Populations, Samples and Estimation.

Consider 3 quite different scenarios.

1) A manufacturer of electronic fuses needs to know the maximum amps at which a any one of a batch of fuses will burn out. Testing every one to destruction will leave none to sell, so a few of them are examined to get an idea of the properties of the batch as a whole.

2) The government of a large country needs to know what proportion of the population will be eligible for benefits in a health program directed at preventing a particular disease. It is not possible to examine everyone in the country for susceptibility to the disease (some of whom will have died and others who will have born during the examination process) so a much smaller group of individuals is examined in order to gage the likely proportion in the population that are susceptible.

3) A local authority, responsible for providing an emergency response team for car crash victims on a collection of highways, needs to know the likely monthly frequency and location of crashes in order to establish the size of, and resources for, the response team. The location and number of crashes per month in recent history is used as a guess of what the likely locations and frequencies would be.

Each of these three cases has all the characteristics of the problem facing a statistician, the need to glean some information about an always obscure and often large collection of things - the Population of Interest - by examining a much smaller and very real sub group of that collection - The Sample.

When every element in the population is identified and the relevant characteristic recorded, the resultant data set is referred to as A Census; it constitutes a complete record of The Population of Interest. It is not necessarily an infinite list (the complete batch of fuses could certainly all be examined, the population of a country could certainly all be counted at a point in time) though sometimes it is (the number of places that accidents could take place in a highway system is infinite and the theoretical frequency with which they occur on average over a given period of time is certainly obscure and unobservable). When for various reasons (economic, feasibility, or practical) a census cannot be taken, the Population of Interest is something about which we can only conjecture. Typically in statistics the characteristic of the population we are interested in is treated as a random variable and a probability density function (p.d.f.) is used to describe its

distribution across the range of potential values. One of the arts of practising applied statistics is that of choosing the right distribution for the problem at hand and estimating the parameters upon which the distribution depends.

The Sample is something tangible that we do observe and use to explore conjectures about the population of interest, i.e. we use the sample to tell us something about a population which we cannot examine directly via a census, specifically we use it to estimate the parameters (or relevant functions of them) of the p.d.f. that describes the population.

The way that the sample is taken will clearly influence the estimates we get. The simplest, most effective form of sampling is simple random sampling wherein all of the elements of the sample are each independently and randomly drawn from the population. Agencies that collect data often collect "representative" samples for reasons of economy and unless great care is taken the results emerging from such samples can be misleading. Two types of "representative sampling" are Cluster sampling - which divides the population into clusters or groups and randomly selects a small set of clusters within which a complete census is taken - and Stratified Sampling - splitting the population into mutually exclusive groups or Strata (by age, location, gender or profession for example) and taking a random sample from each strata. For example if one draws an equal number of people from each of the provinces in Canada (stratification by location) in order to calculate the average height of people in Canada, and if height is related to location so that the further west one goes, the taller people tend to be, then a straight average across all of the samples will misrepresent the average height in Canada. Here we concentrate on the properties of estimators of the population mean based upon a Simple Random Sample.

Properties of the Sample Mean.

Simple random sampling yields well defined and intuitively attractive properties for our estimators. In this instance all the elements in the sample are mutually independent of one another so that the joint distribution of the sample is the product of the individual densities. Furthermore every sample selected should have the same probability of being selected. Consider the following example: 4 students A, B, C and D write a test which the TA has marked. The marks were A 2, B 2, C 3 and D 4 but the instructor does not know this, he does not want to read all of the tests so he takes a random sample of one and takes the mark (which we will interpret as the average of the sample) in order to estimate the class average.

Table 1.

Sample	Estimate	Deviation from mean	Squared Deviation from Expected Value	Probability of Drawing this Sample
A	2	0	0	1/4
B	2	0	0	1/4
C	3	1	1	1/4
D	1	-1	1	1/4
Expected Value	2		1/2	

Notice the expected value of the estimate, which was defined in chapter 3 as:

$$\sum_{allpossiblesamples} \bar{X}_{sample} P(sample)$$

is equal to the true mean of the class , though individual samples will render an estimate different from the true mean (samples C and D for example). This demonstrates that the estimator has the property of Unbiasedness, about which more later. Furthermore the Variance of the sample mean (the expected value of the squared deviation from the expected value), defined in chapter 3 as:

$$\sum_{allpossiblesamples} \left(\bar{X}_{sample} - E\left(\bar{X}_{sample} \right) \right)^2 P(sample)$$

is equal to the true variance of the class. This latter phenomenon arises because the estimator is based on a sample of one.

