PHY 517 / AST 443:
Observational Techniques in Astronomy

Lecture 4:
Statistics

Reminder

- send in your exoplanet transit requests!
- 3 transits per group

- remember to do your weekly check-ins for Lab 0 date with me or the TAs
- my office hours: 2-3pm Tu+We

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 0.05 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

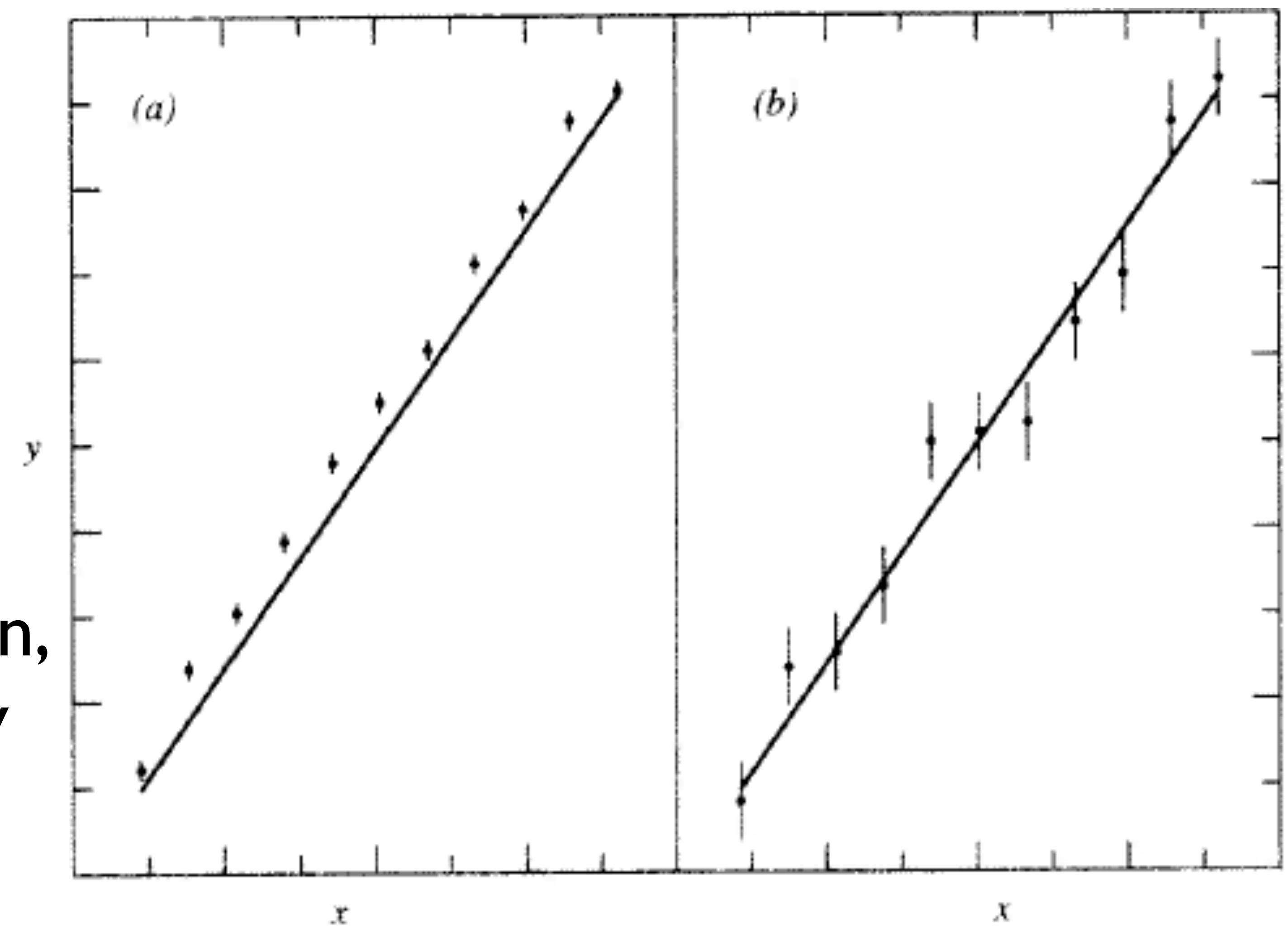
“Error”

- 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

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.

Sample vs. Parent Distribution

- measurement x_i of a quantity x:
  - approximates x
  - not necessarily equal to x because of statistical errors
- 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

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)

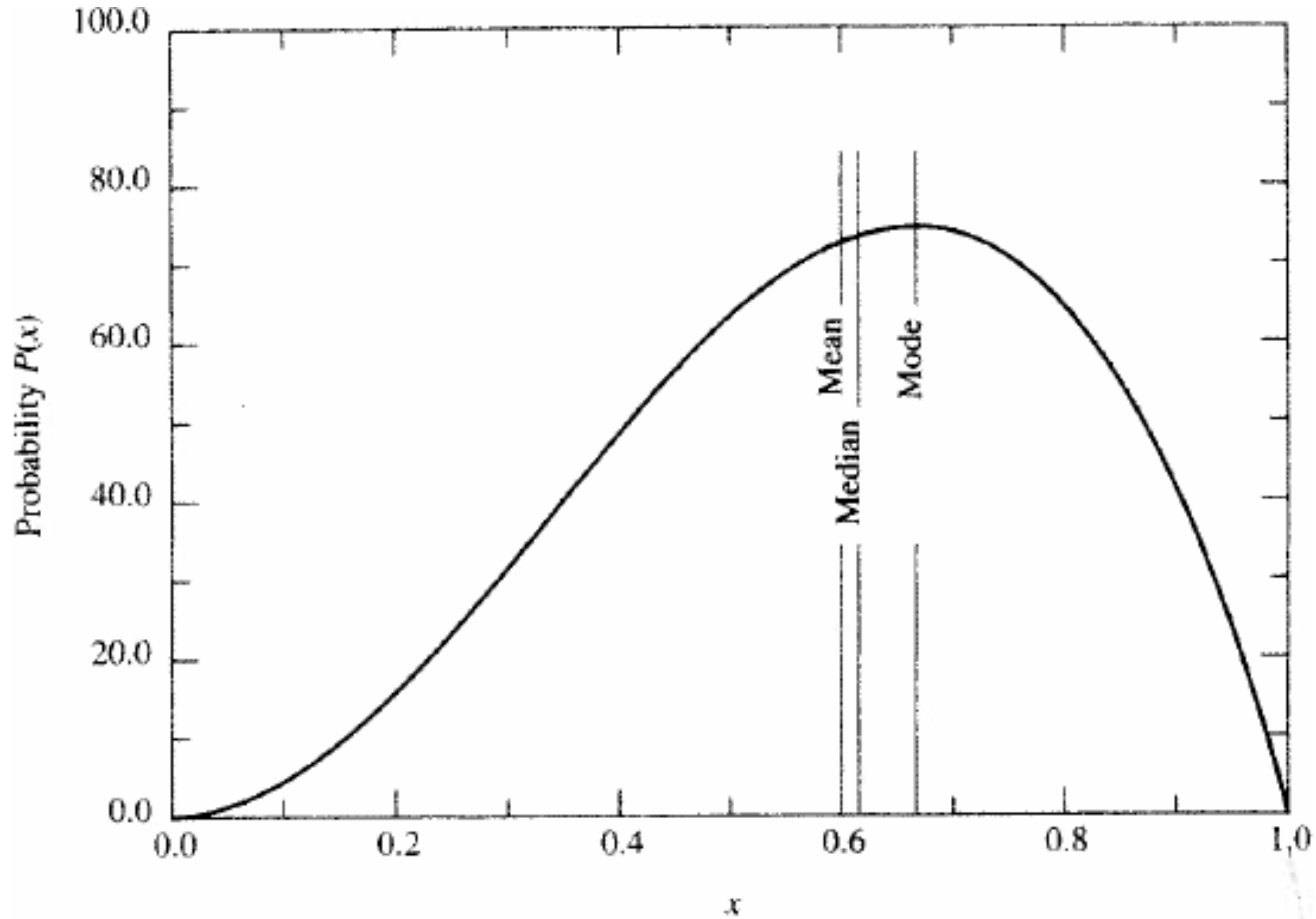
  motivation: less susceptible to "outliers" than the mean

- mode: the most common measurement value

  motivation: the most likely value

Mean, median, and mode

mean, median and mode for an example distribution:



