# Notes for STAT 510

**Jan 23rd, 2020**

Moments: usual moments $\mathbb{E}[x^k]$ and central moments $\mathbb{E}[(x-\mu)^k]$.

Moment generating function $M_{\mathbf{x}}(\mathbf{t}) = \mathbb{E}[e^{\mathbf{t}\cdot\mathbf{X}}]$. Therorem of uniqueness: generating function exists in a small neighborhood around origin, then and two generating functions conincide, the two distributions are the same.

to get moments:

$$\mathbb{E}[X_1^{k_1}\cdots X_p^{k_p}] = \frac{\partial^{k_1+\cdots+k_p}}{\partial t_i^{k_1}\cdots\partial t_p^{k_p}} M_{\mathbf{x}}(\mathbf{t})|_{\mathbf{t}=0}$$

If the generating function is analytic, then it is completely determined by its derivatives and thus all moments. Thus, Corollary: If $X$ and $Y$ agree on all moments, and their mgf (moment generating function) exist then they have the same distribution.

---

**Cumulant (connected correlation function?)**

Cumulant generating function $C_{\mathbf{x}}(\mathbf{t}) = \log(M_{vb}(\mathbf{t}))$. It generates cumulants $\gamma_k = C^{(k)}(0)$.

$$\begin{aligned}
\gamma_1 &= \mu \\
\gamma_2 &= \sigma^2 = \mu_2 \\
\gamma_3 &= \mathbb{E}[(X-\mu)^3] = \mu_3 = \sigma^3 k_3 \\
\gamma_4 &= \mu_4 - 3\sigma^4 = \sigma^4 k_4
\end{aligned}$$

**Question:** For Boltzman distribution, when we do partition function why it looks similar to the above case? It looks like the distribution weights in calculating partition function is 1.

Example 1: For normal distribution, we have $\mu = C'(0) = 0$, $\sigma^2 = C''(0) = 1$ and $k_3 = C^{(3)}(0) = 0$ and $k_4 = C^{(4)}(0) = 0$.

Example 2: Binomial distribution $X \sim Bin(n,p)$. Probability mass function:

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

From the binomial theorem, this is a valid pmf. To get the mgf for it, let's do the calculation:

$$\begin{aligned}
M_X(t) &= \sum_{x=0}^{n} e^{tx} f_X(x) = \sum_{x=0}^{n} (e^t)^x \binom{n}{x} p^x (1-p)^x \\
&= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = (pe^t + 1 - p)^n
\end{aligned}$$

---

**Sum of two independent random variables** Discrete case: $X_1 \sim Unif(\{0,1\})$ and $X_2 \sim Unif(\{0,1,2\})$ and they are independnet. We want to find the distribution of the sum $Y = X_1 + X_2$.

To do this, we first list the event space of $Y$. i.e. Spaces of $Y$ is $\{0,1,2,3\}$. The pmdf of $Y$ is then

$$f_Y(0) = P(Y=0) = P(X_1=0)P(X_2=0) = \frac{1}{2}\frac{1}{3} = \frac{1}{6}$$
$$f_Y(1) = P(Y=1) = P(X_1=0, X_2=1) + P(X_1=1, X_2=0) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$
$$\vdots$$

We then can generalize to get the **Discrete convolution formula**:
Let $X_1$ have pmf $f_1$ with space $\{0,1,\cdots,a\}$ and $X_2$ have pmf $f_2$ with space $\{0,1,\cdots,b\}$. Assume $X_1$ is independent of $X_2$, then the distribution of $Y = X_1 + X_2$ is:
Space of $Y$ is $\{0,1,\cdots,a+b\}$ and $P(Y=y)$ for $y = 0,1,\cdots,a+b$ we need to sum all pairs $(x_1,x_2)$ s.t. $0 \le x_1 \le a$ and $0 \le x_2 \le b$, $x_1 + x_2 = y$. This means $0 \le x_1 \le a$ and $0 \le y - x_1 \le b$ or $\max\{0, y-b\} \le x_1 \le \min\{a,y\}$. A jargon: $\max\{a,b\} = a \vee b$ and $\min\{a,b\} = a \wedge b$. This means

$$P(Y=y) = \sum_{x=0\vee(y-b)}^{a\wedge y} f_1(x_1) f_2(y-x_1) \equiv f_1 * f_2(y)$$

The convolution star $*$ is defined as

$$f * g(x) = \sum_{k=-\infty}^{\infty} f(k) g(x-k)$$

for range $(-\infty, \infty)$. This is done by estending the definition of $f$ and $g$ so that they are zero when outside their original defined domain.
Example: For $X_1 \sim Poisson(\lambda_1)$ and $X_2 \sim Poisson(\lambda_2)$. Then $f_1 = e^{-\lambda}\frac{\lambda^x}{x!}$. Then the sum of two independent Poisson is

$$f_Y(y) = \sum_{x_1=0}^{y} f_1(x_1) f_2(y-x_1) = \sum_{x=0}^{y} e^{-\lambda_1}\frac{\lambda_1^{x_1}}{x_1!} e^{-\lambda_2}\frac{\lambda_2^{y-x_1}}{(y-x_1)!}$$
$$= \frac{1}{y!} e^{-\lambda_1+\lambda_2} \sum_{x=0}^{y} \frac{y!}{x!(y-x)!} \lambda_1^x \lambda_2^{y-x} = \frac{1}{y!} e^{-(\lambda_1+\lambda_2)} (\lambda_1+\lambda_2)^y = Poisson(\lambda_1+\lambda_2)$$

So sum of two Poisson is still Poisson with mean summed.

---

**Convolution for continuous random variables**
Let $(X_1, X_2)$ have joint pdf $f(x_1, x_2)$. Assume the space of $(X_1, X_2)$ is $\mathbb{R}^2$. We would like to find distribution of of $Y = X_1 + X_2$.
We first calculate the cumulative distribution function

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X_1 + X_2 \le y) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{y-x_1} f(x_1, x_2) dx_2 \right\} dx_1$$

Then the pdf of $Y$ is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \int_{-\infty}^{\infty} \left[ \frac{d}{dy} \int_{-\infty}^{y-x_1} f(x_1, x_2) dx_2 \right] dx_1 = \int_{-\infty}^{\infty} f(x_1, y-x_1) dx_1.$$

---

**Jan 28th, 2020**

Last lecture: sums of two variable,i.e. convolution. $Y = X_1 + X_2$ with pmf $f_i$ for $i = 1, 2$. Then

$$f_Y(y) = \sum_{x=0\vee(y-b)} f_1(x) f_2(y-x).$$

with $f_1$ supported on $\{0, \ldots, a\}$ and $f_2$ supported on $\{0, \ldots, b\}$.

Convolution for continuous r.v. we have simimilar expression:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x_1, y - x_1) dx_1$$

If $X_1$ and $X_2$ are independent, we have

$$f_Y(y) = \int_{-\infty}^{\infty} f_1(x_1) f_2(y - x_1) dx_1 \equiv f_1 * f_2(y).$$

---

For example, $X_1, X_2$ are i.i.d. $Exp(\lambda)$ with pdf

$$f_1(x) = f_2(x) = \begin{cases} 0 \\ \lambda e^{-\lambda x} \end{cases} = \lambda e^{-\lambda x} \mathbb{1}(x \geq 0)$$

We then get $f_Y(y) = 0$ for negative $y$. For $y \geq 0$,

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} \lambda e^{-\lambda x_1} \lambda e^{-\lambda(y-x_1)} \mathbb{1}(x_1 \geq 0) \mathbb{1}(y - x_1 \geq 0) dx_1 \\ &= \int_0^y \lambda^2 e^{-\lambda y} dx_1 = \lambda^2 y e^{-\lambda y} = \frac{\lambda^2 y^{2-1)}}{\Gamma(2)} e^{-\lambda y} \end{aligned}$$

This is the Gamma distribution $\Gamma(a = 2, \lambda)$

---

Using hthe moment generating function to calculate the sum of two random variables.

Idea: Find the mgf of $Y$ if we can recognize the mgf of $Y$ as being for a particular distribution. Then by uniqueness of mgf Y has that distribution.

Property: If $X_1, X_2, \ldots, X_k$ are **independent**, and $Y = X_1 + \cdots X_k$, then

$$M_Y(t) = \prod_{j=1}^{k} M_{x_j}(t)$$

Proof: Just plug in the form of $Y$ into the definition of mgf. (Simple)

For example, $X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ and $Y = \sum_j X_j$ with $M_{x_j}(t) = e^{\mu_j t + \frac{1}{2}\sigma_j^2}$ and Then we get

$$M_Y(t) = \prod_{j=1}^{k} e^{\mu_j t + \frac{1}{2} t^2 \sigma_j^2} = e^{(\sum \mu_j) t + \frac{1}{2} t^2 (\sum \sigma_j^2)}$$

This is Normal distribution with mean $\sum \mu_j$ and variance $\sum_j = 1\sigma_j^2$.

For discrete binary distribution with $X_j \sim Bin(n_j, p)$, we calculate $Y = \sum_j X_j$. $M_{x_j} = (pe^t + 1 - p)^{n_j}$ Then

$$M_Y(t) = \prod (pe^t + 1 - p)^{n_j} = (pe^t + 1 - p)^{\sum n_j}$$

. Thus $Y \sim Bin(\sum_j (n_j), p)$. Specially: when $n_i = 1$ for all $i$, then each one is a Bernoulli r.v. and the sum is a $Bin(n, p)$

---

**Transformation** of r.v. (one-dim)
Let $X$ and $Y$ be two r.v.'s in $\mathbb{R}$. s.t. $Y = g(X)$ for some function $g : \mathbb{R} \to \mathbb{R}$. Use $F_X$ or $f_X$ denote the cdf or pdf of $X$ respectively. Then determine $F_Y$ and $f_Y$ of $Y$.
For example: Uniformlly shoot with $\theta \sim unif(0, \pi)$ Then distribution of $X = \cot(\theta)$.
Work with the cdf of $X$. Then $F_X(x) = P(X \leq x) = P(\cot(\theta) \leq x) = P(\theta \geq \text{arccot}(x)) = \frac{1}{\pi}(\pi - \text{arccot}(x))$. Since derivative is $\frac{d}{dx}\text{arccot}(x) = -\frac{1}{1+x^2}$. Thus, $f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. This is Cauchy distribution.

