- If we now also want to estimate $\hat{\sigma}$ at the same time, we want that pair $(\hat{\mu}, \hat{\sigma})$ that maximizes l .
- With $\mu = \bar{X}$, which σ maximizes L ?
- Setting $\frac{\partial L}{\partial \boldsymbol{\theta}} = 0$, $-\Sigma_{j=1}^{n} \frac{1}{2}(X_j - \bar{X})^2 \hat{\sigma}^{-3} \times -2 - n/\hat{\sigma} = 0$, or $\hat{\sigma} = \sqrt{\Sigma_{j=1}^{n}(X_j - \bar{X})^2/n}$.

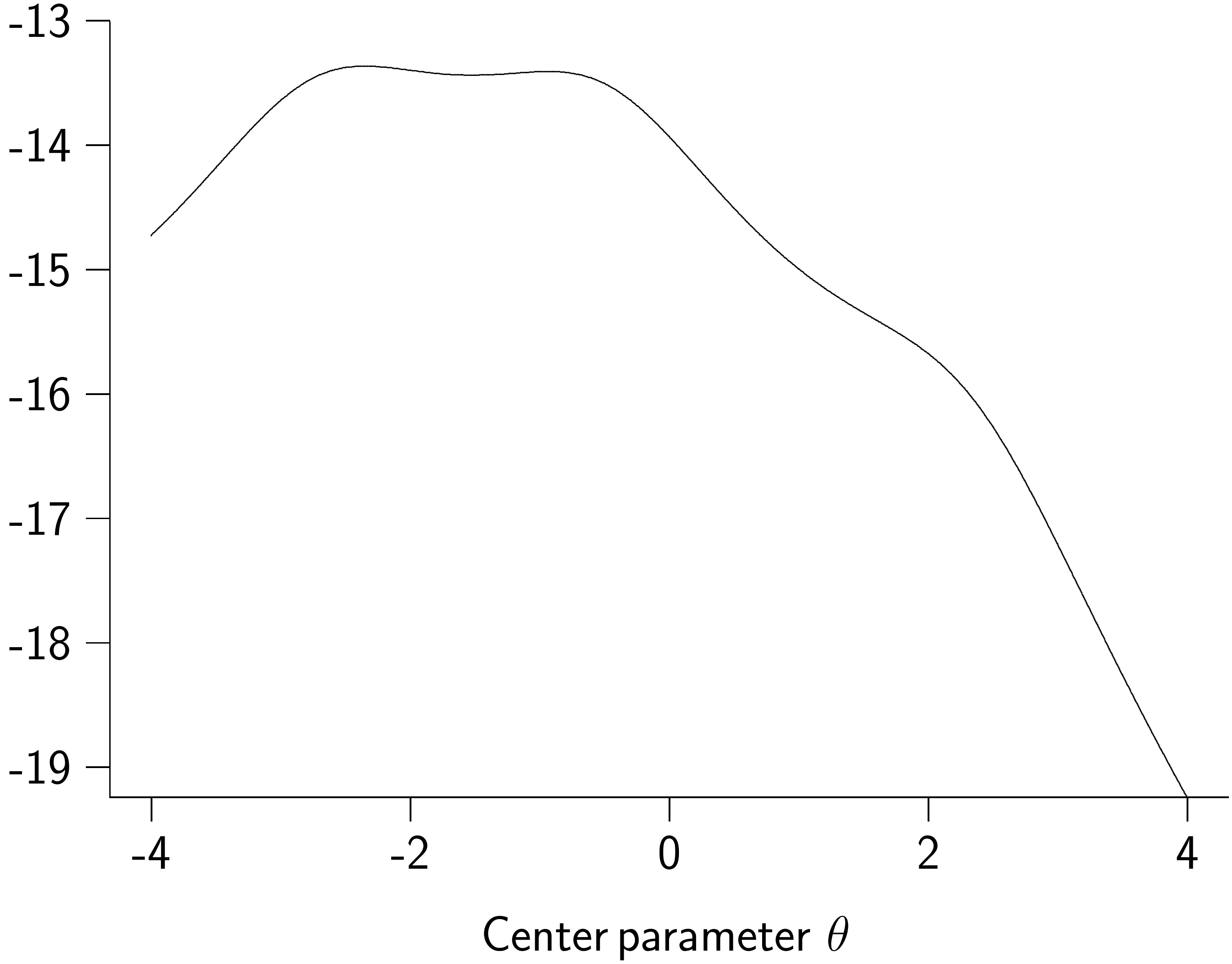
iii. Exponential: $l(\lambda; X) = -\lambda X + \ln(\lambda) \Rightarrow$ likelihood arising from an ind. sample $X_1, \cdots, X_n$ is $l(\lambda; X_1, \cdots, X_n) = -\lambda \, \Sigma_{j=1}^{n} X_j + n \ln(\lambda)$.

