STATISTICS REVIEW

Background Information.

I. Measures of Return and Risk.

A. Operational definition of **Ex Post, Nominal Return**.

r = (ending value – beginning value + cash flows) / (beginning value)

= (ending – beginning)/(beginning) + (cash flows)/(beginning)

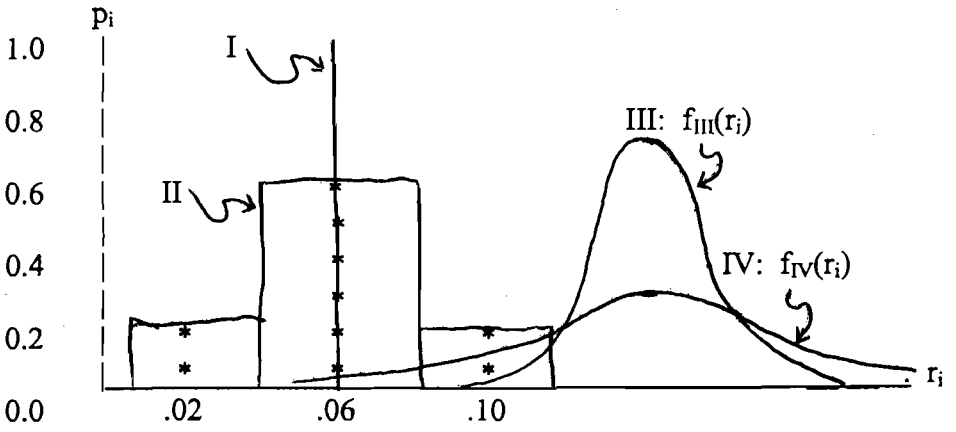
= (% Capital Appreciation) + (% Dividend Income).

1. This is a measure of ex post, actual return earned in the past.
   a. Accounting measure of past performance.
2. For investment decisions by individuals & management, need to consider expected future performance!
3. To get a measure of expected future return, must consider possible future outcomes.
4. This requires the probability distribution function (pdf) that reflects your expectations of possible future outcomes.
5. Consideration of this pdf introduces uncertainty – risk.

B. **Expected Return and Risk.**
– expected outcome & uncertainty about future return from investment.

1. Reflected in investor's probability distribution function (pdf).
2. Consider the pdf's of two possible investments, I & II:

Investment:	I	II		
Possible r (r_i)	.06	.02	.06	.10
Probability (p_i)	1.0	0.2	0.6	0.2



3. Note: I is a T.Bill; r = .06 with certainty (no risk).
   II depends on future states of the world; uncertainty!

4. **Expected Return** = $E(r_i) = \Sigma\, p_i\, r_i$; i indexes states of world.
   a. For investment I, $E(r_i) = (1.0)(.06) = .06$
   b. For investment II, $E(r_i) = (.2)(.02) + (.6)(.06) + (.2)(.10) = .06$

5. Observe, expected return is the same for I & II, but II has more risk.

6. In reality, investor's pdf is continuous, anything between $-\infty$ & $+\infty$.
   a. Thus, pdf is more like smooth curve such as III or IV above.
   b. For continuous pdf's, expected return is defined as the integral:

$$E(r_i) = \int_{-\infty}^{\infty} f(r_i)\, r_i\, dr_i ; \qquad \text{analogous to } \Sigma\, p_i\, r_i .$$

C. How do we measure **Risk**? Several possibilities.

1. Range = $\text{Max}\{r_i\} - \text{Min}\{r_i\}$.

   a. Problem: as sample size N increases, Range increases w/o bound.

2. Semi-Interquartile Range = $(X_{.75} - X_{.25}) / 2$
   = $(75^{\text{th}}$ %-tile - 25^{th} %-tile$) / 2$.

   a. This does not suffer from the problem with the Range.

   b. Used when the variance does not exist.

3. Variance = $\sigma^2 = E[r_i - E(r_i)]^2 = \Sigma\, p_i\, [r_i - E(r_i)]^2$.

   a. For I, $\sigma^2 = (1.0)(.06 - .06)^2 = 0$. No uncertainty, no risk.

   b. For II, $\sigma^2 = (.2)(.06-.02)^2 + (.6)(.06-.06)^2 + (.2)(.10-.06)^2$
   $= .00032 + 0 + .00032$
   $= .00064$

   c. Standard Deviation, σ.
   For I, $\sigma = 0$; For II, $\sigma \approx .0253$

   d. For continuous case, $\sigma^2 = \int_{-\infty}^{\infty} f(r_i)\, [r_i - E(r_i)]^2\, dr_i$.

4. If r_i deviates further from its mean, distribution is more spread out:
   $[r_i - E(r_i)]$ is larger; $[r_i - E(r_i)]^2$ is larger; σ^2 is larger.

D. **Covariance.**

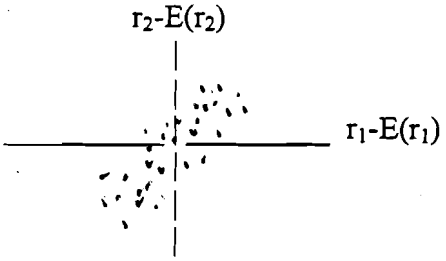
1. Defn: $\sigma_{12} = Cov(r_1, r_2) = E[r_1 - E(r_1)][r_2 - E(r_2)]$

2. Operational Defn: $\sigma_{12} = \sum_{i=1}^{n} [r_{1i} - E(r_1)][r_{2i} - E(r_2)]$

3. Case i: Suppose most combinations, (r_{1i}, r_{2i}) are in 1st & 3rd quadrants.

First quad: $[r_1 - E(r_1)]>0$ | $[r_2 - E(r_2)]>0$ | $\sigma_{12} > 0$

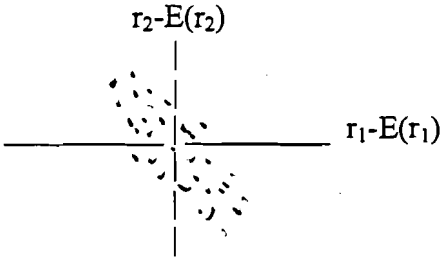
Third quad: $[r_1 - E(r_1)]<0$ | $[r_2 - E(r_2)]<0$ | $\sigma_{12} > 0$



