Final Exam
Math 183, UCSD, Fall 2017
Thursday, December 14th, 11:30am–2:29pm
Instructor: Eddie Aamari

- Write your PID, Name and Section in the spaces provided above.
- Do not unstaple the pages.
- Write your solutions clearly in the spaces provided.
- Answers written outside the answer boxes will not be graded.
- You may use a calculator (any type is fine), but no other electronic devices.
- You may not use your cell-phone, tablet, or computer as a calculator.
- You may use your single sheet of handwritten notes.
- Put away (and silence!) your cell phone and other devices that can be used for communication or can access the Internet.
- Show all of your work; no credit will be given for unsupported answers.
- You may either give exact answers (like 1/3) or round to three decimal places (like 0.333).

DO NOT TURN PAGE UNTIL INSTRUCTED TO DO SO

<table>
<thead>
<tr>
<th>Exercise</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
<th>IX</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Points</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>50</td>
</tr>
</tbody>
</table>
Exercise I (7 points)

1. (½ point each) Indicate whether the following statements are True or False.

(a) A bimodal distribution is symmetric.  
(b) For two events A and B, if \( P(A) = P(A|B) \cdot P(B) \), then A and B are independent. 
(c) The function \( f(x) = \begin{cases} x, & \text{if } -1 \leq x \leq \sqrt{3} \\ 0, & \text{otherwise} \end{cases} \) is a valid probability density function. 
(d) If the points of a scatterplot fall exactly on a line, then the correlation of the two variables must be 1. 
(e) Suppose you are studying a difference in two means, and there really is an effect for what you are studying. If you increase the size of both samples, then the power must increase. 
(f) If you are exploring the mean in two populations and the samples you draw are the same size, then you should use a paired \( t \)-test. 
(g) Increasing the \( \alpha \) level for a test increases the Type I error rate. 
(h) In any process described by a Exponential model, the mean and variance will be equal. 
(i) If \( X \sim N(0, 1) \) and \( Y \sim N(0, 1) \) are independent standard normal random variables, then \( X + Y \sim N(0, 2) \). 
(j) The IQR is a more appropriate measure of spread than the standard deviation in skewed distributions.

2. (½ point each) Does sexual orientation affect how much people prefer a certain color? 
In 2001, researchers explored this question with thousands of college students. Suppose the 95% C.I. for \( P_{\text{LGBT male that likes pink}} - P_{\text{Straight male that likes pink}} \) was calculated as \((-0.03, 0.04)\). Indicate whether the following statements are True or False.

(a) At the level \( \alpha = 1\% \), there is no statistically significant difference in the percent of college-aged straight males and college-aged LGBT males who like pink.  
(b) We are 95% confident that the difference in the observed proportions is in the \((-0.03, 0.04)\). 
(c) The probability the true parameter difference lies in this interval is 0.95. 
(d) The 95% C.I. for difference in the other order \( P_{\text{Straight male that likes pink}} - P_{\text{LGBT male that likes pink}} \) is \((-0.04, 0.03)\).
Exercise II (7 points)

A psychology researcher wants to find out if exercising before taking a quiz affects a student's performance. To test this, he randomly assigns students to either exercise for 10 minutes before taking a short quiz, or to take the quiz without exercising first. 37 students exercise first and average 84% on the quiz with a standard deviation of 7% while 32 students skip exercising and score 81% with a standard deviation of 6%. Neither sample was substantially skewed.

Using $\alpha = 0.05$, conduct a hypothesis test to see if exercising before taking a quiz affects a student's performance. Make sure to: 1) state hypotheses, 2) draw a picture of the sampling distribution, 3) shade a $p$-value, 4) find the $p$-value (or bounds on it), and 5) make a decision about the hypotheses based on the $p$-value.

Let $M_T$ be the average scores of students who took the exercise.
Let $M_C$ be the average scores of students who did not take the exercise.

