LANGEBIO - BIOSTATISTICS

OCTAVIO MARTÍNEZ DE LA VEGA

9. The Central Limit Theorem and an Approximate Confidence Interval for $E[X]$

We are going to see a practical application of what we have learned so far. I will let the more formal definition of statistic and estimator for the next section.

We have seen (without proof, but demonstrating it in R) the Central Limit Theorem, that tell us the approximate distribution of the mean. Now, let's advance one step more in this path, and look for some characteristics of the noble arithmetic average, the mean.

9.1. **Theorem. Expectation and Variance of $\overline{X}$.** Let

$$\{X_1, X_2, \ldots, X_n\}$$

be n independent random variables equally distributed such that

$$E[X_i] = \mu \text{ and } V[X_i] = \sigma^2$$

Then the mean or arithmetic average of those variables is given by

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

Then we will have

$$E[\overline{X}] = \mu$$

$$V[\overline{X}] = \frac{1}{n}\sigma^2$$

Proof. Using the properties of $E[\bullet]$ and $V[\bullet]$ that we have seen we can find

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

$$= \frac{1}{n}E[\sum_{i=1}^{i=n} X_i]$$

Date: April 2012.

$$= \frac{1}{n}\sum_{i=1}^{i=n} E[X_i]$$

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

$$= \frac{1}{n} n\mu = \mu$$

and

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

$$= \frac{1}{n^2} V[\sum_{i=1}^{i=n} X_i]$$

$$= \frac{1}{n^2}\sum_{i=1}^{i=n} V[X_i]$$

(Note: this last step is true only if the variables involved are **independent**).

$$= \frac{1}{n^2}\sum_{i=1}^{i=n} \sigma^2$$

$$= \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2$$

This imply to things; one, the mean is an textbfunbiased estimator of μ (the true expectation) and two, the variance of the mean decreases lineally as function of the sample size (n).

9.2. **Definition. Standard Error of the Mean.** Let $\overline{X}$ be as above, then we define the *Standard Error of the Mean*, denoted as $ee(\overline{X})$ as

$$ee(\overline{X}) = \sqrt{V[\overline{X}]} = \sqrt{\frac{1}{n}\sigma^2} = \frac{1}{\sqrt{n}}\sigma$$

The above theorem is completely general; it works for samples of any distribution (continuous or discrete), given that $E[X_i]$ and $V[X_i]$ are finite, and its very useful because it tell us that with larger samples we have more **precision** (less uncertainty) in our estimation.

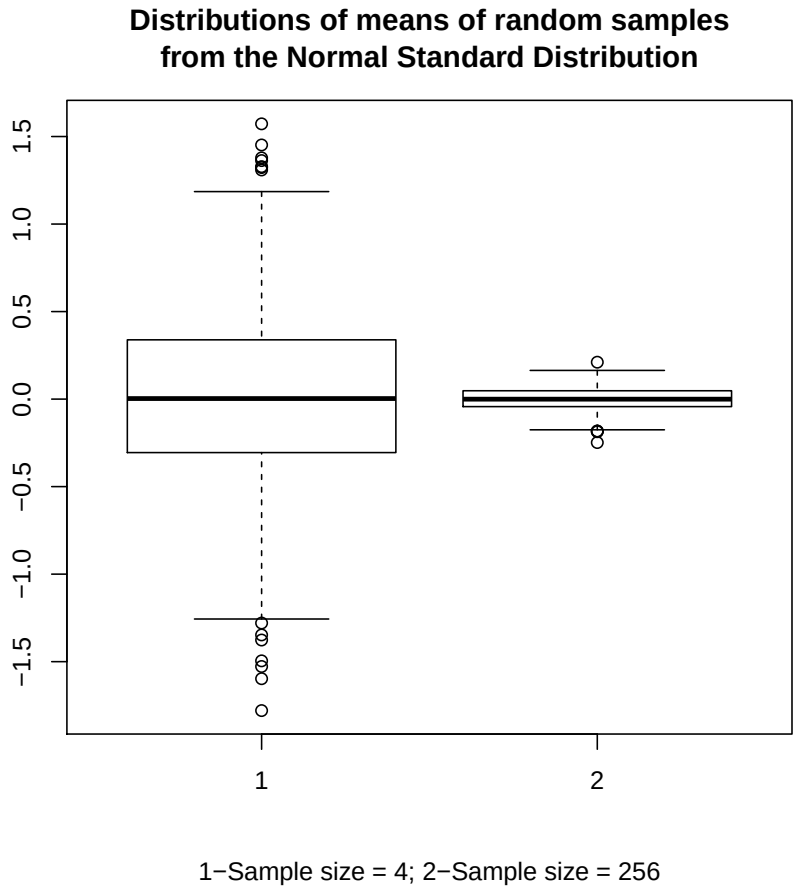
We want to illustrate two facts:

Firstly: In *average* the mean of a set of samples will be close to the **true** mean $E[X_i] = \mu$, and

Secondly: With larger sample size (larger n) we get more precision in the estimation (smaller variance of the mean, $V[\overline{X}] = \frac{1}{n}\sigma^2$).