5. Case ii: Suppose most combinations, (r_{1i}, r_{2i}) are in 2nd & 4th quadrants.

Second quad: $[r_1 - E(r_1)]<0$ | $[r_2 - E(r_2)]>0$ | $\sigma_{12} < 0$

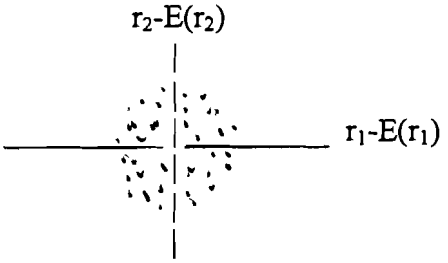
Fourth quad: $[r_1 - E(r_1)]>0$ | $[r_2 - E(r_2)]<0$ | $\sigma_{12} < 0$



6. Case iii: Suppose observations, (r_{1i}, r_{2i}) are scattered in all 4 quadrants.

Terms in 1st & 3rd quadrant $(\sigma_{12} > 0)$;
Terms in 2nd & 4th quadrant $(\sigma_{12} < 0)$;

Altogether, terms cancel out,
so that $\sigma_{12} \approx 0$.



Point: Sign of σ_{12} shows the *nature* of relation between r_1 & r_2.

E. **Correlation** = $\rho_{12} = \sigma_{12} / \sigma_1 \sigma_2$.

1. Note: σ_{12} may vary between $-\infty$ & $+\infty$.

   a. If r_1 and/or r_2 vary more widely (if σ_1 and/or σ_2 larger), then $[r_1 - E(r_1)]$ and/or $[r_2 - E(r_2)]$ are larger in magnitude, and σ_{12} will be larger in mag. (depending on case i, ii, or iii).

2. Thus, magnitude of σ_{12} doesn't tell us about extent of relation.

3. Correlation fixes this problem; adjusts σ_{12} for size of σ_1 and σ_2.

   a. If σ_{12} is larger (because σ_1 and/or σ_2 larger), ρ_{12} corrects for this by dividing σ_{12} by $(\sigma_1 \sigma_2)$.

4. Result: ρ_{12} varies between -1 and +1.

   a. If $\rho_{12} = +1$, r_1 & r_2 are perfectly positively related.

   b. If $\rho_{12} = -1$, r_1 & r_2 are perfectly negatively related.

   c. If $\rho_{12} = 0$, r_1 & r_2 are unrelated.

Statistical Review
[More Tools]

✳ [Throughout this explain with example.

Let $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix}$ be a parameter.

Defn 1: An <u>estimator</u> $\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_p \end{bmatrix}$ of θ is any function of the observed sample.

(It depends on random variables.)

UNBIASED ESTIMATOR

Defn 2: $\hat{\theta}$ is <u>unbiased</u> if $E[\hat{\theta}] = \theta$.

i.e. if $E(\hat{\theta}_i) = \theta_i \quad \forall \quad i = 1, \ldots, p$

Defn 3: The <u>covariance matrix</u> of $\hat{\theta}$ is:

$$Cov(\hat{\theta}) = E\left\{ [\hat{\theta} - E(\hat{\theta})][\hat{\theta} - E(\hat{\theta})]' \right\}$$

note:

$$= E \begin{bmatrix} \hat{\theta}_1 - E(\hat{\theta}_1) \\ \hat{\theta}_2 - E(\hat{\theta}_2) \\ \vdots \\ \hat{\theta}_p - E(\hat{\theta}_p) \end{bmatrix} \begin{bmatrix} \hat{\theta}_1 - E(\hat{\theta}_1) & \cdots & \hat{\theta}_p - E(\hat{\theta}_p) \end{bmatrix}$$

→ $p \times p$ matrix { variances on diag.
covariances off dia

$\text{Cov}(\hat{\theta})_{ij} = ij^{th}$ element

$= E\{[\hat{\theta}_i - E(\hat{\theta}_i)][\hat{\theta}_j - E(\hat{\theta}_j)]\}$

$= \text{Cov}(\hat{\theta}_i, \hat{\theta}_j)$

On diag.; $\text{Var}(\hat{\theta}_i)$

Off diag; $\text{Cov}(\hat{\theta}_i, \hat{\theta}_j)$

EFFICIENT ESTIMATOR

Note: "Efficient" estimators are unbiased and have minimum variance (in some sense).

⇐ We want the $\hat{\theta}$ with min. covariance matrix

Defn 4: An unbiased estimator $\hat{\theta}$ is efficient relative to another unbiased estimator $\tilde{\theta}$ if

$\text{Var}(\lambda'\hat{\theta}) \leq \text{Var}(\lambda'\tilde{\theta})$

for every $p \times 1$ vector λ.

Note: $\lambda'\hat{\theta} = [\lambda_1 \lambda_2 \cdots \lambda_p]\begin{bmatrix}\hat{\theta}_1\\ \hat{\theta}_2\\ \vdots\\ \hat{\theta}_p\end{bmatrix} = \lambda_1\hat{\theta}_1 + \lambda_2\hat{\theta}_2 + \cdots + \lambda_p\hat{\theta}_p$

$= \sum_{i=1}^{p} \lambda_i \hat{\theta}_i$

Defn says that every l.c. of $\hat{\theta}$ has a variance at least as small as every l.c. of $\tilde{\theta}$.

→ Very strong defn.! – ∃ a whole lot of λ's!

e.g. every element of $\hat{\theta}$ must have variance

⇐ ≤ every element of $\tilde{\theta}$.

Note: $\lambda'\hat{\theta}$ is a scalar; ∴ can talk about variance

Lemma: $Var(\lambda'\hat{\theta}) = \lambda' Cov(\hat{\theta})\lambda$ ← ("Prove" with 2×2 case yourself)

Proof: $Var(\lambda'\hat{\theta}) = Var\left[\sum_{i=1}^{p} \lambda_i \hat{\theta}_i\right]$

$$= E\left\{\left[\sum_i \lambda_i \hat{\theta}_i - \sum_i \lambda_i E(\hat{\theta}_i)\right]\right\}^2$$

$$= E\left\{\sum_i \lambda_i \left[\hat{\theta}_i - E(\hat{\theta}_i)\right]\right\}^2$$ (square)

$\lambda_i \neq \lambda_j$ not nec. =, just expanding the square (distributive property)

$$= E\left\{\sum_i \lambda_i \left[\hat{\theta}_i - E(\hat{\theta}_i)\right] \sum_j \lambda_j \left[\hat{\theta}_j - E(\hat{\theta}_j)\right]\right\}$$

(Do yourself)

$$= \sum_i \sum_j \lambda_i \lambda_j E\left[\hat{\theta}_i - E(\hat{\theta}_i)\right]\left[\hat{\theta}_j - E(\hat{\theta}_j)\right]$$

$$= \sum_i \sum_j \lambda_i \lambda_j Cov(\hat{\theta}_i, \hat{\theta}_j)$$

$$= \lambda' Cov(\hat{\theta}) \lambda$$ (quadratic form)

Thm:

An unbiased estimator $\hat{\theta}$ is efficient relative to an unbiased estimator $\tilde{\theta}$ iff $[Cov(\tilde{\theta}) - Cov(\hat{\theta})]$ is p.s.d.