**Hypothesis:**

$H_0 : M_T - M_C = 0$

$H_A : M_T - M_C \neq 0$

**Sampling distribution:**

t-distribution with $df = \min(37-1, 32-1) = 31$.

and $SE = \frac{0.07^2 + 0.06^2}{37} \times 0.0157$

$T = \frac{(0.84 - 0.81) - 0}{0.0157} \approx 1.91$

$T, df = 31$

$P-value = P(|T| > 1.91)$ which is between 0.05 and 0.10 so, $p$-value > 0.05, we reject $H_0$ and conclude there is no statistically difference between taking or not taking the exercise.
Exercise III (7 points)

Suppose $X$ is a continuous random variable with density function

$$f(x) = \begin{cases} 
  c - x, & \text{if } -1/2 \leq x \leq 1/2, \\
  0, & \text{otherwise}
\end{cases}$$

where $c$ is some positive constant.

1. (1 point) What value must $c$ be for $f(x)$ to be a valid density function?

   \[\int_{-\infty}^{\infty} f(x) \, dx = 1\]

   \[\Rightarrow \int_{-\frac{1}{2}}^{\frac{1}{2}} c - x \, dx = c\frac{1}{2} x^2 \bigg|_{-\frac{1}{2}}^{\frac{1}{2}} = 1\]

   \[\Rightarrow \frac{1}{2} c - \frac{1}{2} \cdot \frac{1}{4} - (-\frac{1}{2} c - \frac{1}{2} \cdot \frac{1}{4}) = 1 \quad \Rightarrow \boxed{c = 1}\]

   \[\text{\circ} \quad f(x) > 0 \quad \text{is true for } c = 1\]

2. (2 points) Find the expected value and standard deviation of $X$.

   \[E(X) = \int_{-\infty}^{\infty} x f(x) \, dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x (1-x) \, dx\]

   \[= \int_{-\frac{1}{2}}^{\frac{1}{2}} x - x^2 \, dx = \frac{1}{2} x^2 - \frac{1}{3} x^3 \bigg|_{-\frac{1}{2}}^{\frac{1}{2}} = \boxed{-\frac{1}{12}}\]

   \[E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) \, dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x^2 (1-x) \, dx\]

   \[= \frac{1}{3} x^3 - \frac{1}{4} x^4 \bigg|_{-\frac{1}{2}}^{\frac{1}{2}} = \boxed{\frac{1}{12}}\]

   \[\text{Var}(X) = E(X^2) - (E(X))^2 = \frac{1}{12} - (-\frac{1}{12})^2 = \frac{11}{144}\]

   \[\text{S.D.} = \sqrt{\text{Var}(X)} = \sqrt{\frac{11}{144}} = \boxed{\frac{\sqrt{11}}{12} \approx 0.276}\]
3. (2 points) Compute $P(\mid X\mid < 1/3)$.

$$
P(\mid X\mid < \frac{1}{3}) = P(-\frac{1}{3} < X < \frac{1}{3}) = \int_{-\frac{1}{3}}^{\frac{1}{3}} f(x) \, dx
= \int_{-\frac{1}{3}}^{\frac{1}{3}} 1-x \, dx = \left[ x - \frac{1}{2} x^2 \right]_{-\frac{1}{3}}^{\frac{1}{3}}
= \left\{ \frac{2}{3} \right\}
$$

4. (2 points) Is the median of $X$ positive, negative, or equal to zero? (To answer this question, you can EITHER compute the median explicitly, OR give a qualitative argument)

Suppose $m$ is the median of $X$, then

$$
P(X \geq m) = P(X < m) = .50
$$

So, \( \int_m^{\infty} f(x) \, dx = 0.50 \)

\[ \Rightarrow \int_{-\frac{1}{2}}^{m} 1-x \, dx = 0.50 \]

\[ \Rightarrow \left[ x - \frac{1}{2} x^2 \right]_{-\frac{1}{2}}^{m} = 0.50 \]

\[ \Rightarrow m^2 - 2m - \frac{1}{4} = 0 \]

\[ m = 1 - \frac{\sqrt{5}}{2} \text{ or } 1 + \frac{\sqrt{5}}{2} \]

Since $f(x)$ is skewed to the right (from the plot), the median should be smaller than the mean ($= -\frac{1}{2}$ from part (2)), so, median must be negative $\Rightarrow m = 1 - \frac{\sqrt{5}}{2}$.
Exercise IV (5 points)

