ii. Case with variances not known to be common:

- Let $\sigma^2 = \mathrm{Var}[X_i]$, $\tau^2 = \mathrm{Var}[Y_i]$
- Pivotal quantity $S = (\bar{Y} - \bar{X} - \theta)/\sqrt{\sigma^2/m + \tau^2/n}$ if σ, τ known.
- If σ, τ unknown, estimate by $S_x = \sqrt{\Sigma_{i=1}^m (X_i - \bar{X})/(m-1)}$, $S_y = \sqrt{\Sigma_{i=1}^n (Y_i - \bar{Y})/(n-1)}$ respectively.
- Is $S = (\bar{Y} - \bar{X} - \theta)/\sqrt{S_x^2/m + S_y^2/n}$ pivotal? No.
  - ▷ If $\sigma = 0$, reduces to t_{n-1}.
  - ▷ If $\sigma = \tau$, $m = n$, t_{m+n-2}.
- Standard solution: approximate by t_d, where d is complicated formula of S_x, S_y, m, n.
  - ▷ See Fig. 4.

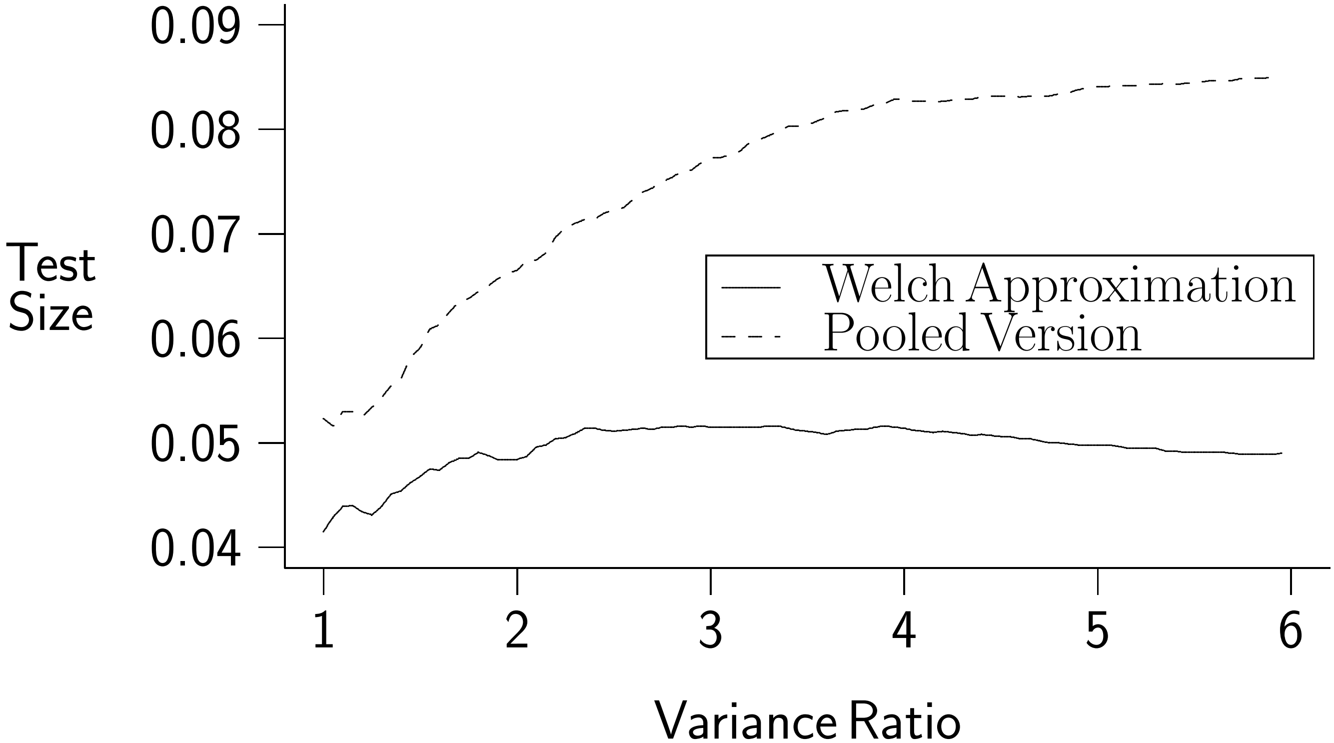
WMS: 9.2

J. Relative Efficiency

1. Definition: The ratio $\mathrm{Var}\left[\hat{\theta}_1\right]/\mathrm{Var}\left[\hat{\theta}_2\right]$ is the *relative efficiency* of $\hat{\theta}_2$ re $\hat{\theta}_1$.
2. Examples:

