Lecture 5: Parameterized probability models, inference, samples, statistics

Jason Mezey
jgm45@cornell.edu
Feb. 9, 2017 (Th) 8:40-9:55
Announcements

• All Piazza and CMS sign-up issues should be resolved (please let us know if you are still having an issue)

• My office hours will STILL be 3-5PM each Tues. 101 Biotech in Ithaca and Large Conference room Dept. Genetic Med. (13th floor Weill-Greenberg) in NYC

• Homework #2 will be posted later this morning

• Computer lab today 5-6PM Ithaca Mann Library and 3-4PM NYC Large Conference Genetic Med. (please go the long way around!!)
Summary of lecture 5

• Last lecture, we discussed functionals take both random variables AND probability models as input to produce useful “summary” (and more!) outputs useful for random variables / probability models in general (expectations, variances)

• In this lecture, we will continue our discussion of such functionals and introduce specific probability models with the concept of parameterized probability distributions, where we also begin our discussion of inference, the concept of a sample (and i.i.d.), and the concept of a statistic
Conceptual Overview

System

Question

Sample

Prob. Models

Inference

Statistics

Assumptions
Review

- **Experiment** - a manipulation or measurement of a system that produces an outcome we can observe
- **Sample Space** ($\Omega$) - set comprising all possible outcomes associated with an experiment
- **Sigma Algebra** or **Sigma Field** ($\mathcal{F}$) - a collection of events (subsets) of the sample space of interest
- **Probability Measure** (=Function) - maps a Sigma Algebra of a sample to a subset of the reals
- **Random Vector** (Variable) - (measurable) function on a sample space
- **Probability Distribution Function** / **Cumulative Distribution Function** (pdf / cdf) - probability mass (discrete) or density (continuous) functions that describes the probability distribution of a discrete OR continuous random variable
- **Expectations and Variances** (Covariances) - functionals that input a random vector and probability measure then output an informative scalar, vector, or matrix

Axioms of Probability

1. Probability functions on sample spaces are probability functions. These rules are called the axiom of probability functions.
2. To be useful, we need some rules for how probability functions are defined (that is, not all those that map sets of numbers are probability functions). We are going to define a function from class.
3. A probability function $P$ is a function, which we could have written $\mathcal{F} \rightarrow [0;1]$ (note that an ability function or mathematical operator that takes an input and produces an output.

Expectations and Variances (Covariances)

- A measure of expectations and variances (covariances) is often introduced to us as the intuitive definition of a function: $\mathcal{F} \rightarrow [0;1]$. For example, we can have the function $f$ that maps a Sigma Algebra of a sample to a subset of the reals.
- $\mathcal{F}$ is a rule that we assume. There is some variation in how these are presented, but we will present them as three axioms:

$$\text{Pr} = 0.$$
$$\text{Pr}(\mathcal{F}) = 1.$$
$$\text{Pr}(\mathcal{S}) = 1.$$

Before we consider the specifics of how we define a probability function, let’s consider the intuitive definition of a function: $\text{Pr}(A) = \mathcal{F}(A)$, if $A$ is a (disjoint) for each $S_i$. For example, we can have the function $f$ such that $f(A) = \mathcal{F}(A)$, where $A$ is a set of numbers.
The connection among these concepts

\[ X = x \quad Pr(X) \]

Random Variable

\[ X(\omega), \omega \in \Omega \]

Pr(\mathcal{F})

Experiment

\( \Omega \)

(Sample Space)

\( \mathcal{F} \)

(Sigma Algebra)
Review: expectations and variances

- As a review we have introduced the fundamental functions of random variables / vectors: **expectations** and **variances**
- These are **functionals** - map a function to a scalar (number)
- The expected value of a discrete and continuous random variables:

  \[
  \text{EX} = \sum_{i=\text{min}(X)}^{\text{max}(X)} (X = i) \Pr(X = i) \quad \text{EX} = \int_{-\infty}^{+\infty} X f_X(x) \, dx
  \]

