Stat 110 Strategic Practice 10, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1 Conditional Expectation & Conditional Variance

1. Show that $E((Y - E(Y|X))^2|X) = E(Y^2|X) - (E(Y|X))^2$, so these two expressions for $\text{Var}(Y|X)$ agree.

2. Prove Eve's Law.

3. Let X and Y be random variables with finite variances, and let $W = Y - E(Y|X)$. This is a *residual*: the difference between the true value of Y and the predicted value of Y based on X.

   (a) Compute $E(W)$ and $E(W|X)$.

   (b) Compute $\text{Var}(W)$, for the case that $W|X \sim \mathcal{N}(0, X^2)$ with $X \sim \mathcal{N}(0, 1)$.

4. Emails arrive one at a time in an inbox. Let T_n be the time at which the nth email arrives (measured on a continuous scale from some starting point in time). Suppose that the waiting times between emails are i.i.d. Expo(λ), i.e., $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are i.i.d. Expo(λ).

   Each email is non-spam with probability p, and spam with probability $q = 1-p$ (independently of the other emails and of the waiting times). Let X be the time at which the first non-spam email arrives (so X is a continuous r.v., with $X = T_1$ if the 1st email is non-spam, $X = T_2$ if the 1st email is spam but the 2nd one isn't, etc.).

   (a) Find the mean and variance of X.

   (b) Find the MGF of X. What famous distribution does this imply that X has (be sure to state its parameter values)?

   Hint for both parts: let N be the number of emails until the first non-spam (including that one), and write X as a sum of N terms; then condition on N.

5. One of two identical-looking coins is picked from a hat randomly, where one coin has probability p_1 of Heads and the other has probability p_2 of Heads. Let X be the number of Heads after flipping the chosen coin n times. Find the mean and variance of X.

2 Inequalities

1. Let X and Y be i.i.d. positive r.v.s, and let $c > 0$. For each part below, fill in the appropriate equality or inequality symbol: write $=$ if the two sides are always equal, $\leq$ if the lefthand side is less than or equal to the righthand side (but they are not necessarily equal), and similarly for $\geq$. If no relation holds in general, write ?.

   (a) $E(\ln(X))$ ____ $\ln(E(X))$

   (b) $E(X)$ ____ $\sqrt{E(X^2)}$

   (c) $E(\sin^2(X)) + E(\cos^2(X))$ ____ 1

   (d) $E(|X|)$ ____ $\sqrt{E(X^2)}$

   (e) $P(X > c)$ ____ $\frac{E(X^3)}{c^3}$

   (f) $P(X \leq Y)$ ____ $P(X \geq Y)$

   (g) $E(XY)$ ____ $\sqrt{E(X^2)E(Y^2)}$

   (h) $P(X + Y > 10)$ ____ $P(X > 5 \text{ or } Y > 5)$

   (i) $E(\min(X, Y))$ ____ $\min(EX, EY)$

   (j) $E(X/Y)$ ____ $\frac{EX}{EY}$

   (k) $E(X^2(X^2 + 1))$ ____ $E(X^2(Y^2 + 1))$

   (l) $E(\frac{X^3}{X^3+Y^3})$ ____ $E(\frac{Y^3}{X^3+Y^3})$

2. (a) Show that $E(1/X) > 1/(EX)$ for any positive non-constant r.v. X.

   (b) Show that for any two positive r.v.s X and Y with neither a constant multiple of the other, $E(X/Y)E(Y/X) > 1$.

3. For i.i.d. r.v.s $X_1, \ldots, X_n$ with mean μ and variance σ^2, give a value of n (as a specific number) that will ensure that there is at least a 99% chance that the sample mean will be within 2 standard deviations of the true mean μ.

4. The famous *arithmetic mean-geometric mean* inequality says that for any positive numbers $a_1, a_2, \ldots, a_n$,

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq (a_1 a_2 \cdots a_n)^{1/n}.$$

Show that this inequality follows from Jensen's inequality, by considering $E\log(X)$ for a r.v. X whose possible values are $a_1, \ldots, a_n$ (you should specify the PMF of X; if you want, you can assume that the a_j are distinct (no repetitions), but be sure to say so if you assume this).

