Probability and Statistics

Exercise sheet 4

Exercise 4.1 (On measurability)

(a) Consider $X_1, ..., X_m$ ($m \geq 1$) random variables defined on some $(\Omega, \mathcal{A}, P)$ and taking values in $(\mathbb{R}, \mathcal{B})$, where $\mathcal{A}$ is a σ-algebra on Ω, P is a probability measure on $\mathcal{A}$ and $\mathcal{B}$ is the Borel σ-field.

Show that

$$-X_1, \max_{1 \leq i \leq m} X_i, \min_{1 \leq i \leq m} X_i$$

are random variables.

(b) Consider now a sequence $(X_n)_{n \geq 1}$ of random variables $X_1, X_2, ...$ defined on $(\Omega, \mathcal{A}, P)$ and taking values in $(\mathbb{R}, \mathcal{B})$ as in (a).

Show that:

$$\sup_{n \geq 1} X_n, \inf_{n \geq 1} X_n, \limsup_{n \to \infty} X_n, \liminf_{n \to \infty} X_n, \lim_{n \to \infty} X_n$$

are all random variables, in the case of the limit assuming that it exists.

Here, recall the definition of

$$\limsup_{n \to \infty} := \inf_{n \geq 1} (\sup_{k \geq n} X_k)$$

and

$$\liminf_{n \to \infty} := \sup_{n \geq 1} (\inf_{k \geq n} X_k).$$

Solution 4.1

(a) Recall that to show that

$$X : (\Omega, \mathcal{A}, P) \mapsto (\mathbb{R}, \mathcal{B})$$

is a random variable, we need to prove that it is measurable, that is $\forall B \in \mathcal{B}$, $\{X \in B\} \in \mathcal{A}$. This is also equivalent to

$$\begin{aligned} & \forall x \in \mathbb{R}\ \{X \leq x\} \in \mathcal{A} \\ \Leftrightarrow & \forall x \in \mathbb{R}\ \{X < x\} \in \mathcal{A} \\ \Leftrightarrow & \forall x \in \mathbb{R}\ \{X \geq x\} \in \mathcal{A} \\ \Leftrightarrow & \forall x \in \mathbb{R}\ \{X > x\} \in \mathcal{A} \end{aligned}$$

since

$$\begin{aligned}\mathcal{B} = \sigma(\mathcal{O}) &= \sigma(\{(-\infty, x] : x \in \mathbb{R}\}) \\ &= \sigma(\{(-\infty, x) : x \in \mathbb{R}\}) \\ &= \sigma(\{[x, +\infty) : x \in \mathbb{R}\}) \\ &= \sigma(\{(x, +\infty) : x \in \mathbb{R}\})\end{aligned}$$

with $\mathcal{O}$ the collection of all open sets and $\sigma(\mathcal{C})$ the minimal σ-algebra generated by a non-empty collection of sets $\mathcal{C}$.

- $\{-X_1 \leq x\} = \{X_1 \geq -x\} \in \mathcal{A}$ $\forall x \in \mathbb{R}$ because X_1 is measurable.
- $$\left\{\max_{1 \leq i \leq m} X_i \leq x\right\} = \bigcap_{i=1}^{m} \{X_i \leq x\} \in \mathcal{A}$$

since $X_1, ..., X_m$ are measurable and $\mathcal{A}$ is closed under finite intersections.

- $$\left\{\inf_{1 \leq i \leq m} X_i \leq x\right\} = \bigcup_{i=1}^{m} \{X_i \leq x\} \in \mathcal{A}$$

for the same reason as above.

Alternatively, we can use the first part to say that $-X_1, ..., -X_m$ are random variables, and so we can argue that

$$\min_{1 \leq i \leq m} X_i = -\max_{1 \leq i \leq m} (-X_i)$$

is a random variable as well by the first and second parts.

