Basics of Probability

Very short introduction

These notes are a summary version of many
Tanvir!?

Contents

4 | Estimation

So far we have been talking mostly about *probability theory*, in particular different objects like *probability space, random variables, distributions, types of distributions, moments and so on....* But we will now start with a different journey, and this will take us to actual realm of *statistics* or more properly we can say *statistical inference*. In many ways these two fields are related but soon you will realize the goals are very different. While the objective of probability theory is to describe and investigate mathematical models of random phenomena, primarily from a theoretical point of view, the goal of statistical inference however is, to propose methods and principles so that from a real world data we can learn about some aspects of the random phenomena that generated the data. So loosely, if probability explains *what is a random variable*, *how we can think about a random variable and "ideally" what data sets one might have from some random variables with the way we have defined*, in statistical inference we come to real world, *we collect the data and model some unknown but useful quantities and then try to guess it with the data*. Since a picture is worth of thousand words, you can try to think the interplay between these two areas as following,

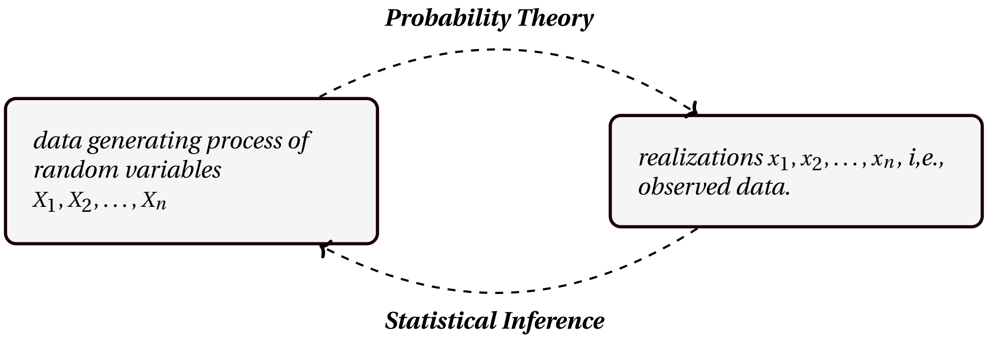
Probability Theory
data generating process of random variables $X_1, X_2, \ldots, X_n$
realizations $x_1, x_2, \ldots, x_n$, i,e., observed data.
Statistical Inference

Figure 4.1: Interplay between Probability theory and Statistical Inference

4.1 Three Approaches: Two Examples

So in statistical inference we are primarily interested to answer following question-

> *Given a data set how can one uncover its underlying probability distribution from a number of random observations?*

Maybe the question is not that clear now, so let us embark on this journey with concrete examples. It turns out there are different ways how we can approach to this problem. In the following first we will give two broad examples that will essentially summarize, what we can say perhaps, a standard way to solve a statistical problem. We will mention three different approaches in statistics, or in statistical inference and then after that we will go into more details.

♣ **Example 4.1** (Rotten Apples).

Suppose an apple seller just received a delivery of $N = 10,000$ apples. It is very natural for him to find out how many of these N apples are already bad. Let us denote the *total number of rotten apples by r*. Then the problem is, the seller doesn't know the actual r, and his quest is finding this r. What he could do is, he could go to each and every apple one by one and check whether each one is good or bad, right?. But the apple seller is a busy man and he realized this is madness! So he came up with a clever idea, that we now call *sampling*. So he collects a sample of $n = 50$ apples. Now assume out of this n, the total number of rotten apple is $x = 25$. So x is simply rotten number in the sample The question is *can the seller say something about r from the rotten number x that he could observe from sample?* We will say, YES! he can, let us explain how,

"$\approx$" means approximately equal

§. Approach 1: Simple ratio calculation

Note we did not explain how the seller does his sampling. In general sampling can be done in different ways, we will come back to this issue shortly!

The simple solution could be if we think - *the proportion of bad apples in the sample is perhaps close to the proportion of bad apples in the total delivery*, in other words we can think

$$x/n \approx r/N.$$

This lead us to think that perhaps $r \approx N \times x/n$. That is, the number $R(x) := N \times (x/n)$ (or more precisely, the nearest integer of this number) is a natural *estimate* of the actual rotten number r. Note, we do not know the actual number r, so we used this technique what in statistics we call *estimation*. So in principle we have come up with a mapping (or a function) R that assigns to the observed value x to an estimate $R(x)$. In statistics such mapping R is called an *estimator*. Later we will formally define what is an estimator but roughly this is a function of the data.

In particular if a box contains r red balls and $N-r$ blue balls (so total N balls), and we select $n \geq 0$ balls randomly without replacement, then if we let X to denote the number of red balls that are obtained. The distribution of X is said to follow hypergeometric distribution with pmf

$$f(x|r, N-r, n) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

for $\max\{0, n-N\} \leq x \leq \min\{n, N-r\}$*. And we write* $X \sim \mathcal{H}_{n;r,N-r}$*.*

Notice x is a random quantity. Why? If the seller would pick a different sample of the same size n, then it is very likely that the number of rotten apples x in that particular sample would be different than this one. So each sample will give us a different estimate.

§. Approach 2: Calculation with a confidence

The first approach is a bit crude one, it just gives us “one” number, now as we have already stated that *different sample will give us different number*, this should give us a sign that maybe just one number is not so reliable. So here is another option, we might go for a probabilistic answer, so that given an observed value x we do not guess a particular value $R(x)$, rather we try to find an interval $C(x)$ (which is a function of x as well), such that the true value r will be within this interval. So for example the interval could be $C(x) = (x-\delta, x+\delta), \delta > 0)$ then with our estimated interval $C(x)$ we would like to have

$$P(x : r \in C(x)) \approx 1 \tag{4.1}$$

in words, we would like to have the probability of being r in the interval that we have constructed using the sample is close to 1. Note, again, since x is random, it is natural that $C(x)$ will also also random. So if the probability calculation in Eq. (4.1) is a correct one and the interval that we have constructed is not that wide, so this means our estimate is not bad! Based on the description of this problem, we can *assume* that x is a realization of the random variable X which follows hypergeometric distributions $\mathcal{H}_{n;r,N-r}$, then the probability calculation in Eq. (4.1) could be carried out with the pmf of X.But note where N, n are known parameters bur r is an *unknown* parameter. So we do not know the exact probability distribution, but we can at least think about what class/family we might want to use. We will come back to this point again when we discuss *statistical modeling*.

§. Approach 3: Making a decision

Now suppose the seller is not thinking about the exact number or a number with a probability, rather he is thinking about two decisions, that is whether the delivery is good or bad. So he thought *if the rotten number of apples in the total is below* 5%*, then it is fine, but if it is above* 5%*, then it is problematic*, since in our example $N = 10,000$, it means he has to decide between two statements,

- all options for r are fine when $r \in \{0,\dots,500\}$.

- none of the options worth looking when $r \in \{501,\dots,10000\}$.

Note, this way of looking at the problem is slightly different than the first two, essentially what we are doing in this case is *we are looking at all possible values r can take and then divide the range of possible values in two parts*, an acceptable part, and and unacceptable part. In statistics these statements that lead us to two decisions are called "hypotheses" (singular 'hypothesis'). So we can write.

- the 'null hypothesis' $H_0 : r \in \{0,\dots,500\}$.

- and the 'alternative' $H_1 : r \in \{501,\dots,10000\}$

So if the seller fails to reject the null hypothesis then the delivery is good, but if he rejects the null then the delivery is bad. Now the question is how can the seller take the decision based on the sample value x? The answer is *the seller has to also find what we say a* ***decision rule*** *of the following type*,

Note, A hypothesis is a statement about a population parameter. The goal of a hypothesis test is to decide, based on a sample from the population, which of two complementary hypotheses is true

- $x \le c \Rightarrow$ the seller supports the null hypothesis, i.e., 'the delivery is good'.
- $x > c \Rightarrow$ he takes the alternative, that is the delivery is bad.

Well we are not done yet!, how to find c?, We can find c probabilistically using following idea.

- We want to make $P(x : x > c)$ (i.e., the probability of the delivery is bad) is small when $r \le 500$, why? Because if we mistakenly conclude that the delivery of apple is a bad one, where in reality it was good, then the seller just concluded wrong. So we minimize this mistake.

- and simultaneously we want to make $P(x : x > c)$ is large when $r > 500$. This means when actually our decision is correct, that is the delivery is truly bad, we want this probability to be large.

Finding a decision rule of this type is known as *testing*. Clearly we need use the distribution of X, so we can use hypergeometric distribution $\mathcal{H}_{n;r,N-r}(x : x > c)$.

(...end of example!)

In this chapter we will try to explain the first two approaches in the last example. The first approach with just one number as our estimate of the unknown quantity is called *point estimation*, the second one where we come up with an interval is called *interval estimation* and the third one is called *hypothesis testing*.

Let us look at another example, and then we go to some details regarding point estimation

♣ **Example 4.2** (Pipe strengths)**.**

Suppose in a power station we would like to have an idea about the strength of the pipe when the pipe gets heated. So the more strength the pipe has, it will last longer. So in a process where different pipe gets heated, we managed to observe n of such pipes and collected n independent measurements of brittleness in heat. Let us denote this n (random) measurements with $x_1,\dots,x_n$. For the moment we will simply assume each of this measurements are realizations of a random variables which are normally distributed. So $X_1, \dots, X_n$ are the random variables that are mutually independent and follow the same distribution that is normal with *unknown mean* $E(X_i) = m, \forall i$ but with *known* variance. Now the way we would like to model the real brittleness is, we can think of the real brittleness as m. So the real brittleness is an expectation of the random variable, but that is unknown to us. Now, the aim is to determine m from the observed point $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. So here x is a n dimensional vector. Again, a statistician can proceed in one of the following three ways:

Note, we made several assumptions in this example. First, we assumed normality. In practice there could be justifications behind this maybe because of our knowledge and prior experience with similar data sets. But the data possibly could have also come from other distributions. Then our normality assumption is wrong. Also note, we assumed the realizations are independent. This is again assumption, maybe in many cases valid, but many cases maybe not.

§. Approach 1: Point Estimation

The first idea that comes to mind is to use the average of the n independent measurements, so we simply calculate

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$$

Note, again like previous example, $\bar{x}$ is random because the observed data x is random. Different data will give us different $\bar{x}$, so it depends on the data. Sometimes for notational clarity we will write $\bar{x}_n$, but you should ALWAYS remember that this depends on n. Different data with same size of n will give us different $\bar{x}$, and moreover changing n will give us different $\bar{x}$ too. So then it means we can write it as a function M, where for this data set $M(x) := \bar{x}$. The value of the function changes with the data set. Here M is an *estimator*, that is a function of the data. This sample mean $M(x)$ might be a good guess but again as as we have discussed before, this estimate is subject to chance, so we cannot trust it too much.

§. Approach 2: Interval Estimation

Now like the previous example we can determine an interval $C(x)$ depending on x, for instance of the form $C(x) = (M(x) - \varepsilon, M(x) + \varepsilon)$ which contains the true value m with sufficient confidence/probability.

$$P(C(\cdot) \ni m) \geq 1 - \alpha$$

for some (small) $\alpha > 0$ and true m. But note this time because of normality assumption this probability can also be calculated using normal distribution, we will explore this in detail in coming sections.

§. Approach 3: Hypothesis Testing

Finally again we can go for taking decisions. Maybe we would like to decide whether the brittleness of the pipe remains below the threshold m_0. If yes we are fine, if not then we have a problem. So again then in theory this means, we can separate the whole range of possible m values in two parts. Then we need to find a decision rule such that

$$\begin{aligned} M(x) \leq c &\Rightarrow \text{ decision for the null hypothesis } H_0 : m \leq m_0, \\ M(x) > c &\Rightarrow \text{ decision for the alternative } H_1 : m > m_0, \end{aligned}$$

for an appropriate threshold level c. And we should choose c such that

$$\{P(M > c) \text{ for } m \leq m_0\} \text{ is small} \qquad \text{and} \qquad \{P(M > c) \text{ for } m > m_0\} \text{ is large}$$

So if we set a small $\alpha > 0$, then we can find c such that $P(M > c) \leq \alpha$ for $m \leq m_0$. Again, making $P(M > c)$ for $m \leq m_0$ small means, we would like to avoid the situation as much as we can when the pipe is good bur we declared it corrupt, and on the other hand the second condition $P(M > c)$ for $m > m_0$ means, we want to make sure that a bad pipe will be recognized with a probability as large as possible.

