EE 178 Probabilistic Systems Analysis
Autumn 2016 Tse Final solutions

There is a total of 7 questions with a total of 100 points. You have a total of 3 hours.

Please write all your answers in the exam booklet.

The exam is closed book but you are allowed two double-sided sheets of notes and a calculator. No other materials are allowed. A table of z-scores can be found at the end of the exam.

Good luck!

Good luck!
Problem 1: Two plus two (4 points)

What are the two key rules and two key theorems we covered in this course?

**Answer:** Bayes’ rule, total probability rule, law of large numbers, central limit theorem.
Problem 2: True or False (14 points)

2 points for each correct answer, 0 point for each incorrect answer, 1 point for each answer left blank.

(a) Let $A, B, C$ be three events. Then:

$$P(A|C) = P(A|B,C)P(B) + P(A|B^c,C)P(B^c).$$

**Answer:** False. Counter example: We choose a fair coin or a biased coin with Heads probability 0.7 (each coin chosen with probability 1/2) and flip the coin once. Let $C$ be the event that the fair coin is chosen, and $A = B$ be the event that the coin flip is Heads. $P(A|C) = 0.5$, and $P(A|B,C)P(B) + P(A|B^c,C)P(B^c) = P(B) = 0.6.$ The conditional version of total probability rule is $P(A|C) = P(A|B,C)P(B|C) + P(A|B^c,C)P(B^c|C)$ instead.

(b) $X$ and $Y$ are independent random variables. Then $X$ and $Y$ are independent conditional on any random variable $Z$.

**Answer:** False. A counter example would be $X, Y$ being two independent fair coin flips, and $Z$ is the number of Heads.

(c) The concept of cumulative distribution function only applies to continuous random variables, not discrete ones.

**Answer:** False.

(d) Let $X$ and $Y$ be two continuous random variables with probability density functions $f_X$ and $f_Y$ respectively. Let $Z = X + Y$. Then the probability density function of $Z$ is $f_Z(z) = f_X(z) + f_Y(z)$.

**Answer:** False. $f_Z$ is the convolution of $f_X$ and $f_Y$.

(e) Let $\Theta$ be a random variable uniformly distributed in $[0, 2\pi]$. Let $X = \sin \Theta$. Then $X$ is uniformly distributed in $[-1, +1]$.

**Answer:** False. $X$ is not uniform in $[-1, +1]$. The transformation from $\Theta$ to $X$ is non-linear.

(f) Let $X$ be Gaussian random variable. Then $Y = -X$ must also be a Gaussian random variable.

**Answer:** True. If $X \sim N(\mu, \sigma^2)$, then $-X \sim N(-\mu, \sigma^2)$.

(g) Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples drawn from the same distribution. Let $\mu$ be the unknown mean of that distribution. Then:

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

is always the maximum likelihood estimate of $\mu$.

**Answer:** False. A counter example is the Laplace distribution with unknown mean in the homework.
Problem 3: Marguerite shuttles (20 points)

Marguerite shuttles arrive at your favorite stop with inter-arrival times as i.i.d. exponentially distributed random variables with mean 10 minutes.

(a) [4] You want to simulate this continuous-time model on the computer but you can only simulate in discrete-time. So you decide to divide time into 1 second intervals and simulate shuttle arrivals on a second by second basis. Find a reasonable probability model for your discrete-time simulation that closely approximates the continuous-time shuttle arrival process. Specify the values of all parameters in your model.

Answer: Let $Z_i$ indicates whether a bus arrives at the $i$-th second ($Z_i = 1$ if there is a bus). Assume $Z_i \sim \text{Bern}(p)$ are independent for different $i$'s. The inter-arrival times would be i.i.d. $\text{Geom}(p)$ (in second). We want the inter-arrival times to be close to $\text{Exp}(\frac{1}{600})$, so we can set $p = \frac{1}{600}$.

(b) [4] Going back to the original continuous-time model: what is the distribution of the number of shuttles that arrive between 1 pm and 2 pm? (Hint: the discrete-time approximation you developed in part (a) may be useful as an intermediate step for thinking about this.)

Answer: $\text{Poiss}(6)$. We can see this from the discrete model that the sum of $60 \cdot 60$ (number of seconds in an hour) different $Z_i$'s follows $\text{Bin}(3600, \frac{1}{600})$, which is approximately $\text{Poiss}(6)$.

(c) [4] You arrive at the stop at 10am on a Wednesday morning. Assuming that the journey to class takes time exponentially distributed with mean 15 minutes, what is the probability that you will be late for your EE 178 class? You can assume that the journey time is independent of the arrival times of the shuttles.

