Practice Final –solutions, 18.05, Spring 2022

Concept Problem 1. *Which of the following represents a valid probability table?*

(i)

outcomes	*1*	*2*	*3*	*4*	*5*
probability	*1/5*	*1/5*	*1/5*	*1/5*	*1/5*

(ii)

outcomes	*1*	*2*	*3*	*4*	*5*
probability	*1/2*	*1/5*	*1/10*	*1/10*	*1/10*

Circle the best choice:

A. *(i)* ***B.*** *(ii)* ***C.*** *(i) and (ii)* ***D.*** *Not enough information*

Solution: C. (i) and (ii)

Concept Problem 2. *True or false: Setting the prior probability of a hypothesis to 0 means that no amount of data will make the posterior probability of that hypothesis the maximum over all hypotheses.*

Circle one: **True** **False**

Solution: True

Concept Problem 3. *True or false: It is okay to have a prior that depends on more than one unknown parameter.*

Circle one: **True** **False**

Solution: True

Concept Problem 4. *Data is drawn from a normal distribution with unknown mean μ. We make the following hypotheses:* H_0*:* $\mu = 1$ *and* H_A*:* $\mu > 1$.

For (i)-(iii) circle the correct answers:

(i) *Is H_0 a simple or composite hypothesis?* ***Simple*** ***Composite***

(ii) *Is H_A a simple or composite hypothesis?* ***Simple*** ***Composite***

(iii) *Is H_A a one or two-sided?* ***One-sided*** ***Two-sided***

Solution: (i) Simple (ii) Composite (iii) One-sided

Concept Problem 5. *If the original data has n points then a bootstrap sample should have*

A. *Fewer points than the original because there is less information in the sample than in the underlying distribution.*
B. *The same number of points as the original because we want the bootstrap statistic to mimic the statistic on the original data.*
C. *Many more points than the original because we have the computing power to handle a lot of data.*

Circle the best answer: ***A*** ***B*** ***C.***

Solution: B.

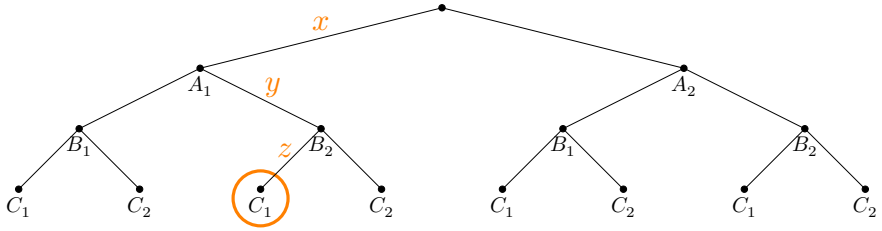
Concept Problem 6. *In 3 tosses of a coin which of following equals the event "exactly two heads"?*

$A = \{THH, HTH, HHT, HHH\}$
$B = \{THH, HTH, HHT\}$
$C = \{HTH, THH\}$

Circle the best answer: ***A*** ***B*** ***C*** ***B and C***

Solution: 2. B

Concept Problem 7. *These questions all refer to the following figure. For each one circle the best answer.*



(i) *The probability x represents* ***A.*** $P(A_1)$ ***B.*** $P(A_1|B_2)$ ***C.*** $P(B_2|A_1)$ ***D.*** $P(C_1|B_2 \cap A_1)$.

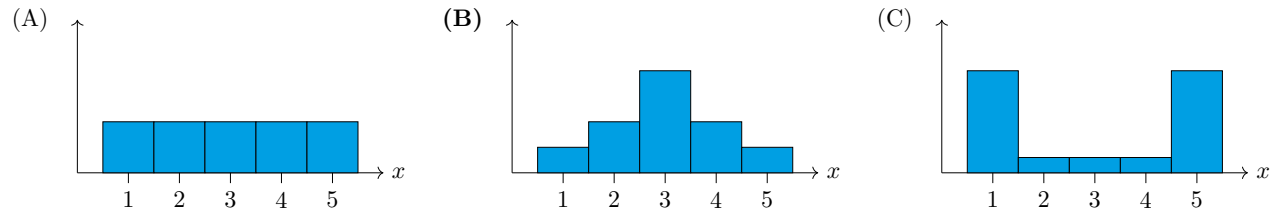
Solution: A. $P(A_1)$.

(ii) *The probability y represents* ***A.*** $P(B_2)$ ***B.*** $P(A_1|B_2)$ ***C.*** $P(B_2|A_1)$ ***D.*** $P(C_1|B_2 \cap A_1)$.
Solution: C. $P(B_2|A_1)$.

(iii) *The probability z represents* ***A.*** $P(C_1)$ ***B.*** $P(B_2|C_1)$ ***C.*** $P(C_1|B_2)$ ***D.*** $P(C_1|B_2 \cap A_1)$.
Solution: D. $P(C_1|B_2 \cap A_1)$.

(iv) *The circled node represents the event* ***A.*** C_1 ***B.*** $B_2 \cap C_1$ ***C.*** $A_1 \cap B_2 \cap C_1$ ***D.*** $C_1|B_2 \cap A_1$.
Solution: C. $A_1 \cap B_2 \cap C_1$.

Concept Problem 8. *The graphs below give the pmf for 3 random variables.*



Circle the answer that orders the graphs from smallest to biggest standard deviation.

