DSCC/CSC 462:
Computational Introduction to Statistics

Midterm Review Part 2

Thursday, November 8, 2022

Important; please read before starting:

- You must **show your work for all questions** in order to receive full credit. Showing your work will also let you receive partial credit for problems.
- For all hypothesis tests and confidence intervals, **you must demonstrate that you know how to solve the question without using any `.test()` functions in `R`**. You may use these functions to check your answers, but writing them down will not count as "showing work."

1. Let A be a random variable with a PDF of $f_A(x) = cx^2$ for $x \in [-1, 1]$.

   (a) Find c.

$$\int_{-1}^{1} cx^2\,dx = 1 \rightarrow \frac{cx^3}{3}\Big|_{-1}^{1} = 1 \rightarrow \frac{2c}{3} = 1 \rightarrow \boxed{c = 3/2}$$

   (b) Find $\mathrm{E}(A)$ and $\mathrm{Var}(A)$.

$\boxed{E(A) = 0}$ by symmetry. Also $\int_{-1}^{1} x \cdot \frac{3}{2}x^2\,dx = \frac{3}{2}\cdot\frac{x^4}{4}\Big|_{-1}^{1} = 0$.

$$Var(A) = E(A^2) - \overbrace{E(A)}^{0}{}^2, \quad E(A^2) = \int_{-1}^{1} x^2 \cdot \frac{3}{2}x^2\,dx = \frac{3}{2}\cdot\frac{x^5}{5}\Big|_{-1}^{1} = \frac{3}{5}$$

$$\rightarrow \boxed{Var(A) = 3/5}$$

   (c) Let B be an independent random variable that follows a Uniform distribution over the domain $[0, 4]$. Now, define $X = 2A + B$. What are $\mathrm{E}(X)$ and $\mathrm{Var}(X)$?

$E(B) = 2$, $Var(B) = \frac{16}{12}$.

$$E(X) = 2E(A) + E(B) = 2\times 0 + 2 = \boxed{2} = \mu_X$$

$$Var(X) = 2^2\,Var(A) + Var(B) = 4\cdot\frac{3}{5} + \frac{16}{12} = \boxed{3.73} = \sigma_X^2$$

   (d) If we were to draw $n = 49$ samples according to X's distribution, what is the distribution of the sample mean?

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \rightarrow \boxed{\bar{X} \sim N(2, 0.276)}$$

(e) We draw a sample of size $n = 49$ according to an unknown distribution. The sample mean is $\overline{x} = 0.5$ and the sample variance is $s^2 = 4$. What is a two-sided 95% confidence interval for the true mean μ?

CI: $\overline{x} \pm t_{0.975,48} \cdot \frac{s}{\sqrt{n}}$ → 0.5 ± 0.574 → $(-0.074, 1.074)$

(f) At $\alpha = 0.1$, is the true mean greater than 0? State which test you should use, your hypotheses, test statistic, p-value, and conclusion.

One-sided t test. $H_0: \mu \leq 0$ vs. $H_1: \mu > 0$.

$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{0.5 - 0}{2/\sqrt{49}} = 1.75$

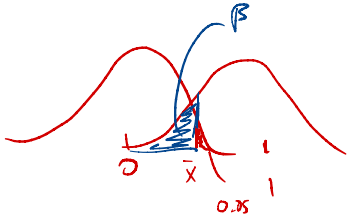
$p = 1 - pt(1.75, df = 48) = 0.043$

$p < \alpha$ → Reject H_0.

Conclude true mean is greater than 0.

(g) What is the power of the test in part (f) if we assume an alternate hypothesis of $\mu_1 = 1$?

Find $\overline{x}$ using H_0 and α: $\frac{\overline{x} - 0}{2/\sqrt{49}} = qt(0.9, df = 48) \rightarrow \overline{x} = 0.371$



Find β using $\overline{x}$ and H_1: $\beta = pt\left(\frac{\overline{x} - \mu_1}{s/\sqrt{n}}, df = 48\right)$

$= pt(-2.2, df = 48) = 0.016$ → Power $= 1 - \beta =$ 0.984

(h) How would increasing the sample size change your answer to part (g)? Explain.

