Distributions

Devan Allen McGranahan & Carissa Wonkka

September 15, 2016

This chapter serves as a basic introduction to data distributions and covers classification by variable types (continuous vs. discrete), plotting distributions of sampled data, and parameterizing Probability Density and Probability Mass Functions by moment-matching and Maximum Likelihood Estimation. The Chapter concludes with a table summarizing common distributions, their data types and support, parameters & moments, and `R` functions for fitting distribution functions and making random draws.

Introduction

All of our statistical analysis consists of matching data and models. Each of these models have *assumptions*, or conditions of the data that must be met before the conclusions drawn by the models can be taken as valid. Simply put, violating assumptions compromises the validity of our analysis and erodes confidence in our conclusions, and when stated as such appears to be a quite critical component of the research process. But while many reviewers are keen to scrutinize sample size and nit-pick sampling methods, the nitty-gritty of statistical modeling is likely to be overlooked by the designers, reporters, and reviewers of ecological research. As researchers, we are often taken at our word that we have done our due diligence in making sure we "did our stats right," and the bar is unfortunately low among ecologists.[1]

Almost all models have one or more assumptions that relate to the *distribution* of the data, which makes it important for ecologists to understand distributions before even considering statistical models, let alone assessing their validity. It is convenient to think of distributions as the *shape of the data*, which is determined by the probability of observing each of the values in the dataset. The more frequently a value occurs in the data, the greater its probability, and the shape refers to the line that connects these probabilities across the range of the data. The mathematics that underlie statistical models assume the data fit a given shape, and the validity of conclusions drawn from these models rests—at least in part—on the distribution of the data fitting the correct shape.

[1] In our defense, many of the quirks that come with the sort of studies ecologists must conduct to study complex, long-term, or broad phenomena in inherently uncontrolled biological systems often render the appropriate model for the data surprisingly complex. More statistical nuance is demanded of ecologists than from their peers in biology, whose controlled petri dishes or tended plots require little more than the most simple statistical tests.

Two major classes of distributions

It should not be surprising that the two major classes of distributions sort out along the same lines as the major classes of variables—continuous and discrete[2]. Within each class, though, there are several options based on the r*support* of the data—the range, or bounds, of possibilities over which the data can occur.

From a practical standpoint the various distributions within each class represent the options available for making sure our data fit model assumptions– matching data shape to a distribution shape is the link between our data and statistical models. Once we are happy we have a distribution that matches the shape of our data, we can proceed with statistical analysis with confidence that distribution assumptions, at least, have been met.

The first step is to become familiar with the various distribution options. We focus here on those (1) commonly encountered by ecologists and (2) used in `R` distribution and statistical analysis functions.

[2] Be careful, though, here we mean the technical definition of discrete—whole numbers, with distinct gaps of impossible non-whole values—not the colloqial discrete used synonymously with "categorical" when we're talking about graphing.

Continuous distributions

The continuous class includes the distribution most frequently-used by "conventional" statistics like ANOVA and linear regression—the normal, or Gaussian, distribution.[3] Again, continuous variables are those that can take on any value between two specified values—any number of decimal points between two whole numbers, depending only on the precision with which we can measure. Given that there are an infinite number of potential values, then, the probability that a random draw—a sample—is of a particular value is... zero. This presents us with a conundrum–since unlike discrete distributions—for which we can calcuate an exact probability for each potential outcome (more on that below)—continuous distributions require us to describe potential outcomes with an equation, known as the *Probability Density Function*, or PDF.

[3] An ecology student who has also taken a basic statistics course, however, might be quite surprised to learn how abnormal it is for the normal distribution to be appropriate for ecological data, but that's why we're here, right?

The Probability Density Function defines the aforementioned "shape" of the data, based on the probability of a given value occurring in a dataset. *Density* effectively refers to how many values—or more accurately, the probability of finding that many values—fall under a certain point on the curve; the more values, or the greater the probability, the higher the density. An important property about PDFs is that the area under the curve between the beginning and ending points always sums to 1: there is 100% probability that a random sample occurs somewhere within those bounds (Fig. 1).

