006 - Statistical Parameters

EPIB 607 - FALL 2020

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

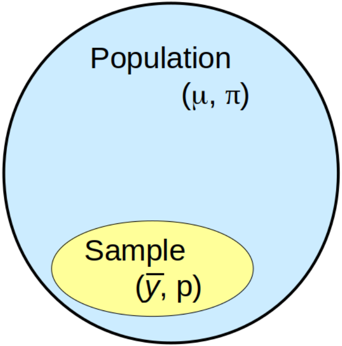
sahir.bhatnagar@mcgill.ca

slides compiled on September 16, 2020

Statistical parameters

Parameters and Statistics

- **Paramter**: An unknown numerical constant pertaining to a population/universe, or in a statistical model.
  - μ: population mean π: population proportion
- **Statistic**: A numerical quantity calculated from a sample. The empirical counterpart of the parameter, used to *estimate* it.
  - $\bar{y}$: sample mean p: sample proportion



The (statistical) parameters we will be concerned with: μ

- 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.

The (statistical) parameters we will be concerned with: π

- 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.

The (statistical) parameters we will be concerned with: λ

- 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.

The (statistical) parameters we will be concerned with

- 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.

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.
- 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.
- 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.

Terminology

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

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).
- In this section, we will limit our attention to ’contrasts’: a compariosn of the parameter values between 2 contexts/locations/sub-populations. Thus the parameter function has just 2 possible ’input’ values.

’Reference’ and ’Index’ categories

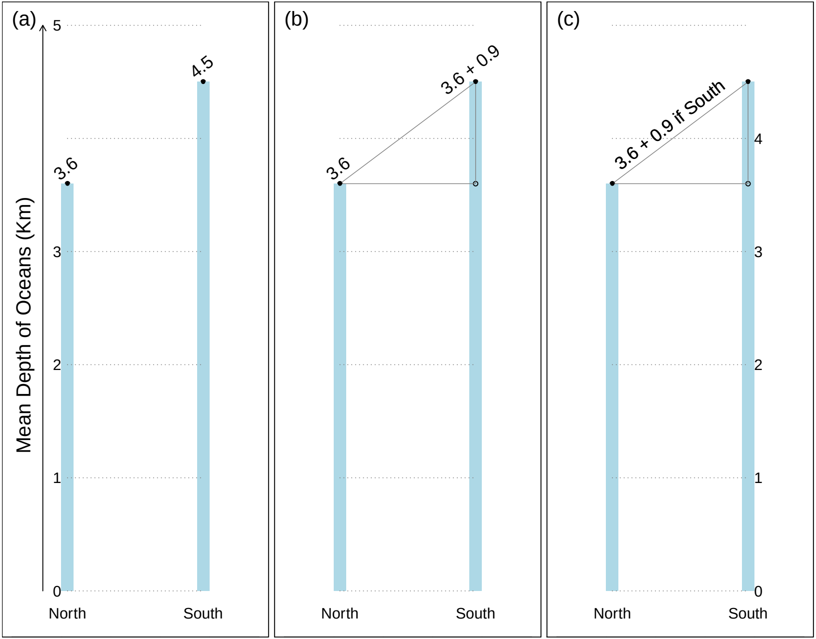
- In many research contexts, the choice of ’reference’ category (starting point, the category against which the other category is compared) will be obvious: it is the **status quo** (standard care, prevailing condition or usual situation, dominant hand, better known category).
- The ’index’ category 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
- 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 chose 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.

’Reference’ and ’Index’ categories

- 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 chose 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.

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-friendly 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.



