STATG006

Introduction to Statistical Data Analysis

Summary Notes 2016/2017

Contents

Chapter 1

Overview of Statistical Data Science

This chapter is intended to present basic concepts of Statistical Data Science, including some preliminary examples of data visualization and analysis. At the end of the chapter, we expect that the student understands the role of Statistics in Data Science, reviewing basic concepts of probability, and getting to appreciate simple examples of statistical modelling. The following notes summarize the chapter presented in the companion slides. Please refer to them for a more complete coverage. The goal of each Summary Note is NOT to repeat or getting into deeper detail on the slides. The slides are always meant to be more comprehensive. Instead, the goal is to just provide a "cheat sheet" of the key ideas scattered through the slides in a more concise way. This first set of Summary Notes is unusually more extended as it contains a review of probability and the corresponding notation that we will make use of in later chapters.

1.1 Key Concepts in Probability

Much in this chapter includes a very brief review of probability. We recommend that each student reads the selected chapters of *STAT1005: Further Probability and Statistics* as available in Moodle. In what follows, we provide the simplest descriptions of the probabilistic concepts mentioned in the companion slides. The goal is mostly to list which concepts have been used and in which context, with more formal definitions relegated to the other learning resources such as slides or companion books/notes. It also establishes some useful notation.

1.1.1 Basics

Statistics can be seen as the study of variability, with probability being the *de facto* language used to represent uncertainty due to randomness, the only game in town for all practical purposes. So before we talk about Statistics, we need to be fluent in its primary language, probability.

The building block of the language of probability is the **random variable**: essentially a quantity that follows a probability distribution. We denote random variables with upper case letters, while lower case symbols are reserved to particular values taken by a random variable. So X might represent a random variable, and x a symbol denoting a particular value that X can take.

Notice that we assume x is a number. Sometimes, for the sake of interpretability and by an abuse of notation, x may informally represent a category. For instance, we may say that a random variable representing gender takes values in the sample space $\{male, female\}$. Formally, though, random variables assume only numerical values, and we may implicitly mean that there is a numerical representation for these values without bothering to mention which (for instance, 0 meaning *male* and 1 meaning *female*).

On the same note, random variables are just one step removed from **events**. To give a cheeky informal definition, events are "things that happen by some random[1] process (which I might not really understand)." For instance, one possible event is "when I toss this coin in my pocket, I will observe tails". We can denote the probability of this event as

$$P(\text{"when I toss this coin in my pocket, I will observe tails"}),$$

which, needless to say, gets really really inconvenient as we describe complicated events. The space of all possible events for a particular phenomenon is called the **sample space**.

Random variables merely provide a way of *encoding* events. So I can speak of the event of my coin showing tails as equivalent of a variable X taking the value of 1. We will use capital "P" to refer to probabilities of events, even those encoded already by random variables. For instance, for the event "$X = 1$", we can speak of $P(X = 1)$.

We now introduce a function to describe how probably a random variable can take values over its sample space (so we do not need to worry about translating events into particular values for random variables all the time). For discrete random variables (where each element of the sample space can be mapped to an unique integer, e.g. 0 for *heads*, 1 for *tails*), we can speak of a **probability mass function** (pmf, for short). So $p(1)$ is a function whose value is the probability of the event "$X = 1$". We often use the generic symbol "p" to denote the pmf of a random variable that is clear from context, although

[1] Notice we do not attempt to define what randomness is. There is a whole philosophical can of worms behind this.

technically we should be less ambiguous and denote that two different distributions should be represented with two different symbols. For instance, if two random variables X and Y follow two different pmfs, we should denote it by using different symbols. Say, $p_X(x)$ and $p_Y(y)$. However, often this just clutters our notation. Implicitly, the variable being alluded to should disambiguate which pmf we are talking about from context, even if we use the same symbols $p(x)$ and $p(y)$.

So if p is a pmf and the **support** of the pmf is $\mathcal{X}$ (the set in which the pmf is positive – say, $\mathcal{X} = \{0,1\}$), we have that p cannot be negative and needs to sum to 1.

$$p(x) > 0, x \in \mathcal{X}$$

$$p(x) = 0, x \notin \mathcal{X}$$

$$\sum_{x\in\mathcal{X}} p(x) = 1$$

For continuous random variables, we run into some technical problems which we will not really get in detail. It suffices to say that if X is continuous, then it is impossible to have both $P(X = x) > 0$ for a non-trivial sample space and $\int P(X = x)\ dx < \infty$, which is kind of a problem if we want our probabilities to add up to 1. So in this case $P(X = x) = 0$ for all x. We get around this problem by specifying the **probability density function** (pdf) of the random variable, which we will typically denote with the same symbol $p(x)$, which should be clear from context. The link between the pdf and probability is given by quantifying the probability of the event "$X \in S$" (for some set S, say, the interval $[1.60, 1.80]$, where X is the height of a person in meters) as

$$P(X \in S) = \int_{x\in S} p(x)\ dx,$$

so if we imagine the graph given by drawing p along all possible values of x, then the corresponding probability will be the area under the curve for that interval. Like the pdf, the pmf has these analogous properties:

$$p(x) > 0, x \in \mathcal{X}$$

$$p(x) = 0, x \notin \mathcal{X}$$

$$\int_{x\in\mathcal{X}} p(x)\ dx = 1$$

One particular useful S in which to assess probabilities, so to speak, is the interval $(-\infty, x]$. In this case, $P(X \in S) = P(X \leqslant x)$. Often we represent this by $F(x)$. That is, we will often use the symbol $F(x)$ to denote a function of x such that $F(x) = P(X \leqslant x)$. We call this function a **cumulative distribution function** (cdf) (or simply **distribution function**). Notice that $F(-\infty) = 0$ and $F(\infty) = 1$. Also, notice this function has to be *monotonic*: that is, if $x' > x$, then it must be the case that $F(x') \geqslant F(x)$.

One of the reasons why this is a nice concept is because it is well-defined for both discrete and continuous variables. A discrete variable, for instance, has a cdf but not a pdf, because the pdf assumes that the corresponding cdf is differentiable. That is, for continuous variables,

$$p(x) = \frac{dF(x)}{dx}.$$

1.1.2 Multivariate Models

In modern Data Science, it is rarely the case we will be interested in isolated random variables. The models we will tackle are often **multivariate models**, meaning probabilities for sets of variables instead of single variables. If X and Y are two random variables, then we can for instance have a **joint (bivariate) pdf** $p(x, y)$. This is again important as we typically encode events with sets of random variables, instead of a single random variable with a very complicated encoding (which is often not even possible). So the event "my height is 1.90m and my weight is 85kg" is more naturally represented by having a random variable quantifying the random outcomes that resulted in my height, and another random variable quantifying the random outcomes that resulted in my weight.