---

Consider general $g : x \to y$ is one-to-one. Then $g$ and $g^{-1}$ is one to one and are strictly either increasing or decreasing. $g$ is not one-to-one Then $g^{-1}(y) = \{x \in X : g(x) = y\}$. In discrete case: $\forall y \in Y$, $f_Y(y) = P(Y = y) = P(x = g^{-1}(y)) = f_X(g^{-1}(y))$.
Continuous case with $g$ one-to-one, then $Y$ is continuous also. Take any point $y$. Let $B_\epsilon = [y - \epsilon/2, y + \epsilon/2]$ for small $\epsilon$. Thus, the interval on $X$ will be $[g^{-1}(y - \epsilon/2), g^{-1}(y + \epsilon/2)]$ because of monotonicity, the order of lower and upper bounds are not changed. Then the probability of $y$ in $B_\epsilon(y)$ would be

$$P(y \in B_\epsilon) \approx f_Y(y)\epsilon$$

. Alternatively $P(Y \in B_\epsilon) = P(X \in g^{-1}(B_\epsilon))$. However

$$g^{-1}(B_\epsilon) = g^{-1}([y - \epsilon/2, y + \epsilon/2]) = \left[g^{-1}(y) - g^{-1\prime}(y)\frac{\epsilon}{2}, g^{-1}(y) + g^{-1\prime}(y)\frac{\epsilon}{2}\right]$$

Then we get

$$P(y \in B_\epsilon) = f_X(g^{-1})g^{-1\prime}(y)\epsilon \quad \to f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|.$$

---

**Jan 30th, 2020**
For some random variable $X$ and $Y = g(X)$ where g is one-to-one. Then

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$$

Example: $X \sim Unif(0, 1)$ and $Y = -\log(X)$. Thus, $X = (0, 1)$ and $Y = (0, \infty)$. Then $g^{-1}(y) = e^{-y}$. The Jacobean is $-e^{-y}$ and therefore $f_Y(y) = e^{-y}$.
Example: [Probability Transform] Let F be some cdf which is strictly increasing. Let $U$ be a uniform in $(0, 1)$. Then let $X = F^{-1}(U)$. Then $f_X(x) = f_U(F(x))|F'(x)| = f(x)$.
The inverse transform sampling: Let $W = F^{-1}(U)$, we claim $W$ has cdf $F$.
proof: $F_W(x) = P(W \leq x) = P(F^{-1}(U) \leq X) = P(U \leq F(x)) = F(x)$.
If $X \sim Exp(\lambda)$, Then the cdf is $1 - e^{-\lambda x}$. Then $F^{-1}(u) = -\log(1 - u)/\lambda$. This is the way to generate exponential random variables. Since $(1 - u)$ is also uniform in $(0, 1)$ and thus, we can directly generate $u' = 1 - u$ and plug in as $-\log(u')/\lambda$.
Example:[location-scale family] (affine transformation) Affine means $Y = aX + b$.
**Lemma**: Let $Z$ have pdf $F$ on $\mathbb{R}$. Then for any $\mu \in \mathbb{R}$ and $\sigma > 0$, $X = \mu + \sigma Z$ has cdf equal to $F((x - \mu)/\sigma)$.

In addition if $Z$ has pdf $f$ on $\mathbb{R}$, Then $X$ has pdf $\frac{1}{\sigma}f((x-\mu)/\sigma)$.
proof: $F_x(x) = P(\mu + \sigma Z \le x) = P(Z \le (x-\mu)/\sigma) = F((x-\mu)/\sigma)$. If $F$ is differentiable, Then $F_X$ is differentiable and we take derivative to get $f_X$ which will be $f_X = \frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$.
When the function $g$ is not monotonic globally, we need to divide the total interval into small ones that are monotonic in each of them and combine all the interval together.

---

Using mdf to calculate the distribution. If $M_Z(t)$ is the mdf of $Z$. Then $X$ has mdf $M_X(t) = \mathbb{E}[e^{(\mu+\sigma Z)t}] = e^{\mu t}M_Z(\sigma t)$
For example $Z \sim \mathcal{N}(0,1)$ Then $M_z(t) = e^{t^2/2}$. Then $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$ if $X = \mu + \sigma Z$.
Example: $U \sim unif(0,1)$ $X = -\log(U)$ Then $M_X(t) = \mathbb{E}[e^{-t\log(U)}]$ Then

$$M_X(t) = \int_0^1 e^{-t\log u}du = \int_0^1 u^{-t}du = \frac{1}{1-t}\, e^{1-t}\Big|_{x=0}^1 = \frac{1}{1-t}$$

Here $t < 1$ should be required to ensure convergence.

---

**Transformations on multivariate**
Def[Jacobian] let $g : x \to y$ be one to one and $g^{-1}$ is continuously differentiable. Then the Jacobian of the transformation $g^{-1}$ is defined as $J_{g^{-1}} : y \to \mathbb{R}$ with

$$J_{g^{-1}}(y) = \begin{vmatrix} \frac{\partial}{\partial y_1}g_1^{-1}(y) & \cdots & \frac{\partial}{\partial y_p}g_1^{-1}(y) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial y_1}g_p^{-1}(y) & \cdots & \frac{\partial}{\partial y_p}g_p^{-1}(y) \end{vmatrix}$$

Theorem suppose this $J_{g^{-1}}(y)$ is continous and nonzero for all $y \in Y$, then $f_Y(y) = f_X(g^{-1}(y))|J_{g^{-1}}(y)|$
Example $X_1 \sim Ga(a,\lambda)$ and $X_2 \sim Ga(\beta,\lambda)$, for $X_1 \perp X_2$. Q: what's the distribution of $Y = X_1/(X_1+X_2)$?
Define the transformation $(X_1, X_2) \to (Y_1, Y_2)$ for $Y_2 = X_1 + X_2$. Then the space will be $X = (0,\infty)\times(0,\infty)$ and $Y = (0,1)\times(0,\infty)$. To find $g^{-1}$ we solve

$$\begin{cases} y_1 = \dfrac{x_1}{x_1+x_2} \\ y_2 = x_1 + x_2 \end{cases} \quad \Rightarrow \quad \begin{cases} x_1 = y_1y_2 \\ x_2 = y_2 - x_1 = y_2(1-y_1) \end{cases}$$

We then get the Jacobian as

$$J_{g^{-1}}(y) = \begin{vmatrix} \frac{\partial}{\partial y_1}y_1y_2 & \frac{\partial}{\partial y_2}y_1y_2 \\ \frac{\partial}{\partial y_1}(y_2 - y_1y_2) & \frac{\partial}{\partial y_2}(y_2(1-y_2)) \end{vmatrix} = \begin{vmatrix} y_2 & y_1 \\ -y_2 & 1-y_1 \end{vmatrix} = y_2$$

Therefore,

$$\begin{aligned} f_Y(y_1,y_2) &= \frac{\lambda^\alpha}{\Gamma(\alpha)}(y_1y_2)^{\alpha-1}e^{-\lambda y_1y_2}\frac{\lambda^\beta}{\Gamma(\beta)}(y_2(1-y_1))^{\beta-1}e^{-\lambda y_2(1-y_1)}y_2 \\ &= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)}y_1^{\alpha-1}(1-y_1)^{\beta-1}y_2^{\alpha+\beta-1}e^{-\lambda y_2} \\ &= \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y_1^{\alpha-1}(1-y_1)^{\beta-1}\right]\left[\frac{\lambda^{\alpha+\beta}}{\Gamma\alpha+\beta}y_2^{\alpha+\beta-1}e^{-\lambda y_2}\right] \end{aligned}$$

This is $Beta(\alpha,\beta) * Gamma(\alpha+\beta,\lambda)$. Thus $Y_1 \perp Y_2$ because they factorize.

Dirichlet distribution (multivariate version of Beta distribution) $X_1 \sim Ga(\alpha, \lambda)$ and $X_2 \sim Ga(\beta, \lambda)$ $X_1 \perp X_2$. Then $Y_1 = X_1/(X_1 + X_2)$ is $beta(\alpha, \beta)$.
General: $X_k \sim Ga(\alpha_k, \lambda)$ for $k = 1, \cdots K$ Let $Y_k = x_k/(x_1 + x_2 + \cdots x_k)$ Then $Y = (Y_1, \cdots Y_{k-1}, Y_k = 1 - \sum) \sim Dirichlet(\alpha_1, \ldots, \alpha_k)$

**Feb 4th, 2020**

Define Dirichlet distribution: $Y = (Y_1, \ldots, Y_{k-1}) \sim Diri(\alpha_1, \ldots, \alpha_k)$ with pdf

$$f_Y(y) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} y_1^{\alpha_1 - 1)} \ldots y_{k-1}^{\alpha_{k-1}-1}(1 - y_1 - \ldots y_{k-1})^{\alpha_k - 1}$$

The distribution of $Y_k$ is $X_k/(X_k + \sum_{i \neq k} \alpha_i$. Since $\sum_{i \neq k} X_k$ is $Gamma(\sum_{i \neq k} \alpha_i), \lambda)$, by the last example in the last lecture we get $Y_k \sim Beta(\alpha_k, \sum_{i \neq k} \alpha_k)$.
As a result, $(Y_1 + Y_2, Y_3, \ldots, Y_{k-1})$ is $Diri(\alpha_1 + \alpha_2, \ldots, \alpha_k)$. More generally,

$$(\sum_{k \in G_1} Y_k, \ldots, \sum_{k \in G_J} Y_k) \sim Diri(\sum_{i \in G_1} \alpha_i, \ldots, \sum_{i \in G_J} \alpha_i)$$

---

**Affine transformation**

$X$ is a $p$ dimensional random vector and we can define $Y = BX + a$ for $B$ as an invertible $p \times p$ random vector. That is $Y = g(X) = BX + a$. Then $g^{-1}(y) = B^{-1}(y - a)$. The Jacobian is just $B^{-1}$. And the determinant is $\det(B)^{-1}$

