Pre-sessional Mathematics for Big Data MSc
Class 3: Probability

Yuri Kalnishkan

September 22, 2018

Probability can be explained at different levels of mathematical rigour. However, the authors of textbooks on probability are faced with an awful trade-off. An intuitive discussion of probability quickly becomes vague and frightfully imprecise. On the contrary, a rigorous presentation of probability quickly becomes horribly formal with no sight of the intuitive picture. You will no doubt see papers and books on probability falling for either of these traps. As long as you stay in the world of discrete probability, it is usually fine (that is why many elementary courses deal only with coins and dice), but we cannot accept this restriction.

I will attempt to stay on the intuitive level while hinting at the heavier machinery behind the scene.

There are many good books on probability. A gentle introduction covering the topics from this tutorial is provided in [Ros06]. A really profound overview of probability theory is contained in lecture notes [Tao] by Terence Tao, but they are far from elementary.

1 A Sample Problem

Attempt this exercise before reading further:

Exercise 1. Alice and Bob play dice. Alice rolls a die once and then Bob rolls a die once. Whoever gets more points wins. What is the probability that Alice wins?

Rolling a die can return one of six scores: 1, 2, 3, 4, 5, or 6. The distribution on them is *uniform*: the probability to get either of these scores equals 1/6.

What is the probability of Alice scoring, say, 5, and Bob scoring 3? These are *independent events*: knowing that Alice scored 5 tells us absolutely

nothing about Bob's score; the probability of him scoring any number is still 1/6. So the probability of the joint event "Alice scores 5 and Bob scores 3" is the product of probabilities: $1/6 \cdot 1/6 = 1/36$.

Now we can proceed as follows. We can enumerate all elementary outcomes where Alice wins. Alice wins if she gets 2 and Bob gets 1; if she gets 3 and Bob gets 2; if she gets 3 and Bob gets 1... Each of these elementary events has probability 1/36. We just need to count them.

The task of counting all outcomes in Alice's favour is doable within reasonable time, but is very boring. There is a better way to approach this problem. Let us look at the situation from a higher level. The elementary outcomes 1 : 1, 2 : 1, 1 : 2, 2 : 2 ... can be grouped into the following *events*: A = "Alice wins", B = "Bob wins", D = "Draw". The probabilities of A and B are equal; mathematically this can be written as $\Pr(A) = \Pr(B)$. Indeed, the game is completely symmetric; there is no *bias* in favour of Alice or Bob. The next important observation is that the probabilities of A, B, and D sum up to 1:

$$\Pr(A) + \Pr(B) + \Pr(D) = 1 \ . \tag{1}$$

Indeed, there is no overlap between A, B, or D (in order words, they are mutually exclusive) and they cover all possible elementary outcomes: each elementary outcome results in Alice winning, Bob winning, or a draw.

We thus get

$$2\Pr(A) + \Pr(D) = 1 \ . \tag{2}$$

It remains to calculate the probability of D and that is very easy. There are six outcomes that result in a draw: 1 : 1, 2 : 2, 3 : 3, 4 : 4, 5 : 5, and 6 : 6. Each has probability 1/36. The probability of D is thus $\Pr(D) = 6 \cdot 1/36 = 1/6$ and this can be plugged into (2). We get

$$\Pr(A) = \frac{1}{2}(1 - \Pr(D)) = \frac{5}{12} \ .$$

2 Probability Space

Now I will introduce a bit of formal maths.

A *sample space* Ω consists of elementary *outcomes* ω. An *event* A is a subset of Ω, i.e., a set of outcomes. In the example above, an elementary outcome consisted of two scores, Alice's and Bob's. The sample space is the set of possible outcomes, 1:1, 1:2, ..., 6:6 (essentially the Cartesian product $\{1, 2, 3, 4, 5, 6\}^2$). We grouped elementary outcomes into events A, B, and C.

