

1. UNIT: AXIOMS AND ELEMENTARY PROPERTIES

- Goal
- What is probability?
- What is a probability space? (mathematical model)

1.1. Axioms of probability.

1.1.1. *Set of outcomes, sample space.*

Examples:

1.1.2. *collection of events:*

Notation (set theory): $EF = E \cap F$, $E \cup F = (E^c F^c)^c$, $\bigcup_n E_n$.

$\Sigma \subset 2^\Omega$ is called an algebra if

- (1) $\Omega \in \Sigma$.
- (2) $E \in \Sigma$ implies $E^c \in \Sigma$.
- (3) $E, F \in \Sigma$ implies $EF \in \Sigma$.

Σ is called a σ -algebra if E_1, E_2, \dots all in Σ implies

$$\bigcup_n E_n \in \Sigma .$$

(Why? later important for continuous random variables)

Examples:

1.1.3. *Probability measure.* A probability measure P is a function $P : \Sigma \rightarrow [0, 1]$ such that

$$P(\Omega) = 1$$

and for a countable collection of disjoint sets (E_n) in Σ we have

$$P\left(\bigcup_n E_n\right) = \sum_n P(E_n) .$$

Question 1: Find the definition of convergence in one of your math books.

Examples: $\Omega = \{1, \dots, m\}$ with $P(A) = \frac{|A|}{m}$. Product spaces, different weights, ordered/unordered pairs, binomial coefficients, card decks, rolling a dice, urn model, more combinatorics.

1.1.4. Elementary properties and not so elementary properties.

Proposition 1.1. (1) $P(\Omega) = 1$

(2) $P(E_1 \cup \dots \cup E_m) = \sum_{j=1}^m P(E_j)$. = holds for mutually disjoint sets.

(3) $E \subset F$ implies $P(E) \leq P(F)$.

(4) If $E_1 \subset E_2 \subset \dots$, then

$$\lim_n P(E_n) = P\left(\bigcup_n E_n\right).$$

(5) If $E_1 \supset E_2 \supset \dots$, then

$$\lim_n P(E_n) = P\left(\bigcap_n E_n\right).$$

Remark 1.2. (4) and (5) seem theoretical now, but are crucial if we want to talk about limit behaviour.

Examples:

1.2. Conditional probability.

Theory: If $P(F) > 0$, we define

$$P(E|F) = \frac{P(EF)}{P(F)}.$$

Then $P(\cdot|F)$ defines a new probability measure on Σ .

Remark: $P(EF) = P(E|F)P(F)$.

Bayes formula:

$$P(E) = P(E \cap F) + P(E \cap F^c) = P(E|F)P(F) + P(E|F^c)P(F^c).$$

(Mathematical simply, but tricky because F must be chosen according to the problem).

Example 1) Three boxes with 2 pieces in it: 2 gold, 2 silver, one box gold and silver. What is the probability of a second gold given a first gold?

Example 2) The king comes from a family of two children. What is the probability of having a sister.

Example 3) No 3a,b, 3j, p=78.

Example 4) No 10 (p=104).

1.2.1. *Independence.* Two events are independent if

$$P(EF) = P(E)P(F).$$

Proposition 1.3. *E and F independent iff E^c and F^c independent.*

Project: Odd's ratio.

Independence comes from independent coordinates

Proposition 1.4. *If E and F are independent, then there is a probability measure \tilde{P} on $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$ such that*

$$P(E) = P(\{(0, 0), (0, 1)\}), P(F) = P(\{(0, 0), (1, 0)\}), P(EF) = P(\{(0, 0)\}).$$

Example: parallel circuits

Example: No 62 b) $p=1/11$.

Problem: Player A plays against the bank. When head shows up he wins a dollar when tail shows up he loses a dollar. He starts with capital N and the game ends if he has lost the money. What is the probability for the game to stop.

The trick: conditional probability and recursion

Example: The gambler's ruin problem.

Example: more points to win (example 4i).

Example: n successes before m failures (example 5b).

Example: Laplace's rule of succession.

Introduction to probability-361

Instructions:

- Question of the day/week-webbased
- Homework (25 %)-individual submission (always Mondays), group work recommended (2-4 students), indicate joint work, ≤ 3 foto-copies. Mark one problem where you wish a detailed discussion in class.
- In groups of 2 students you have to 2 prepare a short essay explaining an application of probability to concrete problems or a theoretical exercise. Part of the grade, posted on web.

2. UNIT 1: AXIOMS AND ELEMENTARY PROPERTIES

2.1. Axioms of probability.

- Goal
- What is probability?
- What is a probability space? (mathematical model)

Axioms of probability:2.1.1. *Set of outcomes, sample space.*

Examples:

2.1.2. *collection of events:***Notation (set theory):** $EF = E \cap F$, $E \cup F = (E^c F^c)^c$, $\bigcup_n E_n$. $\Sigma \subset 2^\Omega$ is called an algebra if

- (1) $\Omega \in \Sigma$.
- (2) $E \in \Sigma$ implies $E^c \in \Sigma$.
- (3) $E, F \in \Sigma$ implies $EF \in \Sigma$.

