Economics 30330: Statistics for Economics
Problem Set 6
University of Notre Dame
Instructor: Julio Garín
Spring 2012

Linear Combinations of Random Variables and Sampling (100 points)

1. Four-part problem. Go get some coffee before starting these.

   (a) On the throw of a fair die, the expected value of the number showing is 3.5 and the standard deviation is 1.71. What is the expected value and standard deviation of the sum of the values from the throw of a pair of dice?

   Let D_i be the value on die i, and T be the sum of both dice, i.e. $T = D_1 + D_2$. Therefore,

   $$\begin{aligned} E(T) &= E(D_1) + E(D_2) \\ &= 3.5 + 3.5 \\ &= 7 \end{aligned}$$

   For the variance we know that both events are independent so $Cov(D_1, D_2) = 0$. By applying the variance for linear combinations of r.v. covered in class (you should know how to derive that by now),

   $$\begin{aligned} Var(T) &= Var(D_1 + D_2) \\ &= Var(D_1) + Var(D_2) \\ &= 1.71^2 + 1.71^2 \\ &= 5.85 \end{aligned}$$

   So the standard deviation is 2.42 ($\sqrt{5.85}$).

   (b) Suppose Y_1 and Y_2 are independent, $\text{Var}(Y_1) = \text{Var}(Y_2) = \sigma_Y^2$ and $Z_1 = Y_1 + Y_2$. What is $\text{Var}(Z_1)$? How does this compare to the result found in part *a*)?

   If Y_1 and Y_2 are independent, then $Cov(Y_1, Y_2) = 0$, and following the same procedure as part a)

   $$\begin{aligned} Var(Z_1) &= Var(Y_1) + Var(Y_2) \\ &= 2\sigma_Y^2 \end{aligned}$$

   (c) Generalize the previous results. Suppose $Y_1, Y_2, \cdots, Y_n$ are independent, $\text{Var}(Y_1) = \text{Var}(Y_2) = \cdots = \text{Var}(Y_n) = \sigma_Y^2$ and $Z_2 = Y_1 + Y_2 + \cdots + Y_n$. What is $\text{Var}(Z_2)$?

   $$\begin{aligned} Var(Z_2) &= Var(Y_1 + Y_2 + ... + Y_n) \\ &= Var(Y_1) + Var(Y_2) + ... + Var(Y_n) \\ &= \sigma_Y^2 + \sigma_Y^2 + ... + \sigma_Y^2 \\ &= n\sigma_Y^2 \end{aligned}$$

   (d) Again, let's keep generalizing these results. Suppose $Y_1, Y_2, \cdots, Y_n$ are independent, $\text{Var}(Y_1) = \text{Var}(Y_2) = \cdots = \text{Var}(Y_n) = \sigma_Y^2$ and $Z_2 = Y_1 + Y_2 + \cdots + Y_n$, and $Z_3 =$

$\left(\frac{1}{n}\right)(Y_1 + Y_2 + \cdots + Y_n)$. What is $\text{Var}(Z_3)$?

$$\begin{aligned} Var(Z_3) &= Var\left[\frac{1}{n}(Y_1 + Y_2 + ... + Y_n)\right] \\ &= \frac{1}{n^2}Var(Y_1) + \frac{1}{n^2}Var(Y_2) + ... + \frac{1}{n^2}Var(Y_n) \\ &= \frac{1}{n^2}\sigma_Y^2 + \frac{1}{n^2}\sigma_Y^2 + ... + \frac{1}{n^2}\sigma_Y^2 \\ &= n\left(\frac{1}{n^2}\sigma_Y^2\right) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Congratulations, you just derived the formula for the variance of the sample mean!

2. In 2003, the average annual salary 10 years after graduation was \$168,000 for men and \$117,000 for women. The standard deviation for male graduate salary is \$40,000 and for female salaries is \$25,000.

   (a) What is the probability that a random sample of 40 males will give a sample mean within \$10,000 of \$168,000?

   Using the CLT, the sampling distribution of $\bar{x}$ is,

   $$\bar{x} \sim \mathcal{N}\left(168000, \frac{40000^2}{40}\right)$$

   Hence the probability can be obtain as usually,

   $$\begin{aligned} P(158000 \leq \bar{x} \leq 178000) &= P\left(\frac{158000 - 168000}{40000/\sqrt{40}} \leq z \leq \frac{178000 - 168000}{40000/\sqrt{40}}\right) \\ &= P(-1.58 \leq z \leq 1.58) \\ &= 1 - 2\Phi(-1.58) \\ &= 1 - 2(0.0571) \\ &= 0.8858 \end{aligned}$$

   (b) What is the probability that a random sample of 40 females will give a sample mean within \$10,000 of \$117,000?

