Lecture 0. Reviews & Distributions

1. Notations and some supplements

Some notations may appear:

- Combination $C_n^r = \binom{n}{r} = \frac{n!}{(n-r)!r!}$; and Permutation $A_n^r = \frac{n!}{(n-r)!}$.
- Indicator function: For any set A, the indicator of A is defined as

$$\mathbb{1}(x \in A) = \begin{cases} 1 & \text{if } x \in A; \\ 0 & \text{if } x \notin A. \end{cases}$$

  or sometimes been shorten as $\mathbb{1}_A(x)$, $\mathbb{1}_A$ or $\mathbb{1}_{\{x \in A\}}$.
- Gamma function $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$, the domain is $\mathbb{R} \backslash \{\{0\} \cup \mathbb{Z}^-\}$ but here we only consider $x > 0$.
  (除去非负整数的实数域)
  (i) For $n \in \mathbb{N}$, then $\Gamma(n) = (n-1)!$.
  (ii) For $z > 0$, $\Gamma(z+1) = z\Gamma(z)$.
  (iii) For $0 < z < 1$, $\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin z\pi}$.
  (iv) Legendre Formula: For $z > 0$, $\Gamma(z)\Gamma(z + \frac{1}{2}) = \frac{\sqrt{\pi}}{2^{2z-1}}\Gamma(2z)$.
- Beta function $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt$, with $x > 0, y > 0$.
  (i) $B(x, y) = B(y, x)$.
  (ii) $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x + y)$.
- Stirling's Formula:

$$\boxed{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}.}$$

$n! \sim n^{n+\frac{1}{2}} e^{-n} \sqrt{2\pi}$

- The maximum and minimum between a and b are denoted as

$$a \vee b = \max\{a, b\}, \quad a \wedge b = \min\{a, b\}.$$

2. A few of the most common distributions

Bernoulli(p) (在一次实验中共取得 x次成功的概率)

pmf: $P(X = x \mid p) = p^x(1-p)^{1-x}; \quad x = 0, 1; \quad 0 \le p \le 1$

mean: $\mathrm{E}X = p$

variance: $\mathrm{Var}\, X = p(1-p)$ $M_X(t) = 1 - p + pe^t$, for all $t \in R$

notes: **1.** A Bernoulli trial (named after James Bernoulli) is an experiment with only two possible outcomes.

2. Bernoulli random variable $X = 1$ if "success" occurs and $X = 0$ if "failure" occurs where the probability of a "success" is p.

Binomial(n, p) (在n次实验中共取得x次成功的概率)

pmf: $P(X = x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}; \quad x = 0, 1, \ldots, n; \quad 0 \le p \le 1$

mean: $\mathrm{E}X = np$

variance: $\mathrm{Var}\, X = np(1-p)$

notes: **1.** A Binomial experiment consists of n independent identical Bernoulli trials, i.e., the experiment consists of a sequence of n trials, where n is fixed in advance of the experiment.

2. The trials are identical, and each trial can result in one of the same two possible outcomes, which we denote by success (S) and failure (F).

3. The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.

4. The probability of success is constant from trial to trial; we denote this probability by p.

5. $X = \sum_{i=1}^{n} Y_i$, where $Y_1, \cdots, Y_n$ are n identical, independent Bernoulli random variables. Hence, the sum of n identical, independent Bernoulli random variables has a Binomial distribution (n, p).

Discrete uniform

pmf: $P(X = x \mid N) = \frac{1}{N}; \quad x = 1, 2, \ldots, N; \quad N = 1, 2, \ldots$

mean: $\mathrm{E}X = \frac{N+1}{2}$

variance: $\mathrm{Var}\, X = \frac{(N+1)(N-1)}{12}$

Poisson(λ) (已知某段时间内随机事件平均发生λ次，求该时间段内发生x次的概率)

pmf: $P(X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}; \quad x = 0, 1, \ldots; \quad 0 \le \lambda < \infty$

mean: $\mathrm{E}X = \lambda$

variance: $\mathrm{Var}\, X = \lambda$

notes: **1.** A Poisson distribution is typically used to model the probability distribution of the number of occurrences (with λ being the intensity rate) per unit time or per unit area.

2. A basic assumption: for small time intervals, the probability of an event occurring is proportional to the length of waiting time.

