Math Review Notes—Probability

Gregory Faletto

Contents

Last updated July 10, 2020

Chapter 1

Probability

These are my notes from taking Math 505A at USC taught by Sergey Lototsky, Math 541A at USC taught by Steven Heilman, ISE 620 at USC taught by Sheldon Ross (as well as the corresponding textbooks *Introduction to Probability Models* [Ross, 2014] and *Stochastic Processes* [Ross, 2008] by Sheldon Ross) and the textbook *Probability and Random Processes* (Grimmett and Stirzaker) 3rd edition [Grimmett and Stirzaker, 2001], Math 541B at USC taught by Stanislav Minsker, Statistics 100B at UCLA taught by Nicolas Christou, as well as a few other sources I cite within the text.

1.1 To Know for Math 505A Midterm 1 (Discrete Random Variables)

1.1.1 Definitions

Definition 1.1.1. The **probability mass function** of a discrete random variable X is the function $f : \mathbb{R} \to [0,1]$ given by $f(x) = \Pr(X = x)$.

Definition 1.1.2. The **(cumulative) distribution function** of a discrete random variable F is given by

$$F(x) = \sum_{i:x_i \leq x} f(x_i)$$

Definition 1.1.3. The **joint probability mass function** $f : \mathbb{R}^2 \to [0,1]$ of two discrete random variables X and Y is given by

$$f(x,y) = \Pr(X = x \cap Y = y)$$

Definition 1.1.4. The **joint distribution function** $F : \mathbb{R}^2 \to [0,1]$ is given by

$$F(x,y) = \Pr(X \leq x \cap Y \leq y)$$

Definition 1.1.5. If $\Pr(B) > 0$ then the **conditional probability** that A occurs given that B occurs is defined to be

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

Definition 1.1.6. (Independent sets.) Let $A_1, A_2, \ldots$ be subsets of a sample space Ω, and let $\mathbb{P}$ be a probability law on Ω. We say that $A_1, A_2, \ldots$ are **independent** if for any finite subset S of $\{1, 2, \ldots\}$, we have

$$\mathbb{P}\Big(\bigcap_{i \in S} A_i\Big) = \prod_{i \in S} \mathbb{P}(A_i)$$

Definition 1.1.7. (Notation.) Let $X : \Omega \to \mathbb{R}$ be a random variable. Let $B \subseteq \mathbb{R}$. We define $\{X \in B\} := \{\omega \in \Omega : X(\omega) \in B\}$.

Definition 1.1.8. (Independence of random variables.) Random variables $X_1, X_2, \ldots$ are **independent** if for every $B_1, B_2, \ldots \subseteq \mathbb{R}$, the events $\{X_1 \in B_1\}, \{X_2 \in B_2\}, \ldots$ are independent; that is,

$$\mathbb{P}\Big(\bigcap_{i=1}^{n} \{X_i \in B_i\}\Big) = \prod_{i=1}^{n} \mathbb{P}(\{X_i \in B_i\})$$

Remark. A more informal definition is as follows: Two random variables X and Y are **independent** if and only if $\Pr(X \cap Y) = \Pr(X)\Pr(Y)$.

Theorem 1.1.1.1. (Law of total probability). If X is a random variable and Y is a discrete random variable taking on values $y_1, y_2, \ldots, y_n$, then $\Pr(X) = \sum_i \Pr(X \mid Y = y_i) \cdot \Pr(Y = y_i)$. (Can be used to prove independence.)

Lemma 1.1.1.2. Let $B_1, A_1, A_2, \ldots, A_m$ be subsets of a sample space Ω and let $\mathbb{P}$ be a probability law on Ω. Let $X_1, X_2, \ldots, X_m$, and Z be random variables. If Z is independent of X_i for all $i \in 1, \ldots, m$, then

$$\mathbb{P}\Big(\bigcap_{i=1}^{m} \{X_i \in A_i\} \cap Z \in B_1\Big) = \mathbb{P}\left(\bigcap_{i=1}^{m} \{X_i \in A_i\}\right) \mathbb{P}(Z \in B_1).$$

That is, $\bigcap_{i=1}^{m} X_i$ is independent of Z.

Proof. Note that by the definition of conditional probability,

$$\mathbb{P}\Big(\bigcap_{i=1}^{m} X_i \in A_i \mid Z \in B_1\Big) = \frac{\mathbb{P}(\bigcap_{i=1}^{m} \{X_i \in A_i\} \cap Z \in B_1)}{\mathbb{P}(Z \in B_1)}$$

$$= \frac{\prod_{j=1}^{m-1} \left[\mathbb{P}(X_j \in A_j \mid \bigcap_{k=j+1}^{m} \{X_k \in A_k\} \cap Z \in B_1)\right] \cdot \mathbb{P}(X_m \in A_m \cap Z \in B_1)}{\mathbb{P}(Z \in B_1)}$$

$$= \frac{\prod_{j=1}^{m-1} \left[\mathbb{P}(X_j \in A_j \mid \bigcap_{k=j+1}^{m} \{X_k \in A_k\}, Z \in B_1)\right] \cdot \mathbb{P}(X_m \in A_m \mid Z \in B_1)\mathbb{P}(Z \in B_1)}{\mathbb{P}(Z \in B_1)}$$

Canceling and using the independence of X_i and Z for all i, we have

$$= \prod_{j=1}^{m-1} \left[\mathbb{P}(X_j \in A_j \mid \bigcap_{k=j+1}^{m} \{X_k \in A_k\}\right] \cdot \mathbb{P}(X_m \in A_m) = \mathbb{P}\left(\bigcap_{i=1}^{m} \{X_i \in A_i\}\right)$$

Then the result follows since

$$\mathbb{P}\left(\bigcap_{i=1}^{m} \{X_i \in A_i\} \cap Z \in B_1\right) = \mathbb{P}\left(\bigcap_{i=1}^{m} X_i \in A_i \mid Z \in B_1\right)\mathbb{P}(Z \in B_1) = \mathbb{P}\left(\bigcap_{i=1}^{m} \{X_i \in A_i\}\right)\mathbb{P}(Z \in B_1).$$

□

Definition 1.1.9. Two random variables X and Y are **uncorrelated** if $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Proposition 1.1.1.3. (a) Two random variables are uncorrelated if and only if their covariance $\text{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$ equals 0.

(b) If X and Y are independent then they are uncorrelated.

Theorem 1.1.1.4. If X and Y are independent and $g, h : \mathbb{R} \to \mathbb{R}$, then $g(X)$ and $h(Y)$ are also independent.

Definition 1.1.10. We say that a nonnegative random variable X is **lattice** with period d if

$$\sum_{n=0}^{\infty} \Pr(X = nd) = 1$$

and d is the largest number that satisfies this equation.

Remark. Most common discrete random variables are lattice, but not all discrete random variables are. For example, the discrete random variable X such that

$$\Pr(X = 1/i) = p_i > 0, \qquad i \geq 1$$

such that $\sum_{i=1}^{\infty} p_i = 1$ is not lattice. Another example is

$$X = \begin{cases} 1 & \text{with probability } 1/2 \\ \sqrt{2} & \text{with probability } 1/2 \end{cases}$$

because there is no d such that there are integers m and n so that

$$1 = md$$

$$\sqrt{2} = nd.$$

(If there were, then we would have

$$\frac{n}{m} = \frac{nd}{md} = \sqrt{2}$$

but there are no integers n and m such that this is true, since $\sqrt{2}$ is irrational.)

1.1.2 Conditioning

Definition 1.1.11. The **conditional distribution function** of Y given $X = x$, written $F_{Y|X}(\cdot \mid x)$, is defined by

$$F_{Y|X}(y \mid x) = \Pr(Y \leq y \mid X = x)$$

Definition 1.1.12. The **conditional probability mass function** of Y given $X = x$, written $f_{Y|X}(\cdot \mid x)$, is defined by

$$f_{Y|X}(y \mid x) = \Pr(Y = y \mid X = x)$$

Definition 1.1.13 (Conditional Expectation (Math 541B definition)). Let X, Y be random variables, and assume that $\mathbb{E}|X| < \infty$. Then the conditional expectation of X given Y, denoted $\mathbb{E}(X \mid Y)$, is a function of Y: $g(Y) = \mathbb{E}(X \mid Y)$ such that for any bounded f,

$$\mathbb{E}(Xf(Y)) = \mathbb{E}(g(Y)f(Y))$$

Now assume that $\mathbb{E}X^2 < \infty$, $\mathbb{E}Y^2 < \infty$. Consider the Hilbert space $(\mathbb{H}, \langle \cdot, \cdot \rangle)$ consisting of all random variables with finite second moment and $\langle X, Y \rangle := \mathbb{E}(XY)$. Consider $L(Y) = \{$all random variables of the form $X = \phi(Y)$ where $\mathbb{E}Z^2 < \infty\}$[1]. Then $\mathbb{E}(X \mid Y)$ is the orthogonal projection of X onto the subspace $L(Y)$. Indeed,

$$\mathbb{E}(X \cdot f(Y)) = \langle X, f(Y) \rangle = \mathbb{E}(\mathbb{E}(X \mid Y)f(Y))) = \langle E(X \mid Y), f(Y) \rangle$$

$$\iff \forall f, \langle X - \mathbb{E}(X \mid Y), f(Y) \rangle = 0.$$

(because $\mathbb{E}\left[(X - \mathbb{E}(X \mid Y)) \cdot f(Y)\right] = \mathbb{E}\left(\mathbb{E}\left[Xf(Y) \mid Y\right] - \mathbb{E}\left[\mathbb{E}(X \mid Y)f(Y) \mid Y\right]\right) = \ldots$.)

$(f(Y) \in L(Y))$.

Remark. Think of expectation as the best summary of a random variable with one value, and conditional expectation as the best summary of one random variable as a function of another random variable.

Remark.

[1] clear that this subspace is linear: any linear combination of square integrable functions of Y is also square integrable. less clear that this is a closed space; that is, that any limit of functions in this space is also in this space. But it turns out that it is, so since this is a closed linear subspace.

Example 1.1.1. Assume (X, Y) has joint pdf $p(x, y)$. Then $\mathbb{E}(X \mid Y = y) = \int_{\mathbb{R}} x p_{X|Y}(x \mid y)\ dx$, where $p_{X|Y}(x \mid y)$ is the conditional density of X given Y; that is,

$$p_{X|Y}(x \mid y) = p(x, y) \Big/ \int_{x \in \mathbb{R}} p(x, y)\ dx.$$

Theorem 1.1.2.1. Iterated expectations:

(i) $\mathbb{E}\big[\mathbb{E}(X \mid Y)\big] = \mathbb{E}(X)$ **(Law of Total Expectation)**

(ii) $\mathbb{E}\big[(X \mid Y) \mid Z\big] = \mathbb{E}(X \mid Y)$

(iii) $\mathbb{E}(E(XY \mid Y)) = \mathbb{E}(Y\mathbb{E}(X \mid Y))$

Proof. (i) Discrete case:

$$\mathbb{E}\big[\mathbb{E}(X \mid Y)\big] = \sum_y \mathbb{E}(X \mid Y = y) \Pr(Y = y) = \sum_y \sum_x x \Pr(X = x \mid Y = y) \Pr(Y = y)$$

$$= \sum_y \sum_x x \Pr(X = x \cap Y = y) = \sum_x x \sum_y \Pr(X = x \cap Y = y) = \sum_x x \Pr(X = x) = \mathbb{E}(X)$$

Continuous case:

$$\mathbb{E}\big[\mathbb{E}(X \mid Y)\big] = \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y) dy = \text{ (by definition 1.75) } \int_{-\infty}^{\infty} \bigg(\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) dx \bigg) f_Y(y) dy$$

$$\text{(by Fubini's Theorem, Theorem ??) } \int_{-\infty}^{\infty} \bigg(\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) f_Y(y) dy \bigg) dx = \int_{-\infty}^{\infty} x f_X(x) dx = \mathbb{E}(X)$$

□

Definition 1.1.14. Conditional Variance: $\mathrm{Var}(X \mid Y) = \mathbb{E}\big[(X - \mathbb{E}(X \mid Y))^2 \mid Y\big]$

Theorem 1.1.2.2 (Conditional Expectation as a Random Variable). (i) Let X, Y be random variables such that (X, Y) is uniformly distributed on the triangle $\{(x, y) \in \mathbb{R}^2 : x \geq 0, y \geq 0, x + y \leq 1\}$. Then

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

(ii) Total Expectation Theorem:

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

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

(iii)

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

(iv)

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

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

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

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

Proof. (i) Note that since $0 \leq x$ and $x + y \leq 1$, conditional on $y = y \geq 0$ x is uniformly distributed on $[0, 1-y]$. That is,

$$\Pr(X \leq x \mid Y = y) = \begin{cases} 0 & x < 0 \\ x/(1-y) & 0 \leq x < 1-y \\ 1 & x \geq 1-y \end{cases}$$

Since the expected value of a random variable uniformly distributed on $[a,b]$ is $(a+b)/2$, it follows that $\mathbb{E}(X \mid Y = y) = (0+1-y)/2 = (1/2)(1-y)$. Therefore $\boxed{\mathbb{E}(X \mid Y) = \frac{1}{2}(1-Y)}$.

(ii)

- Discrete case:

$$\mathbb{E}\big[\mathbb{E}(X \mid Y)\big] = \sum_y \mathbb{E}(X \mid Y = y)\Pr(Y = y) = \sum_y \sum_x x \Pr(X = x \mid Y = y)\Pr(Y = y)$$

$$= \sum_y \sum_x x \Pr(X = x \cap Y = y) = \sum_x x \sum_y \Pr(X = x \cap Y = y) = \sum_x x \Pr(X = x) = \mathbb{E}(X)$$

- Continuous case:

$$\mathbb{E}\big[\mathbb{E}(X \mid Y)\big] = \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y)dy = \text{ (by definition 1.75) } \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)dx \right) f_Y(y)dy$$

$$\text{(by Fubini's Theorem) } \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) f_Y(y)dy \right) dx = \int_{-\infty}^{\infty} x f_X(x)dx = \mathbb{E}(X)$$

- Next we will show that the quantity $\mathbb{E}(X - h(Y))^2$ is minimized among all functions $h : \mathbb{R} \to \mathbb{R}$ when $h(Y) = \mathbb{E}(X \mid Y)$. We seek

$$\arg\min_{\{h:\mathbb{R}\to\mathbb{R}\}} \mathbb{E}(X - h(Y))^2 = \arg\min_{\{h:\mathbb{R}\to\mathbb{R}\}} \big[\mathbb{E}(X^2) - 2\mathbb{E}[h(Y)]\mathbb{E}(X) + \mathbb{E}[h(Y)^2]\big].$$

This expression is quadratic in $\mathbb{E}[h(Y)]$. Differentiating with respect to $\mathbb{E}[h(Y)]$ and setting equal to 0, we have

$$2\mathbb{E}[h(Y)] - 2\mathbb{E}(X) = 0 \iff \mathbb{E}[h(Y)] = \mathbb{E}(X) \implies \boxed{\arg\min_{\{h:\mathbb{R}\to\mathbb{R}\}} \mathbb{E}(X - h(Y))^2 = \mathbb{E}(X \mid Y)}$$

(iii) • Discrete case:

$$\mathbb{E}(Xh(Y)|Y) = \sum_{x\in\mathbb{R}} x \cdot h(Y) \cdot \Pr(X = x \mid Y) = h(Y) \sum_{x\in\mathbb{R}} x \cdot \Pr(X = x \mid Y) = h(Y)\mathbb{E}(X \mid Y).$$

Continuous case:

$$\mathbb{E}(Xh(Y)|Y) = \int_{x\in\mathbb{R}} x \cdot h(Y) \cdot f_{X|Y}(x) = h(Y) \int_{x\in\mathbb{R}} x \cdot f_{X|Y}(x) = h(Y)\mathbb{E}(X \mid Y).$$

• Discrete case: Note that

$$\mathbb{E}[X \mid h(Y)] = \sum_{x\in\mathbb{R}} x \Pr(X = x \mid h(Y)) = \sum_{x\in\mathbb{R}} x\mathbb{E}[\mathbf{1}_{\{X=x\}} \mid h(Y)]$$

where $\mathbf{1}_{\{X=x\}}$ is an indicator variable for X taking on the value x. Note that $\mathbb{E}[\mathbf{1}_{\{X=x\}} \mid h(Y)]$ is a function of Y (and a random variable). Then we have

$$\mathbb{E}([\mathbb{E}(X \mid h(Y)] \mid Y) = \mathbb{E}\left(\sum_{x\in\mathbb{R}} x\mathbb{E}[\mathbf{1}_{\{X=x\}} \mid h(Y)] \mid Y \right) = \sum_{x\in\mathbb{R}} \mathbb{E}[x\mathbb{E}[\mathbf{1}_{\{X=x\}} \mid h(Y)] \mid Y]$$

$$= \text{(by the previous result)} \sum_{x\in\mathbb{R}} \mathbb{E}[\mathbf{1}_{\{X=x\}} \mid h(Y)]\mathbb{E}[x \mid Y] = \sum_{x\in\mathbb{R}} \Pr(X = x \mid h(Y)) \cdot x = \mathbb{E}(X|h(Y)).$$

Continuous case: Note that

$$\mathbb{E}[X \mid h(Y)] = \int_{x\in\mathbb{R}} x f_{X|h(Y)}(x)dx.$$

Note that for a fixed x, $f_{X|h(Y)}(x)$ is a function of Y (and a random variable). Then we have

$$\mathbb{E}([\mathbb{E}(X \mid h(Y)] \mid Y) = \mathbb{E}\left(\int_{x\in\mathbb{R}} x f_{X|h(Y)}(x)dx \mid Y \right) = \int_{x\in\mathbb{R}} \mathbb{E}[x \cdot f_{X|h(Y)}(x) \mid Y]dx$$

$$= \text{(by the previous result)} \int_{x\in\mathbb{R}} f_{X|h(Y)}(x)\mathbb{E}[x \mid Y]dx = \int_{x\in\mathbb{R}} f_{X|h(Y)}(x) \cdot x = \mathbb{E}(X|h(Y)).$$

(iv) • Discrete case:

$$\mathbb{E}(X|X) = \sum_{x\in\mathbb{R}} x \Pr(X = x \mid X)$$

Note that

$$\Pr(X = x \mid X) = \begin{cases} 1 & x = X \\ 0 & \text{otherwise} \end{cases}$$

so we have

$$\mathbb{E}(X|X) = \sum_{x\in\mathbb{R}} x \Pr(X = x \mid X) = \ldots + 0 + 0 + X + 0 + 0 + \ldots = X.$$

Continuous case:

$$\mathbb{E}(X|X) = \int_{x\in\mathbb{R}} x \cdot dF_X$$

Note that

$$\Pr(X \leq x \mid X) = F_{X|X}(x) = \begin{cases} 0 & x < X \\ 1 & x \geq X \end{cases}$$

so we have

$$\mathbb{E}(X|X) = \int_{x\in\mathbb{R}} x \cdot dF_X = X.$$

- Discrete case:

$$\mathbb{E}(X+Y \mid Z=z) = \sum_{x\in\mathbb{Z}}\sum_{y\in\mathbb{Z}}(x+y)\Pr(X=x, Y=y \mid Z)$$

$$= \sum_{x\in\mathbb{Z}}\sum_{y\in\mathbb{Z}} x\Pr(X=x, Y=y \mid Z) + \sum_{x\ \in\mathbb{Z}}\sum_{y\in\mathbb{Z}} y\Pr(X=x, Y=y \mid Z)$$

$$= \sum_{x\in\mathbb{R}} x\Pr(X=x \mid Z) + \sum_{y\in\mathbb{Z}} y\Pr(Y=y \mid Z) = \mathbb{E}(X|Z) + \mathbb{E}(Y|Z).$$

Continuous case:

$$\mathbb{E}(X+Y \mid Z=z) = \int_{x\in\mathbb{R}}\int_{y\in\mathbb{R}}(x+y)\Pr(X=x, Y=y \mid Z)dydx$$

$$= \int_{x\in\mathbb{R}}\int_{y\in\mathbb{R}} x\Pr(X=x, Y=y \mid Z)dydx + \int_{x\in\mathbb{R}}\int_{y\in\mathbb{R}} y\Pr(X=x, Y=y \mid Z)dydx$$

$$= \int_{x\in\mathbb{R}} x\Pr(X=x \mid Z)dx + \int_{y\in\mathbb{R}} y\Pr(Y=y \mid Z)dy = \mathbb{E}(X|Z) + \mathbb{E}(Y|Z).$$

(v) Note by the definition of conditional probability that

$$\mathbb{P}(X=x \mid Y=y \cap Z=z) = \frac{\mathbb{P}(X=x \cap \{Y=y \cap Z=z\})}{\mathbb{P}(Y=y \cap Z=z)}$$

Using the independence of X and Y from Z and Lemma 1.1.1.2, we can express this as

$$= \frac{\mathbb{P}(X=x \cap Y=y)\mathbb{P}(Z=z)}{\mathbb{P}(Y=y)\mathbb{P}(Z=z)} = \frac{\mathbb{P}(X=x \cap Y=y)}{\mathbb{P}(Y=y)} = \mathbb{P}(X=x \mid Y=y)$$

by the definition of conditional probability. So $\mathbb{P}(X=x \mid Y=y, Z) = \mathbb{P}(X=x, Y=y)$. Therefore we have (in the discrete case)

$$\mathbb{E}(X \mid Y=y, Z) = \sum_{x\in\mathbb{R}} x\mathbb{P}(X=x \mid Y=y, Z) = \sum_{x\in\mathbb{R}} x\mathbb{P}(X=x, Y=y) = \mathbb{E}(X \mid Y=y) = g(y)$$

which implies that $\mathbb{E}(X \mid Y, Z) = g(Y) = \mathbb{E}(X \mid Y)$. In the continuous case, note that

$$F_{X|Y,Z}(x) = \mathbb{P}(X \leq x \mid Y, Z) = \mathbb{P}(X \leq x \mid Y) = F_{X|Y}(x)$$

by Lemma 1.1.1.2. Therefore we have

$$\mathbb{E}(X \mid Y, Z) = \int_{x \in \mathbb{R}} x dF_{X|Y,Z}(x) = \int_{x \in \mathbb{R}} x dF_{X|Y}(x) = \mathbb{E}(X \mid Y).$$

□

Corollary 1.1.2.2.1.

$\mathbb{E}(h(Y) \mid Y) = h(Y)$.

Proof. Use the first result in part (iii) of Theorem 1.1.2.2 with $X = 1$ and note that $\mathbb{E}(1 \mid Y) = 1$.

□

Corollary 1.1.2.2.2.

$\mathbb{E}[\mathbb{E}(X \mid Y)) \mid Y] = \mathbb{E}(X \mid Y)$.

Proof. Use the second result in part (iii) of Theorem 1.1.2.2 with $h(Y) = Y$.

□

Lemma 1.1.2.3. If $X \geq Z$ then $\mathbb{E}(X|Y) \geq \mathbb{E}(Z|Y)$.

Proof. Suppose that X and Z are nonnegative. Note that

$$\mathbb{E}(X \mid Y) = \int_0^\infty \Pr(X > t \mid Y) dt \geq \int_0^\infty \Pr(Z > t \mid Y) dt = \mathbb{E}(Z \mid Y)$$

where the third step follows since $X \geq Z$. If X and Z are not nonnegative, then

$$\begin{aligned}
\mathbb{E}(X \mid Y) &= \mathbb{E}(\max\{X, 0\} \mid Y) - \mathbb{E}(\max\{-X, 0\} \mid Y) \\
&= \int_0^\infty \Pr(\max\{X, 0\} > t \mid Y) dt - \int_0^\infty \Pr(\max\{-X, 0\} > t \mid Y) dt \\
&\geq \int_0^\infty \Pr(\max\{Z, 0\} > t \mid Y) dt - \int_0^\infty \Pr(\max\{-Z, 0\} > t \mid Y) dt = \mathbb{E}(Z \mid Y)
\end{aligned}$$

where the inequality follows since $X \geq Z$.

□

1.1.3 Convolution

Theorem 1.1.3.1. Sums of random variables. If X and Y are independent then

$$\Pr(X+Y=z) = f_{X+Y}(z) = \sum_x f_X(x) f_Y(z-x) = \sum_y f_X(z-y) f_Y(y)$$

Remark. Convolution on the integers. Let X, Y be independent integer-valued random variables. Let $t \in \mathbb{Z}$.

$$\Pr(X+Y=t) = \sum_{j,k \in \mathbb{Z}: j+k=t} \Pr(X=j, Y=k) = \sum_{j \in \boldsymbol{I}} \Pr(X=j, Y=t-j) = \sum_{j \in \mathbb{Z}} \Pr(X=j) \Pr(Y=t-j)$$

$$= \sum_{j \in \mathbb{Z}} p_X(j) p_Y(t-j)$$

Definition 1.1.15. Let $g, h : \mathbb{Z} \to \mathbb{R}$ be functions. The **convolution** of g and h, denoted by $g * h$, is a function $g * h : \mathbb{Z} \to \mathbb{R}$ defined by

$$(g*h)(t) = \sum_{j \in \mathbb{Z}} g(j) h(t-j) \quad \forall t \in \mathbb{Z}$$

Definition 1.1.16. (Convolution on the real line.) Let $g, h : \mathbb{R} \to \mathbb{R}$ b functions. The **convolution** of g and h, denoted by $g * h$, is a function $g * h : \mathbb{R} \to \mathbb{R}$ defined by

$$(g*h)(t) = \int_{-\infty}^{\infty} g(x) h(t-x) dx, \quad \forall t \in \mathbb{R}$$

Proposition 1.1.3.2. Let X, Y be two continuous independent random variables such that $\Pr(X+Y \leq t)$ is differentiable with respect to $t \in \mathbb{R}$. Then

$$f_{X+Y}(t) = (f_X * f_Y)(t), \quad \forall t \in \mathbb{R}$$

Proof.

$$\Pr(X+Y \leq t) = \int_{\{(x,y) \in \mathbb{R}^2 : x+y \leq t\}} f_{X,Y}(x,y) dx dy = \int_{x=-\infty}^{x=\infty} \int_{y=-\infty}^{y=t-x} f_X(x) f_Y(y) dy dx$$

Then, since $\Pr(X+Y \leq t)$ is differentiable with respect to t, we have by the Fundamental Theorem of Calculus

$$f_{X+Y}(t) = \frac{\mathrm{d}}{\mathrm{d}t} \Pr(X+Y \leq t) = \int_{x=-\infty}^{x=\infty} f_X(x) \frac{\mathrm{d}}{\mathrm{d}t} \int_{y=-\infty}^{y=t-x} f_Y(y) dy dx = \int_{x=-\infty}^{x=\infty} f_X(x) f_Y(t-x) dx$$

□

Example 1.1.2. Let X, Y be independent standard Gaussian random variables. Then by Proposition 1.1.3.2, $X + Y$ has density

$$f_{X+Y}(t) = \int_{\infty}^{\infty} f_X(x) f_Y(t-x) dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2/2} e^{-(t-x)^2/2} dx$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2 + tx - t^2/2} dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-t/2)^2 - t^2/4 - t^2/2} dx$$

$$= e^{-t^2/4} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-t/2)^2} dx = \frac{1}{2\pi} e^{-t^2/4} \int_{-\infty}^{\infty} e^{-x^2} dx$$

Let $x = y/\sqrt{2}, dx = dy/\sqrt{2}$.

$$= \frac{1}{2\pi} e^{-t^2/4} \int_{-\infty}^{\infty} e^{-y^2/2} \cdot \frac{1}{\sqrt{2}} dy = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2}} e^{-t^2/4} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2}} e^{-t^2/4} \frac{1}{\sqrt{2\pi}}$$

which is the density for a Gaussian random variable distributed as $\mathcal{N}(0, 1)$.

Definition 1.1.17. (Convoluted cdfs; Ross's in-class definition from ISE 620.) Suppose we have random variables X and Y with cdfs F_X and F_Y and pdfs f_X and f_Y. Then

$$(F_Y * F_X)(t) = \Pr(X + Y \leq t) = \int_{-\infty}^{\infty} \Pr(X + Y \leq t \mid X = x) f_X(x) dx$$

$$= \int_{-\infty}^{\infty} F_Y(t - x) f_X(x) dx$$

1.1.4 Compound Random Variables

Definition 1.1.18. Let $\{X_i\}$ be i.i.d. random variables. Let N be a random variable taking on positive integer values. Let

$$S = \sum_{i=1}^{N} X_i$$

Then S is a **compound random variable**.

Proposition 1.1.4.1. (Wald's Equation.) Let $\{X_i\}$ be i.i.d. random variables with mean $\mathbb{E}(X)$. Let N be a random variable taking on positive integer values, and let $S = \sum_{i=1}^{N} X_i$. Then $\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(X)$.

Proof.

$$\mathbb{E}(S \mid N) = \mathbb{E}\Big(\sum_{i=1}^{N} X_i\Big) = \sum_{i=1}^{N} \mathbb{E}(X_i) = N\mathbb{E}(X)$$

$$\mathbb{E}(S) = \mathbb{E}(\mathbb{E}[S \mid N]) = \mathbb{E}(N\mathbb{E}(X)) = \mathbb{E}(N)\mathbb{E}(X)$$

□

Proposition 1.1.4.2. Let $\{X_i\}$ be i.i.d. random variables with mean $\mathbb{E}(X)$. Let N be a random variable taking on positive integer values, and let $S = \sum_{i=1}^{N} X_i$. Then $\text{Cov}(N, S) = \mathbb{E}(X)\text{Var}(N)$.

Proof. Let $\mathbb{E}(X_i) = \mathbb{E}(X)$ for all i. We will use the result from Wald's Equation (Proposition 1.1.4.1: $E(S) = \mathbb{E}(X)\mathbb{E}(N)$). We have

$$\text{Cov}(N, S) = \mathbb{E}(NS) - \mathbb{E}(N)\mathbb{E}(S) = \mathbb{E}[\mathbb{E}(NS \mid N)] - \mathbb{E}(N)\mathbb{E}(N)\mathbb{E}(X) = \mathbb{E}[N\mathbb{E}(S \mid N)] - \mathbb{E}(N)^2\mathbb{E}(X) \quad (1.1)$$

Note that

$$\mathbb{E}(S \mid N) = \mathbb{E}\Big(\sum_{i=1}^{N} X_i\Big) = \sum_{i=1}^{N} \mathbb{E}(X_i) = N\mathbb{E}(X) \implies \mathbb{E}[N\mathbb{E}(S \mid N)] = \mathbb{E}(N^2\mathbb{E}(X)) = \mathbb{E}(X)\mathbb{E}(N^2)$$

Plugging this into (1.1) yields

$$\text{Cov}(N, S) = \mathbb{E}(X)\mathbb{E}(N^2) - \mathbb{E}(N)^2\mathbb{E}(X) = \mathbb{E}(X)\big[\mathbb{E}(N^2) - \mathbb{E}(N)^2\big] = \mathbb{E}(X)\text{Var}(N)$$

□

Definition 1.1.19 (**Compound Poisson random variable**)**.** Let N be a Poisson random variable. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables that are also independent of N. Then

$$S := \sum_{i=1}^{N} X_i$$

is called a **compound Poisson random variable.**

Proposition 1.1.4.3 (**Variance of a compound Poisson random variable, from Ross *Introduction to Probability Models***)**.** For a compound Poisson random variable $S = \sum_{i=1}^{N} X_i$ having $\mathbb{E}(N) = \lambda, \mathbb{E}(X_i) = \mu, \text{Var}(X_i) = \sigma^2$,

$$\text{Var}(S) = \lambda\sigma^2 + \lambda\mu^2 = \lambda\mathbb{E}(X^2).$$

1.1.5 Odds and Ends

Proposition 1.1.5.1. Inclusion-Exclusion Principle:

(a)

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n}(-1)^{k+1}\left(\sum_{1\leq i_1<\ldots<i_k\leq m} \Pr\left(A_{i1}\cap\ldots\cap A_{ik}\right)\right)$$

(b)

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{k=1}^{n}(-1)^{k+1}\left(\sum_{1\leq i_1<\ldots<i_k\leq m} |A_{i1}\cap\ldots\cap A_{ik}|\right)$$

To prove this, we will first prove the Multi-Binomial Theorem.

Lemma 1.1.5.2. (Multi-Binomial Theorem)

$$\prod_{i=1}^{d}(x_i+y_i)^{n_i} = \sum_{k_1=0}^{n_1}\sum_{k_2=0}^{n_2}\cdots\sum_{k_d=0}^{n_d}\binom{n_1}{k_1}x_1^{k_1}y_1^{n_1-k_1}\binom{n_2}{k_2}x_2^{k_2}y_2^{n_2-k_2}\cdots\binom{n_d}{k_d}x_d^{k_d}y_d^{n_d-k_d}$$

Proof. □

We are now ready to prove the Inclusion-Exclusion Principle.

Proof. (Proof of (a).) Begin by noting that

$$\mathbf{1}_{\{\cup_{i=1}^{n}A_i\}} = 1 - \prod_{i=1}^{n}(1-\mathbf{1}_{\{A_i\}}) \tag{1.2}$$

because the expression on the right will equal 1 if at least one term in the product equals 0 (that is, if $\mathbf{1}_{\{A_i\}} = 1$ for some $i \in 1,\ldots,n$) and will equal 0 if every term in the product equals 1 (if $\mathbf{1}_{\{A_i\}} = 0$ for every $i \in 1,\ldots,n$), which is exactly what we want. Expanding the right side of (1.2) using the Multi-Binomial Theorem (Lemma 1.1.5.2), we have

$$= 1 - \prod_{i=1}^{n}(1-\mathbf{1}_{\{A_i\}}) = 1-(1-\mathbf{1}_{\{A_1\}})(1-\mathbf{1}_{\{A_2\}})\cdots(1-\mathbf{1}_{\{A_n\}})$$

$$= 1 - \left[1+\sum_{k=1}^{n}(-1)^k\left(\sum_{1\leq i_1<\ldots<i_k\leq k}\mathbf{1}_{\{A_{i_1}\}}\cdots\mathbf{1}_{\{A_{i_k}\}}\right)\right] = -1\cdot\sum_{k=1}^{n}(-1)^k\left(\sum_{1\leq i_1<\ldots<i_k\leq k}\mathbf{1}_{\{A_{i_1}\cap\ldots\cap A_{i_k}\}}\right)$$

$$= \sum_{k=1}^{n}(-1)^{k+1}\left(\sum_{1\leq i_1<\ldots<i_k\leq k}\mathbf{1}_{\{A_{i_1}\cap\ldots\cap A_{i_k}\}}\right)$$

$$\implies \mathbf{1}_{\{\cup_{i=1}^n A_i\}} = \sum_{k=1}^{n}(-1)^{k+1}\Bigg(\sum_{1\leq i_1<\ldots<i_k\leq k}\mathbf{1}_{\{A_{i_1}\cap\ldots\cap A_{i_k}\}}\Bigg) \tag{1.3}$$

Taking expectations of both sides of (1.3) yields

$$\Pr(\cup_{i=1}^n A_i) = \mathbb{E}\Bigg[\sum_{k=1}^{n}(-1)^{k+1}\Bigg(\sum_{1\leq i_1<\ldots<i_k\leq k}\mathbf{1}_{\{A_{i_1}\cap\ldots\cap A_{i_k}\}}\Bigg)\Bigg]$$

$$= \sum_{k=1}^{n}(-1)^{k+1}\Bigg(\sum_{1\leq i_1<\ldots<i_k\leq k}\mathbb{E}[\mathbf{1}_{\{A_{i_1}\cap\ldots\cap A_{i_k}\}}]\Bigg)$$

$$= \sum_{k=1}^{n}(-1)^{k+1}\Bigg(\sum_{1\leq i_1<\ldots<i_k\leq k}\Pr\left(A_{i_1}\cap\ldots\cap A_{i_k}\right)\Bigg)$$

□

Proposition 1.1.5.3 (**a nice trick for upper bounding binomial sums, from Math 547**). For $d < n$,

$$\sum_{i=0}^{d}\binom{n}{i} \leq \left(\frac{en}{d}\right)^d.$$

Proof.

$$\sum_{i=0}^{d}\binom{n}{i}(d/n)^d \leq \sum_{i=0}^{d}\binom{n}{i}(d/n)^i \leq \sum_{i=0}^{n}\binom{n}{i}(d/n)^i = [1+(d/n)]^n \leq e^d \iff \sum_{i=0}^{d}\binom{n}{i} \leq \left(\frac{en}{d}\right)^d.$$

□

Proposition 1.1.5.4. (Proposition 1.6.1 in Sheldon Ross *A First Course in Probability.*) There are $\binom{n-1}{r-1}$ distinct positive integer-valued vectors $(x_1, x_2, \ldots, x_r), x_i > 0\ \forall\ i$ satisfying the equation $x_1 + x_2 + \ldots + x_r = n$.

Proof. (Not rigorous, but a justification.) Imagine we have n indistinguishable objects to allocate to r people. We lay out the n objects and take $r-1$ sticks to place in the $n-1$ spaces between them. The first person gets all the objects to the left of the leftmost stick, the second person gets the objects between the leftmost and second leftmost stick, and so on, until the last person gets all the objects to the right of the rightmost stick. The constraint that x_i be positive is equivalent to saying that each person must receive at least one object. Therefore we must place each stick in a different place. There are $\binom{n-1}{r-1}$ ways to do this.

□

Proposition 1.1.5.5. (Proposition 1.6.2 in Sheldon Ross *A First Course in Probability.*) There are $\binom{n+r-1}{r-1}$ distinct nonnegative integer-valued vectors $(x_1, x_2, \ldots, x_r), x_i \geq 0\ \forall\ i$ satisfying the equation $x_1 + x_2 + \ldots + x_r = n$.

Proof. We would like to solve the problem

$$x_1 + x_2 + \ldots + x_r = n, x_i \geq 0 \ \forall i$$

Note that we can transform this problem in the following way:

$$x_1 + 1 + x_2 + 1 + \ldots + x_r + 1 = n + 1 \cdot r, x_i + 1 \geq 1 \ \forall i$$

Letting $y_i = x_i + 1$, we have the equivalent system

$$y_1 + y_2 + \ldots + y_r = n + r, y_i \geq 1 \ \forall i$$

Since $y_i \geq 1 \iff y_i > 0$, by Proposition 1.1.5.4, the number of distinct solutions to this equation is $\binom{n+r-1}{r-1}$.

□

Proposition 1.1.5.6. More generally, if we desire solutions to $x_1 + x_2 + \ldots + x_r = n$ such that $x_i \geq k \in \mathbb{N}$, then $\binom{n+r\cdot(1-k)-1}{r-1} = \binom{n+r-1-rk}{r-1}$ solutions are possible.

Proof. We can construct a similar argument to that used in the proof of Proposition 1.1.5.5 by adding $r \cdot (1-k)$ to each side of $x_1 + x_2 + \ldots + x_r = n$:

$$x_1 + 1 - k + x_2 + 1 - k + \ldots + x_r + 1 - k = n + r \cdot (1-k), \ x_i \geq k \ \forall i$$

Then substitute $y_i = x_i + 1 - k$ to yield

$$y_1 + y_2 + \ldots + y_r = n + r \cdot (1-k), \ y_i + k - 1 \geq k \ \forall i$$

then apply Proposition 1.1.5.4 noting that $y_i + k - 1 \geq k \iff y_i \geq 1 \iff x_i \geq k$ to yield the result. □

Proposition 1.1.5.7. Even more generally, suppose we desire solutions to $x_1 + x_2 + \ldots + x_r = n$ such that $x_1 \geq k_1 \in \mathbb{N}, x_2 \geq k_2 \in \mathbb{N}, \ldots, x_r \geq k_r \in \mathbb{N}$. Then

$$\binom{n + \sum_{i=1}^{r}(1-k_i) - 1}{r-1} = \binom{n + r - 1 - \sum_{i=1}^{r} k_i}{r-1}$$

solutions are possible.

Proof. Very similar to the proof of Proposition 1.1.5.6. Add $\sum_{i=1}^{r}(1-k_i)$ to each side of $x_1 + x_2 + \ldots + x_r = n$:

$$x_1 + 1 - k_1 + x_2 + 1 - k_2 + \ldots + x_r + 1 - k_r = n + \sum_{i=1}^{r}(1-k_i), \ x_i \geq k_i \ \forall i$$

Then substitute $y_i = x_i + 1 - k_i$ to yield

$$y_1 + y_2 + \ldots + y_r = n + \sum_{i=1}^{r}(1 - k_i),\ y_i + k_i - 1 \geq k_i\ \forall i$$

Finally, apply Proposition 1.1.5.4 noting that $y_i + k_i - 1 \geq k_i \iff y_i \geq 1 \iff x_i \geq k_i$ to yield the result. □

Proposition 1.1.5.8. Suppose we desire solutions to $x_1 + x_2 + \ldots + x_r = n$ such that $\tau \leq r$ of the $\{x_i\}$ exceed some threshold k. For example, if we have $x_1 \geq k, x_2 \geq k, \ldots, x_\tau \geq k$, with $x_{\tau+1}, \ldots, x_r$ taking on arbitrary values, then the condition is satisfied. (The particular x_i that exceed k does not matter, as long as τ of them exceed k). Then

$$\binom{r}{\tau}\binom{n + r - 1 - k\tau}{r - 1}$$

solutions are possible.

Proof. By Proposition 1.1.5.7, the number of ways this condition can be met for a particular set of τ variables x_i is $\binom{n+r-1-k\tau}{r-1}$. Since there are $\binom{r}{\tau}$ ways to choose which τ variables will exceed k, the result follows. □

1.1.6 Methods for Calculating Quantities

- Expectation

  –

  Definition 1.1.20. (Math 541A definition 1.37.) Let Ω be a sample space, let $\mathbb{P}$ be a probability law on Ω. Let X be a random variable on Ω. Assume that X takes on nonnegative values; that is, $X : \Omega \to [0, \infty)$. We define the **expected value** of X by

  $$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > t)dt$$

  In analytic notation, $\mathbb{E}(X) = \int_\Omega X(\omega)d\mathbb{P}(\omega)$. More generally, if $g : [0 \to \infty) \to [0 \to \infty)$ is a differentiable function such that g' is continuous and $g(0) = 0$, we define

  $$\mathbb{E}(g(X)) = \int_0^\infty g'(t)\mathbb{P}(X > t)dt$$

  For a general random variable X, if $\mathbb{E}(\max\{X, 0\} < \infty)$ and if $\mathbb{E}(\max\{-X, 0\}) < \infty$, we then define $\mathbb{E}(X) = \mathbb{E}(\max\{X, 0\}) - \mathbb{E}(\max\{-X, 0\})$. Otherwise, we say that $\mathbb{E}(X)$ is undefined.