The **Bivariate Normal**: If $X$ is a random vector, Then the covariance matrix is $\Sigma_{ij} = Cov(X_i, X_j) = \mathbb{E}((X - \mathbb{E}(X)(X - \mathbb{E}(X))^T)$. Obviously, $Cov(a + BX) = B \cdot Cov(X) \cdot B^T$.
If $(X_1, X_2)$ are i.i.d. $N(0, 1)$. Then define $Y = BX + a$, and we denote the distribution of $Y$ is $Y \sim N(a, BB^T)$.
First, we get distribution function for $(X_1, X_2)$ as

$$f_X(x) = f_1(x_1)f_2(x_2) = \frac{1}{\sqrt{2\pi}} e^{-x_1^2/2} e^{-x_2^2/x} = \frac{1}{2\pi} e^{-X^T X/2}$$

And then we can get distribution for $Y$ as

$$f_y(y) = f_x(g^{-1}(y))|jacob| = \frac{1}{2\pi} e^{-\frac{1}{2}(B^{-1}(y-a))^T B^{-1}(y-a)} |B|^{-1} = \frac{1}{2\pi|B|} e^{-\frac{1}{2}(y-a)^T (BB^T)^{-1}(y-a)}$$

Since we define $\Sigma = BB^T$, the pdf is then written as

$$f_y(y) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left[-\frac{1}{2}(y-a)^T \Sigma^{-1}(y-a)\right].$$

**Orthogonal transformations**

Define: a $p \times p$ orthogonal matrix is a matrix $\Gamma$ suchthat $\Gamma^T \Gamma = \mathbb{1}_{p \times p}$. If we define $\gamma_i$ as the $i$-th row of $\Gamma$, we get $\gamma_i \cdot \gamma_j = \delta_{ij}$. The two norm is conserved since $(\Gamma x)^T \Gamma x = x^T \Gamma^T \Gamma x = x^T x$.

---

**Feb 6th, 2020**

Affine transformation $\mathbf{Y} = B\mathbf{X} + a$. Orthogonal transformation is $\Gamma^T\Gamma = \Gamma\Gamma^T = \mathbb{I}_p$.

---

**Spherically symmetric distribution**

Def A $p \times 1$ random vector $\mathbf{X}$ has a spherically symmetric distribution if any orhtogonal matrix $\Gamma_{p\times p}$, $X$ and $\Gamma X$ have the same distribution.

If $X$ has pdf $f_X$ then this means $f_X(x) = f_{\Gamma x}(x)$

Remark: if $\Gamma^T\Gamma = \Gamma\Gamma^T = \mathbb{1}_p$, Then $|\Gamma|^2 = 1$. Thus $\det(\Gamma) = \pm 1$. Also, $\Gamma^{-1} = \Gamma^T$. This means the matrix $\Gamma$ is invert

Let $Y = \Gamma X = g(X)$ then $X = g^{-1}(X) = \Gamma^T Y$. Thus the Jacobian is 1. Thus, $f_Y(y) = f_X(\Gamma^{-1}y) = f_X(y)$ . Then $f_X(x) = f_X(\Gamma x)$.

Consider $p = 2$, let $X \in \mathbb{R}^2$ $r = \|x\| = \sqrt{x_1^2 + x_2^2}$. Choose an orthogonal matrix $\Gamma \in O(2)$, such that $\Gamma x = (\|x\|, 0)$ and thus, $f_X(x) = f_X((\|x\|, 0)) = h(\|x\|)$.

For general case, $p \geq 2$, we can always find $\Gamma x = \|X\|\hat{\mathbf{x}}$.

Property: If $X \in \mathbb{R}^p$ is spherically symmetric, and continuous, there exists some function $\mathbb{R}_+ \to \mathbb{R}$ st $f_X(x) = h(\mathbb{X})$.

---

Consider polar coordinates in 2 dimension, $x_1 = r\cos\theta$ and $x_2 = r\sin\theta$, then if $X$ is spherically symmetric, then $R$ and $\Theta$ are independent.

proof: $J_{g^{-1}}(r, \theta) = r$ because $dxdy = d(r\sin\theta)d(r\cos\theta) = rdrd\theta$and therefore $f_{R,\Theta}(r, \theta) = f_X(g^{-1}(r, \theta)) = h(r)r$

Therefore, $f_{R,\Theta}(r, \theta) = 2\pi r h(r)\frac{1}{2\pi}$. Thus, $f_R(r) = 2\pi r h(r)$ and $f_\Theta(\theta) = 1/2\pi$.

If $X$ is bivariate normal, we have $f_X(x) = \frac{1}{2\pi}e^{-(x_1^2+x_2^2)/2}/2\pi$. This means $f(r) = re^{-r^2/2}$. Inverse transform sampling $F_R(r) = 1 - e^{-r^2/2} = u$. Then generate two independent uniform $u_1$ and $u_2$, then

$$\begin{cases} R = \sqrt{-2\log(1-u_1)} \\ \Theta = 2\pi u_2 \end{cases} \Rightarrow \begin{cases} X_1 = \sqrt{-2\log(1-u_1)}\cos(2\pi u_2) \\ X_2 = \sqrt{-2\log(1-u_2)}\sin(2\pi u_2) \end{cases}$$

---

**Order Statistics**

Given sample $X_1, X_2, \ldots X_n$, the order statistics are $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ where

$$\begin{aligned} X_{(1)} &= \min(X_1, \ldots, X_n) \\ X_{(2)} &= \text{second smallest of } X_1, \ldots, X_n \\ &\vdots \\ X_{(n)} &= \max(X_1, \ldots, X_n) \end{aligned}$$

This map is highly non-linear and $n!$ to one map. Let $Y = (X_{(1)}, \ldots, X_{(n)})$ and suppose $X_i$ are i.i.d. Since the map is not one to one, we must use infininitesmal analysis. Assume that $X_i$'s are distinct. Space of $y$ is $\{(y_1, y_2, \ldots, y_n) : y_1 < \cdots < y_n\}$. Fix a $y$, choose $\delta > 0$ small enough such that $(y_1, y_1 + \delta], \ldots (y_n, y_n + \delta]$ are distinct.

For example $n = 2$,

$$\begin{aligned}
&P(y_1 < X_{(1)} < y_1 + \delta, y_2 < x_{(2)} \le y_2 + \delta) \\
=&P((y_1 < x_1 \le y_1 + \delta, y_2 < x_2 \le y_2 + \delta) + P(y_1 < x_2 \le y_2 + \delta, y_1 < x_1 \le y_2 + \delta) \\
=&2P(y_1 < x_1 \le y_1 + \delta)P(y_2 < x_2 \le y_2 + \delta) = 2(F(y_1 + \delta) - F(y_1)))(F(y_2 + \delta) - F(y_2))
\end{aligned}$$

In general,

$$\begin{aligned}
&P(y_1 < x_{(1)} \le y_1 + \delta, \dots, y_2 < x_{(n)} \le y_n + \delta) \\
=&n!P(y_1 < x_1 \le y_1 + \delta, \dots, y_n < x_n \le y_n + \delta) \\
=&n! \prod_{i=1}^{n} [F(y_i + \delta) - F(y_i)]
\end{aligned}$$

---

**Feb 11, 2020**

$X_1, X_2, \cdots, X_n$ i.i.d. cdf $F$ with pdf $f$. $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. And denote $X_{(1)} = Y_1$, $X_{(2)} = Y_2 \cdots$.
Consider the special case $X_i \sim U(0, 1)$. By the general formula, we have $f_Y(y) = n! \prod f(y_i) = n!$

$$y \in y = \{(y_1, \dots, y_m) : 0 < y_1 < \cdots < y_n < 1\}$$

Consider the gap statistics, $G_1 = X_1$, $G_2 = X_{(2)} - X_{(1)}$ and $G_n = X_{(n)} - X_{(n-1)}$.
Then $G_i \in (0, 1)$ and $X_{(k)} = \sum_{i=1}^{k} G_i$ space of $G$ is

$$G = (G_1, \dots, G_n) = \{g \in \mathbb{R}^n : 0 < g_i < 1, 0 < g_1 + \dots g_n < 1\}$$

There is a linear map between $(g_1, \dots, g_n)$ and $y_1, \dots y_n$. That is $g = \phi(y)$

$$\begin{pmatrix} g_1 \\ \vdots \\ \vdots \\ g_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & \cdots & \cdots & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} \tag{0.1}$$

Then, we can inverse the map and get pdf for $G$ and get

$$f_G(g) = f_Y(\varphi^{-1}(g)) = n!$$

Recall the Dirichlet distribution $Diri(a_1, \dots, a_k)$, we have

$$f_X(x) = \frac{\Gamma(a_1 + \cdots + a_k}{\Gamma(a_1) \dots \Gamma(a_k)} x_1^{a_k - 1} \dots x_{k-1}^{a_{k-1} - 1} (1 - x_1 - x_{k-1})^{a_k - 1}$$

Thus, $f_G$ is $Diri(1, 1, \dots, 1)$ with $n + 1$ 1's. From the property of Dirichlet, we have $X_{(k)} = G_1 + \cdots + G_k$ is $Beta(k, n + 1 - k)$. This is the marginal distribution for $y_k$.
For general case, $X_1, \cdots, X_n$ are iid with $F$ or pdf $f$. Then the inverse transform $U_1, \dots, U_n$ iid $U(0, 1)$ , we know

$$F^{-1}(u_1), \dots, F^{-1}(u_n)$$

iid $F$. We let $u_{(1)} \le u_{(2)} \le \cdots \le u_{(n)}$ and let $X_i = F^{(-1)}(u_{(i)})$.