3. It was shown in the Section 2.3 of Casella and Berger (2021) that Binomial pmf approximates Poisson pmf. Poisson pmf is also a limiting distribution of a negative binomial distribution.
4. A useful result: By Taylor series expansion, we have $e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$.

assumptions:

1. Let $X(\Delta)$ be the number of events that occur during an interval Δ.
2. The events are independent; this means if $\Delta_1, \cdots, \Delta_n$ are disjoint intervals then $X(\Delta_1), \cdots, X(\Delta_n)$ are independent.
3. The distribution of $X(\Delta)$ depends only on the length of Δ and not on the time of occurrence.
4. The probability that exactly one event occurs in a small interval of length Δt equals $\lambda \Delta t + o(\Delta t)$ where $o(\Delta t)/\Delta t \to 0$ when $\Delta t \to 0$.
5. The probability that two or more events occur in a small interval of length Δt is $o(\Delta t)$.
6. $X(\Delta)$ satisfies the Poisson distribution $(\lambda\Delta)$

Geometric(p) （在第 x 次实验时恰好取得第一次成功的概率）

pmf: $P(X = x \mid p) = p(1-p)^{x-1}; \quad x = 1, 2, \ldots; \quad 0 \leq p \leq 1$

mean: $\mathrm{E}X = \frac{1}{p}$

variance: $\operatorname{Var} X = \frac{1-p}{p^2}$

notes:

1. The experiment consists of a sequence of independent trials.
2. Each trial can result in either a success (S) or failure (F).
3. The probability of success, p, is constant from trial to trial.
4. The experiment continues (trials are performed) until the first success.
5. $Y = X - 1$ is negative binomial$(1, p)$.
6. The distribution is *memoryless*: $P(X > x \mid X > t) = P(X > s - t)$.

Negative binomial(r, p) （在取得第 r 次成功前恰好失败 x 次实验的概率）

pmf: $P(X = x \mid r, p) = \binom{r+x-1}{x} p^r (1-p)^x; \quad x = 0, 1, \ldots; \quad 0 \leq p \leq 1$

mean: $\mathrm{E}X = \frac{r(1-p)}{p}$

variance: $\operatorname{Var} X = \frac{r(1-p)}{p^2}$

notes:

1. The experiment consists of a sequence of independent trials.

2. Each trial can result in either a success (S) or failure (F).
3. The probability of success, p, is constant from trial to trial.
4. The experiment continues (trials are performed) until a total of r successes have been observed, where r is a specified positive integer.
5. In contrast to the binomial experiment where the number of trials is fixed and the number of successes is random, the negative binomial experiment has the number of successes fixed and the number of trials random.
6. Define X = total number of failures until the r th success. Then $Y = X + r$ is the total number of trials until the r th success and X has a negative binomial distribution. Then, for $y \geq r$,
$$P(Y = y) = P(X = y - r) = \binom{y-1}{r-1} p^r (1-p)^{y-r}.$$
7. We have $P(X = x \mid r, p) = \binom{r+x-1}{x} p^r (1-p)^x = (-1)^x \binom{-r}{x} p^r (1-p)^y, x = 0, 1, \ldots$, hence the name "negative binomial".
8. It is not easy to verify that $\sum_{x=r}^{\infty} \binom{x-1}{r-1} p^r (1-p)^{x-r} = 1$.

Hypergeometric (N个物品中有M个type I，K次抽取抽出 x个type I 的概率)

pmf: $P(X = x \mid N, M, K) = {}^{x}\frac{\binom{M}{\cancel{k}}\binom{N-M}{K-x}}{\binom{N}{K}}; \quad x = 0, 1, 2, \ldots, K;$
$M - (N - K) \leq x \leq M; \quad N, M, K \geq 0$

mean: $\mathrm{E}X = \frac{KM}{N}$

variance: $\mathrm{Var}\, X = \frac{KM}{N} \frac{(N-M)(N-K)}{N(N-1)}$

notes: If $K \ll M$ and N, the range $x = 0, 1, 2, \ldots, K$ will be appropriate.

assumptions:
1. The population or set to be sampled consists of N individuals, objects, or elements (or a finite population).
2. Each individual can be characterized as a success (S) or failure (F), and there are M successes in the population.
3. A sample of K individuals is selected without replacement in such a way that each subset of size N is equally likely to be chosen, i.e., obtain a random sample.
4. Number of success is the random variable.

