# SI140 Probability and Mathematical Statistics: (Fall 2022) Final Exam Solutions

## Problem 1

(10 points) Basic Concepts.

(a) (5 points) Please describe the differences and connections between *Probability* and *Statistics*. Then explain why we say conditioning is the soul of our course.

(b) (5 points) Please describe the pros and cons of *Bayesian statistical inference* and *Classical statistical inference*. Then explain why conjugate priors are important for Bayesian statistical inference.

### Solution

(a)
- *Difference*: Probability deals with the prediction of future events. On the other hand, statistics are used to analyze the frequency of past events. More specifically, probability is a measure of the likelihood of an event to occur. In probability theory we consider some underlying process which has some randomness or uncertainty modeled by random variables, and we figure out what happens. In statistics we observe something that has happened, and try to figure out what underlying process would explain those observations.
- *Connection*: Probability is used to answer questions in the category of Statistics. Most statistical models are based on experiments and hypotheses, and probability is integrated into the theory, to explain the scenarios better with framework and theoretical analysis.

As mentioned in the above, both probability and statistics in the core need to analyze "what have happened" or focus on "some specific randomness". This leads to the concept of conditioning and conditional probability. Conditional probability is essential for scientific, medical, and legal reasoning, since it shows how to incorporate evidence into our understanding of the world in a logical, coherent manner. In fact, a useful perspective is that all probabilities are conditional; whether or not it' s written explicitly, there is always background knowledge (or assumptions) built into every probability.

(b) *Pros of Bayesian statistical inference*:

- It provides a natural and principled way of combining prior information with data, within a solid decision theoretical framework. You can incorporate past information about a parameter and form a prior distribution for future analysis. When new observations become available, the previous posterior distribution can be used as a prior. All inferences logically follow from Bayes' theorem.
- It can calculate the probability that a hypothesis is true, which is generally what the researchers actually want to know.

*Cons of Bayesian statistical inference*:

- It requires you to know or construct a prior, but it does not tell you how to select the prior. If you do not proceed with caution, you can generate misleading results.
- It can produce posterior distributions that are heavily influenced by the priors.

- It may be computationally intensive due to integration over many parameters.

*Pros of classical statistical inference*:

- It tends to be less computationally intensive than the Bayesian statistical inference.

*Cons of classical statistical inference*:

- It does not take into account priors.
- It has only one well-defined hypothesis.

As mentioned above, both the selection of probabilistic prior and the computation to get the posterior are in general hard in practice. The existence of conjugate priors smoothen these processes of selection and computation via a set of well-behaved priors and posteriors, where the priors have good physical meanings for the parameters to estimate (*e.g.*, Beta prior is used for $[0, 1]$ random variables and Gamma prior is used for $[0, \infty)$ random variables, *etc.*) and the posteriors are efficiently computed via conjugacy (*e.g.*, Beta-Binomial, Gamma-Poisson, Normal-Normal, *etc.*).

# Problem 2

(10 points) Let $X$ and $Y$ be two continuous random variables with joint PDF

$$f_{X,Y}(x,y) = \begin{cases} x + cy^2 & \text{if } 0 \le x \le 1, 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) (5 points) Find the value of constant $c$.

(b) (5 points) Find the joint probability $P(0 \le X \le 1/2,\ 0 \le Y \le 1/2)$.

## Solution

(a) Since

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1,$$

we have

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy &= \int_0^1\int_0^1 x + cy^2 dxdy \\
&= \int_0^1 \left(\frac{1}{2}x^2 + cy^2x\right)\Big|_{x=0}^{x=1} dy \\
&= \int_0^1 \frac{1}{2} + cy^2 dy \\
&= \frac{1}{2}y + \frac{1}{3}cy^3\Big|_{y=0}^{y=1} \\
&= \frac{1}{2} + \frac{1}{3}c.
\end{aligned}$$

Therefore, we get $c = \frac{3}{2}$.