  Definition 1.1.21. In the above, taking $g(t) = t^n$ for any positive integer n, for any $t \geq 0$, we have

  $$\mathbb{E}(X^n) = \int_0^\infty nt^{n-1}\mathbb{P}(X > t)dt$$

Remark. If we assume that the expected value and the integral on $\mathbb{R}$ can be commuted, then the following derivation of the formula for $\mathbb{E}(g(X))$ can be given. From the Fundamental Theorem of Calculus, we have

$$g(X) = \int_0^X g'(t)dt = \int_0^\infty g'(t)\mathbf{1}_{\{X>t\}}dt$$

where $\mathbf{1}_{\{X>t\}}$ is an indicator variable. Therefore

$$\mathbb{E}(g(X)) = \int_0^\infty g'(t)\mathbf{1}_{\{X>t\}}dt = \int_0^\infty g'(t)\mathbb{E}(\mathbf{1}_{\{X>t\}}) = \int_0^\infty g'(t)\mathbb{P}(X>t)dt.$$

–

Definition 1.1.22. (Discrete random variables.) $\mathbb{E}(X) = \sum_x x \Pr(X = x)$

–

Theorem 1.1.6.1. (a) $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$
(b) If $X \geq 0$ then $\mathbb{E}(X) \geq 0$

–

Theorem 1.1.6.2. Expectation of sums is sum of expectations if sum is finite or if sum is infinite and all variables are positive, but not necessarily otherwise.

Example 1.1.3 (**Example 4.18 in *Introduction to Probability Models***)**.** Consider a sequence of discrete random variables $X_1, X_2, \ldots$ such that $X_i = 1$ with probability $1/2$ and $X_i = -1$ with probability $1/2$. Let N be a stopping time:

$$N = \min\{n : X_1, + \ldots + X_n = 1\} \implies 1 = X_1 + \ldots + X_N.$$

Let $I\{i \leq N\}$ be an indicator variable for $i \leq N$. So

$$1 = \sum_{i=1}^N X_i = \sum_{i=1}^\infty X_i I\{i \leq N\}.$$

Taking expectations we have

$$1 = \mathbb{E}\sum_{i=1}^\infty X_i I\{i \leq N\}.$$

Note that $N \geq i$ if and only if you have not yet stopped after the first $i-1$ games, so it depends on the results of the previous games but not on X_i. So $I\{i \leq N\}$ and X_i are independent. Therefore

$$\mathbb{E}[X_i I\{i \leq N\} = \mathbb{E}(X_i)\mathbb{E}(I\{i \leq N\}) = 0$$

since $\mathbb{E}(X_i) = 0$. So

$$\sum_{i=1}^\infty \mathbb{E}[X_i I\{i \leq N\}] = \sum_{i=1}^\infty 0 = 0$$

which means that

$$\mathbb{E}\sum_{i=1}^\infty X_i I\{i \leq N\} \neq \sum_{i=1}^\infty \mathbb{E}[X_i I\{i \leq N\}].$$

See also Example **??** in the Stochastic Processes notes.

–

Theorem 1.1.6.3. Law of the Unconscious Statistician: If X has mass function f, and $g : \mathbb{R} \to \mathbb{R}$, then

$$\mathbb{E}(g(X)) = \sum_x g(x)f(x)$$

–

Theorem 1.1.6.4. Expectation is a linear operator: $\mathbb{E}\big(\sum_i X_i\big) = \sum_i \mathbb{E}(X_i)$

–

Theorem 1.1.6.5. (Layer cake formulation.) If N is a discrete random variable taking on non-negative values, then $\mathbb{E}(N) = \sum_{i=1}^{\infty} \Pr(N \geq i)$.

Proof. Let $\mathbf{1}_{\{i \leq N\}}$ be an indicator variable. Then

$$N = \sum_{i=1}^{\infty} \mathbf{1}_{\{i \leq N\}}.$$

Take expectations of both sides to yield the result.

□

Remark. Also note that we can derive this result from Definition 1.1.20:

$$\mathbb{E}(X) = \int_0^{\infty} \Pr(X > t)dt = \sum_{k=1}^{\infty} \int_{k-1}^{k} \Pr(X > t)dt = \sum_{k=1}^{\infty} \Pr(X \geq k) = \sum_{k=0}^{\infty} \Pr(X > k).$$

Using Fubini's Theorem (Theorem **??**) to rearrange the sum, we can arrive at

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} \Pr(X > k) = \sum_{k=0}^{\infty} \sum_{j=k+1}^{\infty} \Pr(X = j) = \sum_{0 \leq k < j < infty} \Pr(X = j)$$

$$= \sum_{j=0}^{\infty} \sum_{k=0}^{j} \mathbb{P}(X = j) = \sum_{j=0}^{\infty} j \Pr(X = j)$$

which is the usual definition for a discrete random variable.

– For the conditional expectation of one Gaussian random variable on another when the covariance or correlation between them is known, see Proposition 1.3.7.1. For the conditional expectation of a set of Gaussian random variables on another set when the covariance matrix is known, see Proposition 1.3.7.2.

- Variance

–

Definition 1.1.23. $\text{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2$

–

Proposition 1.1.6.6. (Useful reformulation:) $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

–

Theorem 1.1.6.7. (Some useful results):

(a) $\text{Var}(aX) = a^2\text{Var}(X)$

(b) $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$

(c) $\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) \pm 2ab\text{Cov}(X,Y)$

(d) **Variance-Covariance Expansion.** $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2\sum_{1 \le i < j \le n} \text{Cov}(X_i, X_j)$

–

Definition 1.1.24. Conditional variance:

$$\text{Var}(X \mid Y) = \mathbb{E}\big[(X - \mathbb{E}(X \mid Y))^2 \mid Y\big]$$

–

Theorem 1.1.6.8. Law of Total Variance: $\text{Var}(X) = \text{Var}\big(\mathbb{E}(X \mid Y)\big) + \mathbb{E}\big(\text{Var}(X \mid Y)\big)$

Proof.

$$\text{Var}\big(\mathbb{E}(X \mid Y)\big) = \mathbb{E}[(\mathbb{E}(X \mid Y))^2] - [\mathbb{E}(\mathbb{E}(X \mid Y))]^2 = \mathbb{E}[(\mathbb{E}(X \mid Y))^2] - \mathbb{E}(X)^2 \tag{1.4}$$

$$\text{Var}(X \mid Y) = \mathbb{E}(X^2 \mid Y) - (\mathbb{E}(X \mid Y))^2 \implies \mathbb{E}[\text{Var}(X \mid Y)] = \mathbb{E}(X^2) - \mathbb{E}[(\mathbb{E}(X \mid Y))^2] \tag{1.5}$$

Adding together (1.4) and (1.5) yields

$$\text{Var}\big(\mathbb{E}(X \mid Y)\big) + \mathbb{E}\big(\text{Var}(X \mid Y)\big) = \mathbb{E}[(\mathbb{E}(X \mid Y))^2] - \mathbb{E}(X)^2 + \mathbb{E}(X^2) - \mathbb{E}[(\mathbb{E}(X \mid Y))^2]$$

$$= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \text{Var}(X)$$

□

Corollary 1.1.6.8.1. (Rao-Blackwell Theorem.) $\text{Var}(X) \ge \text{Var}(\mathbb{E}(X \mid Y))$

Proof. Follows immediately from Theorem 1.1.6.8 by noting that since the variance is nonnegative, $\mathbb{E}\big(\text{Var}(X \mid Y)\big) \ge 0$.

□

–

Proposition 1.1.6.9. If $c \in \mathbb{R}$, then $\text{Var}(c) = 0$.

– For the conditional variance of one Gaussian random variable on another when the covariance or correlation between them is known, see Proposition 1.3.7.1. For the conditional variance of a set of Gaussian random variables on another set when the covariance matrix is known, see Proposition 1.3.7.2.

- Covariance

–

Definition 1.1.25. $\text{Cov}(X,Y) = \mathbb{E}\big[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big]$

–

Proposition 1.1.6.10. (Useful reformulation): $\text{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$

–

Theorem 1.1.6.11. (Some useful results):

(a) $\text{Cov}(aX, BY) = ab\text{Cov}(X, Y)$
(b) $\text{Cov}(X, X) = \text{Var}(X)$
(c) $\text{Cov}(aX + bY) = ac\text{Var}(X) + bd\text{Var}(Y) + (ad + bc)\text{Cov}(X, Y)$
(d) $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$
(e) $\text{Cov}\left(\sum_{i=1}^n X_i, Y\right) = \sum_{i=1}^n \text{Cov}(X_i, Y)$
(f) $\text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j)$

–

Definition 1.1.26. Conditional covariance:

$$\text{Cov}(X, Y \mid Z) = \mathbb{E}(XY \mid Z) - \mathbb{E}(X \mid Z)\mathbb{E}(Y \mid Z) = \mathbb{E}\big[(X - \mathbb{E}(X \mid Z))(Y - \mathbb{E}(Y \mid Z)) \mid Z\big]$$

–

Theorem 1.1.6.12. Law of Total Covariance:

$$\text{Cov}(X, Y) = \mathbb{E}\big(\text{Cov}(X, Y \mid Z)\big) + \text{Cov}\big(\mathbb{E}(X \mid Z), \mathbb{E}(Y \mid Z)\big)$$

1.1.7 Discrete Random Variable Distributions

Binomial: Binomial(n, p) (sum of n Bernoulli random variables)

- Mass function: $\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$
- Distribution: $\Pr(X \leq k) = \sum_{i=0}^k \binom{n}{i} p^i (1 - p)^{n-i}$
- Expectation: $\mathbb{E}(X) = np$
- Variance: $\text{Var}(X) = np(1 - p)$

For an interesting result relating binomial and Poisson random variables, see Proposition 1.1.7.6.

Multinomial:

Definition 1.1.27 (Multinomial$(n, p_1, p_2, \ldots, p_{r-1})$ **distribution**)**.** Suppose that n independent trials, each of which results in either outcome $1, 2, \ldots, r$ with respective probabilities $p_1, p_2, \ldots, p_r$ (with $\sum_i p_i = 1$), are performed. Let N_i denote the number of trials resulting in outcome i. Then the joint distribution of $N_1, \ldots, N_r$ is called the **multinomial distribution.**

- Mass function:

$$\Pr(\boldsymbol{N} = \boldsymbol{\mathcal{N}}) = \binom{n}{\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_r} \prod_{i=1}^r p_i^{\mathcal{N}_i} = \frac{n!}{\mathcal{N}_1! \mathcal{N}_2! \ldots \mathcal{N}_r!} \prod_{i=1}^r p_i^{\mathcal{N}_i}$$

Proof. For $r = 2$, we have the binomial distribution: $\Pr(N_1 = x_1 = \binom{n}{x_1} p_1^{x_1}(1-p_1)^{n-x_1}$; $\Pr(N_2 = x_2 = \binom{n}{x_2} p_2^{x_2}(1-p_2)^{n-x_2}$. But in the general case, it is still true that N_i is a binomial random variable for all i. So in general we have $\Pr(N_i = x_i) = \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n_i - x_i}$, $i\ in\{1, 2, \ldots, r\}$.

We would like the distribution of the random vector $\boldsymbol{N} = (N_1, \ldots, N_r)$. First consider the case where $n = 1$; that is, the Multinouli distribution. Note that in this case, we have $\Pr(\boldsymbol{N} = \boldsymbol{\mathcal{N}}) = \prod_{i=1}^r p_i^{\mathcal{N}_i}$, where $\mathcal{N}_i$ are the entries of the observed vector $\boldsymbol{\mathcal{N}}$ ($\mathcal{N}_i \in \{0, 1\}$.

Now consider an arbitrary $n \in \mathbb{N}$. Then $\boldsymbol{N}$ is the sum of n Multinouli random variables. Just as before, the probability of a particular $\boldsymbol{\mathcal{N}}$ obtained in a particular ordering is $\Pr(\boldsymbol{N} = \boldsymbol{\mathcal{N}}) = \prod_{i=1}^r p_i^{\mathcal{N}_i}$. However, we must also consider the number of possible orderings in which these successes could have occurred. This number of orderings is exactly equal to the multinomial coefficient,

$$\binom{n}{\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_r} = \frac{n!}{\mathcal{N}_1! \mathcal{N}_2! \ldots \mathcal{N}_r!}$$

Therefore the joint probability mass function for $\boldsymbol{N}$ is

$$\Pr(\boldsymbol{N} = \boldsymbol{\mathcal{N}}) = \binom{n}{\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_r} \prod_{i=1}^r p_i^{\mathcal{N}_i} = \frac{n!}{\mathcal{N}_1! \mathcal{N}_2! \ldots \mathcal{N}_r!} \prod_{i=1}^r p_i^{\mathcal{N}_i}$$

□

- Distribution: $\Pr(X \leq k) =$
- Expectation: $\mathbb{E}(X) =$
- Variance: $\text{Var}(X) =$

Proposition 1.1.7.1. $\text{Cov}(N_i, N_j) = -np_ip_j$.

Proof. For a Multinoulli random variable (Multinomial$(1, p_1, \ldots, p_{r-1})$) , we have

$$\text{Cov}(N_i, N_j) = \mathbb{E}(N_iN_j) - \mathbb{E}(N_i)\mathbb{E}(N_j) = 0 - p_ip_j = -p_ip_j$$

(where $\mathbb{E}(N_iN_j) = 0$ because at least one of them must equal 0). Since a multinomial random variable is the sum of n independent Multinoulli random variables, in the general case we have

$$\text{Cov}(N_i, N_j) = -np_ip_j.$$

□

Proposition 1.1.7.2. Let $X = (X_1, \ldots, X_k)$ have multinomial distribution Multinomial$(n; \theta_1, \ldots, \theta_k)$. Then the maximum likelihood estimator of $\theta = (\theta_1, \ldots, \theta_k)$ is

$$\hat{\theta}^{(mle)} = \begin{bmatrix} \hat{\theta}_1^{(mle)} \\ \vdots \\ \hat{\theta}_k^{(mle)} \end{bmatrix} = \begin{bmatrix} x_1/n \\ \vdots \\ x_k/n \end{bmatrix}.$$

Proof. We have for all $\{X \in \mathbb{Z}_+^k : \sum_{i=1}^k X_i = n\}$

$$\Pr(X = (x_1, \ldots, x_k)) = \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^k \theta_i^{x_i} = \frac{n!}{x_1! x_2! \ldots x_k!} \prod_{i=1}^k \theta_i^{x_i}$$

so the log likelihood is

$$\ell_n(\theta) = \log\left(\frac{n!}{x_1! x_2! \ldots x_k!}\right) + \sum_{i=1}^k x_i \log(\theta_i) = \log\left(\frac{n!}{x_1! x_2! \ldots x_k!}\right) + x^T \log(\theta).$$

where $\log(\theta) = (\log(\theta_1), \ldots, \log(\theta_k))$. Then

$$\frac{\mathrm{d}}{\mathrm{d}\theta_i} \ell_n(\theta) = \frac{x_i}{\theta_i}, \qquad i \in [k]$$

But we have the constraint $g(\theta) = \sum_{i=1}^k \theta_i = 1$, so we can use Lagrange multipliers to write

$$\frac{\mathrm{d}}{\mathrm{d}\theta_i} g(\theta) = 1 \implies \begin{bmatrix} x_1/\hat{\theta}_1^{(mle)} \\ \vdots \\ x_k/\hat{\theta}_k^{(mle)} \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \iff \begin{bmatrix} \hat{\theta}_1^{(mle)} \\ \vdots \\ \hat{\theta}_k^{(mle)} \end{bmatrix} = \begin{bmatrix} \lambda x_1 \\ \vdots \\ \lambda x_k \end{bmatrix}$$

for some $\lambda \in \mathbb{R}_+$. Finally,

$$\sum_{i=1}^k \hat{\theta}_i^{(mle)} = 1 \iff \sum_{i=1}^k \lambda x_i = 1 \iff \lambda = \frac{1}{n} \iff \begin{bmatrix} \hat{\theta}_1^{(mle)} \\ \vdots \\ \hat{\theta}_k^{(mle)} \end{bmatrix} = \begin{bmatrix} x_1/n \\ \vdots \\ x_k/n \end{bmatrix}.$$

□

Poisson: Poisson(λ): an approximation of the binomial distribution for n very large, p very small, $np \to \lambda \in (0, \infty)$.

- Mass function:
$$\Pr(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$
- Distribution: $\Pr(X \leq k) = \sum_{i=0}^k \frac{e^{-\lambda} \lambda^i}{i!}$
- Expectation: $\mathbb{E}(X) = \lambda$ (derive from basic definitions)
- Variance: $\mathrm{Var}(X) = \lambda$
- Moment-generating function: $M_X(t) = e^{\lambda(e^t - 1)}$

Proposition 1.1.7.3. Let $X \sim \text{Binomial}(n, p)$. Then

$$\lim_{n\to\infty,\ p\to 0,\ np\to\lambda} X \sim \text{Poisson}(np).$$

Proof.

$$\lim_{n\to\infty,\ p\to 0,\ np\to\lambda} \binom{n}{k} p^k (1-p)^{n-k} = \lim_{n\to\infty,\ p\to 0,\ np\to\lambda} \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}$$

$$= \frac{1}{k!} \cdot \lim_{n\to\infty,\ p\to 0,\ np\to\lambda} (1-p)^{n-k} p^k \prod_{i=0}^{k-1} (n-i) = \frac{1}{k!} \cdot \lim_{n\to\infty,\ p\to 0,\ np\to\lambda} \left(1 - \frac{np}{n}\right)^{n-k} p^k \prod_{i=0}^{k-1} (n-i)$$

Using $\lim_{n\to\infty}(1-\lambda/n)^n = \exp(\lambda)$, and letting $\lambda = np$, we have

$$= \frac{\exp(-np)(np)^k}{k!} = \boxed{\frac{\exp(-\lambda)\lambda^k}{k!}} \sim \text{Poisson}(np)$$

□

Proposition 1.1.7.4. Let $X \sim \text{Poisson}(\lambda)$. If λ is sufficiently large (say $\lambda > 20$, then we can use the approximation

$$X \sim \mathcal{N}(\lambda, \lambda)$$

Proof. (Informal justification.) By Proposition 1.1.7.3, a Poisson distribution can be thought of as a close approximation of a binomial distribution. Since a binomial distribution can be approximated by a normal distribution for n large and np not too large, the same is true of a Poisson distribution. □

Proposition 1.1.7.5. Suppose $X \sim \text{Poisson}(\mu)$, $Y \sim \text{Poisson}(\nu)$, with $X \perp\!\!\!\perp Y$. Then $\{X + Y\} \sim \text{Poisson}(\mu + \nu)$.

Proof. (exercise; use characteristic functions.)

□

Proposition 1.1.7.6. Suppose $X \sim \text{Poisson}(\mu)$, $Y \sim \text{Poisson}(\nu)$, with $X \perp\!\!\!\perp Y$. Then $\{X \mid X + Y = t\} \sim \text{Bin}(t, \mu/(\mu+\nu))$

Proof.

$$\mathbb{P}_{\mu,\nu}(X = x \mid X + Y = t) = \frac{\mathbb{P}_{\mu,\nu}(X = x, Y = t - x)}{\mathbb{P}_{\mu,\nu}(X + Y = t)} = \frac{\mu^x}{x!} e^{-\mu} \cdot \frac{\nu^{t-x}}{(t-x)!} e^{-\nu} \cdot \frac{t! e^{\mu+\nu}}{(\mu+\nu)^t}$$

where we used Proposition 1.1.7.5 to get $\{X + Y\} \sim \text{Poisson}(\mu + \nu)$.

$$= \binom{t}{x} \left(\frac{\mu}{\mu+\nu}\right)^x \left(\frac{\nu}{\mu+\nu}\right)^{t-x}$$

so $\{X \mid T = t\} \sim \text{Bin}(t, \mu/(\mu+\nu))$.

□

Remark. A similar result exists when a sum of three or more Poisson random variables is known; the conditional distribution is multinomial.

Geometric: $\text{G}_1(p)$: the number of Bernoulli trials before the first success.

- Mass function: $\Pr(X = k) = p(1-p)^{k-1}$
- Distribution: $\Pr(X \leq k) = \sum_{i=1}^{k} p(1-p)^{k-1}$
- Expectation: $\mathbb{E}(X) = 1/p$
- Variance: $\text{Var}(X) = (1-p)/p^2$

Negative binomial: $\text{NB}(r, p)$: The number of Bernoulli trials required for r successes. (Can be derived as the sum of r identically distributed geometric random variables.)

- Mass function: $\Pr(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$
- Distribution: $\Pr(X \leq k) = \sum_{i=r}^{k} \binom{i-1}{r-1} p^r (1-p)^{i-r}$
- Expectation: $\mathbb{E}(X) = \frac{r}{p}$
- Variance: $\text{Var}(X) = \frac{r(1-p)}{p^2}$
- Moment-generating function:

$$M_X(t) = \frac{(pe^t)^r}{[1-(1-p)e^t]^r}$$

Hypergeometric: Hypergeometric(N, M, K): When drawing a sample of size K from a group of N items, M of which are special, X is the number of special items retrieved.

- Mass function:

$$\Pr(X = k) = \frac{\binom{M}{k}\binom{N-M}{K-k}}{\binom{N}{K}}$$

- Distribution:

$$\Pr(X \leq k) = \sum_{i=0}^{k} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}}$$

- Expectation: $\mathbb{E}(X) = \frac{MK}{N}$

  Proof. Let Y_j be an indicator variable for special item j being selected. Note that $X = \sum_{j=1}^{M} Y_j$ and that

$$\mathbb{E}(Y_j) = \frac{\binom{1}{1}\binom{N-1}{K-1}}{\binom{N}{K}} = \frac{1 \cdot \frac{(N-1)!}{(N-K)!(K-1)!}}{\frac{N!}{(N-K)!K!}} = \frac{K}{N},$$

so we have

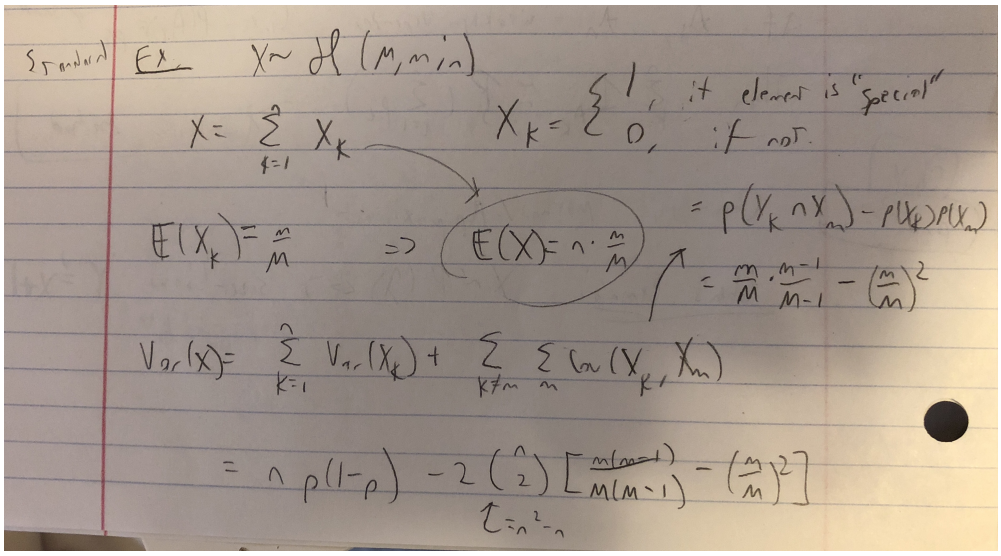
$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{j=1}^{M} Y_j\right) = \sum_{j=1}^{M} \mathbb{E}(Y_j) = \frac{MK}{N}.$$

Alternative proof:

Let Z_i be an indicator variable for the ith selected item to be special. Then $X = \sum_{i=1}^{K} Z_i$ and $\mathbb{E}(Z_i) = \frac{M}{N}$, so we have

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{K} Z_i\right) = \sum_{i=1}^{K} \mathbb{E}(Z_i) = \frac{MK}{N}.$$

Variance: $\mathrm{Var}(X) =$ (find by indicator method. Proof from notes, I think this may have errors:)



□

Proposition 1.1.7.7. Let $X \sim \text{Hypergeometric}(N, M, K)$. Then

$$\lim_{M,N\to\infty, M/N\to p} \Pr(X = k) \sim \text{Binomial}(K, p)$$

Proof.

$$\lim_{M,N\to\infty,M/N\to p} p_k(M,N,K) = \lim_{M,N\to\infty,M/N\to p} \frac{\binom{M}{k}\binom{N-M}{K-k}}{\binom{N}{K}}$$

$$= \lim_{M,N\to\infty,M/N\to p} \frac{M!(N-M)!/[k!(M-k)!(K-k)!(N-M-K+k)!]}{N!/[K!(N-K)!]}$$

$$= \lim_{M,N\to\infty,M/N\to p} \frac{K!}{(K-k)!k!} \cdot \frac{M!(N-M)!(N-K)!}{N!(M-k)!(N-M-K+k)!}$$

$$= \lim_{M,N\to\infty,M/N\to p} \binom{K}{k} \cdot \frac{M!/(M-k)!}{N!/(N-k)!} \cdot \frac{(N-M)!(N-K)!}{(N-k)!(N-M-(K-k))!}$$

$$= \binom{K}{k} \lim_{M,N\to\infty, M/N\to p} \frac{M!/(M-k)!}{N!/(N-k)!} \cdot \frac{(N-M)!/(N-M-(K-k))!}{(N-K+(K-k))!/(N-K)!}$$

$$= \binom{K}{k} \lim_{M,N\to\infty, M/N\to p} \prod_{i=0}^{k-1} \frac{M-i}{N-i} \cdot \prod_{j=0}^{K-k-1} \frac{N-M-j}{N-K+1+j}$$

$$= \binom{K}{k} \left(\frac{M}{N}\right)^k \left(\frac{N-M}{N}\right)^{K-k}$$

$$= \binom{K}{k} \left(\frac{M}{N}\right)^k \left(1-\frac{M}{N}\right)^{K-k} = \binom{K}{k} p^k (1-p)^{K-k}$$

□

1.1.8 Indicator Method

Proposition 1.1.8.1. If $\mathbf{1}_{A_k}$ is an indicator then

(a)

$$\text{Cov}(\mathbf{1}_{A_k}, \mathbf{1}_{A_m}) = \mathbb{E}(\mathbf{1}_{A_k}\mathbf{1}_{A_m}) - \mathbb{E}(\mathbf{1}_{A_k})\mathbb{E}(\mathbf{1}_{A_m}) = \Pr(A_k \cap A_m) - \Pr(A_k)\Pr(A_m)$$

(b)

$$\text{Var}(\mathbf{1}_{A_k}) = \mathbb{E}(\mathbf{1}_{A_k}^2) = \mathbb{E}(\mathbf{1}_{A_k})^2 = \Pr(A_k) - (\Pr(A_k))^2$$

Theorem 1.1.8.2. X is independent of Y if and only if X is independent of $\mathbf{1}_A$, $A \in Y$.

Example problems: 505A Homework 3 problem 9(a)

Worked examples in p. 56 - 59 of Grimmett and Stirzaker 3rd edition.

1.1.9 Linear transformations of random variables

1.1.10 Poisson Paradigm (Poisson approximation for indicator method)

Theorem 1.1.10.1. (Theorem 4.12.9, p. 129 of Grimmett and Stirzaker.) Let A_i be an event. If $X = \sum_{i=1}^m \mathbf{1}_{A_i}$ where $\mathbf{1}_{A_i}$ is an indicator variable for A_i, and the A_i are only weakly dependent on each other, then

$$\text{As } m \to \infty, \quad X \sim \text{Poisson}(\mathbb{E}(X))$$

More specifically, let B_i be n independent Bernoulli random variables with probabilities p_i. If $Y = \sum_{i=1}^n B_i$ then

$$\text{As } n \to \infty, \quad Y \sim \text{Poisson}\left(\mathbb{E}\left(\sum_i B_i\right)\right) = \text{Poisson}\left(\sum_i \mathbb{E}B_i\right) = \text{Poisson}\left(\sum_i p_i\right)$$

Proof. Full proof available in Grimmett and Stirzaker, section 4.12, page 129. A justification of the first claim is as follows: if the A_i are independent and $\Pr(A_i) = p \; \forall i$, then $X \sim \text{Binomial}(m, p)$. Then by Proposition 1.1.7.3, the result follows. It turns out that this result holds up if the probabilities are not necessarily identical (but all small) and the variables are not necessarily independent (but only weakly dependent). □

Solution Alternative Solution to Exercise 1 (Matching Problem):

Let X be the number of matches. Let $P_n = \Pr(X = 0)$ given that there are n people. Let Y be an indicator variable for the first person who receives their sandwich receiving the correct one. Note that

$$P_n = \Pr(X = 0) = \Pr(X = 0 \mid Y = 1)(1/n) + \Pr(X = 0 \mid Y = 0)(n-1)/n.$$

Note that if the first person didn't match ($Y = 0$), we have $n-2$ people with their hats left, but one of the remaining $n-1$ people can't get their sandwich because it was taken by the first person. Therefore

$\Pr(X = 0 \mid Y = 0) = \Pr(\{\text{the extra person selects the first person's hat, and the rest of the people pick the wrong hat}\})$

$+ \Pr(\{\text{the extra person does not select the first person's hat, and the rest of the people don't have any matches}\})$

$$= \frac{1}{n-1} \cdot P_{n-2} + P_{n-1}$$

This yields

$$P_n = \frac{1}{n} P_{n-2} + \frac{n-1}{n} P_{n-1} \iff P_n - P_{n-1} = -\frac{1}{n}(P_{n-1} - P_{n-2})$$

which is a recursive formula. Now we seek to find a closed form solution. We have

$$P_3 - P_2 = -\frac{1}{3}(P_2 - P_1) = -\frac{1}{3!} \iff P_3 = P_2 - \frac{1}{3!} = \frac{1}{2!} - \frac{1}{3!}$$

$$P_4 - P_3 = -\frac{1}{4}(P_3 - P_2) = \frac{1}{4!} \iff P_4 = P_3 + \frac{1}{4!} = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!}$$

$$\vdots$$

$$P_n = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \ldots + \frac{(-1)^n}{n!} = \sum_{k=0}^{n} \frac{(-1)^k}{k!} \to e^{-1} \text{ as } n \to \infty$$

so we see that for large n we approximately have $P_n \sim \text{Poisson}(1)$.

Now we consider $\Pr(X = k)$. Consider a set of k people who all have matches and no one else matches. The probability that everyone in a set of k people match is $(n-k)!/n!$. The probability that none of the other $n-k$ people match is P_{n-k} per above. Since there are $\binom{n}{k}$ ways to chose groups of k people, we have

$$\Pr(X = k) = \binom{n}{k}\frac{(n-k)!}{n!} \cdot P_{n-k} = \frac{P_{n-k}}{k!} \xrightarrow{n\to\infty} \frac{e^{-1}1^k}{k!}$$

so again we see that for large n we approximately have $P_n \sim \text{Poisson}(1)$.

1.1.11 Asymptotic Distributions

Proposition 1.1.11.1.

$$e^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$$

Theorem 1.1.11.2. Stirling's Formula:

$$n! \sim n^n e^{-n}\sqrt{2\pi n}$$

That is,

$$\lim_{n\to\infty}\frac{n^n e^{-n}\sqrt{2\pi n}}{n!} = 1.$$

1.2 Worked problems

1.2.1 Example Problems That Will Likely Appear on Midterm (and Final)

(1) (a) **Fall 2011 Problem 1 (same as HW1 problem 5; similar to HW3 problem 2(5).)** Let A and B be events such that $0 < \Pr(A) < 1$. Show that if $\Pr(B \mid A) = \Pr(B \mid A^c)$, then A and B are independent.

(b) Let X and Y be two discrete random variables, each taking only two possible values. Show that if $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ then X and Y are independent.

Solution.

(a) A and B are independent if and only if

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

We know that

$$\Pr(B) = \Pr(B|A) \cdot \Pr(A) + \Pr(B|A^c) \cdot \Pr(A^c)$$

$$= \Pr(B|A) \cdot \Pr(A) + \Pr(B|A) \cdot (1 - \Pr(A)) = \Pr(B|A) \cdot \Pr(A) + \Pr(B|A) - \Pr(B|A) \cdot \Pr(A)$$

$$= \Pr(B|A)$$

Also, we know that since $\Pr(A) \neq 0$,

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

Per above $\Pr(B|A) = \Pr(B)$, so we have

$$\Pr(B) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

which is what we were trying to prove. So the answer is $\boxed{\text{true.}}$

(b) Without loss of generality, let X and Y have mass functions

$$X = \begin{cases} x_1 & \text{with probability } \Pr(A) \\ x_2 & \text{with probability } \Pr(A^c) \end{cases}$$

$$Y = \begin{cases} y_1 & \text{with probability } \Pr(B) \\ y_2 & \text{with probability } \Pr(B^c) \end{cases}$$

Then $X \perp\!\!\!\perp Y \iff \Pr(A \cap B) = \Pr(A)\Pr(B)$. Let $\alpha = X - x_2$, $\beta = Y - y_2$; that is,

$$\alpha = \begin{cases} x_1 - x_2 & \text{with probability } \Pr(A) \\ 0 & \text{with probability } \Pr(A^c) \end{cases}$$

$$\beta = \begin{cases} y_1 - y_2 & \text{with probability } \Pr(B) \\ 0 & \text{with probability } \Pr(B^c) \end{cases}$$

Then we have

- $\mathbb{E}(\alpha) = (x_1 - x_2)\Pr(A)$
- $\mathbb{E}(\beta) = (y_1 - y_2)\Pr(B)$
- $\mathbb{E}(\alpha\beta) = (x_1 - x_2)(y_1 - y_2)\Pr(A \cap B)$

which we can use to obtain

$$\mathbb{E}(XY) = \mathbb{E}\big[(\alpha + x_2)(\beta + y_2)\big] = \mathbb{E}(\alpha\beta) + y_2\mathbb{E}(\alpha) + x_2\mathbb{E}(\beta) + x_2y_2$$

$$= (x_1 - x_2)(y_1 - y_2)\Pr(A \cap B) + y_2(x_1 - x_2)\Pr(A) + x_2(y_1 - y_2)\Pr(B) + x_2y_2 \tag{1.6}$$

$$\mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(\alpha + x_2)\mathbb{E}(\beta + y_2) = \big[x_2 + (x_1 - x_2)\Pr(A)\big]\big[y_2 + (y_1 - y_2)\Pr(B)\big]$$

$$= x_2y_2 + x_2(y_1 - y_2)\Pr(B) + y_2(x_1 - x_2)\Pr(A) + (x_1 - x_2)(y_1 - y_2)\Pr(A)\Pr(B) \tag{1.7}$$

Using $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, we set (1.6) and (1.7) equal to each other. Canceling terms appearing in both yields

$$(x_1 - x_2)(y_1 - y_2)\Pr(A \cap B) = (x_1 - x_2)(y_1 - y_2)\Pr(A)\Pr(B) \iff \Pr(A \cap B) = \Pr(A)\Pr(B)$$

which proves the independence of X and Y if $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

(2) **Fall 2013 Qual Problem 1.** Consider a sequence of independent tosses of a pair of fair dice. Compute the probability that the sum 4 will occur before the sum 5.

Solution. Let Y_k be the outcome of the kth toss. Let X_{k1} be the number of the first die on the kth toss and X_{k2} be the outcome of the second die. Note that

$$\Pr(Y_k = 4) = \Pr(X_{k1} = 1 \cap X_{k2} = 3) + \Pr(X_{k1} = 2 \cap X_{k2} = 2) + \Pr(X_{k1} = 3 \cap X_{k2} = 1) = 3 \cdot \frac{1}{36} = \frac{1}{12}$$

$$\Pr(Y_k = 5) = \Pr(X_{k1} = 1 \cap X_{k2} = 4) + \Pr(X_{k1} = 2 \cap X_{k2} = 3) + \Pr(X_{k1} = 3 \cap X_{k2} = 2)$$

$$+ \Pr(X_{k1} = 4 \cap X_{k2} = 1) = 4 \cdot \frac{1}{36} = \frac{1}{9}$$

Let A_k be the event that $Y_k = 4$ and $Y_j \neq 4$ or 5, $j = 1, \ldots, k-1$. Note that all A_k are mutually exclusive and

$$\Pr(A_k) = \frac{1}{12} \cdot \left(1 - \frac{3+4}{36}\right)^{k-1} = \frac{1}{12} \cdot \left(\frac{29}{36}\right)^{k-1}.$$

Then

$$\Pr(\{\text{roll a 4 before a 5}\}) = \Pr\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \Pr(A_k) = \frac{1}{12}\sum_{k=1}^{\infty}\left(\frac{29}{36}\right)^{k-1} = \frac{1/12}{1 - 29/36}$$

$$= \frac{3}{36 - 29} = \boxed{\frac{3}{7}}$$

(3) **Spring 2013 Problem 1, in notes from 09/21.** Let X and Y be random variables such that $\mathbb{E}(X \mid Y) = Y, \mathbb{E}(Y \mid X) = X, \mathbb{E}(X^2) < \infty, \mathbb{E}(Y^2) < \infty$. Show that $\mathbb{E}(X - Y)^2 = 0$ (or equivalently, show $\Pr(X = Y) = 1$).

Solution.

$$\mathbb{E}(X - Y)^2 = \mathbb{E}(X^2 - 2XY + Y^2) = \mathbb{E}(X^2) - 2\mathbb{E}(XY) + \mathbb{E}(Y^2)$$

$$\mathbb{E}(XY) = \mathbb{E}(\mathbb{E}(XY \mid Y)) = \mathbb{E}(Y\mathbb{E}(X \mid Y)) = \mathbb{E}(Y \cdot Y) = \mathbb{E}(Y^2)$$

Also,

$$\mathbb{E}(XY) = \mathbb{E}(\mathbb{E}(XY \mid X)) = \mathbb{E}(X\mathbb{E}(Y \mid X)) = \mathbb{E}(X \cdot X) = \mathbb{E}(X^2)$$

Therefore

$$\mathbb{E}(X - Y)^2 = 0$$

(4) **Fall 2016 Problem 4.** Consider a group of $n \geq 4$ people, among whom are Alice, Bob, Charles, and Diana, standing in a row. Assume that all possible orderings of the n people are equally likely.

(a) Compute the probability that Charles stands somewhere between Alice and Bob. (Note: this does not mean that the three are necessarily adjacent; there can be other people between Alice and Bob.)

(b) Compute the probability that Diana stands somewhere between Alice and Bob given that Charles stands somewhere between Alice and Bob.

(c) Let X be the number of people who stand between Alice and Bob. Compute the expected value and the variance of X. (Note: Alice and Bob themselves are not counted in this number.)

Solution.

(a) For this part, n does not make a difference; all we need to know is the ordering of A, B, and C. This is because conditional on a specific ordering of A, B, and C, all arrangements of everyone else are equally likely; and conversely, the a particular ordering of A, B, and C is independent of which three particular slots are made available to them. Examining the permutations of A, B, and C, two of them have C in the middle, so the answer is $2/6 = \boxed{1/3}$.

(b) Similarly, the answer is independent of n, so we work with $n = 4$. All possible orderings with Charles between Alice and Bob are as follows:

$$ACDB,\ ADCB,\ BCDA,\ BDCA,\ ACBD,\ BCAD,\ DACB,\ DBCA$$

The first four of these have Diana between Alice and Bob, so the answer is $4/8 = \boxed{1/2}$.

(c) Let I_k be an indicator variable for the event that person k is between A and B. By the result from part (a), $\mathbb{E}(I_k) = 1/3$. Then we have

$$\text{Var}(I_k) = \mathbb{E}(I_k^2) - \mathbb{E}(I_k)^2 = 1^2 \cdot \Pr(I_k = 1) - \frac{1}{9} = \frac{1}{3} - \frac{1}{9} = \frac{2}{9}$$

Noting that the four arrangements above with Charles and Diana in between Alice and Bob are the only ones where this will be the case of the $4! = 24$ possible orderings, we have

$$\mathbb{E}(I_k I_j) = \frac{4}{24} = \frac{1}{6}$$

so

$$\text{Cov}(I_j, I_k) = \mathbb{E}(I_k I_j) - \mathbb{E}(I_i)\mathbb{E}(I_j) = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}$$

Therefore

$$\mathbb{E}(X) = \mathbb{E}(\sum_{i=1}^{n-2} I_k) = \sum_{i=1}^{n-2} \frac{1}{3} = \frac{n-2}{3}$$

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n-2} I_k\right) = \sum_{i=1}^{n-2} \text{Var}(I_k) + 2 \sum_{1 \leq j < k \leq n-2} \text{Cov}(I_j, I_k) = \frac{2(n-2)}{9} + \left[(n-2)^2 - (n-2)\right] \cdot \frac{1}{18}$$

$$= \frac{4(n-2)}{18} + \frac{(n-2)(n-3)}{18} = \boxed{\frac{(n-2)(n+1)}{18}}$$

(5) **Spring 2015 Problem 2.** A deck of 52 cards is shuffled thoroughly. Someone goes through all 52 cards, scoring 1 point each time 2 cards of the same value are consecutive (that is, when two consecutive cards have the same rank but different suits). For example, the sequence 9H, 8H, 7D, 6C, 7S, 7H, 7C scores 2 points, one for the 7H following the 7S, one for the 7C following the 7H. Let X be the total score.

(a) Compute $\mathbb{E}(X)$.

(b) Compute $\Pr(X = 39)$. (Note that there are 13 different ranks and you cannot score more than 3 per rank.)