Answer: Let $X \sim \text{Exp}(10)$ be the time you have to wait for the next bus (in minutes), and $Y \sim \text{Exp}(15)$ be the time for the journey (in minutes). By total probability rule,

$$P(X + Y \leq 30) = \int_{-\infty}^{\infty} P(X + Y \leq 30 | X = x) f_X(x) \, dx$$

$$= \int_0^{30} P(Y \leq 30 - x) \frac{1}{10} e^{-x/10} \, dx$$

$$= \int_0^{30} \left( 1 - e^{-(30-x)/15} \right) \frac{1}{10} e^{-x/10} \, dx$$

$$= \frac{1}{10} \int_0^{30} \left( e^{-x/10} - e^{-(30-x)/15} \right) \, dx$$

$$= \frac{1}{10} \left( \left( 10 - 10e^{-30/10} \right) - \left( 30e^{-2} - 30e^{-30/30-2} \right) \right)$$

$$= \frac{1}{10} \left( \left( 10 - 10e^{-3} \right) - \left( 30e^{-2} - 30e^{-3} \right) \right)$$

$$= 1 - 3e^{-2} + 2e^{-3}.$$

Hence $P(X + Y > 30) = 3e^{-2} - 2e^{-3} \approx 0.306$.

(d) [4] You arrive at the stop at 10am. Let $T_{\text{prev}}$ be the arrival time of the shuttle you just missed. Let $T_{\text{next}}$ be the arrival time of the next shuttle. Is the expectation of $T_{\text{next}} - T_{\text{prev}}$ smaller, equal or greater than 10 minutes? Do a calculation to justify your answer.

Answer: Since a Poisson process is memoryless, $\mathbb{E}[T_{\text{next}}] = 10 + \frac{1}{6}$ (hours from 12am). Similarly $\mathbb{E}[T_{\text{prev}}] = 10 - \frac{1}{6}$. Hence $\mathbb{E}[T_{\text{next}} - T_{\text{prev}}] = \frac{1}{5}$ (20 minutes) is greater than 10 minutes.
(e) [4] Suppose now Stanford has improved its service and the shuttles now arrive exactly every 10 minutes. Redo part (d) of the question. Can you give some intuition to explain the similarity or difference with your answer to part (d)?

**Answer:** $T_{\text{next}} - T_{\text{prev}}$ is exactly 10 minutes. This is smaller than that in (d) because if the inter-arrival times are random as in (d), it is more likely that you arrive during a longer inter-arrival time interval than a shorter one, and hence the distribution of the length of the inter-arrival time interval in which you arrive is not $\text{Exp}(1)$, but a distribution with a higher mean.
Problem 4: Back to flipping coins (14 points)

You have two coins which look identical, but one has probability of getting a Heads 0.7 and one has probability of getting a Heads 0.3. You randomly pick one coin and flip the same coin \( n \) times.

(a) [4] What is the probability that your second flip is a Heads conditional on the first flip being a Heads?

**Answer:** Let \( X_1, X_2 \ldots \) be your flips, and let \( Y \in \{1, 2\} \) be the coin you pick (\( Y = 1 \) for the coin with Heads probability 0.7). Then

\[
P(X_1 = H) = P(X_1 = H | Y = 1)P(Y = 1) + P(X_1 = H | Y = 2)P(Y = 2)
\]

\[= 0.5.
\]

\[
P(X_1 = X_2 = H) = P(X_1 = X_2 = H | Y = 1)P(Y = 1) + P(X_1 = X_2 = H | Y = 2)P(Y = 2)
\]

\[= 0.7^2 \cdot 0.5 + 0.3^2 \cdot 0.5.
\]

\[= (0.49 + 0.09)/2
\]

\[= 0.29.
\]

\[
P(X_2 = H | X_1 = H) = P(X_1 = X_2 = H) / P(X_1 = H)
\]

\[= 0.58.
\]

(b) [2] Are the first two flips independent? Explain.

**Answer:** Note that \( P(X_2 = H | X_1 = H) = 0.58 \neq 0.5 = P(X_2 = H) \). They are not independent.

(c) [4] You want to guess which coin you have chosen on the basis of the results of the \( n \) flips. Give a reasonable decision rule. What happens to the probability of making a mistake of your decision rule as \( n \to \infty \)?

**Answer:** Yes. The MAP rule is \( \hat{Y} = 1 \) if \( K \geq n/2, \hat{Y} = 2 \) if \( K < n/2\), where \( K \) is the number of Heads in \( X_1, \ldots, X_n \). The probability of incorrect guess is

\[
P(Y \neq \hat{Y}) = \frac{1}{2} (P(Y \neq \hat{Y} | Y = 1) + P(Y \neq \hat{Y} | Y = 2))
\]

\[= \frac{1}{2} (P(K < n/2 | Y = 1) + P(K \geq n/2 | Y = 2))
\]

\[= \frac{1}{2} (P(K/n < 1/2 | Y = 1) + P(K/n \geq 1/2 | Y = 2))
\]

Note that conditioned on \( Y = 1 \), \( K \sim \text{Bin}(n, 0.7) \), hence by law of large numbers, \( K/n \) tends to 0.7. Hence \( P(K/n < 1/2 | Y = 1) \) tends to 0. Similarly \( P(K/n \geq 1/2 | Y = 2) \) tends to 0. Therefore \( P(Y \neq \hat{Y}) \) tends to 0 as \( n \to \infty \).

