Stat 2004

Yutong Ji

August 2022

1 Non-parametric estimate of F_θ given by the empirical distribution function F_n

Let $\boldsymbol{x_1, \ldots, x_n}$ denote as the observed values which corresponding to the random variable $\boldsymbol{X_1, \cdots, X_n}$. They are ***i.i.d.*** with common $\boldsymbol{f(x; \theta)}$ (could be discrete or continuous) we do not know θ yet, but we are about to estimate θ by $\hat{\theta}$

Specifically,

$$\hat{\theta} = T(x_1, \ldots, x_n)$$

where T is some function and we call $T(x_1, \ldots, x_n)$ an estimate and $T(X_1, \ldots, X_n)$ an estimator.

We want the estimation good that $\hat{\theta}$ needs to tend to be close to the true value θ:
Mathematically, we want:

$$E(|T(X_1, \ldots, X_n) - \theta|^\alpha) \tag{1}$$

to be small, where $\alpha > 0$
Aside: when $\alpha = 1$, it is the mean absolute error (MAE) and when $\alpha = 2$, it is the mean squared error (MSE).
MAE: is a measure of errors between paired observation expressing the same phenomenon. It is calculated as the sum of absolute errors (difference between prediction and the true value) divide by the sample size.
MSE: measures the average of the squares of the errors. MSE is a risk function, corresponding to the expected value of the squared error loss. Moreover, MSE is a measure of the quality of an estimator as it is derived from the square of Euclidean distance (the difference) so that it is always a positive value that decreases when the error approaches zero.

On the other hand, we would like the

$$P(|T(X_1, \ldots, X_n) - \theta| < \delta) \tag{2}$$

to be really large (close to 1) for an arbitrary small δ, where $\delta > 0$.

1.1 Empirical distribution function

In statistic, and an *empirical distribution function* also as known as empirical cumulative distribution function (eCDF) is the distribution function associated with the empirical measure of a sample. Interestingly, eCDF is a step function that jumps up by $\frac{1}{n}$ at each of the n data point (the reason can be found from its definition).

Let $(X_1, \ldots, X_n)$ be *i.i.d.* real random variable with common CDF $F(t)$. Then the eCDF can be defined as:

$$\hat{F}_n(t) = \frac{\text{the number of element in the sample} \leqslant \text{t}}{n} = \frac{1}{n}\sum_{I=1}^{n} \mathbb{1}_{X_i \leqslant t} \tag{3}$$

Where $\mathbb{1}_{X_I \leqslant t}$ is the indicator function.

Notice that, the indicator is actually a Bernoulli random variable (it only outputs 1 or 0) with parameter p = $F(t)$; therefore $n\hat{F}_n(t)$ has binomial distribution (sequence if i.i.d. Bernoulli). This implies that $n\hat{F}_n(t)$ is an unbiased estimator.

2 Definition of an unbiased estimator

First of all, *what is bias?* The bias of an estimator (or bias function) is the difference between the estimator's expected value and the true value of the parameter being estimated. So, an **unbiased** estimator is with zero bias.
Aside: in practice, biased estimators are frequently used, since an unbiased estimator does not exit without further assumptions, being unbiased is a too strong condition so that it is not useful...

The definition of the bias:

$$Bias(\hat{\theta}, \theta) = Bias_\theta(\hat{\theta}) = E_{x|\theta}(\hat{\theta} - \theta) \tag{4}$$

Following with the pervious estimator(**T**):

$$Bias(T) = E((X_1, \ldots, X_n)) - \theta$$

3 Method of Moments for estimation of k-dimensional

The method of moment is a method of estimation of population parameters.

Suppose a sample of size n is drawn, resulting in the values $x_1, \ldots, x_n$. For $j = 1, \ldots, k$. We have j-th sample moment, an estimate of μ_j(true population value):

$$\hat{\mu_j} = \frac{1}{n}\sum_{i=1}^{n} x_i^j \tag{5}$$

4 Maximum Likelihood estimation of k-dimensional

The maximum likelihood (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data. This can be achieved by maximising a likelihood function.