Power would increase. Larger n means narrower distributions and smaller $\overline{x}$.

2. Probability and combinatorics.

(a) If $\Pr(D) = 0.5$ and $\Pr(C \cap D) = 0.3$, what is $\Pr(C|D)$?

$$\Pr(C|D) = \frac{\Pr(C \cap D)}{\Pr(D)} = \frac{0.3}{0.5} = 0.6$$

(b) You are given three random variables A, B, and C. If you know that A and B are (pairwise) independent, A and C are (pairwise) independent, and B and C are (pairwise) independent, are A, B, and C all (mutually) independent? Explain.

Not necessarily. See example in class.

(c) If D and E are mutually exclusive events, are they independent? Explain.

No. $\Pr(D \cap E) = 0 \neq \Pr(D) \cdot \Pr(E)$

(d) Suppose candies can either be chocolate or non-chocolate. Alice finds 60% of chocolate candy tasty and 40% of non-chocolate candy tasty. If 30% of all candies are chocolate, what is the probability that a candy is chocolate given that Alice finds it tasty?

$$\Pr(C|T) = \frac{\Pr(T|C)\Pr(C)}{\Pr(T)} = \frac{0.6 \times 0.3}{0.6 \times 0.3 + 0.4 \times 0.7} = 0.39$$

$$\Pr(T|C)\Pr(C) + \Pr(T|C^c)\Pr(C^c)$$

(e) (2 points) A and B are independent events. If $\Pr(A) = 0.4$, $\Pr(B) = 0.5$, what is $\Pr(A \cup B)$?

$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = 0.4 + 0.5 - 0.2 = \boxed{0.7}$

(f) (6 points) There are twenty people in a village. Five of these people own cars. Five randomly picked people will win a bicycle. Due to fairness constraints, the leader must ensure that **exactly** four of the five winners must not own a car. In how many ways can this choice be made?

$\binom{15}{4} \cdot \binom{5}{1} = \boxed{6825}$

(g) (4 points) Given the same setup as above, what if the leader must ensure that **at least** four of the five winners must not own a car? In how many ways can this choice be made now?

$\underbrace{\binom{15}{4}\binom{5}{1}}_{\text{4 w/out cars}} + \underbrace{\binom{15}{5}}_{\text{5 w/out cars}} = \boxed{9828}$

3. Starbucks is very interested in finding out what fraction of people have each season as their favorite so they can release new specialty drinks.

   (a) The CEO of Starbucks believes that 40% of people like summer the most, 30% like fall the most, 10% like winter the most, and 20% like spring the most. The marketing department takes a sample of $n = 300$ people and records their favorite season in the table below. At the $\alpha = 0.01$ significance level, use an appropriate statistical test to determine whether the CEO's belief is correct.

Summer	Fall	Winter	Spring
110	110	20	60

   i. (3 points) State which test you should use, along with your null and alternative hypotheses.

   Test: GoF test

   H_0 : $p_1 = 0.4$, $p_2 = 0.3$, $p_3 = 0.1$, $p_4 = 0.2$

   H_1 : At least one is different.

   ii. (3 points) Calculate the expected number of people who have each season as their favorite:

Summer	Fall	Winter	Spring
120	90	30	60

   iii. (6 points) Calculate the appropriate test statistic. Show your work (i.e., demonstrate that you know the formula for calculating the test statistic).

$$X^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{10^2}{120} + \frac{20^2}{90} + \frac{10^2}{30} + \frac{0^2}{60} = \boxed{8.61}$$

   iv. (3 points) Calculate the p-value. Show your work and write the R code for calculating the p-value.

