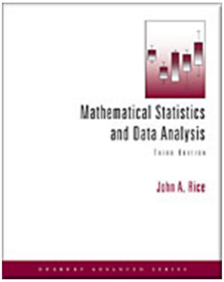
ST122:
Probability and Statistics II

Waleed A. Yousef, Ph.D.,

Human Computer Interaction Lab.,
Computer Science Department,
Faculty of Computers and Information,
Helwan University,
Egypt.

March 24, 2019

Lectures follow Rice, "*Mathematical Statistics and Data Analysis*", 3rd edition, Duxbury:



ISBN 0-534-39942-8

Course Objectives

- Developing rigorous treatment.
- Building intuition and insight.
- Linking to real life problems.
- Coding and scientific computing.

Contents

Introduction: Statistical Inference in a Nutshell

Point estimate - different estimators - assessing estimators - large sample theory
Hypothesis testing.
Interval estimation.
Bayesian approach vs. Frequentist approach

Chapter 6

Distributions Derived from the Normal Distribution

6.1 Introduction

This Chapter discusses 3 probability distributions that frequently occur in Statistics: χ^2, t, and F Distributions.

Remember that if $V \sim Gamma(\alpha, \lambda)$, then

$$
\begin{aligned}
f(v) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} v^{\alpha-1} e^{-\lambda v}, \; v \geq 0, \\
M(t) &= (1 - t/\lambda)^{-\alpha}, \\
E[V] &= \alpha/\lambda, \\
\mathrm{Var}[V] &= \alpha/\lambda^2.
\end{aligned}
$$

And if $V_1, \ldots, V_n$ are i.i.d $Gamma(\alpha, \lambda)$, then

$$
\begin{aligned}
M_{\Sigma_i V_i}(t) &= (1 - t/\lambda)^{-n\alpha}, \\
\Sigma_i V_i &\sim Gamma(n\alpha, \lambda).
\end{aligned}
$$

6.2 χ^2, t, and F Distributions

Definition 1 *If $Z \sim N(0,1)$, then $U = Z^2$ is called chi-square distribution with 1 degree of freedom; i.e., $U \sim \chi_1^2$. It is easy to show that (see Lec. notes Ch. 2):*

$$f_U(u) = \frac{1}{\sqrt{2\pi}} u^{-1/2} e^{-u^2/2}.$$

Notice that:

$$\chi_1^2 \equiv Gamma\left(\frac{1}{2}, \frac{1}{2}\right),$$

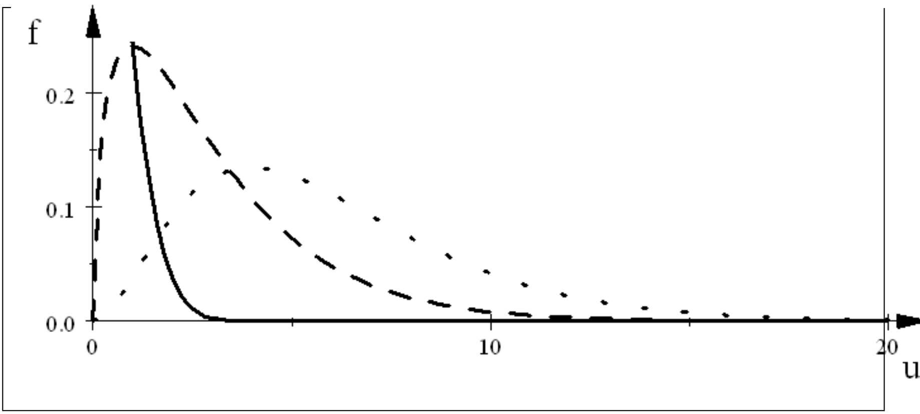
Also:

$$\begin{aligned} X &\sim N\left(\mu, \sigma^2\right), \\ \frac{X-\mu}{\sigma} &\sim N(0,1), \\ \left(\frac{X-\mu}{\sigma}\right)^2 &\sim \chi_1^2. \end{aligned}$$

Definition 2 *If $U_1, \ldots, U_n$ are i.i.d χ_1^2 r.v. then $V = \sum_i U_i$ is called chi-squre distribution with n degrees of freedom; i.e., $V \sim \chi_n^2$.*

Notice that $U_i \sim Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$, then

$$V \sim Gamma(n/2, 1/2),$$
$$f_V(v) = \frac{1}{2^{n/2}\Gamma(n/2)} v^{n/2-1} e^{-v/2},$$
$$E[V] = n, \ \mathrm{Var}[V] = 2n.$$



solid: $n = 1$, dashed: $n = 3$, dotted: $n = 6$

Suppose that U and V are indep, and

$$W = U + V.$$

If $U \sim \chi^2_m$, $V \sim \chi^2_n$ then (obviously)

$$W = \chi^2_m + \chi^2_n = \chi^2_{m+n},$$

Also, if $W \sim \chi^2_k$ and $V \sim \chi^2_n$ then

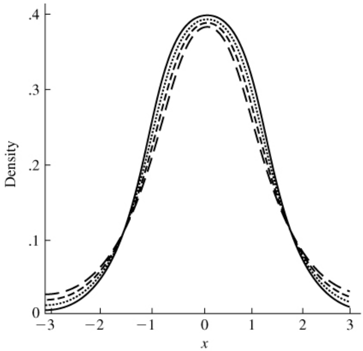
$$\begin{aligned}
\chi^2_k &= U + \chi^2_n \\
M_{\chi^2_k} &= M_U M_{\chi^2_n}, \\
M_U &= \frac{M_{\chi^2_k}}{M_{\chi^2_n}} \\
&= \frac{(1-2t)^{-k/2}}{(1-2t)^{-n/2}} = (1-2t)^{-(k-n)/2} \\
U &\sim \chi^2_{(k-n)}.
\end{aligned}$$

Definition 3 (Student's t Distribution) :
If $Z \sim N(0,1)$, $U \sim \chi^2_n$, and Z, U are indep. then $T = Z/\sqrt{U/n}$ is called t distribution with n degrees of freedom; i.e., $T \sim t_n$. (prove that:)

$$f_T(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2},$$

$$E[T] = 0, \ n \geq 2,$$

$$\mathrm{Var}[T] = \frac{n}{n-2}, \ n \geq 3.$$



- The smaller n the thicker tail.
- The figure shows $t_5, t_{10}, t_{30} (\approx N(0,1))$
- $t_1 \equiv Cauchy(0,1)$.

Definition 4 (Snedecor's F Distribution) *: Let $U \sim \chi^2_m$ and $V \sim \chi^2_n$, and U, V are indep. Then, $W = (U/m)/(V/n)$ is called F distribution with m, n degrees of freedom; i.e., $W \sim F_{m,n}$. (prove that:)*

$$f_W(w) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}} w^{\frac{m}{2}-1}\left(1+\frac{m}{n}w\right)^{-\frac{(m+n)}{2}},$$

$$E[W] = n/(n-2),\ n \geq 3.$$

$$\mathrm{Var}[W] = 2\left(\frac{n}{n-2}\right)^2 \frac{(m+n-2)}{m(n-2)},\ n \geq 5.$$

It is obvious that if $U \sim t_n$, then $U^2 \sim F_{1,n}$.

Also, if $U \sim F_{n,m}$ then $U^{-1} \sim F_{m,n}$.

Summary (with terse notation):

$$\begin{aligned}
N(0,1)^2 &\sim \chi_1^2,\\
\sum_{i=1}^{n} N(0,1)^2 &\sim \chi_n^2,\\
\chi_m^2 + \chi_n^2 &\sim \chi_{m+n}^2,\\
N(0,1)/\sqrt{\chi_n^2/n} &\sim t_n,\\
\left(\chi_m^2/m\right)/\left(\chi_n^2/n\right) &\sim F_{m,n},\\
t_n^2 &\sim F_{1,n}..
\end{aligned}$$

Example 5 *If X_1, X_2, X_3 are iid $N(0,1)$, what is the dist. of*

$$\frac{X_1}{\sqrt{\left(X_1^2 + X_2^2 + X_3^2\right)/3}}$$

6.3 Sample Mean, Sample Variance, and Sampling from Normal Distribution

6.3.1 Basic Concepts of Random Samples

Definition 6 *The r.v. $X_1, \ldots, X_n$ are called a* ***random sample of size*** n ***from the population*** F *if $X_1, \ldots, X_n$ are i.i.d from F; and hence:*
$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \prod_i f(x_i)$.

$$\begin{array}{llllll} & & X_1 & X_2 & \ldots & X_n \\ F & \xrightarrow{Sample_1} & x_1, & x_2, & \ldots & x_n \\ F & \xrightarrow{Sample_2} & x_1, & x_2, & \ldots & x_n \\ & \vdots & & & & \end{array}$$

We focus in our study on infinite populations; Ch. 7 is about finite populations.

Definition 7 *Let $X_1, \ldots, X_n$ be a random sample of size n, and $T(x_1, \ldots, x_n)$ be a real- (or vector-) valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the r.v.*
$Y = T(X_1, \ldots, X_n)$ is called a statistic.

Definition 8 *The sample mean, sample variance, and sample standard deviations are statistics defined as:*

$$\overline{X} = \frac{1}{n}\sum_i X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

$$S = \sqrt{S^2},$$

Observed values will be denoted by $\overline{x}$, s^2, and s.

$$\begin{array}{llllllll}
 & & X_1 & X_2 & \ldots & X_n & & \overline{X} = \frac{1}{n}\sum_i X_i \\
F & \xrightarrow{Sample_1} & x_1, & x_2, & \ldots & x_n & & \overline{x} = \frac{1}{n}\sum_i x_i \\
F & \xrightarrow{Sample_2} & x_1, & x_2, & \ldots & x_n & & \overline{x} = \frac{1}{n}\sum_i x_i \\
 & \vdots & & & & & &
\end{array}$$

Lemma 9 *For any numbers $x_1, \ldots, x_n$:*

$$\min_a \sum_i (x_i - a)^2 = \sum_i \left(x_i - \overline{x}\right)^2,$$

$$\sum_i \left(x_i - \overline{x}\right)^2 = \sum_i x_i^2 - n\overline{x}^2.$$

Proof. : is identical to $\arg\min_c E(Y-c)^2 = E[Y]$.

$$\begin{aligned}
\sum_i (x_i - a)^2 &= \sum_i \left(\left(x_i - \overline{x}\right) + \left(\overline{x} - a\right)\right)^2 \\
&= \sum_i \left(x_i - \overline{x}\right)^2 + \sum_i \left(\overline{x} - a\right)^2 \\
&\quad + 2\sum_i \left(x_i - \overline{x}\right)\left(\overline{x} - a\right) \qquad \left(\textstyle\sum_i x_i = n\overline{x}\right) \\
&= \sum_i \left(x_i - \overline{x}\right)^2 + \sum_i \left(\overline{x} - a\right)^2,
\end{aligned}$$

which is minimized by choosing $a = \overline{x}$.

$$\sum_i (x_i - a)^2 = \sum_i \left(x_i - \overline{x}\right)^2 + \sum_i \left(\overline{x} - a\right)^2$$

$$\sum_i \left(x_i - \overline{x}\right)^2 = \sum_i x_i^2 - n\overline{x}^2. \qquad \left(a \overset{set}{=} 0\right)$$

Notice that: both forms are $O(n)$; however this form requires only one for loop for execution! ■

HW: Write a computer program, and find its complexity (where a step is a multiplication), for calculating

$$S_1 = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j,$$

$$S_2 = \sum_{i=1}^{n} \sum_{j \neq i} x_i x_j.$$

Can you do a mathematical trick to reduce their complexities to $O(n)$. !!!

Theorem 10 (Distribution-Free Properties) :

1. $E\left[\overline{X}\right]=\mu,$

2. $\mathrm{Var}\left[\overline{X}\right]=\sigma^2/n,$

3. $E\left[S^2\right]=\sigma^2.$

Proof. 1 and 2 are proven before. For 3,

$$\begin{aligned}
E\left[S^2\right] &= E\left[\frac{1}{n-1}\sum_i\left(X_i-\overline{X}\right)^2\right]\\
&= \frac{1}{n-1}E\left[\sum_i X_i^2-n\overline{X}^2\right]\\
&= \frac{1}{n-1}\left(\sum_i E\left[X_i^2\right]-nE\left[\overline{X}^2\right]\right)\\
&= \frac{1}{n-1}\left(n\left(\sigma^2+\mu^2\right)-n\left(\frac{\sigma^2}{n}+\mu^2\right)\right)=\sigma^2,
\end{aligned}$$

which is **unbiased estimator** for σ^2. ■

Lemma 11 *Let $X_1, \ldots, X_n$ be a r.s. from a population with mgf $M(t)$, then*

$$M_{\overline{X}}(t) = [M(t/n)]^n.$$

Proof. done before in CLT (just 2 lines). ■

Example 12 *Let $X_1, \ldots, X_n$ be a r.s. from $N\left(\mu, \sigma^2\right)$, then*

$$M(t) = \exp\left(\mu t + \sigma^2 t^2/2\right),$$

$$M_{\overline{X}}(t) = \left[\exp\left(\mu \frac{t}{n} + \sigma^2 \left(\frac{t}{n}\right)^2 /2\right)\right]^n,$$

$$= \exp\left(\mu t + \frac{\sigma^2}{n} t^2/2\right),$$

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

We know that $E\left[\overline{X}\right] = \mu$ and $\mathrm{Var}\left[\overline{X}\right] = \sigma^2/n$. But what is new is that $\overline{X}$ is itself Normal. ***We could have found it by transformation:*** *$Z = X_1 + X_2$. If $X_i \sim Cauchy(0,1)$, prove that $\overline{X} \sim Cauchy(0,1)$ as well!!*

6.3.2 Sampling from the Normal Distribution

Theorem 13 *Let $X_1, \ldots, X_n$ be r.s. form $N\left(\mu, \sigma^2\right)$*

1. $\overline{X} \sim N\left(\mu, \sigma^2/n\right)$,
2. $\overline{X}$ *and* $\left(X_2 - \overline{X}, \ldots, X_n - \overline{X}\right)$ *are indep,*
3. $\overline{X}$ *and* S^2 *are indep,*
4. $(n-1)\, S^2/\sigma^2 \sim \chi^2_{n-1}$.

