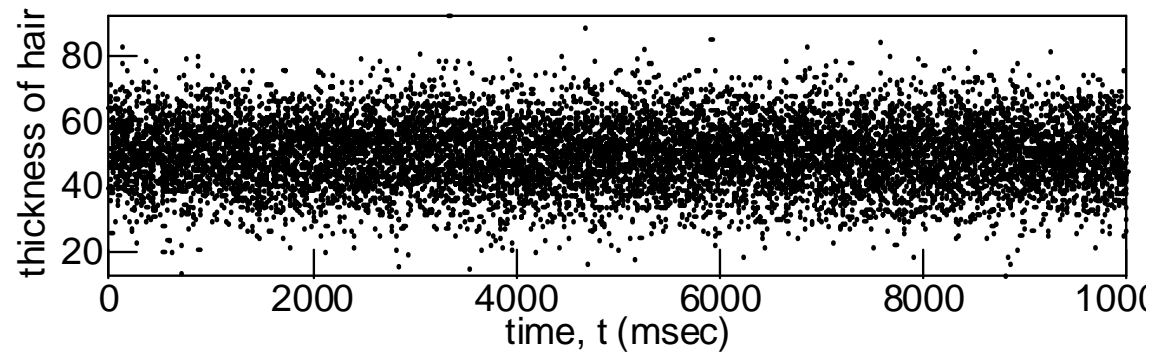


What to do with a large sample of data

- Repeated measurements yield a distribution of data



- Spread in the data has two possible origins
 - fluctuations in the measured quantity
 - random error in the measurement
- Simple measures include the **mean**, **standard deviation**, **standard error of the mean**, and **correlation**.

Mean and Standard Deviation

Defined for any distribution (Gaussian or not)

Mean: $\langle x \rangle = \frac{1}{N} \sum_{j=1}^N x_j$

Mean-square: $\langle x^2 \rangle = \frac{1}{N} \sum_{j=1}^N x_j^2$ (a.k.a. 'second moment')

Variance*: $\sigma_x^2 = \text{average}[(x - \langle x \rangle)^2]$

$$\sigma_x^2 = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 - 2x\langle x \rangle + \langle x \rangle^2 \rangle = \langle x^2 \rangle - 2\langle x \rangle^2 + \langle x \rangle^2$$

$$\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2$$

(mean of the square minus the square of the mean)

* (sometimes the normalization is different: $\sigma_x = \frac{1}{N-1} \sum_{j=1}^N (x_j - \langle x \rangle)^2$.)

Standard deviation: σ_x

A good measure of uncertainty.

Roughly, the half-width of the distribution of values, $P(x)$

Probability distribution function, $P(x)$

$P(x) dx$ = probability that a measurement of x will provide a value in the range of $x \pm dx$. Units of P are $[x]^{-1}$

Quantum mechanics

Heisenberg Uncertainty Principle:

$$\Delta p_x \Delta x \geq \hbar/2$$

But what is Δx ??

$$\Delta x = \sqrt{\langle x^2 \rangle - \langle x \rangle^2},$$

where $\langle x \rangle$ is the expectation value, $\int_{-\infty}^{\infty} \psi^*(x) x \psi(x) dx$,

and $\langle x^2 \rangle = \int_{-\infty}^{\infty} \psi^*(x) x^2 \psi(x) dx$.

(mean, standard deviation, etc. appear in many places)

The Gaussian Probability Distribution Function

“bell curve” or “*normal*” distribution.

The Gaussian probability distribution function (pdf):

$$P(x)dx = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-\langle x \rangle)^2}{2\sigma_x^2}} dx$$

$\langle x \rangle$ = mean (same value as mode and median)

