Amy Braverman

(Jet Propulsion Laboratory)

Basic of Inference
Part 1

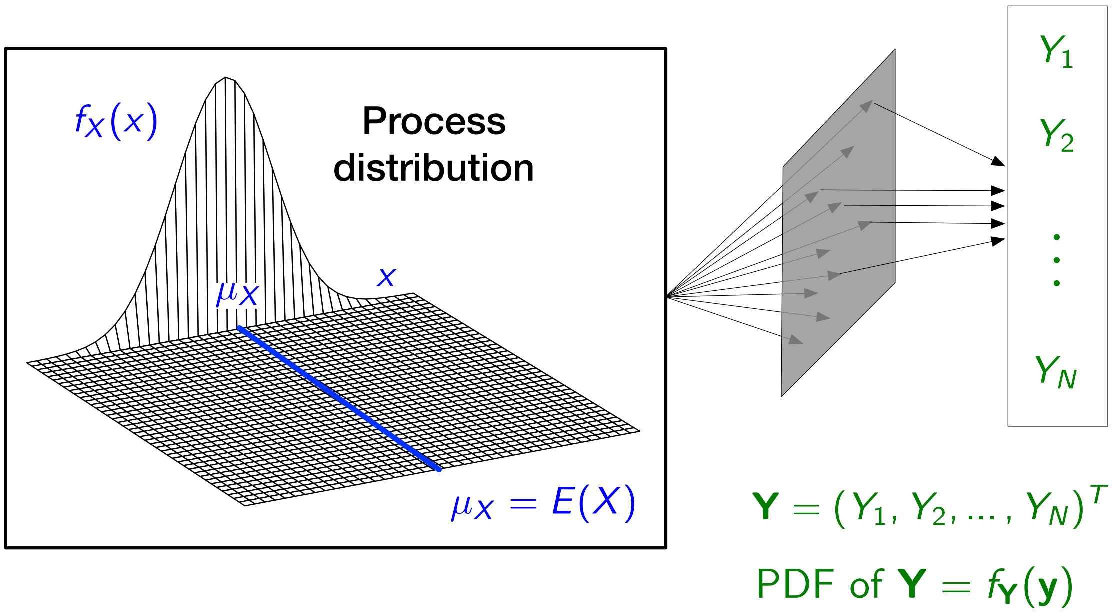
Jet Propulsion Laboratory
Keck Institute for Space Studies
Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology
Caltech

Outline

Introduce inference and the principle of maximum likelihood:

- Basic concepts of inference.
- Maximum likelihood.
- Uncertainty of the estimate.
- Desirable properties of the estimate.