(c) In the line below, circle the number that you think is the closest to the value $\Pr(X = 0)$ and briefly explain your choice:

$$\frac{1}{1000}, \quad \frac{1}{500}, \quad \frac{1}{100}, \quad \frac{1}{50}, \quad \frac{1}{20}, \quad \frac{1}{10}, \quad \frac{1}{5}, \quad \frac{1}{2}$$

Solution.

(a) Start by assuming the permutation is cyclic (that is, after the last card you go back to the beginning). Let Y be the number of matches in this situation. Let A_i be the event that the ith card is followed by a match. Then $\Pr(A_i) = 3/51 = 1/17$, so

$$Y = \sum_{k=1}^{52} \mathbf{1}_{\{A_i\}} \implies \mathbb{E}(Y) = \sum_{i=1}^{52} \mathbb{E}(\mathbf{1}_{\{A_i\}}) = \sum_{i=1}^{52} \Pr(A_i) = 52 \cdot 1/17 = 52/17$$

Note that $\mathbb{E}(Y) = \mathbb{E}(X) + \mathbb{E}(\mathbf{1}_{\{A_i\}})$ because if the permutation is cyclic then you have one extra opportunity to mach at the end.

$$\implies \mathbb{E}(X) = \frac{52}{17} - \frac{1}{17} = \frac{51}{17} = \boxed{3}$$

(b) $\Pr(X = 39) = \Pr(\{\text{all possible matches occur}\})$, so in this event all cards of the same rank are clustered together. There are 13 clusters of 4 cards, so there are 13! ways to order the clusters and 4! ways to order the cards within each cluster. Therefore

$$\boxed{\Pr(X = 39) = \frac{13!(4!)^{13}}{52!}}$$

(c) Because A_i are only weakly dependent and $\Pr(A_i)$ is small for all A_i, we can use the Poisson approximation (see Section 1.1.10); that is, $X \sim \text{Poisson}(\mathbb{E}(X)) = \text{Poisson}(3)$. Therefore

$$\Pr(X = 0) \approx \frac{e^{-3} \cdot 3^0}{0!} = \frac{1}{e^3} \approx \frac{1}{2.8^3} \approx \boxed{\frac{1}{20}}$$

1.2.2 Problems we did in class that professor mentioned

(Fall 2014 Problem 1) (Variation of Midterm problem 1 above) Let A and B be two events with $0 < \Pr(A) < 1,\ 0 < \Pr(B) < 1$. Define the random variables $\xi = \xi(\omega)$ and $\eta = \eta(\omega)$ by

$$\xi(\omega) = \begin{cases} 5 & \text{if } \omega \in A \\ -7 & \text{if } \omega \notin A \end{cases}, \quad \eta(\omega) = \begin{cases} 2 & \text{if } \omega \in B \\ 3 & \text{if } \omega \notin B \end{cases}$$

True or false: the events A and B are independent if and only if the random variables ξ and η are uncorrelated?

Solution. ($\implies$) Suppose A and B are independent. Then ξ and η are uncorrelated if and only if $\mathbb{E}(\xi\eta) = \mathbb{E}(\xi)\mathbb{E}(\eta)$. We can write $\xi = 5 \cdot \mathbf{1}_A - 7 \cdot \mathbf{1}_{A^c}$ and $\eta = 2 \cdot \mathbf{1}_B + 3 \cdot \mathbf{1}_{B^c}$. So we have

$$\xi\eta = (5 \cdot \mathbf{1}_A - 7 \cdot \mathbf{1}_{A^c})(2 \cdot \mathbf{1}_B + 3 \cdot \mathbf{1}_{B^c}) = 10 \cdot \mathbf{1}_{A\cap B} + 15 \cdot \mathbf{1}_{A\cap B^c} - 14 \cdot \mathbf{1}_{A^c\cap B} - 21 \cdot \mathbf{1}_{A^c\cap B^c}$$

$$\implies \mathbb{E}(\xi\eta) = 10\Pr(A\cap B) + 15\Pr(A\cap B^c) - 14\Pr(A^c\cap B) - 21\Pr(A^c\cap B^c)$$

Then

$$\mathbb{E}(\xi)\mathbb{E}(\eta) = (5\Pr(A) - 7\Pr(A^c))(2\Pr(B) + 3\Pr(B^c))$$

$$= 10\Pr(A\cap B) + 15\Pr(A\cap B^c) - 14\Pr(A^c\cap B) - 21\Pr(A^c\cap B^c) = \mathbb{E}(\xi\eta)$$

where the second-to-last step follows from the independence of A and B. Therefore η and ξ are uncorrelated.

($\impliedby$) Now suppose η and ξ are uncorrelated. Then ξ and η are independent if and only if $\Pr(\xi\cap\eta) = \Pr(\xi)\Pr(\eta)$. Define

$$\alpha(\omega) = \xi(\omega) + 7 = \begin{cases} 12 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}, \quad \beta(\omega) = \eta(\omega) - 3 = \begin{cases} -1 & \text{if } \omega \in B \\ 0 & \text{if } \omega \notin B \end{cases}$$

Then we have

$$(\alpha\beta)(\omega) = \begin{cases} -12 & \text{if } \omega \in A\cap B \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\mathbb{E}(\xi\eta) = \mathbb{E}\big[(\alpha - 7)(\beta + 3)\big] = \mathbb{E}(\alpha\beta) + 3\mathbb{E}(\alpha) - 7\mathbb{E}(\beta) - 21$$

$$\mathbb{E}(\xi)\mathbb{E}(\eta) = (\mathbb{E}(\alpha) - 7)(\mathbb{E}(\beta) + 3) = \mathbb{E}(\alpha)\mathbb{E}(\beta) - 7\mathbb{E}(\beta) + 3\mathbb{E}(\alpha) - 21$$

Since by assumption $\mathbb{E}(\xi\eta) = \mathbb{E}(\xi)\mathbb{E}(\eta)$, this yields $\mathbb{E}(\alpha\beta) = \mathbb{E}(\alpha)\mathbb{E}(\beta)$. But

$$\mathbb{E}(\alpha\beta) = -12\Pr(A\cap B), \qquad \mathbb{E}(\alpha)\mathbb{E}(\beta) = 12\Pr(A)(-1)\Pr(B) = -12\Pr(A)\Pr(B)$$

Therefore $\Pr(\xi\cap\eta) = \Pr(\xi)\Pr(\eta)$ and ξ and η are independent.

Exercise 1. Example (Letter/envelope matching problem; sometimes referred to as Montmort's matching problem). An assistant brings n sandwiches for n employees at a company. Each employee ordered a unique sandwich, but unfortunately the assistant forgot to ask that the sandwiches be labeled, so they are all indistinguishable, wrapped in the same paper. The assistant plans to distribute one sandwich to each employee and hope for the best. Let X be the number of sandwiches that are delivered to the correct person.

(a) What is the probability of at least one match; that is, $\Pr(X \geq 1)$?

(b) What is the probability of r correct matches?

(c) What $\mathbb{E}(X)$?

(d) What is $\text{Var}(X)$?

Solution

(a) Let A_k be an indicator variable for the event that sandwich k is matched to the correct employee. Then

$$\Pr(X \geq 1) = \Pr\left(\bigcup_{k=1}^{n} A_k\right)$$

Consider that if there are k correct matches, there are $\binom{n}{k}$ sets of k sandwiches that could be correctly distributed. Also, the probability of a particular set of k sandwiches being correctly distributed is $(n-k)!/n!$. So we have

$$\Pr(X = k) = \binom{n}{k}\frac{(n-k)!}{n!}$$

Therefore by the Inclusion-Exclusion Principle (Proposition 1.1.5.1),

$$\Pr\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n}(-1)^{k-1}\Pr(X=k) = \sum_{k=1}^{n}(-1)^{k-1}\binom{n}{k}\frac{(n-k)!}{n!} = \sum_{k=1}^{n}(-1)^{k-1}\frac{n!}{(n-k)!k!}\frac{(n-k)!}{n!}$$

$$= \sum_{k=1}^{n}\frac{(-1)^{k-1}}{k!} = \frac{(-1)^0}{0!} - \sum_{k=0}^{n}\frac{(-1)^k}{k!} = \boxed{1 - \sum_{k=0}^{n}\frac{(-1)^k}{k!}}$$

As $n \to \infty$, we have

$$1 - \sum_{k=0}^{n}\frac{(-1)^k}{k!} \to 1 - e^{-1} = \boxed{1 - \frac{1}{e}}$$

(b) Clearly there is only one way to match all the sandwiches correctly, so $\Pr(X = r \mid r = n) = 1/n!$. Also, note that it is impossible to match all but one sandwich, so $\Pr(X = r \mid r = n-1) = 0$. Only the cases for $r \leq n-2$ are nontrivial. Using a similar argument as part (a), we see that for any set of m sandwiches, the probability that at least one was correctly distributed is

$$\Pr\left(\bigcup_{k=1}^{m} A_k\right) = \sum_{k=1}^{m} (-1)^{k-1} \frac{(n-k)!}{n!}$$

and that the probability that *any* set of m sandwiches contained at least one correct match is

$$\sum_{k=1}^{m} (-1)^{k-1} \binom{n}{k} \frac{(n-k)!}{n!} = \sum_{k=1}^{m} (-1)^{k-1} \frac{n!}{(n-k)!k!} \frac{(n-k)!}{n!}$$

$$= \sum_{k=1}^{m} \frac{(-1)^{k-1}}{k!} = 1 + \sum_{k=2}^{m} \frac{(-1)^{k-1}}{k!} = 1 - \sum_{k=2}^{m} \frac{(-1)^{k}}{k!}$$

So for $m \geq 2$, the probability of *no* correct matches is $\sum_{k=2}^{m} \frac{(-1)^k}{k!}$ if $n \geq 2$, and of course 0 if $n = 1$. Therefore the probability of r matches is the probability of any one set of r sandwiches all matching and none of the remaining $n - r$ sandwiches matching times the number of sets of r sandwiches; that is,

$$\Pr(X = r \mid r \leq n-2) = \binom{n}{r} \cdot \frac{(n-r)!}{n!} \cdot \left(\sum_{k=2}^{n-r} \frac{(-1)^k}{k!}\right) = \frac{r!}{(n-r)!r!} \cdot \frac{(n-r)!}{n!} \sum_{k=2}^{n-r} \frac{(-1)^k}{k!}$$

$$= \frac{1}{r!} \sum_{k=2}^{n-r} \frac{(-1)^k}{k!}$$

Therefore we have

$$\boxed{\Pr(X = r) = \begin{cases} \frac{1}{r!} \sum_{k=2}^{n-r} \frac{(-1)^k}{k!} & r \leq n-2 \\ 0 & r = n-1 \\ \frac{1}{r!} & r = n \end{cases}}$$

(c)

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} \mathbb{E}(A_k) = \sum_{k=1}^{n} \Pr(A_k = 1) = n \cdot \frac{1}{n} = \boxed{1}$$

(d)

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

$$\mathbb{E}(X^2) = \mathbb{E}\left(\sum_{k=1}^{n} A_k\right)^2 = \mathbb{E}\left(\sum_{k=1}^{n} A_k^2 + 2 \sum_{1 \leq i < j \leq n} A_i A_j\right) = \sum_{k=1}^{n} \mathbb{E}(A_k^2) + 2 \sum_{1 \leq i < j \leq n} \mathbb{E}(A_i A_j)$$

Because

$$\mathbb{E}(A_k^2) = 1^2 \cdot \Pr(A_k = 1) = \frac{1}{n}$$

$$\mathbb{E}(A_i A_j) = \Pr(A_j = 1 \cap A_j = 1) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)} = \frac{1}{2} \cdot \binom{n}{2}^{-1}$$

we have

$$\mathbb{E}(X^2) = \sum_{k=1}^{n} \frac{1}{n} + 2 \sum_{1 \leq i < j \leq n} \frac{1}{n(n-1)} = 1 + 2 \cdot \binom{n}{2} \cdot \frac{1}{2} \cdot \binom{n}{2}^{-1} = 2$$

$$\implies \text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 2 - 1 = \boxed{1}$$

HW3 Problem 2(5). Verify: $\mathbb{E}(X \mid Y) = \mathbb{E}(X)$ if X and Y are independent.

Solution. X and Y are independent if and only if

$$\Pr(X \cap Y) = \Pr(X) \cdot \Pr(Y) \iff \Pr(X = x \cap Y = y) = \Pr(X = x) \Pr(Y = y)$$

$$\iff \Pr(X = x \mid Y = y) \cdot \Pr(Y = y) = \Pr(X = x) \Pr(Y = y) \iff \Pr(X = x \mid Y = y) = \Pr(X = x)$$

$$\implies E(X \mid Y) = \sum_x x \cdot \Pr(X = x \mid Y = y) = \sum_x x \cdot \Pr(X = x) = \mathbb{E}(X)$$

HW3 Problem 2 (parts 1 - 4). Verify:

(1) $\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}(X)$

(2) $\mathbb{E}(g(Y)X \mid Y) = g(Y)\mathbb{E}(X \mid Y)$

(3) $\text{Cov}(\mathbb{E}(X \mid Y), Y) = \text{Cov}(X, Y)$

(4) Y and $X - \mathbb{E}(X \mid Y)$ are uncorrelated.

Solution.

(1)

$$\mathbb{E}\big(\mathbb{E}(X \mid Y)\big) = \sum_y \mathbb{E}(X \mid Y) \Pr(Y = y) = \sum_y \left[\sum_x x \cdot \Pr(X = x \mid Y = y) \Pr(Y = y) \right]$$

$$= \sum_y \left[\sum_x x \cdot \Pr(X = x \cap Y = y) \right] = \sum_y \left[\sum_x x \cdot \Pr(Y = y \mid X = x) \cdot \Pr(X = x) \right]$$

$$= \sum_x \left[x \cdot \Pr(X = x) \cdot \sum_y \big(\Pr(Y = y \mid X = x) \big) \right] = \sum_x \left[x \cdot \Pr(X = x) \cdot 1 \right]$$

$$= \mathbb{E}(X)$$

(2) 2

(3)

$$\text{Cov}(\mathbb{E}(X \mid Y), Y) = \mathbb{E}\Big(\Big[\mathbb{E}(X \mid Y) - \mathbb{E}\big(\mathbb{E}(X \mid Y)\big)\Big]\Big[Y - \mathbb{E}(Y)\Big]\Big)$$

$$= \mathbb{E}\Big(\Big[\mathbb{E}(X \mid Y) - \mathbb{E}(X)\Big]\Big[Y - \mathbb{E}(Y)\Big]\Big) = \mathbb{E}\Big(\mathbb{E}(X \mid Y)Y - \mathbb{E}(X)Y - \mathbb{E}(X \mid Y)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y)\Big)$$

$$= \mathbb{E}\big(\mathbb{E}(X \mid Y)Y\big) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)\mathbb{E}\big(\mathbb{E}(X \mid Y)\big) + \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}\big(\mathbb{E}(X \mid Y)Y\big) - \mathbb{E}(Y)\mathbb{E}(X)$$

$$= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \text{Cov}(X, Y)$$

(4) Y and $X - \mathbb{E}(X \mid Y)$ are uncorrelated if and only if $\text{Cov}\big(Y, X - \mathbb{E}(X \mid Y)\big) = 0 \iff \mathbb{E}(Y \cdot \big[X - \mathbb{E}(X \mid Y)\big]) - \mathbb{E}(Y)\mathbb{E}(X - \mathbb{E}(X \mid Y)) = 0$.

$$\mathbb{E}(Y \cdot \big[X - \mathbb{E}(X \mid Y)\big]) - \mathbb{E}(Y)\mathbb{E}(X - \mathbb{E}(X \mid Y)) = \mathbb{E}(YX - Y\mathbb{E}(X \mid Y)) - \mathbb{E}(Y)\mathbb{E}(X) + \mathbb{E}(Y)\mathbb{E}\big(\mathbb{E}(X \mid Y)\big)\big)$$

$$= \mathbb{E}(YX) - \mathbb{E}\big(Y\mathbb{E}(X \mid Y)\big) - \mathbb{E}(Y)\mathbb{E}(X) + \mathbb{E}(Y)\mathbb{E}(X) = \mathbb{E}(YX) - \mathbb{E}(YX) = 0$$

Spring 2018 Problem 2 (did not complete)

2. Consider positions 1 to n arranged in a circle, so that 2 comes after 1, 3 comes after 2, ..., n comes after $n-1$, and 1 comes after n. Similarly, take 1 to n as values, with cyclic order, and consider all $n!$ ways to assign values to positions, bijectively, with all $n!$ possibilities equally likely. For $i = 1$ to n, let X_i be the indicator that position i and the one following are filled in with two consecutive values in increasing order, and define

$$S_n = \sum_{i=1}^{n} X_i, \quad T_n = \sum_{i=1}^{n} iX_i$$

For example, with $n = 6$ and the the circular arrangement 314562, we get $X_3 = 1$ since 45 are consecutive in increasing order, and similarly $X_4 = X_6 = 1$, so that $S_6 = 3, T_6 = 13$.

a) Compute the mean and the variance of S_n.

b) Compute the mean and the variance of T_n.

Fall 2008 Problem 2 (HW1 Problem 10). Consider a lottery with n^2, tickets, of which only n tickets win prizes. Let p_n be the probability that, out of n randomly selected tickets, at least one wins a prize. Compute $\lim_{n\to\infty} p_n$.

Solution. There are $\binom{n^2}{n}$ possible sets of n tickets. The number of these sets that do not contain at least one winner (that is, they only contain members of the $n^2 - n$ losing tickets) is $\binom{n^2-n}{n}$. Therefore the probability of selecting a set of n tickets that contains at least one winner is

$$p_n = 1 - \binom{n^2-n}{n} \Big/ \binom{n^2}{n} = 1 - \frac{(n^2-n)!}{n!(n^2-n-n)!} \Big/ \frac{(n^2)!}{(n^2-n)!n!} = 1 - \frac{(n^2-n)!}{n!(n^2-2n)!} \cdot \frac{(n^2-n)!n!}{(n^2)!}$$

$$= 1 - \frac{(n^2-n)!}{(n^2-2n)!} \cdot \frac{(n^2-n)!}{(n^2)!} = 1 - \prod_{i=0}^{n-1}(n^2-n-i) \Big/ \prod_{i=0}^{n-1}(n^2-i) = 1 - \prod_{i=0}^{n-1} \frac{n^2-n-i}{n^2-i}$$

$$= 1 - \prod_{i=0}^{n-1}\left(\frac{n^2-i}{n^2-i} - \frac{n}{n^2-i}\right) = 1 - \prod_{i=0}^{n-1}\left(1 - \frac{n}{n^2-i}\right)$$

Therefore

$$\lim_{n\to\infty} p_n = \lim_{n\to\infty}\left[1 - \prod_{i=0}^{n-1}\left(1 - \frac{n}{n^2-i}\right)\right] = 1 - \lim_{n\to\infty}\prod_{i=0}^{n}\left(1 - \frac{n}{n^2-i}\right) = 1 - \lim_{n\to\infty}\prod_{i=0}^{n}\left(1 - \frac{n \cdot \frac{1}{n}}{\frac{n^2}{n} - \frac{i}{n}}\right)$$

$$= 1 - \lim_{n\to\infty}\prod_{i=0}^{n}\left(1 - \frac{1}{n - \frac{i}{n}}\right) = 1 - \lim_{n\to\infty}\prod_{i=0}^{n}\left(1 - \frac{1}{n}\right) = 1 - \lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^n = \boxed{1 - \exp(-1)}$$

1.2.3 Problems we did on homework

Fall 2017 Problem 2 (Homework 3 Problem 6). An urn contains $2n$ balls, coming in pairs: two balls are labeled "1", two balls are labeled "2", ..., two balls are labeled "n". A sample of size n is taken without replacement. Denote by N the number of pairs in the sample. Compute the expected value and the variance of N. **You do not need to simplify the expression for the variance.**

Solution. Let X_k be an indicator variable for both balls labeled k being in the sample. Note that

$$\mathbb{E}(X_k) = \Pr(X_k = 1) = \frac{\binom{2n-2}{n-2}}{\binom{2n}{n}} = \frac{(2n-2)!}{(n-2)!n!} \Big/ \frac{(2n)!}{n!n!} = \frac{(2n-2)!n!}{(2n)!(n-2)!} = \frac{n(n-1)}{2n(2n-1)} = \frac{n-1}{2(2n-1)}$$

Now since $N = \sum_{k=1}^{n} X_k$, we have

$$\mathbb{E}(N) = \mathbb{E}\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} \mathbb{E}(X_k) = \boxed{\frac{n(n-1)}{2(2n-1)}}$$

To obtain the variance, note that

$$\mathbb{E}(N^2) = \mathbb{E}\left(\sum_{k=1}^{n} X_k\right)^2 = \mathbb{E}\left(\sum_{k=1}^{n} X_k^2 + 2\sum_{1\le i<j\le n} X_i X_j\right) = \sum_{k=1}^{n}\mathbb{E}(X_k^2) + 2\sum_{1\le i<j\le n}\mathbb{E}(X_i X_j)$$

Because

$$\mathbb{E}(X_k^2) = 1^2 \cdot \Pr(X_k = 1) = \mathbb{E}(X_k) = \frac{n-1}{2(2n-1)}$$

$$\mathbb{E}(X_iX_j) = \Pr(X_i = 1 \cap X_j = 1) = \frac{\binom{2n-4}{n-4}}{\binom{2n}{n}} = \frac{(2n-4)!}{(n-4)!n!} \Big/ \frac{(2n)!}{n!n!} = \frac{(2n-4)!n!}{(2n)!(n-4)!}$$

$$= \frac{n(n-1)(n-2)(n-3)}{2n(2n-1)(2n-2)(2n-3)} = \frac{(n-1)(n-2)(n-3)}{2(2n-1)(2n-2)(2n-3)}$$

we have

$$\mathbb{E}(N^2) = \sum_{k=1}^{n} \frac{n-1}{2(2n-1)} + 2 \sum_{1 \le i < j \le n} \frac{(n-1)(n-2)(n-3)}{2(2n-1)(2n-2)(2n-3)} = \frac{n(n-1)}{2(2n-1)} + 2\binom{n}{2} \frac{(2n-4)!n!}{(2n)!(n-4)!}$$

$$= \frac{n(n-1)}{2(2n-1)} + \frac{n!}{(n-2)!} \cdot \frac{(2n-4)!n!}{(2n)!(n-4)!} = \frac{n(n-1)}{2(2n-1)} + n(n-1) \cdot \frac{(n-1)(n-2)(n-3)}{2(2n-1)(2n-2)(2n-3)}$$

$$= \frac{n(n-1)}{2(2n-1)} + \frac{n(n-1)^2(n-2)(n-3)}{2(2n-1)(2n-2)(2n-3)}$$

$$\implies \mathrm{Var}(N) = \mathbb{E}(N^2) - \mathbb{E}(N)^2 = \boxed{\frac{n(n-1)}{2(2n-1)} + \frac{n(n-1)^2(n-2)(n-3)}{2(2n-1)(2n-2)(2n-3)} - \frac{n^2(n-1)^2}{4(2n-1)^2}}$$

Fall 2017 Problem 3 (HW3 Problem 8—almost full solution)

Let $U_1, U_2, \ldots$ be iid random variables, uniformly distributed on $[0, 1]$, and let N be a Poisson random variable with mean value equal to 1. Assume that N is independent of $U_1, U_2, \ldots$ and define

$$Y = \begin{cases} 0 & \text{if } N = 0 \\ \max\limits_{1 \le i \le N} U_i & \text{if } N > 0 \end{cases}$$

Compute the expected value of Y.

Solution. Since Y is a function of N, let $Y = y(N)$. By the Law of the Unconscious Statistician,

$$\mathbb{E}(Y) = \mathbb{E}\big(\mathbb{E}(Y \mid N)\big) = \mathbb{E}\big(\mathbb{E}(\max_{1 \le i \le N} U_i \mid N = n)\big)$$

Let $Z_n = \max_{1 \le i \le n} U_i$. The cdf of Z_n can be calculated as follows:

$$\Pr(Z_n \le x) = \Pr(\max_{1 \le i \le n} U_i \le x) = \Pr(U_1 \le x \cap U_2 \le x \cap \ldots \cap U_n \le x) = x^n$$

for $x \in [0, 1]$. Therefore the pdf of Z_n is its derivative, nx^{n-1}. So we have

$$\mathbb{E}(\max_{1\le i\le N} U_i \mid N=n) = \mathbb{E}(Z_n) = \int_0^1 xnx^{n-1}dx = n\int_0^1 x^n dx = n\frac{x^{n+1}}{n+1}\Big|_0^1 = \frac{n}{n+1}$$

Plugging this into the expression for $\mathbb{E}(Y)$ yields

$$\mathbb{E}(Y) = \mathbb{E}\big(\mathbb{E}(Y \mid N)\big) = \sum_{n=0}^{\infty} \frac{n}{n+1}\Pr(N=n) = \sum_{n=1}^{\infty} \frac{n}{n+1}\frac{\exp(-1)1^n}{n!}$$

$$= \frac{1}{e}\sum_{n=1}^{\infty}\frac{n+1-1}{(n+1)!} = \frac{1}{e}\Big(\sum_{n=1}^{\infty}\frac{n+1}{(n+1)!} - \sum_{n=1}^{\infty}\frac{1}{(n+1)!}\Big) = \frac{1}{e}\Big(\sum_{n=1}^{\infty}\frac{1}{n!} - \sum_{m=2}^{\infty}\frac{1}{m!}\Big)$$

$$= \frac{1}{e}\big[e-1-(e-1-1)\big] = \boxed{\frac{1}{e}}$$

Fall 2013 Problem 3/Spring 2011 Problem 2 (HW3 Problem 9; coupon collector problem) Only parts I didn't do: Let D be the event that no box receives more than 1 ball. Fix $a \in (0,1)$. If both $n, d \to \infty$ together, what relation must they satisfy in order to have $\Pr(D) \to a$?

HW3 Problem 9. Consider n (different) balls placed at random in m boxes so that each of m^n configurations is equally likely.

(a) Compute the expected value and the variance of the number of empty boxes.

(b) Show that if $\lim_{m,n\to\infty} m\exp(-n/m) = \lambda \in (0,\infty)$, then, in the same limit, the number of empty boxes has Poisson distribution with parameter λ.

(c) For $k \ge 1$ such that $k+3 \le m$, define the event A_k that the boxes $k, k+1, k+2, k+3$ are empty. Assuming that $m > 8$, compute $\Pr(A_1 \cup A_3 \cup A_5)$. How will the answer change if $m = 8$?

(d) Now imagine that the balls are dropped one-by-one (with each ball equally likely to go into any of the m boxes, independent of all other balls), and denote by N_m the minimal number of balls required to fill all the boxes. Compute $\mathbb{E}(N_m), \mathrm{Var}(N_m)$ and

$$\lim_{m\to\infty}\Pr\left(\frac{N_m - m\log m}{m} \le x\right)$$

(e) Suppose we instead place an unlimited number of balls into the m boxes until we have k consecutive balls land in the same box (it doesn't matter which box). What is the expected number of balls we will drop until this happens?

Solution.

(a) Let A_i be the event that the ith box is empty. Let $\mathbf{1}_{A_i}$ be the indicator for A_i. Then $X = \sum_{i=1}^m \mathbf{1}_{A_i}$.

$$\mathbb{E}(X) = \mathbb{E}\Big(\sum_{i=1}^m \mathbf{1}_{A_i}\Big) = \sum_{i=1}^m \Big(\mathbb{E}\mathbf{1}_{A_i}\Big) = \sum_{i=1}^m \Pr(A_i) = \sum_{i=1}^m \left(\frac{m-1}{m}\right)^n = \boxed{\frac{(m-1)^n}{m^{n-1}}}$$

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{m} \mathbf{1}_{A_i}\right) = \sum_{i=1}^{m} \text{Var}(\mathbf{1}_{A_i}) + 2 \sum_{1 \le i < j \le m} \text{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j})$$

$$\text{Var}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j}) = \mathbb{E}(\mathbf{1}_{A_i}\mathbf{1}_{A_i}) - \mathbb{E}(\mathbf{1}_{A_i})^2 = \Pr(A_i \cap A_i) - \Pr(A_i)^2 = \left(\frac{m-1}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n}$$

$$\text{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j}) = \mathbb{E}(\mathbf{1}_{A_i}\mathbf{1}_{A_j}) - \mathbb{E}(\mathbf{1}_{A_i})\mathbb{E}(\mathbf{1}_{A_j}) = \Pr(A_i \cap A_j) - \Pr(A_i)\Pr(A_j) = \left(\frac{m-2}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n}$$

$$\implies \text{Var}(X) = m \cdot \left[\left(\frac{m-1}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n}\right] + \frac{m!}{(m-2)!}\left[\left(\frac{m-2}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n}\right]$$

$$= \frac{(m-1)^n}{m^{n-1}} - \frac{(m-1)^{2n}}{m^{2n-1}} + (m^2 - m)\left[\left(\frac{m-2}{m}\right)^n - \left(\frac{m-1}{m}\right)^{2n}\right]$$

$$\boxed{\text{Var}(X) = \frac{(m-1)^n}{m^{n-1}} - \frac{(m-1)^{2n}}{m^{2n-1}} + (m-1)\left[\frac{(m-2)^n}{m^{n-1}} - \frac{(m-1)^{2n}}{m^{2n-1}}\right]}$$

(b) Note that

$$X = \sum_{i=1}^{m} \mathbf{1}_{A_i}$$

and that the A_i are only weakly dependent on each other, especially as m and n increase. Therefore as $m, n \to \infty$, the Poisson paradigm (see Section 1.1.10) suggests $X \sim \text{Poisson}(\mathbb{E}(X))$. We have

$$\mathbb{E}(X) = \frac{(m-1)^n}{m^{n-1}}$$

so

$$\lim_{n,m\to\infty} \mathbb{E}(X) = \lim_{n,m\to\infty} m \cdot \left(\frac{m-1}{m}\right)^n = \lim_{n,m\to\infty} m \cdot \left(1 - \frac{1}{m}\right)^n = \lim_{n,m\to\infty} m \cdot \left[\left(1 - \frac{1}{m}\right)^m\right]^{n/m}$$

$$\approx \lim_{n,m\to\infty} m \cdot \left[e^{-1}\right]^{n/m} = \lim_{n,m\to\infty} me^{-n/m}$$

Using

$$\lim_{m,n\to\infty} m \exp(-n/m) = \lambda \in (0, \infty)$$

we have $\boxed{X \sim \text{Poisson}(\lambda) \text{ as } m, n \to \infty}$.

(c)

$$\Pr(A_1 \cup A_3 \cup A_5) = \Pr(A_1)+\Pr(A_3)+\Pr(A_5)-\Pr(A_1 \cap A_3)-\Pr(A_1 \cap A_5)-\Pr(A_3 \cap A_5)+\Pr(A_1 \cap A_3 \cap A_5)$$

We have

$$\Pr(A_1) = \Pr(A_3) = \Pr(A_5) = \left(\frac{m-4}{m}\right)^n$$

$$\Pr(A_1 \cap A_3) = \Pr(A_3 \cap A_5) = \left(\frac{m-6}{m}\right)^n$$

$$\Pr(A_1 \cap A_5) = \Pr(A_1 \cap A_3 \cap A_5) = \left(\frac{m-8}{m}\right)^n$$

Therefore

$$\Pr(A_1 \cup A_3 \cup A_5) = 3\left(\frac{m-4}{m}\right)^n - 2\left(\frac{m-6}{m}\right)^n = \boxed{\frac{3(m-4)^n - 2(m-6)^n}{m^n}}$$

(d) N_m is the minimal number of balls required to fill all the boxes. Let T_i be the number of balls that have to be dropped to fill the ith box after $i-1$ boxes have been filled. The probability of filling a new box after $i-1$ boxes have been filled is $\frac{m-(i-1)}{m}$. (Note that T_1 should be identically 1 regardless of $m \geq 1$; this checks out using this expression.) Therefore T_i has a geometric distribution with $E(T_i) = \frac{m}{m-(i-1)}$. Since $N_m = \sum_{i=1}^m T_i$, we have

$$\mathbb{E}(N_m) = \mathbb{E}\left(\sum_{i=1}^m T_i\right) = \sum_{i=1}^m \mathbb{E}(T_i) = \sum_{i=1}^m \frac{m}{m-(i-1)} = \boxed{m\sum_{i=1}^m \frac{1}{i}}$$

Because the T_i are independent, we have

$$\begin{aligned}\text{Var}(N_m) = \text{Var}\left(\sum_{i=1}^m T_i\right) = \sum_{i=1}^m \text{Var}(T_i) &= \sum_{i=1}^m \left(1 - \frac{m-(i-1)}{m}\right) \Big/ \left(\frac{m-(i-1)}{m}\right)^2 \\ &= \sum_{i=1}^m \frac{i-1}{m} \cdot \left(\frac{m}{m-(i-1)}\right)^2 = \boxed{m\sum_{i=1}^m \frac{i-1}{[m-(i-1)]^2}}\end{aligned}$$

Finally, to find

$$\lim_{m\to\infty} \Pr\left(\frac{N_m - m\log m}{m} \leq x\right)$$

begin by noting that we can also express N_m as

$$\Pr(N_m \leq k) = \Pr\left(X_{m,k} = 0\right)$$

where $X_{m,k}$ is defined as X is in part (b) with k being the number of balls that have been dropped so far, $k \in \mathbb{N} \geq m$. (For $k < m$, $\Pr(N_m \leq k) = 0$.)

Again, let $A_{i,k}$ be the event that the ith box is empty after dropping k balls. Then because $X_{m,k} = \sum_{i=1}^{m} \mathbf{1}_{A_{i,k}}$ and the $A_{i,k}$ are only weakly dependent on each other (especially as m becomes large), the Poisson paradigm (see Section 1.1.10) again suggests that as $m \to \infty$, $X_{m,k} \sim \text{Poisson}(\lambda_k)$ where $\lambda_k = \mathbb{E}(X_{m,k})$ is defined as above. Therefore we have

$$\lim_{m\to\infty} \Pr\left(\frac{N_m - m\log m}{m} \leq x\right) = \lim_{m\to\infty} \Pr\big(N_m \leq xm + m\log m\big) = \lim_{m\to\infty} \Pr\big(X_{m,xm+m\log m}$$

$$= 0) \approx \frac{\exp(-\lambda_{xm+m\log m}) \cdot \lambda^0_{xm+m\log m}}{0!} = \exp\big(-\lambda_{xm+m\log m}\big)$$

And we have

$$\lambda_{xm+m\log m} = \lim_{m\to\infty} m\exp\left(-\frac{xm + m\log m}{m}\right) = \lim_{m\to\infty} m\exp\big(-x-\log m\big) = \lim_{m\to\infty} m/m\exp(-x)$$

$$= \exp(-x)$$

which yields

$$\boxed{\lim_{m\to\infty} \Pr\left(\frac{N_m - m\log m}{m} \leq x\right) = \exp(\exp(-x))}$$

(e) Let $N = N_k$ be the number of balls that are dropped until k consecutive balls land in the same box, and likewise for N_{k-1}. Suppose we have already observed $k-1$ consecutive outcomes (of any kind) in N_{k-1} trials. Then we finish on the next term (by having another consecutive outcome) with probability $1/m$. Otherwise we have a different outcome and then repeat the same process again. So we have

$$\mathbb{E}(N_k \mid N_{k-1}) = N_{k-1} + 1 \cdot \frac{1}{m} + \mathbb{E}(N_k) \cdot \left(1 - \frac{1}{m}\right)$$

Therefore

$$\mathbb{E}(N) = \mathbb{E}(N_k) = \mathbb{E}\big[\mathbb{E}(N_k \mid N_{k-1})\big] = \mathbb{E}(N_{k-1}) + \frac{1}{m} + \left(1 - \frac{1}{m}\right)\mathbb{E}(N_k)$$

$$\iff \frac{1}{m}\mathbb{E}(N_k) = \mathbb{E}(N_{k-1}) + \frac{1}{m} \iff \mathbb{E}(N_k) = m\mathbb{E}(N_{k-1}) + 1$$

We have a recursive formula. Note that $\mathbb{E}(N_1) = 1$ because the number of trials until there is 1 consecutive outcome of any kind is simply 1. We can then calculate as follows:

$$\mathbb{E}(N_2) = m\mathbb{E}(N_{2-1}) + 1 = m + 1$$

$$\mathbb{E}(N_3) = m\mathbb{E}(N_{3-1}) + 1 = m(m+1) + 1 = 1 + m + m^2$$

$$\mathbb{E}(N_4) = m\mathbb{E}(N_{4-1}) + 1 = m(1 + m + m^2) + 1 = 1 + m + m^2 + m^3$$

$$\vdots$$

$$\mathbb{E}(N_k) = \sum_{i=0}^{k-1} m^i = \frac{1 \cdot (1 - m^k)}{1 - m} = \boxed{\frac{m^k - 1}{m - 1}}$$

Fall 2012 Problem 1 (HW2 Problem 10/HW 1 Problem 9) Only part I didn't do: Find the mean and variance of $S_n = X_1 + \ldots + X_n$, the total number of white balls added to the urn up to time n.

HW1 Problem 9. An urn contains b black and w white balls. At each step, a ball is removed from the urn at random and then put back together with one more ball of the same color. Compute the probability p_n to get a black ball on step $n, n \geq 1$.

Solution. Step 1:

$$p_1 = \frac{b}{b + w}$$

Step 2: We need to separately consider the cases where a black ball was selected on step 1 (with probability p_1) or a white ball (with probability $1 - p_1$).

$$p_2 = p_1 \cdot \frac{b+1}{b+w+1} + (1 - p_1) \cdot \frac{b}{b+w+1} = p_1\left(\frac{b+1}{b+w+1} - \frac{b}{b+w+1}\right) + \frac{b}{b+w+1}$$

$$= p_1\left(\frac{1}{b+w+1} + \frac{1}{p_1}\frac{b}{b+w+1}\right) = p_1\left(\frac{1}{b+w+1} + \frac{b+w}{b}\frac{b}{b+w+1}\right)$$

$$= p_1\left(\frac{b+w+1}{b+w+1}\right) = p_1$$

$$\implies p_2 = p_1 = \frac{b}{b+w}$$

Step 3: Regardless of the previous steps, there are now $b + w + 2$ balls in the urn. Since we know that $p_1 = p_2$, the probability that we have selected k black balls so far (and thus, the probability that there are currently $b + k$ black balls in the urn) is given by

$$\Pr(k \text{ balls chosen in first 2 rounds}) = \binom{2}{k} p_1^k (1 - p_1)^{2-k} = \binom{2}{k}\left(\frac{b}{b+w}\right)^k \left(\frac{w}{b+w}\right)^{2-k}$$

$$= \binom{2}{k} \frac{b^k w^{2-k}}{(b+w)^2}$$

for $k \in \{0, 1, 2\}$. Given that we have selected k black balls so far, the probability of selecting a black ball this time is $\frac{b+k}{b+w+2}$. Therefore the probability of selecting a black ball this round is

$$p_3 = \sum_{k=0}^{2} \binom{2}{k} \frac{b^k w^{2-k}}{(b+w)^2} \frac{b+k}{b+w+2} = \frac{1}{(b+w+2)(b+w)^2} \sum_{k=0}^{2} \binom{2}{k} (b+k) b^k w^{2-k}$$

$$= \frac{1}{(b+w+2)(b+w)^2} \left(\binom{2}{0} bw^2 + \binom{2}{1} (b+1)bw + \binom{2}{2} (b+2)b^2 \right)$$

$$= \frac{bw^2 + 2(b+1)bw + (b+2)b^2}{(b+w+2)(b+w)^2} = \frac{b}{b+w} \left(\frac{w^2 + 2bw + 2w + b^2 + 2b}{b^2 + bw + 2b + wb + w^2 + 2w} \right)$$

$$= \frac{b}{b+w} \left(\frac{w^2 + 2bw + 2w + b^2 + 2b}{b^2 + 2bw + 2b + w^2 + 2w} \right) = \frac{b}{b+w} = p_1$$

There seems to be a clear pattern here. Let's find the general formula by induction.

Step $n+1$: Assume that the probability of choosing a black ball on steps $1, 2, \ldots, n$ was $\frac{b}{b+w}$ each time.

(a bunch of boring stuff, then it worked.)

HW2 Problem 10. Random variables $(X_1, \ldots, X_n)$ are called *exchangeable* if $\Pr(X_1 = x_1, \ldots, X_n = x_n) = \Pr(X_{\tau(1)} = x_1, \ldots, X_{\tau(n)} = x_n)$ for all real numbers $x_1, \ldots, x_n$ and every permutation τ of the set $\{1, \ldots, n\}$. In the setting of Problem 9 from Homework 1, let $X_k = 1$ if a white ball is drawn on step k, and $X_k = 0$ otherwise. Show that the random variables $X_1, \ldots, X_n$ are exchangeable for every $n \geq 2$.

