description

These pages are a summary of the formulas used in CSE103. You can use this as a basis for your *handwritten* "cheat-sheet". However, you cannot bring a printout of these pages to the final exam.

Combinatorics

1. The number of different tuples of length k over an alphabet of size n: n^k
2. The number of ways to order n different objects is $n!$
3. The number of sequences (i.e. order is important) of length k that can be created from a set of n different elements $\frac{n!}{(n-k)!}$
4. Binomial coefficients: The number of subsets(i.e. order is not important) of size k in a set of size n: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
5. For any two real numbers $(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}$. To get the binomial distribution let $a = p$, $b = 1-p$.
6. Suppose we have n_1 objects of type 1, n_2 objects of type 2, etc up to n_k so the total number of objects is $n = n_1 + \ldots + n_k$. The number of different ways to order these n objects is $\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$
7. Bounds on the binomial coefficient: $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$

Discrete probability

1. The union bound: for any events $A_1, A_2, \ldots, A_k$; $P(A_1 \cup \ldots \cup A_k) \leq \sum_{i=1}^{k} P(A_i)$
2. Summation rule: If $A_1, A_2, \ldots, A_k$ are a partition of the sample space Ω, i.e. $A_1 \cup \ldots \cup A_k = \Omega$ and $A_i \cap A_j = \emptyset$ if $i \neq j$. Then $\sum_{i=1}^{k} P(A_i) = 1$.
3. If A and B are arbitrary events $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. Conditional probability: $P(A|B) = P(A \cap B)/P(B)$.
5. Independence: A, B are independent events if $P(A \cap B) = P(A)P(B)$, equivalently, if $P(A|B) = P(A)$.
6. Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
7. Conditional summation rule: If A is an event and B_i is a partition of the sample space then $P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$

Series

- Arithmetic sum: $1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$
- Geometric Series $(0 < r < 1)$:

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}; \quad \sum_{i=1}^{\infty} r^i = \frac{r}{1-r}, \quad \sum_{i=1}^{\infty} i r^i = \frac{r}{(1-r)^2}$$

- If we repeatedly flip a coin whose probability of landing heads is p the expected number of flips until the first heads is:

$$\sum_{i=1}^{\infty} i(1-p)^{i-1}p = \frac{p}{1-p}\sum_{i=1}^{\infty} i(1-p)^i = \frac{p}{1-p}\frac{1-p}{p^2} = \frac{1}{p}$$

- Harmonic Sum: $\sum_{i=1}^{n} 1/i \approx \ln n$

Random variables, expectation and Variance

1. A random variable is a function from the instance space Ω to the real line.
2. Two random variables X, Y whose ranges are finite sets A, B are independent if and only if $P(X = a \text{ and } Y = b) = P(X = a)P(Y = b)$ for all $a \in A$ and $b \in B$. More on indpendent rv's in the next section.
3. The expected value (or mean μ) of a random variable X is defined to be

$$\mu \doteq E(X) \doteq \sum_z z P(X = z)$$

properties of the expected value (X, Y are random variables).

   (a) If a, b are constants: $E(aX + b) = aE(X) + b$.

   (b) Linearity of expectations: $E(X + Y) = E(X) + E(Y)$

   (c) Expectation of a product: $E(XY) = E(X)E(Y)$ **if** X and Y are independent.

(d) Linearity of expectations:
$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$

(e) If $X_1, \ldots, X_n$ are independent then $\text{var}(X_1 + \cdots X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n)$

4. The variance and standard deviation

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

$$\sigma = \text{stddev}(X) = \sqrt{\text{var}(X)}$$

- Variance of a sum: $\text{var}\,(X + Y) = \text{var}\,(X) + \text{var}\,(Y)$ **if** X and Y are independent.
- For any constants a, b: $\text{var}(aX + b) = a^2\text{var}(X)$ (adding a constant to a RV has no effect on its variance).
- The standard deviation of n independent identically distributed (IID) random variables:

$$\text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\text{var}\,(X_i)}{n}$$

and therefor

$$\text{stddev}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\text{stddev}(X_i)}{\sqrt{n}}$$

5. Markov Inequality: if X is a non-negative random variable and $a > 0$ is some constant value, then $P(X \geq a) \leq \frac{E(X)}{a}$

6. Chebyshev's inequality: If μ is the mean and σ is the standard deviation of the rv X

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$$

7. Suppose you toss a coin with bias p and code the outcome as either $X = 1$ (with prob. p) or $X = 0$ (with prob. $1-p$). We say that X has a Bernoulli distribution. $E(X) = p$ and $\text{var}(X) = p(1-p)$.