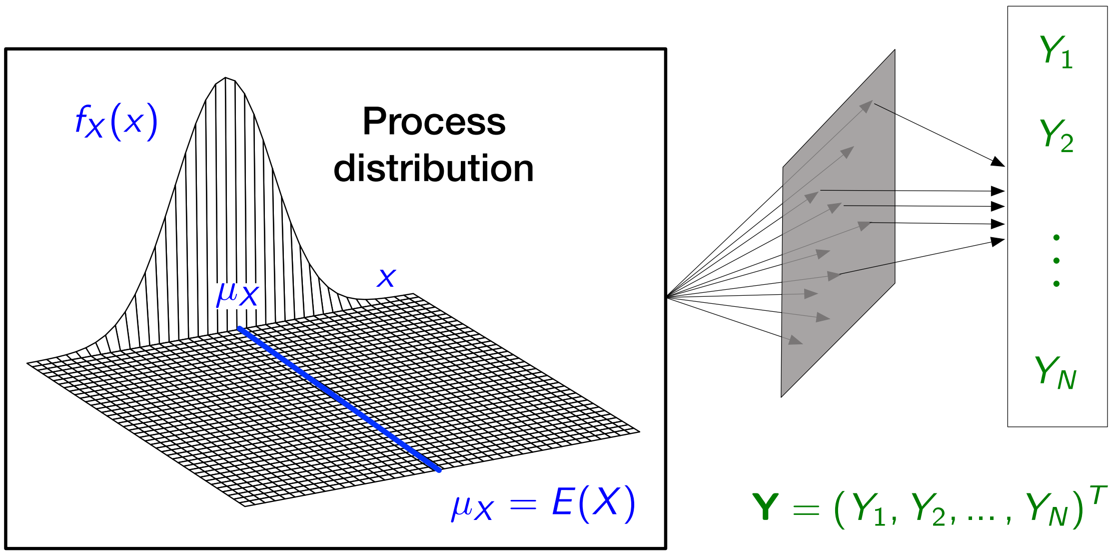
Basic concepts of inference



- X is one-dimensional.
- $\mathbf{Y}$ is N-dimensional.
- The form of $f_{\mathbf{Y}}(\mathbf{y})$ is determined by the form of $f_{\mathbf{X}}(\mathbf{x})$ and the sampling procedure.

- Parameter of interest is $\mu_X = E(X)$.
- Draw a sample of size N from the process distribution. Sample elements represented by random variables $(Y_1, Y_2, \ldots, Y_N)$.

Basic concepts of inference



$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)^T$

PDF of $\mathbf{Y} = f_{\mathbf{Y}}(\mathbf{y})$

- If each draw is independent, the Y_n's are said to be iid (independent and identically distributed).
- Recall: $P(A \cap B) = P(A)P(B)$ if A and B are independent.

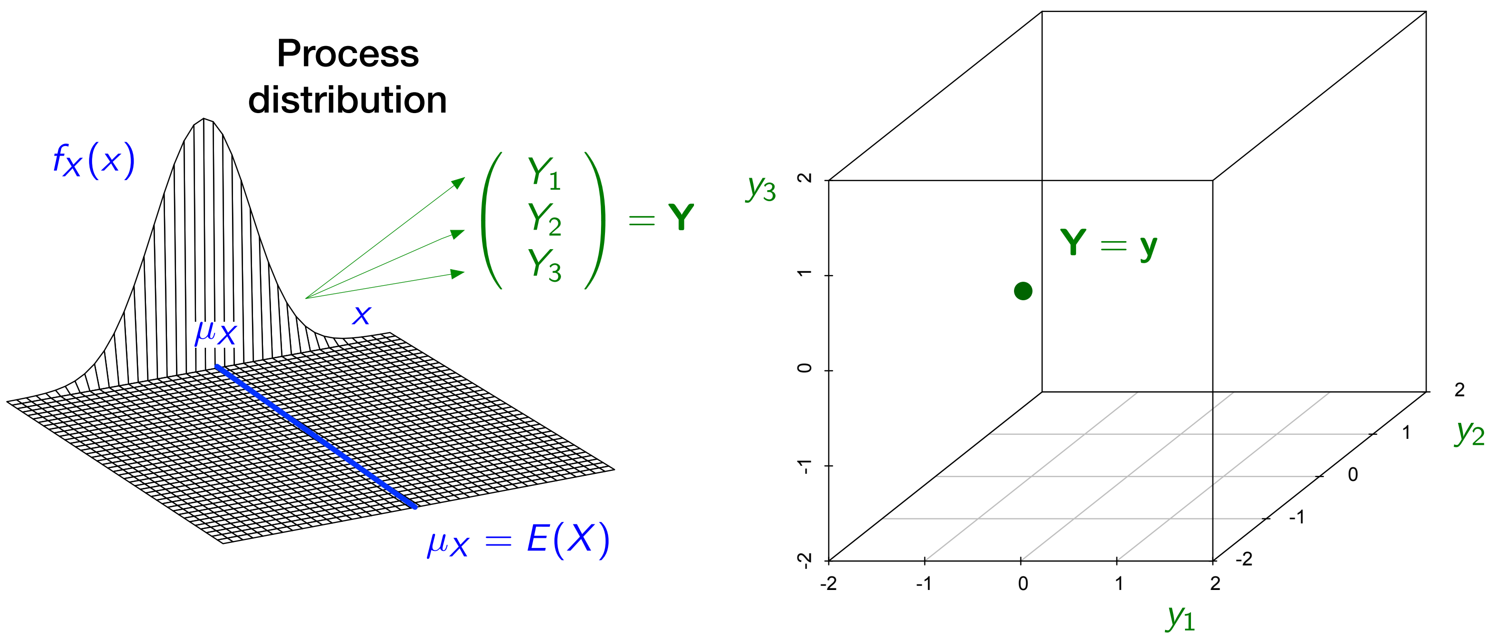
If $Y_1, Y_2, \ldots, Y_N$ are iid, then

