Today: More likelihood and point estimation

Ex: Normal case. Let \( X_1, X_2, \ldots, X_n \) be iid RVs such that \( X_i \sim N(\mu, 1) \). What is the likelihood function for \( \mu \)?

\[
L(\mu | x) = \prod_{i=1}^{n} f(x_i | \mu) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2\right)
\]

In this case, \( \sigma^2 = 1 \)

\[
\prod_{i=1}^{n} f(x_i | \mu) = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2\right) \propto \text{constant}.
\]

As a function of \( \mu \),

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

is constant (\( C(x) \))

\[
\Rightarrow L(\mu | x) = C(x) \exp\left(-\frac{1}{2n} (\mu - \overline{x})^2\right)
\]

Notice that (1) looks like the kernel of a normal density (on \( \mu \!\!\!) with mean \( \overline{x} \) and variance \( 1/n \).

Graph

\[
L(\mu | x)
\]

This likelihood does not integrate to one. Because the constant is \( C(x) \) and not \( \frac{1}{\sqrt{2\pi(1/n)}} \).
Certainly, if we make
\[ L^*(\theta | x) = \frac{c(\theta)^{-1}}{\sqrt{2\pi n}} L(\theta | x) = \frac{1}{\sqrt{2\pi n}} \exp\left(-\frac{1}{2} \frac{(x - \theta)^2}{n}\right) \]

this a PDF on \( \theta \) (the fiducial dist on \( \theta \)).

We could compute things like:
\[ P_\theta \left[ a < x < b \right] = P_\theta \left( \frac{a - X}{\sqrt{n}} < Z < \frac{b - X}{\sqrt{n}} \right); \quad Z \sim \mathcal{N}(0, 1) \]

\[ \theta^* \]

Let \( x_1, x_2 \). \( X_n \) be iid RVs such that each \( X_i \sim \mathcal{U}(0, \theta) \)
\[ \Rightarrow f(x_i | \theta) = \frac{1}{\theta} I(x_i), \Rightarrow L(\theta | x) = \prod_{i=1}^{n} \frac{1}{\theta} I(x_i) = \frac{\theta^{-n}}{\theta^n} \]
\[ = \frac{1}{\theta^n} I(x_{(1)}) \frac{1}{\theta} I(\theta) = c \cdot \frac{1}{\theta^n} I(\theta) \quad \theta \in (0, \infty) \]

Graph of \( L(\theta | x) \)

\[ \begin{array}{c}
\text{TO TRANSFORM TO A DENSITY:} \\
\int_{\theta}^{\theta_n} d\theta = \frac{\theta^{-n}}{(n+1)} \end{array} \]

\[ = \frac{(x_{(n)})^{-n}}{(n+1)} \]

\[ \Rightarrow L^*(\theta | x) = (n+1) (x_{(n)})^{-\frac{n}{n+1}} \frac{1}{\theta} I(\theta) \quad \theta \in (x_{(n)}, \infty) \]

Is the fiducial PDF on \( \theta \).

Do you recognize this PDF? It is a \textit{Pareto pdf}
(with \( x = x_{(n)} \) and \( \beta = n+1 \))
Point Estimation

Idea: Propose a statistic \( W(X_1, X_2, X_n) \) that is a good estimator of the point \( \Theta \). (Notation: \( \hat{\Theta}(X) \), \( \hat{\delta}(X_1, X_n), \hat{\delta} \))

How to propose? What is good?

Methods of Point Estimation

Maximum Likelihood Estimation.

For each sample point \( X \), let \( \hat{\Theta}(X) \) be the value that at maximizes which \( L(\Theta | X) \) attains its maximum. (Global)

The statistic \( T = \hat{\Theta}(X) \) is a maximum likelihood estimator for \( \Theta \).

Ex: For \( X_i \sim U(0, \Theta) \), the maximum likelihood estimator is \( \hat{\Theta}(X) = \max(X_1, X_2, X_n) \).

