VI Frequentist Theories

- start with the basic ingredients
  $(\mathcal{M}, x)$ = (model, data)

- now "imagine" a hypothetical sequence of repeated samples (generated under the same conditions as the observed data x) so these are all i.i.d. f_θ for same θ (the true value) $x_1', x_2', \ldots \overset{iid}{\sim} f_\theta$

eg - Ω = population of students at U of T

- $X(\omega)$ = ht of ω in cm
- f_X = dist of ht over Ω

- $\mathcal{M} = \{f_\theta : \theta \in \Theta\}$ = location-scale normal model

  - take a sample of $n=100$, measure their hts to obtain data $x = (x_1, \ldots, x_{100})$ (sampling without replacement $\approx$ iid sampling)

  - now imagine repeating this process infinitely many times obtaining $x_1', x_2', \ldots$ (so x_i' is comprised of ht measurements of 100 students)

  - then the basic idea behind all frequentist theories of inference is the following.

Principle of Frequentism

Inferences are justified by their good properties in an infinite sequence of performances $x_1, x_2, \ldots$ where these properties should hold for every $\theta \in \Theta$

eg location normal $x = (x_1, \ldots, x_n) \overset{iid}{\sim} N(\mu, 1)$

- inference : estimate μ
- $t_1(x) = \bar{x}$, $t_2(x) = \frac{x_1 + x_n}{2}$
- what is the average value of these estimators in an infinite sequence $x^1, x^2, \ldots$?
- when μ is true $t_1(x) \sim N(\mu, \frac{1}{n})$ and $t_2(x) \sim N(\mu, \frac{1}{2})$
- $E_\mu(t_1(x)) = \mu$, $E_\mu(t_2(x)) = \mu$ for every $\mu \in R^1$, so t_1 and t_2 are unbiased estimators (this answers the question by SLLN)
- so both have this "good" property
- how about comparing them with respect to accuracy

- how do we measure accuracy?
- one way: mean squared error

Def The mean-squared error of estimator t of $\psi = \Psi(\theta) \in \mathbb{R}$ is defined as

$$MSE_\theta(t) = E_\theta\left((t(s) - \Psi(\theta))^2\right)$$

- **note** - when $E_\theta(t(s)) = \Psi(\theta)$ $\forall\theta$ (t is unbiased) then

$$MSE_\theta(t) = Var_\theta(t)$$

eg location model cont'd

$$Var_\mu(t_1) = 1/n$$

$$Var_\mu(t_2) = \frac{1}{2}$$

- so t_1 is better than t_2 whenever $n > 2$ $\forall\mu$.

- why this criterion?
- why not MAD = mean absolute deviation

$$MAD_\theta(t) = E_\theta(|t(x) - \Psi(\theta)|)$$

- theory hurder

location normal

- $t_3(x)$ = sample median

$$= \begin{cases} \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2} & n \text{ even} \\ x_{(\lfloor n/2 \rfloor + 1)} & n \text{ odd} \end{cases}$$

fact

- $MAD_\mu(t_3) \leq MAD_\mu(t_1) \quad \forall \mu$

- questions about the principle of frequentism

(1) - why do we care about the long-run performance of statistical procedures?

- there doesn't seem to be a good answer, generally people who believe that probabilities can only correspond to long-run frequencies

- vague claims about objectivity but model dependent and thus subjective.

(2) - what criteria do we use to compare different procedures?

- we have seen that different criteria lead to different inferences.

* given that we choose the criteria these are subjective but do not satisfy the principle of empirical criticism.

  eg how do we check whether MSE or MAD are appropriate?

- we discuss P-values, confidence regions and optimality theory

(1) P-values (Tests of Significance)

- suppose we have a hypothesis

$H_0 : \Psi(\theta) = \psi_0$

- based on $(\mathcal{M}, x)$ we want to determine whether or not we have evidence that H_0 is true or false?

eg - $x = (x_1, \ldots, x_n) \overset{iid}{\sim}$ Bernoulli(θ)

where $\theta \in [0,1]$

- $H_0 : \theta = \frac{1}{2}$.

- a large # of heads or tails should be evidence against H_0 while about 50% heads should be evidence for H_0.

- what about $H_0 : \theta = \frac{1}{8}$?

- basic p-value idea for assessing $H_0 : \Psi(\theta) = \psi_0$

- find a statistic (any function of the data) T such $T(x)$ measures, in some sense, discrepancies from H_0 and the distribution of T when H_0 is true needs to be p = probability that in a future sample we will obtain a value $T \geq T(x)$ given that we observed $T(x)$ in this sample

known.

- then compute $P_{H_0,T}(T \geq T(x) \mid T(x))$

and interpret small values as evidence <u>against</u> H_0.

eg - $x = (x_1, \dots, x_n)' \sim \text{Bernoulli}(\theta)$

$H_0: \theta = 1/2$

- when H_0 is true

$S = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, 1/2)$

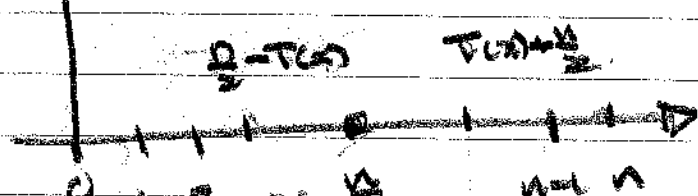
- let $T(x) = |n\bar{x} - n\tfrac{1}{2}|$

- pvalue $= P_{H_0,T}(T \geq T(x))$ p

$$= P_{1/2}\left(S \leq \tfrac{n}{2} - T(x)\right) + P_{1/2}\left(S \geq T(x) + \tfrac{n}{2}\right)$$

$T < T(x)$ iff

$-T(x) < S - \frac{n}{2} < T(x)$



$$= \sum_{k=0}^{\lfloor \frac{n}{2} - T(x) \rfloor} \binom{n}{k} \left(\tfrac{1}{2}\right)^n + \sum_{k=\lceil T(x) + \frac{n}{2} \rceil}^{n} \binom{n}{k} \left(\tfrac{1}{2}\right)^n$$

- what about $H_0 : \theta = \theta_0$?

- $T(x) = |n\bar{x} - n\theta_0|$ $\quad -T(x) \le S - n\theta_0 \le T(x)$

- p-value $= P_{\theta_0}(T \ge T(x))$

$$= \sum_{k=0}^{\lfloor n\theta_0 - T(x) \rfloor} \binom{n}{k} \theta_0^k (1-\theta_0)^{n-k} + \sum_{k=\lceil T(x) + n\theta_0 \rceil}^{n} \binom{n}{k} \theta_0^k (1-\theta_0)^{n-k}$$

$T \le T(x)$

$-\sqrt{n\theta_0(1-\theta_0)}\, T(x) \le S(x) - n\theta_0$
$\le \sqrt{n\theta_0(1-\theta_0)}\, T(x)$

- how small should the p-value be to say we have evidence against H_0? sorites paradox

- big problem: why can't a p-value be viewed as evidence in favor of H_0 when p-value is big?

- <u>recall</u> - suppose $X \sim f_\theta$ with strictly increasing cdf F_θ

  - then put $U = F_\theta(X) \in [0,1]$

  - $P_\theta(U \le u) = P_\theta(F_\theta(X) \le u)$

9

cont case

$= P_\theta(X \le F_\theta^{-1}(u))$ — cdf 1-1 increasing

$= F_\theta(F_\theta^{-1}(u)) = u$

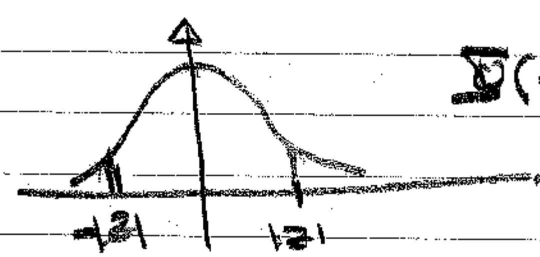
$\therefore U \sim U(0,1)$ independent of θ.

eg <u>location-normal</u>

- $x = (x_1, \dots, x_n)' \sim N(\mu, 1) \quad \mu \in \mathbb{R}^1$
- assess $H_0: \mu = \mu_0$
- $T(x) = \sqrt{n}\,|\bar{x} - \mu_0|$ (Z-statistic)

  and $\bar{x} \overset{H_0}{\sim} N(\mu_0, \frac{1}{n})$ so $Z = \sqrt{n}(\bar{x} - \mu_0) \sim N(0,1)$
- p-value $P_{\mu_0}(T \ge T(x))$

  $= P_{\mu_0}(|Z| \ge \sqrt{n}\,|\bar{x} - \mu_0|)$



$\Phi(-z) = (1 - \Phi(|z|))$

$= 2(1 - \Phi(\sqrt{n}\,|\bar{x} - \mu_0|))$

- what values do we expect when H_0 is true?

- what is the distribution of

$$2(1-\Phi(\sqrt{n}|\bar{X}-\mu_0|))$$

when H_0 is true?

