MATH 5010 –Foundations of Statistical Theory and Probability

❖ More Families of Distributions

Instructor: He Wang
Department of Mathematics
Northeastern University

❖ Outline:

- **More Discrete Distributions**
  - Discrete Uniform Distribution
  - Hypergeometric Distribution
  - Negative Binomial Distribution
- **More Continuous Distributions**

  Lognormal Distribution
  Double Exponential
- **Exponential Families**
- **Location and Scale Families**

Recall:

We have introduced some common families of distributions.

Discrete Distributions
$Bernoulli(p)$
$Binomial(n, p)$
$Geometric(p)$
$Poisson(\lambda)$

Continuous Distributions
$Exponential(\lambda)$
$Gamma(n, \lambda)$
$Uniform(a, b)$
$Normal(\mu, \sigma^2)$
$Beta(\alpha, \beta)$

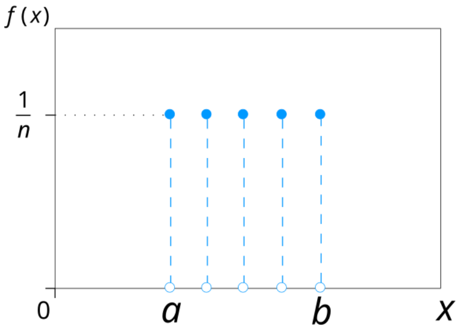
➢ Discrete Distributions:

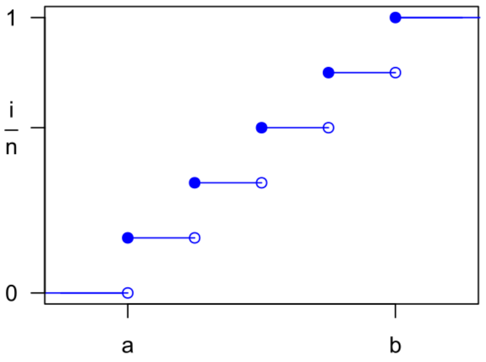
A random variable X is a discrete distribution if the range of X, the sample space, is countable. (E.g., integer-valued outcomes.)

- **Discrete Uniform Distribution**

A random variable X has a discrete uniform $[a, b]$ distribution if

$$P(X = x) = \frac{1}{n} \qquad\qquad n = b - a + 1$$





Example:
throwing a fair six-sided die.

Application: German tank problem in WWII (estimating the maximum $N\ in\ \mathrm{Uniform}[1, N]$)

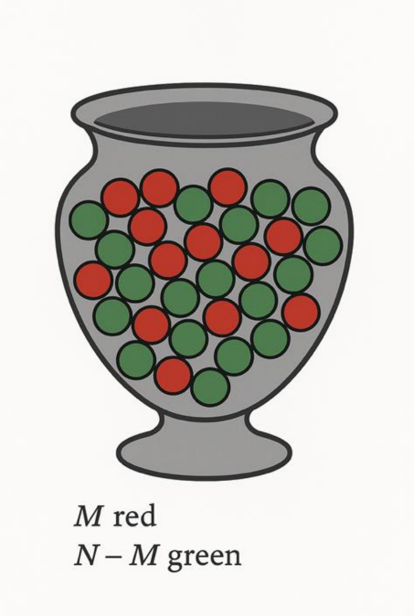
- **Hypergeometric Distribution**

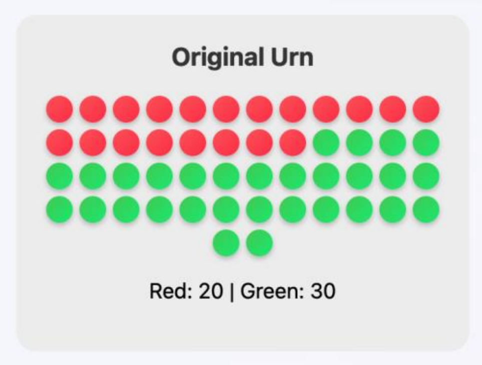
An urn filled with N balls, where M are red and $N - M$ are green.

Take K balls at random (without **replacement**).

