Contents

Chapter 1

Discrete Random Variables, DRVs

Beginning of March 15, 2021

1.1 Geometric Random Variables

Random experiment with repeated, independent Bernoulli trials with the probability of success p.

The **Geometric random variable** X describes the number of trials until (and including) the first success, and we write $X \sim G(p)$.

$R(X) = \{1, 2, 3, \dots\}$ with $P_X(x) = P(X = x) = (1-p)^{x-1}p$. The cumulative distribution function (cdf) is a piecewise, monotone increasing function with limit 1 but never reaches it.

Mean, Variance, and MGF

The mean of $G(p) = 1/p$. Intuitively, if each independent Bernoulli trial has as probability of success p, it takes $1/p$ trials to get one success.

$$\begin{aligned}\mu_X = E[X] &= \sum_{x=1}^{\infty} xP(x) = \sum_{x=1}^{\infty} x(1p-)^{x-1}p = p\sum_{x=1}^{\infty} x(1-p)^{x-1} \\ &= p\sum_{x=1}^{\infty} -\frac{\mathrm{d}}{\mathrm{d}p}(1-p)^x = -p\frac{\mathrm{d}}{\mathrm{d}p}\left[\sum_{x=1}^{\infty}(1-p)^x\right] = -p\frac{\mathrm{d}}{\mathrm{d}p}\left[\frac{1}{1-(1-p)} - 1\right] \\ &= -p\frac{\mathrm{d}}{\mathrm{d}p}\frac{1}{p} = -p(-1/p^2) = \frac{1}{p}.\end{aligned}$$

The variance of $G(p)$ is $(1-p)/p^2$.

$$\begin{aligned}
\sigma_X^2 &= \operatorname{var}[X] = E[X^2] - E[X]^2 \\
&= \sum_{x=1}^{\infty} x^2 P(x) - \frac{1}{p^2} \\
&= \sum_{x=1}^{\infty} x^2 (1-p)^{x-1} p - \frac{1}{p^2} \\
&= p \sum_{x=1}^{\infty} [x(x-1) + x] (1-p)^{x-2} (1-p) - \frac{1}{p^2} \\
&= p(1-p) \sum_{x=1}^{\infty} x(x-1)(1-p)^{x-2} + \sum_{x=1}^{\infty} x(1-p)^{x-1} p - \frac{1}{p^2} \\
&= p(1-p) \sum_{x=1}^{\infty} \frac{\mathrm{d}^2}{\mathrm{d}p^2} [(1-p)^x] + \frac{1}{p} - \frac{1}{p^2} \\
&= p(1-p) \frac{\mathrm{d}^2}{\mathrm{d}p^2} \left[\sum_{x=1}^{\infty} (1-p)^x \right] + \frac{1}{p} - \frac{1}{p^2} \\
&= \frac{1-p}{p^2}
\end{aligned}$$

The MGF is

$$\begin{aligned}
\varphi_X(t) = E[e^{tx}] &= \sum_{x=1}^{\infty} e^{tx} P(x) \\
&= \sum_{x=1}^{\infty} e^{tx} (1-p)^{x-1} p \\
&= p \sum_{x=1}^{\infty} e^{t(x-1)} e^t (1-p)^{x-1} \\
&= p e^t \sum_{x=1}^{\infty} \left[e^t (1-p) \right]^{x-1} \\
&= p e^t \frac{1}{1 - e^t (1-p)} && \text{notice that } |e^t(1-p)| < 1 \text{ locally} \\
&= \frac{p e^t}{1 - e^t (1-p)}.
\end{aligned}$$

1.2 Negative Binomial Random Variables

A generalization of the geometric random variable, not that of a binomial random variable. This describes the random experiment of a repeated, independent Bernoulli trials with probability of success p.

X describes the number of trials to get r successes. We write $X \sim NB(r, p)$.

It follows that $R(X) = \{r, r+1, \dots\}$, and

$$P_X(x) = P(X = x) = \binom{x-1}{r-1} p^{r-1} (1-p)^{(x-1)-(r-1)} p = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

since if we need precisely x trials to get r successes, then there must be precisely $r-1$ successes in the first $x-1$ trials, and the x^{th} trial itself must be a success.

Note that $X \sim NB(1, p) \iff X \sim G(p)$.

Mean, Variance, and MGF

The mean of $NB(r,p)$ is r/p. Let X_i be the number of trials to get another success after already having $i-1$ successes. Then $X_i \sim G(p)$, and more importantly the X_i's are independent. Therefore,

$$\sigma_X = E[X] = E\left[\sum_{i=1}^{r} X_i\right] = \sum_{i=1}^{r} E[X_i] = \frac{r}{p}.$$

The variance of X can be computed similarly, and $\sigma_X^2 = \text{var}[X] = r(1-p)/p^2$.
The MGF is

$$\begin{aligned}\varphi_X(t) &= E[e^{tX}] = E[\exp(t\textstyle\sum)] \\ &= \prod_{i=1}^{r} E[e^{tX_i}] \\ &= \left(\frac{pe^t}{1-e^t(1-p)}\right)^r.\end{aligned}$$

1.3 Poisson Distribution

The Poisson distribution is similar to binomial distribution, but the probability of success becomes a rate applied to a continuum as opposed to discrete selections.
Let a time interval of length 1 be given. Previously, without the concept of continuum, we divide this interval into n subintervals of length $1/n$ and perform independent Bernoulli trails on each one. Let the assumption be that the mean total number of successes among these n Bernoulli trials is λ. Then it immediately follows that the probability of success for each subinterval is λ/n (so that $n \cdot \lambda/n = \lambda$).
We fix this λ but let $n \to \infty$, namely dividing the interval into finer and finer subintervals of length $1/n$. What does the limit mean? It means we are on a continuum $[0,1]$ where each $x \in [0,1]$ resembles a "Bernoulli trial", such that the mean/expected number of total successes for all $x \in [0,1]$ is λ. Of course, as $n \to \infty$, $\lambda/n \to \infty$ and the previous definitions no longer make sense. Therefore, the **Poisson** distribution can be interpreted as a limit that consists of *infinitely many Bernoulli trials*.
Done with the heuristics, now we begin from finite cases (Bernoulli trials) and start the approximation. Suppose we have n independent Bernoulli trails. This naturally gives rise to a binomial distribution with parameters n and λ/n. Let X_n be $B(n, \lambda/n)$. Then

$$\begin{aligned}P_{X_n}(x) &= \binom{n}{x}(\lambda/n)^x(1-\lambda/n)^{n-x} \\ &= \frac{n!}{(n-x)!x!}\frac{\lambda^x}{n^x}(1-\lambda/x)^{n-x} \\ &= \frac{\lambda^x}{x!}(1-\lambda/n)^{-x}\frac{n!}{(n-x)!n^x}(1-\lambda/n)^n.\end{aligned}$$

Notice that $\lim_{n\to\infty}(1-\lambda/n)^n = e^{-\lambda}$:

$$\begin{aligned}\lim_{n\to\infty}\ln(1-\lambda/n)^n &= \lim_{n\to\infty} n\ln(1=\lambda/n) \\ &= \lim_{n\to\infty}\frac{\ln(1-\lambda/n)}{1/n} \\ &\overset{H}{=} \lim_{n\to\infty}\frac{\lambda/n^2\cdot 1/(1-\lambda/n)}{-1/n^2} = -\lambda.\end{aligned}$$

On the other hand,

$$\frac{n!}{(n-x)!n^x} = \frac{n(n-1)\dots(n-x+1)}{n^x} = \frac{n}{n}\cdot\frac{n-1}{n}\cdot\dots\cdot\frac{n-x+1}{n} \to 1 \text{ as } n\to\infty.$$

Therefore,

$$\lim_{n\to\infty} P_{X_n}(x) = \frac{\lambda^x}{x!}\cdot 1\cdot 1\cdot e^{-\lambda} = \frac{e^{-\lambda}\lambda^x}{x!}.$$

Notation wise, $X \sim Pr(\lambda)$. $R(X) = \{0,1,2,\dots\}$ and $P_X(x) = e^{-\lambda}\lambda^x/x!$. Notice that the probabilities add up to 1:

$$\sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1.$$

Mean, Variance, and MGF

The mean is given by

$$\begin{aligned}\mu_X = E[X] &= \sum_{x=0}^{\infty} xP(x)\\ &= \sum_{x=0}^{\infty} xe^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=1}^{\infty}\frac{x\lambda^x}{x!}\\ &= e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^x}{(x-1)!} = \lambda e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda}{x!}\\ &= \lambda e^{-\lambda}e^{\lambda} = \lambda.\end{aligned}$$

> **Remark.** Of course this makes sense! On a continuum with λ being the mean occurrence rate of an event, what else do you expect $E[X]$ to be but λ itself?

The variance is given by

$$\sigma_X^2 = \mathrm{Var}[X] = E[X^2] - \lambda^2 = \sum_{x=0}^{\infty}\frac{x^2\lambda^x e^{-\lambda}}{x!} - \lambda^2 = \lambda.$$

> **Remark.** *A nice problem assigned in HW7 gives the identity $E[X^n] = \lambda E[(X+1)^{n-1}]$ for Poisson distributions. If we apply it to $E[X^2]$, we immediately have*
>
> $$\mathrm{Var}[X] = \lambda E[(X+1)] - E[X]^2 = \lambda(\lambda+1) - \lambda^2 = \lambda.$$
>
> Alternatively, we can again use the idea of "limit of Bernoulli/binomial distributions":
>
> $$\mathrm{Var}[X] = \lim_{n\to\infty}\mathrm{Var}[B(n,\lambda/n)] = \lim_{n\to\infty} n\cdot(\lambda/n)\cdot(1-\lambda/n) = \lambda.$$

The MGF is given by

$$\varphi_X(t) = E[e^{tX}] = e^{\lambda}(e^t - 1).$$

Beginning of March 17, 2021

1.4 Hypergeometric Random Variables

Let N be the size of a population, m a number of "distinguished elements", and n the sample size. For example, let $N = 100, m = 50$, and $n = 10$. Then the **hypergeometric random variable** X gives the probability that, among a sample of n elements, exactly $X = x$ elements are "distinguished" and the remaining not "distinguished".
We write $X \sim H(N, m, n)$ where N, m, n are the parameters.
The range of X is given by $R(X) = \{\max(m - (N - n), 0), \ldots, \min(m, n)\}$, but in this class we only consider the scenario where $m + n \leqslant N$ and so $R(X) = \{0, \ldots, \min(m, n)\}$. Typically this would simply be $\{0, \ldots, n\}$.
It follows naturally from combinatorics that

$$P_X(x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}.$$

The mean and variance are given by

$$\mu_X = E[X] = n \cdot \frac{m}{N} \text{ and } \sigma_X^2 = n \cdot \frac{m}{N}\left(1 - \frac{m}{N}\right)\left(\frac{N-n}{N-1}\right).$$

We will show how to derive these later (once we cover indicator random variable). Now simply notice the similaries of $E[X]$ and $\mathrm{Var}[X]$ of hypergeometric random variables to binomial. The $(N-n)/(N-1)$ is called the **small population correction**. As $N \to \infty$ and n fixed, this correction $\to 1$ and indeed this looks more like a binomial distribution. (Note the only difference is that hypergeometric random variables are without replacements but binomials are with replacement.)

1.5 Discrete Uniform Random Variable

This is a probability distribution with a finite number of values that are equally likely to be observed. If there are k total values then each has a probability $1/k$. In general we write $X \sim \mathrm{unif}(a, b, n)$ but here we first consider the simplest case $X \sim \mathrm{unif}(0, 1, n)$. Then $R(X) = \{1/n, 2/n, \ldots, (n-1)/n\}$ (assuming $n \geqslant 2$). It immediately follows that $P_X(x) = 1/(n-1)$. We write $X \sim \mathrm{DU}(a, b, n)$. (Of course, this can be easily generated to $\mathrm{DU}(a, b, n)$.)

Mean, Variance, and MGF

It's intuitive that the mean is 1/2:

$$\begin{aligned}\mu_X = E[X] &= \sum_x xP(x) = \sum_{i=1}^{n-1} \frac{i}{n}\frac{1}{n-1} \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n-1} i = \frac{1}{n(n-1)}\frac{n(n-1)}{2} = \frac{1}{2}.\end{aligned}$$

The second central moment is given by

$$\begin{aligned}E[X^2] &= \sum_x x^2 P(x) = \sum_{i=1}^{n-1} \frac{i^2}{n^2}\frac{1}{n-1} \\ &= \frac{1}{n^2(n-1)}\sum_{i=1}^{n-1} i^2 = \frac{2n-1}{6n}.\end{aligned}$$

Therefore, the variance of X is given by

$$\sigma_X^2 = \mathrm{Var}[X] = \sigma_X^2 = E[X^2] - E[X]^2 = \frac{n-2}{12n}.$$

Beginning of March 19, 2021

The MGF is given by

$$\begin{aligned}\varphi_X(t) &= E[e^{tX}] = \sum_{x=1}^{n-1} e^{tx/n}\frac{1}{n-1}\\ &= \frac{1}{n-1}\sum_{i=1}^{n-1} e^{ti/n} = \frac{1}{n-1}\sum_{i=1}^{n-1}(e^{t/n})^i\\ &= \frac{1}{n-1}\left[\frac{e^t-1}{e^{t/n}-1}-1\right] = \frac{1}{n-1}\left[\frac{e^t-e^{t/n}}{e^{t/n}-1}\right]\\ &= \frac{e^t-e^{t/n}}{(n-1)(e^{t/n}-1)}.\end{aligned}$$

More generally, we now consider $X \sim \mathrm{DU}(a,b,n)$. Then this is simply $X = a+(b-a)U$, where $U \sim \mathrm{DU}(0,1,n)$. Then it follows that

(1) Range: $R(X) = \{a+(b-a)/n, a+2(b-a)/n, \ldots, a+(n-1)(b-a)/n\}$.

(2) Mean: $E[X] = E[a+(b-a)U] = a+(b-a)E[U] = (a+b)/2$.

(3) Variance: $\mathrm{Var}[X] = \mathrm{Var}[a+(b-a)U] = (b-a)^2\mathrm{Var}[U] = (b-a)^2(n-2)/(12n)$.

(4) MGF: $\varphi_X(t) = \varphi_{a+(b-a)U}(t) = e^{at}\varphi_U[(b-a)t]$.

1.6 Joint Distribution of DRVs

Definition 1.6.1

Given n DRVs, define a *random vector* by $\mathbf{X} = [X_1, \ldots, X_n]$. The **joint probability mass function** (jpmf) is given by

$$P_{\mathbf{X}}(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n).$$

Notice that we cannot observe what happens exactly to one variable while ignoring the rest; most of the times we can only observe the events caused simultaneously by all n random variables, hence the joint pmf.

In the scope of 407, most of the times we focus on cases where $n = 2$ so we only have two random variables, X, Y.

Definition 1.6.2: Marginalization

Assuming $n = 2$ as said above, the **marginal distributions** are given by

$$P_X(x) = \sum_y P_{X,Y}(x,y) \text{ and } P_Y(y) = \sum_x P_{X,Y}(x,y).$$

Indeed, $P_X(x)$ describes the probability of $X = x$ and no restriction is imposed on Y so it can be anything. This approach is called *marginalizing out a subcollection by summing.*

Example 1.6.3. Suppose $n = 4$. If we wanted to find the marginal distribution of X_1 and X_3, then we need to marginalize X_2, X_4 and sum them up:

$$P_{X_1,X_3}(x_1, x_3) = \sum_{X_2}\sum_{X_4} P_{\mathbf{X}}(x_1, x_2, x_3, x_4).$$

Expectation. Given $\mathbf{X} = [X_1, \ldots, X_n]$ and $g : \mathbb{R}^n \to \mathbb{R}$, let

$$Y = g(X_1, \ldots, X_n).$$

Then

$$E[Y] = \sum_{x_1} \cdots \sum_{x_n} g(x_1, \ldots, x_n) P_Y(x_1, \ldots, x_n).$$

If we let $n = 2$, given X, Y and $g : \mathbb{R}^2 \to \mathbb{R}$ and $Z = g(X, Y)$, we have

$$E[Z] = \sum_x \sum_y g(x, y) P_{X,Y}(x, y).$$

Example 1.6.4. Consider $Z = X + Y$. Then

$$\begin{aligned} E[Z] = E[X + Y] &= \sum_x \sum_y (x + y) P_{X,Y}(x, y) \\ &= \sum_x \sum_y x P_{X,Y}(x, y) + \sum_x \sum_y y P_{X,Y}(x, y) \\ &= \sum_x x \sum_y P_{X,Y}(x, y) + \sum_y y \sum_x P_{X,Y}(x, y) \\ &= \sum_x x P_Z(x) + \sum_y y P_Z(y) \\ &= E[X] + E[Y]. \end{aligned}$$

This can be easily generated:

$$E[\sum_{i=1}^n X_i] = \sum_{i=1}^n E[X_i].$$

Also, recall that $E[cX] = cE[X]$. Along with the results shown above, we see that $E[\cdot]$ is a **linear transformation**:

$$E[\sum_{i=1}^n c_i X_i] = \sum_{i=1}^n c_i E[X_i].$$

Variance. We begin by again considering $Z = X + Y$. Recall that $\mathrm{Var}[Z] = E[Z^2] - E[Z]^2$. The latter is

$$E[Z]^2 = (E[X] + E[Y])^2 = E[X]^2 + 2E[X]E[Y] + E[Y]^2.$$

On the other hand,

$$E[Z^2] = E[(X + Y)^2] = E[X^2 + 2XY + Y^2] = E[X^2] + 2E[XY] + E[Y]^2$$

Therefore,

$$\begin{aligned}\sigma_Z^2 = \mathrm{Var}[Z] &= E[X^2] + 2E[XY] + E[Y^2] - E[X]^2 - 2E[X]E[Y] - E[Y]^2 \\ &= (E[X^2] - E[X]^2) + (E[Y^2] - E[Y]^2) + 2(E[XY] - E[X]E[Y]) \\ &= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2(E[XY] - E[X]E[Y]).\end{aligned}$$

If $E[XY] = E[X]E[Y]$, i.e., mean of product = product of means, then

$$\mathrm{Var}[Z] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

Notice that this is not necessarily true at all!! This naturally leads to the following definitions.