A probability measure Pr (or P; I use Pr in this tutorial even though I have to spend an extra letter) assigns probabilities to events. A probability measure should satisfy the following requirements (often called *Kolmogorov axioms* after Andrey Kolmogorov, a Russian mathematician of the 20th century):

1. $\Pr(A) \geq 0$; probabilities cannot be negative;
2. $\Pr(\Omega) = 1$, i.e., the total probability is 1; recall 1 on the right-hand side of equation (1) and the discussion right after it;
3. for a sequence of events $A_1, A_2, \ldots$ that are pairwise *disjoint* or *mutually exclusive* (for $i \neq j$ we have $A_i \cap A_j = \varnothing$, i.e., an empty set) we have
$$\Pr(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} \Pr(A_i) \ .$$

The last axiom of course applies to finite collections of events too. If $C \cap D = \varnothing$, then $\Pr(C \cup D) = \Pr(C) + \Pr(D)$. Having infinity in the statement of the last axiom is important for theoretical derivations. This infinity is not that important for us here.

Recall (1) again; we effectively said that $A \cup B \cup D = \Omega$ and that A, B, and D are mutually exclusive, and wrote $\Pr(A) + \Pr(B) + \Pr(D) = 1$.

In the example above our outcome space Ω was finite (and consisted of pairs of scores). We could write it down on a piece of paper if we wanted to. More importantly, the probability measure Pr came from probabilities of elementary outcomes: each had the probability of 1/36.

Sometimes we have to deal with much trickier spaces when it is hard to keep sight of elementary outcomes. People may even choose not to formally specify them and speak just about events. However the elementary outcomes are lurking somewhere in the background.

Let R be the event "It rains tomorrow in London". What are the elementary outcomes? Quite often we do not care. If we do, we may say that Ω is the set of all states of the universe tomorrow with each ω describing a state of the universe in detail. Then R is the set of such ωs that involve rain in London. Similarly, let S be the set of ωs where the London stock exchange goes up tomorrow.

For the tricky outcomes from the above example, the probabilities of elementary outcomes are quite hard to establish, and usually that is not what we care about. But elementary outcomes combine into our events R and S, which have non-zero probabilities.

What is the event $R \cup S$? We can expand this as "It rains tomorrow *or* London stock market goes up". However we cannot write $\Pr(R \cup S) = \Pr(R) + \Pr(S)$: the events are clearly not disjoint. The event $R \cap S$ is "It rains tomorrow *and* London stock market goes up". Now we can use the formula

$$\Pr(A) + \Pr(B) - \Pr(A \cap B) = \Pr(A \cup B) \ ;$$

it is similar to the inclusion-exclusion formula in combinatorics. We can derive it by considering disjoint events $A \setminus B$, $B \setminus A$, and $A \cap B$. Their union is $A \cup B$, while $A = (A \setminus B) \cup (A \cap B)$ and $B = (B \setminus A) \cup (A \cap B)$.

An important fact is given by *the law of total probability*. Let the events $B_1, B_2, \ldots$ be disjoint ($B_n \cap B_m$ for $n \neq m$) and cover the whole sample space. Then for any event A we have

$$\Pr(A) = \sum_{n=1}^{\infty} \Pr(A \cap B_n) \ .$$

Indeed, $A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots$ and the events $A \cap B_n$ and $A \cap B_m$ remain disjoint for $n \neq m$.

By the law of total probability we can write

$$\begin{aligned} &\Pr(\text{the London stock market goes up tomorrow}) = \\ &\Pr(\text{the London stock market goes up and it rains in London tomorrow}) + \\ &\qquad \Pr(\text{the London stock market goes up and it does not rain} \\ &\qquad\qquad \text{in London tomorrow}) \end{aligned}$$

3 Independence and Conditional Probabilities

An important concept of probability is that of *independence*. Events A and B are independent if $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$. Intuitively independence means that knowing that A has occurred (e.g., Alice has scored 5) tells us nothing of whether B has occurred or not (e.g., whether Bob has scored 3).

Exercise 2. Are events "Alice has scored 5" and "Alice has scored 4" independent?

Note 1. People sometimes confuse independent events with mutually exclusive events. Do not! Those are completely different concepts!

As a mnemonic rule, probabilities of independent events multiply[1].