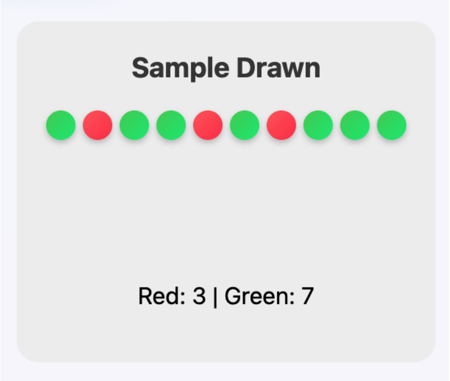
What is the probability that **exactly** x of the balls are red?

This is the hypergeometric distribution.
(What is the distribution if take balls with replacement?)




Original Urn

Red: 20 | Green: 30


Sample Drawn

Red: 3 | Green: 7

Total number of samples:

Number of ways with x red:

Number of ways with $K - x$ green:

Probability mass function of **Hypergeometric Distribution**

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

Verifications of the results for hypergeometric distribution involve combinatorics.

Expected value

$$E[X] = \frac{KM}{N}$$

Variance

$$\operatorname{Var} X = \frac{KM}{N}\left(\frac{(N-M)(N-K)}{N(N-1)}\right).$$

Application (Acceptance sampling)

Suppose a retailer buys machine parts in lots (N) and each item can be either acceptable ($N - M$) or defective M.

Then we can calculate the probability that a sample of size K contains x defectives.

Suppose in total $N = 25$ and select $K = 10$.

Oberved Event: All acceptable.

What is the probability of this event if there are 6 defectives in the lot of 25?

$$P(X = 0) = \frac{\binom{6}{0}\binom{19}{10}}{\binom{25}{10}} = .028,$$

Observed event is quite unlikely if there are six (or more) defectives in the lot.

- **Negative Binomial Distribution**

A sequence of independent Bernoulli trials with two outcomes (T or F) with constant probability p and $1 - p$.

We observe this sequence until a fixed number r of successes occurs.

Random Variable Y denotes number the failures follows the **Negative Binomial Distribution with pmf:**

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

Mean: $E[Y] = \frac{r(1-p)}{p}$

Variance $Var(Y) = \frac{r(1-p)}{p^2}$

The limit of negative binomial distributions is the Poisson distribution:

$$NBinomial(r,p) \to Possion(\lambda)$$

when $r \to \infty$ and $p \to 1, such\ that\ r(1-p) \to \lambda$

Another version of the Negative Binomial Distribution is defined by random variable X: the total number of trails.

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

Application:

- **Modeling the Length of hospital stay**

- **Inverse binomial sampling in sampling biological populations**

If the proportion of individuals possessing a certain characteristic is p and we **sample** until we see r such individuals, then the number of individuals sampled is a negative binomial random variable. We use this to determine how many **samples** we are likely to look at.

- **Survey**

Jim is required survey houses. Jim is not supposed to return home until 30 survey form are filled. He goes door to door, asking people to fill the survey. At each house, there is a 0.6 probability of complete a survey.

What's the probability of filling the last survey at the n-th *house?*

Continuous Distributions:

- **Lognormal Distribution**

Let X be a random variable whose logarithm is normally distributed

$$\log X \sim Normal(\mu, \sigma^2)$$

Then X has a lognormal distribution. By transformation theorem, we have pdf

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

$$0 < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0,$$

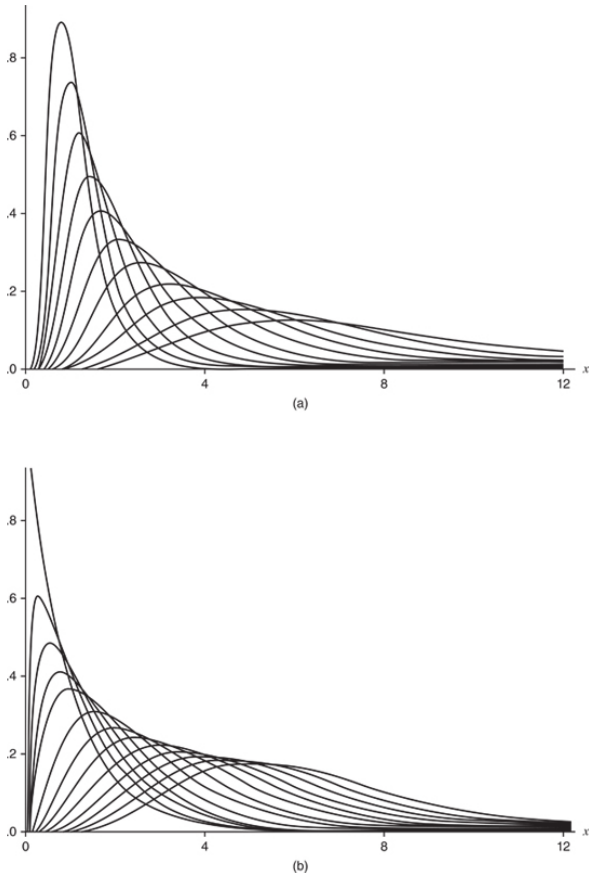
Mean: $E[X] = e^{\mu+(\sigma^2/2)}$

Variance: $Var(X) = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$

Comparing Gamma and Log-Normal Distributions

The lognormal distribution is similar in appearance to the gamma distribution.

The distribution is very popular in modeling applications, when the variable of interest is skewed to the right. (e.g., incomes, length of comments posted in internet, and more)



https://demonstrations.wolfram.com/ComparingGammaAndLogNormalDistributions/

Double Exponential Distribution ((Laplace distribution)

The **double exponential** distribution is formed by reflecting the exponential distribution around its mean.

The pdf is given by

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

for $-\infty < x < \infty, -\infty < \mu < \infty,\ \sigma > 0.$

Mean: $E[X] = \mu$

Variance: $Var(X) = 2\sigma^2$

Exponential Family

Exponential family comprises a set of flexible distribution ranging both continuous and discrete random variables. The members of this family have many important properties which merits discussing them in some general format. Most of the commonly used statistical distributions are members of the **exponential family** of distributions.

- ***Gaussian:*** $\mathbb{R}^{p}$
- ***Bernoulli: binary*** $\{\mathbf{0}, \mathbf{1}\}$
- ***Binomial: counts of success/failure***
- ***Multinomial: categorical***
- ***Poisson:*** $\mathbb{N}^{+}$
- ***Exponential:*** $\mathbb{R}^{+}$
- ***Gamma:*** $\mathbb{R}^{+}$
- ***Laplace:*** $\mathbb{R}^{+}$
- ***Beta:*** $(\mathbf{0}, \mathbf{1})$
- *Von mises: sphere*
- *Dirichlet:* Δ *(Simplex)*
- *Weibull:* $\mathbb{R}^{+}$
- *Weishart: symmetric positive-definite matrices*

A number of common distributions are exponential families, but **only** when certain parameters are fixed and known. For example:

- Binomial (with fixed number of trials)
- Multinomial (with fixed number of trials)
- Negative binomial (with fixed number of failures)

Examples of common distributions that are **not** exponential families

- **Student's t**,
- most **mixture distributions**,
- and even the family of **uniform distributions** when the bounds are not fixed.

Check Wikipedia for each of the distributions.

https://en.wikipedia.org/wiki/Exponential_family

Definition(Exponential Family):

A pdf/pmf of a distribution in d-parameters **exponential family** densities is in the form

$$p(\vec{y};\vec{\eta}) = \frac{1}{Z(\vec{\eta})} h(\vec{y}) \exp[\vec{\eta}^T T(\vec{y})]$$

$$= h(\vec{y}) \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})]$$

Here,

- $\vec{\eta} \in \mathbb{R}^d$ is the **natural parameter** of the distribution.

- $T(\vec{y}) \in \mathbb{R}^d$ is a vector of **sufficient statistics**. In many cases, $T(\vec{y}) = \vec{y}$, then the distribution is said to be in canonical form.

- $h(\vec{y})$ is the is the "**underlying/base measure**", in many cases, $h(\vec{y}) = 1$.