- $P_{\mu_0}(2(1-\Phi(\sqrt{n}\,|\bar{X}-\mu_0|)) \le u) \quad u \in [0,1]$

$$= P_{\mu_0}(\Phi(\quad) \ge 1-\tfrac{u}{2})$$

$$= P_{\mu_0}(\sqrt{n}|\bar{X}-\mu_0| \ge \Phi^{-1}(1-\tfrac{u}{2}))$$

$$= 2(1-\Phi(\Phi^{-1}(1-\tfrac{u}{2})))$$

$$= 2(1-(1-\tfrac{u}{2})) = u.$$

$\therefore$ the p-value is distributed $U(0,1)$ when H_0 is true

- so when H_0 is true all values are equally likely (independent of n)

- note when H_0 is false say $\mu = \mu_1 \ne \mu_0$

$$P_{\mu_1}(2(1-\Phi(\sqrt{n}|\bar{X}-\mu_0|)) \le u)$$

$$= P_{\mu_1}(\sqrt{n}|\bar{X}-\mu_0| \ge \Phi^{-1}(1-\tfrac{u}{2}))$$

$$= 1 - P_{\mu_1}(\sqrt{n}|\bar{X}-\mu_0| \le \Phi^{-1}(1-\tfrac{u}{2}))$$

$$= 1 - P(-\Phi^{-1}(1-\tfrac{u}{2}) \le \sqrt{n}(\bar{X}-\mu_0) \le \Phi^{-1}(1-\tfrac{u}{2})$$

$$= 1 - P_{\mu_{true}}\left(\sqrt{n}(\mu_0 - \mu_{true}) - \Phi^{-1}(1-\tfrac{\alpha}{2}) \le \sqrt{n}(\bar{x} - \mu_{true}) \le \sqrt{n}(\mu_0 - \mu_{true}) + \Phi^{-1}(1-\tfrac{\alpha}{2}) \right)$$

$$= 1 - \left(\Phi\left(\sqrt{n}(\mu_0 - \mu_{true}) + \Phi^{-1}(1-\tfrac{\alpha}{2})\right) - \Phi\left(\sqrt{n}(\mu_0 - \mu_{true}) - \Phi^{-1}(1-\tfrac{\alpha}{2})\right) \right)$$

as $n \to \infty$: first term $\to \begin{cases} 1 & \mu_0 > \mu_{true} \\ 0 & \mu_0 < \mu_{true} \end{cases}$, second term $\to \begin{cases} 1 \\ 0 \end{cases}$

$\to 1$ as $n \to \infty$.

$\therefore$ the p-value converges to 0 as $n \to \infty$.

- so P-value does the right thing when H_0 is false but doesn't when H_0 is true.

= note - the real difference we are interested in is $|\mu_{true} - \mu_0|$ and this is estimated by $|\bar{x} - \mu_0|$

- there is a δ s.t. if $|\mu_{true} - \mu_0| < \delta$ then $H_0: \mu = \mu_0$ is "true"

- but $\bar{x} \sim N(\mu_{true}, \frac{1}{n})$ so we could have $|\mu_{true} - \mu_0| < \delta$ but p-value $2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_0|))$ is very small

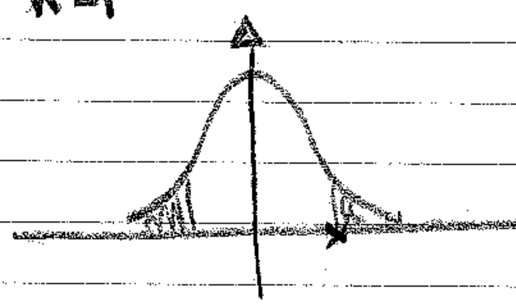
preferred method of avoiding assessment that incorporates δ

- moral: always estimate $|\mu_{true} - \mu_0|$ when p-value small.

eg <u>location-scale normal</u> t-test

- want to assess $H_0: \mu = \mu_0$
- $\sqrt{n}(\bar{x}-\mu_0)/s_x$

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

  $\sim t_{n-1}$ when H_0 is true.
- 
- take $T(x) = |\sqrt{n}(\bar{x}-\mu_0)/s_x|$
- p-value $= P_{H_0}(T \geq T(x))$
- <u>note</u> when $\mu \neq \mu_0$ $\bar{x} \xrightarrow{SLLN} \mu$, $s_x \to \sigma$

  and so $|\sqrt{n}(\bar{x}-\mu_0)/s_x| \to \infty$ while

  $t_n \to N(0,1)$ so p-value $\to 0$

$= 2(1 - G_{n-1}(|\sqrt{n}(\bar{x}-\mu_0)/s_x|))$

where G_{n-1} is the cdf of the t_{n-1} distribution

- again p-value ~ $U(0,1)$ when H_0 is true & and p-value $\to 0$ as $n \to \infty$ when H_0 is false

Eg Likelihood ratio test

- suppose $\psi = \Psi(\theta)$

relative profile likelihood

- put $T(x) = -2\ln L(\theta_{H_0}(x) \mid x) / L(\theta_{MLE}(x) \mid x)$

- so evidence against when $T(x)$ big

- under conditions when $\Psi(\theta) = \psi_0$ $T(x) \xrightarrow{D}$ chi-squared($\dim \Theta - \dim \Psi^{-1}\{\psi_0\}$)
  (dimension of $\{\theta_0\}$ is 0)

- note $\Psi(\theta) = I_{H_0}(\theta)$, $\Psi^{-1}\{1\} = H_0$

$L_\Psi(1 \mid x) = \sup_{\theta \in H_0} L(\theta \mid x) = L(\theta_{H_0}(x) \mid x)$

so just relative profile likelihood

(2.) Rejection Trials and Confidence Regions

- how small does the p-value have to be to say we have evidence against?

- choose $\alpha \in [0,1]$, e.g. $\alpha = 0.05$ and say we have significant results whenever p-value $\leq \alpha$

- suppose we measure hts in cm and want to assess $H_0: \mu = \mu_0$

- what kind of difference are we interested in?

- certainly not mm let $\mu_1 = \mu_0 + \delta$ where $\delta > 0$ is less than the difference we view as being practically significant

- using t-test we know that we will reject H_0 at level α when n is large enough

- so statistical significance is not the same as practical significance

- if you get a statistically significant result estimate the deviation from H_0 to see if you have detected something of practical significance

- so estimate μ by $\bar{x}$ to see if $\bar{x}$ differs from μ_0 by a practically significant amount

- rejection trials

- suppose interest is in $\psi = \Psi(\theta)$ and for each value of ψ_0 we have a procedure such that $R_{\psi_0}(x) \in \{0,1\}$ and $R_{\psi_0}(x) = 1$ means we reject H_0 (and $R_{\psi_0}(x) = 0$ means no rejection) and $P_\theta(R_{\psi_0}(x) = 1) \le \alpha$ whenever $\Psi(\theta) = \psi_0$

eg - $R_{\psi_0}(x) = 1$ whenever a p-value $\le \alpha$ when p-value $\sim U(0,1)$ when H_0 true

- note typically $P_{H_0}(R_{\psi_0}(x) = 1) = \alpha$ in this case due to the uniform distribution of the p-value

eg data snooping

- a test stat. T_n is used to test H_0 and H_0 is rejected when $P_{H_0}(T_n \geq T(x)) \leq \alpha$, where $x = (x_1, \ldots, x_n)$

- investigator takes sample of n and p-value $> \alpha$ but close.

- so decides to take a further sample of m say $x_{n+1}, \ldots, x_{n+m}$ and compute $P_{H_0}(T_{n+m} \geq T(x_1, \ldots, x_{n+m}))$

- what is the probability that H_0 is rejected when H_0 is true?

- two stage procedure

- so P_{H_0}("rejecting H_0")
$= P_{H_0}$("rejecting H_0 at first stage") $+$
P("rejecting H_0 at second stage and not at 1st stage")
$= \alpha + (>0) > \alpha$ so can never reject at level α.

confidence regions

- $\psi = \Psi(\theta)$
- rejection region $R_{\psi_0}(x) = \begin{cases} 1 & \text{reject } H_0: \Psi(\theta) = \psi_0 \\ 0 & \text{accept} \end{cases}$
- $P_\theta(R_{\psi_0}(x)=1) \leq \alpha \quad \forall \theta \in \Psi^{-1}\{\psi_0\}$

- $C(x) = \{\psi_0 : R_{\psi_0}(x) = 0\}$ = set of ψ_0 not rejected when x observed
- C is a random set
- coverage probability

$= P_\theta(\Psi(\theta) \in C(x))$ = prob. C contains true value

$= P_\theta(R_{\Psi(\theta)}(x) = 0) = 1 - P_\theta(R_{\Psi(\theta)}(x) = 1)$

$\geq 1-\alpha \quad \forall \theta$

we call C a $(1-\alpha)$-confidence region for $\psi = \Psi(\theta)$

- <u>note</u> $P_\theta(\Psi(\theta) \in C(x))$ is not the probability that a particular $C(x)$ contains $\Psi(\theta)$ (frequentism)

<u>eg</u> <u>location-scale normal</u>

- $P_{\mu_0}(x) = 2\left(1 - G_{n-1}\left(\left|\frac{\bar{x}-\mu_0}{s_x/\sqrt{n}}\right|\right)\right)$

- $C(x) = \{\mu_0 : P_{\mu_0}(x) > \alpha\}$

$$= \{\mu_0 : G_{n-1}\left(\left|\frac{\bar{x}-\mu_0}{s_x/\sqrt{n}}\right|\right) < 1-\alpha/2\}$$

$$= \{\mu_0 : \left|\frac{\bar{x}-\mu_0}{s_x/\sqrt{n}}\right| < \underbrace{G^{-1}_{n-1}(1-\alpha/2)}_{t_{n-1,1-\alpha/2}}\}$$

$$= \{\bar{x} - s_x/\sqrt{n}\, t_{n-1,1-\alpha/2} \le \mu_0 < \bar{x} + \frac{s_x}{\sqrt{n}} t_{n-1,1-\alpha/2}\}$$

$$= \left(\bar{x} - \frac{s_x}{\sqrt{n}} t_{n-1,1-\alpha/2},\ \bar{x} + \frac{s_x}{\sqrt{n}} t_{n-1,1-\alpha/2}\right)$$

- <u>note</u> $\bar{x} \in C(x)$ and the length of the interval $\frac{2 s_x}{\sqrt{n}} t_{n-1,1-\alpha/2}$ gives an assessment of the accuracy of the estimate.

