CSCI 3022
Final Exam
Fall 2018

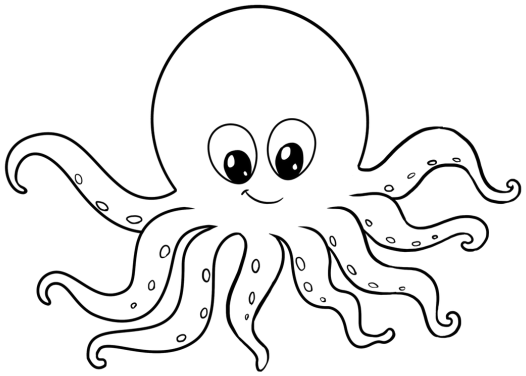
Write **as clearly as you can** and **in the box**:

Name:
Student ID:
Section number:

Read the following:

- **RIGHT NOW**! Write your name on the top of your exam.
- You are allowed **one** 3×5in notecard of **handwritten** notes (both sides). No magnifying glasses!
- You may use a calculator provided that it cannot access the internet or store large amounts of data.
- You may **NOT** use a smartphone as a calculator.
- Clearly mark answers to multiple choice questions on the provided answer line.
- Mark only one answer for multiple choice questions. If you think two answers are correct, mark the answer that **best** answers the question. No justification is required for multiple choice questions.
- If you do not know the answer to a question, skip it and come back to it later.
- For free response questions you must clearly justify all conclusions to receive full credit. A correct answer with no supporting work will receive no credit.
- You have **150 minutes** for this exam.

Page	Points	Score
MC	24	
SA	8	
FR1	17	
FR2	17	
FR3	17	
FR4	17	
Total	100	



Potentially Useful Values and Formulas

Standard Normal Distribution: Here $\Phi(z)$ is the cumulative distribution function for the standard normal distribution evaluated at z. Its equivalent form in Python is $\Phi(z) =$ stats.norm.cdf(z),

$\Phi(4.75) \approx 1.000$	$\Phi(3.00) = 0.999$	$\Phi(2.58) = 0.995$	$\Phi(2.32) = 0.990$	$\Phi(2.00) = 0.977$
$\Phi(1.96) = 0.975$	$\Phi(1.88) = 0.970$	$\Phi(1.80) = 0.964$	$\Phi(1.75) = 0.960$	$\Phi(1.64) = 0.950$
$\Phi(1.44) = 0.925$	$\Phi(1.28) = 0.900$	$\Phi(1.15) = 0.875$	$\Phi(1.04) = 0.850$	$\Phi(1.00) = 0.841$
$\Phi(0.93) = 0.825$	$\Phi(0.84) = 0.800$	$\Phi(0.76) = 0.775$	$\Phi(0.67) = 0.750$	$\Phi(0.60) = 0.725$
$\Phi(0.52) = 0.700$	$\Phi(0.45) = 0.675$	$\Phi(0.39) = 0.650$	$\Phi(0.32) = 0.625$	$\Phi(0.25) = 0.600$
$\Phi(0.19) = 0.575$	$\Phi(0.13) = 0.550$	$\Phi(0.06) = 0.525$	$\Phi(0.00) = 0.500$	

Student's t-Distribution: The following values of the form $t_{\alpha,v}$ are the critical values of the t-distribution with v degrees of freedom, such that the area under the pdf and to the right of $t_{\alpha,v}$ is α. Its equivalent form in Python is $t_{\alpha,v} =$ stats.t.ppf$(1-\alpha, v)$.

$t_{0.05,48}$	$=$	1.677	$t_{0.025,48} = 2.011$
$t_{0.05,44}$	$=$	1.680	$t_{0.025,44} = 2.015$
$t_{0.05,9}$	$=$	1.833	$t_{0.025,9} = 2.262$
$t_{0.05,2}$	$=$	2.920	$t_{0.025,2} = 4.303$

F-Distribution: The following values of the form F_{α,v_1,v_2} are the critical values of the F-distribution with v_1 and v_2 degrees of freedom, such that the area under the pdf and to the right of F_{α,v_1,v_2} is α. Its equivalent form in Python is $F_{\alpha,v_1,v_2} =$ stats.f.ppf$(1-\alpha, v_1, v_2)$.