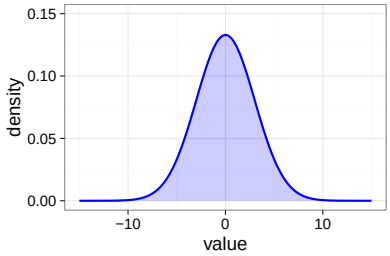

Figure 1: The curved dark line traces plots the Probability Density Function, of PDF, and draws through the infinite possible values for data in this set. The area of the shaded portion below the curve sums to 1. Note the Y axis label, "density;" see the text and become familiar with this concept.

Continuous distributions—Cast of characters Table 1 describes various continuous distributions and the type of data that fit them. Examples of each are plotted in Fig. 2

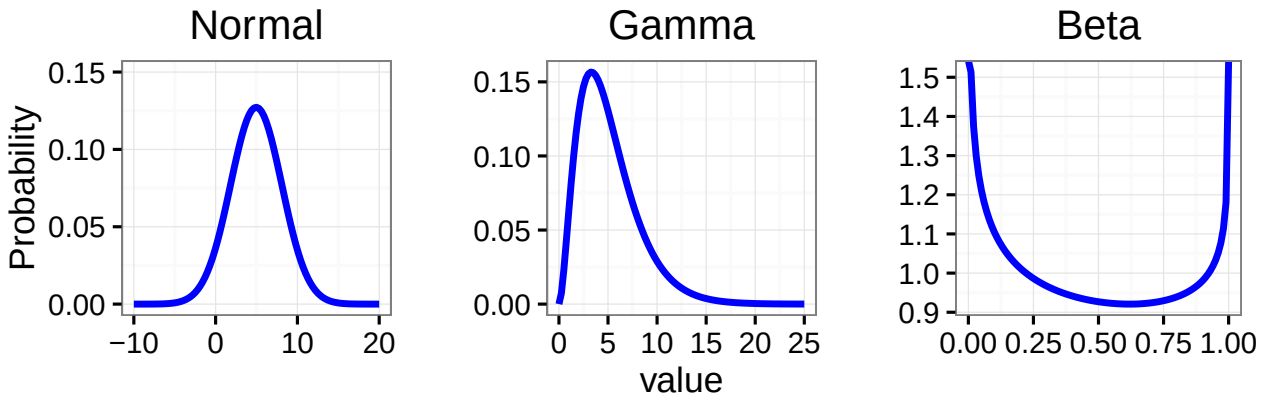

Figure 2: Shapes of three continuous distributions. All have mean=5.

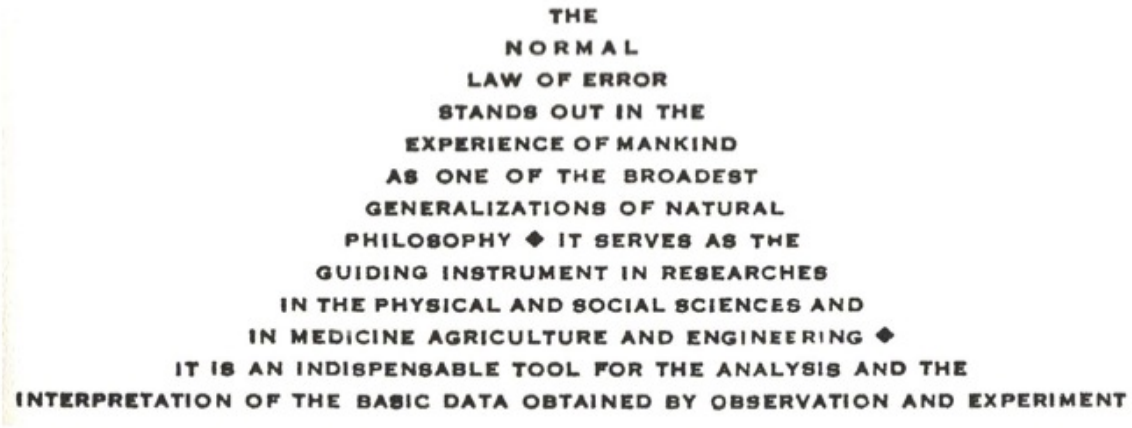

Figure 3: A poetic take on the normal distribution by Jack Youden.