Independence of RVs, Covariance, correlation and anti-correlation

- Two random variables X, Y are independent if and only if for any constants a, b:

$$P\left(X \leq a \wedge Y \leq b\right) = P(X \leq a)P(Y \leq b)$$

If the random variables are integer valued then they are independent if and only if dor any integers i, j

$$P\left(X = i \wedge Y = j\right) = P(X = i)P(Y = j)$$

- The Covariance of two random variables X, Y is

$$\begin{aligned} Cov(X,Y) &= E\left((X - E(X))(Y - E(Y))\right) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

If X, Y are independent then $E(XY) = E(X)E(Y)$ and therefor the covariance is zero. However the other direction is not true: zero covariance does not imply independence.

- If X, Y are *integer valued* random variables

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \sum_{i,j} ijP(X = i \wedge Y = j) - [\sum_i iP(X = i)][\sum_j jP(Y = j)]$$

- The correlation coefficient is a normalized version of the covariance:

$$\text{cor}(X,Y) = \frac{Cov(X,Y)}{\sqrt{\text{var}\,(X)\,\text{var}\,(Y)}}$$

The correlation coefficient varies between -1 and 1. $\text{cor}(X,Y) = 1$ if and only if there are constants $a > 1, b$ such that $P(aX + b = Y) = 1$. Similarly $\text{cor}(X,Y) = -1$ if and only if $P(-aX + b = Y) = 1$. If the correlation (and the covariance) are positive, we say that X and Y are correlated, if it is negative, we say that X and Y are anti-correlated and if it is zero we say that X and Y are uncorrelated (which does *not* imply that they are independent.)

Distributions over the real line

- Suppose X is a random variable defined over the whole real line from $-\infty$ to ∞. The probability distribution for such a random variable can *always* be represented by the Cumulative Distribution Function (CDF) $F(a) \doteq P(x \leq a)$. The CDF is monotone non-decreasing.

- The probability of a segment can be computed as $P(a < X \leq b) = F(b) - F(a)$.

- The uniform distribution on the segment $[a, b]$ corresponds to the CDF that is 0 for $x \leq a$, $\frac{x-a}{b-a}$ for $a \leq x \leq b$ and 1 for $x > b$. We denote such a distribution by $U(a, b)$

- If the derivative of the CDF is defined everywhere then we call the derivative the Probability Density Function (PDF) $f(x)$. The integral of the PDF is the CDF: $F(a) = \int_{-\infty}^{a} f(x)dx$.

- If the density function is defined at a the probability of the event $\{a\}$ is zero. Conversely, if the probability of a single point is non-zero then the the PDF is not defined at that point and we say that the distribution contains a *point mass* at a, denoted $PM(a)$
- If P_1, P_2 are two distributions, and p a constant between 0 and 1, then $pP_1 + (1-p)P_2$ is also a well-defined distribution P_3. We say that P_3 is a mixture of $P1$ and $P2$.

Poisson Distribution

The poisson distribution describes the number of events in a unit time when the events are distributed uniformly in (continuous, non-discretized) time.

- The number of events per unit time is

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

- The expected number of events per unit time is $E(X) = \lambda$.
- The Poisson distribution is the limit of the binomial distribution where $n \to \infty$ and $p = \lambda/n$:

$$\lim_{n\to\infty}\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}$$

- When events are distributed uniformly in time and the expected number of events per unit time is λ, the time between consecutive events $d = t_{i+1} - t_i$ is distributed according to the density $f(d) = \lambda\exp(-\lambda d)$ and the CDF is:

$$P(d \le s) = 1 - e^{-\lambda s}$$

Statistical tests

- Statistical testing is a methodology for quantifying the the significance of conclusions made based on observations. The *null hypothesis* corresponds to the skeptical opinion stating that the observation is explained well by the null distribution H_0. The alternative hypothesis H_1 represents the new explanation that the experiment is intended to confirm. For example, when testing a new drug, the Null hypothesis states that the drug has no effect and the alternative hypothesis states that it does have a beneficial effect.
- The *statistical test* is a function that takes as input the observations and a *significance values* α and outputs either "Reject Null Hypothesis" or "Fail".
- The *significance level* or α-value of a statistical test is (an upper bound on) the probability that the test rejects the null hypothesis when the data if generated according to the null hypothesis, α *is not a random variable*. It is set to a constant value before the observation data is given..
- The p-value of a test *is a random variable*, it is the minimal value of α that would result in rejecting the null hypothesis. In other words, the test rejects the null hypothesis if $p \le \alpha$.
- A type I error is rejecting the null hypothesis when it is correct. The probability of a type I error is bounded by the chosen value of α. A type II error is failing to reject the null hypothesis when the the alternative hypothesis is correct. Usually, we have no control over type II errors. Increasing α increases the probability of type I errors and decreases the probability of type II errors.

