MATH 541A, GRADUATE STATISTICS
SELECTED HOMEWORK SOLUTIONS

STEVEN HEILMAN

Contents

1. Homework 1

Exercise 1.3. Two people take turns throwing darts at a board. Person A goes first, and each of her throws has a probability of $1/4$ of hitting the bullseye. Person B goes next, and each of her throws has a probability of $1/3$ of hitting the bullseye. Then Person A goes, and so on. With what probability will Person A hit the bullseye before Person B does?

Solution. Person A hits the bullseye on her first try with probability $1/4$. If both A and B miss their first throw, then Person A hits the bullseye on her second try with probability $(1-1/4)(1-1/3)(1/4)$. If both A and B miss their first two throws, then Person A hits the bullseye on her third try with probability $(1-1/4)^2(1-1/3)^2(1/4)$. For any positive integer k, let C_k be the event that both A and B miss their first k throws, and Person A hits the bullseye on the $(k+1)^{st}$ try. Then $\mathbf{P}(C_k) = (1-1/4)^k(1-1/3)^k(1/4) = (3/4)^k(2/3)^k(1/4) = (1/2)^k(1/4)$. Let C be the event that person A hits the bullseye before person B. Then $C = \cup_{k\geq 0} C_k$, and $C_k \cap C_{k'} = \emptyset$ if $k \neq k'$. So, from the axioms for a probability law,

$$\mathbf{P}(C) = \mathbf{P}(\cup_{k\geq 0} C_k) = \sum_{k=0}^{\infty} \mathbf{P}(C_k) = (1/4)\sum_{k=0}^{\infty}(1/2)^k = (1/4)(2) = 1/2.$$

□

Exercise 1.4. Two people are flipping fair coins. Let n be a positive integer. Person I flips $n+1$ coins. Person II flips n coins. Show that the following event has probability $1/2$: Person I has more heads than Person II.

Date: May 1, 2023

Solution 1. Let A be the event that Person I has more heads than Person II. Let S_I be the number of heads from the first n coin flips of person I. Let S_{II} be the number of heads from the first n coin flips of person II. Let B_1 be the event that the $(n+1)^{st}$ coin flip of person I is heads. Let B_2 be the event that the $(n+1)^{st}$ coin flip of person I is tails. Then $B_1 \cap B_2 = \emptyset$ since the $(n+1)^{st}$ coin flip of cannot be both heads and tails. And $B_1 \cup B_2 = \Omega$, since the $(n+1)^{st}$ coin flip must be either heads or tails. So, by the total probability theorem,

$$\mathbf{P}(A) = \mathbf{P}(A|B_1)\mathbf{P}(B_1) + \mathbf{P}(A|B_2)\mathbf{P}(B_2).$$

Now, since the $(n+1)^{st}$ coin flip is a fair coin, $\mathbf{P}(B_1) = \mathbf{P}(B_2) = 1/2$. That is,

$$\mathbf{P}(A) = \frac{1}{2}\left(\mathbf{P}(A|B_1) + \mathbf{P}(A|B_2)\right).$$

Given that B_1 occurs, the event A is equal to the event that $S_I \geq S_{II}$. Given that B_2 occurs, the event A is equal to the event $S_I > S_{II}$. So,

$$\mathbf{P}(A) = \frac{1}{2}\left(\mathbf{P}(S_I \geq S_{II}) + \mathbf{P}(S_I > S_{II})\right).$$

Now, $\mathbf{P}(S_I > S_{II}) = \mathbf{P}(S_1 < S_{II})$ by symmetry (with respect to interchanging the roles of person I and person II). So,

$$\mathbf{P}(A) = \frac{1}{2}\left(\mathbf{P}(S_I \geq S_{II}) + \mathbf{P}(S_I < S_{II})\right) = \frac{1}{2}.$$

In the last line, we used that the events $S_I \geq S_{II}$ and $S_I < S_{II}$ are disjoint, and their union is all of Ω, so $\mathbf{P}(S_I \geq S_{II}) + \mathbf{P}(S_I < S_{II}) = 1$.

Solution 2. Let A be the event that Person I has more heads than Person II. Let B be the event that person I has more heads than person II after they both flip n coins. Let C be the event that person I has less heads than person II after they both flip n coins. Let D be the event that person I has the same number of heads as person II after they both flip n coins. Then $B \cap C = C \cap D = B \cap D = \emptyset$, since any such intersection involves mutually exclusive events. Also, $B \cup C \cup D = \Omega$, since after the players each flip n coins, one of the three events B, C, D must occur.

So, by the total probability theorem,

$$\mathbf{P}(A) = \mathbf{P}(A|B)\mathbf{P}(B) + \mathbf{P}(A|C)\mathbf{P}(C) + \mathbf{P}(A|D)\mathbf{P}(D).$$

Given that B has occurred, we already know that A has occurred, so that $\mathbf{P}(A|B) = 1$. Given that C has occurred, it is impossible for A to occur, so that $\mathbf{P}(A|C) = 0$. And given that D has occurred, person I has only one more coin flip; if it is a heads, then A occurs, and if it is tails, then A does not occur. Since the coin is fair, we conclude that $\mathbf{P}(A|C) = 1/2$. That is,

$$\mathbf{P}(A) = \mathbf{P}(B) + \frac{1}{2}\mathbf{P}(C) = \frac{1}{2}(2\mathbf{P}(B) + \mathbf{P}(C)).$$

To conclude, it remains to show that $2\mathbf{P}(B) + \mathbf{P}(C) = 1$. As noted already, $B \cap C = C \cap D = B \cap D = \emptyset$, and $B \cup C \cup D = \Omega$, so Axiom (ii) for Probability Laws says that

$$\mathbf{P}(B) + \mathbf{P}(C) + \mathbf{P}(D) = \mathbf{P}(B \cup C \cup D) = \mathbf{P}(\Omega) = 1.$$

Now, events B and D are symmetric with respect to relabeling the players I and II. Consequently, $\mathbf{P}(B) = \mathbf{P}(D)$. That is, $2\mathbf{P}(B) + \mathbf{P}(C) = 1$, as desired.

Solution 3. Let C_1 be the number of heads of Person I. Let C_2 be the number of heads of Person II. Let $A = \{C_1 > C_2\}$. Since $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$, we have $\mathbf{P}(A)+\mathbf{P}(A^c) = 1$. Note that $A^c = \{C_1 \leq C_2\}$. Since the coins are fair, the probability $\mathbf{P}(A^c)$ can be equivalently stated by relabeling the head and tail of the coin. That is, $\mathbf{P}(A^c)$ is equal to the probability of the event that Person I has less than or equal to the number of tails of Person II. The latter event is equal to $\{C_1 > C_2\}$. That is, $\mathbf{P}(A^c) = \mathbf{P}(C_1 > C_2) = \mathbf{P}(A)$. So, $2\mathbf{P}(A) = 1$, and $\mathbf{P}(A) = 1/2$.

Exercise 1.5. Suppose a test for a disease is 99.9% accurate. That is, if you have the disease, the test will be positive with 99.9% probability. And if you do not have the disease, the test will be negative with 99.9% probability. Suppose also the disease is fairly rare, so that roughly 1 in $20,000$ people have the disease. If you test positive for the disease, with what probability do you actually have the disease?

Solution. Let A be the event that you have the disease, and let B be the event that you test positive for the disease. We want to compute $\mathbf{P}(A|B)$. From Bayes' Rule,

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A)}{\mathbf{P}(B)}\mathbf{P}(B|A) = \frac{\mathbf{P}(A)\mathbf{P}(B|A)}{\mathbf{P}(B|A)\mathbf{P}(A) + \mathbf{P}(B|A^c)\mathbf{P}(A^c)}.$$

It is given that $\mathbf{P}(A) = 2 \cdot 10^{-4}$, $\mathbf{P}(B|A) = .999$, $\mathbf{P}(B|A^c) = .001$. Since $\mathbf{P}(A^c) + \mathbf{P}(A) = 1$, we have $\mathbf{P}(A^c) = 1 - 2 \cdot 10^{-4}$. So,

$$\mathbf{P}(A|B) = \frac{2 \cdot 10^{-4}(.999)}{.999(2 \cdot 10^{-4}) + .001(1 - 2 \cdot 10^{-4})} \approx \frac{2 \cdot 10^{-4}}{.0012} \approx \frac{1}{6}.$$

□

Exercise 1.6 (Inclusion-Exclusion Formula). In the Properties for Probability laws, we showed that $\mathbf{P}(A \cup B) = \mathbf{P}(A)+\mathbf{P}(B)-\mathbf{P}(A \cap B)$. The following equality is a generalization of this fact. Let Ω be a discrete sample space, and let $\mathbf{P}$ be a probability law on Ω. Prove the following. Let $A_1, \ldots, A_n \subseteq \Omega$. Then:

$$\mathbf{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbf{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathbf{P}(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} \mathbf{P}(A_i \cap A_j \cap A_k)$$
$$\cdots + (-1)^{n+1}\mathbf{P}(A_1 \cap \cdots \cap A_n).$$

(Hint: begin with the identity $0 = (1-1)^m = \sum_{k=0}^m (-1)^k \binom{m}{k}$, which follows from the Binomial Theorem. That is, $1 = \sum_{k=1}^m (-1)^{k+1} \binom{m}{k}$. Now, let $x \in \Omega$ such that x is in exactly m of the sets $A_1, \ldots, A_n$. Compute the "number of times" that the element $x \in \Omega$ is counted for both sides of the Inclusion-Exclusion Formula.)

Solution. Let $X := 1_{\cup_{i=1}^n A_i}$ and let $X_i := 1_{A_i}$ for all $1 \leq i \leq n$. We first show that

$$X = 1 - \prod_{i=1}^n (1 - X_i). \qquad (*)$$

To see this, note that $X(\omega) = 1$ if $\omega \in \cup_{i=1}^n A_i$ and $X(\omega) = 0$ otherwise. If $\omega \in \cup_{i=1}^n A_i$, then $\omega \in A_i$ for some $1 \leq i \leq n$, hence $X_i(\omega) = 1$ for some $1 \leq i \leq n$. Therefore the product is equal to zero, so the right-hand side is also equal to 1.

On the other hand, if $\omega \notin \cup_{i=1}^n A_i$, then $X_i(\omega) = 0$ for all $1 \leq i \leq n$. Therefore the product is equal to 1, so the right-hand side is equal to zero.

Since each $\omega \in \Omega$ is contained in either $\cup_{i=1}^n A_i$ or its complement, this shows that $X = 1 - \prod_{i=1}^n (1 - X_i)$.

We deduce the inclusion-exclusion formula by taking the expected value of both sides of $(*)$. On the left side, we get

$$\mathbf{E}X = \mathbf{P}(X = 1) = \mathbf{P}(\cup_{i=1}^n A_i)$$

For the right-hand side, expanding out the product and using the linearity of expectation shows that

$$\mathbf{E}\Big[1 - \prod_{i=1}^n (1 - X_i)\Big] = 1 - \Big[1 - \sum_{i=1}^n \mathbf{E}[X_i] + \sum_{1 \leq i < j \leq n} \mathbf{E}[X_i X_j] + \cdots + (-1)^n \mathbf{E}[X_1 \cdots X_n]\Big]$$

Moreover, $\mathbf{E}[X_i] = \mathbf{P}(X_i = 1) = \mathbf{P}(A_i)$ for all $1 \leq i \leq n$, $\mathbf{E}[X_i X_j] = \mathbf{P}(X_i X_j = 1) = \mathbf{P}(A_i \cap A_j)$ for all $1 \leq i < j \leq n$, and so on. Therefore the above expression simplifies to

$$\sum_{i=1}^n \mathbf{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathbf{P}(A_i \cap A_j) + \cdots + (-1)^{n-1} \mathbf{P}(A_1 \cap \cdots \cap A_n)$$

The proof is therefore concluded by $(*)$.

Exercise 1.7 (**Stein Identity**)**.** Let X be a standard Gaussian random variable, so that X has density $x \mapsto e^{-x^2/2}/\sqrt{2\pi}$, $\forall\ x \in \mathbb{R}$. Let $g \colon \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function such that g and g' have polynomial volume growth. That is, $\exists\ a, b > 0$ such that $|g(x)|, |g'(x)| \leq a(1 + |x|)^b$, $\forall\ x \in \mathbb{R}$. Prove the **Stein identity**

$$\mathbf{E}Xg(X) = \mathbf{E}g'(X).$$

Using this identity, recursively compute $\mathbf{E}X^k$ for any positive integer k.

Alternatively, for any $t > 0$, show that $\mathbf{E}e^{tX} = e^{t^2/2}$, i.e. compute the **moment generating function** of X. Then, using $\frac{d^k}{dt^k}|_{t=0} \mathbf{E}e^{tX} = \mathbf{E}X^k$ and using the power series expansion of the exponential, compute $\mathbf{E}X^k$ directly from the identity $\mathbf{E}e^{tX} = e^{t^2/2}$.

Solution. From the "Change of Variables" formula and integration by parts

$$\begin{aligned}\mathbf{E}Xg(X) &= \int_{-\infty}^{\infty} xg(x)e^{-x^2/2}\frac{dx}{\sqrt{2\pi}} = -\int_{-\infty}^{\infty} g(x)\frac{d}{dx}[e^{-x^2/2}]\frac{dx}{\sqrt{2\pi}} \\ &= \int_{-\infty}^{\infty} \frac{d}{dx}g(x)e^{-x^2/2}\frac{dx}{\sqrt{2\pi}} = \mathbf{E}g'(X).\end{aligned}$$

The fact that no boundary terms arise at $\pm\infty$ follows since g, g' have polynomial volume growth.

For any odd positive integer k, we have $\mathbf{E}X^k = \int_{-\infty}^{\infty} x^k e^{-x^2/2}\frac{dx}{\sqrt{2\pi}} = 0$ since the integrand is odd. When k is an even positive integer, we apply the Stein identity to get

$$\mathbf{E}X^k = \mathbf{E}XX^{k-1} = (k-1)\mathbf{E}X^{k-2} = (k-1)(k-3)\mathbf{E}X^{k-4} = \cdots = (k-1)!!.$$

Alternatively, for any $t \in \mathbb{R}$, we complete the square to get

$$\begin{aligned}\mathbf{E}e^{tX} &= \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} \\ &= \int_{-\infty}^{\infty} e^{t^2/2} e^{-(x-t)^2/2} \frac{dx}{\sqrt{2\pi}} = e^{t^2/2} \int_{-\infty}^{\infty} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = e^{t^2/2}.\end{aligned}$$

We now claim that the derivative of the function on the left exists in $t \in \mathbb{R}$. Fix $t \in \mathbb{R}$ and $h > 0$ and observe

$$h^{-1}[\mathbf{E}e^{(t+h)X} - e^{tX}] = h^{-1} \int_{-\infty}^{\infty} [e^{(t+h)x} - e^{tx}] e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \int_{-\infty}^{\infty} e^{tx} \Big(\frac{e^{hx} - 1}{h} \Big) e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}.$$

By the Dominated Convergence Theorem, the limit exists as $h \to 0$ and

$$\frac{d}{dt} \mathbf{E}e^{tX} = \int_{-\infty}^{\infty} e^{tx} x e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}.$$

(By the Mean Value Theorem, $|\frac{e^{hx}-1}{h}| = \left| x_0 e^{hx_0} \right| \leq |x| \max(e^{hx}, 1)$ for some $x_0 \in [0, x]$, so the Dominated Convergence Theorem applies.) In particular, setting $t = 0$ we get

$$\frac{d}{dt}|_{t=0} \mathbf{E}e^{tX} = \int_{-\infty}^{\infty} x e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \mathbf{E}X.$$

Repeating this argument, we see that $\mathbf{E}e^{tX}$ is infinitely differentiable in t and for any $k > 0$,

$$\frac{d^k}{dt^k}|_{t=0} \mathbf{E}e^{tX} = \int_{-\infty}^{\infty} x^k e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \mathbf{E}X^k.$$

Meanwhile, using the power series expansion of $e^{t^2/2}$, we get

$$\mathbf{E}e^{tX} = e^{t^2/2} = \sum_{k=0}^{\infty} \frac{(t^2/2)^k}{k!}.$$

Equating the k^{th} derivatives at zero of both sides gives

$$\mathbf{E}X^{2k} = \frac{d^{2k}}{dt^{2k}}|_{t=0} \frac{(t^2/2)^k}{k!} = \frac{(2k)!}{2^k k!}.$$

□

Exercise 1.8 (MAX-CUT). The probabilistic method is a very useful way to prove the existence of something satisfying some properties. This method is based upon the following elementary statement: If $\alpha \in \mathbb{R}$ and if a random variable $X \colon \Omega \to \mathbb{R}$ satisfies $\mathbf{E}X \geq \alpha$, then there exists some $\omega \in \Omega$ such that $X(\omega) \geq \alpha$. We will demonstrate this principle in this exercise.

Let $G = (V, E)$ be an undirected graph on the vertices $V = \{1, \ldots, n\}$ so that the edge set E is a subset of unordered pairs $\{i, j\}$ such that $i, j \in V$ and $i \neq j$. Let $S \subseteq V$ and denote $S^c := V \setminus S$. We refer to (S, S^c) as a cut of the graph G. The goal of the MAX-CUT problem is to maximize the number of edges going between S and S^c over all cuts of the graph G.

Prove that there exists a cut (S, S^c) of the graph such that the number of edges going between S and S^c is at least $|E|/2$. (Hint: define a random $S \subseteq V$ such that, for every $i \in V$, $\mathbf{P}(i \in S) = 1/2$, and the events $1 \in S, 2 \in S, \ldots, n \in S$ are all independent. If

$\{i,j\} \in E$, show that $\mathbf{P}(i \in S, j \notin S) = 1/4$. So, what is the expected number of edges $\{i,j\} \in E$ such that $i \in S$ and $j \notin S$?)

Solution. Define S as noted above. (Recall that these random variables can exist by a Corollary of the Kolmogorov Extension Theorem.) If $\{i,j\} \in E$, then by independence we have $\mathbf{P}(i \in S, j \notin S) = \mathbf{P}(i \in S)\mathbf{P}(j \notin S) = (1/2)^2 = 1/4$. Similarly, $\mathbf{P}(i \notin S, j \in S) = 1/4$. And the event that one of i,j is in S and the other is not in S is the disjoint union $\{i \in S, j \notin S\} \cup \{i \notin S, j \in S\}$. So, the probability that one of i,j is in S and the other is not in S is $1/4 + 1/4 = 1/2$. By linearity of expected value, the expected number of cut edges is

$$\mathbf{E} \sum_{\{i,j\} \in E} 1_{\{i \in S, j \notin S\} \cup \{i \notin S, j \in S\}} = \sum_{\{i,j\} \in E} \mathbf{E} 1_{\{i \in S, j \notin S\} \cup \{i \notin S, j \in S\}} = |E| \cdot (1/2).$$

□

Exercise 1.9. Let $n \geq 2$ be a positive integer. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. For any $x, y \in \mathbb{R}^n$, define $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$ and $\|x\| := \langle x, x \rangle^{1/2}$. Let $S^{n-1} := \{x \in \mathbb{R}^n \colon \|x\| = 1\}$ be the sphere of radius 1 centered at the origin. Let $x \in S^{n-1}$ be fixed. Let v be a random vector that is uniformly distributed in S^{n-1}. Prove:

$$\mathbf{E} \left| \langle x, v \rangle \right| \geq \frac{1}{10\sqrt{n}}.$$

Solution. We first claim that it suffices to assume that $x = (1, 0, \ldots, 0)$. To see this, note that the uniform distribution on S^{n-1} is rotation invariant. That is, for any rotation $R \colon \mathbb{R}^n \to \mathbb{R}$, and for any $x \in S^{n-1}$, we have

$$\mathbf{E} \left| \langle x, v \rangle \right| = \mathbf{E} \left| \langle x, Rv \rangle \right|.$$

The Euclidean inner product is itself invariant under rotations, that is

$$\mathbf{E} \left| \langle x, v \rangle \right| = \mathbf{E} \left| \langle x, Rv \rangle \right| = \mathbf{E} \left| \langle R^{-1}x, R^{-1}Rv \rangle \right| = \mathbf{E} \left| \langle R^{-1}x, v \rangle \right|.$$

So, if we choose the rotation R such that $R^{-1}x = (1, 0, \ldots, 0)$, we then

$$\mathbf{E} \left| \langle x, v \rangle \right| = \mathbf{E} \left| \langle (1, 0, \ldots, 0), v \rangle \right|.$$

That is, it suffices to prove the statement when $x = (1, 0, \ldots, 0)$.

Now, using hyperspherical coordinates, we can write this expected value as

$$\mathbf{E} \left| \langle (1, 0, \ldots, 0), v \rangle \right| = \frac{\int_{\phi_{n-1}=0}^{2\pi} \int_{\phi_{n-2}=0}^{\pi} \cdots \int_{\phi_1=0}^{\pi} |\cos \phi_1| \sin^{n-2} \phi_1 \sin^{n-3} \phi_2 \cdots \sin \phi_{n-2} d\phi_1 \cdots d\phi_{n-1}}{\int_{\phi_{n-1}=0}^{2\pi} \int_{\phi_{n-2}=0}^{\pi} \cdots \int_{\phi_1=0}^{\pi} \sin^{n-2} \phi_1 \sin^{n-3} \phi_2 \cdots \sin \phi_{n-2} d\phi_1 \cdots d\phi_{n-1}}.$$

The outermost integrals are the same on the top on bottom, so they cancel and we are left with

$$\mathbf{E} \left| \langle (1, 0, \ldots, 0), v \rangle \right| = \frac{\int_{\phi_1=0}^{\pi} |\cos \phi_1| \sin^{n-2} \phi_1 d\phi_1}{\int_{\phi_1=0}^{\pi} \sin^{n-2} \phi_1 d\phi_1} = \frac{\int_{\phi_1=0}^{\pi/2} \cos \phi_1 \sin^{n-2} \phi_1 d\phi_1}{\int_{\phi_1=0}^{\pi/2} \sin^{n-2} \phi_1 d\phi_1}.$$

The upper integral can be computed exactly by the substitution $u = \sin \phi_1$ so that $du = \cos \phi_1 d\phi_1$ and

$$\int_{\phi_1=0}^{\pi/2} \cos \phi_1 \sin^{n-2} \phi_1 d\phi_1 = \int_0^1 u^{n-2} du = \frac{1}{n-1}.$$

There are many ways to upper bound the lower integral. We use the inequality $\cos(x) \le e^{-x^2/2}$ valid for all $0 \le x \le \pi/2$ [Proof: $-(d/dt)\log(\cos(x)) = \tan(x) \ge x$ by e.g. the power series expansion of $\tan(x) = x + x^3/3 + 2x^5/15 + \cdots$ having all nonnegative coefficients, so $\log(\cos(x)) \le -x^2/2$, so $\cos(x) \le e^{-x^2/2}$ for all $0 \le x \le \pi/2$.], so that

$$\begin{aligned}\int_{\phi_1=0}^{\pi/2} \sin^{n-2}\phi_1 d\phi_1 &= \int_{\phi_1=0}^{\pi/2} \cos^{n-2}\phi_1 d\phi_1 \le \int_{\phi_1=0}^{\pi/2} e^{-\phi_1^2(n-2)/2} d\phi_1 \\ &\le \int_{\phi_1=0}^{\infty} e^{-\phi_1^2(n-2)/2} d\phi_1 = \int_{\phi_1=0}^{\infty} e^{-\phi_1^2(n-2)/2} d\phi_1 \\ &= (n-2)^{-1/2}\int_{x=0}^{\infty} e^{-x^2/2} dx = (n-2)^{-1/2}\sqrt{2\pi}.\end{aligned}$$

Putting everything together,

$$\mathbf{E}\,|\langle (1,0,\ldots,0), v\rangle| \ge \frac{1/(n-1)}{(n-2)^{-1/2}\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}\frac{\sqrt{n-2}}{n-1} \ge \frac{1}{10\sqrt{n}}.$$

Exercise 1.10 (The Power Method). This exercise gives an algorithm for finding the eigenvectors and eigenvalues of a symmetric matrix. In modern statistics, this is often a useful thing to do. The Power Method described below is not the best algorithm for this task, but it is perhaps the easiest to describe and analyze.

Let A be an $n \times n$ real symmetric matrix. Let $\lambda_1 \ge \cdots \ge \lambda_n$ be the (unknown) eigenvalues of A, and let $v_1, \ldots, v_n \in \mathbb{R}^n$ be the corresponding (unknown) eigenvectors of A such that $\|v_i\| = 1$ and such that $Av_i = \lambda_i v_i$ for all $1 \le i \le n$.

Given A, our first goal is to find v_1 and λ_1. For simplicity, assume that $1/2 < \lambda_1 < 1$, and $0 \le \lambda_n \le \cdots \le \lambda_2 < 1/4$. Suppose we have found a vector $v \in \mathbb{R}^n$ such that $\|v\| = 1$ and $|\langle v, v_1\rangle| > 1/n$. (From Exercise 3.1, a randomly chosen v satisfies this property.) Let k be a positive integer. Show that

$$A^k v$$

approximates v_1 well as k becomes large. More specifically, show that for all $k \ge 1$,

$$\left\|A^k v - \langle v, v_1\rangle \lambda_1^k v_1\right\|^2 \le \frac{n-1}{16^k}.$$

(Hint: use the spectral theorem for symmetric matrices.)

