Probability Concepts, Notation, and Distributions

Bayesian Models for Ecologists

Becky Tang

June 3, 2024

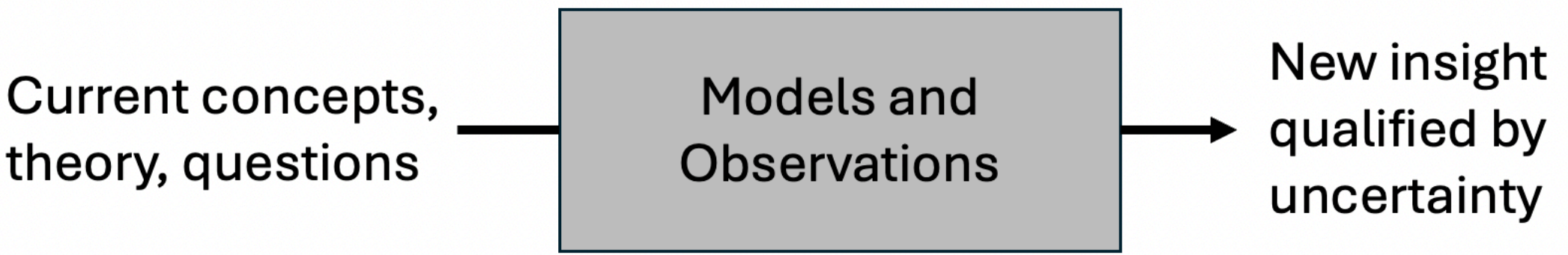
Note!

- Slides adapted from Bailey Fosdick, Chris Che-Castaldo, Mary B. Collins, and N. Thompson Hobbs
- This is a crash course in basic probability concepts. Please stop me at any time if you have questions!

Probability Concepts and Notation

Motivation

A general approach to scientific research:



What you must know and why

Concept	Why learn?
Conditional probability	It is the foundation for Bayes' Theorem and all inferences we will make.
Law of Total Probability	Basis for the denominator of Bayes' Theorem $[y]$.
Factoring joint distributions	Procedure used to build models.
Independence	Allows us to simplify full factored joint distributions.
Probability distributions	Our toolbox for fitting models to data and representing uncertainty.
Moments	A way we can summarize distributions.
Marginal distributions	Bayesian inference is based on marginal distributions of unobserved quantities.
Moment matching	Allows us to embed the predictions of models into any statistical distribution.

More motivation

- **Why Bayes?** Bayesian analysis is the *only* branch of statistics that treats all unobserved quantities as random variables. We seek to understand the characteristic of the probability distributions governing the behavior of these random variables.
- **Why models?** A model of the data describes our ideas of how the data arise.
  - Different types of data (e.g. real numbers, counts, proportions, ordinal categories) and knowledge about the underlying process will require different types of models.
  - Deterministic vs. probabilistic

Topics we will cover now

- Random variables
- Discrete vs. continuous distributions
- Moments (mean and variance)
- Cumulative density function
- Quantile functions
- Working with probability distributions in R
- Monte Carlo integration

Functions and variables

Consider the following function for the equation of a line:

$$y = mx + b$$

- Note $y = f(x)$ is a function of x. For fixed values of m and b, each value x gets mapped to a single $f(x)$.
  - x may be considered the variable of interest.

Random variables

- You have previously learned that the *sample space* is the set of all possible outcomes of a random process.
- A **random variable (R.V. or r.v.)** is a function from a particular sample space to the real numbers
  - Random variables represent the outcome of a random process using values on the real line. They are a mathematical formalization of a quantity or object which depends on random events (Wikipedia)
- Previously you discussed the probability of **events**, which are associated with specific values of a r.v.
- Note: we typically denote random variables by Roman letters (e.g. X or Y for data) and Greek letters (e.g. θ and α for population parameters)

Examples

Random process	Flip a fair coin one time	Flight time of bird recorded
Possible outcomes	Heads or Tails	Any amount of time
Random variable	$X =$ number of Heads	$Y =$ time of flight
Support	$S_X = \{0, 1\}$	$S_Y : y > 0$
Possible probabilities of interest	$\Pr(\text{Heads}) = \Pr(X = 1)$	$\Pr(\text{at least 2 hours}) = \Pr(Y > 2)$

Evaluating these probabilities of interest will require adding some probability concepts to our toolkit!

A note on notation

The notation I use during the board work will feature both:

1. Notation you might find in a probability textbook (functional notation), and
2. Notation used in Hobbs and Hooten (bracket notation)

The notation in (2) is simpler and can be much more intuitive and easy to read, and serves our purposes well when we get to building models.

The notation in (1) is a bit more technical (we explicitly distinguish between random variables and outcomes), which means we can be a bit more specific.

Board work

Probability Distributions and R

Common distributions

- Some distributions are so “common” that we give them specific names. Look to the **distribution sheet**!
- The shape/behavior of each distribution is determined by a specific set of **parameters**.
  - E.g. $X \sim \text{Bernoulli}(\theta) \Rightarrow P(X = x) = [x] = \theta^x(1-\theta)^x$ for $x \in \{0, 1\}$. All Bernoulli distributions have this functional form. But $\theta = 0.25$ versus $\theta = 0.5$ will lead to different outputs for a given value x.
- The parameters specify the exact shape of the distribution, and therefore, affect the *moments* of the distribution.

