PHY 517 / AST 443:
Observational Techniques in Astronomy

Lecture 4:
Statistics

Lab 1

- remember to do your weekly check-ins for Lab 0 date with me or the TAs
- Preparing your lab report as a jupyter notebook:
  - code, plots, etc. can be the same as your lab-mates
  - documentation has to be *in your own words*

Lab preparation

- read, and understand, the lab instructions
- we will quiz you
- if you are not prepared, you will not get to observe

(A brief intro to)

Statistics

Statistics in Astronomy

- we are almost always working in the low signal-to-noise regime
- have to be very careful to make correct inferences from our data!
- robust (and advanced) statistical techniques play a very important role in astronomy

Measurements

- example: 99.123 ± 0.005

- what is 0.005 called?
  - (measurement) uncertainty
  - NOT "error" (inaccurate, though often used)

- what does this mean?
  - if we repeat the measurement many times, in 68% of the cases the true value would fall within the quoted uncertainty interval
  - not-quite-right interpretation: the quoted interval has a 68% chance of containing the true value

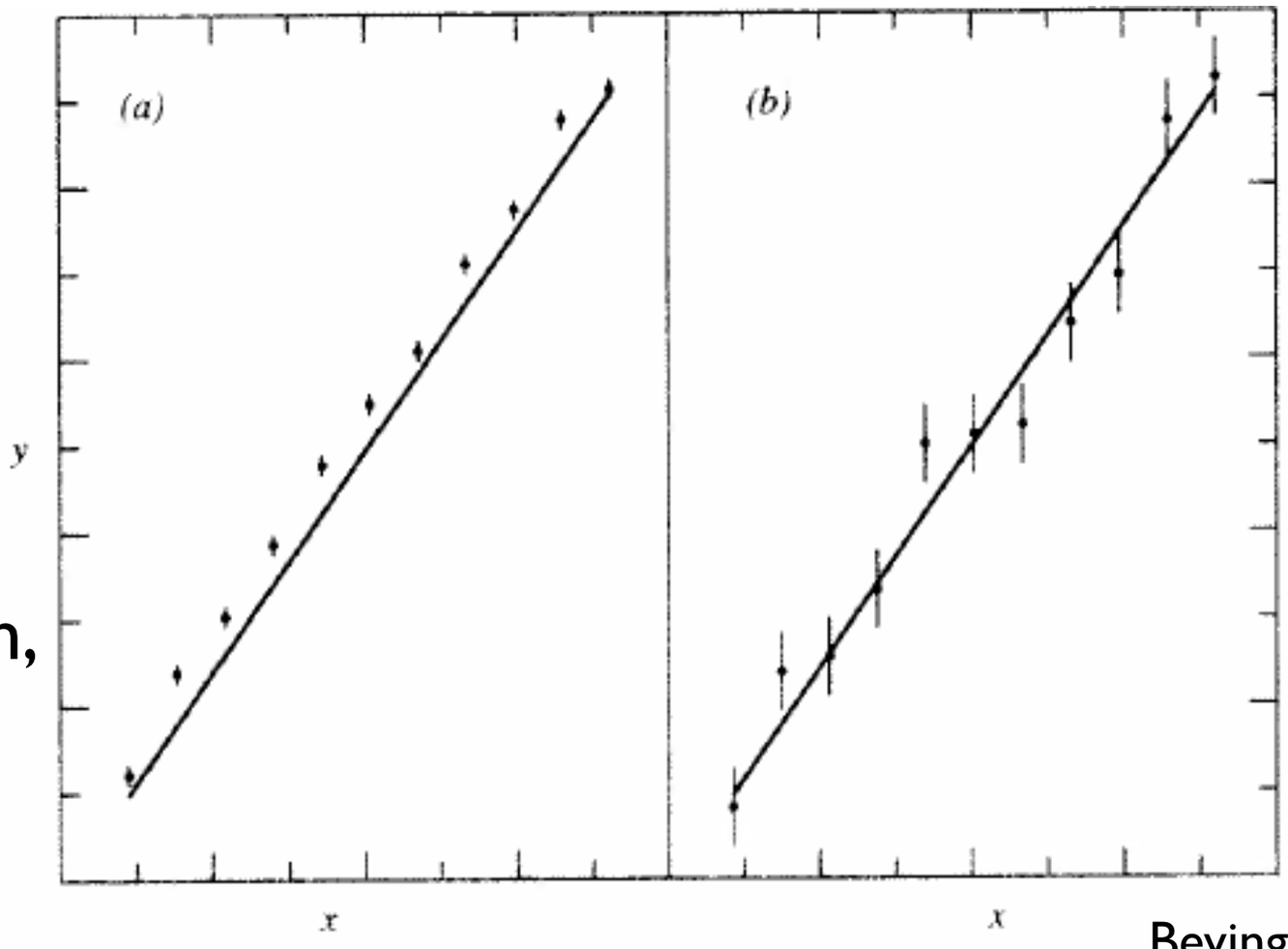
“Error”

- error: difference between *measured* and *true* value
- can be due to:
  - random fluctuations (statistical error)
  - instrumental / algorithmic limitations (systematic error)
  - mistakes (illegitimate error)
- measurements are meaningless if not accompanied by an estimate of the error
- but truth is unknown, have to estimate error indirectly

Accuracy vs. Precision

- accuracy: how close a measurement is to the truth
- precision: size of (statistical) measurement uncertainty

high precision,
low accuracy



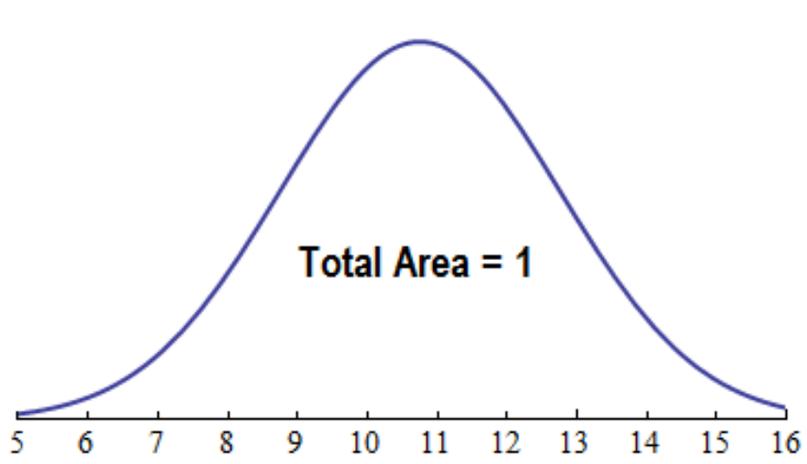
low precision,
high accuracy

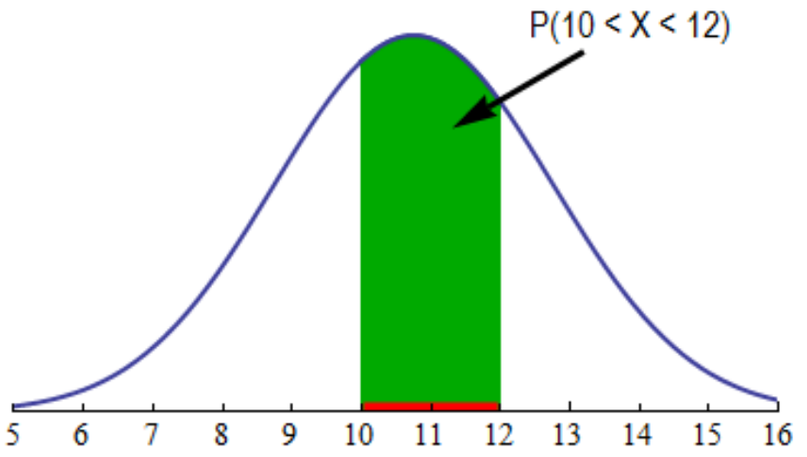
Bevington & Robertson 2003

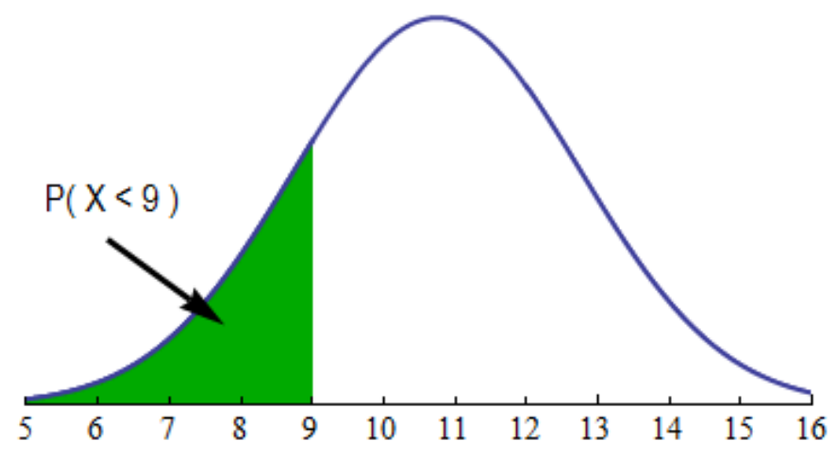
FIGURE 1.1
Illustration of the difference between precision and accuracy. (*a*) Precise but inaccurate data. (*b*) Accurate but imprecise data. True values are represented by the straight lines.

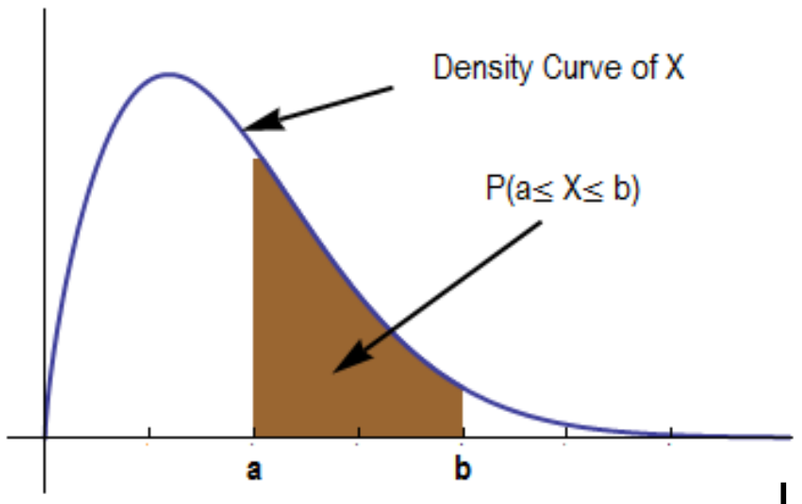
Probability Distributions

- probability distributions: describe expected / measured distributions of measurements
- integrate over range of values to find probability to be in that range









Lumen Learning

Sample vs. Parent Distribution

- measurement x_i of a quantity x:
  - approximates x
  - not necessarily equal to x because of statistical uncertainty
- many measurements x_i:
  - expected to be distributed about true value
  - sample distribution

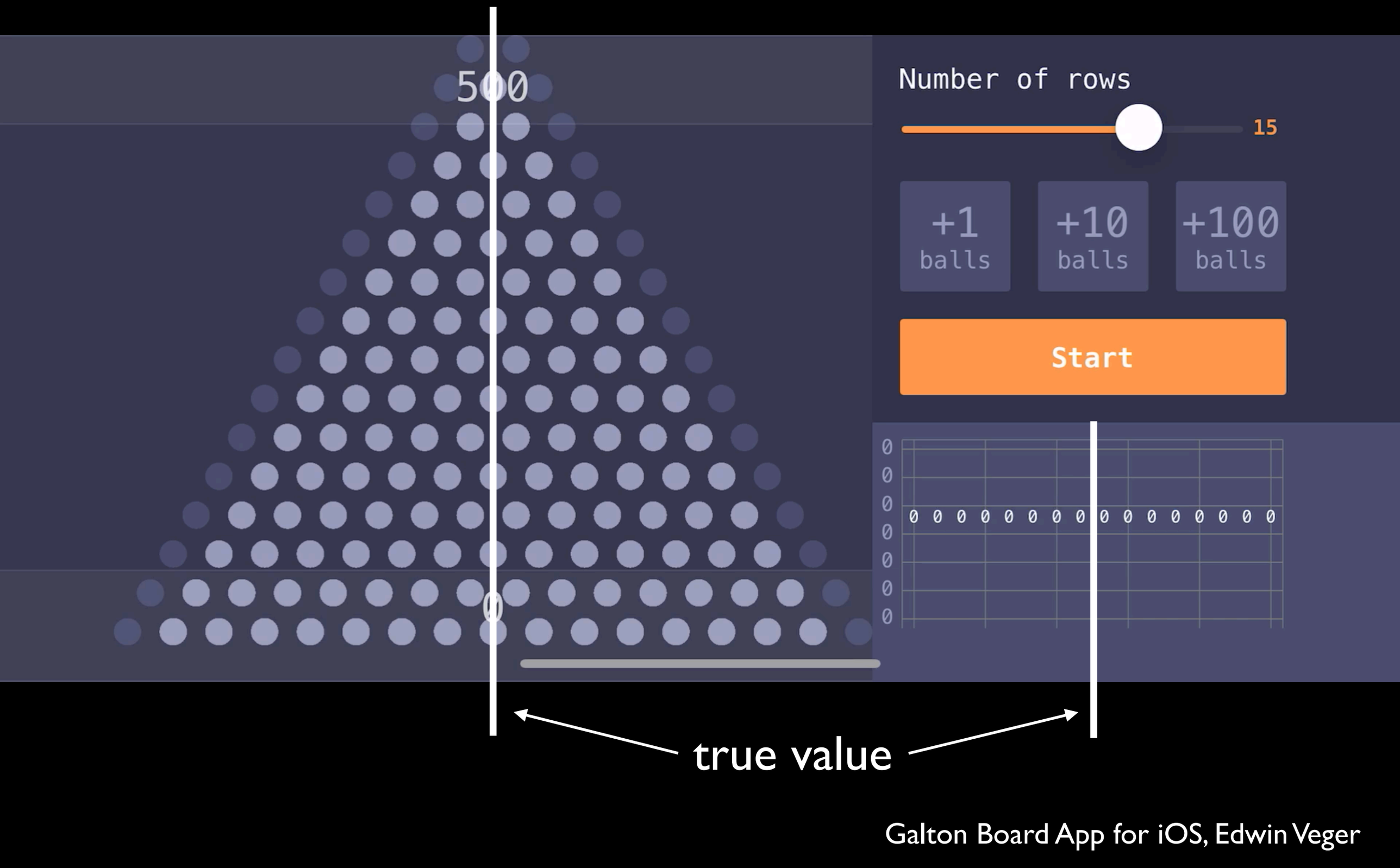
- parent distribution:
  - probability of particular result from single measurement
  - idealized outcome of infinite number of measurements

the sample distribution *samples* the parent distribution

Galton Board:

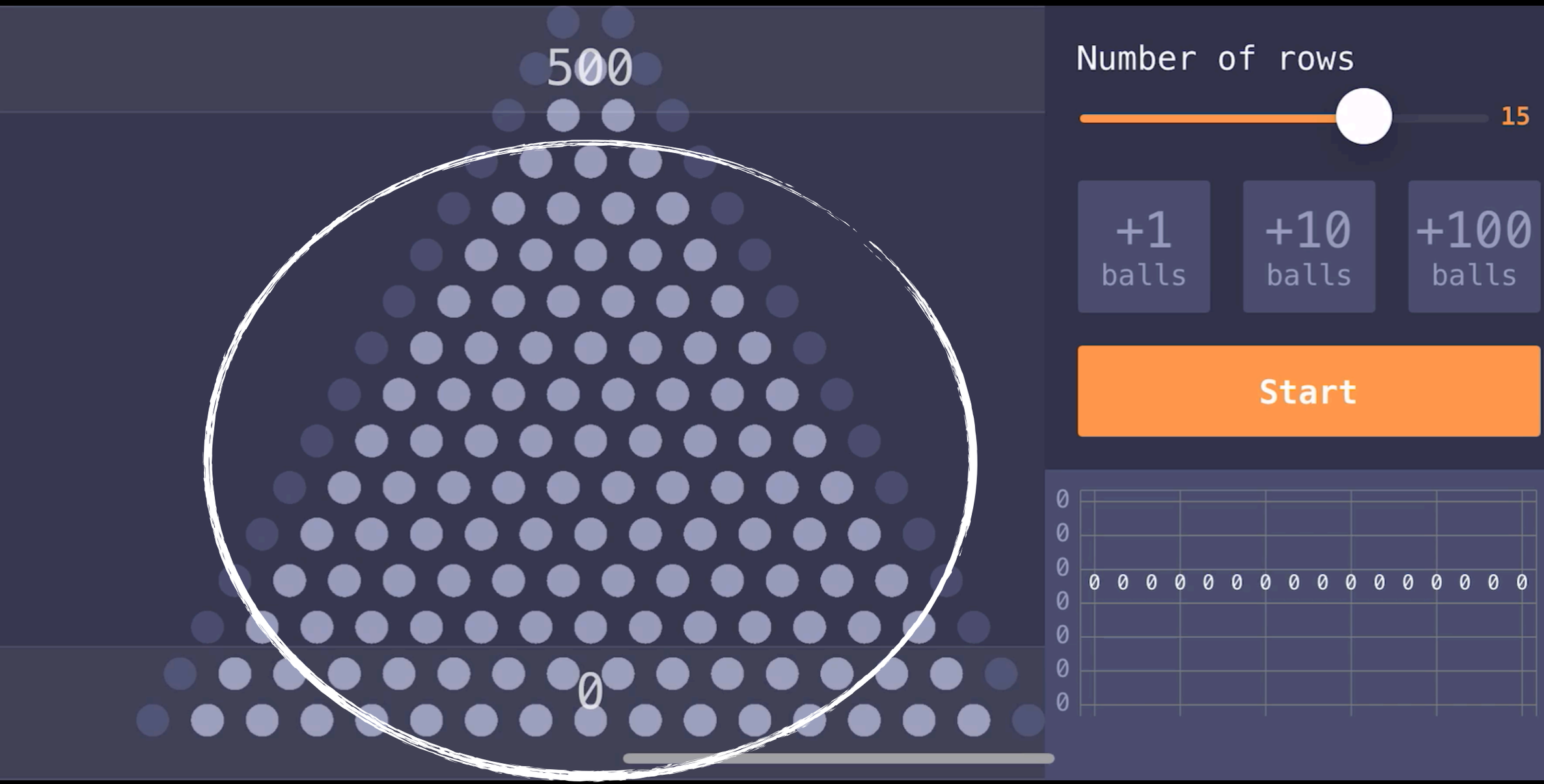
Galton Board App for iOS, Edwin Veger

Galton Board:



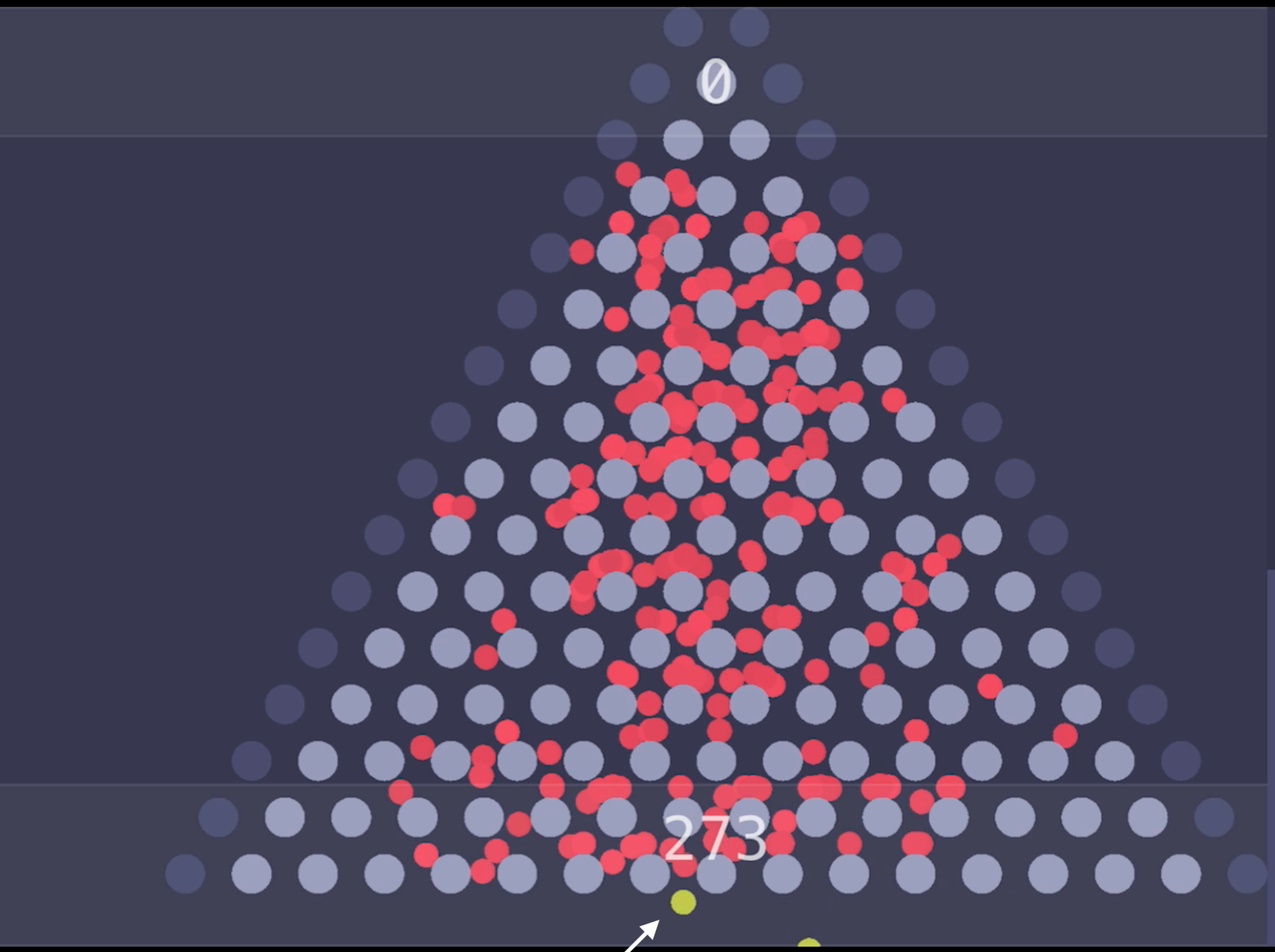
Galton Board App for iOS, Edwin Veger

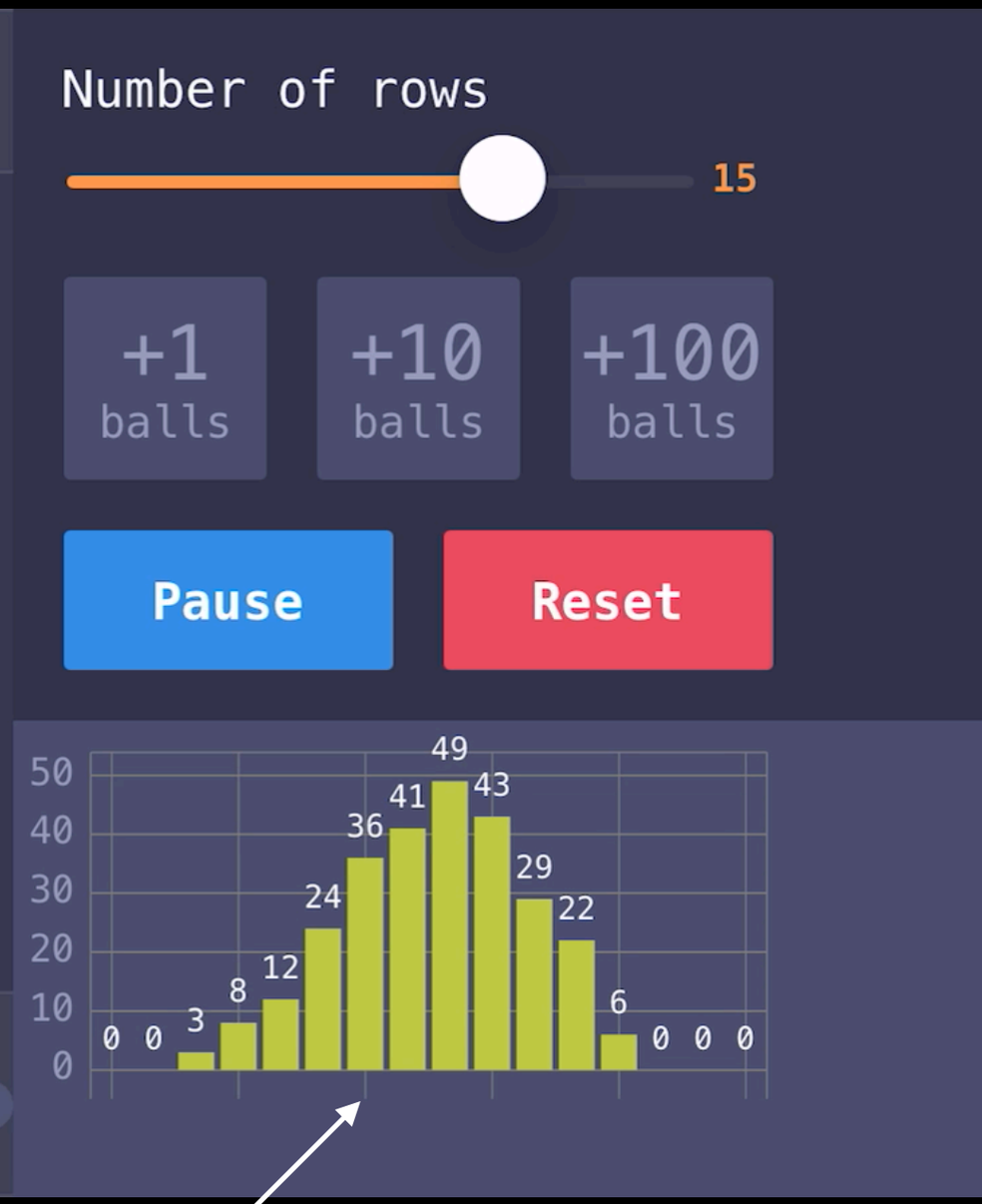
Galton Board:



Galton Board App for iOS, Edwin Veger

Galton Board:



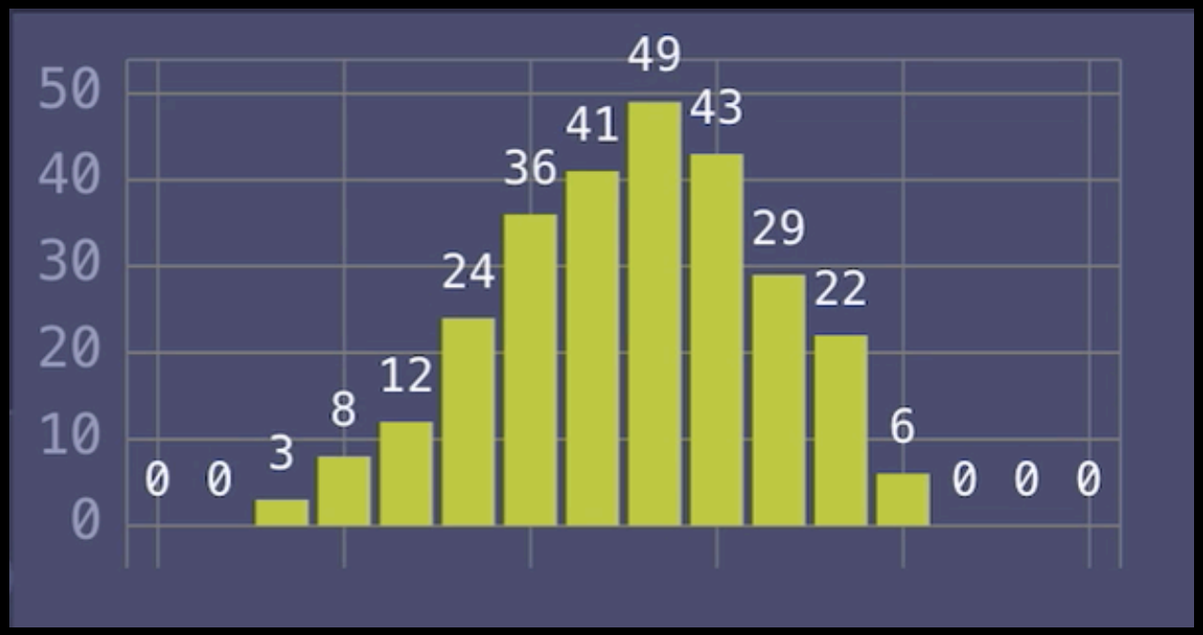


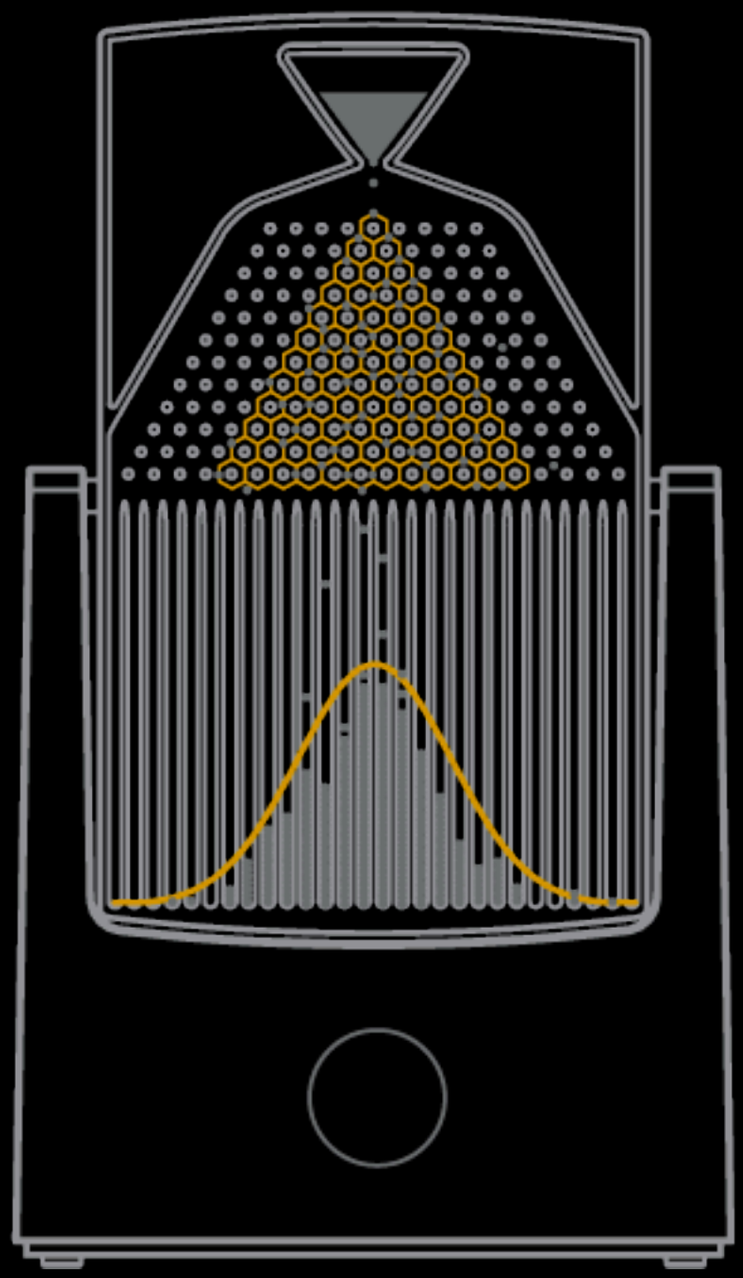
one measurement

many measurements

Galton Board App for iOS, Edwin Veger

Galton Board:





sample population: distribution of many measurements

parent population: Gaussian / normal distribution

Galton Board App for iOS, Edwin Veger

Summary statistics

- only the full sample distribution is the full description of your data
- but usually, it is helpful to describe the sample distribution with a few numbers → summary statistics

Q: can you think of examples?

Mean, median, and mode

- (unweighted) mean of the sample distribution:

$$\bar{x} = \frac{1}{N} \sum_i x_i$$

- IF there are no systematic errors, the mean of the parent population is:

$$\mu = \lim_{N \to \infty} \frac{1}{N} \sum_i x_i$$

Mean, median, and mode

- median: 50th percentile of distribution (half the measurements are smaller, half are greater)

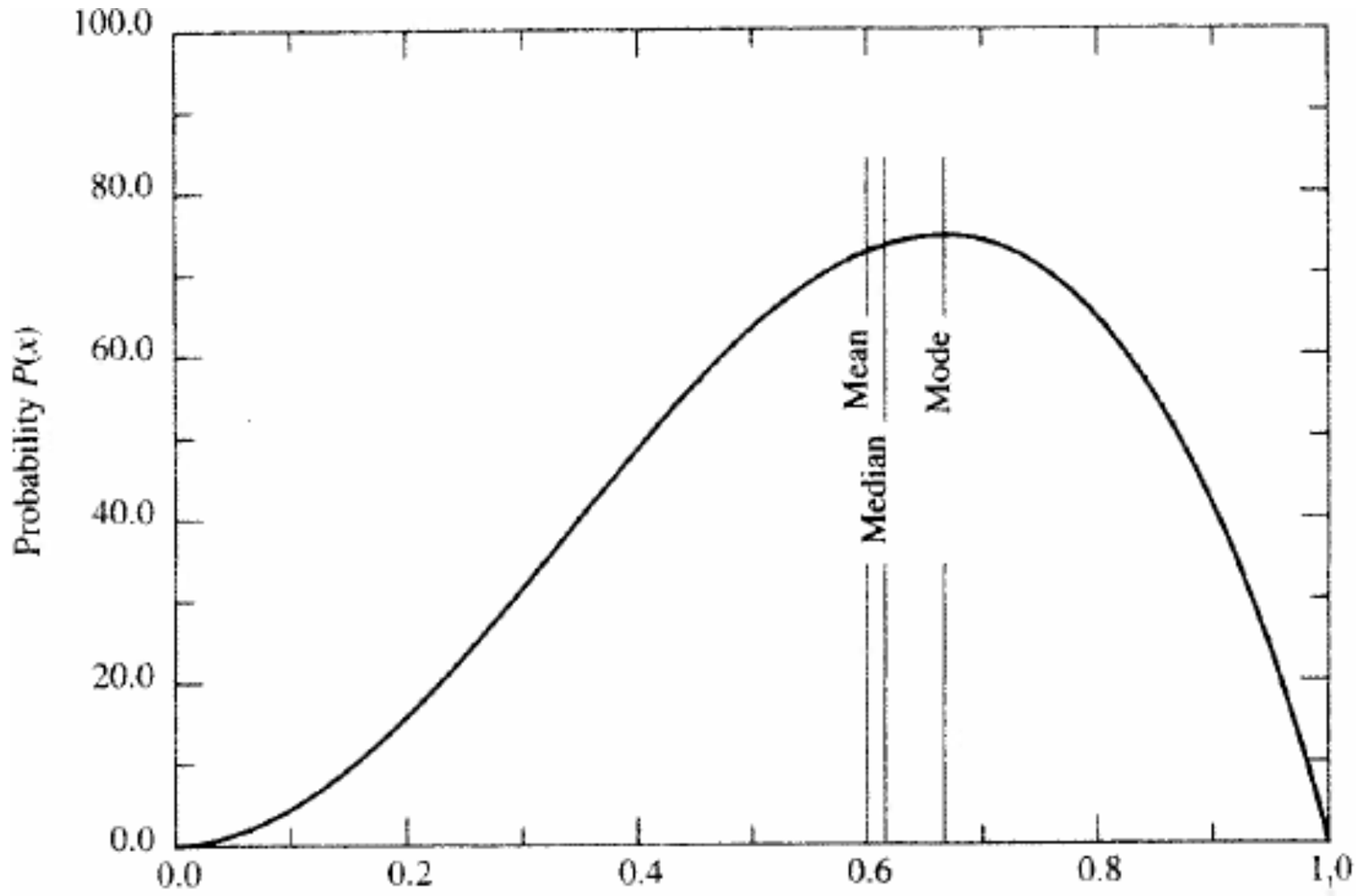
  motivation: less susceptible to "outliers" than the mean

- mode: the most common measurement value

  motivation: the most likely value

Mean, median, and mode

mean, median and mode for an example distribution:



Bevington & Robertson 2003

- generally not equal to each other
- all 3 are useful; which to quote depends on the problem (and personal preference)

Deviation / variance / std. deviation

- deviation of one measurement: $d_i = x_i - \mu$

- sample variance: average of the squares of the deviations

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

- when computing from sample population:

$$s^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2$$

- standard deviation: $\sigma = \sqrt{\text{variance}}$

indicates how much the measurements typically deviate from the mean

How well can we measure the mean?