Ex: For \( X_i \sim N(\mu, \sigma) \) the likelihood is maximized at \( \mu = \overline{X} \).

Another way to show this:

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

\[
\log L(\mu | X) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2
\]

\[
\frac{d \log L(\mu | X)}{d \mu} = 0 + 2 \left( \frac{1}{2} \right) \sum_{i=1}^{n} (x_i - \mu)
\]

Make derivative equal to 0:

\[
\frac{d \log L(\mu | X)}{d \mu} = 0 \Rightarrow \sum_{i=1}^{n} (x_i - \mu) = 0 \Rightarrow \sum_{i=1}^{n} x_i = n\mu
\]

\[
\Rightarrow \hat{\mu} = \overline{X} \text{ is a zero of the first derivative.}
\]
\[
\frac{d^2 \log L(\mu | x)}{d \mu^2} = -n < 0 \implies \hat{\mu} = \bar{x} \text{ is the MLE for } \mu.
\]

In general, if we have a likelihood function \( L(\Theta | x) \) differentiable on \( \Theta \), possible candidates for the MLE are values \((\Theta_1, \Theta_2, \ldots, \Theta_k)\) such that
\[
\frac{\partial^2 L(\Theta | x)}{\partial \Theta_i^2} = 0, \quad i = 1, 2, \ldots, k \quad \text{(you can use } \log L(\Theta | x) \text{ instead of } L(\Theta | x), \text{ since } \log(\cdot) \text{ is a one-to-one function)}.
\]

Notice that in the \( \mu(\Theta) \) case, we cannot apply this, the likelihood function is not differentiable for all values of \( \Theta \).

Ex. Suppose we have an urn with black and red balls and we extract 25 balls with replacement from this urn. And 9 of these balls are black. Let \( p = \text{prob. ball is black in any extraction} \). What is our MLE estimate for \( p ? \) Intuition tells us: \( \hat{p} = 9/25 \)

Formally, \( x \sim \text{Bernoulli}(25, p) \). Observed value of \( x \) is 9.
\[
L(p; x=9) = \binom{25}{9} p^9 (1-p)^{16} \implies \log L(p; x=9) = \log \binom{25}{9} + 9 \log p + 16 \log (1-p)
\]
\[
d log L(p; x=9) = 9 \left( \frac{1}{p} \right) - 16 \left( \frac{1}{1-p} \right) = 0 \quad \implies \quad 9(1-p) = 16p
\]
\[
\implies \quad p = \frac{16}{25} \quad \text{(check that } \frac{d^2 \log L(p; x=9)}{dp^2} < 0 \text{)}
\]

Obs: The maximum likelihood estimator is not unique.
If $X_1, X_2, \ldots, X_n \sim \text{Bin}(\theta)$ then
\[ L(\theta | X) = \theta^\sum X_i (1-\theta)^{(n-\sum X_i)} \quad 0 < \theta < 1 \]
\[ l = \log L(\theta | X) = \sum X_i \log(\theta) + (n-\sum X_i) \log(1-\theta) \]
\[ \Rightarrow \frac{dl}{d\theta} = \frac{\sum X_i}{\theta} - \frac{(n-\sum X_i)}{1-\theta} = 0 \iff \frac{(n-\sum X_i)}{1-\theta} \sum X_i = \theta (n-\sum X_i) \]
\[ \Rightarrow \sum X_i = \theta n \Rightarrow \theta = \frac{\sum X_i}{n} \text{ is the MLE} \]
\[ \frac{d^2 l}{d\theta^2} = -\frac{\sum X_i}{\theta^2} - \frac{(n-\sum X_i)}{(1-\theta)^2} < 0 \quad \hat{\theta} = \frac{\sum X_i}{n} \text{ is the MLE} \]

The MLE is not unique.