Intuition before proof:
Meaning of $\overline{X}$ and $\left(X_2 - \overline{X}, \ldots, X_n - \overline{X}\right)$ are indep?

Suppose $X_i \sim Bernouli\,(1/2)$, and we get a sample where $\overline{X}_{10} = 1$. Obviously, $X_i = 1$.

Aside from normality, observe that

$$\sum_i \left(X_i - \overline{X}\right) = 0,$$

which means we have only $(n-1)$ differences:

$$\left(X_1 - \overline{X}\right) = -\sum_{i=2}^{n} \left(X_i - \overline{X}\right),$$

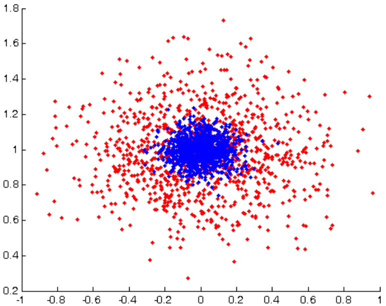
$$\begin{aligned}
S^2 &= \frac{1}{(n-1)} \sum_i \left(X_i - \overline{X}\right)^2 \\
&= \frac{1}{(n-1)} \left[\left(X_1 - \overline{X}\right)^2 + \sum_{i=2}^{n} \left(X_i - \overline{X}\right)^2\right] \\
&= \frac{1}{(n-1)} \left[\left(\sum_{i=2} \left(X_i - \overline{X}\right)\right)^2 + \sum_{i=2}^{n} \left(X_i - \overline{X}\right)^2\right]
\end{aligned}$$

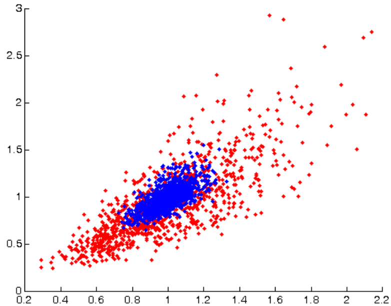
Matlab Code 6.1:

```
figure; hold on;
%Change 'Normal' to 'Exp'

x=random('Normal', 0, 1, 1000, 10);
xbar=mean(x, 2);
s=std(x, 0, 2);
plot(xbar, s, '.r')

x=random('Normal', 0, 1, 1000, 100);
xbar=mean(x, 2);
s=std(x, 0, 2);
plot(xbar, s, '.b')
```





Proof. the mgf is given by

$$
\begin{aligned}
&= M(s, t_2, \ldots, t_n) \\
&= E\left[\exp\left(s\overline{X} + t_2\left(X_2 - \overline{X}\right) + \ldots + t_n\left(X_n - \overline{x}\right)\right)\right] \\
&= E\left[\exp\left(\sum_{i=1}^{n} \frac{s}{n} X_i + \sum_{i=2}^{n} t_i\left(X_i - \overline{X}\right)\right)\right] \\
&= E\left[\exp\left(\sum_{i=1}^{n} \left(\frac{s}{n} + \left(t_i - \overline{t}\right)\right) X_i\right)\right] && (t_1 = 0) \\
&= E\left[\exp\left(\sum_{i=1}^{n} a_i X_i\right)\right] && \left(a_i = \tfrac{s}{n} + \left(t_i - \overline{t}\right)\right) \\
&= \prod_i M_{X_i}(a_i) \\
&= \prod_i \exp\left(\mu a_i + \frac{\sigma^2}{2} a_i^2\right) \\
&= \exp\left[\mu \sum_i a_i + \frac{\sigma^2}{2} \sum_i a_i^2\right] \\
&= \exp\left[\mu s + \frac{\sigma^2}{2}\left(\frac{s^2}{n} + \Sigma_i \left(t_i - \overline{t}\right)^2\right)\right] \\
&= \exp\left(\mu s + \frac{\sigma^2}{2n} s^2\right) \exp\left(\frac{\sigma^2}{2} \sum_i \left(t_i - \overline{t}\right)^2\right),
\end{aligned}
$$

the two factors are the mgf of $\overline{X}$ and $\left(X_2-\overline{X},\ldots,X_n-\overline{X}\right)$. Hence they are independent and since $S=S\left(X_2-\overline{X},\ldots,X_n-\overline{X}\right)$: $\overline{X}$ and S are independent.

Now

$$
\begin{aligned}
\sum_i\left(\frac{X_i-\mu}{\sigma}\right)^2 &= \frac{1}{\sigma^2}\sum_i\left[\left(X_i-\overline{X}\right)+\left(\overline{X}-\mu\right)\right]^2\\
&= \frac{1}{\sigma^2}\sum_i\left(X_i-\overline{X}\right)^2+\frac{1}{\sigma^2}\sum_i\left(\overline{X}-\mu\right)^2\\
&= \frac{1}{\sigma^2}\sum_i\left(X_i-\overline{X}\right)^2+\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2\\
W &= U+V && (U, V\text{ indep.})\\
\chi_n^2 &= U+\chi_1^2\\
U &\sim \chi_{n-1}^2. && (n-1\text{ df})
\end{aligned}
$$

■

Lemma 14

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Proof.

$$\begin{aligned}
\frac{\overline{X}-\mu}{S/\sqrt{n}} &= \frac{\left(\overline{X}-\mu\right)/\left(\sigma/\sqrt{n}\right)}{\left(S/\sqrt{n}\right)/\left(\sigma/\sqrt{n}\right)} \\
&= \frac{\left(\overline{X}-\mu\right)/\left(\sigma/\sqrt{n}\right)}{S/\sigma} \\
&= \frac{\left(\overline{X}-\mu\right)/\left(\sigma/\sqrt{n}\right)}{\sqrt{\left((n-1)\,S^2/\sigma^2\right)/(n-1)}} \\
&= \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} = t_{n-1},
\end{aligned}$$

used for inference about μ when σ is unkown.

$$\frac{\overline{X}-\mu}{\sigma} \sim N(0,1)$$

used for inference about μ when σ is known. ∎

Lemma 15 *If $X \sim N\left(\mu_X, \sigma_X\right)$, $Y \sim N\left(\mu_Y, \sigma_Y\right)$, and we have two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$*

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{m-1,n-1}.$$

Proof.

$$\begin{aligned}
\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} &= \frac{\left((m-1)\,S_X^2/\sigma_X^2\right)/(m-1)}{\left((n-1)\,S_Y^2/\sigma_Y^2\right)/(n-1)} \\
&= \frac{\chi^2_{m-1}/(m-1)}{\chi^2_{n-1}/(n-1)} && \text{(Indep.)} \\
&= F_{m-1,n-1},
\end{aligned}$$

used for inference about σ_X^2/σ_Y^2. ■

Chapter 8

Estimation of Parameters and Fitting of Probability Distributions

8.1 Introduction: Estimation in a Nutshell

- Distributions depend on some population parameters; e.g., $N(\mu,\sigma^2)$, $Exp(\lambda)$, etc. Generally, we should write (e.g.,):

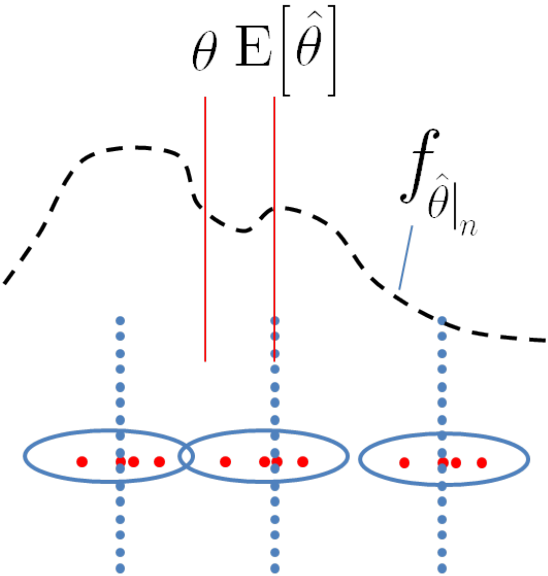
$$f_X(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-1}{2}(x-\mu)^2/\sigma^2\right]$$

- Obtaining data (values of a random sample) allows "estimating" these parameters.

Definition 16 *A* point estimator *is any function* $W(X_1,\ldots,X_n)$ *of a sample; i.e., any statistic is a point estimator.*

- We can choose, e.g., $\widehat{\sigma}^2 = \frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2$ to be an estimator for σ^2.

- $\frac{1}{n}\sum_i (x_i - \overline{x}_i)^2$ is an estimate (realization).



- How to estimate θ "well" ($\widehat{\theta}$)?
- What is $f_{\widehat{\theta}}$ (**sampling distribution**)?
- What is $\mathrm{E}\left[\widehat{\theta}\right], \mathrm{SD}\left[\widehat{\theta}\right]$ (**standard error**),...?
- How to estimate $\tau(\theta)$, e.g.:
  - σ^2, the variance, for $N\left(\mu, \sigma^2\right)$.
  - $\alpha\lambda$, the mean, for $Gamma(\alpha, \lambda)$.

How to decide F_X before estimation?

- From the physics of the problem. E.g., given number of calls in time units, the distribution is known to be $Poisson(\lambda)$.

- Assumption; you need to validate it latter.

Why do we estimate parameters?

- Understanding (interpretation).

- Prediction.

- Simulation and data generation.

How do we choose estimators?

8.2 The Method of Moments

We estimate k^{th} moment by **sample moment**

$$\mu_k = \mathrm{E}\left[X^k\right]$$
$$\widehat{\mu}_k = \frac{1}{n}\sum_i X_i^k.$$

Then for population parameters θ_i, we have

$$\begin{aligned}\mu_1 &= \mu_1(\theta_1,\ldots,\theta_r),\\ &\vdots\\ \mu_r &= \mu_r(\theta_1,\ldots,\theta_r).\end{aligned}$$

We solve

$$\begin{aligned}\theta_1 &= \theta_1\left(\mu_1,\ldots,\mu_r\right),\\ &\vdots\\ \theta_r &= \theta_r\left(\mu_1,\ldots,\mu_r\right).\end{aligned}$$

And

$$\begin{aligned}\widehat{\theta}_1 &= \widehat{\theta}_1\left(\widehat{\mu}_1,\ldots,\widehat{\mu}_r\right),\\ &\vdots\\ \widehat{\theta}_r &= \widehat{\theta}_r\left(\widehat{\mu}_1,\ldots,\widehat{\mu}_r\right).\end{aligned}$$

Motivation behind method of moments

$$\widehat{\mu}_k \xrightarrow{p} \mu_k.$$

Definition 17 *An estimator $\widehat{\theta} = \widehat{\theta}(n)$, which estimates θ, from a sample of size n is said to be consistent in probability if*

$$\widehat{\theta} \xrightarrow{p} \theta.$$

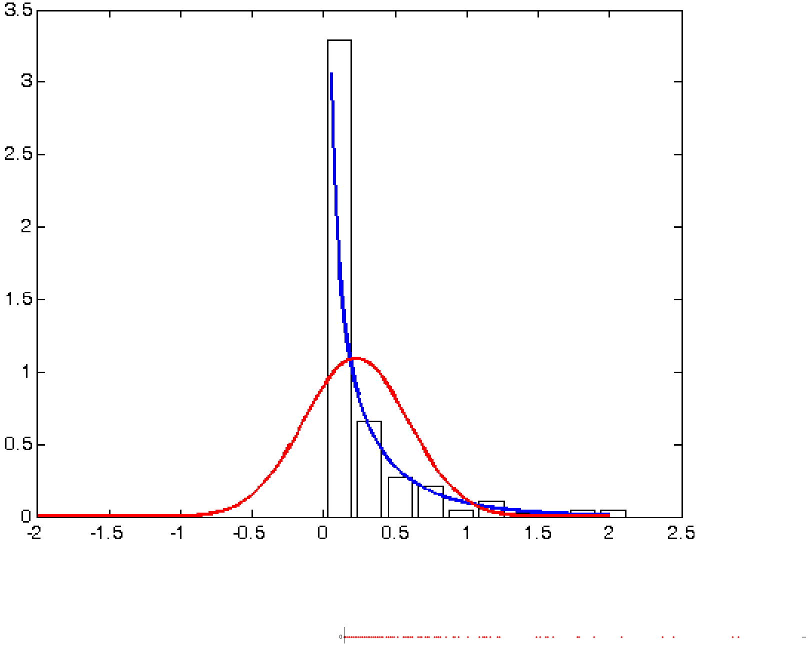
Example 18 *$N(\mu, \sigma^2)$, and the mean and variance of any other distribution:*

$$\begin{aligned}
\widehat{\mu}_1 &= \frac{1}{n}\sum_i X_i = \overline{X}, \\
\widehat{\mu}_2 &= \frac{1}{n}\sum_i X_i^2, \\
\mu_1 &= \mathrm{E}[X] = \mu, \\
\mu_2 &= \mathrm{E}\left[X^2\right] = \mu^2 + \sigma^2, \\
\mu &= \mu_1, \\
\sigma^2 &= \mu_2 - \mu_1^2, \\
\widehat{\mu} &= \widehat{\mu}_1 = \overline{X}, \\
\widehat{\sigma}^2 &= \widehat{\mu}_2 - \widehat{\mu}_1^2 = \frac{1}{n}\sum_i X_i^2 - \overline{X}^2 \qquad (\widehat{\sigma^2}) \\
&= \frac{1}{n}\left(\sum_i X_i^2 - n\overline{X}^2\right) = \frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2 \\
&= \frac{n-1}{n} S^2, \\
\widehat{\mu} &\sim N\left(\mu, \sigma^2/n\right), \\
\frac{n\widehat{\sigma}^2}{\sigma^2} &\sim \chi^2_{n-1}.
\end{aligned}$$

