ULTIMATE COLLECTION OF ALL CHEAT SHEETS & SHORT NOTES BY HARVARD UNIVERSITY, MASSACHUSETTS INSTITUTE OF TECHNOLOGY & CORNELL UNIVERSITY

LINEAR ALGEBRA, STATISTICS & PROBABILITY FOR DATA SCIENCE

10/22/2020

COMPILED BY ABHISHEK PRASAD

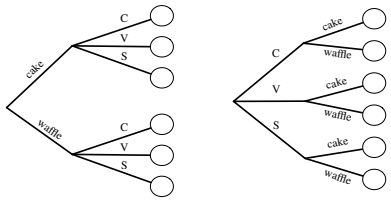
Probability Cheatsheet v2.0

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

Last Updated September 4, 2015

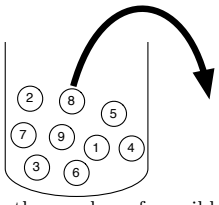
Counting

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 table gives the number of possible samples of size k out of a population of size n, under various assumptions about how the sample is collected.

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

Naive Definition of Probability

If all outcomes are equally likely, the probability of an event A happening is:

$$P_{\text{naive}}(A) = \frac{\text{number of outcomes favorable to } A}{\text{number of outcomes}}$$

Thinking Conditionally

Independence

Independent Events A and B are independent if knowing whether A occurred gives no information about whether B occurred. More formally, A and B (which have nonzero probability) are independent if and only if one of the following equivalent statements holds:

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

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

Unions, Intersections, and Complements

De Morgan's Laws A useful identity that can make calculating probabilities of unions easier by relating them to intersections, and vice versa. Analogous results hold with more than two sets.

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Joint, Marginal, and Conditional

Joint Probability $P(A \cap B)$ or $P(A, B)$ – Probability of A and B.

Marginal (Unconditional) Probability $P(A)$ – Probability of A.

Conditional Probability $P(A|B) = P(A, B)/P(B)$ – Probability of A, given that B occurred.

Conditional Probability *is* Probability $P(A|B)$ is a probability function for any fixed B. Any theorem that holds for probability also holds for conditional probability.

Probability of an Intersection or Union

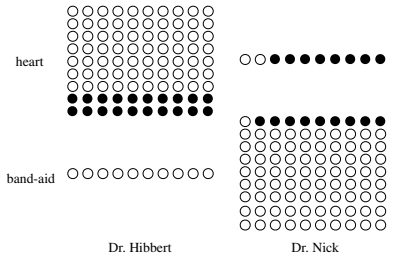
Intersections via Conditioning

$$P(A, B) = P(A)P(B|A)$$
$$P(A, B, C) = P(A)P(B|A)P(C|A, B)$$

Unions via Inclusion-Exclusion

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

Simpson's Paradox



It is possible to have

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

yet also $P(A \mid B) > P(A \mid B^c)$.

Law of Total Probability (LOTP)

Let $B_1, B_2, B_3, ...B_n$ be a *partition* of the sample space (i.e., they are disjoint and their union is the entire sample space).

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)$$
$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)$$

For **LOTP with extra conditioning**, just add in another event C!

$$P(A|C) = P(A|B_1, C)P(B_1|C) + \cdots + P(A|B_n, C)P(B_n|C)$$
$$P(A|C) = P(A \cap B_1|C) + P(A \cap B_2|C) + \cdots + P(A \cap B_n|C)$$

Special case of LOTP with B and B^c as partition:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$
$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Bayes' Rule

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

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

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

We can also write

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

Odds Form of Bayes' Rule

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

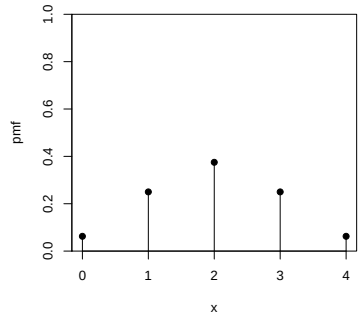
The *posterior odds* of A are the *likelihood ratio* times the *prior odds*.

Random Variables and their Distributions

PMF, CDF, and Independence

Probability Mass Function (PMF) Gives the probability that a *discrete* random variable takes on the value x.

$$p_X(x) = P(X = x)$$

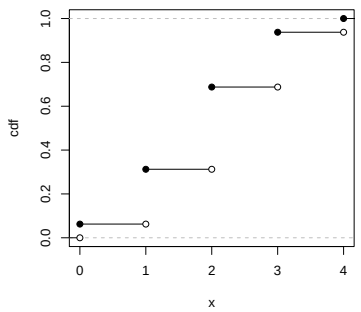


The PMF satisfies

$$p_X(x) \geq 0 \text{ and } \sum_x p_X(x) = 1$$

Cumulative Distribution Function (CDF) Gives the probability that a random variable is less than or equal to x.

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



The CDF is an increasing, right-continuous function with

$$F_X(x) \to 0 \text{ as } x \to -\infty \text{ and } F_X(x) \to 1 \text{ as } x \to \infty$$

Independence Intuitively, two random variables are independent if knowing the value of one gives no information about the other. Discrete r.v.s 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

Expected Value and Linearity

Expected Value (a.k.a. *mean*, *expectation*, or *average*) is a weighted average of the possible outcomes of our random variable. Mathematically, if $x_1, x_2, x_3, \ldots$ are all of the distinct possible values that X can take, the expected value of X is

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

X	Y	$X+Y$
3	4	7
2	2	4
6	8	14
10	23	33
1	−3	−2
1	0	1
5	9	14
4	1	5
...	...	...
$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\frac{1}{n}\sum_{i=1}^{n} y_i$	$\frac{1}{n}\sum_{i=1}^{n}(x_i+y_i)$
$E(X)$	$E(Y)$	$E(X+Y)$

Linearity For any r.v.s X and Y, and constants a, b, c,

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

Same distribution implies same mean If X and Y have the same distribution, then $E(X) = E(Y)$ and, more generally,

$$E(g(X)) = E(g(Y))$$

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

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

Indicator Random Variables

Indicator Random Variable is a random variable that takes on the value 1 or 0. It 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 about counting how many events of some kind occur. Write

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

Note that $I_A^2 = I_A, I_A I_B = I_{A\cap B}$, and $I_{A\cup B} = I_A + I_B - I_A I_B$.

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

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

Variance and Standard Deviation

$$\text{Var}(X) = E\left(X - E(X)\right)^2 = E(X^2) - (E(X))^2$$

$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

Continuous RVs, LOTUS, UoU

Continuous Random Variables (CRVs)

What's the probability that a CRV is in an interval? Take the difference in CDF values (or use the PDF as described later).

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

For $X \sim \mathcal{N}(\mu, \sigma^2)$, this becomes

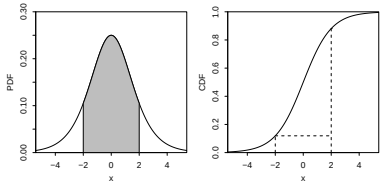
$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

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

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

A PDF is nonnegative and integrates to 1. By the fundamental theorem of calculus, to get from PDF back to CDF we can integrate:

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



To find the probability that a CRV takes on a value in an interval, integrate the PDF over that interval.

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

How do I find the expected value of a CRV? Analogous to the discrete case, where you sum x times the PMF, for CRVs you integrate x times the PDF.

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

LOTUS

Expected value of a function of an r.v. The expected value of X is defined this way:

$$E(X) = \sum_x xP(X = x) \text{ (for discrete } X)$$

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

The **Law of the Unconscious Statistician (LOTUS)** states that you can find the expected value of a *function of a random variable*, $g(X)$, in a similar way, by replacing the x in front of the PMF/PDF by $g(x)$ but still working with the PMF/PDF of X:

$$E(g(X)) = \sum_x g(x)P(X = x) \text{ (for discrete } X)$$

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

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 $g(X) = 2X$ is the number of bike wheels you see in that hour and $h(X) = \binom{X}{2} = \frac{X(X-1)}{2}$ is the number of *pairs* of bikes such that you see both of those bikes in that hour.

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

Universality of Uniform (UoU)

When you plug any CRV into its own CDF, you get a Uniform(0,1) random variable. When you plug a Uniform(0,1) r.v. into an inverse CDF, you get an r.v. with that CDF. For example, let's say that a random variable X has CDF

$$F(x) = 1 - e^{-x}, \text{ for } x > 0$$

By UoU, if we plug X into this function then we get a uniformly distributed random variable.

$$F(X) = 1 - e^{-X} \sim \text{Unif}(0, 1)$$

Similarly, if $U \sim \text{Unif}(0, 1)$ then $F^{-1}(U)$ has CDF F. 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.

Moments and MGFs

Moments

Moments describe the shape of a distribution. Let X have mean μ and standard deviation σ, and $Z = (X - \mu)/\sigma$ be the *standardized* version of X. The kth moment of X is $\mu_k = E(X^k)$ and the kth standardized moment of X is $m_k = E(Z^k)$. The mean, variance, skewness, and kurtosis are important summaries of the shape of a distribution.

Mean $E(X) = \mu_1$

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

Skewness $\text{Skew}(X) = m_3$

Kurtosis $\text{Kurt}(X) = m_4 - 3$

Moment Generating Functions

MGF For any random variable X, the function

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

is the **moment generating function (MGF)** of X, if it exists for all t in some open interval containing 0. The variable t could just as well have been called u or v. It's a bookkeeping device that lets us work with the *function* M_X rather than the *sequence* of moments.

Why is it called the Moment Generating Function? Because the kth derivative of the moment generating function, evaluated at 0, is the kth moment of X.

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

This is true by Taylor expansion of e^{tX} since

$$M_X(t) = E(e^{tX}) = \sum_{k=0}^{\infty} \frac{E(X^k)t^k}{k!} = \sum_{k=0}^{\infty} \frac{\mu_k t^k}{k!}$$

MGF of linear functions If we have $Y = aX + b$, then

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

Uniqueness *If it exists, the MGF uniquely determines the distribution.* This means that for any two random variables X and Y, they are distributed the same (their PMFs/PDFs are equal) if and only if their MGFs are equal.

Summing Independent RVs by Multiplying MGFs. If X and Y are independent, then

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

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

The **joint CDF** of X and Y is

$$F(x, y) = P(X \leq x, Y \leq y)$$

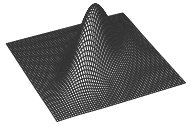
In the discrete case, X and Y have a **joint PMF**

$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

In the continuous case, they have a **joint PDF**

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

The joint PMF/PDF must be nonnegative and sum/integrate to 1.



Conditional Distributions

Conditioning and Bayes' rule for discrete r.v.s

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

Conditioning and Bayes' rule for continuous r.v.s

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

Hybrid Bayes' rule

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

Marginal Distributions

To find the distribution of one (or more) random variables from a joint PMF/PDF, sum/integrate over the unwanted random variables.

Marginal PMF from joint PMF

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

Marginal PDF from joint PDF

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$

Independence of Random Variables

Random variables X and Y are independent if and only if any of the following conditions holds:

- Joint CDF is the product of the marginal CDFs
- Joint PMF/PDF is the product of the marginal PMFs/PDFs
- Conditional distribution of Y given X is the marginal distribution of Y

Write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent.

Multivariate LOTUS

LOTUS in more than one dimension is analogous to the 1D LOTUS. 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 analog of variance for two random variables.

$$\text{Cov}(X, Y) = E\left((X - E(X))(Y - E(Y))\right) = E(XY) - E(X)E(Y)$$

Note that

$$\text{Cov}(X, X) = E(X^2) - (E(X))^2 = \text{Var}(X)$$

Correlation is a standardized version of covariance that is always between -1 and 1.

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

Covariance and Independence If two random variables are independent, then they are uncorrelated. The converse is not necessarily true (e.g., consider $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$).

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

Covariance and Variance The variance of a sum can be found by

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

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

If X and Y are independent then they have covariance 0, so

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

If $X_1, X_2, \ldots, X_n$ are identically distributed and have the same covariance relationships (often by **symmetry**), 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 Properties For random variables W, X, Y, Z and constants a, b:

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

Correlation is location-invariant and scale-invariant For any constants a, b, c, d with a and c nonzero,

$$\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$$

Transformations

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 $Y = g(X)$. Also let $y = g(x)$. If g is differentiable and strictly increasing (or strictly decreasing), then the PDF of Y is

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

The derivative of the inverse transformation is called the **Jacobian**.

Two Variable Transformations Similarly, let's say we know the joint PDF of U and V but are also interested in the random vector (X, Y) defined by $(X, Y) = g(U, V)$. Let

$$\frac{\partial(u, v)}{\partial(x, y)} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}$$

be the **Jacobian matrix**. If the entries in this matrix exist and are continuous, and the determinant of the matrix is never 0, then

$$f_{X,Y}(x, y) = f_{U,V}(u, v)\left|\left|\frac{\partial(u, v)}{\partial(x, y)}\right|\right|$$

The inner bars tells us to take the matrix's determinant, and the outer bars tell us to take the absolute value. In a 2×2 matrix,

$$\left|\left| \begin{matrix} a & b \\ c & d \end{matrix} \right|\right| = |ad - bc|$$

Convolutions

Convolution Integral If you want to find the PDF of the sum of two independent CRVs X and Y, you can do the following integral:

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

Example Let $X, Y \sim \mathcal{N}(0, 1)$ be i.i.d. Then for each fixed t,

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

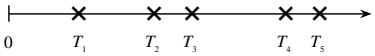
By completing the square and using the fact that a Normal PDF integrates to 1, this works out to $f_{X+Y}(t)$ being the $\mathcal{N}(0, 2)$ PDF.

Poisson Process

Definition We have a **Poisson process** of rate λ arrivals per unit time if the following conditions hold:

1. The number of arrivals in a time interval of length t is Pois(λt).
2. Numbers of arrivals in disjoint time intervals are independent.

For example, the numbers of arrivals in the time intervals $[0, 5]$, $(5, 12)$, and $[13, 23)$ are independent with Pois(5λ), Pois(7λ), Pois(10λ) distributions, respectively.



Count-Time Duality Consider a Poisson process of emails arriving in an inbox at rate λ emails per hour. Let T_n be the time of arrival of the nth email (relative to some starting time 0) and N_t be the number of emails that arrive in $[0, t]$. Let's find the distribution of T_1. The event $T_1 > t$, the event that you have to wait more than t hours to get the first email, is the same as the event $N_t = 0$, which is the event that there are no emails in the first t hours. So

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

Thus we have $T_1 \sim$ Expo(λ). By the memoryless property and similar reasoning, the interarrival times between emails are i.i.d. Expo(λ), i.e., the differences $T_n - T_{n-1}$ are i.i.d. Expo(λ).

Order Statistics

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

Note that the order statistics are *dependent*, e.g., learning $X_{(4)} = 42$ gives us the information that $X_{(1)}, X_{(2)}, X_{(3)}$ are ≤ 42 and $X_{(5)}, X_{(6)}, \ldots, X_{(n)}$ are ≥ 42.

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

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

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

Uniform Order Statistics The jth order statistic of i.i.d. $U_1, \ldots, U_n \sim$ Unif$(0, 1)$ is $U_{(j)} \sim$ Beta$(j, n - j + 1)$.

Conditional Expectation

Conditioning on an Event We can find $E(Y|A)$, the expected value of Y given that event A occurred. A very important case is when A is the event $X = x$. Note that $E(Y|A)$ is a *number*. For example:

- The expected value of a fair die roll, given that it is prime, is $\frac{1}{3} \cdot 2 + \frac{1}{3} \cdot 3 + \frac{1}{3} \cdot 5 = \frac{10}{3}$.
- Let Y be the number of successes in 10 independent Bernoulli trials with probability p of success. Let A be the event that the first 3 trials are all successes. Then

  $$E(Y|A) = 3 + 7p$$

  since the number of successes among the last 7 trials is Bin$(7, p)$.
- Let $T \sim$ Expo$(1/10)$ be how long you have to wait until the shuttle comes. Given that you have already waited t minutes, the expected additional waiting time is 10 more minutes, by the memoryless property. That is, $E(T|T > t) = t + 10$.

Discrete Y	Continuous Y
$E(Y) = \sum_y yP(Y = y)$	$E(Y) = \int_{-\infty}^{\infty} yf_Y(y)dy$
$E(Y\|A) = \sum_y yP(Y = y\|A)$	$E(Y\|A) = \int_{-\infty}^{\infty} yf(y\|A)dy$

Conditioning on a Random Variable We can also find $E(Y|X)$, the expected value of Y given the random variable X. This is *a function of the random variable* X. It is *not* a number except in certain special cases such as if $X \perp\!\!\!\perp Y$. To find $E(Y|X)$, find $E(Y|X = x)$ and then plug in X for x. For example:

- If $E(Y|X = x) = x^3 + 5x$, then $E(Y|X) = X^3 + 5X$.
- Let Y be the number of successes in 10 independent Bernoulli trials with probability p of success and X be the number of successes among the first 3 trials. Then $E(Y|X) = X + 7p$.
- Let $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$. Then $E(Y|X = x) = x^2$ since if we know $X = x$ then we know $Y = x^2$. And $E(X|Y = y) = 0$ since if we know $Y = y$ then we know $X = \pm\sqrt{y}$, with equal probabilities (by symmetry). So $E(Y|X) = X^2, E(X|Y) = 0$.

Properties of Conditional Expectation

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

Adam's Law (a.k.a. Law of Total Expectation) can also be written in a way that looks analogous to LOTP. For any events $A_1, A_2, \ldots, A_n$ that partition the sample space,

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

For the special case where the partition is A, A^c, this says

$$E(Y) = E(Y|A)P(A) + E(Y|A^c)P(A^c)$$

Eve's Law (a.k.a. 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 $X_1, X_2, X_3 \ldots$ be i.i.d. with mean μ. The **sample mean** is

$$\bar{X}_n = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

The **Law of Large Numbers** states that as $n \to \infty$, $\bar{X}_n \to \mu$ with probability 1. For example, in flips of a coin with probability p of Heads, let X_j be the indicator of the jth flip being Heads. Then LLN says the proportion of Heads converges to p (with probability 1).

Central Limit Theorem (CLT)

Approximation using CLT

We use $\dot{\sim}$ to denote *is approximately distributed*. We can use the **Central Limit Theorem** to approximate the distribution of a random variable $Y = X_1 + X_2 + \cdots + X_n$ that is a sum of n i.i.d. random variables X_i. Let $E(Y) = \mu_Y$ and $\text{Var}(Y) = \sigma_Y^2$. The CLT says

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

If the X_i are i.i.d. with mean μ_X and variance σ_X^2, then $\mu_Y = n\mu_X$ and $\sigma_Y^2 = n\sigma_X^2$. For the sample mean $\bar{X}_n$, the CLT says

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

Asymptotic Distributions using CLT

We use $\xrightarrow{D}$ to denote *converges in distribution to* as $n \to \infty$. The CLT says that if we standardize the sum $X_1 + \cdots + X_n$ then the distribution of the sum converges to $\mathcal{N}(0, 1)$ as $n \to \infty$:

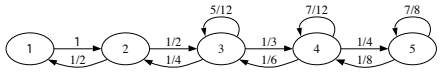
$$\frac{1}{\sigma\sqrt{n}}(X_1 + \cdots + X_n - n\mu_X) \xrightarrow{D} \mathcal{N}(0, 1)$$

In other words, the CDF of the left-hand side goes to the standard Normal CDF, Φ. In terms of the sample mean, the CLT says

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

Markov Chains

Definition



A Markov chain is a random walk in a **state space**, which we will assume is finite, say $\{1, 2, \ldots, M\}$. We let X_t denote which element of the state space the walk is visiting at time t. The Markov chain is the sequence of random variables tracking where the walk is at all points in time, $X_0, X_1, X_2, \ldots$. By definition, a Markov chain must satisfy the **Markov property**, which says that if you want to predict where the chain will be at a future time, if we know the present state then the entire past history is irrelevant. *Given the present, the past and future are conditionally independent.* In symbols,

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

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 positive 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 the possible numbers of steps it would take to return back is $k > 1$.
- Otherwise you are at an **aperiodic state**. The GCD of the possible numbers of steps it would take to return back is 1.

Transition Matrix

Let the state space be $\{1, 2, \ldots, M\}$. The transition matrix Q is the $M \times M$ matrix where element q_{ij} is the probability that the chain goes from state i to state j in one step:

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

To find the probability that the chain goes from state i to state j in exactly m steps, take the (i, j) element of Q^m.

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

If X_0 is distributed according to the row vector PMF $\vec{p}$, i.e., $p_j = P(X_0 = j)$, 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. If a chain (on a finite state space) is irreducible, then all of its states are 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. Examples of reversible chains include any chain with $q_{ij} = q_{ji}$, with $\vec{s} = (\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M})$, and random walk on an undirected network.

Stationary Distribution

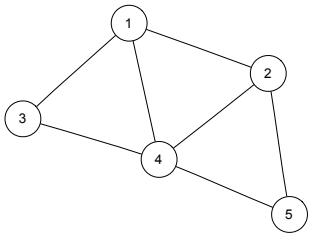
Let us say that the vector $\vec{s} = (s_1, s_2, \ldots, s_M)$ be a PMF (written as a row vector). We will call $\vec{s}$ the **stationary distribution** for the chain if $\vec{s}Q = \vec{s}$. As a consequence, if X_t has the stationary distribution, then all future $X_{t+1}, X_{t+2}, \ldots$ also have 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 to i starting from i is $1/s_i$.

To find the stationary distribution, you can solve the matrix equation $(Q' - I)\vec{s}' = 0$. The stationary distribution is uniform if the columns of Q sum to 1.

Reversibility Condition Implies Stationarity If you have a PMF $\vec{s}$ and a Markov chain with transition matrix Q, then $s_i q_{ij} = s_j q_{ji}$ for all states i, j implies that $\vec{s}$ is stationary.

Random Walk on an Undirected Network



If you have a collection of **nodes**, pairs of which can be connected by undirected **edges**, and a Markov chain is run by going from the current node to a uniformly random node that is connected to it by an edge, then this is a random walk on an undirected network. The stationary distribution of this chain is proportional to the **degree sequence** (this is the sequence of degrees, where the degree of a node is how many edges are attached to it). For example, the stationary distribution of random walk on the network shown above is proportional to $(3, 3, 2, 4, 2)$, so it's $(\frac{3}{14}, \frac{3}{14}, \frac{3}{14}, \frac{4}{14}, \frac{2}{14})$.

Continuous Distributions

Uniform Distribution

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 a draw from any interval within the support is proportional to the length of the interval. See *Universality of Uniform* and *Order Statistics* for other properties.

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 probability of hitting in any specific region is proportional to the length/area/volume of that region, and where the density of occurrence in any one specific spot is constant throughout the whole support.

Normal Distribution

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 the sample mean of i.i.d. r.v.s will approach a Normal distribution as the sample size grows, regardless of the initial distribution.

Location-Scale Transformation Every time we shift a Normal r.v. (by adding a constant) or rescale a Normal (by multiplying by a constant), we change it to another Normal r.v. For any Normal $X \sim \mathcal{N}(\mu, \sigma^2)$, we can transform it to the standard $\mathcal{N}(0, 1)$ by the following transformation:

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

Standard Normal The Standard Normal, $Z \sim \mathcal{N}(0, 1)$, has mean 0 and variance 1. Its CDF is denoted by Φ.

