PROBABILITY APPROXIMATIONS

Janko Gravner (Univ. of California, Davis)

‘‘I think you’re begging the question,’’ said Haydock, ‘‘and I can see looming ahead one of those terrible exercises in probability where six men have white hats and six men have black hats and you have to work it out by mathematics how likely it is that the hats will get mixed up and in what proportion. If you start thinking about things like that, you would go round the bend. Let me assure you of that!’’ (Agatha Christie, The Mirror Crack’d)

Independence

Experiments are *independent* if the outcomes of some of them have no effect on the probabilities of the others.

For example, successive rolls of a fair die, or successive choice of a random card from a full deck *with replacement*, or successive lottery draws, or successive roulette outcomes are are all independent. But drawing cards *without replacement* gives rise to dependent experiments.

Fact: *Probabilities associated with independent experiments multiply*.

Example. Assume that 20% of adults in California are left-handed, and 10% are rich. What percentage of adults are right handed, rich, and their SS# is even? Assuming independence, $0.8 \cdot 0.1 \cdot 0.5 = 0.04$

Example. In an experiment success occurs with probability p. Repeat the experiment independently n times. Compute the probability of exactly k successes.

The answer is $\binom{n}{k} \cdot p^k (1-p)^{n-k}$.

Expectation.

Say each night you play 20 roulette games, each time betting a single dollar on red. Assume that the probability of red is $p = 18/38$. How many times do you win per night, on the average?

Average number of wins is p, so the average number of wins in 20 tries is $20p$.

This is an instance of an *expectation* of a random quantity. We write $E(\text{no. of wins in a single night}) = 20p$. Also $E(\text{no. of wins in 5 nights}) = 100p$.

Important but tricky fact: you can add expectations even though the random quantities are not independent!

Poisson Approximation.

Suppose you have n independent events, each of which has probability p. Assume that n is large, p is small and $\lambda = n \cdot p$ is of moderate size.

The probability that the total number k of of these events happens is

$$\begin{aligned}
&\binom{n}{k} \cdot p^k (1-p)^{n-k} \\
&= \frac{n(n-1)\dots(n-k+1)}{k!} \cdot \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&\approx \frac{\lambda^k}{k!} \cdot e^{-\lambda},
\end{aligned}$$

for $k = 0, 1, \dots$. The expression on the last line is called *Poisson distribution* (or *Law of Rare Events*). In particular, the probability that none of these events happens is about

$$e^{-\lambda}.$$

Important but *very* tricky fact: the above approximation holds even if the events are nearly independent.

History.

This law was derived as above by S. Poisson in 1837, in his book *Research into probabilities in judgements of civil and criminal matter, preceded by general rules for computing probabilities*. Its first use in statistics seems to be by von Bortkewitsch (1898), in his analysis of the number of Prussian soldiers killed each year by horses' kicks.

Montreal Gazette, September 10, 1981.

Boston (UPI) -- Lottery officials say that there is 1 chance in 100 million that the same four digit lottery numbers would be drawn in Massachusetts and New Hampshire on the same night. That's just what happened Tuesday.

The number 8092 came up, paying $5,841 in Massachusetts and $4,500 in New Hampshire. ''There is a 1-in 10,000 chance of any four--digit number being drawn at any time,'' Massachusetts Lottery Commission official David Ellis said, ''but the odds of it happening with two states at any one time is just fantastic,'' he said.

Assuming daily drawings for three years, what really are the odds? Here $n = 1,095$ and $p = 1/10,000$ so $\lambda = 0.1095$ and the probability of at least one such occurrence is about $1 - e^{-0.1095} \approx 0.104$.

The Birthday Problem, revisited.

Assume that a year has d days and k people. The number of unordered pairs of people is $n = k(k-1)/2$. For each fixed pair, the probability they share a birthday is $p = 1/d$. This gives

$$\lambda = \frac{k(k-1)}{2d} \approx \frac{k^2}{2d}$$

Therefore, the probability that no pair shares a birthday is about

$$e^{-k^2/(2d)}$$

Set this equal to 0.5 and solve for k to get

$$k = \sqrt{2 \ln 2 \cdot d} \approx 1.1774 \cdot \sqrt{d}.$$

This gives about 22.5 for $d = 365$.

How many people do we need to be 99% sure of a duplicate birthday?

Answers: 60.

Boston Evening Globe, February 6, 1978.

[...] During the [Massachusetts Lottery's] 22--months existence, no winning number has ever been repeated. Hughes, the expert, doesn't expect to see duplicate winnings until about half of 10,000 possibilities have been exhausted.

