LANGEBIO - BIOSTATISTICS

OCTAVIO MARTÍNEZ DE LA VEGA

10. Statistical Inference: Parameter Estimation

So far we have been learning some concepts that are needed for Statistics, but we have not, I am afraid to say, learning Statistics. Statistics begin when we board a real situation where we do not know the facts; we do not know the distribution of the variables, we do not know the parameters that determine their distributions, we do not know the mean or variance... we know almost nothing. But we have data. Then we must make some "assumptions", try to justify them and then obtain some facts about the phenomenon of interest. Then, we are doing Statistics!

In this section we are going to define a "statistic" and then board the main problem of Statistics, i.e., the estimation of (unknown) parameters. We will quickly review the "Method of Moments" and then, more deeply, the method of Maximum Likelihood for this problem. This will help us to understand Hypothesis Testing and, parallel to this, the estimation of confidence intervals, of which we have seen an example in the previous section.

10.1. **Definition. Statistic.** Let $\{X_i\}_{i=1}^n$ be a set of n independent and identically distributed random variables. Then, a statistic is function of these variables, say

$$f(X_1, X_2, \ldots, X_n)$$

which is itself a random variable and does not contain any unknown parameters.

Examples of statistics are the arithmetic mean,

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

the sampling variance,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}(X_i - \overline{X})^2$$

Date: April 2012.

10.2. **Definition. Sample Moments.** Consider a sample, $\{X_i\}_{i=1}^n$, of some distribution (defined as above). Then we define the $r-th$ moment about 0 as

$$M'_r = \frac{1}{n}\sum_{i=1}^{i=n} X_i^r$$

and the $r-th$ moment about the mean as

$$M_r = \frac{1}{n}\sum_{i=1}^{i=n}(X_i - \overline{X})^r$$

Note that $\overline{X} = M'_1$, the first sampling moment about zero, while $S^2 = \frac{n}{n-1}M_2$.

These function are useful because the expectation of the sampling moments about 0 are equal to the $r-th$ population moment (if that exist). That is, these functions help us to estimate a characteristic of the population.

In particular, $\overline{X}$ and S^2 are **unbiased** estimators of their corresponding population parameters, say

$$E[\overline{X}] = \mu \text{ and } E[S^2] = \sigma^2$$

thus we say that $\overline{X}$ and S^2 are **unbiased** estimators of the respective (population) parameters μ and σ^2.

In general, an *estimator* is a function of the observations that help use to estimate a parameter.

10.3. **Maximum Likelihood.** The information that we have about the unknown random variables come to us trough **data**, that is, from a set of *realization* of the random variable that we are observing; we call that set of realization a sample of the random variable. If we are sure about the process that is generating the random variable, that is if we do not doubt which is the distribution, the *only* thing that we need to know is the parameter or parameters that completely determine such distribution.

How to *estimate* this parameter (or parameters)? A powerful idea is to take as estimates the value of the parameter that are more *likely*. This is simply to accept that our data are "typical" or that they fairly represent the "state of nature" or, to put it in another way, that we did not obtained "strange" or "unlikely" observations of the phenomenon of interest.

Let's exemplify with some real data. In an experiment with *Capsicum*, Garcia-Neria and Rivera-Bustamante (2011; MPMI, Vol. 24, pp. 172-182) crossed plants resistant to a virus with susceptible plants, obtaining 156 resistant and 130 susceptible descendants plants. Without going into the genetics details (we will be back to that later) we ask two questions: 1) How can we statistically model the random variable X : Number of resistant plants (from the same cross) and 2) Which is the

value of the parameter that determine its distribution? It is natural that the answer to our first question is to say that X has a Binomial distribution, and taking the total number of descendants, $k = 156 + 130 = 286$ we can say that

$$X \sim \mathcal{B}(k = 286, p)$$

thus the only unknown parameter is p, the probability of "resistant" (success). Which will be **YOUR** estimate? Very likely you will say

$$\hat{p} = \frac{156}{286}$$

But, why?

Well because *common sense* dictates that it is our best guess, with the available information. In fact, you have done your first *maximum likelihood* estimation!