Marginal $u_{(k)} \sim Beta(k, n+1-k)$ We then have $X_{(k)} \equiv g(u_{(k)}) = F^{-1}(u_{(k)})$ Then $|J_{g^{-1}}(u)| = f(u)$ since $g^{-1} = F$. Thus, the pdf of $X_{(k)}$ is

$$\begin{aligned} f_{X_{(k)}} &= f_{u_{(k)}}(F(x_i))|J_{g^{-1}(x)}| = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} F(x)^{k-1}(1-F(x))^{n-1} f(x) \\ &= \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}(1-F(x))^{n-k} f(x) \end{aligned}$$

---

**Poisson process**

Definition: A Poisson process counts number of random events occurring in a fixed time or space when

(i) the number of events occurring in non-overlapping intervals are independent.

(ii) events occur at a constant rate $\lambda$ per unit time.

(iii) events cannot occur simultaneously.

Example: number of telephone calls occurring during 3 pm and 5 pm.
Let $X$ equal to the number of events in an interval with length $t$. Then we may ask what is the distribution of $X$.
Let $Y_i$ $i = 1, \ldots, n$ be the number of events occurring during the i-th Bin $B_i$ Then $Y_i$ are mutually independent by (i). And by (ii) and (iii) we have $Y_i \sim Bernoulli(\lambda t/n)$ Since we let $n$ is large and no two events can occur in that small interval. Thus, it is a Bernoulli with success p as $\Delta t\lambda = \lambda(t/n)$.
We let $X = Y_1 + \ldots Y_n$ and thus $X \sim Bin(n, \lambda t/n)$.
For $\forall x = 0, 1, \ldots, 2$,

$$\begin{aligned} P(X = x) &= \binom{n}{x} \left(\frac{\lambda t}{n}\right)^x \left(1 - \frac{\lambda t}{n}\right)^{n-x} \\ &= \frac{n!}{(n-x)!x!} \frac{(\lambda t)^x}{n^x} \left(1 - \frac{\lambda t}{n}\right)^{n-x} \\ &= \frac{(\lambda t)^x}{x!} \frac{n(n-1)\ldots(n-x+1)}{n \cdots n} \left[\left(1 - \frac{\lambda t}{n}\right)^{n/\lambda t}\right]^{\lambda t} (1 - \lambda/n)^{-x} \\ \to_{n\to\infty} &\frac{(\lambda t)^x}{x!} 1^x e^{-\lambda t} 1^{-x} = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \end{aligned}$$

This is $X \sim Poisson(\lambda t)$. and $\mathbb{E}[X] = \lambda t$.
Remark: Poisson approximation to Bernoulli distribution If $X_n \sim Bin(n, p)$ where $n$ is large and $p$ is small but keep $\lambda = np$ fixed. Then $X \sim Poisson(\lambda)$.

---

**Feb 13th, 2020**

Poisson process, the number of occurrence in $[0, t]$ is $N_t$ and $N_t \sim Poisson(\lambda t)$.
Also can use Poisson to approximate Binomial distribution.

For example row two dices $n$ times and count the number of occurence of double 6. Thus, $X \sim Bin(n, 1/36)$ and we can approximate with $Poisson(n/36)$. Thus,

$$P(X \leq 2) = e^{-n/36} + \frac{n/36}{1!}e^{-n/36} + \frac{(n/36)^2}{2!}e^{-n/36}$$

Suppose the number of volcano erruption occurring in the next 50 years follow a Poisson process with rate $\lambda$.

Q: When is the next volcano erruption?

Let $T =$ waiting time until the next erruption Clearly, $T \geq 0$, $T$ is a countinuous random variable. And $N_t$ is the number of erruptions before time $t$. $N_t \sim Poisson(\lambda t)$.

We can first calculate the cdf for $T$. Then

$$F_T(t) = P(T \leq t) = P(N_t \geq 0) = 1 - P(N_t = 0) = 1 - e^{-\lambda t}$$

because $T \leq t$ means before time $t$, there is at least one occurrence. This shows $T \sim Exp(\lambda)$. The waiting time is thus expected to be $\mathbb{E}[T] = 1/\lambda$.

Let $T_\alpha$ be the waiting time until the $\alpha$th erruption occurrence.

Q: What is the distribution of $T_\alpha$?

Similarly, we can find

$$F_{T_\alpha} = \mathbb{P}[T_\alpha \leq t] = \mathbb{P}[N_t \geq \alpha] = 1 - \sum_{k=0}^{\alpha-1} \frac{(\lambda t)^k}{k!}e^{-\lambda t}(t > 0)$$

This is the Gamma distribution. More explicitly we can calculate the pdf of $T_\alpha$ which is

$$f_{T_\alpha}(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)}t^{\alpha-1}e^{-\lambda t}$$

Therefore, we see $T_\alpha - T_{\alpha-1}$ should be $T$ because the Poisson process is uniform. Therefore, the sum of $\alpha$ exponential distribution $Exp(\lambda)$ is $Gamma(\alpha, \lambda)$.

Let us define $X_\alpha = T_\alpha - T_{\alpha-1}$ for $\alpha \geq 2$ and $X_1 = T_1$. Then

$$\begin{aligned} P(X_2 > t | X_1 = s) &= P(\text{no occurrence in}(s, s+t] | \text{one occurrence}(0, s]) \\ &= P(\text{no occurrence in } (s, s+t]) = P(N_t = 0) = e^{-\lambda t} \end{aligned}$$

by the homogenity of Poisson.

Gamma is the sume of iid Exp can be shown by moment generating function. $M_x(t) = (1 - t/\lambda)^{-1}$ and $M_{T_\alpha} = (1 - t/\lambda)^{-\alpha}$.

Special case: when $\lambda = 1/2$ and $\alpha = \gamma/2$ Then $T_\alpha \sim \chi^2_\gamma$, the **Chi squared** distribution. with pdf

$$f(x) = \frac{1}{2^{\gamma/2}\Gamma(\gamma/2)}x^{\gamma/2-1}e^{-x/2}$$

**Lemma** (connection to standard model)

If $Z_i$ are iid $N(0, 1)$ for $i = 1, \ldots, r$. Then $X_r = Z_1^2 + \ldots Z_r^2 \sim \chi^2_r$ Can be proved by moment generating function.

---

**Multivariate normal distribution**

Standard multivariate normal $Z = (Z_1, \ldots, Z_p)^T$ where $Z_1 \ldots Z_p$ are iid $N(0,1)$.

Define [Multivariate normal]

$Z$ is $g$dim standard normal. Let $Y = \vec{\mu} + \mathbf{BZ}$. Then $\mathbb{E}[Y] = \vec{\mu}$ and $Cov(Y) = BCov(Z)B^T = BB^T$. We say $Y$ is multivariate normal with mean vector $\vec{\mu}$ and covariance matrix $\Sigma = BB^T$ Write $Y \sim N(\vec{\mu}, \Sigma)$.

Remark (1) $Y$ may not have a density function and if $g < p$, we have degenerate normal.

(2) The distribution of $Y$ depends on $B$ only through $BB^T$.

For example,

$$Y = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} z_1 + z_2 \\ z_2 + 2z_3 \end{pmatrix}$$

is a degenerate bivariate normal.

Def A symmetric $p \times p$ matrix $\Omega$ is positive semi-definite if $Z^T \Omega Z \geq 0$ for $\forall Z \in \mathbb{R}^p$. Positive definite if $Z \in \mathbb{R}^p$ $Z^T \Omega Z > 0$.

Property: $\Sigma$ is psd.

proof: $Z^T \Sigma Z = Z^T BB^T Z = (B^T Z)^T (B^T Z) = \|v\|_2^2 \geq 0$ for $v = B^T Z$.

Theorem: Spectral decomposition for symmetric matrices

Let $\Omega$ be a symmetric $p \times p$ matrix, then there exists a $p \times p$ orthogonal matrix $\Gamma$ and a unique diagonal matrix $\Lambda$ such that $\Omega = \Gamma \Lambda \Gamma^T$

The columns of $\Gamma = (\gamma_1, \ldots, \gamma_p)$ are eigen vectors of $\Omega$ associated with eigenvalues. Also write $\Omega = \sum \lambda \gamma_i \gamma_i^T$.

---

**Feb 18, 2020**

Multivariate normal distribution with $Z = (Z_1, Z_2, \ldots, Z_p)^T$ for $Z_1, \ldots, Z_p$ i.i.d. in $N(0,1)$.

Def $Y = \mu + BZ$ . and say $Y \sim N(\mu, \Sigma)$ for $\Sigma = BB^T$.

For symmetric matrix $\Omega$, which may be also positive semi-definite, we can do spectral decomposition. $\Omega = \Gamma \Lambda \Gamma^T$ for $\lambda_i \geq 0$.

(i) If $\Omega \succeq 0$, we have $\lambda_i \geq 0$.

This is because if $\Omega \succeq 0$, $z^T \Omega z = z^T \Gamma \Lambda \Gamma^T z = \sum_i \lambda_i \|x_j\|^2 \geq 0$. This happens if and only if all $\lambda_i \geq 0$ since this holds for arbitrary $z$ and thus $x_i$.

(ii) $\operatorname{tr} \Omega = \sum_j \Omega_{jj}$ and $\det(\Omega) = \det(\Lambda) = \prod_{j=1}^p \lambda_j$

(iii) $\Omega$ is invertible $\iff$ $\lambda_j \neq 0$ for $j = 1, \ldots, p$.

(iv) $\Omega \succeq 0$ then $\Omega^{1/2} = \Gamma \Lambda^{1/2} \Gamma^T$. Thus, $\Omega^{1/2} \Omega^{1/2} = \Omega^{1/2} (\Omega^{1/2})^T = BB^T$ which means $B = \Omega^{1/2}$ **Remark**: $\Omega^{1/2}$ is the square root of $\Omega$.

If $B$ is not required to be symmetric there are infinitely many $B$ s.t. $BB^T = \Omega$. Since $B = \Omega^{1/2} \Phi$ for $\Phi$ orthogonal.