To deduce the MAP rule formally (not needed for full score):

\[
P(Y = 1 | X_1 = x_1, \ldots, X_n = x_n) \propto P(X_1 = x_1, \ldots, X_n = x_n | Y = 1)P(Y = 1)
\]

\[\propto P(X_1 = x_1, \ldots, X_n = x_n | Y = 1)
\]

\[= 0.7^k \cdot 0.3^{n-k}
\]

where \( k \) is the number of Heads in \( x_1, \ldots, x_n \). Similarly

\[
P(Y = 2 | X_1 = x_1, \ldots, X_n = x_n) \propto 0.3^k \cdot 0.7^{n-k}.
\]
We have
\[
\begin{align*}
\Pr(Y = 1 | X_1 = x_1, \ldots, X_n = x_n) &= \frac{0.7^k \cdot 0.3^{n-k}}{0.3^k \cdot 0.7^{n-k}} \\
&= (7/3)^{2k-n}
\end{align*}
\]
Hence the MAP is \( \hat{Y} = 1 \) if \( k \geq n/2 \), \( \hat{Y} = 2 \) if \( k < n/2 \).

(d) [4] Does the sequence of flips form a Markov chain? Explain. (Hint: the earlier parts of this question may be useful.)

**Answer:** No. Note that by the same argument as (a), \( \Pr(X_n = H | X_{n-1} = H) = 0.58 \). On the other hand, if we know \( X_1 = \cdots = X_{n-1} = H \), then the (posterior) probability that the coin is the one with Heads probability 0.7 is increasing in \( n \), so \( \Pr(X_n = H | X_1 = \cdots = X_{n-1} = H) \) is increasing and tends to 0.7 as \( n \to \infty \). Hence \( \Pr(X_n = H | X_1 = \cdots = X_{n-1} = H) \neq \Pr(X_n = H | X_{n-1} = H) \) when \( n \geq 3 \).
Problem 5: Modeling English (9 points)

We want to model an English sentence as generated by a stationary Markov chain. Each symbol in a sentence is either one of the 26 English letters or a space.

(a) [3] Suppose a sentence is 100 symbols long. Without imposing a Markov chain model, how many parameters are there in the probability model for the sentence?

Answer: There are $27^{100}$ different sentences. Hence there are $27^{100}$ parameters for the probability distribution over sentences (or $27^{100} - 1$ since they sum to 1).

(b) [3] With the Markov model, how many states and how many parameters do you need to specify in the model?

Answer: Each symbol corresponds to a state so there are 27 states. The parameters are the initial distribution (26 parameters) and the transition probabilities ($27 \cdot 26$ parameters). There are $26 + 27 \cdot 26 = 728$ parameters in total.

(c) [3] What data would you use to estimate these parameters, and how would you estimate these parameters?

Answer: We can collect data using an English corpus and breaking it down into sentences. The initial distribution can be estimated from the first symbol of the sentences (e.g. the estimate for $\pi(A)$ is the fraction of sentences starting with symbol $A$). The transition probability $P(B|A)$ can be estimated from extracting all pairs of consecutive symbols among the sentences where the first symbol is $A$, and then compute the fraction of pair $(A, B)$ among those pairs.
Problem 6: Spam detection (18 points)

Consider the spam detection problem we discussed in class. We used one feature, the fraction of letters that are capitalized letters, to predict whether an email is a spam. The model is that if the email is not a spam, this fraction is assumed to be distributed as $N(\mu_0, \sigma^2)$, while if the email is a spam, it is distributed as $N(\mu_1, \sigma^2)$.

To design a spam detector, we use the MAP (maximum a posterior probability) rule, which is to decide an email is a spam if the probability that the email is a spam conditional on the feature value is greater than the probability that the email is normal conditional on the feature value.

Let us assume that it is a priori equally likely for an email to be a spam or not.

(a) [3] Suppose we estimated the parameters exactly, and $\mu_0 = 0.05, \mu_1 = 0.2, \sigma = 0.02$. Give the explicit MAP rule for the spam detector.

