Maths for Economics

10. Introduction to statistics

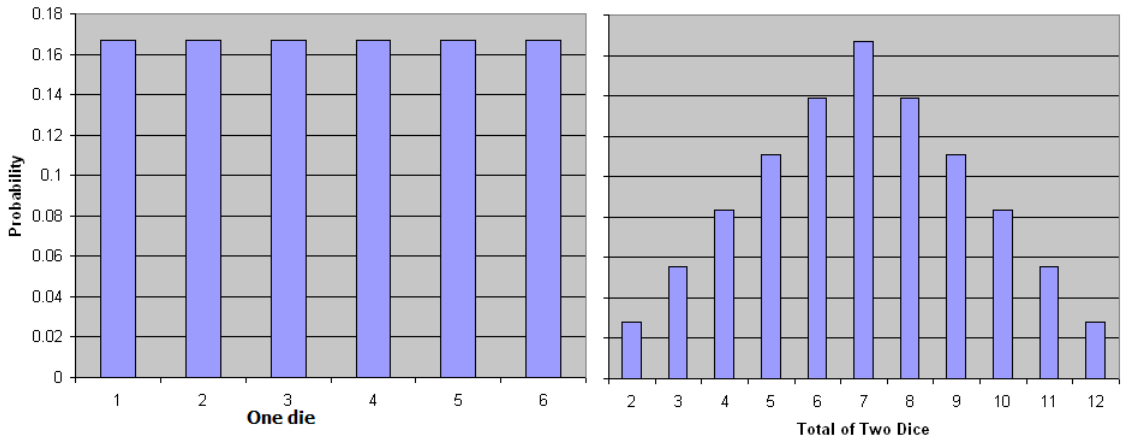
Vera L. te Velde

31. August 2015

First let's introduce a few more useful concepts.

1 Probability distributions

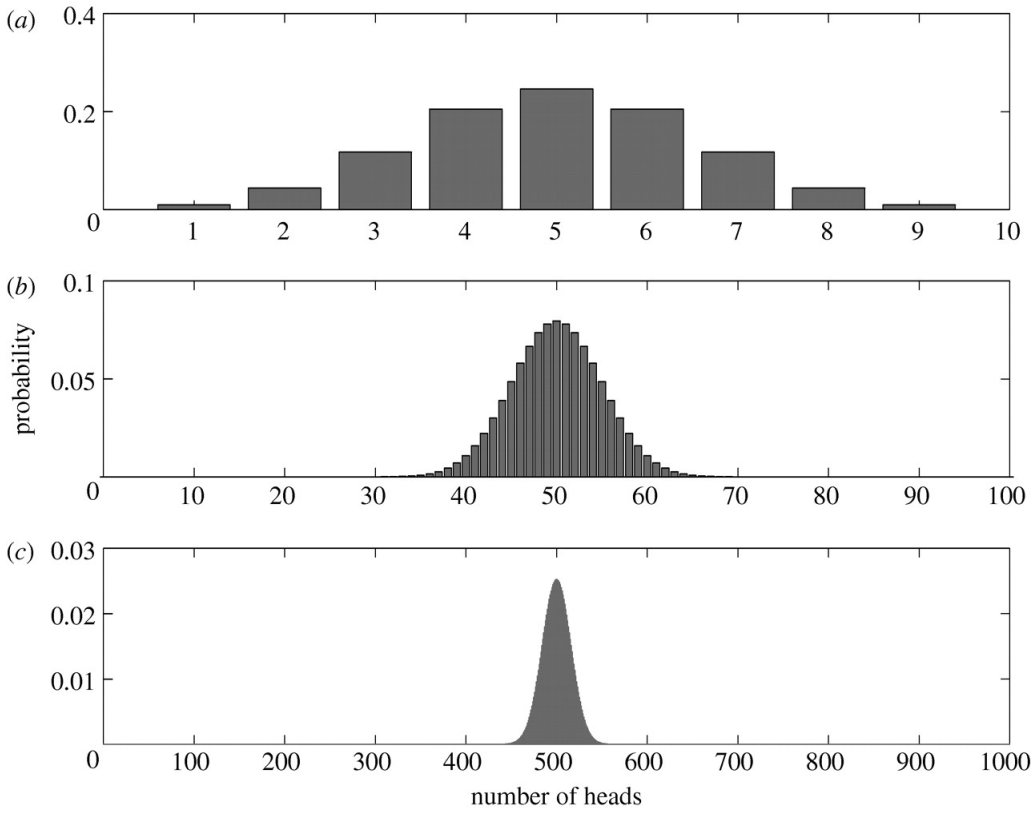
A probability distribution is simply a map from all possible outcomes to probabilities. If we can assign the outcomes numerical values, we can graph a probability distribution by using the outcome value on the horizontal axis and the probability of that outcome on the vertical axis.