Let's see. Given that $X \sim \mathcal{B}(k = 286, p)$ we know that

$$P[X = x] = C_x^k p^x (1-p)^{k-x} = \frac{k!}{x!(k-x)!} p^x (1-p)^{k-x}$$

Now, we know that in fact $k = 286$ and $x = 156$, and then in the previous equation the only unknown is p. Lets rename our probability function thinking of it as a *likelihood* function of p, say

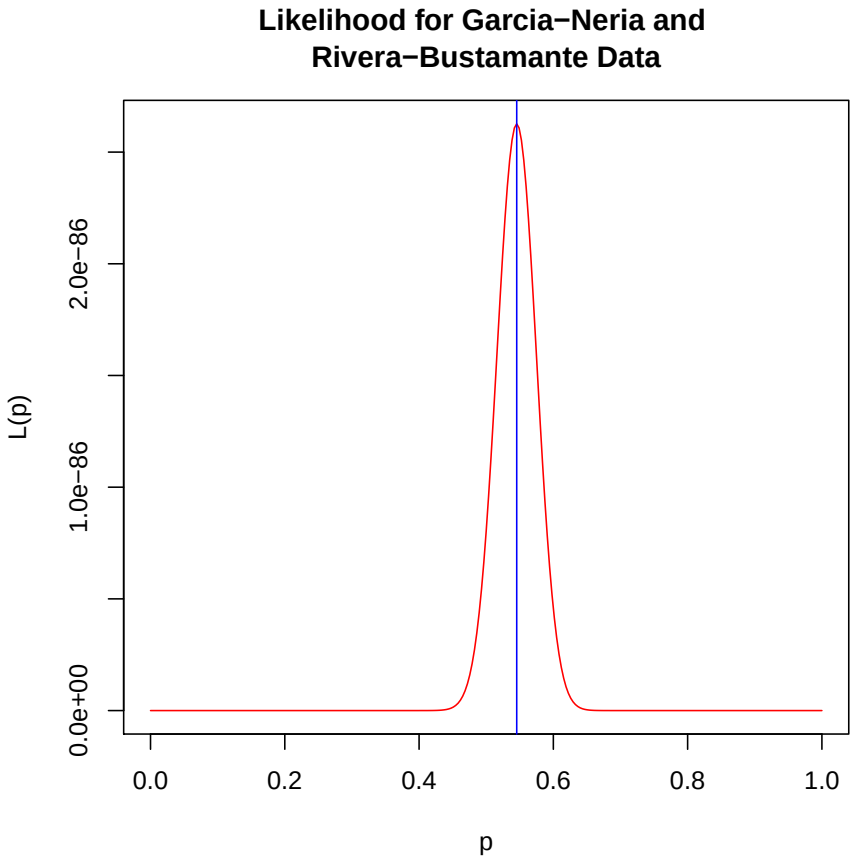
$$L(p) = C_x^k p^x (1-p)^{k-x} = C_{156}^{286} p^{156} (1-p)^{130}$$

Note that C_{156}^{286} is a constant that does not depend on p and thus we can ignore it. Thus

$$L(p) = p^{156} (1-p)^{130}$$

Let's graph this function in R for all possible values of p.

```
> choose(286, 156) # The huge constant that we are ignoring
[1] 1.802136e+84
> p <- c(0:286)/286 # A grid of values of p
> plot(p, (p^156)*((1-p)^130), type="l", col="red",
+ xlab="p", ylab="L(p)", main=
+ "Likelihood for Garcia-Neria and \nRivera-Bustamante Data")
> # Let's put a vertical line in p=156/286
> abline(v=156/286, col="blue")
```



