Events and probability distribution

- An *experiment* is a procedure that yields one of a given set of possible outcomes. The *sample space* of an experiment is the set of possible outcomes. Each outcome in a sample space is also known as a *sample point*. A sample space is called *discrete* if it contains finite or countably infinite number of sample points. The *discrete probability* studies the probability theory when the sample space is discrete.
- Any subset of the sample space is known as an *event*. Each sample point is a *simple (indecomposable) event*; otherwise, it is a *compound (decomposable) event*. Further, an event that has no sample points is denoted by ϕ.

  For any two events E and F of a sample space S, event $E \cup F$ consists of all the sample points that are either in E or in F, event $E \cap F$ consists of all the sample points that are both in E and in F, event $\overline{E}$ consists of all sample points that are part of S but not in E, $E \subseteq F$ denotes that all sample points of E are included in F, etc.
- For a discrete sample space S, consider any function $p : S \rightarrow [0, 1]$ such that $\sum_{s \in S} p(s) = 1$. Then $p(s)$ is said to be the *probability* of sample point s. The function p is called a *probability distribution* over the sample spce S.

  The probability of any event E is defined as $\sum_{\text{sample point } s \in E} p(s)$. And, the probability of an empty event, $p(\phi)$, is equal to zero.
- If S is a finite nonempty sample space of equally likely sample points, and E is an event, that is a subset of S, then the *probability* of E is $p(E) = \frac{|E|}{|S|}$. The *uniform distribution* assigns the probability $\frac{1}{|S|}$ to each outcome in the sample space S. Indeed, when the probability of choosing any element from a sample space S is equal to the probability of choosing any other element of S, then the elements are said to be chosen *uniformly at random*.

* $p(\overline{E}) = 1 - p(E)$.

- proof: since $|\overline{E}| = |S| - |E|$, $p(\overline{E}) = \frac{|S|-|E|}{|S|}$

- The *odds ratio* of any event A is $\frac{p(A)}{p(\overline{A})}$, that is, $\frac{p(A)}{1-p(A)}$.

* If $E \subseteq F$ then $p(E) \leq p(F)$.

* *Inclusion-exclusion principle*: $p(E_1 \cup E_2 \cup \ldots \cup E_n) = \sum_{i=1}^{n} p(E_i) - \sum_{i_1 < i_2} p(E_{i_1} \cap E_{i_2}) + \ldots + (-1)^{r+1} \sum_{i_1 < i_2 < \ldots < i_r} p(E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_r}) + \ldots + (-1)^{n+1} p(E_1 \cap E_2 \cap \ldots \cap E_n)$.

* *Union bound (a.k.a., Boole's inequality)*: $p(\bigcup_{i \geq 1} E_i) \leq \sum_{i \geq 1} p(E_i)$.

- proof: from the inclusion-exclusion principle

- Two events E_1, E_2 are *mutually exclusive (a.k.a., disjoint)* if they cannot occur at the same time, that is, $p(E_1 \cap E_2) = 0$.

  For any finite or countably infinite sequence of *pairwise mutually exclusive events* $E_1, E_2, E_3, \ldots$, the union bound reduces to $p(\bigcup_{i \geq 1} E_i) = \sum_{i \geq 1} p(E_i)$.

- Typical pipeline of problem solving: find the sample space (tree diagrams could help in understanding the sample space), define events of interest, associate probabilities to sample points, and compute event probabilities.

Conditional probability

- Let H be an event with a positive probability. For two arbitrary events A and H, the *conditional probability* of A on the hypothesis H is denoted by $p(A|H)$, and is equal to $\frac{p(A \cap H)}{p(H)}$. That is, H becomes the new sample space with probability of each sample point in H is associated to probability proportional to probability of that sample point in the original sample space, and the sum of new probabilities associated to sample points in H is equal to one.

  The conditional probability $p(B|A)$ is called a *posteriori* if event B precedes event A in time. Otherwise, it is called a *priori*.

* *Product rule of conditional probability*: $p(A \cap H) = p(A|H)p(H)$.

  Generalizing the same yields, $p(E_1 \cap E_2 \cap \ldots \cap E_n) = p(E_1|E_2 \cap E_3 \cap \ldots \cap E_n)p(E_2|E_3 \cap \ldots \cap E_n) \ldots p(E_n)$.

* $p((A \cup B)|H) = p(A|H) + p(B|H) - p((A \cap B)|H)$.

* *The law of total probability*: Let $H_1, \ldots, H_n$ be a set of mutually exclusive events that partition the sample space, then for any arbitrary event A, $p(A) = \sum_{j=1}^{n} p(A|H_j)p(H_j)$.

  - proof: since $p(A) = \sum_{j=1}^{n} p(A \cap H_j)$

* *Bayes' rule*: Let $H_1, \ldots, H_n$ be a set of mutually exclusive events that partition the sample space, then for any arbitrary event A. Then, $p(H_k|A) = \frac{p(A|H_k)p(H_k)}{\sum_j p(A|H_j)p(H_j)}$.

