Level Three Statistics

Alex Elzenaar

September 15, 2019

Contents

Preface

These notes are intended to be a basic introduction to statistics for L3 and Scholarship students. Section 1, on probability measures, follows Rice[1]; the part of section 2 on discrete distributions follows Wackerly, Mendenhall, and Sheaffer; and section 3 on inference follows Efron and Tibsirai.

People who know what they are doing will notice a certain disrespect for problems of finiteness and countability; I have tried to avoid dealing with these problems and have in places given proofs that are either incomplete or not as rigorous as they could be in order to get across the main ideas without getting bogged down in technicalities. For the same reason I have restricted the definitions related to probability density functions, as I do not want to deal with such pathological concepts as integration over the Cantor set!

Prerequisites. A good working knowledge of Level 2 probability is the main prerequisite: although all the results needed are stated precisely and proved, a certain amount of familiarity with what the answers 'should' look like is important. The language of calculus is also occasionally used, but the notes have been written so that such parts can be safely skipped if needed.

[1] The references here are to the bibliography at the end of these notes.

Section 1: Probability

1.1 Foundations

Probability theory is the study of events which occur randomly. The situations within which such events occur are called **experiments**; the set of all possible outcomes of an experiment is called the **sample space** of the experiment. We will generally use Ω to denote sample spaces.

1.1 Example. Flipping a coin three times is an experiment with sample space

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

Particular subsets of sample spaces are called **events**. An event is said to **occur** if one the outcomes contained within it occurs. If A is a subset of Ω we write $A \subseteq \Omega$.

1.2 Example. In the coin-flipping experiment, the event 'heads occurs twice' is the subset

$$A = \{HHT, HTH, THH\}.$$

The event 'at least one tail is flipped' is the subset

$$B = \{HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

1.3 Definition (Set-theoretic operations)**.** If A and B are events in the sample space Ω, then:-

1. The **union** of A and B, denoted $A \cup B$, is the event that either A, or B, or both occurs. In the example above, $A \cup B = \{HHT, HTH, THH, HTT, THT, TTH, TTT\}$ (all outcomes contained within at least one of A and B).
2. The **intersection** of A and B, denoted $A \cap B$, is the event that both A and B occur. In the example above, $A \cap B = \{HHT, HTH, THH\}$ (all outcomes contained within both A and B).
3. The **complement** of A, denoted A^{C}, is the event that A does not occur. In the example above, $A^{\mathsf{C}} = \{HHH, HTT, THT, TTH, TTT\}$ (all outcomes contained in Ω but not A).
4. The **empty set** is the event with no outcomes: $\emptyset = \{\}$. If A and B have no common outcomes then $A \cap B = \emptyset$, and A and B are called **disjoint**.

1.4 Exercise. Draw Venn diagrams to illustrate the following laws:

1. $A \cap B = B \cap A$ and $A \cup B = B \cup A$
2. $(A \cap B) \cap C = A \cap (B \cap C)$ and $(A \cup B) \cup C = A \cup (B \cup C)$
3. $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ and $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

1.5 Definition. A **probability measure** on a sample space Ω is a function P from the subsets of Ω to the real numbers that satisfies the following axioms:

1. $P(\Omega) = 1$
2. If $A \subseteq \Omega$ then $P(A) \geq 0$
3. If A and B are disjoint then $P(A \cup B) = P(A) + P(B)$.

1.6 Proposition (Probability calculus)**.**

1. $P(A^{\mathsf{C}}) = 1 - P(A)$

2. $P(\emptyset) = 0$

3. *If* $A \subseteq B$ *then* $P(A) \leq P(B)$

4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof.

1. A and $A^{\complement}$ are disjoint, and $A \cup A^{\complement} = \Omega$, so $1 = P(\Omega) = P(A \cup A^{\complement}) = P(A) + P(A^{\complement})$ and $P(A^{\complement}) = 1 - P(A)$.
2. $\emptyset^{\complement} = \Omega$ so $P(\emptyset) = 1 - P(\emptyset^{\complement}) = 1 - P(\Omega) = 1 - 1 = 0$.
3. If $A \subseteq B$ then $B = A \cup (B \cap A^{\complement})$ where the two unioned sets are disjoint, and thus $P(B) = P(A) + P(B \cap A^{\complement}) \geq P(A)$ (since $P(B \cap A^{\complement}) \geq 0$).
4. The idea is that if $P(A)$ and $P(B)$ are added together then $P(A \cap B)$ is double-counted. (Draw a Venn diagram). Hence we will write $A \cup B = (A \cap B^{\complement}) \cup (A \cap B) \cup (A^{\complement} \cap B)$ (the unions here are disjoint) and so $P(A \cup B) = P(A \cap B^{\complement}) + P(A \cap B) + P(A^{\complement} \cap B)$. On the other hand, $P(A) = P(A \cap B^{\complement}) + P(A \cap B)$, and $P(B) = P(A \cap B) + P(A^{\complement} \cap B)$. Thus $P(A) + P(B) - P(A \cap B) = P(A \cap B^{\complement}) + P(A \cap B) + P(A^{\complement} \cap B) = P(A \cup B)$.

□

1.7 Example. If a coin is thrown twice then $\Omega = \{HH, HT, TH, TT\}$. Assume that each outcome in Ω is equally likely and has probability $1/4$. Let A denote the event of heads on the first toss, and B denote the event of heads on the second toss. Then $3/4 = P(\{HH, HT, TH\}) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(\{HH, HT\}) + P(\{HH, TH\}) - P(\{HH\}) = 3/4$.

1.8 Remark. A probability measure function is a kind of area: set up a correspondence between the full sample space Ω and a unit square $\square$ of area 1. Then events $A \subseteq \Omega$ correspond exactly to subsets of the square $F \subseteq \square$ such that $P(A)$ equals the area of F.