- $A(\vec{\eta}) = \log \mathrm{Z}(\vec{\eta})$ is called the **log partition function/log normalizer.** $A(\vec{\eta})$ is the normalization constant, to make sure the total probability is 1.

Hence,

$$A(\vec{\eta}) := \log \int h(y) \exp\left(\vec{\eta}^T T(\vec{y})\right) dy$$

Other formats of pdf/pmf of exponential family:

$$p(\vec{y}; \vec{\eta}) = h(\vec{y}) \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})]$$

$$= \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta}) + C(\vec{y})]$$ where $C(\vec{y}) = \log h(\vec{y})$

Sometimes, for GLM construction, we also introduce an extra scale parameter ϕ, called the **dispersion** parameter, to control the shape of $p(\vec{y}\,)$

$$p(\vec{y}; \vec{\eta}, \phi) = \exp\left[\frac{\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})}{\phi} + C(\vec{y}, \phi)\right]$$

This format is better for constructing generalized linear models.

In general, the parameter $\vec{\eta}$ is not the mean of the distribution. We can view $\vec{\eta}$ as a function of the mean $\vec{\mu} = E(\vec{y})$ and write $\vec{\eta} = g(\vec{\mu})$, which is called the **link function.** The inverse $\vec{\mu} = g^{-1}(\vec{\eta})$ is called the **response function**.

A fixed choice of A, h and ϕ defines a family (or set) of distributions that is parameterized by $\vec{\eta}$; as we vary $\vec{\eta}$, we then get different distributions within this family.

An even more general form is

$$p(\vec{y}; \vec{\eta}) = h(\vec{y}) \exp\left[[f(\vec{\eta})]^T T(\vec{y}) - A\big(f(\vec{\eta})\big)\right]$$

Example: (Bernoulli)

The pdf function of Bernoulli distribution:

$$p(y;\mu) = Bern(y;\mu) = \mu^y(1-\mu)^{1-y} \text{ for } y \in \{0,1\}$$

We can write it as

$$p(y;\mu) = \exp(y\log(\mu) + (1-y)\log(1-\mu))$$

$$= \exp\left(y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)\right)$$

Compare to $p(y;\eta) = h(\vec{y})\exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})]$, we have

$$T(y) = y \text{ and } h(\vec{y}) = 1$$

Canonical Link function: $\eta = \log\left(\dfrac{\mu}{1-\mu}\right) \implies \mu = \dfrac{1}{1+e^{-\eta}}$

$$A(\vec{\eta}) = -\log(1-\mu) = \log(1+e^{\eta})$$

Example: (Categorical/Multinomial)

Categorical has a vector of parameters ϕ_k where k goes from 1 to K.

$$p(y;\vec{\phi}) = \phi_1^{\mathbb{I}(\mathrm{y=1})}\phi_2^{\mathbb{I}(\mathrm{y=2})}\ldots\phi_K^{\mathbb{I}(\mathrm{y=K})} = \phi_1^{\mathbb{I}(\mathrm{y=1})}\phi_2^{\mathbb{I}(\mathrm{y=2})}\ldots\phi_K^{\mathbb{I}(\mathrm{y=K})}$$

$$= \exp\left(\sum_{i=1}^{K-1}\mathbb{I}(\mathrm{y=i})\log(\phi_i) + \left(1-\sum_{i=1}^{K-1}\big(\mathbb{I}(\mathrm{y=i})\big)\right)\log(\phi_k)\right)$$

$$= \exp\left(\sum_{i=1}^{K-1}\mathbb{I}(\mathrm{y=i})\log(\phi_i/\phi_k) \;+\log(\phi_k)\right)$$

To express the multinomial as an exponential family distribution, define

$T(i) = \vec{e}_i \in \mathbb{R}^{k-1}$ and $T(k) = \vec{0}$ So, $\mathbb{I}(\mathrm{y=i}) = \mathrm{T(y)_i}$

$p(y;\vec{\phi}) = h(\vec{y})\exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})]$, where

