FINAL EXAM

By taking this exam, you pledge that you will not receive unacknowledged assistance from any source other than myself, the Professor of the course, in completion of the exam.

Please sign here:

**I. Multiple choice questions (40 points):** For each of the following 10 questions, circle one choice that is most suitable to the situation given in the corresponding question *and explain why you think it is the correct answer.*

1. You want to demonstrate that Americans work harder than Europeans, who work 32 hours a week on average. In a sample of 100 Americans, the average weekly hours is 38 hours, with a standard deviation of 12 hours.

a. Null hypothesis: the average American in sample work 38 hours a week or fewer.

b. Null hypothesis: the average American work 32 hours a week or fewer.

c. Null hypothesis: the average American in sample work 32 hours a week or fewer.

d. Null hypothesis: the average American work 38 hours a week or fewer.

2. A sociologist want to demonstrate that there are more females than males in a community based on a random sample of individuals from the community.

a. Under the null hypothesis, the gender of each individual in the sample can be described as a flip of a fair coin.

b. Under the null hypothesis, fraction of female in the population is a nonrandom, unknown figure.

c. The gender of each individual in the sample can be described as a flip of a fair coin.

d. The gender ratio in the population is 50-50.

3. (same experiment as in question 2) A sociologist want to demonstrate that there are more female than male in a community based on a random sample of individuals from the community. In the sample of 200 individuals, 93 are female.

a. P-value: the probability that 93 or fewer of a random sample of 200 turn out to be female if in fact there are no more female than male in the population

b. P-value: the probability that 93 or more of a random sample of 200 turn out to be female if in fact there are no more female than male in the population.

c. P-value: the probability that the fraction of female in the population is as low as 93/200 if in fact there are more female than male.

d. P-value: the probability that there are no more female than male in the population if in fact 93 or more of a random sample of 200 turn out to be female.

4. A pollster want to demonstrate that people in Saigon are more concerned by economic condition than the national rate of 65% by conducting a survey on a random sample of 250 Saigon residents. Hint: think of the % as a unit, not a proportion.

a. The number of people out of the 250 sampled who are concerned by economic condition is a Bernoulli trial for which probability of each success is unknown.

b. Under the null hypothesis, the number of 250 sampled who are concerned by economic condition is a normal distribution for which mean is unknown.

c. The number of people out of the 250 sampled who are concerned by economic condition is a binomial distribution for which probability of each success is unknown.

d. Under the null hypothesis, the number of 250 sampled who are concerned by economic condition is a normal distribution for which mean is known.

5. On average, people spend 20% of their wake time on their electronic devices. You want to demonstrate that teens spend more time than others by randomly sampling 400 teens and recording the amount of time each spends on his/her electronic device.

a. Under the null hypothesis, the average amount of time that teens in sample spend on their devices is approximately a normal distribution with unknown mean.

b. The average amount of time that teens in sample spend on their device is approximately a normal distribution with unknown mean.

c. Under the null hypothesis, the amount of time that a teen in sample spends on his/her devices is approximately a normal distribution with known mean

d. The average amount of time that teens in sample spend on their devices is approximately a normal distribution with known mean.

6. (same experiment as in question 5) On average, people spend 20% of their wake time on their electronic device. You want to demonstrate that teens spend more time than others by randomly sampling 400 teens and recording the amount of time each spends on his/her electronic device. In the sample, the average time spent is 25%; the standard deviation is 10%.

a. p-value = approximately the probability that a normal random variable with mean = 20% and standard deviation = 0.5% results in an outcome of 25% or higher.

b. p-value = the probability that 25% of 400 teens in sample spend more time on time with their devices than average if in fact teens spend the same time as others on electronic devices.

c. p-value = approximately the probability that a normal random variable with mean = 20% and standard deviation = 10% results in an outcome of 25% or higher.

d. p-value = approximately the probability that a normal random variable with mean = 25% and standard deviation = 0.5% results in an outcome of 20% or lower.

7. The average amount of calories that 7-year-old children consume per day is 2500. You want to demonstrate that the average is lower than this for 7-year-olds whose parents live in poverty by observing a random sample of 144 7-year-old from poor background. In your sample, the average is 2400 and the standard deviation is 800.

a. Under the null hypothesis, each sampled 7-year-old’s daily calorie consumption is a continuous random variable with mean equals 2500.

b. Under the null hypothesis, each sampled 7-year-old’s daily calorie consumption is a continuous random variable with standard deviation equals 800/12.

c. Under the null hypothesis, each sampled 7-year-old’s daily calorie consumption is a normal random variable with mean equals 2500.

d. Under the null hypothesis, the average daily calorie consumption of calorie consumption of 7-year-old is a continuous random variable with mean equals 2400.

