1. **Moment Generating Function(MGF)**

MGF is used to generate the moment of a distribution or a random variable.

Definition 1. (Moments of a Ran.Var.) The kth moment of Ran.V Y is defined as $\mathbb{E}(Y^k)$, which exists if the expectation is finite.

Definition 2. Moment Generating Function Let Y be a random variable. The MGF for Y is defined as

$$M_Y(t) = E_Y(e^{tY})$$

if it exists for t in a neighbour of 0 (i.e, for t in the open interval t ∈ (-T,T)).

The way to find a MGF is to compute the expectation by its definition directly. So in the procedure to find a moment is: (1) Find the MGF (2) Take the kth derivative of MGF w.r.t t (3) set t to be 0. Need to consider why around t=0. Make up

Proof.

$$MGF_Y(t) = E(e^{tY}) = E(\sum_{k=0}^{\infty} \frac{t^k Y^k}{k!})$$

the summation in the bracket is the Maclaurin expansion ($x_0 = 0$).

$$= E(1 + ty + \frac{t^2 y^2}{2!} + ...)$$

$$= 1 + tE(y) + \frac{t^2}{2!}E(y^2) + ...$$

Then take the derivative w.r.t t

$$\frac{d}{dt} = E(y) + 2tE(y^2)$$

□

其实对于 moment 的定义有两种，一种是 raw momentum 原点矩，即上述讨论的，另一种是 central momentum 中心距，定义为

$$\mathbb{E}[(X-\mu)^n] = \int_{-\infty}^{\infty} (x-\mu)^n f(x)\, dx$$

相比于原点矩为

$$\mathbb{E}[X^n] = \int_{-\infty}^{\infty} x^n f(x)\, dx$$

注意在 SOA P 的考试中往往使用中心距。

(a) **Useful properties of MGF**

Lemma 1. *MGF of a linear function of a random variable*
If MGF of a Y is $M_Y(t)$, then for Ran.Varb. $Z = a + bY$ has MGF

$$M_Z(t) = e^{at} M_Y(bt)$$

Lemma 2. *MGF of a sum of independent Ran.Varb Suppose Y_i's are independent R.V. Then the MGF of R.V $X = \sum_{i=1}^{all} Y_i$ is*

$$M_X(t) = \prod_{i=1}^{n} M_{Y_i}(t)$$

proof is easy and just notice we need the R.V to be independent in order to use the $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

2. **More advanced Distribution**

Lemma 3. *Chi-square distribution* *Let each* $Z_i \sim N(0,1)$*, then their sum of square distribution*

$$X_d = \sum_{i=1}^{d} Z_i^2 \quad \sim \quad \chi_d^2$$

follows the chi-square distribution with degree of freedom d.

Lemma 4. *t distribution* *If Z ~ N(0,1),* $X_d \sim \chi_d^2$*, then*

$$\frac{Z}{\sqrt{X_d/d}} \quad \sim \quad t_d$$

follows the t distribution with degree of freedom d. 注意分子上的 *Z* 必须与分母中的 *Chi* 是相互独立的。例如假设 Z_i 都是独立的，则

$$Y = \frac{Z_1}{\sqrt{Z_1^2 + Z_2^2/2}}$$

不服从 *t* 分布。

Lemma 5. *F distribution* *Let* $X_{d1} \sim \chi_{d1}^2$*,* $X_{d2} \sim \chi_{d2}^2$ *and they are independent, then*

$$\frac{X_{d1}/d1}{X_{d2}/d2} \quad \sim \quad F_{d1,d2}$$

follows F distribution with df d1 and d2. Notice F distribution has two parameter.

Lemma 6. *Gamma distribution Let each* $Y_i \sim Exp(\lambda_0)$*. Then there summation*

$$\sum_{i=1}^{n} Y_i \sim Gamma(v = n, \lambda = \lambda_0)$$

follows the Gamma distribution with shape parameter n and rate parameter λ_0*.*

3. **Estimation of Normal distribution parameter**

Here we want to make inference of the two parameter that determines the normal distribution which are μ and σ.