**Answer:** Let $X$ be the fraction of letters that are capitalized and $Y$ indicates whether it is a spam for a random email. For an email with fraction of capitalized letters $x$. The posterior probability of $Y = 0$ is proportional to

$$P(Y = 0 | X = x) \propto f_{X|Y}(x|0) P(Y = 0) \quad \text{(Bayes’ rule, and the denominator is common)}$$

$$\propto f_{X|Y}(x|0) \quad \text{(since } P(Y = 0) = P(Y = 1))$$

$$= \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left(-\frac{(x - \mu_0)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left(-\frac{(x - 0.05)^2}{2\sigma^2}\right)$$

Similar for $P(Y = 1 | X = x)$. Hence we guess it is spam (i.e. $\hat{y} = 1$) if $(x - 0.2)^2$ is smaller than $(x - 0.05)^2$, i.e., $x > 0.125$. We guess not spam if $x \leq 0.125$.

To compete with your competitor, you decide that looking at one feature is not enough! Your spam detector needs to look at another feature. You thought maybe checking whether the email contains the word "rich" may be a good idea, since lots of spam try to lure people by promising them to get rich fast.

(b) [2] For simplicity, let’s begin by using this new feature only (and not the old feature) in the spam detector. Two key parameters that need to be estimated are the probability that a spam email contains "rich" and the probability that a normal email contains "rich". How would you estimate these parameters from data?

**Answer:** Let $Z_i \in \{0, 1\}$ indicates whether email $i$ contains “rich” ($Z_i = 1$ if contains “rich”). Then the model is $Z_i \sim \text{Bern}(p_0)$ conditioned on $Y_i = 0$, and $Z_i \sim \text{Bern}(p_1)$ conditioned on $Y_i = 1$. The MLE estimates are

$$\hat{p}_0 = \frac{1}{N_0} \sum_{i: Y_i = 0} Z_i$$

$$\hat{p}_1 = \frac{1}{N_1} \sum_{i: Y_i = 1} Z_i$$

where $N_0$ is the number of emails that are not spam, $N_1$ is the number of emails that are spam.
(c) [4] Suppose these parameters are estimated accurately, and they are 0.3 and 0.1 respectively. What is the MAP rule for deciding whether an email is a spam or not by just looking at this new feature?

Answer:

\[
P(Y = 0 | Z = 0) = \frac{P(Z = 0 | Y = 0) P(Y = 0)}{P(Z = 0 | Y = 0) P(Y = 0) + P(Z = 0 | Y = 1) P(Y = 1)}
\]

\[
= \frac{1 - p_0}{(1 - p_0) + (1 - p_1)}
\]

\[
= \frac{1 - 0.1}{(1 - 0.1) + (1 - 0.3)}
\]

\[
= \frac{9}{16}
\]

\[
> \frac{1}{2}
\]

and hence we guess \( \hat{Y} = 0 \) if \( Z = 0 \). Similarly \( P(Y = 0 | Z = 1) = 0.1/(0.1 + 0.3) = 1/4 < 1/2 \), so we guess \( \hat{Y} = 1 \) if \( Z = 1 \). Hence \( \hat{Y} = Z \).

(d) [3] Now we want to build a spam detector combining this new feature and the old. Given the information in the question so far, do we have a complete probability model for this task? If not, please make further reasonable assumptions to have a complete model.

Answer: We have the distribution of \( Y \), the conditional distribution of \( X \) given \( Y \), and the conditional distribution of \( Z \) given \( Y \). This is not a complete model since we do not know the conditional dependency between \( X \) and \( Z \) given \( Y \). We can further assume \( Z \) and \( X \) are conditionally independent given \( Y \), then we have a complete model.

(e) [6] Using your model, build a spam detector combining the two features, by deriving the explicit MAP rule.

Answer:

\[
P(Y = y | X = x, Z = z) \propto P(Z = z | Y = y) f_{X|Y,Z}(x|y,z) P(Y = y)
\]

\[
\propto P(Z = z | Y = y) f_{X|Y,Z}(x|y,z)
\]

\[
= P(Z = z | Y = y) f_{X|Y}(x|y) \quad \text{(since } X, Z \text{ cond. indep. given } Y) \]

\[
\propto p_{Z|Y}(z|y) \cdot \exp \left( -\frac{(x - \mu_y)^2}{2\sigma^2} \right).
\]

After taking logarithm, we have

\[
\ln p_{Z|Y}(z|y) - \frac{(x - \mu_y)^2}{2\sigma^2}
\]

Hence

\[
\hat{Y} = \arg\min_y \left( \ln p_{Z|Y}(z|y) - \frac{(X - \mu_y)^2}{2\sigma^2} \right).
\]