- suppose C is a $(1-\alpha)$ confidence region for $\psi = \Psi(\theta)$, namely,

$$P_\theta\left(\Psi(\theta) \in C(x)\right) \geq 1-\alpha \quad \forall \theta$$

- put $R_{\psi_0}(x) = 1 - I_{C(x)}(\psi_0)$

- then $P_\theta\left(R_{\Psi(\theta)}(x) = 1\right)$

$$= P_\theta\left(\Psi(\theta) \notin C(x)\right) \leq 1-(1-\alpha) = \alpha.$$

- this gives a size α rejection procedure

- in general - for a particular problem $H_0: \Psi(\theta) = \psi_0$ how do we find an appropriate p-value or rejection trial?

3 Optimality Theory (Frequentist) Chapter 5

(a) Decision Theory

- ingredients : data x , model $\mathcal{M}$
  for inference about $\psi = \Psi(\theta)$ add a
  loss function $L : \Theta \times \Psi \to [0, \infty)$
  where $L(\theta, \psi) = 0$ iff $\Psi(\theta) = \psi$
- the bigger $L(\theta, \psi)$ the greater is the loss
  eg $x = (x_1, \ldots, x_n) \sim N(\mu, 1)$ $\mu \in \mathbb{R}^1$ where $\psi = \Psi(\mu) = \mu$, $L(\mu, \mu') = (\mu - \mu')^2$ quadratic loss.
- we look for a decision function
  $\delta : \mathcal{X} \to \Psi$ that in some sense
  minimizes $L(\theta, \delta(x))$ eg (cont'd) $\delta(x) = \bar{x}$
- how ?
- the risk function associated with δ is defined as

$$R(\theta, \delta) = E_\theta(L(\theta, \delta(x)))$$

$$= \begin{cases} \sum_{x \in \mathcal{X}} L(\theta, \delta(x)) f_\theta(x) & \text{discrete case} \\ \int_{\mathcal{X}} L(\theta, \delta(x)) f_\theta(x)\, dx & \text{cont. case} \end{cases}$$

we want δ s.t. $R_\delta(\theta)$ is uniformly small relative to other δ'

= average loss incurred (eg (cont'd) $R(\mu,\delta) = E_\mu(\bar{x}-\bar{x})^2 = \frac{1}{}$

- we then define an optimal decision function as follows

Def A decision function δ_{opt} is <u>optimal</u> if $R(\theta,\delta_{opt}) \le R(\theta,\delta)$ $\forall\, \theta \in \Theta$

- for each $\tau \in \Psi$ define δ_τ by $\delta_\tau(x) = \tau \quad \forall x$.

- if $\Psi(\theta) = \tau$, then

$$R(\theta,\delta_{\Psi(\theta)}) = E_\theta\left(L(\theta,\delta_{\Psi(\theta)}(x))\right)$$
$$= E_\theta\left(L(\theta,\Psi(\theta))\right) = 0$$

which is the smallest possible risk.

Lemma Suppose $\exists\, \delta_{opt}$. Then $R(\theta,\delta_{opt}) = 0$ $\forall \theta \in \Theta$.

Proof: $R(\theta,\delta_{opt}) \le R(\theta,\delta_{\Psi(\theta)}) = 0$ whenever $\theta \in \Psi^{-1}\{\tau\}$

Theorem If $T(x) \geq 0$ and $E_P(T) = 0$ then $P(\{x : T(x) = 0\}) = 1$

Proof: (discrete case)

$$E_P(T) = \sum_{x \in \mathcal{X}} T(x) P(\{x\}) = \sum_{\substack{x \in \mathcal{X} \\ P(\{x\}) > 0}} T(x) P(\{x\})$$

$= 0$ which says that $T(x) = 0$ whenever $P(\{x\}) > 0$

and note $\sum_{x, P(\{x\}) > 0} P(\{x\}) = 1$.

- so $0 = R(\theta, \delta) = E_\theta(L(\theta, \delta(x)))$
iff $L(\theta, \delta(x)) = 0$ with θ-prob 1
iff $\delta(x) = \Psi(\theta)$ with θ-prob 1

- let $\mathcal{X}_\theta = \{x : P_\theta(x) > 0\}$ = support of P_θ

Lemma If $\exists\ \theta_1, \theta_2 \in \Theta$ s.t. $\Psi(\theta_1) \neq \Psi(\theta_2)$ and $P_{\theta_1}(\mathcal{X}_{\theta_1} \cap \mathcal{X}_{\theta_2}) > 0$, $P_{\theta_2}(\mathcal{X}_{\theta_1} \cap \mathcal{X}_{\theta_2}) > 0$ then $\nexists$ an optimal δ.

Proof: Suppose δ_{opt} exists then when $x \in \mathcal{X}_{\theta_1} \cap \mathcal{X}_{\theta_2}$ we have that $\delta(x) = \Psi(\theta_1)$ and $\delta(x) = \Psi(\theta_2)$ X

- solution :- restrict the search for optimal δ to a subset of all decision functions.

  - so we might require that a decision function have property A and then look for the optimal δ among all those having property A

(b) Optimal Estimation

- suppose $\psi = \Psi(\theta)$ is Euclidean valued so $\Psi \subseteq \mathbb{R}^k$
- many different possible losses L
- "mostly" people use $L(\theta, \psi) = \|\theta - \psi\|^2$

  = squared Euclidean error (standard quadratic loss)
- then $R(\theta, \delta) = E_\theta(\|\theta - \delta(x)\|^2) = MSE_\theta(\delta)$

Theorem (Variance-bias decomposition) When $E_\theta(|\delta_i(x)\delta_j(x)|) < \infty \ \forall i, j, \theta$ then

$$MSE_\theta(\delta) = \underbrace{tr\, Var_\theta(\delta)}_{\text{variance matrix of } \delta} + \|\underbrace{\Psi(\theta) - E_\theta(\delta)}_{\text{bias}}\|^2$$

note① – $\underset{\sim}{X} \in \mathbb{R}^k$ a random vector

– $E(\underset{\sim}{X}) \overset{def}{=} \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_k) \end{pmatrix}$ = mean vector $\in \mathbb{R}^k$

provided $E_\theta(|X_i|) < \infty \quad i = 1, \ldots, k$.

– $Var(\underset{\sim}{X}) \overset{def}{=} \left(E((X_i - E(X_i))(X_j - E(X_j))) \right)$

= variance matrix $\in \mathbb{R}^{k \times k}$

= $(Cov(X_i, X_j))$

provided $E(|X_i X_j|) < \infty \quad \forall i,j$

- note② $E(|Y|) = \sum_y |y| P(Y=y) < \infty$

- $E(Y) = \sum_y y P(Y=y)$

$= \sum_{y \geq 0} y P(Y=y) + \sum_{y<0} y P(Y=y)$

$= \sum_{y \geq 0} |y| P(Y=y) - \sum_{y<0} |y| P(Y=y)$

so $E(Y) \in \mathbb{R}$ iff $E(|Y|) < \infty$

- note③ $|Y| = |Y-c+c| \leq |Y-c| + |c|$

$\leq |Y| + |c| + |c|$

$\therefore E|Y| < \infty$ iff $E|Y-c| < \infty$

<u>note</u> 4 - suppose $E|Y|^k < \infty$ for $k \geq 2$, then $E|Y|^{k-1} < \infty$

Proof: $\sum |y|^{k-1} P(Y=y)$

$$= \sum_{|y| \leq 1} |y|^{k-1} P(Y=y) + \sum_{|y|>1} |y|^{k-1} P(Y=y)$$

$$\leq \sum_{|y| \leq 1} P(Y=y) + \sum_{|y|>1} |y|^k P(Y=y)$$

$$\leq P(|Y| \leq 1) + \sum_y |y|^k P(Y=y)$$

$$= P(|Y| \leq 1) + E|Y|^k < \infty$$

doesn't work in the other direction

<u>note</u> 5 $E|X_i|^2 < \infty \ \forall i$ implies

$E((X_i - E(X_i))(X_j - E(X_j))) = cov(X_i, X_j) \in \mathbb{R}$

Proof: $E|X_i| < \infty$ by note 4 and so $EX_i \in \mathbb{R}$.

