eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

DIRECTIONS

1. Correctly answered questions give one point. There is only one correct answer to each question. Questions not correctly answered carry a penalty of -0.20 points, so it is better to leave a question unanswered rather than giving a wrong answer.
2. Our goal is to gauge your understanding and command of concepts learned during the course, not your visual sharpness. It is a fact, though, that in a multiple choice exam great attention has to be paid to the details. It is quite common that knowledgeable students waste their chances of a good grade because they do not pay sufficient attention to the precise wording of questions.

Please, read carefully before you answer!

3. It will probably help you to discard first answers that are clearly inadequate.
4. Let P_{ex} be the number of points obtained in this exam, and P_{ec} the number of points you may have accumulated along the course (exams and activities). Your course grade is computed as:

   $$\frac{P_{\text{ex}}}{40} \times 7 + P_{\text{ec}}$$

   with the added requirement that you obtain at least 40% of the maximum punctuation in this exam:

   $$\frac{P_{\text{ex}}}{40} \geq 0.40$$

   You may leave after 30 minutes from the start of the exam if you feel you are not going to reach those requirements.
5. The time scheduled for this exam is 2 hours and 40 minutes.

Do not turn the page until advised to do so!
Mark the type of your exam on the orange sheet NOW!!
This exam is of type $\boxed{1}$

Statistics Applied to Economics

Final, May, 31, 2012, Exam version: $\boxed{1}$

Last name: ____________________

First name: ____________________

DNI: ____________________

Group: ____________________

Instructor: ____________________

Section 1. Multiple choice questions

1. Let $F_X(x_i)$ the value of the distribution function evaluated at point x_i of a random variable $X \sim b(p = 0.3, n = 100)$. What would be the preferred approximation of $F_X(x_i)$ among the following?

   (a) $\Phi\left(\frac{x_i - 30}{\sqrt{0.3 \cdot 0.7 \cdot 100}}\right)$

   (b) $\Phi\left(\frac{x_i - 0.5 - 30}{\sqrt{0.3 \cdot 0.7 \cdot 100}}\right)$

   (c) $\Phi\left(\frac{x_i + 0.5 - 30}{\sqrt{0.3 \cdot 0.7 \cdot 100}}\right)$

   (d) $\Phi\left(\frac{x_i + 0.5 - 30}{0.3 \cdot 0.7 \cdot 100}\right)$

   (e) All other answers are false

Start of a question block

All questions up until next horizontal line refer to the situation described next.

Near the end of World War II, V1 bombs began to fall all over London. The British command was uncertain about whether the Germans were aiming at particular targets (in which case impacts would tend to cluster around specific places), or the bombing was just random.

To gain some insight, they divided the area of London in 1000 rectangles of the same surface, and counted how many bombs fell in each rectangle.

2. If all rectangles were equally likely destinations, and one bomb is thrown, what is the probability that it falls in a given rectangle?

   (a) 0.001

   (b) $(0.001)^1(1 - 0.001)^{999}$

   (c) $\binom{1000}{1}(0.001)^1(1 - 0.001)^{999}$

   (d) 0.999

   (e) All other answers are false

3. If all rectangles were equally likely and one thousand bombs are thrown, what is the *exact* distribution of the number of bombs falling in a given rectangle?

   (a) Binomial$(p = 0.001, n = 1000)$

   (b) Exponential with $\lambda = 0.001$

   (c) Exponential with $\lambda = 1$

   (d) $N(m = 1, \sigma^2 = 0.001 \times 0.999)$

   (e) Poisson, $\mathcal{P}(\lambda = 1)$

4. If all rectangles were equally likely, and 1000 bombs were thrown, which do you think would be not the exact but a good approximation to the distribution of the number of bombs falling in a given rectangle?

   (a) Binomial$(p = 0.001, n = 1000)$

   (b) Exponential with $\lambda = 0.001$

   (c) Exponential with $\lambda = 1$

   (d) $N(m = 1, \sigma^2 = 0.001 \times 0.999)$

   (e) Poisson, $\mathcal{P}(\lambda = 1)$

5. Under the assumption of equally likely rectangles, what is the expected number of hits in each rectangle if 1000 bombs are thrown?
   (a) 0.001
   (b) 0.001×0.999
   (c) $0.001^1 \times 0.999^{999}$
   (d) 1000
   (e) 1

6. If 1000 bombs are thrown, what would be the approximate probability that a given rectangle experiences 2 hits?
   (a) 0.16766
   (b) 0.30600
   (c) 0.73576
   (d) 0.18393
   (e) All other answers are false

7. Assume that indeed 1000 bombs were thrown and the 1000 rectangles were classified according to the number of bomb hits as follows:

Number of V1 bomb hits	**Number of rectangles**
0	370
1	365
2	184
3	62
≥ 4	19

   What would you choose to test the hypothesis that all rectangles were equally likely to receive a bomb hit?
   (a) A t-test of equality of means of five different populations.
   (b) A χ^2 goodness-of-fit test.
   (c) A t-test of the hypothesis $p = 0.001$.
   (d) A Snedecor's $\mathcal{F}$ for equality of variances.
   (e) All other answers are false