This has two cases – with known σ and without. If we have already known the population σ then we are ready for estimating the population mean, which simply follows the $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$. If it is unknown, we have to approximate the population variance by

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2$$

the reason why the denominator is (n-1) is because of the unbiased estimation: We can prove that with this (n-1) the $\hat{\sigma}^2$ is an unbiased estimator of σ^2 which satisfies

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

the proof is simple and just remember to use the formula $Var(X) = E(X^2) - E(X)^2$. Then finally with this estimator, we are able to construct C.I and thus ready for inference of mean.

(a) **Distribution of standardized sample mean**
The same logic applies here – if we know the population σ then we are sure that

$$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n}) \implies \frac{\bar{Y}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

the RHS r.v. is the standardized $\bar{Y}$ value. However, if the σ is unknown, we have to use sample variance S^2 to replace σ^2 which will cause the standardized sample mean

$$\frac{\bar{Y}-\mu}{\sqrt{S^2/n}} \sim t_{n-1}$$

which is a t *t distribution* with df (n-1). Proof is on textbook page 2-12.

4. **Convergence in Distribution** 依分布收敛

Definition 3. Let $\{X_n\}$ be a sequence of random variable with CDFs $F_{X_n}(x)$ respectively. Suppose there exists a CDF $F_X(x)$ s.t

$$\lim_{n\to\infty} Pr(X_n \le x) = F_x(x)$$

for all x where $F_x(x)$ is continuous. Then we say the sequences of random variables X_n converges in distribution to $F_X(x)$.

More info maybe covered later

5. **Central limit Theorem** 同分布的中心极限定理

同分布的中心极限定理

Theorem 1. *Let Y_is' be a series of IID r.v with same mean μ and SD σ, and MGF defined in a neighborhood of zero. Define the sum as*

$$X_n = \sum_{i=1}^{n} Y_i$$

Then the standardized sum

$$Z_n = \frac{X_n - n\mu}{\sigma\sqrt{n}}$$

converges in distribution to N(0,1).

注: (1) Y_i 不一定必须要符合正态分布，任何的分布都可以应用中心极限定理。
(2) 标准化中，分子的 $n\mu$ 是 X_n 的均值或期望（容易证明因为所有的 Y_i 都是独立变量，分母同理）。
(3) 另一个版本的 CLT 是去标准化的 r.v，其比上面的版本更好用。即，让所有的信息和要求不变，然后

$$X_n \sim N(n\mu,\ \sigma^2 n)$$

it is simply the application of some characteristic of expectation and variance.

6. **Statistical Estimator**
There are mainly three properties of an estimator: Unbiasedness, Efficiency and Consistency.

(a) **Unbiasedness** 无偏性

Definition 4. Bias *The bias of the estimator $\widetilde{\theta}$ of a parameter θ is*

$$Bias(\widetilde{\theta}) = \mathbb{E}(\widetilde{\theta}) - \theta$$

or rewriting it as

$$Bias(\widetilde{\theta}) = \mathbb{E}(\widetilde{\theta} - \theta)$$

Definition 5. Mean square error(MSE) *The mean square error of* $\widetilde{\theta}$ *as an estimator of* θ *is*

$$MSE(\widetilde{\theta}) = E[(\widetilde{\theta} - \theta)^2] = Var(\widetilde{\theta}) + Bias^2(\widetilde{\theta})$$

proof is on page 3-8, make up later
无偏估计指没有系统误差，即在无数次、多次抽样后必然得到期望值 parameter。所谓无偏性即是

$$\mathbb{E}(\widetilde{\theta}) = \theta$$

由于 MSE 综合考虑了 $Biasness$ 和 $Varience$，所以可以用其作为衡量 $Accuracy$ 的指标。

(b) **Efficiency** 有效性

简而言之，对于两个无偏估计量而言，Variance 小的较 Variance 大的更有效。

(c) **Consistency** 一致性

Definition 6. Consistency The estimator $\widetilde{\theta}_n$ of a parameter θ based on a sample size n is consistent for estimating θ is

$$\lim_{n\to\infty} Pr(|\widetilde{\theta}_n - \theta| < \epsilon) = 1$$

for any fixed error, $\epsilon > 0$.