Since $|\langle v, v_1\rangle|\,\lambda_1^k > 2^{-k}/n$, this inequality implies that $A^k v$ is approximately an eigenvector of A with eigenvalue λ_1. That is, by the triangle inequality,

$$\left\|A(A^k v) - \lambda_1(A^k v)\right\| \le \left\|A^{k+1}v - \langle v, v_1\rangle\lambda_1^{k+1}v_1\right\| + \lambda_1\left\|\langle v, v_1\rangle\lambda_1^k v_1 - A^k v\right\| \le 2\frac{\sqrt{n-1}}{4^k}.$$

Moreover, by the reverse triangle inequality,

$$\left\|A^k v\right\| = \left\|A^k v - \langle v, v_1\rangle\lambda_1^k v_1 + \langle v, v_1\rangle\lambda_1^k v_1\right\| \ge \frac{1}{n}2^{-k} - \frac{\sqrt{n-1}}{4^k}.$$

In conclusion, if we take k to be large (say $k > 10\log n$), and if we define $z := A^k v$, then z is approximately an eigenvector of A, that is

$$\left\|A\frac{A^k v}{\|A^k v\|} - \lambda_1\frac{A^k v}{\|A^k v\|}\right\| \le 4n^{3/2}2^{-k} \le 4n^{-4}.$$

And to approximately find the first eigenvalue λ_1, we simply compute

$$\frac{z^T A z}{z^T z}.$$

That is, we have approximately found the first eigenvector and eigenvalue of A.

Remarks. To find the second eigenvector and eigenvalue, we can repeat the above procedure, where we start by choosing v such that $\langle v, v_1\rangle = 0$, $\|v\| = 1$ and $|\langle v, v_2\rangle| > 1/(10\sqrt{n})$. To find the third eigenvector and eigenvalue, we can repeat the above procedure, where we start by choosing v such that $\langle v, v_1\rangle = \langle v, v_2\rangle = 0$, $\|v\| = 1$ and $|\langle v, v_3\rangle| > 1/(10\sqrt{n})$. And so on.

Google's PageRank algorithm uses the power method to rank websites very rapidly. In particular, they let n be the number of websites on the internet (so that n is roughly 10^9). They then define an $n \times n$ matrix C where $C_{ij} = 1$ if there is a hyperlink between websites i and j, and $C_{ij} = 0$ otherwise. Then, they let B be an $n \times n$ matrix such that B_{ij} is 1 divided by the number of 1's in the i^{th} row of C, if $C_{ij} = 1$, and $B_{ij} = 0$ otherwise. Finally, they define

$$A = (.85)B + (.15)D/n$$

where D is an $n \times n$ matrix all of whose entries are 1.

The power method finds the eigenvector v_1 of A, and the size of the i^{th} entry of v_1 is proportional to the "rank" of website i.

Solution. From the spectral theorem for real symmetric matrices, there exists a basis of $\mathbb{R}^n$ of eigenvectors of A as stated in the exercise. That is, any $v \in \mathbb{R}^n$ can be written as

$$v = \sum_{i=1}^{n} \langle v, v_i\rangle v_i.$$

Since $Av_i = \lambda_i v_i$ for all $1 \le i \le n$, we then have

$$Av = \sum_{i=1}^{n} \langle v, v_i\rangle Av_i = \sum_{i=1}^{n} \langle v, v_i\rangle \lambda_i v_i$$

More generally, for any integer $k \ge 1$,

$$A^k v = \sum_{i=1}^{n} \langle v, v_i\rangle A^k v_i = \sum_{i=1}^{n} \langle v, v_i\rangle \lambda_i A^{k-1} v_i = \cdots = \sum_{i=1}^{n} \langle v, v_i\rangle \lambda_i^k v_i.$$

That is,

$$\left\|A^k v - \langle v, v_1\rangle \lambda_1^k v_1\right\| = \left\|\sum_{i=2}^{n} \langle v, v_i\rangle \lambda_i^k v_i\right\|.$$

Using the stated inequality in the exercise, namely that $|\lambda_i| \le 1/2$, we have $|\lambda_i|^k \le 2^{-k}$ for all $2 \le i \le n$, so that

$$\left\|A^k v - \langle v, v_1\rangle \lambda_1^k v_1\right\|^2 \le 4^{-k} \sum_{i=2}^{n} |\langle v, v_i\rangle| \, \|v_i\| \le 4^{-k} \sum_{i=2}^{n} \|v_i\| = (n-2)4^{-k}.$$

In the middle inequalities, we used the triangle inequality for the norm, and also the Cauchy-Schwarz inequality: $|\langle v, v_i\rangle| \le \|v\| \, \|v_i\| = 1 \cdot 1 = 1$ for all $1 \le i \le n$.

Exercise 1.11. Let X_1, Y_1 be random variables with joint PDF f_{X_1,Y_1}. Let X_2, Y_2 be random variables with joint PDF f_{X_2,Y_2}. Let $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ and let $S\colon \mathbb{R}^2 \to \mathbb{R}^2$ so that $ST(x,y) = (x,y)$ and $TS(x,y) = (x,y)$ for every $(x,y) \in \mathbb{R}^2$. Let $J(x,y)$ denote the determinant of the Jacobian of S at (x,y). Assume that $(X_2, Y_2) = T(X_1, Y_1)$. Using the change of variables formula from multivariable calculus, show that

$$f_{X_2,Y_2}(x,y) = f_{X_1,Y_1}(S(x,y))\,|J(x,y)|\,.$$

Solution. According to the change of variables theorem, if U is a "nice" subset of $\mathbb{R}^2$ and ϕ is an injective differentiable function on U, then

$$\int_{\phi(U)} f(u,v)\; dudv = \int_U f(\phi(x,y))|J\phi(x,y)|dxdy$$

where $J\phi(x,y)$ is the Jacobian of ϕ at (x,y). Since $(X_2,Y_2) = T(X_1,Y_1)$, it follows that

$$\mathbf{P}((X_2,Y_2) \in U) = \mathbf{P}(S(X_2,Y_2) \in SU) = \mathbf{P}((X_1,Y_1) \in S(U)) = \int_{S(U)} f_{X_1,Y_1}(u,v)\; dudv$$

$$= \int_U f_{X_1,Y_1}(S(x,y))|J(x,y)|\; dxdy$$

The last two lines used the definition of the joint density of X_1, X_1. Also, by definition of the joint density of X_2, Y_2, we have

$$\mathbf{P}((X_2,Y_2) \in U) = \int_U f_{X_2,Y_2}(x,y)\; dxdy$$

Combining these observations, we have shown that

$$\int_U f_{X_2,Y_2}(x,y)\; dxdy = \int_U f_{X_1,Y_1}(S(x,y))|J(x,y)|\; dxdy$$

for all "nice" subsets $U \subseteq \mathbb{R}^2$, which implies that $f_{X_2,Y_2} = f_{X_1,Y_1}(S(x,y))|J(x,y)|$ for all $(x,y) \in \mathbb{R}^2$.

Exercise 1.12. Suppose I tell you that the following list of 20 numbers is a random sample from a Gaussian random variable, but I don't tell the mean or standard deviation.

5.1715, 3.2925, 5.2172, 6.1302, 4.9889, 5.5347, 5.2269, 4.1966, 4.7939, 3.7127

5.3884, 3.3529, 3.4311, 3.6905, 1.5557, 5.9384, 4.8252, 3.7451, 5.8703, 2.7885

To the best of your ability, determine what the mean and standard deviation are of this random variable. (This question is a bit open-ended, so there could be more than one correct way of justifying your answer.)

Solution. As stated, this one was open ended, but it was a Gaussian with mean 4.5 and standard deviation 1. One reasonable answer for determining the mean would be to just average all 20 values to get 4.4426. And one reasonable answer for determining the standard deviation is to sum the squares differences of the values to 4.4426, divide by 20, then take the square root, getting an answer of 1.1784. One could also justify dividing by 19 in the last quantity instead of 20. We will discuss that more later in the course.

Exercise 1.13. Suppose I tell you that the following list of 20 numbers is a random sample from a Gaussian random variable, but I don't tell you the mean or standard deviation. Also, around one or two of the numbers was corrupted by noise, computational error, tabulation error, etc., so that it is totally unrelated to the actual Gaussian random variable.

$$-1.2045,\ -1.4829,\ -0.3616,\ -0.3743,\ -2.7298,\ -1.0601,\ -1.3298,\ 0.2554,\ 6.1865,\ 1.2185$$

$$-2.7273,\ -0.8453,\ -3.4282,\ -3.2270,\ -1.0137,\ 2.0653,\ -5.5393,\ -0.2572,\ -1.4512,\ 1.2347$$

To the best of your ability, determine what the mean and standard deviation are of this random variable. Supposing you had instead a billion numbers, and 5 or 10 percent of them were corrupted samples, can you come up with some automatic way of throwing out the corrupted samples? (Once again, there could be more than one right answer here; the question is intentionally open-ended.)

Solution. As stated, this one was open ended, but it was a Gaussian with mean -1 and standard deviation 2. I would say a reasonable answer here is to do the same thing as the previous exercise but then eliminate the two "outliers" which were 6.1865 and -5.5393. Determining whether or not these are outliers could be tricky, and it is an important part of doing statistics in the real world. How do we say for sure what is and is not an outlier, and how can be convince ourselves what to do with such data?

Exercise 1.14. Let $b_1, \ldots, b_n$ be distinct numbers, representing the quality of n people. Suppose n people arrive to interview for a job, one at a time, in a random order. That is, every possible arrival order of these people is equally likely. We can think of an arrival ordering of the people as an ordered list of the form $a_1, \ldots, a_n$, where the list $a_1, \ldots, a_n$ is a permutation of the numbers $b_1, \ldots, b_n$. Moreover, we interpret a_1 as the rank of the first person to arrive, a_2 as the rank of the second person to arrive, and so on. And all possible permutations of the numbers $b_1, \ldots, b_n$ are equally likely to occur.

For each $i \in \{1, \ldots, n\}$, upon interviewing the i^{th} person, if $a_i > a_j$ for all $1 \leq j < i$, then the i^{th} person is hired. That is, if the person currently being interviewed is better than the previous candidates, she will be hired. What is the expected number of hirings that will be made?

Solution. Let $X_i = 1$ if the i^{th} person to arrive is hired, and let $X_i = 0$ otherwise. Person 1 will always be hired, i.e. $\mathbf{P}(X_1 = 1) = 1$, so $\mathbf{E}X_1 = 1$. Since any arrival order is equally likely, $\mathbf{P}(X_2 = 1) = 1/2$. So, $\mathbf{E}X_2 = 1/2$. In general, if i is a positive integer, then $\mathbf{P}(X_i = 1) = 1/i$. This follows since any ordering of the people is equally likely, so there is a probability of $1/i$ of the i^{th} person having the largest number a_i among the numbers $a_1, \ldots, a_i$. So, $\mathbf{E}X_i = 1/i$. (More formally, fix $i \in \{1, \ldots, n\}$, and let $j \in \{1, \ldots, i\}$. Let A_j be the event that $a_j > a_k$ for every $k \in \{1, \ldots, i\}$ such that $k \neq j$. Then $\cup_{j=1}^{i} A_j = \Omega$, and $A_j \cap A_{j'} = \emptyset$ for every $j, j' \in \{1, \ldots, i\}$ with $j \neq j'$. So, $1 = \mathbf{P}(\Omega) = \sum_{j=1}^{i} \mathbf{P}(A_j)$. We now claim that $\mathbf{P}(A_j) = \mathbf{P}(A_{j'})$ for every $j, j' \in \{1, \ldots, i\}$ with $j \neq j'$. Given that this is true, it immediately follows that $\mathbf{P}(A_i) = 1/i$, as desired. To prove our claim, suppose we write any arrival order of the people as $c_1, \ldots, c_n$ where $c_1, \ldots, c_n$ are distinct elements of $\{1, \ldots, n\}$. Then for any $k < i$, any arrival order $c_1, \ldots, c_n$ where a_{c_i} exceeds $a_{c_1}, \ldots, a_{c_{i-1}}$ can be uniquely associated to the arrival order $c_1, \ldots, c_{k-1}, c_i, c_{k+1}, \ldots, c_{i-1}, c_k, c_{i+1}, \ldots, c_n$. That is, the number of orderings where the i^{th} number exceeds the previous ones is equal

to the number of orderings where the k^{th} number exceeds the first i numbers. That is, $\mathbf{P}(A_i) = \mathbf{P}(A_k)$.)

2. Homework 2

Exercise 2.1. You want to complete a set of 100 baseball cards. Cards are sold in packs of ten. Assume that each individual card in the pack has a uniformly random chance of being any element in the full set of 100 baseball cards. (In particular, there is a chance of getting identical cards in the same pack.) How many packs of cards should you buy in order to get a complete set of cards? That is, what is the expected number of cards you should buy in order to get a complete set of cards (rounded up to a multiple of ten)? (Hint: First, just forget about the packs of cards, and just think about buying one card at a time. Let N be the number of cards you need to buy in order to get a full set of cards, so that N is a random variable. More generally, for any $1 \leq i \leq 100$, let N_i be the number of cards you need to buy such that you have exactly i distinct cards in your collection (and before buying the last card, you only had $i-1$ distinct cards in your collection). Note that $N_1 = 1$. Define $N_0 = 0$. Then $N = N_{100} = \sum_{i=1}^{100}(N_i - N_{i-1})$. You are required to compute $\mathbf{E}N$. You should be able to compute $\mathbf{E}[N_i - N_{i-1}]$. This is the expected number of additional cards you need to buy after having already collected $i-1$ distinct cards, in order to see your i^{th} new card.)

Solution. As suggested, consider the random variable $N_i - N_{i-1}$. This random variable is geometrically distributed with parameter $p = \frac{100-(i-1)}{100} = \frac{101-i}{100}$, hence

$$\mathbf{E}[N_i - N_{i-1}] = \frac{100}{101-i}$$

Also, as suggested, note that

$$N = N_{100} = \sum_{i=1}^{100}(N_i - N_{i-1}).$$

So, taking expected values,

$$\mathbf{E}[N_{100}] = \sum_{i=1}^{100}\mathbf{E}[N_i - N_{i-1}] = \sum_{i=1}^{100}\frac{100}{101-i} = 100\sum_{j=1}^{100}\frac{1}{j} \approx 518.7$$

by setting $j = 101 - i$. Finally, to account for the fact that the cards come in packs of 10, round up to the nearest multiple of 10 to obtain 520. □

Exercise 2.2. You are trapped in a maze. Your starting point is a room with three doors. The first door will lead you to a corridor which lets you exit the maze after three hours of walking. The second door leads you through a corridor which puts you back to the starting point of the maze after seven hours of walking. The third door leads you through a corridor which puts you back to the starting point of the maze after nine hours of walking. Each time you are at the starting point, you choose one of the three doors with equal probability.

Let X be the number of hours it takes for you to exit the maze. Let Y be the number of the door that you initially choose.

- Compute $\mathbf{E}(X|Y = i)$ for each $i \in \{1, 2, 3\}$, in terms of $\mathbf{E}X$.
- Compute $\mathbf{E}X$.

Solution. By definition of X and Y, $\mathbf{E}(X|Y=1)=3$, $\mathbf{E}(X|Y=2)=7+\mathbf{E}X$ $\mathbf{E}(X|Y=3)=9+\mathbf{E}X$. From the Total Expectation Theorem,

$$\begin{aligned}\mathbf{E}X &= \mathbf{E}X1_{Y=1}+\mathbf{E}X1_{Y=2}+\mathbf{E}X1_{Y=3}\\ &= \mathbf{E}(X|Y=1)\mathbf{P}(Y=1)+\mathbf{E}(X|Y=2)\mathbf{P}(Y=2)+\mathbf{E}(X|Y=3)\mathbf{P}(Y=3).\end{aligned}$$

That is,

$$3\mathbf{E}X = 3+7+\mathbf{E}X+9+\mathbf{E}X.$$

That is, $\mathbf{E}X=19$. □

Exercise 2.3. Let $X_1,\dots,X_n$ be continuous random variables with joint PDF $f\colon \mathbb{R}^n\to[0,\infty)$. Assume that

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n)=\prod_{i=1}^n f_{X_i}(x_i),\qquad \forall\, x_1,\dots,x_n\in\mathbb{R}.$$

Show that $X_1,\dots,X_n$ are independent.

Solution. Let $A_1,\dots,A_n\subseteq\mathbb{R}$. By the definition of a joint distribution, and then using our assumption,

$$\begin{aligned}\mathbf{P}(X_1\in A_1,\dots,X_n\in A_n) &= \int_{A_1\times\cdots\times A_n} f_{X_1,\dots,X_n}(x_1,\dots,x_n)dx_1\cdots dx_n\\ &= \int_{A_1}\cdots\int_{A_n} f_{X_1}(x_1)\cdots f_{X_n}(x_n)dx_1\cdots dx_n\\ &= \prod_{i=1}^n\int_{A_i} f_{X_i}(x_i)dx_i=\prod_{i=1}^n\mathbf{P}(X_i\in A_i).\end{aligned}$$

Therefore, $X_1,\dots,X_n$ are independent. □

Exercise 2.4. Let $\phi\colon\mathbb{R}\to\mathbb{R}$. We say that ϕ is **convex** if, for any $x,y\in\mathbb{R}$ and for any $t\in[0,1]$, we have

$$\phi(tx+(1-t)y)\le t\phi(x)+(1-t)\phi(y).$$

Let $\phi\colon\mathbb{R}\to\mathbb{R}$. Show that ϕ is convex if and only if: for any $y\in\mathbb{R}$, there exists a constant a and there exists a function $L\colon\mathbb{R}\to\mathbb{R}$ defined by $L(x)=a(x-y)+\phi(y)$, $x\in\mathbb{R}$, such that $L(y)=\phi(y)$ and such that $L(x)\le\phi(x)$ for all $x\in\mathbb{R}$. (In the case that ϕ is differentiable, the latter condition says that ϕ lies above all of its tangent lines.)

(Hint: Suppose ϕ is convex. If x is fixed and y varies, show that $\frac{\phi(y)-\phi(x)}{y-x}$ increases as y increases. Draw a picture. What slope a should L have at x?)

Solution. Assume ϕ is convex.

Now, fix $x\in\mathbb{R}$ and let $b<x<c$. Let $M_R:=\{\frac{\phi(c)-\phi(x)}{c-x}\colon c>x\}$, $M_L:=\{\frac{\phi(x)-\phi(b)}{x-b}\colon b<x\}$ be the slopes of the secant lines through ϕ using points to the right and left of x, respectively. We claim that for any $m\in M_R, p\in M_L$, we have $m\ge p$. Let $m\in M_R, p\in M_L$. By definition, there exist $b<x<c$ such that $m=\frac{\phi(c)-\phi(x)}{c-x}$ and $p=\frac{\phi(x)-\phi(b)}{x-b}$. Let $t\in(0,1)$ such that $tb+(1-t)c=x$. Since ϕ is convex, $t\phi(b)+(1-t)\phi(c)\ge\phi(x)$. The following list of statements

are all equivalent:

$$\begin{aligned} m \geq p &\iff \frac{\phi(c)-\phi(x)}{c-x} \geq \frac{\phi(x)-\phi(b)}{x-b} \\ &\iff (x-b)(\phi(c)-\phi(x)) \geq (c-x)(\phi(x)-\phi(b)) \\ &\iff (x-b)\phi(c)+b\phi(x) \geq c\phi(x)-(c-x)\phi(b) \\ &\iff (1-t)(c-b)\phi(c)+b\phi(x) \geq c\phi(x)-t(c-b)\phi(b) \\ &\qquad\qquad \text{, using } (x-b)=(1-t)(c-b), \quad c-x=t(c-b) \\ &\iff (1-t)(c-b)\phi(c) \geq (c-b)\phi(x)-t(c-b)\phi(b) \\ &\iff (1-t)\phi(c)+t\phi(b) \geq \phi(x) \end{aligned}$$

We verified that the last line holds. We conclude that $m \geq p$. So, if $x \in \mathbb{R}$ is fixed, we can choose some $a \in \mathbb{R}$ such that $p \leq a \leq m$ for all $p \in M_L, m \in M_R$. Consider the function $L(z) = a(z-x)+\phi(x)$, $z \in \mathbb{R}$. Then $L(x) = \phi(x)$. We claim that $L(z) \leq \phi(z)$ for all $z \in \mathbb{R}$. We argue by contradiction. Suppose there is some $z \in \mathbb{R}$ with $L(z) > \phi(z)$. Without loss of generality, $z > x$. By definition of L, $a(z-x)+\phi(x) > \phi(z)$, so $a > \frac{\phi(z)-\phi(x)}{z-x}$. But $z > x$, so $\frac{\phi(z)-\phi(x)}{z-x} \in M_R$, and by choice of a, we must have $a \leq \frac{\phi(z)-\phi(x)}{z-x}$, a contradiction. Having found a contradiction, we conclude that $L(z) \leq \phi(z)$, as desired.

Now, assume: for any $y \in \mathbb{R}$, there exists a constant a and there exists a function $L\colon \mathbb{R} \to \mathbb{R}$ defined by $L(x) = a(x-y)+\phi(y)$, $x \in \mathbb{R}$, such that $L(y) = \phi(y)$ and such that $L(x) \leq \phi(x)$ for all $x \in \mathbb{R}$. We now show that ϕ is convex.

Fix $b, c \in \mathbb{R}$. Let $t \in (0,1)$. Set $y := tb+(1-t)c$. By assumption, there is a function $L(x) = a(x-y)+\phi(y)$ such that $L(x) \leq \phi(x)$ for all $x \in \mathbb{R}$. In particular,

$$a(b-y)+\phi(y) \leq \phi(b), \qquad a(c-y)+\phi(y) \leq \phi(c).$$

Multiplying by $t > 0$ and $(1-t) > 0$ respectively,

$$ta(b-y)+t\phi(y) \leq t\phi(b), \qquad (1-t)a(c-y)+(1-t)\phi(y) \leq (1-t)\phi(c)$$

Note that $t(b-y)+(1-t)(c-y) = tb+(1-t)c-y = 0$ by definition of y. So, adding the inequalities,

$$t\phi(b)+(1-t)\phi(c) \geq \phi(y)+a[t(b-y)+(1-t)(c-y)] = \phi(y) = \phi(tb+(1-t)c).$$

□

Exercise 2.5 (Jensen's Inequality). Let $X\colon \Omega \to [-\infty, \infty]$ be a random variable. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be convex. Assume that $\mathbf{E}\,|X| < \infty$ and $\mathbf{E}\,|\phi(X)| < \infty$. Then

$$\phi(\mathbf{E}X) \leq \mathbf{E}\phi(X).$$

(Hint: use Exercise 2.4 with $y := \mathbf{E}X$.) Deduce the **triangle inequality**:

$$|\mathbf{E}X| \leq \mathbf{E}\,|X|\,.$$

Solution. From Exercise 2.4, for any $y \in \mathbb{R}$, there exists a constant a and there exists a function $L\colon \mathbb{R} \to \mathbb{R}$ defined by $L(x) = a(x-y)+\phi(y)$, $x \in \mathbb{R}$, such that $L(x) \leq \phi(x)$ for all $x \in \mathbb{R}$. So, choose $y := \mathbf{E}X$. Then there exists $a \in \mathbb{R}$ such that

$$a(x-\mathbf{E}X)+\phi(\mathbf{E}X) \leq \phi(x), \qquad \forall\, x \in \mathbb{R}.$$

Taking expected values of both sides in $x = X$, we get

$$\phi(\mathbf{E}X) = \mathbf{E}[a(X - \mathbf{E}X) + \phi(\mathbf{E}X)] \leq \mathbf{E}\phi(X).$$

□

Exercise 2.6 (**Markov's Inequality**)**.** Let $X \colon \Omega \to [-\infty, \infty]$ be a random variable. Then

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|}{t}, \qquad \forall\, t > 0.$$

(Hint: multiply both sides by t and use monotonicity of $\mathbf{E}$.)