(b) • For a sequence $(X_n)_{n \geq 1}$ of random variables, the reasoning is similar:

$$\{\sup_{n \geq 1} X_n \leq x\} = \bigcap_{n \geq 1} \{X_n \leq x\} \in \mathcal{A}$$

since $\mathcal{A}$ is closed under countable intersections, and

$$\inf_{n \geq 1} X_n = -\sup_{n \geq 1} (-X_n)$$

is measurable by the same argument as before: $-X_n$ is measurable, then so is $\sup_{n \geq 1}(-X_n)$ and therefore so is the infimum (or one can argue directly).

- $$\limsup_{n \to \infty} X_n = \inf_{n \geq 1} (\sup_{k \geq n} X_k)$$

Note that each $Y_n := \sup_{k \geq n} X_k$ is measurable as proved above, and so

$$\limsup_{n \to \infty} X_n = \inf_{n \geq 1} Y_n$$

is measurable as well.

- $$\liminf_{n \to \infty} = -\limsup_{n \to \infty} (-X_n)$$

so it must be measurable.

- If $\lim_{n \to \infty} X_n$ exists, then

$$\lim_{n \to \infty} X_n = \limsup_{n \to \infty} X_n$$

and so must be measurable in that case.

Exercise 4.2 (On the cdf of min and max of i.i.d random variables)

Let $X_1, ..., X_n$ be $\overset{\text{iid}}{\sim} F$.

(a) Let $S_n := \max_{1 \le i \le n} X_i$. Find the cdf of S_n as a function of F.

(b) Do the same but for $I_n := \min_{1 \le i \le n} X_i$.

(c) Fix $x \in \mathbb{R}$ such that $F(x) \in (0, 1)$. What is the limit of the cdf of S_n at x as $n \to \infty$? What about the cdf of I_n? How would you interpret these results? What does this mean if $X_1, ..., X_n$ take values in a finite set $\{\xi_1, ..., \xi_k\}$?

Solution 4.2

(a) S_n is a random variable by the previous question.

For $x \in \mathbb{R}$,

$$\begin{aligned} P(S_n \le x) &= P\left(\max_{1 \le u \le n} X_i \le x\right) \\ &= P(X_1 \le x, ..., X_n \le x) \\ &= \prod_{i=1}^{n} P(X_i \le x) \\ &= [F(x)]^n. \end{aligned}$$

(b) For $x \in \mathbb{R}$,

$$\begin{aligned} P(I_n \le x) &= 1 - P(I_n > x) \\ &= 1 - P(X_1 > x, ..., X_n > x) \\ &= 1 - \prod_{i=1}^{n} P(X_i > x) \\ &= 1 - [1 - F(x)]^n. \end{aligned}$$

(c) Let $x \in \mathbb{R}$ be such that $F(x) \in (0, 1)$.

Then

$$\lim_{n\to\infty} P(S_n \le x) = \lim_{n\to\infty} [F(x)]^n = 0$$

and

$$\lim_{n\to\infty} P(I_n \le x) = \lim_{n\to\infty} (1 - [1 - F(x)]^n) = 1 - 0 = 1.$$

This can be interpreted by saying that as n grows, the maximum and minimum are dragged to an extreme value (if they stayed somewhere inside the support of $X_1, ..., X_n$ we would have obtained values of the cdf of S_n or I_n away from 0 and 1, respectively).

In the example where $X_i \in \{\xi_1, ..., \xi_k\}$ with $\xi_1 < ... < \xi_k$, S_n and I_n will converge to ξ_k and ξ_1 respectively, in probability (a concept to be defined later).

Exercise 4.3 (On expectation)

(a) For any cdf F show that

$$P(X \in (a, b]\,) = F(b) - F(a)$$

for any $a < b$ and where X is a random variable with cdf F.

(b) Let X be a nonnegative discrete random variable taking its values in the set $\{x_1, x_2, ...\}$ (possibly countably infinite), where we assume that the values are ordered by $x_1 < x_2 < ....$ Suppose $E(X)$ exists. Show that

$$E(X) = \sum_{j=0}^{\infty}(x_{j+1} - x_j)P(X > x_j)$$

with $x_0 := 0$.