8. *(Refer to table on the previous question. This question requires some computations, you may want to postpone answering.)*
   Do you find evidence that some rectangles got significantly more (or less) impacts than expected under the hypothesis that they were all equally likely receive a bomb hit?
   (a) Yes, at the $\alpha = 0.05$ but not at the $\alpha = 0.01$ significance level.
   (b) Yes, but only at the $\alpha = 0.10$ but not at the $\alpha = 0.05$ significance level.
   (c) Yes, even at the $\alpha = 0.01$ significance level.
   (d) No, the distribution of impact appears well in agreement to what would be expected by mere chance under the hypothesis.
   (e) All other answers are false

End of a question block

---

9. The random variable X is Poisson-distributed. What is the value of $P(X = 4)$ if we know that $P(X = 1) = P(X = 2)$?
   (a) 0.2707
   (b) 0.1465
   (c) 0.0122
   (d) 0.0902
   (e) All other answers are false

10. Sampling n independent observations from a $N(m, \sigma^2)$, we know that $nS^2/\sigma^2 \sim \chi^2_{n-1}$. This implies that S^2 is:
   (a) A consistent estimator of m.
   (b) An unbiased estimator of σ^2.
   (c) An asymptotically unbiased estimator of σ^2.
   (d) An inconsistent estimator of σ^2.
   (e) All other answers are false

11. With a sample of $n = 5$ independent observations from a $N(m, \sigma^2)$ we compute $s^2 = n^{-1}\sum_{i=1}^{5}(x_i - \overline{x})^2 = 33.2$. An approximate 99% confidence interval for σ^2 is:

    (a) Cannot be computed, as we do not know m.

    (b) $33.2 \pm 1.96\sqrt{33.2/5}$.

    (c) $[28.1, 39.4]$.

    (d) $[11.14, 801.93)$.

    (e) All other answers are false

12. When performing a chi-square goodness-of-fit test on a contingency table to test the null hypothesis of independence, it just makes sense to reject it at the α significance level when the test statistic is:

    (a) In the α right tail of the χ^2 distribution with adequate degrees of freedom.

    (b) In the α left tail of the χ^2 distribution with adequate degrees of freedom.

    (c) In either $\alpha/2$ tail of the χ^2 distribution with adequate degrees of freedom.

    (d) Close to the mode of the χ^2 distribution with adequate degrees of freedom.

    (e) All other answers are false

START OF A QUESTION BLOCK

We are interested in estimating the proportion of adult people who had in their childhood a non-contagious disease which strikes randomly. We conduct a survey in three villages, of respective (adult) populations $N_1 = 1123$, $N_2 = 2340$ and $N_3 = 3445$.

We sample *without replacement* $n_1 = n_2 = n_3 = 100$ adults from each village and come up with estimations $\hat{p}_1$, $\hat{p}_2$ and $\hat{p}_3$ of the respective proportions: $\hat{p}_i$ is simply the number of people who caught the disease in village i divided by n_i.

13. If the true proportion of people having suffered the disease is the same in the three villages, which estimate would be more precise?

    (a) $\hat{p}_1$

    (b) $\hat{p}_2$

    (c) $\hat{p}_3$

    (d) All three would be equally precise, as both the sample size and variance of the population are identical.

    (e) All other answers are false

14. If the true proportion of people having suffered the disease in the three villages is the same and you estimate the three proportions *using sampling with replacement,* which estimate would be more precise?

    (a) $\hat{p}_1$

    (b) $\hat{p}_2$

    (c) $\hat{p}_3$

    (d) All three would be equally precise, as both the sample size and variance of the population are identical.

    (e) All other answers are false

END OF A QUESTION BLOCK

---

15. The capital city of Spain is:

    (a) Madrid

    (b) Paris

    (c) Rome

    (d) Berlin

    (e) All other answers are false

16. Consider a random variable whose characteristic function is $e^{0.4(e^{iu}-1)}$ (or, if you prefer, whose moment generating function is $e^{0.4(e^{u}-1)}$). This tells us that X is distributed as:

    (a) Binary with $p = 0.4$

    (b) Binomial with $p = 0.4$

    (c) Exponential with parameter $\lambda = 0.4$

    (d) Poisson with parameter $\lambda = 0.4$

    (e) All other answers are false

17. Sampling n_1 and n_2 independent observations from respectively two populations $N(m_1, \sigma^2)$ and $N(m_2, \sigma^2)$, the statistic

$$\frac{n_1 S_1^2}{n_2 S_2^2}$$

will be distributed as:

(a) Student's t with $n_1 + n_2 - 1$ degrees of freedom.

(b) χ^2 with $n_1 + n_2 - 1$ degrees of freedom.

(c) Snedecor's $\mathcal{F}$ with n_1 and n_2 degrees of freedom.

