Quantitative Methods – Day 1: Inference

Harm H. Schuett

08 January, 2018

Contents

In the previous part we discussed that we are often mostly interested how average outcomes change going from one group of observations to another. But, we have not yet really discussed why we are interested in averages in the first place. That question leads to a host of inference problems and is the focus of this part.

What is statistical inference?

Statistical inference means drawing conclusions based on data, but in a way that explicitely acknowledges uncertainty. For a very simple example, imagine you want to know how much faulty parts a machine produces. You sample 100 parts and you find 10% are faulty. What conclusions do you draw? Is the fault rate really 10%? Or might it be 5% and you simply had a bad sample by chance? Or might it be 20% and we had a luck sample by chance? This is inference: Estimating the fault rate of the machine from a sample and also quantify how uncertain our inferrence is. (e.g, imagine you get a 10% fault rate from a 10.000 parts sample. Are you more confident now that the fault rate is somewhere around 10%?)

What is special about statistical inference is that it tries to draw conclusions by framing the proplem as: **deduce properties of some assumed underlying probability distribution of interest by analysing data.** The distribution of interest describes the population population of interest and its parameters aspects of the population we are interested in, such as mean parameters. In the above example, you could assume that faulty parts follow a poison distribution (because faulty parts are counts) and you want to know the mean of the distribution. It depends a bit on your statistical philosophy what this distribution represents: a) hypothetically repeated sampling (if we sample the machine regularly what would be the distribution of our sampled fault rates?) or b) our uncertainty (the so called bayesian view uses a probability distribution over hypotheses (h1: 10%, h2:, 10.1%, etc. . . .))

The above example might not immediately applicable to scientific questions, but the idea is exactly the same. If you take the standard text book example of "inferring" the returns to schooling (the effect additional years of education on yearly income), we must describe the population of interest and assume some sort of distribution of yearly income. For example we can assume that the Income follows are normal distibution with the mean of the distribution being a function of years of schooling ($Inc \sim N(a_0 + a_1 Scooling, \sigma^2)$). Then we have three aspects of the distribution (a_0, a_1, σ^2) that describe interesting aspects of the population

of interest and we could infer what these parameters might look like from a regression. But we do not know the population, but only have a sample from it.

Inference then encompasses the whole process of deriving estimates and testing hypotheses (or other forms of assessing the uncertainty inherent in the estimates). We always assume that the observed data is somehow sampled from the (larger) population.

Because this is mainstream, we will follow the frequentist approach of thinking about repeated sampling when formulating distributions. **From now on, when talking about a model, we mean all assumptions about the population and distributions to frame the construct of interest**. Always be mindful that are your **inferences are conditional on your assumed model**. The beauty of experiments is that you can often get aways with less assumptions, if done properly.

(Btw: In contrast to inferential statistics, descriptive statistics is solely concerned with properties of the observed data. Descriptive statistics are an often under-appreciated part of the distribution (model) formulation.)

Thinking about Causality and How We use Probability

Nothing in the world is random

> Our goal is to understand the behavioral process that leads to the agent's choice. We take a causal perspective. There are factors that collectively determine, or cause, the agent's choice. Some of these factors are observed by the researcher and some are not. The observed factors are labeled x, and the unobserved factors ϵ. The factors relate to the agent's choice through a function $y = h(x, \epsilon)$. This function is called the behavioral process. It is deterministic in the sense that given x and ϵ, the choice of the agent is fully determined. Since ϵ is not observed, the agent's choice is not deterministic and cannot be predicted exactly. Instead, the probability of any particular outcome is derived. (Train 2009, 3)

The above quote sums up the key perspective that is necessary to understand statistical inference, which is the main mode of thinking when you are an empiricist. You can think of an empirical study as an attempt to estimate an **assumed** relation between to variables Y and X. Often we also **assume** a direction of the relation ($X \rightarrow Y$) But there is a lot of uncertainty about that relation. There may or may not be a relation, the relation might look very different from your assumption, and so on. And most importantly: there are often many more determinants of Y.

