Economics 325–003 **Term I 2016**

Introduction to Empirical Economics **Hiro Kasahara**

Final Exam

You have 2 hours and 30 minutes. When you use your calculator, try to keep the number up to four digits below decimal point. For example, if you will get 6.164414003 in calculator, pelase use 6.1644 for your subsequent calculation. When you refer to the standard normal distribution table, report the number that is closest or report two numbers. Good luck!

1. (20 points) State whether each of the following is true or false. No explanation necessary.

(a) The confidence interval for population parameter θ with confidence level $1-\alpha$ is always given by $\hat{\theta} \pm z_{\alpha/2}\sqrt{Var(\hat{\theta})}$, where $\hat{\theta}$ is an unbiased point estimator of θ and $z_{\alpha/2}$ is the critical value such that $P(Z > z_\alpha) = \alpha$ given $Z \sim N(0,1)$.

Answer: False for two reasons. First, when the population distribution is not normal and the sample size n is not large, the standardized random variable $\frac{\hat{\theta}-\theta}{\sqrt{Var(\hat{\theta})}}$ does not have a standard normal distribution. Second, when we construct the confidence interval for the sample variance under the assumption that the population is normal, the constructed confidence interval is not symmetric around its point estimator.

(b) Let $\{X_1, X_2, ..., X_n\}$ be n observations, each of which is randomly drawn from a distribution with mean μ and variance σ^2. Let $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then, the distribution of a statistic $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ is always given by t-distribution with the degree of freedom $n-1$.

Answer: False — because t statistic does not have t-distribution when X_i is not normally distribution.

(c) If the null hypothesis is not rejected based on sample evidence, the researcher has proven beyond any doubt that the null hypothesis is true.

Answer: False — the rejection of the null hypothesis only implies that it is unlikely that the null hypothesis is true. It is not a proof that the null hypothesis is true.

(d) The central limit theorem cannot be applied to discrete random variables.

Answer: False — the central limit theorem can be applied to discrete random variables, such as Bernouilli random variable.

2. (10 points) Define $W = (X - E(X))/\sqrt{Var(X)}$ and $Z = (Y - E(Y))/\sqrt{Var(Y)}$. (You don't necessarily need to use the summation operator but, if you like, you can use the summation operator.)

   (a) (5 points) Prove that $E[W] = 0$.

   (b) (5 points) Prove that $Cov(W, Z) = Corr(X, Y)$.

   Answer: First, we have

$$E[W] = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{x_i - E(X)}{\sqrt{Var(X)}} P_{ij}^{X,Y} = \frac{1}{\sqrt{Var(X)}} \left(\sum_{i=1}^{n} x_i (\sum_{j=1}^{m} P_{ij}^{X,Y}) - E(X) \right)$$
$$= \frac{1}{\sqrt{Var(X)}} \left(\sum_{i=1}^{n} x_i p_i^X - E(X) \right) = \frac{1}{\sqrt{Var(X)}} (E(X) - E(X)) = 0.$$

   Similarly, we may prove that $E(Z) = 0$. Then,

$$Cov(W, Z) = E\left(\frac{X - E(X)}{\sqrt{Var(X)}} \frac{Y - E(Y)}{\sqrt{Var(Y)}} \right) = \frac{1}{\sqrt{Var(X)}\sqrt{Var(Y)}} E\left((X - E(X))(Y - E(Y))\right)$$
$$= \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = Corr(X, Y).$$

3. (6 points) Let X_1 and X_2 be a pair of random variables. Show that the covariance between the random variables $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$ is 0 if and only if X_1 and X_2 have the same variance.

   Answer: Let $E[X_1] = \mu_1$ and $E[X_2] = \mu_2$. Then,

$$\begin{aligned} Cov(Y_1, Y_2) &= Cov(X_1 + X_2, X_1 - X_2) \\ &= E[\{(X_1 + X_2) - (\mu_1 + \mu_2)\}\{(X_1 - X_2) - (\mu_1 - \mu_2)\}] \quad \text{by definition of covariance} \\ &= E[\{(X_1 - \mu_1) + (X_2 - \mu_2)\}\{(X_1 - \mu_1) - (X_2 - \mu_2)\}] \\ &= E[(X_1 - \mu_1)^2 - (X_2 - \mu_2)^2 + (X_1 - \mu_1)(X_2 - \mu_2) - (X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E[(X_1 - \mu_1)^2 - (X_2 - \mu_2)^2] \\ &= E[(X_1 - \mu_1)^2] - E[(X_2 - \mu_2)^2] \\ &= Var(X_1) - Var(X_2). \end{aligned}$$

   Therefore, $Cov(Y_1, Y_2) = 0$ if and only if $Var(X_1) = Var(X_2)$.