We then find out the explicit decision rule (you can get most points without this calculations). For \( Z = 0 \),

\[
\ln p_{Z|Y}(Z|0) - \frac{(X - \mu_0)^2}{2\sigma^2} = \ln 0.9 - \frac{(X - 0.05)^2}{2 \cdot 0.02^2}
\]
\[
\ln p_{Z|Y}(Z|1) - \frac{(X - \mu_0)^2}{2\sigma^2} = \ln 0.7 - \frac{(X - 0.2)^2}{2 \cdot 0.02^2}
\]

Setting them to be equal to solve the threshold \( t \),

\[
2 \cdot 0.02^2 (\ln 0.9 - \ln 0.7) = (t - 0.05)^2 - (t - 0.2)^2,
\]

\[
t = 0.125 + \frac{0.02^2}{0.15} \ln \frac{0.9}{0.7},
\]

\[
\approx 0.12567
\]

For \( Z = 1 \),

\[
\ln p_{Z|Y}(Z|0) - \frac{(X - \mu_0)^2}{2\sigma^2} = \ln 0.1 - \frac{(X - 0.05)^2}{2 \cdot 0.02^2}
\]

\[
\ln p_{Z|Y}(Z|1) - \frac{(X - \mu_0)^2}{2\sigma^2} = \ln 0.3 - \frac{(X - 0.2)^2}{2 \cdot 0.02^2}
\]

Setting them to be equal to solve the threshold \( t \),

\[
2 \cdot 0.02^2 (\ln 0.1 - \ln 0.3) = (t - 0.05)^2 - (t - 0.2)^2,
\]

\[
t = 0.125 + \frac{0.02^2}{0.15} \ln \frac{0.1}{0.3},
\]

\[
\approx 0.12207
\]

Hence the MAP rule is \( \hat{Y} = 0 \) if \( Z = 0 \) and \( X \leq 0.12567 \) or \( Z = 1 \) and \( X \leq 0.12207 \), \( \hat{Y} = 1 \) otherwise.
Problem 7: RNA-seq again (21 points)

Consider the RNA-seq quantification problem in Lab 3. We like to consider an extension to the more realistic case when the transcripts are of different lengths. Specifically, we will consider a situation when there are two transcripts. Transcript 1 is of length 2099 symbols and transcript 2 is of length 1099 symbols and contain exactly the first 1099 symbols of transcript 1. (This situation is rather common in actual RNA-seq problems because of the underlying biology of the transcription process.) Transcript 1 has abundance $\rho$ and transcript 2 has abundance $1 - \rho$. We sample reads of length 100 symbols in order to estimate $\rho$. As in Lab 3, the sampling process for each read is that we first sample a transcript according to the abundance distribution and then sample the read uniformly among all possible starting positions of that transcript.

We will make the assumption that a read can only come from at most one position in transcript 1.

a) [4] You sample one read $R$. What is the probability of $R = r$?

**Answer:** Let the transcript chosen be $T$. If the read can come from transcript 2, then

$$P(R = r) = \rho P(R = r \mid T = 1) + (1 - \rho) P(R = r \mid T = 2)$$

$$= \frac{\rho}{2099 - 100 + 1} + \frac{1 - \rho}{1099 - 100 + 1}$$

$$= \frac{\rho}{2000} + \frac{1 - \rho}{1000}$$

$$= \frac{1}{1000} - \frac{\rho}{2000}$$

If the read can only come from transcript 1, then

$$P(R = r) = \rho P(R = r \mid T = 1) + (1 - \rho) P(R = r \mid T = 2)$$

$$= \frac{\rho}{2099 - 100 + 1} + 0$$

$$= \frac{\rho}{2000}.$$

(b) [4] You sample $n$ reads $R_1, R_2, \ldots, R_n$ and observe $R_1 = r_1, R_2 = r_2, \ldots, R_n = r_n$. Find the maximum likelihood estimate of $\rho$ from the read data.

**Answer:** Let $K$ be the number of reads in $R_1, \ldots, R_n$ that can only come from transcript 1, and suppose we observe $K = k$. A simple method is to observe that

$$K \sim \text{Bin}(n, a),$$

where $a = \frac{1000\rho}{2000} = \rho/2$. If $a$ can be any number in $[0, 1]$, then its MLE is $K/n$, but since $a \leq 1/2$, and the likelihood is monotonically increasing in the range $[0, 1/2]$ when $K/n \geq 1/2$, we have

$$\hat{a} = \min \left\{ \frac{K}{n}, \frac{1}{2} \right\},$$

$$\hat{\rho} = \min \left\{ \frac{2K}{n}, 1 \right\}.$$

Or we can do this by direct computation. The log-likelihood is

$$\sum_{i=1}^{n} \ln P(R_i = r_i; \rho)$$

$$= (n - k) \ln \left( \frac{1}{1000} - \frac{\rho}{2000} \right) + k \ln \left( \frac{\rho}{2000} \right).$$
Differentiating and setting to 0,

\[- \left( \frac{1}{2000} \right) \frac{n-k}{\rho - 1000} + \left( \frac{1}{2000} \right) \frac{k}{\rho - 2000} = 0,\]

\[\frac{n-k}{\rho - 2} + \frac{k}{\rho} = 0,\]

\[(n-k)\rho = k(2 - \rho),\]

\[\rho = \frac{2k}{n},\]