(...end of example!)

4.2 General Structure of Data and Modeling

If you notice carefully there is a general patterns in the examples that we just discussed, let us know explore these patters and discuss the methods in detail.

4.2.1 The data: outcome of an experiment

We received a piece of a data, in the first example x (the amount of rotten apples) and in the second example $x = (x_1, x_2, \dots, x_n)$ (n measurements). If you think carefully appearing of one data set can be thought of a realization an experiment, where the experiment is simply *"collecting the data"*. So hypothetically every time we go and collect a data from somewhere we perform an experiment and there is a sample space of this experiment, that is the set of all possible data we might get in theory. We can write all possible data in a set amd we will denote this with $\mathcal{X}$.

♣ **Example 4.3** (Sample spaces in Example 4.1 and Example 4.2)**.** For the first example different trials will give us different x (i.e., different number of rotten apples). So here the set of the possible number of rotten apples is $\{0, \dots, n\}$ and it is the *sample space* of this experiment. So we can write $\mathcal{X} = \{0, \dots, n\}$. In the second example, our experiment has given us a data $x_1, x_2, \dots, x_n$. Again you can think in theory "any" random data $x_1, x_2, \dots, x_n$ might be observed. So there is a set of all possible combinations of $x_1, x_2, \dots, x_n$, i.e., $\{(x_1, x_2, \dots, x_n) : x_1, x_2, \dots, x_n) \in \mathbb{R}^n\}$. In this case we can write $\mathcal{X} = \mathbb{R}^n$. So bottom line, there is a set of all possible outcomes of the data, we denote this with $\mathcal{X}$. Below is a vis

(...end of example!)

If we now think in terms of event spaces (i.e., σ-algebra over $\mathcal{X}$), we can write $(\mathcal{X}, \mathscr{F})$, where $\mathcal{X}$ is the our sample space induced from the experiment and $\mathscr{F}$ is a sigma algebra on it where we can assign probabilities. What about the probability distribution (measure)?

4.2.2 The family of distributions and statistical model: a choice

Note that there is a probability distribution P which describes how the data is distributed, or in other words how the elements in $\mathcal{X}$ are distributed. As we have discussed at the beginning *our goal is to know this distribution of the data or sometimes maybe to know just the parameters which will identify this distribution generated the data*. For example in Example 4.1 if true $r = 350$, at least in theory, we would like to know this value. Then again in Example 4.2, if true $m = 100$, we would like to know about it. Now we explain the systematic process to search for this value.

♣ **Example 4.4** (Probability distributions in Example 4.1 and Example 4.2)**.** Actually we have already taken a step in the example. Recall in the examples we have restricted ourselves to only a certain class or family of distributions. In Example 4.1 we fixed ourselves with class of *hypergeometric* distribution. This means we are only looking at the hypergeometric distributions. We don't know all parameters but at least we have now restricted our search. Then Example 4.2 we have assumed $X_1, X_2, \dots, X_n$ are all distributed with "identical" distribution, that is *normal*. And because we have assumed independence, knowing for just one X_i will give us the joint distribution.

(...end of example!)

This means we are to look for the possible distribution that generated the data, we limit ourselves only within certain family or class of distributions. Sometimes this family of distributions, denoted by $\mathcal{P}$ is called the *model* (we will explain model shortly!). But how do we describe this family of distributions. The assumption is we will characterize each of the distributions with its parameters. So for example we will write $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Where Θ is a parameter space and often the assumption is P_θ is fully characterized by θ. It could be that θ is a vector.

♣ **Example 4.5** (Parameter spaces in Example 4.1 and Example 4.2)**.** We can now look at the parameter space and the family of distributions in our two examples. In Example 4.1 $\Theta = \{1,\dots,10000\}$. Then our family of distributions is $\mathcal{P} = \{\mathcal{H}_r : r \in \Theta\}$, where we omitted other parameters because we already know them. In Example 4.2, maybe we can subset of $\mathbb{R}$ to be Θ and then the family of normal distributions are indexed by the expectations which will take value in this subset of $\mathbb{R}$. Note, technically there are more than one parameters in both of the two classes of distributions, but by assumption we already know other parameters, so we can index the family with just one parameter, that is unknown to us. But in general that the family is indexed by vector of parameters, not scalers only.

(...end of example!)

Now we join the pieces together, these three elements, the set $\mathcal{X}$, a σ-algebra $\mathcal{F}$ on it and a class of probability distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ indexed by parameters, together is called a statistical model.

❖ ***Definition* 4.1** (Statistical Model)**.** A statistical model is a triple $(\mathcal{X}, \mathscr{F}, \mathcal{P})$ consisting of a sample space $\mathcal{X}$, a σ -algebra $\mathscr{F}$ on $\mathcal{X}$, and a class $\mathcal{P} := \{P_\vartheta : \vartheta \in \Theta\}$ of (at least two) probability distributions on $(\mathcal{X}, \mathscr{F})$, which are indexed by an index set Θ.

———❖

Other well known terminology of *statistical model* is *statistical experiment*. Since we have to deal with entire class or many (at least two) different probability distributions, we must indicate the respective probability distribution when talking about them. We will denote the distribution corresponding to any particular parameter value ϑ by P_ϑ. Expectations calculated under the assumption that $X \sim P_\vartheta$ (X maybe a random variable or a random vector) will be written E_ϑ. Distribution functions will be denoted by $F(\cdot|\vartheta)$, probability mass functions and probability density functions by $f(\cdot|\vartheta)$. However, these and other subscripts and arguments might be omitted where no confusion can arise. For this handout we will assume we have a parametric model. This means our parameter space is a subset of any Euclidean space, in that case, we will assume *either* of the following options.

- All of the P_ϑ in the class are continuous with densities $f(x|\vartheta)$.

  Or,

- All of the P_ϑ are discrete with probability mass function $f(x|\vartheta)$, and there exists a set $\{x_1, x_2, \dots\}$ that is independent of ϑ such that $\sum_{i=1}^{\infty} p(x_i, \theta) = 1$ for all ϑ

Note we used the same notation for the density and mass functions, but this should be clear from the context whether the random variables in the discussion are discrete or continuous. Although self-evident, it should be emphasized here that the *first basic task of a statistician is to choose the right model!, or in other words the* *right distribution class*. Because that is all she can choose, other things are given.

§. Independent experiments and iid random variables

Finally we need to mention a special case of the experiment that we have been discussing so far but indirectly, specially in Example 4.2 that is independent experiments. This is possibly one of THE most used assumptions that you will come across, if not the only one. In this case our statistical model becomes easy to handle. Recall, in Example 4.2 or the pipe strength example, we assumed each of the realizations of $x_1, \dots, x_n$ is an independent realizations of the experiment of collecting a data. When representing this with random variables this means $(x_1, \dots, x_n)$ are independent realizations of the independent random variables $X_1, \dots, X_n$. Also we assumed identical distributions for $X_1, \dots, X_n$. When this is the case we say the random variables $X_1, \dots, X_n$ are independent and identically distributed random variables, in short *iid* random variables. And a sample is often called a *random sample*. Note, technically speaking a sample can be random even if we do not assume independent experiments that is why there are also other types of random sampling, e.g., simple random sampling which does not assume independent realizations. But in the literature *random sample* often refers to iid experiments or realizations of iid random variables. When we have iid random variables, the joint distribution factored into marginals. So for the real valued random variables, if we denote the cumulative joint distribution function with $F_{X_1,\dots,X_n|\vartheta}$, then we will have

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n; \vartheta) = \prod_{i=1}^{n} F_{X_i}(x_i; \vartheta), \qquad \forall\, (x_1, \dots, x_n) \in \mathbb{R}^n$$

This means we can also write the joint density $f(x_1, \dots, x_n; \vartheta)$ factored in marginals,

$$f(x_1, \dots, x_n; \vartheta) = \prod_{i=1}^{n} f(x_i; \vartheta)$$

♣ **Example 4.6** (Sample pdf-exponential)**.**

Let $X_1, \dots, X_n$ be a random sample from an $Exp(\beta)$ distribution. Specifically, $X_1, \dots, X_n$ might correspond to the times until failure (measured in years) for n identical circuit boards that are put on test and used until they fail. The joint pdf of the sample is

$$f(x_1, \dots, x_n; \beta) = \prod_{i=1}^{n} f(x_i; \beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-(x_1+\cdots+x_n)/\beta}$$

This pdf can be used to answer questions about the sample. For example, what is the probability that all the boards last more than 2 years? We can compute

$$\begin{aligned}
&P(X_1 > 2, \dots, X_n > 2) \\
&\quad = \int_2^\infty \dots \int_2^\infty \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} dx_1 \cdots dx_n \\
&\quad = e^{-2/\beta} \int_2^\infty \cdots \int_2^\infty \prod_{i=2}^{n} \frac{1}{\beta} e^{-x_i/\beta} dx_2 \cdots dx_n \qquad \text{(integrate out } x_1\text{)} \\
&\quad = \vdots \quad \text{(integrate out the remaining } x_i\text{s successively)} \\
&\quad = \left(e^{-2/\beta}\right)^n \\
&\quad = e^{-2n/\beta}
\end{aligned}$$

If β, the average lifelength of a circuit board, is large relative to n, we see that this probability is near 1. The previous calculation illustrates how the pdf of a random sample defined by Example 4.6 can be used to calculate probabilities about the sample. Realize that the independent and identically distributed property of a random sample can also be used directly in such calculations. For example, the above calculation can be done like this:

$$\begin{aligned} &P\left(X_1 > 2, \ldots, X_n > 2\right) \\ &\quad = P\left(X_1 > 2\right) \cdots P\left(X_n > 2\right) \text{(independence)} \\ &\quad = \left[P\left(X_1 > 2\right)\right]^n \text{(identical distribution)} \\ &\quad = \left(e^{-2/\beta}\right)^n \text{(exponential calculation)} \\ &\quad = e^{-2n/\beta} \end{aligned}$$

(...end of example!)

The sample observed assuming iid experiments, or a sample from an iid random variables can be thought of a sample from an *infinite population*. Think of obtaining the values of $X_1, \ldots, X_n$ sequentially. First, the experiment is performed and $X_1 = x_1$ is observed. Then, the experiment is repeated and $X_2 = x_2$ is observed. The assumption of independence in sampling implies that the probability distribution for X_2 is unaffected by the fact that $X_1 = x_1$ was observed first. "Removing" x_1 from the infinite population does not change the population, so $X_2 = x_2$ is still a random observation from the same population. When we are sampling is from a finite population, this assumption may not be relevant. A finite population is a finite set of numbers, we do not have infinite population any more, so clearly taking one sample out of the population will affect the probability of the second sample being chosen. We avoid the details here.....

Notes on notations: At this point we would like to clear about some notations that we will use time to time. Technically for a sample of n observations writing $x_1, \ldots, x_n$ is appropriate, so that no confusion arises about whether we have one sample or n samples. However this increases some notational burden, becasue everytime then we need to write this whole vector. Throughout the handout we will try to write $x_1, \ldots, x_n$ whenever we mention n realizations, however sometimes we will specify it simply with x with defininig $x := (x_1, \ldots, x_n)$, then x is an n dimensional vector. We will try to be explicit if we write x when we are discussing about a random vector, so that no confusion can arise.