These variables, however, are not usually **independent**: knowing somebody's height will allow you better educated guesses about that person's weight when compared to the situation where you know nothing about that particular person. Using bivariate pdfs for the following example (for pmfs, it is all analogous with sums instead of integrals[2]) we define that two random variables X_1 and X_2 with joint pdf $p(x_1, x_2)$ are independent if and only if we can somehow factorize their joint, that is, we can write $p(x_1, x_2)$ as

$$p(x_1, x_2) = h(x_1)g(x_2),$$

where $h(x_1)$ is a function of x_1 only (x_2 does not appear in its definition) and $g(x_2)$ is a function of x_2 only (x_1 does not appear in its definition).

It is not hard to show that if X_1 and X_2 are independent, then the factors you see above are actually

$$p(x_1, x_2) = p(x_1)p(x_2),$$

[2] Sets of discrete variables have joint pmfs. It is of course true that in many problems we have both discrete and continuous variables whose joint random behaviour needs to be described. In this case, it seems we are in a pickle because there is neither a joint pmf nor a joint pdf. However, the **joint cdf** still exists as we can still speak of (say) $P(X_1 \leqslant x_1, \dots, X_p \leqslant x_p)$ for any set of possibly dependent random variables $X_1, \dots, X_p$, regardless which one of these are discrete or continuous. Partial differentiation with respect to any continuous x_i will give a valid function, although its interpretation is not as clear as a joint pdf.

where $p(x_1)$ is the **marginal probability** for X_1 in the model for (X_1, X_2). That is,

$$p(x_1) = \int p(x_1, x_2)\ dx_2.$$

Let us stop for a second to understand how we are overloading symbols here: the symbol "p" used in $p(x_1)$ and $p(x_1, x_2)$ obviously represents two different functions. They do not even share the same arguments! A less ambiguous notation could be

$$p_{12}(x_1, x_2) = p_1(x_1)p_2(x_2)$$

but given the context we will typically ignore this.

Notice that $p(x_1)$ (and $p(x_2)$) are (univariate) pdfs just like we defined before. The link between marginal and joint probabilities is given by the process of **marginalization**. That is, to get the univariate marginal for X_1 in the joint pdf $p(x_1, \dots, x_p)$ we can just integrate away everything but X_1:

$$p(x_1) = \int \dots \int p(x_1, \dots, x_p)\ dx_2 dx_3 \dots dx_p.$$

The **law of total probability** is a somewhat grandiose name for marginalization, although you may find this in textbooks described in terms of events instead of random variables.

Finally, the concept of **conditioning** refers to an operation of *fixing* some random variables at a particular value and calculating how the distribution of the remaining random variables look like. That is, given a joint distribution $p(x_1, x_2)$, we know we can obtain the marginal $p(x_1)$, but how would X_1 behave if we were told that $X_2 = 10$? Conditioning in the bivariate case (referring to a generic value "x_2" instead of any given constant such as 10) is defined by the formula

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)}.$$

The intuition behind it that that we first expand the total probability space (which initially adds up to 1) by a factor of $1/p(x_2)$: a **renormalization** operation. Then we imagine "$p(x_1, x_2)$" as just a function of x_1, as x_2 is a constant now. For instance, if x_2 is weight of a person that we happen to discover, then we look at the slice of the population of that particular weight and what the distribution of heights x_1 is within that slice. But the probabilities within that slice should be renormalized, as we already excluded everybody with a different weight other than x_2.

The concept of conditioning can be defined analogous for pairs of sets of variables, instead of pairs of singleton variables. Moreover, conditioning and marginalization can be combined to find particular conditional pmfs/pdfs starting from a larger pmf/pdf. For instance, to find

$$p(x_1, x_3 \mid x_5, x_6, x_7)$$

starting from a joint over $X_1, \ldots, X_p$, we do marginalization followed by conditioning. That is,

$$p(x_1, x_3, x_5, x_6, x_7) = \int \ldots \int p(x_1, \ldots, x_p) \ dx_2 dx_4 dx_8 dx_9 \ldots x_p.$$

and

$$p(x_5, x_6, x_7) = \int \int p(x_1, x_3, x_5, x_6, x_7) \ dx_1 dx_3,$$

followed by

$$p(x_1, x_3 \mid x_5, x_6, x_7) = \frac{p(x_1, x_3, x_5, x_6, x_7)}{p(x_5, x_6, x_7)}.$$

Finally, the concept of conditioning gives an alternative interpretation of independence, which is perhaps more intuitive. If X_1 and X_2 are independent, then

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)} = \frac{p(x_1)p(x_2)}{p(x_2)} = p(x_1)$$

That is, being told $X_2 = x_2$ brought no information about X_1, so its distribution remains unchanged.

1.1.3 Summarizing Distributions

A pmf/pdf is a whole function, and sometimes we need to summarize them in some useful way. For instance, we may want to find its "centre of mass", or its most probable value, or the probability of a random variable exceeding some particular threshold. By a way of example, let us consider two summaries: the **mean** and and **variance** of a distribution. Both use the concept of **expectation**. If we have a particular transformation of a random variable, say $g(X)$, than that itself is a random variable with a particular distribution. We defined the **expected value** of $g(X)$, denoted as $E[g(X)]$ as the "centre of mass" of that distribution. That is,

$$E[g(X)] = \int g(x)p(x) \ dx.$$

The mean of a random variable is what we get then $g(X) = X$. That is, the mean of X is given by

$$E[X] = \int xp(x) \ dx.$$

We often use the symbol μ to denote a particular expectation, which should be clear from context. So, whenever you see μ, you should know that it should denote $E[X]$ for some random variable clear from context, unless specified otherwise.

The variance of a random variable X, which we will denote as $Var(X)$, is just the expected value of a particular "measure of spread" around its mean. This measure of spread is defined[3] as $(X - E[X])^2$ (or, to use our convention above, $(X - \mu)^2$). That is,

$$Var(X) = E[(X - \mu)^2] = \int (X - \mu)^2 p(x)\ dx = E[X^2] - \mu^2.$$

One particular type of expectation is the **moment** of order d. This merely means $E[X^d]$. So we can see the variance is a function of the first (μ, or $E[X]$) moment and the second moment.

Another useful summary is the **quantile**, the value in the support of a distribution that corresponds to a particular value of the cdf. That is, the inverse of the cdf. The 50% quantile for instance (also called the **median**) is the point "halfway" through the cdf. That is, x such that $F(x) = 0.5$. So for a given probability q in $[0, 1]$, $F^{-1}(q)$ is the quantile corresponding to that probability. The median, therefore, can be written as $F^{-1}(0.5)$. Can you think of situations where the median is a more appropriate summary of a distribution than a mean?

1.1.4 Notes

The above is nothing but a modest summary of basic concepts in probability. I will assume that you will either know this material already, or will put work on learning it by studying the corresponding *STAT1005* notes provided. Some specific topics were not discussed here and will not be strictly required for *STATG006* (such as transformation of random variables), but may be useful in other modules. Books like Rice and Wasserman are also useful resources if you want to get a deeper understanding. Notice that in the summary above I do not mention a single concrete distribution like the Gaussian or the Poisson. This is because I want you to focus on the fundamental concepts. Specific models are discussed within appropriate contexts in our slides and the reading list provided. One common notation to denote that a random variable X follows a particular distribution (say, "D") with a particular set of parameters (say, "θ") is the following:

$$X \sim D(\theta).$$

It is common to describe data as a collection of random variables. So if we have (say) 100 data points, we may distinguish among them by using superscripts. So a dataset could be described as a collection $X^{(1)}, \dots, X^{(100)}$.