1.7 Covariance & Correlation Coefficient

Definition 1.7.1: Covariance

The **covariance** of X and Y is given by

$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Proposition 1.7.2

We have the following properties of $\mathrm{Cov}(\cdot, \cdot)$.

(1) $\mathrm{Cov}(X, Y) = E[XY] - \mu_X \mu_Y$.

(2) ***$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.

(3) ***Covariance is bilinear, i.e.,

$$\mathrm{Cov}(\lambda X + Y, Z) = \lambda \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$$

and same for the other argument since covariance is commutative.

(4) ***$\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

(5) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$, as shown above.

(6) Covariance is invariant under constant addition: $\mathrm{Cov}(X + a, Y) = \mathrm{Cov}(X, Y)$.

Definition 1.7.3: Covariance Matrix

Given $\mathbf{X} = [X_1, \ldots, X_n]$, we define $\Sigma \in \mathbb{R}^{n \times n}$, called the **covariance matrix**, by

$$\Sigma_{i,j} = \mathrm{Cov}(X_i, X_j).$$

Proposition 1.7.4

Immediately following the properties of co variance, the following holds for Σ:

(1) $\Sigma = \Sigma^T$ since covariance is commutative.

(2) Σ is PSD (positive semi-definite), i.e., $a^T \Sigma a \geqslant 0$.

(3) ***Given $(a_1, \ldots, a_n)^T = a \in \mathbb{R}^n$, then

$$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i) = a^T \Sigma a.$$

(*This explains why Σ is PSD.*)

In particular, taking $a = (1, \ldots, 1)^T$ gives

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_i \sum_j \Sigma_{i,j} = \sum_i \mathrm{Var}(X_i) + 2 \sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

The first summation is because $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$ and the second because Σ is symmetric.

(4) For $n = 2$, we have

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(Y, X) & \sigma_Y^2 \end{bmatrix}.$$

Interpretation of Covariance. Take it for granted that $\mathrm{Cov}(X, Y) > 0$ (resp. < 0) if X and Y tend to be on the same (resp. opposite) side(s) of their means with high probability. Why? By definition

$$E[(X - \mu_X)(Y - \mu_Y)],$$

if both terms tend to be positive or both tend to be negative then $E[\cdot]$ tends to be positive; vice versa.

Beginning of March 22, 2021

Proof of $\mathrm{Cov}(X, Y) = E[XY] - \mu_X \mu_Y$. Indeed,

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\ &= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X \mu_Y \\ &= E[XY] - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \\ &= E[XY] - \mu_X \mu_Y. \end{aligned}$$

□

What if $\mathrm{Cov}(X, Y) = 0$? In this case we simply have

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \text{ and } \sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}.$$

Definition 1.7.5

We say two DRVs are **independent** if for any two subsets $A, B \subset \mathbb{R}$,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Remark. We've already shown that if X, Y are independent then

$$P_{X,Y}(x, y) = P(X = x, Y = y) = P(X = x)P(Y = y) = P_X(x)P_X(y),$$

i.e., joint pmf = product of marginal pmf.

Theorem 1.7.6

If X, Y are independent, then $E[XY] = E[X]E[Y] = \mu_X\mu_Y$ and

(1) $\mathrm{Cov}(X, Y) = 0$,

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

More generally, if X_i, X_j are pairwise independent, then

$$\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

A natural question is about the converse: if $E[XY] = E[X]E[Y]$, does it mean X, Y are independent? The answer is no in general.

For example, consider a random variable that takes values $1, 0, -1$, each with a probability of $1/3$. Let Y be $|X|$ so $P_Y(1) = 2/3$ and $P_Y(0) = 1/3$. Then XY can be $1, 0, -1$ for a probability of $2/9, 5/9, 2/9$, respectively. Then $E[XY] = 0$ and so is $E[X]$ and thus $E[X]E[Y] = 0$. However,

$$P_{X,Y}(1, 1) = P(X = 1, |X| = 1) = \frac{1}{3} \text{ and } P_X(1)P_Y(1) = \frac{2}{9}.$$

Definition 1.7.7

The **correlation coefficient** of X, Y is given by

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

Proposition 1.7.8

Properties of correlation coefficient:

(1) If X, Y are independent then $\rho_{X,Y} = 0$.

(2) If $\rho(X, Y) = 0$ then X, Y are **uncorrelated** (not necessarily independent).

(3) $-1 \leqslant \rho(X,Y) \leqslant 1$. *Brief proof: Cauchy-Schwarz!! Let α be a scalar and use the fact that*

$$\operatorname{Var}(X+\alpha Y) = \operatorname{Var}(X) + \alpha^2 \operatorname{Var}(Y) + 2\alpha \operatorname{Cov}(X,Y) \geqslant 0.$$

This gives a nonnegative discriminant, i.e.,

$$4\operatorname{Cov}(X,Y)^2 \geqslant 4\operatorname{Var}(X)\operatorname{Var}(Y)$$

and so factoring and taking square roots gives the claim.

Suppose $Y = aX + b$. Clearly Y and X are dependent (actually, *very, very* dependent). Then

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \operatorname{Cov}(X, aX+b)\\ &= a\operatorname{Cov}(X,X) + \operatorname{Cov}(X,b)\\ &= a\operatorname{Cov}(X,X) = a\sigma_X^2.\end{aligned}$$

Therefore,

$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{a\sigma_X^2}{\sigma_X|a|\sigma_X} = \frac{a}{|a|} = \operatorname{sgn}(a)$$

This is intuitive. If X, Y are linearly related then $\rho(X,Y) = \pm 1$. Recall that $\rho(X,Y) \in [-1,1]$, so $Y = aX + b$ is as correlated as they can get. Indeed, *very, very* positively / negatively correlated.

Beginning of March 24, 2021

Example 1.7.9. Recall: let $X \sim \mathrm{NB}(r,p)$ be a negative binomial r.v. Notice that the relative position of the $(n+1)^{\text{th}}$ appearance does not depend on that of the n^{th} appearance. Therefore

$$X = X_1 + X_2 + \cdots + X_r \text{ where the } X_i\text{'s are i.i.d.} G(p).$$

(i.i.d. stands for *independent and identically distributed*). It follows that

$$E[X] = E[\sum_{i=1}^{r} X_i] = \frac{r}{p},$$

and

$$\operatorname{Var}(X) = \operatorname{Var}(\sum_{i=1}^{r} X_i) = \frac{r(1-p)}{p^2}.$$

Example 1.7.10. Similarly, $X \sim B(n,p)$ can be interpreted as $X = I_1 + \cdots + I_n$ where each I_i is an indicator variable (success gives 1 and failure gives 0).

1.8 Mean and Variance of Hypergeometric R.V. Revisited

Suppose $X \sim H(N, m, n)$ where N is the total size of population, m the number of distinguished objects, and n the sample size. We number the distinguished objects as $1, 2, \ldots, m$ and define some indicator r.v. for $i = 1, 2, \ldots, m$ by

$$I_i = \begin{cases} 1 & i^{\text{th}} \text{ distinguished object chosen;} \\ 0 & \text{otherwise.} \end{cases}$$

It follows that (recall that whether I_n's are independent or not doesn't affect what's below)

$$X = \sum_{i=1}^{m} I_i \implies E[X] = E[\sum_{i=1}^{m} I_i] = \sum_{i=1}^{m} E[I_i].$$

Note that

$$E[I_i] = 1 \cdot P(i^{\text{th}} \text{ distinguished object in sample} + 0 \cdot P(\text{not in}) = \binom{N-1}{n-1}\binom{N}{n}^{-1} = \frac{n}{N}.$$

where the "denominator" describes the total number of options and the numerator describes the total number of samples satisfying the requirement (i^{th} fixed so it's $(N-1)$ choose $(n-1)$). Therefore

$$E[X] = \sum_{i=1}^{m} \frac{n}{N} = \frac{mn}{N} = n(m/N).$$

(Compare this with binomial's mean of np.)

Now we compute the variance of X:

$$\text{Var}(X) = \text{Var}(\sum_{i=1}^{m} I_i) = \sum_{i=1}^{m} \text{Var}(I_i) + 2\sum_{i<j} \text{Cov}(I_i, I_j).$$

The variances of I_i's are easy enough to compute:

$$\text{Var}(I_i) = E[I_i^2] - E[I_i]^2 = \frac{n}{N} - \frac{n^2}{N^2} = \frac{n(N-n)}{N^2}.$$

Now we compute the pairwise covariance (the computation of $E[I_i I_j]$ is similar to that of $E[I_i]$ except now both i and j needs to be picked in order to let the indicator r.v. output 1):

$$\begin{aligned} \text{Cov}(I_i, I_j) &= E[I_i I_j] - E[I_i]E[I_j] = E[I_i I_j] - \frac{n^2}{N^2} \\ &= \binom{N-2}{n-2}\binom{N}{n}^{-1} - \frac{n^2}{N^2} \\ &= \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2}. \end{aligned}$$

Therefore,

$$\begin{aligned} \text{Var}(X) &= \sum_{i=1}^{m} \text{Var}(I_i) + 2\sum_{i<j} \text{Cov}(I_i, I_j) \\ &= \sum_{i=1}^{m} \frac{n(N-n)}{N^2} + 2\sum_{i<j}\left[\frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2}\right] \\ &= \frac{mn(N-n)}{N^2} + m(m-1)\left[\frac{n(n-1)}{N(n-1)} - \frac{n^2}{N^2}\right] \\ &= n \cdot \frac{m}{N} \cdot \left(1 - \frac{m}{N}\right) \cdot \left(\frac{N-n}{N-1}\right). \end{aligned}$$

Compare this with $np(1-p)$ times the small population correction.

Remark. Alternatively, we can define indicator variables $J_1, \dots, J_n$ by

$$J_i = \begin{cases} 1 & i^{\text{th}} \text{ element in sample is distinguished} \\ 0 & \text{otherwise.} \end{cases}$$

Then if $X \sim H(N, m, n)$ we have $X = \sum_{i=1}^{n} J_i$. Analogous computations of $E[X]$ and $\text{Var}[X]$ then follow.

Proposition 1.8.1

If X, Y are independent then, for $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, $f(X), g(X)$ are independent.

$$\begin{aligned} P(f(X) \in A, g(Y) \in B) &= P(X \in f^{-1}(A), Y \in g^{-1}(B)) \\ &= P(X \in f^{-1}(A))P(Y \in g^{-1}(B)) \\ &= P(f(X) \in A) \cdot P(g(Y) \in B). \end{aligned}$$

Proposition 1.8.2

If X, Y are independent then

$$M_{X+Y}(t) = M_X(T)M_Y(t),$$

i.e., the MGF of the sum is the product of the MGFs, a result directly from the property of exponentials:

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}].$$

The last equation is guaranteed by the independence of X and Y which implies the independence of e^{tX} and e^{tY} by the proposition above.
Similarly, if $X_i \sim X_j$ i.i.d. for $i = 1, \dots, n$, then

$$M_{X_1+\cdots+X_n}(t) = M_X(t)^n.$$

For example, if $X \sim G(p)$ (geometric) then $M_X(t) = pe^t/(1-(1-p)e^t)$. Therefore the negative binomial random variable $\text{NB}(r, p)$ is simply the previous MGF raised to the r^{th} power.

1.9 Distribution of a Sum; Convolution "Sneak Peek"

Suppose we have random variables X, Y with pmf P_X and P_Y, respectively. We now compute the pmf of $X + Y$:

$$\begin{aligned} P_{X+Y}(z) &= P(X + Y = z) \\ &= \sum_x P(X = x, Y = z - x) \\ &= \sum_x P(Y = z - x \mid X = x)P(X = x). \end{aligned}$$

If X, Y are independent, then this simply becomes

$$P_{X+Y}(z) = \sum_x P(X = x)P(Y = z - x) =: (P_X * P_Y)(z).$$

This is the **convolution product** of P_X and P_Y. It follows that convolution is commutative.

Example 1.9.1. Suppose $X \sim B(n,p)$ and $Y \sim B(m,p)$. Then it immediately becomes clear that $Z := X+Y$ has range $R(Z) = \{0, 1, \ldots, m+n\}$. Then,

$$\begin{aligned}
P_Z(k) &= \sum_{j=0}^{k} P_X(j)P_Y(k-j) \\
&= \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j} \binom{m}{k-j} p^{k-j}(1-p)^{m-(k-j)} \\
&= \sum_{j=0}^{k} \binom{n}{j}\binom{m}{k-j} p^k (1-p)^{m+n-k} \\
&= p^k(1-p)^{m+n-k} \sum_{j=0}^{k} \binom{n}{j}\binom{m}{k-j} \\
&= \binom{m+n}{k} p^k (1-p)^{m+n-k} \sum_{j=0}^{k} \left[\binom{n}{j}\binom{m}{k-j}\binom{m+n}{k}^{-1} \right].
\end{aligned}$$

Consider $U \sim H(m+n, n, k)$. The total probability of all possible outcomes must be 1, i.e.,

$$\sum_{j=0}^{k} \left[\binom{n}{j}\binom{m}{k-j}\binom{m+n}{k} \right] = 1,$$

and thus $P_Z(k) \sim B(m+n, p)$, which agrees with our intuition.

Beginning of March 26, 2021

Chapter 2

Continuous Random Variables, CRVs

2.0 Introduction (hence the numbering 2.0)

Definition 2.0.1

(Within the scope of 407:) $f : \mathbb{R} \to \mathbb{R}$ is said to be a **probability density function** (pdf) if

(1) $f(x) \geqslant 0$ for all $x \in \mathbb{R}$, and

(2) $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x := \lim_{L\to\infty} \int_{-L}^{L} f(x)\,\mathrm{d}x = 1.$

Definition 2.0.2

Let $\{\Omega, \Sigma, P\}$ be a probability space. $X : \Omega \to \mathbb{R}$ is called a **continuous random variable** if there exists f (or f_X), a pdf, such that

$$P(X \in A) = \int_A f(x)\,\mathrm{d}x.$$

It follows that

$$P(a < X \leqslant b) = \int_a^b f_X(x)\,\mathrm{d}x.$$

Remark. In the scope of 407, the $\leqslant$ and $<$ above can be interchanged (or even replaced by two $<$'s or two $\leqslant$'s) without having any effect on the integral $\int_a^b f_X(x)\,\mathrm{d}x$. Heuristically, note that

$$P(X = a) = \lim_{\epsilon\to 0} P(a - \epsilon < X \leqslant a + \epsilon) = \lim_{\epsilon\to 0} \int_{a-\epsilon}^{a+\epsilon} f_x(X)\,\mathrm{d}x = \int_a^a f_X(x)\,\mathrm{d}x = 0,$$

and

$$P(X \leqslant a) = P(\{X < a\} \sqcup \{X = a\}) = P(X < a) + P(X = a) = P(X < a).$$

Therefore, for a continuous r.v., it makes little sense to ask $P(X = a)$ since $P(x = a) = 0$. This does not mean that $X = a$ never happens; instead, it means that the long-term relative frequency of its occurrence is 0. It is just "*very, very* rare".

Definition 2.0.3

The **cumulative distribution function**, cdf, is defined by

$$F_X(x) = P(X \leqslant x) = \int_{-\infty}^{x} f_X(\tilde{t})\,\mathrm{d}\tilde{t}.$$

Remark. Since

$$P(X \leqslant b) = P(\{X \leqslant a\} \sqcup \{a < X \leqslant b\}) = P(X \leqslant a) + P(a < X \leqslant b),$$

we have

$$P(a < X \leqslant b) = P(X \leqslant b) - P(X \leqslant a) = F(b) - F(a).$$

From the previous remark,

$$P(X \in [a,b]) = P(X \in (a,b]) = P(X \in [a,b)) = P(x \in (a,b)) = F(b) - F(a).$$

Expectation, Variance, and MGF

Expectation. For $X \sim f_X$ and $g : \mathbb{R} \to \mathbb{R}$,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x.$$

If we take $g := \mathrm{id}_x$, we have

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x.$$

Variance. Likewise,

$$\sigma_X^2 = E[(X-\mu_X)^2] = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\,\mathrm{d}x = \mathrm{Var}(X).$$

Notice that

$$\begin{aligned}
\mathrm{Var}(X) = \sigma_X^2 &= \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} (x^2 - 2x\mu_X + \mu_X^2) f_X(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x - 2\mu_X \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x + \mu_X^2 \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x \\
&= E[X^2] - 2\mu_X^2 + \mu_X^2 \\
&= E[X^2] - \mu_X^2.
\end{aligned}$$

Clearly we can relate this to $\mathrm{Var}(X) = E[X^2] - E[X]^2$ for a DRV.

