Probability Theory

Def'd in terms of a **probability space** or **sample space** S (or Ω), a **set** whose elements $s \in S$ (or $\omega \in \Omega$) are called **elementary events**.

View elementary events as possible outcomes of an experiment.

Examples:

- flip a coin: $S = \{\text{head}, \text{tail}\}$
- roll a die: $S = \{1, 2, 3, 4, 5, 6\}$
- pick a random pivot in $A[p \ldots, r]$:
  $S = \{p, p+1, \ldots, r\}$

We're talking only about **discrete** prob. spaces (unlike $S = [0, 1]$), usually **finite**

An **event** is a subset of the prob. space

Examples:

- roll a die; $A = \{2, 4, 6\} \subset \{1, 2, 3, 4, 5, 6\}$ is the event of having an even outcome
- flip two distinguishable coins:
  $S = \{HH, HT, TH, TT\}$, and $A = \{TT, HH\} \subset S$ is the event of having the same outcome with both coins

We say S (the entire sample space) is a **certain event**, and $\emptyset$ (the empty event) is a **null event**

We say events A and B are **mutually exclusive** if $A \cap B = \emptyset$

Axioms

A **probability distribution** $P()$ on S is mapping from events of S to reals s.t.

1. $P(A) \geq 0$ for all $A \subseteq S$
2. $P(S) = 1$ (normalisation)
3. $P(A) + P(B) = P(A \cup B)$ for any two **mutually exclusive** events A and B, i.e., with $A \cap B = \emptyset$.

   Generalisation: for any finite sequence of pairwise mutually exclusive events $A_1, A_2, \ldots$

   $$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

$P(A)$ is called **probability** of event A

A bunch of stuff that follows:

1. $P(\emptyset) = 0$

2. If $A \subseteq B$ then $P(A) \leq P(B)$

3. With $\bar{A} = S - A$, we have $P(\bar{A}) = P(S) - P(A) = 1 - P(A)$

4. For any A and B (**not** necessarily mutually exclusive),

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

Considering discrete sample spaces, we have for any event A

$$P(A) = \sum_{s \in A} P(s)$$

If S is finite, and $P(s \in S) = 1/|S|$, then we have **uniform probability distribution** on S (that's what's usually referred to as "picking an element of S at random")

Conditional probabilities

When you already have partial knowledge

Example: a friend rolls two fair dice (prob. space is $\{(x, y) : \ x, y \in \{1, \ldots, 6\}\}$) tells you that one of them shows a 6. What's the probability for a $6 - 6$ outcome?

Information eliminates outcomes without any 6, i.e., all combinations of 1 through 5. There are $5^2 = 25$ of them. The original prob. space has size $6^2 = 36$, thus we're left with $36 - 25 = 11$ events where at least one 6 is involved.

These are equally likely, thus the sought probability must be $1/11$.

The **conditional probability** of event A given that another event B occurs is

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

given $P(B) \neq 0$

In example:

$$
\begin{aligned}
A &= \{(6,6)\} \\
B &= \{(6,x) : x \in \{1,\ldots,6\}\} \ \cup \\
&\quad\ \{(x,6) : x \in \{1,\ldots,6\}\}
\end{aligned}
$$

with $|B| = 11$ (the $(6,6)$ is in both parts) and thus $P(A \cap B) = P(\{(6,6)\}) = 1/36$ and

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/36}{11/36} = \frac{1}{11}$$

Independence

We say two events are **independent** if

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

equivalent to (if $P(B) \neq 0$) to

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

Events $A_1, A_2, \ldots, A_n$ are **pairwise independent** if

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j)$$

for all $1 \leq i < j \leq n$.

They are **(mutually) independent** if every k-subset $A_{i_1}, \ldots, A_{i_k}$, $2 \leq k \leq n$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ satisfies

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$

Random variables

Reminder: we're talking **discrete** probability spaces
(makes things easier)

A **random variable** (r.v.) X is a function from a probability space S to the reals, i.e., it assigns some value to elementary events

Event "$X = x$" is def'd to be $\{s \in S : \; X(s) = x\}$

Example: roll three dice

- $S = \{s = (s_1, s_2, s_3) \mid s_1, s_2, s_3 \in \{1, 2, \ldots, 6\}\}$
  $|S| = 6^3 = 216$ possible outcomes
- Uniform distribution: each element has prob $1/|S| = 1/216$
- Let r.v. X be sum of dice, i.e.,
  $X(s) = X(s_1, s_2, s_3) = s_1 + s_2 + s_3$

$P(X = 7) = 15/216$ because

115	214	313	412	511
124	223	322	421	
133	232	331		
142	241			
151				

Important: With r.v. X, writing $P(X)$ does **not** make any sense; $P(X = \text{something})$ **does**, though (because it's an **event**)

Clearly, $P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$ (from probability axioms)

If X and Y are r.v. then $P(X = x \text{ and } Y = y)$ is called **joint prob. distribution** of X and Y.

$$
\begin{aligned}
P(Y = y) &= \sum_x P(X = x \text{ and } Y = y) \\
P(X = x) &= \sum_y P(X = x \text{ and } Y = y)
\end{aligned}
$$

R.v. X, Y are **independent** if $\forall x, y$, events "$X = x$" and "$Y = y$" are independent

Recall: A and B are independent iff $P(A \cap B) = P(A) \cdot P(B)$.