Datasets are typically collections of **independent, and identically distributed** (i.i.d) random variables. This means that we can describe the distribution of a generic i-th data point (e.g., the first, second, etc. data point) with the notation

$$X^{(i)} \sim D(\theta),$$

[3] This may look arbitrary, but it is motivated by its interpretation in the Gaussian case.

where, by independence, the pmf/pdf of the joint distribution of a dataset of n points can be written as $p(x^{(1)}, \ldots, x^{(n)}) = \prod_{i=1}^{n} p(x^{(i)})$.

Notice also that I would never require you to memorize the formula of any specific pdf or pmf.

1.2 Key Concepts in Statistics

A probabilistic model describes a way by which data can be potentially generated. **Statistical inference** is the inverse process: from the data, it provides estimates of properties of the process that generated the data. The process is unknown. All we see is data. All we can provide are estimates.

Statistical inference assesses uncertainty about estimates in two main ways: the **frequentist** and **Bayesian** ways. To keep our focus, we will have little to say about Bayesian inference: if you are interested, *STATG004* will go into detail in that way of performing statistical analysis. This is not to say that Bayesian inference is any way less important, but for a variety of cultural and historical reasons it is true that the frequentist approach is more common, and at least as important in practice. Hence, for now on we will focus on the frequentist approach and we will not further discuss the difference, which in practice can be small. In a nutshell, the goal is to provide an understanding of what happens when we consider the long-run performance of our procedures: what if the data had been different from the one we actually had observed? How would my conclusions change? Probability is understood as the limit of a frequency as we get more and more data. This is of course an idealization, as there is no such a thing as infinite data!

1.2.1 Basics

A **sample** is the same as a **data set** (or "**dataset**", which is other common spelling), a collection of records made available as data. In our example with the NHANES data, there is a group of people which we collected data from (i.e., **in-sample** people) and which we did not collect data from (i.e., **out-of-sample** people). All of these recorded and potential people form a **population**. In statistical inference, we want to characterize the population.

To be clear, the name "population" is used to describe the space of data points regardless whether these are people or not. So we may have, for instance, the space of book sale volumes from a particular vendor, the space of cells which react to a particular cancer treatment, and so on.

We can summarize data with **statistics**. A statistic is just a function of the data, that is, any summaries of your sample. It might sound dull, but it emphasizes an important distinction: statistics are based on quantities you directly observe. If a number depends on unknown quantities, then it is NOT a statistic.

1.2.2 Estimation

Models have **free parameters** that can be tweaked to better represent the data. For instance, a model for data based on the Gaussian distribution will have a parameter describing its mean and another describing its variance. We can set the values of the mean parameter and variance parameter to "best" describe the data. There are a variety of ways of quantifying what a good description is, as we will see in later chapters. Sometimes this process of **fitting** the data boils down to solving a straightforward formula, sometimes it is a computationally intensive algorithm that *optimises* a particular objective function.

The result are **estimates** of the "true" unknown parameters. Notice all models are approximations. When we say we can learn the "true" parameters by gathering more and more data, we implicitly assume we are recovering the "best" approximation to it within a family of models, even though this is not spelled out.

One main way of assessing how well a model fits the data is by comparing the fitted model against the **empirical distribution**, the distribution defined by the observed data. Say we have a dataset with observed values $\{x^{(1)}, \dots, x^{(n)}\}$. Its **empirical cdf** is given by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x^{(i)} \leqslant x),$$

where $I(v)$ is 1 if "v" is true, and 0 otherwise. We can compare how the empirical cdf matches the cdf of a particular model. So if we fit a Gaussian model to the data, we can evaluate its cdf at each data point $x^{(i)}$ and contrast it to $F_n(x^{(i)})$. A visual comparison can be constructed, resulting on what is called a **quantile-quantile plot**.

More quantitative and formal ways of performing this comparison will be the main subject of study of Chapter 2. We will see that probabilities take a dual role in Statistics: both as a way of **encoding models** of reality, and as a way of **assessing** how the resulting models explain the data. **Prediction** is one way of assessing models: that is, how out-of-sample records can be predicted from the in-sample ones. Machine Learning, for example, puts a lot of emphasis on prediction. But even prediction can only be estimated, because in practice we have only finitely many data points at any stage where we can test our predictions. So probability is still used to quantify our uncertainty about *estimated* prediction abilities.

1.2.3 Distributions Mentioned in This Chapter

- the Gaussian distribution (also called the Normal);

- the Uniform distribution, in particular the Uniform distribution in the $[0, 1]$ interval. Can you write what the pdf of the Uniform in $[0, 2]$ would look like?

Chapter 2

Statistical Assessment: Hypothesis Testing and Confidence Intervals

This chapter introduces two important ways of assessing assumptions and estimates. Hypothesis testing is primarily used to answer yes/no questions in a statistical way, while confidence intervals provide a range of estimates that in some probabilistic sense are not distinguishable by the data. Both methods are relevant to statistical inference, but can also be easily misused.

2.1 Key Concept in Frequentist Inference

A dataset $\mathbf{X} = \{X^{(1)}, \dots, X^{(n)}\}$ is considered to be a random object. *This can be confusing. Take some time to let it sink in.* But the idea is that, even though you might have right in front of you a set of observed data points $\mathbf{x} = \{x^{(1)}, \dots, x^{(n)}\}$, you might be still concerned about *data that you have not seen*[1]. These hypothetical unseen data points were not in your sample due to all sorts of reasons (for instance, people who refused to respond to a questionnaire; or people who have not been born yet!). However, if we intend to make claims about the population, we should taken into account how "lucky" we were to have observed this particular dataset $\mathbf{x}$.

So any observed statistic s that is a function of $\mathbf{x}$ (say, sample average) is actually a **realization** of a random variable S which is a function of random data $\mathbf{X}$. If s can be used to disprove a theory ("is the Higgs boson there?", "is New Drug better than

[1] Again, there are all sorts of philosophical issues on whether this is the best way of reasoning, which we will avoid discussing at length here. The major alternative is Bayesian inference, but there are other less well known frameworks.

Old Drug?") we have to take into account how S varies given the distribution among hypothetical datasets.

2.2 A Walkthrough of a Hypothesis Testing Example

We have a class of 40 people, 15 of which female. Is there evidence of gender imbalance from this data? One way of starting this analysis is by stating **which hypothesis we want to show**, which we will denote as H_0. We will define it as $\theta = 0.5$, where θ is the probability of a particular student being female. We will define as an **alternative** that $\theta < 0.5$. That is, we will exclude the possibility of $\theta > 0.5$. This will simplify the discussion of what follows.

