# Probability I — Random Processes

Prof. Ned Wingreen

MOL 410/510

## Tossing a coin – the Binomial Distribution

There are two possible outcomes, heads or tails, which we label as H and T. One toss is H or T, an "event" or "sample point." Multiple tosses: HTH... or HHH..., etc.

$\Omega = \{$all possible outcomes$\}$, which we call the "sample space."

Probabilities are assigned to each outcome in the sample space. By convention, the sum of probabilities of all possible outcomes is 1. So for one toss:

$$P(H) + P(T) = 1$$

For a fair coin, $P(H) = P(T) = 1/2$, and for a biased coin $P(H) \neq P(T)$, e.g. $P(H) = 3/5, P(T) = 2/5$.

What does it mean to say $P(H) = 1/2$? For a large number of tosses, 1/2 will be heads. (We assume each toss is independent.)

What is the probability of getting heads twice in two tosses of a fair coin? For a fair coin, all outcomes are equally likely:

$$\Omega = \{\text{HH, HT, TH, TT}\}$$
$$P(\text{HH}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

### What is the probability of $n$ heads in $N$ tosses?

- Fair coin: If all outcomes are equally likely,

  $$P(\text{event}) = \frac{\text{number of ways event can occur}}{\text{total number of outcomes}}$$

  The total number of outcomes is $2^N$, and the number of ways to achieve $n$ heads is

  $$\frac{N!}{(N-n)!n!} = \binom{N}{n}$$

  since $N!/(N-n)!$ is the number of ways to arrange $n$ heads labeled $1, ..., n$ among $N$ events, and $n!$ is the number of permutations of $n$ heads (that is, the factor by which we have overcounted).

  So

  $$P(n) = \binom{N}{n} \frac{1}{2^N}.$$

- Biased coin: $P(H) = p \neq 1/2$. All outcomes are *not* equally likely. The probability of any particular outcome of heads, e.g.

  $$P(\underbrace{\text{HHHHHHH....H}}_{n}\underbrace{\text{TTT...T}}_{N-n}) = \underbrace{p \cdot p \cdot p \cdot p ... \cdot p}_{n}\underbrace{(1-p) \cdot (1-p) \cdot (1-p) ... \cdot (1-p)}_{N-n},$$

  so that

  $$P_{\text{Binomial}}(n) = \binom{N}{n} p^n (1-p)^{N-n}.$$

  This is the "binomial distribution."

# Poisson distribution

This is the limit of the binomial distribution for $p << 1$. For instance: what is the probability of finding zero resistant mutants in a colony of $10^9$ cells, each with a probability of $2 \times 10^{-9}$ of having mutated to resistance:

$$N = 10^9$$
$$p = 2 \times 10^{-9}$$

so that for $\lambda$, the expected number of mutants, we have

$$\lambda = Np = 2.$$

The Poisson distribution expresses the limit of the binomial distribution when $p$ is small and $N$ is large, so that $\lambda = Np$ is moderate. As above, we have

$$P(n) = \binom{N}{n} p^n (1-p)^{N-n} = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}.$$

Consider the first few terms

$$P(0) = \binom{N}{0} (1-p)^N$$
$$\text{Use } (1-p)^N = e^{\log(1-p)^N} = e^{N \log(1-p)} \approx e^{N(-p)} = e^{-Np} = e^{-\lambda}$$
$$P(0) = e^{-\lambda}$$
$$P(1) = \binom{N}{1} (1-p)^{N-1} \approx Npe^{-Np} = \lambda e^{-\lambda}$$

Using Sterling's formula $N! \sim \sqrt{2\pi} N^{N+\frac{1}{2}} e^{-N}$ yields

$$\begin{aligned} P(n) &\approx \frac{N^N}{(N-n)^{N-n}} \sqrt{\frac{N}{N-n}} \frac{e^{-n}}{n!} p^n (1-p)^{N-n} \\ &\approx \left(\frac{N}{N-n}\right)^N N^n \frac{e^{-n}}{n!} p^n (1-p)^{N-n}, \end{aligned}$$

since $\sqrt{\frac{N}{N-n}} \approx 1$. We now observe that

$$\left(\frac{N}{N-n}\right)^N = \left(\frac{1}{1-\frac{n}{N}}\right)^N \approx \left(1 + \frac{n}{N}\right)^N \approx e^{\frac{n}{N}N} = e^n.$$

