INFO 7390 – Advances in Data Sciences and Architecture Exam Solutions

Student Name: _____
Professor: Nik Bear Brown

Rules:
1. NO COMPUTER, NO PHONE, NO DISCUSSION or SHARING.
2. Ask if you don't understand a question.
3. You may three 8½"×11" sheets of notes (you may use both sides, written or printed as small as you like).
4. Time allowed. 1 hour 30 minutes.
5. Bring pen/pencil. The exam will be written on paper.

Q1 (5 Points)

A disease effects 1/100,000 men and 1/50,000 women. Assume there are equal numbers of men and women. A test has been devised to detect this disease. The test has a false negative rate of 5%. The false positive rate is 5%. Suppose a randomly selected person takes the test and tests positive. What is the probability that this person actually has the disease?

Solution:

Want to calculate P(disease|positive)

This equals P(disease|positive man) and P(disease|positive woman)

Plugging in the numbers gives

P(disease|positive man) = (0.00001)(0.95)/((0.00001)(0.95) + (0.99999)(0.05))

P(disease|positive woman) = (0.00002)(0.95)/((0.00001)(0.95) + (0.99999)(0.05))

It is OK to leave like this and average the two numbers since P(man)= P(woman)

But (0.00001)(0.95)/((0.00001)(0.95) + (0.99999)(0.05)) = 0.0001899658
And

(0.00002)(0.95)/((0.00001)(0.95) + (0.99999)(0.05)) = 0.00037993161

(0.00037993161+0.0001899658)/2 = 0.0002849487

Q2 (5 Points)

The probability that Bear parks in a no-parking zone and gets a parking ticket is 0.05. The probability that Bear has to park in a no-parking zone is 0.25. What is the probability that he will get a parking ticket? Assume we know that Bear arrives at school and has to park in a no-parking zone. What is the probability that he will get a parking ticket when he has to park in a no-parking zone?

Solution:

The probability that Bear parks in a no-parking zone and gets a parking ticket is 0.05. The probability that Bear has to park in a no-parking zone (he cannot find a legal parking space) is 0.25. Today, Bear arrives at school and has to park in a no-parking zone. What is the probability that he will get a parking ticket?

Define the events: *N* = Bear parks in a no-parking zone, *T* = Bear gets a parking ticket

Express the given information and question in probability notation:
"probability that Bear parks in a no-parking zone and gets a parking ticket is 0.05"
tells us that *P*(*N* and *T*) = 0.05.

"probability Bear has to park in the no-parking zone is 0.25" tells us that *P*(*N*) = 0.25 (1-0.25)= 0.25*0.75=0.1875

Q3 (5 Points)

If 3 dice are thrown, find the probability that (a) all 3 will show four, (b) all three will be alike, (c) all three will be different (i.e. no two or three are the same)?

Solution:

$$3.\,(a)\ \frac{1}{6^3}=\frac{1}{216}\quad (b)\ \frac{6}{6^3}=\frac{6}{216}=\frac{1}{36}\quad (c)\ \frac{(6)(5)(4)}{6^3}=\frac{(5)(4)}{6^2}=\frac{(5)(4)}{36}=\frac{5}{9}$$

Q4 (5 Points)

There are three groups as Groups A, B, and C. Descriptive statistics are presented below. Randomly draw the scores from populations with absolute homogeneity of variance. We wish to compare these groups using ANOVA. Calculate an F-score for the data below.