Proof: $Var(\lambda'\hat{\theta}) \leq Var(\lambda'\tilde{\theta}) \quad \forall \lambda$ ← (Defn 4) of efficiency

$\rightarrow \lambda' Cov(\hat{\theta}) \lambda \leq \lambda' Cov(\tilde{\theta}) \lambda \quad \forall \lambda$ by Lemma

$\rightarrow 0 \leq \lambda' [Cov(\tilde{\theta}) - Cov(\hat{\theta})] \lambda \quad \forall \lambda$

$\rightarrow [Cov(\tilde{\theta}) - Cov(\hat{\theta})]$ is psd (defn of psd)

Defn: An unbiased estimator $\hat{\theta}$ is efficient if it is efficient relative to all other unbiased estimators of θ.

★ **Thm (Cramer-Rao):** Let $\{x_1, x_2, \ldots, x_T\}$ be a random sample (r.s.) (implies indep) from a population with density, $f(x;\theta)$.

Define the likelihood function

$$L(x_1, \ldots, x_T; \theta) = \prod_{i=1}^{T} f(x_i; \theta)$$

Define the Information Matrix, $\mathcal{I}$, by

$$\mathcal{I}_{ij} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] \quad i, j = 1, \ldots, p$$

(E: expected value)

Then for any unbiased* estimator $\hat{\theta}$ of θ,

$$\left[\mathrm{Cov}(\hat{\theta}) - \mathcal{I}^{-1}\right] \text{ is psd.}$$

expl.: $\mathcal{I}$ is the matrix of 2nd order partia of the log-likelihood function, $\mathcal{I}^{-1}$ is a "lower bound" on all estim in $\mathrm{Cov}(\hat{\theta})$ by this Thm.

(!) → correct: "lower bound" on all estima on the diagonal of $\mathrm{Cov}(\hat{\theta})$ (variance

examples

Corollary: If $Cov(\hat{\theta}) = \mathcal{I}^{-1}$, then $\hat{\theta}$ is efficient.

Note: This is a sufficient condition, not necessary.

i.e. Can calculate $Cov(\hat{\theta})$ and $\mathcal{I}^{-1}$, if elements in $Cov(\hat{\theta})$ = these in $\mathcal{I}^{-1}$, $\hat{\theta}$ is efficient.

--- if $\neq$, $\hat{\theta}$ still may be efficient (i.e. lower-bound, $\mathcal{I}^{-1}$, may be unattainal

Example:

$\{x_1, \ldots, x_T\}$ is random sample from $N(\mu, \sigma^2)$

$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}; \quad f(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$= (2\pi)^{-\frac{1}{2}} (\sigma^2)^{-\frac{1}{2}} \exp\left\{\frac{-1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\therefore L(x_1, \ldots, x_T; \theta) = \prod_{i=1}^{T} (2\pi)^{-\frac{1}{2}} (\sigma^2)^{-\frac{1}{2}} \exp\left\{\frac{-1}{2\sigma^2}(x_i-\mu)^2\right\}$$

$$= (2\pi)^{-\frac{T}{2}} (\sigma^2)^{-\frac{T}{2}} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{T}(x_i-\mu)^2\right\}$$

and $$\boxed{\ln L = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{T}(x_i-\mu)^2}$$

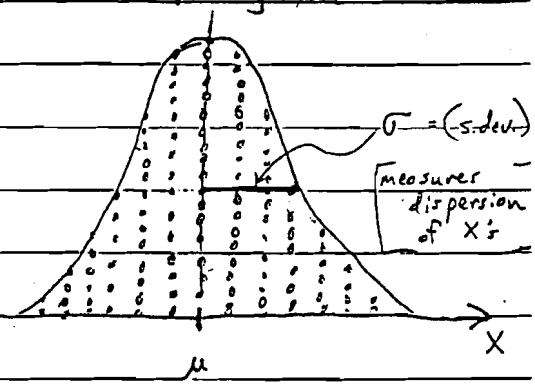
↳ Referred to a lot.

Intuitive expl. of Likelihood function.

e.g. $\{X_1, \ldots, X_T\}$ r.s. from $N(\mu, \sigma^2)$.

What does this mean?
- Draw out T X's, observe their values. $f(X, \theta)$

If you were to plot a frequency distribution or histogram, these T obs. would pile up around μ. →



Can draw a Bell-shaped curve around distribution.
→ Normal distribution

Normal density is equation that gives this curve;

$$f(x;\theta) = (2\pi)^{-1/2}(\sigma^2)^{-1/2} \exp\left\{\frac{-1}{2\sigma^2}(x-\mu)^2\right\}$$

Note: Only information we're given is X's;
don't know true μ and σ^2.
- Want to estimate them from sample obs.!
(Find out mean and variance of sample)

For each obs, X_i, can plug into $f(\theta; X_i)$
and obtain numerical information about the distribution;
presumably gives some point on the Bell-shaped curve.

--- ∴ get T values of density $f(\theta; x_i)$ $i = 1 \ldots T$

— These T values are like "probabilities" — the area under $f(x;\theta)$ is probability.

— 5 (TNICKI B)

How can we put all the information contained in the sample together?

-- The Likelihood function puts all this information together:

$$L = \prod_{i=1}^{T} f(\theta; x_i) = (2\pi)^{-\frac{T}{2}} (\sigma^2)^{-\frac{T}{2}} \exp\left\{\frac{-1}{2\sigma^2} \sum_i (x_i - \mu)^2\right\}$$

This presents a number that depends on the X_i's and the parameter of interest, $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$;

$$L(x_1, \ldots, x_T; \theta),$$

that gives us information about the distribution the X_i's are drawn from.

Presumably the $f(\theta; x_i)$, $i = 1, \ldots, T$, somehow "trace out" the Bell-shaped curve, and the Likelihood function incorporates all this information. [gives some "total probability" or "likelihood" about the dist]

Thus the Likelihood function provides information that suggests to us what kind of distribution it is most likely that these X's would have come from.

i.e. L tells us information about the distribution (μ & σ^2) that would make it most likely that we would observe what we have observed!

