Sampling Theory

Examples - unemployment, household consumption survey

Random sample - set of iid rv's \( x_1, x_2, ..., x_n \); \( x_i \)'s have joint distribution \( f(x, \theta) = [f(x, \theta)]^n \)

\( \theta \) is vector of parameters (e.g., \( \theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \))

Statistic - function that doesn't depend on any of the known parameters; examples:

Sample Mean:  \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \)

Sample Variance:  \( \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \)

Sample Covariance:  \( \hat{\sigma}_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \)

Note for sample variance and covariance, must replace the mean with a function of the observations; can't have a function on the unknown parameters

Theorem: If random sample is from a population with mean \( \mu \) and variance \( \sigma^2 \), then the sample mean is a random variable with mean \( \mu \) and variance \( \sigma^2/n \).

Sample Mean:  \( \bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \)

\[ E(\bar{x}) = \frac{1}{n} (E(x_1) + E(x_2) + \cdots + E(x_n)) = \frac{1}{n} n E(X) = E(X) \]

Sample Variance: assume \( x_i \sim N(\mu, \sigma^2) \)

\[ \text{Var}(\bar{x}) = \left( \frac{1}{n} \right)^2 \left( n \text{Var}(X) \right) = \frac{1}{n} \text{Var}(X) \quad (\text{can sum variances of independent normals}) \]

General case for variance:

\[ \text{Var}(\bar{x}) = E((\bar{x} - \mu)^2) = E\left[ \left( \frac{1}{n} x_1 + \frac{1}{n} x_2 + \cdots + \frac{1}{n} x_n - \mu \right) \left( \frac{1}{n} x_1 + \frac{1}{n} x_2 + \cdots + \frac{1}{n} x_n - \mu \right) \right] = \]

\[ E \left[ \frac{1}{n} (x_i - \mu) + \frac{1}{n} (x_2 - \mu) + \cdots + \frac{1}{n} (x_n - \mu) \right] \left( \frac{1}{n} (x_1 - \mu) + \frac{1}{n} (x_2 - \mu) + \cdots + \frac{1}{n} (x_n - \mu) \right) \]

\[ = E \left[ \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu) \right] \left( \frac{1}{n} \sum_{j=1}^{n} (x_j - \mu) \right) = \]

\[ E \left[ \frac{1}{n^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right] + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} (x_i - \mu)(x_j - \mu) = \]
\[
\frac{1}{n} \sum_{i=1}^{n} E(x_i - \mu)^2 + \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} E[(x_i - \mu)(x_j - \mu)] = \\
\frac{1}{n} \sum_{i=1}^{n} Var(X) + \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} Cov(x_i, x_j) = \frac{1}{n^2} n \sigma^2 + \frac{1}{n^2} n 0 = \frac{\sigma^2}{n}
\]

**Estimation** - want estimators that are unbiased and efficient (small variance)

**Unbiased** - the mean of the distribution of the estimator equals the parameter we’re trying to estimate:

\[
E(\hat{\theta}) = \int_{-\infty}^{\infty} \theta f(\theta) \, d\theta = \theta
\]

Example - Normal Distribution

\[
\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i \sim N(\mu, \sigma^2/n) \text{ (shown last time) so } \hat{\theta}_1 \text{ is unbiased estimate}
\]

\[
\hat{\theta}_2 = \text{median}(x_1, x_2, ..., x_n); \text{ for } N(\mu, \sigma^2/n), E(\hat{\theta}_2) = \mu; \hat{\theta}_2 \text{ is unbiased estimate}
\]

\[
\hat{\theta}_3 = x_5 \sim N(\mu, \sigma^2); E(\hat{\theta}_3) = \mu; \hat{\theta}_3 \text{ is unbiased estimate}
\]

\[
\hat{\theta}_4 = 1/2(x_1 + x_2); E(\hat{\theta}_4) = 1/2[E(x_1) + E(x_2)] = 1/2[\mu + \mu] = \mu; \hat{\theta}_4 \text{ is unbiased estimate}
\]

