

**INSTRUCTIONS**

You have 1 hour and 50 minutes to complete the exam.

- The exam is closed book, closed notes, closed computer, closed calculator, except the provided midterm study guide.
- Mark your answers on the exam itself in the spaces provided. We will not grade answers written on scratch paper or outside the designated answer spaces.
- If you need to use the restroom, bring your phone and exam to the front of the room.

For questions with **circular bubbles**, you should select exactly *one* choice.

- ☐ You must choose either this option
- ☐ Or this one, but not both!

For questions with **square checkboxes**, you may select *multiple* choices.

- ☐ You could select this choice.
- ☐ You could select this one too!

**Preliminaries**

You can complete and submit these questions before the exam starts.

- (a) What is your full name?

- (b) What is your student ID number?

- (c) Who is your lab GSI? You may write *Unknown* if you don't know their name.

- (d) Sign here to confirm that all work on this exam is your own (or type your name if online).

# 1. (42.0 points) BeReal

BeReal is an app that prompts users to post one photo per day. Once prompted, users have 2 minutes to post a photo; after 2 minutes they are late. A user can only post one photo per day, but they can retake the photo before posting. Other users can comment on posts that day. Assume that all comments on a post are made on the same day as the post.

The `posts` table contains one row for each of 528 posts made by Data 8 students. Columns exist for the `user` (`str`) who posted, the `date` (`str`) of their post, the `caption` (`str`) of the posted photo, whether the photo was a `retake` (`bool`), and how many minutes the post was `late` (`float`). The first three rows are:

user	date	caption	retake	late
Wendy	10/13	sunny sun	False	0.0
Wendy	10/14	prof sahai	True	90.2
Nicole	10/13	on the glade	False	35.5

The `comments` table contains one row for each of 1240 comments made on these posts. Columns exist for the `author` (`str`) of the comment, the `comment` (`str`), the date of the `post` (`str`), and the `poster` (`str`) who is the user that posted the photo being commented on. The first three rows are:

author	comment	post	poster
Will	midterm season	10/13	Nicole
Wendy	cute dogs!	10/13	Nicole
Will	prof sighting	10/14	Wendy

## (a) (4.0 points)

This partially completed expression evaluates to the number of posts for which Wendy retook her photo before posting.

```

----- (posts.-----.-----('retake'))
(a)          (b)          (c)

```

i. (1.0 pt) Fill in blank (a).

`sum` or `np.sum` or `np.count_nonzero`

ii. (2.0 pt) Fill in blank (b).

`where('user', 'Wendy')` or `where('user', are.equal_to('Wendy'))`

iii. (1.0 pt) Which of these could fill in blank (c)?

- ☒ `column`
- ☐ `select`
- ☐ `group`
- ☐ `take`

## (b) (9.0 points)

Complete this code to compute two quantities about pictures that aren't retakes:

- `real_fractions` is an **array** with one item for each user that contains the **fraction** (a float) of posts made by that user that are **not retakes**. *The array may be in any order.*
- `good_post_fraction` is the **fraction** (a float) of all posts that have a caption **longer than 5 characters** and a photo that is **not a retake**. Use the `good_post` function to compute it.

**Hints:** The average of an array of `bool` values is the fraction that are `True`. The sum of an array of `bool` values is the number that are `True`.

```
real_fractions = _____ posts.select('user', 'retake')._____.column(_____)
                  (a)                                     (b)          (c)
```

```
def good_post(caption, retake):
    return len(caption) > 5 and _____
                                   (d)
```

```
good_post_fraction = sum(posts._____) / _____
                      (e)          (f)
```

- i. (1.0 pt) Fill in blank (a). If you believe no code is necessary in this blank, write “EMPTY”.

1-

- ii. (2.0 pt) Fill in blank (b). You may not include any `.` in your answer.

`group('user', np.average)`

- iii. (1.0 pt) Fill in blank (c).

`1 or 'retake average' or 'retake mean'`

- iv. (1.0 pt) Fill in blank (d).