一致性的充要条件是 MSE goes to zero $\equiv Var \to$ zero & $Bias \to$ zero as $n \to \infty$. Why? Also, consistency is a special case of *convergence in probability.*

Theorem 2. *(Week law of Large Number)* 弱大数定律 $\equiv$ 样本均值依概率收敛于期望值
Let

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

where all $Y_i s'$ *are* n *independent r.v with same mean* μ *and variance* σ^2 *(both of which must exists). Then for ant* $\epsilon > 0$*,*

$$\lim_{n\to\infty} Pr(|\bar{Y}_n - \mu| < \epsilon) = 1$$

The law of large number concerns about the large sample size n instead of the large observations. proof needed. 注意，此处的加和中的 Y_i 指的并非是样本中的一个个值，而是在每一次增加了样本数目 n 后从新取值得到的新的样本均值——即所取样本不累计。所谓收敛，指一个有关样本均值的数列/函数收敛。

具体一点说——$\bar{Y}_2$ 是当取两个样本时的均值；$\bar{Y}_3$ 是取三个样本时候的均值，注意是从新取三个，与之前取得两个无关。以此类推，可以得到不同样本数目下的关于样本均值的数列，然后通过 $WLLN$ 得到

7. Maximum Likelihood Estimation

Definition 7. (likelihood function 似然函数)

Let $p(x, \theta)$ be the PMF or PDF of a r.v x and θ is the parameter of the distribution. Let $x_1, x_2, ..., x_n$ be the samples from the population. Then

$$\mathbb{L}(\theta) = \mathbb{L}(\theta|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p(x_i|\theta)$$

其实似然函数的表达式与 PMF、PDF 并无不同，只是此时将其看作是关于 θ 的函数。即此时 smaple 信息已经给定，然后通过寻找 θ_m 满足

$$\mathbb{L}(\theta_m) = max\ \mathbb{L}(\theta)$$

则称 θ_m 是 θ 的最大似然估计 (MLE)。

Definition 8. (Observed information)

Letc$f_{Y_1,Y_2,...,Y_n}(y_1,y_2,...,y_n|\theta)$ be the joint PDF or PMF of data $y_1,y_2,...,y_n$. Then the observed information for estimating the parameter θ is

$$I_n(\theta,y_1,y_2,...,y_n) = -\frac{\partial^2}{\partial\theta^2} ln\, f_{Y_1,Y_2,...,Y_n}(y_1,y_2,...,y_n|\theta)$$

and MLE as an method to estimate the real parameter, its estimator $\widetilde{\theta}$ has variance

$$Var(\widetilde{\theta}) \simeq \frac{1}{I_n(\theta,y_1,y_2,...,y_n)}$$

and thus correspondingly an estimate of the var. is

$$\widehat{Var}(\widetilde{\theta}) = \frac{1}{I_n(\hat{\theta},y_1,y_2,...,y_n)} \;\&\; se(\widetilde{\theta}) = \frac{1}{\sqrt{I_n(\hat{\theta},y_1,y_2,...,y_n)}}$$

For the $Var(\widetilde{\theta})$ we use $\simeq$ instead of $=$ since the var, as an approximation, improves as $n\to\infty$. We call this an **asymptotic** result. 渐进性结果。

其实更准确的说，此处的 $Var\widetilde{\theta}$ 是一个渐进无偏估计量（asymptotic unbiased estimator）。即当样本容量 n 趋近于无限大时近似无偏的估计量。用极限表达为

$$\lim_{n\to\infty} \mathbb{E}\hat{\theta}(X_1,X_2,...,X_n) = \theta$$

例如，样本方差的期望 $E(S_n^2) = \frac{n-1}{n}\sigma^2$ 是有偏估计，而当 $n\to\infty$ 后等于 σ，因此称为渐进无偏估计量。但注意在应用时，当 n 很大时也可应用。因此课本上称 *asymptotic* and *large sample* are interchangeable.

Definition 9. (Fisher information) Let $f_{Y_1,Y_2,...,Y_n}(y_1,y_2,...,y_n|\theta)$ be the joint PDF or PMF of data $y_1,y_2,...,y_n$. Then the fisher information for estimating the parameter θ is

$$\mathcal{I}_n(\theta) = \mathbb{E}(I_n) = \mathbb{E}\big(-\frac{\partial^2}{\partial\theta^2} ln\, f_{Y_1,Y_2,...,Y_n}(Y_1,Y_2,...,Y_n\,|\,\theta)\big)$$

Theorem 3. *(Asymptotic normality of the ML estimator) Under regularity conditions on $f_Y(y\,|\,\theta)$, the distribution of ML estimator $\widetilde{\theta}$ of θ has the following limiting standardized distribution*

$$\lim_{n\to\infty} \frac{\widetilde{\theta}-\theta}{1/\sqrt{n\mathcal{I}_1(\theta)}} \longrightarrow N(0,\,1)$$

where the $\longrightarrow$ denotes the convergence in distribution.

