Economics 250 — Exercise 2 **Due 10 November 2016, in class**

Remember to include the departmental cover page. You are free to discuss the questions with classmates, but your answers should be your own.

1 Binomial investment returns [35 points]

An investor is considering the purchase of $1000 of a company's stock. On any given day, this investment has only two possible changes in value: an increase of 5% or a decrease of 4%.

The probabilities of changes in the investment's value are fixed, even if the previous days' changes are known: the investment will increase in value with probability 0.7 and decrease with probability 0.3.

Note: For this question, keep at least 4 significant digits when rounding.

a) What is the mean value (also referred to as the "expected value") of the initial $1000 investment after 1 day? After 2 days?

b) Use the binomial probability formula to find the probability of seeing at least 3 *positive* returns in the first 4 days, and find the probability of seeing at least 3 *negative* returns in the first 4 days.

c) Find the probability of having exactly two positive returns in the first four days using two different methods (show your work):

(*i*) by using your answers to *b*), and

(*ii*) by using the binomial probability formula.

d) If the investment value increases on exactly k of the four days, the value of the investment after four days will be $1000 \times 1.05^k 0.96^{4-k}$. Make a table of three rows (and leave space for a fourth row) showing:

- k, the number of days the investment value increased.
- $P(X = k)$, the probability of each value of k occurring.
- x_k, the investment value after exactly k out of four days of increasing value.

e) Use the values in your table to calculate μ_4, the expected value of the investment after 4 days.

f) Suppose that the investor will invest his money for 60 days. Find the probability of:

(*i*) seeing a positive return on at least half of the days.

(*ii*) seeing a positive return on 45 or more days.

(*iii*) seeing fewer than 18 days with negative returns.

1.1 Answers

a) For one period, this is just the weighted mean: $1000(0.96 \times 0.3 + 1.05 \times 0.7) = 1000(1.023) = 1023$

For two periods, this is the same idea, but now there are four possibilities:

- Event 1: Two "bad" days, where the value will equal $1000 \times .96 \times .96 = 921.6$
- Event 2: One good day followed by a bad day; the final value will equal $1000 \times 1.05 \times .96 = 1008$
- Event 3: One bad day followed by a good day; the final value will equal $1000 \times .96 \times 1.05 = 1008$
- Event 4: Two good days, after which the value equals $1000 \times 1.05 \times 1.05 = 1102.5$

To calculate the expected value, we need to multiple each potential value by the probability of that value occuring. The probabilities are:

- Event 1: Two bad days: $.3 \times .3 = 0.09$
- Event 2: Good then bad: $.7 \times .3 = 0.21$
- Event 3: Bad then good: $.3 \times .7 = 0.21$
- Event 4: Both good days: $.7 \times .7 = 0.49$

(Note that the probabilities add up to 1: always a useful check for a discrete distribution like this).

The expected value of this discrete distribution is therefore:

$$E(X) = \sum_{i=1}^{4} p_i x_i = .09(921.6) + .21(1008) + .21(1008) + .49(1102.5) = 1046.53$$

Since the returns are independent and have the same probabilities, you could also use the binomial formula:

Value:	$1000(0.96^2) = 921.60$	$1000(0.96 \times 1.05) = 1008$	$1000(1.05^2) = 1102.50$
Probability:	$\binom{2}{0}0.3^2 = 0.09$	$\binom{2}{1}(0.7)(0.3) = 0.42$	$\binom{2}{2}0.7^2 = 0.49$

Note, however, that the binomial formula doesn't distinguish between events 2 and 3 (both are 1 increase, 1 decrease): it just has one value with double the probability, but otherwise the probabilities and values are the same, giving us the same answer: $921.60 \times 0.09 + 1008 \times 0.42 + 1102.50 \times 0.49 = 1046.53$.

b) Positive returns:

$P(X \geq 3) = P(X = 3) + P(X = 4) = \binom{4}{3}0.7^3 0.3 + \binom{4}{4}0.7^4 = 0.4116 + 0.2401 = 0.6517$

Negative returns: "at least three negative" means the same thing as "at most 1 positive":