Hence

\[\hat{\rho} = \min \left\{ \frac{2K}{n}, 1 \right\} = \begin{cases} \frac{2K}{n} & \text{if } \frac{2K}{n} \leq 1 \\ 1 & \text{if } \frac{2K}{n} > 1 \end{cases}\]

(c) In a typical RNA-seq experiment, there are hundreds of millions of reads and so it takes huge amount of memory to store the reads. Using your answer to the previous part, do you need to store all the reads themselves to compute the maximum likelihood estimate of \(\rho\)? If so, explain. If not, what is the minimal amount of information from the read data that needs to be stored?

Answer: We only need to store \(n\) and \(k\), where \(k\) is the number of reads that can only come from transcript 1.

(d) How large \(n\) needs to be such that you can guarantee that the true parameter is within \(\pm 0.03\) of the estimate with probability at least 0.95?

Answer: Let \(K\) be the number of reads in \(R_1, \ldots, R_n\) that can only come from transcript 1, then

\[K \sim \text{Bin}(n, \rho/2).\]

\[\mathbb{E}[K] = \frac{n\rho}{2}\]

\[\text{Var}(K) = \frac{n\rho}{2} \left( 1 - \frac{\rho}{2} \right) \leq n/4.\]

\[\mathbf{P}(|\hat{\rho} - \rho| > 0.03) = \mathbf{P} \left( \min \left\{ \frac{2K}{n}, 1 \right\} - \rho > 0.03 \right) \leq \mathbf{P} \left( \left| \frac{2K}{n} - \rho \right| > 0.03 \right) \text{ (since } \rho \leq 1) \leq \frac{\text{Var} \left( \frac{2K}{n} \right)}{0.03^2} \text{ (by Chebyshev since } \mathbb{E} \left[ \frac{2K}{n} \right] = \rho) \leq \frac{1}{n \cdot 0.03^2}\]

To guarantee \(\mathbf{P}(|\hat{\rho} - \rho| > 0.03) \leq 0.05\), we need

\[n \geq 22223.\]
(e) [4] Can you come up with a less conservative estimate of $n$ in the earlier part by using an appropriate approximation?

**Answer:** By central limit theorem,

$$\frac{\frac{2K}{n} - \rho}{\frac{\sqrt{\text{Var}(K)}}{2\sqrt{n}}}$$

tends to $N(0,1)$. Let $Z \sim N(0,1)$,

$$\mathbb{P}(|\hat{\rho} - \rho| > 0.03) \leq \mathbb{P}\left(\left|\frac{\frac{2K}{n} - \rho}{\frac{\sqrt{\text{Var}(K)}}{2\sqrt{n}}}\right| > 0.03\right)$$

$$= \mathbb{P}\left(\left|\frac{2K}{n} - \rho\right| > \frac{0.03}{\frac{\sqrt{\text{Var}(K)}}{2\sqrt{n}}}\right)$$

$$\approx \mathbb{P}\left(\left|Z\right| > \frac{0.03}{\frac{\sqrt{\text{Var}(K)}}{2\sqrt{n}}}\right)$$

$$\leq \mathbb{P}\left(\left|Z\right| > 0.03\sqrt{n}\right)$$

$$= 2Q(0.03\sqrt{n})$$

Setting it to 0.05,

$$0.03\sqrt{n} \approx 1.96$$

$$n \geq 4269.$$

(f) [3] What happens to the 95% confidence interval as $n \to \infty$?

**Answer:** The variance of $\hat{\rho} = \min\left\{\frac{2K}{n}, 1\right\}$ tends to 0, so the length of the confidence interval tends to 0.
Z-Score Table

Here is a table of probabilities associated with the standard Gaussian distribution. It tabulates $P(Z \leq z)$ for various values of $z$, where $Z$ has a standard Gaussian distribution. To look up $P(Z \leq z)$, look up the row of the first two digits of $z$ and then the column of the third. For example, to look up $P(Z \leq 1.34)$, first find the row labeled 1.3 and then the column 0.04. The result is the entry in that cell, i.e., 0.9099.