$F_{0.05,2,9} = 4.256$	$F_{0.05,3,12} = 3.490$
$F_{0.01,2,9} = 8.022$	$F_{0.01,3,12} = 5.953$
$F_{0.025,2,9} = 5.271$	$F_{0.025,3,12} = 5.715$
$F_{0.005,2,9} = 7.226$	$F_{0.005,3,12} = 10.107$

Bayes' theorem	$p(A \mid B) = \dfrac{p(B \mid A)p(A)}{p(B)}$	Law of total probability	$p(E) = \sum_{i=1}^{N} p(E \mid F_i)p(F_i)$
Union of sets	$p(A \cup B) = p(A) + p(B) - p(A \cap B)$	Conditional probability	$p(A \mid B) = \dfrac{p(A \cap B)}{p(B)}$
Sigmoid function	$\text{sigm}(z) = \dfrac{1}{1+e^{-z}}$	Regression	$\hat{\sigma}^2 = \dfrac{SSE}{n-2}, \quad SE(\hat{\beta}) = \dfrac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}$

Three types of F you might use: $F = \dfrac{(SST - SSE)/p}{SSE/(n-p-1)}$ $F = \dfrac{SSB/df_{SSB}}{SSW/df_{SSW}}$ $F = \dfrac{(SSE_{red} - SSE_{full})/(p-k)}{SSE_{full}/(n-p-1)}$

Some confidence intervals: $\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ $\bar{x} \pm t_{\alpha/2,n-1}\dfrac{\sigma}{\sqrt{n}}$ $\left[\dfrac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \dfrac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right]$

Multiple choice problems: Write your answers in the boxes, or they will not be graded!

1. (2 points) You have collected a sample of data from a population of unknown distribution. The sample consists of $n = 10000$ measurements, with sample mean 500 and sample standard deviation 200. Kenzie walks into the lab and exclaims, "Season's Yeetings! It is Wednesday, my dudes. And 494 is the mean of that population!" While Kenzie's memes are always fresh, her information about the mean of the population might be out of date, so you decide to perform a hypothesis test to determine if there is sufficient evidence to conclude that the mean of the population you sampled is greater than 494.

   Which of the following correspond to the p-value for the relevant hypothesis test?

   Let $f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-z^2/2\right]$ represent the standard normal probability density function.

   A. $2\int_{500}^{\infty} f(z)dz$

   B. $\int_{500}^{\infty} f(z)dz$

   C. $\int_{-\infty}^{500} f(z)dz$

   D. $2\int_{3}^{\infty} f(z)dz$

   E. $\int_{3}^{\infty} f(z)dz$

   F. $\int_{-\infty}^{3} f(z)dz$

   G. $f(3)$

   H. $f(500)$

   I. $\Phi(3)$

   J. $\Phi(500)$

   $500 - 494$

   $200/\sqrt{10000}$

   [ ]

2. (2 points) You have collected a sample of size 100 data points and compute a 90% confidence interval for the mean amount of bread gotten by workers in Boulder County. The confidence interval you obtained is $[5, 15]$, where the units are loaves/day. You want to check this result using a larger sample, so you collect a new sample of size 900. The mean of your new, larger sample is 1 loaf/day higher than the smaller sample, but the standard deviation is the same. Using the sample of size 900, what is the 90% confidence interval for mean amount of bread gotten by workers in Boulder County?

   A. $[6, 16]$

   B. $[5, 15]$

   C. $[9.33, 12.67]$

   D. $[8.5, 11.5]$

   E. $[10.44, 11.56]$

   F. $[0.05, 0.95]$

   $CI = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

   $[5, 15] = 10 + 1.64 \cdot \frac{\sigma}{\sqrt{100}}$

   $\frac{10}{1.64}$

   $CI = 11 + 1.64 \frac{\sigma}{30}$

   [ ]

3. (2 points) Suppose you compute a 95% confidence interval using the original sample of size 100 from the previous problem. Which of the following is the best estimate of this 95% confidence interval?

   A. $[4, 16]$

   B. $[6, 14]$

   C. $[5, 15]$

   D. $[6, 16]$

   E. $[2.13, 17.9]$

   [ ]

Multiple choice problems: Write your answers in the boxes, or they will not be graded!