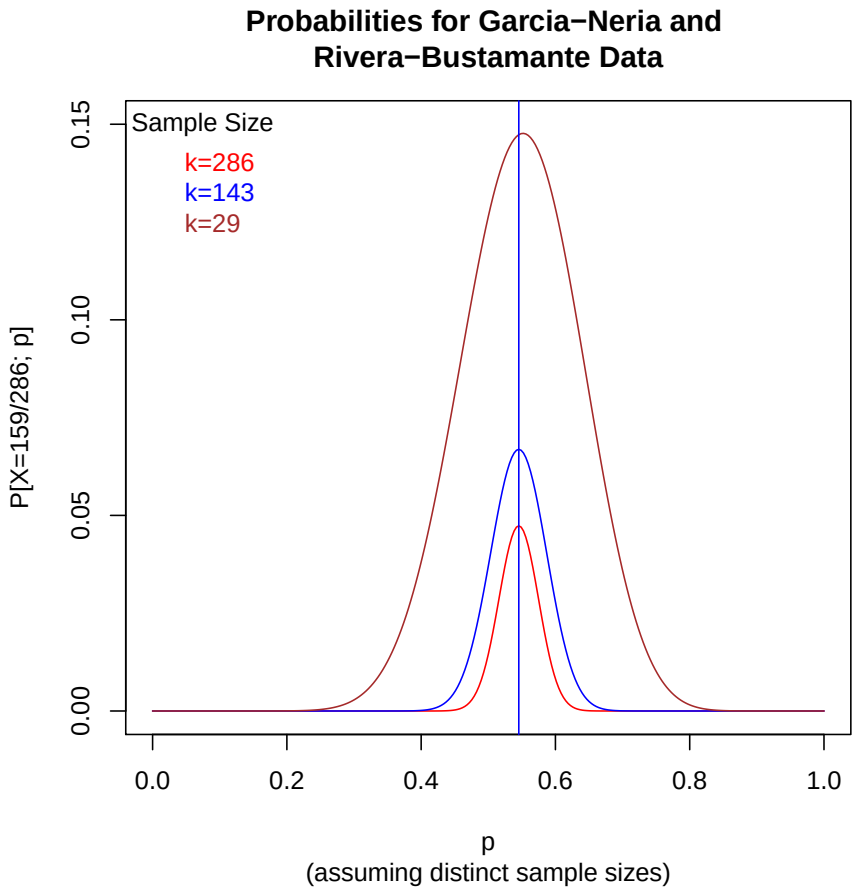
Note how the maximum of the likelihood function coincide, perfectly with our "common sense" estimation of p, say

$$\hat{p} = \frac{156}{286}$$

Also note how the likelihood decreases sharply around the point of maximum. This indicates that there is **very strong** evidence that the true probability is near $156/286 = 0.5454$. Let see how the likelihood function is modified if equivalent results were obtained with less sample size; for example if only $286/2 = 143$ descendants were obtained. We are going to graph the probabilities (that is, we are going to include the constants C_x^{k-x}) to make the graphs comparable.

```
> plot(p, dbinom(x=156, size=286, prob=p),
+ ylim=c(0, 0.15),
+ type="l", col="red",
+ xlab="p", ylab="P[X=159/286; p]", main=
```

```
+ "Probabilities for Garcia-Neria and \nRivera-Bustamante Data",
+ sub="(assuming distinct sample sizes)")
> # Now assuming a sample size of k/2=286/2=143
> points(p, dbinom(x=156/2, size=286/2, prob=p),
+ type="l", col="blue")
> # Now assuming a sample size of k/10=28.6/2=29
> points(p, dbinom(x=round(156/10),
+ size=round(286/10), prob=p),
+ type="l", col="brown")
> abline(v=156/286, col="blue")
> legend("topleft", title="Sample Size",
+ legend=c("k=286","k=143", "k=29"),
+ text.col=c("red", "blue", "brown"),
+ bty="n", title.col="black")
```



Note that for smaller sample size (k) the "large" probabilities are spread in a **longer** rank; that is, we are "less secure" of the true value of p.

There is an analytical way to obtain this estimator. Remember how to obtain the maximum of a function? You can obtain the first derivative of the function and equal it to zero. That is a critical point, and by the criteria of the second derivative you can see if it is a maximum or a minimum. We can use instead of $L(p)$ its natural logarithm, say

$$l(p) = \log(P[X = x]) = \log(C_x^k) + x\log(p) + (k - x)\log(1 - p)$$

Now we obtain

$$\frac{d}{dp}(l(p)) = x\frac{d}{dp}(\log(p)) + (k - x)\frac{d}{dp}(\log(1 - p))$$

