# An introduction to Data Analysis

Krishna N Agaram

# Contents

# Introduction

These notes were made for a full-semester course on Data Analysis And Interpretation at IIT-Bombay. The notes are to a large extent based on the slides for the course and the excellent book **An introduction to Mathematical Statistics** by **Hogg-McKean-Craig**. Any errors are solely my fault. Suggestions are always welcome, and I hope the reader finds it a useful resource in his study.

Krishna N Agaram
September 2022

# Chapter 1

# The Probability measure and Random variables

Section 1.1
## The Probability Set Function

First, some Preliminaries to get us started:

- A **random experiment** is an experiment/procedure whose outcome is uncertain.
- The **sample space** $\Omega$ associated with a random experiment is the set of all possible outcomes of the random experiment.
- An **event** in a random experiment is a subset of the sample space, i.e. a set of "preferred" outcomes of the experiment.
- The **event space** associated with a random experiment is the set of all events in consideration and is (usually) the powerset of the sample space. There are some constraints on a general event space, though, but we will not go into that here.

We now define a very general notion of a **probability function**. One of these functions (as we will later call it, the "frequentist probability function") will turn out to be the notion of probability taught in high school.

**Definition 1**
Let $\Omega$ be a sample space, and $\beta \subseteq \mathcal{P}(\Omega)$ be a corresponding event space. Then a **probability function** or **probability measure** for the random experiment is a function $P : \beta \to [0, 1]$ satisfying the following:

1. $P(\phi) = 0$
2. $P(\Omega) = 1$
3. $P\left(\bigcup A_i\right) = \sum P(A_i)$ for pairwise disjoint events $A_i$ ($A_i \cap A_j = \phi$ for $i \neq j$).

Subsection 1.1.1
### Properties of a probability function

It is easy to see that for events $A, B \in \beta$, we have

**Proposition 1** [Complements]
$P(\neg A) = 1 - P(A)$

**Proposition 2** [2-event PIE]
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

**Proposition 3** [Subsets]
$A \subseteq B \implies P(A) \leq P(B)$. In particular, we get the "triangle inequality" $P(A \cap B) \leq P(A), P(B) \leq P(A \cup B) \leq P(A) + P(B)$

Can we generalize the third defining property of probability measures to the case where the sets $A_i$ are not pairwise disjoint? Indeed!

**Theorem 1** [Principle of Inclusion-Exclusion]

For arbitrary events $A_i \in \beta, 1 \leq i \leq n$, we have

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \dots$$

Proof We induct on the number of sets $n$. We have shown the theorem for the case $n = 2$. Assuming the theorem for $n = k$, we have for $n = k + 1$ the chain of equalities

$$\begin{aligned}
P\left(\bigcup_{i=1}^{k+1} A_i\right) &= P\left(\left(\bigcup_{i=1}^{k} A_i\right) \cup A_{k+1}\right) \\
&= P\left(\bigcup_{i=1}^{k} A_i\right) + P(A_{k+1}) - P\left(\left(\bigcup_{i=1}^{k} A_i\right) \cap A_{k+1}\right) \\
&= P\left(\bigcup_{i=1}^{k} A_i\right) + P(A_{k+1}) - P\left(\bigcup_{i=1}^{k} (A_i \cap A_{k+1})\right) \\
&= \left[\sum_{i=1}^{k} P(A_i) - \sum_{1 \leq i_1 < i_2 \leq k} P(A_{i_1} \cap A_{i_2}) + \sum_{1 \leq i_1 < i_2 < i_3 \leq k} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) + \dots\right] \\
&\quad + P(A_{k+1}) - \left[\sum_{i=1}^{k} P(A_i \cap A_{k+1}) - \sum_{1 \leq i_1 < i_2 \leq k} P((A_{i_1} \cap A_{k+1}) \cap (A_{i_2} \cap A_{k+1})) + \dots\right] \\
&= \sum_{i=1}^{k+1} P(A_i) - \left(\sum_{1 \leq i_1 < i_2 \leq k} P(A_{i_1} \cap A_{i_2}) + \sum_{i=1}^{k} P(A_i \cap A_{k+1})\right) +
\end{aligned}$$

$$\left( \sum_{1 \le i_1 < i_2 < i_3 \le k} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) + \sum_{1 \le i_1 < i_2 \le k} P(A_{i_1} \cap A_{i_2} \cap A_{k+1}) \right) + \dots$$

$$= \left[ \sum_{i=1}^{k+1} P(A_i) - \sum_{1 \le i_1 < i_2 \le k+1} P(A_{i_1} \cap A_{i_2}) + \sum_{1 \le i_1 < i_2 < i_3 \le k+1} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) + \dots \right]$$

as required. **Q.E.D.** □

Section 1.2

## More definitions

Some more definitions are in order.

- **Multiple or Joint experiments**: Performing two random experiments with sample spaces $\Omega_1$ and $\Omega_2$ sequentially gives us a new random experiment with outcome $(s_1 \in \Omega_1, s_2 \in \Omega_2)$, so that its sample space is $\Omega_1 \times \Omega_2$ and so the probability function for the composite experiment maps subsets of $\Omega_1 \times \Omega_2$ to $[0, 1]$.
- **Joint probability**: Given a composite random experiment with sample space $\Omega_1 \times \Omega_2$, we define the **joint probability of events $A \subseteq \Omega_1$ and $B \subseteq \Omega_2$** to be

  $$P(A, B) := P(\{(s_1, s_2), s_1 \in A, s_2 \in B\} \subseteq (\Omega_1 \times \Omega_2)) = P(A \times B)$$

  . It is denoted $P(A, B)$ or $P(A \cap B)$ or $P(A \times B)$ where $\times$ is the cartesian product of sets. Note that the use of the $\cap$ symbol here is merely to say that event $A$ happened in the first experiment **and** event $B$ in the second, it is not an intersection of sets.
- **Conditional probability**: The conditional probability denoted $P(A|B)$ of events $A \subseteq \Omega_1, B \subseteq \Omega_2$ with $P(B) > 0$ is defined by

  $$P(A|B) := \frac{P(A, B)}{P(B)}$$

  This probability is rationalized as the probability of occurrence of the event $A$ when it is known that $B$ has occurred for sure. This is useful since we want to know the probability of $A$ happening with the current info we have, not a global probability of $A$ happening over the full sample space.

  Ex: Consider a pixel in an image. Its true color and perceived (by the camera) color are slightly different. The probability that the perceived color is one of $A \subseteq$ (set of colors) colors conditioned to (or given that) the true color being one of $B$ colors is useful to construct a perturbation or error model for the perceived image. On the other hand, the probability $P(A)$ of the color being in $A$ is heavily dependent on the true color of that pixel and is obviously not a useful metric of perturbation/error.

- **Independence**: Two events $A, B$ are said to be **independent** if the knowledge that one has occurred does not affect the chances of the other event's occurrence (the two events function "independently" of each other). In formal terms, $A$ and $B$ are independent iff

$$P(A|B) = P(A)$$

or in its more symmetric form, iff

$$P(A, B) = P(A)P(B).$$

This definition also allows for $P(A)$ or $P(B)$ to be zero, in which case the events are trivially independent.

- **Conditional independence**: Events $A$ and $B$ are conditionally independent wrt event $C$ if $A$ and $B$ become independent events when the sample space is reduced to only the outcomes for which $C$ is true, that is, when the sample space is just the set $C$ itself. In formal terms, $A, B$ are conditionally independent wrt event $C$ iff

$$P(A|B, C) = P(A|C) \iff P(A, B|C) = P(A|C)P(B|C)$$

or even more readably (here $P_C(*) = P(*|C)$) iff

$$P_C(A, B) = P_C(A)P_C(B).$$

- **Partition of a sample space:** Events $\{A_1, A_2, \ldots A_n\}$ are said to partition the sample space $\Omega$ iff $A_i \cap A_j = \phi$ for $i \neq j$ and $\vee_i A_i = \Omega$. In this case, we get what is called the **law of total probability**.

**Theorem 2** [Law of total probability]

Let $A$ be an event and $\{B_1, \ldots B_n\}$ be a partition of $\Omega$. Then

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$

Proof Note that the events $A \cap B_i$ are pairwise disjoint. If $a \in A$, then since $a \in \Omega = \bigcup_i B_i$, there exists $i$ such that $a \in B_i$, and so there exists $i$ such that $a \in A \cap B_i$, so that $A \subseteq \bigcup_i (A \cap B_i)$. But any element of the RHS is contained in $A$, so that $A = \bigcup_i (A \cap B_i)$ so that $P(A) = P(\bigcup_i (A \cap B_i)) = \sum_i P(A \cap B_i)$. The second inequality in the theorem follows from the definition of conditional probability. □

Finally, note the very useful fact that

$$\sum_i P(B_i) = P(\Omega) = 1$$

for events $\{B_i\}$ partitioning $\Omega$.

Section 1.3
# Random Variables

This has been explained very well in Hogg-Craig. The motivation for random variables has been taken from the book.

A sample space may be tedious to work with if its elements are not numbers. We would like to associate each element of the sample space with a real number and then work with the numbers themselves. Of course, the exact function we use depends heavily on the thing we want to extract from the elements of the sample space. These functions are called **random variables**.

**Definition 2** [Random Variable]
A function $X$ from the sample space $\Omega$ to the real numbers is called a **random variable**. The **space** or **range** of $X$ is the set of real numbers $\mathscr{A} = \{X(s)|s \in \Omega\}$.

Subsection 1.3.1
## The induced probability of a random variable

For a given random variable, every outcome of a random experiment now has a number associated with it. We define the function $P_X : \mathcal{P}(\mathscr{A}) \to [0, 1]$ by

$$P_X(A \subseteq \mathscr{A}) := P(\{s|s \in \Omega \text{ and } X(s) \in A\}).$$

It can be verified that $P_X$ is a valid probability measure on the event space $\mathcal{P}(\mathscr{A})$ - hence it is called the *induced probability*. It is sometimes denoted $\Pr(X \in A)$.

From here on, we denote it as $P(A)$, with no ambiguity to the probability function $P$ associated with the random experiment since the argument to the former $P$ is a subset of the space $\mathscr{A}$ of $X$ while the latter $P$ takes in a subset of the actual sample space $\Omega$. If $A$ is a singleton $\{a\}$, we use $P(a)$ or $P(X = a)$ to denote it instead of $P(\{a\})$ or $P(X \in \{a\})$. Similar notation such as $P(a < X < b)$ is used when the set $A$ is $\{x \in \mathscr{A}|a < x < b\}$.

Subsection 1.3.2
## Types of RVs

**Definition 3** [Discrete random variable]
The random variable $X$ is said to be **discrete** if its space $\mathscr{A}$ is **countable**.

So for discrete $X$ we can write the set $\{s \in \Omega|X(s) \in A\}$ as a disjoint union

$$\{s \in \Omega|X(s) \in A\} = \bigcup_{x \in A} \{s \in \Omega|X(s) = x\}$$

and now since the RHS is a finite union, by the 3rd probability measure axiom we get

$$\Pr(X \in A) = \sum_{x \in A} P(X = x) = \sum_{x \in A} P(x)$$

.

$P(x) = P(X = x)$ is also written $f(x)$, for reasons that will shortly follow.

**Definition 4** [Continuous random variable]
The random variable $X$ is said to be **continuous** if it is not discrete, i.e. its space $\mathscr{A}$ is **uncountable**. It must also satisfy (to be useful) the existence of a function $f(x)$ such that for any $A \subseteq \mathscr{A}$, Pr $(X \in A)$ is given by

$$\text{Pr}\ (X \in A) = P(\{s \in \Omega | X(s) \in A\}) = \int_A f(x)dx$$

Note that the above is not implied by the 3rd probability measure axiom since the union over $x \in A$ is not finite.

Subsection 1.3.3
## The CDF of a random variable

**Definition 5** [Cumulative distribution function for a random variable]
Consider a random variable $X$ with space $\mathscr{A}$. Then the function defined by

$$F(x) := \text{Pr}\ (X \leq x)$$

for all $x \in \mathscr{A}$ is called the **Cumulative Distribution Function** or **CDF** of the random variable $X$.

For $X$ discrete,

$$F(x) = \text{Pr}\ (X \leq x) = \text{Pr}\ (x \in A = e \in \mathscr{A} | e \leq x) = \sum_{e \in A} P(e) = \sum_{e \in \mathscr{A}, e \leq x} P(e)$$

For $X$ continuous,

$$F(x) = \text{Pr}\ (X \leq x) = \text{Pr}\ (x \in A = e \in \mathscr{A} | e \leq x) = \int_A P(e)de = \int_{\{e \in \mathscr{A}, e \leq x\}} P(e)de$$

To make our lives much easier, we now extend the space of continuous random variables to $\mathbb{R}$ by defining $f(x)$ to be 0 on $\mathbb{R} - \mathscr{A}$. So the CDF simplifies to

$$F(x) = \int_{-\infty}^{x} P(x)dx$$

### Properties of the CDF

It is easy to see that for any $x, F(x) = \text{Pr}\ (X \leq x) \in [0, 1]$, since $P$ is a probability measure. Here are some other properties of the CDF.

**Lemma 1**
$F(x)$ is increasing on $\mathscr{A} \subseteq \mathbb{R}$.

Proof | Just note that $F(b) - F(a) = \text{Pr}\ (a < X \leq b) \geq 0 \ \forall \ a < b$ as needed. □

The below properties are very intuitive and are stated without rigorous proof.

**Proposition 4**

$$\lim_{x \to -\infty} F(x) = P(\phi) = 0.$$

Similarly,

$$\lim_{x \to \infty} F(x) = P(\{s \in \Omega | X(s) \in \mathbb{R}\}) = P(\Omega) = 1$$

Combined with the lemma above, this provides an alternate proof of $F(x) \in [0, 1] \ \forall \ x$.

**Proposition 5** [right continuity of a CDF]
For any $x$, $F(x)$ is right-continuous at $x$, i.e.

$$\lim_{\epsilon \to 0^+} F(x + \epsilon) = F(x) \ \forall \ x$$

The above proposition is very intuitive: We have $F(a + \epsilon) - F(a) = \Pr(a < X \leq a + \epsilon)$ which is expected to go to $\Pr(X \in \phi) = 0$ as $\epsilon$ goes to 0.

Note that left continuity is not necessary, though: $F(a) - F(a - \epsilon) = \Pr(\{a - \epsilon < X \leq a\})$ and this set need not be empty. For example, if $P(X = a) > 0$ the RHS will be nonzero for any $\epsilon > 0$. This comes up quite often for discrete random variables where the CDF is (left-)discontinuous at all points in the space of $X$.

Subsection 1.3.4

## The probability mass/density function

Consider a discrete random variable $X$. Suppose we know exactly the function $f(x) = P(X = x)$ for every $x \in \mathscr{A}$. Then for any set $A \subseteq \mathscr{A}$, we have

$$P_X(A) = \sum_{x \in A} f(x)$$

So the function $f(x)$ characterizes the probability measure of the random variable.

Now consider a continuous random variable. Suppose the function $f(x)$ in **Definition 4** is known for every $x \in \mathbb{R}$. Then $P_X(A)$ is known for any subset $A \subseteq \mathbb{R}$ by

$$P(A) = \int_A f(x) dx$$

So here too, $f(x)$ characterizes the probability measure $P_X$ of the random variable.

We call $f(x)$ the **Probability Density function** associated with $X$. It is also called the **distribution of** $X$.