Example - Exponential Distribution

\[
f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}, \ x > 0 \\
F(x, \theta) = 1 - e^{-x/\theta}, \ x > 0 \\
\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i; E \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right) = \frac{1}{n} \sum_{i=1}^{n} E(x_i) = \frac{1}{n} n \theta = \theta \ldots \text{ unbiased}
\]

\[
\hat{\theta}_2 = \text{median}(x_1, x_2, ..., x_n) \ldots F(x, \theta) = 1 - e^{-x/\theta} = 0.5 \Rightarrow \text{median} = \theta \ln(2) \ldots \text{ biased}
\]

\[
\hat{\theta}_3 = x_5; E(\hat{\theta}_3) = \theta; \ldots \text{ unbiased}
\]

\[
\hat{\theta}_4 = 1/2(x_1 + x_2); E(\hat{\theta}_4) = 1/2[E(x_1) + E(x_2)] = 1/2[\theta + \theta] = \theta; \ldots \text{ is unbiased}
\]

**Bias** - \( b(\hat{\theta}) = E[\hat{\theta} - \theta] \); unbiased means \( b(\hat{\theta}) = 0 \)

**Efficiency** - \( \hat{\theta}_1 \) is more efficient than \( \hat{\theta}_2 \) if \( Var(\hat{\theta}_1) < Var(\hat{\theta}_2) \)

Example - Normal Distribution

\[
Var(\hat{\theta}_1) = \sigma^2 / n \ldots \text{ this is the most efficient}
\]

\[
Var(\hat{\theta}_2) = \sigma^2/n
\]

\[
Var(\hat{\theta}_4) = \sigma^2/2
\]

**Purposely Choose Biased Estimator?** - maybe, if you have great gains in efficiency (much lower variance); example: median may be biased, but bias may be small and median has lower variance than mean so the trade-off may be worth while
Mean Squared Error - \( MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + \left[b(\hat{\theta})\right]^2 \); trades off bias and variance; new standard for estimate is lowest \( MSE \) (although sometimes unbiased is more important)

Unbiased - \( MSE(\hat{\theta}) = \text{Var}(\hat{\theta}) \)

Relative Efficiency - \( \hat{\theta}_1 \) is relatively more efficient than \( \hat{\theta}_2 \) if \( \frac{\text{MSE}(\hat{\theta}_1)}{\text{MSE}(\hat{\theta}_2)} = \frac{E[(\hat{\theta}_1 - \theta)^2]}{E[(\hat{\theta}_2 - \theta)^2]} < 1 \)

Unbiased - relatively more efficient if \( \frac{\text{Var}(\hat{\theta}_1)}{\text{Var}(\hat{\theta}_2)} < 1 \)

UMVU - uniformly minimum variance unbiased estimate; unbiased estimator, \( \hat{\theta} \), such that for any other unbiased estimator, \( \hat{\theta}^* \), \( \text{Var}(\hat{\theta}) < \text{Var}(\hat{\theta}^*) \); if we want an unbiased estimator, this will be the best one to use

Sample Variance: \( \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \) or \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \) ... \( \hat{\sigma}^2 \) has smaller \( MSE \), but \( s^2 \) is unbiased

Show \( E(s^2) = \sigma^2 \):

\[
\begin{align*}
    s^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left[ x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right]^2 \\
    &= \frac{1}{n-1} \sum_{i=1}^{n} \left[ x_i^2 - 2x_i \sum_{j=1}^{n} x_j + \left( \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2 \right] \\
    &= \frac{1}{n-1} \sum_{i=1}^{n} x_i^2 - \frac{2}{n} \sum_{j=1}^{n} x_i x_j + \frac{1}{n^2} \sum_{j=1}^{n} \left( \sum_{k=1}^{n} x_k \right) \sum_{j=1}^{n} x_j \\
    &= \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 \right] - \frac{2}{n} \left[ \sum_{i=1}^{n} x_i x_j \right] + \frac{1}{n^2} \left[ \sum_{j=1}^{n} \left( \sum_{i=1}^{n} x_i \right) x_j \right] \\
    &= \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 \right] - \frac{2}{n} \left[ \sum_{i=1}^{n} x_i x_j \right] + \frac{1}{n} \left[ \sum_{i=1}^{n} x_i^2 \right] + \frac{1}{n} \left[ \sum_{j=1}^{n} \sum_{i=1}^{n} x_i x_j \right] \\
    &= \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 \right] - \frac{2}{n} \left[ \sum_{i=1}^{n} x_i x_j \right] + \frac{1}{n} \left[ \sum_{i=1}^{n} x_i^2 \right] - \frac{1}{n} \left[ \sum_{j=1}^{n} \sum_{i=1}^{n} x_i x_j \right] \\
    &= \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 \right] - \frac{1}{n} \left[ \sum_{j=1}^{n} \sum_{i=1}^{n} x_i x_j \right] \\
    &= \frac{1}{n-1} \sum_{i=1}^{n} x_i^2 - \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{i \neq j} x_i x_j \\
    E(x_i) &= \sigma^2 + \mu^2 \quad \text{(comes from assumption that } x_i \text{ iid mean } \mu \text{ and variance } \sigma^2) \\
    \text{Proof: } \text{Var}(x_i) &= E(x_i^2) - [E(x_i)]^2 \\
    &= E(x_i^2) - \mu^2 = (\sigma^2 + \mu^2) - \mu^2 = \sigma^2 \\
    E(x_i x_j) &= \mu^2
Proof: \( \text{Cov}(x_i, x_j) = E(x_i x_j) - E(x_i)E(x_j) = E(x_i x_j) - \mu^2 \)