Solution. Let $t > 0$. Then $t1_{|X|>t} \leq |X|$, so taking expected values of both sides gives

$$t\mathbf{P}(|X| > t) = \mathbf{E}t1_{|X|>t} \leq \mathbf{E}\,|X|\,.$$

□

Exercise 2.7 (**The Chernoff Bound**)**.** Let X be a random variable and let $r > 0$. Define $M_X(t) := \mathbf{E}e^{tX}$ for any $t \in \mathbb{R}$. Show that, for any $t > 0$,

$$\mathbf{P}(X > r) \leq e^{-tr} M_X(t).$$

Consequently, if $X_1, \ldots, X_n$ are independent random variables with the same CDF, and if $r, t > 0$,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i > r\right) \leq e^{-trn}(M_{X_1}(t))^n.$$

For example, if $X_1, \ldots, X_n$ are independent Bernoulli random variables with parameter $0 < p < 1$, and if $r, t > 0$,

$$\mathbf{P}\left(\frac{X_1 + \cdots + X_n}{n} - p > r\right) \leq e^{-trn}(e^{-tp}[pe^t + (1-p)])^n.$$

And if we choose t appropriately, then the quantity $\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - p) > r\right)$ becomes exponentially small as either n or r become large. That is, $\frac{1}{n}\sum_{i=1}^{n} X_i$ becomes very close to its mean. Importantly, the Chernoff bound is much stronger than either Markov's or Cheyshev's inequality, since they only respectively imply that

$$\mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| > r\right) \leq \frac{2p(1-p)}{r}, \quad \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| > r\right) \leq \frac{p(1-p)}{nr^2}.$$

Solution. Since the exponential function is strictly increasing, if $t > 0$, then $X > r$ if and only if $tX > tr$ if and only if $e^X > e^r$. That is,

$$\mathbf{P}(X > r) = \mathbf{P}(tX > tr) = \mathbf{P}(e^{tX} > e^{tr}) \leq e^{-tr}\mathbf{E}e^{tX}.$$

The last line used Markov's inequality. □

Exercise 2.8 (**Confidence Intervals**)**.** Among 625 members of a bank chosen uniformly at random among all bank members, it was found that 25 had a savings account. Give an interval of the form $[a, b]$ where $0 \leq a, b \leq 625$ are integers, such that with about 95% certainty, if we sample 625 bank members independently and uniformly at random (from a very large bank membership), then the number of these people with savings accounts lies in the interval $[a, b]$. (Hint: if Y is a standard Gaussian random variable, then $\mathbf{P}(-2 \leq Y \leq 2) \approx .95$.)

Solution. For any $1 \leq i \leq 625$, let $X_i = 1$ if the i^{th} sampled bank member has a savings account, and $X_i = 0$ otherwise. We assume that $X_1, \ldots, X_{625}$ are iid with $\mathbf{P}(X_1 = 1) = 25/625 = 1/25 =: p$, $\mathbf{E}X_1 = 1/25$ and $\text{var}(X_1) = p(1-p) = (1/25)(1-(1/25)) = 24/625$. Using the Central Limit Theorem as an approximation, we have

$$\mathbf{P}\left(-2 \leq \frac{X_1 + \cdots + X_{625} - 625p}{\sqrt{625p(1-p)}} \leq 2\right) \approx .95$$

That is,

$$\mathbf{P}(25 - 2\sqrt{24} \leq X_1 + \cdots + X_{625} \leq 25 + 2\sqrt{24}) \approx .95$$

So, the number of bank members with a savings account is in the interval $[15.2, 39.4]$ with about 95% certainty. Rounding to integers, we choose $[a, b] = [15, 40]$. □

Exercise 2.9 (**Hypothesis Testing**)**.** Suppose we run a casino, and we want to test whether or not a particular roulette wheel is biased. Let p be the probability that red results from one spin of the roulette wheel. Using statistical terminology, "$p = 18/38$" is the null hypothesis, and "$p \neq 18/38$" is the alternative hypothesis. (On a standard roulette wheel, 18 of the 38 spaces are red.) For any $i \geq 1$, let $X_i = 1$ if the i^{th} spin is red, and let $X_i = 0$ otherwise.

Let $\mu := \mathbf{E}X_1$ and let $\sigma := \sqrt{\text{var}(X_1)}$. If the null hypothesis is true, and if Y is a standard Gaussian random variable

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}\right| \geq 2\right) = \mathbf{P}(|Y| \geq 2) \approx .05.$$

To test the null hypothesis, we spin the wheel n times. In our test, we reject the null hypothesis if $|X_1 + \cdots + X_n - n\mu| > 2\sigma\sqrt{n}$. Rejecting the null hypothesis when it is true is called a type I error. In this test, we set the type I error percentage to be 5%. (The type I error percentage is closely related to the p-value.)

Suppose we spin the wheel $n = 3800$ times and we get red 1868 times. Is the wheel biased? That is, can we reject the null hypothesis with around 95% certainty?

Solution. If the null hypothesis is true, then $\sigma = \sqrt{p(1-p)} = \sqrt{(18/38)(20/38)}$. Plugging in the same values, we have

$$\left|(X_1 + \cdots + X_n - np)/(\sigma\sqrt{n})\right| = \left|(1868 - 3800(18/38))/\sqrt{(18/38)(20/38)}\sqrt{3800}\right| \approx 2.2 > 2.$$

So, we can reject the null hypothesis with above 95% certainty. □

Exercise 2.10. A community has $m > 0$ families. Each family has at least one child. The largest family has $k > 0$ children. For each $i \in \{1, \ldots, k\}$, there are n_i families with i children. So, $n_1 + \cdots + n_k = m$. Choose a child randomly in the following two ways.

Method 1. First, choose one of the families uniformly at random among all of the families. Then, in the chosen family, choose one of the children uniformly at random.

Method 2. Among all of the $n_1 + 2n_2 + 3n_3 + \cdots + kn_k$ children, choose one uniformly at random.

What is the probability that the chosen child is the first-born child in their family, if you use Method 1?

What is the probability that the chosen child is the first-born child in their family, if you use Method 2?

Solution. In Method 1, if the family has i children, then the probability of choosing the first-born is $1/i$. By conditioning on each of the m families being chosen (each being equally likely), the probability that any chosen child is first born is

$$\frac{1}{m}\sum_{i=1}^{k} n_i/i.$$

In Method 2, the probability is the number of families divided by the number of children, i.e.

$$\frac{m}{\sum_{i=1}^{k} in_i}.$$

□

Exercise 2.11. Let $0 < p \leq \infty$. Show that, if $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ converge to $Y : \Omega \to \mathbb{R}$ in L_p, then $Y_1, Y_2, \ldots$ converges to Y in probability.

Then, show that the converse is false.

Solution. Let $\varepsilon > 0$ and let $0 < p < \infty$. From Markov's inequality,

$$\mathbf{P}(|Y_n - Y| > \varepsilon) = \mathbf{P}(|Y_n - Y|^p > \varepsilon^p) \leq \varepsilon^{-p}\mathbf{E}\,|Y_n - Y|^p\,.$$

By assumption, the right side converges to 0. Therefore, $Y_1, Y_2, \ldots$ converges to Y in probability.

The case $p = \infty$ follows from any case $p < \infty$ by Jensen's inequality, since e.g. $\mathbf{E}\,|Y_n - Y|^2 \leq \|Y_n - Y\|_\infty$.

To see that the converse is false, fix $0 < p < \infty$, let $\Omega := [0, 1]$ with $\mathbf{P}$ uniform on Ω and consider $Y_n := n^{1/p} 1_{[0,1/n]}$. Then $Y_1, Y_2, \ldots$ converges in probability to 0, since if $1 > \varepsilon > 0$, then $\mathbf{P}(|Y_n - 0| > \varepsilon) \leq \mathbf{P}([0, 1/n]) = 1/n \to 0$ as $n \to \infty$. However, $Y_1, Y_2, \ldots$ does not converge in L_p to 0 since $\mathbf{E}\,|Y_n - 0|^p = n/n = 1$ for all $n \geq 1$. □

Exercise 2.12. Prove the following statement. Almost sure convergence does not imply convergence in L_2, and convergence in L_2 does not imply almost sure convergence. That is, find random variables that converge in L_2 but not almost surely. Then, find random variables that converge almost surely but not in L_2.

Solution. Let $\Omega := [0, 1]$ with $\mathbf{P}$ uniform on Ω.

For any integer $n \geq 1$, write $n = 2^j + k$, where $0 \leq k < 2^j$, so that the integers j, k are uniquely determined by n. For example, $3 = 2^1 + 1$ and $6 = 2^2 + 2$

Let $Y_1, Y_2, \ldots$ so that for any $n \geq 1$, $Y_n := 1_{[k2^{-j},(k+1)2^{-j}]}$. Then $Y_1, Y_2, \ldots$ converges in L_2 to zero, since $\mathbf{E}(Y_n - 0)^2 = \mathbf{E}Y_n^2 = 2^{-2j} \to 0$ as $n \to \infty$. However, $Y_1, Y_2, \ldots$ does not converge almost surely to 0, since any $\omega \in \Omega$ has infinitely many $n \geq 1$ such that $Y_n(\omega) = 1$.

Finally, let $Z_n := n1_{[0,1/n]}$. Then $Z_1, Z_2, \ldots$ converges almost surely to 0 since $\lim_{n\to\infty} Z_n = 0$ for all $\omega \in (0, 1]$, but it does not converge in L_2, since $\mathbf{E}(Z_n - 0)^2 = \mathbf{E}Z_n^2 = n^2\mathbf{E}1_{[0,1/n]} = n \to \infty$ as $n \to \infty$. □

Exercise 2.13. Estimate the probability that 1000000 coin flips of fair coins will result in more than $501,000$ heads, using the Central Limit Theorem. (Some of the following integrals may be relevant: $\int_{-\infty}^{0} e^{-t^2/2}dt/\sqrt{2\pi} = 1/2$, $\int_{-\infty}^{1} e^{-t^2/2}dt/\sqrt{2\pi} \approx .8413$, $\int_{-\infty}^{2} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9772$, $\int_{-\infty}^{3} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9987$.) (Hint: use Bernoulli random variables.)

Casinos do these kinds of calculations to make sure they make money and that they do not go bankrupt. Financial institutions and insurance companies do similar calculations for similar reasons.

Solution. For any $1 \leq 1$, let $X_i = 1$ if the i^{th} coin flip is heads and $X_i = 0$ otherwise. We assume that $X_1, \ldots$ are iid with $\mathbf{P}(X_1 = 1) = 1/2$, $\mathbf{E}X_1 = 1/2$ and $\text{var}(X_1) = 1/4$. We want to know the probability that

$$X_1 + \cdots + X_{10^7} > 501000.$$

Equivalently, we want the probability of the event

$$\left\{X_1 + \cdots + X_{10^7} - 10^7/2 > 1000\right\} = \left\{\frac{X_1 + \cdots + X_{10^7} - 10^7/2}{\sqrt{10^6}\sqrt{1/4}} > 2\right\} =$$

Using the Central Limit Theorem as an approximation, we have the approximation

$$\begin{aligned}\mathbf{P}\left(\frac{X_1 + \cdots + X_{10^7} - 10^7/2}{\sqrt{10^6}\sqrt{1/4}} > 2\right) &\approx \int_2^\infty e^{-x^2/2} dx/\sqrt{2\pi} \\ &= 1 - \int_\infty^2 e^{-x^2/2} dx/\sqrt{2\pi} \approx 1 - .9772 = .0228.\end{aligned}$$

□

3. Homework 3

Exercise 3.1. Let $n \geq 2$ be a positive integer. Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. For any $x, y \in \mathbb{R}^n$, define $\langle x, y\rangle := \sum_{i=1}^n x_i y_i$ and $\|x\| := \langle x, x\rangle^{1/2}$. Let $S^{n-1} := \{x \in \mathbb{R}^n \colon \|x\| = 1\}$ be the sphere of radius 1 centered at the origin. Let $x \in S^{n-1}$ be fixed. Let v be a random vector that is uniformly distributed in S^{n-1}.

In modern statistics and data science, data can arise as vectors on high-dimensional spheres. A high-dimensional sphere is rather different from a low-dimensional one, so our intuition about the data in low dimensions may not apply any more in high dimensions. For example, any "equator" of the sphere has most of the mass near it, in the following sense:

For any $t > 0$, and for any $x \in S^{n-1}$ that is fixed,

$$\mathbf{P}(v \in S^{n-1} \colon |\langle v, x\rangle| > t/\sqrt{n}) \leq \frac{10}{t}.$$

(Hint: it might be helpful to use Markov's inequality.)

Solution. Argue as in Exercise 3.1 and compute an upper bound for $\mathbf{E}\,|\langle x, v\rangle|$. We first claim that it suffices to assume that $x = (1, 0, \ldots, 0)$. To see this, note that the uniform distribution on S^{n-1} is rotation invariant. That is, for any rotation $R \colon \mathbb{R}^n \to \mathbb{R}$, and for any $x \in S^{n-1}$, we have

$$\mathbf{E}\,|\langle x, v\rangle| = \mathbf{E}\,|\langle x, Rv\rangle|\,.$$

The Euclidean inner product is itself invariant under rotations, that is

$$\mathbf{E}\,|\langle x, v\rangle| = \mathbf{E}\,|\langle x, Rv\rangle| = \mathbf{E}\left|\langle R^{-1}x, R^{-1}Rv\rangle\right| = \mathbf{E}\left|\langle R^{-1}x, v\rangle\right|.$$

So, if we choose the rotation R such that $R^{-1}x = (1, 0, \ldots, 0)$, we then

$$\mathbf{E}\,|\langle x, v\rangle| = \mathbf{E}\,|\langle (1, 0, \ldots, 0), v\rangle|\,.$$

That is, it suffices to compute the expected value when $x = (1, 0, \ldots, 0)$.

Now, using hyperspherical coordinates, we can write this expected value as

$$\mathbf{E}\,|\langle (1,0,\ldots,0), v\rangle| = \frac{\int_{\phi_{n-1}=0}^{2\pi}\int_{\phi_{n-2}=0}^{\pi}\cdots\int_{\phi_1=0}^{\pi}|\cos\phi_1|\sin^{n-2}\phi_1\sin^{n-3}\phi_2\cdots\sin\phi_{n-2}d\phi_1\cdots d\phi_{n-1}}{\int_{\phi_{n-1}=0}^{2\pi}\int_{\phi_{n-2}=0}^{\pi}\cdots\int_{\phi_1=0}^{\pi}\sin^{n-2}\phi_1\sin^{n-3}\phi_2\cdots\sin\phi_{n-2}d\phi_1\cdots d\phi_{n-1}}.$$

The outermost integrals are the same on the top on bottom, so they cancel and we are left with

$$\mathbf{E}\,|\langle (1,0,\ldots,0), v\rangle| = \frac{\int_{\phi_1=0}^{\pi}|\cos\phi_1|\sin^{n-2}\phi_1 d\phi_1}{\int_{\phi_1=0}^{\pi}\sin^{n-2}\phi_1 d\phi_1} = \frac{\int_{\phi_1=0}^{\pi/2}\cos\phi_1\sin^{n-2}\phi_1 d\phi_1}{\int_{\phi_1=0}^{\pi/2}\sin^{n-2}\phi_1 d\phi_1}.$$

The upper integral can be computed exactly by the substitution $u = \sin\phi_1$ so that $du = \cos\phi_1 d\phi_1$ and

$$\int_{\phi_1=0}^{\pi/2}\cos\phi_1\sin^{n-2}\phi_1 d\phi_1 = \int_0^1 u^{n-2}du = \frac{1}{n-1}.$$

There are many ways to lower bound the lower integral. We use the inequality $\cos(x) \geq e^{-x^2}$ valid for all $0 \leq x \leq 1/5$ [Proof: By e.g. Taylor series expansion, $\cos(x) = 1 - x^2/2 + c(x)x^4/24$ where $|c(x)| \leq 1$ and $e^{-x^2} = 1 - x^2 + 2x^4 b(x)$ where $|b(x)| \leq 1$ for all $0 \leq x \leq 1/5$, so $x^2/2 \geq 3x^4$ for all $0 \leq x \leq 1/5$.], so that

$$\begin{aligned}\int_{\phi_1=0}^{\pi/2}\sin^{n-2}\phi_1 d\phi_1 &= \int_{\phi_1=0}^{\pi/2}\cos^{n-2}\phi_1 d\phi_1 \\ &\geq \int_{\phi_1=0}^{1/5} e^{-\phi_1^2(n-2)}d\phi_1 = (n-2)^{-1/2}\int_{\phi_1=0}^{\sqrt{n-2}/5} e^{-\phi_1^2}d\phi_1 \\ &\geq \frac{1}{10}(n-2)^{-1/2}.\end{aligned}$$

Putting everything together,

$$\mathbf{E}\,|\langle (1,0,\ldots,0), v\rangle| \leq \frac{10/(n-1)}{(n-2)^{-1/2}\sqrt{2\pi}} = \frac{10}{\sqrt{2\pi}}\frac{\sqrt{n-2}}{n-1} \leq \frac{10}{\sqrt{n}}.$$

Markov's inequality completes the proof

Exercise 3.2. Let X be uniformly distributed on $[0, 1]$. Show that the location family of X is not an exponential family in the following sense. The corresponding densities $\{f(x+\mu)\}_{\mu\in\mathbb{R}}$ cannot be written in the form

$$h(x)\exp(w(\mu)t(x) - a(w(\mu)))$$

where $h\colon \mathbb{R} \to [0, \infty)$, $w\colon \mathbb{R} \to \mathbb{R}$, $t\colon \mathbb{R} \to \mathbb{R}$, $x \in \mathbb{R}$ and $a(w(\mu))$ is a real number chosen so that the integral of the density is one. (Hint: Argue by contradiction. Assume that the location family is a one-parameter exponential family. Compare where the different densities are zero or nonzero as the parameter changes.)

Solution. The exponential term is positive, so h must be zero whenever $f(x+\mu)$ is zero, for every $\mu \in \mathbb{R}$. So, it is impossible to write the location family in this way.

Exercise 3.3. Suppose we have a k-parameter exponential family in canonical form so that

$$f_w(x) := h(x) \exp\Big(\sum_{i=1}^k w_i t_i(x) - a(w) \Big),$$

$$a(w) := \log \int_{\mathbb{R}^n} h(x) \exp\Big(\sum_{i=1}^k w_i t_i(x) \Big) d\mu(x).$$

$$W := \{ w \in \mathbb{R}^k \colon a(w) < \infty \}.$$

Show that $a(w)$ is a convex function. That is, for any $w_1, w_2 \in \mathbb{R}^k$ and for any $t \in (0,1)$,

$$a(tw_1 + (1-t)w_2) \le t a(w_1) + (1-t) a(w_2).$$

(Hint: use Hölder's inequality of the form $\int |fg| \, d\mu \le (\int |f|^p \, d\mu)^{1/p} (\int |g|^q \, d\mu)^{1/q}$ where $1/p + 1/q = 1$, where $p = t^{-1}$.)

Conclude that the set W is a convex set. (That is, if $w_1, w_2 \in W$ then for any $t \in [0,1]$, $tw_1 + (1-t)w_2 \in W$.)

Solution. Let $p = 1/t$ so that $1/p' = 1 - 1/p = 1 - t$

$$\begin{aligned}
a(tw_1 + (1-t)w_2) &= \log \int_{\mathbb{R}^n} h(x) \exp\Big(\sum_{i=1}^k (t w_{1,i} + ((1-t) w_{2,i})) t_i(x) \Big) d\mu(x) \\
&= \log \int_{\mathbb{R}^n} [h(x)]^{\frac{1}{p} + \frac{1}{p'}} \exp\Big(\sum_{i=1}^k t w_{1,i} t_i(x) \Big) \exp\Big(\sum_{i=1}^k (1-t) w_{2,i} t_i(x) \Big) d\mu(x) \\
&\le \log \Big[\Big(\int_{\mathbb{R}^n} h(x) \exp\Big(p \sum_{i=1}^k t w_{1,i} t_i(x) \Big) d\mu(x) \Big)^{1/p} \\
&\qquad \cdot \Big(\int_{\mathbb{R}^n} h(x) \exp\Big(p' \sum_{i=1}^k (1-t) w_{2,i} t_i(x) \Big) d\mu(x) \Big)^{1/p} \Big] \\
&= \frac{1}{p} \log \Big(\int_{\mathbb{R}^n} h(x) \exp\Big(p \sum_{i=1}^k t w_{1,i} t_i(x) \Big) d\mu(x) \\
&\qquad + \frac{1}{p'} \log \int_{\mathbb{R}^n} h(x) \exp\Big(p' \sum_{i=1}^k (1-t) w_{2,i} t_i(x) \Big) d\mu(x) \Big] \\
&= \frac{1}{p} \log \Big(\int_{\mathbb{R}^n} h(x) \exp\Big(\sum_{i=1}^k w_{1,i} t_i(x) \Big) d\mu(x) \\
&\qquad + \frac{1}{p'} \log \int_{\mathbb{R}^n} h(x) \exp\Big(\sum_{i=1}^k w_{2,i} t_i(x) \Big) d\mu(x) \Big] = t a(w_1) + (1-t) a(w_2).
\end{aligned}$$

Exercise 3.4. Using a two parameter exponential family for a Gaussian random variable (with mean μ and variance σ^2), compute both sides of the following identity in terms of μ

and σ:

$$e^{-a(w)}\frac{\partial^2}{\partial w_i \partial w_j}e^{a(w)} = \int_{\mathbb{R}} t_i(x)t_j(x)h(x)\exp\Big(\sum_{i=1}^{2} w_i t_i(x) - a(w)\Big)d\mu(x), \qquad \forall\, 1 \le i,j \le 2.$$

Recall that in this case,

$$t_1(x) := x, \qquad t_2(x) := x^2, \qquad w_1 := \frac{\mu}{\sigma^2}, \qquad w_2 := -\frac{1}{2\sigma^2},$$

$$a(w) := -\frac{w_1^2}{4w_2} - \frac{1}{2}\log(-2w_2).$$

Solution. For $i = j = 1$, we have

$$e^{-a(w)}\frac{\partial^2}{\partial w_i \partial w_j}e^{a(w)} = \frac{\partial^2}{\partial w_1^2}a(w) + \Big(\frac{\partial}{\partial w_1}a(w)\Big)^2 = -\frac{1}{2w_2} + \frac{w_1^2}{4w_2^2}$$
$$= \int_{\mathbb{R}} x^2 h(x)\exp\Big(\sum_{i=1}^{2} w_i t_i(x) - a(w)\Big)d\mu(x) = \mathbf{E}X^2,$$

So, $\mathbf{E}X^2 = \sigma^2 + \mu^2\sigma^{-4}\sigma^4 = \sigma^2 + \mu^2$, where X is a Gaussian with mean μ and variance σ^2.

For $i = j = 2$, we have

$$e^{-a(w)}\frac{\partial^2}{\partial w_i \partial w_j}e^{a(w)} = \frac{\partial^2}{\partial w_2^2}a(w) + \Big(\frac{\partial}{\partial w_2}a(w)\Big)^2 = -\frac{w_1^2}{2w_2^3} + \frac{1}{2}w_2^{-2} + (\frac{w_1^2}{4w_2^2} - \frac{1}{2}w_2^{-1})^2$$
$$= \int_{\mathbb{R}} x^4 h(x)\exp\Big(\sum_{i=1}^{2} w_i t_i(x) - a(w)\Big)d\mu(x) = \mathbf{E}X^4,$$

So, $\mathbf{E}X^4 = \mu^2\sigma^{-4}4\sigma^6 + 2\sigma^4 + (\mu^2+\sigma^2)^2 = 4\mu^2\sigma^2 + 2\sigma^4 + (\mu^2+\sigma^2)^2 = \mu^4 + 6\sigma^2\mu^2 + 3\sigma^4$, where X is a Gaussian with mean μ and variance σ^2.

For $i = 1, j = 2$, we have

$$e^{-a(w)}\frac{\partial^2}{\partial w_i \partial w_j}e^{a(w)} = \frac{\partial^2}{\partial w_1 \partial w_2}a(w) + \frac{\partial}{\partial w_1}a(w)\frac{\partial}{\partial w_2}a(w) = \frac{w_1}{2w_2^2} - \frac{w_1}{2w_2}\Big(\frac{w_1^2}{4w_2^2} - \frac{1}{2w_2}\Big)$$
$$= \int_{\mathbb{R}} x^3 h(x)\exp\Big(\sum_{i=1}^{2} w_i t_i(x) - a(w)\Big)d\mu(x) = \mathbf{E}X^3,$$

So, $\mathbf{E}X^3 = 2\mu\sigma^2 + \mu(\mu^2+\sigma^2) = \mu^3 + 3\mu\sigma^2$, where X is a Gaussian with mean μ and variance σ^2.

Exercise 3.5. Let $X\colon \Omega \to \mathbb{R}^n$ be a random variable with the **standard Gaussian distribution**:

$$\mathbf{P}(X \in A) := \int_A e^{-(x_1^2+\cdots+x_n^2)/2}dx(2\pi)^{-n/2}, \qquad \forall\, A \subseteq \mathbb{R}^n \text{ measurable.}$$

Let $v_1, \ldots, v_m$ be vectors in $\mathbb{R}^n$. Let $\langle\cdot,\cdot\rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be the standard inner product on $\mathbb{R}^n$, so that $\langle x, y\rangle := \sum_{i=1}^n x_i y_i$ for any $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{R}^n$.

First, let $v \in \mathbb{R}^n$ and show that $\langle X, v\rangle$ is a mean zero Gaussian with variance $\langle v, v\rangle$.

Then, show that the random variables $\langle X, v_1\rangle, \ldots, \langle X, v_m\rangle$ are independent if and only if the vectors $v_1, \ldots, v_m$ are pairwise orthogonal.

(Hint: use the rotation invariance of the Gaussian.)

