NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W4: 9/8-10/15

NC STATE UNIVERSITY

Administrative

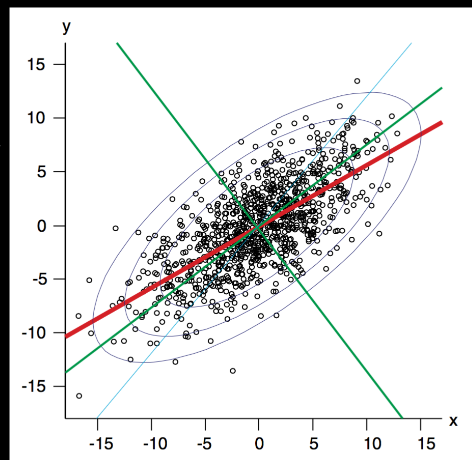
- HW-1: Posted on Moodle
- Due: 9/21/15, 23:55pm

Key points from 9/3

Assume Normal Distribution: $N_2(\mathbf{0}, \mathbf{\Sigma})$; $\sigma_x = 5$, $\sigma_y = 4$, $\rho_{xy} = 0.7$.

Regression line has the slope $\beta \approx \rho_{xy}\sigma_y/\sigma_x$

Direction of first major axis of contour ellipse is determined by the first eigenvector of $\mathbf{\Sigma}$



Solution requires basic understanding of LA/MA and Calculus

Today

- Parameter Estimation

NC STATE UNIVERSITY

So far

- We started with data (samples) drawn from a population
- We wanted to make inference (conclusions) about populations from noisy data that is drawn from it
- The randomness governing our data (samples) is given by densities and mass functions

NC STATE UNIVERSITY

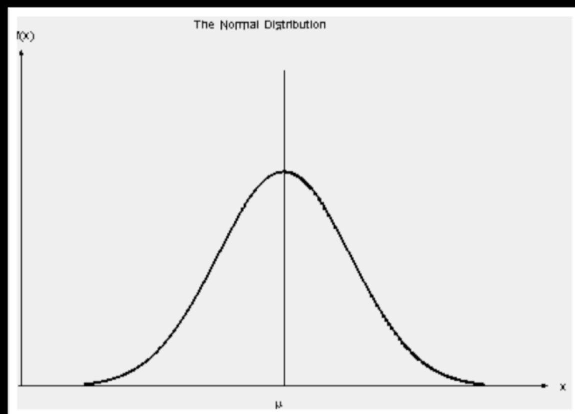
Expected Values

- Expected value of a random variable is intuitively the long-run average value of repetitions of the experiment it represents.
  - E.g., the expected value of a dice roll is 3.5
- The expected value is also known as the expectation, mathematical expectation, EV, mean, or first moment.

NC STATE UNIVERSITY

Expected Values

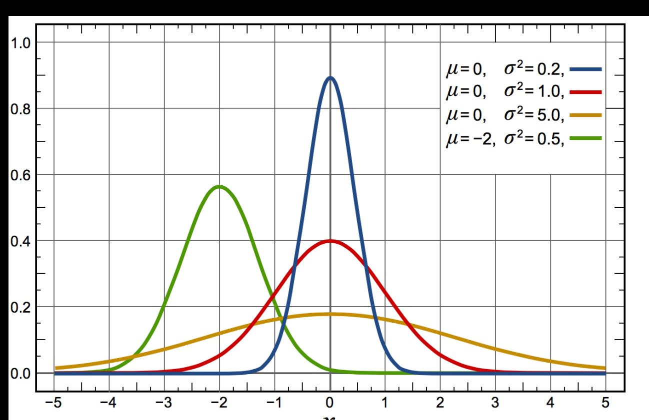
- The mean is the center of the distribution



NC STATE UNIVERSITY

Expected Values

- The variance and standard deviation are measures of horizontal spread or dispersion of the random variable