Example 19 : *Analyzing real dataset for average amount of storms rainfall in Illinois.*

Let's draw data points and normalized histogram (divide by its area):

$$Area = \sum_i \Delta N_i$$
$$= \Delta \sum_i N_i = \Delta n.$$



From the mgf of Gamma we obtained

$$\mathrm{E}[X] = \mu_1 = \frac{\alpha}{\lambda},$$
$$\mathrm{E}\left[X^2\right] = \mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2},$$

Solve both equations for α and λ,

$$\begin{aligned}
\alpha &= \lambda\mu_1 \\
\mu_2 &= \frac{\lambda^2\mu_1^2 + \lambda\mu_1}{\lambda^2}, \\
&= \mu_1^2 + \mu_1/\lambda, \\
\lambda &= \frac{\mu_1}{\mu_2 - \mu_1^2}, \\
\alpha &= \frac{\mu_1^2}{\mu_2 - \mu_1^2}, \\
\widehat{\mu}_1 &= \frac{1}{n}\sum x_i = 0.2244, \\
\widehat{\mu}_2 &= \frac{1}{n}\sum x_i^2 = 0.1836, \\
\widehat{\lambda} &= 1.6842, \\
\widehat{\alpha} &= 0.3779
\end{aligned}$$

$$
\begin{aligned}
f(x) &= \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \\
&= 0.5178 x^{-0.6221} e^{-1.6842x}, \; x \geq 0
\end{aligned}
$$

What would happen have if we fit $N\left(\mu, \sigma^2\right)$?

Matlab Code 8.1:

```
x=[];
x=[x; csvread('illinois60.txt')];
x=[x; csvread('illinois61.txt')];
x=[x; csvread('illinois62.txt')];
x=[x; csvread('illinois63.txt')];
x=[x; csvread('illinois64.txt')];

n=length(x) %will be 227
plot(x, zeros(length(x)), '.r')
[N, xout]=hist(x);
bar(xout, N/(n*(xout(2)-xout(1))), 'w'
    ); %normalize
hold on;
```

```
mu1   = sum(x)/n                    %.2244
mu2   = sum(x.^2)/n                 %.1836
alpha= mu1^2/(mu2-mu1^2)            %.3779
lmda  = mu1/(mu2-mu1^2)             %1.6842

z=0.05:.01:2;
y1=(lmda^alpha)/gamma(alpha) * z.^(
    alpha-1) .* exp(-lmda*z);
plot(z, y1, 'b', 'LineWidth', 2);

z=-2:.01:2;
y2=1/(sqrt(2*pi*(mu2-mu1^2))) *exp(-(z
    -mu1).^2 / (2*(mu2-mu1^2)));
plot(z, y2, 'r', 'LineWidth', 2);
```

Example 20 ($Binomial(n,p)$)

$$
\begin{aligned}
\mu_1 &= np,\\
\mu_2 &= np(1-p) + \left(np\right)^2,\\
p &= \frac{\mu_1}{n},\\
\mu_2 &= \mu_1\left(1-\frac{\mu_1}{n}\right) + \mu_1^2\\
n &= \frac{\mu_1^2}{\mu_1 - \left(\mu_2 - \mu_1^2\right)}\\
p &= \frac{\mu_1 - \left(\mu_2 - \mu_1^2\right)}{\mu_1},\\
\widehat{n} &= \frac{\overline{X}^2}{\overline{X} - \frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2},\\
\widehat{p} &= \frac{\overline{X} - \frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2}{\overline{X}}.
\end{aligned}
$$

- Sometimes the estimate will be negative!!
- In general, method of moments is a good start.

Example 21 (Cov(X, Y)) :

$$\begin{aligned}
\sigma_X^2 &= E\left(X - \mu_X\right)^2 \\
&= E\left(X^2\right) - \mu_X^2 \\
&= \mu_{2X} - \mu_{1X}^2. \\
\mathrm{Cov}(X, Y) &= \mathrm{E}\left(X - \mu_X\right)\left(Y - \mu_Y\right) \\
&= \mathrm{E}[XY] - \mu_X \mu_Y \\
&= \mu_{11} - \mu_{1X}\mu_{1Y}
\end{aligned}$$

$$\begin{aligned}
\widehat{\sigma}_X^2 &= \frac{1}{n}\sum_i X_i^2 - \overline{X}^2 \\
&= \frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2. \\
\widehat{\sigma}_{XY} &= \frac{1}{n}\sum_i X_i Y_i - \overline{XY}. \\
&= \frac{1}{n}\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right).
\end{aligned}$$

Given $x_1, \ldots, x_n$ *and* $y_1, \ldots, y_m$, *what is* $\widehat{\sigma}_{XY}$*?*
What is right (x_i, y_i).

$$
\begin{aligned}
\mathrm{E}\left[X_i Y_i\right] &= \mathrm{Cov}\left(X, Y\right) + \mu_X \mu_Y \\
\mathrm{E}\left[\overline{X}\,\overline{Y}\right] &= \mathrm{Cov}\left(\overline{X}, \overline{Y}\right) + \mathrm{E}\left[\overline{X}\right]\mathrm{E}\left[\overline{Y}\right] \\
&= \mathrm{Cov}\left(\frac{1}{n}\sum_i X_i, \frac{1}{n}\sum_i Y_i\right) + \mu_X \mu_Y \\
&= \frac{1}{n^2}\sum_i \sum_j \mathrm{Cov}\left(X_i, Y_j\right) + \mu_X \mu_Y \\
&= \frac{1}{n}\mathrm{Cov}\left(X, Y\right) + \mu_X \mu_Y
\end{aligned}
$$

$$
\mathrm{E}\textstyle\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) =
$$

$$
\begin{aligned}
&= \mathrm{E}\left[\sum_i X_i Y_i - n\overline{X}\,\overline{Y}\right] \\
&= n\,\mathrm{E}\left[XY\right] - n\,\mathrm{E}\left[\overline{X}\,\overline{Y}\right]. \\
&= n\sigma_{XY} + n\mu_X\mu_Y - \sigma_{XY} - n\mu_X\mu_Y \\
&= (n-1)\,\sigma_{XY}.
\end{aligned}
$$

Therefore, $\frac{1}{n}\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$ is biased for σ_{XY}.

Another proof for $\mathrm{E}\left[\overline{XY}\right]$:

$$
\begin{aligned}
\mathrm{E}\left[\overline{XY}\right] &= \mathrm{E}\left[\left(\frac{1}{n}\sum_i X_i\right)\left(\frac{1}{n}\sum_i Y_i\right)\right] \\
&= \mathrm{E}\left[\frac{1}{n^2}\sum_i\sum_j X_i Y_j\right] \\
&= \frac{1}{n^2}\mathrm{E}\left[\sum_i X_i Y_i + \sum_{i\neq j}\sum X_i Y_j\right] \\
&= \frac{1}{n^2}\left(n\,\mathrm{E}\left[XY\right] + n\left(n-1\right)\mathrm{E}\left[X_i Y_j\right]\right) \\
&= \frac{1}{n}\left(\mathrm{E}\left[XY\right] + \left(n-1\right)\mathrm{E}\left[X_i Y_j\right]\right) \\
&= \frac{1}{n}\left(\mathrm{Cov}\left(X, Y\right) + \mu_X\mu_Y + \left(n-1\right)\mu_X\mu_Y\right) \\
&= \frac{1}{n}\mathrm{Cov}\left(X, Y\right) + \mu_X\mu_Y.
\end{aligned}
$$

8.3 The Method of Maximum Likelihood

Likelihood is a function of parameters:

$$
\begin{aligned}
lik(\theta) &= f_{X_1 \ldots X_n}(x_1, \ldots, x_n | \theta) \\
&= \prod_{i=1}^{n} f(x_i | \theta). \qquad \text{(i.i.d.)}
\end{aligned}
$$

- For given data $x_1, \ldots, x_n$, what is the value of θ that maximizes $lik(\theta)$.
- Remember Example 15, Page 19 in Lecture Notes.
- Much easier, in many cases, to deal with the **log likelihood**:

$$
l(\theta) = \sum_{i=1}^{n} \log f(x_i | \theta).
$$

Example 22 ($Poisson(\lambda)$)

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!},\ 0 \le x.$$

$$lik(\lambda) = p(x_1, \ldots, x_x) = \prod_{i=1}^{n} \left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right),$$

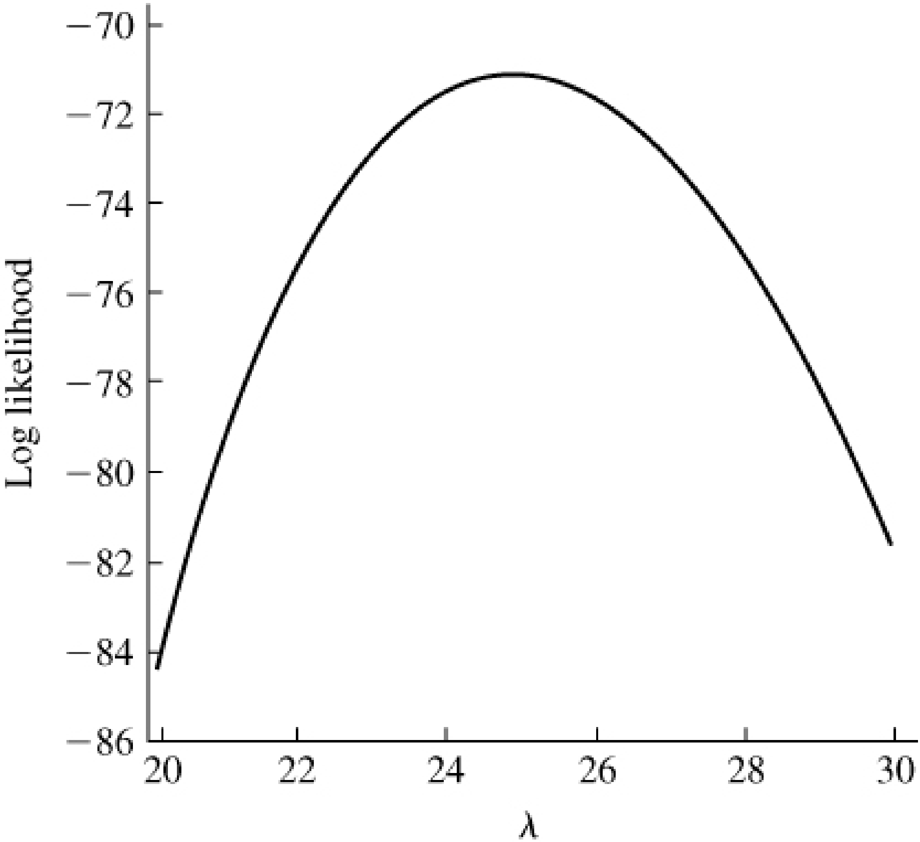
$$\begin{aligned} l(\lambda) &= \sum_{i=1}^{n} \log \left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right) \\ &= \sum_i \left[x_i \log \lambda - \lambda - \log(x_i!) \right] \\ &= \log(\lambda) \sum_i x_i - n\lambda - \sum_i \log(x_i!) \qquad (8.1) \\ l'(\lambda) &= \frac{\sum_i x_i}{\lambda} - n, \qquad (l'(\lambda) \overset{\text{set}}{=} 0) \\ \widehat{\lambda} &= \frac{1}{n} \sum x_i = \overline{X}, \qquad (\text{MoM}) \\ l''(\lambda) &= \frac{-\sum_i x_i}{\lambda^2} \le 0. \qquad (x_i \ge 0) \end{aligned}$$

Therefore, $\widehat{\lambda} = \overline{X}$ *is a point of local maxima; and*

$$\lim_{\lambda \to \infty} l(\lambda) = -\infty,$$

then, $\widehat{\lambda} = \overline{X}$ *is a global maximum as well.*

What does (8.1) mean for `asbestos` dataset?