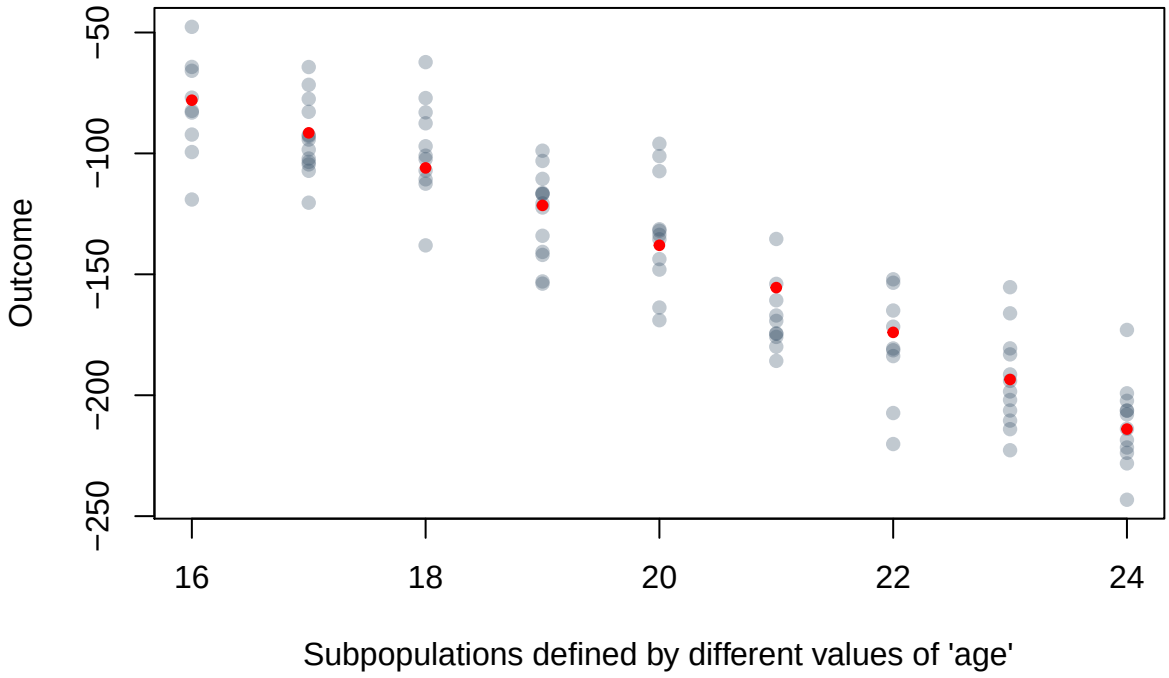
The main parts of this mode of thinking are:

1. The universe is **deterministic**. Everything happens for reasons. Said differently, nothing in the universe is random.
2. We do not know how the universe works. Even though we think the universe is completely deterministic, we do not know the mechanisms that the universe obeys to. As a consequence, things **look random**. *Randomness is thus not a property of the universe but a function of our knowledge of the world.*
3. Whenever we look at data, we need to take our uncertainty (=lack of knowledge of the mechanisms at work) into account. We use randomness as a way of expressing this uncertainty because then we can use probability theory to help us make informed decisions under uncertainty.

Because of all the unmeasured influences, there will always be variation. Remember the example from the first part:

```
set.seed(123)  # fix the random number generator
age_range <- 16:24
age <- sample(x=age_range, size=100, replace=TRUE)  # draw 100 integers randomly
y <- 2 + 3 * age - 0.5*age^2 + rnorm(100, 0, 20)  # y as linear additive function of x
plot(age, y,  # plot x vs y (draws points by default)
     col=rgb(0.2, 0.3, 0.4, 0.3),  # color of the points
```

```
     pch=16,  # point character, 16 is a small dot
     xlab="Subpopulations defined by different values of 'age'",
     ylab="Outcome")
y_ticks <- 2 + 3 * age_range - 0.5 * age_range^2
points(age_range, y_ticks, col="red", pch=20)  # adds average y for evey ticks value
```



Even, though this is simulated, this illustrates the main problem. Suppose for instance the outcome is some kind of happiness index and you get this picture: happiness seems to go down with age. But there is variation around the average. Where would such variation come from? From the myriad other things that influence happiness in addition to age. This is also the reason, we are interested in average effects. We want to know how things change on average as x varies, knowing full well that the outcome depends on tons of other things as well. Sometimes you hear this in workshops: "But does your variable of interest not depend on so many other things?" What is usually behind that question is that the asking person does not believe x to be an important enough variable that would meaningfully shift the average without taking into account many contextual factors. We will talk about how to handle such things in later parts.

