# Stat 23400: Spring 2016 Midterm Exam

**FULL NAME (not nickname)** ________________________________________

**SECTION** □ TR 9:00 (K. Dey) □ TR 10:30 (Dr. Collins) □ TR noon (M. Jahangoshahi)

**PLEASE DO NOT OPEN THIS EXAM**

**UNTIL... THE OFFICIAL EXAM START TIME. THANKS!**

## Exam Non-Disclosure Agreement

I agree that I will not discuss this exam with anyone until after 4pm Friday, May 6, 2016. This restriction includes the exam contents, length, difficulty, style, etc. I understand that if I am found to have disclosed any information about the exam outside of these boundaries, I will be subject to disciplinary action by the University, including the possibility of failing STAT 23400.

**Signature:** ________________________________________

**There are 7 questions: 25 parts (each worth 4 points)**
**...and one extra credit part is also worth 4 points.**

**DO NOT WRITE IN THIS GRADING TABLE:**

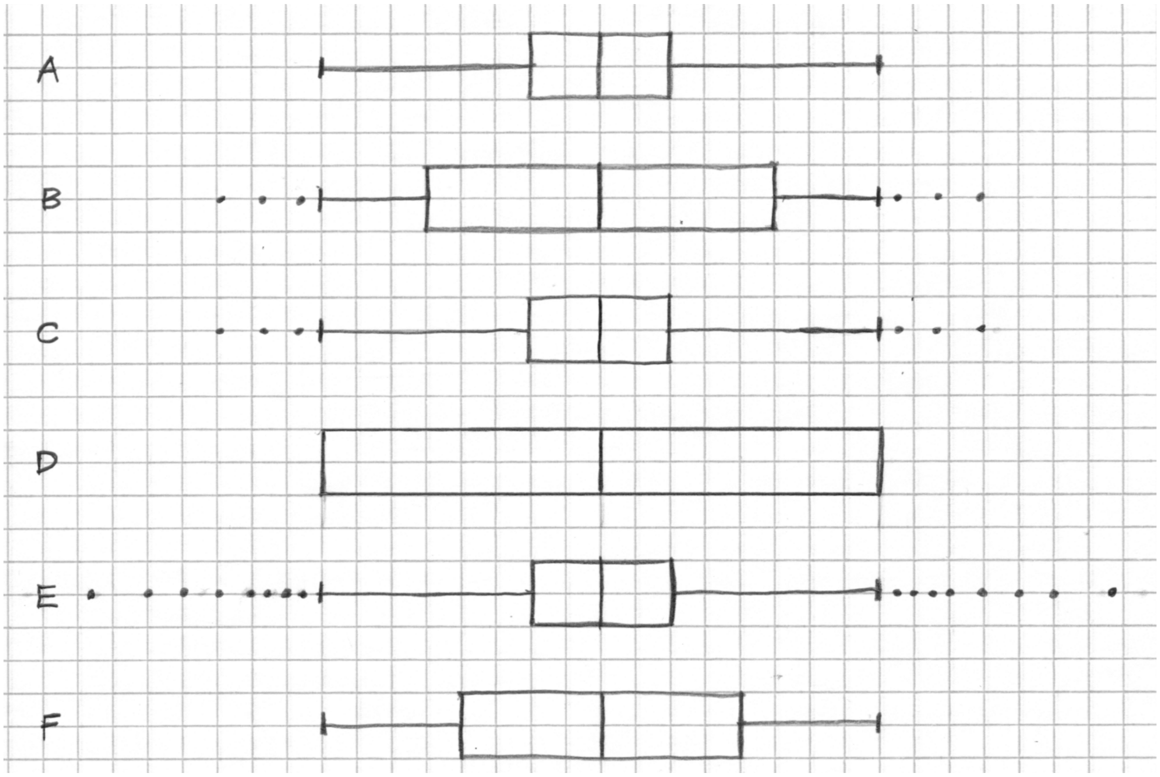
| Question Number | Page Number | Points | Question Parts | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | □ | □ | □ | | | |
| 2 | 3 | □ | □ | □ | □ | □ | |
| 2 | 4 | □ | □ | □ | □ | □ | □ |
| 2 | 5 | □ | □ | □ | | | |
| 3 | 6 | □ | □ | □ | □ | | |
| 4 | 7 | □ | □ | □ | | | |
| 5 | 8 | □ | □ | □ | □ | | |
| 5 | 9 | □ | | | | | |
| 6 | 10 | □ | (Extra Credit) | | | | |
| 6 | 11 | □ | □ | □ | | | |
| 7 | 12 | □ | | | | | |
| Total | | | | | | | |