MGF. The computations are all analogous to those of DRVs:

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,\mathrm{d}x \text{ and } E[X^k] = M_X^{(k)}(0).$$

($M_X^{(k)}(0)$ is called the k^{th} central moment. The 1^{st} central moment is the mean and the 2^{nd} the variance.)

Affine maps. Again, analogously we have

(1) $E[aX+b] = aE[X]+b \implies \mu_{aX+b} = a\mu_X + b$,

(2) $\operatorname{Var}(aX+b) = a^2 \operatorname{Var}(x)$ (recall that variance are invariant under ± constant), and

(3) $\sigma_{aX+b} = \sqrt{\operatorname{Var}(aX+b)} = |a|\sigma_X$.

Why does the MGF work?

For the first central moment (heuristically... DCT taken for granted, of course),

$$\begin{aligned} M_X'(0) &= \frac{\mathrm{d}}{\mathrm{d}t} M_X(t) \text{ at } t=0 \\ &= \frac{\mathrm{d}}{\mathrm{d}t} \int_{-\infty}^{\infty} e^{tx} f_X(x)\,\mathrm{d}x \text{ at } t=0 \\ &= \int_{-\infty}^{\infty} \frac{\partial}{\partial t} e^{tx} f_X(x)\,\mathrm{d}x \text{ at } t=0 \\ &= \int_{-\infty}^{\infty} x e^{tx} f_X(x)\,\mathrm{d}x \text{ at } t=0 \\ &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = E[X]. \end{aligned}$$

Lemma 2.0.4

If X is a CRV with $X \geqslant 0$, then

$$E[X] = \int_0^{\infty} P(X>x)\,\mathrm{d}x = \int_0^{\infty} 1 - F_X(x)\,\mathrm{d}x.$$

Proof. Indeed, first notice that

$$P(X>x) = P(\{X \leqslant x\}^c) = 1 - P(X \leqslant x) = 1 - F_X(x).$$

Therefore,

$$\int_0^{\infty} P(X>x)\,\mathrm{d}x = \int_0^{\infty} 1 - F_X(x)\,\mathrm{d}x.$$

Now, taking Fubini's theorem for granted, we have

$$\begin{aligned} \int_0^{\infty} P(X>x)\,\mathrm{d}x &= \int_0^{\infty}\int_x^{\infty} f_X(y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_0^{\infty}\int_0^{y} f_X(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^{\infty} f_X(y)\int_0^{y} \mathrm{d}x\,\mathrm{d}y \\ &= \int_0^{\infty} y f_X(y)\,\mathrm{d}y = E[X]. \end{aligned}$$

□

Example 2.0.5. Consider

$$f_X(x) = \begin{cases} c(1-x^2) & -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

There is a unique $c \in \mathbb{R}$ that makes f_X a pdf. As long as $c \geqslant 0$, f_X is always nonengative, so it remains to find the c that satisfies the second condition, i.e., integral evaluates to 1. Hence,

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = \int_{-1}^{1} c(1-x)^2\,\mathrm{d}x = \frac{4c}{3} = 1 \implies c = \frac{3}{4}.$$

To express f_X in terms of *indicator functions*, one may use

$$f_X(x) = \frac{3}{4}(1-x^2)\chi_{[-1,1]}(x) \text{ or } \frac{3}{4}(1-x^2)\mathbb{1}_{[-1,1]}(x).$$

For another computation exercise,

$$P(-1.5 < X < 0.5) = \int_{-1.5}^{0.5} f_X(x)\,\mathrm{d}x = \int_{-1}^{0.5} \frac{3}{4}(1-x^2)\,\mathrm{d}x = \frac{27}{32}.$$

Validity check: the answer is indeed between 0 and 1. *Checked.*

Now we compute the cdf F_x of f_x:

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y = \begin{cases} 0 & -\infty < x \leqslant -1 \\ \int_{-1}^{x} 3(1-y^2)/4\,\mathrm{d}y & -1 < x \leqslant 1 \\ 1 & 1 < x < \infty \end{cases}$$

Since f_X is an even function, $E[X]$ *has to be* 0. Indeed,

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_{-1}^{1} \frac{3x}{4}(1-x^2)\,\mathrm{d}x = \int_{-1}^{1} \frac{3x}{4} - \frac{3x^3}{4}\,\mathrm{d}x = 0$$

The variance is

$$\sigma_X^2 = E[X^2] - \mu_X^2 = E[X^2] = \int_{-1}^{1} \frac{3x^2}{4}(1-x^2)\,\mathrm{d}x = \frac{1}{5}.$$

The MGF is disgusting so we will only provide the integral without evaluating it:

$$M_X(t) = E[e^{tX}] = \frac{3}{4}\int_{-1}^{1} e^{tx}(1-x^2)\,\mathrm{d}x.$$

Now we give a *dictionary* of common CRVs before moving into these topics:

(1) Uniform: as suggested by the name.

(2) Exponential: this is related to the Poisson distribution.

(3) *** Normal: the most important one!

(4) Gamma *&* Beta: useful in modeling populations.

(5) χ^2 (Chi squared): extremely important in statistics.

(6) Cauchy *&* Pareto, which we'll both cover (hopefully).

2.1 Uniform Distribution

It is somewhat analogous to the equally likely outcome, except recall that we do not talk about $P(X = a)$. Instead, we say $X \sim U(a,b)$ for $a < b$ if the pdf satisfies

$$f_X(x) = \frac{1}{b-a}\chi_{[a,b]}(x) = \begin{cases} 0 & -\infty < x \leqslant a \\ 1/(b-a) & a < x \leqslant b \\ 0 & b < x < \infty \end{cases}$$

Note that $f_X(x) \geqslant 0$ for all $x \in \mathbb{R}$, and indeed

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = \frac{1}{b-a}\int_{-\infty}^{\infty} \mathrm{d}x = \frac{1}{b-a}\int_a^b \mathrm{d}x = 1.$$

Then,

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}x = \int_{-\infty}^{x} \frac{1}{b-a}\chi_{[a,b]}\,\mathrm{d}x = \begin{cases} 0 & -\infty < x \leqslant a \\ \dfrac{1}{b-a}\displaystyle\int_a^b \mathrm{d}x = \frac{x-a}{b-a} & a < x \leqslant b \\ 1 & b < x < \infty. \end{cases}$$

Beginning of March 29, 2021

Mean, Variance, and MGF

Intuitively the mean should simply be $(a+b)/2$, and indeed

$$\mu_X = E[X] = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x = \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

The variance is given by

$$\sigma_X^2 = \mathrm{Var}[X] = E[X^2] - E[X]^2 = \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}$$

and so $\sigma_X = (b-a)/\sqrt{12}$. Compare these with the mean and variance of a discrete uniform random variable. Recall that if $X \sim \mathrm{DU}(a,b,n)$ then

$$\mathrm{Var}[X] = \frac{(b-a)(n-2)}{12n} \implies \lim_{n\to\infty} \mathrm{Var}[X] = \lim_{n\to\infty} \frac{(b-a)^2(n-2)}{12n} = \frac{(b-a)^2}{12}.$$

Intuitively, $P(X \in (r,s))$ measures the probability that a random point in (a,b) falls within (r,s), so

$$P(X \in (r,s)) = P(r < X < s) = F_X(s) - F_X(r) = \frac{s-a}{b-a} - \frac{r-a}{b-a} = \frac{s-r}{b-a}.$$

The MGF is given by

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx}f(x)\,\mathrm{d}x = \int_a^b \frac{e^{tx}}{b-a}\,\mathrm{d}x = \begin{cases} \dfrac{e^{tb}-e^{ta}}{t(b-a)} & t \neq 0 \\ 1 & t = 0. \end{cases}$$

In fact, one can check that $M_X(t)$ is continuous and differentiable at $t = 0$.

2.2 Exponential Distribution

We write $X \sim \text{Exp}(\lambda)$ for $\lambda > 0$ if X is an exponential r.v. The pdf is given by

$$f_X(x) = \lambda e^{-\lambda x} \chi_{[0,\infty)}(x) = \begin{cases} \lambda e^{-\lambda} x & x \geqslant 0 \\ 0 & x < 0. \end{cases}$$

Heuristically, think of a diminishing tail as $x \to \infty$. Clearly f is always nonnegative. We now check that the improper integral evaluates to 1 so that it makes a pdf:

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_0^{\infty} \lambda e^{-\lambda x}\,\mathrm{d}x = \lim_{L\to\infty} \int_0^L \lambda e^{-\lambda x}\,\mathrm{d}x = \lim_{L\to\infty} -e^{-\lambda L}\Big|_{x=0}^{L} = 1.$$

The cdf is given by

$$F_X(x) = P(\{X \leqslant x\}) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \begin{cases} 0 & x < 0 \\ \displaystyle\int_0^x \lambda e^{-\lambda t}\,\mathrm{d}t = 1 - e^{-\lambda x} & x \geqslant 0 \end{cases} = (1 - e^{-\lambda x})\chi_{[0,\infty)}(x).$$

So, if $0 \leqslant a < b < \infty$ we have $P(X \in (a,b)) = e^{-\lambda a} - e^{-\lambda b}$.

Mean, Variance, and MGF

The MGF is given by

$$\begin{aligned} M_x(T) = E[e^{tX}] &= \int_{-\infty}^{\infty} e^{tx} f(x)\,\mathrm{d}x = \int_0^{\infty} \lambda e^{tx} e^{-\lambda x}\,\mathrm{d}x \\ &= \lambda \int_0^{\infty} e^{(t-\lambda)x}\,\mathrm{d}x = \frac{\lambda}{1-\lambda} e^{t-\lambda x}\Big|_{x=0}^{\infty} \\ &= \frac{\lambda}{\lambda - t} \text{ if } t < \lambda \text{ and } \infty \text{ if } t \geqslant \lambda. \end{aligned}$$

Note that the MGF is defined only for $t < \lambda$.

The mean is given by

$$\begin{aligned} \mu_X = E[X] &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \lambda \int_0^{\infty} x e^{-\lambda x}\,\mathrm{d}x \\ &= [\text{integration by parts}] \\ &= \frac{1}{\lambda}. \end{aligned}$$

Alternatively, we can first compute the MGF and use the fact that $\mu_X = E[X] = M_X'(0) = 1/\lambda$. See below.

The variance is given by

$$\begin{aligned} \sigma_X^2 = \text{Var}[X] &= \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x - \frac{1}{\lambda^2} \\ &= \lambda \int_0^{\infty} x^2 e^{0\lambda x}\,\mathrm{d}x - \frac{1}{\lambda^2} \\ &= [\text{integration by parts}] = \frac{1}{\lambda^2}. \end{aligned}$$

Alternatively, we can again use MGF and get

$$\sigma_x^2 = E[X^2] - E[X]^2 = M_X''(0) - \frac{1}{\lambda^2} = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Therefore the standard deviation is $\sigma_X = 1/\lambda$. Relate this with the Poisson r.v.

2.3 Normal Distribution

We say $X \sim N(\mu, \sigma^2)$ is a **normal distribution** (with mean μ and variance σ^2; some textbooks use μ and σ instead) if X has pdf

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

It follows that the cdf is given by

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-(t-\mu)^2/(2\sigma^2))\, \mathrm{d}t.$$

This integral does not have an elementary anti-derivative, so we'll leave it just like this.

Standard Normal

We say $Z \sim N(0,1) \coloneqq \varphi$ is the **standard normal distribution** if it's the normal distribution with mean 0 and variance 1. In 407, Z is assumed to be denoting the standard normal. It follows that its pdf is

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

Then the cdf of the standard normal distribution is given by (written Φ)

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, \mathrm{d}t.$$

Claim: φ defines a pdf. Clearly $\varphi(x) \geqslant 0$, so it remains to check the integral. Let

$$I = \int_{-\infty}^{\infty} e^{-x^2/2}\, \mathrm{d}x.$$

It follows that

$$\begin{aligned}
I^2 &= \int_{-\infty}^{\infty} e^{-x^2/2}\, \mathrm{d}x \int_{-\infty}^{\infty} e^{-y^2/2}\, \mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2/2} e^{-y^2/2}\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2/2}\, \mathrm{d}x\, e^{-y^2/2}\, \mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\, \mathrm{d}x\, \mathrm{d}y \\
&= \iint_{\mathbb{R}^2} \exp(-r^2(\cos^2\theta + \sin^2\theta)/2) \begin{vmatrix} \partial x/\partial r & \partial x/\partial\theta \\ \partial y/\partial r & \partial y/\partial\theta \end{vmatrix} \mathrm{d}r\, \mathrm{d}\theta \qquad \text{(recall Jacobian; } x = r\cos\theta, y = r\sin\theta.\text{)} \\
&= \iint_{\mathbb{R}^2} e^{-r^2/2} r\, \mathrm{d}r\mathrm{d}\theta = \int_0^{2\pi}\int_0^{\infty} r e^{-r^2/2}\, \mathrm{d}r\, \mathrm{d}\theta \\
&= \int_0^{2\pi} -e^{-u}\Big|_{u=0}^{\infty}\, \mathrm{d}\theta = \int_0^{2\pi} 1 - 0\, \mathrm{d}\theta = 2\pi.
\end{aligned}$$

Therefore $I^2 = 2\pi \implies I = \sqrt{2\pi}$ and $\int_{-\infty}^{\infty} \varphi(x)\, \mathrm{d}x = I/\sqrt{2\pi} = 1$.

Theorem 2.3.1: the most beautiful quote in the history of mathematics

Regarding the identity $\int_{-\infty}^{\infty} e^{-x^2}\, \mathrm{d}x = \sqrt{\pi}/2$, Lord Kelvin once said,

> *A mathematician is one to whom that is as obvious as twice two makes four to you.*

Quoted from Spivak, *Calculus on Manifolds*.

It follows that $P(Z \in (a,b)) = \Phi(b) - \Phi(a)$, and usually this is done by checking a table that has been provided.

Beginning of March 31, 2021

Standard Normal: Mean, Variance, and MGF

Intuitively, the mean should be 0, as the pdf is symmetric along $x = 0$:

$$\begin{aligned}\mu_Z = E[Z] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2} \, \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -\frac{\mathrm{d}}{\mathrm{d}x}\left[e^{-x^2/2}\right] \, \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \left[-e^{-x^2/2}\right]_{x=-\infty}^{\infty} = \frac{1}{\sqrt{2\pi}}(0-0) = 0.\end{aligned}$$

Since the standard normal is defined to be with variance 1, it's no surprise that it indeed is:

$$\begin{aligned}\mathrm{Var}(Z) = E[Z^2] - 0^2 = E[Z^2] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} \, \mathrm{d}x \\ &= [\text{integration by parts}] = 1.\end{aligned}$$

The MGF is given by (surprisingly, this simplifies nicely)

$$\begin{aligned}M_Z(t) = E[e^{tZ}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{x^2/2} \, \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-(x^2 - 2tx)/2) \, \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-(x^2 - 2tx + t^2 - t^2)/2) \, \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-(x-t)^2/2) \exp(t^2/2) \, \mathrm{d}x \\ &= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-(x-t)^2/2) \, \mathrm{d}x \\ &= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} \, \mathrm{d}u = e^{t^2/2}.\end{aligned}$$

Notice that this gives a much nicer way to compute the variance:

$$\mathrm{Var}(Z) = E[Z^2] = M_Z''(0) = \left[te^{t^2/2}t + e^{t^2/2}\right]_{t=0} = 1.$$

General Normal

Notice that if $X \sim N(\mu, \sigma^2)$ then

$$X = \sigma z + \mu.$$

Immediately we have

$$\mu_X = E[X] = E[\sigma Z + \mu] = \mu + \sigma \cdot 0 = \mu \text{ and } \sigma_X^2 = \mathrm{Var}(X) = \mathrm{Var}(\sigma Z + \mu) = \sigma^2 \, \mathrm{Var}(Z) = \sigma^2.$$

The cdf is given by

$$\begin{aligned}F_X(x) = P(X \leqslant x) = P(\sigma Z + \mu \leqslant x) &= P(Z \leqslant (x-\mu)/\sigma) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-t^2/2} \, \mathrm{d}t.\end{aligned}$$

Notice that if we differentiate the cdf, we indeed recover the pdf of a general normal distribution:

$$\begin{aligned} F_X'(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(x-\mu)/\sigma} e^{-t^2/2}\,\mathrm{d}t \\ &= \frac{1}{\sqrt{2\pi}}\frac{\mathrm{d}}{\mathrm{d}x}\int_{-\infty}^{(x-\mu)/\sigma} e^{-t^2/2}\,\mathrm{d}t \\ &= \frac{1}{\sqrt{2\pi}}\cdot\exp[-((x-\mu)/\sigma)^2/2]\cdot\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{x-\mu}{\sigma}\right] \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = f_X(x). \end{aligned}$$

General Normal: Mean, Variance, and MGF

The MGF of $N(\mu,\sigma^2)$ is given by

$$\begin{aligned} M_X(t) &= E[e^{tX}] = E[e^{t(\sigma Z+\mu)}] \\ &= e^{t\mu}E[e^{\sigma tZ}] = e^{t\mu}M_Z(\sigma t) = e^{t\mu}e^{\sigma^2t^2/2} \\ &= \exp\left(t\mu+\frac{\sigma^2t^2}{2}\right). \end{aligned}$$

Not surprisingly, the mean and variance are μ and σ^2, respectively:

$$\mu_X = E[X] = M_x'(0) = \left[(\sigma^2t+\mu)\exp\left(t\mu+\frac{\sigma^2t^2}{2}\right)\right]_{t=0} = \mu,$$

and

$$\sigma_X^2 = \mathrm{Var}(X) = E[X^2]-\mu^2 = M_X''(0)-\mu^2 = \left[(\sigma^2t+\mu)^2\exp\left(t\mu+\frac{\sigma^2t^2}{2}\right)+\sigma^2\exp\left(t\mu+\frac{\sigma^2t^2}{2}\right)\right]_{t=0}-\mu^2 = \sigma^2.$$

The analogue to $P(Z\in(a,b)) = \Phi(b)-\Phi(a)$ for a standard normal distribution is

$$P(X\in(a,b)) = P(Z\in((a-\mu)/\sigma,(b-\mu)/\sigma)) = \Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right).$$

2.4 Gamma Function & Gamma Distribution

The **gamma function** is defined by

$$\Gamma(\alpha) = \int_0^\infty e^{-y}y^{\alpha-1}\,\mathrm{d}y.$$

Note that Γ is defined for all $\alpha>0$ but not $\alpha=0$ (integral starts from 0 and we cannot divide by 0.)
The integrand $\to 0$ as $y\to\infty$; even better: this integral converges for <u>any</u> $\alpha>0$.