- Setting the first derivative $= 0$, $-\Sigma_{j=1}^{n} X_j + n/\hat{\lambda} = 0$, or $\hat{\lambda} = 1/(\Sigma_{j=1}^{n} X_j/n) = 1/\bar{X}$.
- Do we have a maximum? $l''(\lambda; X_1, \cdots, X_n) = -n/\lambda^2$; always negative, and so $\hat{\lambda}$ is a global maximizer.
- Recall that this is not an unbiased estimator; in fact, its expectation is infinite.
- mean is $\mu = 1/\lambda$
  - ▷ Similar calculations say $\hat{\mu} = \bar{X}$.

iv. Harder m.l.e. example: Cauchy distń. Take $X_1, \cdots, X_n \sim$ Cauchy μ ; $f_{X_1,\cdots,X_n}(X_1, \cdots, X_n; \mu) = \Pi \, 1/(1+(X_j-\mu)^2)$. $l(\mu, X_1, \cdots, X_n) = -\Sigma \log(1 + (X_j - \mu)^2)$.

- Likelihood equation is $-\Sigma(\mu - X_j)/(1 + (X_j - \mu)^2) = 0$.
- See Fig. 6.

Fig. 6: Log Likelihood for Cauchy Example



Data are -6.41, -19.83, -2.73, 2.34, -0.48.

v. Uniform Example:

- $X_1, \cdots, X_n \sim \mathcal{U}[0, \theta]$.
- Product of densities is

$$\prod_{i-1}^{n} \begin{cases} 1/\theta & \text{if } X_i \leq \theta \\ 0 & \text{otherwise} \end{cases} = \prod_{i=1}^{n} \begin{cases} 1/\theta & \text{if } \theta \geq X_i \\ 0 & \text{otherwise} \end{cases}$$
$$= \begin{cases} 1/\theta^n & \text{if } \theta \geq X_i \forall i \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1/\theta^n & \text{if } \theta \geq \max X_i \\ 0 & \text{otherwise} \end{cases}$$

- Density is not continuous, and so can't differentiate to maximize.
  - ▷ Also doesn't satisfy requirement of CR lower bound
  - ▷ Density is zero if $\theta < \max X_i$
  - ▷ Density decreases as theta increases if $\theta \geq \max X_i$
  - ▷ Hence MLE is $\hat{\theta} - \max X_i$

e. Invariance property: If $\boldsymbol{\tau} = g(\boldsymbol{\theta})$, for g onto, then $\hat{\boldsymbol{\tau}} = g(\hat{\boldsymbol{\theta}})$.

f. Often easier to consider this function's log $l(\boldsymbol{\theta})$.

i. $\boldsymbol{\theta}$ shows up in the exponents of the normal, exponential, and Poisson distńs, and

ii. In the above-mentioned distńs, and in the binomial distribution, for any value of $\boldsymbol{X}$, $L(\boldsymbol{\theta}) > 0 \forall \boldsymbol{\theta}$ (sound familiar)?

g. Relaxed definition:

i. Since the log likelihood is concerned with relative comparisons of potential parameter values, we can eliminate any terms not containing $\boldsymbol{\theta}$.

ii. Hence we'll also call a log-likelihood function to be that defined above, plus any function of the data **not containing** $\boldsymbol{\theta}$.

