Introduction to Probability Theory and Statistics

August 21, 2008

Contents

Chapter 1

Probability

Probability theory provides a mathematical foundation to concepts such as "probability", "information", "belief", "uncertainty", "confidence", "randomness", "variability", "chance" and "risk". Probability theory is important to empirical scientists because it gives them a rational framework to make inferences and test hypotheses based on uncertain empirical data. Probability theory is also useful to engineers building systems that have to operate intelligently in an uncertain world. For example, some of the most successful approaches in machine perception (e.g., automatic speech recognition, computer vision) and artificial intelligence are based on probabilistic models. Moreover probability theory is also proving very valuable as a theoretical framework for scientists trying to understand how the brain works. Many computational neuroscientists think of the brain as a probabilistic computer built with unreliable components, i.e., neurons, and use probability theory as a guiding framework to understand the principles of computation used by the brain. Consider the following examples:

- You need to decide whether a coin is loaded (i.e., whether it tends to favor one side over the other when tossed). You toss the coin 6 times and in all cases you get "Tails". Would you say that the coin is loaded?

- You are trying to figure out whether newborn babies can distinguish green from red. To do so you present two colored cards (one green, one red) to 6 newborn babies. You make sure that the 2 cards have equal overall luminance so that they are indistinguishable if recorded by a black and white camera. The 6 babies are randomly divided into two groups. The first group gets the red card on the left visual field, and the second group on the right

visual field. You find that all 6 babies look longer to the red card than the green card. Would you say that babies can distinguish red from green?

- A pregnancy test has a 99 % validity (i.e., 99 of of 100 pregnant women test positive) and 95 % specificity (i.e., 95 out of 100 non pregnant women test negative). A woman believes she has a 10 % chance of being pregnant. She takes the test and tests positive. How should she combine her prior beliefs with the results of the test?

- You need to design a system that detects a sinusoidal tone of 1000Hz in the presence of white noise. How should design the system to solve this task optimally?

- How should the photo receptors in the human retina be interconnected to maximize information transmission to the brain?

While these tasks appear different from each other, they all share a common problem: The need to combine different sources of uncertain information to make rational decisions. Probability theory provides a very powerful mathematical framework to do so. Before we go into mathematical aspects of probability theory I shall tell you that there are deep philosophical issues behind the very notion of probability. In practice there are three major interpretations of probability, commonly called the frequentist, the Bayesian or subjectivist, and the axiomatic or mathematical interpretation.

1. **Probability as a relative frequency**

   This approach interprets the probability of an event as the proportion of times such ane event is expected to happen in the long run. Formally, the probability of an event E would be the limit of the relative frequency of occurrence of that event as the number of observations grows large

   $$P(\mathrm{E}) = \lim_{n\to\infty} \frac{n_{\mathrm{E}}}{n} \tag{1.1}$$

   where n_E is the number of times the event is observed out of a total of n independent experiments. For example, we say that the probability of "heads" when tossing a coin is 0.5. By that we mean that if we toss a coin many many times and compute the relative frequency of "heads" we expect for that relative frequency to approach 0.5 as we increase the number of tosses.

This notion of probability is appealing because it seems objective and ties our work to the observation of physical events. One difficulty with the approach is that in practice we can never perform an experiment an infinite number of times. Note also that this approach is behaviorist, in the sense that it defines probability in terms of the observable behavior of physical systems. The approach fails to capture the idea of probability as internal knowledge of cognitive systems.

2. **Probability as uncertain knowledge.**

   This notion of probability is at work when we say things like "I will probably get an A in this class". By this we mean something like "Based on what I know about myself and about this class, I would not be very surprised if I get an A. However, I would not bet my life on it, since there are a multitude of factors which are difficult to predict and that could make it impossible for me to get an A". This notion of probability is "cognitive" and does not need to be directly grounded on empirical frequencies. For example, I can say things like "I will probably die poor" even though I will not be able to repeat my life many times and count the number of lives in which I die poor.

   This notion of probability is very useful in the field of machine intelligence. In order for machines to operate in natural environments they need knowledge systems capable of handling the uncertainty of the world. Probability theory provides an ideal way to do so. Probabilists that are willing to represent internal knowledge using probability theory are called "Bayesian", since Bayes is recognized as the first mathematician to do so.

3. **Probability as a mathematical model.** Modern mathematicians avoid the frequentist vs. Bayesian controversy by treating probability as a mathematical object. The role of mathematics here is to make sure probability theory is rigorously defined and traceable to first principles. From this point of view it is up to the users of probability theory to apply it to whatever they see fit. Some may want to apply it to describe limits of relative frequencies. Some may want to apply it to describe subjective notions of uncertainty, or to build better computers. This is not necessarily of concern to the mathematician. The application of probability theory to those domains will be ultimately judged by its usefulness.

1.1 Intuitive Set Theory

We need a few notions from set theory before we jump into probability theory. In doing so we will use intuitive or "naive" definitions. This intuitive approach provides good mnemonics and is sufficient for our purposes but soon runs into problems for more advanced applications. For a more rigorous definition of set theoretical concepts and an explanation of the limitations of the intuitive approach you may want to take a look at the Appendix.

- **Set**: A set is a collection of elements. Sets are commonly represented using curly brackets containing a collection of elements separated by commas. For example

$$A = \{1, 2, 3\} \tag{1.2}$$

tells us that A is a set whose elements are the first 3 natural numbers. Sets can also be represented using a rule that identifies the elements of the set. The prototypical notation is as follows

$$\{x : x \text{ follows a rule}\} \tag{1.3}$$

For example,

$$\{x : x \text{ is a natural number and } x \text{ is smaller than } 4\ \} \tag{1.4}$$

- **Outcome Space**: The outcome space is a set whose elements are all the possible basic outcomes of an experiment.[1] The sample space is also called **sample space**, **reference set**, and **universal set** and it is commonly represented with the capital Greek letter "omega", Ω. We call the elements of the sample space "outcomes" and represent them symbolically with the small Greek letter "omega", ω.

Example 1: If we roll a die, the outcome space could be

$$\Omega = \{1, 2, 3, 4, 5, 6\} \tag{1.5}$$

In this case the symbol ω could be used to represent either 1,2,3,4,5 or 6.

[1] The empty set is not a valid sample space.

Example 2: If we toss a coin twice, we can observe 1 of 4 outcomes: (Heads, Heads), (Heads, Tails), (Tails, Heads), (Tails, Tails). In this case we could use the following outcome space

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\} \tag{1.6}$$

and the symbol ω could be used to represent either (H, H), or (H, T), or (T, H), or (T, T). Note how in this case each basic outcome contains 2 elements. If we toss a coin n times each basic outcome ω would contain n elements.

- **Singletons:** A singleton is a set with a single element. For example the set $\{4\}$ is a singleton, since it only has one element. On the other hand 4 is not a singleton since it is an element not a set.[2]

- **Element inclusion:** We use the symbol $\in$ to represent element inclusion. The expression $\omega \in A$ tells us that ω is an element of the set A. The expression $\omega \notin A$ tells us that ω is **not** an element of the set A. For example, $1 \in \{1, 2\}$ is true since 1 is an element of the set $\{1, 2\}$. The expression $\{1\} \in \{\{1\}, 2\}$ is also true since the singleton $\{1\}$ is an element of the set $\{\{1\}, 2\}$. The expression $\{1\} \notin \{1, 2\}$ is also true, since the set $\{1\}$ is not an element of the set $\{1, 2\}$.

- **Set inclusion:** We say that the set A is included in the set B or is a **subset** of B if all the elements of A are also elements of B. We represent set inclusion with the symbol $\subset$. The expression $A \subset B$ tells us that both A and B are sets and that all the elements of A are also elements of B. For example the expression $\{1\} \subset \{1, 2\}$ is true since all the elements of the set $\{1\}$ are in the set $\{1, 2\}$. On the other hand $1 \subset \{1, 2\}$ is not true since 1 is an element, not a set.[3]

- **Set equality:** Two sets A and B are equal if all elements of A belong to B and all elements of B belong to A. In other words, if $A \subset B$ and $B \subset A$. For example the sets $\{1, 2, 3\}$ and $\{3, 1, 1, 2, 1\}$ are equal.

- **Set Operations:** There are 3 basic set operations:

[2] The distinction between elements and sets does not exist in axiomatic set theory, but it is useful when explaining set theory in an intuitive manner.

[3] For a more rigorous explanation see the Appendix on axiomatic set theory.

1. **Union**: The union of two sets A and B is another set that includes all elements of A and all elements of B. We represent the union operator with this symbol $\cup$

   For example, if $A = \{1, 3, 5\}$ and $B = \{2, 3, 4\}$, then $A \cup B = \{1, 2, 3, 4, 5\}$. More generally

   $$A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\} \tag{1.7}$$

   In other words, the set $A \cup B$ is the set of elements with the property that they either belong to the set A or to the set B.

2. **Intersection**: The intersection of two sets A and B is another set C such that all elements in C belong to A and to B. The intersection operator is symbolized as $\cap$. If $A = \{1, 3, 5\}$ and $B = \{2, 3, 4\}$ then $A \cap B = \{3\}$. More generally

   $$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\} \tag{1.8}$$

3. **Complementation**: The complement of a set A with respect to a reference set Ω is the set of all elements of Ω which do not belong to A. The complement of A is represented as A^c. For example, if the universal set is $\{1, 2, 3, 4, 5, 6\}$ then the complement of $\{1, 3, 5\}$ is $\{2, 4, 6\}$. More generally

   $$A^c = \{\omega : \omega \in \Omega \text{ and } \omega \notin A\} \tag{1.9}$$

- **Empty set**: The empty set is a set with no elements. We represent the null set with the symbol $\varnothing$. Note $\Omega^c = \varnothing$, $\varnothing^c = \Omega$, and for any set A

  $$A \cup \varnothing = A \tag{1.10}$$
  $$A \cap \varnothing = \varnothing \tag{1.11}$$

- **Disjoint sets**: Two sets are disjoint if they have no elements in common, i.e., their intersection is the empty set. For example, the sets $\{1, 2\}$ and $\{1\}$ are not disjoint since they have an element in common.

- **Collections**: A collection of sets is a set of sets, i.e., a set whose elements are sets. For example, if A and B are the sets defined above, the set $\{A, B\}$ is a collection of sets.

- **Power set**: The power set of a set A is the a collection of all possible sets of A. We represent it as $\mathfrak{P}(A)$. For example, if $A = \{1, 2, 3\}$ then

$$\mathfrak{P}(A) = \{\varnothing, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, A\} \tag{1.12}$$

  Note that 1 is not an element of $\mathfrak{P}(A)$ but $\{1\}$ is. This is because 1 is an element of A, not a set of A.

- **Collections closed under set operations**: A collection of sets is closed under set operations if any set operation on the sets in the collection results in another set which still is in the collection. If $A = \{1, 3, 5\}$ and $B = \{2, 3, 4\}$, the collection $\mathcal{C} = \{A, B\}$ is not closed because the set $A \cap B = \{3\}$ does not belong to the collection. The collection $\mathcal{C} = \{\Omega, \varnothing\}$ is closed under set operations, all set operations on elements of $\mathcal{C}$ produce another set that belongs to $\mathcal{C}$. The power set of a set is always a closed collection.

- **Sigma algebra**: A sigma algebra is a collection of sets which is closed when set operations are applied to its members a countable number of times. The power set of a set is always a sigma algebra.

- **Natural numbers**: We use the symbol $\mathbb{N}$ to represent the natural numbers, i.e., $\{1, 2, 3, \ldots\}$. One important property of the natural numbers is that if $x \in \mathbb{N}$ then $x + 1 \in \mathbb{N}$.

- **Integers**: We use the symbol $\mathbb{Z}$ to represent the set of integers, i.e., $\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$. Note $\mathbb{N} \subset \mathbb{Z}$. One important property of the natural numbers is that if $x \in \mathbb{Z}$ then $x + 1 \in \mathbb{Z}$ and $x - 1 \in \mathbb{Z}$.

- **Real numbers**: We use the symbol $\mathbb{R}$ to represent the real numbers, i.e., numbers that may have an infinite number of decimals. For example, 1, 2.35, $-4/123$, $\sqrt{2}$, and π, are real numbers. Note $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{R}$.

- **Cardinality of sets**:

  - We say that a set is **finite** if it can be put in one-to-one correspondence with a set of the form $\{1, 2, \ldots, n\}$, where n is a fixed natural number.
  - We say that a set is **infinite countable** if it can be put in one-to-one correspondence with the natural numbers.
  - We say that a set is **countable** if it is either finite or infinite countable.

- We say that a set is **infinite uncountable** if it has a subset that can be put in one-to-one correspondence with the natural numbers, but the set itself cannot be put in such a correspondence. This includes sets that can be put in one-to-one correspondence with the real numbers.

1.2 Events

We have defined outcomes as the elements of a reference set Ω. In practice we are interested in assigning probability values not only to outcomes but also to sets of outcomes. For example we may want to know the probability of getting an even number when rolling a die. In other words, we want the probability of the set $\{2, 4, 6\}$. In probability theory set of outcomes to which we can assign probabilities are called **events**. The collection of all events is called the **event space** and is commonly represented with the letter $\mathcal{F}$. Not all collections of sets qualify as event spaces. To be an event space, the collection of sets has to be a sigma algebra (i.e., it has to be closed under set operations). Here is an example:

Example: Consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. Is the collection of sets $\{\{1, 2, 3\}, \{4, 5, 6\}\}$ a valid event space?
Answer: No, it is not a valid event space because the union of $\{1, 2, 3\}$ and $\{4, 5, 6\}$ is the set $\Omega = \{1, 2, 3, 4, 5, 6\}$ which does not belong to $\mathcal{F}$. On the other hand the set $\{\varnothing, \{1, 2, 3\}, \{4, 5, 6\}, \Omega\}$ is a valid event space. Any set operation using the sets in $\mathcal{F}$ results into another set which is in $\mathcal{F}$.

Note: The outcome space Ω and the event space $\mathcal{F}$ are different sets. For example if the outcome space were $\Omega = \{H, T\}$ a valid event space would be $\mathcal{F} = \{\Omega, \varnothing, \{H\}, \{T\}\}$. Note that $\Omega \neq \mathcal{F}$. The outcome space contains the basic outcomes of an experiments. The event space contains sets of outcomes.

1.3 Probability measures

When we say that the probability of rolling an even number is 0.5, we can think of this as an assignment of a number (i.e., 0.5) to a set ,i.e., to the set $\{2, 4, 6\}$. Mathematicians think of probabilities as function that "measures" sets, thus the name **probability measure**. For example, if the probability of rolling an even number on a die is 0.5, we would say that the probability measure of the set

$\{2, 4, 6\}$ is 0.5. Probability measures are commonly represented with the letter P (capitalized). Probability measures have to follow three constraints, which are known as Kolmogorov's axioms:

1. The probability measure of events has to be larger or equal to zero: $P(A) \geq 0$ for all $A \in \mathcal{F}$.

2. The probability measure of the reference set is 1

$$P(\Omega) = 1 \tag{1.13}$$

3. If the sets $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint then

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots \tag{1.14}$$

Example 1: A fair coin. We can construct a probability space to describe the behavior of a coin. The outcome space consists of 2 elements, representing heads and tails $\Omega = \{H, T\}$. Since Ω is finite, we can use as the event space the set of all sets in Ω, also known as the power set of Ω. In our case, $\mathcal{F} = \{\{H\}, \{T\}, \{H, T\}, \varnothing\}$. Note $\mathcal{F}$ is closed under set operations so we can use it as an event space.

The probability measure P in this case is totally defined if we simply say $P(\{H\}) = 0.5$. The outcome of P for all the other elements of $\mathcal{F}$ can be inferred: we already know $P(\{H\}) = 0.5$ and $P(\{H, T\}) = 1.0$. Note the sets $\{H\}$ and $\{T\}$ are disjoint, moreover $\{H\} \cup \{T\} = \Omega$, thus using the probability axioms

$$P(\{H, T\}) = 1 = P(\{H\}) + P(\{T\}) = 0.5 + P(\{T\}) \tag{1.15}$$

from which it follows $P(\{T\}) = 0.5$. Finally we note that Ω and $\varnothing$ are disjoint and their union is Ω, using the probability axioms it follows that

$$1 = P(\Omega) = P(\Omega \cup \varnothing) = P(\Omega) + P(\varnothing) \tag{1.16}$$

Thus $P(\varnothing) = 0$. Note P qualifies as a probability measure: for each element of $\mathcal{F}$ it assigns a real number and the assignment is consistent with the three axiom of probability.

Example 2: A fair die. In this case the outcome space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, the event space is the power set of Ω, the set of all sets of Ω, $\mathcal{F} = \mathfrak{P}(\Omega)$, and $P(\{i\}) = 1/6, \text{ for } i = 1, \ldots, 6$. I will refer to this as the fair die probability space.

Example 3: A loaded die. We can model the behavior of a loaded die by assigning non negative weight values to each side of the die. Let w_i represent the weight of side i. In this case the outcome space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, the event space is the power set of Ω, the set of all sets of Ω, $\mathcal{F} = \mathfrak{P}(\Omega)$, and

$$P(\{i\}) = w_i/(w_1 + \cdots + w_6), \tag{1.17}$$

Note that if all weight values are equal, this probability space is the same as the probability space in Example 2.

1.4 Joint Probabilities

The joint probability of two or more events is the probability of the intersection of those events. For example consider the events $A_1 = \{2, 4, 6\},\ A_2 = \{4, 5, 6\}$ in the fair die probability space. Thus, A_1 represents obtaining an even number and A_2 obtaining a number larger than 3.

$$P(A_1) = P(\{2\} \cup \{4\} \cup \{6\}) = 3/6 \tag{1.18}$$

$$P(A_2) = P(\{4\} \cup \{5\} \cup \{6\}) = 3/6 \tag{1.19}$$

$$P(A_1 \cap A_2) = P(\{4\} \cup \{6\}) = 2/6 \tag{1.20}$$

Thus the joint probability of A_1 and A_2 is $1/3$.

1.5 Conditional Probabilities

The conditional probability of event A_1 given event A_2 is defined as follows

$$P(A_1 \mid A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)} \tag{1.21}$$

Mathematically this formula amounts to making A_2 the new reference set, i.e., the set A_2 is now given probability 1 since

$$P(A_2 \mid A_2) = \frac{P(A_2 \cap A_2}{P(A_2)} = 1 \tag{1.22}$$

Intuitively, Conditional probability represents a revision of the original probability measure P. This revision takes into consideration the fact that we know the event A_2 has happened with probability 1. In the fair die example,

$$P(A_1 \mid A_2) = \frac{1/3}{3/6} = \frac{2}{3} \tag{1.23}$$

in other words, if we know that the toss produced a number larger than 3, the probability that the number is even is 2/3.

1.6 Independence of 2 Events

The notion of independence is crucial. Intuitively two events A_1 and A_2 are independent if knowing that A_2 has happened does not change the probability of A_1. In other words

$$P(A_1 \mid A_2) = P(A_1) \tag{1.24}$$

More generally we say that the events A and A_2 are independent if and only if

$$P(A_1 \cap A_2) = P(A_1)P(A_2) \tag{1.25}$$

In the fair die example, $P(A_1 \mid A_2) = 1/3$ and $P(A_1) = 1/2$, thus the two events are not independent.

1.7 Independence of n Events

We say that the events $A_1, \ldots, A_n$ are independent if and only if the following conditions are met:

1. All pairs of events with different indexes are independent, i.e.,

$$P(A_i \cap A_j) = P(A_i)P(A_j) \tag{1.26}$$

   for all $i, j \in \{1, 2, \ldots, n\}$ such that $i \neq j$.

2. For all triplets of events with different indexes

$$P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k) \tag{1.27}$$

   for all $i, j, k \in \{1, \ldots, n\}$ such that $i \neq j \neq k$.

3. Same idea for combinations of 3 sets, 4 sets, ...

4. For the n-tuple of events with different indexes

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n) \tag{1.28}$$

You may want to verify that $2^n - n - 1$ conditions are needed to check whether n events are independent. For example, $2^3 - 3 - 1 = 4$ conditions are needed to verify whether 3 events are independent.

Example 1: Consider the fair-die probability space and let $A_1 = A_2 = \{1, 2, 3\}$, and $A_3 = \{3, 4, 5, 6\}$. Note

$$P(A_1 \cap A_2 \cap A_3) = P(\{3\}) = P(A_1)P(A_2)P(A_3) = 1/6 \tag{1.29}$$

However

$$P(A_1 \cap A_2) = 3/6 \neq P(A_1)P(A_2) = 9/36 \tag{1.30}$$

Thus A_1, A_2, A_3 are not independent.

Example 2: Consider a probability space that models the behavior a weighted die with 8 sides: $\Omega = (1, 2, 3, 4, 5, 6, 7, 8)$, $\mathcal{F} = \mathfrak{P}(\Omega)$ and the die is weighted so that

$$P(\{2\}) = P(\{3\}) = P(\{5\}) = P(\{8\}) = 1/4 \tag{1.31}$$
$$P(\{1\}) = P(\{4\}) = P(\{6\}) = P(\{7\}) = 0 \tag{1.32}$$

Let the events A_1, A_2, A_3 be as follows

$$A_1 = \{1, 2, 3, 4\} \tag{1.33}$$
$$A_2 = \{1, 2, 5, 6\} \tag{1.34}$$
$$A_3 = \{1, 3, 5, 7\} \tag{1.35}$$

Thus $P(A_1) = P(A_2) = P(A_3) = 2/4$. Note

$$P(A_1 \cap A_2) = P(A_1)P(A_2) = 1/4 \tag{1.36}$$
$$P(A_1 \cap A_3) = P(A_1)P(A_3) = 1/4 \tag{1.37}$$
$$P(A_2 \cap A_3) = P(A_2)P(A_3) = 1/4 \tag{1.38}$$

Thus A_1 and A_2 are independent, A_1 and A_3 are independent and A_2 and A_3 are independent. However

$$P(A_1 \cap A_2 \cap A_3) = P(\{1\}) = 0 \neq P(A_1)P(A_2)P(A_3) = 1/8 \tag{1.39}$$

Thus A_1, A_2, A_3 are not independent even though A_1 and A_2 are independent, A_1 and A_3 are independent and A_2 and A_3 are independent.

1.8 The Chain Rule of Probability

Let $\{A_1, A_2, \ldots, A_n\}$ be a collection of events. The chain rule of probability tells us a useful way to compute the joint probability of the entire collection

$$\begin{aligned} &P(A_1 \cap A_2 \cap \cdots \cap A_n) = \\ &\qquad P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}) \end{aligned} \tag{1.40}$$

Proof: Simply expand the conditional probabilities and note how the denominator of the term $P(A_k \mid A_1 \cap \cdots \cap A_{k-1})$ cancels the numerator of the previous conditional probability, i.e.,

$$P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1}) = \tag{1.41}$$

$$P(A_1)\frac{P(A_2 \cap A_1)}{P(A_1)}\frac{P(A_3 \cap A_2 \cap A_1)}{P(A_1 \cap A_2)} \cdots \frac{P(A_1 \cap \cdots \cap A_n)}{P(A_1 \cap \cdots \cap A_{n-1})} \tag{1.42}$$

$$= P(A_1 \cap \cdots \cap A_n) \tag{1.43}$$

Example: A car company has 3 factories. 10% of the cars are produced in factory 1, 50% in factory 2 and the rest in factory 3. One out of 20 cars produced by the first factory are defective. 99% of the defective cars produced by the first factory are returned back to the manufacturer. What is the probability that a car produced by this company is manufactured in the first factory, is defective and is not returned back to the manufacturer.
Let A_1 represent the set of cars produced by factory 1, A_2 the set of defective cars

and A_3 the set of cars not returned. We know

$$P(A_1) = 0.1 \tag{1.44}$$
$$P(A_2 \mid A_1) = 1/20 \tag{1.45}$$
$$P(A_3 \mid A_1 \cap A_2) = 1 - 99/100 \tag{1.46}$$

Thus, using the chain rule of probability

$$P(A \cap A_2 \cap A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) = \tag{1.47}$$
$$(0.1)(0.05)(0.01) = 0.00005 \tag{1.48}$$

1.9 The Law of Total Probability

Let $\{H_1, H_2, \ldots\}$ be a countable collection of sets which is a partition of Ω. In other words

$$H_i \cap H_j = \varnothing, \text{ for } i \neq j, \tag{1.49}$$
$$H_1 \cup H_2 \cup \cdots = \Omega. \tag{1.50}$$

In some cases it is convenient to compute the probability of an event D using the following formula,

$$P(D) = P(H_1 \cap D) + P(H_2 \cap D) + \cdots \tag{1.51}$$

This formula is commonly known as the law of total probability (LTP)

Proof: First convince yourself that $\{H_1 \cap D, H_2 \cap D, \ldots\}$ is a partition of D, i.e.,

$$(H_i \cap D) \cap (H_j \cap D) = \varnothing, \text{ for } i \neq j, \tag{1.52}$$
$$(H_1 \cap D) \cup (H_2 \cap D) \cup \cdots = D. \tag{1.53}$$

Thus

$$P(D) = P((H_1 \cap D) \cup (H_2 \cap D) \cup \cdots) = \tag{1.54}$$
$$P(H_1 \cap D) + P(H_2 \cap D) + \cdots \tag{1.55}$$

We can do the last step because the partition is countable.

