Probability Cheatsheet v1.1.1

Compiled by William Chen (http://wzchen.com) with contributions from Sebastian Chiu, Yuan Jiang, Yuqi Hou, and Jessy Hwang. Material based off of Joe Blitzstein's (@stat110) lectures (http://stat110.net) and Blitzstein/Hwang's Intro to Probability textbook (http://bit.ly/introprobability). Please share comments, suggestions, and errors at http://github.com/wzchen/probability_cheatsheet.

Last Updated April 22, 2015

Counting

Multiplication Rule - Let's say we have a compound experiment (an experiment with multiple components). If the 1st component has n_1 possible outcomes, the 2nd component has n_2 possible outcomes, and the rth component has n_r possible outcomes, then overall there are $n_1 n_2 \ldots n_r$ possibilities for the whole experiment.

Sampling Table - The sampling tables describes the different ways to take a sample of size k out of a population of size n. The column names denote whether order matters or not.

	Matters	**Not Matter**
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Naïve Definition of Probability - *If the likelihood of each outcome is equal*, the probability of any event happening is:

$$P(\text{Event}) = \frac{\text{number of favorable outcomes}}{\text{number of outcomes}}$$

Probability and Thinking Conditionally

Independence

Independent Events - $\mathbf{A}$ and $\mathbf{B}$ are independent if knowing one gives you no information about the other. $\mathbf{A}$ and $\mathbf{B}$ are independent if and only if one of the following equivalent statements hold:

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

Conditional Independence - $\mathbf{A}$ and $\mathbf{B}$ are conditionally independent given $\mathbf{C}$ if: $P(\mathbf{A} \cap \mathbf{B}|\mathbf{C}) = P(\mathbf{A}|\mathbf{C})P(\mathbf{B}|\mathbf{C})$. Conditional independence does not imply independence, and independence does not imply conditional independence.

Unions, Intersections, and Complements

De Morgan's Laws - Gives a useful relation that can make calculating probabilities of unions easier by relating them to intersections, and vice versa. De Morgan's Law says that the complement is distributive as long as you flip the sign in the middle.

$$(\mathbf{A} \cup \mathbf{B})^c \equiv \mathbf{A^c} \cap \mathbf{B^c}$$
$$(\mathbf{A} \cap \mathbf{B})^c \equiv \mathbf{A^c} \cup \mathbf{B^c}$$

Joint, Marginal, and Conditional Probabilities

Joint Probability - $P(\mathbf{A} \cap \mathbf{B})$ or $P(\mathbf{A}, \mathbf{B})$ - Probability of $\mathbf{A}$ *and* $\mathbf{B}$.

Marginal (Unconditional) Probability - $P(\mathbf{A})$ - Probability of $\mathbf{A}$

Conditional Probability - $P(\mathbf{A}|\mathbf{B})$ - Probability of $\mathbf{A}$ given $\mathbf{B}$ occurred.

Conditional Probability is Probability - $P(\mathbf{A}|\mathbf{B})$ is a probability as well, restricting the sample space to $\mathbf{B}$ instead of Ω. Any theorem that holds for probability also holds for conditional probability.

Simpson's Paradox

$$P(A \mid B, C) < P(A \mid B^c, C) \text{ and } P(A \mid B, C^c) < P(A \mid B^c, C^c)$$
$$\text{yet still, } P(A \mid B) > P(A \mid B^c)$$

Bayes' Rule and Law of Total Probability

Law of Total Probability with partitioning set $\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3, \ldots \mathbf{B}_n$ and with extra conditioning (just add C!)

$$P(\mathbf{A}) = P(\mathbf{A}|\mathbf{B}_1)P(\mathbf{B}_1) + P(\mathbf{A}|\mathbf{B}_2)P(\mathbf{B}_2) + \ldots P(\mathbf{A}|\mathbf{B}_n)P(\mathbf{B}_n)$$
$$P(\mathbf{A}) = P(\mathbf{A} \cap \mathbf{B}_1) + P(\mathbf{A} \cap \mathbf{B}_2) + \ldots P(\mathbf{A} \cap \mathbf{B}_n)$$
$$P(\mathbf{A}|\mathbf{C}) = P(\mathbf{A}|\mathbf{B}_1, \mathbf{C})P(\mathbf{B}_1|\mathbf{C}) + \ldots P(\mathbf{A}|\mathbf{B}_n, \mathbf{C})P(\mathbf{B}_n|\mathbf{C})$$
$$P(\mathbf{A}|\mathbf{C}) = P(\mathbf{A} \cap \mathbf{B}_1|\mathbf{C}) + P(\mathbf{A} \cap \mathbf{B}_2|\mathbf{C}) + \ldots P(\mathbf{A} \cap \mathbf{B}_n|\mathbf{C})$$

Law of Total Probability with $\mathbf{B}$ and $\mathbf{B^c}$ (special case of a partitioning set), and with extra conditioning (just add C!)

$$P(\mathbf{A}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) + P(\mathbf{A}|\mathbf{B^c})P(\mathbf{B^c})$$
$$P(\mathbf{A}) = P(\mathbf{A} \cap \mathbf{B}) + P(\mathbf{A} \cap \mathbf{B^c})$$
$$P(\mathbf{A}|\mathbf{C}) = P(\mathbf{A}|\mathbf{B}, \mathbf{C})P(\mathbf{B}|\mathbf{C}) + P(\mathbf{A}|\mathbf{B^c}, \mathbf{C})P(\mathbf{B^c}|\mathbf{C})$$
$$P(\mathbf{A}|\mathbf{C}) = P(\mathbf{A} \cap \mathbf{B}|\mathbf{C}) + P(\mathbf{A} \cap \mathbf{B^c}|\mathbf{C})$$

Bayes' Rule, and with extra conditioning (just add C!)

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

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

Odds Form of Bayes' Rule, and with extra conditioning (just add C!)

$$\frac{P(\mathbf{A}|\mathbf{B})}{P(\mathbf{A^c}|\mathbf{B})} = \frac{P(\mathbf{B}|\mathbf{A})}{P(\mathbf{B}|\mathbf{A^c})} \frac{P(\mathbf{A})}{P(\mathbf{A^c})}$$

$$\frac{P(\mathbf{A}|\mathbf{B}, \mathbf{C})}{P(\mathbf{A^c}|\mathbf{B}, \mathbf{C})} = \frac{P(\mathbf{B}|\mathbf{A}, \mathbf{C})}{P(\mathbf{B}|\mathbf{A^c}, \mathbf{C})} \frac{P(\mathbf{A}|\mathbf{C})}{P(\mathbf{A^c}|\mathbf{C})}$$

Random Variables and their Distributions

PMF, CDF, and Independence

Probability Mass Function (PMF) (Discrete Only) gives the probability that a random variable takes on the value x.

$$P_X(x) = P(X = x)$$

Cumulative Distribution Function (CDF) gives the probability that a random variable takes on the value x or less.

$$F_X(x_0) = P(X \leq x_0)$$

Independence - Intuitively, two random variables are independent if knowing one gives you no information about the other. X and Y are independent if for ALL values of x and y:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Expected Value and Indicators

Distributions

Probability Mass Function (PMF) (Discrete Only) is a function that takes in the value x, and gives the probability that a random variable takes on the value x. The PMF is a positive-valued function, and $\sum_x P(X = x) = 1$

$$P_X(x) = P(X = x)$$

Cumulative Distribution Function (CDF) is a function that takes in the value x, and gives the probability that a random variable takes on the value at most x.

$$F(x) = P(X \leq x)$$

Expected Value, Linearity, and Symmetry

Expected Value (aka *mean*, *expectation*, or *average*) can be thought of as the "weighted average" of the possible outcomes of our random variable. Mathematically, if $x_1, x_2, x_3, \ldots$ are all of the possible values that X can take, the expected value of X can be calculated as follows:

$$E(X) = \sum_i x_i P(X = x_i)$$

Note that for *any* X and Y, a and b scaling coefficients and c is our constant, the following property of **Linearity of Expectation** holds:

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

If two Random Variables have the same distribution, *even when they are dependent* by the property of **Symmetry** their expected values are equal.

Conditional Expected Value is calculated like expectation, only conditioned on any event A.

$$E(X|A) = \sum_x x P(X = x|A)$$

Indicator Random Variables

Indicator Random Variables is random variable that takes on either 1 or 0. The indicator is always an indicator of some event. If the event occurs, the indicator is 1, otherwise it is 0. They are useful for many problems that involve counting and expected value.