Properties of the Γ function.

(1) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. Indeed,

$$\int_0^\infty e^{-y}y^\alpha\,\mathrm{d}y = -e^{-y}y^\alpha\Big|_{y=0}^\infty + \int_0^\infty e^{-y}\alpha y^{\alpha-1}\,\mathrm{d}y = 0+\alpha\int_0^\infty e^{-y}y^{\alpha-1}\,\mathrm{d}y = \alpha\Gamma(\alpha),$$

where the first equality is from integration by parts: $u=y^\alpha, \mathrm{d}v = e^{-y}\mathrm{d}y$, $\mathrm{d}u = \alpha y^{\alpha-1}\mathrm{d}y$, and $v=-e^{-y}$.

(2) $\Gamma(1)=1$. Clear enough: $\Gamma(1) = \int_0^\infty e^{-y}\,\mathrm{d}y = -e^{-y}\Big|_{y=0}^\infty = -(-1) = 1$.

(3) $\Gamma(n+1) = n\Gamma(n) = \cdots = n!\,\Gamma(1) = n!$ because $\Gamma(1)=1$. This tells us that *the Γ function interpolates factorials.* This gives a much more convenient way to evaluate $\Gamma(x)$ for non-integer $x>0$: for example

$$\Gamma(2.7) = 2.7\cdot1.7\cdot0.7\cdot\Gamma(0.7).$$

Gamma Distribution

We say $X \sim \text{Gamma}(\alpha, \lambda)$ for $\alpha, \lambda > 0$ is a if X has pdf

$$f_X(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} \cdot \chi_{[0,\infty)}(x) \text{ or } f_X(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} \text{ for } x \geqslant 0.$$

It follows that the cdf is given by

$$F_X(x) = \frac{\lambda}{\Gamma(\alpha)} \int_0^\infty e^{-\lambda x}(\lambda x)^{\alpha-1}\, \mathrm{d}x \cdot \chi_{[0,\infty)}(x).$$

Special cases of gamma distribution:

(1) $\alpha = 1, \lambda > 0$: $X \sim \text{Gamma}(1, \lambda) \iff X \sim \text{Exp}(\lambda)$ (recall the exponential distribution).

(2) $\alpha = n/2, \lambda = 1/2$ gives the *Chi-squared distribution with n degree of freedom*: $X \sim \text{Gamma}(n/2, 1/2) \sim \chi^2_n$. If $n = 1$, i.e., degree of freedom is 1, then $X \sim \text{Gamma}(1/2, 1/2) \sim \chi^2$ is simply called the *Chi-squared distribution* (with 1 degree of freedom). Interestingly, if $X_i \sim \chi^2$ are independent and identically distributed (i.i.d.), then

$$X := \sum_{i=1}^n X_i \sim \chi^2_n.$$

Beginning of April 2, 2021

Sum of Normal r.v.'s

Example 2.4.1. If $X_i \sim N(\mu, \sigma^2)$ are i.i.d., then the estimations of **sample mean** and **sample variance** are given by $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. Then $\overline{X} \sim N(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

But how do we compute the mean and variance of a sum of normal r.v.? Easy:

$$E[\overline{X}] = E[\frac{1}{n}\sum_{i=1}^n X_i] = \frac{1}{n}\sum_{i=1}^n E[X_i] = \mu$$

and

$$\text{Var}(\overline{X}) = \text{Var}(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n^2}\text{Var}(\sum_{i=1}^n X_i) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}.$$

But why is this $\overline{X}$ a normal random variable? This is the first time we actually make use of the MGF. Recall that if $Y = aX$ then

$$M_Y(t) = E[E^{tY}] = E[e^{taX}] = E[e^{atX}] = M_X(at),$$

and if X, Y are independent,

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_y(t).$$

Also recall that if $Y \sim N(\mu, \sigma^2)$ then

$$M_Y(t) = \exp\left(t\mu + \frac{\sigma^2 t^2}{2}\right).$$

Therefore,

$$\begin{aligned} M_{\overline{X}}(t) &= E[\exp(t\overline{X})] = E[\exp(\frac{t}{n}\sum_{i=1}^{n} X_i)] \\ &= E[\prod_{i=1}^{n} \exp(\frac{t}{n}X_i)] = \prod_{i=1}^{n} E[\exp(tX_i/n)] \\ &= \prod_{i=1}^{n} \exp\left(\frac{t\mu}{n} + \frac{\sigma^2 t^2}{2n^2}\right) \\ &= \exp\left(t\mu + \frac{\sigma^2 t^2}{2n}\right) = \exp\left(t\mu + \frac{(\mu^2/n)t^2}{2}\right) \end{aligned}$$

and therefore $\overline{X} \sim N(\mu, \sigma^2/n)$.

Back to Gamma: Mean, Variance, and MGF

The MGF of a Gamma of a Gamma r.v. is

$$\begin{aligned} M_X(t) &= E[e^{tX}] = \int_0^\infty \frac{e^{tx}\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}\,\mathrm{d}x \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty e^{(t-\lambda)x}x^{\alpha-1}\,\mathrm{d}x \end{aligned}$$

This integral only makes sense if $t - \lambda < 0$, so assume $t < \lambda$. Let $u = (\lambda - t)x$ and so $\mathrm{d}u = (\lambda - t)\mathrm{d}x$. On the other hand, $x = u/(\lambda - t)$ and $\mathrm{d}x = \mathrm{d}u/(\lambda - t)$. The lower and upper limits of the integral stay the same. Then,

$$\begin{aligned} M_X(t) &= \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty e^{-u}\left(\frac{u}{\lambda - t}\right)^{\alpha-1}\cdot\frac{1}{\lambda - t}\,\mathrm{d}u \\ &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha}\frac{1}{\Gamma(\alpha)}\int_0^\infty e^{-u}u^{\alpha-1}\,\mathrm{d}u \\ &= \left(\frac{\lambda}{\lambda - t}\right)^\alpha \frac{\Gamma(\alpha)}{\Gamma(\alpha)} = \left(\frac{\lambda}{\lambda - t}\right)^\alpha. \end{aligned}$$

The mean is just $M_X'(0)$:

$$\mu_X = M_X'(0) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\lambda}{\lambda - t}\right)^\alpha\Bigg|_{t=0} = \alpha\left(\frac{\lambda}{\lambda - t}\right)^{\alpha-1}\frac{\lambda}{(\lambda - t)^2}\Bigg|_{t=0} = \frac{\alpha}{\lambda}.$$

The variance is

$$\begin{aligned} \sigma_X^2 &= \mathrm{Var}(X) = E[X^2] - \mu_X^2 = M_X''(0) - \left(\frac{\alpha}{\lambda}\right)^2 \\ &= [\cdots] = \frac{\alpha}{\lambda^2}. \end{aligned}$$

Example 2.4.2. Recall we previously stated that if $X_i \sim \chi^2$ are i.i.d. then

$$X := \sum_{i=1}^{n} X_i \sim \chi_n^2 \sim \mathrm{Gamma}(n/2, 1/2).$$

To see this, we compute the MGF:

$$\begin{aligned} M_X(t) &= E[\exp(t\sum_{i=1}^{n} X_i)] \\ &= E[\exp(tX_1)\exp(tX_2)\dots\exp(tX_n)] \\ &= E[e^{tX_1}]E[e^{tX_2}]\dots E[e^{tX_n}] \\ &= M_{X_1}(t)\dots M_{X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) \\ &= \prod_{i=1}^{n}\left(\frac{(1/2)}{1/2-t}\right)^{1/2} = \left(\frac{(1/2)}{1/2-t}\right)^{n/2} \end{aligned}$$

which agrees with the MGF of Gamma$(n/2, 1/2)$.

Example 2.4.3. Suppose $X_i \sim N(\mu,\sigma)^2$ are i.i.d. Recall that we have the sample mean $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ and the sample variance $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Claim: $\overline{X} \sim N(\mu, \sigma^2/n)$ which we have previously shown and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

To simplify the problem a little bit, we consider $\hat{S}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ and we show $\frac{n\hat{S}^2}{\sigma^2} \sim \chi^2_n$. (Notice that we are using the true mean μ, not the estimate mean $\overline{X}$.) Indeed,

$$\frac{n\hat{S}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2$$

where $Z_i \sim N(0,1)$ are i.i.d. standard normal.
Claim: $Z^2 \sim \chi^2$. Indeed, if we look at the cdf of Z^2,

$$\begin{aligned} F_{Z^2}(x) &= P(Z^2 \leqslant x) = P(-\sqrt{x} \leqslant Z \leqslant \sqrt{x}) \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\sqrt{x}}^{\sqrt{x}} e^{-t^2/2}\,\mathrm{d}t = \frac{2}{\sqrt{2\pi}}\int_0^{\sqrt{x}} e^{-t^2/2}\,\mathrm{d}t. \end{aligned}$$

On the other hand, (recall that pdf is the derivative of cdf)

$$\begin{aligned} f_{Z^2}(x) &= \frac{\mathrm{d}}{\mathrm{d}x}F_{Z^2}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\frac{2}{\sqrt{2\pi}}\int_0^{\sqrt{x}} e^{-t^2/2}\,\mathrm{d}t \\ &= \frac{2}{\sqrt{2\pi}}\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{\sqrt{x}} e^{-t^2/2}\,\mathrm{d}t \\ &= \frac{2}{\sqrt{2\pi}}e^{-(\sqrt{x})^2/2}\frac{\mathrm{d}}{\mathrm{d}x}\sqrt{x} \\ &= \frac{2}{\sqrt{2\pi}}e^{-x/2}\frac{1}{2\sqrt{x}} \\ &= \frac{1}{\sqrt{2\pi}}e^{-x/2}x^{1/2-1}. \end{aligned}$$

One more computation before we draw the connection:

$$\begin{aligned}\Gamma(1/2) &= \int_0^\infty e^{-y} y^{1/2-1}\,\mathrm{d}y = \int_0^\infty e^{-y} y^{-1/2}\,\mathrm{d}y \\ &= \int_0^\infty e^{-y}\frac{\sqrt{2}}{\sqrt{2y}}\,\mathrm{d}y = \sqrt{2}\int_0^\infty \frac{e^{-y}}{\sqrt{2y}}\,\mathrm{d}y \\ [u := \sqrt{2y}] &= \sqrt{2}\int_0^\infty e^{-u^2/2}\,\mathrm{d}u \\ &= \frac{\sqrt{2}}{2}\int_{-\infty}^\infty e^{-u^2/2}\,\mathrm{d}u = \sqrt{\pi}.\end{aligned}$$

Therefore (finally!),

$$f_{\chi^2}(x) = \frac{1}{2\sqrt{\pi}}e^{-x/2}\cdot\frac{x^{-1/2}}{2^{-1/2}} = \frac{1}{\sqrt{2\pi}}e^{-x/2}x^{-1/2} = f_{Z^2}(x).$$

Therefore $Z^2 \sim \chi^2$ and since each X_i is i.i.d.,

$$\frac{n\hat{S}^2}{\sigma^2} = \sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \chi^2 = \chi_n^2.$$

2.5 Beta Distribution

A useful distribution with a compact support (along with uniform distribution). We say $X \sim \mathrm{Beta}(\alpha,\beta)$ if

$$f_X(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\chi_{(0,1)}(x) \qquad \text{for } \alpha,\beta>0,$$

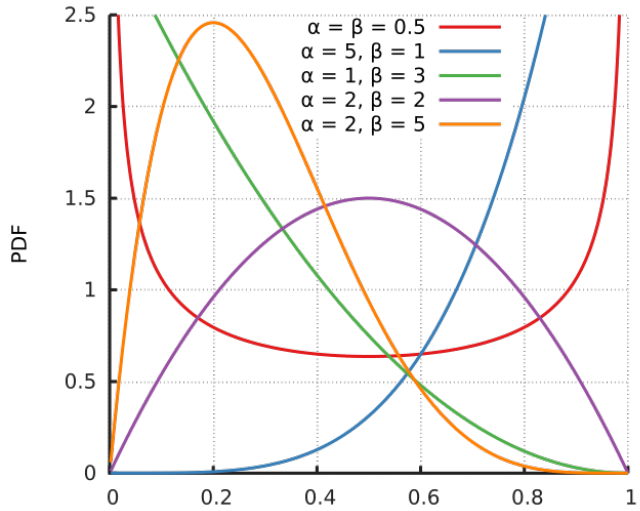
where

$$B(\alpha,\beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x.$$

The factor $1/B(\alpha,\beta)$ ensures that the integral of $f_X(x)$ is 1 so that X makes a random variable.

Beginning of April 5, 2021

Pdfs of some beta distribution from Wikipedia:



Claim. $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. To see this:

$$\begin{aligned}\Gamma(x)\Gamma(y) &= \int_0^\infty e^{-u}u^{x-1}\,\mathrm{d}u \int_0^\infty e^{-v}v^{y-1}\,\mathrm{d}v \\ &= \int_0^\infty \int_0^\infty e^{-u-v}u^{x-1}v^{y-1}\,\mathrm{d}u\,\mathrm{d}v\end{aligned}$$

Now we invoke change of variables $u := zt$ and $v := z(1-t)$ so that as $u, v \to \infty$, $t \to 1$ and $z \to \infty$. The Jacobian is

$$\mathcal{J}(z,t) = \det\begin{bmatrix}\partial u/\partial z & \partial u/\partial t \\ \partial v/\partial z & \partial v/\partial t\end{bmatrix} = \begin{vmatrix} t & z \\ 1-t & -z\end{vmatrix} = z.$$

Then,

$$\begin{aligned}\Gamma(x)\Gamma(y) &= \int_0^\infty \int_0^1 e^{-z}(zt)^{x-1}(z(1-t))^{y-1}|\mathcal{J}(z,t)|\,\mathrm{d}t\,\mathrm{d}z \\ &= \int_0^\infty \int_0^1 e^{-z}(zt)^{x-1}(z(1-t))^{y-1}z\,\mathrm{d}t\,\mathrm{d}z \\ &= \int_0^\infty \int_0^1 e^{-z}z^{x+y-1}t^{x-1}(1-t)^{y-1}\,\mathrm{d}t\,\mathrm{d}z \\ &= \int_0^\infty e^{-z}z^{x+y-1}\,\mathrm{d}z \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t \\ &= \Gamma(x+y)B(x,y).\end{aligned}$$

Since beta distribution is on $(0,1)$, it is often times used to model proportions, for example Bayesian estimation in a binomial distribution (more to come in 408).

There is no closed form expression of MGF. However, it is easy to calculate moment directly:

$$\begin{aligned}E[X^n] &= \int_0^1 x^n \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x \\ &= \frac{1}{B(\alpha,\beta)}\int_0^1 x^{n+\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}B(n+\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(n+\alpha)\Gamma(\beta)}{\Gamma(n+\alpha+\beta)} \\ &= \frac{\Gamma(\alpha+\beta)\Gamma(n+\alpha)}{\Gamma(\alpha)\Gamma(n+\alpha+\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(n+\alpha+\beta)}\cdot\frac{\Gamma(n+\alpha)}{\Gamma(\alpha)} \\ &= \frac{(n+\alpha-1)\cdots(\alpha+1)(\alpha)}{(n+\alpha+\beta-1)\cdots(\alpha+\beta+1)(\alpha+\beta)}.\end{aligned}$$

In particular, the mean of $\mathrm{Beta}(\alpha,\beta)$ is

$$\mu_X = E[X] = \frac{\alpha}{\alpha+\beta}$$

and the second moment is

$$E[X^2] = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}.$$

Therefore, the variance of $\mathrm{Beta}(\alpha,\beta)$ is

$$\sigma_X^2 = E[X^2] - \mu^2 = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.$$

And... a disgusting MGF:

$$M_X(t) = 1 + \sum_{k=1}^\infty \frac{t^k}{k!}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right).$$

2.6 Pareto, Weibull, Cauchy, t, and F Distributions

Pareto Distribution

We say $X \sim \text{Pareto}(\alpha, \beta)$ if the pdf of X is given by

$$f_X(x) = \frac{\beta\alpha^\beta}{x^{\beta+1}}\chi_{[\alpha,\infty)}(x) \qquad \alpha > 0, \beta > 0.$$

The mean is $\mu_X = E[X] = \dfrac{\beta^\alpha}{\beta - 1}$ for $\beta > 1$ (and if $\beta \leqslant 1$ there is no finite mean).

The variance is $\sigma_X^2 = \dfrac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}$ for $\beta > 2$ (and no variance for $\beta \leqslant 2$ similarly).