ABC ***ACB*** ***BAC*** ***BCA*** ***CAB*** ***CBA***

Solution: BAC.

Concept Problem 9. *Suppose you have \$100 and you need \$1000 by tomorrow morning. Your only way to get the money you need is to gamble. If you bet \$k, you either win \$k with probability p or lose \$k with probability $1 - p$. Here are two strategies:*

Maximal strategy: *Bet as much as you can, up to what you need, each time.*

Minimal strategy: *Make a small bet, say \$10, each time.*

Suppose $p = 0.8$.

Circle the better strategy: ***Maximal*** ***2. Minimal***

Solution: $p = 0.8$ use minimal strategy.

If you use the minimal strategy the law of large numbers says your average winnings per bet will almost certainly be the expected winnings of one bet.

Win	-10	10
p	0.2	0.8

The expected value when $p = 0.8$ is 6. Since this is positive you'd like to make a lot of bets and let the law of large numbers (practically) guarantee you will win an average of \$6 per bet. So you use the minimal strategy.

Concept Problem 10. *Consider the following joint pdf's for the random variables X and Y. Circle the ones where X and Y are independent and cross out the other ones.*

A. $f(x,y) = 4x^2y^3$ ***B.*** $f(x,y) = \frac{1}{2}(x^3y + xy^3)$. ***C.*** $f(x,y) = 6e^{-3x-2y}$

Solution:

A. Independent. The variables can be separated: the marginal densities are $f_X(x) = ax^2$ and $f_Y(y) = by^3$ for some constants a and b with $ab = 4$.

B. Not independent. X and Y are not independent because there is no way to factor $f(x,y)$ into a product $f_X(x)f_Y(y)$.

C. Independent. The variables can be separated: the marginal densities are $f_X(x) = ae^{-3x}$ and $f_Y(y) = be^{-2y}$ for some constants a and b with $ab = 6$.

Concept Problem 11. *Suppose $X \sim$ Bernoulli(θ) where θ is unknown. Which of the following is the correct statement?*

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

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

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

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

Circle the letter of the correct statement: ***A*** ***B*** ***C*** ***D***

Solution: B. A Bernoulli random variable takes values 0 or 1. So X is discrete. The parameter θ can be anywhere in the continuous range [0,1]. Therefore the space of hypotheses is continuous.

Concept Problem 12. *Let θ be the probability of heads for a bent coin. Suppose your prior $f(\theta)$ is Beta(6, 8). Also suppose you flip the coin 7 times, getting 2 heads and 5 tails. What is the posterior pdf $f(\theta|x)$? Circle the best answer.*

A. *Beta(2,5)* ***B.*** *Beta(3,6)* ***C.*** *Beta(6,8)* ***D.*** *Beta(8,13)* ***E.*** *Not enough information to say*

Solution: D. By the form of the posterior pdf we know it is Beta$(8, 13)$.

Concept Problem 13. *Suppose the prior has been set. Let x_1 and x_2 be two sets of data. Circle true or false for each of the following statements.*

A. *If x_1 and x_2 have the same likelihood function then they result in the same posterior.* ***True*** ***False***

B. *If x_1 and x_2 result in the same posterior then they have the same likelihood function.* ***True*** ***False***

C. *If x_1 and x_2 have proportional likelihood functions then they result in the same posterior.* ***True*** ***False***

Solution: A. True, B. False C. True

Concept Problem 14. *Each day Jane arrives X hours late to class, with $X \sim$ uniform$(0, \theta)$. Jon models his initial belief about θ by a prior pdf $f(\theta)$. After Jane arrives x hours late to the next class, Jon computes the likelihood function $f(x|\theta)$ and the posterior pdf $f(\theta|x)$.*

Circle the probability computations a frequentist would consider valid. Cross out the others.

A. *prior* ***B.*** *posterior* ***C.*** *likelihood*

Solution: A. Not valid B. not valid C. valid

Both the prior and posterior measure a belief in the distribution of hypotheses about the value of θ. The frequentist does not consider them valid.

The likelihood $f(x|\theta)$ is perfectly acceptable to the frequentist. It represents the probability of data from a repeatable experiment, i.e. measuring how late Jane is each day. Conditioning on θ is fine. This just fixes a model parameter θ. It doesn't require computing probabilities for θ.

Concept Problem 15. *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. unknown. 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 Problem 16. *Consider the following statements about a 95% confidence interval for a parameter θ.*

A. *$P(\theta_0$ is in the CI $\mid \theta = \theta_0) \geq 0.95$*

B. *$P(\theta_0$ is in the CI $) \geq 0.95$*

C. *An experiment produces the CI $[-1, 1.5]$: $P(\theta$ is in $[-1, 1.5] \mid \theta = 0) \geq 0.95$*

Circle the letter of each correct statement and cross out the others:

A **B** **C**

A. **Solution:** Correct, This is the definition of a confidence interval.

B. **Solution:** Incorrect. Frequentist methods do not give probabilities for hypotheses.

C. **Solution:** Correct. Given $\theta = 0$ the probability θ is in [-1, 1.5] is 100%.

Problem 17. ***(a)*** *Let A and B be two events. Suppose that the probability that neither event occurs is 3/8. What is the probability that at least one of the events occurs?*

(b) *Let C and D be two events. Suppose $P(C) = 0.5$, $P(C \cap D) = 0.2$ and $P((C \cup D)^c) = 0.4$. What is $P(D)$?*

(a) Solution: $P((A \cup B)^c) = 3/8 \Rightarrow \boxed{P(A \cup B) = 5/8}$..

