QCB 408 / 508 – Notes on Week 3

Student

2020-03-01

Summary

- Models for RNA-seq data
- Two-step process
- Negative Binomial
- Facts about Random Variables

Models for RNA-seq data

Say we are given m genes (indexed by i) and n observations (indexed by j). These observations could be at any scale – cells, samples, organisms, etc. – but we will assume that they all come from the same biological condition (i.e., the same statistical population). We observe Y_{ij} RNA-seq read counts for gene i in observation j. Y_{ij} is a random variable, and in this week's lectures we saw two alternative ways to model these read counts.

Two-step process

For a gene i, let a_i represent the true proportion of mRNA transcript counts that are from gene i. (Note that because we are considering only a single population, there is only one a_i for each gene.) Because a_i is a proportion, $\sum_{i=1}^{m} a_i = 1$, and in general for RNA-seq data, most a_i are small.

In this highly idealized instructive model, we assume that an RNA-seq experiment is composed of exactly the following two steps:

1. Sample cells and mRNA molecules from the biological sample.

2. Sequence the mRNA molecules to obtain counts.

Step 1: Sample cells and mRNA molecules

Let the unobserved random variable M_j represent the (ground truth) total number of mRNA molecules sampled for observation j. Let the unobserved random variable X_{ij} represent the number of copies of gene i in observation j. Assuming that mRNA molecules are sampled uniformly at random (all cells and all molecules are equally likely to be sampled),

$$X_{ij}|M_j \sim \text{Binomial}(M_j, a_i)$$

Because M_j is large (millions of molecules) and a_i is small, we can approximate this distribution as the following Poisson distribution[1]:

$$X_{ij}|M_j \;\dot{\sim}\; \text{Poisson}(M_j a_i)$$

Observe that both distributions have the same expected value $M_j a_i$, and that the binomial variance $M_j a_i(1-a_i)$ is approximately the Poisson variance $M_j a_i$ because $1 - a_i \approx 1$. See the final section of these notes for a visual explanation of the Poisson approximation for the binomial distribution.

We introduce the unobserved random variable π_{ij} to represent the proportion of molecules in observation j that are from gene i, i.e.,

$$\pi_{ij} = \frac{X_{ij}}{M_j}$$

[1] We use $\dot{\sim}$ to indicate "approximately distributed as."

Clearly, π_{ij} is closely related to our quantity of interest, a_i. Indeed, %TODO: check this derivation with notes when they're posted

$$\begin{aligned}\mathrm{E}\left[\pi_{ij}\right] &= \mathrm{E}\left[\mathrm{E}\left[\pi_{ij}|M_j\right]\right] \\ &= \mathrm{E}\left[\frac{M_j a_i}{M_j}|M_j\right] \\ &= \mathrm{E}\left[\frac{M_j a_i}{M_j}\right] = \mathrm{E}\left[a_i\right] = a_i\end{aligned}$$

That is, the expected value of π_{ij} is a_i. Using the law of total variance, we can write the variance of π_{ij} as

$$\mathrm{Var}(\pi_{ij}) = \mathrm{E}\left[\mathrm{Var}\left(\pi_{ij}|M_j\right)\right] + \mathrm{Var}\left(\mathrm{E}\left[\pi_{ij}|M_j\right]\right)$$

Note that for this equation, all variances are taken over π_{ij} and all expectations are taken over M_j. We can evaluate the second term as follows using the expectation we just saw above:

$$\mathrm{Var}\left(\mathrm{E}\left[\pi_{ij}|M_j\right]\right) = \mathrm{Var}(a_i) = 0$$

where the last equality follows from the fact that a_i is constant with respect to π_{ij}. Then, only the first term remains, so

$$\mathrm{Var}(\pi_{ij}) = \mathrm{E}\left[\mathrm{Var}\left(\pi_{ij}|M_j\right)\right] = \mathrm{E}\left[\mathrm{Var}\left(\frac{X_{ij}}{M_j}|M_j\right)\right] = \mathrm{E}\left[\frac{1}{M_j^2}\,\mathrm{Var}\left(X_{ij}|M_j\right)\right] \approx \mathrm{E}\left[\frac{1}{M_j^2}\left(a_i M_j\right)\right] = \frac{a_i}{M_j}$$

where the $\approx$ corresponds to the Poisson approximation for the binomial distribution. This variance is conceptually similar to the “biological variance," i.e., the variance attributable to biology before any additional variance is introduced by the measurement process.