Does this match with the tail sum seen in the lecture?

(c) Show that if F is the cdf of X (the same X as in (b)), then $E(X)$ can also be given by the formula

$$E(X) = \int_0^{\infty}(1 - F(x))dx. \tag{1}$$

(d) Show that for a general discrete random variable (possibly taking values in $(-\infty, 0)$),

$$E(X) = -\int_{-\infty}^{0} P(X < x)dx + \int_0^{\infty}(1 - F(x))dx \tag{2}$$

provided that $E(X)$ exists.

Remark: Actually, the formulas in 1 and 2 are true in general for any type of nonnegative and general random variables. Also, $\int_{-\infty}^{0} P(X < x)dx$ can be replaced by $\int_{-\infty}^{0} F(x)dx$.

Solution 4.3

(a) We have

$$P(X \in (a, b]) + P(X \leq a) = P(X \in (a, b] \cup (-\infty, a]) = P(X \in (-\infty, b]).$$

Since $P(X \leq a) = F(a)$ and $P(X \leq b) = F(b)$, the result follows.

(b) We have

$$\begin{aligned}
&\sum_{j=0}^{\infty}(x_{j+1}-x_j)P(X>x_j)\\
=&\sum_{j=0}^{\infty}(x_{j+1}-x_j)P(X\geq x_{j+1})\\
=&\sum_{j=0}^{\infty}(x_{j+1}-x_j)\sum_{k=j+1}^{\infty}P(X=x_k)\\
=&\sum_{j=0}^{\infty}(x_{j+1}-x_j)\sum_{k=1}^{\infty}P(X=x_k)1_{k\geq j+1}\\
=&\sum_{k=1}^{\infty}\sum_{j=0}^{\infty}(x_{j+1}-x_j)P(X=x_k)1_{j\leq k-1}\\
=&\sum_{k=1}^{\infty}\sum_{j=0}^{k-1}(x_{j+1}-x_j)P(X=x_k)\\
=&\sum_{k=1}^{\infty}(x_k-x_0)P(X=x_k)\\
=&E(X).
\end{aligned}$$

One must be careful to justify exchanging the summations, but this is correct in this case since all the summands are positive; in particular this can be seen as a discrete version of Fubini's theorem, which in this case gives that

$$\sum_{m,n=0}^{\infty}a_{m,n}=\sum_{m\geq 0}\sum_{n\geq 0}a_{m,n}=\sum_{n\geq 0}\sum_{m\geq 0}a_{m,n}$$

if all the $a_{m,n}\geq 0$.

This matches the tail sum for integer-valued random variables, where $\{x_1,x_2,...\}=\{0,1,...\}$, so that $x_{j+1}-x_j=1$ and $x_j=j$.

(c) Now we check that $E(X)=\int_0^{\infty}(1-F(x))dx$: We have:

$$F(x)=\begin{cases}0 & \text{if } x<x_1\\ P(X=x_1) & \text{if } x_1\leq x<x_2\\ ... & ...\\ \sum_{j=1}^{k}P(X=x_j) & \text{if } x_k\leq x<x_{k+1}\end{cases}$$

Therefore,

$$\begin{aligned}
\int_0^\infty (1-F(x))dx &= \sum_{j=0}^{\infty}\int_{x_j}^{x_{j+1}}(1-F(x))dx \\
&= \sum_{j=0}^{\infty}\int_{x_j}^{x_{j+1}}(1-P(X\le x_j))dx \\
&= \sum_{j=0}^{\infty}(x_{j+1}-x_j)(1-P(X\le x_j)) \\
&= \sum_{j=0}^{\infty}(x_{j+1}-x_j)P(X> x_j) \\
&= E(X).
\end{aligned}$$

The integrals are interpreted as being over $[x_j, x_{j+1})$, though it does not matter whether the intervals are open or closed.

(d) We start by noting that we can always write

$$X = \max(X,0) + \min(X,0) = X^+ - X^-$$

where $X^+ := \max(X,0) \ge 0$, $X^- := -\min(X,0) \ge 0$.