□

Example: A disease called pluremia affects 1 percent of the population. There is a test to detect pluremia but it is not perfect. For people with pluremia, the test is positive 90% of the time. For people without pluremia the test is positive 20% of the time. Suppose a randomly selected person takes the test and it is positive. What are the chances that a randomly selected person tests positive?:

Let D represent a positive test result, H_1 not having pluremia, H_2 having pluremia. We know $P(H_1) = 0.99$, $P(H_2) = 0.01$. The test specifications tell us: $P(D \mid H_1) = 0.2$ and $P(D \mid H_2) = 0.9$. Applying the LTP

$$P(D) = P(D \cap H_1) + P(D \cap H_2) \tag{1.56}$$
$$= P(H_1)P(D \mid H_1) + P(H_2)P(D \mid H_2) \tag{1.57}$$
$$= (0.99)(0.2) + (0.01)(0.9) = 0.207 \tag{1.58}$$

1.10 Bayes' Theorem

This theorem, which is attributed to Bayes (1744-1809), tells us how to revise probability of events in light of new data. It is important to point out that this theorem is consistent with probability theory and it is accepted by frequentists and Bayesian probabilists. There is disagreement however regarding whether the theorem should be applied to subjective notions of probabilities (the Bayesian approach) or whether it should only be applied to frequentist notions (the frequentist approach).

Let $D \in \mathcal{F}$ be an event with non-zero probability, which we will name . Let $\{H_1, H_2, \ldots\}$ be a countable collection of disjoint events, i.e,

$$H_1 \cup H_2 \cup \cdots = \Omega \tag{1.59}$$
$$H_i \cap H_j = \varnothing \text{ if } i \neq j \tag{1.60}$$

We will refer to $H_1, H_2, \ldots$ as "hypotheses", and D as "data". Bayes' theorem says that

$$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{P(D \mid H_1)P(H_1) + P(D \mid H_2)P(H_2) + \cdots} \tag{1.61}$$

where

- $P(H_i)$ is known as the prior probability of the hypothesis H_i. It evaluates the chances of a hypothesis prior to the collection of data.

- $P(H_i \mid D)$ is known as the posterior probability of the hypothesis H_i given the data.
- $P(D \mid H_1), P(D \mid H_2), \ldots$ are known as the likelihoods.

Proof: Using the definition of conditional probability

$$P(H_i \mid D) = \frac{P(H_i \cap D)}{P(D)} \tag{1.62}$$

Moreover, by the law of total probability

$$P(D) = P(D \cap H_1) + P(D \cap H_2) + \cdots = \tag{1.63}$$

$$P(D \mid H_1)P(H_1) + P(D \mid H_2)P(H_2) + \cdots \tag{1.64}$$

□

Example: A disease called pluremia affects 1 percent of the population. There is a test to detect pluremia but it is not perfect. For people with pluremia, the test is positive 90% of the time. For people without pluremia the test is positive 20% of the time. Suppose a randomly selected person takes the test and it is positive. What are the chances that this person has pluremia?:

Let D represent a positive test result, H_1 not having pluremia, H_2 having pluremia. Prior to the the probabilities of H_1 and H_2 are as follows: $P(H_2) = 0.01$, $P(H_1) = 0.99$. The test specifications give us the following likelihoods: $P(D \mid H_2) = 0.9$ and $P(D \mid H_1) = 0.2$. Applying Bayes' theorem

$$P(H_2 \mid D) = \frac{(0.9)(0.01)}{(0.9)(0.01) + (0.2)(0.99)} = 0.043 \tag{1.65}$$

Knowing that the test is positive increases the chances of having pluremia from 1 in a hundred to 4.3 in a hundred.

1.11 Exercises

1. Briefly describe in your own words the difference between the frequentist, Bayesian and mathematical notions of probability.

2. Go to the web and find more about the history of 2 probability theorists mentioned in this chapter.

3. Using diagrams, convince yourself of the rationality of De Morgan's law:

$$(A \cup B)^c = A^c \cap B^c \tag{1.66}$$

4. Try to prove analytically De Morgan's law.

5. Urn A has 3 black balls and 6 white balls. Urn B has 400 black balls and 400 white balls. Urn C has 6 black balls and 3 white balls. A person first randomly chooses one of the urns and then grabs a ball randomly from the chosen urn. What is the probability that the ball be black? If a person grabbed a black ball. What is the probability that the ball came from urn B?

6. The probability of catching Lyme disease after on day of hiking in the Cuyamaca mountains are estimated at less than 1 in 10000. You feel bad after a day of hike in the Cuyamacas and decide to take a Lyme disease test. The test is positive. The test specifications say that in an experiment with 1000 patients with Lyme disease, 990 tested positive. Moreover. When the same test was performed with 1000 patients without Lyme disease, 200 tested positive. What are the chances that you got Lyme disease.

7. This problem uses Bayes' theorem to combine probabilities as subjective beliefs with probabilities as relative frequencies. A friend of yours believes she has a 50% chance of being pregnant. She decides to take a pregnancy test and the test is positive. You read in the test instructions that out of 100 non-pregnant women, 20% give false positives. Moreover, out of 100 pregnant women 10% give false negatives. Help your friend upgrade her beliefs.

8. In a communication channel a zero or a one is transmitted. The probability that a zero is transmitted is 0.1. Due to noise in the channel, a zero can be received as one with probability 0.01, and a one can be received as a zero with probability 0.05. If you receive a zero, what is the probability that a zero was transmitted? If you receive a one what is the probability that a one was transmitted?

9. Consider a probability space $(\Omega, \mathcal{F}, P)$. Let A and B be sets of $\mathcal{F}$, i.e., both A and B are sets whose elements belong to Ω. Define the set operator $''-''$

as follows

$$A - B = A \cap B^c \tag{1.67}$$

Show that $P(A - B) = P(A) - P(A \cap B)$

10. Consider a probability space whose sample space Ω is the natural numbers (i.e., 1,2,3, ...). Show that not all the natural numbers can have equal probability.

11. Prove that any event is independent of the universal event Ω and of the null event $\varnothing$.

12. Suppose ω is an elementary outcome, i.e., $\omega \in \Omega$. What is the difference between ω and $\{\omega\}$?. How many elements does $\varnothing$ have?. How many elements does $\{\varnothing\}$ have?

13. You are a contestant on a television game show. Before you are three closed doors. One of them hides a car, which you want to win; the other two hide goats (which you do not want to win).

    First you pick a door. The door you pick does not get opened immediately. Instead, the host opens one of the other doors to reveal a goat. He will then give you a chance to change your mind: you can switch and pick the other closed door instead, or stay with your original choice. To make things more concrete without losing generality concentrate on the following situation

    (a) You have chosen the first door.

    (b) The host opens the third door, showing a goat.

    If you dont switch doors, what is the probability of wining the car? If you switch doors, what is the probability of wining the car? Should you switch doors?

14. Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which is more probable?

    (a) Linda is a bank teller?

    (b) Linda is a bank teller who is active in the feminist movement?

Chapter 2

Random variables

Up to now we have studied probabilities of sets of outcomes. In practice, in many experiment we care about some numerical property of these outcomes. For example, if we sample a person from a particular population, we may want to measure her age, height, the time it takes her to solve a problem, etc. Here is where the concept of a random variable comes at hand. Intuitively, we can think of a random variable (rav) as a numerical measurement of outcomes. More precisely, a random variable is a rule (i.e., a function) that associates numbers to outcomes. In order to define the concept of random variable, we first need to see a few things about functions.

Functions: Intuitively a function is a rule that associates members of two sets. The first set is called the **domain** and the second set is called the **target** or **codomain**. This rule has to be such that an element of the domain should not be associated to more than one element of the codomain. Functions are described using the following notation

$$f : A \to B \tag{2.1}$$

where f is the symbol identifying the function, A is the domain and B is the target. For example, $h : \mathbb{R} \to \mathbb{R}$ tells us that h is a function whose inputs are real numbers and whose outputs are also real numbers. The function $h(x) = (2)(x)+4$ would satisfy that description. Random variables are **functions** whose domain is the outcome space and whose codomain is the real numbers. In practice we can think of them as numerical measurements of outcomes. The input to a random variable is an elementary outcome and the output is a number.

Example: Consider the experiment of tossing a fair coin twice. In this case the outcome space is as follows:

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}\,. \tag{2.2}$$

One possible way to assign numbers to these outcomes is to count the number of heads in the outcome. I will name such a function with the symbol X, thus $X : \Omega \to \mathbb{R}$ and

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = (T,T) \\ 1 & \text{if } \omega = (T,H) \text{ or } \omega = (H,T) \\ 2 & \text{if } \omega = (H,H) \end{cases} \tag{2.3}$$

In many cases it is useful to define sets of Ω using the outcomes of the random variable X. For example the set $\{\omega : X(\omega) \leq 1\}$ is the set of outcomes for which X associates a number smaller or equal to 1. In other words

$$\{\omega : X(\omega) \leq 1\} = \{(T,T),(T,H),(H,T)\} \tag{2.4}$$

Another possible random variable for this experiment may measure whether the first element of an outcome is "heads". I will denote this random variable with the letter Y_1. Thus $Y_1 : \Omega \to \mathbb{R}$ and

$$Y_1(\omega) = \begin{cases} 0 & \text{if } \omega = (T,T) \text{ or } \omega = (T,H) \\ 1 & \text{if } \omega = (H,H) \text{ or } \omega = (H,T) \end{cases} \tag{2.5}$$

Yet another random variable, which I will name Y_2 may tell us whether the second element of an outcome is heads.

$$Y_2(\omega) = \begin{cases} 0 & \text{if } \omega = (T,T) \text{ or } \omega = (H,T) \\ 1 & \text{if } \omega = (H,H) \text{ or } \omega = (T,H) \end{cases} \tag{2.6}$$

We can also describe relationships between random variables. For example, for all outcomes ω in Ω it is true that

$$X(\omega) = Y_1(\omega) + Y_2(\omega) \tag{2.7}$$

This relationship is represented succinctly as

$$X = Y_1 + Y_2 \tag{2.8}$$

Example: Consider an experiment in which we select a sample of 100 students from UCSD using simple random sampling (i.e., all the students have equal chance of being selected and the selection of each students does not constrain the selection of the rest of the students). In this case the sample space is the set of all possible samples of 100 students. In other words, each outcome is a sample that contains 100 students.[1] A possible random variable for this experiment is the height of the first student in an outcome (remember each outcome is a sample with 100 students). We will refer to this random variable with the symbol H_1. Note given an outcome of the experiment, (i.e., a sample of 100 students) H_1 would assign a number to that outcome. Another random variable for this experiment is the height of the second student in an outcome. I will call this random variable H_2. More generally we may define the random variables $H_1, \ldots, H_{100}$ where $H_i : \Omega \to \mathbb{R}$ such that $H_i(\omega)$ is the height of the subject number i in the sample ω. The average height of that sample would also be a random variable, which could be symbolized as $\bar{H}$ and defined as follows

$$\bar{H}(\omega) = \frac{1}{100}(H_1(\omega) + \cdots + H_{100}(\omega)) \text{ for all } \omega \in \Omega \tag{2.9}$$

or more succinctly

$$\bar{H} = \frac{1}{100}(H_1 + \cdots + H_{100}) \tag{2.10}$$

I want you to remember that all these **random variables are not numbers**, they are functions (rules) that assign numbers to outcomes. The output of these functions may change with the outcome, thus the name random variable.

Definition A random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is a function $X : \Omega \to \mathbb{R}$. The domain of the function is the outcome space and the target is the real numbers.[2]

Notation: By convention random variables are represented with capital letters. For example $X : \Omega \to \mathbb{R}$, tells us that X is a random variable. Specific values of a random variable are represented with small letters. For example, $X(\omega) = u$ tells

[1] The number of possible outcomes is $n!/(100!\,(n-100)!)$ where n is the number of students at UCSD.

[2] Strictly speaking the function has to be Borel measurable (see Appendix), in practice the functions of interest to empirical scientists are Borel measurable so for the purposes of this book we will not worry about this condition.

us that the "measurement" assigned to the outcome ω by the random variable X is u. Also I will represents sets like

$$\{\omega : X(\omega) = u\} \tag{2.11}$$

with the simplified notation

$$\{X = u\} \tag{2.12}$$

I will also denote probabilities of such sets in a simplified, yet misleading, way. For example, the simplified notation

$$P(X = u) \tag{2.13}$$

or

$$P(\{X = u\}) \tag{2.14}$$

will stand for

$$P(\{\omega : X(\omega) = u\}) \tag{2.15}$$

Note the simplified notation is a bit misleading since for example X cannot possibly equal u since the first is a function and the second is a number.

Definitions:

- A random variable X is **discrete** if there is a countable set or real numbers $\{x_1, x_2, \ldots\}$ such that $P(X \in \{x_1, x_2, \ldots\}) = 1$.
- A random variable X is **continuous** if for all real numbers u the probability that X takes that value is zero. More formally, for all $u \in \mathbb{R}$, $P(X = u) = 0$.
- A random variable X is **mixed** if it is not continuous and it is not discrete.

2.1 Probability mass functions.

A probability mass function is a function $p_X : \mathbb{R} \to [0, 1]$ such that

$$p_X(u) = P(X = u) \text{ for all } u \in \mathbb{R}. \tag{2.16}$$

Note: If the random variable is continuous, then $p_X(u) = 0$ for all values of u. Thus,

$$\sum_{u\in\mathbb{R}} p_X(u) = \begin{cases} 0 & \text{if } X \text{ is continuous} \\ 1 & \text{if } X \text{ is discrete} \\ \text{neither } 0 \text{ nor } 1 & \text{if } X \text{ is mixed} \end{cases} \tag{2.17}$$

where the sum is done over the entire set of real numbers. What follows are some important examples of discrete random variables and their probability mass functions.

Discrete Uniform Random Variable: A random variable X is discrete uniform is there is a finite set of real numbers $\{x_1, \ldots, x_n\}$ such that

$$p_X(u) = \begin{cases} 1/n & \text{if } u \in \{x_1, \ldots, x_n\} \\ 0 & \text{else} \end{cases} \tag{2.18}$$

For example a uniform random variable that assigns probability 1/6 to the numbers $\{1, 2, 3, 4, 5, 6\}$ and zero to all the other numbers could be used to model the behavior of fair dies.

Bernoulli Random Variable: Perhaps the simplest random variable is the so called **Bernoulli** random variable, with parameter $\mu \in [0, 1]$. The Bernoulli random variable has the following probability mass function

$$p_X(y) = \begin{cases} \mu & \text{if } y = 1 \\ 1 - \mu & \text{if } y = 0 \\ 0 & \text{if } y \neq 1 \text{ and } y \neq 0 \end{cases} \tag{2.19}$$

For example, a Bernoulli random variable with parameter $\mu = 0.5$ could be used to model the behavior of a random die. Note such variable would also be discrete uniform.

Binomial Random Variable: A random variable X is binomial with parameters $\mu \in [0, 1]$ and $n \in \mathbb{N}$ if its probability mass function is as follows

$$p_X(y) = \begin{cases} \binom{n}{y}\mu^y(1-\mu)^{n-y} & \text{if } u \in \{0, 1, 2, 3, \ldots\} \\ 0 & \text{else} \end{cases} \tag{2.20}$$

Binomial probability mass functions are used to model the probability of obtaining y "heads" out of tossing a coin n times. The parameter μ represents the probability of getting heads in a single toss. For example if we want to get the probability of getting 9 heads out of 10 tosses of a fair coin, we set $n = 10$, $\mu = 0.5$ (since the coin is fair).

$$p_X(9) = \binom{10}{9}(0.5)^9(0.5)^{10-9} = \frac{10!}{9!(10-9)!}(0.5)^{10} = 0.00976 \tag{2.21}$$

Poisson Random Variable A random variable X is Poisson with parameter $\lambda > 0$ if its probability mass function is as follows

$$p_X(u) = \begin{cases} \frac{\lambda^u e^{-\lambda}}{u!} & \text{if } u \geq 0 \\ 0 & \text{else} \end{cases} \tag{2.22}$$

Poisson random variables model the behavior of random phenomena that occur with uniform likelihood in space or in time. For example, suppose on average a neuron spikes 6 times per 100 millisecond. If the neuron is Poisson then the probability of observing 0 spikes in a 100 millisecond interval is as follows

$$p_X(0) = \frac{6^0 e^{-6}}{0!} = 0.00247875217 \tag{2.23}$$

2.2 Probability density functions.

The probability density of a random variable X is a function $f_X : \mathbb{R} \to \mathbb{R}$ such that for all real numbers $a > b$ the probability that X takes a value between a and b equals the area of the function under the interval $[a, b]$. In other words

$$P(X \in [a, b]) = \int_a^b f_X(u)du \tag{2.24}$$

Note if a random variable has a probability density function (pdf) then

$$P(X = u) = \int_u^u f_X(x)dx = 0 \text{ for all values of } u \tag{2.25}$$

and thus the random variable is continuous.

Interpreting probability densities: If we take an interval very small the area under the interval can be approximated as a rectangle, thus, for small Δx

$$P(X \in (x, x + \Delta x]) \approx f_X(x)\Delta x \tag{2.26}$$

$$f_X(u) \approx \frac{P(X \in (x, x + \Delta x])}{\Delta x} \tag{2.27}$$

Thus the probability density at a point can be seen as the amount of probability per unit length of a small interval about that point. It is a ratio between two different ways of measuring a small interval: The probability measure of the interval and the length (also called Lebesgue measure) of the interval. What follows are examples of important continuous random variables and their probability density functions.

Continuous Uniform Variables: A random variable X is continuous uniform in the interval $[a, b]$, where a and b are real numbers such that $b > a$, if its pdf is as follows;

$$f_X(u) = \begin{cases} 1/(b-a) & \text{if } u \in [a, b] \\ 0 & \text{else} \end{cases} \tag{2.28}$$

Note how a probability density function can take values larger than 1. For example, a uniform random variable in the interval $[0, 0.1]$ takes value 10 inside that interval and 0 everywhere else.

Continuous Exponential Variables: A random variable X is called exponential if it has the following pdf

$$f_X(u) = \begin{cases} 0 \text{ if } u < 0 \\ \lambda \exp(-\lambda x) \quad \text{if } u \geq 0 \end{cases} \tag{2.29}$$

we can calculate the probability of the interval $[1, 2]$ by integration

$$P(\{X \in [1, 2]\}) = \int_1^2 \lambda \exp(-\lambda x) dx = \Big[-\exp(-\lambda x) \Big]_1^2 \tag{2.30}$$

if $\lambda = 1$ this probability equals $\exp(-1) - \exp(-2) = 0.2325$.

Gaussian Random Variables: A random variable X is Gaussian, also known as **normal**, with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ if its pdf is as follows

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2) \tag{2.31}$$

where $\pi = 3.1415\cdots$, μ is a parameter that controls the location of the center of the function and σ is a parameter than controls the spread of the function.. Hereafter whenever we want to say that a random variable X is normal with parameters μ and σ^2 we shall write it as $X \sim N(\mu, \sigma^2)$. If a Gaussian random variable X has zero mean and standard deviation equal to one, we say that it is a **standard Gaussian random variable**, and represent it $X \sim N(0, 1)$.

The Gaussian pdf is very important because of its ubiquitousness in nature thus the name "Normal". The underlying reason why this distribution is so widespread in nature is explained by an important theorem known as **the central limit theorem**. We will not prove this theorem here but it basically says that observations which are the result of a sum of a large number of random and independent influences have a cumulative distribution function closely approximated by that of a Gaussian random variable. Note this theorem applies to many natural observations: Height, weight, voltage fluctuations, IQ... All these variables are the result of a multitude of effects which when added up make the observations distribute approximately Gaussian.

One important property of Gaussian random variables is that linear combinations of Gaussian random variables produce Gaussian random variables. For example, if X_1 and X_2 are random variables, then $Y = 2+4X_1+6X_2$ would also be a Gaussian random variable.

2.3 Cumulative Distribution Function.

The cumulative distribution function, of a random variable X is a function $F_X : \mathbb{R} \to [0, 1]$ such that

$$F_X(u) = P(\{X \leq u\}) \tag{2.32}$$

For example, for the random variable described in (2.3)

$$F_X(u) = \begin{cases} 0.0 & \text{if } u < 0.0 \\ 1/4 & \text{if } u \geq 0.0 \text{ and } u < 1 \\ 3/4 & \text{if } u \geq 1 \text{ and } u < 2 \\ 1.0 & \text{if } u \geq 2 \end{cases} \tag{2.33}$$

The relationship between the cumulative distribution function, the probability mass function and the probability density function is as follows:

$$F_X(u) = \begin{cases} \sum_{x \leq u} p_X(u) & \text{if } X \text{ is a discrete random variable} \\ \int_{-\infty}^{u} f_X(x)dx & \text{if } X \text{ is a continuous random variable} \end{cases} \tag{2.34}$$

Example: To calculate the cumulative distribution function of a continuous exponential random variable X with parameter $\lambda > 0$ we integrate the exponential pdf

$$F_X(u) = \begin{cases} \int_0^u \lambda \exp(-\lambda x)dx & \text{if } u \geq 0 \\ 0 & \text{else} \end{cases} \tag{2.35}$$

And solving the integral

$$\int_0^u \lambda \exp(-\lambda x)dx = \Big[-\exp(-\lambda x) \Big]_0^u = 1 - \exp(-\lambda u) \tag{2.36}$$

Thus the cumulative distribution of an exponential random variable is as follows

$$F_X(u) = \begin{cases} 1 - \exp(-\lambda u) & \text{if } u \geq 0 \\ 0 & \text{else} \end{cases} \tag{2.37}$$

Observation: We have seen that if a random variable has a probability density function then the cumulative density function can be obtained by integration. Conversely we can differentiate the cumulative distribution function to obtain the probability density function

$$f_X(u) = \frac{dF_X(u)}{du} \tag{2.38}$$

A Property of Cumulative distribution Functions: Here is a property of cumulative distribution which has important applications. Consider a random variable X with cumulative distribution F_X now suppose we define a new random variable Y such that for each outcome ω

$$Y(\omega) = a + bX(\omega) \tag{2.39}$$

where a and $b \neq 0$ are real numbers. More succinctly we say

$$Y = a + bX \tag{2.40}$$

If we know the cumulative distribution of Y we can easily derive the cumulative distribution of Y.

$$F_Y(u) = P(Y \leq u) = P(a + bX \leq u) = P(X \leq \frac{u-a}{b}\}) = F_X(\frac{u-a}{b}) \tag{2.41}$$

Example: Let X be an exponential random variable with parameter λ, i.e.,

$$F_X(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 - \exp(-\lambda u) & \text{if } u \geq 0 \end{cases} \tag{2.42}$$

Let $Y = 1 + 2X$. In this case $a = 1$ and $b = 2$. Thus the cumulative distribution of Y is

$$F_Y(u) = F_X((u-1)/2) = \begin{cases} 0 & \text{if } (u-1)/2 < 0 \\ 1 - \exp(-\lambda(u-1)/2) & \text{if } (u-1)/2 \geq 0 \end{cases} \tag{2.43}$$

2.4 Exercises

1. Distinguish the following standard symbols P, p_X, F_X, f_X, X, x
2. Find the cumulative distribution function of a Bernoulli random variable X with parameter μ.
3. Find the cumulative distribution of a uniform random variable with parameters a, b.

