All of Statistics

My proposed solutions to the book:

All of Statistics - A Concise Course in Statistical Inference.

Publishing code and pdf to the following GitHub repository.
`https://github.com/CoveredInChocolate/AllStatistics`

Let me know if you find any mistakes! I am sure there are plenty...

Contents

1 Probability

Exercises

1.1
Proving the Continuity of Probabilities.

1.8 Theorem (Continuity of Probabilities). If $A_n \to A$ then

$$\mathbb{P}(A_n) \to \mathbb{P}(A)$$

as $n \to \infty$.

Proof. Suppose that A_n is monotone increasing so that $A_1 \subset A_2 \subset \ldots$. Let $A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$. Define:

$$\begin{aligned}
&B_1 = A_1 \\
&B_2 = \{\omega \in \Omega \ : \ \omega \in A_2, \omega \notin A_1\} \\
&B_3 = \{\omega \in \Omega \ : \ \omega \in A_3, \omega \notin A_2, \omega \notin A_1\} \\
&\quad \vdots
\end{aligned}$$

(1) Showing that B_i are disjoint sets.
We have $B_1 = A_1$. Since $B_2 = \{\omega \in \Omega \ : \ \omega \in A_2, \omega \notin A_1\}$, we can rewrite this as $B_2 = A_2 - A_1$ by definition of set difference.

Since $B_3 = \{\omega \in \Omega \ : \ \omega \in A_3, \omega \notin A_2, \omega \notin A_1\}$ and since $A_1 \subset A_2$, we can rewrite this as $B_3 = A_3 - (A_1 \bigcup A_2)$. In general; $B_k = A_k - (\bigcup_{i=1}^{k-1} A_i)$.

Assuming some random sets B_m, B_p for some $m, p \in \mathbb{N}$. Without loss of generality, we assume $m > p$. Then:

$$\begin{aligned}
B_m \bigcap B_p &= \Big(A_m - (\bigcup_{i=1}^{m-1} A_i)\Big) \bigcap \Big(A_p - (\bigcup_{i=1}^{p-1} A_i)\Big) \\
&= \Big(A_m \bigcap (\bigcup_{i=1}^{m-1} A_i)^c\Big) \bigcap \Big(A_p \bigcap (\bigcup_{i=1}^{p-1} A_i)^c\Big)
\end{aligned}$$

DeMorgan's law

$$= \Big(A_m \cap A_1^c \cap A_2^c \cap \ldots \cap A_p^c \cap \ldots \cap A_{m-1}^c\Big) \bigcap \Big(A_p \cap A_1^c \cap A_2^c \cap \ldots \cap A_{p-1}^c\Big)$$

Reshuffling terms and repeated use of $C \cap C = C$

$$\begin{aligned}
&= A_m \cap A_1^c \cap A_2^c \cap \ldots \cap \underbrace{A_p^c \cap A_p}_{=\emptyset} \cap \ldots \cap A_{m-1}^c\Big) \\
&= \emptyset
\end{aligned}$$

Since $B_m \cap B_p = \emptyset$, they are disjoint. (*Quite certain this is a correct argument, but could have made it a bit easier by going directly to the $A_3 \cap A_2^c \cap A_1^c$ version of the sets.*)

(2) Showing that

$$A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$$

for each n. Showing this with an induction argument. Note that $A_1 = B_1$, and $A_2 \cup A_1 = A_2$ since $A_1 \subset A_2$. Note also that:

$$\begin{aligned} B_2 \cup B_1 &= (A_2 \cap A_1^c) \cup A_1 \\ &= (A_2 \cup A_1) \cap (A_1^c \cup A_1) \\ &= (A_2 \cup A_1) \cap (\Omega) \\ &= A_2 \cup A_1 \\ &= A_2 \end{aligned}$$

So this is true for $n = 2$. Assume that $A_k = \cup_{i=1}^{k} A_i = \cup_{i=1}^{k} B_i$ for $k \in \mathbb{N}$. Showing that this applies to $k + 1$.

$$\bigcup_{i=1}^{k+1} A_i = A_{k+1} \cup \Big(\bigcup_{i=1}^{k} A_i \Big) = A_{k+1} \cup A_k = A_{k+1}$$

$$\begin{aligned} \bigcup_{i=1}^{k+1} B_i &= B_{k+1} \cup \Big(\bigcup_{i=1}^{k} B_i \Big) \\ &= B_{k+1} \cup A_k \\ &= \Big[A_{k+1} \cap \Big(A_k^c \cap \ldots \cap A_1^c \Big) \Big] \cup A_k \\ &= \Big[A_{k+1} \cap \Big(A_k \cup \ldots \cup A_1 \Big)^c \Big] \cup A_k \\ &= \Big[A_{k+1} \cap A_k^c \Big] \cup A_k \\ &= (A_{k+1} \cup A_k) \cap (A_k^c \cup A_k) \\ &= (A_{k+1} \cup A_k) \cap \Omega \\ &= A_{k+1} \cup A_k \\ &= A_{k+1} \end{aligned}$$

Result verified by induction argument.

(3) Showing that

$$\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i. \qquad (\diamondsuit)$$

By definition:

$$A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i.$$

By step (2):

$$A = \lim_{n\to\infty} A_n = \lim_{n\to\infty} \bigcup_{i=1}^{n} A_i = \lim_{n\to\infty} \bigcup_{i=1}^{n} B_i = \bigcup_{i=1}^{\infty} B_i.$$

Hence, $(\diamondsuit)$ is satisfied.

1.2
Proving some well known results by using the axioms:

$$\mathbb{P}(A) \geqslant 0,\ \forall A \subset \Omega \tag{A.1}$$

$$\mathbb{P}(\Omega) = 1, \tag{A.2}$$

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \tag{A.3}$$

where $A_1, A_2, \ldots$ are disjoint in $(A.3)$.

• **Claim**: $\mathbb{P}(\emptyset) = 0$.
Proof. Since $\Omega \cap \emptyset = \emptyset$ and the empty set is disjoint with itself, we can make set $E_1 = \Omega$ and $E_k = \emptyset$ for all $k \geqslant 2$. Now assume for contradiction that $\mathbb{P}(\emptyset) = a > 0$ for some $a \in \mathbb{R}$. Then by $(A.3)$:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) \overset{A.3}{=} \sum_{i=1}^{\infty} \mathbb{P}(E_i) = \mathbb{P}(\Omega) + \sum_{i=2}^{\infty} \mathbb{P}(\emptyset) = 1 + \sum_{i=2}^{\infty} a = \infty$$

Now instead of using $(A.3)$ we use that the infinite set of E_i becomes Ω:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \mathbb{P}(\Omega) \overset{A.2}{=} 1$$

We have reached a contradiction. This shows that $\mathbb{P}(\emptyset) = 0$. $\square$

• **Claim**: $A \cap B = \emptyset \implies \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
Proof. Set $E_1 = A$ and $E_2 = B$ and $E_k = \emptyset$ for all $k \geqslant 3$. Then:

$$\mathbb{P}(A \cup B) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) \overset{A.3}{=} \sum_{i=1}^{\infty} \mathbb{P}(E_i) = \mathbb{P}(A) + \mathbb{P}(B) + \sum_{k=3}^{\infty} \mathbb{P}(\emptyset) = \mathbb{P}(A) + \mathbb{P}(B) + 0 = \mathbb{P}(A) + \mathbb{P}(B)$$

$\square$

With this result, we can apply $(A.3)$ indirectly to any finite sum of disjoint sets. All other sets are set to $\emptyset$ and then (A.3) is applied. Not giving a formal argument, but if A, B, C are mutually disjoint we can set $E_k = \emptyset$ for all $k \geqslant 4$.

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) \overset{A.3}{=} \sum_{i=1}^{\infty} \mathbb{P}(E_i) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) + 0 = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C)$$

• **Claim**: $A \subset B \implies \mathbb{P}(A) \leqslant \mathbb{P}(B)$.
Proof. Since $A \subset B$ we can split B into two disjoint parts: $B = A \cup B - A$. By axiom (A.3):

$$\mathbb{P}(B) = \mathbb{P}(A \cup B - A) \overset{A.3}{=} \mathbb{P}(A) + \mathbb{P}(B - A) \overset{A.1}{\geqslant} \mathbb{P}(A) \implies \mathbb{P}(A) \leqslant \mathbb{P}(B)$$ $\square$

• **Claim**: $0 \leqslant \mathbb{P}(A) \leqslant 1$.
Proof. Since $A \subset \Omega$ and by the previous proof: $\mathbb{P}(A) \leqslant \mathbb{P}(\Omega) = 1$ Combining this with axiom (A.1):

$$0 \overset{A.1}{\leqslant} \mathbb{P}(A) \leqslant \mathbb{P}(\Omega) = 1 \implies 0 \leqslant \mathbb{P}(A) \leqslant 1$$ $\square$

• **Claim**: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
Proof. Since $A \cup A^c$ are disjoint, we get by finite version of (A.3):

$$\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$$

Since $\Omega = A \cup A^c$ we also get by (A.2):

$$\mathbb{P}(A \cup A^c) = \mathbb{P}(\Omega) = 1$$

Putting them together:

$$\mathbb{P}(A) + \mathbb{P}(A^c) = 1 \quad \Longrightarrow \quad \mathbb{P}(A^c) = 1 - \mathbb{P}(A) \qquad \square$$

1.3
Ω is a sample space and $A_1, A_2, \ldots$ are events. We define:

$$B_n = \bigcup_{i=n}^{\infty} A_i, \qquad C_n = \bigcap_{i=n}^{\infty} A_i$$

(a) **Claim**: $B_1 \supset B_2 \supset \ldots$
Proof. This follows directly from the definition of B_n. First we see that:

$$B_1 = \bigcup_{i=1}^{\infty} A_i = A_1 \cup \left(\bigcup_{i=2}^{\infty} A_i \right) \supset \bigcup_{i=2}^{\infty} A_i = B_2, \implies B_1 \supset B_2$$

The same argument can be used for some $k \in \mathbb{N}$.

$$B_k = \bigcup_{i=k}^{\infty} A_i = A_k \cup \left(\bigcup_{i=k+1}^{\infty} A_i \right) \supset \bigcup_{i=k+1}^{\infty} A_i = B_{k+1}, \implies B_k \supset B_{k+1}$$

From this, it follows that $B_1 \supset B_2 \supset \ldots$. $\square$

Claim: $C_1 \subset C_2 \subset \ldots$
Proof. This also follows directly from the definition. First note that:

$$C_1 = \bigcap_{i=1}^{\infty} A_i = A_1 \cap \left(\bigcap_{i=2}^{\infty} A_i \right) \subset \bigcap_{i=2}^{\infty} A_i = C_2 \implies C_1 \subset C_2$$

And, in general, for some $k \in \mathbb{N}$.

$$C_k = \bigcap_{i=k}^{\infty} A_i = A_k \cap \left(\bigcap_{i=k+1}^{\infty} A_i \right) \subset \bigcap_{i=k+1}^{\infty} A_i = C_{k+1} \implies C_k \subset C_{k+1}$$

This shows that $C_1 \subset C_2 \subset \ldots$. $\square$

(b) **Claim**:

$$\omega \in \bigcap_{n=1}^{\infty} B_n \iff \omega \in A_1, A_2, \dots$$

Proof.
$\Rightarrow$) Assume $\omega \in \bigcap_{n=1}^{\infty} B_n$, which means $\omega \in B_k, \forall k \in \mathbb{N}$. Expanding the intersection:

$$\omega \in B_1 \cap B_2 \cap \dots \cap B_k \cap B_{k+1} \cap \dots$$

Now, from the definition of B_1, like we used above:

$$B_1 = \bigcup_{i=1}^{\infty} A_i = A_1 \cup \bigcup_{i=2}^{\infty} A_i = A_1 \cup B_2$$

So, we can write: $B_1 \cap B_2 = (A_1 \cup B_2) \cap B_2 = (A_1 \cap B_2) \cup (B_2 \cap B_2) = (A_1 \cap B_2) \cup B_2$. And in general:

$$B_k = \bigcup_{i=k}^{\infty} A_i = A_k \cup \bigcup_{i=k+1}^{\infty} A_i = A_k \cup B_{k+1},$$

so $B_k \cap B_{k+1} = (A_k \cup B_{k+1}) \cap B_{k+1} = (A_k \cap B_{k+1}) \cup B_{k+1}$. So, each B_i can be decomposed into $(A_i \cap B_{i+1}) \cup B_{i+1}$ and the B_{i+1} is rewritten as its own intersection. Ultimately, we get:

$$\begin{aligned}
&\omega \in B_1 \cap B_2 \cap B_3 \cap \dots \cap B_k \cap B_{k+1} \cap \dots \implies \\
&\omega \in \big[(A_1 \cap B_2) \cup B_2\big] \cap B_3 \cap \dots \cap B_k \cap B_{k+1} \cap \dots \implies \\
&\omega \in \big[(A_1 \cap B_2) \cup (A_2 \cap B_3) \cup B_3\big] \cap \dots \cap B_k \cap B_{k+1} \cap \dots \implies \\
&\omega \in (A_1 \cap B_2) \cup (A_2 \cap B_3) \cup \dots \cup (A_k \cap B_{k+1}) \cup (A_{k+1} \cap B_{k+2}) \cup \dots \implies \\
&\omega \in (A_k \cap B_{k+1}) \forall k \in \mathbb{N} \text{ by assumption that } \omega \in B_k \forall k \in \mathbb{N} \implies \\
&\omega \in A_k, \forall k \in \mathbb{N}
\end{aligned}$$

which means $\omega \in A_1, A_2, \dots, A_k, A_{k+1}, \dots$ proving the first implication.

$\Leftarrow$) Now assume $\omega \in A_1, A_2, \dots, A_k, \dots$ which means that $\omega \in A_k$ for any $k \in \mathbb{N}$. Then,

$$\omega \in A_k \implies \omega \in A_k \cup \left(\bigcup_{i=k+1}^{\infty} A_i \right) = \bigcup_{i=k}^{\infty} A_i = B_k \implies \omega \in B_k,$$

So ω is included in all B_k for $k \in \mathbb{N}$ which means ω is also in the intersection of all B_i.

$$\omega \in B_1, B_2, \dots, B_k, \dots \implies \omega \in \bigcap_{n=1}^{\infty} B_n$$

By implication both ways, we have proved the equivalency. $\square$

(c) *Skipped.*

1.4
Proving DeMorgan's laws. First in the case when $i = \{1, 2, \ldots, n\}$.

Claim:

$$\left(\bigcup_{i=1}^{n} A_i \right)^c = \bigcap_{i=1}^{n} A_i^c$$

PROOF. Starting with $x \in (\cup_{i=1}^{n} A_i)^c$, we have the following equivalencies:

$$\begin{aligned} x \in \left(\bigcup_{i=1}^{n} A_i \right)^c &\iff x \notin \bigcup_{i=1}^{n} A_i \\ &\iff x \notin A_1 \text{ and } x \notin A_2 \text{ and } \ldots \text{ and } x \notin A_n \\ &\iff x \in A_1^c \text{ and } x \in A_2^c \text{ and } \ldots \text{ and } x \in A_n^c \\ &\iff x \in \bigcap_{i=1}^{n} A_i^c \end{aligned}$$

□

Claim:

$$\left(\bigcap_{i=1}^{n} A_i \right)^c = \bigcup_{i=1}^{n} A_i^c$$

PROOF. Starting with $x \in (\cap_{i=1}^{n} A_i)^c$, we have the following equivalencies:

$$\begin{aligned} x \in \left(\bigcap_{i=1}^{n} A_i \right)^c &\iff x \notin \bigcap_{i=1}^{n} A_i \\ &\iff \exists p \in \{1, \ldots, n\} \; x \notin A_p \\ &\iff \exists p \in \{1, \ldots, n\} \; x \in A_p^c \\ &\iff x \in \bigcup_{i=1}^{n} A_i^c \end{aligned}$$

□

Proving these results for a random index set is essentially the same, except finding some way of numerating the index set with $i_1, \ldots, i_n$ (since it has to be countably infinite) which I won't bother doing.

1.5
Tossing a fair coin until we get exactly two heads. Description of the sample space:

$$\Omega = \{(\omega_1, \omega_2, \ldots, \omega_{k-1}, \omega_k) : \omega_i \in \{H, T\}, \omega_{k-1} = \omega_k = H, \omega_{p-1} = H \Rightarrow \omega_p = T, p < k\}$$

Now, to calculate the probability that there will be exactly k tosses. Let us look at a few specific examples:

$$\begin{aligned} k = 2 : \{H, H\} &\implies \mathbb{P}(2) = \frac{1}{2^2} = \frac{1}{4} = 0.25 \\ k = 3 : \{T, H, H\} &\implies \mathbb{P}(3) = \frac{1}{2^3} = \frac{1}{8} = 0.125 \end{aligned}$$

For all tosses after $k > 3$, the last three tosses have to be $\{T, H, H\}$.

$$k=4: \quad \left.\begin{array}{c} H \\ T \end{array}\right| , T, H, H = 2 \cdot \frac{1}{2^4} = \frac{1}{8} = 0.125$$

$$k=5: \quad \left.\begin{array}{c} TT \\ TH \\ HT \end{array}\right| , T, H, H = 3 \cdot \frac{1}{2^5} = \frac{3}{32} = 0.09375$$

$$k=6: \quad \left.\begin{array}{c} TTT \\ TTH \\ THT \\ HTT \\ HTH \end{array}\right| , T, H, H = 5 \cdot \frac{1}{2^6} = \frac{5}{64} = 0.078125$$

$$k=7 \quad \left.\begin{array}{c} TTTT \\ TTTH \\ TTHT \\ THTT \\ HTTT \\ THTH \\ HTHT \\ HTTH \end{array}\right| , T, H, H = 8 \cdot \frac{1}{2^7} = \frac{1}{2^4} = \frac{1}{16} = 0.0625$$

We can see a pattern emerging. The last three tosses are fixed, and for the first $k-3$ tosses, we have to count all the combinations that do not contain 2 simultaneous H. The details are quite interesting, and it turns out that the number after the $k-3$ first tosses follows the Fibonacci sequence (as we can see above). Defining:

$$F(1) = 2, F(2) = 3, F(3) = 5, \ldots$$

In general, we will then get:

$$\mathbb{P}(K=k) = \begin{cases} \dfrac{1}{2^k} & k = 2, 3 \\ \dfrac{F(k-3)}{2^k} & k > 3 \end{cases}$$

1.6

Let the sample space be $\Omega = \{0, 1, 2, \ldots\}$ (which is $\{0\} \cup \mathbb{N}$). We are going to prove that there is no uniform distribution on this sample space. That is, if $\mathbb{P}(A) = \mathbb{P}(B)$ whenver $|A| = |B|$ (the cardinality of the sets) then $\mathbb{P}$ cannot satisfy the axioms of probability.

The first two properties are satisfied. If we take some $k \in \mathbb{N}$ and define $A_1 = \{1, \ldots, k\}$ in such a way that $P(A_1) > 0$, and generally define $A_m = \{mk+1, mk+2, \ldots, mk+k\}$, then $A_n \cap A_m = \emptyset$. (This can be checked by setting $k = 100$ and $n = 5$ and $m = 6$ for instance). Also, $|A_n| = |A_m|$ and $\mathbb{P}(A_n) = \mathbb{P}(A_m) > 0$. This will cause problems in the third axiom:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \infty,$$

since we are summing an infinite number of positive values. The axioms of probability are not satisfied, so we cannot have a uniform distribution on $\{0, 1, 2, \ldots\}$.

1.7

For some events $A_1, A_2, \ldots$, we are going to show that:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leqslant \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Proof. Following the hint, we define:

$$B_n = A_n - \bigcup_{i=1}^{n-1} A_i.$$

Next, we show that any B_m and B_n are disjoint for $n, m \in \mathbb{N}$ and $m > n$.

$$\begin{aligned}
B_m \cap B_n &= \left(A_m - \bigcup_{i=1}^{m-1} A_i\right) \cap \left(A_n - \bigcup_{i=1}^{n-1} A_i\right) \\
&= \left(A_m \cap \left(\bigcup_{i=1}^{m-1} A_i\right)^c\right) \cap \left(A_n \cap \left(\bigcup_{i=1}^{n-1} A_i\right)^c\right) \\
&= \left(A_m \cap \bigcap_{i=1}^{m-1} A_i^c\right) \cap \left(A_n \cap \bigcap_{i=1}^{n-1} A_i^c\right) \\
&= \left(A_m \cap A_{m-1}^c \cap \ldots \cap A_n^c \cap \ldots \cap A_2^c \cap A_1^c\right) \cap \left(A_n \cap A_{n-1}^c \cap \ldots \cap A_2^c \cap A_1^c\right) \\
&= A_m \cap A_{m-1}^c \cap \ldots \cap A_2^c \cap A_1^c \cap A_{n-1}^c \cap \ldots \cap A_2^c \cap A_1^c \cap \underbrace{A_n^c \cap A_n}_{=\emptyset} \\
&= \emptyset
\end{aligned}$$

Hence, the sets B_i are mutually disjoint. Next, we show that:

$$\left(\bigcup_{n=1}^{\infty} A_n\right) = \left(\bigcup_{n=1}^{\infty} B_n\right).$$

We will use an induction argument. For the first element we have: $B_1 = A_1$, and:

$$\begin{aligned}
\cup_{i=1}^{2} B_i &= B_1 \cup B_2 \\
&= A_1 \cup (A_2 - A_1) \\
&= A_1 \cup (A_2 \cap A_1^c) \\
&= (A_1 \cup A_2) \cap (A_1 \cup A_1^c) \\
&= (A_1 \cup A_2) \cap \Omega \\
&= (A_1 \cup A_2) \\
&= \cup_{i=1}^{2} A_i
\end{aligned}$$

So, we assume that for $k \in \mathbb{N}$, we have:

$$\left(\bigcup_{n=1}^{k} A_n\right) = \left(\bigcup_{n=1}^{k} B_n\right),$$

and show that this applies to $k + 1$.

$$
\begin{aligned}
\bigcup_{i=1}^{k+1} B_i &= B_{k+1} \cup \left(\bigcup_{i=1}^{k} B_i \right) \\
&= B_{k+1} \cup \left(\bigcup_{i=1}^{k} A_i \right) \\
&= \left(A_{k+1} - \bigcup_{i=1}^{k} A_i \right) \cup \left(\bigcup_{i=1}^{k} A_i \right) \\
&= \left(A_{k+1} \cap \left(\bigcup_{i=1}^{k} A_i \right)^c \right) \cup \left(\bigcup_{i=1}^{k} A_i \right) \\
&= \left(A_{k+1} \cup \bigcup_{i=1}^{k} A_i \right) \cap \left[\left(\bigcup_{i=1}^{k} A_i \right)^c \cup \left(\bigcup_{i=1}^{k} A_i \right) \right] \\
&= \left(\bigcup_{i=1}^{k+1} A_i \right) \cap \Omega \\
&= \bigcup_{i=1}^{k+1} A_i
\end{aligned}
$$

So, by induction, we have verfied the equality for $k+1$, so it follows that the sets are equal for all indexes.

By definition of B_n, we have $B_n \subset A_n$ which means $\mathbb{P}(B_n) \leqslant \mathbb{P}(A_n)$. Using this, we can prove the main statement:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n \right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n \right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) \leqslant \sum_{n=1}^{\infty} \mathbb{P}(A_n). \qquad \square$$

1.8

Suppose that $\mathbb{P}(A_i) = 1$ for each A_i. We are going to prove that:

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i \right) = 1.$$

Proof. We will prove this by induction. For events A_1 and A_2, we know that $\mathbb{P}(A_1) = 1$ and $\mathbb{P}(A_2) = 1$. By the probability axioms:

$$A_1 \subset A_1 \cup A_2 \subset \Omega \implies \mathbb{P}(A_1) \leqslant \mathbb{P}(A_1 \cup A_2) \leqslant \mathbb{P}(\Omega)$$

Using that $\mathbb{P}(A_1) = 1$ and $\mathbb{P}(\Omega) = 1$:

$$1 \leqslant \mathbb{P}(A_1 \cup A_2) \leqslant 1 \implies \mathbb{P}(A_1 \cup A_2) = 1$$

By Lemma 1.6:

$$
\begin{aligned}
\mathbb{P}(A_1 \cup A_2) &= \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2) \\
1 &= 2 - \mathbb{P}(A_1 \cap A_2) \\
\mathbb{P}(A_1 \cap A_2) &= 1
\end{aligned}
$$

Assuming this holds for $k \in \mathbb{N}$, i.e.:

$$\mathbb{P}\left(\bigcap_{i=1}^{k} A_i\right) = 1.$$

To simplify the notation, we call this intersection W_k, so $\mathbb{P}(W_k) = 1$. Now we will show it also applies to $k+1$.

$$A_1 \cap A_2 \cap \ldots \cap A_k \cap A_{k+1} = W_k \cap A_{k+1}$$

Again, from the probability axioms:

$$W_k \subset W_k \cup A_{k+1} \subset \Omega \quad \Longrightarrow \quad \mathbb{P}(W_k) \leqslant \mathbb{P}(W_k \cup A_{k+1}) \leqslant \mathbb{P}(\Omega)$$

Using that $\mathbb{P}(W_k) = 1$ and $\mathbb{P}(\Omega) = 1$:

$$1 \leqslant \mathbb{P}(W_k \cup A_{k+1}) \leqslant 1 \quad \Longrightarrow \quad \mathbb{P}(W_k \cup A_{k+1}) = 1$$

By Lemma 1.6, and using that $\mathbb{P}(A_i) = 1$ for each i:

$$\begin{aligned}
\mathbb{P}(W_k \cup A_{k+1}) &= \mathbb{P}(W_k) + \mathbb{P}(A_{k+1}) - \mathbb{P}(W_k \cap A_{k+1}) \\
1 &= 2 - \mathbb{P}(W_k \cap A_{k+1}) \\
\mathbb{P}(W_k \cap A_{k+1}) &= 1 \\
\mathbb{P}\left(\bigcap_{i=1}^{k} A_i \cap A_{k+1}\right) &= 1 \\
\mathbb{P}\left(\bigcap_{i=1}^{k+1} A_i\right) &= 1
\end{aligned}$$

Which verifies that equality holds for $k+1$. By induction, we can conclude that:

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = 1. \qquad \square$$

1.9

For a fixed B such that $\mathbb{P}(B) > 0$, we will show that $\mathbb{P}(\cdot|B)$ satisfies the axioms of probability. Recalling the definition of conditional probability:

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

• **Axiom 1**. $\mathbb{P}(A|B) \geqslant 0$ for all A.

Proof. The set $A \cap B$ is an event wrt. $\mathbb{P}(\cdot)$, so $\mathbb{P}(A \cap B) = a \geqslant 0$.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{a}{\mathbb{P}(B)} \geqslant 0,$$

since $\mathbb{P}(B) > 0$. $\square$

• **Axiom 2**. $\mathbb{P}(\Omega|B) = 1$.
PROOF.
$$\mathbb{P}(\Omega|B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1 \qquad \square$$

• **Axiom 3**. If $A_1, A_2, \ldots$ are disjoint, then:
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$$

PROOF. First a supporting argument. If we define $C_i := A_i \cap B$ for each i, then $C_1, C_2, \ldots$ are mutually disjoint sets and by (A.3):
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} C_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(C_i).$$

This means:
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) = \frac{\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \cap B\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\bigcup_{i=1}^{\infty} C_i\right)}{\mathbb{P}(B)} = \frac{\sum_{i=1}^{\infty} \mathbb{P}(C_i)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$$

which proves the final axiom for the conditional probability. $\square$

1.10 Monty Hall
There are three doors, 1, 2, 3, and there is a prize between one of them. We don't know which door contains the prize, so the probability of selecting the correct door is evenly spread out:
$$P(1) = \frac{1}{3}, \quad P(2) = \frac{1}{3}, \quad P(3) = \frac{1}{3}$$

To simplify the calculations, we assume we pick door 1, and that the prize is not behind door 3. Now we will decide if it's better to switch or not. What is the probability that the host will open door III conditioned on where the prize is?
$$\begin{aligned} &P(\mathrm{III}|1) = \frac{1}{2} && \text{(Correct door is selected, door 3 is opened randomly)} \\ &P(\mathrm{III}|2) = 1 && \text{(Host never reveals the prize)} \\ &P(\mathrm{III}|3) = 0 && \text{(Host never reveals the prize)} \end{aligned}$$

Now, what is the probability that the prize is behind door number 2, given that door III was opened? This can be calculated by Bayes Law.
$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

In our case, probability of door 2 containing the prize, given that door III was opened:

$$\begin{aligned}
P(2|\text{III}) &= \frac{\mathbb{P}(\text{III}|2)\mathbb{P}(2)}{\mathbb{P}(\text{III}|1)\mathbb{P}(1)+\mathbb{P}(\text{III}|2)\mathbb{P}(2)+\mathbb{P}(\text{III}|3)\mathbb{P}(3)} \\
&= \frac{(1)(\frac{1}{3})}{(\frac{1}{2})(\frac{1}{3})+(1)(\frac{1}{3})+(0)(\frac{1}{3})} \\
&= \frac{\frac{1}{3}}{\frac{1}{6}+\frac{1}{3}+0} = \frac{1}{3}\Big/\frac{1}{2} \\
&= \frac{2}{3}
\end{aligned}$$

Probability that door 1 contains the prize, given that door III was opened:

$$\begin{aligned}
P(1|\text{III}) &= \frac{\mathbb{P}(\text{III}|1)\mathbb{P}(1)}{\mathbb{P}(\text{III}|1)\mathbb{P}(1)+\mathbb{P}(\text{III}|2)\mathbb{P}(2)+\mathbb{P}(\text{III}|3)\mathbb{P}(3)} \\
&= \frac{(\frac{1}{2})(\frac{1}{3})}{(\frac{1}{2})(\frac{1}{3})+(1)(\frac{1}{3})+(0)(\frac{1}{3})} \\
&= \frac{\frac{1}{6}}{\frac{1}{6}+\frac{1}{3}+0} = \frac{1}{6}\Big/\frac{1}{2} \\
&= \frac{1}{3}
\end{aligned}$$

This shows that we are better off switching doors then remaining in the one we initially selected.

1.11

Claim. If A and B are independent, then A^c and B^c are independent.

Proof. By definition of independence $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. By property of probabilities: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ and $\mathbb{P}(B^c) = 1 - \mathbb{P}(B)$. Applying Lemma 1.6 to $A^c \cup B^c$:

$$\mathbb{P}(A^c \cup B^c) = \mathbb{P}(A^c) + \mathbb{P}(B^c) - \mathbb{P}(A^c \cap B^c) \tag{2.11.1}$$

By property of complements, and using DeMorgan's law:

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}([A \cap B]^c) = 1 - \mathbb{P}(A^c \cup B^c)$$

Which we can rewrite in the following way, and then apply the independence of A and B:

$$\mathbb{P}(A^c \cup B^c) = 1 - \mathbb{P}(A \cap B) = 1 - \mathbb{P}(A)\mathbb{P}(B)$$

By property of complements, we can represent this as:

$$\begin{aligned}
\mathbb{P}(A^c \cup B^c) &= 1 - [1 - \mathbb{P}(A^c)][1 - \mathbb{P}(B^c)] \\
&= 1 - \left[1 - \mathbb{P}(B^c) - \mathbb{P}(A^c) + \mathbb{P}(A^c)\mathbb{P}(B^c)\right] \\
&= \mathbb{P}(A^c) + \mathbb{P}(B^c) - \mathbb{P}(A^c)\mathbb{P}(B^c) \qquad (2.11.2)
\end{aligned}$$

By setting (2.11.1) equal to (2.11.2) we can deduce that $\mathbb{P}(A^c)\mathbb{P}(B^c) = \mathbb{P}(A^c \cap B^c)$ which proves independence. $\square$

Alternative Proof. Slightly more compact argument, but it still relies on Lemma 1.6.

$$\begin{aligned}
\mathbb{P}(A^c)\mathbb{P}(B^c) &= \big[1-\mathbb{P}(A)\big]\big[1-\mathbb{P}(B)\big] \\
&= 1-\mathbb{P}(A)-\mathbb{P}(B)+\mathbb{P}(A)\mathbb{P}(B) \\
&= 1-\big[\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A)\mathbb{P}(B)\big] \\
&\overset{IND}{=} 1-\big[\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A\cap B)\big] \\
&\overset{L1.6}{=} 1-\mathbb{P}(A\cup B) \\
&= \mathbb{P}([A\cup B]^c) \\
&= \mathbb{P}(A^c\cap B^c)
\end{aligned}$$

□

1.12
Initially, we have three cards that we call G (Green both sides), R (Red both sides) and M (mixed colors). Next we define r as 'Red side' and g as 'Green side'. We can define the joint distribution as follows:

	G	M	R	
r		1/6	2/6	1/2
g	2/6	1/6		1/2
	1/3	1/3	1/3	

From the marginal distributions, we can see that the initial probabilities for selecting each card is:

$$\mathbb{P}(G)=\frac{1}{3},\quad \mathbb{P}(R)=\frac{1}{3},\quad \mathbb{P}(M)=\frac{1}{3}$$

And the probabilities for randomly viewing a colored card side:

$$\mathbb{P}(g)=\frac{1}{2},\quad \mathbb{P}(r)=\frac{1}{2}$$

Note that we do not have independene, since $\mathbb{P}(G\cap g)=2/6$ while $\mathbb{P}(G)\mathbb{P}(g)=(1/3)(1/2)=1/6$.

