Homework 7

PSTAT 5A: Spring 2023, with Ethan P. Marzban

Instructions

- Please submit your work to Gradescope by no later than **11:59pm on MONDAY, May 22**. As a reminder, late homework will not be accepted.
- Recall that you will be asked to upload a **single** PDF containing your work for *both* the programming and non-programming questions to Gradescope.
  - You can merge PDF files using either Adobe Acrobat, or using adobe's online PDF merger at this link.

Problem 1: Deriving the Lower-Tailed Hypothesis Test

Consider testing the set of hypothesis

$$\left[\begin{array}{rl} H_0: & p = p_0 \\ H_A: & p < p_0 \end{array}\right.$$

at an arbitrary α level of significance. Define the test statistic TS to be

$$\mathrm{TS} = \frac{\widehat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

a. Show that $\mathrm{TS} \stackrel{H_0}{\sim} \mathcal{N}(0,\ 1)$. If your answer depends on a set of conditions to be true, explicitly state those conditions.

b. Argue, in words, that the test should be of the form

$$\mathtt{decision}(\mathrm{TS}) = \begin{cases} \mathtt{reject}\ H_0 & \text{if } \mathrm{TS} < c \\ \mathtt{fail\ to\ reject}\ H_0 & \text{otherwise} \end{cases}$$

for some constant c. As a hint, look up the logic we used in Lecture 13 to derive the two-tailed test, and think in terms of statements like "$\hat{p}$ is *far away* from p_0". **You do not have to find the value of c in this part**.

c. Now, argue that c must be the α^{th} percentile of the distribution of the standard normal distribution (**not** scaled by negative 1), thereby showing that the full test takes the form

$$\mathtt{decision}(\mathrm{TS}) = \begin{cases} \mathtt{reject}\ H_0 & \text{if } \mathrm{TS} < z_\alpha \\ \mathtt{fail\ to\ reject}\ H_0 & \text{otherwise} \end{cases}$$

where z_α denotes the $(\alpha) \times 100^{\text{th}}$ percentile of the standard normal distribution.

❗ Result: Upper-Tailed Test

When testing the hypotheses

$$\left[\begin{array}{ll} H_0: & p = p_0 \\ H_A: & p > p_0 \end{array}\right.$$

at an α level of significance, the test takes the form

$$\texttt{decision}(\mathrm{TS}) = \begin{cases} \texttt{reject } H_0 & \text{if } \mathrm{TS} > z_{1-\alpha} \\ \texttt{fail to reject } H_0 & \text{otherwise} \end{cases}$$

where:

- $\mathrm{TS} = \dfrac{\widehat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
- $z_{1-\alpha}$ denotes the $(1-\alpha) \times 100^{\text{th}}$ percentile of the standard normal distribution.

provided that

- $np_0 \geq 10$
- $n(1-p_0) \geq 10$

Problem 2: Airplanes (not in the Night Sky)

According to *USAToday*, around 2.75% of flights in 2022 were cancelled. To test this claim, Jaime collects data on a representative sample of 500 flights from 2022 and finds that only 2.01% of these flights were cancelled. Assume that Jaime wishes to perform a two-sided test, at an $\alpha = 0.05$ level of significance.

a. What is the population?
b. What is the sample?
c. Write down the null and alternative hypotheses for this problem. Use mathematical notation.
d. Compute the value of the test statistic.
e. Compute the critical value of the test.
f. Conduct the test, and phrase your conclusions in the context of the problem.

Problem 3: Airplanes (still not in the Night Sky)

Consider again the setup of Problem 2, except now suppose Jaime wishes to conduct an upper-tailed test (still at an $\alpha = 0.05$ level of significance).

a. Does the value of the test statistic change from what you found in Problem 2(d)? If so, provide the new value.
b. Does the critical value change from what you found in Problem 2(e)? If so, provide the new value.

c. Conduct the test, and phrase your conclusions in the context of the problem.

Problem 4: Watch The Time

In a 2015 article, *CBC News* predicted that in 2018 31% of people would wear a watch. Suppose a representative sample of 204 people, taken in 2018, contained 65 people that wore a watch.

a. Assuming *CBC*'s claim is correct, what is the probability that a representative sample (assume it was taken with replacement) contained 65 people that wore a watch? **State your logic clearly, and check all assumptions that may need to be checked.**

b. Assuming *CBC*'s prediction was correct, what is the expected number of people who would be wearing a watch in a sample of 204 people (again, assume the sample was taken with replacement)?

c. Assuming *CBC*'s prediction was correct, what is the variance of the number of people who would be wearing a watch in a sample of 204 people (again, assume the sample was taken with replacement)?

d. Assuming *CBC*'s prediction was correct, what is the probability that between 27.8% and 37.5% of people in a sample of size 204, taken with replacement, wear a watch?

e. Now, assume we wish to test *CBC*'s prediction against the two-sided alternative that the true proportion of people that wore a watch in 2018 was not equal to 31%. State the null and alternative hypotheses for this test in mathematical terms.

f. Conduct a test of the two hypotheses you formulated in part (e) above, using an $\alpha = 0.01$ level of significance.