- 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 deviate from the mean

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:

$$\bar{x} = \frac{\sum_i w_i x_i}{\sum_i w_i}$$

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

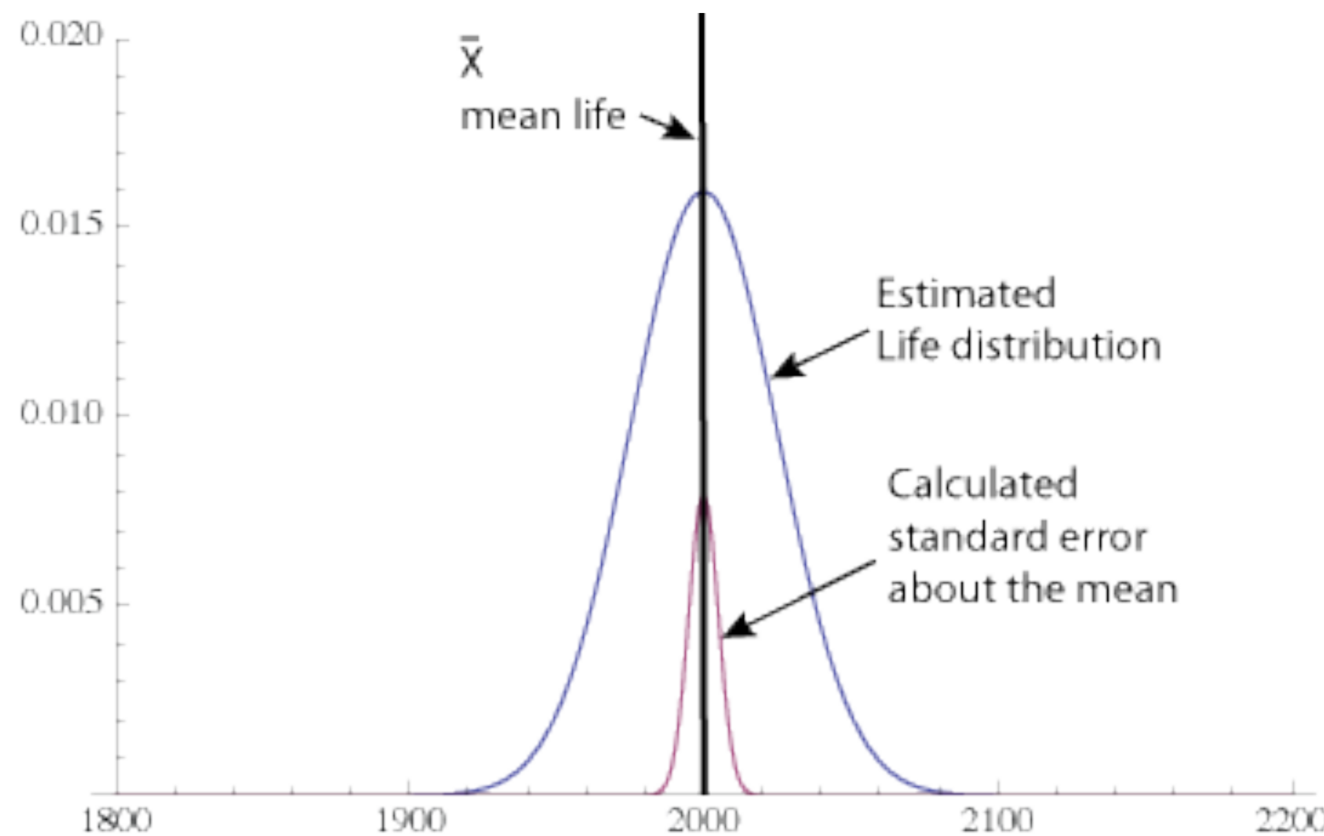
$$\bar{x} = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}$$

Uncertainty on the mean

- variance and std. deviation are measures of the *width* of a distribution
- the location of the mean can be determined more precisely:
- measurement uncertainty on the mean:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{N}$$

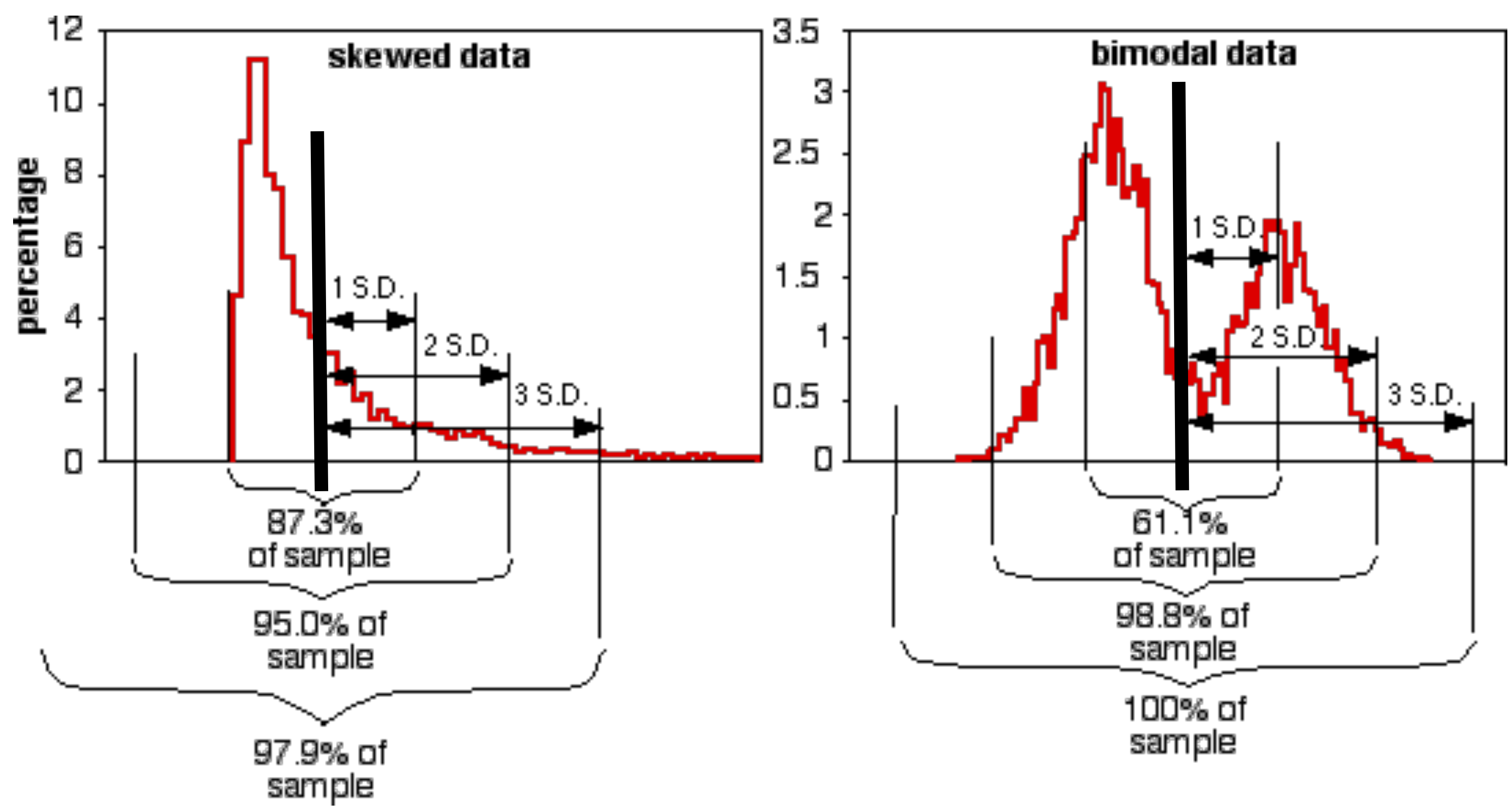
$$\sigma_{\bar{x}}^2 = \frac{1}{\sum_i 1/\sigma_i^2}$$



But what does it all mean?