Now: X, Y are independent iff $\forall x, y$,

$$P(X = x \text{ and } Y = y) = P(X = x) \cdot P(Y = y)$$

Intuition:

$$\begin{aligned} A &= \text{``}X = x'' = \text{``}X = x \text{ and } Y =?'' \\ B &= \text{``}Y = y'' = \text{``}X =? \text{ and } Y = y'' \end{aligned}$$

$$A \cap B = \text{``}X = x \text{ and } Y = y''$$

Welcome to. . . **expected values** of r.v.

Also called **expectations** or **means**

Given r.v. X, its expected value is

$$E[X] = \sum_x x \cdot P(X)$$

Well-defined if sum is finite or converges absolutely

Sometimes written μ_X (or μ if context is clear)

Example: roll a fair six-sided die, let X denote expected outcome

$$\begin{aligned} E[X] &= 1 \cdot 1/6 + 2 \cdot 1/6 + 4 \cdot 1/6 + \\ &\quad 5 \cdot 1/6 + 6 \cdot 1/6 \\ &= 1/6 \cdot (1 + 2 + 3 + 4 + 5 + 6) \\ &= 1/6 \cdot 21 \\ &= 3.5 \end{aligned}$$

Another example: flip three fair coins

For each head you win \$4, for each tail you lose \$3

Let r.v. X denote your win. Then the probability space is

$$\{\text{HHH,HHT,HTH,THH,HTT,THT,TTH,TTT}\}$$

and

$$\begin{aligned} E[X] &= 12 \cdot P(3\text{H}) + 5 \cdot P(2\text{H}) - \\ &\quad -2 \cdot P(1\text{H}) - 9 \cdot P(0\text{H}) \\ &= 12 \cdot 1/8 + 5 \cdot 3/8 - 2 \cdot 3/8 - 9 \cdot 1/8 \\ &= \frac{12 + 15 - 6 - 9}{8} = \frac{12}{8} = 1.5 \end{aligned}$$

which is intuitively clear: each single coin contributes an expected win of 0.5

Important: Linearity of expectations

$$E[X+Y] = E[X] + E[Y]$$

whenever $E[X]$ and $E[Y]$ are defined

True even if X and Y are **not independent**

Some more properties

Given r.v. X and Y with expectations, constant a

- $E[aX] = aE[X]$

  (note: aX is a r.v.)

- $E[aX + Y] = E[aX] + E[Y] = aE[X] + E[Y]$

- if X, Y **independent**, then

$$\begin{aligned} E[XY] &= \sum_x \sum_y xyP(X = x \text{ and } Y = y) \\ &= \sum_x \sum_y xyP(X = x)P(Y = y) \\ &= \left(\sum_x xP(X = x) \right) \left(\sum_y yP(Y = y) \right) \\ &= E[X]E[Y] \end{aligned}$$

Variance

The expected value of a random variable does not tell how "spread out" the variables are.

Example: Two variables X and Y.

P(X=1/4)=P(X=3/4)=1/2
P(Y=0)=P(Y=1)=1/2

Both random variables have the same expected value!

The variance measures the expected difference between the expected value of the variable and an outcome.

$$\begin{aligned} V[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + E^2[X]] \\ &= E[X^2] - E^2[X] \end{aligned}$$

$V[\alpha X] = \alpha^2 V[X]$ and
$V[X + Y] = V[X] + V[Y]$

Standard deviation $\sigma(X) = \sqrt{V[X]}$

Tail Inequalities

Measures the deviation of a random variable from its expected value.

1. **Markov inequality**

   Let Y be a non-negative random variable.Then for all $t > 0$

   $$P[Y \geq t] \leq E[Y]/t \text{ and } P[Y \geq kE[Y]] \leq 1/k.$$

   Proof:Define a function $f(y)$ by $f(y) = 1$ if $y \geq t$ and 0 otherwise.

   Note: $E[f(X)] = \sum_x f(x) \cdot P[X = x]$.

   Hence, $P[Y \geq t] = E[Y]$. Since $f(y) \leq y/t$ for all y we get

   $$E[f(Y)] \leq E[Y/t] = E[Y]/t$$

   This is the best possible bound bound if we only know that Y is non-negative.