NC STATE UNIVERSITY

Expected Value

Outcomes	a_1	a_2	…	a_n
Probability	p_1	p_2		p_n

- $E[X] = a_1p_1 + a_2p_2 + \ldots + a_np_n = \Sigma_i a_i p(a_i)$

DEFINITION. The *expectation* of a discrete random variable X taking the values $a_1, a_2, \ldots$ and with probability mass function p is the number

$$\mathrm{E}[X] = \sum_i a_i \mathrm{P}(X = a_i) = \sum_i a_i p(a_i).$$

9/9/15 © Raju Vatsavai CSC-591. 9

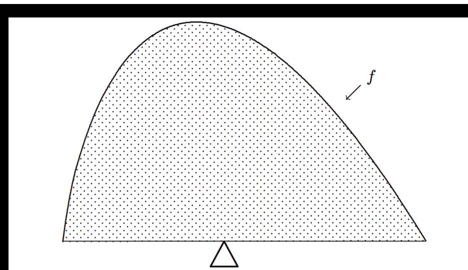
NC STATE UNIVERSITY

Expected Value

DEFINITION. The *expectation* of a continuous random variable X with probability density function f is the number

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x.$$

Expected value as center of gravity



9/9/15 © Raju Vatsavai CSC-591. 10

NC STATE UNIVERSITY

Variance

DEFINITION. The *variance* $\mathrm{Var}(X)$ of a random variable X is the number

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right].$$

AN ALTERNATIVE EXPRESSION FOR THE VARIANCE. For any random variable X,

$$\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2.$$

NC STATE UNIVERSITY

Example

- The random variable is given by the following PDF.

$$f(x) = \begin{cases} 2(1-x) & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise} \end{cases}$$

- First verify that f(x) is valid PDF.
  - 2(1-x) = 2 - 2x ≥ 0 precisely when x ≤ 1; thus f(x) is everywhere nonnegative

NC STATE UNIVERSITY

Few Basic Derivatives

$f(x)$	$f'(x)$
x^n	nx^{n-1}
$\sin x$	$\cos x$
$\cos x$	$-\sin x$
$\tan x$	$\sec^2 x$
e^x	e^x
$\ln x$	$\frac{1}{x}$

$$\frac{d}{dx}(Af(x)+Bg(x)) = Af'(x)+Bg'(x) \quad \text{(the sum rule);}$$
$$\frac{d}{dx}(f(x)g(x)) = f(x)g'(x)+g(x)f'(x) \quad \text{(the product rule);}$$
$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{g(x)f'(x)-f(x)g'(x)}{(g(x))^2} \quad (g(x)\neq 0) \quad \text{(the quotient rule).}$$

NC STATE UNIVERSITY

Few Basic Integrals

$$\int (Cf_1(x)+Df_2(x))\,dx = C\int f_1(x)\,dx + D\int f_2(x)\,dx,$$

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

$$\int \frac{1}{x}\,dx = \ln|x|.$$

$f(x)$	$\int f(x)\,dx$
$x^\alpha \quad (\alpha\neq -1)$	$\frac{x^{\alpha+1}}{\alpha+1}$
x^{-1}	$\ln|x|$
$\cos x$	$\sin x$
$\sin x$	$-\cos x$
$\sec^2 x$	$\tan x$
e^{kx}	$\frac{e^{kx}}{k}$

NC STATE UNIVERSITY

Example

- Check if f(x) has unit area under its graph

$$\int_{-\infty}^{\infty} f(x)dx = 2\int_0^1 (1-x)dx = 2\left(x - \frac{x^2}{2}\right)\Big|_0^1 = 1$$

- Therefore, f(x) is a valid PDF.
- Now compute Expected value

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x.$$

NC STATE UNIVERSITY

Example

- Mean

$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} x f(x)dx \\ &= \int_0^1 x\Big[2(1-x)\Big]dx \\ &= 2\int_0^1 \left(x - x^2\right)dx \\ &= 2\left(\frac{x^2}{2} - \frac{x^3}{3}\right)\Big|_0^1 \\ &= 1/3\end{aligned}$$

- Compute Variance $\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right]$