a. Binomial Distribution.

Fig. 4: Dependence of the Two Sample Test on Variance Ratio



i. Rival Unbiased Estimators of π :

- Suppose $X \sim \mathcal{B}\text{in}(n, \pi)$ and $Y \sim \mathcal{B}\text{in}(m, \pi)$.
- Let $\delta_1(X, Y) = X/n$ and $\delta_2(X, Y) = (X + Y)/(m + n)$.
- By not using some information, δ_1 throws away information. How is this mathematically quantified?

ii. Calculating Relative Efficiency: Note that $\text{Var}\left[\delta_2(X, Y)\right] = \pi(1 - \pi)/(m + n)$ and $\text{Var}\left[\delta_1(X)\right] = \pi(1 - \pi)/n$. Note that $\text{Var}\left[\delta_1(X)\right] > \text{Var}\left[\delta_2(X, Y)\right]$.

b. Estimating a general mean:

i. Consider two ind. measurements X_1 and X_2, with a common mean μ and variance σ^2.

ii. Then $a_1X_1 + a_2X_2$ is unbiased if and only if $a_1 + a_2 = 1$.

iii. The variance is $(a_1^2 + a_2^2)\sigma^2$, which is minimized when $a_1 = a_2 = \frac{1}{2}$.

iv. Relative efficiency of the variance minimizing estimator to the general estimator is $2(a_1^2 + a_2^2)$.

c. Poisson variable.

i. Mean and variance of a $\mathcal{P}(\mu)$ random variable are both μ;

ii. hence an alternate estimator for μ might be the sample variance $\delta(\boldsymbol{X}) = (n-1)^{-1}(\Sigma_{i=1}^{n} X_i^2 - n\bar{X}^2)$.

iii. To see that this is unbiased, refer to discussion about generic variance

iv. Kenney and Keeping (1954) p. 164 show that $\mathrm{Var}\,[\delta(\boldsymbol{X})] \approx \mu(1+2\mu)/n$.

v. sample mean is unbiased and has variance μ/n.

vi. relative efficiency of the sample variance to the sample mean is approximately

$$\frac{\mu(1+2\mu)/n}{\mu/n} = 1 + 2\mu.$$

vii. Here relative efficiency depends on $\boldsymbol{\theta}$.

- This is a relatively simple case, in which one estimator is always better than the other;
- it need not be the case.

WMS: 9.3

K. Consistency.

1. As we saw with our efficiency calculations, $\mathrm{Var}\left[\hat{\theta}\right]$ usually decreases as n increases.

   a. Think of $\hat{\theta}$ as the family of estimators based on various sample sizes,

2. Consistency Definition: An estimator $\hat{\theta}$ is called $consistent$ if

   a. given

      i. any high probability of seeing $\hat{\theta}$ within a certain band, and

      ii. any very small width for this band,

   b. a large enough n ensures that the probability that $\hat{\theta}$ is within the required distance of the true value is as required.

3. $\forall C > 0$ and $\delta > 0 \, \exists M$ possibly depending on δ and C such that $\mathrm{P}\left[\left|\hat{\theta} - \theta\right| \leq C\right] > 1 - \delta$ for any $n > M$.