If \( x_i \)'s are independent \( \text{Cov}(x_i, x_j) = 0 \) so \( E(x_i x_j) = \mu^2 \)

\[
E(s^2) = E\left[ \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} x_i x_j \right] = \frac{1}{n} \sum_{i=1}^{n} E(x_i^2) - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} E(x_i x_j) = \frac{1}{n} \sum_{i=1}^{n} (\sigma^2 + \mu^2) - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mu^2 = \frac{1}{n} n \sigma^2 + \frac{1}{n} n \mu^2 - \frac{1}{n(n-1)} n(n-1) \mu^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2 \text{ ... unbiased} \]

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

\( \text{MSE}(s^2) = \text{Var}(s^2) + [b(s^2)]^2 = 2\sigma^4/(n-1) \)

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

\[
b(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2 = \left(\frac{n-1}{n}\right) E(s^2) - \sigma^2 = \left(\frac{n-1}{n}\right) \sigma^2 - \sigma^2 = -\frac{1}{n} \sigma^2 \]

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

\[
\text{MSE}(\hat{\sigma}^2) = \frac{(2n-1)\sigma^4 / n^2}{2\sigma^4 /(n-1)} = \frac{(2n-1)(n-1)}{2n^2} = \frac{(n-1)}{n} \cdot \frac{(2n-1)}{2n} < 1 \forall n \geq 1 \]

\[
\therefore \hat{\sigma}^2 \text{ is a relatively more efficient estimator of } \sigma^2 \text{ than } s^2 \text{ (but } s^2 \text{ is unbiased and } \hat{\sigma}^2 \text{ is biased... people still debate which is better to use)} \]

**Asymptotic Theory**

What happens to our statistic as the size of our sample increases? Does our statistic converge to the correct value... do \( s^2 \) and \( \hat{\sigma}^2 \) converge to \( \sigma^2 \) as \( n \to \infty \)?

**Non-Stochastic Convergence** - let \( \{b_n\} \) be a sequence of real numbers. If there exists a real number \( b \), and if for every \( \delta > 0 \) there exists an integer \( N(\delta) \) such that for all \( n \geq N(\delta) \)

\[ |b_n - b| < \delta \text{ then } b \text{ is the limit of the sequence } b_n \text{ and write } b_n \to b, \text{ read "} b_n \text{ converges to } b \" \]

**Example**

\[
\{b_n\} = \sum_{i=0}^{n} \left(\frac{1}{2}\right)^n = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots \\
b_1 = 1 \\
b_2 = 1 + \frac{1}{2} = \frac{3}{2} \\
b_3 = 1 + \frac{1}{2} + \frac{1}{4} = \frac{7}{4} \]
Converges to $b = 2$

**Statistical Convergence** - different types; some stronger than others; $1 \Rightarrow 3; 2 \Rightarrow 3; 3 \Rightarrow 4$

1. **Almost Sure Convergence**
2. **Convergence in $r^{th}$ mean**
3. **Convergence in Probability**
4. **Convergence in Distribution**

Let $b_n$ be a statistic based on a random sample of observations

e.g., $b_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ (sample mean) or $b_n = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})$ (sample variance)

use $n$ observations of $x_i$ to construct the statistic $b_n$

**Almost Sure Convergence** - $\{b_n\}$ converges almost surely to $b$ iff there exists a real number $b$ such that $\Pr[b_n \to b] = 1$; written $b_n \xrightarrow{as} b$

