STEVENS INSTITUTE of TECHNOLOGY
School of Business

BIA-652

Multivariate Data Analytics

Estimation

Prof. Feng Mai
School of Business

For academic use only.

Sampling Distribution

Some slides are courtesy of Prof. Andrew Moore and Larry Ruzzo
BIA 652, Multivariate Data Analytics

Statistical Inference

- The purpose of **statistical inference** is to obtain information about a population from information contained in a sample.
- A **population** is the set of all the elements of interest in a study.
- A **sample** is a subset of the population.
- A **parameter** is a characteristic of a population. It is a fixed, unknown number.
- The sample can provide estimates of the values of the population characteristics (parameters).

Sequences of Independent Random Variables

Example:

$X_1, X_2, \ldots, X_n$ = a set of n Normal random variables

- Same (marginal) probability distribution, f(x)
- Identical μ and σ
- μ and σ are NOT random!
- Each X_i is random
- Statistically independent

- IID: independent identically distributed
- This is a “random sample” from the population f(x).

Point Estimation (Undergrad version)

In **point estimation** we use the data from the sample to compute a value of the sample statistic that serves as an estimate of a population parameter.

- We refer to $\overline{x}$ (sample mean) as the *point estimate* of the population mean μ.

s (sample standard deviation) is the point estimator of the population standard deviation σ.

1. How accurate is the estimate?
2. Why?

How accurate is the estimate?

The Sample Mean

Random Variable!!!!

Sample Mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \ldots + \frac{1}{n}X_n$

$$E[\bar{X}] = E\left(\frac{1}{n}X_1\right) + E\left(\frac{1}{n}X_2\right) + \ldots + E\left(\frac{1}{n}X_n\right)$$

$$= \frac{1}{n}E[X_1] + \ldots + \frac{1}{n}E[X_n] = \frac{1}{n}\mu + \ldots + \frac{1}{n}\mu$$

$$= \mu$$

$$\mathrm{Var}[\bar{X}] = \sum_{i=1}^{n} \mathrm{Var}\left[\frac{1}{n}X_i\right] = \sum_{i=1}^{n} \frac{1}{n^2}\mathrm{Var}[X_i] = \sum_{i=1}^{n} \frac{1}{n^2}\sigma^2$$

$$= \frac{\sigma^2}{n}$$

Distribution of $\bar{X}$? Remains to be seen.

Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having mean 0 and variance σ^2 and the common distribution function F and moment-generating function M defined in a neighborhood of zero. Let

