CSC 315, Fall 2022

Exam II Format

- Exam II will be a written exam
- You will <u>not</u> have access to a computer during the exam
- You may bring one page of notes, front and back, any size font
- The formula sheet will be provided for you
- Question format:
    - You may be asked to interpret output from *R* code, such as results from a hypothesis test,
    - You may be asked to add to *R* code (such as adding code to extract or calculate a *p*-value)
    - You may be asked to write *R* code, as required on previous Labs 4 – 7, *but*
        - You will <u>not</u> be asked to write R code for generating any graphs, but
        - You may be asked to interpret an *R* graph that is provided to you
    - You are responsible for all material in Lab #5

Exam II Outline

1. Empirical probability

    a. The empirical probability of an event *E* is the long run proportion of times that the event occurs

    $$\mathrm{P}(E) = \frac{\#\ of\ times\ event\ occurs}{\#\ of\ trials\ or\ experiments}$$

    b. Can be calculated in *R* using the *replicate* function to replicate an event (implemented in a function) such as flipping a coin.

2. Classical probability

    a. Formula: classical probability of an event *E* =

    $$\mathrm{P}(E) = \frac{\#\ of\ ways\ in\ which\ E\ occurs}{\#\ of\ possible\ outcomes\ in\ the\ sample\ space}$$

    when all possible outcomes are equally likely

   b. Can be calculated in *R* using the *permutations* function (when order matters) from the library *gtools* or the *combinations* function (when order does not matter).

3. Probability distributions

   a. The probability that a random variable X is less than the value k can be calculated based on classical probability and is

$$\mathrm{P}(X < k) = \frac{\#\ of\ oservations < k}{total\ \#\ of\ observations\ (sample\ size)}$$

   b. This probability is equivalent to the sum of histogram densities for the bars corresponding to $X < k$.

   c. This probability is equivalent to the area under the curve between $-\infty$ and k for a probability density function.

   d. For any probability density function, the area under the curve between points a and b is equal to $\mathrm{P}(a < X < b)$.

4. The normal probability distribution

   a. $\mathrm{X} \sim \mathrm{N}(\mu, \sigma)$ if it is a unimodal, symmetric, bell-shaped distribution with mean μ and standard deviation σ

   b. Empirical rule: for a normal distribution, approximately 68%, 95% and 99% of observations are within 1, 2, and 3 standard deviations of the mean.

   c. The standard normal distribution is a normal distribution with $\mu = 0$ and $\sigma = 1$.

   d. *R* functions:

      i. *pnorm* –calculates probabilities of the form $\mathrm{P}(X < k)$, equivalent to the area under the curve to the *left* of *k*

      ii. *dnorm* – calculates the probability density (e.g., if plotting the curve)

      iii. *qnorm* – calculates percentiles of the normal distribution

      iv. *rnorm* – random number generation

5. Sampling distribution of the sample mean

    a. The sample mean is a random variable!

    b. Central Limit Theorem: If a distribution X has mean μ and standard deviation σ then the sample distribution of the sample mean $\bar{X}_n$ from a sample of size n has mean μ and standard deviation $\sigma/\sqrt{n}$. Furthermore, the distribution of $\bar{X}_n$ is normal if X is normally distributed, and approximately normal for other distributions if $n >$ 30. The larger the sample size, the closer the distribution is to normality.

6. Hypothesis testing based on a population proportion, a population mean, the difference between population means, or difference between two proportions.

    a. State the null and alternative hypotheses

    b. Specify the distribution of the sample statistic (i.e., a sample proportion or sample mean) under the null hypothesis.

    c. Calculate/find the test statistic (using the *prop.test* or *t.test* functions), or manually using the appropriate formula

    d. Find the p-value

        i. From *prop.test* or *t.test*

        ii. Based on a z- or t- test statistic and using the *pnorm* or *pt* function with appropriate degrees of freedom.

    e. State the conclusion regarding the null hypothesis in the context of the problem, and justify your conclusion based on the p-value.

    f. Interpretation of Type I or Type II errors

Example questions (also see Exam II Practice script)

1. The sample space from flipping a fair coin 3 times is given by the R code below:

```
> permutations(2, 3, c('H','T'), repeats.allowed = TRUE)

     [,1] [,2] [,3]
[1,] "H"  "H"  "H"
[2,] "H"  "H"  "T"
[3,] "H"  "T"  "H"
[4,] "H"  "T"  "T"
[5,] "T"  "H"  "H"
[6,] "T"  "H"  "T"
[7,] "T"  "T"  "H"
[8,] "T"  "T"  "T"
```

Here we will find the probability that the first **or** second coin toss is Heads.

(a) Circle the outcomes where you get heads on the first or second coin toss.

(b) Using the classical definition of probability, write as a fraction the probability that when you flip a coin 3 times, you get heads on the first or second toss.

2. According to data from the CDC, as of 10/28/2021, for individuals 65 and over, 1968 out of 100,000 have been hospitalized with COVID-19; for individuals 18-49, 1050 out of 100,000 have been hospitalized (as of 10/28/2021) Source: https://gis.cdc.gov/grasp/COVIDNet/COVID19_3.html.

The null and alternative hypotheses are given by:

H0: *p_old – p_young* = 0
H1: *p_old – p_young* != 0

Where *p_old* is the proportion of individuals 65 or over who have been hospitalized with COVID-19, and *p_young* is the proportion of 18-49 year olds who have been hospitalized (as of 10/28/2021).

(a) Complete the *prop.test* code that tests against the null hypothesis that there is no difference across age groups in the proportion of individuals who have been hospitalized with COVID-19.

```
prop.test(________________________________________________)
```

(b) Your p-value should be very close to 0 (1.758227e-63). Based on this p-value, state the conclusion regarding the null and alternative hypotheses in the context of this problem.

3. Is the mean body temperature really 98.6 degrees Fahrenheit? Suppose that body temperatures from 109 randomly selected healthy indviduals is stored in the vector *temps* (you can assume that there are no missing values).

(a) Write the R code to calculate the relevant test statistic that would be used to test against the null hypothesis that the mean body temperature is 98.6 degreees Fahrenheit.

(b) Under the null hypothesis, the test statistic in (a) would follow the Student's T distribution with how many degrees of freedom?