4. Example:

a. if $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$,

i. Estimate θ by $\hat{\theta}_n = \bar{X} \sim N(\mu, \sigma^2/n)$.

ii. Then $(\hat{\theta}_n - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$.

iii. Hence $\mathrm{P}\left[\left|\hat{\theta}_n - \mu\right| \le C\right]$

$$= \mathrm{P}\left[\frac{\left|\hat{\theta}_n - \mu\right|}{(\sigma/\sqrt{n})} \le \frac{\sqrt{Cn}}{\sigma}\right] = \Phi\left(\frac{\sqrt{n}C}{\sigma}\right) - \Phi\left(\frac{-\sqrt{n}C}{\sigma}\right),$$

- where Φ is the c.d.f. of a $N(0,1)$ variable.

iv. $\lim_{n\to\infty} \mathrm{P}\left[\left|\hat{\theta}_n - \mu\right| \le C\right] = 1$

- Let $z_{\delta/2}$ satisfy $\Phi(z_{\delta/2}) = 1 - \delta/2$
- For all n such that $\sqrt{n}C/\sigma > z_{\delta/2}$ we have $\mathrm{P}\left[\left|\hat{\theta}_n - \theta\right| \le C\right] > 1 - \delta$.
- Hence $n > z_{\delta/2}^2\sigma^2/C^2 \Rightarrow \mathrm{P}\left[\left|\hat{\theta}_n - \theta\right| \le C\right] > 1 - \delta$.

v. or $n > \ln(\delta)/\ln(1 - C/\delta)$.

5. An inconsistent Estimator: Suppose $f_X(x;\mu) = \pi^{-1}(1 + (x-\mu)^2)^{-1}$.

vi. density of $Z = \frac{1}{2}(X+Y)$ and $W = X$ is

$$f_{W,Z}(w,z;\mu) = \pi^{-2}(1+(w-\mu)^2)^{-1}(1+(2z-w-\mu)^2)^{-1}2$$

vii. Integrate re w:

$$f_Z(z;\mu) = \int_{-\infty}^{+\infty} \frac{2\,dw}{\pi^2(1+(w-\mu)^2)(1+(2z-w-\mu)^2)}.$$

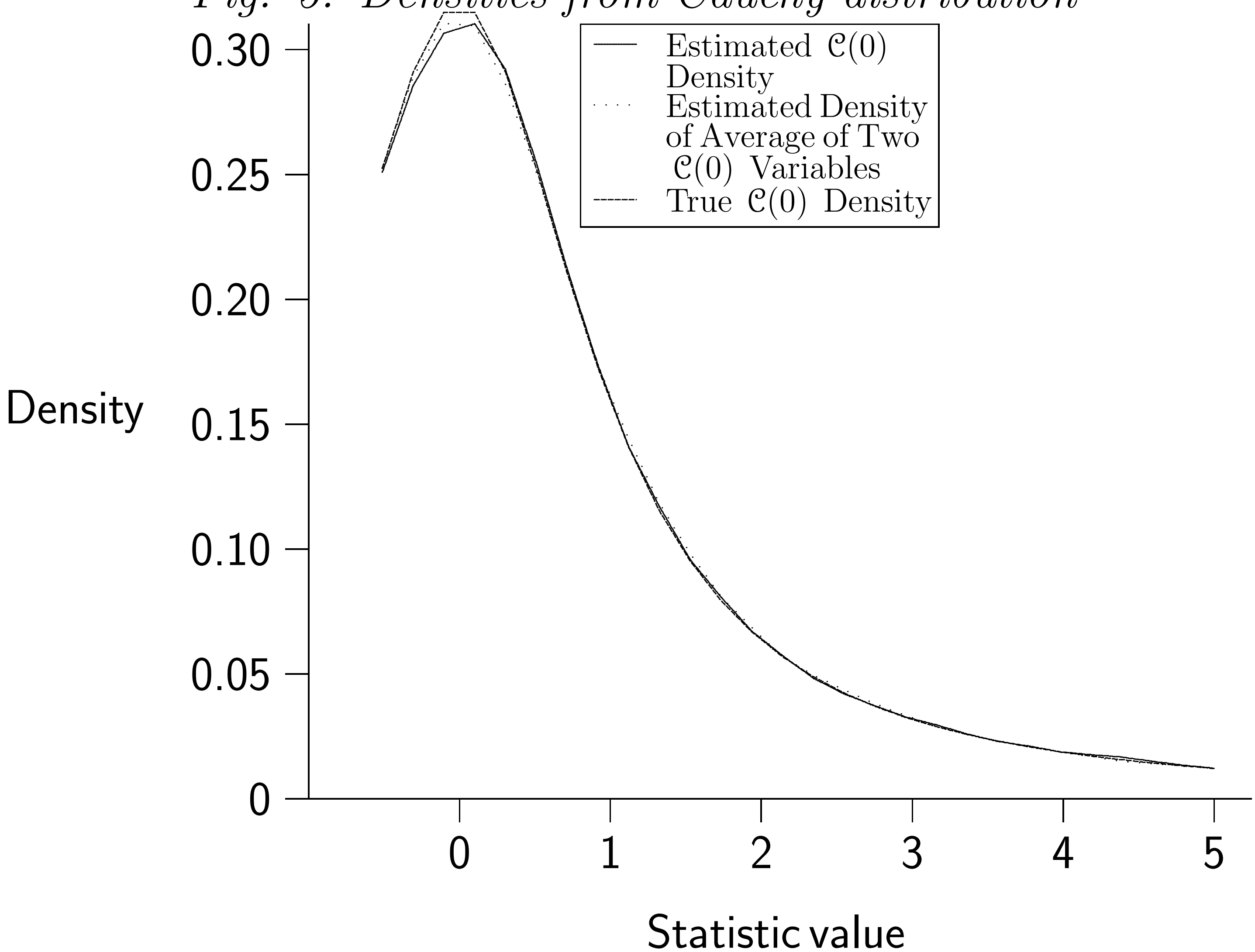
viii. Substitute $w-\mu=v+z$ and using partial fractions:

