IE605: Engineering Statistics

Lecture 06

Manjesh K. Hanawal

Previous Lecture:

- Joint distribution of function of RVs
- Moment Generating Functions (MGFs)
- Conditional PMF and PDF
- Markov's and Chebyshev's inequalities
- Limit theorems: Law of Large Numbers (LLN)
- Limit theorems: Central Limit Theorem (CLT)

This Lecture:

- Exponential Family of Distributions
- Population and Random Sampling
- sample mean, variance and standard deviation
- Sampling from Normal distribution

Parametric Distributions

Discrete Case:

Distribution	PMF: $P(i)$
$Ber(p)$	$p^i(1-p)^{1-i},\ i=0,1$
$Bin(n,p)$	$\binom{n}{i}p^i(1-p)^{n-i},\ 0\leq i\leq n$
$Geo(p)$	$(1-p)^{i-1}p,\ i\geq 1$
$Poi(\lambda)$	$\frac{e^{-\lambda}\lambda^i}{i!},\ i\geq 0$

Continuous Case:

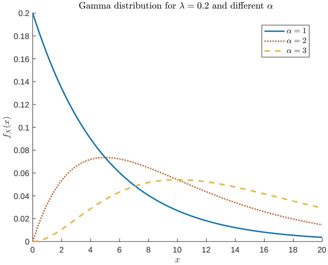
Distribution	PDF $f(x)$
$Uni(a,b)$	$\frac{1}{(b-a)}, x\in(a,b)$
$Exp(\lambda)$	$\lambda e^{-\lambda x}, \forall x>0$
$\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}, \forall x$
$Rayleigh(\sigma)$	$\frac{x}{\sigma^2}e^{-x^2/2\sigma^2}, \forall x>0$

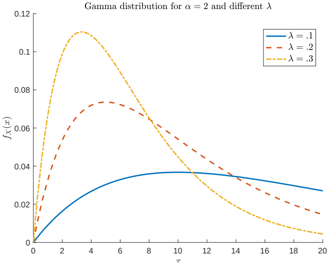
Gamma Distributions

Gamma Distribution:
$X \sim Gamma(\alpha, \lambda)$ for $\alpha, \lambda > 0$

$$f_X(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where α is the **shape parameter** and λ is the **scale parameter**.





Example: Model occurrence of earthquakes in time and magnitude.
Significance: When $\alpha = n$ for some positive integer, then $\sum_{i=1}^n X_i \sim Gamma(n, \lambda)$ where X_is are i.i.d. with $X_i \sim Exp(\lambda)$.

Special cases of Gamma distributions: Chi Square

- *Gamma*$(1/2, 1/2)$:chi-squared distribution with 1 degrees of freedom denoted χ_1^2. Set $\alpha = 1/2$ and $\lambda = 1/2$

$$f_X(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \frac{e^{-x/2}}{\sqrt{x}} & \text{for } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

If $U \sim \mathcal{N}(0, 1)$, $U^2 \sim Gamma(1/2, 1/2)$

- *Gamma*$(n/2, 1/2)$:chi-squared distribution with n degrees of freedom denoted χ_n^2. Set $\alpha = n/2$ and $\lambda = 1/2$

$$f_X(x) = \begin{cases} \frac{(1/2)^{n/2}}{\Gamma(n/2)} x^{n/2-1} e^{-x/2} & \text{for } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

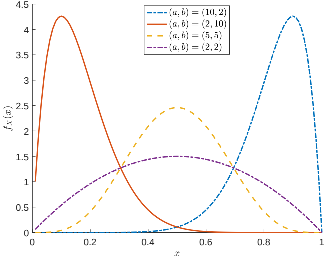
$(U_1, U_2, \ldots, U_n)$ are i.i.d. Now let $U_i \sim \mathcal{N}(0, 1)$. Then $\sum_{i=1}^n U_i^2 \sim \chi_n^2 = Gamma(n/2, 1/2)$. If $U_i \sim Exp(1/2)$, then $\sum_{i=1}^n U_i \sim Gamma(n, 1/2)$.

Beta distributions

Beta Distribution:
$X \sim Beta(a, b)$ for $a, b > 0$

$$f_X(x) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} & \text{for } x \in [0,1] \\ 0 & \text{otherwise,} \end{cases}$$

when $a = b = 1$, X is uniform on $[0, 1]$.



significance: Useful in Bayesian Statistics.