[1] It is hard to show independent events on a Venn diagram. Independence means that $A \cap B$ takes as much area of A as B of the whole Ω; cf. the discussion of conditional probabilities below

Exercise 3. Prove the following important but often overlooked fact. If A and B are independent, then the following pairs of events are also independent:

- A and $\Omega \setminus B$;
- $\Omega \setminus A$ and B;
- $\Omega \setminus A$ and $\Omega \setminus B$.

Rain in London and the state of the stock exchange are more or less independent[2] so we can write $\Pr(R \cap S) = \Pr(R) \Pr(S)$.

Conditional probability $\Pr(A \mid B)$ (reads "the probability of A given B") is the ratio

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr B} \ .$$

The intuitive meaning of conditional probability may be described as follows. Suppose we know that B has occurred. Our probability space Ω effectively contracts to B: anything outside B is no longer possible. As B is the new Ω, its probability is now 1. What is the new probability of A? The conditional probabilities are just these "new" probabilities *given* B.

If A and B are independent, then $\Pr(A \cap B) = \Pr(A) \Pr(B)$ and

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr B} = \frac{\Pr(A) \Pr(B)}{\Pr B} = \Pr(A) \ .$$

Thus stating that $\Pr(A \mid B) = \Pr(A)$ is another way of saying that A and B are independent. The equation $\Pr(A \mid B) = \Pr(A)$ implies $\Pr(A \cap B) = \Pr(A) \Pr(B)$ and $\Pr(B \mid A) = \Pr(B)$. These all are equivalent formulas[3].

The law of total probability can be now written as

$$\Pr(A) = \sum_{n=1}^{\infty} \Pr(A \cap B_n) = \sum_{n=1}^{\infty} \Pr(A \mid B_n) \Pr(B_n) \ .$$

Exercise 4. Write the law of total probability with conditional probabilities for the probability of the stock market going up. What happens if we assume the stock market to be independent of the weather?

[2] A special message for finance students. No, they are in fact not. The performance of many companies is affected by the weather. In the extreme case a torrent in London will send the economy and, consequently, the stock market into a complete collapse. These probabilities are tiny but should not be ignored. Sometimes they may play their part.

[3] Well, as long as $\Pr(B) \neq 0$ and $\Pr(A) \neq 0$. If $\Pr(B) = 0$, difficulties creep in and probability theory becomes really subtle.

Laws of probability can often be conditioned on an event. For example, given an event C we can write the law of total probability as

$$\Pr(A \mid C) = \sum_{n=1}^{\infty} \Pr(A \cap B_n \mid C) = \sum_{n=1}^{\infty} \Pr(A \mid B_n, C) \Pr(B_n \mid C) \ .$$

Indeed, imagine the world where C has occurred and become the new Ω. In the formula above $\Pr(A \mid B_n, C)$ is the probability of A given B_n *and* C (provided B_n and C have occurred). It is the same as $\Pr(A \mid B_n \cap C)$.

4 Bayes Rule

A common error is to confuse $\Pr(A \mid B)$ and $\Pr(B \mid A)$. These two conditional probabilities may be very different! Let us find a formula linking them together.

There are two ways to write the joint probability $\Pr(A \cap B)$:

$$\Pr(A \mid B) \Pr(B) = \Pr(A \cap B) = \Pr(B \mid A) \Pr(A) \ .$$

Therefore

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \Pr(A)}{\Pr(B)} \ .$$

This is the celebrated *Bayes theorem* or *Bayes rule*. (Thomas Bayes lived in the 18th century. He studied in the University of Edinburgh and was a priest in Tunbridge Wells. He formulated a special case of the formula so some say that the complete formula is due to the 19th century French mathematician Laplace. But so many other things are named after Laplace that Bayes gets his credit for the formula.)

If we have disjoint events $A_1, A_2, \ldots$ covering the whole sample space (i.e., $A_n \cap A_m = \varnothing$ for $n \neq m$ and $\bigcup_{n=1}^{\infty} A_n = \Omega$) then by the law of total probability $\Pr(B) = \sum_{n=1}^{\infty} \Pr(B \mid A_n) \Pr(A_n)$. The Bayes formula can be written as

$$\Pr(A_i \mid B) = \frac{\Pr(B \mid A_i) \Pr(A_i)}{\sum_{n=1}^{\infty} \Pr(B \mid A_n) \Pr(A_n)} \ . \tag{3}$$

The immense importance of this formula for data analysis is based on the following interpretation.