- The variance of discrete and continuous random variables:

  \[
  \text{Var}(X) = V(X) = \sum_{i=\text{min}(X)}^{\text{max}(X)} (X = i)^2 \Pr(X = i) \quad \text{Var}(X) = VX = \int_{-\infty}^{+\infty} (X - \text{EX})^2 f_X(x) \, dx
  \]
Review: random vectors
expectations and variances

- Recall that a generalization of a random variable is a random vector, e.g.
  \[ X = [X_1, X_2] \quad P_{X_1,X_2}(x_1, x_2) \text{ or } f_{X_1,X_2}(x_1, x_2) \]

- The expectation (a function of a random vector and its distribution!) is a vector with the expected value of each element of the random vector, e.g.
  \[ \mathbb{E}X = [\mathbb{E}X_1, \mathbb{E}X_2] \]

- Variances also result in variances of each element (and additional terms... see next slide!!)
Random vectors: covariances

- Variances (again a function!) of a random vector are similar producing variances for each element, but they also produce **covariances**, which relate the relationships between random variables of a random vector!!

\[
\text{Cov}(X_1, X_2) = \sum_{i=\text{min}(X_1)}^{\text{max}(X_1)} \sum_{j=\text{min}(X_2)}^{\text{max}(X_2)} ((X_1 = i) - EX_1)((X_2 = j) - EX_2)P_{X_1,X_2}(x_1, x_2)
\]

\[
\text{Cov}(X_1, X_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (X_1 - EX_1)(X_2 - EX_2)f_{X_1,X_2}(x_1, x_2)dx_1dx_2
\]

- Intuitively, we can interpret a positive covariance as indicating “big values of \(X_1\) tend to occur with big values of \(X_2\) AND small values of \(X_1\) tend to occur with small values of \(X_2\)”

- Negative covariance is the opposite (e.g. “big \(X_1\) with small \(X_2\) AND small \(X_1\) with big \(X_2\)”)

- Zero covariance indicates no relationship between big and small values of \(X_1\) and \(X_2\)
An illustrative example

- For example, consider our experiment where we have measured “height” and “IQ” / bivariate normal probability model / identity random variable:
Notes about covariances

- Covariance and independence, while related, are NOT synonymous (!!), although if random variables are independent, then their covariance is zero (but necessarily vice versa!)

- Covariances are symmetric: \( \text{Cov}(X_1, X_2) = \text{Cov}(X_2, X_1) \)

- Other equivalent (and often used) formulations of covariances:

\[
\text{Cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)]
\]

\[
\text{Cov}(X_1, X_2) = \mathbb{E}(X_1X_2) - \mathbb{E}X_1\mathbb{E}X_2
\]

- From these formulas, it follows that the covariance of a random variable and itself is the variance:

\[
\text{Cov}(X_1, X_1) = \mathbb{E}(X_1X_1) - \mathbb{E}X_1\mathbb{E}X_1 = \mathbb{E}(X_1^2) - (\mathbb{E}X_1)^2 = \text{Var}(X_1)
\]
Covariance matrices

- Note that we have defined the “output” of applying an expectation function to a random vector but we have not yet defined the analogous output for applying a variance function to a random vector.

- The output is a covariance matrix, which is a square, symmetric matrix with variances on the diagonal and covariances on the off-diagonals.

- For example, for two and three random variables:

\[
\text{Var}(X) = \begin{bmatrix}
\text{Var}X_1 & \text{Cov}(X_1, X_2) \\
\text{Cov}(X_1, X_2) & \text{Var}X_2
\end{bmatrix}
\]

\[
\text{Var}(X) = \begin{bmatrix}
\text{Var}X_1 & \text{Cov}(X_1, X_2) & \text{Cov}(X_1, X_3) \\
\text{Cov}(X_1, X_2) & \text{Var}X_2 & \text{Cov}(X_2, X_3) \\
\text{Cov}(X_1, X_3) & \text{Cov}(X_2, X_3) & \text{Var}(X_3)
\end{bmatrix}
\]

- Note that not all square, symmetric matrices are covariance matrices (!!), technically they must be positive (semi)-definite to be a covariance matrix.
Covariances and correlations

- Since the magnitude of covariances depends on the variances of $X_1$ and $X_2$, we often would like to scale these such that “1” indicates maximum “big with big / small with small” and “-1” indicates maximum “big with small” (and zero still indicates no relationship).