Uniform(a, b)

pdf: $f(x \mid a, b) = \frac{1}{b-a}, \quad a \leq x \leq b$

mean: $EX = \frac{b+a}{2}$

variance: $\mathrm{Var}\, X = \frac{(b-a)^2}{12}$

cdf: $F(x) = \begin{cases} 0, & \text{for } x \leq a \\ \frac{x-a}{b-a}, & \text{for } a < x < b \\ 1, & \text{for } x \geq b \end{cases}$ (在区间(a,b)上完全随机地选择一点，选到不超过x的概率)

notes: If $a = 0$ and $b = 1$, this is a special case of the beta ($\alpha = \beta = 1$).

Exponential(β) $f(x|\lambda) = \lambda e^{-\lambda x}$

pdf: $f(x \mid \beta) = \frac{1}{\beta} e^{-x/\beta}, \quad 0 \le x < \infty, \quad \beta > 0$

mean: $\mathrm{E}X = \beta$

variance: $\mathrm{Var}\, X = \beta^2$

cdf: $F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{for } x > 0 \\ 0, & \text{for } x \le 0 \end{cases}$ （随机事件在单位时间内平均发生λ次，求该事件相邻两次发生的时间间隔不超过x的概率）

notes:

1. Special case of the gamma distribution.
2. Has the *memoryless* property, *i.e.*, for $s \ge t$,
$$P(X > s \mid X > t) = P(X > s - t).$$
3. If for a continuous nonnegative random variable X that holds $P(X > t + s \mid X > s) = P(X > t)$ for all $s, t \ge 0$, then X must have an exponential distribution.
4. Has many special cases: $Y = X^{1/\gamma}$ is *Weibull*, $Y = \sqrt{2X/\beta}$ is *Rayleigh*, $Y = \alpha - \gamma \log(X/\beta)$ is *Gumbel*.

Gamma(α, β)

pdf: $f(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 \le x < \infty, \quad \alpha, \beta > 0$

mean: $\mathrm{E}X = \alpha\beta$

variance: $\mathrm{Var}\, X = \alpha\beta^2$

cdf: （随机事件在单位时间内平均发生λ次，求该事件相邻α次发生的时间间隔不超过x的概率）

notes:

1. The *gamma function* is defined as: $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$.
2. $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \quad \alpha > 0$.
3. $\Gamma(n) = (n-1)!, \quad$ for any integer $n > 0$.
4. Some special cases are exponential ($\alpha = 1$) and chi squared ($\alpha = p/2\ \beta = 2$).
5. If $\alpha = \frac{3}{2}, Y = \sqrt{X/\beta}$ is *Maxwell*. $Y = 1/X$ has the *inverted gamma distribution*.
6. Can also be related to the Poisson.

Beta(α, β)

pdf: $f(x \mid \alpha, \beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 \le x \le 1, \quad \alpha > 0, \quad \beta > 0$

mean: $\mathrm{E}X = \frac{\alpha}{\alpha+\beta}$

variance: $\mathrm{Var}\, X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

notes:

1. The constant in the beta pdf can be defined in terms of gamma functions $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.
2. The support set of X is $[0, 1]$, thus the beta distribution is one of the few "named" common distributions that give

probability to a finite interval. It can be used to model the proportion.

3. The shape of the beta distribution depends on α and β.

Normal(μ, σ^2) $E|X|^n < \infty$, $\forall n$ (not heavy tailed in terms of moments)

pdf: $f\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/\left(2\sigma^2\right)}, \quad -\infty < x < \infty,$
$-\infty < \mu < \infty, \quad \sigma > 0$

mean: $\mathrm{E}X = \mu$

variance: $\operatorname{Var} X = \sigma^2$

notes:

1. Sometimes called the Gaussian distribution.
2. There are certain reasons for the importance of the normal distribution: it is analytically simple, it has a bell shape, and there is the Central Limit Theorem, which shows that the normal distribution can be used to approximate a large variety of distributions in large samples under some mild conditions. A large portion of statistical theory is built on the normal distribution.
3. μ : location parameter; σ^2 : scale parameter.
4. To prove that the normal density integrates to 1, we need to transform the integral into polar coordinates (see for instance, Casella and Berger (2021), pg.103.).
5. To compute probabilities associated with the normal distribution, we use the standard normal tables.
6. Normal distribution is often used to approximate other probability distributions including the discrete distributions sometimes needing a continuity correction to improve on the approximation. For instance, Binomial (n, p) may be approximated by a normal with $\mu = np$ and $\sigma^2 = np(1-p)$ for large enough n.
7. Poisson (λ) may be approximated by a normal with $\mu = \lambda$ and $\sigma^2 = \lambda$.