It is common to define the problem using this notation:

$$\begin{aligned} H_0 &: \quad \theta = 0.5 \\ H_1 &: \quad \theta < 0.5 \end{aligned}$$

Now, you are the data scientist in charge of assessing the veracity of H_0. Maybe you think that you should assess $P(H_0|\ Data)$, whatever the data is (15 females out of 40 students, in this case). That is, what is the probability of H_0 being true given the data we have seen? **But this requires defining what we mean by "probability of H_0 being true".** This is not necessarily a difficult thing, although in many cases its meaning is not totally clear. Maybe it is could be estimated from past classrooms. Maybe it could be a subjective assessment (explored in more detail in *STATG004*). But whatever it is, quantifying $P(H_0 \mid Data)$ implies H_0 is a random variable. In the traditional framework of hypothesis testing, we will avoid turning H_0 into a random variable: for instance, not everybody feels comfortable in doing a probabilistic assessment of whether two particular drugs are equally good. It is assumed that they either are or they are not, and there might be no clear distribution concerning this claim to justify specifying some $P(H_0)$.

So we need to rethink how to assess this. You come to the conclusion that maybe we can just assess some feature of the data that would be likely to happen if H_0 was true. If this feature turns out to be unlikely, we would reject H_0. The last step requires some judgment: how do we translate "unlikely" outcome to rejection? We will get back to it later on.

Let us decide what could falsify H_0. If we have $Y^{(1)}, Y^{(2)}, \dots, Y^{(40)}$ as our data, where $Y^{(i)} = 1$ if student i is female, 0 otherwise, we have that the statistic

$$X \equiv \sum_{i=1}^{40} Y^{(i)}$$

is a count of how many female students I have. So if all $Y^{(i)}$ are independent Bernoulli (binary) random variables with probability θ of being equal to 1, then X is a binomial random variable with parameters $n = 40$ and θ. In our observed data, $X = 15$.

Now the key point: intuitively we can see that high values of X are at least as likely under $\theta = 0.5$ as they are compared to the alternatives[2]. Conversely, we may then postulate that the event $X \leqslant x$ would not be that likely under H_0 if we make x "small enough". Is 15 small enough? Let us assess it.

We most certainly can compute $P(X \leqslant 15)$ under the assumption that H_0 is true: in this case, $\theta = 0.5$ and $X \sim Binomial(40, 0.5)$. There is no dispute that $P(X \leqslant 15) \approx 0.07$ in this case. Sometimes we can denote is as $P(X \leqslant 15; H_0)$ to make clear what the distribution of X is. This quantity is called the **p-value**.

Now comes the hard part: what is my decision? Do I reject or not H_0? Maybe I decide that 0.07 is not that unlikely. In this case, I "accept" that H_0 is true[3]. Now if this feels somewhat unsatisfying, it is because this does not clarify why 0.07 seems "not strong enough" evidence to reject H_0. You are not alone on thinking this. In practice, we resort to calibrate our decision rule in terms of the **long run frequency of errors that we can make**.

2.2.1 Type I errors

Now, let's apply the frequentist principles mentioned at the beginning. Our data $\{Y^{(1)}, \dots, Y^{(40)}\}$ has been observed, but we may think about other datasets we have not seen but in principle we could have. In our case, $X = 15$, but what could we say about the distribution of our outcomes? The p-value *is* a function of the data, after all. In our example, it is literally given by

$$\sum_{i=0}^{15} \binom{40}{i} 0.5^i (1 - 0.5)^{(40-i)}.$$

If we want assess the long-run performance of our decisions, we need to characterise what the distribution of the p-value is under the assumption H_0 is true ("under the null", as we sometimes we like to say). This is easier to see if we look at the equation above as a function of the statistic we used, X. So, think of the "p-value function", for instance[4]:

$$p_v(X) = \sum_{i=0}^{X} \binom{40}{i} 0.5^i (1 - 0.5)^{(40-i)}$$

[2] Try it with a couple of examples. What is higher: $P(X > 15)$ for $\theta = 0.5$ or for $\theta = 0.1$? You can calculate it in R using `pbinom(15, 40, 0.5, lower.tail = FALSE)` and `pbinom(15, 40, 0.1, lower.tail = FALSE)`.

[3] Formally speaking, the orthodox interpretation is that we "failed to reject" H_0, meaning that we do not have evidence to falsify it. In practice, if you see yourself at the situation that you either have to decide H_0 is true or H_1 is true, it is agreeable to claim we are "accepting" that H_0 is true. Implicit is the fact that all knowledge may be provisional, and this is our conclusion to the best of the evidence we have seen.

[4] I use "p_v" here instead of "p" to avoid confusion with the notation for probability mass functions/density functions

Now, it is no coincidence that this is also a cumulative distribution function. That is, if $X \sim Binomial(40, 0.5)$, the above can also be written as

$$P(X \leqslant x) = F(x) = \sum_{i=0}^{x} \binom{40}{i} 0.5^i (1 - 0.5)^{(40-i)}.$$

for any fixed x. So the "p-value as a random variable" simply means the random variable that comes from a particular transformation of X: $F(X)$.

Still with me? We can show that $F(X)$ a.k.a. $p_v(X)$ follows an uniform distribution between 0 and 1. The proof of this statement is not really important, only the fact that **we know what the distribution of the p-value is, under the null**. Why is this important? Because if we want to make a decision based on "We reject H_0 if and only if $p_v \leqslant t$", for some threshold t, **we can then calculate the probability of erroneously rejecting H_0**. This error is called the **Type I error**, and we can calculate it as follows.

The probability of rejecting H_0 when H_0 is true is the probability $P(p_v(X) \leqslant t; H_0)$. But since $p_v(X) \sim Uniform(0, 1)$ under the null, then $P(p_v(X) \leqslant t; H_0) = t$. So,

- if we decide that we reject the null hypothesis if the p-value is less or equal to 0.05, then our probability of Type I error is 0.05;
- if we decide that we reject the null hypothesis if the p-value is less or equal to 0.07, then our probability of Type I error is 0.07;
- and so on;
- **this probability is the long-run frequency of Type I error** of applying the decision rule "p-value less than or equal to t" to many many datasets.

The upshot? Now we have a less arbitrary justification: we choose our decision rule based on what we think is an acceptable Type I error. Of course, there is still some subjectivity on why a particular threshold t would correspond to an "acceptable Type I error." Many scientific communities for instance regard "0.051" as not **statistically significant** evidence to reject H_0, while 0.05 would be **statistically significant**. This is ridiculous of course, but deciding on statistical significance based on "acceptable Type I error" can only remove so much subjectivity. Your role as a data scientist is to be able to interpret a p-value and to communicate to others what it says about the evidence against H_0. The main message is that your choice of threshold will dictate your Type I error. To come up with a decision threshold will require some domain knowledge of what is reasonable what is not, but the message you want to convey and which you might need to explain to lay people is that this is chosen by the acceptable frequency of making Type I errors in the long run.

2.2.2 Type II errors

What about the opposite mistake? Accepting H_0 when it is false? This is known as a **Type II error**, and we would like to characterise its distribution. But now things get considerably more complicated, as the distribution of the p-value will depend on what the true state of nature is, and in our framework there is no such a thing as one state of nature being more "probable" than another (that is, we have no definition for what a pdf "$p(\theta)$" would mean. θ is just a fixed but unknown state of nature, not a random variable). In our example, what we can do is to characterise what happens for a series of values of θ within the alternative possibilities $\theta < 0.5$ ($\theta \geqslant 0$ is assumed implicitly).