Solution. Without loss of generality, all of the vectors are nonzero. Suppose for now that $m \leq n$. Suppose the vectors $v_1, \ldots, v_m$ are pairwise orthogonal. For any $1 \leq i \leq n$, let $e_i \in \mathbb{R}^n$ be the vector with a 1 in the i^{th} entry and zeros in all other entries. Let Q be any $n \times n$ real orthogonal matrix such that $Qe_i = v_i / \|v_i\|$ for all $1 \leq i \leq m$ (this is possible since $m \leq n$). Specifically, we let the first m columns of Q be $v_1/\|v_1\|, \ldots, v_m/\|v_m\|$. Suppose the remaining columns of Q are $Q_{m+1}, \ldots, Q_n$. Recall that $Q^{-1} = Q^T$ so that $QQ^T = Q^TQ = I_n$, where I_n denotes the $n \times n$ identity matrix. Note the rows of Q^T are $v_1/\|v_1\|, \ldots, v_m/\|v_m\|, Q_{m+1}, \ldots, Q_n$, so

$$Q \begin{pmatrix} \langle X, \frac{v_1}{\|v_1\|} \rangle \\ \vdots \\ \langle X, \frac{v_m}{\|v_m\|} \rangle \\ \langle X, Q_{m+1} \rangle \\ \vdots \\ \langle X, Q_n \rangle \end{pmatrix} = QQ^TX = X. \qquad (*)$$

So, if $A_1, \ldots, A_m \subseteq \mathbb{R}$, we have

$$\begin{aligned}
&\mathbf{P}(\langle X, v_1 \rangle \in A_1, \ldots, \langle X, v_m \rangle \in A_m) = \mathbf{P}(\,(\langle X, v_1 \rangle, \ldots, \langle X, v_m \rangle) \in A_1 \times \cdots \times A_m) \\
&= \mathbf{P}(\,(\langle X, v_1 \rangle, \ldots, \langle X, v_m \rangle, \langle X, Q_{m+1} \rangle, \ldots, \langle X, Q_n \rangle) \in A_1 \times \cdots \times A_m \times \mathbb{R}^{n-m}) \\
&= \mathbf{P}(\,(\langle X, \frac{v_1}{\|v_1\|} \rangle, \ldots, \langle X, \frac{v_m}{\|v_m\|} \rangle, \langle X, Q_{m+1} \rangle, \ldots, \langle X, Q_n \rangle) \in \frac{A_1}{\|v_1\|} \times \cdots \times \frac{A_m}{\|v_m\|} \times \mathbb{R}^{n-m}) \\
&= \mathbf{P}(Q^TQ(\langle X, \frac{v_1}{\|v_1\|} \rangle, \ldots, \langle X, \frac{v_m}{\|v_m\|} \rangle, \langle X, Q_{m+1} \rangle, \ldots, \langle X, Q_n \rangle)^T \in \frac{A_1}{\|v_1\|} \times \cdots \times \frac{A_m}{\|v_m\|} \times \mathbb{R}^{n-m}) \\
&= \mathbf{P}(Q(\langle X, \frac{v_1}{\|v_1\|} \rangle, \ldots, \langle X, \frac{v_m}{\|v_m\|} \rangle, \langle X, Q_{m+1} \rangle, \ldots, \langle X, Q_n \rangle)^T \in Q(\frac{A_1}{\|v_1\|} \times \cdots \times \frac{A_m}{\|v_m\|} \times \mathbb{R}^{n-m})) \\
&= \mathbf{P}(X \in Q(\frac{A_1}{\|v_1\|} \times \cdots \times \frac{A_m}{\|v_m\|} \times \mathbb{R}^{n-m})) \qquad \text{, by } (*) \\
&= \int_{Q(\frac{A_1}{\|v_1\|} \times \cdots \times \frac{A_m}{\|v_m\|} \times \mathbb{R}^{n-m})} e^{-(x_1^2 + \cdots + x_n^2)/2} dx (2\pi)^{-n/2} \quad \text{, by definition of } X \\
&= \int_{\frac{A_1}{\|v_1\|} \times \cdots \times \frac{A_m}{\|v_m\|} \times \mathbb{R}^{n-m}} e^{-x^TQ^TQx/2} dx (2\pi)^{-n/2} \quad \text{, changing variables} \\
&= \int_{\frac{A_1}{\|v_1\|} \times \cdots \times \frac{A_m}{\|v_m\|} \times \mathbb{R}^{n-m}} e^{-\|x\|^2/2} dx (2\pi)^{-n/2} = \prod_{i=1}^m \int_{A_i/\|v_i\|} e^{-x_i^2/2} dx_i/\sqrt{2\pi} \quad \text{, by Fubini's Theorem} \\
&= \prod_{i=1}^m \mathbf{P}(\langle X, \frac{v_i}{\|v_i\|} \rangle \in \frac{A_i}{\|v_i\|}) = \prod_{i=1}^m \mathbf{P}(\langle X, v_i \rangle \rangle \in A_i).
\end{aligned}$$

In the penultimate equality, we used that $\langle X, \frac{v_i}{\|v_i\|} \rangle$ is a standard one-dimensional Gaussian random variable. This follows from the $m = 1$ case of the above.

More generally, for any linearly independent vectors $v_1, \ldots, v_m$, we construct Q as above, where now Q may not be orthogonal, and

$$(QQ^T)^{-1}Q\begin{pmatrix} \langle X, \frac{v_1}{\|v_1\|}\rangle \\ \vdots \\ \langle X, \frac{v_m}{\|v_m\|}\rangle \\ \langle X, Q_{m+1}\rangle \\ \vdots \\ \langle X, Q_n\rangle \end{pmatrix} = (QQ^T)^{-1}QQ^TX = X. \qquad (*)$$

So we then get

$$\begin{aligned}
&\mathbf{P}(\langle X, v_1\rangle \in A_1, \ldots, \langle X, v_m\rangle \in A_m) = \mathbf{P}(\,(\langle X, v_1\rangle, \ldots, \langle X, v_m\rangle) \in A_1 \times \cdots \times A_m) \\
&= \mathbf{P}(\,(\langle X, v_1\rangle, \ldots, \langle X, v_m\rangle, \langle X, Q_{m+1}\rangle, \ldots, \langle X, Q_n\rangle) \in A_1 \times \cdots \times A_m \times \mathbb{R}^{n-m}) \\
&= \mathbf{P}(\,(\langle X, \frac{v_1}{\|v_1\|}\rangle, \ldots, \langle X, \frac{v_m}{\|v_m\|}\rangle, \langle X, Q_{m+1}\rangle, \ldots, \langle X, Q_n\rangle) \in \frac{A_1}{\|v_1\|} \times \cdots \times \frac{A_m}{\|v_m\|} \times \mathbb{R}^{n-m}) \\
&= \mathbf{P}((QQ^T)^{-1}Q(\langle X, \frac{v_1}{\|v_1\|}\rangle, \ldots, \langle X, \frac{v_m}{\|v_m\|}\rangle, \langle X, Q_{m+1}\rangle, \ldots, \langle X, Q_n\rangle)^T \\
&\qquad\qquad \in (QQ^T)^{-1}Q(\frac{A_1}{\|v_1\|} \times \cdots \times \frac{A_m}{\|v_m\|} \times \mathbb{R}^{n-m})) \\
&= \mathbf{P}(X \in (QQ^T)^{-1}Q(\frac{A_1}{\|v_1\|} \times \cdots \times \frac{A_m}{\|v_m\|} \times \mathbb{R}^{n-m})) \qquad \text{, by } (*) \\
&= \int_{(QQ^T)^{-1}Q(\frac{A_1}{\|v_1\|}\times\cdots\times\frac{A_m}{\|v_m\|}\times\mathbb{R}^{n-m})} e^{-(x_1^2+\cdots+x_n^2)/2}dx(2\pi)^{-n/2} \quad \text{, by definition of } X \\
&= |\det(Q)|^{-1} \int_{\frac{A_1}{\|v_1\|}\times\cdots\times\frac{A_m}{\|v_m\|}\times\mathbb{R}^{n-m}} e^{-x^TQ^T(QQ^T)^{-2}Qx/2}dx(2\pi)^{-n/2} \quad \text{, changing variables}
\end{aligned}$$

Letting D be the diagonal matrix with diagonal entries $\|v_1\|^{-1}, \ldots, \|v_m\|^{-1}, 1, \ldots, 1$,

$$\begin{aligned}
&\mathbf{P}(\langle X, v_1\rangle \in A_1, \ldots, \langle X, v_m\rangle \in A_m) \\
&= |\det(Q)|^{-1} \int_{D(A_1\times\cdots\times A_m\times\mathbb{R}^{n-m})} e^{-x^TQ^T(QQ^T)^{-2}Qx/2}dx(2\pi)^{-n/2} \\
&= |\det(Q)|^{-1}\,|\det(D)| \int_{A_1\times\cdots\times A_m\times\mathbb{R}^{n-m}} e^{-x^TDQ^T(QQ^T)^{-2}QDx/2}dx(2\pi)^{-n/2}
\end{aligned}$$

So, the joint density of $\langle X, v_1\rangle, \ldots, \langle X, v_m\rangle$ is

$$|\det(Q)|^{-1}\,|\det(D)| \int_{\mathbb{R}^{n-m}} e^{-x^TDQ^T(QQ^T)^{-2}QDx/2}dx_{n-m+1}\cdots dx_n(2\pi)^{-n/2}.$$

In the special case $n = m = 2$ with $\|v_1\| = \|v_2\| = 1$, the joint density becomes

$$|\det(Q)|^{-1}\, e^{-x^TQ^T(QQ^T)^{-2}Qx/2}(2\pi)^{-n/2}.$$

In this case Q is the matrix whose columns are v_1, v_2. Let $B := Q^T(QQ^T)^{-2}Q$. Suppose we write Q in its singular value decomposition as $Q = UDV$ where U, V are orthogonal and D is diagonal. If $v_1 = \pm v_2$, then $\langle X, v_1\rangle, \langle X, v_2\rangle$ are dependent, so we may assume that $v_1 \neq v_2$ and $v_1 \neq -v_2$. Then Q has rank 2, so D has nonzero diagonal entries. Since $Q^T = V^T D U^T$, using $UU^T = I$ and $VV^T = I$, we get

$$Q^T(QQ^T)^{-2}Q = V^T D^{-2} V.$$

If v_1, v_2 are not orthogonal, $D \neq I$, so $Q^T(QQ^T)^{-2}Q$ is not a diagonal matrix. Therefore, $|\det(Q)|^{-1}\, e^{-x^T Q^T (QQ^T)^{-2} Qx/2}(2\pi)^{-n/2}$ is not a product function, so $\langle X, v_1\rangle, \langle X, v_2\rangle$ are not independent. The general case then follows from the $m = n = 2$ case.

□

Exercise 3.6. Recall that the gamma distribution has two parameters $\alpha, \beta > 0$.

- Show that the gamma distribution is a 2-parameter exponential family.
- Find the mean and variance of a gamma distributed random variable by differentiating the exponential family.
- Find the moment generating function of a gamma distributed random variable, and use it to find the distribution of $\sum_{i=1}^n X_i$ where $X_1, \ldots, X_n$ are independent, and X_i has gamma distribution with parameters α_i and β for all $1 \leq i \leq n$.

  You may use without proof the following uniqueness result about moment generating functions (MGFs): If Y and Z are two random variables whose MGFs coincide in a neighborhood of 0 ($\exists \delta > 0$ for which $M_Y(u) = M_Z(u) < \infty$ for all $u \in [-\delta, \delta]$), then Y and Z have the same distribution.

Solution. Let $\alpha, \beta > 0$. Recall that the gamma distribution has density

$$f(x) := \begin{cases} \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0. \end{cases}$$

We write

$$x^{\alpha-1}e^{-x/\beta} = e^{-x/\beta + (\alpha-1)\log x}.$$

Then, if $h(x) = 1_{x>0}$, we have

$$f(x) = h(x) e^{-x/\beta + (\alpha-1)\log x - \alpha \log \beta - \log \Gamma(\alpha)}.$$

So, if we define $w_1(\alpha, \beta) := -1/\beta$, $w_2(\alpha, \beta) := \alpha - 1$, $t_1(x) := x$, $t_2(x) := \log x$, and $a(w(\alpha, \beta)) := \alpha \log \beta + \log \Gamma(\alpha)$, then we have

$$f(x) = h(x) \exp\Big(\sum_{i=1}^2 w_i(\alpha, \beta) t_i(x) - a(w(\alpha, \beta)) \Big).$$

That is, the Gamma distribution is a two-parameter exponential family, where $\alpha, \beta > 0$

From Example 3.13 in the notes, we have

$$\frac{\partial}{\partial \beta} a(w(\alpha, \beta)) = \mathbf{E}_{\alpha,\beta}(\sum_{i=1}^2 \frac{\partial w_i}{\partial \beta} t_i) = \mathbf{E}_{\alpha,\beta}(\beta^{-2} x)$$

So, the expected value of the Gamma distributed random variable is

$$\beta^2 \frac{\partial}{\partial \beta} a(w(\alpha, \beta)) = \beta^2 \beta^{-1} \alpha = \alpha\beta.$$

Recall that

$$a(w(\alpha,\beta)) = \log \int_{\mathbb{R}} h(x) \exp\Big(\sum_{i=1}^{2} w_i(\alpha,\beta) t_i(x) \Big) d\mu(x).$$

By differentiating $a(w(\alpha,\beta))$ twice, we get

$$\begin{aligned} e^{-a(w(\alpha,\beta))} \frac{\partial^2}{\partial \beta^2} e^{a(w(\alpha\beta))} &= \mathbf{E}_{\alpha,\beta}\Big((\sum_{i=1}^{2} \frac{\partial w_i}{\partial \beta} t_i)^2 + \sum_{i=1}^{2} \frac{\partial^2 w_i}{\partial \beta^2} t_i \Big) \\ &= \mathbf{E}_{\alpha,\beta}\Big((\beta^{-2} x)^2 - 2\beta^{-3} x \Big). \end{aligned}$$

That is,

$$\begin{aligned} \beta^{-4} \mathbf{E}_{\alpha,\beta} x^2 &= e^{-a(w(\alpha,\beta))} \frac{\partial^2}{\partial \beta^2} e^{a(w(\alpha,\beta))} + 2\beta^{-3} \mathbf{E}_{\alpha,\beta} x \\ &= \frac{\partial^2}{\partial \beta^2} a(w(\alpha,\beta)) + \Big(\frac{\partial}{\partial \beta} a(w(\alpha,\beta)) \Big)^2 + 2\beta^{-3}\alpha\beta \\ &= -\alpha\beta^{-2} + \alpha^2\beta^{-2} + 2\beta^{-3}\alpha\beta = \alpha\beta^{-2} + \alpha^2\beta^{-2}. \end{aligned}$$

So, the second moment is $\alpha\beta^2 + \alpha^2\beta^2$. Therefore, the variance is

$$\alpha\beta^2 - (\alpha\beta)^2 = \alpha\beta^2 + \alpha^2\beta^2 - \alpha^2\beta^2 = \alpha\beta^2.$$

We find the moment generating function directly. Let $z \in \mathbb{R}$ with $z < 1/\beta$. Then

$$\begin{aligned} \mathbf{E}_{\alpha,\beta} e^{zX} &= \int_0^\infty \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} e^{zx} dx = \int_0^\infty \frac{x^{\alpha-1} e^{-x(-z+1/\beta)}}{\beta^\alpha \Gamma(\alpha)} dx \\ &= (1-z\beta)^{-\alpha} \int_0^\infty \frac{x^{\alpha-1} e^{-x\frac{1}{\beta/(1-t\beta)}}}{\beta^\alpha (1-z\beta)^{-\alpha} \Gamma(\alpha)} dx = (1-z\beta)^{-\alpha}. \end{aligned}$$

The last equality used that the integral of the gamma density is one.

If $X_1, \ldots, X_n$ are independent and gamma distributed with parameters $\alpha_1, \ldots, \alpha_n > 0$ and $\beta > 0$, then by independence, for any $z < 1/\beta$,

$$\mathbf{E} e^{z(\sum_{i=1}^n X_i)} = \prod_{i=1}^n e^{zX_i} = \prod_{i=1}^n (1-z\beta)^{-\alpha_i} = (1-z\beta)^{-\sum_{i=1}^n \alpha_i}.$$

Inverting the moment generating function (by Theorem 9.2 in the notes), we conclude that $\sum_{i=1}^n X_i$ is gamma distributed with parameters $\sum_{i=1}^n \alpha_i$ and β, since this random variable has this moment generating function.

□

Exercise 3.7. Let $n \geq 2$ be an integer. Let $X_1, \ldots, X_n$ be a random sample of size n. Assume that $\mu := \mathbf{E}X \in \mathbb{R}$ and $\sigma := \sqrt{\mathrm{var}(X)} < \infty$. Let $\overline{X}$ be the sample mean and let S be the sample standard deviation of the random sample. Show the following

- $\mathrm{Var}(\overline{X}) = \sigma^2/n$.
- $\mathbf{E}S^2 = \sigma^2$.

Solution. The first equation follows since the random variables are independent, so the variances add, so that

$$\mathrm{Var}(\overline{X}) = n^{-2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = n^{-2}n\sigma^2 = \sigma^2/n.$$

For the second identity, we have

$$\begin{aligned}
\mathbf{E}S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{E}(X_i-\overline{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{E}X_i^2 - 2\mathbf{E}X_i\overline{X} + \mathbf{E}\overline{X}^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\sigma^2 - (2/n)(\sigma^2+(n-1)\mu^2) + (1/n)^2(n\sigma^2+n(n-1)\mu^2) \\
&= \frac{n}{n-1}\Big(\sigma^2 - (2/n)(\sigma^2+(n-1)\mu^2) + (1/n)^2(n\sigma^2+n(n-1)\mu^2)\Big) \\
&= \frac{n-2+1}{n-1}\sigma^2 - \mu^2 + \mu^2 = \sigma^2.
\end{aligned}$$

□

Exercise 3.8. Let $X\colon \Omega \to \mathbb{R}$ be a random variable with $\mathbf{E}X^2 < \infty$. Show that the quantity $\mathbf{E}(X-t)^2$ is minimized for $t \in \mathbb{R}$ uniquely when $t = \mathbf{E}X$.

Solution. We write

$$\begin{aligned}
\mathbf{E}(X-t)^2 &= \mathbf{E}(X-\mathbf{E}X+\mathbf{E}X-t)^2 = \mathbf{E}(X-\mathbf{E}X)^2 + (\mathbf{E}X-t)^2 + 2\mathbf{E}(X-\mathbf{E}X)(\mathbf{E}X-t) \\
&= \mathbf{E}(X-\mathbf{E}X)^2 + (\mathbf{E}X-t)^2.
\end{aligned}$$

The right quantity is uniquely minimized when $t = \mathbf{E}X$.

Alternatively, note that $(d/dt)\mathbf{E}(X-t)^2 = 2t - 2\mathbf{E}X$ and $(d/dt)^2\mathbf{E}(X-t)^2 = 2$. So, the function $t \mapsto \mathbf{E}(X-t)^2$ is strictly concave with a unique critical point at $t = \mathbf{E}X$, and $\lim_{t\to\pm\infty}\mathbf{E}(X-t)^2 = \infty$, so the critical point $t = \mathbf{E}X$ is the unique global minimum of the function. □

Exercise 3.9. Let X be a chi squared random variables with p degrees of freedom. Let Y be a chi squared random variable with q degrees of freedom. Assume that X and Y are independent. Show that $(X/p)/(Y/q)$ has the following density, known as **Snedecor's f-distribution** with p and q degrees of freedom

$$f_{(X/p)/(Y/q)}(t) := \frac{t^{(p/2)-1}(p/q)^{p/2}\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)}\Big(1+t(p/q)\Big)^{-(p+q)/2}, \qquad \forall\, t > 0.$$

Proof. First note that $f_{(X/p)}(t) = pf_X(tp)$ for all $t \in \mathbb{R}$. Then, starting as in the previous proof we have

$$\begin{aligned} f_{(X/p)/(Y/q)}(t) &= \int_0^\infty a f_{X/p}(at) f_{Y/q}(a) da = \int_0^\infty apf_X(atp) q f_Y(qa) da \\ &= \frac{t^{(p/2)-1} p^{p/2} q^{q/2}}{2^{(p+q)/2}\Gamma(p/2)\Gamma(q/2)} \int_0^\infty a^{p/2} e^{-atp/2} a^{(q/2)-1} e^{-aq/2} da \\ &= \frac{t^{(p/2)-1} p^{p/2} q^{q/2}}{2^{(p+q)/2}\Gamma(p/2)\Gamma(q/2)} \int_0^\infty a^{[(p+q)/2]-1} e^{-a(q+tp)/2} da \end{aligned}$$

The integrand is the density of a gamma distributed random variable with parameters α, β where $\alpha = (p+q)/2$ and $\beta = 2/(q+tp)$; so that if we divide and multiply by $\beta^\alpha \Gamma(\alpha)$, we have

$$\begin{aligned} f_{X/Y}(t) &= \frac{t^{(p/2)-1} p^{p/2} q^{q/2}}{2^{(p+q)/2}\Gamma(p/2)\Gamma(q/2)} \beta^\alpha \Gamma(\alpha) \cdot (1) \\ &= \frac{t^{(p/2)-1} p^{p/2} q^{q/2} \Gamma((p+q)/2)}{2^{(p+q)/2}\Gamma(p/2)\Gamma(q/2)} \Big(\frac{2}{q+tp}\Big)^{(p+q)/2} \\ &= \frac{t^{(p/2)-1} p^{p/2} q^{q/2} \Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)} \Big(q+tp\Big)^{-(p+q)/2} \\ &= \frac{t^{(p/2)-1} (p/q)^{p/2} \Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)} \Big(1+t(p/q)\Big)^{-(p+q)/2} \end{aligned}$$

□

4. Homework 4

Exercise 4.1 (**Order Statistics**)**.** Let $X\colon \Omega \to \mathbb{R}$ be a random variable. Let $X_1, \ldots, X_n$ be a random sample of size n from X. Define $X_{(1)} := \min_{1 \le i \le n} X_i$, and for any $2 \le i \le n$, inductively define

$$X_i := \min\Big\{\{X_1, \ldots, X_n\} \smallsetminus \{X_{(1)}, \ldots, X_{(i-1)}\}\Big\},$$

so that

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)} = \max_{1 \le i \le n} X_i.$$

The random variables $X_{(1)}, \ldots, X_{(n)}$ are called the **order statistics** of $X_1, \ldots, X_n$.

- Suppose X is a discrete random variable and we can order the values that X takes as $x_1 < x_2 < \cdots$. For any $i \ge 1$, define $p_i := \mathbf{P}(X \le x_i)$. Show that, for any $1 \le i, j \le n$,

  $$\mathbf{P}(X_{(j)} \le x_i) = \sum_{k=j}^{n} \binom{n}{k} p_i^k (1-p_i)^{n-k}.$$

  (Hint: Let Y be the number of indices $1 \le j \le n$ such that $X_j \le x_i$. Then Y is a binomial random variable with parameters n and p_i.)

You don't have to show it, but if X is a continuous random variable with density f_X and cumulative distribution function F_X, then for any $1 \leq j \leq n$, $F_{X_{(j)}}$ has density

$$f_{X_{(j)}}(x) := \frac{n!}{(j-1)!(n-j)!} f_X(x)(F_X(x))^{j-1}(1-F_X(x))^{n-j}, \qquad \forall\, x \in \mathbb{R}.$$

(This follows by differentiating the above identity for the cumulative distribution function.)

- Let X be a random variable uniformly distributed in $[0,1]$. For any $1 \leq j \leq n$, show that $X_{(j)}$ is a beta distributed random variable with parameters j and $n-j+1$. Conclude that (as you might anticipate)

$$\mathbf{E}X_{(j)} = \frac{j}{n+1}.$$

- Let $a, b \in \mathbb{R}$ with $a < b$. Let U be the number of indices $1 \leq j \leq n$ such that $X_j \leq a$. Let V be the number of indices $1 \leq j \leq n$ such that $a < X_j \leq b$. Show that the vector $(U, V, n-U-V)$ is a multinomial random variable, so that for any nonnegative integers u, v with $u+v \leq n$, we have

$$\begin{aligned}&\mathbf{P}(U=u, V=v, n-U-V=n-u-v)\\ &\quad = \frac{n!}{u!v!(n-u-v)!} F_X(a)^u (F_X(b)-F_X(a))^v (1-F_X(b))^{n-u-v}.\end{aligned}$$

Consequently, for any $1 \leq i, j \leq n$,

$$\mathbf{P}(X_{(i)} \leq a, X_{(j)} \leq b) = \mathbf{P}(U \geq i, U+V \geq j) = \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \mathbf{P}(U=k, V=m) + \mathbf{P}(U \geq j).$$

So, it is possible to write an explicit formula for the joint distribution of $X_{(i)}$ and $X_{(j)}$ (but you don't have to write it yourself).