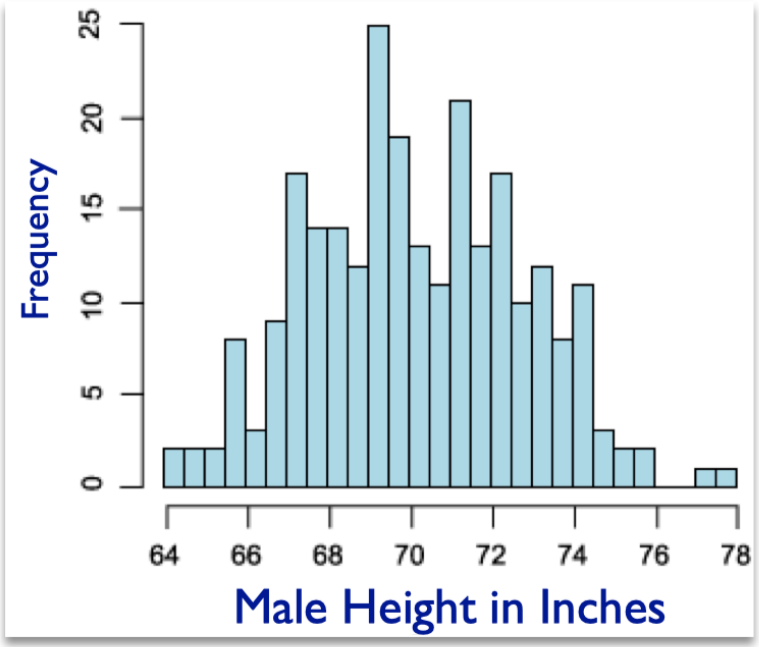
$$S_n = \sum_{i=1}^{n} X_i$$

Then

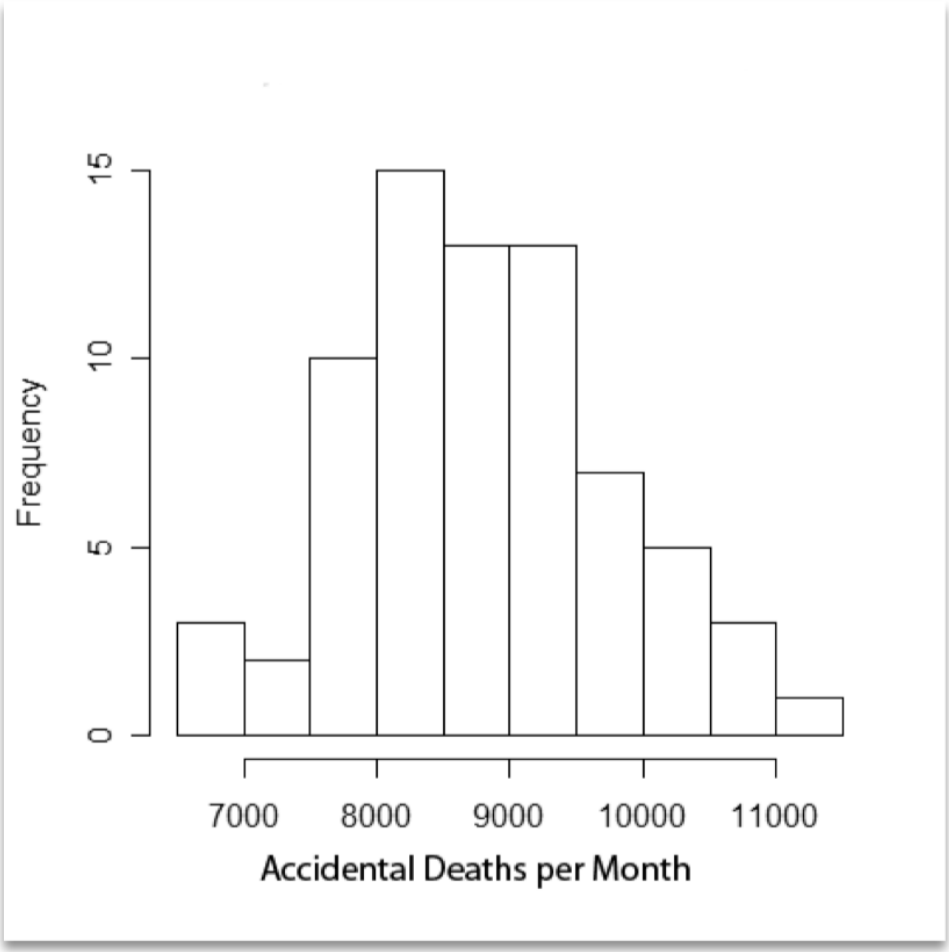
$$\lim_{n\to\infty} P\left(\frac{S_n}{\sigma\sqrt{n}} \leq x\right) = \Phi(x), \qquad -\infty < x < \infty$$

Human height is approximately normal.

Why might that be true?



R.A. Fisher (1918) noted it would follow from CLT if height were the sum of many independent random effects, e.g. many genetic factors (plus some environmental ones like diet). I.e., suggested part of *mechanism* by looking at *shape* of the curve.



BIA 652, Multivariate Data Analytics

In Section 5.3, we considered a sequence of independent and identically distributed (i.i.d.) random variables, $X_1, X_2, \ldots$ having the common mean and variance μ and σ^2. The sample mean of $X_1, X_2, \ldots, X_n$ is

$$\overline{X}_n = \frac{1}{n}\sum^{n} X_i$$

This sample mean has the properties

$$E(\overline{X}_n) = \mu$$

and

$$\text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

The central limit theorem says that, for a fixed number z,

$$P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z) \qquad \text{as } n \to \infty$$

where Φ is the cumulative distribution function of the standard normal distribution. Using a more compact and suggestive notation, we have

$$P\left(\frac{\overline{X}_n - \mu}{\sigma_{\overline{X}_n}} \leq z\right) \to \Phi(z)$$

Central Limit Theorem (Undergrad Version)

Sampling Distribution of $\bar{x}$

The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$.

