SECTION 2 : Statistical Building Blocks

In Section 1 we considered various mathematical aspects of probability theory. We now apply some of those mathematical tools to study the *statistics* of real data samples.

2.1 : The Sampling Distribution

Consider a RV, X, with pdf $p(x)$. Suppose we observe n different *realisations* (values) of X. We call the set $\{X_1, ..., X_n\}$ a **random sample from the population with pdf** $p(x)$. The joint pdf, $g(x_1, ..., x_n)$, is known as the **sampling distribution** of the random sample. We can think of this joint pdf in terms of the 'histogram' picture which we discussed in Section 1; i.e. if we were to repeatedly random sample sets of n numbers from the pdf, $p(x)$, and construct an n-dimensional histogram of the sampled values, then in the limit as the number of samples tends to infinity the 'shape' of the histogram will approximate the sampling distribution, $g(x_1, ..., x_n)$. In this course we will consider only random samples in which all the elements are **independently and identically distributed** (usually written as iid). This means that the sampled value of $X = x_1$ is independent of $X = x_2$ and so on. In other words, the elements of the random sample are statistically independent of each other. It then follows that

$$g(x_1, ..., x_n) \quad = \quad p(x_1)p(x_2)...p(x_n)$$

i.e. the joint pdf of the random sample is product of the individual pdfs.

2.2 : Parameter Estimation

Suppose we wish to study a population which is known (or assumed) to have a pdf, $p(x; \theta)$. This notation indicates that the pdf is dependent upon a (possibly unknown) parameter, θ. If we observe a random sample from the population, $\{X_1, ..., X_n\}$ say, how can we estimate the parameter, θ? How do we decide how 'good' our estimate of θ is (or even what we mean by this question?).

2.2.1 : Statistics

A **statistic** is a function of observable random variables which does **not** depend upon any unknown parameters. Thus if we have a random sample, $\{X_1, ..., X_n\}$, from the population with pdf $p(x; \theta)$ then any function of $\{X_1, ..., X_n\}$ which does **not** depend on θ is an example of a statistic.

Suppose, for example, that $X \sim N(\mu, \sigma)$, where μ and σ are not known *a priori*. Then $X - \mu$ is **not** a statistic, since it depends on the value of the parameter, μ. The key idea in parameter estimation is to use statistics to estimate the unknown parameters of a pdf.

2.2.2 : Estimators

A statistic with which we estimate the value of a parameter is known as an **estimator** of that parameter. Estimators are usually denoted by a caret, or 'hat', e.g. $\hat{\theta}$ is an estimator of θ.

Note that $\hat{\theta}$ is **not** a function of θ (if it depended on the value of θ then it would be redundant as an estimator of θ!). Note that $\hat{\theta}$ is, however, a RV since it is a function of the RVs $\{X_1, ..., X_n\}$. Hence we can (in principle, at least) determine the pdf of $\hat{\theta}$ in terms of the sampling distribution, $g(x_1, ..., x_n)$. This means that the pdf of $\hat{\theta}$ depends upon the **true** value of the parameter, θ. We can therefore write the pdf of $\hat{\theta}$ as $p(\hat{\theta}; \theta)$, and we can use the properties of $p(\hat{\theta}; \theta)$ to decide whether $\hat{\theta}$ is a 'good' estimator.

Consider the following illustrative example. (We take an example from cosmology, although similar examples from any other branch of astronomy could be presented, since it is not the astronomical details but the statistical details which are important here).

Suppose we are measuring the redshift of a nearby galaxy in (say) the Virgo cluster. We do this, of course, by identifying features in the spectrum of the galaxy and comparing their wavelengths with the laboratory values. Thus, if we denote the *true* redshift of the galaxy by z_0, then an estimator of z_0, denoted by $\hat{z}$, will be a function of the observed wavelengths of the (n) identifying spectral features, i.e.

