1 Basics

Random experiment: outcome cannot be predicted with certainty

- Ex: roll a die

Sample space S: collection of all possible outcomes

- We always assume that S is known
- Ex: $S = \{1, 2, 3, 4, 5, 6\}$

Event A: a part of the collection of all possible outcomes

- $A \subset S$
- Event A occurred if the outcome of the experiment is in A
- Ex: $A = \{2, 4, 6\}$

Empty set/null set $\emptyset$ Union $\cup$ Intersection $\cap$ Complement $A^C = \{x \in S \mid x \notin A\} = S \setminus A$ (A' or $\bar{A}$) $S^C = \emptyset$, $\emptyset^C = S$

$A_1, \ldots, A_n$ are mutually exclusive if $A_i \cap A_j = \emptyset$ for all $i, j \in \{1, \ldots, n\}$ where $i \neq j$ $A_1, \ldots, A_n$ are exhaustive if $A_1 \cup \cdots \cup A_n = S$

Commutative laws

- $A \cup B = B \cup A$
- $A \cap B = B \cap A$

Associative laws

- $(A \cup B) \cup C = A \cup (B \cup C)$
- $(A \cap B) \cap C = A \cap (B \cap C)$

Distributive laws

- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

De Morgan's laws:

- $(A \cup B)^C = A^C \cap B^C$
- $(A \cap B)^C = A^C \cup B^C$

Goal: define the probability of an event A Idea: repeat the experiment n times

$\mathcal{N}(A)$: number of times the event A has occurred (frequency of event A) $\frac{\mathcal{N}(A)}{n}$: relative frequency of event A

Idea: relative frequency of event $A \approx$ probability of A for large n We associate A with p, which is the number about which the relative frequency tends to stabilize p is called the probability of event A

Define: a function $P(A)$ that is evaluated for a set is called a set function

Define: probability is a real-valued set function P that assigns to each event A in the sample space S a number $P(A)$ called the probability of event A, such that

- $P(A) \geq 0$ for all $A \subseteq S$
- $P(S) = 1$
- If $A_1, A_2, \ldots$ are events and $A_i \cap A_j = \emptyset$ for $i \neq j$, then $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$ (countably additive)

Theorem:

1. $P(\emptyset) = 0$
2. $P(A_1 + A_2 + \ldots) = P(A_1) + P(A_2) + \ldots$ if $A_1, A_2, \ldots$ are events such that $A_i \cap A_j = \emptyset$ for $i \neq j$ (mutually exclusive)
3. For each event A, $P(A^C) = 1 - P(A)$
4. For events A, B with $A \subseteq B$, one has $P(A) \leq P(B)$
5. For each event A, $0 \leq P(A) \leq 1$
6. If A, B are two events, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   (a) If A, B, C are three events, $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Proofs:

1. Take $A_1 = S, A_2 = \emptyset, A_3\emptyset, \ldots, A_n = \emptyset$ are mutually exclusive, so $1 = P(S) = P(A_1 + \ldots) = P(A_1) + \cdots = 1 + P(A_2) + \cdots \to P(A_2) + \cdots = 0$ so $P(\emptyset) = 0$
2. Take $B_1 = A_1, \ldots, B_k = A_k, B_{k+1} = \emptyset, \ldots,$ if $A_1, \ldots, A_k$ are mutually exclusive, then so are $B_1, \ldots,$ $P(A_1 + \ldots) = P(B_1 + \ldots) = P(B_1) + \cdots = P(B_1) + \cdots + P(B_k) + P(B_{k+1}) + \cdots = P(A_1) + \cdots + P(A_k)$ (finite additive)
3. $1 = P(S) = P(A \cup A^C) = P(A) + P(A^C)$, $P(A^C) = 1 - P(A)$
4. $P(B) = P(A \cup (B \cap A^C)) = P(A) + P(B \cap A^C) \geq P(A)$
5. $P(A) \leq P(S) = 1$ since $A \subseteq S$
6. $A \cup B = A \cup (A^C \cap B)$, $P(A \cup B) = P(A) + P(A^C \cap B)$, $B = (A \cap B) \cup (A^C \cap B)$, $P(B) = P(A \cap B) + P(A^C \cap B)$, $P(A^C \cap B) = P(B) - P(A \cap B)$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2 Methods of enumeration

Define: if each outcome has the same probability of occurring, we say that the outcomes are equally likely, that is $P(\{e_i\}) = \frac{1}{m}$ for all $i \in \{1, \ldots, m\}$, and $P(A) = \frac{|A|}{m}$

Multiplication principle: experiment E_1 has n_1 outcomes, and for each outcome, experiment E_2 has n_2 outcomes, the composite experiment E_1E_2 has n_1n_2 outcomes

Permutation: we fill n positions with n objects

- n choices for the first object
- $n-1$ choices for the second object
- $n-2$ choices for the third object
- $\ldots$
- 1 option for the last object

$n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 1 = n!$ possibilities Each of the $n!$ arrangements of n different objects is called a permutation of the n objects

We fill r positions with n objects

- n choices for the first object
- $n-1$ choices for the second object
- $n-2$ choices for the third object
- $\ldots$
- $n-r+1$ choices for the r-th object

$n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot (n-r+1) = \frac{n!}{(n-r)!} = nPr$ possibilities Each arrangement is called a permutation of n objects taken r at a time

Define: if r objects from a set of n objects are selected and the order is noted, the selected set of r objects is called an ordered sample of size r

Define: sampling with replacement occurs when an object is selected and then replaced before the next object is selected of possible ordered samples of size r taken from a set of n objects is n^r when sampling with replacement

Define: sampling without replacement occurs when an object is not replaced after it has been selected of possible ordered samples of size r taken from a set of n objects is $\frac{n!}{(n-r)!}$ when sampling without replacement

We write $\binom{n}{r} = nCr$ for the number of subsets with r elements from a set with n elements $nCr = \frac{n!}{(n-r)!r!}$ Each of the $\binom{n}{r}$ subsets is called a combination of n objects taken r at a time

	Early	Late	Total
Red	5	8	13
Yellow	3	4	7
Total	8	12	20

Binomial coefficient: $\binom{n}{r}$ Binomial formula: $(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}$