Distribution $I_A \sim \text{Bern}(p)$ where $p = P(A)$

Fundamental Bridge The expectation of an indicator for A is the probability of the event. $E(I_A) = P(A)$. Notation:

$$I_A = \begin{cases} 1 & \text{A occurs} \\ 0 & \text{A does not occur} \end{cases}$$

Variance

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

Expectation and Independence

If X and Y are independent, then

$$E(XY) = E(X)E(Y)$$

Continuous RVs, LotUS, and UoU

Continuous Random Variables

What's the prob that a CRV is in an interval? Use the CDF (or the PDF, see below). To find the probability that a CRV takes on a value in the interval $[a, b]$, subtract the respective CDFs.

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

Note that for an r.v. with a normal distribution,

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$$
$$= \Phi\left(\frac{b-\mu}{\sigma^2}\right) - \Phi\left(\frac{a-\mu}{\sigma^2}\right)$$

What is the Cumulative Density Function (CDF)? It is the following function of x.

$$F(x) = P(X \leq x)$$

What is the Probability Density Function (PDF)? The PDF, $f(x)$, is the derivative of the CDF.

$$F'(x) = f(x)$$

Or alternatively,

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Note that by the fundamental theorem of calculus,

$$F(b) - F(a) = \int_a^b f(x)dx$$

Thus to find the probability that a CRV takes on a value in an interval, you can integrate the PDF, thus finding the area under the density curve.

How do I find the expected value of a CRV? Where in discrete cases you sum over the probabilities, in continuous cases you integrate over the densities.

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Law of the Unconscious Statistician (LotUS)

Expected Value of Function of RV Normally, you would find the expected value of X this way:

$$E(X) = \Sigma_x xP(X = x)$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

LotUS states that you can find the expected value of a *function of a random variable* g(X) this way:

$$E(g(X)) = \Sigma_x g(x)P(X = x)$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

What's a function of a random variable? A function of a random variable is also a random variable. For example, if X is the number of bikes you see in an hour, then $g(X) = 2X$ could be the number of bike wheels you see in an hour. Both are random variables.

What's the point? You don't need to know the PDF/PMF of $g(X)$ to find its expected value. All you need is the PDF/PMF of X.

Universality of Uniform

When you plug any random variable into its own CDF, you get a Uniform[0,1] random variable. When you put a Uniform[0,1] into an inverse CDF, you get the corresponding random variable. For example, let's say that a random variable X has a CDF

$$F(x) = 1 - e^{-x}$$

By the Universality of the the Uniform, if we plug in X into this function then we get a uniformly distributed random variable.

$$F(X) = 1 - e^{-X} \sim U$$

Similarly, since $F(X) \sim U$ then $X \sim F^{-1}(U)$. The key point is that *for any continuous random variable X, we can transform it into a uniform random variable and back by using its CDF.*

Moment Generating Functions (MGFs)

Moments

Moments describe the shape of a distribution. The k^{th} moment of a random variable X is

$$\mu_k' = E(X^k)$$

The mean, variance, and skewness of a distribution can be expressed by its moments. Specifically:

Mean $E(X) = \mu_1'$

Variance $\text{Var}(X) = E(X^2) - E(X)^2 = \mu_2' - (\mu_1')^2$

Moment Generating Functions

MGF For any random variable X, this expected value and function of dummy variable t;

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

is the **moment generating function (MGF)** of X if it exists for a finitely-sized interval centered around 0. Note that the MGF is just a function of a dummy variable t.

Why is it called the Moment Generating Function? Because the k^{th} derivative of the moment generating function evaluated 0 is the k^{th} moment of X!

$$\mu_k' = E(X^k) = M_X^{(k)}(0)$$

This is true by Taylor Expansion of e^{tX}

$$M_X(t) = E(e^{tX}) = \sum_{k=0}^{\infty} \frac{E(X^k)t^k}{k!} = \sum_{k=0}^{\infty} \frac{\mu_k' t^k}{k!}$$

Or by differentiation under the integral sign and then plugging in $t = 0$

$$M_X^{(k)}(t) = \frac{d^k}{dt^k}E(e^{tX}) = E\left(\frac{d^k}{dt^k}e^{tX}\right) = E(X^k e^{tX})$$

$$M_X^{(k)}(0) = E(X^k e^{0X}) = E(X^k) = \mu_k'$$

MGF of linear combinations If we have $Y = aX + c$, then

$$M_Y(t) = E(e^{t(aX+c)}) = e^{ct}E(e^{(at)X}) = e^{ct}M_X(at)$$

Uniqueness of the MGF. *If it exists, the MGF uniquely defines the distribution.* This means that for any two random variables X and Y, they are distributed the same (their CDFs/PDFs are equal) if and only if their MGF's are equal. You can't have different PDFs when you have two random variables that have the same MGF.

Summing Independent R.V.s by Multiplying MGFs. If X and Y are independent, then

$$M_{(X+Y)}(t) = E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY}) = M_X(t) \cdot M_Y(t)$$

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

The MGF of the sum of two random variables is the product of the MGFs of those two random variables.

Joint PDFs and CDFs

Joint Distributions

Review: Joint Probability of events A and B: $P(A \cap B)$
Both the Joint PMF and Joint PDF must be non-negative and sum/integrate to 1. ($\sum_x \sum_y P(X = x, Y = y) = 1$)
($\int_x \int_y f_{X,Y}(x, y) = 1$). Like in the univariate cause, you sum/integrate the PMF/PDF to get the CDF.

Conditional Distributions

Review: By Baye's Rule, $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ Similar conditions apply to conditional distributions of random variables.
For discrete random variables:

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

For continuous random variables:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$$

Hybrid Bayes' Rule

$$f(x|A) = \frac{P(A|X = x)f(x)}{P(A)}$$

Marginal Distributions

Review: Law of Total Probability Says for an event A and partition $B_1, B_2, ...B_n$: $P(A) = \sum_i P(A \cap B_i)$
To find the distribution of one (or more) random variables from a joint distribution, sum or integrate over the irrelevant random variables.
Getting the Marginal PMF from the Joint PMF

$$P(X = x) = \sum_y P(X = x, Y = y)$$

Getting the Marginal PDF from the Joint PDF

$$f_X(x) = \int_y f_{X,Y}(x, y)dy$$

Independence of Random Variables

Review: A and B are independent if and only if either $P(A \cap B) = P(A)P(B)$ or $P(A|B) = P(A)$.
Similar conditions apply to determine whether random variables are independent - two random variables are independent if their joint distribution function is simply the product of their marginal distributions, or that the a conditional distribution of is the same as its marginal distribution.
In words, random variables X and Y are independent for all x, y, if and only if one of the following hold:

- Joint PMF/PDF/CDFs are the product of the Marginal PMF
- Conditional distribution of X given Y is the same as the marginal distribution of X

Multivariate LotUS

Review: $E(g(X)) = \sum_x g(x)P(X = x)$, or $E(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$
For discrete random variables:

$$E(g(X, Y)) = \sum_x \sum_y g(x, y)P(X = x, Y = y)$$

For continuous random variables:

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

Covariance and Transformations

Covariance and Correlation

Covariance is the two-random-variable equivalent of Variance, defined by the following:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

Note that

$$\text{Cov}(X, X) = E(XX) - E(X)E(X) = \text{Var}(X)$$

Correlation is a rescaled variant of Covariance that is always between -1 and 1.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Covariance and Indepedence - If two random variables are independent, then they are uncorrelated. The inverse is not necessarily true.

$$X \perp\!\!\!\perp Y \longrightarrow \text{Cov}(X, Y) = 0$$

$$X \perp\!\!\!\perp Y \longrightarrow E(XY) = E(X)E(Y)$$

Covariance and Variance - Note that

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j)$$

In particular, if X and Y are independent then they have covariance 0 thus

$$X \perp\!\!\!\perp Y \implies \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

In particular, If $X_1, X_2, \ldots, X_n$ are identically distributed have the same covariance relationships, then

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = n\text{Var}(X_1) + 2\binom{n}{2}\text{Cov}(X_1, X_2)$$

Covariance and Linearity - For random variables W, X, Y, Z and constants a, b:

$$\begin{aligned}
\text{Cov}(X, Y) &= \text{Cov}(Y, X) \\
\text{Cov}(X + a, Y + b) &= \text{Cov}(X, Y) \\
\text{Cov}(aX, bY) &= ab\text{Cov}(X, Y) \\
\text{Cov}(W + X, Y + Z) &= \text{Cov}(W, Y) + \text{Cov}(W, Z) + \text{Cov}(X, Y) \\
&\quad + \text{Cov}(X, Z)
\end{aligned}$$

Covariance and Invariance - Correlation, Covariance, and Variance are addition-invariant, which means that adding a constant to the term(s) does not change the value. Let b and c be constants.

$$\text{Var}(X+c) = \text{Var}(X)$$
$$\text{Cov}(X+b, Y+c) = \text{Cov}(X,Y)$$
$$\text{Corr}(X+b, Y+c) = \text{Corr}(X,Y)$$

In addition to addition-invariance, Correlation is *scale-invariant*, which means that multiplying the terms by any constant does not affect the value. Covariance and Variance are not scale-invariant.

$$\text{Corr}(2X, 3Y) = \text{Corr}(X,Y)$$

Continuous Transformations

Why do we need the Jacobian? We need the Jacobian to rescale our PDF so that it integrates to 1.

