Modeling with Probability Distributions - Capturing Noise

Zack Treisman

Spring 2021

Philosophy

Recall that a model fundamentally looks like

$$y = f(x) + \epsilon$$

where $f(x)$ is the **deterministic** part of the model (the signal) and ϵ is the **stochastic** part (the noise).

- This week we are looking at the **noise**.

We'll look at the how to model noise in the general context of studying probability, and we'll lay some groundwork for some additional applications of probability.

The noise is a probability distribution. So far, we have only considered models where the noise follows a normal (Gaussian) distribution. Choosing the best distribution is an important part of the modeling process.

Breaking up the noise

We have already talked about how the noise can be broken into irreducible and reducible error, and the reducible error can be split into bias and variance. This decomposition is model based.

Another way that the noise can be decomposed is more observation based:

- Measurement error - Unavoidable, but hopefully minimal. If it has structure or pattern, this can cause difficulties, some of which can be overcome (eg. distance sampling).
- Process noise - Natural demographic and environmental variability. Minimized with large samples and stable environments. The main input to the stochastic part of a model.

Conditional distributions

A more computationally convenient phrasing and notation than $y = f(x) + \epsilon$ is to describe noise as a **conditional distribution**.

$$Y \sim \mathbb{P}(f(X))$$

- $f(X)$ represents the expected value of Y as a function of X.
- $\mathbb{P}$ can be any distribution.

A model where applying a link function to f makes it linear in its parameters is called a **generalized linear model** (GLM). Somewhat more general f can be fit with a **generalized additive model** (GAM).

The glm and related commands

Fitting generalized linear models in R is done using glm, generalized additive models with gam.

```
?glm

glm(formula, family = gaussian, data,
    na.action, start = NULL, ...)

?family

binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
...
```

Probability

Definitions and notation

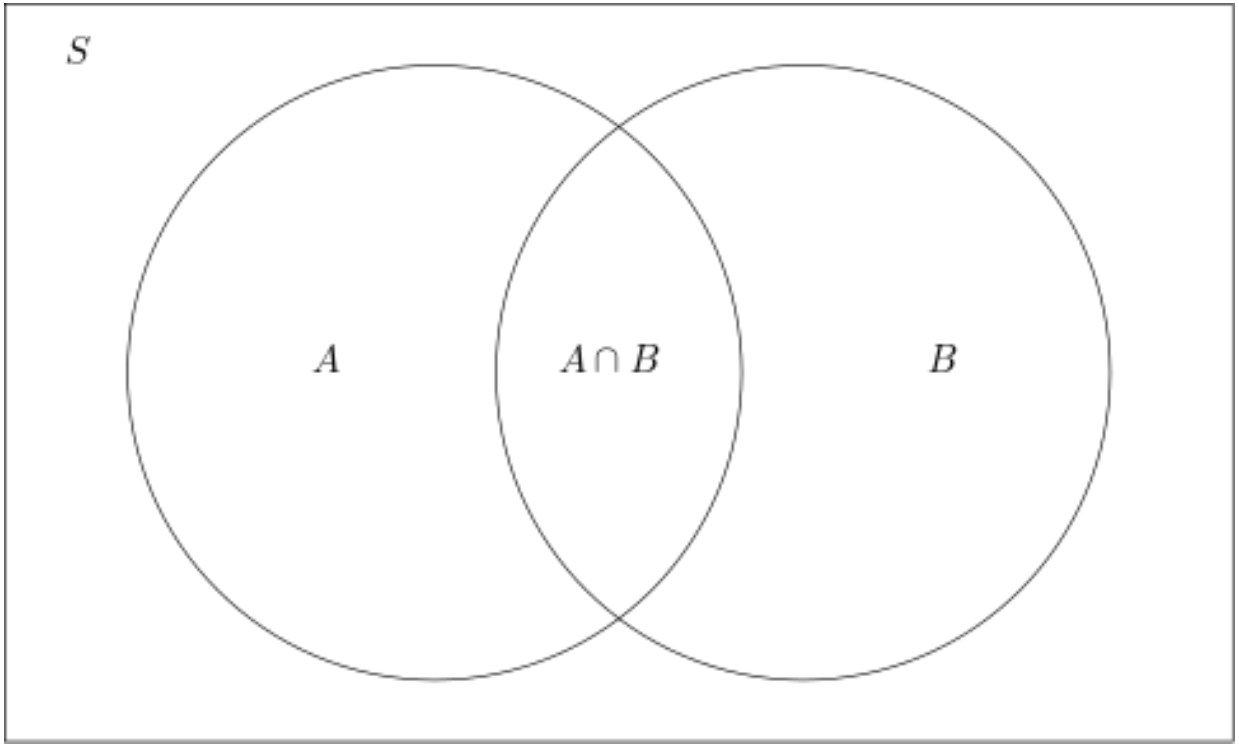
The **sample space** is the set of all possible **outcomes**. Each opportunity for an outcome to occur is a **trial**. Outcomes are collected into **events**. To each event A we assign a number $P(A)$ between 0 and 1 called the **probability** of A representing the frequency with which A occurs.