Suppose now the instructor takes a random sample of two and takes the average mark in order to estimate the class average, Table 2 records the set of possible samples together with their corresponding means etc.

Table 2.

Sample	Average	Deviation from Mean	Squared Deviation from Mean	Probability of drawing this sample

A,B	2	0	0	1/6
A,C	2.5	0.5	.25	1/6
A,D	1.5	-0.5	.25	1/6
B,C	2.5	0.5	.25	1/6
B,D	1.5	-0.5	.25	1/6
C,D	2	0	0	1/6
Expected Value	2		1/6	

Notice that this estimator is also unbiased (its expected value is equal to the population mean) but now its variance is smaller (1/6 as opposed to 1/2). This is a result of the estimator being based upon more information (2 students marks as opposed to one) and reflects the increased precision of the estimator that more information engenders. Another way of interpreting this is that the probability of getting an estimate further than a particular distance from the true value has been reduced by the utilization of more information..

To check this effect of increasing sample size suppose now the instructor takes a sample of 3 and takes the average mark in order to estimate the class average, Table 3 describes the situation.

Table 3.

Sample	Average	Deviation from mean	Squared Deviation from Mean	Probability of Drawing this Sample
ABC	2 1/3	1/3	1/9	1/4
ABD	1 2/3	-1/3	1/9	1/4
ACD	2	0	0	1/4
BCD	2	0	0	1/4
Expected Value	2		1/18	

Indeed the estimator is still unbiased and its variance has been reduced yet again.

What do we take from this example? First the sample mean is a random variable because it is based upon a random sample. Secondly it is an unbiased estimator of the population mean regardless of the sample size. Thirdly its variance diminishes with the sample size reflecting the increased amount of information being employed. In fact in general given a random sample of X_i, i = 1,.., n from a population described by the distribution f(x) such that E(X) = μ and V(X) = σ^2 it is readily seen that:

$$E(\bar{X}) = E(\frac{\sum_{i=1}^{n} X_i}{N}) = \frac{\sum_{i=1}^{n} E(X_i)}{N} = \frac{\sum_{i=1}^{n} \mu}{N} = \mu$$

which is to say that the expected value of the sample mean is the population mean. The
Variance of the sample mean can be deduced in a similar fashion as follows:

$$V(\bar{X}) = V\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2} V(\sum_{i=1}^{n} X_i)$$

this follows from the variance of a constant times a random variable being equal to the constant
squared times the variance of the random variable and that since the Xi's are all independent of
one another the variance of the sum will be equal to the sum of the variances so that:

$$\frac{1}{n^2} V(\sum_{i=1}^{n} X_i) = \frac{1}{n^2} (\sum_{i=1}^{n} V(X_i)) = \frac{1}{n^2} \sum_{i=1}^{n} (\sigma^2) = \frac{\sigma^2}{n}$$

i.e. the variance of the sample mean is equal to the population mean divided by the sample
size so that larger samples imply smaller variances. Given the nature of a simple random sample,
the only requirement for these two properties is that the population has a mean and a variance.
It turns out that there are an infinite number of unbiased estimators of the sample mean based
upon a sample of size n, consider for example X* given by the following formula:

$$X^* = \frac{1}{n+2}(X_1 + X_2 + \sum_{i=1}^{n} X_i)$$

It can be readily shown that E(X*) = μ (demonstrate this as an exercise) but is it "better" than the sample mean? To help us compare and choose between alternative estimators we look for them to have certain qualities or "properties", the properties briefly introduced here are Unbiasedness , Efficiency, Consistency and Sufficiency. We will prefer estimators that possess these qualities over estimators that do not. In this book attention is focused on estimating population means, proportions and variances but there are many other population parameters that are of interest. The following properties can be demanded of an estimator of any unknown parameter so for generality in the following discussion θ denotes the population parameter to be estimated and θ(X,n) denotes the estimator (connoting the fact that the estimator is a function of a sample of the X's of size n).

Unbiasedness.

