Queen's University
Faculty of Arts and Sciences
Department of Economics
Economics 250 2006 Final

Instructions: 3 Hours

READ CAREFULLY. Calculators are permitted (no red stickers). At the end of the exam are several formulae and tables for the binomial, normal and t distributions. Answers are to be written in the examination booklet. Remember most of the grades are awarded for how you set up the problem and NOT for the calculation itself.

Please Note: Proctors are unable to respond to queries about the interpretation of exam questions. Do your best to answer exam questions as written.

You are to answer **ALL** questions. **SHOW ALL YOUR WORK**. There are a total of 100 possible marks to be obtained and marks are indicated for each question.

Answer all 12 questions.

1. (**10 marks**) Relative efficiency can be thought of a relative precision in a confidence interval. You are given the following sample information on 3 independent observations from a normally and identically distributed population with *known* variance of 4:

$$3, 8, 10$$

Consider 2 estimators of the population mean μ

$$\begin{aligned} \ddot{X} &= \frac{1}{2}X_2 + \frac{1}{2}X_3 \\ \bar{X} &= \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3 \end{aligned}$$

(a) Show that each of these is unbiased

$$\begin{aligned} E[\ddot{X}] &= \frac{1}{2}E[X_2] + \frac{1}{2}E[X_3] = \mu \\ E[\bar{X}] &= \frac{1}{3}E[X_1] + \frac{1}{3}E[X_2] + \frac{1}{3}E[X_3] = \mu \end{aligned}$$

(b) What is relative efficiency and calculate it

$$\begin{aligned} V[\ddot{X}] &= \frac{1}{4} \times V[X_2] + \frac{1}{4} \times V[X_2] = \frac{1}{4} \times 4 + \frac{1}{4} \times 4 = 2 \\ V[\bar{X}] &= \frac{\sigma^2}{n} = \frac{4}{3} \\ RE &= \frac{V[\ddot{X}]}{V[\bar{X}]} = \frac{2}{\frac{4}{3}} = 1.5 \text{ (50\% more efficent)} \end{aligned}$$

(c) Calculate the 95% confidence interval for both estimators and explain in what sense the two intervals are different.

$$\begin{aligned} \ddot{X} &= \frac{1}{2} \times 8 + \frac{1}{2} \times 10 = 9 \\ \ddot{X} \pm Z_{\frac{\alpha}{2}} \times \sqrt{2} &= 9 \pm 1.96 \times \sqrt{2} = [7.04 \quad 11.772] \\ \bar{X} &= \frac{1}{3} \times (3 + 8 + 10) = 7 \\ \bar{X} \pm Z_{\frac{\alpha}{2}} \times \sqrt{\frac{4}{3}} &= 7 \pm 1.96 \times \sqrt{\frac{4}{3}} = [\quad 4.736 \quad 9.26] \end{aligned}$$

:Notice that the confidence interval for the sample mean is much more tightly distributed around its estimate of the population mean.

(d) Interpret the intervals: If we do a large numbero fo confidence intervals we know that 95% of them will bracket the true population mean μ

2. (20 marks) A firm wishes to test whether its recent marketing blitz has had an impact on its sales. Before the blitz sales per month were thought to be 12 million dollars. Over the next year 12 monthly observations were taken with a mean sample sales of 12.5 million and a sample standard deviation of 1.

   (a) Develop a formal hypothesis to test and state whether it should be one or two sided

   (b) What assumptions are you making?

   (c) Is the hypothesis retained at the 1% level? What about the 5% level?

   (d) Explain why one or two sided tests can make a difference. Do they have different probabilities of Type I error?

   (e) What is the p-value of this test?