值得注意的是，如果 parameter θ（真实值）在其概率分布所定义的真实值的边界上，则无法预测。例如泊松分布的 $\mu>0$。如果 $\mu=0$ is of interest, 那么将无法预测，因为 C.I 会在负数一边。(actually not so clear here).搞清楚依概率收敛和依分布收敛后补充

注：似然与概率的辨析

概率与似然是对事物发生情况的两种理解视角，两者并列。如果已知总体参数 θ，通过 PDF、PMF 求出的是概率；而如果已知的是样本而未知的是总体参数 θ 则求出的是似然。

8. Bayes' Estimation

The fundamental difference of Bayes and ML estimate is that in Bayesian estimation we treat the parameter as a random variable — instead of constructing the C.I, we are going to find the probability distribution of the parameter.

注:
(1)贝叶斯估计中，综合运用了总体信息、样本信息和先验信息。

(2)Bayes 学派的基本观点是—任意未知的总体参数都可以看做随机变量，并用概率分布描述，且称该分布为先验分布 (prior probability distribution w.r.t θ)。

(3)通过 Bayes 公式综合后得到的新的分布成为后验分布 (posterior distribution)

一言以蔽之，贝叶斯估计的路径就是—结合历史信息得到 θ 先验分布 $\rightarrow$ 整合先验信息到贝叶斯公式 $\rightarrow$ 得到 θ 后验分布进行贝叶斯估计。

(a) **How sample comes – a Baysian perspective** 样本取得途径

Since we treat the parameter as a random variable, we can not think in the same way as the frequency perspective (e.g the sample is taken form a distribution with a fixed parameter). So we think of samples are generated in this way:

(1) 设想从 θ 的先验分布中取得一个样本 θ_0。

(2) 从 $p(\mathbf{X}\,|\,\theta_0)$ 中取得一组样本 $\mathbf{X} = (x_1, x_2, ..., x_n)$。则联合条件概率分布 (CDF) 为

$$p(\mathbf{X}\,|\,\theta_0) = p(x_1, x_2, ..., x_n|\,\theta_0) = \prod_{i=1}^{n} p(x_i\,|\,\theta_0)$$

该分布综合了总体和样本信息

(3) 整合先验信息。在 (2) 中的分布式并未考虑 θ 是一个 r.v，因此结合 θ 的先验分布得到的样本 $\mathbf{X}$ 和参数 θ 的联合分布 (CDF) 为

$$\boldsymbol{h}(\mathbf{X}, \theta) = p(\mathbf{X}\,|\,\theta) \cdot \boldsymbol{\pi}(\theta)$$

$\boldsymbol{\pi}(\theta)$ 是 θ 的先验分布。

(b) **Posterior Probability Distribution** 贝叶斯公式的密度函数形式

为得到后验概率分布，我们将以上 $h(\boldsymbol{X}, \theta)$ 进行分解为

$$\boldsymbol{h}(\mathbf{X}, \theta) = \boldsymbol{\pi}(\theta\,|\,\boldsymbol{X})F_X(\boldsymbol{X}) = p(\mathbf{X}\,|\,\theta)\int_{\Theta} p(\mathbf{X}\,|\,\theta)\boldsymbol{\pi}(\theta)\,d\theta$$

F_X 是 X 的边际分布函数 (CDF)。从而得到

$$\boldsymbol{\pi}(\theta\,|\,\boldsymbol{X}) = \frac{\boldsymbol{h}(\mathbf{X}, \theta)}{F_X(\boldsymbol{X})} = \frac{p(\mathbf{X}\,|\,\theta) \cdot \boldsymbol{\pi}(\theta)}{\displaystyle\int_{\Theta} p(\mathbf{X}\,|\,\theta)\boldsymbol{\pi}(\theta)\,d\theta}$$

该分布即为后验分布。同样如果 θ 的先验是离散型分布，则用 $\sum$ 代替分母的积分。

从上式可以得到一个结论:

$$\underbrace{\boldsymbol{\pi}(\theta\,|\,\boldsymbol{X})}_{Posterior} \propto \underbrace{p(\boldsymbol{X}\,|\,\theta)}_{Likelihod} \times \underbrace{\boldsymbol{\pi}(\theta)}_{Prior}$$

Although sometimes the mathematical form does not support, we can still use this proportional relationship to approximate the posterior probability.