 Σ is called a σ -algebra if E_1, E_2, \dots all in Σ implies

$$\bigcup_n E_n \in \Sigma.$$

(Why? later important for continuous random variables)

Examples:

2.1.3. *Probability measure.* A probability measure P is a function $P : \Sigma \rightarrow [0, 1]$ such that

$$P(\Omega) = 1$$

and for a countable collection of disjoint sets (E_n) in Σ we have

$$P\left(\bigcup_n E_n\right) = \sum_n P(E_n).$$

Question 1: Find the definition of convergence in one of your math books.**Examples:** $\Omega = \{1, \dots, m\}$ with $P(A) = \frac{|A|}{m}$. Product spaces, different weights, ordered/unordered pairs, binomial coefficients, card decks, rolling a dice, urn model, more combinatorics.

2.1.4. Elementary properties and not so elementary properties.

Proposition 2.1. (1) $P(\Omega) = 1$

(2) $P(E_1 \cup \dots \cup E_m) = \sum_{j=1}^m P(E_j)$. = holds for mutually disjoint sets.

(3) $E \subset F$ implies $P(E) \leq P(F)$.

(4) If $E_1 \subset E_2 \subset \dots$, then

$$\lim_n P(E_n) = P\left(\bigcup_n E_n\right).$$

(5) If $E_1 \supset E_2 \supset \dots$, then

$$\lim_n P(E_n) = P\left(\bigcap_n E_n\right).$$

Remark 2.2. (4) and (5) seem theoretical now, but are crucial if we want to talk about limit behaviour.

Examples:

2.2. Conditional probability.

Theory: If $P(F) > 0$, we define

$$P(E|F) = \frac{P(EF)}{P(F)}.$$

Then $P(\cdot|F)$ defines a new probability measure on Σ .

Remark: $P(EF) = P(E|F)P(F)$.

Bayes formula:

$$P(E) = P(E \cap F) + P(E \cap F^c) = P(E|F)P(F) + P(E|F^c)P(F^c).$$

(Mathematical simply, but tricky because F must be chosen according to the problem).

Example 1) Three boxes with 2 pieces in it: 2 gold, 2 silver, one box gold and silver. What is the probability of a second gold given a first gold?

Example 2) The king comes from a family of two children. What is the probability of having a sister.

Example 3) No 3a,b, 3j, p=78.

Example 4) No 10 (p=104).

2.2.1. *Independence.* Two events are independent if

$$P(EF) = P(E)P(F).$$

Proposition 2.3. *E and F independent iff E^c and F^c independent.*

Project: Odd's ratio.

Independence comes from independent coordinates

Proposition 2.4. *If E and F are independent, then there is a probability measure \tilde{P} on $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$ such that*

$$P(E) = P(\{(0, 0), (0, 1)\}), P(F) = P(\{(0, 0), (1, 0)\}), P(EF) = P(\{(0, 0)\}).$$

Example: parallel circuits

Example: No 62 b) $p=1/11$.

Problem: Player A plays against the bank. When head shows up he wins a dollar when tail shows up he loses a dollar. He starts with capital N and the game ends if he has lost the money. What is the probability for the end to stop.

The trick: conditional probability and recursion

Example: The gambler's ruin problem.

Example: more points to win (example 4i).

Example: n successes before m failures (example 5b).

Example: Laplace's rule of succession.

3. UNIT2: DISCRETE RANDOM VARIABLES

Motivation: Random variables are properties of collected data and they also reflect properties of experiments.

Example: Number of red cards received by a player, first time of success in a trial. No 1c), 1e)

3.0.2. Discrete versus nondiscrete sample spaces.

In many applications, we only work with finite sample spaces because we may only be able to collect finitely many data.

Definition 3.1. A set S is countable if we can write the elements in S in a list

$$S = \{s_1, s_2, \dots\}.$$

We also allow finite lists

$$S = \{s_1, \dots, s_m\}$$

and call such sets finite.

Definition 3.2. A measure space (Ω, Σ, μ) is called discrete if Ω is countable (or finite) and $\Sigma = 2^\Omega$.

Example: All possible decks with uniform probability.

Example: $p_n = e^{-\lambda} \frac{\lambda^n}{n!}$, $\Omega = \mathbb{N}_0$

$$P(A) = \sum_{n \in A} p_n.$$

Problem: If we perform infinite many trials for head and tail then

$$\Omega = \{(s_1, s_2, \dots) : s_i \in \{0, 1\}\}$$

is no longer countable. This leads to a mathematical problem and we have to exclude certain (non-constructive) events from being considered.

Solution: Consider events E_n which are constructed using information about n -trials and then pass to the limit.

Proposition 3.3. *Let (Ω, Σ, μ) be a countable probability space, then $\Omega = \{s_1, s_2, \dots, \}$ (or finite) and there exists $(p_n)_{n \geq 1}$ such that*

$$\sum_n p_n = 1$$

and

$$P(A) = \sum_{s_n \in A} p_n.$$

Definition 3.4. *A discrete random variable is given by a function $f : \Omega \rightarrow \mathbb{R}$ defined on a discrete probability space.*

Picture:

Remark 3.5. *We will later extend this definition.*

3.0.3. *distribution of a discrete random variable.*

Let $f : \Omega \rightarrow \mathbb{R}$ be a discrete random variable. The cumulative distribution function is given by

$$F(a) = \sum_{s_n: X(s_n) \leq a} P(\{s_n\}).$$

Remark:* (level sets)

Definition 3.6.

$$Prob(X = x) = \sum_{n, X(s_n) = x} P(\{s_n\}).$$

Lemma 3.7. *$Prob(f = x) = 0$ except for countably many values. Then*

$$F_X(a) = \sum_{x \leq a} Prob(f = x).$$

Examples: Poisson, step functions for tossing coin, urn model.