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 a shooting star is not "due" to come just because you've waited so long. Your waiting time is memoryless; the additional 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) hours. Here $\lambda = 4$ is the **rate parameter**, since shooting stars arrive at a rate of 1 per 1/4 hour on average. The expected time until the next shooting star is $1/\lambda = 1/4$ hour.

Expos as a rescaled Expo(1)

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

Memorylessness The Exponential Distribution is the only continuous memoryless distribution. The memoryless property says that for $X \sim \text{Expo}(\lambda)$ and any positive numbers s and t,

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

Equivalently,

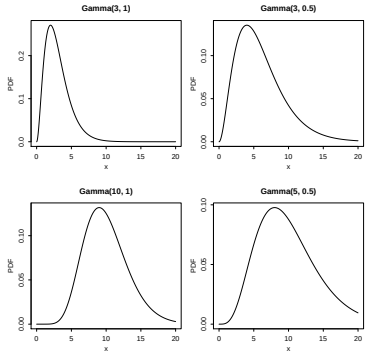
$$X - a | (X > a) \sim \text{Expo}(\lambda)$$

For example, a product with an Expo(λ) lifetime is always "as good as new" (it doesn't experience wear and tear). Given that the product has survived a years, the additional time that it will last is still Expo(λ).

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 $\max(X_1, \ldots, X_k)$ has the same distribution as $Y_1 + Y_2 + \cdots + Y_k$, where $Y_j \sim \text{Expo}(j\lambda)$ and the Y_j are independent.

Gamma Distribution

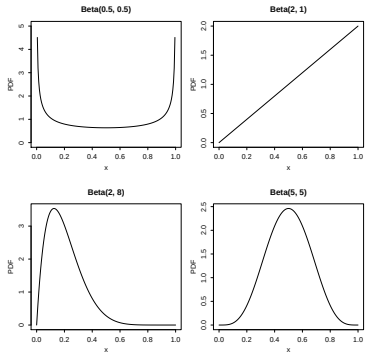


Let us say that X is distributed Gamma(a, λ). We know the following:

Story You sit waiting for shooting stars, where the waiting time for a star is distributed Expo(λ). You want to see n shooting stars before you go home. The total waiting time for the nth shooting star is Gamma(n, λ).

Example You are at a bank, and there are 3 people ahead of you. The serving time for each person is Exponential with mean 2 minutes. Only one person at a time can be served. The distribution of your waiting time until it's your turn to be served is Gamma$(3, \frac{1}{2})$.

Beta Distribution



Conjugate Prior of the Binomial In the Bayesian approach to statistics, parameters are viewed as random variables, to reflect our uncertainty. The *prior* for a parameter is its distribution before observing data. The *posterior* is the distribution for the parameter after observing data. Beta is the *conjugate* prior of the Binomial because if you have a Beta-distributed prior on p in a Binomial, then the posterior distribution on p given the Binomial data is also Beta-distributed. Consider the following two-level model:

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

Then after observing $X = x$, we get the posterior distribution

$$p | (X = x) \sim \text{Beta}(a + x, b + n - x)$$

Order statistics of the Uniform See *Order Statistics*.

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

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

This is known as the **bank–post office result**.

χ^2 (Chi-Square) Distribution

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

Story A Chi-Square(n) is the sum of the squares of n independent standard Normal r.v.s.

Properties and Representations

$$X \text{ is distributed as } Z_1^2 + Z_2^2 + \cdots + Z_n^2 \text{ for i.i.d. } Z_i \sim \mathcal{N}(0,1)$$

$$X \sim \text{Gamma}(n/2, 1/2)$$

Discrete Distributions

Distributions for four sampling schemes

	Replace	**No Replace**
Fixed # trials (n)	Binomial (Bern if $n = 1$)	HGeom
Draw until r success	NBin (Geom if $r = 1$)	NHGeom

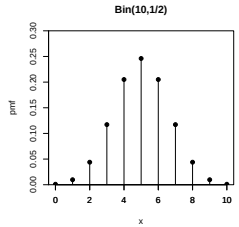
Bernoulli Distribution

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

Story A trial is performed with probability p of "success", and X is the indicator of success: 1 means success, 0 means failure.

Example Let X be the indicator of Heads for a fair coin toss. Then $X \sim \text{Bern}(\frac{1}{2})$. Also, $1 - X \sim \text{Bern}(\frac{1}{2})$ is the indicator of Tails.

Binomial Distribution



Let us say that X is distributed 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 is either a success or a failure, each with the same probability p of success. We can also write X as 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. Then

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

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 Bin(10, $\frac{3}{4}$).

Properties Let $X \sim \text{Bin}(n, p), Y \sim \text{Bin}(m, p)$ with $X \perp\!\!\!\perp Y$.

- **Redefine success** $n - X \sim \text{Bin}(n, 1 - p)$
- **Sum** $X + Y \sim \text{Bin}(n + m, p)$
- **Conditional** $X|(X + Y = r) \sim \text{HGeom}(n, m, r)$
- **Binomial-Poisson Relationship** Bin(n, p) is approximately Pois(λ) if p is small.
- **Binomial-Normal Relationship** Bin(n, p) is approximately $\mathcal{N}(np, np(1-p))$ if n is large and p is not near 0 or 1.

Geometric Distribution

Let us say that X is distributed 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 probability $\frac{1}{10}$ to catch Mew, the number of failed pokeballs will be distributed Geom($\frac{1}{10}$).

First Success Distribution

Equivalent to the Geometric distribution, except that it includes the first success in the count. This is 1 more than the number of failures. If $X \sim \text{FS}(p)$ then $E(X) = 1/p$.

Negative Binomial Distribution

Let us say that X is distributed NBin(r, p). We know the following:

Story X is the number of "failures" that we will have 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 NBin(3, 0.6).

Hypergeometric Distribution

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

Story In a population of w desired objects and b undesired objects, X is the number of "successes" we will have in a draw of n objects, without replacement. The draw of n objects is assumed to be a **simple random sample** (all sets of n objects are equally likely).

Examples Here are some HGeom examples.

- 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.
- The number of Aces in a 5 card hand.
- You have w white balls and b black balls, and you draw n balls. You will draw X white balls.
- You have w white balls and b black balls, and you draw n balls without replacement. The number of white balls in your sample is HGeom(w, b, n); the number of black balls is HGeom(b, w, n).
- **Capture-recapture** A forest has N elk, you capture n of them, tag them, and release them. Then you recapture a new sample of size m. How many tagged elk are now in the new sample? HGeom($n, N - n, m$)

Poisson Distribution

Let us say that X is distributed Pois(λ). 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, it is reasonable to model the number of accidents in a month at that intersection as Pois(2). Then the number of accidents that happen in two months at that intersection is distributed Pois(4).

Properties Let $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$, with $X \perp\!\!\!\perp Y$.

1. **Sum** $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. **Conditional** $X|(X + Y = n) \sim \text{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$
3. **Chicken-egg** If there are $Z \sim \text{Pois}(\lambda)$ items and we randomly and independently "accept" each item with probability p, then the number of accepted items $Z_1 \sim \text{Pois}(\lambda p)$, and the number of rejected items $Z_2 \sim \text{Pois}(\lambda(1-p))$, and $Z_1 \perp\!\!\!\perp Z_2$.

Multivariate Distributions

Multinomial Distribution

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

Story We have n items, which 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 $\vec{p} = (0.25, 0.25, 0.25, 0.25)$. Note that $X_1 + X_2 + \cdots + X_4 = 100$, and they are dependent.

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

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

Marginal PMF, Lumping, and Conditionals Marginally, $X_i \sim \text{Bin}(n, p_i)$ since we can define "success" to mean category i. If you lump together multiple categories in a Multinomial, then it is still Multinomial. For example, $X_i + X_j \sim \text{Bin}(n, p_i + p_j)$ for $i \neq j$ since we can define "success" to mean being in category i or j. Similarly, if $k = 6$ and we lump categories 1-2 and lump categories 3-5, then

$$(X_1 + X_2, X_3 + X_4 + X_5, X_6) \sim \text{Mult}_3(n, (p_1 + p_2, p_3 + p_4 + p_5, p_6))$$

Conditioning on some X_j also still gives a Multinomial:

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

Variances and Covariances We have $X_i \sim \text{Bin}(n, p_i)$ marginally, so $\text{Var}(X_i) = np_i(1-p_i)$. Also, $\text{Cov}(X_i, X_j) = -np_ip_j$ for $i \neq j$.

Multivariate Uniform Distribution

See the univariate Uniform for stories and examples. For the 2D Uniform on some region, probability is proportional to area. Every point in the support has equal density, of value $\frac{1}{\text{area of region}}$. For the 3D Uniform, probability is proportional to volume.

Multivariate Normal (MVN) Distribution

A vector $\vec{X} = (X_1, X_2, \ldots, X_k)$ is Multivariate Normal if every linear combination is Normally distributed, i.e., $t_1X_1 + t_2X_2 + \cdots + t_kX_k$ 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) entry is $\text{Cov}(X_i, X_j)$.

Properties The Multivariate Normal has the following properties.

- Any subvector is also MVN.
- If any two elements within an MVN are uncorrelated, then they are independent.
- The joint PDF of a Bivariate Normal (X, Y) with $\mathcal{N}(0, 1)$ marginal distributions and correlation $\rho \in (-1, 1)$ is

$$f_{X,Y}(x, y) = \frac{1}{2\pi\tau} \exp\left(-\frac{1}{2\tau^2}(x^2 + y^2 - 2\rho xy)\right),$$

with $\tau = \sqrt{1 - \rho^2}$.

Distribution Properties

Important CDFs

Standard Normal Φ

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

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

Convolutions of Random Variables

A convolution of n random variables is simply their sum. For the following results, let X and Y be *independent*.

1. $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2) \longrightarrow X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
2. $X \sim \text{Bin}(n_1, p)$, $Y \sim \text{Bin}(n_2, p) \longrightarrow X + Y \sim \text{Bin}(n_1 + n_2, p)$. $\text{Bin}(n, p)$ can be thought of as a sum of i.i.d. $\text{Bern}(p)$ r.v.s.
3. $X \sim \text{Gamma}(a_1, \lambda)$, $Y \sim \text{Gamma}(a_2, \lambda)$ $\longrightarrow X + Y \sim \text{Gamma}(a_1 + a_2, \lambda)$. $\text{Gamma}(n, \lambda)$ with n an integer can be thought of as a sum of i.i.d. $\text{Expo}(\lambda)$ r.v.s.
4. $X \sim \text{NBin}(r_1, p)$, $Y \sim \text{NBin}(r_2, p)$ $\longrightarrow X + Y \sim \text{NBin}(r_1 + r_2, p)$. $\text{NBin}(r, p)$ can be thought of as a sum of i.i.d. $\text{Geom}(p)$ r.v.s.
5. $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ $\longrightarrow X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Special Cases of Distributions

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. $\text{NBin}(1, p) \sim \text{Geom}(p)$

Inequalities

1. **Cauchy-Schwarz** $|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$
2. **Markov** $P(X \geq a) \leq \frac{E|X|}{a}$ for $a > 0$
3. **Chebyshev** $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$ for $E(X) = \mu, \text{Var}(X) = \sigma^2$
4. **Jensen** $E(g(X)) \geq g(E(X))$ for g convex; reverse if g is concave

Formulas

Geometric Series

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

$$1 + r + r^2 + \cdots = \frac{1}{1 - r} \text{ if } |r| < 1$$

Exponential Function (e^x)

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n$$

Gamma and Beta Integrals

You can sometimes solve complicated-looking integrals by pattern-matching to a gamma or beta integral:

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

Also, $\Gamma(a + 1) = a\Gamma(a)$, and $\Gamma(n) = (n - 1)!$ if n is a positive integer.

Euler's Approximation for Harmonic Sums

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \approx \log n + 0.577\ldots$$

Stirling's Approximation for Factorials

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

Miscellaneous Definitions

Medians and Quantiles Let X have CDF F. Then X has median m if $F(m) \geq 0.5$ and $P(X \geq m) \geq 0.5$. For X continuous, m satisfies $F(m) = 1/2$. In general, the ath quantile of X is $\min\{x : F(x) \geq a\}$; the median is the case $a = 1/2$.

log Statisticians generally use log to refer to natural log (i.e., base e).

i.i.d r.v.s Independent, identically-distributed random variables.

Example Problems

Contributions from Sebastian Chiu

Calculating Probability

A textbook has n typos, which are randomly scattered amongst its n pages, independently. 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 $\boxed{\left(1 - \frac{1}{n}\right)^n}$ probability that there are no typos on your page. For n large, this is approximately $e^{-1} = 1/e$.

Linearity and Indicators (1)

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 I_j be the indicator for the jth day being represented.

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

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

Linearity and Indicators (2)

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

Linearity and First Success

This problem is commonly known as the ***coupon collector problem****.* There are n coupon types. At each draw, you get a uniformly random coupon type. 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 $N = N_1 + \cdots + N_n$, where N_1 is the draws to get our first new coupon, N_2 is the *additional* draws needed to draw our second new coupon and so on. By the story of the First Success, $N_2 \sim \text{FS}((n-1)/n)$ (after collecting first coupon type, there's $(n-1)/n$ chance you'll get something new). Similarly, $N_3 \sim \text{FS}((n-2)/n)$, and $N_j \sim \text{FS}((n-j+1)/n)$. By linearity,

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

This is approximately $n(\log(n) + 0.577)$ by Euler's approximation.

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 are 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. One of these choices has all 3 of the Lyfts arriving first, so 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 (not counting the Ace)? **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 $I_j = 1$ says that j is before all 4 of the Aces in the deck. The probability that this occurs is $1/5$ by symmetry. Let X be the number of cards drawn before the first Ace. Then $X = I_1 + I_2 + \ldots + I_{48}$, where each indicator corresponds to one of the 48 non-Aces. Thus,

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

Minimum and Maximum of RVs

What is the CDF of the maximum of n independent Unif(0,1) random variables? **Answer:** Note that for r.v.s $X_1, X_2, \ldots, X_n$,

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

Similarly,

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

We will use this principle to find the CDF of $U_{(n)}$, where $U_{(n)} = \max(U_1, U_2, \ldots, U_n)$ and $U_i \sim \text{Unif}(0, 1)$ are i.i.d.

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

for $0 < a < 1$ (and the CDF is 0 for $a \leq 0$ and 1 for $a \geq 1$).

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's Law and Eve's Law

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 probability p of solving the cube, independently. Let T be the number of Rubik's Cubes he solves during a day. Find the mean and variance of T.
Answer: Note that $T|N \sim \text{Bin}(N,p)$. So by Adam's Law,

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

Similarly, by Eve's Law, we have that

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

MGF – Finding Moments

Find $E(X^3)$ for $X \sim \text{Expo}(\lambda)$ using the MGF of X. **Answer:** The MGF of an $\text{Expo}(\lambda)$ 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 nth moment of X, so we have $E(X^n) = \frac{n!}{\lambda^n}$ for all nonnegative integers n.

Markov chains (1)

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

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

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 with respect to $\vec{s}$. **Answer:** The equation $\vec{s}Q = \vec{s}$ says that

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

By solving this system of linear equations, we have

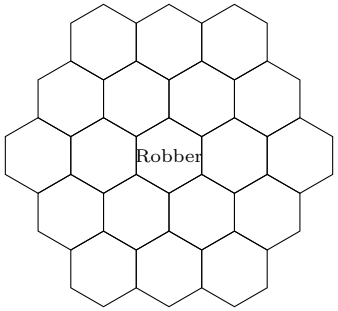
$$\boxed{\vec{s} = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)}$$

To show that the chain is reversible with respect to $\vec{s}$, we must show $s_i q_{ij} = s_j q_{ji}$ for all i, j. This is done if we can show $s_0 q_{01} = s_1 q_{10}$. And indeed,

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

Markov chains (2)

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:** $\boxed{\text{Yes to both.}}$ The Markov chain is irreducible because it can get from anywhere to anywhere else. The Markov chain is aperiodic because the robber can return back to a square in $2, 3, 4, 5, \ldots$ moves, and the GCD of those numbers is 1.

(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, in the long run? **Answer:** By the above, $\boxed{1/14}$.

(d) What is the expected amount of moves it will take for the robber to return to the center tile? **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 relevant events and random variables*. ("Let A be the event that I pick the fair coin"; "Let X be the number of successes.") Clear notion is important for clear thinking! Then decide what it is that you're supposed to be finding, in terms of your notation ("I want to find $P(X = 3|A)$"). Think about what type of object your answer should be (a number? A random variable? A PMF? A PDF?) and what it should be in terms of.
   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 counting principles if the naive definition of probability applies. Is the probability of the complement easier to find? Look for symmetries. Look for something to condition on, then apply Bayes' Rule or the Law of Total Probability.
3. **Finding the distribution of a random variable.** First make sure you need the full distribution not just the mean (see next item). 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 an r.v. with a known distribution, say $Y = g(X)$?
4. **Calculating expectation.** If it has a named distribution, check out the table of distributions. If it's a function of an r.v. with a named distribution, try LOTUS. If it's a count of something, try breaking it up into indicator r.v.s. If you can condition on something natural, consider using Adam's law.
5. **Calculating variance.** Consider independence, named distributions, and LOTUS. If it's a count of something, break it up into a sum of indicator r.v.s. If it's a sum, use properties of covariance. If you can condition on something natural, consider using Eve's Law.
6. **Calculating $E(X^2)$.** Do you already know $E(X)$ or $\text{Var}(X)$? Recall that $\text{Var}(X) = E(X^2) - (E(X))^2$. Otherwise try LOTUS.
7. **Calculating covariance.** Use the properties of covariance. If you're trying to find the covariance between two components of a Multinomial distribution, X_i, X_j, then the covariance is $-np_ip_j$ for $i \neq j$.
8. **Symmetry.** If $X_1, \ldots, X_n$ are i.i.d., consider using symmetry.
9. **Calculating probabilities of orderings.** Remember that all $n!$ ordering of i.i.d. continuous random variables $X_1, \ldots, X_n$ are equally likely.
10. **Determining independence.** There are several equivalent definitions. Think about simple and extreme cases to see if you can find a counterexample.
11. **Do a painful integral.** If your integral looks painful, see if you can write your integral in terms of a known PDF (like Gamma or Beta), and use the fact that PDFs integrate to 1?
12. **Before moving on.** Check some simple and extreme cases, check whether the answer seems plausible, check for biohazards.

Biohazards

Contributions from Jessy Hwang

1. **Don't misuse the naive 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}, even though the latter is six times more likely.
2. **Don't confuse unconditional, conditional, and joint probabilities.** In applying $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, it is *not* correct to say "$P(B) = 1$ because we know B happened"; $P(B)$ is the *prior* probability of B. Don't confuse $P(A|B)$ with $P(A, B)$.
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 $1/n^2$. Binomial and Hypergeometric are often confused; the trials are independent in the Binomial story and dependent in the Hypergeometric story.
4. **Don't forget to do sanity checks.** Probabilities must be between 0 and 1. Variances must be ≥ 0. Supports must make sense. PMFs must sum to 1. PDFs must integrate to 1.
5. **Don't confuse random variables, numbers, and events.** Let X be an r.v. Then $g(X)$ is an r.v. for any function g. 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 $g(E(X))$ 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.

6. **Don't confuse a random variable with its distribution.** To get the PDF of X^2, you can't just square the PDF of X. The right way is to use 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.

7. **Don't pull non-linear functions out of expectations.** $E(g(X))$ does not equal $g(E(X))$ in general. The St. Petersburg paradox is an extreme example. See also Jensen's inequality. The right way to find $E(g(X))$ is with LOTUS.

Recommended Resources

- Introduction to Probability Book (`http://bit.ly/introprobability`)
- Stat 110 Online (`http://stat110.net`)
- Stat 110 Quora Blog (`https://stat110.quora.com/`)
- Quora Probability FAQ (`http://bit.ly/probabilityfaq`)
- R Studio (`https://www.rstudio.com`)
- LaTeX File (`github.com/wzchen/probability_cheatsheet`)

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

Distributions in R

Command	What it does
`help(distributions)`	shows documentation on distributions
`dbinom(k,n,p)`	PMF $P(X=k)$ for $X \sim \text{Bin}(n,p)$
`pbinom(x,n,p)`	CDF $P(X \leq x)$ for $X \sim \text{Bin}(n,p)$
`qbinom(a,n,p)`	ath quantile for $X \sim \text{Bin}(n,p)$
`rbinom(r,n,p)`	vector of r i.i.d. $\text{Bin}(n,p)$ r.v.s
`dgeom(k,p)`	PMF $P(X=k)$ for $X \sim \text{Geom}(p)$
`dhyper(k,w,b,n)`	PMF $P(X=k)$ for $X \sim \text{HGeom}(w,b,n)$
`dnbinom(k,r,p)`	PMF $P(X=k)$ for $X \sim \text{NBin}(r,p)$
`dpois(k,r)`	PMF $P(X=k)$ for $X \sim \text{Pois}(r)$
`dbeta(x,a,b)`	PDF $f(x)$ for $X \sim \text{Beta}(a,b)$
`dchisq(x,n)`	PDF $f(x)$ for $X \sim \chi^2_n$
`dexp(x,b)`	PDF $f(x)$ for $X \sim \text{Expo}(b)$
`dgamma(x,a,r)`	PDF $f(x)$ for $X \sim \text{Gamma}(a,r)$
`dlnorm(x,m,s)`	PDF $f(x)$ for $X \sim \mathcal{LN}(m,s^2)$
`dnorm(x,m,s)`	PDF $f(x)$ for $X \sim \mathcal{N}(m,s^2)$
`dt(x,n)`	PDF $f(x)$ for $X \sim t_n$
`dunif(x,a,b)`	PDF $f(x)$ for $X \sim \text{Unif}(a,b)$

The table above gives R commands for working with various named distributions. Commands analogous to `pbinom`, `qbinom`, and `rbinom` work for the other distributions in the table. For example, `pnorm`, `qnorm`, and `rnorm` can be used to get the CDF, quantiles, and random generation for the Normal. For the Multinomial, `dmultinom` can be used for calculating the joint PMF and `rmultinom` can be used for generating random vectors. For the Multivariate Normal, after installing and loading the `mvtnorm` package `dmvnorm` can be used for calculating the joint PDF and `rmvnorm` can be used for generating random vectors.

