Principles of Statistics for the Natural Sciences
-Notes-

José Miguel Ponciano

February 2, 2016

PREFACE

Add Preface here.

Contents

INTRODUCTION

The contents of these notes are a compendium of examples I have learned through the years as a student and teacher in the fields of Statistical Ecology and Statistical Genetics. An important influence in my formation and in particular in my approach to teaching have been professors Brian Dennis, Paul Joyce and Steve Krone. Brian Dennis has been, for many years now, both a mentor and a friend. Many of the examples at the beginning of these notes are from the courses he teaches at University of Idaho, and that he in turn learned from G.P. Patil at Penn. State, and many others. Some key examples also come from my beautiful time at CIMAT, in Mexico, and in particular, from the course about theoretical likelihood inference taught by Eloísa Díaz-Francés and the late David A. Sprott.

To finish later: Acknowledge examples/teaching material from Paul Joyce, Chris Williams, Ken Newman, Al Manson and Stephen M. Krone

PART 1:

Discrete probability distributions in Ecology and Evolution

1.1 The likelihood function

1.1.1 Two simple mark-recapture models

Sampling with replacement:
Suppose that we are studying a closed population of desert mice. In a first visit to the desert, we trap 49 mice, mark them with a red tag and then release them. After some time, we come back to the study area and trap mice again. Each time we capture a mouse, we record whether it is marked or not and release it. That is, we sample mice *with replacement*. With the recorded data, we seek to estimate the total number of individuals in the population. How do we go about writing a probability model for this experiment? Can we build a statistical model to explain how the data arose? Let

- X be the r.v. that counts the number of marked mice recaptures in the second visit.
- x denote the realized value of X.
- m be the number of marked mice in the population.
- t be the total number of mice in the population.
- n be the total number of mice captured in the second visit (23).

Suppose that the experimental data consist of the following results: $x = 5$, $m = 49$, $n = 23$. Here, t is the only unknown quantity. In what follows, after building a probabilistic model for this experiment we derive the Maximum Likelihood (ML) estimate of t.

In order to build a probabilistic model, first note that the experiment "*recording the number of marked mice among the n captured mice*" can be viewed as a sequence of n trials with binary outcome (marked/not marked or "Success"/"Failure"). Let's assume for now that each of these n trials is independent from each other. Then, the probability of observing a marked mouse (*i.e.* the probability of a success) in one of these trials is $\frac{m}{t}$. Likewise, the probability of observing an unmarked mouse is $\left(1 - \frac{m}{t}\right)$. Hence, the probability of a particular sequence of x successes and $n - x$ failures is $\left(\frac{m}{t}\right)^x \left(1 - \frac{m}{t}\right)^{n-x}$. Noting that the total number of such sequences is equal to

$$\begin{aligned} \frac{\text{\# of ways of assigning } x \text{ marked mice in } n \text{ trials}}{\text{\# of ways that } x \text{ marked mice can be ordered}} &= \frac{n(n-1)(n-2)\ldots(n-x+1)}{x!} \\ &= \frac{n!}{x!(n-x)!} = \binom{n}{x}, \end{aligned}$$

we get that

$$P(X = x) = \binom{n}{x} \left(\frac{m}{t}\right)^x \left(1 - \frac{m}{t}\right)^{n-x}, \quad x \in \{0, 1, 2, \ldots, n\}.$$

This the binomial distribution with parameters n and m/t and from here on we will write $X \sim \text{Bin}\left(n, \frac{m}{t}\right)$. Note that this Binomial distribution is built as a sequence

of n independent binary trials. In probability, a binary random trial with success probability p (m/t in our case) is known as a Bernoulli probability distribution, Be(p). Now, going back to our Binomial distribution, note that the probability of drawing 5 marked mice in 23 trials is then:

$$P(X = 5) = \binom{23}{5} \left(\frac{49}{t}\right)^5 \left(1 - \frac{49}{t}\right)^{23-5}.$$

Since t is an unknown quantity, we can view the right hand side (RHS) of the above equation as function of plausible values of t. This function is plotted in Figure 1.

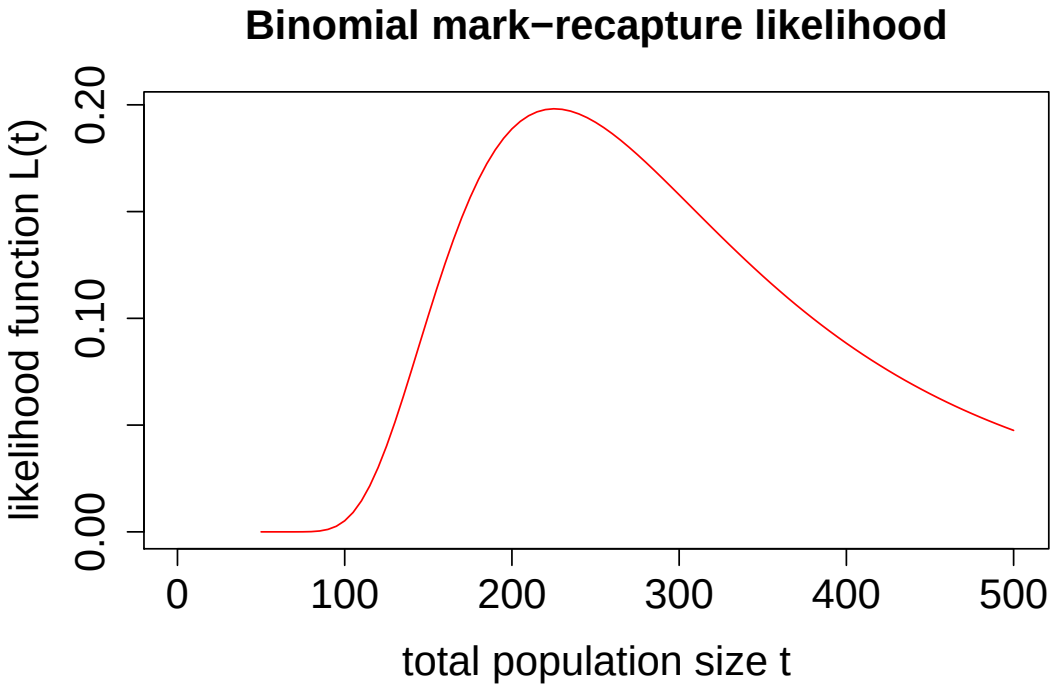

Figure 1: Plot of $P(X = 5)$ as a function of the unknown quantity t.

By doing this exercise, we find that one value around 230 of the unknown quantity t would yield the observed result ($x = 5$) more frequently than any other value. Noting that 'probability' implies a ratio of frequencies and "*about the frequencies of such values we can know nothing whatever*", Fisher (1922) suggested to talk instead of the *likelihood of one value of the unknown parameter being a number of times bigger than the likelihood of another value.* Thus, following Fisher, we refer to the function

$$\ell(t) = \binom{n}{x} \left(\frac{m}{t}\right)^x \left(1 - \frac{m}{t}\right)^{n-x}$$

as the *likelihood function* of t and use it to quantify the relative frequencies with which the values of the hypothetical quantity t would in fact yield the observed sample (Fisher 1922). The value $\hat{t}$ that maximizes this function is called the Maximum Likelihood (ML) estimate of t. Finding this value analytically is straightforward in this case. To do that, we 1) compute $\ln \ell(t)$, 2) find its derivative with respect to t and 3) set it equal to 0 and solve for t:

1)

$$\ln \ell(t) = \ln \binom{n}{x} + x \ln m - x \ln t + (n-x) \ln (t-m) - (n-x) \ln t$$

2)

$$\frac{d \ln \ell(t)}{dt} = -\frac{x}{t} + \frac{(n-x)}{(t-m)} - \frac{n-x}{t},$$

and **3)**

$$\frac{d \ln \ell(t)}{dt} = \frac{n-x}{t-m} - \frac{n}{t} = 0 \Rightarrow \hat{t} = \frac{nm}{x} = 225.4$$

This estimator of t is known as the "Lincoln-Petersen" index in the scientific literature. Finding $\hat{t}$ using `R` is also straightforward. Instead of doing the above calculations in `R`, we will find the integer ML estimate "by hand": First, let's define a function that computes $\ell(t)$ for various values of t, given the (known) values of x, m and n. We can do that using the function `dbinom` that computes the pmf of the Binomial random variable:

```
binom.like<- function(t,n,m,x){
like<- dbinom(x=x,size=n,prob=(m/t),log=FALSE);
return(like)
}
```

Alternatively, instead of using function `dbinom` we could have used the function `lgamma(x)` that computes $\ln(\Gamma(x))$[1] :

```
binom.like<- function(t,n,m,x){
like <- exp(lgamma(n+1)-lgamma(x+1)-lgamma(n-x+1)+x*log(m/t)+(n-x)*log(1-(m/t)));
return(like)
}
```

To do the plot in Figure 1 we type in `R` 's command line :

```
>tvec <- seq(50,500,by=5);
>like.caprecap<- binom.like(t=tvec,n=23,m=49,x=5);
>par(oma=c(1,2,1,1));
>plot(tvec,like.caprecap, col="red",type="l",main="Binomial mark-recapture likelihood",
xlab="total population size t", ylab="likelihood function L(t)",xlim=c(0,501),
cex.main=1.5,cex.lab=1.5,cex.axis=1.5);
```

[1]The gamma function $(\Gamma(x))$ is explained in detail right after the derivation of the gamma distribution. For now, just think of it as the factorial function for real, positive numbers. You might also wonder: why exponentiate and then take the log in this code? Because when dealing with very big and very small numbers, it is numerically more stable to compute sums than multiplications.

Finally, the integer ML estimate of t is found by typing

```
> that<- tvec[which(like.caprecap==max(like.caprecap),arr.ind=T)]
> that
[1] 225
```

Sampling without replacement:

Suppose now that in the second visit we sample n mice *without replacement*. Here again, we let X be the r.v. that counts the number of marked mice recaptures in the second visit. Under this setting we have that

$\binom{t}{n}$ = # of samples of size n from t mice

$1/\binom{t}{n}$ = probability of a particular batch of n mice captured from t mice

$\binom{m}{x}$ = # of ways of choosing x marked mice from m marked mice,

$\binom{t-m}{n-x}$ = # of ways of choosing $n-x$ unmarked mice from $t-m$ unmarked mice and

$\binom{m}{x}\binom{t-m}{n-x}$ = # of ways of choosing x marked and $n-x$ unmarked mice.

Then,

$$P(X=x) = f(x) = \frac{\binom{m}{x}\binom{t-m}{n-x}}{\binom{t}{n}}$$

Hence, X follows the hypergeometric distribution. Note two things: first, if n exceeds $(t-m)$ then some marked animals must appear in the sample. Second, the number of marked animals in the sample cannot exceed m or n. In other words

$$\max(0, m+n-t) \leq x \leq \min(m, n).$$

The ML estimate of t for this setting may be found using four different methods. The first method consists of drawing a picture of the likelihood function and finding graphically $\hat{t}$. The second approach is to take the derivative of $\ln\ell(t)$, set it equal to 0 and solve for t. However no closed form of $\hat{t}$ can be found in this case, and we have to resort to the third approach: numerical maximization of $\ln\ell(t)$. However, before giving up, we can try to find the integer ML estimate analytically. This last approach consists of finding an integer value of t such that $\ell(t) = \ell(t-1)$. Let $[a]$ denote the greatest integer $\leq$ a. Then, first we set $\ell(t) = \ell(t-1)$, solve for t and and take $\hat{t}$ to be $[t]$:

$$\frac{\ell(t-1)}{\ell(t)} - 1 = 0 \Rightarrow \frac{\binom{t-1-m}{n-x}\binom{t}{n}}{\binom{t-m}{n-x}\binom{t-1}{n}} - 1 = 0,$$

and after simplifying (in fact, after some messy algebra) we get

$$(t-m-n+x)t = (t-n)(t-m) \Rightarrow t = \frac{nm}{x}.$$

Rounding to the nearest integer we get $\hat{t} = \left[\frac{nm}{x}\right]$, which is the Petersen index.
It is often the case that multiple independent samples are taken, in which case the setting is:

- $k = \#$ of independent samples taken,
- $t =$ total population size,
- $m_i = \#$ in population that are marked at time of the i^{th} sample,
- $n_i = \#$ captured in the i^{th} sample,
- $x_i = \#$ marked and captured in the i^{th} sample,

and the likelihood function is:

$$\ell(t) = \prod_{i=1}^{k} \frac{\binom{m_i}{x_i}\binom{t-m_i}{n_i-x_i}}{\binom{t}{n_i}}.$$

As an example, consider the following data set: In Alaska, 13 wild goats where captured and marked. Then 3 aerial surveys were done. The results are

flight	**Total # of goats seen**	**Total # of marked goats seen**
1	74	6
2	72	6
3	51	6

Exercises:

1. For the Goats data set in the Hypergeometric distribution example above, write an `R` function that calculates the log-likelihood function and maximize it with respect to t to find $\hat{t}$.

2. Show via simulations that as $t \to \infty$, $m \to \infty$ and $\frac{m}{t} \to p$, where $p \in (0,1)$ is a constant, the hypergeometric distribution approaches the binomial distribution. Illustrate your answers graphically. Type `?rhyper` in `R` to learn how the simulator of random numbers for the hyper-geometric distribution works.

3. **Extra credit:**If $X_1 \sim \text{Bin}(m,p)$ and $X_2 \sim \text{Bin}(t-m,p)$ is independent of X_1 and $Y = X_1 + X_2$, what is $P(X_1 = x|Y = n), \quad 0 \leq n \leq t$? Carefully interpret your result (this is as important as the mathematical derivation you will work out).

4. **Extra credit:** Show (analytically, that is) that as $t \to \infty$, $m \to \infty$ and $\frac{m}{t} \to p$, where $p \in (0, 1)$ is a constant, the hypergeometric distribution approaches the binomial distribution.

Conditional distributions as sampling models: Before moving on with a careful comparison of the sampling with replacement and the sampling without replacement mark-recapture models, let's think a little bit more about the process of eliciting a probability model of how the data arises. In both formulations, we left out one important piece of realism: even if I know that in my study area there are exactly m marked animals from my first visit, setting the traps in the field wouldn't certainly guarantee that all m mice are going to be captured. Let us assume that every marked animal in the population has a probability, say ϕ, of being captured, and thus a probability $1 - phi$ of avoiding capture. Never mind the fact that such probability is the same for every animal, we will deal with biological heterogeneities later. Suppose now that regardless of whether or not the animals are marked, not only all of them have exactly the same chance ϕ of being trapped, but the fate of every mouse is independent from the fate of all the other mice.

Under this setting, the process of attempting to capture all the marked animals in the sample can be thought of as a binomial experiment X_1 with a total number of trials m and success probability ϕ. Likewise, attempting to capture the remaining animals could be modeled with a binomial random variable $X_2 \sim \text{Bin}(t - m, \phi)$. Because all mice have the same probability of being trapped, the total number of marked mice can be modeled with a binomial random variable N defined as the sum of X_1 and X_2, with total number of trials $m + (t - m) = t$ and success probability ϕ. Now, suppose that we go ahead with the trapping and capture n individuals. Then, eliciting a probabilistic model for the *number of marked animals x present in a sample of size n* amounts to ask what is the conditional probability that $X_1 = x$ given that $N = n$. Note that because $N = X_1 + X_2$, if $X_1 = x$, then conditioning on $N = n$ amounts, by necessity, to conditioning on X_2 being equal to $n - x$. Therefore

$$
\begin{aligned}
P(X_1 = x | N = n) &= \tfrac{P(X_1=x, N=n)}{P(N=n)} \\
&= \tfrac{P(X_1=x, X_2=n-x)}{P(N=n)} \text{ and from independence,} \\
&= \tfrac{P(X_1=x)P(X_2=n-x)}{P(N=n)} \\
&= \frac{\binom{m}{x}\phi^x(1-\phi)^{m-x}\binom{t-m}{n-x}\phi^{n-x}(1-\phi)^{t-m-n+x}}{\binom{t}{n}\phi^n(1-\phi)^{t-n}} \\
&= \frac{\binom{m}{x}\binom{t-m}{n-x}}{\binom{t}{n}},
\end{aligned}
$$

which is the hypergeometric probability mass function! This pmf can now be used to make inferences about the unknown value of the total population size! Note that

although we introduced a little bit more of realism by considering the capture probability ϕ, in the end, we didn't need to specify this extra parameter as another unknown parameter to be estimated with the sample at hand. Conditional distributions will often help us to derive the likelihood function for various random sampling settings that are common in ecology and in evolution.

Comparing the two sampling approaches:
Our two models of mark-recapture, the binomial and the hypergeometric models, resulted in the same ML estimator of the total population size: $\hat{t} = \frac{mn}{x}$. So, what difference does it make to assume the first rather simple sampling model (with replacement) vs. the second, more realistic sampling setting (sampling without replacement)? What do we gain by introducing a little more realism in our sampling model? At first, these questions might take us aback because for a given sample, assuming either model leads to the same estimate. Let's try, however, to think a little more about what does it mean to fit both models *via* ML.

When we fit a model using the likelihood function, we are computing a measure of how likely it is to observe the sample at hand for a given parameter value, and an assumed sampling model. Each hypothetical value of the unknown model parameter has an associated value of the likelihood score. Finding the parameter value that maximizes the likelihood function allows one to use the maximized likelihood score $\hat{\ell}$ as an evidence measure for our sampling model. This evidence comes directly from the sample at hand. Comparing the maximized likelihood scores from two different models then amounts to comparing the support in the data for each model. This result is useful, because knowing which model better explains the data is the key to better understand the biological processes generating the data. Thus, because in science we seek to better understand and predict natural phenomena, it seems fair to think that the process of finding the "most likely" value of our sampling model parameter and its associated likelihood score is, as a whole, a scientifically relevant process.

One of the most important factors that may shift the weight of the evidence in favor of one model or the other is how well each sampling model recapitulates the different (biological) dependencies present in the data. A better representation of the structure of variability in the data results in statistical inferences that are more reliable. It turns out that reliability, another desired qualification of proper scientific inference, can be directly expressed as statistical properties of our parameter estimates. So to answer our questions, all we need to do is to elucidate which of the two sampling models leads to more reliable estimators of the parameter of interest (the total population size).

How, then, do we measure the reliability of an estimator? The trick to find out how reliable an estimator can be is to think of these not as point values, but rather, as a realizations of the stochastic sampling process we are using as a model of how the data arises. Consider the binomial mark-recapture sampling scheme, where we sample with replacement. Suppose that we generate many data sets according to such scheme, all under the same setting, *i.e.*, assuming m individuals are initially marked and n individuals are sampled during a second visit. In `R`, a large number of

simulated samples of the number of marked animals captured in the second visit X can be very easily simulated using the binomial random number generator. Just tell R what the setting is (define m, n, t), and the number of simulations (say 20) and in an instant, you can see all the realized values of $\hat{t}$

```
n <- 20;
m <- 49;
tot   <- 225; # True population size
nsims <- 20; # number of simulations
sim.samples <- rbinom(n=nsims, size=n, prob=m/tot);
t.hats <- (m*n)/sim.samples;
print(sim.samples)
print(t.hats)
```

And you'll see something like this

```
> print(sim.samples)
 [1] 6 1 7 2 5 7 3 5 5 3 2 3 4 2 2 8 6 5 6 4
> print(t.hats)
 [1] 163.3333 980.0000 140.0000 490.0000 196.0000 140.0000
 [7] 326.6667 196.0000 196.0000 326.6667 490.0000 326.6667
[13] 245.0000 490.0000 490.0000 122.5000 163.3333 196.0000
[19] 163.3333 245.0000
```

What is important here is to note that each random sample X of the number of marked animals has an associated estimate of t. Because X varies randomly, so does $\hat{t}$. In a very real sense, our estimator 'inherits' its randomness from the random sampling scheme. Suddenly, the sampling process that we designed to learn from the natural world gets back at us with multiple answers!! What do we do as scientists? Well, the first thing to do is to change the focus of our inquiry from 'what is the total population size?' to 'how reliable is it to adopt a particular random sampling scheme and then use it to estimate the total population size?'. And here is where the tools of probability and mathematical statistics become useful, because questioning the reliability of our estimator can be phrased precisely in terms of the mean and the variance of the resulting distribution of $\hat{t}$:

1. Given that the estimator of t would change from sample to sample if I were to repeat the experiment many times, on average how far apart would it be from the true population size? In other words, given that the estimator $\hat{t}$ can be described with a probability distribution, how does the expected value of such distribution compares to the true value of t?

2. What is the average departure of the estimator $\hat{t}$ from t (*e.g.*, what is the variance of $\hat{t}$)? A reliable estimator is an estimator that time after time results in an estimate that isn't too far apart from the true population size.

The first question can be easily addressed using our simulation program. Just make the number of simulations very, very large, compute the average of the resulting $\hat{t}$'s and compare it to the real value of the population size. Now, before we jump into auto-pilot `R`-programming mode, note the following: the samples are being drawn from a binomial probability distribution, and one of the possible outcomes is, of course, 0 successes, or 0 marked animals in the sample. The probability of that event is not negligible:

```
> pbinom(q=0, size=n, prob=m/tot)
[1] 0.007355344
```

Thus, if we do thousands of simulations, every 7000^{th} simulation or so we would get a 0 as an outcome. Then, computing $\hat{t}$ as $(mn)/x$ would return `Inf` in `R`. What do we do? Well, this annoying numerical problem can be avoided altogether by thinking of the task at hand in a slightly different way. In fact, this change of approach lead us directly into an elegant, exact description of the long-run behavior of our estimator. Here's how:

Thus far, we have been concerned with the estimator of the total population size, $\hat{t} = (mn)/X$, where X is random. Equivalently, we could be interested in correctly estimating the proportion of marked animals in the population, m/t. Dividing both sides of the expression for $\hat{t}$ by $\hat{t}$ and multiplying it by X we get that such proportion is

$$\hat{p} = \frac{m}{\hat{t}} = \frac{X}{n}.$$

A direct answer to the questions 1 and 2 above, for each of our sampling settings (with and without replacement) can now be given, when the estimator of interest is not that for the total population size, but the estimator $\hat{p}$ of the proportion of marked animals in the population:

Question 1: On average, how would our estimator compare to the true proportion of marked animals under repeated hypothetical sampling? For either sampling scheme, the answer is the same, and is as follows: First note that X, the number of marked animals in the sample, can be expressed as a sum of n Bernoulli trials, $X_1, X_2, \ldots, X_n$ with success probability m/t:

$$X = \sum_{i=1}^{n} X_i$$

The average, or expected value of these Bernoulli trials is simply computed as: $E[X_i] = (1)\frac{m}{t} + (0)\left(1 - \frac{m}{t}\right) = \frac{m}{t}$. The variance of these trials is in turn given

by [2]:

$$\begin{aligned}\text{Var}[X_i] &= (1 - E[X_i])^2 \tfrac{m}{t} + (0 - E[X_i])^2 \left(1 - \tfrac{m}{t}\right) \\ &= \left(1 - \tfrac{m}{t}\right)^2 \tfrac{m}{t} + \left(0 - \tfrac{m}{t}\right)^2 \left(1 - \tfrac{m}{t}\right) \\ &= \tfrac{m}{t}\left(1 - \tfrac{m}{t}\right).\end{aligned}$$

Regardless of whether these trials are independent (sampling with replacement) or not (sampling without replacement), the following is true:

$$\begin{aligned}E[\hat{p}] = E\left[\tfrac{X}{n}\right] &= E\left[\tfrac{\sum_{i=1}^{n} X_i}{n}\right] \\ &= \tfrac{1}{n}\sum_{i=1}^{n} E[X_i] = \tfrac{1}{n} n E[X_i] = \tfrac{m}{t}.\end{aligned} \tag{1}$$

Wow! That is quite something! Do you see why the excitement? This result, although simple, is very profound because it says that if we are confronted with the same problem over and over, and if we were to estimate the proportion of marked animals in the population, our estimates would be on average identical to the true proportion of marked animals in the population! That's quite a re-assurance for our scientific inquiry, isn't it?