`not retake or retake == False`

- v. (3.0 pt) Fill in blank (e). You may not include any `.` in your answer.

`apply(good_post, 'caption', 'retake')`

- vi. (1.0 pt) Fill in blank (f).

`posts.num_rows`

## (c) (7.0 points)

This partially completed code assigns `sahai_comments` to the total **number of comments** (an integer) made on all of Wendy's posts for which the caption includes the string `sahai` within it.

**Important:** Assume all captions are lowercased already.

**Reminder:** The `posts` table has one row per post and columns `user`, `date`, `caption`, `retake`, and `late`. The `comments` table has one row per comment and columns `author`, `comment`, `post`, and `poster`.

```
sahai_posts = posts.where('user', are.equal_to('Wendy')).where('caption', _____)
                                                    (a)
```

```
sahai_comments = sahai_posts._____(_____, _____, _____).num_rows
                                (b)      (c)      (d)      (e)
```

i. (2.0 pt) Fill in blank (a).

```
are.containing('sahai')
```

ii. (1.0 pt) Which of these could fill in blank (b)?

- ☐ apply
- ☐ select
- ☒ join
- ☐ pivot

iii. (1.0 pt) Fill in blank (c).

```
date
```

iv. (2.0 pt) Fill in blank (d).

```
comments.where('poster', 'Wendy')      or      comments.where('poster',
are.equal_to('Wendy'))
```

v. (1.0 pt) Fill in blank (e).

```
post
```

## (d) (14.0 points)

Write an expression that correctly computes each of the following results. You may abbreviate `posts` as `p` and `comments` as `c`.

`p = posts`  
`c = comments`

**Reminder:** The `posts` table has one row per post and columns `user`, `date`, `caption`, `retake`, and `late`. The `comments` table has one row per comment and columns `author`, `comment`, `post`, and `poster`.

- i. (2.0 pt) The largest **number** (a float) in the `late` column of the `posts` table.

```
max(posts.column('late'))
```

- ii. (2.0 pt) The **number** (an integer) of different authors who commented on one or more post.

```
comments.group('author').num_rows
```

- iii. (2.0 pt) An **array** of the captions of all posts with a positive (non-zero) `late` value. The array can be in any order.

```
posts.where('late', are.above(0)).column('caption')
```

- iv. (3.0 pt) A **table** with one row per user and three columns: the `user`, a column labeled `False` containing the average number of minutes that their non-retake posts were late, and a column labeled `True` containing the average number of minutes that their retake posts were late.

```
posts.pivot('retake', 'user', 'late', np.average)
```

- v. (3.0 pt) The largest **number** (an integer) of comments on any one post. A user can post at most once per day, but there is no limit to the number of times that someone can comment on a post.

```
max(comments.group(['post', 'poster']).column(2))
```

