Common Families of Distributions

The family of distributions: a class of pmfs/pdfs indexed by one or more *parameters*. For example, N(μ,1), Unif(a,b).

- Distributions in one family have a common pdf/pmf form but different parameter values.

For each distribution, we study its mean, variance, and other descriptive measures.

1 Discrete distributions

The range (sample space) of X is countable.

1. Discrete Uniform.
   X: possible values $1, 2, \ldots, N$. Here N is the parameter.

$$P(X = x|N) = 1/N, \quad x = 1, 2, \ldots, N.$$

$\mathrm{E}(X) = \frac{1+N}{2}$
$\mathrm{Var}(X) = \frac{(N+1)(N-1)}{12}$.

More generally, we say $X \sim$ Discrete Unif(N_0,N_1), if X takes all integer values $N_0, N_0 + 1, \ldots, N_1$ with equal probability,

$$P(X = x|N_0, N_1) = 1/(N_1 - N_0 + 1), \quad x = N_0, \ldots, N_1.$$

$\mathrm{E}(X) = \frac{N_0+N_1}{2.}$

2. Hypergeometric.

There is a large urn filled with N balls, M red and $N-M$ green.
Draw K balls at random *without replacement*.
$X =$ the number of red balls in the sampled K balls.
Assume $K \leq M$ and $K \leq (N-M)$, then

$$P(X = x|M, N, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, \quad x = 0, 1, \dots, K.$$

What are the possible value of x?
$\mathrm{E}(X) = \frac{KM}{N}$, $\mathrm{Var}(X) = K\frac{M}{N}(1-\frac{M}{N})\frac{N-K}{N-1}$

$\mathrm{Var}(X) \longrightarrow E(X^2) \longrightarrow E(X(X-1))$

Hint: $x\binom{M}{x} = M\binom{M-1}{x-1}, \quad x(x-1)\binom{M}{x} = M(M-1)\binom{M-2}{x-2}$

$$\binom{N}{K} = \frac{N}{K}\binom{N-1}{K-1}, \quad \binom{N}{K} = \frac{N(N-1)}{K(K-1)}\binom{N-2}{K-2}$$

Hypergeometric Applications:

- *Application 1:* Capture-Recapture method to estimate the total N.

  How many fish in a pond (birds in a forest, bears in a mountain)?

  N unknown.

  Step 1: Catch M fish. Mark them and return them.

  Step 2: Re-catch K fish.

  Let $X = \#$ tagged fish among K caught at the second time.

  We know M, K and also observe X. How to estimate N?

- *Application 2:* Acceptance Sampling.

  There are $N = 25$ machine parts in a lot

  Suppose $M = 6$ parts are defective.

  Randomly sample $K = 10$ parts.

  Let $X = \#$ of defective parts found in this sample.

  What is the probability of having no defective parts in the sample?

3. Binomial.

Repeat a random experiment n times, satisfying four conditions:
(i) Only two possible outcomes: Success (S) or Failure (F);
(ii) The probability of success p is same for each trial;
(iii) The experiments are independent with each other;
(iv) $X = \#$ the total number of successes in n trials.
We call these *Bernoullian trials*. Then $X \sim \mathrm{Bin}(n, p)$, with the pmf

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

$\mathrm{E}(X) = np, \quad \mathrm{Var}(X) = np(1-p), \quad M_X(t) = (1 - p + pe^t)^n.$

REMARK 1: The rth *factorial moment* of X is defined as

$$\mu_{[r]} = E(X(X-1)\cdots(X-r+1)), \quad \forall r \geq 1.$$

It is useful for the calculation of moments of discrete distributions. For example, $\mathrm{Var}(X) = E(X(X-1)) + E(X) - E(X)^2 = \mu_{[2]} + \mu_{[1]} - \mu_{[1]}^2$.

REMARK 2: Central moments have the following recurrence relationship

$$\mu_{r+1} = p(1-p)\left(nr\mu_{r-1} + \frac{d\mu_r}{dp}\right).$$

4. Poisson.

The random variable X takes non-negative integer values

$$P(X = x) = e^{-\lambda}\lambda^x / x!, \quad x = 0, 1, \ldots.$$

$\mathrm{E}(X) = \lambda, \quad \mathrm{Var}(X) = \lambda, \quad M_X(t) = e^{\lambda(e^t - 1)}.$

$\mathrm{Var}(X)$ — $E(X^2)$ — $E[X(X-1)]$.