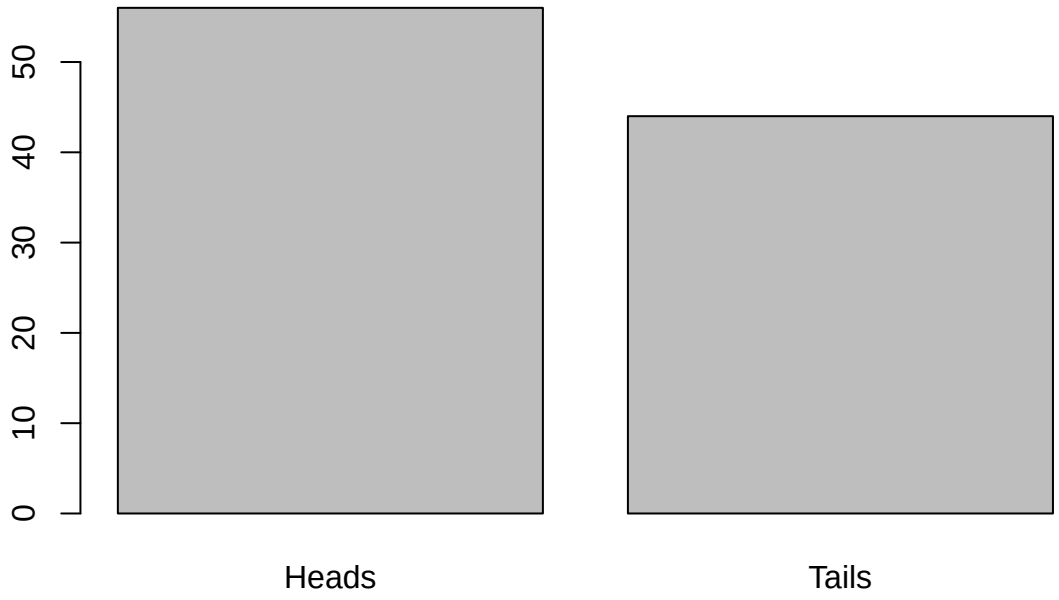
Let's first look at some examples and try to better understand the influence of unmeasured determinants.

The probability of a coin toss is 50%

When you throw a coin in the air a hundred times, how many times does it land heads and how many times tail? You probably have some sort of notion that it should be roughly equal. But it will not be completely 50 / 50. Let's discuss why you have that notion and what's not quite correct about it.

If we would simulate this sequence of coin tosses, we would just draw 100 times from an urn with equal proportions of "head" and "tails" balls:

```r
coin_tosses <- table(rbinom(n=100, size=1, prob=c(0.5,0.5)))
barplot(coin_tosses, names.arg=c("Heads", "Tails"))
```



But when we do this, notice that we have not used any information on the coin tosses. If we would actually perform these tosses and we could measure wind, throwing angle, speed of rotation, etc. we might actually be able to predict quite well how the coin will land. The 50%/50% probability is really a randomness that originates from us not measuring all the determinants of the coin toss. The more we would measure those, the less "random" the coin tosses would be.

Drawing inference from data

Let's use a typical text book question: "What are the returns to schooling?" How much more in wages do you get if you invest into additional education? It is one of the most important labor economics questions because it is another way to frame the question: "how important is education?" with obvious political and regulatory implications.

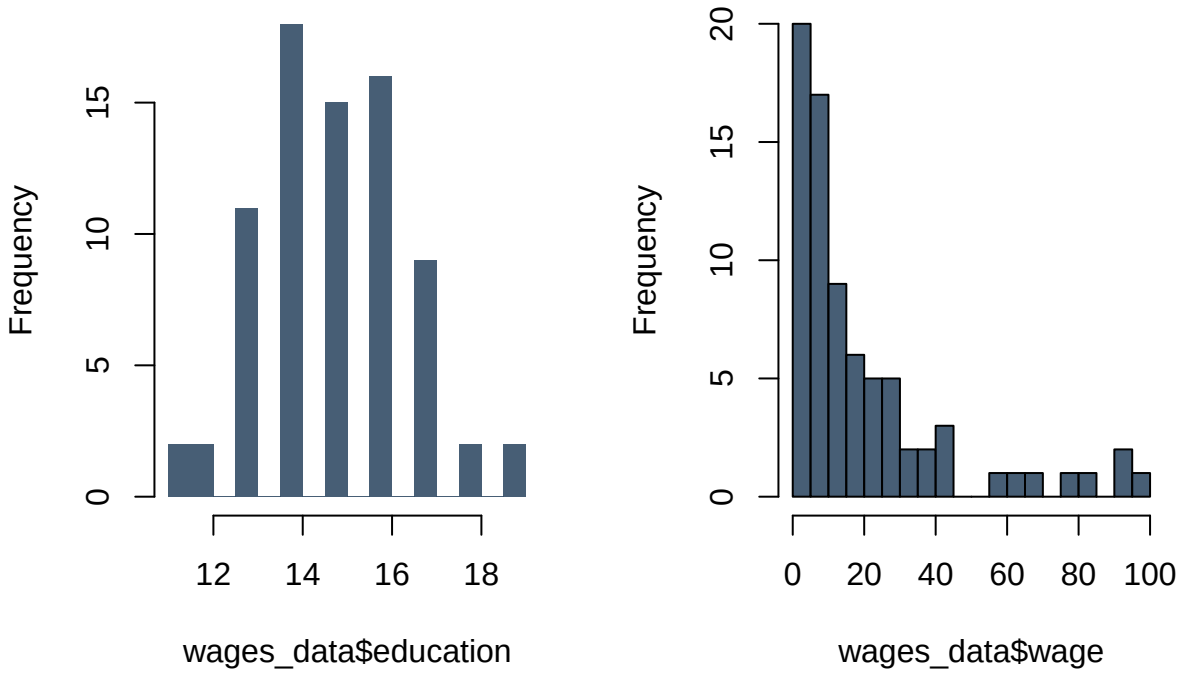
Now, below I give you a data set with wage data. It contains the following variables

```
wage: monthly wage in $100
education: years of education
parent_education: years of education of the least educated parent
school: a code for the high school visited
```