- If we use a large ($n \geq 30$) simple random sample, the **central limit theorem** enables us to conclude that the sampling distribution of $\bar{X}$ can be approximated by a ***normal*** probability distribution.

$$\overline{X}_n \approx \mathrm{N}(\mu, \sigma^2/n),$$

Why are sample mean and sample standard deviation good estimators?

The Method of Moments

The Method of Moments

- The basic idea of this method is to equate certain sample characteristics, such as the mean, to the corresponding population kth moments ($E[X^k]$).
- Then solving these equations for unknown parameter values yields the estimators.
- If there are k parameters, you need at least k equations to solve for the parameters.

The Method of Moments

Definition

Let $X_1, \ldots, X_n$ be a random sample from a pmf or pdf $f(x)$. For $k = 1, 2, 3, \ldots$, the ***k*th population moment,** or ***k*th moment of the distribution *f*(*x*),** is $E(X^k)$. The ***k*th sample moment** is

$$(1/n)\sum_{i=1}^{n} X_i^k.$$

Thus the first population moment is $E(X) = \mu$, and the first sample moment is $\Sigma X_i/n = \overline{X}$.

The second population and sample moments are $E(X^2)$ and $\Sigma X_i^2/n$, respectively. The population moments will be functions of any unknown parameters $\theta_1, \theta_2, \ldots$

The Method of Moments

Definition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pmf or pdf $f(x; \theta_1, \ldots, \theta_m)$, where $\theta_1, \ldots, \theta_m$ are parameters whose values are unknown.

Then the **moment estimators** $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are obtained by equating the first m sample moments to the corresponding first m population moments and solving for $\theta_1, \ldots, \theta_m$.

Under reasonable conditions, method of moments estimates have the desirable property of consistency. An estimate, $\hat{\theta}$, is said to be a **consistent** estimate of a parameter, θ, if $\hat{\theta}$ approaches θ as the sample size approaches infinity. The following states this more precisely.

DEFINITION

Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is said to be consistent in probability if $\hat{\theta}_n$ converges in probability to θ as n approaches infinity; that is, for any $\epsilon > 0$,

$$P(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \qquad \text{as } n \to \infty$$ ■

Example: Exponential Distribution

Let $X_1, X_2, \ldots, X_n$ represent a random sample of service times of n customers at a certain facility, where the underlying distribution is assumed exponential with parameter λ.

Since there is only one parameter to be estimated, the estimator is obtained by equating $E(X)$ to $\overline{X}$.

Since $E(X) = 1/\lambda$ for an exponential distribution, this gives $1/\lambda = \overline{X}$ or $\lambda = 1/\overline{X}$. The moment estimator of λ is then $\hat{\lambda} = 1/\overline{X}$.

Maximum Likelihood Estimation

Assuming sample $x_1, x_2, ..., x_n$ is from a parametric distribution $f(x|\theta)$, estimate θ.

E.g.: Given sample HHTTTTTHTHTTTHH of (possibly biased) coin flips, estimate

θ = probability of Heads

$f(x|\theta)$ is the Bernoulli probability mass function with parameter θ

Likelihood

$P(x \mid \theta)$: Probability of event x given *model* θ

Viewed as a function of x (fixed θ), it's a *probability*

E.g., $\Sigma_x P(x \mid \theta) = 1$

Viewed as a function of θ (fixed x), it's a *likelihood*

E.g., $\Sigma_\theta P(x \mid \theta)$ can be anything; *relative* values of interest.