Let's simulate some numbers for the normal distribution of known parameters $E[X_i] = 0$ and $V[X_i] = 1$

```
> # Lets do two simulations of 1000 samples from the
> # normal distribution with mean=0 and sd = 1
> # The first using n=4 and the second using n=256
> s.n4 <- rep(NA, 1000) # For the n=4 samples
> s.n256 <- rep(NA, 1000) # For the n=256 samples
> for(i in 1:1000){
+         s.n4[i] <- mean(rnorm(4))
+         s.n256[i] <- mean(rnorm(256))
+ }
> mean(s.n4); var(s.n4) # Theoretical values: 0 and 1/4 = 0.25

[1] 0.01241221

[1] 0.2496332

> mean(s.n256); var(s.n256) # Theoretical values: 0 and 1/256 = 0.00390625

[1] 0.002071344

[1] 0.003978513

> # Equivalently, less calculate the (estimated) ee
> # Theoretical values:
> # ee(mean n=4) = 1/sqrt(4) = 1/2
> # ee(mean n=256) = 1/sqrt(256) = 1/16 = 0.0625
> sqrt(var(s.n4)) # Near 0.5?

[1] 0.4996331

> sqrt(var(s.n256)) # Near 0.0625?

[1] 0.06307546

> # Let's see and compare the distributions
> boxplot(list(s.n4, s.n256),
+ main="Distributions of means of random samples\n
+ from the Normal Standard Distribution",
+ sub="1-Sample size = 4; 2-Sample size = 256")
```



Now we will put together the Central Limit Theorem and our knowledge of the expectation and variance of the mean to obtain **approximate** confidence intervals for the mean.

9.3. **Approximate Confidence Intervals for the Mean.** Let

$$\{X_1, X_2, \ldots, X_n\}$$

be a sample of n independent random variables equally distributed such that

$$E[X_i] = \mu \text{ and } V[X_i] = \sigma^2$$

Then

$$\overline{x} \pm 2\widehat{ee}(\overline{x})$$

gives an **approximate** Confidence Interval for the true value of the mean ($E[\overline{X}] = \mu$).

In words, if for a given sample we estimate the mean, $\overline{x}$ and its standard error, $\widehat{ee}(\overline{x})$, then it is *likely* (about 95%) that

$$\mu \in (\overline{x} - 2\widehat{ee}(\overline{x}), \overline{x} + 2\widehat{ee}(\overline{x}))$$

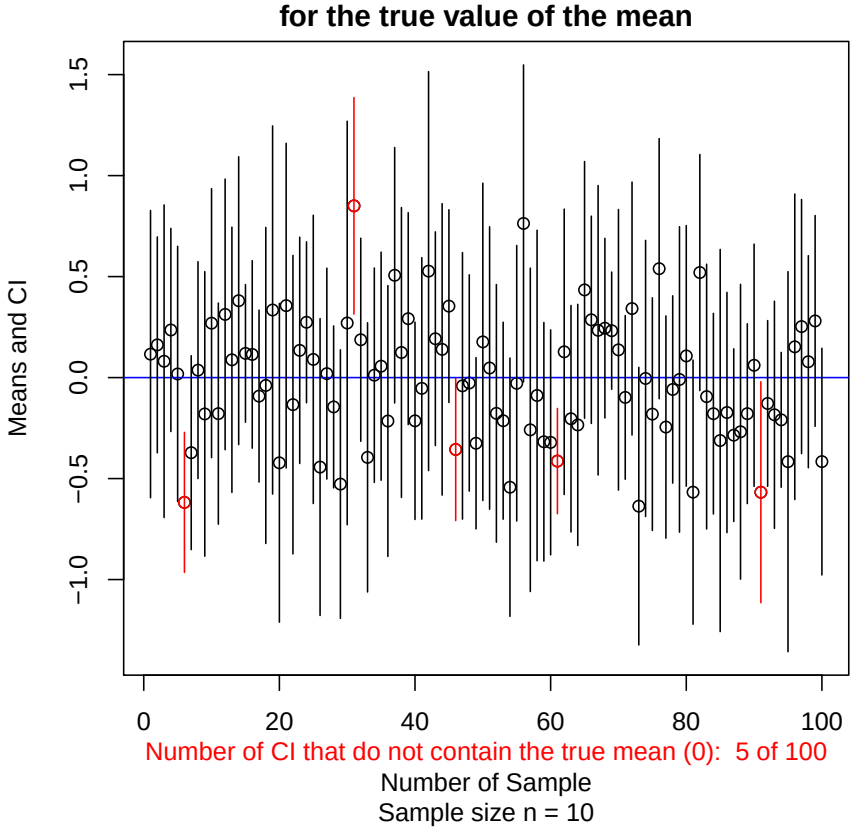
More formally, if we repeat this process many times, then in about 95% of the cases it will be true that

$$\mu \in (\overline{x} - 2\widehat{ee}(\overline{x}), \overline{x} + 2\widehat{ee}(\overline{x}))$$

