Probability and Statistics, From Classical to Bayesian

Week 5

25th May - 1st June 2021

1 Introduction

This week we will be focusing on some important inequalities that are often encountered in probability, we will then move on to a fundamental concept in probability which is convergence and state some of the most famous theorems in Probability and Statistics.

2 Inequalities

Let X be a non-negative integer valued random variable with pmf f(k) for k = 0, 1, 2, ... Fix any number m, say m = 10. Then

$$E[X] = \sum_{k=1}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} 10f(k) = 10P\{X \geq 10\}$$

More generally m $mP\{X \geq m\} \leq E[X]$.

Markov's inequality. Let X be a non-negative random variable with finite expectation. Then, for any t > 0, we have $P\{X \geq t\} \leq \frac{1}{t}E[X]$.

(Note. This inequality might not hold for t=0 in many cases, be careful while setting value of t)

Markov's inequality is simple, but surprisingly useful. Firstly, one can apply it to functions of our random variable and get many inequalities. Here are some. Variants of Markov's inequality.

- (1) If X is a non-negative random variable with finite pth moment, then $P\{X \geq t\} \leq t^{-p}E[X^p]$ for any $t > 0$.
- (2) If X is a random variable with finite second moment and $\mu = E[X]$, then $P\{|X - \mu| \geq t\} \leq t^{-2}Var(X)$ [Chebyshev's inequality]
- (3) If X is a random variable with finite exponential moments, then $P\{X \geq t\} \leq e^{-\lambda t}E[e^{\lambda X}]$ for any $\lambda > 0$. [Chernoff's inequality]

Thus, if we only know that X has finite mean, the tail probability $P\{X \geq t\}$ must decay at least as fast as $\frac{1}{t}$. But, if we knew that the second moment was finite we could assert that the decay must be at least as fast as $\frac{1}{t^2}$, which is better. If $E[e^{\lambda X}] < 1$, then we get much faster decay of the tail, like $e^{-\lambda t}$.
Chebyshev's inequality captures again the intuitive notion that variance measures the spread of the distribution about the mean. The smaller the variance, lesser the spread. An alternate way to write Chebyshev's inequality is

$$P\{\frac{|X-\mu|}{\sigma} \geq t\} \leq t^{-2}$$

Jensen's Inequality. Let I $\subset$ R be an interval and let $\psi . I \to R$ be a twice differentiable function such that its second order derivative ψ'' is continuous on I and $\psi''(x) \geq 0$ $\forall x \in R$ (i.e., ψ is convex). Let X be a random variable with support $S_X \subset I$, and finite expectation. Then,

$$E[\psi(X)] \geq \psi(E[X])$$

This inequality holds for a convex function defined on the whole support set of X. Here is a intuitive proof of the above statement Jensen's.
The reverse holds for concave function, try working that out by yourselves.

3 Convergence in Random Variables

Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables, and X be another random variable defined on the same probability space.
Convergence in quadratic mean (q.m.). Suppose that $\{X_n\}_{n\geq 1}$ and X are such that $E[X_n^2] < \infty$for all $n \geq 1$ and $E[X^2] < \infty$. Then, the sequence $\{X_n\}_{n\geq 1}$ converges to X in q.m. if

$$\lim_{n\to\infty} E[(X_n - X)^2] = 0$$

This is denoted by $X_n \xrightarrow{q.m.} X$ (or, $X_n \xrightarrow{L^2} X$).
By definition, $X_n \xrightarrow{q.m.} X$ if and only if $(X_n - X) \xrightarrow{q.m.} 0$ as $n \to \infty$
Convergence in probability (P). The sequence $\{X_n\}_{n\geq 1}$ converges to X in probability, denoted by $X_n \xrightarrow{P} X$ if

$$\lim_{n\to\infty} P\{\omega . |X_n(\omega) - X(\omega)| > \epsilon\} = 0$$

for every $\epsilon > 0$.
By definition, $X_n \xrightarrow{P} X$ if and only if $(X_n - X) \xrightarrow{P} 0$ as $n \to \infty$
Convergence in q.m. implies convergence in P. By Chebyshev's inequality, for each $\epsilon > 0$

$$0 \leq P\{|X_n - X| > \epsilon\} \leq \frac{E[|X_n - X|^2]}{\epsilon^2}$$

Therefore, $E[|X_n - X|^2] \to 0$ implies that $P|X_n - X| > \epsilon \to 0$ as $n \to \infty$.
The reverse might not always be true, ie. Convergence in P need not imply does not necessarily imply convergence in q.m. . Try to find an example to support this statement.
Let $h . \mathbb{R} \to \mathbb{R}$ be a continuous function If $X_n \xrightarrow{P} X$, then $h(X_n) \xrightarrow{P} h(X)$ as $n \to \infty$.

