Lecture Notes 1
36-705

Our broad goal for the first few lectures is to try to understand the behaviour of sums of independent random variables. We would like to find ways to formalize the fact:

Averages of independent random variables concentrate around their expectation.

We will try to answer this question from the asymptotic (i.e. the number of random variables we average $\to \infty$) and the non-asymptotic viewpoint (i.e. the number of random variables is some fixed finite number). The asymptotic viewpoint is typically characterized by what are known as the Laws of Large Numbers (LLNs) and Central Limit Theorems (CLTs) while the non-asymptotic viewpoint is characterized by concentration inequalities.

We will need to first review what a random variable is, what its expectation is, and what we precisely mean by concentration. This will be fairly terse. See Chapters 1-3 of the book for more details.

Warning: This is a review. We will go quickly because I assume you have taken some probability.

1 Random Variables

Let Ω be a sample space (a set of possible outcomes) with a probability distribution (also called a probability measure) P. A *random variable* is a map $X : \Omega \to \mathbb{R}$. We write

$$P(X \in A) = P(\{\omega \in \Omega : \ X(\omega) \in A\})$$

and we write $X \sim P$ to mean that X has distribution P. The *cumulative distribution function (cdf)* of X is

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

A cdf has three properties:

1. F is right-continuous. At each x, $F(x) = \lim_{n\to\infty} F(y_n) = F(x)$ for any sequence $y_n \to x$ with $y_n > x$.
2. F is non-decreasing. If $x < y$ then $F(x) \leq F(y)$.
3. F is normalized. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

Conversely, any F satisfying these three properties is a cdf for some random variable.

If X is discrete, its *probability mass function (pmf)* is

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

If X is continuous, then its *probability density function function (pdf)* satisfies

$$P(X \in A) = \int_A p_X(x)dx = \int_A p(x)dx$$

and $p_X(x) = p(x) = F'(x)$. The following are all equivalent:

$$X \sim P, \quad X \sim F, \quad X \sim p.$$

Suppose that $X \sim P$ and $Y \sim Q$. We say that X and Y have the same distribution if $P(X \in A) = Q(Y \in A)$ for all A. In that case we say that X and Y are *equal in distribution* and we write $X \stackrel{d}{=} Y$.

Lemma 1 *$X \stackrel{d}{=} Y$ if and only if $F_X(t) = F_Y(t)$ for all t.*

2 Expected Values

The *mean* or expected value of $g(X)$ is

$$\mathbb{E}\left(g(X)\right) = \int g(x)dF(x) = \int g(x)dP(x) = \begin{cases} \int_{-\infty}^{\infty} g(x)p(x)dx & \text{if } X \text{ is continuous} \\ \sum_j g(x_j)p(x_j) & \text{if } X \text{ is discrete.} \end{cases}$$

Recall that:

1. **Linearity of Expectations:** $\mathbb{E}\Big(\sum_{j=1}^k c_j g_j(X)\Big) = \sum_{j=1}^k c_j \mathbb{E}(g_j(X))$.
2. If $X_1, \ldots, X_n$ are independent then
$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_i \mathbb{E}\left(X_i\right).$$
3. We often write $\mu = \mathbb{E}(X)$.

4. $\sigma^2 = \mathsf{Var}\,(X) = \mathbb{E}\,((X-\mu)^2)$ is the **Variance**.

5. $\mathsf{Var}\,(X) = \mathbb{E}\,(X^2) - \mu^2$.

6. If $X_1, \ldots, X_n$ are independent then

$$\mathsf{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_i a_i^2 \mathsf{Var}\,(X_i)\,.$$

7. The covariance is

$$\mathsf{Cov}(X,Y) = \mathbb{E}\Big((X-\mu_x)(Y-\mu_y)\Big) = \mathbb{E}(XY) - \mu_X \mu_Y$$

and the correlation is $\rho(X,Y) = \mathsf{Cov}(X,Y)/\sigma_x\sigma_y$. Recall that $-1 \leq \rho(X,Y) \leq 1$.

The **conditional expectation** of Y given X is the random variable $\mathbb{E}(Y|X)$ whose value, when $X = x$ is

$$\mathbb{E}(Y|X=x) = \int y\; p(y|x)dy$$

where $p(y|x) = p(x,y)/p(x)$.

The *Law of Total Expectation* or *Law of Iterated Expectation*:

$$\mathbb{E}(Y) = \mathbb{E}\big[\mathbb{E}(Y|X)\big] = \int \mathbb{E}(Y|X=x)p_X(x)dx.$$

The *Law of Total Variance* is

$$\mathsf{Var}(Y) = \mathsf{Var}\big[\mathbb{E}(Y|X)\big] + \mathbb{E}\big[\mathsf{Var}(Y|X)\big].$$

The *moment generating function (mgf)* is

$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

If $M_X(t) = M_Y(t)$ for all t in an interval around 0 then $X \stackrel{d}{=} Y$.

The moment generating function can be used to "generate" all the moments of a distribution, i.e. we can take derivatives of the mgf with respect to t and evaluate at $t = 0$, i.e. we have that

$$M_X^{(n)}(t)|_{t=0} = \mathbb{E}\,(X^n)\,.$$

3 Independence

X and Y are *independent* if and only if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

for all A and B.