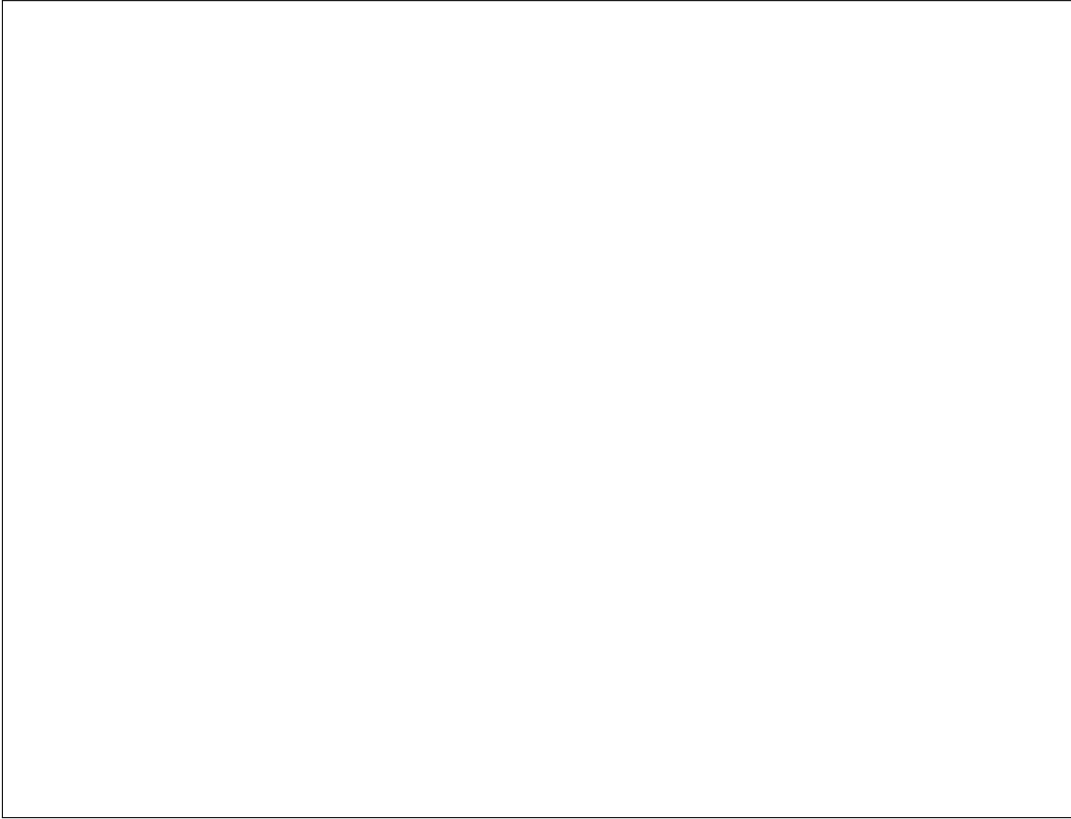
- vi. (2.0 pt) A histogram of the `late` values for the first 20 rows of the `posts` table. You do not need to specify bins or units.

```
posts.take(np.arange(20)).hist('late')
```

