# EPS 659a – Data Analysis in Geoscience

## Problem Set One

*due Friday, September 10, 2021*

*Problem 1: Probability of Discrete Random Variables*

Consider a coin-flip problem. (If you know Pascal's Triangle, this problem will be easier. Pascal's Triangle will be discussed in a class meeting, if you dont know it.) If heads or tails occurs with equal probability for a coin flip, their probabilities are $h = 0.5$ (heads) and $t = 0.5$ (tails). The sum of the probabilities $h + t = 1$, because the coin must land on one of its sides. No edge landings allowed! A coin flip is a "realization" of the random process, that is, the hypothetical coin-flip becomes manifest in The Real World and takes a definite value. If the coin is flipped twice, we assume that the probabilities of heads and tails are independent of the previous flip. In other words, $h = 0.5$ stays the same, no matter whether the previous flip was heads or tails. This means that the coin-flips are statistically independent. You can determine the probabilities of the coin's combined face-landings by multiplying

$$1 = (h + t)(h + t) = h^2 + 2ht + t^2$$

The probability of two heads is $h^2 = 0.5^2 = 0.25$, the probability of two tails is $t^2 = 0.5^2 = 0.25$. There is only one way to obtain two heads, and only one way to obtain two tails. But there are two ways to obtain one heads and one tails, so the probability of that mixed result is $2ht = 2(0.5)(0.5) = 0.5$. The probability of at least one heads is $h^2 + 2ht = 0.75$. In your answers to the questions below, explain your reasoning.

a) if a coin is flipped four times, what is the probability of three heads and one tails?

b) if a coin is flipped four times, what is the probability of at least two heads?

c) if a coin is flipped 8 times, what is the probability of at least 5 tails and at least one heads?

d) if you bet a dollar on every coin flip, what is the probability that you have won at least \$4 after 10 flips? Does this probability change if you vary your choices of heads or tails for each flip?

e) Suppose you have combined a large number of card decks (hearts, spades, clubs, diamonds) so that drawing a few cards from the deck does not change the probabilities of drawing a card significantly. (If you draw cards from a single deck, the probability of drawing a King drops by $\sim$25% after you draw any one King. We dont want to worry about that effect in part (e) here.) If you draw five cards from such a large deck, what is the probability that three or more of the cards will be hearts?

f) If you draw four cards from such a large deck, what is the probability that two or more of the cards will be either a Jack or a Queen?

g) If you draw four cards from a single 52-card deck of cards, what is the probability of drawing two cards that are either a Jack or a Queen?

*Problem 2: Working with a Probability Density Distribution, also known as a PDF*

Consider a random variable $X$ that can take a continuous range of values. The realized values of $X$ follow a probability density function $p(X)$. Probability is different from "probability density" the same way that the density of a rock is different from its mass. Just as one must specify the volume of a rock, even of a tiny rock particle, to determine its mass, one must integrate the PDF $p(X)$ of a continuous random variable over an interval of $X$ to determine a probability. The PDF $p(X) > 0$ at all values where $X$ is possible, $p(X) = 0$ at all values of $X$ that are impossible, and there are no values of $X$ for which $p(X) < 0$. The integrated probability usually is denoted by an upper-case letter, and the probability *density* function conventionally uses a lower-case letter. For instance, the probability that $X_1 \leq X \leq X_2$, that is, that $X$ lies within the interval $[X_1, X_2]$, is denoted by $P(X_1 \leq X \leq X_2)$ and computed with the definite integral

$$P(X_1 \leq X \leq X_2) = \int_{X_1}^{X_2} p(X) dX$$

Just as with the coin-flip probabilities, if we sum the probabilities of all possible values of $X$ we get unity, which in this case means an integral of $p(X)$ over the range of possible $X$.

$$\int_{range} p(X) dX = 1$$

The proper interpretation of the probability density function focusses on the area-under-the-curve that one integrates. The probability of $X$ lying within the interval $[X_o, X_o + dX]$ is $p(X_o)dX$. So the probability corresponds to a narrow sliver of area, with height $p(X_o)$ and width $dX$. If we are stubborn and insist on estimating the probability that $X = X_o$ exactly, then $dX = 0$, our area-sliver has zero width and the probability of $X = X_o$ is zero. Finite probabilities require a finite interval $dX$ of possible values.

The "range of possible $X$" depends on the random variable. Typically we try to describe a random variable with a function that statisticians have studied before. One standard PDF is the Gaussian distribution with mean-value $\mu$ and standard deviation $\sigma$,

$$p(X) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$$

This PDF is defined over all $X \in (-\infty, \infty)$, so that $\int_{-\infty}^{\infty} p(X) dX = 1$. Other PDFs are nonzero over smaller intervals of $X$, as you will see in the exercises.

The expectation value of a function $f(X)$ is denoted by $\langle f(X) \rangle$. The expectation value is determined by summing the values of the function at $X$, multiplied by the probability of $X$. For a continuous random variable $X$, the "probability" in question is $p(X)dX$, corresponding to a thin sliver of area. We execute the sum of probabilities, weighted by $f(X)$, as an integral

$$\langle f(X) \rangle = \int_{range} f(X) p(X) dx$$

The most common expectation values to determine from a PDF are the mean value $\mu = \langle X \rangle$, and the variance about the mean $\sigma^2 = \langle (X - \mu)^2 \rangle = \langle (X - \langle X \rangle)^2 \rangle$. The only requirement for determining an expectation value is that the integral be well-defined. One reason that statisticians like the Gaussian PDF is that its expectation values are well-defined, even for many functions that tend toward infinity as $X \to \pm\infty$.