(d) $N(0, 1)$

(e) All other answers are false

18. You want to test $H_0 : X \sim N(m_1, \sigma_0)$ against an alternative $H_a : X \sim N(m_2, \sigma_0)$, $m_2 > m_1$. Assume you have used the Neyman-Pearson theorem to obtain two critical regions with $\alpha = 0.01$ and $\alpha = 0.02$ respectively. It is clear that the one with $\alpha = 0.01$ has:

(a) Power not greater than that with $\alpha = 0.02$

(b) Power not smaller than that with $\alpha = 0.02$

(c) Power twice than that with $\alpha = 0.02$

(d) Power half than that with $\alpha = 0.02$

(e) All other answers are false

19. Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of random variables which converges in probability to a random variable X. This implies that:

(a) X_n is equal to X with large probability for sufficiently large n.

(b) X_n is close to X with large probability for sufficiently large n.

(c) The *distribution* of X_n is close to that of X for sufficiently large n; but $|X_n - X|$ needs not be small, no matter how large n.

(d) That X_n has distribution which approaches a normal with mean equal to that of X, by the Central Limit Theorem.

(e) All other answers are false

20. Sampling n independent observations from a $N(m, \sigma^2)$ with σ^2 known, produces a $(1 - \alpha)$ confidence interval such as: $\overline{X} \pm z_{\alpha/2}\sqrt{\sigma^2/n}$, where $z_{\alpha/2}$ is the value leaving a tail of probability $\alpha/2$ in a $N(0, 1)$ distribution. This implies, among other things, that $\overline{X}$ is:

(a) A consistent estimator of m.

(b) A consistent estimator of σ^2.

(c) An asymptotically unbiased estimator of σ^2.

(d) An inconsistent estimator of m.

(e) All other answers are false

21. Sampling n independent observations from a $N(m, \sigma^2)$, the statistic

$$\frac{\overline{X} - m}{\sqrt{s^2}}\sqrt{n-1}$$

will be distributed as:

(a) Student's t with $n - 1$ degrees of freedom.

(b) χ^2 with $n - 1$ degrees of freedom.

(c) Snedecor's $\mathcal{F}$ with n and $n - 1$ degrees of freedom.

(d) $N(0, 1)$

(e) All other answers are false

22. You want to test $H_0 : X \sim \mathcal{P}(\lambda = 1)$ against an alternative $H_a : X \sim \mathcal{P}(\lambda = 4)$. Assume that you have a sample of size n and use as a test statistic $\overline{X}$. The largest the value of $\overline{X} \ldots$

(a) The strongest the evidence against H_0.

(b) The weakest the evidence against H_0.

(c) The most likely is H_0.

(d) The least likely is H_a.

(e) All other answers are false

23. Let $X_1, X_2, ..., X_{12}$ be a sample of independent values from a distribution with density function $f(x, \theta) = \frac{1}{\theta}$, $0 \leq x \leq \theta$. If we consider the estimator $\hat{\theta} = k\frac{X_1+X_2+...+X_{12}}{12}$, the value of k which makes $\hat{\theta}$ unbiased for θ is:

(a) $k = 3$

(b) $k = 4$

(c) $k = 1$

(d) $k = 2$

(e) All other answers are false

24. If $X \sim b(p = 0.8, n = 20)$, what is the probability that it takes values less than 10? ($F_X()$ denotes the distribution function of X).

    (a) $P(Y \geq 10)$ with $Y \sim b(p = 0.2, n = 20)$

    (b) $F_X(10)$

    (c) $1 - F_Y(10)$ with $Y \sim b(p = 0.2, n = 20)$

    (d) $F_Y(11)$ with $Y \sim b(p = 0.2, n = 20)$

    (e) All other answers are false

START OF A QUESTION BLOCK

You are to test the hypothesis $H_0 : X \sim N(0, 1)$ versus $H_a : X \sim N(1, 1)$ using the test statistic $\overline{X}$.

25. The critical region would be of the form:

    (a) $[k, \infty)$

    (b) $(-\infty, k]$

    (c) $[k_1, k_2]$

    (d) $[0.5, k_2]$

    (e) All other answers are false

26. One fast and easy way to test the hypothesis given would be to construct a $(1 - \alpha)$ confidence interval for the mean of the distribution. If that interval covers the value 0 (mean under H_0):

    (a) We would reject H_0 at the α significance level.

    (b) We would *not* reject H_0 at the α significance level.

    (c) The sketched procedure is incorrect, because the alternative is one-sided.

    (d) The sketched procedure is incorrect: what it really does is to test H_a against H_0 at the α significance level.

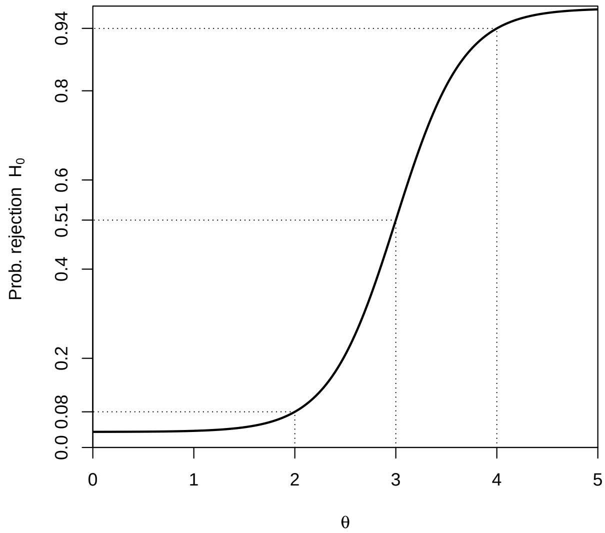
    (e) All other answers are false

END OF A QUESTION BLOCK

START OF A QUESTION BLOCK

All questions up until the next horizontal line refer to information next.