Example: A feeder at my house is visited by various birds and mammals. Each visit is a trial. The set of all critters that visit the feeder is the sample space. A grey jay visiting is an event. By my estimation $P(\text{Grey Jay}) = 0.3$. One of the grey jays that visits I've named June. June visiting the feeder is an outcome.

Notation

Let A and B be events from a sample space S.

- A *or* B is written $A \cup B$. (Inclusive or: A or B or both.)
- A *and* B is written $A \cap B$.
- The **conditional probability** of A given B, written $P(A|B)$, is the probability that A happens if B is known to happen.



Axioms of probability

The mathematics of probability can be derived from the following three algebraic axioms.

1. $P(S) = 1$: Something has to happen.
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$: The probability of either or both of A or B happening is the sum of their individual probabilities, less the probability that both happen (which was counted twice in the sum).
3. $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$: The probability that A happens given that B has happened can be computed by rescaling the probability that both A and B happen by the probability of B.

Algebra of probability

Some immediate consequences of the axioms that are very useful are the following.

- Since $S = A \cup (\text{not } A)$, combining rules 1 and 2 gives that the probability that A doesn't happen is $P(\text{not } A) = 1 - P(A)$.
- More generally, if A and B are any mutually exclusive events, $P(A \cup B) = P(A) + P(B)$.
- The unconditional probability of an event can be computed by making use of known conditional probabilities: $P(A) = P(A|B)P(B) + P(A|\text{not } B)P(\text{not } B)$
- If $P(A) = P(A|B)$ we say that A and B are **independent**. In this situation, rule 3 implies that $P(A \cap B) = P(A)P(B)$.

Application: Zero-inflated distributions

Consider the seed predation example from Bolker (2008):

A feeder has N seeds. The sample space is the number of seeds taken between occasions when the feeder is checked, so the numbers between 0 and N.

On many occasions, no seeds are taken, in which case it is reasonable to assume that the feeder may not have been visited.

- $P(\text{feeder is visited}) = \nu$.

Assume that a visitor to the feeder independently considers taking each seed.

- $P(\text{seed taken}) = p$.

Application: Zero-inflated distributions (cont.)

If no seeds are taken that means that either nobody visited

$$P(\text{no visit}) = 1 - \nu$$

or a visitor came

$$P(\text{visit}) = \nu$$

and decided not to take each seed

$$\begin{aligned} P(\text{not seed } 1 \cap \cdots \cap \text{not seed } N) &= P(\text{not seed } 1) \cdots P(\text{not seed } N) \\ &= (1-p)^N \end{aligned}$$

Putting these together gives

$$P(\text{no seeds taken}) = 1 - \nu + \nu(1-p)^N$$

Application: Zero-inflated distributions (cont. 2)

On the other hand, the event that x seeds are taken consists of

$$\binom{N}{x} = \frac{N!}{x!(N-x)!}$$

different outcomes (one for each way to select x of N seeds) each with probability