4. Consider a uniform random variable in the interval $[0, 1]$.
   (a) Calculate the probability of the interval $[0.1, 0.2]$
   (b) Calculate the probability of 0.5.

5. Let X be a random variable with probability mass function
$$p_X(u) = \begin{cases} 1/5 & \text{if } u \in \{1, 2, 3, 4, 5\} \\ 0 & else \end{cases} \tag{2.44}$$
   (a) Find $F_X(4)$
   (b) Find $P(\{X \leq 3\} \cap \{X \leq 4\})$
   (c) Plot the function $h_X(t) = P(\{X = t\} \mid \{X \geq t\})$ for $t = 1, 2, 3, 4, 5$. This function is commonly known as the "hazard function" of X. If you think of X as the life time of a system, $h_X(t)$ tells us the probability that the system fails at time t given that it has not failed up to time t. For example, the hazard function of human beings looks like a U curve with a extra bump at the teen-years and with a minimum at about 30 years.

6. Let X be a continuous random variable with pdf
$$f_X(u) = \begin{cases} 0.25 & \text{if } u \in [-3, -1] \\ 0.25 & \text{if } u \in [1, 3] \\ 0 & \text{else} \end{cases} \tag{2.45}$$
   (a) Plot f_X. Can it be a probability density function? Justify your response.
   (b) Plot F_X

7. Show that if X is a continuous random variable with pdf f_X and $Y = a+bX$ where and a, b are real numbers. Then
$$f_Y(v) = (1/b) f_X(\frac{u-a}{b}) \tag{2.46}$$
   hint: Work with cumulative distributions and then differentiate them to get densities.

8. Show that if $X \sim N(\mu, \sigma^2)$ then $Y = a + bX$ is Gaussian.

Chapter 3

Random Vectors

In many occasions we need to model the joint behavior of more than one variable. For example, we may want to describe whether two different stocks tend to fluctuate in a somewhat linked manner or whether high levels of smoking covary with high lung cancer rates. In this chapter we examine the joint behavior of more than one random variables. To begin with, we will start working with pairs of random variables.

3.1 Joint probability mass functions

The joint probability mass function of the random variables X and Y is a function $p_{X,Y} : \mathbb{R}^2 \to [0, 1]$, such that for all $(u, v) \in \mathbb{R}^2$

$$p_{X,Y}(u, v) = P(\{X = u\} \cap \{Y = v\}) \tag{3.1}$$

hereafter, we use the simplified notation $P(X = u, Y = u)$ to represent $P(\{X = u\} \cap \{Y = v\})$, i.e., the probability measure of the set

$$\{\omega : (X(\omega) = u) \text{and} (Y(\omega) = v)\} \tag{3.2}$$

3.2 Joint probability density functions

The joint probability density function of the continuous random variables X and Y is a function $f_{X,Y} : \mathbb{R}^2 \to [0, \infty)$ such that for all $(u, v) \in \mathbb{R}^2$ and for all

$(\Delta u, \Delta v) \in R^2$

$$P(X \in [u, u+\Delta u], Y \in [v, v+\Delta v]) = \int_u^{u+\Delta u} \int_v^{v+\Delta v} f_{X,Y}(u,v)\, dv\, du \quad (3.3)$$

Interpretation Note if Δu and Δv are so small that $f_{X,Y}(u,v)$ is approximately constant over the area of integration then

$$P(X \in [u, u+\Delta u], Y \in [v, v+\Delta v]) \approx f_{X,Y}(u,v)\Delta u\, \Delta v \quad (3.4)$$

In other words, the probability that (X, Y) take values in the rectangle $[u, u+\Delta u] \times [v, v+\Delta v]$ is approximately the area of the rectangle times the density at a point in the rectangle.

3.3 Joint Cumulative Distribution Functions

The joint cumulative distribution of two random variables X and Y is a function $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ such that

$$F_{X,Y}(u,v) = P(\{X \leq u\} \cap \{Y \leq v\}) \quad (3.5)$$

note if X and Y are discrete then

$$F_{X,Y}(u,v) = \sum_{x \leq u} \sum_{y \leq v} p_{X,Y}(u,v) \quad (3.6)$$

and if X and Y are continuous then

$$F_{X,Y}(u,v) = \int_{-\infty}^{u} \int_{-\infty}^{v} f_{X,Y}(u,v)\, dv\, du \quad (3.7)$$

3.4 Marginalizing

In many occasions we know the joint probability density or probability mass function of two random variables X and Y and we want to get the probability density/mass of each of the variables in isolation. Such a process is called marginalization and it works as follows:

$$p_X(u) = \sum_{v \in \mathbb{R}} p_{X,Y}(u,v) \quad (3.8)$$

and if the random variables are continuous

$$f_X(u) = \int_{-\infty}^{\infty} f_{X,Y}(u,v)\,dv \tag{3.9}$$

Proof: Consider the function $h : \mathbb{R} \to \mathbb{R}$

$$h(u) = \int_{-\infty}^{\infty} f_{X,Y}(u,v)\,dv \tag{3.10}$$

We want to show that h is the probability density function of X. Note for all $a, b \in \mathbb{R}$ such that $a \leq b$

$$P(X \in [a,b]) = P(X \in [a,b], Y \in \mathbb{R}) = \int_a^b f_{X,Y}(u,v)\,dv\,du = \int_a^b h(u)\,du \tag{3.11}$$

showing that h is indeed the probability density function of X. A similar argument can be made for the discrete case.

3.5 Independence

Intuitively, two random variables X and Y are independent if knowledge about one of the variables gives us no information whatsoever about the other variable. More precisely, the random variables X and Y are independent if and only if all events of the form $\{X \in [u, u+\Delta u]\}$ and all events of the form $\{Y \in [v, v+\Delta v]\}$ are independent. If the random variable is continuous, a necessary and sufficient condition for independence is that the joint probability density be the product of the densities of each variable

$$f_{X,Y}(u,v) = f_X(u) f_Y(v) \tag{3.12}$$

for all $u, v \in \mathbb{R}$. If the random variable is discrete, a necessary and sufficient condition for independence is that the joint probability mass function be the product of the probability mass for each of the variables

$$p_{X,Y}(u,v) = p_X(u) p_Y(v) \tag{3.13}$$

for all $u, v \in \mathbb{R}$.

3.6 Bayes' Rule for continuous data and discrete hypotheses

Perhaps the most common application of Bayes' rule occurs when the data are represented by a continuous random variable X, and the hypotheses by a discrete random variable H. In such case, Bayes' rule works as follows

$$p_{H|X}(i \mid u) = \frac{f_{X|H}(u|i)\, p_H(i)}{f_X(u)} \tag{3.14}$$

where $p_{H|X} : \mathbb{R}^2 \to [0,1]$ is known as the conditional probability mass function of H given X and it is defined as follows

$$p_{H|X}(i \mid u) = \lim_{\Delta u \to 0} P(H = i \mid X \in [u, u + \Delta u]) \tag{3.15}$$

$f_{X|H} : \mathbb{R}^2 \to \mathbb{R}_+$ is known as the conditional probability density function of X given H and it is defined as follows: For all $a, b \in \mathbb{R}$ such that $a \leq b$,

$$P(X \in [a,b] \mid H = i) = \int_a^b f_{X|H}(u|i)\, du \tag{3.16}$$

Proof: Applying Bayes' rule to the events $\{H = i\}$ and $\{X \in [u, u + \Delta u]\}$

$$P(H = i \mid X \in [u, u + \Delta u]) = \frac{P(X \in [u, u + \Delta u] \mid H = i)\, P(H = i)}{P(X \in [u, u + \Delta u])} \tag{3.17}$$

$$= \frac{\int_u^{u+\Delta u} f_{X|H}(x|i)\, dx\, P(H = i)}{\int_u^{u+\Delta u} f_X(x)\, dx} \tag{3.18}$$

Taking limits and approximating areas with rectangles

$$\lim_{\Delta u \to 0} P(H = i \mid X \in [u, u + \Delta u]) = \frac{\Delta u f_{X|H}(u|i) p_H(i)}{\Delta u f_X(u)} \tag{3.19}$$

The Δu terms cancel out, completing the proof.

3.6.1 A useful version of the LTP

In many cases we are not given f_X directly and we need to compute it using $f_{X|H}$ and p_H. This can be done using the following version of the law of total probability:

$$f_X(u) = \sum_{j \in \text{Range}(H)} p_H(j)\, f_{X|H}(u|j) \tag{3.20}$$

Proof: Note that for all $a, b \in \mathbb{R}$, with $a \leq b$

$$\int_a^b \sum_{j \in \text{Range}(H)} p_H(j) f_{X|H}(u|j)\, du = \sum_{j \in \text{Range}(H)} p_H(j) \int_a^b f_{X|H}(u|j)\, du \tag{3.21}$$

$$= \sum_{j \in \text{Range}(H)} p_H(j) P(X \in [a,b] \mid H = j) = P(X \in [a,b]) \tag{3.22}$$

and thus $\sum_{j \in \text{Range}(H)} p_H(j) f_{X|H}(u|j)$ is the probability density of X.

Example: Let H be a Bernoulli random variable with parameter $\mu = 0.5$. Let $Y_0 \sim N(0,1)$, $Y_1 \sim N(1,1)$ and

$$X = (1 - H)(Y_0) + (H)(Y_1) \tag{3.23}$$

Thus,

$$p_H(0) = p_H(1) = 0.5 \tag{3.24}$$

$$f_{X|H}(u \mid i) = \frac{1}{\sqrt{2\pi}} e^{-(u-i)^2/2} \quad \text{for } i \in \{1, 2\}. \tag{3.25}$$

Suppose we are given the value $u \in \mathbb{R}$ and we want to know the posterior probability that H takes the value 0 given that X took the value u. To do so, first we apply LTP to compute $f_X(u)$

$$f_X(u) = (0.5)\frac{1}{\sqrt{2\pi}}\left(e^{-u^2/2} + e^{-(u-1)^2/2}\right) = (0.5)\frac{1}{\sqrt{2\pi}} e^{u^2/2}\left(1 + e^{2u-1}\right) \tag{3.26}$$

Applying Bayes' rule,

$$p_{H|X}(0 \mid u) = \frac{1}{1 + e^{-(1-2u)}} \tag{3.27}$$

So, for example, if $u = 0.5$ the probability that H takes the value 0 is 0.5. If $u = 1$ the probability that H takes the value 0 drops down to 0.269.

3.7 Random Vectors and Stochastic Processes

A random vector is a collection of random variables organized as a vector. For example if $X_1, \ldots, X_n$ are random variables then

$$X = (X_1, Y, \ldots, X_n)' \tag{3.28}$$

is a random vector. A stochastic process is an ordered set of random vectors. For example, if I is an indexing set and X_t, is a random vector for all $t \in I$ then

$$X = (X_i : i \in I) \tag{3.29}$$

is a random process. When the indexing set is countable, Y is called a "discrete time" stochastic process. For example,

$$X = (X_1, Y, X_3, \ldots) \tag{3.30}$$

is a discrete time random process. If the indexing set are the real numbers, then Y is called a "continuous time" stochastic process. For example, if X is a random variable, the ordered set of random vectors

$$X = (X_t : t \in \mathbb{R}) \tag{3.31}$$

with

$$X_t(\omega) = \sin(2\pi t X(\omega)) \tag{3.32}$$

is a continuous time stochastic process.

Chapter 4

Expected Values

The expected value (or mean) of a random variable is a generalization of the notion of arithmetic mean. It is a number that tells us about the "center of gravity" or average value we would expect to obtain is we averaged a very large number of observations from that random variable. The expected value of the random variable X is represented as $E(X)$ or with the Greek letter μ, as in μ_X and it is defined as follows

$$E(X) = \mu_X = \begin{cases} \sum_{u \in \mathbb{R}} p_X(u)u & \text{for discrete ravs} \\ \int_{-\infty}^{+\infty} p_X(u)u \, du & \text{for continuous ravs} \end{cases} \tag{4.1}$$

We also define the expected value of a number as the number itself, i.e., for all $u \in \mathbb{R}$

$$E(u) = u \tag{4.2}$$

Example: The expected value of a Bernoulli random variable with parameter μ is as follows

$$E(X) = (1 - \mu)(0.0) + (\mu)(1.0) = \mu \tag{4.3}$$

Example: A random variable X with the following pdf

$$p_X(u) = \begin{cases} \frac{1}{b-a} & \text{if } u \in [a, b] \\ 0 & \text{if } u \notin [a, b] \end{cases} \tag{4.4}$$

is known as a uniform $[0, 1]$ continuous random variable. Its expected value is as follows

$$E(X) = \int_{-\infty}^{\infty} p_X(u)u\ du = \int_a^b \frac{1}{b-a}u\ du$$
$$= \frac{1}{(b-a)}\Big[\frac{u^2}{2}\Big]_a^b =$$
$$= \frac{1}{(b-a)}(\frac{b^2-a^2}{2}) = (a+b)/2 \quad (4.5)$$

Example: The expected value of an exponential random variable with parameter $\lambda$is as follows (see previous chapter for definition of an exponential random variable)

$$E(X) = \int_{-\infty}^{\infty} up_X(u)du = \int_0^{\infty} u\lambda \exp(-\lambda u)du \quad (4.6)$$

and using integration by parts

$$E(X) = \Big[-u\exp(-\lambda u)\Big]_0^{\infty} + \int_0^{\infty} \lambda \exp(-\lambda u)du = -\Big[\lambda \exp(-\lambda u)\Big]_0^{\infty} = \frac{1}{\lambda} \quad (4.7)$$

4.1 Fundamental Theorem of Expected Values

This theorem is very useful when we have a random variable Y which is a function of another random variable X. In many cases we may know the probability mass function, or density function, for X but not for Y. The fundamental theorem of expected values allows us to get the expected value of Y even though we do not know the probability distribution of Y. First we'll see a simple version of the theorem and then we will see a more general version.

Simple Version of the Fundamental Theorem: Let $X : \Omega \to \mathbb{R}$ be a rav and $h : \mathbb{R} \to \mathbb{R}$ a function. Let Y be a rav defined as follows:

$$Y(\omega) = h(X(\omega)) \text{ for all } \omega \in \mathbb{R} \quad (4.8)$$

or more succinctly

$$Y = h(X) \quad (4.9)$$

Then it can be shown that

$$E(Y) = \begin{cases} \sum_{u\in\mathbb{R}} p_X(u)h(u) & \text{for discrete ravs} \\ \int_{-\infty}^{+\infty} p_X(u)h(u)\,du & \text{for continuous ravs} \end{cases} \tag{4.10}$$

Example: Let X be a Bernoulli rav and Y a random variable defined as $Y = (X-0.5)^2$. To find the expected value of Y we can apply the fundamental theorem of expected values with $h(u) = (u-0.5)^2$. Thus,

$$\begin{aligned} E(Y) = \sum_{u\in\mathbb{R}} p_X(u)(u-0.5)^2 &= p_X(0)(0-0.5)^2 + p_X(1)(1-0.5)^2 \\ &= (0.5)(0.5)^2 + (0.5)(-0.5)^2 = 0.25 \end{aligned} \tag{4.11}$$

Example: The average entropy, or information value (in bits) of a random variable X is represented as $H(X)$ and is defined as follows

$$H(X) = -E(\log_2 p_X(X)) \tag{4.12}$$

To find the entropy of a Bernoulli random variable X with parameter μ we can apply the fundamental theorem of expected values using the function $h(u) = \log_2 p_X(u)$. Thus,

$$\begin{aligned} H(X) = \sum_{u\in\mathbb{R}} p_X(u)\log_2 p_X(u) &= p_X(0)\log_2 p_X(0) + p_X(1)\log_2 p_X(1) \\ &= (\mu)\log_2(\mu) + (1-\mu)\log_2(1-\mu) \end{aligned} \tag{4.13}$$

For example, if $\mu = 0.5$, then

$$H(X) = (0.5)\log_2(0.5) + (0.5)\log_2(0.5) = 1 \text{ bit} \tag{4.14}$$

General Version of the Fundamental Theorem: Let $X_1, \ldots, X_n$ be random variables, let $Y = h(X_1, \ldots, X_n)$, where $h : \mathbb{R}^n \to \mathbb{R}$ is a function. Then

$$E(Y) = \begin{cases} \sum (u_1, \ldots, u_n) \in \mathbb{R}^n p_{X_1,\ldots,X_n}(u_1, \ldots, u_n)h(u_1, \ldots, u_n) & \text{for discrete ravs} \\ \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} p_{X_1,\ldots,X_n}(u_1, \ldots, u_n)h(u_1, \ldots, u_n)du_1 \cdots du_n & \text{for continuous ravs} \end{cases} \tag{4.15}$$

4.2 Properties of Expected Values

Let X and Y be two random variables and a and b two real numbers. Then

$$E(X+Y) = E(X) + E(Y) \tag{4.16}$$

and

$$E(a+bX) = a + bE(X) \tag{4.17}$$

Moreover if X and Y are independent then

$$E(XY) = E(X)E(Y) \tag{4.18}$$

We will prove these properties using discrete ravs. The proofs are analogous for continuous ravs but substituting sums by integrals.

Proof: By the fundamental theorem of expected values

$$\begin{aligned} E(X+Y) &= \sum_{u\in\mathbb{R}}\sum_{v\in\mathbb{R}} p_{X,Y}(u,v)(u+v) \\ &= (\sum_{u\in\mathbb{R}} u \sum_{v\in\mathbb{R}} p_{X,Y}(u,v)) + (\sum_{v\in\mathbb{R}} v \sum_{u\in\mathbb{R}} p_{X,Y}(u,v)) \\ &= \sum_{u\in R} u p_X(u) + \sum_{v\in\mathbb{R}} v p_Y(v) = E(X) + E(Y) \end{aligned} \tag{4.19}$$

where we used the law of total probability in the last step.

Proof: Using the fundamental theorem of expected values

$$\begin{aligned} E(a+bX) &= \sum_{u} p_X(u)(a+bu) \\ &= a\sum_{u\in\mathbb{R}} p_X(u) + b\sum_{u} u p_X(u) = a + bE(X) \end{aligned} \tag{4.20}$$

Proof: By the fundamental theorem of expected values

$$E(XY) = \sum_{u\in\mathbb{R}}\sum_{v\in\mathbb{R}} p_{X,Y}(u,v)uv \tag{4.21}$$

if X and Y are independent then $p_{X,Y}(u,v) = p_X(u)p_Y(v)$. Thus

$$E(XY) = [\sum_{u} u p_X(u)]\,[\sum_{v} v p_Y(v)] = E(X)E(Y) \tag{4.22}$$

□

4.3 Variance

The variance of a random variable X is a number that represents the amount of variability in that random variable. It is defined as the expected value of the squared deviations from the mean of the random variable and it is represented as $\text{Var}(X)$ or as σ_X^2

$$\text{Var}(X) = \sigma_X^2 = E[(X-\mu_X)^2] = \begin{cases} \sum_{u\in\mathbb{R}} p_X(u)(u-\mu_X)^2 & \text{for discrete ravs} \\ \int_{-\infty}^{+\infty} p_X(u)(u-\mu_X)^2 du & \text{for continuous ravs} \end{cases} \tag{4.23}$$

The **standard deviation** of a random variable X is represented as $\text{Sd}(X)$ or σ and it is the square root of the variance of that random variable

$$\text{Sd}(X) = \sigma_X = \sqrt{\text{Var}(X)} \tag{4.24}$$

The standard deviation is easier to interpret than the variance for it uses the same units of measurement taken by the random variable. For example if the random variable X represents reaction time in seconds, then the variance is measured in seconds squares, which is hard to interpret, while the standard deviation is measured in seconds. The standard deviation can be interpreted as a "typical" deviation from the mean. If an observation deviates from the mean by about one standard deviation, we say that that amount of deviation is "standard".

Example: The variance of a Bernoulli random variable with parameter μ is as follows

$$\text{Var}(X) = (1-\mu)(0.0-\mu)^2 + (\mu)(1.0-\mu)^2 = (\mu)(1-\mu) \tag{4.25}$$

Example: For a continuous uniform $[a, b]$ random variable X, the pdf is as follows

$$p_X(u) = \begin{cases} \frac{1}{b-a} & \text{if } u \in [a,b] \\ 0 & \text{if } u \notin [a,b] \end{cases} \tag{4.26}$$

which we have seen as expected value $\mu_X = (a+b)/2$. Its variance is as follows

$$\sigma_X^2 = \int_{-\infty}^{+\infty} u p_X(u) du = \frac{1}{b-a}\int_a^b (u-(a+b)/2)^2 du \tag{4.27}$$

and doing a change of variables $y = (u - (a+b)/2)$

$$\sigma_X^2 = \frac{1}{b-a}\int_{(a-b)/2}^{(b-a)/2} y^2 dy = \frac{1}{b-a}\Big[y^3/3\Big]_{(a-b)/2}^{(b-a)/2} = \frac{b-a}{12} \tag{4.28}$$

Exercise: We will show that in a Gaussian random variable the parameter μ is the mean and σ the standard deviation. First note

$$E(X) = \int_{-\infty}^{\infty} u \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}(\frac{u-\mu}{\sigma})^2) du \tag{4.29}$$

changing to the variable $y = (u - \mu)$

$$E(X) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} y \exp(-\frac{1}{2}(\frac{y}{\sigma})^2) dy + \mu[\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} \exp(-\frac{1}{2}(\frac{y}{\sigma})^2)] = \mu \tag{4.30}$$

The first term is zero because the integrand is an odd function (i.e., $g(-x) = -g(x)$. For the variance

$$\text{Var}(X) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} (u-\mu)^2 \exp(-\frac{1}{2}(\frac{u-\mu}{\sigma})^2) du \tag{4.31}$$

changing variables to $y = (u - \mu)$

$$\text{Var}(X) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} y\ y \exp(-\frac{1}{2}(\frac{y}{\sigma})^2) dy \tag{4.32}$$

and using integration by parts

$$\text{Var}(X) = \frac{1}{\sqrt{2\pi\sigma^2}}\Big(-\sigma^2\Big[y\exp(-\frac{1}{2}(y/\sigma)^2)\Big]_{-\infty}^{+\infty} + \sigma^2\int_{-\infty}^{\infty}\exp(-\frac{1}{2}(\frac{y}{\sigma})^2) dy\Big) = \sigma^2 \tag{4.33}$$

4.3.1 Properties of the Variance

Let X and Y be two random variables and a and b two real numbers then

$$\text{Var}(aX + b) = a^2\text{Var}(X) \tag{4.34}$$

and

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y) \tag{4.35}$$

where $\text{Cov}(X,Y)$ is known as the covariance between X and Y and is defined as

$$\text{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))] \tag{4.36}$$

Proof:

$$\begin{aligned}\text{Var}(aX+b) &= E[(aX+b-E(aX+b))^2] = E[(aX+b-aE(X)-b)^2] \\ &= E[(a(X-E(X))^2] = a^2E[(X-E(X))^2] = a^2\text{Var}(X)\end{aligned} \tag{4.37}$$

Proof:

$$\begin{aligned}\text{Var}(X+Y) &= E(X+Y-E(X+Y))^2 = E((X-\mu_X)+(Y-\mu_Y))^2 \\ &= E[(X-\mu_X)^2] + E[(Y-\mu_Y)^2] + 2E[(X-\mu_X)(Y-\mu_Y)]\end{aligned} \tag{4.38}$$

□

If $\text{Cov}(X,Y) = 0$ we say that the random variables X and Y are **uncorrelated**. In such case the variance of the sum of the two random variables equals the sum of their variances.

