Recall previous lecture:

\[ X \] - binomial random variable
with \( n \) and \( p \), \( \widetilde{p} \) typically known

\[ X = \text{total \# of Heads when an \"unfair\ coin\" is flipped} \ n \ \text{times}. \]

Parameter of interest: \( p \)

Estimator: \( \hat{p} = \frac{X}{n} \)

\[ \text{How good is this estimator?} \]
- We know it is unbiased

\[ \mathbb{E} \left[ \hat{p} \right] = \mathbb{E} \left[ \frac{X}{n} \right] = \frac{1}{n} \mathbb{E} [X] = \frac{1}{n} \cdot np = p \]

- Variance?

\[ \mathbb{V} \left[ \hat{p} \right] = \frac{1}{n^2} \mathbb{V} [X] = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n} \]
But we would not know how big

\[ \sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]

is when we are estimating \( \hat{p} \).

Any thing we can do?

How about we say:

We *estimate* that the standard deviation of the *estimator* \( \hat{p} \) is

\[ \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]

i.e. replace \( \hat{p} \) above by \( \hat{p} \).
Stand error of an estimator $\hat{\theta} = \sqrt{V(\hat{\theta})}$

Similar Example.

Example 6.2.

$x_1, \ldots, x_n \sim N(\mu, \sigma)$

We have random samples $x_1, \ldots, x_n$ but we do not know $\mu$ and $\sigma$.

$\hat{\mu} = \bar{x} = \frac{\sum x_i}{n}$ \hspace{1cm} \leftarrow \text{an unbiased estimator for } \mu$

$\sqrt{V(\hat{\mu})} = \frac{\sigma}{\sqrt{n}}$ \hspace{1cm} \leftarrow \text{standard error}$

$\text{MSE}(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$ \hspace{1cm} q

We don't know it.

So use the sample standard deviation

$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$

$\hat{\sigma} = \text{"Estimated Standard Error of the estimator } \bar{x} \"$
Section 6.2

1. \( f(x) \) — a pdf or pmf

For \( k = 1, 2, 3, \ldots \), the \( k \)th population moment or \( k \)th moment of the distribution \( f(x) \), is

\[
E(X^k) = \begin{cases} 
\int_{-\infty}^{\infty} x^k f(x) \, dx & \text{continuous case} \\
\sum_{x} x^k f(x) & \text{discrete case} 
\end{cases}
\]

2. If \( x_1, \ldots, x_n \) are random sample from the distribution \( f(x) \), then the \( k \)th sample moment is

\[
\frac{1}{n} \sum_{i=1}^{n} x_i^k
\]

\( \Phi \) is a random variable before you “see” the samples

a number after you “see” the samples.
Recall:

In practice, people model quantities of interest using random variables, with a postulated distribution, kind of.

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

Parameters
More generally, a family of pdf/pmf may be of the form

\[ f(x; \theta_1, \ldots, \theta_m) \]

parameters.

Typically:

1. There is good empirical/scientific/mathematical reason, or we blindly believe, that the quantity of interest is well-modeled by such a family of pdf/pmf.

2. But we do not know which choice of \( \theta_1, \ldots, \theta_m \) is the "right" set of parameters to use.

3. We have random samples from this (unknown) distribution \( X_1, \ldots, X_n \).

Q: How to use the random samples to determine the "best" \( \theta_1, \ldots, \theta_m \)?
Moment Estimators:

Solve the system of $m$ eqts:

$k$th population moment

\[ E[X^k] = \text{kth sample moment of the samples.} \]

\[ \int_{-\infty}^{\infty} x^k f(x; \theta_1, \ldots, \theta_m) \, dx \quad \text{for } k = 1, 2, \ldots, m \]

unknowns: $\theta_1, \ldots, \theta_m$

$m$ eqts, $m$ unknowns
Example:

gamma distribution $f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

$(\alpha, \beta > 0)$

$E[X] = \alpha \beta \quad \text{V}[X] = \alpha \beta^2$

$E[X^2] = \text{V}[X] + E[X]^2 = \alpha \beta^2 + \alpha^2 \beta^2$

Solve

\[
\begin{align*}
\alpha \beta &= \frac{1}{n} \sum_{i=1}^{n} x_i \quad \Rightarrow \quad m_1 \\
\alpha \beta^2 + \alpha^2 \beta^2 &= \frac{1}{n} \sum_{i=1}^{n} x_i^2 \quad \Rightarrow \quad m_2 \quad \beta = \frac{m_2 - m_1^2}{m_1} \\
\alpha^2 \beta^2 &= m_1 \\
\alpha \beta^2 + m_1^2 &= m_2, \quad \frac{\alpha \beta^2}{\alpha \beta} = \frac{m_2 - m_1^2}{m_1}
\end{align*}
\]
moment estimators for $\alpha, \beta$

$$\hat{\alpha} = \frac{\bar{x}^2}{\frac{1}{n} \sum x_i^2 - \bar{x}^2}$$

$$\hat{\beta} = \frac{\frac{1}{n} \sum x_i^2 - \bar{x}^2}{\bar{x}}$$

numerical examples: (pg 244)

152 115 109 94 88 137 152 77 160 165
125 40 128 123 136 101 62 153 83 69

$$\bar{x} = \frac{(113.5)^2}{14087.8 - (113.5)^2} \approx 10.7$$

$$\bar{x} = \frac{14087.8 - (113.5)^2}{113.5} \approx 10.6$$
Maximum Likelihood Estimation
- first invented by R.A. Fisher 1920's

Recall the setup: THINK OF JOINT DISTRIBUTION $f(x_1, \ldots, x_n)$

THINK RANDOM VECTOR $\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$

THINK: what is the chance that this random vector is in a neighborhood of $\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$?
The joint pdf gives you this information.

\[ P \left[ \frac{x_i - \frac{S}{2}}{\frac{S}{2}} \leq x_i \leq \frac{x_i + \frac{S}{2}}{\frac{S}{2}}, \text{ for all } i = 1, \ldots, n \right] \]

\[ \lim_{S \to 0} \frac{g_n}{8^n} \]

\[ f(x_1, \ldots, x_n) \]

\[ f \] joint pdf.

Special case

\[ x_1, \ldots, x_n \text{ are independent (not necessarily identical)} \]

\[ f(x_1, \ldots, x_n) = f_1(x_1) \cdot f_2(x_2) \cdots f_n(x_n) \]

\[ \phi \] can be different pdf's

Very special case

\[ f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n) \]

\[ \text{i.i.d.} \]

Now let \( x_1, x_2, \ldots, x_n \) be a random sample (i.e. they are i.i.d.) of a distribution with p.d.f.

\[ f_1(x) = f_1(x; \theta_1, \ldots, \theta_m) \]
Then the joint p.d.f. of \( x_1, \ldots, x_n \) is
\[
f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m) = f_1(x_1) f_2(x_2) \cdots f_m(x_n)
\]

Principle of MLE:
\( x_1, \ldots, x_n \) — observed samples (fixed \#s after you think of them!)

1. Think

\[
f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m)
\]
fixed \#s
free parameters
as a function of \( \theta_1, \ldots, \theta_m \)
usually called the likelihood function.

2. MLE estimates:

Find \( \hat{\theta}_1, \ldots, \hat{\theta}_m \) such that
\[
f(x_1, \ldots, x_n; \hat{\theta}_1, \ldots, \hat{\theta}_m)
\]
fixed, observed samples \( \geq f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m) \)
for all \( \theta_1, \ldots, \theta_m \).