Table of Distributions

Distribution	PMF/PDF and Support	Expected Value	Variance	MGF
Bernoulli Bern(p)	$P(X=1)=p$ $P(X=0)=q=1-p$	p	pq	$q+pe^t$
Binomial Bin(n,p)	$P(X=k)=\binom{n}{k}p^k q^{n-k}$ $k\in\{0,1,2,\dots n\}$	np	npq	$(q+pe^t)^n$
Geometric Geom(p)	$P(X=k)=q^k p$ $k\in\{0,\ 1,\ 2,\ \dots\}$	q/p	q/p^2	$\frac{p}{1-qe^t}$, $qe^t<1$
Negative Binomial NBin(r,p)	$P(X=n)=\binom{r+n-1}{r-1}p^r q^n$ $n\in\{0,\ 1,\ 2,\ \dots\}$	rq/p	rq/p^2	$(\frac{p}{1-qe^t})^r$, $qe^t<1$
Hypergeometric HGeom(w,b,n)	$P(X=k)=\binom{w}{k}\binom{b}{n-k}/\binom{w+b}{n}$ $k\in\{0,1,2,\dots,n\}$	$\mu=\frac{nw}{b+w}$	$\left(\frac{w+b-n}{w+b-1}\right)n\frac{\mu}{n}(1-\frac{\mu}{n})$	messy
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	$e^{t\mu+\frac{\sigma^2 t^2}{2}}$
Exponential Expo(λ)	$f(x)=\lambda e^{-\lambda x}$ $x\in(0,\infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-t}$, $t<\lambda$
Gamma Gamma(a,λ)	$f(x)=\frac{1}{\Gamma(a)}(\lambda x)^a e^{-\lambda x}\frac{1}{x}$ $x\in(0,\infty)$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$	$\left(\frac{\lambda}{\lambda-t}\right)^a$, $t<\lambda$
Beta Beta(a,b)	$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ $x\in(0,1)$	$\mu=\frac{a}{a+b}$	$\frac{\mu(1-\mu)}{(a+b+1)}$	messy
Log-Normal $\mathcal{LN}(\mu,\sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-(\log x-\mu)^2/(2\sigma^2)}$ $x\in(0,\infty)$	$\theta=e^{\mu+\sigma^2/2}$	$\theta^2(e^{\sigma^2}-1)$	doesn't exist
Chi-Square χ^2_n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$ $x\in(0,\infty)$	n	$2n$	$(1-2t)^{-n/2}$, $t<1/2$
Student-t t_n	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+x^2/n)^{-(n+1)/2}$ $x\in(-\infty,\infty)$	0 if $n>1$	$\frac{n}{n-2}$ if $n>2$	doesn't exist

Probability–the Science of Uncertainty and Data

by Fabián Kozynski

Probability

Probability models and axioms

Definition (Sample space) A sample space Ω is the set of all possible outcomes. The set's elements must be mutually exclusive, collectively exhaustive and at the right granularity.

Definition (Event) An event is a subset of the sample space. Probability is assigned to events.

Definition (Probability axioms) A probability law $\mathbb{P}$ assigns probabilities to events and satisfies the following axioms:

Nonnegativity $\mathbb{P}(A) \geq 0$ for all events A.

Normalization $\mathbb{P}(\Omega) = 1$.

(Countable) additivity For every sequence of events $A_1, A_2, \ldots$ such that $A_i \cap A_j = \varnothing$: $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$.

Corollaries (Consequences of the axioms)

- $\mathbb{P}(\varnothing) = 0$.
- For any finite collection of disjoint events $A_1, \ldots, A_n$, $\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i)$.
- $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$.
- $\mathbb{P}(A) \leq 1$.
- If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

Example (Discrete uniform law) Assume Ω is finite and consists of n equally likely elements. Also, assume that $A \subset \Omega$ with k elements. Then $\mathbb{P}(A) = \frac{k}{n}$.

Conditioning and Bayes' rule

Definition (Conditional probability) Given that event B has occurred and that $P(B) > 0$, the probability that A occurs is

$$\mathbb{P}(A|B) \triangleq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Remark (Conditional probabilities properties) They are the same as ordinary probabilities. Assuming $\mathbb{P}(B) > 0$:

- $\mathbb{P}(A|B) \geq 0$.
- $\mathbb{P}(\Omega|B) = 1$
- $\mathbb{P}(B|B) = 1$.
- If $A \cap C = \varnothing$, $\mathbb{P}(A \cup C|B) = \mathbb{P}(A|B) + \mathbb{P}(C|B)$.

Proposition (Multiplication rule)

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdots \mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Theorem (Total probability theorem) Given a partition $\{A_1, A_2, \ldots\}$ of the sample space, meaning that $\bigcup_i A_i = \Omega$ and the events are disjoint, and for every event B, we have

$$\mathbb{P}(B) = \sum_i \mathbb{P}(A_i)\mathbb{P}(B|A_i).$$

Theorem (Bayes' rule) Given a partition $\{A_1, A_2, \ldots\}$ of the sample space, meaning that $\bigcup_i A_i = \Omega$ and the events are disjoint, and if $\mathbb{P}(A_i) > 0$ for all i, then for every event B, the conditional probabilities $\mathbb{P}(A_i|B)$ can be obtained from the conditional probabilities $\mathbb{P}(B|A_i)$ and the initial probabilities $\mathbb{P}(A_i)$ as follows:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_j \mathbb{P}(A_j)\mathbb{P}(B|A_j)}.$$

Independence

Definition (Independence of events) Two events are independent if occurrence of one provides no information about the other. We say that A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Equivalently, as long as $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,

$$\mathbb{P}(B|A) = \mathbb{P}(B) \qquad \mathbb{P}(A|B) = \mathbb{P}(A).$$

Remarks

- The definition of independence is symmetric with respect to A and B.
- The product definition applies even if $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

Corollary If A and B are independent, then A and B^c are independent. Similarly for A^c and B, or for A^c and B^c.

Definition (Conditional independence) We say that A and B are independent conditioned on C, where $\mathbb{P}(C) > 0$, if

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C).$$

Definition (Independence of a collection of events) We say that events $A_1, A_2, \ldots, A_n$ are independent if for every collection of distinct indices $i_1, i_2, \ldots, i_k$, we have

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}).$$

Counting

This section deals with finite sets with uniform probability law. In this case, to calculate $\mathbb{P}(A)$, we need to count the number of elements in A and in Ω.

Remark (Basic counting principle) For a selection that can be done in r stages, with n_i choices at each stage i, the number of possible selections is $n_1 \cdot n_2 \cdots n_r$.

Definition (Permutations) The number of permutations (orderings) of n different elements is

$$n! = 1 \cdot 2 \cdot 3 \cdots n.$$

Definition (Combinations) Given a set of n elements, the number of subsets with exactly k elements is

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

Definition (Partitions) We are given an n–element set and nonnegative integers $n_1, n_2, \ldots, n_r$, whose sum is equal to n. The number of partitions of the set into r disjoint subsets, with the i^{th} subset containing exactly n_i elements, is equal to

$$\binom{n}{n_1, \ldots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}.$$

Remark This is the same as counting how to assign n distinct elements to r people, giving each person i exactly n_i elements.

Discrete random variables

Probability mass function and expectation

Definition (Random variable) A random variable X is a function of the sample space Ω into the real numbers (or $\mathbb{R}^n$). Its range can be discrete or continuous.

Definition (Probability mass funtion (PMF)) The probability law of a discrete random variable X is called its PMF. It is defined as

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) = x\}\right).$$

Properties

$p_X(x) \geq 0$, $\forall\, x$.

$\sum_x p_X(x) = 1$.

Example (Bernoulli random variable) A Bernoulli random variable X with parameter $0 \leq p \leq 1$ ($X \sim \text{Ber}(p)$) takes the following values:

$$X = \begin{cases} 1 & \text{w.p. } p, \\ 0 & \text{w.p. } 1-p. \end{cases}$$

An indicator random variable of an event ($I_A = 1$ if A occurs) is an example of a Bernoulli random variable.

Example (Discrete uniform random variable) A Discrete uniform random variable X between a and b with $a \leq b$ ($X \sim \text{Uni}[a, b]$) takes any of the values in $\{a, a+1, \ldots, b\}$ with probability $\frac{1}{b-a+1}$.

Example (Binomial random variable) A Binomial random variable X with parameters n (natural number) and $0 \leq p \leq 1$ ($X \sim \text{Bin}(n, p)$) takes values in the set $\{0, 1, \ldots, n\}$ with probabilities $p_X(i) = \binom{n}{i} p^i (1-p)^{n-i}$.
It represents the number of successes in n independent trials where each trial has a probability of success p. Therefore, it can also be seen as the sum of n independent Bernoulli random variables, each with parameter p.

Example (Geometric random variable) A Geometric random variable X with parameter $0 \leq p \leq 1$ ($X \sim \text{Geo}(p)$) takes values in the set $\{1, 2, \ldots\}$ with probabilities $p_X(i) = (1-p)^{i-1} p$.
It represents the number of independent trials until (and including) the first success, when the probability of success in each trial is p.

Definition (Expectation/mean of a random variable) The expectation of a discrete random variable is defined as

$$\mathbb{E}[X] \triangleq \sum_x x p_X(x).$$

assuming $\sum_x |x| p_X(x) < \infty$.

Properties (Properties of expectation)

- If $X \geq 0$ then $\mathbb{E}[X] \geq 0$.
- If $a \leq X \leq b$ then $a \leq \mathbb{E}[X] \leq b$.
- If $X = c$ then $\mathbb{E}[X] = c$.

Example Expected value of know r.v.

- If $X \sim \text{Ber}(p)$ then $\mathbb{E}[X] = p$.
- If $X = I_A$ then $\mathbb{E}[X] = \mathbb{P}(A)$.
- If $X \sim \text{Uni}[a, b]$ then $\mathbb{E}[X] = \frac{a+b}{2}$.
- If $X \sim \text{Bin}(n, p)$ then $\mathbb{E}[X] = np$.
- If $X \sim \text{Geo}(p)$ then $\mathbb{E}[X] = \frac{1}{p}$.

Theorem (Expected value rule) Given a random variable X and a function $g:\mathbb{R}\to\mathbb{R}$, we construct the random variable $Y=g(X)$. Then

$$\sum_y yp_Y(y)=\mathbb{E}[Y]=\mathbb{E}[g(X)]=\sum_x g(x)p_X(x).$$

Remark (PMF of $Y=g(X)$) The PMF of $Y=g(X)$ is $p_Y(y)=\sum_{x:g(x)=y}p_X(x)$.

Remark In general $g(\mathbb{E}[X])\neq\mathbb{E}[g(X)]$. They are equal if $g(x)=ax+b$.

Variance, conditioning on an event, multiple r.v.

Definition (Variance of a random variable) Given a random variable X with $\mu=\mathbb{E}[X]$, its variance is a measure of the spread of the random variable and is defined as

$$\mathrm{Var}(X)\triangleq\mathbb{E}\left[(X-\mu)^2\right]=\sum_x(x-\mu)^2p_X(x).$$

Definition (Standard deviation)

$$\sigma_X=\sqrt{\mathrm{Var}(X)}.$$

Properties (Properties of the variance)

- $\mathrm{Var}(aX)=a^2\,\mathrm{Var}(X)$, for all $a\in\mathbb{R}$.
- $\mathrm{Var}(X+b)=\mathrm{Var}(X)$, for all $b\in\mathbb{R}$.
- $\mathrm{Var}(aX+b)=a^2\,\mathrm{Var}(X)$.
- $\mathrm{Var}(X)=\mathbb{E}[X^2]-(\mathbb{E}[X])^2$.

Example (Variance of known r.v.)

- If $X\sim\mathrm{Ber}(p)$, then $\mathrm{Var}(X)=p(1-p)$.
- If $X\sim\mathrm{Uni}[a,b]$, then $\mathrm{Var}(X)=\frac{(b-a)(b-a+2)}{12}$.
- If $X\sim\mathrm{Bin}(n,p)$, then $\mathrm{Var}(X)=np(1-p)$.
- If $X\sim\mathrm{Geo}(p)$, then $\mathrm{Var}(X)=\frac{1-p}{p^2}$

Proposition (Conditional PMF and expectation, given an event) Given the event A, with $\mathbb{P}(A)>0$, we have the following

- $p_{X|A}(x)=\mathbb{P}(X=x|A)$.
- If A is a subset of the range of X, then:
  $p_{X|A}(x)\triangleq p_{X|\{X\in A\}}(x)=\begin{cases}\frac{1}{\mathbb{P}(A)}p_X(x), & \text{if } x\in A,\\ 0, & \text{otherwise.}\end{cases}$
- $\sum_x p_{X|A}(x)=1$.
- $\mathbb{E}[X|A]=\sum_x xp_{X|A}(x)$.
- $\mathbb{E}[g(X)|A]=\sum_x g(x)p_{X|A}(x)$.

Proposition (Total expectation rule) Given a partition of disjoint events $A_1,\dots,A_n$ such that $\sum_i\mathbb{P}(A_i)=1$, and $\mathbb{P}(A_i)>0$,

$$\mathbb{E}[X]=\mathbb{P}(A_1)\mathbb{E}[X|A_1]+\cdots+\mathbb{P}(A_n)\mathbb{E}[X|A_n].$$

Definition (Memorylessness of the geometric random variable) When we condition a geometric random variable X on the event $X>n$ we have memorylessness, meaning that the "remaining time" $X-n$, given that $X>n$, is also geometric with the same parameter. Formally,

$$p_{X-n|X>n}(i)=p_X(i).$$

Definition (Joint PMF) The joint PMF of random variables $X_1,X_2,\dots,X_n$ is $p_{X_1,X_2,\dots,X_n}(x_1,\dots,x_n)=\mathbb{P}(X_1=x_1,\dots,X_n=x_n)$.

Properties (Properties of joint PMF)

- $\sum_{x_1}\cdots\sum_{x_n}p_{X_1,\dots,X_n}(x_1,\dots,x_n)=1$.
- $p_{X_1}(x_1)=\sum_{x_2}\cdots\sum_{x_n}p_{X_1,\dots,X_n}(x_1,x_2,\dots,x_n)$.
- $p_{X_2,\dots,X_n}(x_2,\dots,x_n)=\sum_{x_1}p_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n)$.

Definition (Functions of multiple r.v.) If $Z=g(X_1,\dots,X_n)$, where $g:\mathbb{R}^n\to\mathbb{R}$, then $p_Z(z)=\mathbb{P}(g(X_1,\dots,X_n)=z)$.

Proposition (Expected value rule for multiple r.v.) Given $g:\mathbb{R}^n\to\mathbb{R}$,

$$\mathbb{E}[g(X_1,\dots,X_n)]=\sum_{x_1,\dots,x_n}g(x_1,\dots,x_n)p_{X_1,\dots,X_n}(x_1,\dots,x_n).$$

Properties (Linearity of expectations)

- $\mathbb{E}[aX+b]=a\mathbb{E}[X]+b$.
- $\mathbb{E}[X_1+\cdots+X_n]=\mathbb{E}[X_1]+\cdots+\mathbb{E}[X_n]$.

Conditioning on a random variable, independence

Definition (Conditional PMF given another random variable) Given discrete random variables X,Y and y such that $p_Y(y)>0$ we define

$$p_{X|Y}(x|y)\triangleq\frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

Proposition (Multiplication rule) Given jointly discrete random variables X,Y, and whenever the conditional probabilities are defined,

$$p_{X,Y}(x,y)=p_X(x)p_{Y|X}(y|x)=p_Y(y)p_{X|Y}(x|y).$$

Definition (Conditional expectation) Given discrete random variables X,Y and y such that $p_Y(y)>0$ we define

$$\mathbb{E}[X|Y=y]=\sum_x xp_{X|Y}(x|y).$$

Additionally we have

$$\mathbb{E}[g(X)|Y=y]=\sum_x g(x)p_{X|Y}(x|y).$$

Theorem (Total probability and expectation theorems) If $p_Y(y)>0$, then

$$p_X(x)=\sum_y p_Y(y)p_{X|Y}(x|y),$$
$$\mathbb{E}[X]=\sum_y p_Y(y)\mathbb{E}[X|Y=y].$$

Definition (Independence of a random variable and an event) A discrete random variable X and an event A are independent if $\mathbb{P}(X=x\text{ and }A)=p_X(x)\mathbb{P}(A)$, for all x.

Definition (Independence of two random variables) Two discrete random variables X and Y are independent if $p_{X,Y}(x,y)=p_X(x)p_Y(y)$ for all x,y.

Remark (Independence of a collection of random variables) A collection $X_1,X_2,\dots,X_n$ of random variables are independent if

$$p_{X_1,\dots,X_n}(x_1,\dots,x_n)=p_{X_1}(x_1)\cdots p_{X_n}(x_n),\ \forall\,x_1,\dots,x_n.$$

Remark (Independence and expectation) In general, $\mathbb{E}[g(X,Y)]\neq g(\mathbb{E}[X],\mathbb{E}[Y])$. An exception is for linear functions: $\mathbb{E}[aX+bY]=a\mathbb{E}[X]+b\mathbb{E}[Y]$.

Proposition (Expectation of product of independent r.v.) If X and Y are discrete independent random variables,

$$\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y].$$

Remark If X and Y are independent, $\mathbb{E}[g(X)h(Y)]=\mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]$.

Proposition (Variance of sum of independent random variables) IF X and Y are discrete independent random variables,

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y).$$

Continuous random variables

PDF, Expectation, Variance, CDF

Definition (Probability density function (PDF)) A probability density function of a r.v. X is a non-negative real valued function f_X that satisfies the following

- $\int_{-\infty}^{\infty}f_X(x)\mathrm{d}x=1$.
- $\mathbb{P}(a\le X\le b)=\int_a^b f_X(x)\mathrm{d}x$ for some random variable X.

Definition (Continuous random variable) A random variable X is continuous if its probability law can be described by a PDF f_X.

Remark Continuous random variables satisfy:

- For small $\delta>0$, $\mathbb{P}(a\le X\le a+\delta)\approx f_X(a)\delta$.
- $\mathbb{P}(X=a)=0$, $\forall a\in\mathbb{R}$.

Definition (Expectation of a continuous random variable) The expectation of a continuous random variable is

$$\mathbb{E}[X]\triangleq\int_{-\infty}^{\infty}xf_X(x)\mathrm{d}x.$$

assuming $\int_{-\infty}^{\infty}|x|f_X(x)\mathrm{d}x<\infty$.

Properties (Properties of expectation)

- If $X\ge0$ then $\mathbb{E}[X]\ge0$.
- If $a\le X\le b$ then $a\le\mathbb{E}[X]\le b$.
- $\mathbb{E}[g(X)]=\int_{-\infty}^{\infty}g(x)f_X(x)\mathrm{d}x$.
- $\mathbb{E}[aX+b]=a\mathbb{E}[X]+b$.

Definition (Variance of a continuous random variable) Given a continuous random variable X with $\mu=\mathbb{E}[X]$, its variance is

$$\mathrm{Var}(X)=\mathbb{E}\left[(X-\mu)^2\right]=\int_{-\infty}^{\infty}(x-\mu)^2f_X(x)\mathrm{d}x.$$

It has the same properties as the variance of a discrete random variable.

Example (Uniform continuous random variable) A Uniform continuous random variable X between a and b, with $a<b$, $(X\sim\mathrm{Uni}(a,b))$ has PDF

$$f_X(x)=\begin{cases}\frac{1}{b-a}, & \text{if } a<x<b,\\ 0, & \text{otherwise.}\end{cases}$$

We have $\mathbb{E}[X]=\frac{a+b}{2}$ and $\mathrm{Var}(X)=\frac{(b-a)^2}{12}$.

Example (Exponential random variable) An Exponential random variable X with parameter $\lambda > 0$ ($X \sim Exp(\lambda)$) has PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

We have $E[X] = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$.

Definition (Cumulative Distribution Function (CDF)) The CDF of a random variable X is $F_X(x) = \mathbb{P}(X \leq x)$.
In particular, for a continuous random variable, we have

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\mathrm{d}x,$$

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}.$$

Properties (Properties of CDF)

- If $y \geq x$, then $F_X(y) \geq F_X(x)$.
- $\lim_{x \to -\infty} F_X(x) = 0$.
- $\lim_{x \to \infty} F_X(x) = 1$.

Definition (Normal/Gaussian random variable) A Normal random variable X with mean μ and variance $\sigma^2 > 0$ ($X \sim \mathcal{N}(\mu, \sigma^2)$) has PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

We have $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$.

Remark (Standard Normal) The standard Normal is $\mathcal{N}(0, 1)$.

Proposition (Linearity of Gaussians) Given $X \sim \mathcal{N}(\mu, \sigma^2)$, and if $a \neq 0$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
Using this $Y = \frac{X-\mu}{\sigma}$ is a standard gaussian.

Conditioning on an event, and multiple continuous r.v.

Definition (Conditional PDF given an event) Given a continuous random variable X and event A with $P(A) > 0$, we define the conditional PDF as the function that satisfies

$$\mathbb{P}(X \in B|A) = \int_B f_{X|A}(x)\mathrm{d}x.$$

Definition (Conditional PDF given $X \in A$) Given a continuous random variable X and an $A \subset \mathbb{R}$, with $P(A) > 0$:

$$f_{X|X\in A}(x) = \begin{cases} \frac{1}{\mathbb{P}(A)} f_X(x), & x \in A, \\ 0, & x \notin A. \end{cases}$$

Definition (Conditional expectation) Given a continuous random variable X and an event A, with $P(A) > 0$:

$$\mathbb{E}[X|A] = \int_{-\infty}^{\infty} f_{X|A}(x)\mathrm{d}x.$$

Definition (Memorylessness of the exponential random variable) When we condition an exponential random variable X on the event $X > t$ we have memorylessness, meaning that the "remaining time" $X - t$ given that $X > t$ is also geometric with the same parameter i.e.,

$$\mathbb{P}(X - t > x|X > t) = \mathbb{P}(X > x).$$

Theorem (Total probability and expectation theorems) Given a partition of the space into disjoint events $A_1, A_2, \ldots, A_n$ such that $\sum_i \mathbb{P}(A_i) = 1$ we have the following:

$$F_X(x) = \mathbb{P}(A_1)F_{X|A_1}(x) + \cdots + \mathbb{P}(A_n)F_{X|A_n}(x),$$
$$f_X(x) = \mathbb{P}(A_1)f_{X|A_1}(x) + \cdots + \mathbb{P}(A_n)f_{X|A_n}(x),$$
$$\mathbb{E}[X] = \mathbb{P}(A_1)\mathbb{E}[X|A_1] + \cdots + \mathbb{P}(A_n)\mathbb{E}[X|A_n].$$

Definition (Jointly continuous random variables) A pair (collection) of random variables is jointly continuous if there exists a joint PDF $f_{X,Y}$ that describes them, that is, for every set $B \subset \mathbb{R}^n$

$$\mathbb{P}((X, Y) \in B) = \iint_B f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y.$$

Properties (Properties of joint PDFs)

- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}y$.
- $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \left[\int_{-\infty}^{y} f_{X,Y}(u, v)\mathrm{d}v \right] \mathrm{d}u$.
- $f_{X,Y}(x) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x\, \partial y}$.

Example (Uniform joint PDF on a set S) Let $S \subset \mathbb{R}^2$ with area $s > 0$, then the random variable (X, Y) is uniform over S if it has PDF

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{s}, & (x, y) \in S, \\ 0, & (x, y) \notin S. \end{cases}$$

Conditioning on a random variable, independence, Bayes' rule

Definition (Conditional PDF given another random variable) Given jointly continuous random variables X, Y and a value y such that $f_Y(y) > 0$, we define the conditional PDF as

$$f_{X|Y}(x|y) \triangleq \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Additionally we define $\mathbb{P}(X \in A|Y = y) \int_A f_{X|Y}(x|y)\mathrm{d}x$.

Proposition (Multiplication rule) Given jointly continuous random variables X, Y, whenever possible we have

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

Definition (Conditional expectation) Given jointly continuous random variables X, Y, and y such that $f_Y(y) > 0$, we define the conditional expected value as

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\mathrm{d}x.$$

Additionally we have

$$\mathbb{E}[g(X)|Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\mathrm{d}x.$$

Theorem (Total probability and total expectation theorems)

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x|y)\mathrm{d}y,$$
$$\mathbb{E}[X] = \int_{-\infty}^{\infty} f_Y(y)\mathbb{E}[X|Y = y]\mathrm{d}y.$$

Definition (Independence) Jointly continuous random variables X, Y are independent if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all x, y.