4 Sampling Distributions

This section is named Sampling Distributions as these distributions do come up a lot when you have to "sample" some distributions and proofs of them. You will see more of them as we go along.

χ_r^2 **distribution.** A special case of the gamma distribution in which a = r/2 (where r is a positive integer) and b = 1/2. It is said to have r degree of freedom(s).
. Let $Z_1, \ldots, Z_n$ be i.i.d. $N(0,1)$random variables. Then, $Z_1^2 + \cdots + Z_n^2$ has Gamma(n/2, 1/2) distribution.

t-distribution. Let W denote a random variable that is N(0, 1) and V denote a random variable that is χ_r^2, with W and V independent. Then, the joint pdf of W and V (say, h(w, v)) is

$$h(w,v) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \frac{1}{\Gamma(r/2)2^{r/2}} v^{r/2-1} e^{-v/2} & -\infty < w < \infty, 0 < v < \infty \\ 0 & elsewhere \end{cases}$$

Define a new random variable T by writing

$$T = \frac{W}{\sqrt{V/r}}$$

You can check that distribution of T is indeed a PDF by trying change of multiple variables on above pdf. The distribution of the random variable T is usually called a t-distribution.

5 Weak Law of Large Numbers

Independent and identically distributed random variables

Let $X_1, X_2, \ldots$ be i.i.d random variables. Assume $\mu = E[X_1]$ and $\sigma^2 = Var(X_1)$ are well-defined. Let $S_n = X_1 + \cdots + X_n$ (partial sums) and $\bar{X_n} = \frac{S_n}{n} = \frac{X_1 + \cdots + X_n}{n}$ (sample mean). Then, by the properties of expectation and variance, we have

$$E[S_n] = n\mu, Var(S_n) = n\sigma_1^2,\ E[\bar{X_n}] = \mu, \text{and } Var(\bar{X_n}) = \sigma^2/n.$$

. If we apply Chebyshev's inequality to $\bar{X_n}$, we get for any $\delta > 0$ that

$$0 \le P\{|X_n - \mu| \ge \delta\} \le \frac{\sigma^2}{\delta^2 n}$$

This goes to zero as n tends to infinity(as $n \to \infty$) with delta being fixed. Above equation more or less represents, Weak Law of Large Numbers. If X_k are assumed to have more moments, one can get better bounds (using convergence in q.m. that implies convergence in P for above inequality).

6 Monte Carlo

6.1 Sampling/Stimulating

Probability is applicable to many situations in the real world. As such one may conduct experiments to verify the extent to which theorems are actually valid. For this we need to be able to draw numbers at random from any given distribution.

For example, take the case of Bernoulli(1/2) distribution. One experiment that can give this is that of physically tossing a coin. This is not entirely satisfactory for several reasons. Firstly, are

real coins fair? Secondly, what if we change slightly and want to generate from Ber(0.45)? Instead of going out and looking for physical phenomenons, we prefer to stimulate random variables on computers. Python provides built in libraries to draw numbers from common distributions, even in the case that our distribution is uncommon, there are many ways to deal with it.
The act of generating numbers following a certain distribution is called stimulating/sampling.

6.2 Monte-Carlo method

Till now we have been using integration to answer our questions in probability, now lets do the reverse. Instead of going through the tedious calculations, lets use probability to approximate integrals. The justification for the method comes from the weak law of large numbers (WLLN), which will show that if take simulate our random numbers infinitely many times, then we can get the exact answer. A better justification is from Chebyshev's inequality, which will show us the extent and probability of our error if we sample a large, but finitely many random numbers.

The problem of integration. Let ψ . $[0,\ 1] \to$ R be a continuous function. We would like to compute $I = \int_0^1 \psi(x)dx$. Most often we cannot compute the integral explicitly and for an approximate value we resort to numerical methods. Here is an idea to use random numbers. Let $U_1, U_2, ..., U_n$ be i.i.d. Unif[0, 1] random variables and let $X_1 = \psi(U1), ..., X_n = \psi(U_n)$. Then, X_k are i.i.d. random variables with common mean and variance.

$$\mu = \int_0^1 \psi(x)dx = I \ , \ \sigma^2 = \int_0^1 (\psi(x) - I)^2 dx$$

This gives the following method of finding I. Fix a large number N (appropriately) and pick N uniform random numbers U_k for $1 \le k \le N$. Then, define

$$I_N = \frac{1}{N} \sum_{k=1}^{N} X_k$$

Present I_N as an approximate value of I.