- Poisson distribution is often used for describing the number of occurrences of a certain event in a very large number of observations, assuming that the probability for the event to occur in each observation is very small. Classic examples for Poisson applications include the nuclear decay of atoms, the mutation of DNA.
- Another important application of Poisson is the number of events occurring during a time/space interval of a given length/area, assuming that these events occur with a known average rate and independently in different intervals. Let λ be the rate (the expected number of "events" or "arrivals" that occur per unit), then the number of events occuring in time/space interval t can be modeled as Poisson($t\lambda$).
- Examples: the number of telephone calls arriving in a call center (bus station, bank) during a fixed time interval, the distribution of fish in a lake, the distribution of bombs hits in an area.

EXAMPLE 1. Consider a telephone operator, who, on the average, handles five calls every 3 minutes. What is the probability that there will be no calls in the next minute? At least two calls?

EXAMPLE 2. We can model the number of misprints on pages using Poisson. Assume that on average the number of misprints per page is 0.6.

(a) One page is chosen at random.

i. Compute P(No misprint).

ii. Compute P(At least 2 misprints).

(b) 5 pages are chosen randomly. What is the probability that at least 3 pages will have no misprints?

5. Geometric.

Bernoullian trials, Success or Failure. $P(S) = p$, $q = 1 - p$.

Continue until success. Let $X = \#$ trials needed to get the first S.

$$P(X = x) = pq^{x-1}, \quad x = 1, \ldots.$$

Check that the pmf adds to one.

EXAMPLE: (*Failure times*)

If the probability is 0.001 that a light bulb will fail on any given day, what is the probability that it will last at least 30 days?

Memoryless property: For integers $s > t$,

$$P(X > s | X > t) = P(X > s - t).$$

REMARK:

The geometric distribution "forgets" what has occurred. Prob.(getting additional $s - t$ failures, conditional on having t failures already) = Prob. (getting $s - t$ failures). This is also known as *ageless* or *lack of aging* property.

6. Negative Binomial. (NB)

In Bernoullian trials, continue up to r successes.

Let $X = \#$ trials needed to get r successes. Then $X \sim NB(r,p)$.

$$P(X = x|r,p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \ldots.$$

Alternative definition:

Let $Y = \#$ failures to get r successes. (i.e. $X = Y + r$ or $Y = X - r$)

$$P(Y = y|r,p) = \binom{r+y-1}{r-1} p^r (1-p)^y, \quad y = 0, 1, \ldots.$$

$\mathrm{E}(Y) = r\frac{1-p}{p}$ (*Hint*: $x\binom{n}{x} = (n-x+1)\binom{n}{x-1}$)

$\mathrm{Var}(Y) = r\frac{1-p}{p^2}$. (Find $E(Y(Y-1))$ first).

MGF.

$$M_Y(t) = E(e^{tY}) = \left(\frac{p}{1-(1-p)e^t}\right)^r, \quad \text{if} \quad t < -\log(1-p).$$

Question: Why is the distribution called *negative* binomial?

We can generalize binomial coefficient

$$\binom{a}{k} = \frac{a(a-1)\cdots(a-k+1)}{k!}, a \in \mathbb{R}.$$

by replacing a by $-\alpha$, $\alpha > 0$,

$$\binom{-\alpha}{k} = \frac{(-\alpha)(-\alpha-1)\cdots(-\alpha-k+1)}{k!}$$

$$= (-1)^k \frac{(\alpha+k-1)\cdots\alpha}{k!} = (-1)^k \binom{\alpha+k-1}{k}.$$

Thus

$$P(Y = y|r,p) = \binom{r+y-1}{y} p^r (1-p)^y = (-1)^y \binom{-r}{y} p^r (1-p)^y.$$

Remark 1: We have

$$\sum_{y=0}^{\infty} \binom{r+y-1}{y} (1-p)^y = p^{-r}.$$

$$\sum_{y=0}^{\infty} (-1)^y \binom{-r}{y} (1-p)^y = p^{-r}.$$

Remark 2: Negative binomial can be used to model phenomena in which we are waiting for an occurrence of a specified number of successes.

2 Relationships Between Different Discrete Distributions

(i) Hypergeometric-Binomial relationship

If X has the hypergeometric distribution with parameters N, M, K, where $M/N \to p$, $0 < p < 1$ as $N \to \infty$, and K fixed, then

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}} \to \binom{K}{x} p^x (1-p)^{K-x}.$$

(ii) Hypergeometric-Poisson relationship

Let X have the hypergeometric distribution with parameters N, M, K. If $N, M, K \to \infty$, $M/N \to 0$, and $KM/N \to \lambda$, then

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}} \to e^{-\lambda} \frac{\lambda^x}{x!}.$$

(iii) Convergence of Binomial to Poisson.