Why the word 'density'? $f(x)$ is like a derivative of the probability, given that $f(x)$ is integrated or summed to give the probability. And derivatives are called densities. Why the word 'distribution'? $f(x)$ can be viewed as the 'contribution' of $x$ (In the discrete case, it tells us the probability of $X$ being $x$. In the continuous case, $f(x)dx$ tells us the probability of the random variable being in the range $(x, x + dx]$). In other words, the values of $f(x)$ tell us how the contributions of each $x$ are distributed over the space of the random variable.

It is useful to work with $f(x)$ from which the probability measure can be easily derived. Further, $f$ is not under any significant constraint - *any* $f(x)$ with its sum/integral equalling 1 gives rise to a valid probability measure $P_X$. So we might as well study these functions instead of directly looking at general valid probability measures $P_X$ over the powerset of $\mathscr{A}$.

Subsection 1.3.5

## Computing the pdf from the cdf

It is easy to see that $f(x)$ can be (in all cases covered here) computed from $F(x)$. For example, for a discrete rv on the integers,

$$f(x) = P(X = x) = P(X \geq x) - P(X \geq x - 1) = F(x) - F(x - 1)$$

for any $x$.

And for the reals, this finite difference is replaced by a derivative, in particular

$$f(x) = D_x F(x) \ \forall \ x$$

.

*Remark* From the above, we see that the cdf $F(x)$ can also be used to describe a distribution perfectly. Moreover, *any* function satisfying the properties of the cdf mentioned in the previous section is a valid cdf. The only caveat here is that the pdf has far more freedom of choice, the cdf is a bit more restricted.

Finally, before we move on to the next section, one more definition:

**Definition 6** [Support of a random variable]
The **support** of an rv $X$ is defined to be the set

$$\mathcal{S}(X) := \{x \in \mathscr{A} | f(x) > 0\}$$

Note that this is in general **NOT** the same as $\mathscr{A}$ itself. $\mathscr{A}$ is the set of all values in $\mathbb{R}$ that subsets of the sample space get mapped to by $X$. Some of those events could have a zero probability of happening ($P$ maps them to 0) and so these events' $x$ values might not be in $\mathcal{S}(X)$. It is true, however, from the definition, that $\mathcal{S}(X) \subseteq \mathscr{A}$. The utility of the support of a random variable will be seen in the section on Likelihood functions later on.

# Chapter 2

# Some common distributions

A **distribution** is, as already mentioned, a pdf $f(x)$ for a random variable $X$. In this chapter, we look at some common distributions and their properties.

Section 2.1
## The Uniform Distribution

Consider a discrete RV $X$ with a finite space $\mathscr{A}$. Consider the distribution $f(x)$ for this random variable given by

$$f(x) = \frac{1}{|\mathscr{A}|} \ \forall \ x \in \mathscr{A}$$

This function $f$ is called the **uniform distribution** for the random variable $X$.

For a uniform distribution, we have

$$P(A) = \frac{|A|}{|\mathscr{A}|}$$

for any event $A \subseteq \mathscr{A}$. This is our usual notion of the probability of an event when all outcomes of the rv are equally likely! The number $|A|$ is called the number of elements favorable to the event $A$.

For a continuous random variable, the uniform distribution is found by replacing the size of $\mathscr{A}$ by the length or measure of the interval (or union of intervals) $\mathscr{A}$, and similarly replacing $|A|$ by the length of $A$. Single points thus are given a zero probability - because there are infinitely many of them!

Section 2.2
## The Bernoulli distribution

The simplest distribution out there. It models the **probability of getting a success in one trial of an experiment**. The Bernoulli random variable is either $0$ or $1$, and

the pdf is given by

$$f(x) = \begin{cases} p & x = 0 \\ 1-p & x = 1 \\ 0 & \text{otherwise} \end{cases}$$

where $p$ is said to be a **parameter** for the distribution - each value of the parameter gives a different Bernoulli distribution - though all of them are similar enough. $f(x)$ can be more compactly written as

$$f(x) = p^x(1-p)^{1-x}$$

for $x \in \{0, 1\}$.

Section 2.3

## The Binomial Distribution

This models the **number of successes in $n$ trials of a Bernoulli experiment**. The pdf is taken to be the function

$$f(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

It is easily seen to be a valid (sums to 1) pdf. This function $f = f(x; n, p)$ is called the binomial distribution in parameters $n$ and $p$ for the binomial random variable $X$.

Notice that the case $n = 1$ is exactly the Bernoulli distribution as expected since the Binomial distribution experiment is just $n$ Bernoulli experiments.

A generalization of this distribution is the **Multinomial distribution**, see Hogg-Craig Sec 3.1 for details. The adventurous reader could also look at the **negative Binomial distribution**, which generalizes the geometric distribution described below.

Section 2.4

## The Geometric Distribution

This models the **number of trials required to get a success**. The pdf is defined like so:

$$f(x) = \begin{cases} (1-p)^{x-1}p & x = 1, 2, \ldots \infty \\ 0 & \text{otherwise} \end{cases}$$

It is easily seen to be a valid (sums to 1) pdf. $f(1), f(2), \ldots$ form a G.P., hence the name of this distribution. As usual, $f(x; p)$ is called the geometric distribution in the parameter $p$.

The cdf for this distribution is given by

$$F(x) = \sum_{i=1}^{x} f(i) = p\sum_{i=1}^{x}(1-p)^{i-1} = 1 - (1-p)^x$$

Since in general $x$ could be real in the definition, we say

$$F(x) = \begin{cases} 1 - (1-p)^{\lfloor x \rfloor} & x \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Subsection 2.4.1

## The Memorylessness property

A key property of the geometric distribution is that it is ***memoryless***.

**Definition 7** [Memorylessness]
A distribution on the non-negative integers/reals is said to be **memoryless** if for all $x, y \geq 0$, we have

$$P(X > x + y \mid X > y) = P(X > x)$$

or

$$P(X > x + y) = P(X > x)P(X > y)$$

or in terms of $G(x) = 1 - F(x)$,

$$G(x + y) = G(x)G(y) \; \forall \; x, y \geq 0$$

$x$ and $y$ are usually interpreted as times - so if the probability of an outcome at time t is modeled by a memoryless distribution, what this means is that the probability of an outcome is independent of previous history of the experiment.

Now we show a quite remarkable result. We prove the real number version of this a little later.

**Theorem 3**

Let $X$ be an rv on $\{0, 1, \dots\}$ and $f$. be a pdf for $X$. Then

$$f \text{ is memoryless} \iff f \text{ is a geometric distribution}$$

Proof A geometric distribution has $G(x) = 1 - F(x) = (1-p)^x$. Clearly, $G$ is multiplicative and so the distribution is memoryless. Now, if a distribution is memoryless, we must have $G(x + y) = G(x)G(y)$ for all $x, y \in \mathbb{N}_{\geq 0}$ and it is easy and well-known that $G(x) = a^x$ for $a \geq 0$ is the only solution to this equation. Since $0 \leq G(x) \leq 1 \; \forall \; x$ , $0 < a < 1 (a \neq 0$ since otherwise $F(x)$ doesn't go to 0 at $-\infty$ and $a \neq 1$ since otherwise $F(x)$ doesn't go to 1 at $\infty)$. So $f(x) = F(x) - F(x-1) = a^{x-1}(1-a)$. Replacing $a \in (0, 1)$ with $1 - p$, $p \in (0, 1)$ we get the result. □

Section 2.5

# The Poisson Distribution

Many experiments have a tiny chance of success - a fused bulb, or a car crash, for example. Across a very large number of experiments, though, the number of these

"successes" is finite. This distribution models the **number of successes that occur in a long time duration (aka after a large number of experiments)**. This is computed as a function of a parameter $\lambda$ which, as we show later, turns out to be the mean number of successes obtained.

The pdf is defined like so:

$$f(x;\lambda) = \begin{cases} \frac{\lambda^x}{x!}\exp(-\lambda) & x = 0, 1, \dots \\ 0 & \text{otherwise} \end{cases}$$

The $e^{-\lambda}$ factor is so that the pdf sums to 1.

The reason why this apparently out-of-the-blue pdf models what we said it does follows from the interpretation of the Poisson distribution as the limit of the binomial distribution in the case of an infinitesimal $p$ and a large $n$, with $\lambda := pn$ being finite. Indeed, as $n \to \infty$, note that

$$\begin{aligned} f_{\text{binom}}(x; n, \frac{\lambda}{n}) &= \frac{n!}{(n-x)!x!} \cdot \frac{\lambda^x}{n^x} \cdot (1 - \frac{\lambda}{n})^n \cdot (1 - \frac{\lambda}{n})^{-x} \\ &\to \frac{\lambda^x}{x!} e^{-\lambda} \cdot \frac{n!}{n^x(n-x)!} \cdot 1^{-x} \\ &\to \frac{\lambda^x}{x!} e^{-\lambda} \cdot 1 \\ &= f_{\text{poisson}}(x;\lambda) \end{aligned}$$

where $x \in \{0, 1, \cdots, n\} \to \{0, 1, \dots \infty\}$. This gives us the Poisson distribution.

Subsection 2.5.1

## Poisson Thinning

Consider a Poisson random variable $X$. Now suppose we run the experiment a large number of times and get a finite number $x$ of successes. Now, we select each one with probability $p$ according to a binomial distribution. Let the number of chosen successes be $y$. Hence we define the joint distribution $P(X, Y)$ by

$$P(X = x, Y = y) = P_{\text{poisson}}(X = x) \cdot P_{\text{binom}}(y;\ x, p)$$

for all non-negative numbers $x, y$. Clearly $P(X, Y)$ is 0 when $y > x$.

Now consider the marginal distribution $P_Y(Y = y) := \sum_x P(x, y)$. It is reasonable to expect that this distribution has a mean of $\lambda \cdot p$, since on average (read: mean of ) the number of successes is $\lambda$ and on average we select a proportion $p$ of these $\lambda$, giving $p\lambda$ chosen successes on average as needed.

Can we say something more? Think about it, suppose we had a poisson experiment with an average (as usual, read: mean) number of successes $\lambda = \lambda p$, this does seem similar to the marginal distribution of $Y$. Remarkably, it is indeed so. Notice that the above heuristic explanation isn't enough - the fact that the successes are chosen *binomially* makes all the difference and gives us the result. The heuristic viewpoint was just to make it seem reasonable. Anyway, here's the proof.

$$\begin{aligned}
P_Y(y) &= \sum_{x\geq 0} e^{-\lambda}\frac{\lambda^x}{x!}\cdot P_{\text{binom}}(y;\ x,p)\\
&= \sum_{x=y}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!}\cdot\frac{x!}{y!(x-y)!}p^y(1-p)^{x-y}\\
&= e^{-\lambda}\frac{(\lambda p)^y}{y!}\cdot\sum_{x=y}^{\infty}\frac{\lambda^{x-y}}{(x-y)!}(1-p)^{x-y}\\
&= e^{-\lambda}\frac{(\lambda p)^y}{y!}\cdot e^{\lambda(1-p)}\\
&= e^{-\lambda p}\frac{(\lambda p)^y}{y!} = P_{\text{poisson}}(y;\lambda p)
\end{aligned}$$

as needed.

Section 2.6

## The exponential distribution

Just like the geometric distribution looked at how long it would take for the binomial experiment to return at least one success, here we look at how long (i.e. number of experiments required in units of N, where N is the (large, more precisely infinitely large) number of experiments done in the poisson experiment) it takes to return a success from the poisson experiment. Clearly, this will be modeled by a continuous random variable. Let's see what the pdf would look like. The cdf(or in particular, 1-cdf) is easier to work with here. $G_{\text{exp}}(x) = P_{\text{exp}}(X > x)$ indicates that the first $xN$ experiments didn't give a success. How likely is this? It's exactly the probability of getting $0$ successes over $N$ experiments in a poisson with $\lambda = \lambda x$. So we take

$$G_{\text{exp}}(x) := P_{\text{poisson}}(0;\lambda x) = \exp(-\lambda x)\frac{(\lambda x)^0}{0!} = e^{-\lambda x}$$

Notice that $F = 1 - G$ is indeed a valid cdf!(This can be easily understood from the definition of $G$ as $P_{\text{exp}}(X > x)$ by noting that clearly $0 \leq G \leq 1$ and that as $x \to 0$, $G \to 1$ and as $x \to \infty$, $G$ should obviously go to $0$ - and it does indeed)

From the cdf it now easily follows that

$$f(x) = P_{\text{exp}}(x) = -G'_{\text{exp}}(x) = \lambda e^{-\lambda x},\ x \geq 0.$$

Since $F$ was a valid cdf, $f$ is a valid pdf. Finally, the exponential distribution is named for the fact that $f$ is an exponential.

---

Notice how similar the exponential is to the geometric. In fact, the derivation of

their pdfs is essentially identical -

$$G_{\text{geometric}}(x;p) = P_{\text{binom}}(0;x,p) = \binom{x}{0}p^0(1-p)^{x-0} = (1-p)^x$$

so that

$$f_{\text{geometric}}(x) = -\Delta_1 G(x) = G(x-1) - G(x) = p(1-p)^{x-1}$$

This is exactly identical to the derivation for the exponential!

In fact, we can define the exponential from the geometric, just like Poisson from the binomial!

**Theorem 4**

The exponential cdf $F_{\text{exp}}(x;\lambda)$ is the limit of the geometric cdf $F_{\text{geometric}}(nx; p = \frac{\lambda}{n})$ as $n \to \infty$.

Proof We have

$$G_{\text{geometric}}\left(nx; p = \frac{\lambda}{n}\right) = \left(1 - \frac{\lambda}{n}\right)^{nx} \to \exp(-\lambda x) = G_{\text{exp}}(x;\lambda)$$

so by subtracting both sides from 1 we get the result. Note that $x \in \mathbb{R}_{\geq 0}$ here. $nx$ can be plugged into the geometric since $nx$ is approximately a non-negative integer for $n$ large enough. □

Basically, since the Poisson variable had a "binomial" probability $p = \frac{\lambda}{n}$, the probability of the first success taking longer than $nx$ experiments is exactly $G_{\text{geometric}}\left(nx; p = \frac{\lambda}{n}\right)$, which, in the limit, is precisely the probability that the time it takes in "units" of $n$ to get the first Poisson success is $> x$, and this is exactly what we are trying to model with $G_{\text{exp}}(x;\lambda)$!

Now for a nice surprise - it shouldn't come as a surprise, really - the exponential distribution, just like the geometric, is memoryless. This is easily verified from $G_{\text{exp}}$.

Moreover, the exponential is the *only* distribution of a continuous random variable that is memoryless. This readily follows from the fact that a bounded solution $G$ to $G(x+y) = G(x)G(y)$ over the non-negative reals must be an exponential $e^{ax}$. The condition $a < 0$ follows from $0 \leq G \leq 1$ and the limit as $x \to \infty$, just like in the discrete case.

Section 2.7

## The Normal distribution

Well, here we are. The king of distributions. The normal distribution is the distribution of ideal-gas velocities, of a point-source diffusion, and, as we will see in the Central limit theorem, much more. In fact, it is also the limit(in a certain range of $x$) of the binomial distribution!

The pdf is defined for $x \in \mathbb{R}$ by

$$f(x;\mu,\sigma) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$$

where $t := \frac{x-\mu}{\sigma}$ is the normalized normal random variable. This pdf is usually denoted $G(\mu,\sigma)$ or $\mathcal{N}(\mu,\sigma)$. This distribution is also called the **Gaussian** distribution.