$$p^x(1-p)^{N-x}$$

So for $x > 0$,

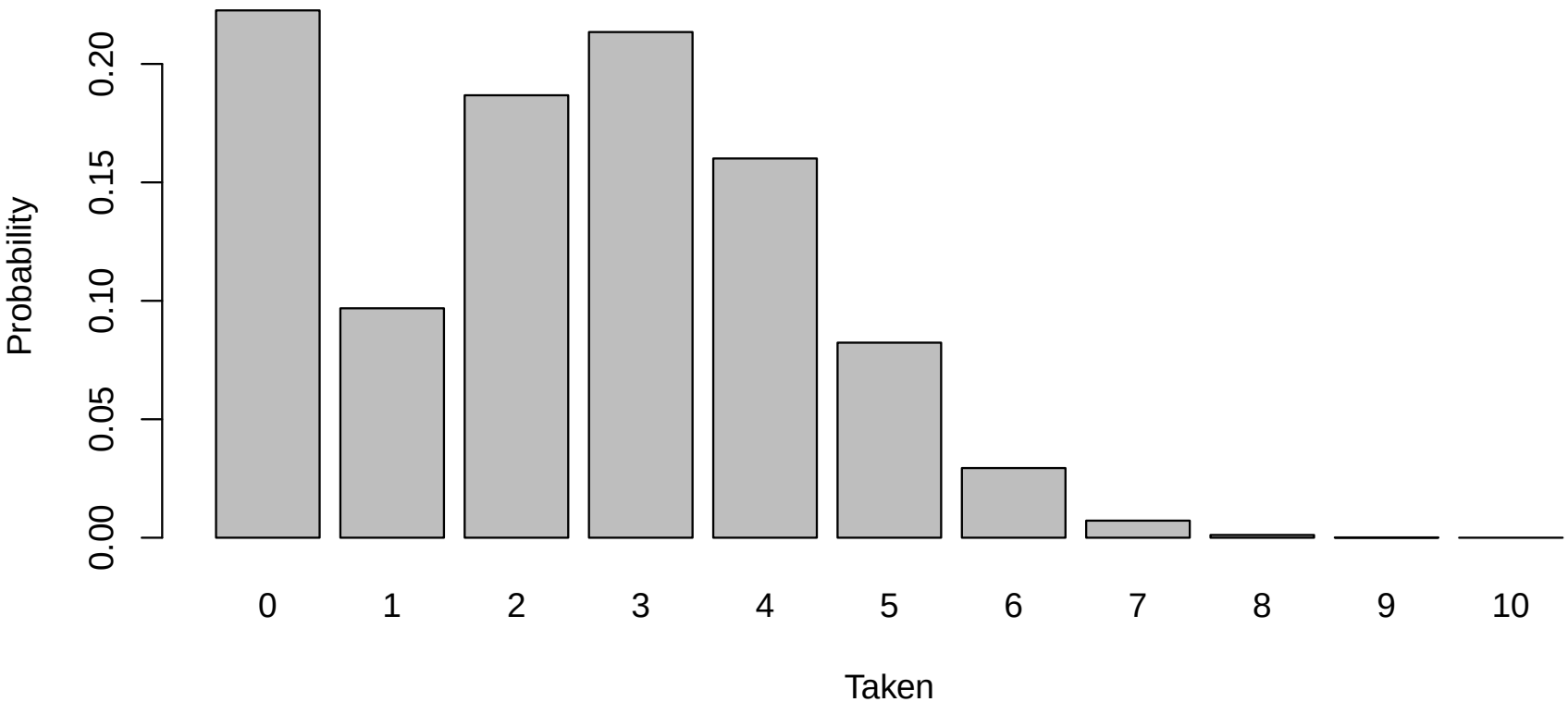
$$P(x \text{ seeds taken}) = \nu \binom{N}{x} p^x(1-p)^{N-x}$$

The distribution that we have just derived is called the **zero-inflated binomial**. Other zero-inflated models are similar, and can be very useful in ecological applications.

Zero-inflated binomial in R

This code defines and plots a zero-inflated binomial model. See Figure 4.1 in Bolker.

```
N <-10 # number of seeds per feeder
nu <- 0.8 # visit probability
p <- 0.3 # probability of taking each individual seed
dzibinom <- numeric(N+1) # Initialize an empty vector of length N+1
dzibinom[1] <- 1-nu+nu*(1-p)^N # Zero seeds taken
for(x in 1:N) { # x seeds taken
  dzibinom[x+1] <-nu*choose(N,x)*p^x*(1-p)^(N-x)}
barplot(dzibinom, names.arg=0:N, xlab="Taken", ylab="Probability")
```



Bayes' Rule and Bayesian Statistics

Bayes Rule

Bayes' Rule allows us to reverse the conditionality in probability calculations. For any events A and B:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Turning the conditionality around is useful for hypothesis testing.

- Hypothesis: A; Observed Data: B
- $P(A|B)$ is the probability that the hypothesis is true given the observed data. (What we want to know.)
- Everything on the right side we can calculate from assumptions in the hypothesis and observations of the data.

Derivation of Bayes' Rule

Before we talk about how to use it, let's take a second to algebraically justify the rule. Going from the axioms of probability to Bayes' Rule is satisfyingly easy:

First, $A \cap B = B \cap A$ are just two ways of writing the same event "A and B". Second, rephrasing axiom 3 above gives

$$P(A \cap B) = P(A|B)P(B)$$

and

$$P(B \cap A) = P(B|A)P(A).$$

Thus

$$P(A|B)P(B) = P(B|A)P(A)$$

and dividing both sides by $P(B)$ gives Bayes' Rule.