Step 2: Sequence mRNA molecules and obtain counts

In this section, we assume that mRNA molecules are sampled uniformly at random for sequencing and subsequent measurement (ignoring any issues like gene length, GC bias, etc.). Let the random variable D_j be the total number of reads we obtain from observation j. Because we observe this quantity, we will write it as d_j and treat it as a constant. Let Y_{ij} be a random variable representing the number of RNA-seq reads from gene i in observation j. While we observe y_{ij} for each RNA-seq experiment, we would like to model the distribution of these counts to obtain population-level information, i.e., a_i.

$$\begin{aligned}Y_{ij}|\pi_{ij}, d_j &\sim \mathrm{Binomial}(d_j, \pi_{ij}) \\ Y_{ij}|\pi_{ij}, d_j &\overset{\cdot}{\sim} \mathrm{Poisson}(d_j \pi_{ij})\end{aligned}$$

The second line is the same Poisson approximation as before, given that d_j is large and π_{ij} is small. We can then compute the expected value and variance of Y_{ij}. The expected value is relatively straightforward:

$$\mathrm{E}\left[Y_{ij}\right] = \mathrm{E}\left[\mathrm{E}\left[Y_{ij}|D_j = d_j, \pi_{ij}\right]\right] = \mathrm{E}\left[d_j \pi_{ij}\right] = d_j\,\mathrm{E}\left[\pi_{ij}\right] = d_j a_i$$

The variance, however, has a few tricks to it (again beginning with the law of total variance):

$$\begin{aligned}\mathrm{Var}(Y_{ij}) &= \mathrm{E}\left[\mathrm{Var}\left(Y_{ij}|D_j = d_j, \pi_{ij}\right)\right] + \mathrm{Var}\left(\mathrm{E}\left[Y_{ij}|D_j = d_j, \pi_{ij}\right]\right) \\ &\approx \mathrm{E}\left[d_j\pi_{ij}\right] + \mathrm{Var}\left(d_j\pi_{ij}\right) \\ &= d_j a_i + d_j^2 a_i \,\mathrm{E}\left[\frac{1}{M_j}\right]\end{aligned}$$

Note that the $\approx$ in the second line again represents the Poisson approximation for the binomial distribution, which allows us to substitute $d_j\pi_{ij}$ into the first term. The second term is then simplified using the expected value we just computed. Because $\mathrm{Var}(Y_{ij}) > \mathrm{E}[Y_{ij}]$, this marginal distribution of Y_{ij} is an example of what is called an *overdispersed* Poisson distribution (compared to a Poisson-distributed random variable X where $\mathrm{Var}(X) = \mathrm{E}[X]$).

Knowing that π_{ij} is closely related to our quantity of interest a_i, we would like to estimate it using the following estimator:

$$\hat{\pi}_{ij} = \frac{Y_{ij}}{d_j}$$

We can see immediately using $\mathrm{E}[Y_{ij}]$ that the expected value of this estimator is π_{ij}, thus it is *unbiased.* The variance of this estimator is as follows:

$$\begin{aligned}\mathrm{Var}(\hat{\pi}_{ij}) = \frac{1}{d_j^2}\,\mathrm{Var}(Y_{ij}) &= \frac{a_i}{d_j} + \mathrm{Var}(\pi_{ij}) \\ &= \frac{a_i}{d_j} + a_i \,\mathrm{E}\left[\frac{1}{M_j}\right]\end{aligned}$$

The second term in this variance the same biological variance – $\mathrm{Var}(\pi_{ij})$ – that we saw above in Step 1. Thus, we can (roughly) refer to the remaining term as the technical variance, i.e., the variance attributable to the measurement process.

Now that we have this estimator for π_{ij}, we would like to use it to estimate a_i. Recall that

- π_{ij} is the proportion of reads from gene i in observation j,
- all of the observations j are from the same biological condition,
- and a_i is the proportion of mRNA molecules from gene i in this biological condition.