$$\frac{1}{4z\left(1+z^2\right)}\left[\frac{2z-v}{\left(1+v^2-2vz+z^2\right)}+\frac{v+2z}{\left(1+v^2+2vz+z^2\right)}\right]$$

ix. Hence Z has same distń as X and Y.

x. See Fig. 5.

Fig. 5: Densities from Cauchy distribution



Calculated from 200000 random draws

xi. Hence mean of 2^k variables has the same distń as X

xii. Hence mean is inconsistent.

6. A general rule

a. Often hard: Usually the bounds on n are not so easily derived explicitly.

b. Use *Chebyshev's inequality* :

i. Relate the probability that a random variable T is farther than a distance C from its mean θ to its variance.

$$\begin{aligned}\mathrm{Var}\left[T\right] &= \sum_t (t-\theta)^2 p_T(t;\theta) \\ &= \sum_{\{t\,||t-\theta|<C\}} (t-\theta)^2 p_T(t;\theta) + \sum_{\{t\,||t-\theta|\geq C\}} (t-\theta)^2 p_T(t;\theta) \\ &\geq 0 + \sum_{\{t\,||t-\theta|\geq C\}} (C)^2 p_T(t;\theta) \\ &= C^2 \sum_{\{t\,||t-\theta|\geq C\}} p_T(t;\theta) = C^2 \mathrm{P}\left[|T-\theta| \geq C\}\right].\end{aligned}$$

ii. So $\mathrm{P}\left[\left|\hat{\theta}-\theta\right| \geq C\right] \leq \mathrm{Var}\left[\hat{\theta}\right]/C^2$.

c. Hence if $\mathrm{E}\left[\hat{\theta}\right] = \theta$ and $\lim_{n\to\infty} \mathrm{Var}\left[\hat{\theta}\right] = 0$, then $\hat{\theta}$ is consistent.

d. Examples

i. If $X \sim \mathcal{B}\text{in}(n,\theta)$, and $\hat{\theta}_n = X/n$, then $\mathrm{Var}\left[\hat{\theta}\right] = \theta(1-\theta)/n$. Then

$$\mathrm{P}\left[\left|\hat{\theta}_n-\theta\right| \geq C\right] \leq \theta(1-\theta)/(C^2 n) \leq 1/(4C^2 n).$$

ii. If $X_1, \cdots, X_n \sim \mathcal{P}(\mu)$, $\hat{\mu} = \bar{X}$

iii. $\mathrm{Var}[\hat{\mu}] = \mu/n$

iv. Applying Chebyshev's inequality, $\mathrm{P}[|\hat{\mu} - \mu| \geq C] \leq \mu/(C^2 n)$ proves consistency,

v. the values of n making the RHS smaller than some limit δ depend on μ.