We want to find the probability that we have a green card given that we observe a green side. This can be calculated by:

$$\mathbb{P}(G|g)=\frac{\mathbb{P}(G\cap g)}{\mathbb{P}(g)}=\frac{2/6}{1/2}=4/6=2/3$$

1.13
A fair coin is thrown until we have both H and T.
(a) The sample space is:

$$\Omega=\{(\omega_1,\ldots,\omega_K):\omega_i\in\{H,T\},1\leqslant i\leqslant K \text{ and } \omega_{K-1}\neq\omega_K, K\geqslant 2\}$$

(b) The probability of getting 3 tosses can only happen in two ways (H,H,T) and (T,T,H). So:

$$P(K=3)=2\cdot\frac{1}{2^3}=\frac{1}{4}$$

(it would be more complicated if it were e.g. $K\leqslant 3$).

1.14
Claim: If $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$, then A is independent of every other event.

Proof. First consider the case $\mathbb{P}(A) = 0$. Obviously, for any $E \in \Omega$:

$$\mathbb{P}(A)\mathbb{P}(E) = 0 \cdot \mathbb{P}(E) = 0.$$

By the axioms of probability: $A \cap E \subset A$ so $\mathbb{P}(A \cap E) \leqslant \mathbb{P}(A) = 0$ and since for any event: $0 \leqslant \mathbb{P}(A \cap E)$ we can conclude that $\mathbb{P}(A \cap E) = 0$. So:

$$\mathbb{P}(A \cap E) = 0 = \mathbb{P}(A)\mathbb{P}(E) \implies \mathbb{P}(A \cap E) = \mathbb{P}(A)\mathbb{P}(E)$$

which proves independence.

Now consider $\mathbb{P}(A) = 1$. Then, for any event $E \in \Omega$:

$$\mathbb{P}(A)\mathbb{P}(E) = 1 \cdot \mathbb{P}(E) = \mathbb{P}(E).$$

Since $A \subset A \cup E \subset \Omega$, then $\mathbb{P}(A) \leqslant \mathbb{P}(A \cup E) \leqslant \mathbb{P}(\Omega)$. We know the probabilities of A and Ω, so $1 \leqslant \mathbb{P}(A \cup E) \leqslant 1$ means that $\mathbb{P}(A \cup E) = 1$. By Lemma 1.6:

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(E) - \mathbb{P}(A \cap E) \\
1 &= 1 + \mathbb{P}(E) - \mathbb{P}(A \cap E) \\
\mathbb{P}(A \cap E) &= \mathbb{P}(E)
\end{aligned}$$

In conclusion:

$$\mathbb{P}(A \cap E) = \mathbb{P}(E) = \mathbb{P}(A)\mathbb{P}(E) \implies \mathbb{P}(A \cap E) = \mathbb{P}(A)\mathbb{P}(E)$$

which proves independence. □

Claim: If A is independent of itself, then $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.

Proof. If A is independent of itself, we can write:

$$\begin{aligned}
\mathbb{P}(A \cap A) &= \mathbb{P}(A)\mathbb{P}(A) \\
\mathbb{P}(A) &= \mathbb{P}(A)^2
\end{aligned}$$

since $A \cap A = A$. We can write this as an equation, by defining $p := \mathbb{P}(A)$.

$$\begin{aligned}
p &= p^2 \\
p^2 - p &= 0 \\
p(p-1) &= 0
\end{aligned}$$

This equation is solved when $p = \mathbb{P}(A) = 1$ or $p = \mathbb{P}(A) = 0$. □

1.15
The probability that a child has blue eyes is 1/4, $\mathbb{P}(B) = 1/4$. We assume independence in the eye color of the children.

(a) Given that one child has blue eyes, what is the probability that 2 or more children have blue eyes? Since it's given that one child has blue eyes, we just need to find the probability that one, or both, of the remaining children have blue eyes. Define B_i: as the event that child i has blue eyes, and $\neg B_i$ the event that child i does NOT have blue eyes. We only need to sum up the following three probabilities:

$$\mathbb{P}(B_1 \cap B_2) = \mathbb{P}(B_1)\mathbb{P}(B_2) = \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{16}$$
$$\mathbb{P}(B_1 \cap \neg B_2) = \mathbb{P}(B_1)\mathbb{P}(\neg B_2) = \left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{3}{16}$$
$$\mathbb{P}(\neg B_1 \cap B_2) = \mathbb{P}(\neg B_1)\mathbb{P}(B_2) = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right) = \frac{3}{16}$$

By summing these up, we find that the probability of two or more children having blue eyes is $\frac{7}{16}$.

(b) Which child has blue eyes is irrelevant. The probability will be the same: $\frac{7}{16}$.

1.16
Proving Lemma 1.14. If A and B are independent events, then $\mathbb{P}(A|B) = \mathbb{P}(A)$. Also, for any pair of events A and B,
$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A).$$

Proof. Suppose A and B are independent. By definition of independence, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. By definition of conditional probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A),$$

which proves the first statement. Again, by definition of conditional probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \Longrightarrow \quad \mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$$

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} \quad \Longrightarrow \quad \mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A),$$

since $A \cap B = B \cap A$ which proves the remaining two statements. $\square$

1.17
Show that

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A|B \cap C)\mathbb{P}(B|C)\mathbb{P}(C).$$

Proof. Define $E := B \cap C$, so $A \cap B \cap C = A \cap E$. By Lemma 1.14:

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A \cap E) = \mathbb{P}(A|E)\mathbb{P}(E) = \mathbb{P}(A|B \cap C)\mathbb{P}(B \cap C) \tag{2.17.1}$$

By Lemma 1.14 again:

$$\mathbb{P}(B \cap C) = \mathbb{P}(B|C)\mathbb{P}(C),$$

which we can replace in equation (2.17.1) and get the desired result:

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A|B \cap C)\mathbb{P}(B|C)\mathbb{P}(C)$$ □

1.18
Suppose k events form a partition of Ω, i.e. they are disjoint and $\bigcup_{i=1}^{k} A_i = \Omega$. Assume that $\mathbb{P}(B) > 0$. If $\mathbb{P}(A_1|B) < \mathbb{P}(A_1)$ then $\mathbb{P}(A_i|B) > \mathbb{P}(A_i)$ for some $i = 2, 3, \ldots, k$.

Proof. By summing over all A_i with and without conditioning on B, we get the following two equalities:

$$\sum_{i=1}^{k} \mathbb{P}(A_i|B) = 1, \qquad \sum_{i=1}^{k} \mathbb{P}(A_i) = 1$$

We can set these equal to each other:

$$\sum_{i=1}^{k} \mathbb{P}(A_i|B) = \sum_{i=1}^{k} \mathbb{P}(A_i)$$

Removing the terms for A_1 from the sums:

$$\mathbb{P}(A_1|B) + \sum_{i=2}^{k} \mathbb{P}(A_i|B) = \mathbb{P}(A_1) + \sum_{i=2}^{k} \mathbb{P}(A_i)$$

Since $\mathbb{P}(A_1|B) < \mathbb{P}(A_1)$, then $\mathbb{P}(A_1) - \mathbb{P}(A_1|B) = a > 0$ for some $a \in \mathbb{R}$. By subtracting $\mathbb{P}(A_1|B)$ from both sides, we can write:

$$\sum_{i=2}^{k} \mathbb{P}(A_i|B) = \mathbb{P}(A_1) - \mathbb{P}(A_1|B) + \sum_{i=2}^{k} \mathbb{P}(A_i)$$

$$\sum_{i=2}^{k} \mathbb{P}(A_i|B) = a + \sum_{i=2}^{k} \mathbb{P}(A_i) \ > \ \sum_{i=2}^{k} \mathbb{P}(A_i)$$

So, we get:

$$\sum_{i=2}^{k} \mathbb{P}(A_i|B) \ > \ \sum_{i=2}^{k} \mathbb{P}(A_i)$$

which means there exists some $i \in \{2, 3, \ldots, k\}$ such that $\mathbb{P}(A_i|B) > \mathbb{P}(A_i)$. □

1.19
Define M as Mac users, W as Windows users, and L as Linux users. We have:

$$\mathbb{P}(M) = 0.3, \quad \mathbb{P}(W) = 0.5, \quad \mathbb{P}(L) = 0.2$$

The computers are infected with a virus, which give the following probabilities:

$$\mathbb{P}(V|M) = 0.65, \quad \mathbb{P}(V|W) = 0.82, \quad \mathbb{P}(V|L) = 0.50$$

A person is selected at random and we learn that her computer has the virus. What is the probability that she is using Windows? Or, what is $\mathbb{P}(W|V)$? We calculate this with Bayes Law.

$$\begin{aligned}
\mathbb{P}(W|V) &= \frac{\mathbb{P}(V|W)\mathbb{P}(W)}{\mathbb{P}(V|M)\mathbb{P}(M) + \mathbb{P}(V|W)\mathbb{P}(W) + \mathbb{P}(V|L)\mathbb{P}(L)} \\
&= \frac{(0.82)(0.5)}{(0.65)(0.3) + (0.82)(0.5) + (0.5)(0.2)} \\
&= 0.5815
\end{aligned}$$

1.20
We have a box containing 5 coins, $C_1, \ldots, C_5$. Each coin is selected at random, so $\mathbb{P}(C_i) = 1/5$ for all coins. They have the following probabilities of getting H when flipped ($p_i = \mathbb{P}(H|C_i)$):

$$p_1 = 0, \quad p_2 = 1/4, \quad p_3 = 1/2, \quad p_4 = 3/4, \quad p_5 = 1$$

(a) We flip a coin and get a head. Calculate $\mathbb{P}(C_i|H)$ for all coins. First an intermediary calculation:

$$\begin{aligned}
\sum_{i=1}^{5} \mathbb{P}(H|C_i)\mathbb{P}(C_i) &= \mathbb{P}(H|C_1)\mathbb{P}(C_1) + \mathbb{P}(H|C_2)\mathbb{P}(C_2) + \mathbb{P}(H|C_3)\mathbb{P}(C_3) + \mathbb{P}(H|C_4)\mathbb{P}(C_4) + \mathbb{P}(H|C_5)\mathbb{P}(C_5) \\
&= (0)\left(\frac{1}{5}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{5}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{5}\right) + \left(\frac{3}{4}\right)\left(\frac{1}{5}\right) + (1)\left(\frac{1}{5}\right) \\
&= 0 + \frac{1}{20} + \frac{2}{20} + \frac{3}{20} + \frac{4}{20} \\
&= \frac{1}{2}
\end{aligned}$$

Calculating the posterior for each C_i.

$$\mathbb{P}(C_1|H) = \frac{\mathbb{P}(H|C_1)\mathbb{P}(C_1)}{\sum_{i=1}^{5} \mathbb{P}(H|C_i)\mathbb{P}(C_i)} = \frac{0}{\frac{1}{2}} = 0$$

$$\mathbb{P}(C_2|H) = \frac{\mathbb{P}(H|C_2)\mathbb{P}(C_2)}{\sum_{i=1}^{5} \mathbb{P}(H|C_i)\mathbb{P}(C_i)} = \frac{\frac{1}{20}}{\frac{1}{2}} = 1/10$$

$$\mathbb{P}(C_3|H) = \frac{\mathbb{P}(H|C_3)\mathbb{P}(C_3)}{\sum_{i=1}^{5} \mathbb{P}(H|C_i)\mathbb{P}(C_i)} = \frac{\frac{2}{20}}{\frac{1}{2}} = 2/10$$

$$\mathbb{P}(C_4|H) = \frac{\mathbb{P}(H|C_4)\mathbb{P}(C_4)}{\sum_{i=1}^{5}\mathbb{P}(H|C_i)\mathbb{P}(C_i)} = \frac{\frac{3}{20}}{\frac{1}{2}} = 3/10$$

$$\mathbb{P}(C_5|H) = \frac{\mathbb{P}(H|C_5)\mathbb{P}(C_5)}{\sum_{i=1}^{5}\mathbb{P}(H|C_i)\mathbb{P}(C_i)} = \frac{\frac{4}{20}}{\frac{1}{2}} = 4/10$$

(b) We toss the coin again. Finding the probability $\mathbb{P}(H_2|H_1)$ (H_i: getting H on toss i). We are still using the same random coin, so we must calculate the probability for all coins together.

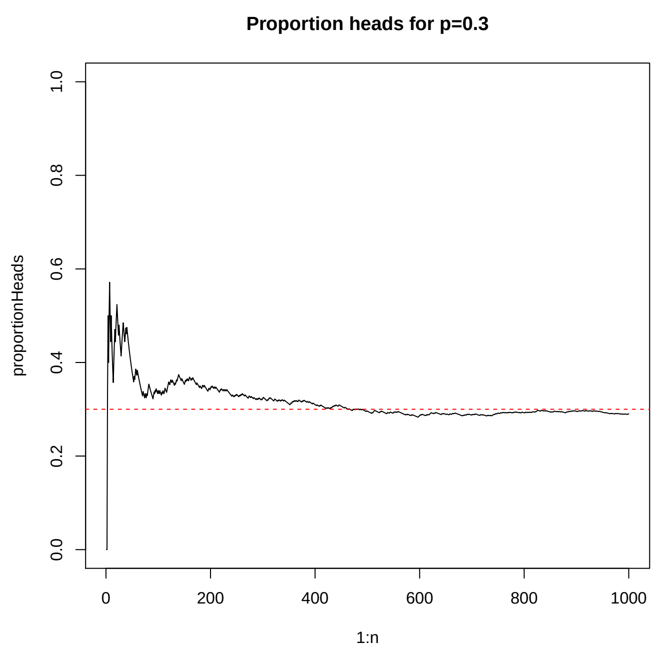
1.21
Proportion of H vs T for $p = 0.3$.

```
# 1.21 - Plotting proportion of H vs T
n = 1000
p = 0.3
coinTosses = sample(c("H","T"), prob = c(p, 1-p), size = n, replace = TRUE)
proportionHeads = rep(0, n)
headCount = 0
for(i in 1:n) {
    if (coinTosses[i] == "H") {
    headCount = headCount + 1
    }
    proportionHeads[i] = headCount/i
}

# PDF
pdf("~/ALLSTAT/ch1_2.21a.pdf")
plot(x = 1:n, y = proportionHeads, type = "l", ylim = c(0,1),
        main = paste0("Proportion heads for p=", p))
abline(h = p, col = "red", lty = "dashed")
dev.off()
```

R

Result
After some initial randomness due to few samples, we can see that the simulated results stabilizes aroud 0.3, as expected.



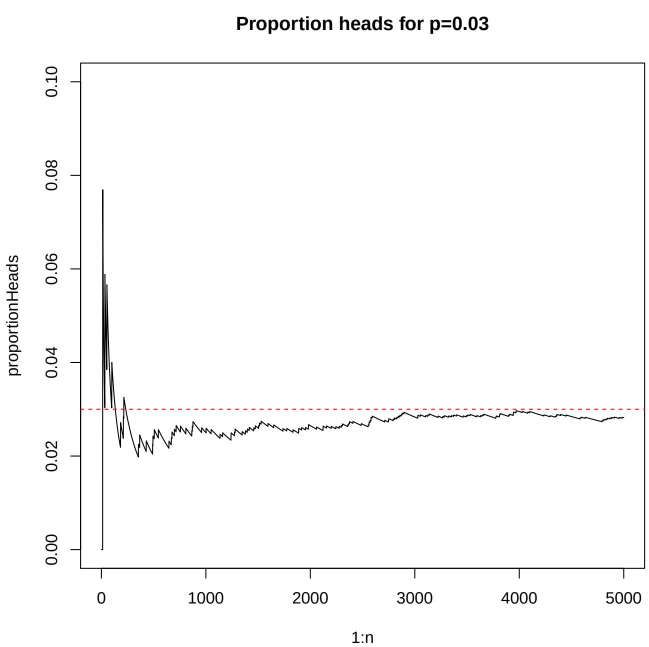
Proportion of H vs T for $p = 0.03$.

```
# 1.21 - Plotting proportion of H vs T
n = 5000
p = 0.03
coinTosses = sample(c("H","T"), prob = c(p, 1-p), size = n, replace = TRUE)
proportionHeads = rep(0, n)
headCount = 0
for(i in 1:n) {
    if (coinTosses[i] == "H") {
    headCount = headCount + 1
    }
    proportionHeads[i] = headCount/i
}

# PDF
pdf("~/ALLSTAT/ch1_2.21b.pdf")
plot(x = 1:n, y = proportionHeads, type = "l", ylim = c(0,1),
        main = paste0("Proportion heads for p=", p))
abline(h = p, col = "red", lty = "dashed")
dev.off()
```

R

Result
After some initial randomness due to few samples, we can see that the simulated results stabilizes aroud 0.03, as expected. Did a simulation of 5000 to make the 'convergence' clearer. Note that the y-axis only goes up to 0.1 in this plot.



1.22

Investigating some properties of Binomial random variables.

```r
# 1.22 - Binomial Random Variables
REP = 10  # Number of simulations per n
p = 0.3   # Probability of H
sim10 = rep(0, REP)
sim100 = rep(0, REP)
sim1000 = rep(0, REP)
# Simulating 10
for(i in 1:REP) {
    # 1 means head
    coinTosses = sample(c(1,0), prob = c(p, 1-p), size = 10, replace = TRUE)
    sim10[i] = sum(coinTosses)
}
# Simulating 100
for(i in 1:REP) {
    # 1 means head
    coinTosses = sample(c(1,0), prob = c(p, 1-p), size = 100, replace = TRUE)
    sim100[i] = sum(coinTosses)
}
# Simulating 1000
for(i in 1:REP) {
    # 1 means head
    coinTosses = sample(c(1,0), prob = c(p, 1-p), size = 1000, replace = TRUE)
    sim1000[i] = sum(coinTosses)
}
df = data.frame(
    SIM10 = sim10,
    SIM100 = sim100,
    SIM1000 = sim1000
)
# Output
df
apply(df, 2, mean)
```

R

As seen in the results, the mean of the 10 simulations is close to np which would be 3, 30 and 300.

```
> df
    SIM10 SIM100 SIM1000
1      1     32     282
2      2     30     298
3      3     33     319
4      3     28     290
5      3     29     284
6      2     34     286
7      4     37     310
8      2     30     294
9      2     30     297
10     2     31     326
> apply(df, 2, mean)
    SIM10  SIM100 SIM1000
    2.4    31.4   298.6
```

1.23

Simulating conditional probabilities. First we simulate an independent experiment.

```
# 1.23 - Simulating a fair die
options(digits=8)
numberOfTosses = 10000
A = c(2, 4, 6)
B = c(1, 2, 3, 4)
AandB = intersect(A, B) # c(2, 4)

dieTosses = sample(1:6, size = numberOfTosses, replace = TRUE)

# Calculating P(A), P(B), P(A)*P(B) and P(A cap B)
PA = sum(dieTosses %in% A)/numberOfTosses
PB = sum(dieTosses %in% B)/numberOfTosses
PAandB = sum(dieTosses %in% AandB)/numberOfTosses

# Output
PA
PB
PA*PB
PAandB
```

R

Results from the calculation. As we can see, the estimates for $\mathbb{P}(A) \approx 1/2$ and $\mathbb{P}(B) \approx 2/3$. Also $\mathbb{P}(A \cap B) \approx \mathbb{P}(A)\mathbb{P}(B) \approx 1/3$. Differences are probably due to rounding errors.

```
> # Output
> PA
[1] 0.5005
> PB
[1] 0.6695
> PA*PB
[1] 0.33508475
> PAandB
[1] 0.3344
```

Now we will construct an experiment with a conditional probability. We will use a fair coin and a die. When we get heads, this will correspond to 1 and the die is unchanged. If we get tails, this will be 2 and will double the die count; so if we get tails and we roll a 2, this will give us a 4.

	1	2	3	4	5	6	7	8	9	10	11	12
H	1/12	1/12	1/12	1/12	1/12	1/12						
T		1/12		1/12		1/12		1/12		1/12		1/12

Define the events: $A = \{2, 3, 4, 5, 6\}$ and $B = \{2, 4, 6, 8\}$ which will give $A \cap B = \{2, 4, 6\}$. The theoretical probabilities are:

$$\mathbb{P}(A) = 8/12 = 2/3 = 0.666$$
$$\mathbb{P}(B) = 7/12 \approx 0.5833$$
$$\mathbb{P}(A)\mathbb{P}(B) = 7/18 \approx 0.3888$$
$$\mathbb{P}(A \cap B) = 1/2 = 0.5$$

Simulating conditional probabilities.

```
# 1.23 - Simulating a conditional probability
options(digits=8)
numberOfTosses = 10000
A = c(2, 3, 4, 5, 6)
B = c(2, 4, 6, 8)
AandB = intersect(A, B) # c(2, 4, 6)

dieTosses = sample(1:6, size = numberOfTosses, replace = TRUE)
coinTosses = sample(1:2, size = numberOfTosses, replace = TRUE)
jointToss = dieTosses*coinTosses

# Calculating P(A), P(B), P(A)*P(B) and P(A cap B)
PA = sum(jointToss %in% A)/numberOfTosses
PB = sum(jointToss %in% B)/numberOfTosses
PAandB = sum(jointToss %in% AandB)/numberOfTosses

# Output
PA
PB
PA*PB
PAandB
```

R

```
> # Output
> PA
[1] 0.6691
> PB
[1] 0.5825
> PA*PB
[1] 0.38975075
> PAandB
[1] 0.4995
```

Repeating the theoretical values from previous page:

$$\begin{aligned}
\mathbb{P}(A) &= 8/12 = 2/3 = 0.666 \\
\mathbb{P}(B) &= 7/12 \approx 0.5833 \\
\mathbb{P}(A)\mathbb{P}(B) &= 7/18 \approx 0.3888 \\
\mathbb{P}(A \cap B) &= 1/2 = 0.5
\end{aligned}$$

The simulated results are very close to the theoretical calculations. We can see that we do not have independence, as expected.

2 Random Variables

Exercises

2.1
Claim: $\mathbb{P}(X = x) = F(x^+) - F(x^-)$. (Discrete)
PROOF. By definition of the CDF:

$$F(x^+) = \lim_{z \downarrow x} F(z) = \lim_{z \downarrow x} \mathbb{P}(X \leqslant z), \qquad F(x^-) = \lim_{y \uparrow x} F(y) = \lim_{y \uparrow x} \mathbb{P}(X \leqslant y)$$

(so $y < x$ and $y \to x$, and $x < z$ and $x \leftarrow z$). By the right continuous property, we can deduce that $z > y$ and we can set $z = x$ and $y = x - 1$.

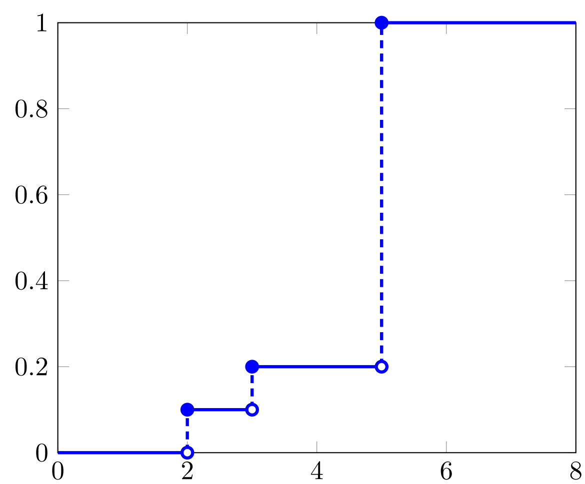
$$\mathbb{P}(X \leqslant x^+) = \mathbb{P}(X \leqslant x) = \mathbb{P}(X = x) + \mathbb{P}(X \leqslant x - 1), \quad \mathbb{P}(X \leqslant x^-) = \mathbb{P}(X \leqslant x - 1)$$

So:

$$\begin{aligned}
\mathbb{P}(X = x) &= \mathbb{P}(X = x) + \mathbb{P}(X \leqslant x - 1) - \mathbb{P}(X \leqslant x - 1) \\
&= \mathbb{P}(X \leqslant x) - \mathbb{P}(X \leqslant x - 1) \\
&= \mathbb{P}(X \leqslant x^+) - \mathbb{P}(X \leqslant x^-) \\
&= F(x^+) - F(x^-)
\end{aligned}$$

□

2.2
Let X be such that $\mathbb{P}(X = 2) = \mathbb{P}(X = 3) = 1/10$ and $\mathbb{P}(X = 5) = 8/10$. Here is a plot of the CDF.



By reading the plot, we can see that:

$$\mathbb{P}(2 < X \leqslant 4.8) = F(4.8) - F(2) = 2/10 - 1/10 = 1/10$$

$$\mathbb{P}(2 \leqslant X \leqslant 4.8) = F(4.8) = 2/10$$

2.3

Lemma 2.15 Let F be the CDF for a random variable X. Then:

1. $\mathbb{P}(X = x) = F(x) - F(x^-)$
2. $\mathbb{P}(x < X \leqslant y) = F(y) - F(x)$
3. $\mathbb{P}(X > x) = 1 - F(x)$
4. If X is continuous, then

$$F(b) - F(a) = \mathbb{P}(a < X < b) = \mathbb{P}(a \leqslant X < b) = \mathbb{P}(a < X \leqslant b) = \mathbb{P}(a \leqslant X \leqslant b)$$

PROOF. We will prove each statement in turn. (1.) was proved in exercise **2.1**. Doing (3) first, since we need it to prove (2).

(3) By definition of complements of sets $A = \{X > x\}$ means $A^c = \{X \leqslant x\}$, and it follows that:

$$\mathbb{P}(X > x) = \mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \mathbb{P}(X \leqslant x) = 1 - F(x).$$

(2) Assume $x < y$. We will need that $\{X > x\} \cup \{X \leqslant y\} = \Omega$, and we will also use Lemma 1.6 (in reverse).

$$\begin{aligned}
\mathbb{P}(x < X \leqslant y) &= \mathbb{P}(\{X > x\} \cap \{X \leqslant y\}) \\
&= \mathbb{P}(X > x) + \mathbb{P}(X \leqslant y) - \mathbb{P}(\{X > x\} \cup \{X \leqslant y\}) \\
&= 1 - F(x) + F(y) - 1 \\
&= F(y) - F(x)
\end{aligned}$$

(4) Similar argument for all cases, so will just do one. We just need to turn the inequalities into strict inequalities. For continuous random variables, pointwise probabilities are 0. Again, we will need to use $\{X > a\} \cup \{X < b\} = \Omega$.

Define $A := \{a \leqslant X\}$ and $B := \{X < b\}$. First, we make the following observation:

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}(\{a \leqslant X\}) \\
&= \mathbb{P}(\{a = X\} \cup \{a < X\}) \\
&= \mathbb{P}(\{a = X\}) + \mathbb{P}(\{a < X\}) - \mathbb{P}(\{a = X\} \cap \{a < X\}) \\
&= 0 + \mathbb{P}(A') - 0 \\
&= \mathbb{P}(A')
\end{aligned}$$

where $A' = \{a < X\}$. We get 0 for the pointwise probability, since this is continuous, and we get 0 because the sets are disjoint. We have shown that $\mathbb{P}(A) = \mathbb{P}(A')$ and can use this to conclude the proof.