Combining these facts brings us to the following estimator:

$$\hat{a}_i = \frac{1}{n}\sum_{j=1}^{n}\hat{\pi}_{ij}$$

Again, because $\mathrm{E}[\hat{\pi}_{ij}] = \mathrm{E}[\pi_{ij}] = a_i$, $\mathrm{E}[\hat{a}_i] = a_i$. The variance of $\hat{a}_i$ is then

$$\mathrm{Var}(\hat{a}_i) = \mathrm{Var}(\frac{1}{n}\sum_{j=1}^{n}\hat{\pi}_{ij}) = \frac{1}{n^2}\,\mathrm{Var}(\sum_{j=1}^{n}\hat{\pi}_{ij}) = \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{Var}(\hat{\pi}_{ij})$$

We can then separate $\mathrm{Var}(\hat{pi}_{ij})$ into the technical component (first term) and the biological component (term corresponding to biological variance (second term) to split $\mathrm{Var}(\hat{a}_i)$ as follows:

$$\mathrm{Var}(\hat{a}_i) = \frac{a_i}{n^2}\sum_{j=1}^{n}\frac{1}{d_j} + \sum_{j=1}^{n}\frac{\mathrm{Var}(\pi_{ij})}{n^2}$$

The first term again corresponds to the technical variance in our estimator $\hat{a}_i$, and the second term corresponds to the biological variance.

We assume that the total number M_j of mRNA molecules for each observation j are independent and identically distributed (iid). Under this assumption, $\mathrm{E}\left[\frac{1}{M_1}\right] = \mathrm{E}\left[\frac{1}{M_2}\right] = \ldots = \mathrm{E}\left[\frac{1}{M_n}\right]$.

In order to summarize the relationship between the mean and the variance in this model of RNA-seq data, we will introduce a few general quantities and define them in the context of this model. The first is the *coefficient of variation* CV, which we define as follows:

$$\mathrm{CV} = \frac{\sqrt{\mathrm{Var}(\pi_{ij})}}{a_i}$$

This quantity is referred to as the *biological* coefficient of variation. Then,

$$(\mathrm{CV})^2 = \frac{\mathrm{Var}(\pi_{ij})}{a_i^2} = \frac{1}{a_i}\,\mathrm{E}\left[\frac{1}{M_j}\right] \equiv \phi_i$$

where the rightmost equals sign denotes a definition, i.e., we define ϕ_i to be $\frac{1}{a_i}\,\mathrm{E}\left[\frac{1}{M_j}\right]$. We define $\mu_{ij} = d_j a_i$, i.e., μ_{ij} is the population mean proportion a_i for gene i times the observed read depth d_j for observation j.

We can then express the variance of Y_{ij} in terms of these quantities:

$$\begin{aligned}\mathrm{Var}(Y_{ij}) &= d_j a_i + d_j^2\,\mathrm{Var}(\pi_{ij}) \\ &= d_j a_i + (d_j a_i)^2 \frac{\mathrm{Var}(\pi_{ij})}{a_i^2} \\ &= \mu_{ij} + \mu_{ij}^2 \phi_i\end{aligned}$$

The parameter phi_i in this model is referred to as the *dispersion parameter*, in that it determines how the variance is scaled by the square of the mean. In practice, it is normally inferred by “borrowing strength" across genes that are assumed to have similar values of ϕ_i. The parameter ϕ_i is also an example of a *nuisance parameter* in the context of statistical inference, meaning that while it is part of the model and thus must be inferred, it is not a quantity of interest in that knowing it does not yield additional insight into the population. One of the key insights of this model is the *mean-variance relationship* – the mean appears in the variance, particularly the square of the mean. In general a strong mean-variance relationship complicates statistical inference, and we will see that this particular relationship with the square of the mean also appears in alternative models of RNA-seq data.

Negative binomial model

Negative binomial model for RNA-seq data

We will now turn to an alternative model for RNA-seq data, which relies on the *negative binomial* distribution. Consider a sequence of Bernoulli trials with a success probability p. Rather than model the number of successes in a fixed number of trials (as in the Binomial distribution), instead we model the number Y of *failures* before the rth success. This value Y is a random varaible distributed according to the negative binomial distribution:

$$Y \sim \mathrm{NegBin}(r, p)$$

$$\Pr(Y = y) = \binom{r+y-1}{y} p^r (1-p)^y$$

Note that this distribution is only defined for non-negative integers y, i.e., $y \in \mathbb{N}$.