$$\hat{z} \quad = \quad \hat{z}(\lambda_1, ..., \lambda_n)$$

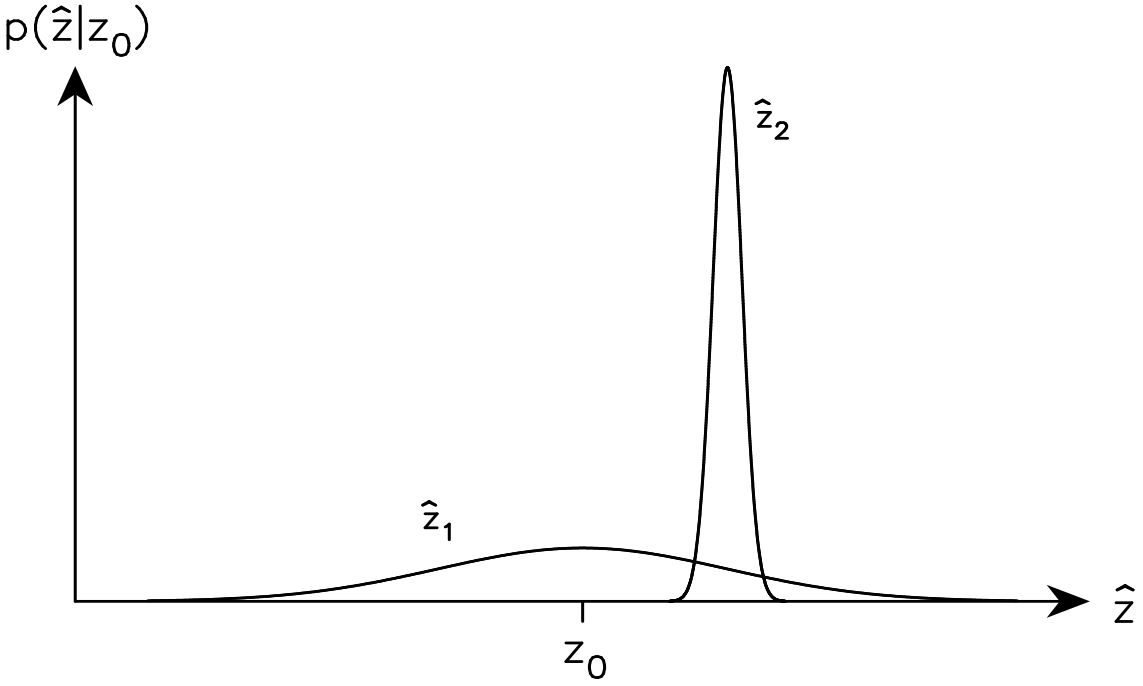
Since the sampling distribution of $\lambda_1, ..., \lambda_n$ depends on z_0, the pdf of $\hat{z}$ also depends on z_0, i.e. $p(\hat{z}) = p(\hat{z}; z_0)$.

We could measure the redshift using, e.g., a 1m-class ground based telescope with a low-resolution spectrograph, but with these data our determination of the redshift will be somewhat inaccurate (since our measured wavelengths of the identifying spectral features will be imprecise). Thus, if we were to repeat our observations with such a telescope a large number of times, a histogram of our estimated redshifts would tend in shape towards the pdf of $\hat{z}_1$, shown in Figure 9. In simple terms, we would say that our observation carried a large **statistical error** but small **systematic error**.

Suppose now we observe the same galaxy with e.g. the high-resolution spectrograph on

HST, and denote by $\hat{z}_2$ our HST estimator of z_0. With HST our wavelength measurements of the galaxy's spectral features will now be much more accurate, leading to a much narrower range of values (i.e. *realisations*) of $\hat{z}_2$, if we were to repeat our HST observations a large number of times. Suppose, however, that – for some reason – we **mis-identify** the features in the galaxy's spectrum, leading to a completely erroneous value of $\hat{z}_2$ in each of these realisations (although, of course, we would only know this if we knew the *true* value of z_0). In this case the pdf of $\hat{z}_2$ would be as shown in Figure 9, and in simple terms we would say that our observation carried a small **statistical error** but a large **systematic error**.

Figure 9: PDF of two estimators of the true redshift, z_0, of a galaxy.



We see from Figure 9 that $p(\hat{z}_1; z_0)$ is much broader than $p(\hat{z}_2; z_0)$, so there is a higher probability that $\hat{z}_1$ will differ considerably from z_0 than for $\hat{z}_2$. However, there **is**, at least, a non-negligible probability that $\hat{z}_1$ lies very close to z_0, whereas the 'narrowness' of $p(\hat{z}_2; z_0)$ means that $\hat{z}_2$ will almost always **systematically underestimate** the true redshift. These two extremes illustrate the essential difficulty in defining one single criterion which determines which estimator is 'best' in a given situation. If one wishes specifically to exclude large statistical errors, but is prepared to tolerate a small systematic 'offset' in the estimator of the parameter (particularly if it is possible to determine the size of that offset, perhaps from independent data, and thus correct for it), then $\hat{z}_2$ would be the better choice. If, on the other hand, even a small systematic error is unacceptable,

then $\hat{z}_2$ would have to be regarded as a 'bad' estimator. In this example we have used the 'closeness' of $\hat{z}_1$ and $\hat{z}_2$ to z_0 as a measure of which estimator is better. We can formalise this idea of 'closeness' of an estimator to the true value of the parameter as follows:-

2.2.3 : Bias of an estimator

We define the **bias**, $B(\hat{\theta};\theta_0)$, of an estimator, $\hat{\theta}$, by

$$\begin{aligned} B(\hat{\theta};\theta_0) &= E(\hat{\theta};\theta_0) - \theta_0 \\ &= \int (\hat{\theta} - \theta_0)\, p(\hat{\theta};\theta)\, d\hat{\theta} \end{aligned}$$

where θ_0 is the true value of the parameter θ. Hence, when an estimator is **unbiased** its expected value is equal to the true value of the parameter.