By linearity of expectation, we have

$$\begin{aligned}
E(X) &= E(X^+) - E(X^-) \\
&= \int_0^\infty (1-F^+(x))dx - \int_0^\infty (1-F^-(x))dx
\end{aligned}$$

where F^- and F^+ denote the cdf of X^- and X^+ respectively. Now:

$$\begin{aligned}
F^+(x) = P(X^+ \le x) &= \begin{cases} 0, & x<0 \\ P(X\le x, X>0) + P(X\le 0), & x\ge 0 \end{cases} \\
&= \begin{cases} 0, & x<0 \\ P(X\in (0,x]) + F(0), & x\ge 0 \end{cases} \\
&= \begin{cases} 0, & x<0 \\ F(x), & x\ge 0. \end{cases}
\end{aligned}$$

Hence,

$$\int_0^\infty (1-F^+(x))dx = \int_0^\infty (1-F(x))dx.$$

Similarly,

$$\begin{aligned}
F^-(x) = P(X^- \le x) &= \begin{cases} 0, & x<0 \\ P(\min(X,0)\ge -x, X>0) + P(X\le 0), & x\ge 0 \end{cases} \\
&= \begin{cases} 0, & x<0 \\ P(X\ge -x), & x\ge 0. \end{cases}
\end{aligned}$$

Thus,

$$\begin{aligned}\int_0^\infty (1-F^-(x))dx &= \int_0^\infty (1-P(X\geq -x))dx\\ &= \int_0^\infty P(X<-x)dx\\ &= \int_{-\infty}^0 P(X<x)dx.\end{aligned}$$

It follows that

$$E(X)=\int_0^\infty (1-F(x))dx - \int_{-\infty}^0 P(X<x)dx$$

as claimed.

Exercise 4.4 (Quantiles)

For a given $0<\alpha<1$, we call the α-quantile of F the quantity

$$q_\alpha = \inf\{x\in\mathbb{R} : F(x)\geq \alpha\},$$

where F is a given cdf.

Remark: The function $\alpha\mapsto q_\alpha$ is also called the generalised inverse of the cdf F. If $\alpha=\frac{1}{2}$, $q_{\frac{1}{2}}$ is called the median.

(a) Show that $\alpha\mapsto q_\alpha$ is non-decreasing on $(0,1)$.

(b) Toss a fair coin n times and record

$$X_i = \begin{cases} 1 & \text{if heads at the } i\text{th toss}\\ 0 & \text{otherwise.}\end{cases}$$

Let

$$Y_n=\sum_{i=1}^n X_i$$

be the number of heads obtained in the n tosses.

Find the cdf of Y_n. Call it F_n.

(c) What is the median of F_n when n is even and when it is odd?

Solution 4.4

(a) For $\alpha'\geq\alpha$ in $(0,1)$ it is clear that

$$S_{\alpha'} := \{x\in\mathbb{R} : F(x)\geq\alpha'\} \subseteq \{x\in\mathbb{R} : F(x)\geq\alpha\} =: S_\alpha.$$

This trivially implies that

$$\begin{aligned}&\inf S_{\alpha'}\geq \inf S_\alpha\\ \Leftrightarrow\ & q_{\alpha'}\geq q_\alpha.\end{aligned}$$

(b) Here we have

$$Y_n = \sum_{i=1}^{n} X_i = \text{“number of 1's among } n \text{ possible occurrences”}$$

Since the n coins are fair and independent, the pmf of Y_n is

$$P(Y_n = j) = \frac{\binom{n}{j}}{2^n},\ j = 0, ..., n.$$

Hence, the cdf F_n is given by

$$F_n(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2^n}, & 0 \leq x < 1 \\ \frac{1+\binom{n}{1}}{2^n}, & 1 \leq x < 2 \\ ... & \\ \frac{1}{2^n}\sum_{k=0}^{j}\binom{n}{k}, & j \leq x < j+1 \\ ... & \\ 1, & x \geq n. \end{cases}$$

Y_n is called a Binomial random variable.

