Open-book, open-notes; two problems (true/false and calculation); each of the 12 true/false questions is worth 10 points, and the calculation problems are worth 160 points, for a total of 280 points. The right answer with no reasoning to support it, or incorrect reasoning, will get half credit, so try to make a serious effort on each part of each problem (this will ensure you at least half credit).

This homework assignment is to be entirely your own efforts; do not collaborate with anyone or get help from anyone but me.

1 True/False

[120 total points] For each statement below, say whether it’s true or false; if true without further assumptions, briefly explain why it’s true (and — extra credit — what its implications are for statistical inference); if it’s sometimes true, give the extra conditions necessary to make it true; if it’s false, briefly explain how to change it so that it’s true.

(A) You’re about to spin a roulette wheel, which will result in a metal ball landing in one of 38 slots numbered \( \Omega = \{0, 00, 1, 2, \ldots, 36\} \); 18 of the numbers from 1 to 36 are colored red, 18 are black, and 0 and 00 are green. You regard this wheel-spinning as fair, by which You mean that all 38 elemental outcomes in \( \Omega \) are equally probable. Under Your assumption of fairness, the classical (Pascal-Fermat) probability of getting a red number on the next spin exists, is unique, and equals \( \frac{18}{38} \).

(B) Under the same conditions as (A), the Kolmogorov (frequentist) probability of getting a red number on the next spin exists, is unique, and equals \( \frac{18}{38} \).

(C) You’re a professor; a new student (whom You’ve never met before) comes to Your office on the day before the quarter begins, saying that she (the student) wants to take a class that You’re about to teach that quarter, but she’s worried she may fail. The Kolmogorov (frequentist) probability that she will fail the class, if she takes it, is undefined, because there’s no unique \( \Omega \) on which to base the Kolmogorov probability calculation.

(D) In the Bernoulli sampling model, in which \( (Y_1, \ldots, Y_n|\theta \mathcal{B}) \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta) \), the sum \( s_n = \sum_{i=1}^{n} y_i \) of the observed data values \( y = (y_1, \ldots, y_n) \) is sufficient for inference about \( \theta \), and this means that in this model You can throw away the data vector \( y \) and focus only on \( s_n \) without any loss of information whatsoever.
The Beta(\(\theta | \alpha, \beta\)) parametric family of distributions is useful as a source of prior distributions when the sampling model is as in (D), because all distributional shapes (symmetric, skewed, multimodal, ...) on (0, 1) are realizable in this family.

Specifying the ingredients \(\{p(\theta | B), p(D | \theta B), A, U(a, \theta)\}\) in Your model for Your uncertainty about an unknown \(\theta\) (in light of background information \(B\) and data \(D\)) is typically easy, because in any given problem there will typically be one and only one way to specify each of these ingredients; an example is the Bernoulli sampling distribution \(p(D | \theta B)\) arising uniquely, under exchangeability, from de Finetti’s Theorem for binary outcomes.

In trying to construct a good uncertainty assessment of the form \(P(A | B)\), where \(A\) is a proposition and \(B\) is a proposition of the form \((B_1 \text{ and } B_2 \text{ and } \ldots \text{ and } B_k)\), You should try hard not to condition on any propositions \(B_i\) that are false, because that would be the probabilistic equivalent of dividing by zero.

Unknown quantities \((Y_1, \ldots, Y_n)\) are IID if and only if they’re exchangeable.

The kind of objectivity in probability assessment sought by people like Venn, in which all reasonable people would agree on the assessed value, is often impossible to achieve, because all such assessments are conditional on the (1) assumptions, (2) judgments and (3) information base of the person making the probability assessment, and different reasonable people can differ along any of those three dimensions.

When making a decision in the face of uncertainty about an unknown \(\theta\), after specifying Your action space \(A\) and utility function \(U(a, \theta)\) and agreeing on the convention that large utility values are to be preferred over small ones, the optimal decision is found by maximizing \(U(a, \theta)\) over all \(a \in A\).

One reason that Bayesian inference was not widely used in the early part of the 20th century was that approximating the (potentially high-dimensional) integrals arising from this approach was difficult in an era when computing was slow and the Laplace-approximation technique had been forgotten.

Jaynes (2003) makes a useful distinction between \{reality\} (epistemology) and \{Your current information about reality\} (ontology); this distinction is useful in probabilistic modeling because \{the world\} does not necessarily change every time \{Your state of knowledge about the world\} changes.

2 Calculation

[20 total points] In Part 2 of the lecture notes (page 3), I mentioned that (strictly speaking) the classical probability \(P_C\) of choosing an odd integer at random from \(\{-2, -1, 0, +1, +2, \ldots\}\) was undefined, because if You tried to apply the definition of classical probability the numerator and denominator would both be infinite. Let’s see if we can get a meaningful answer to this question by a method I highly recommend: sneaking up on infinity (by which I mean: identify an explicit process by which something \((k, \text{ say})\) gets big, and see if the thing of interest to You \((f(k), \text{ say})\) approaches a stable limit as \(k\) increases).