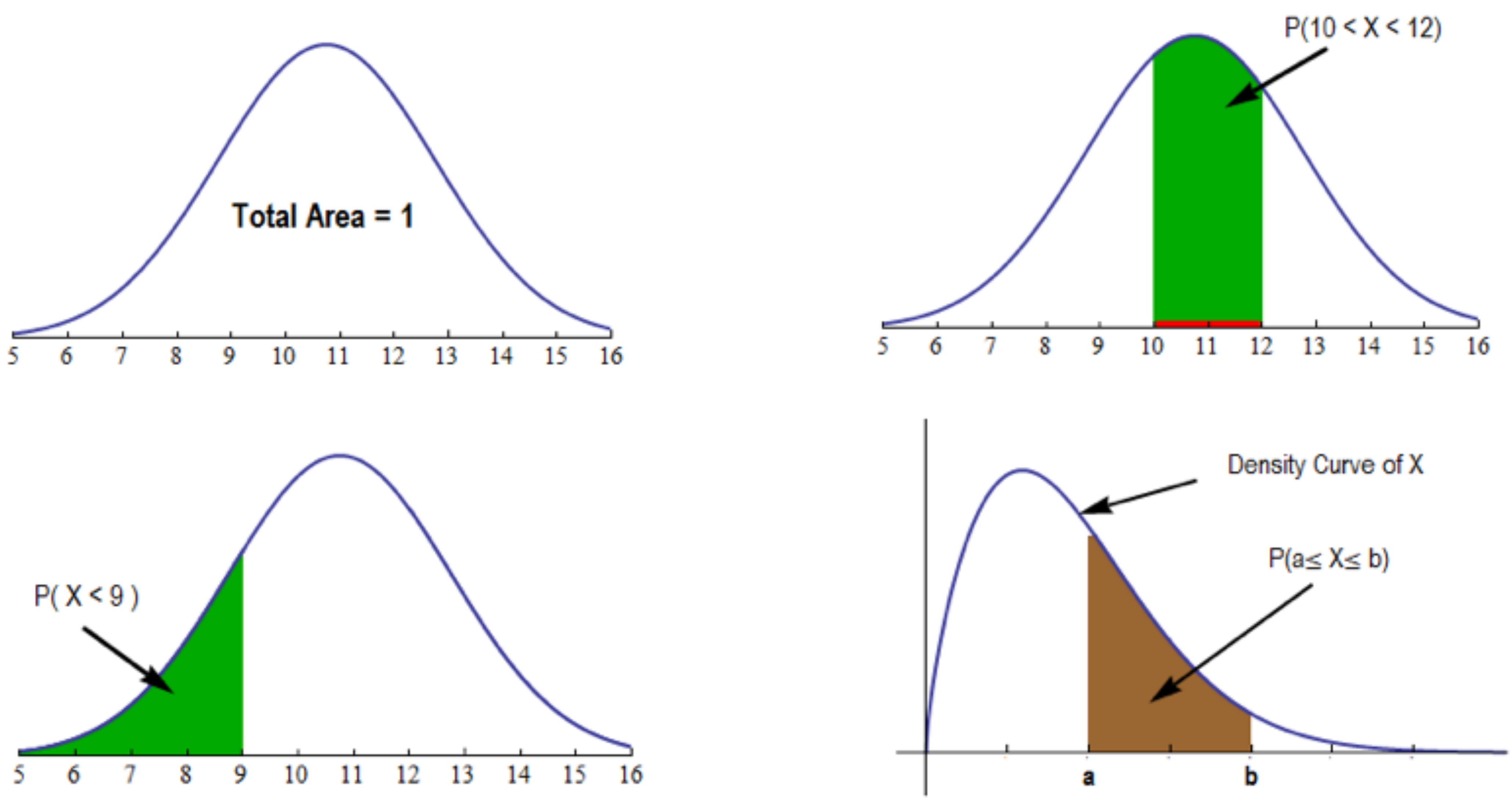
- can calculate mean, variance, etc. for any set of data points

- *that does not guarantee that they are useful descriptions of the distribution !*



Probability Distributions

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



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; $0 \le x \le n$
- $0 < p < 1$

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!)

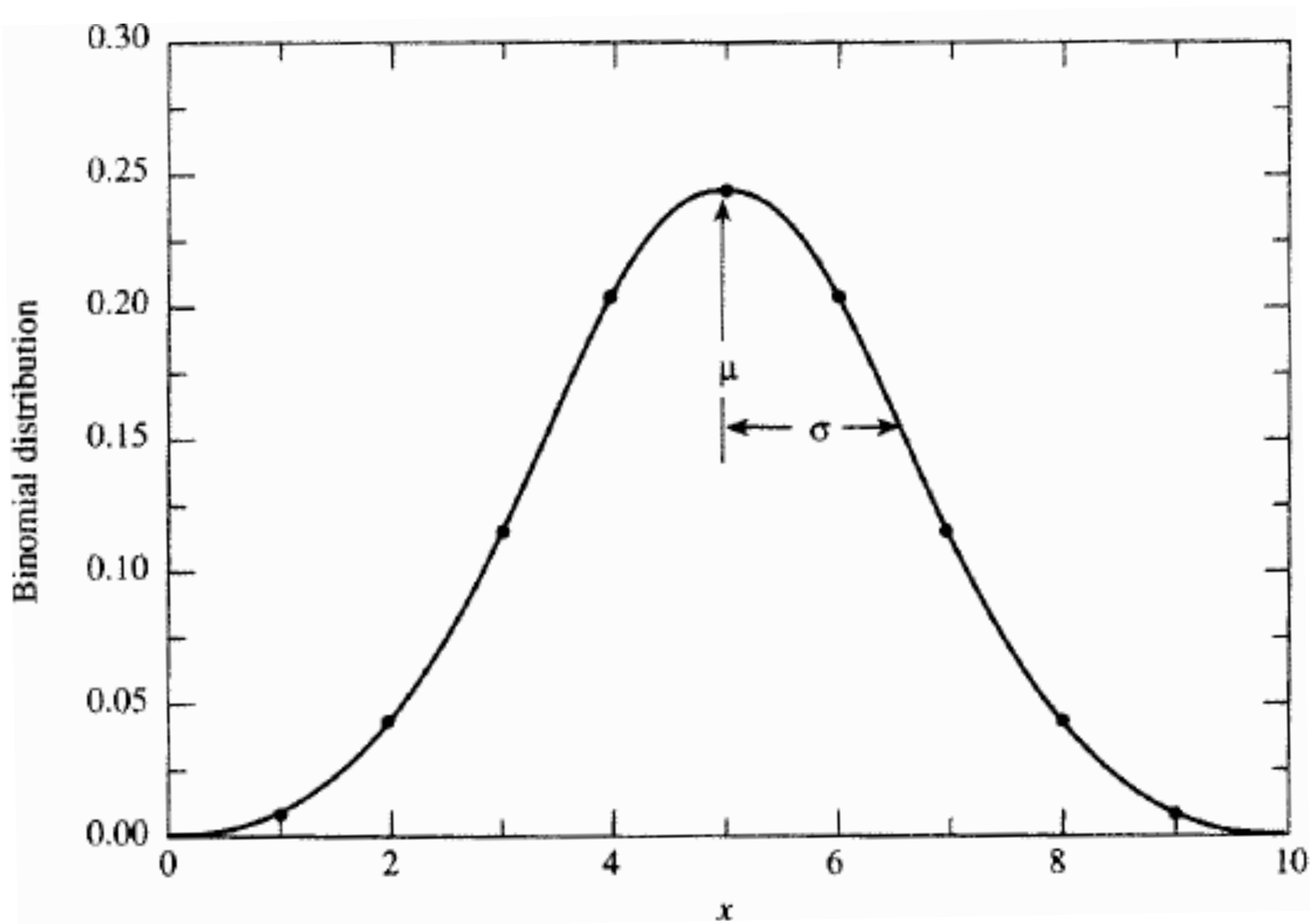
- variance of the binomial distribution:

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

Binomial Distribution - Example (1)

tossing 10 coins, x = number of tails

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



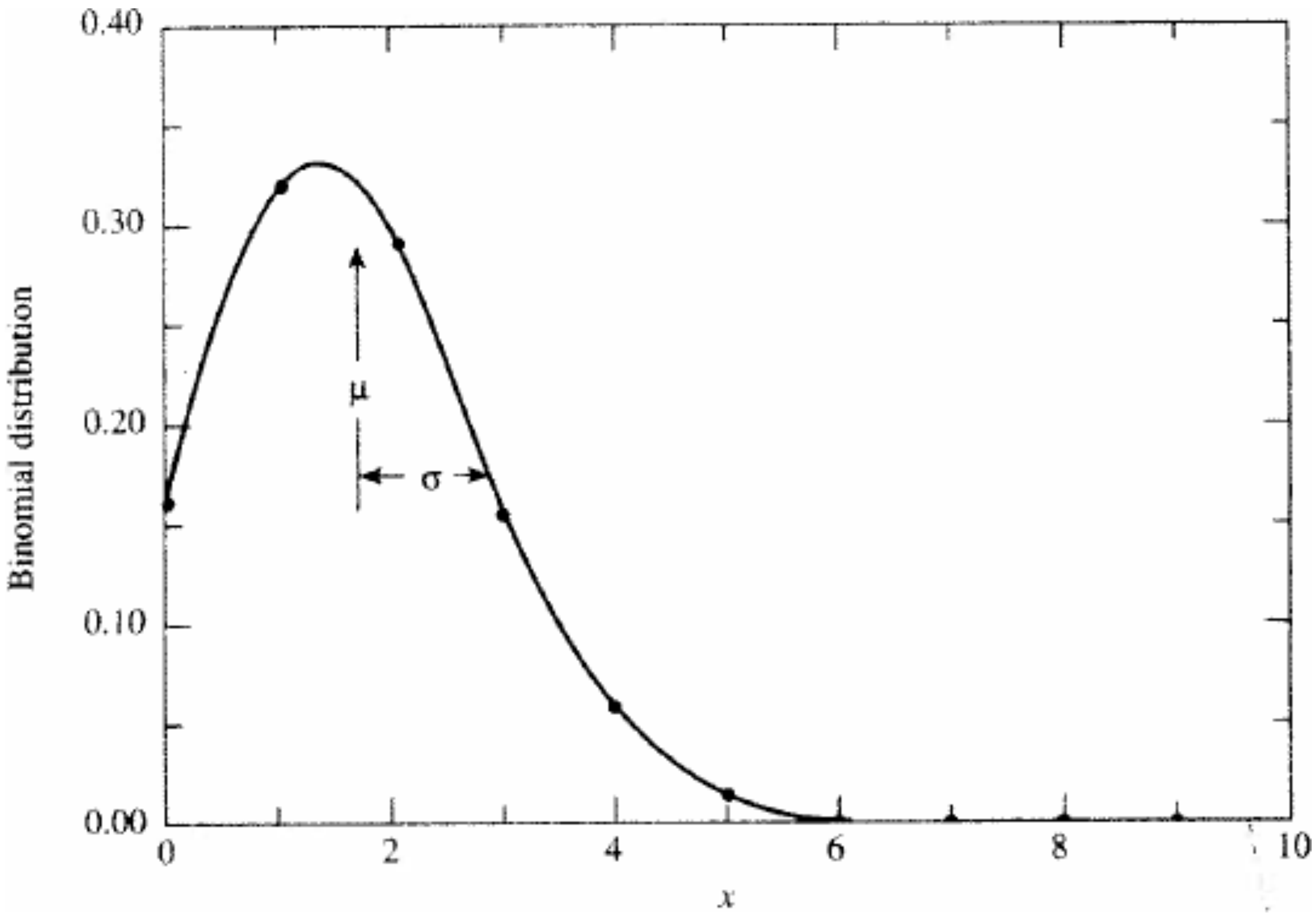
$$\mu = np = 5$$
$$\sigma^2 = np(1-p)$$
$$= 2.5$$

since q=p: distribution is symmetric

Binomial Distribution - Example (2)

