is a large user-community and a tradition of sharing information and ways of doing things. The graphics language contains primitive functions that allow customization, as well as higher-level functions, and is tighly integrated with the statistical routines amd data frame functions. R Markdown helps to foster reproducible research. Shiny apps allow interactivity and visualization, a bit like 'what-ifs' with a spreadsheet.

It takes practice to become comfortable with R. Gor those less mathematical, it is somewhat more cryptic than, and not quite as intuitive as, other packages. For the last several years, the department has offered a 13 hour course introduction to R in first term. Initially the aim was to prepare students for using it in course 621 in second term, but in the Fall 2018 and 2019 offerings of course 607, computing with R and use of R Studio became mandatory. Just as the epidemiology material in the Fall is shared between 2 courses (601 and 602), the aim will be to also spread the statistics material over 607 and 613, and to integrate the two more tightly. As an example, the material on 'descriptive' (i.e., not model-based) statistics and graphical displays will be covered in 613, while 607 will begin with parameters and models. Rather than treat computing as a separate activity, exercises based on 607 material will be carried out as part of 613 classes/tutorials. The statistical material will be used to motivate the computer tasks.

0.7 Appendix:

[Still rough] History of current introductory biostatistics courses

The senior author first taught a 2-course sequence for first year graduate students in epidemiology in 1980, using Colton's Statistics in Medicine as the text for the introductory course (607). He developed his own notes for the second

course, which covered multiple regression for quantitative responses. Over the next 10 years, he continued to teach the first course – first from Colton, but latterly from Moore and McCabe (and undergraduate text) and with epi statistics from Armitage and Berry and some other fundamentals from Freedman (Statistics). Stan S taught the second (621 Data Analysis in the Health Sciences), mostly from Kleinbaum's Applied Regression Analysis and Other Multivariable Methods.

In the 1990s, Lawrence J taught 607, and Michal A 621. Neither used a required textbook. LJ developed an extensive set of written notes (still available on his website) (and contributed a chapter Introduction to Biostatistics: Describing and Drawing Inferences from Data book on Surgical Arithmetic) while MA used transparencies that were widely photocopied. 2000s Robert P 621? LJ 621

Meanwhile JH taught to summer students (mostly medical residents and fellows): 607 and a second course (678, Analysis of Multivariable Data). Both sets of content are available on JH's website. He last taught the Fall version of 607 in 2001, when LJ was on sabbatical.

607: Tina 2006 - 201x Erica M; 20xx - Paramita SC. 2018, 2019 Sahir B 621: Aurelie Alexandra 2020 Shirin

Chapter 1

Introduction

1.1 Goals

Blah

1.2 Structure

Blah Blah

1.3 Attitudes, etc....

Blah Blah Blah

Part I

Part I

Chapter 2

Statistical Parameters

2.1 Parameters

The **objectives** of this chapter are to

- Define what a parameter is in a statistical context
- See examples of such parameters
- Understand the concept of a parameter relation or a parameter equation
- Be able to set up parameter equations that isolate and directly pinpoint parameter differences in both the absolute and relative scales, using a regression equation framework.
- Do so **before** fitting any such (regression) equations to data, so that we can focus on the research objects without having data get in the way.
- See the unity (generality) in what we will be doing in the course, by seeing the big picture, i.e., the forests, not the trees.

We begin by defining is meant by the term parameter **in a statistical context**

> **Parameter** – A constant (of unknown magnitude) in a (statistical) model. [OSM2011, p60]
>
> In Statistics. A numerical characteristic of a population, as distinguished from a statistic obtained by sampling.[OED]

Note that the term can mean other things in other contexts. For example, in **clinical medicine**,

> Parameter – any quantitative aspect/dimension of the client's (patient's) health, subject to measurement (by means of a test).

(Example: systolic blood-pressure.)[OSM, Terms and Concepts of Medicine]

The (statistical) parameters we will be concerned with

- μ The mean level of a quantitative characteristic, e.g. the depth of the earth's ocean or height of the land, or the height / BMI / blood pressure levels of a human population. [One could also think of mathematical and physical constants as parameters, even though their values are effectively 'known.' Examples where there is agreement to many many decimal places include the mathematical constant pi, the speed of light(c), and the gravitational constant G. The speed of sound depends on the medium it is travelling through, and the temperature of the medium. The freezing and boiling points of substances such as water and milk depend on altitude and barometric pressure]. At a lower level, we might be interested in personal characters, such as the size of a person's vocabulary, or a person's mean (or minimum, or typical) reaction time. The target could be a person's 'true score' on some test – the value one would get if one (could, but not realistically) be tested on each of the (very large) number of test items in the test bank, or observed/measured continously over the period of interest.

  Later on we will address sitautions where the mean μ is not the best 'centre' of a distribution, and why we might want to take some other feature, such as the median, or some other quantile, instead.

- π Prevalence or risk (proportion): e.g., proportion of the earth's surface that is covered by water, or of a human population that has untreated hypertension, or lacks internet access, or will develop a new health condition over the next x years. At a lower level, we might be interested in personal proportions, such as what proportion of the calories a person consumes come from fat, or the proportion of the year 2020 the person spent on the internet, or indoors, or asleep, or sedentary.

- λ The speed with which events occur: e.g., earthquakes per earth-day, or heart attacks or traffic fatalities per (population)-year. At a lower level, we might be interested in personal intensities, such as the mean number of tweets/waking-hour a person issued during the year 2020, or the mean number of times per 100 hours of use a person's laptop froze and needed to be re-booted.

Each of these three parameters refers to a characteristic of the overall domain, such as entire surface of the earth, or the entire ocean, or population. There are no indicators for distinguishing among subdomains, so they refer to locations / persons not otherwise specified. We will drill down later.

Especially for epidemiologic research, and also more generally, one can think of π and λ as *parameters of occurrence*. [Although the word occurrence usually has a time element, it can also be timeless: how frequently a word occurs in a static text, or a mineral in a rock.] Prevalence is the proportion in a current

state, and the 5-year risk is the expected proportion or probability of being in a new state 5 years from now. The parameter λ measures the speed with which the elements in question move from the original to the other state.

Even though the depths of the ocean, and blood pressures, are measured on a *quantitative* (rather than on all or none) scale, one can divide the scale into a finite number of bins/caterories, and speak of the prevalence (proportion) in each category. Conversely, one can use a set of descriptive parameters called quantiles, i.e, landmarks such that selected proportions, e.g., 0.05 or 5%, 25%, 50%, 75%, 95% of the distribution are to the left of ('below') these quantiles.

