

Lab 3b: Distribution of the mean

Outline

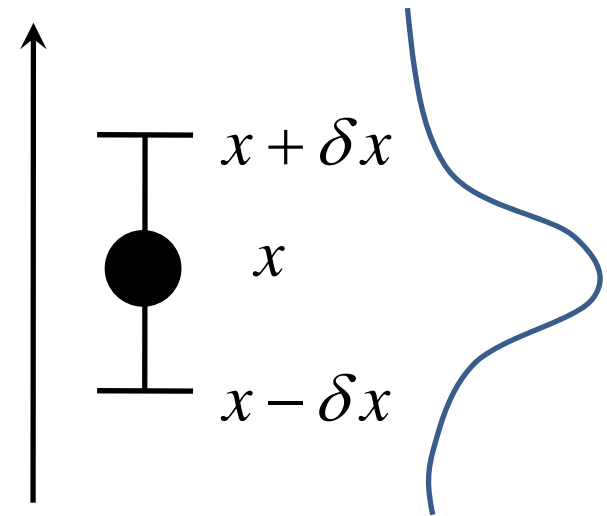
- Distribution of the mean: Normal (Gaussian)
 - Central Limit Theorem
- “Justification” of the mean as the best estimate
- “Justification” of error propagation formula
 - Propagation of errors
 - Standard deviation of the mean

640:104 - Elementary Combinatorics and Probability
960:211-212. Statistics I, II

Finite number of repeated measurements

The average value is the best estimation of true mean value, while the square of standard deviation is the best estimation of true variance.

$$X \approx \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$
$$\sigma^2 \approx \sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$



The *same* set of data also provides the best estimation of the variance of the mean.

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{N} \quad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$$

The Central Limit Theorem

The mean is a sum of N **independent** random variables:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

The PDF of the mean is the PDF of the **sum** of N independent random variables.

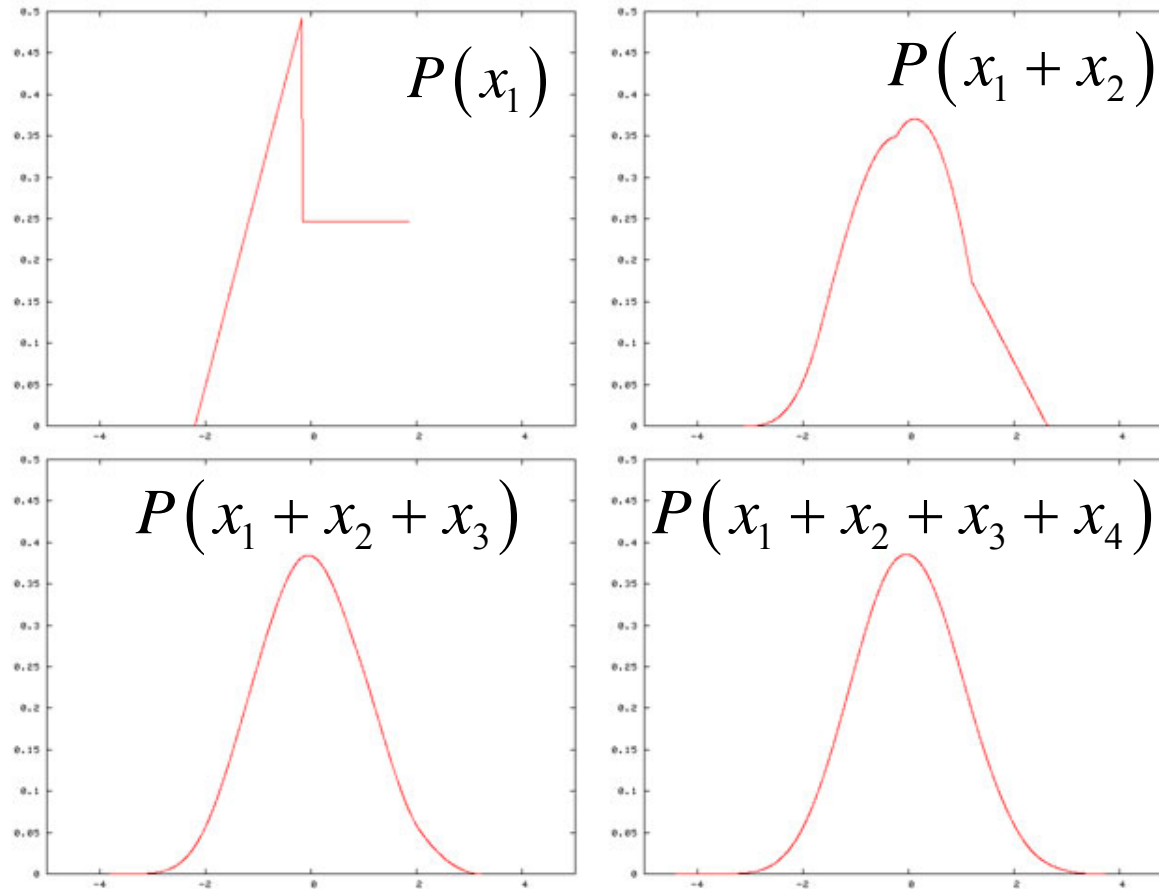
In probability theory, the **central limit theorem (CLT)** states that, given conditions, the mean of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed.