(b)

$$\begin{aligned}
P\left(0 \le X \le \frac{1}{2}, 0 \le Y \le \frac{1}{2}\right) &= \int_0^{\frac{1}{2}}\int_0^{\frac{1}{2}} \left(x + \frac{3}{2}y^2\right) dxdy \\
&= \int_0^{\frac{1}{2}} \left(\frac{1}{2}x^2 + \frac{3}{2}y^2x\right)\Big|_{x=0}^{x=\frac{1}{2}} dy \\
&= \int_0^{\frac{1}{2}} \left(\frac{1}{8} + \frac{3}{4}y^2\right) dy \\
&= \frac{1}{8}y + \frac{1}{4}y^3\Big|_{y=0}^{y=\frac{1}{2}} \\
&= \frac{3}{32}.
\end{aligned}$$

# Problem 3

(10 points) Let $X$ and $Y$ be two continuous random variables with joint PDF

$$f_{X,Y}(x,y) = \begin{cases} 6xy & \text{if } 0 \le x \le 1, 0 \le y \le \sqrt{x}, \\ 0 & \text{otherwise.} \end{cases}$$

(a) (5 points) Find the marginal distributions of $X$ and $Y$. Are $X$ and $Y$ independent?

(b) (5 points) Find $E[X|Y = y]$ and $\text{Var}[X|Y = y]$ for $0 \le y \le 1$.

## Solution

(a) The supports of $X$ and $Y$ are both $[0, 1]$. In this way, we have

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \\ &= \int_0^{\sqrt{x}} 6xydy \\ &= 3xy^2\Big|_{y=0}^{y=\sqrt{x}} \\ &= 3x^2, \end{aligned}$$

and

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx \\ &= \int_{y^2}^{1} 6xydx \\ &= 3yx^2\Big|_{x=y^2}^{x=1} \\ &= 3y - 3y^5. \end{aligned}$$

Therefore,

$$f_X(x) = \begin{cases} 3x^2 & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$f_Y(y) = \begin{cases} 3y - 3y^5 & \text{if } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Since $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$, $X$ and $Y$ are not independent.

(b) Since

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx,$$

to calculate $E[X|Y = y]$, we need to first calculate $f_{X|Y}(x|y)$.
If $y^2 \le x \le 1$,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{2x}{1 - y^4}.$$

In this way,

$$f_{X|Y}(x|y) = \begin{cases} \frac{2x}{1-y^4} & \text{if } y^2 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} E[X|Y=y] &= \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx \\ &= \int_{y^2}^{1} x \frac{2x}{1-y^4} dx \\ &= \frac{2}{3(1-y^4)} x^3 \Big|_{x=y^2}^{x=1} \\ &= \frac{2(1-y^6)}{3(1-y^4)}. \end{aligned}$$

Since

$$\mathrm{Var}[X|Y=y] = E[X^2|Y=y] - (E[X|Y=y])^2,$$

to calculate $\mathrm{Var}[X|Y=y]$, we need to first calculate $E[X^2|Y=y]$.

Since

$$\begin{aligned} E[X^2|Y=y] &= \int_{-\infty}^{\infty} x^2 f_{X|Y}(x|y)dx \\ &= \int_{y^2}^{1} x^2 \frac{2x}{1-y^4} dx \\ &= \frac{1}{2(1-y^4)} x^4 \Big|_{x=y^2}^{x=1} \\ &= \frac{1-y^8}{2(1-y^4)} \\ &= \frac{1+y^4}{2}, \end{aligned}$$

we have,

$$\begin{aligned} \mathrm{Var}[X|Y=y] &= E[X^2|Y=y] - (E[X|Y=y])^2 \\ &= \frac{1+y^4}{2} - \left(\frac{2(1-y^6)}{3(1-y^4)}\right)^2. \end{aligned}$$

# Problem 4

(15 points) Given a coin with the probability $p$ of landing heads, where $p$ is *unknown* and we need to estimate its value through data. In our data collection model, we have $n$ independent tosses, result of each toss is either Head or Tail. Let $X$ denote the number of heads in the total $n$ tosses. Now we conduct experiments to collect data and find out that $X = k$. Then we need to find $\widehat{p}$, the estimation of $p$.

(a) (5 points) Assume $p$ is an unknown constant. Find $\widehat{p}$ through the MLE (Maximum Likelihood Estimation) rule.

(b) (5 points) Assume $p$ is a random variable with a prior distribution Beta$(a, b)$, where $a$ and $b$ are known constants. Find $\widehat{p}$ through the MAP (Maximum a Posterior Probability) rule.

(c) (5 points) Assume $p$ is a random variable with a prior distribution Beta$(a, b)$, where $a$ and $b$ are known constants. Find $\widehat{p}$ through the MMSE (Minimum Mean Square Estimate) rule.

