Lecture 2: Discrete Probability Distributions

IB Paper 7: Probability and Statistics

Carl Edward Rasmussen

Department of Engineering, University of Cambridge

January 25th, 2019

Two useful rules for computing discrete probabilities

When a set of outcomes are all equally likely, the probability of an event is the number of outcomes consistent with the event divided with the total number of possible outcomes.

Example: What is the probability that someone has their birthday on Dec 31st? It's 1/365.

Example: What is the probability that someone has their birthday in April? Roughly 1/12.

The probability that two *independent* events happen simultaneously is given by the product of the probabilities.

Example: What is the probability that two people both have their birthday in January? It's 1/144.

Example: What is the probability that two people both have their birthday in the same month? It's roughly 1/12. There a several ways to think about this.

Computing the probability of more complicated events

What is the smallest number of people you need to ensure that the probability that 2 people have identical birthdays is at least a half?

First, work out the probability q_n that with n people, none share birthdays.

Arbitrarily, order the people.

The birthday of the first person is irrelevant.

The second person must have a birthday different from the first, $p_2 = 364/365$.

The third person must have a birthday different from the first and the second, $p_3 = 363/365$.

Generally, for the i'th person, $p_i = (366 - i)/365$.

Because of independence, $q_n = \prod_{i=1}^{n} p_i = \frac{365}{365}\frac{364}{365}\cdots\frac{366-n}{365} = \frac{365!}{365^n(365-n)!}$.

We're interested in the smallest value of n for which $q_n < 1/2$. A (tedious) calculation shows $n = 23$, $q_{23} = 0.493$.

The Bernoulli Distribution

The simplest distribution is the *Bernoulli* distribution:

$$X \sim \text{Ber}(p), \quad \text{where} \quad 0 \leqslant p \leqslant 1.$$

The random variable X is binary, and takes on the value $p(X = 1) = p$ and $p(X = 0) = 1 - p$.

The probability is sometimes written concisely as

$$p(X = x) = p^x (1 - p)^{(1-x)}.$$

Example: In a simple climate model, you might have the random variable X indicating whether it rains on a particular day. You could then have $X \sim \text{Ber}(p)$, where different values of p would be used for different areas.

Warning: on this page the symbol p (intentionally) has two different meanings: probability or a parameter. Underline all the occurrences where p is a parameter.

Characterising the Bernoulli Distribution

The *mean* or *expectation* of the Bernoulli is

$$\mathbb{E}[X] = \sum_{x \in X} x\, p(x) = 0 \times (1-p) + 1 \times p = p,$$

i.e., the mean of Bernoulli with parameter p is equal to p itself.

The *variance* of a distribution measures how much the outcomes "spread". The variance is defined as

$$\mathbb{V}[X] = \mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{x \in X} (x - \mathbb{E}[X])^2 p(x).$$

For the Bernoulli random variable:

$$\mathbb{V}[X] = \sum_{x \in X} (x-p)^2 p(x) = p^2(1-p) + (1-p)^2 p = ((1-p)+p)p(1-p) = p(1-p).$$

Thus, the maximum variance of the Bernoulli is $1/4$, attained when $p = 1/2$. Does this seem reasonable?

The *entropy* of the Bernoulli $-\mathbb{E}[\log(p)] = -p\log(p) - (1-p)\log(1-p)$ is maximally 1 bit, attained when $p = 1/2$.

Variance and Moments

The variance is always non-negative.

It is sometimes useful to write the variance as

$$\begin{aligned}\mathbb{V}[X] &= \sum_{x \in X} \big(x - \mathbb{E}[X]\big)^2 p(x) = \sum_{x \in X} \big(x^2 + \mathbb{E}[X]^2 - 2x\mathbb{E}[X]\big)p(x) \\ &= \mathbb{E}[X^2] + \mathbb{E}[X]^2 - 2\mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2.\end{aligned}$$

Here, $\mathbb{E}[X^2]$ is called the *second moment*. Similarly, $\mathbb{E}[X]$ is called the *first moment*. The variance is also called the *central second moment*.

Example: Verify the above rule for Bernoulli. The second moment is $0 \times (1-p) + 1^2 p = p$. The square of the first moment is p^2. The variance is thus $p - p^2 = p(1-p)$.