Proposition (Expectation of product of independent r.v.) If X and Y are independent continuous random variables,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Remark If X and Y are independent, $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]$.

Proposition (Variance of sum of independent random variables) If X and Y are independent continuous random variables,

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

Proposition (Bayes' rule summary)

- For X, Y discrete: $p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)}$.
- For X, Y continuous: $f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)}$.
- For X discrete, Y continuous: $p_{X|Y}(x|y) = \frac{p_X(x)f_{Y|X}(y|x)}{f_Y(y)}$.
- For X continuous, Y discrete: $f_{X|Y}(x|y) = \frac{f_X(x)p_{Y|X}(y|x)}{p_Y(y)}$.

Derived distributions

Proposition (Discrete case) Given a discrete random variable X and a function g, the r.v. $Y = g(X)$ has PMF

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x).$$

Remark (Linear function of discrete random variable) If $g(x) = ax + b$, then $p_Y(y) = p_X\left(\frac{y-b}{a}\right)$.

Proposition (Linear function of continuous r.v.) Given a continuous random variable X and $Y = aX + b$, with $a \neq 0$, we have

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right).$$

Corollary (Linear function of normal r.v.) If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$, with $a \neq 0$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Example (General function of a continuous r.v.) If X is a continuous random variable and g is any function, to obtain the pdf of $Y = g(X)$ we follow the two-step procedure:

1. Find the CDF of Y: $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y)$.
2. Differentiate the CDF of Y to obtain the PDF: $f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y}$.

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

$$f_Y(y) = f_X(h(y)) \left| \frac{\mathrm{d}h}{\mathrm{d}y}(y) \right|.$$

where $h = g^{-1}$ in the interval where g is monotonic.

Sums of independent r.v., covariance and correlation

Proposition (Discrete case) Let X, Y be discrete independent random variables and $Z = X + Y$, then the PMF of Z is

$$p_Z(z) = \sum_x p_X(x) p_Y(z - x).$$

Proposition (Continuous case) Let X, Y be continuous independent random variables and $Z = X + Y$, then the PDF of Z is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \mathrm{d}x.$$

Proposition (Sum of independent normal r.v.) Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ independent. Then $Z = X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

Definition (Covariance) We define the covariance of random variables X, Y as

$$\mathrm{Cov}(X, Y) \triangleq \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right].$$

Properties (Properties of covariance)

- If X, Y are independent, then $\mathrm{Cov}(X, Y) = 0$.
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- $\mathrm{Cov}(aX + b, Y) = a\,\mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$.
- $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Proposition (Variance of a sum of r.v.)

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j).$$

Definition (Correlation coefficient) We define the correlation coefficient of random variables X, Y, with $\sigma_X, \sigma_Y > 0$, as

$$\rho(X, Y) \triangleq \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Properties (Properties of the correlation coefficient)

- $-1 \leq \rho \leq 1$.
- If X, Y are independent, then $\rho = 0$.
- $|\rho| = 1$ if and only if $X - \mathbb{E}[X] = c(Y - \mathbb{E}[Y])$.
- $\rho(aX + b, Y) = \mathrm{sign}(a)\rho(X, Y)$.

Conditional expectation and variance, sum of random number of r.v.

Definition (Conditional expectation as a random variable) Given random variables X, Y the conditional expectation $\mathbb{E}[X|Y]$ is the random variable that takes the value $\mathbb{E}[X|Y = y]$ whenever $Y = y$.

Theorem (Law of iterated expectations)

$$\mathbb{E}\left[\mathbb{E}[X|Y]\right] = \mathbb{E}[X].$$

Definition (Conditional variance as a random variable) Given random variables X, Y the conditional variance $\mathrm{Var}(X|Y)$ is the random variable that takes the value $\mathrm{Var}(X|Y = y)$ whenever $Y = y$.

Theorem (Law of total variance)

$$\mathrm{Var}(X) = \mathbb{E}\left[\mathrm{Var}(X|Y)\right] + \mathrm{Var}\left(\mathbb{E}[X|Y]\right).$$

Proposition (Sum of a random number of independent r.v.)
Let N be a nonnegative integer random variable.
Let $X, X_1, X_2, \ldots, X_N$ be i.i.d. random variables.
Let $Y = \sum_i X_i$. Then

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X],$$
$$\mathrm{Var}(Y) = \mathbb{E}[N]\,\mathrm{Var}(X) + (\mathbb{E}[X])^2\,\mathrm{Var}(N).$$

Convergence of random variables

Inequalities, convergence, and the Weak Law of Large Numbers

Theorem (Markov inequality) Given a random variable $X \geq 0$ and, for every $a > 0$ we have

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Theorem (Chebyshev inequality) Given a random variable X with $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$, for every $\epsilon > 0$ we have

$$\mathbb{P}\left(|X - \mu| \geq \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2}.$$

Theorem (Weak Law of Large Number (WLLN)) Given a sequence of i.i.d. random variables $\{X_1, X_2, \ldots\}$ with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, we define

$$M_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

for every $\epsilon > 0$ we have

$$\lim_{n \to \infty} \mathbb{P}\left(|M_n - \mu| \geq \epsilon\right) = 0.$$

Definition (Convergence in probability) A sequence of random variables $\{Y_i\}$ converges in probability to the random variable Y if

$$\lim_{n \to \infty} \mathbb{P}\left(|Y_i - Y| \geq \epsilon\right) = 0,$$

for every $\epsilon > 0$.

Properties (Properties of convergence in probability) If $X_n \to a$ and $Y_n \to b$ in probability, then

- $X_n + Y_n \to a + b$.
- If g is a continuous function, then $g(X_n) \to g(a)$.
- $\mathbb{E}[X_n]$ does not always converge to a.

The Central Limit Theorem

Theorem (Central Limit Theorem (CLT)) Given a sequence of independent random variables $\{X_1, X_2, \ldots\}$ with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, we define

$$Z_n = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu).$$

Then, for every z, we have

$$\lim_{n \to \infty} \mathbb{P}(Z_n \leq z) = \mathbb{P}(Z \leq z),$$

where $Z \sim \mathcal{N}(0, 1)$.

Corollary (Normal approximation of a binomial) Let $X \sim Bin(n, p)$ with n large. Then S_n can be approximated by $Z \sim \mathcal{N}\left(np, np(1 - p)\right)$.

Remark (De Moivre-Laplace 1/2 approximation) Let $X \sim Bin$, then $\mathbb{P}(X = i) = \mathbb{P}\left(i - \frac{1}{2} \leq X \leq i + \frac{1}{2}\right)$ and we can use the CLT to approximate the PMF of X.

Statistics Cheat Sheet

Ch 1: Overview & Descriptive Stats

Populations, Samples and Processes

`Population:` well-defined collection of objects
`Sample:` a subset of the population
`Descriptive Stats:` summarize & describe features of data
`Inferential Stats:` generalizing from sample to population
`Probability:` bridge btwn descriptive & inferential techniques.
In probability, properties of the population are assumed known & questions regarding a sample taken from the population are posed and answered.
`Discrete and Continuous Variables:` A numerical variable is *discrete* if its set of possible values is at most countable.
A numerical value is *continuous* if its set of possible values is an uncountable set.
Probability: pop → sample
Stats: sample → pop

Measures of Location

For observations $x_1, x_2, \cdots, x_n$

`Sample Mean`	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
`Sample Median`	$\tilde{x} = (\frac{n+1}{2})^{\text{nth}}$ observation
`Trimmed Mean`	btwn $\tilde{x}$ and $\bar{x}$, compute by removing smallest and largest observations

Measures of Variability

`Range`	=lgst-smllst observation
`Sample Variance,` σ^2	$= \frac{\sum(x_i - \bar{x}))^2}{n-1} = \frac{S_{xx}}{n-1}$
S_{xx}	$= \sum x_i^2 - \frac{(\sum x_i)^2}{n}$
`Sample Standard Deviation,` σ	$= \sqrt{\sigma^2}$

Box Plots

Order the n observations from small to large. Separate the smallest half from the largest (If n is odd then $\tilde{x}$ is in both halves). The lower fourth is the median of the smallest half (upper fourth..largest..). A measure of the spread that is resistant to outliers is the *fourth spread* f_s given by $f_s =$ upper fourth- lower fourth. Box from lower to upper fourth with line at median. Whiskers from smallest to largest x_i.

Ch 2: Probability

Sample Space and Events

`Experiment`	activity with uncertain outcome
`Sample Space`($\mathcal{S}$)	the set of all possible outcomes
`Event`	any collection of outcomes in $\mathcal{S}$

Axioms, Interpretations and Properties of Probability

Given an experiment and a sample space $\mathcal{S}$, the objective probability is to assign to each event A a number $P(A)$, called the probability of event A, which will give a precise measure of the chance that A will occur. Behaves very much like norm.

Axioms & Properties of Probability:

1. $\forall A \in \mathcal{S}, 0 \le P(A) \le 1$
2. $P(\mathcal{S}) = 1$
3. If $A_1, A_2, \ldots$ is an infinite collection of disjoint events, $P(A_1 \cup A_2 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$
4. $P(\emptyset) = 0$
5. $\forall A, P(A) + P(A') = 1$ from which $P(A) = 1 - P(A')$
6. For any two events $A, B \in \mathcal{S}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
7. For any three events $A, B, C \in \mathcal{S}, P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Equally Likely Outcomes : $P(A) = \frac{N(A)}{N}$

Counting Techniques

`Product Rule for Ordered k-Tuples:` If the first element can be selected in n_1 ways, the second in n_2 ways and so on, then there are $n_1 n_2 \cdots n_k$ possible k-tuples.
`Permutations:` An ordered subset. The number of permutations of size k that can be formed from a set of n elements is $P_{k,n}$

$$P_{k,n} = (n)(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

`Combinations:` An unordered subset.

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

Conditional Probability

$P(A|B)$ is the conditional probability of A given that the event B has occurred. B is the conditioning event.

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

Multiplication Rule: $P(A \cap B) = P(A|B) \cdot P(B)$

Baye's Theorem

Let $A_1, A_2, \ldots, A_k$ be disjoint and exhaustive events (that partition the sample space). Then for any other event B
$P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k)$
$\quad = \sum_{i=1}^{k} P(B|A_i)P(A_i)$

Independence

Two events A and B are `independent` if $P(A|B) = P(A)$ and are `dependent` otherwise.
A and B are `independent` iff $P(A \cap B) = P(A) \cdot P(B)$ and this can be generalized to the case of n mutually independent events.

Random Variables

`Random Variable`: any function $X : \Omega \to \mathbb{R}$

`Prob Dist.`:	describes how the probability of Ω is distributed along the range of X
`Discrete rv`:	rv whose domain is at most countable
`Continuous rv`:	rv whose domain is uncountable and where $\forall c \in \mathbb{R}, P(X = c) = 0$
`Bernoulli rv`:	discrete rv whose range is $\{0, 1\}$

The *probability distribution of X* says how the total probability of 1 is distributed among the various possible X values.

1. Distributions

Discrete RVs

Probabilities assigned to various outcomes in $\mathcal{S}$ in turn determine probabilities associated with the values of any particular rv X.

`Probability Mass Fxn/Probability Distribution,(pmf)`:

$$p(x) = P(X = x) = P(\forall w \in \mathcal{W} : X(w) = x)$$

Gives the probability of observing $w \in \mathcal{W} : X(w) = x$
The conditions $p(x) \ge 0$ and $\sum_{\text{all possible x}} p(x) = 1$ are required for any pmf.
`parameter:` Suppose p(x) depends on a quantity that can be assigned any one of a number of possible values, with each different value determining a different probability distribution. Such a quantity is called a parameter of distribution. The collection of all probability distributions for different values of the parameter is called a family of probability distributions.

Cumulative Distribution Function

(To compute the probability that the observed value of X will be at most some given x)
`Cumulative Distribution Function(cdf):` $F(x)$ of a discrete rv variable X with pmf $p(x)$ is defined for every number x by

$$F(x) = P(X \le x) = \sum_{y : y \le x} p(y)$$

For any number x, $F(x)$ is the probability that the observed value of X will be at most x.
For discrete rv, the graph of $F(x)$ will be a step function- jump at every possible value of X and flat btwn possible values.
For any two number a and b with $a \le b$:
$P(a \le X \le b) = F(b) - F(a^-)$
$P(a < X \le b) = F(b) - F(a)$
$P(a \le X \le a) = F(a) - F(a^-) = p(a)$
$P(a < X < b) = F(b^-) - F(a)$
(where a^- is the largest possible X value strictly less than a)
Taking $a = b$ yields $P(X = a) = F(a) - F(a-1)$ as desired.
`Expected value or Mean Value`

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

Describes where the probability distribution is centered and is just a weighted average of the possible values of X given their distribution. However, the sample average of a sequence of X values may not settle down to some finite number (harmonic series) but will tend to grow without bound. Then the distribution is said to have a *heavy tail*. Can make if difficult to make inferences about μ.
`The Expected Value of a Function`: Sometimes interest will focus on the expected value of some function $h(x)$ rather than on just $E(x)$.
If the RV X has a set of possible values D and pmf $p(x)$, then the expected value of any function $h(x)$, denoted by $E[h(X)]$ or $\mu_{h(X)}$ is computed by

$$E[h(X)] = \sum_{D} h(x) \cdot p(x)$$

`Properties of Expected Value`:

$$E(aX+b) = a \cdot E(X) + b$$

`Variance of X`: Let X have pmf $p(x)$ and expected value μ. Then the $V(X)$ or σ_X^2 is

$$V(X) = \sum_D (x-\mu)^2 \cdot p(x) = E[(X-\mu)^2]$$

The standard deviation (SD) of X is $\sigma = \sqrt{\sigma}$
Alternatively,

$$V(X) = \sigma^2 = \big[\sum_D x^2 \cdot p(x)\big] - \mu^2 = E(X^2) - [E(X)]^2$$

`Properties of Variance`

1. $V(aX+b) = a^2 \cdot \sigma^2$
2. In particular, $\sigma_{aX} = |a| \cdot \sigma_x$
3. $\sigma_{X+b} = \sigma_X$

Continuous RVs

Probabilities assigned to various outcomes in $\mathcal{S}$ in turn determine probabilities associated with the values of any particular rv X. Recall: an rv X is continuous if its set of possible values is uncountable and if $P(X=c) = 0 \quad \forall c.//$

`Probability Density Fxn/Probability Distribution,(pdf)`: $\forall a, b \in \mathbb{R}, a \leq b$

$$P(\forall w \in \mathcal{W} : a \leq X(w) \leq b) = \int_a^b f(x)dx$$

Gives the probability that X takes values between a and b. The conditions $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x) = 1$ are required for any pdf.

`Cumulative Distribution Function(cdf):`

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

For any number $x, F(x)$ is the probability that the observed value of X will be at most x.
By the continuity arguments for continuous RVs we have that

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a < X < b)$$

Other probabilities can be computed from the cdf $F(x)$:

$$P(X > a) = 1 - F(a)$$

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

Furthermore, if X is a cont rv with pdf $f(x)$ and cdf $F(x)$, then at every x at which $F'(x)$ exists, $F'(x) = f(x)$.
`Median`($\tilde{\mu}$): is the 50th percentile st $F(\tilde{\mu}) = .5$. That is half the area under the density curve. For a symmetric curve, this is the point of symmetry.

`Expected/Mean Value`(μ or $E(X)$): of cont rv with pdf $f(x)$

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

If X is a cont rv with pdf $f(x)$ and $h(X)$ is any function of X then

$$E[h(X)] = \mu = \int_{-\infty}^{\infty} h(x) \cdot f(x)dx$$

`Variance:` of a cont rv X with pdf $f(x)$ and mean value μ is

$$\sigma_x^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx = E[(X-\mu)^2]$$

Alternatively,

$$V(X) = E(X^2) - [E(X)]^2$$

Discrete Distributions

The Binomial Probability Distribution

1) The experiment consists of n *trials* where n is fixed
2) Each trial can result in either success (S) or failure (F)
3) The trials are independent
4) The probability of success $P(S)$ is constant for all trials
Note that in general if the sampling is without replacement, the experiment will not yield independent trials. However, if the sample size (number of trials) n is at most 5% of the population, then the experiment can be analyzed as though it were exactly a binomial experiment.
`Binomial rv X:` = no of S's among the n trials
`pmf of a Binomial RV:`,

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x} \quad : x = 0, 1, 2, \ldots$$

`cdf for Binomal RV:` Values in Tble A.1

$$B(x;n,p) = P(X \leq x) = \sum_{y=0}^{x} b(y;n,p)$$

`Mean & Variance of X` If $X \sim Bin(n,p)$ then

$$E(X) = np \quad V(X) = npq$$

Negative Binomial Distribution

1) The experiment consists of independent trials
2) Each trial can result in either Success(S) or Failure(F)
3) The probability of success is constant from trial to trial
4) The experiment continues until a total of r successes have been observed, where r is a specified integer.
`RV Y:` = the no of trials before the rth success.
`Negative Binomial rv:` $X = Y - r$ the number of failures that precede the rth success. In contrast to the binomial rv, the number of successes is fixed while the number of trials is random.
`pmf of the negative binomial rv` : with parameters r = number of S's and $p = P(S)$ is

$$nb(x;r,p) = \binom{x+r-1}{r-1} p^r (1-p)^x \qquad x = 0, 1, 2, \ldots$$

`Mean & Variance of negative binomial rv X:` with pmf $nb(x;r,p)$

$$E(X) = \frac{r(1-p)}{p} \quad V(X) = \frac{r(1-p)}{p^2}$$

Geometric Distribution

`RV X:` = the no of trials before the 1st success.
`pmf of the geometric rv` :

$$p(x) = q^{x-1}p$$

$$E(X) = \sum xq^{x-1}p = 1/p$$

The Poisson Probability Distribution

Useful for modeling rare events
1) independent: no of events in an interval is independent of no of events in another interval
2) Rare: no 2 events at once
3) Constant Rate: average events/unit time is constant ($\mu > 0$)
`RV X=` no of occurrence in unit time interval
`Possion distribution/ Poisson pmf`: of a random variable X with parameter $\mu > 0$ where

$$p(x;\mu) = \frac{e^{-\mu} \cdot \mu^x}{x!} \quad x = 0, 1, 2, \ldots$$

`Binomial Approximation`: Suppose that in the binomial pmf $b(x;n,p)$, we let $n \to \infty$ and $p \to 0$ in such a way that np approaches a value $\mu > 0$. Then $b(x;n,p) \to p(x;\mu)$.
That is to say that in any binomial experiment in which n(the number of trials) is large and p(the probability of success) is small, then $b(x;n,p) \approx p(x;\mu)$, where $\mu = np$.
`Mean and Variance of X:` If X has probability distribution with parameter μ, then $E(X) = V(X) = \mu$

Continuous Distributions

The Normal Distribution, $X \sim N(\mu, \sigma^2)$

`PDF:` with parameters μ and σ where $-\infty < \mu < \infty$ and $0 < \sigma$

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty$$

We can then easily show that $E(X) = \mu$ and $V(X) = \sigma^2$.
`Standard Normal Distribution:` The specific case where $\mu = 0$ and $\sigma = 1$. Then

$$\text{pdf}: \quad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad \text{cdf}: \quad \Phi(z) = \int_{-\infty}^{z} \phi(u)du$$

`Standardization:` Suppose that $X \sim N(\mu, \sigma^2)$. Then

$$Z = (X-\mu)/\sigma$$

transforms X into standard units. Indeed $Z \sim N(0,1)$.

$$P(a \leq X \leq b) = P\Big(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\Big) = \Phi\Big(\frac{b-\mu}{\sigma}\Big) - \Phi\Big(\frac{a-\mu}{\sigma}\Big)$$

`Independence:` If $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$ and X and Y are independent, then $X \pm Y \sim N(\mu_x \pm \mu_y, \sigma_x^2 + \sigma_y^2)$
NOTE: By symmetry of the standard normal distribution, it follows that $\Phi(-z) = 1 - \Phi(z) \quad \forall z \in \mathbb{R}$

`Normal Approx to Binomial Dist:` Let $X \sim Bin(n,p)$. As long as a binomial histogram is not too skewed, Binomial probabilities can be well approximated by normal curve areas.

$$P(X \le x) = B(x;n,p) \approx \Phi\left(\frac{x+0.5-np}{\sqrt{np(1-p)}}\right)$$

As a rule, the approx is adequate provided that both $np \ge 10$ and $n(1-p) \ge 10$.

The Exponential Distribution, $X \sim Exp(\lambda)$

Model for lifetime of firms/products/humans

`Exponential Distribution:` A cont rv X has exp distribution if its pdf is given by

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \ge 0 \quad \lambda > 0$$

$F(x,\lambda)=$ $\quad P(X \le x) = 1 - e^{\lambda x} \quad x \ge 0$
$E(X)=$ $\quad 1/\lambda$
$V(X)=$ $\quad 1/\lambda^2$
`Memoryless Prop`: $P(X > a+x | X > a) = P(X > x)$ for $x \in D, a > 0$

Note: If Y is an rv distributed as a Poisson $p(y;\lambda)$, then the time between consecutive Poisson events is distributed as an exponential rv with parameter λ

Joint Probability Dist

`Joint Range:` Let $X : \mathcal{S} \to \mathbb{D}_1$ and $Y : \mathcal{S} \to \mathbb{D}_2$ be 2 rvs with a common sample space. We define the joint range of the vector (X,Y) of the form

$$\mathbb{D} = \mathbb{D}_1 \times \mathbb{D}_2 = \{(x,y) : x \in \mathbb{D}_1, y \in \mathbb{D}_2\}$$

`Random Vector:` A 2-D random vector(X,Y) is a function from $\mathcal{S} \to \mathbb{R}^2$. It is defined $\forall \omega \in \mathcal{S}$ such that

$$(X,Y)(\omega) = (X(\omega), Y(\omega)) = (x,y) \in \mathbb{D}$$

`Joint Probability Mass Fxn:` For two discrete rv's X and Y. The joint pmf of (X,Y) is defined $\forall (x,y) \in \mathbb{D}$

$$p(x_i, y_j) = P(X = x_i, Y = y_j)$$

It must be that $p(x,y) \ge 0$ and $\sum_i \sum_j p(x_i,y_j) = 1$.

`Marginal Prob Mass Fxn:` of X and of Y, denoted $p_X(x)$ and $p_Y(y)$ respectively,

$$p_X(x) = \sum_{y : p(x,y) > 0} p(x,y) \quad \forall x \in \mathbb{D}_1$$

`Joint Probability Density Fxn:` For two continuous rv's X and Y. The joint pdf of (X,Y) is defined $\forall A \subseteq \mathbb{R}^2$

$$P((X,Y) \in A) = \iint_A f(x,y)dxdy$$

It must be that $f(x,y) \ge 0$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$.
Note also that this integration is commutative.