Let A_n be disjoint hypotheses. The number $\Pr(B \mid A_i)$ is the probability of some B put forward by the hypothesis A_i. The value of $\Pr(B \mid A_i)$ usually follows from the nature of A_i. Different hypotheses give us different values.

The values $\Pr(A_i)$ are the probabilities we assign to the hypotheses. These probabilities are called *prior probabilities*. They represent our trust in the hypothesis *before* we have done any experiments or observations. Now supposed that B has occurred. We need to adjust our beliefs. Using (3) we can calculate *posterior* probabilities $\Pr(A_i \mid B)$, i.e., our trust in the hypotheses *after* B has occurred.

Sometimes we only need to compare $\Pr(A_i \mid B)$ between each other. We may then write the Bayes formula as

$$\Pr(A_i \mid B) \propto \Pr(B \mid A_i) \Pr(A_i)$$

(the sign $\propto$ reads "proportional to"). We have removed the denominator because it is the same anyway and would not matter for the comparison. To compare $\Pr(A_i \mid B)$ we can simply compare $p_i = \Pr(B \mid A_i) \Pr(A_i)$. If we want to calculate true values from the proportional values p_i, we need to *normalise* them:

$$\Pr(A_i \mid B) = \frac{p_i}{\sum_{n=1}^{\infty} p_n} \ .$$

After the normalisation the values sum up to 1 and become probabilities.

The following exercise also shows an important application of the Bayes formula.

Exercise 5. It is known that about two people in a thousand have a disease X. A radio-cardio-blood test has been developed to diagnose the disease. For a person who has the disease the test returns the positive verdict in 99% of cases. For a person who does not have the disease the test returns the positive verdict (this is called *false positive*) in 5% of cases. Bob has done the radio-cardio-blood test and has been diagnosed positive. What is the probability Bob actually has the disease? Contemplate the figures.

Hint: You need to use the Bayes formula and the total probability formula. Consider the events "has disease", "diagnosed positive" etc. Do not forget to transform all percentages into fractions of 1.

Answer for self-checking: About 3.8%.

Although the figures and the disease X are entirely notional, the situation is realistic. The result is so low because there is a significant false positive rate coupled with the large number of people who do not have the disease. One may want to reduce the sensitivity of the test to reduce the false positive rate. But this would probably reduce the true positive rate too, i.e., the number of people who actually have the disease and are diagnosed positive. This is not really what we want. Hence we have to live with 3.8%. It is better than the prior probability 0.2%.

Exercise 6. A team of medical researchers is investigating genetic causes of schizophrenia. They try to find genes linked to the disease. It is estimated that out of the total of 100,000 genes about 10 are linked.

The team studies cases of schizophrenia, looks for suspicious genes, and gathers evidence. After sufficient evidence has been collected, some genes are declared linked to schizophrenia. There is, of course, the possibility of an error. Given that a gene is related to schizophrenia, the probability to find it out is 60%. If a gene is not actually related to schizophrenia, the probability that it is declared related by mistake is 5%.

Calculate the probability that a gene declared linked to schizophrenia by the team is actually linked.

Answer for self-checking: About $1.2 \cdot 10^{-3}$. Yes, it is that low.

This example has been taken from paper [Ioa05], which produced quite a stir in the medical research community. The probability gets even lower if the team is subtly manipulating the data (e.g., by selectively including borderline cases; the diagnosis of schizophrenia is not clear cut and may often be disputed). See the paper for a discussion (and appreciate how the Bayesian framework makes things clearer[4]).

Note 2. In this problem and other problems like this it is particularly difficult to identify the probability space. Is it the set of possible universes where different genes are responsible for schizophrenia? That is not very intuitive. Some people prefer to speak about probabilities as degrees of belief instead. That resolves some logical problems but brings new ones. See Section 1.2.3 of [Bis06] for an interesting discussion. Book [Bis06] is particularly good in its treatment of the Bayesian approach.