Let's see that in R.

```
> # Let's program a function to estimate approx CI:
> approx.ci <- function(x){
+         n <- length(x) # Number of data
+         m <- mean(x) # Estimated mean
+         S <- sd(x) # Estimated Standard Deviation
+         ll <- m - (2*S/sqrt(n)) # Lower limit
+         ul <- m + (2*S/sqrt(n)) # Upper limit
+         res <- c(m, ll, ul)
+         names(res) <- c("mean", "LL", "UL")
+         res
+ }
> approx.ci(rnorm(10)) # Testing the function (m=0)
      mean         LL         UL
 0.2758605 -0.1359509  0.6876719
> # Now, lets obtain 100 CI
> my.ci <- data.frame(rep(NA, 100), rep(NA, 100), rep(NA, 100))
> for(i in 1:100){
+         my.ci[i,] <- approx.ci(rnorm(10))
+ }
> names(my.ci) <- c("mean", "LL", "UL")
> # And let's do a nice graph of each one of the 100 CI
> plot(c(1:100), my.ci$mean,
+ ylim=c(min(my.ci$LL), max(my.ci$UL)),
+ xlab="Number of Sample",
+ ylab="Means and CI",
+ main="Approximate Confidence intervals\n
+ for the true value of the mean",
+ sub="Sample size n = 10")
> abline(h=0, col="blue")
> # Now, let's plot the intervals.
> # Black if they include the true mean (0)
> # and red id they do not.
```

```
> my.out <- 0
> for(i in 1:100){
+ if(my.ci$LL[i]>=0 || my.ci$UL[i]<=0){ # do not include 0
+         my.out <- my.out + 1
+         points(i, my.ci$mean[i], col="red")
+         segments(x0=i, y0=my.ci$LL[i], y1=my.ci$UL[i], col="red")
+ } else { # Interval contains 0
+         segments(x0=i, y0=my.ci$LL[i], y1=my.ci$UL[i])
+ }}
> mtext(paste("Number of CI that do not contain the true mean (0): ",
+ my.out, "of 100"), side=1, line=2, col="red")
```



Where this formula for the approximate confidence intervals comes from?. Remember that we have seen that the Central Limit Theorem states that

$$\overline{X} \Rightarrow \mathcal{N}(E[\overline{X}], V[\overline{X}])$$

and we have seen that

$$E[\overline{X}] = \mu \text{ and } V[\overline{X}] = \tfrac{1}{n}\sigma^2$$

Thus for n "large" we can assume that

$$\overline{X} \sim \mathcal{N}(\mu, \frac{1}{n}\sigma^2)$$

Now, assuming the normal distribution, we can calculate a probability region centered into μ which covers around 95% of the probability. Such region is, precicely

$$P[\overline{x} - 1.96\frac{1}{n}S^2 \leq \mu \leq \overline{x} + 1.96\frac{1}{n}S^2] \approx 0.95$$

Note that the number in the previous equality is 1.96 and NOT 2. I use 2 instead of the (more precise) 1.96 because I want you to **MEMORIZE** that formula. It is quite handy; for example, if in a paper the authors give you the mean and standard error of the mean (or S and n) you can calculate a quick and dirty (approximate) confidence interval for the **TRUE** value of the mean.

How "large" needs n to be for this formula to give good results depends upon the (usually unknown) distribution of the original variables $\{X_i\}$. If this distribution is normal (or near to normal), then the "approximation" is exact. If the distribution is very different from the normal you will need a "larger" sample size (n) for it to work, but in general terms it works fairy well for $n \geqq 20$.

9.3.1. *Homework.* For the distributions

- a) Uniform in $[0, 1]$, $(E[X_i] = 0.5)$,
- b) Binomial with $k = 4$, $p = 0.1$, $(E[X_i] = 0.1 * 4 = 0.4)$ and
- c) Poisson with $\lambda = 4$, $(E[X_i] = 4)$

try the sample sizes $n = 2$, $n = 20$ and $n = 100$ and, in each case make 100 simulations of confidence intervals (you can use the function **approx.ci**) and note how good is our formula for confidence intervals, that is in how many cases the intervals include the real value of $E[X_i]$.

Have fun!