In our example, what happens if $\theta = 0.1$? We have to re-express what our decision rule is in terms of our data X. Recall that the decision rule is based on assuming H_0 is true. Say we set $t = 0.05$. We have to find which value of x is such that $P(X \leqslant x; H_0) = 0.05$, that is, the 0.05 quantile of the distribution of X under the null. Using function `qbinom` from R, for instance, we get[5] 15. So our decision rule, set in stone according to our choice of test statistic X and level 0.05, is "reject H_0 if and only if $X \leqslant 15$". That is, values $X = \{0, 1, 2, \ldots, 15\}$ are the values leading to the rejection of H_0. This set of values is also known as the **critical region** of the test.

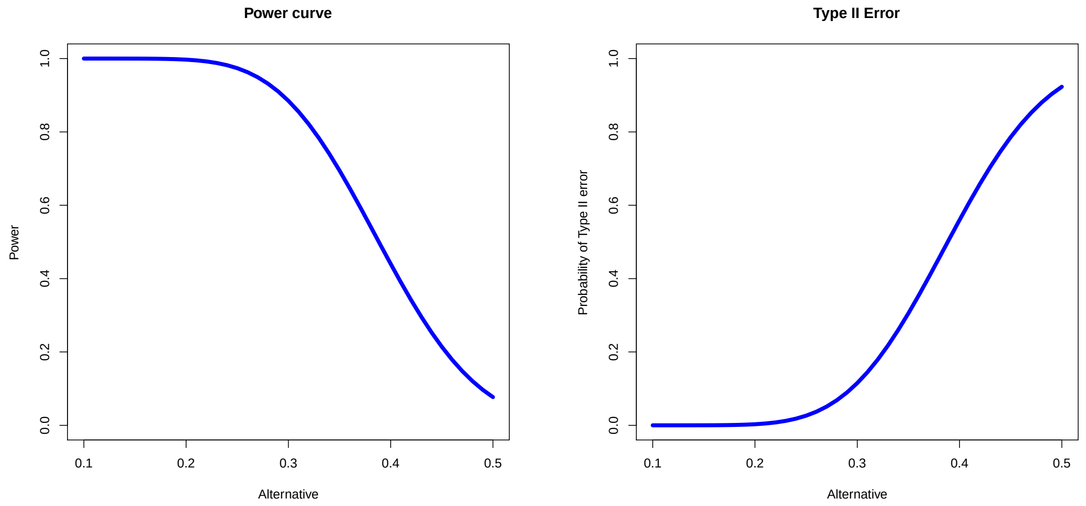
So, we shall ask again, what happens if $\theta = 0.1$? We can derive what happens to the distribution of X: equivalently to our original reasoning, $X \sim \mathit{Binomial}(50, 0.1)$. So, what is the probability of rejecting H_0 in this case? It is $P(X \leqslant 15; \theta = 0.1) \approx 1$. This probability, of correctly rejecting H_0 at the alternative 0.1 is also called the **power** of the test at $\theta = 0.1$. The Type II error probability at 0.1 is just one minus the power. That is, using β to represent Type II error, $\beta(0.1) = 1 - \mathit{power}(0.1) \approx 0$.

We can play this game with other values of the alternative. For instance, what is the power at $\theta = 0.25$? It is $P(X \leqslant 15; \theta = 0.25)$. In R,

```
> pbinom(15, 40, 0.25)
```

and so on. A whole **power curve** can be drawn by combing the $[0, 0.5]$ interval and assessing $P(X \leqslant 15; \theta)$. We get a figure like the one below,

[5] The command is `qbinom(0.05, 40, 0.5)`. You may be scratching your head: just in the initial example we had that $P(X \leqslant 15) \approx 0.07$, now it seems I'm telling you that $P(X \leqslant 15) = 0.05$? I'm not pulling your leg. These are approximations, particularly evident when we have count (integer) data with small numbers. The exact values are not that important although you may want to document how they were computed e.g. which R function was used.



A sensible question is: so what? It looks like we do not have much control here, as the power curve follows immediately from the choice of statistic (X) and level (0.05) in which we fixed our test ("reject when $X \leqslant 15$" follows immediately from these two choice). But it is important to know the power of our test at least for some representative values of the alternative hypothesis, because there are two reasons why we might accept H_0:

1. H_0 is true;

2. H_0 is false, but our power is rubbish;

In our example, if the truth is that $\theta = 0.48$, it looks like that our power is $P(X \leqslant 15; \theta = 0.48) = 0.12$, meaning a probability of Type II error of 0.88 which is rather embarrassing. Pragmatically, maybe it does not matter to distinguish 0.48 from 0.50, if low power at alternatives "not that far from H_0" are not that relevant – but again, this requires some degree of judgment and it is problem dependent.

There are four ways of increasing power:

1. collect more data!

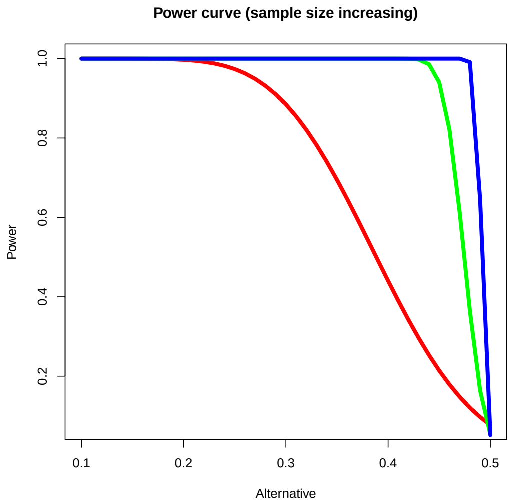
2. allow for a higher Type I error;

3. look for a better test statistic;

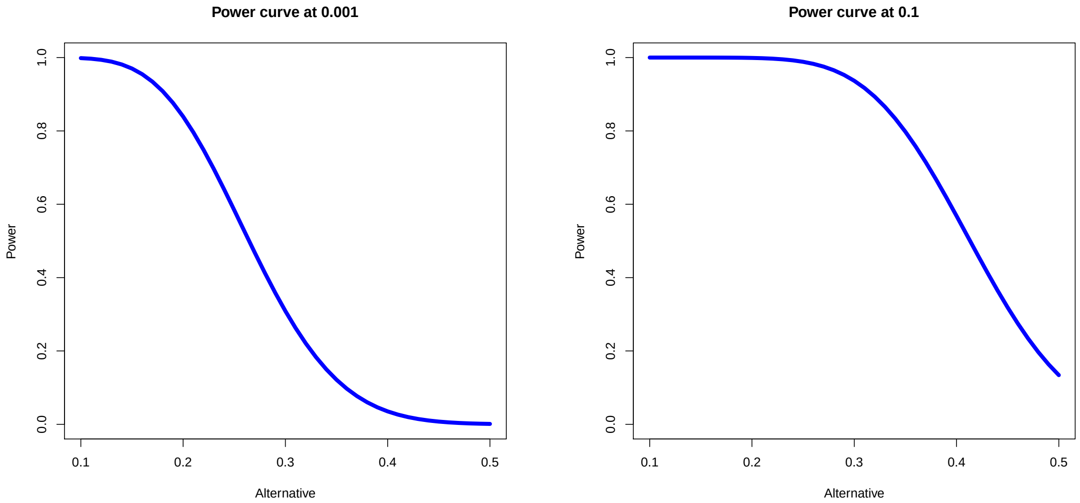
4. make stronger assumptions;