$P(X \leq 1) = P(X = 0) + P(X = 1) = \binom{4}{0}0.3^4 + \binom{4}{1}0.7(0.3)^3 = 0.0081 + 0.0756 = 0.0837$

You could also get this second number by constructing a new binomial variable, let's call it Y, that counts the number of *negative* returns, with a probability of success now being 0.3 instead of 0.7. We would then calculate the second one as:

$P(Y \geq 3) = P(Y = 3) + P(Y = 4) = \binom{4}{3}0.3^3(0.7) + \binom{4}{4}0.3^4 = 0.0756 + 0.0081 = 0.0837$

c) Since "at least 3", "at most 1" (or equivalently, "at least 3 negatives") and "exactly 2" cover every possible outcome, and are all disjoint, we know their probabilities have to add up to 1. Therefore:

$P(X = 2) = 1 - P(X \geq 3) - P(X \leq 1) = 1 - 0.6517 - 0.0837 = 0.2646$

Using the binomial probability formula to get the answer:

$P(X = 2) = \binom{4}{2}0.7^2 0.3^2 = 6(0.0441) = 0.2646$

d) The table is as follows:

k	0	1	2	3	4
x_k	849.35	928.97	1016.06	1111.32	1215.51
$P(X = k)$	0.0081	0.0756	0.2646	0.4116	0.2401

The x_k and $P(X = k)$ values are calculated from:

$$x_0 = (1000)1.05^0 0.96^4 = 849.35 \qquad P(X = 0) = \binom{4}{0}0.7^0 0.3^4 = 1(0.0081) = 0.0081$$

$$x_1 = (1000)1.05^1 0.96^3 = 928.97 \qquad P(X = 1) = \binom{4}{1}0.7^1 0.3^3 = 4(0.0189) = 0.0756$$

$$x_2 = (1000)1.05^2 0.96^2 = 1016.06 \qquad P(X = 2) = \binom{4}{2}0.7^2 0.3^2 = 6(0.0441) = 0.2646$$

$$x_3 = (1000)1.05^3 0.96^1 = 1111.32 \qquad P(X = 3) = \binom{4}{3}0.7^3 0.3^1 = 4(0.1029) = 0.4116$$

$$x_4 = (1000)1.05^4 0.96^0 = 1215.51 \qquad P(X = 4) = \binom{4}{4}0.7^4 0.3^0 = 1(0.2401) = 0.2401$$

e) The 4-period mean return is then just the sum of each probability times its associated value:

$$\begin{aligned}\mu_4 &= 0.0081(849.35) + 0.0756(928.97) + 0.2646(1016.06) \\ &\quad + 0.4116(1111.32) + 0.2401(1215.51) \\ &= 1095.22\end{aligned}$$

f) Since both np and $n(1-p)$ are bigger than 15, we should use the normal approximation to calculate this.

The approximation is: $X \sim \mathrm{N}(np, \sqrt{np(1-p)})$. Plugging in n and p that gives us: $X \sim \mathrm{N}(42, 3.55)$

Thus our calculation involves calculating and evaluating z-scores:

- At least half the days: $P(X \geq 30) = P(Z \geq \frac{30-42}{3.55}) = P(Z \geq -3.38) = 1 - 0.0004 = 0.9996$
- 45 or more days: $P(X \geq 45) = P(z \geq \frac{45-42}{3.55}) = P(Z \geq 0.85) = 1 - 0.8023 = 0.1977$
- Fewer than 18 with negative means 42 or more with positive days. Since 42 is the mean, we can answer this straight away: 0.5. (If you want to work out the calculation: $P(X \geq 42) = P(z \geq \frac{42-42}{3.55}) = P(z \geq 0) = 1 - 0.5 = 0.5$.

