INTRODUCTION AND BACKGROUND RESULTS

$\underset{\sim}{y} \in \underset{\sim}{\mathcal{Y}}$ (Sample space)

$\underset{\sim}{\theta} \in \Omega$ (parameter space)

$p(\underset{\sim}{y} \mid \underset{\sim}{\theta})$ (data density)

$\underset{\sim}{S}_n = \underset{\sim}{S}(y_1, \text{---}, y_n)$ (statistic)

The probability distribution of $\underset{\sim}{S}_n$ is called its "sampling distribution".

Our task is to use the data, via $\underset{\sim}{S}_n$, to draw inferences about $\underset{\sim}{\theta}$.

Estimators and tests are <u>decision rules</u>. The characteristics of the sampling dist'n. of $\underset{\sim}{S}_n$ will essentially determine its quality as a basis for inference.

e.g.: $y_i \sim (\mu, \sigma^2)$ (i.i.d.)

$S_n = \bar{y} = \frac{1}{n}\sum_i y_i \sim (\mu, \sigma^2/n)$

"Moments" of a Distribution:

Let Y be a random variable with density $p(y \mid \theta)$. Then the 'moments' of Y are defined as:

(i) $E(Y^k) = \mu_k' = \int_{-\infty}^{\infty} y^k p(y \mid \theta)\, dy$; $k = 1, 2, \text{---}$

("Raw moments" or "moments about origin")

(ii) $E(Y - \mu_1')^k = \mu_k = \int_{-\infty}^{\infty} (y - \mu_1')^k p(y \mid \theta)\, dy$ $(k = 1, 2, \text{---})$

("moments about the mean")

So: $\mu_0' = E(Y)$ = mean

$\mu_2 = E(Y - \mu_1')^2$ = variance

$= E(Y^2) + (\mu_1')^2 - 2\mu_1' E(Y)$

$= E(Y^2) - (\mu_1')^2$

$= \mu_2' - (\mu_1')^2$

We can always write the moments about mean in terms of moments about zero, & vice versa.

Why are moments important ?

If all of the moments exist, then knowledge of the moments provides knowledge of the underlying distribution.

If all of moments exist, two distributions will be identical if their moments match. (Uniqueness requires existence of moments).

There is a convenient way of generating the moments of a distribution, without any need to integrate. We take the Fourier transform of the density, which is called the Characteristic Function of the random variable.

Definition: The Characteristic Function of Y is $\phi_Y(t) = E(e^{itY})$; $i^2 = -1$

$$= \int_{-\infty}^{\infty} e^{ity} p(y)\, dy$$

Example: $Y \sim N(0,1)$

So, $\phi_Y(t) = E(e^{itY}) = \int_{-\infty}^{\infty} e^{ity} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y+it)^2/2} \cdot e^{-t^2/2} dy$$

$$= e^{-t^2/2} \qquad (= e^{(it)^2/2})$$

Show that:

(i) If $Y \sim N(\mu, \sigma^2)$, then

$$\phi_Y(t) = \exp\left[i\mu t - \frac{\sigma^2 t^2}{2}\right].$$

(ii) If Y_1 & Y_2 are independent, then the c.f. of their sum is the product of their individual c.f.'s.

$$\left[\phi_{Y_1+Y_2}(t) = \phi_{Y_1}(t) \cdot \phi_{Y_2}(t)\right]$$

Example: Let $Z \sim N(0,1)$ & $Y = Z^2$.

$$\phi_Y(t) = \int_{-\infty}^{\infty} e^{itz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2(1-2it)} dz$$

$$= (1-2it)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \frac{1}{(1-2it)^{-\frac{1}{2}}\sqrt{2\pi}} e^{-\frac{z^2}{2(1-2it)^{-1}}} dz$$

$$= (1-2it)^{-p/2}.$$

So, the c.f. for $\chi^2_{(1)}$ is $(1-2it)^{-1/2}$.

Similarly, the c.f. for $\chi^2_{(p)}$ is $(1-2it)^{-p/2}$.