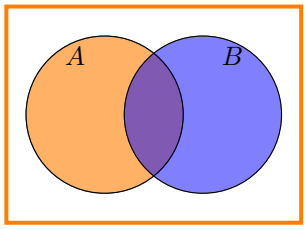


Figure for part (a).

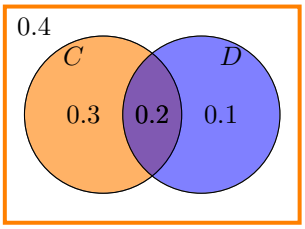


Figure for part (b).

(b) Solution: See the figure: $P((CUD)^c) = 0.4 \Rightarrow P((CUD) = 0.6)$.

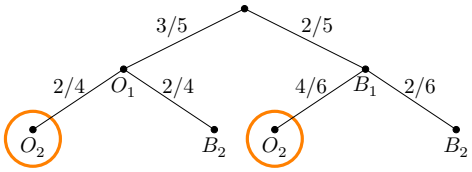
$P(C \cup D) = P(C) + P(D) - P(C \cap D) \Rightarrow 0.6 = 0.5 + P(D) - 0.2 \Rightarrow \boxed{P(D) = 0.3}$.

Problem 18. *An urn contains 3 orange balls and 2 blue balls. A ball is drawn. If the ball is orange, it is kept out of the urn and a second ball is drawn from the urn. If the ball is blue, then it is put back in the urn and an orange ball is added to the urn. Then a second ball is drawn from the urn.*

(a) *What is the probability that both balls drawn are orange?*

(b) *If the second drawn ball is orange, what is the probability that the first drawn ball was blue?*

Solution: Let O_1 be the event the first ball is orange and O_2 that the second ball is orange. Likewise for B_1 and B_2. The following tree captures all the details of the game.



(a) $P(O_1 \cap O_2) = \dfrac{3}{5} \cdot \dfrac{2}{4} = \boxed{\dfrac{6}{20} = 0.3}$.

(b) $P(B_1|O_2) = \dfrac{P(O_2|B_1)P(B_1)}{P(O_2)} = \dfrac{\frac{4}{6} \cdot \frac{2}{5}}{\frac{2}{4} \cdot \frac{3}{5} + \frac{4}{6} \cdot \frac{2}{5}} = \dfrac{8/30}{17/30} = \dfrac{8}{17}$.

Problem 19. *You roll a fair six sided die repeatedly until the sum of all numbers rolled is greater than 6. Let X be the number of times you roll the die. Let F be the cumulative distribution function for X. Compute $F(1)$, $F(2)$, and $F(7)$.*

Solution: $F(1)$: Since you never get more than 6 on one roll we have $\boxed{F(1) = 0}$.

$F(2) = P(X = 1) + P(X = 2)$:

$P(X = 1) = 0$

$P(X = 2) = P(\text{total on 2 dice} = 7,8,9,10,11,12) = \boxed{\frac{21}{36} = \frac{7}{12}}$.

$F(7)$: The smallest total on 7 rolls is 7, so $\boxed{F(7) = 1}$.

Problem 20. *A test is graded on the scale 0 to 1, with 0.55 needed to pass. Student scores are modeled by the following density:*

$$f(x) = \begin{cases} 4x & \text{for } 0 \le x \le 1/2 \\ 4 - 4x & \text{for } 1/2 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) *What is the probability that a random student passes the exam?*

(b) *What score is the 87.5 percentile of the distribution?*

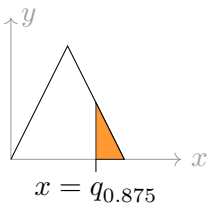
(a) Solution: Let X = score of a random student.

$$P(X \ge 0.55) = \int_{0.55}^{1} f(x)\,dx = \int_{0.55}^{1} 4 - 4x\,dx = 4x - 2x^2\Big|_{0.55}^{1} = 2 - 4 \times 0.55 + 2(0.55)^2 = 0.405$$

(b) Solution: *Geometric method:*

We need the shaded area in the figure to be 0.125
Shaded area = area of triangle = $\frac{1}{2}(1-x)(4-4x) = 0.125$.
Solving for x we get

$$2(1-x)^2 = 0.125 \;\Rightarrow\; (1-x)^2 = \frac{1}{16} \;\Rightarrow\; \boxed{x = \frac{3}{4}}.$$



Analytic mehtod: We want a such that $F(a) = 7/8$. Since $f(x)$ is defined in two pieces we have to compute $F(a)$ in two pieces.

$$F(1/2) = \int_0^{1/2} 4x\,dx = 2x^2\Big|_0^{1/2} = \frac{1}{2}.$$

(Which we knew geometrically already.)

For $a \ge 1/2$ we then have

$$\begin{aligned} F(a) &= \int_0^{1/2} 4x\,dx + \int_{1/2}^{a} 4 - 4x\,dx \\ &= \frac{1}{2} + \int_{1/2}^{a} 4 - 4x\,dx \\ &= \frac{1}{2} + \left[4x - 2x^2\right|_{1/2}^{a} \\ &= 4a - 2a^2 - 1. \end{aligned}$$

Solving for a such that $F(a) = 7/8$ we get

$$4a - 2a^2 - 1 = 7/8 \;\Rightarrow\; 2a^2 - 4a + 15/8 = 0 \;\Rightarrow\; a = \frac{4 \pm \sqrt{1}}{4} = \frac{3}{4}, \frac{5}{4}.$$

Since $\frac{5}{4}$ is not in the range of X we have $\boxed{a = 3/4}$. (The same answer as with the geometric method.)