If the year had $d = 10,000$ days, and $k = 660$ people were chosen at random, then $\lambda = 660^2/20,000 = 21.78$ and the probability of no duplicate birthdays would be about $3.5 \cdot 10^{-10}$!

If fact, after a more careful look, it turned out that there *were* repetitions.

The Birthday Problem, revisited again.

How many people do we need for a significant chance that 3 will share a birthday?

Now the number of triples is $\binom{k}{3} = k(k-1)(k-2)/6 \approx k^3/6$ and $p = 1/d^2$, so that the probability that none of the triples share a birthday is about

$$e^{-k^3/(6 \cdot d^2)}.$$

Setting this equal to 0.5, and solving for k gives

$$k = \sqrt[3]{6 \ln 2 \cdot d^2}.$$

which is about 0.5 when $k = 82$.

The Hat Check Problem.

A large company (of 10,000 employees, say) has a scheme according to which each employee buys a Christmas gift, gifts are then scrambled, put in a large container, and finally each employee gets a random gift from the container. What is the probability that someone gets his or her own gift?

Here $n = 10,000$ and $p = 1/n$, so the approximate probability is simply $1 - e^{-1} \approx 0.632$.

The Socks Problem.

Assume that you have n (say 100) different pairs of socks in a drawer. Assume that you can take out a specified number of socks at random and your object is to get at least one matching pair.

Clearly, if you take out $n+1 = 101$ socks you will be sure to succeed. But this is clearly an overkill. If you select m socks, the probability that a particular pair is among them is

$$p = \frac{\binom{m}{2}}{\binom{2n}{2}} = \frac{m(m-1)}{2n(2n-1)} \approx \frac{m^2}{4n^2}.$$

Therefore, the number of pairs among selected socks is approximately Poisson with $\lambda = np = m^2/(4n)$. For example, if you wish to suceed with 99% probability, select $\sqrt{4n \ln 100} \approx 43$ socks.

Number of successes.

Example 1. Flip a fair coin 100 times. You can expect about 50 heads. What is the probability that you will get more than 55 heads?

Example 2. Playing the roulette, you plan to bet one dollar on red for 200 times. What is the probability you will be ahead at the end?

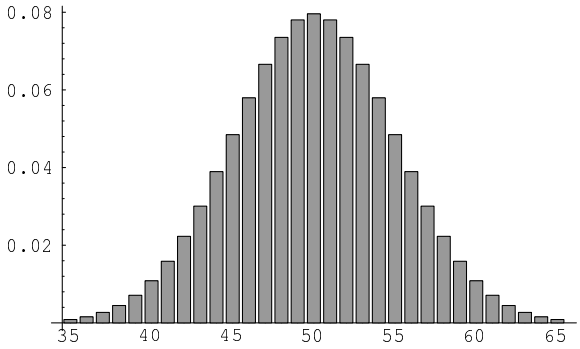
Assume again that there you are performing an experiment independently n times, and success occurs each time with probability p. The number of successes S has the following probabilities:

$$P(S = k) = \binom{n}{k} \cdot p^k (1-p)^{n-k}.$$

The answer to the first example then is

$$P(S > 55) = \sum_{k=56}^{100} \binom{100}{k} \cdot \frac{1}{2^{100}},$$

a little unpleasant to compute. But if you plot the probabilities, you get



which suggests an approximation with a bell–shaped curve and then the huge sums are really close to some integral. This, the most famous of all probability approximations, was first figured out by de Moivre in 1718.

Central limit theorem.

The bell–shaped curve is the one from *normal distribution*

$$\varphi(s) = \frac{1}{\sqrt{2\pi}} e^{-s^2/2},$$

and the approximation works as follows:

$$P\left(\frac{S - np}{\sqrt{np(1-p)}} \leq x \right) \approx \Phi(z) = \int_{-\infty}^{x} \varphi(s)\, ds.$$

Example 1. Here $n = 100$, $p = 1/2$ and $S \leq 55$ occurs exatly when $(S - np)/\sqrt{np(1-p)} \leq (55 - 50)/5 = 1$, so $x = 1$. So, $P(S \leq 54)$ is about $\Phi(1) \approx 0.8413$ (these numbers are obtained from tables in statistics books), and finally $P(S > 55) \approx 0.1587$.

By a similar computation, $P(S > 60) \approx 0.0228$.

Example 2. In order to win, you have to have more successes than failures, i.e., S has to be larger than 100. Since $n = 200$, $p = 9/19$, $x = (100 - np)/\sqrt{np(1-p)} \approx 0.75$, we get

$$P(S \leq 100) \approx \Phi(0.75) \approx 0.7734,$$

so the probability $P(S > 100)$ that you will be ahead is less than 23%. After 1000 games it drops down to 5%.