What is probability?

In 2015 there were about 4 million babies born in the US, and 48.8% of the newborns were girls.

We write: P(newborn is a girl)$= 48.8\%$.

The probability of an event is defined as the proportion of times this event occurs in many repetitions.

This is the standard definition of probility. It requires that it is possible to repeat this chance experiment many times.

While interned during the Second World War, John Kerrich tossed a coin 10,000 times and observed 5,067 tosses resulting in heads.

What is probability?

The long-run interpretation of probability can make it difficult to interpret it for single event:
‘What I was wrong about this year’ by David Leonhardt (12/24/2017)

Sometimes people use a different interpretation:
‘The probability that my best friend calls today is 30%’.

Such a ‘subjective probability’ is not based on experiments, and different people may assign different subjective probabilities to the same event.

Four basic rules

Probabilities are always between 0 and 1. Computing probabilities comes down to applying a few basic rules repeatedly:

We saw that P(newborn is a girl)$= 48.8\%$.
Therefore P(newborn is a boy)$= 51.2\%$.

More formally, write A for an event, such as A='newborn is a girl'.

Complement rule: P(A does not occur) = 1$-$P(A)

Now let's look at a different chance experiment: Rolling a die.

Since each of its six faces is equally likely to come up, each has probability 1/6.

Rule for equally likely outcomes: If there are n possible outcomes and they are equally likely, then

$$\text{P(A)} = \frac{\text{number of outcomes in A}}{n}$$

Four basic rules

The next two rules make it possible to express probabilities of multiple events as those of the individual events.

A and B are *mutually exclusive* if they cannot occur at the same time.

As an example, we roll a die twice.
A= ⚀ on first roll, B= ⚅ on first roll, C= ⚀ on second roll

A and B are mutually exclusive, but A and C are not.

Addition rule: If A and B are mutually exclusive, then

$$\text{P(A or B)} = \text{P(A)} + \text{P(B)}$$

Two events are *independent* if knowing that one occurs does not change the probability that the other occurs.
B and C are independent, but A and B are not.

Multiplication rule: If A and B are independent, then

$$\text{P(A and B)} = \text{P(A)}\,\text{P(B)}$$

Four basic rules

Roll a die three times. What is P(at least one ⚅) ?

We could write 'at least one ⚅' as follows:

⚅ on the first roll or ⚅ on the second roll or ⚅ on the third roll.

But these three events are **not** mutually exclusive, so we can't use the addition rule.

Better way: Start with the complement rule.

P(at least one ⚅ in three rolls)

$= 1-$P(no ⚅ in three rolls)

$= 1-$P((no ⚅ in first roll) and (no ⚅ in second roll) and (no ⚅ in third roll))

$= 1-$P(no ⚅ in first roll)× P(no ⚅ in second roll)×P(no ⚅ in third roll)

$= 1 - \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}$

$= 41.1\%$

Conditional probability

Spam e-mail has a higher chance to contain the word ‘money’ than ham e-mail:

P(‘money’ in e-mail | spam)$= 8\%$, P(‘money’ in e-mail | ham)$= 1\%$.

The *conditional probability of B given A* is

$$\mathsf{P(B|A)} \;=\; \frac{\mathsf{P(A \text{ and } B)}}{\mathsf{P(A)}}$$

General multiplication rule: $\mathsf{P(A \text{ and } B) = P(A)\,P(B|A)}$

In the special case where A and B are independent: $\mathsf{P(A \text{ and } B) = P(A)\,P(B)}$

Conditional probability

Computing probabilities by total enumeration:
P(spam) = 20%. What is the probability that 'money' appears in an e-mail?

From data we know P(money | spam)= 8%, P(money | ham)= 1%.

Idea is to artificially introduce the event 'spam/ham'.

The event 'money appears in the e-mail' can be written as:

money appears and e-mail is spam   or   money appears and e-mail is ham

P(money appears)

=P(money and spam) + P(money and ham)

=P(money | spam) P(spam) + P(money | ham) P(ham)

$= 0.08 \times 0.2 + 0.01 \times 0.8$

$= 2.4\%$

Bayes' rule

From data we know P(money appears in e-mail | e-mail is spam)$= 8\%$, but what we need to build a spam filter is P(e-mail is spam | money appears in e-mail).

$$\text{P(B|A)} = \frac{\text{P(A and B)}}{\text{P(A)}} = \frac{\text{P(B and A)}}{\text{P(A)}} = \frac{\text{P(A|B) P(B)}}{\text{P(A)}}$$

$$\text{P(spam | money)} = \frac{\text{P(money | spam) P(spam)}}{\text{P(money)}} = \frac{0.08 \times 0.2}{0.024} = 67\%$$

Bayes' rule:

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

$$= \frac{\text{P(A|B) P(B)}}{\text{P(A|B)P(B)} + \text{P(A|not B) P(not B)}}$$

Bayesian analysis

The spam filter classifies e-mail as spam via a *Bayesian analysis*:

- Before examining the e-mail, there is a *prior probability* of 20% that it is spam.
- After examining the e-mail for certain keywords such as ‘money’, the filter updates this prior probability using Bayes’ rule to arrive at the *posterior probability* that the e-mail is spam.

Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result (‘false positive’).

Given that a person tests positive, what are the chances that he has the disease?

Know P(D)$= 1\%$, P(+|D)$= 95\%$, P(+|no D)$= 2\%$.

$$\text{Want } \mathsf{P}(\mathsf{D}|+) = \frac{\mathsf{P}(+|\mathsf{D})\ \mathsf{P}(\mathsf{D})}{\mathsf{P}(+)}$$

$$= \frac{\mathsf{P}(+|\mathsf{D})\ \mathsf{P}(\mathsf{D})}{\mathsf{P}(+|\mathsf{D})\ \mathsf{P}(\mathsf{D}) + \mathsf{P}(+|\text{no D})\ \mathsf{P}(\text{no D})}$$

$$= \frac{0.95\times0.01}{0.95\times0.01+0.02\times0.99} \;=\; 32.4\%$$

Case study: Warner’s randomized response model

What percentage of students have cheated during an exam in college?

Problem: Students may be too embarrassed to answer truthfully.

Randomization comes to the rescue:

We do a survey that first instructs students to toss a coin twice. If the student gets ‘tails’ on the first toss, then the student has to answer question 1, otherwise the student answers question 2.

Q1: Have you ever cheated on an exam in college?

Q2: Did you get ‘tails’ on the second toss?

So the answer will be partly random: We don’t know whether a ‘yes’ answer is due to the student cheating or to getting tails on the second toss. This should put the student at ease to answer truthfully.

Key point: While we don't know what an *individual* 'yes' means, we can estimate the proportion of cheaters using all the answers *collectively*:

$$\begin{aligned} \text{P(yes)} &= \text{P(yes and Q1)} + \text{P(yes and Q2)} \\ &= \text{P(yes} \mid \text{Q1) P(Q1)} + \text{P(yes} \mid \text{Q2) P(Q2)} \end{aligned}$$

Solve for $\text{P(yes} \mid \text{Q1)} = \frac{\text{P(yes)}-\text{P(yes} \mid \text{Q2) P(Q2)}}{\text{P(Q1)}}$

In one survey, 27 students answered 'yes' and 30 answered 'no'.

So we estimate $\text{P(yes)} = \frac{27}{27+30} = 47\%$ and get $\text{P(yes} \mid \text{Q1)} = \frac{0.47-0.5\times0.5}{0.5} = 44\%$