Look at the following graph. For a certain hypothesis testing procedure, it gives the probability of rejecting the hypothesis $H_0 : \theta = 2$ as a function of the true value of θ (in the horizontal axis).



27. From the graph above, it is clear the significance level α of the test is:

    (a) 0.08

    (b) 0.05

    (c) 0.92

    (d) 0.96

    (e) All other answers are false

28. From the graph above, it is clear the power is greater as θ:

    (a) Increases

    (b) Decreases

    (c) The power is not affected by θ

    (d) The power is 0.08

    (e) All other answers are false

29. From the graph above, it is clear that as θ increases, β:

    (a) Increases

    (b) Decreases

    (c) Is not affected by θ

    (d) Is less than 0.08

    (e) All other answers are false

30. It is clear that the test was designed to have power against alternatives:

    (a) $H_a : \theta > 2$

    (b) $H_a : \theta < 2$

    (c) $H_a : \theta \neq 2$

    (d) $H_a : \theta = 0$

    (e) All other answers are false

End of a question block

Section 2. Problems

Answer each problem in a different sheet; they will be collected in order, so finish first problem 1, then proceed to problem 2.

1. *(5 points)* A member country of the European Union is considering to opt out of the euro zone and revert to its previous currency. The population seems to be about evenly divided in favor or against of this measure. The Government decides to conduct a sample survey, to estimate the proportion p of people willing to say "yes" to leaving the euro zone.

    (a) If the Government wants to estimate p so that an approximate 95% confidence interval has half-width of 0.01, what is the required sample size?

    (b) The Government thinks the attitude of people is different in large cities and in the rest of the country, and would like to estimate p_1 and p_2 respectively for each part. What sample size is needed for each part (both parts are large, of millions of people)? How does this compare with the sample size for the whole of the country obtained in 1a? Explain.

    (c) According to the last census, N_1 people live in large cities and N_2 in the rest of the country. How would you use your estimates of p_1 and p_2 to obtain an estimate of p for the whole country? How would the precision of that estimate compare with the precision of an estimate of p using a single sample? *(Assume that, indeed, attitudes are very different in large cities and the rest of the country, so estimates of the respective proportions p_1 and p_2 of people are quite different.)* Explain.

    (d) Budget and urgency considerations force to use a sample size of $n = 1000$ respondents. Of those, 630 answer "yes". Give a 95% approximate confidence interval for the proportion of people willing to answer yes in the population.

2. *(5 points, 20 minutes)* Let X_1, X_2, X_3, X_4 be a sample of independent values of the random variable X with exponential distribution

$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{\theta} e^{-\frac{1}{\theta}x} & x \geq 0 \end{cases}$$

Consider the following two estimators of θ:

$$\begin{aligned} \hat{\theta}_1 &= \frac{1}{6}(X_1 + X_2) + \frac{1}{3}(X_3 + X_4) \\ \hat{\theta}_2 &= \frac{X_1 + X_2 + X_3 + X_4}{4} \end{aligned}$$

and answer the following questions:

    (a) What is the bias of $\hat{\theta}_1$ and $\hat{\theta}_2$?

    (b) Which among both has smaller MSE (mean square error, ECM in Spanish)? Reason your answer.

    (c) Indicate, giving the reason, whether $\hat{\theta}_1$ y $\hat{\theta}_2$ are consistent.

eman ta zabal zazu

Universidad del País Vasco Euskal Herriko Unibertsitatea

DIRECTIONS

1. Correctly answered questions give one point. There is only one correct answer to each question. Questions not correctly answered carry a penalty of -0.20 points, so it is better to leave a question unanswered rather than giving a wrong answer.

2. Our goal is to gauge your understanding and command of concepts learned during the course, not your visual sharpness. It is a fact, though, that in a multiple choice exam great attention has to be paid to the details. It is quite common that knowledgeable students waste their chances of a good grade because they do not pay sufficient attention to the precise wording of questions.

Please, read carefully before you answer!

3. It will probably help you to discard first answers that are clearly inadequate.

4. Let P_{ex} be the number of points obtained in this exam, and P_{ec} the number of points you may have accumulated along the course (exams and activities). Your course grade is computed as:

$$\frac{P_{\text{ex}}}{40} \times 7 + P_{\text{ec}}$$

with the added requirement that you obtain at least 40% of the maximum punctuation in this exam:

$$\frac{P_{\text{ex}}}{40} \geq 0.40$$

You may leave after 30 minutes from the start of the exam if you feel you are not going to reach those requirements.

5. The time scheduled for this exam is 2 hours and 40 minutes.

Do not turn the page until advised to do so!
Mark the type of your exam on the orange sheet NOW!!
This exam is of type $\boxed{1}$

Answers for the exam of type $\boxed{1}$

Section 1. Multiple choice questions

1. Let $F_X(x_i)$ the value of the distribution function evaluated at point x_i of a random variable $X \sim b(p = 0.3, n = 100)$. What would be the preferred approximation of $F_X(x_i)$ among the following?

