Chapter 1 - Review of Statistics I

The first semester we laid the groundwork for a full understanding of probability and statistics. Virtually everything that you need to know about what probability is and how it is used to test hypotheses was discussed. Unfortunately, you no doubt remain unclear about many things and certainly you are wondering how we might apply these concepts to real world problems. That is perfectly natural. Everybody feels this way, at first. You should not give up on the subject, since with a little more effort, you will begin to see things much more clearly and you will then understand just how important statistics is to our everyday life. I will try to help you this semester by introducing an incredibly important and useful statistical technique called ***linear regression***.

Linear regression is no doubt the most highly used of all statistical techniques (outside of simple descriptive statistics such as calculating the sample means, variances, and standard deviations). Regression allows us to develop models of how a group of explanatory variables affect - or are linearly related to - a single dependent variable. Unlike many other parts of statistics, linear regression is very intuitive and is extremely useful. Moreover, we have free user-friendly open source software (gretl) that allows us to do all the analysis we will ever need to do. If gretl cannot do the analysis, you probably are looking at your problem in the wrong way.

It may be valuable for us to take a moment and review the basic concepts we introduced last semester. Remember, I know quite well that these are difficult concepts to master and use, but if you are anything like me (and you all are), a little persistence and thought will dissipate the mental clouds and all will be clear.

The foundation of probability and statistics is the idea of a **random variable**. This is merely a variable whose value has yet to be determined. When we consider the sequence of natural numbers 1,2,3,... we all know that the next number in the sequence is 4. It is already determined and therefore we say it is deterministically known. The sequence is deterministic. The same can be said for the sequence of squares 1, 4, 9, 16, ... We all know the next number is 25, since the sequence is formed as 1^2, 2^2, 3^2, 4^2, ... Random numbers are numbers that do not follow **any** formula. We cannot say ahead of time (i.e. before an observation on the random variable is made) exactly what the next number in a sequence of random numbers will be. It has yet to be determined, and therefore is random.

Now, a lot of thought has gone into the next step. Beginning exactly 300 years ago, people started to wonder, quite naturally in a mathematical sense, if some outcomes of a sequence of random numbers were going to be appear more frequently than others. That is, would there be a higher or lower **relative frequency** for some possible outcomes relative to other possible outcomes?[1] This postulated (stable) relative frequency became the definition of probability, and the study of such relative frequencies was called classical probability. Classical probability required us to have a phenomenon which could be repeated over and over, at least in theory.

This big step, in the creation of probability, was nevertheless found to be quite limited. A more inclusive and richer idea was needed. So, classical probability was discarded and, in its place there came the idea of probability as a general measure, governed by a pdf, a "probability density function". The modern concept of probability is that *every* random variable will have associated with it a pdf and all statements about the probability of this random variable can be answered by reference to this pdf.

So, if your putative random phenomenon does not have a clear pdf, you simply cannot use modern probability theory to discuss it.[2] Some people may have heard of the term "nonparametric statistics" and think that this means that the randomness is not subject to

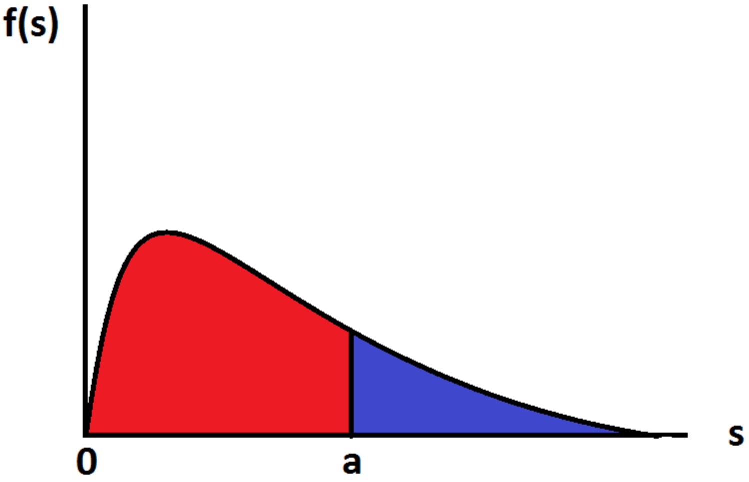
[1] The relative frequency of an observed sequence of data on say N random numbers is the ratio of sucessful events to N. For example, the number of (successful) heads divided by N total flips of a coin.