Theorem 2 *Let (X, Y) be a bivariate random vector with $p_{X,Y}(x, y)$. X and Y are independent iff $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.*

$X_1, \ldots, X_n$ are independent if and only if

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

Thus, $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$.

If $X_1, \ldots, X_n$ are independent and identically distributed we say they are iid (or that they are a random sample) and we write

$$X_1, \ldots, X_n \sim P \quad \text{or} \quad X_1, \ldots, X_n \sim F \quad \text{or} \quad X_1, \ldots, X_n \sim p.$$

4 Transformations

Let $Y = g(X)$ where $g : \mathbb{R} \to \mathbb{R}$. Then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \int_{A(y)} p_X(x)dx$$

where

$$A(y) = \{x : \ g(x) \leq y\}.$$

The density is $p_Y(y) = F'_Y(y)$. If g is strictly monotonic, then

$$p_Y(y) = p_X(h(y)) \left| \frac{dh(y)}{dy} \right|$$

where $h = g^{-1}$.

Example 3 *Let $p_X(x) = e^{-x}$ for $x > 0$. Hence $F_X(x) = 1 - e^{-x}$. Let $Y = g(X) = \log X$. Then*

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(\log(X) \leq y) \\ &= P(X \leq e^y) = F_X(e^y) = 1 - e^{-e^y} \end{aligned}$$

and $p_Y(y) = e^y e^{-e^y}$ for $y \in \mathbb{R}$.

Example 4 Practice problem. *Let X be uniform on $(-1, 2)$ and let $Y = X^2$. Find the density of Y.*

Let $Z = g(X, Y)$. For example, $Z = X + Y$ or $Z = X/Y$. Then we find the pdf of Z as follows:

1. For each z, find the set $A_z = \{(x, y) : g(x, y) \leq z\}$.
2. Find the CDF

$$F_Z(z) \quad = \quad P(Z \leq z) = P(g(X, Y) \leq z) = P(\{(x, y) : g(x, y) \leq z\}) = \int\int_{A_z} p_{X,Y}(x, y)dxdy.$$

3. The pdf is $p_Z(z) = F_Z'(z)$.

Example 5 Practice problem. *Let (X, Y) be uniform on the unit square. Let $Z = X/Y$. Find the density of Z.*

5 Important Distributions

Normal (Gaussian). $X \sim N(\mu, \sigma^2)$ if

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

If $X \in \mathbb{R}^d$ then $X \sim N(\mu, \Sigma)$ if

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right).$$

Chi-squared. $X \sim \chi^2_p$ if $X = \sum_{j=1}^p Z_j^2$ where $Z_1, \ldots, Z_p \sim N(0, 1)$.

Non-central chi-squared (more on this below). $X \sim \chi^2_1(\mu^2)$ if $X = Z^2$ where $Z \sim N(\mu, 1)$.

Bernoulli. $X \sim \text{Bernoulli}(\theta)$ if $\mathbb{P}(X = 1) = \theta$ and $\mathbb{P}(X = 0) = 1 - \theta$ and hence

$$p(x) = \theta^x (1 - \theta)^{1-x} \qquad x = 0, 1.$$

Binomial. $X \sim \text{Binomial}(\theta)$ if

$$p(x) = \mathbb{P}(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad x \in \{0, \ldots, n\}.$$

Uniform. $X \sim \text{Uniform}(0, \theta)$ if $p(x) = I(0 \leq x \leq \theta)/\theta$.

Poisson. $X \sim \text{Poisson}(\lambda)$ if $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $x = 0, 1, 2, \ldots$. The $\mathbb{E}(X) = \mathsf{Var}(X) = \lambda$ and $M_X(t) = e^{\lambda(e^t-1)}$. We can use the mgf to show: if $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$, independent then $Y = X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Multinomial. The multivariate version of a Binomial is called a Multinomial. Consider drawing a ball from an urn with has balls with k different colors labeled "color 1, color 2, ..., color k." Let $p = (p_1, p_2, \ldots, p_k)$ where $\sum_j p_j = 1$ and p_j is the probability of drawing color j. Draw n balls from the urn (independently and with replacement) and let $X = (X_1, X_2, \ldots, X_k)$ be the count of the number of balls of each color drawn. We say that X has a Multinomial (n, p) distribution. The pdf is

$$p(x) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \ldots p_k^{x_k}.$$

Exponential. $X \sim \exp(\beta)$ if $p_X(x) = \frac{1}{\beta}e^{-x/\beta}$, $x > 0$. Note that $\exp(\beta) = \Gamma(1, \beta)$.

Gamma. $X \sim \Gamma(\alpha, \beta)$ if

$$p_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$$

for $x > 0$ where $\Gamma(\alpha) = \int_0^\infty \frac{1}{\beta^\alpha}x^{\alpha-1}e^{-x/\beta}dx$.

Remark: In all of the above, make sure you understand the distinction between random variables and parameters.

