XDAS 2020: Other discrete distributions

Hypergeometric, Geometric, and Negative Binomial

Kris Gunsalus

10/29/2020

Contents

Background

Aho - Chapter 3.3-3.6

Introduction to dnorm, pnorm, qnorm, rnorm - PDF - RMD

Review

- Poisson and Exponential
- Bernoulli distribution
- Binomial distribution

Hypergeometric distribution

The hypergeometric distribution is similar to the *binomial distribution*, except it defines the *probability of obtaining x independent successes* when sampling ***without replacement***. It is classically described in terms of an urn containing some black balls and some white balls. The hypergeometric distribution will tell us how likely it is that a handful of balls picked from the urn contains a certain number of black (or white) balls.

Example: GO-term enrichment

In our field, probably the most common application of the hypergeometric distribution is to test whether a set of genes of interest (e.g. up- or down-regulated genes in a differential gene expression analysis) is enriched for a specific functional annotation (e.g. GO term).

To test this, we need to ask whether the observed overlap is different from what we would expect if we picked the same number of genes at random from the genome. What would this be? We can use set theory to figure this out. Let's visualize the problem:

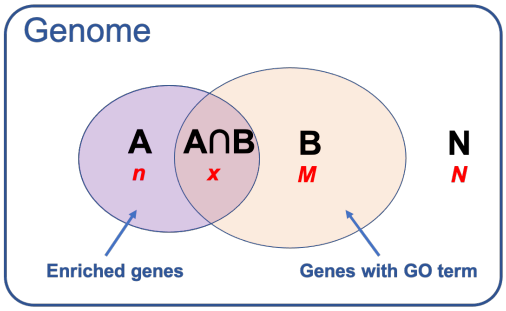


Figure 1: **Overlap between two sets of genes in the genome**

Our question is whether the overlap between the two sets of genes is greater than (or less than) expected by chance. Under independence, we know that:

$$Pr[A \cap B] = Pr[A] * Pr[B]$$

We can use a hypergeometric test to answer this question. The question is formulated in this way:

If we pick n out of N total items (Set A), what is the chance that x of them would also be contained in M (Set B), if the two sets were independent?

To answer this, we need to know just a few things:

- N: the total number of selectable items in the genome
- n: the number of items in Set A
- M: the number of items in Set B
- x: the overlap between Set A and Set B ($A \cap B$)

Hypergeometric PDF

Sampling without replacement means that we are picking a particular set of items from a finite set of total items. Therefore, each trial affects the probability of the next outcome – in other words, we need an equation to find the relative frequency of x in a shrinking sample space. (If the population were infinite, then this would essentially be the same as sampling with replacement, since the sample would not make a dent in the remaining number of individuals to choose from.)

The PDF is defined as:

$$f(x) = P(X = x) = \binom{M}{x}\binom{N-M}{n-x}\Big/\binom{N}{n}$$

where:

- $x \in \{0, 1, 2, ..., n\}$ is the number of "successful" trials
- $N \in \{1, 2, ..., N\}$ is the total number of selectable items
- $M \in \{0, 1, 2, ..., N\}$ is the total number of possible "successful" outcomes (the number of items in the group we are comparing against)
- $n \in \{0, 1, 2, ..., N\}$ is the number of items sampled

The random variable x represents the overlap between the two sets ($A \cap B$), n is Set A, and M is Set B. Here, Set B is defined as "success" and Set A is the sample we are asking about.

The **lower-tail** probability is the probability that ***fewer*** than x overlaps are observed (depletion), and the **upper-tail** probability is the probability that ***more*** than x overlaps are observed (enrichment).