For example, $X_1, \ldots X_n$ are iid normal $\mu, \sigma^2$, then $(X_1, \ldots, X_n)$ are $N(\mu 1_N, \sigma^2 I_n)$. Then define $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i = d^T X$ for $d = (1/n, \ldots, 1/n)^T = \frac{1}{n} 1_n$. Then $\bar{X}_n \sim N(d^T \mu 1_n, d^T (\sigma^2 IN) d)$

$$d^T\mu 1_n = \frac{\mu}{n}1_n^T 1_n = \mu \quad d^T(\sigma^2 I_n)d = \frac{\sigma^2}{n^2}1_n^T 1_n = \frac{\sigma^2}{n}$$

Thus, $\bar{X}_n$ is $N(\mu, \sigma^2/n)$.

---

Marginal:

Let

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

For $Y_1$ as $p_1$ dimensional vector and $p_2$ dimensional vector. Then $Y_1 \sim N(\mu_1, \Sigma_{11})$ and $Y_2 \sim N(\mu_2, \Sigma_{22})$

proof: $Y_1 = (I_p, 0)Y$ thus $Y_1 \sim N\left((I_p, 0) \cdot \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, (I_p, 0)\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\begin{pmatrix} I_p \\ 0 \end{pmatrix}\right)$ and similarly for $Y_2$.

---

Independence:

property: $Y = \binom{Y_1}{Y_2}$ is $p_1 + p_2$ dim multivariate normal then $Y_1$ and $Y_2$ are independent if and only if $\Sigma_{12} = 0$.

proof: Since

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

and

$$B = \begin{pmatrix} \Sigma_{11}^{1/2} & 0 \\ 0 & \Sigma_{22}^{1/2} \end{pmatrix}$$

for $BB^T = \Sigma$. Then

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \Sigma_{11}^{1/2} & 0 \\ 0 & \Sigma_{22}^{1/2} \end{pmatrix}\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$$

for $Z_1$ and $Z_2$ independent standard normal. Thus, $Y_1$ and $Y_2$ are independent as $Z_1$ and $Z_2$ are independent.

Remark: if $\Sigma_{ij} = 0$ for $\forall i \neq j$ then $Y_1, \ldots, Y_k$ are mutually independent. and $\Sigma_{ij} = 0$ then $Y_i$ and $Y_j$ are independent.

---

Sample mean and sample variance:

recall if $X_i$ are iid of $N(\mu, \sigma^2)$ Then $\bar{X}_n = \frac{1}{n}\sum_i x_i$ is $N(\mu, \sigma^2/n)$, and $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ is Normal $(0, 1)$. $Z$ is a pivotal quantity (it has a known distribution).

If $\sigma$ is known then $\mathbb{P}(-Z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{N}} \leq Z_{\alpha/2}) = 1 - \alpha$

what if $\sigma$ is unknown? If $\sigma$ is unknown, then substitute $\sigma$ by estimator $\hat{\sigma}$

$$\hat{\sigma}^2 = S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad (\textbf{MLE})$$

and

$$\hat{\sigma}^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad (\textbf{unbiased})$$

What is the distribution of $\frac{\bar{X}_n - \mu}{\hat{\sigma}/\sqrt{n}}$ ?

Consider joint distribution of $(X_n, U_n)$ for

$$U_n = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = (n-1)S_n^2, \qquad U_n = \left\|(X_1 - \bar{X}_n, X_2 - \bar{X}_n, \ldots, X_n - \bar{X}_n)\right\|^2.$$

Define the vector

$$(X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n) = X - 1_n\bar{X}_n = I_nX - 1_n(\frac{1}{n}1_n^T X) = (I_n - \frac{1}{n}1_n^T 1_n)X,$$

We know that $X = (X_1, \cdots, X_n)^T \sim N(\mu 1_n, \sigma^2 I_n)$.

**Feb 20th, 2020**

Define $H_n = I_n - \frac{1}{n}1_n 1_n^T$, in matrix form it is

$$\begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \ldots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \ldots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & \ldots & \ldots & 1-\frac{1}{n} \end{pmatrix}$$

Properties of $H_n$ :

1. $H_n 1_n = 0_n$
2. $H_n H_n = H_n$ Idenpotent That is $H_n H_n x = H_n x$.

Back to the problem, we have

$$Y_{n+1} = \begin{pmatrix} \bar{X}_n \\ H_n X \end{pmatrix} = \begin{pmatrix} \frac{1}{n}1_n^T \\ H_n \end{pmatrix} X \equiv BX$$

Then we can get the mean of $Y$ as

$$\mu_Y = \mathbb{E}Y = B1_n\mu = \begin{pmatrix} \frac{1}{n}\frac{1}{n}1_n^T \\ H_n \end{pmatrix} 1_n\mu = \begin{pmatrix} \mu \\ 0_n \end{pmatrix}$$

And the covariance of $Y$ is

$$\Sigma_Y = cov(Y) = \begin{pmatrix} \frac{1}{n}1_n^T \\ H_n \end{pmatrix} \sigma^2 I_n \begin{pmatrix} \frac{1}{n}1_n & H_n \end{pmatrix} = \sigma^2 \begin{pmatrix} \frac{1}{n} & \frac{1}{n}1_n^T H_n \\ \frac{1}{n}H_n 1_n & H_n H_n \end{pmatrix}$$

Thus, $Y \sim N\left(\begin{pmatrix} \mu \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{n} & 0_n^T \\ 0_n & H_n \end{pmatrix}\sigma^2\right)$ Thus, $\bar{X}_n$ is independent of $H_n X$ and $\bar{X}_n \sim N(\mu, \sigma^2/n)$ and $\mu_n x \sim N(0, \sigma^2 H_n)$ .

Define Moone-Penrose inverse of a matrix. Let $\Sigma = \Gamma\Lambda\Gamma^T$ be the spectral decomposition of $\Sigma \in bR^{p\times p}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_\nu > 0 = \lambda_{\nu+1} = \ldots \lambda_p$ Then the MP inverse of $\Sigma$ is defined as

$$\Sigma^+ = \Gamma \begin{pmatrix} \Lambda_1^{-1} & 0 \\ 0 & 0 \end{pmatrix} \Gamma^T$$

Quadratic form of a singular multivariate normal $Y \sim N(\mu, \Sigma)$ and $Z^+ = (\Sigma^+)^{1/2}(Y-\mu)$ and $bE[Z] = 0$ and

$$cov(Z) = (\Sigma^+)^{1/2}\Sigma(\Sigma^+)^{1/2} = \Gamma \begin{pmatrix} I_\nu & 0 \\ 0 & 0 \end{pmatrix} \Gamma$$

Therefore $Z = \Gamma^T Z^+$ and then $Z = N(0, diag(I_\nu, 0)) = \Gamma^T(\sigma^+)^{1/2}(Y-\mu)$ This shows $\|Z\|^2 \sim \chi^2_\nu$.
Since $u_n = \|H_n x\|^2$ and $X \sim norm(0, I_n)$ Since $H_n^T = H_n$ and thus $\Lambda^2 = \Lambda$ This shows the eigenvalues are either 0 or 1. Since the eigen values are all ones or zeros, the rank is the trace of $H_n$ which is $n-1$. Thus, $H_n^+ = H_n$. Then $u_n = \|H_n X\|^2 = X^T H_n^2 X = X^T H_n^+ X$ this shows $u_n \sim \sigma^2 \chi^2_{n-1}$
Thm: If $x_i$ are iid $N(\mu, \sigma^2)$ Then $\bar{X}_n$ and $S_n^2$ are independent and

$$\bar{X}_n \sim N(\mu, \sigma^2/n) \quad \text{and} \quad \frac{n-1}{\sigma^2} S_n^2 \sim \chi^2_{n-1}$$

Define student's t-distribution: If $Z \sim N(0,1)$ and $U \sim \chi^2_\nu$ and $Z$ and $U$ are independent then $T = Z/\sqrt{u/\nu}$ is student $t_\nu$ (t with $\nu$ degrees of freedom.)
Then

$$T_n = \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{\sum((x_i - \bar{x}_n)^2}{(n-1)\sigma^2}} \equiv Z_n/\sqrt{V_n/(n-1)}$$

and $Z_n$ and $V_n$ are independent, and $V_n \sim \chi^2_{n-1}$ and $Z_n \sim N(0,1)$ which means $T_n$ is student t with $n-1$ degrees of freedom.

---

**Statistical model** $(X, \chi, P)$
$X$ is the random variable and $\chi$ is the space of $X$ and $P$ is a family of probability distributions. (usually parameterized by $\theta$)

**Feb 25th**

statistical model ($X$, $\chi$, $P$). $X$ is random variable and $\chi$ is the space of $X$ and $P$ is a family of distributions. For example $X \sim N(\mu, 1)$, $\chi = \mathbb{R}$ and $P = \{P_\mu : \mu \in R$ with

$$P_\mu(A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} dx$$

Parametric family: $P = \{p_\theta : \theta \in \Theta\}$ where $\theta$ is the parameter of the model and $\Theta$ is parameter space in $\mathbb{R}^k$ for example $X \sim N(\mu, \sigma^2)$ and $\theta = (\mu, \sigma^2)$
Point estimation:
Goal: estimate $g(\theta)$ where $g : \theta \to \mathbb{R}^g$. For example $P_\theta = N(\mu, \sigma^2)$ then $g(\theta) = \mu$ or $\sigma^2$ or $\frac{\sigma}{\mu}$
Estimator: $\delta : X^n \to \mathbb{R}^q$ where $\{X_1, X_2, \ldots, X_n\}$ a sample of size $n$ from the model $P_\theta$.
$\delta$ only depends on $(X_i)_{i=1}^n$ but cannot depend on $\theta$.
For example

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i = \delta(x_1, \ldots, x_n) \text{ estimates } \mu$$

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \text{ estimates } \sigma^2$$

$\sqrt{S_n^2}\bar{X}_n$ estimates $\sigma/\mu$.

**Plug-in Methods**

Now focus on $q = 1$ let $\theta : F \to \mathbb{R}$ be a functional on family of distributions $F$ $F \leftrightarrow P$. Let $X_1, \ldots, X_n$ iid $\sim F$. Given $F \in \mathcal{F}$, $\theta = \theta(\mathcal{F})$ can be viewed as a parameter to be estimated.