Group	***Mean***	***s^2***	***n***
A	14.7	18.892	60

B	15.6	27.145	30
C	19.9	29.433	10

Solution:

Any form of F-Score= MST/MSE is accepted for full credit. Don't have to calculate final numbers but must have the right numbers in the equations.

https://www.itl.nist.gov/div898/handbook/prc/section4/prc434.htm

Weighted Means Analysis

- The SS_{total} = 2399.
- The grand mean (*GM*) = .6(14.7) + .3(15.6) +.1(19.9) = 15.49.
- The weighted means SS_{among} = 60(14.7-15.49)^2 + 30(15.6-15.49)^2 + 10(19.9-15.49)^2 = 232.29. Notice that each group's deviation from the *GM* is weighted by its sample size.
- The SS_{error} = 2399 – 232.29 = 2166.7.
- $F = \frac{232.29/2}{2166.7/97} = \frac{116.145}{22.337} = 5.20$

Unweighted Means Analysis

Imagine that there is a good reason to believe that the three groups are actually equally represented in the population of interest. In fact, I got data from 100 subjects in each of the three groups. Unfortunately, rats got into my lab and chewed up a bunch of the data records. After losing much data to those dirty rats, I had greatly unequal sample sizes. I am comfortable with the notion that which data records were lost was unrelated to the scores. Accordingly, I decide to weight the groups equally.

- The harmonic mean sample size is $n_h = \frac{3}{1/60 + 1/30 + 1/10} = 20.$
- The variance of the means is VAR(14.7, 15.6, 19.9) = 7.723
- $MS_{among} = n_h \bullet s^2_{means} = 20(7.723) = 154.47$. Notice that this is the same formula we used to compute the among groups mean square with equal sample sizes, except for the use of the harmonic mean sample size.
- SS_{among} = 2(154.47) = 308.94.
- Typically the groups are not equally weighted when estimating the error variance (which is assumed to be constant across populations), so the SS_{error} = same as from the weighted means analysis, 2166.7. To get this by hand you would simply calculate the sums of squares within each group and then sum them. Do note that this will result in the total sum of squares not being equal to the sum of the among groups sums of squares and the error sum of squares.

- $F = \frac{308.94/2}{2166.7/97} = \frac{154.47}{22.337} = 6.915$

Q5 (5 Points)

A lip-reading test is given to normal-hearing subjects (n=22), hearing-impaired subjects (n=20) with pure-tone averages between 75 and 85 dB, and deaf subjects with pure-tone averages above 100 dB (n=21). Between-group variance (s^2_b) = 80, within-group variance (s^2_w) = 23. Is there a significant effect for hearing status (normal hearing vs. hearing impaired vs. deaf)? Stick with *F* values. You do not need to calculate p-values.

Solution:

Any form of F-Score= MST/MSE is accepted for full credit. Don't have to calculate final numbers but must have the right numbers in the equations.

The mean squares are formed by dividing the sum of squares by the associated degrees of freedom.

Let N = Sni. Then, the degrees of freedom for treatment, DFT = k - 1 and the degrees of freedom for error, DFE = N - k

The corresponding mean squares are:

MST = SST / DFT
MSE = SSE / DFE

We are directly given the between-group variance (s^2_b) = 80, within-group variance (s^2_w) = 23.

F = s^2_b / s^2_w = 80/23 = 3.48

Q6 (5 Points)

P (A) = 0.33, P (B) = 0.22, P (C) = 0.44,P (A ∩ B) = 0.11,P (B ∩ C) = 0.11,P (A ∩ C) = 0.11,P (A ∩ B ∩ C) = 0.05

Compute:

1. P ($\bar{A}$ |$\bar{B}$)
2. P (A ∩ B| A ∩ C)
3. P (B | $\bar{C}$)
4. P (B ∩ C | $\bar{A}$)

Solution:

Correct equations with wrong number got -1

$$P(\bar{A}) = 1 - 0.33$$

$$P(\bar{B}) = 1 - 0.22$$

$$P(\bar{C}) = 1 - 0.44$$

1. P $(\bar{A} \mid \bar{B}) = \frac{P(\bar{A} \cap \bar{B})}{P(\bar{B})} = 0.27 + 0.06/0.78$
2. P (A ∩ B| A ∩ C) = $\frac{P((A \cap B) \cap (A \cap C))}{P(A \cap C)} = 0.05/0.11$
3. P (B | $\bar{C}$) = $\frac{P(B \cap \bar{C})}{P(C)} = 0.06/0.56$
4. P (B ∩ C | $\bar{A}$) = $\frac{P(B \cap C \cap \bar{A})}{P(\bar{A})} = 0.06/0.67$

Q7 (5 Points)

Two cards are drawn from a well shuffled deck of 52 cards without replacement. Find the following probabilities.

a. The probability that the second card is a heart given that the first card is a spade.
b. The probability that the first card is a face card and the second card an ace.
c. The probability that one card is a heart and the other a club.

