Introduction to Statistics for Quantitative Analysis

To support P113, P121, P114, P122, P141 labs

Standard deviations like you've never seen them before!

Reviews of Intro to Stats:

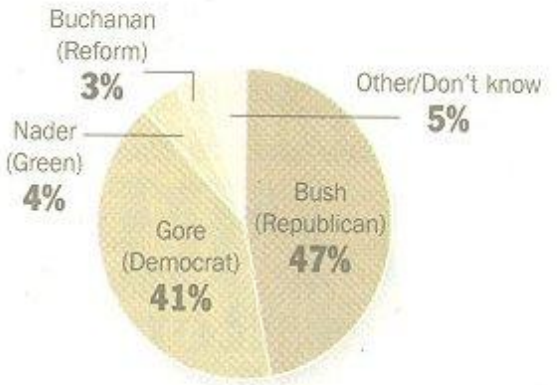
"Spellbinding! The talk at all the parties!" *Jason Huber, 2nd floor Hoeing RA*

"Brilliant! Fun for geeks of all ages!" *Rhett Butler, CNN reviews*

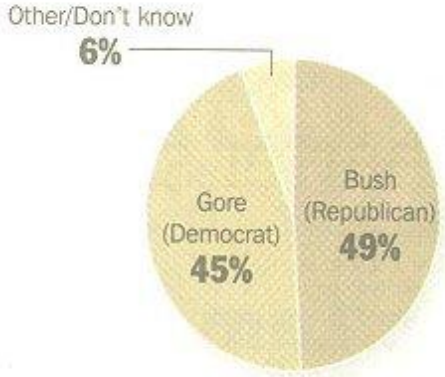
"A seminal work. Brought to you by the author of Solutions to P113 Problem Set #2." *Leona Helmsley, The Shopping Channel*

THE OHIO POLL: Bush leads overall, head-to-head races

Bush leads by six percent when matched against Gore, Nader and Buchanan ...



... and Bush leads Gore by four percent in a head-to-head match-up.



SOURCE: Ohio Poll of 537 likely Ohio voters conducted July 5-13 by the Institute for Policy Research. Margin of error, plus or minus 4.2 percentage points.

1988 US Presidential election

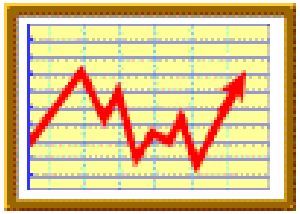
	Month 1	Month 2	
Bush	42%	41%	
Dukakis	40%	43%	
Undecided	18%	16%	±4%

Headline: Dukakis surges past Bush in polls!

Is statistics relevant to you personally?

Global Warming

Analytical medical diagnostics



Effect of EM radiation

What kinds of things can you measure quantitatively?

What kinds of things can you measure qualitatively?

What is the difference between a qualitative and quantitative measurement?

Which of these types of measurement are important in science?

In so far as possible, physics is exact and quantitative … though you will repeatedly see mathematical approximations made to get at the qualitative essence of phenomena.

$$2\frac{1}{2}$$

A quantitative measurement is meaningless without a unit and error.

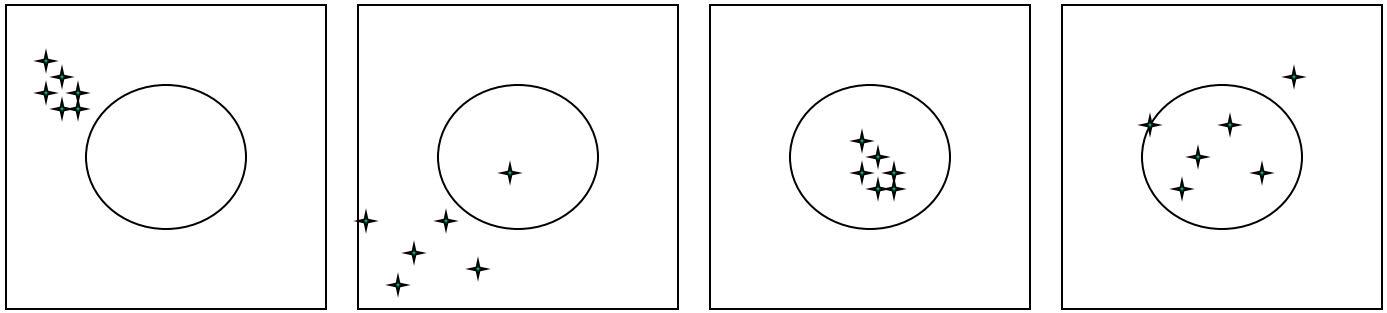
Accuracy:

A measure of closeness to the “truth”.

Precision:

A measure of reproducibility.

Accuracy vs. precision



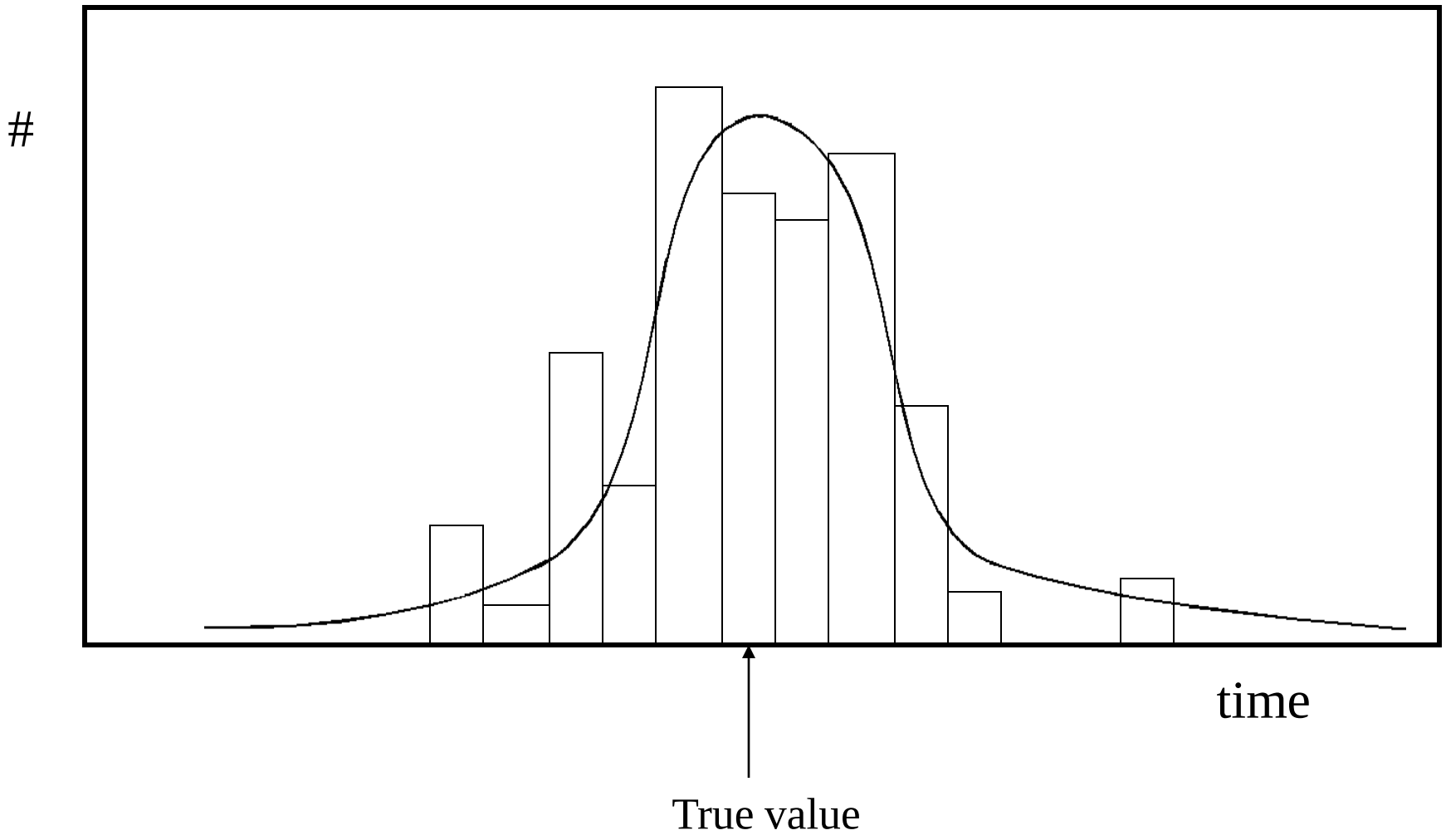
accurate

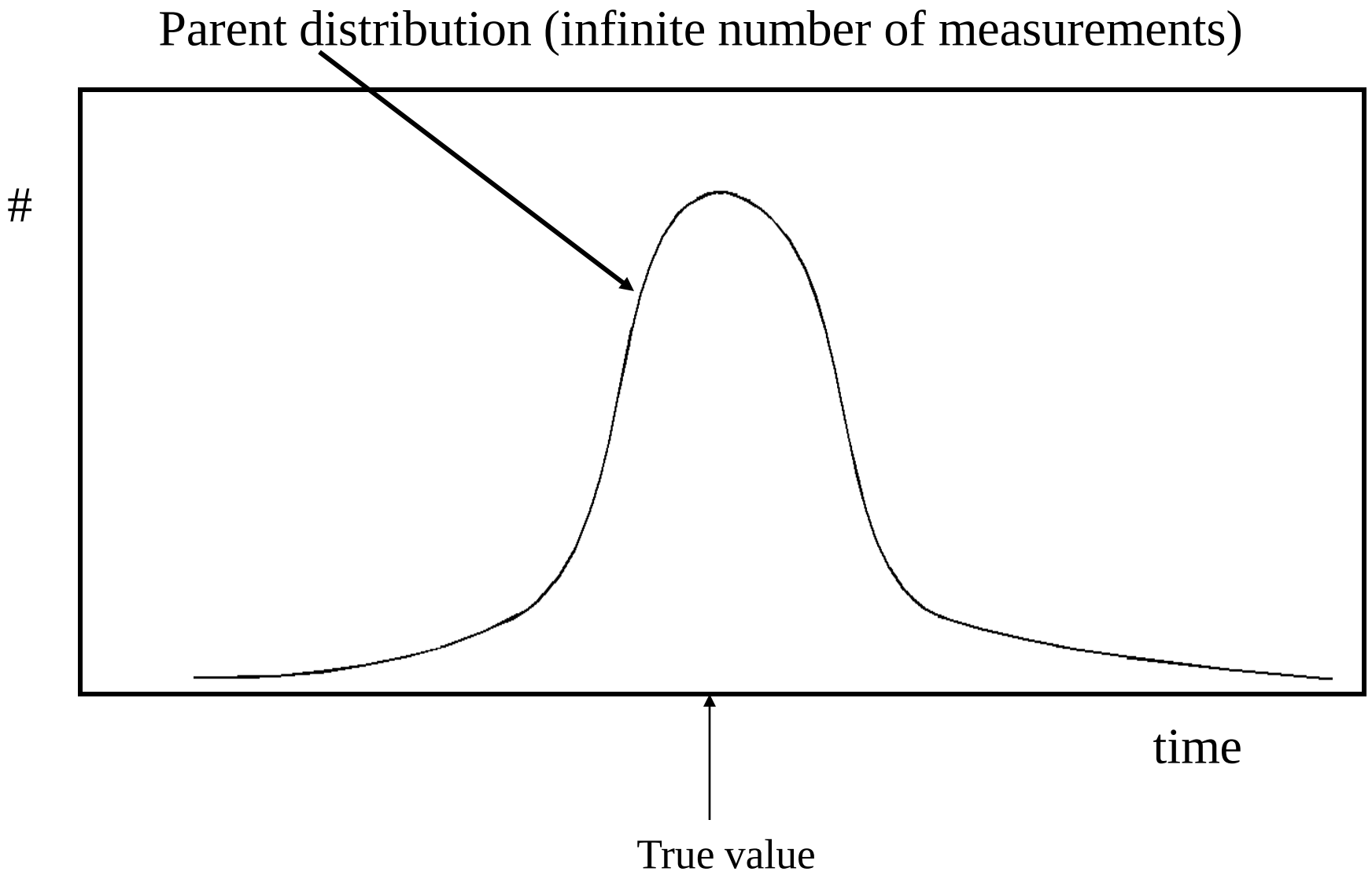
precise

Types of errors

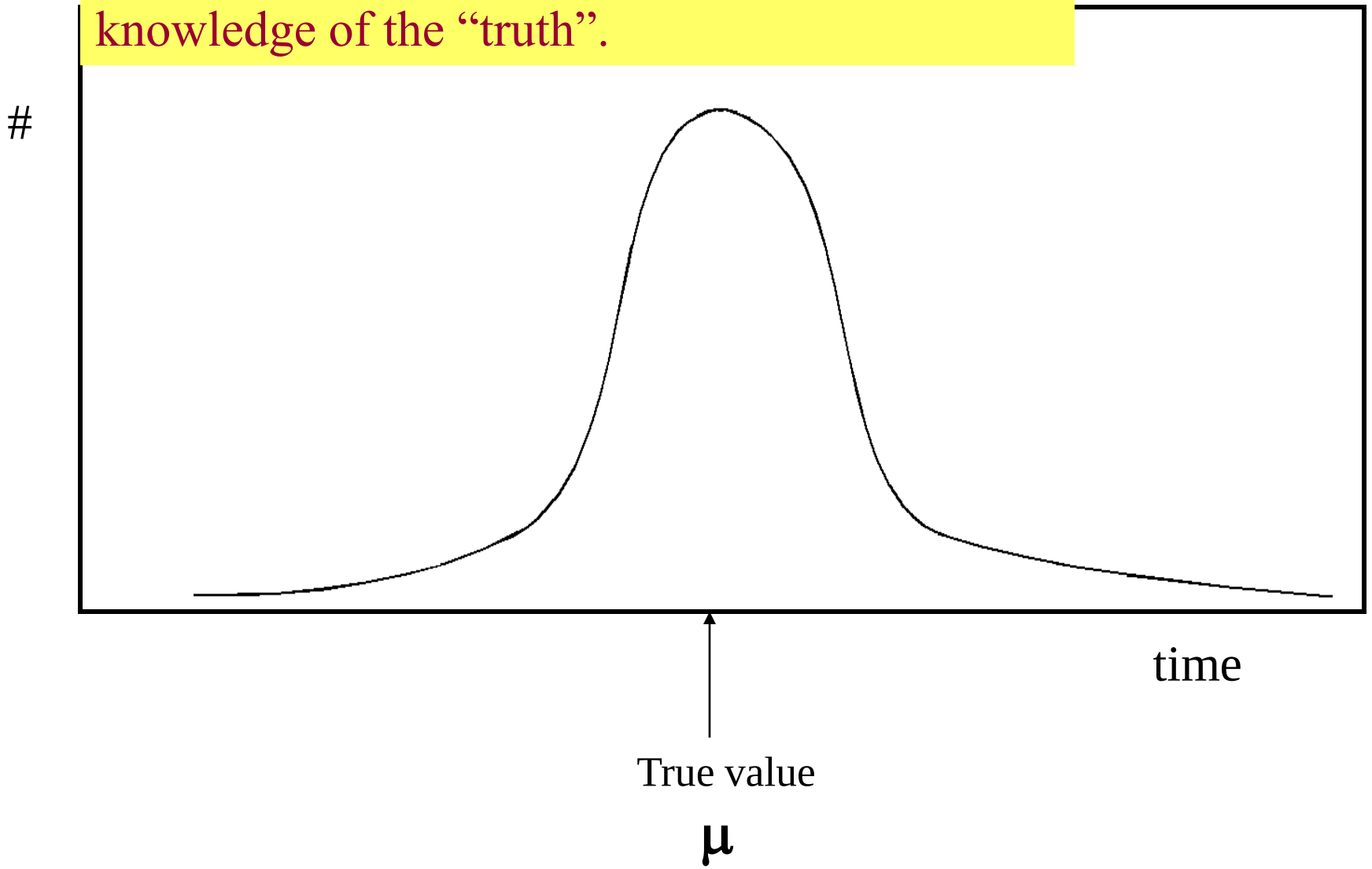
Statistical error: Results from a random fluctuation in the process of measurement. Often quantifiable in terms of "number of measurements or trials". Tends to make measurements less precise.