For example $\theta = \mathbb{E}[X] = \int x dF(x)$ is the mean and $\theta = \inf\{x \in \mathbb{R} : F(x) \geq \frac{1}{2}\}$ is the median.

Those integrals are Lebesgue-Stieltjes integrals allowing both continuous and discrete cdf F.

Assume $X_1 \ldots X_n$ are iid F then $F(x) = \mathbb{P}(X \leq x) = \mathbb{E}[1(X \leq x)]$is a long term frequency of event $\{X \leq x\}$.

Let

$$F_n(x) = \frac{1}{n}\text{num}(X_i \leq x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)$$

be the empirical distribution function of the sample $X_1 \ldots X_n$ that is $F_n(x)$ is the observed frequency of values that are $\leq x$.

Then $P(\lim F_n(x) = F(x)) = 1$ so $F_n(x)$ is a good estimator of $F(x)$.

To estimate $\theta(F)$ we can use the plug-in estimator where we just replace $\theta(F_n)$ by substituting $F$ with $F_n$.

For example

$$\theta = \int x dF(x) \leftarrow \int x dF_n(x) = \theta(F_n)$$

The question is how to evaluate $\int x dF_n(x)$ ?

$F_n$ is piecewise constant function right continuous,

Then

$$\theta(F_n) = \int x dF_n(x) = \int x d(\frac{1}{n}\sum 1(X_i \leq x)) = \sum_{i=1}^{n} \frac{1}{n} \int x d(1(X_i \leq x))$$

Note that $1(X_i \leq x)$ is the cdf of the point mass distribution $\delta_{x_i}$ at $x_i$. Then

$$\int x d1(X_1 \leq x) = E_{\delta_{X_i}}[x] = X_i$$

This means

$$\theta(F_n) = \frac{1}{n}\sum X_i = \bar{X}_n$$

which is the sample mean.

$$\theta(F) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$$

is the variance and plugin estimator:

$$\theta(F_n) = \int x^2 dF_n(x) - \bar{x}_n^2$$

and

$$\int x^2 dF_n(x) = \int x^2 d\left(\frac{1}{n}\sum 1(X_i \leq x)\right) = \frac{1}{n}\sum \int x^2 d1(X_i \leq x) = \frac{1}{n}\sum x_i^2$$

and thus

$$\theta(F_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_i)^2$$

**Method of Moments (Karl Pearson)** Let $X_1 \dots X_n$ iid $f(x|\theta)$ which is pdf if it is continuous and pmf if it is discrete. Let $\theta = (\theta_1, \dots, \theta_k)$ Then we can equate sample moments with true moments

$$\begin{aligned} m_1 = \frac{1}{n}\sum_i x_i \quad &= \mu_1 = \mathbb{E}[x] = \int x f(x|\theta) dx \\ m_2 = \frac{1}{n}\sum_i x_i^2 \quad &= \mu_2 = \mathbb{E}[x^2] = \int x^2 f(x|\theta) dx \\ \vdots \qquad\qquad &\quad \vdots \end{aligned} \tag{0.2}$$

Since there are $k$ moments and $k$ unknowns and we can solve for all $\theta$'s.
For example $X_i$ iid $Bin(k,p)$ and $\theta$ is $k, p$ unknowns. Then $P(X_1 = x) = \binom{k}{x} p^x (1-p)^{k-x}$ and $\mathbb{E}[p] = kp$ and $M(t) = (pe^t + 1 - p)^k$ Thus, $\mathbb{E}[x^2] = k^2p^2 + kp(1-p)$ Then we have two equations

$$\begin{aligned} &\bar{X} = kp \\ &\frac{1}{n}\sum_i x_i^2 = k^2p^2 + kp(1-p) \end{aligned} \tag{0.3}$$

**Feb 27th**

Method of moments, we get samples of $x_1, \dots, x_n \sim f(x|\theta)$ for $\theta = (\theta_1, \dots, \theta_k)$. Then we match

$$\begin{aligned} m_1 &= \frac{1}{n}\sum x_i = \mu(\theta) = \mathbb{E}[x_i] \\ m_2 &= \frac{1}{n}\sum x_i^2 = \mu_2(\theta) = \mathbb{E}[x_i^2] \\ &\vdots \\ m_k &= \frac{1}{n}\sum x_i^k = \mu_k(\theta) = \mathbb{E}[x_i^k] \end{aligned}$$

For example we have $x_i \sim Bin(k.p)$ $\theta = k, p$, we have

$$\begin{cases} \bar{x}_n = kp \\ \frac{1}{n}\sum x_i^2 = k^2p^2 + kp(1-p) \end{cases}$$

The solution is

$$\begin{cases} \hat{p} = 1 - \dfrac{\sum (x_i - \bar{x})^2}{n\bar{x}} \\ \hat{k} = \dfrac{\bar{X}}{\hat{p}} \end{cases}$$

Problem: the $\hat{p}$ may be below 0 and thus it will only give some crude estimate.

---

**Generalized method of moments**
Idea: instead of exact equality, we minimize weighted least square.
For example $x_i \sim Poisson(\lambda)$, we can choose to solve $\lambda$ by sample mean **or** sample variance. by

$$\begin{aligned} &\frac{1}{n}\sum x_i = \lambda \\ &\frac{1}{n}\sum x_i^2 = \lambda + \lambda^2 \end{aligned}$$

which may give unreasonal solution and may not hold simultaneously. Thus, instead of solving exact solution, we minimize the squared residual by some weight matrix $W^{-1}$

$$\hat{\lambda} = \arg\min_{\lambda \geq 0} \begin{bmatrix} \bar{x} - \lambda & \frac{1}{n}\sum x_i^2 - \lambda^2 \end{bmatrix} W^{-1} \begin{bmatrix} \bar{x} - \lambda \\ \frac{1}{n}\sum x_i^2 - \lambda^2 \end{bmatrix}$$

Notice here, we can introduce constraints $\lambda \geq 0$. We thus solve an optimization problem. Here we can set $W = I_2$ and we solve

$$\min(\bar{x} - \lambda)^2 + (\frac{1}{n}\sum x_i^2 - \lambda - \lambda^2)^2$$

In general cases, we can construct different loss function. For example, $l(x, \theta) = (x - \theta)^2$ will give

$$\bar{\mu} = \arg\min \sum_i^n (x_i - \mu)^2 = \bar{x}_n$$

and for $\hat{\mu} = \arg\min \sum |x_i - \mu|$ will give sample median. This is because the derivate of the loss function is sign function. To get the derivative zero, we will need half of sample below $\mu$ and half of sample larger than $\mu$.

**Application: Linear Regression**

Assume $y_i = \alpha + \beta x_i + \epsilon_i$ and assume $\mathbb{E}[y_i|x_i] = \alpha + \beta x_i$ and $\text{Var}(y_i|x_i) = \sigma^2$. or $\mathbb{E}[\epsilon_i|x_i] = 0$ and $\text{Var}(\epsilon_i|x_i) = \sigma^2$ $X$ is called covariance and $Y$ is called response.

Least square estimator:

$$\hat{\alpha}, \hat{\beta} = \arg\min_{\alpha,\beta} \sum_i (y_i - \alpha - \beta)^2$$

To minimize, we need to take derivatives, which will give

$$\frac{\partial}{\partial \alpha} f(\alpha, \beta) = 2\sum (y_i - \alpha - \beta x_i)(-1) = 0$$
$$\frac{\partial}{\partial \beta} = 2\sum (y_i - \alpha - \beta x_i)(-x_i) = 0$$

We thus, need to solve

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \sum x_i \\ \sum x_i y_i \end{pmatrix}$$

The matrix on the left hand side needs to be invertible to give us solution.

$$\det(A) = n\sum x_i^2 - (\sum x_i)^2 = \sum_{i<j} (x_i - x_j)^2 = 0$$

only occurs if $x_i = C$ for all samples. (this is the two D case conclusion where we cannot determine the line if all points are on a point since there will be infinitely many lines passing through $C$.)

---

**Maximum Likelihood estimation**

$X_i \sim f(X|\theta)$ and the likelihood function which is a function of $\theta$ is

$$L(\theta|X) = f(X|\theta) = \prod_{i=1}^n f(x_i|\theta)$$

Remark: for $\theta_1, \theta_2$, if $L(\theta_1|x) > L(\theta_2|x)$ means $\theta_1$ is more plausible (not more probable) than $\theta_2$.
Maximum Likelihood estimator is

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta|X)$$

we can maximize in two senses:

(1) First order condition ($\partial_\theta L(\theta|x) = 0$)

(2) Direct maximization (graph, inequalities)

For (1) we usually take log and then take derivative. because we transfer multiplication to summation which is easier to take derivatives.

$$l(\theta|x) = \log L(\theta|x) = \sum_i \log f(x_i|\theta)$$

we require

$$\frac{\partial\, l(\theta|x)}{\partial \theta_j} = 0$$

but when $\partial^2/\partial\theta^2 < 0$, this is the real local maximum, when $\partial^2/\partial\theta^2 > 0$, this is local minimum and when it is zero, the location is a saddle point.
For example $X_i$ is iid of $N(\theta, 1)$ for $\theta \in \mathbb{R}$. Then

$$L(\theta|x) = \prod_{i=1} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2} \quad \Rightarrow \quad l(\theta|x) = -\frac{1}{2}\sum (x_i - \theta)^2$$

Then taking derivatives, we can set it to zero as

$$\sum x_i - n\theta = 0 \quad \Rightarrow \quad \hat{\theta}_{\text{MLE}} = \bar{x}_n$$