`Marginal Prob Density Fxn:` of X and of Y, denoted $f_X(x)$ and $f_Y(y)$ respectively,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \forall x \in \mathbb{D}_1$$

Note that if $f(x,y)$ is the joint density of the random vector (X,Y) and $A \in \mathbb{R}^2$ is of the form $A = [a,b] \times [c,d]$ we have that

$$P((X,Y) \in A) = \int_c^d \int_a^b f(x,y)dxdy = \int_a^b \int_c^d f(x,y)dxdy$$

`Independence:` Two rvs are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad f(x,y) = f_X(x)f_Y(y)$$

`Conditional Distribution(discrete):` For two discrete rv's X and Y with joint pmf $p(x_i,y_j)$ and marginal X pmf $p_X(x)$, then for any realized value x in the range of X, the conditional mass function of Y, given that $X = x$ is

$$p_{Y|X}(y|x) = \frac{p(x_i,y_j)}{p_X(x)}$$

`Conditional Distribution(cont):` For two continuous rv's X and Y with joint pdf $f(x,y)$ and marginal X pdf $f_X(x)$, then for any realized value x in the range of X, the conditional density function of Y, given that $X = x$ is

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

Expected Values, Covariance & Correlation

`Expected value`: The expected value of a function $h(X,Y)$ of two jointly distributed random variables is

$$E(g(X,Y)) = \sum_{x \in \mathbb{D}_1} \sum_{y \in \mathbb{D}_2} g(x,y)p(x,y)$$

and can be generalized to the continuous case with integrations.//

`Covariance:` Measures the strength of the relation btwn 2 RVs, however very

$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$

Shortcut Formula:

$$Cov(X,Y) = E(XY) - \mu_x\mu_y$$

The defect of the covariance however is that its value depends critically on the units of measurement.

`Correlation:` Cov after standardization. Helps interpret Cov.

$$\rho = \rho_{X,Y} = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

Has the property that $Corr(aX+b, cY+d) = Corr(X,Y)$ and that for any rvs X, Y $-1 \le \rho \le 1$.
Note also that ρ is independent of units, the larger $|\rho|$ the stronger the linear association, considered strong linear relationship if $|\rho| \ge 0.8$.
Caution though: if X and Y are independent then $\rho = 0$ but $\rho = 0$ does not imply that X,Y are independent.
Also that $\rho = 1$ or -1 iff $Y = aX + b$ for some a, b with $a \ne 0$.

`Statistic:` Any quantity whose value can be calculated with sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, a statistic is a random variable and will be denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.

`Sampling Distribution:` probability distribution of a statistic, it describes how the statis- tic varies in value across all samples that might be selected

Stats & Their Distributions

Fxns of Observed Sample Observ

`Obs Sample Mean` $\bar{x} = \frac{1}{n}\sum x_i$
`Obs Sample Var` $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$
`Obs Sample Max` $x_{(n)} = max(x_i)$

A statistic is a random variable and the most common are listed above.

`Simple Random Samples:` The random variables $X_1, \ldots, X_n$ are said to form a simple random sample of size n if each X_i is an independent random variable, every X_i has the same probability distribution.

`Sampling Distribs:` Every statistic has a probability distribution (a pmt or pdf) which we call its sampling distribution. To determine its distrib can be hard but we use simulations and the CLT to do so.

`Simulation Experiments:` we must specify the statistic of interest, the population distribution, the sample size(n) and the number of samples (k). Use a computer to simulate each different simple random sample, construct a histogram which will give approx sampling distribution of the statistic.

The Dist % Sample Mean

Prop: Let $X_1, \ldots, X_n$ be a simple random sample from a distribution with mean μ and variance σ^2. Then $E(\bar{X}) = \mu_{\bar{X}} = \mu$ and $V(\bar{X}) = \sigma^2_{\bar{X}} = \sigma^2/n$. Also if $S_n = X_1 + \ldots + X_n$ then $E(S_n) = n\mu$ and $V(S_n) = n\sigma^2$.

Prop: Let $X_1, \ldots, X_n$ be a simple random sample from a normal distribution with mean μ and variance σ^2. Then for any n, $\bar{X}$ is normal distributed with mean μ and variance σ^2/n. Also S_n is normal distributed with mean $n\mu$ and variance $n\sigma^2$.

Prop: Let $X_1, \ldots, X_n$ be a simple random sample from Bernoulli(p), then $S_n \sim$ Binomial(n,p).

Distribution of The Sample Mean $\bar{X}$

Let $X_1, \ldots, X_n$ be a simple random sample from a distribution with mean μ and variance σ^2. Then $E(\bar{X}) = \mu_{\bar{X}} = \mu$ and $V(\bar{X}) = \sigma^2_{\bar{X}} = \sigma^2/n$

The standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ is often called the *standard error of the mean.*

For a NORMAL random sample with the same mean and std as above, then for any n, $\bar{X}$ is normally distributed with the same mean and std.

`Central Limit Theorem:` Let $X_1, \ldots, X_n$ be a simple random sample from a distribution with mean μ and variance σ^2. Then if n is sufficiently large, $\bar{X}$ has approximately a normal dis with mean μ and variance σ^2/n. Also S_n is normal distributed with mean $n\mu$ and variance $n\sigma^2$. No matter which population we sample from, the probability histogram of the sample mean follow closely a normal curve when n is sufficiently large. *Rule of thumb: if $n \ge 30$ CLT can be used.* It follows from CLT that is $X \sim Bin(n,p)$ and n is large, then n can be distributed by a $N(np, npq)$.

Dist of a Linear Combination

`Linear Comb:` Let $X_1, \ldots, X_n$ be a collxn of n random variables and let $a_1 \ldots a_n$ be n numerical constants. Then the random variable $Y = a_1X_1 + \ldots a_nX_n$ is a linear comb of the $X_i's$.

1. Regardless of whether the $X_i's$ are independent or not
$$E(Y) = a_1E(X_1) + \ldots + a_nE(X_n) = a_1\mu_1 + \ldots + a_n\mu_n$$
2. If $X_1, \ldots, X_n$ are independent
$$V(Y) = V(a_1X_1 + \ldots a_nX_n) + a_1^2\sigma_1^2 + \ldots$$
3. For any $X_1, \ldots, X_n$,
$$V(Y) = \sum_{i=1}\sum_{j=1} a_ia_jCov(X_i, X_j)$$
4. If $X_1, \ldots, X_n$ are independent, normally distributed rvs, then any linear combination of the rvs also has a normal distribution- as does their difference.

$E(X_1 - X_2) = E(X_1) - E(X_2), \forall X, Y$ while
$V(X_1 - X_2) = V(X_1) + V(X_2)$ if X1, X2 independent,

2. Estimators

`Parameter of Interest` (θ) true yet unknown pop parameter
`Point Estimate:` ($\hat{\theta}$) Our guess for θ based on sample data
`Point Estimator:` ($\hat{\theta}$) statistic selected to get a sensible pt est
A sensible way to quantity the idea of $\hat{\theta}$ being close to θ is to consider the least squared error $(\hat{\theta} - \theta)^2$. A good measure of the accuracy is the expected or mean square error MSE = $E[(\hat{\theta} - \theta)^2]$. It is often not possible to find the estimator with the smallest MSE so we often restrict our attention to *unbiased* estimators and find the best estimator of this group.
`Unbiased:` Pt Est $\hat{\theta}$ if $E(\hat{\theta}) = \theta$ for all θ.
Then $\hat{\theta}$ has a prob distribution that is always "centered" at the true θ value.
When choosing estimators, select the unbiased and the one that has the minimum variance.

Estimators

-When $X \sim Bin(n, p)$, the sample proportion $\hat{p} = X/n$ is an unbiased est of p.
- Let $X_1, \ldots, X_n$ be a SRS from a distribution with mean μ and variance σ^2.Then $\hat{\sigma}^2 = S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$ is unbiased for σ^2.
-Let $X_1, \ldots, X_n$ be a SRS from a distribution with mean μ, then $\bar{X}$ is MVUE for μ.
`Standard Error:` of an estimator is its standard deviation $\sigma_{\hat{\theta}} - \sqrt{V(\hat{\theta})}$
`Estimated Standard Error:` If the standard error itself involves unknown parameters whose values can be estimated, substitution of these estimates into $\sigma_{\hat{\theta}}$ yields $\sigma\hat{\theta} = s_{\hat{\theta}}$.

Method of Moments

Let $X_1, \ldots, X_n$ be a SRS from a pdf $f(x)$. For $k = 1, 2, \ldots$ the kth population moment, or kth moment of the distribution $f(x)$, is $E(X^k)$. The kth sample moment is $(1/n)\sum_{i=1}^n X_i^k$.
Let $X_1, \ldots, X_n$ be a SRS from a distribution with pdf $f(x; \theta_1 \ldots \theta_m)$ where θ_i's are unknown. Then the moment estimators $\hat{\theta}_i$'s are obtained from the first m sample moments to the corresponding first m population moments and solving for the θ_i's.

Maximum Likelihood Estimator

Works best when the sample size is large!
Let $X_1, \ldots, X_n$ have joint pmf or pdf
$$f(x_1, \ldots, x_n; \theta_1 \ldots \theta_m)$$
where the θ_i's have unknown values.
When $x_1, \ldots, x_n$ are observed sample values, the above is considered a fxn of the θ_i's and is called the **likelihood function**.
The maximum likelihood estimates (mles) $\hat{\theta}_i$'s are those θ_i's that maximize the likelihood function such that
$$f(x_1, \ldots, x_n; \hat{\theta_1} \ldots \hat{\theta_m}) \geq f(x_1, \ldots, x_n; \theta_1 \ldots \theta_m) \quad \forall \theta_1 \ldots \theta_m$$
When $X_1, \ldots, X_n$ substituted in, the **maximum likelihood estimators** result.

3. Confidence Intervals

Tests in a single sample

When measuring n random variables $Y_i \sim i.i.d.$

Hypotheses about the population mean $E[Y_i]$
`Z-test` (when $n > 40$ or if normality with known variances could be assumed)
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
`CI for Normal Population:` A $100(1-\alpha)\%$ CI for the mean μ of a population when σ is known is
$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$
`T -test` (normality must be assured; for large n this is the same as the z-test). When $\bar{X}$ is the sample mean of a SRS of size n from a Normal(μ, σ^2) population then the RV
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$
has a probability distribution-t with n-1 degrees of freedom.
Note: the density of t_ν is symmetric around 0. t_ν is more spread out than a normal, indeed the few dof the more spread. When dof is large (< 40), the t and normal curve are close. In addition we have that
$$P(|\frac{\bar{X} - \mu}{S/\sqrt{n}}| \leq t_{\alpha/2, n-1}) = 1 - \alpha$$
As a result, the $(1-\alpha)100\%$ CI for the population mean μ under the normal model is
$$\bar{X} \pm t_{\alpha/2, n-1}\frac{S}{\sqrt{n}}$$
Note that here we make the assumption that the observations are realizations of a SRS from a Normal distribution with unknown mean and variance.// `Large Sample Test` for the population proportion (proportions are just means; only valid for $np_0 \geq 10$ and $n(1-p_0) \geq 10$. The $(1-\alpha)$ confidence interval for a population mean μ is
$$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$
For a population proportion
$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \qquad \hat{p} = \bar{X}$$
Hypotheses about the population variance $V[X_i]$
The $(1-\alpha)100\%$ CI for the variance σ^2 of a normal population has a lower limit:
$$(n-1)s^2/\chi^2_{\alpha/2, n-1}$$
and Upper limit:
$$(n-1)s^2/\chi^2_{1-\alpha/2, n-1}$$
A confidence interval for σ has lower and upper limits that are the square roots of the corresponding limits in the interval for σ^2. An upper or a lower confidence bound results from replacing $\alpha/2$ with α in the corresponding limit of the CI.

When measuring two variables for each unit $(X_i, Y_i) \sim i.i.d.$

`Paired t-test` about the difference of population means:
`Test about parameters` β_1 and β_0

Tests in two non-paired, independent samples

4. Hypothesis Testing

In it hard to example the evidence of such a strong count as a lucky draw. The p-value or observed significance level determines whether or not a hypothesis will be rejected- the smaller it is, the stronger evidence against he null hypothesis. The plausibility of statistical models determined by the null hypothesis is based on the sample data and their distributions. The idea is that the null is not rejected unless it is testified implausible overwhelmingly by data.

`Possible Errors:` Type I: reject the null hypothesis when it is true; Type II: fail to reject the null even though it is false.

Power Function

For a given test with critical or rejection region $\{x : T(x) \geq c\}$, the power function is defined as

$$\phi(\theta) = P(T(X_1, \ldots, X_n) \geq c|\theta) = P(T \geq c|\theta)$$

In other words, $\phi(\theta)$ represents the *probability of rejection* H_0 if a particular θ were the true value of parameter of the pmt or pdf $f(x;\theta)$.
In other words, if H_0 is true, $\phi(\theta)$ = Probability of type 1 error. If H_0 is false, $\phi(\theta)$ = 1- Probability of type 2 error.
A court trial, where the null hypothesis is "not guilty" unless there is convincing evidence against it. The aim or purpose of court hearings (collecting data) is to establish the assertion of "guilty" rather than to prove "innocence."
P-value (or observed signicance level) is the probability, calculated assuming that H0 is true, of obtaining a value of the test statistic at least as contradictory to H0 as the value calculated from the available sample. It is also the smallest significance level at which one can reject H0.
In other words, suppose we have observed a realization $x_{obs} = (x1, \ldots, xn)$ of our random sample $X_1, \cdots, X_n \sim f(x, \theta)$. We wish to investigate the compatibility of the null hypothesis, with the observed data. We do so by comparing the probability distribution of the test statistic $T(X1, \cdots, Xn)$ with its observed value $t_{obs} = T(x_{obs})$, assuming H0 to be true. As a measure of compatibility, we calculate

$$p(xobs) = \text{p-value} = P(T(X1, \ldots, X_n) \geq t_{obs}|H0)$$

In general, report the p-value. When it is less than 5% or 1 %, the result is statistically significant.

Hypotheses and Test Procedures

Statistical hypothesis(hypothesis) is a claim or assertion about the value of a single parameter, about the values of several parameters, or about the form of an entire population distribution.
In any hypothesis-testing problem, there are two contradictory hypotheses under consideration.
The **null hypothesis**, denoted H_0 is the claim that is initially assumed to be true (the "prior belief" claim). Often called the hypothesis of no change (from current opinion) and will generally be stated as an equality claim, equal to the *null value*. The **alternative hypothesis** or researcher's hypothesis, denoted by H_a is the assertion that is contradictory to H_0. The alt hypothesis is often the claim that the researcher would really like to validate.
The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0, we will continue to believe in the plausibility of the null hypothesis. The two possible conclusions from a hypothesis-testing analysis are then reject H_0 or fail to reject H_0.
A **test of hypotheses** is a method for using sample data to decide whether the null hypothesis should be rejected.
A **test procedure** is a rule based on sample data, for deciding whether to reject H_0. A procedure has 2 consitutents: 1) a test static, or function of the sample data used to make a decision and 2) a rejection region consisting of those x values for which H_0 will be rejected in favor of H_a.
A test procedure is specified by the following:

1. A **test statistic**, a function of the sample data on which the decision (reject H_0 or do not reject H_0) is to be based
2. A **rejection region**, the set of all test statistic values for which H_0 will be rejected. The basis for choosing a rejection region lies in consideration of the errors that one might be faced with in drawing a conclusion.

The null hypothesis will then be rejected if and only if the observed or computed test statistic value falls in the rejection region.
A **type I error** consists of rejecting the null hypothesis H_0 when it is true- a false negative. A **type II error** involves not rejecting H_0 when H_0 is false- a false positive.
In the best of all possible worlds, test procedures for which neither type of error is possible could be developed. However, this ideal can be achieved only by basing a decision on an examination of the entire population. The difficulty with using a procedure based on sample data is that because of sampling variability, an unrepresentative sample may result, e.g., a value of $\bar{X}$ that is far from μ or a value of $\hat{p}$ that differs considerably from p.
Suppose an experiment and a sample size are fixed and a test statistic is chosen. Then decreasing the size of the rejection region to obtain a smaller value of α results in a larger value of β for any particular parameter value consistent with H_a. In other words, once the test statistic and n are fixed, there is no rejection region that will simultaneously make both α and all β's small. A region must be chosen to effect a compromise between α and β.

Tests About a Population Mean

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

α = P(H_0 is rejected when H_0 is true) = false negative = $P(\bar{X} \leq 70.8$ when $\bar{X} \sim$ normal with $\mu_{\bar{X}} = 75, \sigma_{\bar{X}} = 1.8) = P(Z \geq c$ null when $Z \sim N(0,1))$
β = P(H_0 is accepted when H_0 is false) = false positive = $P(\bar{X} > 70.8$ when $\bar{X} \sim$ normal with $\mu_{\bar{X}} = 72, \sigma_{\bar{X}} = 1.8)$

Tests about a Population Mean

Case1: A Normal Population with a Known σ

Assuming that the sample mean $\bar{X}$ has a normal distribution with $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. When H_0 is true, $\mu_{\bar{X}} = \mu_0$. Consider now the statistic Z obtained by standardizing $\bar{X}$ under the assumption that H_0 is true:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

5. Simple Linear Regression and Correlation

Common theme: to study the relationships among variables.

Model and Summary Statistics

`Bivariate Data:` $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$
`Generic Pair` (X, Y) X- predictor, independent variable, covariate
`Simple Linear Regression:` $Y = \beta_0 + \beta_1 x + \varepsilon$
`Betas` regression coeffs, ε measurement error, cannot be explained by x
The ith observation is given by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and we further assume that ε_i are iid $N(0, \sigma^2)$
`Conditional Expected Value:` For the linear model we have that $E(Y|x) = E(\beta_0 + \beta_1 x + \varepsilon_0) = \beta_0 + \beta_1 x$ which is the average for the group with covariate $\sim x$
`Conditional Standard Deviation:` Similarly we have that $V(Y|x) = \sigma^2$ which is the variance for the group with covariate $\sim x$
`Summary Stats x:` $\bar{x}$ and $SD_x = \sqrt{\frac{S_{xx}}{n-1}}$ or $S_{xx} = \sum(x_i - \bar{x})^2$
`Sum Stats y:` $\bar{y}$ and $SD_y = \sqrt{\frac{S_{yy}}{n-1}}$ or $S_{yy} = \sum(y_i - \bar{y})^2$
`Strength of Linear Assoc:` $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ the sample correlation coeff.

$$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

`Purpose of the Regression:` To quantify the contribution of the predictors $X_1 \ldots X_p$ on the outcome of Y, given $(x_1, \ldots, x_p)$ predict the mean response, quantify the uncertainty in this prediction (with standard error/confidence interval), extrapolate

Estimation of Model Parameters

Data are modeled as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1 \ldots n \quad \varepsilon_i \sim N(0, \sigma^2)$$

How to find good estimates for β_0 & β_1?
- The error between y_i and $\beta_0 + \beta_1 x$ is ε_i and we want to minimize the total "loss"
-In the case of squared-error loss functions, the total loss is $\sum \varepsilon_i^2$
-To minimize, take partial derivatives of SSE wrt each β and set each to zero. Then solve the system of linear equations for each β. In this case

$$\hat{\beta_1} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}} = r\frac{SD_x}{SD_y}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

`Fitted Values:` $\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$ The value y_i predicted based on x_i
`Residuals:` $\hat{\varepsilon_i} = y_i - \hat{y_i}$ Difference between predicted and actual y
`Resdal Sum of Squares:` SSE = SSE $(\hat{\beta_0}, \hat{\beta_1}) = \sum \hat{\varepsilon_i}^2$
`Regression Line:` $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$ Used to predict the mean response $\hat{y}$ for a given x

Estimating σ^2

$$\sigma^2 = \frac{1}{n-2}\sum \hat{\varepsilon}^2 = SSE/2$$

It can be shown that $SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = S_{yy}(1-r^2)$ and hence

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{n-1}{n-2}} SD_y \sqrt{1-r^2}$$

which is smaller that SD_y- the regression has decreased uncertainty about y.

Goodness of fit

`Sum of squares due to regression (SSR)`

$$SS_{reg} = S_{yy} - SSE$$

`Coeff of Determination` R^2: Percentage of variability of Y explained by the regression on X. The larger it is, the better the fit.

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2$$

Inference for Model Parameters

$$\hat{\beta_1} = \frac{S_{xy}}{S_{xx}} \quad \hat{\beta_0} = \bar{y} - \hat{\beta_1} x$$

$$SE(hat)(\hat{\beta_1}) = \hat{\sigma}/\sqrt{S_{xx}} \quad SE(hat)(\hat{\beta_0}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

where the T-statistic is:

$$T = \frac{\hat{\beta_1} - \beta_1}{\hat{SE}(\hat{\beta_1})}$$

`Standard Errors:` SInce the estimators are linear in Y
`Confidence Intervals:` $\hat{\beta_0} \pm t_{\alpha/2, n-2}$

6. Goodness of Fit

Condition: for each cell, the expected count is greater than five
Multinomial dist: probability weights on discrete, unordered possible outcomes
`Homogeneity:` Along the rows we have diff populations and columns are difference categories.
H0: proportion of individuals in category j is the same for each population and that is true for every category. $p_{1j} = \ldots = p_{Ij}$ for $j = 1 \ldots J$
Estimated expected: $\hat{e_{ij}} = \frac{\text{(ith row total)(jth column total)}}{n}$ Test Statistic:

$$\chi^2 = \sum \frac{(ob - estex)^2}{estex} = \sum\sum \frac{(n_{ij} - \hat{e_{ij}})^2}{\hat{e_{ij}}}$$

Rejection Region: $\chi^2 \geq \chi^2_{\alpha(I-1)(J-1)}$
`Independence:` Only one population but looking at the relationship btwn 2 different factors. Each individual in one category associated with first factor and one category associated with second factor.
H0: The null hypothesis here says that an individuals category with respect to factor 1 is independent of the category with respect to factor 2. In symbols, this becomes $p_{ij} = p_i p_j \forall (i,j)$.
Test Statistic, RR and Condition: Same as above
State the uncertainty in a particular estimate of ours.

Basics

The actual sample observations $x1, \ldots, xn$ are assumed to be the result of a random sample $X1, \ldots, Xn$ from a normal distribution with mean value μ and standard deviation σ. We know then(from Ch5) that $\bar{X} \sim N(\mu, \sigma^2/n)$. Standardizing yields

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Obtain an inequality such as

$$P(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95$$

and we manipulate the inequality so that it apprears in the form $l \leq \mu \leq u$ where l,u involve factors save μ.
This interval we now describe is random since the endpoints involve a random variable and centered at $\bar{X}$. It says the probability is .95 that the random interval includes or covers the true value of μ. The confidence level 95% is not so much a statement about any particular interval, instead it pertains to what would happen if a very large number of like intervals were to be constructed using the same CI formula.
`CI for Normal Population:` A $100(1-\alpha)$% CI for the mean μ of a population when σ is known is

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

or equivalently,

$$\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$$

`Necc Sample Size:` for a CI to have width w is $n = (2z_{\alpha/2} \cdot \frac{\sigma}{w})^2$
Note that for sufficiently large n, σ is replaced by S, the sample variance.
`General Large-Sample CI`: Suppose that $\hat{\theta}$ is an estimator approx normal, unbiased, and has an expression for $\sigma_{\hat{\theta}}$. Then standardizing yields

$$P(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}) \approx 1 - \alpha$$

Population Mean (if variance unknown)

With 95% chance the random interval covers μ, population mean.
`Interpretation:` When the estimator is replaced by an estimate, the random interval becomes a realized interval. The word confidence refers to the procedure. If we repeat the experiment many times and construct 95% confidence intervals int he same manner, about 95% of them cover the unknown, but fixed, μ. We don't know whether the current interval covers μ or not but we know that of all the intervals ever constructed 95% will cover.