Let's look at the data:

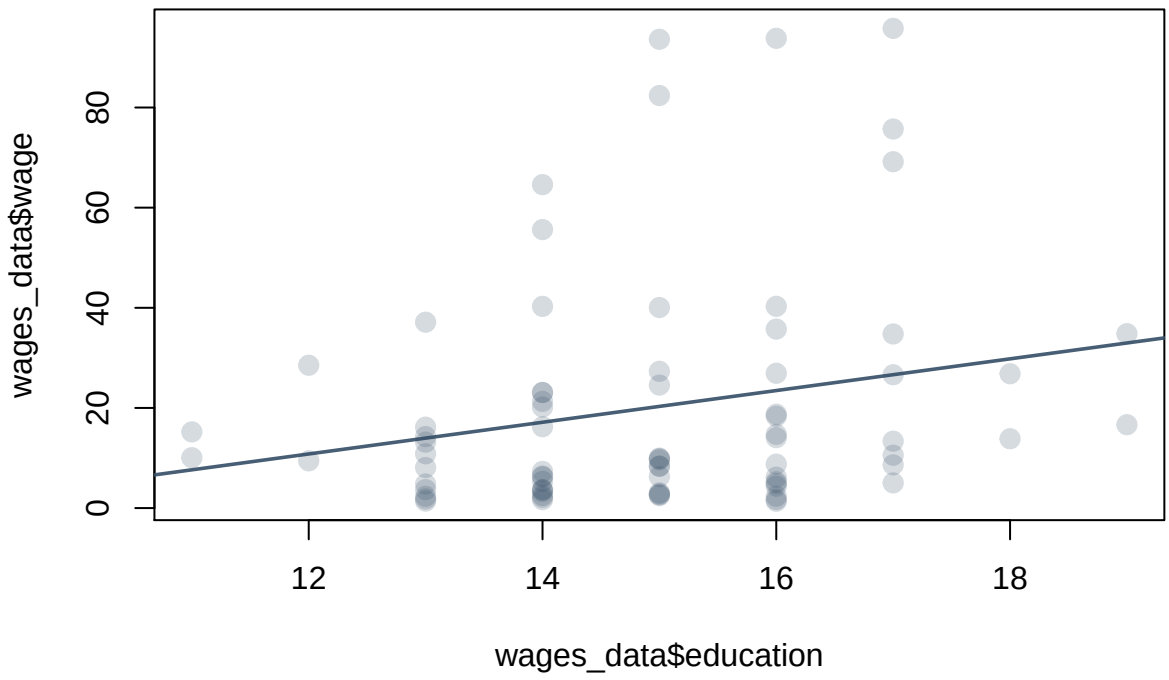
```r
wages_data <- read.csv("../data/wage_sample1.csv")
par(mfrow=c(1,2))  # set graphics options to 1 row and 2 columns
```

```r
cblue <- rgb(0.2, 0.3, 0.4, 0.9)  # define a color
# 1st histogram
hist(wages_data$education,
     breaks=20,  # group datarange into 20 bins
     col=cblue,  # fill color of the bars
     border=NA,  # no border color for the bars
     main=""  # no title above the plot
     )
# 2nd histogram
hist(wages_data$wage, breaks=20, col=cblue, main="")
```



We could run a simple regression to estimate the relation between wages and education Something like:

```r
ablue <- rgb(0.2, 0.3, 0.4, 0.2)
plot(x=wages_data$education, y=wages_data$wage,
     col=ablue,  # last number is the alpha (0.2)
     pch=20,  # plotting 'character'. 20 is solid bullet
     cex=2  # character (or symbol) expansion. Size of the symbol
     )
reg1 <- lm(wage~education, data=wages_data)
abline(reg1, lwd=1.8, col=cblue)
```