Example 23 ($N\left(\mu,\sigma^2\right)$, both are unkown)

$$f\left(x|\mu,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]$$

$$\begin{aligned}
l\left(\mu,\sigma\right) &= \sum_{i=1}^{n}\log f\left(x_i|\mu,\sigma\right)\\
&= \sum_i\left[-\log\sigma-\log\sqrt{2\pi}-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]\\
&= -n\log\sigma-n\log\sqrt{2\pi}-\frac{1}{2\sigma^2}\sum_i\left(x_i-\mu\right)^2
\end{aligned}$$

$$\frac{\partial l}{\partial\mu} = \frac{1}{\sigma^2}\sum_i\left(x_i-\mu\right) \qquad \left(\frac{\partial l}{\partial\mu}\overset{\text{set}}{=}0\right)$$

$$0 = \sum_i x_i - n\widehat{\mu},$$

$$\widehat{\mu} = \frac{1}{n}\sum_i x_i = \overline{X}. \qquad \text{(MoM)}$$

$$\frac{\partial l}{\partial\sigma} = \frac{-n}{\sigma}+\frac{1}{\sigma^3}\sum_i\left(x_i-\mu\right)^2 \qquad \left(\frac{\partial l}{\partial\sigma}\overset{\text{set}}{=}0\right)$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i\left(x_i-\overline{X}\right)^2. \qquad \text{(MoM)}$$

To verify that $\left(\widehat{\mu},\widehat{\sigma}\right)$ is a point of global maxima through calculus we have to satisfy:

First: it is a point of local maxima

- $\frac{\partial l}{\partial \mu}|_{\widehat{\mu}} = \frac{\partial l}{\partial \sigma}|_{\widehat{\sigma}} = 0$ *(satisfied)*
- $\frac{\partial^2 l}{\partial \mu^2}|_{\widehat{\mu}} = 0$ *or* $\frac{\partial^2 l}{\partial \sigma^2}|_{\widehat{\sigma}} = 0$ *(satisfied)*
- $\begin{vmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \sigma} \\ \frac{\partial^2 l}{\partial \mu \partial \sigma} & \frac{\partial^2 l}{\partial \sigma^2} \end{vmatrix}_{\widehat{\mu},\widehat{\sigma}} > 0$ *(needs work).*

Second: there is no maximum at infinity (messy).

Instead, we can use a trick:

$$l(\mu, \sigma) = -n \log \sigma - n \log \sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2$$

is maximized for

$$\sum_i (x_i - \mu)^2 = \sum_i \left(x_i - \overline{X}\right)^2.$$

Then $l\left(\overline{X},\sigma\right)$ *is a function in single variable* σ,

$$\frac{\partial l}{\partial \sigma}=\frac{-n}{\sigma}+\frac{1}{\sigma^3}\sum_i\left(x_i-\overline{X}\right)^2, \qquad \left(\frac{\partial l}{\partial \sigma}\overset{set}{=}0\right)$$

$$\widehat{\sigma}^2=\frac{1}{n}\sum_i\left(x_i-\overline{X}\right)^2$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \sigma^2}&=\frac{n}{\sigma^2}-\frac{3}{\sigma^4}\sum_i\left(x_i-\overline{X}\right)^2\\&=\frac{n}{\sigma^2}\left(1-\frac{3}{n\sigma^2}\sum_i\left(x_i-\overline{X}\right)^2\right),\end{aligned}$$

$$\left.\frac{\partial^2 l}{\partial \sigma^2}\right|_{\widehat{\sigma}}=\frac{n}{\widehat{\sigma}^2}(1-3)<0,$$

which gives a local maximum for $l(\sigma)$. *And*

$$\lim_{\sigma\to\infty} l(\sigma)=-\infty.$$

Hence, $\widehat{\sigma}$ *attains a global maxima.*

Example 24 ($Gamma(\alpha, \lambda)$) :

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x},\ 0 \le x < \infty$$

$$l(\alpha, \lambda) = \sum_{i=1}^{n} \left(\alpha \log \lambda + (\alpha - 1) \log x_i - \lambda x_i - \log \Gamma(\alpha)\right)$$

$$= n\alpha \log \lambda + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \lambda \sum_{i=1}^{n} x_i$$

$$- n \log \Gamma(\alpha)$$

$$\frac{\partial l}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i \qquad \left(\frac{\partial l}{\partial \lambda} \overset{set}{=} 0\right)$$

$$0 = \frac{n\widehat{\alpha}}{\widehat{\lambda}} - \sum_{i=1}^{n} x_i$$

$$\widehat{\lambda} = \frac{\widehat{\alpha}}{\overline{X}}.$$

$$\frac{\partial l}{\partial \alpha} = n \log \lambda + \sum_{i=1}^{n} \log x_i - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \qquad \left(\frac{\partial l}{\partial \alpha} \overset{set}{=} 0\right)$$

$$0 = n \log \left(\frac{\widehat{\alpha}}{\overline{X}}\right) + \sum_{i=1}^{n} \log x_i - n \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})}$$

$$0 = n \log \widehat{\alpha} - n \log \overline{X} + \sum_{i=1}^{n} \log x_i - n \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})},$$

- no closed-form solution.
- solution has to be found either by numerical methods or bootstrap (later)
- more complications for checking the second derivatives.

Example 25

$$
\begin{aligned}
f(x) &= \frac{1}{\theta},\ 0 \le x \le \theta \\
&= \frac{1}{\theta} I_{(0 \le x \le \theta)} \\
l(\theta) &= \sum_{i=1}^{n} -\log\theta,\ x_i \le \theta \\
&= -n\log\theta,\ x^{(n)} \le \theta \\
\widehat{\theta} &= x^{(n)}.
\end{aligned}
$$

- *Intuitively, this is clear.*
- *We know $f_{X^{(n)}}(x)$ for $X \sim Uniform(0,\theta)$.*
- *Compare to MoM:*

$$
\begin{aligned}
\mu_1 &= \frac{\theta}{2} \\
\widehat{\theta} &= 2\overline{X}.
\end{aligned}
$$

Example 26 ($Multinomial(p_1, \ldots, p_m)$) **:** $\sum_{i=1}^{m} p_i = 1,\ \sum_{i=1}^{m} x_i = n$

$$f(x_1, \ldots, x_m) = \frac{n!}{x_1! \ldots x_m!} p_1^{x_1} \ldots p_m^{x_m}$$

$$l(p_1, \ldots, p_m) = \log n! - \sum_{i=1}^{m} \log x_i! + \sum_{i=1}^{m} x_i \log p_i$$

Using Lagrange multiplier

$$L(p_1, \ldots, p_m, \lambda) = \log n! - \sum_{i=1}^{m} \log x_i! + \sum_{i=1}^{m} x_i \log p_i + \lambda \left(\sum_{i=1}^{m} p_i - 1 \right)$$

$$\frac{\partial L}{\partial p_i} = \frac{x_i}{p_i} + \lambda \qquad \left(\frac{\partial L}{\partial p_i} \overset{set}{=} 0\right)$$

$$\widehat{p}_i = \frac{-x_i}{\lambda},$$

$$1 = \sum_i \widehat{p}_i = \sum_{i=1}^{m} \frac{-x_i}{\lambda} = \frac{-n}{\lambda},$$

$$\lambda = -n,$$

$$\widehat{p}_i = \frac{x_i}{n} \qquad \text{(intuitive)}$$

- *A special case is* $Binomial(n, p)$*, where* $m = 2$*,* $p_1 = p$*,* $x_1 = x$*,* n *is known*

$$\widehat{p} = \frac{x}{n},$$

- n *above is a parameter; the number of observations is 1, which is the vector* $(x_1, \ldots, x_m)$

For K observations: $(x_{11},\ldots x_{1m}),\ldots,(x_{K1},\ldots x_{Km})$.

$$f(x_1,\ldots,x_K)=\prod_{k=1}^{K}\frac{n!}{x_{k1}!\ldots x_{km}!}p_1^{x_{k1}}\ldots p_m^{x_{km}}$$

$$L\left(p_1,\ldots,p_m,\lambda\right)=\log(n!)^K-\sum_{i=1}^{m}\sum_{k=1}^{K}\log x_{ki}!$$

$$+\sum_{i=1}^{m}\sum_{k=1}^{K}x_{ki}\log p_i+\lambda\left(\sum_{i=1}^{m}p_i-1\right)$$

$$\frac{\partial L}{\partial p_i}=\frac{\sum_{k=1}^{K}x_{ki}}{p_i}+\lambda,$$

$$\widehat{p}_i=\frac{-\sum_{k=1}^{K}x_{ki}}{\lambda}$$

$$1=\frac{-\sum_{i=1}^{m}\sum_{k=1}^{K}x_{ki}}{\lambda}=\frac{-nK}{\lambda}$$

$$\widehat{p}_i=\frac{\sum_{k=1}^{K}x_{ki}}{nK}=\frac{\overline{X_i}}{n},$$

which for $Binomial\left(n,p\right)$ *will be*

$$\widehat{p}=\frac{\overline{X}}{n},$$

which is very intuitive.

8.3.1 Large Sample Theory for MLE

Reminder:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (\overline{X})$$

$$\widehat{\mu} \xrightarrow{p} \mathrm{E}[X] \qquad \text{(WLLN)}$$

$$\sqrt{n}\frac{\widehat{\mu}-\mu}{\sigma} \xrightarrow{d} N(0,1) \qquad \text{(CLT)}$$

$$\lim_{n\to\infty} \Pr\left(\sqrt{n}\frac{\widehat{\mu}-\mu}{\sigma} \le x\right) = \Pr(N(0,1) \le x)$$

$$\lim_{n\to\infty} \Pr\left(\sqrt{n}\left(\widehat{\mu}-\mu\right) \le \sigma x\right) = \Pr(\sigma N(0,1) \le \sigma x)$$

$$= \Pr\left(N\left(0,\sigma^2\right) \le \sigma x\right)$$

$$\sqrt{n}\left(\widehat{\mu}-\mu\right) \xrightarrow{d} N\left(0,\sigma^2\right) \qquad \text{(CLT')}$$

Definition 27 (Asymptotic Mean and Variance) *: For any statistic (or estimator) T_n, if*

$$k_n\frac{T_n-\mu}{\sigma} \xrightarrow{d} N(0,1), \qquad (k_n \text{ can be } \sqrt{n})$$

we call μ and σ^2 the asymptotic mean and variance (even if $\mathrm{E}[T_n] \ne \mu$ and $\mathrm{Var}[T_n] \ne \sigma^2$).

MoM:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (\overline{X})$$

$$\widehat{\mu} \xrightarrow{p} \mathrm{E}[X] \qquad \text{(WLLN)}$$

$$\sqrt{n}\frac{\widehat{\mu} - \mathrm{E}[X]}{\sqrt{\mathrm{Var}[X]}} \xrightarrow{d} N(0,1) \qquad \text{(CLT)}$$

$$\widehat{\mu}_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r, \qquad \text{(MoM)}$$

$$\widehat{\mu}_r \xrightarrow{p} \mathrm{E}\left[X^r\right] \qquad (\mathrm{E}\left[\widehat{\mu}_r\right] \overset{always}{=} \mathrm{E}[X^r])$$

$$\sqrt{n}\frac{\widehat{\mu}_r - \mathrm{E}[X^r]}{\sqrt{\mathrm{Var}[X^r]}} \xrightarrow{d} N(0,1)$$

Notice that:

- $\mathrm{E}\left[\widehat{\mu}_r\right] = \mathrm{E}[X^r]$ (always unbiased $\forall\, n$)
- the estimated parameters, e.g., $\widehat{\sigma}^2$, may be biased for finite n.

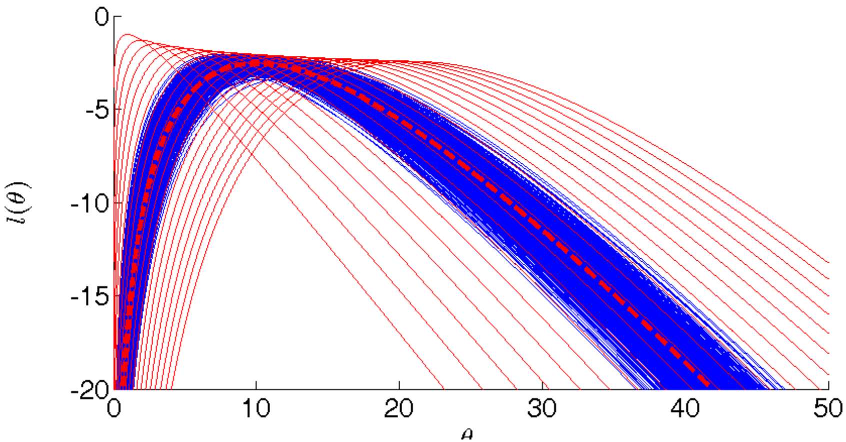
Some Intuition First:

$$l(\theta|X) = X\log\theta - \theta - \log(X!)$$

$$\mathrm{E}[l(\theta|X)] = \mathrm{E}[X]\log\theta - \theta - \mathrm{E}\left[\log(X!)\right]$$

$$l(\theta|X_1,\ldots,X_n) = \sum_i X_i \log\theta - n\theta - \sum_i \log(X_i!)$$

$$\frac{1}{n}l(\theta) \xrightarrow{p} \mathrm{E}\left[\log f(X|\theta)\right]$$



- **Take care:** $\mathrm{E}[X]$ above is $\mathrm{E}_{X|\theta_0}[X]$.

