Summary of Lecture 5

Optional reading: Chapter 3 of OpenIntro Statistics.

Lecture 5 introduced continuous random variables, which are models for uncertainty over continuous outcomes. Unlike discrete random variables, here we cannot list all possible outcomes and assign probability to each possible outcome. Instead we use **probability density function**, pdf, to specify probability of an interval. A probability of any interval (a,b) is equal to the area under the pdf over that interval. The total area under the pdf is equal to 1. Also probability of any particular outcome is equal to zero (for technical reasons).

Uniform distribution gives is used to model uncertainty in cases when any outcome in an interval is equally likely. We write $X \sim U(a,b)$ to denote random variable with the uniform distribution with parameters a and b. The pdf of X is f$f(x) = \frac{1}{b-a}$ if $a < x < b$ and $f(x) = 0$ otherwise.

Normal distribution is the most important distribution in statistics. The pdf is a bell shaped curve. The Normal distribution has two parameters: the mean μ and variance σ^2. If X is a Normal variable with parameters μ and σ^2, we write $X \sim N(\mu, \sigma^2)$. The random variable Z has a standard Normal distribution if $Z \sim N(0,1)$, that is, Z is a Normal random variable with the mean $\mu = 0$ and variance $\sigma^2 = 1$. The mean determines the center of the pdf and the variance the spread of pdf. If $X \sim N(\mu, \sigma^2)$ then the probability that an outcome will be in the interval $\mu \pm 2\sigma$ is roughly 95%. (Probability that X is in the interval $\mu \pm 1.96\sigma$ is exactly 95%, but we will use 2 instead of 1.96.) For a standard normal variable, we have

$$P(-1 < Z < 1) = 0.68$$
$$P(-1.96 < Z < 1.96) = 0.95$$
$$P(-3 < Z < 3) = 0.9974$$

Data arising in the world can often be modeled using an i.i.d. Normal model. To check whether this model is appropriate, you need to make sure that there are no obvious patterns or trends in data and that the histogram looks similar to a pdf of a Normal distribution. In the class, we also saw that some time-series data can be modeled using i.i.d. Normal model after transformation. For example, random walk model assumes that differences between observations are i.i.d. Normal.

Cumulative distribution function provides us another way to compute probability of intervals. The cdf of a random variable X is defined as $F_X(x) = P(X \leq x)$. To compute the probability of an interval (a,b) we can use the cdf as follows $P(a < X < b) = F_X(b) - F_X(a)$.

We can also specify models through linear combinations of random variables. Suppose that we define the random variable Y as $Y = c_0 + c_1 X_1 + c_2 X_2$, then the mean of Y is

$$E[Y] = c_0 + c_1\, E[X_1] + c_2\, E[X_2]$$

and the variance is

$$V[Y] = c_1^2\, V[X_1] + c_2^2\, V[X_2\,] + 2c_1 c_2 \mathrm{cov}[X_1, X_2].$$

Linear combination of independent Normal variables is also normally distributed.

The covariance between two discrete random variables X and Y is given by

$$\sigma_{XY} = \text{cov}(X, Y) = \sum_{all\ (x,y)} p(x, y)\ (x - \mu_x)\big(y - \mu_y\big).$$

The covariance between two continuous random variables is defined similarly, however, you will not need to compute that by hand. In order to compute the covariance, one needs to use marginal probabilities of X and Y to compute the mean.

The correlation between random variables (both discrete and continuous) is given as

$$\rho = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

The correlation always lies between -1 and 1. If ρ is close to 1, there is a line with positive slope such that (X, Y) is "likely" to fall close to it. If ρ is close to -1, it means the same thing but the line has a negative slope. To compute correlation we need to compute the means μ_X and μ_Y (in order to get covariance), and also the variances σ_X^2 and σ_Y^2 (to get the standard deviations).

If two variables are independent, then their correlation is equal to zero. However, the converse is not true. That is, if the correlation is equal to zero, the variables can be dependent.

We standardize variables by subtracting the mean and dividing by the standard deviation. The new variable can be interpreted as the number of standard deviations away from the mean.

For $X \sim N(\mu, \sigma^2)$, the z-value corresponding to a value x is $z = \frac{x-\mu}{\sigma}$. "z-value" or "z-score" represents how many standard deviations a Normal observation is away from the mean.

The **central limit theorem** (CLT) says that the average of a large number of independent random variables is (approximately) Normally distributed. Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables and let

$$Y = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_i X_i.$$

As n gets large, we have that approximately $Y \sim N(\mu_Y, \sigma_Y^2)$.