Systematic error: Results from a bias in the observation due to observing conditions or apparatus or technique or analysis. Tend to make measurements less accurate.


#
time
True value

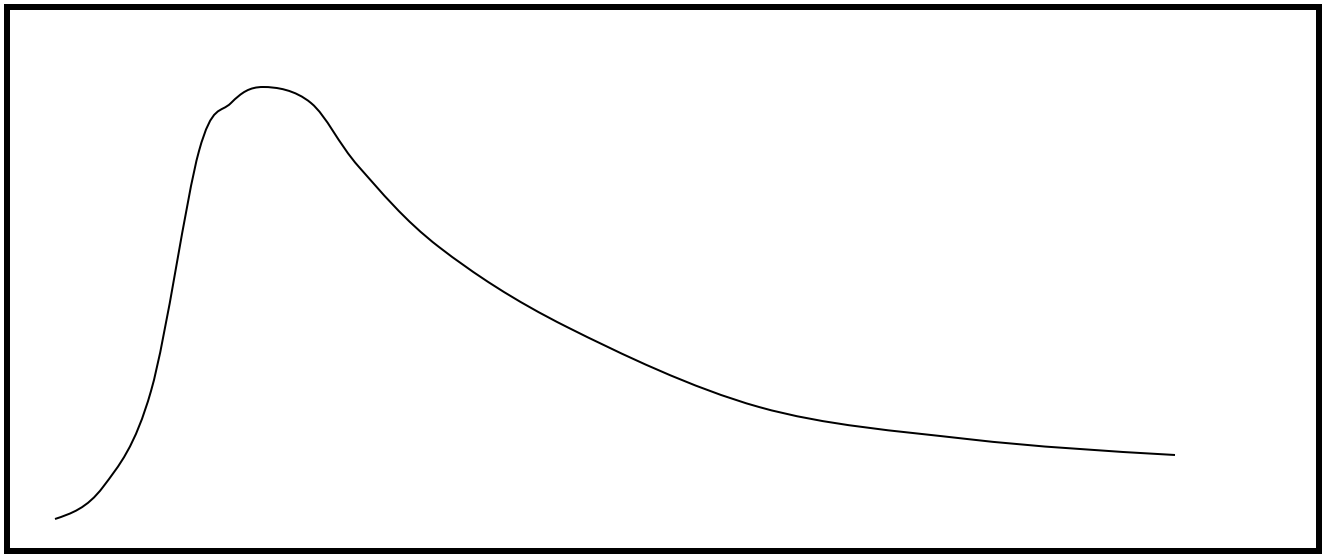

Parent distribution (infinite number of measurements)
#
time
True value

The game: From N (not infinite) observations, determine "μ" and the "error on μ" … without knowledge of the "truth".



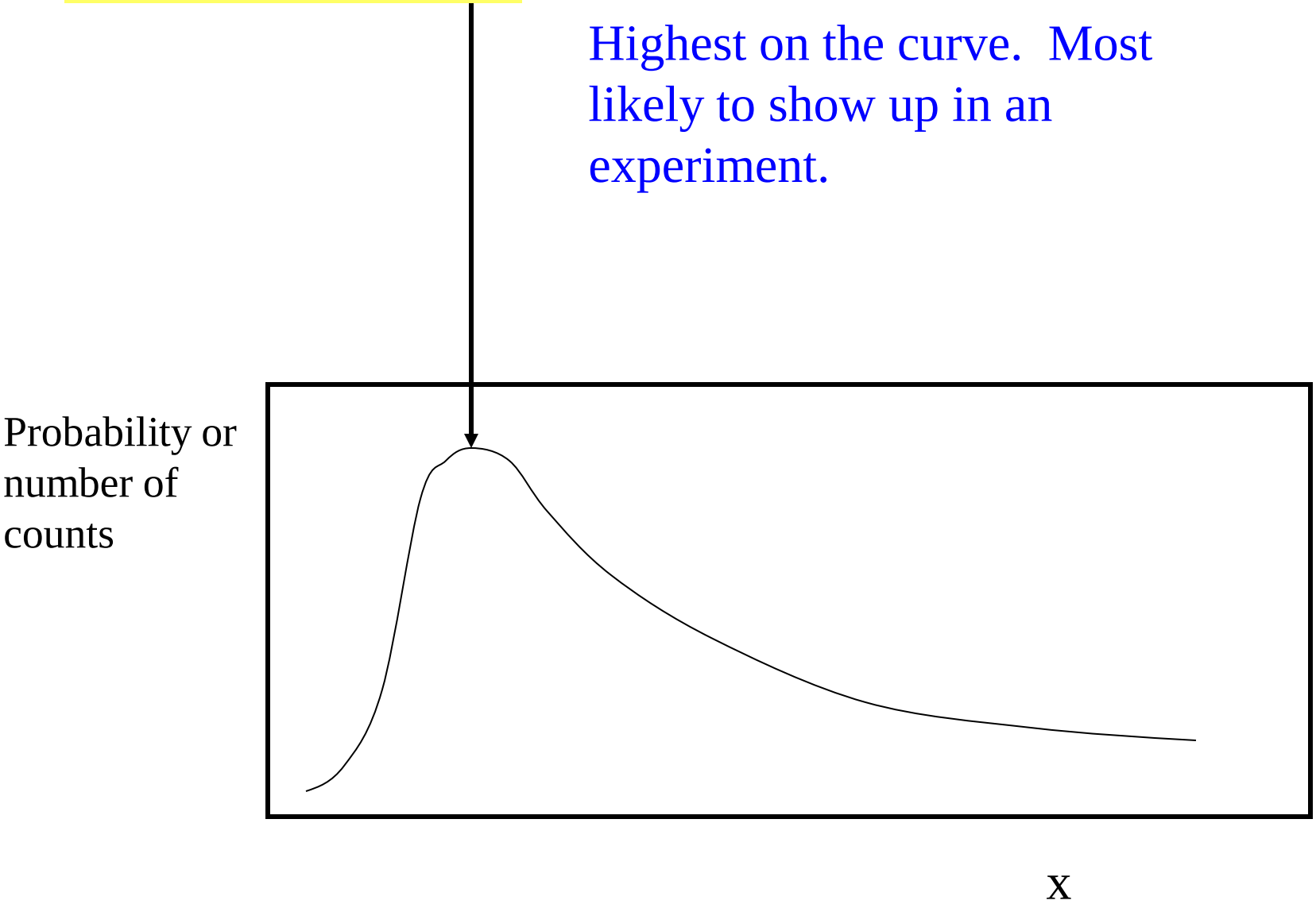
The parent distribution can take different shapes, depending on the nature of the measurement.

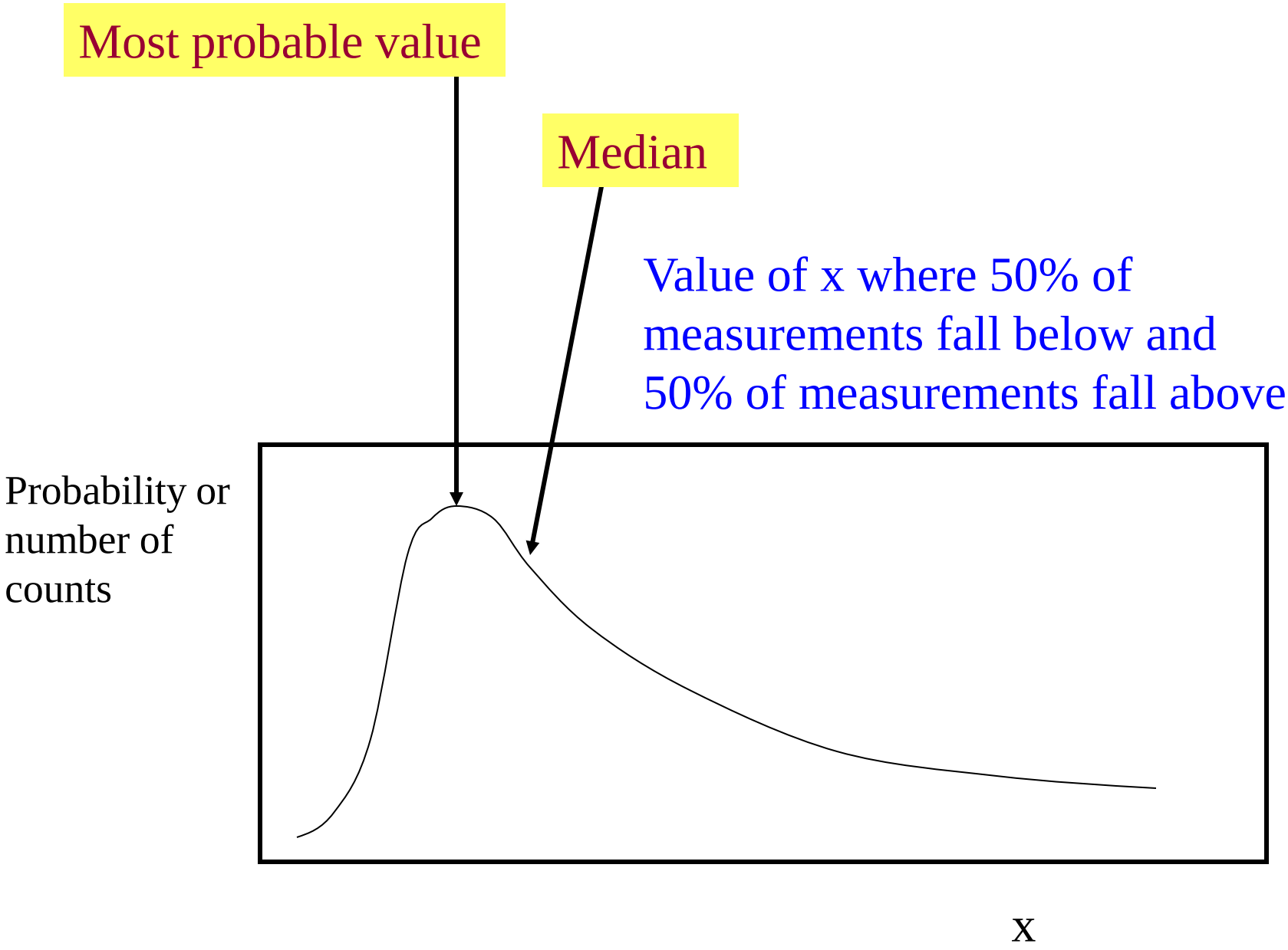
The two most common distributions one sees are the Gaussian and Poisson distributions.