$df = 4 - 1 = 3$. $p = Pr(\chi^2_3 > 8.61) =$ `1 - pchisq(8.61, df=3)` $= \boxed{0.035}$

   v. (2 points) What conclusion do you draw from this test?

$p < \alpha \rightarrow$ Reject H_0. At least one proportion is different.

4. (20 points) A company wants to understand if a person's favorite season is associated with whether or not they are pet owners. The marketing department takes a sample of $n = 250$ people and records both their favorite season and whether or not they are pet owners. The results are in the table below. At the $\alpha = 0.01$ significance level, perform an appropriate statistical test to determine whether or not a person's favorite season is associated with pet ownership.

	Summer	Fall	Winter	Spring	Total
Pet	60	25	15	50	150
No Pet	30	20	15	35	100
Total	90	45	30	85	250

(a) (3 points) State which test you should use, along with your null and alternative hypotheses.

Test: χ^2 Test of Independence

H_0 : Pet ownership and favorite season are independent

H_1 : " " " associated

(b) (3 points) Fill in the following *expected* contingency table:

	Summer	Fall	Winter	Spring	Total
Pet	54	27	18	51	150
No Pet	36	18	12	34	100
Total	90	45	30	85	250

(c) (8 points) Calculate the appropriate test statistic. Show your work (i.e., demonstrate that you know the formula for calculating the test statistic).

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{6^2}{54} + \frac{2^2}{27} + \frac{3^2}{18} + \frac{1^2}{51} + \frac{6^2}{36} + \frac{2^2}{18} + \frac{3^2}{12} + \frac{1^2}{34} = 3.336$$

(d) (4 points) Calculate the p-value. Show your work and write the R code for calculating the p-value.

$df = (2-1)(4-1) = 3$. $p = Pr(\chi^2_3 > \chi^2) = 1 - pchisq(3.336, df = 3) = 0.343$

(e) (2 points) What conclusion do you draw from this test?

$p > \alpha \rightarrow$ Fail to reject H_0. Insufficient evidence of association.

t distribution, df : 40 − 1 = 39.

5. (20 points) Suppose you have a sample of $n = 40$ heights of Canadians. The sample mean is $\overline{x} = 66$ inches and the sample standard deviation is $s = 9$ inches^2.

   (a) (5 points) What is a 95% two-sided confidence interval for the average height of Canadians?

CI : $\overline{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$

$= 66 \pm qt(0.975, df = 39) \times \frac{9}{\sqrt{40}} =$ (63.122, 68.878) inches

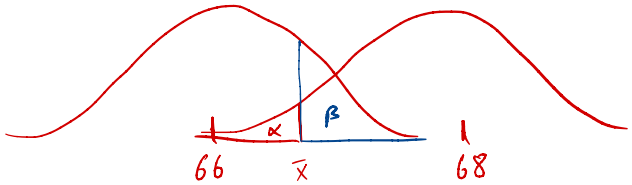
   (b) (5 points) Is the average height of Canadians less than 68 inches? Test at $\alpha = 0.05$. State hypotheses, calculate test statistic, calculate p-value, and state conclusions.

$H_0 : \mu \geq 68$ vs. $H_1 : \mu < 68$.

$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{66 - 68}{9/\sqrt{40}} = -1.405$

$p = Pr(T < t) : pt(-1.405, df : 39) = 0.084$

$p > \alpha : 0.05$. Fail to reject H_0. Insufficient evidence to conclude that the average height of Canadians is less than 68 inches.



(c) (5 points) What is the power of the test in part (b) given alternative hypothesis $H_1 : \mu_1 = 66$?

cutoff $\bar{x}$: $\frac{\bar{x} - 68}{9/\sqrt{40}} = qt(0.05, 39) = 65.602$

$\beta = Pr\left(T > \frac{\bar{x} - \mu_1}{s/\sqrt{n}}\right) = Pr\left(T > \frac{65.602 - 66}{9/\sqrt{40}}\right) = 1 - pt(-0.279, 39)$

$= 1 - 0.391 = 0.609$

Power $= 1 - \beta =$ **0.391** very low!