2 Random Sampling [25 points]

Consider the distribution U(0, 100), which has mean $\mu = 50$.

a) *Use Excel for this question.* Generate 20 rows of 10 values each, where each value is a draw from this distribution. For each of the 20 rows, add an 11th value which equals the mean of the first 10 values. *Tip: you can enter the Excel formula* `=AVERAGE(A1:J1)` *to calculate the mean of cells A1 through J1.*

Plot these 20 sample means on a histogram, choosing an appropriate bin width.

b) If you repeated this process thousands of times to get thousands of sample means (each from 10 draws from U(0, 100)), what would you expect the histogram to look like? Where would you expect the mode to be?

c) If you repeated this process but generated samples of 20 random values instead of 10, how would you expect the histogram shape to change compared to what you described in part *b*)?

d) Suppose you have another distribution, X, for which you do not know the exact distribution, but you do know that X has a mean of 5 and standard deviation of 3. What will be the distribution of $\overline{x}$, the sample mean, for a sample constructed of 25 values drawn from X?

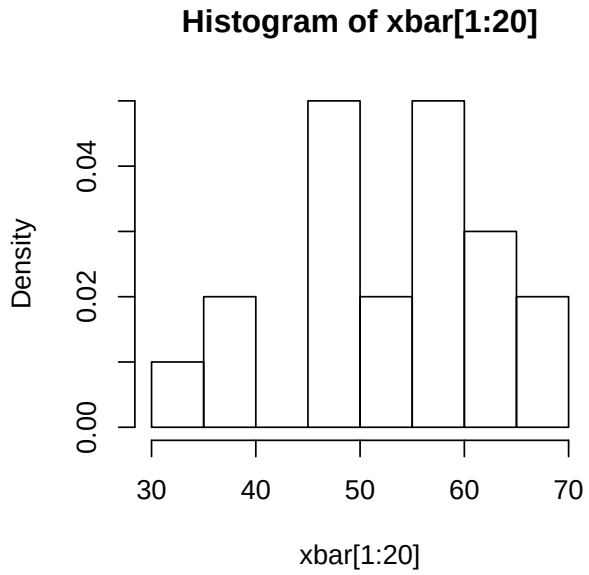
2.1 Answers

a) Answers will vary; it should look vaguely normal, with most of the observations between about 30 and 70, but since there are only 20 samples, any given histogram might not look particularly normal.

The 20 means I got (rounded to 1 decimal places) were:

51.07 38.63 62.42 50.73 58.48 59.95 65.55 67.57 49.81 39.47 63.57 55.08 45.41 49.17 63.69 47.28 57.97 46.66 30.83 56.75

The following is a histogram of these values, using bins of size 5 from 30 to 70.



b) As you add more and more samples, you should, following the central limit theorem, get something that gets closer and closer to a normal distribution, centered on the population mean, $\mu = 50$. Since the mode (the highest point of the density function) for a normal distribution is at the mean, 50 is the mode as well.

It is not necessary for this assignment, but you could calculate the precise distribution parameters. See part *c*), below, for this calculation.

See the answer to the part *c*), below, for a histogram using 10000 samples (*not* required for the assignment).

c) Taking means of 20 values instead of 10 will still be normally distributed and still centered at $\mu = 50$, but with a lower standard deviation: i.e. there will be more height in the middle and less in the tails.

It is not necessary for the assignment, but the precise parameters for the normal distributions can be calculated, from the Central Limit Theorem (p. 271 of the textbook):