Occurence Parameters are not constants of nature [OSM1995]

> It has been noted in the philosophy of science that any science is concerned with functional relations of its objects (Friend and Feibleman, 1937). This proposition is quite evidently tenable for epidemiologic objects of research. Parameters of occurrence, such as the incidence rate for a particular illness, are not constants of nature. Rather, their magnitudes generally depend on — are functions of — a variety of characteristics of individuals — constitutional, behavioral, and/or environmental. Such relations, even if only remotely credible, are generally the objects of medical occurrence research. For example, one is quite usually interested in learning whether the rate of occurrence of some particular illness depends on (is related to or is a function of) gender — regardless of whether there is any express reason to surmise that it might be.
>
> EXAMPLE 1.5 The prevalence of any given blood type based on the ABO antigen system„ while constant over gender and essentially constant over age, is not a constant of nature. It varies by ethnic groupings, for example. Thus the prevalence must be quantified in relation to—as a function of—ethnic group.
>
> EXAMPLE 1.6. For the occurrence of various values of blood pressure among people, one descriptive parameter is the median of the pressure. (This is a value such that the prevalence of its exceedance is 50%.) This parameter, again, is not a constant of nature but depends on age and other characteristics of individuals. For the quantitative nature of the age relation of systolic blood pressure, a rule of thumb used to be that it is, in mm Hg, 100 plus age in years." This rule expresses a regression model - a **regression function** - of the form P = A + B x Age. In this example, P, the occurrence parameter, is the median of systolic blood pressure, A = 100 mm Hg, and B = 1 mm Hg/yr.
>
> The characteristics on which the magnitude of an occurrence parameter depends (causally or otherwise) are **determinants** of the parameter. Thus, in the examples given above, ethnic grouping is

> a determinant of prevalence of any given blood type, and age is a determinant of the median of systolic blood pressure.
>
> "Determinant" has no implication as to causality in science — any more than in everyday locution: the current age of a person is "determined" by his/her year of birth (noncausally), just as the expected outcome of a disease is "determined" by the treatment that is used (causally). The relation of an occurrence measure to a determinant, or a set of determinants, is naturally termed an occurrence relation or an occurrence function. These relations are in general the objects of epidemiologic research. [Even though the general inconstancy of occurrence parameters leads to the consideration of occurrence relations, this latter outlook affords only a partial accommodation of the inconstancy, because occurrence relations the degree also vary according to the type of individual. In particular, measures of a relation (Appendix 2) have determinants of their own.]

Before we start, a comment on terminology

Before we go on, we need to adopt sensible terminology for referring generically to the states, traits, conditions or behaviours whose category-specific parameter values are being compared. Following OSM (see above) we will use the term '**determinant**'. It has several advantages over the many other terms used in different disciplines, such as exposure, agent, independent/explanatory variable, experimental condition, treatment, intervention, factor, risk factor, predictor.

The main advantage is that it is broader, and closer to causally neutral in its connotaion. *Exposure* has environmental connotations, and technically refers to an opportity to injest or mentally take on board a substance or message. *Agent* has causal connotations. The term *independent variable* suggests the investigator has control over it in a laboratory setting. The term *explanatory* is ambiguous as to the mechanism by which the parameter value in the index category got to be different from the value in the index category. Not all contrasts are experimentally formed. The term *factor*, and thus the term *risk factor*, are to be avoided because the word factor derives from the Latin *facere*, (the action of) doing, making, creating. *Predictor* makes one think of the future. The term *regressor* (or its shorthand, the 'X') won't be understood by lay people.

While the word 'determine' can suggest causality (e.g., demand determines the price), it also refers to 'fixing the form, position, or character of beforehand': two points determine a straight line; the area of a circle is determined by its radius.

There is considerable philosophical debate as to whether something 'causes' something else. Some would argue that the extent to which genetics determines one's personality is a causal concept. Others argue that since one cannot cannot consider the alternative, ones biological sex or age can not be considered a causal determinant or a risk factor (in the strict causal meaning of the word). They prefer to refer to them as risk *indicators*.

We now move on to the parameter relations we will be concerned with, beginning with the simplest type.

2.2 Parameter Contrasts

In applied research, we are seldom interested in a single constant. Much more often we are interested in the contrast (difference) between the parameter values in different contexts/locations (Northern hemisphere vs Southern hemisphere), conditions/times (reaction times using the right versus left hand, or behaviour on weekdays versus weekends), or sub-domains or sub-populations (females vs males). Contrasts involving 'persons, places, and times' have a long history in epidemiology.

In this section, we will limit our attention to 'contrasts': a compariosn of the parameter values between 2 contexts/locations/sub-populations. Thus (unlike in OSM's example 1.6) the parameter function has just 2 possible 'input' values. The next section will address more general parameter functions.

'Reference' and 'Index' categories

In many research contexts, the choice of 'reference' category (starting point, the category against which the other category is compared) will be obvious: it is the status quo (standard care, prevailing condition or usual situation, dominant hand, better known category). The 'index' category is the category one is curious about and wishes to learn more about, by contrasting its parameter value with the parameter value for the reference category.

In other contexts, it is less obvious which category should serve as the reference and the index categories, and the choice may be merely a matter of persepctive. If one is more famiar with the Northern hemisphere, it serves as a natural starting point (or 'corner' to use the terminology of Clayton and Hills, or reference category). The choice of reference category in a longevity contrast between males and females, or in-hospital mortality rates or motor vehicle fatality rates during weekends versus weekdays, might depend on what mechanism one wishes to bring out. Or one might close as the reference category the one with the larger amount of experience, or maybe the one with the lower parameter value, so that the 'index minus reference' difference would be a positive quantity, or the 'index: reference ratio' exceeds 1.

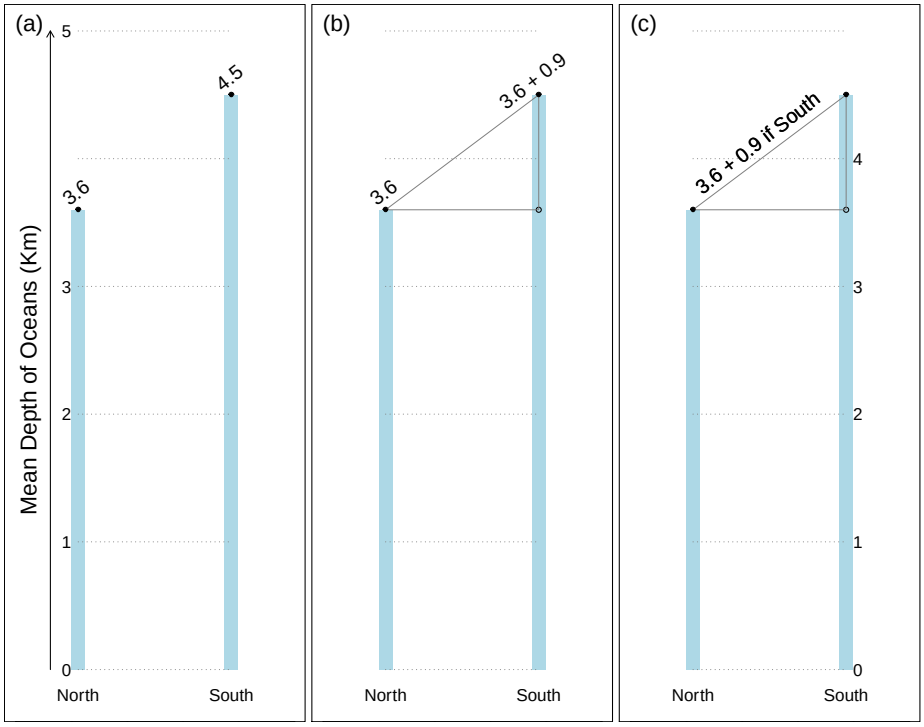
2.2.1 Parameter relations in numbers and words

To make this concrete, we will use hypothetical (and very round) numbers and pretend we 'know' the true parameter values – in our example of the mean depth of the ocean in the Northern hemisphere (reference category) and Southern hemisphere (index category) – to be 3,600 metres (3.6Km) and 4,500 metres (4.5Km) respectively. Thus, the difference (South minus North) is 900 metres or 0.9Km.

If we wished to show the two parameter values graphically, we might do so using the format in panel (a), which shows the 2 hemisphere-specific parameter values – but forces the reader to calculate the difference.

Panel (b) follows a more reader-friendy format, where the difference (the quantity of interest) is isolated: the original 2 parameters are converted to 2 new, more relevant ones.

Panel (c) encodes the relation displayed in panel (b) in a **single phrase** that applies to **both** categories: Onto the 'starting value' of 3.8Km, one **adds** $\Delta\mu$ = 0.9 Km **only if** the resulting parameter pertains to the Southern hemisphere. The 0.9 Km is toggled off/on as one moves from North to South.