4.2.3 Statistics, Estimators: tools for clever statisticians

Now we have come to crux of this chapter, that is *estimators*. If you have noticed, we started the two examples with some questions, in Example 4.1 the question was finding the number of rotten apples. For this we collected a data (sample) x which is the rotten number, then we devised a function $R(x) := N \times (x/n)$ to estimate the parameter r. This function which takes the data and maps into the space where the possible values of the parameter lie, is called an *estimator*. Similarly in Example 4.2, we wrote $M(x) = \bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$, where $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Here M is the function that takes the data as an input and returns a number which is a possible *candidate* for the true parameter m. As we have already mentioned M is also an *estimator* for the population mean m. An estimator is a special case of a more general function known as *statistic*. Roughly a statistic is a function of the data which returns a possible value in the parameter space. Let us see the formal definitions,

❖ ***Definition* 4.2.** Let $(\mathcal{X}, \mathscr{F}, \mathcal{P})$ be a (parametric) statistical model and $(\Sigma, \mathscr{S})$ be an arbitrary event space, then

(a) Any function $T : \mathcal{X} \to \Sigma$ that is measurable with respect to $\mathscr{F}$ and $\mathscr{S}$ is called a *statistic*.

(b) If $\tau : \Theta \to \Sigma$ be a mapping that assigns to each $\vartheta \in \Theta$ to value $\tau(\vartheta) \in \Sigma$ then an *estimator* of $\tau(\vartheta)$ is a statistic T which is constructed to estimate τ.

(c) The distribution of a statistic T is called the *sampling distribution* of the statistic.

---❖

✎ **Remarks 4.1.**

- The definition may look a bit abstract, but note a statistic is simply a function of the random variable (or variables) that generated the data, so its again a random variable.

- Well! an obvious question is why is the notion of a statistic introduced if it is nothing but a random variable? The reason is that although these two notions are mathematically identical, they are interpreted differently. Intuitively, a random variable describes the uncertainty that we are faced with. By way of contrast, a *statistic* is a mapping *a statistician cleverly constructs* to extract some essential information from the observed data.

- Why is the notion of an *estimator* introduced if it is the same as a statistic? And why does the definition of an estimator not mention τ if its related to T? Again, this is *because of the interpretation*. An estimator is simply a statistic that is specifically *tailored* for the task of estimating a parameter τ. For example, in Example 4.2 we used the sample mean which is the estimator for the population mean. The idea of an estimator does not need any more formalization because it is always the function of the data as statistic. So whenever in this handout we will mention an estimator, it will be clear from the context which parameter we are interested in.

- To represent the randomness of T we will follow usual convention. We will write $T(X)$ or $T(X_1, \ldots, X_n)$. So in our second example, we could write $T(X_1, \ldots, X_n) = \bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$.

- Very important! note, $T(X_1, \ldots, X_n)$ and $T(x_1, \ldots, x_n)$ are both conceptually and notationally different. When we write $T(X_1, \ldots, X_n)$, this means T is a function of the random variables, hence it is also a random variable. But if we write $T(x_1, \ldots, x_n)$ this means we have evaluated this estimator for the particular data $(x_1, \ldots, x_n)$ or a realized value, so this is simply a value (i.e., it is already realized and there is no randomness!). In this case when we have a particular value of any estimator, we call this value *an estimate*. So again, an *estimator* is a random variable, and an *estimate* is a realized value of the estimator.

- Finally, note in the definition we used $\tau(\vartheta)$ to represent the parameter that we are interested in. Often it could happen that $\tau(\vartheta) = \vartheta$, on that case we are simply interested in ϑ. Then if the whole parameter space is Θ, the estimator T should also take value on Θ, so in this case we have $T : \mathcal{X} \to \Theta$.

We end this section with a brief summary of what we have outlined till now, and this gives us more or less a general principle (at least for now) to solve a problem. We call this *thought process of a statistician*. Fig. 4.2 describes this thought process. A statistician starts with a unknown quantity ϑ that she would like to know about using data. So she starts thinking that she could model this unknown quantity as a parameter of a distribution P_ϑ. She has to think critically which model to pick, and this lets her to decide the family of distributions. Well now, she is more or less confident that one of the distribution of this family or class has produced the data set she has in her hand, But the statistician does not know exactly which one because she does not know the parameter ϑ. Then after analyzing the structure of P_ϑ, the statistician devises a function T which is an estimator of the parameter ϑ and starts checking whether this was a good assumption or not.

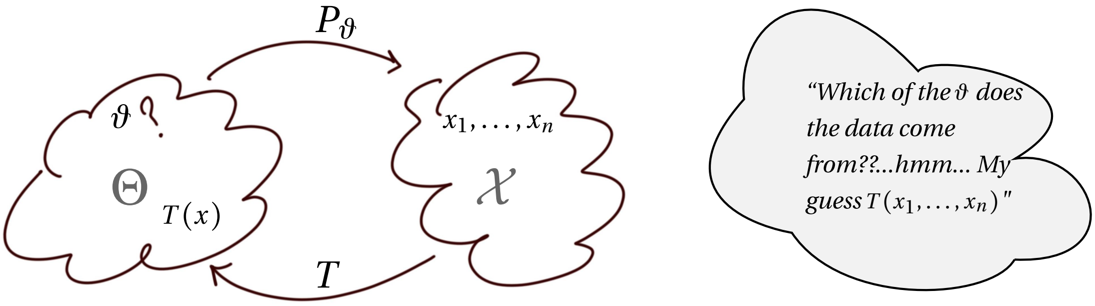
P_ϑ
ϑ ?
Θ
$T(x)$
$x_1, \ldots, x_n$
$\mathcal{X}$
T
"Which of the ϑ does the data come from??...hmm... My guess $T(x_1, \ldots, x_n)$"

Figure 4.2: Thought process of a statistician

Now that we have discussed what is an estimator and how this helps to find the unknown quantity, we will start focusing on following questions.

- What are the possible methods to find an estimator?
- If we find more than one estimator for a parameter how to evaluate them?
- Are there some principles to find an estimator?

In the following sections we will tackle these questions one by one. But before we dive into the details of estimation we will take a short hiatus on some results related to the sample when we have iid random variables and in particular when we have iid observations from a normal distribution. This is important both for practical reasons and also to see some exact sampling distribution of some estimators. After that we will start with *point estimators* and then *interval estimators*.

4.3 Sampling from iid Random Variables

In this short section we collect some results related to *random sampling* (i.e., $x_1, \ldots x_n$ are realizations of iid random variables $X_1, \ldots X_n$). Let us recall the definition of sample mean and variance. So the sample mean from a particular sample is the arithmetic average of the values in that sample, so

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

This is one particular value, or one number and there is no randomness here. But when we treat this arithmetic average as an average of iid random variables, we get a statistic,

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

So $\bar{X}$ is a random quantity which is often used as an estimator of the population mean. And then $\bar{x}$ is *one particular* realized value. Similarly the sample variance is the statistic defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

And finally, the sample standard deviation is the statistic defined by $S = \sqrt{S^2}$.

Since $\bar{X}$ is a random variable, in theory it should have a distribution and there might be a mean and variance. Following theorem is the first result of this kind without assuming any distributions for the random variables $X_1, \ldots, X_n$,

The property that average of the sample mean is the population mean that is written in Theorem 4.1 is known as unbiasedness, we have not formally defined this yet, we will do so in coming sections. But note this is an important property of an estimator.

✣ ***Theorem* 4.1** (Mean and Variance of $\bar{X}$ and S^2).
Let $X_1, \ldots, X_n$ be iid random variables, follow a common distribution with mean μ and variance σ^2, then

(a) $\mathrm{E}(\bar{X}) = \mu$

(b) $\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$

(c) $E(S^2) = \sigma^2$

(b) $\mathrm{Var}(S^2) = \dfrac{1}{n}\left(\mu_4 - \dfrac{n-3}{n-1}\mu_2^2\right)$, where μ_k is called k^{th} central moment, defined as $\mu_k = E\left((X-\mu)^k\right)$, for $k = 1, 2, \ldots$

———✣

Proof. First for (a)

$$\begin{aligned}\mathrm{E}\bar{X} &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\\ &= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right)\\ &= \frac{1}{n} n \mathrm{E} X_i = \mu\end{aligned}$$

Similarly for (b), we have

$$\mathrm{Var}\,\bar{X} = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} n\,\mathrm{Var}\,X_1 = \frac{\sigma^2}{n}$$

For the sample variance, using Theorem 5.2.4, we have

$$\begin{aligned}\mathrm{E}S^2 &= \mathrm{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right)\\ &= \frac{1}{n-1}\left(n\mathrm{E}X_1^2 - n\mathrm{E}\bar{X}^2\right)\\ &= \frac{1}{n-1}\left(n\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) = \sigma^2\end{aligned}$$

■

Note the theorem did not assume any particular distribution for X_is, we assumed only that the random variables are iid and mean and variance exists. Now we extend this result with imposing normality. In this case we will see that we can also derive the the distribution, which is the called the exact sampling distribution of of $\bar{X}$.

4.4 Sampling from iid Normal Random Variables

When we impose normality then naturally we get many nice results related to the sampling distribution of $\bar{X}$ and S^2. But before the results we need to give a short detour about some family of distributions which are derived using Normal distributions.

4.4.1 The Chi-squared distribution

The first one is what we call the family of χ^2 distributions. This is a sub-collection of the family of gamma distributions. So if you have not seen gamma distributions before you can ignore some details that how this is a gamma distribution. We will shortly see that these special gamma distributions arise as sampling distributions of variance estimators based on samples drawn from iid random variables distributed with normal distribution.

Pronunciation of χ^2: add a "k" before you say "I" and then "say squared", often in words its written "Chi-squared".

❖ ***Definition* 4.3** (χ^2 distribution)**.** For each positive number m, the gamma distribution with parameters $\alpha = m/2$ and $\beta = 1/2$ is called the χ^2 distribution with m degrees of freedom. If X has χ^2 distribution with m degrees of freedom we write $X \sim \chi^2_m$.

---❖

✎ **Remarks 4.2** (Moments)**.** $X \sim \chi^2_m$, then $E(X) = m$ and $Var(X) = 2m$. This i probably the most important take away from the definition of χ^2 distribution with m degrees of freedom is it has mean m and variance $2m$.

Gamma distribution: *Let α and β be positive numbers. A random variable X has the gamma distribution with parameters α and β if X has a continuous distribution for which the p.d.f. for $x > 0$,*

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

and for $x \le 0$, $f(x|\alpha,\beta) = 0$

The following is a useful theorem for a sequence of χ^2 distributions with varying degrees of freedom

✣ ***Theorem* 4.2** (Sum of χ^2 distributions)**.** If the random variables $X_1, \dots, X_k$ are independent and if $X_i \sim \chi^2_{m_i}$ for $i = 1, \dots, k$, then the $\sum_i^k X_i \sim \chi^2_{\sum_i^k m_i}$

---✣

The next theorem is our first connection of χ^2 and normal distributions,

✣ ***Theorem* 4.3.** Let $X \sim \mathcal{N}(0,1)$ and define $Y := X^2$. Then has the $Y \sim \chi^2_1$.

---✣

So the last theorem just says that the *square of the standard normal is distributed as χ^2*. Just combining the last two theorems give us the next important corollary

✣ ***Corollary* 4.1.** If the random variables $X_1, \dots, X_m$ are iid with the common distribution $X_i \sim \mathcal{N}(0,1)$, then $\sum_i^k X_i^2 \sim \chi^2_m$.

✎ **Remarks 4.3.**

(a) The last corollary is possibly the most important corollary that you should take away from this part that says the sum of squares of the standard normal is distributed with χ squared distribution where the degrees of freedom is just the number of standard normal distributions in the sum. So if we have sum of squares of the m independent standard normally distributed random variables, the resulting distribution is a χ^2 distribution with mean m and variance $2m$.

(a) The last corollary can also be applied also for sequence of iid random variables which are normal but not standard normal. This is because if $X_i \sim \mathcal{N}(\mu, \sigma^2)$, then we can create a new variable $Z_i := (X_i - \mu)/\sigma$, where $Z_i \sim \mathcal{N}(0,1)$. So if we let $X_1, \ldots, X_n$ be iid random variables that follow the common distribution $X_i \sim \mathcal{N}(\mu, \sigma^2)$, then the sequence $(X_1 - \mu)/\sigma, \ldots, (X_n - \mu)/\sigma$ are independent sequence and each of them is standard normally distributed random variables. Thus we have$\sum_{i=1}^{n} (X_i - \mu)^2/\sigma^2 \sim \chi^2_n$, or we can also write it as,

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n \tag{4.2}$$

.