The goal of maximum likelihood estimation is to determine the parameters for which the observed data have the highest probability.
Suppose that, $X_1, \ldots, X_n$ are and iid sample from a population with pdf or pmfs $f(x|\theta_1, \ldots, \theta_k$, so the likelihood function is defined by:

$$L(\theta|x) = L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k) \tag{6}$$

Aside: sometimes we use ';' (joint condition) instead of '|' (given condition)

Example:

Let $X_1, \ldots, X_n$ be iid N(θ, 1) and let L($\theta|x$) denote the likelihood function. Then we can substitute θ as the μ and 1 as the variance into the likelihood function:

$$\begin{aligned} L(\theta|x) &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}(x_i-\theta)^2} \\ &= \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} \end{aligned}$$

Now we try to maximise the equation by letting its first derivative equals to 0. The trick there is noticing that the power of e is meant to be negative (negative sign in the front), except when $\sum_{i=1}^{n}(x_i - \theta) = 0$.

Therefore the equation (first derivative) $\frac{d}{d\theta}L(\theta|x) = 0$ can be reduce to:

$$\frac{d}{d\theta}L(\theta|x) = 0 \Rightarrow \sum_{i=1}^{n}(x_i - \theta) = 0$$

In conclusion, it is obviously that the solution would be $\hat{\theta} = \overline{x}$, however further proof for it attains the maximum need to be done...

5 TBD

5.1 Score Statistic

The **score** (or **informant**) is the gradient of the log-likelihood function with respect to the parameter vector. Mathematically,

$$s(x_1, \ldots, x_n; \theta) = \frac{d}{d\theta} log(L(\theta)) \tag{7}$$

One interesting result is related to this score statistic is that under certain regularity conditions the expected value (mean) of the score statistic is zero. To see this, we rewrite the likelihood function as a pdf. Then:

$$\begin{aligned} E(s|\theta) &= \int_{\chi} f(x;\theta) \frac{d}{d\theta} log(L(\theta;x)) dx \\ &= \int_{\chi} f(x;\theta) \frac{1}{f(x;\theta)} \frac{d}{d\theta} f(x;\theta) dx, \text{chain rule} \\ &= \int_{\chi} \frac{d}{d\theta} f(x;\theta) dx \end{aligned}$$

Now, we need to assume that the regularity condition allows the interchange of derivative and integral (Leibniz rule) to move further.

$$\frac{d}{d\theta} \int_{\chi} f(x;\theta) dx = \frac{d}{d\theta} 1 = 0$$

5.2 Fisher Information

Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ of a distribution that models X. Specifically, it is the variance of the score statistic (introduced above), or the expected value of the observed information.
Fisher information is defined to be the variance of the score, (under certain regularity conditions) mathematically:

$$\mathcal{F}(\theta) = E((\frac{d}{d\theta} log f(x;\theta))^2|\theta) \tag{8}$$

Also, the fisher information can be expressed as

$$\mathcal{F}(\theta) = E(I(\theta)), \text{ where } I(\theta) = -\frac{d^2}{d\theta^2} log L(\theta) \tag{9}$$

$I(\theta)$ is known as the negative of the Hessian of the log likelihood function. However, if we pass $\hat{\theta}$ (maximum likelihood estimate of θ) into the function I, it is the observed information.

Aside: the observed information (or observed fisher information) is the negative of the second derivative (the Hessian matrix) of the 'log-likelihood'. It is a **sample-based** version of the Fisher information.

Proof:
We are about to prove that the fisher information is equal to the negative Hessian ($I(\theta)$)
In use of the result that:

1. $\frac{d}{d\theta} L(\theta) = (\frac{d}{d\theta} log L(\theta)) L(\theta)$
2. $\int \cdots \int \frac{d}{d\theta} [\{\frac{d}{d\theta} log L(\theta)\} L(\theta)] dx_1 \ldots dx_n = 0$

Before we get into the proof, we first verify these result:
For the **result 1**, we briefly touch it in section 5.1. By chain rule:

$$\frac{d}{d\theta} log(L(\theta)) = \frac{1}{L(\theta)} \frac{d}{d\theta} L(\theta), \text{ by using chain rule}$$