1. The six boxplots below have the following in common:

   (1) all are a plot of $n = 1,000$ data values

   (2) all have the same median

   (3) all are symmetric



(a) Which boxplot shows data most consistent with a uniform density (population)? Explain your reasoning.

(b) Which boxplot shows data most consistent with a normal density (population)? Explain your reasoning.

2. Millenials are adults aged 19-35 in 2016 (born 1981-1997).

Does a larger proportion of male millenials live with their parents than female millenials?
...or is that just a stereotype?

Here are data from a survey of $n = 300$ millenials who graduated from a high school in a Chicago suburb. The school has about 22,000 alumni who are millenials: 10,000 female and 12,000 male.

| | Females | Males | Total |
|---:|---:|---:|---:|
| Yes, I'm living with my parents. | 30 | 70 | 100 |
| Not, thank you very much! | 70 | 130 | 200 |
| Total | 100 | 200 | 300 |

(a) Describe the variable the researchers are considering to be the explanatory variable?

Describe the variable the researchers are considering to be the response variable?

(b) What is the (observed) marginal proportion of alumni living with their parents?

Just write a fraction: $\frac{\text{value A}}{\text{value B}}$

(c) What statistic (a number calculated from the data) would be best to answer the question:

Does a larger proportion of male millenials live with their parents than female millenials?

Write an answer as the difference of two *conditional* proportions: $\frac{\text{value C}}{\text{value D}} - \frac{\text{value E}}{\text{value F}}$

(d) Fill in values (numbers) in the two blanks in the following statement:

Suppose male and female millenials from this high school
actually do live with their parents in equal proportions.

Then of the 100 alumni who reported living with their parents we would have expected...

________ to be females and ________ to be males.

Here is a shuffling simulation (a resampling procedure) that (I think) is a good representation of how the data were actually collected (if male and female millenials from this high school actually do live with their parents in equal proportions.)

(e) Use two stacks of cards: One stack of 10,000 cards. The other: 12,000 cards.

What do these cards represent (in context)? That is, why 10,000 and 12,000 cards?

(f) For the stack of 10,000 cards, use 3,333 red cards and 6,667 black cards.

Why did I choose these particular counts for the red and black cards?

(g) Shuffle all 10,000 cards and select a sample.

How many cards would you select? Explain your reasoning.

(h) Now explain what you would do with the other stack of 12,000 cards.