   But the Markov inequality is quite weak!

   Example: throw n balls into n bins.

Tail Inequalities

1. **Chebyshev's Inequality**

   Let X be a random variable with expectation μ_X and standard deviation σ_X. Then for any $t > 0$,

   $$P[|X - \mu_X| \geq t\sigma_X] \leq 1/t^2.$$

   Proof: First, note that

   $$P[|X - \mu_X| \geq t\sigma_X] = P[(X - \mu_X)^2 \geq t^2\sigma_X^2].$$

   The random variable $Y = (X - \mu_X)^2$ has expectation σ_X^2 (def. of variation). Applying the Markov inequality to Y bounds this probability from above by $1/t^2$.

This bound gives a little bit better results since it uses the "knowledge" of the variance of the variable.

We will use it later to analyze a randomized selection alg.

Chernoff Inequality

The first “good Tail Inequality”.

Assumption: sum X of independent random variables counting variables (binomially distributed X)

Lemma: Let $X_1, X_2 \cdots, X_n$ be independent $0-1$ variables. $P[X_i = 1] = p_i$ with $0 \leq p_i \leq 1$. Then, for $X = \sum_{i=1}^{n} X_i$, $\mu = E[X] = \sum_{i=1}^{n} p_i$, and any $\delta > 0$,

$$P[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}.$$

Proof: Use of the moment generating function.

Proof Chernoff bound

For any positive real t,

$$P[X > (1+\delta)\mu] = P[e^{Xt} > e^{t(1+\delta)\mu}].$$

Applying Markov we get

$$P[X\ (1+\delta)\mu] < \frac{E[e^{tX}]}{e^{t(1+\delta)\mu}}.$$

Bound the right hand side:

$$E[e^{tX}] = E[e^{t\cdot\sum_{i=1}^{n} X_i}] = E\left[\prod_{i=1}^{n} e^{tX_i}\right].$$

Since the X_i are independent variables, the variables e^{tX_i} are also independent. We have

$$E\left[\prod_{i=1}^{n} e^{tX_i}\right] = \prod_{i=1}^{n} E\left[e^{tX_i}\right], \text{ and}$$

$$P[X > (1+\delta)\mu] < \frac{\prod_{i=1}^{n} E[e^{tX_i}]}{e^{t(1+\delta)\mu}}.$$

Proof Chernoff bound II

Now note that e^{tX_i} assumes the value e^t with probability p_i and the value 1 with probability $1 - p_i$. Hence,

$$P[X > (1+\delta)\mu] < \frac{\prod_{i=1}^n p_i e^t + (1-p_i)}{e^{t(1+\delta)\mu}} = \frac{\prod_{i=1}^n 1 + p_i(e^t - 1)}{e^{t(1+\delta)\mu}}$$

Since $1 + x \le e^x$ with $x = p_i(e^t - 1)$ we obtain

$$P[X > (1+\delta)\mu] < \frac{\prod_{i=1}^n e^{p_i(e^t-1)}}{e^{t(1+\delta)\mu}} = \frac{e^{\sum_{i=1}^n p_i(e^t-1)}}{e^{t(1+\delta)\mu}}$$

and finally

$$P[X > (1+\delta)\mu] < \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}}.$$

The above has been proved for any positive real t. We are free to chose the t that results in the best bound. Substituting $t = \ln(1+\delta)$ gives the result.

Coupon Collector Problem

There are n types of coupons and at each trial a coupon is chosen at random.

Each random coupon is equally likley to be any of the n types and the trials are independent (Kinderschokolade!).

Question: How many trials do I need to have at least one copy of each coupon?

Theorem: With a probability of $n^{-\beta+1}$, $\beta \cdot n \ln n$ trials are sufficient.

Proof: Let X be the number of trials required to collect at least one of each coupon.

Let C_i denotes the type of the ith coupon.

We call the ith trial a *success* if $C_i \notin C_1, C_2, \ldots, C_{i-1}$.

Epoch i begins with the trial following the ith success and ends with the trial when the $(i+1)$st success is achieved.

Define X_i, $1 \leq i \leq n-1$, to be the number of trials in the ith epoch. Hence,

$$X = \sum_{i=0}^{n-1} X_i.$$

Coupon Collector II

Let p_i be the probability of a success in epoch i. Then

$$p_i = \frac{n-i}{n}.$$

X_i is geometrically distributed with

$$E[X_i] = \frac{1}{p_i} \qquad \text{and} \qquad V[X_i] = \frac{1-p_i}{p_i}.$$

We have

$$E[X] = E\left[\sum_{i=0}^{n-1} X_i\right] = \sum_{i=0}^{n-1} \cdot \frac{n}{n-i} = n \sum_{i=0}^{n-1} \frac{1}{i} = n \cdot H_n.$$

Note that H_n is the nth Harmonic number. It is asymptotically equal to $\ln n + \Theta(1)$.