The components of the equation are:

- $\binom{M}{x}$: the number of ways to get ($A \cap B$) out of B items
- $\binom{N-M}{n-x}$: the number of ways to get ($A \quad NOT \quad B$) out of ($N \quad NOT \quad B$) items
- $\binom{N}{n}$: the number of ways to get A out of N items

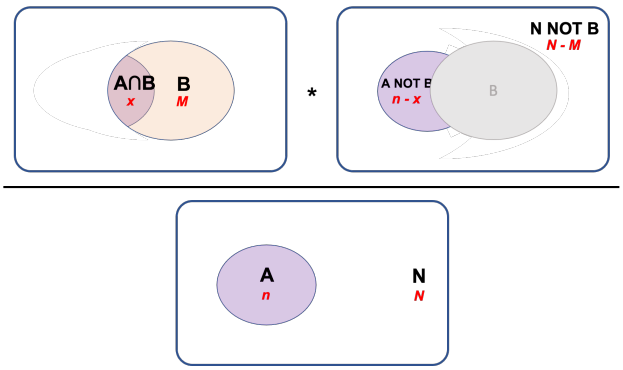


Figure 2: **Hypergeometric formula**

Hypergeometric test

The `hyper` family of functions in R does not observe the usual naming convention, which is a bit confusing at first, but the variables are the same.

For example[1], let's say we have a list of 59 genes that are predicted to be regulated by the mouse E2F transcription factor, and we want to know whether they are enriched for genes involved in the cell cycle. There are 13,588 genes with some GO annotation in the mouse genome, and 611 of them are annotated with the term "cell cycle". We find that 19 out of our 59 genes are annotated with the GO term "cell cycle".

Is the list of predicted E2F targets significantly enriched for the GO term "cell cycle"? We can figure this out in the following way:

```
help(phyper)

# hypergeometric CDF in R is defined as:
#   p(x) = choose(m, x) choose(n, k-x) / choose(m+n, k)
```

[1] http://pedagogix-tagc.univ-mrs.fr/courses/ASG1/practicals/go_statistics_td/go_statistics_td_2015.html

```r
# the formula is:
#   phyper(q, m, n, k, lower.tail = T/F)

# Instead, we will use:
#   phyper(Overlap-1, SetB, N - SetB, SetA, lower.tail= FALSE)
Overlap = 19  # predicted genes with GO term
A = 59     # predicted E2F targets
B = 611    # genes with GO term
N = 13588     # total annotated genes

# Expected overlap based on the null hypothesis:
# [# genes in (A AND B)] = [# genes in A] * [# genes in B] / N
exp.overlap = A * B / N
print(paste("Expected overlap =", round(exp.overlap,2)))
```

```
## [1] "Expected overlap = 2.65"
```

```r
# Fold-enrichment:
fold.enrichment = ( Overlap/A ) / ( B/N )
print(paste("Fold-enrichment =", round(fold.enrichment,2)))
```

```
## [1] "Fold-enrichment = 7.16"
```

```r
# P(enrichment) = upper-tail probability: P(X >= x)
# Note: We use Overlap-1, otherwise we are asking for P(X > x)
#       since this is a discrete distribution
phyper(Overlap-1, B, N - B, A, lower.tail= FALSE)
```

```
## [1] 4.989683e-12
```

```r
# same using PDF instead
sum(dhyper( Overlap:A, B, N - B, A ))
```

```
## [1] 4.989683e-12
```

Fisher's Exact Test

The hypergeometric test is the same as a one-tailed Fisher's exact test:

```r
# set up the contingency table with the overlap in the top left corner
# first row is the predicted targets
A.not.B = A-Overlap
B.not.A = B-Overlap
N.not.AB = N - B - A.not.B
contingency.table = rbind(c(Overlap, A.not.B),
                          c(B.not.A, N.not.AB))
rownames(contingency.table) = c("target","not.target")
colnames(contingency.table) = c("GO","not.GO")
contingency.table
```

```
##             GO not.GO
## target      19     40
## not.target 592  12937
```

```r
# Is the overlap greater than expected by chance?
fisher.test(contingency.table, alternative="greater")
```

```
##
```

```
##  Fisher's Exact Test for Count Data
##
## data:  contingency.table
## p-value = 4.99e-12
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  6.204117      Inf
## sample estimates:
## odds ratio
##   10.37524
```

Geometric distribution

The geometric distribution is similar to the *binomial distribution*, except it gives the probability that x independent Bernoulli **failures** occur **prior to the FIRST success**.