4. (10 points) Suppose that X_1, X_2, X_3 are randomly sampled from a population with mean μ and variance σ^2. Consider the following two point estimators of μ:

$$\hat{\mu}_1 = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3,$$
$$\hat{\mu}_2 = \frac{1}{4}X_1 + \frac{1}{4}X_2 + \frac{1}{2}X_3,$$

(a) (5 points) Prove that both estimators are unbiased.

(b) (5 points) Prove that $\hat{\mu}_1$ is more efficient than $\hat{\mu}_2$.

Answer: (a) $E[\hat{\mu}_1] = (1/3 + 1/3 + 1/3)\mu = \mu$ and $E[\hat{\mu}_2] = (1/4 + 1/4 + 1/2)\mu = \mu$. (b) $Var(\hat{\mu}_1) = (1/3)^2\sigma^2 + (1/3)^2\sigma^2 + (1/3)^2\sigma^2 = \sigma^2/3$ and $Var(\hat{\mu}_2) = (1/4)^2\sigma^2 + (1/4)^2\sigma^2 + (1/2)^2\sigma^2 = (1/16 + 1/16 + 1/4)\sigma^2 = (3/8)\sigma^2$, where the covariance terms are equal to zero because of random sampling. Because $3/8 > 1/3$, $\hat{\mu}_1$ has smaller variance than $\hat{\mu}_2$ and therefore $\hat{\mu}_1$ is more efficient than $\hat{\mu}_2$.

5. (22 points) Let X_1, X_2, ..., X_n be $n = 9$ observations, each of which is randomly drawn from normal distribution with mean μ and variance σ^2. The value of μ is not known while σ^2 is known and equal to 100. We are interested in testing the null hypothesis $H_0 : \mu = 5$ against the alternative hypothesis $H_1 : \mu < 5$. Consider the following two different test statistics (i) $\tilde{X} = 0.6X_1 + \sum_{i=2}^{9} 0.05X_i = 0.6X_1 + 0.05X_2 + 0.05X_3 + ... + 0.05X_9$ and (ii) $\hat{X} = (1/4)(X_1 + X_2 + X_3 + X_4)$. Suppose that the realized values of $\tilde{X}$ and $\hat{X}$ are given by $\tilde{X} = -2.1$ and $\hat{X} = -2.0$, respectively.

(a) (5 points) Compute the variance of $\tilde{X}$.

(b) (5 points) Test the null hypothesis $H_0 : \mu = 5$ agains the alternative hypothesis $H_1 : \mu < 5$ using the test statistic $\tilde{X}$ at the significant level $\alpha = 0.1$.

(c) (6 points) Compute the power of test using the test statistic $\tilde{X}$ at the significance level $\alpha = 0.1$ when the value of μ is equal to -3.

(d) (6 points) Compute the power of test using the test statistic $\hat{X}$ at $\alpha = 0.1$, and discuss which of tests, the test based on $\tilde{X}$ or the test based on $\hat{X}$, is more powerful and which of tests you recommend.

Answer: (a) $Var(\tilde{X}) = (0.6)^2 Var(X_1) + \sum_{i=2}^{9}(0.05)^2 Var(X_i) = (0.6^2 + 8 \times (0.05^2)) \times \sigma^2 = 0.38 \times 100 = 38$.

(b) Under $H_0 : \mu = 5$, $\tilde{X} \sim N(5, 38)$. Therefore, we reject H_0 if $\tilde{X} < 5 - 1.28 \times \sqrt{38} = -2.8904$. Because the realized value of $\tilde{X}$ is -2.1 and does not fall in the rejection region, the test based on $\tilde{X}$ does not reject H_0.

(c) The power of test based on $\tilde{X}$ is $\Pr(\text{Reject } H_0|\mu = -3) = \Pr(\tilde{X} < -2.89|\mu = -3) = \Pr((\tilde{X} - (-3))/\sqrt{38} < (-2.8904 - (-3))/6.1644|\mu = -3) = \Pr(Z < 0.0178) \approx 0.5080$.