Weibull Distribution

We say $X \sim \text{Weibull}(\gamma, \beta)$ if the pdf of X is given by

$$f_X(x) = \frac{\gamma}{\beta}x^{\gamma-1}e^{-x^\gamma/\beta^\gamma}\chi_{[0,\infty)}(x).$$

Note that if $\gamma = 1$ this reduces to an exponential variable (with $\lambda = 1/\beta$).
The mean is $\mu_X = E[X]\Gamma(1 + 1/\gamma)/\beta^\gamma$; the variance is $\sigma_X^2 = \beta^{2/\gamma}\left[\Gamma(1 + 2/\gamma) - \Gamma^2(1 + 1/\gamma)\right]$; and the moments are $E[X^n] = \beta^{n/\gamma}\Gamma(1 + n/\gamma)$.

Cauchy Distribution

We say $X \sim \text{Cauchy}(\theta, \sigma)$ if the pdf is given by

$$f_X(x) = \frac{1}{\pi\sigma} \cdot \frac{1}{1 + \left[\frac{x-\theta}{\sigma}\right]^2} \qquad \text{for } x \in \mathbb{R}, \theta \in \mathbb{R}, \text{ and } \sigma > 0.$$

This is related to the quotient of two standard normals. Mean or variance does not exist (not finite).

Student's t Distribution

We say $X \sim t_\nu$ if

$$f_X(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\frac{1}{\sqrt{\nu\pi}}\frac{1}{(1 + (x^2/\nu))^{(\nu+1)/2}}, \qquad x \in \mathbb{R}, \nu \in \mathbb{Z}^+.$$

(ν is the degree of freedom.) The mean is 0 for $\nu > 1$ and the variance is $\nu/(\nu - 2)$ for $\nu > 2$. The moments is

$$E[X^n] = \begin{cases} \dfrac{\Gamma((n+1)/2)\Gamma((\nu-n)/2)}{\sqrt{\pi}\Gamma(\nu/2)}\nu^{n/2} & \text{if } n < \nu \text{ and even} \\ 0 & \text{if } n < \nu \text{ odd.} \end{cases}$$

In fact, this is related to the standard normal and χ-squared distributions. If $Z \sim N(0,1)$ and $U \sim \chi_n^2$, then

$$Z/\sqrt{U/n} \sim t_n.$$

In statistics, recall if $X_i \sim N(\mu, \sigma^2)$ are i.i.d. then we use $\overline{X} := \dfrac{1}{n}\sum_{i=1}^n X_i$ to estimate the sample mean. As a test hypothesis, suppose the true mean $\mu = \mu_0$. We know if $E[X_i] = \mu_0$ then

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim Z \sim N(0,1).$$

Also recall that we estimate the sample variance by $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Then

$$\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \approx \frac{\overline{X}-\mu_0}{S/\sqrt{n}} = \frac{(\overline{X}-\mu_0)/(\sigma/\sqrt{n})}{(S/\sqrt{n})/(\sigma/\sqrt{n})} - \frac{Z}{S/\sigma} = \frac{Z}{\sqrt{(n-1)S^2/(\sigma^2/(n-1))}} = \frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}} \sim t_{n-1}.$$

F Distribution

We say $X \sim F_{m,n}$ if

$$f_X(x) = \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{m/2} x^{m/2-1}\left(1+\frac{m}{n}x\right)^{-(m+n)/2} \text{ for } x \in (0,\infty).$$

If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ then $X := (U/m)/(V/n) \sim F_{m,n}$.

In statistics, this is used to test σ^2. Suppose we have two populations and we want to see if they have the same variance: let

$$X_i \sim N(\mu, \sigma_1^2), i = 1, 2, \ldots, n \text{ and } X_i \sim N(\mu, \sigma_2^2), i = 1, 2, \ldots, n_2$$

be i.i.d. Hypothesis: $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Let

$$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2 \text{ and } S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(X_i - \overline{X})^2$$

then $\sigma_1^2/\sigma^2 \approx S_1^2/S_2^2 \sim \frac{\chi^2_{n_1}/(n_1-1)}{\chi^2_{n_2}/(n_2-1)} \sim F_{(n_1-1),(n_2-1)}$.

Chapter 3

More on CRVs

3.1 Distribution of a Function of a R.V.

Example 3.1.1. Let $X \sim U(0,1)$ and suppose $Y = X^n$. Then intuitively we have $F_Y(y) = P(Y \leqslant y) = P(X \leqslant y^{1/n}) = F_X(y^{1/n}) = y^{1/n}$. We get the cdf of Y:

$$F_Y(y) = \begin{cases} 0 & y \leqslant 0 \\ y^{1/n} & 0 < y \leqslant 1 \\ 1 & 1 < y \end{cases}$$

and differentiating the cdf w.r.t. y gives

$$f_y(y) = \begin{cases} 0 & y \leqslant 0 \\ \dfrac{y^{1/n-1}}{n} & 0 < y \leqslant 1 \\ 0 & 1 < y. \end{cases}$$

Example 3.1.2. We now generalize the example above using inverse function (since the inverse of x^n is $y^{1/n}$). Suppose $X \sim f_X$ (pdf). Let $Y = X^2$. Then $F_Y(y) = 0$ for $y < 0$ is clear. Now assume $y \geqslant 0$. We see

$$F_Y(y) = P(Y \leqslant y) = P(-\sqrt{y} \leqslant X \leqslant \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

(The above assumed X is a CRV.) Then taking the derivative w.r.t. y gives

$$f_X(y) = \begin{cases} 0 & -\infty < y < 0 \\ \dfrac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}} & y > 0. \end{cases}$$

Theorem 3.1.3

Suppose X is a CRV with pdf $X \sim f_X$. Let $g : \mathbb{R} \to \mathbb{R}$ be either strictly increasing or strictly decreasing. Also assume g is differentiable. (We want g^{-1} to exist.) Then if $Y = g(X)$, the pdf of Y is

$$f_Y(Y) = \begin{cases} 0 & y \neq g(x) \text{ for all } x \\ f_X(g^{-1}(y))\left|\dfrac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right| & y = g(x) \text{ for some } x. \end{cases}$$

where g^{-1} is defined to be such that $g^{-1} \circ g = \mathrm{id}_x$.

Beginning of April 9, 2021

Proof. If g is strictly increasing or decreasing then g^{-1} is well-defined. Notice that if g is increasing then

$$F_Y(y) = P(Y \leqslant y) = P(X \leqslant g^{-1}(y)) = F_X(g^{-1}(y)).$$

Therefore,

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_x(G^{-1}(y)) = f_X(g^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)$$

and $\dfrac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y) = \left|\dfrac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|$. One can show analogously that when g is decreasing the original equation still holds, and this is the case where $|\cdot|$ matters:

$$F_Y(y) = P(X \leqslant y) = P(g(X) \leqslant y) = P(X \geqslant g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

and

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

□

Example 3.1.4. If $X \sim N(\mu, \sigma^2)$ then $Y = e^X$ is said to be **log normal** with parameters μ, σ^2. Notice that $Y \in (0, \infty)$. For $y \geqslant 0$, the pdf is given by

$$f_Y(y) = f_X(\ln y)|1/y| = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right)\frac{1}{y}.$$

3.2 Jointly Continuous R.V.s

Recall that, when talking about discrete r.v.'s, we look at cases where we had multiple discrete r.v.'s associated with the same experiment and we investigated their joint distribution. When the same idea is applied to continuous r.v.'s, instead of a joint pmf, we now have a **joint pdf**.

(*In fact, we can have a mixture of continuous and discrete r.v.'s!*)

Definition 3.2.1

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be a **joint pdf** if

$$f(x) \geqslant 0 \text{ for all } x \text{ and } \int_{\mathbb{R}^n} f(x)\,\mathrm{d}x = 1$$

where $x := (x_1, \ldots, x_n)$ are vectors in $\mathbb{R}^n$.

Definition 3.2.2

We say $X_1, \ldots, X_n$ are **jointly continuous r.v.'s** if there exists $f : \mathbb{R}^n \to \mathbb{R}$ a jpdf such that, if $S \subset \mathbb{R}^n$,

$$P(x \in S) = \int_S f(x)\,\mathrm{d}x.$$

Notice that this is a generalization of the single variable version: $P(X \in A) = \int_A f(x)\,\mathrm{d}x$.

In 407 we will be most concerned with the case $n = 2$. The definition then says $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ is a jpdf if

$$f_{X,Y}(x,y) \geqslant 0 \text{ for all } (x,y) \in \mathbb{R}^2 \text{ and } \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1,$$

and X, Y are jointly continuous if for $S \subset \mathbb{R}^2$,

$$P((X,Y) \in S) = \iint_S f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Definition 3.2.3

It follows naturally that we are able to define the **joint cdf**

$$F_{X,Y}(x,y) = P(X \leqslant x, Y \leqslant y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(\tilde{x},\tilde{y})\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{y}.$$

Remark. Similar to the single-variable case, we may recover the joint pdf by differentiating the joint cdf:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y}\int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(\tilde{x},\tilde{y})\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{y}.$$

In 407 we assume that the differential operator is insensitive to order, i.e., $\dfrac{\partial^2}{\partial x \partial y}$ *and* $\dfrac{\partial^2}{\partial y \partial x}$ *are the same.*

Marginalization

Recall that we can "fix" all but one variables and marginalize the remaining one. Here we apply the same idea.

Definition 3.2.4

For the case of two variables, the ***X*-marginal, *Y*-marginal distributions** are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y \text{ and } f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x.$$

This easily generalizes to the case of n variables: to marginalize $\{X_{k_1}, \ldots, X_{k_m}\} \subset \{X_1, \ldots, X_n\}$,

$$f_{X_{k_1},\ldots,X_{k_m}}(x_{k_1}, \ldots, x_{k_m}) = \int_{\mathbb{R}^{n-m}} f_{X_{k_1},\ldots,X_{k_m}}(x_{k_1}, \ldots, x_{k_m}) \underbrace{\mathrm{d}x_{j_1} \ldots \mathrm{d}x_{j_{m-n}}}_{j_i \notin \{k_1,\ldots,k_m\}}.$$

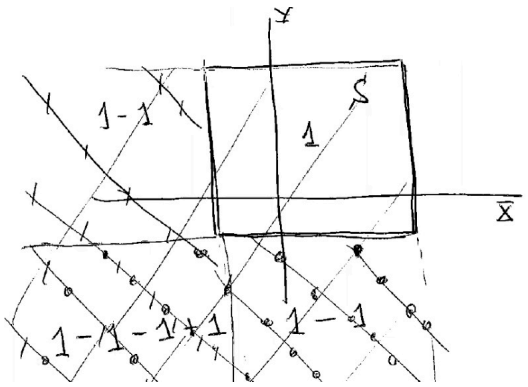
Imagine if we are trying to find $P([X, Y] \in [a, b] \times [c, d])$. By inclusion-exclusion

$$\begin{aligned}[a, b] \times [c, d] &= ((-\infty, b] - (-\infty, a]) \times (([-\infty, d] - (-\infty, c]) \\ &= (-\infty, b] \times (-\infty, d] - (-\infty, b] \times (-\infty, c] - (-\infty, a] \times (-\infty, d] + (-\infty, a] \times (-\infty, c].\end{aligned}$$

Thus

$$P((X, Y) \in [a, b] \times [c, d]) = F_{X,Y}(b, d) - F_{X,Y}(a, d) - F_{X,Y}(b, c) + F_{X,Y}(a, c).$$

An illustration from lecture where the box denotes $[a, b] \times [c, d]$:



Expectation

Suppose $g : \mathbb{R}^n \to \mathbb{R}$ and $X_1, \ldots, X_n$ are jointly continuous with jpdf f, then if $Y = g(X_1, \ldots, X_n)$ (a scalar random variable) we have

$$E[Y] = \int_{\mathbb{R}^n} g(x) f(x)\, \mathrm{d}x \text{ where } x \in \mathbb{R}^n.$$

Intuitively this is just the generalization of $E[X] = \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x$ *in the single variable case where* $g(x) := x$.

Example 3.2.5. Suppose we have two r.v.'s, $g(x, y) = x + y$ and $Z = X + Y$. For simplicity we let $f := f_{X,Y}$. Same thing in the future unless otherwise specified. Then (assuming we can interchange order of integration

in 407)

$$\begin{aligned}
E[Z] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\,\mathrm{d}x\,\mathrm{d}y + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\,\mathrm{d}y\,\mathrm{d}x + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} x\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x + \int_{-\infty}^{\infty} y\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} xf_X(x)\,\mathrm{d}x + \int_{-\infty}^{\infty} yf_Y(y)\,\mathrm{d}y = E[X]+E[Y].
\end{aligned}$$

In general, E is a linear operator, i.e.,

$$E[\sum_{i=1}^{n} c_iX_i] = \sum_{i=1}^{n} c_iE[X_i].$$

But what about variance of sum of jointly continuous r.v.'s? Once again this is analogous to the discrete case.

Covariance, Correlation, & Independence

Definition 3.2.6

The **covariance** of two jointly continuous r.v.'s X and Y is given by

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\
&= \int_{\mathbb{R}^2} (x-\mu_X)(y-\mu_Y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y.
\end{aligned}$$

Just like how in the discrete case $\mathrm{Cov}(X,Y) = E[XY]-\mu_X\mu_Y$ we also have it here:

$$\mathrm{Cov}(X,Y) = \int_{\mathbb{R}^2} xyf(x,y)\,\mathrm{d}x\,\mathrm{d}y - \mu_X\mu_Y.$$

Definition 3.2.7

The **correlation coefficient** of X,Y (again, this should look very familiar!) is given by

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

Once again, Cauchy-Schwarz inequality tells us $\rho(X,Y) \in [-1,1]$, and if $|\rho(X,Y)| = 1$ then $Y = aX+b$ (so they are linearly related and thus *very, very dependent*).

Proposition 3.2.8

Covariance is bilinear and commutative, and $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

Also, as in the discrete case, given $X_1,\dots,X_n$ jointly continuous, we can define the **covariance matrix**

$$\Sigma : \Sigma_{i,j} \coloneqq \mathrm{Cov}(X_i,X_j)$$

and it enjoys all the properties that a discrete r.v. covariance matrix has (symmetric and PSD). In particular, the covariance of a linear combination of X_i's is (just like before)

$$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i) = a^T \Sigma a$$

and thus

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

This naturally leads to the hypothesis that if X, Y are independent then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \implies \mathrm{Cov}(X,Y) = 0$. Indeed, this is true:

Definition 3.2.9

Jointly continuous r.v.'s X, Y are said to be **independent** if

$$P(X \in A, y \in B) = P(X \in A)P(Y \in B).$$

If so, we immediately see $E[XY] = E[X]E[Y]$ and thus $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = 0$. Also recall that $E[XY] = E[X]E[Y]$ is a sufficient but not necessary condition for the independence of X and Y.

Beginning of April 12, 2021

Why X, Y independent $\implies E[XY] = \mu_X \mu_Y$ and $\mathrm{Cov}(X,Y) = 0$?

Proof. On one hand

$$P(X \in A, Y \in B) = \int_{A\times B} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

and on the other hand

$$P(X \in A)P(Y \in B) = \int_A f_X(x)\,\mathrm{d}x \int_B f_Y(y)\,\mathrm{d}y.$$

If X and Y are independent then these two are the same, so (assuming Fubini...)

$$\begin{aligned} P(X \in A, Y \in B) - P(X \in A)P(Y \in B) &= \int_{A\times B} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y - \int_A f_X(x)\,\mathrm{d}x \int_B f_Y(y)\,\mathrm{d}y \\ &= \int_{A\times B} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y - \int_{A\times B} f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{A\times B} f_{X,Y}(x,y) - f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y = 0. \end{aligned}$$

Note that A and B can be chosen arbitrarily, and the above can hold for all A, B only if $f_{X,Y}(x,y) - f_X(x)f_Y(y)$ is uniformly 0.

Then,

$$\begin{aligned} E[XY] &= \int_{\mathbb{R}^2} xy f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathbb{R}^2} xy f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y = E[X]E[Y] \end{aligned}$$

and so $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = 0$. □

Remark. In fact, one can also show the other direction (again, exchanging the integrals, etc.) and obtain the following:

$$X, Y \text{ independent} \iff f_{X,Y}(x,y) = f_X(x) f_Y(y) \text{ for all } x, y.$$

Alternate proof showing X, Y independent $\implies f_{X,Y}(x,y) = f_X(x) f_Y(y)$.
Now we present a slick way to show what has been shown above: if X, Y are independent, then

$$F_{X,Y}(x,y) = P(X \leqslant x, Y \leqslant y) = P(X \leqslant x) P(Y \leqslant y) = F_X(x) F_Y(y).$$

Then

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} \left[F_X(x) F_Y(y) \right] \\ &= \frac{\partial}{\partial x} \left[\frac{\partial}{\partial y} \left(F_X(x) F_Y(y) \right) \right] = \frac{\partial}{\partial x} \left[F_X(x) \frac{\partial}{\partial y} F_Y(y) \right] \\ &= \frac{\partial}{\partial x} \left[F_X(x) f_Y(Y) \right] = f_X(x) f_Y(y). \end{aligned}$$

□

Corollary 3.2.10

Suppose X, Y are jointly continuous. They are independent if and only if

$$f_{X,Y}(x,y) = g(x) h(y)$$

for some functions g, h and <u>all</u> $(x,y) \in \mathbb{R}^2$, i.e., if and only if $f_{X,Y}(x,y)$ can be factored into the product of a function only of x and another one purely of y.