These distribution graphs are another useful way to visualize many concepts in probability and statistics.

1.1 Law of large numbers

With probability distributions, we can visualize the law of large numbers that we learned about in order to understand probabilities as long-run frequencies. Consider the coin flip example. The probability distribution of the fraction of heads when we flip a coin n times looks like this, for different values of n:



As the number of repetitions gets very large, the fraction of heads is almost always very close to 50%. The exact number is still variable, but since a handful of heads or tails doesn't effect the overall total very much, the overall frequency stays close to 50%. We can therefore measure the true probability of a outcome by trying the same test many times and counting the overall frequency with which the event occurs.

2 Statistics

A *statistic* is some piece of information that provides some kind of summary information about a set of data. For example, an *average* is a statistic that tells you what number you can expect the data to be close to usually. Medians, modes, standard deviations, and variances are other common statistics.

2.1 Expected value

The **expected value** of a random event is simply the average outcome that we expect to happen. This relies on assigning numerical values to different outcomes. For example, if we say that flipping a heads is a "0" outcome, and flipping tails is a "1" outcome, the expected value of a coin flip is $0.5 \cdot 0 + 0.5 \cdot 1 = 0.5$. This pattern holds in general. If there

are n possible outcomes in the set X, with values $X = \{x_1, x_2, \cdots, x_n\}$, then

$$E(X) = P(x_1) \cdot x_1 + P(x_2) \cdot x_2 + \cdots + P(x_n) \cdot x_n = \sum_{i=1}^{n} P(x_i) \cdot x_i.$$

For example, the expected value of a die roll is $\frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$. If we flip two coins, there are 3 possible numbers of heads that can come up: 0, 1, and 2. These outcomes have probability .25, .5, and .25 respectively. Therefore, the expected number of heads is $.25 \cdot 0 + .5 \cdot 1 + .25 \cdot 2 = 1$. In three coin flips, the expected number is $.125 \cdot 0 + .375 \cdot 1 + .375 \cdot 2 + .125 \cdot 3 = 1.5$ (even though we can of course never get exactly 1.5 heads). And so on: for any number n coin flips, the expected number of heads is $n/2$.

On a probability distribution graph, the expected value will show up somewhere in the "middle" of all possible outcomes, weighted by probability. Looking at the example above, we can see that 3.5 shows up in the middle of the possible outcomes of a die roll.

2.2 Standard deviation and variance

The expected value of an outcome tells us where the "middle" of all outcomes is. But if all we know is the expected value, we still don't know how far away we can expect to be from that exact number. For example, no matter how many times you flip a coin, you expect that half will turn up heads. But if after hundreds of coin flips your running tally is far from one half, you should be worried that your coin isn't fair. After only a couple of coin flips, however, it isn't surprising to get a percentage of heads that is very far from the expected value.

The standard deviation or variance is what tells us how concerned to be if our measured outcome is a certain distance from the expected outcome. A high variance, or high standard deviation, random outcome is one that jumps all over the place, so it's not surprising to be far away from average. A low variance, or low standard deviation, event stays close to the average.