Exponential families

A family of pdf/pmf is exponential family if

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

- $h(x) \geq 0$ for all x and $c(\theta) \geq 0$
- $w_i(\theta)$ are real valued function of θ (cannot depend on x)
- $t_i(x)$ are real valued function of x (cannot depend on θ)

Discrete distributions

- Binomial
- Poisson
- Negative Binomial

Continuous distributions

- Gaussian
- Gamma
- Beta

Binomial as Exponential family

Fix n. Binomial family parameterized by $p = (0, 1)$

$$
\begin{aligned}
P(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} \\
&= \binom{n}{x} e^{x \log p} e^{(n-x)\log(1-p)} \\
&= \binom{n}{x} e^{x \log p + (n-x)\log(1-p)}
\end{aligned}
$$

Set $\theta = p$. Define:

- $c(\theta) = 1, \qquad h(x) = \begin{cases} \binom{n}{x} & \text{for } x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$
- $w_1(\theta) = \log p, w_2(\theta) = \log(1-p)$
- $t_1(x) = x, t_2(x) = n - x$

$$P(x|\theta) = h(x)c(\theta)\exp\{w_1(\theta)t_1(x) + w_2(\theta)t_2(x)\}$$

Gaussian as Exponential family

$\mathcal{N}(\mu, \sigma^2)$ is parameterized by $\mu \in \mathcal{R}$ and $\sigma^2 > 0$.

$$\begin{aligned} f(x|(\mu,\sigma^2)) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{x\mu}{\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2}\right\} \end{aligned}$$

Set $\theta = (\mu, \sigma^2)$. Define

- $c(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}$. $h(x) = 1$ for all x
- $w_1(\theta) = \frac{1}{2\sigma^2}, w_2(\theta) = \frac{\mu}{\sigma^2}$
- $t_1(x) = -x^2, t_2(x) = x$

$$f(x|\theta) = h(x)c(\theta)\exp\{t_1(x)w_1(\theta) + t_2(x)w_2(\theta)\}$$

Gamma as Exponential family

Gamma(α, λ) is parametrized bt α and λ.

$$\begin{aligned} f(x|(\alpha,\lambda)) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{(\alpha-1)\log x} e^{-\lambda x} \end{aligned}$$

Set $\theta = (\alpha, \lambda)$. Define

- $c(\theta) = \frac{\lambda^\alpha}{\Gamma(\alpha)}$, $h(x) = 1$ for all x
- $w_1(\theta) = (\alpha - 1), w_2(\theta) = -\lambda$
- $t_1(x) = \log x, t_2(x) = x$

$$f(x|\theta) = h(x)c(\theta)\exp\{w_1(\theta)t_1(x) + w_2(\theta)t_2(x)\}$$

Random Sampling

- Samples are used to obtain information about large populations by examining only a small fraction. Examples
  - Who will win the polls?
  - Will there be demand for a new car
  - How many pay taxes
  - Health of people
- How to sample for better results
- Do random sampling for unbiased (to be made precise) results

> Random Variables $X_1, X_2, \ldots, X_n$ are called random samples of size n from population $f(x)$ if they are i.i.d with common distribution with $f(x)$.

if $(x_1, x_2, \ldots, x_n)$ are samples from population $f(x)$

$$f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Sampling with and without replacement

With replacement

- After sampling, the sample is put back before the next sample is drawn randomly.
- Each sample comes from a new fresh experiment
- sampling with replacements gives i.i.d samples (random sample)

Without replacement

- After sampling, the sample is not put back, before the next sample is drawn randomly.
- sampling with replacements can give identical samples but not independent.

Statistic of Random Samples

- When a sample $X_1, X_2, \ldots, X_n$ is drawn, we would be interested in some summary of values
- Any well defined summary may be expressed as a function $T(X_1, X_2, \ldots, X_n)$

> The random variable/vector $Y = T(X_1, X_2, \ldots, X_n)$ is called statistic. The distribution of the statistic Y is called the sampling distribution of Y.

Often used statistics

- Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Sample standard deviation: $S = \sqrt{S^2}$

we will denote the observed values as $\bar{x}, s^2, s$, respectively.