How many cards should be red? __________

How many cards should be black? __________

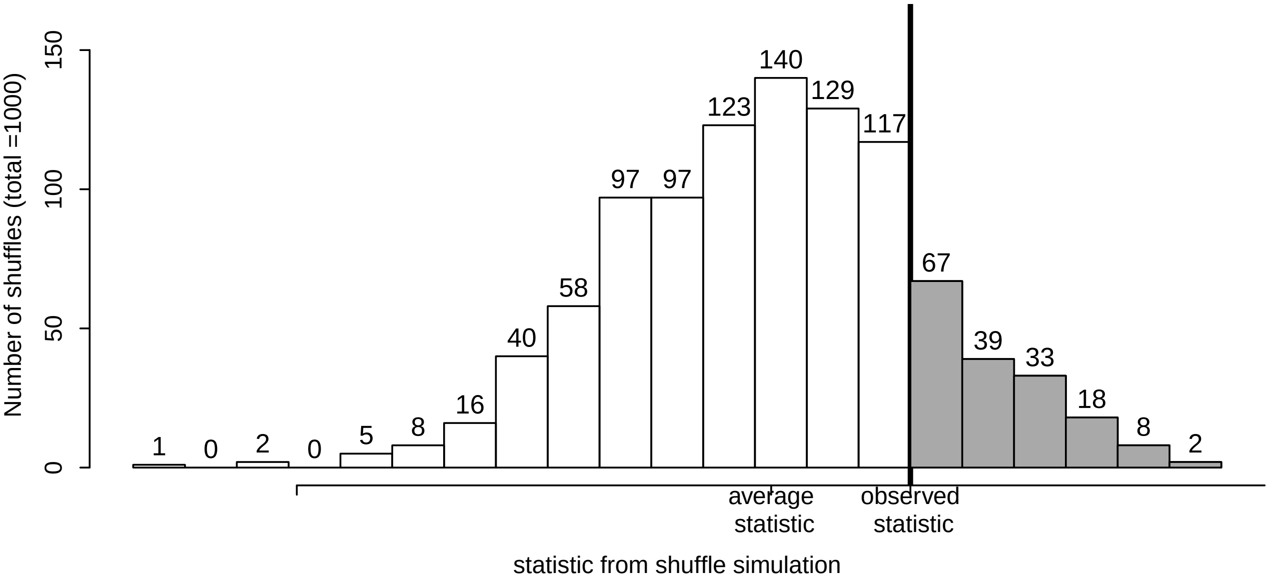
How many cards would you select for a sample? __________

(i) 1,000 times, we plan to select a sample from each stack as in parts (g) and (h).

After each time we shuffle and select a sample of cards from both stacks, we will calculate one statistic (one number) from the sampled cards.

Describe how you would calculate this statistic from the sampled cards.
Hint: See part (c).

Here are the results after repeating the shuffling simulation 1,000 times:



(j) The x-axis values are not shown: just the words "observed statistic" and "average statistic".

What is the value of the observed statistic (from the data)? ____________

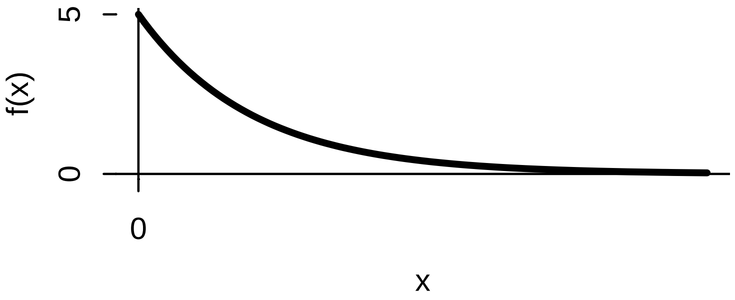
What do you expect to be the value of the average statistic? ____________

(k) The graph provides evidence to help you answer the original question:
Does a larger proportion of male millenials live with their parents than female millenials?

Explain your conclusion and your reasoning.

**NOTE: Explain as if to someone who has not taken a statistics course.**

3. Suppose $X$ is a continuous random variable with the right-skewed probability density function (pdf): $f(x) = 5e^{-5x}$ for $x \geq 0$. Here is a graph of the pdf.



On HW #1 you already calculated the following:

$$\textbf{(1)} \int_0^\infty 5\,e^{-5x}\,dx \;=\; 1 \qquad \textbf{(3)} \int_0^k 5\,e^{-5x}\,dx = 1 - e^{-5k}$$

$$\textbf{(2)} \int_0^\infty 5\,x\,e^{-5x}\,dx = \frac{1}{5} \qquad \textbf{(4)} \int_0^\infty 5\,e^{tx}\,e^{-5x}\,dx \;=\; \frac{5}{5-t} \qquad \text{for } t < 5$$