- A correlation captures this relationship:

$$\text{Corr}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}}$$

- Where we can similarly calculate a correlation matrix, e.g. for three random variables:

$$\text{Corr}(\mathbf{X}) = \begin{bmatrix}
1 & \text{Corr}(X_1, X_2) & \text{Corr}(X_1, X_3) \\
\text{Corr}(X_1, X_2) & 1 & \text{Corr}(X_2, X_3) \\
\text{Corr}(X_1, X_3) & \text{Corr}(X_2, X_3) & 1
\end{bmatrix}$$
Algebra of expectations and variances

- If we consider a function (e.g., a transformation) on \( X \) (a function on the random variable but not on the probabilities directly!), recall that this can result in a different probability distribution for \( Y \) and therefore different expectations, variances, etc. for \( Y \) as well.

- We will consider two types of functions on random variables and the result on expectation and variances: sums \( Y = X_1 + X_2 + \ldots \) and \( Y = a + bX_1 \) where \( a \) and \( b \) are constants.

- For example, for sums, \( Y = X_1 + X_2 \) we have the following relationships:

\[
\begin{align*}
E(Y) &= E(X_1 + X_2) = EX_1 + EX_2 \\
\text{Var}(Y) &= \text{Var}(X_1 + X_2) = \text{Var}X_1 + \text{Var}X_2 + 2\text{Cov}(X_1, X_2)
\end{align*}
\]

- As another example, for \( Y = X_1 + X_2 + X_3 \) we have:

\[
\begin{align*}
E(Y) &= E(X_1 + X_2 + X_3) = EX_1 + EX_2 + EX_3 \\
\text{Var}(Y) &= \text{Var}(X_1 + X_2 + X_3) = \text{Var}X_1 + \text{Var}X_2 + \text{Var}X_3 + 2\text{Cov}(X_1, X_2) + 2\text{Cov}(X_1, X_3) + 2\text{Cov}(X_2, X_3)
\end{align*}
\]
Algebra of expectations and variances

- For the function $Y = a + bX$ we obtain the following relationships:

$$\mathbb{E}Y = a + b\mathbb{E}X$$

$$\text{Var}(Y) = b^2\text{Var}(X)$$

- Finally, note that if we were to take the covariance (or correlation) of two random variables $Y_1$ and $Y_2$ with the relationship:

$$Y_1 = a_1 + b_1X_1, \ Y_2 = a_2 + b_2X_2$$

$$\text{Cov}(Y_1, Y_2) = b_1b_2\text{Cov}(X_1, X_2)$$

$$\text{Corr}(Y_1, Y_2) = \text{Corr}(X_1, X_2)$$
Probability models I

- We have defined $\Pr(X)$, a probability model on a random variable, which technically we produce by defining $\Pr(\mathcal{F})$ and $X(\Omega)$

- So far, we have considered such probability models without defining them explicitly (except for a illustrative few examples)

- To define an explicit model for a given system / experiment we are going to assume that there is a “true” probability model, that is a consequence of the experiment that produces sample outcomes

- We place “true” in quotes since the defining a single true probability model for a given case could only really be accomplished if we knew every single detail about the system and experiment (would a probability model be useful in this case?)

- In practice, we therefore assume that the true probability distribution is within a restricted family of probability distributions, where we are satisfied if the true probability distribution in the family describes the results of our experiment pretty well / seems reasonable given our assumptions
In short, we therefore start a statistical investigation *assuming* that there is a single true probability model that correctly describes the possible experiment outcomes given the uncertainty in our system.