4. (2 points) Dan is adopting pets at random. He will adopt 4 animals, each of which will be an octopus with probability 50% or a buffalo with probability 50%. Out of the following possible outcomes, which is the most likely to occur?

   A. Dan adopts 3 of one type of animal and 1 of the other type (the 3 of one type could be octopuses or buffaloes; we do not know which he will come home with more of)

   B. Dan adopts 4 of one type of animal and none of the other type (the 4 of one type could be octopuses or buffaloes; we do not know which one he will come home with)

   C. Dan adopts 2 octopuses and 2 buffaloes

   D. Dan adopts 3 octopuses and 1 buffalo

   E. Dan adopts 6 penguins

5. (2 points) Consider simulating the roll of a fair, six-sided die. Which of the following quantities does the following function estimate?

```
def rolling_rollers(num_samples):
    rolls = np.random.choice([1,2,3,4,5,6], size=num_samples)
    return np.sum(np.logical_and((rolls % 2)==1),(rolls > 2)) / np.sum((rolls % 2)==1)
```

   A. $P(X \text{ is Odd} \mid X > 2)$

   B. $P(X > 2 \cap X \text{ is Odd})$

   C. $P(X > 2 \mid X \text{ is Odd})$

   D. $P(X > 2 \cup X \text{ is Odd})$

6. (2 points) Claire is in CSEL. It is basically Antarctica in there. Everyone is freezing. On average, the people studying in CSEL get hypothermia at a rate of 3 people per hour. What a mess. Michael tries to adjust the thermostat. Nothing happens. Which expression, below, computes the probability that 2 or more people get hypothermia in CSEL in any particular hour?

   A. $\frac{3^2}{2!}e^{-3}$

   B. $1 - \sum_{k=0}^{2} \frac{3^k}{k!}e^{-3}$

   C. $\frac{2^3}{3!}e^{-2}$

   D. $1 - \sum_{k=0}^{1} \frac{3^k}{k!}e^{-3}$

   E. $\int_{k=2}^{\infty} \frac{3^k}{k!}e^{-3}dk$

Multiple choice problems: Write your answers in the boxes, or they will not be graded!

For problems 7 to 11, consider 3 octopuses: Anders, Becky, and Cthulhu. They live in a beautiful reef, and have lots to eat, but boy, are they moody! On weekdays, the probability that Anders is grumpy is 0.5. On weekdays, the probability that Anders and Becky are *both* grumpy is 0.1. Use the following list to select your answers for all questions on this page. Answers may be used more than once.

A. 0.1	E. 0.4	I. 0.75
B. 0.2	F. 0.5	J. 0.8
C. 0.25	G. 0.6	K. 0.9
D. 0.3	H. 0.7	L. not enough information

7. (2 points) If, on a weekday, neither Anders nor Becky are grumpy with probability 0.2, what is the probability that Becky is grumpy (irrespective of Anders)?

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(B) = P(A \cup B) - P(A) + P(A \cap B) = 0.8 - 0.5 + 0.1 = 0.4$

E

8. (2 points) On a weekday, Cthulhu is grumpy with probability 0.3, and the probability that either Cthulhu or Anders (or both) is grumpy is 0.6. What is the probability that both Cthulhu and Anders are grumpy?

$P(C \cup A) = 0.6$

$P(C \cap A)$?

$P(C \cup A) = P(C) + P(A) - P(C \cap A)$

$P(C \cap A) = P(C) + P(A) - P(C \cup A)$

$0.3 + 0.5 - 0.6$

D

9. (2 points) It turns out that on weekdays, Becky and Cthulhu are never both grumpy at the same time. What is the probability that all 3 octopuses are not grumpy on a weekday?

A

10. (2 points) It's a weekday and you read on Twitter that Anders is grumpy. What is the probability that Becky is grumpy?

$P(B \mid A) = \frac{P(A \mid B) P(B)}{P(A)} = \frac{0.1}{0.5}$

B

11. (2 points) Let's focus on Anders. Anders lives for the weekend. If Anders is grumpy, then the probability that it is a weekday is 5/6. Given that it is the weekend, what is the probability that Anders is grumpy? You may assume that the probabilities of it being any particular day are all equal.