$$= \frac{x}{p} - \frac{k - x}{1 - p}$$

and we set

$$\frac{d}{dp}(l(p)) = \frac{x}{\hat{p}} - \frac{k - x}{1 - \hat{p}} = 0$$

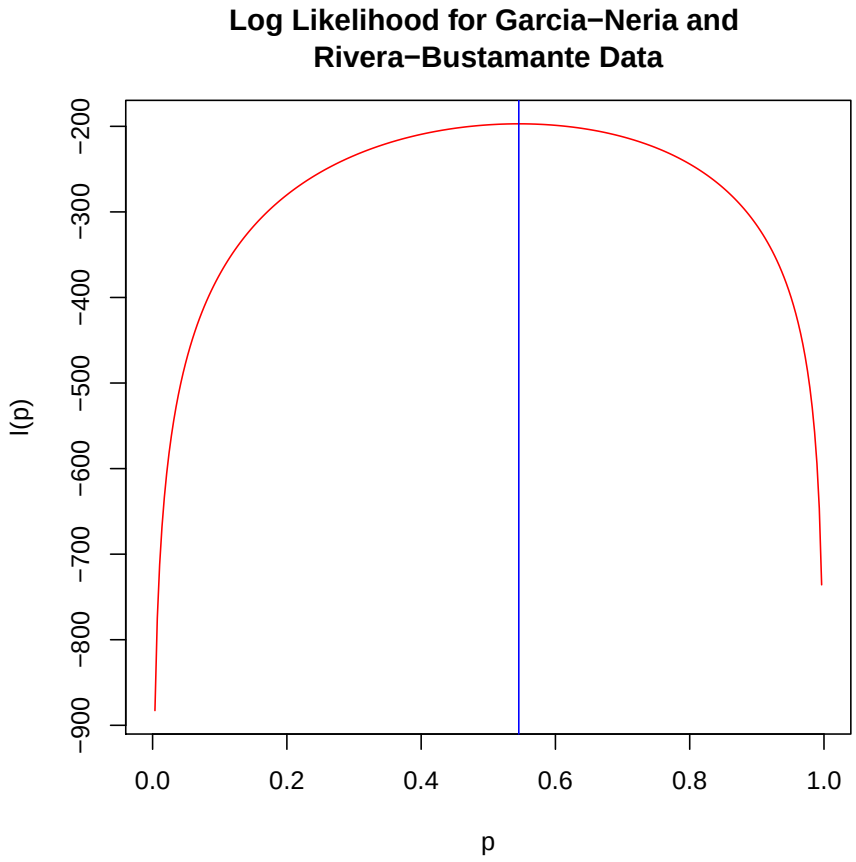
Note that here I use $\hat{p}$ instead of simply p, to denote that the value which will be obtained (the one that satisfy the above equation) is a **maximum likelihood estimator** of p. Of course the only value that satisfy this equation is:

$$\hat{p} = \frac{x}{k}$$

Normally we do not go into the trouble of finding the **maximum likelihood estimator** of a parameter; the Statisticians have done the work for us. However, it is crucial to understand the principles involved.

Let now graph $l(p)$ for the previous data, and see that it maximum again, corresponds to $\hat{p}$.

```
> plot(p, 156*log(p) + 130*log(1-p), type="l", col="red",
+ xlab="p", ylab="l(p)", main=
+ "Log Likelihood for Garcia-Neria and \nRivera-Bustamante Data")
> # Let's put a vertical line in p=156/286
> abline(v=156/286, col="blue")
```



Note that, given the *log* transformation, the maximum of the function is graphically less clear, but still at the same point.

10.4. **Theorem. Properties of Maximum Likelihood Estimators (MLE).** There are different ways to select estimators, but the method of Maximum Likelihood gives estimators which have good properties. Here I will enumerate them without proof.