In general, the starting point of a statistical investigation is to make *assumptions* about the form of this probability model.

More specifically, a convenient assumption is to assume our true probability model is specific model in a family of distributions that can be described with a compact equation.

This is often done by defining equations indexed by *parameters*.
Probability models III

- **Parameter** - a constant(s) \( \theta \) which indexes a probability model belonging to a family of models \( \Theta \) such that \( \theta \in \Theta \)
- Each value of the parameter (or combination of values if there is more than one parameter) defines a different probability model: \( \Pr(X) \)
- We assume one such parameter value(s) is the true model
- The advantage of this approach is this has reduced the problem of using the sample to answer a broad question to the problem of using a sample to make an educated guess at the value of the parameter(s)
- Remember that the foundation of such an approach is still an assumption about the properties of the sample outcomes, the experiment, and the system of interest (!!!)
Discrete parameterized examples

- Consider the probability model for the one coin flip experiment / number of tails.

- This is the Bernoulli distribution with parameter $\theta = p$ (what does $p$ represent!?) where $\Theta = [0, 1]$

- We can write this $X \sim \text{Bern}(p)$ and this family of probability models has the following form:

$$Pr(X = x|p) = P_X(x|p) = p^x(1-p)^{1-x}$$

- For the experiment of $n$ coin flips / number of tails, we can assume the Binomial distribution $X \sim \text{Bin}(n, p)$:

$$Pr(X = x|n, p) = P_X(x|n, p) = \binom{n}{x}p^x(1-p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad n! = n \times (n-1) \times (n-2) \times ... \times 1$$

- There are many other discrete examples: hypergeometric, Poisson, etc.
Continuous parameterized examples

- Consider the measure heights experiment (reals as approximation to the sample space) / identity random variable.

- For this example we can use the family of normal distributions that are parameterized by \( \theta = [\mu, \sigma^2] \) (what do these parameters represent!?) with the following possible values: \( \Theta_\mu = (-\infty, \infty) \), \( \Theta_{\sigma^2} = [0, \infty) \).

- We often write this as \( X \sim N(\mu, \sigma^2) \) and the equation has the following form:

\[
Pr(X = x|\mu, \sigma^2) = f_X(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

- There are many other continuous examples: uniform, exponential, etc.
Example for random vectors

- Since random vectors are the generalization of r.v.'s, we similarly can define parameterized probability models for random vectors

- As an example, if we consider an experiment where we measure “height” and “weight” and we take the 2-D reals as the approximate sample space (vector identity function), we could assume the bivariate normal family of probability models:

\[
f_X(x_1, x_2 | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho}} \exp \left[ -\frac{1}{2(1 - \rho^2)} \left( \frac{(x_1 - \mu_1)^2}{2\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1 \sigma_2} + \frac{(x_2 - \mu_1)^2}{2\sigma_2^2} \right) \right]
\]
Introduction to inference I

- Recall that our eventual goal is to use a sample (generated by an experiment) to provide an answer to a question (about a system).

- So far, we have set up the mathematical foundation that we need to accomplish this goal in a probability / statistics setting (although note we have not yet provided formalism for a sample!!)

- Specifically, we have defined formal components of our framework and made assumptions that have reduced the scope of the problem.

- With these components and assumptions in place, we are almost ready to perform inference, which will accomplish our goal.
Introduction to inference II

- **Inference** - the process of reaching a conclusion about the true probability distribution (from an assumed family probability distributions, indexed by the value of parameter(s) ) on the basis of a sample

- There are two major types of inference we will consider in this course: *estimation* and *hypothesis testing*

- Before we get to these specific forms of inference, we need to formally define: *samples*, *sample probability distributions* (or sampling distributions), *statistics*, *statistic probability distributions* (or statistic sampling distributions)
That’s it for today

- Next lecture, we will introduce samples, statistics, estimators, and begin our discussion of likelihood and maximum likelihood estimators