- Why curves are less than zero?

- We simulated 1000 curves, why few are there

Matlab Code 8.2:

```
theta0=10; theta = (0:.01:50)';
C = 1000;
ltheta = zeros(length(theta), C);

figure1 = figure; fs=20;
set(gcf, 'Units', 'inches');
haxes=axes('Parent',figure1,'YLim'
    ,[-20 0],'XLim',[0 50],'FontSize',
    fs);
xlabel('$\theta$','Interpreter','latex
    ','FontSize',fs, 'Units', '
    normalized');
ylabel('$l(\theta)$','Interpreter','
    latex','FontSize',fs, 'Units', '
    normalized');

hold all;
```

```
n=10;
for c=1:C
    x=random('Poisson',theta0,[n,1]);
    ltheta(:, c)=mean(x)*log(theta)-
      theta-sum(log(factorial(x)))/n;
    plot(theta, ltheta(:, c), 'b');
end;

n=1;
for c=1:C
    x=random('Poisson',theta0,[n,1]);
    ltheta(:, c)=x*log(theta)-theta-
      sum(log(factorial(x)));
    plot(theta, ltheta(:, c), 'r');
end;
plot(theta, mean(ltheta, 2), 'r--', '
  LineWidth', 4);
```

Theorem 28 *Under regularity conditions on f, the MLE estimator is consistent*

Semi-Proof. :Under regularity conditions

$$l(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta),$$

$$\frac{1}{n} l(\theta) \xrightarrow{p} \mathrm{E}\left[\log f(X|\theta)\right], \qquad (\mathrm{E}_{X|\theta_0})$$

$$\arg\max l(\theta) = \arg\max \frac{1}{n} l(\theta) \quad \text{(of course)}$$

$$\overset{I\ hope}{=} \arg\max \mathrm{E}\left[\log f(X|\theta)\right]$$

$$\frac{\partial}{\partial\theta} \mathrm{E}\left[\log f(X|\theta)\right] = \frac{\partial}{\partial\theta} \int \log f(x|\theta)\ f(x|\theta_0)\ dx$$

$$= \int \frac{\partial}{\partial\theta} \log f(x|\theta)\ f(x|\theta_0)\ dx$$

$$= \int \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)} f(x|\theta_0)\ dx$$

$$\left.\frac{\partial}{\partial\theta} \mathrm{E}\left[\log f(X|\theta)\right]\right|_{\theta_0} = \left.\int \frac{\partial}{\partial\theta} f(x|\theta)\, dx\right|_{\theta_0}$$

$$= \left.\frac{\partial}{\partial\theta} \int f(x|\theta)\, dx\right|_{\theta_0}$$

$$= \left.\frac{\partial}{\partial\theta} 1\right|_{\theta_0} = 0$$

■

Lemma 29 *Under regularity conditions:*

$$\mathrm{E}\left[\frac{\partial}{\partial\theta}\log f\left(X|\theta\right)\right]=0 \qquad (\mathrm{E}_{X|\theta})$$

$$\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f\left(X|\theta\right)\right)^2\right]=-\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f\left(X|\theta\right)\right],$$

which is called $I(\theta)$, the Fisher information (information number) of one observation.

- What is the meaning of “Information” here? Let’s see on the figure.

- Meaning of both equations.

Proof.

$$f(x|\theta)\frac{\partial}{\partial\theta}\log f(x|\theta) = f(x|\theta)\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)} = \frac{\partial}{\partial\theta}f(x|\theta)$$

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}(1) = \frac{\partial}{\partial\theta}\int f(x|\theta)\,dx = \int \frac{\partial}{\partial\theta}f(x|\theta)\,dx \\
&= \int f(x|\theta)\frac{\partial}{\partial\theta}\log f(x|\theta)\,dx \qquad\qquad (\mathrm{E}_{X|\theta_0}) \\
&= \frac{\partial}{\partial\theta}\int f(x|\theta)\frac{\partial}{\partial\theta}\log f(x|\theta)\,dx \\
&= \int \frac{\partial}{\partial\theta}f(x|\theta)\frac{\partial}{\partial\theta}\log f(x|\theta)\,dx + \\
&\int f(x|\theta)\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\,dx \\
&= \int f(x|\theta)\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2 dx + \\
&\int f(x|\theta)\left(\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right) dx \\
&= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2\right] + \mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right]
\end{aligned}$$

■

Theorem 30 *Let* $X_1, \ldots, X_n \overset{iid}{\sim} f(X|\theta)$, $\widehat{\theta}$ *is the MLE of* θ. *Then, under regularity conditions*

$$\sqrt{n}\frac{\widehat{\theta}-\theta}{1/\sqrt{I(\theta)}} \xrightarrow{d} N(0,1),$$

$$\sqrt{n}\frac{\tau\left(\widehat{\theta}\right)-\tau(\theta)}{1/\sqrt{I(\theta)}} \xrightarrow{d} N(0,1).$$

That is, any estimator $\tau\left(\widehat{\theta}\right)$ *(or* $\widehat{\theta}$*) is asymptotically unbiased for* $\tau(\theta)$ *(or* θ*) with asymptotic variance of* $1/I(\theta)$. *So, we have* $\xrightarrow{d} N(0,1)$ *in addition to* $\xrightarrow{p} \theta$.

Proof. Suppose that the true value of θ is θ_0

$$l(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$$l'(\theta) = l'(\theta_0) + (\theta-\theta_0)\, l''(\theta_0) + \cdots$$

$$l'\left(\widehat{\theta}\right) = l'(\theta_0) + \left(\widehat{\theta}-\theta_0\right) l''(\theta_0) + \cdots$$

$$\left(\widehat{\theta}-\theta_0\right) \approx -l'(\theta_0)/l''(\theta_0) \qquad \text{(MLE def.)}$$

$$\sqrt{n}\frac{\left(\widehat{\theta}-\theta_0\right)}{\sqrt{1/I(\theta_0)}} \approx \frac{\sqrt{n}\frac{1}{n}l'(\theta_0)/\sqrt{I(\theta_0)}}{\frac{-1}{n}l''(\theta_0)/I(\theta_0)}.$$

$$\frac{1}{n} l'(\theta_0) = \frac{1}{n} \sum_i \frac{\partial}{\partial\theta} \log f(X_i|\theta)\Big|_{\theta_0}$$

$$\mathrm{E}\left[\frac{\partial}{\partial\theta} \log f(X_i|\theta)\Big|_{\theta_0}\right] = 0 \qquad (\mathrm{E}_{X|\theta_0})$$

$$\mathrm{Var}\left[\frac{\partial}{\partial\theta} \log f(X_i|\theta)\Big|_{\theta_0}\right] = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta} \log f(X|\theta)\right)^2\Big|_{\theta_0}\right]$$

$$= I(\theta_0)$$

$$\sqrt{n}\frac{\frac{1}{n} l'(\theta_0) - 0}{\sqrt{I(\theta_0)}} \xrightarrow{d} N(0,1) \qquad (\mathrm{CLT})$$

$$\frac{-1}{n} l''(\theta_0) = \frac{-1}{n} \sum_i \frac{\partial^2}{\partial\theta^2} \log f(X_i|\theta)$$

$$\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(X|\theta)\Big|_{\theta_0}\right] = -I(\theta_0)$$

$$\frac{-1}{n} l''(\theta_0) \xrightarrow{p} I(\theta_0)$$

$$\frac{-1}{n} l''(\theta_0) / I(\theta_0) \xrightarrow{p} 1$$

$$\sqrt{n}\frac{\left(\widehat{\theta} - \theta_0\right)}{\sqrt{1/I(\theta_0)}} \xrightarrow{d} N(0,1).$$

■

Said differently

$$\sqrt{n}\frac{\widehat{\theta}-\theta_0}{\sqrt{1/I(\theta_0)}} \xrightarrow{d} N(0,1),$$

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{d} N(0,1/I(\theta_0)),$$

which means that the MLE $\widehat{\theta}$

- Asymptotically unbiased
- Asymptotic variance $= 1/I(\theta_0)$
- Asymptotically normally distributed.

Why variance decreases with $I(\theta_0)$?

$$I(\theta_0) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\bigg|_{\theta_0}\right]$$

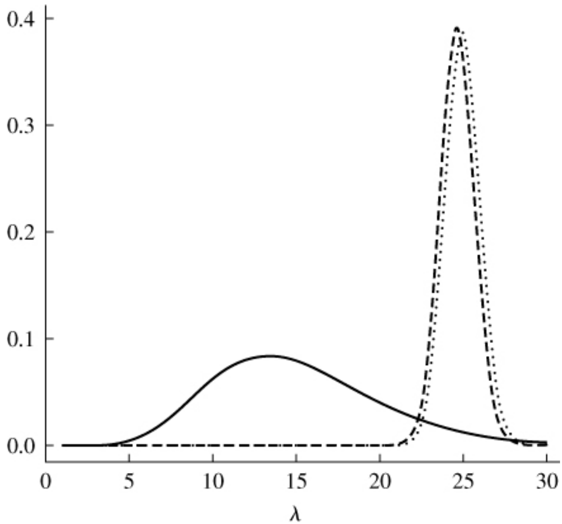
High $I(\theta_0)$ means very sharp curve at θ_0, which means very probable θ_0, which means less likely that the next dataset will not support that inference; and hence less variable the next estimator is.

8.4 The Bayesian Approach to Parameter Estimation

- We treat θ as r.v. with **subjective** prior knowledge f_Θ; as opposed to "Frequentist (or Classical) Approach"

- Data $\mathbf{x} = x_1, \ldots, x_n$ for $\mathbf{X} = X_1, \ldots, X_n$ modifies our belief and produces the posterior $f_{\Theta|\mathbf{X}}$?

- We estimate θ by many criteria; e.g.,:



1. Posterior Mode/Max. A Posteriori (MAP):

$$\widehat{\theta} = \arg\max_{\theta} f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$$

2. Posterior Mean:

$$\widehat{\theta} = \mathop{\mathrm{E}}_{\Theta}[\theta] = \int \theta\, f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta$$

3. Posterior loss function optimization:

$$\begin{aligned}\widehat{\theta} &= \arg\min_{\eta} \mathop{\mathrm{E}}_{\Theta}\left[L\left(\eta,\theta\right)\right]\\ &= \arg\min_{\eta} \int L\left(\eta,\theta\right) f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta\end{aligned}$$

General Framework:

$$\begin{aligned}f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) &= \frac{f_{\mathbf{X},\Theta}(\mathbf{x},\theta)}{f_{\mathbf{X}}(\mathbf{x})}\\ &= \frac{f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\, f_{\Theta}(\theta)}{\int f_{\mathbf{X},\Theta}(\mathbf{x},\theta)\;\, d\theta}\\ &= \frac{f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\, f_{\Theta}(\theta)}{\int f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\, f_{\Theta}(\theta)\;\, d\theta}\\ &= Const(\mathbf{x})\, f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\, f_{\Theta}(\theta)\end{aligned}$$

$$Posterior \propto Likelihood \times Prior.$$

Connection to MLE:

$$
\begin{aligned}
f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) &= Const(\mathbf{x})\, f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\, f_{\Theta}(\theta) \\
&\propto Likelihood \times Prior
\end{aligned}
$$

if we choose an uninformative prior $\Theta \sim U$ to let data speak for themselves:

$$
\begin{aligned}
f_{\Theta|X}(\theta|x) &= Const(x)\, f_{X|\Theta}(x|\theta) \\
&\propto Likelihood
\end{aligned}
$$

Then, if we choose MAP criterion

$$
\widehat{\theta} = \arg\max l(\theta), \qquad \text{(MLE)}
$$

Example 31 (Poisson) $\mathbf{X}$ *denotes* $X_1, \ldots, X_n$*:*

$$
\begin{aligned}
f_{\mathbf{X}|\Lambda} &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!},\ 0 \le x_i,\\
&= \frac{\lambda^{\sum_i x_i} e^{-n\lambda}}{\prod_i x_i!}\\
f_{\Lambda|\mathbf{X}} &= \frac{f_{\mathbf{X}|\Lambda}\left(\mathbf{x}|\lambda\right) f_{\Lambda}\left(\lambda\right)}{\int f_{\mathbf{X}|\Lambda}\left(\mathbf{x}|\lambda\right) f_{\Lambda}\left(\lambda\right)\, d\lambda}\\
&= \frac{\lambda^{\sum_i x_i} e^{-n\lambda} f_{\Lambda}\left(\lambda\right) / \prod_i x_i!}{\int \lambda^{\sum_i x_i} e^{-n\lambda} f_{\Lambda}\left(\lambda\right)\ / \prod_i x_i! d\lambda}\\
&= \frac{\lambda^{\sum_i x_i} e^{-n\lambda} \frac{1}{100}}{\int \lambda^{\sum_i x_i} e^{-n\lambda} \frac{1}{100}\, d\lambda} && \left(\Lambda \sim U\left(0, 100\right)\right)\\
&= \frac{\nu^{\alpha}}{\Gamma\left(\alpha\right)} \lambda^{\alpha-1} e^{-\nu\lambda} && \left(Gamma\left(\alpha, \nu\right)\right)\\
&\sim Gamma\left(S_n + 1, n\right)\\
\widehat{\lambda} &= \mathrm{E}\left[\Lambda\right] = \frac{S_n + 1}{n} = \overline{X} + \frac{1}{n} && \text{(Post. Mean)}\\
\frac{\partial f_{\Lambda|\mathbf{X}}}{\partial \lambda} &= \frac{\nu^{\alpha}}{\Gamma\left(\alpha\right)} \left(\left(\alpha - 1\right) \lambda^{\alpha-2} e^{-\nu\lambda} - \nu \lambda^{\alpha-1} e^{-\nu\lambda}\right)\\
\widehat{\lambda} &= \frac{\alpha - 1}{\nu} = \frac{S_n}{n} = \overline{X} && \text{(MAP}\equiv\text{MLE)}\\
\frac{S_n}{n} &= \frac{573}{23} = 24.9, \qquad \frac{S_n + 1}{n} = 25
\end{aligned}
$$