Solution. We prove the last assertion only. For any $1 \leq j \leq n$, the random vector $(1_{X_j \leq a}, 1_{a < X_j \leq b}, 1_{X_j > b})$ is equal to $(1,0,0)$ with probability $F_X(a)$, it is equal to $(0,1,0)$ with probability $F_X(b) - F_X(a)$, and it is equal to $(0,0,1)$ with probability $1 - F_X(b)$. Also, the set of random vectors $\{(1_{X_j \leq a}, 1_{a < X_j \leq b}, 1_{X_j > b})\}_{j=1}^n$ are all independent, since $X_1, \ldots, X_n$ are independent. It follows from the definition of a multinomial random variable that the random vector

$$\Big(sum_{j=1}^n 1_{X_j \leq a}, \quad \sum_{j=1}^n 1_{a < X_j \leq b}, \quad \sum_{j=1}^n 1_{b < X_j} \Big) = (U, V, n-U-V)$$

is a multinomial random variable with the stated parameters. □

Exercise 4.2. Using Matlab (or any other mathematical system on a computer), verify that its random number generator agrees with the law of large numbers and central limit theorem. For example, average 10^7 samples from the uniform distribution on $[0,1]$ and check how close the sample average is to $1/2$. Then, sum up n samples from the uniform distribution on $[0,1]$, construct this sum n times, make a histogram of the different values of the sum, and check how close the histogram is to a Gaussian (when $n = 10^4$). If you want a challenge, try $n = 10^5$ or $n = 10^6$.

Exercise 4.3. Let $X\colon \Omega \to \mathbb{R}$ be a random variable on a sample space Ω equipped with a probability law $\mathbf{P}$. For any $t \in \mathbb{R}$ let $F(t) := \mathbf{P}(X \le t)$. For any $s \in (0,1)$ define

$$Y(s) := \sup\{t \in \mathbb{R}\colon F(t) < s\}.$$

Then Y is a random variable on $(0,1)$ with the uniform probability law on $(0,1)$. Show that X and Y are equal in distribution. That is, $\mathbf{P}(Y \le t) = F(t)$ for all $t \in \mathbb{R}$.

Solution. For any $t \in \mathbb{R}$, let $A_t := \{s \in (0,1)\colon \sup\{v \in \mathbb{R}\colon F(v) < s\} \le t\}$. And for any $s \in (0,1)$, let $c_s := \sup\{v \in \mathbb{R}\colon F(v) < s\}$. If $a, b \in (0,1)$ satisfy $a < b$, then $c_b \ge c_a$ since F is monotone. So, if $b \in A_t$ and $a < b$, then $a \in A_t$. Therefore, A_t is an interval. Also, since F is monotone and right-continuous, $F^{-1}(F(t))$ is either a single point or an interval that includes its left endpoint. Denote $x(t)$ as the smallest element of $F^{-1}(F(t))$. Then $F(x(t)) = F(t)$ and by definition of $x(t)$, $c_{F(t)} = x(t)$. And if $a < x(t)$, then $c_{F(t)} > c_{F(a)}$, so that $F(a) \in A_t$. So, A_t contains the interval $(0, \lim_{a \to x(t)^-} F(a))$. It could occur that this set is strictly inside $(0, F(t))$. Since F is monotone, this only occurs when $\lim_{a \to x(t)^-} F(a) < \lim_{a \to x(t)^+} F(a) = F(t)$. If y is between these two values, then $y \in A_t$, since $c_y < c_{F(x(t))}$. So, A_t contains $(0, F(t))$. And for any $\varepsilon > 0$, $F(t) + \varepsilon \notin A_t$. So, either $A_t = (0, F(t))$ or $A_t = (0, F(t)]$. In either case, $\mathbf{P}(A_t) = F(t) - 0 = F(t)$. □

Exercise 4.4 (Box-Muller Algorithm). Let U_1, U_2 be independent random variables uniformly distributed in $(0,1)$. Define

$$R := \sqrt{-2\log U_1}, \qquad \Psi := 2\pi U_2.$$
$$X := R\cos\Psi, \qquad Y := R\sin\Psi.$$

Show that X, Y are independent standard Gaussian random variables. So, we can simulate any number of independent standard Gaussian random variables with this procedure.

Now, let $\{a_{ij}\}_{1 \le i,j \le n}$ be an $n \times n$ symmetric positive semidefinite matrix. That is, for any $v \in \mathbb{R}^n$, we have

$$v^T a v = \sum_{i,j=1}^{n} v_i v_j a_{ij} \ge 0.$$

We can simulate a Gaussian random vector with any such covariance matrix $\{a_{ij}\}_{1 \le i,j \le n}$ using the following procedure.

- Let $X = (X_1, \dots, X_n)$ be a vector of i.i.d. standard Gaussian random variables (which can be sampled using the Box-Muller algorithm above).
- Write the matrix a in its Cholesky decomposition $a = rr^*$, where r is an $n \times n$ real matrix. (This decomposition can be computed efficiently with about n^3 arithmetic operations.)
- Let $e^{(1)}, \dots, e^{(n)}$ be the rows of r. For any $1 \le i \le n$, define

$$Z_i := \langle X, e^{(i)} \rangle.$$

Show that $Z := (Z_1, \dots, Z_n)$ is a mean zero Gaussian random vector whose covariance matrix is $\{a_{ij}\}_{1 \le i,j \le n}$, so that

$$\mathbf{E}(Z_i Z_j) = a_{ij}, \qquad \forall\, 1 \le i, j \le n.$$

Solution. We use Exercise 1.11. We have $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ such that $T(U_1, U_2) = (X, Y)$. By Exercise 1.11,

$$f_{X,Y}(x,y) = f_{U_1,U_2}(S(x,y))\,|J(x,y)|\,, \qquad \forall\, x, y \in \mathbb{R},$$

where $S = T^{-1}$ and $J(x,y)$ is the determinant of the Jacobian of S at (x,y). Note that $TS(x,y) = (x,y)$ for all $x, y \in \mathbb{R}$, so solving this equation for S yields

$$S(x,y) = (e^{-\frac{x^2+y^2}{2}}, \frac{1}{2\pi}\tan^{-1}(y/x)), \qquad \forall\, x, y > 0.$$

So, when $x, y > 0$, we have

$$\begin{aligned} 2\pi\,|J(x,y)| &= \left|\det\begin{pmatrix} -xe^{-\frac{x^2+y^2}{2}} & -ye^{-\frac{x^2+y^2}{2}} \\ -\frac{yx^{-2}}{1+(y/x)^2} & \frac{1/x}{1+(y/x)^2} \end{pmatrix}\right| \\ &= \left|\det\begin{pmatrix} -xe^{-\frac{x^2+y^2}{2}} & -ye^{-\frac{x^2+y^2}{2}} \\ -\frac{y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{pmatrix}\right| = e^{-(x^2+y^2)/2}. \end{aligned}$$

A similar calculation holds for all other nonzero x, y. So,

$$f_{X,Y}(x,y) = \frac{1}{2\pi}e^{-(x^2+y^2)/2} f_{U_1,U_2}(S(x,y)) = \frac{1}{2\pi}e^{-(x^2+y^2)/2}, \qquad \forall\, x, y \in \mathbb{R},$$

So, X, Y are independent standard Gaussians since

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-y^2/2} = f_X(x)f_Y(y), \qquad \forall\, x, y \in \mathbb{R},$$

Now

$$\mathbf{E}(Z_i Z_j) = \mathbf{E}\langle X, e^{(i)}\rangle\langle X, e^{(j)}\rangle = \mathbf{E}\sum_{k,\ell=1}^{n} X_k e_k^{(i)} X_\ell e_\ell^{(j)} = \sum_{k=1}^{n} e_k^{(i)} e_k^{(j)} = a_{ij}.$$

□

Exercise 4.5. In the notes we showed that the Delta Method works only assuming that $f'(\theta)$ exists. In fact, the method works even when $f'(\theta)$ does not exist. In this exercise, we assume that

$$f'(\theta^+) := \lim_{y\to\theta^+}\frac{f(y)-f(\theta)}{y-\theta}, \qquad f'(\theta^-) := \lim_{y\to\theta^-}\frac{f(y)-f(\theta)}{y-\theta},$$

exist. For example, consider

$$f(y) := \max(y, 0), \qquad \forall\, y \in \mathbb{R}.$$

Then $f'(0^+) = 1$ while $f'(0^-) = 0$, so $f'(0)$ does not exist.

For simplicity, we assume that $\theta = 0$ and $f(\theta) = 0$.

Let $Y_1, Y_2, \ldots$ be random variables such that $\sqrt{n}(Y_n - \theta)$ converges in distribution to a mean zero Gaussian random variable with variance $\sigma^2 > 0$.

- Argue as in the notes, and show that for all $y \in \mathbb{R}$, there exists a function h with $\lim_{z\to 0} h(z)/z = 0$, and

$$f(y) = f'(0^+)y1_{y>0} + f'(0^-)y1_{y<0} + h(y).$$

- Conclude that

$$\sqrt{n}f(Y_n) = \sqrt{n}\Big(f'(0^+)Y_n 1_{Y_n>0} + f'(0^-)Y_n 1_{Y_n<0} + h(Y_n)\Big).$$

- Deduce that, as $n \to \infty$, $\sqrt{n} f(Y_n)$ converges in distribution to

$$\Big(\sigma f'(\theta^+) 1_{Z>0} + \sigma f'(\theta^-) 1_{Z<0}\Big) Z.$$

(Note that $f'(0^+) Y_n 1_{Y_n>0}$ and $f'(0^-) Y_n 1_{Y_n<0}$ have disjoint supports; this could be useful to prove convergence in distribution as $n \to \infty$.)

Solution. Since $f'(0^+)$ exists, $\lim_{y \to 0^+} \frac{f(y)-f(0)}{y}$ exists. That is, there exists $h_+ \colon \mathbb{R} \to \mathbb{R}$ such that $\lim_{z \to 0^+} \frac{h_+(z)}{z} = 0$ and, for all $y > 0$,

$$f(y) = f(0) + f'(0^+)(y-\theta) + h_+(y-\theta) = f'(0^+)(y) + h_+(y).$$

Similarly, since $f'(0^-)$ exists, $\lim_{y \to 0^-} \frac{f(y)-f(0)}{y}$ exists. That is, there exists $h_- \colon \mathbb{R} \to \mathbb{R}$ such that $\lim_{z \to 0^-} \frac{h_-(z)}{z} = 0$ and, for all $y < 0$,

$$f(y) = f'(0^-)(y) + h_-(y).$$

So, adding these two equalities, we have, for all $y \neq 0$,

$$f(y) = f'(0^+) y 1_{y>0} + f'(0^-) y 1_{y<0} + 1_{y>0} h_+(y) + 1_{y<0} h_-(y).$$

Define $h(y) := 1_{y>0} h_+(y) + 1_{y<0} h_-(y)$. Then the first property holds: for all $y \in \mathbb{R}$,

$$f(y) = f'(0^+) y 1_{y>0} + f'(0^-) y 1_{y<0} + h(y).$$

(Note that both sides are zero when $y = 0$.) Note also that $\lim_{z \to 0} \frac{h(z)}{z} = 0$, since this property holds for h_+, h_- separately.

Plugging in $y = Y_n$, we then obtain

$$\sqrt{n} f(Y_n) = \sqrt{n} \Big(f'(0^+) Y_n 1_{Y_n>0} + f'(0^-) Y_n 1_{Y_n<0} + h(Y_n) \Big). \qquad (*)$$

By assumption, $\forall\ s, t > 0$, $\lim_{n \to \infty} \mathbf{P}(|Y_n - \theta| > st/\sqrt{n}) = 2 \int_{st}^{\infty} e^{-y^2/[2\sigma^2]} \frac{dy}{\sigma\sqrt{2\pi}}$. So, $\forall\ n \geq 1$,

$$\begin{aligned} \mathbf{P}(\sqrt{n}\,|h(Y_n)| > t) &= \mathbf{P}(\sqrt{n}\,|h(Y_n)| > t,\ |Y_n| > st/\sqrt{n}) \\ &\qquad + \mathbf{P}(\sqrt{n}\,|h(Y_n)| > t,\ |Y_n| \leq st/\sqrt{n}) \\ &\leq \mathbf{P}(|Y_n| > st/\sqrt{n}) + \mathbf{P}(\sqrt{n}\,|h(Y_n)| > t,\ |Y_n| \leq st/\sqrt{n}). \end{aligned}$$

As $n \to \infty$, the first term converges to $2 \int_{st}^{\infty} e^{-\frac{y^2}{2\sigma^2}} \frac{dy}{\sigma\sqrt{2\pi}}$, and the second term goes to zero since $\lim_{z \to 0}(h(z)/z) = 0$. So, for any $s, t > 0$, $\lim_{n \to \infty} \mathbf{P}(\sqrt{n}\,|h(Y_n)| > t) \leq 2 \int_{st}^{\infty} e^{-\frac{y^2}{2\sigma^2}} \frac{dy}{\sigma\sqrt{2\pi}}$. Since this holds for any $s > 0$, we can let $s \to \infty$ to get $\lim_{n \to \infty} \mathbf{P}(\sqrt{n}\,|h(Y_n)| > t) = 0$. That is, $\sqrt{n} h(Y_n)$ converges in probability to zero as $n \to \infty$.

So, by Slutksy's Theorem and $(*)$, $\sqrt{n}[f(Y_n)]$ converges in distribution to

$$\Big(\sigma f'(\theta^+) 1_{Z>0} + \sigma f'(\theta^-) 1_{Z<0}\Big) Z,$$

where Z is a standard Gaussian random variable. (Since $\sqrt{n} Y_n$ converges in distribution to a mean zero Gaussian with variance σ^2 as $n \to \infty$, if $t > 0$, we have

$$\lim_{n \to \infty} \mathbf{P}\Big(\sqrt{n}\Big(f'(0^+) Y_n 1_{Y_n>0} + f'(0^-) Y_n 1_{Y_n<0}\Big) \geq t\Big) = \mathbf{P}\Big(\sqrt{n} f'(0^+) Y_n 1_{Y_n>0} \geq t\Big) = \mathbf{P}(\sigma Z > t).$$

Similarly, if $t < 0$

$$\lim_{n\to\infty} \mathbf{P}\Big(\sqrt{n}\Big(f'(0^+)Y_n 1_{Y_n>0} + f'(0^-)Y_n 1_{Y_n<0}\Big) < t\Big) = \mathbf{P}\Big(\sqrt{n} f'(0^-)Y_n 1_{Y_n<0} < t\Big) = \mathbf{P}(\sigma Z < t).$$

It follows that

$$\sqrt{n}\Big(f'(0^+)Y_n 1_{Y_n>0} + f'(0^-)Y_n 1_{Y_n<0}\Big)$$

converges in distribution to

$$\Big(\sigma f'(\theta^+)1_{Z>0} + \sigma f'(\theta^-)1_{Z<0}\Big)Z.$$

) □

Exercise 4.6. Let A, B, Ω be sets. Let $u\colon \Omega \to A$ and let $t\colon \Omega \to B$. Assume that, for every $x, y \in \Omega$, if $u(x) = u(y)$, then $t(x) = t(y)$. Show that there exists a function $s\colon A \to B$ such that

$$t = s(u).$$

Solution. Let a in the range of u. That is, there exists $\omega \in \Omega$ such that $u(\omega) = a$. For any a in the range of u, define then

$$s(a) := t(\omega).$$

Then $s(a) = t(\omega)$ and $s(a) = s(u(\omega))$, so $t(\omega) = s(u(\omega))$, as desired. It remains to show that $s(a)$ is well-defined. To see this, let $\omega' \in \Omega$ such that $u(\omega') = u(\omega)$. We need to show that $s(a)$ is well-defined, i.e. that $t(\omega) = t(\omega')$. This follows immediately from our assumption. So, $s(a)$ is well-defined.

Finally, for any $a \in A$ that is not in the range of u, define $s(a)$ to be an arbitrary element of B. Then $t = s(u)$ still holds. □

Exercise 4.7. Let $\{f_\theta\colon \theta \in \Theta\}$ be a k-parameter exponential family $\{f_\theta\colon \theta \in \Theta,\ a(w(\theta)) < \infty\}$ of probability density functions or probability mass functions, where

$$f_\theta(x) := h(x) \exp\Big(\sum_{i=1}^{k} w_i(\theta)t_i(x) - a(w(\theta))\Big), \qquad \forall\, x \in \mathbb{R}.$$

For any $\theta \in \Theta$, let $w(\theta) := (w_1(\theta), \ldots, w_k(\theta))$. Assume that the following subset of $\mathbb{R}^k$ is k-dimensional:

$$\{w(\theta) - w(\theta') \in \mathbb{R}^k\colon \theta, \theta' \in \Theta\}.$$

That is, if $x \in \mathbb{R}^k$ satisfies $\langle x, y\rangle = 0$ for all y in this set, then $x = 0$. (Note that the assumption of the exercise is always satisfied for an exponential family in canonical form.)

Let $X = (X_1, \ldots, X_n)$ be a random sample of size n from f_θ. Define $t\colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$t(X) := \sum_{j=1}^{n} \left(t_1(X_j), \ldots, t_k(X_j)\right).$$

Show that $t(X)$ is minimal sufficient for θ. (Hint: if you get stuck, look at Example 3.12 in Keener.)

Conclude that if we sample from a Gaussian with unknown mean μ and variance $\sigma^2 > 0$, then $\overline{X}$ is minimal sufficient for μ and $(\overline{X}, S)$ is minimal sufficient for (μ, σ^2).

Warning: the f_θ exponential family mentioned here is a function of one variable. If you use the Theorem from class about checking the ratio of $f_\theta(x)/f_\theta(y)$, the functions there are *joint* density functions (i.e. the product of n copies of the same function).

Optional: If the f_θ functions are always positive, you should be able to change the assumption to the following. For any $\theta \in \Theta$, let $w(\theta) := (w_1(\theta), \dots, w_k(\theta))$. Assume that the following subset of $\mathbb{R}^k$ is k-dimensional:

$$\{w(\theta) \in \mathbb{R}^k \colon \theta, \theta' \in \Theta\}.$$

Solution. The first proof appears in Keener. We instead prove the Optional part.

We use Theorem 5.8 from the notes. Fix $x, y \in \mathbb{R}^n$. Define $w(\theta) := (w_1(\theta), \dots, w_k(\theta))$. We examine the ratio

$$\frac{\prod_{j=1}^n f_\theta(x_j)}{\prod_{j=1}^n f_\theta(y_j)} = \frac{\prod_{j=1}^n f_\theta(x_j)}{\prod_{j=1}^n f_\theta(y_j)} = \exp\Big(\sum_{i=1}^k w_i(\theta) \sum_{j=1}^n [t_i(x_j) - t_i(y_j)]\Big) = \exp\Big(\langle w(\theta), t(x) - t(y)\rangle\Big).$$

If this ratio is constant for all $\theta \in \Theta$, then it follows by our assumption that $t(x) - t(y) = 0$, i.e. $t(x) = t(y)$. Conversely, if $t(x) = t(y)$, then the ratio is constant for all $\theta \in \Theta$. By Theorem 5.8, we conclude that $t(X)$ is a minimal sufficient statistic.

In the Gaussian case, we conclude that $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a minimal sufficient statistic. Since $(\overline{X}, S)$ is a function of $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$, minimal sufficiency of $(\overline{X}, S)$ follows from sufficiency of $(\overline{X}, S)$. Technically we did not show sufficiency of $(\overline{X}, S)$ for (μ, σ^2). Let us show it now. Let $x \in \mathbb{R}^n$. Then

$$\begin{aligned}
f_{\mu,\sigma^2}(x) &= \prod_{i=1}^n \sigma^{-1}(2\pi)^{-1/2} e^{-(x_i-\mu)^2/(2\sigma^2)} \\
&= \sigma^{-n}(2\pi)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 - \frac{n}{2\sigma^2}\mu^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\Big) \\
&= \sigma^{-n}(2\pi)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mathbf{E}X_i + \mathbf{E}X_i)^2 - \frac{n}{2\sigma^2}\mu^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\Big) \\
&= \sigma^{-n}(2\pi)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n \Big[(x_i - \mathbf{E}X_i)^2 + (\mathbf{E}X_i)^2 + 2(x_i - \mathbf{E}X_i)\mathbf{E}X_i\Big] - \frac{n}{2\sigma^2}\mu^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\Big) \\
&= \sigma^{-n}(2\pi)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mathbf{E}X_i)^2 + n\mu^2 + 2\mu\sum_{i=1}^n x_i - 2n\mu^2 - \frac{n}{2\sigma^2}\mu^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\Big)
\end{aligned}$$

We have exhibited $f_{\mu,\sigma^2}(x)$ as a function of $t(x) := (\sum_{i=1}^n x_i, \sum_{i=1}^n (x_i - \mathbf{E}X_i)^2)$, as in the Factorization Theorem. We conclude that $(\overline{X}, S)$ is sufficient for (μ, σ^2). □

Exercise 4.8. Let $\mathbf{P}_1, \mathbf{P}_2$ be two probability laws on the sample space $\Omega = \mathbb{R}$. Suppose these laws have densities $f_1, f_2 \colon \mathbb{R} \to [0, \infty)$ so that

$$\mathbf{P}_i(A) = \int_A f_i(x)dx, \qquad \forall\, i = 1, 2, \qquad \forall\, A \subseteq \mathbb{R}.$$

Show that

$$\sup_{A\subseteq\mathbb{R}} |\mathbf{P}_1(A) - \mathbf{P}_2(A)| = \frac{1}{2}\int_{\mathbb{R}} |f_1(x) - f_2(x)|\, dx.$$

(Hint: consider $A := \{x \in \mathbb{R}\colon f_1(x) > f_2(x)\}$.)

Similarly, if $\mathbf{P}_1, \mathbf{P}_2$ are probability laws on $\Omega = \mathbb{Z}$, show that

$$\sup_{A\subseteq\mathbb{Z}} |\mathbf{P}_1(A) - \mathbf{P}_2(A)| = \frac{1}{2}\sum_{z\in\mathbb{Z}} |\mathbf{P}_1(z) - \mathbf{P}_2(z)|\,.$$

Solution. As suggested in the hint, consider $A := \{x \in \mathbb{R}\colon f_1(x) > f_2(x)\}$. Then

$$\begin{aligned}|\mathbf{P}_1(A) - \mathbf{P}_2(A)| &= \left|\int_A f_1(x)dx - \int_A f_2(x)dx\right| = \int_A (f_1(x) - f_2(x))dx \\ &= \int_A |f_1(x) - f_2(x)|\, dx. \qquad (*)\end{aligned}$$

By definition of A, we have

$$\int_{A^c} |f_1(x) - f_2(x)|\, dx = \int_{A^c} (f_2(x) - f_1(x))dx.$$

So,

$$\begin{aligned}&\int_A |f_1(x) - f_2(x)|\, dx - \int_{A^c} |f_1(x) - f_2(x)|\, dx \\ &\quad= \int_A (f_1(x) - f_2(x))dx - \int_{A^c} (f_2(x) - f_1(x))dx \\ &\quad= \int_{\mathbb{R}} (f_1(x) - f_2(x))dx = 0.\end{aligned}$$

The last equality follows since f_1, f_2 are PDFs. In summary,

$$\int_A |f_1(x) - f_2(x)|\, dx = \int_{A^c} |f_1(x) - f_2(x)|\, dx \qquad (**)$$

So,

$$\int_{\mathbb{R}} |f_1(x) - f_2(x)|\, dx = 2\int_A |f_1(x) - f_2(x)|\, dx \qquad (***).$$

We can then rewrite $(*)$ as

$$|\mathbf{P}_1(A) - \mathbf{P}_2(A)| = \frac{1}{2}\int_{\mathbb{R}} |f_1(x) - f_2(x)|\, dx.$$

That is, we have shown that

$$\sup_{A\subseteq\mathbb{R}} |\mathbf{P}_1(A) - \mathbf{P}_2(A)| \geq \frac{1}{2}\int_{\mathbb{R}} |f_1(x) - f_2(x)|\, dx.$$

It remains to show the reverse inequality. This follows by repeating the above reasoning. For any $A \subseteq \mathbb{R}$ we have

$$\begin{aligned}
&|\mathbf{P}_1(A) - \mathbf{P}_2(A)| \\
&= \left| \int_A f_1(x)dx - \int_A f_2(x)dx \right| \\
&= \left| \int_{\{x \in A:\ f_1(x) > f_2(x)\}} (f_1(x) - f_2(x))dx + \int_{\{x \in A:\ f_1(x) < f_2(x)\}} (f_1(x) - f_2(x))dx \right| \\
&= \left| \int_{\{x \in A:\ f_1(x) > f_2(x)\}} (f_1(x) - f_2(x))dx - \int_{\{x \in A:\ f_1(x) < f_2(x)\}} (f_2(x) - f_1(x))dx \right| \\
&= \left| \int_{\{x \in A:\ f_1(x) > f_2(x)\}} |f_1(x) - f_2(x)|\, dx - \int_{\{x \in A:\ f_1(x) < f_2(x)\}} |f_2(x) - f_1(x)|\, dx \right| \\
&\leq \max\Big(\int_{\{x \in A:\ f_1(x) > f_2(x)\}} |f_1(x) - f_2(x)|\, dx, \quad \int_{\{x \in A:\ f_1(x) < f_2(x)\}} |f_2(x) - f_1(x)|\, dx \Big) \\
&\leq \max\Big(\int_{\{x \in \mathbb{R}:\ f_1(x) > f_2(x)\}} |f_1(x) - f_2(x)|\, dx, \quad \int_{\{x \in \mathbb{R}:\ f_1(x) < f_2(x)\}} |f_2(x) - f_1(x)|\, dx \Big).
\end{aligned}$$

The first inequality used $|a - b| \leq \max(a, b)$, valid for all $a, b > 0$. Both terms in the maximum are equal to each other by $(**)$, so the proof is completed by $(***)$ since we showed that, for any $A \subseteq \mathbb{R}$, we have

$$|\mathbf{P}_1(A) - \mathbf{P}_2(A)| \leq \frac{1}{2} \int_{\mathbb{R}} |f_1(x) - f_2(x)|\, dx.$$

□

Exercise 4.9. Give an example of a statistic Y that is complete and nonconstant, but such that Y is not sufficient.