and $|X_i X_j| = |X_i||X_j| \le |X_i|^2 + |X_j|^2$
so $E X_i X_j \in \mathbb{R}^1$. Thus $Cov(X_i, X_j) =$
$E(X_i X_j - X_i E(X_j) - X_j E(X_i) + E(X_i)E(X_j)) < \infty$

$- Cov(X_i, X_i) = Var(X_i)$

note ⑥ $Var(\underset{\sim}{X}) = \left(Cov(X_i, X_j) \right)$

$= E\left((\underset{\sim}{X} - E(\underset{\sim}{X}))(\underset{\sim}{X} - E(\underset{\sim}{X}))^t \right)$

$= E\left(\underset{\sim}{X}\underset{\sim}{X}^t - \underset{\sim}{X}(E(\underset{\sim}{X}))^t - E(\underset{\sim}{X})\underset{\sim}{X}^t + E(\underset{\sim}{X})(E(\underset{\sim}{X}))^t \right)$

$= E(\underset{\sim}{X}\underset{\sim}{X}^t) - E(\underset{\sim}{X})(E(\underset{\sim}{X}))^t$

note ① $E(\underset{\sim}{x}^t \underset{\sim}{x}) = E(tr\, \underset{\sim}{x}^t \underset{\sim}{x})$

$= E(tr\, \underset{\sim}{x}\underset{\sim}{x}^t) = tr\, E(\underset{\sim}{x}\underset{\sim}{x}^t)$

$\therefore E((\underset{\sim}{x} - E(\underset{\sim}{x}))^t (\underset{\sim}{x} - E(\underset{\sim}{x})))$

$= tr\, E((\underset{\sim}{x} - E(\underset{\sim}{x}))(\underset{\sim}{x} - E(\underset{\sim}{x})))$

$= tr\, Var(\underset{\sim}{x})$

Proof: $MSE_\theta(s)$

$= E_\theta(\|s(x) - \Psi(\theta)\|^2)$

$= E_\theta((s(x) - \Psi(\theta))^t (s(x) - \Psi(\theta)))$

$= tr\, E_\theta((s(x) - \Psi(\theta))(s(x) - \Psi(\theta))^t)$

$= tr\, E_\theta((s(x) - E_\theta(s) + E_\theta(s) - \Psi(\theta))(\quad)^t)$

$= tr\, E_\theta((s(x) - E_\theta(s))(s(x) - E_\theta(s))^t)$

$+ tr\, E_\theta(s(x) - E_\theta(s))(E_\theta(s) - \Psi(\theta))^t$

$+ tr\, E_\theta((E_\theta(s) - \Psi(\theta))(s(x) - E_\theta(s))^t)$

$+ tr\, E_\theta((E_\theta(s) - \Psi(\theta))(E_\theta(s) - \Psi(\theta))^t)$

$= tr\, Var_\theta(s) + \|E_\theta(s) - \Psi(\theta)\|^2$

Theorem (Rao-Blackwell)

If T is a sufficient statistic for model $\mathcal{M}$ and δ_β is an estimator of $\psi = \Psi(\theta) \in \mathbb{R}^k$ then for $\delta_T(x) = E(\delta \mid T(x))$ we have

(i) $E_\theta(\delta_T) = E_\theta(\delta) \quad \forall \theta \in \Theta$

(ii) $MSE_\theta(\delta_T) \le MSE_\theta(\delta)$.

Proof: (i) We have that $E_\theta(\delta)$

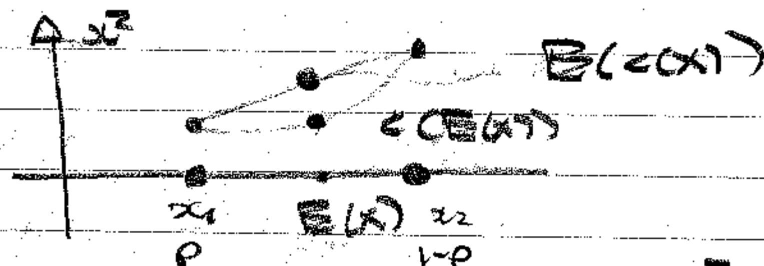
$$\overset{TTE}{=} E_\theta(E_\theta(\delta \mid T)(T(x))) \overset{T\text{ suff}}{=} E_\theta(E(\delta \mid T)(T(x)))$$

$$= E_\theta(\delta_T).$$

(ii) We use Jensen's inequality

Lemma If $E(X)$ exists and c is convex then
$$E(c(X)) \ge c(E(X)).$$

$c(\alpha x_1 + (1-\alpha)x_2)$
$\le \alpha c(x_1) + (1-\alpha)c(x_2)$
$\alpha \in [0,1]$



So $MSE_\theta(\delta) = E_\theta(\|\delta - \Psi(\theta)\|^2)$

$$= E_\theta(E(\|\delta - \Psi(\theta)\|^2 \mid T)(T(x)))$$

$$\overset{\text{Jensen}}{\ge} E_\theta(\|E(\delta \mid T(x)) - \Psi(\theta)\|^2)$$

see for convexity of $\|\cdot\|^2$

$$= MSE_\theta(\delta_T)$$

note - if $T(x_1) = T(x_2)$ then $\delta(x_1) = \delta(x_2)$ and so we restrict to fns of T

- uniformly minimum variance unbiased (UMVU) estimation

- there is no estimator that minimizes MSE_θ uniformly so restrict the s's we consider

Def' s is unbiased for $\Psi(\theta)$ if $E_\theta(s) = \Psi(\theta)\ \forall \theta \in \Theta$

- so $MSE_\theta(s) = \operatorname{tr} Var_\theta(s)$ when s is unbiased

- if T is a statistic then the collection $M_T = \{f_{\theta T} : \theta \in \Theta\}$ is the model for the statistic.

Def A model M is complete if whenever for $h : \mathcal{X} \to \mathbb{R}$ is st $E_\theta(|h|) < \infty$ and $E_\theta(h) = 0\ \forall \theta \in \Theta$ then $P_\theta(\{x : h(x) \neq 0\}) = 0\ \forall \theta \in \Theta$.

- in other words a model is complete if the only unbiased estimator of 0 is 0

eg $x \sim N(\mu, \sigma_0^2)$ $\mu \in \mathbb{R}$, $\sigma_0^2 > 0$ known is complete

$$0 = E_\mu(h) = \int_{-\infty}^{\infty} \frac{h(x)}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2\sigma_0^2}(x-\mu)^2\}\,dx$$

$$= \frac{e^{-\frac{1}{2}\frac{\mu^2}{\sigma_0^2}}}{\sqrt{2\pi}}\Big[\int_{\{h(x)>0\}} \exp\{\tfrac{x\mu}{\sigma_0^2}\}\,|h(x)|\,e^{-\frac{1}{2}\frac{x^2}{\sigma_0^2}}\,dx$$

$$\forall \mu \qquad - \int_{\{h(x)<0\}} \exp\{\tfrac{x\mu}{\sigma_0^2}\}\,|h(x)|\,e^{-\frac{1}{2}\frac{x^2}{\sigma_0^2}}\,dx\Big]$$

So, in particular ($\mu = 0$) $\int_{\{h(x)>0\}} |h(x)|\, e^{-\frac{1}{2}\frac{x^2}{\sigma_0^2}}\,dx = \int_{\{h(x)<0\}} |h(x)|\, e^{-\frac{1}{2}\frac{x^2}{\sigma_0^2}}\,dx$

and so, if these integrals are positive,

$$\frac{\int_{\{h(x)>0\}} \exp\{\mu x\} |h(x)| e^{-\frac{1}{2}x^2} dx}{\int_{\{h(x)>0\}} |h(x)| e^{-\frac{1}{2}x^2} dx} \stackrel{\forall \mu}{=} \frac{\int_{\{h(x)<0\}} \exp\{\mu x\} |h(x)| e^{-\frac{1}{2}x^2} dx}{\int_{\{h(x)<0\}} |h(x)| e^{-\frac{1}{2}x^2} dx}$$

$\forall \mu$. Since the LHS is the mgf of a dist concentrated on $\{h(x)>0\}$ and the RHS is the mgf of a dist concentrated on $\{h(x)<0\}$ this is impossible. and so

$$\int_{\{h(x)>0\}} |h(x)| e^{-\frac{1}{2}x^2} dx = \int_{\{h(x)<0\}} |h(x)| e^{-\frac{1}{2}x^2} dx = 0$$

and so $P_\mu(\{x : h(x) \neq 0\}) = 0 \ \forall \mu$

<u>Theorem</u> (<u>Lehman-Scheffé</u>) If T is a complete mss and S is a function of T with finite 2nd moments then S is UMVU for $E_\theta(S)$.

