Economics 250 — Midterm 2 — Answer guide **19 March 2015**

1. [18] A test for the presence of heart disease gives a numerical score denoted x.

A person receives a positive diagnosis if $x > 90$.

a) Among people *with* heart disease, x follows a normal distribution with mean 105 and standard deviation 8.1. What is the probability that a randomly selected person who has heart disease receives a positive diagnosis?

b) Among people *without* heart disease, x follows a normal distribution with mean 70 and standard deviation 15.0. What is the probability that a randomly selected person who does not have heart disease receives a positive diagnosis?

c) Suppose that it is known that 15% of the population has heart disease. Find the probability that a person who receives a positive diagnosis from this test actually has heart disease.

Answers:

a) If a person has heart disease, $x \sim N(105, 8.1)$, so:

$$\begin{aligned} P(x > 90) &= P\left(z > \frac{90 - 105}{8.1}\right) = P(z > -1.85) = 1 - 0.0322 \\ &= 0.9678 \end{aligned}$$

b) For a person without heart disease, $x \sim N(70, 15)$, so:

$$\begin{aligned} P(x > 90) &= P\left(z > \frac{90 - 70}{15}\right) = P(z > 1.33) = 1 - 0.9082 \\ &= 0.0918 \end{aligned}$$

c) Let H be the event that a person suffers from heart disease, and D being the event that they received a positive diagnosis. The question is then asking us to find $P(H|D)$. Note that the answer to part *a*) is $P(D|H)$ and the answer to part *b*) is $P(D|H^{\mathsf{C}})$, and the question gives $P(H) = 0.15$, from which we can easily calculate $P(H^{\mathsf{C}}) = 1-0.15 = 0.85$.

Thus we have everything we need to apply Bayes' Rule:

$$\begin{aligned} P(H|D) &= \frac{P(D|H)P(H)}{P(D|H)P(H) + P(D|H^{\mathsf{C}})P(H^{\mathsf{C}})} = \frac{(0.9678)(0.15)}{(0.9678)(0.15) + (0.0918)(0.85)} \\ &= 0.65 \end{aligned}$$

2. [20] A professor is interested in the study habits of his students and conducts a random sample of 20 students to determine the mean number of hours spent studying per week. This sample has a mean of 18.4. Suppose that $\sigma = 4.7$ is known and that the population is normally distributed.

All parts of this question should be done at the $\alpha = 0.05$ level.

a) Test the hypothesis that the mean number of hours spent studying per week is 20 against the alternative that the mean is less than 20. What is the p-value of this test?

b) What is the power of this test when the population mean equals 18? What is the power when the population mean equals 20? What is the power when the population mean equals 21.9?

c) What is the probability of making a Type I error with this test? What is the probability of making a Type II error when the population mean equals 18?

Hint: Recall that a Type I error is rejecting the null hypothesis when it is true, and a Type II error is failing to reject the null hypothesis when the alternative hypothesis is true.

Answers:

a) We want to test:

$$H_0 : \mu = 20$$
$$H_a : \mu < 20$$

Since σ is known, we use a z test:

$$\begin{aligned} z &= \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{18.4 - 20}{4.7/\sqrt{20}} \\ &= -1.52 \\ P(z < -1.52) &= 0.0643 \end{aligned}$$

where the last value comes from the table of standard normal probabilities.

0.0643 is the p-value.

Since $0.0643 = p > \alpha = 0.05$, we fail to reject H_0: the data does not give us sufficient evidence to reject that the mean is 20 in favour of the mean being below 20.

b) The power of the test is the probability that we reject H_0 *given* a particular value of the true population parameter, μ.

In order to reject H_0, we need to get $z < -1.645$, where -1.645 is the value of z where the area to the left equals α. (Since this is a one-tail test, we look for α not $\alpha/2$: for a two-tail test at $\alpha = 0.05$, we'd find $z = \pm 1.96$).

So we reject H_0 if $z < -1.645$, and substituting in $z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$, we get:

$$\begin{aligned}
\frac{\overline{x}-20}{4.7/\sqrt{20}} &< -1.645 \\
\overline{x} &< 20 - 1.645(4.7/\sqrt{20}) \\
\overline{x} &< 18.27
\end{aligned}$$

In other words, we will get $z < -1.645$ (and thus reject H_0) whenever $\overline{x} < 18.27$.

Our power is then the probability that a sample actually gives such a sample mean when the population mean equals the given values.

For $\mu = 18$:

$$\begin{aligned}
P(\text{reject } H_0|\mu = 18) = P(\overline{x} < 18.27|\mu = 18) &= P\left(z < \frac{18.27-18}{\sigma/\sqrt{n}}\right) \\
&= P(z < 0.26) \\
&= 0.6026
\end{aligned}$$

When $\mu = 20$, rejecting H_0 means making a Type I error: but this is just α by definition, and so the power in this case is simply $\alpha = 0.05$. If you want, you can calculate it anyway:

$$\begin{aligned}
P(\text{reject } H_0|\mu = 20) = P(\overline{x} < 18.27|\mu = 20) &= P\left(z < \frac{18.27-20}{\sigma/\sqrt{n}}\right) \\
&= P(z < -1.646) \\
&= 0.05
\end{aligned}$$

(slight rounding errors and using the z table will lead to 0.0505 or 0.495, since we don't have exactly -1.645 in the table.)