Compare the properties listed in definition 1.5 and proposition 1.6 with the axioms for a Jordan area function.[2]

1.2 Counting

For finite sample spaces, the following is easy:

1.9 Lemma. *If Ω is finite with N elements, and if every outcome of Ω is equally likely, each outcome has probability $1/N$; if an event $A \subseteq \Omega$ can occur in any one of n mutually exclusive ways then $P(A) = n/N$.*

1.10 Exercise (Simpson's paradox)**.**

1. A black urn contains 5 red and 6 green balls, and a white urn contains 3 red and 4 green balls. You are allowed to choose an urn and then choose a ball at random from that urn. If you choose a red ball, you win a prize. Which urn should you choose to draw from? (Hint: black)
2. Consider another game; this time the black urn has 6 red and 3 green, and the white urn has 9 red and 5 green. Which urn should you choose? (Hint: black)
3. In the final game, the contents of the second black urn are added to the first black urn, and the contents of the second white urn are added to the first white urn. By considering the results to (1) and (2) above, guess which urn should you now pick. Check your answer.

1.11 Proposition (Multiplication principle)**.** *If one experiment has m outcomes, and a second has n outcomes, then there are mn possible outcomes for the two experiments.*

Proof. Denote the outcomes of the first experiment by $a_1, ..., a_m$ and the outcomes of the second by $b_1, ..., b_n$. Then the outcomes of the combination fill exactly an $m \times n$ array in which the pair $a_i b_j$ is in the ith row and jth column. □

1.12 Exercise. An 8-bit binary word is a sequence of eight binary digits. How many different 8-bit words are there?

1.13 Definition. A **permutation** is an ordered arrangement of objects. Suppose from the set $X = \{x_1, ..., x_n\}$ we choose r elements and list them *in order*. If no duplication is allowed, we say we are **sampling without replacement**. If we may choose elements more than once, we say we are **sampling with replacement**.

[2]See, for example, chapter 8 of Lynn Loomis and Schlomo Sternberg, *Advanced calculus* (revised edn.). World Scientific (2014).s

The proof of the following is an exercise.

1.14 Lemma. *The number of orderings of a set with n elements is $n(n-1)\cdots 2\cdot 1$. This number is denoted $n!$ and is called n–**factorial**.*

1.15 Theorem (Elementary counting theorem, part I)**.** *For a set of size n and a sample of size r:*

1. *there are n^r different ordered samples with replacement;*
2. *there are $n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}$ different ordered samples without replacement.*

Proof. We will use the multiplication principle, proposition 1.11. If we allow replacement, the first item in the sample can be chosen n different ways; the second n different ways; and so forth for all r elements. Thus there are $\underbrace{n\cdots n}_{r \text{ times}} = n^r$ samples with replacement.

Without replacement, we may choose the first element n ways; the second $n-1$ ways; and so forth, until we are left with $n-r+1$ choices for the rth element. □

1.16 Exercise.

1. How many ways can five children be lined up?
2. How many ways is it to choose five children from a group of ten, and line them up?
3. How many distinct license plates, made up of three letters followed by three digits, are possible?
4. If all sequences of six characters are equally likely, what is the probability that a license plate for a car contains no duplicate letters or numbers?
5. (Birthday problem) Suppose that a room contains n people. What is the probability that at least two of them share a birth date?
6. How many people must you ask to have a 50:50 chance of finding someone who shares your birthday?

1.17 Theorem (Elementary counting theorem, part II)**.** *The number of unordered samples of r objects selected from n objects without replacement is*

$$\binom{n}{r} := \frac{n!}{(n-r)!r!}.$$

Alternatively, this is the number of subsets of size r that can be found in a set of size n.

*The number $\binom{n}{r}$ is called a **binomial coefficient**.*

Proof. The number of unordered samples of r objects selected from n objects without replacement is just the number of ordered such samples, divided by the number of times each unordered sample occurs (this is just the number of orderings of a sample of size r). Thus we obtain

$$\frac{\frac{n!}{(n-r)!}}{r!} = \frac{n!}{(n-r)!r!}.$$

□

1.18 Corollary (Binomial theorem)**.**

$$(a+b)^n = \binom{n}{0}a^n b^0 + \binom{n}{1}a^{n-1}b^1 + \cdots + \binom{n}{r}a^{n-r}b^r + \cdots + \binom{n}{n}a^0 b^n.$$

1.19 Corollary. *There are 2^n subsets of a set with n elements:*

$$2^n = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n}$$

1.20 Exercise. A monkey at a typewriter types each of the 26 letters of the alphabet exactly once in a random order. What is the probability that the word ‘hamlet’ occurs somewhere in the string of letters?

1.21 Exercise. When performing quality control, only a fraction of the output of a process is sampled and checked. Suppose that n items are produced, and a sample of size r is chosen and examined. Suppose that the entire production contains k defective items. What is the probability that the sample contains exactly m defective items?

1.3 Conditional probability

1.22 Definition. Let A and B be events, and let $P(B) \neq 0$. Then the **conditional probability** of A given B is defined to be

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

The idea is that if we know that B has occured, our sample space is now only the outcomes in B rather than the entirity of Ω.

1.23 Exercise. Is it true that if the occurence of A makes B more likely to occur, then the occurence of B makes A more likely to occur? In other words, if $P(B|A) > P(B)$ does it follow that $P(A|B) > P(A)$?

The following is trivial but often useful.

1.24 Lemma (Multiplication law)**.** $P(A \cap B) = P(A|B)P(B)$.

1.25 Example. An urn contains three red balls and one blue ball. Two balls are selected without replacement. What is the probability that both are red?

Let A denote the event 'the first ball drawn is red', and B denote 'the second ball drawn is red'. Then $P(A) = 3/4$; and $P(B|A) = 2/3$ (if one red ball has been drawn, there are two left out of three). So $P(A \cap B) = (3/4)(2/3) = 1/2$.