Stat 110 Strategic Practice 10 Solutions, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1 Conditional Expectation & Conditional Variance

1. Show that $E((Y - E(Y|X))^2|X) = E(Y^2|X) - (E(Y|X))^2$, so these two expressions for $\text{Var}(Y|X)$ agree.

   This is the conditional version of the fact that

   $$\text{Var}(Y) = E((Y - E(Y))^2) = E(Y^2) - (E(Y))^2,$$

   and so must be true since conditional expectations *are* expectations, just as conditional probabilities are probabilities. Algebraically, letting $g(X) = E(Y|X)$ we have

   $$E((Y-E(Y|X))^2|X) = E(Y^2-2Yg(X)+g(X)^2|X) = E(Y^2|X)-2E(Yg(X)|X)+E(g(X)^2|X),$$

   and $E(Yg(X)|X) = g(X)E(Y|X) = g(X)^2, E(g(X)^2|X) = g(X)^2$ by taking out what's known, so the righthand side above simplifies to $E(Y^2|X) - g(X)^2$.

2. Prove Eve's Law.

   We will show that $\text{Var}(Y) = E(\text{Var}(Y|X))+\text{Var}(E(Y|X))$. Let $g(X) = E(Y|X)$. By Adam's Law, $E(g(X)) = E(Y)$. Then

   $$E(\text{Var}(Y|X)) = E(E(Y^2|X) - g(X)^2) = E(Y^2) - E(g(X)^2),$$

   $$\text{Var}(E(Y|X)) = E(g(X)^2) - (E(g(X))^2 = E(g(X)^2) - (E(Y))^2.$$

   Adding these equations, we have Eve's Law.

3. Let X and Y be random variables with finite variances, and let $W = Y - E(Y|X)$. This is a *residual*: the difference between the true value of Y and the predicted value of Y based on X.

   (a) Compute $E(W)$ and $E(W|X)$.

   Adam's law (iterated expectation), taking out what's known, and linearity give

   $$\begin{aligned} E(W) &= EY - E(E(Y|X)) = EY - EY = 0, \\ E(W|X) &= E(Y|X) - E(E(Y|X)|X) = E(Y|X) - E(Y|X) = 0. \end{aligned}$$

(b) Compute $\text{Var}(W)$, for the case that $W|X \sim \mathcal{N}(0, X^2)$ with $X \sim \mathcal{N}(0,1)$.

Eve's Law gives

$$\text{Var}(W) \;=\; \text{Var}(E(W|X)) + E(\text{Var}(W|X)) = \text{Var}(0) + E(X^2) = 0 + 1 = 1.$$

4. Emails arrive one at a time in an inbox. Let T_n be the time at which the nth email arrives (measured on a continuous scale from some starting point in time). Suppose that the waiting times between emails are i.i.d. Expo(λ), i.e., $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are i.i.d. Expo(λ).

   Each email is non-spam with probability p, and spam with probability $q = 1-p$ (independently of the other emails and of the waiting times). Let X be the time at which the first non-spam email arrives (so X is a continuous r.v., with $X = T_1$ if the 1st email is non-spam, $X = T_2$ if the 1st email is spam but the 2nd one isn't, etc.).

   (a) Find the mean and variance of X.

   Write $X = X_1 + X_2 + \cdots + X_N$, where X_j is the time from the $(j-1)$th to the jth email for $j \geq 2$, and $X_1 = T_1$. Then $N - 1 \sim \text{Geom}(p)$, so

$$E(X) = E(E(X|N)) = E(N\frac{1}{\lambda}) = \frac{1}{p\lambda}.$$

   And

$$\text{Var}(X) = E(\text{Var}(X|N)) + \text{Var}(E(X|N)) = E(N\frac{1}{\lambda^2}) + \text{Var}(N\frac{1}{\lambda}),$$

   which is

$$\frac{1}{p\lambda^2} + \frac{1-p}{p^2\lambda^2} = \frac{1}{p^2\lambda^2}.$$

   (b) Find the MGF of X. What famous distribution does this imply that X has (be sure to state its parameter values)?

   Hint for both parts: let N be the number of emails until the first non-spam (including that one), and write X as a sum of N terms; then condition on N.

   Again conditioning on N, the MGF is

$$E(e^{tX}) = E(E(e^{tX_1}e^{tX_2}\cdots e^{tX_N}|N)) = E\left(E(e^{tX_1}|N)E(e^{tX_2}|N)\ldots E(e^{tX_N}|N)\right) = E(M_1(t)^N),$$

where $M_1(t)$ is the MGF of X_1 (which is $\frac{\lambda}{\lambda-t}$ for $t<\lambda$). By LOTUS, this is

$$p\sum_{n=1}^{\infty} M_1(t)^n q^{n-1} = \frac{p}{q}\sum_{n=1}^{\infty}(qM_1(t))^n = \frac{p}{q}\frac{qM_1(t)}{1-qM_1(t)} = \frac{\frac{p\lambda}{\lambda-t}}{1-\frac{q\lambda}{\lambda-t}} = \frac{p\lambda}{p\lambda-t}$$

for $t<p\lambda$ (as we need $qM_1(t)<1$ for the series to converge). This is the Expo($p\lambda$) MGF, so $X \sim \text{Expo}(p\lambda)$.