The expected value and variance of Y are then

$$\begin{aligned}\mathrm{E}[Y] &= \frac{r(1-p)}{p} \\ \mathrm{Var}(Y) &= \frac{r(1-p)}{p^2}\end{aligned}$$

Let $\mu = \frac{r(1-p)}{p}$, and let $\phi = \frac{1}{r}$. Then, we can express the variance in terms of μ and ϕ as in the previous model to obtain the same mean-variance relationship:

$$\mathrm{Var}(Y) = \mu + \mu^2\phi$$

Thus, we can model RNA-seq data as a negative binomial distribution.

$$Y_{ij} \sim \mathrm{NegBin}(r_i, p_{ij})$$

where $\mu_{ij} = \frac{r_i(1-p_{ij})}{p_{ij}}$ and $\phi_i = \frac{1}{r_i}$. Again, ϕ_i is a nuisance parameter; sometimes it is modeled as a gene-specific parameter as in this formulation (indexed by gene i), and other times it is modeled as a single dispersion parameter that is shared across genes.

Compound gamma-Poisson formulation

The negative binomial distribution is a special case of the gamma[2]-Poisson distribution

$$\begin{aligned}Y|\lambda &\sim \mathrm{Poisson}(\lambda) \\ \lambda &\sim \mathrm{Gamma}(\alpha, \beta)\end{aligned}$$

Under this distribution, in which the random variable Y_{ij} is distributed according to a Poisson distribution paramterized by a gamma random variable λ_{ij}, Y_{ij} is marginally a gamma-Poisson random varaible. Note that the negative binomial distribution is a *special case* of the gamma-Poisson distribution (i.e., for any negative binomial distribution, there exists a specific parameterization of the gamma-Poisson distribution that is equivalent to this negative binomial distribution).

The gamma pdf, expected value, and variance are as follows:

$$f(\lambda; \alpha, \beta) = \frac{\lambda^{\beta-1}e^{-\lambda/\alpha}}{\alpha^\beta\Gamma(\beta)}, \lambda > 0$$

$$\mathrm{E}[\lambda] = \alpha\beta, \mathrm{Var}(\lambda) = \alpha^2\beta$$

The gamma-Poisson pdf, expected value, and variance are as follows:

$$f(y; \alpha, \beta) = \frac{\Gamma(y+\beta)\alpha^y}{\Gamma(\beta)(1+\alpha)^{\beta+y}y!}$$

$$\mathrm{E}[Y] = \alpha\beta, \mathrm{Var}(Y) = \alpha\beta + \alpha^2\beta$$

Let $\mu = \alpha\beta$ and $\phi = \frac{1}{\beta}$. Then, as before, we have

$$\mathrm{Var}(Y) = \mu + \mu^2\phi$$

[2] As we saw previously with the beta distribution, the gamma distribution can take many different shapes as its parameters α and β are varied. However, unlike the beta distribution, the gamma distribution has support over all positive real numbers rather than just $(0, 1)$ as in the beta distribution.

Even under this completely different model, we obtain the same mean-variance relationship as in the two-step model.

Now that we have this result that is identical to the two-step model of RNA-seq data, we can add subscripts and map these variables back to the two-step model. The random variable Y_{ij} represents the number of reads from gene i in observation j, as before.

$$\begin{aligned} Y_{ij}|\lambda_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &\sim \text{Gamma}(\alpha, \beta) \end{aligned}$$

The random variable λ_{ij} corresponds to the quantity $\pi_{ij} d_j$ in the previous model, i.e., $\lambda_{ij} = \pi_{ij} d_j$.[3] Observe that:

- $\text{E}[\pi_{ij} d_j] = a_i d_j$ (from the previous section)
- $\text{E}[\lambda_{ij}] = \alpha\beta$ (by definition of the gamma distribution)
- $\mu_{ij} = a_i d_j$ (from the previous section)

Thus, $\mu_{ij} = a_i d_j = \alpha\beta$. We can also unite the two definitions of ϕ_i to obtain

$$\phi_i = \frac{1}{\beta} = \frac{1}{a_i} \, \text{E}\left[\frac{1}{M_j}\right]$$

Finally, we have that

$$\begin{aligned} \beta_{ij} &= a_i \, \text{E}\left[\frac{1}{M_j}\right]^{-1} \\ \alpha_{ij} &= a_i d_j \cdot \frac{1}{a_i} \, \text{E}\left[\frac{1}{M_j}\right] = d_j \, \text{E}\left[\frac{1}{M_j}\right] \\ \lambda_{ij} &\sim \text{Gamma}(\alpha_{ij}, \beta_{ij}) \end{aligned}$$