Calculate the expected mean $\langle X \rangle$, mean square $\langle X^2 \rangle$, and variance about the mean $\langle (X - \langle X \rangle)^2 \rangle$ for the three distributions in (a), (b), and (c). Verify that each distribution normalizes to unit probability if integrated over its range of applicability.

a) The boxcar (defined for all real $X$)

$$p(X) = \begin{cases} \frac{1}{2\sigma} & \text{for } |X| \le \sigma \\ 0 & \text{for } |X| > \sigma \end{cases}$$

b) The symmetric exponential (defined for all real $X$)

$$p(X) = \frac{1}{2\sigma} \exp(-|X|/\sigma)$$

c) the chi-squared distribution with 4 degrees of freedom (defined only for $X \ge 0$)

$$p(X) = \frac{X}{4\sigma^4} \exp(-X/2\sigma^2)$$

d) What is the probability that $X > \sigma/2$ for the boxcar distribution in part (a)?

e) What is the probability that $X < -\sigma$ for the symmetric exponential distribution in (b)?

*Problem 3: Using Expectation Values*

If random variables $X$ and $Y$ derive from PDFs $p_x(X)$ and $p_y(X)$ that are uncorrelated, the expectation value of their product $XY$, that is, $\langle XY \rangle = \langle X \rangle \langle Y \rangle$. If the expectation values $\langle X \rangle = \langle Y \rangle = 0$, then $\langle XY \rangle = 0$ for uncorrelated $X$ and $Y$. We can estimate the expectations of functions of both variables together using a distribution rule over summed terms. For instance, we can break apart the terms of an $XY$ polynomial in the following manner

$$\langle (X+Y)^2 \rangle = \langle X^2 + 2XY + Y^2 \rangle = \langle X^2 \rangle + 2\langle XY \rangle + \langle Y^2 \rangle$$

Note that the distribution rule for expectation values does not extend to products of random values: $\langle XY \rangle \neq \langle X \rangle \langle Y \rangle$ unless $X$ and $Y$ are uncorrelated. Similarly, $\langle X^2 \rangle \neq \langle X \rangle \langle X \rangle = \langle X \rangle^2$ because $X$ is always correlated with itself.

If we construct a new random variable out of a combination of random variables, e.g., $Z = (X+Y)/2$, its correlations with its component variables is not zero. To make the algebra simpler, assume that $\langle XY \rangle = 0$. If we form the a polynomial with $X$ and $Z$ instead of $X$ and $Y$, the corresponding expectation value for $(X+Z)^2$ is

$$\langle (X+Z)^2 \rangle = \langle X^2 + 2XZ + Z^2 \rangle = \langle X^2 \rangle + 2\langle XZ \rangle + \langle Z^2 \rangle$$

$$= \langle X^2 \rangle + \langle X(X+Y) \rangle + 0.25\langle (X+Y)^2 \rangle = 2.25\langle X^2 \rangle + 0.25\langle Y^2 \rangle$$

We can verify this relationship by generating sample values from a known PDF in R and computing the mean values of these quantities. For instance, the R command `rnorm` returns `n` "realizations" of a normal (Gaussian) PDF with specified `mean` and standard deviation `sd`. NOTE that these means and standard deviations are NOT the same as those in your sample R markdown file from our first lecture.

```
x <- rnorm(n=1000, mean=1, sd=2)
plot(x)
y <- rnorm(n=1000, mean=-1, sd=0.5)
plot(x,y)
z <- (x+y)/2
plot(x,z)
mean(x)
mean(x^2)
mean((x+y)^2)
etc.
```

(a) For `n=1000` and `n=5000` realizations of R-function `rnorm`, compute $\langle X\rangle$, $\langle X^2\rangle$, $\langle Y\rangle$, $\langle Y^2\rangle$, $\langle Z\rangle$, $\langle Z^2\rangle$, $\langle (X+Y)^2\rangle$, and $\langle (X+Z)^2\rangle$ empirically. Compare them with the theoretical expectations (hint: if `sd=1`, then $\langle X^2\rangle$ should be close to one). Which case is closer to the theoretical expectations, `n=1000` or `n=5000`?

(b) Loop the expectations in part (a) over 100 cases. The random-number generator will keep the cases statistically independent. Plot a histogram of the empirical averages $\langle X\rangle$, $\langle X^2\rangle$, $\langle Y\rangle$, $\langle Y^2\rangle$, $\langle Z\rangle$, and $\langle Z^2\rangle$ for `n=1000` cases. Plot these histograms and comment on their means and variances about their means.

*Problem 4: Set Theory Identities*

Use Venn diagrams to explain the following identities, where $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are sets within an event space $\mathcal{S}$. The superscript $C$ refers to the complement of a set. If $\mathcal{A}$ is the set of all Kings in a deck of cards, $\mathcal{A}^C$ is the set of all cards in the deck that are not Kings.

a)

$$(\mathcal{A}\cup\mathcal{B})^C = \mathcal{A}^C \cap \mathcal{B}^C$$

b)

$$(\mathcal{A}\cap\mathcal{B})^C = \mathcal{A}^C \cup \mathcal{B}^C$$

c)

$$(\mathcal{A}\cup\mathcal{B})\cap\mathcal{C} = (\mathcal{A}\cap\mathcal{C})\cup(\mathcal{B}\cap\mathcal{C})$$

d)

$$(\mathcal{A}\cap\mathcal{B})\cup\mathcal{C} = (\mathcal{A}\cup\mathcal{C})\cap(\mathcal{B}\cup\mathcal{C})$$

The symbols $\cup$ and $\cap$ denote union (logical "or") and intersection (logical "and"), respectively. The notation $\mathcal{A}^C$ signifies the complement of the set $\mathcal{A}$.