Tests based on Normality

- **Central Limit Theorem** If $X_1, X_2, ...$ are independent, identically distributed random variables with mean μ and variance σ^2 and

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$$

Then, as $n \to \infty$ the CDF of Y_n converges to the CDF of the standard normal distribution $\mathcal{N}(0,1)$, which has the density distribution:

$$p(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

- **the Normal Approximation:** Given a random variable Y whose distribution is normal $\mathcal{N}(\mu, \sigma^2)$ we can calcuate the probability $P(Y > a) = P(Y \ge a)$ (or $P(Y < a) = P(Y \le a)$) for any value of a. The common way of calculating $P(Y > a)$ (or without a computer is transforming the threshold a to a z-score and then using a table for $\mathcal{N}(0,1)$. The formula for calculating z is $z = \frac{a-\mu}{\sigma}$. And the $P(Y > a) = P(X > z)$ where the distribution of X is $\mathcal{N}(0,1)$.

- A few standard definitions: $Q(z) = P(X > z)$, $\Phi(z) = P(X < z)$, $Q^{-1}(p)$ is the inverse function to Q. In other words: $Q(Q^{-1}(p)) = p$.
- few useful values:

$$Q(1) \approx 15\%, \quad Q(2) \approx 2.5\%,$$

$$Q(3) \approx 0.15\%, \quad Q(4) \approx 0.003\%$$

Randomized Algorithms

- **A Las Vegas Algorithm:** always produces the correct output but the time it takes to produce this output can vary. Let μ be the expected running time.
- **A Monte Carlo Algorithm :** always completes within the same amount of time. However, the output is incorrect with probability $0 < q = 1 - p < 1$.
- **Transforming Las Vegas to Monte Carlo:** We run a timer for time T in parallel with the algorithm. If the algorithm completes before T, we output the output of the algorithm. If the timer's timer reaches time T, we abort the algorithm and output an incorrect output. Using Markov's Inequality we get that the probability of failure is $q \leq \mu/T$
- **Transforming Monte Carlo to Las Vegas:** We assume that we have an efficient way to check whether the output of the algorithm is correct. We repeatedly tune the Monte-Carlo algorithm followed by the checker algorithm until we find a correct output. Suppose the time for one iteration is T, and the probability of success in each iteration is p, then the expected running time until completion is T/p.
- Hashing n elements into a table of size n. The location of each element is distributed uniformly over the n bins, independent of the location of the other elements. The *occupancy* of a bin is the number of elements that it holds. The probability that *"the maximal occupancy (over all bins) is larger than* $\log(n)$*"* goes to zero as n goes to infinity.
- **The power of two:** One very effective method for reducing the maximal occupancy is to use two hash functions instead of one. To add a new element to the table both locations are checked and the new element is added to the location with smaller occupancy. The probability that *"the maximal occupancy is larger than* $\log\log(n)$*"* goes to zero as n goes to infinity.
- **Min-Hash:** Here we are concerned with comparing document. We view each document as a set of words. The Jaccard similarity between two documents A, B, is defined as

$$S(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

  Where A denotes the set words in the document A (without repetitions) and $|A|$ is the number of elements in that set.

  The min-hash method associates with each document a short "signature" so that the similarity between any two documents can be approximated efficiently from their signatures alone.

  The min-hash signature consists of k integers from a very large range (much larger that the set of possible words). Each of the k numbers is computed by using an independent hash function h_i. Each word w is mapped to it's hash value $h_i(w)$. A document is then mapped to the minimum of the hash values of it's words. This gives the ith min-hash value for the document.

  The *probability* that the ith min-hash values of two documents A, B match is equal to the similarity $S(A, B)$. Thus the random variables X_i which are 1 for match, 0 for no-match, are IID binary RV with expected value $S(A, B)$. Using this fact we can compute the minimal value of k required to reach a specified level of accuracy in estimating $S(A, B)$.
- **Bloom filters:** A method for determining whether a given item has been observed in the past. The method consists of a binary vector T of length m which is initialized to all zeros, and k hash functions that map items into the range $1, \ldots, m$. An item x is mapped to the k values $h_1(x), \ldots, h_k(x)$. The corresponding entries in T are checked, if all of these entries are 1 the item is declared to have been observed in the past, then all of the entries are set to 1.
- Bloom filters do not make false negative mistakes - declaring an item to be new when it is not.
- Suppose we have already entered n elements into the filter, The probability of a false positive

(declaring the $n+1$th item as not new when in fact it is new) is $p \approx \left(1 - e^{-kn/m}\right)^k$.

- The optimal choice of k for given values of m, n is $k = \frac{m}{n} \ln 2$
- If we want to have probability at most p of a false positive, we are given the value of n and use the optimal choice for k then the minimal table size we need is: $m \geq (n \ln(1/p))/(\ln 2)^2$