2.2.2 Parameter relations in symbols, and with the help of an index-category indicator

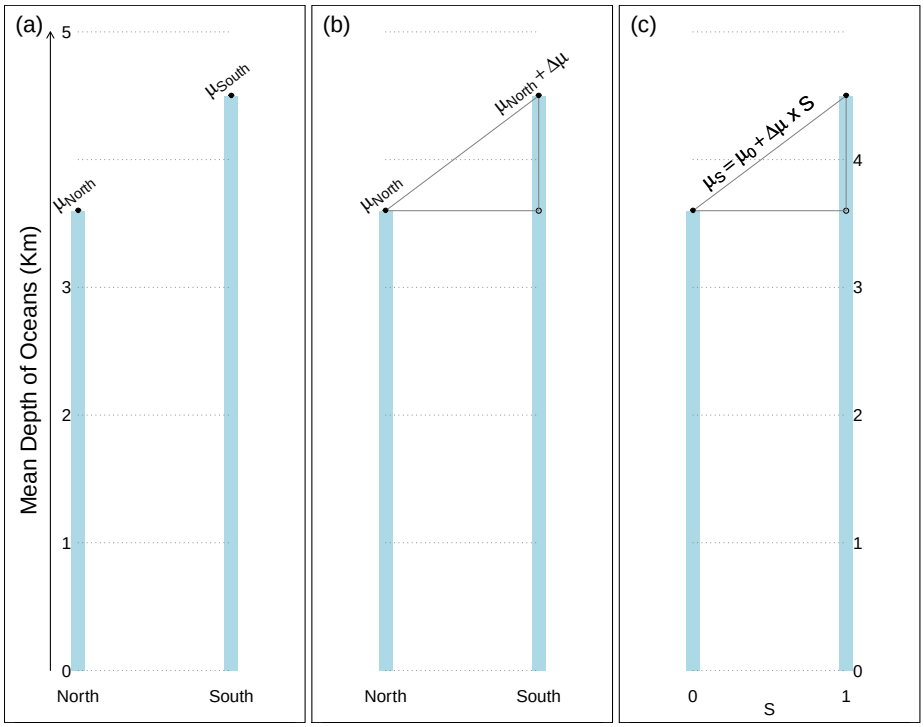
Panels (a) and (b) in the following figure repeat the information in panels (a) and (b) in the preceding Figure, but using Greek letters to symbolically represent the parameters. Just to keep the graphics uncluttered, the labels North and South are abbreviated to N and S and used as subscripts. Also, for brevity, the expression $\Delta\mu$ denotes $\mu_S - \mu_N$.

The relation encoded in a single phrase shown in the previous panel (c) has a compact form suitable for verbal communication. The representation can be

adapted to be more suitable for computer calculations. (The benefit of doing this will become obvious as soon as you try to learn the parameter values by fiiting these models to actual data.) Depending on whether the hemisphere in question is the northern or southern hemisphere, the expression/statement 'the specified hemisphere is the SOUTHERN hemisphere' evaluates to a (logical) FALSE or TRUE. In the binary coding used in computers, it evaluates to 0 or 1, and we call such a 0/1 variable an 'indicator' variable.$\widehat{}$

$\widehat{}$ In 'better families' we speak of INDICATOR variables, not DUMMY variables.
The International Statistical Institute's Dictionary of Statistical Terms objects to the name: the term is 'used, rather laxly, to denote an artificial variable expressing qualitative characteristics ….. [The] word 'dummy' should be avoided.'
Miettinen's Epidemiological Research: Terms and Concepts:- Indicator variate – A variate with 0 and 1 as its (only) realizations, with realization 1 indicating something particular. (Examples: Y = 1 indicating membership in the case series of person-moments and X1 = 1 indicating index category of the etio-genetic determinant in an etiogenetic study – in the logistic model for the object of study.) Dummy variate (synonym: indicator variate) – See 'Indicator variate' in section II – 2. Note: This term is a misnomer: there is nothing dummy about an indicator variate.
We encourage you to use, in your coding, **meaningful variable names** such as `i.South` or `i.Southern` (where `i` stands for indicator of) or `i.Male`. Don't use the name `sex` or `gender`, where the coding is not self evident. If you think `i.Male` is over doing it, then use `Male`.

In panel (c) in the following figure, just to keep the graphics uncluttered, the name of the indicator variable SOUTHERN is abbreviated to S, and μ_S is shorthand for the μ cooresponding to whichever value (0 or 1) of S is specified (we could also write it as $\mu|S$, or μ 'given' S.) Thus, the symbol μ_0 refers to the μ when $S = 0$, or in longerhand, to $\mu \mid S = 0$.



What does the equation in panel (c) remind you of?

Probably the equation of a line. In high school you may have learned it in the form $A + B \times X$ that Miettinen used to describe the relation between median blood pressure and age.

Today, in statistics, these equations are referred to as **regression equations**, and the statistical model is called a regression model. The term 'regression' is unfortunate, since it bears little relation to the original application. It concerned the phenomenon of 'reversion' first described by Charles Darwin. Following his first studies of the sizes of the 'daughter' seeds of sweet peas, his nephew Francis Galton, described the tendency:

> offspring did not tend to resemble their parent seeds in size, but to be always more mediocre ('middling', or closer to the mean) than they — to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small (Galton 1886)

One of the ***first 'regression' lines fitted to human data*** is Galton's line depicting the 'rate of regression on hereditary stature' where, using the term 'deviate' where today we would use 'Z-scores.'

> The Deviates of the Children are to those of their Mid-Parents as 2 to 3. (Galton, 1886)

Because he used z scores (so the means in the parents and in the children were both 0) the equation of the line simplified to

$$\mu(\text{Z-score in children of parents with mean z}) = 0 + (2/3) \times z$$

But don't we need a *cloud of points* to have a regression line?

Although many courses and textbooks introduce regression concepts this way, the answer is **NO.** There is nothing in the regression formulation that specifies at which 'X' values the mean Y values at these X values are to be determined. Unlike many textbboks that start with Xs on a 'continuous' scale, and then later have to deal with a 2-point (binary) X, we are starting with this simplest case, and will move 'up' later.

We are doing this for a few reasons: in epidemiology, the first and simplest contrasts involve just two categories, the reference category and the index category; a simple subtraction of 2 parameter values is easier to do and to explain to a lay person; and there is no argument about how the function behaves at the values between 0 and 1. There are no parameter values at Male = 0.4 or Male = 1.4, they are only at Male=0 and Male=1.

In addition, it is easier to learn the fundamental concepts and principles of regression if we can easily 'see' what exactly is going on. Fewer blackbox formulae mean more transparency and understanding.

Once we see how to represent parameter values in two determinant-categories, we can easily extend it to more than two, such as the ethnic groups in example 1.5 above.
As we will see later on, when we have a value for a dental health parameter (eg the mean number of decayed, missing and filled DMF teeth) at X = 0 parts per million of fluoride in the drinking water, and another parameter value at X = 1 parts per million, we can only look at these 2 parameter values. If this is not enough, we would need to have (obtain) parameter values at the intermediate fluoride levels, or levels beyond 1 ppm, to trace out the full parameter relation, namely how the mean-DMF varies as a function of fluoride levels. If we have large numbers of observations at each level, then the DMF means will trace out a smooth curve. If data are limited, and the trace is jumpy/wobbly, we will probably resort to a sensible smooth function, the coefficients of which will have to be estimated from (fitted to) data.