**March 3rd, 2020**

$X_i \sim f(X|\theta)$, $X = (x_1, \ldots, x_n)$ Then the Liklihood function is $L(\theta|X) = f(x|\theta) = \prod_{i=1}^n f(x_i|\theta)$
The MLE is $\hat{\theta} = \arg\max L(\theta|x)$. The log-likelihood is $l(\theta|x) = \sum_{i=1}^n \log f(x_i|\theta)$.
For example $x_i \sim N(0,1)$ $\theta \in \mathbb{R}$ then $\hat{\theta}_{\text{MLE}} = \bar{X}_n = \frac{1}{n}\sum x_i$
When $\Theta = [0, \infty)$ we have

$$\max_{\theta \geq 0} L(\theta|x) = \max\{-\sum (x_i - \theta)^2\} = \max\{-n(\theta - \bar{x}_n)^2 + n\bar{X}_n^2\}$$

case1: if $\bar{X}_n \geq 0$ then $\hat{\theta} = \bar{x}_n$
case2: if $\bar{X}_n < 0$ then $\hat{\theta} = 0$.
For example $X_i$ is iid $Ber(p)$ , $p \in [0,1]$ then

$$L(p|X) = \prod p^{x_i}(1-p)^{1-x_i}$$

Then if we define $y = \sum x_i$ and we will have

$$l(p|x) = \log L(p|x) = y \log p + (n-y)\log(1-p)$$

Then we can maximize by taking derivative to get

$$\frac{y}{p} - \frac{n-y}{1-p} = 0 \quad \Rightarrow \hat{p} = \frac{y}{n}$$

Consider extreming cases. When $p = 0$, we set $l(0|X) = -\infty$ when $y \neq 0$ and $L(\theta|x) = 0$ when $y = 0$. Thus, $l(\hat{p}|x) \geq l(0|x))$. When $p = 1$, we have $l(1|x) = -\infty$ for $y \neq n$ and 0 when $y = 0$. Then $l(\hat{p}|x) \geq l(1|\theta)$. Thus

$$l(\hat{p}|x) \geq \max\{l(0|x), l(1|x)\}$$

Now consider multi-parameter cases. For example $X_i$ are iid $N(\mu, \sigma^2)$ Then we have

$$L(\theta|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{(x_i-\mu)^2/2\sigma^2}$$

Thus, the loglikelihood is

$$l(\theta|x) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^4}\sum(x_i - \mu)^2$$

Then taking partial differentials, we have

$$\frac{\partial l(\theta|x)}{\partial \mu} = -\frac{1}{2\sigma^2}\sum(x_i - \mu)(-1) = 0$$
$$\frac{\partial l(\theta|x)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^2}\sum(x_i - \mu)^2 = 0$$

Solve it we have

$$\hat{\mu} = n^{-1}\sum x_i \quad \hat{\sigma}^2 = n^{-1}\sum(x_i - \bar{x}_n)^2 = s^2$$

---

**profile likelihood**

$\theta = (\psi, \lambda)$ nuisance parameter is lambda. Fix $\psi$, maximize $L(\psi, \lambda)$. Then we will have a one-dimensional optimization which simplifies the problem. Then $\hat{\lambda}_\psi = \max_\lambda L(\psi, \lambda)$ Then $L(\psi, \hat{\lambda}_\psi)$ is called the profile likelihood.

Then maximize the profile likelihood

$$\hat{\psi} = \arg\max_{\psi} L(\psi, \hat{\lambda}_\psi)$$

Property: $(\hat{\psi}, \hat{\lambda}_{\hat{\psi}}))$ is identical to the MLE $(\hat{\psi}_{\text{MLE}}, \hat{\lambda}_{\text{MLE}})$

---

**Expectation Maximization (EM) algorithm** data contains missing data or latent variables.

for example (missing data): $Y_1, Y2, \ldots, Y_5$. is multinomial$(n, \pi_\theta)$ and $\pi_\theta = (\frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1+\theta), \frac{1}{4}\theta)$. However, the values of $(Y_1, Y_2)$ are missing, but know $(Y_1 + Y_2, Y_3, Y_4, Y_5)$.

for example (latent variable problem)

For each observation $X_i$, the latent variable $Y_i$ is not observed. such that $X_i|Y_i = N(\mu_j, \sigma_j^2)$.

statistical model $(X_i, Y_i)$ and $X_i$ are iid in the marginal distribution of $X$. That is $\sum_j P(x_i|y_i)P(Y_i = j)$ we assume $P(Y = j) = p_j$ and thus the parameter

$$\theta = (\mu_1, \mu_2, \ldots, \mu_m, \sigma_1^2, \ldots, \sigma_m^2, p_1, \ldots, p_m)$$

**March 5th,2020**

Continue from above. We usually, assume $X_i|Y_i \sim N(\mu_k, \sigma_k^2)$ and $P(Y_j = k) = P_k$.
Statistical models: $(X_i, Y_i)$ are iid random variables $P_\theta(x, y)P_k N(\mu_k, \sigma_k^2)$. This will lead to complete data log-likelihood function. $l_\theta = \log P_\theta(x, y)$.
Another models: consider marginal distributions. Then $x_i$ are iid of $P_\theta = \int p_\theta(x, y)dy$. Thus,

$$P_\theta(x) = \sum_{k=1}^{k} P_k \frac{1}{\sqrt{2\pi\sigma_k}} e^{-(x-\mu_k)^2/2\sigma_k^2}$$

,

Although the log likelihood is computable, it is hard to optimize since it contains log of sums. The estimator is

$$\hat{\theta}_{\text{MLE}} = \arg\max \log P_\theta(x)$$

Lemma: (Jensen's inequality)
Let $f(x)$ be a convex function, i.e. $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$ for all $\alpha \in (0, 1]$. Then $\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$.
We would like to maximize

$$\begin{aligned} l(\theta; x) &= \log \int P_\theta(x, y)dy = \log \int_y \frac{P_\theta(x, y)}{q(y)} q(y)dy \\ &= -\left(-\log \mathbb{E}_q[\frac{P_\theta(x, y)}{q(y)}]\right) \geq -\mathbb{E}_q[-\log P_\theta(x, y) + \log q(y)] \\ &= \int q(y) \log P_\theta(x, y)dy - \int q(y) \log q(y)dy \end{aligned}$$

This means we can maximize $\mathbb{E}_q[\log P_\theta(x, y)]$ as a function of $\theta$.
When is the lower bound accurate? we will achieve $=$ if and only if $p_\theta(x, y)/q(y) = c(x)$ this implies, $q(x) = p_\theta(x, y)/c(x)$ and thus, choose $c(x) = p(x)$ and $q(x) = p_\theta(y|x)$.
EM algorithm: Iterative algorithm,
E-step: Define $Q = Q(\theta, \theta_{n-1}) = \mathbb{E}[\log P_\theta(x, y)]|x, \theta_{n-1}] = \int \log P_\theta(x, y) P_{\theta_n}(y|x)$
M-step: we set $\theta_n = \arg\max_\theta Q(\theta, \theta_{n-1})$
This is a special case of minorize-maximization algorithm.

Example: missing data problem.

$$l(\theta; x, y) = C + \log\left[\left(\frac{1}{2}\right)^{y_1} (\frac{1}{4}\theta)^{y_2} \left(\frac{1}{4}(1-\theta)\right)^{y_3} \left(\frac{1}{4}(1-\theta)\right)^{y_4} (\frac{1}{4}\theta)^{y_5}\right]$$

$$l(\theta; x, y) = C' + y_2 \log\theta + (y_3 + y_4)\log(1-\theta) + y_5 \log\theta$$

and

$$p_\theta(y|x) = p_\theta(y_2|y_3, y_4, y_5) = bin(y_1 + y_2, \frac{\theta/4}{\frac{1}{2} + \frac{\theta}{4}}$$

E-step:

$$Q(\theta, \theta_o) = c' + (y_1 + y_2)\frac{\theta_o}{\theta_o + 2} \log\theta + (y_3 + y_4)\log(1-\theta) + y_5 \log\theta$$

we can define $p_{\text{old}} = \theta_o/(2+\theta_o)$ M-step: maximize $Q(\theta, \theta_o)$ to find $\theta$. We thenhave

$$\theta_{\text{new}} = \frac{y_5 + \theta_o(y_1+y_2)}{y_3+y_4+y_5+p_{\text{old}(y_1+y_2)}}$$

The optimized $\theta$ is the fixed point of the above equation.
We can also find $\theta$ by considering maximize $\theta$ for data $(y_1+y_2, y_3, y_4, y_5)$.

**March 26**

**Markov chain Monte Carlo**

Goal: Given a multivariate distribution $\pi(x)$, $x \in X$, generate random samples $x_1, x_2, \ldots, x_m$ from $\pi$

Direct samples (inverse transform sampling),$\pi(x)$. posterior is known up a constant $\pi(x) \sim e^{-u(x)}$.

$$\pi(\theta|x) = e^{-u(x)}/z = \frac{p_\theta(x)\pi(\theta)}{\int P_\theta(x)\pi(\theta)d\theta} = \frac{p_\theta(x)\pi(\theta)}{m(x)}$$

$u(x)$ : negative log-likelihood function + negative log pair. Need to compute $m(x)$ to do inverse sample.
Def (Markov chain) A sequence of r.v. $X_1, X_2, X_3, \ldots$ on a discrete state space $\Omega$ is called a Markov chain if

$$P(X_{t+1}|X_t = x_t, X_{t-1} = x_{t-1}, \ldots X_1 = x_1) = P(X_{t+1} = x_{t+1}|X_t = x_t)$$

Then we have factorization formula:

$$\begin{aligned}
&P(X_1 = x_1, \ldots, X_t = x_t) \\
&= P(X_t = x_t|X_1 = x_1, \ldots, X_{t-1} = x_{t-1}) \to P(X_t = x_t|X_{t-1} = x_{t-1}) \\
&\times P(X_{t-1}|X_1 = x_1, \ldots x_{t-2} = x_{t-2}) \to P(X_{t-1} = x_{t-1}|X_{t-2} = x_{t-2}) \\
&\times \vdots
\end{aligned}$$

We can summarize $P(\cdot|\cdot)$ into a $M \times M$ matrix $\Omega = \{1, 2, \ldots, M\}$

$$P = \begin{pmatrix} p(1|1) & p(2|1) & \ldots & p(M|1) \\ p(1|2) & p(2|2) & \ldots & p(M|2) \\ \vdots & & \ddots & \vdots \\ p(1|M) & p(2|M) & \ldots & p(M|M) \end{pmatrix}$$

start from $i$, $P(X_t = j|X_1 = i) = [P^{t-1}]_{ij}$.