5. One of two identical-looking coins is picked from a hat randomly, where one coin has probability p_1 of Heads and the other has probability p_2 of Heads. Let X be the number of Heads after flipping the chosen coin n times. Find the mean and variance of X.

   The distribution of X is a *mixture* of two Binomials; we have seen before that X is *not* Binomial unless $p_1=p_2$. Let I be the indicator of having the p_1 coin. Then

   $$E(X) = E(X|I=1)P(I=1) + E(X|I=0)P(I=0) = \frac{1}{2}n(p_1+p_2).$$

   Alternatively, we can represent X as $X = IX_1+(1-I)X_2$ with $X_j \sim \text{Bin}(n,p_j)$, and I, X_1, X_2 independent. Then

   $$E(X) = E(E(X|I)) = E(Inp_1+(1-I)np_2) = \frac{1}{2}n(p_1+p_2).$$

   For the variance, note that it is *not* valid to say "$\text{Var}(X) = \text{Var}(X|I=1)P(I=1) + \text{Var}(X|I=0)P(I=0)$"; an extreme example of this mistake would be claiming that "$\text{Var}(I) = 0$ since $\text{Var}(I|I=1)P(I=1) + \text{Var}(I|I=0)P(I=0) = 0$"; of course, $\text{Var}(I) = \frac{1}{4}$). Instead, we can use Eve's Law:

   $$\text{Var}(X) = E(\text{Var}(X|I)) + \text{Var}(E(X|I)),$$

   where $\text{Var}(X|I) = Inp_1(1-p_1)+(1-I)np_2(1-p_2)$ is $np_1(1-p_1)$ with probability $1/2$ and $np_2(1-p_2)$ with probability $1/2$, and $E(X|I) = Inp_1+(1-I)np_2$ is np_1 or np_2 with probability $\frac{1}{2}$ each, so

   $$\text{Var}(X) = \frac{1}{2}\left(np_1(1-p_1)+np_2(1-p_2)\right) + \frac{1}{4}n^2(p_1-p_2)^2.$$

2 Inequalities

1. Let X and Y be i.i.d. positive r.v.s, and let $c > 0$. For each part below, fill in the appropriate equality or inequality symbol: write $=$ if the two sides are always equal, $\leq$ if the lefthand side is less than or equal to the righthand side (but they are not necessarily equal), and similarly for $\geq$. If no relation holds in general, write ?.

   (a) $E(\ln(X)) \leq \ln(E(X))$ (by Jensen: logs are concave)

   (b) $E(X) \leq \sqrt{E(X^2)}$ (since $\text{Var}(X) \geq 0$, or by Jensen)

   (c) $E(\sin^2(X)) + E(\cos^2(X)) = 1$ (by linearity, trig identity)

   (d) $E(|X|) \leq \sqrt{E(X^2)}$ (by (b) with $|X|$ in place of X; here $|X| = X$ anyway)

   (e) $P(X > c) \leq \frac{E(X^3)}{c^3}$ (by Markov, after cubing both sides of $X > c$)

   (f) $P(X \leq Y) = P(X \geq Y)$ (by symmetry, as X, Y are i.i.d.)

   (g) $E(XY) \leq \sqrt{E(X^2)E(Y^2)}$ (by Cauchy-Schwarz)

   (h) $P(X + Y > 10) \leq P(X > 5 \text{ or } Y > 5)$ (if $X + Y > 10$, then $X > 5$ or $Y > 5$)

   (i) $E(\min(X, Y)) \leq \min(EX, EY)$ (since $\min(X, Y) \leq X$ gives $E\min(X, Y) \leq EX$, and similarly $E\min(X, Y) \leq EY$)

   (j) $E(X/Y) \geq \frac{EX}{EY}$ (since $E(X/Y) = E(X)E(\frac{1}{Y})$, with $E(\frac{1}{Y}) \geq \frac{1}{EY}$ by Jensen)

   (k) $E(X^2(X^2+1)) \geq E(X^2(Y^2+1))$ (since $E(X^4) \geq (EX^2)^2 = E(X^2)E(Y^2) = E(X^2Y^2)$, because X^2 and Y^2 are i.i.d. and independent implies uncorrelated)

   (l) $E(\frac{X^3}{X^3+Y^3}) = E(\frac{Y^3}{X^3+Y^3})$ (by symmetry!)