(d) Under $H_0 : \mu = 5$, $\hat{X} \sim N(5, 25)$. Therefore, we reject H_0 if $\hat{X} < 5 - 1.28 \times \sqrt{25} = -1.4$. The power of test based on $\hat{X} = \Pr(\text{Reject } H_0|\mu = -3) = \Pr(\tilde{X} < -1.4|\mu = -3) = \Pr((\tilde{X} - (-3))/5 < (-1.4 - (-3))/5|\mu = -3) = \Pr(Z < 0.32) = 0.6255$, which is larger than 0.5080. Therefore, $\hat{X}$ is more powerful than $\tilde{X}$, and we recommend using $\hat{X}$ over $\tilde{X}$.

Grading: When students make a simple calculation mistake, subtract 2 point out of 5 points. As long as the logic is correct after making a simple calculation mistake, do not penalize students for subsequent questions. Also, students might use different digits below decimal points so that the actual number students get could be slightly different from those in the answer key.

6. (12 points) Table I reports the number of smokers among 400 patients with lung-cancer and the number of smokers among 400 patients with other diseases, which Prof. Kasahara randomly sampled from a population of patients with lung-cancer and from a population of patients with other diseases, respectively. Denote the population proportions of smokers for patients with lung-cancer by p_x and for patients with other diseases by p_y. We test the null hypothesis of $H_0 : p_x \leq p_y$ against $H_1 : p_x > p_y$.

   (a) (6 points) Using the data reported in Table I, test the null hypothesis of $H_0 : p_x = p_y$ against $H_1 : p_x > p_y$ at the significance level $\alpha = 0.05$.

   (b) (6 points) Compute the p-value of the test.

   Answer: (a) Under H_0, let $p_x = p_x = p_0$. Then $\hat{p}_x - \hat{p}_y \sim N(0, p_0(1 - p_0)/n_x + p_0(1 - p_0)/n_y)$, where $n_x = n_y = 400$. Because $\hat{p}_0 = \frac{370+350}{400+400} = \frac{9}{10}$, the estimator for $\sqrt{Var(\hat{p}_x - \hat{p}_y)}$ is given by $\sqrt{\hat{p}_0(1 - \hat{p}_0)/400 + \hat{p}_0(1 - \hat{p}_0)/400} = \sqrt{2 \times \frac{1}{10}(1 - \frac{1}{10})/400} = 0.0212$. We reject H_0 if $\hat{p}_x - \hat{p}_y > 1.645 \times 0.0212 = 0.0349$. The realized value of $\hat{p}_x - \hat{p}_y$ is equal to $370/400 - 350/400 = 20/400 = 0.05$. Because $0.05 > 0.0349$, we reject H_0. [Alternatively, we can compute $\frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}_0(1-\hat{p}_0)/400+\hat{p}_0(1-\hat{p}_0)/400}} = 2.358$, which is larger than 1.64 and therefore we reject H_0.]

   (b) $\frac{\hat{p}_x - \hat{p}_y - 0}{0.0212} = \frac{0.05}{0.0212} = 2.358$. The p-value is given by $Pr(Z > 2.358) \approx 1 - 0.9909 = 0.0081$.

   Grading: When students make a simple calculation mistake, subtract 2 point out of 5 points. As long as the logic is correct after making a simple calculation mistake, do not penalize students for subsequent questions. Also, students might use different digits below decimal points so that the actual number students get could be slightly different from those in the answer key.

Table I: Tobacco Smoked Daily Over the 10 years (Hypothetical Example)

	No. of	No. of
Disease Group	Non-Smokers	Smokers
Men:		
400 lung-cancer patients	30	370
400 patients with other diseases	50	350

7. (10 points) The survey asks randomly sampled eligible voters in the U.S. whether he or she would vote for Trump. An individual i's voting preference is recorded as $X_i = 1$ if she/he would vote for Trump and as $X_i = 0$ otherwise. We also define $Y_i = 1$ if she/he would not vote for Trump and $Y_i = 0$ otherwise. Note that $Y_i = 1 - X_i$.

   Let p_x and p_y represent a fraction of Trump supporters and Trump non-supporters in population so that $p_x = E[X_i]$ and $p_y = E[Y_i]$. We are interested in estimating the population difference $p_x - p_y$.