1.26 Exercise. If it is cloudy, the probability of rain is 0.3. The probability that it is cloudy is 0.2. What is the probability that it is cloudy and raining?

If A is any event, and if we can find n distinct 'pathways' that could get to A such that at least one pathway occurs, we can find the probability of A.

1.27 Theorem. *Law of total probability Let $B_1, B_2, ..., B_n$ be mutually disjoint events such that $B_1 \cup \cdots \cup B_n = \Omega$. Suppose that $P(B_i) > 0$ for all i. If A is any event, we have*

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n).$$

Proof.

$$\begin{aligned} P(A) = P(A \cap \Omega) &= P\left(A \cap (B_1 \cup B_2 \cup \cdots \cup B_n)\right) \\ &= P\left((A \cap B_1) \cup \cdots \cup (A \cap B_n)\right) \\ &= P(A \cap B_1) + \cdots + (A \cap B_n) \\ &= P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n). \end{aligned}$$

□

1.28 Example. An urn contains three red balls and one blue ball. Two balls are selected without replacement. What is the probability that a red ball is selected in the second draw?

Let A be the event "a red ball is selected in the second draw"; let B_1 be the event "a red ball is selected in the first draw"; let B_2 be the event "a blue ball is selected in the first draw".

Then $P(A|B_1) = 2/3$ (since if a red ball was drawn there are two reds and one blue left), $P(A|B_2) = 1/1$, $P(B_1) = 3/4$, and $P(B_2) = 1/4$; thus the law of total probability tells us that $P(A) = (2/3)(3/4) + (1/1)(1/4) = 3/4$.

1.29 Remark. The law of total probability is simply a symbolic version of the method of 'probability trees', taught at Level 2. An annotated probability tree for the previous example is given in figure 1.

1.30 Exercise.

1. Urn A contains four red, three blue, and two green balls. Urn B has two red, three blue, and four green balls. A ball is drawn from urn A and put into urn B, and then a ball is drawn from urn B.
   (a) What is the probability that a red ball is drawn from urn B?
   (b) If a red ball is drawn from urn B, what is the probability that a red ball was drawn from urn A?
2. A couple has two children. What is the probability that both are girls given that

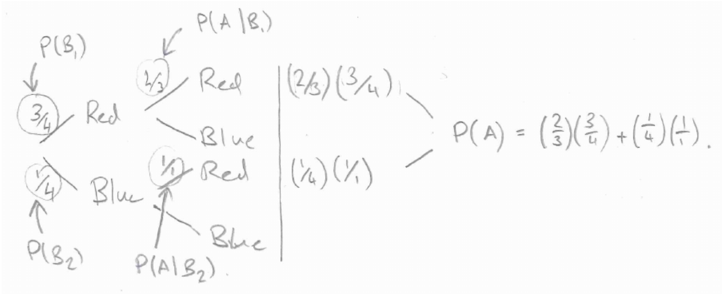

Figure 1: Probability tree for example 1.28.

   (a) the oldest is a girl?
   (b) one of them is a girl?

3. A fair coin is tossed three times.
   (a) Given that there was at least one head, what is the probability that at least two heads were thrown?
   (b) Given that there was at least one *tail*, what is the probability that at least two heads were thrown? (The second 'head' is not a typo.)
4. Adam and Brandi are playing the following game. They write each number from 1 to 100 on seperate pieces of paper; then they randomly select one piece of paper, and then another. They add the two integers that are written on the two pieces of paper. If the sum is even, Adam wins. If the sum is odd, Brandi wins. Is the game fair? Replacing '100' with n, for which values of n is the game fair?
5. Five cards are dealt from a standard 52-card deck, and the first one is a kind. What is the probability of at least one more king?
6. (Monty Hall problem.) You are shown three doors; behind each door lies either a goat or a unicorn. There are two goats, and one unicorn; the object of the game is to end up with the unicorn.

   First, you pick one of the doors, but you are not shown whether a goat or a unicorn is behind it. If only one of the remaining doors hides a goat, it is opened. If both remaining doors hide goats, one is picked at random and opened. This leaves two unopened doors: the one you picked originally, and the door that you did not pick and which has not been opened to reveal a goat.

   You are given a choice: you may either stick with your original choice, or swap to the door that you did not pick and which remains unopened. Does it matter?
7. Show that if $P(A|E) \geq P(B|E)$ and $P(A|E^{\mathsf{C}}) \geq P(B|E^{\mathsf{C}})$ then $P(A) \geq P(B)$.
8. Show that, if the conditional probabilities exist,
$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1}).$$

We have stated laws which allow us to, given conditional probabilities, find probabilities on the entire sample space Ω. As is often the case, the inverse problem is more difficult (division is more difficult than multiplication, antidifferentiation is more difficult than differentiation, ...). One of the most wonderful results of elementary probability, therefore, is the following theorem.

1.31 Theorem (Bayes)**.** *Let $B_1, B_2, ..., B_n$ be mutually disjoint events such that $B_1 \cup \cdots \cup B_n = \Omega$. Suppose that $P(B_i) > 0$ for all i. If A is any event, we have*

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)}$$

Proof. Behold:-

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)}$$

(where the second equality comes from the multiplication law on the top and the law of total probability below). ☐

1.4 Independence

Two events A and B should be independent if the knowledge that B has occured does not provide any information as to the occurance of A (and vice versa); symbolically, if $P(A|B) = P(A)$ and $P(B|A) = P(B)$. By the previous section, the first of these implies that $P(A) = P(A|B) = P(A \cap B)/P(B)$. We will use this as our definition.

1.32 Definition. Two events A and B are called **pairwise independent** if $P(A \cap B) = P(A)P(B)$.

1.33 Example. A card is randomly selected from a deck. Let A be the event that the card is an ace, and let D be the event that the card is a diamond. We have $P(A) = 1/13$, $P(D) = 1/4$, and $P(A \cap D) = 1/52$. But $P(A)P(D) = 1/(13 \cdot 4) = 1/52 = P(A \cap D)$, so the two events are independent.