Discrete distributions

Discrete distributions apply to random variables that have a finite set of possibilities within the support, or range, of the data. There are two subgroups that correspond with the two types of discrete variables:

- Discrete distributions apply to integers—R class `integer`—which are discrete because their values are confined to whole numbers. Such data include counts.
- Categorical distributions are specific to groups—R class `factor`—and describe as probabilities the results of a random event that has one or more possible outcomes. Basic examples are survivorship, presenceabsence, or rolls of a die (six fixed, pre-determined outcomes).

The probabilities of data that follow discrete distribution are described with *probability mass functions*, which gives the probability that the random variable is exactly one of the data values or event outcomes.

Discrete distributions—Cast of characters Table 1 describes various continuous distributions and the type of data that fit them. Some examples are plotted in Fig. 4.

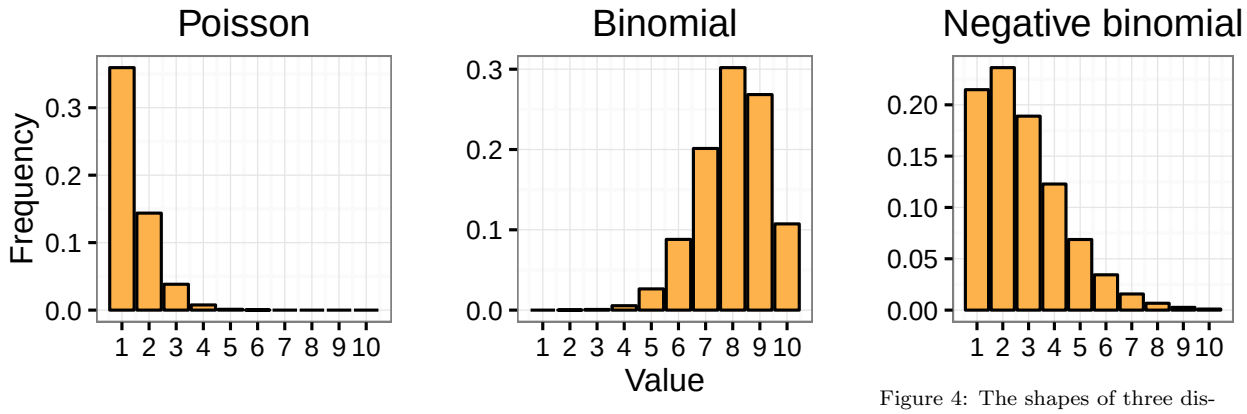

Figure 4: The shapes of three discrete distributions. Parameters set as `size=10` for each; for Poisson, `lambda=0.8` and for both binomial and negative binomial `prob=0.8`.

Fitting data models

Let's review our steps so far in visualizing the distribution of our data:

- Determine if we are dealing with a continuous or discrete distribution
- Plot a histogram or bar graph to show how many observations were recorded across the range of data collected

- For continuous distributions, we also plot a density estimate that begins to connect the distribution of our sample to the population it is meant to represent[4]

[4] Density estimates can also be plotted for discrete variables and might be helpful to see the overall trend of the shape, but this is technically incorrect because it will interpolate between the discrete values where we assume observations cannot occur.

The next step is to parameterize a Probability Density Function or Probability Mass Function from our data to see how well the data fit theoretical distributions. If the PDF or PMF has the proper shape (see Fig. 5) and our data more or less follow the curve, we can proceed knowing that the PDF or PMF represents our data well and we can confidently apply a statistical model that assumes our data have that shape. If either of these are not the case, we have two options: (1) transform our data so that they match the shape of the PDF or PMF, or (2) fit a different theoretical distribution that is closer to the shape of our data.[5]

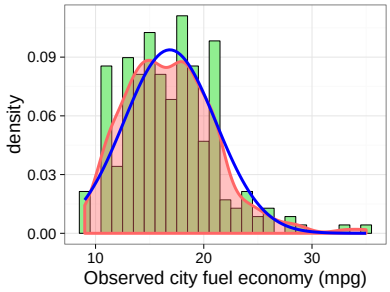

Figure 5: Although the density estimate and normal PDF overlap well, the normal distribution is a poor fit because observations towards the upper bound give the distribution an unsymetrical tail. A log transformation improves symetry.