Now, before we get too carried away, there is a 'little' detail that we need to deal with. Although on average the estimates are identical to the true proportion, we don't know (yet) anything about the variability of such estimates. It may very well be that these estimates, despite being on average identical to the truth, are highly variable. If this is the scenario at hand, then for any given sample our estimate may very well be far away from the true proportion (either to the left or to the right of it). On average, however, these estimates would match the true proportion. This scenario isn't very re-assuring anymore, is it? Therefore, before we can say anything regarding the reliability of our estimates, it seems prudent to elucidate what the average departure of our estimator from its distributional mean is. This is what question 2 above is about.

Before answering this question note the following: because our estimator $\hat{p} = X/n$ inherits its randomness from the sampling scheme that we adopt, it is this sampling scheme what directly determines the probabilistic properties of our estimator. Thus, a change in the sampling scheme can result in a change in the statistical properties of our estimator, like its variability. If the variance of our estimator differs between our two sampling schemes, one being smaller than the other one, then intuitively it makes sense to choose the estimator for which the variance is the smallest. A small variance of our estimator means that, if we were to repeat the experiment over an over, under the same circumstances, all of our estimates of the true proportion of marked animals would be close to each other. Not only that, but equation 1 tells us that these estimates will be on average right smack on the money! Later on we

[2] By the way, now is a good time to check that you are familiar with the expected value and the variance of a probability distribution, and elementary manipulations of these, as described in the Appendix!

will explore these issues all within the framework of maximum likelihood estimation theory. For now, let's have some fun with some simple calculations.

Question 2: What is the average departure of our estimator from the true proportion of marked animals in the population? A straight forward answer to this question is found when we compute the variance of $\hat{p}$, which is, by definition, the average squared departure of the estimator from its mean. Expressing X as the sum of n Bernoulli trials with success probability m/t, as above, we get that

$$\begin{aligned} \mathrm{Var}[\hat{p}] = \mathrm{Var}\left[\tfrac{X}{n}\right] &= \tfrac{1}{n^2}\mathrm{Var}\left[\textstyle\sum_{i=1}^{n} X_i\right] \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}[X_i, X_j] \\ &= \tfrac{1}{n^2}\left(\textstyle\sum_{i=1}^{n}\mathrm{Var}X_i + 2\sum_{1\le i<j\le n}\mathrm{Cov}[X_i, X_j]\right). \end{aligned} \tag{2}$$

Note that in the case of sampling with replacement, the X_i's are not only identical, but necessarily independent. Hence, the covariances for all $i \neq j$ in the last line of equation 2 are null, and we get that

$$\mathrm{Var}[\hat{p}] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}X_i = \frac{\frac{m}{t}\left(1-\frac{m}{t}\right)}{n}, \quad \text{which converges to } 0 \text{ as } n \to \infty.$$

Now, what happens in the sampling without replacement case? Then, the X_i trials are no longer independent, because removing one marked individual obviously affects the number of remaining marked individuals. As a consequence, the remaining proportion of marked animals also changes. Note however that, removing one individual, which ever individual we pick, would result in exactly the same change in the proportion of marked animals. Put in another way, the X_i's are interchangeable. Thus, for every $i \neq j$, the $\mathrm{Cov}[X_i, X_j]$ should be same. Let then, name $\mathrm{Cov}[X_i, X_j]$ for all $i \neq j$ as τ. Before substituting τ into equation 2, note that in the sum $\sum_{1\le i<j\le n}\mathrm{Cov}[X_i, X_j]$ there are a total of $\binom{n}{2}$ summands (convince yourself by using $n = 4$ and writing the double sum in the second line of equation 2 in its entirety!), and hence the last line of equation 2 becomes

$$\mathrm{Var}[\hat{p}] = \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{Var}X_i + n(n-1)\tau\right).$$

Let $i = 1$ and $j = 2$, for instance. Then

$$\begin{aligned} \tau = \mathrm{Cov}[X_1, X_2] &= E[X_1X_2] - E[X_1]E[X_2] \\ &= \textstyle\sum_{x_1=0}^{1}\sum_{x_2=0}^{1} x_1x_2P(X_1 = x_1, X_2 = x_2) - \frac{m}{t}\frac{m}{t} \\ &= (1)(1)P(X_1 = 1, X_2 = 1) - \tfrac{m}{t}\tfrac{m}{t} \\ &= P(X_2 = 1|X_1 = 1)P(X_1 = 1) - \tfrac{m}{t}\tfrac{m}{t} \\ &= \tfrac{m-1}{t-1}\tfrac{m}{t} - \tfrac{m}{t}\tfrac{m}{t}. \end{aligned}$$

Plugging this expression for τ back in our variance calculation above yields (after a little algebraic manipulation -go ahead and try it-):

$$\text{Var}[\hat{p}] = \frac{\frac{m}{t}\left(1-\frac{m}{t}\right)\left(\frac{t-n}{t-1}\right)}{n},$$

which is always smaller than $\frac{\frac{m}{t}\left(1-\frac{m}{t}\right)}{n}$, the variance of $\hat{p}$ under sampling with replacement! Hence, building a more realistic sampling model, one that captures the dependencies among the samples, results in an estimator of the fraction of marked animals in the sample that also takes into account these dependencies and, as a result, has a smaller variance! Even without incorporating extra parameters, a better model of the sampling structure results in a much more reliable (*i.e.*, less variable) estimator of our fraction of marked animals in the sample!

Therefore, even though the mathematical form of the estimator is the same in our two mice sampling models, it is the sampling scheme (with or without replacement) what ultimately dictates a crucial difference in the result of our reliability test. So which estimator do we go with? With the estimator with the smaller variability (*i.e.* greater reliability). The take home message of this exercise is that one should not be fooled into thinking that the ultimate goal of our statistical analysis is parameter estimation per se. Rather, our parameter estimates aren't but the by-product of a careful inferential process, and that we should always inquire about the reliability of the very methods we use to learn from nature.

1.1.2 A trap efficiency model

Fishing weirs are ancient forms of fish traps. Their use has been recorded in many different cultures by archaeologists and historians. Native Americans for instance, used these traps to catch migrating salmonids. Nowadays, biologists use funnel-shapped trap weirs to try to estimate the total number of Sokeye salmons out-migrating in a river. To do that, they set a trap at a given point in the river and catch, mark and release a given number of fish. Call that number m, for 'marked'. Down-river, they set a second trap and make a second catch. Now suppose that out of the total number of fish marked during the first sampling, x are caught in that second catch. What is the efficiency of the trap? If t denotes the total number of fish out-migrating at the time of the first sample, then we can think of measuring trap efficiency by computing the ratio of the number of fish caught in the first sample to the total number of fish, m/t. However another way of assessing trap efficiency would be to calculate the ratio x/m, that is, the number of marked fish caught in the second trap over the total number of fish caught in the first trap. Yet, in this setting, t, the total population size is an unknown constant. Can we find an estimate of t such that our assessment of trap efficiency does not differ between the two ways of calculating it described above? In other words, can we find $\hat{t}$ such that $m/t = x/m$? Solving for t in that equality we see that $\hat{t} = \frac{m^2}{x}$. In what follows, we'll see that $\frac{m^2}{x}$ is in fact identical to the ML estimate of the total population size that results from a probabilistic description of the trapping setting described above.

Let M be a random variable that counts the number of fish caught in the first sample of the fishing experiment described above and let X be the random variable that counts the number of marked fish caught in the second sample. If the probability of capturing a given fish is an unknown constant p, we could assume that:

$$M \sim \text{Bin}\,(t, p)$$

and that

$$(X|M = m) \sim \text{Bin}\,(m, p)\,.$$

By the law of conditional probabilities, the joint distribution of M and X is:

$$\begin{aligned} P(X = x, M = m) &= P(X = x|M = m)P(M = m) \\ &= \tbinom{m}{x} p^x (1-p)^{m-x} \tbinom{t}{m} p^m (1-p)^{t-m} \\ &= \ell(p, t). \end{aligned}$$

The ML estimates of p and t are the values that jointly maximize $\ell(p, t)$. Four questions about those estimates are in order:

1. Are these estimates biased?
2. What kind of distribution do they have?
3. How small is the variance of the distribution of the ML estimates?

4. Does the estimates get closer and closer to the true values as the sample size tends to infinity?

The answers to these questions, reviewed in detail later in the course, are summarized as follows:

1. As sample size tends to infinity, the ML estimates are unbiased.

2. As sample size tends to infinity, the variance of the ML estimates is the smallest possible. Later we'll see the Cramer-Rao inequality which gives a lower bound on the variance of *any* unbiased estimate. An unbiased estimate whose variance achieves this lower bound is said to be **efficient**. Since the asymptotic variance of a ML estimate is equal to the lower bound, these estimates are said to be asymptotically efficient. For finite sample sizes, ML estimates may not be efficient (Rice 1995).

3. The distribution of the ML estimates is assymptotically normal and

4. the estimates are statistically consistent.

To calculate the ML estimates, we first compute the partial derivative of $\ln \ell(p,t)$ with respect to p, set it equal to 0 and solve for p.

$$\begin{aligned} \frac{\partial \ln \ell(p,t)}{\partial p} &= \frac{x}{p} - \frac{m-x}{1-p} + \frac{m}{p} - \frac{t-m}{1-p} = 0 \\ &= (m+x)\frac{1}{p} - (t-x)\frac{1}{1-p} = 0, \end{aligned}$$

and hence

$$(m+x)\frac{1}{p} = (t-x)\frac{1}{1-p} \Rightarrow \hat{p} = \frac{m+x}{m+t}.$$

With the ML estimate of p expressed in terms of the other parameter, we can substitute it in the log-likelihood function and obtain an expression that is now a function of just one unknown parameter, t:

$$\ell(\hat{p},t) = \binom{t}{m}\hat{p}^m(1-\hat{p})^{t-m}\binom{m}{x}\hat{p}^x(1-\hat{p})^{m-x}.$$

Now we can attempt maximizing the log-likelihood function with respect to the parameter t. However, note that just as in the hypergeometric model, a complicated combinatorial term is involved in the derivative of the likelihood function with respect to the parameter of interest t. So, even before substituting p by $\hat{p}$, in order to avoid the hassle of computing such derivative we do the integer trick used in the hypergeometric mark-recapture model. After some algebra, we get that the ML estimate of t is:

$$\hat{t} = \frac{m}{p}.$$

Homework: Your task is to show that $\hat{t} = \frac{m}{p}$ is the integer ML estimate of t. Such estimate is the value of t that satisfies

$$\frac{\ell(\hat{p}, t)}{\ell(\hat{p}, t-1)} = 1.$$

Now that we have the integer ML estimate of t as a function of p, we substitute p by $\hat{p} = \frac{m+x}{t+m}$ and solving for t we get that

$$\hat{t} = \frac{m.m}{x},$$

which matches our initial intuition!

1.1.3 Spatial and temporal randomness

Key ecological processes, such as competition, predation, survival and reproduction occur randomly along the axis of time and space. Accounting for randomness in the spatial component of such process has been key to achieve some understanding of their functioning (Pielou, 1969). In what follows, we derive two models: a simple spatial model of counts of individuals in a given number of sample quadrats and a temporal model to account for the outcome of a fishing experiment.

Spatial model:
Suppose for instance, that we want to make inferences about the mean number of lodgepole pines per unit of area in a forest, and we sample 100 quadrats of size a selected at random within an extense territory of size A. Since each time we sample a quadrat we can potentially count a different number of trees, we will define "$X = \#$ of plants in a randomly located quadrat" as our random variable of interest. Assume further that the location of each quadrat is determined independently form each other. Let n be the total number of plants in the study region. Then, $\frac{a}{A}$ may be taken as the probability that any particular plant is in the sample plot and we can use again the binomial distribution to define a model of spatial randomness, *i.e.*:

$$X \sim \text{Bin}\left(n, \frac{a}{A}\right).$$

Under this model, the mean number of trees per unit area would be $\frac{n}{A}$. As A and n get very large such that $\frac{n}{A} \to \lambda$, a constant,

$$\binom{n}{x}\left(\frac{a}{A}\right)^x\left(1-\frac{a}{A}\right)^{n-x} \to \frac{\mathrm{e}^{-\lambda a}(\lambda a)^x}{x!}, \quad x = 0, 1, 2, 3, \ldots{}^{3}$$

Thus, under the assumption that $p = \frac{a}{A}$ is very small and n is very large, we can approximate the binomial distribution X defined above with a Poisson random variable

[3] It is a straightforward exercise to show that this is true, by taking into consideration the fact that

$$\lim_{n\to\infty}\left(1-\frac{b}{n}\right)^n = \mathrm{e}^{-b}$$

for some finite number b (Go ahead and do it, but no fair peeking *Wikipedia*!).

Table 1: Number of pines present in each of 100 quadrats and estimated expected frequency of pines under the Poisson spatial model.

Data of # of pine trees/quadrat	**observed frequency**	**estimated expected frequency**
0	7	$q\widehat{P}(X=0)$
1	16	$q\widehat{P}(X=1)$
2	20	$q\widehat{P}(X=2)$
3	24	$q\widehat{P}(X=3)$
4	17	$q\widehat{P}(X=4)$
5	9	$q\widehat{P}(X=5)$
6	5	$q\widehat{P}(X=6)$
≥ 7	2	$q\left(1-\sum_{x=0}^{6}\widehat{P(X=x)}\right)$

with mean λa. Suppose a field biologist actually goes out to a lodgepine forest and counts the number of pines present in each of 100 quadrats of size a and then counts the number of quadrats with $0, 1, 2, 3, 4, 5, 6$ and ≥ 7 trees. The obtained counts are summarized in the first two columns of table 1. Because n is potentially large with respect to $p = a/A$, then the binomial model can be approximated with a Poisson distribution. To estimate the parameter of interest λ, we first write down the joint distribution of the data and substitute in the data to obtain the likelihood function:

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \ldots, X_q = x_q) &= \frac{\mathrm{e}^{-\lambda a}(\lambda a)^{x_1}}{x_1!} \cdots \frac{\mathrm{e}^{-\lambda a}(\lambda a)^{x_q}}{x_q!} \\ &= \frac{\mathrm{e}^{-\lambda a q}(\lambda a)^{x_1+x_2+\ldots+x_q}}{x_1!x_2!\ldots x_q!} \\ &= \ell(\lambda). \end{aligned}$$

As before, after doing that we maximize $\ell(\lambda)$ by finding the point where the derivative of $\ln \ell(t)$ is zero:

$$\frac{d \ln \ell(t)}{d\lambda} = -aq + \frac{1}{\lambda}\sum_{i=1}^{q} x_i = 0 \Rightarrow \hat{\lambda} = \frac{1}{aq}\sum_{i=1}^{q} x_i = \frac{\bar{x}}{a},$$

where $\bar{x}$ is the sample mean of the number of trees per cuadrat and a is the cuadrats area. Note that $\hat{\lambda}a = \bar{x}$. With the estimate of $\hat{\lambda}a$ in hand, we can readily calculate the estimated expected frequency of observations of the value x in a sample of size q, as shown in the third column of table 1. Later on we will see how to formally compare the expected vs. the observed frequencies.

Temporal model:
In fisheries, it is a common task to estimate the average catch per unit of effort.

day	Total effort (hrs.)	Total # of Steelheads caught
1	51.85 (t_1)	2 (x_1)
2	48.50 t_2)	2 (x_2)
3	50.20 (t_3)	1(x_3)
4	52.53 (t_4)	1 (x_4)
5	65.37 (t_5)	4 (x_5)
6	70.12 (t_6)	5(x_6)

Table 1.1.3 shows data from a fly-fishing experiment at a single site in LG River in 1988. We can extend the idea of using the Poisson distribution as a model for spatial randomness to model the occurrences of the catches in time as a Poisson process with a given mean per unit of time.

Exercises:

1. Your task is to find the ML estimator of the average catch per unit of effort $\hat{\lambda}$ and compute it using the following Poisson model:

$$P(X = x) = \frac{\mathrm{e}^{-\lambda t}(\lambda t)^x}{x!},$$

   where $X = \#$ of fish caught in t hours.

2. Calculate the ML estimate of λa for the Lodgepole pines example. Once you do that, calculate the expected frequencies of the number of pines in a sample of size q (*i.e.*, fill in with values the third column of table 1) and graph the observed vs. expected frequencies.

1.1.4 The likelihood function for continuous probability models

Modeling waiting times:
A fisherman is sitting in his boat at dawn, in the middle of a calm, small lake just counting the number of fish caught during a period of time t (gosh, I wish I was in this fisherman's boots and $t \approx 12$ hours!!). From our spatial and temporal randomness lecture, we know that we could model the number of fish caught during a period of time t as a Poisson process. That is, we let

$$P(X(t) = x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \ldots$$

The parameter λ is the average number of captures per unit of time. Now, let S be the (random) waiting time until catching the first fish. S is a random variable defined in the positive real line. Let s denote a waiting time value (a positive, real number) and consider the following two events: $[X(s) = 0]$ and $[S > s]$. The first event says that the number of fish caught after waiting a period of time s is equal to 0 and the second event says that the waiting time until catching the first fish is strictly greater

than s. These two events are in fact the same event, so their probability has to be the same. We therefore have that

$$P(X(s) = 0) = P(S > s),$$

which implies that

$$P(X(s) > 0) = 1 - P(X(s) = 0) = P(S \leq s).$$

Using our Poisson model for X(s), we then have that

$$F(s) = P(S \leq s) = 1 - P(X(s) = 0) = 1 - e^{-\lambda s}, \quad (0 < s < \infty).$$

By definition (see Appendix), $F(s)$ is the cumulative distribution function (cdf) of the random variable s. When plotted, it is easy to see that the monotonous, continuous and increasing function $F(s)$ is telling us that if the fisherman waits long enough, he is almost certain to catch a fish. Using the cdf we can answer many questions. For instance, for two positive values a and b such that $a < b$ we could use the cdf and the additivity property of integrals to find out what's the value of $P(a < S \leq b)$:

$$\begin{aligned} P(a < S \leq b) &= P(S \leq b) - P(S \leq a) \\ &= F(b) - F(a) \\ &= 1 - e^{-\lambda b} - (1 - e^{-\lambda a}) \\ &= e^{-\lambda a} - e^{-\lambda b}. \end{aligned}$$

Let Δs represent a small positive change in a realized waiting time, so that $(s, s+\Delta s)$ is a small time interval. Then, according to the above calculation we have that

$$P(s < S \leq s + \Delta s) = F(s + \Delta s) - F(s)$$

and dividing both sides of the equation by the length of this small interval we get a measure of the *density* of probability over the interval $(s, s + \Delta s)$.

$$\frac{P(s < S \leq s + \Delta s)}{\Delta s} = \frac{F(s + \Delta s) - F(s)}{\Delta s}.$$

As $\Delta s \to 0$, the ratio above converges to the derivative of $F(s)$, denoted by $f_S(s)$:

$$\lim_{\Delta s \to 0} \frac{P(s < S \leq s + \Delta s)}{\Delta s} = \frac{dF(s)}{ds} = f_S(s) = \lambda e^{-\lambda s}.$$

The derivative of $F(s)$, $f_S(s)$ is the associated probability distribution function of the random variable S. It is the continuous distribution's equivalent to the probability mass function (see Appendix). Thus, by analogy with the discrete case this is the mathematical object that will be used to define the likelihood function, needed to estimate the parameter λ. This is accomplished as follows:
Suppose that the fisherman in question is on vacations and records the time until catching the first fish in each one of the n fishing occasions he has during his free days (if he is lucky, n is a big number!). At the end of the n sampling (fishing)

occasions he has the following list of waiting times until first capture: $s_1, s_2, \ldots, s_n$ and wishes to use this list to estimate his average number of captures per unit of time. In the discrete case, we could actually plug in the data in the probability mass function and evaluated it for different parameter values to plot the likelihood curve. Here however, note that the probability of a particular observation is 0, since the area under the curve along an interval of length 0 is 0. How do we write down the likelihood function for this continuous case, then? Well, suppose that the precision of the time measuring instrument is given by a small, positive number ϵ. What we could do is to calculate what is known as the *exact likelihood function* for continous data sets, which for a single observation s_1 is the probability measure over a small interval surrounding the observation. If you think of it, considering that many times our data-measuring instruments can be thought of as bounded with a precision ϵ, if the probability of an observation (a point value) is zero, then it does makes sense to calculate the likelihood as the process' probability measure evaluated at the small interval $(\text{observation} - \frac{\epsilon}{2}, \text{observation} + \frac{\epsilon}{2})$. That is,

$$P\left(s_1 - \frac{\epsilon}{2} < S \leq s_1 + \frac{\epsilon}{2}\right) = F\left(s_1 + \frac{\epsilon}{2}\right) - F\left(s_1 - \frac{\epsilon}{2}\right).$$

One of the most important calculus theorems, the mean-value theorem allows us to approximate this exact probability calculation, provided ϵ is small enough with the derivative of $F(s)$, the probability density function as follows:

$$P\left(s_1 - \frac{\epsilon}{2} < S \leq s_1 + \frac{\epsilon}{2}\right) = F\left(s_1 + \frac{\epsilon}{2}\right) - F\left(s_1 - \frac{\epsilon}{2}\right) \approx \epsilon f(s_1).$$

In various mathematical statistical books like Rice (1995) the likelihood function for continuous models is defined as the pdf evaluated at the observations, period. The above argument shows that such definition is in fact an approximation to the *exact* likelihood function defined in terms of the cumulative distribution function. And there are precise cases where this approximation *does not* work and then, this pdf version of the likelihood function has serious mathematical problems (in particular, it may have mathematical singularities). In those cases however, the exact likelihood definition does not have problems, because being defined as a difference of probabilities (evaluations of the cdf function) it is always a number bounded between 0 and 1. This is a very important issue often ignored while doing maximum likelihood, to the extent that some papers have stated that the method of maximum likelihood does not work when in fact, it's the approximation to the exact likelihood function the one that does not work (a detailed account of this topic can be found in Montoya et al 2009).