(d) (5 points) What is the sample size needed in order to achieve $\alpha = 0.05$ and $\beta = 0.05$ given alternative hypothesis $H_1 : \mu_1 = 66$?

One-sided t test.

$\bar{x} = qt_{39}(0.05) \cdot \frac{9}{\sqrt{n}} + 68$

$\beta = 0.05 = Pr\left(T > \frac{\bar{x} - 66}{s/\sqrt{n}}\right)$

$qt_{39}(0.95) = \frac{qt_{39}(0.05) \cdot \frac{9}{\sqrt{n}} + 68 - 66}{9/\sqrt{n}}$

$\frac{9}{\sqrt{n}}\left(-qt_{39}(0.05) + qt_{39}(0.95)\right) = 68 - 66$

$n = \left\lceil \left(\frac{9\left(qt_{39}(0.95) - qt_{39}(0.05)\right)}{68 - 66} \right)^2 \right\rceil$

$n = \lceil 229.943 \rceil =$ **230**

6. (30 points) Suppose you are interested in the proportion of people who like going to the beach. In a sample of $n = 70$ people, 50 of them like going to the beach.

   (a) (10 points) Construct a two-sided 90% (Wald) confidence interval for the true proportion of people who like going to the beach.

↳ $\alpha = 0.1$

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{where} \quad \hat{p} = \frac{50}{70} = 0.714$$

$$\rightarrow \quad 0.714 \pm 1.645\sqrt{\frac{(0.714)(0.286)}{70}} = \boxed{(0.625,\ 0.803)}$$

   (b) (10 points) Is the proportion of people who like going to the beach different from 50%? Use a Normal approximation if applicable. State hypotheses, calculate test statistic, calculate p-value, and state conclusions. Test at $\alpha = 0.05$.

$H_0: p = 0.5$ vs. $H_1: p \neq 0.5$

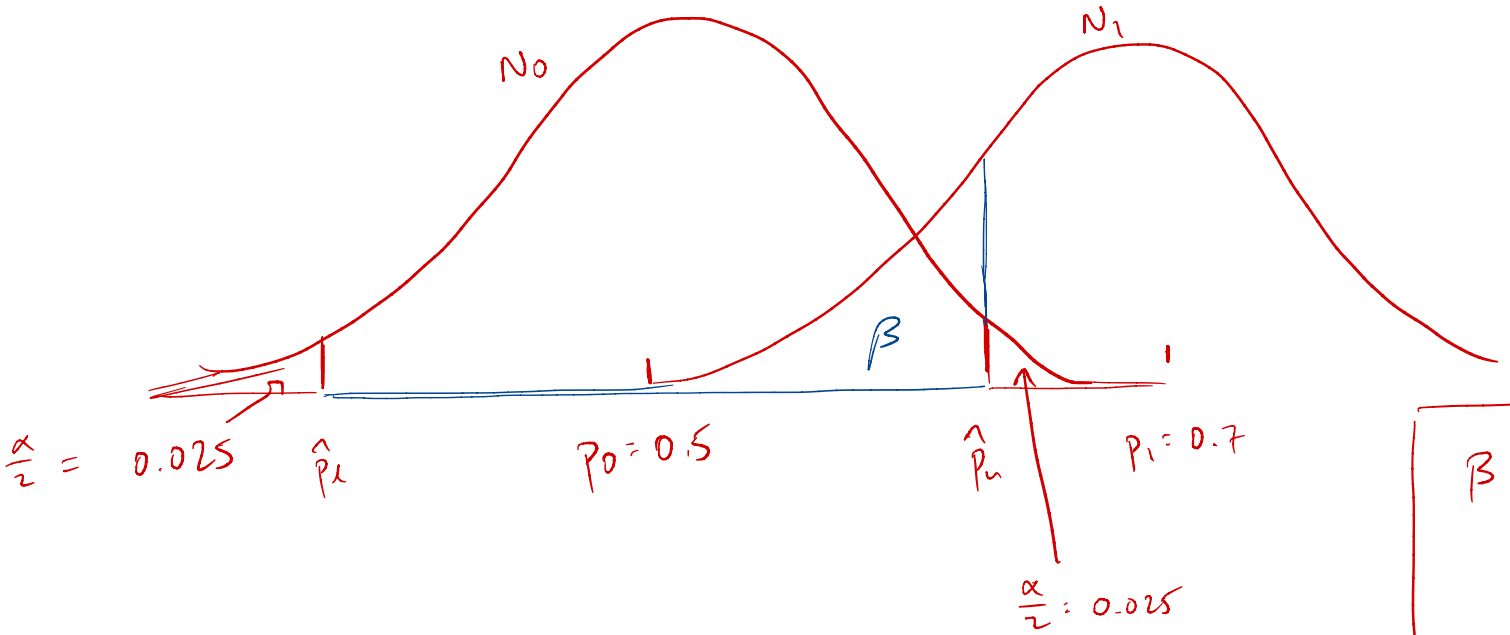
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{5/7 - 1/2}{\sqrt{\frac{1/2 \cdot 1/2}{70}}} = \underline{3.586}$$