Definition 3.8. *(Formal definition of a discrete random variable). Let (Ω, Σ, P) be a probability space. Let function $X : \Omega \rightarrow \mathbb{R}$ is discrete if there are is a sequence $(x_n) \subset \mathbb{R}$ such that*

- (1) $E_n = \{\omega : X(\omega) = x_n\} \in \Sigma$
- (2) For every $a \in \mathbb{R}$

$$P(\{\omega : X(\omega) \leq a\}) = \sum_{x_n \leq a} P(E_n).$$

Definition 3.9. Two random variables f_1, f_2 have the same distribution of their cumulative function coincide. Then for all $x \in \mathbb{R}$

$$\text{Prob}(f_1 = x) = \text{Prob}(f_2 = x).$$

3.0.4. *Expectation and variance.*

Definition 3.10. If $X : \{s_1, s_2, \dots\} \rightarrow \mathbb{R}$ is a discrete random variable, then

$$E(X) = \sum_n X(s_n)P(\{s_n\})$$

is called the expectation. Moreover,

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

is called the variance.

$$SD(X) = \sqrt{\text{Var}(X)}$$

is called the standard deviation.

Example: Tossing a coin, λ -Poisson.

Proposition 3.11. (*Properties of expectation*)

- (1) $E(X_1 + \lambda X_2) = E(X_1) + \lambda E(X_2)$,
- (2) $E(g(X)) = \sum_n g(X(s_n))P(\{s_n\}) = \sum_x g(x)\text{Prob}(X = x)$,
- (3) $a_n = X(s_n)$,

$$E(g(X)) = \sum_n g(a_n)\text{Prob}(X = a_n).$$

Proposition 3.12. *The moment generating function is given by*

$$M_X(t) = E(e^{tX}).$$

Then

$$E(X) = M'_X(0), \quad E(X^2) = M''_X(0).$$

Remark 3.13. *Two meanings of a constant.*

3.0.5. *Bernoulli, binomial, geometric, hypergeometric, Poisson.*

Bernoulli and Binomial: We toss a coin, with probability p , we find head, with probability $1 - p$ we find tail. We define X to be 1 for head and -1 for tail.

$$EX = p + 1 - p = 1 - 2p$$

$$EX^2 = 1, \text{Var}(X) = 1 - (1 - 2p)^2 = 4p - 4p^2.$$

Moreover,

$$Ee^{tX} = pe^t + (1 - p)e^{-t} = e^{-t} + 2p \sinh(t).$$

For $n \in \mathbb{N}$, we denote by X_n the gain after n trials

$$EX_n = \sum_{i=1}^n (1 - 2p) = n(1 - 2p).$$

Also

$$Ee^{tX_n} = (pe^t + (1 - p)e^{-t})^n.$$

Alternatively, we may consider Y being 1 for tail and 0 for head.

$$Ee^{tY} = pe^t + (1 - p) = 1 - p(1 - e^t).$$

$$Ee^{t \sum_{i=1}^n Y_i} = (1 - p(1 - e^t))^n.$$

The distribution of

$$Z_n = \sum_{i=1}^n Y_i$$

is called *binomial with parameters* (n, p)

$$\text{Prob}(Z_n = k) = p^k (1 - p)^{n-k} \binom{n}{k}.$$

Important in many applications.

Geometric: Let us denote by W the waiting time for the first success. Then by independence

$$\text{Prob}(W = k) = p(1 - p)^{k-1}.$$

W is called *geometrically* distributed with parameter p . let $q = 1 - p$, then

$$EW = \sum_{k=1}^{\infty} pk(1 - p)^{k-1} = p \sum_{k=0}^{\infty} (k + 1)q^k$$

With the differentiation trick one can calculate this variable.

$$Ee^{tW} = pe^t \sum_{k=1}^{\infty} e^{t(k-1)} q^{k-1} = \frac{pe^t}{1 - qe^t} = \frac{p}{e^{-t} - q}.$$

Negative binomial: Let T_r be the random variable in a Bernoulli such that $T_r = k$ if the r 'th success is attained in the k -th step. Then $T_r = k$ if and only if

$$Y_1 + \dots + Y_{k-1} = r - 1 \quad \text{and} \quad Y_k = r.$$

By independence

$$\begin{aligned} \text{Prob}(T_r = k) &= p \text{Prob}(Y_1 + \dots + Y_{k-1} = r - 1) \\ &= p \sum_{|A|=r-1, A \subset \{1, \dots, k-1\}} p^{r-1} (1-p)^{k-r} \\ &= \binom{k-1}{r-1} (1-p)^{k-r} p^r. \end{aligned}$$

This is called the *negative binomial* distribution with parameters (r, p) .

hypergeometric:

Poisson:

3.1. Unit 3.

In contrast to discrete random variables for the definition of continuous random variables the σ -algebra Σ matters. Measure theory tells us there is no good measure on the real line which works for all sets.