(c) By definition of $q_{\frac{1}{2}}$ (the median),

$$q_{\frac{1}{2}} = \inf\left\{x : F_n(x) \geq \frac{1}{2}\right\}.$$

- Consider the case where n is even, $n = 2m$. We have:

$$\frac{1}{2^n}\left[\sum_{j=0}^{m}\binom{n}{j} + \sum_{j=m+1}^{n}\binom{n}{j}\right] = 1$$

where

$$\sum_{j=m+1}^{n}\binom{n}{j} = \sum_{j=0}^{m-1}\binom{n}{n-j} = \sum_{j=0}^{m-1}\binom{n}{j}.$$

Thus,

$$\begin{aligned} &\frac{1}{2^n}\left[\sum_{j=0}^{m}\binom{n}{j} + \sum_{j=0}^{m-1}\binom{n}{j}\right] = 1 \\ \Leftrightarrow\ &\frac{1}{2^n}\left[2\sum_{j=0}^{m-1}\binom{n}{j} + \binom{n}{m}\right] = 1 \\ \Leftrightarrow\ &2F(m-1) = 1 - \frac{\binom{n}{m}}{2^n} \\ \Rightarrow\ &F(m-1) < \frac{1}{2}. \end{aligned}$$

On the other hand,

$$
\begin{aligned}
&\frac{1}{2^n}\left[\sum_{j=0}^{m-1}\binom{n}{j}+\sum_{j=m}^{n}\binom{n}{j}\right]=1\\
\Leftrightarrow\ &\frac{1}{2^n}\left[\sum_{j=0}^{m-1}\binom{n}{j}+\sum_{j=0}^{m}\binom{n}{j}\right]=1\\
\Leftrightarrow\ &2F(m)=1+\frac{\binom{n}{m-1}}{2^n}\\
\Rightarrow\ &F(m)>\frac{1}{2}.
\end{aligned}
$$

This implies that $q_{\frac{1}{2}}=m=\frac{n}{2}$.

• If n is odd, $n=2m+1$ then:

$$
\begin{aligned}
&\frac{1}{2^n}\left[\sum_{j=0}^{m}\binom{n}{j}+\sum_{j=m+1}^{n}\binom{n}{j}\right]=1\\
\Leftrightarrow\ &\frac{1}{2^n}\left[\sum_{j=0}^{m}\binom{n}{j}+\sum_{j=0}^{m}\binom{n}{n-j}\right]=1\\
\Leftrightarrow\ &2F(m)=1\\
\Leftrightarrow\ &F(m)=\frac{1}{2}.
\end{aligned}
$$

Since F is a step function, this implies that $q_{\frac{1}{2}}=m=\frac{n-1}{2}$.

Exercise 4.5 (an interesting property of expectations)

(a) Suppose that X is a random variable. Show that $E(X^2)<\infty$ if and only if $\text{var}(X)<\infty$.

(b) Suppose $\text{var}(X)<\infty$. Show that $E(X)$ minimises the function

$$a\mapsto E[(X-a)^2]\quad (a\in\mathbb{R}).$$

Solution 4.5

(a) If $\text{var}(X)<\infty$, then by definition we must have $E(X)$ well-defined. Using the formula

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

we get that

$$E(X^2)=\text{var}(X)+E(X)^2<\infty.$$

Suppose now $E(X^2)<\infty$. By the Cauchy-Schwarz inequality (or Jensen's inequality)

$$
\begin{aligned}
&E(X)^2\le E(X^2)\\
\Rightarrow\ &|E(X)|\le\sqrt{E(X^2)}<\infty
\end{aligned}
$$

Thus, by definition of existence of $E(X)$, the expectation is well-defined.