<table>
<thead>
<tr>
<th>$z$</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.5000</td>
<td>0.5040</td>
<td>0.5080</td>
<td>0.5120</td>
<td>0.5160</td>
<td>0.5199</td>
<td>0.5239</td>
<td>0.5279</td>
<td>0.5319</td>
<td>0.5359</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5398</td>
<td>0.5438</td>
<td>0.5478</td>
<td>0.5517</td>
<td>0.5557</td>
<td>0.5596</td>
<td>0.5636</td>
<td>0.5675</td>
<td>0.5714</td>
<td>0.5753</td>
</tr>
<tr>
<td>0.2</td>
<td>0.5793</td>
<td>0.5832</td>
<td>0.5871</td>
<td>0.5910</td>
<td>0.5948</td>
<td>0.5987</td>
<td>0.6026</td>
<td>0.6064</td>
<td>0.6103</td>
<td>0.6141</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6179</td>
<td>0.6217</td>
<td>0.6255</td>
<td>0.6293</td>
<td>0.6331</td>
<td>0.6368</td>
<td>0.6406</td>
<td>0.6443</td>
<td>0.6480</td>
<td>0.6517</td>
</tr>
<tr>
<td>0.4</td>
<td>0.6554</td>
<td>0.6591</td>
<td>0.6628</td>
<td>0.6664</td>
<td>0.6700</td>
<td>0.6736</td>
<td>0.6772</td>
<td>0.6808</td>
<td>0.6844</td>
<td>0.6879</td>
</tr>
<tr>
<td>0.5</td>
<td>0.6915</td>
<td>0.6950</td>
<td>0.6985</td>
<td>0.7019</td>
<td>0.7054</td>
<td>0.7088</td>
<td>0.7123</td>
<td>0.7157</td>
<td>0.7190</td>
<td>0.7224</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7257</td>
<td>0.7291</td>
<td>0.7324</td>
<td>0.7357</td>
<td>0.7389</td>
<td>0.7422</td>
<td>0.7454</td>
<td>0.7486</td>
<td>0.7517</td>
<td>0.7549</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7580</td>
<td>0.7611</td>
<td>0.7642</td>
<td>0.7673</td>
<td>0.7704</td>
<td>0.7734</td>
<td>0.7764</td>
<td>0.7794</td>
<td>0.7823</td>
<td>0.7852</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7881</td>
<td>0.7910</td>
<td>0.7939</td>
<td>0.7967</td>
<td>0.7995</td>
<td>0.8023</td>
<td>0.8051</td>
<td>0.8078</td>
<td>0.8106</td>
<td>0.8133</td>
</tr>
<tr>
<td>0.9</td>
<td>0.8159</td>
<td>0.8186</td>
<td>0.8212</td>
<td>0.8238</td>
<td>0.8264</td>
<td>0.8289</td>
<td>0.8315</td>
<td>0.8340</td>
<td>0.8365</td>
<td>0.8389</td>
</tr>
<tr>
<td>1.0</td>
<td>0.8413</td>
<td>0.8438</td>
<td>0.8461</td>
<td>0.8485</td>
<td>0.8508</td>
<td>0.8531</td>
<td>0.8554</td>
<td>0.8577</td>
<td>0.8599</td>
<td>0.8621</td>
</tr>
<tr>
<td>1.1</td>
<td>0.8643</td>
<td>0.8665</td>
<td>0.8686</td>
<td>0.8708</td>
<td>0.8729</td>
<td>0.8749</td>
<td>0.8770</td>
<td>0.8790</td>
<td>0.8810</td>
<td>0.8830</td>
</tr>
<tr>
<td>1.2</td>
<td>0.8849</td>
<td>0.8869</td>
<td>0.8888</td>
<td>0.8907</td>
<td>0.8925</td>
<td>0.8944</td>
<td>0.8962</td>
<td>0.8980</td>
<td>0.8997</td>
<td>0.9015</td>
</tr>
<tr>
<td>1.3</td>
<td>0.9032</td>
<td>0.9049</td>
<td>0.9066</td>
<td>0.9082</td>
<td>0.9099</td>
<td>0.9115</td>
<td>0.9131</td>
<td>0.9147</td>
<td>0.9162</td>
<td>0.9177</td>
</tr>
<tr>
<td>1.4</td>
<td>0.9192</td>
<td>0.9207</td>
<td>0.9222</td>
<td>0.9236</td>
<td>0.9251</td>
<td>0.9265</td>
<td>0.9279</td>
<td>0.9292</td>
<td>0.9306</td>
<td>0.9319</td>
</tr>
<tr>
<td>1.5</td>