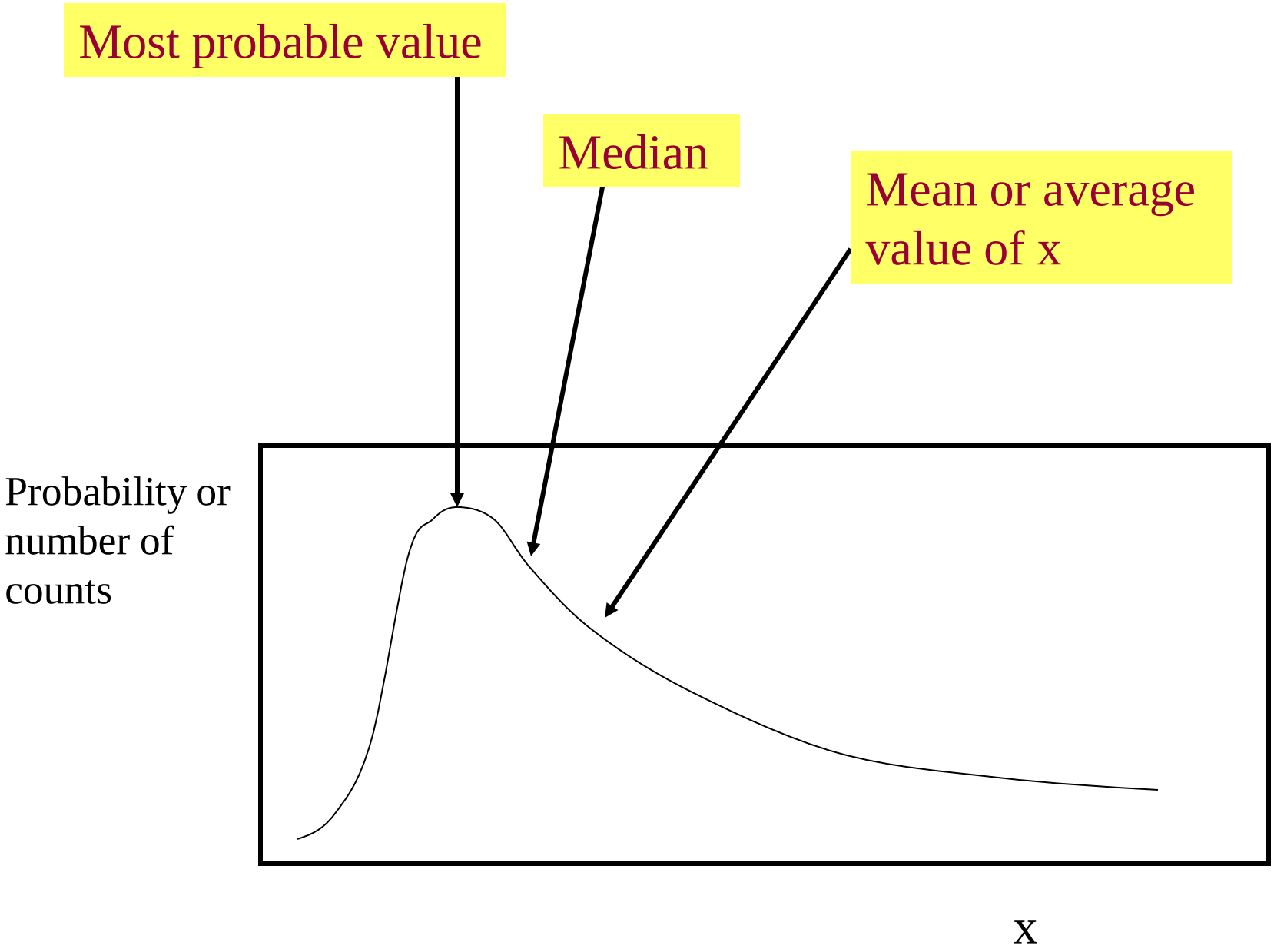


Most probable value

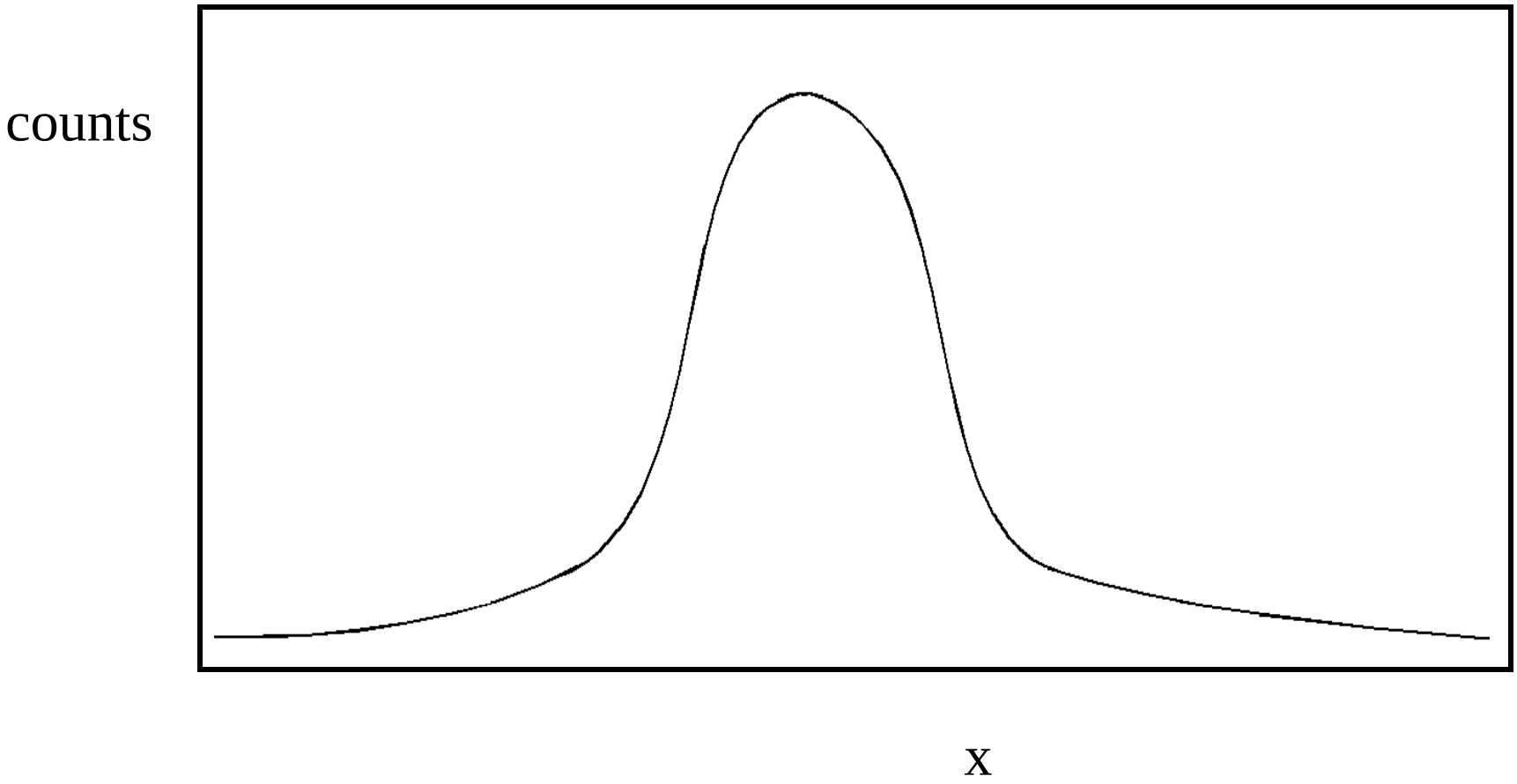
Highest on the curve. Most likely to show up in an experiment.



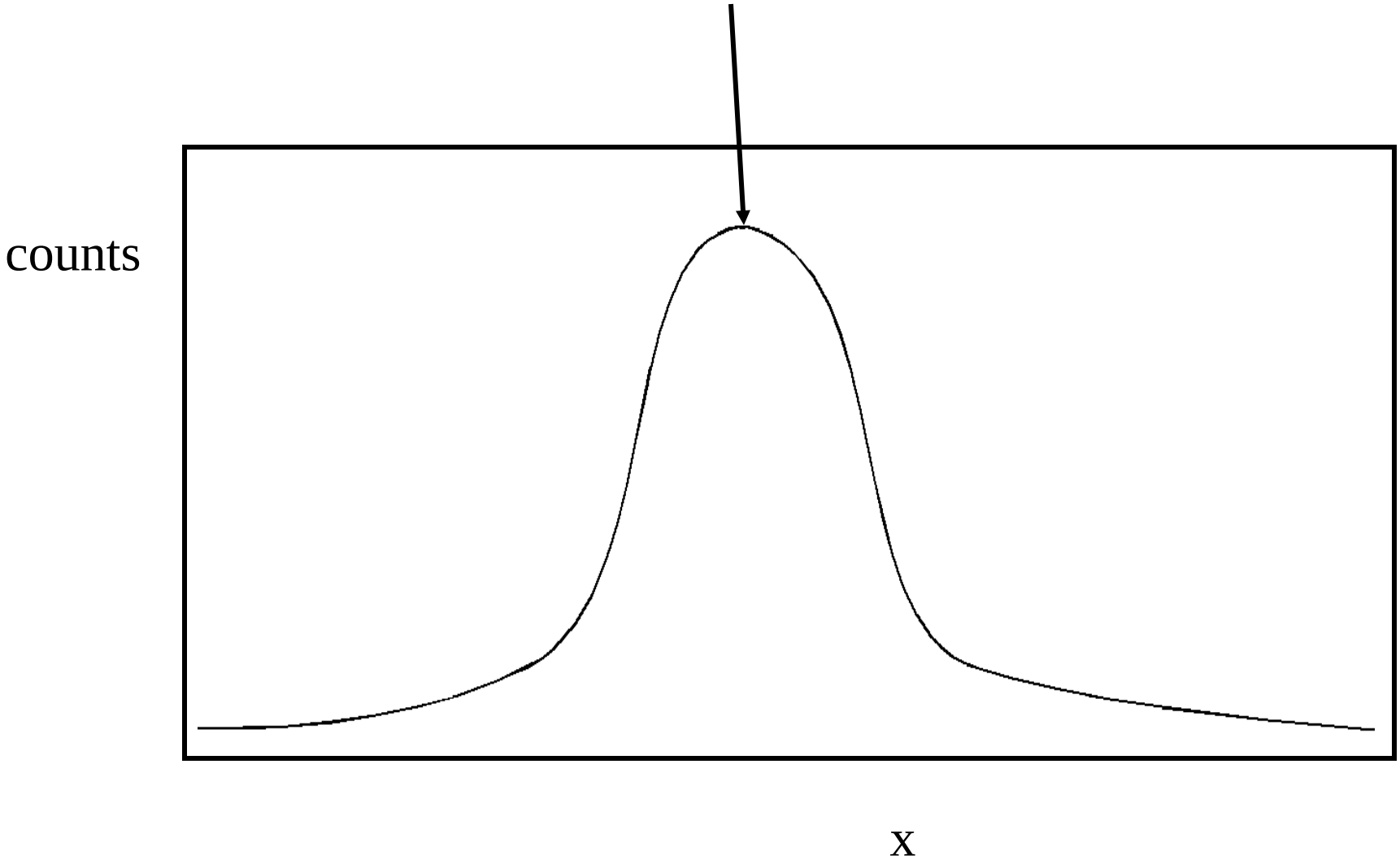

Most probable value
Median
Value of x where 50% of measurements fall below and 50% of measurements fall above
Probability or number of counts
x


Most probable value
Median
Mean or average value of x
Probability or number of counts
x

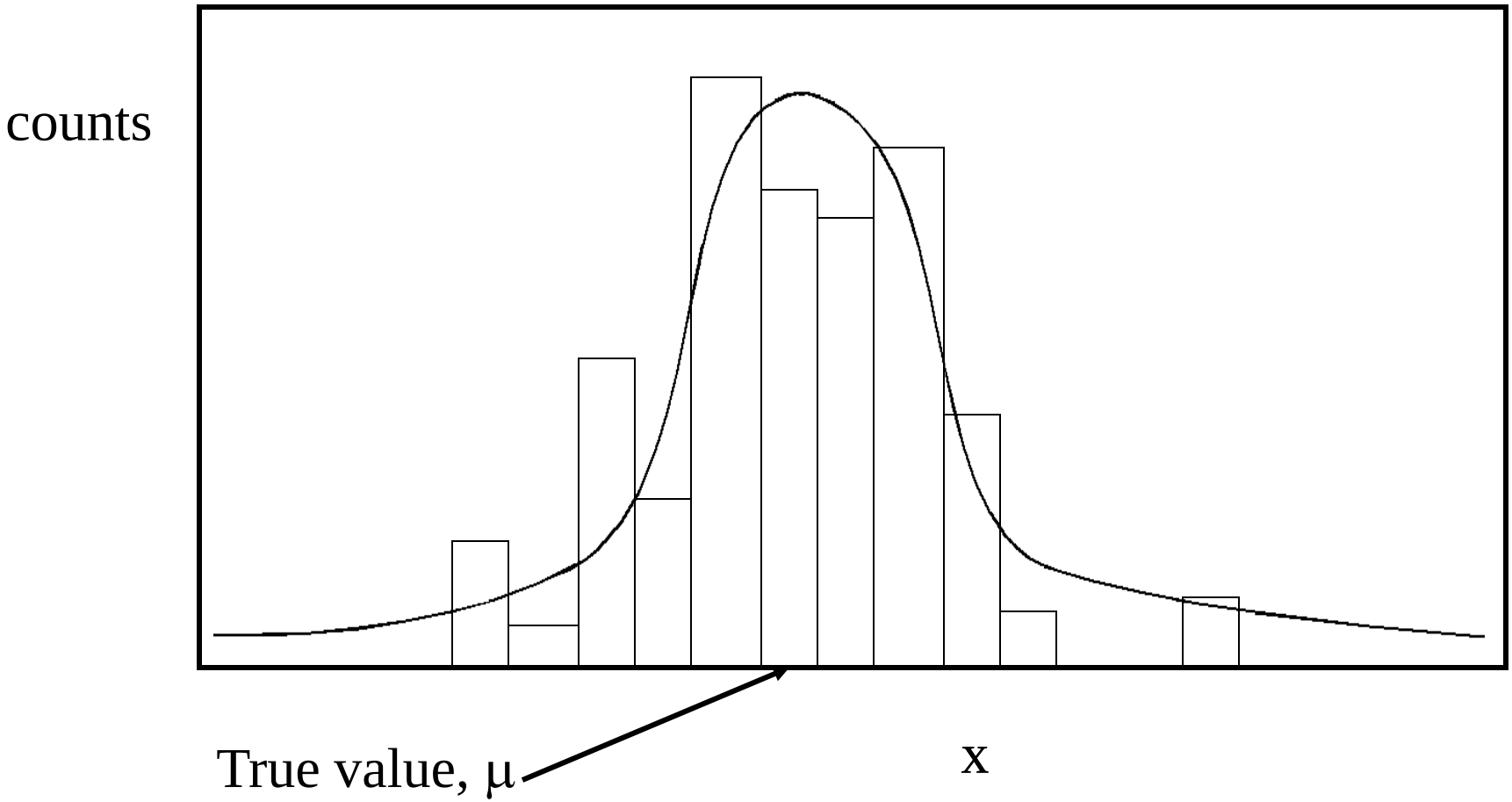
The most common distribution one sees (and that which is best for guiding intuition) is the Gaussian distribution.



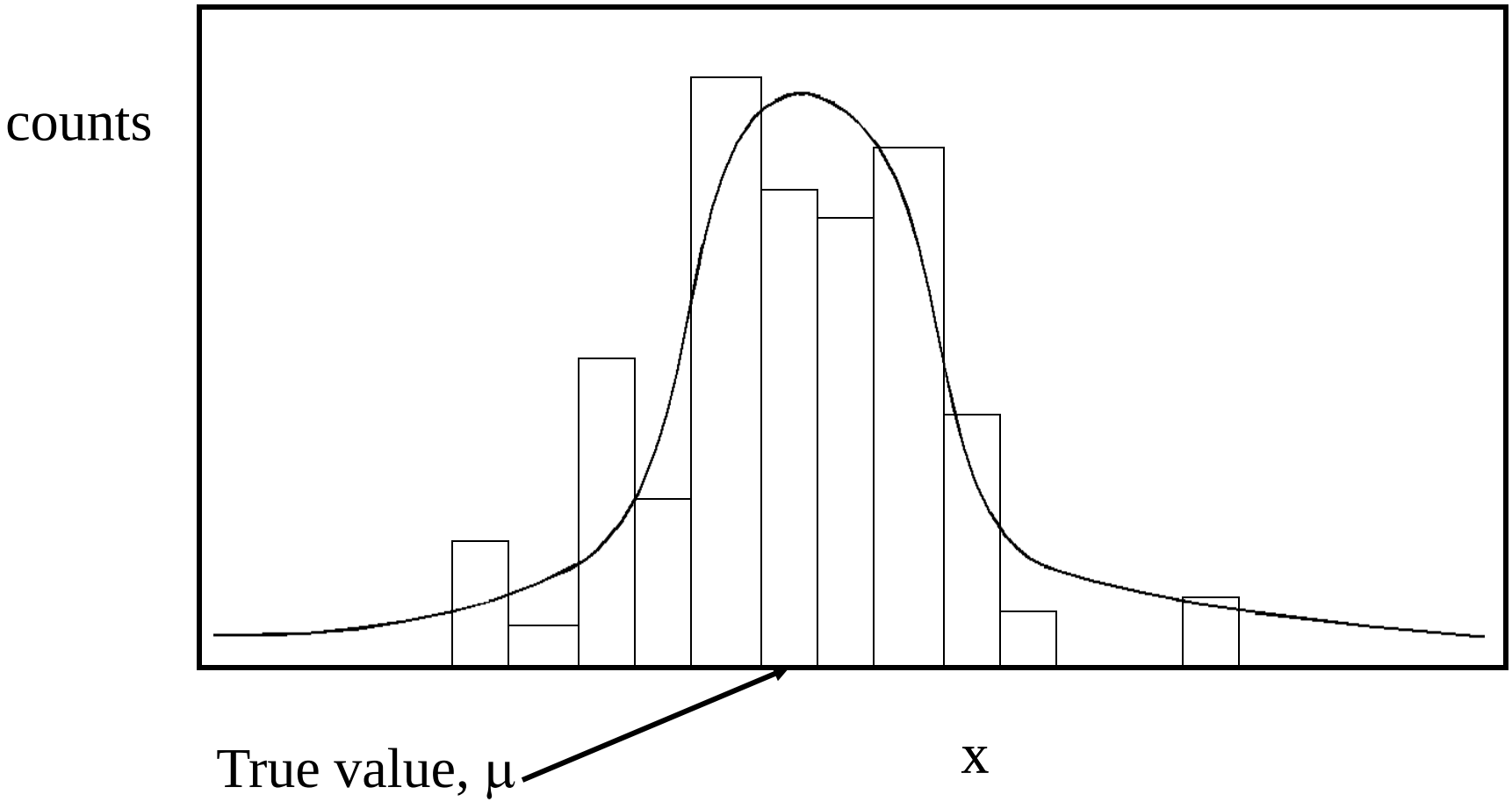
For this distribution, the most probable value, the median value and the average are all the same due to symmetry.