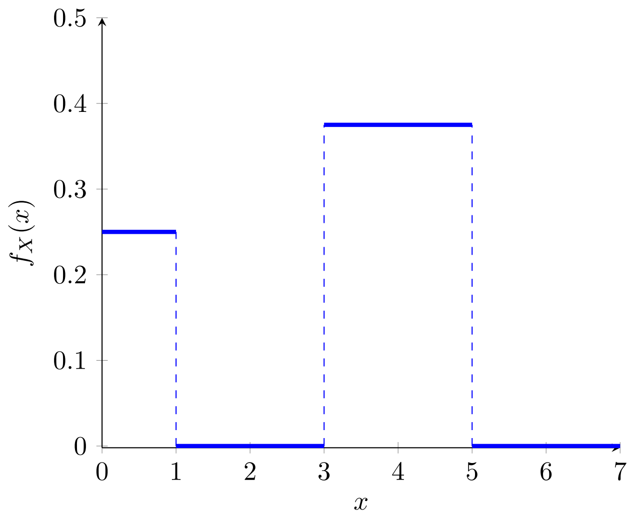
$$\begin{aligned}
\mathbb{P}(a \leqslant X < b) &= \mathbb{P}(\{a \leqslant X\} \cap \{X < b\}) \\
&= \mathbb{P}(A \cap B) \\
&= \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(A \cup B) \\
&= \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(\Omega) \\
&= \mathbb{P}(A') + \mathbb{P}(B) + \mathbb{P}(A' \cup B) \\
&= \mathbb{P}(A' \cap B) \\
&= \mathbb{P}(a < X < b)
\end{aligned}$$

□

2.4
X has the probability density (PDF):

$$f_X(x) = \begin{cases} 1/4 & 0 < x < 1 \\ 3/8 & 3 < x < 5 \\ 0 & \text{otherwise} \end{cases}$$

Plot of the PDF:



From the relatively simple structure, we can easily determine the area under the graph:

$$A = (1)\left(\frac{1}{4}\right) + (2)\left(\frac{3}{8}\right) = \frac{2}{8} + \frac{6}{8} = 1$$

(a) Finding the CDF by integrating the PDF. We will split up the integral in several parts. First for the case when $y \in (0, 1)$:

$$F_X(y) = \int_{-\infty}^{y} f_X(t)dt = \frac{1}{4}\int_0^y 1dt = \frac{1}{4}\Big[t\Big]_0^y = \frac{y}{4}$$

When $y = 1$ we have $F_X(1) = 1/4$. Next, we must consider the case $y \in (1, 3)$. Here the PDF is 0, so it doesn't increase. It remains constant at 1/4 (since the CDF doesn't decrease).

$$F_X(y) = \frac{1}{4}$$

Next is the case $y \in (3, 5)$. Consider the intermediary integral:

$$I_1 = \int_3^y \frac{3}{8}dt = \frac{3}{8}\Big[t\Big]_3^y = \frac{3y - 9}{8}$$

For values $y \in (3, 5)$ we start on 1/4, so the CDF in this region becomes:

$$F_X(y) = \frac{3y - 9}{8} + \frac{1}{4}$$

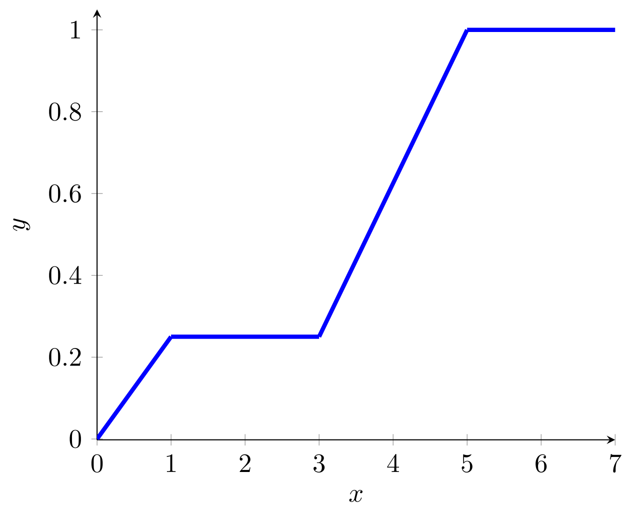
So, the full expression for the CDF becomes:

$$F_X(y) = \begin{cases} y/4 & y \in (0,1) \\ 1/4 & y \in (1,3) \\ \dfrac{3y-9}{8} + \dfrac{1}{4} & y \in (3,5) \\ 1 & y \geqslant 5 \end{cases}$$

Note that when $y = 5$ we get:

$$F_X(5) = \frac{3(5)-9}{8} + \frac{1}{4} = \frac{6}{8} + \frac{2}{8} = 1$$

Plot of the CDF:



(b) Defining $Y = 1/X$ and finding the PDF of Y. Following the hint we are given, we will consider the following three sets:

$$A_1 = \frac{1}{5} \leqslant y \leqslant \frac{1}{3}, \quad A_2 = \frac{1}{3} \leqslant y \leqslant 1, \quad A_3 = y \geqslant 1$$

Where A_1 corresponds to $(3,5)$, A_2 to $(1,3)$ and A_3 to $(0,1)$. We can express the CDF for $F_Y(y)$ in terms of $F_X(x)$:

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leqslant y) = \mathbb{P}(\frac{1}{X} \leqslant y) \\ &= \mathbb{P}(X \geqslant \frac{1}{y}) \\ &= 1 - \mathbb{P}(X \leqslant \frac{1}{y}) \\ &= 1 - F_X(\frac{1}{y}) \end{aligned}$$

First, we consider $A_1 : y \in [1/5, 1/3]$, and when we input $1/y$ to $F_X(\cdot)$, it will be in $(3, 5)$. So:

$$\begin{aligned} F_Y(y) &= 1 - F_X(1/y) \\ &= 1 - \left(\frac{3(\frac{1}{y}) - 9}{8} + \frac{1}{4} \right) \\ &= 1 - \frac{3 - 9y}{8y} - \frac{1}{4} \\ &= \frac{3}{4} + \frac{9y - 3}{8y} \\ &= \frac{15y - 3}{8y} \end{aligned}$$

Next, we consider $A_2 : y \in [1/3, 1]$. The input to $F_X(\cdot)$ will be in $(1, 3)$:

$$\begin{aligned} F_Y(y) &= 1 - F_X(1/y) \\ &= 1 - \frac{1}{4} \\ &= \frac{3}{4} \end{aligned}$$

Next, we consider $A_3 : y \geqslant 1$. The input to $F_X(\cdot)$ will be in $(0, 1)$:

$$\begin{aligned} F_Y(y) &= 1 - F_X(1/y) \\ &= 1 - \frac{\frac{1}{y}}{4} \\ &= 1 - \frac{1}{4y} \end{aligned}$$

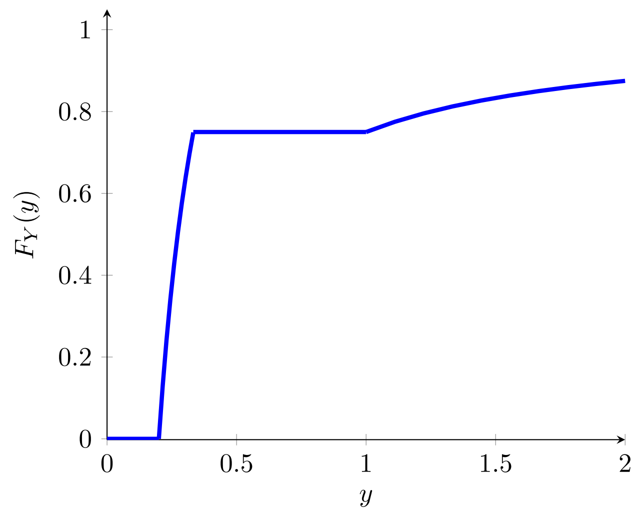
Also, whenever $y < 1/5$, then $1/y > 5$ which means $F_X(\cdot) = 1$, and so:

$$F_Y(y) = 1 - F_X(1/y) = 1 - 1 = 0.$$

This gives a full description of the CDF for $F_Y(y)$.

$$F_Y(y) = \begin{cases} 0 & y < 1/5 \\ \dfrac{15y - 3}{8y} & 1/5 \leqslant y \leqslant 1/3 \\ \dfrac{3}{4} & 1/3 \leqslant y \leqslant 1 \\ 1 - \dfrac{1}{4y} & y \geqslant 1 \end{cases}$$

Plot of CDF:



Finally, we can find the PDF of Y. We differentiate each of the parts in the CDF. When $y \in (1/5, 1/3)$:

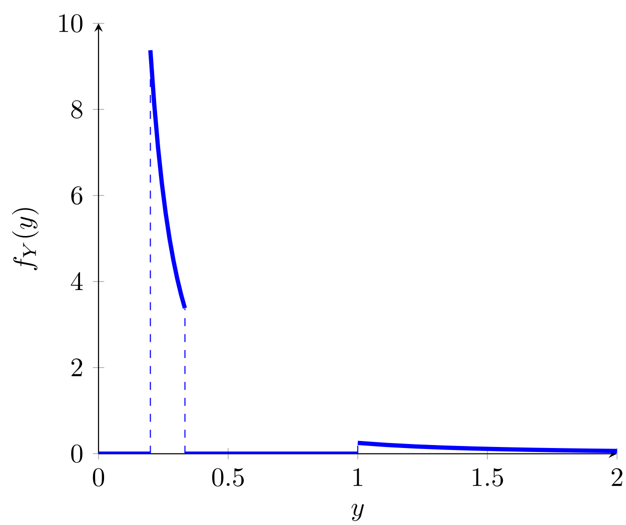
$$\frac{d}{dy}\left(\frac{15y-3}{8y}\right) = \frac{3}{8y^2}$$

When $y \geqslant 1$:

$$\frac{d}{dy}\left(1-\frac{1}{4y}\right) = \frac{1}{4y^2}$$

(All other parts are constant, so they become 0). This gives us the PDF and its plot:

$$f_Y(y) = \begin{cases} 0 & y < 1/5 \\ \dfrac{3}{8y^2} & 1/5 \leqslant y \leqslant 1/3 \\ 0 & 1/3 < y < 1 \\ \dfrac{1}{4y^2} & y \geqslant 1 \end{cases}$$



2.5
Let X and Y be discrete RV. X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x and y.
PROOF.
$\Rightarrow$) Assume that X and Y are independent. That means that for any x, y, we have

$$\mathbb{P}(X = x \cap Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

Starting with the definition of the joint pdf:

$$\begin{aligned} f_{X,Y}(x,y) &= \mathbb{P}(X = x, Y = y) \\ &= \mathbb{P}(X = x \cap Y = y) \\ &= \mathbb{P}(X = x)\mathbb{P}(Y = y) \\ &= f_X(x)f_Y(y) \end{aligned}$$

Which shows that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x and y.

$\Leftarrow$) Assume that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x and y. By definition:

$$\begin{aligned} f_{X,Y}(x,y) &= \mathbb{P}(X = x, Y = y) \\ &= \mathbb{P}(X = x \cap Y = y) \end{aligned}$$

And,

$$f_X(x)f_Y(y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

From our assumption, these are equal, so $\mathbb{P}(X = x \cap Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$ which shows that X and Y are independent.
By implication both ways, the statement is proved. $\square$

2.6
Let X have distribution F and density f, and let A be a subset of the real line, e.g. $A = (a, b)$ for some $a, b \in \mathbb{R}$ and $a < b$. We have the indicator function

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

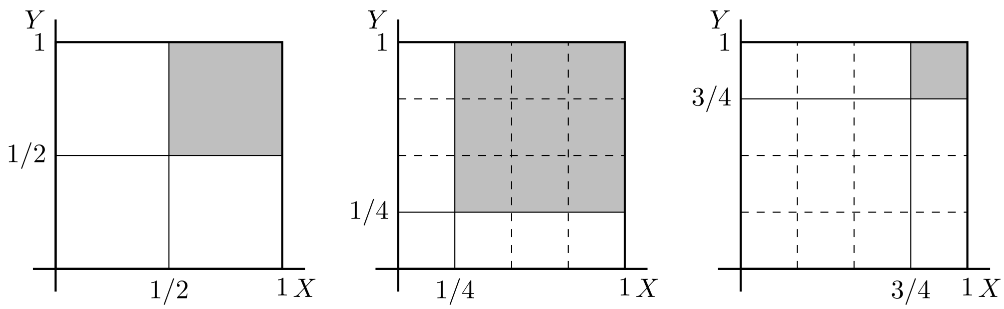
We will set $Y = I_A(X)$ and find the PDF and CDF of Y.

The exercise asks for a probability mass function, but that cannot be correct. Since X has a density f, it is a continuous RV. If $X \sim U(0,1)$ and $A = (0,1)$, then $Y = X$ and it will be a uniform variable with a continuous distribution, so not necessarily discrete.

And if it can be a continuous distribution, what happens if we define $A = \mathbb{Q} \subset \mathbb{R}$? There will be an infinite number of points in any interval with measure 0. Then we cannot define a PDF at all... Poorly formulated exercise in my opinion! Need to fill up with extra assumptions?

Skipping for now.

2.7
Let X and Y be independent and suppose that $X, Y \sim U(0,1)$. For $Z = \min(X, Y)$ we will find the density $f_Z(z)$ for Z. Following the hint, we will first find $\mathbb{P}(Z > z)$. Since any observations of $x, y \in (0,1)$, then we can immediately see that $\mathbb{P}(Z > 0) = 1$ and $\mathbb{P}(Z > 1) = 0$. But what happens for other values? Best way to find out is with some illustrations. Here are plots of the cases $\mathbb{P}(Z > 1/2)$, $\mathbb{P}(Z > 1/4)$ and $\mathbb{P}(Z > 3/4)$.



If we simulate lots of X and Y values, we see that about 1/4th of them will have both X and Y values larger than 1/2, so $\mathbb{P}(Z > 1/2) = 1/4$. Similarly, we get $\mathbb{P}(Z > 1/4) = 9/16$ and $\mathbb{P}(Z > 3/4) = 1/16$. Confirming this with a simulation.

```r
# 2.7 - Simulating U(0,1)
N = 100000;X = runif(N);Y = runif(N)

Z = pmin(X, Y) # This is: Z = min{X, Y}

# Comparing simulated vs. theoretical results
sum(Z > 0.5)/N
1/4
sum(Z > 0.25)/N
9/16
sum(Z > 0.75)/N
1/16
```

```
> # Comparing simulated vs. theoretical results
> sum(Z > 0.5)/N
[1] 0.24888
> 1/4
[1] 0.25

> sum(Z > 0.25)/N
[1] 0.56156
> 9/16
[1] 0.5625

> sum(Z > 0.75)/N
[1] 0.06205
> 1/16
[1] 0.0625
```

By inspecting the images on the previous page, we can determine the 'shape' of the probabilities. For $Z > 1/4$ we remove the union of $X \leqslant 1/4$ and $Y \leqslant 1/4$. We define $A = \{X \leqslant z\}$ and $B = \{Y \leqslant z\}$, and can write the general case as:

$$\begin{aligned}\mathbb{P}(Z > z) &= 1 - \mathbb{P}(\{X \leqslant z\} \cup \{Y \leqslant z\})\\ &= 1 - \mathbb{P}(A \cup B)\\ &= 1 - \big[\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\big]\\ &= 1 - \big[\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B)\big]\end{aligned}$$

Where we used Lemma 1.6, and the fact that X and Y are independent. By using the probability law of complements, we can find the expression for $\mathbb{P}(Z \leqslant z)$.

$$\mathbb{P}(Z \leqslant z) = \mathbb{P}(X \leqslant z) + \mathbb{P}(Y \leqslant z) - \mathbb{P}(X \leqslant z)\mathbb{P}(Y \leqslant z)$$

The CDF for a uniform distribution on $U(a, b)$ is:

$$F(z) = \frac{z-a}{b-a} \quad \Longrightarrow \quad F_X(z) = F_Y(z) = \frac{z-0}{1-0} = z$$

Which means:

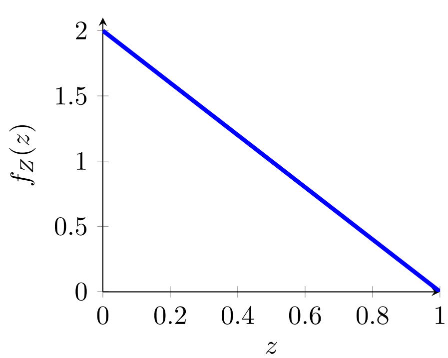
$$F_Z(z) = F_X(z) + F_Y(z) - F_X(z)F_Y(z) = 2z - z^2$$

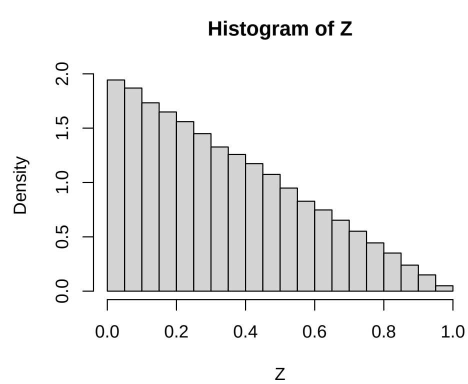
We can confirm our illustrations and simulated examples again by noting that:

$$\begin{aligned}F_Z(1/2) &= \frac{1}{2} + \frac{1}{2} - \left(\frac{1}{2} \cdot \frac{1}{2}\right) = 1 - \frac{1}{4} = \frac{3}{4}\\ F_Z(1/4) &= \frac{1}{4} + \frac{1}{4} - \left(\frac{1}{4} \cdot \frac{1}{4}\right) = \frac{8}{16} - \frac{1}{16} = \frac{7}{16}\\ F_Z(3/4) &= \frac{3}{4} + \frac{3}{4} - \left(\frac{3}{4} \cdot \frac{3}{4}\right) = \frac{24}{16} - \frac{9}{16} = \frac{15}{16}\end{aligned}$$

which gives us the opposite results as expected (since we simulated and illustrated $\mathbb{P}(Z > z)$). The PDF $f_Z(z)$ is the derivative of $F_Z(z)$. Including PDF-plot and histogram of the simulated Zs.

$$f_Z(z) = \frac{d}{dz}\Big(F_Z(z)\Big) = 2 - 2z.$$





2.8
The RV X has CDF F. Finding the CDF of $X^+ = \max\{0, X\}$.

From the definition of CDF:

$$\begin{aligned} F_{X^+}(u) &= \mathbb{P}\big(\max(0,X) \leqslant u\big) \\ &= \mathbb{P}\Big(\{\omega \in \Omega \ : \ X(\omega) \leqslant u \text{ and } u \geqslant 0\}\Big) \end{aligned}$$

We must consider two cases. When $u < 0$:

$$\begin{aligned} F_{X^+}(u) &= \mathbb{P}\Big(\{\omega \in \Omega \ : \ X(\omega) \leqslant u \text{ and } u \geqslant 0\}\Big) \\ &= \mathbb{P}(\emptyset) \\ &= 0 \end{aligned}$$

When $u \geqslant 0$:

$$\begin{aligned} F_{X^+}(u) &= \mathbb{P}\Big(\{\omega \in \Omega \ : \ X(\omega) \leqslant u \text{ and } u \geqslant 0\}\Big) \\ &= \mathbb{P}\Big(\{\omega \in \Omega \ : \ X(\omega) \leqslant u\}\Big) \\ &= \mathbb{P}(X \leqslant u) \\ &= F_X(u) \end{aligned}$$

So in summary:

$$F_{X^+}(u) = \begin{cases} 0 & u < 0 \\ F_X(u) & u \geqslant 0 \end{cases}$$

2.9
We have $X \sim \text{Exp}(\beta)$. The PDF is given by:

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0$$

Finding the CDF by integrating the PDF.

$$\begin{aligned} F(y) &= \int_{-\infty}^{y} f(x) dx \\ &= \frac{1}{\beta} \int_0^y e^{-x/\beta} dx \\ &= \frac{1}{\beta} \big[-\beta e^{-x/\beta} \big]_0^y \\ &= \big[-e^{-x/\beta} \big]_0^y \\ &= 1 - e^{-y/\beta} \end{aligned}$$

To find the inverse $F^{-1}(q)$ we set $q = F(y)$ and solve for y.

$$q = 1 - e^{-y/\beta} \quad \implies \quad y = -\beta \log(1-q) \quad \implies \quad F^{-1}(q) = -\beta \log(1-q)$$

2.10
If X and Y are independent, then $g(X)$ and $h(Y)$ are independent for some functions g and h.

PROOF. Let X and Y be some arbitrary random variables, and let x and y be values in the range of g and h such that $g(X) = x$ and $h(Y) = y$. Then:

$$\mathbb{P}(g(X) = x, h(Y) = y) = \mathbb{P}(X = g^{-1}(x), Y = h^{-1}(y))$$

By independence of X and Y.

$$= \mathbb{P}(X = g^{-1}(x))\mathbb{P}(Y = h^{-1}(y))$$
$$= \mathbb{P}(g(X) = x)\mathbb{P}(h(Y) = y)$$

which shows that g and h satisfies the condition for independence. $\square$

2.11
Tossing a coin which has probability p of getting H. We let X denote the number of heads and Y the number of tails.

(a) Showing that X and Y are dependent. First it will be helpful to consider a simplified case where we have $N = 1$ and $N = 2$ coin tosses, and assuming $p = 1/2$.

$N = 1$ Tosses

	$Y=0$	$Y=1$	
$X=0$	0	1/2	1/2
$X=1$	1/2	0	1/2
	1/2	1/2	1

$N = 2$ Tosses

	$Y=0$	$Y=1$	$Y=2$	
$X=0$	0	0	1/4	1/4
$X=1$	0	1/2	0	1/2
$X=2$	1/4	0	0	1/4
	1/4	1/2	1/4	1

In the case of $N = 2$, we see that $f(0, 2) = 1/4$ while $f_X(0)f_Y(2) = 1/16$. This will be the inspiration for how we show it in the general case with N tosses and probability p of heads.

We only need to show one specific case where $\mathbb{P}(X = x, Y = y) \neq \mathbb{P}(X = x)\mathbb{P}(Y = y)$ to show that these values are dependent. The total number of tosses will be $N = X + Y$ and we compare the cases where we get N heads. In that case, using the Multinomial distribution:

$$\mathbb{P}(X = N, Y = 0) = \binom{N}{N, 0} = p^N(1-p)^0 = p^N,$$

but with the two Binomial distributions:

$$\mathbb{P}(X = N)\mathbb{P}(Y = 0) = \binom{N}{N}p^N(1-p)^0 \times \binom{N}{0}p^0(1-p)^N = p^N(1-p)^N.$$

These are not equal, showing that X and Y are dependent variables.

(b) Now we have $N \sim \text{Poisson}(\lambda)$, where $N = X + Y$ for X heads and Y tails. Show that these values are now independent.

Current solution, but not sure it's correct...
Assuming $X = x$ and $Y = y$. Then:

$$\begin{aligned}
\mathbb{P}(X = x \cap Y = y) &= \mathbb{P}(X = x \cap Y = y \cap N = x + y) \\
&= \mathbb{P}(X = x \cap Y = y \mid N = x + y) \cdot \mathbb{P}(N = x + y) \\
&= \binom{x+y}{x} p^x (1-p)^y \cdot e^{-\lambda} \frac{\lambda^{x+y}}{(x+y)!} \\
&= \frac{\cancel{(x+y)!}}{x!y!} p^x (1-p)^y \cdot e^{-\lambda} \frac{\lambda^x \lambda^y}{\cancel{(x+y)!}} \\
&= e^{-\lambda} \frac{p^x \lambda^x}{x!} \cdot \frac{(1-p)^y \lambda^y}{y!} \\
&= e^{-\lambda p} \frac{p^x \lambda^x}{x!} \cdot e^{-\lambda(1-p)} \frac{(1-p)^y \lambda^y}{y!} \\
&= \mathbb{P}(X = x|\lambda p) \cdot \mathbb{P}(Y = y|\lambda(1-p)) \\
&= \mathbb{P}(X = x)\mathbb{P}(Y = y)
\end{aligned}$$

Which shows we have independence. We used:

$$e^{-\lambda} = e^{-\lambda(p-p+1)} = e^{-\lambda p + \lambda p - \lambda} = e^{-\lambda p} e^{\lambda p - \lambda} = e^{-\lambda p} e^{-\lambda(1-p)}.$$

Also, we made the assumption that $\mathbb{P}(X = x \cap Y = y \mid N = x + y)$ is Binomial. But I think this is only true when we can assume that X and Y are independent... which is what we are trying to show. Will review this later... hopefully!

2.12 Theorem 2.33
Suppose that the range of X and Y is a (possibly infinite) rectangle. If $f(x, y) = g(x)h(y)$ for some functions g and h (not necessarily probability density functions) then X and Y are independent.

Proof. From the joint PDF, we can find the marginal distributions:

$$f_X(x) = \int f(x, y)dy, \quad f_Y(y) = \int f(x, y)dx$$

By applying these integrals to both sides of the equality:

$$\int f(x, y)dy = g(x) \int h(y)dy \implies f_X(x) = g(x)$$

$$\int f(x, y)dx = h(y) \int g(x)dx \implies f_Y(y) = h(y)$$

We get $g(x)$ and $h(y)$ since they are equal to $f(x, y)$ over the entire 'rectangle' and must therefore integrate to 1. This leaves us with: $f(x, y) = g(x)h(y) = f_X(x)f_Y(y)$, and by the results in exercise 2.5, this means that X and Y are independent. $\square$

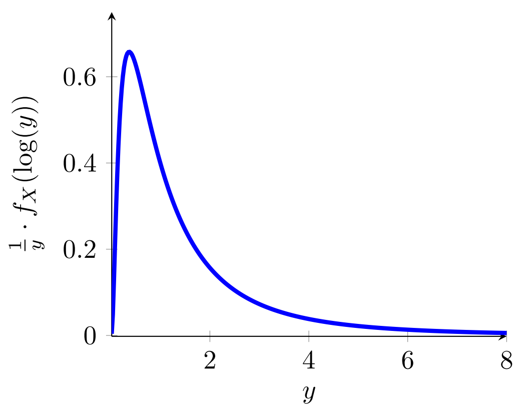
2.13
(a) Finding the PDF of $Y = e^X$ when $X \sim N(0,1)$.

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leqslant y) = \mathbb{P}(e^X \leqslant y) \\ &= \mathbb{P}(X \leqslant \log(y)) = F_X(\log(y)) \end{aligned}$$

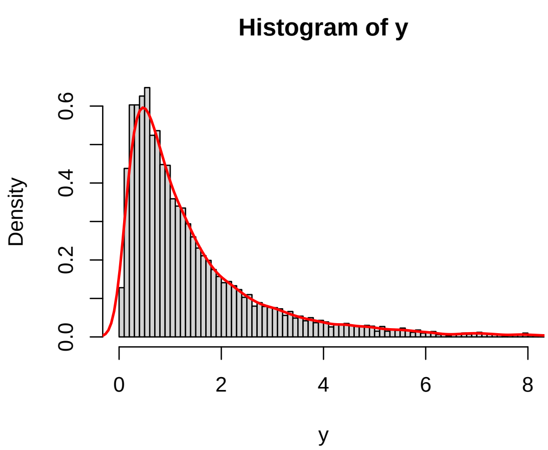
We simply get the standard normal distribution with $\log(y)$ as input. Differentiating to get the PDF:

$$f_Y(y) = \frac{d}{dy} F_X(\log(y)) = f_X(\log(y)) \cdot \frac{1}{y} = \frac{1}{y\sqrt{2\pi}} \exp\left(-\frac{1}{2}\big(\log(y)\big)^2\right)$$

Plotting the function:



(b) Plotting histogram of simulated results. Comparable to the plot above.




```r
x = rnorm(10000)
y = exp(x)
pdf("~/AllStatistics/files/ch2_2.13b.pdf",
      width = 4.7747, height = 4)
d = density(y)
hist(y, breaks = 200, xlim = c(0, 8),
      prob=TRUE)
lines(density(y), col="red", xlim=c(0,8))
dev.off()
```

2.14
We let (X,Y) be uniformly distributed on the unit disk: $\{(x,y) : x^2+y^2 \leqslant 1\}$. Find the CDF and PDF of $R = \sqrt{X^2+Y^2}$.

First, let's write some code in order to simulate the data, which are random points contained within the unit circle. And then use that to simulate R and view the histogram. Plots and code can be found on the next page. By inspection of the histogram it is clear that $f_R(r) = 2r$ which means $F_R(r) = r^2$.

Following the general recipe for transformation of multiple random variables. We will define the function $r(X,Y) = \sqrt{X^2+Y^2}$.

$$F_R(r) = \mathbb{P}(R \leqslant r) = \mathbb{P}(r(X,Y) \leqslant r) = \mathbb{P}(X^2+Y^2 \leqslant r^2) = \int\int_{A_r} f_{X,Y}(x,y)dxdy$$

Finding the set A_r which in this case is still the unit circle, since $\sqrt{x^2+y^2} \leqslant 1 \implies x^2+y^2 \leqslant 1$.
Since the area of the unit circle is π, and since a uniform distribution means the probability is equal everywhere, it follows that:

$$f_{X,Y}(x,y) = \frac{1}{\pi}.$$

As we saw, we must integrate this function over the circle: $\{(x,y) : x^2+y^2 < r\}$ for some $0 \leqslant r \leqslant 1$. This is easiest when calculating the integral in polar coordinates, so we introduce $0 \leqslant \theta \leqslant 2\pi$ and $0 \leqslant \rho \leqslant r$ and change $dxdy = \rho d\rho d\theta$. So:

$$\begin{aligned}
F_R(r) &= \frac{1}{\pi}\int_0^{2\pi}\int_0^r \rho d\rho d\theta \\
&= \frac{1}{\pi}\int_0^{2\pi}\left[\frac{\rho^2}{2}\right]_0^r d\theta \\
&= \frac{1}{\pi}\int_0^{2\pi}\frac{r^2}{2}d\theta \\
&= \frac{r^2}{2\pi}\int_0^{2\pi}(1)d\theta \\
&= \frac{r^2}{2\pi}\left[\theta\right]_0^{2\pi} \\
&= \frac{r^2}{2\pi}\cdot 2\pi \\
&= r^2
\end{aligned}$$

The CDF is $F_R(r) = r^2$ for $0 \leqslant r \leqslant 1$. By differentiating this, we get the PDF.

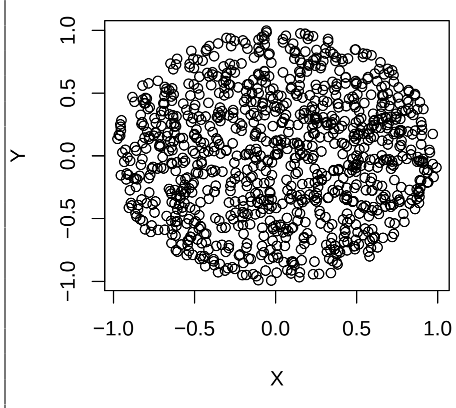
$$f_R(r) = \frac{d}{dr}F_R(r) = 2r,$$

also in the interval $0 \leqslant r \leqslant 1$. This confirms the results we found by simulating which are found on the next page.

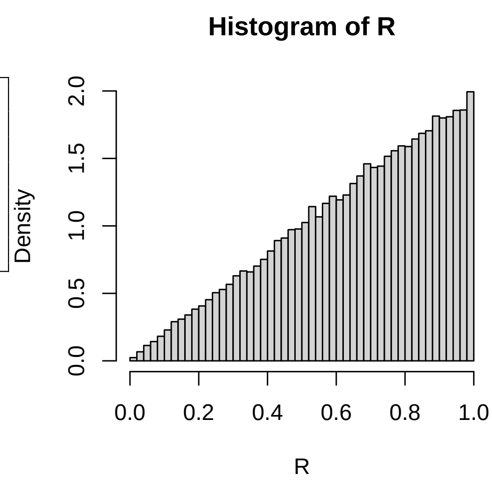
Plots and Code for 2.14

```
simulateUnitCircle = function(N) {
    TR = runif(N)
    UR = runif(N); VR = runif(N)
    t = 2*pi*TR
    u = UR+VR
    u[u > 1] = 2 - u[u > 1]
    r = u
    X = r*cos(t)
    Y = r*sin(t)
    retVal = list()
    retVal$X = X
    retVal$Y = Y
    return(retVal)
}
simVal = simulateUnitCircle(1000)
X = simVal$X
Y = simVal$Y
pdf("~/AllStatistics/files/ch2_2.14.pdf",
    width = 4, height = 4)
plot(X, Y)
dev.off()
```

Result from simulation:




```
simVal = simulateUnitCircle(50000)
R = sqrt(simVal$X^2 + simVal$Y^2)
pdf("~/AllStatistics/files/ch2_2.14b.pdf",
    width = 4, height = 4)
hist(R, breaks = 40, prob = TRUE)
plot(X, Y)
dev.off()
```



2.15

Let X have a continuous, strictly increasing CDF F, and set $Y = F(X)$. Find the density of Y.

By definition of the CDF, $F : \mathbb{R} \to [0,1]$ which will be the domain of Y. Since F is strictly increasing, then $F(X_1) < F(X_2)$ if $X_1 < X_2$, but the important property is that F must be *invertible*. For any $y \in [0,1]$ we get the following:

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leqslant y) &= \mathbb{P}(F(X) \leqslant y) \\ &= \mathbb{P}(X \leqslant F^{-1}(y)) = F(F^{-1}(y)) = y \end{aligned}$$

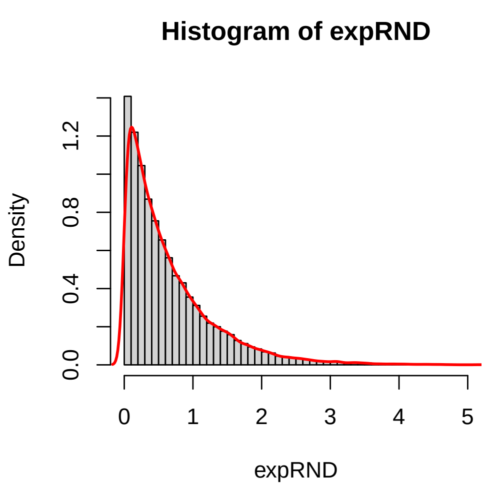
Since $F_Y(y) = y$, then the density is $f_Y(y) = 1$, the derivative. From these observations, we can conclude that $Y \sim U(0,1)$.

Next, if $U \sim U(0,1)$ and $X = F^{-1}(U)$, then X has F as its CDF.

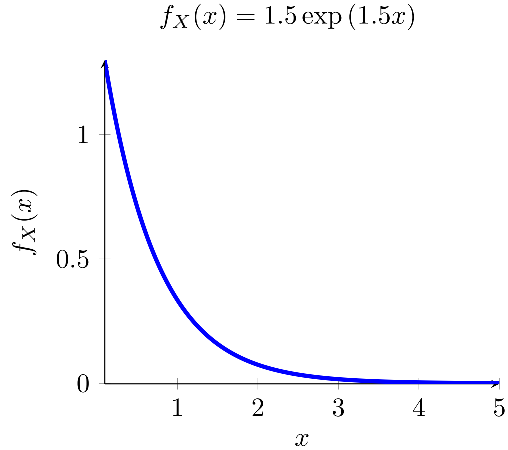
$$\mathbb{P}(F^{-1}(U) \leqslant x) = \mathbb{P}(X \leqslant x) = F(x) \quad \Longrightarrow \quad X \sim F$$

Writing a program that generates $\text{Exp}(\beta)$ values. Using the result found in exercise 2.9.

```
simExp = function(beta, N){
    U = runif(N)
    return(-beta*log(1 - U))
}
# Simulating
expRND = simExp(beta = 2/3, N = 50000)
# Making plot
hist(expRND, breaks=80,
    xlim=c(0, 5), prob=TRUE)
lines(density(expRND), col="red", xlim=c
    (0,5), lwd=2)
```



Plot of the exponential distribution for $\beta = 2/3$.



2.16
If $X \sim \text{Poisson}(\lambda_1)$ and $X \sim \text{Poisson}(\lambda_2)$, and X and Y are independent, then the distribution of X given that $X+Y=n$ is $\text{Binomial}(n,\rho)$ where $\rho = \lambda_1/(\lambda_1+\lambda_2)$. (Changed some variable names).