5 Random Variables

A *random variable* X accepts values at random. We usually argue about a random variable in terms of its distribution, i.e., how often it takes certain values. We may wonder, for example, what the probability is that $X > 0$.

A more precise definition is that a random value $X(\omega)$ is a function on a sample space Ω. Different elementary outcomes lead to different values of X. The event $X > 0$ is really a set of elementary outcomes $\{\omega \mid X(\omega) > 0\} \subseteq \Omega$ and therefore we can speak of its probability.

A *discrete* random variable can take finitely many outcomes. Let a random variable Y be the result of rolling a die. It can take six values 1, 2, 3, 4,

[4] If you disagree with one calculation result in the paper, contact me.

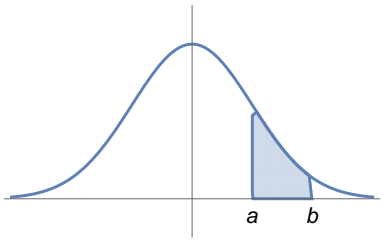

Figure 1: Probability that $a \leq Y \leq b$

5, and 6, each with probability 1/6. This can be written as $\Pr(Y = i) = 1/6$, where $i = 1, 2, 3, 4, 5, 6$. One can say that Y has a uniform distribution over six values $i = 1, 2, 3, 4, 5, 6$.

Some random variables take infinitely many outcomes. Let X be the height in centimetres of a person in the street. It has the finite range (certainly greater than 0 but less than 250) but within that range it can take many values. The distribution can no longer be described by enumerating all possible values with their probabilities.

A *continuous* random variable has a *density function* (sometimes called probability density function and abbreviated *pdf*), which can be used to describe the distribution. Let Y be a random variable taking real values and p_Y be its density. By definition $p_Y(x) \geq 0$ for all $x \in \mathbb{R}$ and the probability $\Pr(a \leq Y \leq b)$ is the integral $\int_a^b p_Y(x)dx$. The integral can be intuitively interpreted as the area bounded by the curve $y = p_Y(x)$ from above, the coordinate axis $y = 0$ from below, and the the lines $x = a$ and $x = b$ on the left and right; see Figure 1.

As an example consider the variable X with the uniform distribution on the segment $[0, 10]$. It has the density

$$p_X(x) = \begin{cases} 0, & \text{if } x < 0 \ ; \\ \frac{1}{10}, & \text{if } 0 \leq x \leq 10 \ ; \\ 0, & \text{if } x > 10 \ . \end{cases} \tag{4}$$

See Figure 2 for a plot. If an interval $[a, b]$ is inside $[0, 10]$, i.e., $0 \leq a < b \leq 10$, then $\Pr(a \leq X \leq b) = (b - a)/10$.

It is customary to describe distributions of random variables by means of distribution functions. A *distribution function* (sometimes called cumulative distribution function and abbreviated *cdf*) of a random variable Y is the

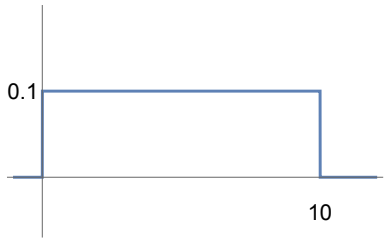

Figure 2: Density function for X

function

$$F_Y(x) = \Pr(Y \leq x) \ .$$

The distribution function can be used to calculate $\Pr(a < Y \leq b) = F_Y(b) - F_Y(a)$.

Exercise 7. Plot the distribution function for X with the density given by (4).

Exercise 8. (Slightly tricky.) Plot the graphs of the distribution functions of Y such that $\Pr(Y = i) = 1/6$, where $i = 1, 2, 3, 4, 5, 6$. *Hint:* The graph consists of steps.

A distribution function is always non-decreasing (if $a < b$ then $\Pr(X \leq a) \leq \Pr(X \leq b)$ and hence $F(a) \leq F(b)$). Its values are always between 0 and 1 (probabilities are always between 0 and 1). As x goes to $-\infty$, $F(x)$ goes to 0 and as x goes to $+\infty$, $F(x)$ goes to 1. In Exercises 7 and 8 both the functions actually reach 0 to the left of 0 and 1 to the right of 6 and 10, respectively.