7. Theorem: If $\hat{\theta}$ consistent for θ, and $g(\theta)$ continuous, then $g(\hat{\theta})$ consistent for $g(\theta)$.

WMS: Question 8.8

L. Variance Bounds: How well can we possibly do?

1. Definition: Define the $expected\, information$ or $Fisher\, information$ $i(\theta) = n\mathrm{E}\left[\partial^2 \ln(f_X(X;\theta))/\partial\theta^2\right]$.
   a. 1st derivative tells how fast density changes with θ.
   b. 2nd derivative tells how fast density curves with θ.
2. Idea:
   a. information about θ depends on how quickly on average $f_X(X;\theta)$ as a function of θ drops away from its peak
   b. This is measured by the inverse of the curvature.
   c. For this course always interpret log as natural logs.

3. Conditions: i.i.d. observations from density smooth in parameter
   a. $f_X(X, \theta)$ is positive on a set func. ind. of θ ,
   b. has two derivatives with respect to θ
   c. and $X_1, \ldots, X_n$ are i.i.d. with p.d.f./p.m.f. $f_X(x, \theta)$
4. Result: A lower bound (the *Cramér-Rao lower bound*) on the variance of an unbiased estimator $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$ of θ is $\mathrm{Var}\left[\hat{\theta}\right] \geq 1/[ni(\theta)]$.
5. Proof: Differentiate identities requiring density to integrate to one and requiring unbiasedness.
   a. Note identity $1 = \int f_X(x;\theta)\,dx$,
      i. differentiate :
      $$0 = \int \frac{\partial(f_X(x;\theta)}{\partial\theta}\,dx = \int \frac{\partial \ln(f_X(x;\theta)}{\partial\theta} f_X(x;\theta)\,dx$$
      ii. differentiate again:
      $$0 = \int \frac{\partial^2(\ln(f_X(x;\theta))}{\partial\theta^2)} f_X(x;\theta)\,dx +$$
      $$\int \frac{\partial(\ln(f_X(x;\theta))}{\partial\theta} \frac{\partial f_X(x;\theta)}{\partial\theta}\,dx$$
      $$0 = \int \frac{\partial^2(\ln(f_X(x;\theta))}{\partial\theta^2} f_X(x;\theta)\,dx +$$
      $$\int \partial(\ln(f_X(x;\theta))/\partial\theta)^2 f_X(x;\theta)\,dx$$
   b. Note identity: $\theta = \int \cdot \int \hat{\theta}(\boldsymbol{x}) f_X(x_1;\theta) \cdots f_X(x_n;\theta)\,d\boldsymbol{x}$.

i. Also differentiate:

$$1 = \int\cdots\int \hat{\theta}(\boldsymbol{x}) \frac{d}{d\theta}[f_X(x_1;\theta)\cdots f_X(x_n;\theta)]\, d\boldsymbol{x}$$

$$= \int\cdots\int \hat{\theta}(\boldsymbol{x}) \sum_{j=1}^{n} \frac{d}{d\theta} f_X(x_j;\theta) \prod_{k\neq j} f_X(x_k;\theta)\, d\boldsymbol{x}$$

$$= \int\cdots\int \hat{\theta}(\boldsymbol{x}) \sum_{j=1}^{n} (d/d\theta)\ln(f_X(x_j;\theta)) \prod_{k} f_X(x_k;\theta)\, d\boldsymbol{x}$$

$$= \mathrm{E}\left[\hat{\theta} \sum_{j=1}^{n} (d/d\theta)\ln(f_X(x_j;\theta))\right].$$

c. Call $U = \Sigma_{j=1}^{n}(d/d\theta)\ln(f_Y(Y_j;\theta))$ the *score statistic* .