  - proof: since $p(H_k|A)p(A) = p(A|H_k)p(H_k)$, and due to law of total probability

- Corollary: Suppose that E and F are events from a sample space S such that $p(E) \neq 0$ and $p(F) \neq 0$. Then, $p(F|E) = \frac{p(E|F)p(F)}{p(E|F)p(F)+p(E|\overline{F})p(\overline{F})}$.

- Let S be the set of balls in a bin, and for any ball $b \in S$ let $p(b)$ be the probability of selecting b from S so that $\sum_{b \in S} p(b) = 1$. A *random sample* S' is a subset of S, wherein each ball in S' is chosen according to probability distribution function p. When the ball selected is not returned to the bin before the next ball is selected then this sampling is termed as *sampling without replacement*. In this case, the probability of selecting a ball from the bin changes from one iteration to another. On the other hand, if the ball selected is returned to the bin before the next ball is selected then this sampling is called *sampling with replacement*. When compared with sampling without replacement, sampling with replacement is often simpler to code and the effect on the probability of making an error is almost negligible, hence the former is a desirable alternative to the latter.

- Two events A and H are said to be *stochastically independent* (or, *independent*) whenever $p(A \cap H) = p(A)p(H)$. For two independent events A and H with $p(H) \neq 0$, it is immediate that $p(A|H) = p(A)$. In other words, A does not depend on the occurance of hypothesis H. The definition of independent events is accepted even when $p(H) = 0$, however, in this case, $p(A|H)$ is not defined.

If A and H are independent events, then so are A and $\overline{H}$, $\overline{H}$ and A, and $\overline{A}$ and $\overline{H}$.

- proof: immediate

We know, for any two disjoint events A and B, $p(A \cap B) = 0$. However, $p(A)p(B) = 0$ only if either $p(A) = 0$ or $p(B) = 0$; that is, two disjoint events are not independent unless one of them has probability zero.

- The set S of events are said to be *mutually independent* whenever for every subset $E'_1, E'_2, \ldots, E'_m$ (with $2 \leq m \leq n$) of events from S, $p(E'_1, E'_2, \ldots, E'_m) = p(E'_1)p(E'_2) \ldots p(E'_m)$. Intuitively, the probability of each event in S is the same no matter which other events has occurred. For any mutually independent events E_1, E_2, and E_3, the following are mutually independent as well: $E_1 \cup E_2, E_3$; $E_1 \cup E_3, E_2$; $E_2 \cup E_3, E_1$.

  The events $E_1, E_2, \ldots, E_n$ are said to be *k-wise independent* whenever every k-set of these events is mutually independent.
  A set S, with $|S| > 2$, of events are mutually independent $\Rightarrow$ events in S are k-wise independent for any fixed integer $k \in [2, |S|]$.
  Whereas, a set S, with $|S| > 2$, of events are k-wise independent for some fixed integer $k \in [2, |S|] \not\Rightarrow$ events in S are mutually independent.

Random variables

- A *random variable* is a (total) function from the sample space of an experiment to the set of real numbers. That is, a random variable assigns a real number to each possible outcome. (Significantly, the random variable is a function, it is not a variable, and it is not random as well!) A discrete random variable is a random variable that takes on only a finite or countably infinite number of values.

  Let X be a random variable from the sample space S, and let $x_1, x_2, \ldots,$ be the values that it assumes. The aggregate of all sample points on which X assumes a fixed value x_j forms the event that $X = x_j$, and the probability of this event is denoted by $p(X = x_j)$. The function $p(X = x_j) = \sum_{s \in S, X(s) = r} p(s)$ is called the *probability distribution of random variable* X or the *probability mass function of* X. Clearly, $\forall_j p(X = x_j) \geq 0$ and $\sum_j p(X = x_j) = 1$. Typically, probabilities are associated directly to the range of X.

- For an event that either happens or does not happen, the former with probability p and the latter with probability $1 - p$, an *indicator random variable* (a.k.a., *Bernoulli random variable*) is a random variable that is equal to 1 if the event happens and 0 otherwise. That is, an indicator random variable maps every outcome of the sample space to $\{0, 1\}$.

- Consider two random variables X and Y defined on the same sample space, and denote the values which they assume, respectively by $x_1, x_2, \ldots,$ and $y_1, y_2, \ldots$. The aggregate of sample points in which the two conditions $X = x_j$ and $Y = y_k$ are satisfied forms an event whose probability is denoted by $p(X = x_j, Y = y_k)$. The function $\forall_{j,k} p(X = x_j, Y = y_k)$ is called the *joint probability distribution* of X and Y. (This defintion can be extended to more than two random variables by obvious means.)

  Clearly, $\forall_{j,k} p(X = x_j, Y = y_k) \geq 0$, and $\sum_{j,k} p(X = x_j, Y = y_k) = 1$.

The $p(X = x_j)$ (resp., $p(Y = y_k)$) is called the *marginal probability* of X, to denote that more than one random variable is involved: For every fixed j, $\sum_k p(X = x_j, Y = y_k) = p(X = x_j)$. Similarly, for every fixed k, $\sum_j p(X = x_j, Y = y_k) = p(Y = y_k)$.