Mean-variance relationships in general

Consider a random variable Y representing count data. Suppose Y is distributed as follows:

$$Y|\lambda \sim \text{Poisson}(\lambda)$$

Here, λ is a positve random variable. The variance of Y can then be computed using the law of total variance:

$$\begin{aligned} \text{Var}(Y) = \text{E}\left[\text{Var}\left(Y|\lambda\right)\right] + \text{Var}\left(\text{E}\left[Y|\lambda\right]\right) \\ \text{Var}(Y) = \text{E}[\lambda] + \text{Var}(\lambda) \end{aligned}$$

Since $\lambda > 0$, the mean $\text{E}[\lambda]$ of its distribution will always be positive, and thus it will always appear in the variance of Y. This implies that any Poisson model for a count variable will have a mean-variance relationship which complicates inference.

[3] Here Y_{ij} is explicitly Poisson-distributed according to λ_{ij}, whereas in the previous model the Poisson relationship between Y_{ij} and $\pi_{ij} d_j$ relied on the Poisson approximation for the binomial distribution.

Facts about random varaibles

Sums of random variables

If X is a random variable and a, b are constants, then

$$\begin{aligned} \mathrm{E}[a+bX] &= a + b\,\mathrm{E}[X] \\ \mathrm{Var}(a+bX) &= b^2\,\mathrm{Var}(X) \end{aligned}$$

Let $X_1, X_2, \ldots, X_n$ be n random variables Then,

$$\begin{aligned} \mathrm{E}\left[\sum_{i=1}^n X_i\right] &= \sum_{i=1}^n \mathrm{E}\left[X_i\right] \\ \mathrm{Var}\left(\sum_{i=1}^n X_i\right) &= \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}(X_i, X_j) \end{aligned}$$

When $X_1, X_2, \ldots, X_n$ are independent[4], then $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$, so

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}\left(X_i\right)$$

Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Suppose $X_1, X_2, \ldots, X_n$ are independent. Then,

$$\mathrm{E}\left[\overline{X}_n\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[X_i\right]$$

Thus, when $\mathrm{E}[X_1] = \mathrm{E}[X_2] = \ldots = \mathrm{E}[X_n] = \theta$, $\mathrm{E}[\overline{X}_n] = \theta$.

$$\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\left(X_i\right)$$

Thus, when $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \ldots = \mathrm{Var}(X_n) = \tau^2$, $\mathrm{Var}(\overline{X}_n) = \tau^2/n$. Roughly speaking, this result indicates that the mean becomes a better estimator (i.e., the variance decreases) of the population mean as the number n of data points increases.

Convergence of random variables

Let $Z_1, Z_2, \ldots$ be a sequence of random variables. For example, Z_n could be the mean of the first n data points, i.e., $Z_n = \overline{X}_n$. Alternatively, $Z_n \sim \mathrm{Binomial}(n, p)$ with some p.

Convergence in Distribution

{Z_n} converges in distribution to the random variable W (written as: $Z_n \xrightarrow{D} W$ as $n \to \infty$) if

$$F_{Z_n}(y) = \Pr(Z_n \leq y) \to \Pr(W \leq y) = F_W(y)$$

for all $y \in \mathrm{R}$, $n \to \infty$.

[4] In this section, when we say a group of random variables are independent, we require only pairwise independence.

Convergence in Probability

$\{Z_n\}$ converges in probability to the random variable W (written as: $Z_n \xrightarrow{P} W$ as $n \to \infty$) if

$$\Pr(|Z_n - W| \leq \epsilon) \to 1$$

as $n \to \infty$ for $\epsilon > 0$.

Note that convergence in probability is a stronger result than convergence in distribution: rather than Z_n converging to a distribution that looks like W, instead the value of Z_n is converging to the value of W. For a fixed number θ, we can also have $Z_n \xrightarrow{P} \theta$.

Almost sure convergence

$\{Z_n\}$ converges "almost surely" (a.s.) or "with probability 1" to W (written as $Z_n \xrightarrow{a.s.} W$) if

$$\Pr(\{\omega : |Z_n(\omega) - W(\omega)| \xrightarrow{n\to\infty} 0\}) = 1$$

This result is again even stronger than the last, saying that there is asymptotically no event ω with positive probability mass where $Z_n(\omega)$ differs from $W(\omega)$.