E.g., if θ = prob of heads in a sequence of coin flips then

$P(HHTHH \mid .6) > P(HHTHH \mid .5)$,

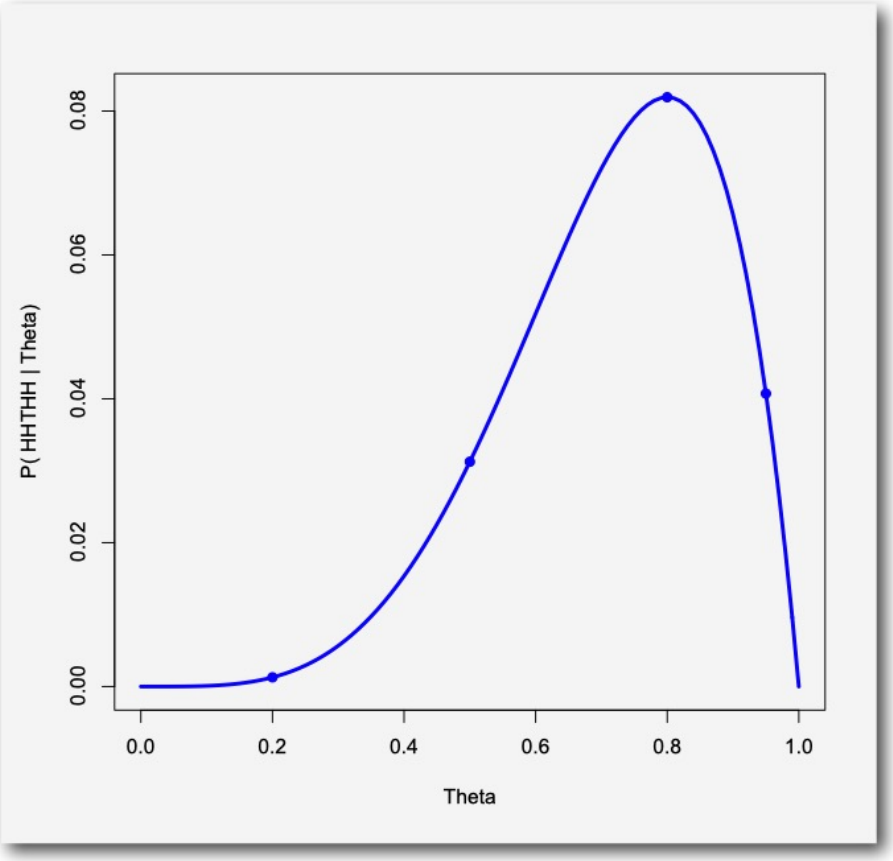
I.e., event HHTHH is *more likely* when $\theta = .6$ than $\theta = .5$

And what θ make HHTHH *most* likely?

Likelihood Function

P(HHTHH | θ):
Probability of HHTHH,
given P(H) = θ:

θ	$\theta^4(1-\theta)$
0.2	0.0013
0.5	0.0313
0.8	0.0819
0.95	0.0407



Maximum Likelihood Parameter Estimation

One (of many) approaches to param. est.
Likelihood of (indp) observations $x_1, x_2, ..., x_n$

$$L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

As a function of θ, what θ maximizes the likelihood of the data actually observed
Typical approach: $\frac{\partial}{\partial\theta} L(\vec{x} \mid \theta) = 0$ or $\frac{\partial}{\partial\theta} \log L(\vec{x} \mid \theta) = 0$

$$\text{Log-Likelihood} = \ln[\, f\,(x_1, \ldots, x_5 \mid \theta)]$$

$$= \ln[\theta^4(1-\theta)^1] = 4\ln(\theta) + \ln(1-\theta)$$

Maximum Likelihood learning of Gaussians

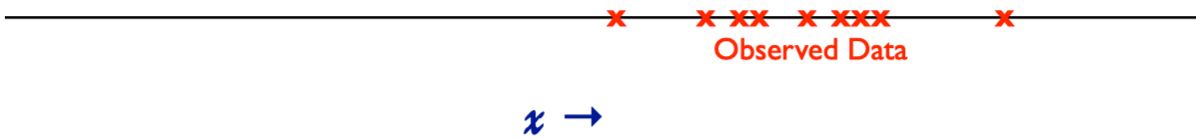
- Why we should care
- Learning Univariate Gaussians (Normal)
- Learning Multivariate Gaussians
- What's a biased estimator?