Problem 21. *Suppose X is a random variable with cdf*

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x(2-x) & \text{for } 0 \le x \le 1 \\ 1 & \text{for } x > 1 \end{cases}$$

(a) *Find $E[X]$.*

(b) *Find $P(X < 0.4)$.*

(a) Solution: $f(x) = F'(x) = 2 - 2x$ on $[0, 1]$. Therefore

$$\begin{aligned} E[X] &= \int_0^1 x f(x)\, dx \\ &= \int_0^1 2x - 2x^2\, dx \\ &= x^2 - \frac{2}{3}x^3 \Big|_0^1 \\ &= \boxed{\frac{1}{3}}. \end{aligned}$$

(b) Solution: $P(X \le 0.4) = F(0.4) = 0.4(2 - 0.4) = 0.4(1.6) = \boxed{0.64}$.

Problem 22. *Compute the mean and variance of a random variable whose distribution is uniform on the interval $[a, b]$.*

It is not enough to simply state these values. You must give the details of the computation.

Solution: Let $X \sim \mathrm{U}(a, b)$. The pdf of X is $f(x) = \dfrac{1}{b-a}$ on the interval $[a, b]$. Thus,

$$E[X] = \int_a^b x f(x)\, dx = \int_a^b \frac{x}{b-a}\, dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \boxed{\frac{b+a}{2}}$$

$$\begin{aligned}
\mathrm{Var}(X) &= \int_a^b (x-\mu)^2 f(x)\,dx \\
&= \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a}\,dx \\
&= \frac{\left(x-\frac{a+b}{2}\right)^3}{3}\frac{1}{b-a}\Bigg|_a^b \\
&= \text{... algebra ...} \\
&= \frac{1}{12}(b-a)^3\frac{1}{b-a} \\
&= \boxed{\frac{(b-a)^2}{12}}.
\end{aligned}$$

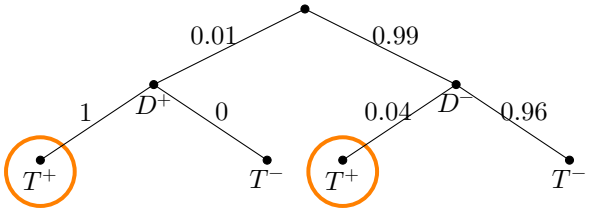
Problem 23. *Defaulting on a loan means failing to pay it back on time. The default rate among MIT students on their student loans is 1%. As a project you develop a test to predict which students will default. Your test is good but not perfect. It gives 4% false positives, i.e. prediciting a student will default who in fact will not. If has a 0% false negative rate, i.e. prediciting a student won't default who in fact will.*

***(a)* Solution:** *Suppose a random student tests positive. What is the probability that he will truly default.*

***(b)* Solution:** *Someone offers to bet me the student in part (a) won't default. They want me to pay them \$100 if the student doesn't default and they'll pay me \$400 if the student does default. Is this a good bet for me to take?*

(a) Solution: We organize the problem in a tree. Here:

D^+ = default, D^- = no default
T^+ = test is positive, T^- = test is negative



$$P(D^+|T^+) = \frac{P(T^+|D^+)P(D^+)}{P(T^+)} = \frac{0.01}{0.01 + 0.99\cdot 0.04} = \boxed{\frac{0.01}{0.0496} = \frac{1}{4.96} \approx 0.2}.$$

(b) Solution: $\mathrm{Odds}(\text{winning}) = \mathrm{Odds}(D^+|T^+) = \dfrac{P(D^+|T^+)}{P(D^-|T^+)} = \dfrac{1/4.96}{3.96/4.96} = \dfrac{1}{3.96}.$

Since the payoff ratio $\dfrac{4}{1}$ is greater than 1/(odds of winning), it is a good bet.

Equivalently we can argue the

$$E[\text{winnings}] = 400\cdot\frac{1}{4.96} - 100\cdot\frac{3.96}{4.96} = \frac{4}{4.96} > 0.$$

A positive expected winnings means it's a good bet.

Problem 24. *Data was taken on height and weight from the entire population of 700 mountain gorillas living in the Democratic Republic of Congo:*

ht\wt	*light*	*average*	*heavy*
short	*170*	*70*	*30*
tall	*85*	*190*	*155*

Let X encode the weight, taking the values of a randomly chosen gorilla: 0, 1, 2 for light, average, and heavy respectively.

Likewise, let Y encode the height, taking values 0 and 1 for short and tall respectively.

(a) *Determine the joint pmf of X and Y and the marginal pmf's of X and of Y.*

(b) *Are X and Y independent?*

(c) *Find the covariance of X and Y.*

For this part, you need a numerical (no variables) expression, but you can leave it unevaluated.

(d) *Find the correlation of X and Y.*

For this part, you need a numerical (no variables) expression, but you can leave it unevaluated.