## Solution

(a) Let $X_i$ be the outcome of $i$th toss. Then $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \text{Bern}(p)$, where $p$ is an unknown constant. The PMF of $X_i$ can be formulated as

$$P_{X_i}(x_i; p) = p^{x_i}(1-p)^{1-x_i}$$

since

$$p^{x_i}(1-p)^{1-x_i} = \begin{cases} p, & \text{if } x_i = 1, \\ 1-p, & \text{if } x_i = 0. \end{cases}$$

The likelihood function is

$$P_X(x; p) = \prod_{i=1}^{n} P_{X_i}(x_i; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^k(1-p)^{n-k}$$

So the corresponding log-likelihood function is

$$g(p) = \log P_X(x; p) = \log p^{S_n}(1-p)^{n-S_n} = S_n \log p + (n - S_n)\log(1-p)$$

Now we try to find $\widehat{p}_{\text{MLE}}$ such that $g(\widehat{p}_{\text{MLE}})$ is the maximum of $g(p)$. We have

$$\begin{aligned} g'(p) = &\ \frac{k}{p} - \frac{n-k}{1-p}, \\ g''(p) = &\ -\frac{k}{p^2} - \frac{n-k}{(1-p)^2} \leq 0 \end{aligned}$$

Let $g'(p) = 0$, we can get $p = \frac{k}{n}$. Since $g''(p) \leq 0$, then we know that

$$\widehat{p}_{\text{MLE}} = \frac{k}{n}$$

is the MLE of $p$.

(b) We know the posterior distribution

$$f_{p|X=k} \propto p^{a+k-1}(1-p)^{b+n-k-1}, \ p \in (0, 1)$$

by Beta-Binomial conjugacy. Then the MAP estimator

$$\widehat{p}_{\text{MAP}} = \arg\max_p f_{\theta|X=k} = \arg\max_p \log(f_{p|X=k})$$

since logarithmic function is monotonically increasing. Let

$$g(p) = \log(f_{p|X=k}) = (a+k-1)\log p + (b+n-k-1)\log(1-p),$$

where we don't consider the proportional constant. Our goal is to find $p^*$ such that $g(p^*)$ is maximum of $g(p)$. We have

$$\begin{aligned} g'(p) = & \ \frac{a+k-1}{p} - \frac{b+n-k-1}{1-p}, \\ g''(p) = & \ -\frac{a+k+1}{p^2} - \frac{b+n-k-1}{(1-p)^2} < 0. \end{aligned}$$

Let $g'(p^*) = 0$. We have $p^* = \frac{a+k-1}{a+b+n-2}$, and $g(p^*)$ is maximum of $g(p)$ since $g''(p) < 0$.

Then we can get the MAP estimate

$$\widehat{p}_{\text{MAP}} = \arg\max_p f_{p|X=k} = \arg\max_p \log(f_{\theta|X=k}) = p^* = \frac{a+k-1}{a+b+n-2}.$$

(c) Since the prior distribution is $p \sim \text{Beta}(a,b)$ and the conditional distribution of $X$ given $p$ is $X|p \sim \text{Bin}(n,p)$, we can get the posterior distribution

$$\Theta|X = k \sim \text{Beta}(a+k, b+n-k)$$

by Beta-Binomial conjugacy. It follows that

$$E(p|X=k) = \frac{a+k}{a+b+n},$$

so the MMSE estimation of $\Theta$ is

$$\widehat{p}_{\text{MMSE}} = E(p|X=k) = \frac{a+k}{a+b+n}.$$

# Problem 5

(10 points) We know that the MMSE of $Y$ given $X$ is given by $g(X) = E[Y|X]$. We also know that the LLSE (Linear Least Square Estimate) of $Y$ given $X$, denoted by $L[Y|X]$, is shown as follows:

$$L[Y|X] = E(Y) + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}\big(X - E(X)\big).$$

Now we wish to estimate the probability of landing heads, denoted by $\theta$, of a biased coin. We model $\theta$ as the value of a random variable $\Theta$ with a known prior with PDF $f_\Theta \sim \text{Unif}(0, 1)$. We consider $n$ independent tosses and let $X$ be the number of heads observed.

(a) (5 points) Show that $E\big[(\Theta - E[\Theta|X])h(X)\big] = 0$ for any real function $h(\cdot)$.

(b) (5 points) Find the MMSE $E[\Theta|X]$ and the LLSE $L[\Theta|X]$.
(Eve's law: $\text{Var}(Y) = E\big[\text{Var}(Y|X)\big] + \text{Var}\big[E(Y|X)\big]$.)