## (e) (8.0 points)

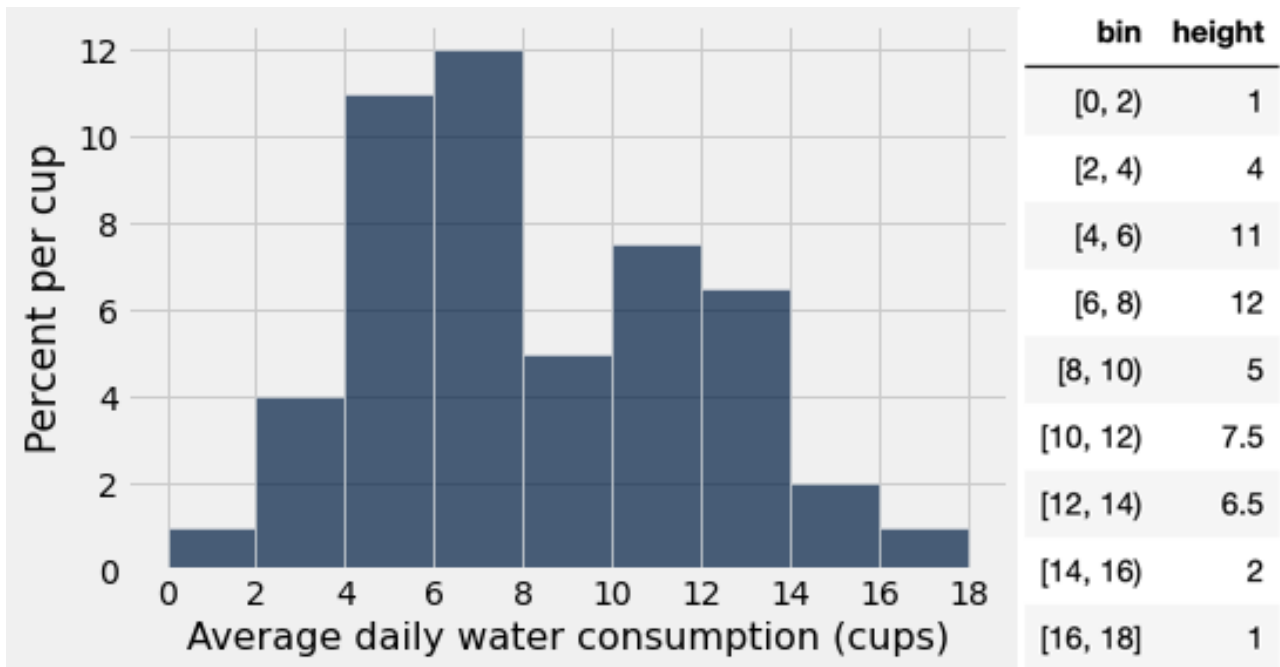
- i. (2.0 pt) What is the best data visualization to compare the distribution of **late** values for posts that are retakes to the distribution of **late** values for posts that are not retakes.
- ☐ Scatter plot
  - ☐ Bar chart
  - ☐ Line plot
  - ☒ Overlaid histograms
- ii. (2.0 pt) Suppose this dataset is a simple random sample of all BeReal posts made by Data 8 students. What null hypothesis would help assess whether, among all Data 8 students' posts, the posts that are retakes have a different distribution of **late** values than the posts that are not retakes? (*Note:* A post is late if it has a non-zero **late** value.)
- ☐ The average **late** value is different for posts that are retakes than posts that are not retakes.
  - ☐ The proportion of **retake** values that are **True** is different for posts that are late than for posts that are not late.
  - ☒ The distribution of **late** values among all Data 8 students is the same for posts that are retakes and posts that are not retakes.
  - ☐ The distribution of **retake** values among all Data 8 students is the same for posts that are late and posts that are not late.
- iii. (2.0 pt) For a permutation test that shuffles the order of **retake** values many times to simulate data under this null hypothesis, what is the purpose of shuffling the order of the column?
- ☒ Shuffling the **retake** values simulates drawing two samples of **late** values from the same distribution.
  - ☐ Shuffling is a form of randomized controlled experiment.
  - ☐ Shuffling is equivalent to taking a random sample with replacement.
  - ☐ Shuffling randomly assigns **retake** conditions to users before they post.
- iv. (2.0 pt) Suppose the absolute difference of means is used as the test statistic, 100,000 simulations are generated under the null hypothesis, and the resulting p-value is zero. What can we conclude from this test? Choose all that apply.
- ☐ For all BeReal users, the distribution of **late** values is different for posts that are retakes and posts that are not retakes.
  - ☒ For all Data 8 students, the distribution of **late** values is different for posts that are retakes and posts that are not retakes.
  - ☐ For all Data 8 students, posts that are retakes have larger **late** values on average than posts that are not retakes.
  - ☐ For all BeReal users, retaking photos causes a change in how late those photos are posted on average.
  - ☐ For all Data 8 students, retaking photos causes a change in how late those photos are posted on average.

(f) (0.0 pt) **OPTIONAL.** Draw or describe the BeReal post that you would make right now.

A large, empty rectangular box with a thin black border, intended for a student to draw or describe a BeReal post. The box is square-shaped and occupies the majority of the lower half of the page.

2. (12.0 points) Histograms

The Mayo Clinic recommends drinking 13.5 cups of water per day. The histogram below describes the average daily water intake during September 2022 (measured in cups of water) for 400 Berkeley students. The exact heights of each bar appear in the adjacent table.



Choose the mathematical expression that evaluates to each quantity described below.

(a) (2.0 pt) The **number** of students who drank an average of at least 8 but less than 10 cups per day.

- ☐ 0.05  
☐  $0.05 * 2$   
☐  $0.05 * 100$   
☐  $0.05 * 2 * 100$   
☐  $0.05 * 400$   
☒  $0.05 * 2 * 400$   
☐ Not enough information provided  
☐ None of these

(b) (2.0 pt) The **proportion** of students who drank an average of at least 13.5 cups per day.