In what sense is this an approximation of I, and why? Indeed, by the WLLN we have $P\{|I_N - I| > \delta\} \to 0$ (mean converges to expectation), and hence we expect I_N to be close to I. How large should N be? For this, we fix two numbers $\epsilon = 0.01$ and $\delta = 0.001$ (you may change the numbers). By Chebyshev's inequality, observe that $P\{|I_N - I| > \delta\} \le \frac{\sigma^2}{N\delta^2}$. Epsilon represents the probability of our approximation being outside the delta neighborhood of I. To get better results we decrease these and get a higher value of N (more the samples we take, the closer we get to the answer).

When the limits are different we use other distributions instead of Uniform. If the limits are a and b of the integral, we use Unif[a,b] when both are finite, shifted exponential when $b \to \infty$ and normal when $a \to -\infty$ and $b \to \infty$. Obviously we have to manipulate our function that we apply on the variables a little bit to make sure the expression for expectation matches the desired integral.

7 Convergence in Distribution

The sequence $\{X_n\}_{n\leq 1}$ is said to converge in distribution to X (written as $X_n \xrightarrow{D} X$) as $n \to \infty$ if

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \forall \quad x \in C_F$$

where, C_F is the set of continuity points of F.

Suppose that there exists $a > 0$ such that the mgfs M and Mn of X and X_n, respectively, are finite on (-a, a) for all n = 1, 2, ... If $\lim_{n\to\infty} M_n(t) = M(t)\ \forall\ t \in (-a, a)$, then $(X_n \xrightarrow{D} X)$ as $n \to \infty$.

- If $X_n \xrightarrow{D} X$, then $h(X_n) \xrightarrow{D} h(X)$ for any continuous function h.
- $X_n \xrightarrow{P} X \to X_n \xrightarrow{D} X$, but the reverse implication is not true in general.

Slutsky's Theorem: If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} c \in \mathbb{R}$, then $Xn+Yn \xrightarrow{D} X+c$ and $XnYn \xrightarrow{D} cX$.

8 Central Limit theorem

Let $X_1, X_2, ...$ be i.i.d. random variables with expectation μ and variance σ^2. We saw that $\bar{X}_n$ has mean μ and standard deviation $= \sigma/\sqrt{n}$. WLLN roughly means that $\bar{X}_n$ is close to μ, within a few multiples of $\sigma/\sqrt{n}$ (as shown by Chebyshev's inequality). Now, we look at $\bar{X}_n$ with a finer microscope. In other words, we ask how exactly does $\bar{X}_n$ behave in an interval around μ, the result turns out to be surprising and remarkable!.