For an estimator to be unbiased it is required that the expected value of the estimator be equal to the parameter being estimated so that:

$$E(\theta(X,n)) = \theta$$

What this means is the average of the estimator over all the possible samples of size n is equal to the value of the parameter being estimated. For estimators that are linear in the random variable (such as the sample mean) unbiasedness is easy to check for given the linearity property of the expectations operator (the expected value of a linear function of a random variable is the same linear function of its expected value). So for example consider the average of a random sample of size n drawn from a normally distributed population such that $X \sim N(\mu,\sigma^2)$ where an element of the random sample is denoted Xi, i = 1,..,n so that E(Xi) = μ and V(Xi) = σ^2.

$$E(\bar{X}) = E(\frac{\sum_{i=1}^{n} X_i}{n}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu .$$

Efficiency

The efficient estimator is the unbiased estimator with the smallest variance for a given sample size. Often we are confronted with two unbiased estimators $\theta_1(X,n)$

and $\theta_2(X,n)$ based upon a sample of size n, if $V(\theta_1(X,n)) < V(\theta_2(X,n))$ then $\theta_1(X,n)$ is said to be relatively more efficient than $\theta_2(X,n)$. Again, given the linearity of the expectations operator, for linear estimators based upon random samples, evaluating their variance are pretty straightforward for example consider the variance of the sample mean.

$$V(\bar{X}) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

With a very large number of estimators a lot of pair-wise comparisons can be tedious. Fortunately we have a criterion known as the Cramer-Rao lower bound which tells us in a particular circumstance what the lowest possible variance for an unbiased estimator is so that all we need to do is to compare the variance of our estimator with the bound. If it is equal to the bound we know we cannot do any better. The formula for this lower bound is:

$$V(\theta(X,n)) \geq (n\int_{\infty}^{-\infty} (\frac{d \ln f(\theta,x)}{d\theta})^2 f(\theta,x)dx)^{-1}$$

which looks much more complicated than it really is since it is just the inverse of n times the expected value of the square of the derivative of the log of the p.d.f. with respect to the parameter being estimated. It turns out that this is often easy to evaluate. In our example of estimating μ by the sample mean from a normal distribution we have $\ln f(\,) = -0.5\ln(2\pi\sigma^2) - (x-\mu)^2/2\sigma^2$, which has a derivative with respect to μ of $(x-\mu)/\sigma^2$ which when squared is $(X-\mu)/\sigma^4$ and since μ = E(X) and V(X) = $E(X-E(X))^2$ it follows that $E((X-\mu)^2/\sigma^4) = \sigma^2/\sigma^4 = 1/\sigma^2$. Multiplying this by n and inverting yields a lower bound for the variance of an unbiased estimator of μ of σ^2/n. Note that this is the same as the variance of our sample mean which we derived above so we've verified that our estimator has as small a variance as is possible for an unbiased estimator.

Consistency

This property or quality depends upon a mathematical notion known as a probability limit one version of which is:

$$\underset{n\to\infty}{PLim}\,\theta(X,n)=\theta \quad when \quad \lim_{n\to\infty} P(\left|\theta(X,n)-\theta\right|>\in)=0$$

where $\in$ is an arbitrarily small number. Intuitively all this says is that, as the sample size grows without bound, the chance that the estimator will be different from what is being estimated tends to zero. So the idea of consistency has to do with what happens to the estimator as the sample size grows without bound. Another way of thinking about it is that if the estimator ceases to be a random variable, becomes as it were a constant which is equal to the true parameter value being estimated, then the estimator is said to be consistent. In mathematical terms the probability limit of the estimator is the true parameter value being estimated or $\underset{n\to\infty}{P\lim}\,\theta(X,n)=\theta$. There are many types of probability limits, the version presented is the simplest. The easiest way of checking for consistency is to check that the estimator is unbiased and that its variance goes to zero as n becomes infinite. So our sample mean, with an Expected Value of μ which is unaffected by n and hence unbiased as n tends to ∞ and a Variance of σ^2 /n which will tend to 0 as n tends to ∞ will indeed be consistent.