2. (a) Show that $E(1/X) > 1/(EX)$ for any positive non-constant r.v. X.

   The function $g(x) = 1/x$ is strictly convex because $g''(x) = 2x^{-3} > 0$ for all $x > 0$, so Jensen's inequality yields $E(1/X) > 1/(EX)$. for any positive non-constant r.v. X.

   (b) Show that for any two positive r.v.s X and Y with neither a constant multiple of the other, $E(X/Y)E(Y/X) > 1$.

The r.v. $W = Y/X$ is positive and non-constant, so (a) yields

$$E(X/Y) = E(1/W) \;\; > \;\; 1/E(W) = 1/E(Y/X).$$

3. For i.i.d. r.v.s $X_1, \ldots, X_n$ with mean μ and variance σ^2, give a value of n (as a specific number) that will ensure that there is at least a 99% chance that the sample mean will be within 2 standard deviations of the true mean μ.

   We have to find n such that

   $$P(|\overline{X}_n - \mu| > 2\sigma) \leq 0.01.$$

   By Chebyshev's inequality (in the form $P(|Y - EY| > c) \leq \frac{\mathrm{Var}(Y)}{c^2}$), we have

   $$P(|\overline{X}_n - \mu| > 2\sigma) \leq \frac{\mathrm{Var}\bar{X}_n}{(2\sigma)^2} = \frac{\frac{\sigma^2}{n}}{4\sigma^2} = \frac{1}{4n}.$$

   So the desired inequality holds if $n \geq 25$.

4. The famous *arithmetic mean-geometric mean* inequality says that for any positive numbers $a_1, a_2, \ldots, a_n$,

   $$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq (a_1 a_2 \cdots a_n)^{1/n}.$$

   Show that this inequality follows from Jensen's inequality, by considering $E\log(X)$ for a r.v. X whose possible values are $a_1, \ldots, a_n$ (you should specify the PMF of X; if you want, you can assume that the a_j are distinct (no repetitions), but be sure to say so if you assume this).

   Assume that the a_j are distinct, and let X be a random variable which takes values from $a_1, a_2, ..., a_n$ with equal probability (the case of repeated a_j's can be handled similarly, letting the probability of $X = a_j$ be m_j/n, where m_j is the number of times a_j appears in the list $a_1, \ldots, a_n$). Jensen's inequality gives $E(\log X) \leq \log(EX)$, since the log function is concave. The left-hand side is $\frac{1}{n}\sum_{i=1}^{n} \log a_i$, while the right hand-side is $\log \frac{a_1+a_2+\cdots+a_n}{n}$. So we have the following inequality:

   $$\log \frac{a_1 + a_2 + \cdots + a_n}{n} \geq \frac{1}{n}\sum_{i=1}^{n} \log a_i$$

   Thus,

   $$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq e^{\frac{1}{n}\sum_{i=1}^{n} \log a_i} = e^{\frac{\log(a_1 \cdots a_n)}{n}} = (a_1 \cdots a_n)^{1/n}.$$

Stat 110 Penultimate Homework, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1. Judit plays in a total of $N \sim \text{Geom}(s)$ chess tournaments in her career. Suppose that in each tournament she has probability p of winning the tournament, independently. Let T be the number of tournaments she wins in her career.

(a) Find the mean and variance of T.

(b) Find the MGF of T. What is the name of this distribution (with its parameters)?

2. Let X_1, X_2 be i.i.d., and let $\bar{X} = \frac{1}{2}(X_1 + X_2)$. In many statistics problems, it is useful or important to obtain a conditional expectation given $\bar{X}$. As an example of this, find $E(w_1 X_1 + w_2 X_2 | \bar{X})$, where w_1, w_2 are constants with $w_1 + w_2 = 1$.

3. A certain stock has low volatility on some days and high volatility on other days. Suppose that the probability of a low volatility day is p and of a high volatility day is $q = 1 - p$, and that on low volatility days the percent change in the stock price is $\mathcal{N}(0, \sigma_1^2)$, while on high volatility days the percent change is $\mathcal{N}(0, \sigma_2^2)$, with $\sigma_1 < \sigma_2$.

Let X be the percent change of the stock on a certain day. The distribution is said to be a *mixture* of two Normal distributions, and a convenient way to represent X is as $X = I_1 X_1 + I_2 X_2$ where I_1 is the indicator r.v. of having a low volatility day, $I_2 = 1 - I_1$, $X_j \sim \mathcal{N}(0, \sigma_j^2)$, and I_1, X_1, X_2 are independent.