[2] This is a tyranny of words. Naturally, in everyday life you may have heard it said that Hillary Clinton will "probably" win the US 2016 presidential election. Some prognosticators may have even created so-called "probabilities" or "odds" that she will win. But, this is more like alchemy than science – something that is obvious as one sees how the numbers are constructed. The US presidential election of 2016 was the first and only one of its kind involving Trump and Clinton and cannot be thought of as a repeated event, even theoretically. This is a case where people clamor for quantification of something that essentially cannot be adequately and satisfactorily quantified. Do not make the mistake of thinking that what they are doing is even close to your study of probability and statistics. It is not, despite their use of similar terms. Considerations such as this led the famous English economist John Maynard Keynes to attempt to extend the field of probability as a branch of logic, where logical propositions are either true or false, or else have an associated level or degree of credibility given the information at hand. If Keynes is right, probability is subjective like beauty, with each person having his own sense of what is probable. Such probability is shared only to the extent that human rationality is shared - your logic is the same as my logic. Followers of this view are called subjectivists or Bayesians, rather than frequentists, who rely on relative frequencies to define probability. See Keynes, Treatise on Probability, (1921).

this pdf issue. That would be wrong. For a simple definition of nonparametric statistics, see this link.

The central relation of probability is that random variables have associated with them uniquely defined pdf's - just like you and your DNA are related. Let's denote this pdf by the symbol $f(s)$. We can now draw an example of a pdf as in Figure 1 below. The red area is the probability that our random variable, X, is less than a, written as $P[X \leq a]$. The red and blue areas must sum to 1, since this is a condition for all pdfs.

Figure 1 Probability Density Function (pdf)



Clearly, it is possible to give probability assessments to nearly any set of points in the above graph.[3] The red area in Figure 1 can be written as

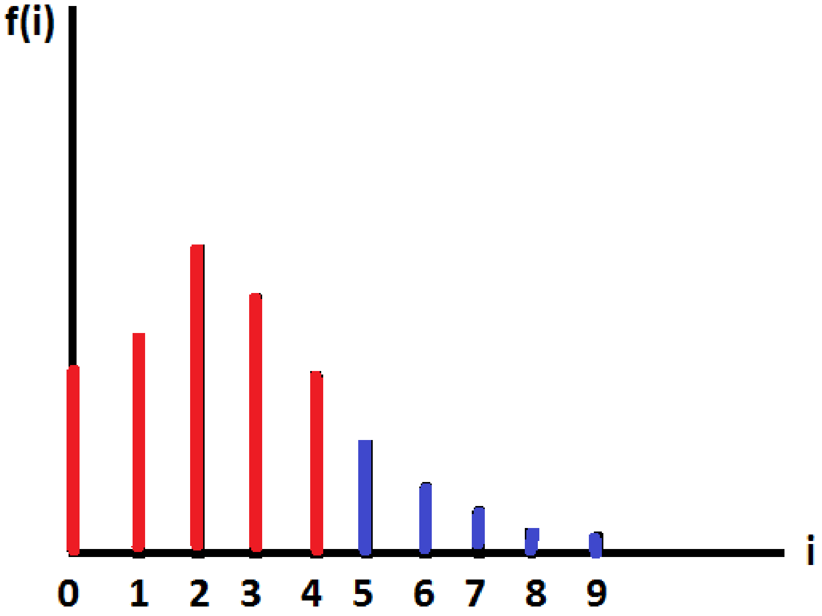
$$P[X \leq a] = \int_0^a f(s)ds$$

Some random variables are discrete in nature and therefore their pdf's (called *probability mass functions*, pmf) are not continuous lines, as shown in Figure 1, but discrete spikes as shown in Figure 2. We can use an analogous method to calculate probabilities for the discrete case. Here we write things as

[3] Purists insist, rightly so, that we must restrict ourselves to sets that possess the countably infinite additivity property. Apparently, there are sets that are so bizarre we cannot infinitely add their probabilities up without getting absurdities, so we restrict ourselves to sets called sigma algebras. No need to worry. Every conceivable set that you will meet in your life will no doubt satisfy this additive property.