Solution. For $n = 2$: There are two cases which we must show are equal to show exchangeability:

$$\Pr(X_1 = 0, X_2 = 1) = \Pr(X_1 = 1, X_2 = 0)$$

First,

$$\Pr(X_1 = 0, X_2 = 1) = \Pr(\text{black first}) \Pr(\text{white second} \mid \text{black first}) = \left(\frac{b}{b+w} \right) \left(\frac{w}{b+w+1} \right)$$

$$\left(\frac{w}{b+w} \right) \left(\frac{b}{b+w+1} \right) = \Pr(X_1 = 1, X_2 = 0)$$

which proves exchangeability for $n = 2$. In the general case, we seek to show that $X_1, \ldots, X_n$ are exchangeable. That is, in all $n+1$ unordered sets $\mathbb{X}_k = \{x_{1k}, x_{2k}, \ldots, x_{nk} \mid x_{ik} \in \{0,1\}, \sum_i x_{ik} = k\}$, in all $\binom{n}{k}$ permutations of $\mathbb{X}_k$,

$$\Pr(\mathbb{X}_{kj} = \Pr(\mathbb{X}_{kj'}$$

where j and j' denote different permutations of $\mathbb{X}_k$. That is,

$$\Pr(X_1 = x_{1k}, X_2 = x_{2k}, \ldots, X_n = x_{nk}) = \Pr(X_{j_1} = x_{1k}, X_{j_2} = x_{2k}, \ldots, X_{j_n} = x_{nk})$$

where $j_1, j_2, \ldots, j_n$ index the permuted variables. Consider $\mathbb{X}_{kj^*}$ where all k white balls are chosen first and all $n-k$ black balls are chosen last. We have

$$\Pr(\mathbb{X}_{kj^*}) = \prod_{i=1}^{k} \left(\frac{w+i-1}{b+w+i-1} \right) \cdot \prod_{i=k+1}^{n} \left(\frac{b+i-k-1}{b+w+i-1} \right)$$

$$= \prod_{i=1}^{n} \left(\frac{1}{b+w+i-1} \right) \cdot \left[\prod_{i=1}^{k} (w+i-1) \prod_{i=k+1}^{n} (b+i-k-1) \right] = \prod_{i=1}^{n} \left(\frac{1}{b+w+i-1} \right) \cdot \left[\prod_{i=1}^{k} (w+i-1) \prod_{i'=1}^{n-k} (b+i'-1) \right]$$

It is easy to see that the leftmost product will always equal the product of the denominators, regardless of the permutation, since one ball is added to the urn after every draw. Similarly, regardless of permutation, the numerator of the probability of drawing the ith white ball will always equal $w+i-1$, the number of white balls already in the urn. Likewise, the numerator of the probability of drawing the i'th black ball is always $b+i'-1$. Because multiplication is commutative, all permutations of these numbers will have equal products. Therefore $\Pr(\mathbb{X}_{kj^*}) = \Pr(\mathbb{X}_{kj}$ for all k. That is,

$$\Pr(X_1 = x_1, \ldots, X_n = x_n) = \Pr(X_{\tau(1)} = x_1, \ldots, X_{\tau(n)} = x_n)$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$, all $n \in \mathbb{Z}$ such that $n \geq 2$, all permutations τ.

Homework 2 Problem 2. Consider the function

$$f(x) = \begin{cases} C(2x - x^2) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Could f be a distribution function? If so, determine C.

(b) Could f be a probability density function? If so, determine C.

Solution.

(a) If f is a distribution function, $\lim_{x \to -\infty} f(x) = 0$, $\lim_{x \to \infty} f(x) = 1$, and $f'(x) \geq 0 \; \forall \; x \in \mathbb{R}$. f clearly does not meet the second or third conditions and is therefore not a distribution function.

(b) If f is a density function then $\int_{-\infty}^{\infty} f(x)dx = 1$ and $f(x) \geq 0 \; \forall x \in \mathbb{R}$.

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^2 C(2x - x^2)dx = C\left[x^2 - \frac{x^3}{3} \right]_0^2 = C\left(4 - \frac{8}{3} - 0 \right) = C \cdot \frac{4}{3}$$

$$= 1 \iff C = \frac{3}{4}$$

Next we check that f is always nonnegative. It equals zero except on $(0, 2)$.

$$\frac{3}{4}(2x - x^2) \geq 0 \iff x(2 - x) \geq 0 \iff x \in (0, 2)$$

Therefore f is nonnegative $\forall x \in \mathbb{R}$, so $\boxed{f \text{ is a probability density function if } C = \frac{3}{4}.}$

HW1 Problem 8. Two people, A and B, are involved in a duel. The rules are simple: shoot at each other once; if at least one is hit, the duel is over, if both miss, repeat (go to the next round), and so on. Denote by p_A and p_B the probabilities that A hits B and B hits A with one shot, and assume that that hitting/missing is independent from round to round. Compute the probabilities of the following events: (a) the duel ends and A is not hit; (b) the duel ends and both are hit; (c) the duel ends after round number n; (d) the duel ends after round number n GIVEN that A is not hit; (e) the duel ends after n rounds GIVEN that both are hit; (f) the duel goes on forever.

Solution.

(a) Let A_k denote the event that the duel is ended by A shooting B in the kth round (with neither person being shot in the first $k - 1$ rounds). Note that $\{A_k | k = 1, 2, \ldots\}$ are all mutually exclusive. Therefore the probability of the duel ending without A being hit is $\sum_{k=1}^{\infty} A_k$. Because the probabilities in each round are constant and independent,

$$A_k = (1 - p_A)^{k-1} p_A (1 - p_B)^k$$

So the probability that the duel ends and A is not hit is

$$\sum_{k=1}^{\infty} A_k = \sum_{k=1}^{\infty} (1 - p_A)^{k-1} p_A (1 - p_B)^k = p_A(1 - p_B) \sum_{k=1}^{\infty} (1 - p_A)^{k-1} (1 - p_B)^{k-1}$$

This is an infinite geometric series. Since the ratio $(1 - p_A)(1 - p_B)$ has absolute value less than 1, the sum can be calculated.

$$\sum_{k=1}^{\infty} A_k = p_A(1 - p_B) \cdot \frac{1}{1 - (1 - p_A)(1 - p_B)} = \frac{p_A(1 - p_B)}{p_A + p_B - p_A p_B} = \boxed{\frac{p_A(1 - p_B)}{p_A(1 - p_B) + p_B}}$$

(b) Similar to part (a). Let C_k denote the event that the duel is ended with both players being shot in the kth round (with neither person being shot in the first $k - 1$ rounds). Again, $\{C_k | k = 1, 2, \ldots\}$ are all mutually exclusive, so the probability of the duel ending in these circumstances is $\sum_{k=1}^{\infty} C_k$. We have

$$C_k = (1 - p_A)^{k-1} p_A (1 - p_B)^{k-1} p_B$$

$$\sum_{k=1}^{\infty} C_k = \sum_{k=1}^{\infty} (1 - p_A)^{k-1} p_A (1 - p_B)^{k-1} p_B = p_A p_B \sum_{k=1}^{\infty} (1 - p_A)^{k-1} (1 - p_B)^{k-1}$$

$$= p_A p_B \cdot \frac{1}{1 - (1 - p_A)(1 - p_B)} = \boxed{\frac{p_A p_B}{p_A + p_B - p_A p_B}}$$

Note that this value is less than the answer from part (a) if $p_B < \frac{1}{2}$ and greater if $p_B > \frac{1}{2}$

(c) Let B_k denote the event that the duel is ended by B shooting A in the kth round (with neither person being shot in the first $k-1$ rounds), with

$$B_k = (1-p_A)^k p_B (1-p_B)^{k-1}$$

Let A_k and C_k be defined as above. Note that $\{A_k | k = 1, 2, \ldots\}, \{B_k | k = 1, 2, \ldots\}, \{C_k | k = 1, 2, \ldots\}$ are all mutually exclusive, and that the event that the duel ends in round n is $\{A_n \cup B_n \cup C_n\}$. So the probability of the duel ending in round n is

$$\Pr(A_n \cup B_n \cup C_n) = \Pr(A_n) + \Pr(B_n) + \Pr(C_n)$$

$$= (1-p_A)^{n-1} p_A (1-p_B)^n + (1-p_A)^n p_B (1-p_B)^{n-1} + (1-p_A)^{n-1} p_A (1-p_B)^{n-1} p_B$$

$$= (1-p_A)^{n-1}(1-p_B)^{n-1}\big[p_A(1-p_B) + (1-p_A)p_B + p_A p_B\big]$$

$$= \boxed{(1-p_A)^{n-1}(1-p_B)^{n-1}(p_A + p_B - p_A p_B)}$$

(d) Let A_k, B_k, C_k be defined as above. The event that the duel ends at round n without A being hit is given by $\{A_n\}$.

$$\Pr(A_n) = \boxed{(1-p_A)^{n-1} p_A (1-p_B)^n}$$

(e) Let A_k, B_k, C_k be defined as above. The event that the duel ends at round n with both players being hit is given by $\{C_n\}$.

$$\Pr(C_n) = \boxed{(1-p_A)^{n-1} p_A (1-p_B)^{n-1} p_B}$$

(f) Let A_k, B_k, C_k be defined as above. The probability that the duel never ends is equal to 1 - the probability that the duel ends at some point, which is $\{A_k | k = 1, 2, \ldots\} \cup \{B_k | k = 1, 2, \ldots\} \cup \{C_k | k = 1, 2, \ldots\}$. Since all of these events are mutually exclusive, we have

$$1 - \Pr(\{A_k | k = 1, 2, \ldots\} \cup \{B_k | k = 1, 2, \ldots\} \cup \{C_k | k = 1, 2, \ldots\}) = 1 - \sum_{k=1}^{\infty} \big(A_k + B_k + C_k\big)$$

$$= 1 - \sum_{k=1}^{\infty} \big((1-p_A)^{k-1} p_A (1-p_B)^k + (1-p_A)^k p_B (1-p_B)^{k-1} + (1-p_A)^{k-1} p_A (1-p_B)^{k-1} p_B\big)$$

$$= 1 - \big[p_A(1-p_B) + (1-p_A)p_B + p_A p_B\big] \sum_{k=1}^{\infty} (1-p_A)^{k-1}(1-p_B)^{k-1}$$

$$= 1 - \big[p_A 1 - p_A p_B) + p_B - p_A)p_B + p_A p_B\big] \cdot \frac{1}{1-(1-p_A)(1-p_B)}$$

$$= 1 - \frac{p_A - p_A p_B + p_B - p_A p_B + p_A p_B}{p_A + p_B - p_A p_B} = 1 - \frac{p_A + p_B - p_A p_B}{p_A + p_B - p_A p_B} = \boxed{0}$$

Homework 1 Problem 1.

(I) Seven different gifts are distributed among 10 children. How many different outcomes are possible if every child can receive (a) at most one gift, (b) at most two gifts, (c) any number of gifts?

(II) Answer the same questions if the gifts are identical (but the children are still different).

Solution.

(I) (a) $\binom{10}{7}7! = \boxed{604,800}$

(b) Clearly all outcomes that satisfy part (I)(a) also satisfy these conditions, so we start with $\binom{10}{7}7! = 604,800$ possible outcomes. In addition, the following outcomes are possible:

(i) **A set of 6 children receive gifts; one child receives two gifts.** There are $\binom{10}{6}$ ways to pick a group of 6 children to receive the gifts. Next, there are $\binom{6}{1} = 6$ ways to choose which child receives two gifts. Finally, there are $7!/2!$ unique ways to distribute the gifts among the children once a particular partition is chosen (since order matters for all of the gifts except for the two that are received by the same child).

(ii) **A set of 5 children receive gifts; two children receive two gifts.** There are $\binom{10}{5}$ ways to pick a group of 5 children to receive the gifts. Next, there are $\binom{5}{2}$ ways to choose which of these children receive one gift and which receive two. Finally, there are $7!/(2!2!)$ unique ways to distribute the gifts among the children once a particular partition is chosen (since order matters for all of the gifts except for the two batches of two gifts that are received by the same child).
(Note that without the restriction that a child can receive at most two gifts, another possibility is that 1 child could receive 3 gifts, but that wouldn't work in this case.)

(iii) **A set of 4 children receive gifts; three children each receive two gifts.** There are $\binom{10}{4}$ ways to pick a group of 4 children to receive the gifts. Next, there are $\binom{4}{3} = 4$ ways to choose which of these children receive one gift and which receive two. Finally, there are $7!/(2!2!2!)$ unique ways to distribute the gifts among the children once a particular partition is chosen (since order matters for all of the gifts except for the three batches of two gifts that are received by the same child).
(Again, there are other possibilities for 4 children to receive 7 gifts, but none that satisfy the condition that no child receives more than 2 gifts.)

Clearly each of these outcomes are mutually exclusive. Therefore the answer is

$$\binom{10}{7}7! + \binom{10}{6}\cdot\binom{6}{1}\cdot\frac{7!}{2!} + \binom{10}{5}\cdot\binom{5}{2}\cdot\frac{7!}{2!2!} + \binom{10}{4}\cdot\binom{4}{3}\cdot\frac{7!}{2!2!2!}$$

$$= 7!\cdot\left(\frac{10!}{3!} + \frac{10!}{6!4!}\cdot 6\cdot\frac{1}{2} + \frac{10!}{5!5!}\cdot\frac{5!}{3!2!}\cdot\frac{1}{4} + \frac{10!}{4!6!}\cdot\frac{4!}{3!}\cdot\frac{1}{8}\right)$$

$$= 7!10!\cdot\left(\frac{1}{3!} + \frac{1}{6!4!}\cdot\frac{6}{2} + \frac{1}{5!}\cdot\frac{1}{3!2!}\cdot\frac{1}{4} + \frac{1}{6!3!\cdot 8}\right)$$

$$= \boxed{7,484,400}$$

(c) $10^7 = \boxed{10,000,000}$

(II) (a) $\binom{10}{7} = \boxed{120}$

(b) Clearly all outcomes that satisfy part (I)(a) also satisfy these conditions, so we start with $\binom{10}{7} = 120$ possible outcomes. In addition, the following outcomes are possible:

(i) A set of 6 children receive gifts; one child receives two gifts (6 distinct ways this could happen for each set of 6 children).

(ii) A set of 5 children receive gifts; two children receive two gifts ($\binom{5}{2}$ distinct ways this could happen for each set of 5 children).

(iii) A set of 4 children receive gifts; three children each receive two gifts (4 distinct ways this could happen for each set of 4 children).

Clearly each of these outcomes are mutually exclusive. Therefore the answer is

$$\binom{10}{7} + \binom{10}{6} \cdot \binom{6}{7-6} + \binom{10}{5} \cdot \binom{5}{7-5} + \binom{10}{4} \cdot \binom{4}{7-4} = \boxed{4,740}$$

(c) By Proposition 1.1.5.5, the number of nonnegative integer-valued vectors $(x_1, x_2, \ldots, x_r)$ satisfying the equation

$$x_1 + x_2 + \ldots + x_r = n$$

is equal to $\binom{n+r-1}{r-1} = \binom{7+10-1}{10-1} = \boxed{11,440}$.

Homework 1 Problem 2.

(I) 20 different gifts are distributed among seven children. How many different outcomes are possible if every child can receive (a) at least one gift, (b) at least two gifts, (c) any number of gifts?

(II) Answer the same questions if the gifts are identical (but the children are still different).

(III) Now try to generalize problems (1) and (2).

Solution.

(I) (a) There are 7^{20} possible allocations of gifts if we have no restrictions. If one child doesn't get a gift, there are $\binom{7}{1}$ ways to choose which child that is and 6^{20} subsequent allocations of gifts. Likewise, there are $\binom{7}{2} \cdot (7-2)^{20}$ ways to allocate the gifts if two children don't receive gifts, $\binom{7}{3} \cdot (7-3)^{20}$ ways if three children don't receive gifts, $\binom{7}{4} \cdot (7-4)^{20}$ ways if four children don't receive gifts, $\binom{7}{5} \cdot (7-5)^{20}$ ways if five children don't receive gifts, and $\binom{7}{6} \cdot (7-6)^{20}$ ways if six children don't receive gifts.

Let A_i denote the number of allocations in which i children do not receive gifts. In order to make the calculation, we must use the Inclusion-Exclusion principle (Proposition 1.1.5.1) (because, for example, some of the allocations in which three children don't receive gifts include allocations where four or more children don't receive gifts, and we don't want to double-count). Therefore the number of ways that at least one child can not receive a gift (i.e. the complement of every child receiving at least one gift) is

$$\left| \bigcup_{i=1}^{6} A_i \right| = \sum_{i=1}^{6} |A_i| - \sum_{1 \le i < j \le 6} |A_i \cap A_j| + \sum_{1 \le i < j < k \le 6} |A_i \cap A_j \cap A_k| - \ldots$$

$$+(-1)^{6-1}\left|A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_6\right|$$

Fortunately, these allocations are nested in the sense that all the allocations where e.g. 5 children do not receive gifts are a subset of all the allocations where 4 children do not receive gifts; that is

$$A_6 \subset A_5 \subset A_4 \subset A_3 \subset A_2 \subset A_1$$

which implies e.g.

$$A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_6 = A_6,$$

$$\sum_{1 \le i < j \le 6} |A_i \cap A_j| = 5|A_6| + 4|A_5| + 3|A_4| + 2|A_3| + |A_2|$$

So we have

$$\left|\bigcup_{i=1}^{6} A_i\right| = |A_6| + |A_5| + |A_4| + |A_3| + |A_2| + |A_1| - (5|A_6| + 4|A_5| + 3|A_4| + 2|A_3| + |A_2|)$$

$$+(4|A_6| + 3|A_5| + 2|A_4| + |A_3|) - (3|A_6| + 2|A_5| + |A_4|) + \ldots - |A_6|$$

$$= |A_1| - |A_2| + |A_3| - |A_4| + |A_5| - |A_6|$$

$$= \binom{7}{1} \cdot 6^{20} - \binom{7}{2} \cdot (7-2)^{20} + \binom{7}{3} \cdot (7-3)^{20} - \binom{7}{4} \cdot (7-4)^{20}$$

$$+\binom{7}{5} \cdot (7-5)^{20} - \binom{7}{6} \cdot (7-6)^{20}$$

The final answer is

$$7^{20} - \left|\bigcup_{i=1}^{6} A_i\right| = 7^{20} - \binom{7}{1} \cdot 6^{20} + \binom{7}{2} \cdot (7-2)^{20} - \binom{7}{3} \cdot (7-3)^{20} + \binom{7}{4} \cdot (7-4)^{20}$$

$$-\binom{7}{5} \cdot (7-5)^{20} + \binom{7}{6} \cdot (7-6)^{20} \approx \boxed{5.616 \cdot 10^{16}}$$

(b) Similar to above, but more complicated. The complement of every child receiving at least two gifts is that at least one child doesn't receive a gift (same as above) or at least one child only receives one gift. So we start from the baseline answer above, and subtract out all the possible allocations in which at least one child receives one gift.

If one child only receives one gift (and the rest receive more than one), there are $\binom{7}{1}$ ways to choose which child that is, $\binom{20}{1}$ ways to choose which gift that child receives, and 6^{20-1} allocations of the remaining gifts. If two children receive only one gift, there are $\binom{7}{2}$ ways to choose which children those are, $\binom{20}{2} \cdot 2!$ ways to choose which gifts those children get and distribute them among those children, and $(7-2)^{20-2}$ ways to allocate the remaining gifts. Likewise, if three children receive only one gift there are $\binom{7}{3}\binom{20}{3} \cdot 3! \cdot (7-3)^{20-3}$ ways to allocate the gifts,

$\binom{7}{4}\binom{20}{4} \cdot 4! \cdot (7-4)^{20-4}$ ways if four children receive only one gift, $\binom{7}{5}\binom{20}{5} \cdot 5! \cdot (7-5)^{20-5}$ ways if five children receive only one gift, and $\binom{7}{6}\binom{20}{6} \cdot 6!(7-6)^{20-6}$ ways if six children don't receive gifts.

Let B_j be the event that j children receive only one gift. Note that $B_1 \cap A_i$ is nonempty $\forall i < 7-1$, $B_2 \cap A_i$ is nonempty $\forall i < 7-2$, and in general, $B_j \cap A_i$ is nonempty $\forall i < 7-j, j \in \{1, 2, \ldots, 6\}$. Applying the Inclusion-Exclusion Principle (Proposition 1.1.5.1) in a similar way as in part (I)(a), the answer is

$$\boxed{7^{20} - \left|\bigcup_{i=1}^{6} A_i\right| - \left|\bigcup_{j=1}^{6} B_j\right| + \sum_{i \in \{1,\ldots,6\}, j \in \{1,\ldots,6\}} \left|A_i \bigcap B_j\right|}$$

Per part (I)(a), the first two terms approximately equal $5.616 \cdot 10^1 6$. Clearly

$$\bigcup_{i \in \{1,\ldots,6\}, j \in \{1,\ldots,6\}} \left(A_i \bigcap B_j\right) \subset \bigcup_{j=1}^{6} B_j$$

which implies

$$-\left|\bigcup_{j=1}^{6} B_j\right| + \left|A_i \bigcap_{i \in \{1,\ldots,6\}, j \in \{1,\ldots,6\}} B_j\right| < 0$$

so the answer to this part will be less than $5.616 \cdot 10^1 6$, which makes sense.

Calculating $\left|\bigcup_{j=1}^{6} B_j\right|$ is not too difficult using Inclusion-Exclusion:

$$\left|\bigcup_{j=1}^{6} B_j\right| = \sum_{j=1}^{6} |B_j| - \sum_{1 \le j < k \le 6} |B_j \cap B_k| + \sum_{1 \le j < k < \ell \le 6} |B_j \cap B_k \cap B_\ell| - \ldots$$

$$+(-1)^{6-1} |B_1 \cap B_2 \cap B_3 \cap \cdots \cap B_6|$$

where since

$$B_6 \subset B_5 \subset B_4 \subset B_3 \subset B_2 \subset B_1$$

which implies e.g.

$$B_1 \cap B_2 \cap B_3 \cap \cdots \cap B_6 = B_6,$$

$$\sum_{1 \le j < k \le 6} |B_j \cap B_j| = 5|B_6| + 4|B_5| + 3|B_4| + 2|B_3| + |B_2|$$

we have

$$\left|\bigcup_{j=1}^{6} B_j\right| = |B_6| + |B_5| + |B_4| + |B_3| + |B_2| + |B_1| - (5|B_6| + 4|B_5| + 3|B_4| + 2|B_3| + |B_2|)$$

$$+(4|B_6| + 3|B_5| + 2|B_4| + |B_3|) - (3|B_6| + 2|B_5| + |B_4|) + \ldots - |B_6|$$

$$= |B_1| - |B_2| + |B_3| - |B_4| + |B_5| - |B_6|$$

$$= \binom{7}{1}\binom{20}{1}\cdot(7-1)^{20-1} - \binom{7}{2}\binom{20}{2}\cdot 2!\cdot(7-2)^{20-2} + \binom{7}{3}\binom{20}{3}\cdot 3!\cdot(7-3)^{20-3} - \binom{7}{4}\binom{20}{4}\cdot 4!\cdot(7-4)^{20-4}$$

$$+\binom{7}{5}\binom{20}{5}\cdot 5!\cdot(7-5)^{20-5} - \binom{7}{6}\binom{20}{6}\cdot 6!(7-6)^{20-6}$$

$$\boxed{\approx 5.846\cdot 10^{16}}$$

However, calculating

$$\sum_{i\in\{1,\ldots,6\}, j\in\{1,\ldots,6\}} \left|A_i \bigcap B_j\right|$$

is very difficult because, for example, $B_2 \cap A_3$ is nonempty but $B_2 \not\subset A_3$ and $A_3 \not\subset B_2$.

(c) $\boxed{7^{20} \approx 7.979\cdot 10^{16}}$

(II) (a) By Proposition 1.1.5.4, there are $\binom{19}{6} = \boxed{27,132}$ ways to do this.

(b) Similar to Problem 1 part (II)(c), if the vector $(x_1, x_2, \ldots, x_7)$ represents the number of gifts given to each child, we would like a solution such that

$$x_1 + x_2 + \ldots + x_7 = 20, x_i \geq 2\ \forall i$$

By Proposition 1.1.5.6, the number of possible allocations under these conditions, is $\binom{20+7\cdot(1-2)-1}{7-1} = \binom{12}{6} = \boxed{924}$.

(c) By Proposition 1.1.5.5, the number of nonnegative integer-valued vectors $(x_1, x_2, \ldots, x_r)$ satisfying the equation

$$x_1 + x_2 + \ldots + x_r = n$$

is equal to $\binom{n+r-1}{r-1}$. In distributing 20 identical gifts to 7 different children, we can imagine the vector $(x_1, x_2, \ldots, x_{10})$ represents the number of gifts given to each child (where x_i is a nonnegative integer for all i). So we have $n = 20$ and $r = 7$. Therefore the number of possible allocations is

$$\binom{20+7-1}{7-1} = \boxed{165,765,600}$$

(III) Generalization of 1(I): If there are g distinguishable gifts and $c \geq g$ children, the number of distinct allocations if each child can receive

(a) at most one gift is $\binom{c}{g}g!$.

(b) at most two gifts is

$$\sum_{i=c-g+1}^{g} \binom{c}{i}\cdot\binom{i}{g-i}\cdot\frac{g!}{(2!)^{g-i}}$$

(c) any number of gifts is c^g.

Generalization of 1(II): If there are g identical gifts and $c \geq g$ children, the number of distinct allocations if each child can receive

(a) at most one gift is $\binom{c}{g}$.

(b) at most two gifts is

$$\sum_{i=c-g+1}^{g} \binom{c}{i} \cdot \binom{i}{g-i}$$

(c) any number of gifts is $\binom{g+c-1}{c-1}$

Generalization of 2(I): If there are g distinguishable gifts and $c \leq g$ children, the number of distinct allocations if each child must receive

(a) at least one gift is

$$c^g - \sum_{i=1}^{c-1} (-1)^{i+1} \binom{c}{i} \cdot (c-i)^g$$

(b) at least two gifts is

$$c^g - \sum_{i=1}^{c-1} (-1)^{i+1} \binom{c}{i} \cdot (c-i)^g - \sum_{i=1}^{c-1} (-1)^{i+1} \binom{c}{i} \binom{g}{i} \cdot i! \cdot (c-i)^{g-i}$$

(c) any number of gifts is c^g

Generalization of 2(II): If there are g identical gifts and $c \leq g$ children, the number of distinct allocations if each child must receive

(a) at least one gift is

$$\binom{g-1}{c-1}$$

(b) at least two gifts is

$$\binom{g-c-1}{c-1}$$

(c) any number of gifts is

$$\binom{g+c-1}{c-1}$$

Homework 1 Problem 4. You have \$20K to invest, and have a choice of stocks, bonds, mutual funds, or a CD. Investments must be made in multiples of \$1K, and there are minimal amounts to be invested: \$2K in stocks, \$2K in bonds, \$3K in mutual funds, and \$4K in the CD. Count the number of choices in each situation: (a) You want to invest in all four, (b) you want to invest in at least three out of four.

Solution.

(a) If the vector $(x_S, x_B, x_{MF}, x_{CD})$ represents the amount of money (in thousands of dollars) invested in each instrument, we would like a solution such that

$$x_S + x_B + x_{MF} + x_{CD} = 20$$

where

$$x_S \geq 2, x_B \geq 2, x_{MF} \geq 3, x_{CD} \geq 4$$

In a way similar to the proof for Proposition 1.1.5.6, note that we can transform this problem in the following way:

$$x_S - 1 + x_B - 1 + x_{MF} - 2 + x_{CD} - 3 = 20 - (1 + 1 + 2 + 3)$$

where

$$x_S - 1 \geq 1, x_B - 1 \geq 1, x_{MF} - 2 \geq 1, x_{CD} - 3 \geq 1$$

Letting $y_S = x_S - 1, y_B = x_B - 1, y_{MF} = x_{MF} - 2, y_{CD} = x_{CD} - 3$, we have the equivalent system

$$y_S + y_B + y_{MF} + y_{CD} = 13, y \geq 1 \ \forall y$$

By Proposition 1.1.5.4, the number of distinct solutions to this equation, and therefore the number of possible allocations under these conditions, is $\binom{13-1}{4-1} = \boxed{220}$.

(b) Enumerate the $\binom{4}{3} = 4$ possibilities.

(i) **Invest in stocks, bonds, and mutual funds.**

$$x_S + x_B + x_{MF} = 202$$

where

$$x_S \geq 2, x_B \geq 2, x_{MF} \geq 3$$

Note that we can transform this problem in the following way:

$$x_S - 1 + x_B - 1 + x_{MF} - 2 = 20 - (1 + 1 + 2)$$

where

$$x_S - 1 \geq 1, x_B - 1 \geq 1, x_{MF} - 2 \geq 1$$

Letting $y_S = x_S - 1, y_B = x_B - 1, y_{MF} = x_{MF} - 2$, we have the equivalent system

$$y_S + y_B + y_{MF} = 16, y \geq 1 \ \forall y$$

Therefore the number of possible allocations under these conditions is $\binom{16-1}{3-1} = \boxed{105}$.

(ii) **Invest in stocks, bonds, and CDs.**

$$x_S + x_B + x_{CD} = 20$$

where

$$x_S \geq 2, x_B \geq 2, x_{CD} \geq 4$$

Note that we can transform this problem in the following way:

$$x_S - 1 + x_B - 1 + x_{CD} - 3 = 20 - (1 + 1 + 3)$$

where

$$x_S - 1 \geq 1, x_B - 1 \geq 1, x_{CD} - 3 \geq 1$$

Letting $y_S = x_S - 1, y_B = x_B - 1, y_{CD} = x_{CD} - 3$, we have the equivalent system

$$y_S + y_B + y_{CD} = 15, y \geq 1 \ \forall y$$

Therefore the number of possible allocations under these conditions is $\binom{15-1}{3-1} = \boxed{91}$.

(iii) **Invest in stocks, mutual funds, and CDs.**

$$x_S + x_{MF} + x_{CD} = 2$$

where

$$x_S \geq 2, x_{MF} \geq 3, x_{CD} \geq 4$$

Note that we can transform this problem in the following way:

$$x_S - 1 + x_{MF} - 2 + x_{CD} - 3 = 20 - (1 + 2 + 3)$$

where

$$x_S - 1 \geq 1, x_{MF} - 2 \geq 1, x_{CD} - 3 \geq 1$$

Letting $y_S = x_S - 1, y_{MF} = x_{MF} - 2, y_{CD} = x_{CD} - 3$, we have the equivalent system

$$y_S + y_{MF} + y_{CD} = 14, y \geq 1 \; \forall y$$

Therefore the number of possible allocations under these conditions is $\binom{14-1}{3-1} = \boxed{78}$.

(iv) **Invest in bonds, mutual funds, and CDs.**

$$x_B + x_{MF} + x_{CD} = 2$$

where

$$x_B \geq 2, x_{MF} \geq 3, x_{CD} \geq 4$$

Note that we can transform this problem in the following way:

$$x_B - 1 + x_{MF} - 2 + x_{CD} - 3 = 20 - (1 + 2 + 3)$$

where

$$x_B - 1 \geq 1, x_{MF} - 2 \geq 1, x_{CD} - 3 \geq 1$$

Letting $y_B = x_B - 1, y_{MF} = x_{MF} - 2, y_{CD} = x_{CD} - 3$, we have the equivalent system

$$y_B + y_{MF} + y_{CD} = 14, y \geq 1 \; \forall y$$

therefore the number of possible allocations under these conditions is $\binom{14-1}{3-1} = \boxed{78}$.

(v) **Invest in all four:** per part 4(a), there are $\boxed{220}$ ways to do this.

Note that all of these possibilities are mutually exclusive. Therefore the total number is

$$\binom{16-1}{3-1} + \binom{15-1}{3-1} + \binom{14-1}{3-1} + \binom{14-1}{3-1} + \binom{13-1}{4-1} = 105 + 91 + 78 + 78 + 220 = \boxed{572}$$

1.2.4 DSO Statistics Group Screening Exam Problems

Exercise 2 (2017 DSO Statistics Group In-Class Screening Exam, Question 1). Let $X_1, X_2, \ldots, X_k$ be independent standard normal random variables and $\gamma_1(t), \ldots, \gamma_k(t)$ infinitely differentiable functions of a real variable defined on a closed, bounded interval, such that $\sum_{i=1}^k \gamma_i^2(t) = 1$ for all t. Let $Z(t) = \sum_{i=1}^k \gamma_i(t) X_i$. Let $\dot{Z}(t), \ddot{Z}(t)$, etc. denote first, second, etc. derivatives of $Z(t)$ with respect to t.

(a) Show that $\text{Cov}(Z(t), \dot{Z}(t)) = 0$.

(b) Evaluate $\mathbb{E}(Z(t) \mid \ddot{Z}(t))$ in terms of $\ddot{Z}(t)$ and expressions of the form

$$\sum_{i=1}^k (\gamma_i(t))^a (\delta^m \gamma_i(t)/\delta t^m)^b,$$

for some a, b, m values.

Solution

(a)

$$\dot{Z}(t) = \frac{\partial}{\partial t} \sum_{i=1}^k \gamma_i(t) X_i = \sum_{i=1}^k \dot{\gamma}_i(t) X_i$$

$$\implies \mathbb{E}\left(\dot{Z}(t)\right) = \sum_{i=1}^k \mathbb{E}\left(\dot{\gamma}_i(t) X_i\right) = 0$$

$$\implies \text{Cov}(Z(t), \dot{Z}(t)) = \mathbb{E}\big[(Z(t) - \mathbb{E}[Z(t)])(\dot{Z}(t) - \mathbb{E}[\dot{Z}(t)])\big] = \mathbb{E}\big[Z(t)\dot{Z}(t)\big]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^k \gamma_i(t) X_i\right)\left(\sum_{i=1}^k \dot{\gamma}_i(t) X_i\right)\right] = \sum_{i=1}^k \mathbb{E}\left(\gamma_i(t)\dot{\gamma}_i(t) X_i^2\right) + 0 = \sum_{i=1}^k \gamma_i(t)\dot{\gamma}_i(t)$$

since $\mathbb{E}(X_i^2) = 1$. But

$$\sum_{i=1}^k \gamma_i \dot{\gamma}_i(t) = 0 \tag{1.8}$$

because

$$\sum_{i=1}^k \gamma_i^2(t) = 1 \iff \frac{\partial}{\partial t}\left(\sum_{i=1}^k \gamma_i^2(t)\right) = 0 \iff 2\sum_{i=1}^k \gamma_i(t)\dot{\gamma}_i(t) = 0,$$

so the conclusion follows.

(b)

$$Z(t) = \sum_{i=1}^k \gamma_i(t) X_i \implies Z(t) \sim \mathcal{N}\left(0, \sum_{i=1}^k \gamma_i^2(t)\right) = \mathcal{N}(0, 1)$$

$$\ddot{Z}(t) = \sum_{i=1}^{k} \ddot{\gamma}_i(t) X_i \implies \ddot{Z}(t) \sim \mathcal{N}\left(0, \sum_{i=1}^{k} \ddot{\gamma}_i^2(t)\right)$$

Also, we have

$$\text{Cov}(Z(t), \ddot{Z}(t)) = \mathbb{E}\big[(Z(t) - \mathbb{E}[Z(t)])(\ddot{Z}(t) - \mathbb{E}[\ddot{Z}(t)])\big] = \mathbb{E}\big[Z(t)\ddot{Z}(t)\big]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^{k} \gamma_i(t) X_i\right)\left(\sum_{i=1}^{k} \ddot{\gamma}_i(t) X_i\right)\right] = \sum_{i=1}^{k} \mathbb{E}\left(\gamma_i(t)\ddot{\gamma}_i(t) X_i^2\right) + 0 = \sum_{i=1}^{k} \gamma_i(t)\ddot{\gamma}_i(t) = -\sum_{i=1}^{k} \dot{\gamma}_i^2(t)$$

because differentiating (1.8) yields

$$\frac{\partial}{\partial t} \sum_{i=1}^{k} \gamma_i \dot{\gamma}_i(t) = 0 \iff \sum_{i=1}^{k} \left(\dot{\gamma}_i^2(t) + \gamma_i(t)\ddot{\gamma}_i(t)\right) = 0 \iff \sum_{i=1}^{k} \gamma_i(t)\ddot{\gamma}_i(t) = -\sum_{i=1}^{k} \dot{\gamma}_i^2(t).$$

Since both distributions are Gaussian, we have

$$\mathbb{E}(Z(t) \mid \ddot{Z}(t)) = \mathbb{E}[Z(t)] + \frac{\text{Cov}[Z(t), \ddot{Z}(t)]}{\text{Var}[\ddot{Z}(t)]}(\ddot{Z}(t) - \mathbb{E}[\ddot{Z}(t)]) = \left[\sum_{i=1}^{k} \ddot{\gamma}_i^2(t)\right]^{-1} \cdot \left[-\sum_{i=1}^{k} \dot{\gamma}_i^2(t)\right] \ddot{Z}(t).$$

Exercise 3 (2018 DSO Statistics Group In-Class Screening Exam, Question 1). (a) For each $x > 0$, let $M(x)$ be a real-valued random variable and set $M(0) = 0$. Assume that the random function $M(x)$ is monotone non-decreasing on $[0, \infty)$. Define $T(y) := \inf\{x \geq 0 : M(x) \geq y\}$. Suppose that $e^{-1}T(y)$ converges in distribution to an Exponential(λ) random variable when $y \to \infty$.

(i) Find non-random $a(x)$ and $b(x) > 0$ such that $(M(x) - a(x))/b(x)$ converges in distribution for $x \to \infty$ to a non-degenerate random variable.

(ii) Provide the distribution function of the limit random variable. What is the name of this distribution?

(b) Let $X \sim \text{Bin}(n, p)$. Find $\mathbb{E}(1/(i + X))$, for $i = 1, 2$. Hint: Recall that $\int x^\alpha dx = x^{a+1}/(a + 1) + C$ for $a \neq -1$.

Solution

(a) (i) We have

$$T(y) = \inf\{x \geq 0 : M(x) \geq y\}; \qquad \lim_{y \to \infty} \Pr\left(e^{-y}T(y) \leq a\right) = 1 - e^{-\lambda a} \qquad \forall a \geq 0.$$

Note that

$$\lim_{y \to \infty} \Pr\left(e^{-y}T(y) \leq a\right) = \lim_{y \to \infty} \Pr(T(y) \leq ae^y)$$

Because $T(y) = \inf\{x \geq 0 : M(x) \geq y\}$ and $M(\cdot)$ is montonically increasing, we have the inequality $M(z) \geq y$ for all $z \geq x$. Therefore $T(y) \leq ae^y \iff M(ae^y) \geq y$. Let $z = ae^y \implies y = \log(z/a)$; then we have

$$\lim_{y\to\infty} \Pr(T(y) \leq ae^y) = \lim_{y\to\infty} \Pr(M(ae^y) \geq y) = \lim_{z\to\infty} \Pr(M(z) \geq \log(z) - \log(a))$$

Let $b = -\log a$ to get

$$= \lim_{z\to\infty} \Pr(M(z) - \log(z) \geq b) = 1 - e^{-\lambda a} \implies \lim_{z\to\infty} \Pr(M(z) - \log(z) \geq b) = 1 - e^{-\lambda e^{-b}}$$

$$\iff \lim_{z\to\infty} \Pr(M(z) - \log(z) \leq b) = e^{-\lambda e^{-b}}$$

(ii) The distribution function is $F(x) = e^{-\lambda e^{-x}}$, a Gumbel distribution with parameter λ. This is a proper cdf because $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.

(b) Using hint:

$$\int_0^1 t^x \, dt = \left[\frac{t^{x+1}}{x+1}\right]_0^1 = \frac{1}{x+1}$$

$$\iff \mathbb{E}\left[\int_0^1 t^X \, dt\right] = \mathbb{E}\left(\frac{1}{X+1}\right) \iff \int_0^1 \mathbb{E}(t^X) \, dt = \mathbb{E}\left(\frac{1}{X+1}\right)$$

$$\iff \int_0^1 \sum_{i=0}^{n} t^i \Pr(X = i) \, dt = \mathbb{E}\left(\frac{1}{X+1}\right) \iff \int_0^1 \sum_{i=0}^{n} t^i \binom{n}{i} p^i (1-p)^{n-i} \, dt = \mathbb{E}\left(\frac{1}{X+1}\right)$$

$$\iff (1-p)^n \int_0^1 \sum_{i=0}^{n} \binom{n}{i} \left(\frac{tp}{1-p}\right)^i \, dt = \mathbb{E}\left(\frac{1}{X+1}\right)$$

$$\iff (1-p)^n \int_0^1 \left(1 + \frac{tp}{1-p}\right)^n \, dt = \mathbb{E}\left(\frac{1}{X+1}\right)$$

$$\iff \int_0^1 (1-p+tp)^n \, dt = \mathbb{E}\left(\frac{1}{X+1}\right)$$

Let $u = 1 - p + tp \implies du = p \, dt$. Then we can write

$$\frac{1}{p}\int_{1-p}^1 u^n \, du = \mathbb{E}\left(\frac{1}{X+1}\right) \iff \frac{1}{p}\left[\frac{u^{n+1}}{n+1}\right]_{1-p}^1 = \mathbb{E}\left(\frac{1}{X+1}\right) \iff \mathbb{E}\left(\frac{1}{X+1}\right) = \frac{1}{p}\left[\frac{1-(1-p)^{n+1}}{n+1}\right]$$

Now we consider $\mathbb{E}\left[\frac{1}{X+2}\right]$.

$$\int_0^1 t^{x+1} \, dt = \left[\frac{t^{x+2}}{x+2}\right]_0^1 = \frac{1}{x+2}$$

$$\iff \mathbb{E}\left[\int_0^1 t^{X+1} \, dt\right] = \mathbb{E}\left(\frac{1}{X+2}\right) \iff \int_0^1 \mathbb{E}(t^{X+1}) \, dt = \mathbb{E}\left(\frac{1}{X+2}\right)$$

$$\iff \int_0^1 \sum_{i=0}^{n} t^{i+1} \Pr(X = i) \, dt = \mathbb{E}\left(\frac{1}{X+2}\right) \iff \int_0^1 \sum_{i=0}^{n} t^{i+1} \binom{n}{i} p^i (1-p)^{n-i} \, dt = \mathbb{E}\left(\frac{1}{X+2}\right)$$

$$\iff (1-p)^n \int_0^1 t \sum_{i=0}^{n} \binom{n}{i} \left(\frac{tp}{1-p}\right)^i \, dt = \mathbb{E}\left(\frac{1}{X+2}\right)$$

$$\iff (1-p)^n \int_0^1 t \left(1 + \frac{tp}{1-p}\right)^n \, dt = \mathbb{E}\left(\frac{1}{X+2}\right)$$

$$\iff \int_0^1 t\,(1-p+tp)^n \, dt = \mathbb{E}\left(\frac{1}{X+2}\right)$$

Let $u = 1-p+tp \implies du = p\,dt$ and $t = (u+p-1)/p$. Then we can write

$$\frac{1}{p^2}\int_{1-p}^1 u^n(u+p-1)\,du = \mathbb{E}\left(\frac{1}{X+2}\right) \iff \frac{1}{p^2}\int_{1-p}^1 [u^{n+1} + (p-1)u^n]\,du = \mathbb{E}\left(\frac{1}{X+2}\right)$$

$$\iff \frac{1}{p^2}\left[\frac{u^{n+2}}{n+2} + (p-1)\frac{u^{n+1}}{n+1}\right]_{1-p}^1 = \mathbb{E}\left(\frac{1}{X+2}\right)$$

$$\iff \mathbb{E}\left(\frac{1}{X+2}\right) = \frac{1}{p^2}\left[\frac{1-(1-p)^{n+2}}{n+2} - (1-p)\frac{1-(1-p)^{n+1}}{n+1}\right]$$

1.3 To Know for Math 505A Midterm 2

1.3.1 Definitions

Definition 1.3.1. A random variable X is **continuous** if its distribution function $F(x) = \Pr(X \leq x)$ can be written as

$$F(x) = \int_{-\infty}^{x} f(u)du$$

for some integrable $f : \mathbb{R} \to [0, \infty)$.