- Let $X = x_j$ be an event with positive probability. The conditional probability of event $Y = y_k$ given $X = x_j$ is denoted by $p(Y = y_k|X = x_j)$, and is equal to $\frac{p(Y=y_k \cap X=x_j)}{p(X=x_j)}$.

  Two random variables X and Y are said to be *(stochastically) independent random variables* whenever $\forall_{j,k}\, p(X = x_j \cap Y = y_k) = p(X = x_j)Y(Y = y_k)$.

  A a set of random variables $X_1, \ldots, X_n$ are said to be *pairwise independent set of random variables* whenever for any pair i, j and for any real numbers α and β, $p((X_i = \alpha) \cap (X_j = \beta)) = p(X_i = \alpha)p(X_j = \beta)$.

  A collection of random variables are said to be *independent and identically distributed (i.i.d.)* whenever those random variables are independent and any two random variables in that collection has the same probability distribution.

- If $X, Y, Z, \ldots$ are random variables defined on the same sample space, then any function $f(X, Y, Z, \ldots)$ is again a random variable. Its distribution can be obtained from the joint distribution of $X, Y, Z, \ldots$.

Expectation of a random variable

- The *expected value*, also called the *expectation* or *mean*, of a random variable X on the sample space S is equal to $E(X) = \sum_{s \in S} X(s)p(s) = \sum_{i \in X(s)} ip(X = i)$. (We assume this series converges, and hence X has finite expectation.)

- If X is a random variable with a uniform distribution on $\{a_1, a_2, \ldots, a_n\}$ then $E(X) = \frac{a_1+a_2+\ldots+a_n}{n}$.

  The expected value of an indicator random variable I for an event is equal to the probability p of that event.

  - proof: $E[I] = 0 \cdot p(I = 0) + 1 \cdot p(I = 1) = p(I = 1) = p$

  Let X be a discrete random variable that takes on only nonnegative integer varlues. Then, $E[X] = \sum_{i=1}^{\infty} p(X \geq i)$.

  - proof: $\sum_{i=1}^{\infty} p(X \geq i) = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} p(X = j) = \sum_{j=1}^{\infty} \sum_{i=1}^{j} p(X = j) = \sum_{j=1}^{\infty} jp(X = j) = E[X]$

- Let $X, X_1, X_2, \ldots, X_n$ be random variables on S. Then, the *linearity of expectations* says the following:
  (a) $E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$,

  - proof: generalize the following, $E(X_1 + X_2) = \sum_{s \in S} p(s)(X_1(s) + X_2(s)) = \sum_{s \in S} p(s)X_1(s) + \sum_{s \in S} p(s)X_2(s)$

  (b) $E(aX + b) = aE(X) + b$, and

  - proof: $E(aX + b) = \sum_{s \in S} p(s)(aX(s) + b) = a \sum_{s \in S} p(s)X(s) + b \sum_{s \in S} p(s) = aE(X) + b$

  (c) $E(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i E(X_i)$ for constants $a_1, a_2, \ldots, a_n \in \mathbb{R}$.

  - proof: from the proofs of above two propositions

- Let X, Y be independent random variables. Also, let f, g be any functions. Then, $f(X)$ and $g(Y)$ are independent.

  - proof: $p((f(X) = a)(g(Y) = b)) = p((X \in f^{-1}(\{a\}))(Y \in g^{-1}(\{b\}))) = p(X \in f^{-1}(\{a\})p(Y \in g^{-1}(\{b\})) = p(f(X) = a)p(g(Y) = b)$

- Let $X, X_1, X_2, \ldots, X_n$ be a collection of independent random variables defined on the sample space S. Then, $E[\Pi_{i=1}^n X_i] = \Pi_{i=1}^n E[X_i]$.

  - proof: for pairwise independent random variables X and Y,
    $E[X \cdot Y] = \sum_i \sum_j (i \cdot j) \cdot p((X = i) \cap (Y = j)) = \sum_i \sum_j (i \cdot j) \cdot p(X = i) \cdot p(Y = j) = (\sum_i i \cdot p(X = i))(\sum_j j \cdot p(Y = j)) = E[X] \cdot E[Y]$.

- *A weaker version of Jensen's inequality*: $E[X^2] \geq (E[X])^2$.

  - proof: consider the random variable $Y = (X - E[X])^2$; since Y is nonnegative, $0 \leq E[Y]$

  Jensen's inequality: For a convex function f, $E[f(X)] \geq f(E[X])$.

  - Two definitions for a function to be a convex:
    A function $f : \mathbb{R} \to \mathbb{R}$ is convex if, for any x_1, x_2 and $0 \leq \lambda \leq 1$, $f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$.
    If f is a twice differentiable function, then f is convex iff $f''(x) \geq 0$.