On the other hand, if we have the prior knowledge that Λ *has* $\mu = 15$ *and* $\sigma = 5$ *then, we can assume that* $\Lambda \sim Gamma(\alpha, \nu)$ *with*

$$\begin{aligned}
\mu &= \alpha/\nu, \\
\sigma^2 &= \alpha/\nu^2, \\
\nu &= \frac{\mu}{\sigma^2} = 0.6 << n && (n = 23) \\
\alpha &= \nu\mu = 9 << S_n, && (S_n = 573) \\
f_{\Lambda|\mathbf{X}} &= \frac{\lambda^{\sum_i x_i} e^{-n\lambda} f_\Lambda(\lambda)}{\int \lambda^{\sum_i x_i} e^{-n\lambda} f_\Lambda(\lambda)\ d\lambda} \\
&= \frac{\lambda^{\sum_i x_i} e^{-n\lambda} \frac{\nu^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\nu\lambda}}{\int \lambda^{\sum_i x_i} e^{-n\lambda} \frac{\nu^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\nu\lambda}\ d\lambda} \\
&= \frac{\lambda^{(S_n+\alpha-1)} e^{-(n+\nu)\lambda}}{\int \lambda^{(S_n+\alpha-1)} e^{-(n+\nu)\lambda} d\lambda} \\
&\sim Gamma(S_n + \alpha, n + \nu) \\
\widehat{\lambda} &= \frac{S_n + \alpha}{n + \nu} = \frac{573 + 9}{23 + .6} = 24.7 && \text{(Post. Mean)} \\
\widehat{\lambda} &= \frac{S_n + \alpha - 1}{n + \nu} = \frac{573 + 9 - 1}{23 + .6} = 24.6 && \text{(MAP)}
\end{aligned}$$

Example 32 ($Ber(p)$) *: n obs., then*

$$
\begin{aligned}
\mu_1 &= p, \\
\widehat{p} &= \overline{X} = \frac{\sum_i x_i}{n} = \frac{\#Heads}{n}, && \text{(MoM)} \\
p_X(x) &= p^x \left(1-p\right)^{1-x}, \ x = 0, 1 \\
l\left(p\right) &= \sum_i x_i \log p + \sum_i (1 - x_i) \log\left(1-p\right) \\
l'\left(p\right) &= \frac{\sum_i x_i}{p} - \frac{\sum_i (1 - x_i)}{1-p} && (l'\left(p\right) \overset{set}{=} 0) \\
\widehat{p} &= \overline{X} = \frac{\sum_i x_i}{n} = \frac{\#Heads}{n}. && \text{(MLE)}
\end{aligned}
$$

Now, if we get 5 heads in 5 trials $\widehat{p}$ will be 1 !!!!

Let's see the Bayesian approach.

$$f_{\mathbf{X}|P} = \prod_{i=1}^{n} p^{x_i} \left(1-p\right)^{1-x_i} = p^{\sum_i x_i} \left(1-p\right)^{\sum_i (1-x_i)}$$

$$f_P\left(p\right) = \frac{\Gamma\left(a+b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)} p^{a-1} \left(1-p\right)^{b-1} \quad (\sim Beta\left(a,b\right))$$

$$\begin{aligned} f_{P|\mathbf{X}} &= \frac{f_{\mathbf{X}|P}\left(\mathbf{x}|p\right) f_P\left(p\right)}{\int f_{\mathbf{X}|P}\left(\mathbf{x}|p\right) f_P\left(p\right)\, dp} \\ &\propto p^{a-1+S} \left(1-p\right)^{b-1+(n-S)} \\ &\sim Beta\left(a+S, b+n-S\right). \end{aligned}$$

$$\begin{aligned} \widehat{p} &= \frac{A-1}{A+B-2} = \frac{a+S-1}{a+b+n-2} && \text{(MAP)} \\ &= \frac{a+S-1}{2a+n-2} && \text{(Symmetric Prior)} \end{aligned}$$

$a = 1$: $U\left(0,1\right)$, $\widehat{p} = \frac{S}{n} \equiv MLE$.

$a = 2$: *not uniform but spread.* $\widehat{p} = \left(S+1\right)/\left(n+2\right)$.

- $S = n$: $\widehat{p} = \left(n+1\right)/\left(n+2\right) \to 1$.
- $S = n/2$: $\widehat{p} = 1/2$ *(of course).*

$a >>$: *insisting on fair coin,* $\widehat{p} \approx a/\left(2a\right) = \frac{1}{2}$

$$f_{P|\mathbf{X}} \sim Beta(a+S, b+n-S)$$

$$\begin{aligned}\widehat{p} &= \frac{A}{A+B}\\ &= \frac{a+S}{a+b+n} \qquad \text{(Posterior Mean)}\end{aligned}$$

8.4.1 Large Sample Theory of Bayesian Inference

X and **x** denote $X_1, \ldots, X_n$ and $x_1, \ldots, x_n$, respectively, to simplify notation.

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \propto f_{\Theta}(\theta)\, f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta),$$

which is dominated by $f_{\mathbf{X}|\Theta}$ as $n \to \infty$.

$$\begin{aligned}
f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) &\propto f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) && (\text{as } n \to \infty)\\
&= \exp\left[\log f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\right]\\
&= \exp\left[l(\theta)\right]\\
&= \exp[l\left(\widehat{\theta}\right) + \left(\theta - \widehat{\theta}\right) l'\left(\widehat{\theta}\right)\\
&+ \frac{1}{2}\left(\theta - \widehat{\theta}\right)^2 l''\left(\widehat{\theta}\right) + \cdots]\\
&\propto \exp\left[-\frac{1}{2}\frac{\left(\theta - \widehat{\theta}\right)^2}{-1/l''\left(\widehat{\theta}\right)}\right] && \left(l'\left(\widehat{\theta}\right) = 0\right)\\
&\sim N\left(\widehat{\theta}, -1/l''\left(\widehat{\theta}\right)\right).
\end{aligned}$$

Do not confuse it with the MLE asymptotic normality.

8.5 Assessing Estimators, Efficiency, and the Cramér-Rao Lower Bound

8.5.1 Mean Squared Error (MSE) Criterion

$$
\begin{aligned}
MSE\left(\widehat{\theta}\right) &= \underset{\mathbf{X}}{\mathrm{E}}\left[\left(\widehat{\theta}-\theta\right)^2\right] \\
&= \underset{\mathbf{X}}{\mathrm{Var}}\left[\widehat{\theta}\right] + \left(\underset{\mathbf{X}}{\mathrm{E}}\,\widehat{\theta}-\theta\right)^2 \\
&= Variance\left(\widehat{\theta}\right) + \left(Bias\left(\widehat{\theta}\right)\right)^2.
\end{aligned}
$$

- Since $MSE = MSE(\theta)$ no best estimator; e.g., $\widehat{\theta} = 12.3$ is the best when $\theta = 12.3$ but terrible otherwise.
- If $Bias\left(\widehat{\theta}\right) = 0$, $\widehat{\theta}$ is unbiased for θ.
- Tradeoff exists between Bias and Variance.
- A biased estimator may has lower MSE.

Example 33 ($\widehat{\sigma}^2$ vs. S^2 for $N(\mu, \sigma^2)$) :

$$
\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2, \\
S^2 &= \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2 \\
\mathrm{E}\left[S^2\right] &= \sigma^2 && \text{(unbiased)} \\
\mathrm{Var}\left[S^2\right] &= \frac{2\sigma^4}{n-1} && \text{(see Extra Materials)} \\
MSE\left(S^2\right) &= \frac{2\sigma^4}{n-1} + \left(\sigma^2 - \sigma^2\right)^2 = \frac{2\sigma^4}{n-1} \\
\mathrm{E}\left[\widehat{\sigma}^2\right] &= \frac{n-1}{n}\sigma^2 && \text{(biased)} \\
\mathrm{Var}\left[\widehat{\sigma}^2\right] &= \mathrm{Var}\left[\frac{n-1}{n}S^2\right] = \left(\frac{n-1}{n}\right)^2 \mathrm{Var}\left[S^2\right] \\
&= \left(\frac{n-1}{n}\right)^2 \left(\frac{2\sigma^4}{n-1}\right) = \frac{2(n-1)\sigma^4}{n^2} \\
MSE\left(\widehat{\sigma}^2\right) &= \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 \\
&= \frac{2n-1}{n^2}\sigma^4 < \frac{2\sigma^4}{n-1} \quad \forall \sigma, n.
\end{aligned}
$$

Remarks:

- Although S^2 is unbiased, $\widehat{\sigma}^2$ has less MSE.
- MSE, for scale parameter, may not be reasonable since $\sigma^2 > 0$.
- $\widehat{\theta}_1$ may be better than $\widehat{\theta}_2$ under some criterion and the other way around and another criterion.

Example 34 ($\widehat{p}$ of $Ber(p)$) :

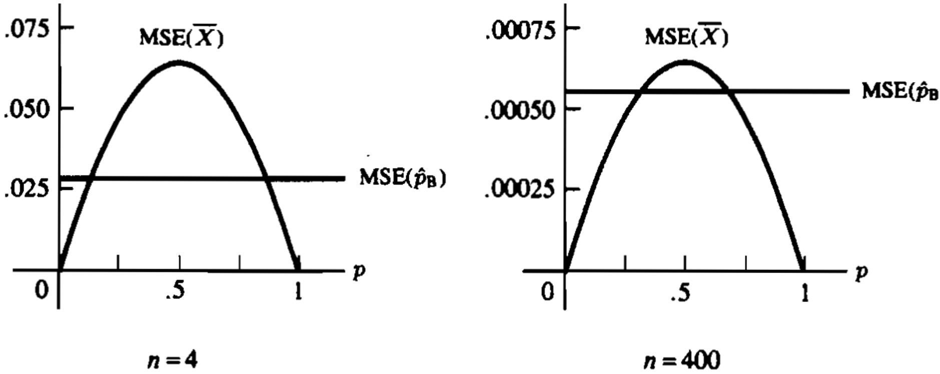
$$\widehat{p}_M = \overline{X} \qquad \text{(MLE)}$$
$$\mathrm{E}\left[\widehat{p}_M\right] = p$$
$$\mathrm{Var}\left[\widehat{p}_M\right] = \frac{1}{n}p(1-p)$$
$$MSE\left(\widehat{p}_M\right) = \frac{1}{n}p(1-p)$$
$$\widehat{p}_B = \frac{S+a}{a+b+n} \qquad \text{(Posterior Mean)}$$
$$\mathrm{E}\left[\widehat{p}_B\right] = \frac{np+a}{a+b+n}$$
$$\mathrm{Var}\left[\widehat{p}_B\right] = \frac{np(1-p)}{(a+b+n)^2}$$
$$MSE\left(\widehat{p}_B\right) = \frac{np(1-p)}{(a+b+n)^2} + \left(\frac{np+a}{a+b+n} - p\right)^2$$

Choosing $a = b = \sqrt{n}/2$ relaxes dependence on p:

$$\widehat{p}_B = \frac{S+\sqrt{n}/2}{n+\sqrt{n}},$$
$$MSE\left(\widehat{p}_B\right) = \frac{n}{4\left(n+\sqrt{n}\right)^2}.$$

$$MSE\left(\widehat{p}_M\right) = \frac{1}{n} p\left(1-p\right)$$

$$MSE\left(\widehat{p}_B\right) = \frac{n}{4\left(n+\sqrt{n}\right)^2}$$



- *For small n, $\widehat{p}_B$ is better unless p is on the boundary.*

- *For large n, $\widehat{p}_M$ is better unless p is in the middle.*

- *Having knowledge about the problem allows choosing the right estimator.*

8.5.2 Best Unbiased Estimator

Definition 35 (UMVUE) *: An estimator* $\widehat{\theta}^*$*, for* θ*, is a best unbiased estimator or uniform minimum variance unbiased estimator (UMVUE) if it satisfies* $\mathrm{E}\left[\widehat{\theta}^*\right] = \theta \ \forall\theta$ *and for any other estimator* $\widehat{\theta}$ *we have* $\mathrm{Var}\left[\widehat{\theta}^*\right] \le \mathrm{Var}\left[\widehat{\theta}\right]$.