$$P(\bar{x}) \xrightarrow{N \rightarrow \infty} P_G(\bar{x})$$

*In practice, N just needs to be large enough.

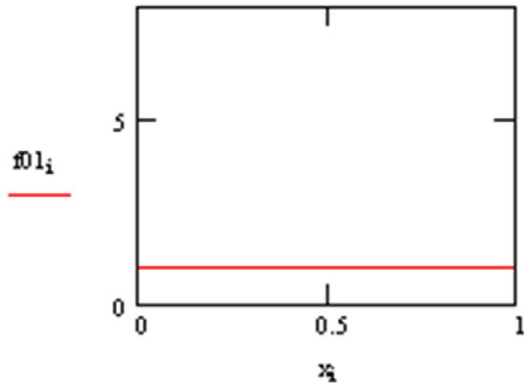
Probability Distribution Function of the mean

An example of CLT

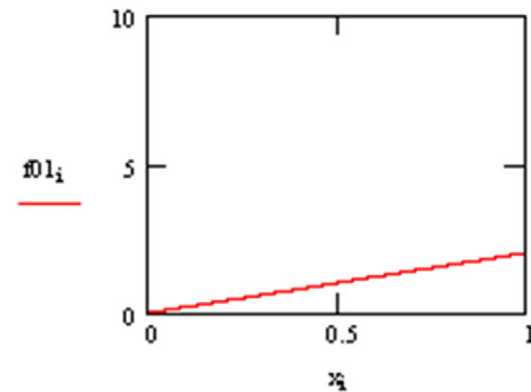


For this example $N > 4$ is “enough”.

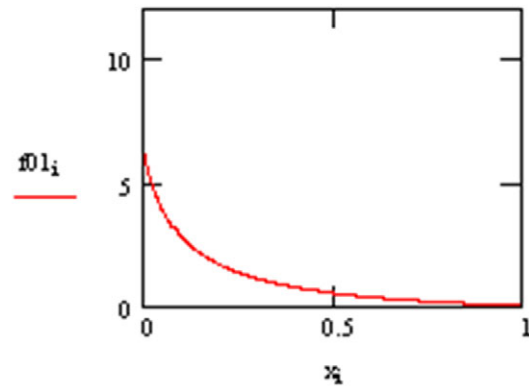
More examples of CLT



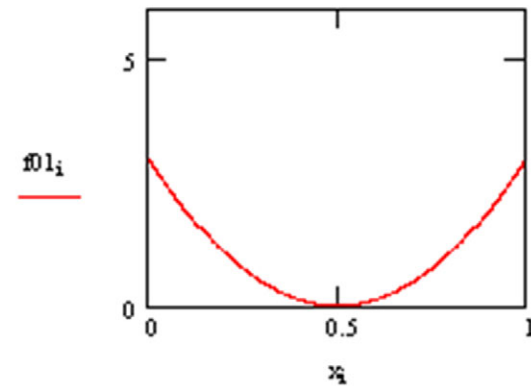
NonNormal Distribution of X



NonNormal Distribution of X



NonNormal Distribution of X



NonNormal Distribution of X

http://www.statisticalengineering.com/central_limit_theorem.html

The significant consequence of CLT

Almost any repeated measurements of independent random variables can be *effectively* described by Normal (Gaussian) distribution.