i. U is the sum of i.i.d. summands;

ii. hence $\mathrm{Var}[U] = n\mathrm{Var}\left[(d/d\theta)\ln(f_Y(Y_j;\theta))\right]$

iii. Since $\mathrm{E}\left[(d/d\theta)\ln(f_Y(Y_j;\theta))\right] = 0$, then $\mathrm{Var}\left[(d/d\theta)\ln(f_Y(Y_j;\theta))\right] = i(\theta)$.

iv. Hence $\mathrm{E}\left[\hat{\theta}U\right] = 1$.

v. By Cauchy–Schwartz, $\mathrm{E}\left[(\hat{\theta}-\theta)^2\right]\mathrm{E}\left[U^2\right] \geq 1$, and $\mathrm{Var}\left[\hat{\theta}\right] \geq 1/[ni(\theta)]$.

d. Cauchy–Schwartz inequality: For any random variables X and Y , $\mathrm{Cov}[X,Y] \leq \sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}$

- Let $U = (X - \mathrm{E}[X])/\sqrt{\mathrm{Var}[X]}$, $V = (Y - \mathrm{E}[Y])/\sqrt{\mathrm{Var}[Y]}$.
- $0 \leq \mathrm{E}\left[(U-V)^2\right] = \mathrm{E}\left[U^2\right] + \mathrm{E}\left[V^2\right] - 2\mathrm{Cov}[U,V] =$

$$1 + 1 - 2\text{Cov}[X, Y] / \sqrt{\text{Var}[X]\,\text{Var}[Y]}$$

Q.E.D

WMS: 9.6-9.7

M. Techniques for generating estimates

1. Method of Moments

a. Definition:

i. Suppose $X_1, \ldots, X_n \sim f_X(x; \theta)$

ii. Law of large numbers tells us that $\Sigma_{j=1}^n X_j/n \approx \text{E}_\theta[X]$

iii. Method of moments says solve $\Sigma_{j=1}^n X_j/n = \text{E}_{\hat\theta}[X]$ for $\hat\theta$.

iv. Expectations above are functions of $\boldsymbol{\theta}$.

v. If there are multiple parameters, might solve $\Sigma_{j=1}^n X_j^2/n = \text{E}_{\hat{\boldsymbol{\theta}}}\left[X^2\right]$, and higher powers

b. Examples:

i. $X_1, \ldots, X_k \sim \mathcal{NB}\text{in}(\theta, m)$

- Number of trials it takes to get m successes, if each has success probability θ
- $\text{E}\left[X_j\right] = m/\theta$ (Theorem 5.6).
- Estimate θ
- $\bar{X} = m/\hat\theta$

- $\hat{\theta} = m/\bar{X}$

ii. The normal distń. $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$.

- Then $\sum_{j=1}^{n} X_j^1/n = \hat{\mu}^1,\ \sum_{j=1}^{n} X_j^2/n = \hat{\mu}^2 + \hat{\sigma}^2$
- Hence $\hat{\mu} = \bar{X}$ and $\Sigma_{j=1}^{n} X_j^2/n = \bar{X}^2 + \hat{\sigma}^2$, or $\hat{\sigma} = \sqrt{\Sigma_{j=1}^{n} X_j^2/n - \bar{X}^2} = \sqrt{\Sigma_{j=1}^{n}(X_j - \bar{X})^2/n}$.
- Recall that this estimate of σ^2 is biased.
- Contrary to what may seem obvious from their definition, these estimators need not be unbiased.

iii. $X_1, \ldots, X_n \sim \mathcal{B}\text{in}(m, \pi)$

- $\mathrm{E}\left[X_j\right] = m\pi$
- $\hat{\pi} = \bar{X}/m = (\Sigma_j X_j)/(nm)$.

iv. Same setup as before

- This time estimate $\psi = \pi/(1-\pi)$
  - ▷ called odds
  - ▷ $\pi = \psi/(1+\psi)$
- $\bar{X} = m\hat{\psi}/(1+\hat{\psi})$
- $\hat{\psi} = \bar{X}/m/(1-\bar{X}/m) = \hat{\pi}/(1-\hat{\pi})$

c. Last example demonstrates equivariance: if you change scale of parameter, you change estimate in exactly the same way.

d. Problems with m.o.m.e.s:

i. No guarantee of near-efficiency.