(a) Find the variance of X in two ways: using Eve's Law, and by calculating $\text{Cov}(I_1 X_1 + I_2 X_2, I_1 X_1 + I_2 X_2)$ directly.

(b) The *kurtosis* of a r.v. Y with mean μ and standard deviation σ is defined by

$$\text{Kurt}(Y) = \frac{E(Y - \mu)^4}{\sigma^4} - 3.$$

This is a measure of how heavy-tailed the distribution of Y. Find the kurtosis of X (in terms of $p, q, \sigma_1^2, \sigma_2^2$, fully simplified). The result will show that even though the kurtosis of any Normal distribution is 0, the kurtosis of X is positive and in fact can be very large depending on the parameter values.

4. We wish to estimate an unknown parameter θ, based on a r.v. X we will get to observe. As in the Bayesian perspective, assume that X and θ have a joint distribution. Let $\hat{\theta}$ be the estimator (which is a function of X). Then $\hat{\theta}$ is said to be *unbiased* if $E(\hat{\theta}|\theta) = \theta$, and $\hat{\theta}$ is said to be the *Bayes procedure* if $E(\theta|X) = \hat{\theta}$.

(a) Let $\hat{\theta}$ be unbiased. Find $E(\hat{\theta}-\theta)^2$ (the average squared difference between the estimator and the true value of θ), in terms of marginal moments of $\hat{\theta}$ and θ.

Hint: condition on θ.

(b) Repeat (a), except in this part suppose that $\hat{\theta}$ is the *Bayes procedure* rather than assuming that it is unbiased.

Hint: condition on X.

(c) Show that it is *impossible* for $\hat{\theta}$ to be both the Bayes procedure and unbiased, except in silly problems where we get to know θ perfectly by observing X.

Hint: if Y is a nonnegative r.v. with mean 0, then $P(Y=0)=1$.

5. The *surprise* of learning that an event with probability p happened is defined as $\log_2(1/p)$ (where the log is base 2 so that if we learn that an event of probability $1/2$ happened, the surprise is 1, which corresponds to having received 1 bit of information). Let X be a discrete r.v. whose possible values are $a_1, a_2, \ldots, a_n$ (distinct real numbers), with $P(X=a_j)=p_j$ (where $p_1+p_2+\cdots+p_n=1$).

The *entropy* of X is defined to be the average surprise of learning the value of X, i.e., $H(X)=\sum_{j=1}^{n} p_j \log_2(1/p_j)$. This concept was used by Shannon to create the field of information theory, which is used to quantify information and has become essential for communication and compression (e.g., MP3s and cell phones).

(a) Explain why $H(X^3)=H(X)$, and give an example where $H(X^2)\neq H(X)$.

(b) Show that the maximum possible entropy for X is when its distribution is uniform over $a_1, a_2, \ldots, a_n$, i.e., $P(X=a_j)=1/n$. (This should make sense intuitively: learning the value of X conveys the most information on average when X is equally likely to take any of its values, and the least possible information if X is a constant.)

Hint: this can be done by Jensen's inequality, without any need for calculus. To do so, consider a r.v. Y whose possible values *are* the probabilities $p_1, \ldots, p_n$, and show why $E(\log_2(1/Y)) \leq \log_2(E(1/Y))$ and how to interpret it.

6. In a national survey, a random sample of people are chosen and asked whether they support a certain policy. Assume that everyone in the population is equally likely to be surveyed at each step, and that the sampling is with replacement (sampling without replacement is typically more realistic, but with replacement will be a good approximation if the sample size is small compared to the population size). Let n be the sample size, and let $\hat{p}$ and p be the proportion of people who support the policy in the sample and in the entire population, respectively. Show that for every $c>0$,

$$P(|\hat{p}-p|>c) \leq \frac{1}{4nc^2}.$$

Stat 110 Penultimate Homework Solutions, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1. Judit plays in a total of $N \sim \text{Geom}(s)$ chess tournaments in her career. Suppose that in each tournament she has probability p of winning the tournament, independently. Let T be the number of tournaments she wins in her career.

(a) Find the mean and variance of T.

We have $T|N \sim \text{Bin}(N, p)$. By Adam's Law,

$$E(T) = E(E(T|N)) = E(Np) = p(1-s)/s.$$

By Eve's Law,

$$\begin{aligned}\text{Var}(T) &= E(\text{Var}(T|N)) + \text{Var}(E(T|N)) \\ &= E(Np(1-p)) + \text{Var}(Np) \\ &= p(1-p)(1-s)/s + p^2(1-s)/s^2 \\ &= \frac{p(1-s)(s+(1-s)p)}{s^2}.\end{aligned}$$

(b) Find the MGF of T. What is the name of this distribution (with its parameters)?