4.4.2 The t distributions

The t distributions are also known as Student's distributions (see Student, 1908), in honor of W. S. Gosset, who published his studies of this distribution in 1908 under the pen name "Student."

The next important family of distributions are called the t distributions, which are also important to understand the random samples from a normal distribution. The t distributions, like the χ^2 distributions, have been widely applied in important problems of statistical inference. The distributions are defined as follows.

❖ ***Definition* 4.4** (t distributions). Consider two independent random variables Y and Z, such that W has the χ^2 distribution with m degrees of freedom and Z has the standard normal distribution. Suppose that a random variable X is defined by the equation

$$X = \frac{Z}{\sqrt{\frac{W}{m}}} \tag{4.3}$$

Then the distribution of X is called the t distribution with m degrees of freedom and we write $X \sim t_m$.

———❖

✎ **Remark 4.1** (Moments of t Distribution). Although the mean of the t distribution does not exist when $m \le 1$, the mean does exist for every value of $m > 1$. Of course, whenever the mean does exist, its value is 0 because of the symmetry of the t distribution. In general, if a random variable $X \sim t_m$ for $(m > 1)$, then it can be shown that $E\left(|X|^k\right) < \infty$ for $k < m$ and that $E\left(|X|^k\right) = \infty$ for $k \ge m$. If m is an integer, the first $m-1$ moments of X exist, but no moments of higher order exist. It can be shown that if $X \sim t_m$ with $(m > 2)$, then $\text{Var}(X) = m/(m-2)$.

4.4.3 The F distributions

F distribution is also known as Snedecor's F, whose derivation is quite similar to that of Student's t. Its named in honor of Ronald Fisher, so we have "F".

Finally we introduce the family, which is known as the family of F distributions. The motivation, however, is somewhat different. The F distribution arises naturally as the distribution of a ratio of variances. Although we will not use this distribution in this chapter, but you will see its very useful in two different hypothesis-testing situations. The first situation is when we wish to test hypotheses about the variances of two different normal distributions. The second situation will arise when we test hypotheses concerning the means of more than two normal distributions. Here is the formal definition,

❖ ***Definition* 4.5** (The F distributions)**.** Let Y and W be independent such that $Y \sim \chi^2_m$ and $W \sim \chi^2_n$. Define a new random variable X as follows:

$$X = \frac{Y/m}{W/n} = \frac{nY}{mW}$$

Then the distribution of X is called the F distribution with m and n degrees of freedom and we write $X \sim F_{m,n}$

---❖

✎ **Remark 4.2.** When we speak of the F distribution with m and n degrees of freedom, the order in which the numbers m and n are given is important, as can be seen from the definition of X When $m \neq n$, the F distribution with m and n degrees of freedom and the F distribution with n and m degrees of freedom are two different distributions. The next theorem gives a result relating the two distributions just mentioned along with a relationship between F distributions and t distributions.

✣ ***Theorem* 4.4.**

(a) If $X \sim F_{m,n}$ then $1/X \sim F_{n,m}$.

(b) If $Y \sim t_n$ then $Y^2 \sim F_{1,n}$.

---✣

4.4.4 Sampling distribution of Mean and Variance

Finally we have come to the important results related to the sampling distributions of $\bar{X}$ and S^2. We can put these results in the following theorem. But first recall

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

✣ ***Theorem* 4.5** (Sampling distributions of $\bar{X}$ and S^2 with normality)**.**
Suppose that $X_1, \dots, X_n$ are *independent random variables* follow the common normal distribution with mean μ and variance σ^2. Then,

$\perp\!\!\!\perp$ notation is used for independence, so when we say X and Y are independent random variables we write $X \perp\!\!\!\perp Y$.

(a)
$$\bar{X} \perp\!\!\!\perp S^2 \tag{4.4}$$

(b)
$$\bar{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right) \quad \left(\text{i.e.,} \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)\right) \tag{4.5}$$

(c)
$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{4.6}$$

---✣

One of the very useful consequences/applications of Theorem 4.5 is

✣ ***Corollary* 4.2.** If $X_1, \dots, X_n$ are *independent random variables* follow the common normal distribution with mean μ and variance σ^2,then

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \tag{4.7}$$

✎ **Remarks 4.4.**

- Part (a) of the Theorem 4.5 says that even if all the random variables are distributed identically with normal distribution with common mean and variance, $\bar{X}$ and variance S^2 are independent.

- Part (b) of the Theorem 4.5 gives us the sampling distribution of $\bar{X}$. Note, again we get this result because of normality assumption. This is often called the *exact* distribution (as opposed to *asymptotic* distribution!). Important is it is not always easy to get the exact distribution but in this case normality helped. Second thing to note is, since we have derived $\bar{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$, this then means $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}\left(0,1\right)$. Last important point is, note that the variance of the distribution is σ^2/n. So we need to know σ^2, if we would like to know the distribution of $\bar{X}$. But this is often an unrealistic assumption.

- Part (c) of the Theorem 4.5 has at least two interesting points to observe. First it gives us the scaled distribution of S^2, where the scaling factor is $(n-1)/\sigma^2$. The second is if you compare this to Eq. (4.2). There we had $\chi^2(n)$ but here the theorem says if we replace the population mean μ with the sample mean $\bar{X}$, the effect is simply to reduce the degrees of freedom in the χ^2 distribution from n to $n-1$

- We can also give a short proof of Corollary 4.2 using Theorem 4.5. First note, we can write,

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\dfrac{\bar{X}-\mu}{\sigma}}{\dfrac{S/\sqrt{n}}{\sigma}} = \frac{\left(\dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)}{\sqrt{\dfrac{S^2}{\sigma^2}}} = \frac{Z}{\sqrt{\dfrac{W}{n-1}}}, \tag{4.8}$$

Where we have defined two new random variables $Z := (\bar{X}-\mu)/(\sigma/\sqrt{n})$ and $W = (n-1)\left(S^2/\sigma^2\right)$. Now applying the result from Theorem 4.5, we can see Z and W are independent and also $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_{n-1}$. But now we can see that this ratio is exactly how we defined t distribution in **Definition 4.4**. So this gives us that $Z/\sqrt{W/(n-1)} \sim t_{n-1}$. One important aspect of this corollary is, now the distribution of $(\bar{X}-\mu)/(S/\sqrt{n})$ does not depend on σ^2. The only unknown element in the ratio is the population mean μ. This is really helpful if we would like to construct confidence intervals or testing for μ based on our sample estimate $\bar{x}$ and we do not know the population variance. This ratio $(\bar{X}-\mu)/(S/\sqrt{n})$ is sometimes called the *t ratio*. Also if we define a random variable $T := (\bar{X}-\mu)/(S/\sqrt{n})$, then we say T has Student's t distribution with $n-1$ degrees of freedom and we write, $T \sim t_{n-1}$

4.5 Point Estimation

Now we start estimation with point estimators. We can start from finding an estimator or we can also start evaluating an estimator. In general these two activities are closely connected. Often the methods of evaluating the estimators suggest how to find a point estimator. However, we have to start from somewhere, so we chose to start from the methods rather than evaluation and principles. For the time being we will only be concerned with the question - *given a data how can we find an estimator? in particular how can we find one number that summarizes the information.* We will be discussing whether how to evaluate them (i.e., estimators are good or bad or desirable) in coming sections, but now let us mechanically see the *methods of point estimation*.

4.5.1 Methods of point estimation

The rationale behind point estimation is quite simple. We have already seen many examples till now, e.g., the sample mean $\bar{X}$ is a point estimator and similarly the variance S^2. The idea is, a point estimator yields a single estimate of some parameter θ (or maybe a function of θ) for every $x \in \mathcal{X}$ instead of an entire confidence region. We may ask, well isn't the idea simple? *to estimate parameters, we can simply use their empirical counterparts*. For example, if our goal is to estimate the parameters of a normal distribution, that is μ and variance σ^2, then we can take their empirical counterparts sample mean and sample variance. This intuition works well, however, it could be the case that the parameters of a distribution do not have such empirical counterparts and neither we are always interested only in mean and variance. So we need some general methods to tackle these cases.

4.5.1.1 Method of Moments Estimators (MOMs)

The method of moments is, perhaps, the oldest method of finding point estimators, dating back at least to Karl Pearson in the late 1800s. It is based on the ideas to use some sort of sample counterparts, however it connects the parameters in a particular way. In method of moments first we equate kth population moment to sample moments and then solve the system of equations to find the estimators.

❖ ***Definition* 4.6** (Method of moments estimator (MOMs))**.** Assume iid random variables $X_1, \ldots, X_n$ follows a common distribution that is indexed by a k -dimensional parameter vector $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$, also assume it has at least k finite moments. For $j = 1, \ldots, k$, we define these moments by $\mu_j(\vartheta) = E_\vartheta\left(X_i^j\right)$. Suppose that the function $\mu(\vartheta) = (\mu_1(\vartheta), \ldots, \mu_k(\vartheta))$ is a one-to-one function of ϑ. Let $M\left(\mu_1, \ldots, \mu_k\right)$ denote the inverse function, that is, for all ϑ

$$\vartheta = M\left(\mu_1(\vartheta), \ldots, \mu_k(\vartheta)\right)$$

Define the sample moments by $m_j = \frac{1}{n}\sum_{i=1}^n X_i^j$ for $j = 1, \ldots, k$. The method of moments estimator of ϑ is $\hat{\vartheta} = M\left(m_1, \ldots, m_j\right)$.

---❖

The definition might be a bit abstract, but note the usual way of finding estimators using MOMs is to set up the k equations $m_j = \mu_j(\theta)$ and then solve for θ. Let us look at some concrete examples.

♣ **Example 4.7** (MOM for Normals)**.** Suppose $X_1, \ldots, X_n$ are iid $\mathcal{N}\left(\zeta, \sigma^2\right)$. In the preceding notation then

$$\begin{aligned}\vartheta_1 &= \zeta\\ \vartheta_2 &= \sigma^2\end{aligned}$$

So $(\vartheta_1, \vartheta_2)$ are our parameters that index the distribution and also we are interested in these parameters. Now note the moments are, $\mu_1 = E(X_i) = \zeta$ and $\mu_2 = E(X_i^2) = Var(X_i) + (E(X_i))^2 = \sigma^2 + \zeta^2$, so we can let

$$\begin{aligned}\mu_1 &= \zeta\\ \mu_2 &= \zeta^2 + \sigma^2\\ m_1 &= \bar{X}\\ m_2 &= (1/n)\sum X_i^2\end{aligned}$$

This gives us to write the system of equations,

$$\boxed{\begin{aligned}\bar{X} &= \zeta\\ \frac{1}{n}\sum X_i^2 &= \zeta^2 + \sigma^2\end{aligned}}$$

Now, solving for ζ and σ^2 yields the method of moments estimators

$$\hat{\theta}_1 = \hat{\zeta} = \bar{X} \quad \text{and} \quad \hat{\theta}_2 = \hat{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$$

(...end of example!)

In the last simple example, the method of moments solution coincides with our intuition and perhaps gives some credibility to both. Also note the method is somewhat more helpful when no obvious estimator suggests itself.

♣ **Example 4.8** (MOM for exponentials). Suppose $X_1, \dots, X_n$ constitutes a random sample from the distribution of X where $X \sim Exp(\lambda)$ (i.e., X is distributed exponentially with parameter λ). So the pdf of X then is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

We will estimate the parameter λ using the MOMs. The first moment of X in this case is

$$\mu_1 = \mathrm{E}(X) = \frac{1}{\lambda}$$

The first empirical moment m_1 is simply $\frac{1}{n}\sum_{i=1}^n X_i = \overline{X}$. Equating the first theoretical moment with the first empirical moment gives

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i$$

Now we solve for λ, and we get the MOMs estimator for the parameter λ,

$$\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^n X_i} = \frac{1}{\bar{X}}$$

(...end of example!)