Example:

- The number of cards you would need to sample (with replacement, i.e. "catch and release") before finding an ace.
- The number of jelly beans you need to sample from a mixed bag before finding a watermelon flavored one (my favorite!)

The PDF is:

$$f(x) = P(X = x) = p(1-p)^x$$

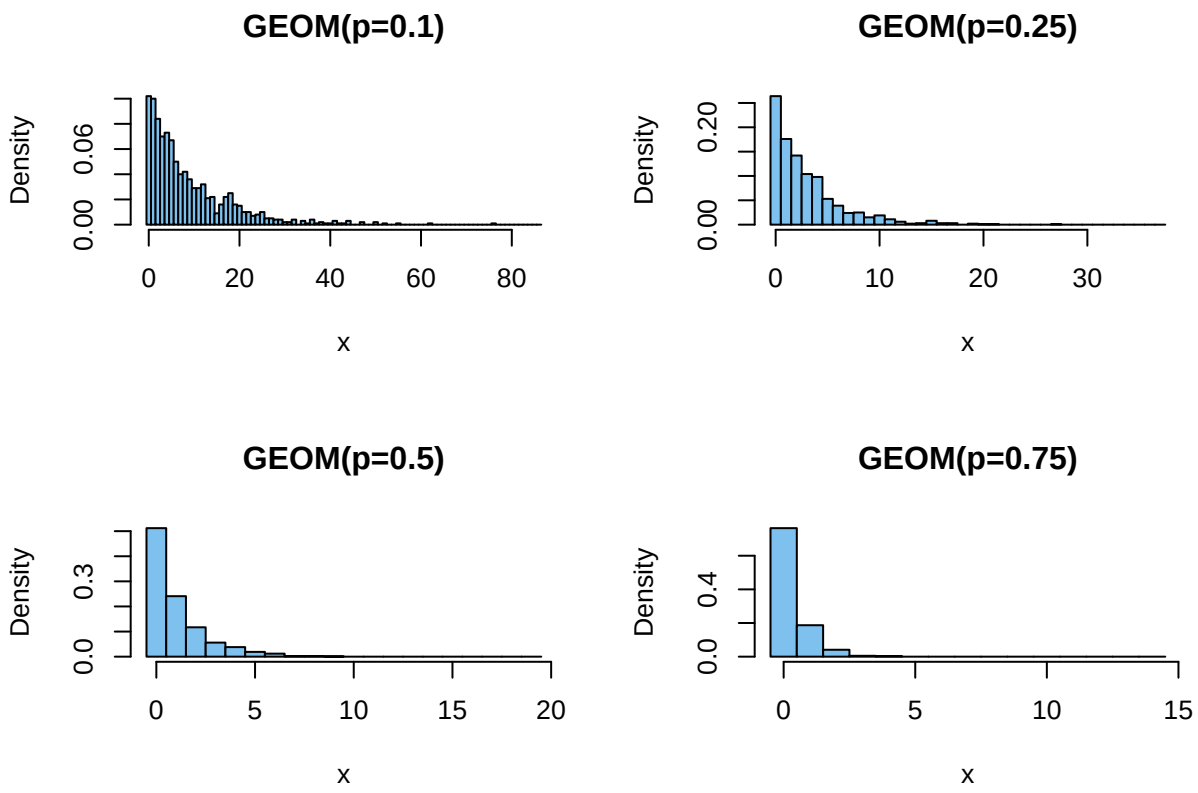
where:

- p is the **unchanging** probability of success in a single Bernoulli trial, and
- $x \in \{0, 1, 2..., n\}$ is the *number of* ***unsuccessful*** trials ***preceding*** the first success.

There is a single parameter p. If a random variable follows the geometric distribution, we write: $X \sim GEO(p)$.