Proof. By definition of conditional probability, by applying hint 1, by independence and by using the fact that $X+Y$ is a Poisson distribution with $\lambda_1+\lambda_2$.

$$\begin{aligned}\mathbb{P}(X=x|X+Y=n) &= \frac{\mathbb{P}(X=x \cap X+Y=n)}{\mathbb{P}(X+Y=n)} \\ &= \frac{\mathbb{P}(X=x \cap Y=n-x)}{\mathbb{P}(X+Y=n)} && \text{(Hint)} \\ &= \frac{\mathbb{P}(X=x)\mathbb{P}(Y=n-x)}{\mathbb{P}(X+Y=n)} && \text{(Independence)}\end{aligned}$$

Each of these terms are known:

$$\begin{aligned}\mathbb{P}(X=x) &= e^{-\lambda_1}\cdot\frac{\lambda_1^x}{x!} \\ \mathbb{P}(Y=n-x) &= e^{-\lambda_2}\cdot\frac{\lambda_2^{(n-x)})}{(n-x)!} \\ \mathbb{P}(X+Y=n) &= e^{-(\lambda_1+\lambda_2)}\cdot\frac{(\lambda_1+\lambda_2)^n}{n!}\end{aligned}$$

We can invert the last equation:

$$\frac{1}{\mathbb{P}(X+Y=n)} = e^{\lambda_1+\lambda_2}\cdot\frac{n!}{(\lambda_1+\lambda_2)^n}$$

Putting it all together, we can complete the justification.

$$\begin{aligned}\mathbb{P}(X=x|X+Y=n) &= \frac{\mathbb{P}(X=x)\mathbb{P}(Y=n-x)}{\mathbb{P}(X+Y=n)} \\ &= \left(e^{-\lambda_1}\cdot\frac{\lambda_1^x}{x!}\right)\left(e^{-\lambda_2}\cdot\frac{\lambda_2^{(n-x)})}{(n-x)!}\right)\left(e^{\lambda_1+\lambda_2}\cdot\frac{n!}{(\lambda_1+\lambda_2)^n}\right) \\ &= \left(\frac{\lambda_1^x}{x!}\right)\left(\frac{\lambda_2^{(n-x)})}{(n-x)!}\right)\left(\frac{n!}{(\lambda_1+\lambda_2)^x(\lambda_1+\lambda_2)^{(n-x)}}\right) \\ &= \left(\frac{n!}{x!(n-x)!}\right)\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{(n-x)} \\ &= \binom{n}{x}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{(n-x)} \\ &= \binom{n}{x}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x\left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{(n-x)} \\ &= \binom{n}{x}\rho^x(1-\rho)^{(n-x)} \sim \text{Binomial}(n,\rho)\end{aligned}$$

for $\rho = \lambda_1/(\lambda_1+\lambda_2)$. $\square$

2.17
The joint probability density is:

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

Calculating $\mathbb{P}(X < 1/2 \,|\, Y = 1/2)$.

We begin by calculating the marginal distribution for $f_Y(y)$.

$$f_Y(y) = \int_0^1 cx + cy^2 dx = c\int_0^1 x dx + c\int_0^1 y^2 dx = c\Big[\frac{x^2}{2}\Big]_0^1 + cy^2\big[x\big]_0^1 = c\left(\frac{1}{2} + y^2\right)$$

Following the steps in example 2.38.

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{c(x+y^2)}{c\left(\frac{1}{2}+y^2\right)} = \frac{x+y^2}{\frac{1}{2}+y^2}$$

Calculating the conditional probability:

$$\begin{aligned}
\mathbb{P}(X < 1/2 | Y = 1/2) &= \int_0^{1/2} f_{X|Y}(1/2|1/2) dx \\
&= \int_0^{1/2} \frac{x + \frac{1}{4}}{\frac{1}{2}+\frac{1}{4}} dx \\
&= \frac{4}{3}\int_0^{1/2} x + \frac{1}{4} dx \\
&= \frac{4}{3}\left[\frac{x^2}{2} + \frac{x}{4}\right]_0^{1/2} = \frac{4}{3}\cdot\frac{1}{4} = \frac{1}{3}
\end{aligned}$$

2.18
Let $X \sim N(3, 16)$. Solve the following using the Normal table and using a computer package.

(a)

$$\mathbb{P}(X < 7) = \mathbb{P}(3 + 4Z < 7) = \mathbb{P}\left(Z < \frac{7-3}{4}\right) = \mathbb{P}(Z < 1) = 0.8413$$

From **R**:

```
> pnorm(1)
[1] 0.84134475
```

```
> 1 - pnorm(-0.25)
[1] 0.59870633
```

(b)

$$\begin{aligned}
\mathbb{P}(X > 2) &= 1 - \mathbb{P}(X \leqslant 2) = 1 - \mathbb{P}(3 + 4Z \leqslant 2) = 1 - \mathbb{P}\left(Z \leqslant \frac{2-3}{4}\right) \\
&= 1 - \mathbb{P}(Z < -1/4) = 1 - (1 - 0.5987) = 0.5987
\end{aligned}$$

(c) Find x such that $\mathbb{P}(X > x) = 0.05$. By finding 0.95 in the table, we see that it is generated by the value 1.645, so:

$$\mathbb{P}(Z \leqslant 1.645) = 0.95 \quad \Longrightarrow \quad \mathbb{P}(Z > 1.645) = 1 - 0.95 = 0.05$$

$$\mathbb{P}(4Z > 6.58) = \mathbb{P}(3 + 4Z > 9.58) = \mathbb{P}(X > 9.58) = 0.05$$

So, $x = 9.58$. From **R**:

```
1 - pnorm(9.5794, mean = 3, sd = 4)
[1] 0.05000037
```

(d)

$$\begin{aligned}
\mathbb{P}(0 \leqslant X < 4) &= \mathbb{P}(0 \leqslant 3 + 4Z < 4) = \mathbb{P}(-3 \leqslant 4Z < 1) \\
&= \mathbb{P}(-3/4 \leqslant Z < 1/4) = \mathbb{P}(Z < 1/4) - \mathbb{P}(Z < -3/4) \\
&= 0.5987 - (1 - 0.7734) = 0.5987 - 0.2266 \\
&= 0.3721
\end{aligned}$$

From **R**:

```
> pnorm(0.25) - pnorm(-0.75)
[1] 0.37207897
#### Alternatively
> CH = rnorm(10000000, mean = 3, sd = 4)
> sum(CH < 4 & CH > 0)/10000000
[1] 0.3719199
```

(e) Find x such that $\mathbb{P}(|X| > |x|) = 0.05$.

We can split this up in two sections if we assume $x > 0$. Using that the sets are disjoint:

$$\begin{aligned}
\mathbb{P}(|X| > |x|) &= \mathbb{P}(X > x \cap X < -x) = \mathbb{P}(X > x) + \mathbb{P}(X < -x) \\
&= 1 - \mathbb{P}(X \leqslant x) + \mathbb{P}(X < -x) = 0.05 \\
\mathbb{P}(X \leqslant x) - \mathbb{P}(X < -x) = 0.95
\end{aligned}$$

Note we get minus. Translating to standard normal:

$$\mathbb{P}\left(Z \leqslant \frac{x-3}{4}\right) - \mathbb{P}\left(Z < \frac{-x-3}{4}\right) = 0.95$$

I am not able to see how to solve this by just looking up values because of the dependence. By using R, we can find that $x = 9.611$ which makes the lookup values 1.65275 and -3.15275. Verifying in **R**.

```
> X = rnorm(10000000, mean = 3, sd = 4)
> x = 9.611
> sum(X < -x | X > x)/10000000
[1] 0.0500769
> sum(abs(X) > abs(x))/10000000
[1] 0.0500769
```

2.19

Proving equation (2.12). Let X be a random variable with pdf $f_X(x)$, and define the random variable $Y = r(X)$, where r is either strictly monotone increasing or decreasing, so $r(\cdot)$ has the inverse function $s(\cdot) = r^{-1}(\cdot)$. Then, we can express the pdf of Y as:

$$f_Y(y) = f_X(s(y))\Big|\frac{ds(y)}{dy}\Big|.$$

Proof. The proof will be in two parts. Assume first that r is strictly increasing, which means that s is strictly increasing and its derivative is positive. From the CDF of Y, we get:

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leqslant y) \\ &= \mathbb{P}(r(X) \leqslant y) \\ &= \mathbb{P}(X \leqslant r^{-1}(y)) \\ &= \mathbb{P}(X \leqslant s(y)) \\ &= F_X(s(y)) \end{aligned}$$

Differentiating the CDF to get the PDF.

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X(s(y)) = f_X(s(y))\frac{ds(y)}{dy}$$

Now assume that r is strictly decreasing, which means s is strictly decreasing and thus has a negative derivative. From the CDF of Y, we now get:

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leqslant y) \\ &= \mathbb{P}(r(X) \leqslant y) \\ &= \mathbb{P}(X \geqslant r^{-1}(y)) \\ &= 1 - \mathbb{P}(X \leqslant s(y)) \\ &= 1 - F_X(s(y)) \end{aligned}$$

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\big[1 - F_X(s(y))\big] = -f_X(s(y))\frac{ds(y)}{dy} = f_X(s(y))\left(-\frac{ds(y)}{dy}\right)$$

Since the derivative will be negative, the negative of the derivative will be positive. Generalizing to both cases, we can write:

$$f_Y(y) = f_X(s(y))\Big|\frac{ds(y)}{dy}\Big| \qquad \square$$

A short note on why we get $1 - \mathbb{P}(X \leqslant s(y))$ when r is strictly decreasing. As an example, say that $X \sim U(0, 5)$ and we define $Y = r(X)$ where $r(x) = 2 - 2x$, which is strictly decreasing. Then $r^{-1}(y) = s(y) = 1 - \frac{y}{2}$, which is also strictly decreasing. Then:

$$\mathbb{P}(r(X) \leqslant y) = \mathbb{P}(2 - 2X \leqslant y) = \mathbb{P}(2 - y \leqslant 2X) = \mathbb{P}(1 - \frac{y}{2} \leqslant X) = \mathbb{P}(X \geqslant s(y))$$

In order to get the expression in terms of $F_X(x)$, we rewrite it:

$$\mathbb{P}(X \geqslant s(y)) = 1 - \mathbb{P}(X \leqslant s(y)) = 1 - F_X(s(y)).$$

2.20

Let $X, Y \sim U(0,1)$ be independent. Find the PDF for $X - Y$ and X/Y. The joint density of X, Y is

$$f(x,y) = 1, \quad x, y \in (0,1)$$

and 0 elsewhere.

We define $Z = X - Y$. We note that Z can assume values between -1 and 1, which are the two extremes. This gives us $A_z = (-1,1)$. We split this into two cases: when $X > Y$, then $A_z = (0,1)$ and when $X < Y$, which means $A_z = (-1,0)$.

When $X > Y$ and $A_z = (0,1)$.

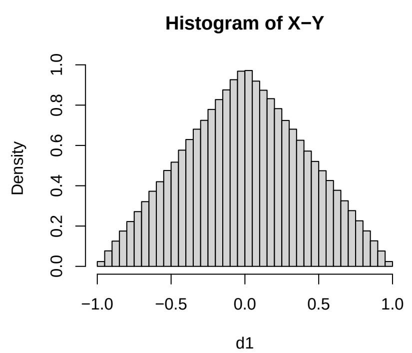
$$\begin{aligned} F_Z(z) &= \int_0^z \int_y^1 1 dx dy = \int_0^z [x]_y^1 dy \\ &= \int_0^z 1 - y dy = \left[y - \frac{y^2}{2} \right]_0^z \\ &= z - \frac{z^2}{2} \end{aligned}$$

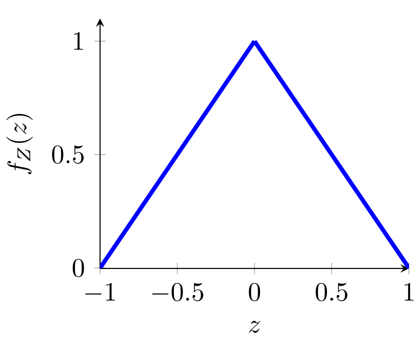
When $X < Y$ and $A_z = (-1,0)$.

$$\begin{aligned} F_Z(z) &= \int_{-1}^z \int_{-1}^y 1 dx dy = \int_{-1}^z [x]_{-1}^y dy \\ &= \int_{-1}^z y + 1 dy = \left[\frac{y^2}{2} + y \right]_{-1}^z \\ &= \frac{z^2}{2} + z + \frac{1}{2} \end{aligned}$$

Summarizing the CDF, and differentiating to find the PDF. (The 1/2 carries over).

$$F_Z(z) = \begin{cases} \frac{z^2}{2} + z + \frac{1}{2} & z \in (-1,0) \\ z - \frac{z^2}{2} + \frac{1}{2} & z \in (0,1) \end{cases} \implies f_Z(z) = \begin{cases} z + 1 & z \in (-1,0) \\ 1 - z & z \in (0,1) \end{cases}$$





Now we define $Z = X/Y$.

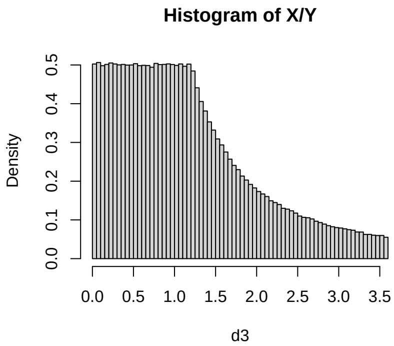
$$F_Z(z) = \mathbb{P}(Z \leqslant z) = \mathbb{P}(X/Y \leqslant z) = \mathbb{P}(X \leqslant zY)$$

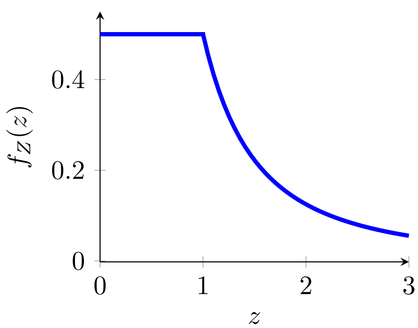
$$F_Z(z) = \int_0^1 \int_0^{\min(z,y)} 1 dx dy = \int_0^1 \min(z,y) dy = \begin{cases} \int_0^{\frac{1}{z}} zy dy + \int_{\frac{1}{z}}^1 1 dy & z \geqslant 1 \\ \int_0^1 zy dy & z < 1 \end{cases}$$

Evaluating the integrals to get the CDF, and differentiating to get the PDF:

$$F_Z(z) = \begin{cases} 1 - \dfrac{1}{2z} & z \geqslant 1 \\ \dfrac{z}{2} & z < 1 \end{cases} \implies f_Z(z) = \begin{cases} \dfrac{1}{2z^2} & z \geqslant 1 \\ \dfrac{1}{2} & z < 1 \end{cases}$$

Plots and simulation code.






```
# Simulating X-Y
N = 10000
X = runif(N);Y = runif(N)
d1 = X - Y
hist(d1, breaks = 50, prob = TRUE,
        main="Histogram of X-Y")

# Simulating X/Y
N = 1000000
X = runif(N);Y = runif(N)
d2 = X/Y
d3 = d2[abs(d2) < 3] # Removing very rare large vals
d3 = d3*length(d2)/length(d3) # Normalizing probability
hist(d3, breaks=60, prob=TRUE,
        main = "Histogram of X/Y")
```

R

2.21
Let $X_1, \dots, X_n \sim \text{Exp}(\beta)$ be IID. Define $Y = \max(X_1, \dots, X_n)$. Find the PDF of Y.

Recall the PDF of the exponential distribution:

$$f_X(x) = \frac{1}{\beta} \exp\left(-x/\beta\right), \quad x \geqslant 0$$

Defining $r(X_1, \dots, X_n) = \max(X_1, \dots, X_n)$. Finding an expression for the CDF of Y.

$$F_Y(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(\max(X_1, \dots, X_n) \leqslant y)$$

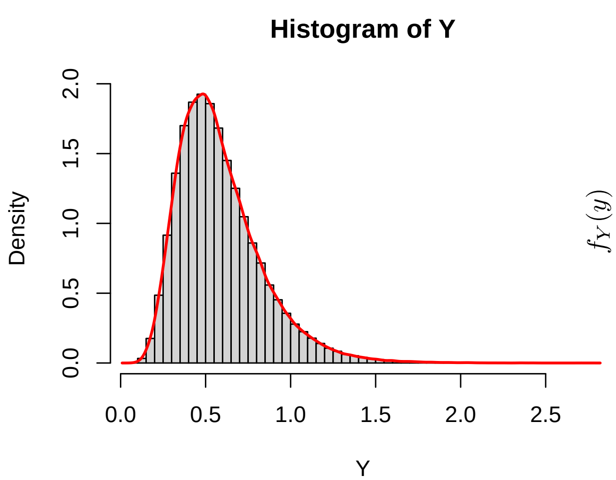
According to the hint, $Y \leqslant y$ iff $X_i \leqslant y$ for all $i \in I$, where $I = \{1, \dots, n\}$ is the index set. We can therefore express the CDF of Y in terms of each individual CDF:

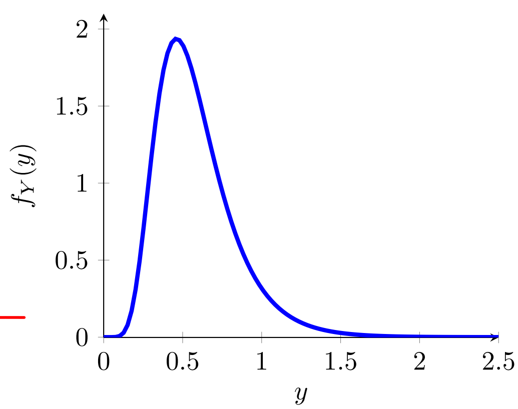
$$F_Y(y) = \mathbb{P}(X_1 \leqslant y) \cap \mathbb{P}(X_2 \leqslant y) \cap \ldots \cap \mathbb{P}(X_n \leqslant y) = \left(F_X(y)\right)^n = \left(1 - \exp(-y/\beta)\right)^n$$

Differentiating wrt. y to find the PDF.

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy}(F_X(y))^n = n(F_X(y))^{n-1} f_X(y) \\ &= n\left(1 - \exp(-y/\beta)\right)^{n-1} \frac{1}{\beta} \exp\left(-y/\beta\right) \\ &= \frac{n}{\beta} \exp\left(-y/\beta\right) \left(1 - \exp(-y/\beta)\right)^{n-1} \end{aligned}$$

Plots and simulation code for $n = 10$ and $\beta = 0.2$.





```
N = 100000
lambda = 5

exp1 = rexp(N, rate = lambda)
exp2 = rexp(N, rate = lambda)
exp3 = rexp(N, rate = lambda)
exp4 = rexp(N, rate = lambda)
exp5 = rexp(N, rate = lambda)
exp6 = rexp(N, rate = lambda)
exp7 = rexp(N, rate = lambda)
exp8 = rexp(N, rate = lambda)
exp9 = rexp(N, rate = lambda)
exp10 = rexp(N, rate = lambda)

Y = pmax(exp1, exp2, exp3, exp4, exp5,
            exp6, exp7, exp8, exp9, exp10)

hist(Y, breaks = 50, prob = TRUE)
lines(density(Y), col="red", xlim=c(0,2.5), lwd=2)
```

R

3 Expectation

Exercises

3.1

Define X as the wealth after n games. The probability of winning and losing is the same for each outcome so $p = 1/2$.

$$\mathbb{E}[X] = \frac{1}{2} \cdot 2c + \frac{1}{2} \cdot \left(\frac{1}{2}\right) c = c + \frac{c}{4} = \frac{5}{4}c$$

We expect to have $5/4 \cdot c$ after n games. We can also verify this result with a simulation in **R**.

```
> games = sample(c(2, 0.5), size = 1000000, replace = TRUE)
> mean(games)
[1] 1.250973
> 5/4
[1] 1.25
```

R

3.2

Claim. $\text{Var}(X) = 0$ if and only if $\mathbb{P}(X = c) = 1$ for some constant c.

Proof.
$\Rightarrow$) Set $c = \mu$ and assume $\text{Var}(X) = 0$, which means that

$$\mathbb{E}[(X - \mu)^2] = 0 \quad \Longrightarrow \quad \int (x - \mu)^2 dF(x) = 0.$$

This can only be 0 when $x = \mu = c$ for the entire domain of X. Hence $\mathbb{P}(X = c) = 1$.

$\Leftarrow$) Assume $\mathbb{P}(X = c) = 1$. When calculating the expectation:

$$\mu = \mathbb{E}[X] = \int c dF(x) = c$$

When calculating the variance:

$$\text{Var}(X) = \int (x - \mu)^2 dF(x) = \int (c - c)^2 dF(x) = 0,$$

since $x = c$ for all x in the domain of X. $\square$

3.3

Let $X_1, \ldots, X_n \sim U(0,1)$ and define $Y = \max(X_1, \ldots, X_n)$. We will calculate $\mathbb{E}[Y]$. It is not stated in the exercise, but we will assume that the X_i are independent. Finding the CDF for Y.

$$\begin{aligned}
F_Y(y) &= \mathbb{P}(Y \leqslant y) \\
&= \mathbb{P}(\max(X_1, \ldots, X_n) \leqslant y) \\
&= \mathbb{P}(X_1 \leqslant y) \cap \ldots \cap \mathbb{P}(X_n \leqslant y) \\
&= \mathbb{P}(X_1 \leqslant y)\mathbb{P}(X_2 \leqslant y) \cdots \mathbb{P}(X_n \leqslant y) && \text{(Independence)} \\
&= (F_X(y))^n
\end{aligned}$$

Differentiating to get the PDF for Y.

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}(F_X(y))^n = n(F_X(y))^{n-1} f_X(y)$$

Since X_i are uniformly distributed, we know that $F_X(y) = y$ and $f_X(y) = 1$, so:

$$f_Y(y) = ny^{n-1}$$

Now we can calculate the expectation of Y.

$$\mathbb{E}[Y] = \int_0^1 y \cdot ny^{n-1} dy = n \int_0^1 y^n dy = n\Big[\frac{y^{n+1}}{n+1}\Big]_0^1 = \frac{n}{n+1}$$

Confirming this result with a numeric simulation in **R**.

```
> # 3.3
> N = 1000000
> U1 = runif(N)
> U2 = runif(N)
> U3 = runif(N)
> U4 = runif(N)
> U5 = runif(N)
> U6 = runif(N)
> U7 = runif(N)
> U8 = runif(N)
> U9 = runif(N)
> U10 = runif(N)
> Y = pmax(U1, U2, U3, U4, U5,
+          U6, U7, U8, U9, U10)
> mean(Y)
[1] 0.9091151
> # Theoretical Result
> 10/11
[1] 0.9090909
```

R

As we can see, the theoretical result is very close to the simulated result for $n = 10$.

3.4 - Random Walk

A particle starts in the origin and jumps left, a step of -1, with probability p and jumps right, a step of 1, with probability $1 - p$. The expected location will be:

$$\mathbb{E}[X] = (-1)p + (1)(1-p) = -p + 1 - p = 1 - 2p$$

To calculate the variance, we start by finding the second moment:

$$\mathbb{E}[X^2] = (-1)^2 p + (1)^2(1-p) = p + 1 - p = 1$$

So the variance is:

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1 - (1-2p)^2 = 1 - (1 - 4p + 4p^2) = 4p - 4p^2$$

3.5

Tossing a fair coin until we get H. Finding the expected number of tosses. The reasoning is as follows. We get H on the first toss with probability 1/2, first H on the second toss with probability $1/2^2 = 1/4$ and so on. The pattern becomes as follows, for the first 7 cases:

Tosses	**Outcome**	**Probability**
1	$\{H\}$	1/2
2	$\{TH\}$	1/4
3	$\{TTH\}$	1/8
4	$\{TTTH\}$	1/16
5	$\{TTTTH\}$	1/32
6	$\{TTTTTH\}$	1/64
7	$\{TTTTTTH\}$	1/128

Define T to be the number of tosses to get H.

$$\begin{aligned}\mathbb{E}[T] &= (1)\left(\frac{1}{2}\right) + (2)\left(\frac{1}{4}\right) + \ldots + (k)\left(\frac{1}{2^k}\right) + \ldots \\ &= \sum_{k=1}^{\infty} \frac{k}{2^k} \\ &= 2\end{aligned}$$

Not delving in to the mathematics of the infinite sum, but it can be shown that this sum becomes 2 which will be the expected number of tosses to get a H. Here is a numeric approximation in **R**.

```
> sumApprox = 0
> for (k in 1:1000) {
+   sumApprox = sumApprox + k/2^k
+ }
> sumApprox
[1] 2
```

R

3.6 Theorem - The Rule of the Lazy Statistician
Proving the following result for the discrete case. Let $Y = r(X)$, then

$$\mathbb{E}[Y] = \mathbb{E}[r(X)] = \sum_x r(x) f_X(x)$$

Proof. Set $\mathcal{X}, \mathcal{Y}$ to be the set of all values for X and Y. We have the functions $r : \mathcal{X} \to \mathcal{Y}$ and $r^{-1} : \mathcal{Y} \to \mathcal{X}$. We will define $s := r^{-1}$ for convenience. By definition of the expectation:

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{y\in\mathcal{Y}} y \cdot f_Y(y) = \sum_{y\in\mathcal{Y}} y \cdot \mathbb{P}(Y = y) \\
&= \sum_{y\in\mathcal{Y}} y \sum_{x\in s(y)} \mathbb{P}(X = x) \\
&= \sum_{y\in\mathcal{Y}} \sum_{x\in s(y)} y\mathbb{P}(X = x) \\
&= \sum_{x\in\mathcal{X}} r(x)\mathbb{P}(X = x) \\
&= \sum_{x\in\mathcal{X}} r(x) f_X(x)
\end{aligned}$$

□

A simplified case to make the proof easier a bit easier to understand. We define the variables X and $Y = r(X) = X^2$, and use the following distributions.

x	$\mathbb{P}(X = x)$
-1	$1/4$
0	$1/4$
1	$1/2$

$Y = r(X) = X^2$

y	$\mathbb{P}(Y = y)$
0	$1/4$
1	$3/4$

Then the calculations above become:

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{y\in\mathcal{Y}} y \cdot f_Y(y) = (0)(1/4) + (1)(3/4) \\
&= \sum_{y\in\mathcal{Y}} y \sum_{x\in s(y)} \mathbb{P}(X = x) = (0)(1/4) + (1)\big[1/4 + 1/2\big] \\
&= \sum_{y\in\mathcal{Y}} \sum_{x\in s(y)} y\mathbb{P}(X = x) = (0)(1/4) + (1)(1/4) + (1)(1/2) \\
&= \sum_{x\in\mathcal{X}} r(x)\mathbb{P}(X = x) = (0)^2(1/4) + (-1)^2(1/4) + (1)^2(1/2) \\
&= \sum_{x\in\mathcal{X}} r(x) f_X(x)
\end{aligned}$$

In both cases, the expectation becomes $3/4$.

3.7
$X \sim F$ is continuous and we suppose that $\mathbb{P}(X > 0) = 1$ and that $\mathbb{E}[X]$ exists. Show that $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)dx$.

PROOF. By definition of the expectation, and since the domain of X must be $(0, \infty)$:

$$\mathbb{E}[X] = \int_0^\infty x \cdot f_X(x)dx$$

Using integration by parts, and setting $u = x$ and $v' = f_X$, which gives $u' = 1$ and $v = F_X$:

$$\int uv' = uv - \int u'v,$$

which gives us:

$$\begin{aligned}
\mathbb{E}[X] = \int_0^\infty x \cdot f_X(x)dx &= xF_X(x)\Big|_0^\infty - \int_0^\infty F_X(x)dx \\
&= xF_X(x)\Big|_0^\infty - \int_0^\infty \mathbb{P}(X \leqslant x)dx \\
&= xF_X(x)\Big|_0^\infty - \int_0^\infty 1 - \mathbb{P}(X > x)dx \\
&= xF_X(x)\Big|_0^\infty - x\Big|_0^\infty + \int_0^\infty \mathbb{P}(X > x)dx \\
&= -\big(\lim_{x\to\infty} x(1 - F_X(x)\big) + \int_0^\infty \mathbb{P}(X > x)dx \\
&= \int_0^\infty \mathbb{P}(X > x)dx
\end{aligned}$$

which proves the result. In the second to last step, we applied the hint regarding the limit. $\square$

3.8 - Theorem 3.17
Let $X_1, \ldots, X_n$ be IID and let $\mu = \mathbb{E}[X_i]$, $\sigma^2 = \text{Var}(X_i)$. Then:

$$\mathbb{E}[\overline{X}] = \mu, \quad \text{Var}(\overline{X}) = \frac{\sigma^2}{n}, \quad \mathbb{E}[S^2] = \sigma^2.$$

PROOF. Start with the expectation of the sample mean:

$$\mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^n \mu = \frac{n \cdot \mu}{n} = \mu.$$

Next, the variance.

$$\text{Var}[\overline{X}] = \text{Var}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n^2}\sum_{i=1}^n \text{Var}[X_i] = \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Finally, the expectation of the sample variance. This calculation is a lot more involved.

$$\mathbb{E}[S^2] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] \implies (n-1)\mathbb{E}[S^2] = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]$$

We will use that $\overline{X} = (1/n)\Sigma X_i$ means that $\Sigma X_i = n\overline{X}$. Running through the calculations:

$$\begin{aligned}
(n-1)\mathbb{E}[S^2] &= \mathbb{E}\left[\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i\overline{X} + \sum_{i=1}^{n}\overline{X}^2)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] - \mathbb{E}\left[2\sum_{i=1}^{n} X_i\overline{X}\right] + \mathbb{E}\left[\sum_{i=1}^{n}\overline{X}^2)\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[2\overline{X}\sum_{i=1}^{n} X_i\right] + \mathbb{E}\left[\overline{X}^2\sum_{i=1}^{n} 1)\right] && (\overline{X} \text{ indp. of } i) \\
&= n\mathbb{E}\left[X_i^2\right] - 2n\mathbb{E}\left[\overline{X}^2\right] + n\mathbb{E}\left[\overline{X}^2\right] && (\Sigma X_i = n\overline{X}) \\
&= n\mathbb{E}\left[X_i^2\right] - n\mathbb{E}\left[\overline{X}^2\right]
\end{aligned}$$

By dividing both sides of the equality by n, we have shown:

$$\frac{n-1}{n}\mathbb{E}[S^2] = \mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\overline{X}^2\right] \tag{3.8.1}$$

We have the second moment for X. From the definition of the variance:

$$\operatorname{Var}(X_i) = \mathbb{E}[X_i^2] - \mu^2 \implies \mathbb{E}[X_i^2] = \operatorname{Var}(X_i) + \mu^2 = \sigma^2 + \mu^2$$

From the previous results:

$$\operatorname{Var}(\overline{X}) = \mathbb{E}[\overline{X}^2] - \mathbb{E}[\overline{X}]^2 \implies \mathbb{E}[\overline{X}^2] = \operatorname{Var}(\overline{X}) + \mathbb{E}[\overline{X}]^2 = \frac{\sigma^2}{n} + \mu^2$$

Replacing each of these into (3.8.1) and isolating $\mathbb{E}[S^2]$.

$$\begin{aligned}
\mathbb{E}[S^2] &= \frac{n}{n-1}\left(\sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right)\right) \\
&= \frac{n}{n-1}\left(\sigma^2 - \frac{\sigma^2}{n}\right) \\
&= \frac{n}{n-1}\left(\frac{n-1}{n}\right)\sigma^2 \\
&= \sigma^2
\end{aligned}$$

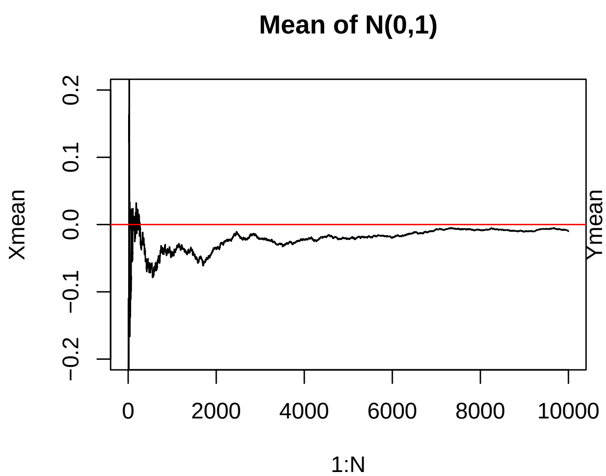
and we have finally shown the final result. $\square$

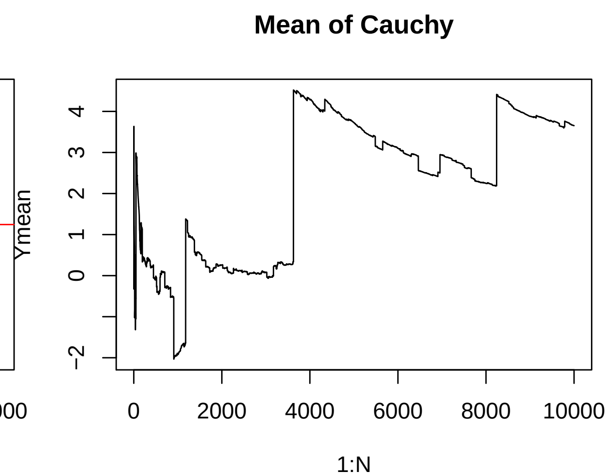
3.9

Computer experiment; studying the effects of the mean over 10.000 simulations from the standard normal distribution and the Cauchy distribution.