✎ **Remark 4.3.** As we have seen from the examples that the moment method is usually easy to apply. However, it does not always provide the best estimates in the statistical sense. Moment estimators do not always possess characteristics such as unbiasedness, efficiency or sufficiency (we will define these objects in next section after MLE). Moreover, the moment estimator does not always exist. For example, the expected value or the first moment of the Cauchy distribution does not exist!, so we cannot use MOMs.

4.5.1.2 Maximum Likelihood Estimators (MLEs)

Maximum likelihood estimators, or often in short we say MLEs, are by far, the most popular estimators. In simple words - under certain distributional assumption this technique gives us an estimator such that the probability of the data is maximized.

❖ ***Definition* 4.7** (Likelihood Function and MLE). Let $f(x_1,\dots,x_n;\vartheta)$ be the joint probability density function (or probability mass function) of the random variables $X_1,\dots,X_n$.

(a) The function $L(\vartheta;x_1,\dots,x_n):\Theta\to[0,\infty)$, where

$$L(\vartheta;x_1,\dots,x_n) := f(x_1,\dots,x_n;\vartheta)$$

is called the *likelihood function for the outcome $x_1,\dots,x_n$*.

(b) An estimator $T^{ML}:\mathcal{X}\to\Theta$ of ϑ is called a *maximum likelihood estimator*, if for each $x_1,\dots,x_n\in\mathcal{X}$

$$L(T^{ML}(x_1,\dots,x_n),x_1,\dots,x_n) = \max_{\vartheta\in\Theta} L(\vartheta;x_1,\dots,x_n)$$

i.e., if the estimate $T^{ML}(x_1,\dots,x_n)$ is a maximizer of the function $L(\vartheta;x_1,\dots,x_n)$ on Θ holding $x_1,\dots,x_n$ fixed. We also write this as $T^{ML}(x_1,\dots,x_n)=\arg\max_{\vartheta\in\Theta} L(\vartheta;x_1,\dots,x_n)$

---❖

Note, the likelihood function is the joint density of the data, but we will regard this as a function of the parameter ϑ. Recall, when we think about the joint density, we think it as a function of the data and treat ϑ as a fixed parameter. Likelihood function is just the opposite, we will now treat the data $x_1,\dots,x_n$ is fixed and this function (calculates same thing as density) is a function of parameter ϑ. That is why if you compare the density $f(x_1,\dots,x_n;\vartheta)$ with $L(\partial;x_1,\dots,x_n)$, you can notice the positions of the arguments of the function have been swapped. ϑ can be scaler or vector.

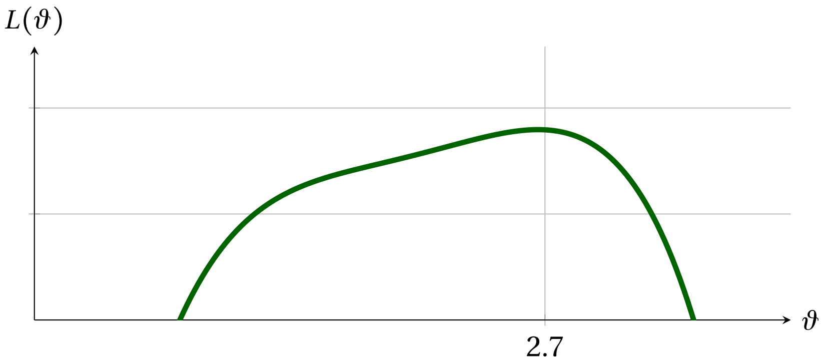

Figure 4.3: Example of a likelihood function

MLE is an estimator which finds this maximal ϑ such that in every data point the likelihood function is maximized. Fig. 4.3 is an example of a likelihood function for a scaler parameter ϑ and it is maximized at the value of 2.7. The idea of the maximum likelihood estimation is by construction it will try to give us the value like 2.7 in Fig. 4.3. So a MLE searches the optimal parameter such that the likelihood of the data is maximized. Following are some important remarks,

✎ **Remarks 4.5.**

- If $X_1, \ldots, X_n$ are iid random variables, which is often the assumption we will make, then we can write,

  $$L(\vartheta; x_1, \ldots, x_n) := \prod_{i=1}^{n} f(x_i; \vartheta)$$

  So likelihood function is then just the product of marginal densities.

- Often we will maximize log-likelihood rather than likelihood. We will denote the log-likelihood with $\ell(\vartheta; x_1, \ldots, x_n)$

  $$\ell(\vartheta; x_1, \ldots, x_n) = \log L(\vartheta; x_1, \ldots, x_n)$$

  This is because log transformation is a monotone transformation (one to one increasing function), so maximizing the log-likelihood leads to the same answer as maximizing the likelihood. Often, it is easier to work with the log-likelihood. Also, note If we multiply $L(\vartheta; x_1, \ldots, x_n)$: by any positive constant c (not depending on ϑ) then this will not change the MLE. Hence, we shall often drop constants in the likelihood function.

- Often too ease the notational burden we will use $\widehat{\theta}$ for an MLE, where $\widehat{\theta} := T^{ML}(X_1, \ldots, X_n)$.

We now give a detailed example, which also explains the intuitions of MLEs.

♣ **Example 4.9** (MLE for normal). Suppose we have observed three realizations of a random variable X, they are $x_1 = 250, x_2 = 258$ and $x_3 = 262$. So this is our data set which consists of these three observations. Now *based on some prior knowledge* of the data we assumed the this data is coming from a normal distribution. So we can think x_1, x_2 and x_3 are realizations of a variable $X \sim \mathcal{N}(\mu, \sigma^2)$, where we know $\sigma^2 = 100$ but the mean μ is unknown to us and our goal is to estimate μ with this observed dataset.

Suppose we have two possible candidates, $\mu_1 = 230$ and $\mu_2 = 257$. Now notice Fig. 4.4 where the left is the density of the normal distribution assuming the mean $\mu_1 = 230$ and the right is the density assuming $\mu_2 = 257$, x_1, x_2 and x_3 are the observed values.

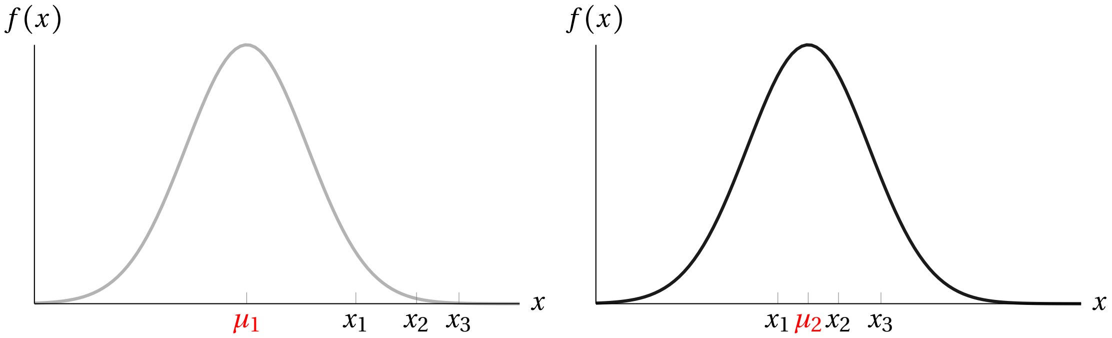

Figure 4.4: Heuristics of MLE

Since we know that for normal distribution, most of the data points will be close to the mean, we can say there is little possibility that the data is coming from the distribution with mean μ_1, rather it is highly likely that the data came from the normal distribution with mean μ_2. The idea of MLE is to use this idea to estimate the unknown parameter μ. So we will select the μ for which the probability of the data is maximized. Or in other words, when we find an MLE we cast this problem into an optimization problem to find an optimal estimate of the parameter given the data.

Now let us solve this problem, note because of normality we can write the density,

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}(10)} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{10}\right)^2\right]$$

Now using this we can write the likelihood function as

$$\begin{aligned}
&L(\mu;x_1,x_2,x_3)\\
&= f(x_1;\mu)\times f(x_2;\mu)\times f(x_3;\mu)\\
&= \frac{1}{\sqrt{2\pi}(10)} \exp\left[-\frac{1}{2}\left(\frac{x_1-\mu}{10}\right)^2\right] \times \frac{1}{\sqrt{2\pi}(10)} \exp\left[-\frac{1}{2}\left(\frac{x_2-\mu}{10}\right)^2\right]\\
&\qquad \times \frac{1}{\sqrt{2\pi}(10)} \exp\left[-\frac{1}{2}\left(\frac{x_3-\mu}{10}\right)^2\right]\\
&= \left(\frac{1}{\sqrt{2\pi}(10)}\right)^3 \exp\left[-\frac{1}{2\times 100}\sum_{i=1}^{3}(x_i-\mu)^2\right] \qquad (4.9)
\end{aligned}$$

Ofcourse we can directly maximize Eq. (4.9) with respect to μ and then solve for the optimal μ, but the algebra is tedious. So we can make our life easier when we see that $L(\mu;x_1,x_2,x_3)$ will be maximized by the value of μ that minimizes $Q(\mu;x_1,x_2,x_3)$, where

$$Q(\mu;x_1,x_2,x_3) = \sum_{i=1}^{3}(x_i-\mu)^2 = \sum_{i=1}^{3}x_i^2 - 2\mu\sum_{i=1}^{3}x_i + 3\mu^2$$

which is simply a quadratic function, we can easily take the derivative, set it equal to 0 and form the first order condition (FOC), then solving for the optimal ϑ, which we denoted by $\hat{\vartheta}$ will be our MLE. So, here is the calculation from the FOC

$$\begin{aligned} &\frac{d}{d\mu}Q(\hat{\mu};x_1,x_2,x_3)=0\\ \implies &-2\sum_{i=1}^{3}x_i+6\hat{\mu}=0\\ \implies &\hat{\mu}=\frac{1}{3}\sum_{i=1}^{3}x_i=\overline{x}\end{aligned}$$

Infact we can apply this result even if we have data sets of n realizations. So on that case $\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$. So the MLE of the population mean μ of the normal distribution is sample mean $\bar{x}$. This makes sense, right!

(...end of example!)

Now we give some important remarks regarding MLEs. Under certain regularity conditions MLEs satisfy some nice properties like, identifiability, existence, consistency and asymptotic normality. We defer this discussion after we introduce some more concepts in coming sections. Now we state a theorem which is useful which is known as the *invariance property of MLE*.

✣ ***Theorem* 4.6** (Invariance Property of MLEs)**.** if $\hat{\vartheta}$ is an MLE of ϑ, and $g(\vartheta)$ be a function of ϑ, then an MLE of $g(\vartheta)$ is $g(\hat{\vartheta})$.

---✣

Note, for a one-to-one function g, this is just a direct extention, but this result holds in general even if g is not one-to-one. For the general case, we need to carefully write the MLE of $g(\vartheta)$.

We now contiue with the last example of Normal distribution but now with both unknown mean and variance, where we will also apply invariance property.

♣ **Example 4.10** (MLE for normals, with unknown meand and variance)**.** Suppose again that $X_1,\dots,X_n$ form a random sample from a normal distribution, but suppose now that both the mean μ and the variance σ^2 are unknown. The parameter vector is then $\theta=\left(\mu,\sigma^2\right)$. For all observed values $x=(x_1,\dots,x_n)$ the likelihood function $L\left(\mu,\sigma^2;x\right)$ can be written as

$$L\left(\mu,\sigma^2;x\right)=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

This function must now be maximized over all possible values of μ and σ^2, where $-\infty<\mu<\infty$ and $\sigma^2>0$ Instead of maximizing the likelihood function $L\left(\mu,\sigma^2;x\right)$ directly, it is actually to maximize $\log L_n\left(\mu,\sigma^2;x\right)$, i.e., the log likelihood function. We will get the same result because, log transformation is a monotone transformation, so this means, maximizing the log likelihood or maximizing the likelihood will give us the same maximum.

$$\begin{aligned} L(\theta)&=\log f_n\left(\boldsymbol{x}|\mu,\sigma^2\right)\\ &=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned}$$

(...end of example!)