Why we should care

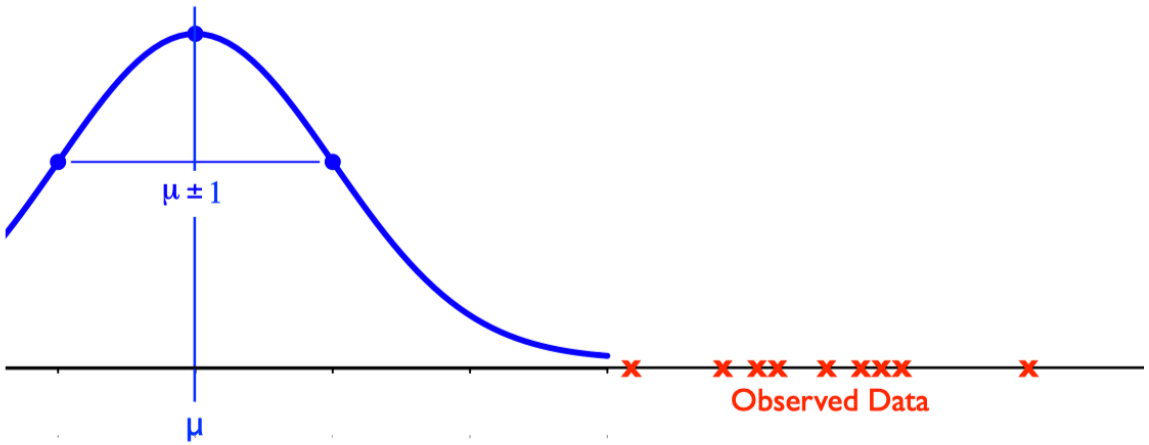
- Maximum Likelihood Estimation is a very very very very fundamental part of data analysis.
- “MLE for Gaussians” is training wheels for future techniques
- Learning Gaussians is more useful than you might guess…

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

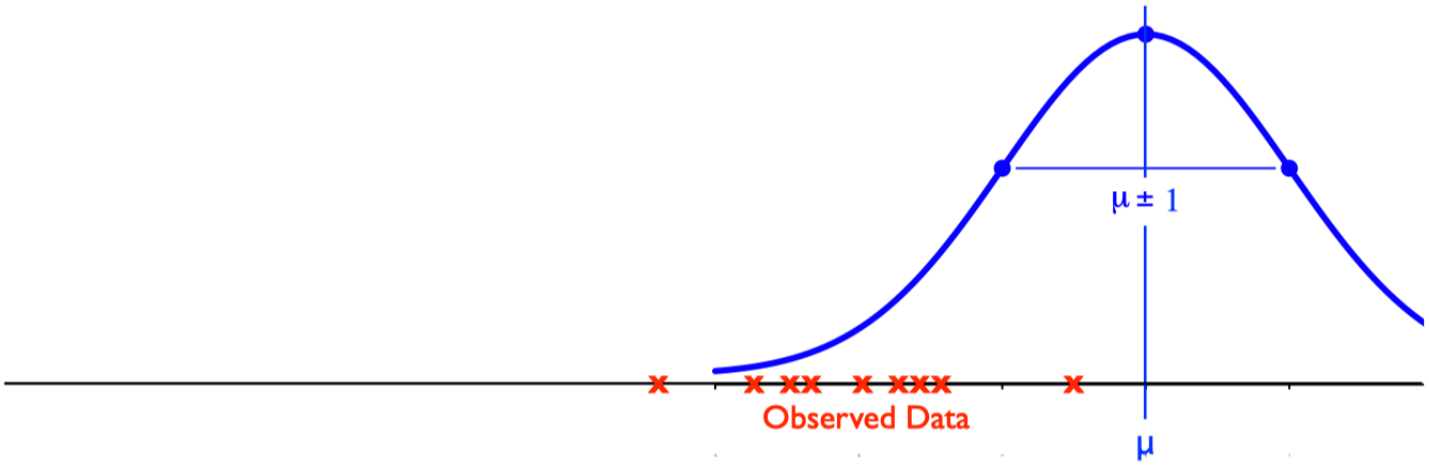
Ex2: I got data; a little birdie tells me it's normal, and promises $\sigma^2 = 1$