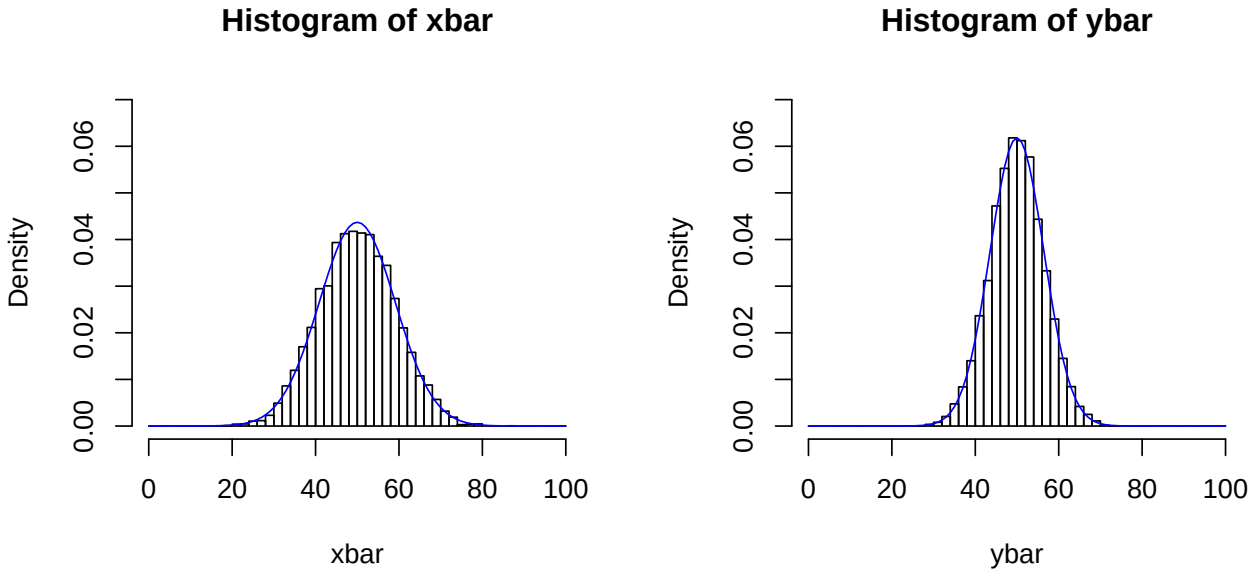
$$\overline{x} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

In this case: $\mu = 50$, $n = 10$ (part d) or 20 (part e). The variance of a $\mathrm{U}(a, b)$ random variable is: $\sigma^2 = \frac{(a-b)^2}{12}$, and so the standard deviation of our $\mathrm{U}(0, 100)$ is: $\sigma = \frac{100-0}{\sqrt{12}}$. Plugging these into the central limit theorem distribution gives us:

$$\overline{x}_{(d)} \sim \mathrm{N}\left(50, \frac{1}{\sqrt{10}}\left(\frac{100}{\sqrt{12}}\right)\right) = \mathrm{N}(50, 9.1287)$$
$$\overline{x}_{(e)} \sim \mathrm{N}\left(50, \frac{1}{\sqrt{20}}\left(\frac{100}{\sqrt{12}}\right)\right) = \mathrm{N}(50, 6.4550)$$

The following two histograms use an identical scale for comparison. The left is a histogram of 10000 sample means of size 10 (for part b)) while the right shows a histogram for 10000 sample means of size 20, for part e. Also drawn on each diagram is a normal density curve using the normal parameters calculated above.

(Note that to be able to draw the density curve and histogram on the same graph, you need to scale the histograms bars to have a total area of 1, measuring densities instead of frequency. The easiest way to do this in Excel is to divide the histogram counts by the bin width times the number of observations—in this case, dividing the counts by 2*10000. Alternatively, some more advanced statistics packages can generate density histograms directly.)



As is apparent, the distribution for part (e) indeed has more probability mass in the central regions and smaller tails. Also apparent is that the central limit theorem is working very well here, even though n is relatively small: the histograms are very close to the expected normal density curves.

d) This question is simply asking you to use the central limit theorem: even without knowing the distribution of X, we typically use the assumption that the mean of a sample from X of size n will be approximately normal, specifically:

$\mathrm{N}(5, 3/\sqrt{25}) = N(5, 0.6)$

3 Shoe sizes [40 points]

Based on your statistics training, a shoe company hires you to perform statistical analysis to help it design and produce a new line of shoe for women. In order to proceed with the

initial production, the company needs to know how many shoes to make of the various sizes. You work with the marketing department to identify the most likely customers of the new shoe and are able to conduct a simple random sample of likely customers whose feet are measured.