Subsection 2.7.1

## The Gaussian as the limit of the Binomial

**Theorem 5**

Let $P_{\text{binom}}(x;n,p)$ be a binomial distribution. Suppose we set $n$ large, but keep $p$ and $q = 1-p$ strictly nonzero and finite - so that $np$ and $nq$ are also large. Then

$$P_{\text{binom}}(x;n,p) \approx P_{\text{Gaussian}}(x;\mu = np, \sigma = \sqrt{npq})$$

for $x$ being at most a constant number of standard deviations from $np$, or $\delta = x - np \leq \mathcal{O}(1)\sqrt{npq} = \mathcal{O}(\sqrt{n})$.

Proof We use **Stirling's formula** for the factorial:

$$n! = \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$$

We first have

$$P_{\text{binom}}(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$
$$\implies \log P_{\text{binom}}(x;n,p) = \log n! - \log x! - \log(n-x)! + x\log p + (n-x)\log(1-p)$$

Now

$$\begin{aligned}\log n! &= n\log n - n + \frac{1}{2}\log 2\pi n + \log\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)\\ &= n\log n - n + \frac{1}{2}\log 2\pi n + \mathcal{O}\left(\frac{1}{n}\right)\end{aligned}$$

where in the second step we used $\log(1+x) = x - \frac{x^2}{2} + \mathcal{O}(x^3)$ for $x \in (-1,1)$
(We will use this again later in the proof).
The main equation then simplifies (where we set $q = 1-p$) to

$$\begin{aligned}\log P_{\text{binom}}(x;n,p) &= n\log n - x\log x - (n-x)\log(n-x) + x\log p + (n-x)\log q\\ &\quad - \log\sqrt{2\pi} + \frac{1}{2}\log\left(\frac{n}{x(n-x)}\right) + \mathcal{O}\left(\frac{1}{n}\right)\\ &= (x+n-x)\log n + x\log\frac{p}{x} + (n-x)\log\frac{q}{n-x}\end{aligned}$$

$$
\begin{aligned}
&- \log \sqrt{2\pi} + \frac{1}{2} \log \left( \frac{np}{x} \cdot \frac{nq}{n-x} \right) - \frac{1}{2} \log (npq) + \mathcal{O}\left(\frac{1}{n}\right) \\
&= x \log \frac{np}{x} + (n-x) \log \frac{nq}{n-x} \\
&- \log \left(\sqrt{2\pi npq}\right) + \frac{1}{2} \log \left( \frac{np}{x} \cdot \frac{nq}{n-x} \right) + \mathcal{O}\left(\frac{1}{n}\right)
\end{aligned}
$$

where in the first step we used the fact that $\mathcal{O}\left(\frac{1}{x}\right) = \mathcal{O}\left(\frac{1}{n-x}\right) = \mathcal{O}\left(\frac{1}{n}\right)$ which follows because $x, n-x$ are $\mathcal{O}(n)$ (since $p$ and $q$ are nonzero and finite).
More simplification incoming. We have

$$
\log \frac{x}{np} = \log \frac{np+\delta}{np} = \log \left(1 + \frac{\delta}{np}\right)
$$

and analogously

$$
\log \frac{n-x}{nq} = \log \frac{nq-\delta}{nq} = \log \left(1 - \frac{\delta}{nq}\right).
$$

Then we have:

$$
\begin{aligned}
&x \log \frac{np}{x} + (n-x) \log \frac{nq}{n-x} = -\left[ (np+\delta) \log \left(1 + \frac{\delta}{np}\right) + (nq-\delta) \log \left(1 - \frac{\delta}{nq}\right) \right] \\
&= -\left[ (np+\delta) \frac{\delta}{np} + (nq-\delta) \frac{-\delta}{nq} - \frac{(np+\delta)}{2} \left(\frac{\delta}{np}\right)^2 - \frac{(nq-\delta)}{2} \left(\frac{-\delta}{nq}\right)^2 + \mathcal{O}\left(n \cdot \frac{\delta^3}{n^3}\right) \right] \\
&= -\left[ 0 \cdot \delta + \delta^2 \left( \frac{1}{np} + \frac{1}{nq} - \frac{1}{2}\left(\frac{1}{np} + \frac{1}{nq}\right)\right) + \mathcal{O}\left(\frac{\delta^3}{n^2}\right) \right] \\
&= -\frac{\delta^2}{2npq} + \mathcal{O}\left(\frac{\delta^3}{n^2}\right)
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\log \left( \frac{np}{x} \cdot \frac{nq}{n-x} \right) &= \log \left(\frac{np}{x}\right) + \log \left(\frac{nq}{n-x}\right) \\
&= -\left[ \log \left(1 + \frac{\delta}{np}\right) + \log \left(1 - \frac{\delta}{nq}\right) \right] \\
&= -\frac{\delta}{n} \left( \frac{1}{p} - \frac{1}{q} \right) + \mathcal{O}\left(\frac{\delta^2}{n^2}\right) \\
&= \mathcal{O}\left(\frac{\delta}{n}\right)
\end{aligned}
$$

Hence

$$
\begin{aligned}
\log P_{\text{binom}}(x; n, p) &= -\frac{\delta^2}{2npq} + \mathcal{O}\left(\frac{\delta^3}{n^2}\right) - \log \left(\sqrt{2\pi npq}\right) + \frac{1}{2}\mathcal{O}\left(\frac{\delta}{n}\right) + \mathcal{O}\left(\frac{1}{n}\right) \\
&= -\frac{\delta^2}{2npq} - \log \left(\sqrt{2\pi npq}\right) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}
$$

since $\delta = \mathcal{O}(\sqrt{n})$, so that indeed, as $n \to \infty$,

$$\log P_{\text{binom}}(x; n, p) \to -\frac{\delta^2}{2npq} - \log\left(\sqrt{2\pi npq}\right)$$

or

$$P_{\text{binom}}(x; n, p) \to \mathcal{N}(np, \sqrt{npq})$$

as needed. □

The interested reader can check out some relatives of the Gaussian distribution - **The $\chi^2-$ and $\Gamma-$distributions** - from Hogg-Craig, Sec 3.3.

# Interlude: Multiple Random variables

Section 2.8

## Functions of a single random variable

**Definition 8**
Let $f$ be a function from $\mathbb{R}$ to $\mathbb{R}$, and $X$ be a random variable. Then the random variable (denoted) $f(X)$ is called a function of random variable $X$ and is defined by

$$(f(X))(s) := f(X(s))$$

for $s \in \Omega$. In other words, if rv $X$ maps $s$ to $x$, then rv $f(X)$ maps $s$ to $f(x)$.

So a function of a random variable $X$ is just another random variable that takes elements in $\Omega$ to $\mathbb{R}$ via the *composed* function $f \circ X$. Also, since a function of a random variable is a random variable, we can have functions of functions of random variables and so on.

For example, if $X$ is a random variable, then $X^2$, $\sin(X)$, $e^{\iota X}$ are all functions of random variable $X$.

Section 2.9

## Multiple Random Variables

Consider two random variables $X$ and $Y$, both on the same space $\Omega$ - but possibly with different pdfs. Then their **sum** $Z$, denoted $X + Y$ is a *random variable on* $\Omega$ defined by

$$Z(s) := X(s) + Y(s) \; \forall \; s \in \Omega$$

The generalization to $n$ variables on $\Omega$ is immediate:

$$Z(s) := \sum_{i=1}^{n} X_i(s) \; \forall \; s \in \Omega$$

Suppose $X$ and $Y$ were from different sample spaces altogether - say from $\Omega_1$ and $\Omega_2$ respectively. Then their sum $Z$ is a random variable from $\Omega_1 \times \Omega_2$ to $\mathbb{R}$ defined by

$$Z(s_1, s_2) := X(s_1) + Y(s_2) \; \forall \; s_1 \in \Omega_1, s_2 \in \Omega_2$$

.

You can probably guess where we're going next.

**Definition 9**
Consider $n$ random variables $X_i$ on sample spaces $\Omega_i$ with arbitrary pdfs. Their sum $Z$ is a random variable on the **joint sample space** $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$, defined by

$$Z(s_1, s_2, \ldots, s_n) := \sum_{i=1}^{n} X_i(s_i) \; \forall \; s_1 \in \Omega_1, \ldots, s_n \in \Omega_n$$

Section 2.10

## Functions of Multiple random variables

Why only the sum? We can define the product of random variables as well, and in general any function you'd like on the random variables. The definition is a natural extension of that for a single random variable.

**Definition 10**
Let $f$ be a function from $\mathbb{R}^n$ to $\mathbb{R}$, and $X_1, X_2, \ldots, X_n$ be $n$ random variables on the sample spaces $\Omega_1, \Omega_2, \ldots, \Omega_n$ respectively. Then the **random variable** (denoted) $f(X_1 \ldots X_n)$ on the joint sample space $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$ is called a function of the random variables $X_i$ and is defined by

$$(f(X_1, \ldots X_n))(s_1, \ldots s_n) := f(X_1(s_1), \ldots X_n(s_n))$$

for $s_i \in \Omega_i$. In other words, if rv $X_i$ maps $s_i \in \Omega_i$ to $x_i \in \mathbb{R}$, then rv $f(X_1, \ldots X_n)$ maps $s_1, \ldots s_n \in \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$ to $f(x_1, \ldots x_n) \in \mathbb{R}$.

Again, note that there is absolutely no restriction on the random variables $X_i$ - some may be discrete, some continuous, some from the same distribution, some from different distributions, some independent, some dependent, doesn't matter.

# Chapter 3

# Measures of a distribution

Section 3.1

## Expectation of a random variable

The expectation function of a random variable $X$ is defined to be

$$\mathbb{E}[X] := \sum_{x \in \mathscr{A}} xP(x)$$

or for a continuous distribution, by

$$\mathbb{E}[X] = \int_{\mathscr{A}} xf(x)\mathrm{d}x$$

It is also called the mean or first moment or center of mass of the distribution. Note that it is not an example of a function of the random variable $X$ as defined above, rather it takes in as argument a random variable, and outputs a number, so in a sense, it is a "function" of the random variable $X$. It is not a map from $\Omega$ to $\mathbb{R}$ though.

Why is the expected value useful? Let's see.

After a long wait, here's the probability measure we almost always work with to model phenomena in real life. Everything should click now!

**Definition 11** [The frequentist probability measure]
The probability measure $P(E \subseteq \Omega)$ is assigned to be the proportion of the number of times '$E$ occurs'(i.e. the outcome of the experiment/phenomenon is $s \in E$) when the random experiment is repeated an infinite number of times.

For the frequentist probability measure, the average value taken on by the random variable $X$ is

$$\begin{aligned}\langle X \rangle &= \sum_{x \in \mathscr{A}} x \cdot (\text{proportion of times X(outcome) = } x) \\ &= x\mathbb{P}(x) \\ &= \mathbb{E}[X]!\end{aligned}$$

And hence, the expectation of the random variable is (when all $P$'s are frequentist) just its average value or mean! And of course, the mean of a distribution is very useful!

- Another (very useful) formula for the expectation:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x\in\mathscr{A}} xP(x)\\ &= \sum_{x\in\mathscr{A}} xP(\{s\in\Omega : X(s)=x\})\\ &= \sum_{x\in\mathscr{A}}\sum_{s\in\Omega, X(s)=x} xP(s)\\ &= \sum_{x\in\mathscr{A}}\sum_{s\in\Omega, X(s)=x} X(s)P(s)\\ &= \sum_{s\in\Omega} X(s)P(s)\end{aligned}$$

Subsection 3.1.1

## Expectation of some common distributions

The proofs are straight from the definition and are left as an exercise.

- Bernoulli: $p$
- Binomial: $np$
- Geometric: $\frac{1}{p} \geq 1$
- Poisson: $\lambda$ - as we had earlier claimed.
- Exponential: $\frac{1}{\lambda}$ - very similar to the geometric.
- Gaussian: $\mu$ - as expected from the graph.
- Uniform in $[a,b]$: $\frac{a+b}{2}$ - as expected from the graph.

Subsection 3.1.2

## Expectation of a function of random variables

Let $f : \mathscr{A}_1 \times \mathscr{A}_2 \times \cdots \times \mathscr{A}_n \to \mathbb{R}$ be a function of $n$ random variables $X_i$ on $\Omega_i$ with space $\mathscr{A}_i$. Then $f$ itself is a random variable and so we have from the definition

$$\begin{aligned}\mathbb{E}[f] &= \sum_{t\in\text{Range}(f)} tP(f=t)\\ &= \sum_{t\in\text{Range}(f)} tP\Big(\{(x_1,x_2,\cdots x_n) : f(x_1,x_2,\cdots x_n)=t\}\Big)\\ &= \sum_{t\in\text{Range}(f)}\sum_{\substack{x_1,x_2,\ldots,x_n\\ f(x_1,x_2,\cdots x_n)=t}} tP(x_1,x_2,\ldots,x_n)\end{aligned}$$

$$= \sum_{t \in \text{Range}(f)} \sum_{\substack{x_1, x_2, \ldots, x_n \\ f(x_1, x_2, \cdots x_n) = t}} f(x_1, x_2, \ldots, x_n) P(x_1, x_2, \ldots, x_n)$$
$$= \sum_{x_1, x_2, \ldots, x_n} f(x_1, x_2, \ldots, x_n) P(x_1, x_2, \ldots, x_n)$$

Here $P$ is the joint probability measure, of course.

### A most useful special case

Consider $n$ **independent** random variables $x_1, \ldots, x_n$ and the following function on $\mathbb{R}^n$:

$$f(x_1, \ldots, x_n) = x_1$$

We would like to find $\mathbb{E}[f]$. What is the expectation over? Over all possible values of $x_i \in \mathbb{R}$. So since the variables are continuous here, $\mathbb{E}[f] = \mathbb{E}[x_1]$ is an $n$-fold integral over all the $x_i$'s:

$$\begin{aligned}
\mathbb{E}[x_1] &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1, x_2, \ldots, x_n) P(x_1, x_2, \ldots, x_n) dx_1 \ldots dx_n \\
&= \mathbb{E}[x_1] = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} x_1 \left[ \prod_{i=1}^{n} P(x_i) \right] dx_1 \ldots dx_n \\
&= \int_{\mathbb{R}} x_1 P(x_1) dx_1 \times \left( \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left[ \prod_{\substack{1 \le j \le n \\ j \ne i}} P(x_j) \right] dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n \right) \\
&= \mathbb{E}_{x_1}[x_1] \cdot \prod_{\substack{1 \le j \le n \\ j \ne i}} \left[ \int_{\mathbb{R}} P(x_j) dx_j \right] \\
&= \mathbb{E}_{x_1}[x_1] \cdot 1^{n-1} \\
&= \mathbb{E}_{x_1}[x_1]
\end{aligned}$$

The latter term in the third step neatly splits into $n-1$ independent integrals, each just being the integral of a pdf - which is precisely 1 - so we only have to worry about the first term, which is the expectation of $f = x_1$ over *only the variables $f$ depends on* (which is just $x_1$ here). This is in fact a general principle (as one can easily see from the above example) - an expectation of a function of only a few random variables taken over the ranges of many **independent** random variables is same as the expectation calculated *assuming the expectation is over only the variables contained in the function.* This is quite useful, especially to compute the expectations of estimators (see Chapter 4)

Subsection 3.1.3

## Linearity of expectation

Let $X, Y : \Omega \to \mathbb{R}$ be random variables, and let $Z : \Omega \to \mathbb{R}$ be defined by $Z := X + Y$. Then

$$\mathbb{E}[X + Y] = \mathbb{E}[Z] = \mathbb{E}[X] + \mathbb{E}[Y]$$

The proof is quite easy, just note

$$\mathbb{E}[Z] = \sum_{s\in\Omega} \Big(X(s) + Y(s)\Big)P(s) = \mathbb{E}[X] + \mathbb{E}[Y]$$

This is called the linearity of expectation - basically, expectation is linear in $X(s)$.