9. **Hypothesis testing**

(a) **Formulation of Hypo. Testing** 假设检验种类形式及所需元素

- Null Hypothesis and Alternative Hypothesis 原假设与备择假设

- Simple Hypothesis v.s Composite Hypothesis 简单假设 v.s 符合假设
- Test statistics and Decision rule 检验统计量与决策规则
- Rejection Region 拒绝域
- Type one Error and Significance level 弃真错误与显著性水平
- Type two Error and Power 取伪错误与功效

—— 简单假设指需要检验的参数在假设中是一个确定的值 (i.e H_0: $\pi = 0.5$, and the corresponding alter.) 而复合假设中的参数是一个范 (i.e H_0: $\pi < 0.5$).

—— Rejection Region 是使得原假设被拒绝的检验统计值的集合。

—— 弃真错误的概率 = 显著性水平 = $Pr(Reject\, H_0 \,|\pi = \pi_0)$

—— Power 功效的定义是

$$Power = Pr(Reject\, H_0 \,|\, \pi = \pi_a)$$

因此取伪错误的概率为 1 - Power。

(b) **Neyman-Pearson Lemma**

Lemma 7. *Let $d(y)$ be a decision rule such that*

$$d(y) = \begin{cases} 1 & (reject\, H_0) \quad if\ \dfrac{f_Y(y|\theta_a)}{f_Y(y|\theta_0)} > c \\ 0 & (do\, not\, reject\, H_0) \quad otherwise \end{cases}$$

for some $c > 0$ and

$$Pr(d(Y) = 1 \,|\, \theta = \theta_0) = \alpha$$

then $d(y)$ is the most powerful test with significant level α.

其中，指定一个显著性水平是为了制定明确的决策规则，也就是为了确定第一个等式中的 c。proof may be made up later to appendix 值得注意的是，Neyman-Pearson Lemma 只适用于单边检测。

In the decision rule, the ratio

$$\frac{f_Y(y|\theta_a)}{f_Y(y|\theta_0)}$$

is called the likelihood ratio. 补：解释当 LHR 是增函数减函数的不同情况。当似然比是关于 y 的单调增函数时，decision rule 应当是 reject H_0 when test statistic $> c$; 如果是减函数则是 reject when test statistics $< c$.

10. **Generalized likelihood ratio test** 广义似然比估计

Definition 10. Let $x_1, x_2, ..., x_n$ i.i.d samples from pdf $p(x, \theta)$, $\theta \in \Theta$, and the Hypothesis testing problem is

$$H_0 : \ \theta \in \Theta_0 \quad v.s \quad H_1 : \theta \in \Theta_a = \Theta - \Theta_0$$

then we define the likelihood ratio as

$$\Lambda(x_1, ..., x_n) = \frac{\sup_{\theta \in \Theta} \ p(x_1, ..., x_n|\theta)}{\sup_{\theta \in \Theta_0} \ p(x_1, ..., x_n|\theta)} \equiv \frac{p(x_1, ..., x_n|\hat{\theta})}{p(x_1, ..., x_n|\hat{\theta}_0)}$$

where hat means the maximum likelihood estimate over the corresponding parameter space.

注: (1)其中分子可以看做基于参数子空间、没有假设时的最大似然估计。如果似然比值很大，说明 $\theta \in \theta_a$ 的可能性要比 $\theta \in \theta_0$ 的可能性大。该比值无法很小，因为子空间中的最大值一定小于等于整个参数空间的最大值，即 $\Lambda \in (1, \infty)$。

(2)sup 与 max 的意思相近。sup（上界）在数学上更严谨。sup 最大即最大似然—-给定 sample 寻找最可能的参数值。