   (a) $\Phi\left(\frac{x_i - 30}{\sqrt{0.3 \cdot 0.7 \cdot 100}}\right)$

   (b) $\Phi\left(\frac{x_i - 0.5 - 30}{\sqrt{0.3 \cdot 0.7 \cdot 100}}\right)$

   (c) $\Phi\left(\frac{x_i + 0.5 - 30}{\sqrt{0.3 \cdot 0.7 \cdot 100}}\right)$

   (d) $\Phi\left(\frac{x_i + 0.5 - 30}{0.3 \cdot 0.7 \cdot 100}\right)$

   (e) All other answers are false

Start of a question block

All questions up until next horizontal line refer to the situation described next.

Near the end of World War II, V1 bombs began to fall all over London. The British command was uncertain about whether the Germans were aiming at particular targets (in which case impacts would tend to cluster around specific places), or the bombing was just random.

To gain some insight, they divided the area of London in 1000 rectangles of the same surface, and counted how many bombs fell in each rectangle.

2. If all rectangles were equally likely destinations, and one bomb is thrown, what is the probability that it falls in a given rectangle?

   (a) 0.001

   (b) $(0.001)^1(1 - 0.001)^{999}$

   (c) $\binom{1000}{1}(0.001)^1(1 - 0.001)^{999}$

   (d) 0.999

   (e) All other answers are false

3. If all rectangles were equally likely and one thousand bombs are thrown, what is the *exact* distribution of the number of bombs falling in a given rectangle?

   (a) **Binomial**$(p = 0.001, n = 1000)$

   (b) Exponential with $\lambda = 0.001$

   (c) Exponential with $\lambda = 1$

   (d) $N(m = 1, \sigma^2 = 0.001 \times 0.999)$

   (e) Poisson, $\mathcal{P}(\lambda = 1)$

4. If all rectangles were equally likely, and 1000 bombs were thrown, which do you think would be not the exact but a good approximation to the distribution of the number of bombs falling in a given rectangle?

   (a) Binomial$(p = 0.001, n = 1000)$

   (b) Exponential with $\lambda = 0.001$

   (c) Exponential with $\lambda = 1$

   (d) $N(m = 1, \sigma^2 = 0.001 \times 0.999)$

   (e) **Poisson,** $\mathcal{P}(\lambda = 1)$

5. Under the assumption of equally likely rectangles, what is the expected number of hits in each rectangle if 1000 bombs are thrown?

   (a) 0.001

   (b) 0.001×0.999

   (c) $0.001^1 \times 0.999^{999}$

   (d) 1000

   (e) **1**

6. If 1000 bombs are thrown, what would be the approximate probability that a given rectangle experiences 2 hits?

   (a) 0.16766

   (b) 0.30600

   (c) 0.73576

   (d) **0.18393**

   (e) All other answers are false

7. Assume that indeed 1000 bombs were thrown and the 1000 rectangles were classified according to the number of bomb hits as follows:

Number of V1 bomb hits	**Number of rectangles**
0	370
1	365
2	184
3	62
≥ 4	19

What would you choose to test the hypothesis that all rectangles were equally likely to receive a bomb hit?

(a) A t-test of equality of means of five different populations.

(b) A χ^2 goodness-of-fit test.

(c) A t-test of the hypothesis $p = 0.001$.

(d) A Snedecor's $\mathcal{F}$ for equality of variances.

(e) All other answers are false

8. *(Refer to table on the previous question. This question requires some computations, you may want to postpone answering.)*
Do you find evidence that some rectangles got significantly more (or less) impacts than expected under the hypothesis that they were all equally likely receive a bomb hit?

(a) Yes, at the $\alpha = 0.05$ but not at the $\alpha = 0.01$ significance level.

(b) Yes, but only at the $\alpha = 0.10$ but not at the $\alpha = 0.05$ significance level.

(c) Yes, even at the $\alpha = 0.01$ significance level.

(d) No, the distribution of impact appears well in agreement to what would be expected by mere chance under the hypothesis.

(e) All other answers are false

End of a question block

9. The random variable X is Poisson-distributed. What is the value of $P(X = 4)$ if we know that $P(X = 1) = P(X = 2)$?

(a) 0.2707

(b) 0.1465

(c) 0.0122

(d) 0.0902

(e) All other answers are false

10. Sampling n independent observations from a $N(m, \sigma^2)$, we know that $nS^2/\sigma^2 \sim \chi^2_{n-1}$. This implies that S^2 is:

(a) A consistent estimator of m.

(b) An unbiased estimator of σ^2.

(c) An asymptotically unbiased estimator of σ^2.

(d) An inconsistent estimator of σ^2.

(e) All other answers are false

11. With a sample of $n = 5$ independent observations from a $N(m, \sigma^2)$ we compute $s^2 = n^{-1}\sum_{i=1}^{5}(x_i - \overline{x})^2 = 33.2$. An approximate 99% confidence interval for σ^2 is:

(a) Cannot be computed, as we do not know m.

(b) $33.2 \pm 1.96\sqrt{33.2/5}$.