For integer \(k \geq 1\), let \(\Omega_k\) be the set of integers \(\{-k, \ldots, -1, 0, +1, \ldots, k\}\).
(i) Find an explicit expression (as a function of $k$) for

$$P_C(\text{odd integer chosen at random from } \Omega_k) \quad (1)$$

for odd $k$, and repeat for even $k$. Do these expressions have well-defined limits as $k \to \infty$, and (if so) do they converge to the same value? Using this as an approach to arrive at an extended definition of classical probability, what is

$$P_C(\text{choosing an odd integer at random from } \{\ldots, -2, -1, 0, +1, +2, \ldots\})? \quad (2)$$

Explain briefly. [10 points]

(ii) Using the same sneaking-up-on-infinity reasoning, what is the frequentist probability

$$P_F(\text{choosing an odd integer at random from } \{\ldots, -2, -1, 0, +1, +2, \ldots\})? \quad (3)$$

Explain briefly. [5 points]

(iii) Repeat (ii) for the Bayesian probability of choosing an odd integer at random from $\{\ldots, -2, -1, 0, +1, +2, \ldots\}$, using the Laplace/Cox/Jaynes information-theoretic definition of Bayesian probability. [5 points]

(B) [30 total points] Consider the HIV screening example we looked at in class, in which $A = \{\text{the patient in question is HIV positive}\}$ and $D = \{\text{ELISA says he’s HIV positive}\}$. Let $p$ stand for the prevalence of HIV among people similar to this patient (recall that in our example $p = 0.01$), and let $\epsilon$ and $\pi$ stand for the sensitivity and specificity of the ELISA screening test, respectively (in our case study $\epsilon = 0.95$ and $\pi = 0.98$).

(i) By using Bayes’s Theorem (in probability or odds form), write down explicit formulas in terms of $p, \epsilon, \pi$ for the positive predictive value (PPV), $P(A|D)$, and negative predictive value (NPV), $P(\text{not } A|\text{not } D)$, of screening tests like ELISA (ELISA’s PPV and NPV with patients like the one in our case study were 0.32 and 0.99948, respectively). These formulas permit analytic study of the tradeoff between PPV and NPV. [10 points]

(ii) [10 points] Interest focused in class on why ELISA’s PPV is so bad for people, like the guy we considered in the case study, for whom HIV is relatively rare ($p = 0.01$).

(a) Holding $\epsilon$ and $\pi$ constant at ELISA’s values of 0.95 and 0.98, respectively, obtain expressions (from those in B(i)) for the PPV and NPV as a function of $p$, and plot these functions as $p$ goes from 0 to 0.1.

(b) Show (e.g., by means of Taylor series) that in this range the NPV is closely approximated by the simple linear function $(1 - 0.056p)$.

(c) How large would $p$ have to be for ELISA’s PPV to exceed 0.5? 0.75?

(d) What would ELISA’s NPV be for those values of $p$?

(e) Looking at both PPV and NPV, would you regard ELISA as a good screening test for subpopulations with (say) $p = 0.1$? Explain briefly.

(iii) [10 points] Suppose now that $p$ is held constant at 0.01 and we’re trying to improve ELISA for use on people with that prevalence of HIV, where “improve” for the sake of this part of the problem means raising the PPV while not suffering too much of a decrease (if any) of the NPV. ELISA is based on the level $L$ of a particular antibody in the blood, and uses a rule of the form {if $L \geq c$ announce that the person is HIV positive}. This means that if you change $c$ the sensitivity and specificity change in a tug-of-war fashion: altering $c$ to make $\epsilon$ go up makes $\pi$ go down, and vice versa.
(a) By using the formulas in B(i), show that as $\epsilon$ approaches 1 with $\pi$ no larger than 0.98, the NPV will approach 1 but the biggest you can make the PPV is about 0.336. Thus if we want to raise the PPV we would be better off trying to increase $\pi$ than $\epsilon$. Suppose there were a way to change $c$ that would cause $\pi$ to go up while holding $\epsilon$ arbitrarily close to 0.95.

(b) Show that $\pi$ would have to climb to about 0.997 to get the PPV up to 0.75.

(c) Is the NPV still at acceptable levels under these conditions? Explain briefly.

(C) [110 total points] As a small part of a study I worked on at the RAND Corporation in the late 1980s, we obtained data on a random sample of $n = 14$ women who came to a hospital in Santa Monica, CA, in 1988 to give birth to premature babies. One outcome of interest was the length of stay (LoS) $y_i$ in the hospital that woman $i$ in this sample experienced, recorded as an integer; it was possible for this variable to be recorded as 0 if the LoS was under 12 hours. The data values were as follows: $y = (y_1, \ldots, y_n) = (1, 2, 1, 1, 4, 1, 2, 2, 0, 3, 6, 2, 1, 3)$. The unknown $\theta$ of principal interest in this problem is the mean LoS for all women giving birth to premature babies at this Santa Monica hospital in 1988.