Proof: Suppose S' also satisfies $E_\theta(S') = E_\theta(S)$. Then $0 = E_\theta(S'-S)$ and so $P_\theta(\{x : S(T(x)) = S'(T(x))\}) = 0 \ \forall \theta$ Note any S' can be improved eg **location - normal model**

: $x = (x_1, \ldots, x_n) \sim N(\mu, \sigma_0^2)$, $\mu \in \mathbb{R}$ where σ_0^2 known.

- $\bar{x}$ is a MSS $\sim N(\mu, \sigma_0^2/n)$

  $\therefore \bar{x}$ is a complete MSS

- $E_\mu(\bar{x}) = \mu \quad \forall \mu$

  $\therefore$ the unique UMVU estimator of μ is $\bar{x}$

- the p-th quantile of a $N(\mu, \sigma_0^2)$ dist. is $\mu + \sigma_0 z_p$ and $E_\mu(\bar{x} + \sigma_0 z_p) = \mu + \sigma_0 z_p$ so $\bar{x} + \sigma_0 z_p$ is UMVU

- $\psi = \Psi(\mu) = \Phi\left(\frac{x_0 - \mu}{\sigma_0}\right)$ = prob. a value $\le x_0$

  problem: $\Phi((x_0 - \bar{x})/\sigma_0)$ is not unbiased

  but $\frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, x_0]}(x_i)$ is unbiased, so

  $E\left(\frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, x_0]}(x_i) \,\middle|\, \bar{x}\right)$ is UMVU.

  $= E\left(I_{(-\infty, x_0]}(x_1) \,\middle|\, \bar{x}\right)$

  $x = \bar{x}\underline{1} + (x - \bar{x})$ $\quad \bar{x} \sim N(\mu, \sigma_0^2/n)$ ind. of $\underline{x} - \bar{x} \sim N_n(\underline{0}, \sigma_0^2(I - \frac{1}{n}\underline{1}\underline{1}'))$

  $e_1 \sim N(0, \sigma_0^2(1 - \frac{1}{n}))$

  $= E\left(I_{(-\infty, x_0]}(\bar{x} + e_1) \,\middle|\, \bar{x}\right)$

  $\bar{x} + e_1 \le x_0$ iff $e_1 \le x_0 - \bar{x}$

$$= \int_{-\infty}^{x_0 - \bar{x}} \frac{1}{\sqrt{2\pi}\sqrt{1+1/n}\,\sigma_0} \exp\left\{-\frac{1}{\sigma_0^2(1+1/n)} z^2\right\} dz$$

$$= \Phi\left(\frac{x_0 - \bar{x}}{\sigma_0\sqrt{1+1/n}}\right)$$

Eg location-scale normal

- $x = (x_1, \ldots, x_n)' \sim N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$, $\sigma^2 > 0$ are unknown
- $(\bar{x}, \|x - \bar{x}1\|^2)$ is a complete MSS (fact)
- so again $\bar{x}$ is UMVU for μ
- $s_x^2 = \frac{1}{n-1}\|x - \bar{x}1\|^2$ is UMVU for σ^2

(note: Ex show $\frac{1}{n+1}\|x - \bar{x}1\|^2$ has a uniformly smaller MSE)

- recall $z = \frac{(n-1)s_x^2}{\sigma^2} \sim$ chi-squared $(n-1)$ $=$ gamma$\left(\frac{n-1}{2}, \frac{1}{2}\right)$ rate $\Rightarrow$

$$E\left(\frac{\sqrt{n-1}\,s_x}{\sigma}\right) = \int_0^\infty z^{1/2} \frac{1}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-1}{2}}} z^{\frac{n-1}{2}-1} e^{-z/2}\, dz$$

$$= \frac{1}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-1}{2}}} \int_0^\infty z^{\frac{n}{2}-1} e^{-z/2}\, dz$$

$$= \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} 2^{1/2}$$

$\therefore \frac{\sqrt{n-1}}{2^{1/2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} s_x$ is UMVU for σ

- Cramer-Rao (information) inequality

- recall: r.v.'s X, Y with $E(X^2) < \infty$, $E(Y^2) < \infty$ then

$$CORR(X,Y) = \frac{COV(X,Y)}{SD(X)\,SD(Y)}$$

$$\text{where } SD(X) = \sqrt{Var(X)}$$

- correlation inequality: $-1 \leq CORR(X,Y) \leq 1$ with $CORR(X,Y) = -1$ iff $Y = a + bX$ (with $b < 0$ and $CORR(X,Y) = 1$ iff $Y = a + bX$ with $b > 0$)

- note when $Y = a + bX$ then

(i) $E(Y) = a + bE(X)$

(ii) $Var(Y) = Var(a + bX) = b^2 Var(X)$

(iii) $COV(X,Y) = E\big((X - E(X))(Y - E(Y))\big)$

$= E\big((X - E(X))\, b(X - E(X))\big) = b\,Var(X)$

$\therefore b = COV(X,Y) / Var(X)$

$a = E(Y) - bE(X) = E(Y) - \frac{COV(X,Y)}{Var(X)} E(X)$

$\therefore Y = E(Y) + \frac{COV(X,Y)}{Var(X)} (X - E(X))$

- <u>recall</u> - for model $\mathcal{M} = \{f_\theta : \theta \in \Theta\}$ and data x

  - likelihood fn $L(\theta|x) = c f_\theta(x)$

  log-likelihood fn $l(\theta|x) = \log c + \log f_\theta(x)$

  score fn. $S(\theta|x) = \frac{\partial l(\theta|x)}{\partial \theta} = \frac{\partial f_\theta(x)/\partial\theta}{f_\theta(x)}$

<u>Lemma</u> "Under conditions" we have

$$E_\theta(S(\theta|x)) = 0$$

Proof: We have that $1 = \int f_\theta(x)\,dx$ so

$$0 = \frac{\partial}{\partial\theta}\int f_\theta(x)\,dx \overset{\text{conditions}}{=} \int \frac{\partial f_\theta(x)}{\partial\theta}\,dx$$

$$= \int \left(\frac{\partial f_\theta(x)/\partial\theta}{f_\theta(x)}\right) f_\theta(x)\,dx = E_\theta(S(\theta|x))$$

- $\frac{\partial S(\theta|x)}{\partial\theta} = \frac{\partial^2 l(\theta|x)}{\partial\theta^2}$

$$= \frac{1}{f_\theta(x)}\frac{\partial^2 f_\theta(x)}{\partial\theta^2} - \frac{1}{f_\theta^2(x)}\left(\frac{\partial f_\theta(x)}{\partial\theta}\right)^2 \quad S^2(\theta|x)$$

- Fisher information $I(\theta) = \text{Var}_\theta(S(\theta|x))$

Under conditions we have

Lemma $I(\theta) = E_\theta\left(-\frac{\partial S(\theta|x)}{\partial \theta}\right)$

Proof: $E_\theta\left(-\frac{\partial S(\theta|x)}{\partial \theta}\right)$

$= E_\theta\left(S^2(\theta|x)\right) - E_\theta\left(\frac{1}{f_\theta(x)}\frac{\partial^2 f_\theta(x)}{\partial \theta^2}\right)$

$Var_\theta(S(\theta|x))$

$\rightarrow = \int \frac{\partial^2 f_\theta(x)}{\partial \theta^2}\,dx \overset{\text{conditions}}{=} \frac{\partial^2}{\partial \theta^2}\int f_\theta(x)\,dx = 0.$

- **note** - under general conditions, for $x = (x_1, \ldots, x_n)$ iid we have

$\theta_{MLE}(x) \xrightarrow{a.s.} \theta$ as $n \to \infty$

$I^{1/2}(\theta_{MLE}(x))(\theta_{MLE}(x) - \theta) \xrightarrow{\mathcal{D}} N(0,1)$ as $n \to \infty$

$\therefore \quad \theta_{MLE}(x) \pm I^{-1/2}(\theta_{MLE}(x))\, z_{1-\alpha/2}$

is an approximate $(1-\alpha)$ CI for θ.