   (f) Calculate power of the test (at 5% level) if the true population 12.5

$$\begin{aligned} H_o &: \mu \leq 12 \\ H_a &: \mu > 12 \end{aligned}$$

   assuming that central limit theorem applies (sample size is very small so that is unlikely). Also assuming each observatioon is independently and identically distributed but since this is sales this is again unllikely since there is seasonal patterns.

$$t_{cal} = \frac{\bar{X} - 12}{\sqrt{\frac{s^2}{n}}} = \frac{12.5 - 12}{\sqrt{\frac{1^2}{12}}} = 1.73$$

   the $t_{11,.01} = 2.718 \Longrightarrow$retain the null hypothesis at the 1% level. $t_{11,.05} = 1.796 \Longrightarrow$just retain. The one sided tests and the two-side have the same probability of Type I error (α) but there are difference critical values. For instance $t_{11,\frac{.05}{2}} = 2.201$ so the null is comfortably retained for the two-sided alternative. There is a difference in power and there is a jchance that the null will be rejected for one-side but retained for a two-sided

$$p - value = P(t > t_{11}) \approx .05$$

$$\begin{aligned} \mu_0 + t_{11,.05} \times \sqrt{\frac{s^2}{n}} &\implies 12 + 1.796 \times \sqrt{\frac{1}{12}} \\ \text{we reject the null of } \mu &\leq 12 \text{ anytime the sample mean is greater than } 12.518 \\ P(X &\bar{>} 12.518 \mid \mu = 12.5) \\ &\approx P(Z > \left(\frac{12.518 - 12.5}{\sqrt{\frac{1}{12}}}\right) \approx .5 \end{aligned}$$

   therefore the power is around 50%.

3. (**10marks**) Suppose their is a convention to select a leader of the party and that there are 4 candidates labelled C_1 C_2, C_3, and C_4candidates. Initially each candidate has the following support

$$P(C_1) = .35 \quad P(C_2) = .28 \quad P(C_3) = .19 \quad P(C_4) = .18$$

(a) There is voting and each time the lowest voting candidate is dropped and we assume that the support for candidate 2 receives 80% of the vote (round to the nearest integer)) from the losing candidate and the remainder are divided evenly between the others.

(b) What is the outcome of this election?

(c) Is this a binomial problem. why or why not?

$$\begin{aligned}
Round \quad 1 &\implies C_4 \text{ drops out} \\
P(C_1) &= .35 + .18 \times .1 = .368 \\
P(C_2) &= .28 + .18 \times .8 = .424 \\
P(C_3) &= .19 + .18 \times .1 = .208 \\
Round \quad 2 &\implies C_3 \text{ drops out} \\
P(C_1) &= .368 + .208 \times .2 = .4276 \\
P(C_2) &= .424 + .208 \times .8 = .5904
\end{aligned}$$

and Candidate 2 wins on the third vote . No it is not a binomial problem as each probability depends on the last outcome. It is not independent.

4. **(5 marks)** A student in 250 remembers that Gregory has told him that more informationin estimating the population mean is preferred to less. The student has 3 observations on X_i available and so just repeats them to give a total of 6 .

(a) Can you illustrate what the student is thinking in the general case when 3 independent observations are augments with another 3 independent observations

(b) Can you show that by simply replicating the 3 observations that there is no fall in the variance of the estimator?

$$\frac{\sigma^2}{3} \text{ vs } .\frac{\sigma^2}{6}$$

$$\begin{aligned}
&\text{The data is } X_1, X_2, X_3, X_1, X_2, X_3 \\
\tilde{X} &= \frac{1}{6}(\ X_1 + X_2 + X_3 + X_1 + X_2 + X_3) \implies E[\tilde{X}] = \mu \\
V[\tilde{X}] &= \frac{\sigma^2}{6} + \frac{1}{36}\left(2C[X_1, X_1] + 2C[X_2, X_2] + 2C[X_3, X_3]\right) \\
&\frac{\sigma^2}{6} + \frac{1}{36}\left(2 \times SD[X_1 1] \times SD[X_1] + 2 \times SD[X_2] \times SD[X_2] + 2 \times SD[X_3] \times SD[X\right. \\
&= \frac{\sigma^2}{3}
\end{aligned}$$

and no gain in efficency

5. (**15marks)** There are two populations of independent data available on literacy and we wish to test whether the two population proportions for literacy are the same. If the first sample, there is 25% cannot read with a sample size of 50 and in the second 20% cannot read with a sample size of 75.

   (a) State the formal hypothesis tests and state any assumptions you are going to make to test this hypothesis?