### Solution

(a) We can see that

$$\begin{aligned}&E[(\Theta - E[\Theta|X])h(X)] = E[\Theta h(X) - E[\Theta|X]h(X)]\\ =&E[\Theta h(X)] - E[E[\Theta|X]h(X)] = E[\Theta h(X)] - E[E[\Theta h(X)|X]].\end{aligned}$$

By Adam's law, we have $E[E[\Theta h(X)|X]] = E[\Theta h(X)]$. Then it follows that

$$E[(\Theta - E[\Theta|X])h(X)] = E[\Theta h(X)] - E[\Theta h(X)] = 0.$$

(b) Firstly, we find the MMSE $E[\Theta|X]$. Since the prior distribution

$$\Theta \sim \text{Unif}(0, 1) = \text{Beta}(1, 1)$$

and given $\Theta = \theta$, the conditional distribution of $X$

$$X|\Theta = \theta \sim \text{Bin}(n, \theta),$$

by Beta-Binomial conjugacy we can get the posterior distribution

$$\Theta|X = k \sim \text{Beta}(k + 1, n - k + 1).$$

It follows that

$$E(\Theta|X = k) = \frac{k + 1}{(k + 1) + (n - k + 1)} = \frac{k + 1}{n + 2},$$

so the MMSE estimator of $\Theta$ is

$$E(\Theta|X) = \frac{X + 1}{n + 2}.$$

Then we find the LLSE $L[\Theta|X]$. Since

$$L[\Theta|X] = E(\Theta) + \frac{\text{Cov}(\Theta, X)}{\text{Var}(X)}(X - E(X)),$$

we just need to calculate these statistics. By $\Theta \sim \text{Unif}(0, 1)$ we have

$$E(\Theta) = \frac{1}{2}, \quad \text{Var}(\Theta) = \frac{1}{12}, \quad E(\Theta^2) = \frac{1}{3}.$$

Problem 5 continued on next page...

We know $X|\Theta = \theta \sim \text{Bin}(n, \theta)$, so

$$E(X|\Theta = \theta) = n\theta \implies E(X|\Theta) = n\Theta,$$

$$\text{Var}(X|\Theta = \theta) = n\theta(1-\theta) \implies \text{Var}(X|\Theta) = n\Theta(1-\Theta).$$

By Adam's law, we have

$$E(X) = E[E(X|\Theta)] = E(n\Theta) = n \cdot E(\Theta) = \frac{n}{2}.$$

By Eve's law, we have

$$\begin{aligned}
\text{Var}(X) = &\ E(\text{Var}(X|\Theta)) + \text{Var}(E(X|\Theta)) \\
= &\ E(n\Theta(1-\Theta)) + \text{Var}(n\Theta) \\
= &\ n(E(\Theta) - E(\Theta^2)) + n^2\text{Var}(\Theta) \\
= &\ n\left(\frac{1}{2} - \frac{1}{3}\right) + n^2 \cdot \frac{1}{12} \\
= &\ \frac{n}{12}(n+2).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\text{Cov}(X, \Theta) = &\ E(\Theta X) - E(\Theta)E(X) \\
= &\ E[E(\Theta X|\Theta)] - E(\Theta)E(X) \\
= &\ E(\Theta E(X|\Theta)) - E(\Theta)E(X) \\
= &\ E(\Theta \cdot n\Theta) - E(\Theta)E(X) \\
= &\ nE(\Theta^2) - E(\Theta)E(X) \\
= &\ n \cdot \frac{1}{3} - \frac{1}{2} \cdot \frac{n}{2} \\
= &\ \frac{n}{12}.
\end{aligned}$$

It follows that

$$\begin{aligned}
L[\Theta|X] = &\ E(\Theta) + \frac{\text{Cov}(\Theta, X)}{\text{Var}(X)}(X - E(X)) \\
= &\ \frac{1}{2} + \frac{\frac{n}{12}}{\frac{n}{12}(n+2)}\left(X - \frac{n}{2}\right) \\
= &\ \frac{1}{2} + \frac{1}{n+2}\left(X - \frac{n}{2}\right) \\
= &\ \frac{X+1}{n+2}.
\end{aligned}$$

We can find in this case $E[\Theta|X] = L[\Theta|X]$.

# Problem 6

(10 points) Laplace's law of succession says that if $X_1, X_2, \dots, X_{n+1}$ are conditionally independent Bern($p$) r.v.s given $p$, but $p$ is given a Unif$(0,1)$ prior to reflect ignorance about its value, then

$$P\left(X_{n+1} = 1 \mid X_1 + \cdots + X_n = k\right) = \frac{k+1}{n+2}.$$

As an example, Laplace discussed the problem of predicting whether the sun will rise tomorrow, given that the sun did rise every time for all $n$ days of recorded history; the above formula then gives $(n+1)/(n+2)$ as the probability of the sun rising tomorrow (of course, assuming independent trials with $p$ unchanging over time may be a very unreasonable model for the sunrise problem).