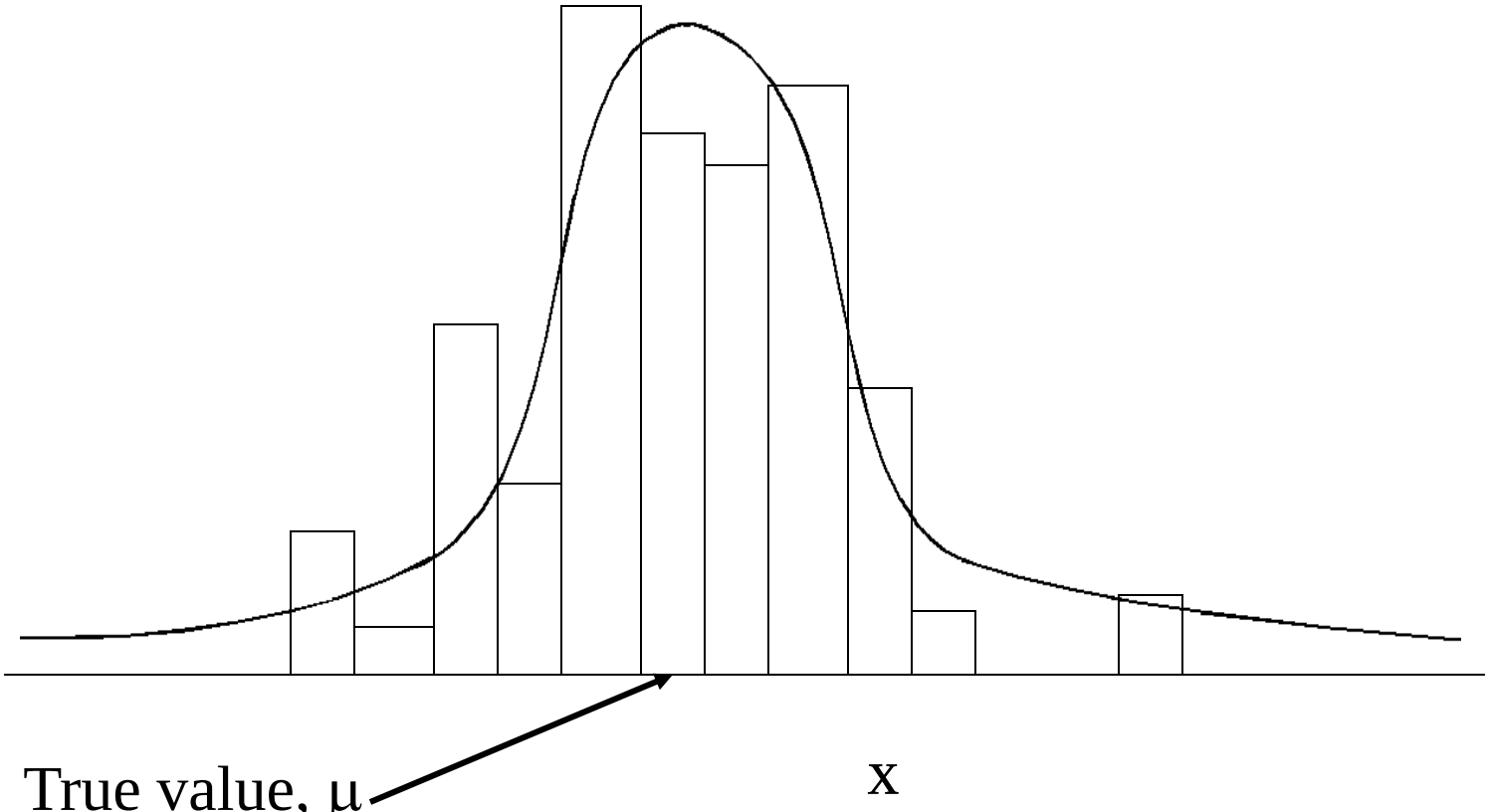


The most probable estimate of μ is given by the mean of the distribution of the N observations



$$"\mu" = \bar{x} = \frac{x_1 + x_2 + \cdots + x_{N-1} + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$



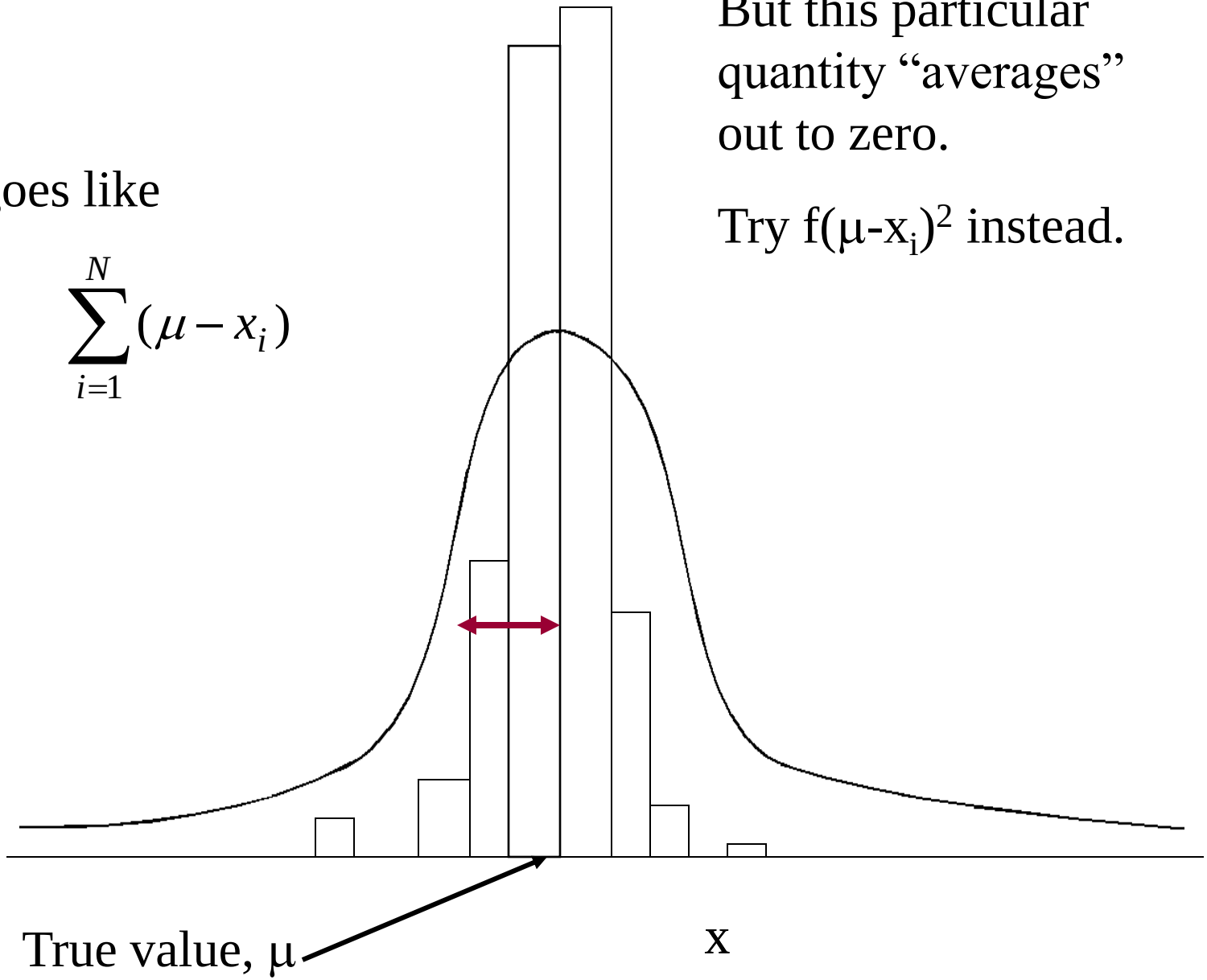

True value, μ
x

Error goes like

$$\sum_{i=1}^{N}(\mu - x_i)$$

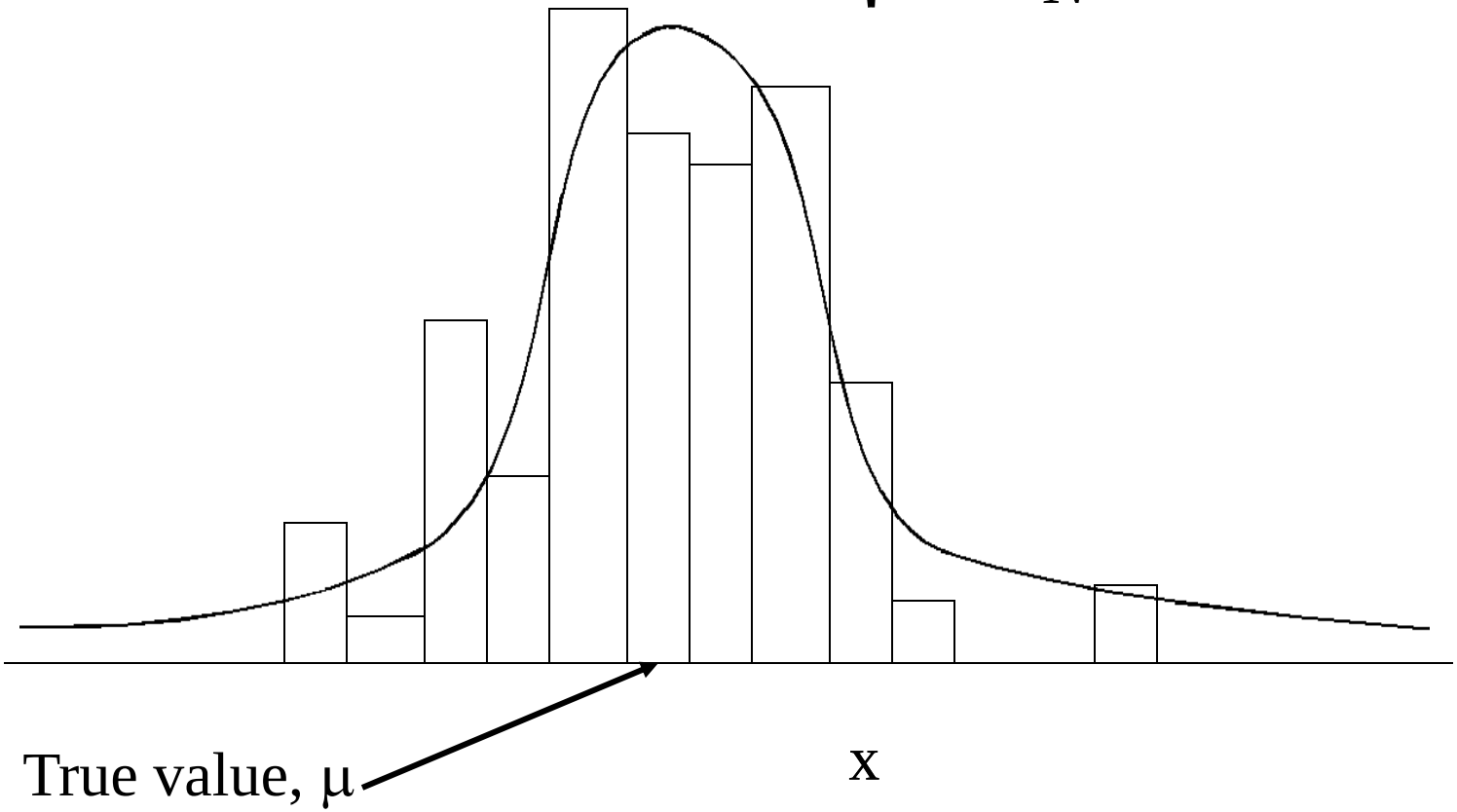
But this particular quantity “averages” out to zero.

Try $f(\mu - x_i)^2$ instead.



The “standard deviation” is a measure of the error in each of the N measurements.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$



μ is unknown. So use the mean (which is your best estimate of μ). Change denominator to increase error slightly due to having used the mean.

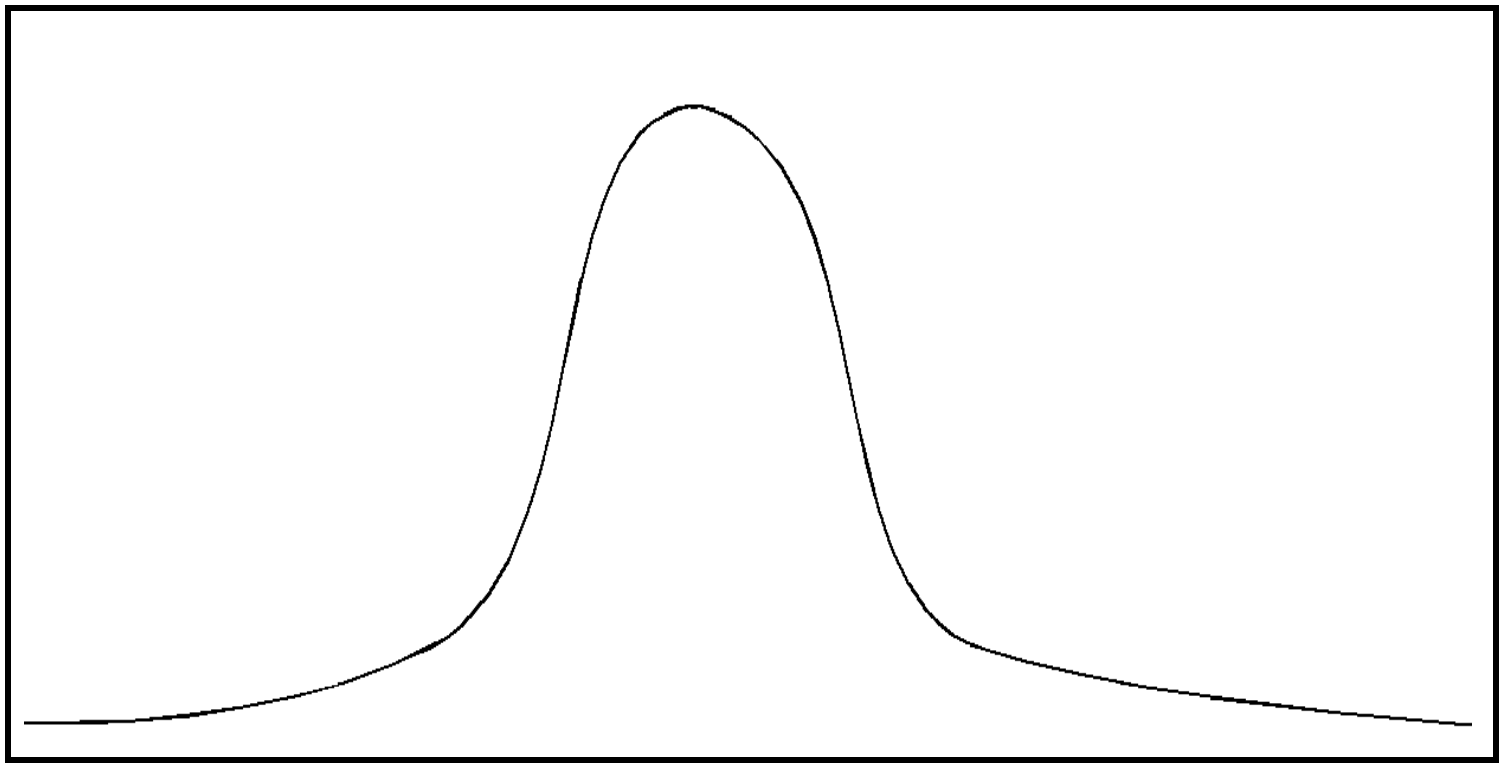
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}}$$

This is the form of the standard deviation you use in practice.

