18.05 Final Exam Spring 2022 Solutions

Part I: Concept questions (70 points)

These questions are all multiple choice or short answer. You don't have to show any work. Work through them quickly.

Concept 1. (4 pts.) *Which of the following is a valid probability table?*

(i)

outcome	*1*	*1.5*	*2*	*2.5*	*3*	*3.5*
probability	*1/5*	*1/5*	*1/5*	*1/5*	*1/5*	*1/5*

(ii)

outcome	*red*	*blue*	*green*	*cyan*	*yellow*
probability	*4/10*	*2/10*	*0*	*1/10*	*3/10*

Circle the best choice:

(i) **(ii)** *(i) and (ii)* *neither (i) nor (ii)*

Solution: (ii)

Concept 2. (4 pts.) *Suppose $P(A) + P(B) > 1$. Consider the following statements.*

(i) $P(A \cup B) = 1$. *(ii)* $P(A \cap B) > 0$.

Which **must** *be true? Circle the best choice below:*

(i) **(ii)** *(i) and (ii)* *neither (i) nor (ii).*

Solution: (ii)

Concept 3. (6 pts.) *Circle* **True** *or* **False** *for each of the following.*

(a) *If A and B are independent then we must have $P(A \cap B) = P(A)P(B)$.* **True** **False**

(b) *If A and B are independent then we must have $P(A \cap B) = P(A) + P(B)$.* **True** **False**

(c) *If A and B are disjoint then A and B must be independent.* **True** **False**

Solution: True False False

Concept 4. (4 pts.) *You believe the MBTA subway arrives late by X hours, where X follows an exponential distribution with unknown parameter λ. To test your hypothesis, you record the lateness of 5 subway trains and get data $x_1, x_2, \dots, x_5$. Which of the following are statistics? Circle the correct answers.*

(a) *The expected value of a sample, namely $1/\lambda$.*

(b) The sample average, $\overline{x} = (x_1 + x_2 + x_3 + x_4 + x_5)/5$.

(c) *The difference between $\overline{x}$ and $1/\lambda$.*

(d) The sample standard deviation.

Solution: (b) and (d)

Concept 5. (3 pts.) *For each of the following, circle it if it is used in Bayesian inference.*

(a) **Likelihood function** ***(b)*** **prior odds** ***(c)*** *p-value*

Solution: Circle **(a) and (b)**

Concept 6. (4 pts.) *Suppose $X \sim$ Bernoulli(θ), where θ is unknown. Complete the following sentence using the words, "discrete," "continuous," or "neither discrete nor continuous."*

The random variable is **discrete**, *the space of hypotheses is* **continuous**.

Solution: discrete, continuous

Concept 7. (2 pts.) *A casino is considering installing a new slot machine. A player who wins is paid \$2 on a \$1 bet. The manufacturer claims that the probability of winning on any play of the slot machine is $p = 0.48$. Before using the machine the casino wants to make sure it will make them money. So they hire you to test the slot machine. Which of the following hypotheses would you use?*

(i) $H_0 : p = 0.48$ *vs* $H_A : p \neq 0.48$

(ii) $H_0 : p = 0.48$ *vs* $H_A : p > 0.48$

(iii) $H_0 : p = 0.48$ *vs* $H_A : p < 0.48$

Circle the best answer: *(i)* **(ii)** *(iii)* *Not enough information.*

Solution: (ii)

Concept 8. (6 pts.) *The following are hypotheses considered in the previous problem. For each hypothesis circle all that apply.*

(a) $H_0 : p = 0.48$	**Simple**	**Composite**	**Two-sided**	**One-sided**
(b) $H_A : p \neq 0.48$	**Simple**	**Composite**	**Two-sided**	**One-sided**
(c) $H_A : p > 0.48$	**Simple**	**Composite**	**Two-sided**	**One-sided**

(a) Solution: Simple
(b) Solution: Composite Two-sided
(c) Solution: Composite One-sided

Concept 9. (5 pts.) *Which of the following are* ***true*** *about p-values? Circle all that apply.*

(a) *The p-value gives the probability of making a type 1 error.*

(b) *The p-value is a measure of how extreme the observed data is.*

(c) *A p-value below the significance level allows us to conclude with certainty that the null hypothesis is false.*

(d) *The p-value is a frequentist concept.*

(e) *If the null hypothesis is true, then the p-value will always be larger than the significance level.*

Solution: (b), (d)

Concept 10. (8 pts.) *A two-sample t-test for equal means of two populations has a p-value of* 0.08.

