Summary of Probability

**Foundations.** Probability space is $(\Omega, \mathcal{F}, \mathbb{P})$; $\Omega$ is the sample space (a set), $\mathcal{F}$ is the collection of events (sub-sets of $\Omega$), $\mathbb{P}$ is the probability measure: a countably additive function from $\mathcal{F}$ to $[0,1]$; $\mathbb{P}(\Omega) = 1$, $\mathbb{P}(\varnothing) = 0$.

Addition rule: $\mathbb{P}(A \bigcup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \bigcap B)$; extends to the inclusion-exclusion principle;
Multiplication rule: $\mathbb{P}(A \bigcap B) = \mathbb{P}(A|B)\mathbb{P}(B)$.

**Random variables.** A random variable $X$ is a (measurable) function from $\Omega$ to $\mathbb{R}$. The cdf of $X$ is $F_X(x) = \mathbb{P}(X \le x)$. A symmetric random variable has symmetric distribution: $F_X(x) = 1 - F_X(-x)$, $x \ge 0$. Discrete random variable takes values in a countable set $\{x_1, x_2, \ldots\}$ and is characterized by the probability mass function $p_X(k) = \mathbb{P}(X = x_k)$. A continuous random variable $X$ has continuous $F_X$ and takes any particular value with probability zero: $\mathbb{P}(X = a) = 0$ for all $a \in \mathbb{R}$. An absolutely continuous random variable $X$ is characterized by the pdf $f_X(x)$ so that $\mathbb{P}(a < X < b) = \int_a^b f(x)dx$.

**Main discrete distributions.**

(1) Uniform on $\{x_1, \ldots, x_n\}$: $p(k) = 1/n$
(2) Bernoulli($p$) or $\mathcal{B}(1,p)$: $p(1) = p,\ p(0) = 1-p$
(3) Binomial($n,p$) or $\mathcal{B}(n,p)$: $p(k) = [n!/k!(n-k)!]p^k(1-p)^{n-k}$
(4) Poisson($\lambda$) or $\mathcal{P}(\lambda)$: $p(k) = e^{-\lambda}\lambda^k/k!$
(5) Geometric($p$) or $G(p)$: $p(k) = p(1-p)^{k-1}$, $k = 1, 2, \ldots$
(6) Negative Binomial($m,p$) or $NB(m,p)$:
$p(k) = [(k-1)!/((m-1)!(k-m)!)]p^m(1-p)^{k-m},\ k = m, m+1, \ldots$
(7) Hypergeometric $\mathcal{H}(N,m,n)$ [$N$ is the total population, $m < N$ is the number of special objects in the population, $n < N$ is the sample size (without replacement)

**Main continuous distributions**.

(1) Uniform $U(a,b)$: $f(x) = 1/(b-a)$, $a < x < b$
(2) Normal $\mathcal{N}(\mu, \sigma^2)$
(3) Exponential $\mathcal{E}(\lambda) = \text{Gamma}(1, \lambda)$
(4) Gamma($\alpha, \lambda$): $f(x) = cx^{\alpha-1}e^{-\lambda x}$, $x > 0$
(5) Beta($\alpha, \beta$): $f(x) = cx^{\alpha-1}(1-x)^{\beta-1}$
(6) $\chi^2_n = \text{Gamma}(n/2, 1/2) = \sum_{k=1}^n X_k^2,\ X_k$ iid $\mathcal{N}(0,1)$
(7) $t_n = \mathcal{N}(0,1)/[\sqrt{\chi^2_n/n}]$
(8) Cauchy$= t_1 = X/Y$, $X, Y$ iid standard normal; pdf is $1/[\pi(1+x^2)]$
(9) $F_{m,n} = [\chi^2_m/m]/[\chi^2_n/n]$

**Characteristics of a random variable.**

The expected value of a discrete random variable is

$$\mu_X = \mathbb{E}(X) = \sum_k x_k p_X(k).$$

The expected value of a continuous random variable is

$$\mu_X = \mathbb{E}(X) = \int_{-\infty}^{+\infty} x f(x) dx.$$

Then $\text{Var}(X) = \sigma_X^2 = \mathbb{E}(X^2) - \mu_X^2$ is the variance of $X$, $[\mathbb{E}(X - \mu_X)^3]/\sigma_X^3$ is skewness, $[\mathbb{E}(X - \mu_X)^4]/\sigma_X^4$ is kurtosis; $\varphi_X(t) = \mathbb{E}e^{itX}$ is the characteristic function of $X$ and $M_X(t) = \mathbb{E}e^{tX}$ is the moment-generating function of $X$; for $n = 1, 2, 3, \ldots,\ \mathbb{E}(X^n)$ is called moment of order $n$ (or $n$-th moment); $\mathbb{E}(X - \mu_X)^n$ is called central moment of order $n$. Note that $\varphi_X$ always exists, even when $\mu_X$ and higher-order moments do not exist.