$h(\vec{y}) = 1; A(\vec{\eta}) = -\log(\phi_K)$

$$\vec{\eta} = \begin{bmatrix}\log(\phi_1/\phi_k)\\ \log(\phi_2/\phi_k)\\ \vdots \\ \log(\phi_{k-1}/\phi_k)\end{bmatrix}$$

Example: (Binomial)

The **binomial distribution ($\boldsymbol{Bi(p,n)}$)** is frequently used to model the number of successes in a sample of size n independent sequences yes-no experiments. The pmf is

$$p(x; n, \mu) = \binom{n}{x} \mu^x (1-\mu)^{n-x}$$

$$= \exp\left(x \log \frac{\phi}{1-\phi} + n \log(1-\phi) - \log \binom{n}{x} \right)$$

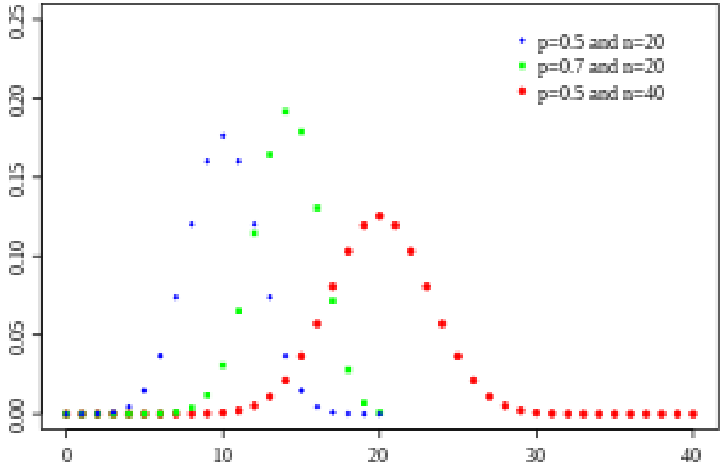
Thus:

$$= \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta}) + C(\vec{y})]$$

Here, $T(\vec{y}) = \mathrm{x}$ $\qquad \vec{\eta} = \log \frac{\phi}{1-\phi}$

$$A(\vec{\eta}) = -n \log(1-\phi) = n \log\left(1 + e^{\phi}\right)$$

$$C(\vec{y}) = -\log \binom{n}{x}$$



Example: (Normal)

The pdf function for normal distributation $Normal(\mu, \sigma^2)$ is

$$p(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{1}{2\sigma^2}\mu^2\right)$$

Compare to $p(y;\eta) = \ h(\vec{y}) \exp[\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})]$, we have

(1) If we treat both (μ, σ^2) as **two parameters**, we need to define

$$h(y) = \sqrt{2\pi}\,\sigma$$

$$\vec{\eta} = \begin{bmatrix} \eta_1 \\ \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ \\ -\frac{1}{2\sigma^2} \end{bmatrix} \qquad T(y) = \begin{bmatrix} y \\ \\ y^2 \end{bmatrix} \qquad A(\vec{\eta}) = (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right)$$

Example: (Normal with known σ^2)

(2) When σ^2 is known (treat as constant), denote $\vec{\theta} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$ it becomes a **one-parameter** exponential family on

$\eta = \frac{\mu}{\sigma^2}$, so $\mu = \sigma^2 \eta$ $\qquad$ $T(y) = y$

$$A(\eta) = \frac{1}{2\sigma^2}\mu^2 = \frac{\sigma^2\eta^2}{2} \qquad h(y) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

In format of $p(y;\eta,\phi) = \exp\left[\frac{\vec{\eta}^T T(\vec{y}) - A(\vec{\eta})}{\phi} + C(\vec{y},\phi)\right]$

$\phi = \sigma^2$, and $\eta = \mu$, and $A(\eta) = \frac{\eta^2}{2}$ $\qquad$ $C(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$

Canonical Link function is identity.