- ☐  $0.02 + 0.01$   
☐  $0.065 + 0.02 + 0.01$   
☐  $0.065 / 2 + 0.02 + 0.01$   
☐  $0.065 / 4 + 0.02 + 0.01$   
☒ Not enough information provided  
☐ None of these



- (c) (2.0 pt) The **proportion** of students who drank an average of at least 4 cups per day.
- ☐  $1 - (0.11)$
  - ☐  $1 - (0.11) * 2$
  - ☐  $1 - (0.01 + 0.04)$
  - ☒  $1 - (0.01 + 0.04) * 2$
  - ☐  $1 - (0.01 + 0.04 + 0.11)$
  - ☐  $1 - (0.01 + 0.04 + 0.11) * 2$
  - ☐ Not enough information provided
  - ☐ None of these
- (d) (2.0 pt) The **proportion** of students who never drank more than 16 cups of water in any single day during September 2022.
- ☐  $1 - (0.01)$
  - ☐  $1 - (0.01) * 2$
  - ☐  $1 - (0.02 + 0.01)$
  - ☐  $1 - (0.02 + 0.01) * 2$
  - ☒ Not enough information provided
  - ☐ None of these
- (e) (2.0 pt) The height (in percent per cup) of the bin from 4 to 8 if this histogram were drawn with `bins=make_array(0, 2, 4, 8, 18)`.
- ☐  $(11 + 12) / 4$  (which equals 5.75)
  - ☒  $(2 * 11 + 2 * 12) / 4$  (which equals 11.5)
  - ☐  $11 + 12$  (which equals 23)
  - ☐  $2 * (11 + 12)$  (which equals 46)
  - ☐ Not enough information provided
  - ☐ None of these
- (f) (2.0 pt) The maximum possible height (in percent per cup) of the bin from 4 to 5 if this histogram were drawn with `bins=np.arange(19)`.
- ☐ 11
  - ☒  $11 * 2$
  - ☐  $11 / 2$
  - ☐  $11 * 4$
  - ☐  $11 / 4$
  - ☐ None of these

### 3. (16.0 points) Chances

Some helpful students decide to pick up trash in a park near their dorm. Half the pieces of trash are bottles, one third are boxes, and one sixth are food.

#### (a) (8.0 points)

For each event below, choose the Python expression that evaluates to the probability of that event when **two** pieces of trash are chosen at random **with replacement**.

i. (2.0 pt) The probability that they are both boxes.

- ☐  $(1 / 3)$
- ☐  $(1 / 3) + (1 / 3)$
- ☒  $(1 / 3) ** 2$
- ☐  $((1 / 3) + (1 / 3)) ** 2$
- ☐  $(1 / 3) ** 2 + (1 / 3) ** 2$

ii. (2.0 pt) The probability that the first one is a box.

- ☒  $(1 / 3)$
- ☐  $(1 / 3) * (2 / 3)$
- ☐  $(1 / 3) ** 2$
- ☐  $1 - (1 / 3)$
- ☐  $1 - (1 / 3) * (2 / 3)$
- ☐  $1 - (1 / 3) ** 2$

iii. (2.0 pt) The probability that one is a box and the other is food. (*Note:* Either the box or food can be chosen first for this event to occur.)

- ☐  $(1 / 3) + (1 / 6)$
- ☐  $(1 / 3) * (1 / 6)$
- ☐  $((1 / 3) + (1 / 6)) ** 2$
- ☐  $((1 / 3) * (1 / 6)) ** 2$
- ☐  $2 * ((1 / 3) + (1 / 6))$
- ☒  $2 * ((1 / 3) * (1 / 6))$

iv. (2.0 pt) The probability that they are both the same kind of trash.