When $\mu = 21.9$:

$$\begin{aligned}
P(\text{reject } H_0|\mu = 21.9) = P(\overline{x} < 18.27|\mu = 21.9) &= P\left(z < \frac{18.27-21.9}{\sigma/\sqrt{n}}\right) \\
&= P(z < -3.45) \\
&= 0.0003
\end{aligned}$$

In other words, when the true value is 21.9, we'd have to be extremely unlucky to get a sample of data that leads us to reject H_0.

c) The probability of making a Type I error with any test is α by definition. Thus $P(\text{Type I}) = \alpha = 0.05$.

The probability of making a Type II error is 1 minus the power, so $P(\text{Type II}|\mu = 18) = 1 - P(\text{reject } H_0|\mu = 18) = 1 - 0.6026 = 0.3974$ where 0.6026 is the power calculated for $\mu = 18$ in the previous part.

3. **[12]** Suppose that 70% of cars contain only one occupant (the driver); the rest contain at least two occupants.

a) If you observe 10 cars, what is the probability that at least 7 of the cars have only one occupant? What is the probability that at least 8 of the cars have only one occupant?

b) If you observe 100 cars, what is the probability that at least 70 have only one occupant? What is the probability that at least 80 of the cars have only one occupant?

Answers:

a) We want to find $P(X \geq 7)$ where X follows a binomial distribution with $n = 10$ and $p = 0.7$:

$$\begin{aligned} P(X \geq 7) &= P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) \\ &= \binom{10}{7}.7^7.3^3 + \binom{10}{8}.7^8.3^2 + \binom{10}{9}.7^9.3^1 + \binom{10}{10}.7^10 \\ &= 0.2668 + 0.2335 + 0.1211 + 0.0282 \\ &= 0.6496 \end{aligned}$$

$P(X \geq 8)$ is the same as the above, but without the $P(X = 7)$ term, which gives $P(X \geq 8) = 0.3828$. If you didn't write down the intermediate calculations, you could also calculate this as $P(X \geq 8) = P(X \geq 7) - P(X = 7) = 0.6496 - \binom{10}{7}.7^7.3^3 = 0.6496 - 0.2668 = 0.3828$.

b) Since with $n = 100$ we have $np = 70 > 15$ and $n(1-p) = 30 > 15$, we can justify using the normal approximation, which is $X \sim N(np, \sqrt{np(1-p)}) \sim N(70, 4.583)$.

$P(X \geq 70) = 0.5$ is trivial, since 70 is the mean of the normal distribution.

For $X \geq 80$ we need to use a z calculation:

$$\begin{aligned} P(X \geq 80) &= P\left(z \geq \frac{80 - 70}{4.583}\right) = P(z \geq 2.18) \\ &= 0.0146 \end{aligned}$$

4. [18] A researcher interested in the incomes of Canadian households has conducted a simple random sample of 45 Canadian households. This sample has a mean of \$84,500 and standard deviation of \$49,000.

a) Assume that the population distribution is normal. Find a 95% confidence interval for the average Canadian household income.

b) Still assuming that the population distribution is normal, perform a two-sided test at the 5% significance level that the population mean equals \$75,000.

The researcher has strong doubts that the population is normal, and is particularly concerned that the data contains two outliers of \$321,000 and \$512,000. In fact, only 12 of the 45 sample values are above \$75,000.

c) Use a sign test to test the two-sided hypothesis that the median equals \$75,000 at the 1% significance level. What is the p-value of this test?

Answers:

a) When dealing with income, distributions are rarely normal, but are typically right-skewed, and so with only 45 observations we might worry about whether a z or t test is suitable. However, since the question explicitly tells us to assume the distribution is normal, and since σ is unknown, we'll use the t confidence interval:

$$\mu \in [\overline{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}]$$

where $df = n - 1 = 44$ and $\alpha/2 = 0.025$ gives us $t^* = 2.021$ (using the $df = 40$ row and $p = .025$ column of the t distribution table).

Plugging in values gives $t^* \frac{s}{\sqrt{n}} = 14762$ and so the confidence interval is:

$$\mu \in [84500 - 14762, 84500 + 14762]$$
$$\mu \in [69738, 99262]$$

b) Since this is a two-tailed test with the same α as part *a*), we can answer the question by just looking at whether \$75,000 is in the confidence interval. It is, and so we fail to reject H_0 in favour of H_a.