**Example** - $\{x_i\}$ is sequence of iid rv's with $E(x_i) = \mu < \infty$ (distribution type doesn’t matter)

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \xrightarrow{as} \mu$$

**Convergence in Probability** - $\{b_n\}$ converges in probability if there exists a real number $b$ such that for every $\varepsilon > 0$, $\Pr[|b_n - b| < \varepsilon] \to 1$ as $n \to \infty$; written $b_n \xrightarrow{p} b$

**Example** - $\{x_i\}$ is sequence of rv's with $E(x_i) = \mu$, $\text{Var}(x_i) = \sigma^2 < \infty$, and $\text{Cov}(x_i,x_j) = 0$; assumptions are much weaker (easier to verify) than almost sure convergence; no assumptions about identical distribution or independence ($\text{Cov} = 0 \neq$ independent; only linear independence)

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \xrightarrow{p} \mu$$

**Convergence in $r^{th}$ Mean** - used in time series; $\{b_n\}$ is sequence of real valued rv’s; if there exists a real number $b$ such that $E[|b_n - b|^r] \to 0$ as $n \to \infty$ for some $r > 0$, then $b_n$ converges in $r^{th}$ mean; written $b_n \xrightarrow{(r)} b$

**Quadratic Mean** - use $r = 2$ (which we usually do); $b_n \xrightarrow{q.m.} b$; note $|b_n - b|^2 = (b_n - b)^2$

so $E[(b_n - b)^2] = E[(b_n - b)^2] = \text{Var}(b_n)$; converges if variance of statistic converges to zero

**Lower $r$** - $b_n \xrightarrow{(r)} b$ for some $r \geq 1$, then $b_n \xrightarrow{(s)} b$ for $0 < s < r$ (i.e. quadratic mean convergence $\Rightarrow$ mean convergence, $r = 1$)

**Convergence in Distribution** - $\{b_n\}$ is sequence of real valued rv’s with distribution functions $\{F_n\}$; if $F_n(x) \to F(x)$ as $n \to \infty$ for every continuity point in $x$, where $F(x)$ is the distribution function of a rv $X$, then $b_n$ converges in distribution to the rv $X$; written $b_n \xrightarrow{d} X$

**Example** - $t_n \xrightarrow{d} N(0,1)$
Consistent Estimator - \( \hat{\theta} \) is consistent estimator of \( \theta \) iff \( \hat{\theta} \to \theta \)

Examples:

\[
\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ is consistent estimator of } \theta \text{ (mean)}
\]

\[
\hat{\theta}_2 = \text{median; if Cov}(x_i, x_j) = 0 \text{ and } E(x_i) = \mu \text{ and } Var(x_i) = \sigma^2 < \infty, \text{ and } x_i \text{ come from a symmetric distribution, then median = mean and } \hat{\theta}_2 \text{ is consistent estimator of } \theta
\]

Exponential Distribution - use median/\( \ln(2) \)

\[
\hat{\theta}_3 = x_i \text{ and } \hat{\theta}_4 = \frac{1}{2}(x_1 + x_2) \text{ are not consistent estimators; they don't change at all as sample size gets bigger}
\]

Weak Law of Large Numbers (WLLN) - as sample size increases, the sample mean converges in probability to population mean; 4 sets of conditions under which WLLN holds

Khinchin - \( \{x_n\} \) is a sequence of iid rv's with finite mean \( \mu \)

\[
\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \to \mu
\]

Chebychev - \( \{x_n\} \) is a sequence of independent rv's with means \( \mu_n \) and variances \( \sigma_n^2 \); if variances are bounded above (i.e., \( \sigma_n^2 < c < \infty \)) and

\[
\bar{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \mu_i, \text{ then } (\bar{x}_n - \bar{\mu}_n) \to 0 \text{ (sample mean minus mean of the means)}
\]

Markov - \( \{x_n\} \) is a sequence of rv's with means \( \mu_n \), if \( Var(\bar{x}_n) \to 0 \) as \( n \to \infty \), then

\[
(\bar{x}_n - \bar{\mu}_n) \to 0 \text{ (doesn't assume independence)}
\]