Circle True or False for each of the following.

(a) *For a significance level of* 0.05*, the null hypothesis of equal means should be rejected.*

True **False**

(b) *A 90% confidence interval for the difference of the means for the two populations includes* 0*.*

True **False**

(c) *A 95% confidence interval for the difference of the means for the two populations includes* 0*.*

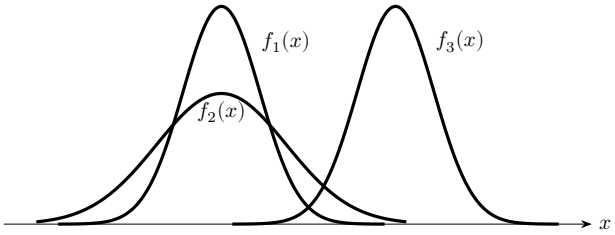
True **False**

(d) *With probability 95% the actual value of the difference of the means is within the 95% t-confidence interval for the difference.*

True **False**

Solution: False False True, False

Concept 11. (2 pts.) *Data is drawn from a* $N(1, \sigma^2)$ *distribution. Let* $\overline{X}_5$ *be the average of 5 data points and* $\overline{X}_{10}$ *the average of 10 data points. Three densities,* f_1*,* f_2*,* f_3 *are shown below. One is the pdf of* $\overline{X}_5$*, one is the pdf of* $\overline{X}_{10}$ *and one is another pdf. Circle the one that is the pdf of* $\overline{X}_{10}$*.*



Solution: f_1

Concept 12. (4 pts.) *Finish the following sentences with "a type I error", "a type II error", or "neither type of error".*

(a) *The rejection of a false null hypothesis is* **neither type of error**

(b) *The rejection of a true null hypothesis is* **a type I error**

Solution: neither, type I

Concept 13. (8 pts.) *Suppose that the data* x_1*,* x_2*, …,* x_n *are drawn from independent, identically distributed, random variables* X_i *with mean* μ *and standard deviation* σ*. Write* $\overline{x} = (x_1 + \cdots + x_n)/n$ *for the sample mean. The Central Limit Theorem states that: (circle all that apply)*

(a) *The distribution of each random variable* X_i *is approximately symmetric around the average* μ*.*

(b) For large n**, the distribution of the sample mean is approximately symmetric** *around the average* μ*.*

(c) For large n**, the average** $\overline{x}$ **approximately follows a normal distribution.**

***(d)** For large n, $(\overline{x}-\mu)/\sigma$ approximately follows a standard normal distribution.*

Solution: (b), (c). (For (d), this follows $N(0, 1/n)$, a normal with mean 0 and variance $1/n$.)

Concept 14. (2 pts.) *Suppose for a certain endeavor θ is the probability of success. Suppose also that our prior for θ is Beta(5,5). We collect data from 30 trials, obtaining 20 successes and 10 failures. What is our posterior pdf $f(\theta|x)$? (Circle the best answer.)*

(i) *Binomial(30, 2/3)*
(ii) Beta(25, 15)
(iii) *Beta(20, 10)*
(iv) *None of the above*

Solution: (ii)

Concept 15. (2 pts.) *Circle* **True** *or* **False**.

Let s be a statistic. If the theoretical distribution of the statistic is hard to compute, then it is not advisable to use bootstrapping to compute confidence intervals for the statistic.

True **False**

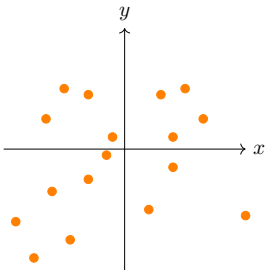
Solution: False, to the contrary

Concept 16. (2 pts.) *Suppose we run a two-sample t-test for equal means with significance level $\alpha = 0.05$. If the data implies we should reject the null hypothesis, then the odds that the two samples come from distributions with the same mean are (circle the best answer)*

***(a)** 19/1* ***(b)** 1/19* ***(c)** 20/1* ***(d)** 1/20* **(e) unknown**

Solution: e. Frequentist methods only give probabilities for data under an assumed hypothesis. They do not give probabilities or odds for hypotheses. So we don't know the odds for distribution means.