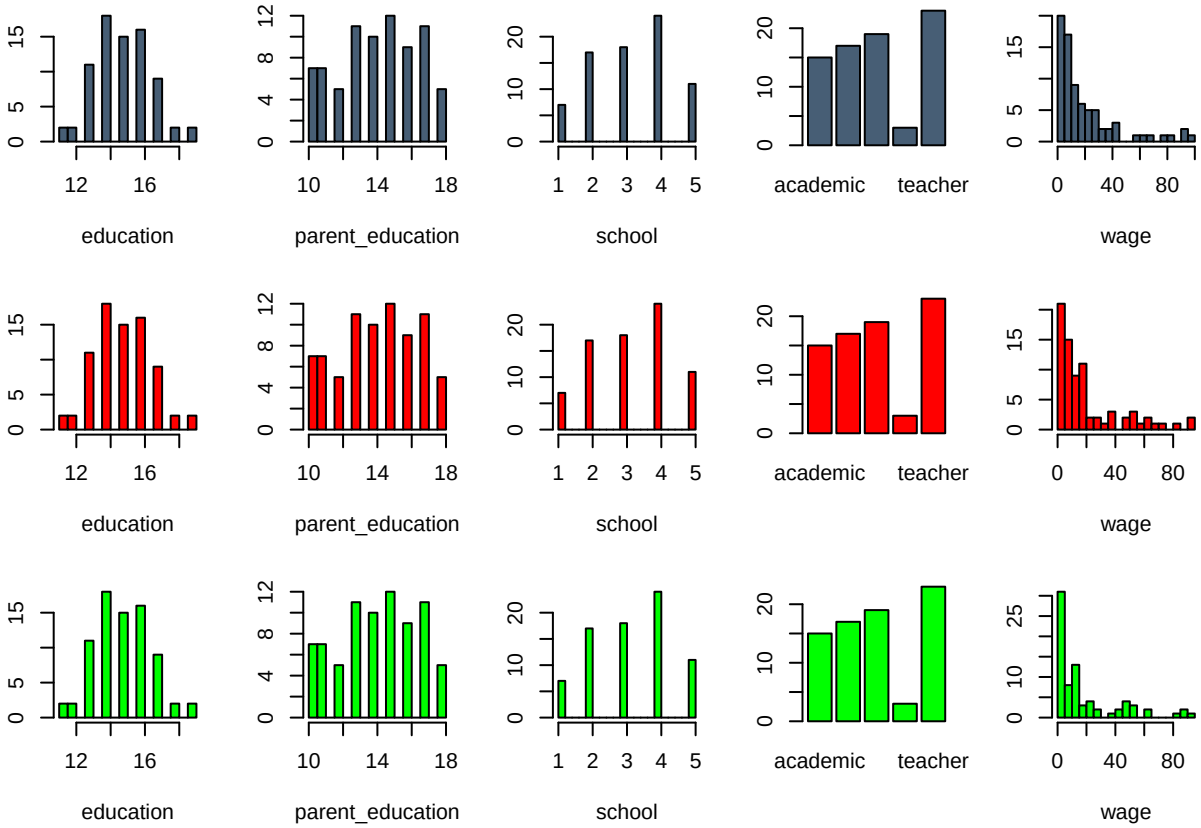


Now, assume that we were able to go into the field and collect new data twice. We didn't find the same people, but we found people with exactly the same characteristics as the ones found in the first sample.

```
wages_data2 <- read.csv("../data/wage_sample2.csv")
wages_data3 <- read.csv("../data/wage_sample3.csv")
reg2 <- lm(wage~education, data=wages_data2)
reg3 <- lm(wage~education, data=wages_data3)
```

Let's compare the newly collected data sets.