General Confidence Intervals

When the sample size is large (> 40), the $(1-\alpha)$ confidence interval for a population mean μ is

$$\bar{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$

For a population proportion

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \qquad \hat{p} = \bar{X}$$

Steps for calculating Confidence Intervals

1) Find an RV having an (approximately) known distribution
2) Cut off tails, that is, select a confidence level $(1-\alpha)$
3) Solve the equation to obtain confidence intervals- isolate the population mean in an approbate string of inequalities.

Intervals Based on a Normal Population

When the sample size is small, we can no longer use the CLT. But maybe we can assume that the data comes from a normal population. In that case we need to account for the uncertainty in estimating σ but by how much?
`T-Statistic:` When $\bar{X}$ is the sample mean of a SRS of size n from a Normal(μ, σ^2) population then the RV

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a probability distribution-t with n-1 degrees of freedom.
Note: the density of t_ν is symmetric around 0. t_ν is more spread out than a normal, indeed the few dof the more spread. When dof is large (< 40), the t and normal curve are close. In addition we have that

$$P(|\frac{\bar{X} - \mu}{S/\sqrt{n}}| \leq t_{\alpha/2, n-1}) = 1 - \alpha$$

As a result, the $(1-\alpha)$% CI for the population mean μ under the normal model is

$$\bar{X} \pm t_{\alpha/2, n-1}\frac{S}{\sqrt{n}}$$

Note that here we make the assumption that the observations are realizations of a SRS from a Normal distribution with unknown mean and variance.

One-Sided Confidence Bounds

`Lower Confidence Bound:` When n is large, then

$$P\left(\frac{\bar{X} - \mu}{S/\sqrt{n}} \leq z_\alpha\right) = 1 - \alpha$$

and solve to find the $(1-\alpha)$ confidence bound $\bar{X} - z_\alpha \frac{s}{\sqrt{n}}$.
`Upper Confidence Bound:` With $(1-\alpha)$ confidence, μ is bounded by $\bar{X} + z_\alpha \frac{s}{\sqrt{n}}$
Note that when n is small, replace z_α by $t_{\alpha, n-1}$.

CI for the Variance of a Normal Population

`Theorem:` Let $X_1, ..., X_n$ be a SRS from a Normal(μ, σ^2) population, where both parameters are unknown. The RV

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum^n (X_i - X)^2}{\sigma^2}$$

has a probability distribution called the χ^2 distribution with n-1 dof.
The density of chi is always positive and has long upper tails. As n increases, the densities become more symmetric.
Furthermore, we have that

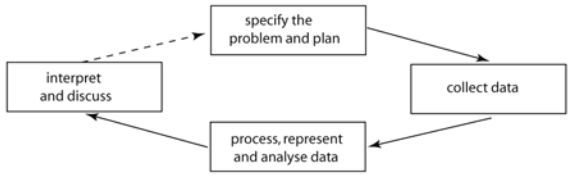
$$P\left(\chi_{1-\alpha/2, n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\alpha/2, n-1}\right) = 1 - \alpha$$

Hence, the $(1-\alpha)$ CI for the population variance σ^2 under the normal model is

$$\left[\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}}, \frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}}\right]$$

The statistical problem solving cycle

Data are numbers in context and the goal of statistics is to get information from those data, usually through *problem solving*. A procedure or paradigm for statistical problem solving and scientific enquiry is illustrated in the diagram. The dotted line means that, following discussion, the problem may need to be re-formulated and at least one more iteration completed.



Descriptive statistics

Given a sample of n observations, $x_1, x_2, \ldots, x_n$, we define the **sample mean** to be

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

and the *corrected* sum of squares by

$$S_{xx} = \sum (x_i - \bar{x})^2 \equiv \sum x_i^2 - n\bar{x}^2 \equiv \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$\frac{S_{xx}}{n}$ is sometimes called the *mean squared deviation*. An **unbiased estimator** of the population variance, σ^2, is $s^2 = \frac{S_{xx}}{(n-1)}$. The **sample standard deviation** is s. In calculating s^2, the divisor $(n-1)$ is called the **degrees of freedom (df)**. Note that s is also sometimes written $\hat{\sigma}$.

If the sample data are ordered from smallest to largest then the:

minimum (Min) is the smallest value;
lower quartile (LQ) is the $\frac{1}{4}(n+1)$-th value;
median (Med) is the middle [or the $\frac{1}{2}(n+1)$ -th] value;
upper quartile (UQ) is the $\frac{3}{4}(n+1)$-th value;
maximum (Max) is the largest value.

These five values constitute a **five-number summary** of the data. They can be represented diagrammatically by a *box-and-whisker plot*, commonly called a *boxplot*.

Grouped Frequency Data

If the data are given in the form of a grouped frequency distribution where we have f_i observations in an interval whose mid-point is x_i then, if $\sum f_i = n$

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{n} \quad \text{and}$$

$$S_{xx} = \sum f_i (x_i - \bar{x})^2 = \sum f_i x_i^2 - \frac{(\sum f_i x_i)^2}{n}.$$

Events & probabilities

The *intersection* of two events A and B is $A \cap B$. The union of A and B is $A \cup B$. A and B are **mutually exclusive** if they cannot both occur, denoted $A \cap B = \emptyset$ where $\emptyset$ is called the **null event**. For an event A, $0 \leq P(A) \leq 1$. For two events A and B

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

If A and B are mutually exclusive then

$$P(A \cup B) = P(A) + P(B).$$

Equally likely outcomes

If a complete set of n elementary outcomes are all equally likely to occur, then the probability of each elementary outcome is $\frac{1}{n}$. If an event A consists of m of these n elements, then $P(A) = \frac{m}{n}$.

Independent events

A, B are *independent* if and only if $P(A \cap B) = P(A)P(B)$.

Conditional Probability of A given B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{if} \quad P(B) \neq 0.$$

Bayes' Theorem: $P(B|A) = \dfrac{P(A|B)P(B)}{P(A)}$.

Theorem of Total Probability

The k events $B_1, B_2, \ldots B_k$ form a *partition* of the sample space S if $B_1 \cup B_2 \cup B_3 \ldots \cup B_k = S$ and no two of the B_i's can occur together. Then $P(A) = \sum_i P(A|B_i)P(B_i)$. In this case Bayes' Theorem generalizes to

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)} \qquad (i = 1, 2, \ldots k)$$

If B' is the *complement* of the event B , $P(B') = 1 - P(B)$ and $P(A) = P(A|B)P(B) + P(A|B')P(B')$ is a special case of the theorem of total probability. The complement of the event B is commonly denoted $\overline{B}$.

43799 D&PS Jan11

mathcentre

For the help you need to support your course

Guide to Statistics:

Probability & Statistics Facts, Formulae and Information

mathcentre is a project offering students and staff free resources to support the transition from school mathematics to university mathematics in a range of disciplines.

www.mathcentre.ac.uk

Permutations and combinations

The number of ways of selecting r objects out of a total of n, where the order of selection is important, is the number of **permutations**: ${}^nP_r = \dfrac{n!}{(n-r)!}$. The number of ways in which r objects can be selected from n when the order of selection is not important is the number of **combinations**: ${}^nC_r = \binom{n}{r} = \dfrac{n!}{r!(n-r)!}$. ${}^nC_n$ must equal 1, so $0! = 1$ and ${}^nC_0 = 1$; ${}^nC_r = {}^nC_{n-r}$. Also

$$ {}^nC_0 + {}^nC_1 + \ldots {}^nC_{n-1} + {}^nC_n = 2^n $$

$$ {}^{n+1}C_r = {}^nC_r + {}^nC_{r-1} $$

Random variables

Data arise from observations on variables that are **measured** on different **scales**. A *nominal* scale is used for named categories (e.g. race, gender) and an *ordinal* scale for data that can be ranked (e.g. attitudes, position) - no arithmetic operations are valid with either. *Interval* scale measurements can be added and subtracted only (e.g. temperature), but with *ratio* scale measurements (e.g. age, weight) multiplication and division can be used meaningfully as well. Generally, random variables are either *discrete* or *continuous*. Note: in reality, all data are discrete because the accuracy of measuring is always limited.

A **discrete** random variable X can take one of the values $x_1, x_2, \ldots,$, the probabilities $p_i = P(X = x_i)$ must satisfy $p_i \geq 0$ and $p_1 + p_2 + \ldots = 1$. The pairs (x_i, p_i) form the **probability mass function** (pmf) of X.

A **continuous** random variable X takes values x from a continuous set of possible values. It has a **probability density function** (**pdf**) $f(x)$ that satisfies $f(x) \geq 0$ and $\int f(x)\mathrm{d}x = 1$, with $P(a < x \leq b) = \displaystyle\int_a^b f(x)\mathrm{d}x$.

Expected values

The expected value of a function $H(X)$ of a random variable X is defined as

$$ E\left[H(X)\right] = \begin{cases} \sum H(x_i)P(X = x_i), & X \text{ discrete.} \\ \int H(x)f(x)\mathrm{d}x, & X \text{ continuous.} \end{cases} $$

Expectation is linear in that the expectation of a linear combination of functions is the same linear combination of expectations. For example,

$$ E[X^2 + \log X + 1] = E[X^2] + E[\log X] + 1 $$

but

$$ E[\log X] \neq \log E[X] \text{ and } E[1/X] \neq 1/E[X] $$

Variance

The variance of a random variable is defined as

$$ \mathrm{Var}(X) = \mathrm{E}[(X - \mu)^2] \equiv \mathrm{E}[X^2] - \mu^2 $$

Properties:

$\mathrm{Var}(X) \geq 0$ and is equal to 0 only if X is a constant.
$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$, where a and b are constants.

Moment generating functions

The moment generating function (mgf) of a random variable is defined as

$$ M_X(t) = \mathrm{E}[\exp(tX)] \qquad \text{if this exists.} $$

$\mathrm{E}[X^k]$ can be evaluated as the:

(i) coefficient of $\dfrac{t^r}{r!}$ in the power expansion of $M_X(t)$.
(ii) r-th derivative of $M_X(t)$ evaluated at $t = 0$.

Measures of location

The **mean** or **expectation** of the random variable X is $\mathrm{E}[X]$, the long-run average of realisations of X. The **mode** is where the **pmf** or **pdf** achieves a maximum (if it does so). For a random variable, X, the **median** is such that $P(X \leq \text{median}) = \frac{1}{2}$, so that 50% of values of X occur above and 50% below the median.

Percentiles

x_p is the 100-p-th percentile of a random variable X if $P(X \leq x_p) = p$. For example, the 5th percentile, $x_{0.05}$, has 5% of the values smaller than or equal to it. The **median** is the 50-th percentile, the **lower quartile** is the 25th percentile, the **upper quartile** is the 75th percentile.

Measures of dispersion

The **inter-quartile range** is defined to be the difference between the upper and lower quartiles, UQ - LQ. The **standard deviation** is defined as the square root of the variance, $\sigma = \sqrt{\mathrm{Var}(X)}$, and is in the same units as the random variable X.

Cumulative Distribution Function

This is defined as a function of any real value t by

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

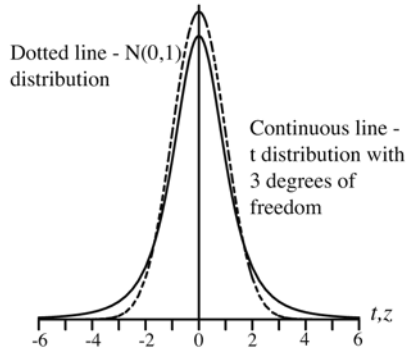
If X is a continuous random variable, F is a continuous function of t; if X is discrete, then F is a step function.

v1. Mar.07

The Central Limit Theorem

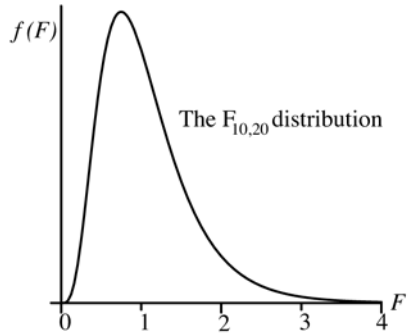
If a random sample of size n is taken from *any* distribution with mean μ and variance σ^2, the sampling distribution of the mean will be *approximately* $\sim \mathrm{N}(\mu, \sigma^2/n)$, where $\sim$ means 'is distributed as'. The larger n is, the better the approximation.

The standard normal and Student's t distributions



If a random variable $X \sim \mathrm{N}(\mu, \sigma^2)$, $z = (X - \mu)/\sigma \sim \mathrm{N}(0, 1)$, the *standard normal distribution*. The t distribution with $(n - 1)$ degrees of freedom is used in place of z for small samples size n from normal populations when σ^2 is unknown. As n increases the distribution of t converges to $\mathrm{N}(0, 1)$. These distributions are used, e.g., for inference about means, differences between means and in regression.

Fisher's F distribution



If $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$ are independent random variables then

$$ \frac{X_1/\nu_1}{X_2/\nu_2} \sim F_{\nu_1, \nu_2} $$

the F distribution with (ν_1, ν_2) degrees of freedom. This distribution is used, for example, for inference about the ratio of two variances, in Analysis of Variance (ANOVA) and in simple and multiple linear regression.

Statistics & Sampling Distributions

Population and samples

A (statistical) **population** is the complete set of all possible measurements or values, corresponding to the entire collection of units, for which inferences are to be made from taking a **sample** - the set of measurements or values that are actually collected from a population.

Simple random sample: every item in the population is equally likely to be in the sample, independently of which other members of the population are chosen.

Parameter: a quantity that describes an aspect of a population, eg. the population mean, μ, or variance, σ^2.

Statistic: a quantity calculated from the sample, e.g. the sample mean, $\bar{x}$, or variance, s^2.

Sampling distributions: The value of a statistic will in general vary from sample to sample, in which case it will have its own probability distribution, called its **sampling distribution**. A statistic used to estimate the value of a *parameter* θ in a distribution is called an **estimator** (the random variable) or an **estimate** (the value).

If $\hat{\theta}$ is an estimator of θ , the mean of its sampling distribution, $\mathbf{E}[\hat{\theta}]$, is called the *sampling mean*. The variance, $\mathrm{Var}(\hat{\theta})$, is called the *sampling variance*.

$\sqrt{\mathrm{Var}(\hat{\theta}\,)}$ is called the *standard error* of $\hat{\theta}$. If $\mathbf{E}[\hat{\theta}] = \theta$, then $\hat{\theta}$ is an unbiased estimator of θ e.g. $\bar{X}$ is an unbiased estimator for μ and has sampling variance $\frac{\sigma^2}{n}$ where $\mathrm{Var}(X_i\,) = \sigma^2$, $(i = 1, 2, \ldots, n)$.

Corrected sum of squares

$$S_{xx} = \sum (x_i - \bar{x})^2 \equiv \sum x_i^2 - n\bar{x}^2 \equiv \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

has expectation $(n-1)\sigma^2$ so that dividing S_{xx} by $(n-1)$ will give an unbiased estimator of σ^2, denoted s^2.

Normal and Chi-squared distributions

If $X_1, X_2, \ldots X_n$ are independently and identically $\sim \mathrm{N}(\mu, \sigma^2)$, then $\sum \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_n$, a Chi-squared distribution with n **degrees of freedom**.

Also $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ independently of $\frac{S_{xx}}{\sigma^2} \sim \chi^2_{(n-1)}$.

Simple Linear Regression

To fit the straight line $y = \alpha + \beta x$ to data (x_i, y_i), $i = 1, 2, \ldots n$ by the method of **least squares** the estimates of slope, β, and intercept, α , are given by:

$$b = \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i \sum y_i\right)}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2} = \frac{S_{xy}}{S_{xx}}, \qquad a = \bar{y} - b\bar{x}$$

If we assume that the x_i are known and that the y_i have normal distributions with means $\alpha + \beta x_i$, and constant variance σ^2, written as $y_i \sim N(\alpha + \beta x_i, \sigma^2)$, then if x_0 is a fixed value

$$b \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

$$a \sim N\left(\alpha, \sigma^2\left\{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right\}\right)$$

$$a + bx_0 \sim N\left(\alpha + \beta x_0, \sigma^2\left\{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right\}\right)$$

A common alternative is to use $\hat{\alpha}$ for a and $\hat{\beta}$ for b.

Correlation

Given observations (x_i, y_i), $i = 1, 2, \ldots, n$ on two random variables X and Y the **Pearson (product moment**) correlation between them is given by:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i \sum y_i\right)}{\sqrt{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}\sqrt{\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2}}$$

We use r to estimate the correlation, ρ , between X and Y. For large n, r is approximately $\sim N\left(\rho, \frac{1}{n-2}\right)$. The (**Spearman**) Rank Correlation Coefficient is given by

$$r_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i is the difference between the *ranks* of (x_i, y_i), $i = 1, 2, \ldots, n$. If ranks are tied, see further reading.

Further reading: Kotz, S., and Johnson,L. (1988) Encyclopedia of Statistical Sciences, Vols.1-9. New York: John Wiley and Sons.

Time Series

A time series Y_t $(t = 1, 2, \ldots, n)$ is a set of n observations recorded through time t, (e.g. days, weeks, months). The arithmetic mean of blocks of k successive values

$$\frac{Y_1 + Y_2 + \ldots + Y_k}{k}, \frac{Y_2 + Y_3 + \ldots + Y_{k+1}}{k}, \ldots$$

is a **simple moving average** of order k, itself a time series which is *smoother* than Y_t and can be used to track, or estimate, the underlying level, μ_t, of Y_t. If $0 < \alpha < 1$ an **exponentially weighted moving average** (EWMA) at time t uses a discounted weighted average of current and past data to estimate μ_t with

$$\hat{\mu}_t = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \ldots$$

This is equivalent to the recurrence relation

$$\hat{\mu}_t = \alpha Y_t + (1-\alpha)\hat{\mu}_{t-1}$$

Moving averages are often plotted on the same graph as Y_t. If Y_t additionally contains trend, R_t, the rate of change of data per unit time, and $\mu_t = \mu_{t-1} + R_{t-1}$, then the recurrence relation is

$$\hat{\mu}_t = \alpha Y_t + (1-\alpha)(\hat{\mu}_{t-1} + \hat{R}_{t-1})$$

If $0 < \beta < 1$ the *trend smoothing equation* is

$$\hat{R}_t = \beta(\hat{\mu}_t - \hat{\mu}_{t-1}) + (1-\beta)\hat{R}_{t-1}$$

known as *Holt's Linear Exponential Smoothing*. If Y_t also contain *seasonality*, S_t, a smoothing constant γ, $(0 < \gamma < 1)$ is used in a *seasonal smoothing equation*, $\hat{S}_t = \gamma Y_t/\hat{\mu}_t + (1-\gamma)\hat{S}_{t-k}$, assuming the periodicity is k, with *multiplicative* seasonality. For monthly data $k = 12$.

Forecasting from time n (now) to time $n+h$ ($h = 1, 2, \ldots$)

Level only, $\hat{Y}_{n+h} = \hat{\mu}_n$, the latest EWMA.

Level and constant trend, $\hat{Y}_{n+h} = a + b(n+h)$, the simple linear regression trend line of Y_t against t.

Level and changing trend, $\hat{Y}_{n+h} = \hat{\mu}_n + h\hat{R}_n$.

Level, changing trend and seasonality $\hat{Y}_{n+h} = \hat{\mu}_n + h\hat{R}_n$, where $\hat{\mu}_n = \alpha Y_n/\hat{S}_{n-12} + (1-\alpha)(\hat{\mu}_{n-1} + \hat{R}_{n-1})$.

A **hypothesis test** involves testing a claim, or **null hypothesis** H_0, about a parameter against an alternative, H_1. A decision to **reject** H_0 or **not reject** H_0 uses sample evidence to *calculate* a **test statistic** which is judged against a **critical value**. H_0 is maintained unless it is made untenable by sample evidence. Rejecting H_0 when we should not is a **Type I error**. The probability (we are prepared to accept) of making a Type I error is called the **significance level** α and yields the critical value. The *smallest* α at which we can just reject H_0 is the p-**value** of the test. Not rejecting H_0 when we should is a **Type II error**, with probability β. The **power** of a hypothesis test is $1-\beta$. An **interval estimate** for a parameter is a *calculated* range within which it is deemed likely to fall. Given α, the set of intervals from infinitely repeated random samples of size n will contain the parameter $(100-\alpha)\%$ of the time: each interval is a $(100-\alpha)\%$ **confidence interval**.

One sample hypothesis tests

1. For $X \sim N(\mu, \sigma^2)$, σ^2 *known; random sample evidence* $\bar{x}$ *and* n. Null hypothesis, $H_0 : \mu = \mu_0$; 2-sided alternative $H_1 : \mu \neq \mu_0$. Test statistic $z_{\text{calc}} = \dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$. Reject H_0 (at the α level) if $|z_{\text{calc}}| \geq z_{\alpha/2}$, the critical value of z.

2. For $X \sim N(\mu, \sigma^2)$, σ^2 *unknown; random sample evidence* $\bar{x}$, s *and* n. Null hypothesis, $H_0 : \mu = \mu_0$; 2-sided alternative $H_1 : \mu \neq \mu_0$. Test statistic $t_{\text{calc}} = \dfrac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t_{(n-1)}$, the t distribution with $(n-1)$ df. For $n > 30$ and if X has *any* distribution, $t \sim N(0,1)$. Reject H_0 if $|t_{\text{calc}}| \geq t_{\alpha/2}$, the critical value of t with $(n-1)$ df.

3. For $X \sim N(\mu, \sigma^2)$, σ^2 *unknown; random sample evidence* s *and* n. Null hypothesis, $H_0 : \sigma^2 = \sigma_0^2$; alternative $H_1 : \sigma^2 > \sigma_0^2$. Test statistic $\chi^2_{\text{calc}} = (n-1)s^2/\sigma_0^2 \sim \chi^2_{n-1}$. Reject H_0 if $\chi^2_{\text{calc}} > \chi^2_\alpha$, the critical value of χ^2 with $(n-1)$ df.

In each case the p-value is the tail area outside the calculated statistic.

Two sample hypothesis tests

For $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, σ_1^2, σ_2^2 *unknown; random sample evidence* $\bar{x}_1$, $\bar{x}_2$, s_1^2, s_2^2, n_1 *and* n_2.

1. Null hypothesis, $H_0 = \mu_1 - \mu_2 = c$; 2-sided alternative $H_1 : \mu_1 - \mu_2 \neq c$. Test statistic $t_{\text{calc}} = \dfrac{(\bar{x}_1 - \bar{x}_2 - c)}{s\sqrt{1/n_1 + 1/n_2}} \sim t_{(n_1+n_2-2)}$, and $s^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}$, assuming $\sigma_1^2 = \sigma_2^2$. Reject H_0 if $|t_{\text{calc}}| \geq t_{\alpha/2}$ the critical value of t with (n_1+n_2-2) df.

2. Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$; alternative $H_1 : \sigma_1^2 > \sigma_2^2$. Test statistic $F_{\text{calc}} = \dfrac{(n_1-1)s_1^2}{(n_2-1)s_2^2} \sim F_{n_1-1, n_2-1}$. Reject H_0 if $F_{\text{calc}} > F_\alpha$ the critical value of F with n_1-1 and n_2-1 df.