Now, this implies that $E(X^2)$ and $E(X)$ are well-defined, and therefore

$$E(X^2) - E(X)^2 = E\left[(X - E(X))^2\right] = \text{var}(X) < \infty.$$

(b) Let $\mu = E[X]$:

$$\begin{aligned} E[(X-a)^2] &= E[(X - \mu + \mu - a)^2] \\ &= E[(X-\mu)^2 + 2(X-\mu)(\mu - a) + (\mu - a)^2] \\ &= \text{var(X)} + (\mu - a)^2. \end{aligned}$$

Thus

$$E[(X-a)^2] \geq \text{var}(X)$$

with equality if and only if $a = \mu$.

Exercise 4.6 (Optional, for the more courageous)

Consider again the birthday problem from another perspective. Suppose that people are coming to a party and you are assigned the mission of writing down the birth date of each guest as they show up.

Let X be the number of people that showed up until you see for the first time a person born on the same day as somebody who showed up earlier.

(a) Find an expression for $E(X)$ (this can be done in two different ways).

(b) Find an expression for $\sigma = \sqrt{\text{var}(X)}$.

(c) The numerical values are given as

$$E(X) \approx 24.62,$$
$$\sigma \approx 12.19.$$

Find an interval $[a, b]$ which satisfies

$$P(a \leq X \leq b) \geq 0.5.$$

Hint: Use Chebyshev's inequality.

Solution 4.6 For $n = 2, ..., N + 1$ with $N = 365$, we have

$$\{X = n\} = \{\text{first } n-1 \text{ people have distinct birthdays, } n^{\text{th}} \text{ person has the same birthday as one of the previous ones}\}$$

Therefore:

$$P(N = n) = \frac{N(N-1)...(N-(n-1)+1)}{N^{n-1}} \times \frac{n-1}{N},$$

and if $n \leq 1$ or $n \geq N + 2$ then $P(X = n) = 0$. Then:

$$\begin{aligned}E(X) &= \sum_{n=2}^{N+1} nP(X=n) \\ &= \sum_{n=2}^{N+1} n\frac{N(N-1)...(N-(n-1)+1)}{N^n}(n-1) \\ &= \sum_{n=1}^{N} n(n+1)\frac{N(N-1)...(N-n+1)}{N^{n+1}}.\end{aligned}$$

We can also use the tail sum formula

$$\begin{aligned}E(X) &= \sum_{n=0}^{\infty} P(X>n) \\ &= 2+\sum_{n=2}^{N} P(X>n) \\ &= 2+\sum_{n=2}^{N} \frac{N(N-1)...(N-n+1)}{N^n}.\end{aligned}$$

(a)

$$\sigma^2 = E(X^2) - E(X)^2$$

where

$$E(X^2) = \sum_{n=2}^{N+1} n(n+1)^2\frac{N(N-1)...(N-n+1)}{N^{N+1}}.$$

Thus,

$$\sigma = \left[\sum_{n=2}^{N+1} \frac{n(n+1)^2N(N-1)...(N-n+1)}{N^{n+1}} - \left(\sum_{n=2}^{N} \frac{n(n+1)N(N-1)...(N-n+1)}{N^{n+1}}\right)^2\right]^{\frac{1}{2}}.$$

(b) Using Chebyshev's inequality, we have that

$$\begin{aligned}&P\left(\frac{|X-\mu|}{\sigma} > k\right) \leq \frac{1}{k^2} \\ \Leftrightarrow\ &P\left(\frac{|X-\mu|}{\sigma} \leq k\right) \geq 1-\frac{1}{k^2}.\end{aligned}$$

If we choose k such that $1-\frac{1}{k^2}=\frac{1}{2}$, that is $k=\sqrt{2}$, then

$$\begin{aligned}&P\left(\frac{|X-\mu|}{\sigma} \leq \sqrt{2}\right) \geq \frac{1}{2} \\ \Leftrightarrow\ &P(\mu-\sigma\sqrt{2} \leq X \leq \mu+\sqrt{2}\sigma) \geq \frac{1}{2} \\ \Leftrightarrow\ &P(X\in[7.37, 41.86]) \geq \frac{1}{2}.\end{aligned}$$

Therefore we can choose the interval $[a,b]=[7.37, 41.86]$.