STAT 6530
Homework 2

Spencer King
811336941

Due: Monday, February 16 via eLC by 11:59 pm

Problems

1. Consider the following study done at the National Institute of Science and Technology. Asbestos fibers on filters were counted as part of a project to develop measurement standards for asbestos concentration. Asbestos dissolved in water was spread on a filter, and 3-mm diameter punches were taken from the filter and mounted on a transmission electron microscope. An operator counted the number of fibers in each of 23 grid squares, yielding the following counts:

   31 29 19 18 31 28 34 27 34 30 16 18
   26 27 27 18 24 22 28 24 21 17 24

   We decide to model the counts as arising from a Poisson(μ) distribution. The probability mass function for this distribution is:

   $$f(y;\mu) = \frac{\mu^y e^{-\mu}}{y!}$$

   for $y = 0, 1, 2, \ldots$, and the distribution has mean and variance both equal to μ, where the rate parameter μ must be a positive real number.

   (a) (5 points) Find the maximum likelihood estimate of μ. Show your work (don't just write the answer, even though we did this in class).

   The likelihood function is

   $$L(\mu) = \prod_{i=1}^{n} f(y_i;\mu) = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}.$$

   The log-likelihood function is

   $$\ell(\mu) = \sum_{i=1}^{n} y_i \log(\mu) - n\mu - \sum_{i=1}^{n} \log(y_i!).$$

To find the maximum likelihood estimate, we take the derivative of the log-likelihood function with respect to μ and set it equal to zero:

$$\frac{d\ell}{d\mu} = \sum_{i=1}^{n} \frac{y_i}{\mu} - n = 0$$

Solving for μ, we get

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{573}{23} \approx 24.913,$$

which is the sample mean of the counts. In this case, since the log-likelihood is concave in μ for $\mu > 0$, the critical point found by setting the first derivative to zero is the unique global maximum, so a second derivative check is unnecessary.

(b) (5 points) Find an approximate 90% confidence interval for μ.

In this case since an approximate confidence interval (CI) will do, we can use the Wald CI. Lets assume $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Poisson}(\mu)$. Then

$$\mathbb{E}[Y_i] = \mu, \qquad \text{Var}(Y_i) = \mu.$$

The sample mean is $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. By independence,

$$\text{Var}(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(Y_i) = \frac{1}{n^2} \cdot n\mu = \frac{\mu}{n}.$$

By the Central Limit Theorem,

$$\bar{Y} \approx N\left(\mu, \frac{\mu}{n}\right),$$

so

$$\frac{\bar{Y} - \mu}{\sqrt{\mu/n}} \approx N(0, 1).$$

Replacing μ in the standard error by $\bar{Y}$ gives the approximate CI

$$\mu \in \bar{Y} \pm z_{0.95}\sqrt{\frac{\bar{Y}}{n}}.$$

Here $n = 23$ and $\bar{Y} = \hat{\mu} = \frac{573}{23} \approx 24.913$. So

$$\sqrt{\frac{\bar{Y}}{n}} = \sqrt{\frac{24.913}{23}} \approx 1.041.$$

Using $z_{0.95} \approx 1.645$ for a 90% two-sided CI, the margin is

$$1.645(1.041) \approx 1.712,$$

hence the approximate 90% CI is

$$24.913 \pm 1.712 = (23.201,\ 26.625).$$

2. Consider an i.i.d. sample of size n from a distribution with probability mass function

$$f(y;p) = p(1-p)^{y-1}$$

for $y = 1, 2, 3, \ldots$, with $0 < p \leq 1$.

(a) (5 points) Find the maximum likelihood estimate of p.