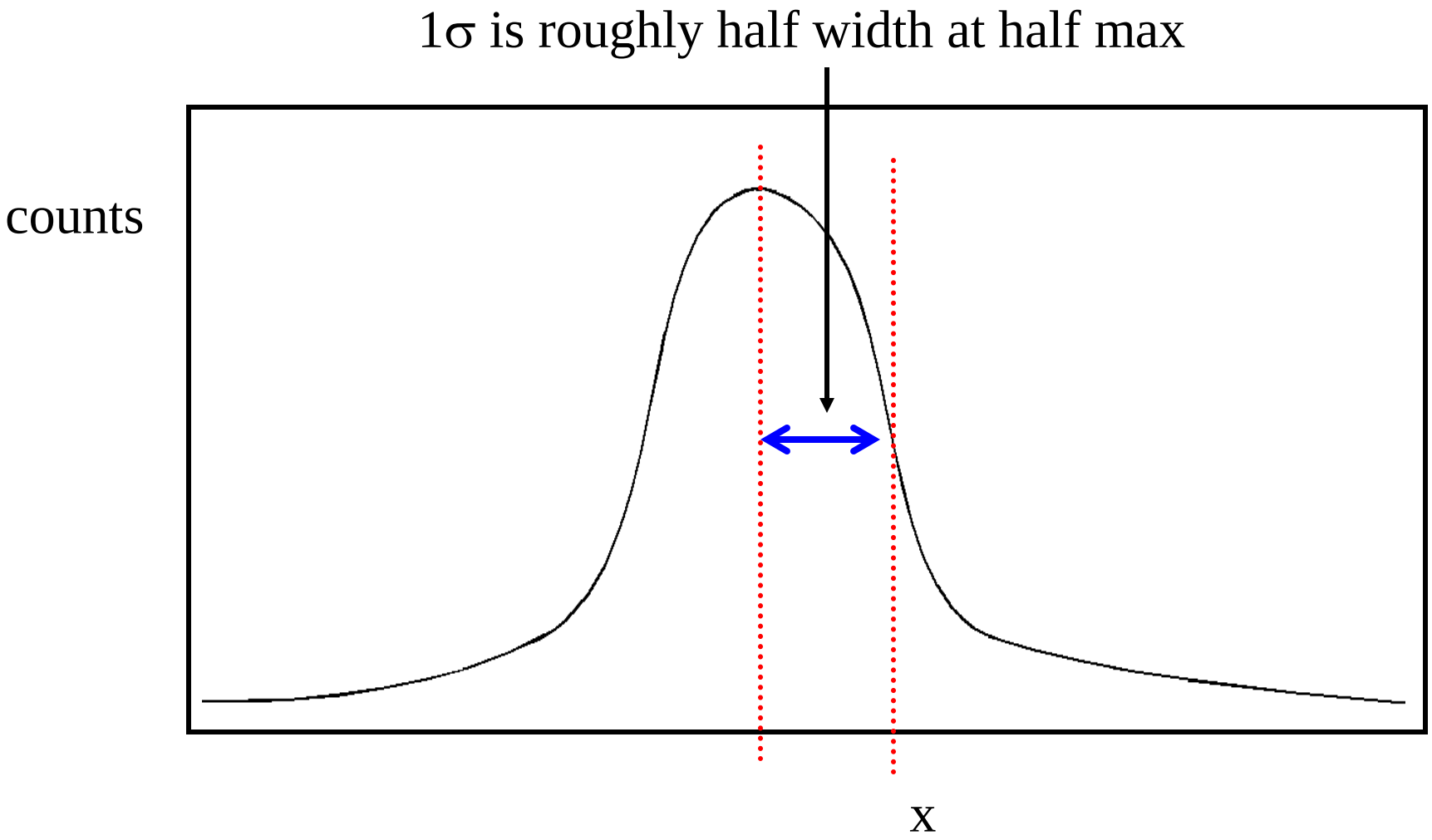
This quantity cannot be determined from a single measurement.

Gaussian distribution

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}}$$

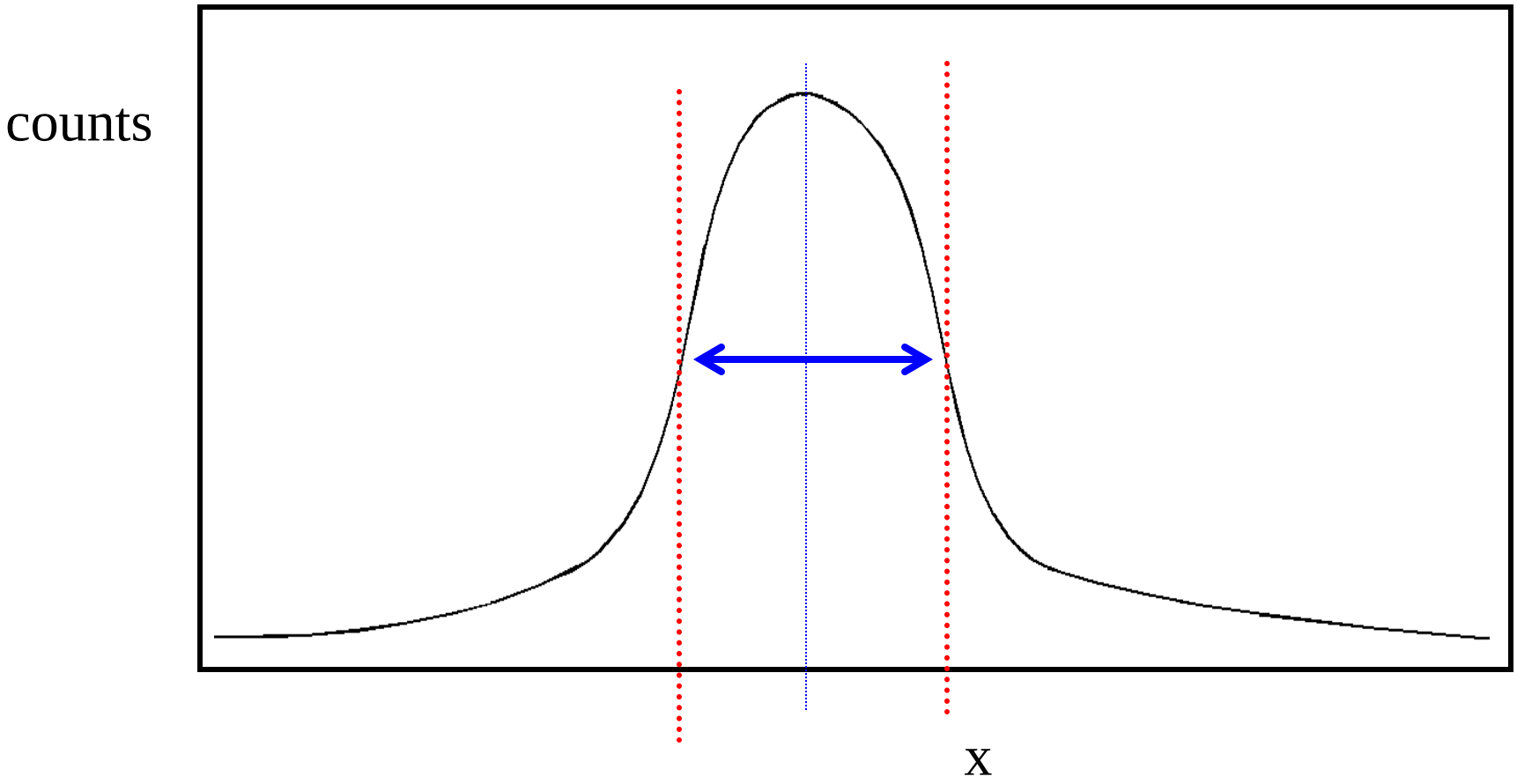


Gaussian distribution intuition



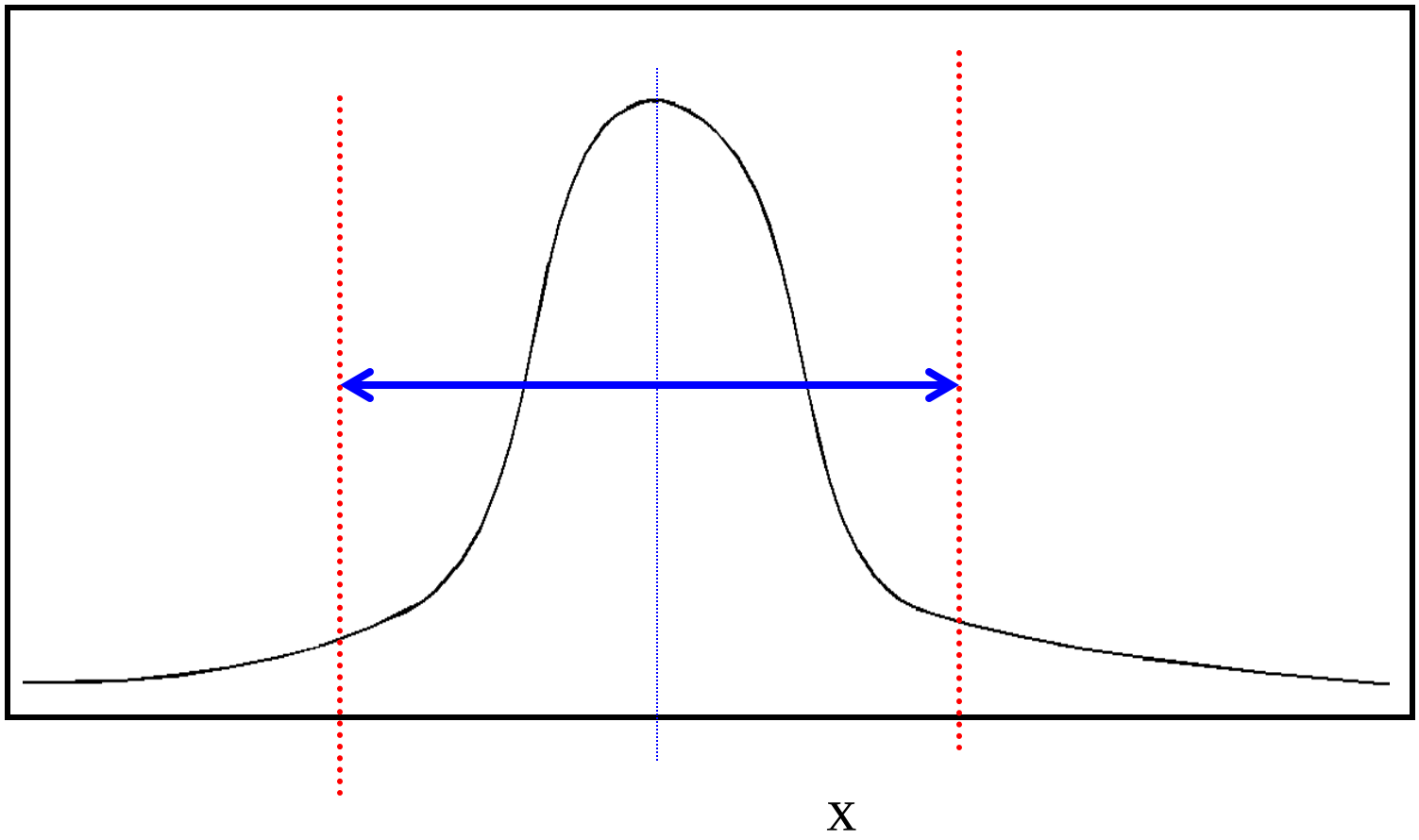
Gaussian distribution intuition

Probability of a measurement falling within $\pm 1\sigma$ of the mean is 0.683



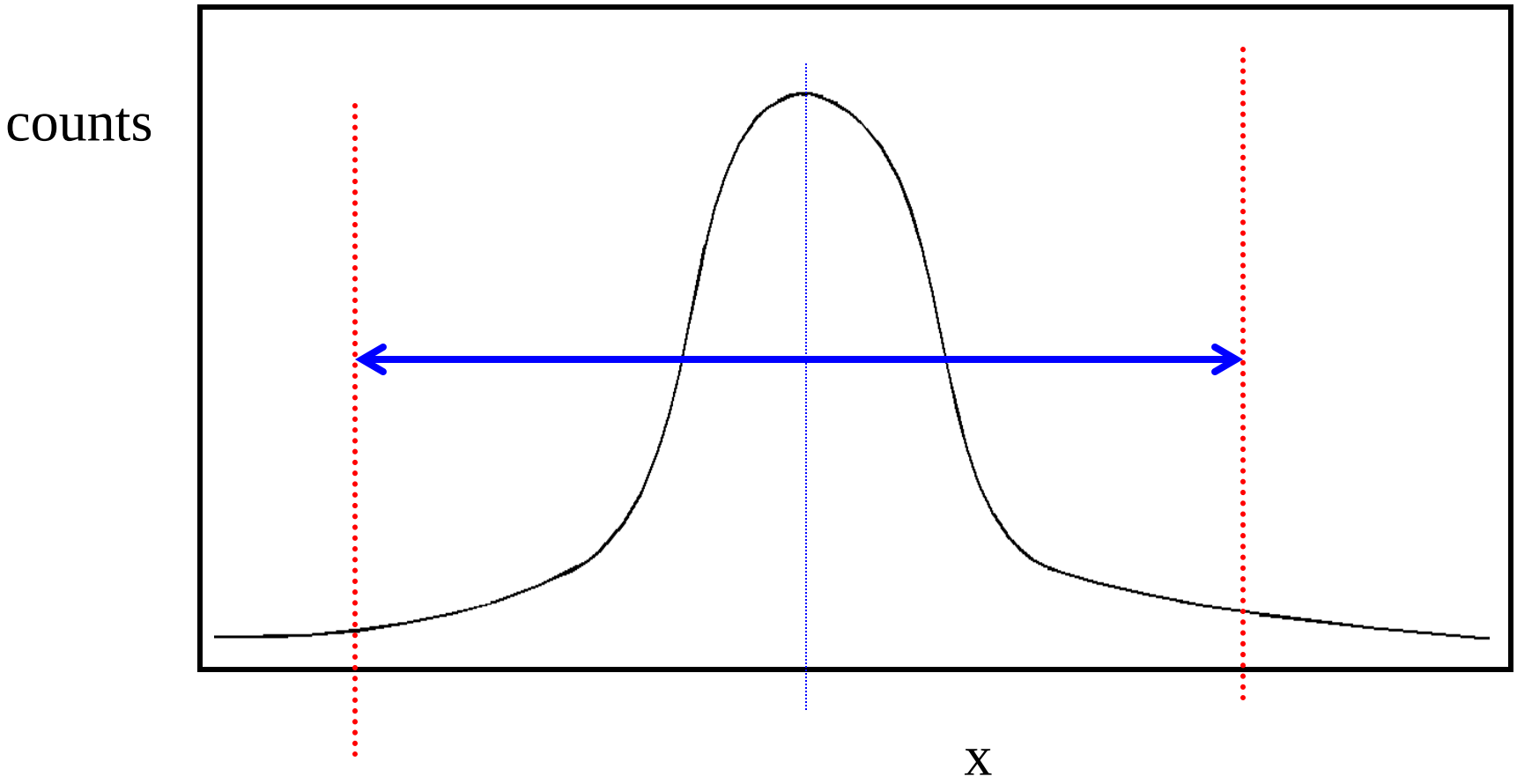
Gaussian distribution intuition

Probability of a measurement falling within $\pm 2\sigma$ of the mean is 0.954



Gaussian distribution intuition

Probability of a measurement falling within $\pm 3\sigma$ of the mean is 0.997



	Month 1	Month 2	
Bush	42%	41%	
Dukakis	40%	43%	
Undecided	18%	16%	±4%

Headline: Dukakis surges past Bush in polls!

The standard deviation is a measure of the error made in each individual measurement.

Often you want to measure the mean and the error in the mean.

Which should have a smaller error, an individual measurement or the mean?