The median of a random variable $X$ is a point $\mathfrak{m}$ such that $\mathbb{P}(X \ge \mathfrak{m}) \ge 1/2$ and $\mathbb{P}(X \le \mathfrak{m}) \ge 1/2$; this point does not have to be unique, but if it is unique (which is often, but not always, the case for continuous distributions), then $\mathbb{P}(X \ge \mathfrak{m}) = \mathbb{P}(X \le \mathfrak{m}) = 1/2$.

For two random variables $X, Y$, $\text{Cov}(X,Y) = \mathbb{E}(XY) - \mu_X\mu_Y$ is covariance and $\text{Cor}(X,Y) = \rho_{X,Y} = \text{Cov}(X,Y)/[\sigma_X\sigma_Y]$ is the correlation coefficient. If $X$ and $Y$ are independent, then $\text{Cov}(X,Y) =$

0, but not, in general, the other ways around [uncorrelated random variables can be dependent]. If $(X, Y)$ is bi-variate normal, then zero correlation implies independence.

Standartization of a random variable $X$ is $Z = \frac{X-\mu_X}{\sigma_X}$, so that $\mathbb{E}(Z) = 0$, $\mathrm{Var}(Z) = 1$, and $Z$ is dimensionless (has no units even when $X$ has).

**Location parameter** $\mu \in \mathbb{R}$ and **scale parameter** $\sigma > 0$ are alternative ways to quantify standartization: if $Z$ is a random variable with pdf $f_Z = f_Z(x)$ and $X = \mu + \sigma Z$, then the pdf $f_X$ of $X$ satisfies

$$f_X(x) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right).$$

The location and scale parameters are especially useful when $\mu_X$ and/or $\sigma_X$ are not defined.

**More on skewness and kurtosis.** Symmetric random variable has zero skewness; skewness is positive when the distribution has a longer tail to the right (skewed to right); skewness is negative when the distribution has a longer tail to the left (skewed to left).

For a unimodal random variable, the distribution function (pdf or probability mass function) has a unique point of maximum, called the mode. Then

- Symmetric distribution has zero skewness and mode=median=mean;
- Skewed-to-right distribution has positive skewness and mode<median<mean;
- Skewed-to-left distribution has negative skewness and mode>median>mean.

Kurtosis is at least 1; it is exactly one for the symmetric Bernoulli distribution (fair coin tossing). Standard normal distribution has kurtosis equal to 3. Distributions with kurtosis equal to 3 are called mesokurtic. Platykurtic distribution has kurtosis<3 and, as a consequence, wider peak and thinner (lighter) tails than the normal distribution. Leptokurtic distribution has kurtosis>3 and, as a consequence, narrower peak and fatter (heavier) tails than the normal distribution.

**Sums of iid random variables.** $X_1, \ldots, X_n$ are iid as $X$ means

$$\mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n) = \prod_{k=1}^{n} F_X(x_k)$$

for all real numbers $x_1, \ldots, x_n$; $S_n = X_1 + \ldots + X_n$; $\bar{X}_n = S_n/n$ is the sample mean. Here are the main examples when the distribution of $S_n$ can be explicitly related to the distribution of $X$:

- If $X$ is Bernoulli($p$), then $S_n$ is Binomial($n, p$).
- If $X$ is Geometric($p$), then $S_n$ is Negative Binomial($n, p$).
- If $X$ is Poisson($\lambda$), then $S_n$ is Poisson($n\lambda$).
- If $X$ is Normal($\mu, \sigma^2$), then $S_n$ is Normal($n\mu, n\sigma^2$).
  Moreover, in this case $[1/\sigma^2]\sum_{k=1}^{n}(X_k - \bar{X}_n)^2$ is independent of $\bar{X}_n$ and has $\chi^2_{n-1}$ distribution.
- If $X$ is Gamma($\alpha, \lambda$), then $S_n$ is Gamma($n\alpha, \lambda$).
- If $X$ is Cauchy, then so is $S_n/n$.

**Limit Theorems.** If $\sigma_X$ exists, then $\mathrm{Var}(\bar{X}_n) = \sigma_X^2/n$, and we have

- LLN (Law of Large Numbers):

  $$\lim_{n\to\infty} \bar{X}_n = \mu_X$$

  (in probability, with probability one, and in $L_1$);
- CLT (Central Limit Theorem):

  $$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sqrt{n}(\bar{X}_n - \mu_X)}{\sigma_X} \le x\right) = \lim_{n\to\infty} \mathbb{P}\left(\frac{(S_n - n\mu_X)}{\sqrt{n}\sigma_X} \le x\right) = \Phi(x),$$

  where $\Phi$ is the cdf of the standard normal distribution;
- LIL (Law of Iterated Logarithm):

  $$\mathbb{P}\left(\limsup_{n\to\infty} \frac{\sqrt{n}(\bar{X}_n - \mu_X)}{\sigma_X\sqrt{2\ln(\ln(n))}} = 1\right) = \mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n - n\mu_X}{\sqrt{n}\sigma_X\sqrt{2\ln(\ln(n))}} = 1\right) = 1,$$

  where $\limsup_{n\to\infty} a_n = \lim_{n\to\infty} \sup_{k\ge n} a_k$.