  - proof: Assuming f has a Taylor expansion and $\mu = E[X]$, there is a value c such that
    $f(x) = f(\mu) + f'(\mu)(x - \mu) + \frac{f''(c)(x-\mu)^2}{2} \geq f(\mu) + f'(\mu)(x - \mu)$, since $f''(c) > 0$ by convexity.
    Then, $E[f(X)] \geq E[f(\mu) + f'(\mu)(X - \mu)] = E[f(\mu)] + f'(\mu)(E[X] - \mu) = f(\mu) = f(E[X])$.

- The *deviation* of X at $s \in S$ is $X(s) - E(X)$. Let X be a random variable with probability distribution $\{f(x_j)\}$, and let $r \geq 0$ be an integer. The r^{th} *moment of X (about the origin)* is $E(X^r)$, which is equal to $\sum_j x_j^r f(x_j)$. (We assume $E(X^r)$ exists.)

  Let $E(X)$ and $E(X^2)$ be the first and second moments of a random variable X. Then, the *variance* (a.k.a., *dispersion*) of X, denoed by $Var[X]$ or σ_X^2, is defined as $E((X - E(X))^2)$. This characterizes how widely a random variable is distributed: small variance indicates large deviations of X from μ are improbable; large variance indicates that not all values assumed by X lie near the mean.

  The *standard deviation* of X, denoted by σ_X, is $\sqrt{Var[X]}$. This measures how spread out the distribution of X around its mean; useful as its units are the same as $E(X)$.

  A few properties of variance:

  (a) $Var[X] = E[X^2] - (E[X])^2$.

  - proof: $Var[X] = E[(X - E[X])^2] = E[X^2 - 2XE[X] + E[X]^2] = E[X^2] - 2E[XE[X]] + E[X]^2 = E[X^2] - 2E[X]E[X] + E[X]^2 = E[X^2] - (E[X])^2$.

  (b) $Var[X + Y] = Var[X] + Var[Y] + 2E[(X - E[X])(Y - E[Y])]$.

  - proof: $Var[X + Y] = E[(X + Y - E[X + Y])^2] = E[(X + Y - E[X] - E[Y])^2] = E[(X - E[X])^2 + (Y - E[Y])^2 + 2(X - E[X])(Y - E[Y])]$

  (c) $Var[aX + b] = a^2 Var[X]$.

  - proof: — homework —

(d) *Bienayme's formula*: If the X_js are pairwise independent, then $Var[X_1+X_2+\ldots+X_n] = Var[X_1]+Var[X_2]+\ldots+Var[X_n]$.

- proof:

  we know $Var[X+Y] = Var[X]+Var[Y]+2E[(X-E[X])(Y-E[Y])] = E[(X-E[X])^2]+E[(Y-E[Y])^2]+2E[(X-E[X])(Y-E[Y])]$;

  since X and Y are independent, $X-E[X]$ and $Y-E[Y]$ are independent as well; hence, $E[(X-E[X])(Y-E[Y])] = E[X-E[X]]E[Y-E[Y]]$;

  for any random variable Z, $E[Z-E[Z]] = E[Z]-E[E[Z]] = 0$;

  therefore, $Var[X+Y] = E[(X-E[X])^2]+E[(Y-E[Y])^2] = Var[X]+Var[Y]$

- If X and Y are two random variables with joint distribution, the *conditional expectation* of Y for a given $X = x_j$, denoted by $E(Y|X=x_j)$, is $\sum_{y_k \in range(Y)} y_k p(Y=y_k|X=x_j)$. Essentially, the expression $E(Y|X)$ is a random variable $f(X)$ that takes on the value $E(Y|X=x_j)$ when $X=x_j$. As the following law depicts, the conditional expectations are quite useful in dividing the expectation calculation into simpler cases.

  The law of total expectations: For any random variables X and Y,
  $E(X) = \sum_{y \in range(Y)} p(Y=y)E(X|Y=y)$.

  - proof: $\sum_y p(Y=y)E[X|Y=y] = \sum_y p(Y=y)\sum_x xp(X=x|Y=y) = \sum_x \sum_y xp(X=x|Y=y)p(Y=y) = \sum_x \sum_y xp(X=x \cap Y=y) = \sum_x xp(X=x) = E[X]$

  For any finite collection of discrete random variables $X_1, X_2, \ldots, X_n$ with finite expectations and for any random variable Y, $E[\sum_{i=1}^n X_i|Y=y] = \sum_{i=1}^n E[X_i|Y=y]$. (That is, the linearity of expectations extends to conditional expectations.)

  - proof: — homework —

  For random variables X and Y, $\quad E[X] = E[E[X|Y]]$.

  - proof: since $E(X|Y)$ is a random variable $f(Y)$ that takes on the value $E(X|Y=y)$ when $Y=y$,
    $E[E[X|Y]] = \sum_{z \in range(Y)} E[X|Y=y]p(Y=y)$, which is equal to $E[X]$ according to the law of total expectations.

Popular distributions

- *Bernoulli distribution*: Repeated independent trials are called *Bernoulli trials* if there are only two possible outcomes (success, failure) for each trial and their probabilities, p and q respectively, remain the same throughout the trials.