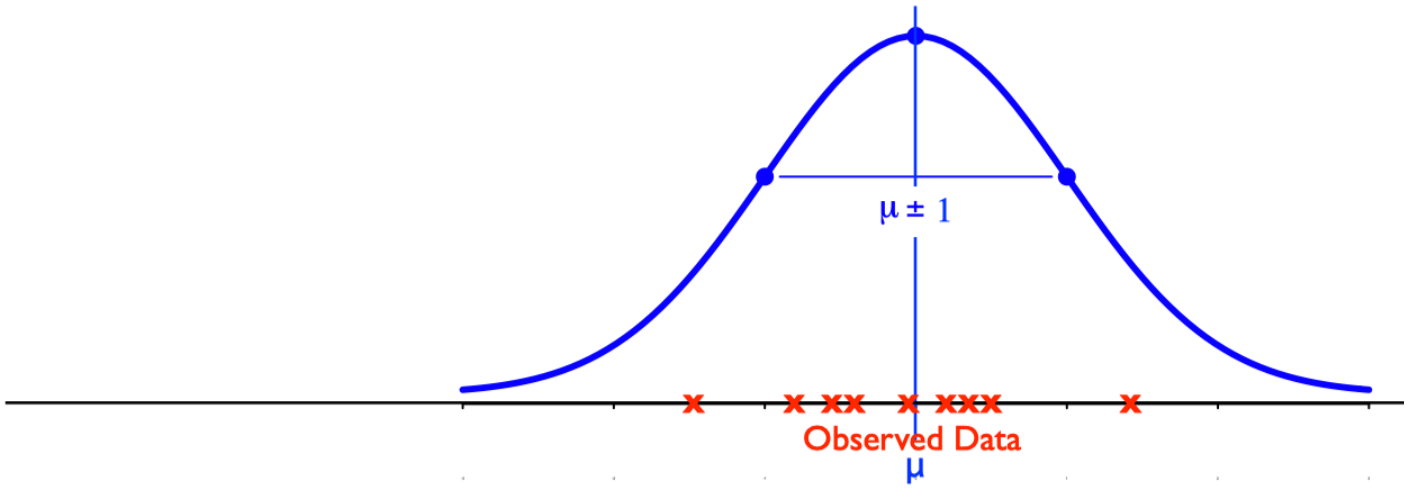
Which is more likely: (a) this?



Which is more likely: (b) or this?



Which is more likely: (c) or *this*?



Gaussian density: $\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$.

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_n \mid \mu, \sigma^2)$$

= (by i.i.d)

= (monotonicity of log)

= (plug in formula for Gaussian)

= (after simplification)

Intermission: A General Scalar MLE strategy

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out $\partial LL/\partial\theta$ using high-school calculus
3. Set $\partial LL/\partial\theta=0$ for a maximum, creating an equation in terms of θ
4. Solve it*
5. Check that you've found a maximum rather than a minimum or saddle-point, and be careful if θ is constrained

*This is a perfect example of something that works perfectly in all textbook examples and usually involves surprising pain if you need it for something new

The MLE μ

$$\hat{\mu}^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_n \mid \mu, \sigma^2)$$

$$= \arg\min_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$= \mu \quad \text{s.t.} \quad 0 = \frac{\partial \text{LL}}{\partial \mu} =$$

= (what?)

BIA 652, Multivariate Data Analytics

$$\hat{\mu}^{mle} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The best estimate of the mean of a normal distribution is the mean of the sample!

A General MLE strategy

Suppose $\mathbf{q} = (q_1, q_2, \ldots, q_n)^T$ is a vector of parameters.

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out $\partial LL/\partial\theta$ using high-school calculus
3. Solve the set of simultaneous equations

$$\frac{\partial LL}{\partial \theta_1} = 0$$
$$\frac{\partial LL}{\partial \theta_2} = 0$$
$$\vdots$$
$$\frac{\partial LL}{\partial \theta_n} = 0$$

4. Check that you're at a maximum

A General MLE strategy

Suppose $\mathbf{q} = (q_1, q_2, \ldots, q_n)^T$ is a vector of parameters.

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out $\partial LL/\partial\theta$ using high-school calculus
3. Solve the set of simultaneous equations

$$\frac{\partial LL}{\partial \theta_1} = 0$$
$$\frac{\partial LL}{\partial \theta_2} = 0$$
$$\vdots$$
$$\frac{\partial LL}{\partial \theta_n} = 0$$

4. Check that you're at a maximum

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_n$ ~(i.i.d) $N(\mu,\sigma^2)$
- But you don't know μ or σ^2
- MLE: For which $\theta = (\mu,\sigma^2)$ is $x_1, x_2, \ldots x_n$ most likely?

$$LL = \log p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$
$$= -n\left(\log\pi + \frac{1}{2}\log\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Take derivative. By the first-order condition (FOC), set derivatives to 0):

$$\frac{\partial LL}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$
$$\frac{\partial LL}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_n$ ~(i.i.d) $N(\mu,\sigma^2)$
- But you don't know μ or σ^2
- MLE: For which $\theta = (\mu,\sigma^2)$ is $x_1, x_2, \ldots x_n$ most likely?

$$\hat{\mu}^{mle} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}^{mle}\right)^2$$

Unbiased Estimators

- An estimator of a parameter is unbiased if the expected value of the estimate is the same as the true value of the parameters.