(c) $[28.1, 39.4]$.

(d) $[11.14, 801.93)$.

(e) All other answers are false

12. When performing a chi-square goodness-of-fit test on a contingency table to test the null hypothesis of independence, it just makes sense to reject it at the α significance level when the test statistic is:

(a) In the α right tail of the χ^2 distribution with adequate degrees of freedom.

(b) In the α left tail of the χ^2 distribution with adequate degrees of freedom.

(c) In either $\alpha/2$ tail of the χ^2 distribution with adequate degrees of freedom.

(d) Close to the mode of the χ^2 distribution with adequate degrees of freedom.

(e) All other answers are false

Start of a question block

We are interested in estimating the proportion of adult people who had in their childhood a non-contagious disease which strikes randomly. We conduct a survey in three villages, of respective (adult) populations $N_1 = 1123$, $N_2 = 2340$ and $N_3 = 3445$.

We sample *without replacement* $n_1 = n_2 = n_3 = 100$ adults from each village and come up with estimations $\hat{p}_1$, $\hat{p}_2$ and $\hat{p}_3$ of the respective proportions: $\hat{p}_i$ is simply the number of people who caught the disease in village i divided by n_i.

13. If the true proportion of people having suffered the disease is the same in the three villages, which estimate would be more precise?

    (a) $\hat{p}_1$

    (b) $\hat{p}_2$

    (c) $\hat{p}_3$

    (d) All three would be equally precise, as both the sample size and variance of the population are identical.

    (e) All other answers are false

14. If the true proportion of people having suffered the disease in the three villages is the same and you estimate the three proportions *using sampling with replacement,* which estimate would be more precise?

    (a) $\hat{p}_1$

    (b) $\hat{p}_2$

    (c) $\hat{p}_3$

    (d) All three would be equally precise, as both the sample size and variance of the population are identical.

    (e) All other answers are false

End of a question block

---

15. The capital city of Spain is:

    (a) Madrid

    (b) Paris

    (c) Rome

    (d) Berlin

    (e) All other answers are false

16. Consider a random variable whose characteristic function is $e^{0.4(e^{iu}-1)}$ (or, if you prefer, whose moment generating function is $e^{0.4(e^{u}-1)}$). This tells us that X is distributed as:

    (a) Binary with $p = 0.4$

    (b) Binomial with $p = 0.4$

    (c) Exponential with parameter $\lambda = 0.4$

    (d) Poisson with parameter $\lambda = 0.4$

    (e) All other answers are false

17. Sampling n_1 and n_2 independent observations from respectively two populations $N(m_1, \sigma^2)$ and $N(m_2, \sigma^2)$, the statistic

$$\frac{n_1 S_1^2}{n_2 S_2^2}$$

will be distributed as:

    (a) Student's t with $n_1 + n_2 - 1$ degrees of freedom.

    (b) χ^2 with $n_1 + n_2 - 1$ degrees of freedom.

    (c) Snedecor's $\mathcal{F}$ with n_1 and n_2 degrees of freedom.

    (d) $N(0, 1)$

    (e) All other answers are false

18. You want to test $H_0 : X \sim N(m_1, \sigma_0)$ against an alternative $H_a : X \sim N(m_2, \sigma_0)$, $m_2 > m_1$. Assume you have used the Neyman-Pearson theorem to obtain two critical regions with $\alpha = 0.01$ and $\alpha = 0.02$ respectively. It is clear that the one with $\alpha = 0.01$ has:

    (a) Power not greater than that with $\alpha = 0.02$

    (b) Power not smaller than that with $\alpha = 0.02$

    (c) Power twice than that with $\alpha = 0.02$

    (d) Power half than that with $\alpha = 0.02$

    (e) All other answers are false

19. Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of random variables which converges in probability to a random variable X. This implies that:

    (a) X_n is equal to X with large probability for sufficiently large n.

    (b) X_n is close to X with large probability for sufficiently large n.

    (c) The *distribution* of X_n is close to that of X for sufficiently large n; but $|X_n - X|$ needs not be small, no matter how large n.

    (d) That X_n has distribution which approaches a normal with mean equal to that of X, by the Central Limit Theorem.

    (e) All other answers are false

20. Sampling n independent observations from a $N(m, \sigma^2)$ with σ^2 known, produces a $(1-\alpha)$ confidence interval such as: $\overline{X} \pm z_{\alpha/2}\sqrt{\sigma^2/n}$, where $z_{\alpha/2}$ is the value leaving a tail of probability $\alpha/2$ in a $N(0,1)$ distribution. This implies, among other things, that $\overline{X}$ is:

    (a) A consistent estimator of m.

    (b) A consistent estimator of σ^2.

    (c) An asymptotically unbiased estimator of σ^2.

    (d) An inconsistent estimator of m.

    (e) All other answers are false

21. Sampling n independent observations from a $N(m, \sigma^2)$, the statistic

$$\frac{\overline{X} - m}{\sqrt{s^2}}\sqrt{n-1}$$

will be distributed as:

    (a) Student's t with $n-1$ degrees of freedom.

    (b) χ^2 with $n-1$ degrees of freedom.