Error in the mean

$$\sigma_m = \frac{\sigma}{\sqrt{N}}$$

Numerical example:

Some say if Dante were alive now, he would describe hell in terms of taking a university course in physics. One vision brought to mind by some of the comments I've heard is that of the devil standing over the pit of hell gleefully dropping young, innocent, and hardworking students into the abyss in order to measure "g", the acceleration due to gravity.

Student 1: 9.0 m/s^2

Student 2: 8.8 m/s^2

Student 3: 9.1 m/s^2

Student 4: 8.9 m/s^2

Student 5: 9.1 m/s^2

$$\bar{a} = \frac{9.0 + 8.8 + 9.1 + 8.9 + 9.1}{5} = 9.0 \frac{m}{s^2}$$

$$\sigma = \sqrt{\frac{(9.0-9.0)^2 + (8.8-9.0)^2 + (9.1-9.0)^2 + (8.9-9.0)^2 + (9.1-9.0)^2}{5-1}}$$

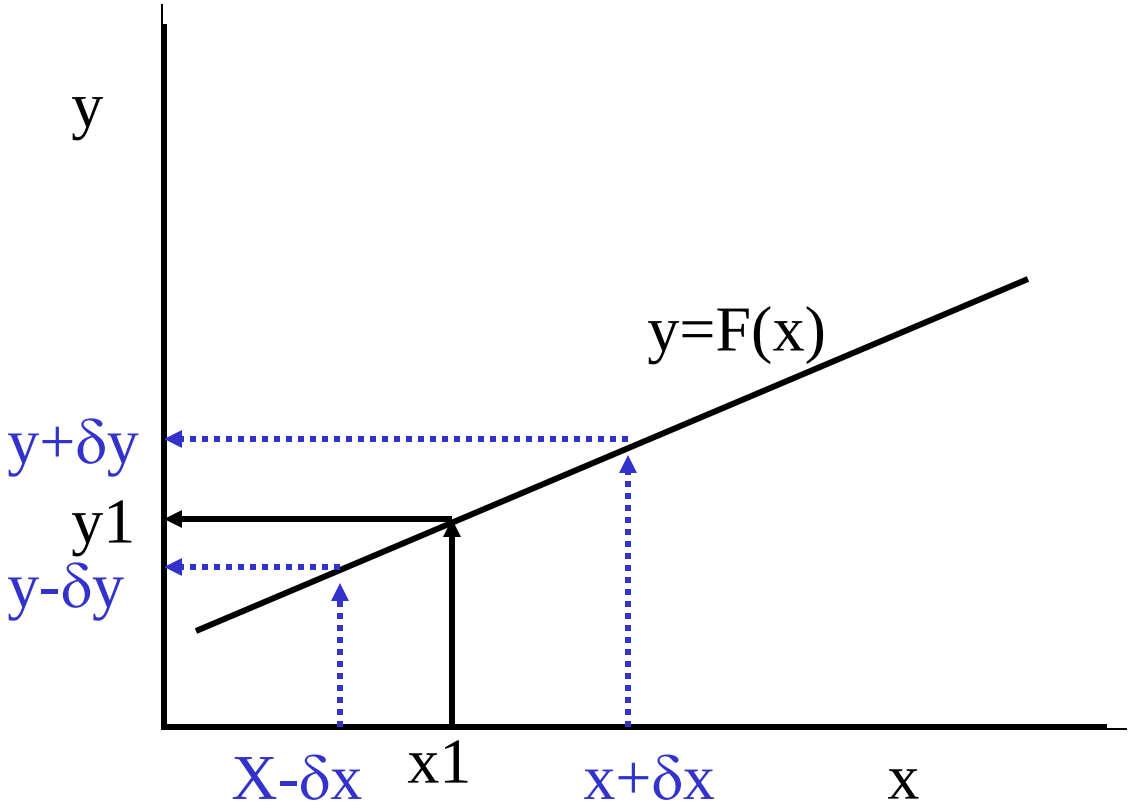
$$= 0.12 \frac{m}{s^2}$$

Error on each individual measurement

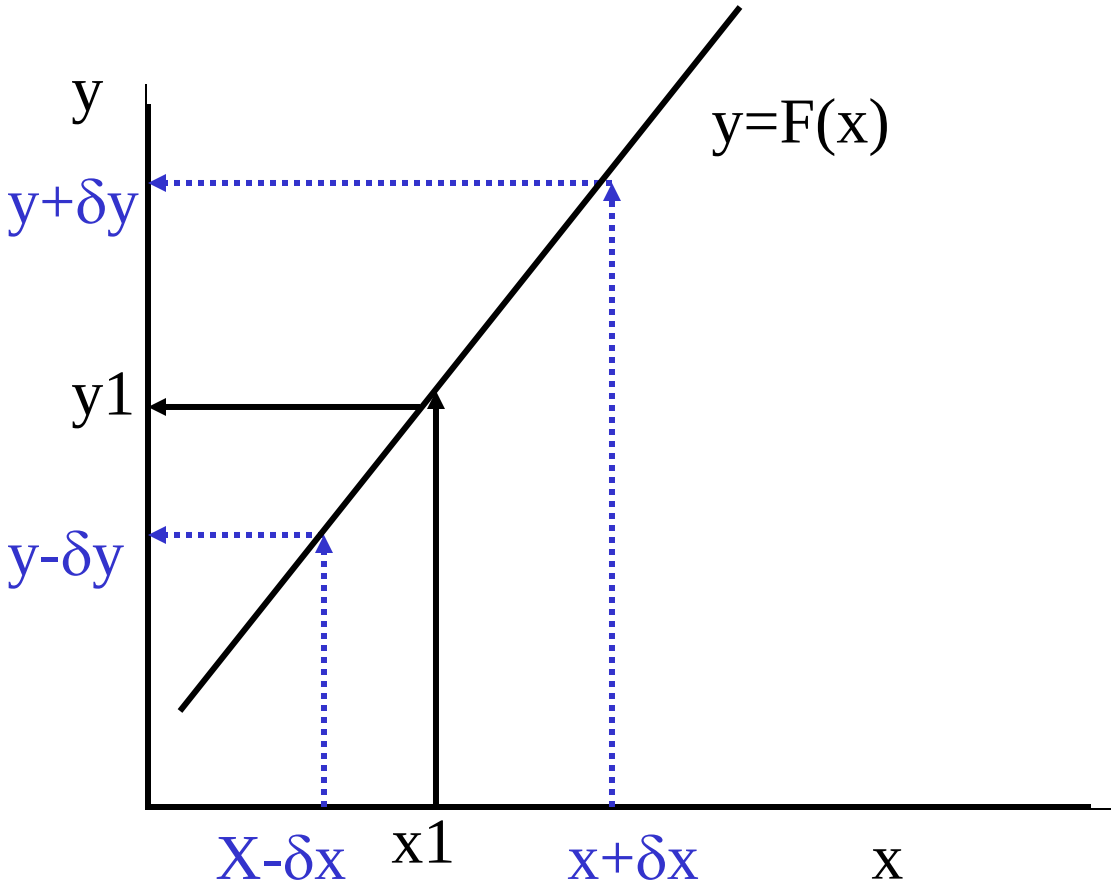
$$\sigma_m = \frac{0.12}{\sqrt{5}} = 0.054 \frac{m}{s^2}$$

$$\bar{a} = 9.00 \pm 0.05 \frac{m}{s^2}$$

How does an error in one measurable affect the error in another measurable?



The degree to which an error in one measurable affects the error in another is driven by the functional dependence of the variables (or the slope: dy/dx)



The complication

$$x = x_o + \mathrm{v}_o t + \frac{1}{2} a t^2$$

$$F = Ma$$

$$P = M\mathrm{v}$$

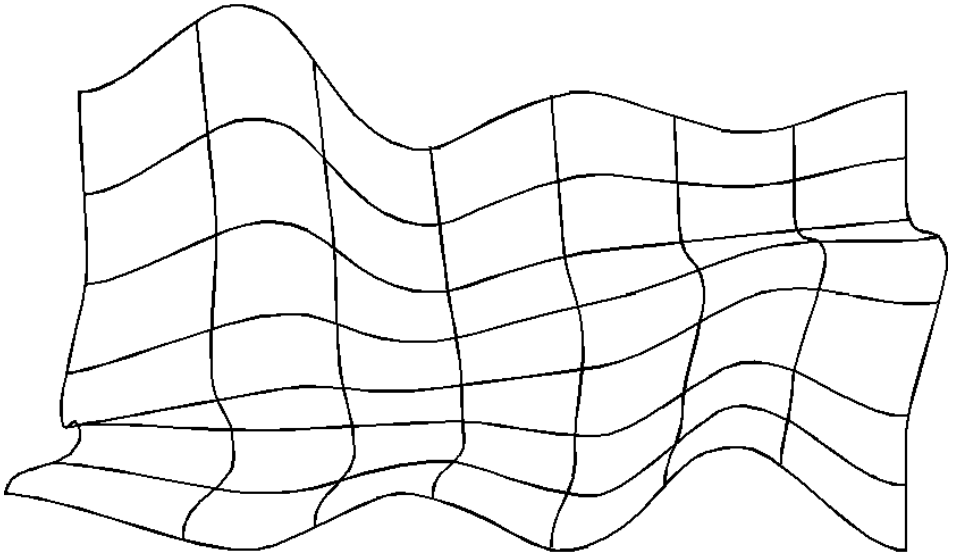
Most physical relationships involve multiple measurables!

$$y = F(x1,x2,x3,\ldots)$$

Must take into account the dependence of the final measurable on each of the contributing quantities.

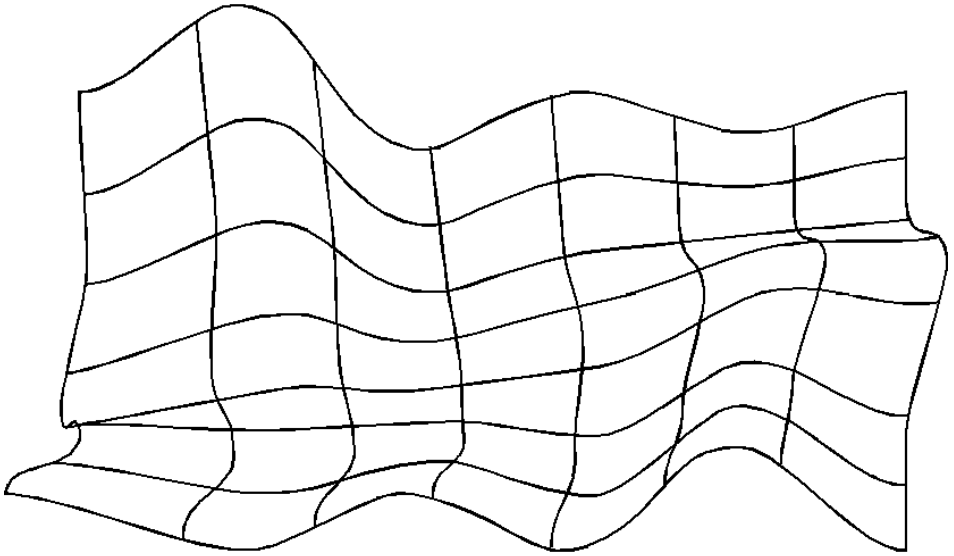
Partial derivatives

What's the slope of this graph??



For multivariable functions, one needs to define a "derivative" at each point for each variable that projects out the local slope of the graph in the direction of that variable … this is the "partial derivative".

Partial derivatives



The partial derivative with respect to a certain variable is the ordinary derivative of the function with respect to that variable where all the other variables are treated as constants.

$$\frac{\partial F(x,y,z,\ldots)}{\partial x} = \frac{dF(x,y,z\ldots)}{dx}\Bigg]_{y,z\ldots const}$$

Example

$$F(x, y, z) = x^2 yz^3$$

$$\frac{\partial F}{\partial x} = 2xyz^3$$

$$\frac{\partial F}{\partial y} = x^2 z^3$$

$$\frac{\partial F}{\partial z} = x^2 y 3z^2$$

Dude! Just give us the freakin' formula!

The formula for error propagation

If f=F(x,y,z…) and you want σ_f and you have σ_x, σ_y, σ_z …, then use the following formula:

$$\sigma_f = \sqrt{\left(\frac{\partial F}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial F}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial F}{\partial z}\right)^2 \sigma_z^2 + \ldots}$$

The formula for error propagation

If f=F(x,y,z…) and you want σ_f and you have σ_x, σ_y, σ_z …, then use the following formula:

$$\sigma_f = \sqrt{\left(\frac{\partial F}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial F}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial F}{\partial z}\right)^2 \sigma_z^2 + \ldots}$$

Measure of error in x

The formula for error propagation

If f=F(x,y,z…) and you want σ_f and you have σ_x, σ_y, σ_z …, then use the following formula:

$$\sigma_f = \sqrt{\left(\frac{\partial F}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial F}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial F}{\partial z}\right)^2 \sigma_z^2 + \ldots}$$

Measure of dependence of F on x

The formula for error propagation

If f=F(x,y,z…) and you want σ_f and you have σ_x, σ_y, σ_z …, then use the following formula:

$$\sigma_f = \sqrt{\left(\frac{\partial F}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial F}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial F}{\partial z}\right)^2 \sigma_z^2 + \ldots}$$

Similar terms for each variable, add in quadrature.

Example

A pitcher throws a baseball a distance of 30±0.5 m at 40±3 m/s (~90 mph). From this data, calculate the time of flight of the baseball.

$$t = \frac{d}{v}$$

$$\frac{\partial F}{\partial d} = \frac{1}{v}$$

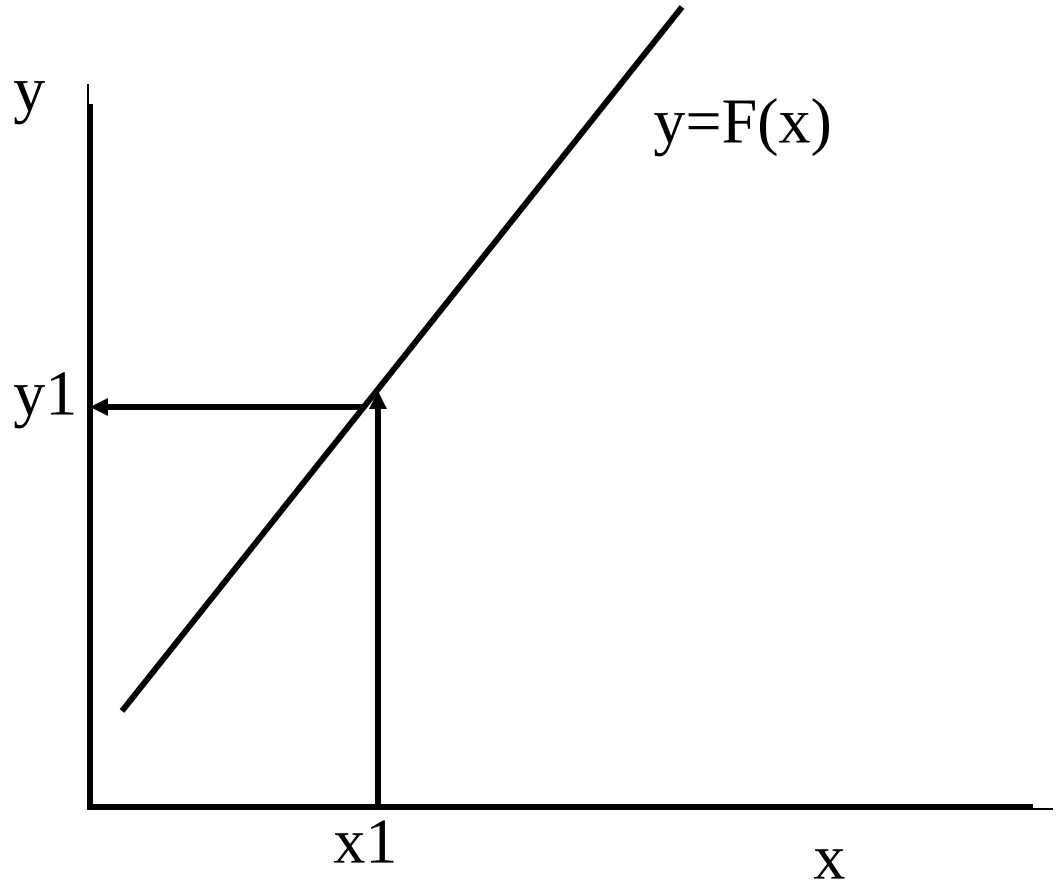
$$\frac{\partial F}{\partial v} = -\frac{d}{v^2}$$