Warning: if more than two events are considered, it is possible that the entire collection is not independent even though every pair is pairwise independent.

1.34 Example. Let a fair coin be tossed twice, and let the events A, B, C be 'heads on the first toss', 'heads on the second toss', and 'exactly one head is thrown' respectively. Then A and B are independent by assumption of fairness; also, $P(C|A) = 0.5 = P(C)$, and $P(C|B) = 0.5 = P(C)$. Thus all three events are pairwise independent. But $P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C)$.

Because of this, we will make a further definition.

1.35 Definition. A collection of events $A_1, ..., A_n$ are called **mutually independent** if $P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n)$.

1.36 Exercise.

1. Show that if A and B are (pairwise) independent, then A and B^{C} are pairwise independent, and A^{C} and B are pairwise independent.
2. If A and B are disjoint, can they be independent? If $A \subseteq B$, can A and B be independent?
3. If a parent has genotype Aa, they transmit either A or a to an offspring (each with a 0.5 chance). The gene transmitted to one offspring is independent of the gene transmitted to another. Consider a parent with three children and the following events: A, 'children 1 and 2 have the same gene'; B, 'children 1 and 3 have the same gene'; and C, 'children 2 and 3 have the same gene. Show that these three events are pairwise independent but not mutually independent.

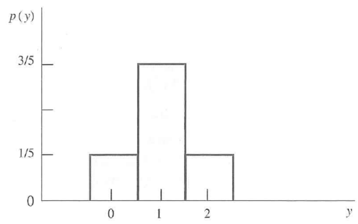

Figure 2: A discrete probability distribution.

Section 2: Random variables

Our discussion above steers clear of 'randomness', despite this being the goal for which we are striving. We would like to define the idea of a variable which can take random values.

We have seen that a probability measure is a function that maps an event in the sample space to a number between 0 and 1; and the random variable will be the description of the set on which the probability measure operates. The values of the random variable will correspond to the outcomes of an experiment.

Our motivating example will be this: consider a set of people; we want to write something like 'the probability that a person has age 13 is p'. Here, our experiment is 'picking a person', and our sample space is the set Ω of people. Our random variable will be the function X that assigns to each person in Ω their age in years; we will define the event 'a random person is assigned the age 13', which we will denote by '$X = 13$', to be the set of all elements x of Ω such that $X(p) = 13$ (i.e. the set of all people with age 13). Then $P(X = 13)$ will be the probability that a randomly chosen person has age 13, and we have sidestepped elegantly the need to have any kind of randomness by operating on the set of all people all at once.

2.1 Discrete random variables

2.1 Definition. Let Ω be a sample space, and let P be a probability measure on Ω. A **random variable** X is a function defined on Ω that can take on values in a set S. The random variable is called **discrete** if the elements of S can be listed in some way such that no element is left off the list.

The event $X = x$ is the set of all ω in Ω such that $X(\omega) = x$; then $P(X = x)$ is simply the probability of the event $X = x$, and we will call $P(X = x)$ 'the probability that X takes the value x'. For conciseness we will often write $p(x) := P(X = x)$ (note the lowercase p) if X is known; the function p is called the **probability function** of X in the presence of Ω and P.

The **probability distribution** of X is simply a graph, formula, or table wich provides $P(Y = y)$ for every value of y.

2.2 Example. A supervisor has three men and three women working for him. He chooses two of them at random for a particular job. Let W denote the number of women he chooses; we will find the probability distribution of W.

The supervisor can select two workers from six in $\binom{6}{2} = 15$ ways. Thus the sample space Ω has 15 outcomes, each of which has probability $1/15$ by assumption. The values for W which have non-zero probability are 0, 1, and 2; the number of ways of selecting $W = 0$ is $\binom{3}{0}\binom{3}{2} = 3$ (choose zero women from three and two men from three), and so $p(0) = P(W = 0) = 3/15$. Similarly, $p(1) = \binom{3}{1}\binom{3}{1}/15 = 9/15$, and $p(2) = 3/15$. Hence $W = 1$ is the most likely outcome, and the probability distribution is in 2.

The following facts are obvious:

2.3 Proposition.

1. $0 \leq P(X = x) \leq 1$ *for any* x.
2. *If* X *can take the values* $x_1, ..., x_n, ...$ *then* $P(X = x_1) + P(X = x_2) + \cdots + P(X = x_n) + \cdots = 1$.

2.4 Definition. If X is a random variable on Ω that takes numerical values, then the subset of elements ω of Ω such that $x \leq X(\omega) \leq y$ is called the event $x \leq X \leq y$. The weak inequalities can be substituted with strong inequalities ($<$) in the obvious fashion.

The following theorem is obvious from the definition.

2.5 Theorem (Discrete integration theorem)**.** *If a probability distribution is graphed, as in 2.2, then $P(x \leq X \leq y)$ is the area of the histogram bounded by x and y on the axis. More formally, $P(x \leq X \leq y) = \int_x^y X(\omega)\,\mathrm{d}\omega$.*

2.6 Exercise. Alice and Bob play a game where each player tosses a fair coin. If the coins both come up tails, Alice wins \$1; if the coins both come up heads, Alice wins \$2; if the coins do not match, Alice loses \$1 (wins −\$1).

Find the probability distribution for X, Alice's expected winnings, after (a) a single play of the game; (b) five consecutive plays.

2.2 Expected values

2.7 Definition. Let X be a discrete random variable, taking on the values $x_1, ..., x_n, ...$, with probability function $p(x)$. Then the **expected value** of X is defined to be $E(X) = x_1 p(x_1) + \cdots + x_n p(x_n) + \cdots$.