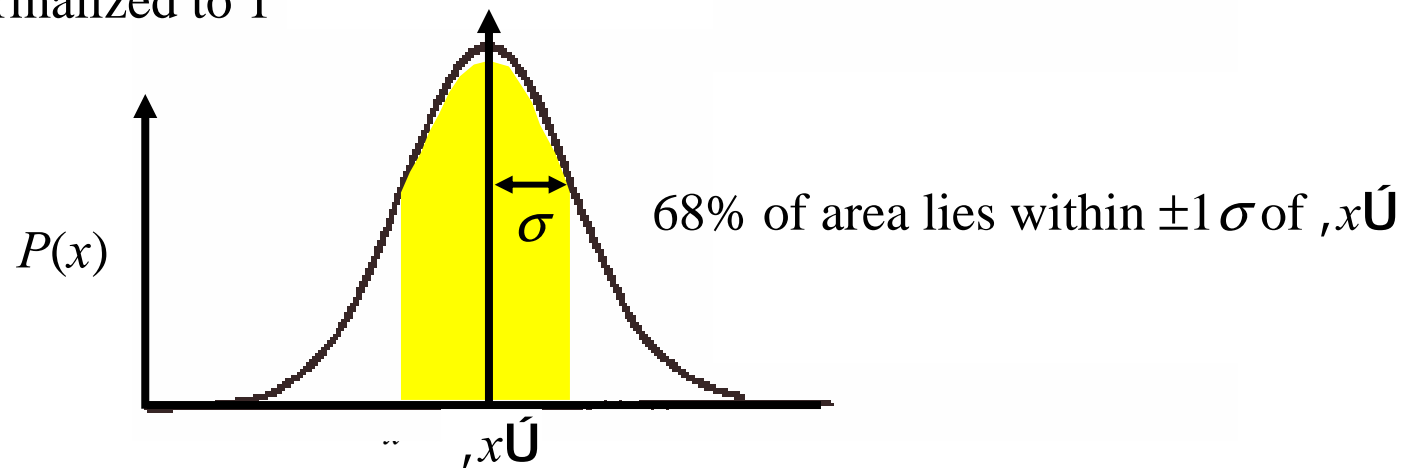
σ_x^2 = variance

σ_x = standard deviation (68% probability that the next measurement of x will be in the range $\langle x \rangle \pm \sigma_x$.)



Karl Friedrich Gauss (1777-1855)

- Normalized to 1



A Bit of Mathematical Detail:

- Probability (P) that x is in the range $[a, b]$ is

$$P(a < x < b) = \int_a^b P(x) dx = \frac{1}{\sigma_x \sqrt{2\pi}} \int_a^b e^{-\frac{(x-\langle x \rangle)^2}{2\sigma_x^2}} dx$$

- The integral for arbitrary a and b cannot be evaluated analytically.
(It is related to the ‘error function’)

We can associate a probability for a measurement to be $|m\sigma_x|$ from the mean just by calculating the area outside of this region.

<u>m</u>	<u>Prob. that x is outside $\pm m\sigma_x$</u>
0.67	0.5
1	0.32
2	0.05
3	0.003
4	0.00006

Why are Gaussians so common?

A crude statement of Central Limit Theorem:

The result of adding many independent (uncorrelated) numbers tends to follow a Gaussian distribution.

a) random walk (displacement = sum of steps)

b) # radioactive decays during a long time interval

c) average value of 10 throws of dice.

A more exact statement:

Let Y_1, Y_2, \dots, Y_n be a sequence of n independent random variables, each with the same probability distribution.

$$\text{Let } x = \sum_{i=1}^n Y_i$$

If $\langle x \rangle$ and σ_x are finite, then as $n \rightarrow \infty$, the distribution of x -values becomes a Gaussian:

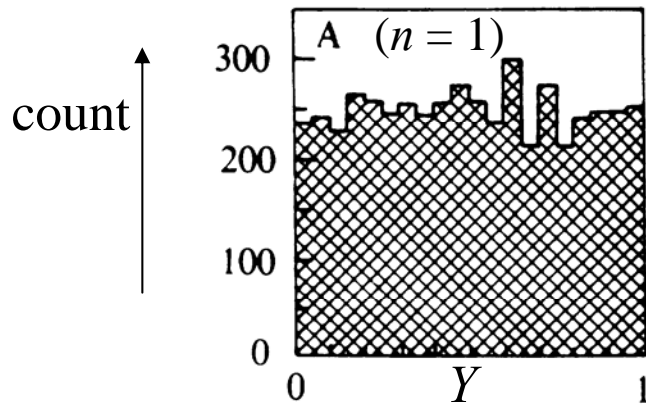
$$P(x)dx = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-\langle x \rangle)^2}{2\sigma_x^2}} dx$$

(but in practice, P is Gaussian even for modest n)