The likelihood function is

$$L(p) = \prod_{i=1}^{n} f(y_i;p) = \prod_{i=1}^{n} p(1-p)^{y_i-1} = p^n(1-p)^{\sum_{i=1}^{n} y_i - 1}$$

If we let $\sum_{i=1}^{n} y_i - 1 = S - n$, then the log-likelihood function is

$$\ell(p) = n\log(p) + (S-n)\log(1-p)$$

Next, we need to take the first and second derivatives

$$\ell'(p) = \frac{n}{p} + \frac{S-n}{1-p}$$

$$\ell''(p) = -\frac{n}{p^2} - \frac{S-n}{(1-p)^2}$$

Now, we know that for $\ell''(p) < 0$ for $(0 < p < 1)$. So, ℓ is strictly concave on $(0,1)$. Therefore, an interior solution to $\ell'(p) = 0$ is the unique global maxmizer on $(0,1)$.

$$\ell'(p) = \frac{n}{p} + \frac{S-n}{1-p} = 0$$

$$p(S-n) = n(1-p)$$

$$pS - pn = n - np$$

$$\hat{p} = \frac{n}{S}$$

We must now check the boundary $p = 1$. If $S > n$ (i.e., at least one $y_i > 1$), then $S - n > 0$ and

$$L(1) = 1^n \cdot 0^(S-n) = 0$$

so the maximizer must be interior and $\hat{p} = \frac{n}{S} \in (0,1)$. If $S = n$ (i.e., all $y_i = 1$), the $L(p) = p^n$, which is increasing on $(0,1]$, so the maximum occurs at $p = 1$. Thus the final MLE is as follows:

$$\hat{p} = \begin{cases} \dfrac{n}{\sum_{i=1}^{n} y_i}, & \text{if } \sum_{i=1}^{n} y_i > n, \\ 1, & \text{if } \sum_{i=1}^{n} y_i = n. \end{cases} = \min\left(1, \frac{n}{\sum_{i=1}^{n} y_i}\right)$$

(b) (5 points) Find the Fisher information $\mathcal{I}(p)$.

We start here by making the following assumptions:

- p is an interior parameter value on the interval $0 < p < 1$ such that derivatives exist and are finite.
- Support, y, is parameter independent and does not depend on p.
- Parameter is smooth such that for each y, ℓ is twice. differentiable
- We can exchange expectation and differentiation in this case.

Now, we can use the following definition for fisher information:

$$\mathcal{I}(p) = -E[\ell''(p)].$$

So

$$\mathcal{I}(p) = -E\left[-\frac{n}{p^2} - \frac{\sum_{i=1}^{n} y_i - 1}{(1-p)^2}\right] = \frac{n}{p^2} + \frac{E[\sum_{i=1}^{n} y_i - 1]}{(1-p)^2}.$$

Since Y_i are i.i.d,

$$E\left[\sum_{i=1}^{n} y_i - 1\right] = nE[y-1] = n(\frac{1}{p} - 1) = \frac{n(1-p)}{p}.$$

Substituting back in

$$\mathcal{I}(p) = \frac{n}{p^2} + \frac{n(1-p)}{p(1-p)^2} = \frac{n}{p^2} + \frac{n}{p(1-p)} = \frac{n}{p^2(1-p)}$$

(c) (5 points) Find the asymptotic variance (meaning, the approximate variance when the sample size n is large) of the maximum likelihood estimate.

Under regularity conditions, we know that

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} N\left(0, \frac{1}{I_1(p)}\right),$$

where $I_1(p)$ is the Fisher information in one observation. Equivalently, since $I_n(p) = nI_1(p)$,

$$\hat{p} \approx N\left(p, \frac{1}{I_n(p)}\right) \text{ for large } n.$$

From this it follows that:

$$Var(\hat{p}_n) \to \frac{1}{I_n(p)} \text{ as } n \to \infty.$$

At this point, we can plug in the Fisher information to obtain the asymptotic variance:

$$Var(\hat{p}_n) \approx \frac{1}{I_n(p)} = \frac{p^2(1-p)}{n}.$$

3. Consider an i.i.d. sample of size n from a $\mathrm{N}(\mu, \sigma^2)$ distribution.

(a) (5 points) Show that the sample mean and sample variance make up a two-dimensional sufficient statistic for (μ, σ^2).