Since we are only concerned with a *single success* following x failures, the above formula fully defines this probability as the intersection of x failures and one success. We do not need to use any binomial coefficients because there is only one way to get this outcome (of course, that is if you stop when you get the first success; it's the same as the old adage, "You always find what you were looking for in the last place you look!")

```r
# rgeom(n,p)

par(mfrow=c(2,2))
prob = c(0.1,0.25,0.5,0.75)
for (i in prob) {
  y = rgeom(1000, i)
  xlimit = max(y) + 10
  title = paste("GEOM(p=",i,")",sep="")
  hist(y, prob=T, br=(-1:xlimit)+.5, col="skyblue2", xlab="x", main=title)
}
```



Negative Binomial

The negative binomial is an extension of the geometric distribution to cases where we are interested in ***more than one success*** after a number of failures. In fact, the geometric distribution is a special case of the negative binomial in which the number of successful trials = 1. So, we can use it to answer questions of the form:

What is the chance that I will get 12 tails before I get at least 2 heads?

Example:

- The probability that you will have a total of 5 failed experiments (e.g. PCR runs, or CRISPR injections of individual *C. elegans* animals) before you will manage to get three experiments to work, given that your success rate is 20% on average (three is the magic number! ;-)
- In another formulation, the total number of experiments you will need to perform in order to achieve three successful experiments.

There are a number of different parameterizations of the NB distribution, depending on exactly which question you are asking. I found a very good discussion of these topics here[2]. This blog also provides insight into the relationship between the negative binomial, the Poisson, the binomial, and the gamma distributions.

PDF

The PDF for the discrete NB distribution is:

[2] https://probabilityandstats.wordpress.com/tag/negative-binomial-distribution/

$$f(x) = P(X_r = x) = \binom{x+r-1}{r-1} p^r (1-p)^x = \binom{x+r-1}{x} p^r (1-p)^x$$
$$= \binom{n-1}{r-1} p^r (1-p)^x = \binom{n-1}{x} p^r (1-p)^x$$

where:

- trials are independent and only two outcomes are possible
- the probability of success p in each trial is constant
- n: total number of trials
- r: number of successes
- x: number of *failures* preceding r successes in n independent trials (note that $x + r = n$)
- X_r: the random variable, representing the number of *unsuccessful trials* before r successes

Since the last (nth) Bernoulli trial is the rth success, the binomial coefficient gives the number of ways to obtain x failures in the preceding $x + r - 1 = n - 1$ trials, which is equivalent to the number of ways to obtain $r - 1$ successes preceding the rth success. The binomial coefficients above are equivalent, since