This discussion leads on naturally to situations where the parameter varies over quantitative levels of a determinant - a topic considered in the next section.

But meantime, we need to answer this question: why limit ourselves to subtraction? why not consider the ratio of the two parameters, rather than their difference?

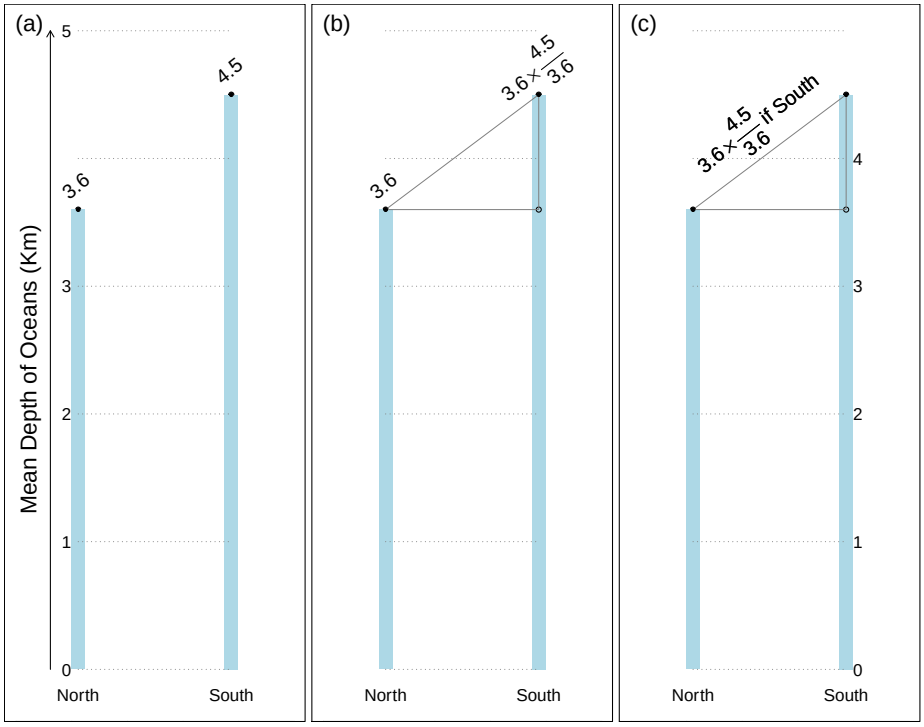
Relative differences (ratios) – *in numbers first*

A ratio can be more helpful than a difference, especially if you are don't have a sense of how large the parameter value is even in the reference category. As an example, on average, how many more red blood cells do men have than women? or how much faster are gamers' reaction times compared with nongamers?

Recall our hypothetical mean ocean depths, 3.6 Km in the oceans in the Northern hemisphere (reference category) and 4.5Km in the oceans of the Southern hemisphere (index category). Thus, the S:N (South divided by North) ratio is 4.5/3.6 or 1.25.

Panel (a) leaves it to the reader to calculate the ratio of the parameter values. In panel (b) the ratio (the quantity of interest) is isolated: again, the original 2 parameters are converted to 2 new, more relevant ones.

Again, panel (c) shows a single master-equation that applies to both hemispheres by togging off/on the ratio of 4.5/3.6.



Relative differences (ratios) – expressed in *symbols* and with the help of the *index-category indicator*

To rewrite these numbers in a symbolic equation suitable for a computer, we again convert the logical 'if South' to a numerical Southern-hemisphere-indicator, using the binary variate S (short for Southern) that takes the value 0 if the Northern hemisphere, and 1 if the Southern hemisphere.

But go back to some long-forgotten mathematics from high school to be able to tell the computer to toggle the ratio off and on. Recall '**powers**' of numbers, where, for example, 'y to the power 2', or y^2 is the square of y. The two powers we exploit are 0 and 1. 'y to the power 1', or y^1 is just y and 'y to the power 0', or y^0 is 1.

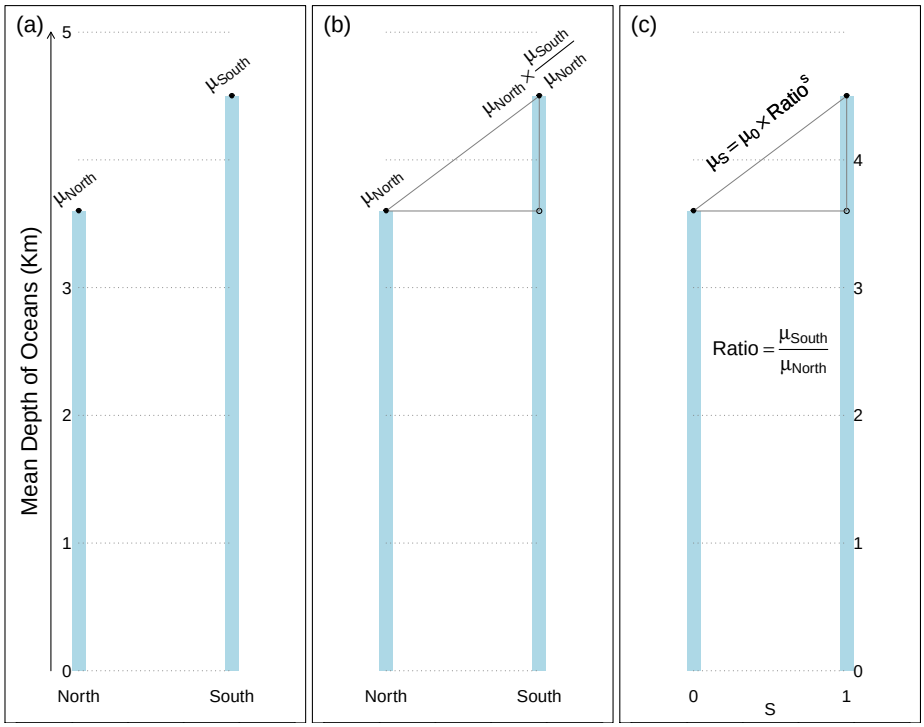
We take advantage of these to write

$$\mu_S = \mu \mid S \ = \mu_0 \ \times \ \left\{\frac{\mu_{South}}{\mu_{North}}\right\}^S = \mu_0 \ \times Ratio^{\ S}.$$

You can check that it works for each hemisphere by setting $S = 0$ and $S = 1$ in turn.

Thus,

$$\log(y^S) = S \times \log(y)$$



Although this is a compact and direct way to express the parameter relation, it is not well suited for fitting these equations to data.

However, in those same high school mathematics courses, you also learned about **logarithms**. For example, that

$$\log(A \times B) = \log(A) + \log(B); \quad \log(y^x) = x \times \log(y).$$

Thus, we can rewrite the equation in panel (c) as

$$\log(\mu_S) = \log(\mu \mid S) \; = \underbrace{\log(\mu_0)}\; + \underbrace{\log(Ratio)}\times S.$$

This has the same 'linear in the two parameters' form as the one for the parameter difference: the parameters are $\underbrace{\log(\mu_0)}$ and $\underbrace{\log(Ratio)}$ and they are made into the following 'linear compound' or 'linear predictor' (see Remarks below) :