As we can see the simulation of the normal distribution quickly stabilizes to a value of around 0 which is the expected value. The Cauchy distribution is a very heavy-tailed distribution and the expectation does not exist. Therefore, there is no mean it can stablize to and it ends up behaving erraticaly even after 10.000 simulations. The jumps that can be seen are extreme outliers that are added to the data, which are relatively common in the Cauchy distribution.

Results:






```
# Setting the number of simulations
N = 10000

# Simulating N standard normal values
X = rnorm(N)
Xmean = cumsum(X)/1:N
plot(1:N, Xmean, type="l", ylim = c(-0.2, 0.2),
main = "Mean of N(0,1)")
abline(h = 0, col="red")

# Simulating N Cauchy values with loc = 0, scale = 1
Y = rcauchy(N)
Ymean = cumsum(Y)/1:N
plot(1:N, Ymean, type="l",
main = "Mean of Cauchy")
```

R

3.10
Let $X \sim N(0,1)$ and define $Y = e^X$. Find $\mathbb{E}[Y]$ and $\text{Var}(Y)$.

One possible way to solve this is defining $r(X) = e^X$ and evaluating:

$$\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int_{\mathbb{R}} e^x f_X(x) dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(x - \frac{1}{2}x^2\right) dx.$$

Completing the square in the exponential term:

$$x - \frac{1}{2}x^2 = -\frac{1}{2}\left(x^2 - 2x\right) = -\frac{1}{2}\left(x^2 - 2x + 1 - 1\right) = -\frac{1}{2}\left(x-1\right)^2 + \frac{1}{2}$$

We can rewrite the integral:

$$\mathbb{E}[Y] = e^{1/2} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}(x-1)^2\right) dx$$

We will integrate this by substitution:

$$u = x - 1 \quad \Longrightarrow \quad du/dx = 1 \quad \Longrightarrow \quad du = dx$$

$$\mathbb{E}[Y] = e^{1/2} \underbrace{\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}u^2\right) du}_{=1} = e^{1/2} \approx 1.6487$$

Confirming with numerical simulation.

```
> N = 10000000; X = rnorm(N); Y = exp(X)
> mean(Y)
[1] 1.648564
> exp(0.5)
[1] 1.648721
> var(Y)
[1] 4.676382
> exp(2) - exp(1)
[1] 4.670774
```

We can calculate the variance by first finding the second moment: $\mathbb{E}[Y^2]$.

$$\mathbb{E}[Y^2] = \mathbb{E}[r(X)^2] = \int_{\mathbb{R}} e^{2x} f_X(x) dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(2x - \frac{1}{2}x^2\right) dx.$$

Just as above, we complete the square and get: $-(1/2)(x-2)^2 + 2$ and move e^2 outside the integral. Then we use integration by substitution: $u = x - 2$ which gives $du = dx$. We end up with a similar integral:

$$\mathbb{E}[Y^2] = e^2 \underbrace{\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}u^2\right) du}_{=1} = e^2$$

Now we can calculate the variance of Y.

$$\text{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = e^2 - (e^{1/2})^2 = e^2 - e^1,$$

and as we can see in the simulation above, this corresponds to the simulated variance.

3.11

Simulating stocks. For each day Y_i can be either -1 or 1, each with probability 1/2. We have $X_n = \sum_{i=1}^n Y_i$ which is the cumulative value of the stock.

(a) Calculating the expectation and variance of X_n. First we calculate it for Y_i.

$$\mathbb{E}[Y_i] = (-1)(1/2) + (1)(1/2) = -1/2 + 1/2 = 0$$
$$\mathbb{E}[Y_i^2] = (-1)^2(1/2) + (1)^2(1/2) = 1/2 + 1/2 = 1$$
$$\text{Var}(Y_i) = \mathbb{E}[Y_i^2] - \mathbb{E}[Y_i]^2 = 1 - 0 = 1$$

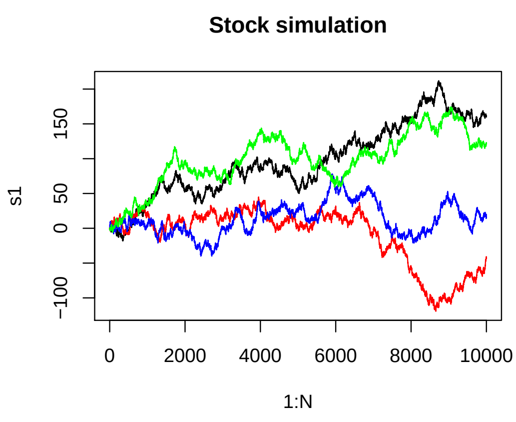
Now for X_n.

$$\mathbb{E}[X_n] = \mathbb{E}\left[\sum_{i=1}^n Y_i\right] = \sum_{i=1}^n \mathbb{E}[Y_i] = 0$$
$$\text{Var}(X_n) = \text{Var}\left[\sum_{i=1}^n Y_i\right] = \sum_{i=1}^n \text{Var}[Y_i] = \sum_{i=1}^n 1 = n$$

(b) Simulating four stocks.

```r
simStock <- function(N) {
    dailyMovements = sample(c(-1, 1), size=N, replace = TRUE)
    totalMovements = cumsum(dailyMovements)
    return(totalMovements)
}
N = 10000
s1 = simStock(N);s2 = simStock(N)
s3 = simStock(N);s4 = simStock(N)
plot(1:N, s1, type="l",
    ylim=c(-119, 212), # Adjust according to simulation results
    main = "Stock simulation")
lines(1:N, s2, type="l", col="red")
lines(1:N, s3, type="l", col="blue")
lines(1:N, s4, type="l", col="green")
```

R



The simulations have mean 0, so they should vary around the x-axis. Since the variance is n, it is dependent on the time. The longer the simulations last, the more they will vary which is what we see.

The standard deviation in this example will be the square root of n, which is about 100, and that is also what we see (and it is expected that they will not be exactly on the standard deviation).

3.12
Deriving the general expression for the expectation and variance for a whole bunch of probability distributions.

- **Bernoulli** with parameter p. Possible values: $\mathcal{X} = \{0, 1\}$.

$$f(x) = p^x(1-p)^{x-1}$$

Expectation.

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot f(x) = (0) + (1)p^1 = p$$

Second moment.

$$\mathbb{E}[X^2] = \sum_{x^2 \in \mathcal{X}} x^2 \cdot f(x) = (0) + (1)^2 p^1 = p$$

Variance.

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$$

- **Poisson** with parameter λ. Possible values: $\mathcal{X} = \mathbb{N} \cup \{0\}$.

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

Expectation. First term disappears. Cancel x against the factorial. Switch to $k = x - 1$.

$$\begin{aligned}
\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x f(x) &= \sum_{x=0}^{\infty} x \cdot e^{-\lambda}\frac{\lambda^x}{x!} \\
&= 0 + \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} \\
&= \lambda e^{-\lambda}e^{\lambda} \\
&= \lambda
\end{aligned}$$

Second moment. We will split the sum into two Poisson sums.

$$\begin{aligned}
\mathbb{E}[X^2] = \sum_{x^2 \in \mathcal{X}} x^2 \cdot f(x) &= \sum_{x=0}^{\infty} x^2 \cdot e^{-\lambda}\frac{\lambda^x}{x!} \\
&= 0 + \lambda e^{-\lambda}\sum_{x=1}^{\infty} x \cdot \frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda e^{-\lambda}\sum_{x=1}^{\infty}(x-1+1) \cdot \frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda e^{-\lambda}\sum_{x=1}^{\infty}\left[(x-1)\frac{\lambda^{x-1}}{(x-1)!} + \frac{\lambda^{x-1}}{(x-1)!}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \lambda e^{-\lambda} \left(\sum_{x=2}^{\infty} (x-1) \frac{\lambda^{x-1}}{(x-1)!} + \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right) \\
&= \lambda e^{-\lambda} \left(\lambda \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right) \\
&= \lambda e^{-\lambda} \left(\lambda \sum_{j=0}^{\infty} \frac{\lambda^{j}}{j!} + \sum_{k=0}^{\infty} \frac{\lambda^{k}}{k!} \right) \\
&= \lambda e^{-\lambda} \left(\lambda e^{\lambda} + e^{\lambda} \right) \\
&= \lambda^2 + \lambda
\end{aligned}
$$

The remaining calculation for the variance is easy.

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

- **Uniform** on (a, b). Possible values: $\mathcal{X} = (a, b)$.

$$
f(x) = \frac{1}{b-a}
$$

Expectation.

$$
\begin{aligned}
\mathbb{E}[X] = \int_{\mathcal{X}} x \cdot f(x) &= \frac{1}{b-a} \int_a^b x dx \\
&= \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b = \frac{1}{b-a} \left(\frac{b^2 - a^2}{2} \right) \\
&= \frac{1}{b-a} \left(\frac{(b+a)(b-a)}{2} \right) \\
&= \frac{b+a}{2}
\end{aligned}
$$

Second moment.

$$
\begin{aligned}
\mathbb{E}[X^2] = \int_{\mathcal{X}} x^2 \cdot f(x) &= \frac{1}{b-a} \int_a^b x^2 dx \\
&= \frac{1}{b-a} \left[\frac{x^3}{3} \right]_a^b = \frac{1}{b-a} \left(\frac{b^3 - a^3}{3} \right) \\
&= \frac{1}{b-a} \left(\frac{(b-a)(b^2 + ab + a^2)}{3} \right) \\
&= \frac{b^2 + ab + a^2}{3}
\end{aligned}
$$

Variance.

$$
\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 &= \frac{b^2 + ab + a^2}{3} - \frac{(b+a)^2}{4} \\
&= \frac{4b^2 + 4ab + 4a^2 - 3b^2 - 6ab - 3a^2}{12} = \frac{b^2 - 2ab + a^2}{12} = \frac{(b-a)^2}{12}
\end{aligned}
$$

• **Exponential** with parameter β. Possible values: $\mathcal{X} = \{x \in \mathbb{R} : x \geqslant 0\}$.

$$f(x) = \frac{1}{\beta}\exp(-x/\beta)$$

Expectation.

$$\mathbb{E}[X] = \int_{\mathcal{X}} x \cdot f(x) = \int_0^\infty x \cdot \frac{1}{\beta} e^{-x/\beta} dx$$

We will integrate this with the substitution $u = x/\beta$, so $du/dx = 1/\beta$ which gives us $dx = \beta du$. Note that the antiderivative of ue^{-u} is $-ue^{-u} - e^{-u}$.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x \cdot \frac{1}{\beta} e^{-x/\beta} dx = \beta \int_0^\infty ue^{-u} du \\
&= \beta \Big[-ue^{-u} - e^{-u} \Big]_0^\infty \\
&= \beta \Big(\lim_{u \to \infty} (-ue^{-u} - e^{-u}) - (-1) \Big) \\
&= \beta
\end{aligned}$$

Skipping the evaluation of the limit. The first term can be shown to converge to 0 by using L'Hôpital's rule. The second term in the limit obviously becomes 0.
Second moment. Will derive the expression for the expectation in the calculations.

$$\mathbb{E}[X^2] = \int_{\mathcal{X}} x^2 \cdot f(x) = \int_0^\infty x^2 \cdot \frac{1}{\beta} e^{-x/\beta} dx$$

Using integration by parts: $u = x^2$ so $u' = 2x$. Setting $v' = (1/\beta)e^{-x/\beta}$ which yields $v = -e^{-x/\beta}$. From the integration-by-parts formula:

$$\int uv' = uv - \int u'v$$

$$\begin{aligned}
\mathbb{E}[X^2] = \int_0^\infty x^2 \cdot \frac{1}{\beta} e^{-x/\beta} dx &= \Big[-x^2 e^{-x/\beta} \Big]_0^\infty - \int_0^\infty -2xe^{-x/\beta} dx \\
&= 0 + \int_0^\infty 2xe^{-x/\beta} dx \\
&= 2\beta \int_0^\infty x \cdot \frac{1}{\beta} e^{-x/\beta} dx \\
&= 2\beta \mathbb{E}[X] \\
&= 2\beta^2
\end{aligned}$$

The limit can be shown to converge to 0 by applying L'Hôpital's rule twice. We use a clever trick and substitute in the expectation and get the result. Finally, we calculate the variance.

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

• **Gamma** with parameters α, β. Possible values: $\mathcal{X} = \{x \in \mathbb{R} : x > 0\}$.

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta)$$

We will use the definition of the Gamma function, and the 'Gamma difference' chaining property:

$$\Gamma(\alpha+1) = \int_0^\infty t^\alpha e^{-t} dt, \quad \Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

In step 3, we make the substitution $t = x/\beta$ which makes $x = \beta t$. $dt/dx = 1/\beta$, so $dx = \beta dt$.

$$\begin{aligned}
\mathbb{E}[X] &= \int_{\mathcal{X}} x \cdot f(x) = \int_0^\infty x \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta) dx \\
&= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^\alpha \exp(-x/\beta) dx \\
&= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} (\beta t)^\alpha \exp(-t) \beta dt \\
&= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} \beta^\alpha t^\alpha \exp(-t) \beta dt \\
&= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} \beta^{\alpha+1} t^\alpha \exp(-t) dt \\
&= \frac{\beta^{\alpha+1}}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty t^\alpha \exp(-t) dt \\
&= \frac{\beta}{\Gamma(\alpha)} \int_0^\infty t^\alpha \exp(-t) dt \\
&= \frac{\beta\Gamma(\alpha+1)}{\Gamma(\alpha)} \\
&= \frac{\beta\alpha\Gamma(\alpha)}{\Gamma(\alpha)} \\
&= \alpha\beta
\end{aligned}$$

For the Gamma distribution, we will not calculate the second moment. Instead we will go straight to the variance calculation and calculate it directly. We will use the same substitution and use the definition of the Gamma function again.

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 &= \int_0^\infty x^2 \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta) dx - \alpha^2\beta^2 \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha+1} \exp(-x/\beta) dx - \alpha^2\beta^2 \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty (\beta t)^{\alpha+1} \exp(-t) \beta dt - \alpha^2\beta^2 \\
&= \frac{\beta^{\alpha+2}}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty t^{\alpha+1} \exp(-t) dt - \alpha^2\beta^2
\end{aligned}$$

Using the definition of the Gamma function.

$$
\begin{aligned}
&= \frac{\beta^2\Gamma(\alpha+2)}{\Gamma(\alpha)} - \alpha^2\beta^2 \\
&= \frac{\beta^2\Gamma(\alpha+2) - \alpha^2\beta^2\Gamma(\alpha)}{\Gamma(\alpha)} \\
&= \frac{\beta^2\big(\Gamma(\alpha+2) - \alpha^2\Gamma(\alpha)\big)}{\Gamma(\alpha)} \\
&= \frac{\beta^2\big((\alpha+1)\Gamma(\alpha+1) - \alpha^2\Gamma(\alpha)\big)}{\Gamma(\alpha)} \\
&= \frac{\beta^2\big((\alpha+1)\alpha\Gamma(\alpha) - \alpha^2\Gamma(\alpha)\big)}{\Gamma(\alpha)} \\
&= \frac{\beta^2\Gamma(\alpha)\big((\alpha+1)\alpha - \alpha^2\big)}{\Gamma(\alpha)} \\
&= \beta^2\big((\alpha+1)\alpha - \alpha^2\big) \\
&= \beta^2\big(\alpha^2\alpha - \alpha^2\big) \\
&= \alpha\beta^2
\end{aligned}
$$

And the calculation for the Gamma distribution has been completed.

- **Beta** with parameter α, β. Possible values: $\mathcal{X} = (0, 1)$.

$$
f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}
$$

In the following, we will use the following definition:

$$
\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt
$$

Expectation.

$$
\begin{aligned}
\mathbb{E}[X] = \int_{\mathcal{X}} x \cdot f(x) &= \int_0^1 x \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1}dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} \\
&= \frac{\alpha}{\alpha+\beta} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}
$$

Variance. Just like for the Gamma distribution, we calculate the second moment directly in the variance calculation. Will use a lot of the same tricks as when calculating the expectation.

$$
\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 &= \int_0^1 x^2 \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} dx - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+1}(1-x)^{\beta-1} dx - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+1}(1-x)^{\beta-1} dx - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \qquad (3.12.1) \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \frac{\alpha^2+\alpha}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{(\alpha^2+\alpha)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{\alpha^3+\alpha^2\beta+\alpha^2+\alpha\beta-\alpha^3-\alpha^2\beta-\alpha^2)}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{\cancel{\alpha^3}+\cancel{\alpha^2\beta}+\cancel{\alpha^2}+\alpha\beta\cancel{-\alpha^3}\cancel{-\alpha^2\beta}\cancel{-\alpha^2})}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}
\end{aligned}
$$

In step (3.12.1) we used the difference/chaining property of the Gamma function.

$$\Gamma(\alpha+2) = (\alpha+1)\Gamma(\alpha+1) = (\alpha+1)\alpha\Gamma(\alpha)$$

and,

$$\Gamma(\alpha+\beta+2) = (\alpha+\beta+1)\Gamma(\alpha+\beta+1) = (\alpha+\beta+1)(\alpha+\beta)\Gamma(\alpha+\beta)$$

This concludes the last calculation. I found this exercise to be a little tedious, to be honest...

3.13

A variable X is generated by assuming a value in $U(0,1)$ if a coin toss is 0 (H), and X is in $U(3,4)$ if the coin toss is 1 (T). The coin is fair, so each of these has a probability of 1/2.

(a) Calculating the mean of X. This will be a conditional expectation, and we define Y to be the coin toss. The marginal distributions might depend on the outcome, but when we calculate them:

$$f_{X|Y}(x|Y=0) = \frac{1}{1-0} = 1, \quad f_{X|Y}(x|Y=1) = \frac{1}{4-3} = 1,$$

we find that it will be $f_{X|Y}(x|y) = 1$ either way.

Now we can calculate the expected value for specific outcomes.

$$\mathbb{E}[X|Y=0] = \int_{\mathcal{X}} x \cdot f_{X|Y}(x|y) = \int_0^1 x dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1-0}{2} = \frac{1}{2}$$

$$\mathbb{E}[X|Y=1] = \int_{\mathcal{X}} x \cdot f_{X|Y}(x|y) = \int_3^4 x dx = \left[\frac{x^2}{2}\right]_3^4 = \frac{16-9}{2} = \frac{7}{2}$$

Using that $p_Y(y)$ is known, we can calculate the expected value of X:

$$\begin{aligned}\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] &= \sum_{y\in\{0,1\}} \mathbb{E}[X|Y=y]\mathbb{P}(Y=y) \\ &= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + \left(\frac{7}{2}\right)\left(\frac{1}{2}\right) \\ &= \frac{1}{4} + \frac{7}{4} = \frac{8}{4} = 2\end{aligned}$$

(b) Calculating the second moment.

$$\mathbb{E}[X^2|Y=0] = \int_{\mathcal{X}} x^2 \cdot f_{X|Y}(x|y) = \int_0^1 x^2 dx = \left[\frac{x^3}{3}\right]_0^1 = \frac{1-0}{3} = \frac{1}{3}$$

$$\mathbb{E}[X^2|Y=1] = \int_{\mathcal{X}} x^2 \cdot f_{X|Y}(x|y) = \int_3^4 x^2 dx = \left[\frac{x^3}{3}\right]_3^4 = \frac{4^3-3^3}{3} = \frac{37}{3}$$

$$\begin{aligned}\mathbb{E}[X^2] = \mathbb{E}[\mathbb{E}[X^2|Y]] &= \sum_{y\in\{0,1\}} \mathbb{E}[X^2|Y=y]\mathbb{P}(Y=y) \\ &= \left(\frac{1}{3}\right)\left(\frac{1}{2}\right) + \left(\frac{37}{3}\right)\left(\frac{1}{2}\right) \\ &= \frac{1}{6} + \frac{37}{6} = \frac{38}{6} \\ &= \frac{19}{3}\end{aligned}$$

Calculating the variance:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{19}{3} - 4 = \frac{7}{3} \approx 2.333$$

And finally the standard deviation:

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{7}{3}} \approx 1.5275$$

To verify these results, we did a numeric simulation.

```
N = 10000; Y = sample(c(0, 1), size=N, replace = TRUE); X = rep(0, N)
for (i in 1:N) {
     if (Y[i] == 1) X[i] = runif(1)
     else X[i] = runif(1, min=3, max=4)
}
> mean(X)
[1] 2.01424
> var(X)
[1] 2.332699
> sd(X)
[1] 1.527318
```

R

3.14

Claim: Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be random variables and $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ be constants. Then:

$$\mathrm{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_i Y_i\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \mathrm{Cov}(X_i, Y_j).$$

Proof. This can be verified with a direct proof. We will do a simplified example with $m = n = 2$, which is not a formal proof, but can easily be extended to one. We will repeatedly use:

$$\mathrm{Cov}(A, B) = \mathbb{E}[AB] - \mathbb{E}[A]\mathbb{E}[B].$$

Starting with the covariance, we apply the identity above. We have, for $m = n = 2$:

$$\begin{aligned}\mathrm{Cov}(a_1X_1 + a_2X_2, b_1Y_1 + b_2Y_2) &= \mathbb{E}[(a_1X_1 + a_2X_2)(b_1Y_1 + b_2Y_2)] - \mathbb{E}[a_1X_1 + a_2X_2]\mathbb{E}[b_1Y_1 + b_2Y_2] \\ &= S_1 - S_2\end{aligned}$$

Multiplying and rewriting the first term:

$$\begin{aligned}S_1 = \mathbb{E}[(a_1X_1 + a_2X_2)(b_1Y_1 + b_2Y_2)] &= \mathbb{E}[a_1b_1X_1Y_1 + a_1b_2X_1Y_2 + a_2b_1X_2Y_1 + a_2b_2X_2Y_2] \\ &= a_1b_1\mathbb{E}[X_1Y_1] + a_1b_2\mathbb{E}[X_1Y_2] + a_2b_1\mathbb{E}[X_2Y_1] + a_2b_2\mathbb{E}[X_2Y_2]\end{aligned}$$

The second term:

$$\begin{aligned}S_2 = \mathbb{E}[a_1X_1 + a_2X_2]\mathbb{E}[b_1Y_1 + b_2Y_2] &= (a_1\mathbb{E}[X_1] + a_2\mathbb{E}[X_2])\,(b_1\mathbb{E}[Y_1] + b_2\mathbb{E}[Y_2]) \\ &= a_1b_1\mathbb{E}[X_1]\mathbb{E}[Y_1] + a_1b_2\mathbb{E}[X_1]\mathbb{E}[Y_2] + a_2b_1\mathbb{E}[X_2]\mathbb{E}[Y_1] + a_2b_2\mathbb{E}[X_2]\mathbb{E}[Y_2]\end{aligned}$$

Subtracting:

$$\begin{aligned}S_1 - S_2 &= a_1b_1\mathbb{E}[X_1Y_1] + a_1b_2\mathbb{E}[X_1Y_2] + a_2b_1\mathbb{E}[X_2Y_1] + a_2b_2\mathbb{E}[X_2Y_2] \\ &\quad - a_1b_1\mathbb{E}[X_1]\mathbb{E}[Y_1] - a_1b_2\mathbb{E}[X_1]\mathbb{E}[Y_2] - a_2b_1\mathbb{E}[X_2]\mathbb{E}[Y_1] - a_2b_2\mathbb{E}[X_2]\mathbb{E}[Y_2] \\ &= a_1b_1(\mathbb{E}[X_1Y_1] - \mathbb{E}[X_1]\mathbb{E}[Y_1]) + a_1b_2(\mathbb{E}[X_1Y_2] - \mathbb{E}[X_1]\mathbb{E}[Y_2]) \\ &\quad + a_2b_1(\mathbb{E}[X_2Y_1] - \mathbb{E}[X_2]\mathbb{E}[Y_1]) + a_2b_2(\mathbb{E}[X_2Y_2] - \mathbb{E}[X_2]\mathbb{E}[Y_2]) \\ &= a_1b_1\mathrm{Cov}(X_1, Y_1) + a_1b_2\mathrm{Cov}(X_1, Y_2) + a_2b_1\mathrm{Cov}(X_2, Y_1) + a_2b_2\mathrm{Cov}(X_2, Y_2) \\ &= \sum_{i=1}^{2}\sum_{i=j}^{2} a_ib_j\mathrm{Cov}(X_i, Y_j)\end{aligned}$$

$\square$

3.15
Defining the joint distribution:

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{3}(x+y) & 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 2 \\ 0 & \text{otherwise} \end{cases}$$

We will calculate $\mathrm{Var}(2X - 3Y + 8)$. Constants disappear in the variance, and by Theorem 2.30:

$$\mathrm{Var}(2X - 3Y + 8) = 4\mathrm{Var}(X) + 9\mathrm{Var}(Y) - 12\mathrm{Cov}(X,Y)$$

Finding the marginal distributions:

$$f_X(x) = \int_0^2 \frac{1}{3}x + \frac{1}{3}y dy = \left[\frac{1}{3}xy + \frac{1}{6}y^2\right]_0^2 = \frac{2x}{3} + \frac{2}{3}$$

$$f_Y(y) = \int_0^1 \frac{1}{3}x + \frac{1}{3}y dx = \left[\frac{1}{6}x^2 + \frac{1}{3}xy\right]_0^1 = \frac{y}{3} + \frac{1}{6}$$

Calculating the expectation of X.

$$\mathbb{E}[X] = \int_0^1 x\left(\frac{2x}{3} + \frac{2}{3}\right) dx = \int_0^1 \left(\frac{2x^2}{3} + \frac{2x}{3}\right) dx = \left[\frac{2x^3}{9} + \frac{2x^2}{6}\right]_0^1 = \frac{5}{9}$$

Second moment:

$$\mathbb{E}[X^2] = \int_0^1 x^2\left(\frac{2x}{3} + \frac{2}{3}\right) dx = \int_0^1 \left(\frac{2x^3}{3} + \frac{2x^2}{3}\right) dx = \left[\frac{2x^4}{12} + \frac{2x^3}{9}\right]_0^1 = \frac{7}{18}$$

Variance of X:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{7}{18} - \frac{25}{81} = \frac{13}{162}$$

Calculating the expectation of Y.

$$\mathbb{E}[Y] = \int_0^2 y\left(\frac{y}{3} + \frac{1}{6}\right) dy = \int_0^2 \left(\frac{y^2}{3} + \frac{y}{6}\right) dx = \left[\frac{y^3}{9} + \frac{y^2}{12}\right]_0^2 = \frac{11}{9}$$

Second moment:

$$\mathbb{E}[Y^2] = \int_0^2 y^2\left(\frac{y}{3} + \frac{1}{6}\right) dy = \int_0^2 \left(\frac{y^3}{3} + \frac{y^2}{6}\right) dx = \left[\frac{y^4}{12} + \frac{y^3}{18}\right]_0^2 = \frac{16}{9}$$

Variance of Y:

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{16}{9} - \frac{121}{81} = \frac{23}{81}$$

From the definition of the covariance - these calculations are pretty large, so skipping the details:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
&= \int_0^2\int_0^1 \left(x-\frac{5}{9}\right)\left(y-\frac{11}{9}\right)\left(\frac{x}{3}+\frac{y}{3}\right)dxdy \\
&= \frac{1}{486}\int_0^2 -9y^2+20y-11dy \\
&= -\frac{1}{81}
\end{aligned}$$

Putting it all together:

$$\begin{aligned}
\mathrm{Var}(2X+3Y-8) &= 4\mathrm{Var}(X)+9\mathrm{Var}(Y)-12\mathrm{Cov}(X,Y) \\
&= 4\left(\frac{13}{162}\right)+9\left(\frac{23}{81}\right)-12\left(-\frac{1}{81}\right) \\
&= \frac{52}{162}+\frac{207}{81}+\frac{12}{81} \;=\; \frac{26}{81}+\frac{207}{81}+\frac{12}{81} \\
&= \frac{245}{81}
\end{aligned}$$

3.16

Let $r(x)$ and $s(y)$ be functions of x and y. Then:

$$\mathbb{E}[r(X)s(Y)|X] = r(X)\mathbb{E}[s(Y)|X], \qquad \mathbb{E}[r(X)|X] = r(X)$$

PROOF. By definition of the conditional expectation in the continuous case.

$$\begin{aligned}
\mathbb{E}[r(X)s(Y)|X=x] &= \int_{\mathcal{Y}} r(x)s(y)f_{Y|X}(y)dy \\
&= r(x)\int_{\mathcal{Y}} s(y)f_{Y|X}(y)dy \\
&= r(X)\mathbb{E}[s(Y)|X]
\end{aligned}$$

For the special case, we can set $s(Y)=1$.

$$\begin{aligned}
\mathbb{E}[r(X)|X=x] &= \int_{\mathcal{Y}} r(x)(1)f_{Y|X}(y)dy \\
&= r(x)\int_{\mathcal{Y}}(1)f_{Y|X}(y)dy \\
&= r(X)\mathbb{E}[(1)|X] \\
&= r(X)
\end{aligned}$$

which proves the statement. It is similar in the discrete case. $\square$

3.17
Proving that
$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X]).$$

Proof. Will disregard the hints.