Theorem 36 (Cramér-Rao Inequality) *: Let* $X_1, \ldots, X_n \overset{i.i.d}{\sim} f(x|\theta)$ *with regularity condition. The for any estimator* $T = T(X_1, \ldots, X_n) = T(\mathbf{X})$

$$\mathrm{Var}(T) \ge \frac{\left(\frac{d}{d\theta}\mathrm{E}[T]\right)^2}{nI(\theta)},$$

$$\mathrm{Var}(T) \ge \frac{1}{nI(\theta)}. \qquad \text{(if } T \text{ is unbiased)}$$

- For all estimators with particular bias: the higher the *information number* the lower the *lower bound.*

- An estimator *attains* (*attainment*) the lower bound is called *efficient.*

Proof. :Since $1 \le \rho = \text{Cov}(T,Z)/\sqrt{\text{Var}(T)\,\text{Var}(Z)}$

$$\text{Var}[T] \ge (\text{Cov}(T,Z))^2/\text{Var}(Z)$$

$$Z = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta)$$

$$\begin{aligned}
\text{Var}[Z] &= n\text{Var}\left[\frac{\partial}{\partial\theta} \log f(X_i|\theta)\right] \\
&= nI(\theta) && \text{(Proof of Th. 30)} \\
\sigma_{TZ} &= \text{E}(Z - \text{E}[Z])(T - \text{E}[T]) = \text{E}[T(Z - \text{E}[Z])] \\
&= \text{E}[ZT] && (\text{E}[Z] = 0) \\
&= \text{E}\left[T\frac{\partial}{\partial\theta} \log \prod_i f(X_i|\theta)\right] \\
&= \text{E}\left[T\frac{\partial}{\partial\theta} \log f(\mathbf{X}|\theta)\right] && (\mathbf{X} = X_1, \ldots, X_n) \\
&= \int T(\mathbf{x}) \frac{\frac{\partial}{\partial\theta} f(\mathbf{x}|\theta)}{f(\mathbf{x}|\theta)} f(\mathbf{x}|\theta)\, d\mathbf{x} \\
&= \frac{\partial}{\partial\theta} \int T(\mathbf{x}) f(\mathbf{x}|\theta)\, d\mathbf{x} \\
&= \frac{\partial}{\partial\theta} \underset{\mathbf{X}}{\text{E}}[T(\mathbf{X})]
\end{aligned}$$

■

Example 37 (Poisson) :

$$
\begin{aligned}
I(\lambda) &= \mathrm{E}\left[\left(\frac{\partial}{\partial\lambda}\log\frac{\lambda^X e^{-\lambda}}{X!}\right)^2\right]\\
&= \mathrm{E}\left[\left(\frac{\partial}{\partial\lambda}\left(X\log\lambda-\lambda-\log X!\right)\right)^2\right]\\
&= \mathrm{E}\left[\left(\frac{X}{\lambda}-1\right)^2\right]\\
&= -\mathrm{E}\left[\frac{\partial^2}{\partial\lambda^2}\log\frac{\lambda^X e^{-\lambda}}{X!}\right] && \text{(easier)}\\
&= -\mathrm{E}\left[\frac{-X}{\lambda^2}\right] = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda},\\
\mathrm{Var}[T] &\geq \frac{\left(\frac{\partial}{\partial\lambda}\mathrm{E}[T]\right)^2}{nI(\lambda)}\\
&= \frac{\lambda}{n} && \text{(for unbiased estimators)}\\
\widehat{\lambda} &= \overline{X} && \text{(MLE)}\\
\mathrm{E}\left[\widehat{\lambda}\right] &= \lambda && \text{(unbiased)}\\
\mathrm{Var}\left[\widehat{\lambda}\right] &= \mathrm{Var}\left[\overline{X}\right] = \frac{1}{n}\mathrm{Var}[X] = \frac{\lambda}{n}, && \text{(attainment)}
\end{aligned}
$$

Example 38 ($U(0,\theta)$) *:$f(x|\theta) = 1/\theta$, then*

$$
\begin{aligned}
I(\theta) &= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log(1/\theta)\right)^2\right]\\
&= \mathrm{E}\left[\left(-\frac{\partial}{\partial\theta}\log\theta\right)^2\right] = 1/\theta^2,\\
\mathrm{Var}\left[\widehat{\theta}\right] &\geq \frac{\left(\frac{\partial}{\partial\theta}\mathrm{E}\left[T\right]\right)^2}{nI(\theta)}\\
&= \frac{\theta^2}{n}, && \text{(for unbiased estimators)}\\
\widehat{\theta} &= 2\overline{X}, && \text{(MoM)}\\
\mathrm{E}\left[\widehat{\theta}\right] &= \theta && \text{(unbiased)}\\
\mathrm{Var}\left[\widehat{\theta}\right] &= \frac{4}{n}\mathrm{Var}\left[X\right] = \frac{4}{n}\frac{\theta^2}{12}\\
&= \frac{\theta^2}{3n} < \frac{\theta^2}{n}. && \text{(!!!where is the problem?)}
\end{aligned}
$$

The regularity condition assumes $(n=1)$*:*

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathrm{E}[T] &= \frac{\partial}{\partial\theta}\int Tf(x|\theta)\,dx \qquad (\mathbf{x}=x)\\ &= \int T\frac{\partial}{\partial\theta}f(x|\theta)\,dx\end{aligned}$$

Let's see

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathrm{E}[T] &= \frac{\partial}{\partial\theta}\int_0^\theta T\frac{1}{\theta}dx\\ &= \frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\int_0^\theta T\,dx\right)\\ &= \left(\frac{\partial}{\partial\theta}\frac{1}{\theta}\right)\int_0^\theta T\,dx + \frac{1}{\theta}\frac{\partial}{\partial\theta}\int_0^\theta T\,dx\\ &= \left(\frac{\partial}{\partial\theta}\frac{1}{\theta}\right)\int_0^\theta T\,dx + \frac{T(\theta)}{\theta}\end{aligned}$$

$$\begin{aligned}\int_0^\theta T\frac{\partial}{\partial\theta}f(x|\theta)\,dx &= \left(\frac{\partial}{\partial\theta}\frac{1}{\theta}\right)\int_0^\theta T\,dx,\\ &\neq \frac{\partial}{\partial\theta}\mathrm{E}[T],\end{aligned}$$

unless $T(\theta)=0\ \forall\theta$.

Homework: repeat with the MLE estimator, scale it to be unbiased, then find its variance.

Loss Function

- Not only for assessment and comparison,
- but also for designing and optimization!

The loss function:

$$
\begin{aligned}
L(\theta, T(\mathbf{X})) &= |\theta - T(\mathbf{X})| && \text{(absolute error (AE))}\\
L(\theta, T(\mathbf{X})) &= (\theta - T(\mathbf{X}))^2 && \text{(squared error (SE))}\\
&\vdots
\end{aligned}
$$

expresses how the estimate $T(\mathbf{X})$ deviates from θ.

The risk:

$$R(\theta, T) = \underset{\mathbf{X}}{\mathrm{E}}\, L(\theta, T(\mathbf{X}))$$

is a function of θ. $R(\theta, T_1)$ may cross with $R(\theta, T_2)$.

MSE (special case):

$$
\begin{aligned}
MSE(\theta) &= R(\theta, T)\\
&= \underset{\mathbf{X}}{\mathrm{E}}\left[L(\theta, T(\mathbf{X}))\right],\\
L(\theta, T(\mathbf{X})) &= (\theta - T(\mathbf{X}))^2.
\end{aligned}
$$

Example 39 (Risk of σ^2 Est.) :

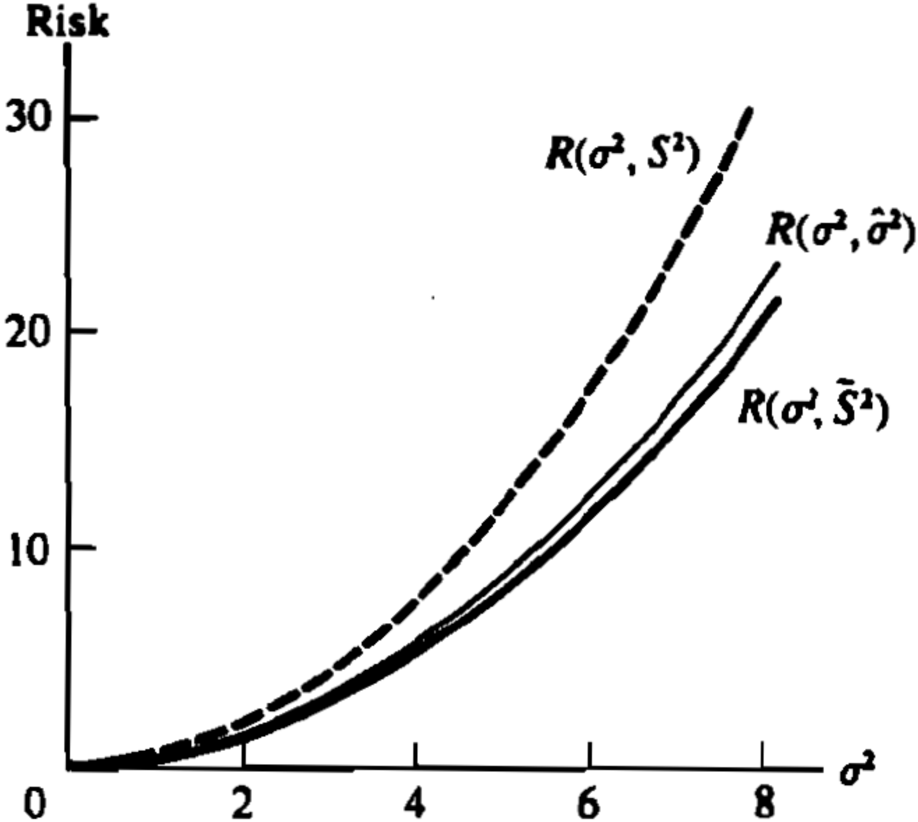
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \quad \left(R\left(\sigma^2, S^2\right) = \frac{2\sigma^4}{n-1}\right)$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \quad \left(R\left(\sigma^2, \widehat{\sigma}^2\right) = \frac{2n-1}{n^2}\sigma^4\right)$$

$$\widetilde{S}^2 = b\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \qquad \left(R\left(\sigma^2, \widetilde{S}^2\right)?\right)$$

$$\begin{aligned}
R\left(\sigma^2, \widetilde{S}^2\right) &= \mathrm{Var}\left[b(n-1)S^2\right] \\
&\quad + \left(\mathrm{E}\left[b(n-1)S^2\right] - \sigma^2\right)^2 \\
&= b^2(n-1)^2\frac{2\sigma^4}{n-1} + (b(n-1)-1)^2\sigma^4 \\
&= \left(2b^2(n-1) + (b(n-1)-1)^2\right)\sigma^4, \\
&= c\sigma^4, \\
c_{\min} &= \frac{2}{n+1} \qquad \left(\text{at } b = \frac{1}{n+1}\right) \\
\widetilde{S}^2 &= \frac{1}{n+1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2
\end{aligned}$$

$$\left(R\left(\sigma^2, \widetilde{S}^2\right) = \frac{2}{n+1}\sigma^4\right)$$



Risk
30
20
10
0
2
4
6
8
σ^2
$R(\sigma^2, S^2)$
$R(\sigma^2, \hat{\sigma}^2)$
$R(\sigma^2, \tilde{S}^2)$

Connection to Cramér-Rao Inequality

$$f\left(x|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{\left(x-\mu\right)^2}{\sigma^2}\right)$$

$$l\left(\theta\right) = -\log\sqrt{2\pi} - \frac{1}{2}\log\theta - \frac{1}{2\theta}\left(x-\mu\right)^2$$

$$\left(\theta = \sigma^2\right)$$

$$l'\left(\theta\right) = \frac{-1}{2\theta} + \frac{\left(x-\mu\right)^2}{2\theta^2}$$

$$l''\left(\theta\right) = \frac{1}{2\theta^2} - \frac{\left(x-\mu\right)^2}{\theta^3}$$

$$\mathrm{E}\left[l''\left(\theta\right)\right] = \frac{1}{2\theta^2} - \frac{\theta}{\theta^3} = \frac{-1}{2\theta^2}$$

$$I\left(\theta\right) = -\mathrm{E}\left[\frac{\partial^2 l\left(\theta\right)}{\partial\theta^2}\right] = \frac{1}{2\sigma^4}$$

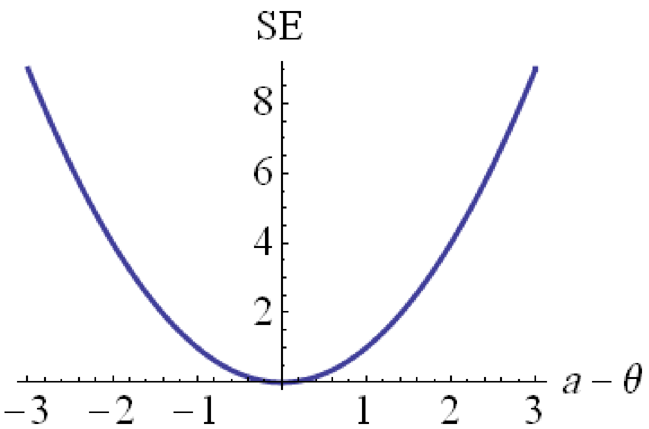
$$\mathrm{Var}\left[T\right] \geq \frac{1}{nI\left(\theta\right)} = \frac{2\sigma^4}{n},$$