For a discrete probability model, likelihood function for the set of observations is expressed as the joint probability of the observations, which from the assumption of *independent samples* can be written as the product of the individual probabilities for each observation. By the same token, the likelihood for the set of recorded waiting times until the first capture is written as the joint probability of the observations $\pm\frac{\epsilon}{2}$, which from the independence assumption becomes the product of the probabilities of each one of these intervals:

$$P\left(s_1 - \frac{\epsilon}{2} < S_1 \leq s_1 + \frac{\epsilon}{2}, s_2 - \frac{\epsilon}{2} < S_2 \leq s_2 + \frac{\epsilon}{2}, \ldots, s_n - \frac{\epsilon}{2} < S_n \leq s_n + \frac{\epsilon}{2}\right) =$$

$$P\left(s_1 - \frac{\epsilon}{2} < S_1 \leq s_1 + \frac{\epsilon}{2}\right) P\left(s_2 - \frac{\epsilon}{2} < S_2 \leq s_2 + \frac{\epsilon}{2}\right) \ldots P\left(s_n - \frac{\epsilon}{2} < S_n \leq s_n + \frac{\epsilon}{2}\right).$$

Using the mean value theorem, this last product can be approximated with

$$f_S(s_1) f_S(s_2) \ldots f_S(s_n) \epsilon^n.$$

Therefore, the likelihood function used to estimate the unknown parameter λ is written as

$$L(\lambda) = f_S(s_1) f_S(s_2) \ldots f_S(s_n) \epsilon^n = \epsilon^n \left(\lambda e^{-\lambda s_1}\right) \left(\lambda e^{-\lambda s_2}\right) \ldots \left(\lambda e^{-\lambda s_n}\right).$$

The ML estimates are then found by maximizing this last expression in terms of the parameter λ. Since ϵ does not depend on the parameter of interest λ, when we compute the log-likelihood and then maximize it, the quantity ϵ^n does not play a role in the estimation process and hence we may write

$$L(\lambda) \propto f_S(s_1) f_S(s_2) \ldots f_S(s_n).$$

Note that this approximation only works when the precision of the measuring instrument does not depend in any way on the parameter of interest (here the parameter λ, which stands for to the mean number of captures per unit of time). It is precisely when a dependence between ϵ and the parameter of interest exists that this mean value theorem approximation breaks down. (In the context of this exponential model, such dependence could occur if the precision of the instrument were to decay a little bit each time an event is recorded. Then, if the events occur at a fast rate, the precision of the instrument decays quickly). In this case, the correct approach is to calculate the exact likelihood function via the cdf exact likelihood. This is a very important problem often obviated in statistical analyses. Now, the log-likelihood function is written as

$$\begin{aligned} \ln L(\lambda) &\propto \ln\left(\lambda^n \exp -\lambda \textstyle\sum_{i=1}^n s_i\right) \\ &= n \ln \lambda - \lambda \textstyle\sum_{i=1}^n s_i, \end{aligned}$$

which allows us to calculate the maximum likelihood estimate of λ after taking the first derivative and set it equal to 0:

$$\begin{aligned} \tfrac{d \ln \ell(\lambda)}{d\lambda} &= \tfrac{n}{\lambda} - \textstyle\sum_{i=1}^n s_i = 0 \\ &\Rightarrow \hat{\lambda} = \tfrac{n}{\sum_{i=1}^n s_i} = \tfrac{1}{\bar{s}}. \end{aligned}$$

In other words, the ML estimate of the capture rate is the inverse of the average waiting time until the first capture computed from the sample, $\bar{s} = \frac{1}{n} \sum_{i=1}^n s_i$.

The preceding approach to model the waiting times until the 1^{st} capture can be readily extended to the case where we are interested in modeling the waiting times until the k^{th} capture: Let $X(t)$ be the random variable that counts the number of fish trapped during t time units. Let S_k be the continuous random variable measuring the waiting time until trapping the k^{th} fish. Again, let's consider the following two

events, $[X(s) < k]$ and $[S_k > s]$. The first event says that the number of fish caught up to time s is strictly less than k and the second event says that the waiting time until the k^{th} capture is strictly greater than s. Thus, these two events are telling the same story, and hence, have the same probability, *i.e.*,

$$P(X(s) < k) = P(S_k > s) \quad \text{and therefore}$$

$1 - P(X(s) < k) = P(S_k \leq s) = F_k(s), \quad$ which is the cdf of the random variable S_k.

Here again, we can use the Poisson random variable as a probabilistic model for $X(t)$, that is

$$P(X(s) = x) = \frac{e^{-\lambda s}(\lambda s)^x}{x!}\, x = 0, 1, 2, 3, \ldots$$

By adopting the Poisson model, we can readily write an expression for $P(X(s) < k)$ and hence solve for the the cdf of S_k:

$$F_k(s) = P(S_k \leq s) = 1 - P(X(s) < k) = 1 - \sum_{x=0}^{k-1} \frac{e^{-(\lambda s)}(\lambda s)^x}{x!}.$$

Just as with the exponential model, we can find the probability density function of the waiting time until capturing the k^{th} by taking the derivative of $F_k(s)$ with respect to s:

$$\begin{aligned}
f_S(s) &= \tfrac{d}{ds}\left[1 - \textstyle\sum_{x=0}^{k-1} \frac{e^{-(\lambda s)}(\lambda s)^x}{x!}\right] \\
&= -\textstyle\sum_{x=0}^{k-1} \frac{d}{ds}\left[\frac{e^{-(\lambda s)}(\lambda s)^x}{x!}\right] \\
&= -\textstyle\sum_{x=0}^{k-1} \left[-\frac{\lambda e^{-(\lambda s)}(\lambda s)^x}{x!} + \frac{e^{-(\lambda s)} x s^{x-1}\lambda^x}{x!}\right] \\
&= \textstyle\sum_{x=0}^{k-1} \left[\frac{\lambda e^{-(\lambda s)}(\lambda s)^x}{x!} - \frac{e^{-(\lambda s)} x s^{x-1}\lambda^x}{x!}\right] \\
&= \textstyle\sum_{x=0}^{k-1} \left[\frac{\lambda^{x+1} e^{-(\lambda s)} s^x}{x!}\right] - \sum_{x=1}^{k-1} \frac{\lambda^x s^{x-1} e^{-\lambda s}}{(x-1)!}
\end{aligned}$$

Now, in the two summation terms above, we can factor out the term $e^{-\lambda s}$ and explicitly write down the sums as follows:

$$f_k(s) = e^{-\lambda s} \left\{ \begin{array}{l} \lambda + \lambda^2 s + \frac{\lambda^3 s^2}{2!} + \ldots + \frac{\lambda^{k-1} s^{k-2}}{(k-2)!} + \frac{\lambda^k s^{k-1}}{(k-1)!} \\ \\ -\lambda - \lambda^2 s - \frac{\lambda^3 s^2}{2!} - \ldots - \frac{\lambda^{k-1} s^{k-2}}{(k-2)!} \end{array} \right.$$

Note that all the terms of the first sum are canceled with one of the terms in the second sum (this is an example of what in mathematics is known as a *telescoping sum*), except for the last term. Therefore, the above equation reduces to

$$f_k(s) = \frac{e^{-\lambda s}\lambda^k s^{k-1}}{(k-1)!}, \quad 0 < s < \infty.$$

The above expression corresponds to the probability density function of a gamma distribution. Thus, the waiting time until the k^{th} capture can be modeled with a gamma distribution. In particular, if $k = 1$, the waiting time until the first capture is exponentially distributed. Now consider the following scenario: starting at time $t = 0$ we start a clock and measure the time until the first fish capture. This waiting time is exponentially distributed. Once the first fish is caught, we re-start the clock and measure the time until the next capture. That time is also exponentially distributed (the reason why this is so is because of the "memory-less" property of the exponential distribution). Iterating this argument, it follows that if the events of interest are the fish captures, then the *inter-event times* for this Poisson process can be modeled with an exponential distribution. The preceding derivation of the gamma distribution shows that the sum of k (independent) exponential distributions follows the gamma distribution.

The use of the gamma distribution is certainly not constrained to modeling waiting times until the k^{th} event. In particular, the value of k can be different than an integer. In that case, the factorial function $(k-1)!$ is not defined and the gamma function $\Gamma(.)$ is used instead. A note about the gamma function is in order. This function can be seen as the continuous version of the factorial function, and indeed, if k is an integer,

$$(k-1)! = \Gamma(k).$$

When k is not an integer, yet still a number in $(0, \infty)$, then the gamma function is given by

$$\Gamma(k) = \int_0^\infty u^{k-1} e^{-u} du. \tag{3}$$

To see why this is so, consider the above expression for $f_k(s)$, the probability density function. Because it is a probability density function, it has to integrate to 1 over the interval of the values allowed for s which is $(0, \infty)$, that is

$$\int_0^\infty f_k(s) ds = \int_0^\infty \frac{e^{-\lambda s} \lambda^k s^{k-1}}{(k-1)!} ds = 1$$

Hence,

$$\int_0^\infty e^{-\lambda s} s^{k-1} ds = \frac{(k-1)!}{\lambda^k}. \tag{4}$$

The integral above can be conveniently manipulated to obtain eq. (4) by making a change of variables. Let $u = \lambda s$. Then, $s = \frac{1}{\lambda} u$ and $ds = \frac{1}{\lambda} du$ and the integral can be re-written as:

$$\begin{aligned} \int_0^\infty e^{-\lambda s} s^{k-1} ds &= \int_0^\infty e^{-u} \left(\tfrac{1}{\lambda} u\right)^{k-1} \left(\tfrac{1}{\lambda}\right) du \\ &= \tfrac{1}{\lambda^k} \int_0^\infty u^{k-1} e^{-u} du, \end{aligned}$$

and it follows that

$$\frac{1}{\lambda^k} \int_0^\infty u^{k-1} e^{-u} du = \frac{(k-1)!}{\lambda^k},$$

which implies that

$$\int_0^\infty u^{k-1}e^{-u}du = (k-1)!.$$

The mathematician Euler used the above equation to define that continuous version of the factorial function and called it the Gamma function:

$$\int_0^\infty u^{k-1}e^{-u}du = (k-1)! = \Gamma(k)$$

Fortunately, we can readily see how this function of k looks like by using `R`, so here's the code. Before you jump in the code, note that I show here how you can include greek letter labels in your plots.

```
### Gamma function plot:
ks <- seq(from=0.05,to=5.2, by=0.001)
gamma.ftn <- gamma(ks)
yylab <- expression(Gamma*"(k)")
mmain <- "Plotting the gamma function"
plot(ks,gamma.ftn,type="l",col="red",xlab="values of k",ylab=yylab,main=mmain)
points(c(0.5,1.0:5.0),gamma(c(0.5,1.0:5.0)), pch=16)
```

The plot is shown in figure 2. One last useful feature of the gamma function is that it can be used to compute the combinatorial terms that occur so often when working with discrete probability distributions. Using the fact that

$$\Gamma(k+1) = k\Gamma(k),$$

and letting a and b be two positive integers, we may write

$$\binom{a}{b} = \frac{a(a-1)(a-2)\dots(a-b+1)}{b!} = \frac{\Gamma(a+1)}{\Gamma(b+1)\Gamma(a-b+1)}. \tag{5}$$

In the mark-recapture examples we mentioned that calculating functions such as $\binom{a}{b}$ using `R` involve multiplying many numbers, something that can be computationally expensive. A better (most efficient computationally) way to compute such functions is to take the logarithm of the expression and then exponentiate it, because then all the products become sums, a far easier task for the computer. For example, to compute $\binom{a}{b}$ we first take the log, which yields

$$\ln\left[\binom{a}{b}\right] = \ln\Gamma(a+1) - \ln\Gamma(b+1) - \ln(a-b+1),$$

and then exponentiate the result. In `R` we would type

```
a <- 30
b <- 5
ln.achooseb <- lgamma(a+1) -lgamma(b+1) - lgamma(a-b+1)
achooseb <- exp(ln.achooseb)
print(achooseb)
```

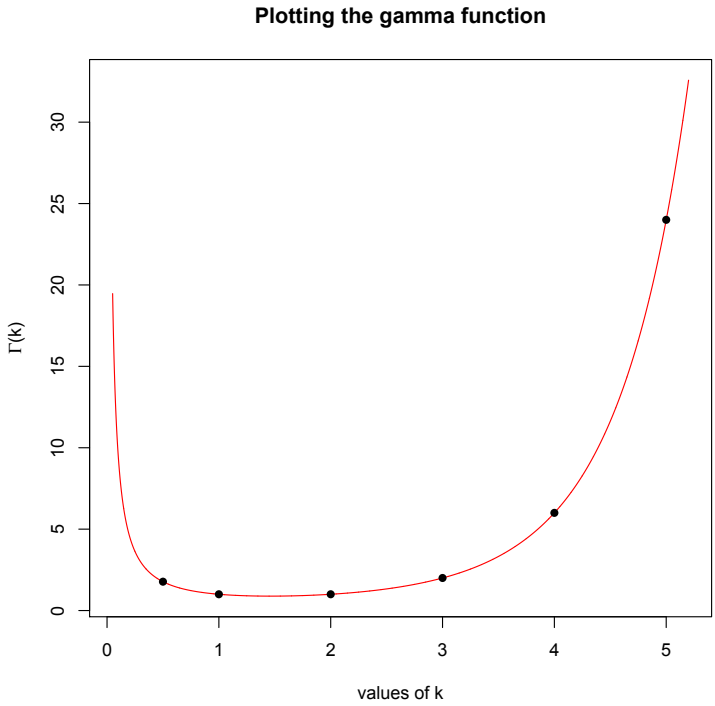

Figure 2: Plot of the gamma function $\Gamma(k)$, where $0 < k < \infty$. Solid dots indicate the values of the gamma function for the integers $1, 2, 3, \ldots$ Also shown with a solid dot is the value of the gamma function for $k = \frac{1}{2}$, which is $\sqrt{\pi} \approx 1.772454$.

With the gamma function in our toolbox, we may now define the general formulation of the gamma distribution's pdf:

$$f(s) = \frac{\lambda^k}{\Gamma(k)} s^{k-1} e^{-\lambda s}, \text{ where}$$

$0 < s < \infty, 0 < \lambda < \infty$ and $0 < k < \infty$. This general version of the gamma distribution is extremely versatile, as shown in figure 3. Here is the code for that plot:

```
svec <- seq(from=0.35,to=40.0, by=0.01)
lambda <- 0.5 # a mean of 0.5 fish per hour
pdfk0.5 <- dgamma(x=svec,shape=0.50,scale=(1/lambda))
svec.exp <- seq(from=0.0,to=40.0, by=0.01)
pdfk1 <- dgamma(x=svec.exp,shape=1,scale=(1/lambda))
pdfk2.5 <- dgamma(x=svec,shape=2.5,scale=(1/lambda))
pdfk15 <- dgamma(x=svec,shape=10,scale=(1/lambda))
plot(svec,pdfk0.5, xlab="s",type="l", ylab="Prob. density")
text(x=2.3,y=0.02,"k=0.5")
points(svec.exp,pdfk1, type="l")
```

```
text(x=5.5,y=0.065,"k=1")
text(x=3.0,y=0.5025,expression(lambda*"=0.5"))
points(svec,pdfk2.5, type="l")
text(x=8.0,y=0.1,"k=2.5")
points(svec,pdfk15, type="l")
text(x=23,y=0.065,"k=10")
points(0.0,0.5,pch=16,cex=0.75)
```

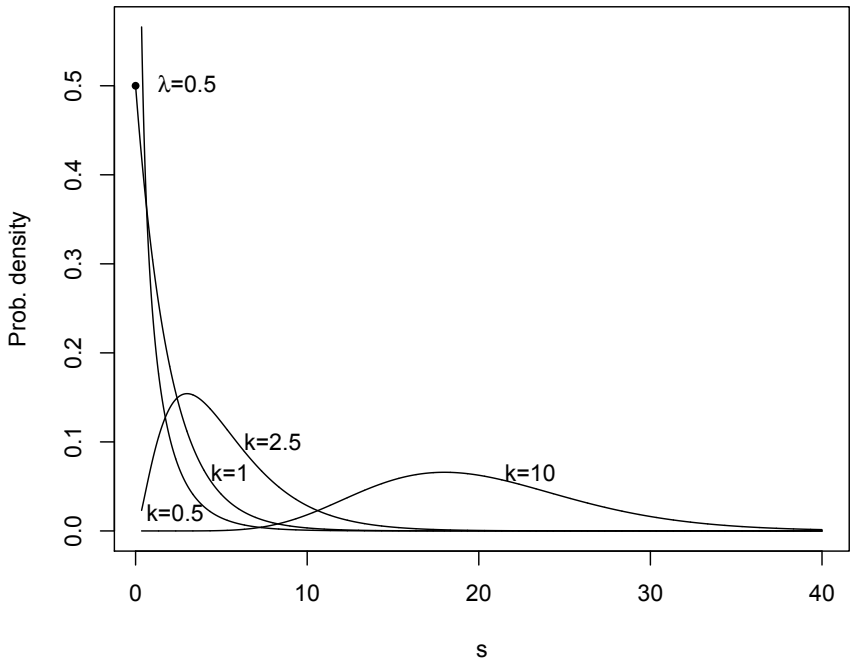

Figure 3: Plot of the probability density function of the gamma distribution for various values of k, the shape parameter.

A couple of points about figure 3 are worth mentioning. First, when $k = 1$, $f(s) = \lambda e^{-\lambda s}$, and so the value of the intercept is $f(0) = \lambda e^{-\lambda 0} = \lambda$. Second, when k is very large, the gamma distribution approaches the normal distribution, and we get a hint at this fact from figure 3. Besides the exponential distribution, another notable particular case of the gamma distribution is the chi-square distribution with r degrees of freedom, which is obtained by making $\lambda = \frac{1}{2}$ and $k = \frac{r}{2}$ in the above pdf. Formally, if S is chi-square distributed with r degrees of freedom, we write

$$S \sim \chi^2_r.$$

Another interesting fact of the gamma distribution is the following: if $k = 1, 2, 3, \ldots$

and if $S \sim \text{gamma}\,(\alpha, k)$ then,

$$\begin{aligned} P(S \leq s) &= \int_0^s \frac{\lambda^k}{\Gamma(k)} s^{k-1} e^{-\lambda s} ds \\ &= 1 - \sum_{x=0}^{k-1} \frac{e^{-s\lambda}(\lambda s)}{x!} \\ &= \sum_{x=k}^{\infty} \frac{e^{-s\lambda}(\lambda s)}{x!}, \quad \text{which is the right tail of the initial Poisson model.} \end{aligned}$$

Thus, the right tail (*i.e.* from k to ∞) of our initial probabilistic model of the number of fish caught during a period of time s is in fact identical to the left tail of the resulting gamma model of the waiting time until catching the k^{th} fish.
Since the publication of a seminal paper by Fisher, Corbet and Williams in 1943 (J. of Animal Ecology, 12:42), the gamma distribution has played a key role in modeling heterogeneity in ecology. In 1969, the statistical ecologist Evelyn Christine Pielou laid out in her book entitled "Introduction to mathematical biology" some of the fundamental concepts and ideas along the formal statistical framework to model heterogeneity in ecology. Of those ideas, some of the most influential come from the spatial and temporal models for randomness. The Poisson model of animal abundance explained above is in fact, one of these influential models. One of her models for heterogeneity is based on the use of the gamma distribution. Explaining and illustrating the details of this model are the contents of the next section.