WMS: 9.4

N. Sufficiency: How much of information do we have to consider, and how much can we toss away as not giving information about the quantity of interest?

1. Example:

a. $X_1, \cdots, X_n \sim \mathcal{B}\text{in}(m, \theta)$ an ind. sample.

b. $\hat{\theta} = \Sigma_i X_i/(mn)$ is an unbiased, consistent, efficient estimator of θ .

c. Is there any other part of the data, other than that summarized by $\hat{\theta}$, that gives information about θ ?

d. The separate p.m.f.s for the variables are

$$\binom{m}{x_i}\pi^{x_i}(1-\pi)^{m-x_i},$$

e. Hence the joint p.m.f. is

$$\begin{aligned} & p_{X_1,\cdots,X_n}(x_1,\cdots,x_n;\pi) \\ &= \prod_{i=1}^{n}\binom{m}{x_i}\pi^{x_i}(1-\pi)^{m-x_i} \\ &= \pi^{\Sigma x_i}(1-\pi)^{mn-\Sigma x_i}\prod_{i=1}^{n}\binom{m}{x_i} \\ &= \pi^{mn\hat{\theta}}(1-\pi)^{mn-mn\hat{\theta}}\prod_{i=1}^{n}\binom{m}{x_i} \end{aligned}$$

and

$$p(\hat{\theta};\pi) = \binom{mn}{mn\hat{\theta}}\pi^{mn\hat{\theta}}(1-\pi)^{mn-mn\hat{\theta}};$$

hence

$$p_{X_1,\cdots,X_n|\hat{\theta}}(x_1,\cdots,x_n|\hat{\theta};\pi) = \frac{\prod_{i=1}^{n}\binom{m}{x_i}}{\binom{mn}{\Sigma_i x_i}}.$$

Hence the additional information given by the X_i after we know their total tells us nothing about π.

2. Definition: $T(X_1,\cdots,X_n)$ is *sufficient* for θ if the distń of $X_1,\cdots,X_n$ conditional on T doesn't depend on θ.

   a. *factorization theorem* : T is sufficient if and only if full p.m.f. can be factored as

$$p_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = g(t(x_1,\cdots,x_n);\theta)u(T,x_1,\cdots,x_n).$$

b. T sufficient $\Rightarrow$ p.m.f. of the data can be written

$$p_{X_1,\cdots,X_n}(x_1,\cdots,x_n;\theta) = p_T(t;\theta)\times$$
$$p_{X_1,\cdots,X_n|T}(x_1,\cdots,x_n|t(x_1,\cdots,x_n))$$

i. the latter factor independent of θ

c. You can also show other direction.

3. The ideas and theorems above also hold for densities.

4. Another example, consider $X_1,\cdots,X_n \sim N(\mu,\sigma^2)$.

a. The joint p.d.f. is

$$\begin{aligned} f_{X_1,\cdots,X_n}(x_1,\cdots,x_n) &= \prod_1^n \frac{\exp(-(x_i-\mu)^2/(2\sigma^2))}{\sigma\sqrt{2\pi}} \\ &= \frac{\exp(-(\Sigma_1^n(x_i-\mu)^2)/(2\sigma^2))}{\sigma^n(2\pi)^{n/2}} \\ &= \frac{\exp\left(\frac{-\Sigma_1^n x_i^2+2\mu\Sigma_1^n x_i-n\mu^2}{2\sigma^2}\right)}{(\sigma^n(2\pi)^{n/2})} \end{aligned}$$

b. If we think we know σ without looking at the data, the model becomes

$$\frac{\exp((2\mu\,\Sigma_1^n x_i - n\mu^2)/(2\sigma^2)) \times \exp((-\Sigma_1^n x_i^2)/(2\sigma^2))}{\sigma^n(2\pi)^{n/2}}.$$

c. Factorization shows that $\Sigma_{i=1}^n X_i$ is sufficient for μ.

i. So is $\hat{\mu} = T/n$.

ii. $\hat{\mu}$ is a good estimator but T is not.