   (b) What is the pooled estimator of the population proportion

   (c) Construct a hypothesis test using the most efficient estimator of the population proportion for the variance at the 10% level

   (d) Construct an hypothesis test using an inefficient (though) unbiased estimator for the variance at the 10% level

   (e) What is the 90% confidence interval in (b) and what can you say about hypothesis tests from this confidence interval

$$\begin{aligned} H_o &: \pi_1 = \pi_2 \\ H_a &: \pi_1 \neq \pi_2 \end{aligned}$$

we are assuming indpedence within and across samples

$$p_{pool} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{.25 \times 50 + .2 \times 75}{50 + 75} = .22$$

$$Z_{cal} = \frac{.25 - .20}{\sqrt{\frac{p_{pool} \times (1 - p_{pool})}{n_1} + \frac{p_{pool} \times (1 - p_{pool})}{n_2}}} = \frac{.25 - .20}{\sqrt{\frac{.22 \times .78}{50} + \frac{.22 \times .78}{75}}} = .66$$

do not reject the null that they are the same

$$Z_{cal} = \frac{.25 - .20}{\sqrt{\frac{p_{pool} \times (1 - p_{pool})}{n_1} + \frac{p_{pool} \times (1 - p_{pool})}{n_2}}} = \frac{.25 - .20}{\sqrt{\frac{.22 \times .78}{50} + \frac{.22 \times .78}{75}}}$$

$$p_{pool} \pm Z_{\frac{\alpha}{2}} \times \sqrt{\frac{p_{pool} \times (1 - p_{pool})}{n_1 + n_2}}$$

$$.22 - 1.645 \times \sqrt{\frac{.22 \times .78}{75}}$$

$$[.141 \quad .299]$$

Any hypothesis test in this interval would be retianed at the 10% level of significance

6. **(5 marks)** When calculating the p-values students in Econ 250 do not have access to the t-distribution for their tests and must use the Z tables.

   (a) Why is that the case?

   (b) Suppose that a p-value with the t-distribution of $n = 5$ for a two-sided test is exactly 0.05. What would the student's use to approximate this from the standard normal table provide

   (c) Explain how this reporting is misleading if someone were testing at the 1% level? assist the explanation.

The t-distribution is indexed by its degress of freedom.

There degrees of freedom are in principle infininite

There would need to be a t-table for each degrees of freedom. Indeed that is alot!

$$\begin{aligned} p-value &= 2 \times (t_4 > tcal) = .05 \\ \text{must be that } tcal &= 2.776 \text{ since } t_{\frac{.05}{2},4} = 2.776 \\ & \text{Using the normal to approximate p-value} \\ p-value &\approx 2 \times (P(Z > 2.776) = 2 \times (1 - .9973) = .0054 \end{aligned}$$

What happens is that with the approximate p-value being so small, this would lead readers to reject null more frequently. :Large test scores using the normal lead to small approximate p-values whereas especially when degrees of freedom are few (like 4) these large test scores can happen more frequently.

7. (**10 marks**) We have said that the normal approximation to the binomial gets more accurate as the sample size increases. This problem explores that phenomenonwith $\pi = .3$..

   (a) Suppose the number of sucesses is 5 with a number of trials of 10 and 20. Calculate the probability for both trial sizes using the binomial distribution and the normal approximation

$$\begin{aligned} P(X &= 5) = \binom{10}{5} .3^5 .7^5 = .10 \\ P(X &= 5) \approx P\left(\frac{4.5-3}{\sqrt{10 \times .3 \times .7}} \leq \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \geq \frac{5.5-3}{\sqrt{10 \times .3 \times .7}} \right) \\ &\approx P(1.03 \leq Z \leq 1.725) \\ F_Z(1.725) - F(1.03) &= .9582 - .8485 = .1097 \end{aligned}$$

$$
\begin{aligned}
P(X &= 5) = \binom{20}{5}.3^5.7^{15} = .179 \\
P(X &= 5) \approx P\left(\frac{4.5-6}{\sqrt{20\times.3\times.7}} \leq \frac{X-n\pi}{\sqrt{n\pi(1-\pi)}} \geq \frac{5.5-6}{\sqrt{20\times.3\times.7}}\right) \\
&\approx P(-.73 \leq Z \leq . - .243) \\
F_Z(.73) - F_Z(.24) &= .7673 - .5948 = .1725
\end{aligned}
$$

8. (**10 marks**) The central limit theorem is a beautiful thing. Suppose we have 20 uniform independent objects over the interval 4 to 20.