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{Y_1}(y_1) \times f_{Y_2}(y_2) \times \ldots \times f_{Y_N}(y_N),$$

$$= f_X(y_1) \times f_X(y_2) \times \ldots \times f_X(y_N) = \prod_{n=1}^{N} f_X(y_n).$$

Basic concepts of inference

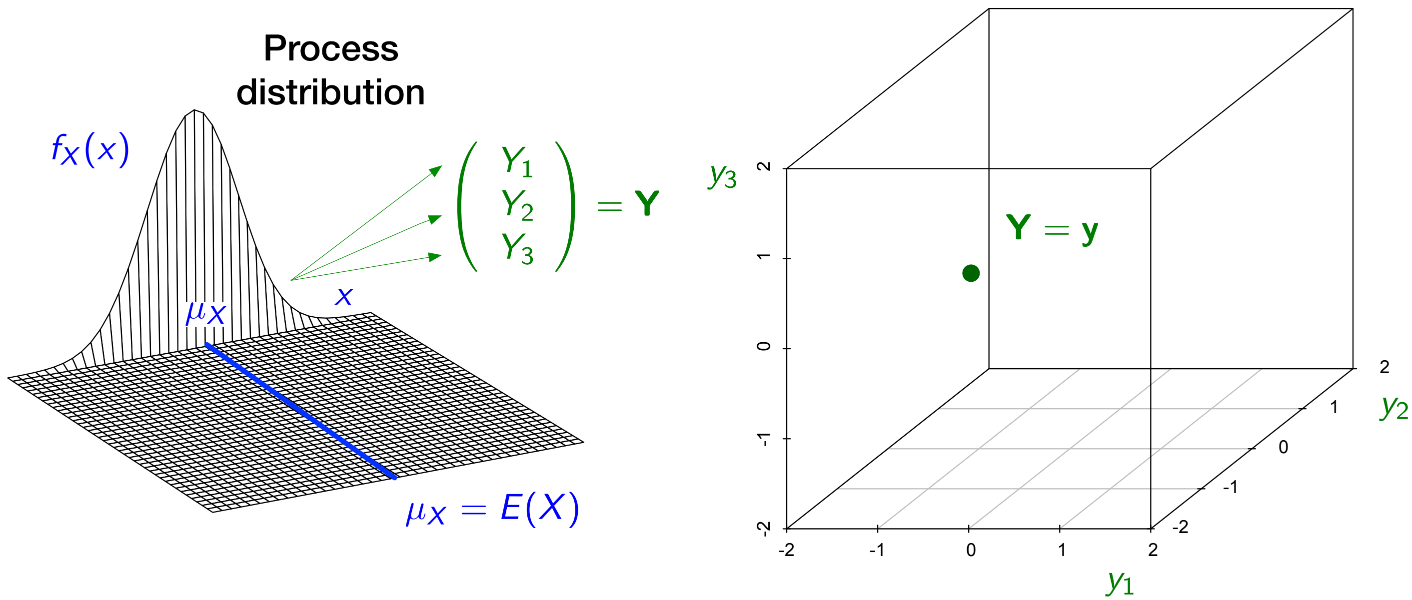
Simple example: sample of size 3 from $f_X(x)$:



- NB: before actually taking the sample, we discuss everything in terms of random variables Y_n. *After* taking the sample we say that the Y_n's have been realized and we denote them by $Y_n = y_n$.