Solution:

Two cards are drawn from a well shuffled deck of 52 cards without replacement. Find the following probabilities.

a. The probability that the second card is a heart given that the first card is a spade.

Without replacement means that the first card is set aside before the second card is drawn and we assume the first card is a spade. There are only 51 cards to choose from for the second card. Thirteen of those cards are hearts.
It's important to notice that the question only asks about the second card.

$$P(\text{2nd heart} \mid \text{1st spade}) = \frac{13}{51}$$

The probability that the second card is a heart given that the first card is a spade is $\frac{13}{51}$.

b. The probability that the first card is a face card and the second card an ace.

Notice that this time the question asks about both of the cards.
There are 12 face cards out of 52 cards when we draw the first card. We set the first card aside and assume that it is a face card. Then there are four aces out of the 51 remaining cards. We want to draw a face card and an ace so use multiplication.

$$P(\text{1st face card and 2nd ace}) = \frac{12}{52} \cdot \frac{4}{51} = \frac{48}{2652} \approx 0.018$$

The probability that the first card is a face card and the second card an ace is approximately 0.018 or 1.8%.

c. The probability that one card is a heart and the other a club.

There are two ways for this to happen. We could get a heart first and a club second or we could get the club first and the heart second.

$$\begin{aligned} P(\text{heart and club}) &= P(\text{heart 1st and club 2nd or club 1st and heart 2nd}) \\ &= P(\text{heart 1st and club 2nd}) + P(\text{club 1st and heart 2nd}) \\ &= \frac{13}{52} \cdot \frac{13}{51} + \frac{13}{52} \cdot \frac{13}{51} \\ &\approx 0.127 \end{aligned}$$

The probability that one card is a heart and the other a club is approximately 0.127 or 12.7%.

Q8 (5 Points)

Calculate the determinant of the following 2x2 matrix. If this were used for a linear transformation, would it scale? What does it mean for the determinant to be negative?

-3	3
0	3

Solution: -3*3 – 3*0 = -9

Yes, it scales by a factor of 9.

For a 2×2 matrix (2 rows and 2 columns):

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

The determinant is:

$$|A| = ad - bc$$

"The determinant of A equals a times d minus b times c"

It is easy to remember when you think of a cross:

- Blue is positive (+ad),
- Red is negative (−bc)

Q9 (5 Points)

You run PCA on data with five independent variables and get five principal components with eigenvalues 19, 15, 11, 4, and 1. What percent of the variance is explained by each principal component?

Solution:

Principal component 1 = 19/(19+15+11+4+1)
Principal component 2 = 15/(19+15+11+4+1)
Principal component 3 = 11/(19+15+11+4+1)
Principal component 4 = 4/(19+15+11+4+1)
Principal component 5 = 1/(19+15+11+4+1)

Q10 (5 Points)

Do men cheat more than women? Suppose you take a group of 1000 randomly selected men and find that 231 had cheated. Suppose in a group of 1200 randomly selected women and find that 176 had cheated. Do the data show that men cheat more than women? Only find the Z-score or F-score. State your assumptions?

Solution:

Do men cheat more than women cheat? Suppose you take a group of 1000 randomly selected men and find that 231 had cheated. Suppose in a group of 1200 randomly selected women, 176 cheated. Do the data show that the men cheat more than women?