Example 3.2.11. We now present a counterexample to the above corollary: if $f_{X,Y}(x,y) = g(x)h(y)$ does <u>not</u> hold for all $(x,y) \in \mathbb{R}^2$ then X, Y may fail to be independent. Consider (for some $c \in \mathbb{R}$ which we'll determine later)

$$f_{X,Y}(x,y) = \begin{cases} 0 & 0 \leqslant y \leqslant x \leqslant 1 \\ cxy & \text{otherwise.} \end{cases}$$

Notice that $0 \leqslant x < y \leqslant 1$ describes the triangle bounded by $(0,0), (1,0)$, and $(1,1)$. Clearly $f_{X,Y}$ is always nonnegative; to make it a joint pdf, we want the double integral to be 1:

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, \mathrm{d}x \, \mathrm{d}y = \int_0^1 \int_0^x cxy \, \mathrm{d}y \, \mathrm{d}x \\ &= c \int_0^1 x \int_0^x y \, \mathrm{d}y \, \mathrm{d}x = c \int_0^1 x \cdot x^2/2 \, \mathrm{d}x \\ &= \frac{cx^4}{8} \Big|_{x=0}^1 \implies c = 8. \end{aligned}$$

Now we find the X-marginal: clearly $f_X(x) = 0$ for $x \notin (0,1)$. For $x \in (0,1)$ we have

$$f_X(x) = \int_0^x 8xy \, \mathrm{d}y = 4x^3 \implies f_X(x) = 4x^3 \chi_{[0,1]}.$$

Likewise for the Y-marginal: it vanishes everywhere but on $[0,1]$:

$$f_Y(Y) = \int_y^1 8xy \, \mathrm{d}x = 4y(1-y^2)\chi_{[0,1]}.$$

It turns out

$$f_X(x)f_Y(y) \neq f_{X,Y}(x,y).$$

Recall we've shown in the previous proof that if X, Y are independent then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Taking the contrapositive here tells us that X and Y are not independent in this example. The problem? Domain! *Alternatively, one can compute the means and variance and show that the covariance is nonzero, but that takes a ridiculous amount of computation so...*

Beginning of April 14, 2021

3.3 Conditional Distributions

Recall that given a probability space $\{\Omega, \Sigma, P\}$ and two events A, B with $P(B) > 0$, we can define a new probability measure $P(A \mid B)$ conditioned on B by

$$P(A \mid B) = \frac{P(AB)}{P(B)}.$$

Note that $\{\Omega, \Sigma, P_B\}$ is also a probability space (where $P_B(A) \coloneqq P(A \mid B)$).

It follows that we can define a probability distributions for r.v.'s conditioned on what another r.v. does.

Discrete Case

Given $\{\Omega, \Sigma, P\}$, $X : \Omega \to \mathbb{R}$, p_X the corresponding pmf, $Y : \Omega \to \mathbb{R}$, and p_Y the corresponding pmf, define

$$p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P_{X,Y}(x,y)}{P_Y(y)}$$

for $x \in \mathbb{R}$ and y such that $P(Y = y) > 0$. If we keep y fixed then we get a pmf of x (conditioned on $Y = y$). We can define a cdf by

$$F_{X|Y}(x \mid y) = P(X \leqslant x \mid Y \leqslant y) = \sum_{z \leqslant x} p_{X|Y}(z \mid y).$$

Notice that, if X, Y are independent, then $P(X = x, Y = y) = P(X = x)P(Y = y)$ and so

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

Example 3.3.1. Let $X \sim \Pr(\lambda_1)$ (Poisson) and $Y \sim \Pr(\lambda_2)$ are independent. What is $p_{X|X+Y}(k \mid n)$? (Of

course we won't be focusing on X conditioned on Y because independence makes this boring.)

$$\begin{aligned} p_{X|X+Y}(k \mid n) &= P(X = k \mid X + Y = n) \\ &= \frac{P(X = k, X + Y = n)}{P(X + Y = n)} = \frac{P(X = k, Y = n - k)}{P(X + Y = n)} \\ [\text{independence}] &= \frac{P(X = k)P(Y = n - k)}{P(X + Y = n)} \\ &\overset{*}{=} \frac{(e^{-\lambda_1}\lambda_1^k/k!)(e^{-\lambda_2}\lambda_2^{n-k}/(n-k)!)}{e^{-\lambda-\lambda_2}(\lambda_1+\lambda_2)^n/n!} \\ &= \frac{n!}{k!(n-k)!}\frac{\lambda_1^k\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} \\ &= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k} \end{aligned}$$

where the denominator after the $\overset{*}{=}$ can be computed using the MGF of Poisson distributions. (If X and Y are independent then $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$, i.e., MGF of sum is product of MGF. One can use this to verify that $X + Y$ in this example is indeed $\Pr(\lambda_1 + \lambda_2)$.) But then the final result rings a bell, does it not?

$$X \mid (X + Y = n) \sim \mathrm{B}(n, \lambda_1/(\lambda_1 + \lambda_2)).$$

Continuous Case

Definition 3.3.2

Given a joint pdf $f_{X,Y}$, we define the **conditional pdf** by

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

for $x \in \mathbb{R}$ and y with $f_Y(y) > 0$.

Where does this definition come from? Heuristically, for *very small* (but positive) $\Delta x, \Delta y \ll 1$ we have

$$\begin{aligned} f_{X|Y}(x \mid y)\Delta x &= \frac{f_{X,Y}(x, y)\Delta x}{f_Y(y)}\frac{\Delta y}{\Delta y} = \frac{f_{X,Y}(x, y)\Delta x\Delta y}{f_Y(y)\Delta y} \\ &\approx \frac{P(x \leqslant X \leqslant x + \Delta x, y \leqslant Y \leqslant y + \Delta y)}{P(y \leqslant Y \leqslant y + \Delta y)} \\ &= P(x \leqslant X \leqslant x + \Delta x \mid y \leqslant Y \leqslant y + \Delta y) \end{aligned}$$

which roughly describes *the probability that X is between x and $x + \Delta x$ given Y is between y and $y + \Delta y$.*

Like the discrete case, $f_{X|Y}(\cdot \mid y)$ is a pdf where the variable is x conditioned on $Y = y$. Indeed this is always

nonnegative, and, given some y,

$$\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)}{f_Y(y)}\,\mathrm{d}x$$
$$= \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x = \frac{f_Y(y)}{f_Y(y)} = 1.$$

Having defined the conditional pdf (the density function), we can now calculate the probability of X being in some $A \subset \mathbb{R}$ given $Y = y$:

$$P(X \in A \mid Y = y) = \int_A f_{X|Y}(x \mid y)\,\mathrm{d}x = \frac{1}{f_Y(y)} \int_A f_{X,Y}(x,y)\,\mathrm{d}x.$$

... and, not surprisingly, we can also compute the **conditional cdf**:

$$F_{X|Y}(x \mid y) = P(X \leqslant x \mid Y = y) = \int_{-\infty}^{x} f_{X|Y}(t \mid y)\,\mathrm{d}t = \frac{1}{f_Y(y)} \int_{-\infty}^{x} f_{X,Y}(t,y)\,\mathrm{d}t.$$

Remark. Note that even though $P(Y = y) = 0$, the definitions above are well-defined because the density $f_Y(y)$ is nonzero and we are conditioning on this.

Example 3.3.3. Consider X, Y jointly continuous with joint pdf

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{e^{-x/y}e^{-y}}{y} & 0 < x, y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

We claim that $f_{X,Y}$ is a joint pdf (and will not verify it). Now we compute $f_{X|Y}(x \mid y)$ and $P(X > 1 \mid Y = y)$. By definition,

$$f_{X|Y} = \begin{cases} \dfrac{f_{X,Y}(x,y)}{f_Y(y)} & 0 < x, y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

We first need to find the Y-marginal $f_Y(y)$ for $y > 0$:

$$f_Y(y) = \int_0^{\infty} \frac{e^{-x/y}e^{-y}}{y}\,\mathrm{d}x = \frac{e^{-y}}{y} \int_0^{\infty} e^{-x/y}\,\mathrm{d}x$$
$$= \frac{e^{-y}}{y}\left[-ye^{-x/y}\right]_{x=0}^{\infty} = e^{-y}.$$

Therefore the conditional pdf actually becomes simpler:

$$f_{X|Y}(x \mid y) = \begin{cases} e^{-x/y}/y & 0 < x, y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

For $P(X > 1 \mid Y = y)$, we simply need to plug in the definition:

$$P(X > 1 \mid Y = y) = \int_1^{\infty} f_{X|Y}(x \mid y)\,\mathrm{d}x = \int_1^{\infty} e^{-x/y}/y\,\mathrm{d}x = e^{-1/y}.$$

Example 3.3.4. In this example we look at the density of a t-distribution with n degrees of freedom. Recall that if $X_i \sim N(\mu, \sigma^2)$ i.i.d. for $i = 1, 2, \ldots, n$ then

$$\frac{(n-1)S^2}{\sigma^2} \sim t_{n-1} \text{ where } S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ and } \overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Also recall that $t_{n-1} \sim Z/\sqrt{\chi^2_{n-1}/(n-1)}$ (where Z is the standard normal $N(0,1)$).
Now we define $t_n \sim Z/\sqrt{Y/n} = \sqrt{n}Z/\sqrt{Y}$. What is $f_{t_n}(t)$?

Since Z, Y are independent,

$$\begin{aligned} F_{t_n|Y}(t \mid y) &= P(t_n \leqslant t \mid Y = y) \\ &= P(\sqrt{n}Z/\sqrt{Y} \leqslant t \mid Y = y) \\ \text{[independence]} &= P(\sqrt{n}Z/\sqrt{y} \leqslant t) = P(Z \leqslant t\sqrt{y/n}). \end{aligned}$$

Therefore,

$$\begin{aligned} f_{t_n|Y}(t \mid Y) &= \frac{\partial}{\partial t}F_Z(t\sqrt{y/n}) = \frac{\partial}{\partial t}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t\sqrt{y/n}} e^{-x^2/2}\,\mathrm{d}x \\ \text{[Leibniz]} &= \frac{1}{\sqrt{2\pi}}e^{-t^2y/(2n)}\sqrt{y/n} = \sqrt{\frac{y}{2\pi n}}\exp\left(-\frac{t^2y}{2n}\right). \end{aligned}$$

Recall that $Y \sim \chi^2_n$ identifies with $\text{Gamma}(n/2, 1/2)$ and

$$f_Y(y) = \frac{e^{-y/2}y^{n/2-1}}{2^{n/2}\Gamma(n/2)} \qquad \text{for } y > 0.$$

Then,

$$f_{t_n,Y}(t,y) = f_{t_n|Y}(t \mid y)f_Y(y) = \frac{y^{(n-1)/2}}{2^{(n+1)/2}\Gamma(n/2)\sqrt{\pi n}}\exp\left(-\frac{(t^2+n)y}{2n}\right) \quad \text{for } t \in (-\infty, \infty), y > 0.$$

Finally, to find $f_{t_n}(t)$, we need to find the t_n-marginal of the above (by integrating!). Let $C := \dfrac{t^2+n}{2n}$. Then

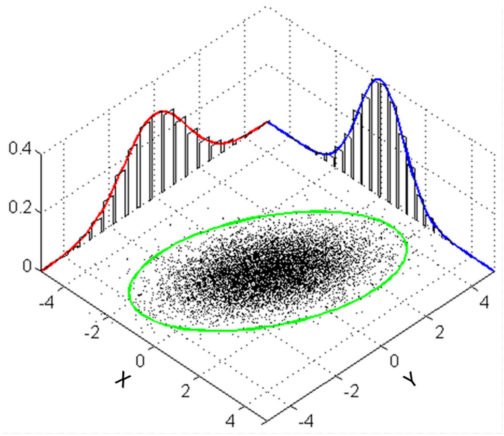
$$\begin{aligned} f_{t_n}(t) &= \int_0^\infty f_{t_n,Y}(t,y)\,\mathrm{d}y \\ &= \frac{1}{2^{(n+1)/2}\Gamma(n/2)\sqrt{\pi n}}\int_0^\infty e^{-cy}y^{(n-1)/2}\,\mathrm{d}y \\ [x := cy \quad \mathrm{d}x = c\mathrm{d}y] &= \frac{C^{-(n+1)/2}}{2^{(n+1)/2}\Gamma(n/2)\sqrt{\pi n}}\int_0^\infty e^{-x}x^{(n-1)/2}\,\mathrm{d}x \\ &= \frac{n^{(n+1)/2}\Gamma((n+1)/2)}{(t^2+n)^{(n+1)/2}\Gamma(n/2)\sqrt{\pi n}} = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\Gamma(n/2)}(1+t^2/n)^{-(n+1)/2} \qquad -\infty < t < \infty. \end{aligned}$$

In fact, as $n \to \infty$, the mess above converges to $\dfrac{1}{\sqrt{2\pi}}e^{-t^2/2}$ the standard normal.

Definition 3.3.5

We say X, Y have **Bivariate Normal Distribution** if, given $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ (both standard deviations > 0), $-1 < \rho := \rho_{X,Y} < 1$ and if their joint pdf is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho)^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right).$$



Beginning of April 16, 2021

If X, Y are uncorrelated then the last term containing 2ρ vanishes and $\sqrt{1-\rho^2}$ becomes just 1. Then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ and hence X, Y are independent. (This is a property unique to normals.)

We can actually find the conditional distributions of a bivariate normal distribution using

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

Given $Y = y$, we have

$$X \sim N(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(x-\mu_X), \sigma_X^2(1-\rho^2))$$

and likewise if we are given $X = x$.

If X, Y are correlated then X conditioned on Y is exactly X and vice versa, so indeed this is another way to see that X, Y are independent in this case. This result generalizes to multi-variate normals (not just bivariate normal).

Continuous r.v. Conditioned on DRV

Suppose X is a CRV with pdf $X \sim f_X$. Let N be a DRV with $N \sim P_N$ its pmf. Heuristically, for $\Delta x \ll 1$,

$$\begin{aligned} f_{X|N}(x \mid n) &\approx \frac{P(x < X < x + \Delta(x) \mid N = n)}{\Delta x} \\ &= \frac{P(x < X < x + \Delta x, N = n)}{P(N = n)\Delta x} \\ &= \frac{P(N = n \mid x < X < x + \Delta x)P(x < X < x + \delta X)}{P(N = n)} \\ [\Delta x \to 0] &\to \frac{P(N = n \mid X = x)}{P(N = n)} f_X(x). \end{aligned}$$

(This is a pdf, not pmf, since f is a CRV.)

Example 3.3.6. Suppose $N \sim B(n+m, P)$ where $P \sim U(0,1)$. Now we compute the distribution of P given $N = n$:

$$f_{P|N}(p \mid n) = \frac{P(N = n \mid P = p)}{P(N = n)} f_P(p)$$
$$= \frac{\binom{n+m}{n} p^n (1-p)^m}{P(N = n)} \chi_{[0,1]}(p) = \begin{cases} Cp^n(1-p)^m & 0 \leqslant p \leqslant 1 \\ 0 & \text{otherwise.} \end{cases}$$

Here C is just the constant determined by $\binom{n+m}{n}/P(N = n)$. Since $f_{P|N}$ is a conditional pdf, the term $p^n(1-p)^m$ cries out for a Beta distribution! Thus $P \mid N = n \sim \text{Beta}(n+1, m+1)$. The integral must evaluate to 1, so we must have

$$C = \frac{1}{B(n+1, m+1)} = \frac{\Gamma(n+m+2)}{\Gamma(n+1)\Gamma(m+1)}.$$

(Recall this from section 2.5 where we discussed the characterization of the normalizing function $B(\cdot, \cdot)$ for Beta distribution.) Therefore,

$$f_{P|N}(p \mid n) = \frac{\Gamma(n+m+2)}{\Gamma(n+1)\Gamma(m+1)} p^n (1-p)^m \chi_{[0,1]}(p)$$

(and also)

$$P(N = n) = \frac{\binom{n+m}{n}}{C} = \frac{\Gamma(n+1)\Gamma(m+1)\binom{n+m}{n}}{\Gamma(n+m+2)}.$$

3.4 Conditional Expectation & Variance

Conditional Expectation

Let X, Y be random variables. We now compute the expectation of X given $Y = y$. Intuitively, this is given by

$$E[X \mid Y = y] = \begin{cases} \sum_x x P_{X|Y}(x \mid y) & (X \text{ discrete}) \\ \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\, dx & (X \text{ continuous}). \end{cases}$$

(Of course, since the conditional pmf/pdf involves division by $P_Y(y)$ and $f_Y(y)$, we have to assume in the first place that these are positive.)

Example 3.4.1. Suppose $X, Y \sim B(n, p)$ independent. We compute $E[X \mid X + Y = M]$. To do so, we need $P_{X|X+Y}$. For $k \leqslant \min(n, m)$ (why so? X cannot be bigger than m if $X + Y = m$, and X cannot be bigger

than n because that's the total number of trials conducted)

$$\begin{aligned} P(X \mid X+Y)(k \mid m) &= \frac{P(X=h, X+Y=m)}{P(X+Y)=M} = \frac{P(X=k, Y=m-k)}{P(X+Y=m)} \\ &= \frac{P(X=k)P(Y=m-k)}{P(X+Y=m)} \\ &= \frac{\binom{n}{k}p^k(1-p)^{n-k}\binom{n}{m-k}p^{m-k}(1-p)^{n-(m-k)}}{\binom{2n}{m}p^m(1-p)^{2n-m}} \\ &= \frac{\binom{n}{k}\binom{n}{m-k}}{\binom{2n}{m}}. \end{aligned}$$

Notice that $k+(m-k)=m$ and $n+n=2n$, so this corresponds to a hypergeometric r.v.! More formally,

$$X \mid X+Y=m \sim H(2n, n, m),$$

i.e., a total of $2n$ elements in which n are distinguished and we are asked to create a sample of size m. (Note that this is a DRV.) Now,

$$E[X \mid X+Y=m] = \frac{mn}{2n} = \frac{m}{2}.$$

Remark. We can think of $E[X \mid Y]$ as a function of Y, a r.v. See the remark below for a direct application.