From Bayes' Rule to Bayesian statistics

Bayes' Rule applied to observed data D and hypothesis H:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

We want the left hand side. What can we do with the three terms on the right?

- $P(D|H)$: Called the **likelihood** of the data.
  - H gives a distribution for calculating $P(\text{datum}|H)$.
  - Independent data means $P(D|H) = \prod_{\text{datum} \in D} P(\text{datum}|H)$.
  - Often calculate *log likelihoods* in practice.
- $P(H)$: This is called the **prior**.
  - Assumed before looking at the data.
  - Choosing appropriate priors can be difficult and controversial.
  - Default to *uninformative* or *uniform* priors.
- $P(D)$: Using a set H_1 to H_N of *exhaustive, mutually exclusive* hypotheses:

$$\begin{aligned} P(D) &= \sum_{j=1}^{N} P(D \cap H_j) \\ &= \sum_{j=1}^{N} P(D|H_j)P(H_j) \end{aligned}$$

Bayes' Rule: Example

- Disease with prevalence of 10 per 100,000.
- Test that is 99% accurate. (For both positive and negative results; these could also be different.)

Two questions: Given a positive test result how likely is it that the subject is sick (sensitivity), and given a negative result, how likely is it that the subject is healthy (specificity)?

- We'll calculate the sensitivity:

$$
\begin{aligned}
P(\text{sick}|\text{test}+) &= \frac{P(\text{test}+|\text{sick})P(\text{sick})}{P(\text{test}+)} \\
&= \frac{P(\text{test}+|\text{sick})P(\text{sick})}{P(\text{test}+|\text{sick})P(\text{sick}) + P(\text{test}+|\text{not sick})P(\text{not sick})} \\
&= \frac{0.99 \times 0.00001}{0.99 \times 0.00001 + 0.01 \times 0.99999} \\
&\approx 0.001
\end{aligned}
$$

So this 99% accurate test is not *nearly* sensitive enough for reliably detecting cases of this disease.

Distributions

Definitions and notation

Let X be a random variable. Traditional notation uses X for the variable and x for particular values.

"I worried for a long time about what the term 'random variable' means. In the end I concluded it means: 'variable'." -J.H.Conway

- The **probability distribution function** of X tells us the probability that X takes a particular value.
  - X discrete: $f(x) = P(X = x)$
  - X continuous: $\int_a^b f(x)dx = P(a \leq X \leq b)$
- The **cumulative distribution function** of X is $F(x) = P(X \leq x)$.

R has many distributions built in. See `?Distributions`.

Moments

A probability distribution defines an **expectation** operation.

$$E[z] = \sum_x zf(x) \text{ or } E[z] = \int zf(x)dx.$$

- $E[x] = \mu = \bar{x}$ is the **mean**.
- $E[(x - \bar{x})^2] = \sigma^2$ is the **variance**.

Continuing in a similar way defines the **skewness** and **kurtosis** (heavy-tailedness). If you need to numerically measure these things you are probably doing something fancy and mathematically impressive.

Method of moments: To choose a particular distribution from an assumed family, calculate moments for data and use these to compute appropriate parameters.

Binomial

X: The number of successes after repeated independent and identical trials.

Parameters are N, the number of trials, and p, the probability of success on each trial.