Example: Suppose $X_1, X_2, \ldots, X_n \sim \text{Unif}(\theta, \theta+1), i=1,2 \ldots,n$
\[ f(x_i | \theta) = \begin{cases} 1 & \theta \leq x_i \leq \theta+1 \\ 0 & \text{otherwise} \end{cases} \]
\[ L(\theta | X) = \prod_{i=1}^{n} f(x_i | \theta) = f(x_{(1)}) f(x_{(2)}) \cdots f(x_{(n)}) = f(\theta) \]

Plot of
\[ L(\theta | X) \]

The MLE for $\theta$ is any value between $(X_{(1)}, X_{(n)})$.

Invariance property

Usually our distribution is indexed by $\theta$ ($f(x | \theta)$)

We might be interested in finding an estimator for $\theta$ (say a function of $\hat{\theta}$).

Inv. property: If $\hat{\theta}$ is the MLE of then $\hat{\theta}(\theta)$ is the MLE for $\theta$. 
Back to Bill’s example. Suppose we want to estimate \( \tau(\theta) = \log \left( \frac{\theta}{1-\theta} \right) \) rather than just \( \theta \). Since \( \hat{\theta} = \frac{\bar{x}}{1+\bar{x}} \) is the MLE for \( \theta \), \( \tau(\hat{\theta}) = \log \left( \frac{\bar{x}}{1+\bar{x}} \right) \) is the MLE for \( \tau(\theta) \).

How do we formulate this?

If \( \tau(\theta) \) is a one-to-one function, there is no problem.

Let \( \eta = \tau(\theta) \iff \theta = \tau^{-1}(\eta) \). The likelihood for \( \eta \) is

\[
L^*(n|\eta) = \prod_{i=1}^n f(x_i|\tau^{-1}(\eta)) = L(\tau^{-1}(\eta)|x) = L(\theta|x)
\]

\[
\sup_{\eta} L^*(n|\eta) = \sup_{\eta} L(\tau^{-1}(\eta)|x) = \sup_{\theta} L(\theta|x)
\]

Thus, the maximum of \( h_\theta \) is attained at \( \eta = \tau(\hat{\theta}) \).

The problem is when we deal with functions that are not one-to-one (e.g., \( \mu^2 \) in the normal case).

If \( \hat{\theta} \) is the MLE for \( \theta \), since \( \tau(\cdot) \) is not one-to-one, we may find another value \( \theta_0 \) such that \( \tau(\hat{\theta}) = \tau(\theta_0) \).

We define the induced likelihood function for \( \tau(\theta) \)

\[
L^*(n|x) = \sup_{\theta} L(\theta|x) \\
\{ \theta: \tau(\theta) = \eta \}
\]

The value \( \hat{\eta} \) that maximizes \( L^*(n|x) \) is called the maximum likelihood estimator for \( \eta \).

By taking \( \sup_{\eta} \), we get out of the problem of lack of uniqueness in the transformation.
We must show that $L^*(\hat{\theta}|x) = L^*(\hat{\tau}(\hat{\theta})|x)$

$\Rightarrow \hat{\tau}(\hat{\theta})$ is an MLE for $\tau$.

$L^*(\hat{\theta}|x) = \sup_{\theta} \sup_{\theta} L(\theta|x) = \sup_{\theta} L(\theta|x) = L(\theta|x)$

Therefore, $\hat{\tau}(\hat{\theta})$ is an MLE.

$L^*[\hat{\tau}(\hat{\theta})|x] = \sup_{\theta} L(\theta|x) = L(\theta|x)$ because

$\theta \in \{ \hat{\theta} : \hat{\tau}(\hat{\theta}) = \hat{\tau}(\hat{\theta}) \}$

Conclusion: $L^*(\hat{\theta}|x) = L^*[\hat{\tau}(\hat{\theta})|x]$, $\hat{\tau}(\hat{\theta})$ is an MLE for $\hat{\tau}(\hat{\theta})$.