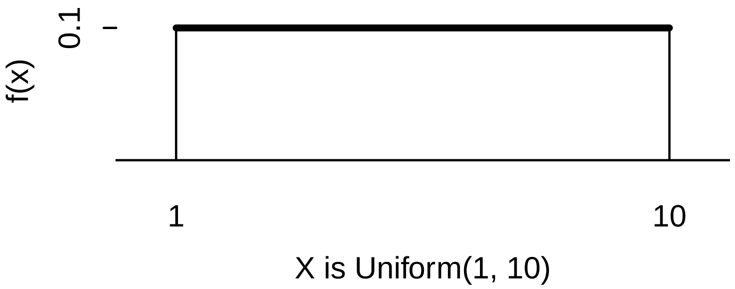
Do not integrate any function for this exercise. You won't have time.

(a) Confirm that $f(x)$ is a valid pdf. Indicate which of the four results you use to decide.

(b) Show how you would calculate $P(X \leq 0.4)$ by plugging the number 0.4 directly into one of the results (1), (2), (3), or (4). Do not calculate.

(c) Use the moment generating function of $X$ to calculate $E(X)$.
Indicate which of the four results you are using and then show your work.

4. Your friend is also a STAT 234 student.
He claims to have simulated 1,000 outcomes from a continuous Uniform(1, 10) density.



He shows you the following summary statistics of his simulated data.

```
  min    Q1 median    Q3   max  mean
1.001 4.258  5.546 7.753 9.989 5.465
```

(a) Of course your friend did the simulation correctly!

But, explain to him which one of the statistics is likely a typographical error and why.

(b) What would you expect a <u>normal</u> quantile plot of these data to look like?

Just show a quick sketch with labels (words) on both axes.
<u>No numbers are needed</u> on either axis.

5. A sample of $n = 25$ is planned to be collected from a population $(X)$ with mean $\mu = 5$.

   (a) Let $a > 0$ be a positive constant and $c = \mu + a$.

   Which will be smaller, $P(\overline{X} > c)$ or $P(X > c)$?
   Choose one and explain. Do not calculate anything.

   (b) We collected one sample of $n = 25$ from the population.

   Normal quantile plots on the next page show the (standardized) data plus 7 simulated samples from a standard normal density: Normal(0,1).

   What do you conclude from these plots? Explain your reasoning
   (Yes, this is a very open-ended question.)

   (c) What probability distribution would you use to estimate $P(\overline{X} > c)$?

   **Do not calculate!**

   Just state what probability distribution you would use
   and how your answer to part (b) supports that choice.

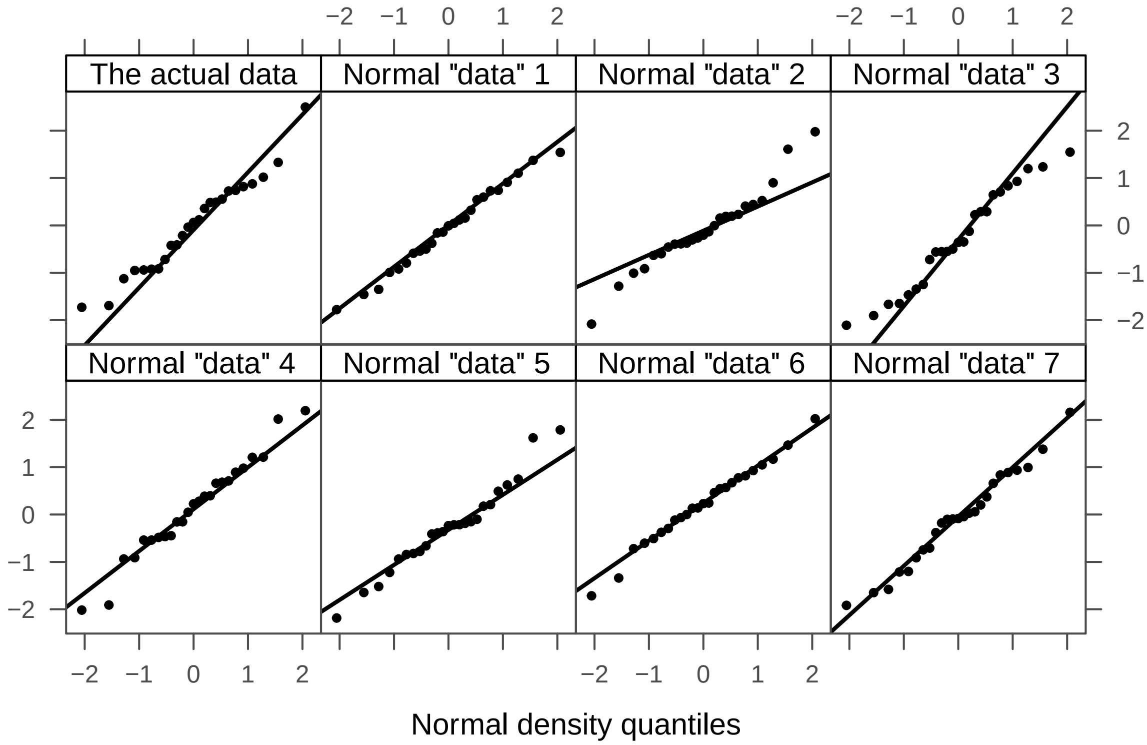
(d) Here are (rounded) summary statistics of the sample.

```
min Q1 median Q3 max mean sd  n missing
  2  4      6  7  11    6  2 25       0
```

