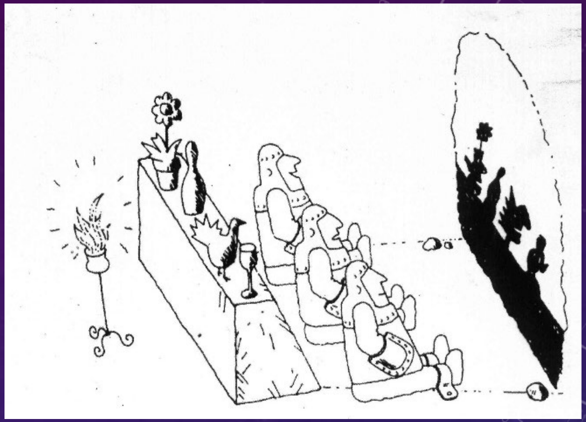
BASIC STATISTICS I

PLATONIC TRUTH & STATISTICS



What are "statistics"?

location – mean, median, mode

spread/scale – rms, scatter, dispersion, sigma, standard deviation, variance (spread^2)

What are these "really"?

data value, error (uncertainty) = estimates of location, spread of underlying "true" distribution from which data point was hypothetically drawn – there are errors on errors!

PLATONIC TRUTH & STATISTICS

What are "statistics"?

location – mean, median, mode

spread/scale – rms, scatter, dispersion, sigma, standard deviation, variance (spread^2)

What are these "really"?

data value, error (uncertainty) = estimates of location, spread of underlying "true" distribution from which data point was hypothetically drawn – there are errors on errors!

"robust" statistics are outlier-resistant:
e.g. biweight for scale
(Beers et al. 1990)
http://adsabs.harvard.edu/abs/1990AJ....100...32B

PROBABILITY DISTRIBUTIONS

- Uniform – used when there is no basis for assuming anything else; may be uniform in a linear or log quantity
- Binomial – used for yes/no analysis of frequency of events/objects (e.g., a galaxy is either an AGN or not an AGN)
- Poisson – used when counting events/objects each of whose existence has a fixed probability (e.g., number of photons incident on a detector, number of galaxies in a volume of space); has special property that variance = mean

  (proof: http://www.proofwiki.org/wiki/Variance_of_Poisson_Distribution)
- Gaussian – used when analyzing values scattered randomly about a mean (the Central Limit Theorem says nearly all other distributions approach a Gaussian in the limit of large numbers)
- Chi-Square – used when comparing data and models; probability distribution describing the relation between data-model residuals and uncertainties as a function of model "degrees of freedom"

PROBABILITY DISTRIBUTION INTEGRALS = 1

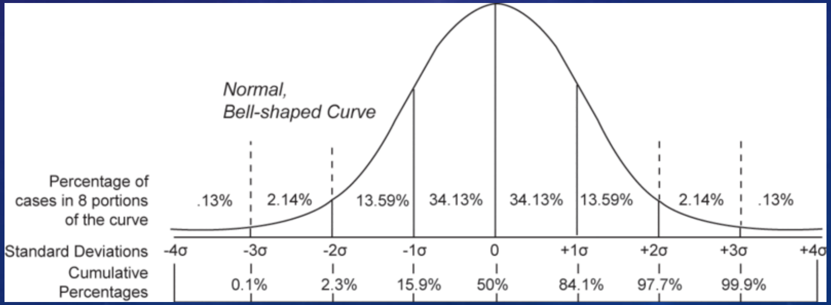
Integral of a Gaussian

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\int_{-\infty}^{\infty} e^{-ax^2} dx \int_{-\infty}^{\infty} e^{-ay^2} dy} = \sqrt{\iint_{0,0}^{2\pi,\infty} e^{-ar^2} r d\varphi dr}$$

$$= \sqrt{2\pi \left({}^{-1}/_{(-2a)} \right)} = \sqrt{\frac{\pi}{a}}$$

Normalize to get total probability = 1:

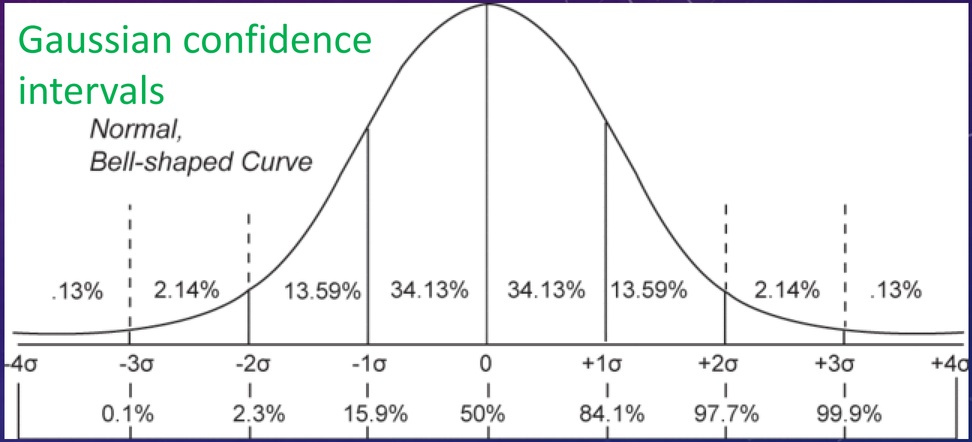
$$P(u) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{u^2}{2\sigma^2}}$$



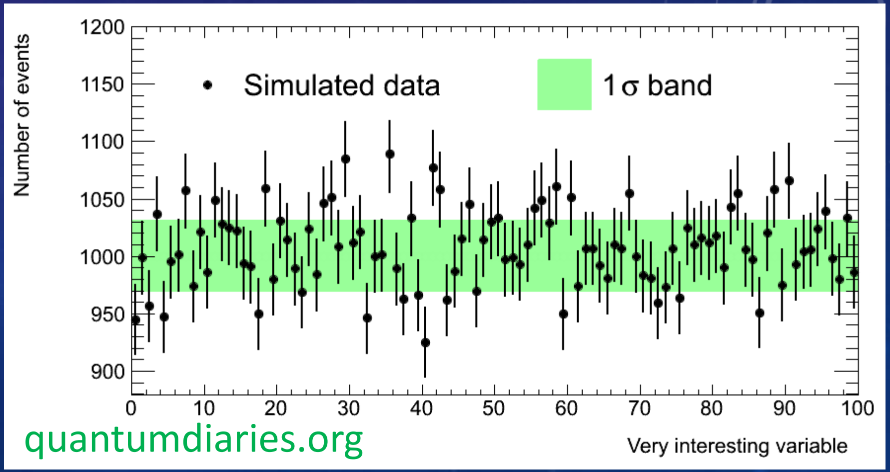
CONFIDENCE INTERVALS

Integrate probability distribution *partially* to get:

- error bars: upper & lower extensions from observed value expected to bound the true value ±X% of the time, e.g. ±1σ errors define 68% "two-sided" confidence interval

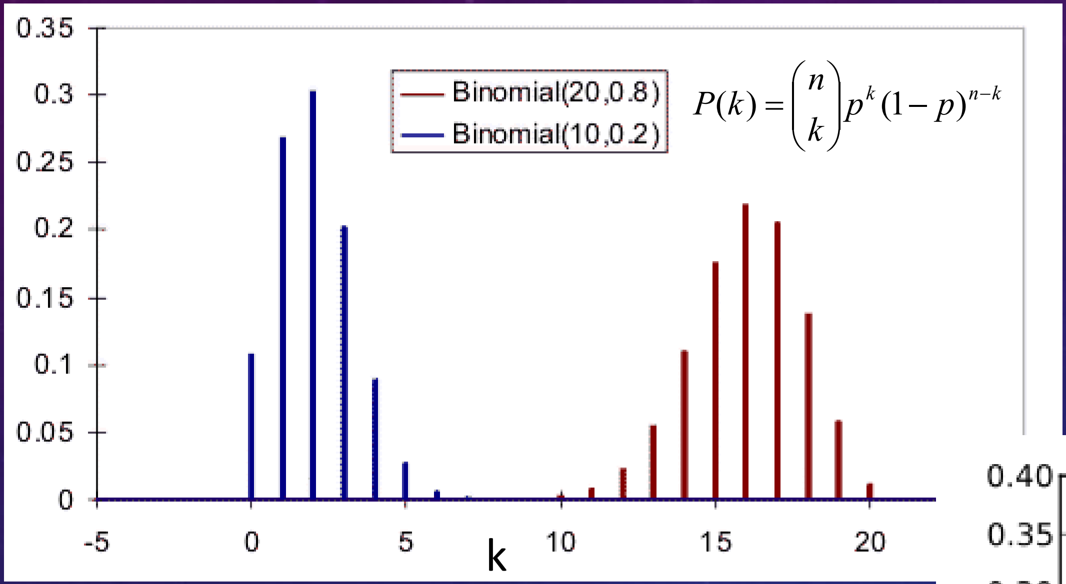


- detection strength: probability that a random value would not exceed the mean by observed value *or more*, e.g., 5σ detection ⇔ "one-sided" % confidence interval integrated from -∞ to 5σ [caveat: % does not really measure confidence – read about p-value crisis!]