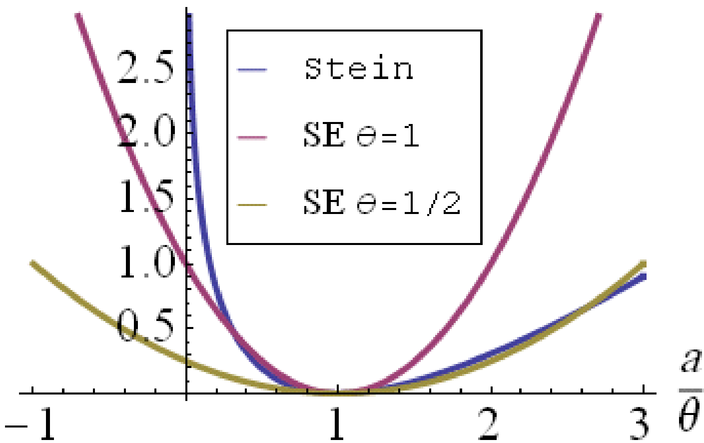
- lower bound of any unbiased estimator of σ^2.
- not attainable by the unbiased version above

Assessing with different Loss Function:

$$L(\theta, a) = (a-\theta)^2 = \theta\left(\frac{a}{\theta} - 1\right)^2 \qquad \text{(SE loss)}$$

$$L(\theta, a) = \frac{a}{\theta} - 1 - \log\left(\frac{a}{\theta}\right) \qquad \text{(Stien's loss)}$$





$$\widetilde{S}^2 = b\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

$$L(\theta, a) = \frac{a}{\theta} - 1 - \log\left(\frac{a}{\theta}\right)$$

$$R\left(\sigma^2, \widetilde{S}^2\right) = \mathrm{E}\left[b(n-1)\frac{S^2}{\sigma^2} - 1 - \log\frac{b(n-1)S^2}{\sigma^2}\right]$$

$$= b\,\mathrm{E}\left[\chi^2_{n-1}\right] - 1 - \log b - \mathrm{E}\log\chi^2_{n-1}$$

$$\frac{\partial R}{\partial b} = \mathrm{E}\left[\chi^2_{n-1}\right] - \frac{1}{b} \qquad (\overset{set}{=} 0)$$

$$b = \frac{1}{\mathrm{E}\left[\chi^2_{n-1}\right]} = \frac{1}{n-1}$$

$$\widetilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = S^2.$$

“Better” in which sense?

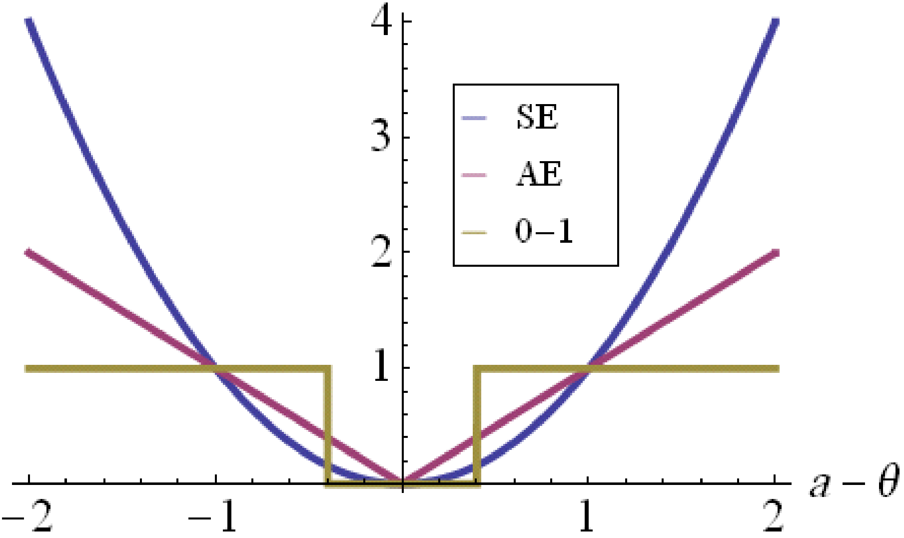
Obtaining Bayesian's Estimator by Loss Function Optimization!

$$
\begin{aligned}
R(\theta, T) &= \underset{\mathbf{X}}{\mathrm{E}}\, L(\theta, T(\mathbf{X})) \\
&= \int L(\theta, T(\mathbf{x}))\, f_{\mathbf{X}}(\mathbf{x}|\theta)\, d\mathbf{x}
\end{aligned}
$$

- no uniformly "best" estimator.
- $R(\theta, T_1)$ may cross with $R(\theta, T_2)$.

$$
\begin{aligned}
\underset{\Theta}{\mathrm{E}}\, R(\theta, T) &= \int_\theta R(\theta, T)\, f_\Theta(\theta)\, d\theta \\
&= \int_\theta \left[\int_{\mathbf{x}} L(\theta, T(\mathbf{x}))\, f_{\mathbf{X}}(\mathbf{x}|\theta)\, d\mathbf{x} \right] f_\Theta(\theta)\, d\theta \\
&= \int_{\mathbf{x}} \left[\int_\theta L(\theta, T(\mathbf{x}))\, f_{\mathbf{X}}(\mathbf{x}|\theta)\, f_\Theta(\theta)\, d\theta \right] d\mathbf{x} \\
&= \int_{\mathbf{x}} \left[\int_\theta L(\theta, T(\mathbf{x}))\, f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \right] f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} \\
T &= \arg\min_T \int_\theta L(\theta, T(\mathbf{x}))\, f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta
\end{aligned}
$$

Solutions under different loss functions:



$$T_1 = \arg\min_T \int_\theta (T-\theta)^2 f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \qquad \text{(SE loss)}$$

$$= \int_\theta \theta f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \qquad \text{(Posterior mean)}$$

$$T_2 = \arg\min_T \int_\theta |T-\theta|\, f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \qquad \text{(AE loss)}$$

$$\begin{aligned}
R &= \int_\theta |T-\theta|\, f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \\
&= \int_{-\infty}^{T} (T-\theta)\, f(\theta)\, d\theta + \int_T^\infty -(T-\theta)\, f(\theta)\, d\theta \\
&= T\int_{-\infty}^{T} f(\theta)\, d\theta - \int_{-\infty}^{T} \theta f(\theta)\, d\theta - \\
&\quad T\int_T^\infty f(\theta)\, d\theta + \int_T^\infty \theta f(\theta)\, d\theta \\
\frac{\partial R}{\partial T} &= \left(\int_{-\infty}^{T} f(\theta)\, d\theta + Tf(T)\right) - Tf(T) - \\
&\quad \left(\int_T^\infty f(\theta)\, d\theta - Tf(T)\right) - Tf(T) \\
&= \int_{-\infty}^{T} f(\theta)\, d\theta - \int_T^\infty f(\theta)\, d\theta \qquad (\stackrel{set}{=} 0) \\
0 &= F_{\Theta|\mathbf{X}}^{-1}(T) - \left(1 - F_{\Theta|\mathbf{X}}^{-1}(T)\right) \\
0.5 &= F_{\Theta|\mathbf{X}}^{-1}(T) \\
T_2 &= F_{\Theta|\mathbf{X}}^{-1}(0.5) \qquad \text{(Posterior median)}
\end{aligned}$$

$$
\begin{aligned}
T_3 &= \arg\min_T \int_\theta I_{0\le|T-\theta|} f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \quad (0-1 \text{ loss})\\
R &= \int_\theta I_{a\le|T-\theta|} f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta\\
&= \int_{a\le|T-\theta|} f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta\\
&= 1-\int_{|T-\theta|<a} f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta\\
&= 1-\int_{T-a}^{T+a} f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta\\
&= 1-\Pr_{\Theta|\mathbf{X}}\left[|\theta-T|<a\right]
\end{aligned}
$$

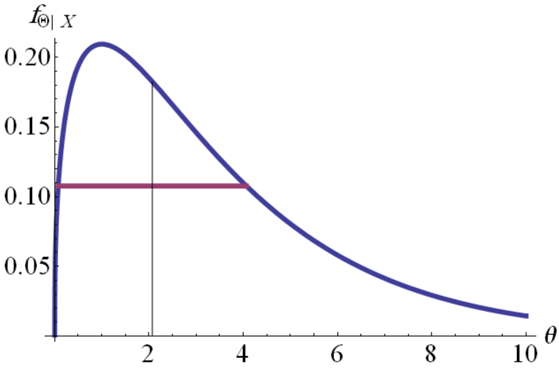
Notice that: we have to maximize the probability $\int_{T-a}^{T+a} f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta$. The period $[T-a, T+a]$ has

- a length of $(T+a)-(T-a)=2a$
- mid point of $\frac{1}{2}\left[(T+a)+(T-a)\right]=T$.
- T and mode do not necessarily coincide.,

which means that T_3 is mid-point of $2a$ modal interval.

$$\frac{\partial R}{\partial T} = f_{\Theta|\mathbf{X}}(T + a|\mathbf{x}) - f_{\Theta|\mathbf{X}}(T - a|\mathbf{x}), \qquad (\overset{set}{=} 0)$$

$$f_{\Theta|\mathbf{X}}(T + a|\mathbf{x}) = f_{\Theta|\mathbf{X}}(T - a|\mathbf{x}).$$



For unimodal symmetric $f_{\Theta|\mathbf{X}}$:
$f_{\Theta|\mathbf{X}}(\theta - M) = f_{\Theta|\mathbf{X}}(\theta + M)$. Therefore,

$$T_3 = Mode. \qquad \text{(MAP)}$$

For $a \to 0$

$$R \approx 1 - f_{\Theta|\mathbf{X}}(T|\mathbf{x}) \cdot 2a,$$
$$T_3 = \arg\max_T f_{\Theta|\mathbf{X}}(T|\mathbf{x}) = Mode \qquad \text{(MAP)}$$

Of course T_3 could have been any point if we starte minimizing the risk from begining not by obtaining the limit:

$$\begin{aligned} R &= 1 - \int_{T-a}^{T+a} f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \\ &= 1 - \int_{T}^{T} f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})\, d\theta \\ &= 1, \end{aligned}$$

unless Θ is discrete or categorical as in Pattern Recognition.

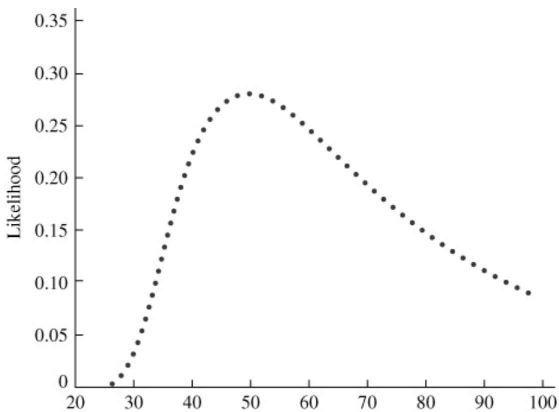
Estimation for Discrete Θ

MLE, Bayesian, Loss Functions have same treatment. However, maximization, expectation,..etc are taken over discrete space. Also, Cramér-Rao Lower Bound is derived for continuous case!

Example 40 (Capture Recapture Method) *: as in Example 15, page 19, first course. x captured animal in a population of θ animals. x was found to be 4 (we renamed variables):*

$$L(\theta) = P(x|\theta) = \frac{\binom{10}{4}\binom{\theta-10}{20-4}}{\binom{\theta}{20}}, \qquad \text{(Likelihood)}$$

$$\widehat{\theta}_{MLE} = 50$$



- maximization is obtained by $L_\theta / L_{\theta+1}$ not by $\frac{\partial L}{\partial \theta}$.

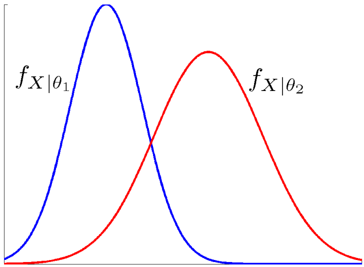
- Bayesian estimation is exactly the same thro defining $f_\Theta(\theta)$.

- However, $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ will be discrete.

Estimation for Categorical Θ (basis for Pattern Recognition)

- $\Theta = \{\theta_1, \ldots, \theta_K\}$, with K categories (classes).
- E.g., $\Theta = \{Male, Female\}$
- MoM is not applicable here (Θ is not numeric).

$$X|\theta_1 \sim N(1.5, .08),$$
$$X|\theta_2 \sim N(1.7, .1).$$



Suppose we got 1.77, 1.58, 1.77, 1.86, 1.75, 1.80, 1.77, 1.67, 1.73, 1.62. Are these readings obtained from Male or Female population?

Bayesian Estimation and MLE

8.5.3 Asymptotic Relative Efficiency (ARE)

Definition 41 *The (sequence of) estimator T_n is said to be asymptotically efficient for θ if*

$$\sqrt{n}\,(T_n - \theta) \xrightarrow{d} N\left(0, \sigma^2\right),$$
$$\sigma^2 = \frac{1}{I(\theta)},$$

which is Cramér-Rao Lower Bound.

It is clear that MLE is asymptotically efficient.