Example: Let $\Omega = \{0, 1\}^{\mathbb{N}} = \{s_1, \dots | s_i \in \{-1, 1\}\}$ the sets of infinite trials of head and tail (Bernoulli experiment) and $0 < p < 1$. Let a_1, \dots, a_m be a fixed finite sequence of outcomes. Certainly, we know that

$$\text{Prob}(\{s_1 = a_1, \dots, s_m = a_m\}) = p^{\text{number of 1's}} (1-p)^{\text{number of 0's}}.$$

Now, we consider Σ to be the smallest σ -algebra containing all the events

$$E_{a_1, \dots, a_m} = \{s_1, \dots, s_m, s_{m+1} : s_1 = a_1, \dots, s_m = a_m\}.$$

Thus Σ contains all the events we want to talk about and is closed under complements and infinite unions.

Example: $\Omega = \mathbb{R}$. We definitely know

$$|[a, b]| = b - a$$

This we define \mathcal{B} be the smallest σ -algebra containing all intervals and find a measure $\lambda : \Sigma \rightarrow [0, \infty)$ such that

$$\lambda([a, b]) = b - a .$$

Here, λ is no longer a probability measure. Moreover, there is a new feature

$$\lambda(\{x\}) = 0$$

for every $x \in \mathbb{R}$.

Example: $\Omega = \mathbb{R}^2$. Same trick:

$$\lambda([a, b] \times [c, d]) = |b - a||d - c| .$$

Short review on measure theory.

Unit 1: Axioms and elementary properties

-What is probability? Axioms of probability. Examples from combinatorics (counting). -Conditional probability and properties. Independence. Recursive calculations and everyday problems.

Unit 2: Discrete random variables

-What are random variables? Discrete probability spaces and discrete random variables, distribution, Expectation, Variation and standard deviation. Moment generating function -Bernoulli, binomial, geometric, negative binomial, hypergeometric, Poisson.

Unit 3: Continuous random variables

-Why do we need sigma-algebras? A minimum of measure and integration. Distribution and density function, expectation, variation and moment generating functions, -Uniform, normal, exponential, Gamma, Beta, Cauchy distribution

Unit 4: Joint distribution, independence and conditional expectation

-How do we handle two variables simultaneously? Joint distribution. What is a conditional expectation (discrete and continuous case)? How can we use conditional

expectations? Building random variables on others. -Expectations and calculating expectations using conditional expectation, best prediction.

Unit 5: Limit theorems

-Central limit theorem and law of large numbers, statement and use, proof using moment generating functions.

Independence comes from independent coordinates

Proposition 3.14. *If E and F are independent, then there is a probability measure \tilde{P} on $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$ such that*

$$P(E) = \tilde{P}(\{(0, 0), (0, 1)\}), P(F) = \tilde{P}(\{(0, 0), (1, 0)\}), P(EF) = \tilde{P}(\{(0, 0)\}).$$

Proof. Define

$$\tilde{P}(\{(0, 0)\}) = P(E)P(F), \tilde{P}(\{(0, 1)\}) = P(E)P(F^c),$$

and

$$\tilde{P}(\{(1, 0)\}) = P(E^c)P(F), \tilde{P}(\{(1, 1)\}) = P(E^c)P(F^c).$$

Let

$$\tilde{E} = \{(0, 0), (0, 1)\}$$

and

$$\tilde{F} = \{(0, 0), (1, 0)\}$$

Then

$$P(EF) = P(\tilde{E}\tilde{F})$$

if and only if

$$P(EF) = P(E)P(F).$$

This proves the assertion. ■

Corollary 3.15. *E and F are independent if and only if E^c and F^c are independent.*

Remark on conditional expectation

If X and Y are variables, then

$$E[X|Y]$$

is a random variable of the form $E[X|Y] = g(Y)$ for a suitable function g .

$$g(y) = E[X|Y = y] = \lim_{\varepsilon \rightarrow 0} \frac{E[X|y - \varepsilon \leq Y \leq y]}{P(y - \varepsilon \leq Y \leq y)}.$$

Similarly, we have formula for the conditional distribution function. For fixed y $E[X|Y = y]$ is a random variable such that

$$\begin{aligned} P(E[X|Y = y] \leq t) &= \lim_{\varepsilon \rightarrow 0} P(X \leq t | y - \varepsilon \leq Y \leq y) \\ &= \lim_{\varepsilon \rightarrow 0} \frac{P(X \leq t, y - \varepsilon \leq Y \leq y)}{P(y - \varepsilon \leq Y \leq y)}. \end{aligned}$$

Example 1: $Y = \sum_k a_k 1_{A_k}$ is discrete, the A_k 's are disjoint and the a_k 's are mutually different. In other words $P(Y = a_k) = P(A_k)$. Then for $t = a_k$

$$\begin{aligned} P(E[X|Y = a_k] \leq t) &= \lim_{\varepsilon \rightarrow 0} \frac{P(X \leq t, a_k - \varepsilon \leq Y \leq a_k)}{P(a_k - \varepsilon \leq Y \leq a_k)} \\ &= \frac{P(X \leq t, Y = a_k)}{P(Y = a_k)} = P(X \leq t | Y = a_k). \end{aligned}$$

For $y \neq a_k$ the limit is undefined. Similarly,

$$E[X|Y = a_k] = \lim_{\varepsilon \rightarrow 0} \frac{E[X|a_k - \varepsilon \leq y \leq a_k]}{P(a_k - \varepsilon \leq Y \leq a_k)} = E[X|A_k] = \frac{E[X 1_{A_k}]}{P(A_k)}.$$