Proposition 3.4.2

$E[E[X \mid Y]] = E[X]$. This provides an alternate way to compute $E[X]$.

Proof. We will prove the continuous case; the discrete case is similar:

$$\begin{aligned} E[E[X \mid Y]] &= \int_{-\infty}^{\infty} f_Y(y)E[X \mid Y=y]\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x\right] f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = E[X]. \end{aligned}$$

□

Example 3.4.3. Consider the expectation of a sum of a random number (N) of random, i.i.d. variables X_i's with means μ. Conditioning the sum on the number N, we have

$$E[\sum_{i=1}^{N} X_i] = E[E[\sum_{i=1}^{N} X_i \mid N]].$$

If we further assume that X_i's and N are independent, then

$$E[\sum_{i=1}^{N} X_i \mid N = n] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = n\mu.$$

Then,

$$E[\sum_{i=1}^{N} X_i \mid N] = NE[X]$$

and

$$E[\sum_{i=1}^{N} X_i] = E[E[\sum_{i=1}^{N} X_i \mid N]] = E[NE[X]] = E[X]E[N]. \qquad \text{(by independence)}$$

Conditional Variance

Having discussed conditional mean, we now consider conditional variance:

$$\mathrm{Var}(X \mid Y) = E[(X - E[X \mid Y])^2 \mid Y].$$

In fact,

$$\mathrm{Var}(X \mid Y) = E[X^2 \mid Y] - E[X \mid Y]^2,$$

highly analogous to $\mathrm{Var}(X) = E[X^2] - E[X]^2$. This tells us that $\mathrm{Var}[X \mid Y]$ itself is a r.v.
The mean of $\mathrm{Var}(X \mid Y)$ is given by

$$\begin{aligned} E[\mathrm{Var}(X \mid Y)] &= E[E[X^2 \mid Y] - E[X \mid Y]^2] \\ &= E[E[X^2 \mid Y]] - E[E[X \mid Y]^2] \\ &= E[X^2] - E[E[X \mid Y]^2]. \end{aligned} \tag{1}$$

On the other hand, by definition of variance (treating $E[X \mid Y]$ as a r.v.)

$$\mathrm{Var}(E[X \mid Y]) = E[E[X \mid Y]^2] - E[E[X \mid Y]]^2 = E[E[X \mid Y]^2] - E[X]^2.$$

Therefore,

$$E[E[X \mid Y]^2] = \mathrm{Var}(E[X \mid Y]) + E[X]^2. \tag{2}$$

Combining (1) and (2) gives

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

so

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = E[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(E[X \mid Y]). \tag{Δ}$$

Proposition 3.4.4

(Recall from before) $E[X] = E[E[X \mid Y]]$ and $\mathrm{Var}(X) = E[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(E[X \mid Y])$.

Example 3.4.5. What is $\operatorname{Var}(\sum_{i=1}^{N} X_i)$? (Same assumptions from the previous example.)

Fist, from the previous example $E[\sum_{i=1}^{N} X_i \mid N = n] = nE[X]$, so

$$\operatorname{Var}(E[\sum_{i=1}^{N} X_i \mid N]) = \operatorname{Var}(NE[X]) = E[X]^2 \operatorname{Var}(N).$$

With the assumption that X_i's are i.i.d. with $\operatorname{Var}(X_i) = \operatorname{Var}(X)$ (treated as constant),

$$E[\operatorname{Var}(\sum_{i=1}^{N} X_i \mid N)] = E[N \operatorname{Var}(X)] = \operatorname{Var}(X)E[N].$$

Therefore (finally!),

$$\begin{aligned}\operatorname{Var}(\sum_{i=1}^{N} X_i) &= E[\operatorname{Var}(\sum_{i=1}^{N} X_i \mid N)] + \operatorname{Var}(E[\sum_{i=1}^{N} X_i \mid N]) \\ &= \operatorname{Var}(X)E[N] + E[X]^2 \operatorname{Var}(N).\end{aligned}$$

Beginning of April 19, 2021

3.5 Convolution: Distribution of Sum of CRVs

suppose $X, Y \sim f(x, y)$ their joint pdf. Let $Z = X + Y$. What is $f_Z(z)$? Intuitively,

$$F_Z(z) = P(Z \leqslant z) = P(X + Y \leqslant Z) = \int_{X+Y \leqslant Z} f(x, y) \, \mathrm{d}y \, \mathrm{d}x.$$

Writing this as a double integral,

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x, y) \, \mathrm{d}y \, \mathrm{d}x.$$

To get the pdf, (taking interchange of limit and integral for granted again)

$$\begin{aligned}f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z} F_Z(z) &= \frac{\mathrm{d}}{\mathrm{d}z} \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x, y) \, \mathrm{d}y \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}z} \int_{-\infty}^{z-x} f(x, y) \, \mathrm{d}y \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} f(x, z - x) \frac{\mathrm{d}}{\mathrm{d}z}(z - x) \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} f(x, z - x) \, \mathrm{d}x.\end{aligned}$$

If X, Y are independent, i.e., $f(x, y) = f_X(x) f_Y(y)$ we can say a little more:

$$f(x, y) = f_X(x) f_Y(y) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \, \mathrm{d}x =: (f_X * f_Y)(z),$$

the **convolution product** of f_X and f_Y. Convolution has a lot of nice algebraic properties: commutativity, distributivity, and associativity, for example.

Heuristically, for commutativity, since $X + Y$ and $Y + X$ define the same random variable, we must have

$$f_{X+Y}(z) = f_{Y+X}(z) \implies (f_X * f_Y)(z) = (f_Y * f_X)(z).$$

Alternatively, one can use u-substitution $u := z - x$ in $\int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x$ and obtain

$$\begin{aligned}(f_X * f_Y)(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x = -\int_{\infty}^{-\infty} f_X(z-u) f_Y(u)\,\mathrm{d}u \\ &= \int_{-\infty}^{\infty} f_X(z-u) f_Y(u)\,\mathrm{d}u = (f_Y * f_X)(z).\end{aligned}$$

But what about more random variables? In particular, what would happen if we add n i.i.d. X_i's up? It turns out we can convolve recursively with the common pdf: let f_k be the density of $S_k := \sum_{i=1}^{k} X_i$. Obviously $S_1 = X_1$ and $f_1(s) = f(s)$. Since $S_2 = X_2 + S_1$ (and they are independent)

$$f_2(s) = (f * f_1)(s)$$

and inductively (we have to traverse through the list one by one...)

$$S_n = X_n + S_{n-1} = f_n(s) = (f * f_{n-1})(s) = \int_{-\infty}^{\infty} f(x) f_{n-1}(s-x)\,\mathrm{d}x.$$

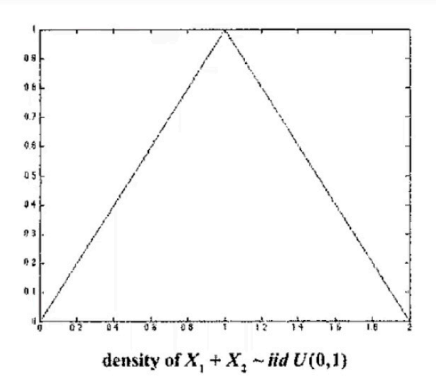
Example 3.5.1. Consider $X \sim U(0,1)$ and so $f_X = \chi_{[0,1]}(x)$. We initialize $f_1 = f$. Then

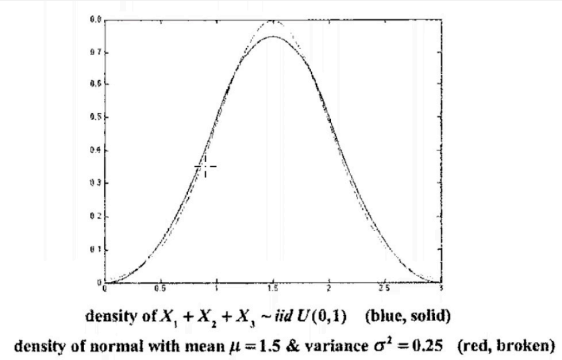
$$\begin{aligned}f_2(s) &= \int_{-\infty}^{\infty} f_1(x) f(s-x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \chi_{[0,1]}(x)\chi_{[0,1]}(s-x)\,\mathrm{d}x \\ &= \int_0^1 \chi_{[0,1]}(s-x)\,\mathrm{d}x = \int_0^1 \chi_{s-1,s}(x)\,\mathrm{d}x = \begin{cases} 0 & s \leqslant 0 \\ s & 0 < s \leqslant 1 \\ 2-s & 1 < s \leqslant 2 \\ 0 & s > 2. \end{cases}\end{aligned}$$

It follows that f_2 has a "triangle density". But what about 3?

$$f_3(s) = \int_{-\infty}^{\infty} \chi_{[0,1]}(x) f_2(s-x)\,\mathrm{d}x = \ldots = \begin{cases} 0 & s \leqslant 0 \\ s^2/2 & 0 < s \leqslant 1 \\ 1 - \dfrac{2-s^2}{2} - \dfrac{1-s^2}{2} & 1 < s \leqslant 2 \\ 1 - (3-s^2)/2 & 2 < s \leqslant 3 \\ 0 & s > 3. \end{cases}$$

This becomes a piecewise quadratic function. See the plots below. Note that the first one (triangle) is nonzero on $[0,2]$ and the one on the right has $[0,3]$.





Plots from lecture on 4/19

One can tell that this starts to look Gaussian – and indeed it does! This leads to the limit theorems, the climax of 407.

3.6 The Limit Theorems :o

Proposition 3.6.1: Markov's inequality

Let $\{\Omega, \Sigma, P\}$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable with $X \geqslant 0$. Then for any $a > 0$,

$$P(X \geqslant a) \leqslant \frac{E[X]}{a}.$$

Proof. For $a > 0$, we define an indicator variable

$$I = \begin{cases} 1 & X \geqslant a \\ 0 & \text{otherwise.} \end{cases}$$

Since $X \geqslant 0$, $I \leqslant X/a$ (clear when $I = 0$ and if $I = 1$ then $X \geqslant a$ — still holds). Thus,

$$E[I] \leqslant E[X/a] = \frac{1}{a}E[X] = \frac{E[X]}{a}.$$

It remains to notice that $E[I] = 1 \cdot P(X \geqslant a) + 0 \cdot P(X < a) = P(X \geqslant a)$. This proves the claim. □

Proposition 3.6.2: Chebyshev's Inequality

If X is a r.v. with $\mu = E[X] < \infty$ and $\sigma^2 = \mathrm{Var}(X) < \infty$ then, for any $k > 0$,

$$P(|X - \mu| \geqslant k) \leqslant \frac{\sigma^2}{k^2}$$

or, equivalently,

$$P(|X - \mu| < k) = 1 - P(|X - \mu| \geqslant k) \geqslant 1 - \frac{\sigma^2}{k^2}.$$

In particular, if $k = j\sigma$ for $j \in \mathbb{N}$ then

$$P(|X-\mu| \geqslant j\sigma) \leqslant \frac{\sigma^2}{j^2\sigma^2} = \frac{1}{j^2} \text{ and } P(|X-\mu| < j\sigma) \geqslant 1 - \frac{1}{j^2}.$$

Thus, for any r.v., X being within 2 standard deviations of μ is $\geqslant 3/4$ and X being within 3σ's is $\geqslant 8/9$.

Proof. Notice that $(X-\mu)^2 \geqslant 0$, so we can apply Markov's inequality with $a = k^2$:

$$P((X-\mu)^2 \geqslant k^2) \leqslant \frac{E[(X-\mu)^2]}{k^2} \iff P(|X-\mu| \geqslant k) \leqslant \frac{\sigma^2}{k^2}.$$ □

Beginning of April 21, 2021

Proposition 3.6.3

If $\text{Var}(X) = 0$ then $P(X = E[X]) = 1$. This follows directly from definition (or from Chebyshev).

Weak Law of Large Numbers (Weak LLN)

Theorem 3.6.4: Weak LLN

Let $\{X_i\}_{i\geqslant 1}^{\infty}$ be i.i.d. with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$ for all i. Then for any $\epsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \epsilon\right) \to 0.$$

Proof. Define $\overline{X}_n := \sum_{i=1}^{n} X_i/n$. Note that this is indeed a random variable — randomness happens. We know

$$E[\overline{X}_n] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu.$$

Also by independence

$$\text{Var}(\overline{X}_n) = \text{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{\sigma^2}{n}.$$

By Chebyshev,

$$P(|\overline{X}_n - \mu| \geqslant \epsilon) \leqslant \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to \infty$. □

Remark. Convergence in probability? What it really means is the uniform convergence in distribution.

Definition 3.6.5

For a sequence of functions $\{f_n\}$ and f on $[0,1]$, we define the following types of convergence:

(1) $f_n \to f$ uniformly if, for all $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $\sup_{x\in[0,1]} |f_n(x) - f(x)| < \epsilon$ for all $n \geqslant N$.

(2) $f_n \to f$ pointwise if for all $\epsilon > 0$ and $x \in [0,1]$, there exists $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \epsilon$ for all $n \geqslant N$. In other words, ϵ in uniform convergence doesn't care about choice of x but its pointwise counterpart does.

(3) $f_n \to f$ almost surely (a.s.) if, for all $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that, if $n \geqslant N$, then $f_n \to f$ pointwise (see below) for all $x \in S \subset [0,1]$ where $P([0,1] \setminus S) = 0$. (Analogous to almost everywhere convergence.)

(4) (Weak LLN) $f_n \to f$ in probability if, given $\epsilon > 0$, the probability that x (any x) is in a set $|f_n(x) - f(x)| > \epsilon$ tends to 0 as $n \to \infty$.

Example 3.6.6. A non-example of functions that converges in probability but nowhere pointwise: consider the following sequence:

$$f_1 = \chi_{[0,1]}$$
$$f_2 = \chi_{[0,1/2]}, f_3 = \chi_{[1/2,1]}$$
$$f_4 = \chi_{[0,1/4]}, f_5 = \chi_{[1/4,1/2]}, f_6 = \chi_{[1/2,3/4]}, f_7 = \chi_{[3/4,1]}$$
$$\dots$$

It follow immediately that any, for any $x \in [0,1]$, for any line as formatted above, (at least) one f_n in that line is 1 at x. Therefore $\{f_n\}$ converges pointwise as nowhere! However, they indeed converge in probability: the "size of the set" decreases by a factor of $1/2$ every time we move down a line.
Convergence hierarchy:

$$\text{unif. conv.} \implies \text{pointwise conv.} \implies \text{a.s. conv.} \implies \text{conv. in probability}$$

Recall that an estimator Y_n (that acts on samples of size n) for a population parameter α is called **unbiased** if $E[Y_n] = \alpha$. Besides that, an estimator is called **consistent** if $Y_n \to \alpha$ as $n \to \infty$.

Example 3.6.7. $\overline{X}_n = \sum_{i=1}^{n} X_i/n$ is unbiased as $E[\overline{X}_n] = \mu$. It is also consistent because of weak LLN.
On the other hand, recall that the sample variance

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ is biased as } E[\hat{S}^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Nevertheless, the weak LLN says it is consistent since as $n \to \infty$, $E[\hat{S}^2] = \sigma^2$.

Beginning of April 23, 2021

Central Limit Theorem

Theorem 3.6.8: Central Limit Theorem, CLT

Suppose $\{X_i\}_{i=1}^{\infty}$ are i.i.d. with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Then

$$\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \to Z \text{ as } n \to \infty$$

(where the convergence refers to convergence in distribution) if and only if

$$P\Big(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant a\Big) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2}\, \mathrm{d}x.$$

Lemma 3.6.9

If $\{X_i\}$ is a sequence of r.v.'s with CDFs $F_1, F_2, \ldots$ and MGFs $M_1, M_2, \ldots$. Let Y be a r.v. with CDF F and MGF M. If $M_i \to M$ with each t (pointwise), then $F_i(x) \to F(x)$ at all x when F is continuous (of course). Therefore, to prove CLT, it suffices to prove that the MGFs of $(\sum X_i - n\mu)/(\sigma\sqrt{n})$ converges to that of Z, $e^{t^2/2}$ (recall this?).