12. (2 points) Let $f(x)$ be the PDF of a normal distribution with mean 1 and variance 0.01. Compute

$$\int_{1.2}^{\infty} f(x)dx$$

A. ~~The limit does not exist.~~

B. 0.1

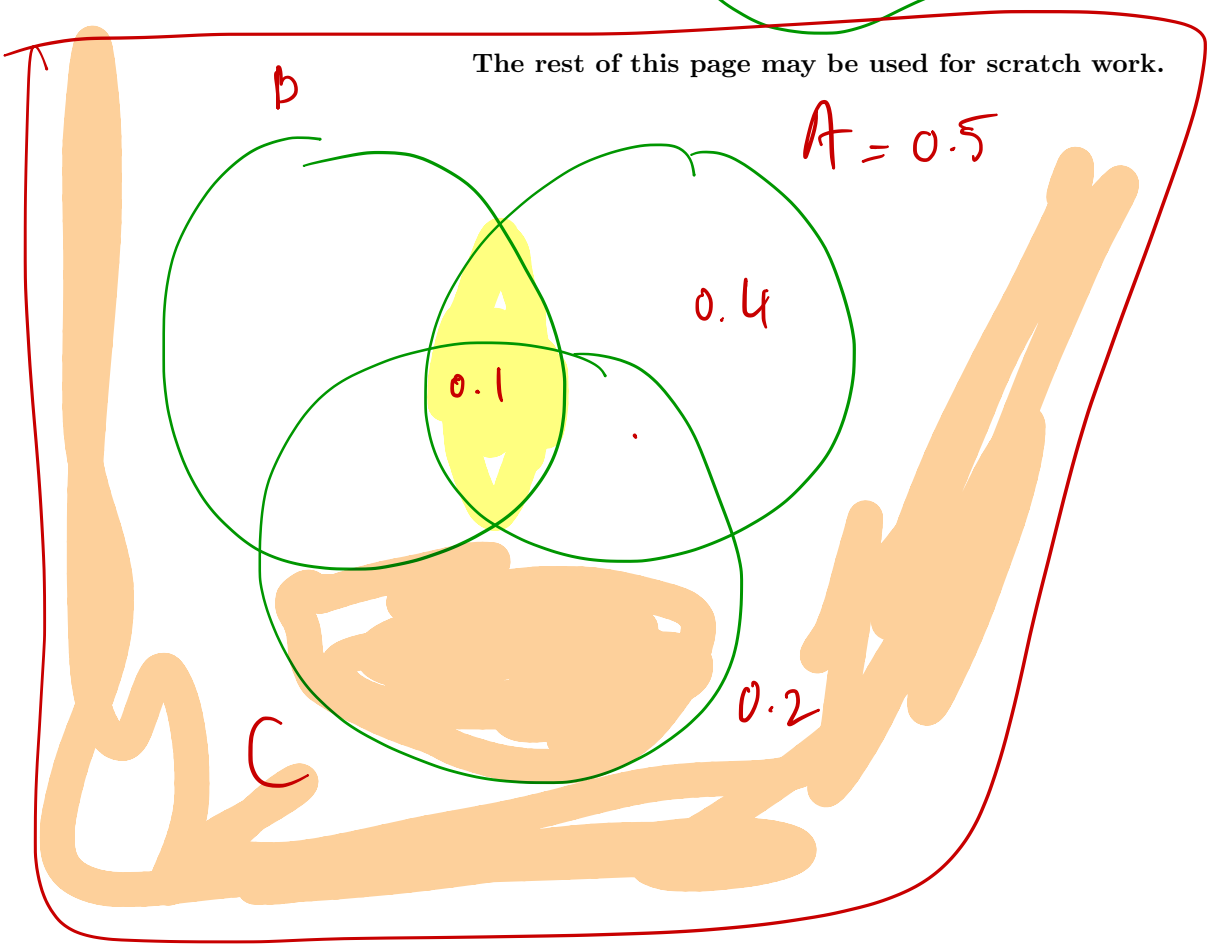
C. 0.036

D. 0.023

E. 0.0115

F. 0.01

The rest of this page may be used for scratch work.



Short answer problems: If you answer does not fit in the box provided, make a note of where it is continued!

13. (4 points) If you're doing quality control for the average strength of carbon fiber that will be used in airplane construction, and your alternative hypothesis is that the strength of the carbon is below tolerance, and therefore unsafe, would you rather have a low Type I error rate or a low Type II error rate? Explain.