Definition 1.3.2. The function f is called the **(probability) density function** of the continuous random variable X.

Proposition 1.3.1.1. If X has pdf $f_X(x)$, then for $\mu \in \mathbb{R}$, $\sigma > 0$,

$$h(x) = \frac{1}{\sigma} f_X\left(\frac{x-\mu}{\sigma}\right)$$

is a pdf. In this setting μ is sometimes called a "location parameter" and σ is called a "scale parameter."

Definition 1.3.3. The **joint distribution function** of X and Y is the function $F : \mathbb{R}^2 \to [0, 1]$ given by

$$F(x, y) = \Pr(X \leq x \cap Y \leq y)$$

Definition 1.3.4. The random variables X and Y are **jointly continuous** with **joint (probability) density function** $f : \mathbb{R}^2 \to [0, \infty)$ if

$$F(x, y) = \int_{v=-\infty}^{y} \int_{u=-\infty}^{x} f(u, v) du dv \text{ for each } x, y \in \mathbb{R}$$

Definition 1.3.5. Two continuous random variables are **independent** if and only if $\{X \leq x\}$ and $\{Y \leq y\}$ are independent events for all $x, y \in \mathbb{R}$.

Ways to show independence:

- Use Definition 1.3.5: show that $\Pr(X \leq x \cap Y \leq y) = \Pr(X \leq x) \Pr(Y \leq y)$ for all $x, y \in \mathbb{R}$.
- **Theorem 1.3.1.2.** The random variables X and Y are independent if and only if $F(x, y) = F_X(x)F_Y(x)$ for all $x, y \in \mathbb{R}$.
- **Proposition 1.3.1.3.** For continuous random variables, the previous condition is equivalent to requiring $f(x, y) = f_X(x)f_Y(y)$.
- **Theorem 1.3.1.4.** If two variables are bivariate normal, they are independent if and only if their covariance

  $$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) dx dy$$

  is equal to 0.

  Theorem 1.3.1.5. (Change of variables.) 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 $T : \mathbb{R}^2 \to \mathbb{R}^2$ and let $S : \mathbb{R}^2 \to \mathbb{R}^2$ 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). Then

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

  Proof. If the transformation from X_1, Y_1 to X_2, Y_2 is given by $S(X_1, Y_1) = (X_2, Y_2)$, then the change of variables formula from calculus is as follows:

  $$\int \int_A f_{X_1,Y_1}(x, y) dx dy = \int \int_B f_{X_1,Y_1}\big(S(x, y)\big) \left| J(x, y) \right| dx dy$$

  where $A \subseteq \text{domain}(f_{X_1,Y_1}(\cdot))$, B is the transformation of the region A under S, and $|J(x, y)|$ is the Jacobian of S at (x, y). It follows from the definition of joint pdfs that the integrand on the right is the joint pdf of (X_2, Y_2); that is,

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

  □

- Characteristic functions:

  Theorem 1.3.1.6. X and Y are independent if and only if $\phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$.

Theorem 1.3.1.7. (Theorem 4.2.3, Grimmett and Stirzaker.) Let X and Y be random variables, and let $g, h : \mathbb{R} \to \mathbb{R}$. If X and Y are independent, then so are $g(X)$ and $h(Y)$.

Definition 1.3.6. The **correlation coefficient** between random variables X and Y is given by

$$\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Theorem 1.3.1.8. The correlation coefficient satisfies $|\rho| \leq 1$.

Proof. Apply the Cauchy-Schwarz Inequality (Theorem **??**) to $X - \mathbb{E}(X)$ and $Y - \mathbb{E}(Y)$:

$$\text{Cov}(X,Y)^2 = \big(\mathbb{E}\big[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\big]^2\big)^2 \leq \mathbb{E}\big[\big(X - \mathbb{E}(X)\big)^2\big]\mathbb{E}\big[\big(Y - \mathbb{E}(Y)\big)^2\big] = \text{Var}(X)\text{Var}(Y)$$

$$\iff \frac{\text{Cov}(X,Y)^2}{\text{Var}(X)\text{Var}(Y)} \leq 1 \iff \rho^2 \leq 1 \iff |\rho| \leq 1$$

□

Theorem 1.3.1.9. (Stein identity or Stein's Lemma.) Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function such that g and g' have polynomial volume growth. That is, $\exists\ a, b > 0$ such that $|g(x)|, |g'(x)| \leq a(1 + |x|)^b$, $\forall\ x \in \mathbb{R}$. Then

$$\mathbb{E}[(X - \mu)g(X)] = \sigma^2 \mathbb{E}g'(X).$$

Proof. Examining the left side, we have

$$\mathbb{E}[(X - \mu)g(X)] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x - \mu)g(x)e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

We will use integration by parts with $u = g(x) \implies du = g'(x)dx$, $dv = (x - \mu)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \implies v = -\sigma^2 \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ to yield the result:

$$\begin{aligned}\mathbb{E}[(X - \mu)g(X)] &= \frac{1}{\sqrt{2\pi\sigma^2}} \left(\left[-g(x) \cdot -\sigma^2 \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right]_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} g'(x) \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \right) \\ &= 0 + \sigma^2 \mathbb{E}(g'(X)).\end{aligned}$$

□

Example 1.3.1. Using Theorem 1.3.1.9, recursively compute $\mathbb{E}X^k$ for any positive integer k. Alternatively, for any $t > 0$, show that $\mathbb{E}e^{tX} = e^{t^2/2}$, i.e. compute the **moment generating function** of X. Then, using $\frac{d^k}{dt^k}|_{t=0}\mathbb{E}e^{tX} = \mathbb{E}X^k$ and using the power series expansion of the exponential, compute $\mathbb{E}X^k$ directly from the identity $\mathbb{E}e^{tX} = e^{t^2/2}$.

Solution We can use this result to recursively calculate $\mathbb{E}(X^k)$ for any positive integer k. Suppose we have $\mathbb{E}(X^{k-1})$. Letting $g(X) = X^{k-1} \implies g'(X) = (k-1)X^{k-2}$, we have

$$\mathbb{E}(X^k) = \mathbb{E}(X \cdot X^{k-1}) = \mathbb{E}(Xg(X)) = \mathbb{E}(g'(X)) \iff \boxed{\mathbb{E}(X^k) = (k-1)\mathbb{E}(X^{k-2})}.$$

Since $X \sim \mathcal{N}(0,1)$, we have $\mathbb{E}(X) = 0$, $\mathbb{E}(X^2) = \text{Var}(X) + \mathbb{E}(X)^2 = 1 + 0 = 1$. Therefore we have

$$\mathbb{E}(X^k) = \begin{cases} \prod_{i=1}^{(k-1)/2}(k-(2i-1))\mathbb{E}(X) & k \text{ is odd} \\ \prod_{i=1}^{k/2}(k-(2i-1))\mathbb{E}(X^2) & k \text{ is even} \end{cases}$$

$$\boxed{\mathbb{E}(X^k) = \begin{cases} 0 & k \text{ is odd} \\ \prod_{i=1}^{k/2}(k-(2i-1)) & k \text{ is even} \end{cases}}$$

1.3.2 Probability-Generating Functions

Definition 1.3.7.

$$G_X(s) = \mathbb{E}(s^X)$$

Theorem 1.3.2.1. Some useful properties:

(a) $\mathbb{E}(X) = G'_X(1)$, $\mathbb{E}[X(X-1)\cdots(X-k+1)] = G^{(k)}(1)$

(b) If X and Y are independent then $G_{X+Y}(s) = G_X(s)G_Y(s)$.

1.3.3 Moment-Generating Functions

Definition 1.3.8.

$$M_X(t) = \mathbb{E}(e^{tX})$$

Theorem 1.3.3.1 (Some useful properties). (a) $\mathbb{E}(X) = M'_X(0)$, $\mathbb{E}(X^k) = M^{(k)}(0)$

(b) If $X_1, X_2, \ldots, X_n$ are independent then $M_{X_1+ldots+X_n}(t) = \prod_{i=1}^n M_{X_i}(t)$.

Proof. (a)

(b)

$$M_{X_1+X_2+\ldots+X_n}(t) = \mathbb{E}\exp\left(t\sum_{i=1}^n X_i\right) = \mathbb{E}\prod_{i=1}^n e^{tX_i} = \prod_{i=1}^n \mathbb{E}e^{tX_i} = \prod_{i=1}^n M_{X_i}(t).$$

□

1.3.4 Characteristic Functions

Definition 1.3.9. The **characteristic function** $\phi_X : \mathbb{R} \to \mathbb{C}$ of a random variable X is defined as

$$\phi_X(t) := \mathbb{E}(e^{itX}) = \int_{\mathbb{R}} e^{itx} dF_X(x) = \int_{\mathbb{R}} e^{itx} f_X(x) dx = \mathbb{E}\cos(tX) + i\mathbb{E}\sin(tX).$$

Note that this formula suggests how we take the expectation of a complex-valued random variable: if $Z \in \mathbb{C}$ we define

$$\mathbb{E}Z = \mathbb{E}(\text{Re}(Z)) + i\mathbb{E}(\text{Im}(Z)).$$

Definition 1.3.10. The **modulus** of the complex number $z = a + bi$ is $|a + bi| = (a^2 + b^2)^{1/2}$.

Theorem 1.3.4.1 (Theorem 3.3.1 in Durrett [2019]). All characteristic functions have the following properties:

(a) $\phi(0) = 1$,

(b) They are Hermitian: $\phi(-t) = \overline{\phi(t)}$,

(c) $|\phi(t)| = \left|\mathbb{E}e^{itX}\right| \leq \mathbb{E}\left|e^{itX}\right| = 1$,

(d) $|\phi(t+h) - \phi(t)| \leq \mathbb{E}\left|e^{ihX} - 1\right|$, so $\phi(t)$ is uniformly continuous on $(-\infty, \infty)$.

(e) $\mathbb{E}e^{it(aX+b)} = e^{itb}\phi(at)$.

Proof. (a) Obvious.

(b)

$$\phi(-t) = \mathbb{E}\cos(-tX) + i\mathbb{E}\sin(-tX) = \mathbb{E}\cos(tX) - i\mathbb{E}\sin(tX) = \overline{\phi(t)}.$$

(c) The inequality follows from Jensen's inequality. For the last equality, note that

$$\mathbb{E}\left|e^{itX}\right| = \mathbb{E}\left|\cos(tX) + i\sin(tX)\right| = 1.$$

(d)

$$\begin{aligned}
|\phi(t+h) - \phi(t)| &= \left|\mathbb{E}\left(e^{i(t+h)X} - e^{itX}\right)\right| \\
&\leq \mathbb{E}\left|e^{i(t+h)X} - e^{itX}\right| \\
&= \mathbb{E}\left|e^{itX}\left(e^{ihX} - 1\right)\right| \\
&= \dots \\
&= \mathbb{E}\left|e^{ihX} - 1\right|.
\end{aligned}$$

Then uniform convergence follows from the bounded convergence theorem (Theorem 1.5.3 in Durrett [2019]).

(e)

$$\mathbb{E}e^{it(aX+b)} = e^{itb}\mathbb{E}e^{itaX} = e^{itb}\phi(at).$$

□

Theorem 1.3.4.2 (**Theorem 3.3.2 in Durrett [2019]**)**.** If X_1 and X_2 are independent and have characteristic functions ϕ_1 and ϕ_2 then $X_1 + X_2$ has characteristic function $\phi_1(t)\phi_2(t)$.

Proof.

$$\mathbb{E}e^{it(X_1+X_2)} = \mathbb{E}\left[e^{itX_1}e^{itX_2}\right] = \mathbb{E}e^{itX_1}\mathbb{E}e^{itX_2}.$$

□

Definition 1.3.11 (**Characteristic function of a random vector**)**.** The **characteristic function** $\phi : \mathbb{R}^n \to \mathbb{C}$ of a random vector $\boldsymbol{u} \in \mathbb{R}^n$ is defined as

$$\phi_{\boldsymbol{u}}(\boldsymbol{s}) = \mathbb{E}\left[\exp\left\{i\boldsymbol{s}^\top \boldsymbol{u}\right\}\right] = \int_{\mathbb{R}} \dots \int_{\mathbb{R}} \exp\left\{i\boldsymbol{s}^\top \boldsymbol{u}\right\} dF_{u_1}(u_1) \dots dF_{u_n}(u_n).$$

Sometimes we may write the **joint characteristic function** of two random variables X and Y (for example) as

$$\phi_{X,Y}(s,t) := \mathbb{E}\exp\left\{i(sX + tY)\right\};$$

note that if and only if X and Y we are independent we have

$$\phi_{X,Y}(s,t) = \mathbb{E}\left[\exp\left\{isX\right\}\exp\left\{itY\right\}\right] = \mathbb{E}\left[\exp\left\{isX\right\}\right]\mathbb{E}\left[\exp\left\{itY\right\}\right] = \phi_X(s)\phi_Y(t).$$

Similarly, the joint characteristic function of two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ is

$$\phi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{s},\boldsymbol{t}) = \mathbb{E}\exp\left\{i(\boldsymbol{s}^\top \boldsymbol{X} + \boldsymbol{t}^\top \boldsymbol{Y})\right\},$$

and if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ we are independent we have

$$\phi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{s},\boldsymbol{t}) = \mathbb{E}\left[\exp\left\{i\boldsymbol{s}^\top \boldsymbol{X}\right\}\exp\left\{i\boldsymbol{t}^\top \boldsymbol{Y}\right\}\right] = \mathbb{E}\left[\exp\left\{i\boldsymbol{s}^\top \boldsymbol{X}\right\}\right]\mathbb{E}\left[\exp\left\{i\boldsymbol{t}^\top \boldsymbol{Y}\right\}\right] = \phi_{\boldsymbol{X}}(\boldsymbol{s})\phi_{\boldsymbol{Y}}(\boldsymbol{t}).$$

(See also Theorem 1.3.4.4 part (d) below.)

Proposition 1.3.4.3. Necessary and sufficient conditions for a function to be a characteristic function:

(a) $\phi_X(0) = 1$

(b) $|\phi(t)| \leq 1 \ \forall \ t$

(c) ϕ is uniformly continuous on $\mathbb{R}$

(d) ϕ is positive semidefinite; that is,

$$\sum_{i,j} \phi(t_j - t_k) z_j \overline{z}_k \geq 0 \text{ for all real } t_1, t_2, \ldots, t_n \text{ and complex } z_1, z_2, \ldots, z_n$$

Or, equivalently, or every set of real numbers $t_1, t_2, \ldots, t_n$, the matrix $\phi(t_i - t_j), i, j \in \{1, 2, \ldots, n\}$ is Hermitian and nonnegative definite.

Remark. Relationship between characteristic functions and probability and moment-generating functions:

$$\phi_X(t) = M_X(it) = G_X(e^{it})$$

Theorem 1.3.4.4 (Some useful properties). (a) $X \perp\!\!\!\perp Y \implies \phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$

(b) $Y = aX + b \implies \phi_Y(t) = e^{itb}\phi_X(at)$

(c) $\phi_X^{(k)}(0) = i^k \mathbb{E}(X^k)$

(d) $\phi_{X,Y}(s,t) = \mathbb{E}(e^{isX}e^{itY})$

(e) $X \perp\!\!\!\perp Y \iff \phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$

Theorem 1.3.4.5. Other facts from notes on course website

(a) If $\phi(t)$ is even, $\phi(0) = 1$, ϕ is convex for $t > 0$, and $\lim_{t\to\infty} \phi(t) = 0$, then ϕ is a characteristic function of an absolutely continuous random variable.

(b) If ϕ is a characteristic function and $\phi(t) = 1 + o(t^2), t \to 0$, then $\phi(t) = 1$ for all t. The random variable with such a characteristic function must have zero mean and zero variance. In particular, if $r > 2$, then $\exp(-|t|^r)$ is not a characteristic function.

(c) If $\phi(t) = e^{p(t)}$ is a characteristic function and $p = p(t)$ is a polynomial, then the degree of p is at most 2. For example, $e^{t^2 - t^4}$ is not a characteristic function.

(d) If ξ is absolutely continuous, then $\lim_{|t|\to\infty} |\phi_\xi(t)| = 0$ (Riemann-Lebesgue).

(e) If $\int_{-\infty}^{\infty} |\phi_\xi(t)|\, dt < \infty$, then ξ is absolutely continuous with pdf

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx)\phi(t)dt$$

1.3.5 Continuous Random Variable Distributions

Uniform: $\mathrm{U}(a, b)$

- Probability density function:

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- Cumulative distribution function:

$$F(x) = \Pr(X \leq x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x \leq b \\ 1 & x > b \end{cases}$$

- Probability-generating function:
- Moment-generating function:

$$M_X(t) = \frac{1}{(b-a)t}\big[\exp(bt) - \exp(at)\big]$$

Proof.

$$M_X(t) = \mathbb{E}(\exp(tX)) = \int_a^b \frac{1}{b-a} \cdot \exp(tx)dx = \frac{1}{b-a}\left[\frac{1}{t}\exp(tx)\right]_a^b = \frac{1}{(b-a)t}\big[\exp(bt) - \exp(at)\big]$$

□

- Characteristic function:

$$\frac{2}{(b-a)t}\sin\left(\frac{1}{2}(b-a)t\right)\exp\left(i(a+b)\frac{t}{2}\right)$$

- Expectation: $\mathbb{E}(X) = (b-a)/2$
- Variance: $\text{Var}(X) = (b-a)^2/12$

Proposition 1.3.5.1. If $X \sim U(0,1)$, then $Y = -\log(X) \sim$ Exponential(1).

Proof.

$$\Pr(Y \leq y) = \Pr(-\log(X) \leq y) = \Pr(\log(X) \geq -y) = \Pr(X \geq e^{-y}) = \int_{\exp(-y)}^{\infty} f_X(t)dt = \int_{\exp(-y)}^{1} dt$$

$$= 1 - e^{-y}$$

which is the cdf for an exponential distribution with mean 1. □

Proposition 1.3.5.2. Let $X_1, \ldots, X_n$ be i.i.d. random variables with $X_1 \sim$ Unif$(0,1)$. Then the pdf of $Q = \prod_{i=1}^n X_i$ is $f_Q(y) = \frac{(-\log y)^{n-1}}{(n-1)!}$.

Proof. By Proposition 1.3.5.1, $-\log(X_i) \sim$ Exponential(1) for all $i \in [n]$. By Corollary 1.3.5.8.1,

$$\sum_{i=1}^n -\log(X_i) \sim \text{Gamma}(n,1)$$

$$\iff \quad -\log\left(\prod_{i=1}^n X_i\right) \sim \text{Gamma}(n,1)$$

That is, $-\log\left(\prod_{i=1}^n X_i\right)$ has pdf

$$f(x) = \frac{1}{1^n\Gamma(n)}x^{n-1}e^{-x/1} = \frac{1}{(n-1)!}x^{n-1}e^{-x}.$$

So for all $t \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left(-\log\left(\prod_{i=1}^n X_i\right) \geq t\right) = \int_t^\infty \frac{1}{(n-1)!}x^{n-1}e^{-x}dx \\
\iff \quad &\mathbb{P}\left(\log\left(\prod_{i=1}^n X_i\right) \leq -t\right) = \int_t^\infty \frac{1}{(n-1)!}x^{n-1}e^{-x}dx \\
\iff \quad &\mathbb{P}\left(\prod_{i=1}^n X_i \leq e^{-t}\right) = \int_t^\infty \frac{1}{(n-1)!}x^{n-1}e^{-x}dx \\
\iff \quad &\mathbb{P}\left(Q \leq y\right) = \int_{-\log y}^\infty \frac{1}{(n-1)!}x^{n-1}e^{-x}dx \\
\iff \quad &F_Q(y) = 1 - \int_0^{-\log y} \frac{1}{(n-1)!}x^{n-1}e^{-x}dx
\end{aligned}$$

where $e^{-t} = y \iff t = -\log(y)$ (so $y \in (0, 1]$). Finally, the pdf of Q is

$$\begin{aligned}
\frac{\partial F_Q(y)}{\partial y} &= \frac{\partial}{\partial y}\left(1 - \int_0^{-\log y} \frac{1}{(n-1)!}x^{n-1}e^{-x}dx\right) \\
&= -\left(\frac{1}{(n-1)!}(-\log y)^{n-1}e^{\log y}\right)\cdot\frac{-1}{y} \\
&= \frac{(-\log y)^{n-1}}{(n-1)!}
\end{aligned}$$

□

Normal (or **Gaussian**): $\mathcal{N}(\mu, \sigma^2)$

- Probability density function:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Cumulative distribution function: $F(x) = \Pr(X \leq x) =$
- Probability-generating function:
- Moment-generating function: $M_X(t) = \exp(\mu t + \sigma^2 t^2/2)$
- Characteristic function: $\phi(t) = \exp(i\mu t - (1/2)\sigma^2 t^2)$. Standard normal: $\phi(t) = \exp((-1/2)t^2)$.
- Expectation: $\mathbb{E}(X) = \mu$

- Variance: $\text{Var}(X) = \sigma^2$

Theorem 1.3.5.3. Let $X := \Omega \to \mathbb{R}^n$ be a random Gaussian variable with the **standard Gaussian distribution**:

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

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 $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{R}^n$.

Let $v \in \mathbb{R}^n$. Then $\langle X, v \rangle$ is a mean zero Gaussian with variance $\langle v, v \rangle$.

Remark. Similar to Exercise 1.1.2. You can then use an induction argument to prove the case for the sum of arbitrarily many Gaussian random variables.

Proof. Let $X = (X_1, X_2, \ldots, X_n)$ and $v = (v_1, v_2, \ldots, v_n)$. Then $\langle X, v \rangle = \sum_{i=1}^n X_i v_i$. Let $Y_i = v_i X_i$, so $\langle X, v \rangle = \sum_{i=1}^n Y_i$. Note that $Y_i \sim \mathcal{N}(0, v_i^2)$. Recall that the moment-generating function of a Gaussian random variable is $\mathbb{E}e^{tX} = e^{\mu t + \sigma^2 t^2/2}$, so we have $M_{Y_i}(t) = \exp(v_i^2 t^2/2)$. Next we seek the distribution of $\sum_{i=1}^n v_i X_i = \sum_{i=1}^n Y_i$. From Proposition 1.67, we have

$$M_{Y_1+Y_2+\ldots+Y_n}(t) = \mathbb{E} \exp\left(t \sum_{i=1}^n Y_i\right) = \mathbb{E} \prod_{i=1}^n e^{tY_i} = \prod_{i=1}^n \mathbb{E} e^{ty_i} = \prod_{i=1}^n M_{Y_i}(t).$$

Then

$$\prod_{i=1}^n M_{Y_i}(t) = \prod_{i=1}^n \exp(v_i^2 t^2/2) = \exp\left(\frac{t^2}{2} \sum_{i=1}^n v_i^2\right) = \exp\left(\frac{t^2}{2} \cdot \langle v, v \rangle\right)$$

which is the same as the moment-generating function for a mean zero Gaussian random variable with variance $\langle v, v \rangle$. By the provided uniqueness result from Problem 6 ("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."), the result follows.

□

Proposition 1.3.5.4. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random sample from $\mathcal{N}(\mu, \sigma)$. Then the sum of these n observations $T = \sum_{i=1}^n X_i$ also follows the normal distribution.

Proof.

$$T = \sum_{i=1}^n X_i = X_1 + X_2 + \ldots + X_n$$

Since X_i is $\mathcal{N}(\mu, \sigma)$, we have

$$M_{X_i}(t) = \exp\left(\mu t + \frac{t^2 \sigma^2}{2}\right)$$

Since all observations are independent, we have

$$M_T(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(\exp\left(\mu t + \frac{t^2\sigma^2}{2}\right)\right)^n = \boxed{\exp\left(n\mu t + \frac{t^2 n\sigma^2}{2}\right)}$$

which is the moment generating function for a normal distribution with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.

□

Lognormal: If the random variable X follows the normal distribution with $\mu = 0$, $\sigma^2 = 1$ and $Y = e^X$, then Y has a lognormal distribution.

- Probability density function:

$$f_X(x) = \frac{1}{x\sqrt{2\pi}} \exp\left(\frac{-1}{2}(\log(x))^2\right)$$

Proof.

$$F_Y(y) = \Pr(Y \leq y) = \Pr(e^X \leq y) = \Pr(X \leq \log(y)) \implies F_Y(y) = F_X(\log(y))$$

$$\implies f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(\log(y)) = f_X(\log(y))\frac{1}{y}$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}x^2} \implies f_Y(y) = \frac{1}{y\sqrt{2\pi}} \exp\left(\frac{-1}{2}(\log(y))^2\right)$$

□

- Cumulative distribution function: $F(x) = \Pr(X \leq x) =$
- Probability-generating function:
- Moment-generating function: $M_X(t) =$
- Characteristic function: $\phi(t) =$.
- Expectation: $\mathbb{E}(X) =$
- Variance: $\mathrm{Var}(X) =$

Gamma: $\Gamma(\alpha, \beta)$

- Probability density function:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} = \frac{1}{\Gamma(\alpha, \beta)} x^{\alpha-1} e^{-x/\beta}$$

- Cumulative distribution function:

$$F(x) = \Pr(X \leq x) =$$

- Probability-generating function:
- Moment-generating function:

$$\left(\frac{1/\beta}{1/\beta - t}\right)^{\alpha} = \left(\frac{1}{1-\beta t}\right)^{\alpha}$$

Proof.

$$\mathbb{E}(e^{tX}) = \int_{\mathbb{R}} e^{tx} f(x)dx = \int_0^{\infty} e^{tx} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} dx = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} x^{\alpha-1} e^{-x(1/\beta - t)} dx$$

Using integration by parts, one can find the identity

$$\int_0^{\infty} x^a e^{-bx} dx = \frac{\Gamma(a+1)}{b^{a+1}}.$$

Using this yields

$$\mathbb{E}(e^{tX}) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(1/\beta - t)^{\alpha}} \boxed{= \left(\frac{1/\beta}{1/\beta - t}\right)^{\alpha} = \left(\frac{1}{1-\beta t}\right)^{\alpha}}$$

□

- Characteristic function:
- Expectation: $\mathbb{E}(X) = \alpha\beta$
- Variance: $\text{Var}(X) = \alpha\beta^2$

Proposition 1.3.5.5. Let $X_i \sim \text{Gamma}(\alpha_i, \beta)$ for $i = 1, 2, \ldots, n$. Then

$$\sum_{i=1}^{n} X_i \sim \text{Gamma}\left(\sum_{i=1}^{n} \alpha_i, \beta\right).$$

Proof. Using the moment-generating function for a Gamma distribution as well as Theorem 1.3.3.1(b), we have

$$M_{X_1+\ldots+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} \left(\frac{1}{1-\beta t}\right)^{\alpha_i} = \left(\frac{1}{1-\beta t}\right)^{\sum_{i=1}^{n} \alpha_i}$$

which is the same as the moment-generating function for a $\Gamma(\sum_{i=1}^{n} \alpha_i, \beta)$ distribution. By the provided uniqueness result ("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."), the result follows. □

Proposition 1.3.5.6. Let $X \sim \text{Gamma}(\alpha, \beta)$. Then as $\beta \to \infty$, $X \xrightarrow{d} \mathcal{N}(\alpha\beta, \alpha\beta^2)$.

Proof. See http://www.math.wm.edu/~leemis/chart/UDR/PDFs/GammaNormal1.pdf. □

Proposition 1.3.5.7 (**Stats 100B homework problem**). The method of moments estimators for α and β are

$$\hat{\alpha} = \frac{n\bar{x}^2}{\sum_{i=1}^n (X_i^2 - \bar{x}^2)}$$

$$\hat{\beta} = \frac{1}{n\bar{x}} \sum_{i=1}^n (X_i^2 - \bar{x}^2)$$

Proof.

$$\mathbb{E}(X) = \alpha\beta \implies \hat{\alpha}\hat{\beta} = \bar{x}$$

$$\hat{\alpha} = \frac{\bar{x}}{\hat{\beta}}$$

$$\text{Var}(X) = \alpha\beta^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

$$\implies \hat{\alpha}\hat{\beta}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{x}^2$$

$$\frac{\bar{x}}{\hat{\beta}}\hat{\beta}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{x}^2$$

$$\bar{x}\hat{\beta} = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{x}^2$$

$$\hat{\beta} = \frac{1}{n\bar{x}} \sum_{i=1}^n X_i^2 - \bar{x} = \frac{1}{n\bar{x}} \sum_{i=1}^n X_i^2 - \frac{n\bar{x}^2}{n\bar{x}}$$

$$\hat{\beta} = \frac{1}{n\bar{x}} \sum_{i=1}^n (X_i^2 - \bar{x}^2)$$

$$\implies \hat{\alpha} = \frac{\bar{x}}{\hat{\beta}} = \frac{n\bar{x}^2}{\sum_{i=1}^n (X_i^2 - \bar{x}^2)}$$

□

Proposition 1.3.5.8 (**Some useful formulae and integrals related to the gamma function**).

- Definition:

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \, dx$$

- Beta function (see also the information on the Beta distribution):

$$B(z_1, z_2) := \int_0^1 t^{z_1-1}(1-t)^{z_2-1}\, dt = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}$$

- Generalization of the factorial function:

$$\Gamma(z+1) = z\Gamma(z)$$

  (in particular, for any integer $n \geq 1$, $\Gamma(n) = (n-1)!$.)

- $\Gamma(1/2) = \sqrt{\pi}$.

- Using integration by parts, one can find the identity

$$\int_0^\infty x^a e^{-bx} dx = \frac{\Gamma(a+1)}{b^{a+1}}.$$

- **Euler's Reflection Formula**: For $z \notin \mathbb{Z}$,

$$\Gamma(1-z)\Gamma(z) = \frac{\pi}{\sin(\pi z)}$$

- **Legendre Duplication Formula:**

$$\Gamma(z)\Gamma\left(z+\frac{1}{2}\right) = 2^{1-2z}\sqrt{\pi}\Gamma(2z)$$

- For all $z \in \mathbb{C}$,

$$\lim_{n\to\infty} \frac{\Gamma(n+z)}{\Gamma(n)n^z} = 1.$$

- From 2018 DSO Statistics Screening Exam:

$$\lim_{x\to 0} x\Gamma(x) = 1$$

- See Theorem 1.1.11.2 (Stirling's Formula)

$$\Gamma(n+1) \sim n^n e^{-n}\sqrt{2\pi n}$$

  That is,

$$\lim_{n\to\infty} \frac{n^n e^{-n}\sqrt{2\pi n}}{\Gamma(n+1)} = 1.$$

χ^2_k **(chi-squared)**: special case of a gamma distribution: $\Gamma(k/2, 2)$. Also the sum of k independent standard normally distributed variables.

- Probability density function:

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} = \frac{1}{\Gamma(k/2, 2)} x^{k/2-1} e^{-x/2}$$

For χ_1^2: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x/2} x^{-1/2}$. For χ_2^2: $f(x) = \frac{1}{2} e^{-x^2}, x > 0$.

Proof. See notes from Math 541A. □

- Cumulative distribution function: $F(x) = \Pr(X \leq x) =$
- Probability-generating function:
- Moment-generating function: $(1-2t)^{-k/2}$ for $t < 1/2$
- Characteristic function:
- Expectation: $\mathbb{E}(X) = k/2 \cdot 2 = k$
- Variance: $\text{Var}(X) = k/2 \cdot 2^2 = 2k$

Exponential: (special case of a gamma distribution: $\Gamma(1, \beta)$. Also a special case of a Weibull distribution with $\beta = 1$.)

- Probability density function: $f(x) = \frac{1}{\beta} \exp(-x/\beta) = \lambda e^{-\lambda x}$
- Cumulative distribution function: $F(x) = \Pr(X \leq x) = 1 - e^{-\lambda x}$
- Probability-generating function:
- Moment-generating function: $\frac{\lambda}{\lambda - t}$
- Characteristic function:
- Expectation: $\mathbb{E}(X) = \beta = \lambda^{-1}$

Proof.

$$\mathbb{E}(X) = \int_0^\infty \overline{F}(t) dt = \int_0^\infty e^{-\lambda t} dt = \frac{1}{\lambda}$$

□

- Variance: $\text{Var}(X) = \beta^2 = \lambda^{-2}$

Proof. See Definition 1.1.21 above for the definition of $E(X^n)$. Then using that:

$$\mathbb{E}(X^2) = 2\int_0^\infty t e^{-\lambda t} dt = \frac{2}{\lambda} \int_0^\infty \lambda t e^{-\lambda t} dt = \frac{2}{\lambda} \mathbb{E}(X) = \frac{2}{\lambda^2}$$

Then use $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ to yield the result. □

Remark. Note also the general case:

$$\mathbb{E}(X^n) = n \int_0^\infty t^{n-1} \overline{F}(t) dt = n \int_0^\infty t^{n-1} e^{-\lambda t} dt = \frac{n}{\lambda} \mathbb{E}(X^{n-1})$$

Corollary 1.3.5.8.1 (**Corollary to Proposition 1.3.5.5**). Let $X_1, \ldots, X_n$ be i.i.d. Exponential(β). Then

$$\sum_{i=1}^{n} X_i \sim \text{Gamma}\left(n, 1/\lambda\right) = \text{Gamma}\left(n, \beta\right).$$

Proof. Since $X_i \sim \text{Exponential}(\beta) = \text{Gamma}(1, \beta)$, by Proposition 1.3.5.5 we have

$$\sum_{i=1}^{n} X_i \sim \text{Gamma}\left(\sum_{i=1}^{n} 1, \beta\right) = \text{Gamma}\left(n, \beta\right).$$

□

For much more on exponential random variables, see Section 1.3.6.

Cauchy:

- Probability density function:

$$f(x) = \frac{1}{\pi(1+x^2)} \text{ (standard Cauchy) }, f(x) = \frac{1}{\pi\sigma(1+(x-\mu)^2/\sigma^2)} \text{ (general)}$$

- Cumulative distribution function: $F(x) = \Pr(X \leq x) =$
- Probability-generating function:
- Moment-generating function:
- Characteristic function:
- Expectation: does not exist
- Variance: does not exist (Cauchy distribution has no moments.)

Beta: Recall:

$$B(\alpha, \beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\implies \frac{B(\alpha+1, \beta)}{B(\alpha, \beta)} = \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{\alpha}{\alpha+\beta}$$

- Probability density function:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(x)^{\alpha-1}(1-x))^{\beta-1}$$

- Cumulative distribution function: $F(x) = \Pr(X \leq x) =$
- Probability-generating function:
- Moment-generating function:

$$1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^k}{k!}$$

- Characteristic function:
- Expectation: $\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$

– Variance:

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Student's t_n: can be defined as

$$T = \frac{Z}{\sqrt{V/v}} \implies T \sim t_v$$

where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_v$, and $Z \perp\!\!\!\perp V$.

– Probability density function:

$$f(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\cdot\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$$

– Cumulative distribution function: $F(x) = \Pr(X \leq x) =$

– Probability-generating function:

– Moment-generating function:

– Characteristic function:

– Expectation: $\mathbb{E}(X) = 0$

– Variance: $\text{Var}(X) = n/(n-2)$

Snedecor's F-distribution: can be defined as

$$X = \frac{U/d_1}{V/d_2} \implies X \sim F_{d_1,d_2}$$

where $U \sim \chi^2_{d_1}$, $V \sim \chi^2_{d_2}$, and $U \perp\!\!\!\perp V$.

– Probability density function:

$$f(x) = \frac{t^{(p/2)-1}(p/q)^{p/2}\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)}\Big(1+t(p/q)\Big)^{-(p+q)/2}, \qquad \forall\, t > 0$$

$$= p^{p/2}q^{q/2}\cdot\frac{\Gamma([p+q]/2)}{\Gamma(p/2)\Gamma(q/2)}\cdot\frac{t^{p/2-1}}{(pt+q)^{(p+q)/2}}$$

– Cumulative distribution function: $F(x) =$

– Probability-generating function:

– Moment-generating function:

– Characteristic function:

– Expectation: $\mathbb{E}(X) =$

– Variance: $\text{Var}(X) =$

Proposition 1.3.5.9. 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, with X and Y independent. Then $(X/p)/(Y/q)$ is an F-distributed random variable with p and q degrees of freedom.

Proof. Let $c = p/2$, $d = q/2$. Note that

$$f_X(x) = \frac{1}{\Gamma(c)2^c} x^{c-1} e^{-x/2}, \quad x > 0$$

$$f_Y(y) = \frac{1}{\Gamma(d)2^d} y^{d-1} e^{-y/2}, \quad x > 0$$

We first seek $F_{X/Y}(t)$. Note that $F_{X/Y}(t) = \Pr(X/Y \leq t) = \Pr(X \leq tY)$. Thinking about this graphically, we can calculate this as

$$\Pr(X \leq tY) = \int_0^\infty \int_0^{ty} f_{X,Y}(x,y) dx dy$$

By assumption, X and Y are independent, so

$$f_{X,Y}(x,y) = \frac{1}{\Gamma(c)\Gamma(d)2^{c+d}} x^{c-1} y^{d-1} e^{-x/2} e^{-y/2}$$

Plugging this in we have

$$\Pr(X \leq tY) = \frac{1}{\Gamma(c)\Gamma(d)2^{c+d}} \int_0^\infty \int_0^{ty} x^{c-1} y^{d-1} e^{-x/2} e^{-y/2} dx dy$$

$$= \frac{1}{\Gamma(c)\Gamma(d)2^{c+d}} \int_0^\infty \int_0^{ty} x^{c-1} e^{-x/2} dx \; y^{d-1} e^{-y/2} dy$$

Rather than solving this integral, we next differentiate with respect to t:

$$f_{X/Y}(t) = \frac{\mathrm{d}}{\mathrm{d}t} F_{X/Y}(t) = \frac{1}{\Gamma(c)\Gamma(d)2^{c+d}} \int_0^\infty y(ty)^{c-1} e^{-ty/2} \; y^{d-1} e^{-y/2} dy$$

$$= \frac{t^{c-1}}{\Gamma(c)\Gamma(d)2^{c+d}} \int_0^\infty y^{c+d-1} e^{-y(t+1)/2} dy \tag{1.9}$$

Compare the integrand of (1.9) to the pdf of a Gamma distributed random variable

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$

to see that with some manipulations we can express the integrand of (1.9) as the pdf of a Gamma distributed random variable with parameters $\alpha = c + d, \beta = 2/(t+1)$:

$$f_{X/Y}(t) = \frac{t^{c-1}\Gamma(c+d)}{\Gamma(c)\Gamma(d)(t+1)^{c+d}} \int_0^\infty \frac{1}{[2/(t+1)]^{c+d}\Gamma(c+d)} y^{c+d-1} e^{-y(t+1)/2} dy$$

Then the integral becomes 1, so we have

$$f_{X/Y}(t) = \frac{t^{c-1}\Gamma(c+d)}{\Gamma(c)\Gamma(d)(t+1)^{c+d}}$$

Substitute back in $c = p/2$, $d = q/2$:

$$f_{X/Y}(t) = \frac{\Gamma([p+q]/2)}{\Gamma(p/2)\Gamma(q/2)} \cdot \frac{t^{p/2-1}}{(t+1)^{(p+q)/2}}$$

Now take into account the constants:

$$f_{(X/p)/(Y/q)}(t) = \frac{p}{q} f_{X/Y}\left(\frac{p}{q}t\right) = \frac{p}{q} \cdot \frac{\Gamma([p+q]/2)}{\Gamma(p/2)\Gamma(q/2)} \cdot \frac{(pt/q)^{p/2-1}}{(pt/q+1)^{(p+q)/2}}$$

$$= \frac{t^{(p/2)-1}(p/q)^{p/2}\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)}\Big(1+t(p/q)\Big)^{-(p+q)/2}$$

(or this can be expressed as)

$$= \left(\frac{p}{q}\right)^{p/2} \cdot \frac{\Gamma([p+q]/2)}{\Gamma(p/2)\Gamma(q/2)} \cdot \frac{t^{p/2-1}q^{(p+q)/2}}{(pt+q)^{(p+q)/2}} = p^{p/2}q^{q/2} \cdot \frac{\Gamma([p+q]/2)}{\Gamma(p/2)\Gamma(q/2)} \cdot \frac{t^{p/2-1}}{(pt+q)^{(p+q)/2}}$$

□

Weibull:

- Probability density function: $f(x) = \alpha\beta x^{\beta-1}\exp(-\alpha x^{\beta})$
- Cumulative distribution function: $F(x) = \Pr(X \leq x) = 1 - \exp(-\alpha x^{\beta})$
- Probability-generating function:
- Moment-generating function:
- Characteristic function:
- Expectation: $\mathbb{E}(X) =$
- Variance: $\text{Var}(X) =$

Pareto: (parameters α, x_m)

- Probability density function: $f(x) = \alpha x_m^{\alpha}/x^{\alpha+1}$ for $x \geq 1, 0$ otherwise.
- Cumulative distribution function: $F(x) = 1 - (x_m/x)^{\alpha}$ for $x \geq 1, 0$ otherwise.
- Probability-generating function:
- Moment-generating function:
- Characteristic function:
- Expectation: $\mathbb{E}(X) = \alpha x_m/(\alpha-1)$ for $\alpha > 1, \infty$ otherwise.
- Variance: $\text{Var}(X) = \frac{x_m^2\alpha}{(\alpha-1)^2(\alpha-2)}$ for $\alpha > 2$, ∞ otherwise.