Note: form of L depends on kind of density the sample is drawn from. Normal* Hypergeometric Binomial Poisson

First Deriv.'s $\quad \ln L = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$

$$\frac{\partial \ln L}{\partial \mu} = \frac{-1}{2\sigma^2}\sum_i 2(x_i-\mu)(-1) = \frac{1}{\sigma^2}\sum_i (x_i-\mu)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{-T}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i (x_i-\mu)^2$$

Check if you want; $E[\ ]$ of each $= 0$.
(This regularity condition holds for nearly all distributions.) - [one exception; the uniform dist.]

Second Deriv.'s

$$\frac{\partial^2 \ln L}{\partial \mu^2} = \frac{1}{\sigma^2}\sum_i(-1) = \frac{-T}{\sigma^2}$$

$$\frac{\partial^2 \ln L}{\partial \sigma^2 \partial \mu} = \frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} = \frac{-1}{(\sigma^2)^2}\sum_i(x_i-\mu) = \frac{-1}{\sigma^4}\sum_i(x_i-\mu)$$

$$\frac{\partial^2 \ln L}{\partial(\sigma^2)^2} = \frac{T}{2(\sigma^2)^2} - \frac{2}{2(\sigma^2)^3}\sum_i(x_i-\mu)^2$$

$$= \frac{T}{2\sigma^4} - \frac{1}{\sigma^6}\sum_i(x_i-\mu)^2$$

Take Expected Values

$$E\left[\frac{\partial^2 \ln L}{\partial \mu^2}\right] = \frac{-T}{\sigma^2}$$

$$E\left[\frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2}\right] = \frac{-1}{\sigma^4}E\left[\sum_i(x_i-\mu)\right] = 0 \qquad (\text{since } E(x_i)=\mu)$$

$$E\left[\frac{\partial^2 \ln L}{\partial(\sigma^2)^2}\right] = \frac{+T}{2\sigma^4} - \frac{1}{\sigma^6}E\left[\sum_i(x_i-\mu)^2\right] \qquad (\text{since } E(x_i-\mu)^2=\sigma^2)$$

$$= \frac{T}{2\sigma^4} - \frac{1}{\sigma^6}\sum_i\sigma^2 = \frac{T}{2\sigma^4} - \frac{T\sigma^2}{\sigma^6} = \frac{-T}{2\sigma^4}$$

$$\begin{bmatrix} -T/\sigma^2 & 0 \\ 0 & -T/2\sigma^4 \end{bmatrix}$$

on board

Thus,

$$\mathcal{I} = -E\begin{bmatrix} \frac{\partial^2 \ln L}{\partial \mu^2} & \frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} \frac{T}{\sigma^2} & 0 \\ 0 & \frac{T}{2\sigma^4} \end{bmatrix}$$

and $\mathcal{I}^{-1} = \begin{bmatrix} \frac{\sigma^2}{T} & 0 \\ 0 & \frac{2\sigma^4}{T} \end{bmatrix}$ ← lower bound by Cramer for $\mathrm{Cov}(\hat{\theta})$

* Consider the estimator, $\hat{\Theta} = \begin{bmatrix} \bar{x} \\ s^2 \end{bmatrix}$ of $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$

where $\bar{x} = \frac{1}{T}\sum_i x_i$ and $s^2 = \frac{1}{T-1}\sum_i (x_i - \bar{x})^2$

and $\bar{x} \sim N(\mu, \frac{\sigma^2}{T})$ and $\frac{(T-1)s^2}{\sigma^2} \sim \chi^2_{T-1}$

Note: $E[\text{any } \chi^2] = \text{its d.f.}$

$\mathrm{Var}[\text{any } \chi^2] = 2 * (\text{its d.f.})$

Thus, $E\left[\frac{(T-1)s^2}{\sigma^2}\right] = T-1$

$\rightarrow E\left[\frac{s^2}{\sigma^2}\right] = 1$; $\rightarrow E(s^2) = \sigma^2$

and $\mathrm{Var}(s^2) = \mathrm{Var}\left[\left(\frac{\sigma^2}{T-1}\right)\left(\frac{T-1}{\sigma^2}\right)s^2\right]$

$= \left(\frac{\sigma^2}{T-1}\right)^2 \mathrm{Var}\left[\frac{(T-1)s^2}{\sigma^2}\right]$

$= \frac{\sigma^4}{(T-1)^2}\, 2(T-1)$

$= \frac{2\sigma^4}{T-1}$

(*)

Hence $E(\hat{\theta}) = E\begin{bmatrix}\bar{x}\\ s^2\end{bmatrix} = \begin{bmatrix}\mu\\ \sigma^2\end{bmatrix} = \theta$ → unbiased

and $Cov(\hat{\theta}) = Cov\begin{bmatrix}\bar{x}\\ s^2\end{bmatrix} = \begin{bmatrix}\frac{\sigma^2}{T} & 0\\ 0 & \frac{2\sigma^4}{T-1}\end{bmatrix}$ X_t is nonstochastic ∴ $cov(\bar{X}, S^2) = 0$

$[Cov(\hat{\theta}) - d^{-1}]$ is p.s.d. for any unbiased estimator!

Compare this with d^{-1}: $\begin{bmatrix}\sigma^2/T & 0\\ 0 & 2\sigma^4/T\end{bmatrix}$

1) $\bar{X}$ is efficient.

$\frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$ } "efficiency measure"

2) s^2 not proved efficient here; Variance not as small as lower bound.

Consistency → $\lim_{n\to\infty} P\{|\hat{\theta}_n - \theta| \le \varepsilon\} = 1$

$\lim_{n\to\infty} Var(\hat{\theta}_n) = 0$

It turns out that s^2 is efficient too.
We need more Thm's to prove it.

Note: The lower bound for s^2 in d^{-1} is unattainable here.

(25) Oh next page

p. 3 Hindi

Defn: Consider a random sample $\{x_1, \ldots, x_T\}$ from $f(x;\theta)$. Let $t(x_1, \ldots, x_T)$ be a function of the X's.
Then t is sufficient for θ if the distribution of $\{x_1, \ldots, x_T\}$ conditional on t does not depend on θ.

If $g_{x_1 \ldots x_T / t}(x_1, \ldots x_T)$ ~~$\neq$~~ θ.