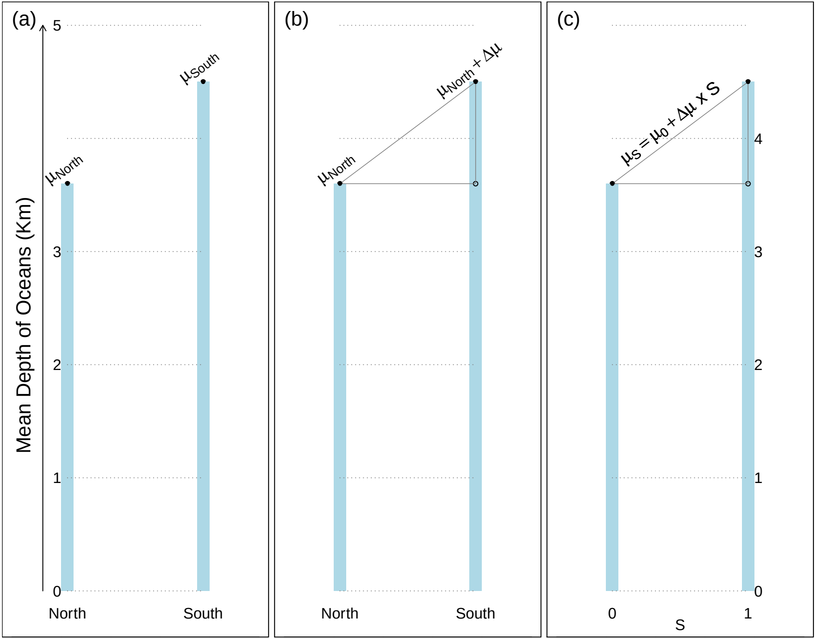
(a)
(b)
(c)
Mean Depth of Oceans (Km)
5
4
3
2
1
0
3.6
4.5
3.6
3.6 + 0.9
3.6 + 0.9 if South
North
South
North
South
North
South

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.

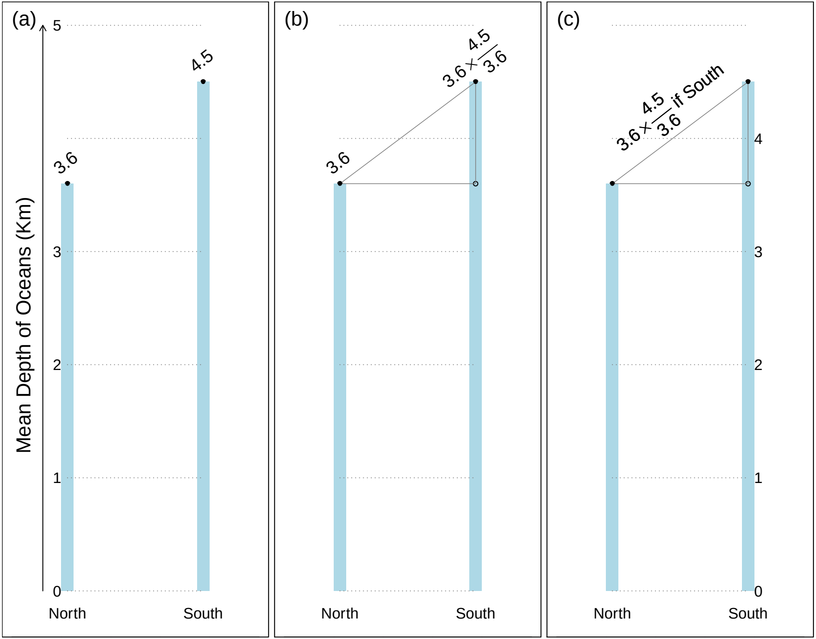
Parameter relations in symbols, and with the help of an index-category indicator

- 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 $\mu \mid S = 0$.


(a)
(b)
(c)
Mean Depth of Oceans (Km)
5
4
3
2
1
0
μ_{South}
μ_{North}
$\mu_{North} + \Delta\mu$
μ_{North}
$\mu_S = \mu_0 + \Delta\mu \times S$
North
South
North
South
0
1
S

Relative differences (ratios) - in numbers

- 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.