1.3.6 More on Exponential Random Variables

Remark. Recall Proposition 1.3.5.1: If $X \sim U(0,1)$, then $Y = -\log(X) \sim \text{Exponential}(1)$.

Proposition 1.3.6.1. Let X be a random variable. Then X is exponentially distributed if and only if X has the **memoryless** property; that is,

$$\Pr(X > s+t \mid X > t) = \Pr(X > s) \;\; \forall\, s,t \geq 0$$

or, equivalently,

$$\Pr(X > s+t) = \Pr(X > s)\Pr(X > t) \;\; \forall\, s,t \geq 0$$

Proof. See Ross *Introduction to Probability Models* section 5.2.2.

□

Remark. Exponential distributions can be derived as follows: Let X be a nonnegative random variable with cdf F and pdf f. Define the **failure** or **hazard rate**

$$\lambda(t) = \frac{f(t)}{1 - F(t)}$$

The intuition for the hazard rate is as follows: think of X as the lifetime and consider the probability that a unit with age t fails within some timespan $(t, t+h)$ with h small; that is, $\Pr(t < X < t+h \mid X > t)$. Letting $1 - F(t) = \overline{F}(t)$, we have

$$\Pr(t < X < t + h \mid X > t) = \frac{\Pr(t < X < t + h)}{\Pr(X > t)} = \frac{\int_t^{t+h} f(s)ds}{\overline{F}(t)} \approx \frac{f(t)}{\overline{F}(t)} \text{ for small h}$$

If this quantity is constant at λ (i.e. the process is "memoryless," see Proposition 1.3.6.1), we have an exponential distribution:

$$\lambda = \frac{f(t)}{\overline{F}(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}}$$

Indeed, a process is memoryless if and only if it is exponential. To see this, note that

$$\int_0^t \lambda(s)ds = \int_0^t \frac{f(s)}{1 - F(s)}ds$$

Letting $u = 1 - F(s) \implies du = -f(s)ds$, we have

$$= \int_{u=1}^{u=\overline{F}(t)} -\frac{du}{u}ds = \int_{\overline{F}(t)}^1 \frac{du}{u}ds = \log(1) - \log(\overline{F}(t)) = -\log \overline{F}(t)$$

$$\implies \int_0^t \lambda(s)ds = -\log \overline{F}(t) \implies \boxed{\overline{F}(t) = \exp\left(-\int_0^t \lambda(s)ds\right)}$$

If the process is memoryless, we must have $\lambda(s) = \lambda$ (constant). Then $\overline{F}(t) = \exp(-\lambda t)$ and we have the exponential distribution. See the Weibull distribution below for a more general case.

Proposition 1.3.6.2. Let X be an exponential random variable. Then

(a) $\mathbb{E}(X - s \mid X > s) = \mathbb{E}(X)$

(b) $\mathbb{E}(X - s) = \Pr(X > s)\mathbb{E}(X)$ and

Proof. (a) By the memoryless property (Proposition 1.3.6.1), $\Pr(X > t + s \mid X > s) = \Pr(X - s > t \mid X > s) = \Pr(X > t)$. Then we have

$$\mathbb{E}(X - s \mid X > s) = \int_0^\infty \Pr(X - s > t \mid X > s)dt = \int_0^\infty \Pr(X > t)dt = \mathbb{E}(X)$$

(b) By the memoryless property (Proposition 1.3.6.1), $\Pr(X > t+s) = \Pr(X - s > t) = \Pr(X > t)\Pr(X > s)$. Then we have

$$\mathbb{E}(X-s) = \int_0^\infty \Pr(X > t)\Pr(X > s)dt = \Pr(X > s)\int_0^\infty \Pr(X > t)dt = \Pr(X > s)\mathbb{E}(X)$$

□

Proposition 1.3.6.3. Let X be an exponential random variable. Then $X - t \mid X > t$ is identically distributed as X.

Proof. Because X is memoryless, we know by Proposition 1.3.6.1 that $\Pr(X > s+t \mid X > t) = \Pr(X > s)$, which is to say $\Pr(X \leq s+t \mid X > t) = \Pr(X \leq s) \iff \Pr(X - t \leq s \mid X > t) = \Pr(X \leq s)$; that is, $X - t \mid X > t$ and X have identical distributions.

□

Proposition 1.3.6.4. Let X be an exponential random variable. Then $\mathbb{E}[X^2 \mid X > t] = \mathbb{E}[(X+t)^2]$.

Proof. By Proposition 1.3.6.3, X and $X - t \mid X > t$ have identical distributions. That means they have identical variances, so

$$\text{Var}(X - t \mid X > t) = \text{Var}(X) \iff \mathbb{E}[(X-t)^2 \mid X > t] - [\mathbb{E}(X - t \mid X > t)]^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.$$

Using $\mathbb{E}(X - t \mid X > t) = \mathbb{E}(X)$ we have

$$\mathbb{E}[(X-t)^2 \mid X > t] - [\mathbb{E}(X)]^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \iff \mathbb{E}[(X-t)^2 \mid X > t] = \mathbb{E}(X^2)$$

$$\mathbb{E}[(X-t)^2 \mid X > t] - [\mathbb{E}(X)]^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \iff \mathbb{E}[X^2 - 2tX + t^2 \mid X > t] = \mathbb{E}(X^2)$$

$$\iff \mathbb{E}[X^2 \mid X > t] - 2t\mathbb{E}(X \mid X > t) + t^2 = \mathbb{E}(X^2) \iff \mathbb{E}[X^2 \mid X > t] = \mathbb{E}(X^2) + 2t\mathbb{E}(X \mid X > t) - t^2$$

Using $\mathbb{E}(X \mid X > t) = t + \mathbb{E}(X)$, we have

$$\mathbb{E}[X^2 \mid X > t] = \mathbb{E}(X^2) + 2t(t + \mathbb{E}(X)) - t^2 = \mathbb{E}(X^2) + 2t\mathbb{E}(X) + t^2 = \mathbb{E}[(X+t)^2]$$

□

Example 1.3.2. (ISE 620): Enter a bank with one teller, 5 present upon your arrival, all service times are exponential with parameter 1. T is time you spend in system.

$$\mathbb{E}(T) = R + \sum_{i=1}^{t} S_i \implies \mathbb{E}(T) = \mathbb{E}(R) + \sum_{i=1}^{5} \mathbb{E}(S_i)$$

$$\mathbb{E}(T) = \mathbb{E}(R) + \frac{5}{\lambda} = \frac{6}{\lambda}$$

Example 1.3.3. (ISE 620): Suppose $X \sim$ Exponential(λ) and Y Exponential(μ). Then

$$\Pr(X < Y) = \int_0^\infty \Pr(Y > X \mid X = x) \cdot \lambda e^{-\lambda x} dx = \int_0^\infty e^{-\mu x} \lambda e^{-\lambda x} dx = \lambda \int_0^\infty e^{-(\lambda+\mu)x} dx = \boxed{\frac{\lambda}{\lambda + \mu}}$$

Remark. (ISE 620). The following is an intuitive explanation for why $\min\{X, Y\}$ is distributed exponentially if X and Y are exponential: Suppose we require two machines with failure times X and Y distributed exponentially, and we need both to work. If at time t they are both still working, the remaining expected time is the same as initially because of the memoryless property (Proposition 1.3.6.1). Thus $\min\{X, Y\}$ is also memoryless, and exponential. Do the math:

$$\Pr(\min\{X, Y\} > t) = \Pr(X > t.Y > t) = e^{-\lambda t} e^{-\mu t} = e^{-(\lambda+\mu)t}$$

The same is not true for maximum:

$$\Pr(\max\{X, Y\} > t) = (1 - e^{-\lambda t})(1 - e^{-\mu t})$$

which is not exponential. Intuitively, if we only need one machine to be working, just because we are working at time t does not mean we are in the same position as we were initially; one machine could have failed. Now the general case: $X_1, \ldots, X_n \sim$ Exponential(λ_i).

$$\Pr(\min\{X_i\} > t) = \Pr(X_1 > t \cap \ldots \cap X_n > t) = \prod_{i=1}^n \Pr(X_i > t) = \exp\left(t \sum_{i=1}^n \lambda_i\right)$$

so the minimum of an arbitrary number of exponential random variables is distributed exponentially. Now

$$\Pr(X_i < \min_{j \neq i}\{X_j\}) = \Pr(X_i < Z)$$

where $Z \sim$ Exponential($\sum_{j \neq i} \lambda_j$. So by the other example, this probability is $\lambda_i / \sum_{j=1}^n \lambda_j$.

Proposition 1.3.6.5. (ISE 620). If X and Y are independent exponential random variables, $\min\{X, Y\}$ is independent of whether X or Y is smaller. That is

$$\Pr(\min\{X, Y\} > t \mid X < Y) = \Pr(\min\{X, Y\} > t).$$

Proof.

$$\Pr(\min\{X, Y\} > t \mid X < Y) = \frac{\Pr(Y > X > t)}{\Pr X < Y)} = \frac{1}{\lambda/(\lambda + \mu)} \cdot \int_0^\infty \Pr(Y > X > t \mid X = s) \lambda e^{-\lambda s} ds$$

$$= \frac{\lambda + \mu}{\lambda} \int_t^\infty e^{-\mu s} \lambda e^{-\lambda s} ds = \int_t^\infty (\lambda + \mu) e^{-(\lambda+\mu)s} ds = e^{-(\lambda+\mu)t} = \Pr(\min\{X, Y\} > t)$$

□

Example 1.3.4. (ISE 620). You have two servers with exponential service times with parameters μ_1 and μ_2. You wait until the first person is done serving a customer, then you are served by that server. T is the time you spend in the system. What is $E(T)$?

$$\mathbb{E}(T) = \mathbb{E}(T \mid \mu_1)\Pr(\mu_1) + \mathbb{E}(T \mid \mu_2)\Pr(\mu_2) = \mathbb{E}(T \mid \mu_1)\frac{\mu_1}{\mu_1+\mu_2} + \mathbb{E}(T \mid \mu_2)\frac{\mu_2}{\mu_1+\mu_2}$$

$$= \left(\frac{1}{\mu_1} + \mathbb{E}(X_1 \mid X_1 < X_2)\right)\frac{\mu_1}{\mu_1+\mu_2} + \left(\frac{1}{\mu_2} + \mathbb{E}(X_2 \mid X_2 < X_1)\right)\frac{\mu_2}{\mu_1+\mu_2}$$

$$= \left(\frac{1}{\mu_1} + \mathbb{E}(\min\{X_1, X_2\})\right)\frac{\mu_1}{\mu_1+\mu_2} + \left(\frac{1}{\mu_2} + \mathbb{E}(\min\{X_1, X_2\})\right)\frac{\mu_2}{\mu_1+\mu_2}$$

$$= \left(\frac{1}{\mu_1} + \frac{1}{\mu_1+\mu_2}\right)\frac{\mu_1}{\mu_1+\mu_2} + \left(\frac{1}{\mu_2} + \frac{1}{\mu_1+\mu_2}\right)\frac{\mu_2}{\mu_1+\mu_2}$$

Proposition 1.3.6.6. (Equation (5.5) in Sheldon Ross *Introduction to Probability Models.*) For independent exponential variables T_1 and T_2 with means $1/\lambda_1$ and $1/\lambda_2$, $\Pr(T_2 < T_1) = \frac{\lambda_2}{\lambda_1+\lambda_2}$.

Proof.

$$\Pr(T_2 < T_1) = \int_0^\infty \Pr(T_2 < T_1 \mid T_2 = t)\lambda_2 e^{-\lambda_2 t}dt = \int_0^\infty \Pr(t < T_1)\lambda_2 e^{-\lambda_2 t}dt = \int_0^\infty e^{-\lambda_1 t}\lambda_2 e^{-\lambda_2 t}dt$$

$$= \int_0^\infty \lambda_2 e^{-(\lambda_1+\lambda_2)t}dt = \frac{\lambda_2}{\lambda_1+\lambda_2}$$

□

1.3.7 Multivariate Gaussian (Normal) Distributions

Definition 1.3.12. From http://pluto.huji.ac.il/~pchiga/teaching/MathStat/SIAnotes2013.pdf (definition 2b6): A random vector $X = (X_1, X_2)$ is Gaussian with mean $\mu = (\mu_1, \mu_2)$ and the covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

if it has a joint pdf of the form

$$f_X(x) = \frac{1}{2\pi\sigma_2\sigma_2\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2}\frac{1}{1-\rho^2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2)} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_1} + \frac{(x_2-\mu_2)^2}{\sigma_2^2)}\right)\right]$$

for $x \in \mathbb{R}^2$.

Definition 1.3.13. (From Heilman Math 541A; Vector-valued Gaussian random variables.) $Z := (Z_1, \dots, Z_d) \in \mathbb{R}^d$ is a **Gaussian random vector** if for all $v \in \mathbb{R}^d$, $\langle v, Z \rangle$ is a Gaussian random variable (in the usual sense). Equivalently, any linear combination of $Z_1, \dots, Z_d$ is a Gaussian random variable. Also, tZ has covariance matrix $a_{ij} = \mathbb{E}\big[(Z_i - \mathbb{E}(Z_i))(Z_j - \mathbb{E}(Z_j)\big], 1 \leq i < j \leq d$.

Definition 1.3.14. A random vector $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ is Gaussian with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)$ and covariance matrix $\boldsymbol{\Sigma} = \big[\boldsymbol{\sigma}_{ij}\big]$ if it has a joint pdf of the form

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right).$$

Definition 1.3.15 (Multivariate Normal distribution, Math 541B definition). $Y \in \mathbb{R}^d$ is multivariate normal if and only if there exist $B \in \mathbb{R}^{d \times k}$ and $C \in \mathbb{R}^d$ such that $Y = BX + C$ where $X = (X_1, \dots, X_k)$ with $X_j \sim \mathcal{N}(0, 1)$ for all $j \in [k]$ and $X_j \perp\!\!\!\perp X_i$ for all $i, j \in [k], i \neq j$.

Proposition 1.3.7.1. [**Conditional distribution of one Gaussian random variable on another. From** http://pluto.huji.ac.il/~pchiga/teaching/MathStat/SIAnotes2013.pdf **(Proposition 3c1)**]

Let X be a Gaussian random variable in $\mathbb{R}^2$ such that

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix}\right)$$

Then $f_{X_1|X_2}(x_1; x_2)$ is Gaussian with the (conditional) mean

$$\mathbb{E}(X_1 \mid X_2 = x_2) = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2 - \mu_2) = \mu_1 + \frac{\sigma_{12}^2}{\sigma_2^2}(x_2 - \mu_2)$$

and the (conditional) variance

$$\mathrm{Var}(X_1 \mid X_2 = x_2) = \sigma_1^2(1 - \rho^2) = \sigma_1^2 - \frac{\sigma_{12}^4}{\sigma_2^2}$$

That is, the conditional distribution of X_1 given $X_2 = x_2$ is

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), (1 - \rho^2)\sigma_1^2\right) = \mathcal{N}\left(\mu_1 + \frac{\sigma_{12}^2}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^4}{\sigma_2^2}\right)$$

Proposition 1.3.7.2 (**Generalization of Proposition 1.3.7.1; from** https://en.wikipedia.org/wiki/Multivariate
Suppose

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right)$$

where $\boldsymbol{\mu}_1 \in \mathbb{R}^q$, $\boldsymbol{\mu}_2 \in \mathbb{R}^{p-q}$, $\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{q \times q}$, $\boldsymbol{\Sigma}_{12} \in \mathbb{R}^{(p-q) \times q}$, and $\boldsymbol{\Sigma}_{22} \in \mathbb{R}^{(p-q) \times (p-q)}$. Then

$$\{\boldsymbol{X}_1 \mid \boldsymbol{X}_2\} \sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T\right) \tag{1.10}$$

Remark. Note that the conditional covariance matrix in (1.10) is the Schur complement (see Sections **??** and **??**) of $\mathbf{\Sigma}_{22}$ in $\mathbf{\Sigma}$.

Remark. A similar Berry-Esseen theorem (Theorem **??**) exists for multivariate Gaussian distributions.

Remark. Note that this matches the OLS coefficients in the univariate case. In other words, the univariate OLS formula can be derived using only this fact.

Recall Theorem 1.3.1.4: if two variables are bivariate normal, they are independent if and only if their covariance

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x, y)dxdy$$

equals 0.

Proposition 1.3.7.3 (Stats 100B homework problem). Let $(X_i, Y_i), i = 1, 2, \ldots, n$, be a random sample from a bivariate normal distribution, where $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are independent.. Then the joint moment generating function of $(\bar{X}, \bar{Y})$ is

$$\exp\left(\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\frac{1}{n}\mathbf{\Sigma}\boldsymbol{t}\right) \sim \boxed{\mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{n}\mathbf{\Sigma}\right)}$$

and $(\bar{X}, \bar{Y})$ is bivariate normal with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\frac{1}{n}\mathbf{\Sigma}$.

Proof. Let $\boldsymbol{W_i} = (X_i, Y_i)$. Then the moment-generating function of $\boldsymbol{W_i} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ is given by

$$M_{\boldsymbol{W_i}}(\boldsymbol{t}) = \exp(t'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\mathbf{\Sigma}\boldsymbol{t})$$

where $\boldsymbol{\mu}$ is a two-dimensional column vector and $\mathbf{\Sigma}$ is a two-by-two matrix. Then

$$(\bar{X}, \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W_i}(\boldsymbol{t}) = \sum_{i=1}^{n} \frac{1}{n}\boldsymbol{W_i}(\boldsymbol{t})$$

Since $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are independent,

$$M_{(\bar{X},\bar{Y})}(\boldsymbol{t}) = \prod_{i=1}^{n} M_{\boldsymbol{W_i}}\left(\frac{1}{n}\boldsymbol{t}\right)$$

$$= \left[\exp\left(\frac{1}{n}\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\frac{1}{n}\boldsymbol{t}'\mathbf{\Sigma}\frac{1}{n}\boldsymbol{t}\right)\right]^n = \exp\left(n\frac{1}{n}\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}n\frac{1}{n^2}\boldsymbol{t}'\mathbf{\Sigma}\boldsymbol{t}\right)$$

$$= \exp\left(\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\frac{1}{n}\mathbf{\Sigma}\boldsymbol{t}\right) \sim \boxed{\mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{n}\mathbf{\Sigma}\right)}$$

Thus $(\bar{X}, \bar{Y})$ is bivariate normal with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\frac{1}{n}\boldsymbol{\Sigma}$.

□

Example 1.3.5 (**Example from 541B**)**.** Let $X \sim \mathcal{N}(0,1)$. Define

$$Y := \begin{cases} X & |X| \leq a \\ -X & |X| > a \end{cases}$$

Then (a) for any $a > 0$, $Y \sim \mathcal{N}(0,1)$, (b) $\exists a > 0$ such that $\mathbb{E}(XY) = 0$. However, X and Y are clearly not independent. In particular the joint distribution of X and Y is not multivariate normal.

1.4 Exponential Families

(For more notes, see Section **??** on generalized linear models.)

Definition 1.4.1. Informally, an **exponential family** is a family of probability distributions that depends on a parameter $w \in \mathbb{R}^k$.

Formally, let n, k be positive integers and let μ be a *measure* on $\mathbb{R}^n$ (that is, a probability law that does not necessarily sum to 1). Let $t_1, \ldots, t_k : \mathbb{R}^n \to \mathbb{R}$. Let $h : \mathbb{R}^n \to [0, \infty]$, and assume h is not identically zero. For any $w = (w_1, \ldots, w_k) \in \mathbb{R}^k$, define

$$a(w) := \log \left[\int_{\mathbb{R}^n} h(x) \exp\left(\sum_{i=1}^{k} w_i t_i(x) \right) d\mu(x) \right], \quad \forall x \in \mathbb{R}^n$$

The set $\{w \in \mathbb{R}^k\}$ is called the **natural parameter space**. On this set, the function

$$f_w(x) := h(x) \exp\left(\sum_{i=1}^{k} w_i t_i(x) - a(w) \right), \quad \forall x \in \mathbb{R}^n$$

satisfies $\int_{\mathbb{R}^n} f_w(x) d\mu(s) = 1$ (by the definition of $a(w)$). (Why?

$$\int_{\mathbb{R}^n} f_w(x) d\mu(s) = \frac{\int_{\mathbb{R}^n} h(x) \exp\left(\int_{i=1}^{k} w_i t_i(x) \right) d\mu(x)}{\int_{\mathbb{R}^n} h(x) \exp\left(\int_{i=1}^{k} w_i t_i(x) \right) d\mu(x)} = 1.)$$

So, the set of functions (which can be interpreted as probability density functions, or as probability mass functions according to μ) $\{f_w : \theta \in \Theta : a(w(\theta)) < \infty\}$ is called a k**-parameter exponential family in canonical form.**

More generally, let $\Theta \in \mathbb{R}^k$ be any set an let $w : \Theta \to \mathbb{R}^k$. We define a **$k$-parameter exponential family** to be a set of functions $\{f_\theta : \theta \in \Theta\}$, where

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

satisfies

Also, an exponential family is called **curved** if the dimension of Θ is less than k.

Example 1.4.1. A Gaussian random variable has mean μ and standard deviation σ, and is in an exponential family:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), \quad \mu \in \mathbb{R}, \sigma > 0$$

In this case, $k = 2$.

$$\vdots$$

(Continuation of example 3.2.) If we instead write this in canonical form, we get

$$a(w) = \frac{\mu^2}{2\sigma^2} + \log(\sigma) = \left(\frac{\mu}{\sigma^2}\right)^2 \left[(-2)\frac{(-1)}{\sigma^2}\right]^{-1} - \frac{1}{2}\log\left((-2)\frac{(-1)}{2\sigma^2}\right) = \frac{w_1^2}{4w_2} - \frac{1}{2}\log(-2w_2)$$

That is, in this case you can get rid of the thetas and write this in canonical form. Define

$$f_w(x) = h(x) \exp\left(\sum_{i=1}^{2} w_i t_i(x) - a(w) \right), \quad \forall x \in \mathbb{R}$$

where w ranges in $\{(w_1, w_2) \in \mathbb{R}^2 : w_2 > 0\}$.

Remark. (notes remark 3.4) **(location family.)** Let X be a random variable with density $f : \mathbb{R} \to \mathbb{R}$. Let $\mu \in \mathbb{R}$. Then the densites $\{f(x + \mu\}_{\mu \in \mathbb{R}}$ are called a **location family** of X. This family may or may not be an exponential family.

Remark. (notes remark 3.6) **(Scale family.)** Let X be a random variable with density $f : \mathbb{R} \to \mathbb{R}$. Let $\sigma > 0$. The family of densities $\{\sigma^{-1} f(x/\sigma)\}_{\sigma > 0}$ is called a **scale family.** Note

$$\int_{\mathbb{R}} \sigma^{-1} f(x/\sigma) dx = \int_{\mathbb{R}} f(y) dy = 1$$

(subsituting $y = x/\sigma, dy = dx/\sigma$).

Remark. (notes remark 3.7) **(Location and scale family.)** The family of densities $\{\sigma^{-1}f((x+\mu)/\sigma)\}_{\sigma>0,\mu\in\mathbb{R}}$ is called a **location and scale family.** This family may or may not be an exponential family (although Gaussian random variables are one example where it is.)

Exercise 4. Try to write a binomial random variable with parameters n and p as a two-parameter exponential family. (Note: it's impossible to do, but instructive to try.)

$$\Pr(X=x)=\binom{n}{x}p^x(1-p)^{n-x}=\frac{n!}{(n-x)!x!}p^x(1-p)^{n-x}=\frac{n!}{(n-x)!x!}\exp\big(x\log(p)\big)(1-p)^{n-x}=\dots$$

It turns out you can do it with one parameter with p, but not with two parameters with n.

Example 1.4.2 (**Example 3.15 in 541A notes**)**.** Write a binomial random variable with parameters n and p as an exponential family (when n is fixed), then take derivatives in p.

Recall

$$\Pr(X=x)=\binom{n}{x}p^x(1-p)^{n-x},\quad 0 \text{ for any other } x.$$

Keep n fixed and look at p.

$$\binom{n}{x}p^x(1-p)^{n-x}=\binom{n}{x}\exp\big(x\log p+(n-x)\log(1-p)\big)=\binom{n}{x}\exp\big[x\log(p/(1-p))-(-1)n\log(1-p)\big]$$

Define $h:\mathbb{R}\to\mathbb{R}$ so that

$$h(x)=\begin{cases}\binom{n}{x} & 0\le x\le n \text{ is an integer}\\ 0 & \text{otherwise}\end{cases}$$

Let $\theta:=p,\Theta:=(0,1)$, $t(x):=x$, $w(\theta):=\log(\theta/(1-\theta)),a(w(\theta))=-n\log(1-\theta)$.

So, $f_\theta(x):=h(x)\exp\big[w(\theta)t(x)-a(w(\theta))\big],\forall x\in\mathbb{R}$.

Now we will take derivatives. As in previous example (last class),

$$e^{-a(w(\theta))}\frac{\partial}{\partial\theta_1}e^{a(w(\theta))}=\mathbb{E}_\theta\bigg(\sum_{i=1}^k\frac{\partial w_i}{\partial\theta_1}t_i\bigg)$$

So in this case

$$e^{-a(w(\theta))}\frac{\partial}{\partial\theta_1}e^{a(w(\theta))}=\mathbb{E}_\theta\bigg(\frac{\mathrm{d}}{\mathrm{d}\theta}w(\theta)t\bigg).$$

Plugging in yields

$$(1-\theta)^n \frac{\mathrm{d}}{\mathrm{d}\theta}(1-\theta)^{-n} = \frac{n}{1-\theta}$$

on the left side and

$$= \left(\frac{1}{\theta} + \frac{1}{1-\theta}\right)\mathbb{E}_\theta(X) = \mathbb{E}_\theta(X)\left(\frac{1}{\theta(1-\theta)}\right)$$

on the right side which yields

$$\mathbb{E}_\theta(X) = \frac{n}{1-\theta}(\theta(1-\theta)) = n\theta \implies \mathbb{E}(X) = pn$$

Remark. This is a probability mass function since it is defined only on integers.

Example 1.4.3 (**GSBA 604**)**.** Suppose $Y \sim \text{Poisson}(\mu)$. Express it as an exponential family.

Solution

$$\mathbb{P}_\mu(Y = y) = \frac{e^{-\mu}\mu^y}{y!}, \qquad y =\in \{0, 1, \ldots\}$$

$$= \frac{e^{y\log\mu - \mu}}{y!}$$

So the natural parameter η is $\log\mu$, the normalizing function $\psi(\eta) = \mu$, and the carrier density $c(y) = 1/y!$.

1.4.1 Differential identities (Generalization of Moment-Generating Functions)

Recall that the moment-generating function for a Gaussian random variable is

$$\mathbb{E}(e^{tX}) = e^{t^2/2} \quad \forall t \in \mathbb{R}$$

Consequently,

$$\left.\frac{\mathrm{d}^m}{\mathrm{d}t^m}\right|_{t=0} \mathbb{E}(e^{tX}) = \mathbb{E}(X^m)$$

for any integer $m > 0$. We can do a similar thing for an exponential family—we can differentiate the parameters of exponential families and find out information about the exponential family. As in Definition 1.4.1, let

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

Define

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

Question: Is $a(w)$ differentiable?

Lemma 1.4.1.1 (Lemma 3.8 in 541A notes)**.** The function $a(w)$ is continuous and has continuous partial derivatives of all orders on the interior of W. Moreover, we can compute these derivatives by differentiating under the integral sign.

Proof. We prove only the case of a first order partial derivative. Consider the case of the partial derivative with respect to w_1 at w in the interior of W. Let $e_1 = (1, 0 \dots, 0) \in \mathbb{R}^k$. Since the exponential function is analytic, it suffices to show that the partial derivative of $e^{a(w)}$ exists in the direction e_1. We form the difference quotient for $e^{a(w)}$ as follows:

$$\frac{\exp[a(w + \epsilon e_1)] - \exp[a(w)]}{\epsilon} = \frac{1}{\epsilon} \int_{\mathbb{R}^n} h(x) \left[\exp\left(\epsilon t_1(x) + \sum_{i=1}^k w_i t_i(x) \right) - \exp\left(\sum_{i=1}^k w_i t_i(x) \right) \right] d\mu(x)$$

$$= \int_{\mathbb{R}^n} h(x) \frac{\exp[\epsilon t_1(x)] - 1}{\epsilon} \exp\left(\sum_{i=1}^k w_i t_i(x) \right) d\mu(x). \tag{1.11}$$

By the Mean Value Theorem,

$$\frac{|e^a - 1|}{a} = \frac{|e^a - e^0|}{a - 0} \leq e^a \text{ if } a > 0, \text{ or } 1 \text{ if } a < 0$$

so we have

$$|e^a - 1| \leq |a| \max\{1, e^a\} \leq |a| e^{|a|} \leq e^{2|a|} \leq e^{2a} + e^{-2a} \quad \forall a \in \mathbb{R}.$$

In particular, if $a = \epsilon t_1(x)$, we have

$$|e^{\epsilon t_1(x)} - 1| \leq e^{2\epsilon t_1(x)} + e^{-2\epsilon t_1(x)} \tag{1.12}$$

Therefore, examining the integrand of (1.11), we have

$$\left| h(x) \frac{\exp[\epsilon t_1(x)] - 1}{\epsilon} \exp\left(\sum_{i=1}^k w_i t_i(x) \right) \right| \leq h(x) \left| \frac{\exp[\epsilon t_1(x)] - 1}{\epsilon} \right| \exp\left(\sum_{i=1}^k w_i t_i(x) \right)$$

$$\leq \text{ (by (1.12))} \; h(x) \big(e^{2\epsilon t_1(x)} + e^{-2\epsilon t_1(x)} \big) \exp\left(\sum_{i=1}^k w_i t_i(x) \right)$$

$$\vdots$$

$$= h(x)|t_1(x)| \exp\left(\epsilon t_1(x) + \sum_{i=1}^{k} w_i t_i(x)\right) + h(x)|t_1(x)| \exp\left(-\epsilon t_1(x) + \sum_{i=1}^{k} w_i t_i(x)\right) \tag{1.13}$$

If both of the expressions in (1.13) always have finite expected value uniformly for all $\epsilon > 0$, we will be done by the Dominated Convergence Theorem (Theorem 1.4.1.2). (Assuming those things are bounded uniformly in expectation then the limit of the integrals is the integral of the limits.)

□

Remark. Notes from proof of Lemma 3.8.

$\left|\frac{e^a - 1}{a}\right| \leq e^a + e^{-a}$. Then

$$\left|h(x) \exp\left(\epsilon t_1(x) + \sum_{i=1}^{k} w_i t_i(x)\right) - \exp\left(\sum_{i=1}^{k} w_i t_i(x)\right)\right| \leq h(x) \exp\left(\sum_{i=1}^{k} w_i t_i(x)\right) |t_1(x)| \Big(\exp(\epsilon t_1(x) + \exp(-\epsilon t_1(x)\Big)$$

Theorem 1.4.1.2 (Dominated convergence theorem (Math 541A))**.** Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ and $Y : \Omega \to [0, \infty)$ such that $|X_i| \leq Y$ for all $i \geq 1$ and $\mathbb{E}(Y) < \infty$. Assume $X_1, X_2, \ldots$ converges almost surely to $X : \Omega \to \mathbb{R}$. Then

$$\lim_{i \to \infty} \mathbb{E}(x_i) = \mathbb{E}\big(\lim_{i \to \infty} X_i\big) = \mathbb{E}(X)$$

Corollary 1.4.1.2.1 (Corollary 3.11 in 541A notes.)**.** Let $\epsilon > 0$. Let $X : \Omega \to \mathbb{R}$ be a random variable such that $\mathbb{E}(e^{wX}) < \infty$ for all $w \in (-\epsilon, \epsilon)$. Then for any integer $n \geq 1$, $\mathbb{E}(X^n)$ exists and

$$\frac{\mathrm{d}^n}{\mathrm{d}w^n}\bigg|_{w=0} e^{wX} = \mathbb{E}(X^n).$$

Proof. Apply Lemma 1.4.1.1 when $\mu = \mathbb{P}, h = 1, k = 1, t(x) = x$: we see that

$$a(w) = \log \int_{\mathbb{R}^n} e^{wx} d\mathbb{P}(x)$$

□

Proof (GSBA 604 proof, just for first two moments). Notation: The density is

$$g_\eta(y) = c(y) \exp\{\eta y - \psi(\eta)\}$$

where $c(y)$ is the **carrier density**, η is the **natural parameter**, and $\psi(\cdot)$ is the **normalizing function**, chosen so that it is a proper density: that is, choose $\psi(\eta)$ such that

$$\int_{\mathbb{R}} g_\eta(y)\ dy = 1,$$

so

$$e^{-\psi(y)} = \left(\int_{\mathbb{R}} c(y) e^{\eta y} \, dy \right)^{-1}$$

Now

$$\int g_\eta(y) dy = 1$$

differentiate with respect to η:

$$\int e^{\eta y \cdot - \psi(\eta)} [y - \frac{\mathrm{d}\psi(\eta)}{\mathrm{d}\eta}] c(y) \, dy = 0$$

$$\implies \int y g_\eta(y) \, dy - \frac{\mathrm{d}\psi(\eta)}{\mathrm{d}\eta} \int e^{\eta y - \psi(\eta)} c(y) \, dy = 0 \quad (1.14)$$

so then $\mathbb{E}(y) - \frac{\mathrm{d}\psi(\eta)}{\mathrm{d}\eta} = 0 \implies \mathbb{E}(y) = \frac{\mathrm{d}\psi(\eta)}{\mathrm{d}\eta}$.

Next,

$$\int y \frac{\mathrm{d}g_\eta(y)}{\mathrm{d}\eta} \, dy - \frac{\mathrm{d}\psi^2(\eta)}{\mathrm{d}\eta^2} = 0$$

$$\implies \int y \cdot e^{\eta y - \psi(\eta)} (y - \frac{\mathrm{d}\psi(\eta)}{\mathrm{d}\eta}) c(y) \, dy - \frac{\mathrm{d}\psi^2(\eta)}{\mathrm{d}\eta^2} = 0$$

$$\implies \int y^2 g_\eta(y) dy - \frac{\mathrm{d}\psi(\eta)}{\mathrm{d}\eta} \underbrace{\int y e^{\eta y - \psi(\eta)} c(y) \, dy}_{\frac{\mathrm{d}\psi(\eta)}{\mathrm{d}\eta}} - \frac{\mathrm{d}\psi^2(\eta)}{\mathrm{d}\eta^2} = 0$$

$$\implies \mathbb{E}(Y^2) - \left[\frac{\mathrm{d}\psi(\eta)}{\mathrm{d}\eta} \right]^2 - \frac{\mathrm{d}\psi^2 \eta)}{\mathrm{d}\eta} = 0$$

$$\implies \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = \frac{\mathrm{d}\psi^2(\eta)}{\mathrm{d}\eta^2}.$$

□

Remark. Notes from example 3.13.

$$e^{-a(w)} \frac{\partial}{\partial w_2} e^{a(w)} = \int_{\mathbb{R}} t_1(x) h(x) \exp \left(\sum_{i=1}^{k} w_i t_i(x) - a(w) \right) dx$$

$$= \int_{\mathbb{R}} t_1(x) f_x(x) dx \quad (1.15)$$

So by chain rule

$$e^{-a(w(\theta))}\frac{\partial}{\partial\theta_1}e^{a(w(\theta))} = e^{-a(w(\theta))}\sum_i \frac{\partial e^{a(w)}}{\partial w_i}\frac{\partial w_i}{\partial\theta_1} \tag{1.16}$$

$$= (\text{by (1.15)})\ \sum_{i=1}^{k}\frac{\partial w_i}{\partial\theta_1}\mathbb{E}_\theta t_i = \mathbb{E}_\theta\Big(\sum_{i=1}^{k} t_i\frac{\partial w_i}{\partial\theta_1}\Big)$$

Example 1.4.4. Recall Example 3.3. Gaussian, mean μ, variance σ^2. $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$, $t_1(x) = x, t_2(x) = x^2$k $w_1(\theta) = \theta_1/\theta_2 = \mu/\sigma^2$, $w_2(\theta) = -1/(2\sigma^2) = -1/(2\theta_2)$, $a(w(\theta)) = \theta_1^2/(2\theta_2) + (1/2)\log(\theta_2) = \mu^2/(2\sigma^2) + \log(\sigma^2)$.

Complete both sides of (1.16).

$$e^{-a(w(\theta))}\frac{\partial}{\partial\theta_1}e^{a(w(\theta))} = \frac{\theta_1}{\theta_2} = \frac{\mu}{\sigma^2}$$

$$\frac{\partial}{\partial\theta_1}e^{a(w(\theta))} = \frac{\mathrm{d}}{\mathrm{d}\theta_1}\sqrt{\theta_2}\exp\left(\frac{\theta_1^2}{2\theta_2}\right) = \frac{\theta_1}{\theta_2}\exp(a(w(\theta))$$

Right side:

$$\frac{\mathrm{d}w_1}{\mathrm{d}\theta_2} = \frac{1}{\theta_2}, \quad \frac{\mathrm{d}w_2}{\mathrm{d}\theta_1} = 0$$

$$\implies \mathbb{E}_\theta\Big(\sum_{i=1}^{k} t_i\frac{\mathrm{d}w_i}{\mathrm{d}\theta_1}\Big) = \mathbb{E}_\theta(t_1\frac{\mathrm{d}w_1}{\mathrm{d}\theta_1} + t_2\frac{\mathrm{d}w_2}{\mathrm{d}\theta_1})$$

$$= \mathbb{E}_\theta(x/\theta_2) \implies \mathbb{E}_\theta(x) = \theta_1 = \mu$$

Theorem 1.4.1.3 (Theorem 3.4.2 from Casella and Berger). If X is a random variable in an exponential family, then

$$\mathbb{E}\left(\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(X)\right) = \frac{\partial}{\partial\theta_j}a(w(\theta)). \tag{1.17}$$

Example 1.4.5. Recall Example 3.3 again: Gaussian, mean μ, variance σ^2. $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$, $t_1(x) = x, t_2(x) = x^2$k $w_1(\theta) = \theta_1/\theta_2 = \mu/\sigma^2$, $w_2(\theta) = -1/(2\sigma^2) = -1/(2\theta_2)$, $a(w(\theta)) = \theta_1^2/(2\theta_2) + (1/2)\log(\theta_2) = \mu^2/(2\sigma^2) + \log(\sigma^2)$.

Complete both sides of (1.17).

$$\frac{\partial}{\partial\theta_1}a(w(\theta)) = \frac{\partial}{\partial\theta_1}\big(\theta_1^2/(2\theta_2) + (1/2)\log(\theta_2)\big) = \frac{\theta_1}{\theta_2}.$$

Left side:

$$\mathbb{E}\left(\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_1}t_i(X)\right) = \mathbb{E}\left(\frac{\partial}{\partial\theta_1}w_1(\theta)t_1(X) + \frac{\partial}{\partial\theta_1}w_2(\theta)t_2(X)\right) = \mathbb{E}\left(\frac{\partial}{\partial\theta_1}\frac{\theta_1}{\theta_2}\cdot x - \frac{1}{2}\frac{\partial}{\partial\theta_1}\frac{1}{\theta_2}x^2\right)$$

$$= \frac{1}{\theta_2}\mathbb{E}(x)$$

$$\implies \frac{1}{\theta_2}\mathbb{E}(x) = \frac{\theta_1}{\theta_2} \implies \mathbb{E}(X) = \theta_1 = \mu.$$

Repeating by taking the partial derivatives with respect to θ_2 instead:

$$\frac{\partial}{\partial\theta_2}a(w(\theta)) = \frac{\partial}{\partial\theta_2}\left(\theta_1^2/(2\theta_2) + (1/2)\log(\theta_2)\right) = \frac{\theta_1^2}{2}\cdot\frac{-1}{\theta_2^2} + \frac{1}{2\theta_2} = \frac{\theta_2 - \theta_1^2}{2\theta_2^2}$$

Left side:

$$\mathbb{E}\left(\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_2}t_i(X)\right) = \mathbb{E}\left(\frac{\partial}{\partial\theta_2}w_1(\theta)t_1(X) + \frac{\partial}{\partial\theta_2}w_2(\theta)t_2(X)\right) = \mathbb{E}\left(\frac{\partial}{\partial\theta_2}\frac{\theta_1}{\theta_2}\cdot x - \frac{1}{2}\frac{\partial}{\partial\theta_2}\frac{1}{\theta_2}x^2\right)$$

$$= \mathbb{E}\left(-\frac{\theta_1}{\theta_2^2}\cdot x + \frac{1}{2\theta_2^2}x^2\right) = -\frac{\theta_1}{\theta_2^2}\mathbb{E}(X) + \frac{1}{2\theta_2^2}\mathbb{E}(X^2) = -\frac{\theta_1^2}{\theta_2^2} + \frac{1}{2\theta_2^2}\mathbb{E}(X^2)$$

$$\implies -\frac{\theta_1^2}{\theta_2^2} + \frac{1}{2\theta_2^2}\mathbb{E}(X^2) = \frac{\theta_2 - \theta_1^2}{2\theta_2^2} \iff -2\theta_1^2 + \mathbb{E}(X^2) = \theta_2 - \theta_1^2 \iff \mathbb{E}(X)^2 = \theta_2 + \theta_1^2 = \sigma^2 + \mu^2.$$

1.5 KL Divergence (DSO 607)

Let $f(\cdot \mid \theta) : \mathbb{R}^k \to \mathbb{R}^n$ be a family of parameterized probability densities with $\theta \in \Theta$. Suppose the true model is parameterized by $\theta_0 \in \mathbb{R}^s$, and we are interested in comparing a different model parameterized by $\theta \in \mathbb{R}^k$ to this model. The likelihood ratio $\tau : \mathbb{R}^n \to \mathbb{R}$ defined by $\tau(x; \theta, \theta_0) := f(x \mid \theta)/f(x \mid \theta_0)$ is a good tool for this comparison. Define the **discrimination** between θ and θ_0 at x to be $\Phi(\tau(x; \theta, \theta_0))$ for some function $\Phi : \mathbb{R} \to \mathbb{R}$, and define the **mean discrimination** $D(\theta; \theta_0) : \mathbb{R}^k \to \mathbb{R}$ between θ and θ_0 as

$$D(\theta; \theta_0) := \int_{\mathbb{R}^n} \Phi(\tau(x; \theta, \theta_0)) \cdot f(x \mid \theta_0)\, dx = \mathbb{E}_{\theta_0}\left[\Phi(\tau(X; \theta, \theta_0))\right].$$

To choose Φ so that $D(\theta; \theta_0)$ behaves like a distance, we would like $\Phi(1) = 0$ since this covers the case where $\theta = \theta_0$, because under the assumption that $f(x \mid \theta) = f(x \mid \theta_0)\ \forall x \in \mathbb{R}^n \iff \theta = \theta_0$, we have $\tau(x; \theta, \theta_0) = 1\ \forall x \in \mathbb{R}^n \iff \theta = \theta_0$. In particular, we would like this to be the minimum of the function; that is, $\tau''(1; \theta, \theta_0) > 0$. Lastly, we would also like $\tau'(x; \theta, \theta_0) > 0\ \forall x \in \mathbb{R}^n$. One such choice is $\Phi(t) := -2\log(t)$. This yields the **Kullback-Leibler (KL) divergence** $I(\theta; \theta_0) : \mathbb{R}^k \to \mathbb{R}$

$$I(\theta;\theta_0) = D(\theta_0,\theta) := -2\int_{\mathbb{R}^n} \log\left(\frac{f(x \mid \theta)}{f(x \mid \theta_0)}\right) \cdot f(x \mid \theta_0)\ dx$$
$$= 2\int_{\mathbb{R}^n} \left[\log\left(f(x \mid \theta_0)\right) - \log\left(f(x \mid \theta)\right)\right] \cdot f(x \mid \theta_0)\ dx = 2\mathbb{E}_{\theta_0}\left[\log\left(f(X \mid \theta_0)\right)\right] - 2\mathbb{E}_{\theta_0}\left[\log\left(f(X \mid \theta)\right)\right]$$
$$= 2\mathbb{E}_{\theta_0} \log\left(\frac{f_{\theta_0}(X)}{f_\theta(X)}\right).$$

Note that it is non-symmetric. It is also non-negative. It equals the expected likelihood ratio. Used in information theory; mutual information relation.