Concept 17. (2 pts.) *For the bivariate data in the following scatter plot, is the correlation between x and y positive or negative? Circle your choice:* **positive** **negative**



Solution: positive

Concept 18. (2 pts.) *Circle* **True** *or* **False**

Linear regression can fit curves other than lines to given data.

True **False**

Solution: True, e.g. polynomials

Part II: Problems (236 points)

Problem 1. (15: 5,5,5)
You roll a fair six-sided die 8 times.

(a) *What is the probability that none of the 8 rolls is a six?*

Solution: $\boxed{(5/6)^8} \approx 1/4$. (Since $5^4 = 625$ and $6^4 = 1296$, we see that $(5/6)^4$ is approximately $1/2$. Squaring…)

(b) *What is the probability that exactly one of the 8 rolls is a six?*

Solution: This is a binomial probability

$$\boxed{\binom{8}{1} \cdot \left(\frac{1}{6}\right) \cdot \left(\frac{5}{6}\right)^7} = \frac{8 \cdot 5^7}{6^8} \approx 2/5$$

(c) *What is the expected number of sixes in the 8 rolls?*

Solution: Expected values add, so **Solution:** $\boxed{8 \cdot 1/6 = 4/3.}$

Problem 2. (16: 4,4,4,4)
Suppose X is a random variable with values in [1,2] and density

$$f(x) = \begin{cases} kx^2 & \text{for } 1 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

where k is a fixed constant.

(a) *What is k?*

Solution: $\int_1^2 x^2\,dx = x^3/3]_1^2 = 8/3 - 1/3 = 7/3$. So $\boxed{k = 3/7}$.

(b) *Find the cdf of X.*

Solution: $\text{cdf}(x) = \int_1^x kt^2\,dt = (3/7)(x^3/3 - 1/3) = \boxed{(x^3-1)/7}$. This calculation is valid for $1 \le x \le 2$; the cdf is $\boxed{0 \textbf{ for } x < 1}$ and $\boxed{1 \textbf{ for } x > 2}$.

(c) *Find $P(X < 3/2)$.*

Solution: This is $\text{cdf}(3/2) = (27/8 - 1)/7 = \boxed{19/56}$.

(d) *Find $E[X]$*

Solution: $E[X] = \displaystyle\int_1^2 x \cdot (3x^2/7)\,dx = 3x^4/28\Big]_1^2 = (3 \cdot 16 - 3 \cdot 1)/28 = \boxed{45/28}$.

Problem 3. (25: 5,5,5,5,5)
Suppose random variables X and Y have units of dollars and:

$$E[X] = 5, \quad \text{Var}(X) = 6^2, \qquad \textit{and} \qquad E[Y] = 10, \quad \text{Var}(Y) = 7^2.$$

Define $W = X + Y$.

(a) *Find $E[W]$. What are the units of $E[W]$?*

Solution: The expected value of a sum of random variables is the sum of the expected values, so

$$E[W] = E[X] + E[Y] = 5 + 10 = \boxed{15}$$

The units are those of X and Y, namely **dollars**: we compute the expected value by adding up values of the variable (which are dollars) multiplied by (dimensionless) probabilities.

(b) *Assume X and Y are independent. Find $\mathrm{Var}(W)$.*

Solution: The magic formula is $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$. Independence implies that the covariance is zero, so

$$\mathrm{Var}(W) = \mathrm{Var}(X) + \mathrm{Var}(Y) = \boxed{6^2 + 7^2 = 36 + 49 = 85}.$$

(c) *Assume $\mathrm{Cov}(X,Y) = 21$. Find $\mathrm{Var}(W)$. (Note: this assumption differs from that in part (b), so $\mathrm{Var}(W)$ can also be different.)*

Solution: Same formula as used in **(b)** gives

$$\mathrm{Var}(W) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y) = \boxed{6^2 + 7^2 + 2\cdot 21 = 36 + 49 + 42 = 127}.$$

The units are square dollars, which seems like a very American concept.

