18.05 Final Exam Solutions

Part I: Concept questions (58 points)

These questions are all multiple choice or short answer. You don't have to show any work. Work through them quickly. Each answer is worth 2 points.

Concept 1. **answer:** C. (i) and (ii)

Concept 2. **answer:** True

Concept 3. **answer:** True

Concept 4. **answer:** (i) Simple (ii) Composite (iii) One-sided

Concept 5. **answer:** B.

Concept 6. **answer:** 2. B

Concept 7. **(i)** **answer:** A. $P(A_1)$.

(ii) **answer:** C. $P(B_2|A_1)$.

(iii) **answer:** D. $P(C_1|B_2 \cap A_1)$.

(iv) **answer:** C. $A_1 \cap B_2 \cap C_1$.

Concept 8. **answer:** BAC.

Concept 9. **answer:** $p = 0.8$ 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 10. **answer:**

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 $f_X(x) = a\mathrm{e}^{-3x}$ and $f_Y(y) = b\mathrm{e}^{-2y}$ for some constants a and b with $ab = 6$.

Concept 11. <u>**answer:**</u> 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 12. <u>**answer:**</u> D. By the form of the posterior pdf we know it is beta(8, 13).

Concept 13. A. True, B. False C. True

Concept 14. <u>**answer:**</u> 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 15. <u>**answer:**</u> 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 16. A. Correct, This is the definition of a confidence interval.

B. Incorrect. Frequentist methods do not give probabilities for hypotheses.

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

Part II: Problems (325 points)

Problem 1. (20) **(a)** $P((A \cup B)^c) = 3/8 \Rightarrow \boxed{P(A \cup B) = 5/8}$..

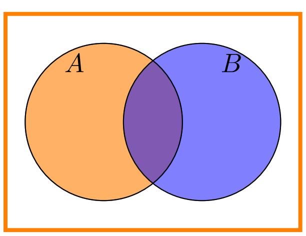


Figure for part (a).

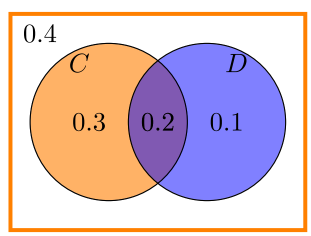
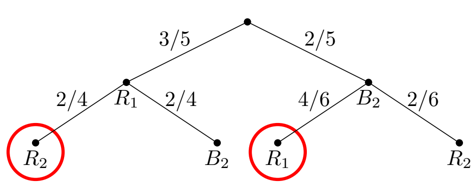


Figure for part (b).

(b) 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 2. (20)



(a) $P(R_1 \cap R_2) = \frac{3}{5} \cdot \frac{2}{4} = \boxed{\frac{6}{20} = 0.3}$.

(b) $P(B_1|R_2) = \frac{P(R_2|B_1)P(B_1)}{P(R_2)} = \frac{\frac{2}{5} \cdot \frac{4}{6}}{\frac{3}{5} \cdot \frac{2}{4} + \frac{2}{5} \cdot \frac{4}{6}} = \frac{8/30}{17/30} = \frac{8}{17}$.

Problem 3. (15) $F(1)$: 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 $\boxed{F(7) = 1}$.

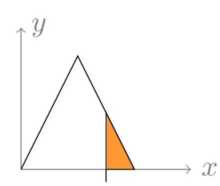
Problem 4. (20) **(a)** Let X = score of a random student.

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

(b) *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 $a \geq 1/2$ 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 5. (15) **(a)** $f(x) = F'(x) = 2 - 2x$ on $[0, 1]$. Therefore

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

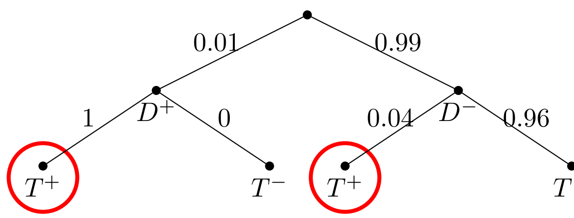
(b) $P(X \le 0.4) = F(0.4) = 0.4(2 - 0.4) = 0.4(1.6) = \boxed{0.64}$.