The **variance** of a random variable is the average squared distance from the actual outcome to the true outcome. If we roll a die, then 1/6 of the time we'll roll a 1, which is 2.5 from the expected value. 1/6 of the time we'll roll 2, which is 1.5 from the expected value. And so on. Squaring these values gives $Var(X) = \frac{1}{6} \cdot 2.5^2 + \frac{1}{6} \cdot 1.5^2 + \frac{1}{6} \cdot 0.5^2 + \frac{1}{6} \cdot 0.5^2 + \frac{1}{6} \cdot 1.5^2 + \frac{1}{6} \cdot 2.5^2 = 2.9$. The formula follows this pattern:

$$\text{Var}(X) = E((X - E(X))^2)$$

This formula is easy to remember as the average squared distance from the true outcome to the expected outcome. But a variation on this formula is often easier to calculate:

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2$$

For example, in the two coin flip example, the three possible values of numbers of heads are 0, 1, and 2, and the square of those are 0, 1, and 4. Using the same probabilities as

before, we have $E(X^2) = .25 \cdot 0 + .5 \cdot 1 + .25 \cdot 4 = 1.5$. And we already know that $E(X) = 1$, so that $(E(X))^2 = 1$, so $Var(X) = 1.5 - 1 = 0.5$.

The **standard deviation** of a random variable is simply the square root of the variance. This gives a better measure of the distance you should expect to be from the expected value, because the square root sort of undoes the squaring of distances in the formula for variance. In the die example, $\sqrt{2.9} = 1.7$, which is about halfway between the expected value and the farthest possible outcome. And for two coins, we should expect that the number of heads we get is approximately $\sqrt{0.5} = .71$ away from the expected number, 1.

$$\text{Stddev}(X) = \sqrt{\text{Var}(X)}$$

Look back at the graph of the probability distributions of the percentage of heads out of a given number of coin flips. Notice that as the number gets larger, the distribution gets concentrated in the middle. This is the same thing as the variance or standard deviation getting smaller. Indeed, the variance of this statistic (percentage heads) is given by $\frac{1}{4n}$, which gets smaller as n gets large.

3 Describing the meaning of measurements

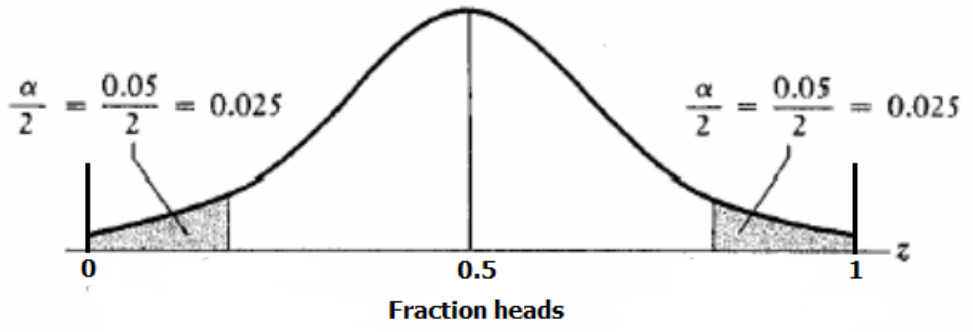
The concepts of expected value and variance tell us a lot about a distribution. In fact, if we run an experiment, we can convey our results simply by stating the average and standard deviation of the outcomes we observed.

For example, if we ask 100 people what their household income is, instead of providing the full list of data, we could simply say that the average is \$10,000 and the standard deviation is \$1,000. This tells us that incomes are clustered fairly closely around \$10,000 each. If instead we reported a standard deviation of \$8,000, this would tell us that there is a large degree of inequality in our polled population. It might be that a few people are very rich compared to everyone else, or it could be that some people are extremely poor and others are doing ok, but in any case, there must be a big mix in order to get a standard deviation so large.