Results regarding random variables

Strong Law of Large Numbers

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with population mean $E[X_i] = \mu$ where $E\left[|X_i|\right] < \infty$. Then

$$\overline{X}_n \xrightarrow{a.s.} \mu, \text{ as } n \to \infty$$

Central limit theorem

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with population mean $E[X_i] = \mu$ and population variance $\mathrm{Var}(X_i) = \sigma^2$. Then, as $n \to \infty$,

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{D} \text{Normal}(0, \sigma^2)$$
$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \text{Normal}(0, 1)$$

Here is the derivation of the second (standard normal) result, using the first result and several of the rules outlined above.

$$\begin{aligned}
\mathrm{Var}\left(\overline{X}_n - \mu\right) &= \mathrm{Var}\left(\overline{X}_n\right) = \frac{\sigma^2}{n} \\
\mathrm{Var}\left(\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}}\right) &= \frac{1}{\sigma^2/n}\,\mathrm{Var}(\overline{X}_n) = 1 \\
\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} &= \sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow{D} \text{Normal}(0, 1)
\end{aligned}$$

Useful facts about normal random variables

Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$. Then

$$\mathrm{E}\left[\overline{X}_n\right] = \mu$$

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

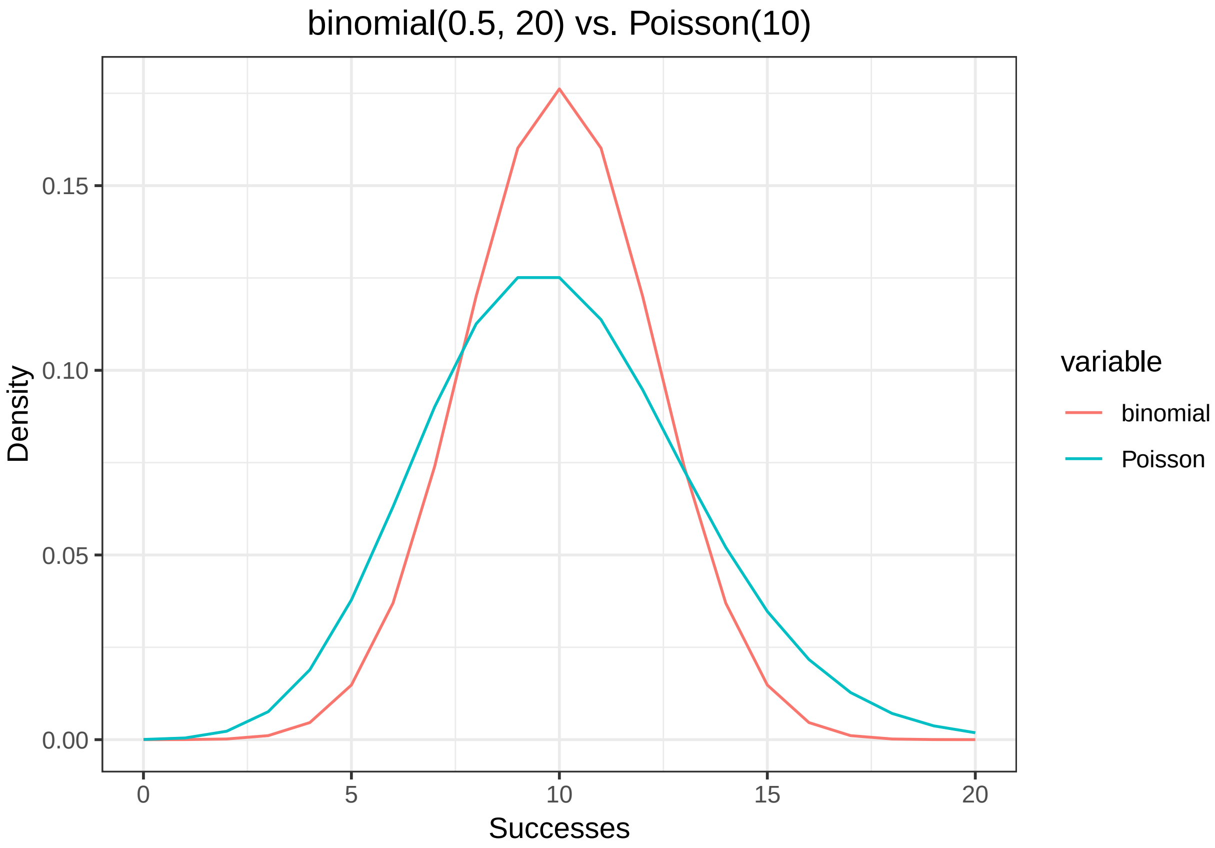
$$\overline{X}_n \sim \mathrm{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

because $X_1 + X_2 + \ldots + X_n \sim \mathrm{Normal}(n\mu, n\sigma^2)$, and $aX_1 + b \sim \mathrm{Normal}(a\mu + b, a^2\sigma^2)$.