✎ **Remark 4.4** (Calculation of ML estimators)**.**

- The *log-likelihood function*, the natural logarithm of the likelihood function ln$(L(\vartheta))$, is a monotone transformation of likelihood. Therefore $L(\vartheta)$ and $\ln L(\vartheta)$ have their maximum at the same position $\vartheta_{\max}$. This property is very helpful because the maximum position of $\ln L(\vartheta)$ is often easier to determine. It applies:

$$\ln(L(\vartheta)) = \ln\left(\prod_{i=1}^{n} f_X(x_i;\vartheta)\right) = \sum_{i=1}^{n} \ln(f_X(x_i;\vartheta)).$$

- If the likelihood function depends on k parameters , then the maximum $(\hat{\vartheta}_1,...,\hat{\vartheta}_k)'$ of the (log) likelihood function is determined by the solution of the k-elementary system of equations

$$\begin{aligned}\frac{\partial}{\partial\vartheta_1}L() &= 0\\ \frac{\partial}{\partial\vartheta_2}L() &= 0\\ &\vdots\\ \frac{\partial}{\partial\vartheta_k}L() &= 0.\end{aligned}$$

A check whether the found places are really maximum places is possible in the following way: Designate with H a $k \times k$ matrix consisting of the partial derivatives of the likelihood function at the place $(\hat{\vartheta}_1,...,\hat{\vartheta}_k)'$

$$H = \begin{pmatrix} h_{11} & \cdots & h_{1k}\\ \vdots & \ddots & \vdots\\ h_{k1} & \cdots & h_{kk}\end{pmatrix} h_{ij} = \left.\frac{\partial L()}{\partial\vartheta_i\partial\vartheta_j}\right| .$$

If the matrix H has a negative definition, i.e. $\sum_{i=1}^{n}\sum_{i=1}^{n} y_i y_j h_{ij} < 0$ for any vector $y = (y_1,...,y_n) \neq (0,...,0) \in \mathbb{R}^k$, then there are maximum digits.

To check whether a matrix is negative definite, the eigenvalues (see chapter matrix calculation can be calculated. If all eigenvalues are negative, the matrix is negatively definite.

- If the parameter ϑ can only assume discrete values, it is useful to examine the monotonicity property of the likelihood function or the quotient $\frac{L(\vartheta)}{L(\vartheta+1)}$. If the value of the quotient changes from a value less than 1 to a value greater than 1, a (local) maximum is reached.

4.6 Evaluation of estimators

The methods we discussed in previous sections gives some reasonable techniques to find point estimators. However the issue is, since it is possible to apply more than one of these methods in a particular situation, we are often faced with the more than one estimators, then how do we choose one?.

Of course, it is possible that different methods of finding estimators will yield the same answer, which makes evaluation a bit easier (e.g., in regression least squares and MOMs) but, in many cases, different methods will lead to different estimators, i.e., different functions of data. So we need to investigate the estimators. This is called *evaluation of the estimators*.

In the following we will discuss some of the criterion to evaluate. For all estimators in the background we always have the statistical model $(\mathcal{X},\mathcal{F},\mathcal{P})$, where $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ is the family of probability distributions.

❖ ***Definition* 4.8** (Bias, Unbiasedness, MSE)**.** Let θ be our parameter of interest, then

- We call an estimator $T : \mathcal{X} \to \mathbb{R}$ of θ *unbiased* if $E_\vartheta(T) = \vartheta$ for all $\vartheta \in \Theta$.
- When T is NOT unbiased, then there is a bias of T at θ, we will denote it by $\text{Bias}_\vartheta(T)$, where $\text{Bias}_\vartheta(T) := E_\vartheta(T) - \vartheta$.
- The mean squared error or in short *MSE* of T at θ is the average of the squared error of T at θ, defined as $\text{MSE}_\vartheta(T) := E_\vartheta\left((T - \vartheta)^2\right)$.

———❖

✎ **Remarks 4.6.**

- An unbiased estimator avoids systematic errors. This is clearly a sensible criterion, but note, it is not automatically compatible with the all estimators, for example there are examples of MLEs and MOMs which are biased (we will see some examples shortly!).
- For an estimator T both the $\text{Bias}_\vartheta(T)$ and $\text{MSE}_\vartheta(T)$ depends on θ, that is they are actually functions of θ. This is why sometimes you will see this is explicitly written in some places, e.g., $\text{Bias}_T(\vartheta)$.
- Also note often T is an estimator of any function of ϑ, e.g., $\tau\vartheta$. So to be more general to define unbiasedness we can write $E_\vartheta(T) = \tau\vartheta$, so it is just a matter of adjusting notations. To have less notational burden we simply used θ.

Why average squared error? In general, any increasing function of the absolute distance $|T - \vartheta|$ would serve to measure the goodness of an estimator (e.g., mean absolute error, $\text{E}_\vartheta(|T - \vartheta|)$ is also a reasonable alternative), but MSE has at least two advantages over other distance measures: First, it is quite tractable analytically and, second, it has a relation with bias and variance, that is $\text{MSE}_\vartheta(T) = \text{Var}_\vartheta(T) + \text{Bias}_\vartheta(T)$, lets see the short proof of this result. We start from the definition, apply the *linearity of expectation* and the fact that θ is a constant.

$$\begin{aligned}
\text{MSE}_\vartheta(T) := E_\theta\left((T - \vartheta)^2\right) &= E_\theta\left(T^2\right) - 2\vartheta E_\theta(T) + \vartheta^2 \\
&= E_\theta\left(T^2\right) - (E_\theta(T))^2 + (E_\theta(T))^2 - 2\vartheta E_\theta(T) + \vartheta^2 \\
&= \text{Var}_\vartheta(T) + (E_\vartheta(T) - \vartheta)^2 \\
&= \text{Var}_\vartheta(T) + \text{Bias}_\vartheta(T)
\end{aligned}$$

In the second equality we added and subtracted a term and in the third equality we have used - for any random variable X, $\text{Var}(X) = E(X^2) - (E(X))^2$. This relation quickly tells us that for an unbiased estimator $\text{MSE}_\vartheta(T) = \text{Var}_\vartheta(T)$, that is MSE is equal to its variance.

Now let us talk about MSE. MSE incorporates two components, first $\text{Var}_\vartheta(T)$, that measures the variability of the estimator (a related terminology is “precision” which is the reciprocal of the variance) and the other is $\text{Bias}_\vartheta(T)$, which measures (accuracy).

An estimator that has good MSE properties has small combined variance and bias. And it is a desirable property because it means we are reducing the error of our approximation. Clearly, unbiased estimators will do a good job of controlling bias, but obviously controlling bias does not guarantee that MSE is controlled because it depends also on bias. So there might be trade off in certain situations. It might be the case that sometimes a small increase in bias can be traded for a larger decrease in variance, resulting MSE to be reduced even more. Let us look at an example, where we have an biased estimator but it reaches to lower MSE than an unbiased estimator.

♣ **Example 4.11.** Recall S^2 (i.e., the sample variance) which is the estimator of σ^2.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

We can look at an alternative estimator

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$$

If you remember this is what we have calculated as a MOMs estimator and also as a MLE of σ^2. Now we can easily calculate,

$$E(\hat{\sigma}^2) = \mathrm{E}\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2$$

so $\hat{\sigma}^2$ is a biased estimator of σ^2. The variance of $\hat{\sigma}^2$ can also be calculated as

$$\mathrm{Var}\,\hat{\sigma}^2 = \mathrm{Var}\left(\frac{n-1}{n}S^2\right) = \left(\frac{n-1}{n}\right)^2 \mathrm{Var}\,S^2 = \frac{2(n-1)\sigma^4}{n^2}$$

and, hence, its MSE is given by

$$\mathrm{E}\left(\hat{\sigma}^2 - \sigma^2\right)^2 = \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4$$

We thus have

$$\mathrm{E}\left(\hat{\sigma}^2 - \sigma^2\right)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4 < \left(\frac{2}{n-1}\right)\sigma^4 = \mathrm{E}\left(S^2 - \sigma^2\right)^2$$

showing that $\hat{\sigma}^2$ has smaller MSE than S^2. Thus, by trading off variance for bias, the MSE is improved.

(...end of example!)

We need to quickly point out that the above example does not imply that S^2 should be abandoned as an estimator of σ^2. The above argument shows that, on the average, $\hat{\sigma}^2$ will be closer to σ^2 than S^2 if MSE is used as a measure. However note, $\hat{\sigma}^2$ is *biased* and will, *on the average underestimate* σ^2. This fact alone may make us uncomfortable about using $\hat{\sigma}^2$ as an estimator of σ^2.

So what can we conclude? We can conclude that there is no absolute answer to obtain an estimator. But what we can do is we can gather different and perhaps more information about the estimators hoping that, for a particular situation we can use these knowledge to choose a good estimator.

In general if we want to choose an estimator solely on the basis of the MSE, then we will literally have to consider all estimators in all points of ϑ in Θ, and maybe choose the "one best" estimator using MSE. But as you can guess we have to deal with a huge class of estimators and in reality there are examples where for certain parts of Θ one estimator has lower MSE than others, and in other parts there are other estimators which have lower MSEs. So rather than comparing all possible estimators in the whole Θ, we can compare two estimators compare their performance and this is what we call *relative efficiency*,

❖ ***Definition* 4.9** (Relative efficiency and admissibility)**.** Let T and $T^{'}$ be two estimators of a parameter ϑ.

- The *relative efficiency*, denoted by $\mathrm{RE}_{\vartheta}\left(T,T^{'}\right)$ of T with respect to $T^{'}$ is
$$\mathrm{RE}_{\vartheta}\left(T,T^{'}\right) := \frac{\mathrm{MSE}_{\vartheta}\left(T^{'}\right)}{\mathrm{MSE}_{\vartheta}(T)} = \frac{E_{\vartheta}\left(\left(T^{'}-\vartheta\right)^{2}\right)}{E_{\vartheta}\left(\left(T-\vartheta\right)^{2}\right)}, \quad \forall\vartheta\in\Theta$$
- T is relatively more efficient than $T^{'}$ if $\mathrm{RE}_{\vartheta}\left(T,T^{'}\right)\geq 1\,\forall\vartheta\in\Theta$ and $\mathrm{RE}_{\vartheta}\left(T,T^{'}\right) > 1$ for some $\vartheta\in\Theta$.
- For T if there exists $T^{'}$ that is relatively more efficient than T, then T is called *inadmissible* for estimating ϑ. Otherwise, T is called *admissible*.

—–❖

Relative efficiency is a good criterion, but it is not uniform in the sense that it gives us an estimator for all ϑ. Maybe we would like to find an estimator in a larger class, which holds for all possible ϑ. As we have discussed only looking at MSE for whole Θ is not a solution. But it turns out, it is possible to narrow our search. One of the ideas is we can narrow our focus to only the class of *unbiased estimators*. So pick a ϑ, and then for all the estimators that you can find at ϑ pick the one which has the lowest MSE. But because it is already unbiased, lowest MSE means lowest variance. Now if we find an unbiased estimator that has lowest variance regardless of which ϑ we pick, we have indeed found solution for all θ but only in this unbiased class of estimators. We call this *uniformly minimum variance unbiased estimator* or UMVUE in short. The word “uniformly” emphasizes that this estimator being minimum in variance for whole ϑ. Here is the formal definition,

❖ ***Definition* 4.10** (Best unbiased estimator or UMVUE)**.** An unbiased estimator T of θ is called a *best unbiased estimator*, also called *uniformly minimum variance unbiased estimator (UMVUE)* if

$$\mathrm{Var}_{\vartheta}(T) \leq \mathrm{Var}_{\vartheta}(T^{'})$$

for any $T^{'}$ which is unbiased at any $\vartheta\in\Theta$.