Kolmogorov - \( \{x_n\} \) is a sequence of independent rv's; \( z_n = \bar{x}_n - \bar{\mu}_n \)

if \( \lim_{n \to \infty} E\left[\frac{z_n^2}{1 + z_n^2}\right] = 0 \) then \( z_n \to 0 \)

Strong Law of Large Numbers - interested in almost sure convergence;

Theorem - \( \{x_n\} \) is a sequence of iid rv's, then \( \bar{x}_n \stackrel{as}{\to} \mu \) iff \( E[|x_n|] < \infty \) (i.e., finite mean)

Kolmogorov II - \( \{x_n\} \) is a sequence of independent rv's with finite variances

if \( \frac{1}{n^2} \sum_{i=1}^{n} Var(x_n) < \infty \), then \( (\bar{x}_n - \bar{\mu}_n) \to 0 \) (allows mean to change over time, but mean of sample means approaches mean of means)

Central Limit Theorem - for a large sample from any distribution, we can approximate the distribution of the sample mean with a normal distribution; \( \{x_n\} \) be a sequence of rv's with

\[
s_n = \sum_{i=1}^{n} x_i \text{ and } \bar{x} = \frac{1}{n} s_n; \text{ the standardized mean } z_n = \frac{\bar{x}_n - E(\bar{x}_n)}{\sqrt{Var(\bar{x}_n)}} = \frac{s_n - E(s_n)}{\sqrt{Var(s_n)}}, \text{ where}
\]

\[
Var(s_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 \text{ and } Var(\bar{x}_n) = \frac{1}{n} \sigma^2; \text{ } z_n \sim N(0,1)
\]
\[ \bar{x}_n \rightarrow N(E(\bar{x}_n), \text{Var}(\bar{x}_n)) = N(\mu, \frac{1}{n} \sigma^2) \]

**De Moivre** - proved CLT where \( x_i \)’s are independent Bernoulli rv’s

**Lindberg-Levy** - proved CLT where \( x_i \)’s are iid with \( \text{Var}(x_i) = \sigma^2 < \infty \)

**Lindberg-Feller** - proved CLT where \( x_i \)’s are independent with \( E(X) = \mu_i \) and \( \text{Var}(x_i) = \sigma^2_i \),

\[ s^2_n \equiv \sum_{i=1}^{n} \sigma^2_i^2, \quad S^2_n \equiv \sum_{i=1}^{n} x_i; \quad \text{for some } \epsilon > 0, \lim_{n \to \infty} \frac{1}{s^2_n} \sum_{i=1}^{n} \int (x - \mu_i)^2 \cdot f_i(x) dx = 0 \]

Only looking at portion of variance that is far away from mean; if that is converging to zero (i.e., tails aren’t too big), the CLT holds

**Multivariate CLT** - \( \{x_n\} \) is sequence of \( k \)-variate iid rv’s (\( x_n \) is \( k \times 1 \) vector) with mean \( \mu \) and variance \( \Sigma \), then \( z_n = \sqrt{n}(\bar{x}_n - \mu) \rightarrow N_k(0, \Sigma) \)

**Chebychev’s Inequality** - for constant \( k > 0 \), \( \Pr(|x - E(X)| \geq k) \leq \frac{\text{Var}(X)}{k^2} \); probability that you’re more than \( k \) away from the mean is less than or equal to variance over \( k^2 \); provides an upper bound on \( \Pr(|x - E(X)| \geq k) \)… usually a very generous upper bound (much higher than the actual probability will be… will see on HW3)

**As Lower Bound** - \( \Pr(|x - E(X)| < k) > 1 - \frac{\text{Var}(X)}{k^2} \)

**Markov’s Inequality** - \( \Pr(x \geq \lambda E(X)) \leq \frac{1}{\lambda} \) for a positive rv \( x \) and \( \lambda > 0 \);

**Example** - look at exponential: \( f(x) = e^{-x/\theta}, E(X) = \theta \)… \( \Pr(x \geq 2\theta) \leq \frac{1}{2} \)… yep

**Likelihood Functions** - how to estimate parameters; let \( x_1, x_2, \ldots, x_n \) be a random sample from a density function \( f(x; \theta) \) where \( \theta \) is vector of distribution parameters; likelihood function

\[ L(\theta; x) = f(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta); \]

looks just like a joint pdf except now the \( x_i \)’s are given and the parameters are the unknowns

