CRAMER-RAO LOWER BOUND: Let \( x_1, \ldots, x_n \) be a sample with pdf \( f(x|\theta) \), and let \( W(x) = W(x_1, \ldots, x_n) \) be any estimator satisfying
\[
\frac{d}{d\theta} \mathbb{E}_\theta [W(x)] = \int \frac{\partial}{\partial \theta} [W(x) f(x|\theta)] \, dx
\]
and
\[
\text{Var}_\theta [W(x)] < \infty
\]
Then,
\[
\text{Var}_\theta (W(x)) \geq \frac{\left( \frac{d}{d\theta} \mathbb{E}_\theta [W(x)] \right)^2}{\mathbb{E}_\theta \left( \left( \frac{\partial}{\partial \theta} \log f(x|\theta) \right)^2 \right)}
\]

The approach in this case is completely different to the one based on complete, sufficient statistics. We try to find a bound for the variance and then see whether we can find an unbiased estimator that achieves this lower bound.

Be careful with the fact that we need to be able to take derivatives inside of the integrals. Michael stated some implications of the regularity conditions.

Be aware of some differences between what is written in the book and what Michael says in the lectures.

Special cases of the C-R formula:
- i.i.d. sample
- exponential family (and some other pdfs → see Lemma 7.3.11, p. 338)
Corollary 7.3.15 (Attainment): Let $X_1, ..., X_n$ be iid $f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the Cramer-Rao Theorem. Let $L(\theta|x) = \frac{1}{n!} f(x|\theta)$ denote the likelihood function. If $W(X) = W(X_1, ..., X_n)$ is an unbiased estimator of $T(\theta)$, then $W(X)$ attains the Cramer-Rao lower bound if

$$ a(\theta) \left[ W(X) - T(\theta) \right] = \frac{d}{d\theta} \log L(\theta|x) $$

for some function $a(\theta)$.

Using some results stated in the book we can also write the following corollary:

Corollary: If $X_1, ..., X_n$ are iid from an exponential family, with pdf $f(x|\theta)$, and if $W(X) = W(X_1, ..., X_n)$ with

$$ \frac{d}{d\theta} E_\theta [W(X)] = \int \frac{d}{d\theta} [W(x)f(x|\theta)] dx $$

then

$$ \text{Var}_\theta [W(X)] < \infty $$

$$ \text{Var}_\theta [W(X)] \geq \frac{\left[ \frac{d}{d\theta} E_\theta (W(X)) \right]^2}{-n E_\theta \left[ \frac{d^2}{d\theta^2} \log f(x|\theta) \right]} $$

Notes:
(a) It is not easy to show that the conditions of the CR thm. hold; you can try to compute the bound and then show that there is an estimator that beats it.
(b) Quick guide: in general, if the range of the pdf depends on the parameter(s), then the theorem will not be applicable.
(c) Nothing guarantees that if the C-R lower bound exists, there is an estimator that attains it.
Example: \( N(\mu, \sigma^2) \) \quad \chi \sim \text{iid} \\

\[
L(\mu, \sigma^2 \mid \chi) = \left( \frac{1}{2 \pi \sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \right\}
\]

\[
\frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2 \mid \chi) = \frac{n}{2 \sigma^4} \left( \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} - \sigma^2 \right)
\]

with \( a(\sigma^2) = \frac{n}{(2 \pi)^{n/2}} \), we have that the best unbiased estimator of \( \sigma^2 \) is \( \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \).

The problem is that you can compute iff \( \mu \) is known. \( \Rightarrow \) in the general case when \( \mu \) is unknown this expression cannot be computed.

What can we do? We can try to find an estimator with a variance "close" to the bound.