Coupon Collector III

Since the X_i's are independent we have

$$V[X] = \sum_{i=0}^{n-1} V[X_i],$$

and

$$V[X] = \sum_{i=0}^{n-1} \frac{ni}{(n-i)^2} = \sum_{i=1}^{n} \frac{n(n-i)}{i^2} = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - nH_n.$$

$\sum_{i=1}^{n} 1/i^2$ converges to a constant and $V[X] = O(n^2 - nH_n)$.

Now we are ready to apply Chebytschev:

$$P[X - E[X] \geq E[X]] \approx P[X - E[X] \geq nH_n] \approx n^2 - H_n/nH_n$$

With $t = \frac{nH_n}{\sqrt{V[X]}}$.

and that is far too weak!

Randomized Selection

We use random sampling to select the kth smallest element of an ordereded set S .

Some definitions:

- $r_s(t)$ is the rank of an element t in set S.
- $S_{(i)}$ is the ith smallest element of S.

We sample with replacement, meaning that we can chose the same element several times.

LazySelect

Input: Ordered set S of n elements and an integer $k \leq n$. **output:** kth smallest element of S.

1. $x = kn^{-1/4}$, $\ell = \max\{\lfloor x - \sqrt{n} \rfloor, 1\}$, and $h = \min\{\lceil x + \sqrt{n} \rceil, n^{3/4}\}$.
2. Pick $n^{3/4}$ elements form S, chosen i.u.r. with replacement. Call this set R.
3. Sort R in time $O(n^{3/4} \log n) = O(n)$.
4. Let $a = R_{(\ell)}$ and $b = R_{(h)}$. Compare a and b to every element of S and compute $r_S(a)$ and $r_S(b)$.
5. Now compute a subset P
   - If $k < n^{1/4}$ then $P = \{y \in S \mid y \leq b\}$,
   - else If $k > n - n^{1/4}$, let $P = \{y \in S \mid y \geq b\}$,
   - else If $k \in [n^{1/4}, n - n^{1/4}]$, let $P = \{y \in S \mid a \leq y \leq b\}$.

   Check whether $S_{(k)} \in P$ and $|P| \leq 4n^{3/4} + 2$. If not, repeat steps 1-4 until such P is found.
6. By sorting P in $O(|P| \log |P|)$ steps, identify $P_{k - r_S(a) + 1}$, which is $S_{(k)}$.

Analysis of LazySelect

The idea of the algorithm is to identify two elements a and b such that both of the following statements hold with high probability $(1 - 1/n^{\alpha})$:

- The element $S_{(k)}$ that we seek is in P.
- The set P of elements between a and b is not very large, so that we can sort it in time $O(n)$.

Theorem With probability $1 - O(n^{-1/4})$, LazySelect finds $S_{(k)}$ on the first pass and thus performs only $2n + o(n)$ comparisions.

We have to consider three cases, here we consider the case $k \in [n^{1/4}, n - n^{1/4}$ and $P = \{y \in S \mid a \leq y \leq b\}$. The alnalysis of the other two cases is similar.

Case 1 We fail 1) if $a > S_{(k)}$ or $b < S_{(k)}$. This means fewer than ℓ samples should be smaller than $S_{(k)}$/ at least h samples should be smaller than $S_{(k)}$.

Let's consider the event $a > S_{(k)}$. Let $X_i = 1$ if the ith random sample is at most $S_{(k)}$, and 0 otherwise (Bernoulli trials).

Let $X = \sum_{i=1}^{n^{3/4}} X_i$.

$$P[X_i = 1] = \frac{k}{n} \text{ and } E[X] = \frac{k}{n^{1/4}}$$

$$\sigma_X^2 = n^{3/4}\left(\frac{k}{n}\right)\left(1 - \frac{k}{n}\right) \leq \frac{n^{3/4}}{4} \text{ and } \sigma_X \leq \frac{n^{3/8}}{2}$$

Analysis of LazySelect II

Now we are ready to apply Chebyshev bounds on X.

$$P[a > S_{(k)}] = P[|X - E[X]| \geq \sqrt{n}] \leq P[|X - E[X]| \geq 2n^{1/8}\sigma_x] \leq \frac{1}{4n^{1/4}}.$$

A similar argument shows that $P[b < S_{(k)}] \leq \frac{1}{4n^{1/4}}$.

Case 2) We have to estimate the probability that P contains more than $4n^{3/4} + 2$ elements. This case can be done very similar to case 1) and is a nice question for your assignments.