It is easy to show that if two random variables are independent they are also uncorrelated. To see why note that

$$\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_Y E(X) \tag{4.39}$$

$$- \mu_X E(Y) + \mu_X\mu_Y = E(XY) - E(X)E(Y) \tag{4.40}$$

we have already seen that if X and Y are independent then $E(XY) = E(X)E(Y)$ and thus X and Y are uncorrelated.

However two random variables may be uncorrelated and still be dependent. Think of correlation as a linear form of dependency. If two variables are uncorrelated it means that we cannot use one of the variables to linearly predict the other variable. However in uncorrelated variables there may still be non-linear relationships that make the two variables non-independent.

Example: Let X, Y be random variables with the following joint pmf

$$p_{X,Y}(u,v) = \begin{cases} 1/3 & \text{if } u = -1 \text{ and } v = 1 \\ 1/3 & \text{if } u = 0 \text{ and } v = 0 \\ 1/3 & \text{if } u = 1 \text{ and } v = 1 \\ 0 & \text{else} \end{cases} \tag{4.41}$$

show that X and Y are uncorrelated but are not independent.

Answer: Using the fundamental theorem of expected values we can compute $E(XY)$

$$E(XY) = (1/3)(-1)(1) + (1/3)(0)(0) + (1/3)(1)(1) = 0 \tag{4.42}$$

From the joint pmf of X and Y we can marginalize to obtain the pmf of X and of Y

$$p_X(u) = \begin{cases} 1/3 & \text{if } u = -1 \\ 1/3 & \text{if } u = 0 \\ 1/3 & \text{if } u = 1 \\ 0 & \text{else} \end{cases} \tag{4.43}$$

$$p_Y(v) = \begin{cases} 1/3 & \text{if } v = 0 \\ 2/3 & \text{if } v = 1 \\ 0 & \text{else} \end{cases} \tag{4.44}$$

Thus

$$E(X) = (1/3)(-1) + (1/3)(0) + (1/3)(1) = 0 \tag{4.45}$$
$$E(Y) = (1/3)(0) + (2/3)(1) = 2/3 \tag{4.46}$$
$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y) = 0 - (0)(2/3) = 0 \tag{4.47}$$

Thus X and Y are uncorrelated. To show that X and Y are not independent note that

$$p_{X,Y}(0,0) = 1/3 \neq p_X(0)p_Y(0) = (1/3)(1/3) \tag{4.48}$$

4.4 Appendix: Using Standard Gaussian Tables

Most statistics books have tables for the cumulative distribution of Gaussian random variables with mean $\mu = 0$ and standard deviation $\sigma = 1$. Such a cumulative distribution is known as the standard Gaussian cumulative distribution and it is represented with the capital Greek letter "Phi", i.e., Φ. So $\Phi(u)$ is the probability that a standard Gaussian random variable takes values smaller or equal to u. In many cases we need to know the probability distribution of a Gaussian random variable X with mean μ different from zero and variance σ^2 different from 1. To do so we can use the following trick

$$F_X(u) = \Phi(\frac{x-\mu}{\sigma}) \tag{4.49}$$

For example, the probability that a Gaussian random variable with mean $\mu = 2$ and standard deviation $\sigma = 4$ takes values smaller than 6 can be obtained as follows

$$P(X \leq 6) = F_X(6) = \Phi(\frac{6-2}{4}) = \Phi(1) \tag{4.50}$$

If we go to the standard Gaussian tables we see that $\Phi(1) = 0.8413$.

Proof: Let Z be a standard Gaussian random variable, and $X = \mu + \sigma Z$. Thus,

$$E(X) = \mu + E(Z) = \mu \tag{4.51}$$

$$\text{Var}(X) = \sigma^2 \text{Var}(Z) = \sigma^2 \tag{4.52}$$

Thus, X has the desired mean and variance. Moreover since X is a linear transformation of Z then X is also Gaussian. Now, using the the properties of cumulative distributions that we saw in a previous chapter

$$F_X(u) = F_Z((u-\mu)/\sigma) = \Phi((u-\mu)/\sigma)) \tag{4.53}$$

□

Example: Let Y be a Gaussian rav with $\mu_Y = 10$ and $\sigma_Y^2 = 100$. Suppose we want to compute $F_Y(-20)$. First we compute $z = (-20 - \mu_Y)/\sigma_Y = -1$. We go to the standard Gaussian tables and find $F_Z(-1) = 0.1587$ Thus $F_Y(20) = F_Z(-1) = 0.1587$.

4.5 Exercises

1. Let a random variable X represent the numeric outcome of rolling a fair die, i.e., $p_X(u) = 1/6$ for $u \in \{1, 2, 3, 4, 5, 6\}$. Find the expected value , standard deviation, and average information value of X.

2. Consider the experiment of tossing a fair coin twice. Let X_1 be a random variable taking value 1 if the first toss is heads, and 0 otherwise. Let X_2 be a random variable taking value 1 if the second toss is heads, and 0 else.

   (a) Find $E(X_1)$, $E(X_2)$

   (b) Find Var(X_1), Var(X_2)

   (c) Find Var$(X_1/2)$ and Var$(X_2/2)$

   (d) Find Var$(X_1 + X_2)$ and Var$[(X_1 + X_2)/2]$

3. Find the information value of a continuous random variable in the interval $[a, b]$.

4. Show that the variance of an exponential random variable with parameter λ is $1/\lambda^2$.

5. Let X be a standard Gaussian random variable. Using Gaussian tables find:

   (a) $F_X(-1.0)$

   (b) $F_X(1.0)$

   (c) $P(-1.0 \leq X \leq 1.0)$

   (d) $P(X = 1.0)$

Chapter 5

The precision of the arithmetic mean

Consider taking the average of a number of observations all of which are randomly sampled from a large population. Intuitively it seems that as the number of observations increases this average will get closer and closer to the true population mean. In this chapter we quantify how close the sample mean gets to the population mean as the number of observations in the sample increases. In other words we want to know how good the mean of a sample of noisy observations is as an estimate of the "true" value underlying those observations.

Let's formalize the problem using our knowledge of probability theory. We start with n *independent* random variables variables $X_1, \ldots, X_n$. We will also assume that all these random variables have the same mean, which we will represent as μ and the same variance, which we will represent as σ^2. In other words

$$E(X_1) = E(X_2) = \cdots = E(X_n) = \mu \tag{5.1}$$

and

$$\text{Var}(X_1) = \text{Var}(X_2) = \cdots = \text{Var}(X_n) = \sigma^2 \tag{5.2}$$

This situation may occur, when our experiment consists of randomly selecting n individuals from a population (e.g., 20 students from UCSD). In this case the outcomes of the random experiment consists of a sample of n subjects. The outcome space is the set of all possible samples of n subjects. The random variable X_i would represent a measurement of subject number i in the sample. Note that in this case all random variables have the same mean, which would be equal to the average observation for the entire population. They also have the same variance, since all subjects have an equal chance of being in any of the n possible positions of the samples.

Given a sample ω we can average the n values taken by the random variables for that particular sample. We will represent this average as $\bar{X}_n$,

$$\bar{X}_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} X_i(\omega); \quad \text{for all } \omega \in \Omega \tag{5.3}$$

More concisely,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{5.4}$$

Note that $\bar{X}_n$ is itself a random variable (i.e., a function from the outcome space to the real numbers). We call this random variable the sample mean.

One convenient way to measure the precision of the sample mean is to compute the expected squared deviation of the sample mean from the true population mean, i.e.,

$$\sqrt{E[(\bar{X}_n - \mu)^2]} \tag{5.5}$$

For example, if $\sqrt{E[(\bar{X}_n - \mu_Y)^2]} = 4.0secs$ this would tell us that on average the sample mean deviates from the population mean by about 4.0 secs.

5.1 The sampling distribution of the mean

We can now use the properties of expected values and variances to derive the expected value and variance of the sample mean .

$$E(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu \tag{5.6}$$

thus the expected value of the sample mean is the actual population mean of the random variables. A consequence of this is that

$$E[(\bar{X}_n - \mu)^2] = E[(\bar{X}_n - E(\bar{X}_n))^2] = \text{Var}(\bar{X}_n) \tag{5.7}$$

This takes us directly to our goal. Using the properties of the variance, and considering that the random variables $X_1, \ldots, X_n$ are independent we get

$$Var(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) = \frac{\sigma^2}{n} \tag{5.8}$$

The standard deviation of the mean is easier to interpret than the variance of the mean. It roughly represents how much the means of independent randomly obtained samples typically differ from the population mean. Since the standard deviation of the mean is so important, many statisticians give it a special name: the **standard error of the mean**. Thus the standard error of the mean simply is the square root of the variance of the mean, and it is represented as $\text{Sd}(\bar{X}_n)$ or $Se(\bar{X}_n)$. Thus,

$$\text{Sd}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} = \sqrt{E[(\bar{X}_n - \mu)^2]} \tag{5.9}$$

Finally we are done! Equation 5.9 tells us that the uncertainty of the sample mean (its standard deviation) increases proportionally to the uncertainty about individual observations (σ) and decreases proportionally to the square root of the number of observations. Thus, if we want to double the precision of the sample mean (reduce its standard deviation by half) we need to quadruple the number of observations in the sample.

Example: Let X_1 and X_2 be independent Bernoulli random variables with parameter $\mu = 0.5$. Thus $E(X_1) = E(X_2) = \mu = 0.5$ and $\text{Var}(X_1) = \text{Var}(X_2) = \sigma^2 = 0.25$, moreover

$$E(\bar{X}_2) = (0)(0.25) + (0.5)(0.5) + (1)(0.25) = 0.5 = \mu \tag{5.10}$$

$$\text{Var}(\bar{X}_2) = \sigma^2/2 = 0.125 \tag{5.11}$$

5.1.1 Central Limit Theorem

This is a very important theorem which we will not prove here. The theorem tells us that as n goes to infinity, the *cumulative distribution* of $\bar{X}_n$ is closely approximated by that of a Gaussian random variable with mean $E(\bar{X}_n)$ and standard deviation $\text{Sd}(\bar{X}_n)$. In practice for $n \geq 30$ the Gaussian cumulative distribution provides very good approximations.

Example: We toss a fair coin 100 times. What is the probability that the proportion of "heads" be smaller or equal to 0.45?

Answer: Tossing a fair coin 100 times can be modeled using 100 independent identically distributed Bernoulli random variables each of which has parameter

$\mu = 0.5$. Let's represent these 100 random variables as $X_1, \ldots, X_{100}$ where X_i takes the value 1 if the i^{th} time we toss the coin we get heads. Thus the proportion of heads is the average of the 100 random variables, we will represent it as $\bar{X}$.

$$\bar{X} = \frac{1}{100}(X_1 + \cdots + X_{100}) \tag{5.12}$$

We know

$$\mu = E(X_1) = \cdots = E(X_{100}) = 0.5 \tag{5.13}$$

$$\sigma^2 = \text{Var}(X_1) = \cdots = \text{Var}(X_{100}) = 0.25 \tag{5.14}$$

Thus

$$E(\bar{X}) = 0.5 \tag{5.15}$$

$$\text{Var}(\bar{X}) = \sigma^2/100 = 0.0025 \tag{5.16}$$

Since $n = 100 > 30$ the cumulative distribution of $\bar{X}$ is approximately Gaussian. Thus

$$P(\bar{X} \leq 0.45) = F_{\bar{X}}(0.45) \approx \Phi(\frac{0.45 - 0.5}{\sqrt{0.0025}}) = \Phi(-1) = 0.1586 \tag{5.17}$$

5.2 Exercises

1. **The Central Limit Theorem**

   (a) Goto the book Web site and click on LinuStats → Coin Simulator

   (b) Do 1000 replications of a coin tossing experiment. Each experiment should consists of 10 coin tosses (10 observations per experiment). The result should be 1000 numbers each of which represents the proportion of tails obtained in a particular experiment. If we encoded Heads as "0" and tails as "1", the outcome of a coin toss is an outcome from a Bernoulli 0,1 random variable with parameter $\mu = 0.5$. The proportion of tails obtained in 10 tosses is the average of 10 independent identically distributed Bernoulli random variables.

   $$\bar{X} = \frac{\sum_{i=1}^{10} X_i}{10} \tag{5.18}$$

   This average is itself a random variable. Moreover the central limit theorem tell us that it should be approximately Gaussian. We will check that now:

(c) Use the Edit menu in your browser to copy the 1000 numbers obtained in the coin tossing experiments.

(d) Go back to LinuStats and choose the descriptive statistics page.

(e) Use the Edit menu in your browser to paste the 1000 numbers into the data window.

(f) Choose a minimum value of 0, maximum value of 1, and window of 0.1

(g) Analyze the data by clicking the appropriate button.

(h) Copy the resulting relative frequency table and transform it into a cumulative distribution function

(i) Copy the Mean and SD (Standard Deviation) of the 1000 numbers. Explain what this standard deviation means. I'll call these numbers $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$

(j) Compare the obtained cumulative distribution function with that predicted by a Gaussian distribution with mean $\mu_{\bar{X}}$ and standard deviation $\sigma_{\bar{X}}$.

2. **The sampling distribution of the mean**

(a) Go to the coin simulator and do 1000 experiments each with 1 toss (i.e., one observation per experiment). Using the same procedure as in the previous exercise, calculate the mean and standard deviation of the 1000 experiments. Plot the relative frequency density polygon. Interpret your results.

(b) Same as the previous question but now toss the coin 2 times per experiment. Obtain the new mean and standard deviation and interpret what they mean.

(c) Same with 4, 16 and with 32 observations per experiment.

(d) Plot a graph with the obtained means as a function of the number of tosses.

(e) Plot a graph with the obtained standard deviations as a function of the number of tosses.

(f) Plot a graph with the obtained variances as a function of the number of tosses.

(g) Explain your results. How does the mean, variance, and standard deviation change as the number of observations per experiment increases. How does the relative frequency density polygon change as the number of observations per experiment increases? Does it make sense?

3. What is less probable: 1) to get 80 or more tails out of tossing a fair coin 100 times, 2) to get 30 or more tails out of tossing a fair coin 50 times. NOTE: Since $n > 30$ you can assume that the cumulative distribution of the mean is approximately Gaussian.

4. An experimenter inserts intracellular electrodes on a large sample of randomly selected neurons in primary visual cortex (V1). Intracellular electrodes are designed to measure the response of single neurons. The researcher finds the average response of the neurons is 10 mVolts and the standard deviation 1 mVolt. On a subsequent experiment the researcher inserts extracellular electrodes on randomly selected location in V1. Extracellular electrodes compute the average response of a number of neurons around the tip of the electrode. The researcher finds that when the electrode is extracellular the average response is still 10 mVolts but the standard deviation goes down to 0.01 mVolts. Assume the neurons in V1 are independent and have identical response distributions.

   (a) Explain how the standard deviation of the mean of n independent random variables changes as the number of random variables increases.

   (b) Estimate how many neurons are having an effect on the extracellular electrode. Justify your response.

Chapter 6

Introduction to Statistical Hypothesis Testing

The goal of statistical hypothesis testing is to provide a rational basis for making inferences and decisions about hypotheses based on uncertain data. For example based on the results of a experiment with a limited sample of individuals we may want to decide whether there is enough evidence to say that smoking causes cancer. There are two major approaches to statistical hypothesis testing: 1) The classic approach, 2) The Bayesian approach. The classic approach is the standard used when analysing scientific experiments. The Bayesian approach is dominant in machine perception, artificial intelligence and engineering applications.

6.1 The Classic Approach

Classic statisticians (also known as frequentists) view scientific hypotheses as propositions with a fixed truth value. Within this approach a hypothesis is either true or false and thus it makes no sense to talk about its probability. For example, for a frequentist the hypothesis "Smoking causes cancer" is either true or false. It makes no sense to say that there is a 99 % chance that smoking causes cancer. In this approach we perform experiments, gather data and if the data provide sufficient evidence against a hypothesis we reject the hypothesis.

Here is an example that illustrates how the classic approach works in practice. Suppose we want to disprove the hypothesis that a particular coin is fair. We toss the coin 100 times and get 74 % tails. We model our observations as a random variable $\bar{X}$ which is the average of 100 iid Bernoulli ravs with unknown parameter

μ

$$\bar{X} = \frac{1}{100}\sum_{i=1}^{100} X_i \tag{6.1}$$

The statement that we want to reject is called **the null hypothesis**. In this case the null hypothesis is that $\mu = 0.5$, i.e., that the coin is fair. We do not know $E(\bar{X})$ nor Var$(\bar{X})$. However we know what these two values would be if the null hypothesis were true. We represent the conditional expected value of $\bar{X}$ assuming that the null hypothesis is true as $E(\bar{X} \mid H_n \text{ true})$. The variance of $\bar{X}$ assuming that the null hypothesis is true is represented as Var$(\bar{X} \mid H_n \text{ true})$. In our example

$$E(\bar{X} \mid H_n \text{ true}) = 0.5 \tag{6.2}$$

$$\text{Var}(\bar{X} \mid H_n \text{ true}) = 0.25/100 \tag{6.3}$$

Moreover, due to the central limit theorem the cumulative distribution of $\bar{X}$ should be approximately Gaussian. We note that a proportion of 0.74 tails corresponds to the following standard score

$$Z = \frac{0.74 - 0.5}{\sqrt{0.25/100}} = 4.8 \tag{6.4}$$

Using the standard Gaussian tables we find

$$P(\{\bar{X} \geq 4.8\} \mid H_n \text{ true}) < 1/10^6 \tag{6.5}$$

If the null hypothesis were true the chances of obtaining 74 % tails or more are less than one in a million. We now have two alternatives: 1) We can keep the null hypothesis in which case we would explain our observations as an extremely unprobable event due to chance. 2) We can reject the null hypothesis in view of the fact that if it were correct the results of our experiment would be an extremely improbable event. Sir Ronald Fisher, a very influential classic statistician, explained these options as follows:

The force with which such conclusion is supported logically is that of a simple disjunction: Either an exceptionally rare chance event has occurred, or the theory or random distribution [the null hypothesis] is not true [**?**]

In our case obtaining 74 % tails would be such a rare event if the null hypothesis were true that we feel compelled to reject the null hypothesis and conclude that the coin is loaded. But what if we had obtained say 55 % tails? Would that be enough evidence to reject the null hypothesis? What should we consider as enough evidence? What standards should we use to reject a hypothesis? The next sections explain the classic approach to these questions. To do so it is convenient to introduce the concept of Type I and Type II errors.

6.2 Type I and Type II errors

Jerzy Neyman and E.S. Pearson developed the bulk of the classic approach to hypothesis testing between 1928 and 1933. Neyman and Pearson emphasized hypothesis testing as a procedure to make decisions rather than as a procedure to falsify hypothesis. In their own words

Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run experience, we shall not be too often wrong. [**?**]

They viewed the task of statistical hypothesis testing as similar to the detection of signals in the presence of noise. Let me illustrate this concept with the following analogy. Suppose you are a fire detector. You are hanging up there in the ceiling of a house and your task is to decide whether the house is on fire. Once in a while you measure the amount of smoke passing through your sensors. If the amount is beyond a critical value you announce the house residents that the house is on fire. Otherwise you stay quiet.

In this analogy the null hypothesis is the theoretical possibility that the house is not on fire and the alternative hypothesis is that the house is on fire. Measuring how much smoke there is out there is the equivalent of conducting an experiment and summarizing the results with some statistic. The information available to us is imperfect and thus we never know for sure whether the house is or is not on fire. There is a myriad of intervening variables that may randomly change the amount of smoke. Sometimes there is a lot of smoke in the house but there is no fire, sometimes, due to sensor failure, there may be fire but our sensors do not activate. Due to this fact we can make two types of mistakes:

1. We can have **false alarms**, situations where there is no fire but we announce that there is a fire. This type of error is known as **a Type I error**. In scientific research, type I errors occur when scientists reject null hypotheses which in fact are true.

2. We can also **miss** the fire. This type of error as **Type II error**. Type II errors occur when scientists do not reject null hypothesis which in fact are false.

Note that for type I errors to happen two things must occur:

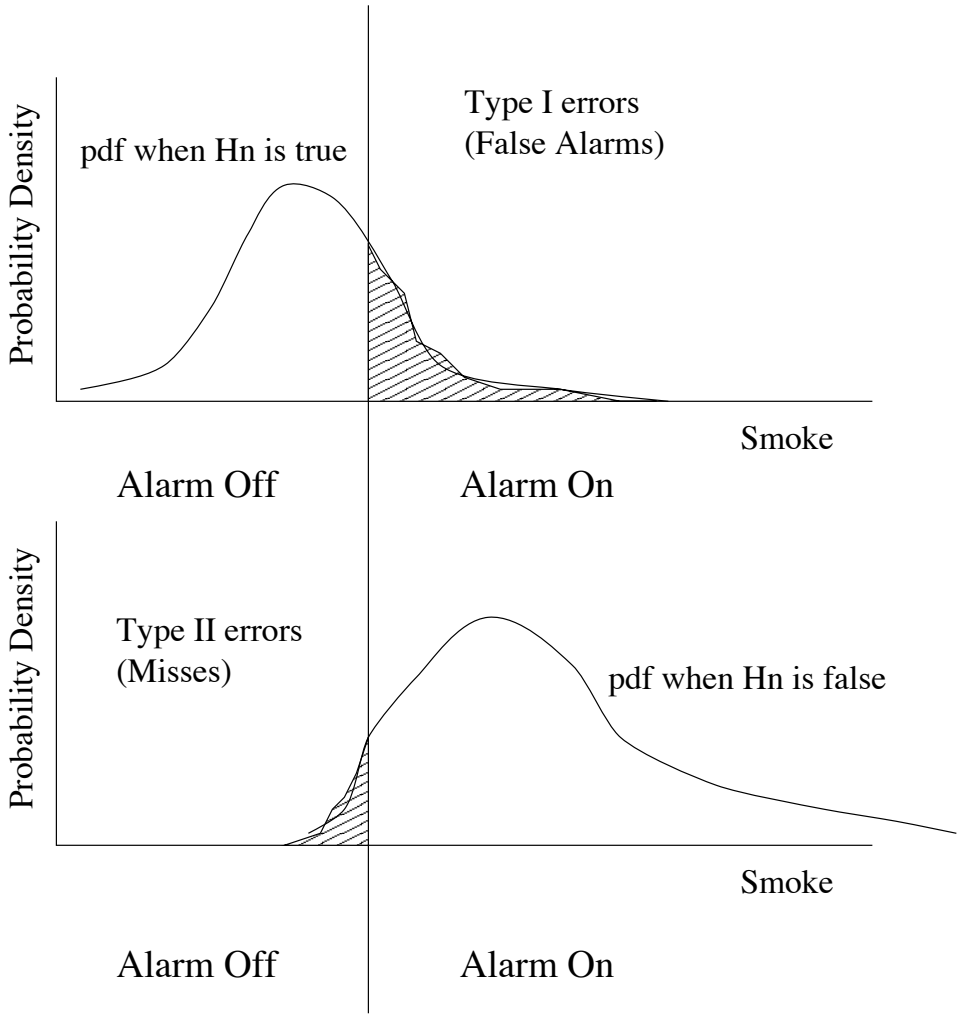

Figure 6.1: An illustration of the process of statistical hypothesis testing. The upper figure shows the distribution of smoke when there is no fire. The lower figure shows the distribution when there is fire. It can be seen that on average there is more smoke when there is fire but there is overlap between the two conditions. The task of hypothesis testing is to decide whether there is a fire based only on the amount of smoke measured by the sensor.

(1) *The null hypothesis must be true.*
2) *We reject it.*

(6.6)

and for type II errors to happen two things must occur:

(1) *The null hypothesis must be false.*

(2) *We do not reject it.* (6.7)

6.2.1 Specifications of a decision system

The performance of a decision system can be specified in terms of its potential to make errors when the null hypothesis is true and when the null hypothesis is false.