- In the special case $X$ is Bernoulli: $\mathbb{P}(X=1)=p$, $\mathbb{P}(X=0)=1-p$, so that $\mu_X=p$, $\sigma_X^2=p(1-p)$, and $S_n$ is Binomial$(n,p)$, we also have the Local Limit Theorem (LLT), which gives an approximation by the normal pdf as opposed to cdf:
$$\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}\approx\frac{1}{\sqrt{2\pi}\sqrt{np(1-p)}}\exp\left(-\frac{(k-np)^2}{2np(1-p)}\right);$$
this is equivalent to the Stirling formula $n!\approx\sqrt{2\pi n}(n/e)^n$.
- Other special limit theorems: $\mathcal{H}(N,m,n)\approx\mathcal{B}(n,m/N)$ for large $N,m$ and fixed $n$; $\mathcal{B}(n,p)\approx\mathcal{P}(np)$ for large $n$ and small $p$; $\mathcal{P}(\lambda)\approx\mathcal{N}(\lambda,\lambda)$ for large $\lambda$.

**Conditional expectation.** Assuming $\sigma_X$ is finite, $g(Y)=\mathbb{E}(X|Y)$ is the function of $Y$ such that $g(Y)$ minimizes $\mathbb{E}(X-f(Y))^2$ among all possible functions $f$; $\mathbb{E}g(Y)=\mathbb{E}(X)=\mu_X$, $\mathbb{E}\big((f(Y)X|Y)=f(Y)\mathbb{E}(X|Y)$. If $X$ and $Y$ are independent, then $\mathbb{E}(X|Y)=\mu_X$. Conditional variance is
$$\mathrm{Var}(X|Y)=\mathbb{E}\Big(\big(X-\mathbb{E}(X|Y)\big)^2|Y\Big);\quad \mathrm{Var}(X)=\mathrm{Var}\big(\mathbb{E}(X|Y)\big)+\mathbb{E}\mathrm{Var}(X|Y).$$

**Bi-variate Gaussian distribution.** $[(X,Y)$ is bi-variate Gaussian$]\leftrightarrow[aX+bY$ is Gaussian for every real $a,b]$. In this case, $\mathbb{E}(X|Y)=\mu_X+\rho_{X,Y}(\sigma_X/\sigma_Y)(Y-\mu_Y)$ (normal correlation theorem).

**Inequalities.**

(1) Markov: if $X\ge 0$, then $\mathbb{P}(X>c)\le[\mathbb{E}X]/c$.
(2) Chebyshev: $\mathbb{P}(|X-\mu_X|>c)\le[\mathrm{Var}(X)]/c^2$.
(3) Jensen: if $f=f(x)$ is convex ($f''(x)\ge 0$), then $\mathbb{E}f(X)\ge f(\mu_X)$. For example, $M_X(t)\ge e^{t\mu_X}$.
(4) Cauchy-Schwartz: $|\mathbb{E}(XY)|\le\sqrt{\mathbb{E}X^2}\sqrt{\mathbb{E}Y^2}$.
(5) Hölder: $|\mathbb{E}(XY)|\le\big(\mathbb{E}|X|^p\big)^{1/p}\big(\mathbb{E}|Y|^q\big)^{1/q}$, $(1/p)+(1/q)=1$.
(6) Lyapunov: if $p>q$, then $\big(\mathbb{E}|X|^p\big)^{1/p}\ge\big(\mathbb{E}|X|^q\big)^{1/q}$.

**Types of convergence for random variables.** The sequence of random variables $\xi_n, n\ge 1$, converges to random variable $\xi$, as $n\to\infty$

(1) With probability one, if $\mathbb{P}(\lim_{n\to\infty}\xi_n=\xi)=1$;
(2) In probability, if, for every $\varepsilon>0$, $\lim_{n\to\infty}\mathbb{P}(|\xi_n-\xi|>\varepsilon)=0$;
(3) In distribution, if $\lim_{n\to\infty}F_{\xi_n}(x)=F_\xi(x)$ for all $x$ where $F_\xi$ is continuous. Equivalently, $\lim_{n\to\infty}\varphi_{\xi_n}(t)=\varphi_\xi(t)$ for all $t$.
(4) In $L_p$, $p>0$, if $\lim_{n\to\infty}\mathbb{E}|\xi_n-\xi|^p=0$.

"With probability one" implies "In probability", and "In probability" implies "In distribution". "In $L_p$" implies "In probability". If the limit $\xi$ is constant (non-random), then "In distribution" implies "In probability".