  For a Bernoulli random variable X that has value 1 if the result is heads and 0 otherwise, $E[X] = p$ and $Var[X] = p(1-p)$.

  - proof:

    $E[X] = p \cdot 1 + (1-p) \cdot 0 = p = p(X=1)$

    $Var[X] = E[X^2] - (E[X])^2 = (0^2 \cdot p(X=0) + 1^2 \cdot p(X=1)) - (p)^2$

  A sequence of independent Bernoulli trials with success probability associated to each specific trial is called a *Poisson trial*. This is different from the Poisson approximation/distribution, which is described later.

- *Binomial distribution*: The probability that n Bernoulli trials with probabilities p for success and $q = 1-p$ for failure result in k successes and $n-k$ failures, denoted by $b(k;n,p)$, is $\binom{n}{k}p^k q^{n-k}$. (This assumes order of successes in n trials does not matter.)

  Since $\sum_{i=0}^{n} b(i;n,p) = 1$, $b(k;n,p)$ is a probability distribution.

  For a binomial random variable X that has value i if there are i successes, $E[X] = np$ and $Var[X] = npq$.

  - proof:

    $E[X] = \sum_{j=0}^{n} j\binom{n}{j}p^j(1-p)^{n-j} = \sum_{j=0}^{n} j\frac{n!}{j!(n-j)!}p^j(1-p)^{n-j} = \sum_{j=1}^{n}\frac{(n-1)!}{(j-1)!((n-1)-(j-1))!}p^{j-1}(1-p)^{(n-1)-(j-1)}$
    $= np\sum_{k=0}^{n-1}\frac{(n-1)!}{k!((n-1)-k)!}p^k(1-p)^{(n-1)-k} = np\sum_{k=0}^{n-1}\binom{n-1}{k}p^k(1-p)^{(n-1)-k} = np$

    $E[X^2] = \sum_{j=0}^{n}\binom{n}{j}p^j(1-p)^{n-j}j^2 = \sum_{j=0}^{n}\frac{n!}{(n-j)!j!}p^j(1-p)^{n-j}((j^2-j)+j) = \sum_{j=0}^{n}\frac{n!(j^2-j)}{(n-j)!j!}p^j(1-p)^{n-j} +$
    $\sum_{j=0}^{n}\frac{n!j}{(n-j)!j!}p^j(1-p)^{n-j} = n(n-1)p^2\sum_{j=2}^{n}\frac{(n-2)!}{(n-j)!(j-2)!}p^{j-2}(1-p)^{n-j} + np\sum_{j=1}^{n}\frac{(n-1)!}{(n-j)!(j-1)!}p^{j-1}(1-p)^{n-j} =$
    $n(n-1)p^2 + np$

    $Var[X] = E[X^2] - (E[X])^2 = n(n-1)p^2 + np - n^2p^2 = np - np^2 = np(1-p)$

- *Poisson approximation to binomial distribution*: Let X_n be a binomial random variable with parameters n and p, where p is a function of n and $\lim_{n\to\infty} np = \lambda$ is a constant that is independent of n. Then, for any fixed k, $\lim_{n\to\infty} p(X_n = k) = \frac{e^{-\lambda}\lambda^k}{k!}$.

  - proof:

    $p(X_n = k) = \binom{n}{k}p^k(1-p)^{n-k} \le \frac{n^k}{k!}p^k\frac{(1-p)^n}{(1-p)^k} \le \frac{(np)^k}{k!}\frac{e^{-pn}}{1-pk} = \frac{e^{-pn}(np)^k}{k!}\frac{1}{1-pk}$

    $p(X_n = k) = \binom{n}{k}p^k(1-p)^{n-k} \ge \frac{(n-k+1)^k}{k!}p^k(1-p)^n \ge \frac{((n-k+1)p)^k}{k!}e^{-pn}(1-p^2)^n \ge e^{-pn}\frac{((n-k+1)p)^k}{k!}(1-p^2n)$

    hence, $\lim_{n\to\infty}\frac{e^{-pn}(np)^k}{k!}\frac{1}{1-pk} \ge \lim_{n\to\infty} p(X_n = k) \ge \lim_{n\to\infty} e^{-pn}\frac{((n-k+1)p)^k}{k!}(1-p^2n)$

    since, as n approaches infinity, p approaches zero, $\frac{e^{-\lambda}\lambda^k}{k!} \ge \lim_{n\to\infty} p(X_n = k) \ge \frac{e^{-\lambda}\lambda^k}{k!}$

  And, the empirical evidence shows Poisson's distribution very well approximates the binomial distribution.

  Since $\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = 1$, $p(k;\lambda)$ can be conceived as the probability of exactly k successes, called as *Poisson distribution*.

  For a Poisson random variable X that has value i if there are i successes, $E[X] = \lambda$ and $Var[X] = \lambda$.