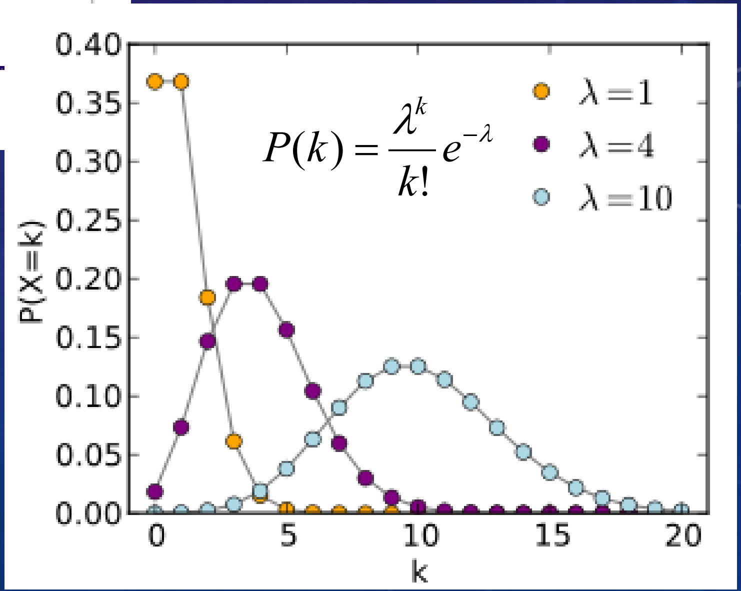
IN GENERAL ERRORS MAY BE ASYMMETRIC (NON-GAUSSIAN DISTRIBUTIONS, LOG(X), ETC.)

www.epixanalytics.com



Binomial:
classification problems;
errors on “yes” inverse
to errors on “no”

Poisson:
counting problems;
mean count λ,
lower bound zero



wikipedia.org

FREQUENTIST & BAYESIAN STATISTICS

Two approaches to statistics:

1) Frequentist – if you roll a fair die 100 times, what fraction of rolls should give a six? → probability distribution of possible outcomes

2) Bayesian – observe a six rolled 15 times out of 100, what is the likelihood that the die is fairly weighted? → likelihood distribution of possible models (fairly weighted, unfairly toward/away from six, etc.)

FREQUENTIST & BAYESIAN STATISTICS

Two approaches to statistics:

1) Frequentist – if you roll a fair die 100 times, what fraction of rolls should give a six? → probability distribution of possible outcomes

2) Bayesian – observe a six rolled 15 times out of 100, what is the likelihood that the die is fairly weighted? → likelihood distribution of possible models (fairly weighted, unfairly toward/away from six, etc.)

Astronomers get *one* universe to constrain many theories…

BASICS OF PROBABILITY

Frequentists define probabilities as relative frequencies of specific outcomes. Bayesians define probabilities as numerical formalizations of our degree of belief that specific outcomes will occur. Both accept the same mathematical axioms governing probabilities.

Kolmogorov axioms of probability

- any random event A has prob(A) between 0 and 1
- the sure event has prob(A)=1
- if A and B are exclusive events, then

  prob(A or B) = prob(A) + prob(B)

and it follows that

- If A and B are independent events, then

  prob(A and B) = prob(A) × prob(B)

CONDITIONAL PROBABILITY & MARGINALIZATION

Conditional probability

- wish to know probability of A, given that we know B
- definition of this is

  prob(A|B) = prob(A and B) / prob(B)
- note this equals p(A) iff A and B are independent

Marginalization

- if want the total probability of A "marginalized" (summed) over all values of a "nuisance parameter":

  prob(A) "marginalized over B" = $\sum_i$ prob(A|B_i) × prob(B_i)
- not necessarily equal to the total probability of A, since we may not have marginalized over other parameters of interest

BAYES' THEOREM

by symmetry prob(A and B) = prob(B and A)

use prob(A|B) = prob(A and B) / prob(B)

and equivalently prob(B|A) = prob(B and A) / prob(A)

algebra yields prob(A|B) × prob(B) = prob(B|A) × prob(A)

or $\boxed{\text{prob}(B|A) = \text{prob}(A|B) \times \text{prob}(B) / \text{prob}(A)}$

- ➔ math holds for *both frequentist & Bayesian statistics*
- ➔ defines Bayesian approach when A="data" and B="model"
- ➔ LHS is "posterior probability" of the model
- ➔ prob(A|B) is the "likelihood" of the data given the model, also used by frequentists
- ➔ prob(B) is "prior" probability of the model, only used by Bayesians