(a) Solution: Probability table:

$Y \backslash X$	0	1	2	marginal for Y
0	170/700	70/700	30/700	270/700
1	85/700	190/700	155/700	430/700
marginal for X	255/700	260/700	185/700	1

(b) Solution: We check if $P(X=0, Y=0) = P(X=0)P(Y=0)$.

$$\frac{170}{700} \stackrel{?}{=} \frac{255}{700}\frac{270}{700}.$$

Cross-multiply and do a little algebra

$$170 \cdot 700 \stackrel{?}{=} 255 \cdot 270 \quad \Leftrightarrow \quad 11900 \stackrel{?}{=} \quad \Leftrightarrow \; 11900 \stackrel{?}{=} 68850$$

Since they are not equal $\boxed{X \text{ and } Y \text{ are not independent.}}$

(c) Solution:

$$E[X] = \frac{260}{700} + 2 \cdot \frac{185}{700} = \frac{630}{700} = \frac{9}{10}$$
$$E[Y] = \frac{430}{700} = \frac{43}{70}$$
$$E[XY] = \frac{190}{700} + 2 \cdot \frac{155}{700} = \frac{500}{700} = \frac{5}{7}$$
$$\mathrm{Cov}(X,Y) - E[XY] - E[X]E[Y] = \frac{5}{7} - \frac{9}{10} \cdot \frac{43}{70} = \frac{113}{700}$$

(d) The definition of correlation is $\text{Cor}(X,Y) = \dfrac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$. So we first need to compute the variances of X and Y.

$$E[X^2] = \frac{260}{700} + 4 \cdot \frac{185}{700} = \frac{1000}{700} = \frac{10}{7}$$

Thus,

$$\begin{aligned} \text{Var}(X) &= E[X^2] - E[X]^2 = \frac{10}{7} - \frac{81}{100} = \frac{433}{700} \\ E[Y^2] &= \frac{43}{70} \\ \text{Var}(Y) &= E[Y^2] - E[Y]^2 = \frac{43}{70} - \left(\frac{43}{70}\right)^2 = \frac{43 \cdot 27}{70^2} \end{aligned}$$

therefore

$$\text{Cor}(X,Y) = \frac{113/700}{\sqrt{433/700}\,\sqrt{43 \cdot 27/70^2}}$$

Note: We would accecpt –even encourage solutions– that left the fractions uncomputed, e.g. $\sigma_Y = \sqrt{43/70 - (43/70)^2}$.

Problem 25. *A political poll is taken to determine the fraction p of the population that would support a referendum requiring all citizens to be fluent in the language of probability and statistics.*

(a) *Assume $p = 0.5$. Use the central limit theorem to estimate the probability that in a poll of 25 people, at least 14 people support the referendum.*

Your answer to this problem should be a decimal.

Solution: Let $X \sim \text{binomial}(25, 0.5)$ = the number supporting the referendum. We know that

$$E[X] = 12.5, \qquad \text{Var}(X) = 25 \cdot \frac{1}{4} = \frac{25}{4}, \qquad \sigma_X = \frac{5}{2}.$$

Standardizing and using the CLT we have $Z = \dfrac{X - 12.5}{5/2} \approx \text{N}(0,1)$ Therefore,

$$P(X \geq 14) = P\left(\frac{X - 12.5}{5/2} \geq \frac{14 - 12.5}{5/2}\right) \approx P(Z \geq 0.6) = \Phi(-0.6) = \boxed{0.2743},$$

where the last probability was looked up in the Z-table.

(b) *With p unknown and n the number of random people polled, let $\overline{X}_n$ be the fraction of the polled people who support the referendum.*

What is the smallest sample size n in order to have a 90% confidence that $\overline{X}_n$ is within 0.01 of the true value of p?

Your answer to this problem should be an integer.

Solution: The rule of thumb CI is

$$\overline{x} \pm z_{0.05} \cdot \frac{1}{2\sqrt{n}}.$$

So we want $\dfrac{z_{0.05}}{2\sqrt{n}} \leq 0.01$.

From the table $z_{0.05} = \Phi(-0.05) = 1.65$. So we want

$$\frac{1.65}{2\sqrt{n}} \leq 0.01 \quad \Rightarrow \quad \sqrt{n} \geq \frac{165}{2} \quad \Rightarrow \quad n > (82.5)^2 = 6806.25$$

Solution: $\boxed{n = 6807}$

Problem 26. *Suppose a researcher collects $x_1, \ldots, x_n$ i.i.d. measurements of the background radiation in Boston. Suppose also that these observations follow a Rayleigh distribution with parameter τ, with pdf given by*

$$f(x) = x\tau e^{-\frac{1}{2}\tau x^2}.$$

Find the maximum likelihood estimate for τ.

Solution: For a fixed τ the pdf for x_i is $f(x_i \,|\, \tau) = x\tau e^{-\frac{1}{2}\tau x^2}$. Therefore the likelihood function of the data is

$$f(\text{data} \,|\, \tau) \;=\; x_1 x_2 \cdots x_n \tau^n \, e^{-\frac{1}{2}\tau \sum x_i^2}.$$

The log likelihood is

$$\ln(f(\text{data} \,|\, \tau)) \;=\; \ln(x_1 x_2 \cdots x_n) + n \ln(\tau) - \frac{1}{2}\tau \sum x_i^2.$$

We find the MLE for τ by taking a derivative of the log likelihood with respect to τ and setting equal to 0.

$$\frac{d \, \ln(f(\text{data} \,|\, \tau))}{d\tau} = \frac{n}{\tau} - \frac{1}{2}\sum x_i^2 = 0 \;\Rightarrow\; \frac{n}{\tau} = \frac{1}{2}\sum x_i^2 \;\Rightarrow\; \boxed{\tau = \frac{2n}{\sum x_i^2}.}$$

Problem 27. *Bivariate data $(4, 10), (-1, 3), (0, 2)$ is assumed to arise from the model $y_i = b|x_i - 3| + e_i$, 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 $y = b|x - 3|$ to the data?*

(b) *Given the above data, determine the least squares estimate for b.*

For this problem we want you to calculate all the way to a fraction $b = \frac{r}{s}$, where r and s are integers.