14. (4 points) In their free time, Stella and Ferdinand O'Flaherty often flip coins to determine if the coins are fair. They are assessing the bias of a coin with an octopus on one side and the formula for integration by parts on the other side. "Weird!" they think. Stella and Ferdinand flip the coin 100 times, and decide that if ≥ 60, or ≤ 40 flips result in the octopus side, they will deem the coin unfair. What significance level are they using in this hypothesis testing experiment?

Write your answer for the significance level in this box:

and show your work in the larger box below.

$H_0 : p = 0.5$

$H_1 : \hat{p} \geq 60$ or $\hat{p} \leq 40$

$$z_{\alpha/2} = \frac{0.6 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = 2$$

$$\frac{\alpha}{2} = 1 - \Phi(2) = 0.023 \Rightarrow \alpha = 0.046$$

Free response problems: If your answers do not fit on this page and the back of it, make a note of where the work is continued!

15. (17 points) Suppose you have collected a sample from a population whose probability density function is governed by some unknown parameter θ. You compute a test statistic, X, which is defined for values in $[0, 1]$ and has the following probability density function.

$$f(x) = 2\theta x + 1 - \theta$$

Please answer the following questions, and be sure to show your work—for sufficient space, a blank page follows this one.

(a) (2 points) You want to see if there is sufficient evidence to conclude that the true parameter value is *greater than* $\theta = 1/2$. State the relevant null and alternative hypotheses.

(b) (7 points) Sketch the probability density function $f(x)$, assuming that the null hypothesis from part (a) is true. Label your axes, the density function f, a few important $x-$ and $y-$tick marks along your axes.

Suppose the test statistic that you compute is $X = 3/4$. Depict this in your sketch as a vertical line, and clearly mark in your sketch what the area/value/point that gives the p-value associated with the hypothesis test from part ().

(c) (6 points) Compute the p-value associated with the hypothesis test from part (a), assuming you have the test statistic given in part (b). You must perform this calculation by hand and show all work to receive credit.

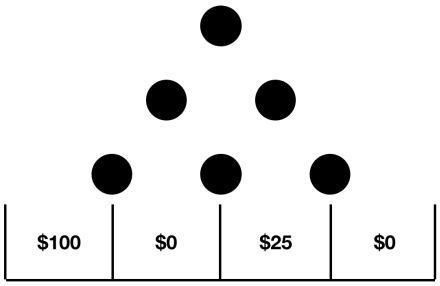
(d) (2 points) What can we conclude, at the 10% significance level?

(a) $H_0: \theta_0 = 1/2$

$H_1: \theta_1 \geq 1/2$

(b) $f(x) = x + 1/2$

Additional Workspace



16. (17 points) A game of Plinko is to be played on the board above. You notice that the \$100 bin is on the left so you tilt the board such that the probability that the Plinko disk moves left is $\frac{3}{4}$ and probability that the disk moves to the right is $\frac{1}{4}$. As usual, movements at all pegs are independent, from each other, and there is no funny business except for your board tilting, which no one seems to have noticed.

    (a) (3 points) What is the PMF for your winnings for a single Plinko disk?

    (b) (4 points) You drop 5 disks. What is the probability that your total winnings are \$100?

    (c) (3 points) You drop 5 disks. What is the probability that your total winnings are \$125?

    (d) (4 points) What are your expected winnings with 5 disks?

    (e) (3 points) You decide that the board is tilted a little too much, and endeavor to balance it such that for a single disk, $P(\text{win } \$100) = P(\text{win } \$25) > 0$. Analytically compute the required bias p, where p is the probability that a disk moves to the right at each peg.

Additional Workspace

17. (17 points) You are literally stuck in the sewer. UGH! How does this sort of thing always happen during exams?? You are trying to find your way out by walking along a wide tunnel. You arrive at a junction, and see, with your flashlight, that there are three possible paths. You look left, and see that there are lights spaced every so often along the ceiling. *Lit*, you think to yourself. You look right, and see that there is water dripping from the ceiling and pooling on the ground. *Dank*, you think to yourself. You look straight ahead and see that tree roots have broken through some of the walls and are twisted and curled across the ground. *Gnarly*, you think to yourself. Now you must decide: Lit, Dank, or Gnarly?