Example: (Poisson)

Poisson is a discrete distribution defined to express the **number events** that occur in a **unit** of time or space. This distribution, which is similar to Gaussian distribution but for count data, is given by

$$p(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = \frac{1}{y!}\exp(y \log \lambda - \lambda)$$

$$E(Y) = \lambda, \mathrm{Var(Y)} = \lambda$$

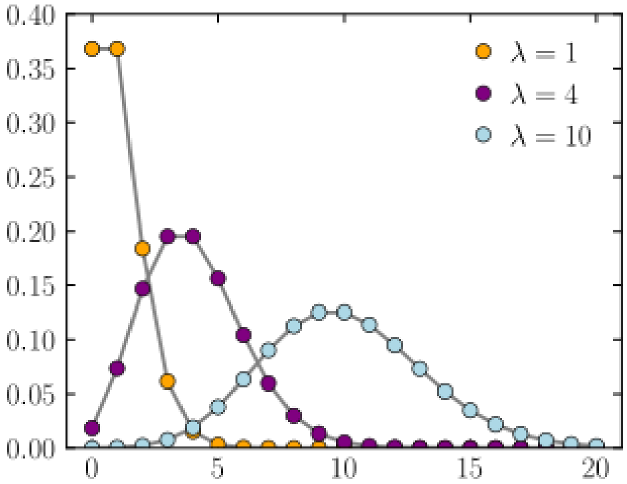
Compare to exponential family,

$$\eta = \log \lambda$$

$$T(y) = y$$

$$h(y) = \frac{1}{y!}$$

$$A(\eta) = \lambda = e^{\eta}$$



Example: (Exponential Distribution)

The exponential distribution is a distribution that models the **independent arrival** time. Its distribution (the probability density function, pdf) is given as

$$P(y;\lambda) = \lambda e^{-\lambda y} \text{ for y} \geq 0$$

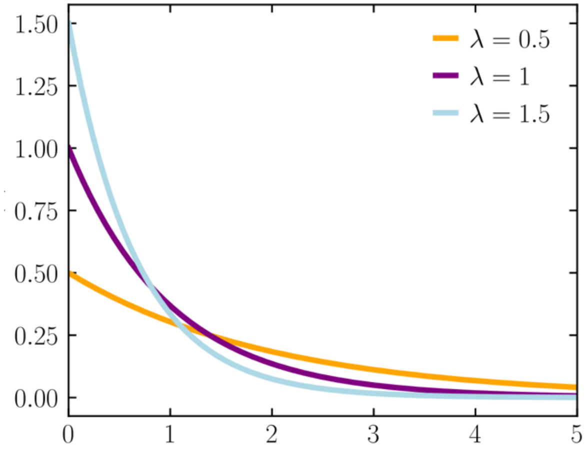
$$E(Y) = \frac{1}{\lambda}, \mathrm{Var(Y)} = \frac{1}{\lambda^2}$$

Compare to exponential family,

$$\eta = \lambda$$

$$T(y) = -y$$

$$h(y) = \mathbb{I}(y \geq 0)$$

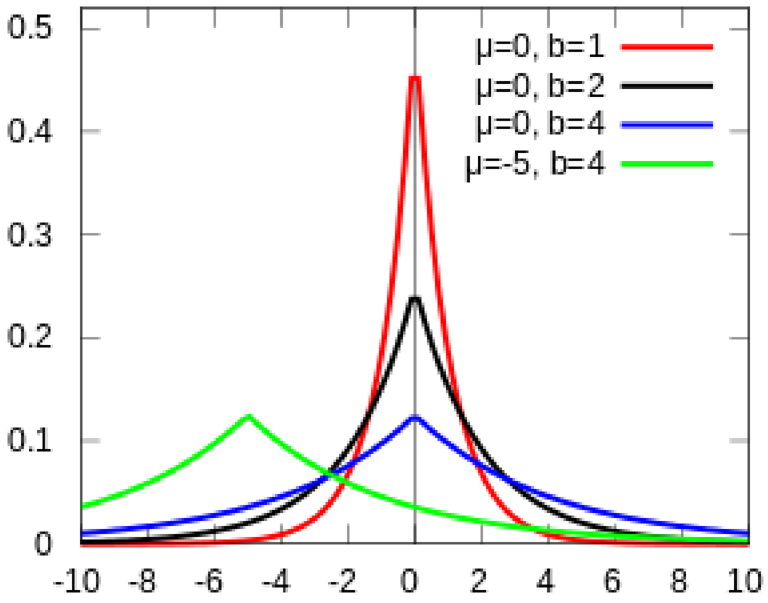


$$Z(\lambda) = \frac{1}{\lambda}$$ So, $A(\lambda) = \log Z(\lambda) = -\log \lambda$