$$\binom{a}{b} = \binom{a}{a-b} \quad for \quad 0 \leq b \leq a.$$

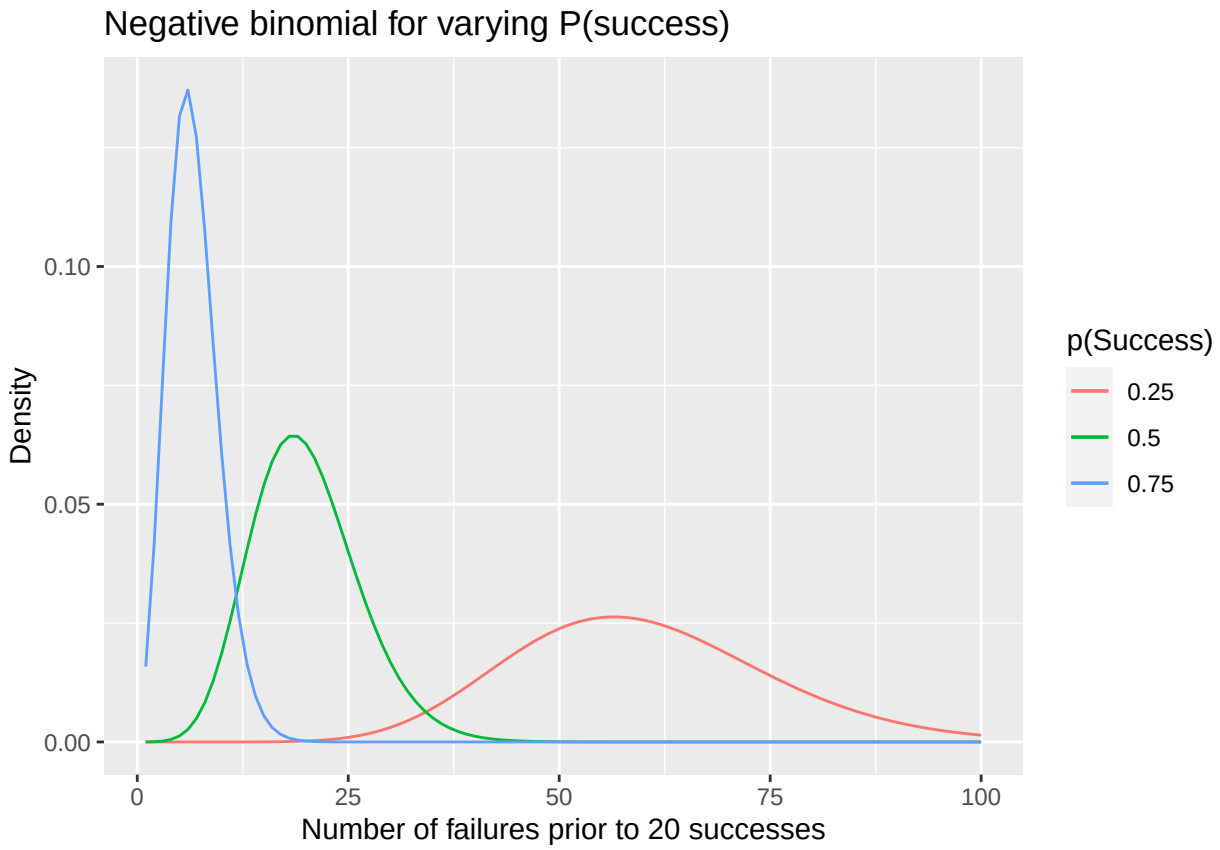
If a random variable follows a negative binomial distribution, we write $X \sim NB(r, p)$. There are two parameters: the ***number of successes*** desired, and the ***probability of success***. It is important to remember that X, the random variable, represents the number of *failures* before achieving the desired number of successes. So, the distribution represents the probability of obtaining different numbers of failures before a total of r successes is obtained.

The **geometric** distribution is a special case of the NB, where $r = 1$: $X \sim GEO(p)$ is equivalent to $X \sim NB(1, p)$