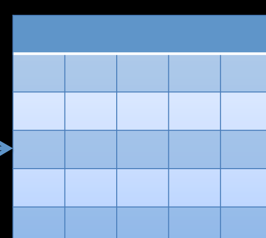
Example

- Variance

$$
\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f(x)dx \\
&= \int_0^1 (x - 1/3)^2 \cdot 2(1-x)dx \\
&= 2\int_0^1 (x^2 - \frac{2}{3}x + \frac{1}{9})(1-x)dx \\
&= 2\int_0^1 (-x^3 + \frac{5}{3}x^2 - \frac{7}{9}x + \frac{1}{9})dx \\
&= 2(-\frac{1}{4}x^4 + \frac{5}{9}x^3 - \frac{7}{18}x^2 + \frac{1}{9}x)\Big|_0^1 \\
&= 2(-\frac{1}{4} + \frac{5}{9} - \frac{7}{18} + \frac{1}{9}) \\
&= \frac{1}{18}
\end{aligned}
$$

Parameter Estimation

i.i.d
sample



~320 Million
Life Expectancy: 79.56 Y/M; 77.11 Y/FM
Sex Ratio: 0.97 M/FM

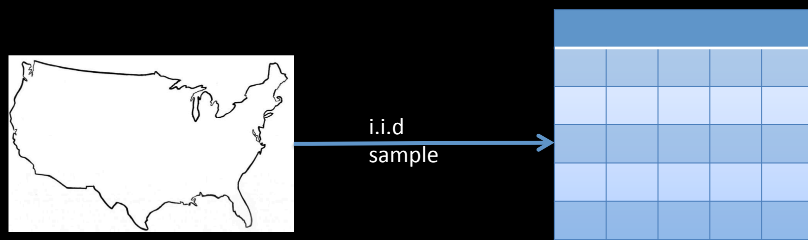
$f(x_i|p)$ is prob dist for M/FM

$$f(x_i|p) = p^{x_i} (1-p)^{1-x_i}$$

X_i = 1 for Male; 0 Otherwise

NC STATE UNIVERSITY

Parameter Estimation



~320 Million
Life Expectancy: 79.56 Y/M; 77.11 Y/FM
Sex Ratio: 0.97 M/FM

$f(x_i|p)$ is prob dist for M/FM

$f(x_i|p) = p^{x_i} (1 - p)^{1-x_i}$

X_i = 1 for Male; 0 Otherwise

For one sample
$f(1|p) = p^1(1-p)^{1-1} = p$
$f(0|p) = p^0(1-p)^{1-0} = 1-p$

What if we have N samples?
$f(x_1, x_2,,, x_N|p)$ (iid) = $p^{x1} (1 - p)^{1-x1} p^{x2} (1 - p)^{1-x2} \dots p^{xN} (1 - p)^{1-xN}$

(please note sub/super scripts correctly)

9/9/15 © Raju Vatsavai CSC-591. 19

NC STATE UNIVERSITY

Parameter Estimation

For one sample
$f(1|p) = p^1(1-p)^{1-1} = p$
$f(0|p) = p^0(1-p)^{1-0} = 1-p$

What if we have N samples?
$f(x_1, x_2,,, x_N|p)$ (iid) = $p^{x1} (1 - p)^{1-x1} p^{x2} (1 - p)^{1-x2} \dots p^{xN} (1 - p)^{1-xN}$
$f(x_1, x_2,,, x_N|p)$ (iid) = $\Pi^N_{i=1} p^{x_i} (1 - p)^{1-x_i}$ (Likelihood)
(joint probability)
In general we don't know what "p" is, so we need to estimate it from the data.
MLE: for which parameter "p", $x_1, x_2,,, x_N$ is most likely?
(please note sub/super scripts correctly)

9/9/15 © Raju Vatsavai CSC-591. 20

Maximum Likelihood Estimation

- General outline for single parameter
  - Write down the likelihood function: L(θ)
  - Maximize likelihood (difficult due to product)
    - Log-likelihood (monotonic)
    - Take ln (natural log)
    - Differentiate l(θ) with respect to the parameter (θ)
    - Set derivative 0 and solve resulting equation
    - Check this is maximum (by taking 2nd derivative) (generally we don't need, e.g., uni-modal Gaussian)