Each of the $\binom{n}{r}$ arrangements of n objects, r of type A, $n-r$ of type B, is called a distinguishable permutation

If $a = b = 1$, $(1+1)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r} = \sum_{r=0}^{n} \binom{n}{r} = 2^n$

3 Conditional probability

Ex: 20 tulip bulbs, 8 bloom early and 12 bloom late, 13 are red and 7 are yellow, 5 are early and red Probability of red given early $P(R|E) = \frac{5}{8} = \frac{N(R \cap E)}{N(E)} = \frac{N(R \cap E)/20}{N(E)/20} = \frac{P(R \cap E)}{P(E)}$

Define: the conditional probability of an event A, given that event B has occurred, is defined by $P(A|B) = \frac{P(A \cap B)}{P(B)}$ provided that $P(B) > 0$

Ex: two fair 4-sided dice are rolled $A = \{\text{the sum is 3}\}$ $B = \{\text{the sum is 3 or 5}\}$ $A \subseteq B$ $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{2/16}{6/16} = \frac{1}{3}$

Theorem: $P(\cdot|B)$ is a probability measure if $P(B) > 0$, as in

- $P(A|B) \geq 0$
- $P(S|B) = 1$ if $B \subseteq S$
- $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B)$ if A_1, A_2 are mutually exclusive

Implications:

- $P(A^C|B) = 1 - P(A|B)$
- $P(A \cup D|B) = P(A|B) + P(D|B) - P(A \cap D|B)$

Multiplication rule: for two events A, B:

- $P(A \cap B) = P(A) \cdot P(B|A)$ provided $P(A) > 0$
- $P(A \cap B) = P(B) \cdot P(A|B)$ provided $P(B) > 0$

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

Ex: 25 balloons (10 yellow, 8 red, 7 green), hit one uniformly with a dart two times $A = \{\text{first is yellow}\}$ $B = \{\text{second is yellow}\}$

$P(A) = \frac{10}{25}$ $P(B|A) = \frac{9}{24}$ $P(A \cap B) = P(A) \cdot P(B|A) = \frac{10}{25} \cdot \frac{9}{24} = \frac{3}{20}$

Ex: two cups with marbles Cup 1: 3 blue, 2 white Cup 2: 1 blue, 2 white Transfer one marble from cup 1 to cup 2 uniformly at random, then draw one marble uniformly at random from cup 2

$B1 = \{\text{draw blue from cup 1}\}$ $W1 = \{\text{draw white from cup 1}\}$ $B2 = \{\text{draw blue from cup 2}\}$ $W2 = \{\text{draw white from cup 2}\}$

$P(B1 \cap B2) = P(B1) \cdot P(B2|B1) = \frac{3}{5} \cdot \frac{2}{4}$ $P(B2) \cdot P(B1|B2)$ gives the same result but also way more complicated

$P(B2) = P(B2 \cap (B1 \cup W1)) = P((B1 \cap B2) \cup (W1 \cap B2)) = P(B1 \cap B2) + P(W1 \cap B2) = \frac{3}{5} \cdot \frac{2}{4} + \frac{2}{5} \cdot \frac{1}{4} = \frac{4}{10}$

Independent events: A, B are events, we say that A, B are independent if $P(A \cap B) = P(A) \cdot P(B)$, otherwise, they are called dependent

Observation: A, B are independent if and only if $P(B|A) = P(B)$, in which case$P(B|A) = \frac{P(A \cap B)}{P(A)} = P(B)$

Ex: flip a fair coin twice, the sample space is $S = \{HH, HT, TH, TT\}$ $A = \{HH, HT\}$ (heads on 1st) $B = \{HT, TT\}$ (tail on 2nd) $C = \{TT\}$ (both tails)

$D \subseteq S, P(D) = \frac{|D|}{4}$

$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(\{HT\})}{1/2} = \frac{1}{2} = P(B)$ A, B are independent

$P(C|B) = \frac{P(B \cap C)}{P(B)} = \frac{P(\{TT\})}{1/2} = \frac{1}{2} \neq P(C)$ B, C are dependent

$P(A \cap C) = 0 \neq P(A) \cdot P(C)$ A, C are dependent

Note: if $P(A) = 0$, then A, B are independent $0 \leq P(A \cap B) \leq P(A) = 0$, so $P(A \cap B) = 0$ $P(A \cap B) = 0 \cdot P(B) = P(A) \cdot P(B)$

Theorem: if A, B are independent, then so are

- A^C, B
- A, B^C
- A^C, B^C

We prove the first line, other cases work the same Clear if $P(B) = 0$, since it does not matter what A is For $P(B) > 0$, $P(A^C \cap B) = P(A^C|B) \cdot P(B) = (1 - P(A|B)) \cdot P(B) = (1 - P(A)) \cdot P(B) = P(A^C) \cdot P(B)$

Corollary: if $P(A) = 1$, and A, B are events, then A, B are independent A^C, B are independent as $P(A^C) = 1 - P(A) = 0$, then apply above theorem $A = (A^C)^C, B$ are independent

Define:

- $A_1, \ldots, A_k$ are called pairwise independent if $P(A_i \cap A_j) = P(A_i) \cdot P(A_j)$ for all $i \neq j$
- $A_1, \ldots, A_k$ are called mutually independent if $P(\cap A_i) = \prod P(A_i)$ for all combinations of A_i

Law of total probability: let $B_1, \ldots, B_m$ be mutually exclusive and exhaustive, with $P(B_i) > 0$ for all i, then $P(A) = \sum P(B_i) \cdot P(A|B_i)$

Theorem: let $B_1, \ldots, B_m$ be a partition of S and $P(B_i) > 0$ for all i, if $P(A) > 0$, then $P(B_k|A) = \frac{P(B_k \cap A)}{P(A)} = \frac{P(B_k) \cdot P(A|B_k)}{\sum P(B_i) \cdot P(A|B_i)}$

4 Random variables

Define: given a random experiment with sample space S, a function $X : S \to \mathbb{R}$ is called a random variable The space of X is the set $X(S) = \{x \in \mathbb{R} : X(s) = x \text{ for some } s \in S\}$