Sufficiency

A sufficient estimator is one that uses all of the information in the sample effectively. This is most easily checked for via what is called the factorization theorem which simply states that the estimator θ(X,n) is sufficient if the joint p.d.f. of the sample can be factorized into the product of two functions one of which contains only the estimator and what is being estimated and the other which only contains the data. That is to say θ(X,n) is sufficient for θ if we can write:

$$(f(X_1,\theta)f(X_2,\theta)f(X_3,\theta)...f(X_n,\theta)) = g(\theta(X,n),\theta)h(X_1,X_2,..,X_n)$$

Example

To exemplify these ideas we consider four estimators of the mean of a normal population μ_1,μ_2,μ_3 *and* μ_4 which are respectively of the form:

$$\mu_1=\frac{1}{n}\sum_{i=1}^{n}X_i\,;\,\mu_2=\frac{1}{n+1}\left(X_1+\sum_{i=1}^{n}X_i\right);\mu_3=\frac{1}{n}\left(1+\sum_{i=1}^{n}X_i\right);\mu_4=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i \;\; where\; n_1<n$$

μ_1 is the sample mean, μ_2 is the sample mean with the first observation counted twice, μ_3 is the sample mean where the researcher has inadvertently included the number 1 in the summation and μ_4 is where the (lazy) researcher has decided to average only the first n1 of his observations.

1. Unbiasedness. Since these are all linear estimators examining them for unbiasedness is a straightforward matter of taking expectations as we did earlier. It is readily observed that μ_1, μ_2 and μ_4 are all unbiased since their expectation is μ, the expected value of is $\mu_3 = \mu + 1/n$ so this is not an unbiased estimator unless the sample size becomes infinite.

2. Efficiency. The variances of the four estimators are respectively $\sigma^2/n, \sigma^2\left[1/n + (n-1)/(n(n+1)^2\right], \sigma^2/n \text{ and } \sigma^2/n_1$ so that only μ_1 attains the Cramer Rao lower bound (μ_3 is not unbiased so its variance should not be compared to the bound) and is thus an efficient estimator.

3. Consistency. Since as n tends to infinity $\mu_1, \mu_2 \text{ and } \mu_3$ are all unbiased with variances that go to 0 they are all consistent. With respect to μ_4, although it is unbiased is not consistent because as n tends to infinity its variance remains constant at σ^2/n1.

4. Sufficiency. Given independent X's and normality, the joint density of the sample may be written as:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i-\mu)^2} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{\sum_{i=1}^{n}-(X_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i^2+2X_i\mu+\mu^2\right)} = g(\mu_1,\mu_2)h(X)$$

Where

$$g(\mu_1,\mu_2) = \left(\frac{1}{2\pi\sigma^2}\right)^n e^{\frac{-n}{2\sigma^2}\left(-2\mu_1,\mu_2+\mu^2\right)} \text{ and } h(X) = e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}X_i^2}$$

Note that the joint density can only be factorized with respect to μ_1 in this fashion none of the other estimators are sufficient statistics.

The Central Limit Theorem

For a random sample $X_i, i = 1,....,n$ from any population with a p.d.f. f(X) such that E(X) = μ and V(X) = σ^2 the sample mean $\bar{X}$ is distributed as N (μ , σ^2 /n) for n sufficiently large enough. This important result removes concern about the underlying distribution of X and shifts attention to the normal distribution and its close relatives when the mean of the population is of interest.

More on Estimation.

In the previous chapter we looked at the properties of estimators and the criteria we could use to choose between types of estimators. Here we examine more closely some very popular basic estimation techniques, two of which focus on the estimation of parameters of a pre-specified probability density function (Maximum Likelihood and Method of Moments techniques) and the third which focuses on the estimation of the shape of an un-specified probability density function (kernel estimation). In all cases we are confronted with a random sample $X_i, i = 1, 2, ..., n.$ and in the first two cases we know the form of the p.d.f. f (X, θ) but not the value of the parameter θ (often there will be more than one parameter, the techniques are readily extended to deal with this situation) in the third case we do not know the form of f all we are trying to calculate is the value of f ( ) for a given x. The third case relates solely to continuous random variables, the first two cases relate to both discrete and continuous random variables, in our discussion we refer only to the continuous case though we will give examples of discrete random variable problems.