Likelihood function

Let $X_1, \ldots, X_n$ have joint pmf or pdf

$$f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_m) \tag{7.6}$$

where the parameters $\theta_1, \ldots, \theta_m$ have unknown values. When $x_1, \ldots, x_n$ are the observed sample values and (7.6) is regarded as a function of $\theta_1, \ldots, \theta_m$, it is called the **likelihood function**. The maximum likelihood estimates $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are those values of the θ_i's that maximize the likelihood function, so that

$$f(x_1, x_2, \ldots, x_n; \hat{\theta}_1, \ldots, \hat{\theta}_m) \geq f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_m) \text{ for all } \theta_1, \ldots, \theta_m$$

When the X_i's are substituted in place of the x_i's, the **maximum likelihood estimators** (mle's) result.

Poisson Example

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

For $X_1, X_2, \ldots, X_n$ iid Poisson random variables will have a joint frequency function that is a product of the marginal frequency functions, the log likelihood will thus be:

$$\begin{aligned} l(\lambda) &= \textstyle\sum_{i=1}^n (X_i log\lambda - \lambda - logX_i!) \\ &= log\lambda \textstyle\sum_{i=1}^n X_i - n\lambda - \sum_{i=1}^n logX_i! \end{aligned}$$

We need to find the maximum by finding the derivative:

$$l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^n x_i - n = 0$$

Exponential Distribution

Suppose $X_1, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ. Because of independence, the likelihood function is a product of the individual pdf's:

$$f(x_1, \ldots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdot \cdots \cdot (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \Sigma x_i}$$

The ln(likelihood) is

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

Equating $(d/d\lambda)$[ln(likelihood)] to zero results in $n/\lambda - \Sigma x_i = 0$, or $\lambda = n/\Sigma x_i = 1/\bar{x}$. Thus the mle is $\hat{\lambda} = 1/\overline{X}$;

Normal Distribution

If $X_1, X_2, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ random variables their density is written:

$$f(x_1, \ldots, x_n|\mu, \sigma) = \prod_i^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}[\frac{x_i - \mu}{\sigma}]^2\right)$$

Regarded as a function of the two parameters, μ and σ this is the likelihood:

$$\ell(\mu, \sigma) = -nlog\sigma - \frac{n}{2}log2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$$

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu)$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3}\sum_{i=1}^n (x_i - \mu)^2$$

so setting these to zero gives $\bar{X}$ as the mle for μ, and $\hat{\sigma}^2$ as the usual.

Why MLE

- Widely regarded as the best method of point estimation
- MLE provides a feasible solution
- MLE has nice asymptotic properties
  - Consistent
  - Asymptotically normal
  - Efficiency

NC STATE UNIVERSITY

Properties

- Consistency: means that having a sufficiently large number of observations n, it is possible to find the value of θ_0 with arbitrary precision.
- Asymptotically Normal: means that the estimated parameter is equal to the true parameters plus a random error that is approximately normal (given sufficient data), and the error's variance decays as 1/n

9/9/15 © Raju Vatsavai CSC-591. 27

NC STATE UNIVERSITY

Least Squares Estimators

- Instead of maximizing the likelihood of the observed outcome, we can construct an estimator by looking at the distance of the observed outcome and the outcome that we would expect with a particular parameter value.
- The value that minimizes this distance is then an estimate for the parameter.

9/9/15 © Raju Vatsavai CSC-591. 28

NC STATE UNIVERSITY

Confidence Intervals

- Properties of estimators provide valuable information for comparing and choosing an estimator, but no quality information
  - E.g., consistency guarantees that the estimated value will approach the true value in the limit, but does not give information on how close it is to the true value, given some data with a certain number of samples.
- For quantifying the quality of a particular estimate, we can compute a *confidence interval* (*CI*) around the estimate $\hat{\theta}_n$, such that this interval covers the true value θ with some high probability 1 − α. The narrower this interval, the closer we are to the true value, with high probability.

9/9/15 CSC-591. 29

NC STATE UNIVERSITY

Acknowledgements

- Vipin Kumar (Minnesota)
- Jiawei Han (UIUC)
- Hanspeter Pfister (Harvard)
- Larry Wasserman (CMU)

9/9/15 CSC-591. 30