Then we multiply both side with $L(\theta)$ we got the result 1.
For the **result 2**, we substitute the result 1 into the score statistic under regularity condition (expectation is zero). Details also are shown in section 5.1. Following from the result 2, we have:

$$\int \cdots \int \frac{d}{d\theta} [\{\frac{d}{d\theta} log L(\theta)\} L(\theta)] dx_1 \ldots dx_n = 0$$

$$\int \cdots \int ([\frac{d^2}{d\theta^2} log L(\theta)] L(\theta) + [\frac{d}{d\theta} log L(\theta)]^2 L(\theta)) dx_1 \ldots dx_n = 0$$

$$\textit{using the result 1 to transfer it back:} E([\frac{d^2}{d\theta^2} log L(\theta)] + [\frac{d}{d\theta} log L(\theta)]^2) = 0$$

$$\textit{by the linearity of expectation:} \underbrace{E([\frac{d^2}{d\theta^2} log L(\theta)])}_{\text{the negative Hessian}} + \underbrace{E([\frac{d}{d\theta} log L(\theta)]^2)}_{\text{Fisher information}} = 0$$

$$\textit{with the definition of } I(\theta) \textit{ in (9):} \quad -E(I(\theta)) + \mathcal{F}(\theta) = 0$$

Finally we have proven that $\mathcal{F}(\theta) = E(I(\theta))$

6 The regularity conditions

In the parameter estimation, the likelihood function is usually assumed to obey certain conditions - regularity condition.

- $f(x;\theta)$ have common support that the set $x : f(x;\theta) > 0$ does not depend on θ
- sample size is an open interval defined in $\mathbb{R}$(the partial derivative of $f(x;\theta)$ almost exist everywhere)
- the first derivative exists for all $\theta \in \Omega$
- the fisher information is greater than zero for all $\theta \in \Omega$
- the integral and derivative may be interchangeable for $L(\theta; x_1, \ldots, x_n)$ or $T(x_1, \ldots, x_n)L(\theta; x_1, \ldots, x_n)$ (illustrated the reason in 5.1)

The above conditions are sufficient, but not necessary. So, the model does not meet these regularity conditions may or may not have a maximum likelihood estimator.

7 Derivation of the Cramér-Rao lower bound

The Cramer-Rao bound (**CRB**) express a lower bound on the variance of unbiased estimators of a fixed and unknown parameter. The variance of any such estimator is at least as high as the inverse of the *Fisher information*. In other words, it expresses an upper bound on the precision (the inverse of variance) of unbiased estimators - *the precision of any such estimator is at most the Fisher information.*

Suppose that $T = t(X)$ is an unbiased estimator with expectation $g(\theta)$ which $E(T) = g(\theta)$. We are about to prove that for all θ, we have:

$$var(t(X)) \geq \frac{(g'(\theta))^2}{\mathcal{F}(\theta)} \tag{10}$$

Let X be a random variable with pdf $f(x;\theta)$ so $T = t(X)$ is a statistic used as an estimator for $g(\theta)$. Moreover, we define S as the score (see section 5.1).
If we consider the covariance of T and S:

$$\begin{aligned}
cov(S,T) &= E(ST) - E(S)E(T) \\
&= E(ST), \textit{ since under regularity condition } E(S) = 0 \textit{ (see section 5.1):} \\
&= E(T * [\frac{1}{f(X;\theta)}\frac{d}{d\theta}f(X;\theta)]), \textit{ following the definition of the score statistic} \\
&= \int t(x)[\frac{1}{f(x;\theta)}\frac{d}{d\theta}f(x;\theta)]f(x;\theta)dx \\
&= \frac{d}{d\theta}[\int t(x)f(x;\theta)dx], \textit{ the regularity allows the interchange of derivative and integral} \\
&= \frac{d}{d\theta}E(T) \\
&= g'(\theta), \textit{ where } E(T) = g(\theta) \textit{ is given}
\end{aligned}$$

The Cauchy-Schwarz inequality shows that: $\sqrt{var(T)var(S)} \geq |cov(T,S)|$
Therefore, $var(T) \geq \frac{(g'(\theta))^2}{var(S)}$
Finally, since the variance of score statistic is the Fisher information so that we have the result (10).

8 Regular Exponential Family

In this section, we provides two form of how to define the exponential family.