Thus $g(a_k) = E[X|A_k]$ satisfies

$$E[X|Y] = g(Y).$$

Example 2: $\Omega = \mathbb{R}^2$.

$$P(A) = \int_A f_{X,Y}(x,y) dx dy.$$

$X(x,y) = x, Y(x,y) = y$. Then we see that

$$\begin{aligned}
P(X \leq t|Y = y) &= \lim_{\varepsilon \rightarrow 0} \frac{\int_{-\infty}^t \int_{y-\varepsilon}^y f_{X,Y}(x, y) dy dx}{\int_{-\infty}^{\infty} \int_{y-\varepsilon}^y f_{X,Y}(x, y) dy dx} = \lim_{\varepsilon \rightarrow 0} \frac{\frac{1}{\varepsilon} \int_{-\infty}^t \int_{y-\varepsilon}^y f_{X,Y}(x, y) dy dx}{\frac{1}{\varepsilon} \int_{-\infty}^{\infty} \int_{y-\varepsilon}^y f_{X,Y}(x, y) dy dx} \\
&= \frac{\int_{-\infty}^t f_{X,Y}(x, y) dx}{\int_{-\infty}^{\infty} f_{X,Y}(x, y) dx} = \int_{-\infty}^t f_{X|Y}(x|y) dx .
\end{aligned}$$

Therefore the density of $E[X|Y = y]$ is given by $f_{X|Y}(x|y)$

$$P(E[X|Y = y] \in A) = \int_A f_{X|Y}(x|y) dx .$$

Similarly,

$$\begin{aligned}
E(X|Y = y) &= \lim_{\varepsilon \rightarrow 0} \frac{\int_{-\infty}^{\infty} \int_{y-\varepsilon}^y x f_{X,Y}(x, y) dy dx}{\int_{-\infty}^{\infty} \int_{y-\varepsilon}^y f_{X,Y}(x, y) dy dx} \\
&= \lim_{\varepsilon \rightarrow 0} \frac{\frac{1}{\varepsilon} \int_{-\infty}^{\infty} \int_{y-\varepsilon}^y x f_{X,Y}(x, y) dy dx}{\frac{1}{\varepsilon} \int_{-\infty}^{\infty} \int_{y-\varepsilon}^y f_{X,Y}(x, y) dy dx} \\
&= \frac{\int_{-\infty}^{\infty} x f_{X,Y}(x, y) dx}{\int_{-\infty}^{\infty} f_{X,Y}(x, y) dx} = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx .
\end{aligned}$$

Therefore

$$g(y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx .$$

Conclusion: If X, Y are arbitrary random variables with joint distribution function $f_{X,Y}$ and

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

then

$$E(X|Y)(\omega) = g(Y(\omega)) .$$

Application: How to recover X from the conditional distribution: In the discrete case, we have

$$P(X \leq t) = \sum_k P(X \leq t, Y = a_k) = \sum_k P(X \leq t|A_k)P(A_k)$$

In the continuous case, we fix $\varepsilon > 0$

$$\begin{aligned} P(X \leq t) &= \sum_{k \in \mathbb{Z}} P(X \leq t, (k-1)\varepsilon < Y \leq k\varepsilon) \\ &= \sum_{k \in \mathbb{Z}} P(X \leq t|(k-1)\varepsilon < Y \leq k\varepsilon)P((k-1)\varepsilon < Y \leq k\varepsilon) \\ &= \int_{\mathbb{R}} \sum_{k \in \mathbb{Z}} 1_{[(k-1)\varepsilon, k\varepsilon]}(y) P(X \leq t|(k-1)\varepsilon < Y \leq k\varepsilon) f_Y(y) dy \\ &\rightarrow_{\varepsilon \rightarrow 0} \int_{\mathbb{R}} P(X \leq t|Y = y) f_Y(y) dy \end{aligned}$$

This implies

$$P(X \in A) = \int_{\mathbb{R}} P(X \in A|Y = y) f_Y(y) dy .$$

Similarly,

$$P(X \in A) = E(1_A(X)) = E(P(X \in A)|Y) \quad \text{and} \quad E(X) = E(E[X|Y]) .$$

Theorem: $\min_h E|X - h(Y)|^2 = E|X - E[X|Y]|^2$

Proof: Fix y and write $g(y) = E[X|Y = y]$. Then

$$\begin{aligned} E[|X - h(Y)|^2|Y = y] &= E[|X - h(y)|^2|Y = y] = E[|X - g(y) + g(y) - h(y)|^2|Y = y] \\ &= E[|X - g(y)|^2] + 2E[(X - g(y))(g(y) - h(y))|Y = y] + (g(y) - h(y))^2 \\ &= E[|X - g(y)|^2] + 2E[X - g(y)|Y = y](g(y) - h(y)) + (g(y) - h(y))^2 \\ &= E[|X - g(y)|^2] + 2(E[X|Y = y] - g(y))(g(y) - h(y)) + (g(y) - h(y))^2 \\ &= E[|X - g(y)|^2] + (g(y) - h(y))^2 . \end{aligned}$$

Integrating over $y = Y(\omega)$ yields

$$E(|X - h(Y)|^2) = E(|X - g(Y)|^2) + E(g(Y) - h(Y))^2 \geq E(|X - g(Y)|^2) .$$

The minimum is attained for $h = g$. ■

4. INDEPENDENT DISCRETE RANDOM VARIABLES

We say that discrete random variables X and Y are independent if

$$\text{Prob}(X = x, Y = y) = \text{Prob}(X = x) \text{Prob}(Y = y)$$

holds for all $x, y \in \mathbb{R}$.