(a) Solution: We assume the random error terms e_i are independent, have mean 0 and all have the same variance (homoscedastic).

(b) Solution:

$$\begin{aligned} E[b] &= \text{sum of the squared errors} \\ &= \sum (y_i - b|x_i - 3|)^2 \\ &= (10 - b)^2 + (3 - 4b)^2 + (2 - 3b)^2 \end{aligned}$$

The least squares fit is found by setting the derivative (with respect to b) to 0,

$$\frac{d\,E[b]}{db} \;=\; -2(10-b)-8(3-4b)-6(2-3b) \;=\; 52b-56=0.$$

Therefore the least squares estimate of b is $\boxed{\hat{b}=\dfrac{56}{52}=\dfrac{14}{13}}$.

Problem 28. Taxi problem *Data is collected on the time between arrivals of consecutive taxis at a downtown hotel. We collect a data set of size 45 with sample mean $\overline{x}=5.0$ and sample standard deviation $s=4.0$.*

(a) *Assume the data follows a normal random variable.*

(i) Find an 80% confidence interval for the mean μ of X.

(ii) Find an 80% χ^2-confidence interval for the variance?

(b) *Now make no assumptions about the distribution of of the data. By bootstrapping, we generate 500 values for the average inter-arrival time $\overline{x}^*$. The smallest and largest 150 are written in non-decreasing order on the next page.*

Use this data to find an 80% percentile bootstrap confidence interval for μ.

(c) *We suspect that the time between taxis is modeled by an exponential distribution, not a normal distribution. In this case, are the approaches in the earlier parts justified?*

(d) *When might method (b) be preferable to method (a)?*

(a) Solution: Since σ is unknown we use the Studentized mean

$$t=\frac{\overline{x}-\mu}{s/\sqrt{n}}\sim t(44)$$

which follows a t distribution with 44 degrees of freedom.

(i) The 80% CI is $\overline{x}\pm t_{0.1}\,\dfrac{s}{\sqrt{n}}$. From the t-table we get $t_{0.1}$ with $df=44$ is approximately 1.3. Thus,

$$80\%\text{ CI }=\boxed{5\pm\frac{4}{\sqrt{45}}\cdot 1.3}$$

(ii) We use the statistic $\dfrac{(n-1)s^2}{\sigma^2}\sim\chi^2(44)$. The 80% confidence interval for σ^2 is

$$\left[\frac{(n-1)s^2}{c_{0.9}},\frac{(n-1)s^2}{c_{0.1}}\right],$$

where $c_{0.9}$ and $c_{0.1}$ are the right critical values from the chi-square distribution with 44 degrees of freedom.

$$80\%\text{ CI for }\sigma^2 \;=\; \left[\frac{(n-1)s^2}{56.37},\frac{(n-1)s^2}{32.49}\right] \;=\; \boxed{\left[\frac{44\cdot 16}{56.37},\frac{44\cdot 16}{32.49}\right]}$$

(b) Solution: The 80% percentile bootstrap CI is $[c^*_{0.1},c^*_{0.9}]$, where $c^*_{0.1}$ and $c^*_{0.9}$ are empirical quantiles for $\overline{x}^*$

$c^*_{0.1}$ is the 50th element = 4.800. (Really it is interporlated between the 49th and 50th element, fortunately both are 4.800)

$c^*_{0.9}$ is the 450th element = 5.169. (Really it is interporlated between the 449th and 450th element, fortunately both are 5.169)

So the 80% CI = $\boxed{[4.800, 5.169].}$

(c) Solution: The approach in (b) is fine since it makes no assumptions about the underlying distribution. The approach in (a) is more problematic since $\dfrac{\overline{x}-\mu}{s/\sqrt{n}}$ does not follow a Student-t distribution. However for an exponential distribution and $n = 45$ the approximation is not too bad.

(d) Solution: Method (b) is preferable if the sample mean $\overline{x}$ is not drawn from a normal distribution.

The 150 smallest and 150 largest values of $\overline{x}^*$ for taxi problem are given in the following table.