(d) *Assume $\mathrm{Cov}(X,Y) = 21$. Compute $Cor(X,Y)$. What are the units of $Cor(X,Y)$?*

Solution: The correlation is by definition $\mathrm{Cor}(X,Y) = \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$. The numerator has units $(\$)^2$. What's under the square root has units $(\$)^4$, so the denominator has units $(\$)^2$, and the correlation is **dimensionless**. The value is

$$\mathrm{Cor}(X,Y) = \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} = \boxed{21/\sqrt{6^2\cdot 7^2} = 21/(6\cdot 7) = 1/2}.$$

(e) *Define $Z = \dfrac{(X - E[X])}{\sqrt{\mathrm{Var}(X)}} = (X-5)/6$. Find $E[Z]$ and $\mathrm{Var}(Z)$.*

Solution: This is the standardization of X, which is supposed to have **mean 0 and variance 1**. Here's why. The linearity of expectation $E[aX+b] = aE[X] + b$ (for constants a and b) gives

$$E[Z] = E[X]/6 - 5/6 = \boxed{5/6 - 5/6 = 0}.$$

The corresponding fact for variance is $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$, so

$$\mathrm{Var}(Z) = \mathrm{Var}(X)/6^2 = \boxed{6^2/6^2 = 1}.$$

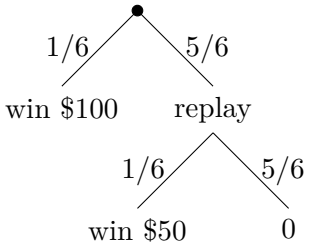
Problem 4. (20: 10,5,5)
You roll two fair (6-sided) dice. If the sum of the dice is greater than 9, you win \$100. If the sum is 9 or less you get to roll again. On the second roll, if your sum of dice is greater than 9, you win \$50, otherwise you win nothing. Let X be the random variable of how much money you win by playing this game.

(a) *Construct a probability model for X, i.e. make a probability table.*

Solution:

sum of dice	2	3	4	5	6	7	8	9	10	11	12
probability	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

So the probability of rolling less than or equal to 9 is $30/36 = 5/6$, and the probability of 10 or more is $6/36 = 1/6$. The probabilities for various outcomes of the game can be computed from the following tree.



The table of probabilities and outcomes is

outcome	$100	$50	0
prob.	1/6	5/36	25/36

(b) *What is the expected amount you will win playing the game?*

Solution: This can be read from the last table in **(a)** : it's

$$\text{Expected payoff} = (25/36)\cdot\$0 + (5/36)\cdot\$50 + (1/6)\cdot\$100 = \boxed{\$850/36 = \$23\frac{11}{18} \approx \$23.61}$$

(c) *What would you be willing to pay to play this game? (Justify your answer).*

Solution: If you **pay less than** $23.61, you'll make money in the long run. But you need a big stake to be able to be sure you can keep playing long enough if to win: since you get nothing two-thirds of the time, you have a one in ten chance of winning nothing in ten games, for instance; so you clearly need to have a stake (much) more than ten times the price to be even 90% sure of winning.

So I don't think I'd pay much more than **$10 a game**.

Problem 5. (10)

The Pareto distribution is used in economics modeling. To keep it simple we'll use the Pareto distribution that takes values $x \geq 1$ and has pdf

$$f(x \mid \theta) = \theta x^{-\theta-1} \quad \text{for } x \geq 1.$$

It's defined whenever $\theta > 0$. Assume $x_1, \dots, x_n$ are n independent samples from a Pareto(θ) distribution, find the maximum likelihood estimate of θ.

Solution: The likelihood for these n independent samples is the product of the densities at all the x_i:

$$[\theta x_1^{-\theta-1}][\theta x_2^{-\theta-1}]\cdots[\theta x_n^{-\theta-1}] = \theta^n (x_1 \cdots x_n)^{-\theta-1}$$

The log likelihood is

$$n\log(\theta) + \log(x_1 \cdots x_n)(-\theta - 1).$$

The MLE is the choice of $\theta > 0$ that maximizes this function. The function approaches $-\infty$ as θ approaches 0 or ∞ (because all the x_i need to be at least 1 to have a chance of being modelled by Pareto; this makes $\log(x_1 \cdots x_n) > 0$). So any maximum must occur when the derivative with respect to θ is zero. That is,

$$n/\theta - \log(x_1 \cdots x_n) = 0,$$

which gives

$$\boxed{\theta = \frac{n}{\log(x_1 \cdots x_n)} = \frac{n}{\sum_i \log(x_i)}}.$$

Problem 6. (15: 5,10)
A certain medical condition exists in 1% of the population. A screening test for the condition has a 4% false positive rate and a 0% false negative rate.

(a) *What are the odds that a random person has the condition?*

Solution: Let's let $D+$ = having the condition; $D-$ = not having it; $T+$ = testing positive.