Exercise 9. Let $p(x)$ be the density function of a random variable. What is the value of the integral $\int_{-\infty}^{+\infty} p(x)dx$? *Hint:* Very little knowledge of integration is required to answer this.

Two random variables X and Y are called *independent* if the events $X \in A$ and $X \in B$ are independent for all suitable[5] sets $A, B \subseteq \mathbb{R}$. In other words, X and Y are independent if

$$\Pr(X \in A \text{ and } Y \in B) = \Pr(X \in A) \Pr(Y \in B)$$

for all A and B.

[5]Technically speaking, we must exclude subsets $A \subseteq \mathbb{R}$ such that $\Pr(X \in A)$ cannot be measured. You will most certainly never meet such sets in practice.

6 Gaussian Distribution

Gaussian (or *normal*) distribution is very important. The Gaussian distribution with the mean μ and standard deviation σ (or variance σ^2, which means the same thing) has the density function

$$p_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \ ,$$

where $\pi \approx 3.14159$ and $e \approx 2.71828$ is Euler's constant. The graph of this function has a bell shape with the peak at μ. The statement that X has a Gaussian distribution with the mean μ and variance σ^2 can be written as $X \sim \mathcal{N}(\mu, \sigma^2)$.

The distribution function of a Gaussian distribution cannot be written explicitly in terms of standard functions of analysis.

The *standard normal distribution* has $\mu = 0$ and $\sigma = 1$. Its density is often denoted by ϕ (a Greek letter called "phi"):

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

and the distribution function by Φ (capital phi):

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \ .$$

There is no explicit formula for Φ. Old probability textbooks often contain extensive tables of values of Φ in a small font.

Exercise 10. One well-known finance textbook has a table of values of $\Phi(x)$ with negative values of x on the left page and positive values of x on the right page. Explain why one page is redundant. In other words, explain how we can calculate $\Phi(-x)$ if we know $\Phi(x)$. *Hint:* Plot the graph of the density function and mark the relevant areas.

A question about the distribution of an arbitrary Gaussian variable can be answered in terms of Φ using the following fact. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$ (this fraction is called the *standard score* or *Z-score* in statistics). (The fact can be shown using elementary properties of integrals, but I omit the proof here.)

Exercise 11. Let $Y \sim \mathcal{N}(2, 100)$. Express the probability that $Y \leq 12$ in terms of Φ. Express the probability that $2 \leq Y \leq 12$ in terms of Φ.

The importance of the Gaussian distribution for probability and statistics is based upon the central limit theorem. If $\xi_1, \xi_2, \ldots$ are independent random variables (ξ is a Greek letter called "xi") then under very general assumptions the average $\frac{1}{n}\sum_{i=1}^{n}\xi_n$ approximately has a Gaussian distribution for large n.

A value such as the height of a person is influenced by a large number of small independent factors (different genes (human height is a so called *polygenic trait*), childhood conditions including nutrition etc) and therefore is approximately Gaussian[6].

7 Expectation and Variance

The following parameters are very important for describing and analysing distributions.

The *expectation* (or *average* or *mean value*) of a random variable is defined as follows. If Y is discrete and takes values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$, respectively, then its expectation is

$$\mathrm{E}\, Y = \sum_{i=1}^{n} p_i x_i \ .$$

If Y is continuous and has density $p_Y(x)$, then its expectation is

$$\mathrm{E}\, Y = \int_{-\infty}^{+\infty} x p_Y(x) dx \ .$$

There is some similarity in the formulas. We can consider a grid of values x_i on $\mathbb{R}$ situated at the distance Δ from each other and approximate the integral by the sum $\sum x_i p_Y(x_i)\Delta$. Here $p_Y(x_i)\Delta$ is approximately the probability of Y being close to x_i. As a matter of fact, both the formulas are special cases of the *Lebesgue integral*.

The following properties of expectation are important:

1. The expectation of a constant is the constant. If a random variable X takes only one value c, then $\mathrm{E}\, X = c$.