1- 10	4.466	4.506	4.509	4.515	4.578	4.597	4.618	4.635	4.653	4.664
11- 20	4.670	4.672	4.685	4.696	4.703	4.707	4.713	4.721	4.727	4.727
21- 30	4.729	4.731	4.738	4.738	4.740	4.743	4.744	4.745	4.751	4.752
31- 40	4.759	4.760	4.768	4.774	4.775	4.777	4.778	4.780	4.784	4.784
41- 50	4.787	4.789	4.789	4.790	4.791	4.791	4.792	4.796	4.800	4.800
51- 60	4.800	4.802	4.805	4.807	4.808	4.808	4.811	4.812	4.812	4.817
61- 70	4.818	4.818	4.819	4.821	4.821	4.822	4.824	4.825	4.826	4.830
71- 80	4.830	4.834	4.836	4.837	4.837	4.838	4.838	4.840	4.840	4.841
81- 90	4.841	4.841	4.842	4.843	4.844	4.844	4.845	4.845	4.846	4.846
91- 100	4.847	4.848	4.849	4.849	4.850	4.852	4.852	4.854	4.855	4.855
101- 110	4.856	4.858	4.858	4.858	4.862	4.863	4.865	4.865	4.866	4.866
111- 120	4.867	4.869	4.871	4.872	4.876	4.876	4.876	4.877	4.877	4.881
121- 130	4.882	4.886	4.886	4.886	4.888	4.889	4.891	4.892	4.892	4.893
131- 140	4.895	4.897	4.897	4.897	4.898	4.899	4.901	4.902	4.902	4.903
141- 150	4.903	4.904	4.905	4.905	4.905	4.907	4.907	4.907	4.907	4.907

351-360	5.073	5.074	5.075	5.075	5.077	5.077	5.077	5.077	5.078	5.079
361-370	5.079	5.079	5.080	5.081	5.081	5.082	5.083	5.084	5.085	5.085
371-380	5.087	5.087	5.088	5.091	5.091	5.091	5.092	5.092	5.093	5.093
381-390	5.094	5.094	5.096	5.097	5.100	5.100	5.101	5.101	5.102	5.103
391-400	5.104	5.104	5.106	5.106	5.108	5.108	5.108	5.108	5.108	5.110
401-410	5.110	5.111	5.112	5.112	5.112	5.112	5.113	5.114	5.114	5.115
411-420	5.118	5.122	5.122	5.123	5.127	5.129	5.129	5.132	5.134	5.134
421-430	5.134	5.135	5.136	5.136	5.137	5.140	5.141	5.142	5.142	5.143
431-440	5.143	5.145	5.146	5.147	5.147	5.148	5.151	5.151	5.154	5.155
441-450	5.156	5.162	5.163	5.164	5.164	5.165	5.166	5.168	5.169	5.169
451-460	5.170	5.172	5.172	5.175	5.178	5.179	5.180	5.181	5.182	5.182
461-470	5.182	5.186	5.195	5.202	5.202	5.205	5.206	5.210	5.216	5.219
471-480	5.220	5.220	5.221	5.222	5.224	5.225	5.232	5.232	5.236	5.236
481-490	5.243	5.244	5.245	5.251	5.253	5.258	5.261	5.263	5.266	5.273
491-500	5.274	5.288	5.288	5.291	5.307	5.312	5.314	5.316	5.348	5.488

Problem 29. **Note.** *In this problem the geometric(p) distribution is defined as the total number of trials to the first failure (the value includes the failure), where p is the probabilitiy of success.*

(a) *What sample statistic would you use to estimate p?*

(b) *Describe how you would use the parametric bootstrap to estimate a 95% basic confidence interval for p. You can be brief, but you should give careful step-by-step instructions.*

(a) Solution: Since $\mu = 1/(1-p)$, so $p = 1 - 1/\mu$, we should use the approximation $\boxed{\hat{p} = 1 - 1/\overline{x}}$.

(b) Solution: Step 1. Approximate p by $\hat{p} = 1 - 1/\overline{x}$.

Step 2. Generate a bootstrap sample $x_1^*, \dots, x_n^*$ from geom$(\hat{p})$.

Step 3. Compute $p^* = 1 - 1/\overline{x}^*$ and $\delta^* = p^* - \hat{p}$.

Repeat steps 2 and 3 many times (say 10^4 times.

Step 4. List all the δ^* and find the critical values.

Let $\delta^*_{0.025} = 0.025$ critical value $= 0.975$ quantile.

Let $\delta^*_{0.975} = 0.975$ critical value $= 0.025$ quantile.

Step 5. The basic bootstrap confidence interval is $[\hat{p} - \delta^*_{0.025},\ \hat{p} - \delta^*_{0.975}]$.

Problem 30. *You independently draw 100 data points from a normal distribution.*

(a) *Suppose you know the distribution is $N(\mu, 4)$ ($4 = \sigma^2$) and you want to test the null hypothesis $H_0 : \mu = 3$ against the alternative hypothesis $H_A : \mu \neq 3$.*

If you want a significance level of $\alpha = 0.05$. What is your rejection region?

You must clearly state what test statistic you are using.

(b) *Suppose the 100 data points have sample mean 5. What is the p-value for this data? Should you reject H_0?*

(c) *Determine the power of the test using the alternative $H_A : \mu = 4$.*

(a) Solution: We will use the standardized mean based on H_0 as a test statistic:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{x} - 3}{2/10} = 5(\overline{x} - 3).$$

At $\alpha = 0.05$ we reject H_0 if

$$z < z_{0.975} = -1.96 \quad \text{or} \quad z > z_{0.025} = 1.96.$$

(Or we could have used $\overline{x}$ as a test statistic and got the corresponding rejection region.)

(b) Solution: With this data we have $z = \frac{5-3}{2/10} = 10$. The rejection region is two sided so

$$p = P(|Z| > |z|) = P(|Z| > 10) = 0.$$

Yes, since $p < \alpha$ you should reject H_0.

