COMP 4441 Midterm Examination

Troy Jennings

QUESTION 1

An team has 8 members. Denote them by $\{1, 2, 3, 4, 5, 6, 7, 8\}$. Construct a reasonable, standard model for selecting a team member in such a way that any member is equally likely to be selected, recording the member selected, and repeating this process one more time using the remaining set of seven team members. Thus outcomes will be pairs of values (a, b) with $a, b \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ and $a \neq b$. You don't have to explain the model, just provide the values requested below.

What is the probability of the outcome $(5, 3)$? (5 points)

For picking the first element, we have $\binom{8}{1}$ ways to choose the first element (8 possibilities), and $\binom{7}{1}$ ways to choose the second element (7 possibilities), for a total of 56 possible ways to choose 2 elements without replacement. This gives the probability of any event (a, b) where $a \neq b$ and $a, b \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ as:

$$P(a, b) = \frac{1}{56}$$

What is the probability of the event $\{(a, b)|a < b\}$? (5 points)

If we pick 1 as the first element (a), then we have $\binom{7}{1}$ ways to pick b such that $a < b$. If we pick 2 as the first element (a), then we have $\binom{6}{1}$ ways to pick b such that $a < b$. We would continue this process down to selecting $a = 8$. We can generalize this process as $\sum_{k=1}^{8-1} k$ ways to pick a pair $(a, b)|a < b$. Given previous information, this gives the probability of the event $\{(a, b)|a < b\}$ as:

$$P\{(a, b)|a < b\} = \frac{28}{56} = \frac{1}{2}$$

QUESTION 2

Consider a continuous random variable X with the probability density function defined by $f(x) = \frac{3}{4}(1 - x^2)$ for $x \in [-1, 1]$ and $f(x) = 0$ otherwise. What is the probability of the event consisting of the interval $[-\frac{1}{3}, 0]$? (10 points)

```
integrand <- function(x){ (.75)*(1 - x^2) }
integrate(integrand, lower= -1/3, upper= 0)
```

```
## 0.2407407 with absolute error < 2.7e-15
```

Using integration, we find that the probability that event is in the interval $[-\frac{1}{3}, 0]$ is approximately 0.2407407.

QUESTION 3

Construct a reasonable model for rolling a fair die twice and recording the results in order. You don't have to explain the model, just provide the information requested below. Let X be the event that the first number is odd. Let Y be the event that the second number is even. Let Z be the event that both numbers have the same parity, that is, both are even or both are odd.

Let $P(X) = \frac{|X|}{|S|} = \frac{3}{6} = \frac{1}{2}$, $P(Y) = \frac{|Y|}{|S|} = \frac{3}{6} = \frac{1}{2}$, and $P(Z) = \frac{9}{36}$.

Are the events X and Z independent? (5 points)

We will let S be the set of all outcomes, (a, b), such that S includes all 36 possible pairs of roll combinations. Here, $P(X) = \frac{1}{2}$, since there are 18 ways to roll an odd number for the first roll. Then $P(Z) = \frac{9}{36}$ since there are 9 ways to roll an odd number for the second roll for which the first roll was an odd number. The intersection, $P(X \cap Z)$, is $\frac{9}{36]}$ since there are 9 combinations in Z that are also in X. Since $P(X \cap Z) = \frac{9}{36}$ is not equal to $P(X)P(Z) = \frac{1}{6}$, events X and Z are dependent.

Are the events $X \cap Y$ and Z independent? (5 points)

Here, $P(X) = \frac{1}{2}$ and $P(Y) = \frac{1}{2}$, and $P(X \cap Y) = \frac{9}{36}$ since there are 9 ways to roll any odd on the first roll and any even on the second roll. Again $P(Z) = \frac{9}{36}$ and since $P(X \cap Y)$ is a subset of rolls with different parity, the intersection with Z will be the empty set, or $\frac{0}{36}$. Since 0 is not equal to $P(X \cap Y)P(Z) = \frac{1}{9}$, the events $X \cap Y$ and Z are also dependent.