Solution:

1. State the random variables and the parameters in words.
   x_1 = number of men who cheat
   x_2 = number of women who cheat
   p_1 = proportion of men who cheat
   p_2 = proportion of women who cheat

2. State the null and alternative hypotheses and the level of significance
   $H_o : p_1 = p_2$ or $H_o : p_1 - p_2 = 0$
   $H_A : p_1 > p_2$ $H_A : p_1 - p_2 > 0$
   $\alpha = 0.05$

3. State and check the assumptions for a hypothesis test
   a. A simple random sample of 1000 responses about cheating from men is taken. This was stated in the problem. A simple random sample of 1200 responses about cheating from women is taken. This was stated in the problem.
   b. The samples are independent. This is true since the samples involved different genders.
   c. The properties of the binomial distribution are satisfied in both populations. This is true since there are only two responses, there are a fixed number of trials, the probability of a success is the same, and the trials are independent.
   d. The sampling distributions of $\hat{p}_1$ and $\hat{p}_2$ can be approximated with a normal distribution. $x_1 = 231$, $n_1 - x_1 = 1000 - 231 = 769$, $x_2 = 176$, and $n_2 - x_2 = 1200 - 176 = 1024$ are all greater than or equal to 5. So both sampling distributions of $\hat{p}_1$ and $\hat{p}_2$ can be approximated with a normal distribution.

4. Find the sample statistics, test statistic, and p-value

Sample Proportion:

$$n_1 = 1000 \qquad n_2 = 1200$$

$$\hat{p}_1 = \frac{231}{1000} = 0.231 \qquad \hat{p}_2 = \frac{176}{1200} \approx 0.1467$$

$$\hat{q}_1 = 1 - \frac{231}{1000} = \frac{769}{1000} = 0.769 \qquad \hat{q}_2 = 1 - \frac{176}{1200} = \frac{1024}{1200} \approx 0.8533$$

Pooled Sample Proportion, $\overline{p}$:

$$\overline{p} = \frac{231 + 176}{1000 + 1200} = \frac{407}{2200} = 0.185$$

$$\overline{q} = 1 - \frac{407}{2200} = \frac{1793}{2200} = 0.815$$

Test Statistic:

$$z = \frac{(0.231 - 0.1467) - 0}{\sqrt{\frac{0.185 * 0.815}{1000} + \frac{0.185 * 0.815}{1200}}}$$

$$= 5.0704$$

Q11 (5 Points)

Answer the following about the t-SNE cost function?

I. Describe the t-SNE cost function.

II. Does t-SNE create a probability distribution, if so which one(s)?

III. Is the cost function convex?

Solution:.

Describe the t-SNE cost function.

You need to mention that it is using a probabilistic distance between points to measure cost.

t-SNE – at a high level – basically works like this:

Step 1: In the high-dimensional space, create a probability distribution that dictates the relationships between various neighboring points

Step 2: It then tries to recreate a low dimensional space that follows that probability distribution as best as possible.

I. Does t-SNE create a probability distribution, if so which one(s)?

- t-SNE creates a **probability distribution** using the **Gaussian** distribution that defines the relationships between the points in high-dimensional space.

  t-SNE uses the **Student t-distribution** to **recreate** the probability distribution in low-dimensional space.

I. Is the cost function convex?

No

- t-SNE optimizes the embeddings directly using gradient descent. The cost function is **non-convex** though, meaning there is the risk of getting stuck in local minima.

  As explained in StatQuest: t-SNE, Clearly Explained https://youtu.be/NEaUSP4YerM

Q12 (5 Points)

What are the loadings (or factor loadings) in PCA (Principal Component Analysis)? How are they useful?

Solution:

Factor **loadings** (factor or component coefficients) : The factor **loadings**, also called component **loadings**in **PCA**, are the correlation coefficients between the variables (rows) and factors (columns). Analogous to Pearson's r, the squared factor **loading** is the percent of variance in that variable explained by the factor.

The indicate the importance of a feature in an eigenvector.

Q13 (5 Points)

What's the difference between principal component analysis and multidimensional scaling? Does one or the other use eigen-decomposition?