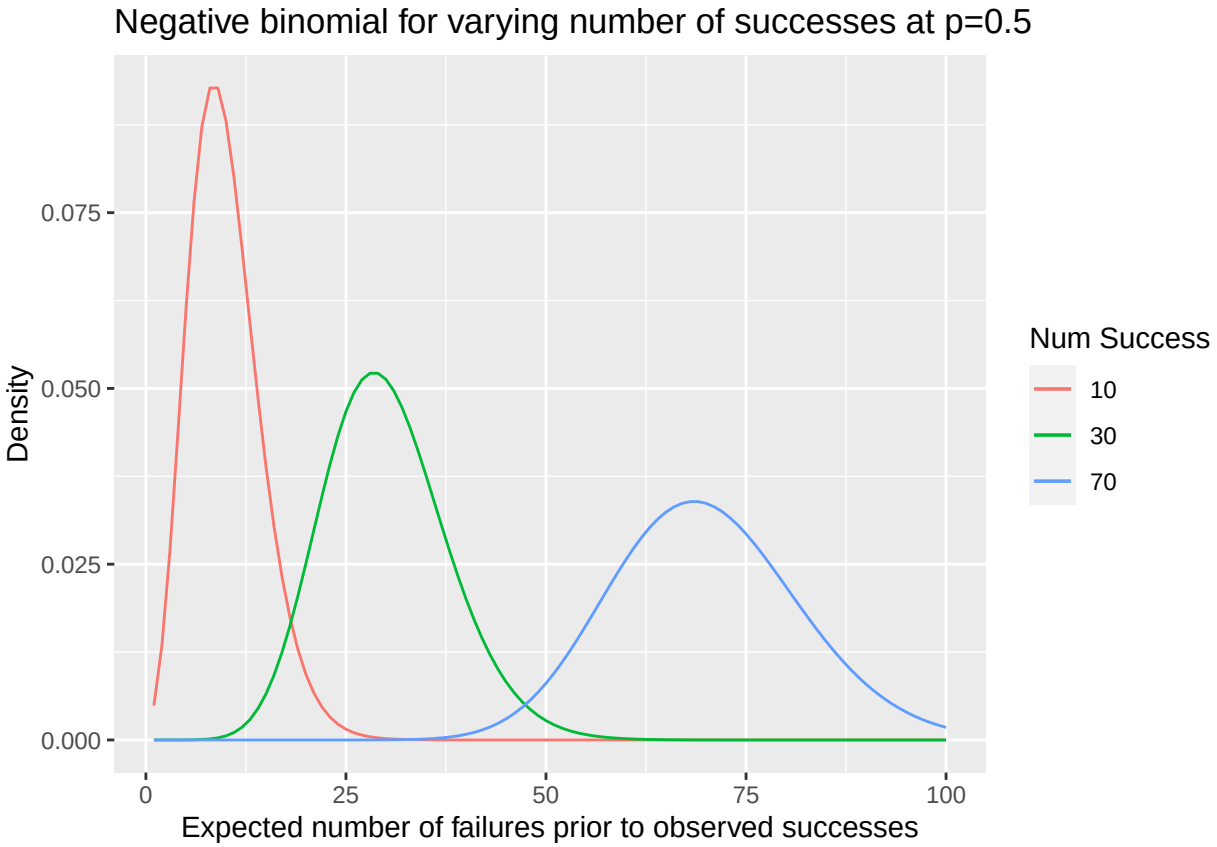
The mean and variance of the negative binomial distribution are: $E(X) = \frac{r}{p}$ and $Var(X) = \frac{r(1-p)}{p^2}$.

Below are density plots for the NB showing how its shape changes as the probability changes or the number of successes changes.

```
library(ggplot2)

# The plot below shows varying the probability
nbinomdata = data.frame(x=1:100,
                        y=dnbinom(1:100,size=20,0.25),
                        z=dnbinom(1:100,20,0.5),
                        w=dnbinom(1:100,20,0.75))

ggplot(nbinomdata) +
  geom_line(aes(x=x, y=y, col="0.25")) +
  geom_line(aes(x=x, y=z, col="0.5")) +
  geom_line(aes(x=x, y=w, col="0.75")) +
  ggtitle("Negative binomial for varying P(success)") +
  xlab("Number of failures prior to 20 successes") +
  ylab("Density")+
  labs(col = "p(Success)")
```




```
# The plot below shows varying the number of successes
nbinomdata = data.frame(x=1:100,
                        y=dnbinom(1:100,10,0.5),
                        z=dnbinom(1:100,30,0.5),
                        w=dnbinom(1:100,70,0.5))

ggplot(nbinomdata) +
  geom_line(aes(x=x, y=y, col="10")) +
  geom_line(aes(x=x, y=z, col="30")) +
  geom_line(aes(x=x, y=w, col="70")) +
  ggtitle("Negative binomial for varying number of successes at p=0.5") +
  xlab("Expected number of failures prior to observed successes") +
  ylab("Density") +
  labs(col = "Num Success")
```