Example: Laplace Distribution (double exponential distribution)

The Laplace distribution $(Laplace(\mu, b))$

$$f(x|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$



With known μ

$$\eta = -\frac{1}{b} \qquad T(y) = |x-\mu|$$

In general, it is not.

$$E(X) = \mu \qquad Var(X) = 2b^2$$

The(MLE) estimator of μ is the sample median.

- Laplace distribution relates to other distributions, e.g., the difference between two iid exponential random variables is governed by a Laplace distribution.
- Laplace distribution has been used in speech recognition and in JPEG image compression.

Example: Beta Distribution

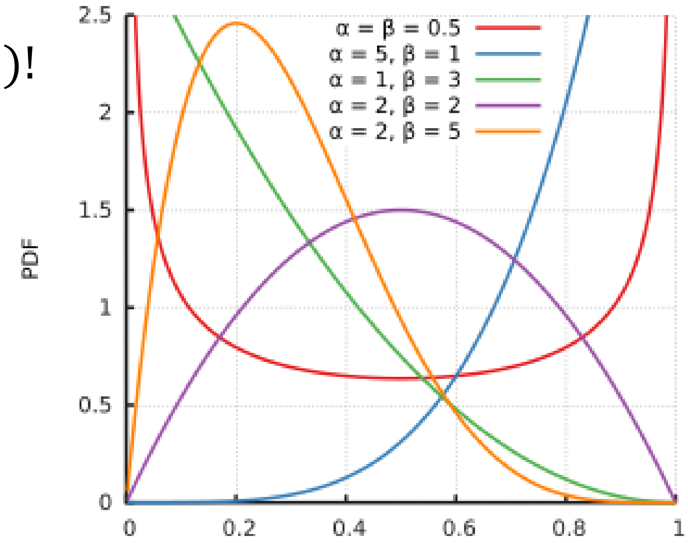
Beta Distribution ($Beta(\alpha,\beta)$**)** is often used as **prior** on Binomial distributions (it is a conjugate prior).

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \exp\Big((\alpha-1)\log \mathrm{x} + (\beta-1)\log(1-\mathrm{x}) + \log\big(\mathrm{B}(\alpha,\beta)\big)\Big)$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the Gamma function.

If $z = k$ is an positive integer, then $\Gamma(k) = (k-1)!$

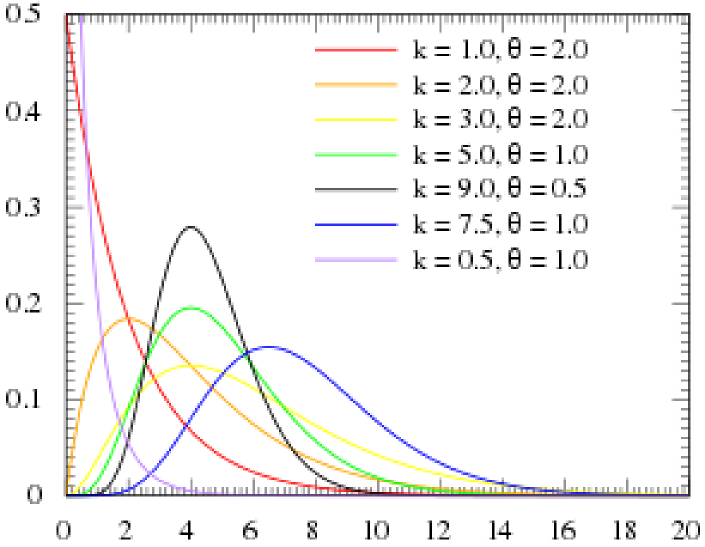


Example: Gamma Distribution

Gamma Distribution $(Gamma(k, \theta))$

$$p(x; k, \theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)} = \exp\left((k-1)\log x - \frac{x}{\theta} - k\log\theta - \log\Gamma(k)\right)$$

Here, $\Gamma(k)$ is the gamma function.