One Variable Transformations Let's say that we have a random variable X with PDF $f_X(x)$, but we are also interested in some function of X. We call this function $Y = g(X)$. Note that Y is a random variable as well. If g is differentiable and one-to-one (every value of X gets mapped to a unique value of Y), then the following is true:

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| \qquad f_Y(y)\left|\frac{dy}{dx}\right| = f_X(x)$$

To find $f_Y(y)$ as a function of y, plug in $x = g^{-1}(y)$.

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$$

The derivative of the inverse transformation is referred to the **Jacobian**, denoted as J.

$$J = \frac{d}{dy}g^{-1}(y)$$

Convolutions

Definition If you want to find the PDF of a sum of two independent random variables, you take the convolution of their individual distributions.

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_x(x) f_y(t-x)dx$$

Example Let $X, Y \sim$ i.i.d $N(0,1)$. Treat t as a constant. Integrate as usual.

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-(t-x)^2/2} dx$$

Poisson Processes and Order Statistics

Poisson Process

Definition We have a Poisson Process if we have

1. Arrivals at various times with an average of λ per unit time.
2. The number of arrivals in a time interval of length t is Pois(λt)
3. Number of arrivals in disjoint time intervals are independent.

Count-Time Duality - We wish to find the distribution of T_1, the first arrival time. We see that the event $T_1 > t$, the event that you have to wait more than t to get the first email, is the same as the event $N_t = 0$, which is the event that the number of emails in the first time interval of length t is 0. We can solve for the distribution of T_1.

$$P(T_1 > t) = P(N_t = 0) = e^{-\lambda t} \longrightarrow P(T_1 \leq t) = 1 - e^{-\lambda t}$$

Thus we have $T_1 \sim \text{Expo}(\lambda)$. And similarly, the interarrival times between arrivals are all Expo(λ), (e.g. $T_i - T_{i-1} \sim \text{Expo}(\lambda)$).

Order Statistics

Definition - Let's say you have n i.i.d. random variables $X_1, X_2, X_3, \ldots X_n$. If you arrange them from smallest to largest, the ith element in that list is the ith order statistic, denoted $X_{(i)}$. $X_{(1)}$ is the smallest out of the set of random variables, and $X_{(n)}$ is the largest.

Properties - The order statistics are dependent random variables. The smallest value in a set of random variables will always vary and itself has a distribution. For any value of $X_{(i)}$, $X_{(i+1)} \geq X_{(j)}$.

Distribution - Taking n i.i.d. random variables $X_1, X_2, X_3, \ldots X_n$ with CDF $F(x)$ and PDF $f(x)$, the CDF and PDF of $X_{(i)}$ are as follows:

$$F_{X_{(i)}}(x) = P(X_{(j)} \leq x) = \sum_{k=i}^{n} \binom{n}{k} F(x)^k (1-F(x))^{n-k}$$

$$f_{X_{(i)}}(x) = n\binom{n-1}{i-1} F(x)^{i-1}(1-F(X))^{n-i} f(x)$$

Universality of the Uniform - We can also express the distribution of the order statistics of n i.i.d. random variables $X_1, X_2, X_3, \ldots X_n$ in terms of the order statistics of n uniforms. We have that

$$F(X_{(j)}) \sim U_{(j)}$$

Beta Distribution as Order Statistics of Uniform - The smallest of three Uniforms is distributed $U_{(1)} \sim \text{Beta}(1,3)$. The middle of three Uniforms is distributed $U_{(2)} \sim \text{Beta}(2,2)$, and the largest $U_{(3)} \sim \text{Beta}(3,1)$. The distribution of the the j^{th} order statistic of n i.i.d Uniforms is:

$$U_{(j)} \sim \text{Beta}(j, n-j+1)$$
$$f_{U_{(j)}}(u) = \frac{n!}{(j-1)!(n-j)!} t^{j-1}(1-t)^{n-j}$$

Conditional Expectation and Variance

Conditional Expectation

Conditioning on an Event - We can find the expected value of Y given that event A or $X = x$ has occurred. This would be finding the values of $E(Y|A)$ and $E(Y|X=x)$. Note that conditioning in an event results in a *number*. Note the similarities between regularly finding expectation and finding the conditional expectation. The expected value of a dice roll given that it is prime is $\frac{1}{3}2 + \frac{1}{3}3 + \frac{1}{3}5 = 3\frac{1}{3}$. The expected amount of time that you have to wait until the shuttle comes (assuming that the waiting time is $\sim \text{Expo}(\frac{1}{10})$) given that you have already waited n minutes, is 10 more minutes by the memoryless property.

Discrete Y	Continuous Y
$E(Y) = \sum_y yP(Y=y)$	$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy$
$E(Y\|X=x) = \sum_y yP(Y=y\|X=x)$	$E(Y\|X=x) = \int_{-\infty}^{\infty} y f_{Y\|X}(y\|x)dy$
$E(Y\|A) = \sum_y yP(Y=y\|A)$	$E(Y\|A) = \int_{-\infty}^{\infty} y f(y\|A)dy$

Conditioning on a Random Variable - We can also find the expected value of Y given the random variable X. The resulting expectation, $E(Y|X)$ is *not a number but a function of the random variable* X. For an easy way to find $E(Y|X)$, find $E(Y|X=x)$ and then plug in X for all x. This changes the conditional expectation of Y from a function of a number x, to a function of the random variable X.

Properties of Conditioning on Random Variables

1. $E(Y|X) = E(Y)$ if $X \perp\!\!\!\perp Y$
2. $E(h(X)|X) = h(X)$ (taking out what's known). $E(h(X)W|X) = h(X)E(W|X)$
3. $E(E(Y|X)) = E(Y)$ (**Adam's Law**, aka Law of Iterated Expectation of Law of Total Expectation)

Law of Total Expectation (also Adam's law) - For any set of events that partition the sample space, $A_1, A_2, \ldots, A_n$ or just simply A, A^c, the following holds:

$$E(Y) = E(Y|A)P(A) + E(Y|A^c)P(A^c)$$
$$E(Y) = E(Y|A_1)P(A_1) + \cdots + E(Y|A_n)P(A_n)$$

Conditional Variance

Eve's Law (aka Law of Total Variance)

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))$$

MVN, LLN, CLT

Law of Large Numbers (LLN)

Let us have $X_1, X_2, X_3 \ldots$ be i.i.d.. We define $\bar{X}_n = \frac{X_1+X_2+X_3+\cdots+X_n}{n}$ The Law of Large Numbers states that as $n \longrightarrow \infty$, $\bar{X}_n \longrightarrow E(X)$.

Central Limit Theorem (CLT)

Approximation using CLT

We use $\dot{\sim}$ to denote *is approximately distributed*. We can use the central limit theorem when we have a random variable, Y that is a sum of n i.i.d. random variables with n large. Let us say that $E(Y) = \mu_Y$ and $\text{Var}(Y) = \sigma_Y^2$. We have that:

$$Y \dot{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$$

When we use central limit theorem to estimate Y, we usually have $Y = X_1 + X_2 + \cdots + X_n$ or $Y = \bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. Specifically, if we say that each of the iid X_i have mean μ_X and σ_X^2, then we have the following approximations.

$$X_1 + X_2 + \cdots + X_n \dot{\sim} \mathcal{N}(n\mu_X, n\sigma_X^2)$$
$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \dot{\sim} \mathcal{N}(\mu_X, \frac{\sigma_X^2}{n})$$

Asymptotic Distributions using CLT

We use $\xrightarrow{d}$ to denote *converges in distribution to* as $n \longrightarrow \infty$. These are the same results as the previous section, only letting $n \longrightarrow \infty$ and not letting our normal distribution have any n terms.

$$\frac{1}{\sigma\sqrt{n}}(X_1 + \cdots + X_n - n\mu_X) \xrightarrow{d} \mathcal{N}(0,1)$$

$$\frac{\bar{X}_n - \mu_X}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Markov Chains

Definition

A Markov Chain is a walk along a (finite or infinite, but for this class usually finite) discrete **state space** {1, 2, ..., M}. We let X_t denote which element of the state space the walk is on at time t. The Markov Chain is the set of random variables denoting where the walk is at all points in time, $\{X_0, X_1, X_2, \ldots\}$, as long as if you want to predict where the chain is at at a future time, you only need to use the present state, and not any past information. In other words, the *given the present, the future and past are conditionally independent.* Formal Definition:

$$P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \ldots, X_n = i) = P(X_{n+1} = j | X_n = i)$$

State Properties

A state is either recurrent or transient.