The odds of having the condition are $\dfrac{P(D+)}{P(D-)} = \dfrac{1/100}{99/100} = \boxed{\dfrac{1}{99}.}$

(b) *Suppose a random person tests positive for the condition. What are the odds they have the condition?*

Solution: The odds after the evidence are the prior odds multiplied by the likelihood ratio. The likelihoods are

$$P(T+|D+) = 1.00, \quad P(T+|D-) = 0.04.$$

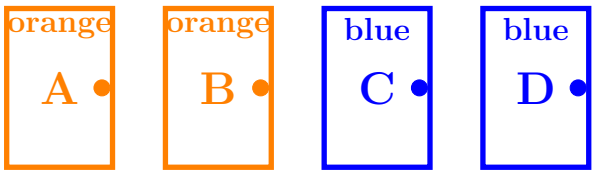
So, the likelihood ratio is $\dfrac{P(T+|D+)}{P(T+|D-)} = \dfrac{1}{0.04} = 25$. Thus, the posterior odds are

$$\text{prior odds} \cdot \text{likelihood ratio} = \boxed{\frac{1}{99} \cdot 25 = \frac{25}{99}} \approx 1/4.$$

After a positive test, the odds are about $\boxed{\textbf{1 to 4}}$ that the person has the condition (or about a 20% chance).

Problem 7. (30: 5,10,5,10)
This question is about a robot, Bayz-E, who tries to figure out where it is located using Bayesian updating. Bayz-E is randomly placed in front of one of four doors, A, B, C, or D. At every time step, Bayz-E scans the color of the door in front of it. The outcome will be either orange of blue. However, its color scanner is not entirely accurate. Bayz-E scans the door's color correctly with a probability 0.7, and scans the color incorrectly with a probability 0.3.

Note. *To avoid confusion for you, the color of the door is written above its letter. (Bayz-E has not yet learned to read, though its machine learning algorithm is working on it.)*



(a) *Bayz-E is placed in front of door A, B, C, or D at random (i.e. the probability of each door is the same). What is the prior probability it is in front of door B?*

Solution: There are four hypotheses, labeled A, B, C, D according to the door in front of which Bayz-E starts. Each of the four is equally likely, so the prior for B (or for any other door) is $\boxed{1/4}$.

(b) *Bayz-E scans the door color and detects "orange." What is the posterior probability it is in front of door B?*

Solution: Time to start making tables: We'll abbreviate Bayes numerator as BN.

		likelihood		
hypothesis	prior	P(orange | hypoth.)	BN	posterior
A (orange)	1/4	7/10	7/40	7/20
B (orange)	1/4	7/10	7/40	$\boxed{7/20}$
C (blue)	1/4	3/10	3/40	3/20
D (blue)	1/4	3/10	3/40	3/20
sum	1		1/2	1

So the posterior probability of B is $\boxed{P(B \,|\, \text{orange}) = 7/20}$.

(c) *After computing the posterior probabilities, Bayz-E moves to the next door to the right. (if it was at door D, it moves to door A). What is the (new) prior probability it is now in front of door C? (That is, after detecting orange in part (b) and after moving one to door to the right.)*

Solution: To keep things straight, we start a new table for the door Bayz-E is currently in front of, i.e. after moving. The priors are found by rotating the posteriors from the table in part (b), e.g. the part (b) posterior probability for A becomes the prior probability for B etc.

hypothesis	prior	likelihood	BN 2	posterior 2
A (orange)	3/20 –part (b) posterior for D			
B (orange)	7/20 –part (b) posterior for A			
C (blue)	7/20 –part (b) posterior for B			
D (blue)	3/20 –part (b) posterior for C			
sum	1			

$$\boxed{P(\text{now in front of } C \,|\, \text{orange on first scan}) = 7/20}.$$

(d) *What is the predictive probability that the color scan of this door (after the move in part (c)) will detect "blue"?*

Solution: We continue the table started in part (c).

hypothesis	prior	likelihood P(2nd is blue | hypoth.)	BN 2	posterior 2
A (orange)	3/20	3/10	9/200	
B (orange)	7/20	3/10	21/200	
C (blue)	7/20	7/10	49/200	
D (blue)	3/20	7/10	21/200	
sum	1		$\boxed{100/200}$	

The predictive probability is the total probability of blue, i.e. the sum of the Bayes numerator column:

$$\boxed{P(\text{blue on second door} \,|\, \text{orange on first door}) = 1/2}.$$

Problem 8. (20)

According to the Mars website, each packet of Milk Chocolate M&M's should contain 20% blue, 20% brown, 20% green, 15% orange, 15% red, and 10% yellow M&M's.