**Example** - \( x_1, x_2, \ldots, x_n \) are iid \( N(\mu, \sigma^2) \) so

\[ f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2 \pi \sigma^2}} e^{-\frac{1}{2 \sigma^2}((x - \mu)^2)} \]

\[ L(\mu, \sigma^2; x) = (2\pi \sigma^2)^{-n/2} e^{-\frac{1}{2 \sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2} \]

**Log-Likelihood Function** - natural logarithm of the likelihood function; will be negative because you’re taking \( \ln \) of probabilities (< 1)

**Example** - \( \ln[L(\mu, \sigma^2; x)] = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{-1}{2 \sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \)

**Information Matrix** - estimate parameters for distribution from sample (e.g., \( \mu \) and \( \sigma^2 \) for normal distribution); need to estimate covariance matrix for unbiased estimators

**Single Parameter** - \( I(\theta) = -E \left[ \frac{\partial^2 \ln L(\theta; x)}{\partial \theta^2} \right] = E \left[ \left\{ \frac{\partial \ln L(\theta; x)}{\partial \theta} \right\}^2 \right] \)
2 Or More Parameters - \( I_{i,j} = -E \left[ \frac{\partial^2 \ln L(x; \theta)}{\partial \theta_i \partial \theta_j} \right] = E \left[ \frac{\partial \ln L(x; \theta)}{\partial \theta_i} \frac{\partial \ln L(x; \theta)}{\partial \theta_j} \right] \)

Information Matrix - \( I(\theta) = \begin{bmatrix} I_{11} & I_{12} & \cdots & I_{1k} \\ I_{21} & I_{22} & \cdots & I_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ I_{k1} & I_{k2} & \cdots & I_{kk} \end{bmatrix} \); \( I_{ij} = I_{ji} \) so it's symmetric matrix

Example - \( x_1, x_2, \ldots, x_n \) are iid \( \text{N}(\mu, \sigma^2) \) so

\[
\ln[L(\mu, \sigma^2; x)] = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
\]

\[
\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)
\]

\[
\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2}
\]

\[
E \left[ \frac{\partial^2 \ln L}{\partial \mu^2} \right] = E \left[ \frac{-n}{\sigma^2} \right] = -\frac{n}{\sigma^2}
\]

\[
\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2
\]

\[
E \left[ \frac{\partial \ln L}{\partial \sigma^2} \right] = E \left[ \frac{-n}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} \sum_{i=1}^{n} (x_i - \mu)^2 \right] = \frac{n}{2(\sigma^2)^2} + \frac{-n}{(\sigma^2)^3} \sum_{i=1}^{n} [x_i - \mu]^2 = 0
\]

\[
\frac{\partial^2 \ln L}{\partial \sigma^2 \partial \mu} = -\frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)
\]

\[
E \left[ \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \mu} \right] = E \left[ \frac{-1}{(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu) \right] = -\frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} (E(x_i) - \mu) = -\frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} (\mu - \mu) = 0
\]

\[
\frac{\partial^2 \ln L}{\partial \sigma^2} = -\frac{1}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} \sum_{i=1}^{n} (x_i - \mu)^2
\]

\[
E \left[ \frac{\partial^2 \ln L}{\partial \sigma^2} \right] = E \left[ \frac{-n}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} \sum_{i=1}^{n} (x_i - \mu)^2 \right] = \frac{n}{2(\sigma^2)^2} + \frac{-n}{(\sigma^2)^3} \sum_{i=1}^{n} E[(x_i - \mu)^2] = 0
\]

\[
\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} \sum_{i=1}^{n} (x_i - \mu)^2
\]

\[
E \left[ \frac{\partial \ln L}{\partial \sigma^2} \right] = E \left[ \frac{-n}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} \sum_{i=1}^{n} (x_i - \mu)^2 \right] = \frac{n}{2(\sigma^2)^2} + \frac{-n}{(\sigma^2)^3} \sum_{i=1}^{n} E[(x_i - \mu)^2] = 0
\]

\[
I(\mu, \sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2(\sigma^2)^2} \end{bmatrix}
\]

Regularity Conditions -

(1) \( \frac{\partial \ln L}{\partial \theta} \) exists for all \( x \) and \( \theta \)