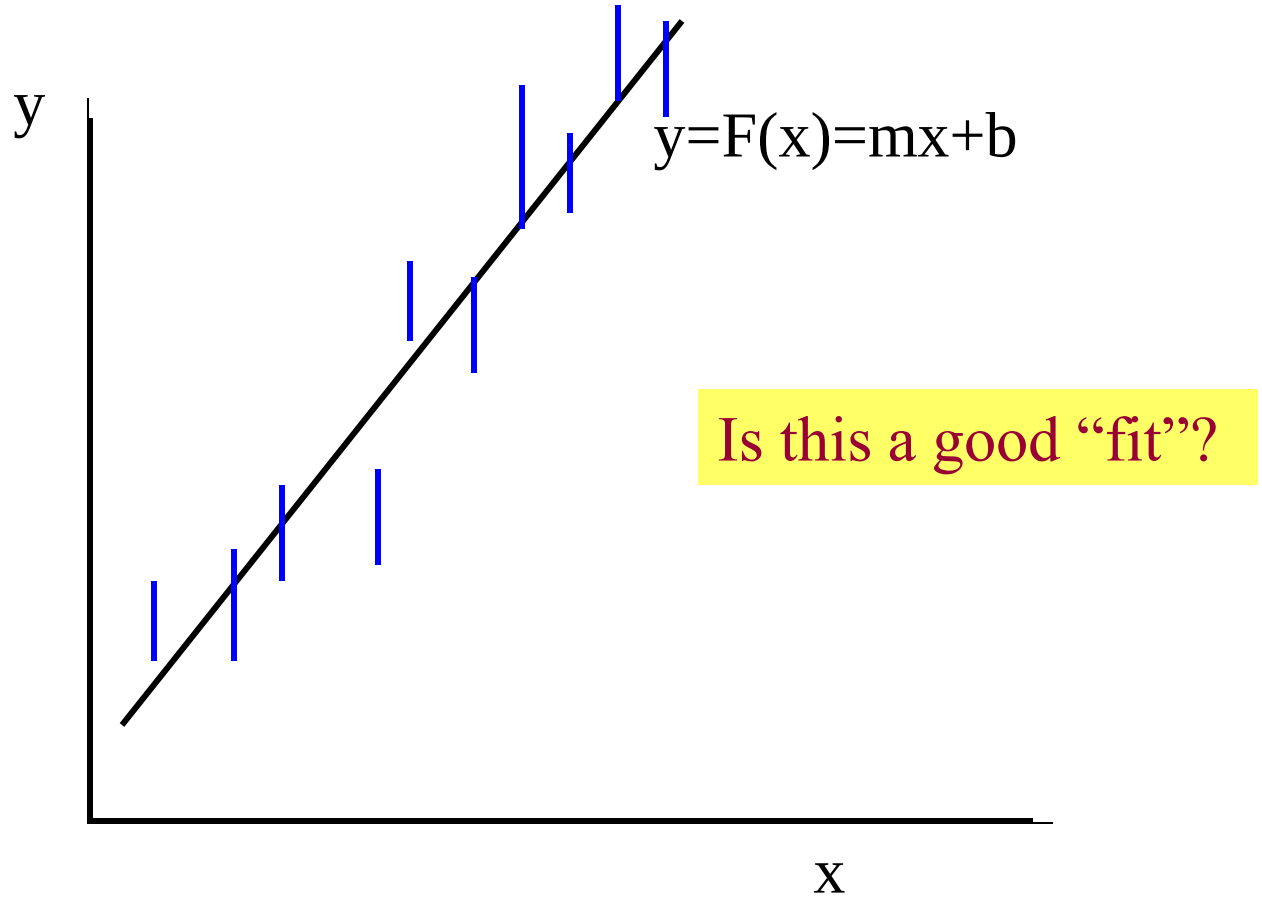
$$\sigma_t = \sqrt{\left(\frac{1}{v}\right)^2 \sigma_d^2 + \left(-\frac{d}{v^2}\right)^2 \sigma_v^2}$$

$$\sigma_t = \sqrt{\left(\frac{0.5}{40}\right)^2 + \left(\frac{30}{40^2}\right)^2 3^2} = 0.058$$

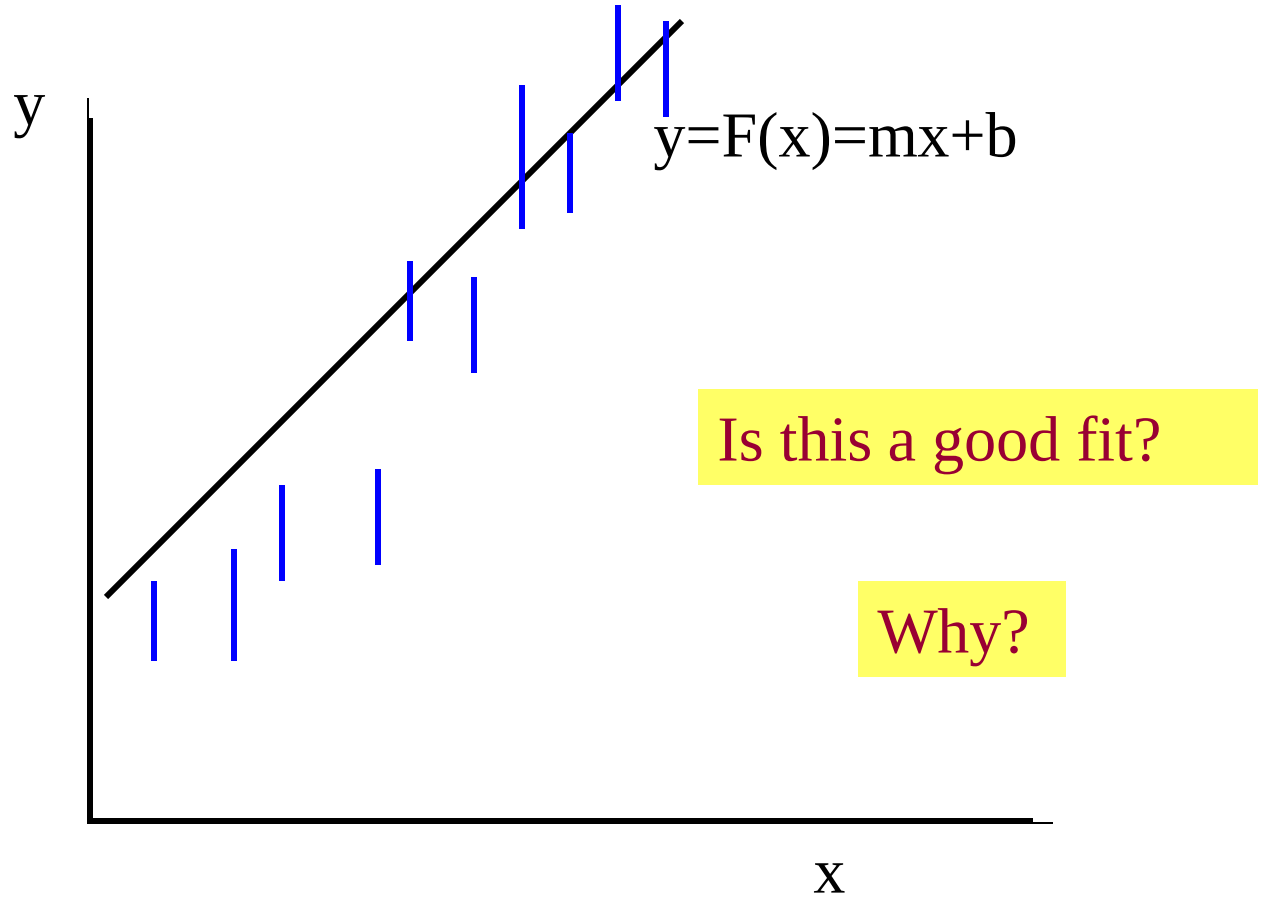
$$t = 0.75 \pm 0.058 s$$

Why are linear relationships so important in analytical scientific work?



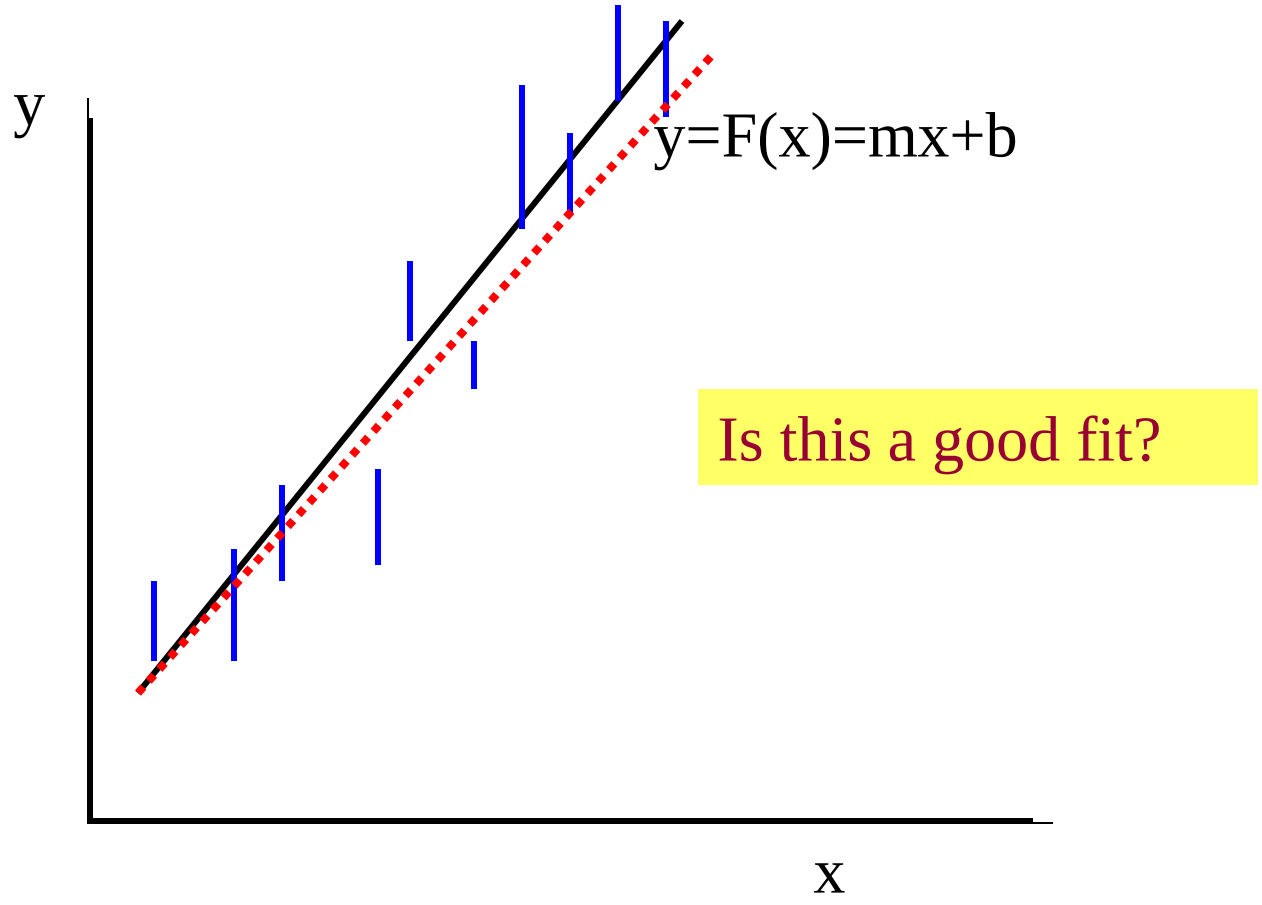


Is this a good “fit”?



Is this a good fit?

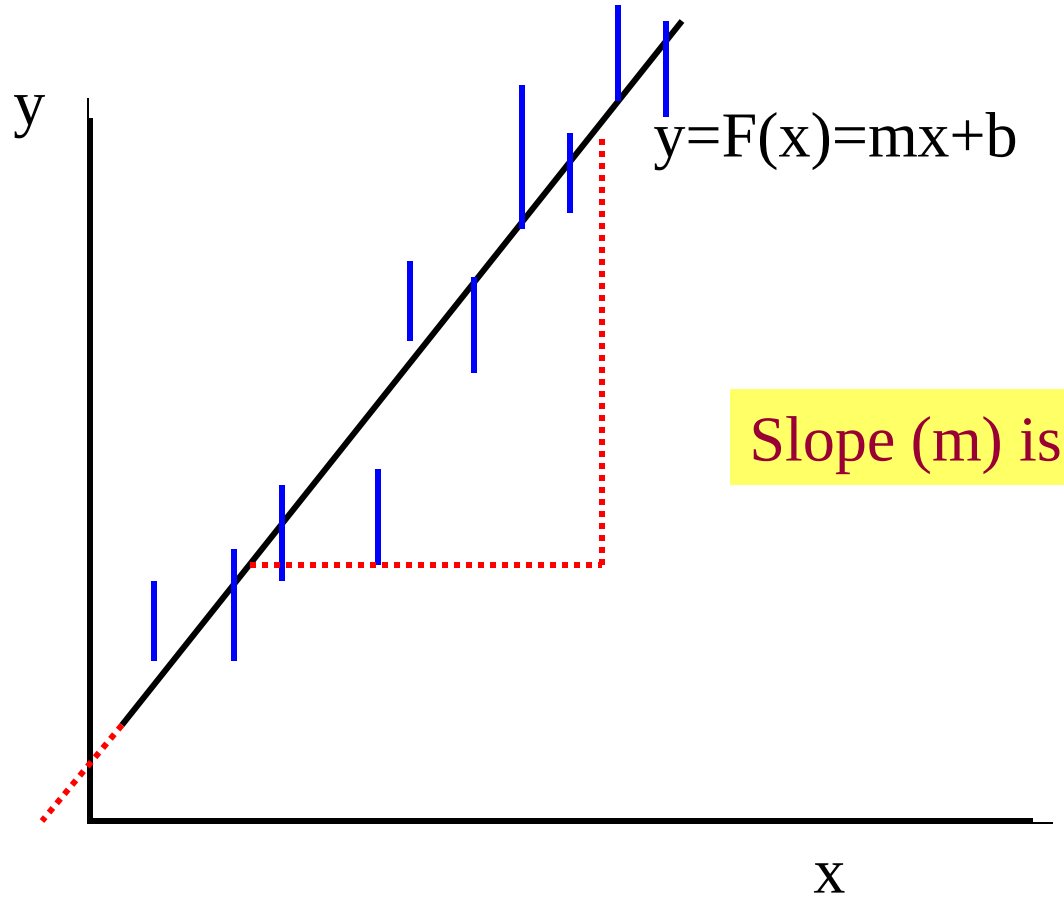
Why?