So

$$P(n) \approx e^n (pN)^n \frac{e^{-n}}{n!}(1-p)^{N-n} = (pN)^n \frac{1}{n!}(1-p)^{N-n}.$$

In a similar way,

$$(1-p)^{N-n} \approx e^{-p(N-n)} \approx e^{-pN} = e^{-\lambda},$$

so that

$$P_{\text{Poisson}}(n) = e^{-\lambda}\frac{\lambda^n}{n!},$$

the Poisson distribution.

Since $\lambda$ is the expected number of resistant mutants, it is easy to find the probability of no mutants:

$$P(0) = e^{-\lambda} = e^{-2} \approx 0.135.$$

# Hypergeometric distribution

The hypergeometric distribution describes the probability of finding $k$ "special elements" in a randomly chosen group of $r$ elements. For instance, what is the random probability that in a gene expression experiment 20 of the top 100 upregulated genes will be involved in cell cycle if 200 out of 5000 total genes are cell-cycle related?

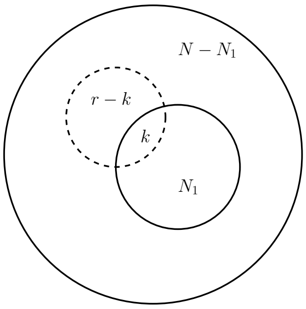

Figure 1: Hypergeometric distribution schematic

$$\begin{aligned} N &= \text{total elements (5000)} \\ N_1 &= \text{special elements (200)} \\ r &= \text{number chosen (100)} \\ k &= \text{number special (20)} \end{aligned}$$

so that $r-k$ is the number of non-special elements. We seek

$$P(k) = \frac{\text{ways of choosing exactly } k \text{ special and } r-k \text{ non-special elements}}{\text{total ways of choosing } r \text{ elements}}$$

Special elements can be chosen in $\binom{N_1}{k}$ ways, non-special elements can be chosen in $\binom{N-N_1}{r-k}$ ways, while total ways of choosing $r$ elements is $\binom{N}{r}$.

Since any choice of $k$ special elements can be combined with any choice of $r-k$ non-special elements, we get

$$P_{\text{Hypergeometric}}(k) = \frac{\binom{N_1}{k}\binom{N-N_1}{r-k}}{\binom{N}{r}}$$

which can be rearranged to yield the canonical expression

$$\begin{aligned} P(k) &= \frac{N_1!}{k!(N_1-k)!}\frac{(N-N_1)!}{(r-k)!(N-N_1-r+k)!}\frac{r!(N-r)!}{N!} \\ &= \frac{\binom{r}{k}\binom{N-r}{N_1-k}}{\binom{N}{N_1}} \end{aligned}$$

For our example:

$$P(k=20) = \frac{\binom{100}{20}\binom{4900}{180}}{\binom{5000}{200}}$$

How surprised should one be by finding 20 out of 100 genes involved in the cell cycle? Perhaps it's better to ask for the probability of seeing 20 *or more* out of 100 genes:

$$P_+(20) = \sum_{k=20}^{100} P(k) \approx 1.37 \times 10^{-9}$$

So it is very unlikely to find 20 cell cycle genes out of the top 100 by random chance. This probability is usually reported as a so-called "p-value" $= -\log_{10} P_+(k)$. In our case, the p-value is 8.86.

## Null models and "confidence"

The above is an example of the use of a null model in statistics. From the data, the best we can do is to report our confidence that the data is not just a random outcome. So we assume a null model (in this case, randomly picked genes) and calculate the probability of the occurrence by chance of an outcome at least as biased as the observed data. Usually, if this probability is $< 5\%$, we report the result as significant. But beware, if we are testing multiple hypotheses (e.g. not just overlap with cell-cycle genes, but with other gene categories as well) then it becomes more likely that at least one apparent correlation will occur just by chance.