- If you start at a **Recurrent State**, then you will always return back to that state at some point in the future. ♪*You can check-out any time you like, but you can never leave.* ♪
- Otherwise you are at a **Transient State**. There is some probability that once you leave you will never return. ♪*You don't have to go home, but you can't stay here.* ♪

A state is either periodic or aperiodic.

- If you start at a **Periodic State** of period k, then the GCD of all of the possible number steps it would take to return back is > 1.
- Otherwise you are at an **Aperiodic State.** The GCD of all of the possible number of steps it would take to return back is 1.

Transition Matrix

Element q_{ij} in square transition matrix Q is the probability that the chain goes from state i to state j, or more formally:

$$q_{ij} = P(X_{n+1} = j | X_n = i)$$

To find the probability that the chain goes from state i to state j in m steps, take the $(i,j)^{\text{th}}$ element of Q^m.

$$q_{ij}^{(m)} = P(X_{n+m} = j | X_n = i)$$

If X_0 is distributed according to row-vector PMF $\vec{p}$ (e.g. $p_j = P(X_0 = i_j)$), then the PMF of X_n is $\vec{p}Q^n$.

Chain Properties

A chain is **irreducible** if you can get from anywhere to anywhere. An irreducible chain must have all of its states recurrent. A chain is **periodic** if any of its states are periodic, and is **aperiodic** if none of its states are periodic. In an irreducible chain, all states have the same period.
A chain is **reversible** with respect to $\vec{s}$ if $s_i q_{ij} = s_j q_{ji}$ for all i, j. A reversible chain running on $\vec{s}$ is indistinguishable whether it is running forwards in time or backwards in time. Examples of reversible chains include random walks on undirected networks, or any chain with $q_{ij} = q_{ji}$, where the Markov chain would be stationary with respect to $\vec{s} = (\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M})$.
Reversibility Condition Implies Stationarity - If you have a PMF $\vec{s}$ on a Markov chain with transition matrix Q, then $s_i q_{ij} = s_j q_{ji}$ for all i, j implies that s is stationary.

Stationary Distribution

Let us say that the vector $\vec{p} = (p_1, p_2, \ldots, p_M)$ is a possible and valid PMF of where the Markov Chain is at at a certain time. We will call this vector the stationary distribution, $\vec{s}$, if it satisfies $\vec{s}Q = \vec{s}$. As a consequence, if X_t has the stationary distribution, then all future $X_{t+1}, X_{t+2}, \ldots$ also has the stationary distribution.
For irreducible, aperiodic chains, the stationary distribution exists, is unique, and s_i is the long-run probability of a chain being at state i. The expected number of steps to return back to i starting from i is $1/s_i$ To solve for the stationary distribution, you can solve for $(Q' - I)(\vec{s})' = 0$. The stationary distribution is uniform if the columns of Q sum to 1.

Random Walk on Undirected Network

If you have a certain number of nodes with edges between them, and a chain can pick any edge randomly and move to another node, then this is a random walk on an undirected network. The stationary distribution of this chain is proportional to the **degree sequence**. The **degree sequence** is the vector of the degrees of each node, defined as how many edges it has.

Continuous Distributions

Uniform

Let us say that U is distributed Unif(a, b). We know the following:

Properties of the Uniform For a uniform distribution, the probability of an draw from any interval on the uniform is proportion to the length of the uniform. The PDF of a Uniform is just a constant, so when you integrate over the PDF, you will get an area proportional to the length of the interval.

Example William throws darts really badly, so his darts are uniform over the whole room because they're equally likely to appear anywhere. William's darts have a uniform distribution on the surface of the room. The uniform is the only distribution where the probably of hitting in any specific region is proportion to the area/length/volume of that region, and where the density of occurrence in any one specific spot is constant throughout the whole support.

Normal

Let us say that X is distributed $\mathcal{N}(\mu, \sigma^2)$. We know the following:

Central Limit Theorem The Normal distribution is ubiquitous because of the central limit theorem, which states that averages of independent identically-distributed variables will approach a normal distribution regardless of the initial distribution.

Transformable Every time we stretch or scale the normal distribution, we change it to another normal distribution. If we add c to a normally distributed random variable, then its mean increases additively by c. If we multiply a normally distributed random variable by c, then its variance increases multiplicatively by c^2. Note that for every normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, we can transform it to the standard $\mathcal{N}(0, 1)$ by the following transformation:

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Example Heights are normal. Measurement error is normal. By the central limit theorem, the sampling average from a population is also normal.

Standard Normal - The Standard Normal, denoted Z, is $Z \sim \mathcal{N}(0, 1)$

CDF - It's too difficult to write this one out, so we express it as the function $\Phi(x)$

Exponential Distribution

Let us say that X is distributed Expo(λ). We know the following:

Story You're sitting on an open meadow right before the break of dawn, wishing that airplanes in the night sky were shooting stars, because you could really use a wish right now. You know that shooting stars come on average every 15 minutes, but it's never true that a shooting star is ever ''due" to come because you've waited so long. Your waiting time is memorylessness, which means that the time until the next shooting star comes does not depend on how long you've waited already.

Example The waiting time until the next shooting star is distributed Expo(4). The 4 here is λ, or the rate parameter, or how many shooting stars we expect to see in a unit of time. The expected time until the next shooting star is $\frac{1}{\lambda}$, or $\frac{1}{4}$ of an hour. You can expect to wait 15 minutes until the next shooting star.

Expos are rescaled Expos

$$Y \sim \text{Expo}(\lambda) \to X = \lambda Y \sim \text{Expo}(1)$$

Memorylessness The Exponential Distribution is the sole continuous memoryless distribution. This means that it's always "as good as new", which means that the probability of it failing in the next infinitesimal time period is the same as any infinitesimal time period. This means that for an exponentially distributed X and any real numbers t and s,

$$P(X > s + t | X > s) = P(X > t)$$

Given that you've waited already at least s minutes, the probability of having to wait an additional t minutes is the same as the probability that you have to wait more than t minutes to begin with. Here's another formulation.

$$X - a | X > a \sim \text{Expo}(\lambda)$$

Example - If waiting for the bus is distributed exponentially with $\lambda = 6$, no matter how long you've waited so far, the expected additional waiting time until the bus arrives is always $\frac{1}{6}$, or 10 minutes. The distribution of time from now to the arrival is always the same, no matter how long you've waited.

Min of Expos If we have independent $X_i \sim \text{Expo}(\lambda_i)$, then $\min(X_1, \ldots, X_k) \sim \text{Expo}(\lambda_1 + \lambda_2 + \cdots + \lambda_k)$.

Max of Expos If we have i.i.d. $X_i \sim \text{Expo}(\lambda)$, then $\max(X_1, \ldots, X_k) \sim \text{Expo}(k\lambda) + \text{Expo}((k-1)\lambda) + \cdots + \text{Expo}(\lambda)$

Gamma Distribution

Let us say that X is distributed Gamma(a, λ). We know the following:

Story You sit waiting for shooting stars, and you know that the waiting time for a star is distributed Expo(λ). You want to see "a" shooting stars before you go home. X is the total waiting time for the ath shooting star.

Example You are at a bank, and there are 3 people ahead of you. The serving time for each person is distributed Exponentially with mean of 2 time units. The distribution of your waiting time until you begin service is Gamma$(3, \frac{1}{2})$

Beta Distribution

Conjugate Prior of the Binomial A prior is the distribution of a parameter before you observe any data ($f(x)$). A posterior is the distribution of a parameter after you observe data y ($f(x|y)$). Beta is the *conjugate* prior of the Binomial because if you have a Beta-distributed prior on p (the parameter of the Binomial), then the posterior distribution on p given observed data is also Beta-distributed. This means, that in a two-level model:

$$X | p \sim \text{Bin}(n, p)$$
$$p \sim \text{Beta}(a, b)$$

Then after observing the value $X = x$, we get a posterior distribution $p|(X = x) \sim \text{Beta}(a + x, b + n - x)$

Order statistics of the Uniform See *Order Statistics*

Relationship with Gamma This is the bank-post office result. See *Reasoning by Representation*

χ^2 Distribution

Let us say that X is distributed χ^2_n. We know the following:

Story A Chi-Squared(n) is a sum of n independent squared normals.