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. The joint density of $x = (x_1, \ldots, x_n)$ is

$$f_{\mu,\sigma^2}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right).$$

Using the identity

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2,$$

and noting that

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)s^2, \quad \text{where} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

we obtain

$$f_{\mu,\sigma^2}(x_1, \ldots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{(n-1)s^2 + n(\bar{x} - \mu)^2}{2\sigma^2}\right).$$

This can be written in the factorized form

$$f_{\mu,\sigma^2}(x_1, \ldots, x_n) = \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right)}_{g_{\mu,\sigma^2}(\bar{x}, s^2)} \cdot \underbrace{1}_{h(x_1, \ldots, x_n)}.$$

By the Neyman–Fisher factorization theorem, the statistic

$$T(X) = (\bar{X}, S^2)$$

is sufficient for (μ, σ^2).

(b) (5 points) Suppose we know that $\sigma^2 = 4$ but μ unknown. Find a sufficient statistic for μ.

The joint density of $x = (x_1, \ldots, x_n)$ is

$$f_{\mu}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \cdot 4}} \exp\left(-\frac{(x_i - \mu)^2}{2 \cdot 4}\right) = (2\pi \cdot 4)^{-n/2} \exp\left(-\frac{1}{8} \sum_{i=1}^{n} (x_i - \mu)^2\right).$$

Again using the identity

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2,$$

we obtain

$$f_\mu(x_1,\dots,x_n) = (2\pi\cdot 4)^{-n/2}\exp\left(-\frac{n(\bar{x}-\mu)^2}{8}\right)\exp\left(-\frac{1}{8}\sum_{i=1}^{n}(x_i-\bar{x})^2\right).$$

Hence

$$f_\mu(x_1,\dots,x_n) = \underbrace{(2\pi\cdot 4)^{-n/2}\exp\left(-\frac{n(\bar{x}-\mu)^2}{8}\right)}_{g_\mu(\bar{x})}\cdot\underbrace{\exp\left(-\frac{1}{8}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)}_{h(x_1,\dots,x_n)}.$$

By the Neyman–Fisher factorization theorem, $T(X)=\bar{X}$ is sufficient for μ.

(c) (5 points) Suppose we know that $\mu=-3$ but σ^2 is unknown. Find a sufficient statistic for σ^2.

The joint density of $x=(x_1,\dots,x_n)$ is

$$f_{\sigma^2}(x_1,\dots,x_n) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i+3)^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i+3)^2\right).$$

Let

$$T(X)=\sum_{i=1}^{n}(X_i+3)^2.$$

Then

$$f_{\sigma^2}(x_1,\dots,x_n) = \underbrace{(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}T(x)\right)}_{g_{\sigma^2}(T(x))}\cdot\underbrace{1}_{h(x_1,\dots,x_n)}.$$

By the Neyman–Fisher factorization theorem, $T(X)=\sum_{i=1}^{n}(X_i+3)^2$ is sufficient for σ^2.

4. (10 points) A social scientist wanted to estimate the proportion of school children in Boston who live in a single-parent family. She decided to use a sample size such that, with probability 0.95, the error would not exceed 0.05. How large a sample size should she use, if she has no idea of the size of that proportion?

We want $P\left(|\hat{p}-p|\le 0.05\right)=0.95$. For a proportion, the margin of error is

$$E = z_{0.975}\sqrt{\frac{p(1-p)}{n}},$$

with $z_{0.975}=1.96$ for 95% confidence. Thus we need

$$1.96\sqrt{\frac{p(1-p)}{n}}\le 0.05.$$

Since p is unknown, we use the worst-case value $p(1-p) \leq 0.25$ (attained at $p = 0.5$). Therefore,

$$1.96\sqrt{\frac{0.25}{n}} \leq 0.05.$$

Solving for n:

$$\sqrt{\frac{0.25}{n}} \leq \frac{0.05}{1.96}, \qquad \frac{0.25}{n} \leq \left(\frac{0.05}{1.96}\right)^2,$$

$$n \geq \frac{0.25}{(0.05/1.96)^2} = 0.25\left(\frac{1.96}{0.05}\right)^2 = 384.16.$$