Alessandre decides to test this claim. She buys 20 packets of Milk Chocolate M&M's. Each packet has 50 M&M's, so Alessandre has a total of 1000 M&M's. She counts each color and observes the following counts of M&M's.

	blue	*brown*	*green*	*orange*	*red*	*yellow*	*total*
Observed count	*180*	*190*	*185*	*160*	*165*	*120*	*1000*

Run a hypothesis test at the 0.05 significance level to test whether the published Mars color distribution is correct. Carefully state what you are testing and your conclusion.

(Write down the full numerical expression for your test statistic. In order to use the tables you will need to estimate the value of the test statistic. There is no need to compute it in full precision.)

Solution: We want to do a **chi-squared test for goodness of fit**: we are testing whether the distribution published on the Mars website is a good fit with the observed counts. H_0 is that the data was drawn from the Mars' distribution and H_A is that it is drawn from another distribution.

Here's a table of the data and a bit of the computation.

	blue	brown	green	orange	red	yellow
obs. count	180	190	185	160	165	120
publ. freq.	0.20	0.20	0.20	0.15	0.15	0.10
exp. count	200	200	200	150	150	1-0
$(O-E)^2/E$	$\frac{20^2}{200} = 2$	$\frac{10^2}{200} = 0.5$	$\frac{15^2}{200} = 1.125$	$10^2/150 = 0.67$	$\frac{15^2}{150} = 1.5$	$\frac{200^2}{100} = 4$

The χ-squared statistic is the sum of the entries in the last row, namely

$$X^2 = 2 + 0.5 + 1.125 + 0.67 + 1.5 + 4 \approx \boxed{9.8}.$$

The number of degrees of freedom is the number of colors minus 1 (because we used a known total number of M&M's), or $\boxed{\mathbf{df} = 5}$. Using the table at the end of the test, we see that 9.8 corresponds to a p-value between 0.05 and 0.10. (Interpolating it is about 0.085, using R we get 0.081.) Since this is greater than 0.05, we do not reject the null hypothesis that the published distribution is correct. We will not blog that their website gives misleading information.

Problem 9. (20: 5,10,5)
Jerry wants to brag to his non-MIT colleagues about how smart MIT students are. To give himself credibility, he decides to run a statistical test comparing the IQ scores of MIT students and Harvard students.

He collects IQ scores from 11 MIT students. The data has a sample mean of 115, with a sample standard deviation of 8.

He then collects IQ scores from 11 Harvard students. Their scores have a sample mean of 110, with a sample standard deviation of 6.

(a) *Which test should he run to compare the IQ scores from the two schools? What assumptions will he need to make? What are the null and alternative hypotheses?*

Solution: We can run a two-sample t test for comparing means. This test assumes that the two sets of data are **drawn from normal distributions** (a reasonable assumption for the results of IQ tests, which are designed that way) with **equal variances** (a reasonable assumption since the sample standard deviations are not far apart). The null hypothesis is **means are equal**. The alternative hypothesis is **means are unequal**. (Can also argue that H_A should be one-sided.)

(b) *Run the test with a significance level of $\alpha = 0.05$. Should Jerry reject the null hypothesis or not?*

(In this problem there is some arithmetic. You will want to use $\sqrt{100/11} \approx \sqrt{9} = 3$.)

Solution: The t statistic is $t = (\overline{x} - \overline{y})/s_p$, where s_p^2 is the pooled variance

$$s_p^2 = \frac{10 \cdot 64 + 10 \cdot 36}{11 + 11 - 2}\left(\frac{1}{11} + \frac{1}{11}\right) = \frac{100}{11}.$$

According to the hint $s_p \approx 3$. The difference in means is 5, so $t \approx 5/3 = 1.67$. The number of degrees of freedom is $11+11-2 = 20$. According to the table, the right tail probability for the t distribution with 20 degrees of freedom at 1.67 is between 0.05 and 0.10, (interpolating it's about 0.058). The two-tailed p-value is twice the right tail probability, (about 0.116), which is clearly greater than $\alpha = 0.05$. So, we **do not reject** the null hypothesis, i.e., the difference between the mean IQ scores of MIT and Harvard students is not signifcant at significance level 0.05.