   (a) Explain the central limit theorem in this context and illustrate

   (b) If the sample size is 25 what is the distribution of the sample mean

   (c) What is the probability of a single observation being 2 standard deviations around the mean

   (d) What is the probability of the sample mean being plus or minus 2 standard deviations around its mean

   The central limit is a theorem that says as the sample size gets large for practically any distribution that is independently and identically distributed the sample mean willbe approximately normally distributed in large samples.

$$
X \sim U[4 \quad 20] \Rightarrow \mu = \frac{a+b}{2} = 12 \quad \sigma^2 = \frac{(b-a)^2}{12} = \frac{(20-4)^2}{12} = 21.3
$$

$$
\bar{X}_{25} \sim N(12, \frac{21.3}{25})
$$

$$
\begin{aligned}
\sigma &= \sqrt{21.3} = 4.6 \Longrightarrow 2 \times 4.6 = 9.2 \\
P(12 - 9.2 &< X < 12 + 9.2) = 100\%
\end{aligned}
$$

   since we are looking at normal distribution we can go immediately to standard deviationand see that

$$
P - 2 < Z < 2) = .95
$$

9. (**10 marks**) Suppose we are hypothesis testing at the 5% level,

   (a) With 20 true independent hypothesis, how often should we reject the null hypothesis?

   (b) Another way to think of this is what is the probability of not rejecting the null hypothesis in 20 successive true tests.

(c) Suppose there are 2 hypothesis tests that are true and independent, what is the probability of rejecting either of them at the 5% level?

$$\begin{aligned} X &= \text{number of rejections} \\ E[X] &= n \times .05 = 1 \end{aligned}$$

$$P(X = 0) = \binom{20}{0} .05^0 .95^{20} = .359$$

$$\begin{aligned} P(R_1 \cup R_2) &= P(R_1) + P(R_2) - P(R_1) \times P(R_2) \\ .05 + .05 - .05 \times .05 &= .0975 \end{aligned}$$

Formula Sheet

Statistics Formulas

Notation

- All summations are for $i = 1, \ldots, n$ unless otherwise stated.
- $\sim$ means 'distributed as'

Population Mean

$$\mu \ = \ \frac{1}{N} \sum_{i=1}^{N} x_i$$

Sample Mean

$$\bar{X} \ = \ \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Population Variance

$$\begin{aligned} \sigma^2 &= \frac{1}{N} \sum_{i=1}^{N} [x_i \ - \ \mu]^2 \\ &= \frac{1}{N} \sum_{i=1}^{N} {x_i}^2 - \mu^2 \end{aligned}$$

Sample Variance

$$s^2 \ = \ \tfrac{1}{n-1} \textstyle\sum_{i=1}^{n} [x_i \ - \ \bar{X}]^2$$

Alternatively

$$s^2 = \frac{1}{n-1} [\sum_{i=1}^{n} x_i^2 \ - \ n\bar{X}^2]$$

Grouped Data (with k classes)

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{j=1}^{k}\nu_j f_j \quad \text{where } \nu_j \text{ is the class mark for class } j \\ s^2 &= \frac{1}{n-1}\sum_{j=1}^{k} f_j(\nu_j - \bar{X})^2 \end{aligned}$$

Probability Theory

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \quad \text{(additive law)} \\ P(A \cap B) &= P(B)P(A \mid B) \quad \text{(multiplicative law)} \end{aligned}$$

If the E_i are mutually exclusive and exhaustive events for $i = 1, \ldots, n$, then

$$\begin{aligned} P(A) &= \sum_{i}^{n} P(A \cap E_i) = \sum_{i}^{n} P(E_i)P(A \mid E_i) \\ P(E_i \mid A) &= \frac{P(E_i)P(A \mid E_i)}{P(A)} \quad \text{(Bayes' Theorem)} \end{aligned}$$