3.1 Confidence intervals

Another way to summarize measurements is with confidence intervals. This tells us a little bit more than even the variance.

The confidence interval simply tells us the upper and lower bounds of a statistic that we can be 95% sure that we're between. 1-95%=5% is called α and can be set to whatever level you like, but 5% is the most common. We split up the extra 5% of possibilities equally above and below the interval: We are sure that 97.5% of the time the outcome will be less than the upper bound, and sure that 97.5% of the time the outcome will be greater than the lower bound.



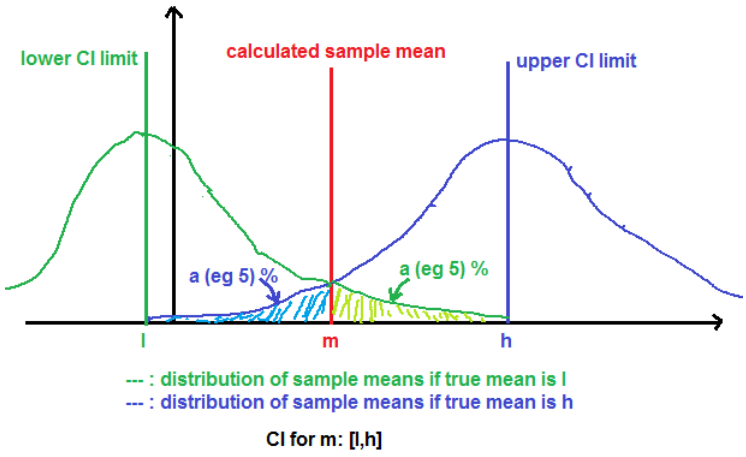
If we flip a coin 100 times, we can be 95% sure that the total number of heads will be between about 40 and 60, so the 95% confidence interval for the fraction of heads is $(.4, .6)$. If we flip it 1,000,000 times, the confidence interval is $(.499, .511)$. This is yet another illustration of the law of large numbers.

Confidence intervals don't have to be symmetric around the expected value, though. If we have a coin that comes up heads 90% of the time, then the expected fraction of heads is .9, but the 95% CI after 100 flips is approximately $(.83, .95)$.

3.2 Confidence intervals in reverse

This definition of confidence intervals lets you say what range of outcomes you can be reasonably sure of getting if you know what the probability distribution is. But even more often confidence intervals are used in the reverse: to say what the probability distribution likely is,is if you know the outcome.

For example, knowing that a coin will come up heads 50% of the time will let us calculate the range of values we can be 95% sure of getting if we flip it 100 times. But if we flip it 100 times without knowing whether the coin is fair, we can figure out what range of probabilities we can be 95% sure that the coin is truly driven by.



This graph might represent the average height of 100 people that we surveyed on the street. If the average height is 150cm, and if we know that the standard deviation is approximately 10cm, we can figure out that the lowest *average* height that might reasonably be measured as 150 is 130cm, and that the highest average that might reasonably be measured is 170cm. We would say that we are 95% sure that the true average height is between 130 and 170cm.

3.3 p-values

A last statistic that tells us information about a distribution without having to see the entire distribution is the p-value. This is used very widely when someone wants to distinguish their measured value from some comparison value. Especially, if someone measures a non-zero outcome value, they might use a p-value to show how sure they are that the underlying probability is not exactly zero.

Let's go back to the height example above. If we measure an average height of 150cm, how sure can we be that the average height in the entire population is actually at least 130cm? The p-value is the probability of measuring 150cm *given* that the true average height is 130cm.

Since we already calculated a 95% CI for the true height of $[130, 170]$, we know that the p-value for this comparison is 5%. The p-value of the statement "the true average height is no more than 50%" is approximately 0.5. And the p-value of the statement "the true average height is no more than 0" is 0.