8. (same experiment as in question 7) The average amount of calories that 7-year-old children consume per day is 2500. You want to demonstrate that the average is lower than this for poor 7-year-olds by observing a random sample of 144 7-year-old from poor background. In your sample, the average is 2400 and the standard deviation is 800.

a. p-value = the probability that a normal random variable with mean = 2400 and standard deviation = 800 results in an outcome of 2500 or higher.

b. p-value = the probability that we’d obtain a sample of 144 individuals with mean as high as 2500 if in fact the population mean is 2400.

c. p-value = the probability that a normal random variable with mean = 2400 and standard deviation = 800/12 results in an outcome of 2500 or lower.

d. p-value = the probability that we’d obtain a sample of 144 individuals with mean as low as 2400 if in fact the population mean is 2500.

**II. Short answer questions (40 points)**

1. It is reported that in a country, 80% of all males think that poor people are lazy. Among females, only 30% think that poor people are lazy. Exactly half of people in the population think that poor people are lazy. What is the probability that a randomly selected person is male conditional on he/she believes that poor people are lazy?
2. In a population, the distribution of chocolate consumption per year (in oz) is a uniform distribution between 0 and 200.
3. If a person is randomly drawn from the population, calculate the probability that his chocolate consumption is between 20 and 60 oz.
4. Calculate the probability that a random person consumes less than 80 oz. conditional on him consuming more than 30 oz.
5. An independent random sample of 400 individuals is drawn from the population. Show the steps you’d take to calculate the probability that, on average, they consume less than 45 oz.

A picture always helps. Here are some useful formulas:

|  |  |
| --- | --- |
| **Notation** | \mathcal{U}(a, b) or \mathrm{unif}(a,b) |
| **Parameters** | -\infty < a < b < \infty \, |
| [**Sample**](https://en.wikipedia.org/wiki/Support_(mathematics)) **space** | x \in [a,b] |
| [**PDF**](https://en.wikipedia.org/wiki/Probability_density_function) | \begin{cases}                   \frac{1}{b - a} & \text{for } x \in [a,b]  \\                   0               & \text{otherwise}                 \end{cases} |
| [**CDF**](https://en.wikipedia.org/wiki/Cumulative_distribution_function)  **F(x)** | \begin{cases}                   0               & \text{for } x < a \\                   \frac{x-a}{b-a} & \text{for } x \in [a,b) \\                   1               & \text{for } x \ge b                 \end{cases} |
| [**Mean**](https://en.wikipedia.org/wiki/Expected_value) | \tfrac{1}{2}(a+b) |
| [**Variance**](https://en.wikipedia.org/wiki/Variance) | \tfrac{1}{12}(b-a)^2 |

1. To demonstrate that Unicorn’s IQ scores are, on average, higher than human average of 110, you randomly sample 64 unicorns and measure their IQ (note that 64 = 8^2). In your sample, the average IQ score is 115 and the standard deviation is 20. Complete the 4 steps of hypothesis testing to make a conclusion based on your sample:
2. State the null and alternative hypotheses.
3. Describe the test statistics based on the null hypothesis.
4. Briefly describe the sample outcome.
5. Define in your own words the p-value. Graphically describe how you’d calculate the p-value (as in it’s a probability that what happens with respect to what random variable) and briefly write down the command that you’d calculate it with a computer program. Suppose that the p-value turns out to be less than 5%. State your conclusion at the statistical significance of 5%.

**Part III: Data analysis (20 pts)**

This part of the exam is based on the data set house-price, attached here: <https://drive.google.com/open?id=0B5mG_WzHwxj8a3pTOTBQb0pnTlE>

1. Identify the zip code (under variable Zip of the data) that is associated with the highest house price in the data set. To do this you will need to create appropriate statistics to compare house prices between different zip codes in the data set. You can further clarify your findings by creating a table or a graph that summarizes your comparison results.
2. After identifying the zip code with the highest house price in the data set, you want to conduct an analysis to explain whether the higher price level in that zip code is associated with larger size, the number of bathrooms, the number of garages, or other factors we do not observe in the data. This analysis is important because it helps homebuyers to correctly determine prices of their purchases based on the aforementioned characteristics.

To conduct such an analysis, run the appropriate regression for home price separately for the zip code you identified in question 1, and another, similar regression for other districts. Compare your finding in words to answer the question.

1. Assume that the data you work with is obtained from a random sample; the regression results that Excel created for you is calculated based entirely on this data. This question requires thinking: Are the regression results non-random numbers, random variables, or outcomes of random variables? Based on your understanding, do the answers you obtained from questions 1 and 2 also apply for the true house prices in the city where this data come from? Why or why not?