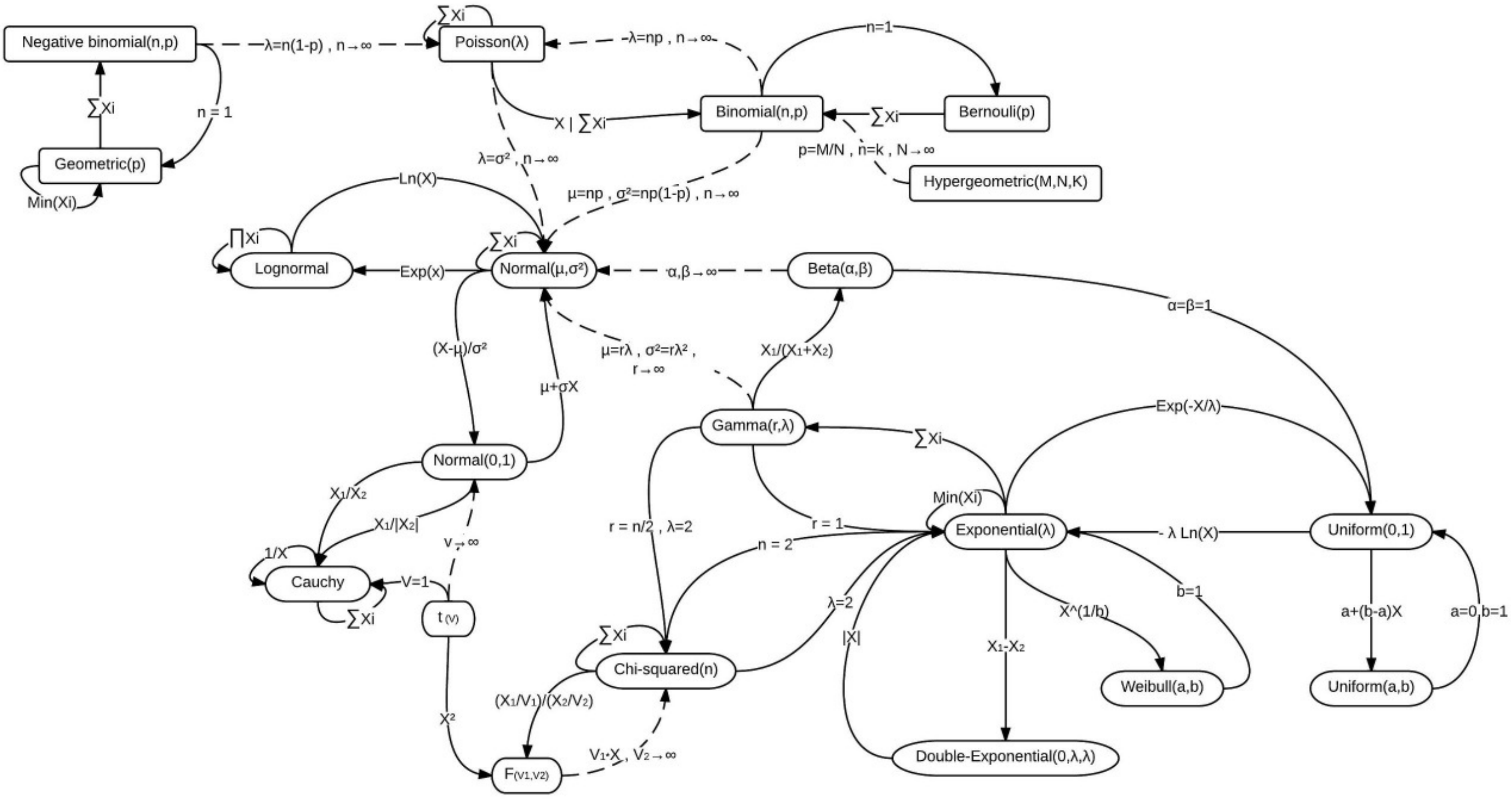
Central limit theorem. Let $X_1, X_2, ...$ be i.i.d. random variables with expectation μ and variance σ^2. We assume that the variance is finite and non zero. Define $Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$, and let Z be a N(0, 1) random variable. Then,

$$Z_n \xrightarrow{D} Z \text{ as } n \to \infty$$

What is remarkable about this? The end result does not depend on the distribution of $X_i s$ at all! Only the mean and variance of the distribution were used! We will get to see more of this result next week when we start statistics.

9 Generally used Results

These results are generally used while simplification, instead of applying change of variable every time, you can refer them from here. Obviously you can check all this is correct using change of variables, though that is not recommended. (Dotted line represents approximate results)



Negative binomial(n,p)
λ=n(1-p) , n→∞
Poisson(λ)
∑Xi
λ=np , n→∞
n=1
Binomial(n,p)
Bernouli(p)
∑Xi
n = 1
∑Xi
Geometric(p)
Min(Xi)
X | ∑Xi
λ=σ² , n→∞
p=M/N , n=k , N→∞
Hypergeometric(M,N,K)
Ln(X)
μ=np , σ²=np(1-p) , n→∞
∏Xi
∑Xi
Lognormal
Exp(x)
Normal(μ,σ²)
α,β→∞
Beta(α,β)
α=β=1
(X-μ)/σ²
μ=rλ , σ²=rλ² , r→∞
X₁/(X₁+X₂)
μ+σX
Exp(-X/λ)
Gamma(r,λ)
∑Xi
Normal(0,1)
X₁/X₂
X₁/|X₂|
Min(Xi)
r = n/2 , λ=2
r = 1
Exponential(λ)
- λ Ln(X)
Uniform(0,1)
1/X
v→∞
n = 2
Cauchy
V=1
∑Xi
t(V)
λ=2
X^(1/b)
b=1
a+(b-a)X
a=0,b=1
|X|
X₁-X₂
∑Xi
Chi-squared(n)
Weibull(a,b)
Uniform(a,b)
(X₁/V₁)/(X₂/V₂)
X²
V₁·X , V₂→∞
F(V1,V2)
Double-Exponential(0,λ,λ)