Let $I_j \sim \text{Bern}(p)$ be the indicator of Judit winning the jth tournament. Then

$$\begin{aligned}E(e^{tT}) &= E(E(e^{tT}|N)) \\ &= E((pe^t + q)^N) \\ &= s\sum_{n=0}^{\infty}(pe^t + 1 - p)^n(1-s)^n \\ &= \frac{s}{1-(1-s)(pe^t+1-p)}.\end{aligned}$$

This is reminiscent of the Geometric MGF used on HW 9. The famous discrete distributions we have studied whose possible values are $0, 1, 2, \ldots$ are the Poisson, Geometric, and Negative Binomial, and T is clearly not Poisson since its variance doesn't equal its mean. If $T \sim \text{Geom}(\theta)$, we have $\theta = \frac{s}{s+p(1-s)}$, as found by setting $E(T) = \frac{1-\theta}{\theta}$ or by finding $\text{Var}(T)/E(T)$. Writing the MGF of T as

$$E(e^{tT}) = \frac{s}{s+(1-s)p-(1-s)pe^t} = \frac{\frac{s}{s+(1-s)p}}{1-\frac{(1-s)p}{s+(1-s)p}e^t},$$

we see that $T \sim \text{Geom}(\theta)$, with $\theta = \frac{s}{s+(1-s)p}$. Note that this is consistent with (a).

The distribution of T can also be obtained by a story proof. Imagine that just before each tournament she may play in, Judit retires with probability s (if she retires, she does not play in that or future tournaments). Her tournament history can be written as a sequence of W (win), L (lose), R (retire), ending in the first R, where the probabilities of W, L, R are $(1-s)p, (1-s)(1-p), s$ respectively. For calculating T, the losses can be ignored: we want to count the number of W's before the R. The probability that a result is R given that it is W or R is $\frac{s}{s+(1-s)p}$, so we again have $T \sim \text{Geom}(\frac{s}{s+(1-s)p})$.

2. Let X_1, X_2 be i.i.d., and let $\bar{X} = \frac{1}{2}(X_1 + X_2)$. In many statistics problems, it is useful or important to obtain a conditional expectation given $\bar{X}$. As an example of this, find $E(w_1X_1 + w_2X_2|\bar{X})$, where w_1, w_2 are constants with $w_1 + w_2 = 1$.

By symmetry $E(X_1|\bar{X}) = E(X_2|\bar{X})$ and by linearity and taking out what's known, $E(X_1|\bar{X}) + E(X_2|\bar{X}) = E(X_1 + X_2|\bar{X}) = X_1 + X_2$. So $E(X_1|\bar{X}) = E(X_2|\bar{X}) = \bar{X}$ (this was also derived in class). Thus,

$$E(w_1X_1 + w_2X_2|\bar{X}) = w_1E(X_1|\bar{X}) + w_2E(X_2|\bar{X}) = w_1\bar{X} + w_2\bar{X} = \bar{X}.$$

3. A certain stock has low volatility on some days and high volatility on other days. Suppose that the probability of a low volatility day is p and of a high volatility day is $q = 1 - p$, and that on low volatility days the percent change in the stock price is $\mathcal{N}(0, \sigma_1^2)$, while on high volatility days the percent change is $\mathcal{N}(0, \sigma_2^2)$, with $\sigma_1 < \sigma_2$.

Let X be the percent change of the stock on a certain day. The distribution is said to be a *mixture* of two Normal distributions, and a convenient way to represent X is as $X = I_1X_1 + I_2X_2$ where I_1 is the indicator r.v. of having a low volatility day, $I_2 = 1 - I_1$, $X_j \sim \mathcal{N}(0, \sigma_j^2)$, and I_1, X_1, X_2 are independent.

(a) Find the variance of X in two ways: using Eve's Law, and by calculating $\text{Cov}(I_1X_1 + I_2X_2, I_1X_1 + I_2X_2)$ directly.