Collecting more data seems an easy thing to say: in our example, we don't really have a choice as the number of students enrolled is a natural phenomenon - you wouldn't control it for the sake of running a hypothesis test! But sometimes it is part of the problem, e.g., of surveying people about their voting intentions: how many people should I interview? By computing power curves using different plausible statistics that you think you may end up observing when the data is actually collected, you can get a sense whether your sample size is large enough. If you want a power of 0.8 to distinguish 0.48 from 0.50 at a Type I error of 0.05, maybe you should restrict your study for classes with more than 40 students, for instance.

For some tests, there will be software that calculates sample sizes for given desirable levels of Type I and Type II errors. Sometimes a numerical algorithm might be necessary, but the main message is that you should understand the logic of power calculations, and why it is useful to understand the power of your method as a function of the sample size. The figure below illustrates our example when as n changes from 40 (red) to 1000 (green) to 10000 (blue):



If sample sizes cannot be chosen, the other possibility is to allow for a higher Type I error. For instance, if we set the level of our test to be some rather stringent such as as 0.001, the critical region will be $X \in \{0, 1\}$. Now the power our test will suffer a big hit! On the other hand, if we set the level to be 0.1 (critical region: $X \leqslant 16$), power will increase at the expense of a Type I error of 0.1 instead of 0.05. The two figures below illustrate these two power curves at 0.001 and 0.1:



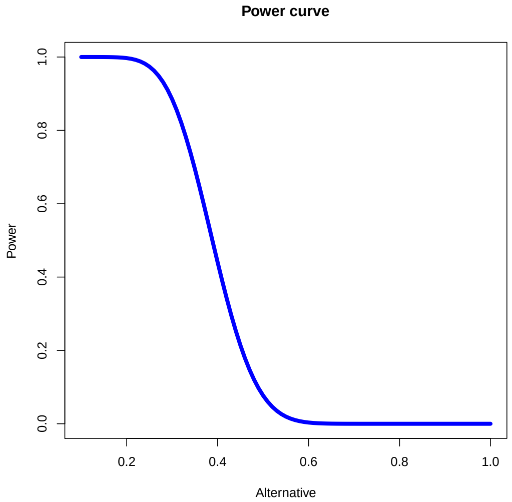
2.2.3 The choice of test statistic, and more on assumptions

Finally, the more complicated change is looking for a test statistic of higher power. This can be fairly complicated, and we will not discuss it in detail. Sometimes (as in our example), the test statistic is quite obvious. One (ridiculous) alternative statistic, for instance, would be using the number of females among the first 20 students that registered. There is no reason whatsoever to choose this, but if you look only in terms of Type I error, this *will* allow you to design a test with the correct Type I property. But as this is equivalent to having half of the sample size, you *will* lose power. So it is good to keep in mind that picking a test with the correct Type I error control may not be enough in practice.

One rule of thumb is: for a fixed Type I error, choose a test which maximises power. In many situations, there is no dominant test: some tests give more power for particular values of the alternative, but not for others. If there is indeed a test which is guaranteed to have power as good as anything else for any value of the alternative, it will be a **uniformly most powerful test**. In some cases, we can find these beasts, but we will not emphasise this in *STATG006*. It suffices to say that many standard tests (like those illustrated in the slides) have good power, and the real things to look for (if you can) is sample size.

Also, some tests will have more power if you make more assumptions. For instance, if our alternative assumption in the example above was $H_1 : \theta \neq 0.5$ instead of $H_1 : \theta < 0.5$, then we would have to consider cases where the proportion of females would be higher than 0.5. We will "lose power" in this case, in the sense that we will have "more" alternatives with "low" power.

Why? Now we have to consider what will happen for $\theta \in [0, 1]$. Let's see how the power curve behaves if we keep evaluating it now over the entire $[0, 1]$ space (we don't care what happens precisely at 0.5):



Of course the curve did not change in the $[0, 0.5]$ interval, but now it highlights that the test is a disaster for anything greater than 0.5. We can choose a different critical region: find two boundaries c_1 and c_2 such that $P(X \in [0, c_1] \cup [c_2, 1]; H_0) = 0.05$. There are infinitely many ways of choosing c_1 and c_2. It is not unreasonable to pick c_1 as the 0.025 quantile of the distribution of X, and c_2 as the 0.975 quantile of the distribution[6] of X. In our example, for $\theta = 0.5$ and $n = 40$, these quantiles are:

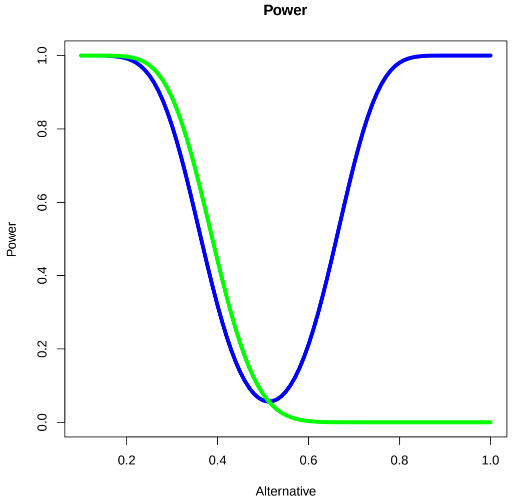
```
> qbinom(p = 0.025, size = 40, prob = 0.5)
[1] 14
> qbinom(p = 0.975, size = 40, prob = 0.5)
[1] 26
```

This is called a **two-tailed test**, to distinguish it from the **one-tailed test** of the original example. With critical region $\{0, 1, \ldots, 14, 26, 27, \ldots, 40\}$ we will get the same Type I error by construction, and the following power curve:

$$power(\theta) = P(X \leqslant 14; \theta) + P(X \geqslant 26; \theta)$$

The figure below shows the power curve for the new test (blue curve) against the original one-tailed test (green curve, which was not meant to cover cases where $\theta > 0.5$). Both have the same Type I error, but we gain power by using our first (one-sided) test if we assume that the alternative is only $\theta < 0.5$.

[6] To see that, remember that for two disjoint events A and B, $P(A \cup B) = P(A) + P(B)$. So $P(X \in [0, c_1] \cup [c_2, 1]; H_0) = P(X \in [0, c_1]; H_0) + P(X \in [c_2, 1]; H_0) = 0.025 + (1 - 0.975) = 0.5$



2.3 Hypothesis Testing: General Concepts and Critical Assessment

Now we go briefly over the most basic concepts in a more formal way, and discuss the shortcomings of this framework for assessing evidence.