Example The sum of squared errors are distributed χ^2_n

Properties and Representations

$$E(\chi^2_n) = n, Var(X) = 2n, \chi^2_n \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$$

$$\chi^2_n = Z_1^2 + Z_2^2 + \cdots + Z_n^2, Z \sim^{i.i.d.} \mathcal{N}(0, 1)$$

Discrete Distributions

DWR = Draw w/ replacement, DWoR = Draw w/o replacement

	DWR	**DWoR**
Fixed # trials (n)	Binom/Bern (Bern if $n = 1$)	HGeom
Draw 'til k success	NBin/Geom (Geom if $k = 1$)	NHGeom (see example probs)

Bernoulli

The Bernoulli distribution is the simplest case of the Binomial distribution, where we only have one trial, or $n = 1$. Let us say that X is distributed Bern(p). We know the following:

Story. X "succeeds" (is 1) with probability p, and X "fails" (is 0) with probability $1 - p$.

Example. A fair coin flip is distributed Bern$(\frac{1}{2})$.

Binomial

Let us say that X is distributed $\text{Bin}(n, p)$. We know the following:

Story X is the number of "successes" that we will achieve in n independent trials, where each trial can be either a success or a failure, each with the same probability p of success. We can also say that X is a sum of multiple independent $Bern(p)$ random variables. Let $X \sim \text{Bin}(n, p)$ and $X_j \sim \text{Bern}(p)$, where all of the Bernoullis are independent. We can express the following:

$$X = X_1 + X_2 + X_3 + \cdots + X_n$$

Example If Jeremy Lin makes 10 free throws and each one independently has a $\frac{3}{4}$ chance of getting in, then the number of free throws he makes is distributed $\text{Bin}(10, \frac{3}{4})$, or, letting X be the number of free throws that he makes, X is a Binomial Random Variable distributed $\text{Bin}(10, \frac{3}{4})$.

Binomial Coefficient $\binom{n}{k}$ is a function of n and k and is read n *choose* k, and means out of n possible indistinguishable objects, how many ways can I possibly choose k of them? The formula for the binomial coefficient is:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Geometric

Let us say that X is distributed $\text{Geom}(p)$. We know the following:

Story X is the number of "failures" that we will achieve before we achieve our first success. Our successes have probability p.

Example If each pokeball we throw has a $\frac{1}{10}$ probability to catch Mew, the number of failed pokeballs will be distributed $\text{Geom}(\frac{1}{10})$.

First Success

Equivalent to the geometric distribution, except it counts the total number of "draws" until the first success. This is 1 more than the number of failures. If $X \sim FS(p)$ then $E(X) = 1/p$.

Negative Binomial

Let us say that X is distributed $\text{NBin}(r, p)$. We know the following:

Story X is the number of "failures" that we will achieve before we achieve our rth success. Our successes have probability p.

Example Thundershock has 60% accuracy and can faint a wild Raticate in 3 hits. The number of misses before Pikachu faints Raticate with Thundershock is distributed $\text{NBin}(3, .6)$.

Hypergeometric

Let us say that X is distributed $\text{HGeom}(w, b, n)$. We know the following:

Story In a population of b undesired objects and w desired objects, X is the number of "successes" we will have in a draw of n objects, without replacement.

Example 1) Let's say that we have only b Weedles (failure) and w Pikachus (success) in Viridian Forest. We encounter n Pokemon in the forest, and X is the number of Pikachus in our encounters. 2) The number of aces that you draw in 5 cards (without replacement). 3) You have w white balls and b black balls, and you draw n balls. You will draw X white balls. 4) Elk Problem - You have N elk, you capture n of them, tag them, and release them. Then you recollect a new sample of size m. How many tagged elk are now in the new sample?

PMF The probability mass function of a Hypergeometric:

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$$

Poisson

Let us say that X is distributed $\text{Pois}(\lambda)$. We know the following:

Story There are rare events (low probability events) that occur many different ways (high possibilities of occurences) at an average rate of λ occurrences per unit space or time. The number of events that occur in that unit of space or time is X.

Example A certain busy intersection has an average of 2 accidents per month. Since an accident is a low probability event that can happen many different ways, the number of accidents in a month at that intersection is distributed $\text{Pois}(2)$. The number of accidents that happen in two months at that intersection is distributed $\text{Pois}(4)$

Multivariate Distributions

Multinomial

Let us say that the vector $\vec{\mathbf{X}} = (X_1, X_2, X_3, \ldots, X_k) \sim \text{Mult}_k(n, \vec{p})$ where $\vec{p} = (p_1, p_2, \ldots, p_k)$.

Story - We have n items, and then can fall into any one of the k buckets independently with the probabilities $\vec{p} = (p_1, p_2, \ldots, p_k)$.

Example - Let us assume that every year, 100 students in the Harry Potter Universe are randomly and independently sorted into one of four houses with equal probability. The number of people in each of the houses is distributed $\text{Mult}_4(100, \vec{p})$, where $\vec{p} = (.25, .25, .25, .25)$. Note that $X_1 + X_2 + \cdots + X_4 = 100$, and they are dependent.

Multinomial Coefficient The number of permutations of n objects where you have $n_1, n_2, n_3 \ldots, n_k$ of each of the different variants is the **multinomial coefficient**.

$$\binom{n}{n_1 n_2 \ldots n_k} = \frac{n!}{n_1! n_2! \ldots n_k!}$$

Joint PMF - For $n = n_1 + n_2 + \cdots + n_k$

$$P(\vec{X} = \vec{n}) = \binom{n}{n_1 n_2 \ldots n_k} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Lumping - If you lump together multiple categories in a multinomial, then it is still multinomial. A multinomial with two dimensions (success, failure) is a binomial distribution.

Variances and Covariances - For $(X_1, X_2, \ldots, X_k) \sim \text{Mult}_k(n, (p_1, p_2, \ldots, p_k))$, we have that marginally $X_i \sim \text{Bin}(n, p_i)$ and hence $\text{Var}(X_i) = np_i(1 - p_i)$. Also, for $i \neq j$, $\text{Cov}(X_i, X_j) = -np_ip_j$, which is a result from class.

Marginal PMF and Lumping

$$X_i \sim \text{Bin}(n, p_i)$$

$$X_i + X_j \sim \text{Bin}(n, p_i + p_j)$$

$$X_1, X_2, X_3 \sim \text{Mult}_3(n, (p_1, p_2, p_3)) \to X_1, X_2 + X_3 \sim \text{Mult}_2(n, (p_1, p_2 + p_3))$$

$$X_1, \ldots, X_{k-1} | X_k = n_k \sim \text{Mult}_{k-1}\left(n - n_k, \left(\frac{p_1}{1 - p_k}, \ldots, \frac{p_{k-1}}{1 - p_k}\right)\right)$$

Multivariate Uniform

See the univariate uniform for stories and examples. For multivariate uniforms, all you need to know is that probability is proportional to volume. More formally, probability is the volume of the region of interest divided by the total volume of the support. Every point in the support has equal density of value $\frac{1}{\text{Total Area}}$.

Multivariate Normal (MVN)

A vector $\vec{X} = (X_1, X_2, X_3, \ldots, X_k)$ is declared Multivariate Normal if any linear combination is normally distributed (e.g. $t_1X_1 + t_2X_2 + \cdots + t_kX_k$ is Normal for any constants $t_1, t_2, \ldots, t_k$). The parameters of the Multivariate normal are the mean vector $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$ and the covariance matrix where the $(i, j)^{\text{th}}$ entry is $\text{Cov}(X_i, X_j)$. For any MVN distribution: 1) Any sub-vector is also MVN. 2) If any two elements of a multivariate normal distribution are uncorrelated, then they are independent. Note that 2) does not apply to most random variables.

Distribution Properties

Important CDFs

Exponential $F(X) = 1 - e^{-\lambda x}, x \in (0, \infty))$

Uniform(0, 1) $F(X) = x, x \in (0, 1)$

Poisson Properties (Chicken and Egg Results)

We have $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$ and $X \perp\!\!\!\perp Y$.

1. $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. $X|(X + Y = k) \sim \text{Bin}\left(k, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)$
3. If we have that $Z \sim \text{Pois}(\lambda)$, and we randomly and independently "accept" every item in Z with probability p, then the number of accepted items $Z_1 \sim \text{Pois}(\lambda p)$, and the number of rejected items $Z_2 \sim \text{Pois}(\lambda q)$, and $Z_1 \perp\!\!\!\perp Z_2$.

Convolutions of Random Variables

A convolution of n random variables is simply their sum.