Ex: $S = \{1, 2, \ldots, 6\}$ $X(S) = S$ but also $X(S) = S^2$

Ex: $S = \{H, T\}$ $X(H) = 1$ and $X(T) = 0$

Let S be a discrete sample space (finite or countable), X is called a random variable of the discrete type (or a discrete random variable), and X is said to have a distribution of the discrete type

Note: S being finite means $|S| < \infty$ S being countable means there exists a bijection $\phi : S \to \mathbb{N}$ Countable: $S = \mathbb{N}$, $S = \mathbb{Z}$, $S = \mathbb{Q}$ Not countable: $S = \mathbb{R}$, $S = [0, 1]$

For a discrete random variable, the probability $P(X = x) = P(\{s \in S : X(s) = x\})$ is denoted by $f(x)$ and called the probability mass function (pmf)

We define $S_X = \{x \in \mathbb{R} : f(x) > 0\}$ as the support of X

The pmf of a discrete random variable is a function $f : \mathbb{R} \to [0, 1]$ that satisfies the following properties

1. $f(x) > 0$ for $x \in S_X$
2. $f(x) = 0$ for $x \notin S_X$
3. $\sum_{x \in S_X} f(x) = 1$
4. $P(X \in A) = \sum_{x \in A} f(x)$ for all $A \subseteq S_X$
    (a) $P(X \in A) = \sum_{x \in A \cap S_X} f(x)$ for all $A \subseteq \mathbb{R}$

Ex: X corresponds to a fair 6-sided die $S_X = \{1, \ldots, 6\}$ $f(x) = \frac{1}{6}$ for $x \in \{1, \ldots, 6\}$ and 0 otherwise

$P(X \in [-5, 3]) = \sum_{x \in [-5,3] \cap S_X} f(x) = \sum_{x \in \{1,2,3\}} f(x) = \frac{3}{6}$

Define: if $f(x)$ is constant over S_X, we say that X is uniform over S_X (uniformly distributed)

Define: the (cumulative) distribution function (cdf) of a random variable X is defined by $F(x) = P(X \leq x) = P(X \in (-\infty, x]) = \sum_{a \in (-\infty, x] \cap S_X} f(a)$

Ex: $S_X = \{1, \ldots, m\}$, $f(x) = \frac{1}{m}$ for $x \in S_X$ $F(x) = P(X \leq x) = 0$ for $x < 1$, $\frac{k}{m}$ for $k \leq x < k + 1$ and $k = 1, \ldots, m$, and 1 for $x \geq m$

Ex: roll a fair 4-sided die twice $S = \{1, \ldots, 4\}^2$ $X((S_1, S_2)) = max(S_1, S_2)$ $S_X = \{1, \ldots, 4\}$

$P(X = 1) = \frac{1}{16}$ $P(X = 2) = \frac{3}{16}$ $P(X = 3) = \frac{5}{16}$ $P(X = 4) = \frac{7}{16}$

Probability histogram is only for $S_X \subseteq \mathbb{Z}$ For all $x \in S_X$, draw a rectangle of height $f(x)$ and width 1, centered at x

Note: pmf is the area of the rectangle

Let $X_1, X_2, \ldots$ be independent random variables with pmf $f(x) = \frac{4-x}{6}$ with $S_X = \{1, 2, 3\}$ We take n random samples and get the empirical average $\frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n} \sum_{i=1}^{3} i |\{k \in \{1, \ldots, n\} : X_k = i\}| = \sum_{i=1}^{3} i \frac{|\{k \in \{1, \ldots, n\} : X_k = i\}|}{n}$

$\frac{|\{k\in\{1,\dots,n\}:X_k=i\}|}{n}$ is the relative frequency, so it intuitively converts into the probability $\sum_{i=1}^{3} if(i) = 1\cdot\frac{3}{6}+2\cdot\frac{2}{6}+3\cdot\frac{1}{6}=\frac{10}{6}$

Define: let X be a discrete random variable with pmf $f(x)$, the expectation (or expected value) of X is denoted by $E(x)=\sum_{x\in S_X} x\cdot f(x)$ whenever this sum converges absolutely ($\sum_{x\in S_X}|x\cdot f(x)|<\infty$)

If $u:\mathbb{R}\to\mathbb{R}$ is a function, we define the expectation of $u(x)$ by $E(u)=\sum_{x\in S_X} u(x)\cdot f(x)$ whenever this sum converges absolutely

Theorem: whenever it exists, the following properties are true

1. If c is a constant, then $E(c)=c$
2. $E(c\cdot a(x))=c\cdot E(a(x))$
3. $E(u_1(x)+u_2(x))=E(u_1(x))+E(u_2(x))$

Ex: X is a random variable with pmf $f(x)=\frac{1}{3}$ with $S_X=\{-1,0,1\}$ $E(X)=(-1)\cdot f(-1)+0\cdot f(0)+1\cdot f(1)=0$ $E(X^2)=(-1)^2\cdot f(-1)+0^2\cdot f(0)+1^2\cdot f(1)=\frac{2}{3}$ $E(10X^2+7X)=10\cdot E(X^2)+7\cdot E(X)=\frac{20}{3}$

An experiment is a success with probability $p\in(0,1)$ and a failure with probability $1-p$, the experiment is repeated until the first success occurs, say this happens on the trial X $f(k)=P(X=k)=(1-p)^{k-1}\cdot p$ $k=1,2,3,\dots$

Note that $\sum_{k=1}^{\infty} f(k)=\sum_{k=1}^{\infty}(1-p)^{k-1}\cdot p=p\cdot\sum_{k=0}^{\infty}(1-p)^k=p\cdot\frac{1}{1-(1-p)}=1$ $E(X)=\sum_{k=1}^{\infty} k\cdot f(k)=\sum_{k=1}^{\infty} k\cdot p\cdot(1-p)^{k-1}=p\cdot\sum_{k=1}^{\infty} k\cdot(1-p)^{k-1}=p\cdot\frac{1}{(1-(1-p))^2}=\frac{1}{p}$

If X has pmf $f(k)=(1-p)^{k-1}\cdot p$ where $k\in\{1,2,3,\dots\}$, we say that X has a geometric distribution