Solution. Let $X_1, \ldots, X_n$ be Bernoulli random variables with unknown parameter $0 < \theta < 1$. Define $t(x_1, \ldots, x_n) := x_1$. We claim that $Y := t(X_1, \ldots, X_n)$ is complete but not sufficient. Completeness follows since if $f \colon \mathbb{R} \to \mathbb{R}$ satisfies $\mathbf{E}_\theta f(Y) = 0$ for all $0 < \theta < 1$, then $\theta f(1) + (1 - \theta) f(0)$ for all $0 < \theta < 1$, so that $f(0) = f(1) = 0$, so that $f(Y) = 0$, i.e. Y is complete. However, Y is not sufficient since $X_1, \ldots, X_n$ conditioned on X_1 does depend on θ. For example,

$$\mathbf{P}(X_1 = 1, X_2 = 1 | X_1 = 1) = \mathbf{P}(X_2 = 1) = \theta.$$

□

Exercise 4.10. This exercise shows that a complete sufficient statistic might not exist.

Let $X_1, \ldots, X_n$ be a random sample of size n from the uniform distribution on the three points $\{\theta, \theta + 1, \theta + 2\}$, where $\theta \in \mathbb{Z}$.

- Show that the vector $Y := (X_{(1)}, X_{(n)})$ is minimal sufficient for θ.
- Show that Y is not complete by considering $X_{(n)} - X_{(1)}$.
- Using minimal sufficiency, conclude that any sufficient statistic for θ is not complete.

Solution. We apply Theorem 5.8 from the notes. Suppose $x, y \in \mathbb{Z}^n$ satisfy $f_\theta(x) = c(x, y) f_\theta(y)$ for some $c(x, y)$ and for all $\theta \in \mathbb{Z}$, and there exists $\theta_1, \theta_2 \in \mathbb{Z}$ such that

$f_{\theta_1}(x) > 0$ and $f_{\theta_2}(y) > 0$. Then $c(x,y) > 0$, $x_1, \dots, x_n \in \{\theta_1, \theta_1+1, \theta_1+2\}$ and $y_1, \dots, y_n \in \{\theta_2, \theta_2+1, \theta_2+2\}$. Since $f_\theta(x) = c(x,y) f_\theta(y)$ for all $\theta \in \Theta$ (including θ_1), we may assume that $y_1, \dots, y_n \in \{\theta_1, \theta_1+1, \theta_1+2\}$. Note that the number of $\theta \in \mathbb{Z}$ such that $f_\theta(x) > 0$ is equal to 4 minus the maximum of x plus the minimum of x, and similarly for y. So, $f_\theta(x) = c(x,y) f_\theta(y)$ for all $\theta \in \mathbb{Z}$ implies that

$$\max_{1\le i\le n} x_i - \min_{1\le i\le n} x_i = \max_{1\le i\le n} y_i - \min_{1\le i\le n} y_i, \qquad (*)$$

If $\max_{1\le i\le n} x_i > \max_{1\le i\le n} y_i$, then $f_\theta(x) > 0$ while $f_\theta(y) = 0$ for $\theta = \max_{1\le i\le n} x_i$. We therefore conclude that $\max_{1\le i\le n} x_i \le \max_{1\le i\le n} y_i$. Interchanging the roles of x, y we conclude also that $\max_{1\le i\le n} x_i \ge \max_{1\le i\le n} y_i$. That is,

$$\max_{1\le i\le n} x_i = \max_{1\le i\le n} y_i, \qquad (**)$$

Then $(*)$ implies that

$$\min_{1\le i\le n} x_i = \min_{1\le i\le n} y_i, \qquad (***)$$

This equality and $(**)$ imply that $t(x) = t(y)$. This argument can be reversed. If $t(x) = t(y)$, then $(**)$ and $(***)$ hold, so $(*)$ holds, so $f_\theta(x) = c(x,y) f_\theta(y)$ for some $c(x,y)$ and for all $\theta \in \mathbb{Z}$. So, by Theorem 5.8, Y is minimal sufficient.

Now, Y is not complete since if $f(y_1, y_2) = y_2 - y_1$ for all $y_1, y_2 \in \mathbb{R}$, then $f(Y) \neq 0$ but $f(Y)$ does not depend on θ, so there exists $c \in \mathbb{R}$ such that $\mathbf{E}_\theta[f(Y) - c] = 0$ for all $\theta \in \mathbb{Z}$, so Y is not complete.

Let Z be any sufficient statistic for θ. Since Y is minimal sufficient, there exists a function r such that $Y = r(Z)$. Therefore, with f as just defined, we have $\mathbf{E}_\theta[f(r(Z)) - c] = 0$ for all $\theta \in \mathbb{Z}$, so Z is not complete. □

Exercise 4.11 ((Optional) This exercise requires some measure theory so it is optional.)**.** Let $\{f_\theta \colon \theta \in \Theta\}$ be a k-parameter exponential family $\{f_\theta \colon \theta \in \Theta, a(w(\theta)) < \infty\}$ of **joint** probability density functions or probability mass functions in canonical form, where

$$f_w(x) := h(x) \exp\Big(\sum_{i=1}^k w_i t_i(x) - a(w) \Big), \qquad \forall\, x \in \mathbb{R}^n, \qquad \forall\, w \in \{w \in \mathbb{R}^k \colon a(w) < \infty\}.$$

Assume that the following subset of $\mathbb{R}^k$ contains an open set in $\mathbb{R}^k$:

$$\{w \in \mathbb{R}^k \colon a(w) < \infty\}.$$

Assume also that there is no redundancy in the functions $t_1, \dots, t_k$, i.e. assume: if $\exists$ $\alpha_1, \dots, \alpha_k \in \mathbb{R}$ such that $\sum_{i=1}^k \alpha_i t_i(x) = 0$ for all $x \in \mathbb{R}^n$, then $\alpha_1 = \cdots = \alpha_k = 0$.

Let X be a random sample **of size** 1 from f_θ (so $X = (X_1, \dots, X_n)$, and $X_1, \dots, X_n$ are all real valued). Define $t \colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$t(X) := (t_1(X), \dots, t_k(X)).$$

Show that $t(X)$ is complete for θ.

Hint: if you get stuck, look at Theorem 4.3.1 in Lehmann-Romano. An early step in the proof uses the change of variables formula for the pushforward measure.

Once we know the above statement, we can deduce the following about repeated random samples from a single variable exponential family.

Let $\{f_\theta \colon \theta \in \Theta\}$ be a k-parameter exponential family $\{f_\theta \colon \theta \in \Theta, a(w(\theta)) < \infty\}$ of probability density functions or probability mass functions in canonical form, where

$$f_w(x) := h(x) \exp\Big(\sum_{i=1}^{k} w_i t_i(x) - a(w) \Big), \qquad \forall\, x \in \mathbb{R}, \qquad \forall\, w \in \{w \in \mathbb{R}^k \colon a(w) < \infty\}.$$

Assume that the following subset of $\mathbb{R}^k$ contains an open set in $\mathbb{R}^k$:

$$\{w \in \mathbb{R}^k \colon a(w) < \infty\}.$$

Assume also that there is no redundancy in the functions $t_1, \ldots, t_k$, i.e. assume: if $\exists$ $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ such that $\sum_{i=1}^{k} \alpha_i t_i(x) = 0$ for all $x \in \mathbb{R}$, then $\alpha_1 = \cdots = \alpha_k = 0$.

Let $X_1, \ldots, X_n$ be a random sample **of size** n from f_θ. Define $t \colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$t(X) := \sum_{j=1}^{n} \left(t_1(X_j), \ldots, t_k(X_j) \right).$$

Show that $t(X)$ is complete for θ.

5. Homework 5

Exercise 5.1 (**Conditional Expectation as a Random Variable**)**.** Let $X, Y, Z \colon \Omega \to \mathbb{R}$ be discrete or continuous random variables. Let A be the range of Y. Define $g \colon A \to \mathbb{R}$ by $g(y) := \mathbf{E}(X|Y = y)$, for any $y \in A$. We then define the **conditional expectation** of X given Y, denoted $\mathbf{E}(X|Y)$, to be the random variable $g(Y)$.

(i) Let X, Y be random variables such that (X, Y) is uniformly distributed on the triangle $\{(x, y) \in \mathbb{R}^2 \colon x \geq 0, y \geq 0, x + y \leq 1\}$. Show that

$$\mathbf{E}(X|Y) = \frac{1}{2}(1 - Y).$$

(ii) Prove the following version of the Total Expectation Theorem

$$\mathbf{E}(\mathbf{E}(X|Y)) = \mathbf{E}(X).$$

- If X is a random variable, and if $f(t) := \mathbf{E}(X - t)^2$, $t \in \mathbb{R}$, then the function $f \colon \mathbb{R} \to \mathbb{R}$ is uniquely minimized when $t = \mathbf{E}X$. A similar minimizing property holds for conditional expectation. Let $h \colon \mathbb{R} \to \mathbb{R}$. Show that the quantity $\mathbf{E}(X - h(Y))^2$ is minimized among all functions $h \colon \mathbb{R} \to \mathbb{R}$ when $h(Y) = \mathbf{E}(X|Y)$. (Hint: use the previous item.)

(iii) Show the following:

$$\mathbf{E}(Xh(Y)|Y) = h(Y)\mathbf{E}(X|Y).$$
$$\mathbf{E}([\mathbf{E}(X|h(Y))]\,|Y) = \mathbf{E}(X|h(Y)).$$

(iv) Show the following

$$\mathbf{E}(X|X) = X.$$
$$\mathbf{E}(X + Y|Z) = \mathbf{E}(X|Z) + \mathbf{E}(Y|Z).$$

(v) If Z is independent of X and Y, show that

$$\mathbf{E}(X|Y, Z) = \mathbf{E}(X|Y).$$

(Here $\mathbf{E}(X|Y, Z)$ is notation for $\mathbf{E}(X|(Y, Z))$ where (Y, Z) is interpreted as a random vector, so that X is conditioned on the random vector (Y, Z).)

Solution. **(i)** If $y \in [0,1]$,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_{x=0}^{x=1-y} 2dx = 2(1-y).$$

Otherwise, $f_Y(y) = 0$. So, if $y \in [0,1]$

$$\mathbf{E}(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y}(x,y)dx = \int_{-\infty}^{\infty} x\frac{f_{X,Y}(x,y)}{f_Y(y)}dx = \int_{x=0}^{x=1-y} \frac{1}{1-y}xdx = \frac{1}{2}(1-y).$$

And $\mathbf{E}(X|Y=y)$ is undefined when $y \notin [0,1]$, since $f_Y(y) = 0$ when $y \notin [0,1]$.

Then, by definition of $\mathbf{E}(X|Y)$, we have

$$\mathbf{E}(X|Y) = \frac{1}{2}(1-Y).$$

Below, we only consider discrete random variables, the discrete case being similar.

(ii) Using our definitions, $\mathbf{E}(X|Y)$ takes the value $y \in \mathbb{R}$ with probability $\mathbf{P}(Y=y)$, so that

$$\mathbf{E}(\mathbf{E}(X|Y)) = \sum_{y\in\mathbb{R}} \mathbf{E}(X|Y=y)\mathbf{P}(Y=y) = \sum_{y\in\mathbb{R}} \mathbf{E}X1_{Y=y} = \mathbf{E}X\sum_{y\in\mathbb{R}} 1_{Y=y} = \mathbf{E}X.$$

Using Property (iii) below,

$$\begin{aligned}\mathbf{E}(X-h(Y))^2 &= \mathbf{E}\mathbf{E}[(X-h(Y))^2|Y] = \mathbf{E}\mathbf{E}[X^2 - 2Xh(Y) + (h(Y))^2|Y] \\ &= \mathbf{E}X^2 - 2\mathbf{E}[h(Y)\mathbf{E}(X|Y)] + \mathbf{E}(h(Y))^2.\end{aligned}$$

The last two terms can be written as

$$\sum_{y\in\mathbb{R}}[-2h(y)\mathbf{E}(X|Y=y) + (h(y))^2]\mathbf{P}(Y=y)$$

For fixed $y \in \mathbb{R}$, the quantity $[-2h(y)\mathbf{E}(X|Y=y) + (h(y))^2]$ is minimized when $h(y) = \mathbf{E}(X|Y=y)$. So, the quantity $-2\mathbf{E}[h(Y)\mathbf{E}(X|Y)] + \mathbf{E}(h(Y))^2$ is minimized when $h(y) = \mathbf{E}(X|Y=y)$ for all $y \in \mathbb{R}$. That is, $h(Y) = \mathbf{E}(X|Y)$.

(iii) Let $y \in \mathbb{R}$. We are required to show that

$$\mathbf{E}(Xh(y)|Y=y) = h(y)\mathbf{E}(X|Y=y)$$

This is a property of conditional expectation from elementary probability. Now, let $y \in \mathbb{R}$. Recall that $W := \mathbf{E}(X|h(Y))$ takes the value $\mathbf{E}(X|\,h(Y)=z)$ with probability $\mathbf{P}(h(Y)=z)$. In particular, $\mathbf{E}(X|h(Y))$ is constant on any set of the form $\{h(Y)=z\}$, with $z \in \mathbb{R}$ fixed. So, $\mathbf{E}(X|h(Y))$ is constant on any set of the form $\{Y=y\}$, with $y \in \mathbb{R}$ fixed. Similarly, $\mathbf{E}(W|Y)$ is constant on any set of the form $\{Y=y\}$, with $y \in \mathbb{R}$ fixed. So, it suffices to show that both constants are the same, i.e. that $\mathbf{E}(W|Y=y) = \mathbf{E}(X|h(Y)=h(y))$. But by definition of W, we have

$$\mathbf{E}(W|Y=y) = \frac{1}{\mathbf{P}(Y=y)}\mathbf{E}W1_{Y=y} = \mathbf{E}(X|h(Y)=h(y))$$

(iv) For any $x \in \mathbb{R}$, we know that $\mathbf{E}(X|X=x) = x$, so $\mathbf{E}(X|X) = X$, by definition of conditional expectation. Also, we take it as given that $\mathbf{E}(X+Y|Z=z) = \mathbf{E}(X|Z=z) + \mathbf{E}(Y|Z=z)$, so that, by the definition of conditional expectation (as a random variable),

$$\mathbf{E}(X+Y|Z) = \mathbf{E}(X|Z) + \mathbf{E}(Y|Z).$$

(v) If Z is independent of X and Y, we know that

$$\mathbf{E}(X|Y=y, Z=z) = \mathbf{E}(X|Y=y).$$

Therefore, by definition of conditional expectation (as a random variable),

$$\mathbf{E}(X|Y,Z) = \mathbf{E}(X|Y).$$

□

Exercise 5.2 (**Conditional Jensen Inequality**)**.** Prove Jensen's inequality for the conditional expectation. Let $X, Y\colon \Omega \to \mathbb{R}$ be random variables that are either both discrete or both continuous. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be convex. Then

$$\phi(\mathbf{E}(X|Y)) \leq \mathbf{E}(\phi(X)|Y)$$

If ϕ is strictly convex, then equality holds only if X is constant on any set where Y is constant. That is, (by an Exercise from the previous homework) equality holds only if X is a function of Y.

(Hint: first show that if $X \geq Z$ then $\mathbf{E}(X|Y) \geq \mathbf{E}(Z|Y)$.)

Proof. If $X \geq Z$, then $X1_{Y=y} \geq Z1_{Y=y}$, so $\mathbf{E}(X|Y=y) \geq \mathbf{E}(Z|Y=y)$, so $\mathbf{E}(X|Y) \geq \mathbf{E}(Z|Y)$. Let $L\colon \mathbb{R} \to \mathbb{R}$ be any linear function such that $L(x) \leq \phi(x)$ for all $x \in \mathbb{R}$. Then

$$L(X) \leq \phi(X), \qquad \mathbf{E}(L(X)|Y) \leq \mathbf{E}(\phi(X)|Y).$$

Since L is linear, $\mathbf{E}(L(X)|Y) = L(\mathbf{E}(X|Y))$. So, for any linear L satisfying $L(x) \leq \phi(x)$ for all $x \in \mathbb{R}$,

$$L(\mathbf{E}(X|Y)) \leq \mathbf{E}(\phi(X)|Y). \qquad (*)$$

Since ϕ is convex, $\phi(x)$ is the supremum of $L(x)$ over all linear functions L satisfying $L(y) \leq \phi(y)$ for all $y \in \mathbb{R}$. So, taking the supremum of both sides of $(*)$ over all L shows that

$$\phi(\mathbf{E}(X|Y)) \leq \mathbf{E}(\phi(X)|Y).$$

□

Exercise 5.3. Let Y, Z be a statistics, and suppose Z is sufficient for $\{f_\theta\colon \theta \in \Theta\}$. Show that $W := \mathbf{E}_\theta(Y|Z)$ does not depend on θ. That is, there is a function $t\colon \mathbb{R}^n \to \mathbb{R}$ that does not depend on θ such that $W = t(X)$, where X is the random sample.

Solution. We consider the discrete case only. Since Z is sufficient for θ, the conditional distribution of the sample X conditioned on Z does not depend on θ. So, if $h\colon \mathbb{R}^n \to \mathbb{R}$ and if $Y = h(X)$, then

$$\mathbf{E}_\theta(h(X)|Z=z) = \sum_{x\in\mathbb{R}} h(x)\mathbf{P}_\theta(X=x|Z=z) =: g(z),$$

and the expression on the right does not depend on θ, by assumption. So, $\mathbf{E}_\theta(h(X)|Z) = g(Z)$, and since Z is a statistic, we can write $Z = f(X)$, so that $\mathbf{E}_\theta(h(X)|Z) = g(f(X))$, where both g, f have no dependence on θ. □

Exercise 5.4. Let $X_1, \ldots, X_n$ be a random sample of size n, so that X_1 is a sample from the uniform distribution on the interval $[\theta - 1/2, \theta + 1/2]$, where $\theta \in \mathbb{R}$ is unknown.

- Show that $(X_{(1)}, X_{(n)})$ is minimal sufficient but not complete.

- The sample mean $\overline{X}$ might seem to be a reasonable estimator for θ, but it is not a function of the minimal sufficient statistic, so maybe it is not so good. Find an unbiased estimator for θ with smaller variance than $\overline{X}$ (for all θ). Then, examine the ratio of the variances (i.e. relative efficiency) for $\overline{X}$ and your estimator. (Don't try to find a UMVU; it does not exist! We will show this on the next homework.)

Solution.

Minimal sufficiency follows by Theorem 5.8 in the notes. Note that

$$f_\theta(x_1, \ldots, x_n) = 1_{x_1,\ldots,x_n \in [\theta-1/2, \theta+1/2]} = 1_{x_{(1)}, x_{(n)} \in [\theta-1/2, \theta+1/2]}.$$

So, $f_\theta(x) = c(x,y) f_\theta(y)$ for all $\theta \in \mathbb{R}$ if and only if, for all $\theta \in \mathbb{R}$, we have the dichotomy that either

- $x_{(1)}, x_{(n)}, y_{(1)}, y_{(n)} \in [\theta - 1/2, \theta + 1/2]$, or
- $x_{(1)}, x_{(n)}, y_{(1)}, y_{(n)} \notin [\theta - 1/2, \theta + 1/2]$.

The last condition holds if and only if $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$. (The converse is clear, and for the forward direction, note e.g. if $x_{(1)} < y_{(1)}$, then choosing $\theta := y_{(1)} + 1/2$, we get $x_{(1)} \notin [\theta - 1/2, \theta + 1/2]$ but $y_{(1)} \in [\theta - 1/2, \theta + 1/2]$.)

For the unbiased estimator, consider

$$Y := \frac{X_{(1)} + X_{(n)}}{2}.$$

Note that the sample mean has variance $\frac{1}{12n}$. To compute the variance of Y, note that

$$\mathbf{P}_\theta(X_{(n)} < t) = [\mathbf{P}_\theta(X_1 < t)]^n = \begin{cases} 0 & \text{if } t < \theta - 1/2 \\ (t - \theta + 1/2)^n & \text{if } t \in [\theta - 1/2, \theta + 1/2] \\ 1 & \text{if } t > \theta + 1/2 \end{cases}.$$

$$\mathbf{P}_\theta(X_{(1)} > t) = [\mathbf{P}_\theta(X_1 > t)]^n = \begin{cases} 1 & \text{if } t < \theta - 1/2 \\ (\theta + 1/2 - t)^n & \text{if } t \in [\theta - 1/2, \theta + 1/2] \\ 0 & \text{if } t > \theta + 1/2 \end{cases}..$$

$$\begin{aligned} \mathbf{E}_\theta X_{(n)} &= \int_0^{\theta+1/2} \mathbf{P}_\theta(X_{(n)} > t) dt = \theta + 1/2 - \int_{\theta-1/2}^{\theta+1/2} \mathbf{P}_\theta(X_{(n)} < t) dt \\ &= \theta + 1/2 - \int_{\theta-1/2}^{\theta+1/2} (t - \theta + 1/2)^n dt = \theta + 1/2 - \frac{1}{n+1}. \end{aligned}$$

$$\begin{aligned} \mathbf{E}_\theta X_{(1)} &= \int_0^{\theta+1/2} \mathbf{P}_\theta(X_{(1)} > t) dt = \theta - 1/2 + \int_{\theta-1/2}^{\theta+1/2} \mathbf{P}_\theta(X_{(1)} > t) dt \\ &= \theta - 1/2 + \int_{\theta-1/2}^{\theta+1/2} (\theta + 1/2 - t)^n dt = \theta - 1/2 + \frac{1}{n+1}. \end{aligned}$$

Therefore,

$$\mathbf{E}_\theta(X_{(1)} + (X_{(n)}))/2 = \theta, \qquad \forall\, \theta \in \Theta.$$

For the variance of Y, we compute

$$\begin{aligned}
\mathbf{E}_\theta X_{(n)}^2 &= \int_0^{\theta+1/2} 2t\mathbf{P}_\theta(X_{(n)} > t)dt = (\theta+1/2)^2 - \int_{\theta-1/2}^{\theta+1/2} 2t\mathbf{P}_\theta(X_{(n)} < t)dt \\
&= (\theta+1/2)^2 - \int_{\theta-1/2}^{\theta+1/2} 2t(t-\theta+1/2)^n dt \\
&= (\theta+1/2)^2 - \frac{2(\theta+1/2)}{n+1} + \int_{\theta-1/2}^{\theta+1/2} \frac{2}{n+1}(t-\theta+1/2)^{n+1}dt \\
&= (\theta+1/2)^2 - \frac{2(\theta+1/2)}{n+1} + \frac{2}{(n+1)(n+2)}.
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}_\theta X_{(1)}^2 &= \int_0^{\theta+1/2} 2t\mathbf{P}_\theta(X_{(1)} > t)dt = (\theta-1/2)^2 + \int_{\theta-1/2}^{\theta+1/2} 2t\mathbf{P}_\theta(X_{(1)} > t)dt \\
&= (\theta-1/2)^2 + \int_{\theta-1/2}^{\theta+1/2} 2t(\theta+1/2-t)^n dt \\
&= (\theta-1/2)^2 + \frac{2(\theta-1/2)}{n+1} + \int_{\theta-1/2}^{\theta+1/2} \frac{2}{n+1}(\theta+1/2-t)^n dt \\
&= (\theta-1/2)^2 + \frac{2(\theta-1/2)}{n+1} + \frac{2}{(n+1)(n+2)}.
\end{aligned}$$

From Theorem 5.4.6 in the book, the joint density of $X_{(1)}$ and $X_{(n)}$ is (when $n > 1$)

$$f_{X_{(1)},X_{(n)}}(u,v) = n(n-1)1_{u\in[\theta-1/2,\theta+1/2]}1_{v\in[\theta-1/2,\theta+1/2]}1_{u<v}[F_{X_1}(v) - F_{X_1}(u)]^{n-2},$$

where F_X is the cumulative distribution function of X_1, so that $F_{X_1} = \mathbf{P}(X_1 > t) = 1_{t<\theta+1/2}\min(1,(\theta+1/2-t))$. So,

$$\begin{aligned}
\mathbf{E}X_{(1)}X_{(n)} &= n(n-1)\iint_{\mathbb{R}^2} uvf_{X_{(1)},X_{(n)}}(u,v)dudv \\
&= n(n-1)\int_{u=\theta-1/2}^{u=\theta+1/2}\int_{v=\theta-1/2}^{v=\theta+1/2} uv1_{u<v}[1_{v<\theta+1/2}\min(1,(\theta+1/2-v)) \\
&\qquad - 1_{u<\theta+1/2}\min(1,(\theta+1/2-u))]^{n-2}dvdu \\
&= n(n-1)\int_{u=\theta-1/2}^{u=\theta+1/2}\int_{v=\theta-1/2}^{v=\theta+1/2} uv1_{u<v}[(\theta+1/2-v)-(\theta+1/2-u)]^{n-2}dvdu \\
&= n(n-1)\int_{u=\theta-1/2}^{u=\theta+1/2}\int_{v=\theta-1/2}^{v=\theta+1/2} 1_{u<v}uv[v-u]^{n-2}dvdu
\end{aligned}$$

According to the computer, this integral is

$$\theta^2 - \frac{n-2}{4n+8}.$$

In summary,

$$\begin{aligned}
\mathrm{Var}_\theta \frac{X_{(1)}+X_{(n)}}{2} &= \mathbf{E}_\theta\Big(\frac{X_{(1)}+X_{(n)}}{2}-\theta\Big)^2 \\
&= \frac{1}{4}\mathbf{E}_\theta X_{(1)}^2 + \frac{1}{4}\mathbf{E}_\theta X_{(1)}^2 + \frac{1}{2}\mathbf{E}_\theta X_{(1)}X_{(n)} - \theta^2 \\
&= \frac{1}{4}\Big((\theta-1/2)^2 + \frac{2(\theta-1/2)}{n+1} + \frac{2}{(n+1)(n+2)}\Big) \\
&\quad + \frac{1}{4}\Big((\theta+1/2)^2 - \frac{2(\theta+1/2)}{n+1} + \frac{2}{(n+1)(n+2)}\Big) \\
&\quad + \frac{1}{2}(\theta^2 - \frac{n-2}{4n+8}) - \theta^2 \\
&= -\frac{4}{8}\frac{1}{n+1} + \frac{1}{8} - \frac{1}{8}\frac{n-2}{n+2} + \frac{1}{(n+1)(n+2)} \\
&= -\frac{4}{8}\frac{1}{n+1} + \frac{1}{8}\frac{4}{n+2} + \frac{1}{(n+1)(n+2)} \\
&= \frac{1}{2}\frac{-(n+2)+(n+1)}{(n+1)(n+2)} + \frac{1}{(n+1)(n+2)} = \frac{1}{2}\frac{1}{(n+1)(n+2)}.
\end{aligned}$$

□

Exercise 5.5. Let $X_1, \ldots, X_n$ be a random sample of size $n = 2$, so that X_1 is a sample from exponential distribution with unknown parameter $\theta > 0$, so that X_1 has density $\theta e^{-x\theta}1_{x>0}$.