Properties of statistics $\bar{X}$ and S^2

$X_1, X_2, \ldots, X_n$ is random sample from a population with mean μ and variance σ^2

- $\mathbb{E}(\bar{X}) = \mu$

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu$$

- $Var(\bar{X}) = \sigma^2/n$

$$\begin{aligned} Var(\bar{X}) = Cov(\bar{X}, \bar{X}) &= Cov\left(\frac{1}{n}\sum X_i, \frac{1}{n}\sum X_j\right) \\ &= \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right)\left(\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)\right)\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left((X_i - \mu)^2\right) = \frac{\sigma^2}{n} \end{aligned}$$

- $\mathbb{E}(S^2) = \sigma^2$

$$
\begin{aligned}
\mathbb{E}(S^2) &= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}(X_i + \mu - \mu - \bar{X})^2\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_{i}\left((X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu)\right)\right) \\
&= \frac{1}{n-1}\left(\sum_{i} Var(X_i) + \sum_{i} Var(\bar{X}) - \frac{2}{n}\sum_{i}\mathbb{E}((X_i - \mu)^2)\right) \\
&= \frac{1}{n-1}\left(n\sigma^2 + n\frac{\sigma^2}{n} - \frac{2}{n}n\sigma^2\right) = \frac{1}{n-1}(n\sigma^2 - \sigma^2) = \sigma^2
\end{aligned}
$$

- $\mathbb{E}(\bar{X}) = \mu$: Statistic $\bar{X}$ is unbiased estimator of μ
- $\mathbb{E}(S^2) = \sigma^2$: Statistic S^2 is unbiased estimator of σ^2

Sampling from Gaussian distribution

$X_1, X_2, \ldots, X_n$ is a random sample from population $\mathcal{N}(\mu, \sigma^2)$. Then, $\bar{X}$ and S^2 are such that

- $\bar{X}$ has a $\mathcal{N}(\mu, \sigma^2/n)$ distribution
- $\bar{X}$ and S^2 are independent
- $(n-1)S^2/\sigma^2$ has chi-square distribution with $n-1$ degree of freedom, i.e, $\sim Gamma((n-1)/2, 1/2)$.

Proof: workout!

Student's t-distributions

Random sample $X_1, X_2, \ldots X_n$ is drawn form population $\mathcal{N}(\mu, \sigma^2)$

- $\frac{\bar{X}-\mu}{\sigma^2/n} \sim \mathcal{N}(0,1)$
- If σ^2 is known $\frac{\bar{X}-\mu}{\sigma^2/n}$ can infer μ as it is the only unknown
- In most cases σ^2 is not known. How to infer about μ?
- G.S. Gosset (published under pseudonym student) introduced

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

> Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Then the quantity $(\bar{X}-\mu)/(S/\sqrt{n})$ has Student's t- distribution with $n-1$ degrees of freedom.

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}}$$

- Define $U = (\bar{X}-\mu)/(\sigma/\sqrt{n})$ and $V = (n-1)S^2/\sigma^2$
- $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{n-1}$ (chi-squared with $n-1$ degree of freedom)
- Random variables U and V are independent (check!)
- The distibution of $\frac{U}{\sqrt{V/n-1}}$ gives student's t-distribution

PDF of Student's t-distribution

- t_p denotes Student's t-distribution with p degrees of freedom
- If $X \sim t_p$, for all $-\infty < x < \infty$

$$f_X(x) = \frac{\Gamma\left(\frac{p-1}{2}\right)}{\Gamma(\frac{p}{2})} \frac{1}{\sqrt{p\pi}} \frac{1}{\left(1 + \frac{t^2}{p}\right)^{\frac{p+1}{2}}}$$

- Special case. Set $p = 1$ (corresponding to $n = 2$ samples)

$$f_X(x) = \frac{1}{\pi} \frac{1}{1 + t^2} \quad \text{(Cauchy Distribution)}$$

Derivation of Student's t-distribution

- $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{n-1}$
- Joint distribution of (U, V) for all $-\infty < u < \infty$ and $v > 0$

$$f_{UV}(u,v) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \frac{(1/2)^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} v^{\frac{n-1}{2}-1} e^{-v/2}$$

- Define transformation $X = \frac{U}{\sqrt{V/(n-1)}}$ and $Y = V$.
- Find Joint distribution $f_{XY}(x,y)$
- Find marginal $f_X(x)$