1.1.5 Heterogeneity in ecology: a spatial model (Pielou, 1969)

Recall our Poisson spatial model of animal or plant abundances, where it is assumed that if all spatial units available to them were identical, then their distribution pattern would be homogeneous. Then, the total number of individuals appearing in a randomly located quadrat of size s can be modeled with a Poisson random variable $X(s)$ with mean λs. That is,

$$P(X(s) = x) = \frac{e^{-\lambda s}(\lambda s)^x}{x!},\ x = 0, 1, 2, 3, \ldots$$

Now, suppose that habitable units are dissimilar among them. The dissimilarities between these spatial units could arise simply because some of these units provide more favorable environments than others. As a consequence, the mean number of individuals per unit area λ may change dramatically from one spatial unit to the other. Thus, accounting for spatial heterogeneity while modeling animal or plant abundances, amounts to specifying a quantitative framework for these changes in λ. A first, naive approach would be to try to estimate as many values of λ as habitable units are available for sampling. Such scheme however can easily lend itself to a case where we have more parameters than data available for estimation. Instead, we consider a *probabilistic mechanism* modeling the patterns of variation of the mean number of individuals per unit area. Such task involves specifying a probability distribution for the mean number of individuals per unit area. To be consistent with our

previous notation, we let Λ (capital λ) denote the random variable modeling the values of the mean number of individuals per unit area. Based on biological grounds, it is difficult to justify the use of one particular probability distribution in this case. The only thing we know is that such random variable should take on positive, real values. When confronted with such problem, Pielou (1969) proposed using the gamma distribution (also known as Pearson type III distribution). The reason for choosing this distribution, she wrote, "is that whatever λ's true distribution it is likely that some type III curve can be found to approximate it closely". Indeed, as it is shown in figure 3, the gamma distribution can adopt many different shapes. The argument brought forth by Pielou had been used before by Fisher, Corbet and Williams' 1943 paper. Such modeling argument seems at first a somewhat phenomenological approach because we try to match a pattern in nature (the pattern of variation of the mean number of individuals per unit area) based on a particular mathematical model's shape. Whenever possible, the modeling efforts should instead be directed towards the formulation of a probabilistic model starting from first biological principles. This second approach focuses on modeling the (biological) processes that generate variation, not on the patterns. Accordingly, in our example such approach would consist of modeling the variability in the mean number of individuals per unit area as a result of a hypothetical biological mechanism that generates heterogeneity in the habitability of the spatial units. Such mechanisms could be very diverse and one must decide before hand which alternatives are of more interest, given the system, data and nature of the problem. If however, one can think of a myriad of biological mechanisms (or interactions between them) that may generate heterogeneity in habitability, then the approach taken by Pielou is a nice compromise between getting a somewhat realistic representation of the patterns of variability in λ and model simplicity.
Let the gamma probability distribution function (pdf) of Λ be

$$g_\Lambda(\lambda) = \frac{\alpha^k}{\Gamma(k)}\lambda^{k-1}e^{-\alpha\lambda},\ 0 < \lambda < \infty.$$

If $d\lambda$ denotes a small (positive) change in the value of λ, then the probability of picking a value of the mean number of individuals per unit area between λ and $\lambda + d\lambda$ is given by

$$P(\lambda < \Lambda \leq \lambda + d\lambda) \approx g_\Lambda(\lambda)d\lambda.$$

For a particular habitable unit, once we draw a value of λ at random from our gamma model above, the number of individuals in a quadrat can be modeled with a Poisson distribution. Mathematically, this is expressed by writing the conditional distribution of $X(s)$ given $\Lambda = \lambda$ as a Poisson distributed random variable, *i.e.*

$$P(X(s) = x|\Lambda = \lambda) = f(x|\lambda) = \frac{e^{-\lambda s}(\lambda s)^x}{x!}.$$

However, given a data set like the Lodgepole pines quadrat samples, the way to connect our probabilistic model with the data is by means of the marginal, unconditional distribution of $X(s)$. To find it, we average $(X(s) = x|\Lambda = \lambda)$ over all the possible

values that λ can take on. To write down this average, we use the definition of the expected value of a function of a random variable (see appendix), which is the following integral:

$$\begin{aligned} P(X(s)=x)=f(x) &= \int_0^\infty f(x|\lambda)g(\lambda)d\lambda \\ &= E_g\left[f(x|\lambda)\right]. \end{aligned}$$

Thus, the integral above says that in order to find the probabilistic law of the number $X(s)$ of individuals found in a sample quadrat that takes into account the heterogeneity in λ, we have to *average* each of the Poisson probabilities over all the λ values. The average, or expectation, is taken with respect to the distribution $g(\lambda)$ of the mean number of individuals per unit area (This is why we write $E_g[\ldots]$). The integral above is readily solved, by considering eq(4) re-written using the gamma function. Let κ and β be two positive constants. Then,

$$\int_0^\infty x^{\kappa-1}e^{-\frac{x}{\beta}}dx=\Gamma(\kappa)\beta^\kappa,$$

which allows us to solve the integral directly as follows:

$$\begin{aligned} P(X(s)=x) &= \int_0^\infty f(x|\lambda)g(\lambda)d\lambda \\ &= \int_0^\infty \frac{e^{-\lambda s}(\lambda s)^x}{x!}\frac{\alpha^k}{\Gamma(k)}\lambda^{k-1}e^{-\alpha\lambda}d\lambda \\ &= \frac{\alpha^k s^x}{\Gamma(k)x!}\int_0^\infty e^{-\lambda(\alpha+s)}\lambda^{x+k-1}d\lambda \qquad (6) \\ &= \frac{\alpha^k s^x}{\Gamma(k)x!}\Gamma(x+k)\left(\frac{1}{\alpha+s}\right)^{x+k} \\ & \quad \frac{\Gamma(x+k)}{\Gamma(k)x!}\left(\frac{s}{s+\alpha}\right)^x\left(\frac{\alpha}{s+\alpha}\right)^k,\ x=0,1,2,3,\ldots \end{aligned}$$

Note that (see eq. (5))

$$\begin{aligned} \frac{\Gamma(x+k)}{\Gamma(k)x!} &= \frac{(x+k-1)\Gamma(x+k-1)}{\Gamma(k)x!} \\ &= \frac{(x+k-1)(x+k-2)\Gamma(x+k-2)}{\Gamma(k)x!} \\ &= \frac{(x+k-1)(x+k-2)\Gamma(x+k-2)\ldots k\Gamma(k)}{\Gamma(k)x!} \\ &= \frac{(x+k-1)((x+k-1)-1)((x+k-1)-2)\ldots((x+k-1)-(x-1))}{x!} \\ &= \frac{(x+k-1)((x+k-1)-1)((x+k-1)-2)\ldots((x+k-1)-x+1)((x+k-1)-x)!}{x!((x+k-1)-x)!} \\ &= \binom{k+x-1}{x}, \end{aligned}$$

and therefore

$$P(X(s)=x)=\binom{k+x-1}{x}\left(\frac{s}{s+\alpha}\right)^x\left(\frac{\alpha}{s+\alpha}\right)^k,\ x=0,1,2,3,\ldots \qquad (7)$$

Expression (7) is the probability mass function of the Negative Binomial distribution. The parameters of the Negative binomial distribution are the over-dispersion parameter k and the fraction

$$p = \frac{\alpha}{s+\alpha} = 1 - q,$$

where $q = s/(s+\alpha)$. Formally we write

$$X(s) \sim \text{NegBin}\left(k, p = \frac{\alpha}{s+\alpha}\right), \text{ and } 0 < p < 1.$$

The properties of the Negative Binomial distribution of $X(s)$ are determined by the values of α and k, which are the two parameters of the gamma distribution used to model heterogeneity in the values of λ. To see why, consider first the expressions for the mean and the variance of the gamma distribution Λ, which are:

$$\text{E}[\Lambda] = \frac{k}{\alpha} \quad \text{and} \quad \text{Var}[\Lambda] = \frac{k}{\alpha^2}. \tag{8}$$

Now, let's take a look at the mean and variance of the Negative Binomial distribution:

$$\text{E}[X(s)] = \frac{kq}{p} = \left(\frac{k}{\alpha}\right) s = \text{E}[\Lambda]s, \quad \text{and} \tag{9}$$

$$\text{Var}[X(s)] = \frac{kq}{p^2} = \frac{ks(s+\alpha)}{\alpha^2} = \left(\frac{k}{\alpha^2}\right) s^2 + \left(\frac{k}{\alpha}\right) s = \text{Var}[\Lambda]s^2 + \text{E}[\Lambda]s. \tag{10}$$

From equations (8), (9) and (10) we may conclude the following:

1. In the Negative Bionomial distribution, the variance is greater than the mean. The expression for the variance of the Negative Binomial distribution is composed of two terms. The first one, $\frac{ks^2}{\alpha^2}$, is positive. The second one is identical to the expression for the mean eq. (9). Thus, the variance of the Negative Binomial distribution is equal to the mean plus an extra, positive quantity. If this term is null, then the expression for the variance eq. (10) is identical to the expression for the mean eq. (9). When this term is not null and the variance is greater than the mean, we say that the counts modeled are *over-dispersed* (In general, when dealing with random variables in the positive integers, in any instance were the variance of the process of interest is bigger than the mean it is said that the process is overdispersed).

2. The amount of over-dispersion is given by the size of the term $\frac{ks^2}{\alpha^2}$ which is in fact s^2 times the variance of the gamma distribution used to model heterogeneity in the values of λ, $\frac{k}{\alpha^2}$. As the variance of this gamma distribution converges to 0, the amount of heterogeneity in the values of λ decreases, which results in a decrease in the over-dispersion of the counts. In the limit, a variance of 0 in the gamma distribution is equivalent to assume that there is no heterogeneity and thus, that the Poisson model with a fixed value of λ can be used to model the counts $X(s)$.

3. As k and α grow large so that k/α converges to a finite constant, k/α^2 converges to 0 and the Negative Binomial distribution converges to a Poisson distribution. In fact, it can be easily shown that as $k \to \infty$, $\alpha \to \infty$, $p = \frac{\alpha}{s+\alpha} \to 1$ (hence $q \to 0$) and $kq \to \beta$ (where β is a positive quantity), then the probability mass function
$$f(x) = \binom{k+x-1}{x} p^k q^x$$
of the Negative Binomial distribution converges to the probability mass function of the Poisson distribution
$$f(x) = \frac{e^{-\beta}\beta^x}{x!}.$$

In her book, Pielou presented yet another way in which over-dispersion can arise in a spatial model of animal/plant abundance. Pielou asked the following question: what if the spatial units in the study area could be split into habitable and non-habitable units (habitable/non-habitable with respect of the species of interest)? In the habitable units, the counts $X(s)$ observed in a quadrat of size s could be modeled with our (homogeneous) Poisson model with mean λs. Let θ be the proportion of non-habitable units in the study area. So, if we place a quadrat at random in this study area and count the number of individuals observed in the quadrat, there are two ways in which these counts can be 0: first, a sampled quadrat may simply fall in an un-habitable unit and hence, no individual is going to appear in the sample. This event happens with probability θ. Second, a quadrat may fall into a habitable unit but just by chance, we do not observe any individual in the sample. Thus, if X denotes the random variable counting the number of individuals in a quadrat of size s placed at random in such study area,

$$\begin{aligned} P(X=0) &= P(\text{quadrat falls in un-habitable unit}) \\ &\quad + P(\text{quadrat falls in habitable unit and Poisson count} = 0) \\ &= \theta + (1-\theta)P(X(s)=0) \\ &= \theta + (1-\theta)e^{-\lambda s}, \end{aligned}$$

and for all $x > 0$ we have that

$$\begin{aligned} P(X=x) &= P(\text{quadrat falls in habitable unit and Poisson count} = x) \\ &= (1-\theta)\tfrac{e^{-\lambda s}(\lambda s)^x}{x!} \end{aligned}$$

Such model is known in the statistics literature as a "Poisson model with added zeros", or the "Zero Inflated Poisson (ZIP)" distribution. Later on, we will formally confront the Poisson, Negative-Binomial and Poisson model with added zeros to a data set were we will model horseshoe crab abundances.

1.1.6 Probability generating functions: a brief overview

A probability generating function is yet another very special expected value. As its name implies it, this expectation *generates* the probabilities for any discrete probability distribution. It is also a great instrument used to compute quite easily and

without painstaking algebraic manipulations the mean and variance of any discrete probability distribution. In Ecology and Evolution, we often model a natural process, such as population growth, or the evolution of allele frequencies in a population, as a stochastic process. Stochastic processes are easily amenable to model the departure of the population trend from an expectation elicited from biological principles. It is in these situations that the probability generating function can guide our calculations. The coolest (by far) use of probability generating functions, or pgf's as we will call them from now on, consists of deriving probability distributions from other probability distributions when it is not obvious at first which distribution should be used. Starting with the derivation of its definition, we will illustrate some of the basic properties of the pgf, followed by a series of examples.

We start with a population of an herbivorous insect that lay eggs in the leafs of the plant its larvae feeds on. Furthermore, assume that such insect possess a characteristic quantitative trait of interest, which happens to be the focus of our study. Suppose that such trait is encoded by two alleles A and a. The genotypic proportion of the genotypes AA and Aa is z, while that of the genotype aa is $1 - z$. That is:

Genotype:	$\underbrace{AA \quad Aa}$	$\underbrace{aa}$
Phenotype:	z	$1 - z$

In the field, we count the number of insects per leaf.

Figure 4: Insect eggs on a leaf (that would be my garden's lettuces, by the way).

Now, assume that the number of insects on a leaf, X, can be modeled with a Poisson random variable. Suppose, then, that $X \sim Poisson(\lambda)$. Now we ask ourselves: what is the probability of finding 0 individuals with the genotype aa on one leaf? Using the law of total probability, such probability is simply the sum of the probabilities of all the ways that event can happen. Because 0 insects can be found in any one leaf if by chance, the Poisson count is 0, *or* if that count is exactly 1 but that insect is not of genotype aa, *or* if that count is 2 but both insects are not of type

aa, etc ..., if follows that

P(No insects of type aa are found on a leaf)

$= P$(No insects on the leaf or 1 insect of type z or 2 insects of type $z \ldots$)

$= P(X=0) + P(X=1)z + P(X=2)z^2 + P(X=3)z^3 + \ldots$

$= f(0) + f(1)z + f(2)z^2 + \ldots = \sum_{x=0}^{\infty} f(x)z^x$

$= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} z^x$

$= \frac{e^{-\lambda}}{e^{-\lambda z}} \sum_{x=0}^{\infty} \frac{e^{-\lambda z}(\lambda z)^x}{x!}$

$= e^{-\lambda(1-z)}$.

Note that the resulting function $e^{-\lambda(1-z)}$ is in fact, another special expected value, it's $E[z^X]$, where z is a number between 0 and 1. This is the probability generating function of the random variable X. In general, if X is any discrete random variable taking on values in the set S, then the probability generating function, denoted as $\phi(z)$ is:

$$E\left[z^X\right] = \phi(z) = \sum_{x \in S} f(x)z^x, \quad (0 \leq z \leq 1).$$

In particular, if X is a random variable defined in the non-negative integers, then,

$$\begin{aligned} \phi(z) = E\left[z^X\right] &= z^0 f(0) + z^1 f(1) + z^2 f(2) + z^3 f(3) + \ldots \\ &= \sum_{x=0}^{\infty} f(x)z^x. \end{aligned}$$

We now list some of its properties, which result from evaluating the pgf at 0 or 1, or taking its derivative and evaluating it at 0 or 1.

- **Property 1:**

$$\begin{aligned} \phi(0) &= f(0) \\ \phi(1) &= f(0) + f(1) + f(2) + \ldots = 1 \end{aligned}$$

- **Property 2:**

$$\begin{aligned} \phi'(z) &= f(1) + 2zf(2) + 3z^2 f(3) + 4z^3 f(4) + \ldots \\ &= \sum_{x=1}^{\infty} x z^{x-1} f(x) \\ \phi'(0) &= f(1). \\ \phi'(1) &= 0 + 1f(1) + 2f(2) + 3f(3) + \ldots = E[X] = \mu'_{(1)}. \end{aligned}$$

Thus, the first derivative of the pgf evaluated at 0 is equal to $P(X = 1) = f(1)$ and the first derivative evaluated at 1 is equal to the expected value of the random variable X. The expected value is also called the "first factorial moment", or "first moment", for simplicity.

- **Property 3:** When we evaluate this second derivative at 0, we find $2P(X = 2)$. Also, the "second factorial moment", defined as $E[X(X-1)]$ is found by evaluating the second derivative at 1. Together with the first factorial moment, the second factorial moment help us find very easily the variance of X. Thus,

$$\begin{aligned}
\phi''(z) &= 2f(2) + 3 \times 2zf(3) + 4 \times 3z^2 f(4) + \ldots \\
&= \textstyle\sum_{x=2}^{\infty} x(x-1)z^{x-2} f(x), \text{ and it follows that} \\
\phi''(0) &= 2f(2), \text{ and} \\
\phi''(1) &= 0(-1)f(0) + 1(0)f(1) + 2(1)(1)f(2) + 3(2)f(3) + 4(3)f(4) + \ldots \\
&= \textstyle\sum_{x=0}^{\infty} x(x-1)f(x) = E[X(X-1)] = \mu'_{(2)}.
\end{aligned}$$

Since $\phi'(1) = E[X] = \mu'_{(1)}$ and $\phi''(1) = E[X(X-1)] = \mu'_{(2)}$,

$$\begin{aligned}
\text{Var}[X] &= E[X^2] - E[X] + E[X] - \{E[X]\}^2 \\
&= E[X(X-1)] + E[X] - \{E[X]\}^2 \\
&= \phi''(1) + \phi'(1) - [\phi'(1)]^2.
\end{aligned}$$

In general the r^{th} moment of X can be easily found using

$$\phi^{(r)}(1) = \sum_{x=0}^{\infty} x(x-1)\ldots(x-r+1)f(x) = \mu'_{(r)}.$$

- **Property 4:** The pgf "generates" the probabilities for any discrete random variable X: Doing a Taylor Series expansion of $\phi(z)$ around 0 we get:

$$\phi(z) = \phi(0) + z\phi'(0) + \tfrac{z^2}{2!}\phi''(0) + \tfrac{z^3}{3!}\phi'''(0) + \ldots$$

Because by definition, $\phi(z)$ is also equal to

$$\phi(z) = z^0 f(0) + z^1 f(1) + z^2 f(2) + z^3 f(3) + \ldots,$$

it follows that we can equate each one of the terms in both sums, *i.e.*:

$$\begin{aligned}
\phi(0) &= f(0) \\
\phi'(0) &= f(1) \\
\tfrac{\phi''(0)}{2!} &= f(2) \\
&\vdots \\
\tfrac{\phi^{(x)}(0)}{x!} &= f(x),
\end{aligned}$$

and it also follows that we can write $\phi(z) = f(0) + zf(1) + z^2 f(2) + z^3 f(3) + \ldots$ and $f(x) = P(X = x) = \frac{\phi^{(x)}(0)}{x!}$. So yes indeed, $\phi(z)$ *generates* the probabilities $f(x) = P(X = x)$!!

- **Property 5:** The moment generating function of the random variable X is given by $m(t) = E[e^{tX}]$, and is defined not only for discrete distributions but also for continuous distributions. For any random variable X, if it exists, this special expected value is related to the probability generating function $\phi(z)$ in the following way:

$$\begin{aligned} \phi(z) &= E(z^X) \\ &= E\left[\left(e^{\ln z}\right)^X\right] \\ &= E\left[\left(e^t\right)^X\right] \\ &= m(\ln z). \end{aligned}$$

  Hence, the pgf of a random variable X is equal to its moment generating function, evaluated at $\ln z$. Likewise,

$$m(t) = E\left[\left(e^t\right)^X\right] = E[z^X] = \phi(e^t).$$

  So the moment generating function is equal to the pgf evaluated at e^t. Finally, as its name implies, the moment generating function (mgf) is very useful to derive the moments in a very similar way (see Rice, 1995): the moments are given by evaluating the derivatives of the mgf at 0 (not 1, as in the pgf. This fact becomes evident when you think of the transformation between the mgf and the pgf stated above -indeed-, $e^{(0)} = 1$!!.

Below are some examples using the five properties above.

Example 1.1. The pgf of a Poisson random variable: Recall that we found that the pgf for the Poisson distribution with parameter λ is $\phi(z) = e^{-\lambda(1-z)}$. Then, according to the calculations above, we can easily find the mean and the variance of the Poisson distribution using the first two factorial moments. Accordingly, we set

$$\begin{aligned} \phi(z) &= e^{-\lambda(1-z)}. \quad \text{The first derivative of the pgf is then} \\ \phi'(z) &= \lambda e^{-\lambda(1-z)}, \quad \text{and the second derivative is} \\ \phi''(z) &= \lambda^2 e^{-\lambda(1-z)}. \quad \text{Iterating the derivatives we get that the } x^{\text{th}}\text{derivative is} \\ &\vdots \\ \phi^{(x)} &= \lambda^x e^{-\lambda(1-z)} \end{aligned}$$

and

$$f(x) = \frac{\phi^{(x)}(0)}{x!} = \frac{\lambda^x e^{-\lambda(1-z)}}{x!}\Big|_{z=0} = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Thus, we recovered the Poisson probabilities form its pgf. Also, from these calculations it follows that both, the mean and the variance of the Poisson random variable can be obtained in a single line each:

$$\begin{aligned} E[X] &= \phi'(1) = \lambda e^{-\lambda(0)} = \lambda \\ \text{Var}[X] &= \phi''(1) + \phi'(1) - (\phi'(1))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. \end{aligned}$$

Example 1.2. The sum of two independent Poisson r.v.'s is also Poisson distributed:
Suppose that X_1 and X_2 are two independent Poisson random variables such that $X_1 \sim Poisson(\lambda_1)$ and $X_2 \sim Poisson(\lambda_2)$. What is the distribution of $Y = X_1 + X_2$? By definition, the pgf of Y, $\phi(z)$ is given by:

$$\phi(z) = E[z^Y] = E[z^{X_1+X_2}] = E[z^{X_1}]E[z^{X_2}].$$

The last equality results from the independence between X_1 and X_2. Also, note that, from the calculations above, $E[z^{X_i}] = e^{-\lambda_i(1-z)}$. Hence,

$$\phi(z) = e^{-\lambda_1(1-z)}e^{-\lambda_2(1-z)} = e^{-(\lambda_1+\lambda_2)(1-z)},$$

which corresponds exactly to the pgf of a Poisson random variable with parameter $\lambda_1 + \lambda_2$. Therefore,

$$Y \sim \text{Poisson}(\lambda_1 + \lambda_2).$$

In population dynamics this trick is used to compute the probability distribution of the total number of offspring produced in a single generation: suppose that in a population of birds that reproduces once a year, we model the number of offspring produced by a single female with a Poisson random variable with parameter λ. If there are n females present, then the total number of offspring produced can be modeled with another Poisson random variable whose mean is

$$\underbrace{\lambda + \lambda + \ldots + \lambda}_{n \text{ times}} = n\lambda.$$

This result assumes that, independently from each other, all females have the same offspring distribution.