QUESTION 4

Consider a continuous random variable X with the probability density function defined by $f(x) = c(x + 2)$ for $x \in [-2, 0]$, $f(x) = c(2 - x)$ for $x \in [0, 2]$ and $f(x) = 0$ otherwise. What is value of c? (5 points)

First, let

$$F(x) \int_{-2}^{0} c(x + 2)\, dx = c \cdot \int_{-2}^{0} (x + 2)\, dx$$

. Then, we can rewrite the integrals, evaluate them, and solve for c.

Rewriting the integral, we have

$$c \cdot \left[\int_{-2}^{0} x\, dx + 2 \cdot \int_{-2}^{0} 1\, dx) \right]$$

Evaluating the integrals independently, we are left with

$$c \cdot \left[\frac{x^2}{2} \bigg|_{-2}^{0} + 2 \cdot (x \bigg|_{-2}^{0}) \right] = c \cdot \left[(0 - \tfrac{4}{2}) + 2(0 + 2) \right] = c \cdot [2]$$

Setting our equation to 1, we have $1 = c \cdot 2$ and solving for c, $c = \frac{1}{2}$.

QUESTION 5

If you model the data below as the result of 8 independent Bernoulli trials with probability of success equal to p, what is the maximum likelihood estimate of p? (5 points)

```
b <- c("success", "success", "failure", "success", "success", "success","success", "failure")
mle <- sum(b == 'success') / length(b)
print(mle)
```

```
## [1] 0.75
```

From class, we know that the maximum likelihood estimate of a binomial: $MLE(\theta) = \frac{x}{n}$, where $x =$ the number of successes and $n =$ the number of trials. Here, we have $x = 5$ and $n = 8$, so the MLE is $\frac{6}{8} = .75$.

QUESTION 6

If you model the data below as a sample from a Normal distribution, what is the maximum likelihood estimate for the σ^2 in the density? (5 points)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

```
v <- c(0.59, -0.55, 1.95, 1.02, 0.30)
## calculate x-bar
mle.mu <- mean(v)
## calculate sigma^2
sum((v - mle.mu)^2)/length(v)
```

```
## [1] 0.678456
```

QUESTION 7

If you perform a test of the null hypothesis that the values in v from question 6 are from a population for which the mean of 5 observations is Normally distributed with mean 0 and variance $\frac{1}{5}$, what is the two-sided p-value? Please use the two-sided p-value understood as the probability of observing a value of the mean of v under the null hypothesis with density less than or equal to that of the observed test statistic. Is the result strong evidence against the null hypothesis? (10 points)

```
## population parameters
pop.mean = 0
pop.sd = sqrt(1/5)
## v is our sample data

## test stat is the observed mean
obs.mean <- mean(v)
n <- length(v)
```

```r
## run simulation with the sample size as the sample
iter = n
samp.means = rep(NA, iter)
for (i in 1:iter){
  samp.means[i] = mean(rnorm(n= n, mean= pop.mean, sd= pop.sd))
}

## compute the 2-sided p-value
mean(samp.means >= obs.mean)*2
```

```
## [1] 0
```

From the two-sided p-value, all of the simulated sample means are greater than the observed statistic. Therefore, this is strong evidence in support of the null hypothesis.

QUESTION 8

In the data “demog_data.csv”, if you model the “INCTOT” variable as a linear function of the “AGE” variable, what slope and intercept give the line with the least squares best fit? (You may use a built-in function for this or the formulas derived in class) (10 points)

This question uses 2018 data primarily for Denver county accessed through IPUMS-USA, University of Minnesota, www.ipums.org , Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 9.0 [dataset]. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D010.V9.0 The PUMA-to-county restriction was done using MABLE, http://mcdc.missouri.edu/websas/geocorr12.html.