Rounding up,

$$n = 385.$$

5. Consider collecting a sample of size $n = 1497$ from a population with the goal of estimating the proportion of the population that prefers coffee rather than tea. Suppose that the true proportion is $p = 0.53$.

```
### Code shared between part (a) and part (b)

set.seed(123)

n <- 1497
p_true <- 0.53
s_vals <- c(5, 10, 100, 1000)

# function to estimate coverage for given confidence level
   and number of simulations s
coverage_score_ci <- function(s, conf_level, n, p_true) {
  alpha <- 1 - conf_level
  z <- qnorm(1 - alpha/2)

  contain <- logical(s)

  for (i in 1:s) {

    x <- rbinom(1, size = n, prob = p_true)
    phat <- x / n

    se <- sqrt(phat * (1 - phat) / n)
    lower <- phat - z * se
    upper <- phat + z * se

    contain[i] <- (lower <= p_true && p_true <= upper)
```

```
  }

  mean(contain) * 100  # percent coverage
}
```

(a) (10 points) Use R to to simulate s different samples of size n from the population (using the true value of p to simulate the data). For each sample, construct a 95% score confidence interval for p. Report the percentage of the s confidence intervals that contain the true value of p. Do this for $s = 5, 10, 100, 1000$.

```
### 95% score (Wald) intervals
cat("95% score CI coverage (%)\n")
for (s in s_vals) {
  cov <- coverage_score_ci(s, conf_level = 0.95, n
      = n, p_true = p_true)
  cat("s =", s, ":", round(cov, 1), "%\n")
}

### 95% score CI coverage (%)
# s = 5 : 80 %
# s = 10 : 80 %
# s = 100 : 96 %
# s = 1000 : 95.4 %
```

(b) (10 points) Repeat part [(a)], but this time construct 70% score confidence intervals.

```
### 70% score (Wald) intervals
cat("70% score CI coverage (%)\n")
for (s in s_vals) {
  cov <- coverage_score_ci(s, conf_level = 0.70, n
      = n, p_true = p_true)
  cat("s =", s, ":", round(cov, 1), "%\n")
}

### 70% score CI coverage (%)
# s = 5 : 80 %
# s = 10 : 90 %
# s = 100 : 70 %
# s = 1000 : 71.2 %
```

6. Consider a sample $\mathbf{Y} = (Y_1, \ldots, Y_n)$ that we wish to use to estimate the parameter θ. Suppose that $\theta_1(\mathbf{Y})$ is an estimator of θ with $E[(\theta_1(\mathbf{Y}))^2] < \infty$ for all θ, suppose that $T(\mathbf{Y})$ is a sufficient statistic for θ, and let $\theta_2(\mathbf{Y}) = E[\theta_1(\mathbf{Y})|T(\mathbf{Y})]$. Then, for all θ,

$$E[(\theta_2(\mathbf{Y}) - \theta)^2] \leq E[(\theta_1(\mathbf{Y}) - \theta)^2],$$

and the inequality is strict unless $\theta_1(\mathbf{Y}) = \theta_2(\mathbf{Y})$.

(a) (5 points) Note that $\theta_1(\mathbf{Y})$ and $\theta_2(\mathbf{Y})$ are both estimators of θ. Interpret the statement above in the context of comparing the two estimators and what it implies about using sufficient statistics to construct estimators.

The construction

$$\theta_2(\mathbf{Y}) = \mathbb{E}[\theta_1(\mathbf{Y}) \mid T(\mathbf{Y})]$$

takes an arbitrary estimator $\theta_1(\mathbf{Y})$ and *Rao–Blackwellizes* it by averaging out all sample-to-sample variation that is *irrelevant* once we know the sufficient statistic $T(\mathbf{Y})$. The displayed inequality

$$\mathbb{E}\big[(\theta_2(\mathbf{Y}) - \theta)^2\big] \le \mathbb{E}\big[(\theta_1(\mathbf{Y}) - \theta)^2\big] \quad \text{for all } \theta$$

says that θ_2 has *no larger mean squared error* (MSE) than θ_1, uniformly over the parameter space. In other words, conditioning on a sufficient statistic cannot make an estimator worse in MSE, and typically makes it strictly better. Practically, this implies:

- Sufficiency means $T(\mathbf{Y})$ retains all information about θ in the sample.
- Therefore, any dependence of $\theta_1(\mathbf{Y})$ on $\mathbf{Y}$ beyond what is captured by $T(\mathbf{Y})$ is "noise" with respect to estimating θ.
- Replacing θ_1 by $\theta_2 = \mathbb{E}[\theta_1 \mid T]$ removes that extra noise, yielding an estimator at least as accurate (in MSE), with strict improvement unless θ_1 was already a function of T a.s.