(i) Make a histogram or stem-and-leaf plot of the observed data values. [5 points]

(ii) [105 total points] One possible sampling model for non-negative integer-valued variables is the Poisson distribution with mean $\theta$: You could take the $Y_i$ (given $\theta$) as conditionally IID Poisson($\theta$), where the marginal sampling distribution for observation $i$ would then be

$$P(Y_i = y_i | \theta) = \left\{ \begin{array}{ll} \frac{\theta^{y_i}e^{-\theta}}{y_i!} & \text{for } y_i = 0, 1, \ldots \\ 0 & \text{otherwise} \end{array} \right\}. \tag{4}$$

(a) Verify that, if the Poisson distribution is parameterized in this way, $\theta$ is indeed the mean of this distribution; in other words, show that if $(Y_i | \theta) \sim \text{Poisson}(\theta)$ then $E(Y_i | \theta) = \theta$. Use this to show that the method-of-moments estimator $\hat{\theta}_{MoM}$ of $\theta$ is the sample mean $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$. [10 points]

(b) Show that the maximum-likelihood estimator $\hat{\theta}_{MLE}$ of $\theta$ in this model is also $\bar{Y}$. [10 points]

(c) To get an informal idea of whether the Poisson sampling model fits this data set reasonably well, complete the following table:

<table>
<thead>
<tr>
<th>$y_i$</th>
<th>Empirical $\hat{P}(Y_i = y_i)$</th>
<th>Best-Fitting Poisson $\hat{P}(Y_i = y_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.126</td>
<td>0.126</td>
</tr>
<tr>
<td>1</td>
<td>0.357</td>
<td>0.357</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\geq 7$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In this table, the empirical $\hat{P}(Y_i = y_i)$ values are just the observed relative frequencies, and the best-fitting Poisson values are obtained by computing probabilities.
from the Poisson distribution with $\theta = \bar{y}$. Informally, does the fit look good to You? Explain briefly. [10 points]

(d) Verify that in this model $\theta$ is also the variance of the distribution, and use this to create another informal check on the Poisson sampling model. [10 points]

(e) Compute a large-sample standard error for $\hat{\theta}_{MLE}$ using observed Fisher information, and use this to construct an approximate 95% confidence interval for $\theta$. [15 points]

(f) Show that the conjugate prior for $\theta$ in this Poisson sampling model is the Gamma distribution $\Gamma(\alpha, \beta)$:

$$p(\theta|\alpha \beta B) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta \theta}. \quad (5)$$

(NB This differs from the parameterization of the Gamma distribution in Casella and Berger.) [10 points]

(h) Show that the conjugate updating rule in this model is

$$\begin{cases} 
(\theta|\alpha \beta B) \sim \Gamma(\alpha, \beta) \\
(Y_i|\theta B) \sim \text{Poisson}(\theta) \\
(i = 1, \ldots, n)
\end{cases} \rightarrow (\theta|s_n \alpha \beta B) \sim \Gamma(\alpha + s_n, \beta + n), \quad (6)
$$

where $s_n = \sum_{i=1}^n y_i$ is sufficient for $\theta$ in this model. [10 points]

(i) With this parameterization of the Gamma distribution, it turns out that

$$\text{if } (\theta|\alpha \beta B) \sim \Gamma(\alpha, \beta) \text{ then } E(\theta|\alpha \beta B) = \frac{\alpha}{\beta} \text{ and } V(\theta|\alpha \beta B) = \frac{\alpha}{\beta^2}. \quad (7)$$

Use the mean expression in (7) to show that the posterior mean is a weighted average of the prior mean and the sample mean, in which the prior mean gets $\beta$ votes and the sample mean gets $n$ votes; this identifies the prior sample size in this model as $n_0 = \beta$. [10 points]

(j) Suppose that, before this study was conducted, not much was known external to the data set about $\theta$; this suggests a diffuse prior in which the prior sample size $\beta$ is small, for example $\Gamma(0.01, 0.01)$. Plot the prior, likelihood and posterior on the same graph with this $\Gamma(0.01, 0.01)$ prior and the data set given in this problem. Compute the posterior mean and SD (to get the posterior SD, use the variance expression in (7)), and compare with the MLE and its standard error in part (e). Use $\texttt{R}$ or $\texttt{Maple}$ (or some other computing environment) to compute a 95% posterior interval for $\theta$, and compare with Your approximate interval based on the MLE; briefly discuss any differences You find between the maximum-likelihood and Bayesian inferential answers. [20 points]