Define A stationary distribution of a Markov chain is a distribution $\pi$ on $\Omega$ such that $\pi(y) = \sum_{x\in\Omega} p(y|x)\pi(x)$
Suppose we start from a random initial state $X_1 \sim \pi$

$$P(x_2 = y) = \sum_{x\in\Omega} p(x_2 = y, x_1 = x) = \sum_{x\in\Omega} P(x_2 = y|x_1 = x)p(x_1 = x) = \pi(y)$$

here the last equal sign is from the definition of stationary distribution. $X_2$ also has distribution of $\pi$. More generally $X_t \sim \pi$ for $\forall t \geq 1$.

Def: A Markov chain is called ergodic if there exists $r \in \mathbb{N}$ s.t. $P^r > 0$ (elementwise compare).
Check two points: 1. Irreducible condition: $\forall x, y \in \Omega$ , $\exists r(x, y)$, s.t. $P^{r(x,y)}(x|y) > 0$. 2. Aperiodic condition: $\forall x \in \Omega$. $GCD\{r : p^r(x, x) > 0\}$.
Theorem: a finite ergodic markov chain has a unique stationary distribution $\pi$. In addition for $x_0 \in \Omega$, $P(X_{t+1} = \cdot|X_1 = x_0) = p^t(\cdot, x_0) \to \pi(\cdot)$. as $t \to \infty$.

**March 31st, 2020**
Def: A Markov Chain is reversible if exists distribution $\pi$ on $\Omega$ s.t.

$$p(x|y)\pi(y) = p(y|x)\pi(x)$$

this is the detailed balance equation.
property: such a $\pi$ must be a stationary distribution of the Markov chain.
proof: $\sum_y p(x|y)\pi(y) = \sum_y p(y|x)\pi(x) = \pi(x) \sum_y p(y|x) = \pi(x)$.
In equilibrium, total mass from $x \to y$ $(\pi(x) \cdot p(y|x))$ is equal to total mass from $y \to x$ $(\pi(y) \cdot p(x|y))$.
Not all Markov Chain is detailed balanced. However, we can design detailed balance Markov Chain to get stationary distribution $\pi(x)$.
Question: How can we design a Markov Chain transition probabilities such that detailed balanced equation holds with our target $\pi$?

---

**Metropolis-Hasting Algorithm**
Let $\pi(x) = e^{-u(x)}/Z$. Construct a reversible Markov Chain as follows: Let $x_t = x$ be the current state, we will perform the following two steps repeatedly:

- Generate $Y = y \sim Q(y|x)$ for some Markov transition matrix $Q$, where $Q(\cdot|x)$ as the proposal distribution.
- Set $x_{t+1} = y$ with probability

$$\alpha(y|x) = \min\{1, \frac{\pi(y)Q(x|y)}{\pi(x)Q(y|x)}\}$$

  $\alpha$ is acceptance probability. $\frac{\pi(y)Q(x|y)}{\pi(x)Q(y|x)}$ is acceptance ratio.

Remark: The acceptance ratio does not depend on normalization constant $Z$, that is we can run this algorithms without knowing $Z$.
In practice, in step 2, we can generate $U \sim unif(0, 1)$ and set $X_{t+1} = y$ if $u < \frac{\pi(y)Q(x|y)}{\pi(x)Q(y|x)}$ and set to $x$ otherwise.
typical choise of $Q(y|x)$ is the normal distribution $Q(\cdot|x) = \mathcal{N}(x, \sigma^2)$.
If $\alpha(y|x) \ll 1$, then the Markov chain moves very slowly. On the contrary, if $\alpha(y|x) \sim 1$, then the step size is usually very small. Gelman Gilks and Roberts show 0.234 is usually the optimal acceptance rate.
Theorem: $\pi$ is the stationary distribution of the resulting Markov chain. proof: We have

$$p(y|x) = \begin{cases} Q(y|x)\alpha(y|x) & y \neq x \\ 1 - \sum_{z \neq x} Q(z|x)\alpha(z|x) & y = x \end{cases}$$

when $y = x$, the detailed balance is obvious.
If $y \neq x$, $p(y|x)\pi(x)$ as a function of $(x, y)$ is symmetric in them. Then

$$\begin{aligned}\pi(x)p(y|x) &= \pi(x)Q(y|x)\min\left\{\frac{\pi(y)Q(x|y)}{\pi(x)Q(y|x)}, 1\right\} \\ &= \min\{\pi(y)Q(x|y), \pi(x)Q(y|x)\} \\ &= \min\left\{\frac{\pi(x)Q(y|x)}{\pi(y)Q(x|y)}, 1\right\}\pi(y)Q(x|y) \\ &= p(x|y)\pi(y)\end{aligned}$$

detailed balance equation holds for $\pi(x) \Rightarrow \pi$ is the stationary distribution of the Markov Chain.

---

**Gibbs Sampling**
Consider $\Omega = \mathbb{R}^2$. $\pi(x_1, x_2)$ will the stationary distribution we want to approximate. Consider, $X_t = (x_t^1, x_t^2)$ repeatedly perform the following step:

1. sample $x_{t+1}^1 \sim \pi(x^1|x^2 = x_t^2)$
2. sample $x_{t+1}^2 \sim \pi(x^2|x^1 = x_{t+1}^1)$

This can be viewed as a special case of Metropolis algorithm when the acceptance probability is always one.
rough proof: proposal $Q((x_{t+1}^1, x_t^2)|(x^1, x^2)) = \pi(x_{t+1}^1|x^2)$. Then

$$\alpha = \frac{\pi(x_{t+1}^1, x^2)\pi(x^1|x^2)}{\pi(x^1, x^2)\pi(x_{t+1}^1|x^2)} = \frac{\pi(x_{t+1}^1, \pi^2)\pi(x^1, x^2)/\pi(x^2)}{\pi(x^1, x^2)\pi(x_{t+1}^1, x^2)/\pi(x^2)} = 1$$

---

**April 2nd, 2020**

**Hypothesis testing**
$(X, \chi, P)$ parametric $P = \{p_\theta : \theta \in \Theta\}$ with $\theta$ parameter. Partition $P$ into $P_1 \& P_2$ $P_1 \cap P_2 = \emptyset$ and $P_1 \cup P_2 = P$.
Example: in physics, if we want to test whether a theory is correct, we divide $H_1 : p \in P_1$ ( the theory is not correct) and $H_0 : p \in P_2$ (null hypothesis, theory is correct).
Let $X_1, \ldots, X_n \sim P$, Goal: test $H_0 : p \in P_1$, null hypothesis, and $H_1 : p \in P_2$, alternative. In parametric case, $P_1 = \{P_\theta : \theta \in \Theta_1\}$ and $P_2 = \{P_\theta : \theta \in \Theta_2\}$.
There are two ways to do hypothesis testing:

1. Frequentist framework: come up with a test statistic T. find some cut off value t reject $H_0$ if $T \geq t$.
2. Bayesian: prior probability on $H_0$ and $H_1$. In parametric case, consider a prior $\Pi$ on $\Theta$. Then $\pi(H_0) = \pi(\theta \in \Theta_1)$ and $\pi(H_1) = \pi(\theta \in \Theta_2)$.
   Or we specify weights on $H_0$ and $H_1$ and use $p(\theta|H_0)$ and $p(\theta|H_1)$ to get posterior prob. on $H_0$ and $H_1$. $\pi(H_0|x)$ and $\pi(H_1|x)$.

Example: $X_1, \ldots, X_n$ are iid $N(\mu, \sigma_0^2)$. $\sigma_0^2$ known $\theta = \mu$. $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$
Recall the estimator for $\mu$ is $\bar{X}_n = \frac{1}{n}\sum x_i \sim N(\mu, \sigma^2/n)$. $T(X) = |Z|$ for $Z = \frac{\bar{x}_n - \mu_0}{\sigma_0/\sqrt{n}}$ (z-statistics). under $H_0$, $Z \sim N(0,1)$ reject $H_0$ if $T(x) = |Z| > c$ for some cut-off value $c \geq 0$.

| truth | Accept $H_0$ | Reject $H_0$ |
|---|---|---|
| $H_0$ | good ✓ | false positive (Type I error) |
| $H_1$ | false negative Type II error | good ✓ |

Thus, Large value of $C$ then less rejections for $H_0$ and thus more false negatives.
small value of $C$ then more rejection for $H_0$ then more false positive,
Idea: need to control the type I error while minimize the type II error.
Def (level and size): A hypothesis test has level $\alpha \in (0,1)$ if $p_\theta(T(x) > c) \leq \alpha$ (reject null). for all $\theta \in \Theta_1$.
This is an upper bound to type I error under $H_0$.
The size of a test is the smallest $\alpha$ for which it is level $\alpha$.

$$\text{size} = \sup_{\theta \in \Theta_0} P_\theta(T(X) > c).$$

Remark:

1. A test with level 0.05 has also level 0.1.

2. level and size are often used interchangeably.

Def (power): The power of a test is a function of $\theta \in \Theta_2$ that power$(\theta) = 1 - P_\theta(T(x) \leq C)$. So to find a good test, we constrain the size of all such tests, while maximizing the power.
Example: Z-statistics, $Z = \frac{\bar{X}_n - \mu_0}{\sigma_0/\sqrt{n}}$. Under $H_0$, $Z \sim N(0,1)$ and $T(Z) = |Z|$. cutoff = c.
size $= P_{\mu_0}(T(z) > c) = P_{\mu_0}(|Z| > c) = 2[1 - \phi(c))$ where $\phi$ is the cdf of $N(0,1)$. constraint size as $\alpha$ then $c = \phi^{-1}(1 - \alpha/2) = Z_{\alpha/2}$.
Power analysis: Under $H_1 : \mu \neq \mu_0$,

$$Z = \frac{\bar{x}_n - \mu}{\sigma_0/\sqrt{n}} + \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}$$

Reject $H_0$ if $\frac{\bar{x}_n - \mu}{\sigma_0/\sqrt{n}} > Z_{\alpha/2} - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}$