Let $\{X_i\}$ a sample of size n of a distribution that depends on a parameter θ, and let $\hat{\theta}$ be the MLE of the parameter. Then under regular conditions the MLE has the following properties:

- 1. Consistency: $\hat{\theta}$ converges in probability to the true value of the parameter θ, that is, for a small constant ε

$$\lim_{n\to\infty} P[\hat{\theta} - \varepsilon \leq \theta \leq \hat{\theta} + \varepsilon] = 1$$

- 2. Asymptotic normality and *small* variance: As the sample size n increases, $\hat{\theta}$ tend to have a normal distribution with mean θ and a variance which will be the "smallest possible", say $v = V[\hat{\theta}]$ will be smaller that the variance of "other" estimators. In symbols, as n increases:

$$\hat{\theta} \sim \mathcal{N}(\theta, v)$$

10.5. **Example. MLE of $E[X]$.** Let $\{X_i\}$ a sample of size $n = 10$ of the normal distribution of unknown μ and $\sigma^2 = 1$, that is

$$X_i \sim \mathcal{N}(\mu, 1)$$

It happens that the MLE of μ in this case is the arithmetic mean, $\overline{X}$. In this excercise we will try to find a better estimate of μ than $\overline{X}$; of course, we will not, but possibly we will learn something.

Here is my list of putative estimators:

$$g_1(X_1, \ldots, X_{10}) = \overline{X}$$

$$g_2(X_1, \ldots, X_{10}) = X_1$$

$$g_3(X_1, \ldots, X_{10}) = median\{X_1, \ldots, X_{10}\}$$

$$g_4(X_1, \ldots, X_{10}) = \frac{max\{X_1, \ldots, X_{10}\} + min\{X_1, \ldots, X_{10}\}}{2}$$

$$g_5(X_1, \ldots, X_{10}) = \frac{1X_1 + 2X_2 + \ldots + 10X_{10}}{55} = \frac{\sum_{i=1}^{10} iX_i}{\sum_{i=1}^{10} i}$$

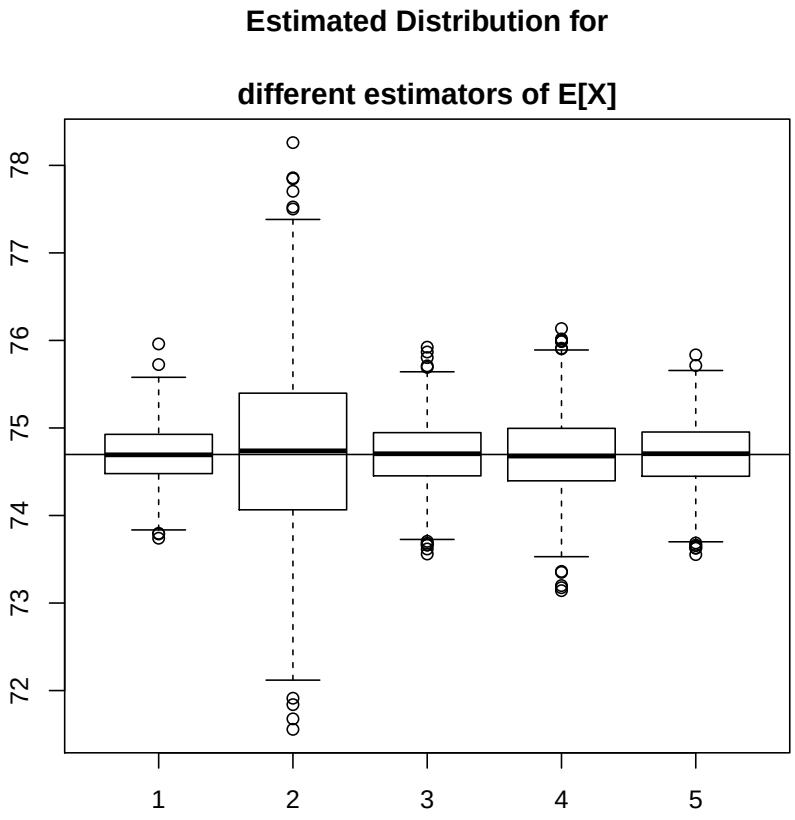
Do you have any objections against my estimators?. Discuss.