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}[(Y - \mathbb{E}[Y])^2] \\
&= \mathbb{E}\big[((Y - \mathbb{E}[Y|X]) + (\mathbb{E}[Y|X] - \mathbb{E}[Y]))^2\big] \\
&= \mathbb{E}\Big[(Y - \mathbb{E}[Y|X])^2 + 2(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - \mathbb{E}[Y]) + (\mathbb{E}[Y|X] - \mathbb{E}[Y])^2\Big] \\
&= \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] + 2\mathbb{E}[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - \mathbb{E}[Y])] + \mathbb{E}[(\mathbb{E}[Y|X] - \mathbb{E}[Y])^2] \\
&= E_1 + 2E_2 + E_3
\end{aligned}$$

Consider each of these in turn. Using the definition of the conditional variance, with $\mu(x) = \mathbb{E}[Y|X]$.

$$E_1 = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] = \mathbb{E}\big[\mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X]\big] = \mathbb{E}\Big[\int_{\mathcal{Y}} (y - \mu(x))^2 f(y|x)dy\Big] = \mathbb{E}[\mathrm{Var}(Y|X)]$$

For E_3, we define $M = \mathbb{E}[Y|X]$, and $\mu_M = \mathbb{E}[M] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$. On this form, we see that we can use the definition of variance.

$$E_3 = \mathbb{E}[(\mathbb{E}[Y|X] - \mathbb{E}[Y])^2] = \mathbb{E}[(M - \mu_M)^2] = \mathrm{Var}(M) = \mathrm{Var}(\mathbb{E}[Y|X])$$

Finally, we consider E_2. Using iterated expectation and conditioning on X, then $\mathbb{E}[Y|X]$ and $\mathbb{E}[Y]$ become constants wrt. expectation. Define $C := \mathbb{E}[Y|X] - \mathbb{E}[Y]$ in order to clarify.

$$\begin{aligned}
\mathbb{E}[E_2] &= \mathbb{E}\Big(\mathbb{E}\big[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - \mathbb{E}[Y])\big|X\big]\Big) \\
&= \mathbb{E}\Big(\mathbb{E}\big[(Y - \mathbb{E}[Y|X])C\big|X\big]\Big) \\
&= \mathbb{E}\Big(C\mathbb{E}\big[(Y - \mathbb{E}[Y|X])\big|X\big]\Big) \\
&= \mathbb{E}\Big(C(\mathbb{E}[Y|X] - \mathbb{E}[Y|X])\Big) \\
&= \mathbb{E}\Big(C \cdot 0\Big) \\
&= 0
\end{aligned}$$

Collecting all the terms:

$$\begin{aligned}
\mathrm{Var}(Y) &= E_1 + 2E_2 + E_3 \\
&= \mathbb{E}[\mathrm{Var}(Y|X)] + 2(0) + \mathrm{Var}(\mathbb{E}[Y|X]) \\
&= \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])
\end{aligned}$$

and the result has been proved. $\square$

3.18
If $\mathbb{E}[X|Y=y]=c$ for some constant c, then X and Y are uncorrelated.

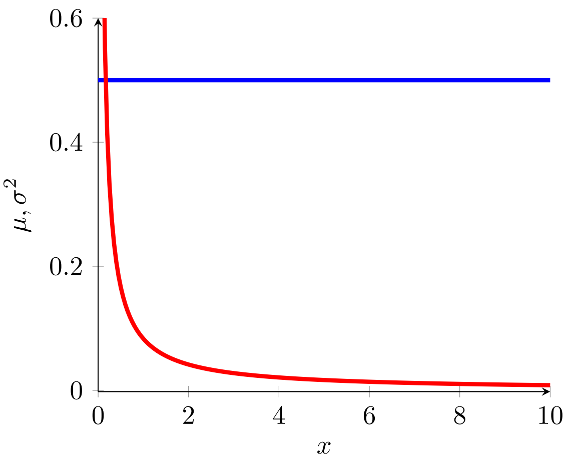
Proof. Assuming $\mathbb{E}[X|Y=y]=c$ for a constant c. Calculating the covariance:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]\\ &= \mathbb{E}[\mathbb{E}[XY|Y]]-\mathbb{E}[\mathbb{E}[X|Y]]\mathbb{E}[Y]\\ &= \mathbb{E}[\mathbb{E}[X|Y]Y]-\mathbb{E}[\mathbb{E}[X|Y]]\mathbb{E}[Y]\\ &= \mathbb{E}[cY]-\mathbb{E}[c]\mathbb{E}[Y]\\ &= c\mathbb{E}[Y]-c\mathbb{E}[Y]\\ &= 0\end{aligned}$$

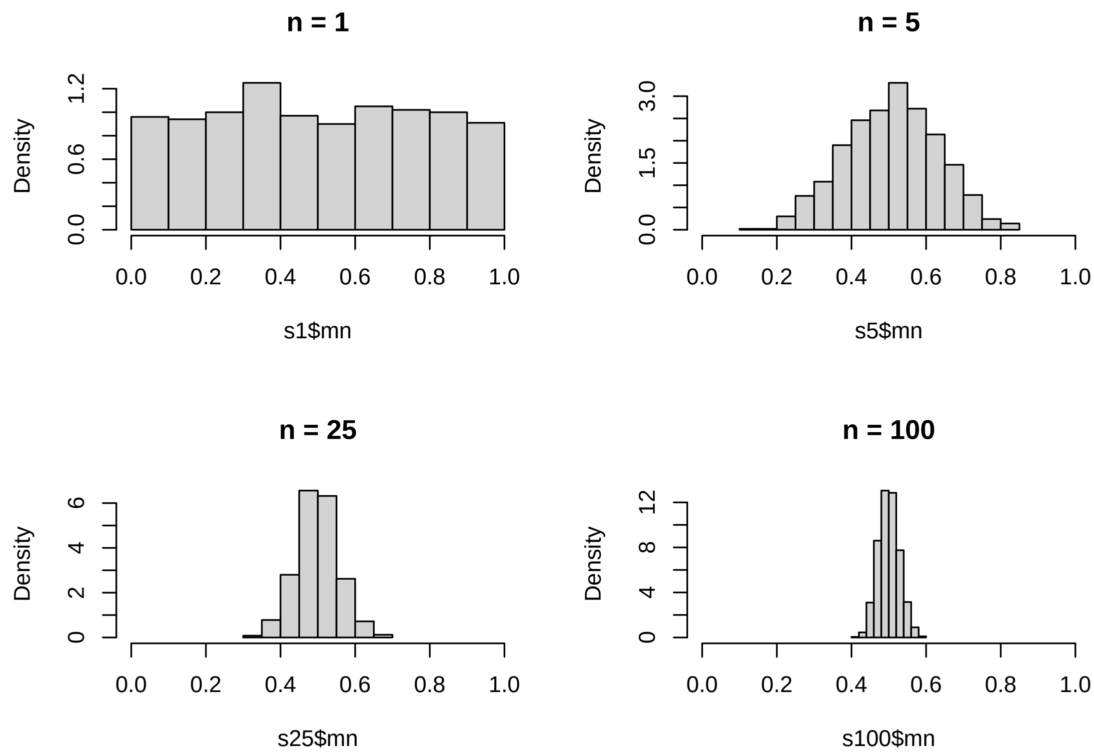
Since the covariance is 0, then it follows that the correlation is 0. □

3.19
Studying the sample distribution for the Uniform(0,1) distribution. If $X\sim U(0,1)$ then $\mathbb{E}[X]=1/2$ and $\mathrm{Var}(X)=1/12$. Will not plot f_X since it will simply be the straight line at $y=1$.

From Theorem 3.17, we found that $\mathbb{E}[barX_n]=\mu=1/2$ and $\mathrm{Var}(\overline{X}_n)=\sigma^2/n=1/(12n)$. Plotting these as functions of n, mean as blue and variance as red. The mean remains at 1/2, but the variance will become a lot smaller as n increases.



For each $n=1,5,25,100$ we do a 1000 simulations of each n and make a histogram of the results that can be seen on the next page. The results are centered around the mean, 1/2, but the spread becomes smaller as we increase n. We are basically seeing the central limit theorem in action.




```
simSampDist <- function(numSim) {
  retMean = rep(0, 1000); retVar = rep(0, 1000)
  for(i in 1:1000) {
    sim = runif(numSim)
    retMean[i]  = mean(sim)
    retVar[i]  = var(sim)
  }
  retVal = list()
  retVal$mn = retMean
  retVal$vr = retVal
  return(retVal)
}
s1 = simSampDist(1)
s5 = simSampDist(5)
s25 = simSampDist(25)
s100 = simSampDist(100)
# Creating plots
par(mfrow = c(2,2))
hist(s1$mn, main = "n = 1", xlim=c(0,1), prob = TRUE)
hist(s5$mn, main = "n = 5", xlim=c(0,1), prob = TRUE)
hist(s25$mn, main = "n = 25", xlim=c(0,1), prob = TRUE)
hist(s100$mn, main = "n = 100", xlim=c(0,1), prob = TRUE)
```

R

3.20
If a is a vector and X is a random vector with mean μ and variance Σ, then $\mathbb{E}[a^T X] = a^T \mu$ and $\mathrm{Var}(a^T X) = a^T \Sigma a$. If A is a matrix then $\mathbb{E}[AX] = A\mu$ and $\mathrm{Var}(AX) = A\Sigma A^T$.

Proof.

$$\mathbb{E}[a^T X] = \begin{bmatrix} \mathbb{E}[a_1 X_1] \\ \mathbb{E}[a_2 X_2] \\ \vdots \\ \mathbb{E}[a_n X_n] \end{bmatrix} = \begin{bmatrix} a_1 \mathbb{E}[X_1] \\ a_2 \mathbb{E}[X_2] \\ \vdots \\ a_n \mathbb{E}[X_n] \end{bmatrix} = \begin{bmatrix} a_1 \mu_1 \\ a_2 \mu_2 \\ \vdots \\ a_n \mu_n \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}^T \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = a^T \mu$$

The variance of a vector is the covariance matrix.

$$\mathrm{Var}(a^T X) = \begin{bmatrix} \mathrm{Var}(a_1 X_1) & \mathrm{Cov}(a_1 X_1, a_2 X_2) & \cdots & \mathrm{Cov}(a_1 X_1, a_n X_n) \\ \mathrm{Cov}(a_2 X_2, a_1 X_1) & \mathrm{Var}(a_2 X_2) & \cdots & \mathrm{Cov}(a_2 X_2, a_n X_n) \\ \vdots & \vdots & \ddots & \\ \mathrm{Cov}(a_n X_n, a_1 X_1) & \mathrm{Cov}(a_m X_m, a_2 X_2) & \cdots & \mathrm{Var}(a_n X_n) \end{bmatrix}$$

$$= \begin{bmatrix} a_1 \mathrm{Var}(X_1) a_1 & a_1 \mathrm{Cov}(X_1, X_2) a_2 & \cdots & a_1 \mathrm{Cov}(X_1, X_n) a_n \\ a_2 \mathrm{Cov}(X_2, X_1) a_1 & a_2 \mathrm{Var}(X_2) a_2 & \cdots & a_2 \mathrm{Cov}(X_2, X_n) a_n \\ \vdots & \vdots & \ddots & \\ a_n \mathrm{Cov}(X_n, X_1) a_1 & a_n \mathrm{Cov}(X_n, X_2) a_2 & \cdots & a_n \mathrm{Var}(X_n) a_n \end{bmatrix}$$

$$= \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}^T \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Var}(X_n) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a^T \Sigma a$$

The same general idea applies to matrices as well, execpt the notation becomes more extensive. Will skip this for brevity. □

3.21
Let X and Y be random variables. If $\mathbb{E}[Y|X] = X$, then $\mathrm{Cov}(X, Y) = \mathrm{Var}(X)$.

Proof. Applying the covariance identity and using iterated expectation:

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[\mathbb{E}[XY|X]] - \mathbb{E}[X]\mathbb{E}[\mathbb{E}[Y|X]] \\ &= \mathbb{E}[X\mathbb{E}[Y|X]] - \mathbb{E}[X]\mathbb{E}[X] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \mathrm{Var}(X) \end{aligned}$$

and the result has been proved. □

3.22

Let $X \sim U(0,1)$ and $0 < a < b < 1$. Define:

$$Y = \begin{cases} 1 & 0 < x < b \\ 0 & \text{otherwise} \end{cases}, \qquad Z = \begin{cases} 1 & a < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Are Y and Z independent? One way of thinking about independence is whether knowing something about Y tells us anything about Z, or vice-versa. In this case it doesn't seem to be that way. If we only know that Z is one, there is no way to determine whether $x < b$ or $x > b$, so we don't know what Y will be. By this reasoning, Y and Z are independent. To calculate it, let's set $a = 1/4$ and $b = 3/4$. Then:

	$Y=0$	$Y=1$	
$Z=0$	$1/16$	$3/16$	$1/4$
$Z=1$	$3/16$	$9/16$	$3/4$
	$1/4$	$3/4$	1

To determine e.g. $\mathbb{P}(Y = 0 \cap Z = 0)$ we first determine $\mathbb{P}(Y = 0|Z = 0)$ which is $1/4$, and then we determine $\mathbb{P}(Z = 0)$ which is $1/4$ which makes $\mathbb{P}(Y = 0 \cap Z = 0) = 1/16$. Now we calculate that $\mathbb{P}(Y = 0)\mathbb{P}(Z = 0) = (1/4)(1/4) = 1/16 = \mathbb{P}(Y = 0 \cap Z = 0)$ etc. which verifies that we have independence.

(b) Finding $\mathbb{E}[Y|Z]$. By definition of the iterated expectation and using independence ($\mathbb{P}(Y = y|Z = z) = \mathbb{P}(Y = y)$).

$$\mathbb{E}[Y|Z = 0] = \sum_y y \cdot \mathbb{P}(Y = y|Z = 0) = 0 + (1)\mathbb{P}(Y = 1|Z = 0) = \mathbb{P}(Y = 1) = b$$
$$\mathbb{E}[Y|Z = 1] = \sum_y y \cdot \mathbb{P}(Y = y|Z = 1) = 0 + (1)\mathbb{P}(Y = 1|Z = 1) = \mathbb{P}(Y = 1) = b$$

The last step follows because we know $x \sim U(0,1)$, so the probability is equal to the length of the interval.

Because of independence, Z does not affect the outcome. And in fact, there is a result that says $\mathbb{E}[Y|Z] = \mathbb{E}[Y]$ when Y is independent of Z, which we have almost proved above.

3.23

Finding the moment generating function for the Poisson, Normal and Gamma distributions. Recall the definition of the moment generating function:

$$M(t) = \mathbb{E}[e^{tX}] = \begin{cases} \displaystyle\sum_x e^{tx}p(x) & \text{discrete with pmf } p(x) \\ \displaystyle\int_{-\infty}^{\infty} e^{tx}f(x)dx & \text{continuous with pdf } f(x) \end{cases}$$

As usual with the Poisson distribution, the following identity is useful:

$$\sum_{k=0}^{\infty} \frac{a^n}{n!} = e^a$$

• **Poisson**. Assuming some arbitrary λ.

$$\begin{aligned}
M(t) &= \mathbb{E}[e^{tX}] \\
&= \sum_{n=0}^{\infty} e^{-\lambda} \frac{e^{tn}\lambda^n}{n!} \\
&= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} \\
&= e^{-\lambda} e^{\lambda e^t} \\
&= e^{\lambda e^t - \lambda} \\
&= e^{\lambda(e^t - 1)} \\
&= \exp\big(\lambda(e^t - 1)\big)
\end{aligned}$$

• **Normal**. Assuming some arbitrary μ and σ^2. The usual way of deriving the MGF for a normal distribution is to first do it for the standard normal distribution, since this is a lot easier. Denote this by $M_Z(t)$. We will 'compelte the square'.

$$\begin{aligned}
M_Z(t) &= \mathbb{E}[e^{tZ}] \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{1}{2}x^2} \, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(tx) \exp\left(-\frac{1}{2}x^2\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}x^2 + tx\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}x^2 + tx - \frac{t^2}{2} + \frac{t^2}{2}\right) dx \\
&= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x^2 - 2tx + t^2)\right) dx \\
&= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x - t)^2\right) dx \\
&= e^{t^2/2}
\end{aligned}$$

The integration is done by substituting $u = x - t$ which leads to the usual standard pdf, which integrates to 1. Hence, we have the MGF for the standard normal distribution. Using the MGF for the standard normal distribution, we can find it for a general normal distribution.

$$\begin{aligned}
M_X(t) &= \mathbb{E}[e^{tX}] = \mathbb{E}[e^{t(\mu + \sigma Z)}] \\
&= \mathbb{E}[e^{t\mu + t\sigma Z)}] = e^{t\mu}\mathbb{E}[e^{t\sigma Z)}] \\
&= e^{t\mu} M_Z(t\sigma) = e^{t\mu} e^{(t\sigma)^2/2} \\
&= \exp\left\{\frac{\sigma^2 t^2}{2} + \mu t\right\}
\end{aligned}$$

So tedious!!

- **Gamma**. Assuming some arbitrary α and β.

$$\begin{aligned}
M(t) &= \mathbb{E}[e^{tX}] \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty e^{tx} x^{\alpha-1} \exp(-x/\beta) dx \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} \exp(tx - x/\beta) dx \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} \exp(-(\frac{1}{\beta} - t)x) dx \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot \frac{\Gamma(\alpha)}{\left(\frac{1}{\beta} - t\right)^\alpha} \\
&= \frac{1}{\left[\beta\left(\frac{1}{\beta} - t\right)\right]^\alpha} \\
&= \frac{1}{[1 - \beta t]^\alpha} \\
&= (1 - \beta t)^{-\alpha}
\end{aligned}$$

where we used the definition of the Gamma function like in an earlier exercise.

3.24

Let $X_1, \dots, X_n \sim \text{Exp}(\beta)$. Find the MGF of X_i. Prove that $\sum_{i=1}^n X_i \sim \text{Gamma}(n, \beta)$.

Finding the MGF for the exponential distribution with parameter β.

$$\begin{aligned}
M(t) &= \mathbb{E}[e^{tX}] \\
&= \frac{1}{\beta} \int_0^\infty \exp(tx) \exp(-x/\beta) dx \\
&= \frac{1}{\beta} \int_0^\infty \exp(-(1/\beta - t)x) dx \\
&= \frac{1}{\beta} \left(\frac{1}{\frac{1}{\beta} - t} \right) \\
&= \frac{1}{\beta} \left(\frac{\beta}{1 - \beta t} \right) \\
&= \frac{1}{1 - \beta t}
\end{aligned}$$

All X_i have the same MGF. If we now assume they are independent, and define $Y = \sum_{i=1}^n X_i$, then according to Lemma 3.31:

$$M_Y(t) = \prod_{i=1}^n M_X(t) = \prod_{i=1}^n \left(\frac{1}{1 - \beta t} \right) = \frac{1}{(1 - \beta t)^n} = (1 - \beta t)^{-n}$$

which is the MGF for a Gamma distribution with parameters n and β, as we can see from the previous exercise. Hence, $Y \sim \text{Gamma}(n, \beta)$.

4 Inequalities

Practical Example

Example 1 - Markov's Inequality.
You hear that the mean age of NYU students is 20 years, but you know quite a few students that are older than 30. You decide to apply Markov's inequality to bound the fraction of students above 30 by modeling age as a nonnegative random variable A.

$$\mathbb{P}(A > 30) \leqslant \frac{\mathbb{E}[A]}{30} = \frac{2}{3},$$

so at most two thirds of the students are over 30. This isn't a very precise bound, but then we also only use the expectation.

Example 2 - Chebyshev's Inequality.
The previous bound is a little too weak. After investigating we discover that the standard deviation of student age is actually just 3 years. Applying Chebyshev's inequality to this information and obtain

$$\mathbb{P}(|A - \mathbb{E}[A]| > 10) \leqslant \frac{\mathrm{Var}(A)}{100} = \frac{9}{100}.$$

So at least 91% of the students are under 30 years old (and above 10).

Exercises

4.1
Let $X \sim \mathrm{Exp}(\beta)$. As we found in exercise 3.12, $\mu = \mathbb{E}[X] = \beta$ and $\sigma^2 = \mathrm{Var}(X) = \beta^2$. By applying Chebyshev's inequality:

$$\mathbb{P}(|X - \beta| \geqslant t) = \mathbb{P}(|X - \mu| \geqslant t) \leqslant \frac{\sigma^2}{t^2} = \frac{\beta^2}{t^2}$$

We are going to compare Chebyshev's inequality with the following inequality. Since $\sigma \geqslant 0$ and $k > 1$, we can apply Markov's inequality to get:

$$\mathbb{P}(|X - \mu| \geqslant k\sigma) = \mathbb{P}((X - \mu)^2 \geqslant k^2\sigma^2) \leqslant \frac{\mathbb{E}[(X - \mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

We recognize this as the corollarly to the Chebyshev inequality. Comparing, and using that $\sigma = \beta$, and using t instead of k:

$$\mathbb{P}(|X - \beta| \geqslant t) \leqslant \frac{\beta^2}{t^2}$$
$$\mathbb{P}(|X - \beta| \geqslant t\beta) \leqslant \frac{1}{t^2}$$

If $\beta > 1$, the first bound becomes 'weaker' which makes sense since $\beta t > t$ and we are considering a smaller interval. If $\beta < 1$ the reverse is true. Also interesting to note that by scaling up the bounded interval by β gives a β^2 increase in the bound, provided that $\beta > 1$. Shows there is a nonlinear relationship for the bound when using the Exponential distribution.

4.2
Let $X \sim \text{Poisson}(\lambda)$. As seen, $\mathbb{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$. Applying Chebyshev's inequality to show that $\mathbb{P}(X \geqslant 2\lambda) = 1/\lambda$.

$$\mathbb{P}(X \geqslant 2\lambda) = \mathbb{P}(X - \lambda \geqslant \lambda) \leqslant \mathbb{P}(|X - \lambda| \geqslant \lambda) \leqslant \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}.$$

Which shows that $\mathbb{P}(X \geqslant 2\lambda) \leqslant 1/\lambda$.

4.3
Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and define

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Bound $\mathbb{P}(|\overline{X} - p| > \epsilon)$ with Chebyshev's and Hoeffding's inequality. Show that when n is large, the Hoeffding's bound is smaller than Chebyshev's bound. (Assuming $\epsilon > 0$).

For the Bernoulli distribution, $\mathbb{E}[X_i] = p$ and $\text{Var}(X_i) = p(1-p)$. Since we are using a statistic, $\mathbb{E}[\overline{X}] = p$ and $\text{Var}(\overline{X}) = p(1-p)/n$ as per Theorem 3.17. Applying Chebyshev's inequality:

$$\mathbb{P}(|\overline{X} - p| > \epsilon) \leqslant \frac{p(1-p)/n}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leqslant \frac{1/4}{n\epsilon^2} = \frac{1}{4n\epsilon^2}$$

(Using that the max value of $p - p^2$ is $1/4$ for $p \in (0,1)$). Applying Hoeffding's inequality via Theorem 4.5.

$$\mathbb{P}(|\overline{X} - p| > \epsilon) \leqslant 2e^{-2n\epsilon^2} = \frac{2}{e^{2n\epsilon^2}}$$

To examine what happens as n grows, we divide the Hoeffding bound by the Chebyshev bound and call it L_n:

$$L_n = \left(\frac{2}{e^{2n\epsilon^2}}\right) \Big/ \left(\frac{1}{4n\epsilon^2}\right) = \frac{8n\epsilon^2}{e^{2n\epsilon^2}}$$

If we take the limit when $n \to \infty$, both the numerator and denominator diverges. So we apply L'Hôpital's rule and differentiate the numerator and denominator wrt. n:

$$\lim_{n\to\infty} L_n = \lim_{n\to\infty} \frac{8n\epsilon^2}{e^{2n\epsilon^2}} = \lim_{n\to\infty} \frac{8\epsilon^2}{2\epsilon^2 e^{2n\epsilon^2}} = 0$$

Since we lose the dependence on n in the numerator, the fraction goes to 0 as n grows. So as n becomes large, L_n becomes 0, which means that the Hoeffding bound becomes smaller than the Chebyshev bound.

4.4
Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$.

(a) For some $\alpha > 0$, define:

$$\epsilon = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)} \implies \epsilon^2 = \frac{1}{2n}\log\left(\frac{2}{\alpha}\right)$$

and set $\widehat{p} = n^{-1}\sum_{i=1}^n X_i$. Now we define $C_n = (\widehat{p} - \epsilon, \widehat{p} + \epsilon)$ and will show that the probability that C_n contains $\widehat{p}$ is $\alpha - 1$, using Hoeffding's inequality.

C_n containing p means that:

$$p \in C_n \implies \widehat{p} - \epsilon \leqslant p \leqslant \widehat{p} + \epsilon \implies -\epsilon \leqslant p - \widehat{p} \leqslant \epsilon \implies |\widehat{p} - p| \leqslant \epsilon$$

where we used that $|\widehat{p} - p| = |p - \widehat{p}|$ in the final step. So:

$$\mathbb{P}(p \in C_n) = \mathbb{P}(|\widehat{p} - p| \leqslant \epsilon) = 1 - \mathbb{P}(|\widehat{p} - p| > \epsilon)$$

(the probability that p is in the interval C_n is equal to 1 minus the probability that is outside the interval, basically). We can use the bound found in Theorem 4.5 again.

$$\begin{aligned}
\mathbb{P}(|\widehat{p} - p| > \epsilon) &\leqslant 2\exp\left\{-2n\epsilon^2\right\} \\
&= 2\exp\left\{-2n \cdot \frac{1}{2n}\log\left(\frac{2}{\alpha}\right)\right\} \\
&= 2\exp\left\{-\log\left(\frac{2}{\alpha}\right)\right\} \\
&= 2\exp\left\{\log\left(\frac{\alpha}{2}\right)\right\} \\
&= 2\left(\frac{\alpha}{2}\right) \\
&= \alpha
\end{aligned}$$

Rewriting this inequality in terms of $1 - \alpha$.

$$\begin{aligned}
\mathbb{P}(|\widehat{p} - p| > \epsilon) &\leqslant \alpha \\
1 + \mathbb{P}(|\widehat{p} - p| > \epsilon) &\leqslant 1 + \alpha \\
1 - \alpha &\leqslant 1 - \mathbb{P}(|\widehat{p} - p| > \epsilon)
\end{aligned}$$

Which shows the desired result:

$$1 - \alpha \leqslant 1 - \mathbb{P}(|\widehat{p} - p| > \epsilon) = \mathbb{P}(p \in C_n).$$

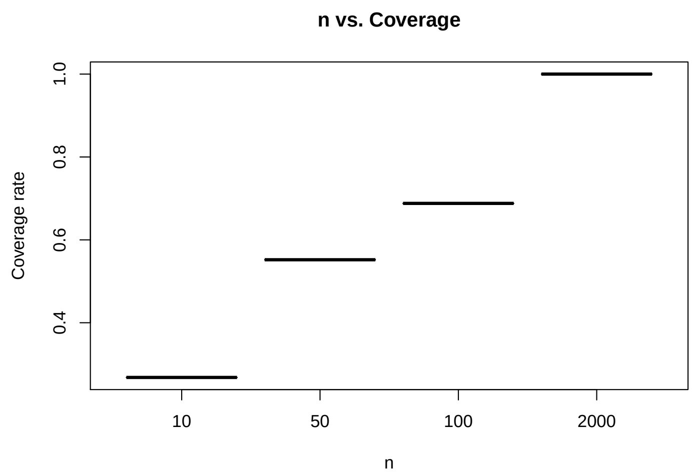
(b) Running some simulations. Doing 1000 simulations for $n = 10, 50, 100, 2000$ and checking how often the p lies in the range $\widehat{p} \pm \epsilon$.

See code and plots on the next page.

```
# 4.4(b): p = 0.4, alpha = 0.05
simBern <- function(n) {  # Simulate Bernoulli
    NSIM = 1000 # Do 1000 simulations per n
    simulationList = matrix(rep(0, n*NSIM), nrow = NSIM)
    for (i in 1:NSIM) {
        simulationList[i,1:n] = sample(c(1, 0), size=n,
                                       replace = TRUE, prob = c(0.4, 0.6))
    }
    return(simulationList)
}
countCoverage <- function(sim, alpha) {
    N = ncol(sim); NSIM = nrow(sim)
    covRate = rep(0, NSIM)
    for (i in 1:NSIM) {
        mn = mean(sim[i, 1:N])
        if(abs(0.4 - mn) < 0.05) {
            covRate[i] = 1
        }
    }
    hitRate = sum(covRate)/NSIM;print(hitRate)
}
# Bernoulli simulations
bsim10 = simBern(10);bsim50 = simBern(50)
bsim100 = simBern(100);bsim2000 = simBern(2000)
> countCoverage(bsim10, alpha=0.05)
[1] 0.268
> countCoverage(bsim50, alpha=0.05)
[1] 0.552
> countCoverage(bsim100, alpha=0.05)
[1] 0.684
> countCoverage(bsim2000, alpha=0.05)
[1] 1
```

R

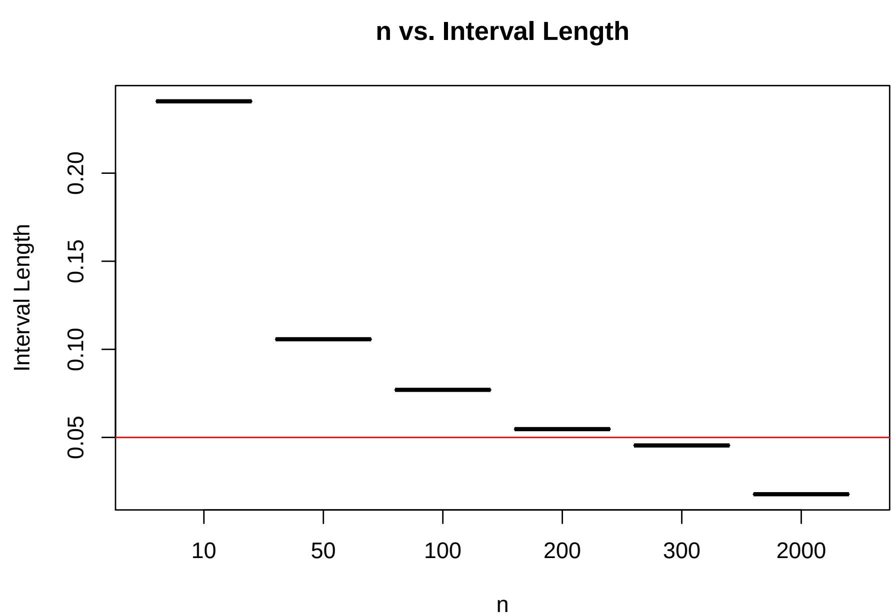
Plotting the coverage rate vs. n. As n increases, the interval will contain p more often.



(c) Investigating the relationship between n and the interval length: $2|\widehat{p}-p|$. When does it become smaller than 0.05? According to the simulation results, that happens roughly when n passes 200, based on our 1000-per-n simulation.

```
# 4.4(c): Plotting length of interval
IntervalLength <- function(sim) {
    N = ncol(sim); NSIM = nrow(sim)
    intLength = rep(0, NSIM)
    for (i in 1:NSIM) {
        mn = mean(sim[i, 1:N])
        intLength[i] = 2*abs(0.4 - mn)
    }
    print(mean(intLength))
}
> IntervalLength(bsim10)
[1] 0.2408
> IntervalLength(bsim50)
[1] 0.10572
> IntervalLength(bsim100)
[1] 0.077
> IntervalLength(bsim200)
[1] 0.0547
> IntervalLength(bsim300)
[1] 0.04543333
> IntervalLength(bsim2000)
[1] 0.017746
```

R



4.5 - Theorem 4.7: Mill's Inequality
Let $Z \sim N(0,1)$ and $t > 0$. Then,

$$\mathbb{P}(|Z| > t) \leqslant \sqrt{\frac{2}{\pi}} \cdot \frac{e^{-t^2/2}}{t}.$$

Proof. From the hint, we can use the symmetry property of the normal distribution, and when $t > 0$:

$$\mathbb{P}(|Z| > t) = 2\mathbb{P}(Z > t) \quad \Longrightarrow \quad \frac{1}{2}\mathbb{P}(|Z| > t) = \mathbb{P}(Z > t) \quad \Longrightarrow \quad \frac{t}{2}\mathbb{P}(|Z| > t) = t\mathbb{P}(Z > t)$$

As laid out in the proof of the Markov inequality, we get:

$$\frac{t}{2}\mathbb{P}(|Z| > t) = t\mathbb{P}(Z > t) = t\int_t^\infty f(x)dx \leqslant \int_t^\infty xf(x)dx = I$$

Calculating the integral I:

$$I = \int_t^\infty xf(x)dx = \frac{1}{\sqrt{2\pi}}\int_t^\infty x \cdot e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}\left[-e^{-x^2/2}\right]_t^\infty = \frac{1}{\sqrt{2\pi}}\Big(0 - (-e^{-t^2/2})\Big) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$$

Collecting our findings:

$$\frac{t}{2}\mathbb{P}(|Z| > t) \leqslant \frac{1}{\sqrt{2\pi}}e^{-t^2/2} \quad \Longrightarrow \quad \mathbb{P}(|Z| > t) \leqslant \sqrt{\frac{2}{\pi}}\frac{e^{-t^2/2}}{t}$$

which concludes the proof. $\square$

An additional note on the fraction manipulation.

$$\frac{2}{\sqrt{2\pi}} = \frac{\sqrt{2}\sqrt{2}}{\sqrt{2}\sqrt{\pi}} = \frac{\sqrt{2}}{\sqrt{\pi}} = \sqrt{\frac{2}{\pi}}$$

4.6
Let $Z \sim N(0,1)$. We simulate the probability that $\mathbb{P}(|Z| > t)$ and plot the results. See code and plot on next page.