8.1 statistic interface book

In the book (Statistic Interface) we define exponential family as *a family of pdfs or pmfs that can be expressed as:*

$$f(x|\theta) = h(x)c(\theta)exp(\sum_{i=1}^{k} w_i(\theta)t_i(x)) \tag{11}$$

Where,
$h(x) \geq 0$ and $t_1(x), \ldots, t_k(x)$ are functions in range of all real value. Notice that they required the observations instead of unknown parameters, so they also called scalar function.
$c(\theta) \geq 0$ and $w_1(\theta), \ldots, w_k(\theta)$ are functions in range of all real value.

To verify a family of pdfs or pmfs belong to exponential family, we must identify those functions. We will go through an example of normal exponential family.

Example (1 observation normal) : let $f(x|\mu, \sigma^2)$ be the $N(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2)$. Then, we have:

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

Define:

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

Thus,

$$f(x|\mu, \sigma^2) = h(x)c(\mu, \sigma)exp(w_1(\mu, \sigma)t_1(x) + w_2(\mu, \sigma)t_2(x))$$

Where it is the exponential form with $k = 2$ with defined in (11).

8.2 Lecture notes: n-dimensional case

In the case of $x_1, \ldots, x_n$ observations, the likelihood function $L(\theta) = f(x_1, \ldots, x_n; \theta)$ has the following form, if it belongs to the exponential family:

$$f(x_1, \ldots, x_n; \theta) = b(x_1, \ldots, x_n)\frac{exp(c(\theta)^T T(x_1, \ldots, x_n))}{\alpha(\theta)} \tag{12}$$

Where,

- $T(x_1, \ldots, x_n)$ is a q-dimensional sufficient statistic
- $b(x_1, \ldots, x_n)$ and $\alpha(\theta)$ are non-negative scalar function
- $c(\theta)$ is a $q \times 1$ vector function of d-dimensional parameter vector θ

Note that if $q = d$ and the Jacobian of $c(\theta)$ is full rank, then it is said to be from a regular exponential family. The coefficient $c(\theta)$ is call the natural or canonical parameter vector.

8.3 One important takeaway in regular exponential family: ML Estimates of Parameters

It can be shown that the maximum likelihood estimate $\hat{\theta}$ of θ is the value of θ that satisfies the equation:

$$E_{\hat{\theta}}(T(X_1, \ldots, X_n)) = T(x_1, \ldots, x_n) \tag{13}$$

Proof:
Suppose that the joint distribution of the data $X_1, \ldots, X_n$ belongs to the regular exponential family (in form of (12)).

Firstly, the log likelihood function of $L(\theta)$ is:

$$\begin{aligned} logL(\theta) &= logb(x_1, \ldots, x_n) + log(\frac{exp(c(\theta)^T T(x_1, \ldots, x_n))}{\alpha(\theta)}) \\ &= logb(x_1, \ldots, x_n) + log(exp(c(\theta)^T T(x_1, \ldots, x_n))) - log(\alpha(\theta)) \\ &= logb(x_1, \ldots, x_n) + c(\theta)^T T(x_1, \ldots, x_n) - log(\alpha(\theta)) \end{aligned}$$

Next, on differentiation of the log likelihood function with respect to θ:

$$\begin{aligned} \frac{d}{d\theta} logL(\theta) &= \frac{d}{d\theta}[(logb(x_1, \ldots, x_n) + c(\theta)^T T(x_1, \ldots, x_n) - log(\alpha(\theta)))] \\ &= T(x_1, \ldots, x_n) - \frac{d}{d\theta} log(\alpha(\theta)) \end{aligned}$$

Notice that, under regular condition that the expectation of the score statistic is zero, where $E(\frac{d}{d\theta} logL(\theta)) = 0$. We take the expectation of both side:

$$E(\frac{d}{d\theta} logL(\theta)) = E(T(x_1, \ldots, x_n) - \frac{d}{d\theta} log(\alpha(\theta))) = 0$$

That is: $T(x_1, \ldots, x_n) - \frac{d}{d\theta} log(\alpha(\theta)) = 0 \Rightarrow T(x_1, \ldots, x_n) = \frac{d}{d\theta} log(\alpha(\theta))$
Now, we take the expectation again we get:

$$E(T(x_1, \ldots, x_n)) = \frac{d}{d\theta} log(\alpha(\theta)), \textit{since } \alpha(\theta)) \textit{ is a scalar}$$