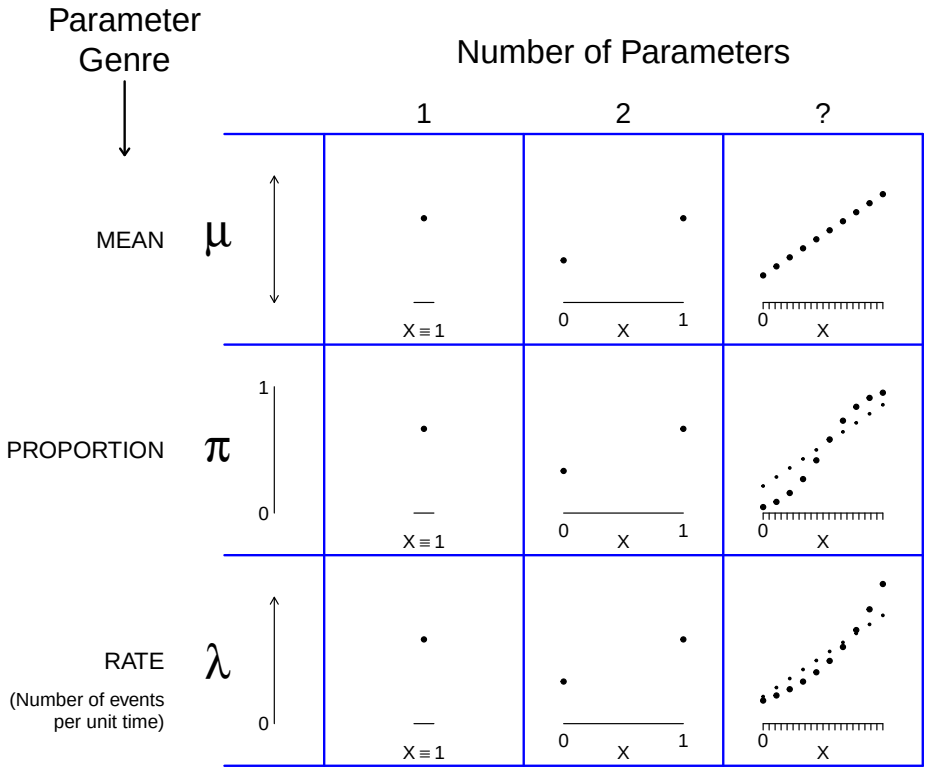
$$\log(\mu_S) = \log(\mu \mid S) \; = \underbrace{\log(\mu_0)}\times 1 \; + \underbrace{\log(Ratio)}\times S.$$

The course is concerned with using 'regression' software to 'fit'/'estimate' these 2 parameters from n depth measurements indexed by S.

2.3 Parameter functions

A very simple example of a function that describes how parameter values vary over quantitative levels of a determinant is the straight line shown in the upper right panel of the next figure. Here the determinant has the generic name X, and the equation is of the $A+B\times X$ or $B_0+B_1\times X$ or $\beta_0+\beta_1\times X$ straight line form. Miettinen used the convention that the upper case letters A and B are used to denote the (true but unknown) coefficient values, whereas the lower case leters a and b are used to denote their empirical counterparts, sometimes called estimated coefficients or fitted coefficients. This sensible and simple convention also avoids the need, if one uses Greek letters for the theoretical coefficient values, to put 'hats' on them when we refer to their empirical counterparts, or 'estimate/fit' them. Fortunately, journals don't usually allow investigators to use 'beta-hats'; but this means that the investigators have to be more careful with their words and terms.

As we go left to right in the following grid, the models become more complex. The simplest is the one of the left, in column 1, the one JH refers to as 'the mother of all regression models.' It refers to a *single* or *overall* situation/population/domain, so $X \equiv 1$, it takes on the value 1 in/for every instance/member. So the parameter equation is $\mu_X = \mu \mid X \; = \mu \times 1$. In column 2, there are 2 subdomains, indexed by the 2 values of the determinant (here generically called 'X'), namely $X = 0$ and $X = 1$. In the 3rd column, the number of of parameters is left unspecified, since the numbers of coefficients to specify a line/curve might vary from as few as 1 (if we were describing how the volume of a cube dependeds or, was is function of, its radius) to 2 (for a straight line that did not go through the origin, or for a symmetric S curve) to *more than 2* (e.g., for a non-symmetric S curve, or a quadratic shape).



A few more remarks on the panels in this Figure

- The 3 rows refer to the 3 core parameters we have given examples of above. All 3 are governed by the same principles, although there are more possibilities of different possible scales for some parameters.

- In setting (column) 1: there is just 1 parameter (value shown as a dot) corresponding to the 'overall' population or the entire domain. You can think of it as the limiting or 'degenerate' case of the columns to its right. One can still write it in a 'regression' model.

  It is of the form P(arameter) = B, involving no indicators for distinguishing among subdomains of the referent domain of the distribution, say adults not otherwise specified. [MSH2018, p63] It is sometimes referred to as a null or 'incercept only' regression model.

  We will exploit this idea to take a more holistic/general and economical approach to this introductory course. Many text-books/courses do not mention regression models until quite late, and spend a lot of time on '1-sample' (and even '2-sample') problems without pointing out that these are merely sub-cases of regression models. This 'silos' practice of promoting/learning a separate

software routine for dealing with a 1 sample problem, when one can get the same answer from a regression routine, leads to dead ends and wastes time.

Once we get to fitting/estimating a mean (or proportion or rate) parameter to/from data, we will encourage doing so within a regression framework.

- In setting (column) 2, there are 2 parameter values, one for the reference category and one for the index category of the determinant. As we have seen, how they relate to each other can be can be expressed in a number of different ways. A common and useful way is via a parameter equation that contains a parameter for the reference category and a comparative parameter (some measure of the difference between the two parameters) – the latter is often of most interest.

- In setting (column) 3, the parameter equation traces the parameter over a continuum of possible values of the determinant, using as many coefficients as are needed. In this particular diagram, the values of the determinant (X) are shown starting at X = 0, but this does not have to be. In data analysis, one often shifts the X origin, so that the 'intercept' makes more sense. For example, if one was plotting world temperatures, or ice-melting dates (see Chapter on Computning) against calendar year, it would be better to have the incercept refer to the fitted temperature for when the series begins, rather than when our current Western calendat begins (at the year 0 AD). Likewise, if we were describing the relation between ideal weight and height it is good to start near where people's heights are. Thus, '100 pounds for a **height of 5 feet**, with five additional pounds for each added inch of height' for women, and '106 pounds for a **height of 5 feet**, and six additional pounds for every added inch of height.' for men. Of course, if you wish, for women you could use the mathematically equivalent '-300 pounds for a **height of 0 feet**, with five additional pounds for each added inch of height' but it is not that easy to remember, and doesn't apply for much of the (unspecified) height range!

A few remarks on associated terminology

Instead of **'regression models'**, some textbooks and courses refer to **'linear models'** :

> **Linear model**: Formulation of the mean/'expectation' of (the distribution of) a random variate (Y) as a linear compound of a set $B_0, B_1, B_2, ...$ of parameters: as $B_0 + B_1X_1 + B_2X_2 + ...$ [Miettinen 2011, p54]
>
> **The meaning of 'linear'** in the appellation of this model has nothing to do with straight lines; it refers to the mathematical concept of 'linear compound': given quantities Q_1, Q_2, etc., a linear compound of these is the sum $C_1Q_1 + C_2Q_2 + ...$, where C_1 etc. are the 'coefficients' that define a particular linear compound of the set

> of quantities constituted by the Qs. So, the 'general linear model' is linear in the sense that the dependent parameter, M, is formulated as a linear compound of the independent parameters B0, B1, etc., the coefficients in this linear compound being 1, X1, etc. The model is, in this way, 'linear in the parameters.' [MSH2018, p65]

Statistics courses in the social sciences, the biological laboratory sciences, and other experimentally-based sciences, typically move on from 1- and 2-sample procedures (unfortunately, mainly focusing on statistical *tests*) to **'analysis of variance' models**

Miettinen explains an **'analysis of variance models** this way:

> In the 'analysis of variance model,' the random variate at issue – Gaussian – has a mean whose value depends on a *nominal-scale determinant*, a nominal scale being characterized by discrete categories without any natural order among them. The names of the (nominal) categories, some N in number, could be Category 1, Category 2, … , Category N. The term for the model is a misnomer. For, at issue is not analysis but synthesis of data, and the synthesis is not directed to learning about the variance of the random variate; it focuses on the mean, the relation of the mean to the (nominal-scale) determinant of it.
>
> A simple example of these models might address the mean of systemic blood-pressure – defined as the weighted average of the diastolic and systolic pressures with weights 2/3 and 1/3, respectively – in relation to ethnicity, represented by three categories. An 'analysis-of- variance' model would define a random variate (Y) as representing the numerical value of the pressure (statistical variates inherently being numerical) and having a Gaussian distribution with means M1, M2, and M3 in those ethnicity categories 1, 2, and 3, respectively, with the variance of the distribution invariant among them. The random variate (Y) is the '*dependent*' variate in the meaning that the value of its mean depends on ethnicity; and the ethnicity categories are represented in terms of suitably-defined '*independent*' – non-random – variates (Xs).
>
> The form of the 'analysis-of-variance' model in this simple example is: $M = B_0 + B_1X_1 + B_2X_2$, where M is the mean of Y and the two independent variates are indicators of two particular ones of the three ethnic categories. One possibility in this framework is to take X_1 and X_2 to be indicators of Category 2 and Category 3, respectively – an indicator variate being one that takes on the value 1 for the category it indicates, 0 otherwise.
>
> In terms of this model, B_0 is the value of M when $X_1 = X_2 = 0$, that is, for Category 1 (i.e., $B_0 = M_1$); and for Category 2 and Category 3 the values of M are represented by $B_0 + B_1$ and $B_0 + B_2$,

respectively (i.e., $M_2 = B_0 + B_1$, and $M_3 = B_0 + B_2$). Thus, the difference between M_1 and M_2 is represented by B_1; B_2 represents the difference between M_1 and M_3; and the difference between M_2 and M_3 is the difference between B_1 and B_2.

In this 'analysis-of-variance' framework it is feasible to accommodate, jointly, whatever number of nominal-scale determinants of the magnitude of the mean of the dependent variate. A simple example of this is the addition of the two categories of gender for consideration jointly with the three categories of ethnicity. These two determinants jointly imply a single nominal-scale determinant with six categories (as each of the three categories of ethnicity is split into two subcategories based on gender).

When involved in the definition of the independent variates is only a single determinant, the model is said to be for 'one-way analysis of variance'; with two determinants the corresponding term (naturally) is 'two-way analysis of variance'; etc.

2.4 Phraseology to avoid

It is quite common to hear a regression coefficient (fitted or theoretical) interpreted this way:

> "For every 1 unit increase in X, the 'Y' parameter increases by β_X units."

or as follows

> "As (when) you increase X by 1 unit, you increase the Y parameter by β_X units."

We pick up this terminology very early, maybe even back in high school, and from other people around us. But, in interpreting the B = 1 mm Hg/yr in Miettenen's example (100 plus age in years), should we use such phrases?

Or, since you don't know the source of, or the data behind this rule, you can take a look at the distributions of some anthropometric chacacteristics (height, weight, forced expiratory volume, FEV) measured cross-sectionally, in different populations – Busselton, Australia and rural Southwest Ethiopia – in 1972 and 1992. By eye, try to estimate the slope you would get if you regressed the age-and sex-specific means or medians on the ages. and then summarize the gradient across age.

Remember that these these subjecsts aren't aging or going anywhere, and nobody was watching them age.

It is more accurate to say:

> People who were aged a+1 years at the time of the survey had heights/weights/FEVs that were t.tt units higher/lower than people who were aged a years.

or

> The mortality rate was u.uu units higher/lower (or u.u times higher/lower) in the experience in the index category than the reference category.

This way, you are telling the reader that this is a static source, and not a dynamic situation where conditions are being manipulated by the investigators, or the subjects being watched as their ages go up [for many readers, the word 'increased' implies that some human force deliberately changed the dial, and turned the X up or down, as one could do with temperature or humidity in a laboratory.]

One of JH's favourite examples of people being misled into thinking that a cross-sectional dataset allows you to say that 'as people get older, they ...' is the McGill epidemiology department's studies, in the 1960s, on the health of the more than 10,000 millers and miners of asbestos. These workers were born between 1890 and 1920. In cross-sectional studies, there were gradients in mean height across attained age. It would be easy to give them a 'as people get older, they shrink in height' or they 'gain in height' interpretation. It is easy to overlook the fact that some of these were children and adolescents during the Depression.

Next Chapter

The next chapter will begin in the upper left corner of the grid, and address situations where the **estimand** (the parameter to be estimated) is μ. It will describe how we 'estimate' / 'fit' a single μ parameter from/to a finite number of observations, and how we quantify and report how far off the target our method of estimation can/might be.

2.5 SUMMARY

2.6 Exercises

So far, we have only dealt with equations involving a difference and the ratio of two μ parameters.

1. Extend the graphs and the equations (for the difference of means and the ratio of means) to the π parameter. Use as an example the proportions of the surfaces of the Northern (reference category) and Southern hemisphere (index category) covered by water, i.e, π_{North} and π_{South}. Use the hypothetical values $\pi_{North} = 0.65$ and $\pi_{South} = 0.75$.
2. Instead of focusing on the proportions covered by *water*, focus on the proportions covered by *land*.

- How does the **difference** of the two proportions relate to the difference calculated in 1?
- How does the **ratio** of the two proportions relate to the ratio calculated in part 1? i.e., is one the reciprocal of the other?
- Can you think of different scale, where the ratio when the focus is *land* IS just the reciprocal of the ratio when the focus is *water*.?
- If you can, show that the log of the ratio when the focus is *land* IS just the negative of the log of the ratio when the focus is *water*.?

3. Extend the graphs and the equations (for the difference of means and the ratio of means) to the λ parameter. Use as an example the mean number of earthquakes per year in the Northern (reference category) and Southern hemisphere (index category) , i.e, λ_{North} and λ_{South}. Use the hypothetical values $\lambda_{North} = 5.0$ and $\lambda_{South} = 7.5$

2.7 References

Olli S. Miettinen. Theoretical epidemiology: Principles of Occurrence Research in Medicine. Wiley, New York, 1985. Chapter 1: The study of occurrence patterns in medicine. Introduction.

David Clayton and Michael Hills. Statistical Models for Epidemiology. Oxford University Press, 1993. Chapter 22: Introduction to regression models.

Kenneth J. Rothman. Epidemiology: An introduction. Oxford University Press, 2012. Chaper 12: Using regression models in epidemiologic analysis.

Olli S. Miettinen, Johann Steurer, Albert Hofman. Clinical Research Transformed. Springer, 2019. Chapter 7: The Logistic Regression Model (The Precursors of the General Linear Model; The General Linear Model; The Generalized Linear Model; The Logistic Regression Model)

Chapter 3

The parameter μ

[and other location (and spread and shape) parameters]

The **objectives** of this chapter are to

-
-

Although few textbooks do so, we think it is worth distinguishing two contexts.