Example 1.3. Suppose that $X \sim Binomial(n, p)$. That is, $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$, $x = 0, 1, 2, \ldots, n$. Then, using the binomial theorem, we find that the pgf of X is given by

$$\phi(z) = E[z^X] = \sum_{x=0}^{n} z^x \binom{n}{x} p^x (1-p)^{n-x} = (zp + (1-p))^n.$$

To find the mean and variance of the Binomial distribution, it is enough to take the first and second derivatives of this pgf and evaluate it at 1. That is

$$\phi'(z) = n(q + zp)^{n-1} p.$$

Hence, $E[X] = \phi'(1) = n(q+p)^{n-1}p = np$.
The second derivative of the pgf is $\phi''(z) = n(n-1)(q+zp)^{n-2}p^2$, and therefore

$$E[X(X-1)] = \phi''(1) = n(n-1)p^2, \text{from which we get that}$$

$$\begin{aligned} Var(X) &= \phi''(1) + \phi(1) - [\phi'(1)]^2 \\ &= n(n-1)p^2 + np - [n^2p^2] \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np(1-p). \end{aligned}$$

Example 1.4. The negative binomial distribution as a sum of Geometric distributions:
In probability, the Negative Binomial distribution is commonly employed to count the (random) number of failures before the k^{th} success. Here we'll see that this distribution arises as a sum of k geometric distributions. It is important to mention that the geometric distribution, in turn, can be written in two different ways: as the number of independent Bernoulli trials needed to get one success, or as the number of failures before the first success. Let p be the probability of success of such Bernoulli trials and $q = 1-p$. Let X denote the first parameterization of the geometric distribution. That is, $X =$ the number of Bernoulli trials needed to get 1 success. X can take on the values $\{1, 2, 3, \ldots\}$ and

$$\begin{aligned} P(X=1) &= p, \\ P(X=2) &= (1-p)p, \\ P(X=3) &= (1-p)^2p, \\ &\vdots \\ P(X=x) &= (1-p)^{x-1}p. \end{aligned}$$

We write $X \sim Geo_1(p)$. Now, let $Y = X - 1$ be the number of failures before the first success. Then, the random variable Y takes on values in $\{0, 1, 2, 3, \ldots\}$ and

$$\begin{aligned} P(Y=0) &= p, \\ P(Y=1) &= (1-p)p, \\ P(Y=2) &= (1-p)^2p, \\ &\vdots \\ P(Y=y) &= (1-p)^yp. \end{aligned}$$

We write $Y \sim Geo_2(p)$. Suppose we are running a series of Bernoulli trials with probability of success p and count the number of failures before the first success. This random variable, call it Y_1 is distributed $Geo_2(p)$. However, we don't stop the experiment there and, as soon as we get the first success, we start the counting again and record the number of failures until the second success. This second count, denoted Y_2, is also $Geo_2(p)$. Continuing with the experiment, we get that the k^{th} count, Y_k, is also distributed $Geo_2(p)$. Therefore, the random variable counting the total number of failures before the k^{th} success, call it W, is found by summing all the Y_i counts. That is,

$$W = Y_1 + Y_2 + \ldots + Y_k.$$

Knowing the pgf of the Y_i and this result, we can easily find the pgf of W, and see if it matches the form of the pgf of a known random variable. To do that, first note that since

$$P(Y_i = y_i) = (1-p)^{y_i}p,$$

it follows that the Y_i's pgf $\alpha(z)$ is given by

$$\alpha(z) = p\sum_{x=0}^{\infty}(zq)^x = p\frac{1}{1-qz} = \frac{1-q}{1-qz}.$$

The above equality results from noting that, if $q + p = 1$,

$$\begin{array}{rcl} \sum_{x=0}^{\infty} q^x p & = & p(1+q+q^2+q^3+\ldots) = 1 \\ & \Leftrightarrow & 1+q+q^2+q^3+\ldots = \frac{1}{p} = \frac{1}{1-q}. \end{array}$$

Next, since W is the sum of the Y_i's, we can write it's pgf $\phi(z) = E[z^W]$ as

$\phi(z) = E[z^{Y_1+Y_2+\ldots+Y_k}]$, which from the independence assumption becomes

$$E[z^{Y_1}]E[z^{Y_2}]\ldots E[z^{Y_k}] = \left(\frac{1-q}{1-qz}\right)\left(\frac{1-q}{1-qz}\right)\ldots\left(\frac{1-q}{1-qz}\right) = \left(\frac{1-q}{1-qz}\right)^k,$$

which matches, as it turns out, the pgf of a Negative Binomial distribution with parameters k and p. That is,

$$P(W = w) = \binom{k+w-1}{w}p^k q^w, w = 0, 1, 2, 3, \ldots$$

Example 1.5. Clustering: The pgf and compound distributions
Let's return for a moment to our example where we count the number of insects on a leaf. Look at figure 4, which depicts a cluster of eggs on a leaf. Had we taken a picture of the entire leaf, we would have seen many such clusters. Now, within each cluster of eggs, it is possible that not all of them hatch, just a few. It is even possible that out of each bunch of eggs, only one successfully hatches. Suppose that we are interested in counting the total number of larvae per leaf, T. Let X_i be the random variable counting the number of larvae hatching from egg mass i. Let N denote the random variable counting the total number of egg masses per leaf. Assume that $X_1, X_2, \ldots, X_N$ are independent and identical. Then, T can be expressed as a **randomly stopped sum**:

$$T = X_1 + X_2 + \ldots + X_N$$

This sum is *randomly stopped* because the total number of egg masses found on a leaf is itself a random variable. Let the pmf and pgf of X_i be given by

$$P(X_i = x) = f(x), \quad \text{and}\, \alpha(z) = E[z^{X_i}] = \sum_{x=0}^{\infty} z^x f(x), \text{ respectively.}$$

Likewise, let the pmf and pgf of N be given by

$$P(N = n) = g(n), \quad \text{and}\, \phi(z) = E[z^N] = \sum_{n=0}^{\infty} z^n g(n), \text{ respectively.}$$

Suppose now that we specify the distribution of X_i and the distribution of N (*i.e.* suppose that we know the explicit forms of $f(x), g(x), \alpha(z)$ and $\phi(z)$). Finally, let $h(t) = P(T = t)$ and $\psi(z) = E[z^T] = \sum_{t=0}^{\infty} z^t h(t)$ denote the pmf and the pgf of T. Using the same population genetics interpretation of the pgf, we have that the P(one egg mass does not have any bugs with genotype aa) is given by

$$\alpha(z) = f(0)z^0 + f(1)z^1 + f(2)z^2 + \ldots, \text{ just as before.}$$

Furthermore, P(a leaf has no aa bugs) $= \psi(z)$, which by definition is written as

$$\psi(z) \;\; = \;\; h(0)z^0 + h(1)z^1 + \ldots$$

Note however that $\psi(z)$ can also be written as

$$\psi(z) \;\; = \;\; g(0) + g(1)\alpha(z) + g(2)[\alpha(z)]^2 + g(3)[\alpha(z)]^3 + \ldots,$$

because $g(0) =$ the probability of finding no egg masses in the leaf, $g(1)\alpha(z)$ is the probability of finding 1 egg mass on the leaf and that such egg mass doesn't contain a single aa bugs, $g(2)[\alpha(z)]^2$ is in fact the probability of finding 2 egg masses on the leaf, but none of these contain aa bugs, and so on. This sum can be written compactly as

$$\psi(z) \;\; = \;\; \textstyle\sum_{n=0}^{\infty} g(n)[\alpha(z)]^n,$$

which it is immediately recognizable as the pgf of N evaluated at $\alpha(z)$. Therefore, we get that

$$\psi(z) \;\; = \;\; \textstyle\sum_{n=0}^{\infty} g(n)[\alpha(z)]^n \;\; = E[(\alpha(z))^N] = \phi(\alpha(z)).$$

If, for instance, each egg mass produces at most one larvae, then the number of larva per egg mass can be modeled with a Bernoulli distribution. That is,

$$P(X_i = x) = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } q = 1 - p. \end{cases}$$

In that case $\alpha(z) = z^0 f(0) + z^1 f(1) = q + zp$. If, we also assume that $N \sim \text{Poisson}(\lambda)$ so that $\phi(z) = \exp^{-\lambda(1-z)}$, then it immediately follows that

$$\begin{aligned} \psi(z) = E[z^T] = \phi(\alpha(z)) \;\; &= \;\; e^{-\lambda(1-(q+zp))} \\ &= \;\; e^{-\lambda+q\lambda-\lambda zp} \\ &= \;\; e^{-\lambda p(1-z)}, \text{ i.e. } T \sim \text{Poisson}(\lambda p) \end{aligned}$$

So in an instant, because the pgf uniquely determines the pmf, we can identify the distribution of T. The random variable T is an example of a "contagion" or "compound" distribution.

Assume now that, as opposed to the previous example where at most one egg hatches, *at least* one egg hatches. Thus, X_i is distributed according to a random variable that takes on integers, excluding the 0. One such distribution can be obtained

from the Taylor Series expansion around 0 of $-\ln(1-q)$, where $q \in (0,1)$. Such expansion is given by

$$-\ln(1-q) = q + \frac{q^2}{2} + \frac{q^3}{3} + \dots$$

Dividing both sides of the equation by $-\ln(1-q)$ we get that

$$1 = \frac{q^1}{-1\ln(1-q)} + \frac{q^2}{-2\ln(1-q)} + \frac{q^3}{-3\ln(1-q)} + \dots$$

Since every term on the right hand side corresponds to an integer, starting at 1, and that the sum of the terms is equal to 1, we can define a discrete probability distribution X_i having as pmf

$$P(X_i = x) = \frac{q^x}{-x\ln(1-q)}, \quad x = 1, 2, 3, \dots$$

This distribution is known as the log-series distribution. The pgf of such distribution is given by

$$\alpha(z) = \frac{\ln(1-qz)}{\ln(1-q)}.$$

Now that we have a probability distribution for the case where at least one egg hatches from the egg mass, let's assume again that $N \sim \text{Poisson}(\lambda)$ so that $\phi(z) = \exp^{-\lambda(1-z)}$. Then, the pgf of T is found to be

$$\begin{aligned}
\psi(z) &= \phi(\alpha(z)) = e^{-\lambda(1-\alpha(z))} \\
&= e^{-\lambda\left(1-\frac{\ln(1-qz)}{\ln(1-q)}\right)} \\
&= \left(\frac{1-q}{1-qz}\right)^{-\lambda/\ln(1-q)} \\
&= \left(\frac{1-q}{1-qz}\right)^{k}, \text{ where } k = -\lambda/\ln(1-q) > 0
\end{aligned}$$

The resulting expression is immediately recognizable as the pgf of a Negative Binomial random variable (see example above with the geometric distribution). Therefore,

$$T \sim \text{NegBin}(k, p = 1-q),$$

and we found yet another way of deriving the Negative Binomial probability distribution. In Ecology, contagion distributions as sampling models of counts are useful because many spatial phenomena match the general setting of the eggs within an egg mass, and egg masses within leafs.

1.1.7 The multinomial distribution

The multinomial distribution is a generalization of the Binomial distribution for categorical variables with more than two response types. Suppose that our statistical population of interest is composed of $k = 5$ different types. Accordingly, let π_1 be the proportion of individuals of type 1 in the population, π_2 the proportion of individuals of type 2 in the population, and so on. Then, necessarily, $\pi_1 + \pi_2 + \ldots + \pi_k = 1$. Suppose further that we take a random sample of size n from the population. Such sample would be composed of a random number of individuals of each type. Let

$$Y_1 = \text{the random number of individuals of type 1 in the sample,}$$

$$Y_2 = \text{the random number of individuals of type 2 in the sample,}$$

$$Y_3 = \text{the random number of individuals of type 3 in the sample,}$$

$$Y_4 = \text{the random number of individuals of type 4 in the sample,}$$

$$Y_5 = \text{the random number of individuals of type 5 in the sample.}$$

Because necessarily $Y_1 + Y_2 + Y_3 + Y_4 + Y_5 = n$ for any n, the random variables Y_i are dependent. The joint probability distribution of the $Y_i's$ is given by

$$P(Y_1 = y_1, Y_2 = Y_2, \ldots, Y_k = y_k) = \frac{n!}{y_1!y_2!\ldots y_k!}\pi_1^{y_1}\pi_2^{y_2}\ldots\pi_k^{y_k},$$

where the y_i's are non-negative integers that add up to n. This is the pmf of the multinomial distribution. If we let $k = 2$ so that only two possible types exist, then $Y_2 = n - Y_1$, $\pi_2 = 1 - \pi_1$, and

$$P(Y_1 = y_1, Y_2 = y_2) = \frac{n!}{y_1!(n - y_1)!}\pi_1^{y_1}\pi_2^{n-y_1}.$$

Thus, the binomial distribution is indeed a special case of the Multinomial distribution.

- **Reduced parameter multinomial models**
- **Generalized likelihood ratio tests**
- **Samuel Wilks' Generalized likelihood ratio tests for reduced parameter multinomial models**
- **Confidence intervals and likelihood ratio tests**
- **Abraham Wald's theorem**

1.1.8 Hypothesis tests: a review of basic concepts

Some of the very basic concepts and ideas about hypothesis tests can be reviewed by means of simple examples, without dwelling into likelihood theory. This is the purpose of this section. In the next section, we will present the theoretical details of likelihood inference along with detailed examples of biological relevance. These lectures will be the founding blocks of the rest of our course.

Fisher's tea lady:

In R.A. Fisher's experimental designs book there is a ten pages account of an experiment where he basically laid out the most important principles of experimentation. The experiment is known as "Fisher's tea lady experiment". This experiment was also later described by Fisher's daughter, who wrote his biography. A pdf file of this textbook fragment is posted in the course web page. This account tells the story of a lady that claimed to be able to distinguish between a tea cup which was prepared by pouring the tea first and then the milk and another tea cup where the milk was poured first. Fisher then wonders if there is there a good experiment that could be devised in order to formally test the lady's claim. The null hypothesis of this purported experiment would then be that the lady has no selection ability whatsoever. A logical experiment would consist of offering the lady a set of "tea-first" cups and another set of "milk-first" cups and let her guess the tea cup type (milk-first or tea-first) of each one. The question is -Fisher noted- that it is not evident how many of each type and in what order shall this be done in order to carry a convincing experiment. Fisher begins by noting that, the more cups are offered to the lady, the harder it is to achieve a perfect classification of all the tea cups. Also, note that by giving her the same number of tea-first cups than milk-first cups we would allow each of the 2 types to get the same simultaneous presentation (*i.e.* opportunity to be chosen). Suppose that we ask the lady to select 4 milk-first from a total of 8 cups (That is, we offer her 4 milk-first and 4 tea-first cups). In how many ways can she make the 4 choices? Fisher noted that for the first cup there are 8 choices, for the second there are 7 choices, 6 choices for the third and finally, 5 choices for the fourth milk-first cup. Therefore, this succession of choices can be made in $8\times7\times6\times5 = 1680$ number of ways. But this takes into account not only every possible set of 4, but also every possible set in every possible order. Now, 4 objects can be arranged in order in $4\times3\times2\times1 = 24$ ways and therefore, since the 4 cups are assumed to be identical in every respect and we do not care about the order in which these 4 cups were given, then the number of ways of picking 4 cups out of 8 is

$$\begin{aligned}\frac{\#\text{ number of ways of assigning 4 cups as milk-first among the 8 cups}}{\#\text{ of ways that 4 cups can be ordered}} &= \frac{8\times7\times6\times5}{4\times3\times2\times1}\\ &= \frac{8!}{4!(8-4)!},\end{aligned}$$

which is $\binom{8}{4} = 70$. So if the lady was picking purely at random and didn't have any distinguishing ability whatsoever, she would have a probability of 1/70 of picking

up a particular sequence of cups assigned by her as milk-first that happens to be the correct one. What if we set 3 milk-first cups and 3 tea-first cups? Then, since $\binom{6}{3} = 20$ the lady would have a 1/20 probability of picking the correct sequence just by chance. Fisher decided to go for the harder test and decided to give her 4 milk-first and 4 tea-first cups. Now that we have decided on the number of cups, we can compute the probability of each possible outcome if the lady was picking purely at random (that is, if she had no ability to distinguish between a tea-first and a milk-first cup). The possible outcomes of the experiment are the following: the lady could pick 4 right out of the 4 of one type and therefore get 0 wrong out of the other type. We will denote this event $4R/0W$. She could also get three right of the first type and one wrong of the second type. This event will be denoted by $3R/1W$. According to this notation scheme, the other possible events are $2R/2W$, $1R/3W$ and $0R/4W$. Computing the probabilities of each of these events is a straightforward counting exercise. Considering the event $3R/1W$ for instance, we note that there are $\binom{4}{3}$ number of ways of picking 3 right out of 4 of the first type and independently of that, there are $\binom{4}{1}$ ways of choosing 1 wrong out of the other 4 cups of the second type. Iterating this argument for the other events we get that,

$$P(3R/1W) = \frac{\binom{4}{3} \times \binom{4}{1}}{\binom{8}{4}} = \frac{16}{70}.$$

Likewise,

$$P(4R/0W) = \frac{\binom{4}{4} \times \binom{4}{0}}{\binom{8}{4}} = \frac{1}{70},$$

$$P(2R/2W) = \frac{\binom{4}{2} \times \binom{4}{2}}{\binom{8}{4}} = \frac{36}{70},$$

$$P(1R/3W) = \frac{\binom{4}{1} \times \binom{4}{3}}{\binom{8}{4}} = \frac{16}{70},$$

and

$$P(0R/4W) = \frac{\binom{4}{0} \times \binom{4}{4}}{\binom{8}{4}} = \frac{1}{70}.$$

These probabilities completely specify the probability mass function of the outcomes of the experiment where the picking was done purely at random, that is, assuming that the lady has no detection ability. Therefore, this is the distribution of outcomes under the null hypothesis. Suppose that the experiment is carried and the lady picks 3 right of the first type and 1 wrong of the second type ($3R/1W$). Is this evidence enough to convince ourselves that she is not picking the cups at random and that she indeed has a detection ability? So we ask ourselves, if the null hypothesis is correct and the lady is picking purely at random, how unlikely it is to get an outcome as extreme or more more than the one we actually observed. This amounts to specify the probability of making only one error or less by pure dumb luck. According to the calculations above, that probability is

$$P(3R/1W) + P(4R/0W) = \frac{17}{70} \approx 0.24.$$

So if the null hypothesis is true, then there is a chance that we would have observed a choice as good or better than the one we saw about 24% (about a fifth) of the time! That chance is way too big to convince our skeptic (Fisher) that his null hypothesis is wrong. 0.24 is in fact, the p-value of the test of the lady's claim. Compare that value to what we are used to think of what a good skeptic's convincing threshold is: 0.05 (or 5% of the time). Hence, here we blatantly failed to reject the null hypothesis! Fisher's account is important in many ways and the most notable is the description of the value of randomization in experimentation (which I explicitly left out in here) as well as his careful elaboration of the logics of hypothesis testing.

Exercise 1.1. Suppose we ask the lady to select 4 milk-first from a total of 8 cups (that is, we offer her 4 milk-first and 4 tea-first cups). In how many ways can she make the four choices? Fisher noted that for the first cup there are 8 choices, for the second there are 7 choices, 6 choices for the third and finally, 5 choices for the fourth milk-first cup. Therefore, this succession of choices can be made in $8\times7\times6\times5 = 1680$ number of ways. But this takes into account not only every possible set of 4, but also every possible set in every possible order. Now, 4 objects can be arranged in order in $4\times3\times2\times1 = 24$ ways and therefore, since the 4 cups are assumed to be identical in every respect and we do not care about the order in which these 4 cups were given, then the number of ways of picking 4 cups out of 8 is

$$\begin{aligned}\frac{\#\text{ number of ways of assigning 4 cups as milk-first among the 8 cups}}{\#\text{ of ways that 4 cups can be ordered}} &= \frac{8\times7\times6\times5}{4\times3\times2\times1}\\ &= \frac{8!}{4!(8-4)!},\end{aligned}$$

which is $\binom{8}{4} = 70$. So if the lady is picking purely at random, she can assign four cups as "milk-first" in 70 different ways.

1. What is the probability that the lady doesn't make any mistake and correctly chooses the 4 milk-first cups?

2. Had Fisher given her 3 milk-first cups and 3 tea-first cups, what would the probability of correctly picking the 3 milk-first cups had been?

3. In fact, when he was thinking how to design the experiment, Fisher chose the number of cups after computing the probability of making no mistakes in two cases: when she is given to select 4 cups out of a total of 8 and when she is given 3 cups out of a total of 6. Given your answers to the two questions above, which number of cups do you think Fisher picked: 4 and 4 or 3 and 3? Why?

4. Enumerate all the possible outcomes of the experiment when a total of 8 cups are given to her (4 of each type). Hint: for instance, one outcome is as follows: she can pick 4 right out of the 4 of one type and therefore get 0 wrong out of the other type. Denote this event as $4R/0W$, where R stands for 'right' and W for 'wrong'. Use the same notation for all the other events.

5. Compute the relative frequency with which every single one of these possible outcomes occurs.

Hypothesis test for the sample mean (known variance):

Suppose that an education researcher suspects that college students at UF have a higher IQ than the population at large. The average IQ score from the population at large is 100. Because the IQ score can be though as a continuous phenotypic trait, to model its distributional properties this researcher should use a continuous random variable. In particular, here we'll use a Normal distribution, which is symmetric around the mean. Now, suppose that the standard deviation σ of the IQ scores distribution is known and equal to 15. To confront his suspicion with data, the researcher takes a *random sample* of $n = 30$ IQ tests from the population of UF students and obtains a sample mean score $\bar{x}$ equal to 105.3. A colleague of the education researcher is very skeptic of this suspicion and in fact, tells him that an IQ sample mean of 105.3 is not really an unlikely outcome if these 30 samples really came from a population of scores that is normally distributed around a mean $\mu = \mu_0 = 100$. The value $\mu_0 = 100$ embodies the skeptic's point of view, it corresponds to his hypothesized value of the mean of the distribution of IQ scores from which our random sample was drawn. In statistical terms, this is called the *null* hypothesis. Conducting a hypothesis test in this case amounts to convincing the skeptic that the researcher's suspicion (that is, the *alternative* hypothesis) that $\mu > 100$ is indeed supported by the data. In response to his colleague's questioning, the researcher starts by asking himself how unusual a sample mean of 105.3 would be if it really came from the IQ distribution of the population at large. Repeated *independent random sampling* from the population at large of IQ scores generates a series of sample means. Each time a sample of IQ scores is taken, a new sample mean is obtained. Thus, the computed sample mean IQ score can be considered as the outcome of a random variable. Let's denote this random variable $\bar{X}$ (remember that capital letters in this notes denote a random variable, unless otherwise specified). From the Appendix review (and a bit of common sense) we know that if the samples are really random, independent and drawn from a population with mean $\mu_0 = 100$ (the skeptic's hypothesis), the distribution of the sample mean is again Normal, with mean equal to $\mu = \mu_0$ and variance σ^2/n. We write:

$$\bar{X} \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{n}\right).$$

Asking how unusual would a sample mean of 105.3 be if the null hypothesis were to be true then amounts to compute an integral, the area to the right of $\bar{x} = 105.3$ under a normal curve whose mean is $\mu_0 = 100$ and variance is $\frac{\sigma^2}{n} = \frac{15^2}{30}$. This area is in fact a probability. It is the probability that $\bar{X} \geq 105.3$ which is given by

$$P(\bar{X} \geq 105.3) = \int_{105.3}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp -\frac{(\bar{x} - \mu_0)^2}{2\sigma^2/n} d\bar{x}.$$

Fortunately, we can ask `R` to compute that integral for us with the following line:

```
> 1-pnorm(q=105.3, mean=100, sd= 15/sqrt(30))
[1] 0.02647758
```

Alternatively, we could go the old ways and standardize our normal distribution of sample means $\bar{X}$ and get the equivalent quantile value of $\bar{x} = 105.3$ in the standard normal distribution Z. Because $\bar{X}$ can be thought of as the following linear transformation of the standard normal distribution

$$\bar{X} = \frac{\sigma}{\sqrt{n}} Z + \mu_0,$$

solving for Z in this equation allows us to find the standardized value of $\bar{x} = 105.3$. That is, since

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

the standardized value of $\bar{x} = 105.3$ is found to be

$$z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{105.3 - 100}{15/\sqrt{30}} = 1.935280.$$

Then, knowing that $P(\bar{X} \geq 105.3) = P(Z \geq 1.935280)$ we just have to do a table look up to find out this probability, or, if we don't have our "Z-table" at hand, just ask `R` again:

```
> 1-pnorm(q=1.935280, mean=0, sd= 1)
[1] 0.02647797
```

which happily corresponds to the previous value found before (besides some numerical round-off error). What does the 0.02647797 means? It simply means that if the skeptic's hypothesis was true and the 30 sampled IQ scores came from a distribution with mean 100, then the probability of observing a sample mean *as big or bigger than* 105.3 is slightly less than 0.03. In other words, if the skeptic's hypothesis was true and we were to repeat the experiment of drawing a sample of size $n = 30$ student's IQ scores and each time compute the sample mean, less than 3% of the time we would actually observe sample means as high or higher than 105.3. So our researcher now has computed a value, 0.02647797, that makes his colleague's hypothesis untenable. Given the evidence against his hypothesis, the skeptic concedes and admits to be convinced. How small has the value of $P(\bar{X} \geq \bar{x})$ to be in order to convince a skeptic? Well, in a typical statistical analysis, the threshold to reject the skeptic's null hypothesis is set to be less than 5%, or 0.05. Very serious scientific experiments set the convincing threshold to 0.01. In any case however, that threshold is what is known as α and the probability of observing a test statistic as extreme (extreme in the direction of the research hypothesis) or more than the value actually observed is known as the *p-value*. So this skeptic vs. researcher argument is really where the famous quasi-robotic "**Decision rule:** Reject H_0 if p-value $< \alpha$" comes from. Also, note that whenever a decision is made, two possible errors arise: first, the null hypothesis could be true, but it is rejected. Since we reject the null hypothesis whenever we observe a p-value less than α, given that the null hypothesis is true, that probability is just given by α. Second, it may be possible that we fail to reject H_0 even if it is false. The probability of that happening is denoted by

β. $1-\beta$ is therefore the probability of making the correct choice and it is known is statistics as the *power* of the test. In future lectures we will deal with trying to compute the power for our ANOVAS. Finally, note that I've written "Failing to reject H_0". Why? Why not simply "accepting H_0"? Well, it can be argued that accepting the null hypothesis may suggest that it has been proved simply because it has not been disproved yet. This is a logical fallacy known as "the argument from ignorance".