Problem 6. (15) Let $X \sim \mathrm{U}(a, b)$. The pdf of X is $f(x) = \dfrac{1}{b-a}$ on the interval $[a, b]$. Thus,

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

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

Problem 7. (20) **(a)** 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) $\text{Odds(winning)} = \text{Odds}(D^+|T^+) = \dfrac{P(D^+|T^+)}{P(D^-|T^+)} = \dfrac{1/4.96}{3.96/4.96} = \dfrac{1}{3.96}$.

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 8. (30) **(a)** 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) 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)

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

(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 9. (20) **(a)** 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) 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

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

<u>**answer:**</u> $\boxed{n = 6807}$

Problem 10. (10 pts)

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

$$f(\text{data}\,|\,\tau) \;=\; x_1x_2\cdots x_n\tau^n\, e^{-\frac{1}{2}\tau\sum x_i^2}.$$

The log likelihood is

$$\ln(f(\text{data}\,|\,\tau)) \;=\; \ln(x_1x_2\cdots x_n) + n\ln(\tau) - \frac{1}{2}\tau\sum x_i^2.$$

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

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

Problem 11. (15) **(a)** We assume the random error terms e_i are independent, have mean 0 and all have the same variance (homoscedastic).

(b)

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

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 $\boxed{\hat{b} = \frac{56}{52} = \frac{14}{13}}$.

Problem 12. (30)

(a) 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 $\overline{x} \pm t_{0.1}\,\frac{s}{\sqrt{n}}$. 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 $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(44)$. 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) The 80% bootstrap CI is $[\overline{x} - \delta^*_{0.1}, \overline{x} - \delta^*_{0.9}]$, where $\delta^*_{0.1}$ and $\delta^*_{0.9}$ are empirical right tail critical points for δ^*

$\delta^*_{0.1}$ = 450th element = 0.169
$\delta^*_{0.9}$ = 50th element = -0.2

So the 80% CI $= [5 - 0.169,\ 5 + 0.2] = [4.831,\ 5.2]$.

(c) 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) Method (b) is preferable if the underlying distribution is highly asymmetric.

Problem 13. (15) **(a)** Since $\mu = 1/p$ we should use the approximation $\boxed{\hat{p} = 1/\overline{x}}$.

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

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

Step 3. Compute $p^* = 1/\overline{x}^*$ 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 bootstrap confidence interval is $[\hat{p} - \delta^*_{0.025},\ \hat{p} - \delta^*_{0.975}]$.

Problem 14. (30) **(a)** 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) 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) 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 $\mu = 4$

$$
\begin{aligned}
&= 1 - P\left(\frac{2.61-4}{2/10} < Z < \frac{-.61}{2/10}\right) \\
&= 1 - P(-6.9 < Z < -3.05) \\
&= 1 - \Phi(-3.05) + \Phi(-6.9) \\
&= 1 - 0.0011 + 0 = \boxed{0.9989}.
\end{aligned}
$$

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 15. (30) **(a)** This is a normal/normal conjugate prior/likilihood update.

Hypothesis	Prior	Likelihood	Posterior
θ	N(80, 16)	$f(x\|\theta) \sim \mathrm{N}(\theta, 0.01)$	$\mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$

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) 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

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

Problem 16. (20) **(a)** Let θ represent the number of sides to the die. The data is $x_1 = 7$

Hypothesis θ	prior $p(\theta)$	likelihood $p(x_1 = 7 \,\|\, \theta)$	unnorm. post. $p(\theta)p(x_1 = 7 \,\|\, \theta)$	posterior $p(\theta \,\|\, x_1 = 7) = \frac{p(\theta)p(x_1 = 7 \,\|\, \theta)}{p(x_1 = 7)}$
$\theta = 6$	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) Odds $= \dfrac{p(\theta = 12 \mid x_1 = 7)}{p(\theta \neq 12 \mid x_1 = 7)} = \dfrac{2/5}{3/5} = \boxed{\dfrac{2}{3}}$.

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

θ	$p(\theta \mid x_1 = 7)$	$p(x_2 = 7 \mid \theta)$	$p(\theta \mid x_1 = 7)p(x_2 = 7 \mid \theta)$
$\theta = 6$	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 $p(x_2 = 7 \mid x_1 = 7) = \boxed{\dfrac{13}{120}}$.