1. $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2)$, $X \perp\!\!\!\perp Y \longrightarrow X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. $X \sim \text{Bin}(n_1, p)$, $Y \sim \text{Bin}(n_2, p)$, $X \perp\!\!\!\perp Y \longrightarrow X + Y \sim \text{Bin}(n_1 + n_2, p)$ Note that Binomial can thus be thought of as a sum of iid Bernoullis.
3. $X \sim \text{Gamma}(n_1, \lambda)$, $Y \sim \text{Gamma}(n_2, \lambda)$, $X \perp\!\!\!\perp Y \longrightarrow X + Y \sim \text{Gamma}(n_1 + n_2, \lambda)$ Note that Gamma can thus be thought of as a sum of iid Expos.
4. $X \sim \text{NBin}(r_1, p)$, $Y \sim \text{NBin}(r_2, p)$, $X \perp\!\!\!\perp Y \longrightarrow X + Y \sim \text{NBin}(r_1 + r_2, p)$
5. All of the above are approximately normal when λ, n, r are large by the Central Limit Theorem.
6. $Z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $Z_1 \perp\!\!\!\perp Z_2 \longrightarrow Z_1 + Z_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Special Cases of Random Variables

1. $\text{Bin}(1, p) \sim \text{Bern}(p)$
2. $\text{Beta}(1, 1) \sim \text{Unif}(0, 1)$
3. $\text{Gamma}(1, \lambda) \sim \text{Expo}(\lambda)$
4. $\chi^2_n \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$
5. $\text{NBin}(1, p) \sim \text{Geom}(p)$

Reasoning by Representation

Beta-Gamma relationship If $X \sim \text{Gamma}(a, \lambda)$, $Y \sim \text{Gamma}(b, \lambda)$, $X \perp\!\!\!\perp Y$ then

- $\frac{X}{X+Y} \sim \text{Beta}(a, b)$
- $X + Y \perp\!\!\!\perp \frac{X}{X+Y}$

This is also known as the **bank-post office result**.

Binomial-Poisson Relationship $\text{Bin}(n, p) \to \text{Pois}(\lambda)$ as $n \to \infty$, $p \to 0$, $np = \lambda$.

Order Statistics of Uniform $U_{(j)} \sim \text{Beta}(j, n - j + 1)$

Universality of Uniform For any X with CDF $F(x)$, $F(X) \sim U$

Formulas

In general, remember that PDFs integrated (and PMFs summed) over support equal 1.

Geometric Series

$$a + ar + ar^2 + \cdots + ar^{n-1} = \sum_{k=0}^{n-1} ar^k = a\frac{1 - r^n}{1 - r}$$

Exponential Function (e^x)

$$e^x = \sum_{n=1}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n$$

Gamma and Beta Distributions

You can sometimes solve complicated-looking integrals by pattern-matching to the following:

$$\int_0^\infty x^{t-1}e^{-x}\,dx = \Gamma(t) \qquad \int_0^1 x^{a-1}(1-x)^{b-1}\,dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Where $\Gamma(n) = (n-1)!$ if n is a positive integer

Bayes' Billiards (special case of Beta)

$$\int_0^1 x^k(1-x)^{n-k}\,dx = \frac{1}{(n+1)\binom{n}{k}}$$

Euler's Approximation for Harmonic Sums

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \approx \log n + 0.57721\ldots$$

Stirling's Approximation

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Miscellaneous Definitions

Medians A continuous random variable X has median m if $P(X \le m) = 50\%$
A discrete random variable X has median m if $P(X \le m) \ge 50\%$ and $P(X \ge m) \ge 50\%$

Log Statisticians generally use log to refer to ln

i.i.d random variables Independent, identically-distributed random variables.

Example Problems

Contributions from Sebastian Chiu

Calculating Probability (1)

A textbook has n typos, which are randomly scattered amongst its n pages. You pick a random page, what is the probability that it has no typos? **Answer -** There is a $\left(1-\frac{1}{n}\right)$ probability that any specific typo isn't on your page, and thus a $\left(1-\frac{1}{n}\right)^n$ probability that there are no typos on your page. For n large, this is approximately $e^{-1} = \boxed{1/e}$ by a definition of e^x.

Calculating Probability (2)

In a group of n people, what is the expected number of distinct birthdays (month and day). What is the expected number of birthday matches? **Answer -** Let X be the number of distinct birthdays, and let I_j be the indicator for whether the j^{th} days is represented.

$$E(I_j) = 1 - P(\text{no one born day } j) = 1 - (364/365)^n$$

By linearity, $\boxed{E(X) = 365\left(1-(364/365)^n\right)}$. Now let Y be the number of birthday matches and let J_i be the indicator that the i^{th} pair of people have the same birthday. The probability that any two people share a birthday is 1/365 so $\boxed{E(Y) = \binom{n}{2}/365}$.

Linearity of Expectation

*This problem is commonly known as the **hat-matching problem**.* n people have n hats each. At the end of the party, they each leave with a random hat. What is the expected number of people that leave with the right hat? **Answer -** Each hat has a $1/n$ chance of going to the right person. By linearity of expectation, the average number of hats that go to their owners is $\boxed{n(1/n) = 1}$.

First Success and Linearity of Expectation

*This problem is commonly known as the **coupon collector problem**.* There are n total coupons, and each draw, you get a random coupon. What is the expected number of coupons needed until you have a complete set? **Answer -** Let N be the number of coupons needed; we want $E(N)$. Let $N = N_1 + \cdots + N_n$, N_1 is the draws to draw our first distinct coupon, N_2 is the *additional* draws needed to draw our second distinct coupon and so on. By the story of First Success, $N_2 \sim FS((n-1)/n)$ (after collecting first coupon type, there's $(n-1)/n$ chance you'll get something new). Similarly, $N_3 \sim FS((n-2)/n)$, and $N_j \sim FS((n-j+1)/n)$. By linearity,

$$E(N) = E(N_1) + \cdots + E(N_n) = \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} = \boxed{n\sum_{j=1}^{n}\frac{1}{j}}$$

Which is approximately $n\log(n)$ by Euler's approximation for harmonic sums.

First Step Conditioning

In every time period, Bobo the amoeba can die, live, or split into two amoebas with probabilities 0.25, 0.25, and 0.5, respectively. All of Bobo's offspring have the same probabilities. Find $P(D)$, the probability that Bobo's lineage eventually dies out. **Answer -** We use law of probability, and define the events B_0, B_1. and B_2 where B_i means that Bobo has split into i amoebas. We note that $P(D|B_0) = 1$ since his lineage has died, $P(D|B_1) = P(D)$, and $P(D|B_2) = P(D)^2$ since both lines of his lineage must die out in order for Bobo's lineage to die out.

$$\begin{aligned} P(D) &= 0.25P(D|B_0) + 0.25P(D|B_1) + 0.5P(D|B_2) \\ &= 0.25 + 0.25P(D) + 0.5P(D)^2 \end{aligned}$$

Solving the quadratic equation, we get that $P(D) = 0.5$ or 1. We dismiss 1 as an extraneous solution since the expected number of Bobos increase every generation. Thus our answer is $\boxed{P(D) = 0.5}$

Orderings of i.i.d. random variables

I call 2 UberX's and 3 Lyfts at the same time. If the time it takes for the rides to reach me is i.i.d., what is the probability that all the Lyfts will arrive first? **Answer -** since the arrival times of the five cars are i.i.d., all 5! orderings of the arrivals are equally likely. There are 3!2! orderings that involve the Lyfts arriving first, so the probability that the Lyfts arrive first is $\boxed{\frac{3!2!}{5!} = 1/10}$. Alternatively, there are $\binom{5}{3}$ ways to choose 3 of the 5 slots for the Lyfts to occupy, where each of the choices are equally likely. 1 of those choices have all 3 of the Lyfts arriving first, thus the probability is $\boxed{1/\binom{5}{3} = 1/10}$

Expectation of Negative Hypergeometric

What is the expected number of cards that you draw before you pick your first Ace in a shuffled deck? **Answer -** Consider a non-Ace. Denote this to be card j. Let I_j be the indicator that card j will be drawn before the first Ace. Note that if j is before all 4 of the Aces in the deck, then $I_j = 1$. The probability that this occurs is 1/5, because out of 5 cards (the 4 Aces and the not Ace), the probability that the not Ace comes first is 1/5. 1/5 here is the probability that any specific non-Ace will appear before all of the Aces in the deck. (e.g. the probability that the Jack of Spades appears before all of the Aces). Thus let X be the number of cards that is drawn before the first Ace. Then $X = I_1 + I_2 + ... + I_{48}$, where each indicator correspond to one of the 48 not Aces. Thus,

$$E(X) = E(I_1) + E(I_2) + ... + E(I_{48}) = 48/5 = \boxed{9.6}$$

.