Finally, we take a step back: we substitute $\frac{d}{d\theta} log(\alpha(\theta) = E(T(x_1, \ldots, x_n))$ back to $T(x_1, \ldots, x_n) = \frac{d}{d\theta} log(\alpha(\theta)$:

$$[E(T(X_1, \ldots, X_n))]_{\theta=\hat{\theta}} = T(x_1, \ldots, x_n))$$

9 Minimum Variance Bound Estimators

In section 7, we have seen that the condition needed for an unbiased estimator T attains the Cramer-Rao lower bound. Interestingly, if there is equality in the Cauchy-Schwarz inequality applied to $\frac{d}{d\theta}logL(\theta)$ (Score statistic) and T. The equality holds if and only if there is a linear function of the other:

$$\frac{d}{d\theta}logL(\theta) = k(\theta)T + l(\theta) \tag{14}$$

Then, T is the minimum variance bound (MVB). *We will explore more to see what* $k(\theta)$ *and* $l(\theta)$ *are.*

When we take the expectation of both side.

$$\begin{aligned} E(\frac{d}{d\theta}logL(\theta)) &= E(k(\theta)T + l(\theta)) \\ 0 &= k(\theta)E(T) + l(\theta) \\ 0 &= k(\theta)g(\theta) + l(\theta), \textit{perviously, it is known that } E(T) = g(\theta) \end{aligned}$$

This leads to $l(\theta) = -k(\theta)g(\theta)$, so we substitute it into (14):

$$\frac{d}{d\theta}logL(\theta) = k(\theta)T - k(\theta)g(\theta) \tag{15}$$

$$= k(\theta)(T - g(\theta)) \tag{16}$$

Now we have a neat version of the required linear function. It is saying that the score statistic must factor into $(T - g(\theta))$ times a function of the unknown parameter θ only and not depend on the observation.

On the other hand, in order to know whether an estimator T attains the MVB is to calculate its variance and see if it is equal to (needs to ask):

By using the result from (16), if we square both sides and take their expectations, we can find that:

square both side:

$$[\frac{d}{d\theta}logL(\theta)]^2 = [k(\theta)(T - g(\theta))]^2 = k(\theta)^2 * [T^2 - 2Tg(\theta) + g(\theta)^2]$$

take expectations of both side:

$$\begin{aligned} E([\frac{d}{d\theta}logL(\theta)]^2) &= E(k(\theta)^2 * [T^2 - 2Tg(\theta) + g(\theta)^2]) \\ &= k(\theta)^2 E([T^2 - 2Tg(\theta) + g(\theta)^2]), \textit{ the linearity of the expectation} \end{aligned}$$

Now before we take any step further, recall that $E(T) = g(\theta)$ (seen section 7)

$$E([\frac{d}{d\theta}logL(\theta)]^2) = k(\theta)^2(E(T^2) - E(T)^2) \tag{17}$$

$$= k(\theta)^2 var(T) \tag{18}$$

$$\mathcal{F}(\theta) = \frac{k(\theta)^2 g'(\theta)^2}{\mathcal{F}(\theta)}, \textit{ Since, } var(t(X)) \geq \frac{(g'(\theta))^2}{\mathcal{F}(\theta)} \tag{19}$$

We get intuition from (19) that it explains why if there is an equality in the Cauchy-Schwarz, the (14) holds and we have the minimum variance bound.

Following from (19), it shows that for now, $\mathcal{F}(\theta) = |k(\theta)g'(\theta)|$ so continuing from (18),

$$var(T) = \frac{\mathcal{F}(\theta)}{k(\theta)^2} = \frac{k(\theta)g'(\theta)}{k(\theta)^2} = \frac{g'(\theta)}{k(\theta)}$$

In this case, since $g(\theta)$ ($g(\theta) = E(T)$) is just a number so, $g'(\theta) = 1$. Recall that:
$\mathcal{F}(\theta) = |k(\theta)g'(\theta)| \Rightarrow \mathcal{F}(\theta) = k(\theta)$, so the derivation is finished with:

$$var(T) = \frac{1}{\mathcal{F}(\theta)} \tag{20}$$

Aside: this is also called the Scalar unbiased case of Cramér–Rao bound.

