Lecture 2: Probing the power of probability density functions

Announcements: auditors, please make sure that I have your UCSD e-mail.
Copyrighted material on Canvas

Recap

Lecture 1 provided a broad overview of the course and key statistical concepts. This lecture will review core concepts in probability. Those who took SIOC 221A might have seen some aspects of this, and I've inserted the relevant notes into a side bar at the end for reference.

Uncertainties and data We started class with an example of fitted parameters from a published paper. The parameters differed by a factor of 2 or so, and we wondered if the differences were statistically significant. That led to some consideration of uncertainties. Here are some ways that data values can be plotted:

- Data points
- Data points with standard error bars (e.g. $\pm 2\sigma$, where σ is one standard deviation). This works well when data are normally distributed.
- Data points with box and whiskers presentation indicating 25th percentile, median, 75th percentile, and extrema.
- Probability density functions (pdfs). Hard to plot, but show full distribution of values.

To get started, we'll talk about pdfs and cumulative distribution functions.

Cumulative distributions and probability density functions

The complete description of a single random variable is its **distribution function** (or **cumulative distribution function**) D or, equivalently, its **probability density function** (pdf) F. These are defined by

$$D_x(r) \;=\; \text{Fraction of occurrences with } \; x < r, \tag{1}$$

where the fraction of occurrences is a probability. The notation might seem a little unusual: the random variable x is a subscript, and the distribution function depends on the deterministic variable r. Some properties of the distribution function are:

$$\begin{aligned} D_x(-\infty) &= 0, & (2)\\ D_x(\infty) &= \int_{-\infty}^{\infty} dr\; F(r) = 1, & (3)\\ D_x(r) &\leq D_x(s) \text{if} r \leq s. & (4) \end{aligned}$$

The **probability density function** (pdf) is

$$F_x(r) \;=\; \frac{d}{dr} D_x(r) \quad \text{so that} \quad F_x(r)\, dr \;=\; \text{Fraction of occurrences with } \; r < x < r + dr. \tag{5}$$

Some properties of the pdf are

$$\begin{aligned} F_x(r) &\geq 0 & (6)\\ \int_{-\infty}^{\infty} F_x(r)\, dr &= 1 & (7)\\ D_x(r) &= \int_{-\infty}^{r} F_x(r)\, ds & (8) \end{aligned}$$

Since the pdf is continuous, it is the limit of a histogram describing the number of occurrences in each of several "bins" of x, normalized so that the area under the curve is 1.

In the deterministic case that every realization produces the constant $x = A$, then $F_x(r) = \delta(r - A)$ and D would be a Heaviside unit step function with its edge at $r = A$. A handy representation of the probability density function is

$$F_x(r) = \langle \delta(r - x) \rangle. \tag{9}$$

To see why this is so, let N be an effectively infinite number of realizations of x and M be the number of realizations with $x < r$. Then, by definition,

$$D_x(r) \; = \; \frac{M}{N} \; = \; \frac{1}{N} \sum_{n=1}^{N} \int_{-\infty}^{r} dy \; \delta(y - x_n) \; = \; \int_{-\infty}^{r} dy \langle \delta(y - x) \rangle \tag{10}$$

from which (9) follows by differentiation with respect to r. The delta function is introduced into the third term to generate 1 when $x_n > r$ and 0 when it is not. The pdf is the complete description of x (taken alone) because it immediately gives the average of any function, say G, of the random variable. This is easy to see using (9):

$$\langle G(x) \rangle = \langle \int dr \, G(r) \delta(r - x) \rangle = \int dr G(r) \langle \delta(r - x) \rangle = \int dr \, G(r) \, F_x(r) \tag{11}$$

Let's think about empirical pdfs. We can start with the distribution of random data. Suppose we use a standard random number generator in Matlab:

```
% generate a histogram of random numbers
hist(rand(10000,1),25)
```

This will produce a histogram of uniform data, with 25 bins. If we want the area under the curve to be 1, we need to normalize by the total number of points and by bin width, or give Matlab instructions to plot as a pdf rather than a raw histogram:

```
% generate a histogram of random numbers
histogram(rand(10000,1),'Normalization','pdf')
```

Geophysical variables are probably usually closer to Gaussian than uniform in distribution. If we want to look at a uniform distribution we can use "randn" instead of "rand":

```
% generate a histogram of random numbers
histogram(randn(10000,1),'Normalization','pdf')
```

This produces random numbers with a Gaussian distribution. Recall that a Gaussian distribution can be written:

$$F(r) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(r - \mu)^2}{2\sigma^2} \right], \tag{12}$$

where μ is the mean, and σ is the standard deviation. As a refresher, we like Gaussian distributions, because they are easy to calculate, and have well defined properties. We know that 68% of measurements will be within $\pm\sigma$ of the mean, and 95% of measurements will be within $\pm 2\sigma$ of the mean.

Of course, as we noted in class, when we look at physical variables, many are decidedly non-Gaussian, and that will lead us to ask how we can characterize non-Gaussian variables.

Side bar: Statistical terminology Here's a quick review of some definitions. A **random variable** is a variable whose value is determined by a **random process**. If a process produces values that are not perfectly predictable from what is known, it is a random process. A poorly controlled experiment might be considered a random process. Any function of a random variable is also a random variable. A **realization** of the process produces one random value of the variable. A large collection of realizations produced under statistically identical conditions (the same deterministic parameters) is an **ensemble** of "identically prepared" observations. The act of flipping a coin is a random process. If we let $x = 1$ for "heads" and -1 for "tails", x is a random variable. A single flip produces a realization of x and an afternoon of flipping using the same technique would yield an ensemble of realizations.

A central concept is the **average** or **expected value** of a random variable. The average of x is

$$\langle x \rangle = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} x_n \tag{13}$$

where x_n is the value of x in the n^{th} realization. Throughout these notes the symbol $\langle \cdot \rangle$ will be reserved for the ideal average requiring an infinite number of realization.

It is helpful to think of $\langle \cdot \rangle$ as a linear operator which can be applied to random variables. It is a linear operator because

$$\langle X + Y \rangle = \langle Y + X \rangle = \langle X \rangle + \langle Y \rangle, \quad \langle aX \rangle = \langle Xa \rangle = a\langle X \rangle, \quad \langle XY \rangle = \langle YX \rangle \neq \langle X \rangle \langle Y \rangle \tag{14}$$

when a is a constant.