Minimum and Maximum of Random Variables

What is the CDF of the maximum of n independent Uniformly-distributed random variables? **Answer -** Note that

$$P(\min(X_1, X_2, \ldots, X_n) \ge a) = P(X_1 \ge a, X_2 \ge a, \ldots, X_n \ge a)$$

Similarily,

$$P(\max(X_1, X_2, \ldots, X_n) \le a) = P(X_1 \le a, X_2 \le a, \ldots, X_n \le a)$$

We will use that principal to find the CDF of $U_{(n)}$, where $U_{(n)} = \max(U_1, U_2, \ldots, U_n)$ where $U_i \sim \text{Unif}(0,1)$ (iid).

$$\begin{aligned} P(\max(U_1, U_2, \ldots, U_n) \le a) &= P(U_1 \le a, U_2 \le a, \ldots, U_n \le a) \\ &= P(U_1 \le a)P(U_2 \le a)\ldots P(U_n \le a) \\ &= \boxed{a^n} \end{aligned}$$

Pattern Matching with e^x Taylor Series

For $X \sim \text{Pois}(\lambda)$, find $E\left(\frac{1}{X+1}\right)$. **Answer -** By LOTUS,

$$E\left(\frac{1}{X+1}\right) = \sum_{k=0}^{\infty}\frac{1}{k+1}\frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-\lambda}}{\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{k+1}}{(k+1)!} = \boxed{\frac{e^{-\lambda}}{\lambda}(e^\lambda - 1)}$$

Adam and Eve's Laws

William really likes speedsolving Rubik's Cubes. But he's pretty bad at it, so sometimes he fails. On any given day, William will attempt $N \sim \text{Geom}(s)$ Rubik's Cubes. Suppose each time, he has a independent probability p of solving the cube. Let T be the number of Rubik's Cubes he solves during a day. Find the mean and variance of T. **Answer -** Note that $T|N \sim \text{Bin}(N, p)$. As a result, we have by Adam's Law that

$$E(T) = E(E(T|N)) = E(Np) = \boxed{\frac{p(1-s)}{s}}$$

Similarly, by Eve's Law, we have that

$$\begin{aligned} \text{Var}(T) &= E(\text{Var}(T|N)) + \text{Var}(E(T|N)) = E(Np(1-p)) + \text{Var}(Np) \\ &= \frac{p(1-p)(1-s)}{s} + \frac{p^2(1-s)}{s^2} = \boxed{\frac{p(1-s)(p+s(1-p))}{s^2}} \end{aligned}$$

MGF - Distribution Matching

(Referring to the Rubik's Cube question above) Find the MGF of T. What is the name of this distribution and its parameter(s)? **Answer -** By Adam's Law, we have that

$$\begin{aligned} E(e^{tT}) = E(E(e^{tT}|N)) = E((pe^t + q)^N) &= s\sum_{n=0}^{\infty}(pe^t + 1 - p)^n(1-s)^n \\ &= \frac{s}{1-(1-s)(pe^t+1-p)} = \frac{s}{s+(1-s)p-(1-s)pe^t} \end{aligned}$$

Intuitively, we would expect that T is distributed Geometrically because T is just a filtered version of N, which itself is Geometrically distributed. The MGF of a Geometric random variable $X \sim \text{Geom}(\theta)$ is

$$E(e^{tX}) = \frac{\theta}{1-(1-\theta)e^t}$$

So, we would want to try to get our MGF into this form to identify what θ is. Taking our original MGF, it would appear that dividing by $s + (1-s)p$ would allow us to do this. Therefore, we have that

$$E(e^{tT}) = \frac{s}{s+(1-s)p-(1-s)pe^t} = \frac{\frac{s}{s+(1-s)p}}{1-\frac{(1-s)p}{s+(1-s)p}e^t}$$

By pattern-matching, it thus follows that $\boxed{T \sim \text{Geom}(\theta)}$ where

$$\boxed{\theta = \frac{s}{s+(1-s)p}}$$

MGF - Finding Momemts

Find $E(X^3)$ for $X \sim \text{Expo}(\lambda)$ using the MGF of X. **Answer -** The MGF of an Expo(λ) is $M(t) = \frac{\lambda}{\lambda - t}$. To get the third moment, we can take the third derivative of the MGF and evaluate at $t = 0$:

$$\boxed{E(X^3) = \frac{6}{\lambda^3}}$$

But a much nicer way to use the MGF here is via pattern recognition: note that $M(t)$ looks like it came from a geometric series:

$$\frac{1}{1 - \frac{t}{\lambda}} = \sum_{n=0}^{\infty} \left(\frac{t}{\lambda}\right)^n = \sum_{n=0}^{\infty} \frac{n!}{\lambda^n} \frac{t^n}{n!}$$

The coefficient of $\frac{t^n}{n!}$ here is the n^{th} moment of X, so we have $E(X^n) = \frac{n!}{\lambda^n}$ for all nonnegative integers n. So again we get the same answer.

Markov Chains

Suppose X_n is a two-state Markov chain with transition matrix

$$Q = \begin{matrix} \\ 0 \\ 1 \end{matrix} \begin{matrix} \begin{matrix} 0 & 1 \end{matrix} \\ \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \end{matrix}$$

Find the stationary distribution $\vec{s} = (s_0, s_1)$ of X_n by solving $\vec{s}Q = \vec{s}$, and show that the chain is reversible under this stationary distribution. **Answer -** By solving $\vec{s}Q = \vec{s}$, we have that

$$s_0 = s_0(1-\alpha) + s_1\beta \text{ and } s_1 = s_0(\alpha) + s_0(1-\beta)$$

And by solving this system of linear equations it follows that

$$\boxed{\vec{s} = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)}$$

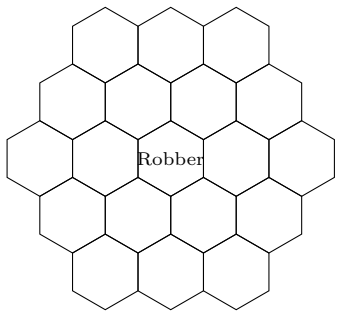
To show that this chain is reversible under this stationary distribution, we must show $s_i q_{ij} = s_j q_{ji}$ for all i, j. This is done if we can show $s_0 q_{01} = s_1 q_{10}$. Indeed,

$$s_0 q_{01} = \frac{\alpha\beta}{\alpha+\beta} = s_1 q_{10}$$

thus our chain is reversible under the stationary distribution.

Markov Chains, continued

William and Sebastian play a modified game of Settlers of Catan, where every turn they randomly move the robber (which starts on the center tile) to one of the adjacent hexagons.



a) Is this Markov Chain irreducible? Is it aperiodic? **Answer -** Yes to both The Markov Chain is irreducible because it can get from anywhere to anywhere else. The Markov Chain is also aperiodic because the robber can return back to a square in $2, 3, 4, 5, \ldots$ moves. Those numbers have a GCD of 1, so the chain is aperiodic.

b) What is the stationary distribution of this Markov Chain? **Answer -** Since this is a random walk on an undirected graph, the stationary distribution is proportional to the degree sequence. The degree for the corner pieces is 3, the degree for the edge pieces is 4, and the degree for the center pieces is 6. To normalize this degree sequence, we divide by its sum. The sum of the degrees is $6(3) + 6(4) + 7(6) = 84$. Thus the stationary probability of being on a corner is $3/84 = 1/28$, on an edge is $4/84 = 1/21$, and in the center is $6/84 = 1/14$.

c) What fraction of the time will the robber be in the center tile in this game? **Answer -** From above, $\boxed{1/14}$.

d) What is the expected amount of moves it will take for the robber to return? **Answer -** Since this chain is irreducible and aperiodic, to get the expected time to return we can just invert the stationary probability. Thus on average it will take $\boxed{14}$ turns for the robber to return to the center tile.