$L(x_1 \ldots x_T) = g(t,\theta) \cdot h(x_1 \ldots x_T)$

not a function of θ.

Intuition:

If t is sufficient, it contains all the information in the sample about θ, and it wouldn't make any difference if we were given (information about) θ or not.
(explicit)

* In other words, if t is a sufficient statistic for θ, it contains <u>all the information</u> wrt θ contained in the sample.

* <u>Blackwell-Rao Thm:</u>

A necessary and sufficient condition for $t(x_1, \ldots, x_T)$ to be <u>sufficient</u> for θ is that

$$h(x_1, \ldots, x_T; \theta) = g(t, \theta) \cdot h(x_1, \ldots, x_T)$$

where g does not depend on $\{x_1, \ldots, x_T\}$ except thru t, and h does not depend on θ.

→ i.e You must be able to factor the likelihood function in this manner.

expl. with example! ☺

Example: $\{x_1, \ldots, x_T\}$ is r.s. from $N(\mu, \sigma^2)$.

$$L = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^T \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right\}$$

Note: $\sum_i (x_i-\mu)^2 = \sum_i (x_i-\bar{x})^2 + T(\bar{x}-\mu)^2$

Substituting,

$$L = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^T \exp\left\{\frac{-1}{2\sigma^2}\left[\sum_i (x_i-\bar{x})^2 + T(\bar{x}-\mu)^2\right]\right\}$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^T \exp\left\{\frac{-1}{2\sigma^2}\left[(T-1)s^2 + T(\bar{x}-\mu)^2\right]\right\}$$

Observe, L = function of $\theta = \begin{bmatrix}\mu \\ \sigma^2\end{bmatrix}$ & $t = \begin{bmatrix}\bar{x} \\ s^2\end{bmatrix}$

(*) If g = the entire function, L, and $h = 1$, then t is sufficient for θ.

Hence we have factored L into 2 functions;

$g(t, \theta)$ which doesn't depend on $\{x_1, \ldots, x_T\}$ except thru t;

and $h(x_1, \ldots, x_T)$ which doesn't depend on θ.

* The only place the x_i show up is in t, ($\bar{x}$ & s^2)

[redacted]: (Min-variance unbiased estimator) Unbiased estimator that is a function of a minimal sufficient stat.

Note: Many other t's work here;

e.g. $t' = \begin{bmatrix} \bar{x} \\ \sum_i (x_i - \bar{x})^2 \end{bmatrix}$, $t'' = \begin{bmatrix} \sum_i x_i \\ \sum_i x_i^2 \end{bmatrix}$, etc.

--- As long as the t's contain all the information about θ in the sample, L can be factored into some $g(t,\theta)$ * some h.

But also Note: If only one sufficient statistic, t, that is unbiased.

This brings up the following Thm.

Method of Lehmann & Scheffé: Say $(X_1 \ldots X_n)$ and $(Y_1 \ldots Y_n)$ comes from same dist.

If $\frac{L(X_1 \ldots X_n)}{L(Y_1 \ldots Y_n)}$ = free of θ then $g(Y_1 \ldots Y_n)$ is the minimal sufficient stat for θ.

(or OH)

Thm: An unbiased estimator which is a function of a sufficient statistic is efficient.

→ sufficient; uses all information

→ efficient; min. variance of all poss. estimates

Recall, showed that $\bar{x}$ is efficient; its variance = lower bound in $\mathcal{I}^{-1}$

Now, know that s^2 is efficient by above Thm.

Sometimes must settle for less than an efficient estimator.

OH

Defn: An unbiased estimate of θ is the *Best Linear Unbiased Estimator* (BLUE) if

① it is a linear function of the sample

and ② it is efficient relative to any other linear unbiased estimate.

Example: $\{x_1, \ldots, x_T\}$ is r.s. from *any* distribution with finite mean and variance. Then $\bar{x}$ is BLUE of μ.

Note: Already shown, if $\{x_1, \ldots, x_T\}$ is *Normal*, $\bar{x}$ is efficient estimator of μ.

$\Leftarrow$ (even beats nonlinear estimates)

The above statistical concepts have been dealing with "*finite* sample properties of estimators."

Now consider the Asymptotic (large sample) Properties of Estimators.

Asymptotic Properties hold only for ∞ samples, But as T ↑es, the estimators behave more closely to their As. Properties.

Let $\hat{\theta}_T$ be an estimator of θ based on a sample of T observations.

~~Useless~~ Defn: $\hat{\theta}_T$ is asymptotically unbiased if

$$\lim_{T \to \infty} E(\hat{\theta}_T) = \theta.$$

- Not very useful since we are often interested in estimators where the moments don't exist (e.g. Cauchy dist.).

ex. Consider $\hat{\theta}_T = \frac{T-1}{T}\bar{X}$. — Biased (est of μ), but As. Unbiased.

More Useful

Defn: Consider a sequence of random variables e.g. $\{\hat{\theta}_1, \hat{\theta}_2, \ldots\}$

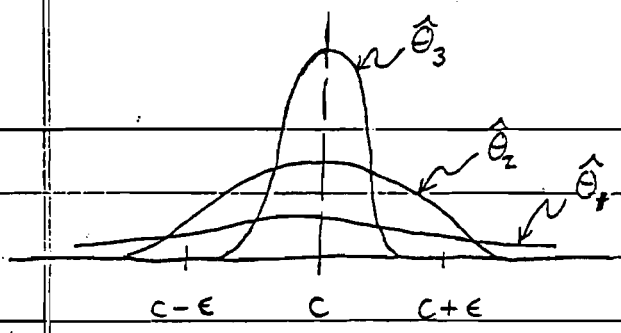
Then if $\exists$ a number c $\ni$

$$\lim_{T \to \infty} P(|\hat{\theta}_T - c| > \epsilon) = 0 \quad \forall \epsilon > 0$$

then the sequence converges in probability to c, and we write

$$\operatorname{plim}_{T \to \infty} \hat{\theta}_T = c.$$

expl. ↓



The probability (area under pdf's, outside of $c \pm \epsilon$) approaches 0 as $T \to \infty$.

..... The distribution of the $\hat{\theta}$'s converges to a "spike" at c.
-- The variance of $\hat{\theta}_T \to 0$ as $T \to \infty$. ✡

(Note:) the above 2 Defn's say almost the same thing. The second is more useful since we don't need to take any expected values to make use of it.
- plim also → variance goes to zero!