Suppose we want to estimate the mean

$$g(\theta) := 1/\theta.$$

- Using the Rao-Blackwell Theorem (or any other method), find the UMVU for $g(\theta)$.
- Show that $\sqrt{X_1X_2}$ has smaller mean squared error than the UMVU.
- Find an estimator with even smaller mean squared error, for all $\theta \in \Theta$.

Solution.

We try to write the joint density

$$f_\theta(x_1,x_2) = \theta^2 e^{-\theta(x_1+x_1)}1_{x_1>0,x_2>0} = 1_{x_1>0,x_2>0}e^{-\theta(x_1+x_2)-(-2\log\theta)}$$

as an exponential family, in order to find the complete sufficient statistics. To this end, let $t(x_1,x_2) := x_1+x_2$. Then $t(X_1,X_2) = X_1+X_2$ is complete and sufficient for θ by Exercises 4.7 and 4.11 (since $\theta \mapsto 1/\theta$ is a bijection on the domain $\theta \in (0,\infty)$, X_1+X_2 is also complete and sufficient for $1/\theta$). The statistic X_1+X_2 is also unbiased if we divide by 2, so that $Y := (X_1+X_2)/2$ is unbiased, complete and sufficient for $1/\theta$. So, by Lehman-Scheffé's Theorem, the UMVU for $1/\theta$ is

$$\mathbf{E}_\theta(Y|Y) = Y = \frac{1}{2}(X_1+X_2).$$

The variance of Y is

$$\mathrm{Var}_\theta(Y) = \frac{1}{4}\mathrm{Var}_\theta(X_1) + \frac{1}{4}\mathrm{Var}_\theta(X_2) = \frac{1}{2}\mathrm{Var}_\theta(X_1) = \frac{1}{2\theta^2}.$$

On the other hand, $Z := \sqrt{X_1 X_2}$ satisfies

$$\begin{aligned}\mathbf{E}_\theta(Z-\theta^{-1})^2 &= \mathbf{E}_\theta X_1X_2 - 2\theta^{-1}\mathbf{E}_\theta\sqrt{X_1}\sqrt{X_2} + \theta^{-2}\\ &= (\mathbf{E}_\theta X_1)^2 - 2\theta^{-1}(\mathbf{E}_\theta\sqrt{X_1})^2 + \theta^{-2}\\ &= \theta^{-2} - 2\theta^{-1}(\theta^{-1/2}\sqrt{\pi}/2)^2 + \theta^{-2} = \theta^{-2}[1-(\pi/2)+1] = \theta^{-2}(2-\pi/2) < \theta^{-2}/2.\end{aligned}$$

Here we used

$$\mathbf{E}_\theta\sqrt{X_1} = \int_0^\infty x^{1/2}\theta e^{-\theta x}dx = \theta^{-1/2}\int_0^\infty x^{1/2}e^{-x}dx = \theta^{-1/2}\sqrt{\pi}/2.$$

There are a few ways to get a further improvement. Let $t > 0$ and consider $t\sqrt{X_1X_2}$. Then

$$\begin{aligned}\mathbf{E}_\theta(t\sqrt{X_1X_2}-\theta^{-1})^2 &= t^2\mathbf{E}_\theta X_1X_2 - 2t\theta^{-1}\mathbf{E}_\theta\sqrt{X_1}\sqrt{X_2} + \theta^{-2}\\ &= t^2(\mathbf{E}_\theta X_1)^2 - 2t\theta^{-1}(\mathbf{E}_\theta\sqrt{X_1})^2 + \theta^{-2}\\ &= t^2\theta^{-2} - 2t\theta^{-1}(\theta^{-1/2}\sqrt{\pi}/2)^2 + \theta^{-2} = \theta^{-2}[t^2 - t(\pi/2)+1].\end{aligned}$$

The minimum value of $t^2 - t\pi/2 + 1$ occurs when $t = \pi/4$, and in this case the mean squared error is

$$\theta^{-2}(1-\pi^2/16) \approx (.3831)\theta^{-2}.$$

□

6. Homework 6

Exercise 6.2. Let $X_1, \dots, X_n$ be a random sample of size n, so that X_1 is a sample from the uniform distribution on the interval $[\theta - 1/2, \theta + 1/2]$, where $\theta \in \mathbb{R}$ is unknown. From a previous homework, we tried to find a low variance estimator for θ, but the UMVU seemed to not exist. In this exercise, you are asked to show that a UMVU does not exist, using the following outline, in the case $n = 1$. Moreover, if $g(\theta)$ is a nonconstant differentiable function of $\theta \in \mathbb{R}$, show that no UMVU of $g(\theta)$ exists when $n = 1$:

- Let $U = u(X_1)$ be an unbiased estimator of 0, where $u\colon \mathbb{R} \to \mathbb{R}$. By differentiating the definition of unbiasedness with respect to θ, conclude that
$$u(x+1) = u(x), \qquad \text{for a.e. } x \in \mathbb{R}.$$
Give an example of an unbiased estimator U of 0 such that $u(x) \neq 0$ for all $x \in \mathbb{R}$.
- Argue by contradiction. Assume that W is UMVU for $g(\theta)$. Using the characterization from class, conclude that $\mathbf{E}_\theta WU = 0$, so that if $W = w(X_1)$ with $w\colon \mathbb{R} \to \mathbb{R}$, then
$$w(x+1)u(x+1) = w(x)u(x), \qquad \text{for a.e. } x \in \mathbb{R}.$$
Then conclude that
$$w(x+1) = w(x), \qquad \text{for a.e. } x \in \mathbb{R}.$$
- To complete the exercise, what can you say about the condition that W is unbiased for $g(\theta)$?

(Optional) Can you make the same conclusion for a sample of size 2? Hint: Fourier series.

Solution. We have $\int_{\theta-1/2}^{\theta+1/2} u(x)dx = 0$ for all $\theta \in \mathbb{R}$. Differentiating this condition and applying the Fundamental Theorem of Calculus, we get $u(\theta + 1/2) = u(\theta - 1/2)$ for a.e. $\theta \in \mathbb{R}$. (It is assumed that $\mathbf{E}_\theta |U| < \infty$ for all $\theta \in \mathbb{R}$ in order to be an unbiased estimator of 0.) For an example of an unbiased estimator of 0, consider e.g. $u(x) := \operatorname{sign}(\sin(2\pi x))$ for all $x \in \mathbb{R}$.

Assume that W is UMVU. By Theorem 6.18, an Alternate Characterization of UMVU, we must have $\mathbf{E}_\theta WU = 0$ for all $\theta \in \mathbb{R}$ when U is unbiased for 0. That is, $\mathbf{E}_\theta WU = 0$ for all $\theta \in \mathbb{R}$. By step one, we conclude that

$$w(x+1)u(x+1) = w(x)u(x), \qquad \text{for a.e. } x \in \mathbb{R}.$$

Using the u we produced there, we can divided both sides by u to conclude that

$$w(x+1) = w(x), \qquad \text{for a.e. } x \in \mathbb{R}.$$

Now, since W is unbiased for $g(\theta)$, we have $g(\theta) = \int_{\theta-1/2}^{\theta+1/2} w(x)dx$. Differentiating both sides, we get

$$g'(\theta) = w(\theta + 1/2) - w(\theta - 1/2) = 0,$$

so that $g(\theta)$ is constant in θ. □

Exercise 6.3. Let $X_1, \ldots, X_n$ be a random sample of size n, so that X_1 is a sample from the uniform distribution on the interval $[\theta - 1/2, \theta + 1/2]$, where $\theta \in \mathbb{R}$ is unknown. Although the UMVU for θ does not exist and unbiased estimators do exist, if we instead restrict to *location equivariant* estimators, then there *is* a minimum variance estimator of θ among this class. We say that an $Y := t(X_1, \ldots, X_n)$ with $t\colon \mathbb{R}^n \to \mathbb{R}$ is location equivariant if

$$t(x_1, \ldots, x_n) + a = t(x_1 + a, \ldots, x_n + a), \qquad \forall\, (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad \forall\, a \in \mathbb{R}.$$

- Using location equivariance for the density $f := 1_{[-1/2,1/2]}$, and letting $f_\theta(x) := f(x - \theta)$, show that

$$\mathbf{E}_\theta(W - \theta)^2 = \int_{\mathbb{R}^n} [t(x)]^2 \prod_{i=1}^n f(x_i)dx_1 \cdots dx_n, \qquad \forall\, \theta \in \Theta.$$

  (Note that the expression on the right does not depend on θ.)
- Let $H := \{x \in \mathbb{R}^n\colon \langle x, (1, \ldots, 1)\rangle = 0\}$, where as usual $\langle (x_1, \ldots, x_n), (y_1, \ldots, y_n)\rangle = \sum_{i=1}^n x_i y_i$. Using location equivariance again, show that

$$\mathbf{E}(W - \theta)^2 = \int_H \Big(\int_{\mathbb{R}} |t(x) - a|^2 \prod_{i=1}^n f(x_i - a)da \Big) dH(x).$$

  (Here $dH(x)$ denotes integration on the hypersurface H, i.e. $dH(x)$ is not the same as $dx_1 \cdots dx_n$)
- So, to minimize $\mathbf{E}(W - \theta)^2$, it suffices to minimize $\int_{\mathbb{R}} [a - t(x)]^2 \prod_{i=1}^n f(x_i - a)da$, for any fixed $x \in H$. What choice of $t(x)$ minimizes $\int_{\mathbb{R}} [a - t(x)]^2 \prod_{i=1}^n f(x_i - a)da$, when $x \in H$ is fixed?
- Conclude that the W minimizing $\mathbf{E}(W - \theta)^2$ for all $\theta \in \mathbb{R}$, over all location equivariant estimators satisfies

$$W = \frac{\int_{\mathbb{R}} a \prod_{i=1}^n f(X_i - a)da}{\int_{\mathbb{R}} \prod_{i=1}^n f(X_i - a)da}.$$

- So, in our original example when $f = 1_{[-1/2,1/2]}$, show that $W = \frac{X_{(1)}+X_{(n)}}{2}$ achieves the minimum variance among location equivariant estimators, despite the UMVU not existing. This estimator is also unbiased, but this was not guaranteed to occur in our construction.
- (Optional) Perform the above analysis for $f_\theta(x) := \theta^{-1} f(x/\theta)$, $\theta > 0$ to find the variance minimizer among *scale-equivariant* estimators

$$t(ax_1, \ldots, ax_n) = at(x), \qquad \forall\, x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad \forall\, a > 0.$$

You should find the optimal estimator to be

$$t(x) := \frac{\int_{\mathbb{R}} a^n \prod_{i=1}^n f(ax_i) da}{\int_{\mathbb{R}} a^{n+1} \prod_{i=1}^n f(ax_i) da}$$

Solution. Applying the equivariance property, changing variables, using Fubini's Theorem, then using equivariance again

$$\begin{aligned}
\mathbf{E}_\theta (W-\theta)^2 &= \int_{\mathbb{R}^n} |t(x) - \theta|^2 \prod_{i=1}^n f(x_i - \theta) dx_1 \cdots dx_n \\
&= \int_{\mathbb{R}^n} |t(x_1 - \theta, \ldots, x_n - \theta)|^2 \prod_{i=1}^n f(x_i - \theta) dx_1 \cdots dx_n \\
&= \int_{\mathbb{R}^n} |t(x)|^2 \prod_{i=1}^n f(x_i) dx_1 \cdots dx_n
\end{aligned}$$

We now change variables, so that $y_1 = x_1 + a, y_2 = x_2 + a, \ldots, y_{n-1} = x_{n-1} + a$ and $a = x_n$. Then the Jacobian determinant of this change of variables is 1, so

$$\mathbf{E}_\theta (W-\theta)^2 = \int_{\mathbb{R}^n} |t(y_1 - a, \ldots, y_{n-1} - a, a)|^2 \prod_{i=1}^{n-1} f(y_i - a) f(a) dy_1 \cdots dy_{n-1} da$$

Changing variables again so that $z = x_n - a$

$$\begin{aligned}
&= \int_H \Big(\int_{\mathbb{R}} |t(x_1 - z, \ldots, x_{n-1} - z, z)|^2 \prod_{i=1}^{n-1} f(x_i - z) f(z) dz \Big) dH(x) \\
&= \int_H \Big(\int_{\mathbb{R}} |t(x_1 - a, \ldots, x_n - a)|^2 \prod_{i=1}^{n} f(x_i - a) da \Big) dH(x) \\
&= \int_H \Big(\int_{\mathbb{R}} |t(x) - a|^2 \prod_{i=1}^{n} f(x_i - a) da \Big) dH(x).
\end{aligned}$$

Now, write

$$\int_{\mathbb{R}} |t(x) - a|^2 \prod_{i=1}^n f(x_i - a) da = \int_{\mathbb{R}} ([t(x)]^2 - 2at(x) + a^2) \prod_{i=1}^n f(x_i - a) da.$$

With all other quantities fixed, the value of $t(x)$ minimizing this expression is

$$t(x) := \frac{\int_{\mathbb{R}} a \prod_{i=1}^n f(x_i - a) da}{\int_{\mathbb{R}} \prod_{i=1}^n f(x_i - a) da}, \qquad \forall\, x \in H.$$

We can write any $y \in \mathbb{R}^n$ as $y = b(1,\dots,1) + (y - b(1,\dots,1))$ where $b := (\sum_{i=1}^n y_i)$ and $x := (y - b(1,\dots,1)) \in H$ so that, by the equivariance property,

$$\begin{aligned} t(y) &= t(x + (b,\dots,b)) = t(x) + b = \frac{\int_{\mathbb{R}} a \prod_{i=1}^n f(y_i - a - b)da}{\int_{\mathbb{R}} \prod_{i=1}^n f(y_i - a - b)da} + b \\ &= \frac{\int_{\mathbb{R}} (a-b) \prod_{i=1}^n f(y_i - a)da}{\int_{\mathbb{R}} \prod_{i=1}^n f(y_i - a)da} + b = \frac{\int_{\mathbb{R}} a \prod_{i=1}^n f(y_i - a)da}{\int_{\mathbb{R}} \prod_{i=1}^n f(y_i - a)da} + b - b \\ &= \frac{\int_{\mathbb{R}} a \prod_{i=1}^n f(y_i - a)da}{\int_{\mathbb{R}} \prod_{i=1}^n f(y_i - a)da}. \end{aligned}$$

When $f = 1_{[-1,2/1/2]}$, we have

$$\begin{aligned} \prod_{i=1}^n f(x_i - a) &= 1_{x_1,\dots,x_n \in [a-1/2, a+1/2]} = 1_{x_{(1)}, x_{(n)} \in [a-1/2, a+1/2]} \\ &= 1_{a-1/2 \le x_{(1)} \le x_{(n)} \le a+1/2} = 1_{a \le x_{(1)}+1/2 \le x_{(n)}+1/2 \le a+1}. \end{aligned}$$

So,

$$\begin{aligned} \int_{\mathbb{R}} a \prod_{i=1}^n f(x_i - a)da &= \int_{\mathbb{R}} a 1_{a \le x_{(1)}+1/2 \le x_{(n)}+1/2 \le a+1} da \\ &= \int_{x_{(n)}-1/2}^{x_{(1)}+1/2} ada = \frac{[x_{(1)} + 1/2]^2 - [x_{(n)} - 1/2]^2}{2}. \end{aligned}$$

$$\int_{\mathbb{R}} \prod_{i=1}^n f(x_i - a)da = \int_{\mathbb{R}} 1_{a \le x_{(1)}+1/2 \le x_{(n)}+1/2 \le a+1} da = \int_{x_{(n)}-1/2}^{x_{(1)}+1/2} da = x_{(1)} - x_{(n)} + 1.$$

$$t(x) = \frac{\int_{\mathbb{R}} a \prod_{i=1}^n f(x_i - a)}{\int_{\mathbb{R}} \prod_{i=1}^n f(x_i - a)} = \frac{1}{2} \frac{x_{(1)}^2 - x_{(n)}^2 + x_{(1)} + x_{(n)}}{x_{(1)} - x_{(n)} + 1} = \frac{1}{2}[x_{(1)} + x_{(n)}].$$

□

Solution. [Optional]

Applying the equivariance property, changing variables, changing to hyperspherical coordinates, then using equivariance again

$$\begin{aligned}
\mathbf{E}_\theta(W-\theta)^2 &= \int_{\mathbb{R}^n} |t(x)-\theta|^2 \prod_{i=1}^n \theta^{-1} f(x_i/\theta) dx_1 \cdots dx_n \\
&= \int_{\mathbb{R}^n} |t(\theta x)-\theta|^2 \prod_{i=1}^n f(x_i) dx_1 \cdots dx_n \\
&= \theta^2 \int_{\mathbb{R}^n} |t(x)-1|^2 \prod_{i=1}^n f(x_i) dx_1 \cdots dx_n \\
&= \theta^2 \int_{S^{n-1}} \Big(\int_0^\infty a^{n-1} |t(ax)-1|^2 \prod_{i=1}^n f(ax_i) da \Big) dS(x) \\
&= \theta^2 \int_{S^{n-1}} \Big(\int_0^\infty a^{n-1} |at(x)-1|^2 \prod_{i=1}^n f(ax_i) da \Big) dS(x)
\end{aligned}$$

Now, write

$$\int_0^\infty a^{n-1} |at(x)-1|^2 \prod_{i=1}^n f(ax_i) da = \int_0^\infty a^{n-1} [a^2 [t(x)]^2 - 2at(x) + 1] \prod_{i=1}^n f(ax_i) da.$$

With all other quantities fixed, the value of $t(x)$ minimizing this expression is

$$t(x) := \frac{\int_{\mathbb{R}} a^n \prod_{i=1}^n f(ax_i) da}{\int_{\mathbb{R}} a^{n+1} \prod_{i=1}^n f(ax_i) da}, \qquad \forall\, x \in H.$$

We can write any $y \in \mathbb{R}^n$ as $y = \|y\| [y/\|y\|]$ so that, by the equivariance property

$$\begin{aligned}
t(y) &= \|y\| \, t(y/\|y\|) = \|y\| \, \frac{\int_{\mathbb{R}} a^n \prod_{i=1}^n f(a y_i/\|y\|) da}{\int_{\mathbb{R}} a^{n+1} \prod_{i=1}^n f(a y_i/\|y\|) da} \\
&= \frac{\int_{\mathbb{R}} a^n \prod_{i=1}^n f(a y_i) da}{\int_{\mathbb{R}} a^{n+1} \prod_{i=1}^n f(a y_i) da}.
\end{aligned}$$

□

Exercise 6.4. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a convex function. Let $x \in \mathbb{R}^n$ be a local minimum of f. Show that x is in fact a global minimum of f.

Now suppose additionally that f is a C^1 function (all derivatives of f exist and are continuous), and $x \in \mathbb{R}^n$ satisfies $\nabla f(x) = 0$. Show that x is a global minimum of f.

Solution. Let $y \in \mathbb{R}^n$ with $y \neq x$. Let $t \in (0,1)$. Assume for the sake of contradiction that $f(y) < f(x)$. By convexity of f,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) < (t + (1-t))f(x) = f(x).$$

Letting $t \to 1^-$ shows that points near x have smaller f values than x, contradicting the local minimality of x. We conclude that $f(y) \geq f(x)$ for all $y \in \mathbb{R}^n$, so that x is a global minimum of f.

In the case that f is C^1, $\nabla f(x) = 0$ implies that x is a local minimum of f, so that the first assertion implies that x is a global minimum of f. To see this, note that the definition of convexity of f implies that the function f lies above the horizontal tangent plane of f at x. $\square$

Exercise 6.5. Let A be a real $m \times n$ matrix. Let $x \in \mathbb{R}^n$ and let $b \in \mathbb{R}^m$. Show that the function $f\colon \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \frac{1}{2}\|Ax - b\|^2$ is convex. Moreover, show that

$$\nabla f(x) = A^T(Ax - b), \qquad D^2 f(x) = A^T A.$$

(Here $D^2 f$ denotes the matrix of second derivatives of f.)

So, if $\nabla f(x) = 0$, i.e. if $A^T Ax = A^T b$, then x is the global minimum of f. And if A has full rank, then $A^T A$ is invertible, so that $x = (A^T A)^{-1} A^T b$ is the global minimum of f.

Solution. We have

$$f(x) = \frac{1}{2}(x^T A^T Ax - x^T A^T b - b^T Ax + b^T b) = \frac{1}{2}\sum_{i,k=1}^{n}\sum_{j=1}^{m} x_i a_{ji} a_{jk} x_k - \sum_{i=1}^{n}\sum_{j=1}^{m} x_i a_{ji} b_j + \frac{1}{2}\sum_{i=1}^{m} b_i^2.$$

So convexity follows since, if $t \in (0,1)$ and $x, y \in \mathbb{R}^n$,

$$\begin{aligned}
&2[tf(x) + (1-t)f(y) - f(tx + (1-t)y)] \\
&\quad = tx^T A^T Ax + (1-t)y^T A^T Ay - (tx + (1-t)y)^T A^T A(tx + (1-t)y) \\
&\quad = (t - t^2)x^T A^T Ax + [(1-t) - (1-t)^2]y^T A^T Ay - 2t(1-t)x^T A^T Ay \\
&\quad = t(1-t)\Big(x^T A^T Ax + y^T A^T Ay - 2x^T A^T Ay\Big) \\
&\quad = t(1-t)(x-y)^T A^T A(x-y) \geq 0.
\end{aligned}$$

The last inequality uses that $A^T A$ is a positive semidefinite matrix, so $(x-y)^T A^T A(x-y) \geq 0$.