Flexibility in analysis

It's up to you to determine/justify if your random variable's behavior is characterized by one of these common distributions.

- How? Narrow down choices via continuous vs discrete, looking at support, think about what the random process.
- Once you've selected a distribution, don't forget to identify the parameters.

Probability model	**Support for random variable**
Normal	real numbers
Multivariate normal	vector of real numbers
Lognormal	positive real numbers
Gamma	positive real numbers
Beta	real numbers on [0,1] or (0,1)
Bernoulli	0 or 1
Binomial	counts in two categories with an upper bound
Poisson	counts
Multinomial	counts in more than two categories
Negative binomial	counts
Dirichlet	proportions in two or more categories
t	real numbers
Cauchy	real numbers

Working with distributions in R

Distribution	Functions			
Beta	pbeta	qbeta	dbeta	rbeta
Binomial	pbinom	qbinom	dbinom	rbinom
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy
Chi-Square	pchisq	qchisq	dchisq	rchisq
Exponential	pexp	qexp	dexp	rexp
F	pf	qf	df	rf
Gamma	pgamma	qgamma	dgamma	rgamma
Geometric	pgeom	qgeom	dgeom	rgeom
Hypergeometric	phyper	qhyper	dhyper	rhyper
Logistic	plogis	qlogis	dlogis	rlogis
Log Normal	plnorm	qlnorm	dlnorm	rlnorm
Negative Binomial	pnbinom	qnbinom	dnbinom	rnbinom
Normal	pnorm	qnorm	dnorm	rnorm
Poisson	ppois	qpois	dpois	rpois
Student t	pt	qt	dt	rt
Studentized Range	ptukey	qtukey	dtukey	rtukey
Uniform	punif	qunif	dunif	runif
Weibull	pweibull	qweibull	dweibull	rweibull
Multinomial	pmultinom	qmultinom	dmultinom	rmultinom

https://www.stat.umn.edu/geyer/old/5101/rlook.html

- `p***(x,...)`: CDF function; returns cumulative probability $\leq$ `x`
- `q***(p,...)`: quantile function; returns value x such that probability of $\leq x$ is `p` (or at least `p` if discrete)
- `d***(x,...)`: PDF/PMF: returns density at or probability of `x`
- `r***(n,...)`: random generator; returns `n` random values from the distribution
- Must specify the specific parameter values in `...` Type `?function` into Console for details.

Distributions in R: example

Let X represent the observed temperature of a certain stream. Someone tells you that $X \sim \text{Normal}(50, 25)$.

- Take a look at the distribution sheet. What do you think the values 50 and 25 represent?
- What R code would you type to evaluate $\Pr(X < 45)$?
- What R code would you type to evaluate $\Pr(40 < X < 45)$?

Distributions in R: example (cont.)

Let X represent the observed temperature of a certain stream. Someone tells you that $X \sim \text{Normal}(50, 25)$.

- What R code would you type to evaluate $\Pr(X < 45)$?

```
1  pnorm(q = 45, mean = 50, sd = sqrt(25))
```

```
[1] 0.1586553
```

- What R code would you type to evaluate $\Pr(40 < X < 45)$?

```
1  pnorm(q = 45, mean = 50, sd = sqrt(25)) - pnorm(q = 40, mean = 50, sd = sqrt(25))
```

```
[1] 0.1359051
```

Monte Carlo Integration

Basic idea: we can estimate any property of a distribution using a large number of random samples from the distribution.

Question: what is each line of code doing?

```
n <- 1000
samps <- rpois(n, lambda = 4)
samps[1:5]
mean(samps)
var(samps)
sum(samps > 7)/n
```

Monte Carlo Integration (cont.)

```r
n <- 1000
samps <- rpois(n, lambda = 4)   # simulate n = 1000 Poisson(4) random variables
samps[1:5]                      # take a look at the first five samples
```

```
[1] 3 3 4 7 2
```

```r
mean(samps)                     # estimate the mean of Poisson(4) dist.
```

```
[1] 4.018
```

```r
var(samps)                      # estimate the variance of Poisson(4) dist
```

```
[1] 4.057734
```

```r
sum(samps > 7)/n                # estimate Pr(X > 7) where X ~ Poisson(4)
```

```
[1] 0.054
```

```r
1 - ppois(q = 7, lambda = 4)    # compare to true probability
```

```
[1] 0.05113362
```

Your turn!

Barn swallows form pair bonds (male/female pairings) in the spring before mating season. Each male/female pair shares a nest and cares for the offspring of the female. Often a number of the female's offspring were sired by males other than her mate. Suppose previous literature suggests the probability that an offspring's father is the female's bond mate is 0.8.

Task: Calculate the probability that a female with five offspring will have less than or equal to one offspring whose father is the female's mate.

Probability problem solving recipe:

1. Define the random variable(s), and write the desired probability or quantity in terms of random variable(s)
2. Identify the distribution of the r.v.'s and any relevant parameters
3. Draw a picture (if applicable)
4. Calculate the desired probability or quantity (possibly using R)