2. Expectations are additive. For all random variables X and Y we have $\mathrm{E}(X + Y) = \mathrm{E}\, X + \mathrm{E}\, Y$.

[6] Well, nearly. There is one factor making a relatively big contribution, and that factor is the person's sex. If you consider men's and women's heights separately, you will get closer approximations by two different Gaussian distributions. Their combination is a *Gaussian mixture*; they play a very important part in statistics.

3. For a random variable X and a constant c we have $\mathrm{E}(cX) = c\,\mathrm{E}\,X$.

4. If random variables X and Y are independent, then $\mathrm{E}\,XY = \mathrm{E}\,X\,\mathrm{E}\,Y$.

For discrete variables these properties can be checked rather easily. For continuous this is a bit more difficult. I omit the proofs.

It is important that $\mathrm{E}\,XY = (\mathrm{E}\,X)(\mathrm{E}\,Y)$ only holds for independent random variables. For example, $\mathrm{E}\,X^2 = \mathrm{E}\,XX$ rarely equals $(\mathrm{E}\,X)^2$ since X is not independent of itself.

Knowing the expectation tells you what values *to expect* from X. However there are many properties of distributions that $\mathrm{E}\,X$ does not capture. For example, one may wonder if values of X are tightly grouped around $\mathrm{E}\,X$ or can deviate by a lot.

The variance answers this question. *The variance* of a random variable X is defined as

$$\mathrm{var}\,X = \mathrm{E}((X - \mathrm{E}\,X)^2) \ .$$

A small variance implies that the distribution is concentrated near its mean and a large variance means that it is spread around.

The *standard deviation* is often more intuitive:

$$\sigma_X = \sqrt{\mathrm{var}\,X} \ .$$

For the Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ with the mean of μ and standard deviation σ we have

$$\mathrm{E}\,X = \mu$$
$$\sigma_X = \sigma$$

as the names of the parameters suggest. (Note that this is not a trivial statement though. We defined μ and σ as some parameters with no particular meaning. Showing that they are actually the mean and the standard deviation requires complicated integration beyond the scope of this tutorial.)

Here are important properties of variances:

1. The variance of a constant is zero. If a random variable X is constant and takes only one value c, then $\mathrm{var}\,X = 0$.

2. For every random variable X we have $\mathrm{var}\,X = \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2$.

3. For every random variable X and a constant c we have $\mathrm{var}(cX) = c^2\,\mathrm{var}\,X$.

4. The variance is additive for independent variables. If random variables X and Y are independent, then $\operatorname{var}(X+Y) = \operatorname{var} X + \operatorname{var} Y$.

All these properties can be derived from the properties for expectations. Let us check the second one.

$$\operatorname{var} X = \mathrm{E}((X - \mathrm{E}\, X)^2)$$

(let us open up the brackets)

$$= \mathrm{E}(X^2 - 2X(\mathrm{E}\, X) + (\mathrm{E}\, X)^2)$$

(by the additivity of expectations)

$$= \mathrm{E}\, X^2 + \mathrm{E}(-2X(\mathrm{E}\, X)) + \mathrm{E}((\mathrm{E}\, X)^2)$$

(taking the multiplicative constant $(-2(\mathrm{E}\, X))$ out)

$$= \mathrm{E}\, X^2 - 2\,\mathrm{E}\, X\,\mathrm{E}\, X + \mathrm{E}((\mathrm{E}\, X)^2)$$

(recalling that the expectation of a constant is the constant)

$$\begin{aligned} &= \mathrm{E}\, X^2 - 2\,\mathrm{E}\, X\,\mathrm{E}\, X + (\mathrm{E}\, X)^2 \\ &= \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2 \ . \end{aligned}$$

Exercise 12. Prove the other properties in the same fashion.

References

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[Ioa05] J. P. A. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8), 2005.

[Ros06] S. M. Ross. *A first course in probability*. Pearson Prentice Hall, 7th edition, 2006.

[Tao] T. Tao. Lecture notes for Math 275A. Available at https://terrytao.wordpress.com/category/teaching/275a-probability-theory/.