Alternatively (though it isn't necessary) we could perform a t-test:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{84500 - 75000}{49000/\sqrt{45}} = 1.301.$$

Looking along the row for $df = 40$ (since we don't have $df = 44$ in the table), we see that 1.301 is very close to the critical values for 0.1: slightly below it for $df = 40$ and slightly above it for $df = 50$. Since this is a two-tailed test, we need to double this value, and so can safely conclude that our p-value is approximately equal to 0.2—which is definitely larger than $\alpha = 0.05$, and so we fail to reject H_0 at the $\alpha = 0.05$ significance level.

c) To do a sign test, we treat the data as a binomial distribution with $n = 45$ and $p = 0.5$. Since $np = n(1-p) = 22.5 > 15$, we can use the normal approximation.

The probability of seeing 12 or fewer observations above the hypothesized median is:

$$P(X \leq 12) = P\left(z \leq \frac{12 - 22.5}{\sqrt{45(0.5)(0.5)}}\right) = P(z \leq -3.13) = .0009$$

However, since this is a two-sided test, we need to double this value[1] to get a p-value of .0018.

5. **[12]** A researcher wishes to test whether house prices in a particular city changed from 2013 to 2014. She uses a random sample of the selling prices of 32 houses in 2013; the sample has a mean of \$430,000 and standard deviation of \$94,000. A similar sample of 40 houses sold in 2014 has a mean of \$485,000 with sample standard deviation of \$115,000. The researcher checks the data and verifies that neither sample contains any significant outliers.

a) Write down the null and alternative hypotheses and perform the test at the 95% confidence level. What is the p-value of your test?

b) Calculate a 99% confidence interval for the change in house prices from 2013 to 2014.

Answers:

a) The null and alternative hypotheses are:

$$H_0 : \mu_{2014} - \mu_{2013} = 0$$
$$H_a : \mu_{2014} - \mu_{2013} \neq 0$$

This could also be written as:

$$H_0 : \mu_{2014} = \mu_{2013}$$
$$H_a : \mu_{2014} \neq \mu_{2013}$$

To test this, since the σ values are unknown and we don't have a good reason to suppose that the two σ values are equal, we need to use the usual t test for a difference in means without pooling:

$$\begin{aligned} t &= \frac{(\overline{x}_A - \overline{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{55000 - 0}{\sqrt{\frac{94000^2}{32} + \frac{115000^2}{40}}} \\ &= \frac{55000}{24632} = 2.23 \end{aligned}$$

[1] The reason we double this is because, if the median really is 75000, seeing a sample with 33 or more of the 45 values *above* the mean is just as extreme as seeing a sample of 12 or fewer *below* the mean. We could instead also add $P(X \geq 33) = P(z \geq 3.13) = .0009$ which of course equals the $P(z \leq -3.13) = .0009$ value we got above: hence simplying doubling one of these values is a useful shortcut for symmetric distributions.

For converting this into a p-value we need to know the degrees of freedom. As discussed in class, we have three choices: we can simply use the smaller n minus 1 (so $df = 31$); or we can note that the n and s values are not too dissimilar and so we can use $df = n_A + n_B - 2 = 70$; or we can use the Satterthwaite approximation from the formula sheet, which gives $df = 69.96$.

Given our available df values, this means we'll either look at the $df = 30$ of $df = 60$ rows. In both cases, 2.23 lies between the $p = .01$ and $p = .02$ critical values. Since we are conducting a two-tailed test, however, we need to double these, and conclude that our p-value is between .02 and .04.

Since this range of values is smaller than $\alpha = .05$, we reject H_0 in favour of H_a; the data gives us statistically significant evidence that the mean house price increased from 2013 to 2014.

b) Our confidence interval is:

$$\begin{aligned}\mu_{2014} - \mu_{2013} \in [&(\overline{x}_{2014} - \overline{x}_{2013}) - t_{df,\alpha/2} SE(\overline{x}_{2014} - \overline{x}_{2013}),\\ &(\overline{x}_{2014} - \overline{x}_{2013}) + t_{df,\alpha/2} SE(\overline{x}_{2014} - \overline{x}_{2013})]\end{aligned}$$

The critical t value depends on our degrees of freedom, either 70 or 31 (see the justification given in the previous part of the question), and so we'll use either the $df = 30$ or $df = 60$ rows.

For $df = 30$, we get a 99% critical value of 2.750; for $df = 60$ we get a critical value of 2.660.

The $SE(\cdot)$ term we already calculated in the denominator of the t statistic above—it equals:

$$\sqrt{\frac{94000^2}{32} + \frac{115000^2}{40}} = 24632$$

Our 99% confidence interval is thus:

$$\mu_{2014} - \mu_{2013} \in [55000 - 2.750(24632), 55000 + 2.750(24632)] = [-12738, 122738]$$

if we use the $df = 30$ critical value, and:

$$\mu_{2014} - \mu_{2013} \in [55000 - 2.660(24632), 55000 + 2.660(24632)] = [-10521, 120521]$$

if using the $df = 60$ critical value.