5. Example $X, Y \sim \mathcal{P}(\theta)$

a. $\hat{\mu} = \frac{1}{3}X + \frac{2}{3}Y$

i. $\hat{\mu} = \frac{2}{3} \Rightarrow X = 2$ and $Y = 0$ or $X = 0$ and $Y = 1$

ii. $\mathrm{P}\left[X = 2|\hat{\mu} = \frac{2}{3}\right] =$

$$\frac{\exp(-\mu)\mu^2/2!\exp(-\mu)}{\exp(-\mu)\mu^2/2!\exp(-\mu) + \exp(-\mu)\exp(-\mu)\mu^1/1!} = \frac{\mu^2}{\mu^2 + 2\mu},$$

iii. depends on μ: $\hat{\mu}$ not sufficient

b. $\hat{\mu} = \frac{1}{2}X + \frac{1}{2}Y$

i. $\mathrm{P}\left[X = x|\hat{\mu} = u\right] =$

$$\frac{\exp(-\mu)\mu^x/x!\exp(-\mu)\mu^{2u-x}/(2u-x)!}{\exp(-2\mu)\mu^{2u}/(2u)!} = \frac{2u!}{x!(2u-x)!},$$

ii. does not depend on μ: sufficient

6. Hence entire data set $X_1, \cdots, X_n$ is sufficient.

a. For independent data, so is ordered data set.

7. Example where sufficient statistic doesn't tell the whole story:

a. A collection of cars is inspected for defective wheels

b. Estimate the proportion π of wheels which are defective.

c. Under the binomial model, the sample proportion is sufficient for inference on π.

d. Consider two scenarios:

Scenario 1: # of wheels defective	# of times observed	Scenario 2: # of wheels defective	# of times observed
0	5	0	44
1	19	1	0
2	36	2	0
3	27	3	0
4	13	4	56
Total	100	Total	100

i. Both scenaria give the same estimate of π

ii. the second case gives strong evidence that the binomial model is wrong.

iii. This demonstrates that the sufficient statistic tells about the parameters in the model; remainder tells about the suitability of the model itself.

WMS: 9.5

O. Rao Blackwell Theorem: Reduce the variance of an unbiased estimate by conditioning on a sufficient statistic.

1. Suppose

   a. $\tilde{\theta}$ unbiased for θ

   b. U sufficient for θ

2. Let $\hat{\theta} = \mathrm{E}\left[\tilde{\theta}|U\right]$
   a. Then $\mathrm{Var}\left[\hat{\theta}\right] = \mathrm{Var}\left[\mathrm{E}\left[\hat{\theta}|U\right]\right] + \mathrm{E}\left[\mathrm{Var}\left[\hat{\theta}|U\right]\right] \geq \mathrm{Var}\left[\tilde{\theta}\right]$.
3. Hence can find another estimator with often smaller variance.
4. Example: $X_1, \cdots, X_n \sim \mathcal{U}[0, \theta]$.
   a. $\tilde{\theta} = 2X_1$ unbiased.
   b. $U = \max X_j$ sufficient.
   c. Applying the Rao-Blackwell procedure,

$$\mathrm{E}\left[X_1|U\right] = U\mathrm{P}\left[X_1 = U|U\right] + \mathrm{E}\left[X_1 I(X_1 < U)|U\right]\mathrm{P}\left[X_1 < U\right]$$

$$= U/n + ((n-1)/n)U/2$$

   d. $\hat{\theta} = U(1 + 1/n)$.