  - proof:

    $E[X] = \sum_{k\ge 0} k\frac{1}{k!}\lambda^k e^{-\lambda} = \lambda e^{-\lambda}\sum_{k\ge 1}\frac{1}{(k-1)!}\lambda^{k-1} = \lambda e^{-\lambda}\sum_{j\ge 0}\frac{\lambda^j}{j!} = \lambda e^{-\lambda}e^{\lambda} = \lambda$

    $E[X^2] = \sum_{k\ge 0} k^2\frac{1}{k!}\lambda^k e^{-\lambda} = \lambda e^{-\lambda}\sum_{k\ge 1} k\frac{1}{(k-1)!}\lambda^{k-1} = \lambda e^{-\lambda}(\sum_{k\ge 1}(k-1)\frac{1}{(k-1)!}\lambda^{k-1} + \sum_{k\ge 1}\frac{1}{(k-1)!}\lambda^{k-1}) =$
    $\lambda e^{-\lambda}(\lambda\sum_{k\ge 2}\frac{1}{(k-2)!}\lambda^{k-2} + \sum_{k\ge 1}\frac{1}{(k-1)!}\lambda^{k-1}) = \lambda e^{-\lambda}(\lambda\sum_{i\ge 0}\frac{1}{i!}\lambda^i + \sum_{j\ge 0}\frac{1}{j!}\lambda^j) = \lambda e^{-\lambda}(\lambda e^{\lambda} + e^{\lambda}) = \lambda(\lambda+1)$

    $Var[X] = \lambda^2 + \lambda - \lambda^2 = \lambda$

- *Geometric distribution*: A geometric random variable X with parameter p is given by the following probability distribution: $p(X = n) = (1-p)^{n-1}p$. That is, for the geometric random variable X to be equal to n, there must be $n-1$ failures followed by a success. Here, n is any positive integer.

  For a geometric random variable X, $\sum_{n\ge 1} p(X = n) = 1$. That is, success is guaranteed to occur eventually, considering there are infinitely many trials.

  - proof: $p\sum_{n\ge 1}(1-p)^n = p((1-p)^0 + (1-p)^1 + \ldots) = 1$

Memoryless (a.k.a., Markov's) property: for a geometric random variable X with $n, k > 0$, $p(X = n + k|X > k) = p(X = n)$.

- proof: $p(X = n+k|X > k) = \frac{p((X=n+k)\cap(X>k))}{p(X>k)} = \frac{p(X=n+k)}{p(X>k)} = \frac{(1-p)^{n+k-1}p}{\sum_{i=k}^{\infty}(1-p)^i p} = \frac{(1-p)^{n+k-1}p}{(1-p)^k} = (1-p)^{n-1}p = p(X = n)$

For a geometric random variable X that has value i if the success occurs at i^{th} trial, $E[X] = \frac{1}{p}$ and $Var[X] = \frac{p}{(1-p)^2}$.

- proof:
  since $p(X \geq i) = \sum_{n=i}^{\infty}(1-p)^{n-1}p = (1-p)^{i-1}$, $E[X] = \sum_{i=1}^{\infty} p(X \geq i) = \sum_{i=1}^{\infty}(1-p)^{i-1} = \frac{1}{1-(1-p)}$
  $E[X^2] = \sum_{k\geq 1} k^2 qp^k = p\sum_{k\geq 1} k^2 qp^{k-1} = p(\frac{2}{q^2} - \frac{1}{q})$
  $Var[X] = E[X^2] - (E[X])^2 = p(\frac{2}{q^2} - \frac{1}{q}) - (\frac{1}{1-(1-p)})^2 = \frac{p}{(1-p)^2}$

- *Negative binomial distribution (a.k.a., Pascal distribution)*: The probability distribution of the number X of Bernoulli trials needed to find the r^{th} success at the trial number $n = r + k$, when the success probability of a Bernoulli trial is p. The probability, $p(X = r + k)$, the r^{th} success occurs at the trial number $n = r + k$ (where $k = 0, 1, 2, \ldots$), denoted by $f(k; r, p)$, is $\binom{r+k-1}{k}p^r(1-p)^k$. Since this is equal to $\binom{-r}{k}p^r(-q)^k$, this distribution is named so.

  - proof: $\binom{-n}{k} = \frac{(-n)(-n-1)\ldots(-n-k+1)}{k!} = (-1)^k\frac{n(n+1)\ldots(n+k-1)}{k!} = (-1)^k\frac{(n+k-1)!}{k!(n-1)!} = (-1)^k\binom{n+k-1}{k}$

  Observation: For $r = 1$, any negative binomial distribution is essentially a geoemtric distribution.

  For $p > 0$, since $\sum_{k=0}^{\infty} f(k; r, p) = 1$, the success is guaranteed to occur eventually.

  - proof: due to binomial theorem, $p^r\sum_{k=0}^{\infty}\binom{-r}{k}(-q)^k = p^r(1-q)^{-r} = p^r p^{-r} = 1$

  The sum of geometric random variables follows the negative binomial distribution.