How are we going to judge if an estimator is better than other?. We can study their distribution, in particular their expectation and variance to see which one is "better". Of course, the sensible thing is to try this problem analytically, but let's play with simulation in R.

```
> # Programing the estimators
> my.estim <- function(x){
+ if(length(x) != 10){ stop("I want a vector of size 10!")}
+ res <- c(mean(x), x[1], median(x), (max(x)+min(x))/2, sum(c(1:10)*x/55))
+ names(res) <- paste("g", c(1:5), sep="")
+ res
+ }
> # Trying my function
> my.estim(c(1:10))
 g1  g2  g3  g4  g5
5.5 1.0 5.5 5.5 7.0
```

```
> # Now lets simulate 1000 instances of samples
> # of size n=3 from the normal distribution,
> # with a "secret" mean and try our estimators.
> secret <- runif(1, min=0, max=100) # secret E[X]
> my.sim <- data.frame(rep(NA,5),
+ rep(NA,5),rep(NA,5),rep(NA,5),rep(NA,5))
> for(i in 1:1000){
+ my.sim[i,] <- my.estim(rnorm(10, mean=secret, sd=1))
+ }
> names(my.sim) <- paste("g", c(1:5), sep="")
> # Now, let's see which estimator is "better":
> secret # The secret value of the parameter: E[X_i]
[1] 74.69716
> # Let's see the means of the 1000 realizations
> apply(my.sim, 2, mean)
      g1       g2       g3       g4       g5
74.69959 74.73830 74.69947 74.70425 74.70075
> # And the absolute difference with the true value
> abs(apply(my.sim, 2, mean) - secret)
         g1          g2          g3          g4          g5
0.002423715 0.041139311 0.002302841 0.007089737 0.003584555
> # The rank
> apply(my.sim, 2, max) - apply(my.sim, 2, min)
      g1       g2       g3       g4       g5
2.219892 6.703011 2.363119 2.994035 2.281532
> # The (sampling) variance
> apply(my.sim, 2, var)
       g1        g2        g3        g4        g5
0.1040039 1.0037653 0.1364391 0.2068284 0.1353217
> # And it's square root
> apply(my.sim, 2, sd)
       g1        g2        g3        g4        g5
0.3224964 1.0018809 0.3693766 0.4547839 0.3678610
> # And finally a summary of their distributions
> boxplot(list(my.sim$g1, my.sim$g2, my.sim$g3,
+ my.sim$g4, my.sim$g5),
+ main="Estimated Distribution for\n
```

```
+ different estimators of E[X]",
+ sub="(sample size n = 10)")
> abline(h=secret)
```

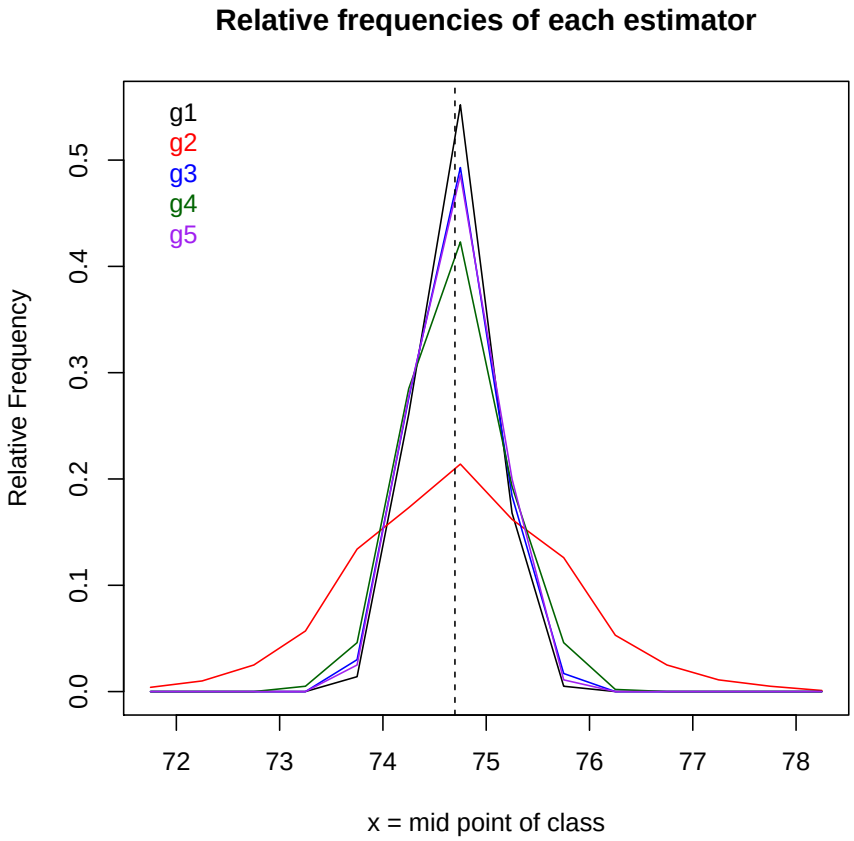


In average, which estimator is closest to the true value of the parameter?