Poisson approximation for the binomial distribution

In this section, we illustrate the Poisson approximation for the binomial distribution using some visual examples. This approximation works better for large n (many Bernoulli trials) and small p (low probability of success in each trial). To see this, first we look at an example where n is relatively small and p relatively large. Note that the expected value of both distributions is 10.

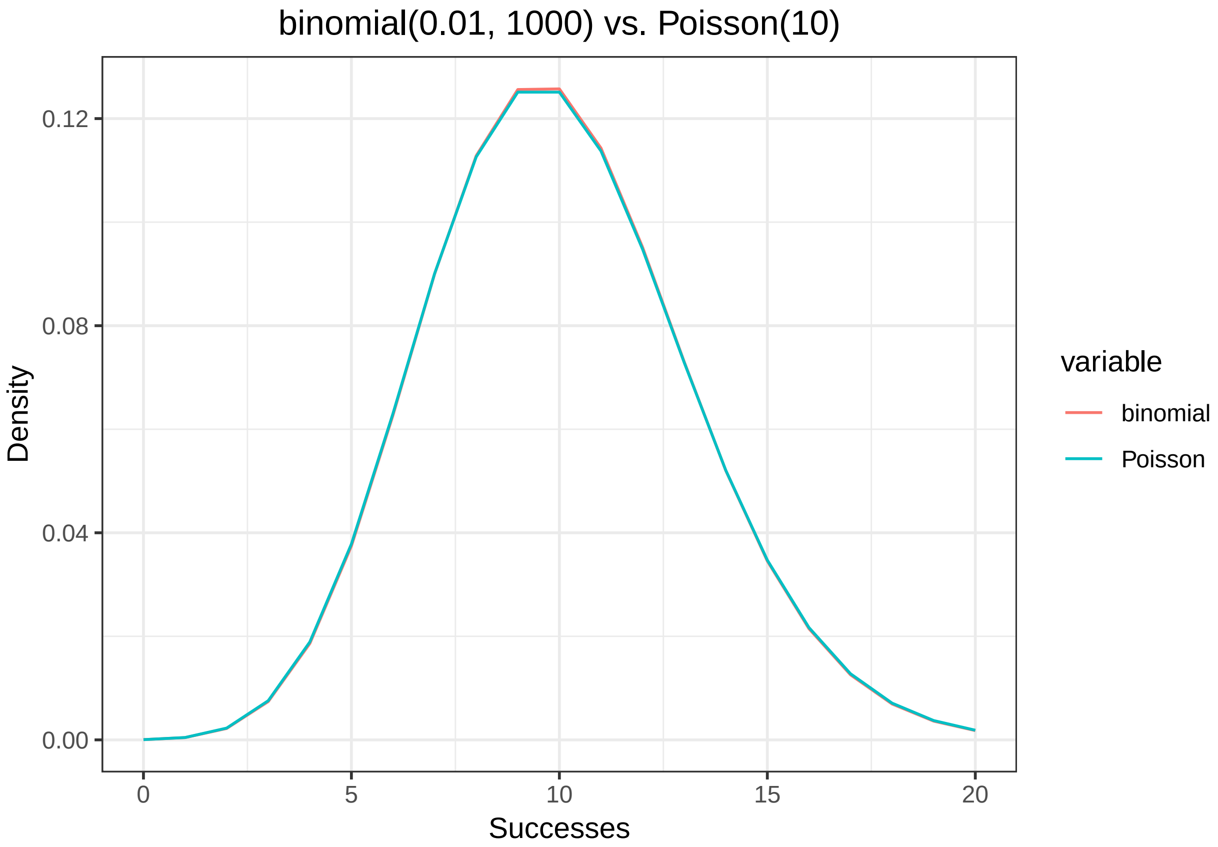
```
> library(ggplot2)
> library(reshape2)

Attaching package: 'reshape2'
The following object is masked from 'package:tidyr':

    smiths
> xmax <- 20
> x <- seq(0, xmax, 1)
> density_binom1 <- dbinom(x = x, 20, 0.5)
> density_pois1 <- dpois(x = x, 10)
> df1 <- data.frame(x=x, binomial = density_binom1, Poisson = density_pois1)
> plot1 <- ggplot(dat = melt(df1, id.var="x"), aes(x=x, y=value)) +
+   geom_line(aes(colour=variable, group=variable)) +
+   ggtitle("binomial(0.5, 20) vs. Poisson(10)") +
+   xlab("Successes") + ylab("Density") + xlim(c(0, xmax))
> plot1
```