(c) *Estimate the 95% confidence interval for the IQ of MIT students.*

(Your answer should be a numerical expression. There is no need to work it out all the way to a decimal answer.)

Solution: The 95% confidence intervals is

$$\overline{x} \pm t_{0.025} \cdot \frac{s}{\sqrt{n}} = 115 \pm 2.23 \cdot \frac{8}{\sqrt{11}}.$$

Here $t_{0.025}$ is the right t-critical value with 10 degrees of freedom.

Problem 10. (20: 10,10)
MIT has decided to form a new Department of Statistics and Probability. In a vote for the new head of this department, suppose 50% of the MIT population supports Sarah, 20%

supports So Hee, and the remaining 30% is split evenly among Jerry, Jen, Alessandre and Gabe.

(a) *A poll asks 100 random people whom they support. Estimate the probability that at least 45% of those polled support Sarah.*

Solution: Write S for the fraction of the sample that supports Sarah.

$$E[S] = 0.5, \quad \sigma_S = \frac{1}{2\sqrt{n}} = \frac{1}{20}$$

Since S is an average of 100 Bernoulli(0.5) variables, the CLT says it's approximately normal. Standardizing gives

$$P(S > 0.45) = P\left(\frac{S - 0.5}{1/20} > \frac{0.45 - 0.50}{1/20}\right) \approx P(Z > -1) \approx \boxed{0.84.}$$

(b) *A poll of n people reports that 53% $\pm$ 5% support Sarah at the 95% confidence level. What is the value of n?*

Solution: The rule of thumb for 95% polling confidence intercals tells us that

$$0.05 = \frac{1}{\sqrt{n}} \;\Rightarrow\; \sqrt{n} = 20 \;\Rightarrow\; \boxed{n = 400.}$$

If we use the more precise critical z of 1.96 (for 95%) instead of the 2 in the rule of thumb, we get $\boxed{n = 385}$.

Problem 11. (10)
You independently draw 100 data points from a $N(\mu, 1)$ distribution, where μ is unknown. Suppose you test the null hypothesis $H_0 : \mu = 0$ against the alternative hypothesis $H_A : \mu \neq 0$ using a significance level of $\alpha = 0.05$. What is the power of the test for the alternative $H_A : \mu = 0.4$?

Solution: Power $= P(\text{reject} \mid \mu = 0.4) = 1 - P(\text{don't reject} | \mu = 0.4)$.

Rejection region for $\overline{x}$ is

$$z_{0.975} \cdot \frac{\sigma}{\sqrt{n}} \leq \overline{x} - \mu_0 \leq z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}.$$

We have $\mu_0 = 0, \quad \sigma = 1, \quad n = 100, \quad \alpha = 0.05$. So, the rejection region is

$$-1.96 \cdot \frac{1}{10} \leq \overline{x} \leq 1.96 \cdot \frac{1}{10} \;\Leftrightarrow\; -0.196 \leq \overline{x} \leq 0.196.$$

So, $P(\text{don't reject} | \mu = 0.4) = P(-0.196 \leq \overline{x} \leq 0.196 \,|\, \mu = 0.4)$.

Standardizing $\overline{x}$ this becomes

$$\begin{aligned} P(\text{don't reject} \,|\, \mu = 0.4) &= P\left(\frac{-0.196 - 0.4}{1/10} \leq \frac{\overline{x} - 0.4}{1/10} \leq \frac{0.196 - 0.4}{1/10} \,|\, \mu = 0.4\right) \\ &= P(-5.96 \leq Z \leq -2.04). \end{aligned}$$

In the last expression Z is a standard normal random variable. Using the standard normal table, we get

$$P(\text{don't reject}) = \Phi(-2.04) - \Phi(-5.96) \approx 0.0207 - 0.$$

So $\boxed{\text{Power} = 1 - 0.0207 \approx 0.98.}$

Problem 12. (20: 5,10,5)
Bivariate data $(4,1), (-2,1.5), (0,0.5)$ is assumed to arise from the model $y_i = b|x_i - 2| + e_i$, where b is a constant and e_i are independent random variables.

(a) *What assumptions are needed on e_i so that it makes sense to do a least squares fit of a curve $y = b|x-2|$ to the data?*

Solution: We need to know that the error terms e_i have mean 0. It is good for them to be independent and to have the same variance (homoscedastic). It is wonderful if they are independent identically distributed normal of mean zero.