The Binomial Distribution

The *Binomial* is the number of successes r in n independent, identical Bernoulli trials, each with probability of success p.

We write

$$X \sim B(n, p), \quad \text{where} \quad n = 0, 1, 2, \ldots, \quad \text{and} \quad 0 \leqslant p \leqslant 1.$$

The value of the probability function is

$$p(X = r) = {}_nC_r \, p^r (1 - p)^{(n-r)}.$$

The notation ${}_nC_r$ are the *combinations* n choose r, the number of ways one can choose r out of n when the order doesn't matter:

$${}_nC_r = \binom{n}{r} = \frac{n!}{(n-r)!r!}.$$

The Bernoulli is a special case of the Binomial: $\text{Ber}(p) = B(1, p)$.

Refresher on Permutations and Combinations

If you don't remember this, refer to part 1A math notes for more details.

The number of ways in which you can *permute* r elements from a total of n (when order *is* important) is

$$ {}_nP_r = \frac{n!}{(n-r)!} = n(n-1)(n-2)\cdots(n-r+1), $$

since, for the first item you can choose between n, for the second $n-1$ and so on.

The number of ways in which you can *choose* r elements from a total of n (when order *is not* important) is

$$ {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{(n-r)!r!}, $$

as there are $r!$ orderings of the *permutations* which gives rise to the same *choices*.

Warning: $n!$ is difficult to compute with for large n. Stirling's approximation

$$ n! \simeq \sqrt{2\pi n}\exp(-n)n^n. $$

Binomial Example

A semiconductor factory produces CPUs on silicon wafers, 100 CPU's on each wafer. Some of the CPUs don't work. Cutting the chips from the wafer is expensive so wafers with many failing units are discarded.

Assume that the probability of a CPU failing is independent and $p = 0.05$. What is the probability $p(\geqslant 5)$ that 5 or more CPUs fail?

Answer: Let the random variable X be the number of failures on a wafer:

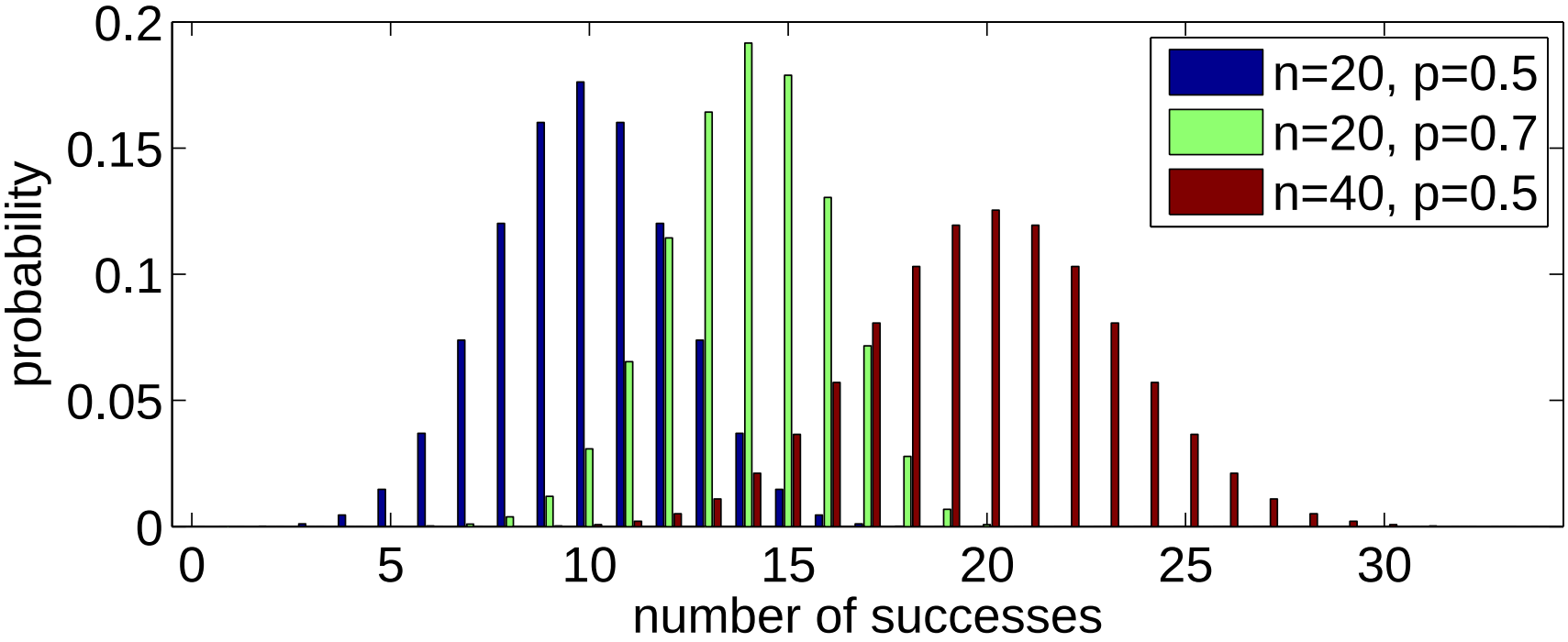
$$p(X \geqslant 5) = 1 - p(X \leqslant 4) = 1 - \sum_{i=0}^{4} B(i|n,p) = 1 - \sum_{i=0}^{4} {}_nC_i \, p^i (1-p)^{100-i}$$
$$\simeq 1 - 0.006 - 0.031 - 0.081 - 0.140 - 0.178 = 0.564,$$