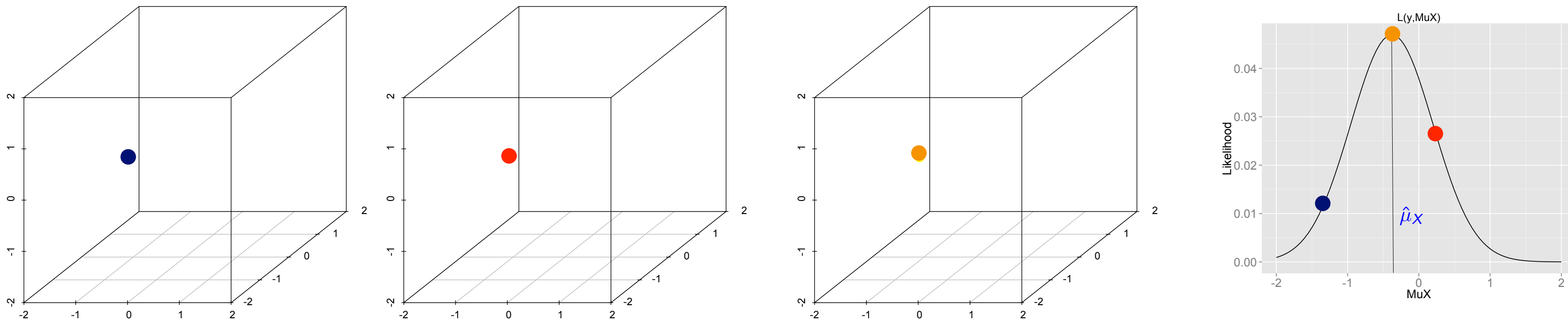
Basic concepts of inference

What can we learn about μ_X from the sample $\mathbf{Y}$?



- Think of the PDF of the sample as a function of μ_X.
- Which value of μ_X makes the value of $\mathbf{y}$ we actually realized, most probable?

Maximum likelihood



- The PDF viewed as a function of μ_X is called the likelihood function of μ_X: $L(\mu_X, \mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}, \mu_X)$.

- The <u>maximum likelihood estimate</u> (MLE) maximizes $L(\mu_X, \mathbf{y})$ for the realized sample, $\mathbf{y}$.

Example:

If $f_X(x)$ is the Gaussian distribution with expected value μ_X and variance $\sigma_X^2 = 1$,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu_X)^2}{2} \right\},$$

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(y_1-\mu_X)^2}{2} \right\} \times \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(y_2-\mu_X)^2}{2} \right\}$$
$$\times \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(y_3-\mu_X)^2}{2} \right\} = \left[\frac{1}{\sqrt{2\pi}} \right]^3 \exp\left\{ -\frac{1}{2} \sum_{n=1}^{3} (y_n-\mu_X)^2 \right\},$$

$= f_{\mathbf{Y}}(\mathbf{y}, \mu_X)$ to emphasize functional dependence on μ_X.

Example (cont'd):

- In cases where we know the form of $f_{\mathbf{Y}}(\mathbf{y})$, it's possible we can solve for the MLE analytically by finding where the derivatives of $L(\mu_X, \mathbf{y})$ equal zero.

- Often easier to solve for the maximum of $\log L(\mu_X, \mathbf{y})$:

$$L(\mu_X, \mathbf{y}) = \left[\frac{1}{\sqrt{2\pi}}\right]^3 \exp\left\{-\frac{1}{2}\sum_{n=1}^{3}(y_n - \mu_X)^2\right\},$$