(c) Solution: Power $= P(\text{reject} \mid \mu = 4)$

Our z-statistic is $z = \dfrac{\overline{x} - 3}{2/10}$ and we don't reject if

$$-1.96 \le z \le 1.96 \quad \Leftrightarrow \quad -1.96 \le \frac{\overline{x} - 3}{2/10} \le 1.96 \quad \Leftrightarrow \quad 2.61 \le \overline{x} \le 3.39$$

So,

$$\begin{aligned} \text{Power} &= P(\text{reject} \mid \mu = 4) \\ &= 1 - P(\text{don't reject} \mid \mu = 4) \\ &= 1 - P(2.61 < \overline{x} < 3.39 \mid \mu = 4) \end{aligned}$$

We standardize using the given mean $\mu = 4$

$$\begin{aligned} &= 1 - P\left(\frac{2.61 - 4}{2/10} < Z < \frac{-0.61}{2/10}\right) \\ &= 1 - P(-6.9 < Z < -3.05) \\ &= 1 - \Phi(-3.05) + \Phi(-6.9) \\ &= 1 - 0.0011 + 0 = \boxed{0.9989}. \end{aligned}$$

The probabilities were looked up in the z-table. We used $\Phi(-6.9) \approx 0$.

(We could have used much less calculation to find that the non-rejection range is $\overline{x}$ between $-7\sigma_{\overline{x}}$ and $-3\sigma_{\overline{x}}$ from the mean $\mu = 4$.)

Problem 31. *Suppose that you have molecular type with unknown atomic mass θ. You have an atomic scale with normally-distributed error of mean 0 and variance 0.5.*

(a) *Suppose your prior on the atomic mass is N(80, 4). If the scale reads 85, what is your posterior pdf for the atomic mass?*

(b) *With the same prior as in part (a), compute the smallest number of measurements needed so that the posterior variance is less than 0.01.*

(a) Solution: This is a normal/normal conjugate prior/likilihood update.

Hypothesis	Prior	Likelihood	Posterior
θ	$\mathrm{N}(80, 4)$	$f(x\|\theta) \sim \mathrm{N}(\theta, 0.5)$	$\mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$

We have

$$a = \frac{1}{\sigma^2_{\text{prior}}} = \frac{1}{4}, \qquad b = \frac{1}{\sigma^2} = \frac{1}{0.5} = 2.$$

For the update

$$\begin{aligned} \mu_{\text{post}} &= \frac{a\mu_{\text{prior}} + bx}{a + b} \\ &= \frac{80/4 + 170}{1/4 + 2} = \frac{760}{9} \approx 84.44 \\ \sigma^2_{\text{post}} &= \frac{1}{a + b} \\ &= \frac{1}{1/4 + 2} = \frac{4}{9} \approx 0.4444 \end{aligned}$$

So, the posterior is

$$f(\theta \,|\, x = 84) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}}) = \mathrm{N}(84.44, 0.4444)$$

(b) Solution: In this case $a = 1/4$, $b = n/0.5 = 2n$. We know

$$\sigma^2_{\text{post}} = \frac{1}{a+b} = \frac{1}{1/4 + 2n} = \frac{4}{8n+1}$$

Now $\sigma^2_{\text{post}} \leq 0.01$ gives us

$$\frac{4}{8n+1} \leq 0.01 \quad \Rightarrow \quad 400 \leq 8n+1 \quad \Rightarrow \quad \frac{399}{8} \leq n \quad \textbf{Solution:} \ \boxed{n = 50}.$$

Problem 32. *Your friend grabs a die at random from a drawer containing two 6-sided dice, one 8-sided die, and one 12-sided die. She rolls the die once and reports that the result is 7.*

(a) *Make a discrete Bayes table showing the prior, likelihood, and posterior for the type of die rolled given the data.*

(b) *What are your posterior odds that the die has 12 sides?*

(c) *Given the data of the first roll, what is your probability that the next roll will be a 7?*

(a) Solution: Let θ represent the number of sides to the die. The data is $x_1 = 7$

Hypothesis θ	prior $p(\theta)$	likelihood $p(x_1 = 7 \,|\, \theta)$	Bayes numer. $p(\theta)p(x_1 = 7 \,|\, \theta)$	posterior $p(\theta \,|\, x_1 = 7) = \dfrac{p(\theta)p(x_1 = 7 \,|\, \theta)}{p(x_1 = 7)}$
$\theta = 6$	$1/2$	0	0	0
$\theta = 8$	$1/4$	$1/8$	$1/32$	$3/5$
$\theta = 12$	$1/4$	$1/12$	$1/48$	$2/5$

(b) Solution: Odds $= \dfrac{p(\theta = 12 \,|\, x_1 = 7)}{p(\theta \neq 12 \,|\, x_1 = 7)} = \dfrac{2/5}{3/5} = \boxed{\dfrac{2}{3}}$.

(c) We extend the table in order to compute the posterior predictive probability.

θ	$p(\theta \,|\, x_1 = 7)$	$p(x_2 = 7 \,|\, \theta)$	$p(\theta \,|\, x_1 = 7)p(x_2 = 7 \,|\, \theta)$
$\theta = 6$	0	0	0
$\theta = 8$	$3/5$	$1/8$	$3/40$
$\theta = 12$	$2/5$	$1/12$	$2/60$
Total			**13/120**

The total probability $p(x_2 = 7 \,|\, x_1 = 7) = \boxed{\dfrac{13}{120}}$.

18.05 Introduction to Probability and Statistics
Spring 2022