Your sample contains the following sizes (in centimeters):

26.1	25.9	26.2	23.3	25.0	26.2	23.3	26.1
23.5	22.4	25.0	26.1	25.7	22.3	24.9	

From this data you calculate the sample mean $\overline{x} = 24.8$ and sample standard deviation $s = 1.4467$.

a) The designer of the shoes claims that the mean female foot size equals 24.0cm. The head of the marketing department claims that the mean female foot size is *at least* 10 inches, that is, 25.4cm. For each of these two claims, write down the null and alternative hypotheses, using the usual notation.

b) Assuming that you know the true population standard deviation is $\sigma = 1.27$, for each of the two claims:

(*i*) Calculate a test statistic that will allow you to test the statistical significance of this claim. What is the distribution of this statistic?

(*ii*) Find the p-value associated your test statistic. Is this a one-tailed or two-tailed test?

(*iii*) Can you reject the claim at the 95% confidence level? At the 99% confidence level? What is the highest confidence level at which you could reject the claim?

(*iv*) If you had obtained a test statistic of *positive* 4.7 instead of what you obtained for (*i*), would you reject H_0 at the 99% confidence level?

c) Still assuming $\sigma = 1.27$, find a 97% confidence interval for the mean shoe size.

d) Now suppose that you do not know σ. For each of the two claims:

(*i*) Calculate a test statistic that will allow you to test the statistical significance this claim. What is the distribution of this statistic? *Remember to include any relevant parameters of the distribution in your answer, if appropriate.*

(*ii*) Can you reject the claim at the 90% confidence level? At the 95% confidence level?

e) Still supposing σ is unknown, find a 95% confidence interval for the mean shoe size.

f) Calculate the power of the test: $H_0 : \mu = 25.0, H_a : \mu \neq 25.0$ with $n = 15$ and $\sigma = 1.27$ at a significance level of $\alpha = 0.05$ when the true population mean equals 24.0cm.

3.1 Answers

a) For the "mean is 24cm" claim:

$$H_0 : \mu = 24$$
$$H_a : \mu \neq 24$$

and for the "at least 25.4cm" claim:

$$H_0 : \mu \geq 25.4$$
$$H_a : \mu < 25.4$$

(You could alternatively write $H_0 = 25.4$ for the latter, as long as the alternative is still $<$).

b) (*i*) 24cm:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{24.8 - 24}{1.27/\sqrt{15}} = 2.44$$

25.4cm:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{24.8 - 25.4}{1.27/\sqrt{15}} = -1.83$$

Both test statistics follow the standard normal distribution, that is, $N(0, 1)$. (This is sometimes also called the "Z" distribution).

(*ii*) 24cm:

$$p = P(Z > 2.44) + P(Z < -2.44) = 2 \times P(Z < -2.44) = 2(0.0073)$$
$$= 0.0146$$

This is a two-tailed test, hence the two probabilities in the above equation.

25.4cm:

$$p = P(Z < -1.83) = 0.0336$$

This is a one-tailed test.

(*iii*) We can reject both claims at the 95% but not 99% confidence level. For 24cm, we could reject H_0 at up to the 98.54% confidence level, and for the 25.4cm claim we could reject H_0 at up to the 96.64% confidence level.

(*iv*) For the 24cm case, yes, we would reject H_0 because $z = 4.7$ represents a value far out on the right tail of the z-distribution.

For the second hypothesis we would *not* reject H_0 because 4.7, even though it is a large test statistic, it is a large *positive* test statistic, but for this one-sided

test, we only reject for large *negative* test statistics. If you work out the p-value, you would get a p-value very close to 1, which clearly won't be rejected for any sensible value of α.

Another way to think of this: the alternative is about finding evidence of a mean significantly below 25.4, but getting we can only get a large positive test statistic if we have a sample mean significantly *above* 25.4; values above 25.4, however, don't give us any evidence in support of the alternative that the population mean is *below* 25.4.

c) We need to find the z^* value that gives us an area of 0.015 in each tail so that the area of both tails adds up to 0.03.