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

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



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

$$\sigma^2 = np(1-p) = 1.39$$

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

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

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

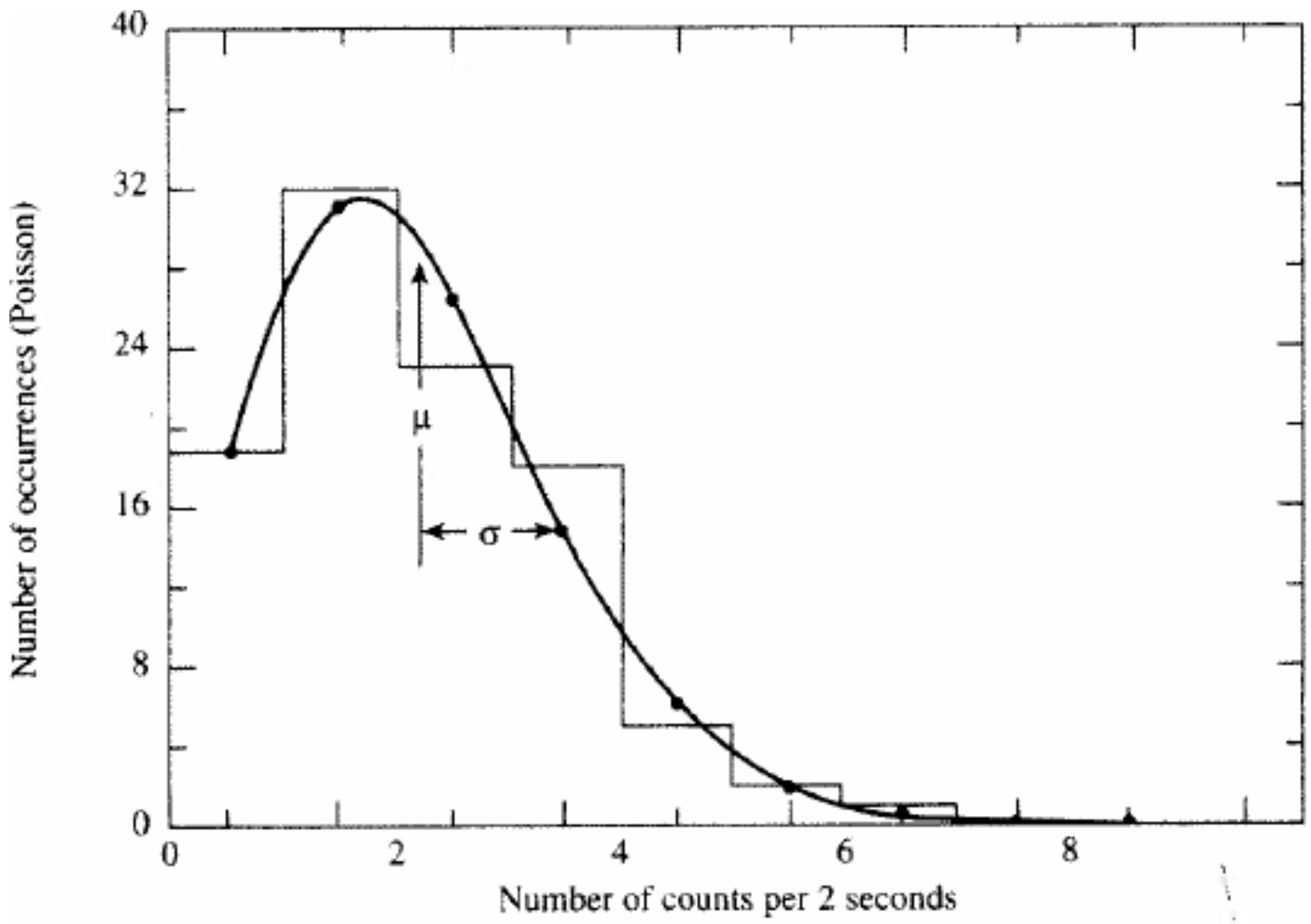
Poisson Distribution

- 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:

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

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

Gaussian / Normal Distribution

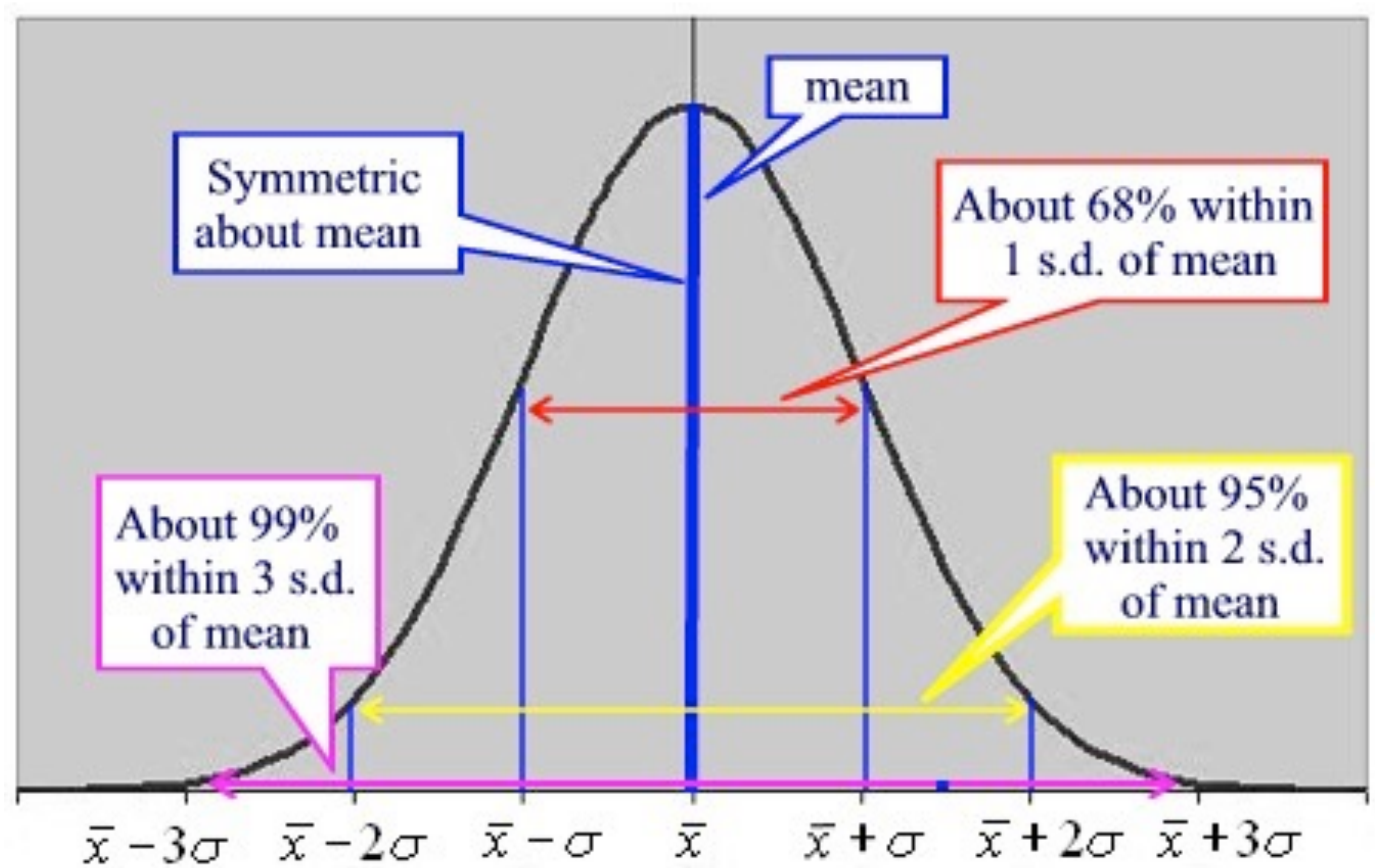
- the most commonly used probability distribution