10 Sample Variance

10.1 Pre-known sample variance (biased)

In pervious sections, we know that we can derive sample variance from observations, especially the method of moment tells us that we can get the variance if you manipulate the second moment. That is:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Where, we have $x_1, \ldots, x_n$ as the observed data, and $\bar{x}$ is the first moment (sample mean): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

10.2 Unbiased sample variance

We still can obtain unbiased sample variance, if we correct the bias yields. What we do here is taking the expectation of the biased sample variance, then there is a special relation between them.

$$E(\hat{\sigma}^2) = E[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2]$$

Before we actually go to the derivation, let's state some handy results which can help us to understand:

- true mean is μ and true variance is σ^2
- $var(x_i) = \sigma^2$, $var(\bar{x}) = \frac{\sigma^2}{n}$, $E(x_i) = E(\bar{x}) = \mu$
- $E(x^2) = var(x) + (E(x))^2$
-

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\end{aligned}$$

Continuing from $E[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2]$:

$$\begin{aligned}
E[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2] &= \frac{1}{n}E(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2) \\
&= \frac{1}{n}(\sum_{i=1}^{n}E(x_i^2) - nE(\bar{x}^2)) \\
&= \frac{1}{n}(\sum_{i=1}^{n}[var(x_i)+E(x_i)^2] - n[var(\bar{x})+E(\bar{x})^2]) \\
&= \frac{1}{n}(n(\sigma^2+\mu^2) - n(\frac{\sigma^2}{n}+\mu^2)) \\
&= \frac{1}{n}(n-1)\sigma^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

The amazing result shows that the sample variance is expected to be the $\frac{n-1}{n}$ of true (population) variance, so that we can donate $\mathcal{S}^2$ as the unbiased sample variance that is:

$$\mathcal{S}^2 = \frac{n}{n-1}\hat{\sigma}^2 \tag{21}$$

Moreover, since we can have multiple observations, sample variance itself can be treated as a random variable. In case of each observation is independent with each other then $\mathcal{S}^2$ follows a chi-squared distribution:

$$\frac{n-1}{\sigma^2}\mathcal{S}^2 \sim \chi^2_{n-1} \tag{22}$$

With that, we can investigate some basic information of $\mathcal{S}^2$.
Under **normal distribution** circumstances:

$$E(\mathcal{S}^2) = E(\frac{\sigma^2}{n-1}\chi^2_{n-1}) = E(\frac{\sigma^2}{n-1}(n-1)) = \sigma^2 \tag{23}$$

Recall that: $var(aX) = a^2var(X)$

$$var(\mathcal{S}^2) = var(\frac{\sigma^2}{n-1}\chi^2_{n-1}) = \frac{\sigma^4}{(n-1)^2}var(\chi^2_{n-1}) = \frac{\sigma^4}{n-1} \tag{24}$$

11 Sufficient Statistic

A statistic is said to be sufficient if there is no other statistic that can be calculated from the same sample that carries any additional information.

we will continue to use T statistic here as conversion. So mathematically, a statistic $t = T(x)$ is sufficient for unknown parameter θ if the conditional probability distribution of the data X, given the statistic does not dependent on the unknown parameter.

11.1 Fisher-Neyman factorisation theorem

Fisher's factorisation theorem provides a characterisation of sufficient statistic that we can use the equation to determine if a statistic is sufficient.

A statistic T is said to be sufficient for unknown parameter (θ) if and only if the joint pdf can be factored as:

$$f(x_1, \ldots, x_n; \theta) = h_1(T(x_1, \ldots, x_n); \theta) h_2(x_1, \ldots, x_n), \forall \theta \in \Omega \tag{25}$$

Where we can see here h_1 and h_2 are non-negative functions and h_2 only depends on the observed data.

11.2 Minimal sufficiency

A sufficient statistic is said to be *minimal sufficient* if it can be represent as a function of any other sufficient statistic. Specifically, if $T(X)$ is minimal statistic if and only if:

- $T(X)$ is sufficient
- suppose that $S(X)$ is another sufficient statistic, then there exists a function f that $T(X) = f(S(X))$ (it is a function of every other statistic)

11.3 Completeness

Completeness is a property of a statistic in relation to a model for a set of observed data. Essentially, it ensures that the distributions corresponding to different values of the parameters are distinct.
A sufficient statistic T is complete if:

$$E_\theta(w(T)) = 0, \forall \theta \in \Omega \tag{26}$$

Where w is all possible measurable function.
Takeaway: a complete sufficient statistic is always the minimal sufficient statistic.