Ex: let $b\in\mathbb{R}$ $E((X-b)^2)=E(X^2-2bX+b^2)=E(X^2)-2bE(X)+b^2=E(X^2)+b(b-2E(X))$ Over all $b\in\mathbb{R}$, this is minimized by $b=E(X)$

$E(X^k)$ is called the k-th moment of X $E((X-E(X))^2)$ is called the variance of X, often denoted by $\sigma^2=var(X)$

Notes:

- $E((X-E(X))^2)=E(X^2-2E(X)X+E(X)^2)=E(X^2)-2E(X)^2+E(X)^2=E(X^2)-E(X)^2$
- $\sigma^2=var(X)\geq 0$
- $var(aX)=E((aX)^2)-E(aX)^2=a^2E(X^2)-a^2E(X)^2=a^2var(X)$
- $var(X+b)=E((X+b-E(X+b))^2)=E((X-E(X))^2)=var(X)$
- $\sigma=\sqrt{\sigma^2}$ is called the standard deviation of X

Define: let X be a discrete random variable with pmf $f(x)$ and space S_X, if there exists $h>0$ such that $M(t)=E(e^{tx})=\sum_{x\in S_X} e^{tx}\cdot f(x)$ exists and is finite for all $t\in(-h,h)$, the function M is called the moment-generating function of X

$M(0)=\sum_{x\in S_X} 1\cdot f(x)=1$

$M'(t) = \frac{d}{dt}M(t) = \frac{d}{dt}\sum_{x\in S_X} e^{tx} \cdot f(x) = \sum_{x\in S_X} xe^{tx} \cdot f(x)$ $M'(0) = \sum_{x\in S_X} x \cdot 1 \cdot f(x) = E(X)$

$M''(t) = \frac{d}{dt}M'(t) = \frac{d}{dt}\sum_{x\in S_X} xe^{tx} \cdot f(x) = \sum_{x\in S_X} x^2e^{tx} \cdot f(x)$ $M''(0) = \sum_{x\in S_X} x^2 \cdot 1 \cdot f(x) = E(X^2)$

$M^{(k)}(0) = E(X^k)$

Ex: let X have uniform distribution on $\{1,\ldots,m\}$, X has pmf $f(x) = \frac{1}{m}$ for $x \in \{1,\ldots,m\}$ $E(X) = \sum_{x\in S_X} x\cdot\frac{1}{m} = \sum_{x=1}^{m} x\cdot\frac{1}{m} = \frac{1}{m}\cdot\sum_{x=1}^{m} x = \frac{1}{m}\cdot\frac{m(m+1)}{2} = \frac{m+1}{2}$ $E(X^2) = \sum_{x\in S_X} x^2\cdot\frac{1}{m} = \frac{1}{m}\cdot\sum_{x=1}^{m} x^2 = \frac{1}{m}\cdot\frac{m(m+1)(2m+1)}{6} = \frac{(m+1)(2m+1)}{6}$ $var(X) = E(X^2) - E(X)^2 = \frac{(m+1)(2m+1)}{6} - (\frac{m+1}{2})^2 = \frac{m^2-1}{12} = E((X-E(X))^2)$

5 Binomial distribution

A Bernoulli experiment is a random experiment with two possible outcomes

A sequence of Bernoulli trials occurs when a Bernoulli experiment is performed several independent times and the probability of success p remains the same for each trial

Let X be the random variable defined by $X(\text{success}) = 1$ and $X(\text{failure}) = 0$, we say that X has Bernoulli distribution

In a sequence of Bernoulli trials, we write X_i for the random variable associated with the i-th trial, $(X_1,\ldots,X_n)$ is called a random sample of size n from a Bernoulli distribution

We are interested in the number of successes of a Bernoulli trial of size n, as in $X = \sum_{i=1}^{n} X_i$

X is said to have a binomial distribution with parameters n and p

- Bernoulli performed n times $\to$ independent trials
- Probability of success is p (constant) $\to$ X is the number of successes

Let $k \in \{0,\ldots,n\}$, then $X = k$ if and only if there exists $A \subseteq \{1,\ldots,n\}$ such that $|A| = k, X_i = 1$ for all $i \in A$, and $X_i = 0$ for all $i \in \{1,\ldots,n\}\backslash A$

$P(X = k) = \binom{n}{k} p^k(1-p)^{n-k}$

Note: $\sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k} = (p+1-p)^n = 1$ and $E(X) = \sum_{i=1}^{n} E(X_i) = np$

Note: if X has a binomial distribution with parameters n and p, then $n - X$ has a binomial distribution with parameters n and $1-p$

$M(t) = E(e^{tx}) = \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k}e^{tk} = \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k(1-p)^{n-k} = (pe^t + 1 - p)^n$ $M'(t) = n(pe^t + 1 - p)^{n-1}pe^t$ $M'(0) = np = E(X)$

Let X be the number of trials needed until we see the r-th success, we say that X has a negative binomial distribution

$X = k$ means $r-1$ successes after $k-1$ trials and success on the k-th trial $P(X = k) = \binom{k-1}{r-1} p^{r-1}(1-p)^{k-1-(r-1)}p = \binom{k-1}{r-1} p^r(1-p)^{k-r}$

If $r = 1$, this is a geometric distribution, as $P(X = k) = (1 - p)^{k-1}p$ For $r = 1$, we computed $E(X) = \frac{1}{p}$ For $r \geq 1$, let y_r be the number of repeats until the r-th success, so $X = y_1 + \sum_{k=2}^{r}(y_k - y_{k-1})$ and $E(X) = \frac{r}{p}$

6 Poisson distribution

Let the number of occurrences of some event in a given continuous interval be counted, then we have an approximate Poisson process with parameter $\lambda > 0$ if the following conditions are satisfied:

1. The number of occurrences in non-overlapping sub-intervals are independent
2. The probability of exactly one occurrence in a sufficiently short sub-interval of length n is approximately λn
3. The probability of two or more occurrences in a sufficiently short sub-interval is essentially 0