(a) (5 points) Find the posterior distribution of $p$ given $X_1 = x_1, \dots, X_n = x_n$, and show that it only depends on the sum of the $x_j, j \in \{1, \dots, n\}$.

(b) (5 points) Prove Laplace's law of succession, using a form of LOTP to find

$$P\left(X_{n+1} = 1 \mid X_1 + \cdots + X_n = k\right)$$

by conditioning on $p$.

## Solution

(a) Let $m = \sum_{i=1}^n x_i$ (so $m$ is the number of successes) and $\mathbf{X} = (X_1, \dots, X_n)$. We are required to find $f_{p|\mathbf{X}}(p \mid \mathbf{x})$ where $p \in (0,1)$ and $\mathbf{x} = (x_1, \dots, x_n)$. Using Bayes' rule, we have that

$$f_{p|\mathbf{X}}(p \mid \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x} \mid p) f(p)}{P(\mathbf{X} = \mathbf{x})}.$$

Observe that for arbitrary $i = 1, \dots, n$ we can write that probability function of $X_i$ given $p$ is

$$P(X_i = x_i \mid p) = p^{x_i}(1-p)^{1-x_i}.$$

Using the fact that these variables $X_i$ are independent given $p$, we have that

$$P(\mathbf{X} = \mathbf{x} \mid p) = P(X_1 = x_1 \dots X_n = x_n \mid p) = \prod_{i=1}^n P(X_i = x_i \mid p) = p^m (1-p)^{n-m}.$$

In order to determine $P(\mathbf{X} = \mathbf{x})$ we are going to use LOTP conditioning on $p$. We have that

$$P(\mathbf{X} = \mathbf{x}) = \int_0^1 P(\mathbf{X} = \mathbf{x} \mid p) f(p) dp = \int_0^1 p^m (1-p)^{n-m} dp = \beta(m+1, n-m+1)$$

where $\beta$ is Beta function. Finally, we have that

$$f_{p|\mathbf{X}}(p \mid \mathbf{x}) = \frac{p^m (1-p)^{n-m}}{\beta(m+1, n-m+1)}.$$

Observe that it does not depend on random vector $\mathbf{X}$, it only depends on statistics $M$ which is the sum of random variables $X_1, \dots, X_n$.

(b) From part (a), $p \mid \sum_{i=1}^n X_i = k \sim \text{Beta}(k+1, n-k+1)$. Conditioning on $p$, by LOTP

$$\begin{aligned}
P\left(X_{n+1} = 1 \mid \sum_{i=1}^n X_i = k\right) &= \int_0^1 P\left(X_{n+1} = 1 \mid p, \sum_{i=1}^n X_i = k\right) f\left(p \mid \sum_{i=1}^n X_i = k\right) \mathrm{d}p \\
&= \int_0^1 p \cdot \frac{p^k (1-p)^{n-k}}{\beta(k+1, n-k+1)} \mathrm{d}p \\
&= \mathbb{E}\left(p \mid \sum_{i=1}^n X_i = k\right) \\
&= \frac{k+1}{(k+1) + (n-k+1)} = \frac{k+1}{n+2}.
\end{aligned}$$

# Problem 7

(15 points) A handy rule of thumb in statistics and life is as follows: *Conditioning often makes things better.* This problem explores how the above rule of thumb applies to estimating unknown parameters.
Let $\theta$ be an unknown parameter that we wish to estimate based on data $X_1, \dots, X_n$ (these are r.v.s before being observed, and then after the experiment they "crystallize" into data).
In this problem, $\theta$ is viewed as an unknown constant, and is *not* treated as an r.v. as in the Bayesian statistical inference. Let $T_1$ be an estimator for $\theta$ (this means that $T_1$ is a function of $X_1, \dots, X_n$ which is being used to estimate $\theta$).
A strategy for improving $T_1$ (in some problems) is as follows. Suppose that we have an r.v. $R$ such that $T_2 = E(T_1|R)$ is a function of $X_1, \dots, X_n$ (in general, $E(T_1|R)$ might involve unknowns such as $\theta$ but then it couldn't be used as an estimator). Also suppose that $P(T_1 = T_2) < 1$, and that $E(T_1^2)$ is finite.