Notice that no assumptions have been taken on $X, Y$ - other than that they are both on $\Omega$ (we shall lift this soon) - they could be dependent, functions of each other, or even identical. The result is always true.

The same proof works to show that for variables $X_1, \dots X_n$ on $\Omega$,

$$\boxed{\mathbb{E}\left[\sum_i X_i\right] = \sum_i \mathbb{E}[X_i]} \quad (\star)$$

We can further generalize. Suppose $X : \Omega_1 \to \mathbb{R}$ and $Y : \Omega_2 \to \mathbb{R}$. Then the sum $Z = X + Y : \Omega_1 \times \Omega_2 \to \mathbb{R}$. Now we have

$$\begin{aligned}
\mathbb{E}[Z] &= \sum_{s_1,s_2} \Big(X(s_1) + Y(s_2)\Big)P(s_1, s_2) \\
&= \sum_{s_1} X(s_1)\left(\sum_{s_2} P(s_1, s_2)\right) + \sum_{s_2} Y(s_2)P_Y(s_2)\left(\sum_{s_1} P(s_1, s_2)\right) \\
&= \sum_{s_1} X(s_1)P_X(s_1) + \sum_{s_2} Y(s_2)P_Y(s_2) \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

in this case as well. Further, it is easy - by a direct summation from the definition or induction on the number of random variables $n$ - to see that $(\star)$ holds when the variables $X_i$ are from different distributions as well.

**Theorem 6**

If $X, Y$ are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

Proof

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x,y} xyP(X = x, Y = y) \\
&= \sum_{x,y} xyP_X(X = x)P_Y(Y = y) \\
&= \left[\sum_x xP_X(x)\right]\left[\sum_y yP_Y(y)\right] \\
&= \mathbb{E}[X]\,\mathbb{E}[Y]
\end{aligned}$$

as required. □

Notice $X$ and $Y$ need not be from the same distribution. The sum over $x, y$ instead of $(s_1, s_2)$ abstracts the fact that $X$ and $Y$ are from different distributions. The values

$x, y$ are always real and can be manipulated in the usual ways without worrying about the underlying sample spaces.

In general, if $\{X_i\}$ are independent, then the same proof or induction gives that

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}\left[X_i\right]$$

which is intuitively expected as well - the mean of the product is the product of the means when the rvs don't talk to each other.

Section 3.2

# The median and mode

Subsection 3.2.1

## Quantiles and the Median

**Definition 12** [Quantile]
The $q^{th}$ quantile is defined to be any real number $x_q$ satisfying

$$P(X < x_q) \leq q \leq P(X \leq x_q)$$

or equivalently,

$$P(X \leq x_q) \geq q \text{ and } P(X \geq x_q) \geq 1 - q$$

The **first quartile** is defined to be the $\frac{1}{4}$-th quantile. Similarly the second, third and fourth quartiles.

A **median** is any **2nd quartile**. Imagine a continuous distribution. Then the line $x$ = median divides the area bound by the pdf in half (and similarly the $q^{\text{th}}$ quantile divides it in the ratio $q : 1 - q$).

For a continuous distribution, $P(X < x_q) = P(X \leq x_q)$ so that the $q^{\text{th}}$ quantile is any number $x_q$ satisfying $P(X \leq x_q) = q$ or $F(x_q) = q$. If $F$ is strictly increasing, then any $q^{\text{th}}$-quantile or in particular the median is unique.

Subsection 3.2.2

## The mode

For a discrete random variable, the mode of the distribution is any $x \in \mathscr{A}$ such that $f(x) = \max_{y \in \mathscr{A}} f(y)$.

For a continuous distribution, the mode is defined to be the value of $x$ at any local maximum of the pdf $f(x)$.

Reasonably enough, any distribution with a unique mode is said to be **unimodal** otherwise it is **multimodal**.

**Theorem 7**

Let $f$ be unimodal and symmetric about $x = a$. Then

$$\textit{mean}(f) = \textit{median}(f) = \textit{mode}(f) = a$$

Proof Suppose a mode of $f$ is $x \neq a$. Then $2a-x$ is also a maximum for $f$ and so $2a-x \neq x$ is also a mode for $f$, a contradiction. So the only possible mode for $f$ is $a$.
Now

$$F(a) = \int_{-\infty}^{a} f(x)\mathrm{d}x = \frac{1}{2}\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \frac{1}{2}$$

so that $a$ is a median for $f$. Is the median unique for $f$?
Finally, we have

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} xf(x)\mathrm{d}x \\
&= \int_{-\infty}^{a} xf(x)\mathrm{d}x + \int_{a}^{\infty} xf(x)\mathrm{d}x \\
&= \int_{-\infty}^{a} xf(x)\mathrm{d}x + \int_{-\infty}^{a} (2a-x)f(2a-x)\mathrm{d}x \\
&= 2a\int_{-\infty}^{a} f(x)\mathrm{d}x \\
&= a
\end{aligned}$$

so the mean is also $a$, as required. □

Section 3.3

## Variance of a random variable

It is a measure of the spread of the distribution.

$$\mathrm{var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \geq 0$$

It can be rewritten as

$$\mathrm{var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

*Remark* We get the RMS-AM inequality from the rewritten expression for variance! Notice this is exactly the same as the usual proof of the inequality.

The **deviation** or **standard deviation** of the distribution is defined to be

$$\sigma(X) := \sqrt{\mathrm{var}(X)}.$$

It is especially useful when we might scale our data, in which case the variance goes as the square but the deviation scales linearly with the input.

Subsection 3.3.1

### Variance of some common distributions

Left as an exercise.

- Bernoulli: $pq$ - so the variance is maximized when $p = q = 0.5$.
- Binomial: $npq$ - again, the variance is maximized when $p = q = 0.5$.
- Geometric: $\frac{1}{p^2}$ - and so the mean and standard deviation are identical.

- Poisson: $\lambda$ - and not $\lambda^2$, note.
- Exponential: $\frac{1}{\lambda^2}$ - just like the geometric! Sometimes written in terms of $\beta := \frac{1}{\lambda}$
- Gaussian: $\sigma^2$ - and so the symbol given to the parameter $\sigma$ in the pdf is justified.
- Uniform in $[a, b]$: $\frac{(b-a)^2}{12}$ - obviously it should increase with the size of the interval. A squared relationship is a bonus.

Finally, note that the binomial $P_{\text{binom}}(x; n, p)$ with $p, q$ non-zero, finite and $n \to \infty$ gives $npq$ as the variance, and indeed, its limit (the Gaussian) also has the same variance!

Subsection 3.3.2

## Properties

- $\text{var}(X + c) = \text{var}(X)$
- $\text{var}(cX) = c^2\text{var}(X)$

And so $\text{var}(cX + d) = \text{var}(cX) = c^2\text{var}(X)$.

**Theorem 8**

*var*$(X + Y) =$ *var*$(X) +$ *var*$(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y])$

Proof By linearity we have $\mathbb{E}\left[(X+Y)^2\right] = \mathbb{E}\left[X^2\right] + \mathbb{E}\left[Y^2\right] + 2\mathbb{E}[XY]$ so

$$\begin{aligned}\text{var}(X+Y) &= \mathbb{E}\left[(X+Y)^2\right] - \mathbb{E}[X+Y]^2 \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 + \mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2 + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) \\ &= \text{var}(X) + \text{var}(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y])\end{aligned}$$

as needed. □

It is easy to see (by direct expansion of the LHS or induction on the above theorem) that the following generalization holds:

$$\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}(X_i) + 2\sum_{1 \le i < j \le n}\left(\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]\right)$$

*Corollary* If $X, Y$ are independent (possibly from different sample spaces), then

$$\textit{var}(X+Y) = \textit{var}(X) + \textit{var}(Y)$$

By using induction or the generalization of the above theorem, we get the useful result that for $X_1, X_2, \ldots X_n$ **pairwise** independent, we have

$$\text{var}\left(\sum_i X_i\right) = \sum_i \text{var}(X_i)$$

Notice that the variance $npq$ of the binomial distribution follows immediately from the above corollary and the Bernoulli variance $pq$ - this is perhaps the quickest proof of the same.

Section 3.4

# Some useful Inequalities

Subsection 3.4.1

## Markov's inequality

**Theorem 9**

Let $X$ be a r.v. with pdf $P$, and let $u : \mathscr{A} \to \mathbb{R}$ be non-negative. Finally, let $c$ be positive. Then we have,

$$P(u(X) \geq c) \leq \frac{\mathbb{E}\left[u(X)\right]}{c}$$

Proof

$$\begin{aligned}
\mathbb{E}\left[u(X)\right] &= \sum_{x \in \mathscr{A}} u(x)P(x) \\
&\geq \sum_{\substack{x \in \mathscr{A} \\ u(x) \geq c}} u(x)P(x) \\
&\geq c \sum_{\substack{x \in \mathscr{A} \\ u(x) \geq c}} P(x) \\
&= cP(u(X) \geq c)
\end{aligned}$$

as required. □

Subsection 3.4.2

## Chebyshev's inequality

**Theorem 10**

Let random variable $X$ have pdf $P$, mean $\mu$ and finite variance $\sigma^2$. Let $c$ be positive. Then

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

Proof

$$\begin{aligned}
P(|X - \mu| \geq c) &= P((X - \mu)^2 \geq c^2) \\
&\leq \frac{\mathbb{E}\left[(X - \mu)^2\right]}{c^2} = \frac{\sigma^2}{c^2}
\end{aligned}$$

as required. □

Verifying plausibility: The theorem tells us that the probability of the outcome of a random variable deviating by $c$ from the mean is directly proportional to the variance (as one would expect, yes) and decreases with an increase in $c$ (again, as one would expect).

**Note:** $X$ may be a function of some other random variable(s) as well (What is the expectation over in this case?). Indeed, we use this fact in the proof of the law of large numbers.

*Corollary* Let $k$ be positive. Then

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Of course this is only useful when $k > 1$.

Subsection 3.4.3

## Jensen's inequality

**Theorem 11**

Let $X$ be any random variable, $f$ be any **convex** function. Then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Proof Watch carefully, for this looks too easy. We have that the (or 'a', in case $f$ is not differentiable at the mean) tangent line at $x = \mathbb{E}[X]$ lies below the function, that is, is the tangent line is $y = ax + b$, then

$$\mathbb{E}[f(x)] \geq \mathbb{E}[ax + b] = a\mathbb{E}[x] + b = f(\mathbb{E}[X])$$

and we're done. □

Another perspective on the theorem: The theorem is just saying that the centroid of the points $x_i, f(x_i)|x_i \in \mathscr{A}$ lies above the value of the function at the same x coordinate. Where did we use this in the proof?

Similarly, one has the following:

**Theorem 12**

Let $X$ be any random variable, $f$ be any **concave** function. Then

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$$

Proof Let a tangent line to $f$ at $x = \mathbb{E}[X]$ be $y = ax + b$. Then by assumption $\mathbb{E}[f(x)] \leq \mathbb{E}[ax + b]$ so that

$$\mathbb{E}[f(x)] \leq \mathbb{E}[ax + b] = a\mathbb{E}[x] + b = f(\mathbb{E}[X])$$

as needed. □

Finally, note that the inequalities in both theorems are strict if $f$ is strictly convex or concave, respectively.

Subsection 3.4.4
### Minimizer of the $\mathcal{L}^1$ norm

**Theorem 13**

$$\mathbb{E}\left[|X-c|\right] = \min_{a\in\mathbb{R}} \mathbb{E}\left[|X-a|\right] \iff c \text{ is a median.}$$

Proof For the discrete case, let

$$\Phi(x) := \mathbb{E}\left[|X-x|\right] = \sum_{t\in\mathscr{A}} |t-x|P(t)$$

then we have TODO □

Subsection 3.4.5
### The law of large numbers

**Theorem 14** [The Law of Large Numbers]

Let random variables $X_1, X_2, ..., X_n$ be independent (but not necessarily identically distributed), each with the same mean $\mu = E[X_i]$ and finite variance $\sigma^2 = \text{var}\,(X_i)$. Let the random variable $\overline{X}$ be defined by

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

Then, for any fixed $k > 0$,

$$\lim_{n\to\infty} P(|\overline{X} - \mu| \geq k) = 0$$

or $\overline{X}$ converges in probability to its expectation $\mathbb{E}[\overline{X}] = \frac{n\mu}{n} = \mu$

Proof We have $\mathbb{E}[\overline{X}] = \frac{n\mu}{n} = \mu$. We also have $\text{var}(\overline{X}) = \frac{1}{n^2}\sum_i \text{var}\,(X_i) = \frac{\sigma^2}{n}$ since the $X_i$'s are independent. So by Chebyshev on the random variable $\overline{X} : \Omega^n \to \mathbb{R}$ we have

$$0 \leq P(|\overline{X} - \mu| \geq k) \leq \frac{\sigma^2}{nk^2}$$

Taking the limit as $n \to \infty$ yields the result. □

This theorem tells us that as we take a huge number of measurements $X_i$, the mean of all these is *always* approximately equal to the mean of the whole distribution.

**Intuiting the theorem:** The $X_i$'s are identically distributed and independent, so for a large sample size, we can expect (approximately) that "every $x \in \mathscr{A}$ appears in the sample according to the probability distribution" so their average is the same in

every sample - variance $\to 0$ - and is equal to the mean of the distribution itself - The LLN. Chebyshev gives us the formal proof of the same.

Section 3.5

# Joint distributions: The Covariance and Correlation

We now look at some measures that tell us how two random variables relate to each other.

Subsection 3.5.1

## Covariance

**Definition 13**
Let $X, Y$ be two random variables with joint distribution $P(X, Y)$ and expectations $\mu_X, \mu_Y$. Then the **covariance of $X$ and $Y$** is defined by:

$$\begin{aligned} cov(X, Y) &:= \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right] \\ &= \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \end{aligned}$$

Basically, if the covariance is large and positive, then if $x$ was larger than $\mu_X$ or $(x - \mu_X) > 0$, we would expect $(y - \mu_Y) > 0$ as well - else the contribution of the product to the expectation would be large and negative which is unlikely - that is, $X$ and $Y$ usually vary in the same direction from their means. We say $X$ and $Y$ vary "positively".

As expected, the covariance of two independent variables is 0. But note that **the covariance being zero does not imply that the variables are independent**. Covariance is bilinear.

Subsection 3.5.2

## Standardized Random Variables and the Correlation coefficient

Remember the $t$ in the definition of the Gaussian pdf? We revisit that.

**Definition 14** [Standardized RVs]
Let $X$ be a random variable with finite mean $\mu$ and finite variance $\sigma^2$. Then the random variable $X^*$ defined by

$$X^* := \frac{X - \mu}{\sigma}$$

is called the standardized form of $X$.

Clearly, $\mathbb{E}\left[X^*\right] = 0$ and $\text{var}\left(X^*\right) = 1$. Note that **the standardized variable is unaffected by a scaling of the whole data**.