Ex: people arriving at the post office between 10AM and 11AM Say we have an interval of length L, the number of occurrences in this interval has a Poisson distribution with parameters λ and L

n is the number of sub-intervals λ is the expected number of occurrences over interval L X_n is the number of sub-intervals in which an occurrence happens X is the total number of occurrences X_n is binomially distributed with parameters n and $\lambda\frac{1}{n}$ As $n \to \infty$, $P(X_n = k) \to P(X = k) = e^{-n\lambda\frac{1}{n}}\frac{(n\lambda\frac{1}{n})^k}{k!} = e^{-\lambda}\frac{\lambda^k}{k!}$

7 Bivariate distribution

Define: let X and Y be two random variables, let $S = S_X \times S_Y$, the function $f : \mathbb{R}^2 \to [0, 1]$ defined by $f(x, y) = P(X = x, Y = y)$ is called the joint pmf of X and Y, it satisfies

1. $f(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$
2. $\sum_{(x,y)\in S} f(x, y) = 1$
3. $P((X, Y) \in A) = \sum_{(x,y)\in A} f(x, y)$
4. $f(x, y) = 0$ for all $(x, y) \in \mathbb{R}^2 \backslash S$

Define: let X and Y have joint pmf f, the marginal probability mass functions f_X (resp. f_Y) of X (resp. of Y) are defined by $f_X(x) = P(X = x) = \sum_{y\in S_Y} f(x, y)$ and $f_Y(y) = P(Y = y) = \sum_{x\in S_X} f(x, y)$

X and Y are called independent if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all $(x, y) \in S_X \times S_Y$ and dependent otherwise This is equivalent to $f(x, y) = f_X(x)f_Y(y)$ for all $(x, y) \in \mathbb{R}^2$

Define: let X and Y be two discrete random variables with joint pmf f and $u : S_X \times S_Y \to \mathbb{R}$, then the expectation $E(u(X,Y))$ is defined by $E(u(X,Y)) = \sum_{(x,y)\in S} u(x,y)f(x,y)$ whenever the sum converges absolutely

Remarks: if $u(x,y) = x$, $E(u(X,Y)) = \sum_{(x,y)\in S} xf(x,y) = \sum_{x\in S_X} xf_X(x) = E(X)$ If $u(x,y) = (x - E(X))^2$, $E(u(X,Y)) = var(X)$ If $u(x,y) = x + y$, $E(u(X,Y)) = E(X+Y) = E(X) + E(Y)$

Theorem: if X and Y are independent, then $E(XY) = E(X)E(Y)$ $E(XY) = \sum_{x\in S_X}\sum_{y\in S_Y} xyf(x,y) = \sum_{x\in S_X}\sum_{y\in S_Y} xyf_X(x)f_Y(y) = \sum_{x\in S_X} xf_X(x)\sum_{y\in S_Y} yf_Y(y) = E(X)E(Y)$

Remark: X and Y are independent implies $E(XY) = E(X)E(Y)$, but $E(XY) = E(X)E(Y)$ does not imply X and Y are independent

Define: let X and Y be two discrete random variables with $\mu_X = E(X)$ and $\mu_Y = E(Y)$, the covariance of X and Y is defined by $\sigma_{X,Y} = Cov(X,Y) = E((X-\mu_X)(Y-\mu_Y))$

The correlation coefficient of X and Y is defined by $\rho = \frac{\sigma_{X,Y}}{\sigma_X\sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$

Note: one can show that $p \in [-1,1]$

$Cov(X,Y) = E(XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y) = E(XY) - E(X)E(Y)$ X and Y are independent implies $\sigma_{X,Y} = 0$

Best approximation: $Y = aX + b$ $a = \rho\frac{\sigma_Y}{\sigma_X}$ $b = E(Y) - aE(X)$

8 Conditional distribution

$P(Y = y \mid X = x) = \frac{P(X=x,Y=y)}{P(X=x)}$

Define: let X and Y be two random variables with joint pmf f, the conditional probability mass function of Y given X is given by $g(y \mid x) = \frac{f(x,y)}{f_X(x)}$

Define: the conditional expectation of Y given X is defined by $E(Y \mid X = x) = \sum_{y\in S_Y} y \cdot P(Y = y \mid X = x)$ $E(u(Y) \mid X = x) = \sum_{y\in S_Y} u(y)\cdot P(Y = y \mid X = x)$

Law of total expectation: assume that $E(Y)$ exists, then $E(Y) = E(E(Y \mid X)) = \sum_{x\in S_X} E(Y \mid X = x)P(X = x)$

Proof: $\sum_{x\in S_X} E(Y \mid X = x)P(X = x) = \sum_{x\in S_X}\sum_{y\in S_Y} yP(Y = y \mid X = x)P(X = x) = \sum_{x\in S_X}\sum_{y\in S_Y} yf(x,y) = E(Y)$

9 Continuous random variables

The distribution of the random variable is defined through its cumulative distribution function (cdf) F, defined as $F(x) = P(X \le x)$ and its probability density function (pdf) defined by $f(x) = F'(x)$

Note:

- $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$
- $F(x)$ is weakly increasing, so $f(x) = F'(x) \ge 0$
- $F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$

- $P(a < X \leq b) = F(b) - F(a) = \int_a^b f(t)dt$
- $P(X \in A) = \int_A f(t)dt$

Define: if X is a random variable and there exists f such that $P(X \in A) = \int_A f(t)dt$, we say that X is of the continuous type, or a continuous random variable

Interpretation: $P(X \in [x - \frac{\epsilon}{2}, x + \frac{\epsilon}{2}]) = \int_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} f(t)dt \approx \epsilon f(x)$

Define: let X be a continuous random variable with cdf F and pdf f

- The expectation of X is defined by $E(X) = \int_{-\infty}^{\infty} xf(x)dx$ if the integral exists
- For a function $u : \mathbb{R} \to \mathbb{R}$, the expectation of $u(X)$ is defined by $E(u(X)) = \int_{-\infty}^{\infty} u(x)f(x)dx$ if the integral exists
- We define the variance of X to be $\sigma^2 = Var(X) = E((X - E(X))^2) = \int_{-\infty}^{\infty}(x - E(X))^2 f(x)dx$
- For $p \in (0, 1)$, we define the p-th percentile as the number π_p such that $F(\pi_p) = \int_{-\infty}^{\pi_p} f(t)dt = p$
  - $\pi_{0.25}$ is the first quartile
  - $\pi_{0.5}$ is the second quartile/median
  - $\pi_{0.75}$ is the third quartile