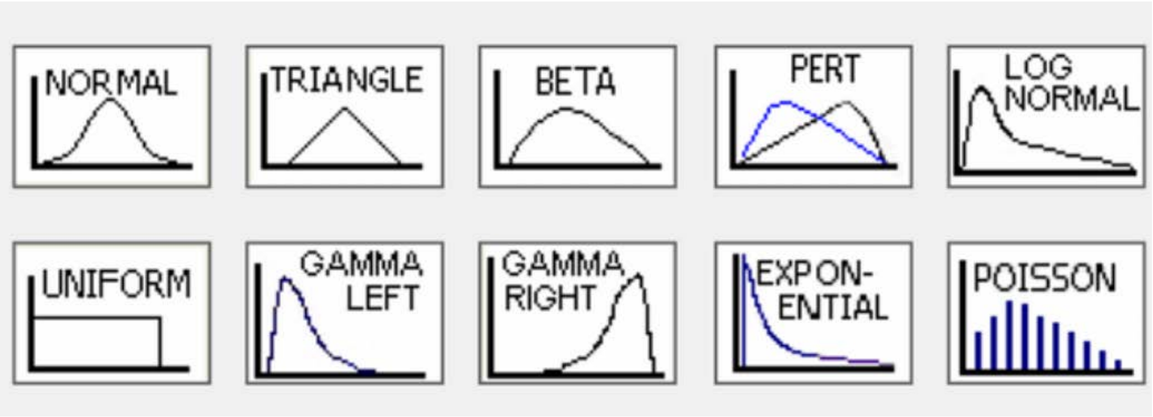
$$P[X \leq K] = \sum_{i=0}^{K} f(i)$$

Figure 2 Probability Mass Function (pmf)



Once again, to get the probability that X is less than or equal to 4, you add up the red spikes in Figure 2. Nothing could be easier. The hard part is trying to decide which type of pdf your random phenomenon has. Sometimes, the nature of the random phenomenon will tell us what is the pdf (like flipping a coin). At other times, we look for a reasonable pdf

Figure 3 - Some Basic pdfs and one pmf



that will model the random phenomenon (like a normal distribution). In a sense, probaility theory is just humans' best efforts to grapple with seeming disorder and uncertainty. It is our systematic effort to try to bring a little sense to the recurring chaos we are observing in the world.[4]

In most cases we consider, the pdf will have **unknown** parameters. These paramters are not only unknown, they are unknowable. However, with data we can guess them. This is the subject of estimation. For example, the negative exponential (useful in testing the lifespan of lightbulbs) can be written as

$$f(s) = \lambda e^{-\lambda s} \quad \text{for } 0 \le s < \infty \ .$$

The theoretical mean of of this random variable can be shown to be $E[X] = \frac{1}{\lambda}$. We can therefore estimate the mean by $\hat{E}[X] = \overline{x} = \frac{(x_1 + x_2 + \cdots + x_N)}{N}$ assuming we have data or (observations on random X) and hence we can estimate λ by $\hat{\lambda} = \frac{1}{\overline{x}}$.

The grand scheme of things is that (i) we have a random variable, X, which naturally has (ii) a pdf. The pdf is a formula used to (iii) calculate all probailities of X. In order to use the pdf, we must have (iv) guesses or estimates for the values of all unknown parameters. Getting (v) data or observations on random X, we then (vi) combine these data to form estimates of the parameters in the pdf. Finally, we may want to (vii) test hypotheses about the unknown parameters in the pdf. We will review testing hypotheses in a moment.