Solution:

MDS is done by transforming distances into similarities and performing PCA (eigen-decomposition or singular-value-decomposition) on those. PCA might be called the algorithm of the simplest MDS.

Q14 (5 Points)

Assume PCA is being used for data reduction on one-hundred independent variables. Will PCA (Principal Component Analysis) perform well when one eigenvalue equals ten times the sum of the other ninety-nine eigenvalues? Is this possible?

Solution:

Yes. Most of the variance is being captured by one eigenvalue.

Q15 (5 Points)

Describe the k-means algorithm. The problem is computationally difficult (NP-hard). What cost function does it minimize? Is k-means guaranteed to reach a global optimum? Is k-means guaranteed to converge?

Solution: ,

k-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

The algorithm works as follows:

1. First we initialize k points, called means, randomly.
2. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
3. We repeat the process for a given number of iterations and at the end, we have our clusters.

What cost function does it minimize?

The Euclidean distance between the centroid and the points.

Is k-means guaranteed to reach a global optimum?

No

Is k-means guaranteed to converge?

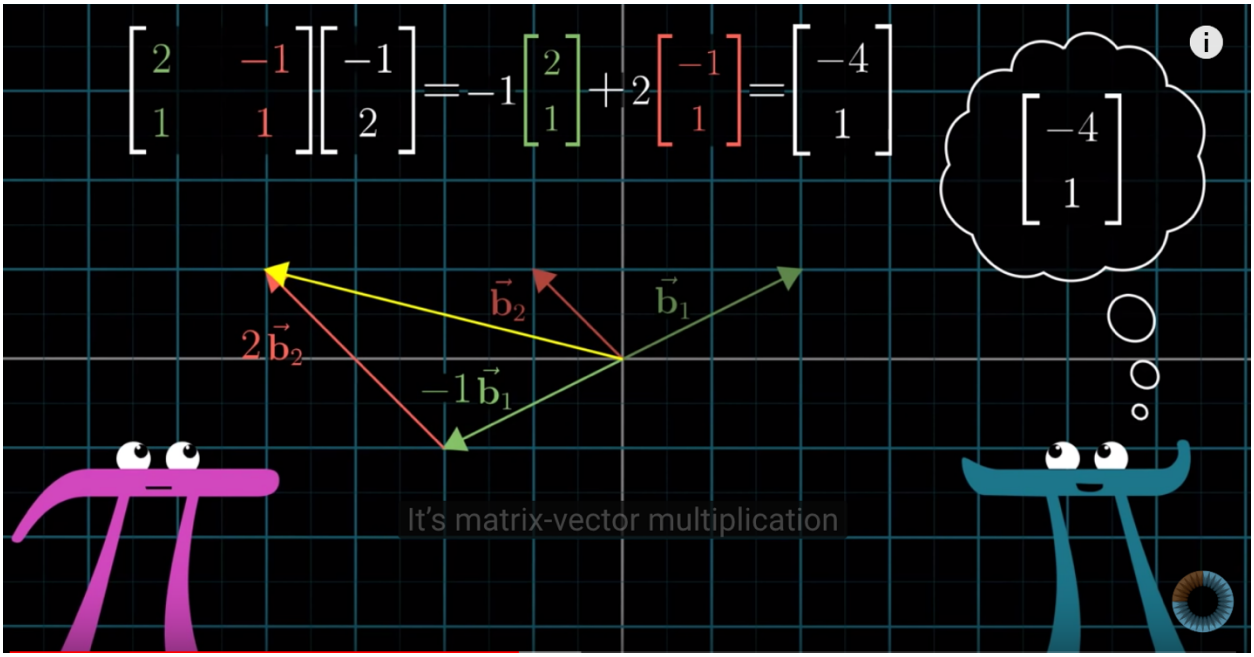
Yes

Q16 (5 Points)

Assume a vector space has the basis vectors *v*1=(2,1) and *v*2=(-1,1). What is the change of basis that transforms a point using the unit vectors for 2D Cartesian coordinates to a point in that space?