Ex: For the normal $(N, 1)$ model the MLE for

$\mu^2$ is $\hat{\mu}^2 = \bar{x}^2$.

For the Bernoulli model the MLE of $\theta = (1 - \theta)^{1/2}$

is $\bar{x}^{1/2}(1 - \bar{x})^{1/2}$.

Method of Moments.

Match moments.

Let $X_1, X_2$, $X_n$ be a sample with $k$-population moments

and source for $\theta$.

Let $X_1, X_2, X_n$ be a sample with pdf or pmf

$f(x; \theta_1, \theta_2, \theta_k)$, $\theta = (\theta_1, \theta_2, \theta_k)$

defines

$M_1 = \frac{1}{n} \sum_{i=1}^{n} x_i^1$, $\mu_1 = \mu(x^1)$

$M_2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2$, $\mu_2 = \mu(x^2)$

$M_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$, $\mu_k = \mu(x^k)$
Note that each population moment $\mu_j'$ is a function of $\Theta$, i.e., $\mu_j' = \mu_j'(\Theta_1, \Theta_2, \ldots, \Theta_K)$.

The methods of moment estimator ($\hat{\Theta}$) is obtained by solving the equations for $(\Theta_1, \Theta_2, \ldots, \Theta_K)$ in terms of $(m_1, m_2, \ldots, m_K)$.

\[
\begin{align*}
m_1 &= \mu_1'(\Theta_1, \Theta_2, \ldots, \Theta_K) \\
m_2 &= \mu_2'(\Theta_1, \Theta_2, \ldots, \Theta_K) \\
&\vdots \\
m_K &= \mu_K'(\Theta_1, \Theta_2, \ldots, \Theta_K)
\end{align*}
\]

Example: $X_1, X_2, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$, $\Theta = (\mu, \sigma^2)$

\[
\mu_1' = E(X) = \mu; \quad \mu_2' = E(X^2) = \mu^2 + \sigma^2
\]

Eqs.

\[
\begin{align*}
\mu &= \frac{1}{n} \sum_{i=1}^{n} x_i \quad (1) \\
\mu^2 + \sigma^2 &= \frac{1}{n} \sum_{i=1}^{n} x_i^2 \quad (2)
\end{align*}
\]

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

Example: $X_1, X_2, \ldots, X_n$ i.i.d. $X_i \sim U(\Theta_1, \Theta_2)$

Methods of moment estimator for $\Theta$

\[
\begin{align*}
F(x|\Theta) &= \frac{1}{\Theta} I(x) \\
E(x) &= \int_{\Theta}^\Theta x \frac{1}{\Theta} \, dx = \frac{1}{\Theta} \left[ \frac{x^2}{2} \right]_0^\Theta = \frac{\Theta}{2} \\
\hat{\Theta} = 2\bar{x}
\end{align*}
\]

(Consistent to MLE $\hat{\Theta} = \max X_1, X_2, \ldots, X_n$)
\[ x_1, x_2, \ldots, x_n \text{ iid observations such that } x \sim N(0, \sigma^2) \]
\[ E(x) = 0, \quad m_1 = \bar{x} \Rightarrow \bar{x} = 0? \]
\[ E(x^2) = \sigma^2, \quad m_2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \]

If \( x \sim \text{Cauchy} (\theta) \Rightarrow f(x | \theta) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2} \)

**Mean and Variance do not exist**

**mgf does not exist \Rightarrow The Method of Moments estimator does not exist. (Rest in Parameter Space)**

**Bayes Estimators:**
In the Bayesian approach, we consider \( \theta \) to be a random quantity subject to probability assessment (the prior distribution) \( \pi(\theta) \)

The prior reflects the subjective knowledge that a researcher has about \( \theta \) and is not based on data, but on previous experience or history.