Another area of statistics which we discussed last semester is the ideas of mean and variance. The theoretical mean and variance of a random variable, X, having pdf $f(s)$, are defined as, respectively,

$$E[X] = \int_{-\infty}^{\infty} sf(s)ds$$

[4] The fact is that statistics is merely a very sophisticated collection of methods for detecting and verifying *patterns* in observations or data. There is always the possibility that any pattern seen in the data is actually an artifact (something made by humans and not having a separate, objective existence apart from humans). This is what underlies the problem of induction.

$$\text{var}(X) = \int_{-\infty}^{\infty} (s - E[X])^2 f(s) ds$$

Note that these formulas are theoretical, depend only on the pdf, and have nothing to do with data. However, we can certainly estimate these two theoretical objects using data. The typical way this is does is as follows -

$$\hat{E}[X] = \overline{x}$$

$$\text{v}\hat{\text{a}}\text{r}(X) = \sum_{i=1}^{N} (x_i - \overline{x})^2 / (N-1)$$

Why are we interested in means and variances? The main function of the mean and variance is to summarize the pdf in just two numbers. They give us a quick and ready way of knowing where and how spread out the pdf is.[5] The mean summarizes where the pdf is generally located (and is referred to as a measure of central tendency). The variance summarizes how spread out the pdf is; how dispersed the probability is about the mean (and is called a measure of dispersion).

We also considered certain laws or rules governing means and variances. These are easily proven using the definitions of mean and variance above and the property of integrals. For example, we know the following regarding means and variances.

(1) $E[\alpha X_1 + \beta X_2 + \gamma] = \alpha E[X_1] + \beta E[X_2] + \gamma$ for any constants $\alpha, \beta,$ and γ.

(2) $E[XY] = E[X]E[Y]$ if the random variables X and Y are statistically independent[6]

(3) $\text{var}(\alpha X + \beta) = \alpha^2 \text{var}(X)$ for any constant (including negative numbers) α and β

(4) $\text{var}(X) = E[(X - E[X])^2] = E[X^2] - \{E[X]\}^2$

(5) $\text{var}(X) = 0$ if and only if X = constant and is therefore not random

There are other rules as well and a good student of statistics will memorize these first and then prove them to be true over and over. They are very useful in understanding the subject better.

[5] The mean and variance are also important in deriving certain formulas for test statistics. A good example of this is the Central Limit Theorem that utilizes both the mean and the variance of the random variable X.

[6] The random variables X and Y will be indpendent if the value of X cannot affect the probability of Y and thevalue of Y cannot affect the probability of X. In this case the joint pdf of X and Y can be factored into two parts multiplicatively as f(s,w) = g(s)h(w). This is called the factorization theorem.

The last important discussion we had last semester concerned the Central Limit Theorem (CLT) and how to use it to test hypotheses. As we have noted in class, the CLT is probably the most important theorem in all of statistics.[7] This means we better know how to use it.

To begin with we can state the theorem as follows,

> **<u>CLT:</u>** Let $X_1, X_2, \ldots, X_N$ be a random sample that are independent and identically distributed with mean μ and variance σ^2. Form the random sample mean using these random variables to get $\bar{X} = \frac{X_1 + X_2 + \ldots + X_N}{N}$. Then $\frac{\sqrt{N}(\bar{X} - \mu)}{\sigma} \to N(0,1)$ as N $\to \infty$.

As an example of how to use the CLT, we can consider testing a hypothesis on the Negative Exponential distribution.

Suppose we have a random variable X distributed as a Negative Exponential with pdf equal to $f(s) = \lambda e^{-\lambda s}$ for $0 \le s < \infty$ and zero elsewhere. Next, lets suppose that for whatever reason we want to test the hypothesis H$_o$: $\lambda = 4$. We proceed to go out and sample X 100 times, meaning that our number of observations is N = 100. We get data $x_1, \ldots, x_{100}$ and form a sample mean from the data equal to $\bar{x} = 0.32$. Can we reject the null hypothesis H$_o$ at the 5% level of significance? To test this using the Neyman--Pearson method of testing hypotheses, we must follow several steps. Here they are.

(step i) Get the null hypothesis, H$_o$.

That is easy because we have already done this. H$_o$ is H$_o$: $\lambda = 4$.

(step ii) Assume the null is true.

That is even easier because all we have to do is assume $\lambda = 4$. We need to do this to use the CLT as you will see.

(setp iii) Get a test statistic, T().

