MGA 60204 Prof. Kasey Buckles

Statistics Boot Camp

(compiled by Prof. Patty Anderson, Dartmouth College)

This sheet reviews some of the probability and statistics that I will assume you still know from your previous statistics course. If none of this looks familiar, you have a lot of work to do to prepare for this class! You need to understand these concepts well. If necessary, pull out your old stats text and notes, or refer to the Appendices in your text to help you further review this material.

Independence

For 2 random variables X and Y, if the outcome of Y is completely unrelated to the outcome of X, then X and Y are said to be independent.

Expected Value

The expected value of a random variable X, E(X), is a weighted average of the possible realizations of the random variable, where the weights are the probabilities of occurrence. It is also called μ_X, or the population mean. More concretely,

for a discrete random variable, $E[X] \equiv \sum_{j=1}^{k} x_j f(x_j)$ and

for a continuous random variable $E[X] \equiv \int_{-\infty}^{\infty} xf(x)dx$.

Important Properties of the Expectations Operator

1. E[a] = a
2. E[aX] = aE[X]
3. E[aX + b] = aE[X] + b
4. E[X + Y] = E[X] + E[Y]
5. $E[(aX)^2] = a^2E[X^2]$
6. If X and Y are independent, then E[XY] = E[X]E[Y]

Variance and Standard Deviation

The variance of a random variable X is a measure of its dispersion around its mean, E(X) and is defined as:

$$\mathrm{Var}[X] = \sigma_X^2 = E[(X - E[X])^2] = E[(X - \mu_x)^2] = E[X^2] - \mu_X{}^2$$

Since variance is measured in units that are the square of the units in which X is measured, the standard deviation, which is the positive square root of the variance, is often reported since it is measured in the same units as X.

$$\text{Std. Dev.}[X] = \sigma_X$$

Important Properties of Variance

1. $\text{Var}[a] = 0$
2. $\text{Var}[aX+b] = a^2\text{Var}[X]$
3. $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$
4. $\text{Var}[X - Y] = \text{Var}[X] + \text{Var}[Y] - 2\text{Cov}[X, Y]$
5. $\text{Var}[aX + bY] = a^2 \text{Var}[X] + b^2 \text{Var}[Y] + 2ab\text{Cov}[X, Y]$

Covariance and Correlation

The covariance is a measure of (linear) association between two random variables. Let W and Z be random variables, then the covariance between W and Z is defined as

$$\text{Cov}[W,Z] = E[(W - E[W])(Z - E[Z])] = E[(W - \mu_W)(Z - \mu_Z)]$$

where μ_W and μ_Z are the expected values of W and Z, respectively. Note that using the properties of the expections operator, and some algebra, we can also write:

a. $\text{Cov}\ [W, Z]= E[WZ]- E[W]E[Z] = E[WZ] - \mu_W\mu_Z$
b. $\text{Cov}\ [W, Z]= E[(W - E[W])Z] = E[(W - \mu_W)Z]$
c. $\text{Cov}\ [W, Z]= E[(Z- E[Z])\ W] = E[Z - \mu_Z)W]$

Just as Var[X] is measured in units of X squared, Cov[W,Z] is measured in units that are the product of the units of W and of Z. This can be confusing – if W is dollars and Z is education, Cov[W,Z] is measured in education-dollars. A useful transformation is the correlation coefficient, ρ, which is unit free. It is always between –1 and +1. The correlation coefficient between W and Z is defined as

$$\rho[W,Z] = \text{Cov}[W,Z]/(\sigma_W\sigma_Z)$$

Important Properties of Covariance

1. $\text{Cov}\ [X, X]= \text{Var}[X]$
2. $\text{Cov}\ [aX+b, cY+d]= ac\text{Cov}[X,Y]$
3. If X and Y are independent, then $\text{Cov}[X,Y]=0$

Conditional Expectations and Variance

The expectation of Y conditional on X (the conditional expectation, or conditional mean) is written as $E[Y|X]$ and allows for us to characterize the relationship between X and Y, even if it is nonlinear. We can also substitute a conditional mean into the variance formula to obtain the conditional variance.

Important Properties of Conditional Expectations and Variance

1. If X and Y are independent, then $E[Y|X] = E[Y]$
2. If X and Y are independent, then $\text{Var}[Y|X] = \text{Var}[Y]$

Sample Moments

For a given sample, we can estimate our population moments using the following estimators:

The sample mean: $\overline{X} = n^{-1}\sum X_i$

The sample variance: $S^2 = \frac{1}{n-1}\sum\left(X_i - \overline{X}\right)^2$

The sample covariance: $S_{XY} = \frac{1}{n-1}\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$

The Law of Large numbers implies that the sample mean is always a consistent estimator for the population mean. For large samples, an alternative for the variance and covariance that replaces n – 1 with n is also consistent.

The Central Limit Theorem

The Central Limit Theorem is a key result from statistics. It essentially says that if you draw a large random sample $\{Y_1, Y_2, \ldots Y_3\}$ from a distribution with mean μ and variance σ^2, then you can act as if you drew from a normal distribution with mean μ and variance σ^2. More precisely, we can say that $Z = \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}}$ has an asymptotic standard (i.e. mean 0, variance 1) normal distribution and so does the sample counterpart (which substitutes S for σ). Note also that any linear combination of independent, normally distributed random variables will also be normally distributed.

Sampling Distributions

Contrary to what the name might suggest to you, a sampling distribution is **not** the distribution from which your sample is drawn. Instead, it is defined as "the probability distribution of an estimator over all possible sample outcomes." To think about what this really means, consider the estimator for the population mean, μ, which as noted above is the sample mean $\overline{X}$. Imagine drawing a random sample of size n from the population and calculating $\overline{X}$. Now draw a different random sample of size n and calculate $\overline{X}$ again. Do this over and over and over and over, etc. You would not expect to calculate the same $\overline{X}$ each time. Instead, if you plotted all of the calculated sample means, you would get the sampling distribution. We can describe the mean and variance of this distribution, as follows:

$$E[\overline{X}] = \mu$$
$$Var[\overline{X}] = \sigma^2/n$$

Given the Central Limit Theorem, we can say that $\overline{X}$ is asymptotically normally distributed with mean μ and variance σ^2/n. That is, we can treat the sampling distribution of the estimator as asymptotically normal.