**Definition 15** [Correlation]

The **correlation coefficient** of two random variables $X$ and $Y$ is defined to be

$$cor(X,Y) := \mathbb{E}[X^*Y^*] = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

The correlation also has the following nice form as a covariance:

$$\text{cor}(X,Y) = \mathbb{E}[X^*Y^*] = \mathbb{E}[X^*Y^*] - \mathbb{E}[X^*]\mathbb{E}[Y^*] = \text{cov}(X^*,Y^*)$$

**Lemma 2**
Let $X, Y$ be random variables. Then

$$-1 \leq cor(X,Y) \leq 1$$

Proof This is basically the exact same proof that $-1 \leq \mathbf{a}\cdot\mathbf{b} \leq 1$ for **unit** vectors **a**, **b**. Here $\|\cdot\|^2$ is replaced by $\mathbb{E}[\cdot^2]$, that's all. Note that $\mathbb{E}[(X^*)^2] = \mathbb{E}[(Y^*)^2] = 1$. We have

$$0 \leq \mathbb{E}\left[(X^* + Y^*)^2\right] = 2 + 2\text{cor}(X,Y)$$

and

$$0 \leq \mathbb{E}\left[(X^* - Y^*)^2\right] = 2 - 2\text{cor}(X,Y)$$

completing the proof. □

**The curious case of linearly dependent variables**

Let $X, Y$ be related by $Y = mX + c$. Also let $\mathbb{E}[X] = \mu$ and $\text{var}(X) = \sigma^2$. Then we have the following results.

**Theorem 15**

We have the following:

- $\text{cov}(X,Y) = m\sigma^2$.
- $\text{cor}(X,Y) = \text{sgn}(m)$. In fact $Y^* = \text{sgn}(m)X^*$.

Proof It is easy to see that $\mathbb{E}[Y] = m\mu + c$ and $\text{var}(Y) = m^2\sigma^2$. Hence

$$Y^* = \frac{mX + c - (m\mu + c)}{|m|\sigma} = \text{sgn}(m)X^*$$

Hence $\text{cor}(X,Y) = \mathbb{E}[X^*Y^*] = \text{sgn}(m)\mathbb{E}[(X^*)^2] = \text{sgn}(m)$. Also $\text{cov}(X,Y) = \sigma_X\sigma_Y\text{cor}(X,Y) = \text{sgn}(m)|m|\sigma\cdot\sigma = m\sigma^2$ as needed. □

The proofs can also be done straight from the definition in each case. This is left as an exercise.

So if $X, Y$ are linearly related, then $|\text{cor}(X,Y)| = 1$. Remarkably, the converse is also true.

**Theorem 16**

$$|cor(X,Y)| = 1 \iff X, Y \text{ are linearly related almost everywhere.}$$

Proof We have already shown the only if direction. For the if direction, we again use the unit vector analogy. We are trying to show that unit vectors **a** and **b** are parallel if their dot product is $\pm 1$. We proceed in the same way as we would to prove this. If $\text{cor}\,(X,Y) = +1$, then we have $\mathbb{E}\left[(X^* - Y^*)^2\right] = 0$, so that $(X^* - Y^*)^2 = 0$ almost everywhere or $X^* = Y^*$ almost everywhere. Expanding $X^*$ and $Y^*$ in terms of $X$ and $Y$ gives us that $X$ and $Y$ are linearly related, as needed. The case $\text{cor}\,(X,Y) = -1$ is handled similarly. □

We end this section with a nice way to find the exact linear relationship between linearly related variables $X$ and $Y$. We have $Y^* = \text{sgn}(m)X^*$

# Chapter 4

# Estimation

**Definition 16** [**i.i.d**]
The random variables $X_1, \dots X_n$ are said to be independent and identically distributed or i.i.d if they have the same pdf and are pairwise independent.

Note that the random variables are usually from the same sample space but they need not be so. Their distribution should be identical and they must be independent, that's all.

Example: Let a coin toss (modeled as a Bernoulli distribution) be performed $100$ times, and the number on the $i^{\text{th}}$ toss be represented by the random variable $X_i$. Then clearly, the variables $X_i$ are independent, and since the same coin is tossed each time, obey the same (Bernoulli) distribution, hence they are i.i.d.

**Definition 17**
If random variables $X_1, \dots, X_n$ are **i.i.d**, then they constitute a **random sample** of size $n$ from their common distribution. $n$ is called the sample size.

A bit of abuse of notation here: The *set* $\{X_1, \dots X_n\}$ is called a sample, the set $\{x_1, x_2, \dots X_n\}$ where each rv has been assigned a number is also a sample, and each element of either set is also called a sample. Which sample is referred to is usually clear from the context.

Section 4.1
## Statistics

This common distribution from which data was "drawn" is usually unknown.

Usually, from a sample, we would like to estimate something about the distribution so that we can predict the same about new data from the distribution.

A statistic does exactly this - estimates something about the distribution.

**Definition 18** [statistic]
Let $X_1, ..., X_N$ denote a sample from a distribution with generic rv random variable $X$ (i.e., all of $X_1, ..., X_N$ have the same distribution as $X$).
Let $T(X_1, ..., X_N)$ be a **function of the sample**(function of the $N$ RVs in the sam-

ple). Then, random variable $T$ is called a **statistic**.

Given a sample (usually with $N$ reasonably large) we would like to guess the distribution the sample was drawn from so that we **can now make predictions** about new data that might come from that distribution. Hence we define:

**Definition 19** [statistical model]
A probabilistic description of real-life data. Description typically involves a distribution with some parameters to be determined.

A model is a data-generator - give input, will give you the output.

The *type* of the distribution is one that the statistician chooses (and fixes) - the parameters that this instance of the distribution takes are called the (to be found) parameters of the model.

We must measure and calculate these parameters to get a good approximation to the actual distribution. This is what Estimation theory does. It is a branch that deals with estimating the values of parameters underlying a statistical model based on measured-empirical data.

While data-generation gets you data from parameter values (i.e. from the actual distribution), Estimation gets you the right parameter values for the actual distribution from measured/provided data.

So we want the parameters. We have some data, let's try to glean the parameter values from them.

An estimator is a statistic whose sole goal in life is to estimate the parameters of a parametrized distribution straight from the data.

**Definition 20** [Estimator]
A deterministic statistic that estimates a certain parameter of a distribution is called an estimator for that parameter of the distribution.

Subsection 4.1.1
### Performance of an estimator

Let $T$ be an estimator.

- The **mean** of an estimator is defined to be $\mathbb{E}[T]$.
- The **variance** of an estimator is defined to be $\mathbb{E}\left[(T - \mathbb{E}[T])^2\right]$.
- The **bias** of an estimator is defined to be $\mathbb{E}[T] - \theta$, where $\theta$ is the real parameter of the distribution. An estimator is unbiased if it has zero bias. Else it is biased.
- The **Mean Squared Error / MSE** of an estimator is defined to be the expected value of the squared error as the sample varies:

$$\mathrm{MSE}(T) := \mathbb{E}\left[(T - \theta)^2\right]$$

This is what we usually try and minimize for an estimator.

A useful result:

**Theorem 17** [Bias-Variance decomposition]

$$MSE(T) = var(T) + bias(T)^2$$

Proof $(T-\theta)^2 = (T - \mathbb{E}[T] + \mathbb{E}[T] - \theta)^2 = (X+b)^2$, where $X = T - \mathbb{E}[T]$ is a variable and $b = \mathbb{E}[T] - \theta$ is the bias of the estimator and is a constant independent of the sample. Now we have

$$\begin{aligned} \text{MSE}(T) &= \mathbb{E}\left[(T-\theta)^2\right] \\ &= \mathbb{E}\left[(X^2 + 2bX + b^2)\right] \\ &= \text{var}(T) + 2b(0) + b^2 \\ &= \text{var}(T) + \text{bias}(T)^2 \end{aligned}$$

as required. □

**Definition 21** [Consistent estimator]
An estimator $T$ for parameter $\theta$ is **consistent** if $T_N = T(X_1, \dots X_N)$ converges in probability to the true value $\theta_{\text{true}}$ of the distribution. In other words, for any $k > 0$,

$$\lim_{N\to\infty} P(|T_N - \theta_{\text{true}}| \geq k) = 0$$

A good estimator must be consistent - I'm feeding it an infinitely large sample (albeit just countably infinite) to get the parameter from, and it better do so, otherwise how will it even get close with just a finite sample size!

All we have to do now is figure what a good estimator should look like. We use the ingenious Likelihood Functions to do so.

Section 4.2

## The Likelihood Function

**Definition 22**
Let $\theta$ be a parameter for the distribution, and $S = \{X_1, \dots X_n\}$ be a sample from the distribution. Then the **likelihood function for** $\theta$ as a function of the sample $S$ is defined by

$$\mathcal{L}(\theta; X_1, \dots X_n) := \frac{1}{N}\sum_{i=1}^{N} \log(P(X_i; \theta))$$

where $P(\star; \theta)$ denotes the pdf of the distribution with parameter $\theta$. Note that the Likelihood is defined to be $-\infty$ if any of the $P(X_i; \theta)$ terms are $0$.

The likelihood function defined above is also called the log-likelihood function. Alternate definitions for the likelihood function do exist, a common one being the

geometric mean (or even just the product) of the $P(X_i;\theta)$'s in which case $-\infty$ is replaced by 0, we will find it convenient to work with the given definition. In Chapter 6 on Bayesian Estimation, we will use the likelihood defined above without the $\frac{1}{N}$ term and this will also be called the log-likelihood function.

Subsection 4.2.1

## The Maximum Likelihood Estimation (MLE) Theorem

We must look at how we might go about deciding the distribution parameters for a given (fixed in our discussion) sample.

**Theorem 18**

Let $\theta_{\text{true}}$ be the parameter/parameters that **led to the sample** $X_1, X_2 \dots X_n$ (i.e. the actual parameter/parameters of the distribution). Also suppose that any member of the parametrized distribution $P(\star,\theta)$ ($P$ is the pdf, note) in consideration has the same support $\mathscr{S}$. Finally suppose that $\mathbb{E}\left[\frac{P(X;\theta)}{P(X;\theta_{\text{true}})}\right]$ is finite. Then

$$\lim_{n\to\infty} P\Big(\mathcal{L}\left(\theta_{\text{true}}; X_1, X_2, \dots X_n\right) > \mathcal{L}\left(\theta; X_1, X_2, \dots X_n\right); \theta_{\text{true}}\Big) = 1$$

for all $\theta \neq \theta_{\text{true}}$.

Proof First note that the event

$$\mathcal{L}\left(\theta_{\text{true}}; X_1, X_2, \dots X_n\right) > \mathcal{L}\left(\theta; X_1, X_2, \dots X_n\right)$$

is equivalent to

$$S = \frac{1}{N}\sum_{i=1}^{n} \log\left(\frac{P(X_i;\ \theta)}{P(X_i;\ \theta_{\text{true}})}\right) < 0$$

Let $Y_i$ be random variables defined by

$$Y_i = \log\left(\frac{P(X_i;\ \theta)}{P(X_i;\ \theta_{\text{true}})}\right)$$

for $i \in \{1, 2, \dots, N\}$. Also, let $Y$ be a generic element of the $Y_i$'s. Note that $\mathbb{E}[Y]$ will have the actual distribution $P(\star;\theta_{\text{true}})$ as its underlying pdf since that's where each $X_i$ came from.
As $N \to \infty$, we have that

$$\begin{aligned} S &= \frac{\sum_{i=1}^{N} Y_i}{N} \\ &\to \mathbb{E}[Y] \text{ (LLN)} \\ &= \mathbb{E}\left[\log\left(\frac{P(X;\ \theta)}{P(X;\ \theta_{\text{true}})}\right)\right] \\ &< \log\left(\mathbb{E}\left[\frac{P(X;\ \theta)}{P(X;\ \theta_{\text{true}})}\right]\right) \text{ (Jensen)} \end{aligned}$$

$$= \log\left(\sum_{x\in\mathscr{S}} \chi(x)P(x;\theta_{\text{true}})\right) \text{ ($\mathbb{E}[Y]$ was wrt the actual distribution)}$$
$$= \log\left(\sum_{x\in\mathscr{S}} \frac{P(x;\ \theta)}{P(x;\ \theta_{\text{true}})}P(x;\ \theta_{\text{true}})\right)$$
$$= \log\left(\sum_{x\in\mathscr{S}} P(x;\theta)\right)$$
$$= \log 1$$
$$= 0$$

which concludes the proof. □

The main thing in the proof was the way the likelihood was defined - as a sum of logarithms of probabilities (or equivalently, as a product of probabilities). But a sum of probabilities etc would not work, for example.

Subsection 4.2.2

## What does the theorem tell us?

Let $S = \{X_1, ..., X_n\}$ be a **fixed** large sample. The size $n$ of S can be taken to be large enough that the preceding theorem holds for this $n$. Now, for this sample, suppose

$$\theta_0 = \arg\sup_{\theta} \mathcal{L}(\theta; S)$$

Then is it possible that $\theta_0 \neq \theta_{\text{true}}$ for sample S? No! Since otherwise we get $P(\mathcal{L}(\theta_{\text{true}};\ \text{sample}) > \mathcal{L}(\theta_0;\ \text{sample})) < 1$ - since when the sample = $S$, the inequality is false. (But what if the set of samples isn't finite, then $P(\dots)$ is still 1, there is no direct contradiction). Hence, $\theta_0$ must be $\theta_{\text{true}}$, and we have found the true parameter! In actuality, this will be only a good approximation to the true parameter since $n$ is finite (and hence LLN, used in the proof, will only be approximate), but a rather good estimate anyway.

Hence we get the so-called Maximum Likelihood or ML estimator:

$$T(S) := \arg\max_{\theta} \mathcal{L}(\theta; S)$$

is called the ML estimator for the parameter $\theta$. It can easily be calculated from the sample using analytic techniques or numerical methods like Gradient Ascent.

### Consistency of the ML estimator

The ML estimator may not exist in some cases. But when it does exist, it can be shown to be consistent.

**Theorem 19**

The ML estimator is consistent. That is, if the estimator is $T_n =$

$T(X_1, X_2, \ldots X_n) = \theta$, then for any $k > 0$,

$$\lim_{n\to\infty} P(|T_n - \theta_{true}| \geq k) = 0$$

Proof Let $n$ be finite. Then by the MLE theorem we have that

$$P\Big(\mathcal{L}\left(\theta_{\text{true}}; X_1, X_2, \ldots X_n\right) > \mathcal{L}\left(\theta; X_1, X_2, \ldots X_n\right); \theta_{\text{true}}\Big) \approx 1$$

which is approximate since the LLN is approximate for $n$ finite. The above equation means that there are some samples $S = \{X_1, X_2, \ldots X_n\}$ which violate this, but there aren't too many of them (The probability of the inequality being obeyed is very likely). So the event that $|T_n(S) - \theta_{\text{true}}| \geq k$ has a non-zero but small probability. As $n \to \infty$, this small probability vanishes and so the estimator is consistent. Basically, when $n \to \infty$, there are (in probability) no samples S where the maximum likelihood is achieved at a $\theta \neq \theta_{\text{true}}$. □

Subsection 4.2.3

## ML estimators for some common distributions

Let $S = \{X_1, X_2, \ldots, X_n\}$ be a sample from the distribution. We estimate the parameters of the distribution from this sample by maximizing the likelihood corresponding to this sample. The maximization is easily done using standard derivative and gradient tests. Note that in each case we verify the consistency of the estimator. The results are stated below and are left as an easy exercise in calculus.