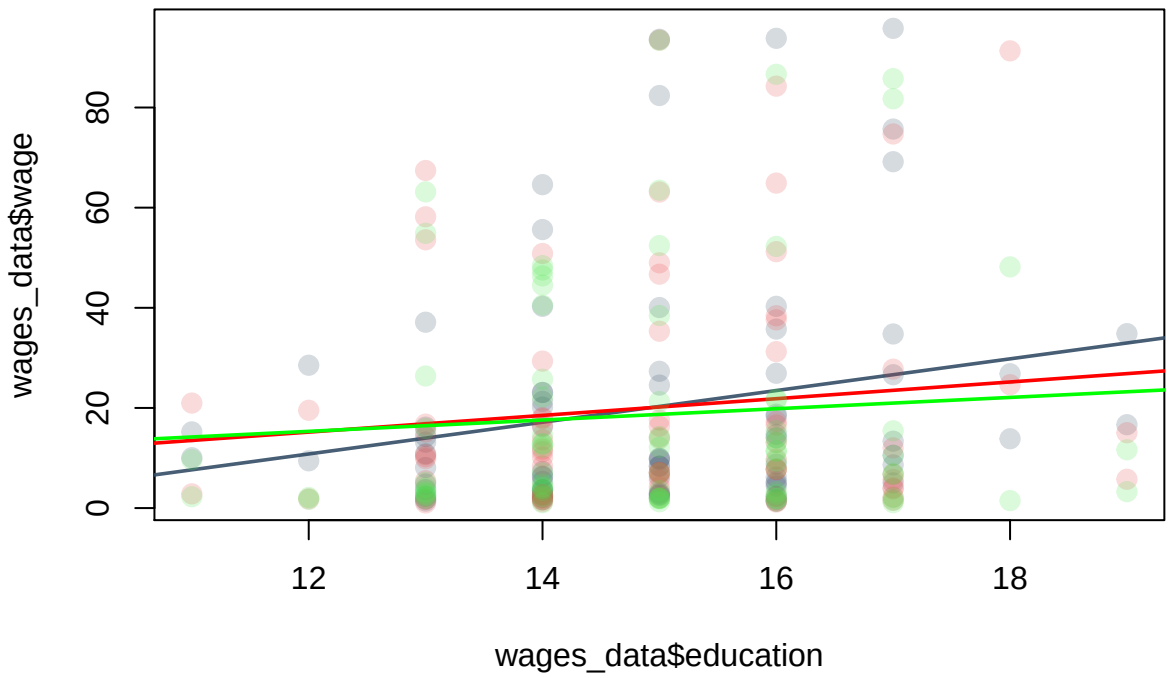
```
par(mfrow=c(3,5),  # 3 rows, 2 columns
    mar=c(4,2,2,2)  # slightly adjust margins
    )
draw_hist_row <- function(data, color, nbreaks=20) {
   with(data, hist(education, main="", breaks=nbreaks, col=color))
   with(data, hist(parent_education, main="", breaks=nbreaks, col=color))
   with(data, hist(school, main="", breaks=nbreaks, col=color))
   tab_occ <- table(data[, "occupation"])
   barplot(tab_occ, main="", col=color)
   with(data, hist(wage, main="", breaks=nbreaks, col=color))
}
draw_hist_row(data=wages_data, color=cblue)
draw_hist_row(data=wages_data2, color="red")
draw_hist_row(data=wages_data3, color="green")
```



As you can see, the survey service took great pains to get as close to collecting the same sample as possible. The plots show that people with exactly equal characteristics were interviewed. However, wage, while showing a similar distribution differs. Why? Because, in addition to the observed variables (education, parent_education, school, and occupation) the wage is probably determined by many other things that they did/could not measure in the first and subsequent surveys. Things that we did not know could matter, or could not measure, or where to expensive to measure, etc. Thus we can never truly replicate a study. Even if we would collect hundred more samples with the same observables, we will always get a slightly different wage distribution.

Of course, this will have an impact on our conditional expectation estimate:

```
plot(x=wages_data$education, y=wages_data$wage,col=ablue, pch=20, cex=2)
abline(reg1, lwd=1.8, col=cblue)
points(x=wages_data2$education, y=wages_data2$wage,col=rgb(.9,.3,0.3, 0.2), pch=20, cex=2)
abline(reg2, lwd=1.8, col="red")
points(x=wages_data3$education, y=wages_data3$wage, col=rgb(.3,.9,0.3, 0.2), pch=20, cex=2)
abline(reg3, lwd=1.8, col="green")
```



As you can see, the regression lines differ. Which one is the correct one? None of the three is. Why? Think about how wage is determined:

$$wage_i = f\left(education_i, parent\_education_i, school\_quality_i, occupation\_wage\_level_i\right) + u_i$$

We *assume* that education, parent_education, school, and occupation matter. We don't know for sure and we don't quite know how either. Said differently, we don't know the functional form of $f(\cdot)$ with certainty. And there are likely many things that also play a role ,like person i's ability. Such additional determinants are captured in the u_i term. It is often called the *error term*, but is really only a catch-all term for everything unmeasured.

So, going back to our free samples, all samples are likely equal in the $f\left(education_i, parent\_education_i, school\_quality_i, occupat$ part, but differ in the u_i part! Frequentists statisticians think of this part as the "random part". **Taken together with our uncertainty about the true mechanism, this part induces uncertainty in our estimates.** Because u_i differs from sample to sample, it induces variation in estimates. This is a critical point that you should really strive to understand.

Imagine, a fully deterministic mechanism:

$$y_i = 1 + 2 * x_i + 3 * z_i - 2 * v_i$$

But you don't know that. You only hypothesize that y and x are related, but you don't know that it is a linear relation, you don't know the effect of x on y is stable, with magnitude 2, and you don't know about z and v. We might assume that the relation is of the general form:

$$y_i = a_0 + a_1 * x_i + u_i$$

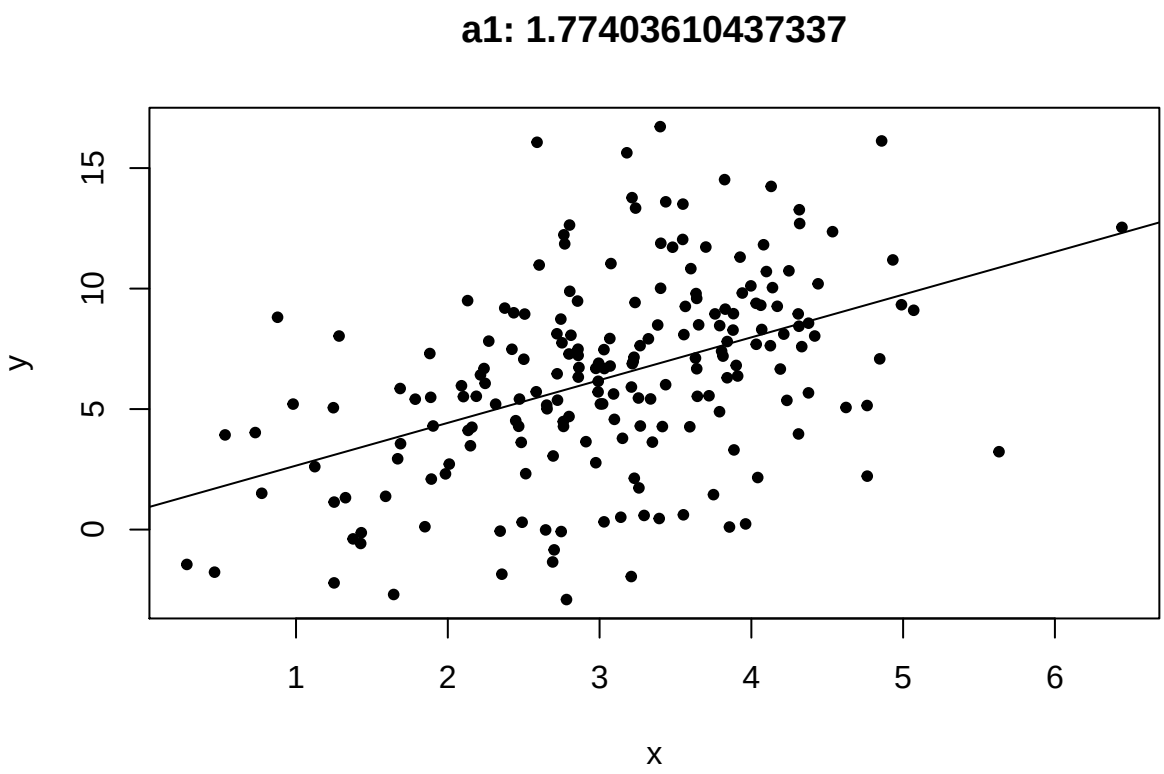
where $u_i = 3 * z_i - 2 * v_i$. We can fit this with a regression.

$$y_i = \hat{a}_0 + \hat{a}_1 * x_i + \hat{u}_i$$

The regression tries to find an $\hat{a}_0$ and $\hat{a}_1$ that minimizes the sum of squared error – tries to explain as much as possible of y. That is simply another way of saying it minimizes the sum of of squared residuals $\hat{u}_i$. Or that it estimates the conditional expectation. How the average y changes with changes in x. Let's simulate this case to make it more accessible.

```
set.seed(9870137)
n <- 200
x <- rnorm(n=n, mean=3, sd=1)
z <- rnorm(n=n, mean=1, sd=1)
v <- rnorm(n=n, mean=2, sd=1)
y = 1 + 2 * x + 3 * z - 2 * v

reg_ex <- lm(y~x)
x_coeff <- coef(reg_ex)["x"]
plot(x, y, pch=20, main=paste("a1:", x_coeff))
abline(reg_ex)
```



You can see that the estimate $\hat{a}_1$ is not exactly equal $a_1 = 2$. Since we simulated this, we know the average of y changes its magnitude by 2, as we changes x by one unit. However, our estimate is only 1.77. Why? This is because of all the unmeasured stuff (v and z in u). v and z affect y, but since we did not measure them, their effect is "random". That random interference makes the estimated relation between y and x differ from 2.

If we would control the effects of these two variables in a multivariate regression, we would not have that

problem

```
coef(lm(y ~ x + v + z))
```

```
## (Intercept)           x           v           z
##           1           2          -2           3
```

You see, if we put v and z into the equation, we get exactly the coefficients. There is no uncertainty left (Except for functional form. Here, we know the functional form – linear, additive – as well)