12 Rao-Blackwell Theorem

12.1 Theorem 1

Suppose that $U(X_1, \ldots, X_n)$ is an unbiased estimator of the unknown parameter θ and on top of that $T(X_1, \ldots, X_n)$ is a sufficient statistic. If there is a function W that:

$$W(T(X_1, \ldots, X_n)) = E[U(X_1, \ldots, X_n) | T(X_1, \ldots, X_n)] \tag{27}$$

Then we can conclude:

1. The function $W(T(X_1, \ldots, X_n))$ is a function of T that it does not depend on unknown parameter θ
2. $W(T(X_1, \ldots, X_n))$ is also an unbiased estimator of $\theta \Rightarrow E[W(T(X_1, \ldots, X_n))] = \theta$
3. $var(W(T)) < var(U)$, except when $U = W(T)$

Proof:

1. Recall that T is sufficient statistic of θ so that T does not depend on unknown parameters θ (definition of sufficient statistic). Therefore, any function with given T does not depend on θ as well. Hence, W does not depend on θ

2. Taking the expectation of (27) and by using of the tower property:

$$\begin{aligned} E[W(T(X_1,\ldots,X_n))] &= E[E[U(X_1,\ldots,X_n)|T(X_1,\ldots,X_n)]] \\ &= E(U(X_1,\ldots,X_n)) \\ &= \theta, \text{ since U is an unbiased estimator} \end{aligned}$$

3. Taking the variance of U:

$$\begin{aligned} var(U) &= var(E(U|T)) + E(var(U|T)) \\ &= var(W) + E(var(U|T)) \\ &> var(W(T)) \end{aligned}$$

Since variance is generally greater than zero

The main takeaway from this theorem is that if we have an unbiased estimator and a sufficient statistic T of unknown parameters θ, we can generate another unbiased estimator with a smaller variance. *So, if we have any kind of estimator, for instance, $g(\theta)$, we just compute the expectation of $g(\theta)$ conditionally given on a sufficient statistic $T(x)$, then we have a better estimator and it never worse, since the variance of the estimator has been reduced.* **Note: the precondition is that the original estimator is not a function of T alone.**

12.2 Theorem 2: Lehmann-Scheffé

Suppose that T now is a **complete statistic** and $U(T)$ is an unbiased estimator of θ which having finite variance. Then $U(T)$ is an uniformly minimum-variance unbiased estimator (UMVUE) of θ and it is unique.

13 Asymptotic Evaluations

13.1 Consistency

The property of consistency is a fundamental one which requiring that the estimator converges to the **true value** as simple size becomes infinite. Here we concerns a sequence of estimators rather than a single one.

Suppose that we observe $X_1, X_2, \ldots, X_n$ with respect to a distribution $f(x;\theta)$. Then we construct a sequence of estimators $T_n = T_n(X_1, X_2, \ldots, X_n)$:

T_n is said to be consistency if it converges to the true value that for every $\varepsilon > 0$:

$$\lim_{n\to 0} Pr(|T_n - \theta| \geq \varepsilon) \to 0 \tag{28}$$

Alternatively, if T_n is a sequence of estimators of a parameter θ that having $\lim_{n\to 0} var(T_n) \to 0$ and $\lim_{n\to 0} bias(T_n) \to 0$, then the sequence of estimates T_n is consistent.

(29)

Proof: We are about to the alternative statement above directly by using the result of Chebychev's Inequality which states that:

$$P(|T_n - \theta| \geq \varepsilon) \leq \frac{E((T_n - \theta)^2)}{\varepsilon^2}$$

From (28), we know that $P(|T_n - \theta| \geq \varepsilon)$ has to converge to 0 to be able to be a consistent estimator, so the only way to achieve that is making $\frac{E((T_n-\theta)^2)}{\varepsilon^2} = 0$, which is equivalence as making $E((T_n - \theta)^2)$ converges to 0.
And from (1), we know that $E((T_n - \theta)^2)$ is MSE that $E((T_n - \theta)^2) = var(T_n) + [bias(T_n)]^2$
Therefore, we complete the proof that the variance of the bias of T_n need to be 0 to make the sequence of T_n to be consistent.