This 'expectation' is a generalisation of the concept of **mean** that we have worked with since primary school. If we consider the set $\Omega = \{\omega_1, ..., \omega_n\}$ then the mean of the set is

$$\mu(\Omega) = \frac{1}{n}(\omega_1 + \cdots + \omega_n).$$

Now suppose we define a function X from Ω such that $X(\omega_i) = \omega_i$ for each i, and we equip Ω with the uniform probability function: that is, we set $P(\omega_i) = 1/n$ for each i. I claim that $E(X) = \mu(\Omega)$. Indeed,

$$E(X) = \omega_1 P(\omega_1) + \cdots + \omega_n P(\omega_n) = \omega_1(1/n) + \cdots + \omega_n(1/n) = \frac{1}{n}(\omega_1 + \cdots + \omega_n) = \mu(\Omega).$$

Because of this, we will often write mean rather than expected value, and $\mu(X)$ for $E(X)$.

We are frequently interested in the average value (or more properly the expectation value) of some function of a random variable. For example, molecules in space move at varying velocities; the velocity V is a random variable. Further, the kinetic energy $E = \frac{1}{2}mV^2$ is a function of the random variable V, and is therefore itself a random variable. Consequently, to find the mean amount of energy we need to find the mean value of V^2.

2.8 Theorem. *Let X be a real-valued discrete random variable, taking the values $x_1, ..., x_n$, with probability function $p(x)$. Let g be a real-valued function that is defined on all the values of X (we will often abuse language and say that g is a function defined on X). Then the expected value of $g(X)$ is*

$$E[g(X)] = g(x_1)p(x_1) + \cdots + g(x_n)p(x_n)$$

Proof (optional). We will prove this in the case that X takes on only finitely many values, $x_1, ..., x_n$. Suppose $g(X)$ takes on the distinct values $g_1, ..., g_m$, where $m \leq n$ (since two different values of X might produce the same value of $g(X)$). It follows that $g(X)$ is a random variable satisfying $P[g(y) = g_i] = \sum_{\substack{\text{all } x_j \text{ such that} \\ g(x_j) = g_i}} p(x_j)$ (where Σ is used to denote a sum); denote the probability function of $g(X)$ by p^*. Then

$$\begin{aligned} E[g(X)] &= g_1 p^*(g_1) + \cdots + g_m p^*(g_m) \\ &= g_1 \left(\sum_{\substack{\text{all } x_j \text{ such that} \\ g(x_j) = g_i}} p(x_j) \right) + \cdots + g_n \left(\sum_{\substack{\text{all } x_j \text{ such that} \\ g(x_j) = g_i}} p(x_j) \right) \\ &= g(x_1)p(x_1) + \cdots + g(x_n)p(x_n). \end{aligned}$$

$\square$

2.9 Remark. The above theorem is actually true for *every* discrete random variable; however, the proof becomes fiddly when the sums become infinite (one needs to check that they exist).

Using our generalised mean, we will define a generalised standard deviation. Recall that the **variance** of the set $\Omega = \{\omega_1, ..., \omega_n\}$ is the number

$$\sigma^2(\Omega) = \frac{1}{n-1}\left[(\omega_1 - \mu(\Omega))^2 + \cdots + (\omega_n - \mu(\Omega))^2\right]$$

(that is, the sum of the squares of the differences between the measurements and their mean, divided by $n-1$) and the **standard deviation** of Ω is just $\sigma = \left|\sqrt{\sigma^2(\Omega)}\right|$ (that is, the positive square root of the deviance).

2.10 Definition. If X is a standard variable with expectation $E(X) = \mu$, the **variance** of X is defined to be the expected value of $(X-\mu)^2$. That is,

$$V(X) = E[(X-\mu)^2].$$

The **standard deviation** of X is defined to be the positive square root of $V(X)$.

2.11 Example. A random variable X has the following probability distribution.

x	$P(X = x)$
0	1/8
1	1/4
2	3/8
3	1/4

Then

$$E(X) = 0(1/8) + 1(1/4) + 2(3/8) + 3(1/4) = 1.75,$$

and

$$\begin{aligned} V(X) &= E((X-1.75)^2) \\ &= (0-1.75)^2(1/8) + (1-1.75)^2(1/4) + (2-1.75)^2(3/8) + (3-1.75)^2(1/4) \\ &= 0.9375; \end{aligned}$$

thus the mean is $\mu = 1.75$ and the standard deviation is $\sigma = \sqrt{0.9375} = 0.968$.

Note that $\mu \pm \sigma$ is the range $0.782 < X < 2.718$, which includes 1 and 2 (i.e. over 62% of the total probability distribution).

We will now prove some simple rules for calculating combinations of expectation values.

2.12 Proposition. *Let X be a random variable, and let $p(x)$ be the function $P(X = x)$. Let $g, g_1, ..., g_n$ be functions of X, and let λ be a constant. Then:*

1. $E(\lambda) = \lambda$
2. $E(\lambda g(X)) = \lambda E(g(X))$
3. $E(g_1(X) + \cdots + g_n(X)) = E(g_1(X)) + \cdots + E(g_n(X))$

The final two statements together are called ***linearity****.*

Proof. Suppose X can take the values $x_1, ...$ Then:-

1. Note that λ is trivially a function of X, and so $E(\lambda) = \lambda P(x_1) + \cdots = \lambda(P(x_1) + \cdots) = \lambda$.
2. $E(\lambda g(X)) = \lambda g(x_1)P(x_1) + \cdots = \lambda(g(x_1)P(x_1) + \cdots) = \lambda E(g(X))$.
3. $E(g_1(X) + \cdots + g_n(X)) = [g_1(x_1) + \cdots + g_n(x_1)]P(x_1) + \cdots = [g_1(x_1)P(x_1) + \cdots] + \cdots + [g_n(x_1)P(x_1)] = E(g_1(X)) + \cdots + E(g_n(X))$.

$\square$

We may now give an easier method for calculating the variance of a discrete random variable.