Ex: let X have uniform distribution on $[0, m]$ $E(X) = \int_0^m \frac{1}{m}xdx = \frac{1}{m}(\frac{x^2}{2})|_0^m = \frac{m}{2}$ $E(X) = \int_0^m \frac{1}{m}x^2dx = \frac{1}{m}(\frac{x^3}{3})|_0^m = \frac{m^2}{3}$ $Var(X) = \frac{m^2}{3} - \frac{m^2}{4} = \frac{m^2}{12}$ $\pi_{0.25}$ is such that $F(\pi_{0.25}) = \frac{1}{m}\pi_{0.25} = \frac{1}{4}$, so $\pi_{0.25} = \frac{m}{4}$

Ex: $f(x) = 2x$ for $x \in (0, 1)$, $f(x) = 0$ otherwise $F(x) = x^2$ for $x \in (0, 1)$, 1 for $x \geq 1$, and 0 for $x \leq 0$ $P(X \in (0.5, 2]) = F(2) - F(0.5) = 1 - \frac{1}{4} = \frac{3}{4}$

10 Exponential distribution

Consider an approximate Poisson process with rate λ, let W be the time until the first arrival, $W \in \mathbb{R} \geq 0$ $P(W > t) = P(\text{no arrival in } (0, t]) = e^{-\lambda t}\frac{(\lambda t)^0}{0!} = e^{-\lambda t}$ W is said to have an exponential distribution with parameter $\lambda > 0$ For $t > 0$, $F(t) = 1 - P(W > t) = 1 - e^{-\lambda t}$ and $f(t) = \lambda e^{-\lambda t}$ For $t \leq 0$, $F(t) = f(t) = 0$ $E(W) = \int_0^{\infty} x\lambda e^{-\lambda x}dx = (-\frac{e^{-\lambda x}(\lambda x+1)}{\lambda})|_0^{\infty} = \frac{1}{\lambda}$

11 Normal distribution

We say that a continuous random variable has a normal (or Gaussian) distribution if its pdf is of the form $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$ for $x \in \mathbb{R}$ and some $\mu \in \mathbb{R}$, $\sigma > 0$

$f(x) \geq 0$ $\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = I$ (can use variable change to get rid of μ and σ) $I^2 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(-\frac{x^2+y^2}{2})dxdy = \frac{1}{2\pi} \int_0^{\infty} \int_0^{2\pi} e^{-\frac{r^2}{2}} r d\theta dr = \int_0^{\infty} e^{-\frac{r^2}{2}} r dr = (-e^{-\frac{r^2}{2}})|_0^{\infty} = 1$

One can show that $E(X) = \mu$ and $var(X) = \sigma^2$

In the case that $\mu = 0$ and $\sigma^2 = 1$, we say that X has a standard normal distribution We write $N(\mu, \sigma^2)$ for this distribution

For a standard normal distribution, we have (by symmetry) $-\pi_\alpha = \pi_{1-\alpha}$ for all $\alpha \in (0, 1)$

If X has distribution $N(\mu_1, \sigma_1^2)$ and Y has distribution $N(\mu_2, \sigma_2^2)$, and X and Y are independent, then $X + Y$ has distribution $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

If X has distribution $N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$ has distribution $N(0, 1)$ Proof: $P(Z \leq z) = P(\frac{X-\mu}{\sigma} \leq z) = P(X \leq z\sigma + \mu) = \int_{-\infty}^{z\sigma+\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{z\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$

Ex: if $X = N(3, 16)$, then $P(4 \leq X \leq 8) = P(\frac{4-3}{4} \leq \frac{X-3}{4} \leq \frac{8-3}{4}) = P(\frac{1}{4} \leq Z \leq \frac{5}{4}) = \Phi(1.25) - \Phi(0.25) \approx 0.296$

Remarks: values of $P(Z \leq z) = \Phi(z)$ are not easy to compute and approximations are often given in old textbooks

If Z has a standard normal distribution, then Z^2 is chi-squared distribution with 1 degree of freedom

If X and Y are independent and are normally distributed with expectations μ_X and μ_Y and variance σ_X^2 and σ_Y^2, then $\alpha X + \beta Y$ is normally distributed with expectation $\alpha\mu_X + \beta\mu_Y$ and variance $\alpha^2\sigma_X^2 + \beta^2\sigma_Y^2$

12 Continuous bivariate distributions

Define: let X and Y be two random variables of the continuous type, we say that X and Y have joint pdf f, if $P(X \leq a, Y \leq b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x, y)dxdy$ for all $a, b \in \mathbb{R}$

Example: $f(x, y) = \frac{1}{2}$ for $0 \leq x \leq 2, 0 \leq y \leq 1$, and 0 otherwise $P(X \leq a, Y \leq b) = \int_0^b \int_0^a \frac{1}{2} dxdy = \frac{ab}{2}$

If X and Y have joint pdf f:

1. The set $S = \{(x, y) \in \mathbb{R}^2 : f(x, y) > 0\}$ is called the support of X and Y
2. $f(x, y) \geq 0$ for all $x, y \in \mathbb{R}$
3. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1$
4. For $A \subseteq \mathbb{R}^2$, $P((X, Y) \in A) = \iint_A f(x, y)dxdy$

Ex: X and Y have joint pdf $f(x, y) = \frac{4}{3}(1 - xy)$ for $x, y \in [0, 1]$ and 0 otherwise $P(X \leq \frac{1}{2}, Y \leq \frac{1}{4}) = P((X, Y) \in A)$ where $A = \{(x, y) \in \mathbb{R}^2 : x \leq \frac{1}{2}, y \leq \frac{1}{4}\} = \iint_A f(x, y)dxdy = \int_0^{\frac{1}{4}} \int_0^{\frac{1}{2}} \frac{4}{3} - \frac{4}{3}xydxdy = \frac{4}{3}\frac{1}{4}\frac{1}{2} - \frac{4}{3} \int_0^{\frac{1}{4}} \int_0^{\frac{1}{2}} xydxdy = \frac{11}{64}$