Remark 4.1. Let X, Y be independent and $g, h : \mathbb{R} \rightarrow \mathbb{R}$ functions. Then $g(X), h(Y)$ are independent.

Proof. Let $z \in \mathbb{R}$ and $L_z(g) = \{x \in \mathbb{R} : g(x) = z\}$. Let $w \in \mathbb{R}$ and $L_w(h) = \{y \in \mathbb{R} : h(y) = w\}$. (Think of $g(X) = X^2$ then you have two solutions with $x^2 = z$). Then we have

$$\begin{aligned} \text{Prob}(g(X) = z, h(Y) = w) &= \sum_{x \in L_z(g), y \in L_w(h)} \text{Prob}(X = x, Y = y) \\ &= \sum_{x \in L_z(g), y \in L_w(h)} \text{Prob}(X = x) \text{Prob}(Y = y) \\ &= \sum_{x \in L_z(g)} \text{Prob}(X = x) \sum_{y \in L_w(h)} \text{Prob}(Y = y) \\ &= \text{Prob}(g(X) = z) \text{Prob}(h(Y) = w). \quad \blacksquare \end{aligned}$$

Proposition 4.2. Let X and Y be independent. Then

$$EXY = EXEY.$$

Proof. Indeed,

$$\begin{aligned} EXY &= \sum_{z \in \mathbb{R}} z \text{Prob}(XY = z) = \sum_{z \in \mathbb{R}} z \sum_{x, y \in \mathbb{R}, xy=z} \text{Prob}(X = x, Y = y) \\ &= \sum_{z \in \mathbb{R}} z \sum_{x, y \in \mathbb{R}, xy=z} \text{Prob}(X = x) \text{Prob}(Y = y) \\ &= \sum_{x, y \in \mathbb{R}} xy \text{Prob}(X = x) \text{Prob}(Y = y) \\ &= \sum_x x \text{Prob}(X = x) \sum_{y \in \mathbb{R}} y \text{Prob}(Y = y) \\ &= EX EY. \quad \blacksquare \end{aligned}$$

Let us recall the moment generating function

$$M_X(t) = Ee^{tX}.$$

Then we have

$$M_X^{(k)}(t) = EX^k e^{tX}.$$

for all $k \in \mathbb{N}$. In particular

$$M_X^{(k)}(0) = EX^k.$$

Application: *If X and Y are independent, then*

$$Ee^{t(X+Y)} = Ee^{tX}e^{tY} = Ee^{tX}Ee^{tY}.$$

Therefore

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

By induction we have for independent random variables X_1, \dots, X_n

$$M_{X_1+\dots+X_n}(t) = \prod_{j=1}^n M_{X_j}(t).$$

This is in particular interesting if

$$\text{Prob}(X_j = x) = \text{Prob}(X = x)$$

is given by the point mass function of a fixed random variable X . Then, we get

$$(4.1) \quad M_{X_1+\dots+X_n}(t) = M_X(t)^n.$$

For example the moment generating function for a geometric random variable with parameter p is given by

$$M_X(t) = \frac{e^t p}{1 - e^t(1-p)} = \frac{p}{e^{-t} - (1-p)}.$$

We get

$$M'_X(t) = \frac{pe^{-t}}{(e^{-t} - (1-p))^2},$$

and

$$M''_X(t) = \frac{2pe^{-2t}}{(e^{-t} - (1-p))^3} - \frac{pe^{-t}}{(e^{-t} - (1-p))^2}.$$

In particular

$$EX = M'_X(0) = \frac{1}{p}$$

and

$$EX^2 = M_X''(0) = \frac{2p}{p^3} - \frac{p}{p^2} = \frac{2}{p^2} - \frac{1}{p}.$$

The variance is

$$\text{Var}(X) = EX^2 - (EX)^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1}{p^2} - \frac{1}{p} = \frac{1-p}{p^2}.$$

The negative binomial distribution is given by

$$W_r = \sum_{j=1}^r X_j$$

where X_j are independent geometrically distributed with parameter p . We have

$$\text{Prob}(W_r = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

for $k = r, r+1, \dots$. For the moment generating function we get

$$M_{W_r}(t) = \left[\frac{p}{e^{-t} - (1-p)} \right]^r.$$

This gives

$$M'_{W_r}(t) = r \left[\frac{p}{e^{-t} - (1-p)} \right]^{r-1} \frac{pe^{-t}}{(e^{-t} - (1-p))^2},$$

and

$$\begin{aligned} M''_{W_r}(t) &= r(r-1) \left[\frac{p}{e^{-t} - (1-p)} \right]^{r-1} \frac{p^2 e^{-2t}}{(e^{-t} - (1-p))^4} \\ &+ r \left[\frac{p}{e^{-t} - (1-p)} \right]^{r-1} \frac{2pe^{-2t}}{(e^{-t} - (1-p))^3} - \frac{pe^{-t}}{(e^{-t} - (1-p))^2}. \end{aligned}$$