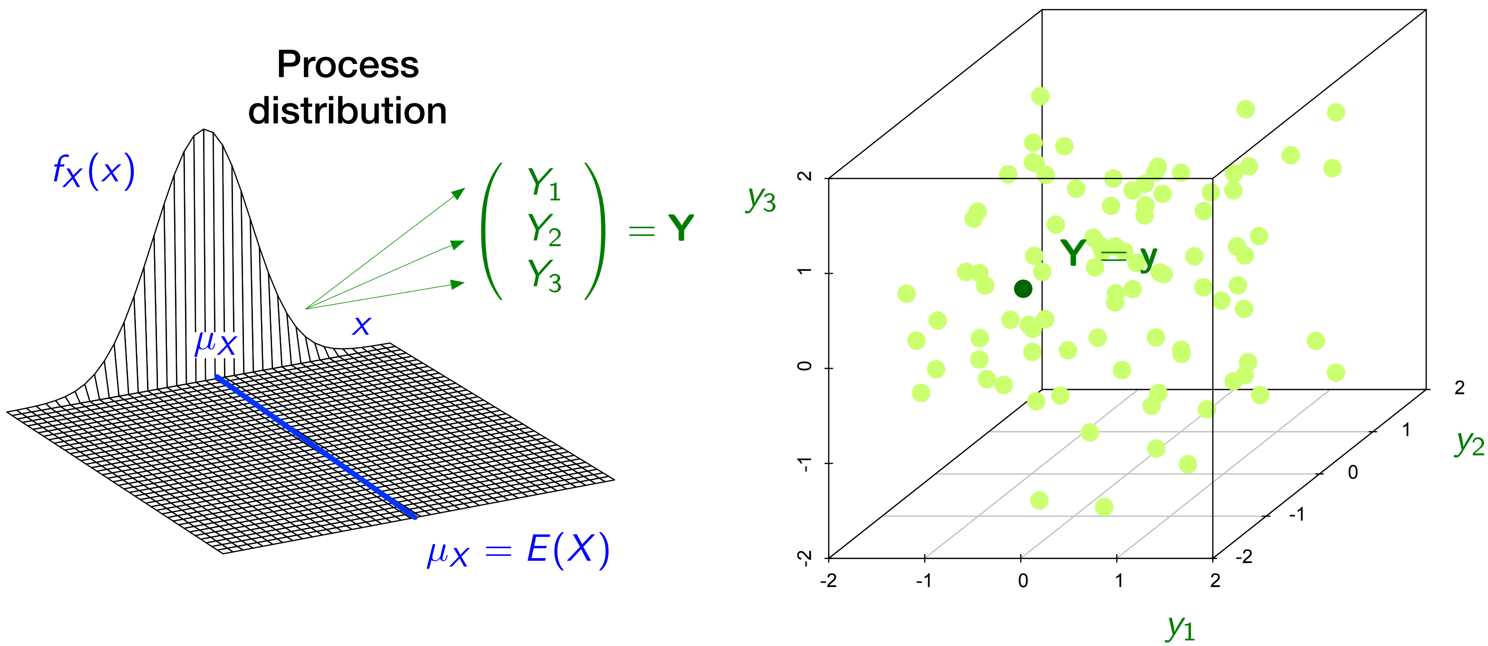
$$\log L(\mu_X, \mathbf{y}) = \frac{3}{2}\log 2\pi - \frac{1}{2}\sum_{n=1}^{3}(y_n - \mu_X)^2,$$

$$\frac{\partial}{\partial \mu_X}\log L(\mu_X, \mathbf{y}) = -\sum_{n=1}^{3}(y_n - \mu_X) = 0 \Rightarrow \sum_{n=1}^{3} y_n = 3\mu_X \Rightarrow \frac{1}{3}\sum_{n=1}^{3} y_n = \hat{\mu}_X,$$