   Now the sampling distribution is given by

   $$\bar{x} \sim \mathcal{N}\left(117000, \frac{25000^2}{40}\right)$$

   $$\begin{aligned} P(107000 \leq \bar{x} \leq 127000) &= P\left(\frac{107000 - 117000}{25000/\sqrt{40}} \leq z \leq \frac{127000 - 117000}{25000/\sqrt{40}}\right) \\ &= P(-2.52 \leq z \leq 2.52) \\ &= 1 - 2\Phi(-2.52) \\ &= 1 - 2(0.00059) \\ &= 0.9882 \end{aligned}$$

   (c) What do you prefer: giraffes or rhinos?

   Rhinos.

(d) What is the probability that a random sample of 100 males will give a sample mean less than \$164,000?

The sampling distribution is

$$\bar{x} \sim \mathcal{N}\left(168000, \frac{40000^2}{100}\right)$$

$$\begin{aligned} P(\bar{x} \leq 164000) &= P\left(z \leq \frac{164000 - 168000}{40,000/\sqrt{100}}\right) \\ &= P(z \leq -1) \\ &= \Phi(-1) \\ &= 0.1587 \end{aligned}$$

3. Suppose in a population the math (M) and verbal (V) SAT scores have the following moments: $\mathbb{E}(M) = 510$, $\mathbb{E}(V) = 475$, $\text{Var}(M) = 750$, $\text{Var}(V) = 610$, and $\rho_{M,V} = 0.4$. What is the expected value and variance of the total SAT, $T = M + V$?

$$\begin{aligned} E(T) &= E(M + V) \\ &= E(M) + E(V) \\ &= 510 + 475 \\ &= 985 \\ Var(T) &= Var(M) + Var(V) + 2Cov(M, V) \\ &= \sigma_M^2 + \sigma_V^2 + 2\rho_{MV}\sigma_M\sigma_V \\ &= 750 + 610 + 2(0.4)(\sqrt{750})(\sqrt{610}) \\ &= 1901.2 \end{aligned}$$

4. A random sample of size N is selected from a population with $\sigma = 10$. What is the standard error of the mean if

(a) $N = 500$?

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{10}{\sqrt{500}} \\ &= 0.447 \end{aligned}$$

(b) $N = 5,000$?

$$\begin{aligned} \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{10}{\sqrt{5000}} \\ &= 0.1414 \end{aligned}$$

(c) $N = \infty$? Compare this result with the previous parts.

As $n \to \infty$, $\sigma_{\bar{x}} \to 0$. This was illustrated in the parts (a) and (b), where the standard error decreased as the sample size increased. Put differently, our estimate for the population mean becomes more precise as the size of the sample is increases.

5. Final grades in a class are a weighted average of the midterm (25%) and final (75%) exams. Each exam has 100 possible points. Suppose the average and standard deviation of scores on the midterm were 71 and 19 respectively, while the values for the final exam were 69 and 23. Suppose further that the correlation coefficient between the two exams is 0.50.

   (a) What is the mean and standard deviation on the final grades in class?
   We have that $G = 0.25M + 0.75F$, hence the expected value is given by

$$\begin{aligned} E(G) &= E(0.25M + 0.75F) \\ &= 0.25E(M) + 0.75E(F) \\ &= 0.25\mu_M + 0.25\mu_F \\ &= 0.25(71) + 0.75(69) \\ &= 69.5 \end{aligned}$$

   Similarly, for the variance,

$$Var(G) = 0.25^2\sigma_M^2 + 0.75^2\sigma_F^2 + 2(0.25)(0.75)\rho_{F,M}$$

   Although we do not know $Cov(F, M)$, we know that $\sigma_{FM} = \rho_{FM}\sigma_F\sigma_M$. Therefore we can substitute for those values:

$$\begin{aligned} \sigma_G^2 &= 0.25^2(19^2) + 0.75^2(23) + 2(0.25)(0.75)(0.5)(19)(23) \\ &= 402 \end{aligned}$$

   Which gives us a standard deviation of $\sigma_G = \sqrt{\sigma_G^2} = \sqrt{402} = 20.049$.

   (b) Suppose the final grades are normally distributed with mean and variance found in part *a*). What fraction of students will get an A if they need more than 93 points to obtain that grade?
   Since a linear combination of r.v. that are normally distributed is also normally distributed, we have that the distribution of final grades, X, is given by

$$X \sim \mathcal{N}(69.5, 402)$$

   The fraction can be, hence, calculated as a probability:

$$\begin{aligned} P(x \geq 93) &= 1 - P(x \leq 93) \\ &= 1 - P\left(z \leq \frac{93 - 69.5}{\sqrt{402}}\right) \\ &= 1 - \Phi(1.172) \\ &= 1 - 0.8790 \\ &= 0.1210 \end{aligned}$$

   12.1% of students will get an A in this class if 93 is the score needed for that.