    (c) Snedecor's $\mathcal{F}$ with n and $n-1$ degrees of freedom.

    (d) $N(0,1)$

    (e) All other answers are false

22. You want to test $H_0 : X \sim \mathcal{P}(\lambda = 1)$ against an alternative $H_a : X \sim \mathcal{P}(\lambda = 4)$. Assume that you have a sample of size n and use as a test statistic $\overline{X}$. The largest the value of $\overline{X}\ldots$

    (a) The strongest the evidence against H_0.

    (b) The weakest the evidence against H_0.

    (c) The most likely is H_0.

    (d) The least likely is H_a.

    (e) All other answers are false

23. Let $X_1, X_2, ..., X_{12}$ be a sample of independent values from a distribution with density function $f(x, \theta) = \frac{1}{\theta}$, $0 \le x \le \theta$. If we consider the estimator $\hat{\theta} = k\frac{X_1+X_2+...+X_{12}}{12}$, the value of k which makes $\hat{\theta}$ unbiased for θ is:

    (a) $k = 3$

    (b) $k = 4$

    (c) $k = 1$

    (d) $k = 2$

    (e) All other answers are false

24. If $X \sim b(p = 0.8, n = 20)$, what is the probability that it takes values less than 10? ($F_X()$ denotes the distribution function of X).

    (a) $P(Y \ge 10)$ with $Y \sim b(p = 0.2, n = 20)$

    (b) $F_X(10)$

    (c) $1 - F_Y(10)$ with $Y \sim b(p = 0.2, n = 20)$

    (d) $F_Y(11)$ with $Y \sim b(p = 0.2, n = 20)$

    (e) All other answers are false

START OF A QUESTION BLOCK

You are to test the hypothesis $H_0 : X \sim N(0, 1)$ versus $H_a : X \sim N(1, 1)$ using the test statistic $\overline{X}$.

25. The critical region would be of the form:

    (a) $[k, \infty)$

    (b) $(-\infty, k]$

    (c) $[k_1, k_2]$

    (d) $[0.5, k_2]$

    (e) All other answers are false

26. One fast and easy way to test the hypothesis given would be to construct a $(1-\alpha)$ confidence interval for the mean of the distribution. If that interval covers the value 0 (mean under H_0):

    (a) We would reject H_0 at the α significance level.

    (b) We would *not* reject H_0 at the α significance level.

    (c) The sketched procedure is incorrect, because the alternative is one-sided.

    (d) The sketched procedure is incorrect: what it really does is to test H_a against H_0 at the α significance level.

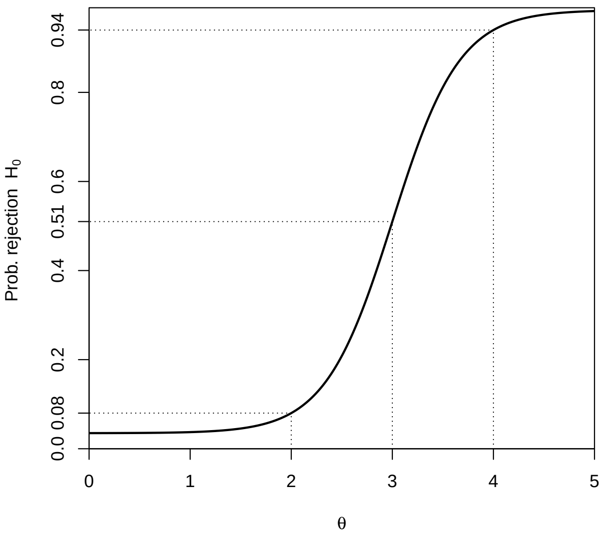
    (e) All other answers are false

End of a question block

Start of a question block

All questions up until the next horizontal line refer to information next.

Look at the following graph. For a certain hypothesis testing procedure, it gives the probability of rejecting the hypothesis $H_0 : \theta = 2$ as a function of the true value of θ (in the horizontal axis).



27. From the graph above, it is clear the significance level α of the test is:

    (a) 0.08

    (b) 0.05

    (c) 0.92

    (d) 0.96

    (e) All other answers are false

28. From the graph above, it is clear the power is greater as θ:

    (a) Increases

    (b) Decreases

    (c) The power is not affected by θ

    (d) The power is 0.08

    (e) All other answers are false

29. From the graph above, it is clear that as θ increases, β:

    (a) Increases

    (b) Decreases

    (c) Is not affected by θ

    (d) Is less than 0.08

    (e) All other answers are false

30. It is clear that the test was designed to have power against alternatives:

    (a) $H_a : \theta > 2$

    (b) $H_a : \theta < 2$

    (c) $H_a : \theta \neq 2$

    (d) $H_a : \theta = 0$

    (e) All other answers are false

End of a question block

Section 2. Problems

Answer each problem in a different sheet; they will be collected in order, so finish first problem 1, then proceed to problem 2.

1. *(5 points)* A member country of the European Union is considering to opt out of the euro zone and revert to its previous currency. The population seems to be about evenly divided in favor or against of this measure. The Government decides to conduct a sample survey, to estimate the proportion p of people willing to say "yes" to leaving the euro zone.