Definition 1.5.1 (**Mutual information; from GSBA 604**). Suppose we have two independent variables X and Y. If they are independent, we have $P_{(X,Y)} = P_X \cdot P_Y$. Then the **mutual information** between X and Y is $D(P_{(X,Y)}, P_X \cdot P_Y)$.

Also, the **entropy** of X with density f is equal to $\int f(x) \log f(x)\ dx = \mathbb{E}_X \log(f(x))$.

Of course, in practice we will estimate θ_0 as best as we can, by maximizing the **probabilistic negentropy**

$$\mathbb{E}_Z I(\theta;\hat{\theta}_0(Z)) := 2\mathbb{E}_{\theta_0}\left[\log\left(f(X \mid \theta_0)\right)\right] - 2\mathbb{E}_{\theta_0,Z}\left[\log\left(f\left(X \mid \hat{\theta}_0(Z)\right)\right)\right]$$

where Z describes the probability distribution of the sample data and $\hat{\theta}_0(Z)$ is our estimate of θ_0 from the data.

Way we wrote this in DSO 607: KL Divergence of density f (estimated) from density g (true model/distribution):

$$I(g;f) = \int [\log g(x)]g(z)dz - \int [\log f(x)]g(z)dz = \mathbb{E}_g \log(g(z)) - \mathbb{E}_g(\log(f(z))$$

GSBA 604: Notice that for exponential families,

$$\frac{1}{2}D(\theta_0,\theta) = (\theta_0-\theta)\mathbb{E}_{\theta_0}(X) - [\psi(\theta_0)-\psi(\theta)] = (\theta_0-\theta)\frac{\mathrm{d}}{\mathrm{d}\theta_0}\psi(\theta_0) - [\psi(\theta_0)-\psi(\theta)]$$
$$\implies \psi(\theta_0) + [\theta-\theta_0]\frac{\mathrm{d}}{\mathrm{d}\theta_0}\psi(\theta_0) + \frac{1}{2}D(\theta_0,\theta) = \psi(\theta).$$

where $\mathbb{E}_{\theta_0}(X) = \frac{\mathrm{d}}{\mathrm{d}\theta_0}\psi(\theta_0) = \mu(\theta_0)$. For a visual image of this formula, see Figure 1.1 (similar figure in p. 16 of exponential family notes).

One application of KL Divergence is AIC for model selection; see Section **??**. See also Section **??**.

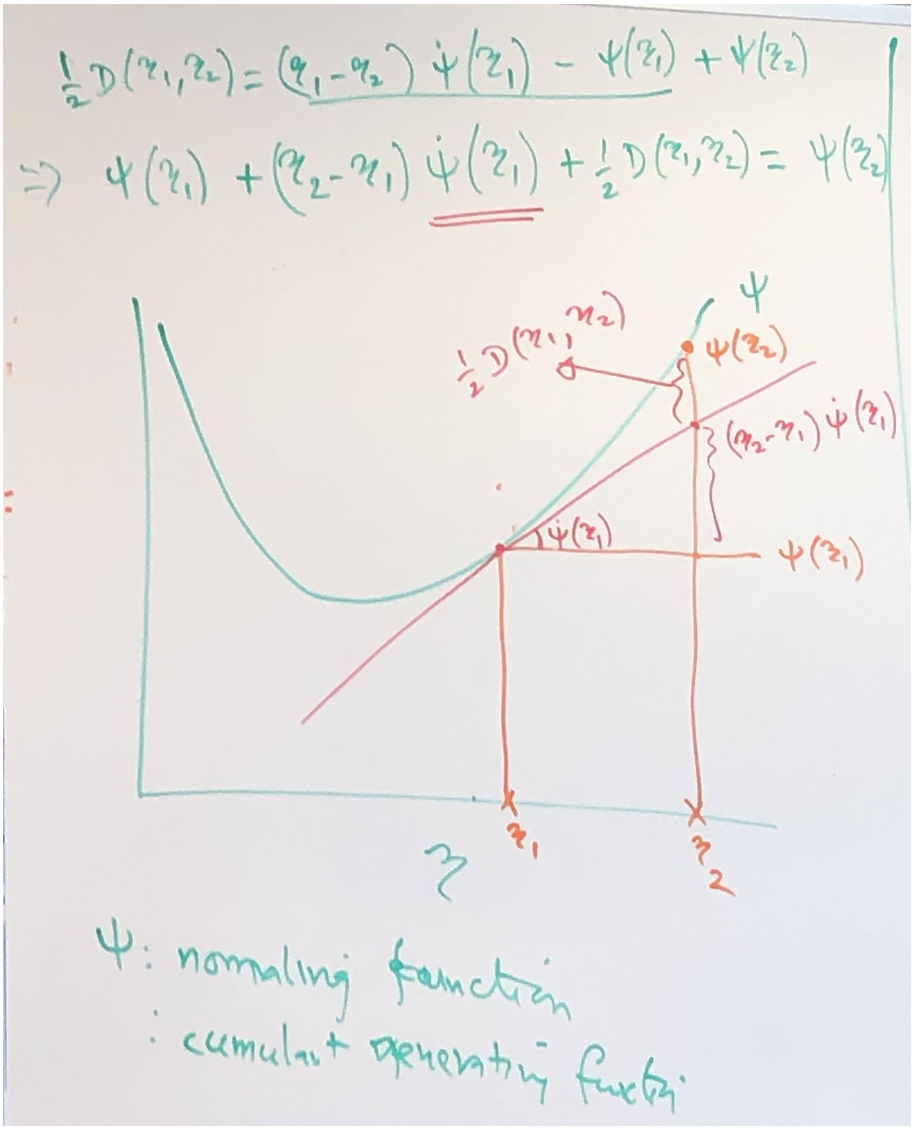


Figure 1.1: Photo of KL divergence (with different notation; $\theta_0 = \eta_1$, $\theta = \eta_2$).

1.6 Worked problems

1.6.1 Example Problems That Will Likely Appear on Midterm (and Final)

(1) Let X be uniform on $[1, 5]$, let Y be uniform on $[0, 1]$, and assume that X and Y are independent.

(a) Compute the probability density function of the product XY.

(b) **Only part included on midterm.** Compute the cumulative distribution function of the ratio X/Y.

(c) Compute the characteristic function of the sum $X + Y$.

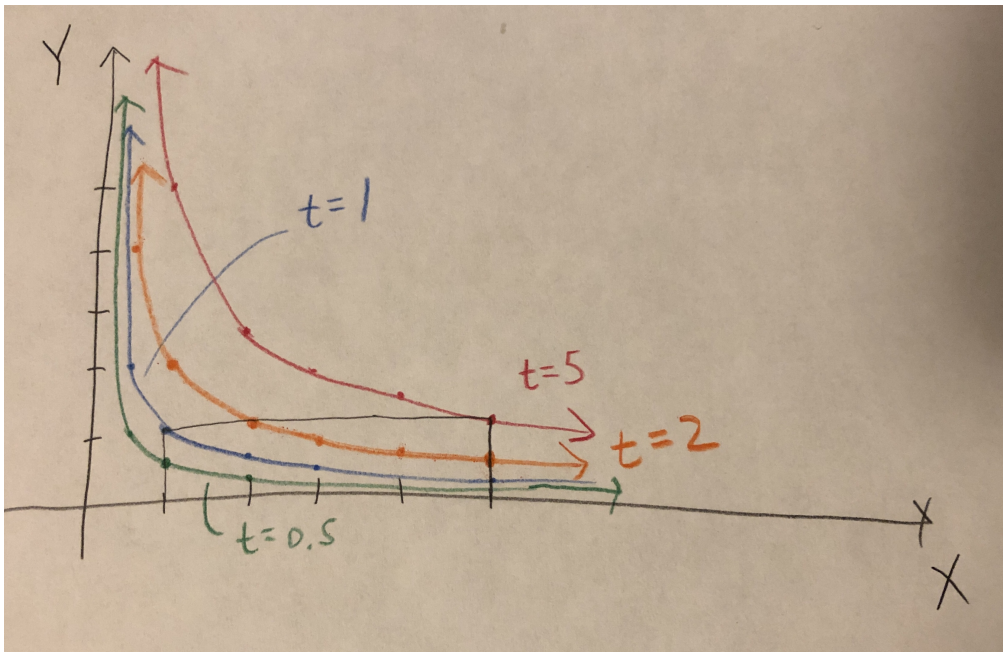
(d) Compute the moment-generating function of the random variable $X - \ln(Y)$.

Solution.

(a) We will find the cdf and then differentiate to yield the pdf. Observe that

$$F_{XY}(t) = \Pr(XY \leq t) = \Pr(Y \leq tX^{-1})$$

Plotting the density function of XY along with plots of $F_{XY}(t)$ as a function of X for various values of t, we have the following:



Now since both X and Y are distributed uniformly, for a given t, $\Pr(Y \leq tX^{-1})$ is the area under the curve and in the rectangle, weighted by $1/4$ since the rectangle has total area 4 but total probability 1. It is clear that the four regimes we need to consider are (1) $t < 0$, (2) $0 \leq t < 1$, (3) $1 \leq t < 5$, and (4) $t \geq 5$.

(1) $t < 0$: The curve lies below the rectangle, so there is no area below the curve and in the rectangle. Therefore $\boxed{\Pr(Y \leq tX^{-1} \mid t < 0) = 0}$. (This is also clear since tX^{-1} would be a negative number and Y is nonnegative.)

(2) $0 \leq t < 1$: Integrating the relevant area, we have

$$\Pr(Y \leq tX^{-1} \mid 0 \leq t < 1) = \frac{1}{4}\int_1^5 \frac{t}{x}dx = \frac{t}{4}\big[\log(x)\big]_1^5 = \boxed{\frac{t}{4}\log(5)}$$

(3) $1 \leq t < 5$: In this case, the area is a rectangle of height 1 and width $t-1$ plus the area under the curve from t to 5.

$$\Pr(Y \leq tX^{-1} \mid 1 \leq t < 5) = \frac{1}{4}\left(1\cdot(t-1)+\int_t^5 \frac{t}{x}dx\right) = \frac{1}{4}\left(t-1+t\big[\log(x)\big]_t^5\right) = \boxed{\frac{1}{4}\big[t\big(1+\log(5/t)\big)-1\big]}$$

(4) $t \geq 5$: In this case, the entire rectangle lies below the curve. Therefore $\boxed{\Pr(Y \leq tX^{-1} \mid t \geq 5) = 1}$.

So we have

$$F_{XY}(t) = \begin{cases} 0 & t < 0 \\ \frac{t}{4}\log(5) & 0 \leq t < 1 \\ \frac{1}{4}\big[t\big(1+\log(5/t)\big)-1\big] & 1 \leq t < 5 \\ 1 & t \geq 5 \end{cases}$$

Finally, differentiating yields

$$\boxed{f_{XY}(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{4}\log(5) & 0 \leq t < 1 \\ \frac{1}{4}\log\left(\frac{5}{t}\right) & 1 \leq t < 5 \\ 0 & t \geq 5 \end{cases}}$$

since

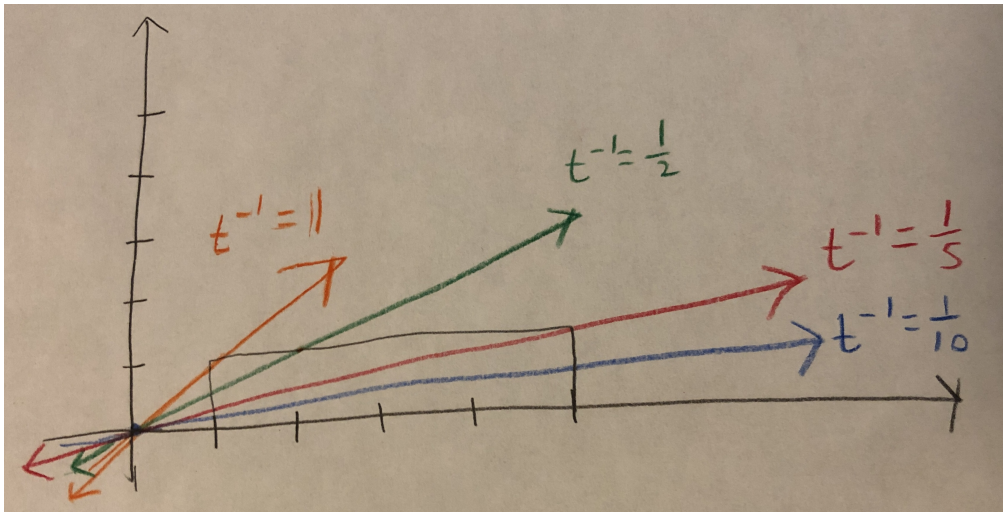
$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{4}\big[t\big(1+\log(5/t)\big)-1\big]\right) = \frac{1}{4}\left[1+\log(5/t)+t\left(\frac{1}{5/t}\cdot -5\cdot t^{-2}\right)\right] = \frac{1}{4}\left[1+\log(5/t)+t\left(t\cdot -1\cdot t^{-2}\right)\right]$$

$$= \frac{1}{4}\log\left(\frac{5}{t}\right)$$

(b) **Only part included on midterm.** We will proceed in a similar way as part (a). Observe that

$$F_{X/Y}(t) = \Pr\left(\frac{X}{Y} \leq t\right) = \Pr(Y \geq X/t)$$

Plotting the density function of X/Y along with plots of $F_{X/Y}(t)$ as a function of X for various values of t, we have the following:



Now since both X and Y are distributed uniformly, for a given t, $\Pr(Y \geq X/t)$ is the area above the curve and in the rectangle, weighted by $1/4$ since the rectangle has total area 4 but total probability 1. It is clear that the three regimes we need to consider are (1) $t^{-1} \geq 1 \iff t \leq 1$, (2) $1/5 \leq t^{-1} < 1 \iff 1 < t \leq 5$, and (3) $0 < t^{-1} < 1/5 \iff t > 5$.

(1) $t \leq 1$: The curve lies above the rectangle, so there is no area above the curve and in the rectangle. Therefore $\Pr(Y \geq X/t \mid t \leq 1) = 0$. (This is also clear since X/t would have to be greater than 1 and Y is less than or equal to 1.)

(2) $1 < t \leq 5$: The relevant area is the triangle above the green line in the rectangle. Note that it intersects the vertical line at $Y = 1/t$ and the horizontal line at $X = t$.

$$\Pr(Y \geq X/t \mid 1 < t \leq 5) = \frac{1}{4} \cdot \frac{1}{2}\left(1 - \frac{1}{t}\right)(t-1) = \frac{1}{8}\left(t - 1 - 1 + \frac{1}{t}\right) = \frac{1}{8}\left(t - 2 + \frac{1}{t}\right)$$

(3) $t > 5$: In this case, the area is a trapezoid above the blue line and in the rectangle. Note that the blue line intersects the left vertical line at $Y = 1/t$ and the right vertical line at $Y = 5/t$.

$$\Pr(Y \geq X/t \mid t > 5) = \frac{1}{4} \cdot \frac{1}{2} \cdot \left(1 - \frac{1}{t} + 1 - \frac{5}{t}\right) \cdot 4 = \frac{1}{2} \cdot \left(2 - \frac{6}{t}\right) = 1 - \frac{3}{t}$$

So we have

$$\boxed{F_{XY}(t) = \begin{cases} 0 & t \leq 1 \\ \frac{1}{8}\left(t - 2 + \frac{1}{t}\right) & 1 < t \leq 5 \\ 1 - \frac{3}{t} & t > 5 \end{cases}}$$

(c) The characteristic function for a uniform distribution on $[a, b]$ is

$$\frac{2}{(b-a)t} \sin\left(\frac{1}{2}(b-a)t\right) \exp\left(i(a+b)\frac{t}{2}\right).$$

Using the fact that $X \perp\!\!\!\perp Y \implies \phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$, we have

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) = \frac{2}{(5-1)t} \sin\left(\frac{1}{2}(5-1)t\right) \exp\left(i(5+1)\frac{t}{2}\right) \cdot \frac{2}{t} \sin\left(\frac{1}{2}t\right) \exp\left(i\frac{t}{2}\right)$$

$$= \frac{1}{2t}\sin(2t)\exp(3it)\cdot\frac{2}{t}\sin\left(\frac{1}{2}t\right)\exp\left(i\frac{t}{2}\right) = \boxed{\frac{1}{t^2}\exp\left(\frac{7}{2}it\right)\cdot\sin(2t)\sin\left(\frac{1}{2}t\right)}$$

(d) The moment-generating function for a uniform distribution on $[a,b]$ is

$$M_X(t) = \mathbb{E}(\exp(tX)) = \int_a^b \frac{1}{b-a}\cdot\exp(tx)dx = \frac{1}{b-a}\left[\frac{1}{t}\exp(tx)\right]_a^b = \frac{1}{(b-a)t}\big[\exp(bt)-\exp(at)\big]$$

Therefore the moment-generating function for X is $t^{-1}\big[\exp(t)-1\big]$. Note that

$$\Pr(Y\le y) = \Pr(-\log(X)\le y) = \Pr(\log(X)\ge -y) = \Pr(X\ge e^{-y}) = \int_{\exp(-y)}^{\infty} dt = \int_{\exp(-y)}^{1} dt$$

Substituting $t = e^{-u}$ (so that we have $u = -\log(t)$, $dt = -e^{-u}du$, we have

$$\Pr(Y\le y) = -\int_y^0 e^{-u}du = \big[e^{-u}\big]_y^0 = 1-e^{-y}$$

which is the cdf for an exponential distribution with mean 1. Therefore $Y = -\log(X)\sim$ Exponential(1), so

$$M_Y(t) = \frac{1}{1-t}$$

Using the fact that if X and Y are independent then $M_{X+Y}(t) = M_X(t)M_Y(t)$, we have

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \frac{\exp(t)-1}{t}\cdot\frac{1}{1-t} = \boxed{\frac{\exp(t)-1}{t-t^2}}$$

(2) **Fall 2016 Problem 2.** Let X and Y be i.i.d. exponential with mean 1. Show that for every $t>0$ the events $\{\omega : \min\{X,Y\} > t\}$ and $\{\omega : X < Y\}$ are independent.

Solution. Note that $\min\{X,Y\} > t \iff X > t \cap Y > t$.

- $\Pr(\min\{X,Y\}>t) = \Pr(X>t\cap Y>t) = \Pr(X>t)\Pr(Y>t)$

$$= \int_t^\infty e^{-x}dx\int_t^\infty e^{-y}dy = -e^{-x}\big|_t^\infty - e^{-y}\big|_t^\infty = \boxed{e^{-2t}}$$

- $\Pr(X<Y)$: Note that in Figure 2, the region that satisfies this condition is region G_1 plus G_2 (note that X and Y are nonnegative). Therefore we can find this probability by integrating the joint pdf over that region.

$$\Pr(X<Y) = \iint\limits_{\{G_1+G_2\}} f_{X,Y}(x,y)dxdy$$

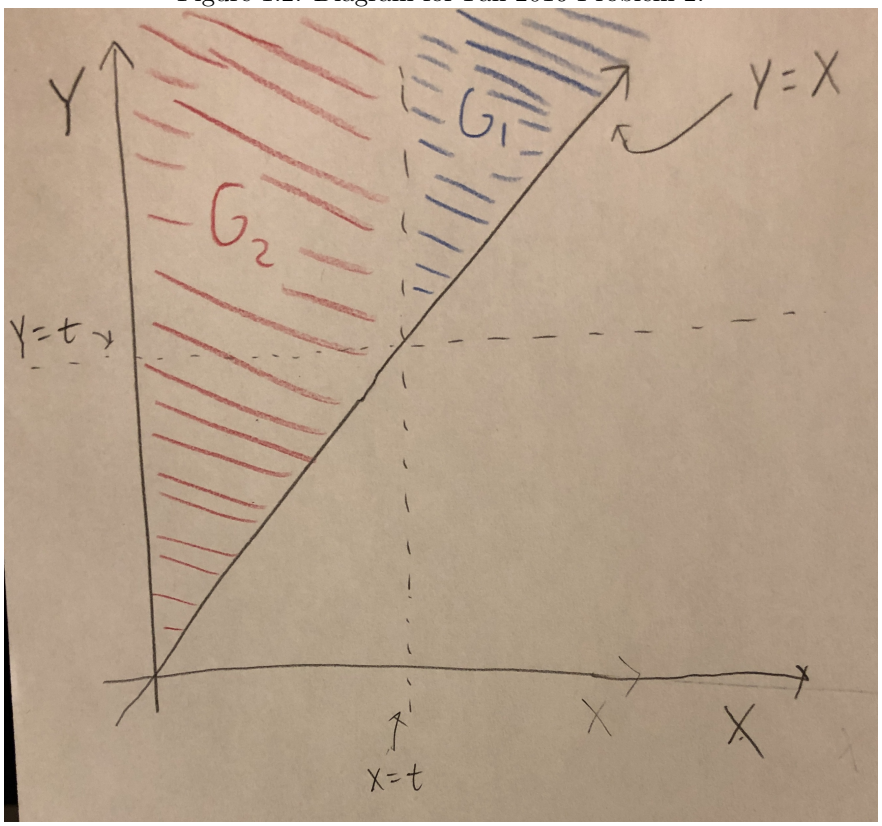
Note that the joint pdf is the probability of the marginal pdfs since X and Y are independent.

$$= \int_0^\infty\int_x^\infty e^{-x-y}dydx = \int_0^\infty e^{-x}\int_x^\infty e^{-y}dydx$$

$$\int_x^\infty e^{-y}dy = -e^{-y}\big|_x^\infty = e^{-x}$$

$$\implies \Pr(X<Y) = \int_0^\infty e^{-2x}dx = -\frac{1}{2}e^{-2x}\big|_0^\infty = \boxed{\frac{1}{2}}$$

Figure 1.2: Diagram for Fall 2016 Problem 2.



- $\Pr(X < Y \cap \min\{X,Y\} > t)$: Note that in Figure 2, the region that satisfies this condition is region G_1. Therefore we can find this probability by integrating the joint pdf over that region.

$$\Pr(X < Y \cap \min\{X,Y\} > t) = \int\int_{G_1} f_{X,Y}(x,y)dxdy = \int_t^\infty \int_t^y e^{-x-y}dxdy = \int_t^\infty e^{-y}\int_t^y e^{-x}dxdy$$

$$\int_t^y e^{-x}dx = -e^{-x}\big|_t^y = -e^{-y} + e^{-t}$$

$$\implies \Pr(X < Y \cap \min\{X,Y\} > t) = \int_t^\infty e^{-y}\big(-e^{-y} + e^{-t}\big)dy = \int_t^\infty \big(e^{-t-y} - e^{-2y}\big)dy$$

$$= \frac{1}{2}e^{-2y} - e^{-t}e^{-y}\big|_t^\infty = -\frac{1}{2}e^{-2t} - e^{-2t} = \boxed{\frac{1}{2}e^{-2t}}$$

Note that

$$\Pr(X < Y \cap \min\{X,Y\} > t) = \frac{1}{2}\cdot e^{-2t} = \Pr(X < Y)\Pr(\min\{X,Y\} > t)$$

Therefore the events $\{\omega : \min\{X,Y\} > t\}$ and $\{\omega : X < Y\}$ are independent for every $t > 0$.

(3) In a certain area, earthquakes happen at a frequency of one every four days. What is the probability that more than 100 earthquakes will occur in this area in one year (365 days)?

Solution. We can think of this as a Poisson process (see section **??**) with $\lambda = 1/4$. Then there are two ways to obtain the answer: we can either examine the number of earthquakes in a 365 day period $N(365)$ and find the probability that $N(365) > 100$, or we can examine the number of days until the 101st earthquake T_{101} and find the probability that $T_{101} < 365$.

(i) **Number of earthquakes in 365 days:** Let $N(t)$ be the number of earthquakes that occur in t days after the start of this process. By Theorem **??**, $N(t) \sim \text{Poisson}(t \cdot 1/4)$. Then

$$\Pr(N(t) > 100) = \sum_{j=101}^{\infty} \frac{(365\cdot 1/4)^j \exp(-365\cdot 1/4)}{j!}$$

To obtain an answer for this, we can use the normal approximation to a Poisson distribution (Proposition 1.1.7.4):

$$N(t) \sim \mathcal{N}(t/4, t/4) \implies \Pr(N(365) > 100) \approx \Pr\left(\mathcal{N}(0,1) > \frac{100.5 - 365/4}{\sqrt{365/4}}\right)$$

$$= \Pr\left(\mathcal{N}(0,1) > \frac{100.5 - 91.25}{\sqrt{91.25}}\right) \approx \Pr\left(\mathcal{N}(0,1) > \frac{9.25}{9.1}\right) \approx \boxed{0.1664}$$

(ii) **Number of days before 100th earthquake:** Let T_n be the number of days until the nth earthquake happens. By Corollary **??**, $T_n \sim \text{Gamma}(n, 4)$. Then

$$\Pr(T_{101} < 365) = \int_0^{365} \frac{1}{\Gamma(101,4)} x^{101-1}e^{-x/4}dx$$

To obtain an answer for this, we can use the normal approximation to a Gamma distribution (Proposition 1.3.5.6):

$$T_n \sim \mathcal{N}(404, 1616) \implies \Pr(T_{101} < 365) \approx \Pr\left(\mathcal{N}(0,1) < \frac{365-404}{\sqrt{1616}}\right) \approx \Pr\left(\mathcal{N}(0,1) < \frac{-40}{40}\right)$$

$$= \Pr\big(\mathcal{N}(0,1) < -1\big) \approx \boxed{0.1660}$$

1.6.2 More Problems From Homework

Homework 5 Problem 4.

Let $X_1, X_2, \ldots$ be i.i.d. having moment-generating functions $M_X = M_X(t), t \in (-\infty, \infty)$. Let N be an integer-valued random variable with moment-generating function $M_N = M_N(t), t \in (-\infty, \infty)$. Assume that N is independent of all X_k and define $S = \sum_{k=1}^{N} X_k$. Confirm that the random variable S has the moment-generating function $M_S = M_S(t)$ defined for all $t \in (-\infty, \infty)$ and

$$M_S(t) = M_N\big(M_X(t)\big)$$

Then use the result to derive the formulae

$$\mathbb{E}(S) = \mu_N \mu_X, \mathrm{Var}(S) = (\sigma_N^2 - \mu_N)\mu_X^2 + \mu_N \sigma_X^2$$

where $\mu_N = \mathbb{E}(N)$, $\mu_X = \mathbb{E}(X_1)$, $\sigma_N^2 = \mathrm{Var}(N)$, and $\sigma_X^2 = \mathrm{Var}(X_1)$. How will the above computations change if we use the characteristic function ϕ_X instead of the moment-generating function M_X?

Solution.

$$M_S(t) = \mathbb{E}(e^{tS}) = \mathbb{E}[\mathbb{E}(e^{tS} \mid N)] = \sum_{n=0}^{\infty} \mathbb{E}(e^{tS} \mid N = n) \Pr(N = n) = \sum_{n=0}^{\infty} \mathbb{E}(e^{t(X_1 + X_2 + \ldots + X_n)} \mid N = n) \Pr(N = n)$$

$$= \sum_{n=0}^{\infty} \mathbb{E}(e^{tX_1} e^{tX_2} \cdots e^{tX_n}) \Pr(N = n)$$

By independence of the X_i we have

$$= \sum_{n=0}^{\infty} \mathbb{E}(e^{tX_1}) \mathbb{E}(e^{tX_2}) \cdots \mathbb{E}(e^{tX_n}) \Pr(N = n)$$

which, since the X_i are i.i.d., can be written as

$$= \sum_{n=0}^{\infty} \mathbb{E}(e^{tX_1})^n \Pr(N = n) = \sum_{n=0}^{\infty} (M_X(t))^n \Pr(N = n)$$

But since $G_N(s) = \mathbb{E}(s^N) = \sum_{n=0}^{\infty} s^n \Pr(N = n)$, this can be written as

$$M_S(t) = G_N(M_X(t))$$

as desired. Note that

$$M_S'(t) = G_N'\big(M_X(t)\big)M_X'(t)$$

$$M_S''(t) = G_N''\big(M_X(t)\big)(M_X'(t))^2 + G_N'\big(M_X(t)\big)M_X''(t)$$

So we have

- $\mathbb{E}(S) = M_S'(0) = G_N'\big(M_X(0)\big)M_X'(0) = G_N'(1)\mathbb{E}(X_1) = \mathbb{E}(N)\mathbb{E}(X_1) = \mu_N\mu_X$
- $\text{Var}(S) = \mathbb{E}(S^2) - \mathbb{E}(S)^2 = M_S''(0) - (M_S'(0))^2$

$$= G_N''\big(M_X(0)\big)(M_X'(0))^2 + G_N'\big(M_X(0)\big)M_X''(0) - \mu_N^2\mu_X^2 = G_N''\big(1\big)\mathbb{E}(X_1)^2 + G_N'\big(1\big)\text{Var}(X_1) - \mu_N^2\mu_X^2$$

$$= \mathbb{E}[N(N-1)]\mathbb{E}(X_1)^2 + \mathbb{E}(N)\text{Var}(X_1) - \mu_N^2\mu_X^2 = \mathbb{E}[N^2 - N]\mathbb{E}(X_1)^2 + \mathbb{E}(N)\text{Var}(X_1) - \mu_N^2\mu_X^2$$

$$= [\mathbb{E}(N^2) - \mathbb{E}(N)^2 + \mathbb{E}(N)^2 - \mathbb{E}(N)]\mathbb{E}(X_1)^2 + \mathbb{E}(N)\text{Var}(X_1) - \mu_N^2\mu_X^2 =$$

$$= [\text{Var}(N) + \mathbb{E}(N)^2 - \mathbb{E}(N)]\mathbb{E}(X_1)^2 + \mathbb{E}(N)\text{Var}(X_1) - \mu_N^2\mu_X^2 = (\sigma_N^2 + \mu_N^2 - \mu_N)\mu_X^2 + \mu_N\sigma_X^2 - \mu_N^2\mu_X^2$$

$$= \boxed{(\sigma_N^2 - \mu_N)\mu_X^2 + \mu_N\sigma_X^2}$$

To use the characteristic function ϕ_X instead of the moment generating function M_X, we would do the following:

$$\phi_S(t) = \mathbb{E}(e^{itS}) = \mathbb{E}[\mathbb{E}(e^{itS} \mid N)] = \sum_{n=0}^{\infty} \mathbb{E}(e^{itS} \mid N = n)\Pr(N = n) = \sum_{n=0}^{\infty} \mathbb{E}(e^{it(X_1+X_2+\ldots+X_n)} \mid N = n)\Pr(N = n)$$

$$= \sum_{n=0}^{\infty} \mathbb{E}(e^{itX_1}e^{tX_2}\cdots e^{itX_n})\Pr(N = n)$$

By independence of the X_i we have

$$= \sum_{n=0}^{\infty} \mathbb{E}(e^{itX_1})\mathbb{E}(e^{itX_2}) \cdots \mathbb{E}(e^{itX_n}) \Pr(N = n)$$

which, since the X_i are i.i.d., can be written as

$$= \sum_{n=0}^{\infty} \mathbb{E}(e^{itX_1})^n \Pr(N = n) = \sum_{n=0}^{\infty} (\phi_X(t))^n \Pr(N = n)$$

But since $G_N(s) = \mathbb{E}(s^N) = \sum_{n=0}^{\infty} s^n \Pr(N = n)$, this can be written as

$$\phi_S(t) = G_N(\phi_X(t))$$

Homework 5 Problem 7.

(a) Let $X_1, X_2, \ldots, X_n$ be independent with mean zero and finite third moment. Prove that

$$\mathbb{E}(X_1 + \ldots + X_n)^3 = \mathbb{E}X_1^3 + \ldots + \mathbb{E}X_n^3$$

Solution.

(a) Let $\mathbb{E}(\exp(it_iX_i) = \phi_{X_i}(t_i)$. Let $S_n = \sum_{i=1}^n X_i$. Then by independence the characteristic function for S_n is

$$\mathbb{E}(\exp(itS_n)) = \phi_{S_n}(t) = \prod_{i=1}^{n} \phi_{X_i}(t)$$

Then

$$\mathbb{E}(X_1 + X_2 + \ldots + X_n)^3 = \mathbb{E}(S_n^3) = \phi_{S_n}^{(3)}(0)$$

$$= \sum_{i=1}^{n} \phi_{X_i}^{(3)}(0) \cdot \Big(\prod_{j \in \{1,\ldots,n\}, j \neq i} \phi_{X_j}(0) \Big) + C \Big[\sum_{i=1}^{n} \cdot \Big(\sum_{j \in \{1,\ldots,n\}, j \neq i} \phi_{X_i}^{(2)}(0) \phi_{X_j}^{(1)}(0) \Big) \cdot \Big(\prod_{k \in \{1,\ldots,n\}, k \neq i, j} \phi_{X_k}(0) \Big) \Big]$$

where C is some coefficient resulting from the multinomial expansion of S_n after repeated differentiation product rules. But because $\mathbb{E}(X_i) = 0$, $\phi_{X_i}^{(1)}(0) = 0 \ \forall i$, so the second term goes to 0. Therefore we have

$$\mathbb{E}(X_1 + X_2 + \ldots + X_n)^3 = \sum_{i=1}^{n} \phi_{X_i}^{(3)}(0) \cdot \Big(\prod_{j \in \{1,\ldots,n\}, j \neq i} \phi_{X_j}(0) \Big) = \sum_{i=1}^{n} \mathbb{E}(X_i^3) \cdot 1^{n-1} = \sum_{i=1}^{n} \mathbb{E}(X_i^3)$$

as desired.

Homework 6 Problem 10.

(a) For $p \in (0,1)$, let $x(p)$ be the smallest number of people so that there is a better than $100 \cdot p\%$ chance to have at least two born on the same day. Find an approximate expression for $x(p)$, and sketch the graph of the function $x = x(p)$.

(b) Repeat part (a) when you want at least three people to share a birthday.

Solution.

(a) Let $f(x)$ be the probability of no matches in birthdays in a group of x people; that is,

$$f(x) = \frac{365 \cdot 364 \cdot 363 \cdots (365 - x + 1)}{365^x} = \frac{1}{365^x} \cdot \frac{365!}{(365-x)!} = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{x-1}{365}\right)$$

Using the first order Taylor approximation $\exp(-k/x) \approx 1 - k/x$, we have

$$f(x) = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{x-1}{365}\right) \approx \exp(-1/365)\exp(-2/365) \cdots \exp(-(x-1)/365)$$

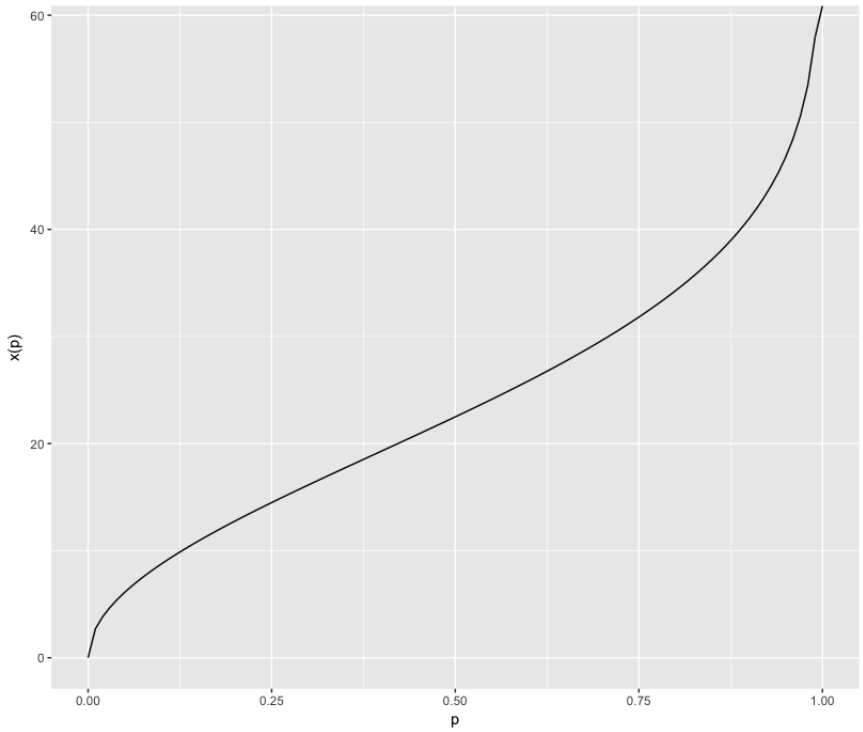
$$= e^{-(x^2-x)/(2 \cdot 365)}$$

We want the probability of a match to be at least p; that is, $f(x) \leq 1 - p$. Setting this equal to $q = 1 - p$, we have

$$e^{-(x^2-x)/(2 \cdot 365)} = q \iff -\frac{x^2 - x}{730} = \log(q) \iff x^2 - x + 730\log(q) = 0$$

$$\implies x = 0.5 + \sqrt{1/4 + 730\log(1/q)} \approx \boxed{\sqrt{2 \cdot 365 \log(1/q)}}$$

where we discard the negative root because we have to have a nonnegative number of people, and we don't worry about the decimals since this is an approximation and we have to round up to the nearest whole person anyway.



(b) For a group of three people, the Poisson approximation (see Section 1.1.10) is more convenient. The number of groups of 3 people in a room of x people is $\binom{x}{3}$. For a group of three people, the probability that all three have the same birthday is $1 \cdot 1/365 \cdot 1/365 = 365^{-2}$. Therefore we can think of the number of matches of three people as distributed Poisson with expectation $\binom{x}{3} \cdot 365^{-2}$. Then we have the probability of at least one "success" (triplet with three matched birthdays) is

$$1 - \frac{\exp(-\lambda)\lambda^0}{0!} = 1 - \exp\left(-\binom{x}{3} \cdot 365^{-2}\right)$$

We set this equal to p and solve:

$$p = 1 - \exp\left(-\binom{x}{3} \cdot 365^{-2}\right) \iff -\binom{x}{3} \cdot 365^{-2} = \log(1-p) \iff \frac{x!}{(x-3)!3!} = 365^2 \cdot \log\left(\frac{1}{1-p}\right)$$

$$\iff x(x-1)(x-2) = 6 \cdot 365^2 \cdot \log\left(\frac{1}{1-p}\right) \iff (x^2-x)(x-2) = x^3 - 3x^2 + 2x = 6 \cdot 365^2 \cdot \log\left(\frac{1}{1-p}\right)$$

This has a unique real solution, but it is hard to find.

Exercise 5. Let X, Y, Z be independent uniform on (0, 1). Compute the cdfs of XY, X/Y, and XY/Z.