Chi squared(p)

pdf: $f(x \mid p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}; \quad 0 \leq x < \infty; \quad p = 1, 2, \ldots$

mean: $\mathrm{E}X = p$

variance: $\operatorname{Var} X = 2p$

notes: Special case of the gamma distribution.

Cauchy(θ, σ)

pdf: $f(x \mid \theta, \sigma) = \frac{1}{\pi\sigma} \frac{1}{1+\left(\frac{x-\theta}{\sigma}\right)^2}, \quad -\infty < x < \infty; \quad -\infty < \theta <$

$\infty, \quad \sigma > 0$

mean: does not exist

variance: does not exist

notes:
1. Special case of Student's t, when degrees of freedom $= 1$.
2. If X and Y are independent $\mathrm{n}(0,1)$, X/Y is Cauchy.
3. Cauchy has the same shape as normal density but with thicker tails.
4. Mgf does not exist so all moments do not exist.

Lognormal(μ, σ^2)

pdf: $f\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \frac{e^{-(\log x - \mu)^2/\left(2\sigma^2\right)}}{x}, \quad 0 \le x < \infty, \quad -\infty < \mu < \infty$
$\sigma > 0$

mean: $\mathrm{E}X = e^{\mu + \left(\sigma^2/2\right)}$

variance: $\operatorname{Var} X = e^{2\left(\mu+\sigma^2\right)} - e^{2\mu+\sigma^2}$

notes: If X is random variable such that $\log X \sim n\left(\mu, \sigma^2\right)$, then X has a lognormal distribution.

Double exponential(μ, σ)

pdf: $f(x \mid \mu, \sigma) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$

mean: $\mathrm{E}X = \mu$

variance: $\operatorname{Var} X = 2\sigma^2$

notes:
1. Also known as the Laplace distribution.
2. The double exponential distribution has fatter tails than the normal distribution but still remains all of its moments.
3. The double exponential distribution is not bell-shaped and has a peak at $x = \mu$. At the peak, f_X is not differentiable.

t

pdf: $f(x \mid \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(1+\left(\frac{x^2}{\nu}\right)\right)^{(\nu+1)/2}}, \quad -\infty < x < \infty, \quad \nu = 1, \ldots$

heavy tailed (in terms of moments):

$E|t_3|^{3+\delta} \begin{cases} < +\infty & \forall \delta < 0 \\ = +\infty & \forall \delta \geq 0 \end{cases}$

mean: $\mathrm{E}X = 0, \quad \nu > 1$

variance: $\operatorname{Var} X = \frac{\nu}{\nu-2}, \quad \nu > 2$

notes: Related to F $\left(F_{1,\nu} = t_\nu^2\right)$

F

pdf: $f\left(x \mid \nu_1, \nu_2\right) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{(\nu_1-2)/2}}{\left(1+\left(\frac{\nu_1}{\nu_2}\right)x\right)^{(\nu_1+\nu_2)/2}}$

$0 \le x < \infty; \quad \nu_1, \nu_2 = 1, \ldots$

mean: $\mathrm{E}X = \frac{\nu_2}{\nu_2-2}, \quad \nu_2 > 2$

variance: $\mathrm{Var}\, X = 2\left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)}, \quad \nu_2 > 4$

notes: Related to chi squared ($F_{\nu_1,\nu_2} = \left(\frac{\chi^2_{\nu_1}}{\nu_1}\right) / \left(\frac{\chi^2_{\nu_2}}{\nu_2}\right)$, where the χ^2 s are independent) and t $\left(F_{1,\nu} = t^2_\nu\right)$.