6. Show that the sample variance is an unbiased estimator of the population variance.

We want to show that $E(s^2) = \sigma^2$, where $s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$.

$$
\begin{aligned}
E(s^2) &= E\left[\frac{\sum_i (x_i - \bar{x})^2}{n-1}\right] \\
\implies (n-1)E(s^2) &= E\left[\sum_i (x_i - \bar{x})^2\right] \\
&= E\left[\sum_i (x_i^2 - 2x_i\bar{x} + \bar{x}^2)\right] \\
&= E\left[\sum_i x_i^2 - 2\bar{x}\sum_i x_i + \sum_i \bar{x}^2\right] \\
&= E\left[\sum_i x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right] \\
&= E\left[\sum_i x_i^2\right] - 2nE(\bar{x}^2) \\
&= \sum_i E(x_i^2) - 2nE(\bar{x}^2)
\end{aligned}
$$

Recall that $Var(X) = E[X^2] - [E(X)]^2 \implies E[X^2] = Var(X) + [E(X)]^2$.

It follows that, $E[x_i^2] = Var(x_i) + [E(x_i)]^2 = \sigma^2 + \mu^2$ *because* $Var(x_i) = \sigma^2$ *and* $E(x_i) = \mu$.

Similarly, $E[\bar{x}^2] = Var(\bar{x}) + [E(\bar{x})]^2 = \frac{\sigma^2}{n} + \mu^2$ *because* $Var(\bar{x}) = \frac{\sigma^2}{n}$ *and* $E(\bar{x}) = \mu$.

$$
\begin{aligned}
\implies (n-1)E(s^2) &= \sum_i (\sigma^2 + \mu^2) - 2n\left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= n(\sigma^2 + \mu^2) - 2n\left(\frac{\sigma^2}{n}\right) + n\mu^2 \\
&= n\sigma^2 + n\mu^2 - 2n\frac{\sigma^2}{n} - n\mu^2 \\
&= (n-1)\sigma^2 \\
\implies E(s^2) &= \sigma^2
\end{aligned}
$$

Therefore, s^2 *is an unbiased estimator of the population variance.*

7. A paint manufacturer advertises that their exterior paint will last 5 years. Assume paint life is normally distributed with a standard deviation of 0.5 years.

   (a) Suppose a local TV reporter tests this claim, paints one house, and notices that the house paint only lasts 4.5 years. Would you consider this evidence against the manufacturer's claim?
   We first need to describe the sampling or probability distribution of the sample mean:

$$\bar{x} \sim \mathcal{N}\left(5, 0.15/\sqrt{1}\right)$$

Now we are ready to obtain the probability:

$$\begin{aligned} P(x \leq 4.5) &= P\left(z \leq \frac{4.5-5}{0.5/\sqrt{1}}\right) \\ &= P(z \leq -1) \\ &= \Phi(-1) \\ &= 0.1587 \end{aligned}$$

There is a 15.85% chance that the paint on one one house, chosen randomly, will last less than 4.5 years. This is not strong evidence against the manufacturer.

(b) Suppose instead of the TV report's test, a consumer magazine paints 10 houses and finds the average life of the paint is 4.5 years. Would you consider this evidence against the manufacturer's claim?

Since the sample size changed, so did the sampling distribution which is now more precise given the reduction in the standard error, $\sigma_{\bar{x}}$, due to the larger sample. We have that

$$\bar{x} \sim \mathcal{N}\left(5, 0.15/\sqrt{10}\right)$$

$$\begin{aligned} P(x \leq 4.5) &= P\left(z \leq \frac{4.5-5}{0.5/\sqrt{10}}\right) \\ &= P(z \leq -3.16) \\ &= 0.0007 \end{aligned}$$

The manufacturer's claim is probably false, since it is very unlikely the paint on 10 houses, selected at random, will last less than 5 years.

8. A water bottler sells spring water in 1 liter bottles. The machines are set such that on average, 1.02 liters are dispensed with a known standard deviation of 0.06 liters. The firm routinely collects a random sample of bottles and tests whether the machine is dispensing correctly. Suppose the machine is working properly (μ and σ are known). What is the chance that in a random sample of 16 bottles, the sample mean will be within 0.05 liters of the specified level?