We will also include the Markov bounds. Since the Markov inequality only applies to nonnegative random variables, we use it on $|Z|$. Finding $\mathbb{E}[|Z|]$ by using that the standard normal distribution is symmetric, so we multiply the expected value for $\mathbb{E}[Z > 0]$ by 2.

$$\begin{aligned}
\mathbb{E}[|Z|] &= 2\left(\frac{1}{\sqrt{2\pi}}\int_0^\infty x \cdot e^{-x^2/2}dx\right) \\
&= \sqrt{\frac{2}{\pi}}\Big[-e^{-x^2/2}\Big]_0^\infty \\
&= \sqrt{\frac{2}{\pi}}\Big(0 - (-1)\Big) \\
&= \sqrt{\frac{2}{\pi}}
\end{aligned}$$

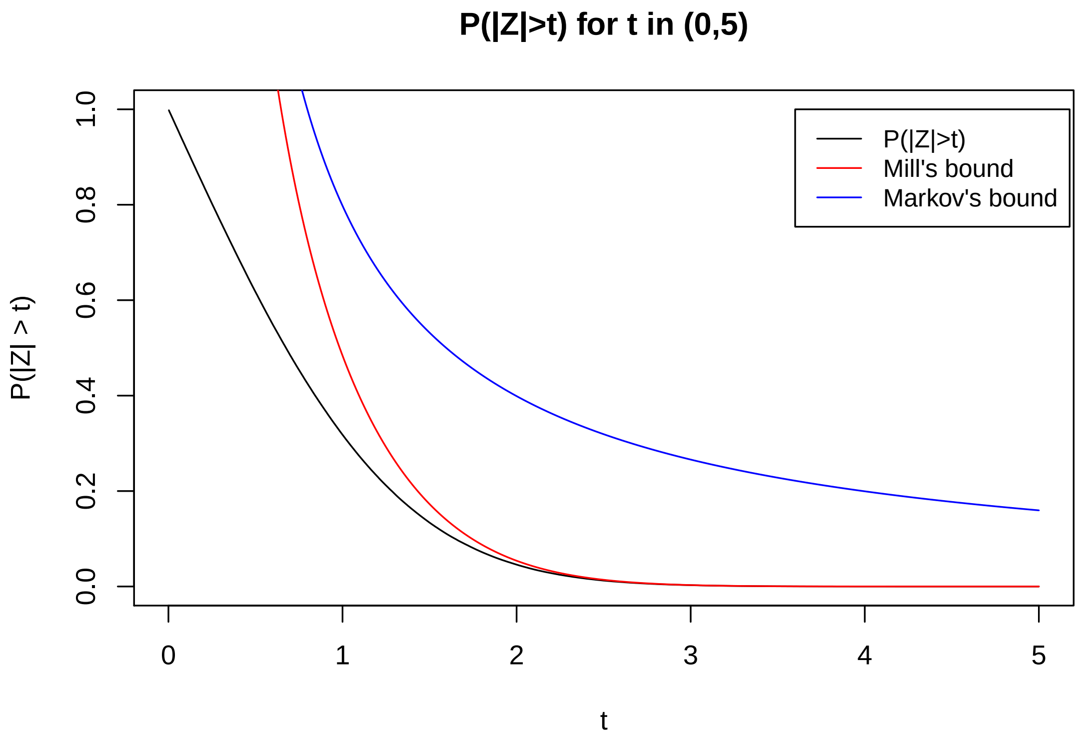
(This can easily be verified numerically).

```
# Plotting P(Z > t) as a function of t
Z = rnorm(1000000)
t = 1:2000/400 # t goes from 0 to 5
P_Zgt = rep(0, length(t))

for (i in 1:length(t)) {
    P_Zgt[i] = sum(abs(Z) > t[i])
}
plot(t, P_Zgt/1e6, type="l",
     main="P(|Z|>t) for t in (0,5)")
```

R

Results from plot, when including bounds for the Markov and Hill inequalities:



The Markov inequality gives a very lose bound, but as we can see the Mill inequality gives a bound that really hugs the $\mathbb{P}(|Z| > z)$ curve.

4.7
Let $X_1, \dots, X_n \sim N(0,1)$. Bound $\mathbb{P}(|\overline{X}| > t)$ using Mill's inequality, where $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ and compare to Chebyshev's bound.

We begin by applying the Chebyshev's inequality. The sampling distribution of $\overline{X}$ gives $\mathbb{E}[\overline{X}] = 0$ and $\text{Var}(\overline{X}) = \sigma^2/n = 1/n$.

$$\mathbb{P}(|\overline{X}| > t) = \mathbb{P}(|\overline{X} - \mu| > t) \leqslant \frac{\sigma^2}{t^2} = \frac{1}{nt^2}$$

Now, the variable $\overline{X}$ no longer has a standard normal distribution, as we noted above. We have $\overline{X} \sim N(0, 1/n)$. But we can do the following, since we only require $t > 0$:

$$\begin{aligned}\mathbb{P}(|X| > t) &= \mathbb{P}(|0 + (1/n)Z| > t)\\ &= \mathbb{P}(1/n|Z| > t)\\ &= \mathbb{P}(|Z| > nt)\end{aligned}$$

So, by defining $k = nt > 0$, we get:

$$\mathbb{P}(|X| > t) = \mathbb{P}(|Z| > k) \leqslant \sqrt{\frac{2}{\pi}}\frac{e^{-k^2/2}}{k} = \sqrt{\frac{2}{\pi}}\frac{e^{-(nt)^2/2}}{nt}$$

No particular observations, except that we can note Hill's inequality relies on the square of n, while Chebyshev's bound only uses n. By taking the limit and applying L'Hôpital's rule, we can easily see that Hill's bound will be smaller as n grows.

5 Convergence of Random Variables

Definition 5.1 Types of convergence
Let $X_1, X_2, \dots$ be a sequence of random variables and let X be another random variable. Let F_n denote the CDF of X_n and let F denote the CDF of X.

1. X_n converges to X in *probability*, $X_n \xrightarrow{P} X$, if for all $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0$$

as $n \to \infty$.

2. X_n converges to X in *distribution*, $X_n \rightsquigarrow X$, if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

at all t for which F is continuous.

Definition 5.2
X_n converges to X in *quadratic mean* (L_2), written $X_n \xrightarrow{qm} X$, if

$$\mathbb{E}[(X_n - X)^2] \to 0$$

as $n \to \infty$.

Relationship in convergence types.

$$\xrightarrow{qm} \implies \xrightarrow{P} \implies \rightsquigarrow$$

Exercises

5.1
Let $X_1, \dots, X_n$ be IID with finite mean and variance μ and σ^2. Let $\overline{X}$ be the sample mean and S_n^2 be the sample variance.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}, \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

(a) Showing that $\mathbb{E}[S_n^2] = \sigma^2$ was done in exercise 3.8.

(b) Showing that $S_n^2 \xrightarrow{P} \sigma^2$. Directly, this means that for any $\epsilon > 0$,

$$\mathbb{P}(|S_n^2 - \sigma^2| > \epsilon) = 0 \text{ as } n \to \infty.$$

But we can do it in a different way. Following the hint, we want to rewrite S_n^2 so we can apply the law of large numbers.

Doing the calculations on the next page.

$$
\begin{aligned}
S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left[X_i^2 - 2X_i\overline{X} + \overline{X}^2\right]\right)
\end{aligned}
$$

Distributing the sum.

$$
\begin{aligned}
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) + \overline{X}^2\right) \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X}^2 + \overline{X}^2\right) \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2\right)
\end{aligned}
$$

WLLN on $\overline{X}^2$ with $g(x) = x^2$.

$$
\begin{aligned}
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \mu^2\right) \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right)
\end{aligned}
$$

The last rewrite is a standard identity. Now, we define $Y_i = X_i - \mu$. Then $\overline{Y} \to \mathbb{E}[(X_i - \mu)]$ by the WLLN. By Theorem 5.5(f), we can apply $g(x) = x^2$ and get $\overline{Y}^2 \to \mathbb{E}[(X_i - \mu)^2] = \mathrm{Var}(X_i) = \sigma^2$. Finally, we use the hint with $c_n = n/(n-1)$ which obviously tends to 1, so we can apply Theorem 5.5(d) and hence we have proved that $S_n^2 \xrightarrow{P} \sigma^2$. (The hint says using 5.5(e), but that is convergence with distribution, which I don't think is correct). (Update: this is a misprint which is noted in the errata2.pdf for the book).

5.2

Let $X_1, X_2 \ldots$ be a sequence of random variables. Show that $X_n \xrightarrow{qm} b$ if and only if

$$
\lim_{n\to\infty} \mathbb{E}[X_n] = b, \quad \text{and} \quad \lim_{n\to\infty} \mathrm{Var}[X_n] = 0.
$$

Proof.

$\Rightarrow$) We assume $X_n \xrightarrow{qm} b$, which means

$$
\mathbb{E}[(X_n - b)^2] \to 0, \quad n \to \infty.
$$

We can rewrite the expression:

$$
\begin{aligned}
\mathbb{E}[(X_n - b)^2] &= \mathbb{E}[X_n^2 - 2bX_n + b^2] \\
&= \mathbb{E}[X_n^2] - 2b\mathbb{E}[X_n] + b^2 \\
&= \mathbb{E}[X_n^2] - \mathbb{E}[X_n]^2 + \mathbb{E}[X_n]^2 - 2b\mathbb{E}[X_n] + b^2 \\
&= \mathrm{Var}(X_n) + (\mathbb{E}[X_n] - b)^2
\end{aligned}
$$

By our assumption $\mathbb{E}[(X_n - b)^2] \to 0$ as $n \to \infty$, and so $\text{Var}(X_n) + (\mathbb{E}[X_n] - b)^2 \to 0$ as $n \to \infty$. Since $\text{Var}(X_n) \geqslant 0$ and $(\mathbb{E}[X_n] - b)^2 \geqslant 0$, then we can conclude that

$$\text{Var}(X_n) \to 0, \qquad (\mathbb{E}[X_n] - b)^2 \to 0 \quad \Longrightarrow \quad \mathbb{E}[X_n] \to b$$

as $n \to \infty$.
$\Leftarrow$) We assume

$$\lim_{n\to\infty} \mathbb{E}[X_n] = b, \quad \text{and} \quad \lim_{n\to\infty} \text{Var}[X_n] = 0.$$

Then $\lim_{n\to\infty} \text{Var}(X_n) + (\mathbb{E}[X_n] - b)^2 = 0$. By reversing the calculations from the first part:

$$\lim_{n\to\infty} \text{Var}(X_n) + (\mathbb{E}[X_n] - b)^2 = 0 \quad \Longrightarrow \quad \lim_{n\to\infty} \mathbb{E}[(X_n - b)^2] = 0 \quad \Longrightarrow \quad X_n \xrightarrow{qm} b.$$

By showing implication both ways, the result is proved. $\square$

5.3
Let $X_1, \dots, X_n$ be IID and let $\mu = \mathbb{E}[X_i]$ with finite variance. Show that $\overline{X} \xrightarrow{qm} \mu$.

Proof. Define:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

By the WLLN, $\overline{X}_n \xrightarrow{P} \mu$, which we can also express as

$$\lim_{n\to\infty} \mathbb{E}[\overline{X}_n] = \mu.$$

The variance is finite, so $\text{Var}(X_i) = \sigma^2 < \infty$. This means that for the sample variance:

$$\lim_{n\to\infty} \text{Var}(\overline{X}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$$

With this, we can simply apply the result from exercise 5.2 which shows that $\mathbb{E}[(\overline{X} - \mu)^2] \to 0$ as $n \to \infty$ which proves $\overline{X} \xrightarrow{qm} \mu$. $\square$

5.4
Let $X_1, X_2, \dots$ be a sequence of random variables such that

$$\mathbb{P}\left(X_n = \frac{1}{n}\right) = 1 - \frac{1}{n^2}, \quad \text{and} \quad \mathbb{P}(X_n = n) = \frac{1}{n^2}.$$

Just noting the probabilities for each outcome as n grows.

	$n=1$	$n=2$	$n=3$	$n=4$	$\dots$
$p(x) = \mathbb{P}(X_n = \frac{1}{n})$	$p(1) = 0$	$p(\frac{1}{2}) = \frac{3}{4}$	$p(\frac{1}{3}) = \frac{8}{9}$	$p(\frac{1}{4}) = \frac{15}{16}$	$\dots$
$p(x) = \mathbb{P}(X_n = n)$	$p(1) = 1$	$p(2) = \frac{1}{4}$	$p(3) = \frac{1}{9}$	$p(4) = \frac{1}{16}$	$\dots$

As n becomes large, the probability that we get n and not $1/n$ becomes very small. But as n becomes large, it will also yield extreme outliers in the sequence with a non-zero probability.

Starting by calculating the mean, second moment, and variance.

$$\begin{aligned}\mathbb{E}[X_n] &= \left(\frac{1}{n}\right)\left(1-\frac{1}{n^2}\right)+(n)\left(\frac{1}{n^2}\right)\\ &= \frac{1}{n}-\frac{1}{n^3}+\frac{1}{n}\\ &= \frac{2}{n}-\frac{1}{n^3}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X_n^2] &= \left(\frac{1}{n^2}\right)\left(1-\frac{1}{n^2}\right)+(n^2)\left(\frac{1}{n^2}\right)\\ &= \frac{1}{n^2}-\frac{1}{n^4}+1\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(X_n) &= \mathbb{E}[X_n^2]-\mathbb{E}[X_n]^2\\ &= \frac{1}{n^2}-\frac{1}{n^4}+1-\left(\frac{2}{n}-\frac{1}{n^3}\right)^2\\ &= \frac{1}{n^2}-\frac{1}{n^4}+1-\left(\frac{4}{n^2}-\frac{4}{n^4}+\frac{1}{n^6}\right)\\ &= 1-\frac{3}{n^2}+\frac{3}{n^4}-\frac{1}{n^6}\end{aligned}$$

Results verified by simulation. (See code in `5.4.R`). E.g. for $n = 200$ and 10M simulations:

```
# Simulated vs. Theoretical
> mean(Xn)
[1] 0.009879878
> 2/n - 1/n^3
[1] 0.009999875
> var(Xn)
[1] 0.9759275
> 1 - 3/n^2 + 3/n^4 - 1/n^6
[1] 0.999925
```

R

From these expressions we can see that $\lim_{n\to\infty}\mathbb{E}[X_n] = 0$, but $\lim_{n\to\infty}\mathrm{Var}(X_n) = 1$. Since the variance does not become 0, we know by the result in exercise 5.2 that this does NOT converge in quadratic mean.

Checking if X_n converges in probability. If we fix some $\epsilon > 0$, we can apply the Chebyshev inequality. Set $\mu = \mathbb{E}[X_n]$ and $\sigma^2 = \mathrm{Var}(X_n)$:

$$\mathbb{P}(|X_n-\mu|>\epsilon) \leqslant \frac{\sigma^2}{\epsilon^2} = \frac{1-\frac{3}{n^2}+\frac{3}{n^4}-\frac{1}{n^6}}{\epsilon^2}$$

By taking the limit $n\to\infty$ on both sides, we get:

$$\lim_{n\to\infty}\mathbb{P}(|X_n-\mu|>\epsilon) \leqslant \lim_{n\to\infty}\frac{1-\frac{3}{n^2}+\frac{3}{n^4}-\frac{1}{n^6}}{\epsilon^2} = \frac{1}{\epsilon^2}.$$

Since this does not tend to 0, we do NOT have convergence in probability.

5.5
Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Prove that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} p, \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{qm} p.$$

Proof. Recalling how to calculate the expectation and second moment for a Bernoulli(p) variable:

$$\mathbb{E}[X] = (1)p + (0)(p-1) = p, \qquad \mathbb{E}[X^2] = (1)^2 p + (0)^2(p-1) = p,$$

With these, we can calculate the variance.

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1-p).$$

We define $Y_i := X_i^2$ and exploit the simplicity of the Bernoulli distribution. In this case:

$$\mathbb{E}[Y] = \mathbb{E}[X^2] = (1)^2 p + (0)^2(p-1) = p, \qquad \mathbb{E}[Y^2] = \mathbb{E}[X^4] = (1)^4 p + (0)^4(p-1) = p,$$

With these, we can calculate the variance.

$$\text{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = p - p^2 = p(1-p).$$

So we have $\mu = \mathbb{E}[Y] = p$ and $\sigma^2 = p(1-p)$. We can now define the sample mean and variance:

$$\mathbb{E}[\overline{Y}] = p, \qquad \text{Var}(\overline{Y}) = \frac{\sigma^2}{n} = \frac{p(1-p).}{n}.$$

Since the variance tends to 0 as $n \to \infty$ we can use the results in exercise 5.3 to conclude that $\overline{Y} \xrightarrow{qm} p$ and by how we defined Y, $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{qm} p$. Since we have convergence in quadratic mean, it follows that we also have convergence in probability. $\square$

5.6
The height of men has mean 68 inches and standard deviation 2.6 inches. We have $n = 100$. Finding the approximate probability that the average height in the sample will be at least 68 inches.

We can approximate this probability with the CLT (central limit theorem). We want to find the probability that the sample height $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is at least as big as the population mean: $\mathbb{P}(\overline{X} > \mu)$.

$$\mathbb{P}(\overline{X} > \mu) = 1 - \mathbb{P}(\overline{X} \leqslant \mu)$$

By the central limit theorem, using that $n = 100$, $\mu = 68$ and $\sigma = 2.6$:

$$\mathbb{P}(\overline{X} \leqslant 68) = \mathbb{P}(\overline{X} - 68 \leqslant 0) = \mathbb{P}\left(\frac{10(\overline{X} - 68)}{2.6} \leqslant 0\right) \approx \mathbb{P}(Z \leqslant 0) = 0.5$$

(Since the standard normal distribution is symmetric and centered at 0). So we get:

$$\mathbb{P}(\overline{X} > 68) = 1 - \mathbb{P}(\overline{X} \leqslant 68) = 1 - 0.5 = 0.5.$$

5.7

Let $\lambda_n = \frac{1}{n}$ for all n and let $X_n \sim \text{Poisson}(\lambda_n)$.

(a) Showing that $X_n \xrightarrow{P} 0$. By properties of the Poisson distribution, we have

$$\mu = \mathbb{E}[X_n] = \frac{1}{n}, \qquad \sigma^2 = \text{Var}(X_n) = \frac{1}{n}.$$

By fixing some $\epsilon > 0$ and applying Chebyshev's inequality, we get:

$$\mathbb{P}(|X_n - \frac{1}{n}| > \epsilon) = \mathbb{P}(|X_n - \mu| > \epsilon) \leqslant \frac{\sigma^2}{\epsilon^2} = \frac{1}{n\epsilon^2}$$

By taking the limit on both sides:

$$\lim_{n\to\infty} \mathbb{P}(|X_n - \frac{1}{n}| > \epsilon) =\leqslant \lim_{n\to\infty} \frac{1}{n\epsilon^2} = 0,$$

and since $\mu = 1/n \to 0$ we have shown that $X_n \xrightarrow{P} 0$.

(b) We define $Y_n = nX_n$ and will show that $Y_n \xrightarrow{P} 0$. Finding the mean and variance:

$$\mathbb{E}[Y_n] = \mathbb{E}[nX_n] = n\mathbb{E}[X_n] = n\left(\frac{1}{n}\right) = 1$$

$$\text{Var}(Y_n) = \text{Var}(nX_n) = n^2\text{Var}(X_n) = n^2\left(\frac{1}{n}\right) = n$$

From the Chebyshev inequality we can only really conclude that Y_n does NOT converge to 1, but we can't use it for determining the asymptotic behavior of Y_n at 0. Instead we will use Theorem 5.5(f):$X_n \xrightarrow{P} 0$ then $g(X_n) \xrightarrow{P} g(X)$. Here, Y_n is a function of X_n, defined as: $Y_n = nX_n$ where $g(x) = nx$, which is a continuous function. In this case we get that $X_n \xrightarrow{P} 0$ implies $nX_n \xrightarrow{P} n \cdot 0$ so $Y_n \xrightarrow{P} 0$.

5.8

A program has $n = 100$ pages of code. Let $X_i \sim \text{Poisson}(1)$ be iid and denote the number of errors on page i. Let $Y = \sum_{i=1}^{n} X_i$ denote the total number of errors. Use the CLT to approximate $\mathbb{P}(Y < 90)$.

The mean and variance are $\mu = \mathbb{E}[X_i] = 1$ and $\sigma^2 = \text{Var}(X_i) = 1$. An important observation: the CLT applies to sample means, but in this case we are simply summing up 100 independent Poisson variables, and so $Y \sim \text{Poisson}(100)$, i.e. $\mathbb{E}[Y] = 100$ and $\text{Var}(Y) = 100$.

Let us define $W = \frac{1}{n}Y$, which means:

$$W = \frac{1}{n}Y = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

then by the CLT, $W \sim N(1, 1/100)$, so $\mathbb{E}[W] = 1$ and $\text{Var}(W) = 1/100$.

By going the other way, we can find a normal approximation for Y by using that $Y = nW$ where $n = 100$.:

$$\mathbb{E}[Y] = \mathbb{E}[nW] = n\mathbb{E}[W] = n(1) = 100,$$

$$\text{Var}(Y) = \text{Var}(nW) = n^2\text{Var}(W) = (100)^2 \left(\frac{1}{100}\right) = 100$$

The standard deviation is: $\sqrt{\text{Var}(Y)} = 10$. Approximating the probability.

$$\mathbb{P}(Y < 90) = \mathbb{P}(Y - 100 < -10) = \mathbb{P}\left(\frac{Y - 100}{10} < -1\right) \approx \mathbb{P}(Z \leqslant -1) \approx 0.1586$$

Note: this exercise demonstrates that the CLT is just an approximation, and under certain conditions it might not be a very accurate approximation. In `5.8.R` a simulation repeating the conditions were done one million times, and numerically, the probability that $\mathbb{P}(X < 90)$ turns out to be about 0.1467.

```
> # Approximating answer numerically
> length(Y)
[1] 1000000
> sum(Y < 90)/length(Y)
[1] 0.146773
```

R

5.9

Suppose that $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$ and define:

$$X_n = \begin{cases} X & \text{with probability} \quad 1 - \frac{1}{n} \\ e^n & \text{with probability} \quad \frac{1}{n} \end{cases}$$

We will determine what kinds of convergences it satisfies.

For starters, we can get some information about X. The expectation:

$$\mathbb{E}[X] = (-1)\left(\frac{1}{2}\right) + (1)\left(\frac{1}{2}\right) = 0$$

Then, the second moment and the variance.

$$\mathbb{E}[X^2] = (-1)^2\left(\frac{1}{2}\right) + (1)^2\left(\frac{1}{2}\right) = 1$$

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1 - 0 = 1$$

Next we can find the expectation of X_n. Since we are using X which is itself a random variable, we can't calculate it directly, instead we will use iterated expectation. Calculations are on the next page.

$$\mathbb{E}[X_n] = X\left(1 - \frac{1}{n}\right) + e^n\left(\frac{1}{n}\right)$$

$$\begin{aligned}\mathbb{E}[X_n] = \mathbb{E}[\mathbb{E}[X_n|X]] &= \mathbb{E}[X_n|X=1]\mathbb{P}(X=1) + \mathbb{E}[X_n|X=-1]\mathbb{P}(X=-1) \\ &= \left[(1)\left(1-\frac{1}{n}\right) + (e^n)\left(\frac{1}{n}\right)\right]\left(\frac{1}{2}\right) + \left[(-1)\left(1-\frac{1}{n}\right) + (e^n)\left(\frac{1}{n}\right)\right]\left(\frac{1}{2}\right) \\ &= \frac{e^n}{n}\end{aligned}$$

Calculating the second moment.

$$\begin{aligned}\mathbb{E}[X_n^2] = \mathbb{E}[\mathbb{E}[X_n^2|X]] &= \mathbb{E}[X_n^2|X=1]\mathbb{P}(X=1) + \mathbb{E}[X_n^2|X=-1]\mathbb{P}(X=-1) \\ &= \left[(1)^2\left(1-\frac{1}{n}\right) + (e^n)^2\left(\frac{1}{n}\right)\right]\left(\frac{1}{2}\right) + \left[(-1)^2\left(1-\frac{1}{n}\right) + (e^n)^2\left(\frac{1}{n}\right)\right]\left(\frac{1}{2}\right) \\ &= 1 - \frac{1}{n} + \frac{e^{2n}}{n}\end{aligned}$$

Variance.

$$\begin{aligned}\mathrm{Var}(X_n) &= \mathbb{E}[X_n^2] - \mathbb{E}[X_n]^2 \\ &= 1 - \frac{1}{n} + \frac{e^{2n}}{n} - \frac{e^{2n}}{n^2} \\ &= 1 - \frac{1}{n} + \left(\frac{1}{n} - \frac{1}{n^2}\right)e^{2n}\end{aligned}$$

Verified by simulation for $n = 5$. (See `5.9.R`).

```
> # Theoretical vs. Simulated
> mean(Xn)
[1] 29.72278
> exp(n)/n
[1] 29.68263
> var(Xn)
[1] 3528.798
> (1 - 1/n) + (1/n - 1/n^2)*exp(2*n)
[1] 3525.035
```

R

Both the expectation and variance will tend to infinity as $n \to \infty$, so we can conclude that X_n does NOT converge in probability, and it does NOT converge in quadratic mean. Even just the oscillating X wouldn't converge. The variance would be constant, so there would be no quadratic mean convergence by the results in exercise 5.2, and we would not be able to apply Chebyshev's inequality to prove convergene in probability.

Evaluating the density convergence later.

5.10
Let $Z \sim N(0,1)$ and let $t > 0$ Show that for any $k > 0$

$$\mathbb{P}(|Z| > t) \leqslant \frac{\mathbb{E}[|Z|^k]}{t^k}.$$

PROOF. Following the method used in the proof of Markov's inequality. Will also use the hint from 4.5: $\mathbb{P}(|Z| > t) = 2\mathbb{P}(Z > t)$ and what was found in 4.6 that the expected value of $|Z|$ is two times the integral from 0 to ∞.

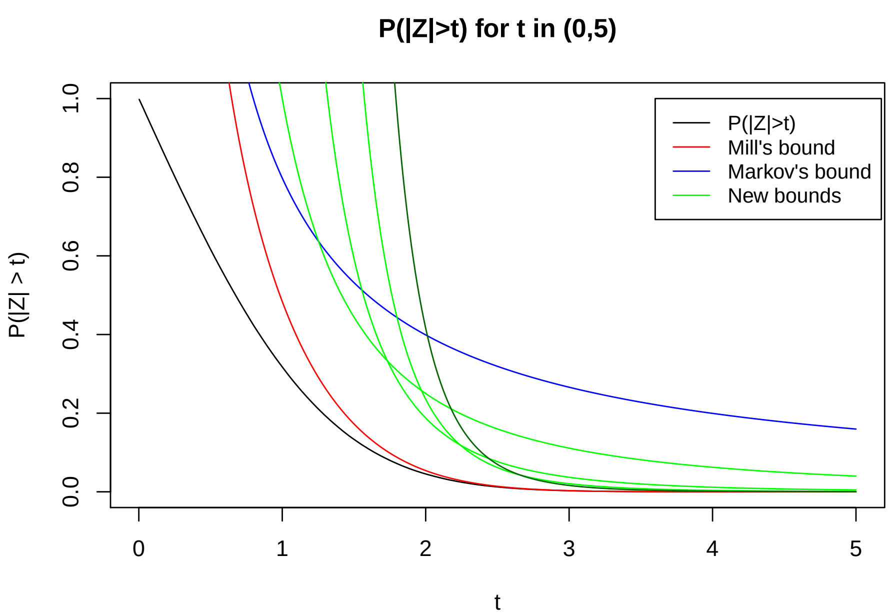
For any $t > 0$ and $k > 0$:

$$\begin{aligned}\mathbb{E}[|Z|^k] &= 2\int_0^\infty z^k f(z)dz = 2\int_t^\infty z^k f(z)dz + 2\int_0^t z^k f(z)dz \\ &\geqslant 2\int_t^\infty z^k f(z)dz \geqslant 2t^k \int_t^\infty f(z)dz = 2t^k\mathbb{P}(Z > t) = t^k\mathbb{P}(|Z| > t)\end{aligned}$$

Which implies,

$$\mathbb{P}(|Z| > t) \leqslant \frac{\mathbb{E}[|Z|^k]}{t^k} \qquad \square$$

Comparing to Mill's inequality and Markov's inequality by extending the plot used in exercise 6.4. Plotting the bounds for $k = 2, 4, 6, 8$. This method is not quite as precise as Mill's inequality, but it gets very close after $t > 3$. The bound is not as good for smaller t, where even the Markov bound is better. (The dark green line is $k = 8$, and the Markov bound corresponds to $k = 1$)



Code for making this plot is in `5.10.R`.

5.11
Let $X_n \sim N(0, 1/n)$ and let X have the following CDF:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geqslant 0 \end{cases}$$

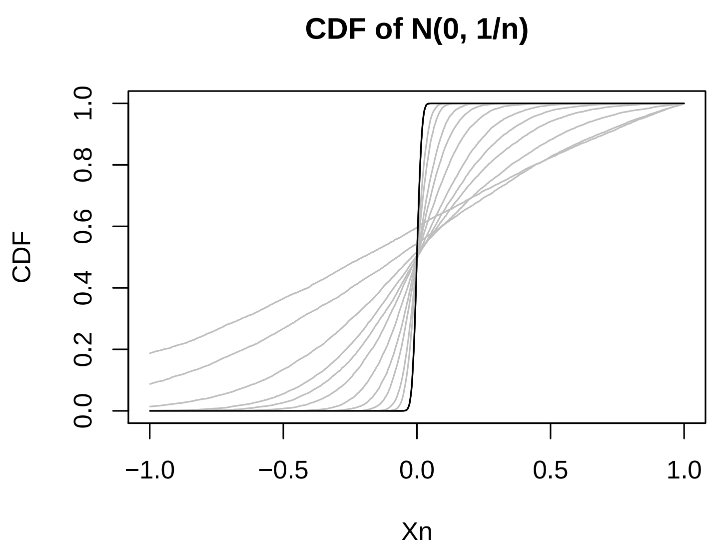
The variable X has a point mass distribution at 0, and we want to see if X_n converges. Note that $\mu = \mathbb{E}[X_n] = 0$ and $\sigma^2 = \text{Var}(X_n) = 1/n$ and $\lim_{n\to\infty}(1/n) = 0$, so by 5.2, this actually converges in quadratic mean to 0, which implies convergence in distribution.

Also showing it directly by using Chebyshev's inequality. For any $\epsilon > 0$:

$$\mathbb{P}(|X_n - 0| > \epsilon) \leqslant \frac{\sigma^2}{\epsilon^2} = \frac{1}{n\epsilon^2} \to 0$$

when $n \to \infty$. In conclusion, $X_n \xrightarrow{P} X$.

Next is investigating convergence in distribution which must be true, since convergence in probability implies convergence in distribution. We can also see the trend by plotting the CDF of X_n for $n = 1, 2, 5, 10, \ldots, 5000$ where $n = 5000$ is in black. Code for generating plot in `5.11.R`.



To show convergence in distribution, we find an expression for the CDF of X_n. By using the definition of the CDF and translating to the standard normal distribution:

$$F_n(z) = \mathbb{P}(X \leqslant z) = \mathbb{P}\left(\frac{X_n - 0}{1/\sqrt{n}} \leqslant \frac{z - 0}{1/\sqrt{n}}\right) = \mathbb{P}(\sqrt{n}X_n \leqslant \sqrt{n}z) = \mathbb{P}(Z \leqslant \sqrt{n}z),$$

where $Z \sim N(0, 1)$. As $n \to \infty$ then for any $z < 0$, $F_n(z) = 0$ and for $z \geqslant 0$, $F_n(z) = 1$ so it assumes the exact same form as $F(x)$, so $F_n \to F$ and $X_n \rightsquigarrow X$ as $n \to \infty$.

5.12
Let $X_1, X_2, \ldots$ and X be random variables that are positive integers. Show that $X_n \rightsquigarrow X$ if and only if, for every integer k,

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k).$$

Proof. This is a discrete random variable.
$\Rightarrow$). Assuming that $X_n \rightsquigarrow X$. By definition of convergence in distribution, for any k,

$$\lim_{n\to\infty} F_n(k) = F(k) \quad \text{and} \quad \lim_{n\to\infty} F_n(k-1) = F(k-1)$$

Since we have a discrete distribution,

$$\begin{aligned}
F(k) - F(k-1) &= \sum_{p\leqslant k} \mathbb{P}(X = p) - \sum_{p\leqslant k-1} \mathbb{P}(X = p) \\
&= \mathbb{P}(X = k) + \sum_{p\leqslant k-1} \mathbb{P}(X = p) - \sum_{p\leqslant k-1} \mathbb{P}(X = p) \\
&= \mathbb{P}(X = k)
\end{aligned}$$

The same relation is true for F_n, but here we also apply our assumption of convergence.