- width of distribution (standard deviation) remains the same
- the more measurements, the closer the mean to the true value

Galton Board App for iOS, Edwin Veger

Uncertainty on the mean

- variance and std. deviation are measures of the *width* of a distribution

- with increasing number of measurements, the typical deviation of measured mean and true value decrease

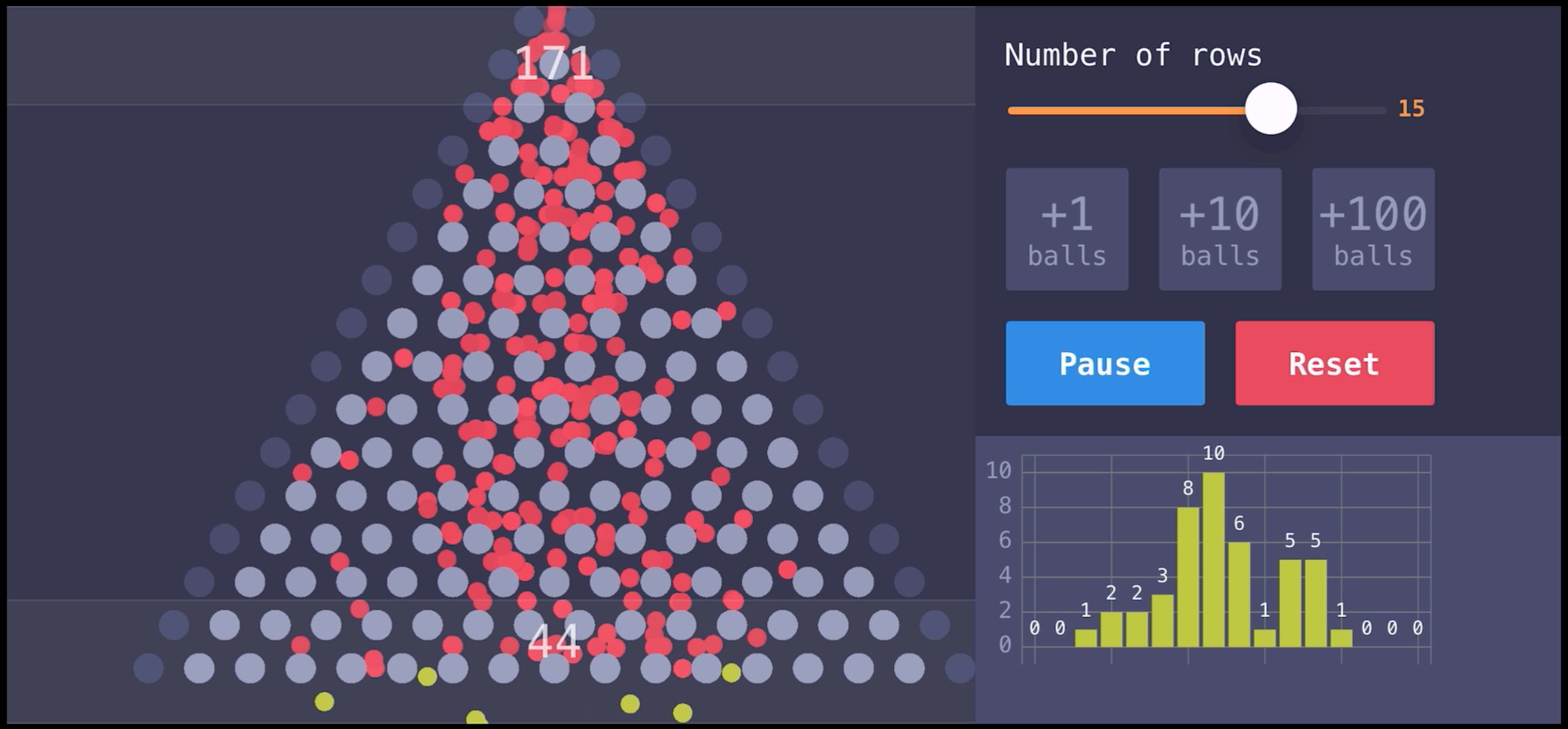
- measurement uncertainty on the mean:

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

- width of distribution of repeated measurements of the mean

- σ: distribution of single measurements around the true value

- σ_x: distribution of means of N measurements around the true value

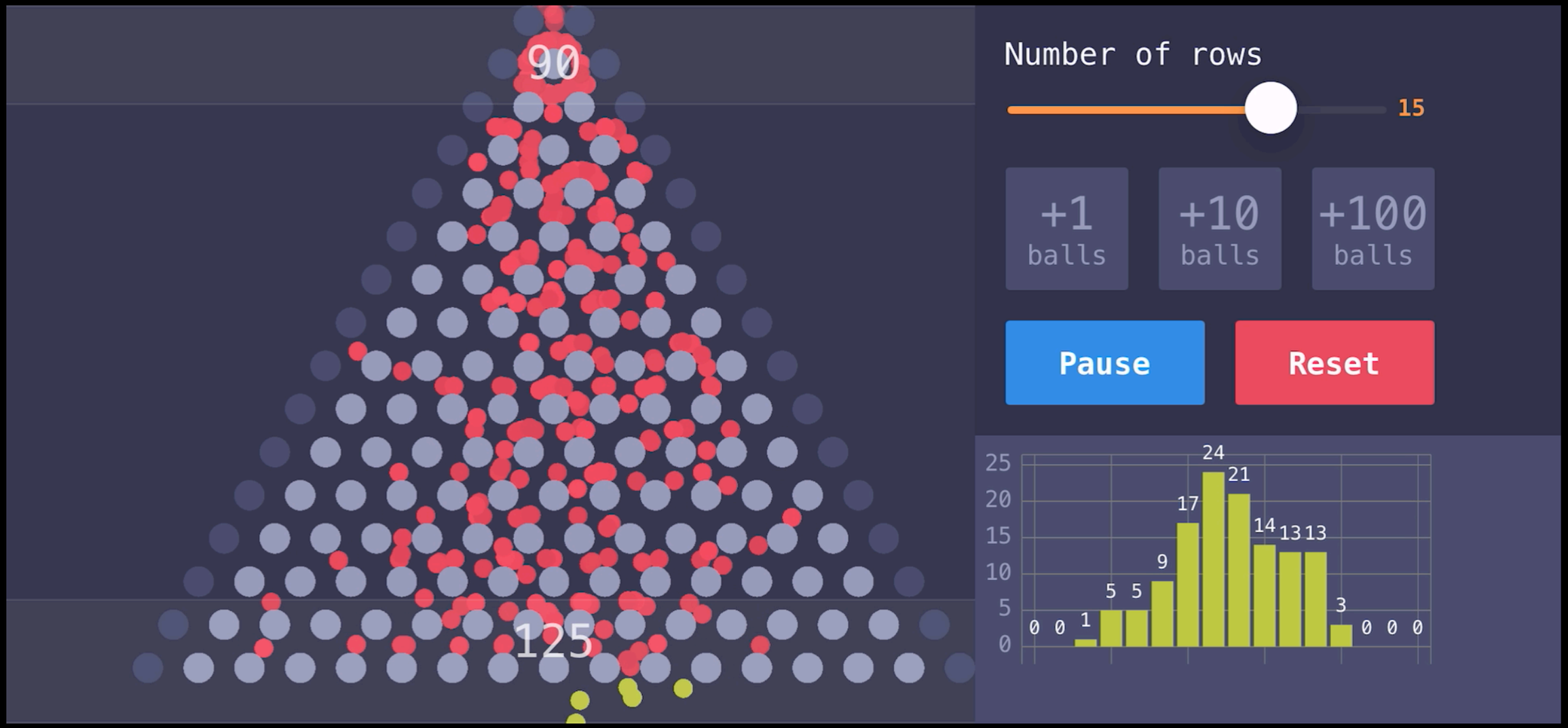


many measurements:

value: $$\bar{x} = \frac{1}{44}\sum_{i=1}^{44} x_i$$

$$\sigma = \sqrt{\frac{1}{43}\sum_{i=1}^{44}(x_i - \bar{x})^2}$$

uncertainty: $$\sigma_{\bar{x}} = \sigma/\sqrt{44}$$

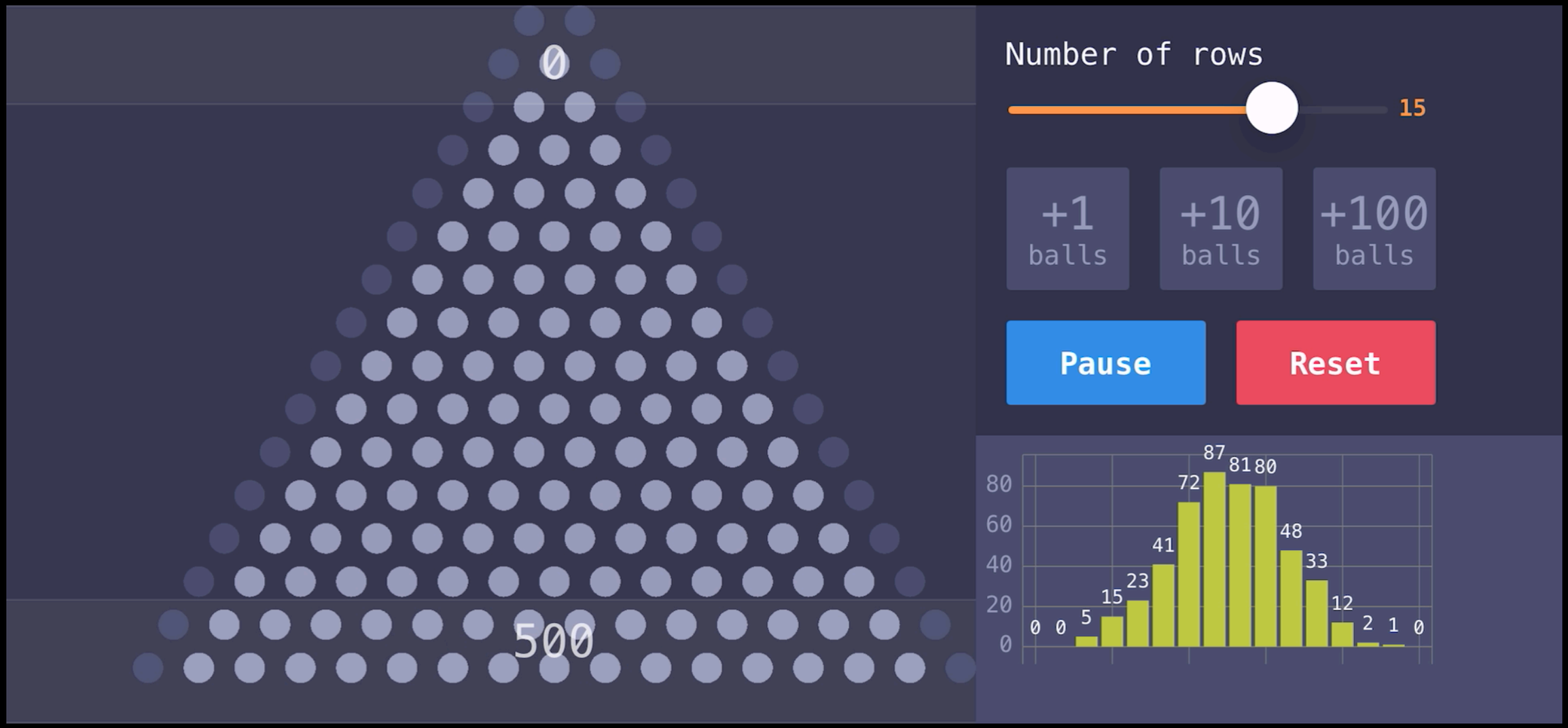


many measurements:

value: $$\bar{x} = \frac{1}{125}\sum_{i=1}^{125} x_i$$

$$\sigma = \sqrt{\frac{1}{124}\sum_{i=1}^{44}(x_i - \bar{x})^2}$$

uncertainty: $$\sigma_{\bar{x}} = \sigma/\sqrt{125}$$



many measurements:

value: $$\bar{x} = \frac{1}{500}\sum_{i=1}^{500} x_i$$

$$\sigma = \sqrt{\frac{1}{499}\sum_{i=1}^{500}(x_i - \bar{x})^2}$$

uncertainty: $$\sigma_{\bar{x}} = \sigma/\sqrt{500}$$

Weighted mean

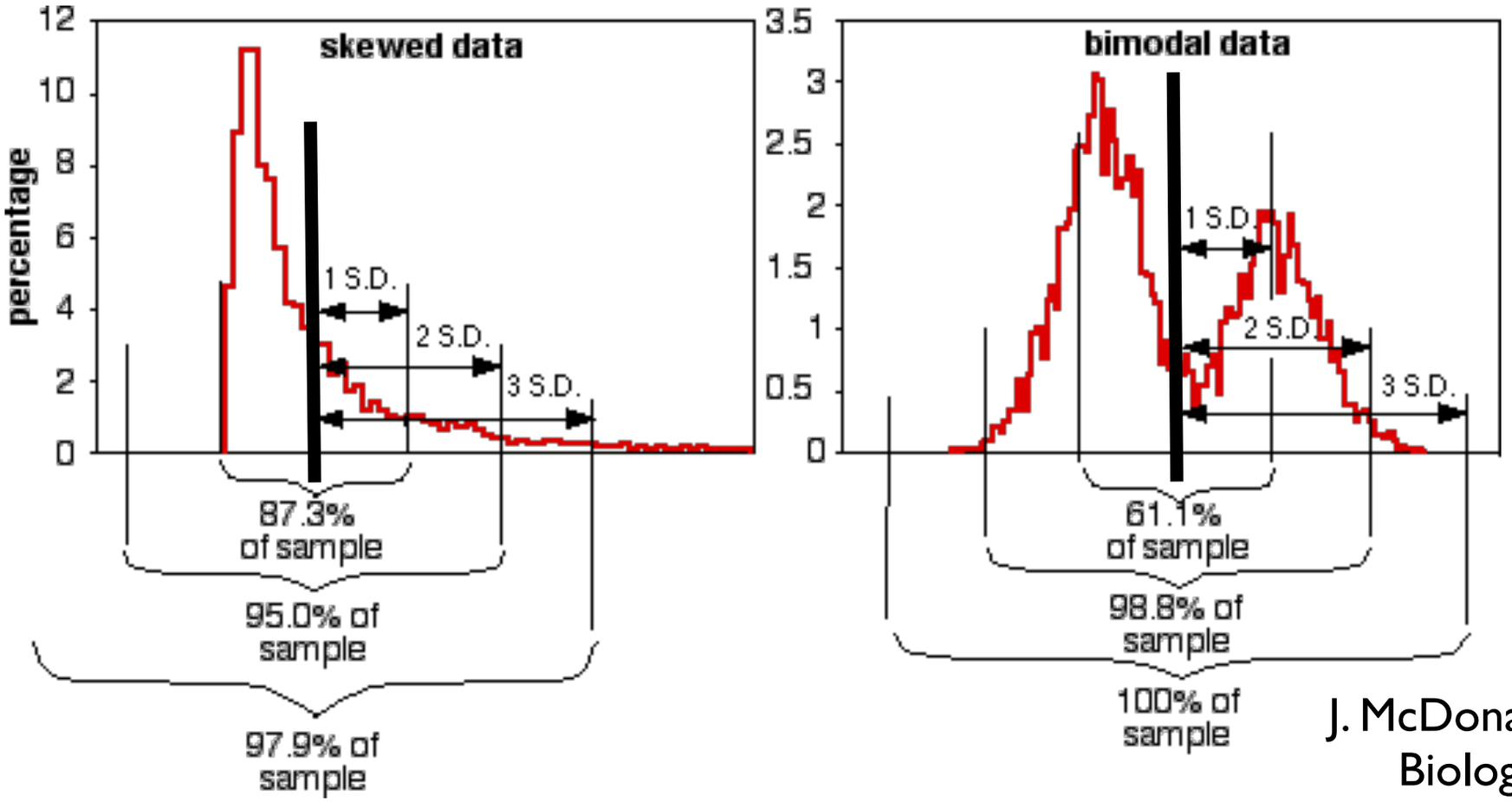
- previously, all measurements had equal weight
- some measurements are more precise than others; can assign weight w_i to each measurement x_i
- weighted mean:

$$\bar{x} = \frac{\sum_i w_i x_i}{\sum_i w_i} \qquad \sigma^2_{\bar{x}} = \frac{1}{\sum_i w_i}$$