By Eve's Law,

$$\text{Var}(X) = E(\text{Var}(X|I_1)) + \text{Var}(E(X|I_1)) = E(I_1^2\sigma_1^2 + (1-I_1)^2\sigma_2^2) + \text{Var}(0) = p\sigma_1^2 + (1-p)\sigma_2^2,$$

since $I_1^2 = I_1, I_2^2 = I_2$. For the covariance method, expand

$$\text{Var}(X) = \text{Cov}(I_1X_1 + I_2X_2, I_1X_1 + I_2X_2) = \text{Var}(I_1X_1) + \text{Var}(I_2X_2) + 2\text{Cov}(I_1X_1, I_2X_2).$$

Then $\text{Var}(I_1X_1) = E(I_1^2X_1^2) - (E(I_1X_1))^2 = E(I_1)E(X_1^2) = p\text{Var}(X_1)$ since $E(I_1X_1) = E(I_1)E(X_1) = 0$. Similarly, $\text{Var}(I_2X_2) = (1-p)\text{Var}(X_2)$. And $\text{Cov}(I_1X_1, I_2X_2) =$

$E(I_1 I_2 X_1 X_2) - E(I_1 X_1)E(I_2 X_2) = 0$ since $I_1 I_2$ always equals 0. So again we have $\text{Var}(X) = p\sigma_1^2 + (1-p)\sigma_2^2$.

(b) The *kurtosis* of a r.v. Y with mean μ and standard deviation σ is defined by

$$\text{Kurt}(Y) = \frac{E(Y-\mu)^4}{\sigma^4} - 3.$$

This is a measure of how heavy-tailed the distribution of Y. Find the kurtosis of X (in terms of $p, q, \sigma_1^2, \sigma_2^2$, fully simplified). The result will show that even though the kurtosis of any Normal distribution is 0, the kurtosis of X is positive and in fact can be very large depending on the parameter values.

Note that $(I_1 X_1 + I_2 X_2)^4 = I_1 X_1^4 + I_2 X_2^4$ since the cross terms disappear (because $I_1 I_2$ is always 0) and any positive power of an indicator r.v. is that indicator r.v.! So

$$E(X^4) = E(I_1 X_1^4 + I_2 X_2^4) = 3p\sigma_1^4 + 3q\sigma_2^4.$$

Alternatively, we can use $E(X^4) = E(X^4|I_1 = 1)p + E(X^4|I_1 = 0)q$ to find $E(X^4)$. The mean of X is $E(I_1 X_1) + E(I_2 X_2) = 0$, so the kurtosis of X is

$$\text{Kurt}(X) = \frac{3p\sigma_1^4 + 3q\sigma_2^4}{(p\sigma_1^2 + q\sigma_2^2)^2} - 3.$$

This becomes 0 if $\sigma_1 = \sigma_2$, since then we have a Normal distribution rather than a mixture of two different Normal distributions. For $\sigma_1 < \sigma_2$, the kurtosis is positive since $p\sigma_1^4 + q\sigma_2^4 > (p\sigma_1^2 + q\sigma_2^2)^2$, as seen by a Jensen's inequality argument, or by interpreting this as saying $E(Y^2) > (EY)^2$ where Y is σ_1^2 with probability p and σ_2^2 with probability q.

4. We wish to estimate an unknown parameter θ, based on a r.v. X we will get to observe. As in the Bayesian perspective, assume that X and θ have a joint distribution. Let $\hat{\theta}$ be the estimator (which is a function of X). Then $\hat{\theta}$ is said to be *unbiased* if $E(\hat{\theta}|\theta) = \theta$, and $\hat{\theta}$ is said to be the *Bayes procedure* if $E(\theta|X) = \hat{\theta}$.

(a) Let $\hat{\theta}$ be unbiased. Find $E(\hat{\theta} - \theta)^2$ (the average squared difference between the estimator and the true value of θ), in terms of marginal moments of $\hat{\theta}$ and θ.

Hint: condition on θ.

Conditioning on θ, we have

$$\begin{aligned} E(\hat{\theta}-\theta)^2 &= E(E(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2|\theta)) \\ &= E(E(\hat{\theta}^2|\theta)) - E(E(2\hat{\theta}\theta|\theta)) + E(E(\theta^2|\theta)) \\ &= E(\hat{\theta}^2) - 2E(\theta E(\hat{\theta}|\theta)) + E(\theta^2) \\ &= E(\hat{\theta}^2) - 2E(\theta^2) + E(\theta^2) \\ &= E(\hat{\theta}^2) - E(\theta^2). \end{aligned}$$

(b) Repeat (a), except in this part suppose that $\hat{\theta}$ is the *Bayes procedure* rather than assuming that it is unbiased.

Hint: condition on X.