Information

Theorem (Cramer-Rao Inequality)

Under conditions

$$Var_\theta(s) \ge \left(\frac{\partial E_\theta(s)}{\partial \theta}\right)^2 I^{-1}(\theta)$$

with equality iff

$$s(x) = E_\theta(s) + \frac{\partial E_\theta(s)}{\partial \theta} I^{-1}(\theta)\, S(\theta \mid x)$$

Proof: We have $Corr_\theta^2(s(x), S(\theta \mid x)) \le 1$

or $Cov_\theta^2(s(x), S(\theta \mid x)) \le Var_\theta(s)\, I(\theta)$

and $Cov_\theta(s(x), S(\theta \mid x)) = E_\theta(s(x)\, S(\theta \mid x))$

$= \int s(x) \frac{\partial f_\theta(x)}{\partial \theta}\, dx = \frac{\partial}{\partial \theta} \int s(x) f_\theta(x)\, dx$

$= \frac{\partial}{\partial \theta} E_\theta(s)$. We have equality iff

$$s(x) = E_\theta(s) + \frac{\partial E_\theta(s)}{\partial \theta} I^{-1}(\theta)\, S(\theta \mid x).$$

- so we have a lower bound on the variance of the estimator
- so if $Var_\theta(s)$ attains the lower bound it is UMVU for $E_\theta(s)$

eg $x = (x_1, \dots, x_n) \overset{iid}{\sim} \text{Bernoulli}(\theta)$

- $S(x) = \bar{x}$, $E_\theta(\bar{x}) = \theta$
- so $\dfrac{\partial E_\theta(S)}{\partial \theta} = 1$
- $\ell(\theta|x) = \log \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}$

$$= n\bar{x}\log\theta + n(1-\bar{x})\log(1-\theta)$$

$$S(\theta|x) = \frac{n\bar{x}}{\theta} - \frac{n(1-\bar{x})}{1-\theta} = \left(\frac{n}{\theta} + \frac{n}{1-\theta}\right)\bar{x} - \frac{n}{1-\theta}$$

$$= \frac{n}{\theta(1-\theta)}\bar{x} - \frac{n}{1-\theta}$$

- $\text{Var}_\theta(S(\theta|x)) = \dfrac{n^2}{\theta^2(1-\theta)^2}\text{Var}_\theta(\bar{x})$

$$= \frac{n^2}{\theta^2(1-\theta)^2}\frac{\theta(1-\theta)}{n} = \frac{n}{\theta(1-\theta)} = I(\theta)$$

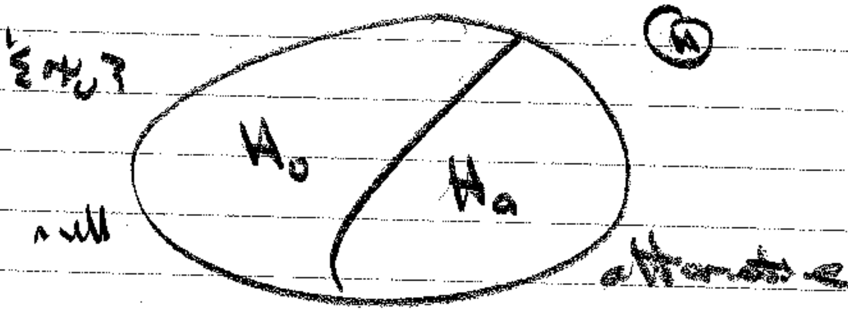
- so lower bound is $\dfrac{\theta(1-\theta)}{n}$

$\therefore$ since $\text{Var}_\theta(\bar{x})$ attains the lower bound it is UMVU for θ.

(c) Optimal Hypothesis Testing

- suppose we have $\Theta = H_0 \cup H_a$ where $H_0 \cap H_a = \emptyset$

eg $H_0 = \psi^{-1}\{\psi_0\}$



- want a decision (test) function $\delta : \mathcal{X} \to \{0,1\}$ where $\delta(x) = 1$ means we reject H_0 and $\delta(x) = 0$ means we accept H_0

- equivalently state a critical region

$$C_\delta = \{x : \delta(x) = 1\}$$ = set of data where H_0 is reject

- so $\delta(x) = I_{C_\delta}(x)$

- how do we choose δ?

- specify a loss function.

- $$L(\theta, d) = \begin{cases} 0 & \theta \in H_0, d = 0 \text{ or } \theta \in H_a, d = 1 \\ l_1 & \theta \in H_0, d = 1 \quad \text{type I error} \\ l_2 & \theta \in H_a, d = 0 \quad \text{type II error} \end{cases}$$

when $l_1 = l_2 = 1$ we call L 0-1 loss.

- risk function $R_\delta(\theta) \overset{\text{0-1 loss}}{=} E_\theta(L(\theta, \delta(x)))$

$$= \begin{cases} P_\theta(H_0 \text{ rejected}) & \theta \in H_0 \\ P_\theta(H_0 \text{ accepted}) & \theta \in H_a \end{cases}$$

- as we know there is no optimal δ
- so restrict the possible δ
- How?
- <u>size α δ</u> : put an upper bound $\alpha \in [0,1]$ on the prob. of making a type I error and look for optimal δ in this class
- so δ is size α if $P_\theta(H_0 \text{ rejected}) \le \alpha$ $\forall \theta \in H_0$
- we want to minimize $P_\theta(H_0 \text{ accepted})$ for each $\theta \in H_a$ or equivalently maximize the power function

$$\beta_\delta(\theta) = P_\theta(H_0 \text{ rejected})$$

for each $\theta \in H_a$.

- optimal test is called <u>uniformly most powerful (UMP) size α test</u>

- we consider the simplest problem where $\Theta = \{\theta_1, \theta_2\}$ with $H_0 = \{\theta_1\}$ and $H_a = \{\theta_2\}$

- Fact - in some problems there is still no optimal size α test

  - so enlarge the class of δ to allow for randomized decision (test) functions, namely, $\delta : \mathcal{X} \to [0,1]$

  - so $\delta(x)$ = prob. of rejecting H_0 when x is observed.

- so $R_\delta(\theta) = \begin{cases} E_\theta(\delta) & \theta \in H_0 \\ 1 - E_\theta(\delta) & \theta \in H_a \end{cases}$

$$= I_{H_0}(\theta) E_\theta(\delta) + I_{H_a}(\theta)(1 - E_\theta(\delta))$$

<u>Theorem</u> (Neyman Pearson or Fundamental Lemma)

When $\Theta = \{\theta_1, \theta_2\}$ the test of $H_0 = \{\theta_1\}$ vs $H_a = \{\theta_2\}$ of the form

$$\delta_0(x) = \begin{cases} 1 & \text{if } f_{\theta_2}(x)/f_{\theta_1}(x) > k_0 \\ \gamma & = k_0 \\ 0 & < k_0 \end{cases}$$

where $k_0 \in [0,\infty]$ and $\gamma \in [0,1]$ are determined by $E_{\theta_1}(\delta) = \alpha$ is UMP size α.

note ① - $f_{\theta_2}(x)/f_{\theta_1}(x)$ is the likelihood ratio of θ_2 to θ_1

② - define $\delta_\alpha(x) \equiv \alpha$ so

$$E_\theta(\delta_\alpha) = \alpha \quad \forall \theta$$

- this implies that the UMP size α test δ satis

$$\beta_\delta(\theta_2) \geq \alpha$$

- we say δ is unbiased (prob. of rejecting H_0 when it is false is no smaller than the prob of rejecting H_0 when it is true)

Proof: Consider, for some $k \in [0,\infty]$, $\gamma \in [0,1]$

$$\delta_0(x) = \begin{cases} 1 & f_{\theta_2}(x)/f_{\theta_1}(x) > k \\ \gamma & \quad = k \\ 0 & \quad < k. \end{cases}$$

We show we can choose k and γ to make size α.

First note that $\{x : f_{\theta_1}(x) = f_{\theta_2}(x) = 0\}$ has P_{θ_1} and P_{θ_2} measure 0. So for each $x \in \mathcal{X}$ we assume $f_{\theta_1}(x) > 0$ or $f_{\theta_2}(x) > 0$ and so $f_{\theta_2}(x)/f_{\theta_1}(x)$ is always defined (could be ∞).

Case 1 Suppose $\alpha = 1$. Then set $k = 0$ and $\gamma = 1$ so $\delta(x) = 1$ and note $E_{\theta_1}(\delta_0) = 1$ so δ is of size α. Also $E_{\theta_2}(\delta_0) = 1$ so δ_0 is UMP size α (no test can have power greater than 1).

Case 2 Suppose $\alpha = 0$. Then put $k = \infty$ and $\gamma = 0$. Then $\delta_0(x) = 0$ iff $f_{\theta_1}(x) > 0$

which implies $E_{\theta_1}(\delta_0) = 0$ so δ_0 is of size α. Suppose δ' is also of size $\alpha = 0$. Then $\delta'(x) = 0$ whenever $f_{\theta_1}(x) > 0$ which implies $\delta'(x) \le \delta_0(x)$ which implies $E_{\theta_2}(\delta') \le E_{\theta_2}(\delta_0)$ and δ_0 is UMP size α.

<u>Case 3</u> Suppose $0 < \alpha < 1$. Consider

$1 - \alpha^*(k) = P_{\theta_1}(f_{\theta_2}(x)/f_{\theta_1}(x) \le k)$

which is the cdf of the likelihood ratio. So $1 - \alpha^*(k)$ is nondecreasing in k with $1 - \alpha^*(-\infty) = 0$ and $1 - \alpha^*(\infty) = 1$

Let k_0 be the smallest value such that $1 - \alpha \le 1 - \alpha^*(k_0)$ (recall a cdf is right continuous so k_0 exists). Now

$P_{\theta_1}(f_{\theta_2}(x)/f_{\theta_1}(x) = k_0) = (1 - \alpha^*(k_0)) - (1 - \alpha^*(k_0 - 0))$

$= \alpha^*(k_0 - 0) - \alpha^*(k_0)$ where $1 - \alpha^*(k_0 - 0) = \lim_{\varepsilon \downarrow 0} 1 - \alpha^*(k_0 - \varepsilon)$

Now put

$$\gamma = \begin{cases} \dfrac{\alpha - \alpha^*(k_0)}{\alpha^*(k_0 - 0) - \alpha^*(k_0)} & \alpha^*(k_0 - 0) \neq \alpha^*(k_0) \\ 0 & \text{otherwise} \end{cases}$$

and note $1 - \alpha^*(k_0 - \varepsilon) \leq 1 - \alpha \quad \forall\, \varepsilon > 0$

and so $1 - \alpha^*(k_0 - 0) \leq 1 - \alpha$ which implies

$\alpha^*(k_0 - 0) \geq \alpha$ which implies $\gamma \in [0, 1]$.