The duality between Confidence Intervals and Hypothesis Tests:

Suppose we were conducting a hypothesis test of

$$H_0 : \mu = \mu_0 = 500$$
$$H_a : \mu \neq \mu_0.$$

We go out and take a random sample of size $n = 60$ knowing that $\sigma = 100$. Look at the graph below and locate the rejection region and the acceptance region for this example of a two-sided hypothesis test.

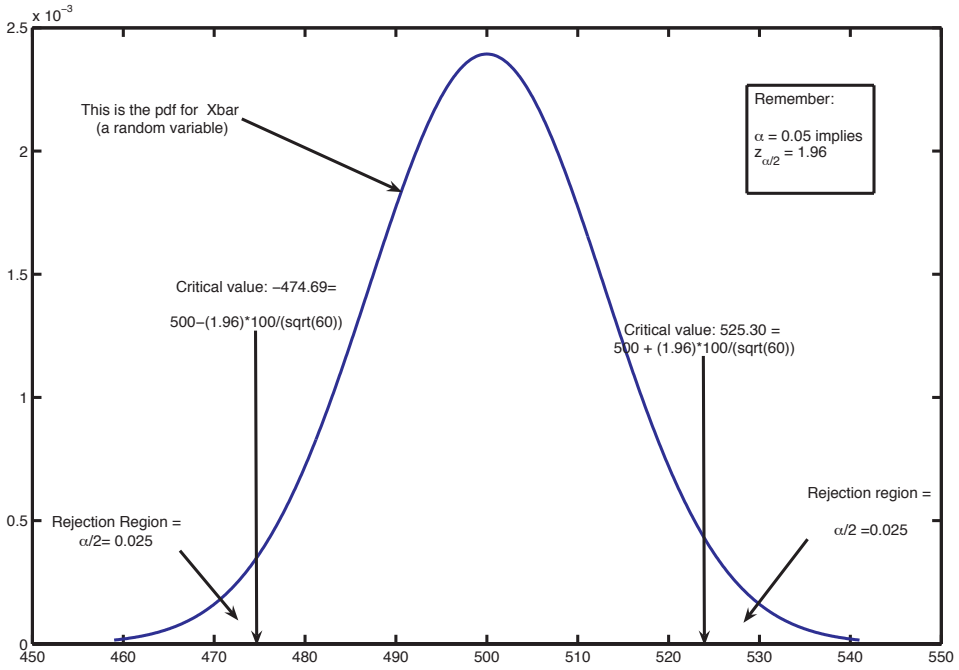

Figure 5: Probability distribution for $\bar{X}$: Depicted are the rejection and the acceptance regions for the two-sided hypothesis test.

What's the size of the acceptance region? That's a probability, it's the area between the two critical quantile values of the distribution of $\bar{X}$. This area is found

to be:

$$\begin{aligned} & P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ = \; & P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad \text{(subtracting} \quad \mu \quad \text{everywhere)} \\ = \; & P\left(-\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad \text{(subtracting} \quad \bar{X} \quad \text{everywhere)} \\ = \; & P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad (\times -1). \end{aligned} \tag{11}$$

So what this is showing is that, in fact, the confidence interval for μ is just the set of all values of μ_0 for which the null hypothesis $\mu = \mu_0$ would not be rejected in a test against the alternative hypothesis $\mu \neq \mu_0$. Note that because $\bar{x}$ is the *realized* value (that is, one fixed quantity) from the probability distribution of $\bar{X}$, it doesn't make sense to ask

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) =?$$

For a particular random sample, the realized confidence interval

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

either contains or does not contains the true mean μ and we actually do not know which of these two outcomes occurred. If we were to repeat the experiment many many times however, and each time after taking a random sample of size n, we computed the sample mean and its realized confidence interval, then $(1-\alpha)\times 100\%$ of the time the realized confidence interval would contain the true mean μ. For each individual confidence interval, the true mean would either be inside or it would not. Thus, repeating this experiment many many times and computing a confidence interval each time can be though of as a horse shoe game where we have our eyes closed. We shoot the horse shoe many many times (*i.e.* we get the random sample, compute its mean and confidence interval) and each time we either make a stake (*i.e.* the true mean is contained in our realized confidence interval) or we miss the stake (the true mean is not contained in our realized confidence interval), but we do not know for sure what happened (we have our eyes closed!). The only thing we know from the probability calculations above is that $(1-\alpha)\times 100\%$ ($= 95\%$ if $\alpha = 0.05$) of the time the true mean value will be contained in the confidence interval. This is a very hard concept to understand and it is not commonly understood properly.

1.2 An introduction to some theoretical properties of Maximum Likelihood estimation and testing

Parameter estimation is really the beginning point (if not a by-product) of a much more rich activity in terms of science: trying to distinguish between competing models -tentative explanations- of natural phenomena. Mathematical statistics theory gives us a very elegant (-sorry, I *had* to use the word 'elegant' here, however I promise I'll never use the word "obvious" while describing how I go from one equation to another-) framework to achieve such undertaking. In this section, I'll summarize some of the most important statistical results useful to carry inference by maximum likelihood.

1.2.1 Fisher's Information

Let $X_1, X_2, \ldots, X_n$ be a sample of size n modeled using a discrete probability distribution. The likelihood function for the *realized* observations is the joint pmf evaluated at the observations $x_1, x_2, \ldots, x_n$:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1, x_2, \ldots, x_n; \theta),$$

where θ is the parameter characterizing the probability mass function of X_i, $i = 1, \ldots, n$. When Fisher defined the likelihood, he noted that when this joint pmf evaluated at the observations is graphed as a function of the parameter of interest, not only can we find the most likely value of the parameter given the data and probabilistic sampling model at hand, but also, that the steepness of the likelihood function around the ML estimate is a surrogate of how much information the data contains about θ. Indeed, the steepness of the likelihood function around the ML estimate is in fact telling us how fast the verisimilitude of one value of the parameter of interest decays with respect to the verisimilitude of the ML estimate, as we move further apart from it. A steep decrease in likelihood $f(\underline{x}; \theta)$ will result in a small set of parameter values around the ML estimate $\hat{\theta}$ being favored as "very likely", while the rest of the parameter space is deemed as "not very likely" (relative the ML estimate). Therefore, highly peaked likelihoods are in fact telling us that given the data and the probabilistic model at hand, there is enough information in the data to narrow down quite precisely our inferences about the parameter of interest. Mathematically, the amount of information in the likelihood function for both, discrete and continuous probability models, is quantified using Fisher's information $\mathcal{I}(\theta)$:

$$\mathcal{I}(\theta) = E_{\underline{X}}\left(\left[\frac{\partial}{\partial\theta}\ln f(\underline{x}; \theta)\right]^2\right). \tag{12}$$

Under adequate smoothness conditions on f (we will discuss this later), Fisher's information can be written as

$$\mathcal{I}(\theta) = -E_{\underline{X}}\left(\left[\frac{\partial^2}{\partial\theta^2}\ln f(\underline{x}; \theta)\right]\right). \tag{13}$$

A short proof of why equations 12 and 13 can be written is shown towards the end of this lecture. However, before dealing with such proof, it is important to

note that the definition of Fisher's information using a second derivative (eq. 13) is readily interpretable: the information in the likelihood function can be measured as the average of the rate of change of the slope of the likelihood function. Since the likelihood is a function of the data, which are realizations of the random sampling scheme, this rate of change of the slope of the likelihood function also changes from one data set to another. Therefore, Fisher's definition is in fact proposing to use this *average curvature* as a measure of how much information about the parameter of interest is conveyed by a random sampling experiment. In particular, for a particular data set, the information conveyed by the likelihood function about the parameter θ can be estimated by plugging-in the definition of $\mathcal{I}(\theta)$ the ML estimate $\hat{\theta}$. Because any function of a ML estimate is also a ML estimate, then the resulting estimate of Fisher's information, $\widehat{\mathcal{I}(\theta)}$ is also a ML estimate. $\widehat{\mathcal{I}(\theta)}$ is needed in order to compute what are known as "Wald's confidence intervals". We now state Abraham Wald's theorem (1948, The Annals of Mathematical Statistics 19(1):40-46):

Theorem 1.1. Under regularity conditions on the likelihood function $f(\underline{x};\theta)$, the random variable $\widehat{\Theta}$ (the ML estimate) converges in distribution to a Normal random variable with mean θ and variance $\mathcal{I}(\theta)^{-1}$. A $100 \times (1-\alpha)\%$ confidence interval for θ is given by

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\mathcal{I}(\hat{\theta})^{-1}},$$

where $\mathcal{I}(\hat{\theta})^{-1}$ is the inverse of Fishers' information (eqs 12 and 13) evaluated at the ML estimate $\hat{\theta}$ of θ and the observed data.

The regularity conditions above roughly say that θ cannot lie on the boundary of the parameter space for the inference to be valid and that the range of the X_i's cannot depend on θ. Also, the appearance of multi-modal likelihoods at low sample sizes compromises the validity of these results (*more on these reg. conditions later*).

Example 1.6. Suppose $X_1, X_2, \ldots, X_n$ are iid random samples from a Normal distribution with known variance σ^2. Then, the likelihood function is written as

$$\ell(\mu) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \quad \text{and the log likelihood is}$$

$$\ln\ell(\mu) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2} - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Taking the derivative of the log-likelihood with respect to μ, setting it equal to 0 and solving for μ gives

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

Having found that the ML estimate of μ is the sample mean, we set out to compute the confidence interval for the sample mean given by Wald's theorem above. Accordingly,

we need to evaluate the expectation of the second derivative of the log-likelihood and evaluate such expression at the ML estimate:

$$\begin{aligned}
\tfrac{d^2}{d\mu}[\ln \ell(\mu)] &= \tfrac{d^2}{d\mu}\left[-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^n (x_i-\mu)^2\right] \\
&= -\tfrac{d}{d\mu}\left[\tfrac{d}{d\mu}\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^n (x_i-\mu)^2\right] \\
&= \tfrac{d}{d\mu}\left[\tfrac{1}{\sigma^2}\textstyle\sum_{i=1}^n (x_i-\mu)\right] \\
&= \tfrac{d}{d\mu}\left[(\textstyle\sum_{i=1}^n x_i)\,\tfrac{1}{\sigma^2} - \tfrac{n\mu}{\sigma^2}\right] \\
&= -\tfrac{n}{\sigma^2}.
\end{aligned}$$

Therefore $\mathcal{I}(\hat{\theta}) = -E_{\underline{X}}\left(-\frac{n}{2}\right) = \frac{n}{\sigma^2}$ and Wald's $100(1-\alpha)\%$ confidence interval for the mean would be given by

$$\bar{x} \pm z_{\alpha/2}\sqrt{\left(\mathcal{I}(\hat{\theta})\right)^{-1}} \Rightarrow \bar{x} \pm z_{\alpha/2}\sqrt{\left(\frac{\sigma^2}{n}\right)},$$

which is exactly the confidence interval shown in the simple hypothesis test of UF student's IQ. This example is very revealing, since we can see right away that the inverse of Fisher's information is none other than the variance characterizing the distribution of the ML estimate!!! Indeed, the random variable $\bar{X}$ is the ML estimate of the mean μ and we've seen before that under this normal sampling scheme, the sample mean $\bar{X}$ is normally distributed with mean μ and variance σ^2/n.

To see why equations 12 and 13 are equivalent, first note that the joint pmf (pdf) of the observations has to integrate to 1 by definition, that is

$$\int f(\underline{x};\theta)d\underline{x} = 1.$$

Furthermore, the expected value of any function $h(\underline{X})$ of the joint random vector of the observations $\underline{X}$ is written as

$$E_f[h(\underline{X})] = \int h(\underline{X})f(\underline{x};\theta)d\underline{x}.$$

Finally, we will define the score function $u(\underline{x};\theta)$

$$u(\underline{x};\theta) = \frac{\partial}{\partial\theta}\ln f(\underline{x};\theta) = \frac{1}{f(\underline{x};\theta)}\frac{\partial}{\partial\theta}f(\underline{x};\theta).$$

With these definitions in hand, we note that

$$0 = \frac{\partial}{\partial\theta}\int f(\underline{x};\theta)d\underline{x} = \int \frac{\partial}{\partial\theta}f(\underline{x};\theta)d\underline{x}. \tag{14}$$

The integration and the differentiation in the right hand side of the above equation have been interchanged (which requires some continuity assumptions about $f(.)$). Using the definition of $u(\underline{x};\theta)$ we write $\frac{\partial}{\partial\theta}f(\underline{x};\theta)$ as

$$\frac{\partial}{\partial\theta}f(\underline{x};\theta) = \left[\frac{\partial}{\partial\theta}\ln(f(\underline{x};\theta))\right] f(\underline{x};\theta) \tag{15}$$

and substitute the right hand side into equation (14) to get

$$0 = \int \frac{\partial}{\partial\theta}f(\underline{x};\theta)d\underline{x} = \int \left[\frac{\partial}{\partial\theta}\ln(f(\underline{x};\theta))\right] f(\underline{x};\theta)d\underline{x}.$$

Derivating a second time we get

$$\begin{aligned} 0 &= \tfrac{\partial}{\partial\theta}\int \left[\tfrac{\partial}{\partial\theta}\ln(f(\underline{x};\theta))\right] f(\underline{x};\theta)d\underline{x} \\ &= \int \tfrac{\partial}{\partial\theta}\left[\tfrac{\partial}{\partial\theta}\ln(f(\underline{x};\theta))\right] f(\underline{x};\theta)d\underline{x} \\ &= \int \left[\tfrac{\partial^2}{\partial\theta^2}\ln(f(\underline{x};\theta))\right] f(\underline{x};\theta)d\underline{x} + \int \tfrac{\partial}{\partial\theta}f(\underline{x};\theta)\tfrac{\partial}{\partial\theta}\ln(f(\underline{x};\theta))d\underline{x}. \end{aligned}$$

The integrand of the rightmost term in the RHS (Right Hand Side) of the last line can be re-written using eq. 15 as

$$0 = \int \left[\frac{\partial^2}{\partial\theta^2}\ln(f(\underline{x};\theta))\right] f(\underline{x};\theta)d\underline{x} + \int \left[\frac{\partial}{\partial\theta}\ln(f(\underline{x};\theta))\right]^2 f(\underline{x};\theta)d\underline{x},$$

or

$$-\int \left[\frac{\partial^2}{\partial\theta^2}\ln(f(\underline{x};\theta))\right] f(\underline{x};\theta)d\underline{x} = \int \left[\frac{\partial}{\partial\theta}\ln(f(\underline{x};\theta))\right]^2 f(\underline{x};\theta)d\underline{x}. \tag{16}$$

Both sides of equation 16 are integrals of a function of $\underline{x}$ times the joint pdf $f(\underline{x};\theta)$, and thus correspond to the definition of the expectation of those functions with respect to the distribution of the random vector of observations $\underline{X}$ (see definition of the expected value of a function of a random variable in the appendix). It follows that equation can be re-written as

$$-E_{\underline{X}}\left[\frac{\partial^2}{\partial\theta^2}\ln(f(\underline{x};\theta))\right] = E_{\underline{X}}\left[\left[\frac{\partial}{\partial\theta}\ln(f(\underline{x};\theta))\right]^2\right]$$

which is the equality we wanted to prove.

Appendix 1:

a basic probability review

1.3 Basic Probability review: Random experiments and events

The idea of probability, chance or randomness is very old and is deeply rooted in the analysis of gambling games. The models derived from probability theory are used for any situation for which the outcomes occur randomly. A **random experiment** is a process whose outcome is not known in advance (yes, -flipping a coin- is one of them...).The **sample space** associated with an experiment is the set of all possible outcomes in the sample, often denoted Ω. An **event** is a subset of the sample space.

Example 1.7. Flipping a coin once, $\Omega = \{H, T\}$. H is an event.

Example 1.8. To come to class, you drive through a sequence of 3 intersections with traffic lights, each time you either stop (s), or continue (c).

$$\Omega = \{sss, ssc, scc, ccc, ccs, css, scs, csc\}.$$

Let A be the event "stopping at the first light" and B the event "stopping at the the third light".

A **probability** is a number between 0 and 1 associated with a particular event in the sample space of a random experiment. This number measures the chance that the event will occur. If A is an event, we write $P(A)$ for the probability of the event A. There are various operational interpretations of probability. The **classical interpretation** arose from games of chance. If a sample space consist of **equally likely** outcomes, then the probability of an event is defined as the ratio of the number of outcomes favorable to the event to the total number of possible outcomes. These equally likely events are also known as "elementary events".

$$P(A) = \frac{\text{number of outcomes favorable to the event}}{\text{Total number of possible outcomes}}$$

Example 1.9. Conduct the random experiment of flipping a coin 4 times and list all the possible outcomes. There are 16 possible outcomes. All these events are equally likely, so the probability of any of these events is simply 1/16. If the event A is defined as "getting a head in both the first and second flip", then the outcomes favorable to the event A are

$$HHHH,\ HHHT,\ HHTH,\ HHTT.$$

All the other outcomes are not favorable to the event A. Thus, $P(A) = 4/16 = 1/4$.

The **frequency interpretation** of probability is as follows: if the random process is hypothetically repeated, then the long-run proportion of times an event occurs is the probability of the event.

Example 1.10. Consider the experiment of flipping a coin $1, 2, 3, \ldots, 10000$ times using the software `R`. For each coin flip, we record the proportion of heads. Then plot the proportion of heads vs. the number of trials. Here is the code:

```
nflips <- 10000
Head <- TRUE
Tail <- FALSE
flips <- sample(c(Tail,Head),size=nflips,replace=T,prob=c(0.5,0.5))
length(flips)
tot.trials <- 1:nflips # or seq(1,nflips,1)
prop.heads <- rep(0,nflips)
for(i in 1:nflips){

outcomes <- flips[1:i]
num.heads<- sum(outcomes)
flips.sofar <- i
prop.heads[i] <- num.heads/flips.sofar

}

plot(1:nflips,prop.heads, ylab="Proportion of heads", xlab="Number of trials",
col="red",type="l",main="Proportion of heads vs. the number of trials",ylim=c(0,1))
abline(h=0.5)
```

And here is the figure

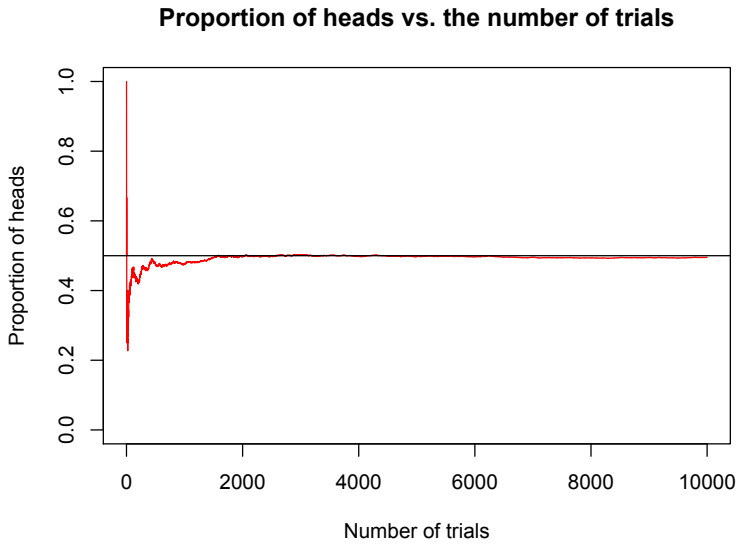

Figure 6: Illustration of the frequency interpretation of probability. The probability of a head is conceived as the long-run relative frequency of heads in a very large number of trials. As the number of trials approaches infinity, the relative frequency of heads approaches exactly 1/2

The **subjective interpretation** of probability involves personal statements of belief regarding the chance of a given event. Subjective probabilities vary from individual to individual. A betting scenario is an ideal representation for this interpretation:

Example 1.11. UF basketball team will play this saturday. If you bet a dollars to my b dollars that UF will win, your probability that UF wins is

$$P(\text{UF wins}) = \frac{a}{a+b}$$

The **odds** refers to the ratio of a probability to 1 minus that probabiity

$$\text{your odds in favor of UF} = \frac{P(\text{UF wins})}{1 - P(\text{UF wins})}$$

If an event has probability 2/3 of happening, the odds are $(2/3)/(1/3) = 2$. Usually this is reported as "the odds of the event happening are 2 to 1". Later on, when studying bayesian statistics, we will talk more about odds.

The different interpretations of probability briefly described above lead to intrinsically different ways of carrying a statistical analysis in science.

1.3.1 Probability properties

1. For any event A, $0 \leq P(A) \leq 1$.
2. If Ω is the sample space, $P(\Omega) = 1$.

Before stating the third, we need one definition: the **complement** of the event A, denoted A^c is defined as the event that A *does not occur*. A^c includes all the elementary events that are not in A. If $\Omega = \{A, A^c\}$ then, the thrid property is

3. $P(A^c) = 1 - P(A)$ or $P(A^c) + P(A) = P(\Omega) = 1$

The next item is the definition of **disjoint events**: two events A and B are disjoint if they have no outcomes in common. Disjoint events are also defined as "mutually exclusive" events because the occurrence of one of the events excludes the possibility of the occurrence of the other event.

Example 1.12. Consider the experiment: "rolling two dice". Let A be the event: a total of 7 shows. Let B be the event: a total of 11 shows. Then, A and B are mutually exclusive. If A occurs, B cannot occur (and viceversa): if you observe event A, a total of 7, you could not at the same time observe event B, a total of 11.

Note: The case above where $\Omega = \{A, A^c\}$ is a special situation where the events A and $B = A^c$ are both complementary to each other and disjoint, because the sample space consists of only two events. That is not always true: in the example above, A and B are disjoint but not complementary to each other.
We now move to the next property:

4. **addition rule for disjoint events:** if A and B are disjoint events, then

$$P(A \text{ or } B) = P(A) + P(B)$$

Example 1.13. Consider the example above with elementary events, where we flipped a coin four times in a row. One event consisting of 4 disjoint, elementary outcomes is $A = \{HHHT,\ HHTH,\ HTHH,\ THHH\}$. Then, by the addition rule

$$P(A) = \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{4}{16}.$$

Another event, B is defined as $B = \{HHTT,\ HTHT,\ HTTH,\ THHT\, THTH,\ TTHH\}$. Then

$$P(B) = \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{6}{16}$$

and using the addition rule again

$$P(A \text{ or } B) = \frac{4}{16} + \frac{6}{16} = \frac{10}{16}$$

Suppose that the probability of event A is the same whether event B has or has not occurred, that is

$$P(A \text{ given } B) = P(A|B) = P(A).$$

Then we say that the occurrence of event A is not dependent on the occurrence of event B, that is, A and B are **independent events**.