- The **type I error specification** is the probability of making errors when the null hypothesis is true. This specification is commonly represented with the symbol α. For example if we say that a test has $\alpha \leq 0.05$ we guarantee that if the null hypothesis is true the test will not make more than 1/20 mistakes.
- The **type II error specification** is the probability of making errors when the null hypothesis is false. This specification is commonly represented with the symbol β. For example if we say that for a test β is unknown we say that we cannot guarantee how it will behave when the null hypothesis is actually false.
- The **Power specification** is the probability of correctly rejecting the null hypothesis when it is false. Thus the power specification is 1.0 - β.

The current standard in the empirical sciences dictates that **for a scientific test to be acceptable the type I error specification has to be smaller or equal to 1/20** (i.e., $\alpha \leq 0.05$). The standard does not dictate what the type II error specification should be. In the next chapter we will see examples of statistical tests that meet this type I error specification.

6.3 The Bayesian approach

The essence of the Bayesian approach can be summarized as follows:

- A recognition of the fact that humans not only make binary decisions about hypotheses but also want to assign degrees of belief to the different hypotheses
- An assumption that probability is useful to describe beliefs not just relative frequencies.
- An emphasis on the problem of how to combine external data with prior knowledge to modify beliefs.

To Fisher's statement that

> *"It should be possible to draw conclusions from the data alone, without apriori assumptions."* [**?**],

L. Savage, a well known Bayesian replies that

> "*We had a slogan about letting the data speak for themselves, but when they do, they tell us how to modify our opinions, not what opinion is justifiable.*" [**?**]

In practice the main difference between Bayesians and frequentists is that Bayesians treat hypotheses as if they were probabilistic events and thus are willing to assign them probability values. Here is an example of how the Bayesian approach would work in practice.

Example: A doctor believes that a patient has a 10% chance of having Lyme disease. She gives the patient a blood test and the test comes out positive. The manual for this test says that that out of 100 patients with Lyme disease, 80 % test positive. Moreover, out of 100 patients with no Lyme disease 30 % test positive. What is the probability that the patient has Lyme disease?

Answer: If you were a pure classic you would be unwilling to answer this question. You would simply say that probabilities should not be applied to empirical hypotheses. This person either has Lyme disease or he does not. You would simply say that your tools do not apply to this problem. If you were Bayesian you would be willing to use probability theory to play with knowledge and internal beliefs so this question makes sense to you. You could represent the hypothesis

that the patient has Lyme disease as an event H_1 which has a prior probability of 10 %.

$$P(H_1) = 0.1 \tag{6.8}$$
$$P(H_0) = 1 - P(H_1) = 0.9 \tag{6.9}$$

where H_0 represents the hypothesis that the patient does not have Lyme disease. The positive test is a data event D with the following characteristics

$$P(D \mid H_1) = 0.8 \tag{6.10}$$
$$P(D \mid H_0) = 0.3 \tag{6.11}$$

Now you could apply Bayes' theorem to compute the probability of the hypotheses given the data

$$P(H_1 \mid D) = \frac{(0.8)(0.1)}{(0.8)(0.1) + (0.3)(0.9)} = 0.23 \tag{6.12}$$

After seeing the results of the test, it would be rational for the Doctor to update her beliefs and give her patient a 23 % probability of having Lyme disease. Note the emphasis here is on upgrading beliefs based on empirical data. The emphasis is not on deciding whether a hypothesis is true or false. Of course it is now up to the doctor and her patient to use the 23 % probability figure to perhaps get a better test or to evaluate the costs and benefits of treatments.

6.4 Exercises

1. Urn A contains 50% black balls 50 % white balls. Urn B contains 45 % black balls 55 % white balls. You get a sample of n randomly selected balls. All balls in the sample belong to the same urn but you do not know which one. Your task is to decide which urn the sample belongs to. Let the null hypothesis be the idea that the sample comes from Urn A.

   (a) Suppose there are 10 balls in the sample (i.e., n = 10).

      i. What would the critical value be for a Type I error specification of 1/20 ?

      ii. What would the power specification be? NOTE: in this case we can calculate the power because it is possible to find the distribution of the mean assuming the alternative hypothesis is true. In general this may not be possible.

iii. What would the critical value be for a Type I error specification of 1/100 ?

iv. What would the power specification be if we use a Type I error specification of 1/100?

(b) Same as the previous 4 questions but using $n = 20$ instead of $n = 10$. What did you learn out of this exercise?

2. True or false (justify your response):

(a) Due to the current standards we should expect one out of 20 scientific studies to be wrong.

(b) An experiment results in a p-value of 0.03, another experiment in a p-value of 0.0001. Using the current 0.05 standard in both experiments we reject the null hypothesis. Moreover, in the second experiment there is a smaller chance of being wrong by rejecting the null hypothesis.

(c) An experiment results in a p-value of 0.00001. Therefore the type I error specification is 0.00001.

(d) An experiment results in a p-value of 0.5. The probability of making a type I error in this experiment is unknown.

(e) An experiment results in a p-value of 0.5. The power specification of this experiment is 0.5.

(f) An experiment results in a p-value of 0.01. The probability of making a type II error in this experiment is unknown.

(g) An experiment results in a p-value of 0.01. The probability of making a type I error in this experiment is unknown.

(h) An experiment results in a p-value of 0.01. The type I error specification is 0.05.

(i) An experiment results in a p-value of 0.01. The probability of making a type II error in this experiment zero.

(j) An experiment results in a p-value of 0.01. The probability of making a type I error in this particular experiment is either 1 or 0.

Chapter 7

Introduction to Classic Statistical Tests

In the previous chapter we saw that the current standards of empirical science use classical statistical tests with a type I error specification of 5 %. In this chapter we see two classic statistical tests: The Z test and the T test. Of the two the second one is more general and by far more important. The Z test is introduced for historical reasons and because it serves as a nice steping stone to the T test.

7.1 The Z test

This test is **used when the null hypothesis specifies the mean and the variance of the observations**. Let's see how the test works using the following example. We toss a coin 100 times and we obtain 55 heads. Is this enough evidence to say that the coin is loaded?

7.1.1 Two tailed Z test

1. We have n random variables

$$X_1, \ldots, X_n \tag{7.1}$$

where X_i commonly represents a measurements from subject i in a sample of n subjects. We assume the following:

- The random variables are independent and identically distributed.

- If $n < 30$ the random variables are Gaussian.

In our example $X_1, \ldots, X_n$ are Bernoulli random variables. If the outcome of toss number i is heads then X_i takes value 1 otherwise it takes value 0.

2. Formulate a null hypothesis that specifies the mean and standard deviation of the mean. We represent these as $E(\bar{X} \mid H_n \text{ true})$ and $\text{Sd}(\bar{X} \mid H_n \text{ true})$

   In our example the null hypothesis H_n is that the coin is fair. In such case $E(\bar{X} \mid H_n \textit{ true}) = E(X_i \mid H_n \textit{ true}) = 0.5$ and

$$\textit{Sd}(\bar{X} \mid H_n \textit{ true}) = \frac{\textit{Sd}(X_i \mid H_n \textit{ true})}{\sqrt{n}} = \frac{0.5}{10} = 0.05 \tag{7.2}$$

3. Define the random variable Z as follows

$$Z = \frac{\bar{X} - E(\bar{X} \mid H_n \text{ true})}{\text{Sd}(\bar{X} \mid H_n \text{ true})} = \frac{\bar{X} - E(\bar{X} \mid H_n \text{ true})}{\text{Sd}(X_i \mid H_n \text{ true})/\sqrt{n}} \tag{7.3}$$

4. Compute, the value taken by the random variable Z,

   In our example

$$Z = \frac{0.55 - 0.5}{0.05} = 1 \tag{7.4}$$

5. If $Z \notin [-1.96, 1.96]$ reject H_n and report that the results were **statistically significant**. Otherwise withhold judgment and report that the resuts **were not statistically significant**.

   In our example $Z \in [-1.96, 1.96]$ so we withhold judgment. We do not have enough evidence to say that the coin is loaded.

Proof: Since this is a classical test we just need to show that with this procedure the type I error specification is no larger than 5 %. In other words $P(H_n\text{rejected}|H_n\text{true}) \leq 0.05$. To do so first we will show that if H_n is true then Z is an standard Gaussian random variable. First note that Z is just a linear combination of $\bar{X}$

$$Z = a + b\bar{X} \tag{7.5}$$

$$a = -\frac{E(\bar{X} \mid H_n \text{ true})}{\text{Sd}(\bar{X} \mid H_n \text{ true})} \tag{7.6}$$

$$b = \frac{1}{\text{Sd}(\bar{X} \mid H_n \text{ true})} \tag{7.7}$$

If $X_1, \ldots, X_n$ are Gaussian then $\bar{X}$ is also Gaussian (because the sum of Gaussian is Gaussian). If $X_1, \ldots, X_n$ are not Gaussian but $n > 30$, by the Central limit theorem, then the cumulative of $\bar{X}$ is approximately Gaussian. Under these conditions Z is also Gaussian (because Z is a linear transformation of $\bar{X}$ and a linear transformation of a Gaussian rav is also Gaussian). Moreover

$$E(Z \mid H_n \text{ true}) = a + bE(\bar{X} \mid H_n \text{ true}) = 0 \tag{7.8}$$

$$\text{Var}(Z \mid H_n \text{ true}) = b^2\text{Var}(\bar{X} \mid H_n \text{ true}) = 1 \tag{7.9}$$

Thus, if the null hypothesis is true and the assumptions of the Z test are met, then Z is a standard Gaussian random variable and

$$P(H_n \text{ rejected} \mid H_n \text{ true}) = P(Z \notin [-1.96, 1.96]) = 2\Phi(-1.96) = 1/20 \tag{7.10}$$

□

7.1.2 One tailed Z test

In this case the null hypothesis includes an entire range of values for the sample mean. For example, the null hypothesis could be that the expected value of the mean is smaller than 0. Or that the expected value of the mean is smaller than 3. The first thing we do is to pick the extreme case proposed by the null hypothesis and which is finite. For example if the null hypothesis is that the expected value of the sample mean is larger than 3, then the extreme case is that the expected value is exactly 3. I'm going to represent this extreme case of the null hypothesis as H_x to distinguish it from the more general null hypothesis that we want to reject. For example if the null hypothesis says $E(\bar{X} \mid H_n \text{ true}) \leq 3$ the the extreme case of the null hypothesis would claim $E(\bar{X} \mid H_x \text{ true}) = 3$.

1. Formulate the null hypothesis and the extreme hypothesis.
2. Compute the value taken by the random variable Z as in the two-tailed case, only in this case use expected values given by H_x instead of H_n.
3. If positive values of Z are the only ones inconsistent with H_n and if $Z > 1.64$ reject H_n. Report that the results were **statistically significant**. If negative values of Z are the only ones inconsistent with H_n and if $Z < -1.64$ then reject the null hypothesis. Report that the results are **statistically significant**. Otherwise withhold judgment. Report that the results **were not statistically significant**.

To see that this test also has the desired type I error specification note that

$$P(H_n \text{ rejected} \mid H_n \text{ true}) \leq P(H_n \text{ rejected} \mid H_x \text{true}) \tag{7.11}$$
$$= P(Z > 1.64) = \Phi(-1.64) = 1/20 \tag{7.12}$$

Thus we have proven that the type I error specification is not larger than 1/20, making the test acceptable from a classical point of view.

Example *We want to show that bees prefer red flowers rather than yellow flowers. We construct 100 plastic flowers 50 of which are yellow and 50 red. Other than the color the flowers are indistinguishable in appearance. We then measure for 200 bees the amount of time they spend in yellow versus red flowers. We find that 160 out of the 200 bees spent more time in the red flowers. Is this enough evidence to say that bees prefer red flowers?*

We can model the preference of each bee as independent Bernoulli random variables $X_1, \ldots, X_{200}$, where $X_i = 1$ represents the event that bee number i in the sample spent more time on the red flowers. In our case the sample mean takes the following value $\bar{X} = 160/200 = 0.8$. The null hypothesis says that $E(\bar{X} \mid H_n \text{ true}) \leq 0.5$. This hypothesis covers an entire range of values so a one tailed test is appropriate. The extreme case of the null hypothesis, symbolized as H_x says that $E(\bar{X} \mid H_x \text{ true}) = 0.5$. Thus,

$$\bar{X} = 0.8 \tag{7.13}$$
$$E(\bar{X} \mid H_x) = 0.5 \tag{7.14}$$
$$\text{Sd}(\bar{X} \mid H_x) = \text{Sd}(X_i \mid H_x)/\sqrt{200} = 0.035 \tag{7.15}$$
$$Z = \frac{0.8 - 0.5}{0.035} = 8.57 \tag{7.16}$$

Since positive values of Z contradict H_n and $Z > 1.64$ we reject the null hypothesis and conclude that bees prefer red flowers over yellow flowers.

7.2 Reporting the results of a classical statistical test

First let us concentrate on the things you should avoid when reporting results:

- Avoid statements like "The null hypothesis was" or "The alternative hypothesis was". These terms are just used to understand the logic underlying statistical tests. You should assume that the readers already know about this logic.
- Avoid statements like "The results prove the idea that", or "The results disprove the idea that". As we know, statistical test are not infallible so scientists avoid strong statements like "prove" and "disprove".
- Avoid statements like "The data shows that very probably there is a difference between the experimental and the control groups...". You cannot say this because we do not know the probability of the null hypothesis. Remember we are using a classical test and thus the null hypothesis is either true or false.
- Avoid statements like "The results are statistically insignificant". The word insignificant is misleading. Simply say that the results are not statistically significant.
- Avoid statements like "The results were very significant". This suggests that your results are very important or that your type I error specification is different from that of other scientists. This is misleading since you would have published your results if your test passed with your type I error specification.

7.2.1 Interpreting the results of a classical statistical test

If the results are not significant, it basically means that we do not have enough evidence, for now, to reject the null hypothesis. Here is an example interpretation:

- There is not enough evidence to support the idea that the drug has an effect on the subjects' reaction time.

If the results are significant we can say that the data are not consistent with the null hypothesis. Here are some standard interpretations of significant results. Note how we avoided strong words like "prove" in favor of softer words like "support"

- The data support the idea that the drug has an effect on the subjects' reaction time.
- The results of this experiment support the idea that drug X has an effect on reaction time.

7.3 The T-test

One major problem with the Z test is that the null hypothesis needs to specify the value of $\text{Var}(\bar{X})$. This value is known for Bernoulli random variables, but in general it is unknown. In such a case, we may estimate the variance of the mean using our sample information. To do so first define the sample variance as

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{7.17}$$

note that the sample variance is a random variable. In the Appendix to this Chapter we prove that the sample variance is an unbiased estimate of the variance of the observations, i.e.,

$$E(S_X^2) = \text{Var}(X_i) \tag{7.18}$$

Since S_X^2 is an unbiased estimate of the variance of the observations, we can divide it by n to get an unbiased estimate of the variance of the sample mean. We represent this estimate as $S_{\bar{X}}^2$

$$S_{\bar{X}}^2 = \frac{S_{X_i}^2}{n} \tag{7.19}$$

Now remember the Z random variable was defined as follows

$$Z = \frac{\bar{X} - E(\bar{X} \mid H_n \text{ true})}{\text{Sd}(\bar{X} \mid H_n \text{ true})} = \frac{\bar{X} - E(\bar{X} \mid H_n \text{ true})}{\text{Sd}(X \mid H_n \text{ true})/\sqrt{n}} \tag{7.20}$$

The T random variable is very similar to the Z random variable except for the fact that in the denominator it uses an estimate of the variance of the mean instead of the true variance

$$T = \frac{\bar{X} - E(\bar{X} \mid H_n \text{ true})}{S_{\bar{X}}} = \frac{\bar{X} - E(\bar{X} \mid H_n \text{ true})}{S_X/\sqrt{n}} \tag{7.21}$$

7.3.1 The distribution of T

The distribution of the T random variable was discovered by William Sealy Gosset (1876–1937). Gosset was a chemist and mathematician hired by the Guinees brewery in 1899. Brewing is a process in which statistical analysis had great potential since brewers have to deal with variable materials, temperature changes and so on. The story goes that at the end of the 19^{th} century scientific methods and statistical modeling were just beginning to be applied to brewing but the methods available at the time required large numbers of observations. Gosset had to work with experiments with small samples, for which the Z test did not work very well. As a statistician Gosset liked to do everything starting from first principles and disliked the use of recipes and tabulations. This gave him great flexibility and power when tackling new problems. He once said "Doing it from first principles every time preserves mental flexibility". Gosset was also a very good carpenter, an activity to which he also applied his devotion to first principles; he disliked the use of complicated tools and liked doing as much as possible with a pen-knife. In 1908 Gosset published a paper entitled "The probable error of the mean", where he described the distribution of the T random variable, as a statistic applicable to experiments with small numbers of observations. He published this paper under the name "Student". His desire to remain anonymous gave him a romantic, unassuming reputation. His theoretical distribution became known as the Student-distribution, or simply the T-distribution. In fact there is an infinite number of *T-distributions*, with each member of the family identified by a parameter known as the **degrees of freedom** (df). The probability density function of T is given by the following formula:

$$f_T(u) = \frac{1}{K(df)}(1 + \frac{u^2}{df})^{-(df+1)/2} \tag{7.22}$$

$$K(df) = \int_{-\infty}^{\infty}(1 + \frac{x^2}{df})^{-(df+1)/2}\, dx \tag{7.23}$$

In practice, we do not need to worry about the formula for $f_T(u)$ since the values for the cumulative distribution

$$F_T(v) = \int_{\infty}^{v} f_T(x)dx \tag{7.24}$$

appears in most statistics textbooks. Figure 3 shows the shape of the T-distribution with 1 df. The distribution is bell shaped but it has longer tails than the Gaus-

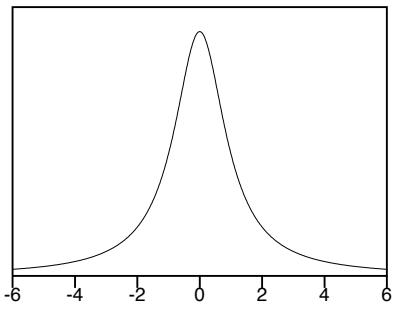

Figure 7.1: The t-distribution with one degree of freedom.

sian distribution. When the number of degrees of freedom is very large, the T-distribution closely approximates the Gaussian distribution.

7.3.2 Two-tailed T-test

Here is the procedure to perform a two-tailed T-test

1. Figure out the expected value of the sample mean if the null hypothesis is true $E(\bar{X} \mid H_n \text{ true})$

2. Compute T the value taken by T for your specific sample.

3. Get the degrees of freedom (i.e., number of independent observations) in $S_{\bar{X}}$. In our case the number of degrees of freedom is $n - 1$.

4. Go to T tables for $n - 1$ degrees of freedom, and compute the value c such that
$$P(T > c \mid H_n \text{ true}) = 1/40 \tag{7.25}$$
The number c is called the critical value.

5. If $T \notin [-c, c]$ reject the null hypothesis, otherwise withhold judgment.

Example: We get 3 bags of potato chips from a specific brand. The company that makes these chips says that on average there are 40 chips per bag. Let X_i measure the number of chips in bag i. In our sample we have $X_1 = 30$, $X_2 = 10$

and $X_3 = 20$. Do we have enough evidence to reject the idea that on average there are 40 chips per bag?

The null hypothesis is that there are 40 chips per bag. Thus $E(\bar{X} \mid H_n \text{ true}) = 40$. Moreover

$$\bar{X} = (30 + 10 + 20)/3 = 20 \tag{7.26}$$

$$S_X^2 = \frac{(30-20)^2 + (10-20)^2 + (20-20)^2)}{3-1} = 100 \tag{7.27}$$

$$S_{\bar{X}} = \sqrt{S_X^2/n} = \sqrt{100/3} = 5.77 \tag{7.28}$$

$$T = \frac{20-40}{5.77} = -3.46 \tag{7.29}$$

$$df = n - 1 = 2 \tag{7.30}$$

$$c = 4.30 \tag{7.31}$$

I obtained the critical value $c = 4.30$ using the tables of the T statistic with $df = 2$. Since $T \in [-4.30, 4.30]$ we withhold judgment. We do not have enough evidence to reject the null hypothesis.

Proof: This procedure has the correct type I error specification since

$$P(H_n \text{ rejected} \mid H_n \text{ true}) = P(T \notin [-c, c]) \tag{7.32}$$

$$= 2P(T > c \mid H_n \text{ true}) = 1/20 \tag{7.33}$$

Procedure for one-tailed T-test In this case the null hypothesis takes an entire range. For example, it could say that the expected value of the mean is smaller than 20, not just 20. As with the Z test I will use H_n to represent the null hypothesis and H_x to represent the extreme case in this null hypothesis. For example if the null hypothesis says that the expected value of the mean is smaller than 20, then H_x says that the expected value of the mean is exactly equal to 20. The procedure to test one-tailed hypotheses is as follows

1. Formulate the null hypothesis and the extreme hypothesis.
2. Compute the value of T and df as in the two-tailed case, only in this case use expected values given by H_x instead of H_n.
3. Using tables find the value c such that $P(T > c \mid H_x) = 1/20$

4. If positive values of T are the only ones inconsistent with H_n and if $T > c$ reject H_n. If negative values of T are the only ones inconsistent with H_n and if $T < -c$ reject the hypothesis. Otherwise withhold judgment.

For example if we want to show that the consumers are being cheated. Then we want to reject the hypothesis that the average number of chips is larger or equal to 40. In such case the null hypothesis is $E(\bar{X} \mid H_n) \geq 40$. The extreme case of the null hypothesis says $E(\bar{X} \mid H_x) = 40$. Moreover only negative values of T are inconsistent with the null hypothesis. We go to tables and find that for 2 degrees of freedom $P(T > -2.91) = 1/20$. Since $T = -3.46$ is smaller than -2.91, we have enough evidence to reject the null hypothesis.

Proof: This procedure has the desired type I error specification since

$$P(H_n \text{ rejected} \mid H_n \text{ true}) \leq P(T \notin [-c, c] \mid H_x \text{ true}) \tag{7.34}$$

$$= 2P(T > c \mid H_x \text{ true}) = 1/20 \tag{7.35}$$

7.3.3 A note about LinuStats

LinuStats, and some T-tables provide probabilities of the absolute value of T, i.e.,

$$P(|T| \geq c) = P(\{-T \leq c\} \cup \{T \geq c\}) = 2P(T \geq c) \tag{7.36}$$

Thus, if you want to obtain the critical value c such that $P(T \geq c) = x$ then you need to give LinuStats the value $2x$. If you do that LinuStats will give you the value c such that

$$P(|T| \geq c) = 2x \tag{7.37}$$

from which it follows that

$$P(T \geq c) = 2x/2 = x \tag{7.38}$$

which is what you wanted to begin with.

7.4 Exercises

1. Which of the following is more likely not to belong to a population with $\mu = 0$ and $\sigma^2 = 100$. Explain why. (5 points)

(a) A single subject with a value of 20.

(b) A random, independent sample of 100 subjects with a sample mean of -2.0

(c) A random, independent sample of 10000 subjects with a sample mean of 1.0

2. *A research team is testing whether newborns have color vision. Ten infants were tested at the UCSD hospital within their first 3 hours of life. Experimenters showed the infants a card with 2 color patches: red and green. The card was positioned so that the distance between the two color patches was about 15 degrees of visual angle. Each color patch was circular with a diameter of 7 degrees of visual angle. For half the infants red was on the right side, of the card and for the other half red was on the left side. The two colors were calibrated to have equal luminance, so that they could not be distinguishable by a black-and-white vision system. After parental consent was granted, the infants were tested as follows. Once it was decided that the infant was in an alert state, the card was presented in front of him/her and an observer recorded, during a 30 second period how long infants looked to the left and right side of the card. The observer did not know whether left corresponded to red or to the green patch. The dependent variable was the total time in seconds looking to the red patch minus the time looking to the green patch. The results were as follows:* $\{-2, -1, 10, 8, 4, 2, 11, 1, -3, 4\}$.