(b) **This question is only required for students enrolled in STAT 6530.** (10 points) Provide a proof of the statement above. Hints: First, show that the two estimators have the same means, so it suffices to compare their variances. To compare the variances of the estimators, recall the following result: if X and Z are random variables and X has finite variance, then $Var(X) = E[Var(X|Z)] + Var(E[X|Z])$.

Let θ be fixed. Let $T = T(\mathbf{Y})$ and define

$$\theta_2(\mathbf{Y}) = \mathbb{E}[\theta_1(\mathbf{Y}) \mid T].$$

First note that θ_1 and θ_2 have the same mean:

$$\mathbb{E}[\theta_2(\mathbf{Y})] = \mathbb{E}[\mathbb{E}[\theta_1(\mathbf{Y}) \mid T]] = \mathbb{E}[\theta_1(\mathbf{Y})],$$

by the tower property. Now compare the MSEs. Use the conditional-variance decomposition (law of total variance):

$$\mathrm{Var}(\theta_1(\mathbf{Y})) = \mathbb{E}[\mathrm{Var}(\theta_1(\mathbf{Y}) \mid T)] + \mathrm{Var}(\mathbb{E}[\theta_1(\mathbf{Y}) \mid T])\,.$$

But $\mathbb{E}[\theta_1(\mathbf{Y}) \mid T] = \theta_2(\mathbf{Y})$, so

$$\mathrm{Var}(\theta_1(\mathbf{Y})) = \mathbb{E}[\mathrm{Var}(\theta_1(\mathbf{Y}) \mid T)] + \mathrm{Var}(\theta_2(\mathbf{Y})) \quad \ge \quad \mathrm{Var}(\theta_2(\mathbf{Y})).$$

Thus $\mathrm{Var}(\theta_2) \leq \mathrm{Var}(\theta_1)$. Next show that the two estimators have the same bias (hence the same squared bias). Since their expectations coincide,

$$\mathrm{Bias}(\theta_2) = \mathbb{E}[\theta_2(\mathbf{Y})] - \theta = \mathbb{E}[\theta_1(\mathbf{Y})] - \theta = \mathrm{Bias}(\theta_1).$$

Therefore,

$$\mathrm{MSE}(\theta_i) = \mathbb{E}[(\theta_i(\mathbf{Y}) - \theta)^2] = \mathrm{Var}(\theta_i(\mathbf{Y})) + \mathrm{Bias}(\theta_i)^2, \quad i = 1, 2,$$

and since the biases are equal while $\mathrm{Var}(\theta_2) \leq \mathrm{Var}(\theta_1)$, we conclude

$$\mathbb{E}[(\theta_2(\mathbf{Y}) - \theta)^2] \leq \mathbb{E}[(\theta_1(\mathbf{Y}) - \theta)^2].$$

Finally, determine when the inequality is strict. From

$$\mathrm{Var}(\theta_1(\mathbf{Y})) - \mathrm{Var}(\theta_2(\mathbf{Y})) = \mathbb{E}[\mathrm{Var}(\theta_1(\mathbf{Y}) \mid T)],$$

we have equality if and only if

$$\mathbb{E}[\mathrm{Var}(\theta_1(\mathbf{Y}) \mid T)] = 0 \quad \Longleftrightarrow \quad \mathrm{Var}(\theta_1(\mathbf{Y}) \mid T) = 0 \text{ a.s.}$$

But $\mathrm{Var}(\theta_1 \mid T) = 0$ a.s. holds if and only if $\theta_1(\mathbf{Y})$ is (a.s.) a function of T, equivalently

$$\theta_1(\mathbf{Y}) = \mathbb{E}[\theta_1(\mathbf{Y}) \mid T] = \theta_2(\mathbf{Y}) \quad \text{a.s.}$$

Hence the inequality is strict unless $\theta_1(\mathbf{Y}) = \theta_2(\mathbf{Y})$ almost surely. $\square$