(2) \( \frac{\partial}{\partial \theta} \int L(\theta, x) dx = \int \frac{\partial}{\partial \theta} \ln L(\theta, x) dx \) (i.e., can switch order of integration an differentiation)

(3) \( \frac{\partial}{\partial \theta} \int \hat{\theta}(x) \ln L(\theta, x) dx = \int \hat{\theta}(x) \frac{\partial \ln L(\theta, x)}{\partial \theta} dx \)
\(0 < E \left[ \left( \frac{\partial \ln L}{\partial \theta} \right)^2 \right] < \infty \) for all \( \theta \)

(5) \( \frac{\partial \ln L}{\partial \theta} \) and \( \frac{\partial^2 \ln L}{\partial \theta^2} \) exist and are continuous for all \( \theta \)

**Cramer-Rao Lower Bound (CRLB)**

**Single Parameter** - if regularity conditions hold, if \( \hat{\theta} \) is an unbiased estimator of the parameter \( \theta \) then \( \text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)} \)

**Multiple Parameters** - if \( \hat{\theta}_j(x) \) is an unbiased estimator of \( \theta_j \) for \( j = 1, \ldots, k \) (k parameters).

Assuming regularity conditions hold \( \text{Cov}(\hat{\theta}) - [I(\theta)]^{-1} \) is a positive semi-definite matrix (i.e., \( [I(\theta)]^{-1} \) is a lower bound for the covariance matrix of an unbiased estimator)

**Significance** - if you have an unbiased estimator that attains the CRLB of the variance, then you know that this is the most efficient unbiased estimator (UMVU); it's not always possible to find an unbiased estimator that attains the CRLB

**Example** - consider a random sample from a normal distribution \( N(\mu, \sigma^2) \)

Unbiased estimators:
\[
\bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

Information matrix (did this last time)
\[
I(\mu, \sigma^2) = \begin{bmatrix}
\frac{n}{\sigma^2} & 0 \\
0 & \frac{n}{2(\sigma^2)^2}
\end{bmatrix}
\]

Take inverse of this
\[
[I(\mu, \sigma^2)]^{-1} = \begin{bmatrix}
\frac{\sigma^2}{n} & 0 \\
0 & \frac{2\sigma^4}{n}
\end{bmatrix}
\]

so the CRLB for the variance of \( \bar{\mu} \) is \( \sigma^2/n \)

\[
\text{Var}(\bar{\mu}) = \text{Var} \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(x_i) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n} \quad \text{so} \quad \bar{\mu} \quad \text{is the CRLB}
\]

\[
\text{Var}(s^2) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sigma^2} \sim \chi^2_{(n-1)} \quad \Rightarrow \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}
\]

\[
\text{Var}(s^2) = \frac{(\sigma^2)^2}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{n-1} \quad \text{so} \quad s^2 \quad \text{is NOT the CRLB} \quad = \frac{2\sigma^4}{n}
\]

**Cramer-Rao Inequality** - let \( x_i \sim f(x; \theta) \) and \( T = T(x_1, \ldots, x_n) \) be a statistic such that \( E(T) = u(\theta) \) (some function of \( \theta \)). Assume regularity conditions. Then \( \text{Var}(T) \geq \left[ u'(\theta) \right]^2 / I(\theta) \)
**Maximum Likelihood Estimator (MLE)** - choose values of the parameters that maximize the likelihood function (or the log-likelihood function). Take all the partial derivatives of $\ln L(\theta; x)$ and set them equal to zero and solve for $\hat{\theta}$

**Score** - derivative of $\ln L(\theta, x)$

**Score Vector** $S(\hat{\theta})$ - comprised of all partial derivatives of $\ln L(\theta; x)$

**Example** - Let $x_1, \ldots, x_n$ be a random sample from a $N(0, \theta)$ distribution

$$L(\theta; x) = (2\pi \theta)^{-n/2} \prod_{i=1}^{n} \exp \left\{ -\frac{x_i^2}{2\theta} \right\} = (2\pi \theta)^{-n/2} \exp \left\{ -\frac{1}{2\theta} \sum_{i=1}^{n} x_i^2 \right\}$$

$$\ln L(\theta; x) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\theta) - \frac{1}{2\theta} \sum_{i=1}^{n} x_i^2$$

$$\frac{\partial \ln L}{\partial \theta} = -\frac{n}{2\theta} + \frac{1}{2\theta^2} \sum_{i=1}^{n} x_i^2 = 0$$

$$-n + \frac{1}{\theta} \sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$ is the MLE of $\theta$