This gives

$$EW_r = rp$$

and

$$EW_r^2 = r(r-1) \frac{p^2}{p^4} + r \left(\frac{2}{p^2} - \frac{1}{p} \right).$$

This is not a coincidence!

$$EW_r = \sum_{j=1}^r EX_j = \frac{r}{p}$$

Moreover, by independence

$$\begin{aligned}
 EW_r^2 &= E\left(\sum_{j=1}^r X_j\right)^2 = \sum_{j,k=1}^r EX_j X_k \\
 &= \sum_{j=1}^r EX_j^2 + \sum_{j \neq k \in \{1, \dots, r\}} EX_j EX_k \\
 &= rEX^2 + r(r-1)(EX)^2 \\
 &= r\left(\frac{2}{p^2} - \frac{1}{p}\right) + r(r-1)\frac{1}{p} \\
 &= \frac{2r}{p^2} + \frac{r(r-2)}{p}.
 \end{aligned}$$

Both methods give the result and that minimizes the chance that we make a mistake.

Crash course on integration

5. RIEMANN INTEGRAL

Let $f : [a, b] \rightarrow \mathbb{R}$ be a function. For any partition $\Delta = \{a = a_0 < \dots < a_n = b\}$, the lower sum is defined by

$$s(\Delta, f) = \sum_{i=0}^{n-1} \min\{f(x) : a_i \leq x \leq a_{i+1}\}(a_{i+1} - a_i)$$

and the upper sum is defined by

$$S(\Delta, f) = \sum_{i=0}^{n-1} \max\{f(x) : a_i \leq x \leq a_{i+1}\}(a_{i+1} - a_i).$$

If f is integrable then

$$s(\Delta, f) \leq \int_a^b f(x)dx \leq \inf_{\Delta} S(\Delta, f)$$

and for every $\varepsilon > 0$ there exists a partition such that

$$S(\Delta, f) - s(\Delta, f) < \varepsilon.$$

Example: Consider $f(x) = \frac{1}{x}$ and $\Delta_n = \{1, 2, \dots, n-1, n\}$. Then we have

$$\ln n = \int_1^n f(x)dx \leq S(\Delta_n, f) = \sum_{k=1}^{n-1} \frac{1}{k}.$$

But also, we have

$$\ln n = \int_1^n f(x)dx \geq S(\Delta_{n-1}, f) = \sum_{k=1}^{n-1} \frac{1}{k+1} = \sum_{k=1}^n \frac{1}{k} - 1.$$

This gives

$$\ln n \leq \sum_{k=1}^n \frac{1}{k} \leq 1 + \ln n.$$

This tells us that $\sum_{k=1}^n \frac{1}{k}$ grows like $\ln n$ and in particular $\sum_k 1/k$ is ∞ .

If we have an interval $[a, b] \subset [0, 1]$, we can define the function

$$1_{[a,b]}(x) = \begin{cases} 1 & \text{if } a \leq x \leq b \\ 0 & \text{else} \end{cases}.$$

You will be not surprised to see that

$$\int f(x)dx = b - a$$

and $b - a$ is the length of the interval. Therefore for a subset $A \subset [0, 1]$, we could try the same: Define

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases}.$$

Although this is a good idea for intervals and we get

$$P(A) = L(A) = b - a$$

for intervals, we have a problem with $A = \mathbb{Q} \cap [0, 1]$. There we have

$$\min\{1_A(x) : a_i \leq x \leq a_{i+1}\} = 0$$

and

$$\max\{1_A(x) : a_i \leq x \leq a_{i+1}\} = 1.$$

This gives

$$s(\Delta, 1_A) = 0 \quad \text{and} \quad S(\Delta, 1_A)$$

for any partition. This means that 1_A is not integrable in the Riemann sense.

Riemann integrals don't provide a good answer for $P(\mathbb{Q} \cap [0, 1])$

The right answer follows from the axioms of probability. Let $a \in [0, 1]$ be a point.

Then, we have

$$P(\{a\}) = P\left(\bigcap_n [a, a + \frac{1}{n}]\right) = \lim_n P([a, a + \frac{1}{n}]) = \lim_n \frac{1}{n} = 0.$$

Since $\mathbb{Q} \cap [0, 1]$ is a countable set there is a list r_1, r_2, \dots such that

$$\{r_1, r_2, \dots\} = [0, 1] \cap \mathbb{Q}.$$

Then the second axiom gives

$$P([0, 1] \cap \mathbb{Q}) = \sum_j P(\{r_j\}) = 0.$$

So there are many more irrationals than rational numbers. (Let's hope there are more rational students than irrational students though).

6. GENERAL INTEGRAL

In the following (Ω, Σ, P) is a probability space. We may consider for example $\Omega = [0, 1]$ and

$$P([c, d]) = d - c$$

and Σ the smallest σ -algebra which contains all the intervals. Or we may have $\Omega = [a, b]$ and

$$P([c, d]) = \frac{d - c}{b - a}.$$

The third example is by far the most important one $\Omega = \mathbb{R}$ and

$$P([c, d]) = \int_c^d \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx.$$

Here the function

$$f(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

is called the density function for the normal distribution (see details) later. In the following, we want to define how to define integration in this setting.