(a) (5 points) Use Jensen's inequality and Adam's Law to show that $T_2$ is better than $T_1$ in the sense that the mean squared error is less, *i.e.*,

$$E\big[(T_2 - \theta)^2\big] < E\big[(T_1 - \theta)^2\big].$$

(b) (5 points) The bias of an estimator $T$ for $\theta$ is defined to be $b(T) = E(T) - \theta$. An important identity in statistics, a form of the *bias-variance* trade-off, is that the mean squared error equals the variance plus the squared bias:

$$E\big[(T - \theta)^2\big] = \mathrm{Var}(T) + \big(b(T)\big)^2.$$

Use this identity and Eve's law to give an alternative proof of the result from (a).

(c) (5 points) Now suppose that $X_1, \dots, X_n$ are *i.i.d.* with mean $\theta$, and consider the special case $T_1 = X_1, R = \sum_{j=1}^{n} X_j$. Find $T_2$ in a simplified form, and check that it has a lower mean squared error than $T_1$ for $n \geq 2$. Also, explain what happens to $T_1$ and $T_2$ as $n \to \infty$.

## Solution

(a) Consider right side of the inequality that has to be proved. Apply Adam's law conditioning on $\theta$. We have that

$$E(T_1 - \theta)^2 = E(T_1^2) - 2\theta E(T_1) + \theta^2 = E(E(T_1^2 \mid R)) - 2\theta E(E(T_1 \mid R)) + \theta^2.$$

Apply Jensen's inequality on function $t \mapsto t^2$ on the first term in the expression above, *i.e.*, we have that $E(E(T_1^2 \mid R)) > E(E(T_1 \mid R)^2)$. We can write strict inequality since the quadratic function is strictly convex. Plug the definition of $T_2 = E(T_1 \mid R)$ so we get that the expression above is strictly greater of

$$E(T_2^2) - 2\theta E(T_2) + \theta^2 = E(T_2 - \theta)^2$$

Hence, we have proved that

$$E(T_1 - \theta)^2 > E(T_2 - \theta)^2$$

(b) Plugging $T_1$ in the given relation and using the definition of bias of estimator, we have that

$$E(T_1 - \theta)^2 = \mathrm{Var}(T_1) + (E(T_1) - \theta)^2 \tag{1}$$

Use Eve's law on the variance to get that

$$\mathrm{Var}(T_1) = \mathrm{Var}(E(T_1 \mid R)) + E(\mathrm{Var}(T_1 \mid R))$$

But, we have that $E(T_1 \mid R) = T_2$ and that $\mathrm{Var}(T_1 \mid R)$ is random variable that values at strictly positive real numbers. (In general, the variance is non-negative, but it is equal to zero if and only if the random variable is equal to constant with the probability one. But here, there is no such a trivial case). So, the expectation of $\mathrm{Var}(T_1 \mid R)$ is strictly greater than zero. Hence, we have that

$$\mathrm{Var}(T_1) = \mathrm{Var}(E(T_1 \mid R)) + E(\mathrm{Var}(T_1 \mid R)) = \mathrm{Var}(T_2) + E(\mathrm{Var}(T_1 \mid R)) > \mathrm{Var}(T_2)$$

Use the fact that $E(T_1) = E(E(T_1 \mid R)) = E(T_2)$ and plug all these information in (1) to obtain that

$$\mathrm{Var}(T_1) + (E(T_1) - \theta)^2 > \mathrm{Var}(T_2) + (E(T_2) - \theta)^2 = E(T_2 - \theta)^2$$

Hence we have proved that

$$E(T_1 - \theta)^2 > E(T_2 - \theta)^2$$

(c) Let's find what in fact is $T_2$. We have that

$$T_2 = E(T_1 \mid R) = E(X_1 \mid \sum_j X_j)$$

But, we have that

$$R = \sum_j X_j = E(\sum_i X_i \mid \sum_j X_j) = \sum_i E(X_i \mid \sum_j X_j)$$

Since variables $X_i$ are all equally distributed and independent, we have the symmetry and that

$$\sum_j X_j = n \cdot T_2 \Rightarrow T_2 = \frac{\sum_j X_j}{n}$$

Now we have that

$$E(T_2 - \theta)^2 = E\left(\frac{\sum_j X_j}{n} - \theta\right)^2$$

but since the expectation of $T_2$ is equal to $\theta$, the expression above is in fact

$$\mathrm{Var}\left(\frac{\sum_j X_j}{n}\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_j X_j\right) < \frac{1}{n^2} n \mathrm{Var}(T_1) < \mathrm{Var}(T_1) = E(T_1 - \theta)^2$$

so we have proved that

$$E(T_2 - \theta)^2 < E(T_1 - \theta)^2$$

Finally, if $n \to \infty$, variable $T_1 = X_1$ remains the same because it does not depend on $n$. On the other hand, using the law of the large numbers, we have that

$$\lim_{n\to\infty} T_2 = \lim_{n\to\infty} \frac{\sum_j X_j}{n} = E(X_1) = \theta$$

with the probability one.