Is this a good fit?

Graphical analysis

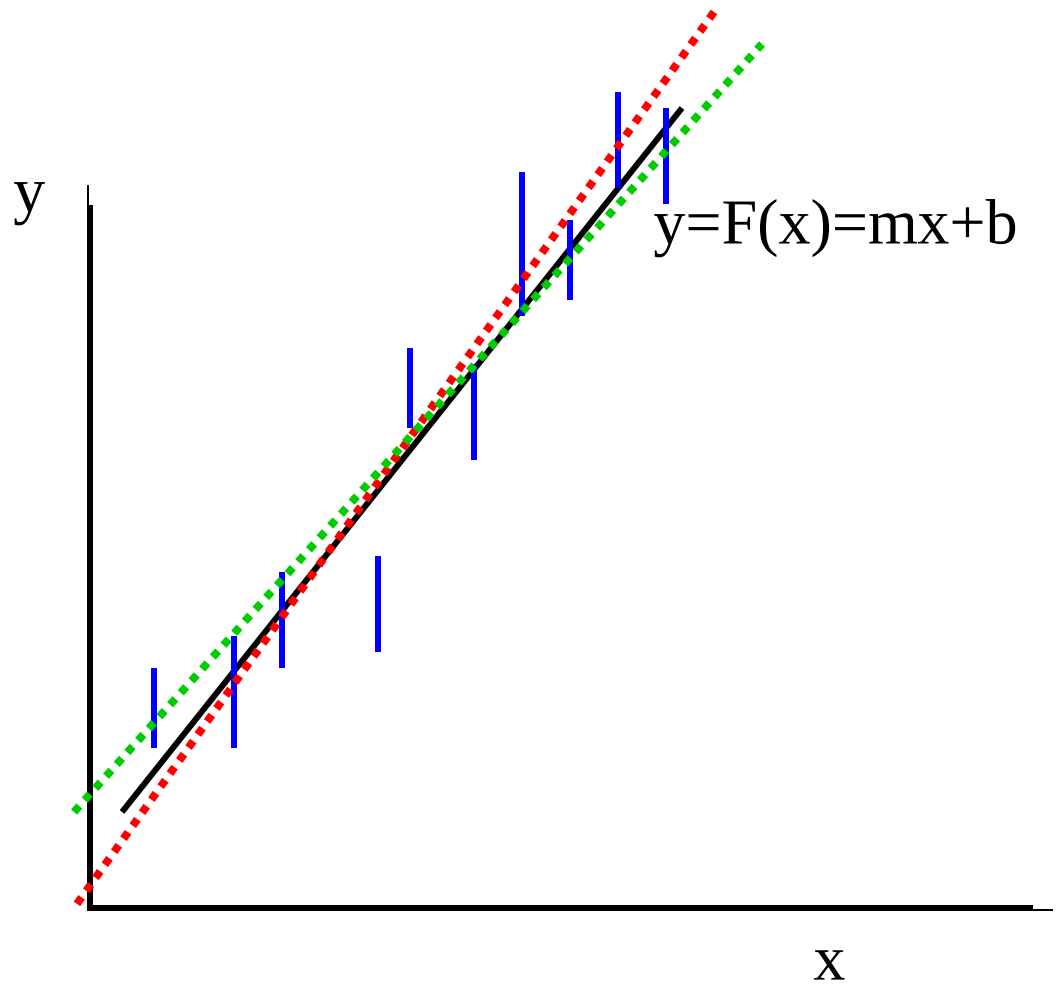
pencil and paper still work!



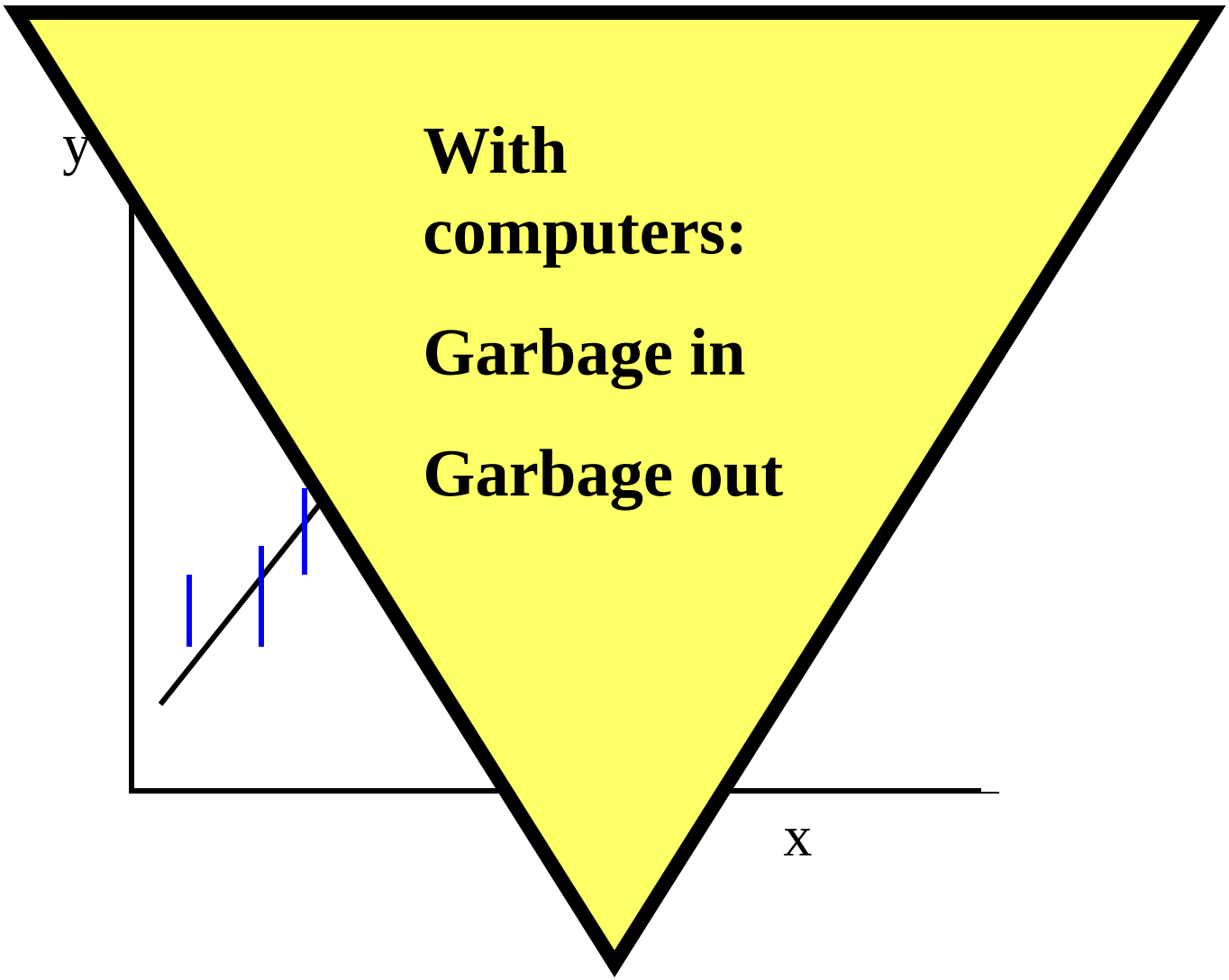
Slope (m) is rise/run

b is the y-intercept

Graphical determination of error in slope and y-intercept

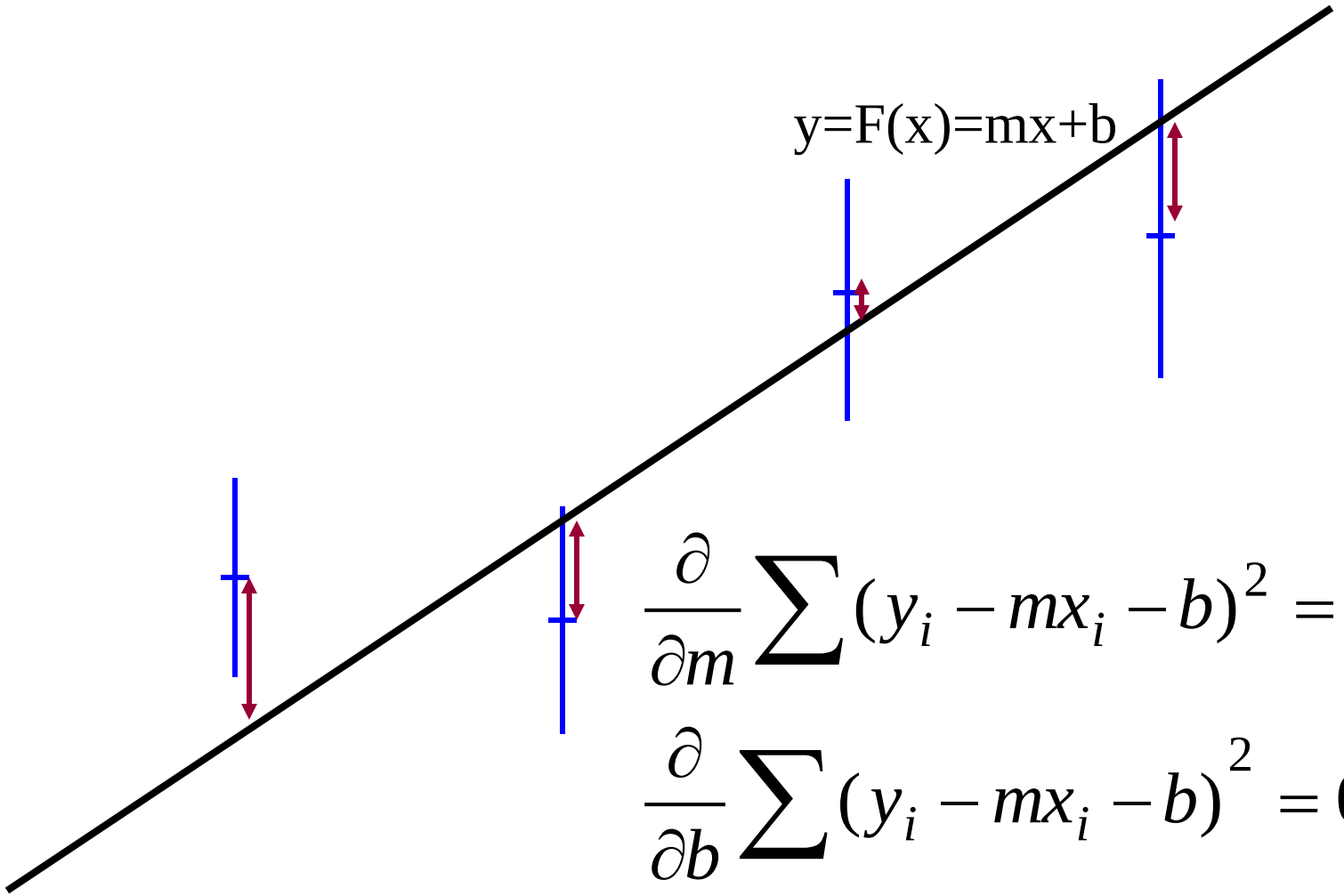


Linear regression



Simple linear regression

Hypothesize a line



$$\frac{\partial}{\partial m}\sum (y_i - mx_i - b)^2 = 0$$

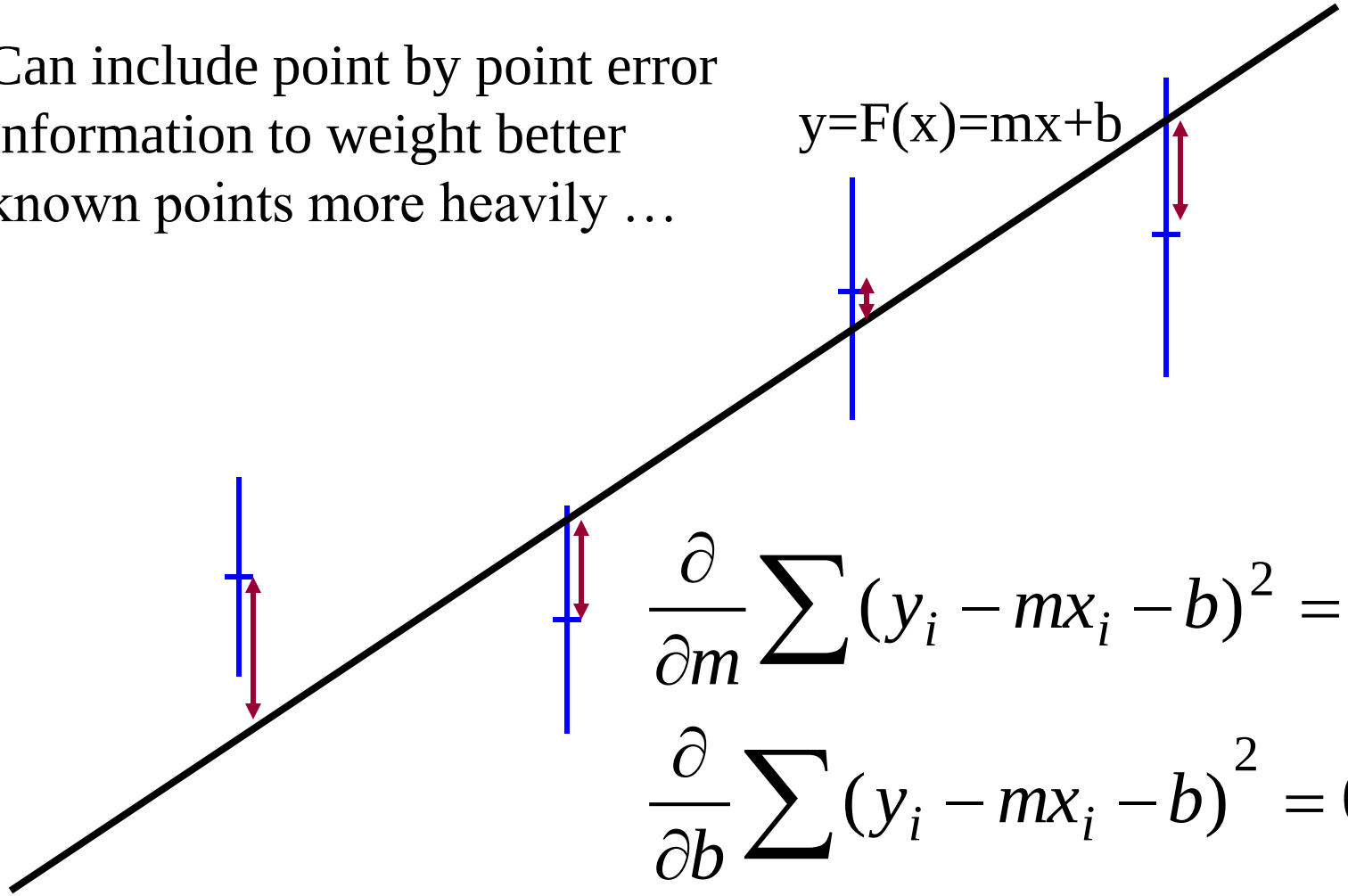
$$\frac{\partial}{\partial b}\sum (y_i - mx_i - b)^2 = 0$$

Minimize sum of squared “residuals” of data with respect to the line (m and b).

Simple linear regression

Hypothesize a line

Can include point by point error information to weight better known points more heavily …



$$\frac{\partial}{\partial m}\sum (y_i - mx_i - b)^2 = 0$$

$$\frac{\partial}{\partial b}\sum (y_i - mx_i - b)^2 = 0$$