But first: what does it mean to "model" data? We really just mean, "fit a curve to the data," or find the PDF that fits our data the best. Once we know what equation best describes the shape of our data, we can:

- Find an appropriate statistical test by matching the shape of our data to the shape the equations within the test assume our data fit—*i.e.* meet the assumptions of the statistical model.
- Generate a set of random numbers based on our observed distribution.

Random number generation is the basis of *data simulation* and serves several purposes:

- Estimate a confidence interval, *e.g.* to determine the probability the data do or do not overlap a particular value, such as 0.
- Increase sample size to boost statistical power, based on the theory that if one had been able to collect more data, they would fall along the curve described by the PDF.
- Fit data for "custom" models that base statistical tests on the moments of a specific dataset. Most off-the-shelf statistical functions in `R` and other software are pre-programmed with math and make assumptions about the shape of our data, as described above. But one can program this math and make it specific to the data. This is one of the first steps in Bayesian statistics and begins with moments.

[5] There can be disadvantages to transformations that merit consideration before application. Many researchers go to great lengths to apply transformations in search of normality in their data to satisfy the assumptions of ANOVA or Ordinary Least-Squares regression when a non-parametric test or Generalised Linear Model might fit the data without transformation.

Parameter estimation

Two options for parameter estimation include (1) Moment-matching and (2) Maximum Likelihood Estimation.[6]

[6] For a more detailed description of each and a lot of `R` code for both, check out V. Ricci's manual *Fitting Distributions with* `R` on r-cran: `ftp://cran.r-project.org/pub/R/doc/contrib/Ricci-distributions-en.pdf`

Moment-matching In addition to assumptions, another important concept from statistics is necessary to *use* distributions: Moments. From our friends at Wikipedia:

> . . . a moment is a specific quantitative measure, used in both mechanics and statistics, of the shape of a set of points. . . . If the points represent probability density, then the zeroth moment is the total probability (i.e. one), the first moment is the mean, the second central moment is the variance, the third moment is the skewness, and the fourth moment (with normalization and shift) is the kurtosis.

We'll get to each moment in turn, but for now we focus on first and second moments: the mean and variance of the data (since we've already covered the zeroth moment–total probability). These moments are important because they "drive" the shape of the curve—the math that plots the PDF makes the calculations based on the mean and variance of the data.

> *If one knows the mean and variance of a dataset, one can fit a PDF and model those data.*

Remember mean is denoted as μ and variance as σ^2 but beware the difference between standard deviation, σ, and variance, σ^2, which is the square of standard deviation.

This is usually a pretty simple approach; notice how in Table 1 all of the distribution parameters can be expressed mathematically in terms of just the mean (μ) and variance (σ^2), which are easy to calculate from a data vector. *But this approach assumes that the mean and variance of the sample are equal to corresponding values in the broader population*, which might not be an assumption we can take at face value. And either way, plugging μ and σ^2 into some of the distribution equations—*e.g.*, the beta distribution (Table 1)—can make for some long strings of code and increase the probability of committing a programming or arithmetic error.

```
> gg1 <- ggplot(msleep, aes(x=sleep_total)) +
+           geom_histogram(aes(y=..density..),
+             binwidth=0.5, color="black",
+             fill="lightgreen") +
+           geom_density(alpha=0.3, color="#FF6666",
+                        fill="#FF6666")
> mean(msleep$sleep_total)

[1] 10.43373

> sd(msleep$sleep_total)

[1] 4.450357

> gg1 <- gg1 + stat_function(data=msleep, fun=dnorm,
+                 colour="blue", size=1.3,
+                 args=list(mean=10.43, sd=4.45))
```

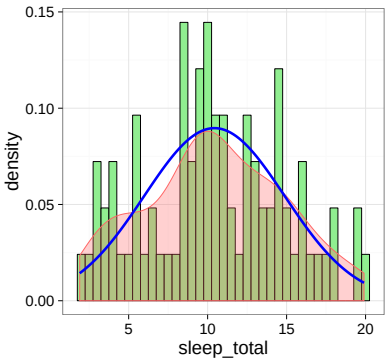

Figure 6: A PDF based on the normal distribution fit by moment-matching μ and σ from mammal sleep data.