Let $X \sim$Bin(n, p). If $n \to \infty$, $p \to 0$, $np \to \lambda$ (i.e., n large, p small, but np is moderate), then Bin$(n, p) \to$ Poisson(λ).

Direct Proof:

We have seen the proof with MGF.

Illustration: Poisson approximates a good approximation to Binomial if n is large and p is small.

EXAMPLE: A typesetter on average makes one error in every 500 words. Each page contains 300 words. What is the chance that there will be no more than 2 errors in 5 pages?

Thumb of rule: $n \geq 20$ and $p \leq 0.05$.

(iv) Convergence of NB to Poisson.

If $r \to \infty$, $p \to 1$ so that $r(1-p) \to \lambda$, then negative binomial goes to Poisson (λ). And

$$\mathrm{E}(Y) = \frac{r(1-p)}{p} \to \lambda,$$

$$\mathrm{Var}(Y) = \frac{r(1-p)}{p^2} \to \lambda.$$

(v) Connection of NB with Geometric.

NB$(1, p)$ is the same as Geometric(p).

General r: Geometric in r stages, independent.

Let X_j = the number of trials needed to get jth success after getting $(j-1)$th success. Then $X_j \sim$ Geometric(p), for each j. Define

$$X = X_1 + \cdots + X_r.$$

Show $X \sim$NB(r,p).

3 Continuous distributions.

7. Continuous Uniform. We say $X \sim \text{Unif}(a, b)$ if it has the pdf

$$f(x|a, b) = \frac{1}{b - a}, \quad a < x < b.$$

$E(X) =$

$E(X^k) =$

$\text{Var}(X) =$

$E(e^{tX}) =$

8. Exponential Family. We say $X \sim \text{Exp}(\beta)$, where $\beta > 0$, if it has the pdf

$$f(x|\beta) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0.$$

CDF $F(x) =$

$M(t) =$

Memoryless:

$$P(X > s | X > t) = P(X > s - t), \quad s > t > 0.$$

9. Gamma Family

Gamma function. Consider the integral $\int_0^\infty e^{-x}x^{\alpha-1}dx$. It can be shown that for all $\alpha > 0$, the integral is finite. Define the gamma function

$$\Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}dx.$$

The gamma function has the following properties:

- $\Gamma(1) = \int_0^\infty e^{-x}dx = 1$.
- If $\alpha > 1$, then by integration by parts,

$$\Gamma(\alpha+1) = \left[x^\alpha(-e^{-x})\right]_0^\infty - \int_0^\infty (-e^{-x})\alpha x^{\alpha-1}dx = \alpha\Gamma(\alpha).$$

- In particular, for any positive integer $n > 1$,

$$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) = \cdots = (n-1)!.$$

- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

We say $X \sim \text{Gamma}(\alpha, \beta)$, where $\alpha, \beta > 0$, if it has the pdf

$$f(x|\alpha,\beta) = \frac{1}{\beta^\alpha\Gamma(\alpha)}e^{-x/\beta}x^{\alpha-1}, \quad x > 0.$$

α shape parameter, which most influences the peakedness of the distribution. β scale parameter, which influences the spread of the distribution.

Graphs

$\mathrm{E}(X) =$

$\mathrm{E}(X^k) =$

$\mathrm{Var}(X) =$

$M_X(t) =$

CDF. No closed form. Have to use tables of incomplete gamma integrals $\int_0^x e^{-t} t^{\alpha-1} dt$.

REMARK 1: β is the scalar parameter.
(i) $aX \sim \mathrm{gamma}(\alpha, a\beta)$.

(ii) X/β is gamma $(\alpha, 1)$ — standardized.

Remark 2: Two special cases of Gamma

(1) If $\alpha = 1$, Gamma$(1, \beta)$ is the same as Exp(β).

(2) If $\alpha = p/2$, $\beta = 2$, becomes chi-square with p degrees of freedom, denoted as χ^2_p.

$$f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, \quad 0 < x < \infty.$$

Example: If Z is $N(0, 1)$, then $Y = Z^2$ is distributed as χ^2_1 =Gamma$(\frac{1}{2}, 2)$.

Remark 3: **Gamma-Poisson relationship**:

If X is Gamma(α, β), where α is an integer, then for any x,

$$P(X \leq x|\alpha, \beta) = P(Y \geq \alpha), \quad \text{where} \quad Y \sim \text{Poisson}(x/\beta).$$

10. Weibull.
Define $Y = X^{1/\gamma}$, where $\gamma > 0$ and $X \sim$exp(β), then

$$f(y|\beta,\gamma) = \frac{\gamma}{\beta} y^{\gamma-1} e^{-y^\gamma/\beta}, \quad y > 0.$$

Shapes

CDF

$$P(Y \le y) = P(X^{1/\gamma} \le y) = P(X \le y^\gamma) = 1 - e^{-y^\gamma/\beta}$$

Moment: $E(Y^k) =$

REMARK: Useful for modeling *hazard* function.