More on the Multivariate Normal. Let $Y \in \mathbb{R}^d$. Then $Y \sim N(\mu, \Sigma)$ if

$$p(y) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right).$$

Then $\mathbb{E}(Y) = \mu$ and $\mathsf{cov}(Y) = \Sigma$. The moment generating function is

$$M(t) = \exp\left(\mu^T t + \frac{t^T\Sigma t}{2}\right).$$

Theorem 6 *(a). If $Y \sim N(\mu, \Sigma)$, then $E(Y) = \mu$, $cov(Y) = \Sigma$.*

(b). If $Y \sim N(\mu, \Sigma)$ and c is a scalar, then $cY \sim N(c\mu, c^2\Sigma)$.

(c). Let $Y \sim N(\mu, \Sigma)$. If A is $p \times n$ and b is $p \times 1$, then $AY + b \sim N(A\mu + b, A\Sigma A^T)$.

Theorem 7 *Suppose that $Y \sim N(\mu, \Sigma)$. Let*

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11}\, \Sigma_{12} \\ \Sigma_{21}\, \Sigma_{22} \end{pmatrix}.$$

where Y_1 and μ_1 are $p \times 1$, and Σ_{11} is $p \times p$.

(a) $Y_1 \sim N_p(\mu_1, \Sigma_{11}), Y_2 \sim N_{n-p}(\mu_2, \Sigma_{22})$.

(b) Y_1 and Y_2 are independent if and only if $\Sigma_{12} = 0$.

(c) If $\Sigma_{22} > 0$, then the condition distribution of Y_1 given Y_2 is

$$Y_1 | Y_2 \sim N_p(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}). \tag{1}$$

Lemma 8 *Let $Y \sim N(\mu, \sigma^2 I)$, where $Y^T = (Y_1, \ldots, Y_n), \mu^T = (\mu_1, \ldots, \mu_n)$ and $\sigma^2 > 0$ is a scalar. Then the Y_i are independent, $Y_i \sim N_1(\mu, \sigma^2)$ and*

$$\frac{\|Y\|^2}{\sigma^2} = \frac{Y^T Y}{\sigma^2} \sim \chi_n^2 \left(\frac{\mu^T \mu}{\sigma^2} \right).$$

Theorem 9 *Let $Y \sim N(\mu, \Sigma)$. Then:*
(a) $Y^T \Sigma^{-1} Y \sim \chi_n^2(\mu^T \Sigma^{-1} \mu)$.
(b) $(Y - \mu)^T \Sigma^{-1} (Y - \mu) \sim \chi_n^2(0)$.

6 Sample Mean and Variance

Let $X_1, \ldots, X_n \sim P$. The sample mean is

$$\overline{X}_n = \widehat{\mu}_n = \frac{1}{n} \sum_i X_i$$

and the sample variance is

$$S_n^2 = \widehat{\sigma}_n^2 = \frac{1}{n-1} \sum_i (X_i - \widehat{\mu}_n)^2.$$

Some authors instead define the sample variance as

$$\widehat{\sigma}_n^2 = \frac{1}{n} \sum_i (X_i - \widehat{\mu}_n)^2.$$

The *sampling distribution* of $\widehat{\mu}_n$ is

$$G_n(t) = \mathbb{P}(\widehat{\mu}_n \leq t).$$

Practice Problem. Let $X_1, \ldots, X_n$ be iid with $\mu = \mathbb{E}(X_i) = \mu$ and $\sigma^2 = \mathsf{Var}(X_i) = \sigma^2$. Then

$$\mathbb{E}(\widehat{\mu}_n) = \mu, \quad \mathsf{Var}(\widehat{\mu}_n) = \frac{\sigma^2}{n}, \quad \mathbb{E}(\widehat{\sigma}_n^2) = \sigma^2.$$

Theorem 10 *If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ then*

(a) $\widehat{\mu}_n \sim N(\mu, \frac{\sigma^2}{n})$.

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

(c) $\widehat{\mu}_n$ *and* $\widehat{\sigma}_n^2$ *are independent.*

7 A preview of the next few lectures

Let us consider a simple experiment. I toss a fair coin n times, and if the outcome is heads I record $X_i = +1$, and if the outcome is tails I record $X_i = -1$. These are called *Rademacher* random variables. Now, let us consider the average:

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

It is easy to see that $\mathbb{E}[\widehat{\mu}_n] = 0$, and we would like to know how far $\widehat{\mu}_n$ is from its expectation. When all the $X_i = +1$ (for instance), we have have that $\widehat{\mu}_n = 1$. There is however a remarkable phenomenon, known as the concentration of measure phenomenon, that asserts that $\widehat{\mu}_n$ "concentrates" much closer to $\mathbb{E}[\widehat{\mu}_n]$.

The average of n i.i.d random variables concentrates within an interval of length roughly $1/\sqrt{n}$ around the mean.

The basic intuition is that in order to for a sample average to be far from the expectation, many *independent* random variables need to work together simultaneously, which is extremely unlikely. Moreover, $\hat{\mu}_n$ has, approximately, a Normal distribution. These seemingly simple facts underly essentially all of statistics and machine learning.