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

- 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



Central Limit Theorem

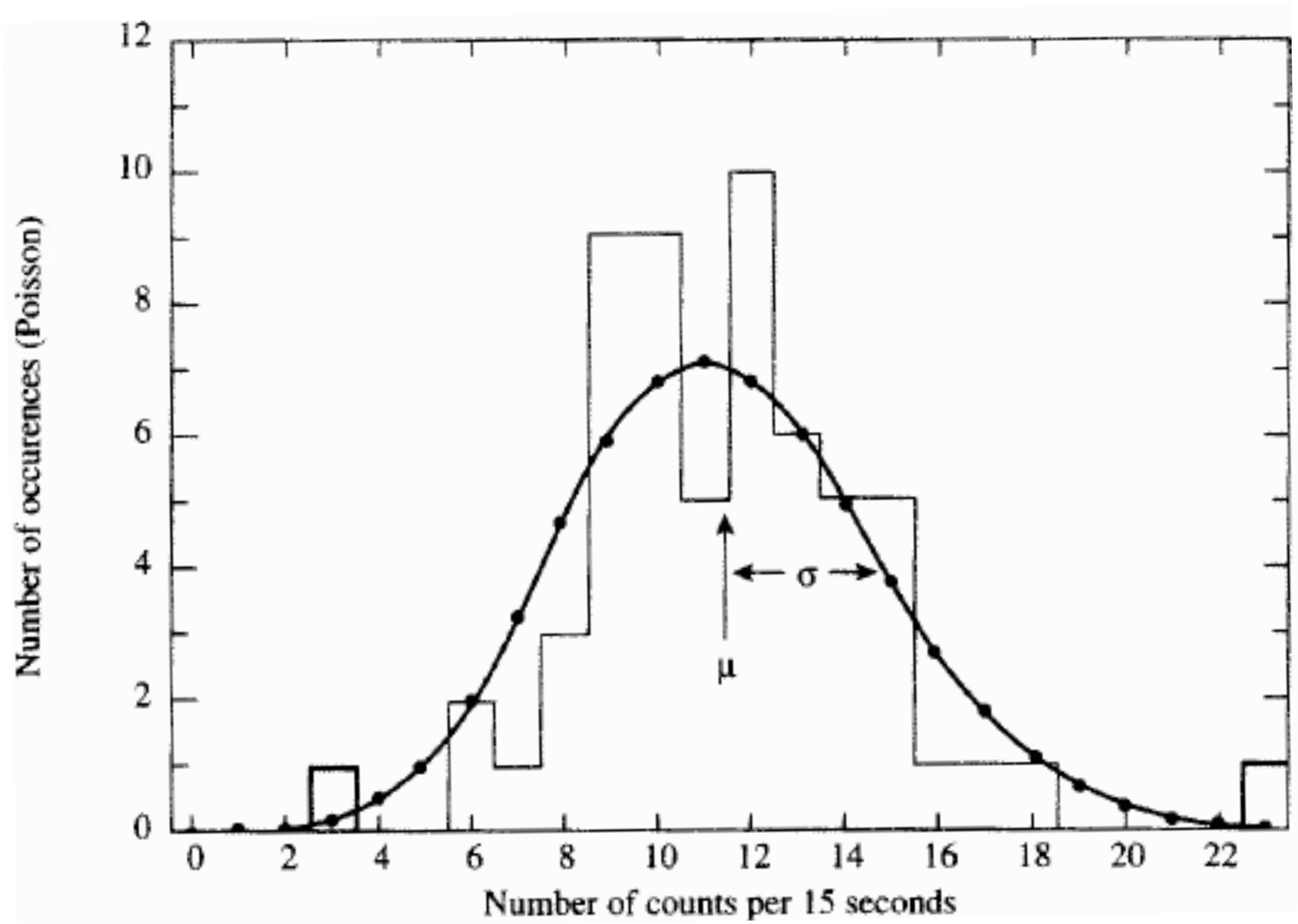
- "the sum of n random values drawn from a probability distribution function of finite variance, σ^2, tends to be Gaussian distributed about the expectation value for the sum, with variance $n\sigma^2$"

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

- many processes in nature described by a normal (or log-normal) distribution

Gauss Distribution - Example

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



measured mean:

$$\bar{x} = 11.48$$

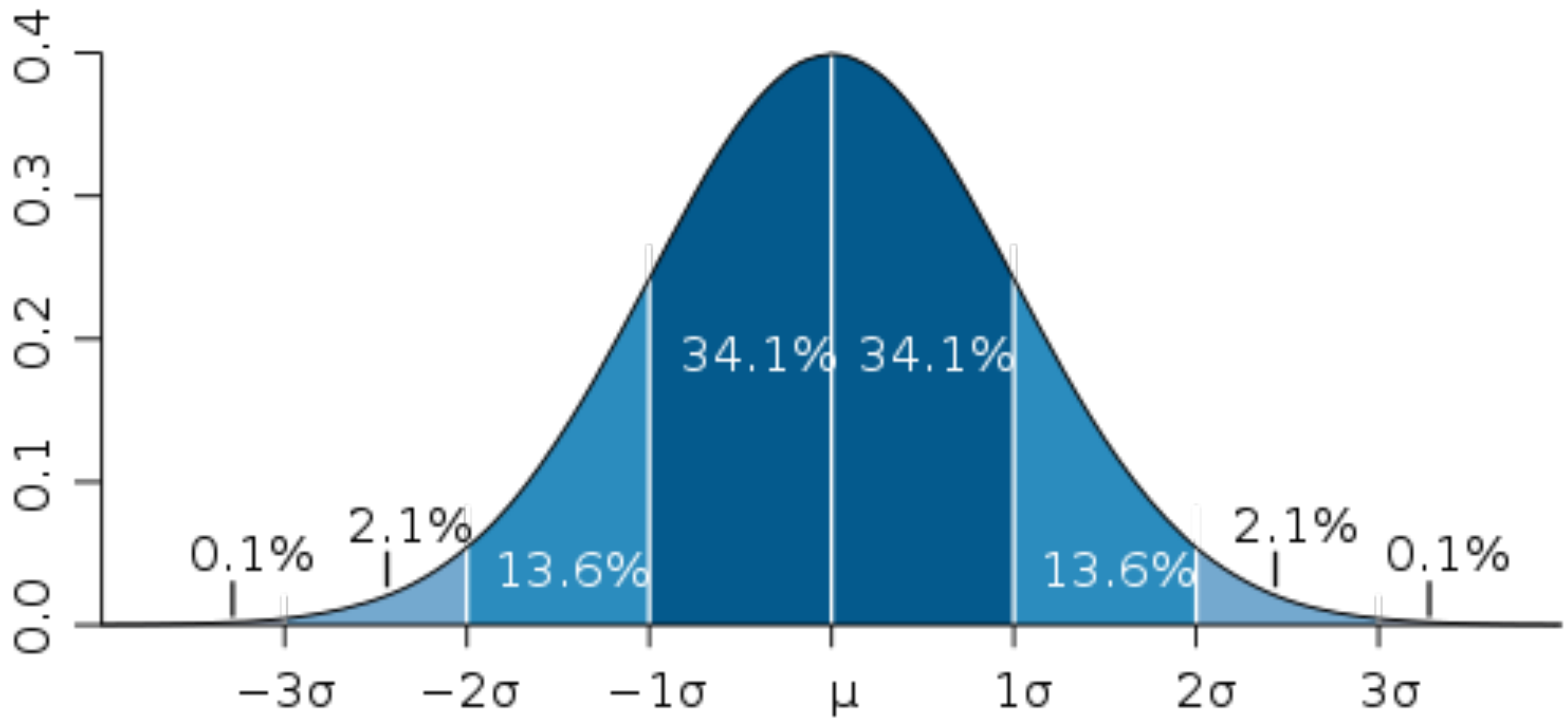
solid curve:

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

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

Gaussian / Normal Distribution

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



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

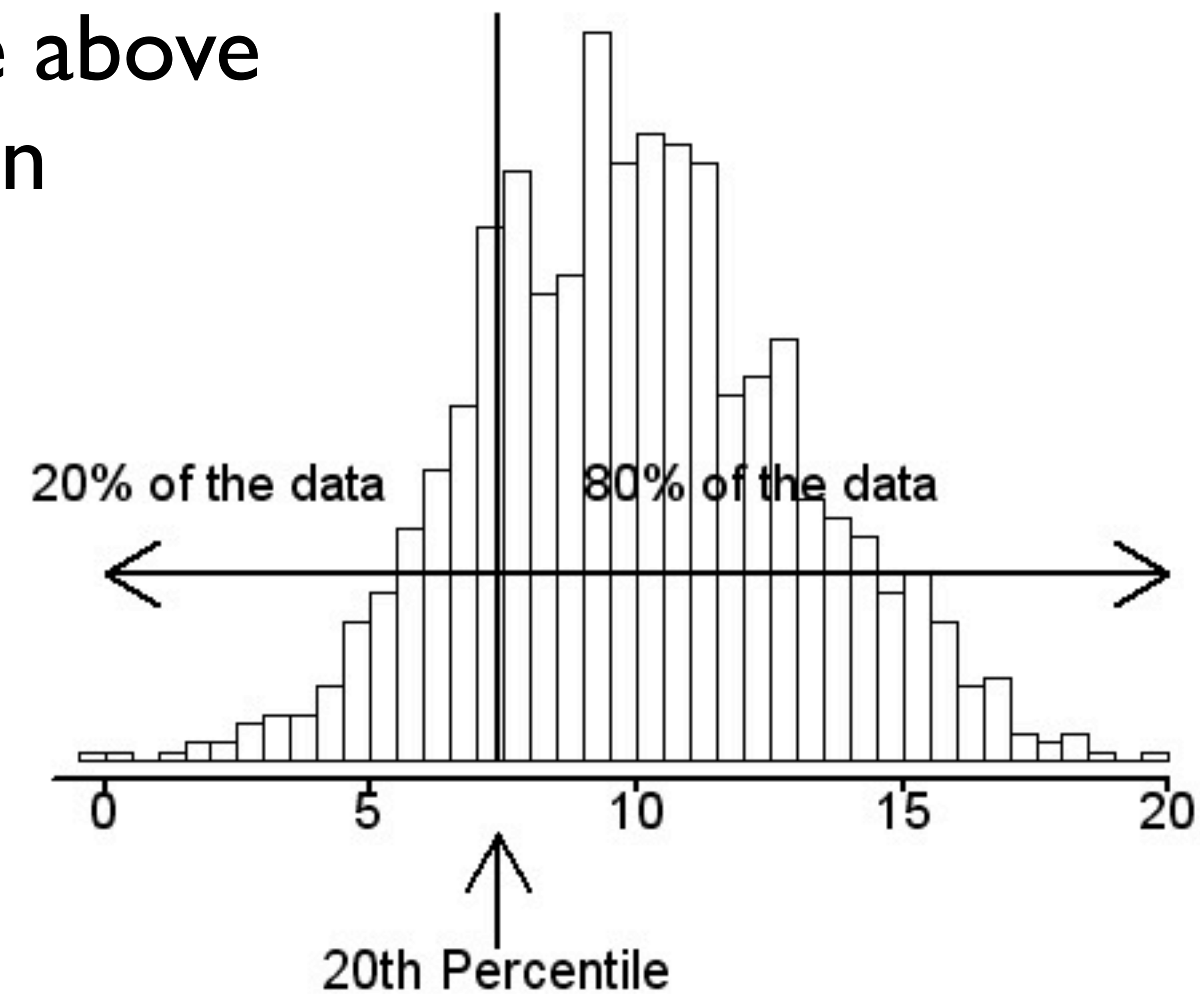
- only a valid measure of probability if the background distribution is Gaussian!

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

Non-Gaussian distributions

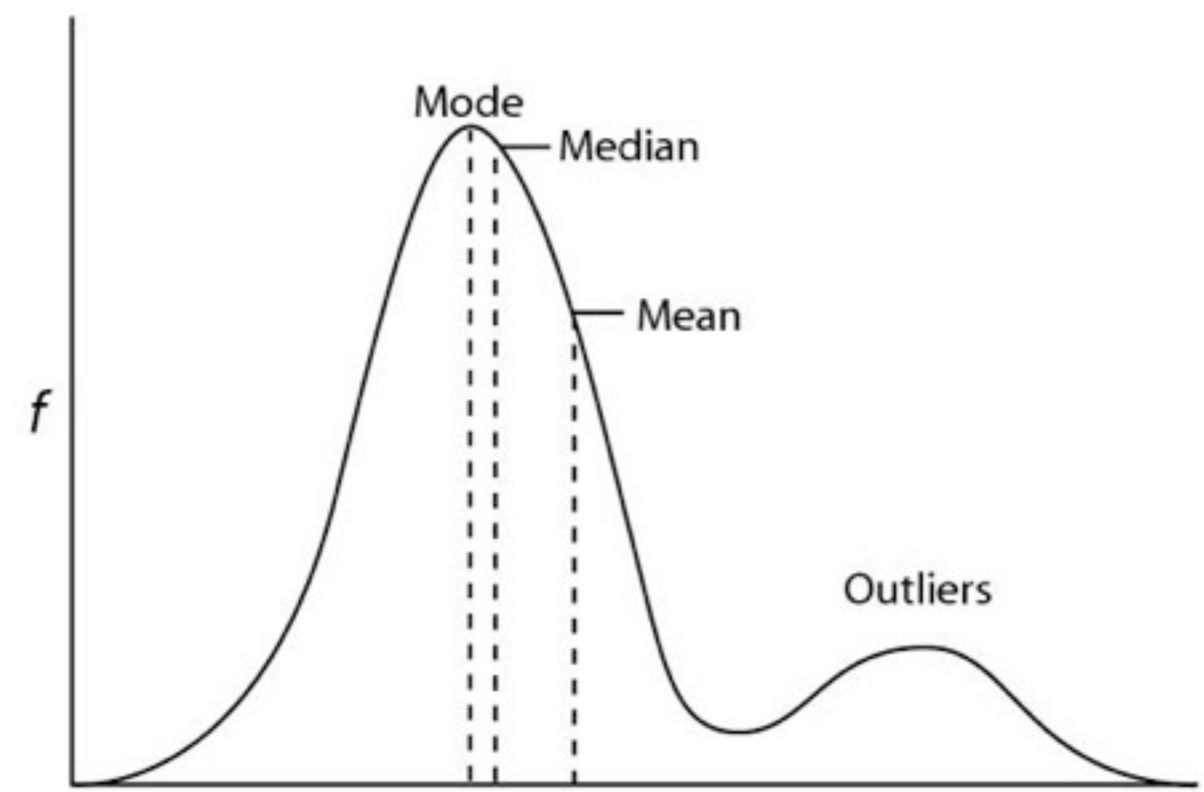
- what if your distribution is non-Gaussian?
- have to decide on case-by-case basis
- percentiles (quartiles): can always sort your data, quote values that are above certain percentage of population
- median: 50th percentile; half the data above, half below
- can quote measurement + uncertainty with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



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. $x = a\,u\,v$

with a = constant: $\frac{\sigma_x^2}{x^2} = \frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2}$

Resampling

- what about errors on dependent quantities in non-Gaussian case?

- can use resampling methods (bootstrap, jackknife, ...)

- e.g. bootstrap: resampling with replacement

- N measurements
- draw from your measurements N times (can draw same measurement more than once)
- determine derived quantity
- repeat n times
- quantify the distribution up from n iterations (e.g. percentiles)

Model fitting

- to fit a model to a dataset, need to quantify how good the fit describes the data

- if errors are Gaussian, optimal statistic is χ^2 ("chi-squared")

$$\chi^2 = \sum_i \frac{(D[x_i] - M[x_i])^2}{\sigma_i^2}$$

$D[x_i]$ are the data values; $M[x_i]$ are the values of the model evaluated at positions x_i

(note similarity to normal probability distribution!)

Model fitting

- the best-fitting model is the one that minimizes the χ^2 value

$$\chi^2_{\min} = \sum_i \frac{(D[x_i] - M_{\mathrm{best}}[x_i])^2}{\sigma_i^2}$$

how to find the best-fit model:

- brute force: make a grid of parameter values, calculate χ^2 for each
- use a minimization algorithm

Model fitting

- you have found the "best-fit" parameters of the model that minimize the χ^2, but is that model actually a good fit?

$$\chi^2_\nu = \frac{\chi^2_{\rm min}}{\nu}$$

- reduced chi-square: scale best-fit chi-square by ν, the number of free parameters = number of data points minus the number of free model parameters