# Problem 8

(10 points) Let $X$ and $Y$ be two independent random variables satisfying first success distribution FS($p$).

(a) (5 points) Define $Z_1 = X - Y$. Find the PMF of $Z_1$ and $E(Z_1)$.

(b) (5 points) Define $Z_2 = \frac{X}{Y}$. Find the PMF of $Z_2$ and $E(Z_2)$.

## Solution

Let $q = 1 - p$.

(a)

$$\begin{aligned}
P(Z_1 = k) &= P(X - Y = k) \\
&= P(X = Y + k) \\
&= \sum_{j=1}^{\infty} P(X = Y + k | Y = j) P(Y = j) \\
&= \sum_{j=1}^{\infty} P(X = j + k | Y = j) P(Y = j) \\
&= \sum_{j=1}^{\infty} P(X = j + k) P(Y = j), k \in \mathbb{Z}.
\end{aligned}$$

- Case 1: $k \geq 0$.

$$\begin{aligned}
P(Z_1 = k) &= \sum_{j=1}^{\infty} pq^{j+k-1} pq^{j-1} \\
&= p^2 q^{k-2} \sum_{j=1}^{\infty} q^{2j} \\
&= p^2 q^k \frac{1}{1-q^2} \\
&= p^2 q^{k-2} \frac{q^2}{1-q^2} \\
&= \frac{p(1-p)^k}{2-p}.
\end{aligned}$$

- Case 2: $k < 0$:

$$\begin{aligned}
P(Z_1 = k) &= \sum_{j=-k+1}^{\infty} pq^{j+k-1} pq^{j-1} \\
&= p^2 q^{k-2} \sum_{j=-k+1}^{\infty} q^{2j} \\
&= p^2 q^{k-2} \frac{q^{-2(k-1)}}{1-q^2} \\
&= p^2 q^k \frac{q^{-2k}}{1-q^2} \\
&= \frac{p(1-p)^{-k}}{(2-p)}.
\end{aligned}$$

Therefore, we have

$$P(Z_1 = k) = \begin{cases} \frac{p(1-p)^{|k|}}{2-p} & \text{if } k \in \mathbb{Z}, \\ 0 & \text{otherwise.} \end{cases}$$

Problem 8 continued on next page... 

For the expectation, we have

$$E[Z_1] = E[X] - E[Y] = 0.$$

where the series in the last equality is due to the mean of random variable Geom($p$).

(b) Let $m$, $n \in \mathbb{N}^+$, $(m, n) = 1$.

$$\begin{aligned}
P(Z_2 = \frac{n}{m}) = &\ P\left(\frac{X}{Y} = \frac{n}{m}\right) \\
= &\ \sum_{k=1}^{\infty} P(X = nk)P(Y = mk) \\
= &\ \sum_{k=1}^{\infty} pq^{nk-1}pq^{mk-1} \\
= &\ p^2 q^{-2} \sum_{k=1}^{\infty} q^{(n+m)k} \\
= &\ p^2 \frac{q^{m+n-2}}{1 - q^{m+n}} \\
= &\ \frac{p^2(1-p)^{m+n-2}}{1-(1-p)^{m+n}}.
\end{aligned}$$

$$\begin{aligned}
E(Z_2) = &\ \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} \frac{n}{m} P(X = n, Y = m) \\
= &\ \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} \frac{n}{m} p^2 q^{m+n-2} \\
= &\ \sum_{m=1}^{\infty} \frac{1}{m} p^2 q^{m-1} \sum_{n=1}^{\infty} nq^{n-1} \\
= &\ \sum_{m=1}^{\infty} \frac{1}{m} p^2 q^{m-1} \frac{1}{(1-q)^2} \\
= &\ \sum_{m=1}^{\infty} \frac{1}{m} q^{m-1} \\
= &\ \frac{1}{q} \sum_{m=1}^{\infty} \frac{q^m}{m} \\
= &\ \frac{1}{q}(-\ln(1-q)) \\
= &\ -\frac{1}{1-p} \ln p
\end{aligned}$$

# Problem 9

(10 points) A scientist makes two measurements $X, Y$, considered to be *i.i.d.* random variables.