As usual, you can take the point of view that \( \hat{\theta}_1, \ldots, \hat{\theta}_m \) are r.v.'s because the samples are, before you use them, random variables!
Ex: \( X_1, \ldots, X_n \) random sample from an exponential distribution \( X_i \sim \text{Exp}(\lambda) \)

\[
f(x_1, \ldots, x_n; \lambda) = \begin{cases} \left( \prod e^{-\lambda x_i} \right) & \text{all } x_i > 0 \\ 0 & \text{otherwise} \end{cases}
\]

Since \( \ln \) is an increasing function

\[\ln(x)\]

maximizing \( \ln f(x_1, \ldots, x_n; \lambda) \) is the same as maximizing

\[
\ln f(x_1, \ldots, x_n; \lambda) = n \ln \lambda - \lambda \sum x_i
\]

(when \( x_i > 0 \))

Calculus I:

\[
\hat{\lambda} = \frac{\partial}{\partial \lambda} \ln f(x_1, \ldots, x_n; \lambda) = \frac{n}{\lambda} - \frac{\sum x_i}{\lambda}
\]

so \( \hat{\lambda} = \frac{n}{\sum x_i} \) (\( = \frac{1}{\bar{x}} \))
Comment: (Recall p.177) 
\[ E \left[ a \text{ r.v. with exponential distribution of parameter } \lambda \right] \]
\[ = \frac{1}{\lambda} \]

So \( \lambda = \frac{1}{\text{(True expected value)}} \)

Hence using \( \frac{1}{\overline{x}} \) (sample mean) to estimate \( \lambda \)

seems very reasonable.

However: this estimator is actually biased, as

\[ E \left[ \frac{1}{x} \right] \neq \frac{1}{E[x]} = \lambda \]

\[ \uparrow \]

what we want

\[ \uparrow \]

what we use to estimate what we want.
Example

\( X_1, \ldots, X_n \) random sample from a normal distribution.

\[
\begin{align*}
f(x_1, \ldots, x_n; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \quad \text{for} \quad i = 1, \ldots, n \quad \text{and} \quad \sigma^2 > 0 \\
&= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \\
\ln f(x_1, \ldots, x_n; \mu, \sigma^2) &= -\frac{n}{2} \left(\ln 2\pi + \ln \sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
\end{align*}
\]

In the following calculation, we think of \( \sigma^2 \) as a variable by itself, i.e. we pretend that we forget the fact that \( \sigma^2 = \sigma \cdot \sigma \).

This will make the calculation a little simpler.

\[
\begin{align*}
\frac{\partial f}{\partial \mu} &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \mu)(-1) = \frac{1}{\sigma^2} \left[\sum_{i=1}^{n} x_i\right] - n\mu \\
\frac{\partial f}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} - \frac{1}{2} (-1)(\sigma^2)^{-2} \sum_{i=1}^{n} (x_i - \mu)^2 \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2
\end{align*}
\]
Set \( \frac{\partial f}{\partial \mu} = 0 \) \quad \frac{\partial f}{\partial \sigma^2} = 0

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

MLE estimators are:

\[
\hat{\mu} = \overline{X} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{X})^2
\]

Recall that \( \hat{\sigma} \) is a biased estimator for \( \sigma \).
Example

\[ X = \text{ waiting time for, say, a bus, uniformly distributed on } [0, \theta] \]

\[ f(x) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases} \]

Random samples: \( X_1, \ldots, X_n \)

\[ f(x_1, \ldots, x_n; \theta) = \begin{cases} \frac{1}{\theta^n}, & 0 \leq x_i \leq \theta \text{ for all } i = 1 \ldots n \\ 0, & \text{otherwise} \end{cases} \]

What is the \( \theta \) that maximizes the likelihood function \( f(x_1, \ldots, x_n; \theta) \)?

Assume \( \theta \) is a fixed variable, there is only one variable \( \theta \), use calculus? \( \frac{\partial}{\partial \theta} \)

Set \( \frac{\partial f}{\partial \theta} = 0 \) then solve for \( \theta \)?

Not this time, the likelihood function is not differentiable, standard calculus is not (directly) applicable.

But the maximizer is easy to get if you think about the likelihood function:
\[ f(x_1, ..., x_n; \theta) = \begin{cases} 1 & \text{if } \theta \leq \max(x_1, ..., x_n) \\ 0 & \text{if } \theta > \max(x_1, ..., x_n) \end{cases} \]

Note:

\[ \hat{\theta} = \max(x_1, ..., x_n) \]

Once again, the MLE estimator is biased.