**Example** - Let $x_1, \ldots, x_n$ be a random sample from a Pareto distribution

$$f(x; \theta) = \theta x^{-\theta - 1} \text{ for } 1 \leq x < \infty$$

$$L(\theta; x) = \prod_{i=1}^{n} \theta x_i^{-\theta - 1} = \theta^n \prod_{i=1}^{n} x_i^{-\theta - 1}$$

$$\ln L(\theta; x) = n \ln(\theta) - (\theta + 1) \sum_{i=1}^{n} \ln(x_i)$$

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} \ln(x_i) = 0$$

$$\frac{n}{\theta} = \sum_{i=1}^{n} \ln(x_i)$$

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \ln(x_i)}$$

**Example** - Let $x_1, \ldots, x_n$ be a random sample from a distribution with density

$$f(x; \theta) = \frac{1}{\theta} \text{ for } 0 \leq x \leq \theta, \text{ 0 otherwise}$$

$$L(\theta; x) = \prod_{i=1}^{n} \frac{1}{\theta} = \theta^{-n} \text{ for } 0 \leq x_i \leq \theta, i = 1, \ldots, n, \text{ 0 otherwise}$$

$$\ln L(\theta; x) = -n \ln(\theta) \text{ for } 0 \leq x_i \leq \theta, i = 1, \ldots, n, \text{ 0 otherwise}$$

$$\frac{\partial \ln L}{\partial \theta} = \frac{-n}{\theta} = s(\theta; x) - n \ln(\theta) \text{ for } 0 \leq x_i \leq \theta, i = 1, \ldots, n, \text{ 0 otherwise}$$

**Problem** - there is a discontinuity at $\theta$; $s(\theta; x)$ is not a function of $x$; can’t set $s(\theta; x) = 0$
Solution - want to make \( s(\theta; x) \) as close to zero as possible so we want to make \( \theta \) as large as possible; we know that \( \theta \geq x \), \( \therefore \) min possible value of \( s(\theta; x) \) is at \( \hat{\theta} = \max(x_1, \ldots, x_n) \)

Multiple Parameter Example - Let \( x_1, \ldots, x_n \) be a random sample from \( N(\mu, \sigma^2) \)

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

\[
\ln L(\mu, \sigma^2; x) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 
\]

\[
\frac{\partial \ln L}{\partial \mu} = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \mu) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) 
\]

\[
\frac{\partial \ln L}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 
\]

\[
s(\mu, \sigma^2; x) = \left[ \begin{array}{c} \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{array} \right] = \left[ \begin{array}{c} 0 \\ 0 \end{array} \right] 
\]

\[
\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \Rightarrow \sum_{i=1}^{n} (x_i - \mu) = 0 \Rightarrow \left( \sum_{i=1}^{n} (x_i) \right) - n\mu = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} 
\]

Substitute into the second equation

\[
-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0 \Rightarrow \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = n \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 
\]

NOTE: This is a biased estimator

Properties of MLEs

Invariance - Let \( \hat{\theta} \) be a MLE of \( \theta \). If \( g(\bullet) \) is a function of \( \theta \) then the MLE of \( g(\theta) \) exists and is given by \( g(\hat{\theta}) \)

Consistency and Uniqueness - under regularity conditions for Cramer-Rao lower bound, there exists a solution vector to the likelihood equations that is consistent

\[
\lim_{n \to \infty} \text{Pr} \left[ \text{S}^2(\hat{\theta}_n) = 0 \right] = 1 \quad (\text{existence}) \quad \text{and} \quad \hat{\theta}_n \to \theta \quad (\text{consistency}) 
\]

Asymptotic Normality - \( \lim_{n \to \infty} \text{I}(\theta) / n = \Sigma(\theta) \) under the regularity conditions of the CRLB

\[
\sqrt{n} (\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \Sigma^{-1}(\theta)) \quad \text{i.e., MLEs have normal distribution at the limit} 
\]

Asymptotic Efficiency - asymptotic variance of \( \hat{\theta}_n \) equals the limit of the CRLB

Inverse of 2x2 matrix
\[
\begin{align*}
\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} &= \frac{1}{ad - cb} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}
\end{align*}
\]