E.g. N measurements x_i ($i=1, 2, \dots, N$) of random variable x can be divided to m subsets of measurements $\{n_j\}, j=1, 2, \dots, m$.

$$N = n_1 + n_2 + \dots + n_m$$

Define the means of each subset as new measurements of new random variable y . If n_j are large enough, the PDF of y_j is approximately a Normal (Gaussian) one.

$$y_j = \bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i$$

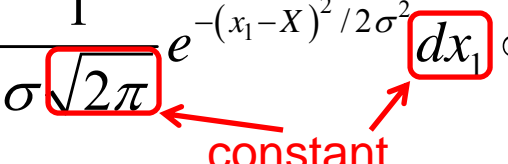
“Justification” of the mean as the best estimate

Assuming all the measured values follow a normal distribution $G_{X,\sigma}(x)$ with **unknown** parameters X and σ .

$$G_{X,\sigma} = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2}$$

The probability of obtaining value x_1 in a small interval dx_1 is:

$$\text{Prob}(x \text{ between } x_1 \text{ and } x_1 + dx_1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_1-X)^2/2\sigma^2} dx_1 \propto \frac{1}{\sigma} e^{-(x_1-X)^2/2\sigma^2}$$


constant

Similarly, the probability of obtaining value x_i ($i=0, \dots, N$) in a small interval dx_i is:

$$\text{Prob}(x \text{ between } x_i \text{ and } x_i + dx_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-X)^2/2\sigma^2} dx_i \propto \frac{1}{\sigma} e^{-(x_i-X)^2/2\sigma^2}$$

Justification of the mean as best estimate (cont.)

The probability of N measurements with values: x_1, x_2, \dots, x_N :

$$\begin{aligned}\text{Prob}_{X,\sigma}(x_1, \dots, x_N) &= \text{Prob}(x_1) \times \text{Prob}(x_2) \times \dots \times \text{Prob}(x_N) \\ &\propto \frac{1}{\sigma^N} e^{-\sum_{i=1}^N (x_i - X)^2 / 2\sigma^2}\end{aligned}$$

Here x_1, x_2, \dots, x_N are **known** results because they are measured values. On the other hand, X and σ are **unknown** parameters. We want to find out the best estimate of them from $\text{Prob}_{X,\sigma}(x_1, x_2, \dots, x_N)$. One of the reasonable procedures is the *principle of maximum likelihood*, i.e. maximizing $\text{Prob}_{X,\sigma}(x_1, x_2, \dots, x_N)$ respect to $X \rightarrow$ minimizing

$$\begin{aligned}&\sum_{i=1}^N (x_i - X)^2 / 2\sigma^2 \\ \Rightarrow \sum_{i=1}^N (x_i - X) &= 0 \Rightarrow (\text{best estimate of } X) = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x}\end{aligned}$$

Best estimate of variance

Similarly, we could maximize $\text{Prob}_{X,\sigma}(x_1, x_2, \dots, x_N)$ respect to σ to get the best estimate of σ .

$$-\frac{N}{\sigma^{N+1}} \times e^{-\sum (x_i - X)^2 / 2\sigma^2} + \frac{1}{\sigma^N} \times \left[-\sum_{i=1}^N (x_i - X)^2 \right] \times \left(-\frac{1}{\sigma^3} \right) \times e^{-\sum (x_i - X)^2 / 2\sigma^2} = 0$$

$$\Rightarrow -N + \frac{1}{\sigma^2} \sum_{i=1}^N (x_i - X)^2 = 0$$

$$\Rightarrow (\text{best estimate of } \sigma) = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - X)^2}$$

However, we don't really know X . In practice, we have to replace X with the mean value, which reduces the accuracy. To account for this factor, we have to replace N with $N-1$.

$$(\text{best estimate of } \sigma) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

Justification of error propagation formula (I)

Measured quantity plus a constant:

$$q = x + A \quad \Rightarrow \quad \bar{q} = \bar{X} + A, \quad \sigma_q = \sigma_x$$

Measured quantity multiplies a fixed #:

$$q = Bx \quad \Rightarrow \quad \bar{q} = B \cdot \bar{X}, \quad \sigma_q = B\sigma_x$$

Sum of two measured quantities: $q = x + y$

$$P(q) = \int_{q=x+y} P(x) P(y) dx dy = \int_{-\infty}^{\infty} P(x) P(q-x) dx$$

Assuming $X=Y=0$,

$$P(x) = \frac{e^{-x^2/2\sigma_x^2}}{\sigma_x \sqrt{2\pi}} \quad P(y) = \frac{e^{-y^2/2\sigma_y^2}}{\sigma_y \sqrt{2\pi}}$$