Example: $\{z_1, z_2, \ldots, z_T\}$ a r.s. from $N(\mu, \sigma^2)$

$$\bar{z}_T = \frac{1}{T}\sum_1 z_i \quad ; \quad \bar{z}_T \sim N\left(\mu, \frac{\sigma^2}{T}\right)$$

The variance $\to 0$ as $T \to \infty$.

$$\lim_{T\to\infty} P(|\bar{z}_T - \mu| > \epsilon) = 0,$$

or $\underline{\text{plim}\ \bar{z}_T = \mu}$.

This ex. reflects the Law of Large Numbers, and shows the As. Properties of $\bar{z}_T$.

Defn: An estimator $\hat{\theta}_T$ is consistent if

$$\operatorname{plim}_{T\to\infty} \hat{\theta}_T = \theta.$$

— implies 2 conditions:

1) $E(\hat{\theta}_T) \to \theta$
2) $\operatorname{Var}(\hat{\theta}_T) \to 0$

Defn: The Mean Square Error of $\hat{\theta}_T$ is

$$MSE(\hat{\theta}_T) = E(\hat{\theta}_T - \theta)^2$$

and note ...

$$= E\left[(\hat{\theta}_T - E(\hat{\theta}_T)) + (E(\hat{\theta}_T) - \theta)\right]^2$$

$$= E[\hat{\theta}_T - E(\hat{\theta}_T)]^2 + [E(\hat{\theta}_T) - \theta]^2 + 2[E(\hat{\theta}_T) - \theta]\underbrace{E[\hat{\theta}_T - E(\hat{\theta}_T)]}_{0}$$

$$= E[\hat{\theta}_T - E(\hat{\theta}_T)]^2 + [E(\hat{\theta}_T) - \theta]^2$$

$$= \operatorname{Var}(\hat{\theta}_T) + [\text{Bias}]^2$$

Note: The $MSE(\hat{\theta}_T) = \operatorname{Var}(\hat{\theta}_T)$ if $\hat{\theta}_T$ is unbiased.

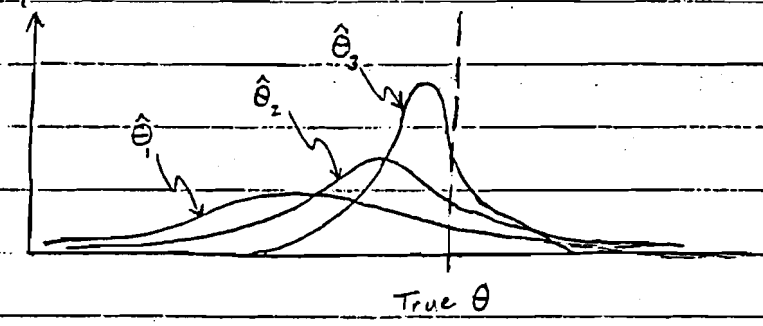
Thus $MSE(\hat{\theta}_T)$ is a criterion used to judge how good an estimator is, if it is biased. — just like $\operatorname{Var}(\hat{\theta}_T)$ is used to judge unbiased estimates.

Thm: If $\lim_{T\to\infty} MSE(\hat{\theta}) = 0$,

then $\hat{\theta}_T$ is consistent.

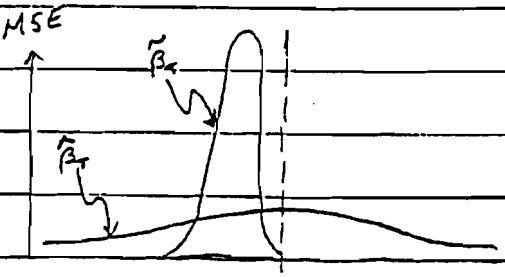
Proof: meets 2 conditions —
Variance & Bias go to zero.

picture:



Note: "Bias" is not always a terrible monster
to avoid at all costs.
In some cases a Biased estimate $\tilde{\beta}$
may be preferable to an unbiased est, $\hat{\beta}$

e.g.



→ here $MSE(\tilde{\beta}) < MSE(\hat{\beta})$

[note here that $\tilde{\beta}$ will never yield true β !!]

— Better to get close to true β
with high degree of precision,
than to have unbiased estimate
with huge variance (low precision)

On handout

Thm: (Slutsky)

If $\text{plim } X_T = c$, and if g is any continuous function, then

$$\text{plim } g(X_T) = g(\text{plim } X_T) = g(c)$$

Examples: ① $\text{plim } (\hat{\theta}_T)^2 = (\text{plim } \hat{\theta}_T)^2$

② $\hat{\theta}, \tilde{\theta}$; $\text{plim}\left(\frac{\hat{\theta}}{\tilde{\theta}}\right) = \frac{\text{plim } \hat{\theta}}{\text{plim } \tilde{\theta}}$

Note: This does <u>not</u> hold in general for Expected Values.

e.g. $[E(x)]^2 \neq E(x^2)$. → unless $\text{Var}(x) = 0$

& ie. $E(x^2) = [E(x)]^2 + \text{Var}$

Discussion:

① The plim is the (large sample) counterpart to expected value wrt finite samples.

② Many textbook treatments of asymptotic distributions are wrong because they frequently use expectations (moments), which do not always exist.

verify → (e.g. Johnston, Kmenta)

Defn: Let $\hat{\theta}_T$ be a consistent estimator of θ & $\sqrt{T}(\hat{\theta}_T - \theta)$ converges in distribution to $N(0, \Omega)$. Then the Asymptotic Distribution of $\hat{\theta}$ is $N(\theta, \frac{1}{T}\Omega)$

(where Ω is the covariance matrix of $\hat{\theta}$.)

Intuitive expl.:

** $(\hat{\theta}_T - \theta) \to 0$ as fast as $\sqrt{T}$ changes as $T \to \infty$;

--- Instead of the probability density functions converging to a point (as with plim's),

their cumulative distribution functions converge to a known or fixed cum. dist. function.

Shady Comment:

Given $E(\hat{\theta}_T) \to \theta$ and $Cov(\hat{\theta}_T) \to \frac{1}{T}\Omega$,

it is often true that

$$\frac{\hat{\theta}_T - \theta}{\frac{\sqrt{\Omega}}{\sqrt{T}}} \to N(0,1)$$

— by Central Limit Thm.

Explain with examples!

Trivial Example: $\{x_1, \ldots, x_T\}$ is a r.s. from any distribution, with mean μ and variance σ^2.