(a) (5 points) If $X, Y \sim \mathcal{N}(0,1)$. Find the correlation between the larger and smaller of the values, *i.e.*, $\mathrm{Corr}\big(\max(X,Y), \min(X,Y)\big)$.

(b) (5 points) If $X, Y \sim \mathrm{Unif}(0,1)$. Find the correlation between the larger and smaller of the values, *i.e.*, $\mathrm{Corr}\big(\max(X,Y), \min(X,Y)\big)$.

## Solution

(a) Let $M = \max(X,Y), L = \min(X,Y)$. First, we compute $\mathrm{Cov}(M,L)$ using

$$\mathrm{Cov}(M,L) = E(ML) - E(M)E(L).$$

We have

$$E(ML) = E(XY) = E(X)E(Y) = 0.$$

By $X - Y \sim \mathcal{N}(0,2)$, we can get $E(|X-Y|) = \frac{2}{\sqrt{\pi}}$. Combine

$$E(M) + E(L) = E(M+L) = E(X+Y) = E(X) + E(Y) = 0$$

and

$$E(M) - E(L) = E(M-L) = E(|X-Y|) = \frac{2}{\sqrt{\pi}},$$

we get

$$E(M) = \frac{1}{\sqrt{\pi}},\ E(L) = -\frac{1}{\sqrt{\pi}}.$$

Thus we have

$$\mathrm{Cov}(M,L) = \frac{1}{\pi}.$$

Now we compute $\mathrm{Var}(M)$ and $\mathrm{Var}(L)$. By symmetry of Normal distribution, $(-X,-Y)$ has the same distribution as $(X,Y)$, so $\mathrm{Var}(M) = \mathrm{Var}(L)$. Since

$$E[(X-Y)^2] = \mathrm{Var}(X-Y) = 2$$

and

$$E[(X-Y)^2] = E[(M-L)^2] = E(M^2) + E(L^2) - 2E(ML) = E(M^2) + E(L^2),$$

we get $E(M^2) + E(L^2)$. Then we have

$$\mathrm{Var}(M) = \mathrm{Var}(L) = 1 - \left(-\frac{1}{\sqrt{\pi}}\right)^2 = 1 - \frac{1}{\pi}.$$

Finally, we can get

$$\mathrm{Corr}(M,L) = \frac{\mathrm{Cov}(M,L)}{\sqrt{\mathrm{Var}(M)\mathrm{Var}(L)}} = \frac{1}{\pi - 1}.$$

(b) Let $M = \max(X,Y), L = \min(X,Y)$, then we have

$$\begin{aligned} F_M(m) = &\ P(\max(X,Y) \le m) = P(X \le m, Y \le m) = m^2 \\ F_L(l) = &\ P(\min(X,Y) \le l) = 1 - P(X \ge l, Y \ge l) = 1 - (1-l)^2. \end{aligned}$$

Problem 9 continued on next page. . . 

Further, we have

$$\begin{aligned} E[M] &= \int_0^1 m \cdot 2m dm = \frac{2}{3}, \\ E[M^2] &= \int_0^1 m^2 \cdot 2m dm = \frac{1}{2}. \end{aligned}$$

Since

$$E[M] + E[L] = E[M + L] = E[X + Y] = E[X] + E[Y] = 1,$$

we have

$$E[L] = \frac{1}{3}.$$

Note that

$$\begin{aligned} \mathrm{Var}[L] &= \mathrm{Var}[1 - L] \\ &= \mathrm{Var}[1 - \min(X, Y)] \\ &= \mathrm{Var}[\max(1 - X, 1 - Y)] \\ &= \mathrm{Var}[\max(X, Y)] \\ &= \mathrm{Var}[M], \end{aligned}$$

we have

$$\mathrm{Var}[L] = \mathrm{Var}[M] = E[M^2] - E[M]^2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}.$$

Next, we compute

$$\mathrm{Cov}(M, L) = E[ML] - E[M]E[L] = E[XY] - E[M]E[L] = E[X]E[Y] - E[M]E[L] = \frac{1}{4} - \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{36}.$$

Therefore, we have

$$\mathrm{Corr}(M, L) = \frac{\mathrm{Cov}(M, L)}{\sqrt{\mathrm{Var}[M]\mathrm{Var}[L]}} = \frac{\frac{1}{36}}{\sqrt{\frac{1}{18} \cdot \frac{1}{18}}} = \frac{1}{2}.$$