We know that according to the central limit theorem, $\bar{x}$ is normally distributed with a mean of μ and a variance of σ^2/n. When the machine is working properly, $\mu = 1.025$ and $\sigma = 0.06$. In this case, $n = 16$ so $\sqrt{n} = 4$. In other words, by the CLT, $\bar{x} \sim \mathcal{N}(1.05, 0.06^2/16)$.

$$\begin{aligned} P(0.97 \leq \bar{x} \leq 1.07) &= P\left(z \leq \frac{1.07-1.05}{0.06/4}\right) - P\left(z \leq \frac{0.97-1.05}{0.06/4}\right) \\ &= \Phi(3.33) - \Phi(-3.33) \\ &= 0.9996 - 0.004 \\ &= 0.9992 \end{aligned}$$

9. A student surveys 60 undergraduates to determine the average number of drinks consumed over the past two weeks ($\bar{x}$). Based on previous surveys, the student believes the standard deviation of drinks per week in the population is 7. The researcher would like to reduce the standard error of by increasing the sample size.

(a) If the student's beliefs are correct, what will be the standard error of $\bar{x}$?

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{7}{\sqrt{60}} \\ &= 0.904\end{aligned}$$

(b) How many students would the student have to survey to cut the standard error of in half?

We would like to reduce the standard error to 0.452 and since the population standard deviation remains the same we have that

$$\frac{7}{\sqrt{n}} = 0.452$$

So that $7/\sqrt{n} = 0.452 \implies \sqrt{n} = 7/0.452 \implies n = (7/0.452)^2$. *Hence,*

$$n = 239.8$$

The researcher will have to survey at least 240 students to cut the standard error in half.

(c) How many students would the research have to survey to cut the standard error by 75%?
The reduction of 75% is equivalent to saying that the new standard error has to be 25% of the original value, or $0.25(0.904) = 0.226$. *As in part b),* $7/\sqrt{n} = 0.226 \implies \sqrt{n} = 7/0.226 \implies n = (7/0.226)^2$. *Hence,*

$$n = 959.35$$

The student needs to survey at least 960 students to cut the standard error by 75%.

10. The covariance between x and y represents the average of the products of the deviations of x and y from their respective means. In other words, it represents the average of the sum of the cross-products. Show that the sum of cross-product can be written either as $\sum_i (x_i - \bar{x})\, y_i$ or $\sum_i (y_i - \bar{y})\, x_i$.

$$\begin{aligned}\sum_i (x_i - \bar{x})(y_i - \bar{x}) &= \sum_i (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}) \\ &= \sum_i x_i (y_i - \bar{y}) - \bar{x} \sum_i y_i + \sum_i \bar{x}\bar{y} \\ &= \sum_i x_i (y_i - \bar{y}) - n\bar{x}\bar{y} + n\bar{x}\bar{y} \\ &= \sum_i x_i (y_i - \bar{y})\end{aligned}$$

Extra Problems

1. In 2000, a *Time*/CNN polled 589 voters. If the population proportion for a candidate is $p = 0.5$ and $\bar{p}$ is the sample proportion of likely voters that favor this candidate,

   (a) Show the sampling distribution of $\bar{p}$

   $$\sigma_{\bar{p}}^2 = \frac{\sigma^2}{n} = \frac{p(1-p)}{n} = \frac{(0.5)(0.5)}{589} = 0.00042$$
   $$\implies \bar{p} \sim \mathcal{N}(0.5, 0.00042)$$

   (b) What is the probability that the poll will provide a sample proportion within ± 0.04 of the population proportion?

   $$\begin{aligned} P(0.5 - 0.4 \leq \bar{p} \leq 0.5 + 0.4) &= P\left(\frac{-0.4}{\sigma/\sqrt{n}} \leq z \leq \frac{0.4}{\sigma/\sqrt{n}}\right) \\ &= P(-1.942 \leq z \leq 1.942) \\ &= 1 - 2P(z \leq -1.942) \\ &= 1 - 2\Phi(-1.942) \\ &= 1 - 2(0.0262) \\ &= 0.9476 \end{aligned}$$

2. A random sample of size $N = 1,000$ is drawn from a population with $p = 0.4$, where p is the proportion of the population that has a certain characteristic.

   (a) What is the expected value of $\bar{p}$ and its standard error?

   $$\begin{aligned} E(\bar{p}) &= p = 0.4 \\ \sigma_{\bar{p}} &= \sqrt{\frac{p(1-p)}{n}} \\ &= \sqrt{\frac{(0.4)(0.6)}{1000}} \\ &= 0.01549 \end{aligned}$$

   (b) Illustrate the sampling distribution of $\bar{p}$.

   $$\bar{p} \sim \mathcal{N}(0.4, 0.01549^2)$$