For ex: in the problem of estimating \( \theta \) in a Bernoulli model a possible prior could be a pdf like:

\[ \pi(\theta) \text{ saying that } \theta \text{ is around } \frac{1}{2} \]

For another experiment, the prior could be

\[ \text{favoring } \theta \text{ around } \frac{2}{3} \]
Someone that does not know much about $\theta$ could use:

$$\pi(\theta) \sim U(0,1)$$

Anyway, they're all valid statements since prices are subjective.

When the data is collected $X = (x_1, x_2, \ldots, x_n)$ we update the information of the prior $\pi(\theta)$ with Bayes theorem to obtain the posterior distribution $\pi(\theta | x)$.

In fact, Bayes theorem establishes:

$$\pi(\theta | x) = \frac{f(x | \theta) \pi(\theta)}{m(x)}; \quad m(x) = \int f(x | \theta) \pi(\theta) d\theta$$

Where $m(x)$ is the marginal distribution of $X$. A Bayes estimator is a summary of $\pi(\theta | x)$, a median or a mean of this distribution.

Ex: $x_1, x_2, \ldots, x_n$ be iid Bernoulli $(\theta)$ Rvs.

Suppose the prior distribution of $\theta$ is a Beta $(\alpha, \beta)$, $\alpha, \beta$ fixed.

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}; \quad 0 < \theta < 1$$

$$f(x | \theta) = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i}$$

$$m(x) = \int \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} \pi(\theta) d\theta \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta$$
Inside the integral we recognize the form of a Beta pdf.

\[ \int_0^1 \theta^{x + \sum_i x_i - 1} (1 - \theta)^{\beta + n - \sum_i x_i - 1} d\theta = \frac{\Gamma(x + \sum_i x_i) \Gamma(\beta + n - \sum_i x_i)}{\Gamma(x + \beta + n)} \]

\[ \int_0^1 \theta^x (x + \sum_i x_i, \beta + n - \sum_i x_i) \theta^{\beta - 1} \]

\[ \Gamma(\alpha + \beta) \cdot \Gamma(\alpha + \beta + n) \]

\[ m(x) = \frac{\Gamma(\alpha + \sum_i x_i) \Gamma(\alpha + n - \sum_i x_i)}{\Gamma(\alpha + \beta + n)} = \frac{\Gamma(\alpha') \Gamma(\beta')}{\Gamma(\alpha' + \beta')} \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha' + \beta')} \cdot \frac{1}{\Gamma(\alpha' + n)} \cdot \frac{1}{\Gamma(\beta') + \beta'} \]

\[ \Rightarrow \pi(\theta | x) = \theta^{\sum_i x_i - \sum_i x_i - 1} (1 - \theta)^{(n - \sum_i x_i) - 1} \]

\[ \pi(\theta | x) = \frac{\sum_i x_i}{\Gamma(\alpha' + \beta')} \cdot \frac{\Gamma(\alpha' + \beta')}{\Gamma(\alpha + \beta + n)} \cdot \frac{\Gamma(\alpha' + \beta')}{\Gamma(\alpha' + \beta')} \cdot \frac{1}{\Gamma(\alpha' + \beta')} \cdot \frac{1}{\Gamma(\beta')} \cdot \frac{1}{\Gamma(\beta') + \beta'} \]

\[ \Rightarrow \pi(\theta | x) = \frac{\Gamma(\alpha' + \beta')}{\Gamma(\alpha' + \beta')} \cdot \theta^{\sum_i x_i - 1} (1 - \theta)^{(n - \sum_i x_i) - 1} \]

A Bayes point estimator could be the mean of this Beta

\[ \hat{\theta} = \frac{\sum_i x_i}{\sum_i x_i + \alpha'} = \frac{\sum_i x_i}{\sum_i x_i + \beta'} \]

FUNCTION OF $\alpha$, $\beta$ AND THE DATA $x$

Notice that:

\[ \hat{\theta} = \frac{\alpha + \sum_i x_i}{\alpha + \sum_i x_i + \beta + n} \]

\[ \frac{\alpha'}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \bar{x} \]

"Pooling" information from the prior and the data.
Notice that we started with a beta prior and finished with a beta posterior.