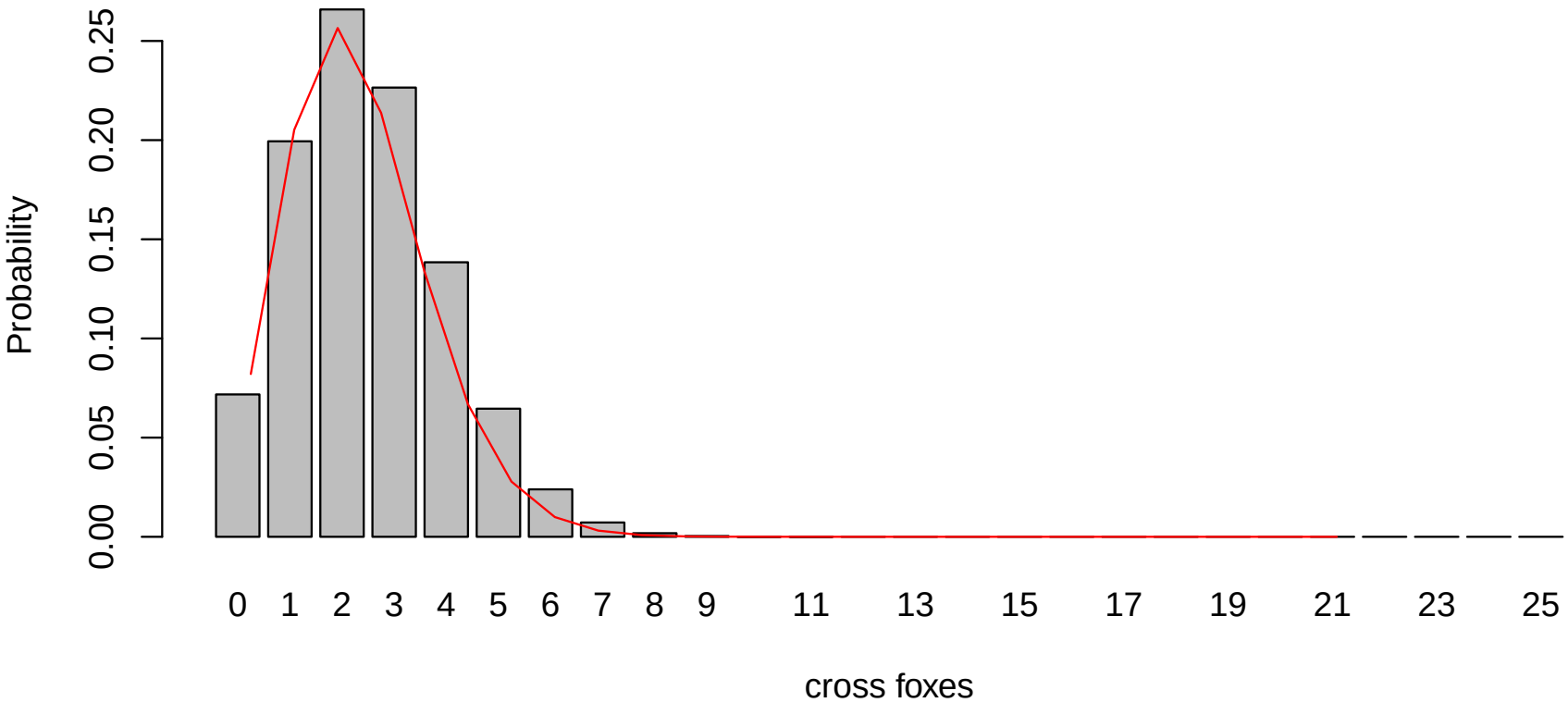
- Discrete, defined for $0 \leq x \leq N$
- $f(x) = \binom{N}{x} p^x (1-p)^{N-x}$
- $\mu = Np$, $\sigma^2 = Np(1-p)$

Approximately normal for large N, intermediate p. Approximately Poisson for large N, small p.

Binomial example

Suppose 10% of red foxes have the cross fox color variation. How many cross foxes would we expect in a sample of size N?

```
N <- 25
p <- 0.1
barplot(dbinom(0:N, N, p),
        names.arg=0:N, xlab="cross foxes", ylab="Probability")
lines(dpois(0:N, N*p), col="red") # Poisson approximation
```