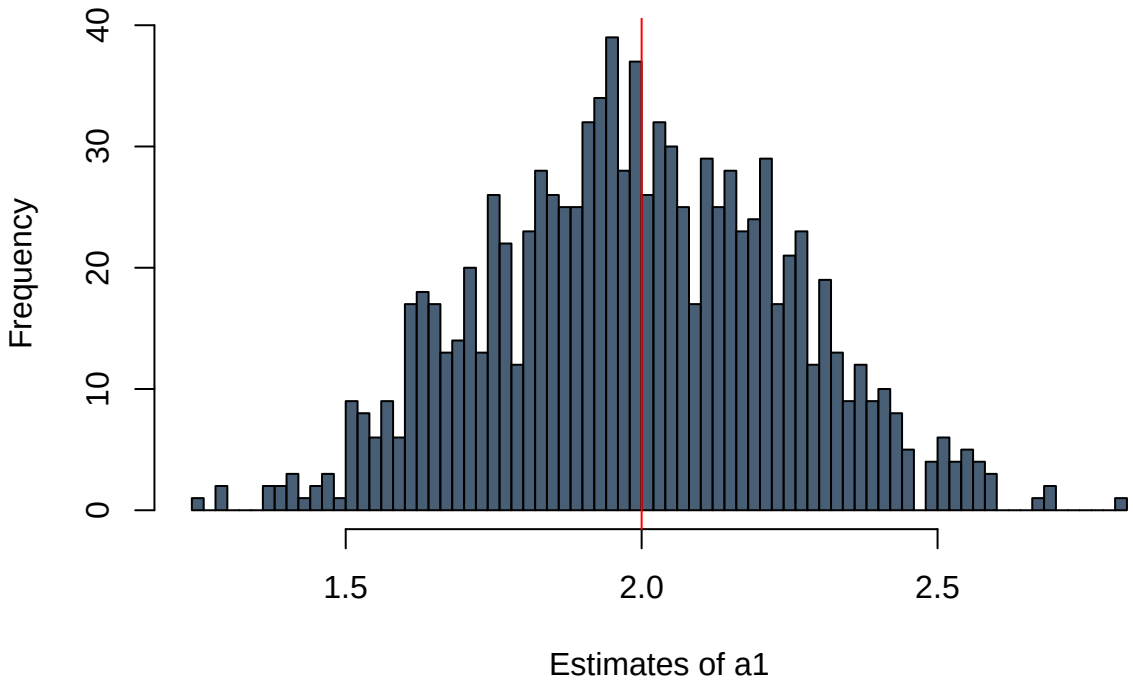
Quantifying Uncertainty

Sampling uncertainty

We understand now where uncertainty comes from and what randomness really means. When we have a research question and a sample that we want to use, we face model uncertainty and uncertainty from unmeasured influences. In frequentist terms, we now make a thought experiment: what would happen to our estimates, if we could replicate our sample, with exactly the same measured, observable variables? Of course, we won't be able to replicate the unmeasured determinants, so our outcome variable will be different, leading to different estimates. **But, if we could do such replications, we can use that variation in estimates to measure uncertainty in estimates due to the variation in unmeasured variables**. This is again something we can simulate:

We keep the same x. Because we draw 1000 different z and v and recompute y.

```
reps <- 1000
a1_sim <- vector(mode="numeric", length=reps)
for (i in 1:reps){
  z <- rnorm(n=n, mean=1, sd=1)
  v <- rnorm(n=n, mean=2, sd=1)
  y = 1 + 2 * x + 3 * z - 2 * v
  a1_sim[i] <- coef(lm(y~x))["x"]
}
hist(a1_sim, breaks=100, main="", xlab="Estimates of a1", col=cblue)
abline(v=2,col="red")
```



This variation tells us how much uncertainty we have in our estimate. This is a gross over-simplified simulation of course. In a normal situation:

- we don't know that a_1 is actually 2
- we only have one sample
- we do not know whether the assumed functional form (linear, additive) is correct

We cannot easily incorporate the last point. An often overlooked caveat is that everything we will discuss in a bit (standard errors, p-values, etc.) is conditional on us correctly guessing the overall functional form (But in some situations, there are ways like matching to reduce model dependence).

From simulations to standard errors

If we have only one sample, how can we use this thought experiment? Remember that the root of the idea is to expressing uncertainty about our estimates based on the question: "How would my estimate vary, if I could draw many samples?" Of course, this is only an epistemological crutch. We need to draw inferences on one sample alone. With sufficient assumptions however, we can *estimate* how much our estimates would vary *if we could replicate our sample a million times*. We will discuss this in the next section on regression basics.

Excercises

1. Simulation. Try to set up your own fake data, simulate it, run regressions and see how your coefficient of interest varies. This will not only give you a better feeling for where the variation comes from, but also help you get familiar with R. Set up the simulation so that you can change:
    - The number of unmeasured determinants in the error term.

- The amount of unmeasured error
- The true functional form and the assumed functional form.

2. Please open the data set "Nutrition_Physical_Activity_and_Obesity.csv" and start exploring it. In particular:
   - What data is in it?
   - is this data set in a state that you can do easy inference on?
   - filter and transform the data so as to be useful for analysis
   - explore the noisiness of the data.

References

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation.* Cambridge university press.