Consider $\bar{X}_T = \frac{1}{T}\sum x_i$; $E(\bar{x}) = \mu$, $Var(\bar{x}) = \frac{\sigma^2}{T}$.

$\text{plim } \bar{x} = \mu$ by the Law of Large Numbers (shown before $\bar{x}$ consistent)

Then $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{T}}} \to N(0,1)$ by Central Limit Thm.

Note: $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{T}}} = \frac{\sqrt{T}(\bar{x}-\mu)}{\sigma} \to N(0,1)$.

$\star\star$ ..... Thus $\sqrt{T}(\bar{x}-\mu) \to N(0, \sigma^2)$.

Then it follows from our defn., that

$$\bar{x} \to N\left(\mu, \frac{1}{T}\sigma^2\right),$$

& we know this is true even for small samples.

"$\sqrt{T}$" is like a magic normalization factor.

Normally, wrt this example, we don't talk about $\bar{x}$ converging in distribution to this because this dist. is a $f(\frac{1}{T})$ & this is messy as $T \to \infty$. ~~~ ∴ Typically standardize $\bar{x}$ & then work with $N(0,1)$ which is not a $f(\frac{1}{T})$

Useful Fact:

A statistic $\sim N(\mu, \sigma^2)$ multiplied by a constant, c, is distributed $N(c\mu, c^2\sigma^2)$

i.e. $c * N(\mu, \sigma^2) \sim N(c\mu, c^2\sigma^2)$

Now

A less trivial example.

$\{X_1, \ldots, X_T\}$ a r.s. from $N(\mu, \sigma^2)$

$s^2 = \frac{1}{T-1} \sum_i (X_i - \bar{X})^2$ is a consistent estimator of σ^2.

Recall $\frac{(T-1)s^2}{\sigma^2} \sim \chi^2_{T-1}$ Mean: $T-1$ Var: $2(T-1)$

Consider

$$\sqrt{T}(s^2 - \sigma^2) = \sqrt{2}\sigma^2 \sqrt{\frac{T}{T-1}} \left[\frac{\frac{(T-1)s^2}{\sigma^2} - (T-1)}{\sqrt{2(T-1)}} \right] \downarrow$$

(Algebra) →

(Take $\lim_{T\to\infty}$) → $\sqrt{2}\sigma^2 \sqrt{\frac{T}{T+}} \cdot \frac{\frac{(T-1)(s^2-\sigma^2)}{\sigma^2}}{\sqrt{2}\sqrt{T-1}}$

$\sqrt{2}\sigma^2$ (1) $\to N(0,1)$

Hence, $\sqrt{T}(s^2 - \sigma^2) \longrightarrow \sqrt{2}\sigma^2 * N(0,1)$

or $\sqrt{T}(s^2 - \sigma^2) \longrightarrow N(0, 2\sigma^4)$ [by useful fact]

or The Asymptotic Dist of s^2 is $N(\sigma^2, \frac{2\sigma^4}{T})$

[by Defn, p. 18].

☆ Nice to know!

As. Variance of s^2 is Cramer-Rao lower bound $I(\theta)^{-1}$ → As. eff!

Useful Fact Restated: (more general form)

If A is a matrix } $\operatorname{plim} A$ exists,
& if ξ is a vector } $\xi \to N(0, \Omega)$,

then $A\xi \to N[0, (\operatorname{plim} A)\,\Omega\,(\operatorname{plim} A)']$.

Note: $\operatorname{plim} A \Rightarrow$ plim of each element in

In this example,
let $A = \sqrt{2}\,\sigma^2\sqrt{\frac{T}{T-1}}$; ξ = our standardized χ^2 statistic;

then from the Algebra, $\sqrt{T}A\xi = \sqrt{T}(s^2 - \sigma^2)$, previous page !!!

and $A\xi \to N[0, (\operatorname{plim} A)\,\Omega\,(\operatorname{plim} A)']$

or $\to N[0, (\sqrt{2}\,\sigma^2)\,1\,(\sqrt{2}\,\sigma^2)]$

or $\to N(0, 2\sigma^4)$.

*Here $\operatorname{plim} A = A = \sqrt{2}\,\sigma^2$

Defn:

If $\hat{\theta}$ & $\tilde{\theta}$ are consistent estimators of θ with asymptotic covariance matrices $\frac{1}{T}\Omega$ & $\frac{1}{T}\Psi$, respectively, then

$\hat{\theta}$ is <u>asymptotically efficient relative to</u> $\tilde{\theta}$ if $(\Psi - \Omega)$ is psd.

---(analagous to finite sample defn)

Defn: A consistent estimator $\hat{\theta}$ is <u>asymptotically efficient</u> if it is asymptotically efficient relative to every other consistent estimator of θ.

Defn: Let $\{x_1, \ldots, x_T\}$ be a r.s. from a population with density, $f(x;\theta)$. Then the <u>Maximum Likelihood Estimator</u> (<u>MLE</u>) of θ is the value that maximizes the Likelihood function, $L = \prod_{t=1}^{T} f(x_t;\theta)$, wrt θ.

<u>Intuition</u>: Picking the value of $\hat{\theta}$ that makes the sample observations most likely;
i.e. the value of $\hat{\theta}$ that makes it most likely that you would observe what you have observed!

Example: $\{x_1, \ldots, x_T\}$ is a r.s. from $N(\mu, \sigma^2)$.

Recall, $L = (2\pi)^{-\frac{T}{2}} (\sigma^2)^{-\frac{T}{2}} \exp\left\{\frac{-1}{2\sigma^2} \sum_i (x_i - \mu)^2\right\}$

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum (x_i - \mu)^2$$

maximize this wrt μ & σ^2.

Note: since the ln function is monotonic, the values that max. $(\ln L)$ also max. (L).

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\hat\sigma^2}\sum (x_i - \hat\mu) = 0$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{-T}{2\hat\sigma^2} + \frac{1}{2\hat\sigma^4}\sum (x_i - \hat\mu)^2 = 0$$

Convention: When we set 1st deriv.'s $= 0$, the parameters become <u>MLE's</u>, so put little hats on them ($\hat{}$).

Solving, we get:

$$\hat\mu = \bar X$$
$$\hat\sigma^2 = \frac{1}{T}\sum (x_i - \bar X)^2 \qquad \left[= \left(\frac{T-1}{T}\right) s^2\right]$$

<u>Very Important</u>

<u>Thm:</u> Under general conditions*, the MLE $\hat\theta$ of a parameter θ is consistent and asymptotically efficient. Furthermore,

$$\sqrt{T}\left(\hat\theta - \theta\right) \to N\left[0, \lim_{T\to\infty}\left(\frac{\mathcal{I}}{T}\right)^{-1}\right].$$

*See Appendix of Schmidt.