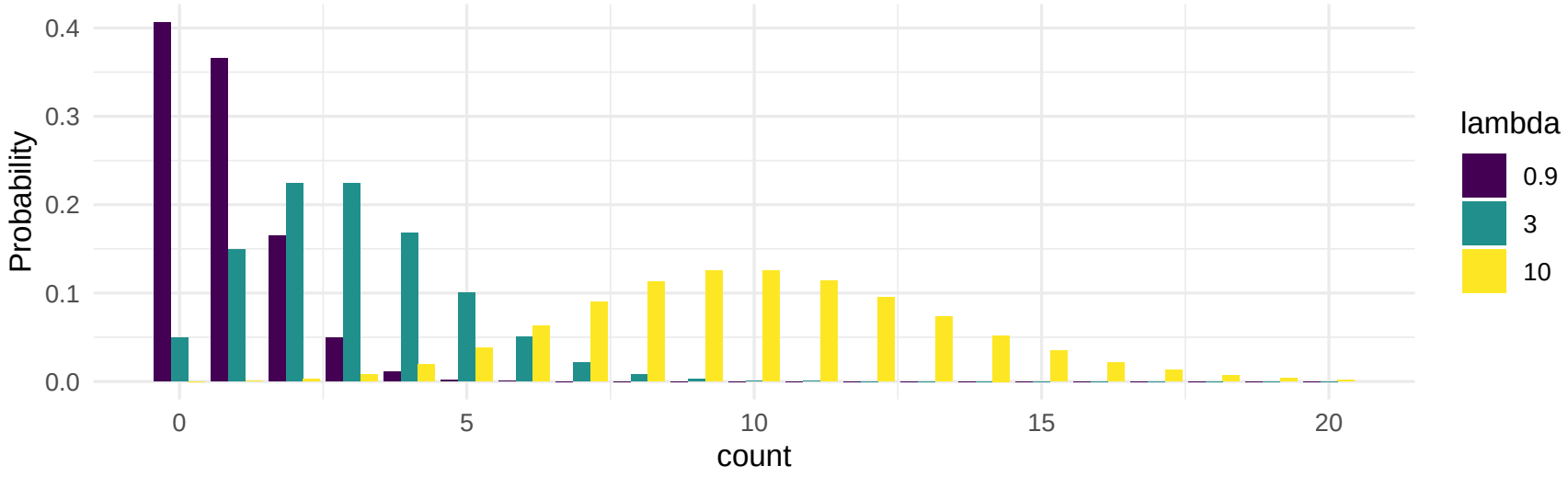
Poisson

X: The count of observations of an evenly distributed event in a given time/space/unit of counting effort.

Parameter is λ, the expected count.

- Discrete, defined for $0 \leq x$
- $f(x) = \dfrac{e^{-\lambda}\lambda^{x}}{x!}$
- $\mu = \lambda$, $\sigma^2 = \lambda$

Right skewed. Approximately normal for large λ.

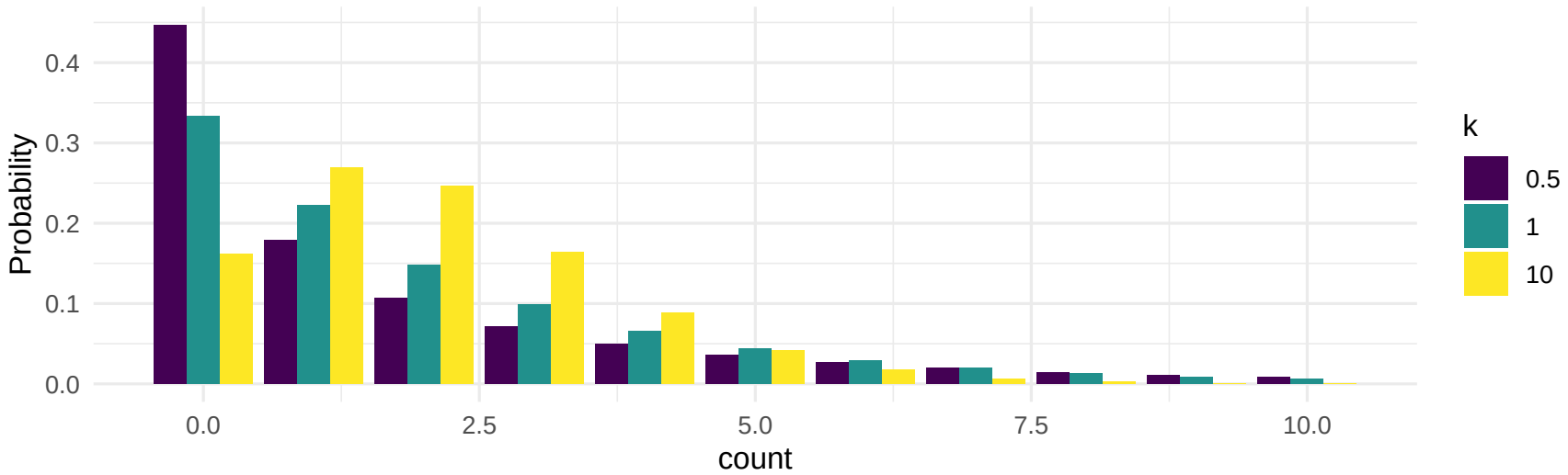


Negative Binomial

X: Similar to Poisson, but the events can be clustered.

Parameters are μ, the expected count, and k, the overdispersion parameter. Smaller k means more clustering.

- Discrete, defined for $0 \leq x$
- $f(x) = \dfrac{\Gamma(k+x)}{\Gamma(k)x!}\left(\dfrac{k}{k+\mu}\right)^k\left(\dfrac{\mu}{k+\mu}\right)^x$
- $\mu = \mu$, $\sigma^2 = \mu + \mu^2/k$



Alternatively X: count of failures before a fixed number of successes.

Uniform

The most boring distribution. X from $U(a, b)$ means all values between the lower limit a and upper limit b are equally likely.