Now we can see a couple of examples of consistent estimators to make our understanding concrete.

Lemma:
$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$ are consistent estimators of σ^2.

Demonstration:
We can take advantages of the speciality of $\sum_{j=1}^{n}(x_j - \bar{x})^2$. Recall that $\frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{\sigma^2} \sim \chi^2_{n-1}$. So,

$$E[\frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{\sigma^2}] = n - 1$$

And

$$var[\frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{\sigma^2}] = \sigma^{-4}2(n-1)$$

Therefore,

$$E[\sum_{j=1}^{n}(x_j - \bar{x})^2] = (n-1)\sigma^2$$

$$var[\sum_{j=1}^{n}(x_j - \bar{x})^2] = 2(n-1)\sigma^4$$

Hence,
For s^2:

$$E(s^2) = E[\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

$$var(s^2) = var[\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2] = (\frac{1}{n-1})^2 2(n-1)\sigma^4 = 0, \text{ as n} \to 0$$

$E(s^2) = \sigma^2$ shows s^2 is unbiased and $\lim_{n\to 0} var(s^2) = 0$. So by using the alternative statement (29) we can determine that s^2 is consistent estimator.
For $\hat{\sigma}^2$:

$$\begin{aligned} E(\hat{\sigma}^2) &= E[\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2] \\ &= \frac{1}{n}(n-1)\sigma^2 \\ &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \sigma^2, \text{ as } n \to 0 \end{aligned}$$

$$
\begin{aligned}
var(\hat{\sigma}^2) &= E[\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2] \\
&= (\frac{1}{n})^2 2(n-1)\sigma^4 \\
&= \frac{2(n-1)\sigma^4}{n^2} \\
&= 0, \text{ as } n \to 0
\end{aligned}
$$

So, similarly with s^2, but $\hat{\sigma}$ is unbiased only when the sample size is approaching infinite (large sample).

14 Asymptotic efficiency

14.1 Efficiency

Before we go into the asymptotic efficiency, let's have a look at efficiency in general.
In statistic, the efficiency is a measure of quality of an estimator - *how good is our estimator, which we normally prefer the one has less variance.*

Mathematically, the efficiency of an unbiased estimator is defined as:

$$
e(T) = \frac{1/\mathcal{F}(\theta)}{var(T)} \tag{30}
$$

Where, $\mathcal{F}(\theta)$ is the Fisher information.
Recall that the section 7 about Cramér-Rao lower bound: $var(t(X)) \geq \frac{(g'(\theta))^2}{\mathcal{F}(\theta)}$,

$$
e(T) = \frac{1/\mathcal{F}(\theta)}{var(T)} \leq \frac{1/\mathcal{F}(\theta)}{\frac{(g'(\theta))^2}{\mathcal{F}(\theta)}} \leq \frac{1}{(g'(\theta))^2}
$$

since, $var(T)$ could be larger as it only states the possible lower bound.
As $(g'(\theta))^2$ is positive, we can conclude that $e(T) \leq 1$.

Aside: in general, the variance (spread of the an estimator) is a measure of the estimator efficiency and performance. More formally, the mean square error (MSE) is the proper measure of the quality of the estimator (refer to section 1). For instance, if T_1 performs better than T_2, then $MSE(T_1) < MSE(T_2)$. However, in the unbiased estimator condition, it would depend only on the variance only.
Proof:
The MSE can be expressed as $MSE(T) = var(T) + [E(T) - \theta]^2$. Since T is unbiased, it makes $E(T) - \theta = 0$. So under unbiased estimators scenarios, $MSE(T) = var(T)$.

If $e(T)$ attains to 1, we would call the estimator T an efficient estimator. More interestingly, an efficient estimator is just the minimum variance unbiased estimator (MVUE, section 9).

14.2 asymptotically efficiency

Some estimator can attain efficiency asymptotically (when the sample turns to infinite), so that called asymptotically efficient estimators.
One important theorem is: let $X_1, X_2, \ldots$ be idd $f(x;\theta)$