2.13 Proposition. *Let X be a random variable, and let $p(x)$ be the function $P(X = x)$. Let $\mu = E(X)$, and let $\sigma = +\sqrt{V(X)}$. Then:*

$$V(X) = \sigma^2 = E[(X - \mu)^2] = E(X^2) - \mu^2.$$

Proof. The first two equalities are by definition, so we need to show that $E[(X - \mu)^2] = E(X^2) - \mu^2$. Applying the rules from proposition 2.12 we find

$$E[(X - \mu)^2] = E(X^2 - 2\mu X + \mu^2) = E(X^2) + E(-2\mu X) + E(\mu^2) = E(X^2) - 2\mu E(X) + \mu^2$$

and since $E(X) = \mu$, $E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2$. □

2.14 Exercise. Recall from exercise 2.6 that Alice and Bob were playing a game. Calculate the mean and variance of the random variable X, Alice's winnings, after one play. How much should Alice pay to play a round of the game if her net winnings, the difference between the payoff and cost of playing, are to have mean 0?

2.15 Exercise. Let X be a discrete random variable with mean $E(X) = \mu$ and variance $V(X) = \sigma^2$. Define $Y = X + 1$.

1. Do you expect $E(Y)$ to be larger than, equal to, or smaller than $E(X)$? Why?
2. Check your answer to 1. by writing $E(Y + 1)$ in terms of $E(X)$.
3. Recalling that the variance is a measure of spread, do you expect $V(Y)$ to be larger than, equal to, or smaller than $V(X)$? Why?
4. Use 2. to verify your expectation from 3.

2.3 Binomial distributions

The distribution we will consider now models a situation where a process is repeated, and each time only two outcomes are possible. A production line produces a series of items, each either defective or nondefective; a coin is flipped several times; a sequence of rocket launches either succeeds or fails.

2.16 Definition. A **binomial distribution** is an experiment that consists of a fixed number n of identical trials, such that:

1. Each trial results in one of two outcomes, success or failure.
2. The probability of success for each trial is equal to some value p that is constant across the experiment.

2.17 Theorem. *If an experiment satisfies the three conditions for a binomial distribution, with n trials and success probability p, then the random variable X defined to be the number of successful trials, has the following probability function:*

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

This is called the probability function of the distribution.

Proof. Let Ω be the sample space of the experiment; Ω is the set of all strings of length n consisting of the characters S (success) and F failure. Then $X(\omega)$ is the function that counts the number of instances of S in ω.

We need to count the number of items in Ω with x instances of S. This is the same as picking, without replacement, x numbers from 1 to n (the numbers of the trials which succeed); i.e. there are $\binom{n}{x}$ such items. Further, the probability of each such item showing up is $p^x(1 - p)^{n-x}$. The result follows. □

The binomial distributions for $n = 100$ and various success probabilities p are graphed in figure 3.

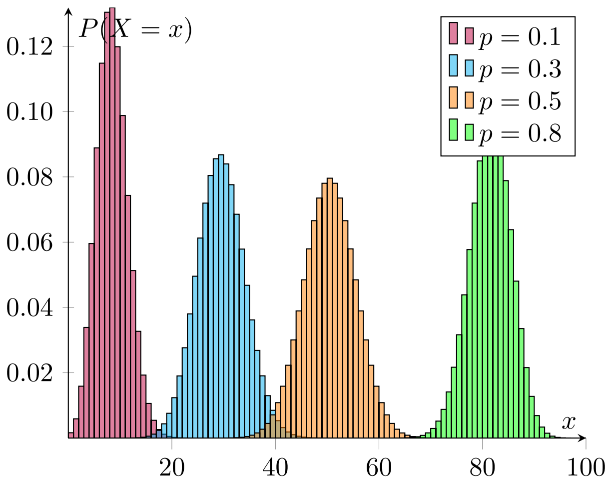

Figure 3: Various binomial distributions, $n = 100$.

2.18 Example. If binary information is transmitted across an unreliable or noisy channel, we can model the probability that a single bit is incorrectly transmitted as a constant, p. To improve reliability, we repeat each bit n times (where n is odd) and decide that the bit is 1 if a majority of the repeated digits arrive with value 1. If the probability p is constant over time, we have a binomial distribution with n trials and success probability p (where 'success' means incorrect transmission).

For a given message, we transmit each bit five times and there is a 10% chance that any received bit has been changed in transmission (so $n = 5$ and $p = 0.1$). Then the probability a given bit is correctly received is the probability of at least three correct repeated bits; that is, the probability of at most two errors.

This probability is just

$$P(X = 0) + P(X = 1) + P(X = 2) = \binom{5}{0} 0.1^0 0.9^{5-0} + \binom{5}{1} 0.1^1 0.9^{5-1} + \binom{5}{2} 0.1^2 0.9^{5-2} = 0.9914$$

which is much better than p.

2.19 Exercise.

1. Which is more likely: 9 heads in 10 tosses of a fair coin, or 18 heads in 20 tosses?
2. Consider the binomial distribution of n trials and success probability p, with probability function $P(X = x)$. For what value k is $P(X = k)$ maximised? (Hint: consider $P(X = k)/P(X = k + 1)$.)
3. What are the mean and standard deviation for a binomial distribution?

2.4 Continuous distributions

We have considered discrete probability distributions, where a given random variable X can take on discrete values $x_1, ..., x_n, ...$, and the probability that X takes on a particular value is predetermined: the function $P(X = x)$ is well-defined. We would now like to extend this to cover situations where the probability of any particular value is tiny. For example, consider the set of all heights: since heights are continuous, the probability that any particular chosen height is exactly 2.00 m (say) is essentially zero.

The most profitable method for this generalisation is given by proposition 2.5, the discrete integration theorem. Instead of viewing our probability functions as giving probabilities of particular values, we will view probability functions as giving probabilities of *ranges* of values.