where the notation $B(i|n,p)$ is the probability of getting i successes in n trials in a Binomial with probability p.

This shows that just over half the wafers will have 5 failures or more.

Some Binomial Distributions

Below is an illustration of the Binomial probability distribution for various settings of the parameters n and p.



Characterising the Binomial Distribution

The expectation of the Binomial is

$$\begin{aligned}\mathbb{E}[X] &= \sum_{r=0}^{n} r\, p(X=r) = \sum_{r=0}^{n} r\, {}_nC_r\, p^r (1-p)^{n-r} = \sum_{r=1}^{n} \frac{rn!}{(n-r)!r!} p^r (1-p)^{n-r} \\ &= np \sum_{r=1}^{n} \frac{(n-1)!}{(n-r)!(r-1)!} p^{r-1} (1-p)^{n-r} \\ &= np \sum_{\tilde{r}=0}^{\tilde{n}} \frac{\tilde{n}!}{(\tilde{n}-\tilde{r})!\tilde{r}!} p^{\tilde{r}} (1-p)^{\tilde{n}-\tilde{r}} = np,\end{aligned}$$

where $\tilde{n} = n-1$ and $\tilde{r} = r-1$, and using the fact that the Binomial normalizes to one.

In fact, the result is not surprising, since the Binomial gives the number of successes in n independent Bernoulli trials; the Bernoulli has expectation p.

Variance of the Binomial

We can compute the variance in an analogous way to the expectation, by doing the substitution twice.