The associated z^* value is 2.17. The confidence interval is thus:

$$\begin{aligned} & \left[\overline{x} - z^* \frac{\sigma}{\sqrt{n}}, \overline{x} + z^* \frac{\sigma}{\sqrt{n}}\right] \\ = & \left[24.8 - 2.17\frac{1.27}{\sqrt{15}}, 24.8 + 2.17\frac{1.27}{\sqrt{15}}\right] \\ = & \ [24.09, 25.51] \end{aligned}$$

d) With σ unknown, we instead use s, which gives t statistics instead of z statistics:

(*i*) 24cm:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{24.8 - 24}{1.4467/\sqrt{15}} = 2.142$$

25.4cm:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{24.8 - 25.4}{1.4467/\sqrt{15}} = -1.606$$

Both statistics follow the t-distribution with $n - 1 = 14$ degrees of freedom.

(*ii*) 24cm: Comparing 2.142 to the critical values in the textbook's "Table D" with 14 degrees of freedom, we see that it bigger than the critical value 1.761 associated with $p = .05$, but smaller than the 2.145 critical value for $p = 0.025$. Since this is a two-sided test, we have to double these p-values and conclude that our p value is between 0.05 and 0.1. Thus we can reject H_0 at a 90% confidence level but not at a 95% (or higher) confidence level.

25.4cm: Table D gives us the critical value for upper-tail tests: the lower-tail critical values are simply the negatives. Since we're doing a *one-tailed* test this time, the critical value we care about for a 90% confidence level is the one for $p = .10$, again with 14 degrees of freedom. Since our t-statistic of -1.606 is a larger negative number than the critical value of -1.345, we can reject the null hypothesis the 90% confidence level. Since it is not bigger than the $p = .05$ critical value of -1.761, we fail reject the null hypothesis at the 95% confidence level.

e) The confidence interval formula is almost the same as the z version, above, but with $t^* = 2.145$ critical value (from the t-distribution with 14 degrees of freedom, using $p = .025$) and using s in place of σ:

$$\begin{aligned}
&\left[\overline{x} - t^*\frac{s}{\sqrt{n}}, \overline{x} + t^*\frac{s}{\sqrt{n}}\right] \\
&= \left[24.8 - 2.145\frac{1.4467}{\sqrt{15}}, 24.8 + 2.145\frac{1.4467}{\sqrt{15}}\right] \\
&= [24.00, 25.60]
\end{aligned}$$

f) Since this is a two-sided test, we reject H_0 if we get a z value above 1.96 or below -1.96. Plugging in the null hypothesis μ, and the given σ and n values this means we reject when:

$$\begin{aligned}
\frac{\overline{x}^*_{high} - 25}{1.27/\sqrt{15}} &> 1.96 \\
\overline{x}^*_{high} &> 25.64
\end{aligned}$$

or when:

$$\begin{aligned}
\frac{\overline{x}^*_{low} - 25}{1.27/\sqrt{15}} &< -1.96 \\
\overline{x}^*_{low} &< 24.36
\end{aligned}$$

The power is then the probability of getting a value of $\overline{x}$ that is outside these critical values when μ equals the given alternative, $\mu_a = 24$:

$$\begin{aligned}
Power &= P(\overline{x} \geq \overline{x}^*_{high}) + P(\overline{x} \leq \overline{x}^*_{low}) \\
&= P(\frac{\overline{x} - \mu_a}{\sigma/\sqrt{n}} \geq \frac{25.64 - 24}{1.27/\sqrt{15}}) + P(\frac{\overline{x} - \mu_a}{\sigma/\sqrt{n}} \leq \frac{24.36 - 24}{1.27/\sqrt{15}}) \\
&= P(Z \geq 5.00) + P(Z \leq 1.10) \\
&= P(Z \leq -5.00) + P(Z \leq 1.10) \\
&= 0.0000 + 0.8643 \\
&= 0.8643
\end{aligned}$$

In other words, when the true value is actually 24, we have about an 86% chance that we will reject the (false) null hypothesis that the true value equals 25.

Note that we use the null hypothesis value to calculate the $\overline{x}^*$ values; when we calculate the power, we use the given alternative value, μ_a: this gives us the probability of getting a rejected value when the *actual* population mean equals $\mu_a = 24$, but the *hypothesis* test is for $\mu = 25$.