Solution

Using the information from part (a), and the fact that $f_X(x) = 1$ (for $x \in [0, 1]$) and likewise for $f_Y(y)$:

- XY:

$$F_{XY}(z) = \int_0^\infty f_X(x) \int_{-\infty}^{z/x} f_Y(y) dy dx - \int_{-\infty}^0 f_X(x) \int_\infty^{z/x} f_Y(y) dy dx$$

$$= \int_0^1 \big[(z/x)\mathbf{1}_{\{0<z/x\leq 1\}} + \mathbf{1}_{\{z/x>1\}}\big] dx = \int_0^1 \big[(z/x)\mathbf{1}_{\{z\leq x\}} + \mathbf{1}_{\{z>x\}}\big] dx = \int_0^z dx + \int_z^1 (z/x) dx$$

$$= z + z\log(x)\Big|_z^1 = z + z\log(1) - z\log(z) = z(1-\log(z))$$

$$\implies \boxed{F_{XY}(z) = \begin{cases} 0 & z \leq 0 \\ z(1-\log(z)) & 0 < z \leq 1 \\ 1 & z > 1 \end{cases}}$$

- X/Y:

$$F_{X/Y}(z) = \int_0^\infty f_Y(y) \int_{-\infty}^{zy} f_X(x) dx dy - \int_{-\infty}^0 f_Y(y) \int_\infty^{zy} f_X(x) dx dy$$

$$= \int_0^1 \big[zy\mathbf{1}_{\{0<zy\leq 1\}} + \mathbf{1}_{\{zy>1\}}\big] dy = \int_0^1 \big[zy\mathbf{1}_{\{y>0 \cap y\leq 1/z\}} + \mathbf{1}_{\{y>1/z\}}\big] dy = \int_0^{1/z} zy \cdot dy + \int_{1/z}^1 dy$$

$$= \frac{zy^2}{2}\Big|_0^{1/z} + (1 - 1/z) = \frac{z}{2z^2} + 1 - \frac{2}{2z} = 1 - \frac{1}{2z}$$

$$\implies F_{XY}(z) = \begin{cases} 0 & z \leq 0 \\ 1 - \frac{1}{2z} & 0 < 1/z \leq 1 \end{cases}$$

$$= \boxed{\begin{cases} 0 & z \leq 0 \\ z/2 & 0 < z \leq 1 \\ 1 - \frac{1}{2z} & z > 1 \end{cases}}$$

- XY/Z: Consider this the cdf of the quotient of $W = XY$ and Z.

$$F_U(u) = \int_0^\infty f_Z(z) \int_{-\infty}^{uz} f_W(w) dw dz - \int_{-\infty}^0 f_Z(z) \int_\infty^{uz} f_W(w) dw dz$$

$$= \int_0^1 \int_0^{uz} -\log(w)\mathbf{1}_{\{0<uz\leq 1\}} dw dz = \int_0^1 -\big[w\log(w) - w\big]_0^{uz} \mathbf{1}_{\{0<z\leq 1/u\}} dz$$

$$= \int_0^{1/u} uz\big[1 - \log(uz)\big] dz = \frac{u}{4} z^2 \big(3 - 2\log(uz)\big)\Big|_0^{1/u} = \frac{u}{4u^2}\big(3 - 2\log(1)\big) - 0 = \frac{3}{4u}$$

$$\implies \boxed{F_{XY/Z}(u) = \begin{cases} 0 & u \leq 0 \\ \frac{3}{4u} & 0 < u \leq 3/4 \\ 1 & u > 3/4 \end{cases}}$$

1.7 Random Matrix Theory

Definition 1.7.1 (**Wigner-type random matrix**). $H \in \mathbb{R}^{N\times N}$ is a **Wigner-type random matrix** if each entry H_{ij} equals either 1 or -1 with equal probability and H is symmetric ($H = H^+$. Other than the symmetry restriction, the entries are independent.

Remark. Because a Wigner-type random matrix is symmetric, the eigenvalues are real.

Example 1.7.1. 10 i.i.d. random variables $X_1, \dots, X_{10}$, where $\Pr(X_1 = m) = 1/3$ for $m \in \{2, 3, 4\}$. What is the distribution of $S = \sum_k X_k$?

This is not an interesting problem since a computer can easily calculate the answer, even if the number of random variables is in the thousands.

Example 1.7.2. n i.i.d. random variables $X_1, \dots, X_n$, where n is large. Find $f(n)$ and $g(n)$ such that

$$Z_n = \frac{S_n - f(n)}{g(n)}$$

has a nontrivial limiting distribution.

Solution

We know if $\mathbb{E}X_k^2 < \infty$, choosing $f(n) = n\mathbb{E}X_1$ and $g(n) = \sqrt{n\text{Var}(X_1)}$ allows convergence to a standard Gaussian random variable under the Central Limit Theorem. Also note that if n is large, $S_n \approx n\mathbb{E}X_1$ and $(S_n - n\mathbb{E}X_1) = \mathcal{O}\left(\sqrt{n\text{Var}(X_1)}\right)$.

Let $\hat{S}_n$ be the largest eigenvalue of a Wigner random matrix H_n. It turns out that $\hat{S}_n \approx 2\sqrt{n}$ and $\hat{S} - 2\sqrt{n} = \mathcal{O}(n^{1/6})$. Lastly,

$$\frac{\hat{S} - 2\sqrt{n}}{n^{1/6}}$$

converges in distribution to a Tracy-Widom distribution [Johnstone, 2001].

Exercise 6. Suppose $X_1, \dots, X_n$ are i.i.d. uniform on $[0, 1]$. Let $Y_1, \dots, Y_n$ be the order statistics $X_{(1)}, \dots, X_{(n)}$. Let $Z_n = Y_{n/2+1} - Y_{n/2}$. Find $f(n)$, $g(n)$ such that

$$\frac{Z_n - f(n)}{g(n)} \xrightarrow{d} \text{non-trivial distribution.}$$

Example 1.7.3. Let $X_n = \sum_{k=1}^n X_k$ where $X_k \sim \text{Ber}(1/2)$. We know that

$$\frac{S_n - n/2}{1/2\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

However, what is $\mathbb{P}(S_n > 0.6n)$? It turns out that $\mathcal{P}(S_n > 0.6n) \sim e^{-cn}$.

For a Gaussian Orthogonal Ensemble, eigenvalues tend to be between -2 and 2. Difference between adjacent eigenvalues $\lambda_i - \lambda_{i+1}$ is of interest. Tends to be of order $1/n$, so after mutliplying by n, tends to be of order 1.

$$n(\lambda_i - \lambda_{i+1}) \xrightarrow{d} \text{Gaudin-Mehta distribution.}$$

Approximately given by Wigner's

$$\mathbb{P}(s) = \frac{\pi s}{2} e^{-\frac{\pi}{4}s^2}.$$

Bulk Universality Conjecture: Gap distribution for the bulk eigenvalues for a large generic d-regular graph converges to this distribution.

Universality Conjecture (Wigner-Dyson-Mehta): bulk eigenvalue statistics are universal, depending only on the symmetry class of the matrix ensemble, regardless of the distribution of individual matrix entries. Proven by Erdos, Yau, and Tau and Vu (under 4th moment condition)

1.7.1 Large Deviation Theory

Wigner-type matrix H. Let the eigenvalues be $\lambda_1, \dots, \lambda_n \in \mathbb{R}$. Suppose we normalize H by dividing it by the square root of n; that is, let $H_1 = H/\sqrt{n}$. Define a measure

$$\mu_n := \frac{1}{n} \sum_{k=1}^{n} \delta_{\lambda_k}$$

so for $I \subset \mathbb{R}$,

$$\mu_n(I) = \#\{i : \lambda_i \in I\}.$$

Note that μ_n is random because H is random. We will show that

$$\mu_n \xrightarrow{a.s.} \mu.$$

Proposition 1.7.1.1. For any continuous smooth function g,

$$\int g(t) d\mu_n(t) \xrightarrow{a.s.} \int g(t) d\mu(t).$$

We can also write this as

$$\langle \mu_n, g \rangle \xrightarrow{a.s.} \langle \mu, g \rangle.$$

In other words, for any interval $[a, b] \subset \mathbb{R}$,

$$\mu_n([a,b]) \xrightarrow{a.s.} \mu([a,b]).$$

where

$$d\mu = \frac{1}{2\pi}\sqrt{(4-x^2)_+}\,dx$$

where

$$(4-x^2)_+ = \begin{cases} 4-x^2 & 4-x^2 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Roughly speaking, for a Wigner random matrix where $\mathbb{E}H_{ij} = 0$ and $\mathbb{E}H_{ij}^2 = 1$ and $\mathbb{E}H_{ij}^p$ exists and is finite for all p, if n is large then the number of eigenvalues of H_n in the interval $[a,b]$ divided by n is close to the number

$$\int_a^b \frac{1}{2\pi}\sqrt{(4-x^2)_+}\,dx.$$

Let $\lambda_n^{(n)}$ be the largest eigenvalue of H_n. We have that

$$\frac{\lambda_n^{(n)} - 2}{n^{2/3}} \xrightarrow{d} \text{Tracy-Widom distribution.}$$

Proposition 1.7.1.2. Suppose $k = \lfloor n/2 \rfloor$. Let $\lambda_k^{(n)}$ be the k-th smallest eigenvalue of H_n. γ_k is defined as

$$\int_{-2}^{\gamma_k} \frac{1}{2\pi}\sqrt{(4-x^2)_+} = \frac{k}{n}$$

Then

(1) $\lambda_k^{(n)} \approx \gamma_k$

(2) $\lambda_k^{(n)} - \gamma_k = \mathcal{O}\left(\frac{1}{2}\sqrt{\log n}\right)$.

(3)

$$\frac{\lambda_k^{(n)} - \gamma_k}{n^{-1}\sqrt{\log n}} \xrightarrow{d} \mathcal{N}(0,1).$$

These results hold regardless of the exact distribution of H_{ij}. This property is referred to as **universality.** (Like the Central Limit Theorem.)

Proposition 1.7.1.3. Fix $a, b \in \mathbb{R}$, $E_0 \in \mathbb{R}$..

$$I_{E_0,a,b} = [E_0 + a/n, E_0 + b/n].$$

Then the number of eigenvalues of H in I $J_{E_0,a,b}$ is $\mathcal{O}(1)$. (Regardless of the exact distribution of H_{ij}.)

Proposition 1.7.1.4. Let $\lambda_k^{(n)}$ be an eigenvalue of H_n. Let $u_k^{(n)}$ be the corresponding unit-length eigenvector; that is,

$$Hu_k^{(n)} = \lambda_k^{(n)} u_k^{(n)}, \qquad \|u_k^{(n)}\|_2 = 1.$$

Then

1. For all fixed $w \in \mathbb{R}^n$ such that $\|wrVert_2 = 1$,

$$\frac{|\langle w, u_k^{(n)}\rangle|^2}{1/n} \xrightarrow{d} (\mathcal{N}(0,1))^2 .$$

2. Let $g_k^w := |\langle w, u_k^{(n)}\rangle|^2$. Then for $k \neq \ell$,

$$(g_k^w, g_\ell^w) \xrightarrow{d} \left([\mathcal{N}(0,1)]^2, [\mathcal{N}(0,1)]^2\right)$$

   and the two vectors are independent.

3. Let w, w' be two orthogonal unit vectors ($\langle w, w'\rangle = 0$). Then

$$(g_k^w, g_k^{w'}) \xrightarrow{d} \left([\mathcal{N}(0,1)]^2, [\mathcal{N}(0,1)]^2\right)$$

   and the two vectors are independent.

Let X have i.i.d. entries but it is not necessarily symmetric. We will investigate the eigenvalues of $X^T X$, which converge to a Marchenko–Pastur distribution.

Let X have eigenvalues $\lambda_k \in \mathbb{C}$. Let $\mu_n = n^{-1} \sum_{k=1}^n \delta_{\lambda_k}$. Then $n\mu_n(A)$ is the number of eigenvalues of X in A for $A \subset \mathbb{C}$.

$$\mu(A) = \frac{\text{Area}(A \cap D)}{\pi}, \qquad D := \{z : |z| \leq 1\}.$$

This is called a **circular law** (density $1/\pi$ in D, 0 outside of D). Results exist if $\mathbb{E}X_i = 0$, $\mathbb{E}X_{ij}^2 = 1$.

Proposition 1.7.1.5. $z \in \mathbb{C}$, $B_n = z^\ell$, $\ell = N^{-1/2+\delta}$, then the number of eigenvalues of X in B_n is roughly $\pi^{-1}\ell^2 n$. where B_n is a box with width ℓ.

Proposition 1.7.1.6. Suppose u_k is an eigenvector of X with $\|u_k\|_2 = 1$. Let $u_k(m)$ be the mth component of X. Note $\|u_k\|_2^2 = 1$, u_k has n components. With high probability,

$$\max_m |u_k(m)|^2 \leq n^{-1+\epsilon} \qquad \forall \epsilon > 0.$$

Delocalization: eigenvectors are spread out fairly uniformly over a range of possible values.

Opposite case: Suppose we have a diagonal random matrix $\tilde{X}_{ij}$. Then the eigenvectors are the unit vectors e_i, so the eigenvectors are very localized.

Some other Wigner-type random matrices:

(1) H, $\mathbb{E}H_{ij} = 0$, $\mathbb{E}H_{ij}^2 = \mathcal{O}(1)$, $\sum \mathbb{E}(H_{ij})^2 = N$. This is a generalized Wigner random matrix.

(2) Remove the condition that $\sum \mathbb{E}(H_{ij})^2 = N$. This is a **universal Wigner matrix.** In this case, the eigenvectors may have some favored direction.

(3) $\mathbb{E}H_{ij}^2 = 1$ but mean is not 0. Then similar results.

(4) Suppose X is Wigner and T is a fixed matrix. Then there are results for TX.

(5) H_{ij} does not have order 1 all the time; that is, some entries are on different scales than others (some much larger). Examples:

(a) Sparse matrix, like Erdos-Renyi matrix. E-R graph: line connecting V_i, V_j if and only iff $A_{ij} = 1$. $\mathbb{P}(A_{ij} = 1) = p(n)$, some function of n. All A_{ij} are independent of each other besides the fact that A is symmetric.
$A = H + A_0$, where A_{ij} is defined as above, $\mathbb{E}H_{ij} = 0$, $\mathbb{E}H_{ij}^2 = p_n(1 - p_n)$. Note that if $p_n \ll 1$,

$$\vdots$$

we have

$$\text{Var}\left(\frac{1}{\sqrt{p_n}} A_{ij}\right) \approx 1, \qquad \mathbb{E}\left(\frac{1}{\sqrt{p_n}} A_j\right) \approx 0,$$

but $\hat{A} = p_n^{-1/2} A$, $\hat{A}_{ij} \gg 1$; that is, this is a sparse matrix with mostly 0 entries, and the entries that are nonzero are large.

(b) **Band matrix.** $H_{ij} \cdot f(|i - j|)$, where

$$f(m) = \begin{cases} 1 & m \leq W \\ 0 & m > W. \end{cases}$$

The interesting case is when $W \ll n$.

Topics:

- **Topic 1:** For a matrix A, find optimal m and M such that for all x, $m\|x\| < \|Ax\| < M\|x\|$ (i.e., m is the smallest singular value and M is the largest). Then the condition number is $\kappa(A) = M/m$.

  Suppose $A \in \mathbb{R}^{n \times m}$, $a_{ij} = g_{ij}$, $g_{ij} \sim \mathcal{N}(0, 1)$. Marchenko–Pastur Law:

  $$\|A\| \approx \sqrt{n} + \sqrt{m}$$

  $$\sigma_{min}(A) \approx \sqrt{n} - \sqrt{m}$$

  If A is a tall matrix, $n \gg m$, then $\kappa(A)$ is constant. In square case, $\sigma_{min}(A) < 0 \iff A$ is invertible.

- **Least Singular Value of square matrices:** Intuition: eigenvector should not have a preference—"uniform on the sphere." True for Gaussian matricies.

  Typical $X \sim \text{Uniform}(S^{n-1})$:

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

$$\mathbb{E}|X_i|^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|X_j|^2 = \frac{1}{n}$$

where we used the fact that X is a unit vector. Therefore $|X_i| \sim n^{-1/2}$.

- $\|X\|_\infty = \max_j |X_j|$:
  $\mathbb{P}(|X_i| > t/\sqrt{n}) \leq \exp(-C_1 t^2)$ for $t > C_1$.

$$\mathbb{P}\left(|X_i| > \frac{R\sqrt{\log n}}{\sqrt{n}}\right) \leq \exp(-CR^2 \log n)$$

$$\mathbb{P}\left(\exists j \text{ such that } |X_j| > \frac{R\sqrt{\log n}}{\sqrt{n}}\right) < n\exp(-CR^2 \log n)$$

$$\|X\|_\infty \leq R\sqrt{\frac{\log n}{n}} \text{ with probability } \geq 1 - n^{1-CR^2}.$$

- **Topic 2:** ℓ_∞ delocaliztion. Suppose we have A with $\mathbb{E}a_{ij} = 0$, $\mathbb{E}A_{ij}^2 = 1$. Eigenector v, $\|v\|_2 = 1 \implies \|v\|_\infty < \frac{\log C}{\sqrt{n}}$.
- **Topic 3:** No-gap delocalization of eigenvectors.

$$X = \begin{pmatrix} \sqrt{2/n} \\ \vdots \\ \sqrt{2/n} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

  (first $n/2$ entries are $\sqrt{2/n}$, rest are 0.)

  For every $I \subset \{1, \ldots, n\}$ with $|I| = \epsilon n$, let I index over the nonzero entries of an eigenvector. Want to show that $\|V_i\|_2 \geq \epsilon^c$.

- **Topic 0:** Subgaussian random matrices

Linear algebra: $A \in \mathbb{R}^{n \times m}$. Singular value decomposition:

$$A = U\Sigma V^T$$

where $UU^T = I_n$, $VV^T = I_m$, and Σ is a diagonal matrix with diagonal entries equal to the singular values.

1.7.2 Band Matrix

$n \times n$ random matrix with bandwidth w. $H_{ij} \neq 0 \implies |i-j| \leq w$. Let $w = n^a$ for $0 < a < 1$. We have for the nonzero entries, $H_{ij} = \pm 1$, $\mathbb{E}H_{ij} = 0, \mathbb{E}H_{ij}^2 = 1$.

Let H be a Wigner band matrix. Since H is symmetric, decompose H as $H = UDU^+$, where D is diagonal. Then $H^m = UD^mU^+$. When H is a band matrix with reasonably small w, then for any eigenvector u of H, usually $|u(k)|^2$ is contained within a narrow range.

Roughly speaking, for each eigenvector u_k of H there exists an integer a_k, where $1 \leq a_k \leq n$, such that $u(k) \approx 0$ if $|k \cdot a| \gg w^2$.

From linear algebra: if λ_ℓ and u_ℓ are corresponding eigenvalues and eigenvectors of H,

$$\left[H^m\right]_{bc} = \sum_\ell \lambda_\ell^m u_\ell(b) u_\ell^+(c).$$

As you raise the matrix to higher and higher powers, the bandwidth will initially increase but eventually won't.

$$u_\ell(b) \neq 0, u_\ell(c) \neq 0 \implies |b-c| = o(w^2).$$

This is so-called **localization** of a band matrix. **Conjecture:** this occurs for $w = n^a$ when $a < 1/2$. When $a > 1/2$ we have delocalization.

For $d = 3$, a band matrix is delocalized as long as $a > 0$.

1. **GOE/GUE:** Wigner-type random matrix (e.g., $\mathbb{E}H_{ij} = 0, \mathbb{E}H_{ij}^2 = 1$). If $H_{ij} \sim \mathcal{N}(0,1)$, then it's called GOE (Gaussian orthogonal *). If H is GOE, then for any orthogonal matrix O ($OO^+ = I$), then OHO^+ has the same distribution as H. Also, the probability density of $\lambda_1, \ldots, \lambda_n$ of H is

   $$\prod e^{-\sum_k 1/4\lambda_k^2} \prod_{i<j} |\lambda_i - \lambda_j|.$$

   this is an **orthogonal polynomial.**

   Recall important properties of Wigner matrix: independent entries, orthogonal invariance. For a symmetric matrix, the only solution is a GOE matrix. For Hermitian, must be GUE.

2. **Comparison Method:** unknown X, known $\tilde{X}$, for Wigner, $\mathbb{E}\lambda_n^{(n)} \approx \mathbb{E}_n^{(n)}$.

   Look at Central Limit Theorem again: X_i i.i.d., $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = 1$, then

   $$\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \xrightarrow{d} \mathcal{N}(0,1).$$

   Special case: if $X_i \sim \mathcal{N}(0,1)$, result holds exactly even in finite samples.

   Need to prove:

   $$\mu_n \to \mathcal{N}(0,1) = \mu$$

Need to prove: for all $f \in C_{b,c}^{\infty}$ (smooth functions), $\langle f, \mu_n \rangle \to \langle f, \mu \rangle$. For all f,

$$\mathbb{E}f\left(\frac{\sum_i X_i}{\sqrt{n}}\right) \to \mathbb{E}f(Z),$$

where $Z \sim \mathcal{N}(0,1)$. So need to prove

$$\mathbb{E}f\left(\frac{\sum_i X_i}{\sqrt{n}}\right) - \mathbb{E}f\left(\frac{\sum_i X_i^G}{\sqrt{n}}\right) = o(1) \tag{1.18}$$

where X_i^G is Gaussian. (show the thing we want to study and something we know about (Gaussian R.V.s) are close together)

Let

$$S_n = \sum_{i=1}^n X_i, \qquad S_n^G = \sum_{i=1}^n X_i^G, \qquad \tilde{S}_n^k = \sum_{i=1}^k X_i + \sum_{i=k+1}^n X_i^G.$$

Note that $\tilde{S}_n^n = S_n$ and $\tilde{S}_n^0 = S_n^G$. We can express the left side of (1.18) as

$$\sum_k \left[\mathbb{E}f\left(\frac{\tilde{S}^{k+1}}{\sqrt{n}}\right) - \mathbb{E}f\left(\frac{\tilde{S}^k}{\sqrt{n}}\right)\right]$$

1.7.3 Subgaussian Random Variable

$g \sim \mathcal{N}(0,1)$, $t \to \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$. Fact

1. $$\mathbb{P}(|g| > t) \le 2\exp(-Ct^2)$$

2. $$G = \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix}$$

   $g_i \sim \mathcal{N}(0,1)$, with density function $x \to (2\pi)^{-n/2}\exp\left(-\sum_i X_i^2/2\right)$.

   So the density only depends on $\|X\|_2$, density is invariant under rotation, which means if U is an orthogonal matrix, then $UG \sim G$.

Definition 1.7.2 (**Subgaussian random variable**)**.** A random variable X is called **subgaussian** if

$$\|X\|_\Psi = \inf\left\{s > 0 : \mathbb{E}F\left(\frac{|X|}{s}\right) \le 1\right\} < \infty$$

("find s such that $\mathbb{E}F(|X|/s) = 1$.")

Remark. $\|\cdot\|_\Psi$ is a norm (the **subgaussian norm**); one can check that it satisfies the triangle inequality, homogeneity, and it is nonnegative.

Proposition 1.7.3.1. The following are equivalent:

1. There exists a κ such that $\|X\|_\Psi \leq \kappa$.
2. There exists a κ such that for all $t > 0$, $\mathbb{P}(|X| > t) \leq 2\exp(-t^2/\kappa^2)$.
3. There exists a κ such that $\sup_{p>1}\left\{\frac{\mathbb{E}(|X|^p)^{1/p}}{\sqrt{p}}\right\} \leq \kappa$.
4. There exists a κ such that $\mathbb{E}\exp(\lambda X) \leq 2\exp(\lambda^2\kappa^2)$.
5. (only when $\mathbb{E}X = 0$) $\mathbb{E}\exp(\lambda X) \leq \exp(C\lambda^2\kappa^2)$.

Proof. First we show that (1) implies (2). We will use exponential Markov's inequality: if Z is nonnegative,

$$\mathbb{P}(Z < t) \leq \frac{\mathbb{E}Z}{t}$$

$$\mathbb{P}(|X| > t) = \mathbb{P}\left(\exp(|X|^2/\kappa^2) > \exp(t^2/\kappa^2)\right) \leq \frac{\mathbb{E}\exp(|X|^2/\kappa^2)}{\exp(t^2/\kappa^2)} = \frac{\mathbb{E}[f(|X|/\kappa)+1]}{\exp(t^2/\kappa^2)}$$

$$\leq \frac{2}{\exp(t^2/\kappa^2)} = 2\exp(-t^2/\kappa^2).$$

which implies g is subgaussian. Now we show that 2 implies 3.

$$\mathbb{E}|X|^p = \mathbb{E}\left(\int_0^{|X|^p} 1\ dt\right) = \int_0^\infty \mathbf{1}_{[t,\infty)}(|X|^p)dt = \int_0^\infty \mathbb{P}(|X|^p < Mt)\ dt$$

Let $u^p = t$. We will use integration by parts

$$= \int_0^\infty pu^{p-1}\mathbb{P}(|X| > u)\ du \leq \int_0^\infty pu^{p-1}2\exp(-u^2/\kappa^2)\ du = \int_0^\infty \underbrace{pu^{p-2}}_{w}\cdot\underbrace{u2\exp(-u^2/\kappa^2)}_{v'}du$$

$$= [wv]_0^\infty - \int_0^\infty w'vdu = pu^{p-2}(-\kappa^2\exp(-u^2/\kappa^2)]_0^\infty + \int_0^\infty p(p-2)u^{p-3}\kappa^2\exp(-u^2/\kappa^2)\ du$$

$$= p(p-2)(p-4)C^p\kappa^p \leq C^p 9p^{p/2}\kappa^p$$

pth root $\implies \sqrt{p}C\kappa$.

Next we will show 3 implies 1.

$$\vdots$$

□

Proposition 1.7.3.2. Consider $a \in S^{n-1}$. Suppose $X_1, \ldots, X_n$ are mean 0 with $\|X_i\|_\Psi \leq \kappa$ for all i. Then $\sum_i a_i X_i$ is subgaussian (for all $n \in \mathbb{Z}_{++}$).

Proof.

$$\mathbb{E}\exp\left(\lambda \sum_i a_i X_i\right) = \mathbb{E}\prod_i \exp(\lambda a_i X_i) = \prod_i \mathbb{E}\exp(\lambda a_i X_i) \leq \prod_i \mathbb{E}\exp(C\lambda^2 a_i^2 \kappa^2)$$

$$= \mathbb{E}\exp\left(C\lambda^2 \left(\sum_i a_i^2\right)\kappa^2\right) = \mathbb{E}\exp\left(C_1 \lambda^2 \kappa^2\right).$$

Therefore $\sum_i a_i X_i$ is subgaussian (and $\|\sum_i a_i X_i\|_\Psi \leq C_1 \kappa$).

□

Examples of subgaussian random variables: Gaussian, Bernoulli, bounded.

Definition 1.7.3 (**Hilbert-Schmidt Norm**)**.**

$$\|A\|_{HS} = \sqrt{\sum_{i,j} a_{ij}^2}.$$

Remark. Note

$$\mathrm{Tr}(A^* A) = \|A\|_{HS}^2 = \sum_i s_i^2(A)$$

(the sum of the squared singular values of A).

Proposition 1.7.3.3 (**Hanson-Wright Inequality**)**.**

Proof. □

1.7.4 Topic 1: Invertibility of Matrices

Very tall case: Suppose $A \in \mathbb{R}^{n \times m}$, with $n \gg m$. Entries a_{ij} are independent with $\mathbb{E}a_{ij} = 0$, $\mathbb{E}a_{ij}^2 = 1$, and $\|a_{ij}\|_\Psi \leq \kappa$. A left inverse of A, A^+ exists if the smallest singular value of A is positive; that is, $\delta_m(A) > 0$. Then the columns of A span $\mathbb{R}^m$, and there exists a matrix A^+ such that $A^+ A = I_m$.

Proposition 1.7.4.1 (**Operator norm of** A)**.**

$$\mathbb{P}\left(\|A\| \geq R(\sqrt{n} + \sqrt{m})\right) \leq \exp(-C_2 R^2 (\sqrt{n} + \sqrt{m})^2$$

for $R > C_1$.

Proof. •

$$\|A\| = \max_{x \in S^{n-1}} \|Ax\|_2 = \max_{x \in S^{n-1}} \max_{y \in S^{n-1}} \langle y, Ax \rangle$$

- For fixed $y \in S^{n-1}$, $x \in S^{n-1}$,

$$\langle y, .Ax \rangle = \sum_{i,j} y_i a_{ij} x_j$$

Using the subgaussianity of a_{ij},

$$\left\| \sum_{i,j} x_i x_j a_{ij} \right\|_\Psi \leq C \sqrt{\sum_{i,j} (y_i x_j)^2 \kappa} = CK$$

Using the second definition of the subgaussian norm,

$$\implies \mathbb{P}\left(|\langle y, Ax \rangle| > t \right) \leq 2 \exp(-Ct^2/\kappa^2)$$

We need an upper bound for all $x \in S^{m-1}, y \in S^{n-1}$.

Definition 1.7.4 (ϵ**-net**)**.** $\mathcal{M} \subseteq S^{m-1}$ is an ϵ**-net** of S^{m-1}, if for all $x \in S^{m-1}$, there exists an $x' \in \mathcal{M}$ such that $\|x - x'\|_2 \leq \epsilon$.

Take a 1/4-net in S^{m-1}, called $\mathcal{M}$. Likewise, take a 1/4-net in S^{n-1}, called $\mathcal{N}$.

$$\mathbb{P}\left(\exists x \in \mathcal{M}, y \in \mathcal{N} \text{ s.t. } |\langle y, Ax \rangle| > t \right) \leq |\mathcal{M}||\mathcal{N}| \exp(-Ct^2).$$

Suppose for all $x \in \mathcal{M}, y \in \mathcal{N}$, we have $|\langle y, Ax \rangle| \leq t$. Let $x_0 \in S^{m-1}, y_0 \in S^{n-1}$ such that $\|A\| = \langle y_0, Ax_0 \rangle$. Take $y \in \mathcal{N}, x \in \mathcal{M}$ such that $\|y - y_0\| \leq 1/4, \|x - x_0\| \leq 1/4$. Then

$$\|A\| = \langle y, Ax_0 \rangle + \langle y_0 - y, Ax_0 \rangle \leq \langle y, Ax_0 \rangle + \|y_0 - y\| \|A\|$$

$$= \langle y, Ax \rangle + \langle y, A(x_0 - x) \rangle + \frac{1}{4}\|A\| \leq \langle y, Ax \rangle + \|y\| \|A\| \|x_0 - x\| + \frac{1}{4}\|A\|$$

$$\implies \|A\| \leq \langle y, Ax \rangle + \frac{1}{2}\|A\|$$

$$\iff \|A\| \leq 2\langle y, Ax \rangle \leq 2t$$

since we already have $|\langle y, Ax \rangle| \leq t$.

Then

$$\mathbb{P}\left(\forall x \in \mathcal{M}, \forall y \in \mathcal{N}, |\langle y, Ax \rangle| \leq t \right) \geq 1 - |\mathcal{M}||\mathcal{N}| \exp(-Ct^2/\kappa^2)$$

therefore with probability at least $1 - |\mathcal{M}||\mathcal{N}| \exp(-Ct^2/\kappa^2)$, we have $\|A\| \leq 2t$.

□

Next, adjust t according to $|\mathcal{M}||\mathcal{N}|$ to maximize the probability.

Proposition 1.7.4.2. For $\epsilon < 1$, let $\mathcal{N}$ be an ϵ-net of S^{n-1}. Then $\mathcal{N}$ satisfies $|\mathcal{N}| < (3/\epsilon)^n$.

Proof. Let B_2^n denote the Euclidean unit ball in $\mathbb{R}^n$. Then $x + \epsilon B_2^n = \{x + \epsilon y : y \in B_2^n\}$. We will show that for all $x, y \in \mathcal{N}$,

$$(x + \epsilon/2B_2^n) \bigcap (y + \epsilon/2B_2^n) = \emptyset.$$

Construct $\mathcal{N}$ be picking a point $x_1 \in S^{n-1}$, x_2 so that $|x_1 - x_2| > \epsilon$, keep choosing points until there isn't any more empty space. Once we stop, there are no more points in S^{n-1} such that its epsilon distance away from all points in $\mathcal{N}$ which means $\mathcal{N}$ is an epsilon net. Then

$$\bigcup_{x \in \mathcal{N}} (x + \epsilon/2B_2^n) \subset (1 + \epsilon/2)B_2^n,$$

so

$$\left| \bigcup_{x \in \mathcal{N}} (x + \epsilon/2B_2^n) \right| < |(1 + \epsilon/2)B_2^n| \implies |\mathcal{N}| \cdot |\epsilon/2Bz_2^n| < |1 + \epsilon/2B_2^n|$$

$$\implies |\mathcal{N}|(\epsilon/2)^n < (1 + \epsilon/2)^n \implies |\mathcal{N}| < (2/\epsilon + 1)^n < (3/\epsilon)^n.$$

□

Proposition 1.7.4.3 (**Paley-Zygmund Inequality**)**.** If Z is a nonnegative random variable with $\mathbb{E}Z = 1, \mathbb{E}Z^2 < K$, then for $\lambda \in (0, 1)$, there exists $p > 0$ such that $\mathbb{P}(Z > \lambda) > p$.

Proof.

$$1 = \mathbb{E}Z = \mathbb{E}Z\mathbf{1}_{\{Z \in [0,\lambda]\}} + \mathbb{E}Z\mathbf{1}_{\{Z \in (\lambda,\infty)\}} \leq \lambda + \sqrt{\mathbb{E}Z^2\mathbb{E}(\mathbf{1}_{\{Z \in (\lambda,\infty)\}})} \leq \lambda + \sqrt{K\mathbb{P}(Z > \lambda)}$$

$$\iff \frac{(1 - \lambda)^2}{K} \leq \mathbb{P}(Z > \lambda),$$

which proves the statement for $p = \frac{(1-\lambda)^2}{K}$.

□

1.7.5 Random Graphs

1. **Erdos-Renyi graph:** $\mathcal{G}(N,p)$ line connecting V_i, V_j if and only iff $A_{ij} = 1$. $\mathbb{P}(A_{ij} = 1) = p(n)$, some function of n. All A_{ij} are independent of each other besides the fact that A is symmetric. connected with probability p.
2. **Random d-regular graph model:** $\mathcal{G}_{n,d}$: set of d-regular graphs with n vertices.

These are used for statistical inference, particularly community detection.

$$|G_{nd}| = \Pr(G \text{ simple }) \cdot \frac{(nd-1)!!}{(d!)^n}$$

Theorem 1.7.5.1. $\mathbb{P}(G \text{ is simple}) = e^{-(d^2-1)/4}$.

Remark. It doesn't go to 0 as $n \to \infty$.

1. $$|G_{nd}| = (1+o(1))e^{-(d^2-1)/2} \cdot \frac{(nd-1)!!}{(d!)^n}$$
2. efficient way to sample a d-regular graph. Expected waiting time is $e^{(d^2-1)/4}$.
3. If you can probe property A is true for the configuration model with probability $1 - o(1)$, then the same thing is true for the uniform model.

$$o(1) = \mathbb{P}(A \text{ is not true}) \geq \mathbb{P}(\mathcal{G} \text{ is simple, } A \text{ is not true for } \mathcal{G})$$

$$= \mathbb{P}(A \text{ is not true} \mid \mathcal{G} \text{ is simple})\mathbb{P}(\mathcal{G} \text{ simple})$$

$$\approx \mathbb{P}_{\text{uniform}}(A \text{ is not true})e^{-(d^2-1)/4}$$

$$\mathbb{P}_{\text{uniform}}(A \text{ is not true}) = o(1) \text{ as } n \to \infty.$$

1.8 Distance Correlation

Definition 1.8.1 ($\|\cdot\|_w$**-norm for complex functions; Definition 1 in Székely et al. [2007]).** For complex functions γ defined on $\mathbb{R}^p \times \mathbb{R}^q$, the $\|\cdot\|_w$-norm in the weighted L_2 space of functions on $\mathbb{R}^{p+q}$ is defined by

$$\|\gamma(t,s)\|_w^2 := \int_{\mathbb{R}^{p+q}} |\gamma(t,s)|^2 \, w(t,s)dt \, ds,$$

where $w(t,s)$ is an arbitrary positive weight function for which the integral above exists.

The distance covariance is the $\|\cdot\|_w$ distance between the characteristic functions of two random variables with a suitably chosen weight function. That is, the distance covariance between X and Y is

$$\begin{aligned}\mathcal{V}^2(X,Y;w) &:= \|\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)\|_w^2 \\ &= \int_{\mathbb{R}^{p+q}} |\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)|^2 \, w(t,s)dt \; ds,\end{aligned}$$

where $\phi_X(t)$ is the characteristic function of X, $\phi_Y(s)$ is the characteristic function of Y, and $\phi_{X,Y}(t,s)$ is the joint characteristic function of X and Y (see Definition 1.3.11). We will also define a kind of distance variance

$$\mathcal{V}^2(X;w) = \int_{\mathbb{R}^{p+q}} |\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)|^2 \, w(t,s)dt \; ds.$$

This distance will be 0 if and only if X and Y are independent. Then we will construct distance correlation from distance covariance in an analogous way to how Pearson correlation is defined from covariance.

We want to choose $w(t,s)$ in a way so that the distance correlation is scale invariant (i.e., doesn't change if the units of the variables change). We will also want the distance correlation to be positive for dependent variables. Finally, we will also want $w(t,s)$ not to be integrable over $\mathbb{R}^{p+q}$. The reason why is that if $w(t,s)$ is integrable and both X and Y have finite variance, one can show using Taylor expansions of the underlying characteristic functions that

$$\lim_{\epsilon \to 0} \frac{\mathcal{V}^2(\epsilon X, \epsilon Y; w)}{\sqrt{\mathcal{V}^2(\epsilon X; w)\mathcal{V}^2(\epsilon Y; w)}} = \rho^2(X,Y),$$

where $\rho^2(X,Y)$ is the square of the Pearson correlation between X and Y (the coefficient of determination). Of course, ρ may equal 0 even if X and Y are dependent. So if w is integrable, the distance correlation can be arbitrarily close to zero even if X and Y are dependent. It turns out the below choice of w does what we want and yields relatively simple resuls.

Definition 1.8.2 (Distance Covariance; Definition 2 in Székely et al. [2007]). Let $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$ be random vectors. The **distance covariance** $\mathcal{V}(X,Y)$ between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined by

$$\begin{aligned}\mathcal{V}^2(X,Y) &:= \|\phi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{t},\boldsymbol{s}) - \phi_{\boldsymbol{X}}(\boldsymbol{t})\phi_{\boldsymbol{Y}}(\boldsymbol{s})\|_w^2 \\ &= \int_{\mathbb{R}^{p+q}} \|\phi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{t},\boldsymbol{s}) - \phi_{\boldsymbol{X}}(\boldsymbol{t})\phi_{\boldsymbol{Y}}(\boldsymbol{s})\|^2 \, w(\boldsymbol{t},\boldsymbol{s})d\boldsymbol{t}d\boldsymbol{s}\end{aligned}$$

(that is, $\mathcal{V}(X,Y)$ is the square root of this quantity) where

$$w(t,s) := \frac{1}{c_p c_q |t|_p^{1+p} |s|_q^{1+q}}$$

where

$$c_d := \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}.$$

Distance variance is defined as the square root of

$$\mathcal{V}^2(X) = \int_{\mathbb{R}^{p+q}} |\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)|^2 \, w(t,s) dt \, ds$$

using the same choice of w.

Definition 1.8.3 (**Distance correlation; Definition 3 in Székely et al. [2007]**). The **distance correlation** (dCor) between random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ with finite first moments is the nonnegative number $\mathcal{R}(\boldsymbol{X}, \boldsymbol{Y})$ defined by

$$\mathcal{R}^2(\boldsymbol{X}, \boldsymbol{Y}) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}} & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0, \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0. \end{cases}$$

In Remark 3, Székely et al. [2007] show that if $\mathbb{E}\|\boldsymbol{X}\|_p^2 < \infty$ and $\mathbb{E}\|\boldsymbol{Y}\|_q^2 < \infty$,

$$\mathcal{V}^2(\boldsymbol{X}, \boldsymbol{Y}) = S_1 + S_2 - 2S_3,$$

where

$$\begin{aligned} S_1 &:= \mathbb{E}\left[\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_p \|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|_q\right], \\ S_2 &:= \mathbb{E}\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_p \mathbb{E}\|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|_q, \quad \text{and} \\ S_3 &:= \mathbb{E}\left(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_p \|\boldsymbol{Y}_1 - \boldsymbol{Y}_3\|_q\right), \end{aligned}$$

where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are two independent draws of $\boldsymbol{X}$ (and likewise for $\boldsymbol{Y}$). Given a random sample $\{(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)$, they suggest the moment estimator

$$\hat{\mathcal{V}}^2(\boldsymbol{X}, \boldsymbol{Y}) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3$$

where

$$\begin{aligned} \hat{S}_1 &:= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|\boldsymbol{X}_i - \boldsymbol{X}_j\|_p \|\boldsymbol{Y}_i - \boldsymbol{Y}_j\|_q, \\ \hat{S}_2 &:= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|\boldsymbol{X}_i - \boldsymbol{X}_j\|_p \cdot \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|\boldsymbol{Y}_i - \boldsymbol{Y}_j\|_q, \quad \text{and} \\ \hat{S}_3 &:= \frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{\ell=1}^n \|\boldsymbol{X}_i - \boldsymbol{X}_j\|_p \|\boldsymbol{Y}_i - \boldsymbol{Y}_\ell\|_q. \end{aligned}$$

Li et al. [2012] propose using distance correlation for screening, noting "it allows for arbitrary regression relationship of $\boldsymbol{y}$ onto $\boldsymbol{X}$, regardless of whether it is linear or nonlinear. The distance correlation also permits univariate and multivariate responses, regardless of whether it is continuous, discrete, or categorical. In addition, it allows for groupwise predictors. Thus, this distance correlation-based screening procedure is completely model-free."

Bibliography

R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, USA, 5th edition, 2019. URL https://services.math.duke.edu/~rtd/PTE/pte.html.

G. Grimmett and D. Stirzaker. *Probability and random processes*, volume 80. Oxford university press, 2001. URL http://scholar.google.com/scholar.bib?q=info:xzStZXK20NkJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&ct=citation&cd=0.

I. M. Johnstone. On The Distribution of the Largest Eigenvalue in Principal Components Analysis. *The Annals of Statistics*, 29(2):295–327, 2001. URL https://projecteuclid.org/download/pdf{_}1/euclid.aos/1009210544.

R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012. ISSN 01621459. doi: 10.1080/01621459.2012.695654. URL https://www.tandfonline.com/action/journalInformation?journalCode=uasa20.

S. Ross. *Stochastic Processes*. Wiley series in probability and statistics. Wiley India Pvt. Limited, 2nd ed. edition, 2008. ISBN 9788126517572. URL https://books.google.com/books?id=HVHqPgAACAAJ.

S. Ross. *Introduction to Probability Models*. Academic Press, Boston, eleventh edition edition, 2014. ISBN 978-0-12-407948-9. doi: https://doi.org/10.1016/B978-0-12-407948-9.00012-8. URL http://www.sciencedirect.com/science/article/pii/B9780124079489000128.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, dec 2007. ISSN 00905364. doi: 10.1214/009053607000000505.