$$p = 2 \cdot Pr(Z > 3.586) = 2 \cdot (1 - pnorm(3.586)) = \underline{0.00034}$$

$p < \alpha = 0.05$, so reject H_0.

Conclude proportion of people who like going to the beach is different than 0.5.

(c) (10 points) What is the power of the test in part (b) if the alternative hypothesis is that the true $p_1 = 0.7$?



Find $\hat{p}_u$ that corresponds to $\alpha = 0.05$.

$$\frac{\hat{p}_u - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \text{qnorm}(0.975) \rightarrow \hat{p}_u = 0.617$$

$$\frac{\hat{p}_\ell - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \text{qnorm}(0.025) \rightarrow \hat{p}_\ell = 0.383$$

$\mu = 0.7$

$\sigma = \sqrt{\frac{0.7 \times 0.3}{70}}$

β = area under N_1 between $\hat{p}_\ell$ and $\hat{p}_u$

$= \text{pnorm}\left(\hat{p}_u, 0.7, \sqrt{\frac{0.7 \times 0.3}{70}}\right) - \text{pnorm}\left(\hat{p}_\ell, 0.7, \sqrt{\frac{0.7 \times 0.3}{70}}\right)$

$= 0.0648$

Power $= 1 - \beta = \boxed{0.935}$

7. (10 points) χ^2 tests and nonparametric inference.

(a) (5 points) Consider a setting where you are running a goodness-of-fit test for a categorical variable that can take on four different values. What is the p-value of this test associated with a test statistic of $X^2 = 5$?

$df = 4 - 1 = 3$

$Pr(\chi^2_3 > 5) = 1 - \text{pchisq}(5, df = 3) = \boxed{0.172}$

(b) (5 points) Consider a setting where you have $n = 15$ paired data points and you know these data do not come from a normal distribution. When running a two-sided Wilcoxon signed-rank test where the null hypothesis is that the median difference between paired data points is zero, and given that the sum of positive ranks $T^+ = 74$, calculate the test statistic $T = T^+ - T^-$ and the associated p-value using a Normal approximation.

$T^+ + T^- = \sum_{i=1}^{15} i = \frac{15 \cdot 16}{2} = 120 \rightarrow T^- = 46$

$T = T^+ - T^- = 74 - 46 = 28$

$z_T = \frac{T - \mu_T}{\sigma_T} \sim N(0,1)$ where $\mu_T = 0$, $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} = 35.214$

$z_T = \frac{28}{35.214} = 0.795$

$p = 2 \cdot Pr(z > z_T) = 2 \cdot (1 - \text{pnorm}(0.795)) = \boxed{0.427}$

Blank page for scratch work

Blank page for scratch work