Normal distribution: C-R lower bound

\[
L(\mu, \sigma^2 \mid \chi) = \left( \frac{1}{2 \pi \sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} \right\}
\]

Compute the CR lower bound using the formula for random sample from exponential family:

\[
\frac{\partial^2}{\partial \sigma^2} \log L = -\frac{1}{2 \sigma^4} - \frac{(x - \mu)^2}{\sigma^6}
\]

\[
-\mathbb{E} \left( \frac{\partial^2}{\partial \sigma^2} \log L \right) = -\mathbb{E} \left( -\frac{1}{2 \sigma^4} - \frac{(x - \mu)^2}{\sigma^6} \right) = \frac{1}{2 \sigma^4}
\]

So any unbiased estimator of \( \sigma^2 \) satisfies

\[
\text{Var}(\hat{\sigma})(\hat{\sigma}) \geq \frac{2 \sigma^4}{n}
\]
Now, we know that

\[ \text{Var}(s^2) = \frac{2\sigma^4}{n-1} \]

Consider an alternative estimator:

\[ \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n+1} = \frac{1}{n+1} \left( \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \]

We know that

\[ Y = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \sim \chi^2_{n-1} \]

\[ \text{Var}(Y) = 2(n-1) \]

\[ \hat{\sigma}^2 = \frac{\sigma^2}{n+1}, \quad Y \Rightarrow \text{Var}(\hat{\sigma}^2) = \frac{\sigma^4}{(n+1)^2}, \quad \text{Var}(Y) \]

\[ = 2\sigma^4 \cdot \frac{n-1}{(n+1)^2} \]

\[ \text{Var}(\hat{\sigma}^2) < C-\text{R lower bound} ! \]

\[ \frac{n-1}{n+1} \cdot \frac{1}{n+1} \Rightarrow \frac{n-1}{(n+1)^2} < \frac{1}{n} \]

Why is this happening?

We are looking at estimators with different expected values.
HYPOTHESIS TESTING

- Null hypothesis: what we are assuming to be true.
- Alternative hypothesis: the alternative, literally...
- p-value: the probability of obtaining the test statistic we computed or one more extreme given our null is true.
- significance level: before we decide to reject or accept, we set up a significance level (normally 5%). This is the probability of rejecting the null when the null is true. We reject $H_0$ if $p$-value $< \text{sig. level}$
- power: $P(\text{rejecting } H_0 \mid \Theta \in \Theta_1)$. The power is a function of $\Theta$. At $\Theta = \Theta_0$ (if $\Theta_0$ is a singleton), the power is just the significance level. As $\Theta$ moves away from $\Theta_0$, the power should increase (greater chance of rejecting $\Theta = \Theta_0$).
- Type I error: Rejecting $H_0$ when $H_0$ is true.
- Type II error: $H_0$ when $H_1$ is true.
- Power = \begin{cases} P(\text{Type I}) \text{ when } \Theta \in \Theta_0 \\ 1 - P(\text{Type II}) \text{ when } \Theta \in \Theta_1 \end{cases}

Example — Intuition:
The nature of testing is to use our sample to ask questions about how likely something is in the population, and to draw conclusions based on that likelihood.

→ You have a population of men. You believe their mean height is $\Theta_0 = 68''$. The variance is known to be $\sigma^2 = 16''$. You sample 100 men, they have an average height of 69.2''.

(a) Test if the population mean $\neq 68$
(b) Test if the population mean $> 68$
Under the null hypothesis (of $\theta_0 = 68^\circ$),

$$Z = \frac{69.2 - 68}{4/10} = 3 \quad \text{has} \quad Z \sim N(0,1)$$

because $\bar{x} \sim N(68, \frac{\sigma^2}{n})$

(a) If the null is true, there is a 1% chance of observing what we observed.
(b) If the null is true, there is a 0.5% chance of observing what we observed.

Both probabilities are small - given that we observed $\bar{x} = 69.2$, it is unlikely that $\theta_0 = 68$.