   (a) Formulate the null hypothesis

   (b) Should the test be 1-tail or 2-tails?

   (c) Do you have enough evidence to reject the null hypothesis?

   (d) What is the probability that you made a type I error?

   (e) What is the probability that you made a type II error?

   (f) What is the type I error specification?

   (g) What is the power specification?

   (h) What are the assumptions needed to guarantee a 5% type I error specification?

3. You are given 3 randomly selected kittens of the same age. Their weights are 3, 3 and 2 pounds. Your biology book says that 1 year old kittens weigh 7.0 pounds. Your goal is to prove that these kittens are not 1 year old.

(a) What is the null hypothesis?

(b) Do you have enough evidence to say that the kittens are not 1 year old?

(c) What is the probability that you made a type I error?

(d) What is the probability that you made a type II error?

(e) What is the type I error specification?

(f) What is the power specification?

(g) What are the assumptions needed to guarantee a 5% type I error specification?

7.5 Appendix: The sample variance is an umbiased estimate of the population variance

Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean μ_X and variance σ_X^2. To ease the presentation I will prove that $E(S_X^2) = \sigma_X^2$ for the case in which $\mu_X = 0$. Generalizing the proof for arbitrary values of μ_X is mechanical once you know the proof for $\mu_X = 0$. First note if $\mu_X = 0$ then

$$\text{Var}(X_i) = E(X_i - \mu_X)^2 = E(X_i^2) = \sigma_X^2 \tag{7.39}$$

$$\text{Var}(\bar{X}) = E[(\bar{X} - E(\bar{X}))^2] = E(\bar{X}^2) = \frac{\sigma_X^2}{n} \tag{7.40}$$

and since X_i, X_j are independent when $i \neq j$

$$E(X_i X_j) = E(X_i)E(X_j) = 0 \tag{7.41}$$

Using the definition of S_X^2, and considering that $X_1, \ldots, X_n$ have the same distribution, we get

$$E(S_X^2) = \frac{1}{n-1} \sum_{i=1}^{n} E(X_i - \bar{X})^2 = \frac{n}{n-1} E(X_1 - \bar{X})^2 \tag{7.42}$$

Moreover

$$E(X_1 - \bar{X})^2 = E(X_1^2) + E(\bar{X}^2) - 2E(X_1\bar{X}) = \sigma_x^2 + \frac{\sigma_X^2}{n} - 2E(X_1\bar{X}) \tag{7.43}$$

Finally

$$E(X_1\bar{X}) = E(X_1\frac{1}{n}\sum_{i=1}^{n} X_i) \tag{7.44}$$

$$= \frac{1}{n}\big(E(X_1^2) + \sum_{i=2}^{n} E(X_1X_j)\big) = \frac{\sigma_X^2}{n} \tag{7.45}$$

Thus

$$E(S_Y^2) = \frac{n}{n-1}\big(\sigma_X^2 + \frac{\sigma_X^2}{n} - 2\frac{\sigma_X^2}{n}\big) = \sigma_X^2 \tag{7.46}$$

□

Chapter 8

Intro to Experimental Design

The goal of experimentation is to study causal relationships between physical events. Behind every scientific experiment there are individuals and indirectly entire societies trying to figure out whether two or more phenomena are related (e.g., are lung cancer and smoking related?, does a drug improve the life-expectancy of patients?, does an education program help economically disadvantaged students?). Different experimenters have to tackle issues specific to their disciplines but generally speaking empirical scientists share many deal of common problems: 1) They need to deal with the variability and uncertainty inherent in natural data, 2) They need to organize and communicate the obtained data in an efficient fashion, 3) They need to make inferences based on limited samples of data, 4) They need to design experiments carefully to leave as few possible interpretations of the results as possible.

8.1 An example experiment

In this chapter we will discuss general concepts of experimental design. For concreteness, I will introduce these concepts in relation to the following experiment which was part of J. Ridley Stroop's [**?**] doctoral dissertation. Stroop was intrigued by the well known fact that it takes longer to name colors than to read color names (i.e., it takes longer to say that a red patch is red than to read the word "RED"). Theories of the time proposed explanations based on interference effects and thus Stroop decided to study the effect of interfering color names upon naming colors. The results, which were rather spectacular, are nowadays known as the "Stroop effect". To accommodate our needs I have modified Stroop's orig-

inal experiment while maintaining the overall spirit of his work. You can read his original experiment at the 1935 issue of the Journal of Experimental Psychology [**?**].

- **Materials**:

  There were 2 lists of stimuli:

  - One list of 100 stimuli each of which consisted of four capital X letters "XXXX". The color of the ink varied randomly on each stimulus. These lists were used for the "Neutral" test condition.
  - One list of 100 stimuli consisting of color names (e.g. "GREEN") but with each word printed with ink of color different from that of the color named by the word. These lists were used for the "Interference" test condition. The colors used were red, blue, green, brown, and purple. The colors were arranged randomly while making sure that no color would immediately follow itself. The words were printed an equal number of times on each color (except for the color they named).

- **Subjects and Procedure**:

  Twenty volunteer college undergraduate students (10 males and 10 females) participated in the experiment. All participants were tested individually. They were seated near the window so as to have good daylight illumination from the left side. A ten-word sample of each test was read before reading the test the first time. The instructions were to name the colors as they appeared in regular reading line as quickly as possible and to correct all errors. On the signal "Ready! Go!" the sheet which the subjects held face down was turned by the participant and read aloud. The words were followed on another sheet by the experimenter and the time was taken with a stop watch to a fifth of a second. Within each sex category half of the participants were randomly assigned to the Neutral condition and half to the Interference condition.

- **Results**

  Means and standard deviations of the time it takes to name 100 stimuli appear in Table I.

Sex	Neutral		Interference	
male	111.1	[21.6]	69.2	[10.8]
female	107.5	[17.3]	61	[10.5]

Table 8.1: Mean Time and Standard Deviation [in square brackets] per 100 stimuli. Time measured in seconds.

8.2 Independent, Dependent and Intervening Variables

By convention the variables manipulated by the experimenter are known as the **independent variables** (e.g, the interference level). The different values of the independent variable are called **treatment levels**, or simply **treatments**. For example, in Stroop's experiment the treatments were the "Neutral" and "Interference". Experiments are designed to test whether the different levels of the independent variable have an effect on another variable known as the **dependent variable**. In Stroop's experiment the dependent variable is the reaction time.

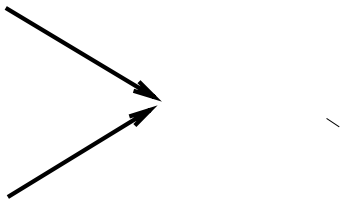

Figure 8.1: Independent and intervening variables have a potential effect on the dependent variable.

Intervening variables are variables other than the independent variable that have a potential effect on the dependent variable. Even though we are interested on the relationship between the independent and the dependent variables, in practice there are many other variables that are also having an effect on the dependent variable. For example, the temperature of the room, the age of the subject, the particular mood of the subject at test time, may also have an effect on the depen-

dent variable. Some experimenters emphasize the fact that they are not interested in the effects of some intervening variables, by calling them **nuisance variables**.

Good experimental designs control for all possible intervening variables so that the only systematic source of variation that can explain the results is the independent variable. If a study is poorly designed and intervening variables may actually explain our results we say that the design lacks **internal validity**, or that it has **confounding variables**. For example, if we choose all the subjects in the neutral condition to be males and all the subjects in the interference condition to be females, we would say that sex is a confounding variable. The design lacks internal validity and we would not be able to tell whether the effects on the dependent variable are due to the variable "sex" or to the treatment conditions (Neutral vs. Interference).

The art of experimental design consists of making sure intervening variables distribute in a "fair" fashion amongst the different treatment conditions so that the only systematic differences between the different treatment groups can be attributed to the independent variable. The concept of "fair" distribution of the intervening variables is very tricky and we should return to it later. For now we'll just use the concept intuitively and say that when an intervening variable is distributed in a "fair" fashion amongst the treatment conditions we say that the variable has been **controlled**. If an intervening variable has been controlled then it cannot be used to explain away our results. For example, if all our subjects are females, and we still find a difference between the two neutral and the interference conditions, clearly sex cannot explain this difference. By holding the sex of all subjects constant we have controlled its possible effect. In scientific parlance we'd say that sex is a controlled variable. Only when all possible intervening variables are controlled we can proceed to analyze whether the independent variable X has an effect on the dependent variable Y. In the next section we discuss the most common methods of control available to scientists.

8.3 Control Methods

1. **Holding constant:** The idea is simple. We make sure an intervening variable is held constant in all our observations. This disqualifies it as a potential explanation of our results. For example if all subjects were treated in the same room at the same temperature, it follows that temperature cannot explain the obtained differences between treatment conditions. Holding constant is a rather dramatic method of control but it has its problems. For

example, if we decide to hold "sex" constant by studying only male subjects, we no longer know whether the results are also applicable to female subjects. In the lingo of experimental design, we say that holding things constant reduces the **external validity** of the experiment. External validity simply means the power to generalize our results to populations larger than the specific set of observations obtained in the experiment (e.g., our ability to say that the Stroop effect occurs to subjects other than the 20 subjects investigated in our experiment).

2. **Blocking:** The idea in blocking methods is to basically replicate the study with different levels of the blocked intervening variable. For example we may "block" age by categorizing subjects into less than 15, 15-30, 30-40, 40-50, 50-60, more than 60). Then we can systematically study the experimental results within each of the different blocks. An important form of blocking is **blocking by subject**, which indicates that we actually block the variable "subject identity", thus controlling all intervening variables, known and unknown, that make subjects different from each other. When we block by subject, each subject is considered as a block and he/she goes through all the treatment conditions in the experiment. The idea is to study the effect of a treatment within each subject (e.g. study the difference between the neutral and interference conditions on a subject by subject basis). In this case the experiment is said to be **within subjects**, or **repeated measures** as opposed to **between subjects**. Within subject designs are a bit trickier to analyze and thus we will later dedicate them a special chapter. For now we will concentrate on between subject designs.

3. **Probabilistic Control:** This is by far the most important methods of control and it is arguably the one form of control that made rigorous experiments in the social and biological sciences possible. The fact is that even though we may try to control for known intervening variables, there will always be intervening variables unknown to us that may have a potential effect on our results. Hey, may be the position of Jupiter at the time of the experiment has an effect on some people's capacity to complete the Stroop task. Who knows?

   So here is the problem: How can we possibly control for intervening variables we cannot even think of? Scientists have found a very ingenious way to solve this problem. The trick is to give up equal distribution of intervening variable in *specific experiments*, instead what we maintain equal

amongst the treatment conditions is the **probability** distribution of the variable. We think of an ensemble of replications of the same experiment and guarantee that over this entire ensemble, the intervening variables will be distributed evenly across the different treatment conditions. When we do that, we say that the intervening variables have been **randomized**.

We have actually seen control by randomization in previous methods even though I avoided calling it by its name. For example, in control by matching we first clustered subjects in attempt to distribute intervening variable equally. But since we had differences between the subjects within each cluster, we then randomly assigned subjects to each treatment condition. In **completely randomized designs** we simply assign subjects randomly to each condition without any previous matching.

We can think of the subject identity as an intervening variable that includes all the peculiarities, known and unknown to us, that make each subject different from everybody else at test time. We then randomize the effects of all these known and unknown variables by assigning each subject randomly across the different experimental conditions. For example, we could put as many cards in a hat as the number of different treatment conditions. Then we can ask each subject in our experiment to draw a card identifying his/her treatment condition. Note that by using probabilistic control, experiments become **a game of chance**, thus the importance of probability theory and statistics in the analysis of experiments.

Note again that randomization does not guarantee equal distribution of intervening variables in single experiments, however it guarantees that on average, if we were to repeat the same experiment many many times, the intervening variables will be equally distributed among the different treatment groups. Technically speaking, when a variable is randomized, we say that it is a **random variable**, a mathematical object studied in probability theory and statistics. We will finish up this chapter with some useful concepts.

4. **Matching:** You can think of matching as controlled randomization. For example, if we simply randomly assign subjects to different treatment conditions, it is always possible that in specific experiments exceptional subjects accumulate disproportionally in some treatment conditions. Matching techniques try to control for these possible imbalances in the following way. First we cluster cluster our subjects in terms of their similarity with respect to a particular variable. Each cluster must has as many subjects as treatment

conditions. For example, if we wanted to "match by age" in the Stroop experiment, we would put together groups of 2 people of similar age. If we had 4 subjects with ages 18, 49, 30, and 16, we would make a cluster with the younger subjects (18, 16) and another cluster with the older subjects (49,30). Once the clusters are formed, subjects within each cluster are **randomly assigned** to different treatment groups.

5. **Statistical Control:** In many cases we may not be able to guarantee a uniform distribution of the intervening variables but we may at least try to make some aspect of the distribution of variables be the same in all the treatment groups. For example, it would be impossible for the room temperature to be exactly the same all the time but at least we may want to guarantee than on average the room temperature was the same across treatment conditions. We may also want to make sure that the average age of the subjects is about the same across all treatment groups. Note that equalizing averages does not guarantee equal distributions. For example, treatment group 1 could have 50 % 5 year old subjects and 50% 35 year old subjects. Group 2 could have just 20 year old subjects. Both groups have the same average age but obviously the distribution is completely different. Although this form of control is rather weak, in some cases it may be all we can do. In such cases we report in a crystal clear way our method of control and leave it it is up to the scientific community to decide on the validity of our results.

8.4 Useful Concepts

1. **Experiments vs. Observational studies:**

   Scientific studies are commonly classified into observational and experimental depending on whether the independent variable is observational or experimental. An independent variable is **Experimental** if its value is assigned by the experimenter to the subjects.[1] For example if experimenters assign different subjects different dosages of a drug then the drug's dose is an experimental variable. The crucial point here is that the experimenter, not the subjects, decides which treatment each subject will receive.

[1] Note that being "experimental" is a property of the independent variable, not a property of the dependent variable. If you want to know whether a study is experimental you need to focus on the independent variable; do not worry about the dependent variable.

If an independent variable is not experimental then it is called **observational**. For example, age does not qualify as an experimental variable since age cannot be assigned to subjects. All experimenters can do is to classify subjects according to their age.

The goal of experimental independent variables is to study cause and effect relationships (e.g., to establish whether a drug is or is not a beneficial to treat a particular sickness). Observational variables on the other hand can only be used to study relationships, without establishing causes and effects. For example, early observational studies about the relationship between smoking and cancer showed that smokers tended to have a higher cancer rate than non-smokers. The dependent variable was whether the subject had cancer. The independent variable was whether the person smoked. Note in this case the independent variable is observational, since it would be unethical for experimenters to force a group of subject to smoke. The experimenters could only classify subjects into smokers or non-smokers. Unfortunately, the results of observational studies like this were inconclusive since they could be due to an intervening variable that is systematically different between the two groups. For example highly stressed subjects could have a propensity to smoke and a higher cancer risk. According to this view, stress, and not smoking per se, could cause cancer. It could also be the case that subjects with cancer have a higher propensity to smoke just because cancer stresses you out. According to this view it is cancer that causes smoking!

A classic example of observational studies involves the analysis of differences between men and women on some dependent variable of interest. Note that since these these studies are observational. We can investigate whether men and women are different on some dependent variable (e.g., memory, IQ, or income) but we cannot assess what the causes are for the obtained differences. These causes could be due to trivial variables, such as the fact that on average men tend to be heavier, or they could be due to complex socio-political variables (e.g., the underlying cause for the observed differences could be that the educational system treats men and women differently causing).

2. **Exploratory vs. Confirmatory Studies:** The distinction between these two different types of research is somewhat subtle, yet important. Confirmatory studies tend to have only a small and well defined set of treatments and behaviors under study. Clear a-priori hypotheses are made about what

to expect in the data and only those a-priori hypothesis are tested. For example, a confirmatory study may investigate whether taking an aspirin a day reduces the chances of heart attack. In exploratory studies the experimenters may have a general idea about what to expect in the data but they are willing to analyze the data in a very wide variety of ways with the hope of finding potentially interesting results. Exploratory studies are extremely important in science and they are the source in many cases of new unexpected results. However, the results in exploratory studies tend to be inconclusive basically because by analyzing the data in many different ways it is always possible to find some sort of pattern that may just be due to chance. You can see this effect in sports analysis when all sorts of different statistics are used to explain after the fact why a team has lost or won. In general, results of exploratory analysis are treated with caution and attempts are made to replicate these results with well defined confirmatory experiments.

3. **Random Assignment vs. Random Selection:** This is a very important distinction that novices sometimes confuse. Random selection refers to the process of selecting a sample from a population of subjects. In many experiments this selection does not need to be random. For example, in psychophysical experiments may times the experimenter himself and his colleagues are the subjects of the study, and thus they are not "randomly selected". However these non-randomly selected subjects may still be randomly assigned to the different treatment conditions. Random selection has an effect on the external validity of the experiment, our power to generalize to a larger population. Random assignment has an effect on the internal validity of the experiment, our power to infer that the independent variable is the one causing the observed results.

8.5 Exercises

1. The following questions refer to the experiment outlined at the beginning of this chapter:

   (a) What are the independent and dependent variables?

   (b) Find intervening variables that where held constant, matched and randomized.

   (c) Is the experiment experimental or observational? Explain why.

(d) Is the experiment within subjects or within subjects?

2. Design a between subject experiment to test whether bees can distinguish red from green.

3. Design a between subject experiment to test whether human newborns can tell the difference between male and female faces.

4. Design an experiment to test whether caffeine has an effect on the subjective feeling of Nervousness. Make sure you specify the following:

   (a) Independent variable and its implementation in the different treatment groups.

   (b) Dependent variable and how you are going to measure it.

   (c) Intervening variables and how you will control them.

5. Consider the following hypothetical experiment inspired on research by Baker and Theologus [**?**] on the effects of caffeine on visual monitoring.

   The purpose of the was to asses the effect of caffeine on a visual monitoring task that simulated automobile night driving. Subjects sat in a semi-darkened room approximately 12 oft from a visual display that consisted of two 1-inch red lights spaced 6 in. apart. At random intervals ranging from 1.5 to 3.5 min., the red lights were driven apart at a rate of 12.56 in/min for 30 seconds. The geometry of the viewing condition corresponded to a driving situation in which one vehicle followed 60 yd. behind a lead vehicle at night driving. The speed at which the lights separated simulated what would be perceived if the lead vehicle reduced its velocity by 0.92 mph. Subjects had to continuously monitor the two red-lights and press a button whenever they detected any separation of the two lights. The reaction time, from the beginning of the light separation to the initiation of the subject;s response was automatically recorded. Testing was conducted over a 4-hr period with 20 trials being administered each hour. for a total of 80 responses per subject. During the 4 hour testing period an FM radio tuned to a local station was played to avoid excessive sensory deprivation. There were 10 paid male volunteers drawn from universities in the Washington DC area. Average age was 29.91 yrs. Subjects were matched by age (e.g. divided into 5 couples with no more than a year difference). Each member of the age-matched couple was randomly assigned to a different treatment

group. There were two such treatment groups: 1) A control group of subjects given placebo tablets just before testing began and 2 hours before the end of testing. 2) An experimental group of subjects that were given a tablet with 200 mg. of caffeine just before testing began, and another tablet with 200 mg. of caffeine 2 hours before the end of testing. Drug and placebo tablet administration was conducted according to a double-blind paradigm, with tablets placed in coded envelopes in a manner unknown to the subjects and test administrators.

(a) What are the independent and dependent variables?

(b) Find intervening variables that where held constant, matched and randomized.

(c) Is the study experimental or observational? Explain why.

(d) Is the study within subjects or within subjects?

6. What is the difference between external and internal validity?

7. A friend of yours observes that swimmers nicer bodies than other athletes. She concludes that swimming helps have a good body. What's wrong with this conclusion? Design an experiment to test your friend's hypothesis.

8. Read the Abstract, Introduction and Methods sections of the article in Reference [**?**]. Describe the control methods used in the study.

Chapter 9

Experiments with 2 groups

In this chapter we learn how to apply the T test to experiments with 2 groups. For example one group may receive a drug and another group may receive a placebo. The form of the test depends on whether the experiment is performed with different subjects in each group or whether the same subjects are used for the two groups. The first type of experiment is called **between subjects** or **randomized**. The second is called **within subjects** or **repeated measures**.

9.1 Between Subjects Experiments

In this case we have two sets of random variables. The first set describes measurements in a group of subjects, the second set describes measurements in a different group of subjects. The number of subjects per group is the same, and we will represent it with the letter n. Thus, the total number of subjects in the experiment is $2n$. Let $X_{1,1}, \ldots, X_{1,n}$ represent the measurements for the first group and $X_{2,1}, \ldots, X_{2,n}$ represent the measurements for the second group. Let $\bar{X}_1$ represent the average of the observations in the first group and $\bar{X}_2$ the average of the

observations in group 2. We thus have

$$\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1,i} \tag{9.1}$$

$$\bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2,i} \tag{9.2}$$

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{1,i} - \bar{X}_1)^2 \tag{9.3}$$

$$S_{\bar{X}_1}^2 = \frac{1}{n}S_1^2 \tag{9.4}$$

$$S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{2,i} - \bar{X}_2)^2 \tag{9.5}$$

$$S_{\bar{X}_2}^2 = \frac{1}{n}S_2^2 \tag{9.6}$$

The null hypothesis specifies the expected value of the difference between the sample means of the two groups, i.e, $E(\bar{X}_2 - \bar{X}_1 \mid H_n \text{ true})$. In most experiments the null hypothesis says, that this expected value is zero, i.e., that on average the two groups are no different. The variable of interest in this case is $\bar{X}_2 - \bar{X}_1$. We assume that the random variables $X_{1,1}, \ldots, X_{1,n}$ and the random variables $X_{2,1}, \ldots, X_{2,n}$ are independent and Gaussian. Moreover we assume that within each group all the random variables have the same expected value

$$E(X_{1,1}) = E(X_{1,2}) = \cdots = E(X_{1,n}) \tag{9.7}$$

$$E(X_{2,1}) = E(Y_{2,2}) = \cdots = E(Y_{2,n}) \tag{9.8}$$

$$\tag{9.9}$$

and that all the random variables have the same variance

$$\text{Var}(X_{1,1}) = \cdots = \text{Var}(X_{1,n}) = \text{Var}(X_{2,1}) = \cdots = \text{Var}(X_{2,n}) = \sigma_X^2 \tag{9.10}$$

where we do not the value of σ_X^2. Note in single group experiments the variable of interest is the sample mean for that group and we use an estimate of the standard deviation of the sample mean. In two group experiments the random variable of interest is $\bar{X}_2 - \bar{X}_1$, i.e, the difference between two group means, and thus we need to come up with an estimate of the standard deviation of $(\bar{X}_2 - \bar{X}_1)$. Note

that since $\bar{X}_1$ and $\bar{X}_2$ are independent ravs

$$\text{Var}(\bar{X}_2 - \bar{X}_1) = \text{Var}(\bar{X}_2) + \text{Var}(-\bar{X}_1) = \text{Var}(\bar{X}_2) + \text{Var}(\bar{X}_1) \tag{9.11}$$

$$= 2\text{Var}(\bar{X}_1) = \frac{2}{n}\text{Var}(X_{1,1}) \tag{9.12}$$

The last two steps are valid because we assume $\text{Var}(X_{1,1}) = \cdots = \text{Var}(X_{2,n})$. Thus $\text{Var}(\bar{X}_1) = \text{Var}(\bar{X}_2)$. We can get an umbiased estimate of the variance of the observations by averaging S_1^2 and S_2^2. We represent this pooled estimate as S^2

$$S^2 = \frac{1}{2}(S_1^2 + S_2^2) \tag{9.13}$$

An unbiased estimate of the variance of the difference between sample means is as follows

$$S^2_{\bar{X}_1 - \bar{X}_2} = \frac{2}{n}S^2 = S^2_{\bar{X}_2} + S^2_{\bar{X}_1} \tag{9.14}$$

and the T random variable is defined as follows

$$T = \frac{\bar{X}_2 - \bar{X}_1 - E(\bar{X}_2 - \bar{X}_1 \mid H_n true)}{S_{\bar{X}_2 - \bar{X}_1}} = \frac{\bar{X}_2 - \bar{X}_1 - E(\bar{X}_2 - \bar{X}_1 \mid H_n true)}{\sqrt{S^2_{\bar{X}_2} + S^2_{\bar{X}_1}}} \tag{9.15}$$

In this case the number of degrees of freedom is $2(n-1)$ since $S_{\bar{X}_2 - \bar{X}_1}$ is based on $2(n-1)$ independent observations, i.e., $n-1$ independent observations in S_1^2 and $n-1$ independent observations in S_2^2.