  - proof:
    say, for any i, the iterations following the i^{th} success and ends with the $(i+1)^{th}$ success are said to be in an *epoch*; that is, an experiment is partitioned into epochs;
    hence, the overall number of trials needed to achieve k successes is the count of trials in each epoch;
    since the number of trials in an epoch is a geometric random variable, leading to the observation stated

  For a negative binomial random variable X that has value n if the r^{th} success occurs at the trial number $n = r + k$, $E[X] = \frac{r}{p}$ and $Var[X] = \frac{rq}{p^2}$.

  - proof:
    since $p(X = k) = q^k p$ where $k = 0, 1, 2, \ldots$, $E[X] = qp(1 + 2q + 3q^2 + \ldots) = qp(1-q)^{-2} = \frac{q}{p}$
    let X_i be a random variable denoting the number of Bernoulli trials to be performed to succeed for the i^{th} time after having succeeded $i-1$ times; then, X_i is a geometric random variable with probability of success p; hence, the variance of X_i is $\frac{1-p}{p^2}$; since X_is are all independent $Var[X] = Var[\sum_{i=1}^{r} X_i] = \sum_{i=1}^{r} Var[X_i] = r \cdot \frac{1-p}{p^2}$

  Let X be a negative binomial random variable with parameters r and p. Let Y be a binomial random variable with parameters n and p. Then, $p(X > n) = p(Y < r)$. (This is quite useful since finding the right tail of a negative binomial distribution directly from its definition is difficult.)

  - proof: since X is the minimum number of trials needed to get r successes and Y is the number of successes in n trials, $X > n$ iff $Y < r$

- *Hypergeometric distribution*[1]: Suppose in a population of n balls, n_1 are white and $n_2 = n - n_1$ are black. The probability of choosing r balls from these n balls without replacement so that k ($\min(0 \leq k \leq n_1, r)$) of them are black balls is $\frac{\binom{n_1}{k}\binom{n-n_1}{r-k}}{\binom{n}{r}}$. We note that $\sum_{k=0}^{n} \frac{\binom{n_1}{k}\binom{n-n_1}{r-k}}{\binom{n}{r}} = 1$.

  When n_1 and n are large in relation to r, hypergeometric distribution is same as $b(k; r, \frac{n_1}{n})$, that is, choosing with or without replacement does not matter.

  - proof: $\frac{\binom{n_1}{k}\binom{n-n_1}{r-k}}{\binom{n}{r}} \approx \binom{r}{k}(\frac{n_1}{n})^k(1 - \frac{n_1}{n})^{r-k}$

  For a hypergeometric random variable X that has value k if in r balls k are black, $E[X] = \frac{rn_1}{n}$ and $Var[X] = \frac{n-r}{n-1}rp(1-p)$.

  - proof:

    $E[X] = \sum_{k=0}^{n} \frac{k\binom{n_1}{k}\binom{n-n_1}{r-k}}{\binom{n}{r}} = \frac{rn_1}{n} \sum_{k=1}^{n} \frac{\binom{n_1-1}{k-1}\binom{(n-1)-(n_1-1)}{(r-1)-(k-1)}}{\binom{n-1}{r-1}} = \frac{rn_1}{n} \sum_{t=0}^{n-1} \frac{\binom{n_1-1}{t}\binom{(n-1)-(n_1-1)}{(r-1)-t}}{\binom{n-1}{r-1}} = \frac{rn_1}{n}$

    $Var[X] = \sum_{k=0}^{n} \frac{(k-\frac{rn_1}{n})^2\binom{n_1}{k}\binom{n-n_1}{r-k}}{\binom{n}{r}} = \sum_{k=0}^{n} \frac{k^2\binom{n_1}{k}\binom{n-n_1}{r-k}}{\binom{n}{r}} - \frac{2rn_1}{n} \sum_{k=0}^{n} \frac{k\binom{n_1}{k}\binom{n-n_1}{r-k}}{\binom{n}{r}} + \frac{r^2n_1^2}{n^2} \sum_{k=0}^{n} \frac{\binom{n_1}{k}\binom{n-n_1}{r-k}}{\binom{n}{r}} =$
    $\sum_{k=1}^{n} \frac{k^2\binom{n_1}{k}\binom{n-n_1}{r-k}}{\binom{n}{r}} - \frac{r^2n_1^2}{n^2} = \frac{rn_1}{n} \sum_{k=1}^{n} \frac{(k-1)\binom{n_1-1}{k-1}\binom{n-n_1}{r-k}}{\binom{n-1}{r-1}} + \frac{rn_1}{n} \sum_{k=1}^{n} \frac{(k-1)\binom{n_1-1}{k-1}\binom{n-n_1}{r-k}}{\binom{n-1}{r-1}} - \frac{r^2n_1^2}{n^2} = \frac{rn_1(r-1)(n_1-1)}{n(n-1)} +$
    $\frac{rn_1}{n} - \frac{r^2n_1^2}{n^2}$

- *Multinomial distribution*: Consider events $E_1, E_2, \ldots, E_r$ whose probabilities are respectively $p_1, p_2, \ldots p_r$. The probability that in n independent trials, E_1 occurs k_1 times, E_2 occurs k_2 times, etc., for $k_1 + k_2 + \ldots + k_r = n$, denoted by $m(k_1, k_2, \ldots, k_r; n, p_1, p_2, \ldots, p_r)$, is $\frac{n!}{k_1!k_2!\ldots k_r!}p_1^{k_1}p_2^{k_2}\ldots p_r^{k_r}$.