The result is

$$\mathbb{V}[X] = np(1-p).$$

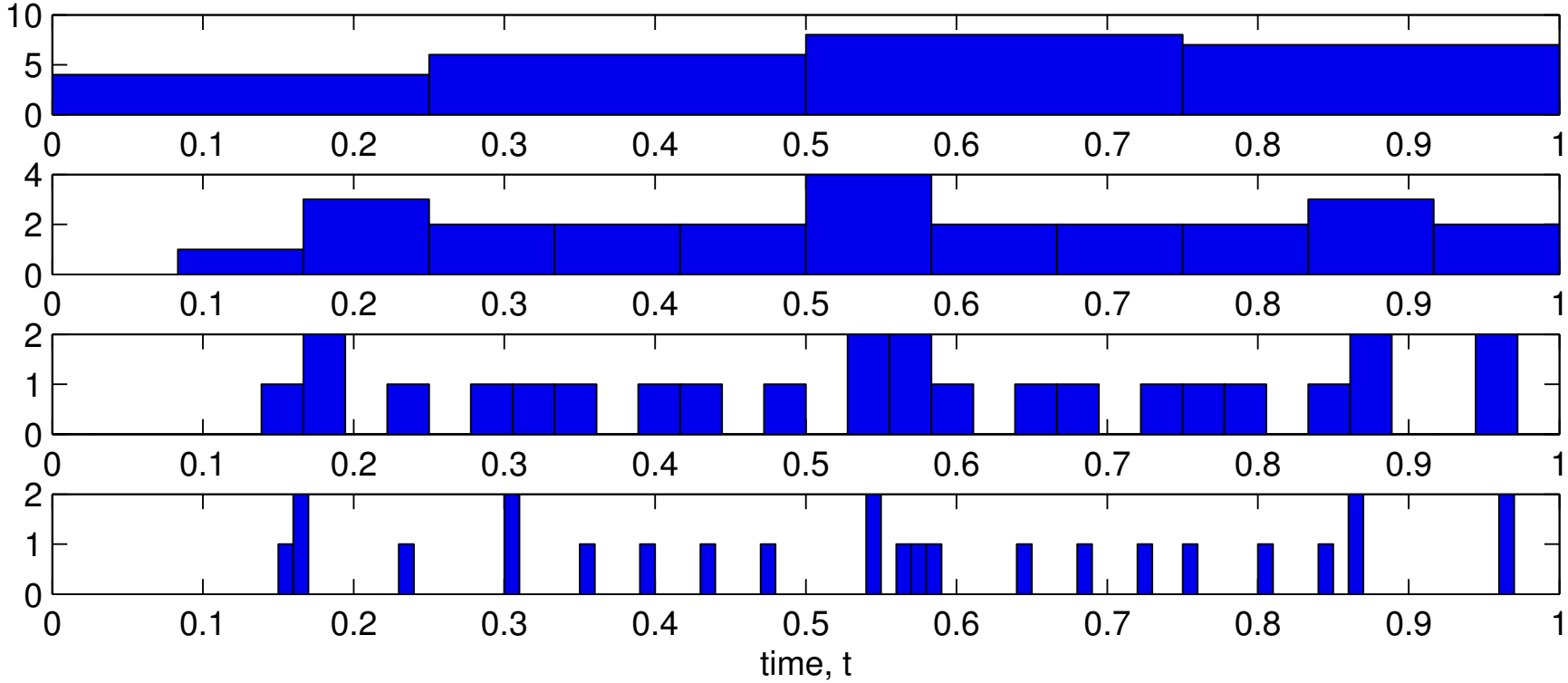
These are instances of two general rules which we will see later:

When you add *independent* variables

- the *means* add up, and
- the *variances* add up.

Independent Event Arrivals

Imagine phone calls arriving randomly and independently to a phone exchange, with an *average rate* (or *intensity*) λ.



Above are 4 histograms of the same 25 arrivals at different time resolutions.

What is the probability distribution over 'number of arrivals per second'?

Note: as the histogram bins get smaller, the probability of multiple arrivals in a bin decreases and *each bin tends to a Bernoulli event*.

Examining a Sequence of Bernoulli Events

Therefore, we look at getting r arrivals in a Binomial $B(n, p)$, where the number of bins n grows.

When the bins get smaller, the probability of an arrival within a bin falls *proportional to the bin size*, $p = \lambda/n$.

So, we want: $p(X = r) = B(n, \lambda/n)$ as $n \to \infty$.

$$\begin{aligned} p(X = r) &= \lim_{n\to\infty} B(n, \lambda/n) = \lim_{n\to\infty} \frac{n!}{(n-r)!r!}\left(\frac{\lambda}{n}\right)^r\left(1-\frac{\lambda}{n}\right)^{n-r} \\ &= \lim_{n\to\infty} \underbrace{\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-r+1}{n}}_{=1}\left(\frac{\lambda^r}{r!}\right)\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{=\exp(-\lambda)}\underbrace{\left(1-\frac{\lambda}{n}\right)^{-r}}_{=1} \\ &= \frac{\lambda^r \exp(-\lambda)}{r!} \end{aligned}$$

This is called the *Poisson* distribution, with intensity λ.

The Poisson Distribution

If events happen randomly at independent times, with an average intensity (number of events per unit time) λ, then the probability that a certain number of events r within a time interval is Poisson distributed, $\mathrm{Po}(\lambda)$.