If H_n is true and the assumptions are met, then T follows the distribution described by Gosset, which is available in tables. The procedure to do one tailed and two tailed tests is identical as for one group experiments provided the new formulas for T and for df are used.

Example *A group of 6 migrane patients was gathered and randomly assigned to one of two conditions. In the experimental group subject were given 500 mg of a new migrane drug whenever they had a migrane access. In the control group subjects were given two aspirins. One hour after the access, subjects were asked to rank from 0 to 100 how bad the headache was. 0 meaning, "headache? what headache?", and 100 meaning "I cant take it anymore". The results were as follows*
Experimental Group =$[10, 30, 20]$
Control Group = $[56, 64, 60]$

First of all we compute the values taken by the relevant random variables in our experiment

$$\bar{X}_1 = (10 + 30 + 20)/3 = 20 \tag{9.16}$$
$$\bar{X}_2 = (56 + 64 + 60)/3 = 60 \tag{9.17}$$
$$S_1^2 = ((10 - 20)^2 + (30 - 20)^2 + (20 - 20)^2)/2 = 100 \tag{9.18}$$
$$S_2^2 = ((56 - 60)^2 + (64 - 60)^2 + (60 - 60)^2)/2 = 16 \tag{9.19}$$
$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{100/3 + 16/3} = 6.21 \tag{9.20}$$

In our case it makes sense to use a one tailed test, since we are trying to reject the hypothesis that the aspirin works better or equal to the new drug. Thus the null hypothesis tells us that $E(\bar{X}_2 - \bar{X}_1 \mid H_n \text{ true}) \leq 0$. The extreme case, i.e., what we call H_x says $E(\bar{X}_1 - \bar{X}_2 \mid H_x \text{ true}) = 0$. Thus,

$$T = \frac{\bar{X}_2 - \bar{X}_1 - E(\bar{X}_2 - \bar{X}_1 \mid H_x \text{ true})}{S_{\bar{X}_2 - \bar{X}_1}} = \frac{(60 - 40) - 0}{6.21} = 3.22 \tag{9.21}$$

The number of degrees of freedom is $(2)(3-1) = 4$. We go to tables and find that for 4 degrees of freedom $P(T > 2.13 \mid H_x \text{ true}) = 0.05$. Thus the critical value is 2.13. Since the value taken by T is larger than 2.13 we have enough evidence to reject the null hypothesis. We can say that the new drug works better than aspirin.

9.1.1 Within Subjects Experiments

In the previous experiments we assumed all the measurements were independent. In within subjects experiments this assumption is clearly incorrect: if we test the same subjects in the two groups, then the random variables in the first group will be related to the random variables in the second group. This type of experiments are called "repeated measurements" or within subjects. In this case we work with the differences between the observations obtained in the two groups on a subject by subject basis

$$D_j = X_{2,j} - X_{1,j} \tag{9.22}$$

and the sample average of the difference would be

$$D = \frac{1}{n} \sum_{j=1}^{n} D_j = \bar{X}_2 - \bar{X}_1 \tag{9.23}$$

In most cases the null hypothesis says that there is no differences in the expected values of the two groups, i.e., $E(\bar{D} \mid H_n \text{ true}) = 0$. The T random variable follows

$$T = \frac{\bar{D} - E(\bar{D} \mid H_n \text{ true})}{S_{\bar{D}}} \tag{9.24}$$

This effectively transforms a 2 group experiment into a single group experiment that can be analyzed with the techniques we saw in the previous chapter.

Example *Three undergraduate students from UCSD volunteered to participate on an experiment to test the effect of caffeine on reaction time. Each student was tested in two different conditions with the tests being separated in time by one week. In one condition the subjects drunk a double espresso. In the other conditions the subjects drunk a double decaffeinated espresso. The order of the two conditions was randomly assigned. On each condition, the subjects were tested 15 minutes after drinking the espresso. Subjects were tested on a task that required rapid reaction to a stimulus. The average reaction time in milliseconds of each subject was recorded. The results were as follows*
Caffeine Condition =$[10, 30, 20]$
Decaf Condition = $[56, 64, 60]$
The first number on each condition corresponds to the first subject, second number to second subject and third number to the third subject. Is the difference in reaction time between the two conditions statistically significant?

This is a repeated measures experiment so the T-test would be based on the difference measurements obtained for each of the three subjects.

$$D_1 = 10 - 56 = -46 \tag{9.25}$$
$$D_2 = 30 - 64 = -34 \tag{9.26}$$
$$D_3 = 20 - 60 = -40 \tag{9.27}$$
$$\bar{D} = (-46 - 34 - 40)/3 = -40 \tag{9.28}$$
$$S_D^2 = \frac{1}{2}((6^2) + (6^2) + 0^2) = 18 \tag{9.29}$$
$$S_{\bar{D}}^2 = 18/3 = 6 \tag{9.30}$$
$$T = \frac{18 - 0}{\sqrt{6}} = 7.35 \tag{9.31}$$
$$df = (3 - 1)8 = 2 \tag{9.32}$$
$$P(T \mid H_x > 2.91 \text{ True}) = 1/20 \text{ Obtained using T-tables} \tag{9.33}$$

Thus the critical value is 2.91. Since 7.35 is larger than 2.91. The difference in reaction time between the two groups is statistically significant.

9.2 Exercises

1. *A San Diego based biotek company is testing whether a new drug has a good effect on amnesia, the inability to remember recently presented materials. The sample consisted of 8 amnesic patients that volunteered to participate in the experiment. The subjects were randomly assigned to either the experimental group (drug) or the control group (placebo). Subjects were treated with 125 mg a day of the drug or the placebo for 30 days. After one month subjects were given a list of 100 training pictures one at a time at 2 second intervals. Half an hour later subjects were presented 200 new pictures: the 100 training pictures plus 100 new ones. Subjects were asked to discriminate whether the pictures were old or new. The dependent variable was the percentage of correctly discriminated pictures. The results in the experimental group were as follows:* $\{61, 60, 59, 60\}$. *The results in the control group were as follows:* $\{50, 51, 50, 49\}$.

   (a) Is the experiment randomized or within subjects?

   (b) Specify the null hypothesis if you were to use a two tailed test.

   (c) Do you have enough evidence to reject this null hypothesis?

   (d) Specify the null hypothesis if you were to use a one tailed test.

   (e) Do you have enough evidence to reject this null hypothesis?

   (f) What assumptions did you make?

   (g) A replication of the same experiment with a different random sample of subjects results in a control group mean of 10 instead of 50? Do you find this surprising? Justify your response.

   (h) Repeat the analysis assuming the same subjects are used in both groups. For example, Subject 1 gets a score of 61 when he/she takes a drug and a score of 50 when he/she takes the placebo.

Chapter 10

Factorial Experiments

NOTE FROM JAVIER R. Movellan: This chapter may be particularly buggy. Use at your own risk.

10.1 Experiments with more than 2 groups

In this chapter we will apply the T-test to experiments that involve more than two treatment groups. The basic ideas are the same as in single group and two group experiments. The only difference is that we use information from all the groups to get better estimates of the variance. The experimental design we will address is called "between subjects" indicating that each of the treatment groups is made of different randomly assigned subjects.

Suppose we want to study the joint effects of caffeine and alcohol on the time it takes people to respond to a red light. Eight volunteer students from the Cognitive Science Department at UCSD were randomly assigned to one of four treatment groups: Group 1) Each subject in this group has a non-alcoholic beer and a decaffeinated coffee. Group 2) Each subject in this group has an alcoholic beer and a decaffeinated coffee. Group 3) Each subject in this group has a non-alcoholic beer and a regular coffee. Group 4) Each subject in this group has an alcoholic beer and a regular coffee. Each of the 8 subjects in our sample was measured in terms of the times it took them to do a test that involved pushing a pedal in response to a red light. Table 11 shows the results of the experiment

Suppose we want to know whether alcohol has an effect on people that take regular coffee. In this cases the null hypothesis would be that the alcohol has no effect, i.e.,

$$E(\bar{X}_4 - \bar{X}_3 \mid H_n \text{ true}) = 0 \tag{10.1}$$

where $\bar{X}_i$ represents the sample mean of group i in the experiment. We could test this hypothesis restricting our analysis to groups 3 and 4. This would simply be a two group experiment that we already know how to analyze. However there is a better way to do things. The idea is to use a pooled estimate of the variance based on the 4 groups, instead of just two groups. The advantage of such an estimate is that it is based on more observations. If our experiment has a different groups, the pooled estimate would be

$$S^2 = \frac{1}{a}(S_1^2 + S_2^2 + S_3^2 + \cdots + S_a^2) \tag{10.2}$$

and it would have $(a)(n-1)$ degrees of freedom. We know

$$\text{Var}(\bar{X}_4 - \bar{X}_3) = 2\text{Var}(\bar{X}_1) = \frac{2}{n}\text{Var}(X_{1,1}) \tag{10.3}$$

thus our estimate of this variance will be

$$S^2_{\bar{X}_4 - \bar{X}_3} = \frac{2}{n}S^2 \tag{10.4}$$

and the T random variable would be

$$T = \frac{\bar{X}_4 - \bar{X}_3}{\sqrt{(2/n)S^2}} \tag{10.5}$$

Table 10.1:

	$NoAlcohol$		$Alcohol$	
$NoCaffeine$	3	1	7	5
$Caffeine$	1	1	10	16

In our example

$$\bar{X}_1 = 2 \tag{10.6}$$
$$\bar{X}_2 = 6 \tag{10.7}$$
$$\bar{X}_3 = 1 \tag{10.8}$$
$$\bar{X}_4 = 13 \tag{10.9}$$
$$S_1^2 = \frac{1}{2-1}((3-1)^2 + (1-1)^2) = 2 \tag{10.10}$$
$$S_2^2 = \frac{1}{2-1}((7-6)^2 + (5-6)^2) = 2 \tag{10.11}$$
$$S_3^2 = \frac{1}{2-1}((1-1)^2 + (1-1)^2) = 0 \tag{10.12}$$
$$S_4^2 = \frac{1}{2-1}((10-13)^2 + (13-16)^2) = 18 \tag{10.13}$$
$$S^2 = \frac{1}{4}(2+2+0+18) = \frac{11}{2} \tag{10.14}$$
$$T = \frac{13-1-0}{\sqrt{\frac{2}{2}\frac{11}{2}}} = 5.11 \tag{10.15}$$
$$df = 4; \tag{10.16}$$

we see in tables that

$$P(T > 2.77 \mid H_n \text{ true}) = 1/20 \tag{10.17}$$

10.2 Interaction Effects

In many cases it is important to know whether the effect of a treatment changes when administered in combination with other treatments. For example in many cases we need to know whether a drug has the same effect when combined with other drugs. This concept is so important that statisticians gave it a special name: "interaction effects". In general, if the effect of a treatment changes when combined with other treatments, we say that **the treatments interact**, or that the effects are **non-additive**. Otherwise we say that **the treatments are additive**, or that there is no interactions.

The simplest experimental design that allows to study whether treatments interact consists of 4 groups formed by the combination of two treatments of one

factor (e.g., caffeine versus no caffeine) and two treatments of another factor (e.g., alcohol versus no alcohol). This type of design is called a 2×2 factorial design.

Now we need to translate the concept of interaction into a specific testable hypothesis. Let's take the example experiment we have analyzed so far. If there is no interaction between alcohol and caffeine, then the effect of alcohol should be the same regardless of whether a person takes regular or decaffeinated coffee. We measure the effect of alcohol when combined with decaffeinated coffee as $E(\bar{X}_2 - \bar{X}_1)$ and the effect of alcohol when combined with regular coffee as $E(\bar{X}_4 - \bar{X}_3)$. Thus, if there is no interactions, these two effects have to be the same. In other words, The null hypothesis of no interaction effects says that

$$E[(\bar{X}_2 - \bar{X}_1) - (\bar{X}_4 - \bar{X}_3) \mid H_n \text{ true}] = 0 \tag{10.18}$$

or equivalently

$$E[\bar{X}_2 + \bar{X}_3 - \bar{X}_1 - \bar{X}_4 \mid H_n \text{ true}] = 0 \tag{10.19}$$

So the random variable of interest is $\bar{X}_2 + \bar{X}_3 - \bar{X}_1 - \bar{X}_4$. Since we know that

$$\text{Var}(\bar{X}_2 + \bar{X}_3 - \bar{X}_1 - \bar{X}_4) = 4\text{Var}(\bar{X}_1) = \frac{4}{n}\text{Var}(X_{1,1}) \tag{10.20}$$

then our estimate of the standard deviation of $\bar{X}_2 + \bar{X}_3) - \bar{X}_1 - \bar{X}_4$ is

$$S^2_{\bar{X}_2+\bar{X}_3-\bar{X}_1-\bar{X}_4} = \frac{4}{n}S^2 \tag{10.21}$$

and it has $(a)(n-1)$ degrees of freedom. The T random variable follows

$$T = \frac{\bar{X}_2 + \bar{X}_3 - \bar{X}_1 - \bar{X}_4 - E(\bar{X}_1 - \bar{X}_4 \mid H_n \text{ true})}{S_{\bar{X}_2-\bar{X}_3-\bar{X}_1-\bar{X}_4}} \tag{10.22}$$

In our example,

$$T = \frac{6 + 1 - 2 - 13}{\sqrt{(4/2)(11/2)}} = 2.41 \tag{10.23}$$

$$2df = (4)(2-1) = 4 \tag{10.24}$$

For 4 df the critical value is 2.71 and thus we do not have enough evidence to reject the null hypothesis. We cannot say that alcohol and caffeine interact.

Chapter 11

Confidence Intervals

NOTE FROM JAVIER: This chapter has not been properly debugged. Use at your own risk.

In this chapter we study a method to bracket expected values based on sample information. We may want to bracket expected values of the mean of a group , or the difference between sample means of two groups, or the difference of the difference between sample means of 4 groups (we'd do so to analyze interactions). In all cases the basic ideas are the same. While hypothesis testing and interval estimation are closely related, the later method is in general more informative for the following two reasons: 1) Classical hypothesis testing does not tell us the probability that our decisions be correct. Classical interval estimation does tell us about this. 2) Classical interval estimation methods are easier to relate to Bayesian methods so they provide a better foundation for a way of analyzing experiments consistent with classical and Bayesian approaches.

We will see an example of interval estimation using the data from an experiment we used in a previous chapter. Here is a re cap of the experiment at hand:

Suppose we want to study the joint effects of caffeine and alcohol on the time it takes people to respond to a red light. Eight volunteer students from the Cognitive Science Department at UCSD were randomly assigned to one of four treatment groups: Group 1) Each subject in this group has a non-alcoholic beer and a decaffeinated coffee. Group 2) Each subject in this group has an alcoholic beer and a decaffeinated coffee. Group 3) Each subject in this group has a non-alcoholic beer and a regular coffee. Group 4) Each subject in this group has an alcoholic beer and a regular coffee. Each of the 8 subjects in our sample was measured in terms of the times it took them to do a test that involved pushing a pedal in response to a red light. Table 11 shows the results of the experiment

Suppose we are interested in bracketing the value of $E(\bar{X}_3)$. To do so consider the following T random variable

$$T = \frac{\bar{X}_3 - E(\bar{X}_3)}{S_{\bar{X}_3}} \tag{11.1}$$

Note that now we do not condition on the null hypothesis being true. We simply treat $E(\bar{X}_3)$ as a fixed, yet unknown number. Thus, since we are using the correct expected value, T will follow Gosset's t-distribution. In previous chapters we saw that for this particular experiment,

$$\bar{X}_3 = 1 \tag{11.2}$$
$$S^2 = 11/2 \tag{11.3}$$
$$df = 4 \tag{11.4}$$
$$c = 2.77 \tag{11.5}$$

where c was obtained using T-tables so that $P(T \geq c) = 1/40$. Moreover. since $\bar{X}_3$ is based on two observations

$$S_{\bar{X}_3} = \sqrt{S^2/2} = \sqrt{(1/2)(11/2)} = 1.65 \tag{11.6}$$

Now notice the following,

$$P(-c \leq T \leq c) = 0.95 \tag{11.7}$$
$$= P(-c \leq \frac{\bar{X}_3 - E(\bar{X}_3)}{S_{\bar{X}_3}} \leq c) \tag{11.8}$$
$$= P(-cS_{\bar{X}_3} \leq \bar{X}_3 - E(\bar{X}_3) \leq cS_{\bar{X}_3}) \tag{11.9}$$
$$= P(-\bar{X}_3 - cS_{\bar{X}_3} \leq -E(\bar{X}_3) \leq -\bar{X}_3 + cS_{\bar{X}_3}) \tag{11.10}$$
$$= P(\bar{X}_3 + cS_{\bar{X}_3} \geq E(\bar{X}_3) \geq c\bar{X}_3 - cS_{\bar{X}_3}) \tag{11.11}$$
$$\tag{11.12}$$

Table 11.1:

	$NoAlcohol$		$Alcohol$	
$NoCaffeine$	3	1	7	5
$Caffeine$	1	1	10	16

Now lets define the random variables L and U to represent the upper and lower limits of our intervals,

$$L = \bar{X}_3 - cS_{\bar{X}_3} \tag{11.13}$$
$$U = \bar{X}_3 + cS_{\bar{X}_3} \tag{11.14}$$
$$\tag{11.15}$$

We have learned that

$$P(E(\bar{X}_3) \in [L, U]) = 0.95 \tag{11.16}$$

In other words, there is a 95 % chance that the random interval $[L, U]$, will cover the true value of $E(\bar{X}_3)$. Note that the interval $[L, U]$ is random, i.e., it changes with the outcome of the experiment, and $E(\bar{X}_3)$ is not random, i.e., it is a fixed yet unknown number.

Example In our experiment,

$$L = \bar{X}_3 - cS_{\bar{X}_3} = 1 - (2.77)(1.65) = 3.57 \tag{11.17}$$
$$U = \bar{X}_3 - cS_{\bar{X}_3} = 1 + (2.77)(1.65) = 5.57 \tag{11.18}$$

In our particular experiment, the confidence interval for $E(\bar{X}_3)$, the expected value of the mean of the third group, is $[3.57, 5.57]$ we do not know whether in this particular experiment the confidence interval brackets the expected value, however we know that in 19/20 experiments, the procedure we used will be correct.

Appendix A

Useful Mathematical Facts

1. **Symbols**

   (a) $\triangleq$ "Is defined as"

   (b) $n!$ "n-factorial"

$$n! = \begin{cases} 1 & \text{if } n = 0 \\ (1)(2)(3)\cdots(n) & \text{if } n \neq 0 \end{cases} \tag{A.1}$$

   (c) Sterling's approximation

$$n! \approx \sqrt{2\pi n}(\frac{n}{e})^n \tag{A.2}$$

   (d) $e = 2.718281828\cdots$, the natural number

   (e) $\ln(x) = \log_e(x)$ The natural logarithm

   (f) 1_A The indicator (or characteristic) function of the set A. It tells us whether or not an element belongs to a set. It is defined as follows, $1_A : \Omega \to \{0, 1\}$.

$$1_A(\omega) = \begin{cases} 1 & \text{for all } \omega \in A \cap \Omega \\ 0 & \text{for all } \omega \in A^c \cap \Omega \end{cases} \tag{A.3}$$

   Another common symbol for the indicator function of the set A is ξ_A

2. **The Greek alphabet**

A	α	alpha	I	ι	iota	P	ρ	rho
B	β	beta	K	κ	kappa	Σ	σ	sigma
Γ	γ	gamma	Λ	λ	lambda	T	τ	tau
Δ	δ	delta	M	μ	mu	Υ	υ	upsilon
E	ϵ	epsilon	N	ν	nu	Φ	ϕ	phi
Z	ζ	zeta	Ξ	ξ	xi	X	χ	chi
H	η	eta	O	o	omicron	Ψ	ψ	psi
Θ	θ	theta	Π	π	pi	Ω	ω	omega

3. **Series**

$$1 + 2 + 3 + \cdots + n = \frac{(n)(n+1)}{2} \tag{A.4}$$

$$a^0 + a^1 + a^2 + \cdots + a^{n-1} = \frac{1 - a^n}{1 - a} \tag{A.5}$$

$$1 + a + a^2 + \cdots = \frac{1}{1 - a}, \text{for} |a| < 1 \tag{A.6}$$

$$a + 2a^2 + 3a^3 + \cdots = \frac{a}{(1 - a)^2}, \text{for} 0 < a < 1 \tag{A.7}$$

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \tag{A.8}$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \tag{A.9}$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \tag{A.10}$$

$$e^{jx} = \cos(x) + j \sin(x) \text{ where } j \triangleq \sqrt{-1} \tag{A.11}$$

4. **Binomial Theorem**

$$(a+b)^n = \sum_{m=0}^{n} \binom{n}{m} a^{n-m}\ b^m \tag{A.12}$$

where

$$\binom{n}{m} \triangleq \frac{n!}{(m!)\ (n-m)!} \tag{A.13}$$

Note from the binomial theorem it follows that

$$2^n = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} \tag{A.14}$$

5. **Exponentials**

$$a^0 = 1 \tag{A.15}$$

$$a^{m+n} = a^m\ a^n \tag{A.16}$$

$$a^{-n} = \frac{1}{a^n} \tag{A.17}$$

$$(ab)^n = a^n b^n \tag{A.18}$$

6. **Logarithms**

$$a^{(\log_a(x))} = x \tag{A.19}$$

$$\log_a(x\ y) = \log_a(x) + \log_a(y) \tag{A.20}$$

$$\log_a(x^y) = y \log_a(x) \tag{A.21}$$

$$\log_a(1) = 0 \tag{A.22}$$

$$\log_a(a) = 1 \tag{A.23}$$

$$\log_a(x) = (\log_b(x))/\log_b(a) \tag{A.24}$$

7. **Quadratic formula**
   The roots of the equation $ax^2 + bx + c = 0$ are

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{A.25}$$

8. **Factorizations**

$$a^2 - b^2 = (a - b)(a + b) \tag{A.26}$$

9. **Trigonometry**

$$\tan(\alpha) = \frac{\sin(\alpha)}{\cos(\alpha)} \tag{A.27}$$

$$\sin^2(\alpha) + \cos^2(\alpha) = 1 \tag{A.28}$$

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta) \tag{A.29}$$

$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta) \tag{A.30}$$

$$\tan(\alpha + \beta) = \frac{\tan(x) + \tan(y)}{1 + \tan(x)\tan(y)} \tag{A.31}$$

10. **Hyperbolics**

$$\sinh(x) = \frac{e^x - e^{-x}}{2} \tag{A.32}$$

$$\cosh(x) = \frac{e^x + e^{-x}}{2} \tag{A.33}$$

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} \tag{A.34}$$

11. **Complex Numbers**

We use the convention $j \triangleq \sqrt{-1}$. There are three ways to represent a complex number

(a) Cartesian representation

$$x = (x_r, x_i) = x_r + jx_i \tag{A.35}$$

where x_r and x_i are the real and imaginary components of x.