3.1 Two genres

- The *first* is where, *if* there were no measurement issues, we would 'see' / 'get' / 'observe' the same constant every time we made a determination, but where, because of unavoidable measurement variations, there is a statistical distribution of 'measurements' around that constant. Examples include measurements of constants such as the speed of light, or of physical constants, such as a standard weight (e.g. 1 Kg) or of a fixed distance measured by a smart phone app or a fixed number of steps measured by a step-counter. Examples of 'personal' constants that are constant – at least in the short term – but not easily or reproducibly measured might be the size of a person's vocabulary, or a person's mean (or minimum, or typical) reaction time. Or, the target could be a person's 'true score' on some test – the value one would get if one (could, but not realistically) be tested on each of the (very large) number of test items in the test bank, or observed/measured continously over the period of interest.

  Starting with this simpler 'measurement variation only' context makes it easier to master the statistical laws that govern the variation of values derived from a combination of measurements, the variation of *statistics*.

- The *second* is where the variation is primarily (or in a few deluxe cases where there are no measurement issues, entirely) due to genuine – e.g., biological – variation. Examples of such (often effectively infinite in size) biological distributions include the depths of the ocean, or the heights or weights or blood pressures of a specific population.

  In this context, the distribution is less likely to display the symmetry observed when the variation is entirely due to measurement variations around some constant. Thus, there may be several possible choices of the 'centre' of the distribution. So, in addition to pursuing the 'mean' parameter μ we will also pursue other numerical parameters for the centre.

No matter which of the two genres we are dealing with, it may be important to quantify the spread (and maybe the shape) of the distribution. Even if this aspect may be of secondary interest, it has a bearing on what we can say about how far off the target (off the parametr) our estimators might be.

We will begin with the first genre, where, *if* there were no measurement issues, we would 'see' / 'get' / 'observe' the same constant every time we made a determination, but where, because of unavoidable measurement variations, there is a statistical distribution of 'measurements' around that constant. Because its estimation involves the same statistical laws as when the distribution/variation is biological, we will refer to this elusive 'constant' as μ.

3.2 Fitting these to data / Estimating them from data

Experiments to Determine the Density of the Earth. By Henry Cavendish, Esq. F.R.S. and A.S. Philosophical Transactions of the Royal Society of London, Vol. 88. (1798), pp. 469-526.

http://www.medicine.mcgill.ca/epidemiology/hanley/bios601/Mean-Quantile/Cavendish1798.pdf

The following Table contains the Results of the Experiments

```
density = c(
5.50, 5.61, 4.88, 5.07, 5.26, 5.55, 5.36, 5.29, 5.58, 5.65,
5.57, 5.53, 5.62, 5.29, 5.44, 5.34, 5.79, 5.10, 5.27, 5.39,
5.42, 5.47, 5.63, 5.34, 5.46, 5.30, 5.75, 5.68, 5.85)

round(mean(density),2)
```

```
## [1] 5.45
```

```
lm.fit =  lm(density ~ 1)
print(summary(lm.fit),digits=1)
```

```
##
```

```
## Call:
## lm(formula = density ~ 1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -0.57  -0.15   0.01   0.16   0.40
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.45       0.04     133   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2 on 28 degrees of freedom
```

```
round(confint(lm.fit),2)
```

```
##             2.5 % 97.5 %
## (Intercept)  5.36   5.53
```

```
library(mosaic)
bootstrap.fits <- do(1000) * lm( resample(density) ~1)
round(confint(bootstrap.fits$Intercept),2)
```

```
##            2.5% 97.5%
## percentile 5.36  5.52
```

```
round(sd(bootstrap.fits$Intercept),2)
```

```
## [1] 0.04
```

- Metrics (criteria) for measuring (best) fit

Chapter 4

The (proportion) parameter π

4.1 Example one

etc

4.2 Example two

etc

Chapter 5

The (event rate) parameter λ

5.1 Etc

-

5.2 ETC

-

Chapter 6

Contrast: 2 μ parameters

6.1 Estimand, estimator, estimate

Chapter 7

Contrast: 2 π parameters

7.1 Estimand, estimator, estimate

Chapter 8

Contrast: 2 λ parameters

8.1 Estimand, estimator, estimate

Part II

Part II

Chapter 9

Probability

9.1 Conditional – forwards

•

9.2 Conditional – reverse

- Application: medical diagnostic tests

Exercises: Efron, Monty Hall, Economist, Wald : CF screening

Chapter 10

Distributions /Random Variables

Chapter 11

Statistical Inference

11.1 2 schools

11.1.1 Bayesian

11.1.2 Frequentist

Part III

Part III

Chapter 12

Mathematics

12.1 Notation

- **Variables and Subscripts**
  Variable Y with n sample values denoted $y_1, y_2, ..., y_n$ in order of entry; The "1", "2", … are called subscripts or indices. We use the letter i (or j) and the range "1 to n" to denote the n different y values and refer to the value of the ith y as "y_i".
- **Summation**
  The term Σy (spoken: "sigma y" or "sum of y's") is used as a shorthand for the sum $y_1 + y_2 + \cdots + y_n$.

12.2 Powers, Logarithms and Anti–logarithms

- The term $y^{1/2}$ is shorthand for the square root of y or $\sqrt{y}$. Likewise, $y^{1/n}$ denotes the n-th root of y.
- $\ln(y)$ denotes the "natural log of y" or "log of y to the base e" i.e. $\log_e(x)$, where e is 2.718.
  Note: y must be positive; ln (y) ranges from -inf to +inf.

  ln (0.1) = -2.30; ln (1) = 0; ln (2) = 0.69; ln (10) =2.30
- $\ln(A \times B) = \ln(A) + \ln(B)$; $\ln(\frac{A}{B}) = \ln(A) - \ln(B)$
- $\exp(y)$ is shorthand for e^y or "exponential of y" or the natural anti-log of y. y ranges from -∞ to +∞. and $\exp(y)$ yields a positive value. eg. exp(-1) = 0.36; exp(0) =1; exp(.5) =1.64; exp(1) = 2.71…

Chapter 13

Computing Week 1

The **'computing' objectives** are to learn how to use `R` to put series of observations into vectors, and how to plot one series against another.

The **'statistical' objective** of this exercise is to understand the concept of a distribution of a numerical characteristic (here an amount of elapsed time), and the various numbers describing its 'central' location and spread, and other 'landmarks'. You will also be introduced (in the next section) to 2 functions that give a more complete description of a distribution.

13.1 Biological background

Later on we will examine climate trends using unusual datasets, which suggest that over the last few centuries, winter tends to end earlier, and plants tend to flower earlier.

One such dataset arose as part of a long-running contest, the Nenana Ice Classic More here

13.2 Statistical Task

You are asked to approximate and carefully examine the distribution of guesses in 2018, contained in the Book of Guesses for that year.

For now, we will measure the guesses (and eventually the actual time) as the numbers of days since the beginning of 2018. Thus a guess of Tuesday April 17 5:20 p.m. would be measured as 31 + 28 + 31 + 16 + (16 + 20/60)/24 = 106.6806 days since the beginning of 2018.

It would be tedious to try to apply optical character recognition (OCR) to each of the 1210 pages in order to be able to computerize all of the almost 242,000

guesses. Instead, you are asked to reconstruct the distribution of the guesses in two more economical ways:

1. By determining, for each of the 36 x 2 = 72 half-days days from April 10 to May 15 inclusive, the proportion, p, of guesses that are earlier than midnight on that date. [In `R`, if p = 39.6% of the guesses were below xy.z days, we would write this as pGuessDistribution(xy.z) = 0.396. Thus, if we were dealing with the location of a value in a Gaussian ('normal') distribution, we would write `pnorm(q=110, mean = , sd = )`] Once you have determined these 72 proportions (p's), plot them on the vertical axis against the numbers of elapsed days since the beginning of the year on the horizontal axis. Thus the horizontal axis runs from 92 + 10 = 102 days to 92 + 30 + 15 = 137 days.