<td>0.9332</td>
<td>0.9345</td>
<td>0.9357</td>
<td>0.9370</td>
<td>0.9382</td>
<td>0.9394</td>
<td>0.9406</td>
<td>0.9418</td>
<td>0.9429</td>
<td>0.9441</td>
</tr>
<tr>
<td>1.6</td>
<td>0.9452</td>
<td>0.9463</td>
<td>0.9474</td>
<td>0.9484</td>
<td>0.9495</td>
<td>0.9505</td>
<td>0.9515</td>
<td>0.9525</td>
<td>0.9535</td>
<td>0.9545</td>
</tr>
<tr>
<td>1.7</td>
<td>0.9554</td>
<td>0.9564</td>
<td>0.9573</td>
<td>0.9582</td>
<td>0.9591</td>
<td>0.9599</td>
<td>0.9608</td>
<td>0.9616</td>
<td>0.9625</td>
<td>0.9633</td>
</tr>
<tr>
<td>1.8</td>
<td>0.9641</td>
<td>0.9649</td>
<td>0.9656</td>
<td>0.9664</td>
<td>0.9671</td>
<td>0.9678</td>
<td>0.9686</td>
<td>0.9693</td>
<td>0.9699</td>
<td>0.9706</td>
</tr>
<tr>
<td>1.9</td>
<td>0.9713</td>
<td>0.9719</td>
<td>0.9726</td>
<td>0.9732</td>
<td>0.9738</td>
<td>0.9744</td>
<td>0.9750</td>
<td>0.9756</td>
<td>0.9761</td>
<td>0.9767</td>
</tr>
<tr>
<td>2.0</td>
<td>0.9772</td>
<td>0.9778</td>
<td>0.9783</td>
<td>0.9788</td>
<td>0.9793</td>
<td>0.9798</td>
<td>0.9803</td>
<td>0.9808</td>
<td>0.9812</td>
<td>0.9817</td>
</tr>
<tr>
<td>2.1</td>
<td>0.9821</td>
<td>0.9826</td>
<td>0.9830</td>
<td>0.9834</td>
<td>0.9838</td>
<td>0.9842</td>
<td>0.9846</td>
<td>0.9850</td>
<td>0.9854</td>
<td>0.9857</td>
</tr>
<tr>
<td>2.2</td>
<td>0.9861</td>
<td>0.9864</td>
<td>0.9868</td>
<td>0.9871</td>
<td>0.9875</td>
<td>0.9878</td>
<td>0.9881</td>
<td>0.9884</td>
<td>0.9887</td>
<td>0.9890</td>
</tr>
<tr>
<td>2.3</td>
<td>0.9893</td>
<td>0.9896</td>
<td>0.9898</td>
<td>0.9901</td>
<td>0.9904</td>
<td>0.9906</td>
<td>0.9909</td>
<td>0.9911</td>
<td>0.9913</td>
<td>0.9916</td>
</tr>
<tr>
<td>2.4</td>
<td>0.9918</td>
<td>0.9920</td>
<td>0.9922</td>
<td>0.9925</td>
<td>0.9927</td>
<td>0.9929</td>
<td>0.9931</td>
<td>0.9932</td>
<td>0.9934</td>
<td>0.9936</td>
</tr>
<tr>
<td>2.5</td>
<td>0.9938</td>
<td>0.9940</td>
<td>0.9941</td>
<td>0.9943</td>
<td>0.9945</td>
<td>0.9946</td>
<td>0.9948</td>
<td>0.9949</td>
<td>0.9951</td>
<td>0.9952</td>
</tr>
<tr>
<td>2.6</td>
<td>0.9953</td>
<td>0.9955</td>
<td>0.9956</td>
<td>0.9957</td>
<td>0.9959</td>
<td>0.9960</td>
<td>0.9961</td>
<td>0.9962</td>
<td>0.9963</td>
<td>0.9964</td>
</tr>
<tr>
<td>2.7</td>
<td>0.9965</td>
<td>0.9966</td>
<td>0.9967</td>
<td>0.9968</td>
<td>0.9969</td>
<td>0.9970</td>
<td>0.9971</td>
<td>0.9972</td>
<td>0.9973</td>
<td>0.9974</td>
</tr>
<tr>
<td>2.8</td>
<td>0.9974</td>
<td>0.9975</td>
<td>0.9976</td>
<td>0.9977</td>
<td>0.9977</td>
<td>0.9978</td>
<td>0.9979</td>
<td>0.9979</td>
<td>0.9980</td>
<td>0.9981</td>
</tr>
<tr>
<td>2.9</td>
<td>0.9981</td>
<td>0.9982</td>
<td>0.9982</td>
<td>0.9983</td>
<td>0.9984</td>
<td>0.9984</td>
<td>0.9985</td>
<td>0.9985</td>
<td>0.9986</td>
<td>0.9986</td>
</tr>
<tr>
<td>3.0</td>
<td>0.9987</td>
<td>0.9987</td>
<td>0.9988</td>
<td>0.9988</td>
<td>0.9988</td>
<td>0.9989</td>
<td>0.9989</td>
<td>0.9989</td>
<td>0.9990</td>
<td>0.9990</td>
</tr>
</tbody>
</table>