There are various ways of getting the moments of the distribution from the c.f.:

(a) Differentiation –

$$\mu_k' = \left[\frac{d^k \phi(t)}{dt^k} \Big|_{t=0} \right] / i^k$$

e.g.: $Y \sim \chi^2_{(p)}$; $\phi_Y(t) = (1-2it)^{-p/2}$

$$\phi_Y'(t) = (-p/2)(-2i)(1-2it)^{-p/2-1}$$

$$= ip(1-2it)^{-(p/2+1)}$$

$$\therefore \mu_1' = [\phi_Y'(t)|_{t=0}]/i = p. \quad \#$$

$$\phi_Y''(t) = -(ip)(p/2+1)(-2i)(1-2it)^{-(p/2+2)}$$

$$\therefore \mu_2' = [\phi_Y''(t)|_{t=0}]/i^2 = 2p(p/2+1)$$

$$= 2p(p+2)/2 = p(p+2)$$

So, var.$(Y) = \mu_2 = \mu_2' - (\mu_1')^2 = 2p$. #

(b) Taylor-Series Expansion –

μ_k' = coefficient of $(it)^k/k!$ in series expansion.

e.g. $Y \sim N(0,1)$; $\phi_Y(t) = e^{-t^2/2}$

$$= e^{(it)^2/2}$$

Now, $e^x = 1 + x + \frac{x^2}{2!} + \cdots\cdots$

So, $\phi_Y(t) = 1 + \frac{(it)^2}{2} + \left[\frac{(it)^2}{2}\right]^2/2! + \left[\frac{(it^2)}{2}\right]^3/3!$

$$+ \left[\frac{(it^2)}{2}\right]^4/4! + \cdots$$

$$= 1 + \frac{(it)^2}{2!} + \frac{(it)^4}{8} + \frac{(it)^6}{48} + \cdots$$

So, $\mu_k' = 0$; for all odd k.

$\mu_2' = 1$

$\mu_4' = 3$ $\quad (\frac{1}{8} = 3/4!)$

i.e.: $E(Y) = 0$; var.$(Y) = 1 - 0^2 = 1$.

Definition: Skewness $= \dfrac{\mu_3}{(\mu_2)^{3/2}}$

$$\therefore \mu_3 = E(Y-\mu)^3 = \mu_3' + 2(\mu_1')^3 - 3\mu_1'\mu_2'$$

e.g. $Y \sim N(0,1)$

$\mu_1' = \mu_3' = 0 \quad ; \quad \mu_2' = 1 \quad ; \mu_2 = 1$

So, Skewness $= 0$

Definition: Kurtosis $= [\mu_4/(\mu_2)^2]$

$\mu_4 = E(Y-\mu)^4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$

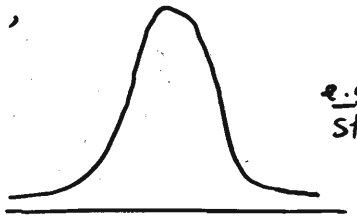
e.g. $Y \sim N(0,1)$

$\mu_4 = 3 - 0 + 0 - 0 = 3$

(Also true for $N(\mu, \sigma)$.)

Normal Dist'n. is "Mesokurtic"

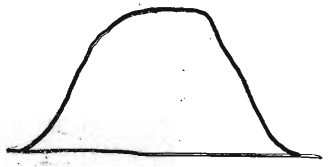
If kurtosis > 3, "Leptokurtic"



e.g. Student 't'

If kurtosis < 3, "Platykurtic"

e.g. Uniform



So, we often report the "excess kurtosis":

$[\mu_4/(\mu_2)^2] - 3$.

Summary:

* Moments of a distribution fully describe that dist'n. (if they exist).
* Characteristic fctn. provides a painless way of obtaining the moments.
* The moments are used to determine properties of a statistic when used as estimator or test statistic.

Estimator Properties:

Focus on "exact" (finite-sample) properties. Consider "asymptotic properties after introducing Maximum Likelihood estimation.

(i) Bias:

$\text{Bias}(\hat{\underline{\theta}}_n) = E(\hat{\underline{\theta}}_n) - \underline{\theta}$

("Unbiased" if Bias $= 0$)

Compares first moment of $\hat{\underline{\theta}}_n$ with location

(ii) (Relative) Efficiency:

Let $\hat{\underset{\sim}{\theta}}_n$ & $\tilde{\underset{\sim}{\theta}}_n$ be 2 unbiased estimators of $\underset{\sim}{\theta}$. Then $\hat{\underset{\sim}{\theta}}_n$ is efficient relative to $\tilde{\underset{\sim}{\theta}}_n$ if $[V(\tilde{\underset{\sim}{\theta}}_n) - V(\hat{\underset{\sim}{\theta}}_n)]$ is p.s.d.