Maximum Likelihood Estimation is useful when one needs a more robust estimator of moments in the population and/or wants to shorten code a bit by plugging parameters directly into distribution equations with minimal programming.

```
> # Load Gerlough and Schuhl (1955) traffic data
> turns <- c(rep(0,14),rep(1,30),rep(2,36),rep(3,68),
+           rep(4,43),rep(5,43),rep(6,30),rep(7,14),
+           rep(8,10),rep(9,6),rep(10,4),rep(11,1),rep(12,1))
> traffic <- data.frame(turns)
> gg2 <- ggplot(traffic, aes(x=turns)) + theme_bw() +
+           geom_histogram(aes(y=..density..),
+                     binwidth=1, color="black",
+                     fill="#feb24c")
```

`R` includes several options for Maximum Likelihood Estimation. Perhaps the most basic one—in that it only works for single variables at a time—is `fitdistr()` from the `MASS` package, which comes with your `R` installation but needs to be specifically called even though it is already installed:

Other MLE functions include `stats::nlm()` and `stats4::mle()`.

```
> library(MASS)
> fitdistr(traffic$turns, "Poisson")

    lambda
  3.893333
 (0.113920)

> mean(traffic$turns)

[1] 3.893333

> gg.pn <- gg2 + geom_point(data=transform(data.frame(x=0:12),
+                                  y=dpois(x, lambda = 3.89)),
+                      aes(x, y), stat="identity", pch=21,
+                      bg="#2b8cbe", col="white", size=4)
> fitdistr(traffic$turns, "negative binomial")

      size          mu
  11.1555359    3.8933229
 ( 3.6876098) ( 0.1323139)

> gg.nb <- gg2 + geom_point(data=transform(data.frame(x=0:12),
+                                  y=dnbinom(x, size=11.16,
+                                            mu=3.89)),
+                      aes(x, y), stat="identity", pch=21,
+                      bg="#2b8cbe", col="white", size=4)
```

Note how we sort of trick `ggplot` into taking our PMF as a `data.frame` and simultaneously create a vector `x` for `dpois()`. This `transform(data.frame...` scheme works well for plotting discrete distributions with `ggplot`.