People often wonder if the quality of car parts really depends on their price. To find out, *Consumer Reports* listed the cost (in dollars) and power (in cold cranking amps) of auto batteries. We want to know if more expensive batteries are generally better in terms of starting power. Below is the scatterplot of the data:

1. (1 point) State the null and alternative hypotheses (in words) for checking for an association between the cost and the power of auto batteries.

   \[ H_0: \beta_1 = 0; \quad \text{where } \beta_1 \text{ is the slope of the fitted line for Cost and Power of auto batteries have no association.} \]

   \[ H_A: \beta_1 \neq 0; \quad \text{Cost and Power of auto batteries have some association.} \]

2. (1 point) What value should the true slope of the regression model have if no association exists? Answer with mathematical symbols.

   \[ \beta_1 = 0 \text{ if no association cause we will get a horizontal line.} \]

3. (3 points) For the sample of 30 auto batteries they studied, the authors report a slope of 4.14 and a standard error (for the slope) of 1.282. Create a 99% confidence interval for the slope of the true line.

   After this, decide (using your C.I.) if you should stick with the null hypothesis or move to the alternative hypothesis, at the level \( \alpha = 1\% \).

   \[ b_1 = 4.14, \quad SE_{b_1} = 1.282 \]

   \[ \text{CI: } b_1 \pm t_{n-2} SE_{b_1} \quad \text{For } n=30, \ 99\% \ C.I \text{ gives } t_{28} = 2.76 \]

   \[ \text{CI} = 4.14 \pm 2.76 \cdot (1.282) \approx (0.602, 7.678) \]

   Since 0 is not in this 99% CI, we conclude, there is some association between the cost and the power of auto batteries. (Move to the alternative hypothesis.)
Exercise V (7 points)

Suppose that patients arrive at an emergency room independently of one another at a constant rate of four per hour. Assume that on average, 20% of patients are children, and that the ages of different patients are independent of one another.

1. (2 points) Find the probability that exactly 3 of the next 8 patients who arrive will be children.

Let $X$ denote number of pa child patients. $X \sim \text{Bin}(8, 0.2)$

$$P(X = 3) = \binom{8}{3} \cdot 0.2^3 \cdot (1-0.2)^5 = 0.147$$

2. (2 points) Find the probability that exactly 4 patients arrive in the next half hour.

Let $Y$ be the number patients arrive

$Y \sim \text{Poisson}\left(\frac{4}{2}\right)$ since patient come at $\frac{4}{2}$/half hour

$$P(Y = 4) = \frac{e^{-2} \cdot 4^4}{4!} \approx 0.090$$

3. (3 points) Suppose you arrive in the emergency room at 2:00 PM. What is the probability that it will be at least 2:20 PM before the next patient arrives?

Let $Z$ be the time between two patients arrival.

$Z \sim \text{Exp}(4)$.

$$P(Z > \frac{20}{60}) = P(Z > \frac{1}{3}) = \int_{\frac{1}{3}}^{\infty} 4e^{-4x} \, dx$$

$$= -e^{-4x}\bigg|_{\frac{1}{3}}^{\infty} = \left[-e^{-\frac{4}{3}} \times 0.263\right]$$
Exercise VI (4 points)

You walk into a cheese shop looking to buy some cheese. The shop has cheeses from France and Switzerland. Two-thirds of the cheeses are French, while the rest are Swiss. Suppose 10% of the French cheeses are pasteurized, while 20% of the Swiss cheeses are pasteurized. If you randomly taste a cheese and it turns out to be pasteurized, what is the probability this cheese comes from Switzerland?

Let $F$ denote cheeses from France, $S$ denote cheeses from Swiss, $P$ denote cheeses are pasteurized.

Tree diagram.

```
all cheeses
\[ \begin{array}{c}
F \frac{2}{3} \leftarrow \\
S \frac{1}{3}
\end{array} \]
```

\[
\begin{align*}
\text{PC s | P} &= \frac{\text{PC s} \cdot \text{PC s}}{\text{PC P}} \\
&= \frac{\text{PC s} \cdot \text{PC s}}{\text{PC s} \cdot \text{PC s} + \text{PC P}} \\
&= \frac{20\% \cdot \frac{1}{3}}{20\% \cdot \frac{1}{3} + 10\% \cdot \frac{2}{3}} = \left[ \frac{1}{2} \right]
\end{align*}
\]
Exercise VII (4 points)

Suppose SAT scores are normally distributed with a mean of 1000 (out of 1600) and a standard deviation of 200.