Counting Formulae

$$P_R^N = \frac{N!}{(N-R)!}$$

$$C_R^N = \binom{N}{R} = \frac{N!}{(N-R)!R!}$$

Random Variables

Let X be a discrete random variable, then:

$$\begin{aligned}
\mu_x &= E[X] = \sum_x xP(X=x) \\
\sigma_x^2 &= V[X] = E[(x-\mu_x)^2] = \sum_x (x-\mu_x)^2 P(X=x) \\
\sigma_x^2 &= E[X^2] - (E[X])^2
\end{aligned}$$

The **covariance** of X and Y is

$$\begin{aligned}
Cov[X,Y] &= \sigma_{xy} = E\left[(X-\mu_x)(Y-\mu_y)\right] \\
&= E[XY] - E[X] \times E[Y]
\end{aligned}$$

The **correlation coefficient** of X and Y

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

If X and Y are independent random variables and a, b, and c are constants, then:

$$\begin{aligned}
E[a+bX+cY] &= a + b\mu_x + c\mu_y \\
V[a+bX+cY] &= b^2\sigma_x^2 + c^2\sigma_y^2
\end{aligned}$$

If X and Y are correlated then

$$V[a+bX+cY] = b^2\sigma_x^2 + c^2\sigma_y^2 + 2bc\sigma_{xy}$$

Coefficient of Variation (CV)

$$\begin{aligned} CV &= \frac{\sigma}{\mu} \times 100\% \text{ for population} \\ CV &= \frac{s}{\bar{X}} \times 100\% \text{ for sample} \end{aligned}$$

Univariate Probability Distributions

Binomial Distribution: For $x = 0, 1, 2, ..., n$ and :

$$Pr[X = x] = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

$$E[X] = n\pi$$
$$V[X] = n\pi(1-\pi)$$

Uniform Distribution: For $a < x < b$**:**

$$f(x) = \frac{1}{b-a}$$

$$E[X] = \frac{a+b}{2}$$
$$V[X] = \frac{(b-a)^2}{12}$$

Normal Distribution:For $-\infty < x < \infty$:

$$f(x) = (2\pi\sigma^2)^{-1/2} exp[\frac{-(x-\mu)^2}{2\sigma^2}]$$

$$E[X] = \mu$$
$$V[X] = \sigma^2$$

$$\begin{aligned} X &\sim N(\mu_X, \sigma_X^2) \\ Z &= \frac{X - \mu_X}{\sigma_X} \sim N(0,1) \end{aligned}$$

Estimators in General

If $\hat{\theta} \sim N(\theta, V[\hat{\theta}])$, say, then under appropriate conditions:

$$Z = \frac{\hat{\theta} - \theta}{SD[\hat{\theta}]} \sim N(0,1)$$

Confidence Intervals:

(a) $100(1-\alpha)\%$ confidence interval for $\theta : \hat{\theta} \pm Z_{\alpha/2} SD[\hat{\theta}]$, with known variance.

(b) If $V[\hat{\theta}]$ is unknown, then

$$t = \frac{\hat{\theta} - \theta}{SE[\hat{\theta}]} \sim t$$

with appropriate degrees of freedom. $100(1-\alpha)\%$ confidence interval for $\theta : \hat{\theta} \pm t_{\alpha/2} SE[\hat{\theta}]$, where $SE[\hat{\theta}]$ is an estimator of $SD[\hat{\theta}]$.

Estimating Means and Proportions

$$\bar{X} = \frac{1}{n}\sum X_i; \quad s^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$$

$$E[\bar{X}] = \mu; \quad V[\bar{X}] = \frac{\sigma^2}{n}$$

$$p = \frac{X}{n}; \quad E[p] = \pi; \quad V[p] = \frac{\pi(1-\pi)}{n}$$

Differences of Means and Proportions for Independent Samples

$$s^2_{pool} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2; \quad V[\bar{X}_1 - \bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

If variances are assumed (ie. $\sigma_1^2 = \sigma_2^2$) to be the same we may , estimate

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s_{pool}^2 (\frac{1}{n_1} + \frac{1}{n_2})}$$

$$\begin{aligned} p_{pool} &= \frac{X_1 + X_2}{n_1 + n_2}; \\ E[p_1 - p_2] &= \pi_1 - \pi_2; \\ V[f_1 - f_2] &= \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2} \end{aligned}$$