This Thm means that the Asymptotic Covariance Matrix of $\hat{\theta}$ is

$$\frac{1}{T}\lim_{T\to\infty}\left(\frac{\mathcal{I}}{T}\right)^{-1} = \frac{1}{T}\lim_{T\to\infty} T(\mathcal{I}^{-1})$$

$$\downarrow$$

$$= \mathcal{I}^{-1} \quad \text{[if } T \text{ is sufficiently large.]}$$

-- Simply the Information Matrix!
Hence MLE's are asymptotically efficient.

Thus we'll often work with MLE's --- they're good!
---- Many computer programs present information about the Likelihood function, e.g. $\ln L$.
MLE's will be the estimates that maximize this.
Hence, can change model as desired, & choose the one that reaches a higher point on the Likelihood surface...

Thm: Let $\hat{\theta}$ be any consistent estimator of θ.
Define the Linearized MLE as

(Convert inefficient estimator, $\hat{\theta}$ to efficient one $\tilde{\theta}$)

$$\tilde{\theta} = \hat{\theta} - \left[\frac{\partial^2 \ln L}{\partial\theta\,\partial\theta'}\right]^{-1}\left[\frac{\partial \ln L}{\partial\theta}\right]$$

$p\times 1$ $\quad$ $p\times 1$ $\quad$ $p\times p$ $\quad$ $p\times 1$

($p\times p$: similar to $\mathcal{I}$)

with all derivatives evaluated at $\hat{\theta}$.
Then $\tilde{\theta}$ is consistent and asymptotically efficient

This is a way of converting an asymptotically inefficient estimator into an asymptotically efficient one. ($\hat{\theta}$ into $\tilde{\theta}$)

Called the Newton-Rathson Algorithm or the Method of Scoring.
Used in computer programs;
spares the more costly work of evaluating all the MLE's.

[Intuition: get functional form of 1st & 2nd derivatives in the Algorithm; plug in values of $\hat{\theta}$ to get $\tilde{\theta}$.

It is good if $\hat{\theta}$ is close to begin with, not good if not...

$\hat{\theta}_T$ is <u>asymptotically unbiased</u> if

$$\lim_{T\to\infty} E(\hat{\theta}_T) = \theta.$$

Defn: Consider a sequence of random variabl
e.g. $\{\hat{\theta}_1, \hat{\theta}_2, \ldots\}$.
Then if $\exists$ a number c $\ni$

$$\lim_{T\to\infty} P(|\hat{\theta}_T - c| > \epsilon) = 0 \quad \forall \epsilon > 0$$

then the sequence <u>converges in probability</u> to c, and we write

$$\text{plim } \hat{\theta}_T = c.$$

Defn: An estimator $\hat{\theta}_T$ is <u>consistent</u> if
$\text{plim } \hat{\theta}_T = \theta$.

Defn: The <u>Mean Square Error</u> of $\hat{\theta}_T$ is

$$MSE(\hat{\theta}_T) = E(\hat{\theta}_T - \theta)^2.$$

Thm: If $\lim_{T\to\infty} MSE(\hat{\theta}) = 0$,
then $\hat{\theta}_T$ is consistent.

Thm: (Slutsky)
If $\text{plim } x_T = c$, and if g is any
continuous function, then

$$\text{plim } g(x_T) = g(\text{plim } x_T) = g(c)$$

Consider 2 Well-known Tests of Hypotheses.

i) Likelihood Ratio Test (Neyman-Pearson Application)

← Consider some function of θ, $f(\theta) = 0$. ← H_0 (null hypoth
$(m \times 1)$

a) Maximize L wrt θ;
observe L_u = maximized value of L, unrestricted.

$= L(X_1, \ldots, X_T;\ \theta = \hat{\theta} = \overset{\text{unrestr.}}{\text{MLE of } \theta})$

b) Maximize L s.t. $f(\theta) = 0$;
observe L_r = maximized value of L, restricted.
e.g. [get $\hat{\theta}$ s.t. H_0: $\theta = c$]

Obviously $L_u \geq L_r$.
The magnitude of their difference indicates whether the hypothesis that $f(\theta) = 0$ [H_0] can be accepted or rejected.
If $H_0: f(\theta) = 0$ is true,
L_u should approximately = L_r;
i.e. the constraint should not be severely binding.

$\therefore$ Under H_0, $-2 \ln\left(\frac{L_r}{L_u}\right) \rightarrow \chi^2_m$

Reject H_0 if $-2 \ln\left(\frac{L_r}{L_u}\right) = 2(\ln L_u - \ln L_r)$ is large.

ii) The Wald Test

Let $\hat{\theta}$ = the unrestricted MLE of θ.

As a test statistic, consider $f(\hat{\theta})$; is it close to 0?

H_0: for $\theta = \alpha$, $f(\theta) = \theta - \alpha$.

Suppose we know the asymptotic dist. of $\hat{\theta}$:

$$\sqrt{T}\, f(\hat{\theta}) \rightarrow N(0, \Omega)$$

"Suppose" ← It's not. To comes from the property of MLE.

$\sqrt{T}(\hat{\theta} - \theta) \rightarrow N(0, \lim_{T\to\infty}(\frac{I}{T})^{-1})$

Under H_0: $f(\theta)$

[i.e. suppose we can prove the Central Limit Thm in this case.]

Then

$$T\left[f(\hat{\theta})' \, \Omega^{-1} \, f(\hat{\theta})\right] \rightarrow \chi^2_m$$

$1 \times m \quad m \times m \quad m \times 1$

* Same dist. as Likelihood Ratio Test.

* In calculating this test statistic, evaluate Ω^{-1} at $\hat{\theta}$ (where-ever θ appears).

* Advantage: doesn't require construction of constrained MLE.

* See Judge, p.757

EXAM 2

Consider a random sample $\{X_1, \ldots, X_T\}$

from a population with density

$$f(X) = \frac{1}{\theta} e^{-\frac{1}{\theta} X} \qquad \text{for } X \geq 0$$

A. Find the Maximum Likelihood Estimator of θ.

B. Prove that it is unbiased.

C. Prove that it is efficient.

D. How would you test the null hypothesis,

$$H_0: \theta - 3 = 0 \quad ?$$

Useful Fact: The mean of the above distribution is θ, and the variance is θ^2.