Proof of CLT. For simplicity, assume $\mu = 0$ and $\sigma^2 = 1$ (for now). Since X_i's are i.i.d., all of them have the same MGF; call it $M(t)$. For this special case, we want to show that the MGF of $\sum X_i/\sqrt{n}$ converges to $e^{t^2/2}$.
First note that

$$M_{x_i/\sqrt{n}}(t) = E[e^{tX_i/\sqrt{n}}] = E[e^{t\sqrt{n}X_i}] = M(t/\sqrt{n}).$$

Therefore, since the X_i's are independent,

$$M_{\sum X_i/\sqrt{n}}(t) = M(t/\sqrt{n})^n.$$

Now we define $L(t) := \log(M(t))$ [note that $M(t)$ is positive and this is well-defined]. Then

$$L(0) = \log(M(0)) = \log E[e^0] = 0.$$

Also,

$$\begin{aligned} L'(0) &= \frac{\mathrm{d}}{\mathrm{d}t}\log(M(t)) = \frac{1}{M(t)}M'(t)\Big|_{t=0} \\ &= \frac{M'(t)}{N(t)}\Big|_{t=0} = \frac{M'(0)}{M(0)} = \frac{\mu}{1} = \mu = 0, \end{aligned}$$

and

$$\begin{aligned} L''(0) &= \frac{M''(t)M(t) - M'(t)^2}{M(t)^2}\Big|_{t=0} \\ &= \frac{M''(0)M(0) - M'(0)^2}{M(0)^2} = M''(0) = E[X_i^2] - E[X_i]^2 = \sigma^2 = 1. \end{aligned}$$

Recall that we want to show

$$(M(t/\sqrt{n}))^n \to e^{t^2/2} \quad \text{as } n \to \infty,$$

or, equivalently, taking log on both sides,

$$n\log(M(t/\sqrt{n})) = nL(t/\sqrt{n}) \to t^2/2.$$

Indeed,

$$\begin{aligned}\lim_{n\to\infty} nL(t/\sqrt{n}) &= \lim_{n\to\infty} \frac{L(t/\sqrt{n})}{1/n} \\ [\text{L'Hop}] &= \lim_{n\to\infty} \frac{L'(t/\sqrt{n})n^{-3/2}t}{-2/n^2} = \lim_{n\to\infty} \frac{L'(t/\sqrt{n})t}{2/\sqrt{n}} \\ [\text{L'Hop}] &= \lim_{n\to\infty} \frac{-L''(t/\sqrt{n})n^{-3/2}t^2}{-2n^{-3/2}} = \lim_{n\to\infty} L''(t/\sqrt{n})\frac{t^2}{2} \\ &= L''(0)t^2/2 = t^2/2,\end{aligned}$$

and we are done! (Notice that L'Hop is done differentiating both sides by n, not t.)

For the more general case, we will apply the special case ($\mu = 0, \sigma^2 = 1$) to $Y_i := (X_i - \mu)/\sigma$. Then $E[Y_i] = 0$ and $\text{Var}(Y_i) = 1$. Then,

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^n (X_i - \mu)/\sigma}{\sqrt{n}} = \sum_{i=1}^n X_i/\sqrt{n} \to Z,$$

as claimed. Now we are actually done! This marks the end of the climax of 407. □

3.7 Applications of the CLT

Binomial vs. Bernoulli vs. Normal

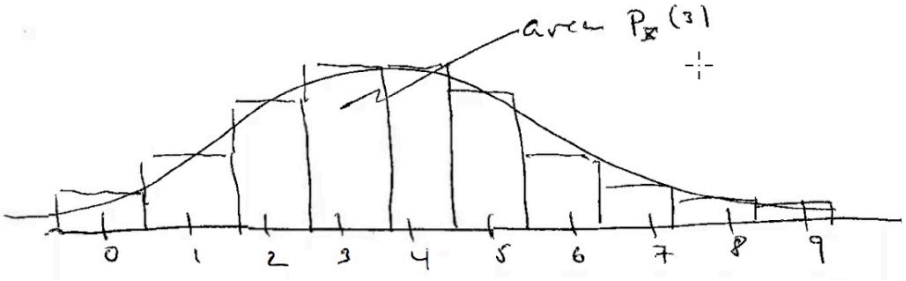
Let $X_i \sim B(1,p)$ be i.i.d. with $i = 1, 2, ..., n$. Recall that

$$E[X_i] = p \qquad \text{Var}(E_i) = p(1-p).$$

Now we define $X := \sum_{i=1}^n X_i$. It follows that

$$\begin{aligned}P(X \leqslant x) = P(\sum_{i=1}^n X_i \leqslant x) &= P\left(\frac{\sum_{i=1}^n X_i - np}{\sqrt{p(1-p)}\sqrt{n}} \leqslant \frac{x-np}{\sqrt{p(1-p)}\sqrt{n}}\right) \\ [\text{CLT}] \approx P\left(Z \leqslant \frac{x-np}{\sqrt{p(1-p)}\sqrt{n}}\right) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\cdots} e^{-t^2/2}\, dt.\end{aligned}$$

(See figure from lecture below.) The larger the n, the better this approximation for obvious reasons. After all, L in CLT stands for large!



Here we are approximating a discrete random variable that only gives integer values. To balance the errors it is better to set the integer values as the center of each box and let the endpoints be .5's. Approximating a discrete distribution gives rise to **continuity correction**. For example,

$$P(X \leqslant 3) \approx P\left(\frac{3.5 - np}{\sqrt{np(1-p)}}\right), P(X < 3) \approx P\left(Z \leqslant \frac{2.5 - np}{\sqrt{np(1-p)}}\right),$$

and $X = 3$ corresponds to the entire rectangle of length 1, so

$$P(X = 3) \approx P\left(\frac{2.5 - np}{\sqrt{np(1-p)}} \leqslant Z \leqslant \frac{3.5 - np}{\sqrt{np(1-p)}}\right).$$

Election Polling

These materials are originally prepared for fall 2020's class so... not that interesting now (spring 2021).

Suppose we have a poll that says the following:

$$\begin{bmatrix} \text{Biden} & \hat{p} \\ \text{Trump} & 1-\hat{p} \end{bmatrix} \quad \begin{matrix} \text{MARGIN OF ERROR } \pm 3\% (\pm 0.03) \\ \text{SURVEY OF 1200 LIKELY VOTERS} \end{matrix}$$

(*Of course this is idealized as they are not the only candidates.*) Let the true proportion that supports Biden be p. Ideally, we want $\hat{p} = p$ (or close enough). Now define

$$X_i := \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ voter supports Biden} \\ 0 & \text{if the } i^{\text{th}} \text{ voter supports Trump.} \end{cases}$$

We can think of X_i as $B(1, p)$. Based on what we know (of $\hat{p}$), we can define

$$\hat{p}_n := \frac{\sum_{i=1}^n X_i}{n} = \frac{1}{n}\sum_{i=1}^n X_i.$$

Here's where the famous **95% confidence interval** for p comes into play. In a standard normal,

$$P(-1.96 \leqslant Z \leqslant 1.96) = Z,$$

so by CLT we have

$$\begin{aligned} 0.95 &\approx P\left(-1.96 \leqslant \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} \leqslant 1.96\right) \\ &= P\left(-1.96 \leqslant \frac{\left[\sum_{i=1}^n X_i - p\right]/n}{\sqrt{p(1-p)/n}} \leqslant 1.96\right) \\ &= P(-1.06 \leqslant (\hat{p} - p)/\sqrt{p(1-p)/n} \leqslant 1.96) \\ &= P(\hat{p} - 1.96\sqrt{p(1-p)/n} \leqslant p \leqslant \hat{p} + 1.96\sqrt{p(1-p)/n}). \end{aligned}$$

Therefore, there is a 95% chance that the true ratio p lies within this interval, i.e., 95% probability that

$$|p - \hat{p}| \leqslant 1.96\sqrt{\frac{p(1-p)}{n}}.$$

With some conservative estimation ($1.96 < 2$ and $p(1-p) \leqslant 1/4$ for $p \in [0, 1]$), taking $n = 1200$ gives

$$|p - \hat{p}| \leqslant 2 \cdot \sqrt{\frac{1}{4n}} \approx 0.029 < 0.03.$$

This explains where the 1200 and 0.03 comes from.

Beginning of April 26, 2021

Monte-Carlo Integration (Application of LLN)

Suppose $g:[0,1]\to\mathbb{R}$. What is $\int_0^1 g(x)\,\mathrm{d}x$? How do we compute it in the 407 way?

Claim. We can let $X\sim U(0,1)$, $X_i\sim X$ be i.i.d. Then

$$\int_0^1 g(x)\,\mathrm{d}x=\int_{-\infty}^{\infty} g(x)\chi_{[0,1]}(x)\,\mathrm{d}x=E[g(X)]$$
$$[\text{LLN}]=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

It follows that

$$\int_0^1 g(x)\,\mathrm{d}x\approx\frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

In pre-analysis style integration, a 1-fold integral takes n evaluations, a 2-fold integral takes n^2 evaluations, and a k-fold integration takes n^k integrations. This number grows exponentially[!]

Also, X need not to be uniform. Suppose $X\sim f$ and $X_i\sim X$ i.i.d., then we can choose g that overweights the important region of f (so g vanishes on where f is insignificant) with

$$\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x=\int_{-\infty}^{\infty}\frac{g(x)}{f(x)}f(x)\,\mathrm{d}x=E[(g/f)(X)]$$
$$[\text{LLN}]=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{g}{f}(X_i).$$

This is known as the **importance sampling**.

3.8 Simulating Randomness

Question. How to simulate randomness?

Answer. What's so-called a *pseudo-random number generator* that generates a number between 0 and 1. Nevertheless these numbers are still from a formula, except the formula is huge and we will almost never notice the pattern. In MATLAB this is the `rand()` function.

Since numbers are stored discretely in computers, we can only approximate the uniform distribution by discrete ones. In particular, the computer would use a number from the following list:

$$\frac{1}{2^{53}-1},\frac{2}{2^{53}-1},\dots,\frac{2^{53}}{2^{53}-1}.$$

($2^{53}-1$ is the current number MATLAB uses; machine eps?)

Beginning of April 28, 2021

For example, if we were to simulate rolling a dice, instead of actually asking the compute to roll a dice, we would call the `rand()` function and the computer would return a (more or less) random number between 0 and 1 (specifically, from the list above).

More generally, if $\Omega=\{\omega_1,\dots,\omega_n\}$ where $P(\{\omega_i\})=P_i$, we let

$$r_0=0,r_1=p_1,\dots,r_n=\sum_{i=1}^{n} r_i.$$

It is clear that these r_i's partition the interval $[0,1]$ — and we can simulate a random experiment using this trick!

Say we called `rand()`. For example,

$$P(\{\omega_4\}) = P(r_3 < \texttt{rand()} < r_4).$$

We cannot really simulate an experience with an infinite sample space, but we can "truncate" Ω and lump all the discarded terms into the last term that is not discarded. A concrete example: consider our same old example of flipping a coin until head comes up:

$$\Omega = \{H, TH, TTH, TTTH, ...\}$$

so $P_1 = \frac{1}{2}, P_2 = \frac{1}{4}, \ldots, P_n = \frac{1}{2^n}$. However, since we discard all the remaining ones, we add everything else to P_n, namely

$$P_n = \frac{1}{2^n} + \sum_{i=n+1}^{\infty} \frac{1}{2^i} = \frac{1}{2^{n-1}}.$$

Pseudo RNG

How do we generate these random numbers then? It is done by a formula such that every time we call `rand()` it produces a number between 0 and 1. The numbers should appear *equally likely* and *independently*. The algorithm is supported by statistical tests to ensure that they are "random" enough (e.g. histogram / chi-squared test).

Definition 3.8.1

A **Mersenne prime** is a prime number of form $2^p - 1$ where p itself is a prime. There are infinitely many Mersenne primes and the current biggest one is $2^{82,589,933} - 1$, found in Dec 2018.

Definition 3.8.2: PRNG

Let p be a large prime number. We will construct a number generator (formula) that prodocues a number from the set $\{1, 2, ..., p_1\}$. Then the PRNG will return this number divided by p. There will be three parameters involved:

(1) p a sufficiently large,

(2) R_0 the seed, a number from $\{1, 2, ..., p_1\}$, and

(3) a, the multiplier, a number from $\{2, 3, ..., p-1\}$.

The seed is generated by

$$R_n = a^n R_0 \pmod{p} = aR_{n-1} \pmod{p}.$$

If p is not large enough, we will eventually run into a pattern once the list of numbers has been exhausted and it starts over again!).

The seed only determines where on the list will we begin; what determines the list is a and p.

For example, if $p = 7, R_0 = 5, a = 2$:

R_n	$2R_n$	$[2R_n]$
5	10	3
3	6	6
6	12	5
5	10	3

and it becomes clear that this choice does not exhaust the list of all possible numbers between 1 and $p-1$, i.e., $R_0\langle 2\rangle \neq \mathbb{Z}/p\mathbb{Z}$. This is bad because the only numbers it will generate are $3/7, 5/7$, and $6/7$, where numbers like $1/7, 2/7$, and $4/7$ did not appear at all (but we want them for a PRNG). For this specific case, letting $a = 3$ fixes the problem: $5-1-3-2-6-4-5$. Then PRNG gives these numbers divided by 7, and that's a nice PRNG for this particular p. Therefore if we could find a nice a for $p = 2^{53}-1$, we would have a pretty nice PRNG. In `Matlab` currently this a is 7^5.

Simulating Continuous Random Variables

Suppose we can given a strictly (or not) increasing function $F(x)$ (so that it admits an inverse). Suppose it goes from 0 to 1 on some specified interval. We want to produce a r.v. X with cdf F using a random variable $U \sim U(0,1)$. For $X = F^{-1}(U)$, we have

$$P(X \leqslant x) = P(F^{-1}(U) \leqslant x) = O(U \leqslant F(x)) = F(x).$$

This shows that X has the desired cdf F.

Example 3.8.3. Consider an exponential r.v. with $f_X(x) = \alpha e^{-\alpha x}$ for $x \geqslant 0$ and some $\alpha > 0$. The cdf is

$$F_X(x) = \int_0^x \alpha e^{-\alpha t}\,\mathrm{d}t = 1 - e^{-\alpha x}.$$

The inverse of this cdf is

$$F^{-1}(u) = -\frac{\ln(1-u)}{\alpha}.$$

What if F^{-1} does not have a simple form, for example that of a normal distribution, which we cannot even write down the cdf explicitly?

Rejection Sampling

Definition 3.8.4

A $\mathbb{R}^2$ analogue of uniform random variables: let $B \subset \mathbb{R}^2$ and let X, Y be random variables. We say they have a uniform distribution on B, i.e., $(X,Y) \sim U(B)$, if, for every $A \subset B$,

$$P((X,Y) \in A) = \frac{\text{area}(A)}{\text{area}(B)}.$$

The workaround? Instead of trying hard to find F^{-1}, now we consider the graph of f and the area under f, which forms a subset of $\mathbb{R}^2$. Put formally, define $B := \{(x,y) : 0 \leqslant y \leqslant f(x)\}$ and let (X,Y) be a sample from $U(B) : (X,Y) \sim U(B)$. Define $A_{\tilde{x}}$ to be the subset of B with $x \leqslant \tilde{x}$, i.e., $A_{\tilde{x}} := \{(x,y) \in B : x < \tilde{x}\}$. Then,

$$P(X \leqslant \tilde{x}) = P((X,Y) \in A_{\tilde{x}}) = \frac{\text{area}(A_{\tilde{x}})}{\text{area}(B)} = \text{area}(A_{\tilde{x}}) = \int_{-\infty}^{\tilde{x}} f(x)\,\mathrm{d}x = F(\tilde{x}),$$

thanks to the fact that area of $B = 1$ as it's the area under a pdf.

So, the question becomes: how to generate a sample from $U(B)$? This is done by **rejection sampling**. Suppose we have some larger set $C \supset B$ that we already know how to sample (a nice rectangle, for example, for f with compact support). Then we can performance what's known as the rejection sampling algorithm.

```
1  Choose (X', Y') ~ C
2  if (X', Y') ∈ B
3     X = X'
4     break
5  else go to line 1 % reject
```

Then,

$$\begin{aligned} P(X \leqslant \tilde{x}) &= P(X \leqslant \tilde{x} \mid (X', Y') \in B) \\ &= \frac{P((X', Y') \in A_{\tilde{x}} \cap B)}{P((X', Y') \in B)} \\ &= \frac{\text{area}(A_{\tilde{x}} \cap B)/\text{area}(C)}{\text{area}(B)/\text{area}(C)} = \frac{\text{area}(A_{\tilde{x}} \cap B)}{\text{area}(B)} = \text{area}(A_{\tilde{x}}) = F(\tilde{x}), \end{aligned}$$

again, thanks to the fact that $\text{area}(B) = 1$.

Example 3.8.5. Suppose $f(x) = 0$ unless $x \in [a, b]$ and f is uniformly bounded by M. Then we can simply take C to be $[a, b] \times [0, M]$. (We choose the box as small as possible to maximize efficiency.) To sample from $U(C)$, we can use $U_1, U_2 \sim U(0, 1)$ i.i.d. Then

$$(X', Y') = (a + (b - a)U, MU_2) \sim U(C).$$

Example 3.8.6. But what if f is not compactly supported? We can still apply the same idea but with some modifications. Instead of covering the region by a box (which we can't in this case), we can dominate the cdf by another distribution, of which we know the distribution! One nice candidate is the exponential distribution — as shown above, we can find explicit formula for its inverse, and it's defined on $[0, \infty)$. We can extend it into a "doubly exponential distribution" with domain $(-\infty, \infty)$. Then we multiply it by some constant M to bound our not so nicely-behaved cdf of f.

To put formally, let $X \sim h$ and, given X, select $Y \sim U(0, Mh(X))$. Then $(X, Y) \sim U(C)$ and we can do rejection sampling once more. Notice that $(X, Y) \sim U(C)$. Indeed,

$$\begin{aligned} f_{X,Y}(x, y) &= f_X(x) f_{Y|X}(y \mid x) \\ &= \frac{h(x)\chi_{(0,Mh(x))}}{Mh(x)} = \frac{\chi_C}{M} \end{aligned}$$

which is uniform on C.

End of Course