- for reasonable (Gaussian) distributions, optimal weight is the inverse of the variance of each measurement:

$$\bar{x} = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2} \qquad \sigma^2_{\bar{x}} = \frac{1}{\sum_i 1/\sigma_i^2}$$

- can calculate mean, variance, etc. for any set of data points
- *that does not guarantee that they are useful descriptions of the distribution !*



J. McDonald, Handbook of Biological Statistics

- if we know the shape of the parent distribution, we know which summary statistics to use

Common Probability Distributions

- many, many possible distributions have been quantified; here, consider 3 particularly important ones:
  - Binomial distribution: for experiments with only a small number of possible final states (e.g. coin toss)
  - Poisson distribution: counting experiments for discrete events (e.g. photon counts)
  - Gaussian (or Normal) distribution: distribution of events about the mean for a wide variety of processes; limiting case of binomial and poisson distributions

Binomial Distribution

- experiment with only two possible outcomes:
  - state 0: probability p
  - state 1: probability q = (1-p)
- n realizations
- probability that x of the n realizations are in state 0:

$$P_{\mathrm{B}}(x|n,p) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

- x: positive integers; $0 \le x \le n$
- $0 < p < 1$
- $\sum_{x=0}^{n} P_{\mathrm{B}}(x|n,p) = 1$

Binomial Distribution

- mean of the binomial distribution:

$$\mu = \sum_{x=0}^{n} \left(x \cdot P_{\mathrm{B}}(x|n,p) \right) = np$$

(agrees with intuition!)

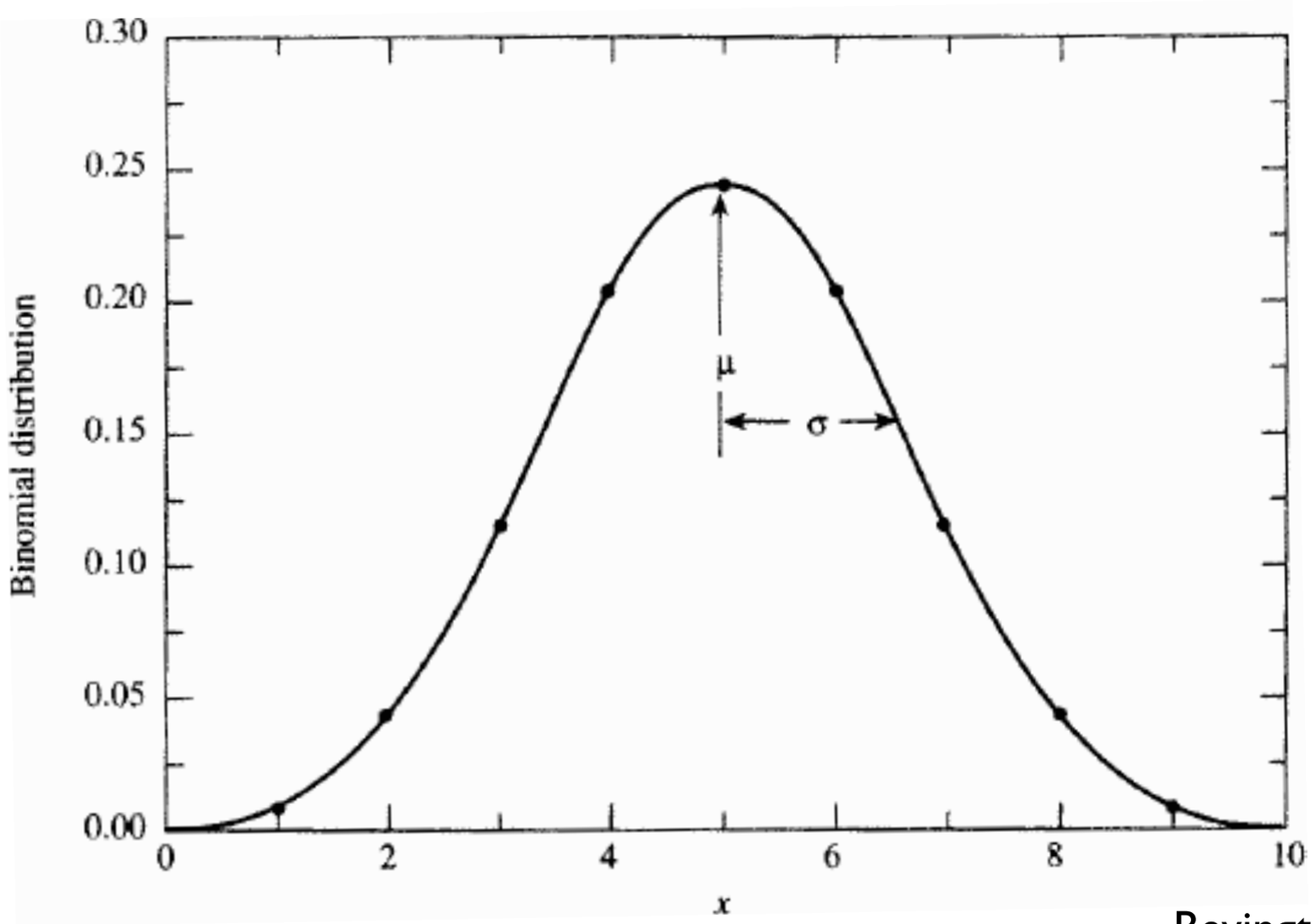
- variance of the binomial distribution:

$$\sigma^2 = \sum_{x=0}^{n} \left((x-\mu)^2 \cdot P_{\mathrm{B}}(x|n,p) \right) = np(1-p)$$

Binomial Distribution - Example (1)

tossing 10 coins, x = number of tails

$n = 10\,,\; p = 0.5 \quad \rightarrow \quad P(x) = P_{\mathrm{B}}(x|10, 0.5)$



$$\mu = np = 5$$

$$\sigma^2 = np(1-p) = 2.5$$

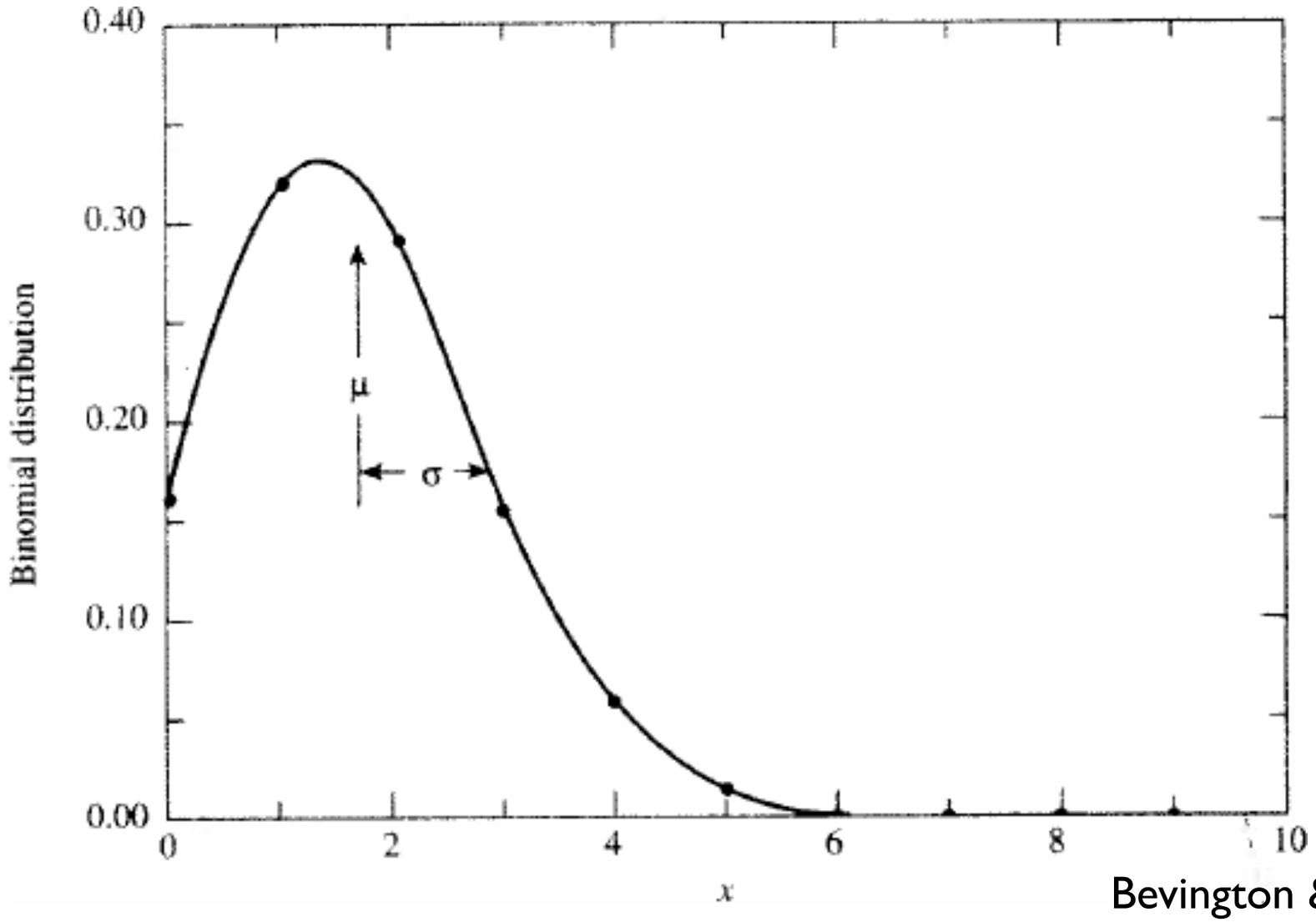
Bevington & Robertson 2003

since q=p: distribution is symmetric

Binomial Distribution - Example (2)

roll 10 dice, x = number of rolls with 6 eyes

$$n = 10\,,\; p = 1/6 \quad \rightarrow \quad P(x) = P_{\mathrm{B}}(x|10, 1/6)$$



$$\mu = np = 5/3$$

$$\sigma^2 = np(1-p)$$

$$= 1.39$$

note: mean is not the mode

Bevington & Robertson 2003

since q≠p: distribution is not symmetric

Poisson Distribution

- limit of binomial distribution if number of trials is large, and probability of "success" in a given trial is small, while the mean $\mu = np$ remains finite

$$P_{\mathrm{P}}(x|\mu) = \lim_{\substack{n\to\infty \\ p\to 0}} P_{\mathrm{B}}(x|n,p) = \frac{\mu^x}{x!}e^{-\mu}$$

- for experiments where a mean number of events, μ, can be measured
- neither the number of realizations n, nor the probability of "success" p need to be known
- example: "counting" experiments, e.g. flux measurements