- **Bernoulli:**
$$p_{\text{mle}} = \frac{\sum_{i=1}^{n} X_i}{n}$$
$p$ is the MLE-calculated parameter. And indeed, as $n \to \infty$, $p_{\text{mle}} \to \mu = p_{\text{actual}}$. For $n$ finite, again, note that the MLE $p$ is but an approximation (LLN is an approximation for finite $n$ so the MLE theorem is approximate for $n$ finite)
- **Binomial:**
$$p_{\text{mle}} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$
As $n \to \infty$, $p_{\text{mle}} \to \frac{np_{\text{actual}}}{n} = p_{\text{actual}}$.
- **Geometric:** $p_{\text{mle}} = \frac{1}{\overline{X}}$, $p_{\text{mle}} \to \frac{1}{\mu_{\text{geom}}} = p_{\text{actual}}$ as $n \to \infty$
- **Poisson:** $\lambda_{\text{mle}} = \overline{X} \to \mu_{\text{poisson}} = \lambda_{\text{actual}}$
- **Exponential:** $\lambda_{\text{mle}} = \frac{1}{\overline{X}} \to \frac{1}{\mu_{\text{exp}}} = \lambda_{\text{actual}}$
- **Normal:**
$$\mu_{\text{mle}} = \overline{X} \to \mu_{\text{gaussian}} = \mu_{\text{actual}}$$
$$\sigma^2_{\text{mle}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \to \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] = \sigma^2_{\text{actual}}$$

- **Half-Normal:** This distribution is 0 for $x \leq 0$, and $2 \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$ for $x > 0$.
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - 0)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$
- **Uniform continuous:** Let $X_1 \leq X_2 \cdots \leq X_n$. We have the pdf
$$P(x; a, b) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$
for $\mathscr{A} = [a, b]$. Also $\mathcal{L}(a, b; S) = \frac{1}{n}\sum_{i=1}^{n} P(X_i;\ a, b)$ is $-\infty$ if $X_1 < a$ or $b < X_n$, hence for max-likelihood $a \leq X_1 \leq X_2 \cdots \leq X_n \leq b$. In this case $\mathcal{L}(a, b; S) = -\log(b - a)$ is maximized when $b - a$ is minimized, that is, when
$$a = X_1 = \min_{i=1}^{n} X_i$$
$$b = X_n = \max_{i=1}^{n} X_i$$
which is our required ML estimate for the parameters $a$ and $b$ of the continuous uniform distribution.

Section 4.3

## Sample estimators

We've seen ML estimators. They are really good at estimation. **Sample estimators** are another kind of estimator - they assume the given input data is all there is in the distribution. That is, the values $\{X_1, X_2, \ldots X_n\}$ are taken to be exactly the values that the random variable whose distribution we are looking for takes. Thus the sample estimator for the mean argues that the mean is close to

$$\overline{X}(S = \{X_1, X_2, \ldots, X_N\}) := \frac{1}{N}\sum_{i=1}^{N} X_i$$

Similarly, the sample variance estimator is

$$\widehat{\sigma^2}(X_1, \ldots, X_N) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} X_i^2\right) - \overline{X}^2$$

Essentially, all expectations are taken to be the average of the values at each datapoint, that is, $\mathbb{E}[f(X)]$ in the actual expression of the parameter is replaced by $\frac{1}{N}\sum_{i=1}^{N} f(X_i)$ in the estimator.

We also have the very useful sample covariance estimator

$$\widehat{C}(\{X_1, Y_1\}, \ldots, \{X_N, Y_N\}) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$= \left(\frac{1}{N}\sum_{i=1}^{N} X_i \cdot Y_i\right) - \overline{X} \cdot \overline{Y}$$
$$= \overline{XY} - \overline{X} \cdot \overline{Y}$$

Since we have $\frac{1}{N}\sum_{i=1}^{N} f(X_i) \to \mathbb{E}[f(X)]$ as $n \to \infty$ by the LLN, the mean, variance and covariance (and any others which are representable as some combination of expectations) estimators are consistent (when the actual distribution satisfies the assumptions of the LLN, of course).

Note that the sample mean estimator is sometimes also called the mean of the data (since it is indeed a mean - but not quite an expectation!). Similarlythe other two estimators are called the variance and covariance of the data, respectively.

Subsection 4.3.1

## Bias of the Sample Estimators

1. The sample mean estimator $\overline{X}(X_1, \dots X_N)$ is unbiased.

Proof We have $\mathbb{E}\left[\overline{X}\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[X_i] = \frac{1}{N}N\mu = \mu$ where $\mu$ is the true mean, hence the estimator $\overline{X}$ is unbiased. □

Note that here we have used the linearity of expectation for the many random variables case.

We can easily generalize the above to the sample mean estimator for any function of the random variable:

**Theorem 20**

Let $\{X_1, X_2, \dots, X_n\}$ be a sample from a distribution with rv $X$, then the estimator $\overline{f(X)}(X_1, X_2, \dots, X_n) := \frac{1}{N}\sum_{i=1}^{N} f(X_i)$ is unbiased.

Proof Say $\mathbb{E}[f(X)] = \mu$. Now simply expand:

$$\mathbb{E}\left[\overline{f(X)}\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[f(X_i)] = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu$$

and hence, the bias $= \mathbb{E}\left[\overline{f(X)}\right] - \mu = 0$ □

2. The sample variance estimator is **biased**.

Proof We have

$$\mathbb{E}\left[\widehat{\sigma^2}\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\overline{X}^2\right]$$

We find

$$\mathbb{E}\left[X_i^2\right] = \text{var}(X_i) + \mathbb{E}[X_i]^2 = \sigma^2 + \mu^2$$

where $\sigma^2$ is the true variance. Similarly, we have

$$\mathbb{E}\left[\overline{X}^2\right] = \text{var}\left(\overline{X}\right) + \mathbb{E}\left[\overline{X}\right]^2 = \text{var}\left(\overline{X}\right) + \mu^2$$

Now

$$\text{var}\left(\overline{X}\right) = \frac{1}{N^2}\text{var}\left(\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}\sum_{i=1}^{N}\text{var}\left(X_i\right) = \frac{\sigma^2}{N}$$

where we have used the fact that the variance of the sum is the sum of the variances when the variables are pairwise independent. Putting this together we get

$$\mathbb{E}\left[\widehat{\sigma^2}\right] = \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{N} + \mu^2\right) = \sigma^2\left(1 - \frac{1}{N}\right)$$

$$\implies \text{bias} = \mathbb{E}\left[\widehat{\sigma^2}\right] - \sigma^2 = \frac{-\sigma^2}{N} \neq 0$$

□

Note that the bias of the sample variance estimator decreases with an increase in the sample size.

3. The sample covariance estimator is also biased.

Proof Let $\mathbb{E}\left[X_i\right] = \mu_X, \mathbb{E}\left[Y_i\right] = \mu_Y$ and $\mathbb{E}\left[X_iY_i\right] = \mu_{XY}$. Note that for $i \neq j$, $X_i$ and $Y_j$ are also independent. Now we compute the expectation of the sample covariance:

$$\mathbb{E}\left[\widehat{C}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i \cdot Y_i\right] - \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) \cdot \left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right)\right]$$

The first term is essentially the sample mean estimator for the random variable $XY$ and so has expectation $\mathbb{E}\left[XY\right] = \mu_{XY}$. The second term expands to

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) \cdot \left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right)\right] &= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbb{E}\left[X_iY_i\right] - \sum_{i\neq j}\mathbb{E}\left[X_iY_j\right]\right) \\
&= \frac{1}{N^2}\left(N\mu_{XY} - \sum_{i\neq j}\mathbb{E}\left[X_i\right]\mathbb{E}\left[Y_j\right]\right) \\
&= \frac{(N\mu_{XY} - N(N-1)\mu_X\mu_Y)}{N^2}
\end{aligned}$$

so we get

$$\mathbb{E}\left[\widehat{C}\right] = (\mu_{XY} - \mu_X\mu_Y)\left(1 - \frac{1}{N}\right)$$

and hence

$$\text{bias} = \mathbb{E}\left[\widehat{C}\right] - (\mu_{XY} - \mu_X\mu_Y) = \frac{-(\mu_{XY} - \mu_X\mu_Y)}{N} \neq 0$$

where $(\mu_{XY} - \mu_X \mu_Y)$ is the true covariance of $X$ and $Y$. □

*Remark* It is quite remarkable that the bias is exactly proportional to the true covariance meaning a simple scaled version of the sample covariance estimator will be unbiased. Similarly for the variance estimator. For example,

$$\widehat{\sigma^2}' = \frac{N}{N-1}\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

$$\widehat{C}' = \frac{N}{N-1}\widehat{C} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

are unbiased estimators for the variance and covariance of the data. Nevertheless, when we refer to the variance and covariance sample estimators, we mean the original biased ones.

Section 4.4

## The Central Limit Theorem

Consider any *arbitary* distribution and consider a sample of size $N$ from this distribution. Consider the sample mean function $\overline{X}$ of these $N$ RVs. It is also a random variable with some distribution. But what distribution? We do know that the expectation of the distribution is $\mathbb{E}\left[\overline{X}\right] = \mathbb{E}\left[X_1\right] = \mu$ so it's the same as that of the original distribution, and similarly the variance of this distribution is $\frac{\text{var}(X_1)}{N} = \frac{\sigma^2}{N}$. Okay, so we know something about the distribution of the sample mean. What else? Here's where the central limit theorem comes in. It tells us that as $N$ grows larger (the distribution at $N$ = 10-20 is good enough a Gaussian in everyday life), the distribution of $\overline{X}$ is a **Gaussian**, with the mean and variance as calculated earlier.

Notice the power of this theorem. Excepting the fact that the distribution had a finite mean and variance, nothing else is assumed about the distribution. And yet, always, the sample mean has a Gaussian distribution about the actual mean $\mu$ with a smaller and smaller standard deviation $\frac{\sigma}{\sqrt{N}}$ as the sample size $N$ grows larger.

Note that the sample mean as the number of draws from the original distribution (the sample size) increases to $\infty$ is expected to go exactly to the mean $\mu$ of the original distribution - and it does! As $N$ increases, indeed the deviation about the peak $\mu$ of the Gaussian reduces to 0. (Dirac Delta about the mean, anybody?)

This theorem is why most distributions that we don't know are labeled Gaussian - given any sample from a source distribution, we choose to work with the sample mean as our random variable from that source instead, and we *know* this rv's distribution - a Gaussian! The parameters of the Gaussian are then easily found by ML estimation.

**Theorem 21** [Central Limit Theorem]

Let $X_1, X_2, \ldots X_N$ be a sample from a distribution with finite mean and variance.

Let the random variable $\overline{X}$ be defined by

$$\overline{X} := \frac{\sum_{i=1}^{N} X_i}{N} \text{ (Sample Mean)}$$

Then as $N \to \infty$, $\overline{X}$ has a Gaussian distribution (pdf) with mean $\mu$ and variance $\frac{\sigma^2}{N}$, where $\mu, \sigma$ are the mean and std. deviation of the original distribution.

Note that the distribution of the sample mean converges very quickly to a normal (with variance dependent on the sample size, of course) distribution as $N$ increases. A sample size of about 10 or 20 is more than enough and has a tiny Skew and Kurtosis (measures of difference from a perfect Normal distribution). But of course, don't abuse this - $N$ = 1, 2 or 3 usually don't work well at all.

Section 4.5

## Linear Regression

The idea is to fit data that is approximately linear to a line.

- **Given:** Data $\{(x_i, y_i)\}_{i=1}^{n}$, where *we choose the $x_i$'s as we like.* So essentially the data is just the set of $y_i$'s.
- **Distribution assumed:** *[The linear model]* We assume that the data $\{y_i\}$ comes from a family of distributions parametrized by $n + 3$ constants - $\alpha$, $\beta$, $\sigma$, $x_1$, $x_2$, …, $x_n$. The distribution is on $n$ variables $y_1$, $y_2$, …, $y_n$ with the following form:

  $$P(y_1, \ldots, y_n;\ \alpha, \beta, \sigma, x_1, \ldots, x_n) = \prod_{i=1}^{n} \mathcal{N}(y_i - (\alpha + \beta x_i);\ 0, \sigma^2)$$

  The data generated when we plug in some values for the $x_i$'s is assumed to be generated from the above family at some constant $\alpha$, $\beta$ and $\sigma$. We wish to estimate the above three constants (and in practice, we usually only care about $\alpha$ and $\beta$) in terms of the data $y_i$ received and the parameters $x_i$ we chose.
- **Goal:** From the given data, estimate $\alpha$ and $\beta$. We do this here using the ML estimator for the two parameters.

Of course, we start the problem by computing the likelihood for this distribution. The likelihood is

$$\mathcal{L}\,(\alpha, \beta;\ \text{data}) = \log P(\text{data};\ \alpha, \beta, \sigma, \{x_i\})$$

we maximize $\frac{1}{N}$ times this:

$$\begin{aligned}\mathcal{L} &= \frac{1}{N} \sum_i \log \mathcal{N}(\eta_i = y_i - \beta x_i - \alpha;\ 0, \sigma^2) \\ &= -\left( \frac{1}{2\sigma^2} \sum_i \frac{(y_i - \beta x_i - \alpha)^2}{N} \right) - \log \sigma - \log \sqrt{2\pi}\end{aligned}$$

which is just (up to some constants) the negative of the $\mathcal{L}^2$ cost function from machine learning! That's why it is used - it is minimized at the Ml estimates for the parameters, perfect! Similarly, the $\mathcal{L}^1$ norm comes up instead if the $\eta_i$'s were assumed to come from a Laplace distribution ($f(x;\, b) \propto \exp\left(-\frac{|x|}{b}\right)$).

It is now a simple matter to find $\alpha$ and $\beta$:
At maximum likelihood, $\nabla P = 0$, or in particular $\frac{\partial P}{\partial \alpha} = \frac{\partial P}{\partial \beta} = 0$.
Thus we get

$$0 = \frac{\partial P}{\partial \alpha} = \frac{1}{\sigma^2 N}\sum_{i=1}^{N}(y_i - \beta x_i - \alpha) \implies \overline{y} = \alpha + \beta\overline{x}$$

$$0 = \frac{\partial P}{\partial \alpha} = \frac{1}{\sigma^2 N}\sum_{i=1}^{N} x_i(y_i - \beta x_i - \alpha) \implies \overline{xy} = \alpha\overline{x} + \beta\overline{x^2}$$

Solving these we get

$$\boxed{\widehat{\beta} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\widehat{C}(\{x_i, y_i\})}{\widehat{\sigma^2}(\{x_i\})}}$$

$$\boxed{\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} = \frac{\overline{x^2}\cdot\overline{y} - \overline{x}\cdot\overline{xy}}{\widehat{\sigma^2}(\{x_i\})}}$$

The line $y = \widehat{\alpha} + \widehat{\beta}x$ is the line predicted by the ML Estimator and is called the **best-fit line** for the data.

Subsection 4.5.1
### Analysis of the estimators for $\alpha$ and $\beta$

1. The estimator $\widehat{\beta}$ for $\beta$ is **unbiased**.

Proof We must find $\mathbb{E}[\widehat{\beta}]$, yes. But what is the expectation over? Over all possible values of $y_i \in \mathbb{R}$. Okay. The denominator of $\widehat{\beta}$ is a constant, then. We have

$$\mathbb{E}[\widehat{\beta}] = \frac{\mathbb{E}\left[\overline{xy}\right] - \overline{x}\cdot\mathbb{E}\left[\overline{y}\right]}{\overline{x^2} - \overline{x}^2}$$

First, it is easy to show that $\mathbb{E}\left[y_i\right] = \alpha + \beta x_i$ for each $i$ (Use $\mathbb{E}\left[y_i\right] = \mathbb{E}_{y_i}[y_i]$). Now, we compute the expectations we need:

$$\mathbb{E}\left[\overline{y}\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[y_i\right] = \frac{1}{N}\sum_{i=1}^{N}(\alpha + \beta x_i) = \alpha + \beta\overline{x}$$

$$\mathbb{E}\left[\overline{xy}\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i y_i\right] = \frac{1}{N}\sum_{i=1}^{N} x_i\mathbb{E}\left[y_i\right] = \frac{1}{N}\sum_{i=1}^{N} x_i(\alpha + \beta x_i) = \alpha\overline{x} + \beta\overline{x^2}$$

Substituting, we get $\mathbb{E}[\widehat{\beta}] = \beta$, as required. □

2. $\widehat{\alpha}$ is also unbiased.

