# Modeling with Probability Distributions - Capturing Noise

Zack Treisman

Spring 2021

WESTERN COLORADO UNIVERSITY
MATHEMATICS & COMPUTER SCIENCE DEPARTMENT

# 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 $\epsilon$ 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 discussed 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 probability distribution.

A model where applying a link function to $f$ makes it linear in its parameters is called a **generalized linear model** (GLM).

- e.g. $f(x) = e^{\beta_0 + \beta_1 x}$

Somewhat more general $f$ can be fit with a **generalized additive model** (GAM).

- e.g. splines. local regression

# The glm and related commands

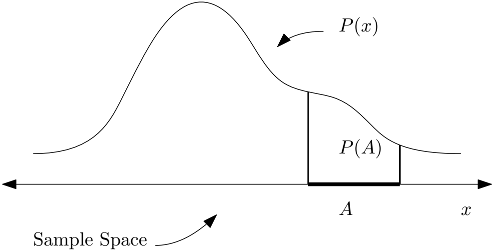
Fitting generalized linear models in R is done using `glm`, generalized additive models with `gam`.

```
?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.

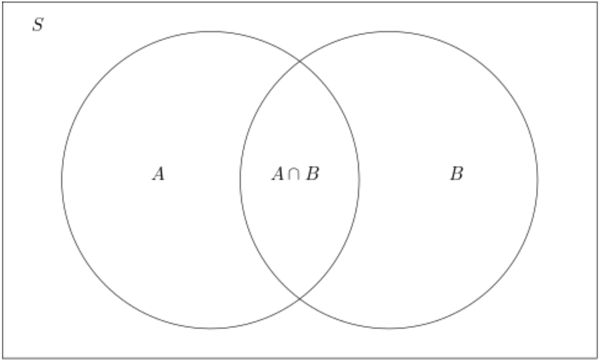
Example: A feeder at my house is visited by various birds and mammals. Each visit is a trial. The set of all critters that visit the feeder is the sample space. A grey jay visiting is an event. By my estimation $P(\text{Grey Jay}) = 0.3$. One of the grey jays that visits I've named June. June visiting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



# More 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**, $P(A \cup B) = P(A) + P(B)$.
- An **unknown 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 $x > 0$,

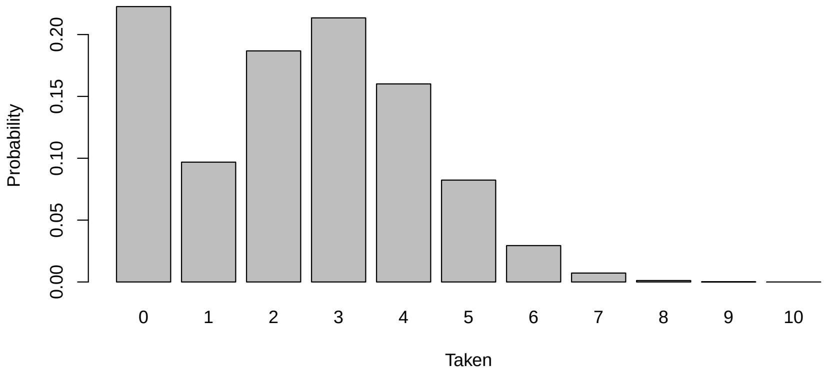
$$P(x \text{ seeds taken}) = \nu \binom{N}{x} p^x (1-p)^{N-x}$$

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 myriad 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")
```



# More 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 $f(x)$ on a sample space $S$ defines an **expectation** operation.

$$E[z] = \sum_{x \in S} z f(x) \text{ or } E[z] = \int_{x \in S} z f(x) dx.$$

- $E[x] = \mu = \bar{x}$ 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.

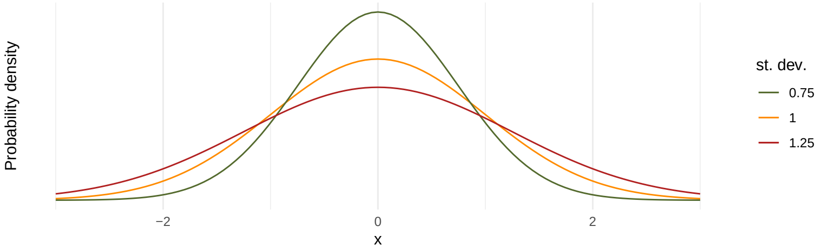
**Method of moments**: To choose a particular distribution from an assumed family, calculate moments for data and use these to compute appropriate parameters.

# Normal

$X$: The sum of many independent samples.

Parameters mean $\mu$ and variance $\sigma^2$ (or standard deviation $\sigma$). 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.