Now $E_{\theta_1}(\delta_0) = \gamma P_{\theta_1}(f_{\theta_2}(x)/f_{\theta_1}(x) = k_0)$

$+ P_{\theta_1}(f_{\theta_2}(x)/f_{\theta_1}(x) > k_0)$

$= \alpha - \alpha^*(k_0) + 1 - (1 - \alpha^*(k_0)) = \alpha$ so

δ has exact size α.

Now let δ' be any other size α test function and $E_{\theta_2}(\delta') \geq E_{\theta_2}(\delta_0)$.

We write $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \cup \mathcal{X}_2$ so

$\mathcal{X}_i \cap \mathcal{X}_j = \phi$ when $i \neq j$ where

$$\mathcal{X}_0 = \{x : s_0(x) - s'(x) = 0\}$$

$$\mathcal{X}_1 = \{x : s_0(x) - s'(x) < 0\}$$

$$\mathcal{X}_2 = \{x : s_0(x) - s'(x) > 0\}$$

Note that $\mathcal{X}_1 = \{x : s_0(x) - s'(x) < 0,\ f_{\theta_2}(x)/f_{\theta_1}(x) \leq k\}$

(since $s'(x) \in [0,1]$ and $s_0(x) = 1$ when LR $> k_0$)

and $\mathcal{X}_2 = \{x : s_0(x) - s'(x) > 0,\ f_{\theta_2}(x)/f_{\theta_1}(x) \geq k_0\}$

Therefore,

$$0 \geq E_{\theta_2}(s_0) - E_{\theta_2}(s') = E_{\theta_2}(s_0 - s')$$

$$= E_{\theta_2}\left(I_{\mathcal{X}_0}(x)(s_0(x) - s'(x))\right) + E_{\theta_2}\left(I_{\mathcal{X}_1}(x)(s_0(x) - s'(x))\right)$$

$$+ E_{\theta_2}\left(I_{\mathcal{X}_2}(x)(s_0(x) - s'(x))\right)$$

$$\geq k_0 E_{\theta_1}\left(I_{\mathcal{X}_0}(x)(s_0(x) - s'(x))\right)$$

$$+ k_0 E_{\theta_1}\left(I_{\mathcal{X}_1}(x)(s_0(x) - s'(x))\right)$$

$$+ k_0 E_{\theta_1}\left(I_{\mathcal{X}_2}(x)(s_0(x) - s'(x))\right)$$

$$= k_0\left(E_{\theta_1}(s_0) - E_{\theta_1}(s')\right) = k_0(\alpha - E_{\theta_1}(s'))$$

≥ 0. Therefore $E_{\theta_2}(s) = E_{\theta}(s')$

and this proves that δ is UMP size α.

Corollary If δ is UMP size α then $\delta(x) = \delta_0(x)$ whenever $f_{\theta_2}(x)/f_{\theta_1}(x) \neq k_0$. Furthermore δ has exact size α unless $\exists$ UMP size α test with power 1

Theorem If T is a sufficient statistic and δ and $\delta'(x) = E(\delta \mid T(x))$ then $E_\theta(\delta) = E_\theta(\delta')$ $\forall$ θ

Proof: $E_\theta(\delta') \overset{def}{=} E_\theta(E(\delta \mid T(x)))$
$\overset{TTE}{=} E_\theta(\delta)$.

- so in looking for UMP size α test we can restrict attention to tests that depend on the data only through a MSS.

- operationally - $1-\alpha^*(k_0) = P_{\theta_1}(f_{\theta_2}(x)/f_{\theta_1}(x) \leq k_0)$

- find smallest k_0 s.t. $1-\alpha \leq 1-\alpha^*(k_0)$ or $\alpha^*(k_0) \leq \alpha$
$\alpha^*(k_0) = P_{\theta_1}(f_{\theta_2}(x)/f_{\theta_1}(x) > k_0) \leq \alpha$ and put
$\gamma = \frac{\alpha - \alpha^*(k_0)}{\ldots}$

Eg $x = (x_1, \ldots, x_n) \overset{iid}{\sim} N(\mu, \sigma_0^2)$ σ_0^2 known

- $H_0 : \mu = \mu_0$ vs $H_a : \mu = \mu_1$

- $T(x) = \bar{x}$ is sufficient for model with $\Theta = \{\mu_1, \mu_2\}$

- $\bar{x} \sim N(\mu, \sigma_0^2/n)$

- $f_{\mu_2, T}(\bar{x}) / f_{\mu_1, T}(\bar{x})$

$$= \frac{(2\pi)^{-\frac{1}{2}} \left(\frac{\sigma_0^2}{n}\right)^{-\frac{1}{2}} \exp\left\{-\frac{n}{2\sigma_0^2}(\bar{x}-\mu_1)^2\right\}}{(2\pi)^{-\frac{1}{2}} \left(\frac{\sigma_0^2}{n}\right)^{-\frac{1}{2}} \exp\left\{-\frac{n}{2\sigma_0^2}(\bar{x}-\mu_0)^2\right\}}$$

$$= \exp\left\{-\frac{n}{2\sigma_0^2}\left((\bar{x}-\mu_1)^2 - (\bar{x}-\mu_0)^2\right)\right\}$$

$$= \exp\left\{-\frac{n}{2\sigma_0^2}\left(\cancel{\bar{x}^2} - 2\mu_1\bar{x} + \mu_1^2 - \cancel{\bar{x}^2} + 2\mu_0\bar{x} - \mu_0^2\right)\right\}$$

$$> k_0$$

iff $-\frac{n}{\sigma_0^2}(\mu_0 - \mu_1)\bar{x} > \log k_0 + \frac{n}{2\sigma_0^2}(\mu_0^2 - \mu_1^2)$

iff $\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \underset{\mu_1 > \mu_2, \, <}{\overset{\mu_0 < \mu_1}{>}} -\frac{\sqrt{n}\mu_0}{\sigma_0} - \frac{\sigma_0}{\sqrt{n}} \frac{\log k_0}{(\mu_0 - \mu_1)} + \frac{\sqrt{n}}{\sigma_0} \frac{(\mu_0 + \mu_1)}{2} = k_0'$

- $\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$

- note $P_{\mu_0}\left(\frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}} = k_0'\right) = 0$ so $\gamma = 0$
- also $k_0' = z_{1-\alpha}$ implies

$$\delta_1(x) = \begin{cases} 1 & \frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}} > z_{1-\alpha} \quad \mu_0 < \mu_1 \\ & \left(< z_\alpha \quad \mu_0 > \mu_1\right) \\ 0 & \text{otherwise} \end{cases}$$

is the UMP size α test for
$H_0: \mu = \mu_0$ vs $H_a: \mu = \mu_1$

- note - test does not involve μ_1 so δ_1
is UMP size α for
$H_0: \mu = \mu_0$ vs $H_a: \mu > \mu_0$
(or $H_a: \mu < \mu_0$)

- also - when $\mu < \mu_0$ $P_\mu\left(\frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}} > z_{1-\alpha}\right)$

$$= P\left(\frac{\bar{x}-\mu}{\sigma_0/\sqrt{n}} > z_{1-\alpha} + \frac{\mu_0-\mu}{\sigma_0/\sqrt{n}}\right) < \alpha$$

and so $\delta_1(x) = 1$ when $\frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}} > z_{1-\alpha}$

is size α for $H_0: \mu \le \mu_0$ vs
$H_a: \mu > \mu_0$

- now if δ' is size α for $H_0: \mu \le \mu_0$
vs $H_a: \mu > \mu_0$ then δ' is size α for
$H_0: \mu = \mu_0$ vs $H_a: \mu > \mu_0$ and so
power of δ' $\le$ power δ_1 for $\mu > \mu_0$

- this proves that δ_1 is UMP size α for $H_0: \mu \le \mu_0$ vs $H_a: \mu > \mu_0$
  - similarly $\delta_2(x) = 1$ when $(\bar{x}-\mu_1)/\sigma_0/\sqrt{n} < z_\alpha$ is UMP size α for $H_0: \mu \ge \mu_0$ vs $H_a: \mu < \mu_0$
- what about $H_0: \mu = \mu_0$ vs $H_a: \mu \ne \mu_0$ (two-sided problem)?
- suppose δ is UMP size α for the problem and only depends on $\bar{x}$
- now δ_1 and δ_2 are both size α for this problem which implies

$$E_\mu(\delta) = E_\mu(\delta_1) \quad \text{when } \mu > \mu_1$$
$$= E_\mu(\delta_2) \quad \text{when } \mu < \mu_1$$

- but the completeness of $\bar{x}$ tells us that $\delta(\bar{x}) = \delta_1(\bar{x})$ and $\delta(\bar{x}) = \delta_2(\bar{x})$ which is a contradiction.