- *If* $x_1, x_2, \ldots x_n$ ~(i.i.d) $N(\mu,\sigma^2)$ then

$$E[\hat{\mu}^{mle}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \mu$$

$\hat{\mu}^{mle}$ is unbiased

Biased Estimators

- An estimator of a parameter is biased if the expected value of the estimate is different from the true value of the parameters.
- *If* $x_1, x_2, \ldots x_n$ ~(i.i.d) $N(\mu,\sigma^2)$ then

$$E\left[\hat{\sigma}^2_{mle}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}^{mle}\right)^2\right] = E\left[\frac{1}{n}\left(\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2\right] \neq \sigma^2$$

$\hat{\sigma}^2_{mle}$ is **biased**

MLE Variance Bias

If $x_1, x_2, \ldots x_n$ ~(i.i.d) $N(\mu,\sigma^2)$ then

$$E\left[\hat{\sigma}^2_{mle}\right] = E\left[\frac{1}{n}\left(\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2\right] = \left(1 - \frac{1}{n}\right)\sigma^2 \neq \sigma^2$$

Intuition check: consider the case of =1

Why should our guts expect that s^2_{mle} would be an underestimate of true s^2?

See reading for a proof.

Unbiased estimate of Variance

Define

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{\hat{\sigma}^2_{mle}}{\left(1 - \frac{1}{n}\right)} \qquad \text{So } E\left[\hat{\sigma}^2_{\text{unbiased}}\right] = \sigma^2$$

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \hat{\mu}^{mle}\right)^2$$

Unbiased discussion

- *Which is best?*

$$\hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}^{mle})^2$$

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}^{mle})^2$$

Answer:

It depends on the task

And doesn't make much difference once --> large

Don’t get too excited about being unbiased

- *Assume* $x_1, x_2, \ldots x_n$ ~(i.i.d) $N(\mu,\sigma^2)$
- Suppose we had these estimators for the mean

- $\hat{\mu}^{bad} = \frac{1}{n+7\sqrt{n}} \sum_i x_i$

$$\hat{\mu}^{crap} = x_1$$

Are either of these unbiased?

Will either of them asymptote to the correct value as n gets large?

If X follows a Poisson distribution with parameter λ, that is,

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

If $X_1, \ldots, X_n$ are i.i.d. and Poisson, what is the MLE for λ?

Nice Properties of the MLE

MLE $\hat{\theta}$ has the following nice properties:

- Consistency: $P_{\theta_0}(\hat{\theta}_{MLE} = \theta_0) \to 1 \qquad n \to \infty$
- Asymptoticly Normal: $\sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, \sigma^2_{MLE})$
- Asymptotic optimality: MLE has the smallest variance
- Invariance Property

9.14 Theorem. *Let $\tau = g(\theta)$ be a function of θ. Let $\widehat{\theta}_n$ be the* MLE *of θ. Then $\widehat{\tau}_n = g(\widehat{\theta}_n)$ is the* MLE *of τ.*

Confidence intervals

How accurate do we expect $\hat{\mu}^{mle}$ and $\hat{\sigma}^2_{mle}$ to be, and how to estimate these accuracies from data?

• Analytically (see Rice 8.5.2 and 8.5.3 and Fisher information)

Define $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$ then $\mathrm{Var}(\hat{\theta} - \theta_0) \approx \frac{1}{nI(\theta_0)}$

• Numerically (observed Fisher information)

$$l_n''(\theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2}\log f_\theta(X_i)$$

• Non-parametrically (using randomization and bootstrapping)

What is the confidence interval for Poisson MLE?

If $x_1 \ldots x_n$ are iid samples from continuous Uniform(0, θ). What is the MLE for θ?

STEVENS
INSTITUTE of TECHNOLOGY
School of Business

Thank you!

Prof. Feng Mai
School of Business

For academic use only.