(b) *Given the above data and the assumptions from part (a), determine the least squares estimate for b.*

Solution: The sum of the squares of the errors is

$$T = \sum_{i=1}^{3} (y_i - b|x_i - 2|)^2.$$

So, setting the derivative to 0 give

$$\frac{dT}{db} = \sum_{i=1}^{3} -2|x_i - 2|(y_i - b|x_i - 2|) = 0.$$

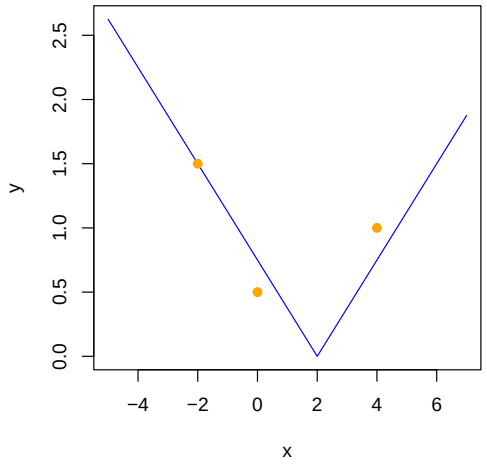
Solving we get

$$b\sum_{i=1}^{3} |x_i - 2|^2 = \sum_{i=1}^{3} y_i|x_i - 2| \;\Leftrightarrow\; b = \frac{\sum_{i=1}^{3} y_i|x_i - 2|}{\sum_{i=1}^{3} |x_i - 2|^2}.$$

Plugging in numbers we find $\boxed{b = 3/8.}$

(c) *Make a graph showing the data points and your least squares fit curve.*

Solution:



Problem 13. (15)

Data is collected on the time between trades at a stock exchange. We collect a data set of size 36 *with sample mean* $\bar{x} = 7.0$ *and sample standard deviation* $s = 0.84$.

Make no assumptions about the distribution of the data. By bootstrapping, we generate 500 bootstrap means $\overline{x}^*$. *The smallest 50 and largest 50 are written in non-decreasing order below, e.g. the 12th smallest value is 6.672.*

Use this data to find an 90% percentile bootstrap confidence interval for μ.

Solution: The 90% percentile bootstrap CI is $[q^*_{0.05}, q^*_{0.95}]$, where $q^*_{0.05}$ and $q^*_{0.95}$ are empirical quantiles for $\overline{x}^*$.

$q^*_{0.05}$ = 25th element = 6.740. $q^*_{0.95}$ = 475th element = 7.224

So $\boxed{\textbf{the 90\% CI } = [6.740,\ 7.224]}$.

1- 10	*6.466*	*6.506*	*6.509*	*6.515*	*6.578*	*6.597*	*6.618*	*6.635*	*6.653*	*6.664*
11- 20	*6.670*	*6.672*	*6.685*	*6.696*	*6.703*	*6.707*	*6.713*	*6.721*	*6.727*	*6.727*
21- 30	*6.729*	*6.731*	*6.738*	*6.738*	*6.740*	*6.743*	*6.744*	*6.745*	*6.751*	*6.752*
31- 40	*6.759*	*6.760*	*6.768*	*6.774*	*6.775*	*6.777*	*6.778*	*6.780*	*6.784*	*6.784*
41- 50	*6.787*	*6.789*	*6.789*	*6.790*	*6.791*	*6.791*	*6.792*	*6.796*	*6.798*	*6.800*
451- 460	*7.170*	*7.172*	*7.172*	*7.175*	*7.178*	*7.179*	*7.180*	*7.181*	*7.182*	*7.182*
461- 470	*7.182*	*7.186*	*7.195*	*7.202*	*7.202*	*7.205*	*7.206*	*7.210*	*7.216*	*7.219*
471- 480	*7.220*	*7.220*	*7.221*	*7.222*	*7.224*	*7.225*	*7.232*	*7.232*	*7.236*	*7.236*
481- 490	*7.243*	*7.244*	*7.245*	*7.251*	*7.253*	*7.258*	*7.261*	*7.263*	*7.266*	*7.273*
491- 500	*7.274*	*7.288*	*7.288*	*7.291*	*7.307*	*7.312*	*7.314*	*7.316*	*7.348*	*7.488*

18.05 Introduction to Probability and Statistics
Spring 2022