Maximum Likelihood Estimation

The intuition behind this technique is to choose a value for the unknown θ that will make the chance of us having obtained the sample we did obtain as big as possible. The rationale for this is that any sample we get is going to be a more likely to be a high probability sample than a low probability sample. Imagine we wish to estimate the average height of males and we randomly sample 4 males from off the street, we would be surprised if all 4 were above 7 feet and similarly we would be surprised if they were all below 4 feet. This is because they are unlikely samples.
We would be a lot less surprised if their heights were between 5 and 6 feet because that would constitute a more likely sample. Thus it makes sense to choose a value for θ which maximizes the probability of having got the sample that we got. Given f(x, θ) and independently drawn Xi;s, the joint density of the sample which is referred to as L, the likelihood, is given by:

$$L = \prod_{i=1}^{n} f\left(X_i, \theta\right)$$

and the estimation technique simply amounts to deriving the formula for θ in terms of the Xi;s which maximizes this function with respect to θ. For technical

reasons (i.e. the algebra is usually easier!) we usually maximize the log of the likelihood. When there is more than one parameter the first order conditions are simply solved simultaneously (see the examples below).

Method of Moments Estimation

The motivation here is quite different from, and somewhat more straightforward than, that for the maximum likelihood method, it relies on common sense. We have seen in an earlier chapter that given f(x,θ) we can obtain a formula for the theoretical mean or expected value of x and we can similarly obtain a formula for the theoretical variance and any other moments of x, for example if x is a continuous random variable we have:

$$E(X) = \int_{-\infty}^{\infty} Xf(X,\theta)\,dx$$

$$V(X) = \int_{-\infty}^{\infty} (X - E(X))^2 f(X,\theta)\,dx$$

The sample mean and sample variance are estimates of E(X) and V(X) respectively so all that is needed is to set one of the formulae to its corresponding sample equivalent and then solve for the value of θ, when there is more than one parameter we choose as many moments as we have parameters to solve for.

Some Examples.

1. The Poisson Distribution provides us with an example of a discrete distribution with one parameter. In this case Xi is an integer ≥ 0 and f(x,λ) is of the form:

$$f(x,\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

and E(x) = λ. In this case the Method of Moments technique is very straight forward, since we simply set our estimator of λ to the sample mean $\bar{X}$. For finding the formula for the Maximum Likelihood estimator the logarithm of the likelihood is given by:

$$\ln L = \ln \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} = \sum_{i=1}^{n} \left(X_i \ln \lambda - \ln(X_i!) \right) - n\lambda$$

and taking the derivative w.r.t. λ and setting it to $\bar{X}$ yields:

$$n = \sum_{i=1}^{n} \frac{X_i}{\lambda}$$

Solving this for λ yields the sample mean of X, (note that in this case the Maximum Likelihood Estimator and the Method of Moments Estimator are the same).

2. The Power Function Distribution provides us with an example of a one parameter continuous distribution. In this case Xi is a number between 0 and 1 and f(x, θ) is of the form:

$$f(x,\theta) = \theta x^{\theta-1}$$

and E(x) = θ/ (θ+1). For the Method of Moments estimator we simply set the formula for E(x) equal to $\bar{X}$ and solve for θ so that our estimator for θ will be $\bar{X} \big/ (1-\bar{X})$. For the Maximum Likelihood estimator the logarithm of the likelihood is given by:

$$\ln L = \ln \prod_{i=1}^{n} \theta X_i^{\theta-1} = \sum_{i=1}^{n} (\ln \theta + (\theta - 1) \ln X_i)$$

Taking the derivative w.r.t. θ and setting to zero yields:

$$\frac{n}{\theta} = \sum_{i=1}^{n} -\ln X_i$$

Re-arranging in terms of θ yields a Maximum Likelihood estimator n/Σ -ln(Xi) which of course is very different from the Method of Moments estimator above.

3. The Normal Distribution provides us with a continuous random variable example of a two parameter problem where the unknown parameters in the distribution are μ and σ^2, the mean and variance respectively. The p.d.f. in this case is given by:

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

Where E(x) = μ and V(x) = σ^2. Again the Method of Moments estimators are trivial, we simply set the estimators of μ and σ^2 equal to the sample mean and sample variance respectively. For the Maximum Likelihood estimators the logarithm of the likelihood is given by:

$$lnL = \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = -\frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x-\mu)^2}{2\sigma^2}$$

Now we have to take the derivatives with respect to both μ and σ^2, set them both to zero and solve the equations simultaneously. Taking the derivatives and setting them to zero after some cancellations yields:

$$n\mu = \sum_{i=1}^{n} X_i$$

$$\frac{n}{\sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^4}$$

Solving these simultaneously yields the sample mean as the Maximum Likelihood estimator for μ (the same as the Method of Moments estimator) and $\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$ /n as the Maximum Likelihood estimator of σ^2 (which is different from the method of moments estimator).