Justification of error propagation formula (II)

Sum of two measured quantities: $q = x + y$

$$\begin{aligned} P(q) &= \int_{-\infty}^{\infty} P(x) P(q-x) dx = \int_{-\infty}^{\infty} \frac{e^{-x^2/2\sigma_x^2}}{\sigma_x \sqrt{2\pi}} \frac{e^{-(q-x)^2/2\sigma_y^2}}{\sigma_y \sqrt{2\pi}} dx \\ &= \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma_x^2} - \frac{(q-x)^2}{2\sigma_y^2}} dx \\ &= \dots \\ &= \frac{1}{\sqrt{2\pi(\sigma_x^2 + \sigma_y^2)}} e^{-\frac{q^2}{2(\sigma_x^2 + \sigma_y^2)}} \end{aligned}$$

$$\sigma_q = \sqrt{\sigma_x^2 + \sigma_y^2}$$

Justification of error propagation formula (III)

The general case: $q = q(x)$.


Assuming σ_x are small enough so that we are concerned only with values of x close to X .

Using Taylor expansion:

$$q(x) \approx q(X) + \left(\frac{\partial q}{\partial x} \right) (x - X)$$

$$= \left[q(X) - \left(\frac{\partial q}{\partial x} \right) X \right] + \left(\frac{\partial q}{\partial x} \right) x$$

constant



$$\sigma_q = \sqrt{\left(\frac{\partial q}{\partial x} \sigma_x \right)^2} = \left| \frac{\partial q}{\partial x} \sigma_x \right|$$

Justification of error propagation formula (IV)

The more general case: $q = q(x, y)$.

Assuming σ_x and σ_y are small enough so that we are concerned only with values of x close to X and y close to Y .

Using Taylor expansion:

$$\begin{aligned} q(x, y) &\approx q(X, Y) + \left(\frac{\partial q}{\partial x}\right)(x - X) + \left(\frac{\partial q}{\partial y}\right)(y - Y) \\ &= \left[q(X, Y) - \left(\frac{\partial q}{\partial x}\right)X - \left(\frac{\partial q}{\partial y}\right)Y \right] + \left(\frac{\partial q}{\partial x}\right)x + \left(\frac{\partial q}{\partial y}\right)y \end{aligned}$$

$$\sigma_q = \sqrt{\left(\frac{\partial q}{\partial x} \sigma_x\right)^2 + \left(\frac{\partial q}{\partial y} \sigma_y\right)^2}$$

Standard deviation of the mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Assuming they are normally distributed and have the same μ and σ_x .

$$\Rightarrow \sigma_{\bar{x}} = \sqrt{\left(\frac{\partial q}{\partial x_1} \sigma_x\right)^2 + \left(\frac{\partial q}{\partial x_2} \sigma_x\right)^2 + \cdots + \left(\frac{\partial q}{\partial x_N} \sigma_x\right)^2}$$

$$\Rightarrow \sigma_{\bar{x}} = \sqrt{\left(\frac{1}{N} \sigma_x\right)^2 + \left(\frac{1}{N} \sigma_x\right)^2 + \cdots + \left(\frac{1}{N} \sigma_x\right)^2} = \sqrt{\frac{1}{N} \sigma_x^2} = \frac{\sigma_x}{\sqrt{N}}$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}$$

Lab 3b: test the CLT

1. One run of decay rate ($R=N/\Delta t$) measurement
 - a. Sampling rate: 0.5 second/sample (Δt)
 - b. Run time: 30 seconds (# of measurements: $n=60$)
 - c. calculate \bar{N} , σ_N , $\sigma_{\bar{N}}$, plot histogram.
2. Repeat the 30-sec run 20 times (a set of runs)
 - a. Remember to save data
 - b. Plot histograms of 21 means with proper bin widths
 - c. **Make sure you answer all the questions**
3. Sample size ($n=T/\Delta t$) dependence
 - Keep the same configuration, vary run time
 - Record \bar{N} and σ_N , plot \bar{N} vs. $\log n$ with error bar = $\sigma_{\bar{N}}$.

* “Origin” (OriginLab®) is more convenient than Matlab.

<http://www.physics.rutgers.edu/rcem/~ahogan/326.html>