$P(Y \leq \frac{X}{2}) = \int_0^1 \int_0^{\frac{x}{2}} \frac{4}{3}(1-xy)dydx = \int_0^1 \frac{4}{3}(y - \frac{xy^2}{2})|_0^{\frac{x}{2}} dx = \frac{4}{3}\int_0^1 \frac{x}{2} - \frac{x^3}{8} dx = \frac{4}{3}(\frac{x^2}{4} - \frac{x^4}{32})|_0^1 = \frac{7}{24}$

Define: let X and Y be two continuous random variables with pdf f, the marginal density function of X and Y is given by $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$ and $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$

f_X and f_Y are the density functions of X and Y For $A \subseteq \mathbb{R}$, $P(X \in A) = P((X,Y) \in A \times \mathbb{R}) = \int_A \int_{\mathbb{R}} f(x,y)dydx = \int_A f_X(x)dx$

Ex: $f(x,y) = \frac{4}{3}(1-xy)$ for $x, y \in [0,1]$ and 0 otherwise $f_X(x) = \int_0^1 \frac{4}{3} - \frac{4}{3}xydy = \frac{4}{3} - \frac{2}{3}x$

Define: for a function $u : \mathbb{R}^2 \to \mathbb{R}$, the expectation $E(u(x,y))$ is defined by $\int_{\mathbb{R}} \int_{\mathbb{R}} u(x,y)f(x,y)dxdy$

Ex: $E(Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} yf(x,y)dxdy = \int_{\mathbb{R}} y \int_{\mathbb{R}} f(x,y)dxdy = \int_{\mathbb{R}} yf_Y(y)dy$

Ex: $E(X) = \int_0^1 \int_0^1 x\frac{4}{3}(1-xy)dydx = \int_0^1 \int_0^1 \frac{4}{3}xdydx - \frac{4}{3}\int_0^1 x^2 \int_0^1 ydydx = \frac{4}{3}\int_0^1 xdx - \frac{4}{3}\int_0^1 \frac{x^2}{2}dx = \frac{4}{6} - \frac{4}{6}(\frac{1}{3}x^3)|_0^1 = \frac{4}{9}$ $E(X^2) = \int_0^1 \int_0^1 x^2 \frac{4}{3}(1-xy)dydx = \int_0^1 \frac{4}{3}x^2dx - \int_0^1 \frac{4}{3}x^3 \int_0^1 ydydx = \frac{5}{18}$

Theorem: if X and Y are random variables with joint pdf f, then X and Y are independent if and only if $f(x,y) = f_X(x)f_Y(y)$

Ex: $f(x,y) = \frac{4}{3}(1-xy)$ for $x, y \in [0,1]$ $f_X(x) = \frac{4}{3}(1 - \frac{x}{2})$ for $x \in [0,1]$ $f_Y(y) = \frac{4}{3}(1 - \frac{y}{2})$ for $y \in [0,1]$

Consider $x = y = \frac{1}{4}$ $f(x,y) = \frac{5}{4}$ $f_X(x)f_Y(y) = \frac{49}{36} \neq f(x,y)$

13 Functions of a random variable

Theorem: if X is a discrete random variable with pmf f, and $u : \mathbb{R} \to \mathbb{R}$ is a function, then $Y = u(X)$ has pmf $g(y) = P(Y = y) = P(u(X) = y) = P(X \in u^{-1}(y)) = \sum_{x \in u^{-1}(y)} f(x)$ where $u^{-1}(y) = \{x \in \mathbb{R} : u(x) = y\}$

Ex: $S_X = \{-2, -1, \ldots, 5\}$, $f(x) = \frac{1}{8}$ for all $x \in S_X$, $u(x) = x^2$ $g(y) = P(u(X) = y) = P(X^2 = y) = \frac{2}{8}$ for $y = 1, 4$ and $\frac{1}{8}$ for $y = 0, 9, 16, 25$

Let X be a continuous random variable with state space S_X, let $u : S_X \to S_Y$ be a function, the inverse function of u is the function $v : S_Y \to S_X$ such that $u(v(y)) = y$ for all $y \in S_Y$ and $v(u(x)) = x$ for all $x \in S_X$

Ex: $S_X = S_Y = \mathbb{R} \geq 0$, $u(x) = x^2$, $v(y) = \sqrt{y}$ $S_X = \mathbb{R}$, $S_Y = \mathbb{R} \geq 0$, $u(x) = x^2$, the inverse does not exist as $v(u(1)) = v(1)$ needs to be 1 and $v(u(-1)) = v(1)$ needs to be -1

Ex: let X be a continuous random variable with $S_X = \mathbb{R}$ and pdf f, and $u : \mathbb{R} \to \mathbb{R}$ be monotone, increasing, and invertible, let v be the inverse of u $P(u(X) \leq t) = P(v(u(X)) \leq v(t)) = P(X \leq v(t)) = \int_{-\infty}^{v(t)} f(x)dx$ The pdf of $Y = u(X)$ is $\frac{d}{dt}P(u(X) \leq t) = \frac{d}{dt}\int_{-\infty}^{v(t)} f(x)dx = \frac{d}{dt}\int_{-\infty}^{t} f(v(t))v'(t)dt = f(v(t))v'(t)$

Theorem: if X is a continuous random variable with state space S_X and $u : S_X \to S_Y$ is a continuous function with inverse v, the $Y = u(X)$ has pdf $g(y) = f(v(y))|v'(y)|$

Ex: $S_X = \mathbb{R}_{\geq 0}$, $f(x) = e^{-x}$, $u : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 1}$, $u(x) = e^x$, $v(y) = ln(y)$, and if X has pdf f, then $Y = u(X)$ has pdf $g(y) = f(v(y))|v'(y)| = e^{-ln(y)}\frac{1}{y} = \frac{1}{y^2}$