4 Exercises

1. A bag of marbles contains 6 marbles, labeled 1, 1, 1, 2, 2, 2. You draw two marbles from the bag. Draw the probability distribution of the possible sums of the two marbles.

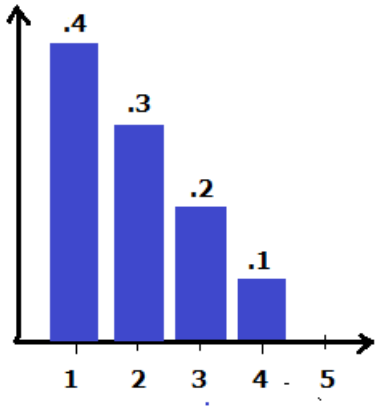
2. Now you repeat this game many times, writing down the sum you get each time. Draw the approximate probability distribution of the *fraction* of times you get a sum of 3, after 5 repetitions, 100 repetitions, and 1000 repetitions.

3. What is the expected value, variance, and standard deviation of each random outcome?

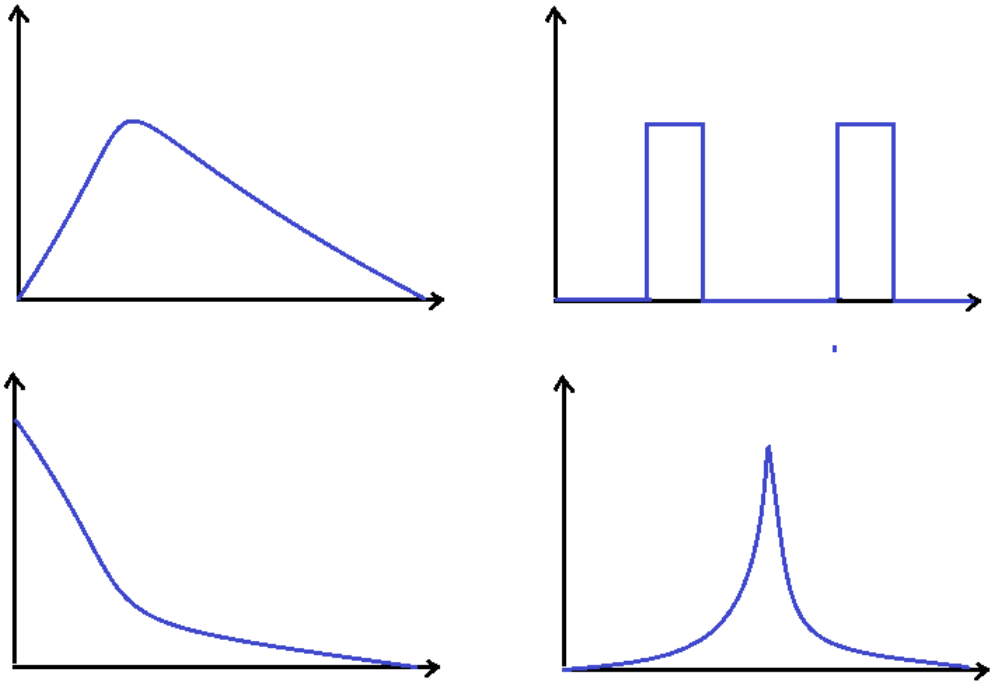
   (a) The sum of two rolled dice.

   (b) A natural number chosen randomly (each with equal probability) between 1 and 100.

   (c) The following distribution:



4. Draw approximately where the expected value, standard deviation, and 95



5. Some flips a coin 100 times and gets 55 heads. The 95% confidence interval for the true probability of flipping heads is calculated to be [.46, .65]. Illustrate the situation in a graph: What does the probability distribution of the fraction of heads look like if the true probability is .45 or .65?

6. A bag contains three marbles, all of which are either blue or red. Someone draws a red marble from the bag. What is the p-value for the statement "The number red marbles is greater than 1"?