The data are restricted to women in the age range 30-80 in PUMAs primarily in Denver.

```r
## read in the data for the problems
dat <- read_csv('demog_data.csv')
```

```
## Warning: Missing column names filled in: ’X1’ [1]

##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double()
## )
## i Use ‘spec()‘ for the full column specifications.
```

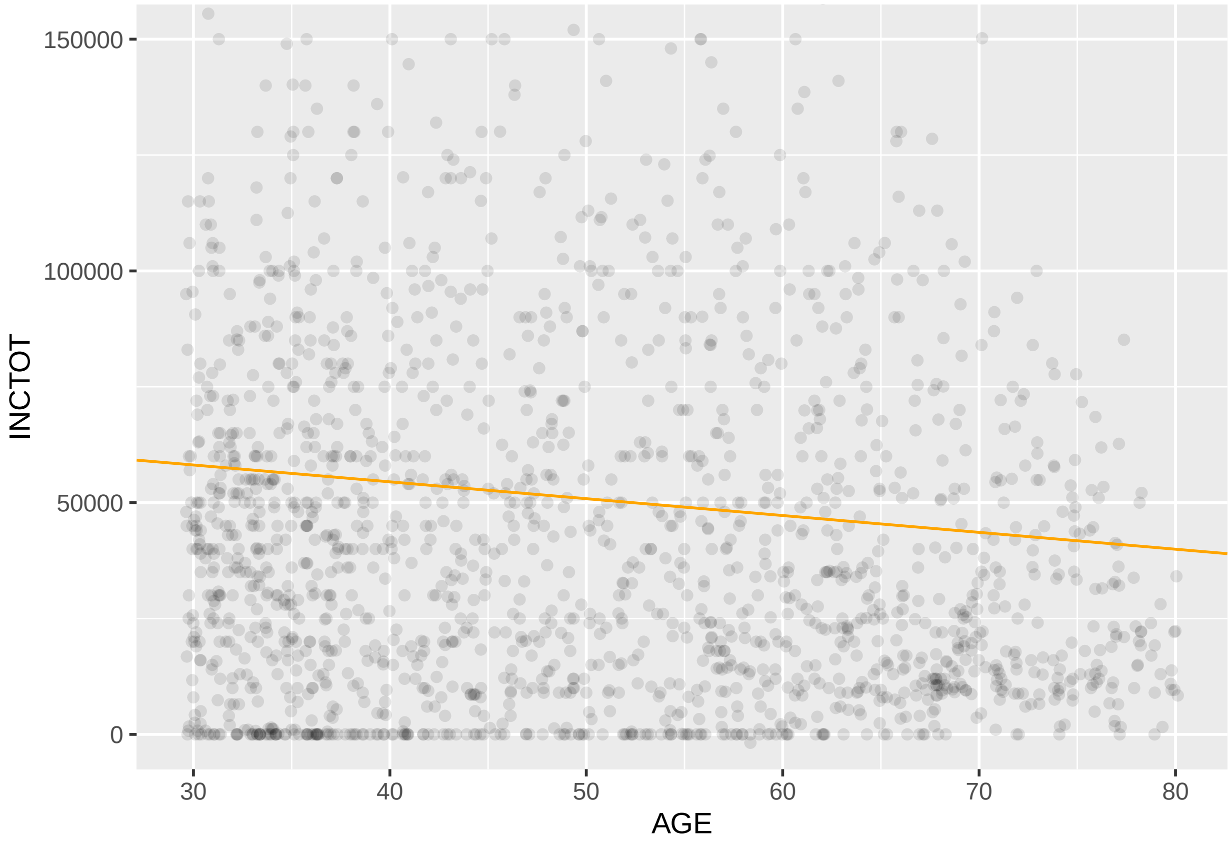
```r
## calculate the linear model and grab the coefficients
coeffs <- lm(INCTOT~AGE, data= dat)$coefficients
## calculate the change in income associated with an increase of 1 in "EDUC"
print(coeffs)
```

```
## (Intercept)         AGE
##  69025.2720   -363.3782
```