(scalar: $\text{var.}(\tilde{\theta}_n) \geq \text{var.}(\hat{\theta}_n)$)

If one or both of the estimators is biased, $\hat{\underset{\sim}{\theta}}_n$ is relatively more efficient than $\tilde{\underset{\sim}{\theta}}_n$ if $[M(\tilde{\underset{\sim}{\theta}}_n) - M(\hat{\underset{\sim}{\theta}}_n)]$ is p.s.d., where $M(\underset{\sim}{\theta}_n^*) = V(\underset{\sim}{\theta}_n^*) + \text{Bias}(\underset{\sim}{\theta}_n^*)\text{Bias}(\underset{\sim}{\theta}_n^*)'$ is the matrix MSE. (1st. & 2nd. moments)

(iii) Sufficiency:

If we can find a statistic, $\underset{\sim}{S}_n$, that tells us as much about the population parameter(s) as does the full sample, this statistic will be "sufficient" for estimation purposes. More formally -

Let $\{\underset{\sim}{y}_i\}_{i=1}^n$ be a random sample from $p(\underset{\sim}{y}|\underset{\sim}{\theta})$. Then $\underset{\sim}{S}_n$ is a "sufficient statistic" iff $p(\underset{\sim}{y}|\underset{\sim}{S}_n)$ does not depend on $\underset{\sim}{\theta}$.

(If you know $\underset{\sim}{S}_n$, the sample values themselves add no information about $\underset{\sim}{\theta}$.)

This idea extends to "jointly sufficient statistics". And, any 1-1 transformation of a set of jointly sufficient statistics is also jointly sufficient.

Example: If $\sum_i y_i$ & $\sum_i y_i^2$ are jointly sufficient, then so are $\bar{y}$ and $\sum_i (y_i - \bar{y})^2$,

Now, how do we find a sufficient statistic in practice?

Factorization Theorem:

$\underset{\sim}{S}_n$ is sufficient iff

$$p(y_1, \dots, y_n \mid \underset{\sim}{\theta}) = g(\underset{\sim}{S}_n \mid \theta)\, h(y_1, \dots, y_n)$$

where $h(\cdot) \geq 0$ & does not involve $\underset{\sim}{\theta}$, and $g(\cdot) \geq 0$ & depends on the y_i's only through $\underset{\sim}{S}_n$.

For jointly sufficient statistics, this becomes:

$$p(y_1, \dots, y_n \mid \underset{\sim}{\theta}) = g(\underset{\sim}{S}_{1n}, \dots, \underset{\sim}{S}_{pn} \mid \underset{\sim}{\theta})\, h(y_1 \dots y_n)$$

Example: Let $y_1, \dots, y_n$ be drawn randomly from a Bernoulli density:

$$p(\underset{\sim}{y} \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta)$$

$$= \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \quad ; \quad 0 \leq \theta \leq 1$$

$$= \theta^{\Sigma y_i} (1-\theta)^{n - \Sigma y_i} \cdot 1$$

So, $g(\cdot) = \theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i}$; $h(\cdot) = 1$

& $\Sigma_i y_i$ is sufficient for θ (so is $\bar{y}$)

Example: $y_i \sim$ iid $N(\mu, 1)$

$$f(\underset{\sim}{y} \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}(y_i - \mu)^2\right]$$

$$= \frac{1}{(2\pi)^{n/2}} \exp\left[-\tfrac{1}{2} \Sigma_i (y_i - \mu)^2\right]$$

$$= \frac{1}{(2\pi)^{n/2}} \exp\left[-\tfrac{1}{2}(\Sigma_i y_i^2 + 2\mu \Sigma_i y_i + n\mu^2\right.$$

$$= \frac{1}{(2\pi)^{n/2}} \exp\left[\mu \Sigma_i y_i - \tfrac{n}{2}\mu^2\right] \exp\left[-\tfrac{1}{2}\Sigma_i y_i^2\right]$$

Let $g(\cdot) = \exp\left[\mu \Sigma_i y_i - \tfrac{n}{2}\mu^2\right]$

$$h(\cdot) = \frac{1}{(2\pi)^{n/2}} \exp\left[-\tfrac{1}{2}\Sigma_i y_i^2\right]$$

Then, $\Sigma_i y_i$ is a sufficient statistic, & so is $\bar{y}$.

Intuitively, it looks as if sufficient stats. (or simple fcns. of them) may lead to familiar & sensible estimators — this will indeed turn out to be the case.

Now let's pursue the idea of estimation as a decision-making procedure & introduce some ideas from statistical decision theory.