In **hypothesis testing**, we are typically contrasting a **null hypothesis** H_0 against some **alternative hypothesis** H_1. In many cases, the alternative hypothesis is something as quaint as "H_0 is false", so sometimes we will not bother to make it explicit (as in most of the slides of Chapter 2). The rule of the game is usually introduced as: do I have enough evidence to falsify ("**reject**" is the statistical jargon) H_0? Sometimes *failing to reject* H_0 is interpreted to be equivalent to *accepting* H_0 *is true*. This is an interpretation that is not universal, or even incentivised: if we cannot falsify H_0, we may *provisionally* assume it is true as we have no evidence to say otherwise. Or we might also come to the conclusion that we still have not enough evidence pro or against H_0. Subjective judgment takes place here.

The elephant in the room is that in most cases the null hypothesis makes a very precise claim, such as "the probability of people living for at least 5 years given cancer treatment A is the same as the probability of people living for at least 5 years given cancer treatment B". It might be a stretch of imagination to believe both probabilities to be exactly the same, *but statements as precise as these are the typical statements tested in hypothesis testing.* One way of conciliating these two disparate thoughts (that we know H_0 is false and yet we test it(!)) is that hypothesis testing can proceed *as if* the real difference is just practically zero if the amount of data we collected cannot distinguish the difference from zero.

This raises another question: do we have enough data? If we do not reject H_0, it may be because i) H_0 is very true indeed or ii) we do not have enough **power**. That is, we do not have enough data to falsify, with high probability, the null hypothesis if the null is indeed false (notice that the "high" in "high probability" is also subjective. Ideally, we would like it to be as high as possible – and in some special cases we can find **uniformly most powerful tests**, but this only applies to very special situations). With low power, comes great irresponsibility.

However, this highlights another way one can get useful conclusions out of hypothesis testing: do we have enough data to proceed with complicated models based on H_1? If not, we may proceed by assuming H_0 is true and do the best we can, because otherwise the effort of adopting the more complicated H_1 seems futile (for instance, we might fail to falsify the hypothesis that our data follows the Gaussian distribution – so for the sake of doing the best we can with our data, using complicated non-Gaussian models might be a waste of effort even if we do not believe Gaussianity holds in reality). Hypothesis testing in this case becomes a "test of sample size".

On the other hand, it is true that, in most practical cases, any H_0 will be falsified given large enough sample sizes. This is not the same as saying that it is of practical significance to adopt H_1 instead of H_0: the smallest deviance from H_0 can be detected if power is high, but this deviance might be negligible (in the A vs B treatment comparison, if the difference of efficacy is something of the order of say 10^{-5}, then it is hard to justify the better treatment if it has other downsides such as cost and side effects). In a situation like this, hypothesis testing is useless and we need other ways of assessing how H_0 is failing to be satisfied.

2.3.1 Other Topics on Hypothesis Testing

In the slides, we discuss the difference between **Type I** error and **Type II** error - basically, the two ways of making mistakes (rejecting H_0 when it is true, not rejecting H_0 when it is false). Typically **test statistics** are designed so that we know their distribution under H_0, so we can control Type I error. Knowing power is harder, since power will depend on which way H_0 is false. For a range of possible alternative hypothesis we may be able to find the power of the test at a particular sample size and as such find the minimal sample size we should collect in order to confidently falsify H_0. This is best covered in *STATG002*, although we showed the basics in our example.

Here are some specific tests that we used to illustrate the machinery of hypothesis testing in the slides: the t-test; the Wald test; the chi-squared (χ^2) for **goodness-of-fit**; and paired tests. You can find a few more examples in the *STAT1005* notes. If you are taking *STATG002* or *STATG003* you will come across others too. For further details, see Rice or Wasserman. The main important lesson here is to understand the logic of hypothesis testing:

1. invent a test statistic that can falsify H_0 if it assumes particular values;
2. derive its (approximate) distribution under H_0, so if H_0 is true we can find a region under which H_0 will be rejected with a particular probability (and if this was not complicated enough, this probability – the **p-value** – is itself a random variable since it is a function of our data. So if we had different data, the p-value would be different);
3. find the power of this test, either under all possible alternatives or for a range of useful ones.

Step 2 is sometimes difficult. Step 3 can be even harder. It is not our goal here to ask you to derive these, but to understand through some examples what the logic is. For many common tasks, such as comparing two treatments ("A/B testing" in business jargon), off-the-shelf tests exist.

2.4 Confidence Intervals

Like any statistic, an estimate of a quantity of interest will depend on the data seen. For instance, consider some $\hat{\theta}$ as an estimate of a parameter of the distribution of some random variable X (for instance, the sample average as an estimate of the population mean). This is sometimes called a **point estimate**, as it boils down our guesses about θ to a single point.

But the fact that an estimate is a function of data means, once again, that had the data been different, the estimate would have been different too. The distribution of the data is carried to the distribution of the statistic. Therefore, we may want to summarize what we can tell about θ with more than a point. More precisely, we may want to have some confidence in this summary, using probability as the language to communicate it.

A **confidence interval** provides essentially this: an interval that will include θ with some probability. *IMPORTANTLY, IT IS THE INTERVAL WHICH IS RANDOM, NOT THE PARAMETER. THINK OF THE LOWEST POINT AND THE HIGHEST POINT IN THIS INTERVAL AS THEMSELVES BEING STATISTICS.* So the lower/upper bound in the interval are themselves functions of the data. The parameter of interest (e.g., population mean) is not a function of the data, it is a constant that happens to be of an unknown value.

In the example given in the slides, we introduced the following recipe:

1. Find a statistic whose distribution depends on the parameter of interest;
2. Find a transformation of it which will follow a distribution whose parameters we will know (e.g., standard Gaussian). This transformation can, and typically will, depend on unknown parameters;

3. Use the known distribution to define an interval which will contain the transformation with a pre-defined probability;

4. Re-express this interval so that it can be easily read as an interval for the parameter of interest.

In practice, it might be enough to guarantee that the interval will contain the true parameter with at least probability p. We call this probability the **coverage** of the interval, which will hold *regardless* of the value of the true parameter. The ideal is that, as you present more and more of such intervals, the frequency in which the interval contains the true parameter will converge to p (although you can never say in which situations you failed and in which you succeeded, unless you get further data to complement your inference). In reality, you should not of course believe you will get this exact coverage since you will be relying on approximate models: the goal is to be “less wrong”, not to be perfect.

2.4.1 Approximations and Computationally-Intensive Approaches

While the first step might be simple, the second can typically be complicated. The **Central Limit Theorem**, explained in the slides using the Gaussian distribution to approximate the distribution of averages, can be used to approximate the distribution of many statistics. One problem with this is that even finding the variance of some statistics is not that easy.

An approximation that can be used to calculate variances is the **bootstrap**: its strange name cames from the idea of “pulling yourself by the bootstraps” by using your data as if it was the population, then drawing samples from it to mimic what variability introduced by a sampling mechanism. A computer-intensive algorithm is typically necessary for this.