Let (Ω, Σ, P) be a probability space. Our motivation is again simple. For a set $A \in \Sigma$, we consider again

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases}$$

and require

$$E1_A = \int_A dP = P(A).$$

A random variable $f : \Omega \rightarrow \mathbb{R}$ is called simple function ($f \in S$) if there exists A_1, \dots, A_n and r_1, \dots, r_n such that

$$f(\omega) = \sum_{i=1}^n r_i 1_{A_i}(\omega).$$

For a simple function

$$Ef = \int_{\Omega} f dP = \sum_{i=1}^n r_i P(A_i).$$

Lemma 6.1. *Let $f, g \in S$, then*

$$\int (f + tg) dP = \int f dP + t \int g dP.$$

Proof: Assume $f = \sum_{i=1}^n r_i 1_{A_i}$, $g = \sum_{j=1}^n s_j 1_{B_j}$, then

$$f + tg = \sum_{i=1}^n r_i 1_{A_i} + \sum_{j=1}^n t s_j 1_{B_j}$$

and thus

$$E(f + tg) = \sum_{i=1}^n r_i P(A_i) + t \sum_{j=1}^n s_j P(B_j) = E(f) + tE(g).$$

The real problem is to show that $\int f dP$ is well-defined, but that's not your problem. ■

Proposition 6.2. *The map $\int : S \rightarrow \mathbb{R}$ has the following properties*

- (1) $\int (f + tg) dP = \int f dP + t \int g dP$,
- (2) $|\int f dP| \leq \int |f| dP$,
- (3) $\min(f)P(\Omega) \leq \int f dP \leq \max(f)P(\Omega)$.

Now, we want to define the integral for arbitrary rv's. We assume that $f : \Omega \rightarrow \mathbb{R}$ is measurable. Let $\varepsilon > 0$ and for $k \in \mathbb{Z}$ we define

$$A_{k,\varepsilon} = \{\omega \in \Omega : k\varepsilon < f(\omega) \leq (k+1)\varepsilon\} \in \mathcal{F}$$

We say that f is *integrable* if

$$\sum_{k \in \mathbb{Z}} \max\{1, |k\varepsilon|\} P(A_{k,\varepsilon})$$

is finite for all ε . In that case, we define

$$\int f dP = \lim_{\varepsilon \rightarrow 0} \sum_{k \in \mathbb{Z}} k\varepsilon P(A_{k,\varepsilon}).$$

Remark 6.3. 1) For simple functions this gives the same value.

2) (For experts) For every integrable functions there exists a sequence (f_n) of simple functions such that

$$\lim_n \int |f_n - f| dP = 0$$

and

$$\int f dP = \lim_n \int f_n dP.$$

3) Let \mathcal{B} the smallest σ -algebra generated by intervals in \mathbb{R} . Then there exists a measure $\mu : \mathcal{B} \rightarrow [0, \infty]$ such that

$$\mu(I) = |I|$$

holds for all intervals. The same definition replacing P with μ applies to measurable functions $f : \mathbb{R} \rightarrow \mathbb{R}$.

Let us show 1) for convenience. Let $f : \Omega \rightarrow \mathbb{R}$ a simple function. Then $f(\Omega) = \{a_1, \dots, a_n\}$ has only finitely many values. Define $B_i = \{\omega : f(\omega) = a_i\}$. Let $\varepsilon_0 = \min_{i \neq j} |a_i - a_j|$. If $\varepsilon < \varepsilon_0$, then every a_i belongs to at most one interval $(k\varepsilon, (k+1)\varepsilon)$. Therefore, we have exactly n elements $k_1, \dots, k_n \in \mathbb{Z}$ such that

$$A_{k_i,\varepsilon} = B_i$$

for all $i = 1, \dots, n$. This implies

$$\left| \sum_{k \in \mathbb{Z}} k\varepsilon P(A_{k,\varepsilon}) - \sum_{i=1}^n a_i P(B_i) \right| = \left| \sum_{i=1}^n (k_i\varepsilon - a_i) P(B_i) \right|$$

$$\begin{aligned} &\leq \sum_{i=1}^n (a_i - k_i \varepsilon) P(B_i) \\ &\leq \varepsilon \sum_{i=1}^n P(B_i) \leq \varepsilon . \end{aligned}$$

Of course, we get

$$\lim_{\varepsilon \rightarrow 0} \sum_k k \varepsilon P(A_{k\varepsilon}) = \sum_{i=1}^n a_i P(A_i) = E(f) .$$

In general, we have the following theorem.

Theorem 6.4. *Let $L^1(\Omega, P)$ the set of integrable functions. Let $f, g \in L^1(\Omega, P)$.*

Then

(1) $f + tg \in L^1(\Omega, P)$ and

$$\int (f + tg) dp = \int f dp + t \int g dp,$$

(2) $|\int f dp| \leq \int |f| dp,$

(3) $\min(f)P(\Omega) \leq \int f dp \leq \max(f)P(\Omega),$

(4) *Let $h \in L^1(\Omega, P)$ be positive, f and (f_n) rv's such that $|f_n| \leq h$ for all $n \in \mathbb{N}$ and there exists a set $A \subset \Omega$ such that $P(A) = 0$*

$$f(\omega) = \lim_n f_n(\omega)$$

for all $\omega \in A^c$. Then

$$\lim_n \int f_n dP = \int f dP .$$

One of the important properties is that continuous function on intervals are integrable and

$$\int_0^1 f(x) dx = \int f dP = E(f) .$$