Problem Solving Strategies

Contributions from Jessy Hwang, Yuan Jiang, Yuqi Hou

1. **Getting Started.** Start by *defining events* and/or *defining random variables*. ("Let A be the event that I pick the fair coin"; "Let X be the number of successes.") Clear notion = clear thinking! Then decide what it is that you're supposed to be finding, in terms of your location ("I want to find $P(X = 3|A)$"). Try simple and extreme cases. To make an abstract experiment more concrete, try drawing a picture or making up numbers that could have happened. Pattern recognition: does the structure of the problem resemble something we've seen before.
2. **Calculating Probability of an Event.** Use combinatorics if the naive definition of probability applies. Look for symmetries or something to condition on, then apply Bayes' rule or LoTP. Is the probability of the complement easier to find?
3. **Finding the distribution of a random variable.** Check the *support* of the random variable: what values can it take on? Use this to rule out distributions that don't fit. - Is there a *story* for one of the named distributions that fits the problem at hand? - Can you write the random variable as a function of a r.v. with a known distribution, say $Y = g(X)$? Then work directly from the definition of PDF or PMF, expressing $P(Y \leq y)$ or $P(Y = y)$ in terms of events involving X only. - For PDFs, find the CDF first and then differentiate. - If you're trying to find the joint distribution of two *independent* random variables, just multiple their marginal probabilities - Do you need the distribution? If the question only asks for the expected value of X, you might be able to find this without knowing the entire distirbution of X. See the next item.
4. **Calculating Expectation.** If it has a named distribution, check out the table of distributions. If its a function of a r.v. with a named distribution, try LotUS. If its a count of something, try breaking it up into indicator random variables. If you can condition on something, consider using Adam's law. Also consider the variance formula.
5. **Calculating Variance.** Consider independence, named distributions, and LotUS. If it's a count of something, break it up into a sum of indicator random variables. If you can condition on something, consider using Eve's Law.
6. **Calculating $E(X^2)$** - Do you already know $E(X)$ or $\text{Var}(X)$? Remember that $\text{Var}(X) = E(X^2) - E(X)^2$.
7. **Calculating Covariance** If it's a count of something, break it up into a sum of indicator random variables. If you're trying to calculate the covariance between two components of a multinomial distribution, X_i, X_j, then the covariance is $-np_ip_j$.
8. If X and Y are i.i.d., have you considered using symmetry?
9. **Calculating Probabilities of Orderings of Random Variables** Have you considered looking at order statistics? - Remember any ordering of i.i.d. random variables is equally likely.
10. Is this the birthday problem? Is this a multinomial problem?
11. **Determining Independence** Use the definition of independence. Think of extreme cases to see if you can find a counterexample.
12. Does something look like Simpson's Paradox? make sure you're looking at 3 events.
13. **Find the PDF.** If the question gives you two r.v., where you know the PDF of one r.v. and the other r.v. is a function of the first one, then the problem wants you to use a transformation of variables (Jacobian). You can also find the pdf by differentiating the CDF.
14. **Do a painful integral.** If your integral looks painful, see if you can write your integral in terms of a PDF (like Gamma or Beta), so that the integral equals 1.
15. **Before moving on.** Plug in some simple and extreme cases to make sure that your answer makes sense.

Biohazards

Section author: Jessy Hwang

1. **Don't misuse the native definition of probability** - When answering "What is the probability that in a group of 3 people, no two have the same birth month?", it is *not* correct to treat the people as indistinguishable balls being placed into 12 boxes, since that assumes the list of birth months {January, January, January} is just as likely as the list {January, April, June}, when the latter is fix times more likely.
2. **Don't confuse unconditional and conditional probabilities, or go in circles with Baye's Rule** - $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. It is *not* correct to say "$P(B) = 1$ because we know that B happened."; P(B) is the probability *before* we have information about whether B happened. It is *not* correct to use $P(A|B)$ in place of $P(A)$ on the right-hand side.
3. **Don't assume independence without justification** - In the matching problem, the probability that card 1 is a match and card 2 is a match is not $1/n^2$. - The Binomial and Hypergeometric are often confused; the trials are independent in the Binomial story and not independent in the Hypergeometric story due to the lack of replacement.
4. **Don't confuse random variables, numbers, and events.** - Let X be a r.v. Then $f(X)$ is a r.b. for any function f. In particular, X^2, $|X|$, $F(X)$, and $I_{X>3}$ are r.v.s. $P(X^2 < X|X \geq 0), E(X), \text{Var}(X)$, and $f(E(X))$ are numbers. $X = 2$ and $F(X) \geq -1$ are events. It does not make sense to write $\int_{-\infty}^{\infty} F(X)dx$ because $F(X)$ is a random variable. It does not make sense to write $P(X)$ because X is not an event.
5. **A random variable is not the same thing as its distribution** - To get the PDF of X^2, you can't just square the PDF of X. The right way is to use one variable transformations - To get the PDF of $X + Y$, you can't just add the PDF of X and the PDF of Y. The right way is to compute the convolution.
6. $E(g(X))$ **does not equal** $g(E(X))$ **in general.** - See the St. Petersburg paradox for an extreme example. - The right way to find $E(g(X))$ is with LotUS.

Recommended Resources

- Introduction to Probability (`http://bit.ly/introprobability`)
- Stat 110 Online (`http://stat110.net`)
- Stat 110 Quora Blog (`https://stat110.quora.com/`)
- Stat 110 Course Notes (`mxawng.com/stuff/notes/stat110.pdf`)
- Quora Probability FAQ (`http://bit.ly/probabilityfaq`)
- LaTeX File (`github.com/wzchen/probability_cheatsheet`)

Please share this cheatsheet with friends!
`http://wzchen.com/probability-cheatsheet`

Distributions

Distribution	PDF and Support	Expected Value	Variance	MGF
Bernoulli Bern(p)	$P(X=1)=p$ $P(X=0)=q$	p	pq	$q+pe^t$
Binomial Bin(n,p)	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ $k\in\{0,1,2,\dots n\}$	np	npq	$(q+pe^t)^n$
Geometric Geom(p)	$P(X=k)=q^kp$ $k\in\{0,\ 1,\ 2,\ \dots\}$	q/p	q/p^2	$\frac{p}{1-qe^t}, qe^t<1$
Negative Binom. NBin(r,p)	$P(X=n)=\binom{r+n-1}{r-1}p^rq^n$ $n\in\{0,\ 1,\ 2,\ \dots\}$	rq/p	rq/p^2	$(\frac{p}{1-qe^t})^r, qe^t<1$
Hypergeometric HGeom(w,b,n)	$P(X=k)=\binom{w}{k}\binom{b}{n-k}/\binom{w+b}{n}$ $k\in\{0,1,2,\dots,n\}$	$\mu=\frac{nw}{b+w}$	$\frac{w+b-n}{w+b-1}n\frac{\mu}{n}(1-\frac{\mu}{n})$	–
Poisson Pois(λ)	$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ $k\in\{0,\ 1,\ 2,\ \dots\}$	λ	λ	$e^{\lambda(e^t-1)}$
Uniform Unif(a,b)	$f(x)=\frac{1}{b-a}$ $x\in(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Normal $\mathcal{N}(\mu,\sigma^2)$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$ $x\in(-\infty,\infty)$	μ	σ^2	$e^{t\mu+\frac{\sigma^2t^2}{2}}$
Exponential Expo(λ)	$f(x)=\lambda e^{-\lambda x}$ $x\in(0,\infty)$	$1/\lambda$	$1/\lambda^2$	$\frac{\lambda}{\lambda-t}, t<\lambda$
Gamma Gamma(a,λ)	$f(x)=\frac{1}{\Gamma(a)}(\lambda x)^ae^{-\lambda x}\frac{1}{x}$ $x\in(0,\infty)$	a/λ	a/λ^2	$\left(\frac{\lambda}{\lambda-t}\right)^a, t<\lambda$
Beta Beta(a, b)	$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ $x\in(0,1)$	$\mu=\frac{a}{a+b}$	$\frac{\mu(1-\mu)}{(a+b+1)}$	–
Chi-Squared χ^2_n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$ $x\in(0,\infty)$	n	$2n$	$(1-2t)^{-n/2}, t<1/2$
Multivar Uniform A is support	$f(x)=\frac{1}{\lvert A\rvert}$ $x\in A$	–	–	–
Multinomial Mult$_k(n,\vec{p})$	$P(\vec{X}=\vec{n})=\binom{n}{n_1\dots n_k}p_1^{n_1}\dots p_k^{n_k}$ $n=n_1+n_2+\dots+n_k$	$n\vec{p}$	$\mathrm{Var}(X_i)=np_i(1-p_i)$ $\mathrm{Cov}(X_i,X_j)=-np_ip_j$	$\left(\sum_{i=1}^k p_ie^{t_i}\right)^n$

Inequalities

Cauchy-Schwarz

$$|E(XY)|\le\sqrt{E(X^2)E(Y^2)}$$

Markov

$$P(X\ge a)\le\frac{E|X|}{a}$$

Chebychev

$$P(|X-\mu_X|\ge a)\le\frac{\sigma_X^2}{a^2}$$

Jensen

g convex: $E(g(X))\ge g(E(X))$

g concave: $E(g(X))\le g(E(X))$