11. Standard Normal and Normal DistributionGaussian.

(1) Z follows standard normal distribution if

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

This implies that $\int_{-\infty}^{\infty} e^{-z^2/2} dz = \sqrt{2\pi}$. (check and remember it!)

By symmetry, $E(Z) = 0$. More generally,
$M_Z(t) = e^{t^2/2}$
$E(Z^{2k+1}) = 0$.
$E(Z^{2k}) = (2k-1)!!$
Variance, skewness, kurtosis: $\text{Var}(Z) = 1, \gamma_1 = 0, \gamma_2 = 0$.

CDF

$$F_Z(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

$\Phi(0) = 1/2$, $\Phi(-\infty) = 0$, $\Phi(\infty) = 1$, $\Phi(-z) = 1 - \Phi(z)$.
$\Phi(-4) \approx 0$, $\Phi(4) \approx 1$.
Tables of $\Phi(t)$ are available.

(2) Let $X = \mu + \sigma Z$, then $X \sim N(\mu, \sigma^2)$ and its pdf is

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

$E(X) = \mu$, $\text{Var}(X) = \sigma^2$.
$M_X(t) = E(e^{tX}) =$

Central moments of X

CDF $F_X(x) = P(X \leq x) = P(Z \leq (x-\mu)/\sigma) = \Phi(\frac{x-\mu}{\sigma})$,

(3) If $Y = aX + b$, then $Y \sim N(a\mu + b, a^2\sigma^2)$ and

$$f_Y(y) = \frac{1}{|a\sigma|\sqrt{2\pi}} e^{-(y-a\mu-b)^2/2(a\sigma)^2}.$$

(4) **Standardization Technique**: Taking $a = 1/\sigma$, $b = -\mu/\sigma$, then

$$(X - \mu)/\sigma \sim N(0, 1)$$

(5) **Normal approximation to binomial.**

Assume $X \sim \text{bin}(n, p)$. If n is large (at least 30 or 40), p neither too small nor too large, then X is approximately $N(np, np(1-p))$.

Eg. 50 attempts, probability of success 0.3.

$P(X \leq 13) =$

Continuity correction:

$$P(a \leq X \leq b) \approx \Phi((b + 0.5 - \mu)/\sigma) - \Phi((a - 0.5 - \mu)/\sigma).$$

Remark 1: Most common distribution, most loved.

Remark 2: Used to approximate distribution of a variable that is the sum of independent variables— central limit theorem.

12. Beta.

$X \sim Beta(\alpha, \beta)$, where $\alpha, \beta > 0$, if $x \in (0, 1)$ and has the pdf

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1,$$

where $B(\alpha, \beta)$ is the so-called beta function

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = B(\beta, \alpha).$$

Shapes

Moments

Remark 1: The special case Beta$(1, 1)$ is the same as Unif$(0, 1)$.

Remark 2: If $X \sim$ beta(α, β), then $1 - X \sim$ beta (β, α).

13. Lognormal(μ, σ^2).
X is lognormal(μ, σ^2) if $\log X$ is distributed as $N(\mu, \sigma^2)$.
cdf is $\Phi((\log x - \mu)/\sigma)$.

pdf

Moments.

Shape: Very skew, model for income distribution

14. Cauchy.

$$f(x|\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}, \quad -\infty < x < \infty.$$

$\theta \in R$ is the location parameter.
CDF

$$F(x|\theta) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x-\theta).$$

Symmetric about θ, so median is θ.
No mean exists. $E(|X|) = \infty$. No higher moments.

15. Double exponential. Also known as Laplace distribution.

$$f(x|\mu, \sigma) = \frac{1}{2\sigma} e^{-|x-\theta|/\sigma}, \quad -\infty < x < \infty.$$

$\theta \in R$ is the location parameter, $\sigma > 0$ is the scale parameter.

Standard: $\mu = 0$, $\sigma = 1$.
CDF:

Moments:

MGF:

4 Exponential Family.

Many common families take the form

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta})\exp(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)). \tag{1}$$

EXAMPLE. Binomial.

EXAMPLE. Poisson.

EXAMPLE. Normal.

NON-EXPONENTIAL FAMILY EXAMPLE.