$$\Rightarrow \hat{\mu}_X = \bar{y}.$$

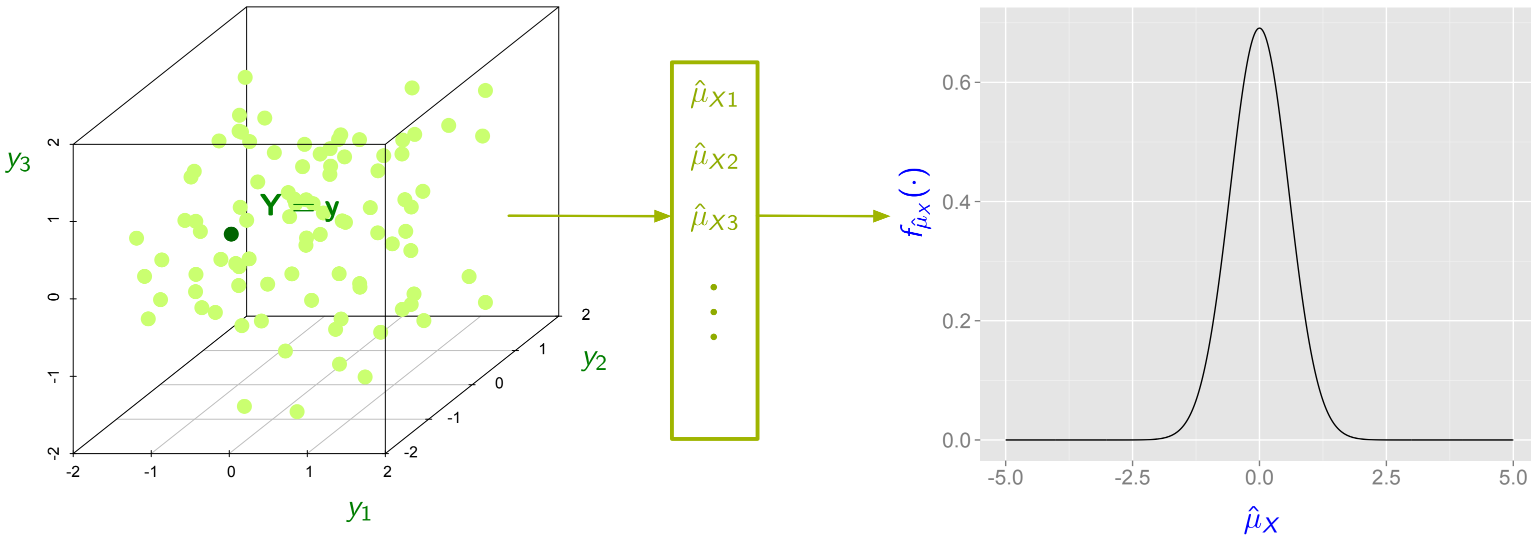
Uncertainty of the estimate

What is the uncertainty of the estimate?



- There are many other samples we *could* have gotten. There is a $\hat{\mu}_X$ associated with each one of them.
- The distribution over all possible samples is a full description of the uncertainty of $\hat{\mu}_X$ as an estimate of μ_X.

Desirable properties of the estimate



- The PDF of a statistic ($\hat{\mu}_X$) is called its sampling distribution.
- Desirable properties:
  1. Unbiasedness: $E(\hat{\mu}_X - \mu_X) = 0$,
  2. Minimum variance: $E(\hat{\mu}_X - E(\hat{\mu}_X))^2 \leq E(\tilde{\mu}_X - E(\tilde{\mu}_X))^2$ for any other estimate, $\tilde{\mu}_X$.

An estimate with both properties is said to be MVUE.

Desirable properties

We can propose *any* function of the sample, $g(\mathbf{y}) = \hat{\theta}$, as an estimator of θ.

- Unless we are careful, $\hat{\theta}$ may not have desirable statistical properties (unbiasedness, minimum variance).
- Typically, we choose $g(\cdot)$ so that $\hat{\theta}$ is the "sample version" of θ.
- When θ is a moment of $f_X(x)$ this usually isn't too far off, but can be suboptimal.

Think of $g(\mathbf{y}) = \hat{\theta}$ as *dimension reduction*: its sampling distribution lives in a low dimensional space.

If $\hat{\theta}$ contains the all the information about θ that is contained in $\mathbf{y}$, then $\hat{\theta}$ is called a sufficient statistic for θ.

Desirable properties

Computability: it may be hard to solve for the parameter value that makes the sample most likely in more general situations:

- Unknown or non-Gaussian PDF's,
- Dependent sampling,
- $N >> 3$,
- Arbitrary process distribution parameters of interest, θ.

Bibliography

- *Mathematical Statistics and Data Analysis* by John Rice, Wadsworth, 1995.
- *Statistical Inference* by George Casella and Roger L. Burger, Wadsworth, 1990.

Next

So how do we carry out inference then? The next module presents some solutions.