Solution:

This is directly from the video. The fact that we are starting from unit vectors for 2D Cartesian coordinates makes the transformation matrix obvious.



Change of basis | Essence of linear algebra, chapter 13 https://youtu.be/P2LTAUO1TdA

Q17 (5 Points)

Assume two classes have average GPAs as follows

Class 1	Class 2
*n*1=17	*n*2=35

$x1=3.51$	$x2=3.24$
$s1=0.51$	$s2=0.52$

n is the number of students
x is the mean GPA
s is the standard deviation for the GPA

Construct a point estimate and a 99% confidence interval for $\mu 1-\mu 2$

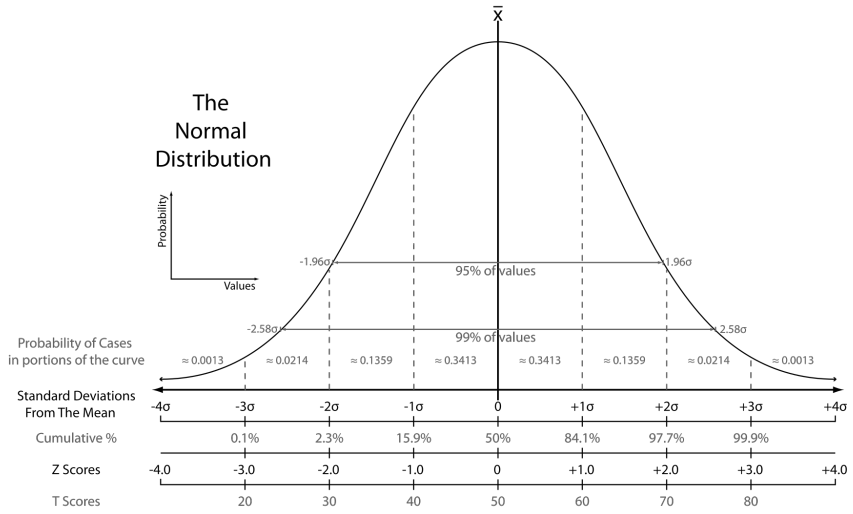
Solution:

Create a novel homework problem with solutions for example Inference for Two Means: Introduction. See the example Inference for Two Means: Introduction https://youtu.be/86ss6qOTfts

$x1-x2=3.51-3.24=0.27$
$a = 1- .99 =0.01$

so we need $Za/2=Z.005 = 2.576$
$(x1-x2)\pm Z\alpha/2\sqrt{(s2^2/n1+s2^2/n2)}=0.27\pm 2.576\sqrt{(0.51^2/174+0.52^2/355)} =0.27\pm 0.12$

Use the following probability table for questions 18 through 20



Q18 (5 Points)

Given H_0: $\mu=0$ and H_a: $\mu<-2$ and n = 36. Assume at our significance level we reject the null hypothesis at Z-values less than -2.

At what values do we reject the null hypothesis? What is the probability of a Type I error?

Solution:

Calculating Power and the Probability of a Type II Error (A One-Tailed E...
https://youtu.be/BJZpx7Mdde4

Q19 (5 Points)

Assume the true value of μ is -1 in Q18.

What is the power of the test in Q18?

Solution:

Calculating Power and the Probability of a Type II Error (A One-Tailed E...
https://youtu.be/BJZpx7Mdde4

Q20 (5 Points)

Using the standard score graph above give values for the following confidence intervals

A. 90% Two-sided
B. 95% Two-sided
C. 99% Two-sided
D. 95% One-sided μ<mean
E. 90% One-sided μ>mean

Solution:

Any explanation that shows ROUGHLY where 90% 95% 99% are on the chart as well as the two-sided is above and below and one-sided is one-sided were accepted for full credit.

90% 95% 99%

1.645 1.96 2.576