- Since $\mathrm{Var}\left[\Sigma_{j=1}^{n} X_j^k/n\right]$ generally large k large, them in the estimating equations may add a lot of variability to $\hat{\theta}$.

ii. They may not even exist: cf. Cauchy distń.

e. Main advantage:

i. Intuitive.

ii. Generally speaking consistent.

2. Extensions

a. Can equate other sample quantities with population quantities

i. Ex., median

ii. Works better for some distributions like Cauchy

3. Likelihood methods.

a. Definition: The joint p.d.f. for all of the observations is known as the *likelihood function* $L(\boldsymbol{\theta})$.

i. with the observed data substituted in and

ii. viewed as a function of $\boldsymbol{\theta}$,

iii. $L(\boldsymbol{\theta})$ arose earlier when talking about the Cramér-Rao bound.

iv. Heuristically $L(\boldsymbol{\theta})$ measures the relative likelihood of various potential values for $\boldsymbol{\theta}$.

b. Parameter Estimation:

i. Take that value that is most likely in the sense described here; that is, maximize the likelihood function, or equivalently, maximize the log likelihood.

ii. Value of $\boldsymbol{\theta}$ where $L(\boldsymbol{\theta})$ is maximized is called the $m.l.e.$ (m.l.e.) and is usually written $\hat{\theta}$.

c. For ind. observations,

$$f_{X_1,\cdots,X_n}(x_1,\cdots,x_n;\boldsymbol{\theta}) = \prod_1^n f_{X_j}(x_j;\boldsymbol{\theta})$$

i. Hence

$$L(\boldsymbol{\theta};X_1,\cdots,X_n) = \prod_1^n L(\boldsymbol{\theta};X_j)$$

ii. and

$$l(\boldsymbol{\theta};X_1,\cdots,X_n) = \sum_{j=1}^n l(\boldsymbol{\theta};X_j)$$

iii. so the log likelihood for a collection of ind. random variables is the sum of the ind. log likelihoods.

d. Examples:

i. Poisson: $l(\lambda;X) = \log(\Pi_{i=1}^n \exp(-\lambda)\lambda_j^X/X_j! = \Sigma_{i=1}^n \log(\exp(-\lambda)\lambda_j^X/X_j!) = \Sigma_{i=1}^n -\lambda + X_j\log(\lambda) - \log(/X_j!)$

- Setting the first derivative $= 0$, $-\Sigma_{i=1}^{n}[-1 + X_i/\hat{\lambda}] = 0$, or $\hat{\lambda} = \bar{X}$.
- Do we have a maximum? $l''(\lambda; X_1, \cdots, X_n) = -\Sigma_{i=1}^{n} X_i/\lambda^2$; always negative, and so $\hat{\lambda}$ is a global maximizer.

ii. Normal $l(\mu, \sigma; X) = -(X-\mu)^2/(2\sigma^2) - \ln(\sigma) - \frac{1}{2}\ln(2\pi)$ $\Rightarrow$ likelihood arising from an ind. sample $X_1, \cdots, X_n$ is

$$l(\mu, \sigma; X_1, \cdots, X_n) = -\frac{\Sigma_j (X_j - \mu)^2}{2\sigma^2} - n\ln(\sigma) - \frac{n}{2}\ln(2\pi).$$

- Setting the first derivative with respect to μ to 0, $-\Sigma_{j=1}^{n}(X_j - \hat{\mu})/(\hat{\sigma}^2) = 0,$ and $\Sigma_{j=1}^{n}(X_j - \hat{\mu}) = 0$, and $\hat{\mu} = \bar{X}$.
- $\frac{\partial^2 l}{\partial \mu^2} = -\sigma^{-2} < 0 \, \forall \mu, \sigma$; hence we have a minimum regardless of σ