Proof | $\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}$, so $\mathbb{E}[\widehat{\alpha}] = \mathbb{E}[\overline{y}] - \mathbb{E}[\widehat{\beta}]\overline{x} = (\alpha + \beta\overline{x}) - \beta\overline{x} = \alpha$, as required. □

3. The point $(\overline{x}, \overline{y})$ lies on the best-fit line. Proof is left to the reader.

4. The sum of the residuals $\eta_i = y_i - \alpha - \beta x_i$ is 0. Proof is left to the reader.

Subsection 4.5.2

## Why Linear Regression

Why are these assumptions on our distribution reasonable? Why did we take care in checking how good our estimators are?

Consider an experiment as follows, from which we collect our data. We measure the temperature ($y_i$'s) at 6 PM, 7 PM, 8 PM, ... 12 AM (the choosable constant $x_i$'s). It seems (we plotted the points, and they seemed to be approximately on a line) that temperature varies linearly with time.

But the temperature each day depends on other things as well and so is not the same every single time at 6 PM. It can be sometimes more than usual, sometimes less. Due to this randomness in the data $y_i$, we adjust the linear relation to $y_i = \alpha + \beta x_i + \eta_i$ for each $i$, where the "error terms" $\eta_i$'s are **zero-mean i.i.d gaussians** with variance $\sigma^2$. The zero mean is because we assume that errors may be positive or negative with no bias towards either side. The Gaussian is just because, well, CLT, plus it also gives the likelihood a very simple form.

---

Finally, note that there are usually some so-called **outliers** in our data - points where $\eta_i$ was unnaturally high - possible, yes - so that these points lie much farther from the predicted line. More outliers worsen our approximation of the actual distribution.

Outliers could also tell us whether our assumption on the distribution was a good one: If these outliers are reasonably small in number, then yes, that means our assumption that $Y$ was linear in $X$ is alright; otherwise, it means we have to cook up some other distribution for $X$ and $Y$.

# Interlude: More tools of the trade

Section 4.6

## Transformation of Random Variables

If I know the pdf $p$ of a random variable $X$, can I find the pdf $q$ of the random variable $g(X)$? This is what we address in this short but most useful chapter.

Let's take it step by step. Suppose $g$ was **strictly increasing**. Also let $Y = g(X)$. Then we have

$$P(a < Y < b) = P(g^{-1}(a) < X < g^{-1}(b))$$

so that for any $a, b$ we have

$$\int_a^b q(y)dy = \int_{g^{-1}(a)}^{g^{-1}(b)} p(x)dx = \int_a^b p(g^{-1}(t))(g^{-1}(t))'dt$$

so that

$$\boxed{q(y) \equiv p(g^{-1}(y)) \cdot (g^{-1}(y))'}$$

which is the equation we need. There are other similar ways to get the result (substituting $y = g(x)$ in the left integral, for example, or using the cdf (basically $a = -\infty, b = x$)), we here give another proof using the Leibniz rule.

We have $F(y) = P(Y \leq y) = P(X \leq g^{-1}(y)) = \int_{-\infty}^{g^{-1}(y)} p(x)dx$. So by the Leibniz rule $q(y) = F'(y) = g^{-1}(y)'p(g^{-1}(y))$ as required.

For $g$ **strictly decreasing**, we have a similar relation, except that

$$\int_a^b q(y)dy = P(a < Y < b) = P(g^{-1}(a) > X > g^{-1}(b)) = -\int_{g^{-1}(a)}^{g^{-1}(b)} p(x)dx$$

so that

$$q(y) \equiv -p(g^{-1}(y)) \cdot (g^{-1}(y))'$$

in this case. Note that $q$ is non-negative as we expect. In any case (increasing or decreasing) we have

$$q(y) \equiv p(g^{-1}(y)) \cdot |(g^{-1}(y))'|$$

Now for the general case where $g$ is neither strictly increasing nor decreasing. Here $Y = y = g(x)$ could be obtained at multiple $x$, so that there are possibly more terms to $P(a < Y < b)$. The terms must be added and then we can do the usual $u$-subsitution to find $q(y)$.

In practice, we do the following:

- Split the domain of $g$ into intervals where $g$ is just increasing and where $g$ is just decreasing.
- Compute the partial pdfs $q_i(y)$ for each interval.
- The required pdf is $q(y) = \sum_i q_i(y)$.

We now take a very useful application of random variable transformation, colloquially called the **inverse cdf** method.

Subsection 4.6.1

### The inverse cdf sampling method

This method solves the very important problem of being able to artificially sample from an arbitrary distribution that you know the pdf (or cdf) of.

We focus on continuous distributions. With discrete distributions you know the probabilities of each value for the random variable so it is quite easy to sample from (how exactly might one use this to sample?), so our main focus is on sampling from continuous distributions.

Let $X$ have the cdf $g$. Then $g$ is increasing (we assume strictly increasing, this is usually the case for continuous random variables)

Here's the trick: define $Y := g(X) \in [0, 1]$. So $Y$ is a random variable measuring the area under the pdf. We then have

$$F(y) = \int_{-\infty}^{g^{-1}(y)} p(x)dx = g(g^{-1}(y)) = y$$

so that the pdf $q(y) \equiv 1$, that is, $Y$ is uniformly distributed in $[0, 1]$! That is,

**For any random variable $X$ with a cdf $g$, $g(X)$ is uniform in $[0, 1]$.**

Then, to sample $X$, sample $Y$ from a uniform distribution (this is very easily done by a computer) and return $x = g^{-1}$(the sample)! This simulates the distribution of $X$. Why? Well, it's trivial: pdf (our sample) $= |(g^{-1})^{-1}(x)'|p_Y((g^{-1})^{-1}(x)) = |g'(x)| \cdot 1 = p_X(x)$, so indeed, we sample from $X$'s distribution!

Section 4.7

## Multidimensional random variables

We've seen random variables. Functions of one random variable. Functions of many random variables. What were their codomains? the set $\mathbb{R}$ of real numbers. These were, in the most general sense, one-dimensional multivariate functions. Here, we look at functions and random variables with a higher-dimensional codomain. Enter the Multidimensional random variable!

A random variable (in the most general sense) is a function $X : \Omega \to \mathbb{R}^d$.

$$X(s \in \Omega) = \begin{bmatrix} X_1(s) \\ X_2(s) \\ \vdots \\ X_d(s) \end{bmatrix}$$

where the functions $X_i$ are usual one-dimensional random variables $\Omega \to \mathbb{R}$. We shall denote a multidimensional random variable usually by $X$ itself, but when care is required as $\vec{X}$. Multidimensional random variables are also called random vectors for obvious reasons. We could, if we needed, similarly define random matrices and random tensors and so on, but since you can always flatten these into random vectors the theory is essentially complete using vectors alone.

The pdf of a random vector is a function $p : \mathbb{R}^d \to \mathbb{R}$. Like in the one-variable case, it is implicitly defined as a function $p$ such that for any region $A \subseteq \mathbb{R}^d$,

$$P(X \in A) = \int_{\vec{x} \in A} p(\vec{x}) d\tau$$

The cdf is a bit troublesome here, though, and we shall not be using it in the multidimensional setting.

Also, as with one-dimensional random variables, we shall be interested only in the distribution of the random variable in $\mathbb{R}^d$ as described by its pdf $p : \mathbb{R}^d \to \mathbb{R}$ and not on the particular sample space $\Omega$ whose distribution the rv is capturing.

Functions on multidimensional random variables are like before, albeit much more general. Let $X \in \mathbb{R}^n$ be a random variable and $f : \mathbb{R}^n \to \mathbb{R}^m$ be a function. Then $f(X)$ is defined just as you would expect, and is a random variable $\Omega \to \mathbb{R}^m$.

The mean of $\vec{X}$ is a vector, with the $i^{\text{th}}$ component being $\mathbb{E}[X_i]$.

$$\mathbb{E}[X] := \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_d] \end{bmatrix}$$

What should we define as the variance of $X$? Well, it should be the mean of the distance from the means. But what is the distance of $\vec{X}$ from $\mathbb{E}\left[\vec{X}\right]$? Let's use the standard euclidean norm. So we have

$$\text{var}(X) := \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] = \mathbb{E}\left[(X - \mathbb{E}[X])^T (X - \mathbb{E}[X])\right]$$

We can further simplify this as for one-dimensional rvs:

$$\begin{aligned} \text{var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^T (X - \mathbb{E}[X])\right] \\ &= \mathbb{E}\left[(X^T - \mathbb{E}[X]^T)(X - \mathbb{E}[X])\right] \\ &= \mathbb{E}\left[X^T X - 2X\mathbb{E}[X]^T + \mathbb{E}[X]^T \mathbb{E}[X]\right] \\ &= \mathbb{E}\left[X^T X\right] - \mathbb{E}[X]^T \mathbb{E}[X] \end{aligned}$$

Now, each $X_i$ is a one-dimensional random variable. So we can talk about the covariance of random variables $X_i$ and $X_j$. What is the value of the covariance? Precisely $\mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$. Suppose we define the covariance of the random vector $X$ to be a matrix of all the covariances between its components, like so:

$$C \in \mathcal{M}_{d \times d}(\mathbb{R})$$

$$C_{ij} = \text{cov}\left(X_i, X_j\right)$$

Let's try to find an expression for the matrix $C$. Suppose we define the random variable $X'$ by $X' = X - \mathbb{E}\left[X\right]$. Then

$$C_{ij} = \text{cov}\left(X_i, X_j\right) = \mathbb{E}\left[X_i' X_j'\right] = \mathbb{E}\left[(X'X'^T)_{ij}\right] = \mathbb{E}\left[X'X'^T\right]_{ij}$$

so that the covariance matrix $C$ is given by

$$C = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(X - \mathbb{E}\left[X\right])^T\right]$$

Notice the wonderful similarity between the variance and covariance - one is the expectation of an inner product, the other of an outer product. Quite neat, indeed! Just like the variance, the covariance can also be written in a simplified form as:

$$C = \mathbb{E}\left[XX^T\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[X\right]^T$$

Note also that the covariance is positive semidefinite (just like the variance is non-negative) since

$$v^T C v = \mathbb{E}\left[v^T X' X'^T v\right] = \mathbb{E}\left[\left\|X'^T v\right\|^2\right] \geq 0$$

for any $v \in \mathbb{R}^d$. Notice also that $\text{var}\left(X\right) = \text{trace}\left(\text{covar}(X)\right)$.

Subsection 4.7.1

### Change of variables in a multidimensional setting

Suppose we had a multidimensional rv $X \in \mathbb{R}^n$ with pdf $p : \mathbb{R}^n \to \mathbb{R}$. Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be invertible, and let random variable $Y$ be defined by $Y = g(X) \in \mathbb{R}^m$. Then what is the pdf $q$ of $Y$? We have for any $A \subseteq \mathbb{R}^m$ that $P(Y \in A) = P(X \in g^{-1}(A))$ so that

$$\int_A q(y) d\tau_{\mathbb{R}^m} = \int_{g^{-1}(A)} p(x) d\tau_{\mathbb{R}^n} = \int_A p(g^{-1}(t)) |\det\Big(Dg^{-1}(t)\Big)| d\tau_{\mathbb{R}^m}$$

which gives

$$\boxed{q(y) \equiv p(g^{-1}(y)) |\det\Big(Dg^{-1}(y)\Big)|}$$

which is a nice generalization of the corresponding result for a 1D random variable.

Section 4.8

## A quick introduction to Matrix calculus

As we delve into multidimensional random variables and the super useful multivariate gaussian, we need to learn some useful mathematical tools. We assume that the reader is familiar with basic multivariable calculus. In what follows, the term "object" is a placeholder for a scalar/vector/matrix (and in the most general case a tensor - but we will not be needing tensors here).

We define the derivative of an object $A$ with respect to a scalar $x$ denoted $\frac{\partial A}{\partial x}$ to be the object obtained by replacing every element of $A$ with its derivative wrt $x$. For example, if $A$ is a $m \times n$ matrix $A_{ij}$, then $\frac{\partial A}{\partial x}$ is an $m \times n$ matrix with

$$\left(\frac{\partial A}{\partial x}\right)_{ij} = \frac{\partial}{\partial x} A_{ij}$$

We define the derivative of an object $A$ "with respect to the vector of parameters $\mathbf{x}$", denoted $\frac{\partial A}{\partial \mathbf{x}}$ to be the following:

$$\frac{\partial A}{\partial \mathbf{x}} := \begin{pmatrix} \frac{\partial A}{\partial x_1} & \frac{\partial A}{\partial x_2} & \cdots & \frac{\partial A}{\partial x_n} \end{pmatrix}$$

This coincides with the gradient of $A$ in the case of $A$ being a scalar function of $x_1, \ldots, x_n$ and $\mathbf{x} = (x_1, \ldots, x_n)$. An important thing to note, here: we called $\mathbf{x}$ a "vector". It is rather more like a "list" of parameters here, but we shall use the terms interchangeably.

Now for an example. Let's try to calculate the derivative of the scalar $\mathbf{x}^T A \mathbf{x}$ wrt $\mathbf{x}$. First, before we calculate any derivatives, let's expand $\mathbf{x}^T A \mathbf{x}$.

$$\mathbf{x}^T A \mathbf{x} = \sum_j x_j (A\mathbf{x}) j = \sum_j x_j \sum_k A_{jk} x_k = \sum_{j,k} A_{jk} x_j x_k.$$

Noting that $\frac{\partial x_i}{\partial x_j} = \delta_{ij}$ (where $\delta_{ij}$ is the Kronecker delta which is 1 if $i = j$ and 0 otherwise), we have

$$\frac{\partial \mathbf{x}^T A \mathbf{x}}{\partial x_i} = \sum_{j,k} A_{jk} \frac{\partial x_j x_k}{\partial x_i} = \sum_{j,k} A_{jk} (x_j \delta_{ki} + x_k \delta_{ji}) = \sum_j A_{ji} x_j + \sum_k A_{ik} x_k$$

Next, note that $\sum_k A_{ik} x_k$ is just the $i^{\text{th}}$ component of the row vector $\mathbf{x}^T A^T$. Similarly, $\sum_j A_{ji} x_j = \sum_j (A^T)_{ij} x_j$ is just the $i^{\text{th}}$ component of the row vector $\mathbf{x}^T (A^T)^T = \mathbf{x}^T A$. Thus we get $\frac{\partial \mathbf{x}^T A \mathbf{x}}{\partial x_i} = \left[\mathbf{x}^T (A + A^T)\right]_i$ for each $i$, so that

$$\boxed{\frac{\partial \mathbf{x}^T A \mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^T (A + A^T)}$$

It is left as an exercise to the reader to show that $\frac{\partial A\mathbf{x}}{\partial x_i} = i^{\text{th}}$ column of $A$ (yes, $\frac{\partial A}{\partial x_i}$ can be a vector or matrix as well), which gives

$$\boxed{\frac{\partial A\mathbf{x}}{\partial \mathbf{x}} = A}$$

# Chapter 5

# The Multivariable Gaussian

**Definition 23**
The random vector $X \in \mathbb{R}^D$ is a multivariate gaussian if there exists a finite set $\{W_1, W_2, \dots W_N\}$ of i.i.d $G(0,1)$ Gaussian random variables with $N \geq D$ such that each $X_i$ can be represented as

$$X_d = \mu_d + \sum_{n=1}^{N} A_{dn} W_n$$

for some constants $A_{dn}$ and $\mu_d$, that is, there exist matrix $A \in \mathcal{M}_{D\times N}(\mathbb{R})$ and vector $\mu \in \mathbb{R}^D$ with

$$X = AW + \mu$$

So in essence, a linear transformation and a translation applied to a standard Gaussian random vector (which is what we'll call a random vector with components being i.i.d $G(0,1)$ RVs) gives you a general multivariate gaussian random vector.