Let X has the pdf

$$f(x|\theta) = \frac{1}{\theta} e^{1-x/\theta}, \quad 0 < \theta < x < \infty.$$

4.1 Full/Curved exponential family

In the exponential family form (1),

if $d = k$, we say the family is a *full exponential family*;

Otherwise, if $d < k$, we say the family is a *curved exponential family.*

EXAMPLES:

$N(\mu, \sigma^2)$.

$N(\mu, \mu^2)$.

4.2 Expectation and Variance Calculation

Take log of the pdf, differentiate wrt θ_j and then integrate/sum wrt x. This leads to the expectation-variance identities.

THEOREM.

$$E\left(\sum_{i=1}^{k}\frac{\partial w_i(\boldsymbol{\theta})}{\partial\theta_j}t_i(X)\right)=-\frac{\partial}{\partial\theta_j}\log c(\boldsymbol{\theta}).$$

$$\mathrm{var}\left(\sum_{i=1}^{k}\frac{\partial w_i(\boldsymbol{\theta})}{\partial\theta_j}t_i(X)\right)=-\frac{\partial^2}{\partial^2\theta_j}\log c(\boldsymbol{\theta})-E\left(\sum_{i=1}^{k}\frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial^2\theta_j}t_i(X)\right).$$

Joint/marginal MGF of $(t_1(X),\ldots,t_k(X))$.

Application. Poisson.

5 Location-scale family.

THEOREM: Let $f(x)$ be any pdf and let μ and $\sigma > 0$ be any given constants, then the function

$$g(x|\mu, \sigma) = \frac{1}{\sigma} f((x - \mu)/\sigma)$$

is a pdf.

Proof:

DEFINITION: Let $f(x)$ be any pdf. Then

(i) the family of pdfs of $f(x - \mu)$, indexed by $-\infty < \mu < \infty$, is called the *location* family, and μ is called the *location* parameter of the family.

(ii) the family of pdfs of $\frac{1}{\sigma} f(x/\sigma)$, indexed by $\sigma > 0$, is called the *scale* family, and σ is called the *scale* parameter of the family.

(iii) the family of pdfs of $\frac{1}{\sigma} f((x - \mu)/\sigma)$, indexed by (μ, σ) with $-\infty < \mu < \infty$ and $\sigma > 0$, is called the *location-scale* family; μ is called the location parameter and σ is called the scale parameter.

EXAMPLE: Normal. $Z = (X - \mu)/\sigma \sim f$.

Probability calculations through Z.

Expectation variance through Z.

6 Probability Inequalities

Chebychev/Markov Inequality.

If g is non-negative, then

$$P(g(X) \geq r) \leq E(g(X))/r, \quad \forall r > 0.$$

Applications: Assume $E(X) = \mu$ and $Var(X) = \sigma^2$, then

$$P(|X - \mu| \geq t\sigma) \leq 1/t^2, \quad \forall t > 0.$$

Another Probability Inequality:

$$P(X \geq a) \leq e^{-at}M(t) \quad \text{for any} \quad t > 0.$$

Differentiation of an integral.

- **Leibnitz's rule.** Assume that $f(x, \theta)$ differentiable in θ and $a(\theta) < b(\theta)$ are differentiable. Then

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)dx = f(b(\theta),\theta)b'(\theta) - f(a(\theta),\theta)a'(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta} f(x,\theta)dx.$$

  To prove this, consider a function $F(u, v, \theta) = \int_u^v f(x, \theta)dx$ and apply chain rule to $F(a(\theta), b(\theta), \theta)$.

  In particular,

$$\frac{d}{d\theta}\int_a^b f(x,\theta)dx = \int_a^b \frac{\partial}{\partial\theta} f(x,\theta)dx.$$

  Thus one can change differentiation and integration whenever the limits are finite.

- If the integration range is infinite, can you change limit and integration?

  Not always!

  Dominated convergence theorem.

  If $|f(x, y)| \le g(x)$ for all $|y - y_0| < \delta$, where $g(x)$ is integrable (satisfying $\int_{-\infty}^{\infty} g(x)dx < \infty$), then

$$\lim_{y\to y_0}\int_{-\infty}^{\infty} f(x,y)dx = \int_{-\infty}^{\infty} \lim_{y\to y_0} f(x,y)dx.$$

  An analogous result holds for infinite sums.

  As a corollary, we can interchange differentiation and integration by the following result.

  If $f(x, \theta)$ is continuously differentiable near θ_0 and the derivative is bounded uniformly by an integrable function $g(x)$, then

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f(x,\theta)dx.$$