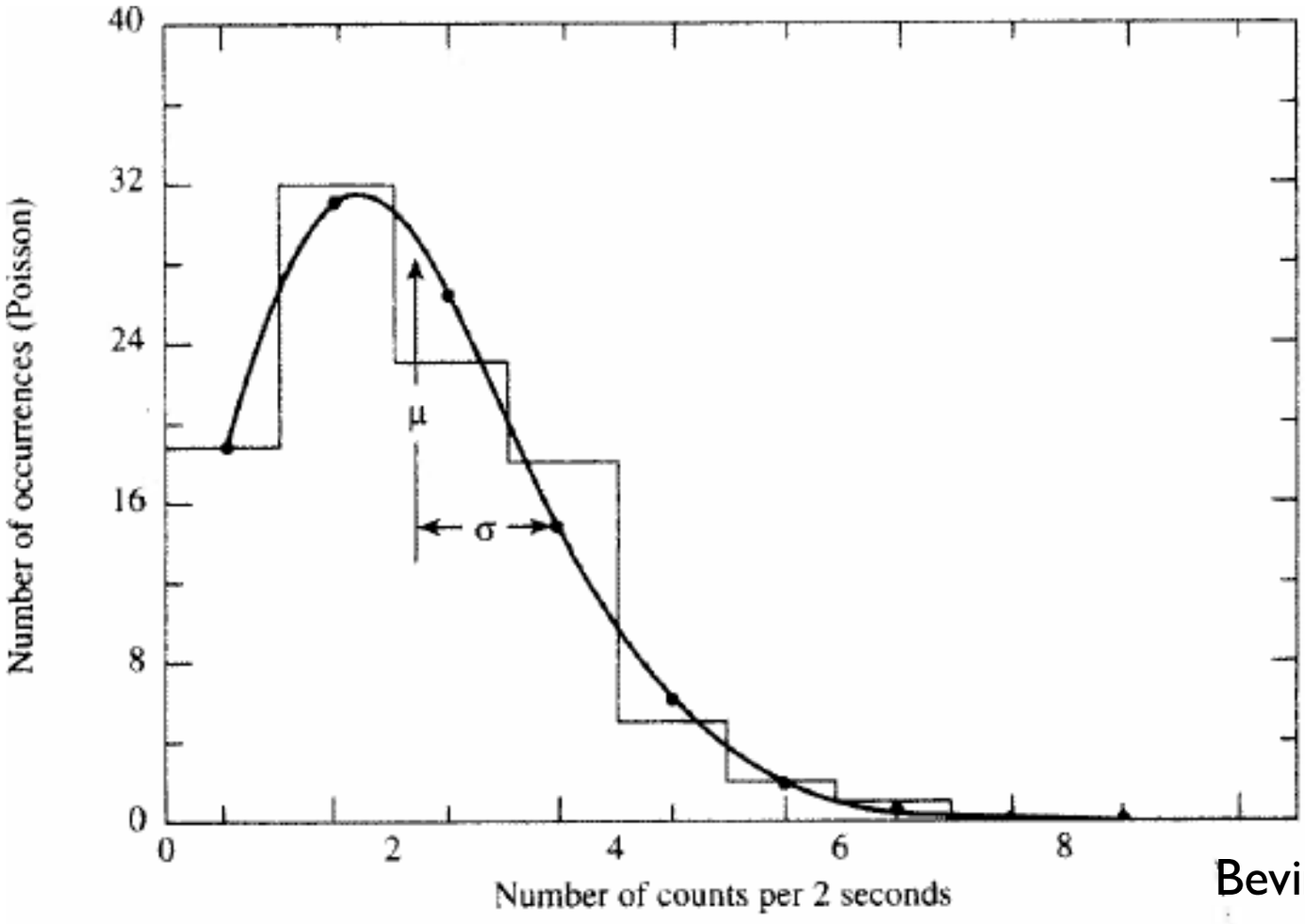
Poisson Distribution

- mean of the Poisson distribution: μ
- variance of the Poisson distribution: $\sigma^2 = \mu$
- standard deviation: $\sigma = \sqrt{\mu}$

- a flux measurement typically consists of measuring a number of events, N, per time interval Δt, with $\mu = N/\Delta t$
- assuming that the time interval is precisely known, the uncertainty on the mean follows from $\sqrt{N}$

Poisson Distribution - Example

a detector measures the number of gamma-ray photons per 2 second interval, making 100 measurements:



measured mean:

$$\bar{x} = 1.69$$

solid curve:

$$P_{\mathrm{P}}(x|1.69)$$

Bevington & Robertson 2003

note: Poisson distribution is defined for positive, integer values of x

Poisson Distribution - Example

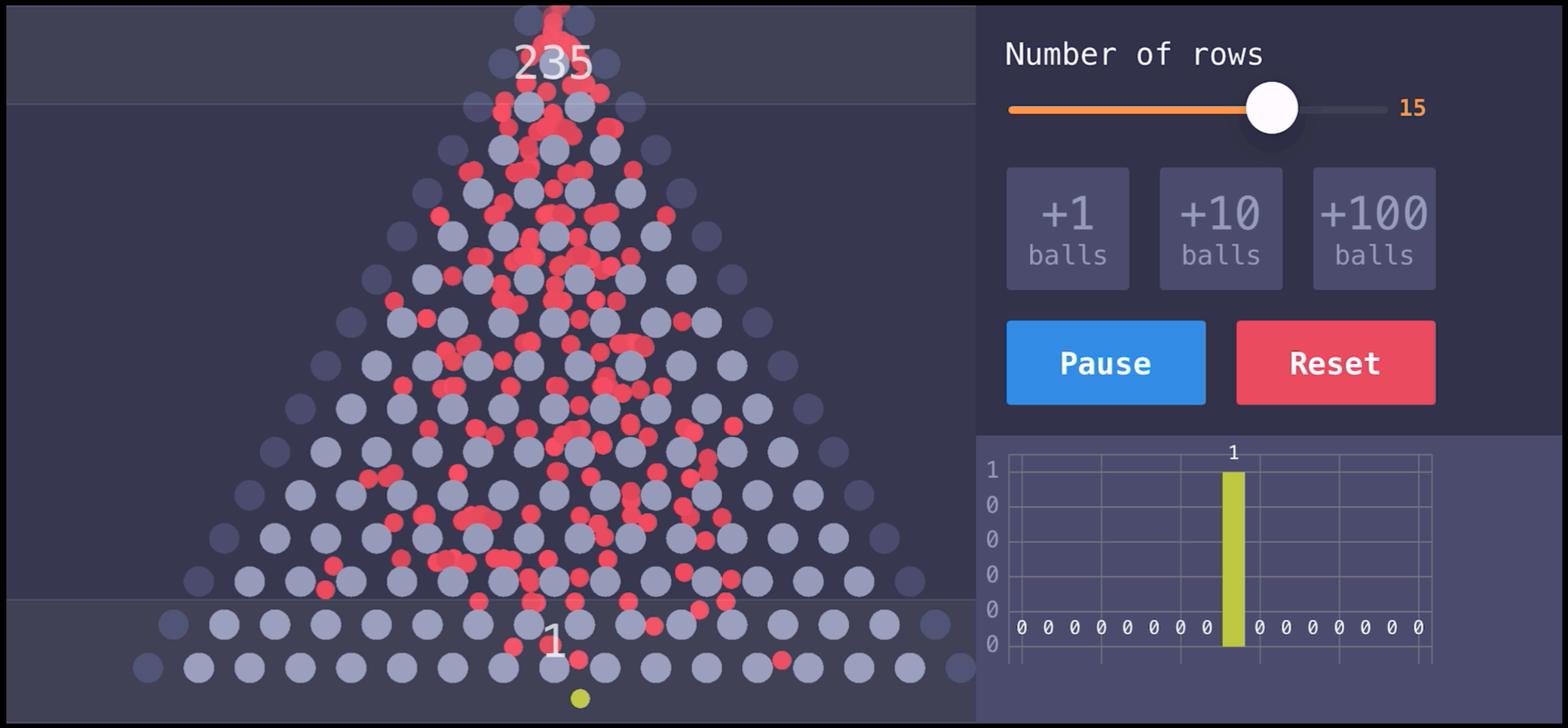
On your CCD, you count a flux of 100 photons from a star.

Q: Report an estimate of the flux (in number of photons) and the uncertainty.

Only 1 measurement → flux estimate (μ) is 100 photons.

Counting experiment: standard deviation is sqrt(μ).

Measurement: F = 100 +/- 10



value: x_1

one
measurement:

uncertainty: $\sqrt{x_1}$ (if Poisson-distributed)

Gaussian / Normal Distribution

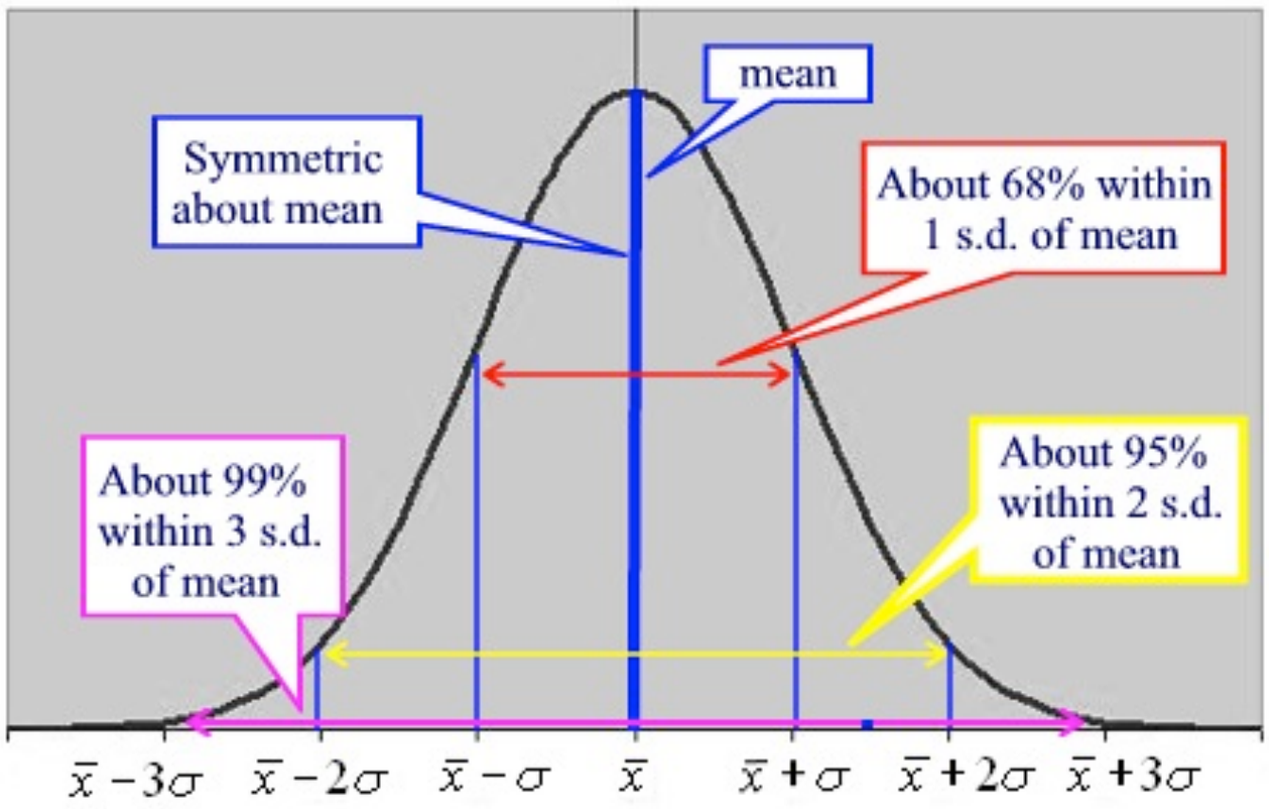
- the most commonly used probability distribution

$$P_{\mathrm{G}}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- mean: μ , standard deviation: σ

- can be derived as limit of the Poisson distribution for large values of the mean, μ≳30

