Preliminary Statistics September 2013

Answers to class exercise

Problem Set 4: Estimation and Confidence Intervals

1. (a) $Y \sim IN(\mu, \sigma^2)$ means the random variable Y is distributed independently normal with expected value μ and variance σ^2. Each realisation of Y does not carry any information about the other realisations. The realisations are independent from one another, hence their covariances are zero.

   (b) The sample mean, $\hat{\mu} = \sum_{i=1}^{n} \frac{Y_i}{n}$ is an unbiased estimator of the expected value μ. The sample variance, $s^2 = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n-1}$ is an unbiased estimator of the variance σ^2. An estimator of a parameter is unbiased if the expected value equals the true value of the parameter, e.g. $E(\hat{\mu}) = \mu$.

   (c) The standard Error of $\hat{\mu}$ is: $SE(\hat{\mu}) = \sigma/\sqrt{n}$. Since σ is an unknown parameter we can estimate it using *s*. Hence, an estimate of the standard error would be $\widehat{SE(\hat{\mu})} = s/\sqrt{n}$.

   (d) If the distribution was highly skewed, it would not be symmetrical. For instance, income is skewed to the right i.e. there is a long tail to the right covering high incomes. Consequently the mean (due to its sensitivity to outliers) is greater than the median which is greater than the mode. Since mean is driven by the extreme high values, the median would be a better measure of central tendency. With most people earning below the average (mean) income, the median would better represent the population income.

2. a) We need to find the expected value and variance of each of the three estimators.
   (i) The first estimator for the mean is unbiased as $E(\hat{\mu}) = \mu$.

$$E(\hat{\mu}) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n}E(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}[E(X_1) + E(X_2) + \cdots + E(X_n)]$$
$$= \frac{1}{n}n\mu = \mu$$

   (ii) The second estimator is downward biased as it underestimates the true value of the population mean with $E(\hat{\mu}) < \mu$.

$$E(\tilde{\mu}) = E\left(\frac{\sum X_i}{n+1}\right) = \frac{1}{n+1}E(X_1 + X_2 + \cdots + X_n)$$
$$= \frac{1}{n+1}[E(X_1) + E(X_2) + \cdots + E(X_n)] = \frac{1}{n+1}n\mu < \mu$$

(iii) The third estimator is unbiased as $E(\hat{\mu}) = \mu$.

$$E(\check{\mu}) = E(X_1) = \mu$$

b) Regarding the estimators' variances:

(i) The firth estimator is biased and has the smallest variance among the unbiased estimators:

$$Var(\hat{\mu}) = Var\left(\frac{\sum X_i}{n}\right) = Var\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}Var(X_1 + X_2 + \cdots + X_n)$$
$$= \frac{1}{n^2}[Var(X_1) + Var(X_2) + \cdots + Var(X_n)] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

(ii) The second estimator has a small variance however is biased:

$$Var(\tilde{\mu}) = Var\left(\frac{\sum X_i}{n+1}\right) = Var\left(\frac{1}{n+1}\sum X_i\right) = \frac{1}{(n+1)^2}Var(X_1 + X_2 + \cdots + X_n)$$
$$= \frac{1}{(n+1)^2}[Var(X_1) + Var(X_2) + \cdots + Var(X_n)] = \frac{n}{(n+1)^2}\sigma^2$$
$$< \sigma^2$$

(iii) The third estimator is unbiased but it has a larger variance then the first, hence it is not efficient.

$$Var(\check{\mu}) = Var(X_1) = \sigma^2$$

To sum up, the first and the last estimators are unbiased, though the second one is not as $\tilde{\mu}$ systematically underestimated the true population mean. Only the first estimator is BLUE because among the unbiased ones (i.e. the first and the third) it has the smallest variance with $\frac{\sigma^2}{n} < \sigma^2$.The last estimator is hence unbiased, but not efficient.

3. Define Y, the random variable, as weekly sales of the particular record. From a sample of n=100 we have calculated the average weekly sale of the record i.e. $\bar{Y} = 260$, with a standard deviation of $s = 96$.

   As we have a large sample (100) from which we calculate our mean, we can assume that Y follows a normal distribution (instead of a t-distribution).

   NOTE: This is because for df.=100 and more of a t-distribution, the t-distribution follows a normal distribution. Hence we can use the critical values of a normal one (which equal the critical values for the t_{100}-distribution. However, this is a simplification done for estimating confidence intervals. If doing hypothesis testing we would normally go with the t-distribution.

   Exploit the symmetry of the normal distribution so that

$$P(-Y_{0.025} < Y < Y_{0.025}) = 95\% = 0.95$$

and hence for the standard normal

$$P(-z_{0.025} < Z < z_{0.025}) = 0.95.$$

Where $z_{0.025}$ is the critical value above which 2.5% (0.025) of the probability space is located. Better, the realisations of the Z value have a 0.025 probability to be equal or greater than $z_{0.025}$.

You know that the critical value for the Z-distribution is 1.96 as $P(-1.96 < z < 1.96) = 95\% = 0.95$. Trough standardisation one gets:

$$z_{0.025} = \frac{Y_{0.025} - \bar{Y}}{s}$$

which is:

$$1.96 = \frac{Y_{0.025} - 260}{9.6}$$

and hence $Y_{0.025} = 1.96 * 9.6 + 260 = 260 + 1.96 * 9.6$
and $-Y_{0.025} = -1.96 * 9.6 + 260 = 260 - 1.96 * 9.6$ respectively.

NOTE: 9.6 is the standard error of the sample mean $s/\sqrt{n} = 96/\sqrt{100} = 9.6$.

The 95% confidence interval for the true mean is hence given by:

$$P(260 - 1.96 * 9.6 < \mu_Y < 260 + 1.96 * 9.6) = 0.95$$
$$P(241.184 < \mu_Y < 278.816) = 0.95$$

Hence we are 95% confident that the true mean will lie between the interval [271.184; 278.816].

An interval estimate is better than the point estimate since it is more accurate by providing an interval and the confidence that we have that the true mean would lie within this interval. The width of the interval will depend on the standard error (note that for estimators we say standard error instead of standard deviation) of the estimator, and consequently on the sample size (remember formula for standard error $s.e. = s/\sqrt{n}$, which declines with n increasing) and the variance (standard deviation) of the sample. It also depends on the confidence coefficient (if we estimate the 95% interval or the 99% interval; obviously the 99% interval would be larger than the 95% interval).