As you can see, while the distributions roughly place probability mass in similar places, they differ in structure: the Poisson distribution has a larger variance than the binomial distribution. The Poisson variance is the same as the expected value, so 10, whereas the binomial variance is $20 \cdot 0.5 \cdot 0.5 = 5$.

Now, we consider a binomial distribution with small $p = 0.01$ and large $n = 1000$. Note that the expected value is again 10 for both distributions, but now the binomial variance is $1000 \cdot 0.01 \cdot 0.99 = 9.9$ which is very close to the Poisson variance of 10.

```
> xmax <- 20
> x <- seq(0, xmax, 1)
> density_binom2 <- dbinom(x = x, 1000, 0.01)
> density_pois2 <- dpois(x = x, 10)
> df2 <- data.frame(x=x, binomial = density_binom2, Poisson = density_pois2)
> plot2 <- ggplot(dat = melt(df2, id.var="x"), aes(x=x, y=value)) +
+   geom_line(aes(colour=variable, group=variable)) +
+   ggtitle("binomial(0.01, 1000) vs. Poisson(10)") +
+   xlab("Successes") + ylab("Density") + xlim(c(0, xmax))
> plot2
```



As you can see, the distributions appear virtually identical.

Let $X_1 \sim \text{binomial}(p, n)$, and $X_2 \sim \text{Poisson}(pn)$. Then, consider the ratio between the variance of X_1 and the variance of X_2:

$$\frac{\text{Var}(X_1)}{\text{Var}(X_2)} = \frac{np(1-p)}{np} = 1 - p$$

Thus, the Poisson distribution has a larger variance than the Poisson distribution by a factor of $1 - p$. In our first example, $1 - p = 1 - 0.5 = 0.5$, so the Poisson distribution had double the variance of the binomial distribution. However, in the second example, $1 - p = 0.99$, so the variance of the Poisson distribution was off by only 1% from that of the binomial distribution. Note that this ratio does not depend on n, so a larger number of trials will not improve the approximation.

Session Information

```
> sessionInfo()
R version 3.6.0 (2019-04-26)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS  10.15.3

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
```

```
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] reshape2_1.4.3  forcats_0.4.0   stringr_1.4.0   dplyr_0.8.1
 [5] purrr_0.3.2     readr_1.3.1     tidyr_0.8.3     tibble_2.1.1
 [9] ggplot2_3.1.1   tidyverse_1.2.1 knitr_1.22

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.1        cellranger_1.1.0  pillar_1.4.0      compiler_3.6.0
 [5] plyr_1.8.4        tools_3.6.0       digest_0.6.18     lubridate_1.7.4
 [9] jsonlite_1.6      evaluate_0.13     nlme_3.1-140      gtable_0.3.0
[13] lattice_0.20-38   pkgconfig_2.0.2   rlang_0.3.4       cli_1.1.0
[17] rstudioapi_0.10   yaml_2.2.0        haven_2.1.0       xfun_0.7
[21] withr_2.1.2       xml2_1.2.0        httr_1.4.0        hms_0.4.2
[25] generics_0.0.2    grid_3.6.0        tidyselect_0.2.5  glue_1.3.1
[29] R6_2.4.0          readxl_1.3.1      rmarkdown_1.12    modelr_0.1.4
[33] magrittr_1.5      backports_1.1.4   scales_1.0.0      htmltools_0.3.6
[37] rvest_0.3.3       assertthat_0.2.1  colorspace_1.4-1  labeling_0.3
[41] stringi_1.4.3     lazyeval_0.2.2    munsell_0.5.0     broom_0.5.2
[45] crayon_1.3.4
```