- ☐  $(1 / 2) * (1 / 3) * (1 / 6)$
- ☐  $1 - ((1 / 2) * (1 / 3) * (1 / 6))$
- ☒  $(1 / 2) ** 2 + (1 / 3) ** 2 + (1 / 6) ** 2$
- ☐  $(1 / 2) ** 2 * (1 / 3) ** 2 * (1 / 6) ** 2$

## (b) (8.0 points)

There are 6,000 pieces of trash in the park (3,000 bottles, 2,000 boxes, 1,000 pieces of food). Adrian collects 20 pieces at random and observes that A of them are bottles. Bala collects 100 pieces at random and observes that B of them are bottles.

i. (3.0 pt) Which of the following are more probable than not? Choose all that apply.

- ☐ A is larger than B
- ☒ B is larger than A
- ☐  $(A / 20)$  is larger than  $(B / 100)$
- ☐  $(B / 100)$  is larger than  $(A / 20)$
- ☒  $\text{abs}(A / 20 - 0.5)$  is larger than  $\text{abs}(B / 100 - 0.5)$
- ☐  $\text{abs}(B / 100 - 0.5)$  is larger than  $\text{abs}(A / 20 - 0.5)$
- ☐ None of these

ii. (3.0 pt) Fill in the blank so that calling `bottles(n)` simulates collecting n pieces of trash at random with replacement from the park and returns the **number** that are bottles. For credit, your answer must fit on one line.

```
trash = make_array('bottle', 'bottle', 'bottle', 'box', 'box', 'food')
```

```
def bottles(n):
    "Return the number of bottles in a random sample of n pieces of trash."

    return _____
```

```
sum(np.random.choice(trash, 20) == 'bottle')
```

iii. (2.0 pt) Fill in the blank so that p is an empirical estimate of the probability that 20 pieces of trash chosen at random with replacement from the park will contain **at least 12** bottles.

```
results = make_array()
for i in np.arange(10000):
    results = np.append(results, bottles(20))
```

```
p = _____
```

```
sum(results >= 12) / len(results)
```

#### 4. (30.0 points) Stars

You read online that the Milky Way galaxy contains 100 billion active stars, 10 billion white dwarf stars, and 1 billion neutron stars. You also find the following *near-Earth dataset* online: within 300 trillion kilometers (10 parsecs) of Earth, there are 392 active stars and 21 white dwarf stars, but 0 neutron stars.

(a) (2.0 pt) Which of the following are true about the near-Earth dataset? Choose all that apply.

- ☒ It describes a sample of the stars in the galaxy.
- ☐ It describes a simple random sample of the stars in the galaxy.
- ☐ It describes a randomly selected region of the galaxy.
- ☒ It describes the population of all stars within 300 trillion kilometers of Earth.
- ☐ None of these.

(b) (8.0 points)

You decide to use these data to investigate whether the distribution of stars (active stars, white dwarf stars, and neutron stars) near Earth is the same as the distribution across the galaxy or whether it's unusual.

i. (2.0 pt) Complete this **null** hypothesis: *The distribution of stars...*

- ☐ near Earth has exactly the same proportions as the galaxy's distribution.
- ☒ near Earth is like a random sample from the galaxy's distribution.
- ☐ in a randomly selected region of the Milky Way is the same as the galaxy's distribution.
- ☐ in every sample of stars is the same as the galaxy's distribution.

ii. (2.0 pt) Complete this **alternative** hypothesis: *The distribution of stars near Earth...*

- ☐ is the same as the galaxy's distribution.
- ☐ is consistent with the galaxy's distribution.
- ☒ is not like a random sample from the galaxy's distribution.
- ☐ is not the same as the distribution from a randomly selected region of the galaxy.

iii. (2.0 pt) Which test statistic is best for choosing between these null and alternative hypotheses?

- ☒ Total variation distance from the galaxy's distribution
- ☐ Difference between the number of neutron stars in the sample and the galaxy
- ☐ Difference between the proportion of neutron stars in the sample and the galaxy
- ☐ Difference between the sizes of the samples

iv. (2.0 pt) Under the null hypothesis, what would best explain the fact that near Earth there are 0 neutron stars?