$\therefore \nexists$ a UMP size α test for $H_0: \mu = \mu_0$ vs $H_a: \mu \ne \mu_0$

- fact - there does exist a UMPU (uniformly most powerful unbiased) size α test for the two-sided problem given by

$$\delta(x) = \begin{cases} 1 & \left|\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right| > z_{1-\alpha/2} \\ 0 & \text{otherwise} \end{cases}$$

Eg $x = (x_1, \ldots, x_n) \overset{iid}{\sim}$ Bernoulli(θ), $\theta \in [0,1]$

- $H_0 : \theta = \theta_0$ vs $H_a : \theta = \theta_1$
- $T(x) = n\bar{x}$ is sufficient for model with $\Theta = \{\theta_1, \theta_2\}$
- $n\bar{x} \sim$ Binomial (n, θ)
- $f_{\theta_2, T}(n\bar{x}) \,/\, f_{\theta_1, T}(n\bar{x})$

$$= \binom{n}{n\bar{x}} \theta_1^{n\bar{x}} (1-\theta_1)^{n(1-\bar{x})} \Big/ \binom{n}{n\bar{x}} \theta_0^{n\bar{x}} (1-\theta_0)^{n(1-\bar{x})}$$

$$= \left(\frac{\theta_1}{\theta_0}\right)^{n\bar{x}} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n(1-\bar{x})}$$

$$> k_0$$

iff $n\bar{x} \log\left(\frac{\theta_1}{\theta_0}\right) + n(1-\bar{x}) \log\left(\frac{1-\theta_1}{1-\theta_0}\right)$

$$= n\bar{x} \log\left(\frac{\theta_1}{1-\theta_1} \Big/ \frac{\theta_0}{1-\theta_0}\right) + n \log\left(\frac{1-\theta_1}{1-\theta_0}\right) > \log k_0$$

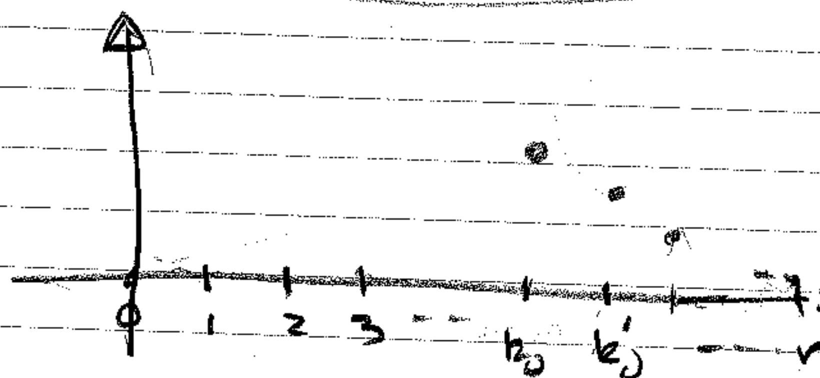
iff $n\bar{x} \underset{\theta_1 < \theta_0}{\overset{\theta_1 > \theta_0}{\gtrless}} \log\left(\frac{\theta_1}{1-\theta_1} \Big/ \frac{\theta_0}{1-\theta_0}\right) \left(\log k_0 - n \log\left(\frac{1-\theta_1}{1-\theta_0}\right)\right)$ < 0 when $\theta_1 > \theta_0$

since $\left(\frac{\theta}{1-\theta}\right)' = \frac{1}{1-\theta} + \frac{\theta}{(1-\theta)^2} > 0$

- so when $\theta_1 > \theta_0$ find k_0' st

$P_{\theta_0}(n\bar{x} > k_0') \leq \alpha$, $P_{\theta_0}(n\bar{x} > k_0' - 1) > \alpha$

and put $\gamma = \dfrac{\alpha - P_{\theta_0}(n\bar{x} \geq k_0)}{P_{\theta_0}(n\bar{x} \geq k_0 - 1) - P_{\theta_0}(n\bar{x} \geq k_0)} \neq 0$



- **note** - typically we will have to randomize

- **note** UMP size α for

$$H_0: \theta \leq \theta_1 \quad \text{vs} \quad H_a: \theta > \theta_1$$

can prove $\displaystyle\sum_{k=0}^{x} \binom{n}{x} \theta^k (1-\theta)^{n-k}$

$$= \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x)} \int_\theta^1 t^x (1-t)^{n-x-1}\, dt$$

so decreasing in θ

- again no UMP size α for $H_0: \theta = \theta_0$ vs $H_a: \theta \neq \theta_0$ but there is a UMPU sat

$$\varphi(x) = \begin{cases} 1 & \bar{x} \notin (k_1, k_2) \\ \gamma_1 & \bar{x} = k_1 \\ \gamma_2 & \bar{x} = k_2 \\ 0 & \text{otherwise} \end{cases}$$

where k_1, k_2 are chosen to satisfy

$E_{\theta_0}(\varphi) = \alpha$

$E_{\theta_0}(\bar{x}\varphi(\bar{x})) = \alpha\,\theta_0$

(c) Optimal Confidence Regions

- suppose we have a size α test S_{ψ_0} for each $\psi_0 \in \Psi$ each with no randomization.

- put $C(x) = \{\psi_0 : S_{\psi_0}(x) = 0\}$

  $=$ set of ψ_0 accepted after observing x.

- coverage probability
$$P_\theta(\Psi(\theta) \in C(x)) = P_\theta(S_{\Psi(\theta)}(x) = 0)$$
$$= 1 - P_\theta(S_{\Psi(\theta)}(x) = 1) \geq 1-\alpha \quad \forall \theta.$$

  $\therefore$ $C(x)$ is a $(1-\alpha)$ confidence region for $\psi = \Psi(\theta)$

- alternatively if C is a $(1-\alpha)$ CR for ψ then put
$$S_{\psi_0}(x) = \begin{cases} 1 & \psi_0 \notin C(x) \\ 0 & \psi_0 \in C(x) \end{cases}$$
so when $\Psi(\theta) = \psi_0$ we have
$$P_\theta(S_{\psi_0}(x) = 1) = 1 - P_\theta(S_{\psi_0}(x) = 0)$$
$$= 1 - P_\theta(\psi_0 \in C(x)) \leq 1 - (1-\alpha) = \alpha.$$
and so S_{ψ_0} is size α.

- **note** - when $\delta_{\tau_0}(x) \in [0,1]$ then C is randomized in that given x $\tau_0 \in C(x)$ with prob $1-\delta_{\tau_0}(x)$

- then $P_\theta(\Psi(\theta) \in C(x))$
TTP
$= E_\theta(1 - \delta_{\Psi(\theta)}(x))$ prob in given x
$= 1 - E_\theta(\delta_{\Psi(\theta)}(x)) \geq 1-\alpha.$

- what does power correspond to with confidence regions?

- suppose θ' is s.t. $\Psi(\theta') \neq \Psi(\theta)$ then

$P_\theta(\Psi(\theta') \in C(x))$ = prob C covers false value $\Psi(\theta')$ when θ true

$= E_\theta(1 - \delta_{\Psi(\theta')}(x)) = 1 - E_\theta(\delta_{\Psi(\theta')}(x))$

$\therefore$ if δ_{H_0} is UMP size α for $\beta_{\delta_{\Psi(\theta')}}(\theta)$ each τ_0 we have that C is a uniformly most accurate $1-\alpha$ CR for $\tau = \Psi(\theta)$ as it uniformly minimizes the prob of covering false values.

- suppose each S_{H_0} is unbiased then

$$P_\theta\left(\Psi(\theta') \in C(x)\right)$$

$$= 1 - E_\theta\left(S_{\Psi(\theta')}\right) \begin{cases} \geq 1-\alpha & \Psi(\theta') = \Psi(\theta) \\ \leq 1-\alpha & \Psi(\theta') \neq \Psi(\theta) \end{cases}$$

$\therefore$ prob. of covering a false value is less than or equal to the probability of covering true value

- also if the S_{H_0} are UMPU size α then $C(x)$ is a UMAU $1-\alpha$ CR for $\Psi(\theta)$

- sometimes people look for confidence regions with smallest expected value for each Θ namely $C(x)$ minimize

$$E_\theta(\mathrm{vol}\, C(x))$$

among all γ-CR's for $\psi = \Psi(\theta)$

- note $E_\theta(\mathrm{vol}\, C(x))$

$$= \int_{\mathcal{X}} \mathrm{vol}\, C(x)\, f_\theta(x)\, dx$$

$$= \int_{\mathcal{X}} \left(\int_{C(x)} d\psi \right) f_\theta(dx)$$

$$= \int_{\mathcal{X}} \int_{\Psi} I_{C(x)}(\psi)\, d\psi\, f_\theta(x)\, dx$$

positive integrals

$$= \int_{\Psi} \int_{\mathcal{X}} I_{C(x)}(\psi)\, f_\theta(x)\, d\psi$$

$$= \int_{\Psi} P_\theta(\psi \in C(x))\, d\psi$$

$$= \int_{\Psi \setminus \{\Psi(\theta)\}} P_\theta(\psi \in C(x))\, d\psi$$

$\therefore$ a UMA γ-CR also has smallest expected volume (not conversely since you can find γ-CR's with smallest expected volumes that are not UMA)

* note invariant CR's.