Kernel Estimation.

The issue to be addressed here is the estimation of some unknown density function f(x) which underlies a sample of observations on x. Given such a sample x_i, i = 1,..,n, the generation of a naive estimate of f(x) is straightforward. If the random variable X has the density f(x) then:

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P\left(x - h < h < x + h\right)$$

for given h we can estimate P(x-h < X < x+h) by the proportion of observations falling into the interval x-h, x+h. Letting I( ) be an indicator function where I(z) = 1 if z is true and 0 otherwise, our estimator of f(x) (call it $f_e(x)$) may be written as:

$$f^e(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2h} I\left(\left|\frac{x - x_i}{h}\right| < 1\right)$$

There is a connection with histograms, suppose no x_i lands exactly on the boundary of a bin, then this estimator corresponds to splitting the range of the random variable into bins of width 2h

allowing x to be the "center" of each bin and treating $f_e(x)$ as the ordinate of the histogram. The problem with this type of estimator is that it is not "smooth" but consists of a sequence of jumps at x ± h with a zero derivative everywhere else. Kernel estimators get around this problem by replacing .5I( ) in the above formula by a **kernel function** K( ) with certain desirable properties that to some degree resolve the "smoothness" problem. So that our estimator looks like:

$$f^e(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right)$$

Where h is usually referred to as the band width, window width or smoothing parameter.

The kernel function.

Generally a kernel function would be selected so that it satisfies:

$$\int_{-\infty}^{\infty} K(y)dy$$

making the vast array of continuous density functions suitable candidates. Provided K() is everywhere non negative (making it a density function) $f_e(\ )$ will itself be a density function and will inherit all of the continuity and differentiability properties of K.

Three Examples:
In each of the following h is the bandwidth and $t_i = (x\text{-}X_i)/h$. The **Epanechnikov** Kernel $K_E(t_i)$ is of the form:

$$K_e(t_i) = \frac{.75(1 - .2t_i^2)}{\sqrt{5}} \quad t_i^2 < 5$$
$$= 0$$

The **Gaussian** Kernel $K_G(t_i)$ is of the form:

$$K_g(t_i) = \frac{1}{(2\pi)^{.5}} e^{\frac{t_i^2}{2}}$$

The **Biweight** Kernel $K_B(t_i)$ is of the form:

$$K_B(t_i) = \frac{15}{16}\left(1 - t^2\right)^2 \quad |t| < 1; 0 \text{ otherwise}$$

Note also the original naive estimator is like a rectangular kernel with $K(t_i) = .5|t_i| < 1, 0$ otherwise.

Choosing the 'h' and the Kernel.

Think about the mean integrated squared error of the estimator defined by:

$$MISE(f_e) = E\int\left(f^e(x) - f(x)\right)^2 dx$$

Since the integrand is non-negative the order of integration and expectation can be reversed, note also that

$$E(f_e - f)^2 = E(f_e - E(f_e) + E(f_e) - f)^2 = \left(E(f_e) - f\right)^2 + E(f_e - E(f_e))^2 = bias^2 + \text{var}(f_e)$$

yielding:

$$MISE(f_e) = \int (E(f^e(x)) - f(x))^2 dx + \int \text{var}(f_e(x))dx$$

which is the integrated squared bias plus the integrated variance. This would conceptually be a useful thing to minimize in choosing h and the Kernel. Rewriting K* = K/h it can be noted that:

$$E(f^e) = \frac{1}{n}\sum_{i=1}^{n} E(K^*(t_i)) = \int K^*(t) f(x)dx$$

which, for given f, does not depend upon n but only on K and h. This indicates that taking larger samples alone will not reduce the bias; attention has to be focused on the choice of h and K! Confining attention to Kernels symmetric about zero with continuous derivatives at all orders with a variance v_k, it can be shown that (see Silverman pages 39 to 40) that the optimal h is equal to:

$$v_k^{\frac{-2}{5}} \left(\int K(t)^2 dt\right)^{\frac{1}{5}} \left(\int f''\left(x^2\right)dx\right)^{\frac{-1}{5}} n^{\frac{-1}{5}}$$