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(X_n = k) &= \lim_{n\to\infty} F_n(k) - \lim_{n\to\infty} F_n(k-1) \\
&= F(k) - F(k-1)
\end{aligned}$$

By combining these together, we have shown, for any k,

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k).$$

$\Leftarrow$) Now we assume that the limit of $\mathbb{P}(X_n = k)$ is equal to $\mathbb{P}(X = k)$ for all k, which means they have equal probability mass function. Then, for any k:

$$\begin{aligned}
F(k) = \mathbb{P}(X \leqslant k) &= \sum_{z=1}^{k} \mathbb{P}(X = z) \\
&= \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \ldots + \mathbb{P}(X = k) \\
&= \lim_{n\to\infty} \mathbb{P}(X_n = 1) + \lim_{n\to\infty} \mathbb{P}(X_n = 2) + \ldots + \lim_{n\to\infty} \mathbb{P}(X_n = k) \\
&= \sum_{z=1}^{k} \lim_{n\to\infty} \mathbb{P}(X_n = z) \\
&= \lim_{n\to\infty} \sum_{z=1}^{k} \mathbb{P}(X_n = z) \\
&= \lim_{n\to\infty} \mathbb{P}(X_n \leqslant k) = \lim_{n\to\infty} F_n(k)
\end{aligned}$$

(We can change the order of the sum and limit since it is a finite sum). In conclusion, this shows that $\lim_{n\to\infty} F_n(k) = F(k)$ and so $X_n \rightsquigarrow X$. $\square$

5.13
Let $Z_1, Z_2, \ldots$ be IID random variables with density f. Suppose $\mathbb{P}(Z_i > 0) = 1$ (the Z_i are nonnegative) and that $\lambda = \lim_{x \downarrow 0} f(x) > 0$. Define:

$$X_n = n \min\{Z_1, \ldots, Z_n\}.$$

Show that $X_n \rightsquigarrow Z$ where $Z \sim \text{Exp}(\lambda)$.

First, let's look at the CDF for a min function. Assume that we have some $V_1, \ldots, V_n$ that are IID and $V_i \sim F_V$, then define

$$W_n = \min(V_1, \ldots, V_n).$$

What is the CDF of F_W of W? By definition, we have:

$$F_W(y) = \mathbb{P}(W \leqslant y) = \mathbb{P}(\min(V_1, \ldots, V_n) \leqslant y)$$

This means that for at least one i, we have $V_i \leqslant y$. Another way of expressing this is to look at the complement, which is equivalent. $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$, which is the case when ALL $V_i > y$. Using this, and using independence:

$$\begin{aligned}
F_W(y) = \mathbb{P}(W \leqslant y) &= 1 - \mathbb{P}(V_1 > y, V_2 > y, \ldots, V_n > y) \\
&= 1 - \mathbb{P}(V_1 > y)\mathbb{P}(V_2 > y) \cdots \mathbb{P}(V_n > y) \\
&= 1 - [1 - F_V(y)][1 - F_V(y)] \cdots [1 - F_V(y)] \\
&= 1 - [1 - F_V(y)]^n
\end{aligned}$$

We will use this result in the following.

Returning to the exercise. Define $Z_m := \min(Z_1, \ldots, Z_n)$. Then:

$$F_n(y) = \mathbb{P}(X_n \leqslant y) = \mathbb{P}(nZ_m \leqslant y) = \mathbb{P}\left(Z_m \leqslant \frac{y}{n}\right) = 1 - \left[1 - F\left(\frac{y}{n}\right)\right]^n$$

From using the result above.

For the CDF, which is an integral, we can make an approximation with a step function (which can be thought of as approximating the curve with a square),

$$F\left(\frac{y}{n}\right) = \int_0^{y/n} f(z)dz \approx f(c_j)\left(\frac{y}{n} - 0\right) = f(c_j)\frac{y}{n}$$

where c_j is e.g. the midpoint in the interval, such as $c_j = \frac{y}{2n}$. When letting $n \to \infty$ then $f(c_j) \to f(0) = \lambda$. So:

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} 1 - \left[1 - F\left(\frac{x}{n}\right)\right]^n = \lim_{n\to\infty} 1 - \left[1 - \frac{f(c_j)x}{n}\right]^n = 1 - \exp(-f(0)x) = 1 - \exp(-\lambda x)$$

We recognise this as the exponential CDF with parameter λ, so if we set $Z \sim \text{Exp}(\lambda)$ with CDF F_Z, then we have shown that $F_n \to F_Z$ as $n \to \infty$, so $X_n \rightsquigarrow Z$.

5.14
Let $X_1, \dots, X_n \sim U(0,1)$ and let $Y_n = \overline{X}_n^2$. We will find the limiting distribution of Y_n.

By the CLT, $\overline{X} \sim N(\mu, \sigma^2)$, and from the uniform distrbution this is:

$$\mu = \mathbb{E}[X] = \frac{1+0}{2} = \frac{1}{2}$$

$$\sigma^2 = \text{Var}(X) = \frac{(1-0)^2}{12} = \frac{1}{12}$$

And so, $\overline{X} \sim N(1/2, 1/12)$. To find the distribution of $\overline{X}^2$, we apply the delta method with $g(x) = x^2$, which gives us $N(g(\mu), (g'(\mu))^2\sigma^2)$. Calculating:

$$g(\mu) = (1/2)^2 = 1/4$$

$$(g'(\mu))^2\sigma^2) = (2(1/2))^2(1/12) = 1/12$$

Which means that $\overline{X}^2 \sim N(1/4, 1/12)$. Verified with a simulation in `5.14.R`.

5.15
Let

$$\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix}, \dots, \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}$$

be IID random vectors with mean $\mu = (\mu_1, \mu_2)$ and variance Σ with

$$\overline{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i}, \qquad \overline{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i}.$$

We will find the limiting distribution of $Y_n = \overline{X}_1/\overline{X}_2$.

Following example 5.16 closely. $Y_n = g(\overline{X}_1, \overline{X}_2)$ where

$$g(s_1, s_2) = \frac{s_1}{s_2}.$$

By the central limit theorem:

$$\sqrt{n}\begin{pmatrix} \overline{X}_1 - \mu_1 \\ \overline{X}_2 - \mu_2 \end{pmatrix} \rightsquigarrow N(0, \Sigma).$$

Finding the gradient of $g(s_1, s_2)$.

$$\nabla g(s_1, s_2) = \begin{pmatrix} \dfrac{\partial g}{\partial s_1} \\ \dfrac{\partial g}{\partial s_2} \end{pmatrix} = \begin{pmatrix} \dfrac{1}{s_2} \\ -\dfrac{s_1}{s_2^2} \end{pmatrix}$$

Calculating the deltra transformed covariance matrix.

$$
\begin{aligned}
\nabla g(\mu_1,\mu_2)^T \Sigma \nabla g(\mu_1,\mu_2) &= \begin{pmatrix} \dfrac{1}{\mu_2} \\ -\dfrac{\mu_1}{\mu_2^2} \end{pmatrix}^T \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \dfrac{1}{\mu_2} \\ -\dfrac{\mu_1}{\mu_2^2} \end{pmatrix} \\
&= \begin{pmatrix} \dfrac{1}{\mu_2} \\ -\dfrac{\mu_1}{\mu_2^2} \end{pmatrix}^T \begin{pmatrix} \dfrac{\sigma_{11}}{\mu_2} - \dfrac{\sigma_{12}\mu_1}{\mu_2^2} \\ \dfrac{\sigma_{21}}{\mu_2} - \dfrac{\sigma_{22}\mu_1}{\mu_2^2} \end{pmatrix} \\
&= \frac{\sigma_{11}}{\mu_2^2} - \frac{\sigma_{12}\mu_1}{\mu_2^3} - \frac{\sigma_{21}\mu_1}{\mu_2^3} + \frac{\sigma_{22}\mu_1^2}{\mu_2^4} \\
&= \frac{\mu_2^2\sigma_{11} - \mu_1\mu_2\sigma_{12} - \mu_1\mu_2\sigma_{21} + \mu_1^2\sigma_{22}}{\mu^4}
\end{aligned}
$$

With this we have found the limiting distribution.

$$
\sqrt{n}\big(\overline{X}_1/\overline{X}_2 - \mu_1/\mu_2\big) \rightsquigarrow N\Big(0, \frac{\mu_2^2\sigma_{11} - \mu_1\mu_2\sigma_{12} - \mu_1\mu_2\sigma_{21} + \mu_1^2\sigma_{22}}{\mu^4}\Big)
$$

5.16

Finding an example where $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$, but $X_n + Y_n \not\rightsquigarrow X + Y$.

Set $X \sim U(0,1)$ and $Y \sim U(-1,0)$. Then we define $X_n = U(0,1)$ and $Y_n = -X_n$ and we denote the CDFs as $X \sim F$, $Y \sim G$, $X_n \sim F_n$ and $Y_n \sim G_n$.

Then:

$$
\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} x = x = F(x)
$$

$$
\lim_{n\to\infty} G_n(x) = \lim_{n\to\infty} -x = -x = G(x)
$$

so $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$. However,

$$
\lim_{n\to\infty} F_n(x) + G_n(x) = \lim_{n\to\infty} x - x = \lim_{n\to\infty} 0 = 0
$$

so $X_n + Y_n \rightsquigarrow 0$, but $X + Y \neq 0$. (This can be seen in exercise 2.20 with $X - Y$, which corresponds to this one if we instead use $X + Z$ with $Z = -Y$. Also see `5.16.R`).

6 Models, Statistical Inference and Learning

Definitions

The bias of an estimator is defined as

$$\texttt{bias}(\widehat{\theta}_n) = \mathbb{E}_\theta[\widehat{\theta}_n] - \theta.$$

The standard error, denoted by $\texttt{se}$ and is defined as:

$$\texttt{se}(\widehat{\theta}_n) = \sqrt{\mathrm{Var}(\widehat{\theta}_n)}.$$

The MSE, mean squared error, is defined as:

$$\mathrm{MSE} = \mathbb{E}_\theta[(\widehat{\theta}_n - \theta)^2].$$

The MSE can according to Theorem 6.9 be written as:

$$\mathrm{MSE} = \texttt{bias}^2(\widehat{\theta}_n) + \mathrm{Var}_\theta(\widehat{\theta}_n).$$

Exercises

6.1

Let $X_1, \ldots, X_n \sim \mathrm{Poisson}(\lambda)$ and let $\widehat{\lambda} = n^{-1} \sum_{i=1}^n X_i$. We will calculate the bias, se and MSE of the estimator.

For the Poisson distribution, we know that $\mathbb{E}[X_i] = \lambda$ and $\mathrm{Var}(X_i) = \lambda$.

Calculating the bias.

$$\texttt{bias}(\widehat{\lambda}) = \mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^n X_i\right] - \lambda = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta[X_i] - \lambda = \lambda - \lambda = 0$$

which shows that the mean is an unbiased estimator. Calculating the variance.

$$\mathrm{Var}(\widehat{\lambda}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\lambda}{n}$$

The standard error is the square root of the variance.

$$\texttt{se}(\widehat{\lambda}) = \sqrt{\frac{\lambda}{n}}$$

Finally, the MSE.

$$\mathrm{MSE} = 0^2 + \frac{\lambda}{n} = \frac{\lambda}{n}.$$

6.2
Let $X_1, \ldots, X_n \sim U(0, \theta)$ and let $\widehat{\theta} = \max(X_1, \ldots, X_n)$. Find `bias`, `se` and MSE.

For the $U(0, \theta)$ uniform distribution, $\mathbb{E}[X_i] = \theta/2$ and $\mathrm{Var}(X_i) = \theta^2/12$. However, we are using the maximum value, which we don't know the expectation of. In order to calculate it, we will need the PDF, which requires the CDF. For each X_i the CDF is $F(x) = x/\theta$. We assume independence and will find the CDF of $\widehat{\theta}$ which we will denote as F_θ.

$$
\begin{aligned}
F_\theta(x) = \mathbb{P}(\widehat{\theta} \leqslant x) &= \mathbb{P}(X_1 \leqslant x, X_2 \leqslant x, \ldots, X_n \leqslant x) \\
&= \mathbb{P}(X_1 \leqslant x)\mathbb{P}(X_2 \leqslant x) \cdots P(X_n \leqslant x) \\
&= \left(\frac{x}{\theta}\right)^n \\
&= \frac{x^n}{\theta^n}
\end{aligned}
$$

We find the PDF by differentiating the CDF.

$$
f_\theta(x) = F'(x) = \frac{n\theta^{n-1}}{\theta}
$$

Now we can calculate the expectation. (A simulated verification is in `6.2.R`).

$$
\begin{aligned}
\mathbb{E}[\widehat{\theta}] &= \int_0^\theta x \cdot \frac{nx^{n-1}}{\theta} dx = \frac{n}{\theta^n} \int_0^\theta x^n dx \\
&= \frac{n}{\theta^n} \left[\frac{x^{n+1}}{n+1}\right]_0^\theta = \frac{n}{n+1} \frac{1}{\theta^n} \theta^{n+1} \\
&= \frac{n}{n+1} \theta
\end{aligned}
$$

We will also need the variance, so we calculate the second moment.

$$
\begin{aligned}
\mathbb{E}[\widehat{\theta}^2] &= \int_0^\theta x^2 \cdot \frac{nx^{n-1}}{\theta} dx = \frac{n}{\theta^n} \int_0^\theta x^{n+1} dx \\
&= \frac{n}{\theta^n} \left[\frac{x^{n+2}}{n+2}\right]_0^\theta = \frac{n}{n+2} \frac{1}{\theta^n} \theta^{n+2} \\
&= \frac{n}{n+2} \theta^2
\end{aligned}
$$

Now we can calculate the variance (skipping the algebra). (A simulated verification is in `6.2.R`).

$$
\begin{aligned}
\mathrm{Var}(\widehat{\theta}) &= \mathbb{E}[\widehat{\theta}^2] - \mathbb{E}[\widehat{\theta}]^2 \\
&= \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\right)^2 \theta^2 \\
&= \frac{n}{(n+2)(n+1)^2}\theta^2
\end{aligned}
$$

Finding the `bias`.

$$\texttt{bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta = \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\theta$$

This estimator is not unbiased, because we will often end up with a slightly smaller value. As we have shown, we could use the unbiased estimator $\frac{n+1}{n}\max(X_1, \ldots, X_n)$ instead.

Finding the standard error, which is just the square root of the variance.

$$\texttt{se}(\widehat{\theta}) = \sqrt{\mathrm{Var}(\widehat{\theta})} = \frac{\sqrt{n}}{\sqrt{n+2}(n+1)}\theta$$

And finally, the MSE.

$$\begin{aligned}
\mathrm{MSE} &= \texttt{bias}^2(\widehat{\theta}) + \mathrm{Var}(\widehat{\theta}) \\
&= \frac{1}{(n+1)^2}\theta^2 + \frac{n}{(n+2)(n+1)^2}\theta^2 \\
&= \frac{n+2}{(n+2)(1+n)^2}\theta^2 + \frac{n}{(n+2)(n+1)^2}\theta^2 \\
&= \frac{2n+2}{(n+2)(n+1)^2}\theta^2
\end{aligned}$$

6.3

Let $X_1, \ldots, X_n \sim U(0, \theta)$ and let $\widehat{\theta} = 2\overline{X}_n$. Find the bias, se and MSE.

Here the estimator is two times the mean, or written more explicitly.

$$\widehat{\theta} = \frac{2}{n}\sum_{i=1}^{n} X_i$$

Now we can use the properties of the uniform distribution:

$$\mathbb{E}[X_i] = \frac{\theta}{2} \quad \text{and} \quad \mathrm{Var}(X_i) = \frac{\theta^2}{12}$$

Finding the bias.

$$\texttt{bias}(\widehat{\theta}) = \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n} X_i\right] - \theta = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] - \theta = 2 \cdot \frac{\theta}{2} - \theta = \theta - \theta = 0$$

In this case, we have an unbiased estimator. Calculating the variance:

$$\mathrm{Var}(\widehat{\theta}) = \mathrm{Var}\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) = \frac{4}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right) = \frac{4}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

(Simulations for expectation and variance confirmed in `6.3.R`).

The standard error, se, is just the square root of the variance.

$$\texttt{se}(\widehat{\theta}) = \sqrt{\mathrm{Var}(\widehat{\theta})} = \frac{\theta}{\sqrt{3n}}$$

Calculating the MSE.

$$\begin{aligned}
\mathrm{MSE} &= \texttt{bias}^2(\widehat{\theta}) + \mathrm{Var}(\widehat{\theta}) \\
&= 0^2 + \frac{\theta^2}{3n} \\
&= \frac{\theta^2}{3n}
\end{aligned}$$

7 Estimating the CDF and Statistical Functionals

Definition 7.1 The empirical distribution function $\widehat{F}_n$ is the CDF that puts mass 1/n at each data point X_i.

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leqslant x)}{n}$$

where $I(\cdot)$ is the indicator function.

Exercises

7.1 - Theorem 7.3

$$\mathbb{E}[\widehat{F}_n(x)] = F(x)$$

Proof.

$$\begin{aligned}
\mathbb{E}[\widehat{F}_n(x)] = \mathbb{E}\left[\frac{\sum_{i=1}^n I(X_i \leqslant x)}{n}\right] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[I(X_i \leqslant x)] \\
&= \frac{1}{n}\sum_{i=1}^n \Big(1 \cdot \mathbb{P}(X_i \leqslant x) + 0 \cdot \mathbb{P}(X_i > x)\Big) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{P}(X_i \leqslant x) = \frac{1}{n} \cdot n\mathbb{P}(X_i \leqslant x) \\
&= \mathbb{P}(X_i \leqslant x) \\
&= F(x) \qquad \square
\end{aligned}$$

The second identity.

$$\mathrm{Var}(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$$

Proof. Calculating the variance of $I(X_i \leqslant x)$, starting with the second moment.

$$\mathbb{E}[I(X_i \leqslant x)^2] = (1)^2 \cdot \mathbb{P}(X_i \leqslant x) + (0)^2 \cdot \mathbb{P}(X_i > x) = \mathbb{P}(X_i \leqslant x)$$

The variance.

$$\mathrm{Var}(I(X_i \leqslant x)) = \mathbb{E}[I(X_i \leqslant x)^2] - \mathbb{E}[I(X_i \leqslant x)]^2 = \mathbb{P}(X_i \leqslant x) - \mathbb{P}(X_i \leqslant x)^2 = \mathbb{P}(X_i \leqslant x)(1-\mathbb{P}(X_i \leqslant x))$$

$$\begin{aligned}
\mathrm{Var}(\widehat{F}_n(x)) = \mathrm{Var}\left[\frac{\sum_{i=1}^n I(X_i \leqslant x)}{n}\right] &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[I(X_i \leqslant x)] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{P}(X_i \leqslant x)(1-\mathbb{P}(X_i \leqslant x)) \\
&= \frac{\mathbb{P}(X_i \leqslant x)(1-\mathbb{P}(X_i \leqslant x))}{n} \\
&= \frac{F(x)(1-F(x))}{n} \qquad \square
\end{aligned}$$

MSE.

$$\text{MSE} = \frac{F(x)(1-F(x))}{n} \to 0$$

Proof. First we need the bias, but as the first result showed, this is an unbiased estimator.

$$\texttt{bias}(\widehat{F}_n) = \mathbb{E}[\widehat{F}_n] - F_n = F_n - F_n = 0$$

The MSE is the sum of the squared bias and the variance, but since the bias is 0, it becomes equal to the variance.

$$\text{MSE} = \frac{F(x)(1-F(x))}{n}$$

As, $n \to \infty$, this will tend to 0.

$$\lim_{n\to\infty} \text{MSE} = \lim_{n\to\infty} \frac{F(x)(1-F(x))}{n} = 0 \qquad \square$$

The empirical distribution converges in probability to $F(x)$.

$$\widehat{F}_n(x) \xrightarrow{P} F(x).$$

Proof. Recalling that $\mu = \mathbb{E}[\widehat{F}_n] = F_n$ and $\sigma^2 = (F(x)(1-F(x)))/n$. Using Chebyshev's inequality. For any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\widehat{F}_n(x) - F_n(x)| > \epsilon) \leqslant \lim_{n\to\infty} \frac{\sigma^2}{\epsilon^2} = \lim_{n\to\infty} \frac{F(x)(1-F(x))}{n\epsilon^2} = 0.$$

This shows that $\widehat{F}_n(x) \xrightarrow{P} F(x)$. $\square$

7.2

For $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and $Y_1, \ldots, Y_m \sim \text{Bernoulli}(q)$, we will find the plug-in estimator and estimated standard error for p, an approximate 90 percent confidence interval for p, the plug-in estimator and standard error for $p - q$ and approximate 90 percent confidence interval for $p - q$. (Oh, is that all?)

The plug-in estimator is really just a term for using the $\widehat{F}_n$ instead of F for calculating functionals, such as the mean, median etc. For estimating p in a Bernoulli distribution, it is simply the sample mean.

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Finding the standard error by e.g. using $\widehat{p}$ in the variance formula.

$$\texttt{se}(\widehat{p}) = \frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}}$$

By consluting a standard normal table we can see that 1.65 is the value that gives 0.95 of the data, which will correspond to a 90% two-sided test.

$$\widehat{p} \pm z_{.10/2}\texttt{se}(\widehat{p}) = 1.65 \cdot \texttt{se}(\widehat{p})$$

Finding an estimate for $p - q$ with a plug-in estimator is simply subtracting the sample means.

$$\widehat{p} - \widehat{q} = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_j$$

The standard error for $\widehat{q}$ is:

$$\mathtt{se}(\widehat{q}) = \frac{\sqrt{\widehat{q}(1-\widehat{q})}}{\sqrt{m}}$$

And the combined standard error is:

$$\mathtt{se} = \sqrt{(\mathtt{se}(\widehat{p}))^2 + (\mathtt{se}(\widehat{q}))^2},$$

and in the same way, we can calculate a 90% confidence interval by using the same value from the standard normal table.

$$\widehat{p} - \widehat{q} \pm 1.65 \cdot \mathtt{se}$$

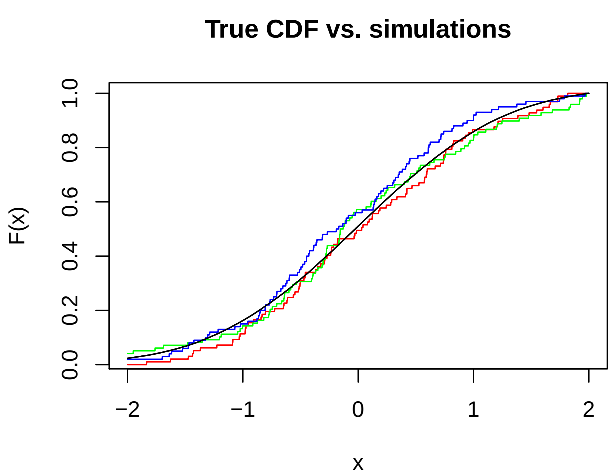
7.3

Program code found in `7.3.R`. There is no technical appendix, so the confidence intervals are meant to be made with the Dvoretzky-Kiefer-Wolfowitz inequality.

For a 95% confidence interval, we define

$$\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$$

where $\alpha = 0.05$ and specify the bounds as $(\widehat{F}_n(x) - \epsilon_n, \widehat{F}_n(x) + \epsilon_n)$. We simulate 100 $X \sim N(0, 1)$ variables a total of 1000 times and calculate the confidence intervals for each bound, then we count how many of the 1000 simulations make bounds that do not contain the true CDF. Illustration of what the simulations look like. (Black is the true CDF, the colored are simulations).



On running the simulation, approximately 95% of the simulation bounds contained the true CDF.
When redoing for the Cauchy distribution, the results are approximately 95% as well.

```
[1] "Proportion of normal distributions containing the CDF"
> sum(checkNorm)/length(checkNorm)
[1] 0.952
[1] "Proportion of Cauchy distributions containing the CDF"
> sum(checkCauchy)/length(checkCauchy)
[1] 0.949
```

R

7.4

Let $X_1, \ldots, X_n \sim F$ and let $\widehat{F}_n(x)$ be the empirical distribution function. For a fixed x, we will use the CLT to find the limiting distribution of $\widehat{F}_n(x)$.

We make the definiton:

$$\overline{Y}_n = \frac{\sum_{i=1}^n I(X_i \leqslant x)}{n}$$

to highlight the fact that $\widehat{F}_n(x) = \overline{Y}_n$ is a sample mean. By the CLT:

$$\frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma} \rightsquigarrow \mathcal{N}(0,1)$$

where $\mu = \mathbb{E}[\overline{Y}_n]$ and $\sigma^2/n = \mathrm{Var}(\overline{Y}_n)$, or the equivalent statement:

$$\overline{Y}_n \rightsquigarrow \mathcal{N}(\mu, \sigma^2/n) = \mathcal{N}(\mathbb{E}[\overline{Y}_n], \mathrm{Var}(\overline{Y}_n))$$

Since x is fixed, we can apply Theorem 7.3:

$$\mathbb{E}[\overline{Y}_n] = \mathbb{E}[\widehat{F}_n(x)] = F(x)$$

$$\mathrm{Var}(\overline{Y}_n) = \mathrm{Var}(\widehat{F}_n(x)) = \frac{F(x)[1-F(x)]}{n}$$

In summary, when replacing $\widehat{F}_n(x)$ for $\overline{Y}_n$, we have shown:

$$\widehat{F}_n(x) \approx \mathcal{N}\left(F(x), \frac{F(x)[1-F(x)]}{n}\right)$$

We can actually apply Chebyshev's inequality in this case. For any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\widehat{F}_n(x) - F(x)| > \epsilon) \leqslant \lim_{n\to\infty} \frac{F(x)[1-F(x)]}{n\epsilon^2} = 0,$$

which shows that $\widehat{F}_n(x) \xrightarrow{P} F(x)$ which again implies $\widehat{F}_n(x) \rightsquigarrow F(x)$.

7.5
Let x and y be two distinct points. Find $\mathrm{Cov}\big(\widehat{F}_n(x), \widehat{F}_n(y)\big)$.

First of all, introducing the short-hand notation:

$$I_x := I(X_i \leqslant x), \qquad I_y := I(X_i \leqslant y)$$

Assuming independence of the X_i. Then, $\mathrm{Cov}(I_x, I_y) = 0$ whenever $i \neq j$. By definition:

$$\begin{aligned}
\mathrm{Cov}\big(\widehat{F}_n(x), \widehat{F}_n(y)\big) &= \mathrm{Cov}\left(\frac{\sum_{i=1}^n I_x}{n}, \frac{\sum_{j=1}^n I_y}{n}\right) \\
&= \frac{1}{n^2}\mathrm{Cov}\left(\sum_{i=1}^n I_x, \sum_{j=1}^n I_y\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Cov}(I_x, I_y)
\end{aligned}$$

The last part follows from the identity proved in exercise 3.14, and from independence, which leads to all terms with different indices to become 0, hence leaving just a single sum. Now we will do an intermediary calculation, and we will assume that $x < y$:

$$\begin{aligned}
\mathrm{Cov}(I_x, I_y) &= \mathbb{E}[I_x I_y] - \mathbb{E}[I_x]\mathbb{E}[I_y] \\
&= \mathbb{E}[I_x] - \mathbb{E}[I_x]\mathbb{E}[I_y] && \text{(Since } x < y\text{)} \\
&= \mathbb{E}[I_x](1 - \mathbb{E}[I_y]) \\
&= \mathbb{P}(X_i \leqslant x)(1 - \mathbb{P}(X_i \leqslant y)) \\
&= F(x)\big[1 - F(y)\big]
\end{aligned}$$

Returning to the main calculation:

$$\begin{aligned}
\mathrm{Cov}\big(\widehat{F}_n(x), \widehat{F}_n(y)\big) &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Cov}(I_x, I_y) \\
&= \frac{1}{n^2}\sum_{i=1}^n F(x)\big[1 - F(y)\big] \\
&= \frac{1}{n^2} \cdot n \cdot F(x)\big[1 - F(y)\big] \\
&= \frac{F(x)\big[1 - F(y)\big]}{n}.
\end{aligned}$$

This is provided that $x < y$. But since the covariance is symmetric, then the smallest probability can always be put in as x. Since F is unknown, the plug-in estimator is:

$$\frac{\widehat{F}_n(x)\big[1 - \widehat{F}_n(y)\big]}{n}$$

Note that this is very similar to the expression for the variance.

$$\mathrm{Var}(\widehat{F}_n) = \frac{F(x)(1 - F(x))}{n}$$

7.6

Let $X_1, \ldots, X_n \sim F$ and let $\widehat{F}$ be the empirical distribution function. For $a < b$ define $\theta = T(F) = F(b) - F(a)$ and set $\widehat{\theta} = T(\widehat{F}_n) = \widehat{F}_n(b) - \widehat{F}_n(a)$. Find the estimted standard error of $\widehat{\theta}$ and an expression for the $1-\alpha$ confidence interval for θ.

Following example 7.15, the standard error of a difference is the square root of the following variances:

$$
\begin{aligned}
\mathtt{se}(\widehat{\theta}) &= \sqrt{\mathrm{Var}(\widehat{F}_n(b) - \widehat{F}_n(a))} \\
&= \Big(\mathrm{Var}(\widehat{F}_n(b)) + \mathrm{Var}(\widehat{F}_n(b)) - 2\mathrm{Cov}(\widehat{F}_n(a), \widehat{F}_n(b))\Big)^{\frac{1}{2}} && \text{(Prop. of var.)} \\
&= \frac{1}{\sqrt{n}} \left(F(b)[1-F(b)] + F(a)[1-F(a)] - 2F(a)[1-F(b)]\right)^{\frac{1}{2}} && \text{(Thm 7.3, Ex. 7.5)} \\
&= \sqrt{\frac{(F(b)-F(a))[1-(F(b)-F(a))]}{n}} && \text{(See notes.)} \\
&= \sqrt{\frac{(\widehat{F}_n(b)-\widehat{F}_n(a))[1-(\widehat{F}_n(b)-\widehat{F}_n(a))]}{n}} && \text{(Plug-in est.)} \\
&= \sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}} && \text{(Def. of } \widehat{\theta}\text{)}
\end{aligned}
$$

Finding the confidence interval is easy. Just find the corresponding $z_{\alpha/2}$ level and calculate:

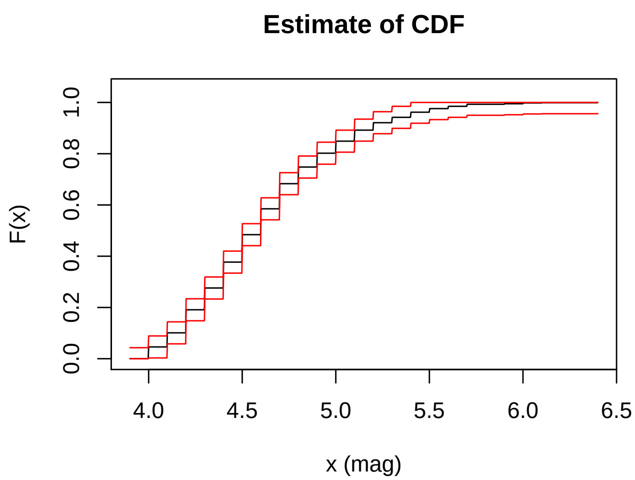
$$\widehat{\theta} \pm z_{\alpha/2}\mathtt{se}(\widehat{\theta}).$$

For the sake of clarity, here are the notes for the algebra steps used in the step above. Instead of $F(b)$ and $F(a)$ we just write b and a.

$$
\begin{aligned}
b(1-b) + a(1-a) - 2a(1-b) &= b - b^2 + a - a^2 - 2a + 2ab \\
&= b - b^2 - a - a^2 + 2ab \\
&= b - a - b^2 + 2ab - a^2 \\
&= b - a - (b^2 - 2ab + a^2) \\
&= b - a - (b-a)^2 \\
&= (b-a)\big[1-(b-a)\big]
\end{aligned}
$$

7.7

Code found in `7.7.R`. Plot of empirical distribution and bounds:



(The CDF jumps because there are only 1000 points and jumps in the data itself).

Finding the confidence interval for $F(4.9) - F(4.3)$.

```
'95% Confidence Interval for F(4.9) - F(4.3):'
( 0.495 , 0.557 )
```

R

7.8