Logistic(μ, β)

pdf: $f(x \mid \mu, \beta) = \frac{1}{\beta} \frac{e^{-(x-\mu)/\beta}}{\left[1+e^{-(x-\mu)/\beta}\right]^2}, \quad -\infty < x < \infty,$

$-\infty < \mu < \infty, \quad \beta > 0$

mean: $\mathrm{E}X = \mu$

variance: $\mathrm{Var}\, X = \frac{\pi^2\beta^2}{3}$

notes: The cdf is given by $F(x \mid \mu, \beta) = \frac{1}{1+e^{-(x-\mu)/\beta}}$

extreme value distributions

Pareto(α, β)

pdf: $f(x \mid \alpha, \beta) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \quad a < x < \infty, \quad \alpha > 0, \quad \beta > 0$

mean: $\mathrm{E}X = \frac{\beta\alpha}{\beta-1}, \quad \beta > 1$

variance: $\mathrm{Var}\, X = \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \quad \beta > 2$

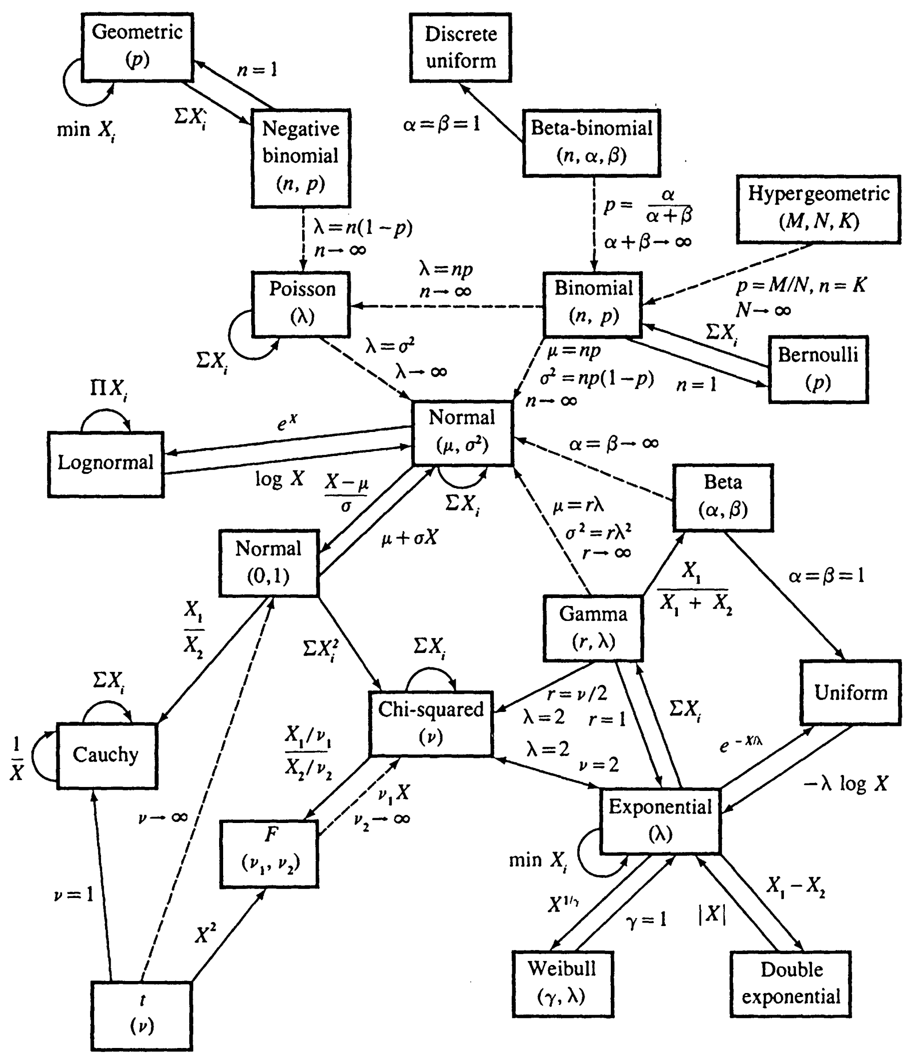
Weibull(γ, β)

pdf: $f(x \mid \gamma, \beta) = \frac{\gamma}{\beta} x^{\gamma-1} e^{-x^\gamma/\beta}, \quad 0 \le x < \infty, \quad \gamma > 0, \quad \beta > 0$

mean: $\mathrm{E}X = \beta^{1/\gamma}\Gamma\left(1 + \frac{1}{\gamma}\right)$

variance: $\mathrm{Var}\, X = \beta^{2/\gamma}\left[\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma^2\left(1+\frac{1}{\gamma}\right)\right]$

♣ Relationship among Common distributions



References

Casella, G., & Berger, R. L. (2021). Statistical inference. Cengage Learning.