Interestingly, we can do more than estimate variances and plugging them in the Gaussian approximation. We can use bootstrap samples directly to generate confidence intervals that bypass the Gaussian step, using a point estimate as a point of comparison to derive an interval, obtaining a so-called **pivotal interval**.

With the understanding of the basic tools for statistical assessment, we will see how they can be applied in the context of specific models. Next, linear regression.

Chapter 3

Linear Regression

This chapter discusses what might be the most popular statistical modelling tool of all. Like everything else in modelling, linear regression has strengths and weaknesses, and a good data scientist should understand those. Also, what we saw in Chapter 2 can be applied in the context of regression. In the Supervised Learning course, you will also see more about linear regression, with an emphasis on its use in prediction, which I do not emphasise here.

3.1 Core Idea

You should be aware by now of conditional probability. Linear regression does not necessarily concern itself with conditional probabilities, but more generally with distributions of some outcome variable Y that will depend on the value of some other observed variables[1]. We will call these informative variables **covariates** or sometimes **input variables**. The symbol $\mathbf{x}$ will be used to denote the values taken by (a vector of) covariates.

More precisely, the whole distribution of Y is not the primary object of study in linear regression. Its primary goal is to model how its expectation changes as a function of $\mathbf{x}$, which follows from the following assumption:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where in the regression equation we use lower case x to denote that these are constants, whether or not the outcome of a random process. The coefficients $\{\beta_i\}$ in this equation are parameters that need to be estimated, with coefficient β_0 playing the role of the

[1] Variables which are not necessarily random variables – we could study study how an outcome changes with time, where time is not random, for instance.

intercept of the equation on the right hand side[2]. Moreover, all the randomness in Y comes from the randomness in the zero-mean random variable ϵ, which is sometimes called an **error term** and which is uncorrelated with the covariates. Therefore,

$$E[Y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

In general, it is not necessary to assume a model for ϵ, but in many application it is taken to be such that $\epsilon \sim N(0, \sigma^2)$, where estimating σ^2 may or may not be relevant.

3.2 Fitting

To fit a model that is a regression line, one alternative is just to write a "fitness function" to be optimised. The most common one is the **mean squared error**:

$$MSE(\beta) = \sum_{d=1}^{n} \left(y^{(d)} - \sum_{i=0}^{p} \beta_i x_i^{(d)} \right)^2,$$

for $x_0 \equiv 1$. We can then optimize this function by taking the derivative of it with respect to all coefficient parameters, setting it to zero, and solving the system[3]. This type of estimator is sometimes known as the **least-squares** estimator.

An alternative is to write a **likelihood** function: to specify a full probabilistic model for Y. The likelihood is just the same expression as the probability density of the data. The reason for the name "likelihood" instead of "probability of the data" as that we see the expression as a function of the parameters instead of a function of the data (and, hence, we typically ignore in the likelihood function any factors which are not a function of the parameters).

By far, the most common likelihood used in practice is the Gaussian likelihood, hinted at the previous section. Even though the Gaussian model might look restrictive, it is less of a bad assumption within a regression model, as some of the variability of Y was removed by fixing $\mathbf{x}$. The likelihood is a product of the probability of each outcome variable given its inputs, as the samples $\{y^{(1)}, \ldots, y^{(n)}\}$ as independent (even though they are not i.i.d.). As the product of small numbers goes to zero very fast, it is more common to think in terms of **log-likelihoods**. For the Gaussian case, we have

$$l(\beta) \equiv -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{d=1}^{n} \left(y^{(d)} - \sum_{i=0}^{p} \beta_i x_i^{(d)} \right)$$

[2] It is not uncommon to refer to an artificial x_0 that is equal to 1, so that the equation can be written more simply as $\sum_{i=0}^{p} \beta_i x_i$.

[3] This is not that straightforward if matrix notation is not used, which we omit. You can try it for the simpler case where $p = 1$.

Notice that optimising this function with respect to $\{\beta_0, \dots, \beta_p, \sigma^2\}$ will give the same solution for the coefficients as the least-squares estimator. This is *not* the same as assuming Gaussianity, just as it would be somewhat silly to say that using the empirical average (the solution to $\sum_{d=1}^{n}(y^{(d)} - \mu)^2$) implies a Gaussianity assumption.

3.3 Diagnostics

Although under mild conditions we do not need to make assumptions about Gaussianity in order to fit a linear regression model[4], problems with major deviances from Gaussianity might require a large sample size to give reliable estimates – in practice, plotting the **residuals** of the regression can be very informative about the adequacy of the model, the residuals being just the difference between the observed outcomes and the fitted expectations. Things to look for are:

- **heteroskedasticity**: as we plot the residuals against the fitted outcomes, do we see their spread around zero in a homogeneous way? If not, this is what we call heteroskedasticity (as opposed to **homoskedasticity**). The reason why this violates the regression model is the fact that the model postulates i.i.d error terms;
- **outliers**: points which in the outcome space are far from the bulk of the outcome variables. This is evidence of non-linearity, measurement error, or high variance of the error terms incompatible with Gaussian errors;
- **leverage points**: points which in the input space are far from the bulk of the covariates (not easy to visualise without special transformations). Again, this is not necessarily a problem, but they provide some evidence of possible instability of the estimates, had your data been different.

Most software packages will provide diagnostic plots based on residuals. Definitely something to be used in practice.

None of these diagnostics provide solid evidence about the interpretation of the covariates as causes of the outcome. Assumptions for causal inference are a whole different can of worms, which are better covered in *STATG002*.

3.4 Inference

As the outcome of linear regression analysis is an estimate, we are interested also in understanding its statistical properties. Slides and textbooks provide details, but the

[4] Least-squares is **consistent** under the assumption of linearity sans Gaussianity, under some "reasonable" conditions about the error term ϵ. That is, $\hat{\beta}$ "converges" to β, in a sense we will not make precise here, but which can be found in books like Rice or Wasserman.

main lessons here are:

- **we may want to test whether some coefficients are zero**, even if believe that a "perfectly zero" coefficient is non-sense. Sometimes measuring the covariates can cost both time and money, or they make the model unnecessarily difficult to understand. So assessing if we have no evidence for the contribution of a covariate is also a useful piece of information.

- **we may want to assess confidence intervals for coefficients**, as one of the main advertising features of linear regression is still relative simplicity of interpretation. Remember: confidence intervals is about what we can honestly report as estimates. If a large range of coefficient values is compatible with the data, we should make this clear.

3.5 Other Issues

The most obvious shortcoming of linear regression is (unsurprisingly) the linearity assumption. This is not necessarily a fatal flaw: linearity is an assumption with respect to the way we represent our covariates. We may want to make some non-linear transformations of the input which might lead to a good linear fit after all For instance, sometimes pairwise products (also known as pairwise **interactions**) of covariates (e.g., x_1x_2) can improve much of the fit while retaining interpretability. If we have many covariates to begin with, it might actually be wishful thinking to fit a complex non-linear model to it, as we shall see in Chapter 5.

Another shortcoming is the error additivity assumption, and the possible Gaussianity that goes with it. We will see in Chapter 4 how to go beyond the Gaussian model to represent other important distributions in practice.