Example 1.14. Consider the experiment "draw two balls from a urn". Let $B =$ red on first draw and $A =$ red on second draw. If we draw two with replacement -that is returning each ball to the urn after drawing them-, then A and B are independent. If we draw two without replacement then A and B are dependent (not independent). When sampling without replacement, if the event B occurs, then the ratio of the number of ways we could obtain A to the total number of possible draws changes. Thus, in that case the event B changes the probability of the event A.

The next property of probability refers to independent events:

5. **Multiplication rule for independent events:** if A and B are independent events, then

$$P(A \text{ and } B) = P(A)P(B)$$

Example 1.15. Flip a coin 2 times and let $A = H$ on first flip, $B = H$ on second flip. Then

$$P(HH) = P(A \text{ and } B) = P(A)P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Example 1.16. flip a coin n times

$$P(\text{all heads}) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\ldots\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^n$$

Example 1.17. Suppose we have a urn with three red balls and one blue ball and draw 2 with replacement. Let A = red on the first draw, B = blue on the second draw

$P(A \text{ and } B) = P(\text{picking red\#1 or \#2 or \#3 on 1st draw}) \text{ and } P(\text{picking blue on the 2nd draw}),$

so

$$P(A \text{ and } B) = \left(\frac{1}{4} + \frac{1}{4} + \frac{1}{4}\right) \times \left(\frac{1}{4}\right) = \frac{3}{4} \times \frac{1}{4} = \frac{3}{16}$$

If two events are not disjoint, we use the

6. **Addition rule for any two events**:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Example 1.18. Flip a coin 3 times. Let A = 2 or 3 H and B = 1 or 2 T. Then $P(A \text{ or } B) = \frac{4}{8} + \frac{6}{8} - \frac{3}{8} = \frac{7}{8}$

Example 1.19. A = 2 or more H, B = 3 T. A and B are disjoint. Hence $P(A \text{ or } B) = \frac{4}{8} + \frac{1}{8} - 0 = \frac{5}{8}$.

Just as we extended the addition rule for any two events, we can extend the multiplication rule for any two events:

7. **Multiplication rule for any two events**

$$P(A \text{ and } B) = P(B)P(A|B) = P(A)P(B|A),$$

and

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(B)P(A|B)}{P(A)},$$
$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}.$$

The last two equations are known as **Bayes rule**.

Example 1.20. Consider the following experiment, where we flip three identical coins and record the outcome. There are 8 equally likely outcomes of this coin flipping experiment:

$$HHH,\ HHT,\ HTH,\ THH,\ HTT,\ THT,\ TTH,\ TTT.$$

Let A be the event "one or more H" were obtained and B "one or more T" were obtained. Then, $P(A) = 7/8$ and $P(B) = 7/8$, and

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{6/8}{7/8} = 6/7$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{6/8}{7/8} = 6/7$$

Table 2:

	BS	Master	Professional	phD	Total
Female	616	194	30	16	856
Male	529	171	44	26	770
Total	1145	365	74	42	1626

Exercise 1.2. If the events A and B in the coin flipping experiment above are now defined as $A =$ "one or more heads" and $B =$ "on tail exactly", show that $P(B|A) = 3/7$ and $P(A|B) = 1$

Example 1.21. Consider the data cross-tabulating the different educational degrees vs. the gender of the degree recipient (see table 2)

Following the definition of conditional probability we have that,

$$P(\text{Female}) = \frac{856}{1626} \approx 0.53.$$

$$P(\text{Master}|\text{Female}) = \frac{P(\text{Master and Female})}{P(Female)} = \frac{194/1626}{856/1626} = \frac{194}{856},$$

which you could have guessed by intuition. To check if the events "Professional" and "Female" are independent, we do

$$P(\text{Professional}|\text{Female}) = \frac{30}{856} \approx 0.0350467$$

$$P(Professional) = \frac{74}{1626} \approx 0.0455104.$$

Hence, since $P(\text{Professional}|\text{Female}) \neq P(Professional)$ these two events are not independent!

Example 1.22. Testing the test: Medical diagnosis tests are rigorously tested themselves using conditional probabilities. The objective then is to compute the probabilities of an incorrect diagnosis as well as the probabilities of a correct diagnosis. A given diagnosis can be wrong if the patient is healthy but the test yields a positive (**a false positive**) or if the patient has the disease but the test yields a negative (**a false negative**). On the other hand, there are two ways in which a correct diagnosis can happen: if the patient has the disease and the test yields a positive, or if the patient is healthy and the test yields a negative. The probabilities of each one of these correct diagnosis types are called the **test sensitivity** and the **test specificity** respectively. Using conditional probabilities and simple notation, we can easily keep track of all the probabilities involved. Denote the outcome of the diagnosis test as $\mathcal{P}$ for positive and

$\mathcal{N}$ for negative, and the state of the patient as $\mathcal{D}$ when the disease is present and as $\mathcal{D}^c$ when it is not. Then, we can easily write down the correct diagnosis probabilities as a function of the probabilities of a false negative and of a false positive:

$$\text{test sensitivity} = P(\mathcal{P}|\mathcal{D}) = 1 - P(\mathcal{N}|\mathcal{D}), \quad \text{and the}$$

$$\text{test specificity} = P(\mathcal{N}|\mathcal{D}^c) = 1 - P(\mathcal{P}|\mathcal{D}^c).$$

Here's an example: suppose that we are testing a new test that attempts to diagnose if a pregnant woman is carrying a fetus with the Down Syndrome. 5282 pregnant women where tested using both, a costly yet perfect testing method and the test of interest. Using a completely reliable test and the test of interest allows evaluating the quality of this new test. Below (see table 3) I cross-tabulated the test diagnosis and the real Down Syndrome status for these 5282 women and their child: The conditional

Table 3:

	$\mathcal{P}$	$\mathcal{N}$	Total
$\mathcal{D}$	48	6	54
$\mathcal{D}^c$	1307	3921	5228
Total	1355	3927	5282

probabilities of interest can then be quickly computed using `R` :

```
# Entering the data, with row and column names
test.results <- matrix(c(48,6,1307,3921), nrow=2,ncol=2,byrow=TRUE);
row.names(test.results) <- c("D", "Dc");
colnames(test.results) <- c("Pos", "Neg");
print(test.results)

# Computing totals
rowtotals <- apply(test.results,2,sum);
coltotals <- apply(test.results,1,sum);
grand.tot <- sum(test.results);
print(rowtotals)
print(coltotals)
print(grand.tot)

# Notation:
# Pos = positive test; Neg = negative test
# D = Disease present; Dc = Disease absent

# False Positive rate: P(Pos|Dc) = P(Pos AND Dc)/P(Dc)
Ppos.G.Dc <- test.results[2,1]/coltotals[2]
```

```
# False Negative rate: P(Neg|D) = P(Neg AND D)/P(D)
Pneg.G.D <- test.results[1,2]/coltotals[1]

# True Negative rate, Specificity = P(Neg|Dc) = 1- P(Pos|Dc)
Pneg.G.Dc <- 1 -  Ppos.G.Dc
# True Positive rate, Sensitivity = P(Pos|D) = 1- P(Neg|D)
Ppos.G.D <- 1 - Pneg.G.D

print(Pneg.G.Dc)
print(Ppos.G.D)
```

8. **Law of total probability:** In what follows we use the symbol $\cup$ to the denote the **union of two events**, that is, the set of all outcomes that are included in either the first or the second event. Hence $P(A \,\text{or}\, B)$ is written as $P(A \cup B)$. The **intersection of two events** is the set of all outcomes that are included in both events and is the noted by the symbol $\cap$. Hence $P(A \,\text{and}\, B)$ is written as $P(A \cap B)$. Let $B_1, B_2, \ldots, B_n$ be n mutually exclusive events such that

$$\bigcup_{i=1}^{n} B_i = \Omega \quad \text{and} \quad P(B_i) > 0 \quad \text{for} \quad \text{all} \quad i.$$

Furthermore, let A be any event in Ω. Then:

$$P(A) = \sum_{1=1}^{n} P(A \cap B_i) = \sum_{1=1}^{n} P(A|B_i)P(B_i).$$

It follows that Bayes rule can be extended to the law of total probability:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{1=1}^{n} P(A|B_i)P(B_i)}.$$

Example 1.23. First, here's a simple example that illustrates the law of total probability: A drawer has 2 white socks and 2 blue socks. Daniel reaches in and draws out 2 socks in succession without replacement. Let $W1 =$ "white on the first draw", $W2 =$ "white on the second draw. Likewise, let $B1 =$ "Blue on the first draw" and $B2 =$"Blue on the second draw". A tree diagram illustrating the events and conditional probabilities is shown in Figure 1.23

Then,

$$P(\text{two white socks}) = P(W1 \,\text{and}\, W2) = P(W2|W1)P(W1) = \frac{1}{3}\frac{1}{2} = \frac{1}{6},$$

$$P(\text{two socks of the same color}) = P(W1 \,\text{and}\, W2) + P(B1 \,\text{and}\, B2) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3},$$

$$P(\text{second sock is blue}) = P(B2) = P(B2|B1)P(B1) + P(B2|W1)P(W1) = \frac{1}{3}\frac{1}{2} + \frac{1}{2}\frac{2}{3} = \frac{1}{2}.$$

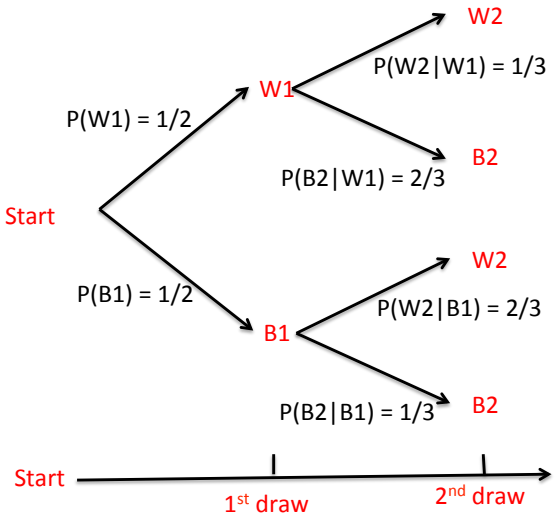

Figure 7: Tree diagram for the experiment of drawing 2 socks in succession from a drawer

Finally, using the law of total probability we can compute the probability of getting a white sock on the first draw given that on the second draw we got a blue sock:

$$P(W1|B2) = \frac{P(B2|W1)P(W1)}{P(B2)} = \frac{P(B2|W1)P(W1)}{P(B2|B1)P(B1)+P(B2|W1)P(W1)}$$

$$= \frac{\frac{2}{3}\frac{1}{2}}{\frac{1}{3}\frac{1}{2}+\frac{1}{2}\frac{2}{3}} = \frac{1/3}{1/2} = \frac{2}{3}$$

Exercise 1.3. Recall the Down syndrome example from above. Solve for the specificity or true negative rate ($P(\mathcal{N}|\mathcal{D}^c)$) and the sensitivity or true positive rate ($P(\mathcal{P}|\mathcal{D})$) using the law of total probability (I want a formula, not a number). The R code for this calculation is below. Use it to find these formulae

```
# Another way of doing these calculations:

# Get the table of probabilities by
# elementwise division of all counts by the grand total:
probs.table <- test.results/grand.tot
print(probs.table)

# False Positive rate  P(Pos | DisAb) = P(Pos AND DisAb)/P(DisAb)
P.posGdisab <- probs.table[2,1]/(probs.table[2,1]+ probs.table[2,2])

# False Negative rate
P.NegGdis <- probs.table[1,2]/(probs.table[1,1]+probs.table[1,2])
```

```
# True Positive rate: P(Pos|Dis) = P(Pos AND Dis)/P(Dis)
P.posGdis <- probs.table[1,1]/(probs.table[1,1]+probs.table[1,2])

# True Negative rate: P(N|DisAb) = P(N AND DisAb)/P(DisAb)
P.NegGdisab <- probs.table[2,2]/(probs.table[2,1]+probs.table[2,2])

# And just to check:
print(P.posGdis)
print(1-P.NegGdis)

print(P.NegGdisab)
print(1-P.posGdisab)
```

Exercise 1.4. Suppose that the company that produces these tests makes publicly available these results. Now, suppose also that a government inspector must decide whether a new, randomly picked positive sample of the test belongs to a patient with the disease. That is, for the sample at hand, he needs to know $P(\mathcal{D}|\mathcal{P})$. How would you compute such probability using the law of total probability?

1.4 Probability distributions

1.4.1 discrete case

A **random variable** is a numerical outcome of a random experiment (usually denoted with an upper case letter, like X). Examples of random variables are:

- Roll two dice and let $X =$ the sum of the two numbers that appear.
- 1 day's growth in dry weight of a plant.
- Number of Democrats in a random sample of voters (small here in Idaho!).

The **Probability Distribution** of a random variable is the collection of all of its possible outcomes and their associated probabilities *i.e.*, for all possible outcomes x (lower case for outcomes) we give a value to $P(X = x)$. It is said to be **a discrete random variable** if there is a finite or countable sequence of possible values x.

Example 1.24. Flip three coins and let X be the number of heads (H) that we see. There are 8 equally likely outcomes of this coin flipping experiment:

$$HHH,\ HHT,\ HTH,\ THH,\ HTT,\ THT,\ TTH,\ TTT.$$

Out of these 8 outcomes, only one contains no heads (TTT). The event TTT thus occurs with probability 1/8. Likewise, the event HHH of having 3 heads in a row, occurs with probability 1/8. The event of having only 1 head occurs in 3 instances (HTT, THT, TTH) and hence, $P(\text{one head}) = 3/8$. Finally, the event of having two heads also occurs in 3 out of the 8 outcomes and thus, $P(\text{two heads}) = 3/8$. Thus, the number of heads after flipping a coin 3 times in a row can be modeled with a random variable X that can take on the values $0, 1, 2$ and 3 heads with probabilities $1/8, 3/8, 3/8$ and $1/8$ respectively. The probability mass function for the random variable X is therefore given by:

x	0	1	2	3
$P(X = x)$	1/8	3/8	3/8	1/8

The **cumulative distribution function** of any random variable (discrete, continuous or in between) is given by:

$$F(x) = P(X \leq x).$$

In this case we have:

$$F(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1/8 & \text{if} \quad 0 \leq x < 1 \\ 4/8 & \text{if} \quad 1 \leq x < 2 \\ 7/8 & \text{if} \quad 2 \leq x < 3 \\ 1 & \text{if} \quad 3 \leq x \end{cases}$$

To check this, note for example that for $1 \leq x < 2$, $P(X \leq x) = P(X \in \{0, 1\}) = 1/8 + 3/8$.

Often, more than one random variable can be defined on the same sample space. Suppose that two discrete random variables X and Y are defined on the same sample space and that they take on values $x_1, x_2 \ldots$ and $y_1, y_2 \ldots$ Their **joint frequency function** $p(x_i, y_j)$, or joint probability mass function is defined as

$$p(x_i, y_j) = P(X = x_i \,\text{and}\, Y = y_j) = P(X = x_i, Y = y_j).$$

The following example illustrate the calculation of such joint probability mass function.

Example 1.25. A fair coin is tossed 3 times. Let X be the number of heads on the first toss and Y the total number of heads. Then, the sample space is:

$$\Omega = \{HHH,\, HHT,\, HTH,\, THH,\, HTT,\, THT,\, TTH,\, TTT\}$$

and the joint frequency function of X and Y is given by:

	y			
x	0	1	2	3
0	1/8	2/8	1/8	0
1	0	1/8	2/8	1/8

Thus, for example $P(X = 0, Y = 2) = 1/8$. If we wish to find the frequency function of Y from the joint frequency function, then in general, we simply have to sum down the appropriate column of the table.$p_Y(y) = P(Y = y)$ is called the **marginal frequency function** of Y:

$$\begin{array}{rcrcrcl}
P(Y=0) &=& P(Y=0,X=0) + P(Y=0,X=1) &=& \frac{1}{8} + 0 = \frac{1}{8} \\
P(Y=1) &=& P(Y=1,X=0) + P(Y=1,X=1) &=& \frac{2}{8} + \frac{1}{8} = \frac{3}{8} \\
P(Y=2) &=& P(Y=2,X=0) + P(Y=2,X=1) &=& \frac{1}{8} + \frac{2}{8} = \frac{3}{8} \\
P(Y=3) &=& P(Y=3,X=0) + P(Y=3,X=1) &=& 0 + \frac{1}{8} = \frac{1}{8}
\end{array}$$

Two discrete random variables X and Y are said to be **independent** if:

$$P(X = x, Y = y) = P(X = x) \times P(Y = y).$$

Example 1.26. Consider the following joint distribution of X and Y:

	Y=		
X=	1	2	Tot
1	.3	.1	.4
2	.4	.2	.6
Tot	.7	.3	1

Are X and Y independent? Well, let's check if the definition above holds: First, note that the probability that $X = 1$ and that at the same time $Y = 2$ is 0.1. That is, $P(X = 1, Y = 2) = 0.1$. Now, $P(X = 1) \times P(Y = 2) = 0.4 \times 0.3 = 0.12 \neq 0.1$. Therefore we conclude that the random variables are not independent.

Exercise 1.5. Now let the distribution function be:

	Y=		
X=	1	2	Tot
1	.28	.12	.4
2	.42	.18	.6
Tot	.7	.3	1

Are X and Y independent?

The independence concept for a collection of events can be extended when we are dealing with jointly distributed random variables. Consider the first example above (example 1.26) with two jointly distributed random variables X and Y. Then the multiplication rule for any two events written in the context of these two jointly distributed random variables is written as:

$$P(X = x, Y = y) = P(X = x|Y = y)P(Y = y) = P(Y = y|X = x)P(X = x).$$

Example 1.27.

$$P(X = 1, Y = 2) = P(Y = 2|X = 1)P(X = 1) = \frac{0.1}{0.4} \times 0.4 = 0.1.$$

convince yourselves that using $P(X = x, Y = y) = P(X = x|Y = y)P(Y = y)$ with $x = 1$ and $y = 2$ leads to the same result.

If we fix y and look at $P(X = x|Y = y)$ as a function of x, what we have is the conditional distribution of X given that $Y = y$.

Example 1.28. Consider the following joint probability distribution for the random variables X and Y:

	Y=				
X=	100	125	150	175	Tot
5	.08	.08	.06	0	.22
5.5	.08	.16	.16	.08	.48
6	0	.08	.10	.12	.30
Tot	.16	.32	.32	.2	1

Then,

$$P(X = 5|Y = 150) = \frac{P(X = 5, Y = 150)}{P(Y = 150)} = \frac{0.06}{0.32},$$

$$P(X = 5.5|Y = 150) = \frac{P(X = 5.5, Y = 150)}{P(Y = 150)} = \frac{0.16}{0.32},$$

$$P(X = 6|Y = 150) = \frac{P(X = 6, Y = 150)}{P(Y = 150)} = \frac{0.10}{0.32},$$

or in other words we take the third column of the table and divide it by its sum to make it a probability distribution.

1.4.2 Mean and variance of a discrete random variable

Suppose X is a discrete random variable that takes on the values in $S = 0, 1, 2, 3, \ldots, n$ with probabilities $P(X = 0), P(X = 1), \ldots, P(X = n)$. Then, the mean of X, denoted $E[X]$ (**Expected value** of X) is defined to be:

$$\begin{aligned} E[X] &= \sum_{x \in S} xp(x) \\ &= 0 \times P(X = 0) + 1 \times P(X = 1) + \ldots \times P(X = n). \end{aligned} \tag{17}$$

Example 1.29. For the random variable $X =$ "Flip three coins and let x be the number we get", the mean is computed as follows:

$$E[X] = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{12}{8} = \frac{3}{2}.$$

Generally, if $h(X)$ is a function of a discrete random variable X, then

$$E[h(X)] = \sum_{x \in S} h(x) P(X = x). \tag{18}$$

In particular, if $h(X) = X^2$, then $E[X^2] = \sum_{x \in S} x^2 P(X = x)$. (Watch out, $E[X^2] \neq E[X]^2$!).

The variance of X, Var$[X]$ it's a measure of the average square departures from the mean of X, $E[X] = \mu$. That is, it is the $E[h(X)]$ where $h(X) = (X - \mu)^2$.

$$\begin{aligned} \text{Var}[X] &= E\left[(X - E[X])^2\right] \\ &= \sum_{x \in S} (x - \mu)^2 P(X = x). \end{aligned} \tag{19}$$

Example 1.30. For the above example, we had $E[X] = \mu = \frac{3}{2}$, so the variance of X is:

$$\begin{aligned} \text{Var}[X] &= \left(0 - \tfrac{3}{2}\right)^2 \tfrac{1}{8} + \left(1 - \tfrac{3}{2}\right)^2 \tfrac{3}{8} + \left(2 - \tfrac{3}{2}\right)^2 \tfrac{3}{8} + \left(3 - \tfrac{3}{2}\right)^2 \tfrac{1}{8} \\ &= 2\left(\tfrac{3}{2}\right)^2 \tfrac{1}{8} + 2\left(\tfrac{1}{2}\right)^2 \tfrac{1}{8} = \tfrac{3}{4} \end{aligned}$$

Now, in the above definition of the variance, replace μ by $E[X]$ and develop

the square in the sum to get:

$$\begin{aligned}
\mathrm{Var}[X] &= \sum_{x\in S} \left(x - \mathrm{E}[X]\right)^2 P(X=x) \\
&= \sum_{x\in S} \left(x^2 P(X=x) - 2x\mathrm{E}[X]P(X=x) + \mathrm{E}[X]^2 P(X=x)\right) \\
&= \sum_{x\in S} x^2 P(X=x) - \sum_{x\in S} 2x\mathrm{E}[X]P(X=x) + \sum_{x\in S} \mathrm{E}[X]^2 P(X=x) \\
&= \overbrace{\sum_{x\in S} x^2 P(X=x)}^{\mathrm{E}[X^2]} - 2\mathrm{E}[X]\overbrace{\sum_{x\in S} xP(X=x)}^{\mathrm{E}[X]} + \mathrm{E}[X]^2 \overbrace{\sum_{x\in S} P(X=x)}^{1} \\
&= \mathrm{E}\left[X^2\right] - 2\mathrm{E}[X].\mathrm{E}[X] + \mathrm{E}[X]^2 \\
&= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2.
\end{aligned} \tag{20}$$

Hence, another way to compute the variance of X, if it exists, is:

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2. \tag{21}$$

Exercise 1.6. Use eq. 21 to compute the variance of X.