$E[X] = k\theta$

$Var[X] = k\theta^2$

Mode $[X]$= $(k-1)\theta$ if $k \geq 1$

$Gamma(1, \lambda) = \mathrm{Exp}(\lambda)$ the exponential distribution.

If k is a positive integer, $\Gamma(k) = (k-1)!$ the distribution $Gamma(k, \theta)$ is called Erlang distribution.

Gamma Distribution is popular as a prior on coefficients. Obtained from integral over waiting times in Poisson distribution

Erlang distribution is the distribution of the time until the kth event of a Poisson process with a rate of λ.

Example of waiting times in Queueing Systems: Erlang distribution models number of telephone calls which might be made at the same time to the operators of the switching stations.

Another popular parameterizations $Gamma(\alpha, \beta)$ with $\alpha = k, and\ \beta = \frac{1}{\theta}$

Moments of Exponential Family

In the family of exponential distributions, the $A(\vec{\eta})$ function is in fact the **Moment generating function of** $T(Y)$.

That is gradient $\nabla_{\vec{\eta}}\big(A(\vec{\eta})\big) = E\big(T(\vec{y})\big)$, Hessian matrix $H\big(A(\vec{\eta})\big) = Cov\big(T(\vec{y})\big)$

We show this by derivatizing this term:

$$A(\vec{\eta}) := \log \int h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy$$

Let us compute the one dimensional case:

$$\frac{d(A(\eta))}{d\,\eta} = \frac{\frac{d}{d\eta} \int h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy}{\int h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy} = \frac{\int T(\vec{y}) h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy}{\int h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy}$$

$$= \frac{\int T(\vec{y}) h(y) \exp\big(\vec{\eta}^T T(\vec{y})\big)\, dy}{\exp(A(\eta))} = \int T(\vec{y}) h(y) \exp(\vec{\eta}^T T(\vec{y}) - A(\eta))\, dy = E\big(T(\vec{y})\big)$$

Similarly, for the second derivative,

$$\frac{d^2(A(\eta))}{d\eta^2} = Var\big(T(\vec{y})\big)$$

Remark: Here our calculation is for one dimension η. In general, when $\vec{\eta} \in \mathbb{R}^{\boldsymbol{d}}$, we only need to change differential $\frac{d(A(\eta))}{d\,\eta}$ to gradient $\nabla_{\vec{\eta}}\big(A(\vec{\eta})\big)$.

$\mathrm{A}(\eta)$ is a convex function, since (co)variance matrix is positive semi-definite.

Example: (Bernouli)

$$A(\vec{\eta}) = \log(1 + e^{\eta})$$

So,

$$\frac{d(A(\eta))}{d\,\eta} = \cdots = \frac{1}{1 + e^{-\eta}} = \mu = E(Y)$$

➢ Location and Scale Families

The **location–scale family** is a class of probability distributions that can be obtained from a fixed "standard" distribution by **shifting** (location) and **scaling** (scale) its variable.

Let X be a random variable with a known CDF $F(x)$ and PDF $f(x)$.

Then, for constants $\mu \in \mathbb{R}$ (location) and $\sigma > 0$ (scale), the random variable:

$$Y = \mu + \sigma X$$

is said to belong to the **location–scale family** of the distribution of X.

The PDF for $Y\ is$

$$f_Y(y) = \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right)$$

Examples

- Normal distribution
- Cauchy distribution
- Uniform distribution (continuous)
- Uniform distribution (discrete)
- Logistic distribution
- Laplace distribution
- Student's t-distribution

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition
- **Book 2. [W]: All of Statistics: Larry Wasserman**
- **Book 3. Introduction to Probability**. C.M. Grinstead and J.L. Snell. American Mathematical Society, 2012
- **Book 4. Introduction to Probability Models**, S. Ross, 12th edition (published by Academic Press).

Online books:

https://www.probabilitycourse.com/