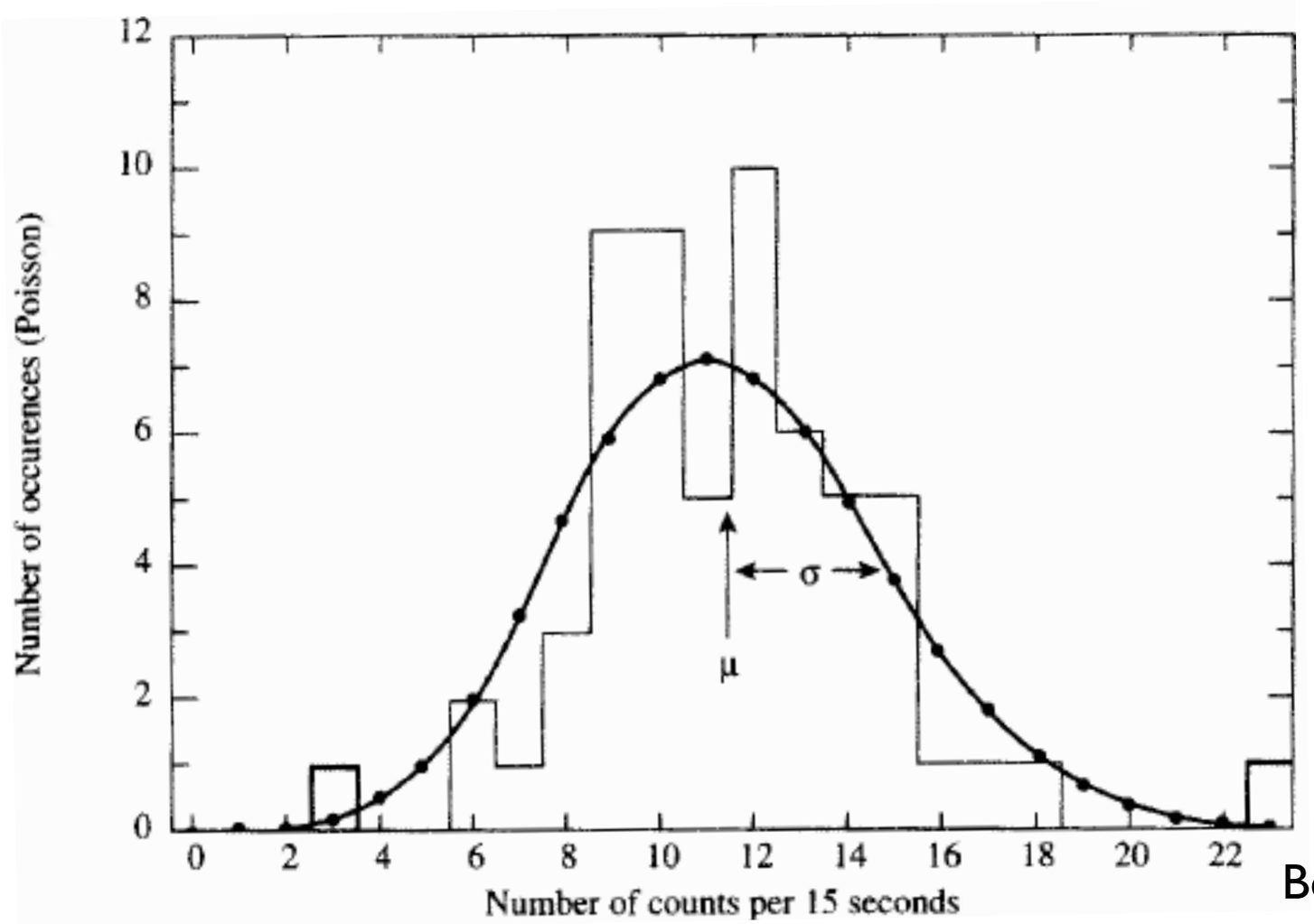
- can also be derived as limit of many other distributions



Regents Prep

Gauss Distribution - Example

a detector measures the number of gamma-ray photons per 15 second interval, making 60 measurements:



measured mean:

$$\bar{x} = 11.48$$

solid curve:

$$P_{\mathrm{P}}(x|11.48)$$

Bevington & Robertson 2003

note: unlike Poisson distribution, Gaussian is continuous and defined for all x

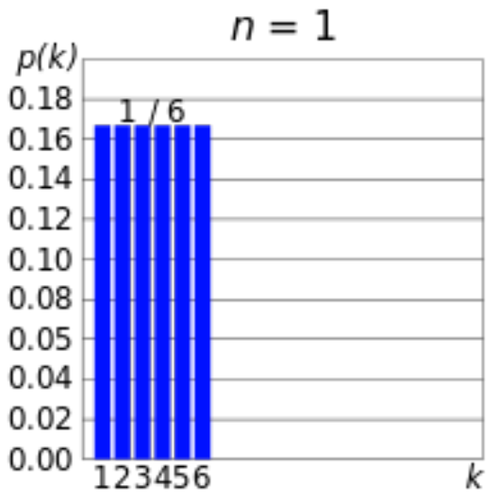
Central Limit Theorem

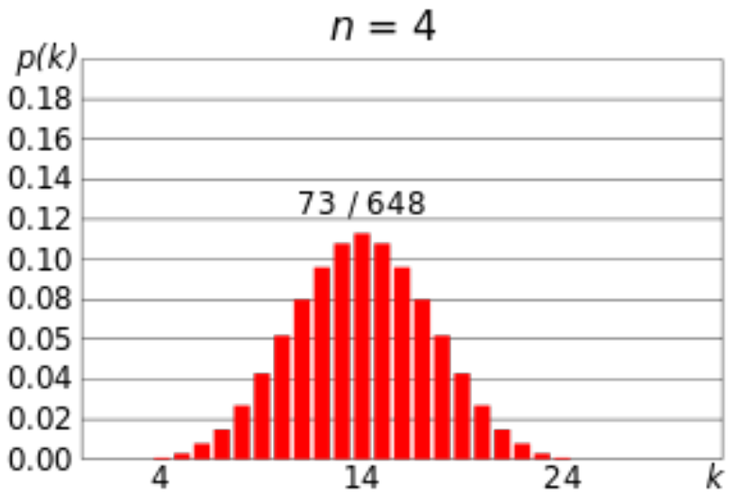
- "the sum of n random values drawn from a probability distribution function of finite variance, σ^2, tends to be Gaussian distributed about the expectation value for the sum, with variance $n\sigma^2$"

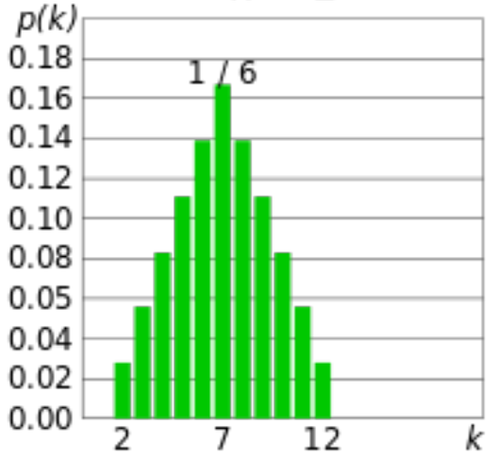
- in other words: the distribution of *the mean of a large number* of *random, independent draws* will tend to a normal distribution

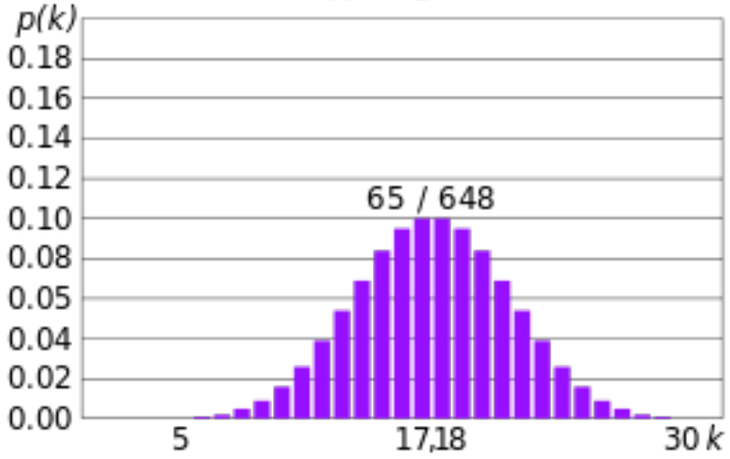
- many processes in nature (that are based on sums or means) described by a normal (or log-normal) distribution

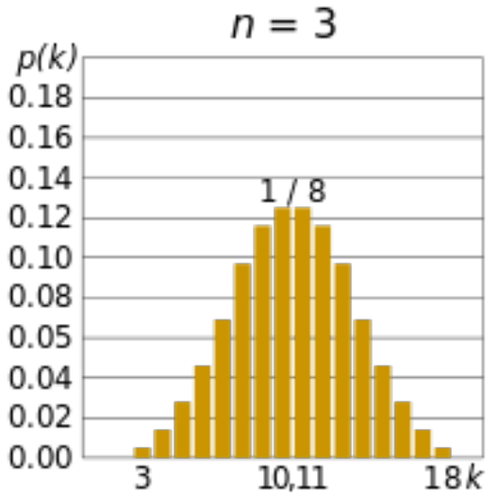
Central Limit Theorem

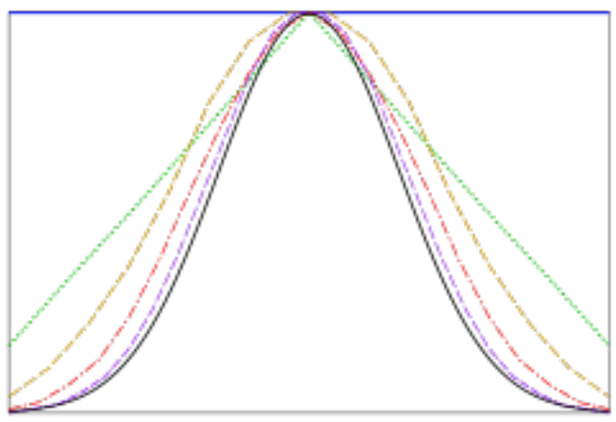










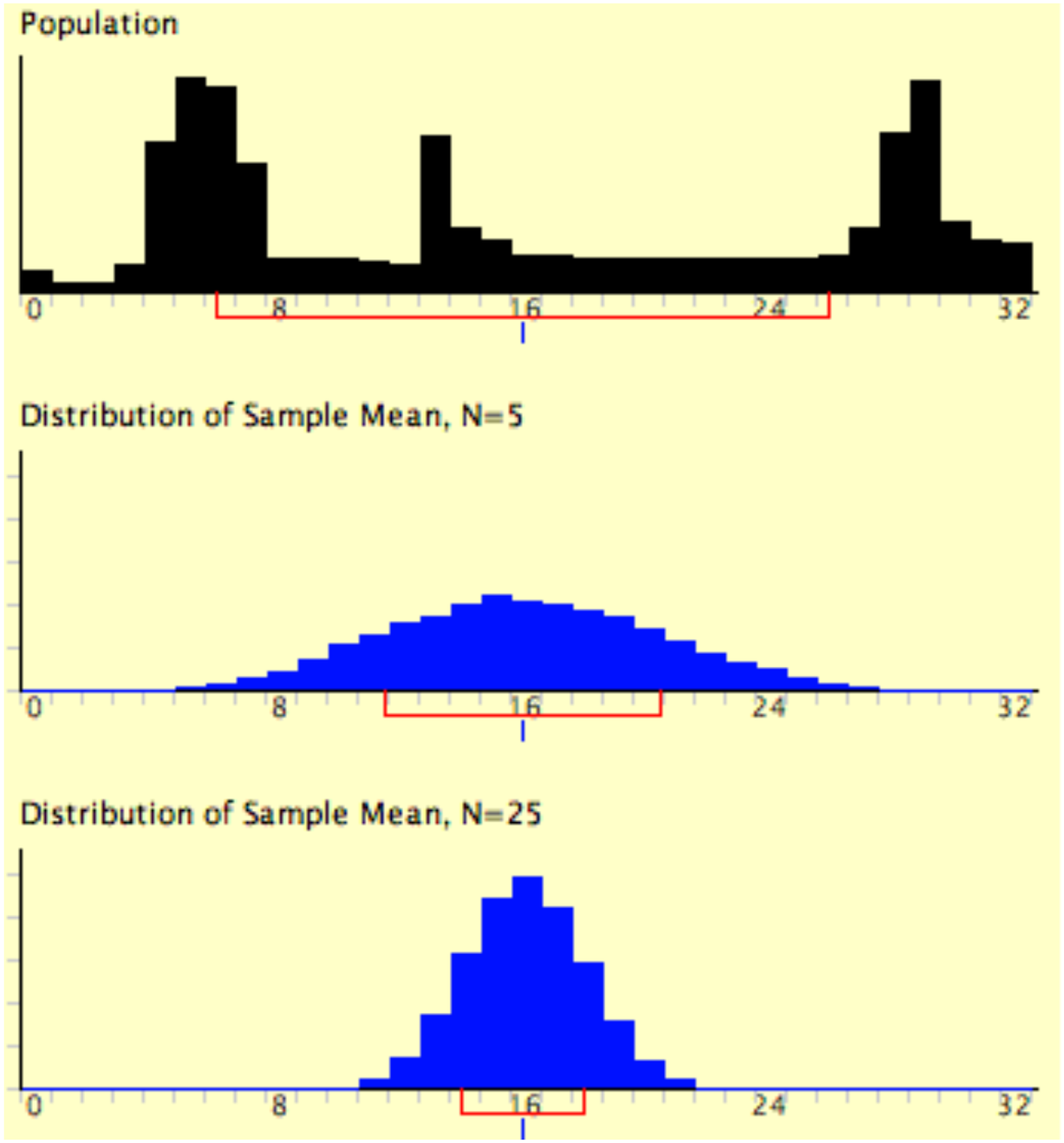


wikimedia

Example: the sum of n dice rolls

Note: individual dice rolls are NOT described by a Gaussian distribution - but the sum is, if n is large

CLT and Uncertainty on the Mean



David Lane

Recall: uncertainty on the mean of N measurements

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

Std. dev. of many measurements of the mean, with N samples each

CLT tells us that distribution of means is normally distributed, even if parent population is not

Photon Counting?