An alternative formulation, which gives the total number of trials that must be performed in order to achieve the rth success, is:

$$f(n; r, p) = P(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

where $n = \{r, r+1, r+2, ...\}$

CDF and Survival function

The CDF of the NB is:

$$P(X_r \leq x) = \sum_{j=0}^{k} \binom{j+r-1}{j} p^r (1-p)^j \quad for \quad k = 0, 1, 2, ...$$

It is common to encounter the *survival function*, which is the complement of the CDF of a particular distribution. If the CDF is $F(X) = P(X \leq x)$, then the *survivorship* is $P(X > x) = 1 - F(X)$

And the survival function is:

$$P(X_r > x) = \sum_{j=k+1}^{\infty} \binom{j+r-1}{j} p^r (1-p)^j \quad for \quad k = 0, 1, 2, ...$$

Relationship between the NB and other distributions

Geometric distribution When $r = 1$, the NB becomes the geometric distribution, which is therefore a special case of the NB: $GEO(p) = NB(1, p)$.

Therefore, the *negative binomial* is to the *geometric* distribution as the *binomial* is to the *Bernoulli* distribution.

Poisson distribution The Poisson distribution is a special case of the NB when *the number of successes is very large* and the *probability of success* is very small (in order to keep the mean of the distribution constant).

$$POIS(\lambda) = lim_{r->\infty} NB(r, r/(\lambda + r)))$$

Binomial distribution The negative binomial takes its name from the defining equation for the binomial coefficient with negative numbers.

The negative binomial distribution describes the *waiting time* before the rth success in n independent Bernoulli trials, where as the binomial distribution describes the *success rate* in n trials.

If there are k failures before r successes, then there are at most $r-1$ successes in $n = k+r$ trials. Consequently, *the survival function of the NB is the same as the CDF of the binomial distribution* with parameters $n = k+r$ and p, where k is the number of failures and r is the number of successes.

Equivalently, *the CDF of the NB is the survival function of the binomial distribution.*

Negative binomial in analysis of deep sequencing data

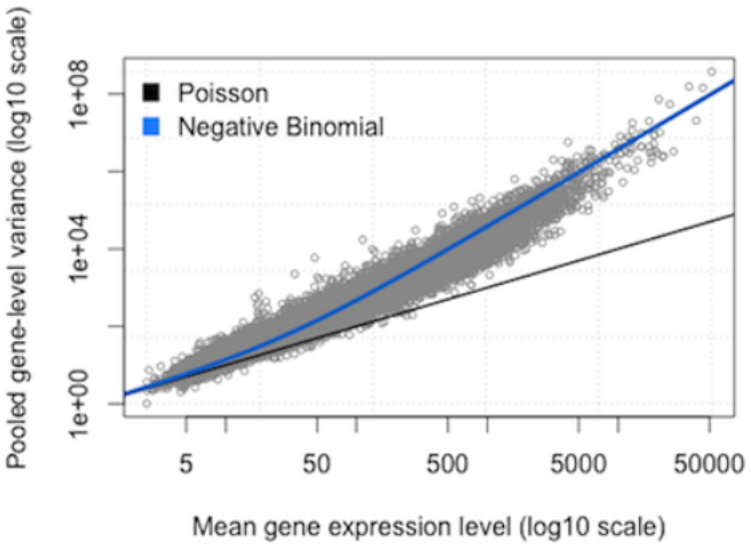
The negative binomial has become popular in recent years as a way to model the distribution of read counts in deep sequencing data. In brief, count data for technical replicates can be modeled by a Poisson distribution. Recall that the mean and the variance for a Poisson are given by the same parameter, λ. Thus, the variance is proportional to the mean.

When considering biological replicates, it turns out that the variation in the counts varies with the number of counts per feature (e.g. expression level); it is *overdispersed*. In such cases, the Poisson is no longer the best model for the data.

Instead, the NB is used to model the uncertainty in the variance. In this case, the variation is proportional to the mean, with an added term to account for the dispersion:

$$\sigma^2 = \mu + \alpha\mu^2$$

where α is the dispersion parameter. For $\alpha > 1$, the dispersion is greater than the mean; as α goes to 0, the NB converges on a Poisson distribution.