QUESTION 9

Make a scatterplot with the "AGE" variable on the horizontal axis and the "INCTOT" variable on the vertical axis. Include the line computed in question 8. (If you used the built-in function for question 8, please extract the values of the slope and intercept from the fitted model object, rather than using copy-paste.) (10 points)

```
## plot the data
g <- ggplot(dat, aes(x= AGE,y= INCTOT)) + geom_jitter(alpha= .1) +
    coord_cartesian(ylim= c(0, 1.5e5)) +
    geom_abline(slope= coeffs[2], intercept= coeffs[1], color = 'orange')
g
```



QUESTION 10

In the data set "demog_data.csv" what is the mean value of INCTOT for women in each category of EDUC? (For full credit, please do this without looping through the EDUC values. You can use "summarize" from dplyr, with group_by.) (5 points)

```
## find the average income by education group
dat %>% group_by(EDUC) %>% summarize(avg.income= mean(INCTOT))
```

```
## # A tibble: 11 x 2
##     EDUC avg.income
##     <dbl>      <dbl>
##   1     0     15528.
##   2     1      7990
##   3     2     11606.
##   4     3     13692.
##   5     4     17440
##   6     5     14967.
##   7     6     25262.
##   8     7     36395.
##   9     8     41731.
##  10    10     62338.
##  11    11     76741.
```

QUESTION 11

Please restrict the data set "demog_data.csv" to cases in which "INCTOT" is greater than 0 and "EDUC" is greater than 4. What is the mean value of "INCTOT" for the remaining observations? (5 points)

```
## filter for INCTOT  0 and EDUC > 4
dat.ftr <- filter(dat, INCTOT > 0, EDUC > 4)
## find the mean of the remaining observations
print(mean(dat.ftr$INCTOT))
```

```
## [1] 57781.92
```

QUESTION 12

Suppose that each of the 6 members of a jury returns a "guilty" vote independently with probability 0.3. What is the probability of 4 or more "guilty" votes? (5 points)

```
## calculate the probability of 4 guilty votes + the probability of more than 4 guilty votes
dbinom(4, size= 6, prob= 0.3) + pbinom(4, size= 6, prob= 0.3, lower.tail = F)
```

```
## [1] 0.07047
```

QUESTION 13

At what value x does the cumulative distribution ϕx satisfy $\phi(x) = .95$ if $\phi(x)$ is the cumulative distribution of a Normal random variable with $\mu = 70$ and $\sigma^2 = 49$, so $\sigma = 7$? (5 points)

```
## calculate the quantile at which the Z-score for phi(x) = 0.95
qnorm(p= 0.95, mean= 70, sd= 7)
```

```
## [1] 81.51398
```

QUESTION 14

Set the random seed to 12345. Create a vector of 60 samples from the exponential distribution (“rexp”) with rate equal to 1/(1,000,000). Create a matrix of these values with 10 rows and 6 columns in such a way that the vector of the first 10 integers in the sample equals the first column in the matrix, the vector of the second 10 integers in the sample equals the second column in the matrix, and so on. Use the “apply” function to find the median value in each row.(5 points)

```
## set the seed
set.seed(12345)
## sample 60 values from the exponential distribution
samp <- rexp(n= 60, rate=(1/1000000))
## generate a matrix with 10 rows and 6 columns
mat <- matrix(data= samp, nrow= 10, ncol= 6, byrow= FALSE)
## find the median of each row
med <- apply(X= mat, MARGIN =1, FUN= function(x) median(x))
dat <- data.frame(row.num= seq(1, 10, 1), row.median= med)
print(dat)
```

```
##    row.num row.median
## 1        1   873397.4
## 2        2   453153.4
## 3        3  1030464.2
## 4        4   496973.2
## 5        5   271763.1
## 6        6   586052.6
## 7        7  1260217.7
## 8        8  1572977.2
## 9        9   806750.9
## 10      10   953688.1
```