One image:

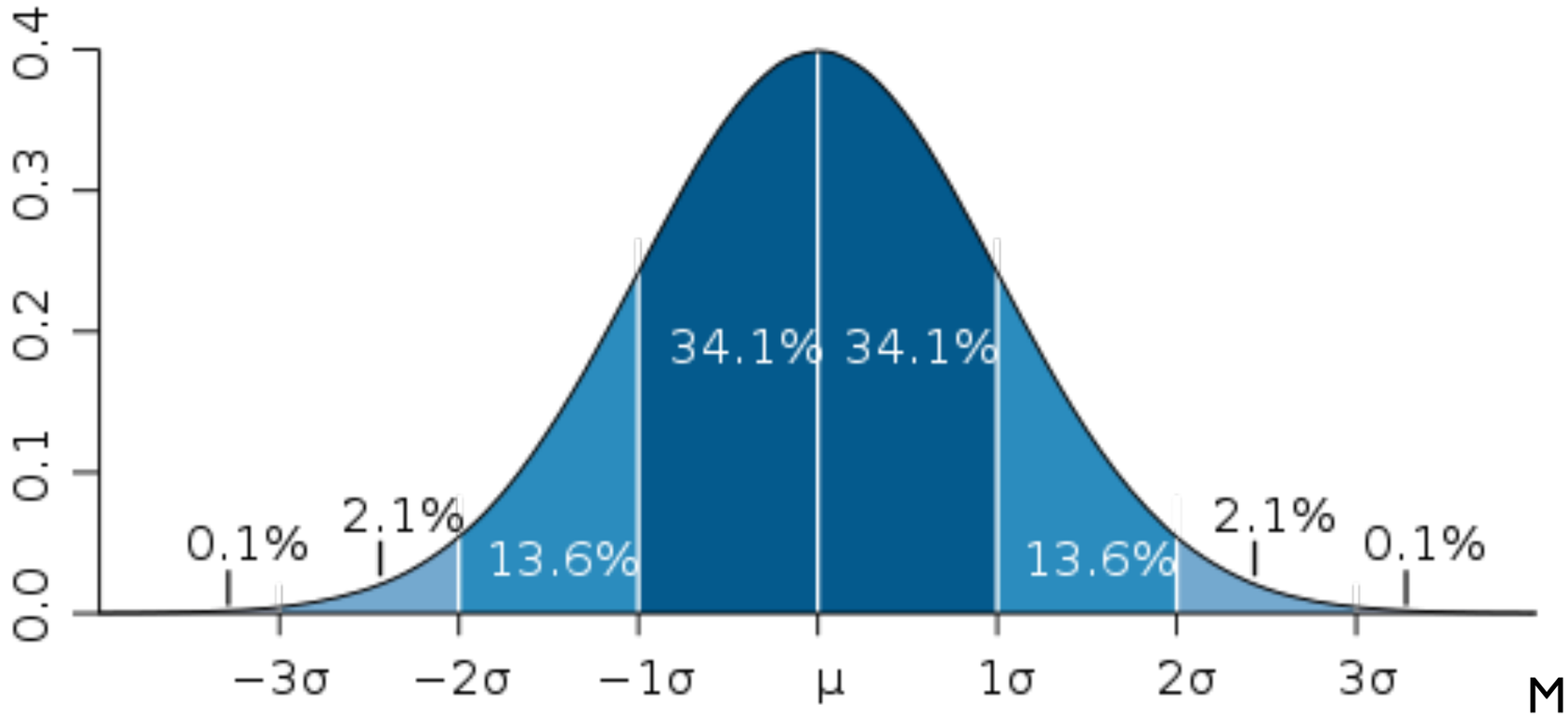
- Flux = $N_{photons}$ / Δt from star - sum of many "draws" of single photons hitting CCD, "success" = detection
- Poisson uncertainty: sqrt($N_{photons}$)

Repeated measurements of the same star:

- Flux measurements are Gaussian distributed (as long as brightness of star does not change)

Gaussian / Normal Distribution

relation between the probability of occurrence and number of standard deviations away from mean:



M. Toews

measurements should fall:

- within 1σ of the mean 68.3% of the time
- within 2σ of the mean 95.4% of the time
- within 3σ of the mean 99.73% of the time

Significant detections?

- the significance of a detection is often quoted in "sigmas" to indicate the probability that the signal is (in)consistent with a random fluctuations

- only a valid measure of probability if the background distribution is Gaussian!

- in particle physics: need >5σ to claim detection
- in astronomy: detections are claimed at >3σ
- don't trust claims below 3σ

Significance of differences

significance of a deviation from the "expected" value: $\frac{|x-\mu|}{\sigma_x}$

|measurement - expectation| / uncertainty = number sigmas

more likely, you are comparing your measurement x_1 to somebody else's measurement x_2

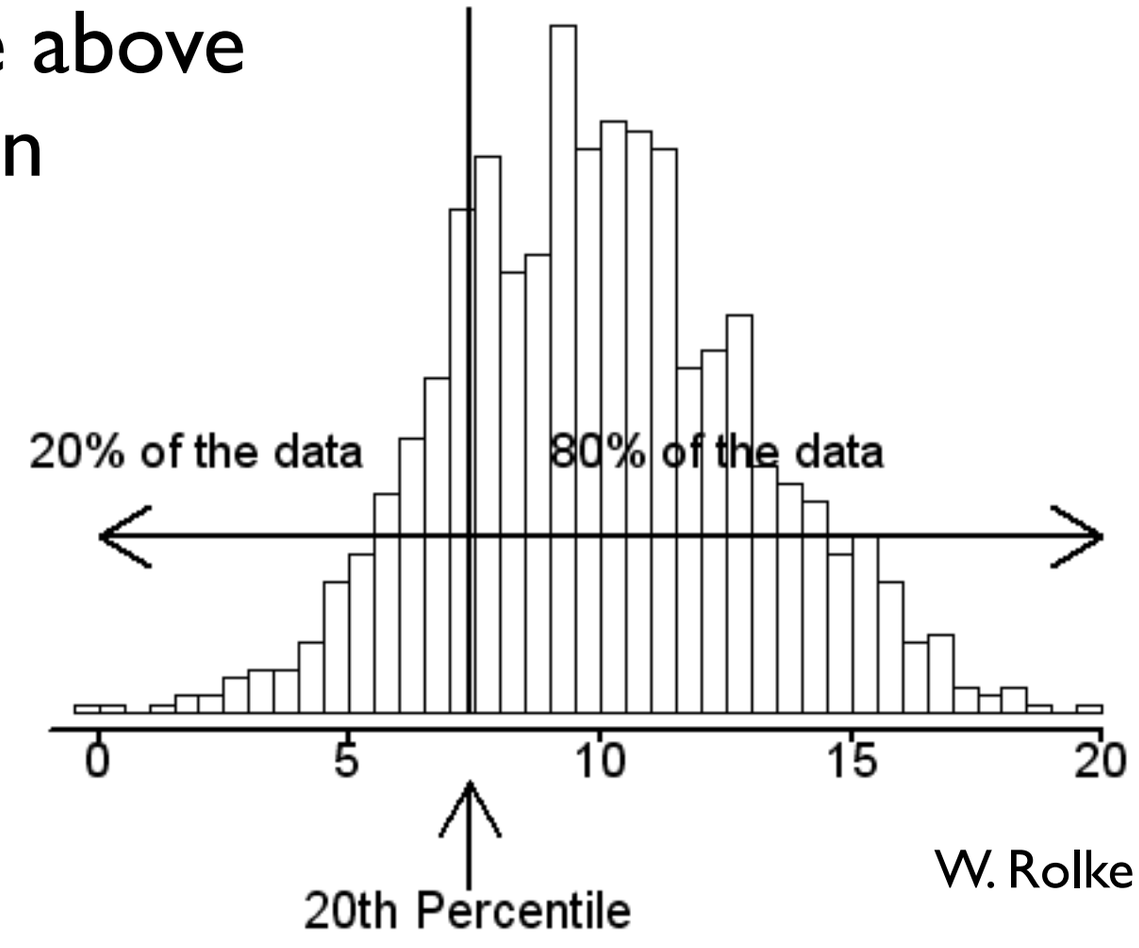
have to take both measurement uncertainties into account:

$$\frac{|x_1 - x_2|}{\sigma_{\mathrm{tot}}} = \frac{|x_1 - x_2|}{\sqrt{\sigma_{x_1}^2 + \sigma_{x_2}^2}}$$

Non-Gaussian distributions

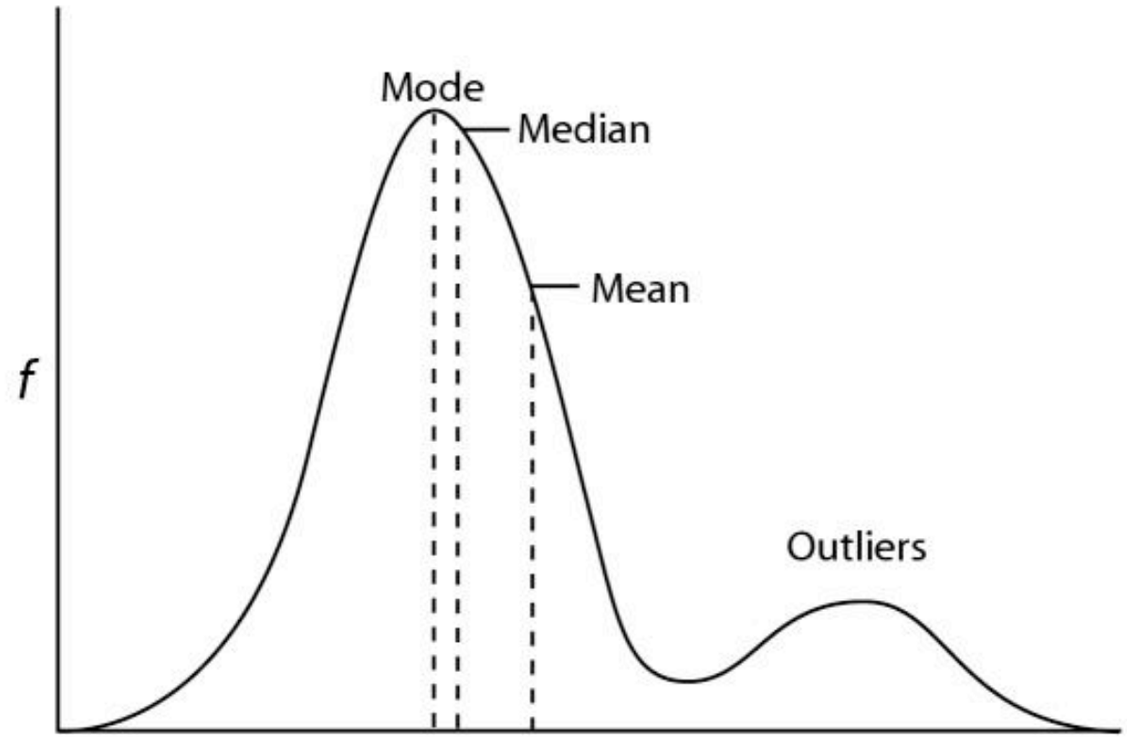
- what if your distribution is non-Gaussian?
- have to decide on case-by-case basis
- percentiles (quartiles): can always sort your data, quote values that are above certain percentage of population
- median: 50th percentile; half the data above, half below
- can quote measurement + uncertainty with percentiles, e.g.:

$$99.123^{+0.005}_{-0.004}$$



Outliers

- for normal distribution, median = mean
- what if distribution is "almost" normal, but has a few outliers? *e.g. cosmic rays in dark frame*
- mean: significantly affected by outliers
- median: robust against (small number of) outliers
- sometimes, it's ok to remove gross outliers ("sigma-clipping"), but need to make sure not to bias your results!



Hedges & Shah 2003

Error propagation

- often, want to determine dependent variable *x* that is a function of one or more measurements

e.g. $x = f(u, v)$ *u* and *v* have (measured variances):

$$\sigma_u^2 = \frac{1}{N-1}\sum_i (u_i - \bar{u})^2 \qquad \sigma_v^2 = \frac{1}{N-1}\sum_i (v_i - \bar{v})^2$$

covariance between *u* and *v*:

$$\sigma_{uv}^2 = \frac{1}{N-1}\sum_i (u_i - \bar{u})(v_i - \bar{v})$$

note: if *u* and *v* are independent, covariance vanishes for large N

Error propagation

- Gaussian case: variance in *x* can be expressed in terms of variance in *u* and *v*, and the covariance between them:

$$\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v}\right)^2 + 2\sigma_{uv}^2 \left(\frac{\partial x}{\partial u}\right)\left(\frac{\partial x}{\partial v}\right)$$

- if u and v independent:

$$\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v}\right)^2$$

e.g. $x = a\,u\,v$ $\qquad \dfrac{\sigma_x^2}{x^2} = \dfrac{\sigma_u^2}{u^2} + \dfrac{\sigma_v^2}{v^2}$

with a = constant:

Note: if *f* is non-linear in *x* or *y*, and uncertainty is large, this approximation breaks down