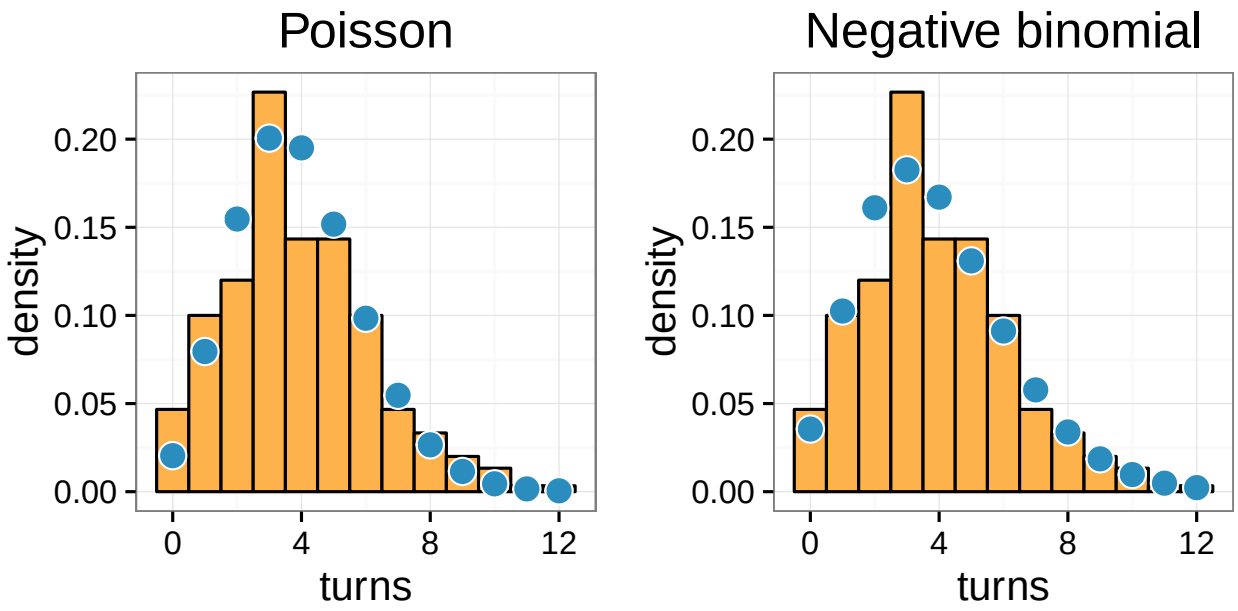

Figure 7: Comparison of two PMFs fit by calls to `MASS::fitdistr()`. Note that the PMF is indicated by points, not a curve as in the PDF, since the PMF makes explicit probability predictions for each point on the X axis.

Distribution	Example data	Parameters & Moments	R functions
Continuous			
Normal	Any continuous quantity, positive or negative. Sums, measurements, etc.	μ, σ^2	`dnorm(x, mean, sd)`, `rnorm(n, mean, sd)`
Lognormal	Continuous, non-negative quantities. *Note:* Specific properties can apply to log values.	α (μ on log scale), β (σ^2 on log scale)	`dlnorm(x, meanlog, sdlog)`, `rlnorm(n, meanlog, sdlog)`
Gamma	Continuous, positive (>0) quantities.	α = shape: $\frac{\mu^2}{\sigma^2}$ β = rate: $\frac{\mu}{\sigma^2}$	`dgamma(x, shape, rate)`, `rgamma(n, shape, rate)`
Beta	Continuous quantities between 0 and 1; proportions and probabilities.	$\alpha = \frac{(\mu^2-\mu^3-\mu\sigma^2)}{\sigma^2}$ $\beta = \frac{\mu-2\mu^2+\mu^3-\sigma^2+\mu\sigma^2}{\sigma^2}$	`dbeta(x, shape1, shape2)`, `rbeta(n, shape1, shape2)`
Uniform	Any real number. Creates a flat line.	α = lower limit of range β = upper limit of range	`dunif(x, min, max)`, `runif(n, min, max)`
Discrete			
Poisson	Counts of things that occur randomly over space.	λ, which = μ	`dpois(x, lambda)`, `rpois(n, lambda)`
Binomial	Number of successes, or hits, in a number of trials; *e.g.* survivors in a group of organisms, plots containing an endangered species, pixels in a raster file that meet a condition.	η, number of trials (size) ϕ, probability of success. $\phi = 1 - \sigma^2/\mu$ $\eta = \mu^2/(\mu - \sigma^2)$	`dbinom(x, size, prob)`, `rbinom(n, size, prob)`
Negative binomial	Number of successes in a sequence of independent Bernoulli trials before a specified (non-random) number of failures occurs. Robust alternative to Poisson.	η, number of trials (size) ϕ, probability of success, or μ; only one of ϕ or μ required. $\phi = 1 - \sigma^2/\mu$ $\eta = \mu^2/(\mu - \sigma^2)$	`dnbinom(x, size, prob/mu)`, `rnbinom(n, size, prob/mu)`
Bernoulli	Special case of binomial where number of trials = 1 and data range between 0 and 1; used in survival analysis and occupancy models.	ϕ, probability that the random variable = 1. $\phi = \mu$ $\phi = 1/2 + 1/2\sqrt{1 - 4\sigma^2}$	`dbinom(x, size=1, prob)`, `rbinom(n, size=1, prob)`
Multinomial	Counts that fall into more than two categories.	$\mathbf{z}$, a vector of number of counts/category ϕ, vector of probabilities per category.	`dmultinom(x, size, prob)`, `rmultinom(n, size, prob)`

Table 1: Summary of distributions, parameters and moments, and R functions. μ = mean and σ^2 = variance. R functions beginning with `d` plot a density function along the vector `x` and those with an `r` generate a random sample of length `n`.