**Likelihood ratio test**

$$TLR = 2 \left[ \log L(\theta_0 | x_1, \ldots, x_n) - \log L(\theta_1 | x_1, \ldots, x_n) \right]$$

→ Neyman-Pearson lemma: if this exists, it's the best.

example: $x_1, \ldots, x_n$ iid

$$f(x | \theta) = \begin{cases} e^{-(x-\theta)} & \text{if } x \geq \theta \\ 0 & \text{if } x < \theta \end{cases}$$

$$L(\theta | x) = \begin{cases} e^{-\sum x_i + n \theta} & \theta \leq x_{(n)} \\ 0 & \theta > x_{(n)} \end{cases}$$
\[ H_0: \theta = \theta_0 \quad \text{vs.} \quad H_1: \theta > \theta_0 \]

\[
\log L(\theta|x) = \begin{cases} 
-\sum x_i + n\theta & \theta \leq x(w) \\
-\infty & 0 > x(w)
\end{cases}
\]

\[ T_{LR} = \begin{cases} 
2 \left[ -\sum x_i + n x(w) - (-\sum x_i + n \theta_0) \right] & \text{if } x(w) > \theta_0 \\
0 & \text{if } x(w) \leq \theta_0
\end{cases}
\]

\[ T_{LR} = 2 \left( n(x(w) - \theta_0) \right) \]

Intuitively, you will reject for \( x(w) \) "sufficiently" far to the right ... \( x(w) > \theta_0 \) implies that you reject \( H_0 \).

\underline{Lagrange multiplier test}

\[ T_{LM}(x_1, \ldots, x_n) = \frac{\left[ \frac{\partial}{\partial \theta} \ell(\theta_0|x_1, \ldots, x_n) \right]^2}{\frac{\partial^2}{\partial \theta^2} \ell(\theta_0|x_1, \ldots, x_n)} \]

Why is this useful?
- No need to actually compute any estimator
- Has a clear interpretation in terms of a constrained optimization problem.

\[ \frac{\partial}{\partial \theta} \ell(\theta_0|x_1, \ldots, x_n) = \lambda \]

where \( \lambda \) is the Lagrange multiplier in the constrained maximization problem.
Wald test statistic

\[ T_w (x_1, ... , x_n) = \frac{(\hat{\theta}_m - \theta_0)^2}{\left[-\frac{\partial^2}{\partial \theta^2} l(\hat{\theta}_m | x_1, ... , x_n)\right]^{-1}} \]

where \[ \left[-\frac{\partial^2}{\partial \theta^2} l(\hat{\theta}_m | x_1, ... , x_n)\right]^{-1} \approx \text{Var}(\hat{\theta}_m) \]

This \( T_w \) is a different approximation to the LRT. The Wald principle is far more general and applies in many other situations.

Wald principle: reject \( H_0 \) when \(|\hat{\theta} - \theta_0| \) is "large", where \( \hat{\theta} \) is some estimator of \( \theta \).

Notes:
- if \( x_i \) iid \( N(\mu, \sigma^2) \) \( \sigma^2 \) known, all the 3 tests are numerically equivalent!

Generalization of the LRT:

\( H_0: \theta \in \Theta_0 \ vs \ H_1: \theta \in \Theta_1 \)

LR test statistic:

\[ T_{LR} (x_1, ... , x_n) = 2 \left[ l(\hat{\theta}_m | x_1, ... , x_n) - l(\hat{\theta}_0 | x_1, ... , x_n) \right] \]

where \( \hat{\theta}_m = \arg\max_{\theta \in \Theta_m} l(\theta | x_1, ... , x_n) \)
\( \hat{\theta}_0 = \arg\max_{\theta \in \Theta_0} l(\theta | x_1, ... , x_n) \)
Exercises

1. \( y_i = \beta x_i + \epsilon_i \quad \text{\( x_i \) constants (known)} \)
   \( \epsilon_i \sim N(0, \sigma^2) \quad \sigma^2 \text{ known iid} \)
   Want to test \( H_0: \beta = 0 \) vs. \( H_1: \beta \neq 0 \).

- Construct the LRT
- Show it is numerically equivalent to the traditional \( z \)-test. (the same as a \( t \)-test but it has a \( N(0,1) \) distribution because \( \sigma^2 \) is known).