(a)
5
4.5
3.6
3
2
1
0
Mean Depth of Oceans (Km)
North
South
(b)
3.6
$3.6 \times \frac{4.5}{3.6}$
North
South
(c)
$3.6 \times \frac{4.5}{3.6}$ if South
4
3
2
1
0
North
South

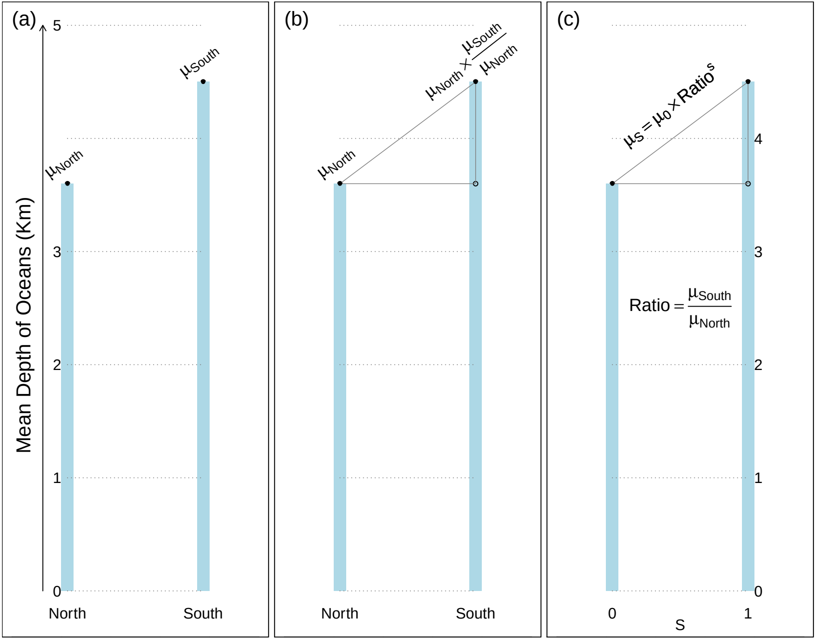
Relative differences (ratios) - expressed in symbols

- 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,

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



(a)
Mean Depth of Oceans (Km)
5
3
2
1
0
μ_{North}
μ_{South}
North
South
(b)
μ_{North}
$\mu_{North} \times \frac{\mu_{South}}{\mu_{North}}$
North
South
(c)
$\mu_S = \mu_0 \times Ratio^s$
$Ratio = \frac{\mu_{South}}{\mu_{North}}$
4
3
2
1
0
0
1
S

- 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

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

- 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.

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.

Parameter Functions

- 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).

Parameter
Genre
↓

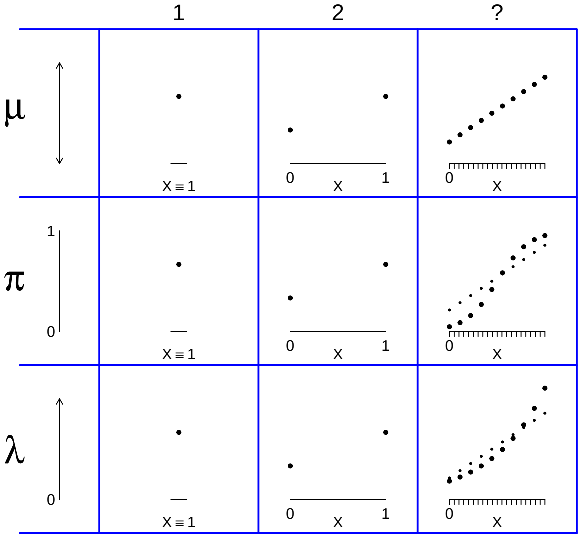
Number of Parameters

MEAN

PROPORTION

RATE

(Number of events
per unit time)


1
2
?
μ
π
λ
1
0
0
X ≡ 1
0
X
1
0
X
X ≡ 1
0
X
1
0
X
X ≡ 1
0
X
1
0
X