We will confine ourselves to sample spaces which are **open intervals**: that is, if a and b are real numbers such that $a < b$, the open interval (a, b) is defined to be the set of all x such that $a < x < b$. (We will allow a or b to take the values $\pm\infty$.)

To avoid pathologies, we will restrict our definition of an 'event' in Ω to be a *finite* union of sub-intervals of Ω. That is, if $\Omega = (0, 1)$, the set of all rational numbers in $(0, 1)$ will not be considered an event by us. We will even modify the notation $A \subseteq \Omega$ to mean 'A is a subset of Ω such that A is a finite union of subintervals'.

It is possible to make the theory work if we remove these restrictions, but the author is not a measure theorist.

If $X \subseteq \Omega$ can be written as the finite union $X = (\alpha_1, \beta_1) \cup \cdots \cup (\alpha_n, \beta_n)$ then we will write

$$\int\limits_X \rho(x)\,\mathrm{d}x := \int\limits_{\alpha_1}^{\beta_1} \rho(x)\,\mathrm{d}x + \cdots + \int\limits_{\alpha_n}^{\beta_n} \rho(x)\,\mathrm{d}x.$$

(Recall that $\int_\alpha^\beta \rho(x)\,\mathrm{d}x$ is the **definite integral** of ρ from α to β, and measures 'the area under the graph of $y = \rho(x)$ between the bounds $x = \alpha$ and $x = \beta$'.)

2.20 Definition. Let $\Omega = (a, b)$ be a sample space, and let ρ be a function from (a, b) taking real values such that:

1. Whenever α and β satisfy the inequality $a \leq \alpha \leq \beta \leq b$, $\int_\alpha^\beta \rho(x)\,\mathrm{d}x$ exists and is non-negative, and
2. The identity $\int_a^b \rho(x)\,\mathrm{d}x = 1$ holds.

Then ρ is called a **probability density** (or simply **density**) on Ω. If ρ is continuous (that is, if $\lim_{x\to x_0} \rho(x) = \rho(x_0)$ for all x_0 in Ω), then ρ is called a **continuous probability density**.

As is often the case, we will prove a big 'this definition actually makes sense' theorem.

2.21 Theorem. *Let $\Omega = (a, b)$ be a sample space and let ρ be a probability density on Ω. Then the function P defined on the subsets $A \subseteq \Omega$ by*

$$P(A) = \int\limits_A \rho(x)\,\mathrm{d}x$$

is a probability measure on Ω.

Proof. We need to check the three criteria of definition 1.5:

1. $P(\Omega) = 1$: indeed, $\int_\Omega \rho(x) = \int_a^b \rho(x)\,\mathrm{d}x = 1$ (guaranteed by definition of a density function).
2. If $A \subseteq \Omega$ then $P(A) \geq 0$: this is guaranteed by the non-negativity condition in criterion 1 of the density function definition.
3. If A and B are disjoint then $P(A \cup B) = P(A) + P(B)$: this follows by additivity of integrals.

$\square$

Note that our definition of a continuous random variable will be simply a density function; the probability measure on the sample space will be very tightly coupled with the density function (instead of sticking a random variable into an already existing probability space, we are defining the probability in terms of the random variable).

2.22 Definition. A **continuous random variable** X on a sample space Ω is defined to consist of the following data:

1. A sample space, $\Omega = (a, b)$.
2. A continuous probability density ρ on Ω.
3. The probability measure P induced by ρ, as defined by theorem 2.21.

We will write $\alpha < X < \beta$ for the event $(\alpha, \beta) \subseteq \Omega$. Hence

$$P(\alpha < X < \beta) := \int\limits_\alpha^\beta \rho(x)\,\mathrm{d}x.$$

We can generalise our concept of expected value.

2.23 Definition. If X is a continuous random variable on (a, b) with density ρ, the **expected value** X is defined to be

$$E(X) := \frac{1}{b-a} \int_a^b x\rho(x)\,\mathrm{d}x.$$

Let g be a continuous function on $(a, b) = \Omega$; if X is a continuous random variable on Ω, it would be nice to define some continuous random variable $g(X)$ in analogy with the discrete case. In order to do this, we will need to equip $g(X)$ with a sample space and a density function. If Ω is the sample space of X then the natural sample space is $g(\Omega)$ (i.e. for every x in Ω, $g(x)$ is in the sample space of $g(X)$); we just need to pick a density function.

It will turn out that the function g needs to be *differentiable*, not just continuous, so that we can apply the inverse chain rule (also known as u–substitution). We also need the function g' to be continuous (recall that it is possible for a differentiable function to have a discontinuous derivative; e.g. $x \mapsto \int_0^x |t|\,\mathrm{d}t$ has derivative $x \mapsto |x|$ by the fundamental theorem of calculus). A differentiable function whose derivative is continuous is called **continuously differentiable**.

We will first deal with the case where g is strictly increasing with no stationary points: whenever $\alpha < \beta$ in Ω, $g(\alpha) < g(\beta)$, and $g'(x) \neq 0$ for all x in Ω.

2.24 Remark. It is implicitly assumed that if Ω is an open interval then $g(\Omega)$ is an open interval; this follows by the strictly increasing assumption (it is just the inverse function theorem) but is not true in general (e.g. if $f(x) = x^2$ then $f((-1, 1)) = [0, 1)$ which is not open).

2.25 Theorem. *If X is a continuous random variable with sample space Ω and density ρ, and if g is a strictly increasing real-valued continuously differentiable function on Ω, then the real-valued function π from $g(\Omega)$ defined by $\pi(g(x)) = \rho(x)/g'(x)$ is a continuous probability density on $g(\Omega)$; hence the object $g(X)$ consisting of the space $g(\Omega)$, the function π, and the associated probability measure Q on $g(\Omega)$, is a continuous random variable.*

Proof. We need to check that $\pi : g(x) \mapsto \rho(x)/g'(x)$ is a continuous probability density: it is continuous since ρ and g' are continuous, so we need to check the two conditions of definition 2.20.