—–❖

How do we go to find an UMVUE estimator, check one by one? It turns out its not an easy problem. Even if you compare any two unbiased estimators, you can then combine them to create another set of estimators, so at the end you have too many comparisons to make. Here is an example,

♣ **Example 4.12.** Let $X_1,\ldots,X_n$ be iid Poisson (λ), and let $\bar{X}$ and S^2 be the sample mean and variance, respectively. Recall that for the Poisson pmf both the mean and variance are equal to λ, and because we know that the sample mean and variance are unbiased regardless of the distributions, so

$$\begin{aligned} \mathrm{E}_{\lambda}(\bar{X}) &= \lambda, \quad \text{for all } \lambda \\ \mathrm{E}_{\lambda}(S^2) &= \lambda, \quad \text{for all } \lambda \end{aligned}$$

so both $\bar{X}$ and S^2 are unbiased estimators of λ To determine the better estimator, $\bar{X}$ or S^2, we should now compare variances. In this case it is possible to show that $\mathrm{Var}_{\lambda}\left(\bar{X}\right)\leq\mathrm{Var}_{\lambda}\left(S^2\right)$ for all λ. So now we can establish that $\bar{X}$ is better than S^2. Now, consider the class of estimators which are convex combinations of the two, so

$$T_a\left(\bar{X}, S^2\right) = a\bar{X} + (1-a)S^2$$

Now for every constant a we also have $E_\lambda(T_a\left(\bar{X}, S^2\right)) = \lambda$, so now we have infinitely many unbiased estimators of λ. Even if $\bar{X}$ is better than S^2, is it better than every $T_a\left(\bar{X}, S^2\right)$? Furthermore, how can we be sure that there are not other, better, unbiased estimators around?

(...end of example!)

So the example suggests perhaps we need a more comprehensive approach. Here is one such approach, suppose that, for estimating a parameter ϑ of a distribution $f(x|\vartheta)$, we can specify a lower bound, say $B(\vartheta)$, on the variance of *any unbiased estimator* of ϑ. So if we can find an unbiased estimator T satisfying $\mathrm{Var}_\theta(T) = B(\theta)$, we have found a UMVUE. In this approach this bound is called *Cramér-Rao Lower Bound* or CRLB in short. To establish CRLB, we need some assumptions on the family of the distributions. These assumptions together are often called, *regularity conditions*.

❖ ***Definition* 4.11** (Regularity Conditions)**.** A one parameter family $(\mathcal{X}, \mathscr{F}, \mathcal{P})$, where $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$, is called regular if

(r1) Θ is an open interval on $\mathbb{R}$

(r2) The likelihood function on $\mathcal{X} \times \Theta$ is strictly positive and continuously differentiable in ϑ. So, in particular, there exists the *score* function $U_\vartheta(x)$, where

$$U_\vartheta(x) := \frac{d}{d\vartheta} \log f(x, \vartheta) = \frac{f_x'(\vartheta)}{f_x(\vartheta)}$$

(r3) $f(x, \vartheta)$ allows to interchange the differentiation in ϑ and the integration over x

$$\int \frac{d}{d\vartheta} f(x, \vartheta) dx = \frac{d}{d\vartheta} \int f(x, \vartheta) dx$$

(r4) For each $\vartheta \in \Theta$, the variance $I(\vartheta) := V_\vartheta\left(U_\vartheta\right)$ exists and is non-zero. The function $I : \vartheta \to I(\vartheta)$ is then called the *Fisher information* of the model.

———❖

✎ **Remarks 4.7.**

- The significance of (r3) comes when we see,

$$E_\vartheta\left(U_\vartheta\right) = \int \frac{d}{d\vartheta} f(x, \vartheta) dx = \frac{d}{d\vartheta} \int f(x, \vartheta) dx = \frac{d}{d\vartheta} 1 = 0$$

  which says that the score function of ϑ is centered with respect to P_ϑ. In particular,

$$I(\vartheta) = \mathbb{E}_\vartheta\left(U_\vartheta^2\right)$$

  because $\left(E_\vartheta\left(U_\vartheta\right)\right)^2 = 0$.

4.6.1 Principles

In this section we mention certain principles to find a good estimators.

4.6.2 Sufficiency Principle

❖ ***Definition* 4.12** (Sufficient Statistic). A statistic $T(X_1,\ldots,X_n)$ is a *sufficient statistic* for ϑ if the conditional distribution of the sample $X_1,\ldots,X_n$ given the value of $T(X_1,\ldots,X_n)$ does not depend on ϑ.

–––❖

✣ ***Lemma* 4.1** (Sufficiency Principle). If $T(X)$ is a sufficient statistic for ϑ, then any inference about ϑ should depend on the sample X only through the value $T(X)$. That is, if x and y are two sample points such that $T(x) = T(y)$, then the inference about ϑ should be the same whether $X = x$ or $X = y$ is observed.

Lets elaborate on the Definition 4.12 a bit assuming the $T(X_1,\ldots,X_n)$ has a discrete distribution. To ease the notational burden we will denote $X = (X_1,\ldots,X_n)$, so X is simply a n dimensional random vector and the realizations will be written as x.

So we are interested in conditional distribution of the sample given the value of a statistic. That is for example, a particular value $T(X) = t$ of the statistic can result for a set of value $A_t = \{x : T(x) = t)\}$ from the distribution of X, so we will have the distribution for $P_\vartheta(X = x|T(X) = t)$. If x is not a sample point from that set, than $T(x) \neq t$, then clearly $P_\vartheta(X = x|T(X) = t) = 0$.

Thus we can restrict our interest in $P(X = x|T(X) = T(x))$.

By the definition, if $T(X)$ is a sufficient statistic, this conditional probability is the same for all values of ϑ so we can omit the subscript ϑ. A sufficient statistic captures all the information about ϑ in this sense.

Let us look at an example. Consider two statisticians, 1st-statistician observes the value of $X = x$ and can compute the value $T(x)$, of the statistic $T(X)$ at x. Now consider 2nd-statistician only a value for the statistic $T(X) = T(x)$ but does not know from any sample point. 2nd-statistician knows $P(X = y|T(X) = T(x)$ which is the probability distribution on $A_{T(x)} = \{y : T(y) = T(x))\}$. This is because this can be computed from the model without knowledge of the true value of ϑ just by looking at the conditional distribution. Thus, 2nd-statistician can use this distribution to generate an observation Y satisfying $P(Y = y|T(X) = T(x)) = P(X = y|T(X) = T(x))$. It turns out that, for each value of ϑ, X and Y we have the same unconditional probability distribution, as we shall see below. So 1st-statistician who knows X, and 2nd-statistician who knows Y have equivalent information about ϑ. All his knowledge about ϑ is contained in the knowledge that $T(X) = T(x)$. So 2nd-statistician who knows only $T(X) = T(x)$, has just as much information about ϑ as does 1st-statistician , who knows the entire sample $X = x$

To complete the above argument, we need to show that X and Y have the same unconditional distribution, that is, $P_\vartheta(X = x) = P_\vartheta(Y = x)$ for all x and ϑ. Note that the events $\{X = x\}$ and $\{Y = x\}$ are both subsets of the event $\{T(X) = T(x)\}$. Also recall that

$$P(X = x|T(X) = T(x)) = P(Y = x|T(X) = T(x))$$

and these conditional probabilities do not depend on ϑ. Thus we have

$$
\begin{aligned}
&P_\vartheta(X=x)\\
&=P_\vartheta(X=x \text{ and } T(X)=T(x))\\
&=P(X=x|T(X)=T(x))P_\vartheta(T(X)=T(x)) && \left(\begin{array}{c}\text{definition of}\\ \text{conditional probability}\end{array}\right.\\
&=P(Y=x|T(X)=T(x))P_\vartheta(T(X)=T(x))\\
&=P_\vartheta(Y=x \text{ and } T(X)=T(x))\\
&=P_\vartheta(Y=x)
\end{aligned}
$$

To use Definition 4.12, to verify that a statistic $T(X)$ is a sufficient statistic for ϑ, we must verify that for any fixed values of x and t, the conditional probability $P_\vartheta(X=x|T(X)=t)$ is the same for all values of ϑ. Again, this probability is 0 for all values of ϑ if $T(x)\neq t$. So, we must verify only that $P_\vartheta(X=x|T(X)=T(x))$ does not depend on ϑ. But since $\{X=x\}$ is a subset of $\{T(X)=T(x)\}$

$$
\begin{aligned}
P_\vartheta(X=x|T(X)=T(x)) &= \frac{P_\vartheta(X=x \text{ and } T(X)=T(x))}{P_\vartheta(T(X)=T(x))}\\
&= \frac{P_\vartheta(X=x)}{P_\vartheta(T(X)=T(x))}\\
&= \frac{f(x|\vartheta)}{g(T(x)|\vartheta)}
\end{aligned}
$$

where $f(x|\vartheta)$ is the joint pmf of the sample X and $g(t|\vartheta)$ is the pmf of $T(X)$. Thus, $T(X)$ is a sufficient statistic for ϑ if and only if, for every x, the above ratio of pmfs is constant as a function of ϑ . Although we showed this for the discrete distributions, but this can be extended to continuous cases, this gives us following **lemma**

✣ ***Lemma* 4.2.** If $f(x|\vartheta)$ is the joint pdf or pmf of X and $g(t|\vartheta)$ is the pdf or pmf of $T(X)$, then $T(X)$ is a sufficient statistic for ϑ if, for every x in the sample space, the ratio $f(x|\vartheta)/g(T(x)|\vartheta)$ is constant as a function of ϑ.

Now, we will verify sufficiency for two common statistics with Lemma 4.2.

♣ **Example 4.13** (Binomial sufficient statistic)**.** Let $X_1,\ldots,X_n$ be iid Bernoulli RVs with parameter $p, 0<p<1$. Then their joint pmf is given by,

$$
f(x|p) := \prod_i p^x \cdot (1-p)^{1-x_i}
$$

We will show that $TX = \sum_{i=1}^n X_i = X_1+\cdots+X_n$ is a sufficient statistic for p. Note that $T(X)$ gives the total number of X_i's that are 1. Therefore, $T(X)$ has a Bin(n,p) distribution and the pmf is given by

$$
g(T(x)|p) := \binom{n}{T(x)} p^{T(x)}(1-p)^{n-T(x)} = \binom{n}{\sum_{i=1} x_i} p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}
$$

The ratio of pmfs is then

$$\begin{aligned}
\frac{f(x|p)}{g(T(x)|p)} &= \frac{\prod_i p^x \cdot (1-p)^{1-x_i}}{\begin{pmatrix} n \\ \sum_i x_i \end{pmatrix} p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}} \\
&= \frac{p^{\sum_i x_i}(1-p)^{\sum_i (1-x_i)}}{\begin{pmatrix} n \\ \sum_i x_i \end{pmatrix} p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}} \\
&= \frac{p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}}{\begin{pmatrix} n \\ \sum_i x_i \end{pmatrix} p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}} \\
&= \frac{1}{\begin{pmatrix} n \\ \sum_i x_i \end{pmatrix}} \\
&= \frac{1}{\begin{pmatrix} n \\ T(x) \end{pmatrix}}
\end{aligned}$$

(...end of example!)

Since this ratio does not depend on ϑ, by Lemma 4.2, $T(X)$ is a sufficient statistic for ϑ. The interpretation is this: The total number of 1 s in this Bernoulli sample contains all the information about ϑ that is in the data. Other features of the data, such as the exact value of X_2, contain no additional information.