Confidence interval for a population mean - σ^2 unknown

If X has mean μ and variance σ^2, with $n > 30$ an approximate $100(1-\alpha)\%$ confidence interval for μ is $\bar{x} - \dfrac{t_{\alpha/2}s}{\sqrt{n}}$ to $\bar{x} + \dfrac{t_{\alpha/2}s}{\sqrt{n}}$.

If $X \sim N(\mu, \sigma^2)$ the interval is exact for all n.

Standard statistical distributions

Name/parameters	Conditions/application	pdf/pmf	Mean	Variance	mgf	Notes
Binomial Bin(n,p) Positive integer n Probability p, $0 \leq p \leq 1$	n independent success/fail trials each with probability p of success. X = number of successes.	$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \ldots, n$	np	$np(1-p)$	$\left(1-p+pe^t\right)^n$	$X \sim \text{Bin}\ (n,p)$ $\Rightarrow n - X \sim \text{Bin}(n, 1-p)$
Geometric Geom(p) Probability p, $0 \leq p \leq 1$	Repeated independent success/fail trials each with probability p of success. X = number of trials up to and including the first success.	$P(X=x) = (1-p)^{x-1}p$ $x = 1, 2, \ldots$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$	$\dfrac{pe^t}{1-(1-p)e^t}$	Has the "lack of memory" property $P(X > a+b \mid X > b) = P(X > a)$
Poisson Po(λ) λ a positive number	Events occur randomly at a constant rate. X = number of occurrences in some interval. λ is the expected number of occurrences	$P(X=x) = e^{-\lambda}\dfrac{\lambda^x}{x!}$ $x = 0, 1, 2, \ldots$	λ	λ	$\exp(\lambda(e^t-1))$	Useful as approximation to Bin(n,p) if n is large and p is small
Normal $N(\mu, \sigma^2)$ μ, σ both real; $\sigma > 0$	A widely used distribution for symmetrically distributed random variables with mean μ and standard deviation σ.	$f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$ all real x	μ	σ^2	$\exp\left(\mu t + \dfrac{1}{2}\sigma^2 t^2\right)$	Can approximate Binomial, Poisson Pascal and Gamma distributions (see Central Limit Theorem)
Exponential Expon(θ)	Events are occurring at rate θ per unit time. X = time to first occurrence.	$f(x) = \theta \exp(-\theta x)$ $x > 0$	$\dfrac{1}{\theta}$	$\dfrac{1}{\theta^2}$	$\dfrac{\theta}{\theta - t}$, $t < \theta$	Has the "lack of memory"property $P(X > a+b \mid X > b) = P(X > a)$
Negative-binomial or Pascal Pasc(r,p) Positive integer n Probability p, $0 \leq p \leq 1$	Repeated independent success/fail trials each with probability p of success. X = number of trials up to and including the r-th success.	$P(X=x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$ $x = r, r+1, r+2, \ldots$	$\dfrac{r}{p}$	$\dfrac{r(1-p)}{p^2}$	$\left(\dfrac{pe^t}{1-(1-p)e^t}\right)^r$	Pasc$(1,p) \equiv$ Geom(p)
Gamma Ga(α, β) $\alpha, \beta > 0$	Generalization of the exponential distribution; if α is an integer it represents the waiting time to the α-th occurrence of a random event where β is the expected number of events.	$f(x) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ $x > 0$	$\dfrac{\alpha}{\beta}$ $\alpha > 1$	$\dfrac{\alpha}{\beta^2}$	$\left(\dfrac{\beta}{\beta - t}\right)^\alpha$, $t < \beta$	Ga$(1, \lambda) \equiv$ Expon(λ) If ν is an integer, Ga$(\nu/2, 2)$ is χ^2_ν, the Chi-squared distribution with ν df.

Thomas Finley, tomf@cs.cornell.edu

Linear Algebra

A *subspace* is a set $S \subseteq \mathbb{R}^n$ such that $\mathbf{0} \in S$ and $\forall \mathbf{x}, \mathbf{y} \in S, \alpha, \beta \in \mathbb{R}$. $\alpha\mathbf{x} + \beta\mathbf{y} \in S$.

The *span* of $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is the set of all vectors in $\mathbb{R}^n$ that are linear combinations of $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

A *basis* B of subspace S, $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \subset S$ has $Span(B) = S$ and all $\mathbf{v}_i$ linearly independent.

The *dimension* of S is $|B|$ for a basis B of S.

For subspaces S, T with $S \subseteq T$, $dim(S) \leq dim(T)$, and further if $dim(S) = dim(T)$, then $S = T$.

A *linear transformation* $T : \mathbb{R}^n \to \mathbb{R}^m$ has $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \alpha, \beta \in \mathbb{R}$. $T(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha T(\mathbf{x}) + \beta T(\mathbf{y})$. Further, $\exists A \in \mathbb{R}^{m \times n}$ such that $\forall \mathbf{x}$. $T(\mathbf{x}) \equiv A\mathbf{x}$.

For two linear transformations $T : \mathbb{R}^n \to \mathbb{R}^m$, $S : \mathbb{R}^m \to \mathbb{R}^p$, $S \circ T \equiv S(T(\mathbf{x}))$ is linear transformation. $(T(\mathbf{x}) \equiv A\mathbf{x}) \wedge (S(\mathbf{y}) \equiv B) \Rightarrow (S \circ T)(\mathbf{x}) \equiv BA\mathbf{x}$.

The matrix's *row space* is the span of its rows, its *column space* or *range* is the span of its columns, and its *rank* is the dimension of either of these spaces.

For $A \in \mathbb{R}^{m \times n}$, $rank(A) \leq \min(m, n)$. A has *full row* (or *column*) *rank* if $rank(A) = m$ (or n).

A *diagonal matrix* $D \in \mathbb{R}^{n \times n}$ has $d_{j,k} = 0$ for $j \neq k$. The diagonal *identity matrix* I has $i_{j,j} = 1$.

The *upper* (or *lower*) *bandwidth* of A is $\max |i - j|$ among i, j where $i \geq j$ (or $i \leq j$) such that $A_{i,j} \neq 0$.

A matrix with lower bandwidth 1 is *upper Hessenberg*.

For $A, B \in \mathbb{R}^{n \times n}$, B is A's *inverse* if $AB = BA = I$. If such a B exists, A is *invertible* or *nonsingular*. $B = A^{-1}$.

The inverse of A is $A^{-1} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ where $A\mathbf{x}_i = \mathbf{e}_i$.

For $A \in \mathbb{R}^{n \times n}$ the following are equivalent: A is nonsingular, $rank(A) = n$, $A\mathbf{x} = \mathbf{b}$ has a solution $\mathbf{x}$ for any $\mathbf{b}$, if $A\mathbf{x} = \mathbf{0}$ then $\mathbf{x} = \mathbf{0}$.

The *nullspace* of $A \in \mathbb{R}^{m \times n}$ is $\{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \mathbf{0}\}$.

For $A \in \mathbb{R}^{m \times n}$, $Range(A)$ and $Nullspace(A^T)$ are *orthogonal complements*, i.e., $\mathbf{x} \in Range(A), \mathbf{y} \in Nullspace(A^T) \Rightarrow \mathbf{x}^T\mathbf{y} = 0$, and for all $\mathbf{p} \in \mathbb{R}^m$, $\mathbf{p} = \mathbf{x} + \mathbf{y}$ for unique $\mathbf{x}$ and $\mathbf{y}$.

For a *permutation matrix* $P \in \mathbb{R}^{n \times n}$, PA permutes the rows of A, AP the columns of A. $P^{-1} = P^T$.

Gaussian Elimination

GE produces a factorization $A = LU$, GEPP $PA = LU$.

Plain GE

1: **for** $k = 1$ **to** $n - 1$ **do**
2: **if** $a_{kk} = 0$ **then** stop
3: $\ell_{k+1:n,k} = a_{k+1:n,k}/a_{kk}$
4: $a_{k+1:n,k:n} = a_{k+1:n,k:n} - \ell_{k+1:n,k}a_{k,k:n}$
5: **end for**

Backward Substitution

1: $\mathbf{x} = zeros(n, 1)$
2: **for** $j = n$ **to** 1 **do**
3: $x_j = \dfrac{w_j - u_{j,j+1:n}x_{j+1:n}}{u_{j,j}}$
4: **end for**

GEPP

1: **for** $k = 1$ **to** $n - 1$ **do**
2: $\gamma = \underset{i \in \{k+1,\ldots,n\}}{\operatorname{argmax}} |a_{ik}|$
3: $a_{[\gamma,k],k:n} = a_{[k,\gamma],k:n}$
4: $\ell_{[\gamma,k],1:k-1} = \ell_{[k,\gamma],1:k-1}$
5: $p_k = \gamma$
6: $\ell_{k:n,k} = a_{k:n,k}/a_{kk}$
7: $a_{k+1:n,k:n} = a_{k+1:n,k:n} - \ell_{k+1:n,k}a_{k,k:n}$
8: **end for**

To solve $A\mathbf{x} = \mathbf{b}$, factor $A = LU$ (or $A = P^TLU$), solve $L\mathbf{w} = \mathbf{b}$ (or $L\mathbf{w} = \hat{\mathbf{b}}$ where $\hat{\mathbf{b}} = P\mathbf{b}$) for $\mathbf{w}$ using forward substitution, then solve $U\mathbf{x} = \mathbf{w}$ for $\mathbf{x}$ using backward substitution. The complexity of GE and GEPP is $\frac{2}{3}n^3 + O(n^2)$. GEPP encounters an exact 0 pivot iff A is singular.

For banded A, $L + U$ has the same bandwidths as A.

Norms

A *vector norm* function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ satisfies:

1. $\|\mathbf{x}\| \geq 0$, and $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \vec{0}$.
2. $\|\gamma\mathbf{x}\| = |\gamma| \cdot \|\mathbf{x}\|$ for all $\gamma \in \mathbb{R}$, and all $\mathbf{x} \in \mathbb{R}^n$.
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, for all $x, y \in \mathbb{R}^n$.

Common norms include:

1. $\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$
2. $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$
3. $\|\mathbf{x}\|_\infty = \lim_{p \to \infty} (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}} = \max_{i=1..n} |x_i|$

An *induced matrix norm* is $\|A\|_\square = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_\square}{\|\mathbf{x}\|_\square}$. It satisfies the three properties of norms.

$\forall \mathbf{x} \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}$, $\|A\mathbf{x}\|_\square \leq \|A\|_\square \|\mathbf{x}\|_\square$.

$\|AB\|_\square \leq \|A\|_\square \|B\|_\square$, called *submultiplicativity*.

$\mathbf{a}^T\mathbf{b} \leq \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$, called *Cauchy-Schwarz inequality*.

1. $\|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^n |a_{i,j}|$ (max row sum).
2. $\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^m |a_{i,j}|$ (max column sum).
3. $\|A\|_2$ is hard: it takes $O(n^3)$, not $O(n^2)$ operations.
4. $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m a_{i,j}^2}$. $\|\cdot\|_F$ often replaces $\|\cdot\|_2$.

Numerical Stability

Six sources of error in scientific computing: modeling errors, measurement or data errors, blunders, discretization or truncation errors, convergence tolerance, and rounding errors.

$$\underbrace{\pm}_{\text{sign}} \underbrace{d_1.d_2d_3\cdots d_t}_{\text{mantissa}} \times \underbrace{\beta}_{\text{base}} \overbrace{e}^{\text{exponent}}$$

For single and double:
$t = 24$, $e \in \{-126, \ldots, 127\}$
$t = 53$, $e \in \{-1022, \ldots, 1023\}$

The *relative error* in $\hat{\mathbf{x}}$ approximating $\mathbf{x}$ is $\frac{|\hat{\mathbf{x}} - \mathbf{x}|}{|\mathbf{x}|}$.

Unit roundoff or *machine epsilon* is $\epsilon_{mach} = \beta^{-t+1}$. Arithmetic operations have relative error bounded by ϵ_{mach}.

E.g., consider $z = x - y$ with input x, y. This program has three roundoff errors. $\hat{z} = ((1 + \delta_1)x - (1 + \delta_2)y)(1 + \delta_3)$, where $\delta_1, \delta_2, \delta_3 \in [-\epsilon_{mach}, \epsilon_{mach}]$.

$$\frac{|z - \hat{z}|}{|z|} = \frac{|(\delta_1 + \delta_3)x - (\delta_2 + \delta_3)y + O(\epsilon_{mach}^2)|}{|x - y|}$$

The bad case is where $\delta_1 = \epsilon_{mach}$, $\delta_2 = -\epsilon_{mach}$, $\delta_3 = 0$:

$$\frac{|z - \hat{z}|}{|z|} = \epsilon_{mach}\frac{|x + y|}{|x - y|}$$

Inaccuracy if $|x + y| \gg |x - y|$ called *catastrophic calcellation*.

Conditioning & Backwards Stability

A problem instance is *ill conditioned* if the solution is sensitive to perturbations of the data. For example, $\sin 1$ is well conditioned, but $\sin 12392193$ is ill conditioned.

Suppose we perturb $A\mathbf{x} = \mathbf{b}$ by $(A + E)\hat{\mathbf{x}} = \mathbf{b} + \mathbf{e}$ where $\frac{\|E\|}{\|A\|} \leq \delta, \frac{\|\mathbf{e}\|}{\|\mathbf{b}\|} \leq \delta$. Then $\frac{\|\hat{\mathbf{x}} + \mathbf{x}\|}{\|\mathbf{x}\|} \leq 2\delta\kappa(A) + O(\delta^2)$, where $\kappa(A) = \|A\|\|A^{-1}\|$ is the *condition number* of A.

1. $\forall A \in \mathbb{R}^{n \times n}$, $\kappa(A) \geq 1$.
2. $\kappa(I) = 1$.
3. For $\gamma \neq 0$, $\kappa(\gamma A) = \kappa(A)$.
4. For diagonal D and all p, $\|D\|_p = \max_{i=1..n} |d_{ii}|$. So, $\kappa(D) = \frac{\max_{i=1..n} |d_{ii}|}{\min_{i=1..n} |d_{ii}|}$.

If $\kappa(A) \geq \frac{1}{\epsilon_{mach}}$, A may as well be singular.

An algorithm is *backwards stable* if in the presence of roundoff error it returns the exact solution to a nearby problem instance.

GEPP solves $A\mathbf{x} = \mathbf{b}$ by returning $\hat{\mathbf{x}}$ where $(A + E)\hat{\mathbf{x}} = \mathbf{b}$. It is backwards stable if $\frac{\|E\|_\infty}{\|A\|_\infty} \leq O(\epsilon_{mach})$. With GEPP, $\frac{\|E\|_\infty}{\|A\|_\infty} \leq c_n\epsilon_{mach} + O(\epsilon_{mach}^2)$, where c_n is worst case exponential in n, but in practice almost always low order polynomial.

Combining stability and conditioning analysis yields $\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq c_n \cdot \kappa(A)\epsilon_{mach} + O(\epsilon_{mach}^2)$.

Determinant

The *determinant* $det : \mathbb{R}^{n \times n} \to \mathbb{R}$ satisfies:

1. $det(AB) = det(A)\, det(B)$.
2. $det(A) = 0$ iff A is singular.
3. $det(L) = \ell_{1,1}\ell_{2,2}\cdots\ell_{n,n}$ for triangular L.
4. $det(A) = det(A^T)$.

To compute $det(A)$ factor $A = P^TLU$. $det(P) = (-1)^s$ where s is the number of swaps, $det(L) = 1$. When computing $det(U)$ watch out for overflow!

Orthogonal Matrices

For $Q \in \mathbb{R}^{n \times n}$, these statements are equivalent:

1. $Q^TQ = QQ^T = I$ (i.e., Q is *orthogonal*)
2. The $\|\cdot\|_2$ for each row and column of Q. The inner product of any row (or column) with another is 0.
3. For all $\mathbf{x} \in \mathbb{R}^n$, $\|Q\mathbf{x}\|_2 = \|\mathbf{x}\|_2$.

A matrix $Q \in \mathbb{R}^{m \times n}$ with $m > n$ has *orthonormal columns* if the columns are orthonormal, and $Q^TQ = I$.

The product of orthogonal matrices is orthogonal.

For orthogonal Q, $\|QA\|_2 = \|A\|_2$ and $\|AQ\|_2 = \|A\|_2$.

QR-factorization

For any $A \in \mathbb{R}^{m \times n}$ with $m \geq n$, we can factor $A = QR$, where $Q \in \mathbb{R}^{m \times m}$ is orthogonal, and $R = \begin{bmatrix} R_1 & 0 \end{bmatrix}^T \in \mathbb{R}^{m \times n}$ is upper triangular. $rank(A) = n$ iff R_1 is invertible.

Q's first n (or last $m - n$) columns form an orthonormal basis for $span(A)$ (or $nullspace(A^T)$).

A *Householder reflection* is $H = I - \frac{2\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}$. H is symmetric and orthogonal. Explicit H.H. QR-factorization is:

1: **for** $k = 1 : n$ **do**
2: $\mathbf{v} = A(k : m, k) \pm \|A(k : m, k)\|_2\mathbf{e}_1$
3: $A(k : m, k : n) = \left(I - \frac{2\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}\right) A(k : m, k : n)$
4: **end for**

We get $H_nH_{n-1}\cdots H_1A = R$, so then, $Q = H_1H_2\cdots H_n$. This takes $2mn^2 - \frac{2}{3}n^3 + O(mn)$ flops.

Givens requires 50% more flops. Preferable for sparse A.

The Gram-Schmidt produces a *skinny/reduced* QR-factorization $A = Q_1R_1$, where $Q_1 \in \mathbb{R}^{m \times n}$ has orthonormal columns. The *Gram-Schmidt* algorithm is:

Left Looking

1: **for** $k = 1 : n$ **do**
2: $\mathbf{q}_k = \mathbf{a}_k$
3: **for** $j = 1 : k - 1$ **do**
4: $R(j, k) = \mathbf{q}_j^T\mathbf{a}_k$
5: $\mathbf{q}_k = \mathbf{q}_k - R(j, k)\mathbf{q}_j$
6: **end for**
7: $R(k, k) = \|\mathbf{q}_k\|_2$
8: $\mathbf{q}_k = \mathbf{q}_k/R(k, k)$
9: **end for**

Right Looking

1: $Q = A$
2: **for** $k = 1 : n$ **do**
3: $R(k, k) = \|\mathbf{q}_k\|_2$
4: $\mathbf{q}_k = \mathbf{q}_k/R(k, k)$
5: **for** $j = k + 1 : n$ **do**
6: $R(k, j) = \mathbf{q}_k^T\mathbf{q}_j$
7: $\mathbf{q}_j = \mathbf{q}_j - R(k, j)\mathbf{q}_k$
8: **end for**
9: **end for**

In left looking, let line 6 be $\mathbf{q}_j^T\mathbf{q}_{j-1}$ for modified G.S. to make it backwards stable.

Positive Definite, $A = LDL^T$

$A \in \mathbb{R}^{n \times n}$ is *positive definite* (PD) (or *semidefinite* (PSD)) if $\mathbf{x}^TA\mathbf{x} > 0$ (or $\mathbf{x}^TA\mathbf{x} \geq 0$).

When LU-factorizing symmetric A, the result is $A = LDL^T$; L is unit lower triangular, D is diagonal. A is SPD iff D has all positive entries. The *Cholesky factorization* is $A = LDL^T = LD^{1/2}D^{1/2}L^T = GG^T$. Can be done directly in $\frac{n^3}{3} + O(n^2)$ flops. If G has all positive diagonal A is SPD.

To solve $A\mathbf{x} = \mathbf{b}$ for SPD A, factor $A = GG^T$, solve $G\mathbf{w} = \mathbf{b}$ by forward substitution, then solve $G^T\mathbf{x} = \mathbf{w}$ with backwards substitution, which takes $\frac{n^3}{3} + O(n^2)$ flops.

For $A \in \mathbb{R}^{m \times n}$, if $rank(A) = n$, then A^TA is SPD.

Basic Linear Algebra Subroutines

0. Scalar ops, like $\sqrt{x^2 + y^2}$. $O(1)$ flops, $O(1)$ data.
1. Vector ops, like $\mathbf{y} = a\mathbf{x} + \mathbf{y}$. $O(n)$ flops, $O(n)$ data.
2. Matrix-vector ops, like rank-one update $A = A + \mathbf{x}\mathbf{y}^T$. $O(n^2)$ flops, $O(n^2)$ data.
3. Matrix-matrix ops, like $C = C + AB$. $O(n^2)$ data, $O(n^3)$ flops.

Use the highest BLAS level possible. Operators are architecture tuned, e.g., data processed in cache-sized bites.

Linear Least Squares

Suppose we have points $(u_1, v_1), \ldots, (u_5, v_5)$ that we want to fit a quadratic curve $au^2 + bu + c$ through. We want to solve for

$$\begin{bmatrix} u_1^2 & u_1 & 1 \\ \vdots & \vdots & \vdots \\ u_5^2 & u_5 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} v_1 \\ \vdots \\ v_5 \end{bmatrix}$$

This is *overdetermined* so an exact solution is out. Instead, find the *least squares* solution $\mathbf{x}$ that minimizes $\|A\mathbf{x} - \mathbf{b}\|_2$.

For the *method of normal equations*, solve for $\mathbf{x}$ in $A^TA\mathbf{x} = A^T\mathbf{b}$ by using Cholesky factorization. This takes $mn^2 + \frac{n^3}{3} + O(mn)$ flops. It is conditionally but not backwards stable: A^TA doubles the condition number.

Alternatively, factor $A = QR$. Let $\mathbf{c} = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 \end{bmatrix}^T = Q^T\mathbf{b}$. The least squares solution is $\mathbf{x} = R_1^{-1}\mathbf{c}_1$.

If $rank(A) = r$ and $r < n$ (rank deficient), factor $A = U\Sigma V^T$, let $y = V^Tx$ and $c = U^Tb$. Then, $\min \|A\mathbf{x} - \mathbf{b}\|_2 = \min \sqrt{\sum_{i=1}^r (\sigma_i y_i - c_i)^2 + \sum_{i=r+1}^m c_i^2}$, so $y_i = \frac{c_i}{\sigma_i}$. For $i = r + 1 : n$, y_i is arbitrary.

Singular Value Decomposition

For any $A \in \mathbb{R}^{m \times n}$, we can express $A = U\Sigma V^T$ such that $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, and $\Sigma = diag(\sigma_1, \cdots, \sigma_p) \in \mathbb{R}^{m \times n}$ where $p = \min(m, n)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. The σ_i are singular values.

1. Matrix 2-norm, where $\|A\|_2 = \sigma_1$.
2. The condition number $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$, or rectangular condition number $\kappa_2(A) = \frac{\sigma_1}{\sigma_{\min(m,n)}}$. Note that $\kappa_2(A^TA) = \kappa_2(A)^2$.
3. For a rank k approximation to A, let $\Sigma_k = diag(\sigma_1, \cdots, \sigma_k, \mathbf{0}^T)$. Then $A_k = U\Sigma_kV^T$. $rank(A_k) \leq k$ and $rank(A_k) = k$ iff $\sigma_k > 0$. Among rank k or lower matrices, A_k minimizes $\|A - A_k\|_2 = \sigma_{k+1}$.
4. Rank determination, since $rank(A) = r$ equals the number of nonzero σ, or in machine arithmetic, perhaps the number of $\sigma \geq \epsilon_{mach} \times \sigma_1$.

$$A = U\Sigma V^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma(1 : r, 1 : r) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}$$

See that $range(U_1) = range(A)$. The SVD gives an orthonormal basis for the range and nullspace of A and A^T.