(b) Polar representation:

$$|x| \triangleq \sqrt{x_r^2 + x_i^2} \tag{A.36}$$

is called the magnitude of x.

$$\angle x \triangleq \text{arc tan} \frac{x_i}{x_r} \tag{A.37}$$

is called the phase of x.

(c) Exponential representation

$$x = |x|e^{j\angle x} = |x|(\cos(\angle x), \sin(\angle x)) \tag{A.38}$$

Operation on complex numbers:

(a) Addition/Substraction:

$$(x_r, x_i) + (y_r, y_i) = (x_r + y_r, x_i + y_i) \tag{A.39}$$

(b) Multiplication

$$|(xy)| = |x||y| \tag{A.40}$$

$$\angle(xy) = \angle x + \angle y \tag{A.41}$$

(c) Conjugation

The complex conjuage of $x = (x_r, x_i)$ is $\tilde{x} = (x_r, -x_i)$. Note $|x| = |\tilde{x}|$ and $\angle(\tilde{x}) = -\angle(x)$.

(d) Inner Product

Let $x = (x_1, \cdots, x_n)$ and $y = (y_1, \ldots, y_n)$ be complex vectors (i.e., each component of x and of y is a complex number). The inner prouct of x and y is defined as follows

$$< x, y >= x \cdot y = \sum_{i=1}^{n} x_i \tilde{y}_i \tag{A.42}$$

12. **Derivatives**

Let $y = f(x)$

$$\frac{dy}{dx} = \lim_{\Delta_x \to 0} (f(x + \Delta x) - f(x))/\Delta x \tag{A.43}$$

Here are some alternative representations of the derivative

$$f'(x) = y' = \frac{dy}{dx} = \frac{d}{dx}y = \frac{df(x)}{dx} = \frac{d}{dx}f(x) = \frac{df(u)}{du}|_{u=x} \tag{A.44}$$

- Exponential:

$$\frac{d}{dx}\exp(x) = \exp(x) \tag{A.45}$$

- Polynomial:

$$\frac{d}{dx}x^m = mx^{m-1} \tag{A.46}$$

- Logarithm

$$\frac{d}{dx}\ln x = \frac{1}{x} \tag{A.47}$$

- Sine

$$\frac{d}{dx}\sin x = \cos x \tag{A.48}$$

- Cosine

$$\frac{d}{dx}\cos x = -\, sinx \tag{A.49}$$

- Linear combinations

$$\frac{d}{dx}((a)f(x) + (b)g(x)) = (a)\frac{d}{dx}f(x) + (b)\frac{d}{dx}g(x) \tag{A.50}$$

- Products

$$\frac{df(x)g(x)}{dx} = f(x)\frac{dg(x)}{dx} + g(x)\frac{df(x)}{dx} \tag{A.51}$$

- Chain Rule
  Let $y = f(x)$ and $z = g(y)$

$$\frac{dy}{dx} = \frac{dz}{dy}\frac{dy}{dx} \tag{A.52}$$

  You can think of the chain rule in the following way: x changes y which changes z. How much z changes when x changes is the product of how much y changes when x changes times how much z changes when y changes. Here is a simple example that uses the chain rule

$$\frac{d\exp(ax)}{dx} = \frac{d\exp(ax)}{dax}\frac{ax}{dx} = \exp(ax)(a) \tag{A.53}$$

13. **Indefinite Integrals**

The **indefinite integral** of the function f is a function whose derivative is f (i.e., the antiderivative of is f). This function is unique up to addition of arbitrary constant. The expression

$$\int f(x)dx = F(x) + C \tag{A.54}$$

means that $F'(x) = f(x)$. The C reminds us that the derivative of $F(x)$ plus any arbitrary constant is also $f(x)$.

- Linear Combinations

$$\int af(x) + bg(x)dx = a\int f(x)dx + b\int g(x)dx \tag{A.55}$$

- Polynomials

$$\int x^m dx = \frac{x^{m+1}}{m+1} + C \tag{A.56}$$

- Exponentials

$$\int \exp(x)dx = \exp(x) + C \tag{A.57}$$

- Logarithms

$$\int \frac{1}{x}dx = \ln(x) + C \tag{A.58}$$

- Integration by parts

$$\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx \tag{A.59}$$

  The formula for integration by parts easily follows from the formula for the derivative of the product of $f(x)g(x)$.

History

- The first version of this document was written by Javier R. Movellan in 1994. The document was 6 pages long.
- The document was made open source under the GNU Free Documentation License Version 1.1 on August 9 2002, as part of the Kolmogorov project.

Appendix B

Set Theory

Intuitively we think of sets as collections of elements. The crucial part of this intuitive concept is that we are willing to treat sets as entities distinguishable from their elements. Sometimes we identify sets by enumeration of their elements. For example, we may talk about the set whose elements are the numbers 1, 2 and 3. In mathematics such sets are commonly represented by embracing the elements of the set using curly brackets. For example, the set $\{1, 2, 3\}$ is the set whose elements are the numbers 1, 2 and 3. Sometimes sets are defined using some property that identifies their elements. In such case it is customary to represent the sets using the following formula

$$\{x : x \text{ has a particular property}\} \tag{B.1}$$

For example, the set $\{1, 2, 3\}$ can be represented as

$$\{x : x \text{ is a natural number smaller than } 4\}. \tag{B.2}$$

The intuitive concept of sets as collections of elements is useful but it can only take us so far. You may complain that we have not really defined what a set is since we have not defined collections and we have not specified what qualifies as an element. We have not specified either what qualifies as a property. Consider for example the proposition $\{x : x \text{ is not an element of } x\}$, i.e., the set of sets which are not elements of themselves. We can prove by contradiction that such a set does not exist. Let's assume that this set exists and lets represent it with the symbol y. If y is an element of y then, since all the elements of y are not an element of themselves it follows that y is not an element of y. Moreover, if y is not an element of y then, y must be an element of y. In other words, if we assume

the set y exists we get a contradiction. Therefore we have to conclude that y does not exist. Using similar reasoning one can also show that the set of all sets does not exist either (see proof later in this document). But this raises deep questions:

1. What does it mean to say that a set exists or does not exist? For example Leopold Kronecker, a German mathematician born in 1823, claimed that the only numbers that assuredly exist are the natural numbers (1,2,3 ...). According to him the set of real numbers are just a fantasy that does not exist. But think about it, what criteria, other than authority, can we use to decide whether the natural numbers or the real numbers exist?

2. How can we tell whether something is or is not a set?

3. What are valid elements of a set?

Axiomatic set theory was developed to provide answers to such questions. In axiomatic set theory:

1. A set exists if the proposition that asserts its existence is logically true. Moreover within this theory there are only sets so if a formal object is not a set, it does not exist.

2. If the assumption that an object exists leads to a contradiction we can assert that that object does not exist, or equivalently, that it is not a set.

3. There are no atomic elements: An object exists if and only if it is a set. Of course sets can have elements but those elements must be sets themselves otherwise they would not exist.

One "annoying" aspect of axiomatic set theory is that sets become a logical abstraction detached from our everyday experience with collections of physical objects. You should think of mathematical sets as logical "objects" which are part of a formal structure. Within this theory to say that an object exists is the same as saying that it is a set. To say that an object does not exist is the same as saying that it is not a set. Everyday collection of physical objects are no longer sets in this theory and thus they do not exist. While this approach may strike you as peculiar, it turns out to be extremely powerful and in fact it has become the foundation of mathematics. The formal structure of set theory while independent from the physical world provides very useful tools to model the world itself. The key is to develop set structures constrained in ways that mirror essential properties of the

physical world. For example, the properties of the set of natural numbers (i.e., 1,2,3, . . .) mirrors our experience counting collections of physical objects.

Axiomatic set theory is a first order logical structure. First order logic works with propositions, i.e., logical statements constructed according to the rules of logic and that can take two values. For convenience we call these two values "True" and "False". Set theory, and thus the entire body of mathematics reduces to logical propositions that use the following elements:

1. Variables (e.g., $a, b, \ldots x, y, z$) which stand for sets.

2. The predicate $\in$, which stands for element inclusion. For example, if the proposition $(x \in y)$ takes the value true, we know that both x and y are sets and that x is an element of y. For example, the proposition

$$\{1, 2, 3\} \in \{\{1, 2\}, \{4, 5\}, \{1, 2, 3\}\} \tag{B.3}$$

   takes the value "True".

3. Logical operators

   (a) $\neg P$, where $\neg$ is the logical "negation" operator.

   (b) $P \wedge P$, where $\wedge$ is the logical "and" operator.

   (c) $P \vee P$, where $\vee$ is the logical "or" operator.

   (d) $P \rightarrow P$, where $\rightarrow$ is the logical "implication" operator.

   (e) $P \leftrightarrow P$, where $\leftrightarrow$ is the logical "bijection" operator.

   (f) $\forall x P$ is the logical "for-all" quantifier.

   (g) $\exists x P$ is the logical "exists" quantifier.

The names of the different operators (i.e., "negation", "and", "or", "implication" ...) are selected for convenience. We could have given them completely different names, all we really need to know is how they operate on propositions.

All propositions in set theory are built out of atomic propositions of the form $(x \in y)$ connected using the logical operators. If P and Q are propositions, e.g., P could be $(x \in y)$ and Q could be $(y \in z\ )$ then $\neg P$, $P \wedge P$, $P \vee Q$ $P \rightarrow Q$, $P \leftrightarrow Q$, $\forall x P$ and $\exists x P$ are also propositions.

The effect of the connectives on the truth value of propositions is expressed in Table B . Thus, if the proposition P takes value "True" and the proposition Q takes the value "False" then the proposition $(P \wedge Q)$ takes the value "False". The

P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$	$P \rightarrow Q$	$P \leftrightarrow Q$
T	T	F	T	T	T	T
T	F	F	F	T	F	F
F	T	T	F	T	T	F
F	F	T	F	F	T	T

Table B.1: The truth tables of the logical operators. T stands for "True" and F for "False".

propositions $\forall x P$ and $\exists x P$ tell us that x is a variable that can take as value any formal object that qualifies as a set. It also tells us that P is a proposition whose truth value depends on x. For example, P could be $(x \in y) \vee (x \in z)$, where y and z are fixed sets and x acts as a variable. The proposition $\forall x P$ takes the value "True" if P takes the value "True" for all sets. The proposition $\exists x P$ takes the value "True" if there is at least one set for which P takes the value "True". Remember when we say for all sets we do not mean sets of physical objects. In fact we still have to define what we mean by set.

B.1 Proofs and Logical Truth

Sometimes we treat propositions as formulas whose truth value depends on the truth values taken by variables in the proposition. For example if P and Q are propositional variables then the $P \wedge Q$ is a propositional formula whose truth value depends on the specific truth values taken by P and Q. We say that a propositional formula is **logically true** if for all valid combinations of truth values of the elements in the formula, the formula can only take the value "True". For example for the formula $(P \vee \neg P)$ there is only two valid combination of truth values for P and $\neg P$: "True, False" and "False, True". In both case the formula $(P \vee (\neg P))$ takes the value "True" and thus we say that it is logically true. Similarly if a propositional formula can only take "False" values we say that it is **logically false**. For example $(P \wedge (\neg P))$ is logically false. A **proof** is a process that shows a propositional formula is logically true.

B.2 The Axioms of Set Theory

To simplify the presentation of axiomatic set theory I will use "pseudocode", i.e., a combination of logical propositions, mathematical symbols, and English statements. I do so under the understanding that all these statements can be written as pure logical propositions.

I will use the symbol $\notin$ in propositions of the form $(x \notin y)$ as an alternative notation to $\neg(x \in y)$. I will use the formula

$$\exists\{x : P\} \tag{B.4}$$

as an alternative notation to the propositional formula

$$\exists y \forall x P \tag{B.5}$$

This formula simply says that there is a set of elements that satisfy the proposition P. If the formula takes the value "True" then the symbols $\{x : P\}$ refers to a set that make the proposition $\forall x P$ "True". When a set x makes the proposition P true, I will say that x satisfies P. For example the set 1 satisfies the propositional formula $(x \in \{1, 2\})$.

In set theory all existing objects are sets. If an object exists it is a set otherwise it does not exist. To remind us of the fact that sets include elements we sometimes refer to sets as a collection of sets, or as a families of sets. This is just a "human factors" trick since the theory makes no distinction between sets, families, collections or elements. In axiomatic set theory elements, collections, and families are just sets.

Axiomatic set theory is commonly presented using 9 redundant axioms, which are the foundation of all mathematical statements.

B.2.1 Axiom of Existence:

An axiom is a restriction in the truth value of a proposition. The axiom of existence forces the proposition

$$\exists y \forall x (x \notin y) \tag{B.6}$$

to take the value "True". We call call the sets that satisfy $\forall x(x \notin y)$ **empty sets**. Thus the axiom of existence tells us that there is at least one empty set, we will see later that in fact there is only one empty set.

B.2.2 Axiom of Equality:

This axiom is also called the axiom of extensionality and it defines the predicate "$=$". For mnemonic convenience when the proposition $(x = y)$ takes the value "True" we say that the sets x and y are equal. In order to define how the symbol "$=$" works it is convenient to create a new predicate, which we will symbolize as $\subset$. The new predicate works as follows: For all sets u and v if the proposition

$$\forall x(x \in u) \rightarrow (x \in v) \tag{B.7}$$

is true then the proposition $(u \subset v)$ is true. For mnemonic convenience if the proposition $(u \subset v)$ takes the value "True" we say that u is a subset of v.

The axiom of equality says that if the proposition $(u \subset v) \wedge (v \subset u)$ is true then the proposition $(u = v)$ is true. In other word, the proposition

$$\forall u(u \in x \leftrightarrow u \in y) \rightarrow (x = y) \tag{B.8}$$

takes the value "True". The formula $(x \neq y)$ is used as an alternative notation to $\neg(x = y)$. We will now use the axiom of equality to prove that there is only one empty set.

Theorem: The empty set is unique.

Proof: Let x and y be empty sets, then $u \in y$ and $u \in x$ are always false for all sets u. Thus $(u \in y \leftrightarrow u \in x)$ is true for all sets u and since by the axiom of equality

$$(\forall u(u \in x \leftrightarrow u \in y)) \rightarrow (x = y) \tag{B.9}$$

is true then it follows that $(x = y)$ must be true. Hereafter we identify the empty set with the symbols $\varnothing$ or alternatively with the symbol $\{\}$.

□

B.2.3 Axiom of Pair:

So far set theory has only given us one set: the empty set. The axiom of pair brings new sets to life. The axiom of pair says that if x and y exist (i.e., if they are sets) there also exists a set whose only elements are x and y. We will represent such set as $\{x, y\}$. The axiom of pair forces the proposition

$$\forall x \forall y \exists \{x, y\} \tag{B.10}$$

to take the value "True". The set made out of the sets a and a is symbolized as $\{a, a\}$ or $\{a\}$ and is called the singleton whose only element is a. So starting with the empty set $\varnothing$, it follows that the set $\{\varnothing\}$ exists. Note that $\varnothing$ and $\{\varnothing\}$ are different since the first has no element and the second has one element, which is the empty set.

Ordered pairs: The ordered pair of the sets x and y is symbolized (x, y) and defined as follows

$$(x, y) \triangleq \{\{x\}, \{x, y\}\} \tag{B.11}$$

where $\triangleq$ stands for "equal by definition".

Exercise: Prove that two ordered pairs (a, b) and (c, d) are equal if and only if $a = b$ and $c = d$.

Ordered sequences: Let $x_1, \ldots, x_n$ be sets. The ordered sequence $(x_1, \ldots, x_n)$ is recursively defined as follows

$$(x_1, \cdots, x_n) = ((x_1, \cdots, x_{n-1}), x_n) \tag{B.12}$$

Exercise: Prove that two n-tuples pairs $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ are equal if and only if $a_1 = b_1$ and $a_2 = b_2$ and ... $a_n = b_n$.

B.2.4 Axiom of Separation:

This axiom tells us how to generate new sets out of elements of an existing set. To do so we just choose elements of an existing set that satisfy a proposition. Consider a proposition P whose truth value depends on the sets u and v, for example, P could be $(u \in v)$. The axiom of separation forces the proposition

$$\exists\{x : (x \in u) \wedge P\} \tag{B.13}$$

to take the value "True" for all sets u, v and for all propositions P with truth values dependent on u and v.

Fact: There is no set of all sets.
The proof works by contradiction. Assume there is a set of all sets, and call it u. Then by the axiom of separation the following set r must exist

$$r = \{x : (x \in u) \wedge (x \notin x)\} \tag{B.14}$$

and since $(x \in u)$ is always true, this set equals the set

$$\{x : x \notin x\} \tag{B.15}$$

Then $(r \in) r \leftrightarrow (r \notin r)$ which is a logically false proposition. Thus the set of all sets does not exist (i.e., it is not a set).

Intersections: The intersection of all the sets in the set s, or simply the intersection of s is symbolized as $\cap s$ and defined as follows:

$$\cap s = \{x : \forall y(\ (y \in s) \rightarrow (x \in y)\ )\} \tag{B.16}$$

For example, if $s = \{\{1,2,3\},\{2,3,4\}\}$ then $\cap s = \{2,3\}$. The axiom of separation tells us that if s exists then $\cap s$ also exists. We can then use the axiom of equality to prove that $\cap s$ is in fact unique. For any two sets x and y, we represent their intersection as $x \cap y$ and define it as follows

$$x \cap y \triangleq \cap \{x, y\} \tag{B.17}$$

For example, if $x = \{1,2,3\}$ and $y = \{2,3,4\}$ then

$$x \cap y = \cap\{\{1,2,3\},\{2,3,4\}\} = \{2,3\} \tag{B.18}$$

B.2.5 Axiom of Union:

It tells us that for any set x we can make a new set whose elements belong to at least one of the elements of x. We call this new set the union of x and we represent it as $\cup x$. For example, if $x = \{\{1,2\},\{2,3,4\}\}$ then $\cup x = \{1,2,3,4\}$. More formally, the axiom of union forces the proposition

$$\forall s \exists \cup s \tag{B.19}$$

to be true. Here $\cup s$ is defined as follows

$$\cup s \triangleq \{x : \exists y (y \in s) \wedge (x \in y)\} \tag{B.20}$$

For example, if $x = \{\{1, 2, 3\}, \{2, 3, 4\}\}$ then $\cup x = \{1, 2, 3, 4\}$. Using the axiom of equality $\cup x$ can be shown to be unique. For any two sets x and y, we define the union of the two sets as follows For example,

$$\{1, 2, 3\} \cup \{2, 3, 4\} = \{1, 2, 3, 4\} \tag{B.21}$$

$$x \cup y \triangleq \cup \{x, y\} \tag{B.22}$$

B.2.6 Axiom of Power:

This axiom tells us that for any set x the set of all subsets of x exists. We call this set the power set of x and represent it as $\mathfrak{P}(x)$. More formally, the axiom of power forces the proposition

$$\forall s \exists \{x : x \subset u\} \tag{B.23}$$

to take the value "True". For example, if $s = \{1, 2\}$ then

$$\mathfrak{P}(s) = \{\{1\}, \{2\}, \varnothing, \{1, 2\}\}. \tag{B.24}$$

Cartesian Products: The Cartesian product of two sets u and v, is symbolized $a \times b$ and defined as follows

$$a \times b = \{(x, y) : (x \in a) \wedge (y \in b)\} \tag{B.25}$$

Using the axioms of separation, union and power, we can show that $x \in y$ exists because it is a subset of $\mathfrak{P}(\mathfrak{P}(x \cup y))$. Using the axiom of identity we can show that it is unique.

Functions: A function f with domain u and target v is a subset of $u \times v$ with the following property: If (a, c) and (b, c) are elements of f then $a = b$. More formally, if the proposition

$$\forall a \forall b \forall c(((a, c) \in f) \wedge ((b, c) \in f) \rightarrow (a = b)) \tag{B.26}$$

takes the value "True" then we say that the set f is a function.

The following formulae are alternative notations for the same proposition:

$$(x, y) \in f \tag{B.27}$$

$$y = f(x) \tag{B.28}$$

$$x \mapsto f(x) \tag{B.29}$$

$$x \mapsto y \tag{B.30}$$

B.2.7 Axiom of Infinity:

This axiom forces the proposition

$$\exists s \forall x (x \in s) \rightarrow (\{x, \{x\}\} \in s) \tag{B.31}$$

to take the value "True". In English this axiom tells us that there is a set s such that if x is an element of s then the pair $\{x, \{x\}\}$ is also an element of s. Sets that satisfy the proposition

$$\forall x (x \in s) \rightarrow (\{x, \{x\}\} \in s) \tag{B.32}$$

are called inductive (or infinite) sets.

Natural numbers: The natural numbers plus the number zero are defined as the intersection of all the inductive sets and are constructed as follows:

$$\begin{aligned}
0 &\triangleq \{\} && \text{(B.33)}\\
1 &\triangleq \{0, \{0\}\} = \{\{\}, \{\{\}\}\} && \text{(B.34)}\\
2 &\triangleq \{1, \{1\}\} = \{\{\{\}, \{\{\}\}\}, \{\{\{\}, \{\{\}\}\}\}\} && \text{(B.35)}\\
&\vdots
\end{aligned}$$

The axiom of existence in combination with the axiom of infinity guarantee that these sets exist. Note that the symbols $1, 2, \cdots$ are just a mnemonic convenience. The bottom line is that numbers, and in facts all sets, are just a bunch of empty curly brackets!

B.2.8 Axiom of Image:

Let $f : u \rightarrow v$ be a function (i.e, a subset of $u \times b$). Define the image of u under f as the set of elements for which there is an element of u which projects into that element. We represent that set as $I_f(u)$. More formally

$$I_f(u) = \{y : \exists x (x \in u) \wedge (f(x) = y)\} \tag{B.36}$$

The axiom of image, also called the axiom of replacement, tells us that for all sets u and for all functions f with domain u the set $I_f(u)$ exists.

B.2.9 Axiom of Foundation:

This axiom prevents the existence of sets who are elements of themselves.

B.2.10 Axiom of Choice:

This axiom tells us that every set with no empty elements has a choice function. A choice function for a set s is a function with domain s and such that for each $x \in s$ the function takes a value $f(x) \in x$ which is an element of x. In other words, the function f picks one element from each of the sets in s, thus the name "choice function". For example, For the set $s = \{\{1,2,3\},\{2,5\},\{2,3\}\}$ the function $f : s \to \{1,2,3\}$ such that

$$f(\{1,2,3\}) = 3 \tag{B.37}$$
$$f(\{2,5\}) = 2 \tag{B.38}$$
$$f(\{2,3\}) = 2 \tag{B.39}$$

is a choice function since for each set in s the function f picks an element of that set. The axiom of choice is independent of the other axioms, i.e., it cannot be proven right or wrong based on the other axioms. The axiomatic system presented here is commonly symbolized as ZFC (Zermelo-Fraenkel plus axiom of Choice), the axiomatic system without the axiom of choice is commonly symbolized as ZF.

History

- The first version of this document was written by Javier R. Movellan in 1995. The document was 8 pages long.
- The document was made open source under the GNU Free Documentation License Version 1.1 on August 9 2002, as part of the Kolmogorov project.

History

- The first version of this document was written by Javier R. Movellan in 1996.
- Javier taught an undergraduate course on Probability and Experimental Design at the Department of Cognitive Science at UCSD. He used this document as the main textbook in 1996, 1997, and 1998.
- The document was made open source under the GNU Free Documentation License Version 1.1 on August 9 2002, as part of the Kolmogorov project. At the time of the release the document had 128 pages and included the following Chapters: (1) Probability; (2) Random Variables; (3) Random Vectors; (4) Expected Values; (5) The precision of the arithmetic mean; (6) Introduction to Statistical Hypothesis Testing; (7) Introduction to Classic Statistical Tests; (8) Introducation to Experimental Design; (9) Experiments with 2 groups; (10) Factorial Experiments; (11) Confidence Intervals; (12) Appendix I: Useful Mathematical Facts; (13) Appendix II: Set Theory.
- October 9, 2003. Javier R. Movellan changed the license to GFDL 1.2 and included an endorsement section.