1.4.3 Rules for expected values and variances

Here are some useful facts about means and variances. Most of them are illustrated using discrete random variables for simplicity, but these rules are also valid for continuous random variables. If a and b are constants, then:

$$E[aX + b] = aE[X] + b. \tag{22}$$

For example, if $E[X] = 0$ and $Y = X+b$, then $E[Y] = E[X]+b = b$. In general, if $X_i, i = 1, 2, \ldots n$ are jointly distributed random variables with expectations $E[X_i]$ and Y is a linear function of the X_i, $Y = a + \sum_{i=1}^{n} b_i X_i$, then

$$E[Y] = a + \sum_{i=1}^{n} b_i E\left[X_i\right]. \tag{23}$$

Using the same arguments to derive eq. (21), one can show that, if a and b are constants, then

$$\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]. \tag{24}$$

Proof: Let $Y = aX + b$. Since $E[Y] = aE[X] + b$, then

$$\begin{aligned}
\mathrm{E}[(Y - \mathrm{E}[Y])^2] &= \mathrm{E}\left\{\left[aX + b - a\mathrm{E}[X] - b\right]^2\right\} \\
&= \mathrm{E}\left\{a^2\left[X - \mathrm{E}[X]\right]^2\right\} \\
&= a^2\mathrm{E}\left\{\left[X - \mathrm{E}[X]\right]^2\right\} \\
&= a^2\mathrm{Var}[X]
\end{aligned}$$

Now, let X and Y be two random variables with means μ_X and μ_Y respectively. Furthermore, let σ_X^2 and σ_Y^2 denote their variances. Then, if X and Y are independent,

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y]. \tag{25}$$

However, if they are not independent,

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X,Y], \tag{26}$$

where

$$\text{Cov}[X,Y] = \text{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \tag{27}$$

is the **covariance** of the two random variables X and Y.The covariance between these two random variables is a measure of their joint variability, or their degree of association. Eq.27 can be expanded and re-written as

$$\text{Cov}[X,Y] = \text{E}[XY] - \text{E}[X]\text{E}[Y], \tag{28}$$

where $\text{E}[XY] = \sum_x \sum_y xyP(X=x, Y=y)$. In particular, note that if X and Y are independent, then $\text{E}[XY] = \text{E}[X]\text{E}[Y]$ and the covariance is 0. The converse however is not true in general.

The **correlation coefficient**, usually denoted ρ, is a dimensionless quantity that varies between -1 and 1 that is defined in terms of the covariance. From the rules from variances and expected values, it is easily seen that if X and Y are both subjected to linear transformations (such as changing their units from inches to meters), the covariance value can change but the correlation coefficient does not change, since it does not depend on the units of measurement. It follows that ρ is in many cases a more useful measure of association than the covariance (Rice 1995). ρ is defined to be:

$$\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}}. \tag{29}$$

It then follows that eq. (26) can be re-written as:

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2\rho\sigma_X\sigma_Y. \tag{30}$$

If we have more than two random variables and we need to compute the covariance of a linear combination of those random variables, the following facts are needed: first, suppose that $U = a + \sum_{i=1}^n b_i X_i$, $V = c + \sum_{j=1}^n d_j Y_i$, where X_i's and the Y_j's are random variables and a, the b_j's and the d_j's are all constants. Then,

$$\text{Cov}(U,V) = \sum_{i=1}^n \sum_{j=1}^n b_i d_j \text{Cov}(X_i, Y_j).$$

The formula above leads to a general way to compute the variance of a linear combination of random variables:

$$\text{Var}\left(a + \sum_{i=1}^n b_i X_i\right) = \sum_{i=1}^n \sum_{j=1}^n b_i b_j \text{Cov}(X_i, X_j).$$

Finally, if the X_i are all independent, then $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$ and

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

1.4.4 Elements of counting

First, we do a little bit of counting to understand the binomial formula

Example 1.31. A man has 4 pairs of pants and 6 shirts. In how many ways can he get dressed? 4×6. To see this, just make a 2 by 2 table listing all the possible outcomes.

Multiplication rule: In general, suppose that m experiments are performed in order and that no matter what the outcomes of experiments $1, \ldots, k-1$ are, experiment k has n_k possible outcomes. Then, the total number of outcomes is $n_1 \times n_1 \times \ldots \times n_k \ldots \times n_m$.

Exercise 1.7. A restaurant offers soup or salad to start, and has 11 entrees to choose from, each of which is served with rice, baked potato or zuchini. How many meals can you have if you can choose to eat one of their 4 desserts or have no dessert?

First of all, note that we have here 4 experiments: choosing a start, an entrée, a side dish and a dessert. So, to get the answer we just apply the multiplication rule

$$\text{number of meals} = 2 \times 11 \times 3 \times 5 = 330.$$

Q: How many ways can 5 people stand in line?

To answer this question, we think about building the line up one person at a time starting from the front. There are 5 people we can choose to put at the front of the line. Having made the first choice, we have 4 possible choices for the second position. Similarly, we have 3 possible choices for the third position, 2 for the fourth and 1 for the last. Invoking the multiplication rule we get:

$$5 \times 4 \times 3 \times 2 \times 1 = 5!$$

and in general, with n items we define n **factorial** as

$$n(n-1).(n-2)\ldots 2.1 = n!$$

Example 1.32. Twelve people belong to a club. How many ways can they pick a president, vice-president, secretary and treasurer? We have 12 choices for president. Once this choice has been done, we have 11 choices for vice-president, then we have 10 positions for secretary and finally only 9 left for treasurer. Therefore:

$$\text{number of ways to pick the 4 offices} = 12 \times 11 \times 10 \times 9.$$

In general, if we have k offices and n club members, then the answer is:

$$n.(n-1).(n-2)\dots(n-k+1) \tag{31}$$

Multiplying and dividing the above by $(n-k)!$ we have:

$$n.(n-1).(n-2)\dots(n-k+1)\frac{(n-k)!}{(n-k)!} = \frac{n!}{(n-k)!},$$

which is the expession for the permutation of n things taken k at a time. We're almost there. Here is an exercise to practice the above formula:

Exercise 1.8. In a horse race, the first three finishers are said to "Win", "Place" and "Show". How many finishes are possible for a race with 11 horses?

In the problem above, the order in which the choices were made was important (show example). Next we consider the case in which the choice made is not important:

Example 1.33. A club has 23 members. How many ways can they pick 4 people to be on a committee to plan a party?

By the previous example we imagine making the committee members stand in line which by eq. 31 can be done in 23.22.21.20 ways. However the number of different committees is less than that, because here we are not interested in the individual positions. Hence, noting that each committee can stand in line in 4! ways, to get the total number of committees we divide 23.22.21.20 by 4!, that is:

$$\frac{23.22.21.20}{4!} = 23.11.7.5 = 8855.$$

In general, suppose we want to pick k people out of a group of n, then, the number of **combinations of n things taken k at a time** is

$$\frac{n.(n-1).(n-2)\dots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!} \tag{32}$$

1.4.5 Continuous random variables

Although in this course a lot of time is spent developing the most important continuous distributions and its applications through examples in ecology and genetics, in what follows some of the most basic facts about two continuous probability distributions are outlined. These distributions are the uniform distribution and the normal distribution.

A **continuous random variable X is real-valued**. The possible values for Y, form intervals of real numbers (which are uncountable infinite sets of possible values).

Probability density function (pdf, or probability curve): The probability that the continuous random variable X takes a value in the interval (a, b) is the area under the pdf $f(x)$ between a and b:

$$P(a \le Y \le b) =) \int_a^b f(x)dx = \text{area under the pdf curve between } a \text{ and } b.$$

Note two things. First, the area under the entire pdf is 1. If the random variable X has support from $-\infty$ to $+\infty$, then

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Second, $P(X = x) = 0$ for any x in the support of X.
The expected value of mean of X is a measure of the center of the distribution of X. It is a constant indicating where the pdf would balance on a fulcrum, where the value of x acts as weights. If X is a continuous random variable with support in $-\infty, \infty$, then

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} xf(x)dx. \end{aligned} \tag{33}$$

The variance of a continuous random variable X is also an expectation: it measures the spread of the distribution and is computed as the expected value of a squared deviation of X from its mean $\mu = \mathrm{E}[X]$:

$$\begin{aligned} \mathrm{Var}[X] &= \int_{\infty}^{\infty} (x-\mu)^2 f(x)dx \\ &= \mathrm{E}[(X-\mu)^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2. \end{aligned} \tag{34}$$

The concept of expectation is also important to describe other two well known quantities describing the shape of a continuous distribution: the **skewness** and the **kurtosis**. The skewness is a measure of asymmetry of the distribution and the kurtosis is a measure of the peakedness of the distribution. If $\mu = \mathrm{E}[X]$ and
$sigma^2 = \mathrm{V}[X]$, then the skewness is given by

$$\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

and the kurtosis is given by

$$\frac{\mathrm{E}\left[X^4\right]}{\sigma^2}.$$

Example 1.34. The pdf of the Uniform distribution on (a, b) is given by

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b \text{ and } 0 \text{ otherwise.}$$

Since $\frac{dx^2}{dx} = 2x$, it is easy to see that the expected value of the uniform distribution X between 0 and 1 is:

$$\mathrm{E}(X) = \int_0^1 xf(x)dx = \int_0^1 x(1)dx = \left[\frac{1}{2}x^2\right]\Big|_0^1 = \left[\frac{1}{2}(1)^2 - \frac{1}{2}(0)^2\right] = \frac{1}{2}.$$

If $X \sim \text{Unif}(0,1)$, then $Y = (b-a)X + a$ is uniformly distributed between a and b. From the expected value of a uniform distribution in $(0,1)$ we can easily move to the expected value of a uniform distribution Y between a and b:

$$\text{E}(Y) = \text{E}\left[(b-a)X + a\right] = (b-a)\text{E}[X] + a = (b-a)\frac{1}{2} + a = \frac{a+b}{2}.$$

The variance of the uniform distribution in $(0,1)$ is given by

$$\begin{aligned}
\text{Var}[X] &= \int_0^1 \left(x - \frac{1}{2}\right)^2 (1)dx \\
&= \int_0^1 \left[x^2 - x + \left(\frac{1}{2}\right)^2\right] dx \\
&= \int_0^1 x^2 dx - \int_0^1 x dx + \int_0^1 \frac{1}{4} dx \\
&= \left[\frac{1}{3}x^3\right]\Big|_0^1 - \left[\frac{1}{2}x^2\right]\Big|_0^1 + \left[\frac{1}{4}x\right]\Big|_0^1 \\
&= \frac{1}{3} - \frac{1}{2} + \frac{1}{4} = \frac{1}{12}.
\end{aligned} \tag{35}$$

In the homework we will work with the Normal distribution.

1.4.6 Sampling distributions

Definitions and concepts:

1. **A statistical universe** is a defined set of elements in which you are interested. Ex.: Adult women in the US.

2. **A statistical population** is the set of values associated with each of the elements of the universe. Ex. the height of women in the US. Note that this is a collection of values, not individuals.

3. A probability distribution serves as a model for a population of quantities. Ex. The normal distribution serves as a model for the collection of all heights of adult women in the US:

4. **Parameters** are fixed quantities (usually unknown) that characterize the properties of the distribution. Ex. The true mean μ and variance σ^2 of the heights of adult women in the US. Usually a researcher is interested in making the best guess about the value of these parameters.

5. It is often impossible to measure all the values of the population of interest. **A sample** is a collection of measured values from that population. A **Simple Random Sample** (**SRS** hereafter) is a sample selected in such a way that all possible samples were equally likely to be selected.

Ex.: Let our universe be the last 6 individuals of a bird population. Let our statistical population (the quantities associated with the birds that we are interested in) be their tail length in cms. Let these lengths be, for the sake of simplicity, $1, 2, 3, 4, 5, 6$ cms. Assume each individual value is equally likely to be sampled and its tail can be measured **exactly**. Then, since each value occurs once, the population of tail lengths is described by the distribution Y below:

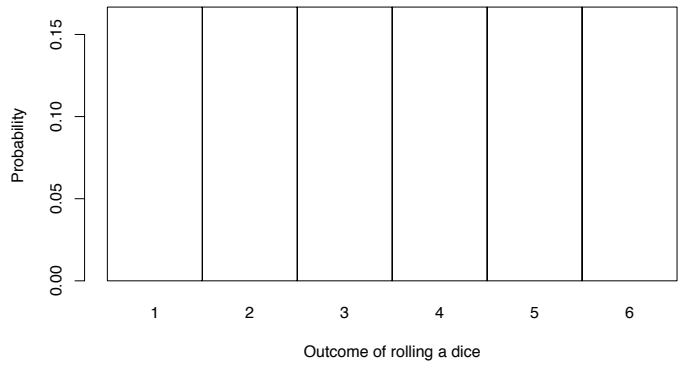

Figure 8: Probability distribution for Y: this is identical to the pmf of the outcome of rolling a dice.

What are the Expected value and variance of Y?

$$\begin{aligned} E[Y] &= \sum_{k=1}^{6} k \times P(Y = k) \\ &= 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6) \\ &= 3.5 \end{aligned}$$

and

$$\begin{aligned} Var[Y] &= \sum_{k=1}^{6}(k-3.5)^2 \times P(Y=k) \\ &= (1-3.5)^2(1/6) + (2-3.5)^2(1/6) + \ldots + (6-3.5)^2(1/6) \\ &= 2.916667 \end{aligned}$$

Then $E[Y] = \mu = 3.5$ and $Var[Y] = \sigma^2 = 2.916667$.

Now, suppose we take a SRS of size n. Then, the experiment "I capture a bird in the i^{th} occasion, measure (exactly) its tail length and put it back where it was", will be an outcome from the random variable Y_i, and the outcome of sampling n times will be the outcome of the random variables $Y_1, Y_2, \ldots, Y_n$, which are pairwise independent and identically distributed to Y above.

Now suppose that we take k samples of size 2. That is, we reach into our universe and sample 2 individual values, one after the other, k times. Each time I repeat the experiment, we expect to get different outcomes. Then, how many different samples of size $n = 2$ can I take? Well, I have 6 possible choices for the first sample and 6 possible choices for the second sample, then the answer is 36. Suppose that for each of these 36 different samples of size 2 we average the measurements. Then we have:

$\frac{1+1}{2} = 1$	$\frac{2+1}{2} = 1.5$	$\frac{3+1}{2} = 2$	$\frac{4+1}{2} = 2.5$	$\frac{5+1}{2} = 3$	$\frac{6+1}{2} = 3.5$
$\frac{1+2}{2} = 1.5$	$\frac{2+2}{2} = 2$	$\frac{3+2}{2} = 2.5$	$\frac{4+2}{2} = 3$	$\frac{5+2}{2} = 3.5$	$\frac{6+2}{2} = 4$
$\frac{1+3}{2} = 2$	$\frac{2+3}{2} = 2.5$	$\frac{3+3}{2} = 3$	$\frac{4+3}{2} = 3.5$	$\frac{5+3}{2} = 4$	$\frac{6+3}{2} = 4.5$
$\frac{1+4}{2} = 2.5$	$\frac{2+4}{2} = 3$	$\frac{3+4}{2} = 3.5$	$\frac{4+4}{2} = 4$	$\frac{5+4}{2} = 4.5$	$\frac{6+4}{2} = 5$
$\frac{1+5}{2} = 3$	$\frac{2+5}{2} = 3.5$	$\frac{3+5}{2} = 4$	$\frac{4+5}{2} = 4.5$	$\frac{5+5}{2} = 5$	$\frac{6+5}{2} = 5.5$
$\frac{1+6}{2} = 3.5$	$\frac{2+6}{2} = 4$	$\frac{3+6}{2} = 4.5$	$\frac{4+6}{2} = 5$	$\frac{5+6}{2} = 5.5$	$\frac{6+6}{2} = 6$

Each of these means is a **statistic**: a quantity calculated from the sample, usually for the purpose of estimating an unknown parameter. Furthermore, note that each possible value of the statistic has an associated probability of occurrence. For example, the value 1.5 occurs 2 out of the 36 total number of different outcomes. So its associated probability is 2/36. Hence, **the statistics, such as the sample mean are themselves random variables!!!**. The probability distribution defined by the possible values of the sample mean and their associated probabilities define the **exact sampling distribution of the sample mean** $\bar{Y_n}$:
This is why using capital letter for random variables is important! Not doing it keeps us from fully understanding that a statistic is a random variable!!!
Examples of statistics: If a sample of size n is modeled using n independent and identically distributed (iid) random variables $Y_1 \ldots Y_n$, then

- The sample mean of size n:

$$\bar{Y_n} = \frac{1}{n}\left(Y_1 + Y_2 + \ldots + Y_n\right).$$

$\bar{Y_n}$	$P(\bar{Y_n} = \bar{y})$
1	1/36
1.5	2/36
2	3/36
2.5	4/36
3	5/36
3.5	6/36
4	5/36
4.5	4/36
5	3/36
5.5	2/36
6	1/36

- The sample variance for a size n:

$$S^2 = \frac{1}{(n-1)}\left[\left(Y_1 - \bar{Y_n}\right)^2 + \left(Y_2 - \bar{Y_n}\right)^2 + \ldots + \left(Y_n - \bar{Y_n}\right)^2\right]$$

Note that, here, because our population was very small, we could easily list **all the possible values of a sample mean of size** $n = 2$. However, it is often the case that the population is large enough so that it is virtually impossible to list all the possible values of the sample mean, for a particular n. In that case, we can at best make an **Estimate of the sampling distribution of the sample mean**. A very important theorem, the Law of Large Numbers, tells us that the estimation of that sampling distribution improves as the sample size increase.

Now, because we are trying to make inferences about the population, it is of interest to know the properties of the sampling distribution of the sample mean.

Exercise 1.9. Compute the expected value and the variance of $\bar{Y_n}$ above. These values should be equal to $\mathrm{E}[\bar{Y_n}] = 3.5 = \mu$ and $\mathrm{Var}[\bar{Y_n}] = 1.4583 = \frac{2.9166667}{2} = \frac{\sigma^2}{n}$.

Here are the 4 most important properties of the sampling distribution of the sample mean:

1. Let $Y_1, Y_2, \ldots, Y_n$ be *iid* random variables with mean μ and variance σ^2. Let $\bar{Y_n} = \frac{Y_1+Y_2+\ldots+Y_n}{n}$, then $E\left[\bar{Y_n}\right] = \mu$:

$$\begin{aligned} E\left[\bar{Y_n}\right] &= E\left[\frac{Y_1 + Y_2 + \ldots + Y_n}{n}\right] \\ &= \frac{1}{n}E\left[Y_1 + Y_2 + \ldots + Y_n\right] \\ &= \frac{1}{n}E\left[nY_i\right] \\ &= \frac{1}{n}nE\left[Y_i\right] \\ &= \mu. \end{aligned}$$

2.

$$\text{Var}\left[\bar{Y}_n\right] = \frac{\sigma^2}{n}.$$

$$\begin{aligned}
\text{Var}\left[\bar{Y}_n\right] &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&= \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^{n} Y_i\right) \\
&= \left(\frac{1}{n}\right)^2 n\text{Var}\left(Y_1\right) \\
&= \frac{\sigma^2}{n}
\end{aligned} \tag{36}$$

3. **The Law of large numbers** says that the sample mean $\bar{X}_n$ is close to the mean μ of the underlying population when the sample size n is large. This implies, that, if we want to determine the average height of the 21 year old males in the US, we do not have to measure the height of all of the more than 1 million people in that category, but instead, we can **estimate** this quantity by measuring the heights of say, 1000 individuals. How close will the mean of a sample of size 1000 will be to the mean of the underlying population? We will deal with this question in a bit. Formally, the **Weak Law of Large Numbers** is:

   Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with common mean $EX_i = \mu$, and $E|X_i| < \infty$. Then, as $n \to \infty$,

$$P\left(|\bar{X}_n - \mu| > \epsilon\right) \to 0, \quad \text{for} \quad \text{all} \quad \epsilon > 0.$$

   (We say then that $\bar{X}_n$ converges to μ in probability). If $\bar{X}_n$ is thought of as an estimate of μ, this property is called **statistical consistency**.

   Now, **Chebyshev's inequality** let us put a bound on the probability that the sample mean will be ϵ units apart from the true population mean given a sample of size n:

$$P\left(|\bar{X}_n - \mu| \geq \epsilon\right) \leq \frac{\text{Var}\bar{X}_n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

   Example 1.35. How large does my sample size needs to be? Chebyshev's inequality let us find the answer to that very common question in biological experiments. Suppose we flip a coin $n = 10000$ times and let X_i be 1 if the i^{th} toss is a Head or 0 otherwise, so that $\bar{X}_n$ is the fraction of the first n tosses that result in Heads. The X_i are Bernoulli trials with success probability $P(X_i = 1) = p = \frac{1}{2}$ mean and variance:

$$\mathrm{E}[X_i] = P(X_i = 1) = \frac{1}{2};\ \mathrm{Var}[X_i] = p(1-p) = \frac{1}{4}.$$

Taking $\epsilon = 0.01$ and using Chebyshev's inequality we get:

$$P\left(\left|\bar{X}_n - \frac{1}{2}\right| \geq 0.01\right) \leq \frac{1/4}{(0.01)^2 10^4} = \frac{1}{4}.$$

Since a fourth seems too high for a bound on the probability that the sample mean will be ϵ units apart from the true population mean given a sample of size n, we have to keep flipping coins to get a tighter distribution of the sample mean! In particular, since that bound is exactly $\frac{\sigma^2}{n\epsilon^2}$, and if we want to have a probability bound of, say $\delta = 0.10$ that the sample mean will be within ϵ units apart from the true population mean, we can easily solve for what the sample size needs to be in order to reach that bound:

$$n = \frac{\sigma^2}{\delta\epsilon^2} = \frac{1/4}{(0.10)(0.01)^2} = 25000$$

The next theorem is motivated by the question: If $X_1, X_2 \ldots$ are *iid* random variables, what's the distribution of $X_1 + X_2$? of $\frac{X_1+X_2}{2}$?

4. **The Central Limit Theorem**

*Suppose $X_1, X_2, \ldots, X_n$ are **any** independent and identically distributed random variables with common mean $EX_i = \mu$, and variance $\sigma^2 < \infty$. Then, as $n \to \infty$,*

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) \to P(Z \leq z), \quad \text{where} \quad Z \sim N(0,1).$$

In other words,$\bar{X}_n$ will converge to a normal distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$.

Let $S_n = X_1 + X_2 + \ldots + X_n$. Then, $E[S_n] = n\mu$ and $Var[S_n] = n\sigma^2$. Then, the CLT also says that, if X_i has **any** distribution, then

$$S_n \to N(n\mu, n\sigma^2).$$

In fact, note that often n does not have to be very large for that convergence to occur. If $\bar{X}_n$ is thought of as an estimate of μ, this property is called asymptotic normality