Which of the estimators gives the most compact distribution? That is, which is less variable?

```
> # A different plot
> temp <- hist(my.sim$g2, plot=FALSE)
> my.classes <- temp$breaks
> n.classes <- length(my.classes) - 1
> my.freqs <- matrix(NA, nrow=n.classes, ncol=5)
> for(i in 1:n.classes){
+ for(j in 1:5){
+ my.freqs[i, j] <-
```

```
+ sum(1*(my.sim[,j]>=my.classes[i])&
+ (my.sim[,j]<my.classes[i+1]))
+ }}
> attributes(my.freqs)$dimnames[[2]] <- names(my.sim)
> attributes(my.freqs)$dimnames[[1]] <- temp$mids
> plot(temp$mids, my.freqs[,1]/1000, type="l",
+ ylim=c(0, max(my.freqs/1000)),
+ ylab="Relative Frequency", xlab="x = mid point of class",
+ main="Relative frequencies of each estimator")
> points(temp$mids, my.freqs[,2]/1000, type="l", col="red")
> points(temp$mids, my.freqs[,3]/1000, type="l", col="blue")
> points(temp$mids, my.freqs[,4]/1000, type="l", col="darkgreen")
> points(temp$mids, my.freqs[,5]/1000, type="l", col="purple")
> legend("topleft", bty="n",
+ legend=names(my.sim),
+ text.col=c("black","red","blue","darkgreen","purple"))
> abline(v=secret, lty=2)
```



What is your conclusion about which is the "best" estimator of $E[X_i]$?

10.6. **Summary and Comments.** One of the central goals of Statistics is to make inferences using our data. To begin we can have a model for a random variable, that is sensible because it is based on reasonable assumptions. But then in general this model will have "unknown" parameters which is necessary to "estimate". Here we have seen a method for estimation, Maximum Likelihood, that is intuitively appealing; we can summarise this principle as

> *Take as estimation of the parameters the values that make it more* ***likely*** *that the data that you have, have arisen.*

To do this we look at the probability (or density) function, $f_X(x, \theta)$, not in its normal role but we assume that the value of X is known and look what happens if we vary the parameter, θ. That is we look at the likelihood function, $L(\theta) = f_X(x, \theta)$, and we take has estimate of the parameter, say $\hat{\theta}$, the value that maximize that

function, say

$$\hat{\theta} = max(L(\theta))$$

In many cases this can be done explicitly and thus we get a *formula* for the MLE which is obtained equaling to zero the first derivative of $L(\theta)$. Basically the same can be done when the distribution depends upon more than one parameter.

10.6.1. *Homework.*

- 1. Show that S^2, the sampling variance, can also be written as

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

- 2. For the estimator g_2 show that $V[g_2] > V[g_1] = V[\overline{X}]$.
- 3. Find $V[g_5]$ (use the properties of the variance seen before) and see if it is true that $V[g_5] > V[g_1] = V[\overline{X}]$.
- 4. Think about the general case of the estimators in Example 10.5, where n, the sample size, can take any value, and as in the example $\sigma^2 = 1$. Try to prove that, for any value of n, $V[g_5] > V[g_1] = V[\overline{X}]$. You can find useful the Jensen's inequality (see *http* : *//en.wikipedia.org/wiki/Jensen_Inequality*). You can possibly illustrate this fact plotting the following function:
  `> squares <- function(n=1){((sum(c(1:n)^2)/(sum(c(1:n)))^2))-(1/(n^2))}`
  Find if it true that

$$V[g_5] = V[g_1]$$

  when $n = 1$ (trivial case) and, more interesting if

$$\lim_{n\to\infty}(V[g_5]) = V[g_1] = V[\overline{X}]$$

Have fun!