Of course, this is just one of many possible samples of $n = 25$ from the population. There is a distribution of all possible $\overline{X}$'s from all possible samples.

Calculate an estimate for $P(\overline{X} > 5.8)$.

Hint: Use what you know about the sampling distribution of $\overline{X}$'s, the population, and, as needed, the summary statistics from this one sample.



6. Let $X_1, X_2, \ldots, X_n$ be random variables representing a possible random sample from a population with mean $\mu$ and variance $\sigma^2$.

   (a) **Part (a) is EXTRA CREDIT**. Just skip this page until end of exam.

   Show that $\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2 = \left[\sum_{i=1}^{n}(X_i - \mu^2)\right] - n(\overline{X} - \mu)^2$

   90% of this proof is just algebra. Please do not explain algebra. But...
   **Clearly indicate any statistical formula you use and where you use it.**

   I'll get you started with a little trick: subtracting and then adding back $\mu$.

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2$$

$$= \sum_{i=1}^{n}[\,(X_i - \mu) - (\overline{X} - \mu)\,]^2$$

$=$

$=$

$=$

$=$

$=$

$=$

$=$

$$= \left[\sum_{i=1}^{n}(X_i - \mu^2)\right] - n(\overline{X} - \mu)^2$$

(b) Use the claim in part (a) as a starting point to show that $E\left[\sum_{i=1}^{n}(X_i-\overline{X})^2\right]=(n-1)\sigma^2$

$\boxed{\textbf{Just use the claim in part (a) even if you did not prove it.}}$

**Clearly indicate any statistical formula you use and where you use it.**

Show your work here.

$$E\left[\sum_{i=1}^{n}(X_i-\overline{X})^2\right]=$$

$$=$$

$$=$$

$$=$$

$$=$$

$$=(n-1)\sigma^2$$

(c) Use the claim in part (b) as a starting point to show that the sample variance is an unbiased estimator of $\sigma^2$.

$\boxed{\textbf{Just use the claim in part (b) even if you did not prove it.}}$

**Clearly indicate any statistical formula you use and where you use it.**

7. You plan to collect a random sample of $n = 300$ students in the College to estimate $p =$ the proportion of students who have taken (or plan to take) a STAT course.

   Dr. Collins makes an educated guess that $p = 3/4 = 0.75$.

   Use the information above and a normal approximation to estimate the probability that you will observe a $\widehat{p}$ within 5% (0.05) of the true (but unknown) value $p$.

   This is a very open-ended question, so show and explain/justify your work.