Several independent random variables

Setup: $X_1, \ldots, X_n$ independent random variables, $u_1, \ldots, u_n$ functions from $\mathbb{R} \to \mathbb{R}$, $X_1, \ldots, X_n$ is called a random sample of size n, Y is a combination of $u_1(X_1), \ldots, u_n(X_n)$, for example $Y = \sum_{i=1}^n u_i(X_i)$ or $Y = \prod_{i=1}^n u_i(X_i)$, if X_i has pdf $f_i(x_i)$, then $(X_1, \ldots, X_n)$ has pdf $f : \mathbb{R}^n \to \mathbb{R}$ $f(x) = f_1(x_1) \cdot \cdots \cdot f_n(x_n)$

Theorem: if $X_1, \ldots, X_n$ are independent, then when all expectations exist, $E(u_1(X_1) \cdot \cdots \cdot u_n(X_n)) = E(u_1(X_1)) \cdot \cdots \cdot E(u_n(X_n))$

If $X_1, \ldots, X_n$ are random variables (independent or not), then $E(u_1(X_1) + \cdots + u_n(X_n) = E(u_1(X_1)) + \cdots + E(u_n(X_n))$

Proof: discrete case only $Y = (X_1, \ldots, X_n)$ has pmf $f(x_1, \ldots, x_n) = P(Y = (x_1, \ldots, x_n)) = \prod_{i=1}^n P(X_i = x_i) = \prod_{i=1}^n f_i(x_i)$ $E(u_1(X_1) \cdot \cdots \cdot u_n(X_n)) = \sum_{x_1} \cdots \sum_{x_n} u_1(x_1) \cdot \cdots \cdot u_n(x_n) f(x_1, \ldots, x_n) = \sum_{x_1} u_1(x_1) f_1(x_1) \cdot \cdots \cdot \sum_{x_n} u_n(x_n) f_n(x_n)$ $= E(u_1(X_1)) \cdot \cdots \cdot E(u_n(X_n))$

$E(u_1(X_1) + \cdots + u_n(X_n)) = \sum_{x_1} \cdots \sum_{x_n} (u_1(x_1) + \cdots + u_n(x_n)) f(x_1, \ldots, x_n)$ $= \sum_{x_1} \cdots \sum_{x_n} u_1(x_1) f(x_1, \ldots, x_n) + \cdots + \sum_{x_1} \cdots \sum_{x_n} u_n(x_n) f(x_1, \ldots, x_n) = E(u_1(X_1)) + \cdots + E(u_n(X_n))$

If $X_1, \ldots, X_n$ are independent and $E(X_i) = \mu_i$, $Var(X_i) = \sigma_i^2$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, then $E(\sum_{i=1}^n \alpha_i X_i) = \sum_{i=1}^n \alpha_i \mu_i$ and $Var(\sum_{i=1}^n \alpha_i X_i) = \sum_{i=1}^n \alpha_i^2 \sigma_i^2$

Proof: Let $Y = \sum_{i=1}^n \alpha_i X_i$ and $\mu_Y = E(Y)$ $Var(Y) = E((Y - \mu_Y)^2) = E((\sum_{i=1}^n \alpha_i X_i - \sum_{i=1}^n \alpha_i \mu_i)^2) = E((\sum_{i=1}^n \alpha_i (X_i - \mu_i))^2) = \sum_{i=1}^n \sum_{j=1}^n E(\alpha_i (X_i - \mu_i) \alpha_j (X_j - \mu_j)) = \sum_{i=1}^n E(\alpha_i^2 (X_i - \mu_i)^2) + 2 \sum_{1 \leq i < j \leq n} E(a_i (X_i - \mu_i) a_j (X_j - \mu_j))$ Since $X_1, \ldots, X_n$ are independent, $E(u_1(X_1) \cdot \cdots \cdot u_n(X_n)) = E(u_1(X_1)) \cdot \cdots \cdot E(u_n(X_n)) = \sum_{i=1}^n \alpha_i^2 E((X_i - \mu_i)^2) + 2 \sum_{1 \leq i < j \leq n} \alpha_i \alpha_j E(X_i - \mu_i) E(X_j - \mu_j)$ $E((X_i - \mu_i)^2) = \sigma_i^2$, $E(X_i - \mu_i) = 0 = \sum_{i=1}^n \alpha_i^2 \sigma_i^2$

Markov's inequality: let X be a non-negative random variable and let $a > 0$, then $P(X \geq a) \leq \frac{E(X)}{a}$ Proof: define Y by $Y = a$ for $x \geq a$ and 0 otherwise, so $Y \leq X$ $E(X) = E(X - Y + Y) = E(X - Y) + E(Y) \geq E(Y)$, so $E(Y) \leq E(X)$ $E(Y) = a \cdot P(X \geq a)$ therefore $P(X \geq a) \leq \frac{E(X)}{a}$

Chebyshev's inequality: let X be a random variable with $E(X) = \mu$, then for $t \geq 0$, $P(|X - \mu| \geq t) \leq \frac{Var(X)}{t^2}$ Proof: $P(|X - \mu| \geq t) = P((X - \mu)^2 \geq t^2) \leq \frac{E((X-\mu)^2)}{t^2} = \frac{Var(X)}{t^2}$

Law of large numbers (weak): let $X_1, \ldots, X_n$ be independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for all i, define $S_n = \frac{1}{n} \sum_{i=1}^n X_i$, then for every $\epsilon > 0$, $P(|S_n - \mu| \geq \epsilon) \to 0$ as $n \to \infty$ In particular, this holds if $X_1, \ldots, X_n$ are independent and identically distributed, we say that S_n converges to μ in probability Proof: $P(|S_n - \mu| \geq \epsilon) = P(|X_1 + \cdots + X_n - n\mu| \geq n\epsilon) \leq \frac{Var(X_1 + \cdots + X_n)}{n^2 \epsilon^2} = \frac{n\sigma^2}{n^2\epsilon^2} - \frac{\sigma^2}{n\epsilon^2} \to 0$ as $n \to \infty$

Note: for $X_1, \ldots, X_n$ independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, $P(|S_n - \mu| \geq \frac{c}{\sqrt{n}}) = P(|X_1 + \cdots + X_n - n\mu| \geq c\sqrt{n}) \leq \frac{Var(X_1 + \cdots + X_n)}{c^2 n} = \frac{n\sigma^2}{c^2 n} = \frac{\sigma^2}{c^2}$