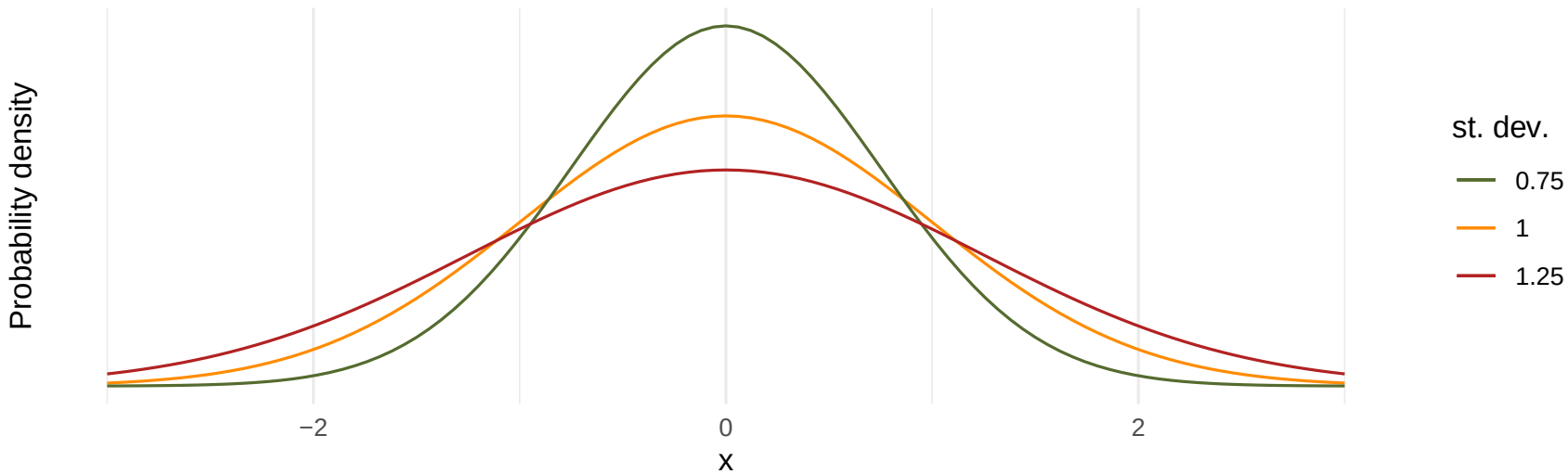
- Generally continuous. Could be discrete depending on context.

Normal

X: The sum of many independent samples.

Parameters mean μ and variance σ^2 (or standard deviation σ). Write $N(\mu, \sigma^2)$.

- Continuous, defined for all real numbers x
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$



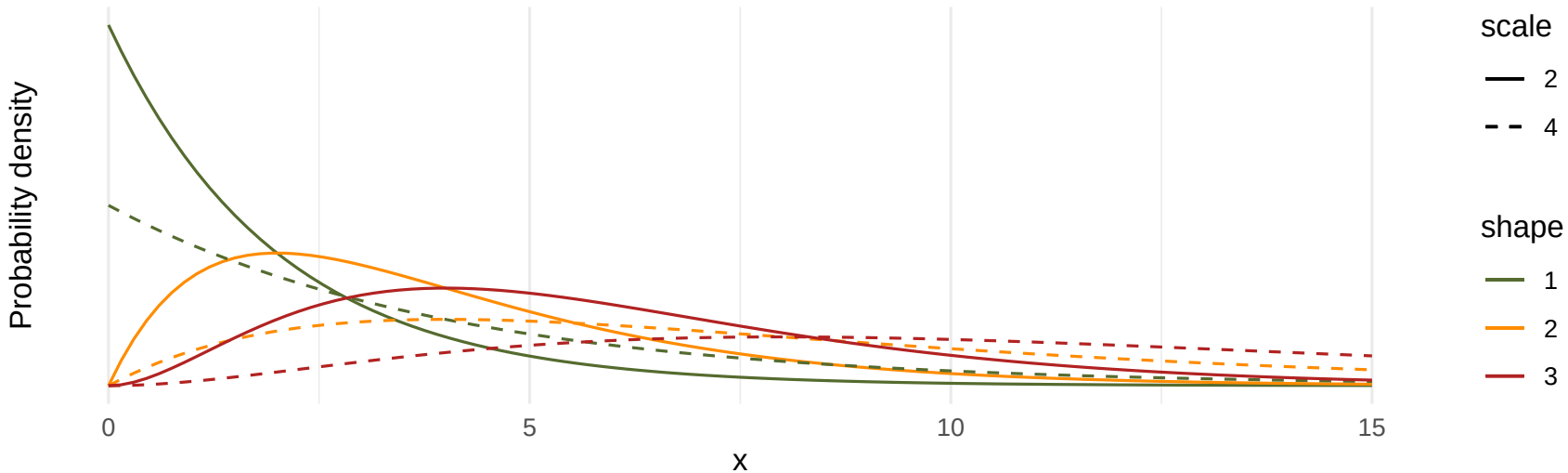
The gold standard for noise. Most classical statistical techniques rely on assuming that the noise is normal.

