STAT371: Introductory Applied Statistics for the Life Sciences **Fall 2017**

Chapter 4 — Random Variables and Distributions

1. Random Variables

   - **Random variable** is the name given to a conceptual entity that represents a random process. RVs are usually denoted by capital letters, like X, Y, etc.
   - Once the random process that defines an RV is performed, we call the result a **realization** of the RV. Realizations of RVs are not themselves random, but the process by which an RV is realized is random. Realizations of RVs are usually denoted by lower-case letters, like x, y, etc.
   - The **probability distribution** of a random variable consists of a description of the possible values that the RV can realize to, along with the probabilities that each realization will occur. Depending on the type of RV, the descriptions of the possible values and probabilities can take different forms.
     - **Discrete** RVs take only certain values. Probability distributions for discrete RVs are called **probability mass functions**, or **pmfs**, and consist of lists of the values that can be taken by the RV, together with the probabilities of each value. The pmf for an RV X is usually denoted $p(x)$.
     - **Continous** RVs take values in specified ranges. Probability distributions for continous RVs are called **probability density functions**, or **pdfs**, and consist of ranges of values that can be taken by the RV, together with a function that lives on those ranges. The area under the function between any two possible realizations of the RV determines the probability that the RV will realize to a value in that range. The pdf for an RV X is usually denoted $f(x)$.
   - Think of an RV as representing a population, and a collection of realizations of that RV as a sample.
   - RVs have two useful properties:
     - The **expectation** of an RV X, denoted $E(X)$ or μ_X, is like the mean of the population. Theoretically, it represents the mean of an infinite number of realizations of X. The expectation of a discrete RV X is:
     $$\mu_X = E(X) = \sum_x x * p(x).$$
     - The **variance** of an RV X, denoted $VAR(X)$, or σ_X^2 is like the variance of the population. $SD(X) = \sqrt{VAR(X)}$. Theoretically, it represents the variance

of an infinite number of realizations of X. The variance of a discrete RV X is:

$$\sigma_X^2 = VAR(X) = \sum_x p(x) * (x - E(X))^2.$$

2. The Bernoulli Distribution

   - We call a RV a **Bernoulli** RV if it can only realize to the values 0 or 1, and the probability that it realizes to 1 is called π. A Bernoulli RV X can be denoted as $X \sim Bern(\pi)$.
   - If $X \sim Bern(\pi)$, then $E(X) = \pi$, and $VAR(X) = \pi(1 - \pi)$.

3. The Binomial Distribution

   - A **binomial random process** has the following properties:
     - The random process consists of n identical sub-processes, called **trials**.
     - Each trial results in one of two possible outcomes. One is usually called a **success**, and the other a **failure**.
     - The probability of a success on any single trial is the same for every trial, and is denoted π.
     - The trials are independent, in that the outcome of any trial does not affect the outcome of any other trial.
   - If all of the above are true, then a **binomial** RV, call it B, is the total number of successes achieved in n trials of a binomial random process with probability π of success on any given trial. We denote such an RV as $B \sim Bin(n, \pi)$, and the probability of observing b sucesses is:

     $$p(b) = \frac{n!}{b!(n-b)!}\pi^b(1 - \pi)^{n-b},$$

     and the expectation and variance are:

     $$E(B) = n\pi, \text{ and } VAR(B) = n\pi(1 - \pi).$$

   - Since the individual trials in a binomial random process are Bernoulli RVs, where a success is defined as a 1, and a failure as a 0, a Binomial RV can be thought of as the sum of n independent Bernoulli RVs that all have the same probability of success, π.

4. Continuous Random Variables

   - A continuous RV can be though of as the limit of a discrete RV as the number of posssible outcomes becomes infinite.

- For a **density histogram**, the y-axis is scaled so that the total area of all the bars equals one.
- As the bins of a density histogram get narrower, and the sample size increases, eventually the tops of the bars can be approximated by a smooth curve, the pdf.
- For any RV X, we define the **cumulative distribution function** or **CDF** to be the function that returns the probability of the RV X realizing to a value that is less than or equal to any given value x. The CDF is usually denoted by $F(x)$. In symbols, $F(x) = P(X \leq x)$.

5. The Normal Distribution

   The normal distribution has the following properies:

   - The normal can realize to any value between $-\infty$ and ∞.
   - The normal is symmetric around the mean, μ.
   - The inflection points (points where the curve moves from concave downward to concave upward) are at $\mu \pm \sigma$.
   - The total area under the curve is 1.
   - The area under the curve between $\mu - \sigma$ and $\mu + \sigma$ is about 0.68; the area under the curve between $\mu - 2\sigma$ and $\mu + 2\sigma$ is about 0.95. Very little of the area is farther than two sds from the mean.
   - If X is a normal RV, the pdf of X is:

   $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(\frac{-(x-\mu)^2}{2\sigma^2})}$$

   - A normal RV X is denoted $X \sim N(\mu, \sigma^2)$, and the expectation and variance are:

   $$E(X) = \mu \text{ and } VAR(X) = \sigma^2 \text{ (so } SD(X) = \sigma).$$

   - If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.
   - If $Z \sim N(0, 1)$, then $X = Z\sigma + \mu \sim N(\mu, \sigma^2)$.
   - The z such that $P(Z \geq z) = \alpha$ is called z_α. We call this a z critical value.