(iv) Risk:

Suppose we have a Loss function,

$$L(\hat{\underset{\sim}{\theta}}_n ; \underset{\sim}{\theta}) \geq 0 \quad ; \text{ all } \underset{\sim}{\theta}, \hat{\underset{\sim}{\theta}}_n$$

$$= 0 \quad ; \text{ if } \hat{\underset{\sim}{\theta}}_n = \underset{\sim}{\theta}.$$

(N.B. : Loss is random)

To get a single value, average:

$$\text{Risk}(\hat{\underset{\sim}{\theta}}_n ; \underset{\sim}{\theta}) = R(\hat{\underset{\sim}{\theta}}_n ; \underset{\sim}{\theta})$$

$$= E[L(\hat{\underset{\sim}{\theta}}_n ; \underset{\sim}{\theta})]$$

$$= \int_{\underset{\sim}{y}} L(\hat{\underset{\sim}{\theta}}_n ; \theta)\, p(\underset{\sim}{y} | \underset{\sim}{\theta})\, d\underset{\sim}{y}$$

e.g. Quadratic Loss -

(i) Scalar: $L = c(\hat{\theta}_n - \theta)^2 \quad ; c > 0$

$R = MSE$ if $c = 1$.

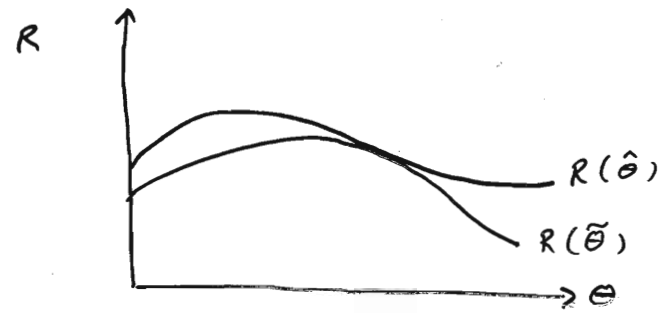
(ii) Vector: $L = (\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta})' W (\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}) \quad ; W$ p.d.

$R = \text{tr.}(MMSE)$, if $W = I$.

(v) Admissibility -

Low risk seems desirable. $\hat{\underset{\sim}{\theta}}_n$, an estimator (decision rule), is Inadmissible if $\exists$ any $\tilde{\underset{\sim}{\theta}}_n$ such that $R(\tilde{\underset{\sim}{\theta}}_n) \leq R(\hat{\underset{\sim}{\theta}}_n)$ for all $\underset{\sim}{\theta}$, and $R(\tilde{\underset{\sim}{\theta}}_n) < R(\hat{\underset{\sim}{\theta}}_n)$ for some $\underset{\sim}{\theta}$.

An Admissible estimator is one which is not inadmissible.

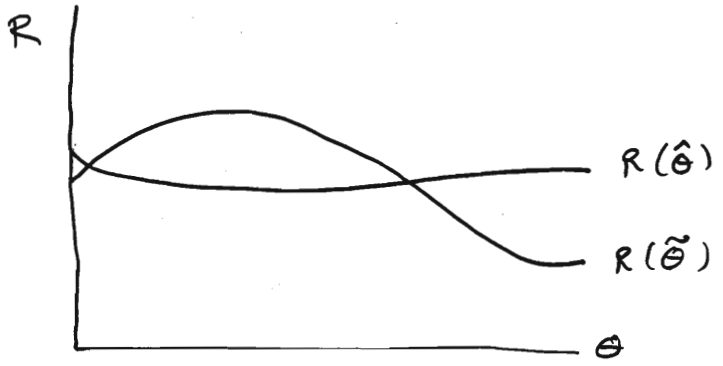


Many common estimators are inadmissible!

e.g. usual OLS estimators of β & σ^2 in linear regression, under quadratic loss, if $K \geq 3$! (Stein, 1956)

(vi) Mini-Max –

Typically, risk fctns. of estimators of interest cross –



Which estimator is preferred? We might average the risk (over θ), but what weights? (Return to this as Bayesians!)

One (conservative) option –

$\hat{\theta}$ is Mini-max within family of estimators, for specific loss fctn. if

$$\max. R(\hat{\theta}) = \min. \{ \max. R(\theta^*) \},$$

$\forall\ \theta^*$ in family of interest.

Some Questions:

* Are mini-max. estimators admissible?
* Are admissible estimators mini-max.?
* Are there any systematic, general, principles for constructing estimators that ensure that the estimators will have "good" properties of the types discussed so far?

We'll be considering 3 general principles –

(a) Maximum Likelihood Estimation
(b) Method of Moments Estimation
(c) Bayesian Estimation

(& the associated testing principles).

As a second-best solution, if it is hard to find principles guaranteed to yield "good" estimators in the above senses, maybe we can find principles that do well for large 'n