2. By determining the 1st, 2nd, … , 98th, 99th percentiles. These are specific examples of 'quantiles', or q's. The q-th quantile is the value (here the elapsed number of days since the beginning of 2018) such that a proportion q of all guesses are below this value, and 1-q are above it. [In `R`, if 40% of the guesses were below 110.2 days, we would write this as qGuessDistribution(p=0.4) = 110.2 days. Thus, if we were dealing with the 40th percentile of a Gaussian distribution with mean 130 and standard deviation 15, we would write `qnorm(p=0.4, mean = 130, sd = 15).`] Once you have determined them, plot the 99 p's (on the vertical axis) against the 99 (elapsed) times on the horizontal axis.

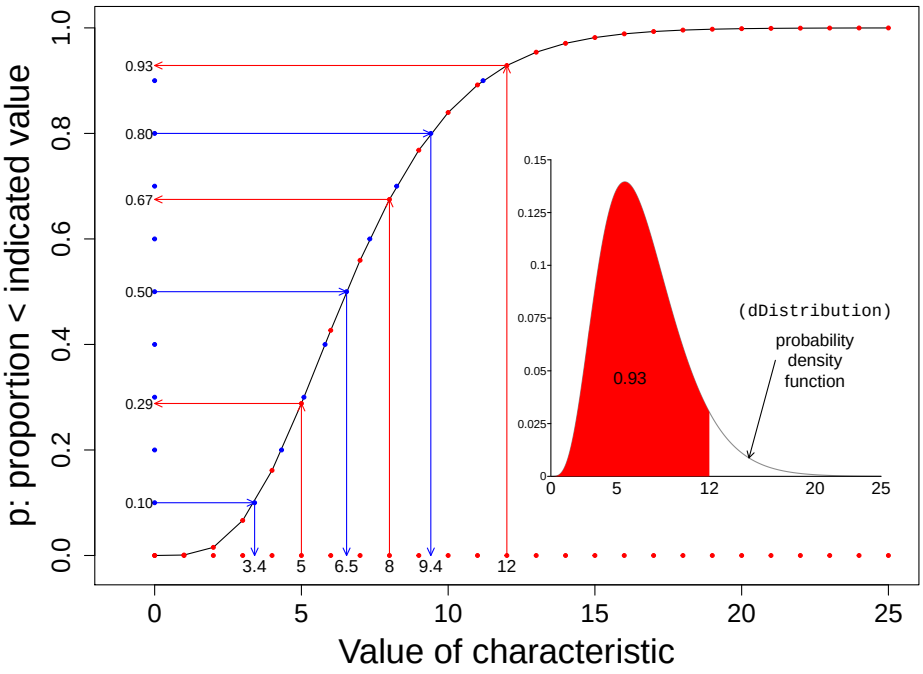
13.2.1 The p and q functions: an orientation

The '**p**' function tells us, for a given value of the characteristic, what **p**roportion of the distribution lies to the left of this specified value.

The '**q**' (or quantile) function tells us, for a given proportion p, what is the value of the characteristic such that that specified proportion p of the distribution lies to the left of this 'q' value.

In the plot below, the values of the **p** function are shown on the vertical axis, in red, against the (in this case, equally-spaced) values of the characteristic, shown on the horizontal axis. You enter on the horizontal axis, and exit with an answer on the vertical axis.

The **q** function (in blue) goes into the opposite direction. You enter at some proportion on the vertical axis, and exit with a value of the characteristic (a quantile) on the horizontal axis. In our plot, the proportions on the vertical axis are equally-spaced. Percentiles and quartiles are a very specific sets of quantiles: they are obtained by finding the values that divide the distribution into 100 or into 4.



13.2.2 Exercises

1. Once you have determined the 72 (cumulative) proportions (p's) associated with the 72 half-days, plot them on the vertical axis against the numbers of elapsed days since the beginning of the year on the horizontal axis. Thus the horizontal axis runs from 92 + 10 = 102 days to 92 + 30 + 15 = 137 days.

2. The 1st, 2nd, … , 98th, 99th percentiles are not so easy to determine since you have to locate the 2419th, 4839th, 7258t, … entries in the 1201-page Book of Guesses and plot the 99 p's (on the vertical axis) against the 99 (elapsed) times (q's) on the horizontal axis. Instead, use the first entry on each of pages 11, 21, … in this excerpt. Using a different colour, plot these slightly-more-dense quantiles on the horizontal axix against the following percentages:

```
entries = 200*seq(10,1200,10) + 1
percent = 100 * entries/241929
noquote( paste(head(round(percent,1),10),collapse="%, ") )
```

```
## [1] 0.8%, 1.7%, 2.5%, 3.3%, 4.1%, 5%, 5.8%, 6.6%, 7.4%, 8.3
```

```
tail(round(percent,1),10)
```

```
##  [1] 91.8 92.6 93.4 94.2 95.1 95.9 96.7 97.5 98.4 99.2
```

3. Compare the Q_{25}, Q_{50}, and Q_{75} obtained directly with the ones obtained by interpolation of the curve showing the results of the other method.

4. Compare the directly-obtained proportions of guesses that are before (the end of) April 20, April 30, and May 10 with the ones obtained by interpolation of the curve showing the results of the other method.

5. By successive subtractions, calculate the numbers of guesses in each 1/2 day bin, and make a histogram of them. From them, calculate the mean, the mode, and the standard deviation.

6. (For a future assignment, but you can start thinking about how) From a random sample of 100 guesses from the book, estimate how many guesses in the entire book are PM.

```
my.id = 800606
set.seed(my.id)
n = 50
sample.entry.numbers = sample(x = 1:241929, size=n)
sorted.sample.entry.numbers = sort(sample.entry.numbers)
head(sorted.sample.entry.numbers,10)
```

```
##  [1] 10542 17437 18351 21113 24086 28782 30055 32220 33162 36443
```

```
page.number = ceiling(sorted.sample.entry.numbers/200)
within.page = sorted.sample.entry.numbers-200*(page.number-1)
column.number = ceiling(within.page/100)
row.number = within.page - 100*(column.number-1)

dataset = data.frame(page.number,column.number,row.number)
head(dataset)
```

```
##   page.number column.number row.number
## 1          53             2         42
## 2          88             1         37
## 3          92             2         51
## 4         106             2         13
## 5         121             1         86
## 6         144             2         82
```

```
tail(dataset)
```

```
##    page.number column.number row.number
## 45        1087             1         80
## 46        1097             2          3
## 47        1121             1         16
## 48        1131             1         55
## 49        1175             2         52
## 50        1181             2         30
```

- How far off was the median guess in 2018 from the actual time? Answer in days, and (with reservations stated) as a percentage? {see the 2020 brochure }
- Why did the experts at the country fair do so much better?
- Where were the punters in 2019 with respect to the actual time?
- Instead of measuring the guessed times from the beginning of the year, suppose that, as Fonseca et al did, we measure the guessed times from the spring equinox in Alaska, i.e. from 8:15 a.m. on Tuesday, March 20, 2018, Alaska time. In this scale, compute the mean guess, and the SD of the guesses.
- Suppose, again, we measure the guessed times from the spring equinox, but in weeks. In this scale, compute the mean guess, and the SD of the guesses.

Some links on the 'Wisdom of Crowds'

https://www.technologyreview.com/s/528941/forget-the-wisdom-of-crowds-neurobiologists-reveal-the-wisdom-of-the-confident/

https://www.all-about-psychology.com/the-wisdom-of-crowds.html

http://galton.org/essays/1900-1911/galton-1907-vox-populi.pdf

- How much warmer/colder in Nov-April is Monreal than Nenana?