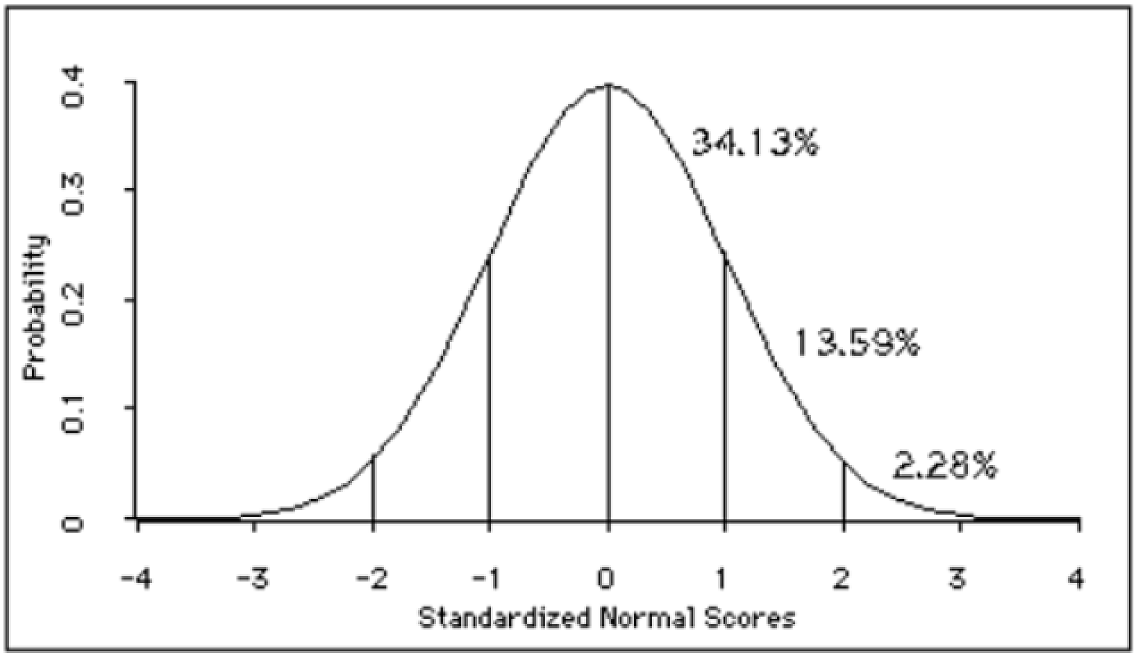
This is normally a very hard part of the testing procedure. Usually, statisticians have alreeady done the heavy lifting. It is theor job to help us get these test statistics. No one would ever expect you to be able to find such test statistics on your own. Because we will use the CLT we have a beautiful test statistic already. Our test statistic is as follows - T(X) = $\frac{\sqrt{N}(\bar{X} - \mu)}{\sigma}$

[7] Of course, the CLT is proven using still deeper theorems, but these are of less value in applied work. The CLT allows us to develop wonderful and quick ways to go about testing hypotheses when we have large amounts of data.

However, there are some additional things we must do. **First**, we note that for the negative exponential, the mean $\mu = \frac{1}{\lambda}$. But, we are assuming $\lambda = 4$ (Remember step (ii) above) . Therefore, $\mu = \frac{1}{4} = 0.25$. **Second**, the variance of a Negative Exponential random variable is $\sigma^2 = \frac{1}{\lambda^2} = \frac{1}{4^2} = \frac{1}{16} = 0.0625$. The square root of this is $\sigma = 0.25$. Also, remember that N = 100, so that $\sqrt{100} = 10$.

(step iii) Get the pdf of the test statistic.

The CLT tells us that our test statistic T(X) = $\frac{\sqrt{100}(\bar{X} - 0.25)}{0.25}$ is a approximately a standard normal random variable. This means the pdf can be graphed as



(step iv) Substitute the sample mean $\bar{x}$ = 0.32 in for the random variable $\bar{X}$ in the formula for our test statistic. If the absolute value of this test statistic is greater than 1.96 then reject Hₒ, Otherwise do NOT reject Hₒ. In our case the observed test statistic is

$$T_o = \left| \frac{\sqrt{100}(0.29 - 0.25)}{0.25} \right| = 1.6 < 1.96 \text{ do NOT reject } \mathrm{H}_o : \lambda = 4$$