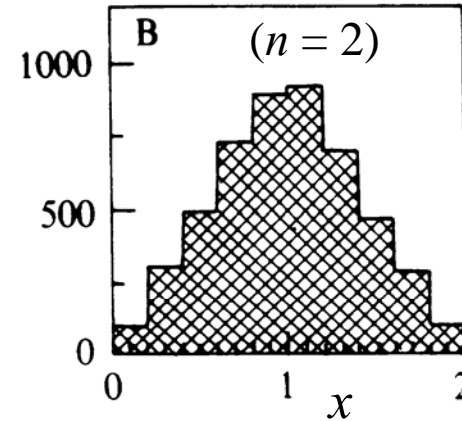
Illustration of the Central Limit Theorem

Histograms:

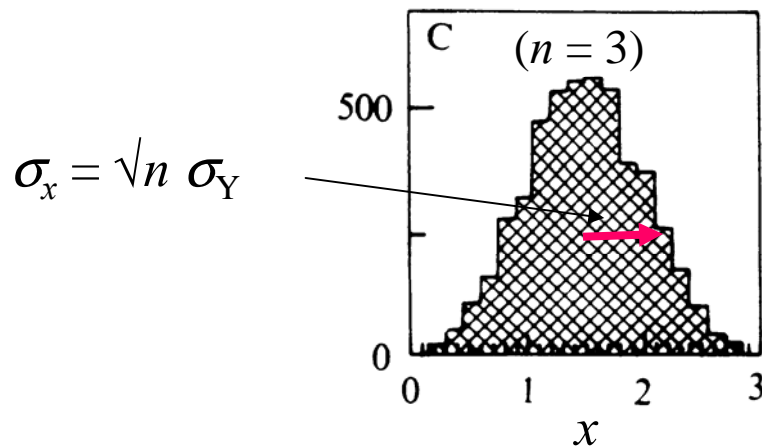
A) 5000 random numbers



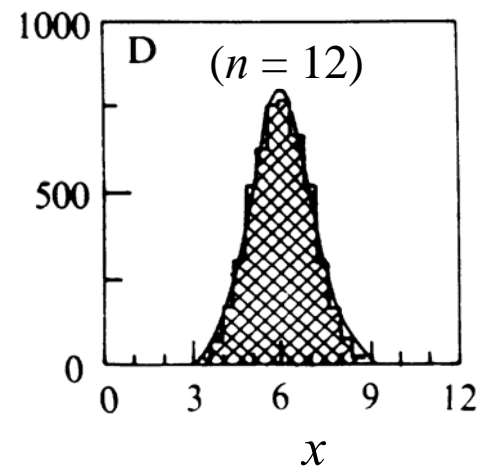
B) 5000 pairs: $x = Y_1 + Y_2$



C) 5000 triplets: $x = Y_1 + Y_2 + Y_3$



D) 5000 12-plets



$$\sigma_x = \sqrt{n} \sigma_Y$$

Another example: A Random Walk, x and σ_x

Take steps of 1m to right or left. That is, $Y_i = \pm 1$

Q: What is the displacement, x , after n steps?

A: $x = \sum_{i=1}^n Y_i$, but this is not terribly useful because the answer is different for each measurement.

A better question: “what’s the average of x over many 25-step walks?”

$$\langle x \rangle = \left\langle \sum_{i=1}^n Y_i \right\rangle = \sum_{i=1}^n \langle Y \rangle_i = 0$$

Q: If we measure x again, how close will it be to the average?

A: Within the std. deviation of x (with 68% prob.)

$$\begin{aligned} \sigma_x^2 &= \langle x^2 \rangle - \langle x \rangle^2 = \left\langle \sum_{i=1}^n Y_i^2 \right\rangle = \left\langle \sum_{i=1}^n Y_i \sum_{j=1}^n Y_j \right\rangle \\ &= \left\langle \sum_{i=1}^n Y_i \sum_{j=1}^n Y_j \right\rangle = \sum_{\substack{i,j=1 \\ (i \neq j)}}^n Y_i Y_j + \sum_{\substack{i,j=1 \\ (i=j)}}^n Y_i Y_i = n \end{aligned}$$

=0 because different steps are uncorrelated.