Gamma

X: The waiting time until a set number of events take place.

Parameters scale s, the length per event, or rate $r = 1/s$, the rate at which events occur, and shape a, the number of events.

- Continuous, $x \geq 0$
- $f(x) = \frac{1}{s^a \Gamma(a)} x^{a-1} e^{-x/s}$
- $\mu = as$, $\sigma^2 = as^2$



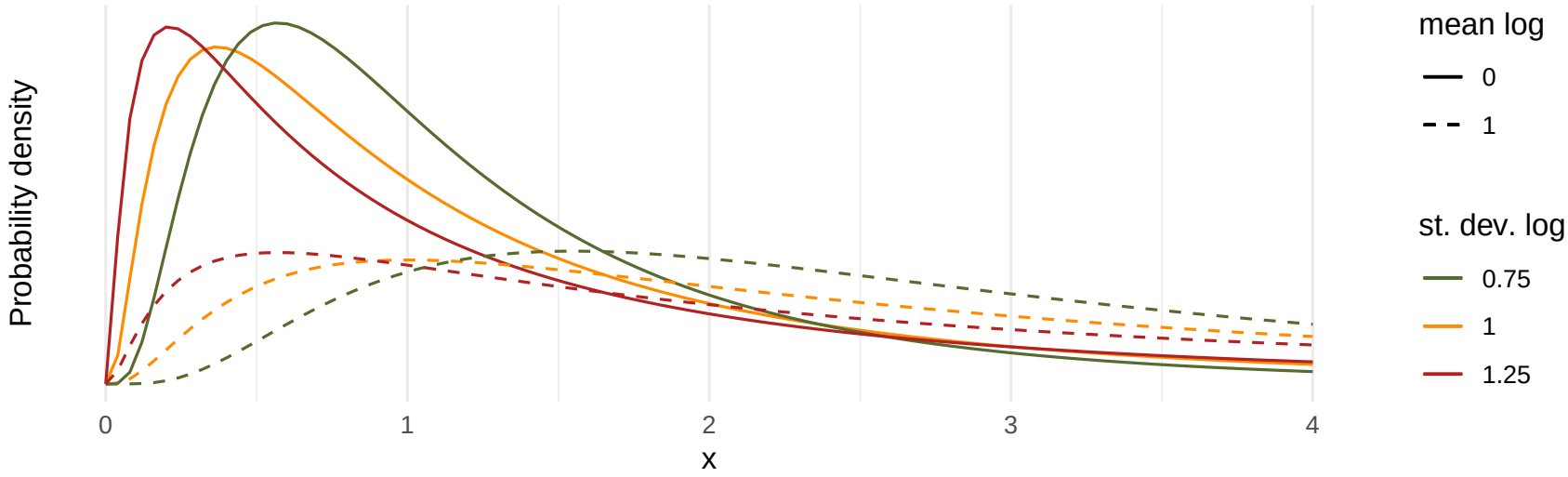
Along with the log-normal, also used for models needing a continuous, right skewed, non-negative distribution without necessarily having a mechanistic reason.

Log-normal

X: The product of many independent samples.

Parameters mean of the log μ and standard deviation of the log σ.

- Continuous, $x > 0$
- $X \sim \exp(\mu + \sigma Z)$ for $Z \sim N(0, 1)$.



Recall that the mean is $\exp(\mu + \sigma^2/2)$.

Mixtures and compounded distributions

Sometimes it is useful to combine distributions or allow the parameters of a distribution be drawn from another distribution. For example, the effects of unknown or unmeasured variables, can potentially be captured by such a varying parameter.

Combining a finite number of distributions into a single distribution is called a **mixture distribution**. The zero-inflated binomial that we created earlier is an example.

Drawing a parameter of one distribution from a second is called a **compound distribution**. Drawing the rate parameter λ for a Poisson distribution from a Gamma distribution gives a negative binomial distribution.

References

Bolker, Benjamin M. 2008. *Ecological Models and Data in R*. Princeton University Press.