♣ **Example 4.14** (Normal sufficient statistic)**.**

Let $X_1, \ldots, X_n$ be iid $\mathcal{N}\left(\mu, \sigma^2\right)$ where σ^2 is known. We want to show that the sample mean, $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$, is a sufficient statistic for μ. The joint pdf of the sample X is

$$\begin{aligned}
f(x|\mu) &= \prod_{i=1}^n \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\left(x_i-\mu\right)^2/\left(2\sigma^2\right)\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\sum_{i=1}^n \left(x_i-\mu\right)^2/\left(2\sigma^2\right)\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\sum_{i=1}^n \left(x_i-\bar{x}+\bar{x}-\mu\right)^2/\left(2\sigma^2\right)\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\left(\sum_{i=1}^n \left(x_i-\bar{x}\right)^2 + n(\bar{x}-\mu)^2\right)/\left(2\sigma^2\right)\right)
\end{aligned}$$

The last equality is true because the cross-product term $\sum_{i=1}^n \left(x_i-\bar{x}\right)\left(\bar{x}-\mu\right)$ may be rewritten as $\left(\bar{x}-\mu\right)\sum_{i=1}^n \left(x_i-\bar{x}\right)$, and $\sum_{i=1}^n \left(x_i-\bar{x}\right) = 0$. Recall that the sample mean $\bar{X}$ has a not $\left(\mu, \sigma^2/n\right)$ distribution with pdf

$$g(\bar{x}|\mu) = \left(2\pi\sigma^2/n\right)^{-1/2} \exp\left(-n(\bar{x}-\mu)^2/\left(2\sigma^2\right)\right)$$

Then, the ratio of the pdfs is

$$\begin{aligned}
\frac{f(x|\mu)}{g(\bar{x}|\mu)} &= \frac{\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\left(\sum_{i=1}^n \left(x_i-\bar{x}\right)^2 + n(\bar{x}-\mu)^2\right)/\left(2\sigma^2\right)\right)}{\left(2\pi\sigma^2/n\right)^{-1/2} \exp\left(-n(\bar{x}-\mu)^2/\left(2\sigma^2\right)\right)} \\
&= n^{-1/2}\left(2\pi\sigma^2\right)^{-(n-1)/2} \exp\left(-\sum_{i=1}^n \left(x_i-\bar{x}\right)^2/\left(2\sigma^2\right)\right)
\end{aligned} \qquad (4.10)$$

which does not depend on μ. By Theorem 4.2, the sample mean is a sufficient statistic for μ.

(...end of example!)

To find a sufficient statistic for a particular model using the definition is usually tedious. One must guess a statistic $T(X)$ to be sufficient, find its pmf or pdf, and check that the ratio of pdfs or pmfs does not depend ob ϑ. The following theorem makes the task easier by simply allowing to find a sufficient statistic by simple inspection of the pdf or pmf of the sample. It says, if the pdf or pm of X can be factored into a product such that one factor, does not depend on ϑ and the other factor, which does depend on ϑ, but depends on sample X only through $T(X)$.

✣ ***Theorem* 4.7** (Factorization Theorem)**.** Let$f(x|\vartheta)$ be the joint pdf or pmf of a sample X. A statistic $T(X)$ is a sufficient statistic for ϑ if and only if there exist functions $g(t|\vartheta)$ and $h(x)$ such that, for all x and for all ϑ

$$f(x|\vartheta) = g(T(x)|\vartheta)h(x) \tag{4.11}$$

Proof. Here we give the proof only for discrete distributions. Suppose $T(X)$ is a sufficient statistic. Choose $g(t|\vartheta) = P_\vartheta(T(X) = t)$ and $h(x) = P(X = x|T(X) = T(x))$. Because $T(X)$ is sufficient, the conditional probability defining $h(x)$ does not depend on ϑ. Thus this choice of $h(x)$ and $g(t|\vartheta)$ is legitimate, and for this choice we have

$$\begin{aligned} f(x|\vartheta) &= P_\vartheta(X = x) \\ &= P_\vartheta(X = x \text{ and } T(X) = T(x)) \\ &= P_\vartheta(T(X) = T(x))P(X = x|T(X) = T(x)) \quad \text{(sufficiency)} \\ &= g(T(x)|\vartheta)h(x) \end{aligned}$$

So factorization in (4.11) has been exhibited. We also see from the last two lines above that

$$P_\vartheta(T(X) = T(x)) = g(T(x)|\vartheta)$$

so $g(T(x)|\vartheta)$ is the pmf of $T(X)$.

Now for the other direction of the proof, we assume the factorization in (4.11) exists. Let $q\,(t|\vartheta)$ be the pmf of $T(X)$. To show that $T(X)$ is sufficient we examine the ratio $f(x|\vartheta)/q(T(x)|\vartheta)$. Define $A_{T(x)} = \{y : T(y) = T(x)\}$. Then

$$\begin{aligned} \frac{f(x|\vartheta)}{q(T(x)|\vartheta)} &= \frac{g(T(x)|\vartheta)h(x)}{q(T(x)|\vartheta)} \\ &= \frac{g(T(x)|\vartheta)h(x)}{\sum_{A_{T(x)}} g(T(y)|\vartheta)h(y)} \quad \text{(definition of the pmf of } T\text{)} \\ &= \frac{g(T(x)|\vartheta)h(x)}{g(T(x)|\vartheta)\sum_{A_{T(x)}} h(y)} \quad \text{(since } T \text{ is constant on } A_{T(x)}\text{)} \\ &= \frac{h(x)}{\sum_{A_{T(x)}} h(y)} \end{aligned}$$

Since the ratio does not depend on ϑ, by Lemma 4.2, $T(X)$ is a sufficient statistic for ϑ. ■

———✣

The Factorization Theorem is also known as *Neyman-Fisher Factorization Criterion*. To use the Factorization Theorem to find a sufficient statistic, we factor the joint pdf of the sample into two parts, with one part not depending on ϑ. The part that does not depend on ϑ constitutes the $h(x)$ function. The other part, the one that depends on ϑ, usually depends on the sample x only through some function $T(x)$ and this function is a sufficient statistic for ϑ. This is illustrated in the following example.

♣ **Example 4.15** (Continuation of Example 4.14)**.** For the normal model described earlier, we saw that the pdf could be factored as

$$f(x|\mu) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\sum_{i=1}^{n} (x_i-\bar{x})^2/\left(2\sigma^2\right)\right) \exp\left(-n(\bar{x}-\mu)^2/\left(2\sigma^2\right)\right) \tag{4.12}$$

We can define

$$h(x) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\sum_{i=1}^{n} (x_i-\bar{x})^2/\left(2\sigma^2\right)\right)$$

which does not depend on the unknown parameter μ. The factor in (4.12) that contains μ depends on the sample x only through the function $T(x) = \bar{x}$, the sample mean. So we have

$$g(t|\mu) = \exp\left(-n(t-\mu)^2/\left(2\sigma^2\right)\right)$$

and note that

$$f(x|\mu) = h(x)g(T(x)|\mu)$$

Thus, by the Factorization Theorem, $T(X) = \bar{X}$ is a sufficient statistic for μ.

(...end of example!)

In all the previous examples, the sufficient statistic is a real-valued function of the sample. All the information about ϑ in the sample x is summarized in the single number $T(x)$. Sometimes, the information cannot be summarized in one number and several numbers are required instead. In such cases, a sufficient statistic is a vector, say $T(X) = (T_1(X), \ldots, T_r(X))$. This situation often occurs when the parameter is also a vector, say $\vartheta = (\vartheta_1, \ldots, \vartheta_s)$. The Factorization Theorem may be used to find a vector-valued sufficient statistic.

♣ **Example 4.16** (Normal sufficient statistic, both parameters unknown)**.** Again assume that $X_1, \ldots, X_n$ are iid $\mathcal{N}\left(\mu, \sigma^2\right)$ but, unlike Example 4.14 assume that both μ and σ^2 are unknown so the parameter vector is $\vartheta = \left(\mu, \sigma^2\right)$. Now when we use the Factorization Theorem, any part of the joint pdf that depends on either μ or σ^2 must be included in the g function. From (4.10) it is clear that the pdf depends on the sample x only through the two values $T_1(x) := \bar{x}$ and $T_2(x) := s^2 = \sum_{i=1}^{n} (x_i-\bar{x})^2/(n-1)$. Thus we can define $h(x) = 1$ and

$$\begin{aligned} g(t|\theta) &= g\left(t_1, t_2|\mu, \sigma^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\left(n\left(t_1-\mu\right)^2 + (n-1)t_2\right)/\left(2\sigma^2\right)\right) \end{aligned}$$

Then it can be seen that

$$f\left(x|\mu, \sigma^2\right) = g\left(T_1(x), T_2(x)|\mu, \sigma^2\right) h(x)$$

Thus, by the Factorization Theorem, $T(X) = (T_1(X), T_2(X)) = \left(\bar{X}, S^2\right)$ is a sufficient statistic for $\left(\mu, \sigma^2\right)$ in this normal model.

(...end of example!)

❖ ***Definition* 4.13** (Complete Statistic)**.** Let $T(X)$ be a statistic with a family of pdfs or pmfs of $f(t|\vartheta)$ indexed by $\vartheta \in \Theta$. If the condition, $E_\vartheta(g(T)) = 0$ for all $\vartheta \in \Theta$, implies $P_\vartheta(g(T) = 0) = 1$ for all $\vartheta \in \Theta$, then the family $f(t|\vartheta)$, $\vartheta \in \Theta$ is called a *complete family of pdfs or pmfs* and $T(X)$ is called a *complete statistic*.

–––❖

Notice that completeness is a property of a family of probability distributions, not of a particular distribution. For example, if $T(X)$ has a $\mathcal{N}(0,1)$ distribution, then defining $g(t) = t$, we have that $E(g(T(X))) = E(T(X)) = 0$. But the function $g(t) = t$ satisfies $P(g(T(X)) = 0) = P(T(X) = 0) = 0$, not 1. However, this should not give the idea that the pdf is not complete because this is a particular pdf, not a family of pdfs. If X has a $\mathcal{N}(\vartheta, 1)$ distribution, $-\infty < \vartheta < \infty$, we shall see that **no function of X, except one that is 0 with probability 1 for all** ϑ, satisfies $E_\vartheta(g(X)) = 0$ for all ϑ. Thus, the family of $\mathcal{N}(\vartheta, 1)$ distributions, $-\infty < \vartheta < \infty$, is complete.

♣ **Example 4.17** (Binomial complete sufficient statistic)**.** Suppose that has T a Bin(n,p) distribution, $0 < p < 1$. Let g be a function such that $E_p(g(T)) = 0$.

Then

$$\begin{aligned} 0 = E_p(g(T)) &= \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} \\ &= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t \end{aligned}$$

for all $p, 0 < p < 1$. The factor $(1-p)^n$ is not 0 for any p in this range. Thus it must be that

$$0 = \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \sum_{t=0}^{n} g(t) \binom{n}{t} r^t$$

for all $r, 0 < r < \infty$. But the last expression is a polynomial of degree n in r, where the coefficient of r^t is $g(t)\binom{n}{t}$. For the polynomial to be 0 for all r, each coefficient must be 0. since none of the $\binom{n}{t}$ terms is 0, this implies that $g(t) = 0$ for $t = 0, 1, \ldots, n$. since T takes on the values $0, 1, \ldots, n$ with probability 1, this yields that $P_p(g(T) = 0) = 1$ for all p, the desired conclusion. Hence, T is a complete statistic.

(...end of example!)

♣ **Example 4.18** (Completeness of Normal with known Mean and Unknown variance)**.** !!! This example is from old handout!!!

Given is the family of distributions $\mathcal{N}(\mu_0, \sigma^2)$ indexed by $\sigma^2 \in \mathbb{R}^+$ with known mean $\mu_0 \in \mathbb{R}$. Let $X \sim N(\mu_0, \sigma^2)$. Select $h(X) = X - \mu_0$, then the following applies

$$E_{\sigma^2}(h(X)) = E_{\sigma^2}(X - \mu_0) = E_{\sigma^2}(X) - \mu_0 = \mu_0 - \mu_0 = 0.$$

However, this results in

$$P_{\sigma^2}(H(X) = 0) = P_{\sigma^2}(X = \mu_0) = 0,$$

because X is a constant random variable.

The family of densities of the normal distribution with known expected value and unknown variance is therefore not complete.

(...end of example!)

4.6.3 Asymptotic Considerations

❖ ***Definition* 4.14** (Asymptotically unbiased)**.** Let $\{\{T_\vartheta(X_1,...,X_n)\}_n, n \in \mathbb{N}$,be a sequence of point estimators. It's called *asymptotically unbiased* estimator of ϑ, if it holds

$$\lim_{n\to\infty} \mathrm{E}_\vartheta(T_\vartheta(X_1,...,X_n)) = \vartheta.$$

The expected value thus converges to the true parameter value for $n \to \infty$.

---❖