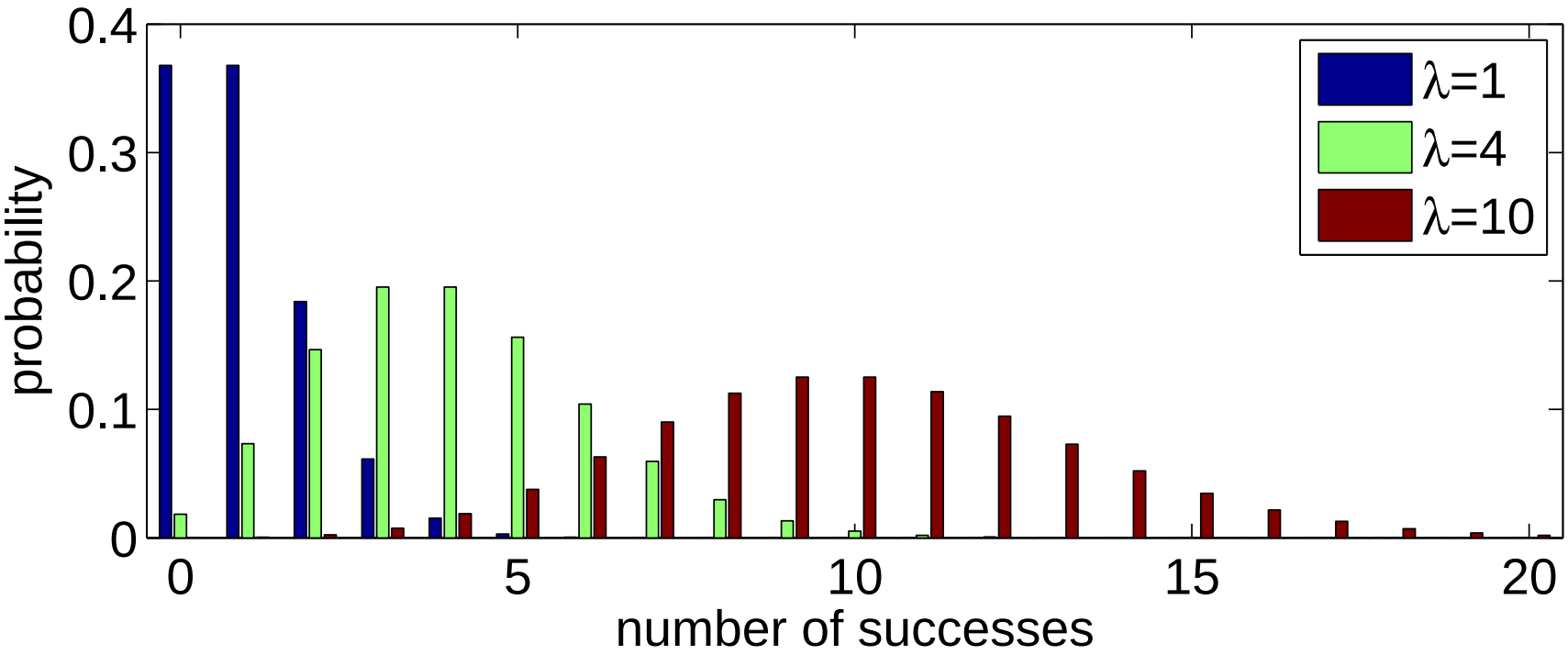
$$p(X = r) = \frac{\lambda^r \exp(-\lambda)}{r!}, \quad \text{where} \quad \lambda \geqslant 0.$$

Examples: radioactive decay, network traffic, customer arrival, etc.

The expectation and variance of the Poisson are both $\mathbb{E}[X] = \mathbb{V}[X] = \lambda$, which can be derived from the limiting Binomial.

Some Poisson Distributions

Below are some Poisson distributions with different intensities



Because Binomial distributions with large n are tedious to compute with, one can approximate the Binomial with the Poisson with the same mean, i.e. $B(n,p) \simeq Po(np)$. The approximation is good when n is large (say $n > 50$) and p small (say $p < 0.1$).