We use the coordinate expression to differentiate f, so that

$$\begin{aligned}
\frac{\partial}{\partial x_{i'}} f(x) &= \sum_{j=1}^{m} x_{i'} a_{ji'}^2 + \frac{\partial}{\partial x_{i'}} \sum_{k \in \{1,\ldots,n\}\colon k \neq i'} \sum_{j=1}^{m} x_{i'} a_{ji'} a_{jk} x_k - \frac{\partial}{\partial x_{i'}} \sum_{j=1}^{m} x_{i'} a_{ji'} b_j \\
&= \sum_{j=1}^{m} x_{i'} a_{ji'}^2 + \sum_{k \in \{1,\ldots,n\}\colon k \neq i'} \sum_{j=1}^{m} a_{ji'} a_{jk} x_k - \sum_{j=1}^{m} a_{ji'} b_j \\
&= \sum_{k=1}^{n}\sum_{j=1}^{m} a_{ji'} a_{jk} x_k - \sum_{j=1}^{m} a_{ji'} b_j \\
&= (A^T Ax)_{i'} - (A^T b)_{i'} = [A^T(Ax - b)]_{i'}.
\end{aligned}$$

Therefore, $\nabla f(x) = A^T(Ax - b)$. Similarly,

$$\frac{\partial}{\partial x_{i'}}\frac{\partial}{\partial x_{j'}} f(x) = \frac{\partial}{\partial x_{j'}}\Big(\sum_{k=1}^{n}\sum_{j=1}^{m} a_{ji'} a_{jk} x_k - \sum_{j=1}^{m} a_{ji'} b_j\Big) = \sum_{j=1}^{m} a_{ji'} a_{jj'} = (A^T A)_{i'j'}$$

Therefore, $D^2 f(x) = A^T A$. $\square$

Exercise 6.6 (Least Squares/ Ridge Regression). Let $Z_1, \ldots, Z_n$ be independent identically distributed Gaussian random variables with zero mean and known variance $\sigma^2 > 0$. Suppose $w \in \mathbb{R}^k$ is an unknown vector, and for all $1 \leq i \leq n$, there are known vectors $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^k$. Our observed data are

$$X_i := \langle x^{(i)}, w \rangle + Z_i, \qquad \forall\, 1 \leq i \leq n.$$

Here $Z_1, \ldots, Z_n$ represent experimental noise. The goal is to determine w.

So, our data are $X = (X_1, \ldots, X_n)^T$. In this exercise we restrict attention to *linear estimators*, i.e. we only consider statistics of the form

$$Y := BX,$$

where B is a $k \times n$ real matrix.

- Let A be the $n \times k$ matrix so that the i^{th} row of A is the row vector $x^{(i)}$) . Assume that $k \leq n$ and the matrix A has full rank. Find the value of $w \in \mathbb{R}^k$ that minimizes the quantity

$$\sum_{i=1}^{n} (X_i - \langle x^{(i)}, w \rangle)^2$$

  (considering w as a variable, with all other quantities fixed.)
- Find the unbiased estimator of w with minimal variance, among all linear estimators. That is, minimize

$$\mathbf{E} \|Y - w\|^2 = \mathbf{E} \sum_{j=1}^{k} (Y_j - w_j)^2$$

  over all choices of B such that $\mathbf{E}Y = w$, where Y is an arbitrary linear estimator. (Hint: If $\mathbf{E}Y = w$, what does the matrix B satisfy? Can you come up with some B that satisfies $\mathbf{E}BX = w$? Also, it might be easier to first compute the expected value of a matrix $\mathbf{E}(Y - w)(Y - w)^T$)

Solution. Define $f(w) := \sum_{i=1}^{n} (X_i - \langle x^{(i)}, w \rangle)^2 = \|X - Aw\|^2$. From Exercise 6.5, the global minimum of f is $w = (A^T A)^{-1} A^T X$.

$$X_i := \langle x^{(i)}, w \rangle + Z_i, \qquad \forall\, 1 \leq i \leq n.$$

Since $\mathbf{E}Y = w$, and $\mathbf{E}X = Aw$, we have $w = \mathbf{E}Y = \mathbf{E}BX = B\mathbf{E}X = BAw$, so that BA is the identity matrix I. (Since A is $n \times k$ and $k \leq n$, this does not imply that B is the inverse of A; indeed the inverse of a matrix is only formally defined for square matrices, though we could say that B is a left inverse in this case.) Note that we can write B as $B = (A^T A)^{-1} A^T + C$, where C satisfies $CA = 0$. Now,

$$\begin{aligned} \mathbf{E}(Y - w)(Y - w)^T &= \mathbf{E}(BX - w)(BX - w)^T = \mathbf{E}BXX^T B^T - \mathbf{E}BXw^T - \mathbf{E}w(BX)^T + w^T w \\ &= B(Aww^T A + I)B^T - w^T w = BB^T. \end{aligned}$$

The last equality used $BA = I$. Also, since $X = Aw + Z$, we have $\mathbf{E}XX^T = E(Aw + Z)(Aw + Z)^T = Aww^TA^T + I$. Then

$$\begin{aligned}\mathbf{E}(Y - w)(Y - w)^T = BB^T &= ((A^TA)^{-1}A^T + C)((A^TA)^{-1}A^T + C)^T \\ &= (AA^T)^{-1} + (A^TA)^{-1}A^TC^T + CA(AA^T)^{-1} + CC^T \\ &= (AA^T)^{-1} + CC^T.\end{aligned}$$

The last equality used $CA = 0$. So, we have

$$\mathbf{E}\left\|Y - w\right\|^2 = \mathrm{Tr}\mathbf{E}(Y - w)(Y - w)^T = \mathrm{Tr}((AA^T)^{-1}) + \mathrm{Tr}(CC^T) \geq \mathrm{Tr}((AA^T)^{-1}).$$

with equality when $C = 0$, i.e. when $B = (A^TA)^{-1}A^T$, as desired. The inequality used that CC^T is positive semidefinite. □

Exercise 6.7. Let $X_1, \ldots, X_n$ be a random sample of size n, so that X_1 has the Poisson distribution with parameter θ, i.e.

$$\mathbf{P}_\theta(X_1 = x) = \theta^x e^{-\theta}/x!, \qquad \forall \text{ nonnegative integers } x.$$

Suppose we want to estimate $\mathbf{P}_\theta(X_1 = 0) = e^{-\theta}$.

- One way we can try to estimate $e^{-\theta}$ is to count the fraction of zeros in the sample of size n. Define
$$Y_n := \frac{1}{n}\left|\{1 \leq i \leq n \colon X_i = 0\}\right|.$$
Find the limiting distribution of Y_n as $n \to \infty$.
- Give an explicit formula for the MLE Z_n of $e^{-\theta}$. Find the limiting distribution of Z_n as $n \to \infty$.
- Compute the relative efficiency of these two estimators as $n \to \infty$.

Solution. Since $Y_n = \frac{1}{n}\sum_{i=1}^n 1_{X_i=0}$ and $\mathbf{P}(X_i = 0) = e^{-\theta}$, the random variables $1_{X_i=0}$ are Bernoulli random variables with parameter $e^{-\theta}$, so that Y_n satisfies the conclusion of the LLN and CLT, i.e. Y_n converges almost surely to $e^{-\theta}$ as $n \to \infty$, and $\sqrt{n}(Y_n - e^{-\theta})$ converges to a mean zero Gaussian with variance $e^{-\theta}(1 - e^{-\theta})$ as $n \to \infty$.

The MLE of θ is a value of θ maximizing

$$\log \prod_{i=1}^n \theta^{X_i} e^{-\theta}/X_i! = \log\left(\theta^{\sum_{i=1}^n X_i} e^{-n\theta} \prod_{i=1}^n [X_i!]\right) = \sum_{i=1}^n \log(X_i!) - n\theta + \log\theta \sum_{i=1}^n X_i.$$

Taking a derivative in θ, we get $-n + \frac{1}{\theta}\sum_{i=1}^n X_i$. From the first derivative test, there is a unique maximum value of θ when $\theta = \frac{1}{n}\sum_{i=1}^n X_i$, so the MLE for θ is $\frac{1}{n}\sum_{i=1}^n X_i$. By the functional equivariance property of the MLE, the MLE for $e^{-\theta}$ is then

$$Z_n = e^{-\frac{1}{n}\sum_{i=1}^n X_i}.$$

Denote $f(\theta) := e^{-\theta}$. Note that $\sqrt{n}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i\right) - \theta\right]$ converges in distribution by the CLT to a mean zero Gaussian with variance θ as $n \to \infty$. From the Delta Method, $\sqrt{n}(Z_n - f(\theta))$ converges in distribution as $n \to \infty$ to a mean zero Gaussian with variance

$$\theta(f'(\theta))^2 = \theta e^{-2\theta}.$$

So, the relative efficiency of these estimators as $n \to \infty$ is the ratio of the variances, i.e.

$$\frac{\theta e^{-2\theta}}{e^{-\theta}(1-e^{-\theta})} = \frac{\theta e^{-\theta}}{1-e^{-\theta}} = \frac{\theta}{e^{\theta}-1}.$$

This quantity is less than 1 when $\theta > 0$, so that the MLE has strictly smaller variance (better efficiency) as $n \to \infty$. □

Exercise 6.8. Let $X_1, \ldots, X_n$ be a random sample of size n, so that X_1 has the Laplace density $\frac{1}{2}e^{-|x-\theta|}$ for all $x \in \mathbb{R}$, where $\theta \in \mathbb{R}$ is unknown. Find the MLE of θ.

Solution. The log likelihood satisfies

$$\log \ell(\theta) = \log \prod_{i=1}^{n} (1/2) e^{-|x_i - \theta|} = n \log 2 - \sum_{i=1}^{n} |x_i - \theta| .$$

For each $1 \le i \le n$, the function $\theta \mapsto |x_i - \theta|$ is convex in θ. Since a sum of convex functions is convex, the function $\theta \mapsto \sum_{i=1}^{n} |x_i - \theta|$ is also a convex function in θ. Moreover, $(d/d\theta) \log \ell(\theta) < 0$ as $\theta \to \infty$ and $(d/d\theta) \log \ell(\theta) > 0$ as $\theta \to \infty$. Since additionally $\log \ell(\theta)$ is concave, it has a maximum value.

Let Y_n be any median of $X_1, \ldots, X_n$, so that half of the values of $X_1, \ldots, X_n$ are at least Y_n, and half of the values of $X_1, \ldots, X_n$ are at most Y_n. At values of θ where $\log \ell(\theta)$ is differentiable, we have

$$\frac{d}{d\theta} \log \ell(\theta) = -\sum_{i=1}^{n} \operatorname{sign}(x_i - \theta) = \#\{1 \le i \le n \colon x_i < \theta\} - \#\{1 \le i \le n \colon x_i > \theta\}.$$

Consequently, $(d/d\theta) \log \ell(\theta) \le 0$ for $\theta \le Y_n$ and $(d/d\theta) \log \ell(\theta) \ge 0$ for $\theta \ge Y_n$. It follows that Y_n is an MLE for θ. □

7. Homework 7

Exercise 7.1. Consistency of a continuous method of moments estimator follows from the following statement, which you are required to prove.

Fix $k \ge 1$. For any $1 \le j \le k$, let $M_{j,1}, M_{j,2}, \ldots$ be real-valued random variables that converge in probability to a constant $c_j \in \mathbb{R}$. Let $h \colon \mathbb{R}^k \to \mathbb{R}$ be continuous. Then, as $n \to \infty$,

$$h(M_{1,n}, \ldots, M_{j,n})$$

converges in probability to the constant $h(c_1, \ldots, c_j)$.

Solution. Since h is continuous, for any $\varepsilon > 0$, there exists $\delta > 0$ such that, if $\sum_{i=1}^{k} |c_i - m_i| < \delta$, then $|h(c_1, \ldots, c_k) - h(m_1, \ldots, m_k)| < \varepsilon$.

Let $\varepsilon > 0$, and then let $\delta > 0$ as above. By assumption, for any $1 \le i \le k$, we have

$$\lim_{n \to \infty} \mathbf{P}(|M_{i,n} - c_i| > \delta) = 0. \qquad (*)$$

We therefore write

$$\begin{aligned}
&\mathbf{P}(|h(c_1,\dots,c_k) - h(M_{1,n},\dots,M_{k,n})| > \varepsilon) \\
&= \mathbf{P}(|h(c_1,\dots,c_k) - h(M_{1,n},\dots,M_{k,n})| > \varepsilon, \sum_{i=1}^{k} |c_i - M_{i,n}| < \delta) \\
&\qquad + \mathbf{P}(|h(c_1,\dots,c_k) - h(M_{1,n},\dots,M_{k,n})| > \varepsilon, \sum_{i=1}^{k} |c_i - M_{i,n}| \geq \delta) \\
&\leq \mathbf{P}(|h(c_1,\dots,c_k) - h(M_{1,n},\dots,M_{k,n})| > \varepsilon, \sum_{i=1}^{k} |c_i - M_{i,n}| < \delta) \\
&\qquad + \mathbf{P}(\sum_{i=1}^{k} |c_i - M_{i,n}| \geq \delta)
\end{aligned}$$

The first probability is zero, by definition of δ (since h is continuous, this event is the empty set). The second probability is bounded by $\sum_{i=1}^{k} \mathbf{P}(|c_i - M_{i,n}| \geq \delta)$, which goes to zero as $n \to \infty$ by $(*)$. Therefore, for any $\varepsilon > 0$, we have shown that

$$\lim_{n\to\infty} \mathbf{P}(|h(c_1,\dots,c_k) - h(M_{1,n},\dots,M_{k,n})| > \varepsilon) = 0.$$

□

Exercise 7.2. This exercise demonstrates that the MLE might not be consistent.

Let Z be a Gaussian random variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Then $X := e^Z$ has the lognormal distribution with parameters μ and σ^2. Let $\gamma \in \mathbb{R}$ and define

$$X' := \gamma + e^Z.$$

In this case X' is said to have the three-parameter lognormal distribution with parameters $\gamma, \mu \in \mathbb{R}$, and $\sigma^2 > 0$. Let $X_1, \dots, X_n$ be i.i.d. from this three-parameter lognormal distribution.

- Find the density of X_1.
- Suppose γ is known. Find the maximum likelihood estimator (M, T) of (μ, σ^2). (Assume $\gamma < X_{(1)}$.)
- Let $\ell(\gamma, \mu, \sigma^2)$ denote the log-likelihood function. The MLE of (γ, μ, σ^2) if it exists, will maximize $\ell(\gamma, M, T)$ over γ. Determine

$$\lim_{\gamma \uparrow X_{(1)}} \ell(\gamma, M, T).$$

Hint: Show first that as $\gamma \uparrow X_{(1)}$,

$$M = M(\gamma) \sim \frac{1}{n} \log(X_{(1)} - \gamma), \qquad \text{and } T = T(\gamma) \sim \frac{n-1}{n^2} \log^2(X_{(1)} - \gamma),$$

where the notation $f(\gamma) \sim g(\gamma)$ means $f(\gamma)/g(\gamma) \to 1$ as $\gamma \uparrow X_{(1)}$.

Solution. The density of X_1 is

$$\frac{d}{dx}\mathbf{P}(\gamma + e^Z \le x) = \frac{d}{dx}\mathbf{P}(Z \le \log(x-\gamma)) = \frac{d}{dx}\int_{-\infty}^{\log(x-\gamma-)} e^{-(y-\mu)^2/2\sigma^2} dy/\sqrt{2\sigma^2\pi}$$
$$= \frac{1}{x-\gamma} e^{-[\log(x-\gamma)-\mu]^2/2\sigma^2} \frac{1}{\sigma\sqrt{2\pi}}.$$

So, the log likelihood for n samples is

$$\log \prod_{i=1}^n \frac{1}{X_i-\gamma} e^{-[\log(X_i-\gamma)-\mu]^2/2\sigma^2} \frac{1}{\sigma\sqrt{2\pi}} = \sum_{i=1}^n -\log(X_i-\gamma) - \frac{(\log(X_i-\gamma)-\mu)^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi}).$$

The μ derivative is

$$\sum_{i=1}^n \frac{\log(X_i-\gamma)-\mu}{\sigma^2}.$$

The σ derivative is

$$\sum_{i=1}^n \sigma^{-3}(\log(X_i-\gamma)-\mu)^2 - \frac{1}{\sigma}.$$

Solving for μ, σ when both derivatives are zero gives

$$n\sigma^2 = \sum_{i=1}^n (\log(X_i-\gamma)-\mu)^2, \qquad \sum_{i=1}^n \log(X_i-\gamma) = n\mu,$$

$$\mu = \frac{1}{n}\sum_{i=1}^n \log(X_i-\gamma), \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^n \Big(\log(X_i-\gamma) - \frac{1}{n}\sum_{j=1}^n \log(X_j-\gamma)\Big)^2.$$

As argued in the notes, this is the unique global maximum of the likelihood function. As suggested in the hint, these formulas imply that as $\gamma \uparrow X_{(1)}$,

$$M = M(\gamma) \sim \frac{1}{n}\log(X_{(1)}-\gamma), \quad \text{and } T = T(\gamma) \sim \frac{n-1}{n^2}\log^2(X_{(1)}-\gamma),$$

We then examine the asymptotic behavior of ℓ as $\gamma \uparrow X_{(1)}$. The first term in the definition of ℓ behaves like $-\log(X_{(1)}-\gamma)$. The second term is of constant order times n. The last term behaves like $\log n - 2\log\log(X_{(1)}-\gamma)$. So, the first term dominates the other two (for fixed n as $\gamma \uparrow X_{(1)}$), and the log likelihood converges to plus infinity. In particular, $\mathbf{E}\sup_{(\gamma',\mu',\sigma^{2\prime})\in\Theta}\left|\log f_{(\gamma',\mu',\sigma^{2\prime})}(X_1,\dots,X_n)\right| = \infty$, i.e. this assumption of the consistency theorem is violated. Also, an MLE does not exist (since no maximum of ℓ exists), so we cannot assert that an MLE converges in probability to $(\gamma, \mu\sigma^2)$. □

Exercise 7.3 (**Least Squares/ Ridge Regression, Part 2**)**.** Suppose $w \in \mathbb{R}^k$ is an unknown vector, and for all $1 \le i \le n$, there are known vectors $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^k$. Our observed data are $X_1, \dots, X_n \in \mathbb{R}$. In linear least squares regression, we try to determine the best linear relationship between the vectors $x^{(1)}, \dots, x^{(n)}$ and the data $X_1, \dots, X_n$. Let A be the $n \times k$ matrix so that the i^{th} row of A is the row vector $x^{(i)}$. Assume that $k \le n$ and the matrix A has full rank. In a previous homework, we found $w \in \mathbb{R}^k$ that minimizes the quantity

$$\sum_{i=1}^n (X_i - \langle x^{(i)}, w\rangle)^2$$

We also interpreted the minimal w as an estimator. In some cases, the estimator for w could have large variance, which is undesirable. To deal with this issue, let $c > 0$ and consider the quantity

$$\sum_{i=1}^{n}(X_i - \langle x^{(i)}, w\rangle)^2 + c\,\|w\|^2\,. \qquad (*)$$

Find the value of $w \in \mathbb{R}^k$ that minimizes this quantity.

The term $\|w\|^2$ penalizes w from having large entries. By Lagrange Multipliers, a critical point w of the constrained minimization problem

$$\text{minimize } \sum_{i=1}^{n}(X_i - \langle x^{(i)}, w\rangle)^2 \quad \text{subject to } \|w\|^2 \le 1$$

is equivalent to the existence of a $c \in \mathbb{R}$ such that w is a critical point of $(*)$.

The L_2 penalization term in $(*)$ sometimes still allows w to have large entries. So, let $c > 0$ and consider the quantity

$$\sum_{i=1}^{n}(X_i - \langle x^{(i)}, w\rangle)^2 + c\sum_{i=1}^{n}|w_i|\,. \qquad (**)$$

Prove that there exists a $w \in \mathbb{R}^k$ that minimizes this quantity (this w is known as the LASSO, or least absolute shrinkage and selection operator). The L_1 penalization term in $(**)$ is better at penalizing large entries of w (a similar observation applies in the compressed sensing literature). Unfortunately, there is no closed form solution to $(**)$ in general. The constrained minimization problem

$$\text{minimize } \sum_{i=1}^{n}(X_i - \langle x^{(i)}, w\rangle)^2 \quad \text{subject to } sum_{i=1}^{n}|w_i| \le 1$$

is morally equivalent to $(**)$, but technically Lagrange Multipliers does not apply since the constraint is not differentiable everywhere.

Solution. Define $f(w) := \sum_{i=1}^{n}(X_i - \langle x^{(i)}, w\rangle)^2 = \|X - Aw\|^2 + \|w\|^2$. From Exercise 6.5, $\nabla f = A^T(Ax - b) + cx$, so if $\nabla f = 0$, $(A^TA + cI)x = A^Tb$the global minimum of f is $w = (A^TA + cI)^{-1}A^TX$.

$$X_i := \langle x^{(i)}, w\rangle + Z_i, \qquad \forall\, 1 \le i \le n.$$

□

Exercise 7.4 (Second Order Jackknife). Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}^n$ be i.i.d random variables so that X_1 has distribution $f_\theta \colon \mathbb{R}^n \to [0, \infty)$, $\theta \in \Theta$. Let $Y_1, Y_2, \ldots$ be a sequence of estimators for θ so that for any $n \ge 1$, $Y_n = t_n(X_1, \ldots, X_n)$ for some $t_n \colon \mathbb{R}^{n^2} \to \Theta$. For any $n \ge 1$, define the **second order jackknife estimator** of Y_n to be

$$\begin{aligned} Z_n := \frac{n^2}{2}Y_n - \frac{(n-1)^2}{n}\sum_{i=1}^{n} t_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\ + \frac{(n-2)^2}{n(n-1)}\sum_{1\le i<j\le n} t_{n-2}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n). \end{aligned}$$

Assume that $Y_1, Y_2, \ldots$ are asymptotically unbiased, so that there exists $a, b, c, d \in \mathbb{R}$ such that

$$\mathbf{E}Y_n = \theta + a/n + b/n^2 + \frac{c}{n^3} + \frac{d}{n^4} + O(1/n^5), \qquad \forall\, n \geq 1. \qquad (*)$$

Show that

$$\mathbf{E}Z_n = \theta + O(1/n^3).$$

And if $c = d = 0$ and the $O(1/n^5)$ term is zero in $(*)$, then Z_n is unbiased.

$$\begin{aligned}
\mathbf{E}Z_n &\overset{(*)}{=} n^2\theta/2 + na/2 + b/2 + c/(2n) + d/(2n^2) + O(1/n^5) \\
&\quad - (n-1)^2\Big(\theta + a/(n-1) + b/(n-1)^2 + \frac{c}{(n-1)^3} + \frac{d}{(n-1)^4} + O(1/n^5)\Big) \\
&\quad + (n-2)^2/2\Big(\theta + a/(n-2) + b/(n-2)^2 + \frac{c}{(n-2)^3} + \frac{d}{(n-2)^4} + O(1/n^5)\Big) \\
&= n^2\theta/2 + na/2 + b/2 + c/(2n) + d/(2n^2) + O(1/n^3) \\
&\quad - (n-1)^2\theta - a(n-1) - b - c/(n-1) - d/(n-1)^2 + O(1/n^3) \\
&\quad + (n-2)^2\theta/2 + a(n-2)/2 + b/2 + c/2(n-2) + d/2(n-2)^2 + O(1/n^3) \\
&= \theta + c\Big(\frac{1}{2n} - \frac{1}{n-1} + \frac{1}{2(n-2)}\Big) + d\Big(\frac{1}{2n^2} - \frac{1}{(n-1)^2} + \frac{1}{2(n-2)^2}\Big) + O(1/n^3) \\
&= \theta + c\Big(\frac{(n-1)(n-2) - 2n(n-2) + n(n-1)}{2n(n-1)(n-2)}\Big) \\
&\quad + d\Big(\frac{(n-1)^2(n-2)^2 - 2n^2(n-2)^2 + n^2(n-1)^2}{2n^2(n-1)^2(n-2)^2}\Big) + O(1/n^3) \\
&= \theta + c\Big(\frac{-(n+1)(n-2) + n(n-1)}{2n(n-1)(n-2)}\Big) \\
&\quad + d\Big(\frac{-2n^2(n-2)^2 + [n^2 + (n-2)^2](n-1)^2}{2n^2(n-1)^2(n-2)^2}\Big) + O(1/n^3)
\end{aligned}$$

Exercise 7.5. Do Question 1 on the Fall 2011 qualifying exam here:
https://dornsife.usc.edu/mgsa/statistics-a/

Solution. We have $f_{Y_1}(y) = \frac{2y}{\theta^2} 1_{y \in [0,\theta]}$ for all $y \in \mathbb{R}$, for all $\theta > 0$, where $\theta > 0$ is an unknown parameter that we would like to estimate. Denote $Y = (Y_1, \ldots, Y_n)$. Both $I_Y(\theta)$ and $I_{Y_1}(\theta)$ are not well-defined, since the region where the PDF of Y_1 is nonzero is a function of θ. So, we could state a Cramér-Rao inequality here, but since the Fisher information is not well-defined, the Cramér-Rao inequality is vacuous in this case (the inequality does not apply since the Fisher information is not well-defined).

When $n = 1$, note that $\mathbf{E}Y_1 = \theta^{-2} \int_0^\theta 2y^2 dy = (2/3)\theta$, so $\mathbf{E}(3/2)Y_1 = \theta$, i.e. $(3/2)Y_1$ is an unbiased estimator of θ, which is also UMVU for θ by Lehmann-Scheffé when $n = 1$ since Y_1 is sufficient for θ. So, at least when $n = 1$, we have a lower bound for unbiased estimators Z of θ of the form

$$\mathrm{Var}(Z) \geq \mathrm{Var}(Y_1) = \theta^2/18.$$

And if we try to naïvely compute the Fisher information of Y_1, we will get a different inequality than this one when $n = 1$. □

USC Department of Mathematics, Los Angeles, CA
E-mail address: stevenmheilman@gmail.com