Then, the beta distribution is conjugate for Bernoulli model.

\[
\begin{align*}
\text{Beta prior: } & \quad \pi(\theta) \\
\text{Beta posterior: } & \quad \pi(\theta | x_1, x_2)
\end{align*}
\]

\[
\alpha' = \alpha + \sum x_i; \quad \beta' = \beta + n - \sum x_i \quad \text{(simple updating for the parameters)}
\]

Suppose \( x_1, x_2 \), \( x_i \) are iid \( \text{Exp} (\theta) \), where \( \theta = 1/\lambda \)

\[
f(x | \lambda) = \lambda e^{-\lambda x}, \quad x > 0
\]

Find a conjugate prior for \( \lambda \).

Determined by the likelihood function.

\[
f(x | \lambda) = \pi(\lambda) e^{-\lambda x} = \lambda^n e^{-\lambda \sum x_i} \quad \text{(sufficient statistic)}
\]

Notice that as a function of \( \lambda \), this looks like a gamma pdf with \( (\alpha = n+1; \quad \beta = 1/\sum x_i) \)

So Gamma should work as a conjugate model or family.

\[
\pi(\lambda) = \frac{1}{\beta^n \Gamma(n)} \lambda^{n-1} e^{-\lambda/\beta}, \quad \lambda > 0, \quad \alpha > 0, \beta > 0
\]

(\( \alpha/\beta \) fixed by the experimenter)
\[ m(x) = \int_0^\infty \frac{x^n e^{-\frac{\sum x_i}{\gamma}}}{\Gamma(n) \gamma^n} x^{\alpha - 1} e^{-\frac{x}{\beta}} \, dx \]

\[ = \frac{1}{\Gamma(n) \gamma^n} \int_0^\infty x^{\alpha + n - 1} e^{-\frac{\sum x_i + \gamma}{\gamma}} \, dx ; \text{ MAKE } \beta = \left( \frac{1}{\beta + \sum x_i} \right)^{-1} \]

\[ = \frac{\Gamma(n + \beta)}{\Gamma(n) \beta^n} \frac{1}{\sum x_i} \Gamma(n, \beta) \]

\[ = \frac{\Gamma(n + \beta)}{\Gamma(n) \beta^n} \frac{1}{\sum x_i} \int_0^\infty \text{Gamma}(x, \beta) \, dx = \left( \frac{\beta}{\beta + \sum x_i} \right)^\alpha \frac{\Gamma(n + \beta)}{\Gamma(n) \beta^n} \]

Posterior
\[ p(x|\lambda) = \frac{\lambda^n e^{-\lambda \sum x_i} \frac{1}{\Gamma(n) \beta^n} \lambda^{\alpha - 1} e^{-\beta \lambda}}{(\beta + \sum x_i)^{n+\alpha}} = \frac{1}{(\beta + \sum x_i)^n \Gamma(n) \beta^n} \lambda^{\alpha - 1} e^{-\beta \lambda} \]

A Bayesian estimator could be the mean of \( \lambda \) with respect to the posterior (Remember the mean = \( a/\beta \))

\[ \hat{\lambda}_B = \left( \frac{1}{\beta + a} \right) \]

Advantages of Bayesian

\[ \text{1) RESULTS ARE PRIOR DEPENDENT.} \]

\[ \text{2) CALCULATIONS OUTSIDE CONJUGATE FAMILIES MAY BE COMPLEX AND MAY REQUIRE NUMERICAL INTEGRATION.} \]

For \( m(x) \) (NOT GOING TO COVER THIS IN THIS CLASS)

Skip section 7.2.4 (EM algorithm)

Friday starts at Section 7.3. Evaluation of estimators