Unfortunately "optimal h" here depends upon knowledge of the unknown f( ) we are attempting to estimate, however it does tell us that the optimal window gets smaller as the sample size grows (last term) and as the degree of fluctuation of the unknown function increases (penultimate term). Substituting the value of the optimal h back into the formula for the mean integrated squared error and minimizing with respect to K results in the Epanechnikov Kernel. The relative efficiencies of other kernels can be shown to be .9512 for the Gaussian kernel, .9939 for the Biweight kernel and .9295 for the rectangular suggesting that there is little to choose between kernels on efficiency grounds.

Choosing h.

Referring to the normal family of distributions with a variance σ^2 , yields a value of "optimal h" of 1.06 $\sigma^2 n^{-2}$. One could then estimate σ from the data and, on the presumption that the distribution being estimated was like the normal, use this as the value for h. When the underlying distribution is not normal this tends to result in over-smoothing (especially when bi-modality is present). A safe alternative, based upon a sample standard deviation of σ and a sample interquartile range of ξ , the bandwidth 'h' is specified as:

$$h = \frac{.9\min\left(\sigma, \frac{\xi}{1.34}\right)}{n^{\frac{1}{5}}}$$

Least Squares Cross Validation.

Noting that the integrated squared error may be written as:
and that the last term does not depend upon f_e and hence h interest focuses on minimizing an

approximation to the first two terms with respect to h. After some tedious argument it may be shown that a good approximation to these two terms is:
where K_c is defined as:
and $K_2(t)$ is the convolution of the Kernel with itself. Numerical methods for minimizing this w.r.t. h can easily consume inordinate amounts of time however fourier transform methods can be used to substantially reduce computations (see Silverman P61-66). Non the less the computational burden remains considerable.
Likelihood Cross Validation
Let $f_{-i}(\ )$ be the Kernel estimate calculated by missing out observation x_i then $\ln f_{-1}(x_i)$ is the loglikelihood
of f as the density underlying the independent additional observation x_i. We can think
of maximizing this with respect to h, indeed why not maximize:
This is related to the Kullback-Leibler Information distance $I(f,f_e)$ where:
Thinking of $E(CV(h))$ as the expectation of f_{-j}
$_e$ for some arbitrary j we have:
thus we are minimizing the Kullback - Leibler information distance plus a constant.
Alternatively Silverman suggests eyeballing the problem. Plot out a selection of curves based upon different h's and choose the one that best suits ones priors.

A variable bandwidth h: The Adaptive Kernel.

One of the problems with the above estimators is that the degree of smoothing is constant over all x, the same value in regions densely populated with x as it is in regions sparsely populated with x. This can lead to over-smoothing in the dense areas (taking out "bumps" that should be there) and / or under-smoothing in sparse areas (leaving in bumps that should not be there). To solve this problem a variable bandwidth estimator has been developed. Essentially in this case our estimator f_{ae} is of the form:
Where h_i is usually referred to as the local band width, window width or smoothing parameter. Various methods are available for estimation in this case, one of the simplest, most practical and most effective is the following. Compute f_e in one of the preceding methods which yields a fixed bandwidth h. Calculate f_{gm} the geometric mean of f_e for the sample given by:
set h_i as:
where a is a sensitivity parameter chosen by the investigator. Generally 0 # a # 1 with a=0 returning us to the fixed bandwidth estimator. Most applications seem to choose a = 0.5.

Consistency

Under apparently very mild conditions on the kernel namely,
together with $*tK(t) * 6\ 0$ as $* t* 64$ and a window width h_n satisfying $h_n 6\ 0$ as $nh_n 6\ 4$ convergence in probability of $f_e(x)$ to f(x) (convergence at a point) can be established. Essentially the requirement on h is that it does not converge to 0 as rapidly as n^{-1} ensuring the expected number of points in $x \pm h_n$ tends to infinity with n. Further, and more importantly, under similar conditions $\sup_x * f_e(x) - f(x) *$ can also be shown to converge to 0. A note of caution, the rate of convergence is often very slow so that in this instance consistency is by no means a warranty for good estimates!