2.2.4 : Risk of an estimator

We define the **risk**, $R(\hat{\theta};\theta_0)$, of an estimator, $\hat{\theta}$, by

$$\begin{aligned} R(\hat{\theta};\theta_0) &= E\left[(\hat{\theta} - \theta_0)^2;\theta_0)\right] \\ &= \int (\hat{\theta} - \theta_0)^2\, p(\hat{\theta};\theta)\, d\hat{\theta} \end{aligned}$$

The risk of an estimator is also known as the **mean squared error**. Note that when an estimator is unbiased then the risk is identically equal to the **variance** of the estimator.

In the example of Figure 9, $\hat{z}_1$ is an unbiased estimator with a large risk (and variance), whereas $\hat{z}_2$ is negatively biased, but has smaller risk (and very small variance).

Note that the bias of $\hat{z}_2$ is itself a function of z_0. This fact indicates two fundamental difficulties:-

- If we apply a correction to remove the bias of $\hat{z}_2$ at z_0 it does not follow in general that this correction will leave $\hat{z}_2$ unbiased *for all* true values of z; indeed the correction may *increase* the bias of the estimator for other true redshifts.
- In any case, to completely remove the bias of $\hat{z}$ at z_0 **strictly speaking** we need to already know the value of z_0 – if we knew that, then we would have no need to estimate the parameter!

Fortunately, in practice one can frequently define estimators which are unbiased for a wide range of, or indeed all, values of the unknown parameter, so that in particular we don't need to know the true value of the unknown parameter to know that its estimator is unbiased. The simplest example of such an estimator is the **sample mean**.

2.2.5 : The sample mean

Let $\{X_1, ..., X_n\}$ denote a random sample drawn from a population with pdf $p(x)$, mean value μ and finite variance σ^2. We define the **sample mean** as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Clearly $\hat{\mu}$ is an estimator. If each X_i is independently and identically distributed (iid), then $\hat{\mu}$ is an **unbiased** estimator of μ, for all values of μ. (For proof of this result see handout and lectures).

The **variance**, $\sigma^2_{\hat{\mu}}$, of the sample mean is given by

$$\sigma^2_{\hat{\mu}} = \sigma^2/n$$

(For proof of this result see the handout: this proof is not examinable)

This result is extremely important in statistics, since it implies that, whatever the underlying population (provided it has finite variance) the distribution of the sample mean becomes increasingly concentrated near the population mean as the sample size increases. Thus, the larger the sample, the more sure we can be that $\hat{\mu}$ is a good estimator of μ. This idea is formalised quantitatively in the **law of large numbers**.

2.2.6 : The law of large numbers

Let $p(x; \mu, \sigma^2)$ be the pdf of a RV, X, with mean, μ, and finite variance, σ^2. Let $\hat{\mu}$ be the sample mean of a random sample of size n drawn from $p(x; \mu, \sigma^2)$. Let ϵ and δ be two specified small numbers such that $\epsilon > 0$ and $0 < \delta < 1$. If n is any integer such that $n > \sigma^2/\epsilon^2\delta$, then

$$\text{Prob}[\,|\hat{\mu} - \mu| < \epsilon\,] \quad \geq \quad 1 - \delta$$

Thus, we can make the probability that $\hat{\mu}$ lies within ϵ of μ arbitrarily close to unity, simply by taking a large enough sample of data.The proof of this theorem is, again, non-examinable, but is provided on a handout for completeness.

What is striking about the law of large numbers is the fact that we made no assumptions about the form of the pdf of X (apart from its finite variance), and yet we can *still* make precise statements about the probable 'closeness' of $\hat{\mu}$ and μ for a given sample size.

In fact, we can go much further than this in determining the properties of the sample mean, by using one of the most important theorems in statistics.

2.2.7 : The Central Limit Theorem

Let $p(x; \mu, \sigma^2)$ denote a pdf with mean μ and finite variance σ^2. Let $\hat{\mu}$ denote the sample mean of the iid random sample $\{X_1, ..., X_n\}$, of size n. Then as $n \to \infty$, the pdf of $\hat{\mu}$ approaches a normal pdf with mean value μ and variance σ^2/n.

We will not prove the central limit theorem in this course. We do, however, highlight its importance. The CLT states that, no matter what pdf our random sample is drawn from, the sample mean will have an approximately normal distribution as the sample size increases. The CLT justifies the importance of the normal distribution – in applied statistics in general, and in astronomy in particular. Astronomy is filled with situations where one 'bins' or groups sets of observational data. The CLT tells us that, when we bin data with a sufficiently large sample, the fluctuations in the average of the binned data will look approximately normally distributed. Figure 10 illustrates this, for random samples drawn from an exponential distribution – i.e. the underlying pdf is very different from a normal pdf, and yet the distribution of the sample mean very closely approximates a normal pdf as the sample size increases.

The sample mean is, thus, defined according to an intuitively simple expression, is unbiased, and has very special asymptotic properties which are almost independent of the pdf of the underlying population. This is rarely the case with other parameters of a pdf, however, and we require in statistics more general methods for finding estimators – methods which take account of the form of the underlying pdf from which our sample data are drawn.

Figure 10: Histograms of the sample mean of sample of size n drawn from an exponential distribution, for $n = 10$, $n = 20$, $n = 40$ and $n = 100$. Note the increasingly close approximation to a normal distribution.