Practice Final
18.05, Spring 2022

The test is divided into two parts. The first part is a series of concept questions. You don't need to show any work on this part. The second part consists of standard problems. You need to show your work on these.

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

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**

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**

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**

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.

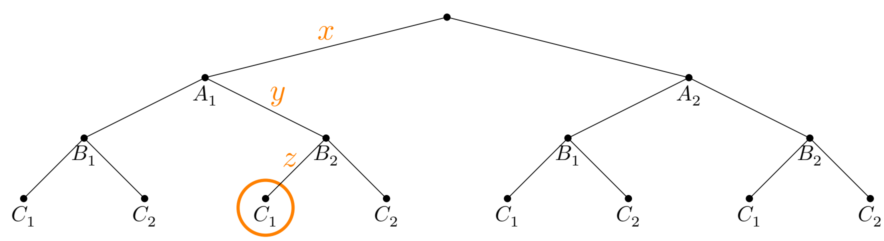
Circle the best answer: **A** **B** **C.**

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**

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



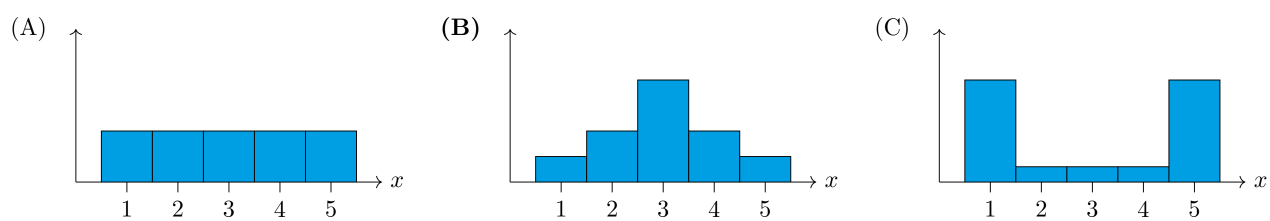
(i) The probability x represents **A.** $P(A_1)$ **B.** $P(A_1|B_2)$ **C.** $P(B_2|A_1)$ **D.** $P(C_1|B_2 \cap A_1)$.

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

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

(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.** $C_1|B_2 \cap A_1$.

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**

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**

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.** $f(x, y) = 6e^{-3x-2y}$

Concept Problem 11. Suppose $X \sim \text{Bernoulli}(\theta)$ 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**

Concept Problem 12. Let θ be the probability of heads for a bent coin. Suppose your prior $f(\theta)$ is $\text{Beta}(6, 8)$. 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

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**

Concept Problem 14. Each day Jane arrives X hours late to class, with $X \sim \text{uniform}(0, \theta)$. 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

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

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

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

B. $P(\theta_0 \text{ is in the CI }) \geq 0.95$

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

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

A **B** **C**

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 $P(D)$?

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?

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 $F(7)$.

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:

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

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

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

Problem 21. Suppose X is a random variable with cdf

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

(a) Find $E[X]$.

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

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.

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?

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.

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.

(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.

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

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

Find the maximum likelihood estimate for τ.

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.

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)?

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.

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

(a) Suppose you know the distribution is $\mathrm{N}(\mu, 4)$ $(4 = \sigma^2)$ 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 $H_A : \mu = 4$.

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.

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?