Compute the SVD by using shifted QR on A^TA.

Information Retrival & LSI

In the *bag of words* model, $\mathbf{w}_d \in \mathbb{R}^m$, where $\mathbf{w}_d(i)$ is the (perhaps weighted) frequency of term i in document d. The *corpus* matrix is $A = [\mathbf{w}_1, \cdots, \mathbf{w}_n] \in \mathbb{R}^{m \times n}$. For a query $\mathbf{q} \in \mathbb{R}^m$, rank documents according to a $\frac{\mathbf{q}^T\mathbf{w}_d}{\|\mathbf{w}_d\|_2}$ score.

In *latent semantic indexing*, you do the same, but in a k dimensional subspace. Factor $A = U\Sigma V^T$, then define $A^* = \Sigma_{1:k,1:k}V_{:,1:k}^T \in \mathbb{R}^{k \times n}$. Each $\mathbf{w}_d^* = A_{:,d}^* = U_{:,1:k}^T\mathbf{w}_d$, and $\mathbf{q}^* = U_{:,1:k}^T\mathbf{q}$.

In the Ando-Lee analysis, for a corpus with k topics, for $t \in 1:k$ and $d \in 1:n$, let $R_{t,d} \geq 0$ be document d's relevance to topic t. $\|R_{:,d}\|_2 = 1$. *True document similarity* is $RR^T = \mathbb{R}^{n\times n}$, where entry (i,j) is relevance of i to j. Using LSI, if A contains information about RR^T, then $(A^*)^T A^*$ will approximate RR^T well. LSI depends on even distribution of topics, where distribution is $\rho = \frac{\max_t \|R_{t,:}\|_2}{\min_t \|R_{t,:}\|_2}$. Great for ρ is near 1, but if $\rho \gg 1$, LSI does worse.

Complex Numbers

Complex numbers are written $z = x + iy \in \mathbb{C}$ for $i = \sqrt{-1}$.

The *real part* is $x = \Re(z)$. The *imaginary part* is $y = \Re(z)$.

The *conjugate* of z is $\overline{z} = x - iy$. $\overline{A\overline{x}} = (\overline{Ax})$, $\overline{A}\,\overline{B} = (\overline{AB})$

The *absolute value* of z is $|z| = \sqrt{x^2+y^2}$.

The *conjugate transpose* of $\mathbf{x}$ is $\mathbf{x}^H = (\overline{\mathbf{x}})^T$. $A \in \mathbb{C}^{n\times n}$ is *Hermitian* or *self-adjoint* if $A = A^H$.

If $Q^HQ = I$, Q is *unitary*.

Eigenvalues & Eigenvectors

For $A \in \mathbb{C}^{n\times n}$, if $A\mathbf{x} = \lambda\mathbf{x}$ where $\mathbf{x} \neq 0$, $\mathbf{x}$ is an *eigenvector* of A and λ is the corresponding *eigenvalue*.

Remember, $A - \lambda\mathbf{x}$ is singular iff $det(A - \lambda I) = 0$. With λ as a variable, $det(A-\lambda I)$ is A's *characteristic polynomial*.

For nonsingular $T \in \mathbb{C}^{n\times n}$, $T^{-1}AT$ (the *similarity transformation*) is *similar* to A. Similar matrices have the same characteristic polynomial and hence the same eigenvalues (though probably different eigenvectors). This relationship is reflexive, transitive, and symmetric.

A is *diagonalizable* if A is similar to a diagonal matrix $D = T^{-1}AT$. A's eigenvalues are D's diagonals, and the eigenvectors are columns of T since $AT_{:,i} = D_{i,i}T_{:,i}$. A is diagonalizable iff it has n linearly independent eigenvectors.

For symmetric $A \in \mathbb{R}^{n\times n}$, A is diagonalizable, has all real eigenvalues, and the eigenvectors may be chosen as the columns of an orthogonal matrix Q. $A = QDQ^T$ is the *eigendecomposition* of A. Further for symmetric A:

1. The singular values are absolute values of eigenvalues.
2. Is SPD (or SPSD) iff eigenvalues > 0 (or ≥ 0).
3. For SPD, singular values equal eigenvalues.
4. For $B \in \mathbb{R}^{m\times n}$, $m \geq n$, singular values of B are the square roots of B^TB's eigenvalues.

For any $A \in \mathbb{C}^{n\times n}$, the *Schur form* of A is $A = QTQ^H$ with unitary $Q \in \mathbb{C}^{n\times n}$ and upper triangular $T \in \mathbb{C}^{n\times n}$.

In this sheet I denote $\lambda_{|\max|} = \max_{\lambda\in\{\lambda_1,\ldots,\lambda_n\}} |\lambda|$.

For $B \in \mathbb{C}^{n\times n}$, then $\lim_{k\to\infty} B^k = 0$ if $\lambda_{|\max|}(B) < 1$.

Power Methods for Eigenvalues

$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)}$ converges to $\lambda_{|\max|}(A)$'s eigenvector.

Once you find an eigenvector $\mathbf{x}$, find the associated eigenvalue λ through the *Raleigh quotient* $\lambda = \frac{\mathbf{x}^{(k)T}A\mathbf{x}^{(k)}}{\mathbf{x}^{(k)T}\mathbf{x}^{(k)}}$.

The *inverse shifted power method* is $\mathbf{x}^{(k+1)} = (A - \sigma I)^{-1}\mathbf{x}^{(k)}$. If A has eigenpairs $(\lambda_1, \mathbf{u}_1), \ldots, (\lambda_n, \mathbf{u}_n)$, then $(A - \sigma I)^{-1}$ has eigenpairs $\left(\frac{1}{\lambda_1-\sigma}, \mathbf{u}_1\right), \ldots, \left(\frac{1}{\lambda_n-\sigma}, \mathbf{u}_n\right)$.

Factor $A = QHQ^T$ where H is upper Hessenberg.

To factor $A = QHQ^T$, find successive Householder reflections $H_1, H_2, \ldots$ that zero out rows 2 and lower of column 1, rows 3 and lower of column 2, etc. Then $Q = H_1^T \cdots H_{n-2}^T$.

$A^{(0)} = A$
for $k = 0, 1, 2, \ldots$ **do**
  Set $A^{(k)} - \sigma^{(k)}I = Q^{(k)}R^{(k)}$
  $A^{(k+1)} = R^{(k)}Q^{(k)} + \sigma^{(k)}I$
end for

$A^{(k)}$ is similar to A by orthog. trans. $U^{(k)} = Q^{(0)}\cdots Q^{(k+1)}$. Perhaps choose $\sigma^{(k)}$ as eigenvalues of submatrices of A.

Arnoldi and Lanczos

Given $A \in \mathbb{R}^{n\times n}$ and unit length $\mathbf{q}_1 \in \mathbb{R}^n$, output Q, H such that $A = QHQ^T$. Use Lanczos for symmetric A.

Arnoldi

for $k = 1 : n-1$ **do**
  $\tilde{\mathbf{q}}_{k+1} = A\mathbf{q}_k$
  for $\ell = 1 : k$ **do**
    $H(\ell,k) = \mathbf{q}_\ell^T\tilde{\mathbf{q}}_{k+1}$
    $\tilde{\mathbf{q}}_{k+1} = \tilde{\mathbf{q}}_{k+1} - H(\ell,k)\mathbf{q}_\ell$
  end for
  $H(k+1,k) = \|\tilde{\mathbf{q}}_{k+1}\|_2$
  $\mathbf{q}_{k+1} = \frac{\tilde{\mathbf{q}}_{k+1}}{H(k+1,k)}$
end for

Lanczos

$\beta_0 = \|\mathbf{w}_0\|_2$
for $k = 1, 2, \ldots$ **do**
  $\mathbf{q}_k = \frac{\mathbf{w}_{k-1}}{\beta_{k-1}}$
  $\mathbf{u}_k = A\mathbf{q}_k$
  $\mathbf{v}_k = \mathbf{u}_k - \beta_{k-1}\mathbf{q}_{k-1}$
  $\alpha_k = \mathbf{q}_k^T\mathbf{v}_k$
  $\mathbf{w}_k = \mathbf{v}_k - \alpha_k\mathbf{q}_k$
  $\beta_k = \|\mathbf{w}_k\|_2$
end for

For Lanczos, the α_k and β_k are diagonal and subdiagonal entries of the Hermitian tridiagonal T_k, and we have H in Arnoldi. After very few iterations of either method, the eigenvalues of T_k and H will be excellent approximations to the "extreme" eigenvalues of A.

For k iterations, Arnoldi is $O(nk^2)$ times and $O(nk)$ space, Lanczos is $O(nk)+k\cdot\mathcal{M}$ time ($\mathcal{M}$ is time for matrix-vector multiplication) and $O(nk)$ space, or $O(n+k)$ space if old $\mathbf{q}_k$'s are discarded.

Iterative Methods for $A\mathbf{x} = \mathbf{b}$

Useful for sparse A where GE would cause fill-in.

In the *splitting method*, $A = M - N$ and $M\mathbf{v} = \mathbf{c}$ is easily solvable. Then, $\mathbf{x}^{(k+1)} = M^{-1}\left(N\mathbf{x}^{(k)} + \mathbf{b}\right)$. If it converges, the limit point $\mathbf{x}^*$ is a solution to $A\mathbf{x} = \mathbf{b}$.

The error is $\mathbf{e}^{(k)} = (M^{-1}N)^k\mathbf{e}_0$, so splitting methods converge if $\lambda_{|\max|}(M^{-1}N) < 1$.

In the *Jacobi method*, consider M as the diagonals of A. This will fail of A has any zero diagonals.

Conjugate Gradient

Conjugate gradient iteratively solve $A\mathbf{x} = \mathbf{b}$ for SPD A. It is derived from Lanczos and takes advantage of if A is SPD then T is SPD. It produces the exact solution after n iterations. Time per iteration is $O(n) + \mathcal{M}$.

$\mathbf{x}^{(0)} =$ arbitrary ($\mathbf{0}$ is okay)
$\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}^{(0)}$
$\mathbf{p}_0 = \mathbf{r}_0$
for k=0,1,2,… **do**
  $\alpha_k = (\mathbf{r}_k^T\mathbf{r}_k)/(\mathbf{p}_k^TA\mathbf{p}_k)$
  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{p}_k$
  $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_kA\mathbf{p}_k$
  $\beta_{k+1} = (\mathbf{r}_{k+1}^T\mathbf{r}_{k+1})/(\mathbf{r}_k^T\mathbf{r}_k)$
  $\mathbf{p}_{k+1} = \mathbf{r}_{k+1} - \beta_{k+1}\mathbf{p}_k$
end for

Error is reduced by $(\sqrt{\kappa(A)}-1)/(\sqrt{\kappa(A)}+1)$ per iteration. Thus, for $\kappa(A) = 1$, CG converges after 1 iteration. To speed up CG, use a *preconditioner* M such that $\kappa(MA) \ll \kappa(A)$ and solve $MA\mathbf{x} = M\mathbf{b}$ instead.

Multivariate Calculus

Provided $f : \mathbb{R}^n \to \mathbb{R}$, the gradient and Hessian are

$$\nabla f = \begin{bmatrix} \frac{\delta f}{\delta x_1} \\ \vdots \\ \frac{\delta f}{\delta x_n} \end{bmatrix}, \nabla^2 f = \begin{bmatrix} \frac{\delta^2 f}{\delta x_1^2} & \frac{\delta^2 f}{\delta x_1 \delta x_2} & \cdots & \frac{\delta^2 f}{\delta x_1 \delta x_n} \\ \vdots & & & \vdots \\ \frac{\delta^2 f}{\delta x_n \delta x_1} & \frac{\delta^2 f}{\delta x_n \delta x_2} & \cdots & \frac{\delta^2 f}{\delta x_n^2} \end{bmatrix}$$

If f is c^2 (2^{nd} partials are all continuous), $\nabla^2 f$ is symmetric.

The Taylor expansion for f is

$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \mathbf{h}^T\nabla f(\mathbf{x}) + \frac{1}{2}\mathbf{h}^T\nabla^2 f(\mathbf{x})\mathbf{h} + O(\|\mathbf{h}\|^3)$

Provided $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, the Jacobian is

$$\nabla\mathbf{f} = \begin{bmatrix} \delta f_1/\delta x_1 & \cdots & \delta f_1/\delta x_n \\ \vdots & \ddots & \vdots \\ \delta f_m/\delta x_1 & \cdots & \delta f_m/\delta x_n \end{bmatrix}$$

$\mathbf{f}$'s Taylor expansion is $\mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{f}(\mathbf{x}) + \nabla\mathbf{f}(\mathbf{x})\mathbf{h} + O(\|\mathbf{h}\|^2)$.

A *linear* (or *quadratic*) *model* approximates a function $\mathbf{f}$ by the first two (or three) terms of $\mathbf{f}$'s Taylor expansion.

Nonlinear Equation Solving

Given $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, we want $\mathbf{x}$ such that $\mathbf{f}(\mathbf{x}) = \mathbf{0}$.

In *fixed point iteration*, we choose $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{x}^{(k+1)} = \mathbf{g}(\mathbf{x}^{(k)})$. If it converges to $\mathbf{x}^*$, $\mathbf{g}(\mathbf{x}^*) - \mathbf{x}^* = \mathbf{0}$.

$\mathbf{g}(\mathbf{x}^{(k)}) = \mathbf{g}(\mathbf{x}^*) + \nabla\mathbf{g}(\mathbf{x}^*)(\mathbf{x}^{(k)} - \mathbf{x}^*) + O(\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2)$ For small $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}^*$, ignore the last term. If $\nabla\mathbf{g}(\mathbf{x}^*)$ has $\lambda_{|\max|} < 1$, then $\mathbf{x}^{(k)} \to \mathbf{x}^*$ as $\|\mathbf{e}^{(k)}\| \leq c^k\|\mathbf{e}^{(0)}\|$ for large k, where $c = \lambda_{|\max|} + \epsilon$, where ϵ is the influence of the ignored last term. This indicates a *linear rate of convergence*.

Suppose for $\nabla\mathbf{g}(\mathbf{x}^*) = QTQ^H$, T is *non-normal*, i.e., T's superdiagonal portion is large relative to the diagonal. Then this may not converge as $\|(\nabla\mathbf{g}(\mathbf{x}^*))^k\|$ initially grows!

In *Newton's method*, $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - (\nabla\mathbf{f}(\mathbf{x}^{(k)}))^{-1}\mathbf{f}(\mathbf{x}^{(k)})$. This converges *quadratically*, i.e., $\|\mathbf{e}^{(k+1)}\| \leq c\|\mathbf{e}^{(k)}\|^2$.

Automatic differentiation takes advantage of the notion that a computer program is nothing but arithmetic operations, and one can apply the chain rule to get the derivative. This may be used to compute Jacobians and determinants.

Optimization

In continuous optimization, $f : \mathbb{R}^n \to \mathbb{R}$ is the *objective function*, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$ holds *equality constraints*, $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^p$ holds *inequality constraints*.

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{g}(\mathbf{x}) = \mathbf{0} \\ & \mathbf{h}(\mathbf{x}) \geq \mathbf{0} \end{array}$$

We did unrestricted optimization $\min f(\mathbf{x})$ in the course.

A *ball* is a set $B(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{y}\| < r\}$.

We have *local minimizers* $\mathbf{x}^*$ which are the best in a region, i.e., $\exists r > 0$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in B(\mathbf{x}^*, r)$. A *global minimizer* is the best local minimizer.

Assume f is c^2. If $\mathbf{x}^*$ is a local minimizer, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*)$ is PSD. Semi-conversely, if $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*)$ is PD, then $\mathbf{x}^*$ is a local minimizer.

Steepest Descent

Go where the function (locally) decreases most rapidly via $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k\nabla f(\mathbf{x}^{(k)})$. α_k is explained later. SD is stateless: depends only on the current point. Too slow.

Newton's Method for Unconstrained Min.

Iterate by $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - (\nabla^2 f(\mathbf{x}^{(k)}))^{-1}\nabla f(\mathbf{x}^{(k)})$, derived by solving for where $\nabla f(\mathbf{x}^*) = \mathbf{0}$. If $\nabla^2 f(\mathbf{x}^{(k)})$ is PD and $\nabla f(\mathbf{x}^{(k)}) \neq \mathbf{0}$, the step is a descent direction.

What if the Hessian isn't PD? Use (a) secant method, (b) direction of *negative curvature* where $\mathbf{h}^T\nabla^2 f(\mathbf{x}^{(k)})\mathbf{h} < 0$ where $\mathbf{h}$ or $-\mathbf{h}$ (doesn't work well in practice), (c) *trust region* idea so $\mathbf{h} = -(\nabla^2 f(\mathbf{x}^{(k)}) + tI)^{-1}\nabla f(\mathbf{x}^{(k)})$ (interpolation of NMUM and SD), (d) factor $\nabla^2 f(\mathbf{x}^{(k)})$ by Cholesky when checking for PD, detect 0 pivots, modify that diagonal in $\nabla^2 f(\mathbf{x}^{(k)})$ and keep going (unjustified by theory, but works in practice).

Line Search

Line search, given $\mathbf{x}^{(k)}$ and step $\mathbf{h}$ (perhaps derived from SD or NMUM), finds a $\alpha > 0$ for $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha\mathbf{h}$.

In *exact line search*, optimize $\min f(\mathbf{x}^{(k)} + \alpha\mathbf{h})$ over α. Frowned upon because it's computationally expensive.

In *Armijo* or *backtrack line search*, initialize α. While $f(\mathbf{x}^{(k)} + \alpha\mathbf{h}) > f(\mathbf{x}^{(k)}) + 0.1\alpha\nabla f(\mathbf{x}^{(k)})^T\mathbf{h}$, halve α.

Secant/quasi Newton methods use an approximate always PD $\nabla^2 f$. In Broyden-Fletcher-Goldfarb-Shanno:

$B_0 =$ initial approximate Hessian {OK to use I.}
for $k = 0, 1, 2, \ldots$ **do**
  $\mathbf{s}_k = -B_k^{-1}\nabla f(\mathbf{x}^{(k)})$
  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k\mathbf{s}_k$ {Use special line search for α_k!}
  $\mathbf{y}_k = \nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\mathbf{x}^{(k)})$
  $B_{k+1} = B_k + \frac{\mathbf{y}_k\mathbf{y}_k^T}{\alpha\mathbf{y}_k^T\mathbf{s}_k} - \frac{B_k\mathbf{s}_k\mathbf{s}_k^TB_k}{\mathbf{s}_k^TB_k\mathbf{s}_k}$
end for

By maintaining B_k in factored form, can iterate in $O(n^2)$ flops. B_k is SPD provided $\mathbf{s}_k^T\mathbf{y} > 0$ (use line search to increase α_k if needed). The secant condition $\alpha_kB_{k+1}\mathbf{s}_k = \mathbf{y}_k$ holds. If BFCS converges, it converges superlinearly.

Non-linear Least Squares

For $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$, $m \geq n$, we want the $\mathbf{x}$ for $\min \|\mathbf{g}(\mathbf{x})\|_2$.

In the *Gauss-Newton* method, $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{h}$ where $\mathbf{h} = (\nabla\mathbf{g}(\mathbf{x})^T\nabla\mathbf{g}(\mathbf{x}))^{-1}\nabla\mathbf{g}(\mathbf{x})^T\mathbf{g}(\mathbf{x})$. Note that $\mathbf{h}$ is a solution to a linear least squares problem $\min\|\nabla\mathbf{g}(\mathbf{x}^{(k)})\mathbf{h} - \mathbf{g}(\mathbf{x}^{(k)})\|$! GN is derived by applying NMUM to to $\mathbf{g}(\mathbf{x})^T\mathbf{g}(\mathbf{x})$, and dropping a resulting *tensor* (derivative of Jacobian). You keep the quadratic convergence when $\mathbf{g}(\mathbf{x}^*) = \mathbf{0}$, since the tensor $\to 0$ as $k \to \infty$.

Ordinary Differential Equations

ODE (or PDE) has one (or multiple) independent variables.

In *initial value problems*, given $\frac{d\mathbf{y}}{dt} = f(\mathbf{y}, t)$, $\mathbf{y}(t) \in \mathbb{R}^n$, and $\mathbf{y}(0) = \mathbf{y}_0$, we want $\mathbf{y}(t)$ for $t > 0$. Examples include:

1. Exponential growth/decay with $\frac{d\mathbf{y}}{dt} = a\mathbf{y}$, with closed form $\mathbf{y}(t) = \mathbf{y}_0e^{at}$. Growth if $a > 0$, decay if $a < 0$.
2. Ecological models, $\frac{dy_i}{dt} = f_i(y_1, \ldots, y_n, t)$ for species $i = 1, \ldots, n$. y_i is population size, f_i encodes species relationships.
3. Mechanics, e.g. wall-spring-block models for $F = ma$ ($a = \frac{d^2x}{dt^2}$) and $F = -kx$, so $\frac{d^2x}{dt^2} = \frac{-kx}{m}$. Yields $\frac{d[x,v]^T}{dt} = \begin{bmatrix} v & \frac{-kx}{m} \end{bmatrix}^T$ with $\mathbf{y}_0$ as initial position and velocity.

For *stability of an ODE*, let $\frac{d\mathbf{y}}{dt} = A\mathbf{y}$ for $A \in \mathbb{C}^{n\times n}$. The *stable* or *neutrally spable* or *unstable* case is where $\max_i \Re(\lambda_i(A)) < 0$ or $= 0$ or > 0 respectively.

In *finite difference methods*, approximate $\mathbf{y}(t)$ by discrete points $\mathbf{y}_0$ (given), $\mathbf{y}_1, \mathbf{y}_2, \ldots$ so $\mathbf{y}_k \approx \mathbf{y}(t_k)$ for increasing t_k.

For many IVPs and FDMs, if the *local truncation error* (error at each step) is $O(h^{p+1})$, the *global truncation error* (error overall) is $O(h^p)$. Call p the *order of accuracy*.

To find p, substitute the exact solution into FDM formula, insert a remainder term $+R$ on RHS, use a Taylor series expansion, solve for R, keep only the leading term.

In *Euler's method*, let $\mathbf{y}_{k+1} = \mathbf{y}_k + \mathbf{f}(\mathbf{y}_k, t_k)h_k$ where $h_k = t_{k+1} - t_k$ is the *step size*, and $\mathbf{y}' = \mathbf{f}(\mathbf{y}, t)$ is perhaps computed by finite difference. $p = 1$, very low. Explicit!

A *stiff problem* has widely ranging time scales in the solution, e.g., a transient initial velocity that in the true solution disappears immediately, chemical reaction rate variability over temperature, transients in electical circuits. An explicit method requires h_k to be on the smallest scale!

Backward Euler has $\mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{f}(\mathbf{y}_{k+1}, t_{k+1})$. BE is *implicit* ($\mathbf{y}_{k+1}$ on the RHS). If the original program is stable, any h will work!

Miscellaneous

$\sum_{k=1}^{n\pm\text{constant}} k^p = \frac{n^{p+1}}{p+1} + O(n^p)$

$ax^2 + bx + c = 0$. $r_1, r_2 = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$. $r_1r_2 = \frac{c}{a}$

Exact arithmetic is slow, futile for inexact observations, and NA relies on approximate algorithms.