- ☐ The number of stars near Earth is small.
- ☐ Neutron stars are dim and hard to observe.
- ☐ Neutron stars are mostly near the center of the galaxy and Earth is not.
- ☒ Due to random chance, there happen to be 0 neutron stars near Earth.

**(c) (10.0 points)**

Fill in the blanks of the simulation below to calculate a p-value for this hypothesis test by sampling 10000 times under the null hypothesis. (*Note: 1e11 is a valid way to express 100 billion in Python.*)

```
def half_abs_sum(a):
    "Return half the sum of the absolute values of a single array."
    return sum(np.abs(a)) / 2

obs = make_array(392, 21, 0) # An array of counts
dist = make_array(1e11, 1e10, 1e9) / (1e11 + 1e10 + 1e9) # An array of proportions
observed_stat = half_abs_sum(_____)
                                (a)

stats = make_array()
for i in _____:
    (b)

    sample = sample_proportions(_____, dist)
                                (c)

    stats = np.append(stats, half_abs_sum(_____))
                                (d)

p_value = _____ / 10000
          (e)
```

i. (2.0 pt) Fill in blank (a).

```
obs / sum(obs) - dist
```

ii. (2.0 pt) Fill in blank (b).

```
np.arange(10000)
```

iii. (1.0 pt) Which of these could fill in blank (c)?

- ☐ obs
- ☒ sum(obs)
- ☐ stats
- ☐ sum(stats)

iv. (2.0 pt) Fill in blank (d).

```
sample - dist
```

v. (3.0 pt) Fill in blank (e).

```
sum(stats >= observed_stat)
```

## (d) (10.0 points)

Interpret the hypothesis test you just implemented.

i. (4.0 pt) What does the p-value represent in this hypothesis test? Choose all that apply.

- ☐ The theoretical probability that the alternative hypothesis is true.
- ☐ An empirical estimate of the probability that the alternative hypothesis is true.
- ☐ The theoretical probability that the null hypothesis is true.
- ☐ An empirical estimate of the probability that the null hypothesis is true.
- ☐ The theoretical probability under the null hypothesis that the test statistic would be at least as large as the observed test statistic.
- ☒ An empirical estimate of the probability under the null hypothesis that the test statistic would be at least as large as the observed test statistic.
- ☐ The fraction of all possible samples under the null hypothesis that have a test statistic at least as large as the observed test statistic.
- ☒ The fraction of simulated samples under the null hypothesis that had a test statistic at least as large as the observed test statistic.
- ☐ None of these

ii. (2.0 pt) What does it mean to describe the p-value cutoff as an error probability?

- ☐ The p-value cutoff is the chance that the null hypothesis is true.
- ☐ The p-value cutoff is the chance that the null hypothesis is false.
- ☒ The p-value cutoff is the chance that the null hypothesis will be rejected if it were true.
- ☐ The p-value cutoff is the chance that the null hypothesis will not be rejected if it were false.

iii. (2.0 pt) Assuming everything that you read online is accurate and the computed `p_value` from your code is 0.0015. What can you conclude using a p-value cutoff of 1%? Choose all that apply.

- ☐ The data are consistent with the null hypothesis.
- ☒ The data are **not** consistent with the null hypothesis.
- ☒ The null hypothesis can be rejected.
- ☐ The null hypothesis **cannot** be rejected.

iv. (2.0 pt) Which of the following facts, if true, should cause you to question the conclusions you drew from this hypothesis test? Choose all that apply.

- ☒ The total number of active stars in the Milky Way is unknown, but estimated to be between 100 billion and 400 billion.
- ☒ White dwarf stars are difficult to observe, so scientists believe there are some near Earth that have not yet been counted.
- ☐ Many of the active stars near Earth are much younger than the Sun.
- ☐ The rotational speed of the Milky Way implies that 90% of the mass in the galaxy is dark matter.
- ☐ None of these