By the same argument as (a) except now conditioning on X, we have

$$\begin{aligned} E(\hat{\theta}-\theta)^2 &= E(E(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2|X)) \\ &= E(E(\hat{\theta}^2|X)) - E(E(2\hat{\theta}\theta|X)) + E(E(\theta^2|X)) \\ &= E(\hat{\theta}^2) - 2E(\hat{\theta}^2) + E(\theta^2) \\ &= E(\theta^2) - E(\hat{\theta}^2). \end{aligned}$$

(c) Show that it is *impossible* for $\hat{\theta}$ to be both the Bayes procedure and unbiased, except in silly problems where we get to know θ perfectly by observing X.

Hint: if Y is a nonnegative r.v. with mean 0, then $P(Y = 0) = 1$.

Suppose that $\hat{\theta}$ is both the Bayes procedure and unbiased. By the above, we have $E(\hat{\theta}-\theta)^2 = a$ and $E(\hat{\theta}-\theta)^2 = -a$, where $a = E(\hat{\theta}^2) - E(\theta^2)$. But that implies $a = 0$, which means that $\hat{\theta} = \theta$ (with probability 1). That can only happen in the extreme situation where the observed data reveal the true θ *perfectly*; in practice, nature is much more elusive and does not reveal its deepest secrets with such alacrity.

5. The *surprise* of learning that an event with probability p happened is defined as $\log_2(1/p)$ (where the log is base 2 so that if we learn that an event of probability $1/2$ happened, the surprise is 1, which corresponds to having received 1 bit of information). Let X be a discrete r.v. whose possible values are $a_1, a_2, \ldots, a_n$ (distinct real numbers), with $P(X = a_j) = p_j$ (where $p_1 + p_2 + \cdots + p_n = 1$).

The *entropy* of X is defined to be the average surprise of learning the value of X, i.e., $H(X) = \sum_{j=1}^{n} p_j \log_2(1/p_j)$. This concept was used by Shannon to create the field of information theory, which is used to quantify information and has become essential for communication and compression (e.g., MP3s and cell phones).

(a) Explain why $H(X^3) = H(X)$, and give an example where $H(X^2) \neq H(X)$.

Note that the definition of $H(X)$ depends only on the *probabilities* of the distinct values of X, not on what the values are. Since the function $g(x) = x^3$ is one-to-one, $P(X^3 = a_j^3) = p_j$ for all j. So $H(X^3) = H(X)$. For a simple example where $H(X^2) \neq H(X)$, let X be a "random sign," i.e., let X take values -1 and 1 with probability 1/2 each. Then X^2 has entropy 0, whereas X has entropy $\log_2(2) = 1$.

(b) Show that the maximum possible entropy for X is when its distribution is uniform over $a_1, a_2, \ldots, a_n$, i.e., $P(X = a_j) = 1/n$. (This should make sense intuitively: learning the value of X conveys the most information on average when X is equally likely to take any of its values, and the least possible information if X is a constant.)

Hint: this can be done by Jensen's inequality, without any need for calculus. To do so, consider a r.v. Y whose possible values *are* the probabilities $p_1, \ldots, p_n$, and show why $E(\log_2(1/Y)) \leq \log_2(E(1/Y))$ and how to interpret it.

Let Y be as in the hint. Then $H(X) = E(\log_2(1/Y))$ and $E(1/Y) = \sum_{j=1}^n \frac{p_j}{p_j} = n$. By Jensen's inequality, $E(\log_2(T)) \leq \log_2(E(T))$ for any positive r.v. T, since $\log_2$ is a concave function. Therefore,

$$H(X) = E(\log_2(1/Y)) \leq \log_2(E(1/Y)) = \log_2(n).$$

Equality holds if and only if $1/Y$ is a constant, which is the case where $p_j = 1/n$ for all j. This corresponds to X being equally likely to take on each of its values.

6. In a national survey, a random sample of people are chosen and asked whether they support a certain policy. Assume that everyone in the population is equally likely to be surveyed at each step, and that the sampling is with replacement (sampling without replacement is typically more realistic, but with replacement will be a good approximation if the sample size is small compared to the population size). Let n be the sample size, and let $\hat{p}$ and p be the proportion of people who support the policy in the sample and in the entire population, respectively. Show that for every $c > 0$,

$$P(|\hat{p} - p| > c) \leq \frac{1}{4nc^2}.$$

We can write $\hat{p} = X/n$ with $X \sim \text{Bin}(n, p)$. So $E(\hat{p}) = p, \text{Var}(\hat{p}) = p(1-p)/n$. Then by Chebyshev's inequality,

$$P(|\hat{p} - p| > c) \leq \frac{\text{Var}(\hat{p})}{c^2} = \frac{p(1-p)}{nc^2} \leq \frac{1}{4nc^2},$$

where the last inequality is because $p(1-p)$ is maximized at $p = 1/2$.