By remark 2.24, $g(\Omega) = (g(a), g(b))$. Pick $g(\alpha)$ and $g(\beta)$ in $g(\Omega)$ such that $g(a) \leq g(\alpha) \leq g(\beta) \leq g(b)$. Then:

$$0 \leq \int_\alpha^\beta \rho(x)\,\mathrm{d}x = \int_\alpha^\beta \frac{\rho(x)}{g'(x)} g'(x)\,\mathrm{d}x = \int_\alpha^\beta \pi(g(x))g'(x)\,\mathrm{d}x = \int_{g(\alpha)}^{g(\beta)} \pi(y)\,\mathrm{d}y,$$

and

$$1 = \int_a^b \rho(x)\,\mathrm{d}x = \int_a^b \frac{\rho(x)}{g'(x)} g'(x)\,\mathrm{d}x = \int_a^b \pi(g(x))g'(x)\,\mathrm{d}x = \int_{g(a)}^{g(b)} \pi(y)\,\mathrm{d}y.$$

□

We can extend this now to the case where g is not too wiggly.

2.26 Definition. A function g is called **hemiwiggly** on an interval (a, b) if it is continuously differentiable and has only finitely many stationary points on that interval.

Section 3: Inference

The primary reference for this section is Efron and Tibshirani. There are three basic questions to be asked in statistical investigations:

1. How should I collect my data?
2. How should I summarise the data I have collected?
3. How accurate are my summaries?

One tool to help answer the third question is the bootstrap, a comparatively modern methord for making particular kinds of statistical inference. The calculations involved are intricate enough that they are virtually impossible without a computer; the iNZight software can perform the relevant computations.[3]

A study[4] was performed in 1987 to see whether small aspirin doses would prevent heart attacks in healthy middle-aged men. The data were collecred by a controlled, randomised, double-blind study. One half of the subjects reveived aspirin, the other half placebo; the subjects were randomly assigned to the groups, and the supervising doctors were not told which group each subject was assigned to. These precautions guard against seeing benefits that don't exist, while maximising the chance of detecting a positive effect.

Here is a summary of the statistics that were measured:

	heart attacks	**subjects**
aspirin	104	11037
placebo	189	11034

We can calculate the relative risk that a person has a heart atach in the aspirin group versus the placebo group fairly easily:

$$\hat{\rho} = \frac{104/11037}{189/11034} - 0.55$$

Thus if the study is reliable, the aspirin-takers have only 55% as many heart attacks as the placebo-takers.

In reality, we are not interested in the relative risk $\hat{\rho}$ above — it is only the *estimated* risk for this particular sample. Rather, we are interested in the relative risk ρ for the entire population. Indeed, the sample here is large (over 22,000 subjects); but the number of observed heart attacks is only 293.

It is possible to compute, using bootstraps, a confidence interval for the real value of ρ: we find that, with 95% condidence, the real value lies within the interval

$$0.43 < \rho < 0.70.$$

The importance of confidence intervals is illustrated by another example from the same study. Here is the data for strokes, measured for the same subjects as the heart attack table above.

	heart attacks	**subjects**
aspirin	119	11037
placebo	98	11034

Now the relative risk is estimated to be $\hat{\rho} = 1.21$: that is, taking aspirin seems be harmful! On the other hand, if the confidence interval is calculated we find that $0.93 < \rho < 1.59$. This range includes the neutral value $\rho = 1$, and so it is impossible to say with any certainty whether aspiring is beneficial or harmful.

A bootstrap is a data-based simulation method for producing confidence intervals like these. The term comes from the phrase 'to pull oneself up by one's bootstraps'.[5]

We will outline the process of boostrapping as it pertains to the stroke example. We create two populations: the first consisting of 119 ones and 11037 zeroes, and the second consisting of 98 ones and 11034 zeroes. We draw, with replacement, a sample of 11037 items from the first population (sample A), and 11034 items from the second population (sample B). From these, we derive the **bootstrap replicant** of $\hat{\rho}$:

$$\hat{\rho}^* = \frac{\text{proportion of ones in sample A}}{\text{proportion of ones in sample B}}.$$

[3]A guide can be found at this URL: `https://www.stat.auckland.ac.nz/~wild/d2i/exercises/6.5%20exercise-bootstrap-re-sampling_16.pdf`.

[4]Efron and Tibshirani, pp.2–4

[5]OED: To make use of existing resources or capabilities to raise (oneself) to a new situation or state; to modify or improve by making use of what is already present.

We repeat this process a large number of times — say 1000 times — and end up with 1000 bootstrap replicants. These replicates contain a large amount of information about the 'shape' of the original data. For example, the standard deviation of the 1000 replicants generated from the stroke data turned out to be 0.17; this is an estimate of the standard error of the relative risk $\bar{\rho}$, and indicates that the observed ratio of $\hat{\rho} = 1.21$ is only a little more than one standard error larger than 1; thus a value of $\rho = 1$ cannot be ruled out.

To produce a rough 95% confidence interval like $0.93 < \rho < 1.59$, we simply take the 25th and 75th largest of the replicates (which turned out to be 0.93 and 1.59).

(Why does this witchcraft work?!)

Bibliography and further reading

See references in text, as well as:

- Bradley Efron and Robert J. Tibshirani, *An introduction to the bootstrap.* Chapman & Hall/CRC (1998).
- Miklós Bóna, *A walk through combinatorics* (3e). World Scientific (2013).
- John A. Rice, *Mathematical statistics and data analysis* (3e). Thomson Brooks/Cole (2007).
- Dennis D. Wackerly, William Mendenhall III, and Richard L. Sheaffer, *Mathematical statistics with applications* (7e). Thomson Brooks/Cole (2008).