There are various explanations for modeling sequence count data with the NB distribution, and a relatively simple one may be found here: https://bioramble.wordpress.com/2016/01/30/why-sequencing-data-is-modeled-as-negative-binomial/

The authors of DESeq2 provide a more detailed discussion[3].

Technically, the NB is a *Poisson-Gamma mixture distribution*: a mixture of Poisson distributions where the uncertainty in the various λs follows a Gamma distribution. The details are beyond our pay grade for the purposes of this class, but you may want to file this for future reference.

Examples

Hypergeometric

What is the probability of holding two aces in a hand of five cards? (There are 52 cards in a full deck.)

$$P(X = x) = \binom{M}{x}\binom{N-M}{n-x}\Big/\binom{N}{n} = \binom{4}{2}\binom{52-4}{5-2}\Big/\binom{52}{5}$$
$$= \frac{\frac{4!}{2!2!}\frac{48!}{45!3!}}{\frac{52!}{47!5!}} = \frac{6 * (48 * 47 * 46/6)}{(52 * 51 * 50 * 49 * 48)/120} = 0.04$$

```
dhyper(2,4,48,5)
```

```
## [1] 0.03992982
```

```
choose(4,2) * choose(48,3) / choose(52,5)
```

```
## [1] 0.03992982
```

Geometric

You are performing an exit poll at a polling station in a district where the proportion of independent voters is 20%.

[3]https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r106

$$f(x) = P(X = x) = p(1-p)^x$$

How likely is it that you will have to interview at least 10 people before finding one that voted independent? No more than 10? No more than 5? Exactly 5?

n: total number interviewed $x = n - 1$: number of failures $p = 0.2$: probability of success

```
pgeom(9,0.2, lower.tail = F)  # 10 or more: P(X > 9)
## [1] 0.1073742
pgeom(10,0.2, lower.tail = T) # 10 or fewer: P(X <= 10)
## [1] 0.9141007

pgeom(5,0.2, lower.tail = T) # P(X <= 5)
## [1] 0.737856
sum(dgeom(0:5,0.2))          # same using PDF
## [1] 0.737856
dgeom(5,0.2)                 # P(X = 5 failures): the 6th person is independent
## [1] 0.065536
0.2*(1-0.2)^5                # same using equation
## [1] 0.065536
```

Negative Binomial

What is the chance that you will have to interview exactly 15 people in total order to find 5 that voted independent? No more than 15 (i.e. 15 or fewer)? More than 15? At least 15 (i.e. 15 or more)?

n: total interviewed r: number of successes $x = n - r$: number of failures $p = 0.2$: probability of success

$$P(X = 10) = \binom{x+r-1}{r-1} p^r (1-p)^x = \binom{14}{4} (0.2)^5 (0.8)^{10} = 0.034$$

```
# for n=15, x=n-r (# of failures = total - # of successes)
dnbinom(10, size=5, prob=0.2)      # P(X = 10) exactly 10
## [1] 0.0343941
pnbinom(10, 5, 0.2, lower.tail=T) # P(X <= 10) 10 or fewer
## [1] 0.1642337
pnbinom(10, 5, 0.2, lower.tail=F)  # P(X > 10) more than 10
## [1] 0.8357663
pnbinom(9, 5, 0.2, lower.tail=F)  # P(X >= 10) 10 or more
## [1] 0.8701604
```

What is the chance you'll interview 5 or less before finding one independent? More than 5? Exactly 5?

```
# same as geometric function
pnbinom(5, size=1, prob=0.2) # P(X <= 5, r=1)
## [1] 0.737856
pnbinom(5, size=1, prob=0.2, lower.tail=F) # P(X > 5, r=1)
## [1] 0.262144
dnbinom(5, size=1, prob=0.2) # P(X = 5,  r=1)
## [1] 0.065536
```