Section 5.1
## pdf of the multivariate Gaussian

From hereon in this chapter, we work only for Gaussians with $N = D$, that is, A is square. We also assume that $A$ is invertible. At the end of the chapter, we briefly allude to the case when $N > D$.

What is the pdf $p : \mathbb{R}^D \to \mathbb{R}$ of $W$? We have

$$\begin{aligned} p(\vec{w}) = p(w_1, w_2, \dots w_D) &= P(W = (w_1, w_2, \dots w_D)) \\ &= P(W_1 = w_1, W_2 = w_2, \dots W_D = w_D) \\ &= \prod_{d=1}^{D} P(W_d = w_d) \\ &= \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{w^T w}{2}\right) \end{aligned}$$

We now use the transformation $X = g(W) = AW + \mu$. We have

$$p_X(x) = p_W(g^{-1}(x)) \cdot |\det\Big(Dg^{-1}(x)\Big)'|$$

Here $g^{-1}(X) = A^{-1}(X - \mu)$, so $Dg^{-1} = A^{-1}$. The pdf thus becomes

$$\begin{aligned} p_X(x) &= \frac{1}{(2\pi)^{D/2}|\det(A)|} \exp\left(-\frac{(A^{-1}(X-\mu))^T(A^{-1}(X-\mu))}{2}\right) \\ &= \frac{1}{(2\pi)^{D/2}|\det(A)|} \exp\left(-\frac{(X-\mu)^T(A^T)^{-1}A^{-1}(X-\mu)}{2}\right) \\ &= \frac{1}{(2\pi)^{D/2}|\det(A)|} \exp\left(-\frac{(X-\mu)^T C^{-1}(X-\mu)}{2}\right) \end{aligned}$$

where $C = AA^T$ is positive definite (since $A$ is invertible).

SUBSECTION 5.1.1

## Level sets of the multivariate Gaussian pdf

The level sets of the pdf $p_X$ are given by

$$\{x : (x-\mu)^T C^{-1}(x-\mu) = c\} \iff \{y + \mu : y^T C^{-1} y = c\}$$

Now $y^T C^{-1} y$ can be written (where $C$ is diagonalized as $VDV^T$)

$$y^T C^{-1} y = y^T (VD^{-1}V^T)y = (V^T y)^T D^{-1}(V^T y)$$

so setting $z := Vy$ the equation $y^T C^{-1} y = c$ reduces to

$$y^T C^{-1} y = z^T D^{-1} z = \sum_{i=1}^{D} \frac{z_i^2}{\lambda_i} = c \iff \sum_{i=1}^{D} \left(\frac{z_i}{\sqrt{c\lambda_i}}\right)^2 = 1$$

where $z = (z_1, \ldots, z_n)$ and $\{\lambda_i\}_{i=1}^D$ are the (positive) eigenvalues of $C$. So the level sets are hyper-ellipsoids in the $z$-system with axes along $e_i$ (the cardinal directions). Now, since $y = Vz$, in the $y$-system the ellipsoid's axes are along $y = Vz = Ve_i = v_i$ where $\{v_i\}_{i=1}^D$ are the eigenvectors of $C$. Note that these directions are independent of the constant $c$.

So finally in the $x$-system, the level sets of the pdf are hyper-ellipsoids centered at $x = \mu$ with axes along $v_i$, where $\{v_i\}_{i=1}^D$ are the eigenvectors of the covariance $C$.

SUBSECTION 5.1.2

## Mean, variance and covariance of a multivariate Gaussian random variable

We have $\mathbb{E}[W] = \vec{0}$ since each component of $W$ has mean 0. Further, since each component has variance 1 and the components have pairwise covariance 0 (they are independent) we have $\text{cov}(W) = I_{D\times D}$. Now let $X = AW + \mu$. We obtain

$$\mathbb{E}[X] = \mathbb{E}[AW] + \mu = A \cdot \mathbb{E}[W] + \mu = A\vec{0} + \mu = \mu$$

Onto the covariance.

$$\begin{aligned}\operatorname{cov}(X) &= \operatorname{cov}(AW+\mu)\\ &= \mathbb{E}\left[(AW+\mu-\mu)(AW+\mu-\mu)^T\right]\\ &= \mathbb{E}\left[AWW^TA^T\right]\\ &= A\left(\mathbb{E}\left[WW^T\right]\right)A^T\\ &= AIA^T = AA^T\end{aligned}$$

and that's why we decided to denote $AA^T$ in the pdf by $C$!

The variance now follows immediately: $\operatorname{var}(X) = \operatorname{trace}(C) = \operatorname{trace}(AA^T)$.

Section 5.2

## ML Estimates for the multivariate Gaussian

Consider the *vector* data $\{\vec{X}_i\}_{i=1}^N$ (could be pixels in an image, etc).

Done in notes, p55-56

Section 5.3

## Marginals and Conditionals

Gaussian-ness is invariant under all arithmetic operations. Which means, that the marginals and Conditionals (under arbitrary constraint matrix) are also gaussians.

Section 5.4

## Principal Component Analysis

In this section, our 'data' is a sample of $N$ vectors, each vector representing an observation. Suppose the vectors are in $D$-dimensional space. The space of the sample is the region of space that the sample points occupy in $\mathbb{R}^D$.

In typical data, a fair number of components of each vector in the sample are usually dependent, meaning that the space occupied by the sample is not thrown all around $\mathbb{R}^D$ arbitrarily, rather, most (except some outliers due to imperfect data etc) vectors are contained in a particular subspace of $\mathbb{R}^D$. Finding this subspace is especially useful when $D$ and $N$ are large (as is usually the case, for example in the MNIST dataset there are $N = 60000$ samples, each one in $D = 28^2 = 784$ dimensional space), to greatly speed up computation.

Principal Component Analysis does exactly this. The idea is the following. todo:

- max dispersion along one direction (1d subspace). then max variance if orth to $v_1$, so on.
- Max D-dimensional subspace using $D = 2$ as nice example.

Section 5.5

## The case N > D

Suppose we had a Gaussian $X_{D\times 1} = A_{D\times N}W_{N\times 1} + \mu_{D\times 1}$ where $N > D$. We show that you can infact choose a "better" $A$ which is $D \times D$ instead. The idea is to use the SVD on $A$. $A$ can be written $A_{D\times N} = U_{D\times D}S_{D\times N}V^T_{N\times N}$ where $U, V$ are orthogonal and $D$ is rectangular diagonal with non-negative reals on the first $D < N$ diagonal elements, all other entries are $0$. So $D$ can be written $D = \left[\Lambda_{D\times D}|O\right]$. Now consider $AW$. We have

$$AW = USV^TW = USX^1 = U\left[\Lambda_{D\times D}|O\right]\begin{bmatrix} X^2_{D\times 1} \\ \\ X^3 \end{bmatrix} = U\Lambda X^2 = A^1X^2$$

where $X^1 = V^TW$ is a $N$-dimensional gaussian, $X^2$ = first $D$ rows of $X^1$ ($D$-dimensional marginal) is also a gaussian but in $D$ dimensions, that is, $X^2 = A^2W_0 + \chi_D$ for a $D \times D$ matrix $A^2$ and vector $\chi_D$ where $W_0$ is the $D$-dimensional standard gaussian. Also, $A^1$ is $D \times D$. All this put together gives that

$$X = AW + \mu = A^1(A^2W_0 + \chi_D) + \mu = A'W_0 + \mu'$$

which means that $X$ can be represented as a $D$-dimensional gaussian with $A$ being $D \times D$. Hence, the case $N > D$ is essentially (after slicing $X^1$ to $X^2$) the case of $N = D$. So given a multivariate Gaussian in $\mathbb{R}^D$, we may as well set it to be equal to $AW_0 + \mu$ for $A$ being $D \times D$. Oh, and also, since in the $N = D$ case our analysis relies on the fact that $A$ is invertible, we must have that rank $A_{D\times N} = D$ so that $A'$ is invertible.

Just for fun, let's calculate the mean and covariance for the case $N > D$.

# Chapter 6

# Bayesian Analysis

This might be the most fun part of the whole course. In this chapter, we look at another, totally different way of estimating parameters of a distribution - estimating parameters based on what you believe they are.

The key idea is the following: The parameter $\theta$ for a distribution is some number that we don't know and want to estimate. How would we do so? Well, Maximum Likelihood estimation is one way, where we treat $\theta$ as an unknown constant and the MLE theorem gave us a way (an approximate way, since $N$ is finite) to find a good (and consistent) estimate for the parameter from the data. Suppose now that I instead tell you that $\theta$ itself has a distribution - that is, $\theta$ is *more likely to be* $\theta_1$ than $\theta_2$, and the like. And so you would expect - if what you think the distribution for $\theta$ is actually not that bad - that $\theta_{\text{true}}$ is the mode of the distribution for $\theta$, since that's what $\theta$ is most likely to be.

Take a moment for this to sink in. We are saying the following:

> $\theta$ has an associated distribution describing how likely it is for $\theta_{\text{true}}$ to be equal to a particular value $\theta$ for each value $\theta \in \mathbb{R}$. In other words, $\theta$ is now a random variable with its own probability distribution.

Section 6.1

## Bayes' Rule

The rule is quite trivial just by itself, but it is extremely useful.

**Theorem 22**

Consider random variables $X$ and $Y$ with a joint distribution $P(X, Y)$. Let the notation $P(X|Y)$ refer to the conditional of $X$ wrt $Y$ on this joint distribution $P(X, Y)$. Then we have the **Bayes' Rule –**

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} = \frac{P(Y|X)P(X)}{\sum_{x \in \mathscr{A}_X} P(Y|x)P(x)}$$

Proof The conditional (did we define it before?) is defined to be

$$P(X|Y) := \frac{P(X,Y)}{P(Y)}$$

where $P(Y) := \sum_{x \in \mathscr{A}_X} P(x,Y)$ is the marginal of $Y$ in the joint probability distribution. Similarly $P(Y|X) := \frac{P(X,Y)}{P(X)}$. The theorem follows by noting that $P(X,Y) = P(Y)P(X|Y) = P(X)P(Y|X)$. □

There is a slight subtlety here. Note that the conditional is only defined when the variable we are conditioning with respect to has a nonzero probability (since it is in the denominator). So Bayes' rule in the form as shown is applicable only when $P(X), P(Y) \neq 0$.

Now we look at some standard terms used in Bayesian Analysis.

- **The random variable $X$:** The random variable representing the parameters we want to find. Often we replace the symbol $X$ by a capital version of the symbol usually used to denote the parameters for the distribution we are trying to fix.
- **The random variable $Y$:** The random variable representing the data we receive from the unknown distribution.
- **The prior distribution $P(x)$:** The probability distribution of the parameter $X$ *before* we receive any data.
- **The likelihood distribution $P(y|x)$:** The conditional probability of the data being $y$ given a particular value $x$ of the parameter. Essentially, this tells you how likely it is that your value $x$ for the parameter is actually the right one. This is a similar idea to that of the likelihood function seen before.
- **The evidence distribution $P(y)$:** The probability distribution of the data across values of the parameter - that is, for each $y \in \mathscr{A}_Y$, $P(y) = \sum_x P(x,y)$.
- **The posterior distribution $P(x|y)$:** The conditional probability of the parameter being $x$ given that the data received is $y$. This is what we usually take to be the "correct distribution for the parameter $X$" till more data comes in.

Bayes' rule can then be more succintly rephrased as

$$\text{Posterior}(x,y) = \frac{\text{Likelihood}(x,y) \times \text{Prior}(x)}{\text{Evidence}(y)}$$

Notice that the evidence is not a function of the parameter $x$. So, leveraging that the posterior $P(x|y)$ is a distribution (for any given $y$), we can simply use

$$\text{Posterior}(x,y) \propto \text{Likelihood}(x,y) \times \text{Prior}(x)$$

and then normalize it over $x$ at the end. (Indeed, the evidence is the product of the likelihood and the prior across all values of the parameter $X$).

Section 6.2

# Bayesian Inference 1 - MAP estimation

This is where we first see why all of this is useful. Consider tossing a biased coin where you don't know $p \in [0, 1]$. Before tossing it, you have a **prior** belief that $p = 0.5$ 80% of the time and $p = 0.7$ otherwise - a discrete distribution on $p$. Why this belief? Well, it's your choice. All of estimation is your choice. Of course, to do a good job of estimating from the posterior, one needs to choose a good prior - more on that later. But right now, this is what you think the distribution of random variable $X$ (represents the distribution of parameter $p$) is.

Now you toss the coin $N = 50$ times to "test your prior". Suppose you obtained $y = 31$ heads. What does this tell you? Of course, it tells you that your prior is not quite correct. The cool part is that since you now have the data $y$, you can compute the posterior distribution $P(x|y)$. And since this accounts for the data as well, you have another distrbution for the parameter $p$ - one that used both your original prior and the data for its construction.

Let's work this out concretely. First, $X$ represents the distribution for parameter $p$, and $Y$ is a binomial random variable at each $x$. We now choose the joint distribution $P(x, y)$ to be

$$P(x, y) = P_X(x) \cdot P_{\text{binom}}(y; N = N, p = x)$$

where $P_X$ is our prior function and $P_{\text{binom}}$ is the familiar binomial distribution that we place on phenomena like coin tosses. The parameter $p$ of the binomial distribution is to be determined. Note that $P(x) := \sum_y P(x, y) = P_X(x) \sum_y P_{\text{binom}}(y; N, x) = P_X(x)$ here. So our prior $P_X(x)$ is exactly what we plug in for $P(x)$ in the Bayes' rule. This will be a general idea - we will not explicitly specify the joint distribution, that the joint distribution is exactly this - the prior times the distribution parametrized with parameter $x$ - is understood.

Here our prior was taken to be

$$P_X(x) = \begin{cases} 0.8 & x = 0.5 \\ 0.2 & x = 0.7 \\ 0 & \text{otherwise} \end{cases}$$

for $x \in [0, 1]$ (range of the parameter $p$).

Similarly, we have the likelihood

$$P(y|x) = P_{\text{binom}}(y; N, x)$$

from the joint distribution and the evidence

$$P(y) = \sum_x P(x, y) = P_X(0.5) \cdot P_{\text{binom}}(y; N, 0.5) + P_X(0.7) \cdot P_{\text{binom}}(y; N, 0.7)$$

Notice the evidence is just a product of the prior and the likelihood at each $x$. Thanks to this, we shall not need to worry about the evidence and can straightaway compute the product of the prior and the likelihood, and later normalize, taking care of the evidence term.

Hence, plugging in $y = 31$ we have the posterior

$$P_X(x|31) \propto P_X(x)\binom{50}{31}x^{31}(1-x)^{19}$$

for $x = 0.5, 0.7$ and zero otherwise. Computing we get

$$P(0.5|31) =$$