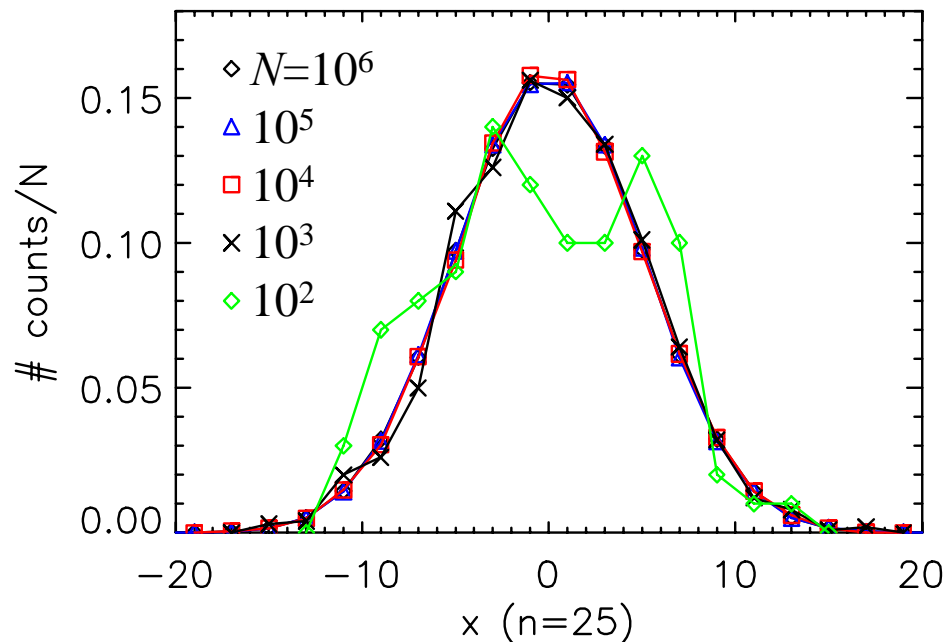
More generally, $\sigma_x = \sqrt{n} \sigma_Y$

Also apparent from the rule for propagating error when adding.

...Random Walks: a numerical example

$x = \sum_{i=1}^n Y_i$, where $Y_i = \pm 1$. Here I chose $n = 25$ (i.e., 25 steps per trajectory)

Calculate x many times (N times) and histogram the x -values:



N	the mean, \bar{x}	σ_x^2	Std error of mean
100	-0.35	30.69	0.55
10^3	0.087	25.63	0.16
10^4	0.037	24.86	0.051
10^5	0.0032	24.81	0.016
10^6	-0.0039	25.06	0.005

Conclusions:

when $N \sim 10^3$ or larger, $P(x)$ looks Gaussian (for $n=25$).

$\sigma_x^2 \sim n\sigma_Y^2 = 25$, as expected.

$\bar{x} = 0$, within the std. error of the mean.

Standard error of the mean

- **Measure $\langle x \rangle$** from N measurements of x (*i.e.*, many random walks):

$$\langle x \rangle = \frac{1}{N} \sum_{j=1}^N x_j$$

- How precise is this answer? Would another N trajectories give the same $\langle x \rangle$?

Std. error of the mean of $x = \frac{\sigma_x}{\sqrt{N}} = \frac{\text{[std. dev. of one measurement of } x\text{]}}{\sqrt{\text{[number of independent measurements of } x\text{]}}^*$

What does it mean?

The next time someone measures x (just one measurement), there is a 68% probability that it lies in the range $\langle x \rangle \pm \sigma_x$

The next time someone measures $\langle x \rangle$ there is a 68% probability that the answer will lie in the range $\langle x \rangle \pm \sigma_x/\sqrt{N}$.

To measure $\langle x \rangle$ the more measurements the better!

- (This formula requires that the measurements must be uncorrelated and that N is large.)

Correlations

The **correlation of two parameters** (a, b) has a specific meaning in statistics:

$$C_{ab} = \frac{\frac{1}{N} \sum_{i=1}^N (a_i - \langle a \rangle)(b_i - \langle b \rangle)}{\sigma_a \sigma_b} \quad 1 \quad C_{ab} \quad -1$$

If $C_{ab} = 1$, they are perfectly correlated.

(e.g. diameter and circumference of perfect circles. Or C_{aa})

If $C_{ab} = 0$, they are uncorrelated, or completely independent.

(e.g. number of letters in a person's name and the day of the month on which they were born.)

If $C_{ab} = -1$, they are anticorrelated: if one increases, the other decreases.

(e.g., the heights of two children on a see-saw.)

ID# of person	Height (inches)	Weight (lb)
1	65.7	124.5
2	52.9	116.9
3	64.1	204.0
4	79.1	185.5
5	72.0	190.6

Example: does a person's weight correlate with their height?

a = height.

b = weight.

From the given (fake) data, I get $C_{ab} = 0.50$
(yes, they are correlated, but not perfectly)