# 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.

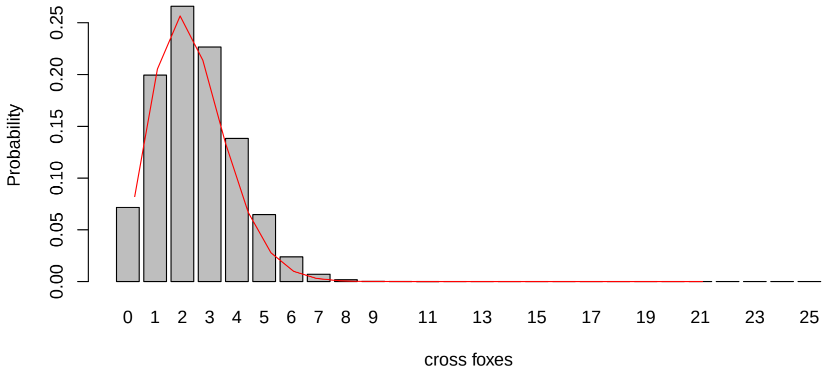
- Discrete, defined for $0 \leq x \leq N$
- $f(x) = \binom{N}{x} p^x (1-p)^{N-x}$
- $\mu = Np$, $\sigma^2 = Np(1-p)$
- Logistic regression estimates the probability of $y =$ success based on predictors $x$: `glm(p~x, family=binomial)`
  - Default link is `logit`.

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
```



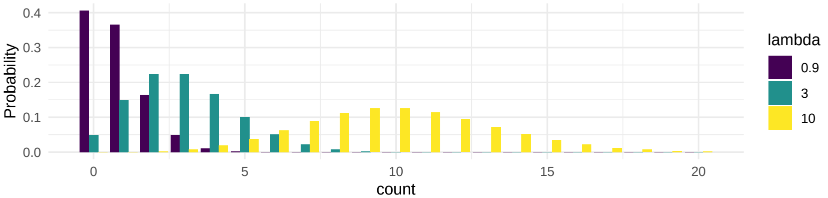
# Poisson

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

Parameter is $\lambda$, the expected count.

- Discrete, defined for $0 \leq x$
- $f(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$
- $\mu = \lambda$, $\sigma^2 = \lambda$
- Poisson regression estimates a count $y$ based on predictors $x$: `glm(y~x, family=poisson)`
  - Default link is `log`.

Right skewed. Approximately normal for large $\lambda$.

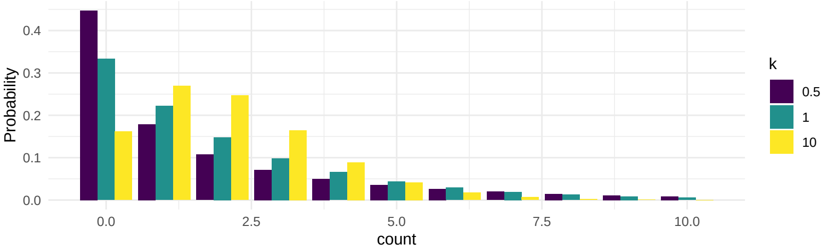


# Negative Binomial

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

Parameters are $\mu$, the expected count, and $k$, the overdispersion parameter. Smaller $k$ means more clustering.

- Discrete, defined for $0 \leq x$
- $f(x) = \dfrac{\Gamma(k+x)}{\Gamma(k)x!} \left(\dfrac{k}{k+\mu}\right)^k \left(\dfrac{\mu}{k+\mu}\right)^x$
- $\mu = \mu$, $\sigma^2 = \mu + \mu^2/k$
- Negative binomial regression: `glm.nb(y~x)`
  - Default link is `log`.
  - `glm.nb` is in `MASS`. Other options exist.

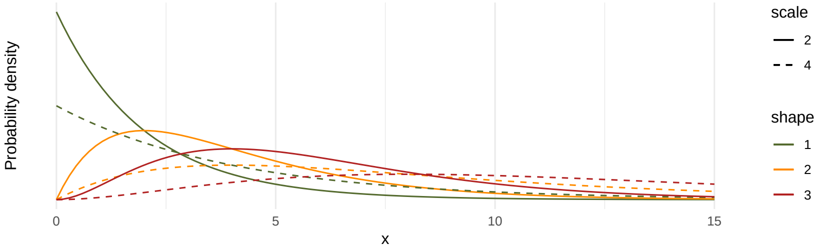


# Gamma

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



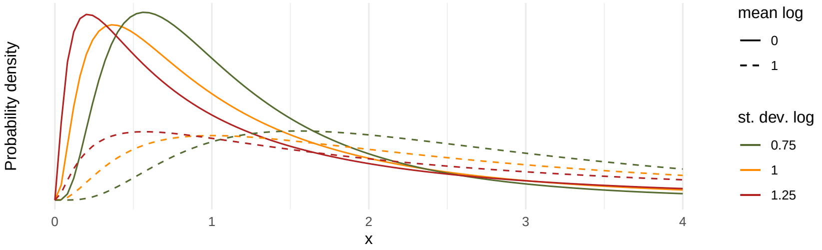

Along with the log-normal, also used for models needing a continuous, right skewed, non-negative distribution without necessarily having a mechanistic reason.

# Log-normal

$X$: The product of many independent samples.

Parameters mean of the log $\mu$ and standard deviation of the log $\sigma$.

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