   Suppose that we randomly sample $n = 400$ voters and we construct two data sets, $\{X_1, X_2, ..., X_n\}$ and $\{Y_1, Y_2, ..., Y_n\}$, based on the same sample of $n = 400$ voters. Let $\hat{p}_x = \frac{1}{n}\sum_{i=1}^n X_i$ and $\hat{p}_y = \frac{1}{n}\sum_{i=1}^n Y_i$.

   (a) (5 points) Derive the variance of the difference between $\hat{p}_x$ and $\hat{p}_y$ in terms of p_x and p_y and n.

   (b) (5 points) Suppose that $\hat{p}_x = 0.52$ and $\hat{p}_y = 0.48$. Construct the 95 percent confidence interval for $p_x - p_y$, assuming that n is large enough to apply the Central Limit Theorem.

   Answer: (a) $Var(\hat{p}_x - \hat{p}_x) = Var(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n Y_i) = Var(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n (1 - X_i)) = Var(\frac{1}{n}\sum_{i=1}^n (2X_i - 1)) = \frac{1}{n^2}\sum_{i=1}^n Var(2X_i - 1) = \frac{1}{n} 4Var(X_i) = \frac{4p_x(1-p_x)}{n} = \frac{4p_y(1-p_y)}{n}$.
   (b) $\hat{p}_x - \hat{p}_y = 0.04$ and the estimate for $\sqrt{Var(\hat{p}_x - \hat{p}_y)}$ is given by $\sqrt{\frac{4\hat{p}_x(1-\hat{p}_x)}{400}} = 0.050$. Therefore, the 95 percent confidence interval for $p_x - p_y$ is given by $0.04 \pm 1.96 \times 0.05$ or $[-0.058, 0.138]$.

   Grading: If the answer to (a) is $Var(\hat{p}_x - \hat{p}_x) = \frac{p_x(1-p_x)}{n} + \frac{p_y(1-p_y)}{n}$, students will get zero point for both (a) and (b).

8. (10 points) Suppose that we have 2 data sets, each of which is randomly sampled. For each data set, we construct a 90 percent confidence interval for the population parameter θ. For $i = 1, 2$, define a Bernoulli random variable X_i, where $X_i = 1$ if the confidence interval constructed from the i-th data set contains the population parameter θ and $X_i = 0$ if the confidence does not contain the population parameter θ. Let $\Pr(X_i = 1) = p$ for $i = 1, 2$. Define $\bar{X} = \frac{1}{2}\sum_{i=1}^2 X_i$.

(a) (5 points) What is the probability mass function of $\bar{X} = \frac{1}{2}\sum_{i=1}^{2} X_i$ if $p = 0.9$? [Hint: What are the possible values $\bar{X}$ can take? What is the probability that each value of $\bar{X}$ happens?]

(b) (5 points) Suppose that the realized value of $\bar{X}$ is equal to 0 because neither of two confidence intervals contains the population parameter θ.[1] Test the null hypothesis of $H_0 : p \geq 0.9$ against the alternative hypothesis of $H_1 : p < 0.9$ at the significance level $\alpha = 0.05$ by (i) constructing the rejection region and (ii) examining if the realized value of $\bar{X} = 0$ is in the rejection region or not.

Answer: (a) $\bar{X}$ takes the value of 0, 1/2, and 1. $\Pr(\bar{X} = 0) = \Pr((X_1, X_2) = (0, 0)) = (1 - p)^2 = 0.01$, $\Pr(\bar{X} = 1/2) = \Pr((X_1, X_2) = (0, 1)) + \Pr((X_1, X_2) = (1, 0)) = 2p(1 - p) = 0.18$, and $\Pr(\bar{X} = 1) = \Pr((X_1, X_2) = (1, 1)) = p^2 = 0.81$.

(b) We reject $H_0 : p = 0.9$ if $\bar{X} = 0$ at significance level $\alpha = 0.05$ because $\Pr(\bar{X} = 0) = 0.01$ and $\Pr(\bar{X} \leq 1) = 0.01 + 0.18 > 0.05$. Therefore, the rejection region is $\{\bar{X} = 0\}$. The realized value of $\bar{X}$ is zero and falls in the rejection region. Therefore, we reject H_0.

Grading: for (b), give 1 point for explicitly stating rejection region and 4 points for rejecting H_0 under the correct reasoning that $\Pr(\bar{X} = 0) = 0.01 < \alpha = 0.05$.

[1] Here, we assume that we know the true value of θ so that we know whether each of confidence interval contains the true parameter value or not.