1. (2 points) Suppose the SAT offers certificates of excellence to any student with a score above 1250. If you know that a randomly-chosen student has been awarded a certificate of excellence, what is the probability the student has a score above 1400?

\[
P(\text{Score} > 1400 | \text{Certificate of Excellence}) = \frac{P(\text{Score} > 1400)}{P(\text{Score} > 1250)}
\]

\[
\text{Score} \sim N(1000, 200)
\]

\[
= \frac{P(Z > \frac{1400 - 1000}{200})}{P(Z > \frac{1250 - 1000}{200})}
\]

\[
= \frac{1 - 0.9972}{1 - 0.8944} \approx 0.216
\]

2. (2 points) You decide to pick random students one-at-a-time until you finally encounter a student with an SAT score below 700. On average, how many students must you pick before finding such a student?

\[
\text{Prob of finding a student with score} < 700 = P(\text{score} < 700)
\]

\[
= P\left(Z < \frac{700 - 1000}{200}\right) = P(Z < -1.5) = 1 - 0.9332 = 0.0668
\]

The event of picking a student with score < 700 is a Bernoulli trial with a success probability \(p = 0.0668\), \(X \sim \text{Bin}(p)\). Here, \(X\) is the random variable for the number of Bernoulli trials for the first success while picking students at random. The expected value is \(E(X) = \frac{1}{p} \approx 14.97\)

On average, 14.97 students must be picked before finding a student with score < 700.
Exercise VIII (2 points)

Below is a comic from xkcd.com by Randall Munroe. Discuss the statistical ideas brought forward by the comic and why the comic is funny.

originally, the test was to find a link between jelly beans and acne with \( \alpha = 0.05 \). Consequently, there were 20 tests conducted for each color of the jelly beans. We shouldn't use the same \( \alpha \) for each of the tests for colored jelly beans because if \( X \) is the number of reasons we do in our 20 tests, the prob of error

\[
P(X > 1) = 1 - P(X = 0) = 1 - (1 - \alpha)^{20} = 1 - (1 - 0.05)^{20} = 0.6415
\]

we should use the **Benferroni correction** for the value of \( \alpha \), i.e., \( \alpha' = \frac{\alpha}{20} = 0.0025 \). With the correction the probability of error should decrease

\[
P(X > 1) = 1 - (1 - \alpha')^{20} = 1 - (1 - 0.0025)^{20} = 0.049
\]
Exercise IX (7 points)

Your sporty friend claims that surfing increases people's self-reported happiness scores. You are curious if this is true and decide to conduct a study. You plan to ask $n$ random people who don't surf how happy they are (on a 0 - 10 scale). You read online that the average happiness level of surfers is 7.3 with a standard deviation of 0.9. Assume this standard deviation is the same for people who don't surf. If you want your study to have $\alpha = 0.10$ and a power of 0.85, what is the smallest size you could use for $n$ if the smallest difference that interests you is 0.1 point on the happiness scale? Include a picture with your solution.

Let $\mu$ be the avg happiness level of non-surfers.

$H_0: \mu = 7.3$

$H_a: \mu < 7.3$

Computing the no of SEs below 7.3 we have the rejection fence

$P(Z < Z_{0.1}) = 0.1$

$Z_{0.1} = -1.28$ (from the 2 tail)

Power = 0.85

$1.28 \text{ SE below 7.3 is where we have the rejection fence.}$

Taking the power (0.85) & the alt. hypothesis into account, we compute the distance between 7.2 & the rejection fence.

$P(Z < Z_{0.85}) = 0.85; \quad \therefore \quad Z_{0.85} = 1.04$ (from Z table)

$1.04 \text{ SE above 7.2 is where we have the rejection fence.}$

$1.28 \text{ SE} + 1.04 \text{ SE} = 0.1$ (as per the diff of interest)

$SE = 0.1 / 2.32 \approx 0.04310$

$0.9 / 0.04310 \equiv n \approx 43.6044$

$\therefore n = 437$

The smallest size we could use for $n$ is 437.