Problem 5: Random Variables

Let X be a random variable with probability mass function

k	-3.1	0	0.7	1.2
$\mathbb{P}(X = k)$	a	0.19	0.21	0.48

a. What is the value of a?

b. Compute $\mathbb{P}(\{X = -3\} \cup \{X = 0.7\})$.

c. Compute $\mathbb{P}(X \leq 1)$.

d. Compute $\mathbb{E}[X]$, the expected value of X.

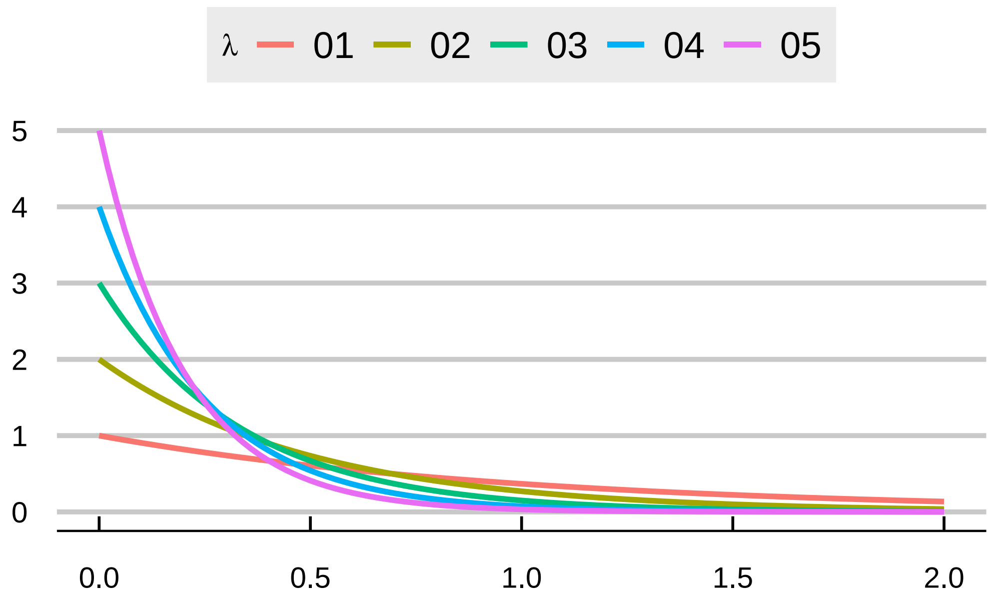
e. Compute $\text{SD}(X)$, the standard deviation of X.

Problem 6: Programming: The Exponential Distribution

Another continuous distribution that we haven't discussed thus far is the so-called **Exponential distribution**. It takes a single parameter, called the *rate* parameter (denoted λ) and has probability density function (p.d.f.):

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

We use the notation $X \sim \text{Exp}(\lambda)$ to denote the fact that a random variable X follows the Exponential distribution with parameter λ. The density curves of the $\text{Exp}(\lambda)$ distribution look like:



The Exponential distribution is often used for modeling lifetimes; e.g. the lifetime of a lightbulb, etc. It turns out that there is a nice closed-form expression for the area underneath a portion of an Exponential density curve: if $X \sim \text{Exp}(\lambda)$, then

$$\mathbb{P}(a \leq X \leq b) = e^{-a \cdot \lambda} - e^{-b \cdot \lambda}$$

assuming $0 < a < b < \infty$. For example, if $X \sim \text{Exp}(1)$, then $\mathbb{P}(1 \leq X \leq 2) = e^{-1 \cdot 1} - e^{-2 \cdot 1} = e^{-1} - e^{-2} \approx 0.2325$.

❗ Task 1

Write a function called `d_exp()` that takes in two arguments, `x` and `lam`, and returns the value of the p.d.f. of the Exp(`lam`) distribution at the point `x`. Your function should:

- have a default `lam` value of 1
- return zero for any negative values of `x`

Check that your function behaves as follows:

```
d_exp(3.5, 2.31) # specify both arguments
```

```
0.00071177231822478
```

```python
d_exp(3.5) # use default lam value
```

```
0.0301973834223185
```

```python
d_exp(-2, 4) # return, due to negative input
```

```
0
```

❗ Task 2

Write a function called `p_exp()` that takes in three arguments: , `a`, `b`, and `lam`, and returns the probability that an Exp(`lam`)-distributed random variable lies between `a` and `b`. Set `lam` to have a default value of `1`. **Think very carefully about any cases you might need to consider!** (You may assume that `a` is always less than `b`.)

Check that your function behaves as follows:

```python
p_exp(1, 2, 1) # specify all three arguments
```

```
0.23254415793482963
```

```python
p_exp(1, 2) # use default lam value
```

```
0.23254415793482963
```

```python
p_exp(-1, 2) # specify negative `a` value
```

```
0.8646647167633873
```

NOTE: One quirk of python is that, when defining a function with multiple arguments, only *some* of which have default values, you must place the arguments with default values *after* those that do not. I think you will see what I mean when you try to define your `p_exp()` function above!