  - proof: obvious from lectures on counting

  The expected number of times the outcome i was observed over n trials is $E[X_i] = np_i$ and its variance $Var[X_i]$ is $np_i(1-p_i)$.

Tail bounds

The tail inqualities help in deriving bounds on probabilities when only the mean and variance of a probability distribution are known.

- *Markov's*: Let X be a random variable that assumes only nonnegative values. Then, for all $a > 0$, $p(X \geq a) \leq \frac{E[X]}{a}$.

  - proof: let Y be $I\{X \geq a\}$, where I is an indicator random variable;
    then, $p(X \geq a) = E[Y] \leq E[\frac{X}{a}] = \frac{E[X]}{a}$;
    here, $Y \leq \frac{X}{a}$ since if $\frac{X}{a} \geq 1$, then $Y = 1$; otherwise, $0 \leq \frac{X}{a} < 1$ and $Y = 0$

  Corollary: If X is a nonnegative random variable, then for all $c \geq 1$, $\quad p(X \geq cE(X)) \leq \frac{1}{c}$.

- *Chebyshev's*: For any $a > 0$, $\quad p(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2}$.

  - proof: $p(|X - E[X]| \geq a) = p((X - E[X])^2 \geq a^2) \leq \frac{E[(X-E[X])^2]}{a^2} = \frac{Var[X]}{a^2}$

[1] not presented in lectures and hence not included in exam syllabus

Corollary: For any $a > 0$, $\quad p(|X - E[X]| \geq c\sigma_X) \leq \frac{1}{c^2}$.

Here is a nice application, the weak law of large numbers: Let $X_1, X_2, \ldots, X_n$ be pairwise independent variables with the same mean, μ, and deviation, σ (that is, they are identically distributed). Then, for any $\epsilon > 0$, $p(|\frac{X_1+X_2+\ldots+X_n}{n} - \mu| \geq \epsilon) \leq \frac{1}{n}(\frac{\sigma}{\epsilon})^2$. (Significantly, as $n \to \infty$, the RHS tends to 0, saying, the average of independent samples approaches the mean.)

- proof: since $E[\frac{X_1+X_2+\ldots+X_n}{n}] = \mu$ and $Var(\frac{X_1+X_2+\ldots+X_n}{n}) = \frac{\sigma^2}{n}$

- *Chernoff's*:
(a) Let X be a random variable.
For any $t > 0$, $p(X \geq a) \leq \frac{E[e^{tX}]}{e^{ta}}$.
In particular, $p(X \geq a) \leq \min_{t>0} \frac{E[e^{tX}]}{e^{ta}}$.

  - proof: $p(X \geq a) = p(e^{tX} \geq e^{ta}) \leq \frac{E[e^{tX}]}{e^{ta}}$

  For any $t < 0$, $p(X \leq a) \leq \frac{E[e^{tX}]}{e^{ta}}$.
  In particular, $p(X \leq a) \leq \min_{t<0} \frac{E[e^{tX}]}{e^{ta}}$.

  - proof: $p(X \leq a) = p(e^{tX} \geq e^{ta}) \leq \frac{E[e^{tX}]}{e^{ta}}$

  While the value of t that minimizes $\frac{E[e^{tX}]}{e^{ta}}$ gives the best possible bounds, often one chooses a value of t that gives a convenient form.

  (b) Let $X_1, X_2, \ldots, X_n$ be independent Poisson trials such that $p(X_i = 1) = p_i$. Also, let $X = \sum_{i=1}^n X_i$ and $\mu \geq E[X]$.
  For any $\delta > 0$, $\quad p(X \geq (1+\delta)\mu) < (\frac{e^\delta}{(1+\delta)^{(1+\delta)}})^\mu$.
  For any $0 < \delta < 1$, $\quad p(X \leq (1-\delta)\mu) < (\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}})^\mu$.

  - proof: — not proved in class though used them in solving a couple of problems —

- For large n, Chebyshev's tail inequality yields tighter bounds to Markov's. And, Chernoff bounds yield tight (exponentially small) bounds as compared to the polynomially small bounds via Markov's or Chebyshev's tail inequalities.

- *Hoeffding's*: Let $X_1, X_2, \ldots, X_n$ be independent random variables with $E[X_i] = \mu_i$ and $p(a_i \leq X_i \leq b_i) = 1$ for constants a_i and b_i. Then, $p(|\sum_{i=1}^n X_i \sum_{i=1}^n \mu_i| \geq \epsilon) \leq 2e^{-2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2}$.

  - proof: — not proved in class though used it in solving two problems —

References:

* An Introduction to Probability Theory and its Applications vol 1 by W. Feller.

* Probability and Computing by M. Mitzenmacher and E. Upfal.