(a) If the Government wants to estimate p so that an approximate 95% confidence interval has half-width of 0.01, what is the required sample size?

Answer: The half-width of the confidence interval assuming $p = q = 0.5$ is $z_{\alpha/2}\sqrt{0.25/n}$. If $1-\alpha = 0.95$, $z_{\alpha/2} \approx 1.96$ and we need to solve for n the equation:

$$z_{\alpha/2}\sqrt{0.25/n} = 0.01;$$

this leads to $n \approx 9604$.

(b) The Government thinks the attitude of people is different in large cities and in the rest of the country, and would like to estimate p_1 and p_2 respectively for each part. What sample size is needed for each part (both parts are large, of millions of people)? How does this compare with the sample size for the whole of the country obtained in 1a? Explain.

Answer: Both subpopulations are very large ("of millions of people"). Therefore, the finite population correction will be very small, and to a very good approximation we can say that we need *the same* sample size for each; that is 9604 observations for the large cities and 9604 for the rest of the country. Therefore, to estimate the proportion in both subpopulations will cost about twice as much as estimating with the same precision in the whole population.

(c) According to the last census, N_1 people live in large cities and N_2 in the rest of the country. How would you use your estimates of p_1 and p_2 to obtain an estimate of p for the whole country? How would the precision of that estimate compare with the precision of an estimate of p using a single sample? *(Assume that, indeed, attitudes are very different in large cities and the rest of the country, so estimates of the respective proportions p_1 and p_2 of people are quite different.)* Explain.

Answer: A sensible estimate of p would be

$$\hat{p} = \frac{N_1}{N_1+N_2}\hat{p}_1 + \frac{N_2}{N_1+N_2}\hat{p}_2.$$

This would be unbiased if both p_1 and p_2 are (trivial to show) and would use our knowledge of the relative sizes of the two subpopulations. This can be seen as a case of stratified sampling, and the situation described (very different strata) is the one where we would stand to gain the most. You might have invoked here the formula giving the variance gain of stratified vs. simple random sampling.

(d) Budget and urgency considerations force to use a sample size of $n = 1000$ respondents. Of those, 630 answer "yes". Give a 95% approximate confidence interval for the proportion of people willing to answer yes in the population.

Answer: That would be easily built as $0.63 \pm 1.96\sqrt{\frac{0.63\times0.37}{1000}}$; an overly conservative confidence interval would be $0.63 \pm 1.96\sqrt{\frac{0.50\times0.50}{1000}}$.

2. *(5 points, 20 minutes)* Let X_1, X_2, X_3, X_4 be a sample of independent values of the random variable X with exponential distribution

$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{\theta}e^{-\frac{1}{\theta}x} & x \geq 0 \end{cases}$$

Consider the following two estimators of θ:

$$\begin{aligned} \hat{\theta}_1 &= \frac{1}{6}(X_1+X_2) + \frac{1}{3}(X_3+X_4) \\ \hat{\theta}_2 &= \frac{X_1+X_2+X_3+X_4}{4} \end{aligned}$$

and answer the following questions:

(a) What is the bias of $\hat{\theta}_1$ and $\hat{\theta}_2$?

Answer: The bias is zero in both cases, as shown by taking mean value of $\hat{\theta}_1$ and $\hat{\theta}_2$.

(b) Which among both has smaller MSE (mean square error, ECM in Spanish)? Reason your answer.

Answer: Since the bias is zero in both cases, MSE is coincident with variance, so the question boils down to finding the estimator with the smallest variance.

That can be obtained easily (e.g., $\mathrm{Var}(\hat{\theta}_2) = \frac{1}{16}\sum_{i=1}^{4}\mathrm{Var}(X_i]$. The variance of X_i happens to be θ^2, but knowledge of this is not required; replacing by σ^2 is enough to perform the comparison with the variance of $\hat{\theta}_1$. The variance of $\hat{\theta}_2$ happens to be smaller.

Another valid answer that would have saved even that small computation would be: "$\hat{\theta}_2$ is the MLE of θ, that we have seen in class attains the Cramer-Rao lower bound for unbiased estimators; hence it has the smallest variance." (I would have been more than happy with that answer although it only rules out the possibility of $\hat{\theta}_1$ having smaller variance than $\hat{\theta}_2$; you should still consider the possibility $\mathrm{Var}(\hat{\theta}_2) = \mathrm{Var}(\hat{\theta}_1)$.)

(c) Indicate, giving the reason, whether $\hat{\theta}_1$ y $\hat{\theta}_2$ are consistent.

Answer: This is a question that has caused some spur, and in retrospect not one that I am fully happy with. I have accepted just about any answer which shows you know what consistency is. Examples of valid answers would be:

i. "Consistency is an asymptotic notion, that describes what happens to a sequence of estimators when the sample size tends to infinity. As both estimators given here are based on a finite sample size, this notion is not applicable."

ii. "They are not consistent; however, estimators of the same *form* with increasing sample size might be. For instance, the like of $\hat{\theta}_2$ defined as

$$\hat{\theta}_{2,n} = \frac{X_1 + \ldots + X_n}{n}$$

can be immediately shown to be consistent (unbiased with variance $\theta^2/n \to 0$).