    Luckily, this is the sort of decision that you are good at because: Data Science. The sewer smells bad, but perhaps one of these paths smells *less* bad than the others. You conduct four independent samples of the air quality in each of the three directions, measuring the air's bad smell in units of *anti-freshness* (AF). You record your data and write it into the table below:

Lit (AF)	Dank (AF)	Gnarly (AF)
0	3	5
1	4	5
2	6	2
1	3	4

    (a) (4 points) Name a statistical test that you can use to compare the mean air qualities in the three directions to determine whether or not the three directions are equivalent. Then, clearly state the null and alternative hypotheses for your test.

    (b) (9 points) Compute the relevant **test statistic** to test the hypotheses from part (a). Put a **box** around your answer for the test statistic. Show all calculations if you wish to receive credit.

    (c) (4 points) For a test at the $\alpha = 0.05$ significance level, perform a rejection region test for your test statistic from part (b). Be sure to clearly state (i) the distribution you are referencing (including any degrees of freedom), (ii) the critical value to which you are comparing your test statistic, and (iii) the conclusion of your test.

Additional Workspace

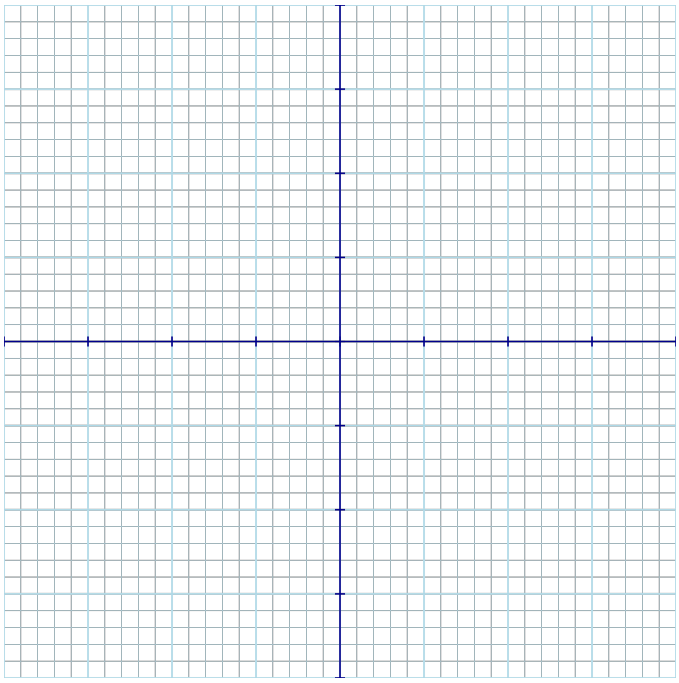
18. (17 points) You wake up in the morning, and rub the sleep from your eyes. The sun is shining. *Let's get this bread,* you think to yourself.

You go down to the kitchen to acquire bread. Unfortunately, the bread that you have in your pantry is beginning to go stale. Naturally, you collect data and fit a logistic regression model to predict whether a piece of bread is stale or edible, given two features: x_1 and x_2. You fit a model of the form

$$P(\text{edible} \mid x_1, x_2) = \text{sigm}(\beta_0 + \beta_1 x_1 + \beta_2 x_2).$$

(a) (2 points) In general, (i) what is the meaning of the term decision boundary and (ii) what is the shape of the decision boundary in a logistic regression with two features?

(b) (4 points) Suppose from here on out that $\beta_0 = 2$, $\beta_1 = 1$, and $\beta_2 = -2$. Show the decision boundary on the axes below. Be sure to label your axes. Label three regions on the plot as "Classify as edible", "Classify as stale", "Unable to classify."



(c) (8 points) Use your results from part (b) to classify the following pieces of bread. You may mark on the plot from part (b) if you wish, provided that the decision boundary and previous labels are not obscured.

x_1	x_2	**Classification** (edible, stale, or unable to classify)
-2	-2	
-4	0	
0	3	
2	2	
0	0	
1	2	
3	0	
4	2	

(d) (3 points) Still attempting to acquire the bread, you go to the very back of the pantry and find a piece of bread with features $(x_1, x_2) = (-2, \ln \frac{1}{2})$. What are the odds that it is edible? State those odds using the conventions discussed in class.

Additional Workspace