- example: fitting a line: two model parameters (slope and intercept)

$$\nu = \text{number of data points} - 2$$

Model fitting

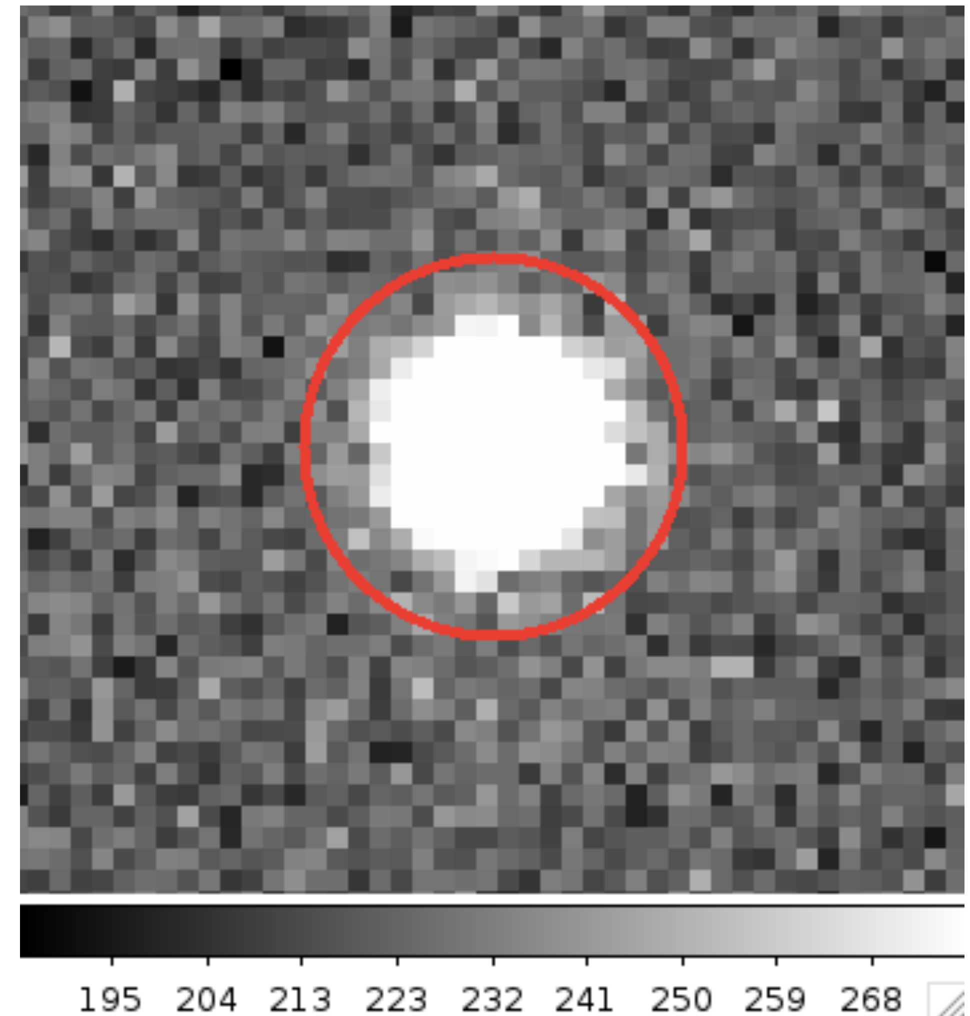
- given a random realization of an experiment with ν degrees of freedom, the probability to obtain $\chi^2_{\min}/\nu$ or larger is described by the chi-squared distribution
- comparing the measured reduced chi-squared to the expectation (from the chi-squared distribution) is an indication whether the model is an acceptable fit to the data
- for an acceptable model, the remaining deviations should be well described by a random (Gaussian) process

$\chi^2_{\min}/\nu \approx 1$	model is a good fit
$\chi^2_{\min}/\nu \gg 1$	model is a bad fit
$\chi^2_{\min}/\nu \ll 1$	model is a overfitting the data

Signal-to-Noise

flux measured in an aperture:

total electrons = electrons from object + electrons from background (sky, dark current, etc.)



signal = total electrons - background electrons

noise: counting statistics (i.e. Poisson distribution, $\sqrt{N}$), have to consider noise from all sources

Signal-to-Noise

signal: $$N_{\rm object} = N_{\rm total} - N_{\rm background}$$

noise: if the noise contributions are independent of each other, can add quadratically:

$$\sigma = \sqrt{\sum_{i \,\in\, \rm noise\, terms} \sigma_i^2}$$

Note: the "counting" processes apply to the number of registered electrons. The counts reported in the image have been rescaled by the gain, $N_{\rm counts} = N_{\rm electrons}/G$

Signal-to-Noise

noise contributions:

- shot noise from source

$$\sigma_{\rm object} = \sqrt{N_{\rm object}}$$
$$= \sqrt{S_{\rm object} \times t}$$

- sky noise

$$\sigma_{\rm sky} = \sqrt{N_{\rm sky}}$$
$$= \sqrt{s_{\rm sky} \times n_{\rm pix} \times t}$$

- dark current noise

$$\sigma_{\rm dk} = \sqrt{N_{\rm dk}}$$
$$= \sqrt{s_{\rm dk} \times n_{\rm pix} \times t}$$

- read-out noise

$$\sigma_{\rm ro} = {\rm RON} \times \sqrt{n_{pix}}$$

Signal-to-Noise

total signal-to-noise: can add noise components quadratically

$$
\begin{aligned}
SNR &= \frac{N_{\rm object}}{\sqrt{\sum_{\rm noise} \sigma_i^2}} \\
&= \frac{N_{\rm object}}{\sqrt{N_{\rm object} + N_{\rm sky} + N_{\rm dk} + N_{\rm ro}}} \\
&= \frac{s_{\rm object} \times t}{\sqrt{s_{\rm object} \times t + n_{\rm pix} \times s_{\rm sky} \times t + n_{\rm pix} \times s_{\rm dk} \times t + n_{\rm pix} \times {\rm RON}^2}}
\end{aligned}
$$

"CCD signal-to-noise equation"

Signal-to-Noise

in general, you do not want to be limited by dark current and read-out noise!

limiting case1: very bright object $N_{\rm object} >> N_{\rm other}$

$$SNR = \frac{N_{\rm object}}{\sqrt{N_{\rm object}}} = \frac{s_{\rm object} \times t}{\sqrt{s_{\rm object} \times t}}$$

$$\propto \sqrt{t}$$

Signal-to-Noise

in general, you do not want to be limited by dark current and read-out noise!

limiting case 2: faint objects $N_{\text{sky}} >> N_{\text{other}}$

$$SNR = \frac{N_{\text{object}}}{\sqrt{N_{\text{sky}}}} = \frac{s_{\text{object}} \times t}{\sqrt{s_{\text{sky}} \times n_{\text{pix}} \times t}}$$

$$\propto \sqrt{t}$$

Sky Background

twilight:

Sun at -6° : “civil twilight”, still bright

Sun at -12° : “nautical twilight”, can see bright stars

Sun at -18° : “astronomical twilight”

twilight is scattered light (blue)

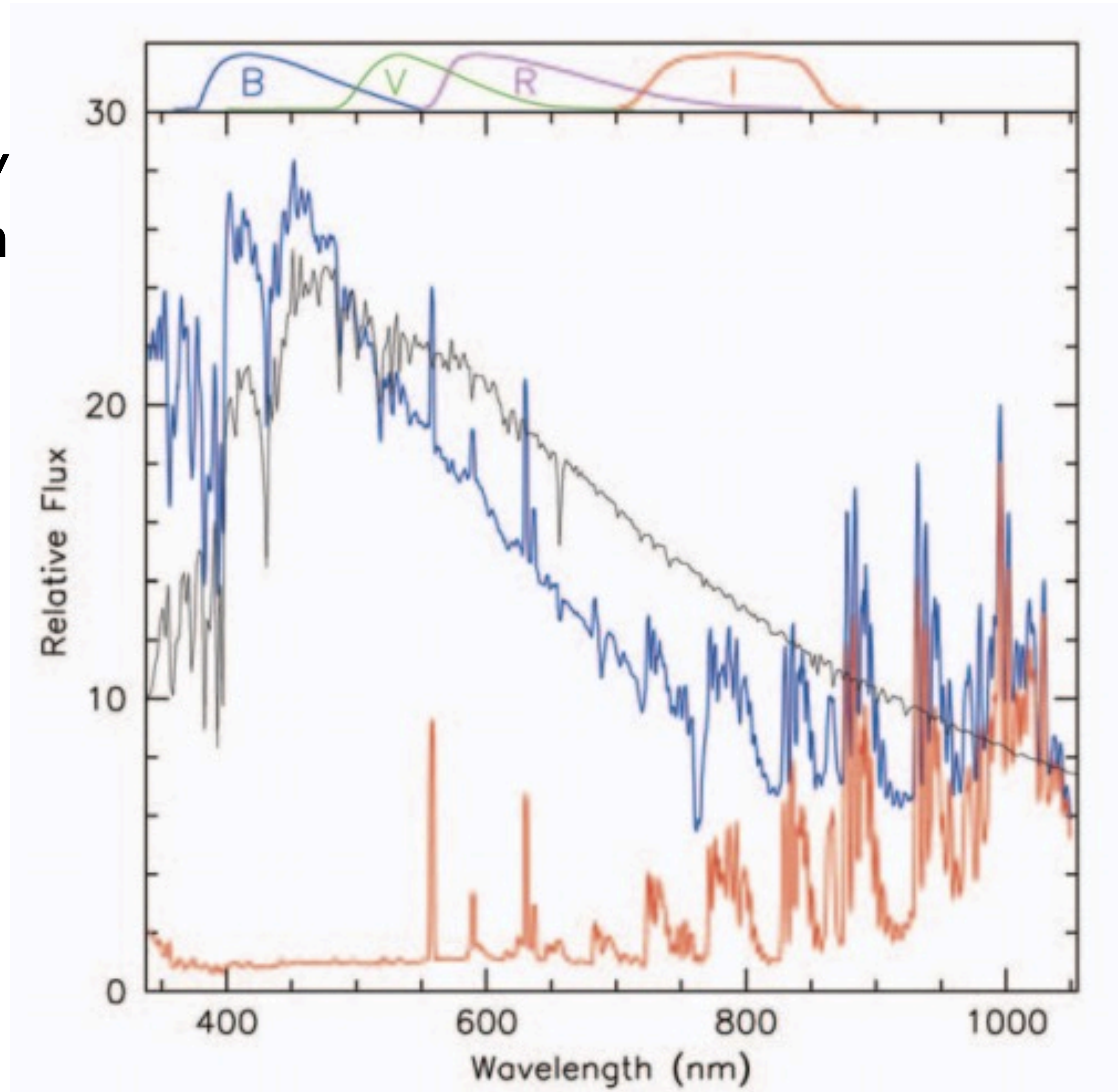
observations in different filters are affected differently

sky is “dark” in red filters before -18°

Sky Background

moonlight:

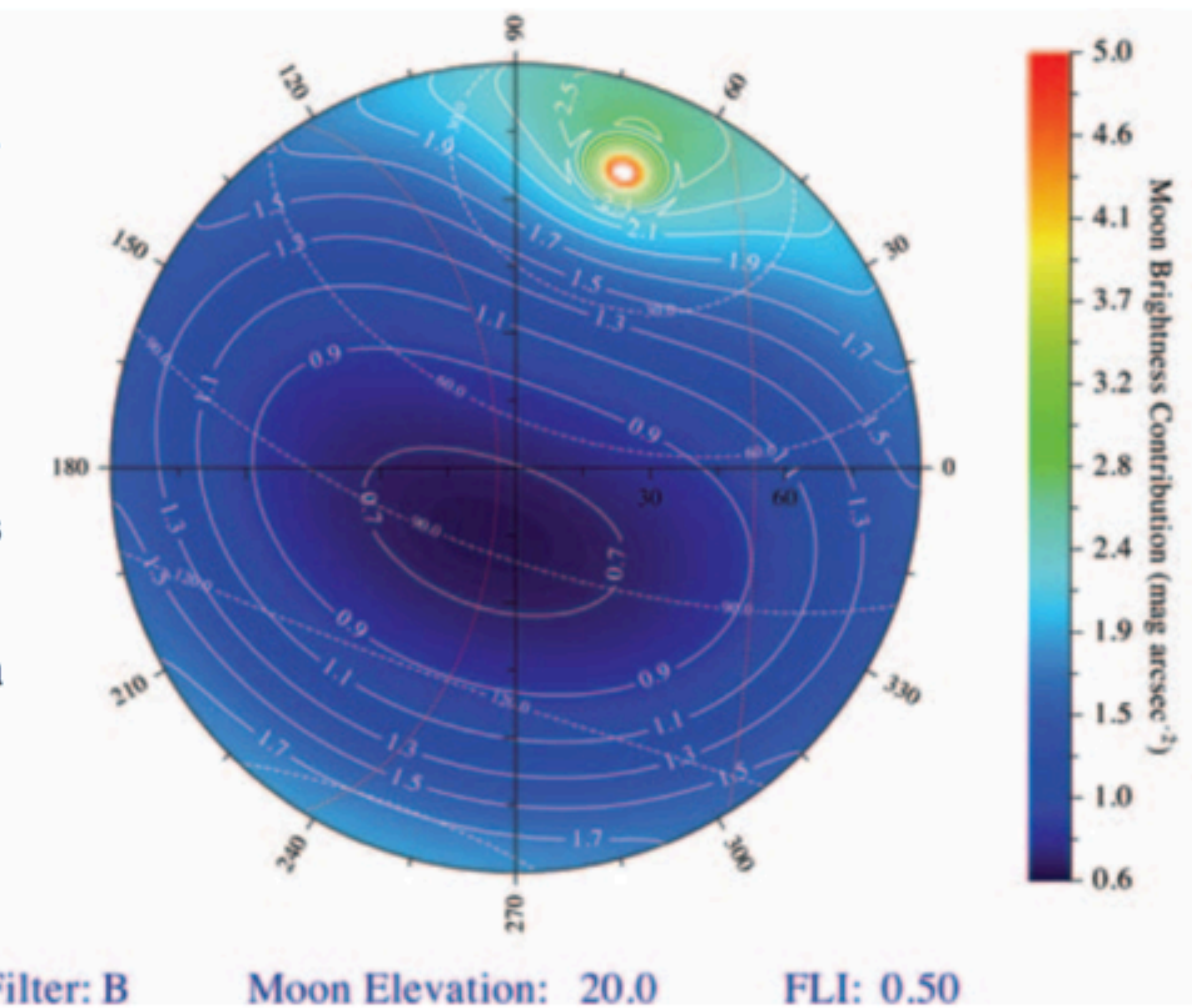
detrimental in the very blue; not a big problem in the infrared



Sky Background

sky brightness from moonlight depends on distance from it, and from horizon

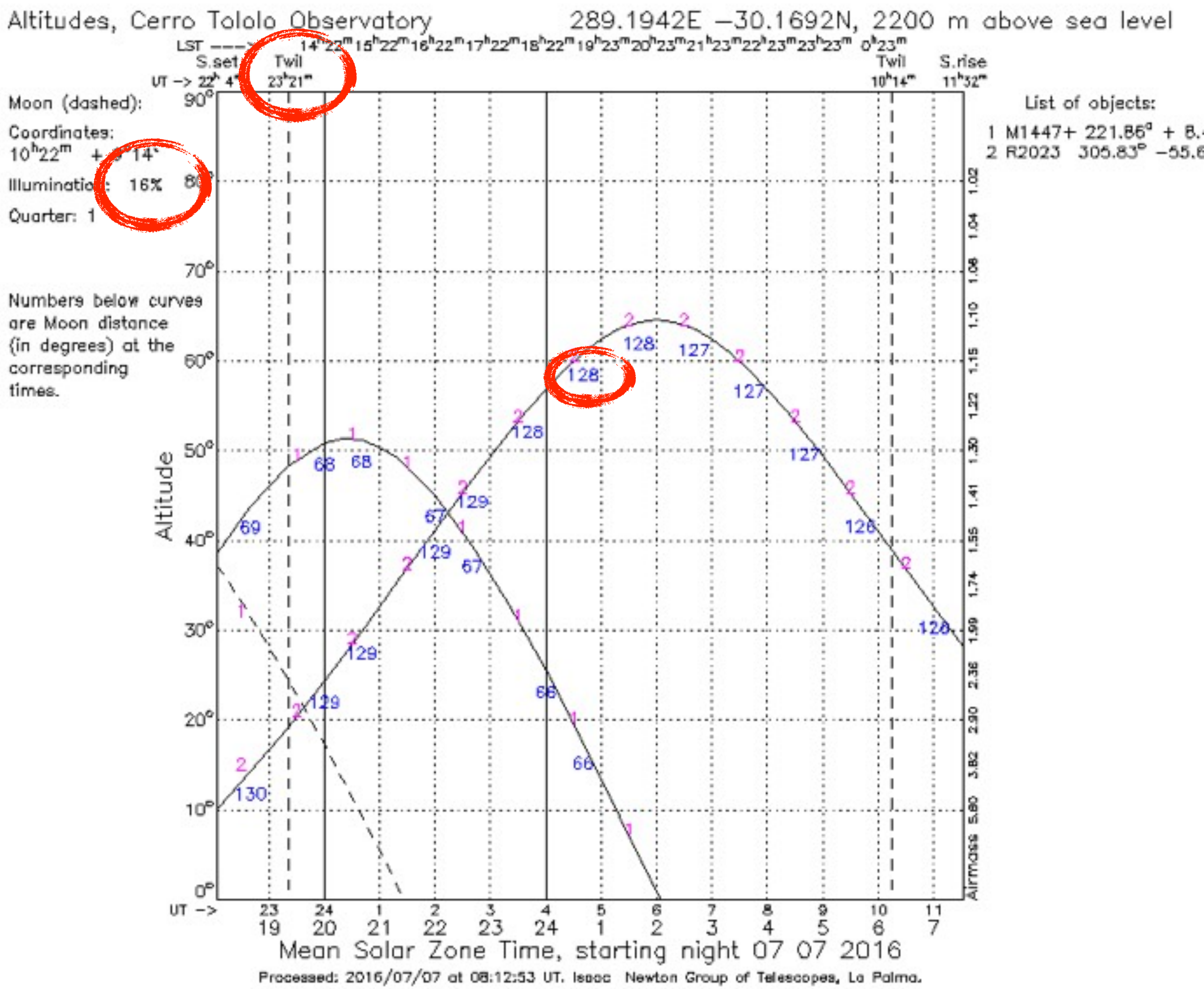
Figure 5: Example isophotal *alt-az* map for the expected moonlight contribution. The dashed white lines trace the loci at constant angular distance from the moon, while the two dotted red lines indicate the extreme apparent lunar paths during a full Saros cycle.



in addition: moonlight can cause reflections inside telescope, from dome, etc.

StarAlt

- indicates times of astronomical twilight
- indicates lunar illumination
- indicates distance to the Moon

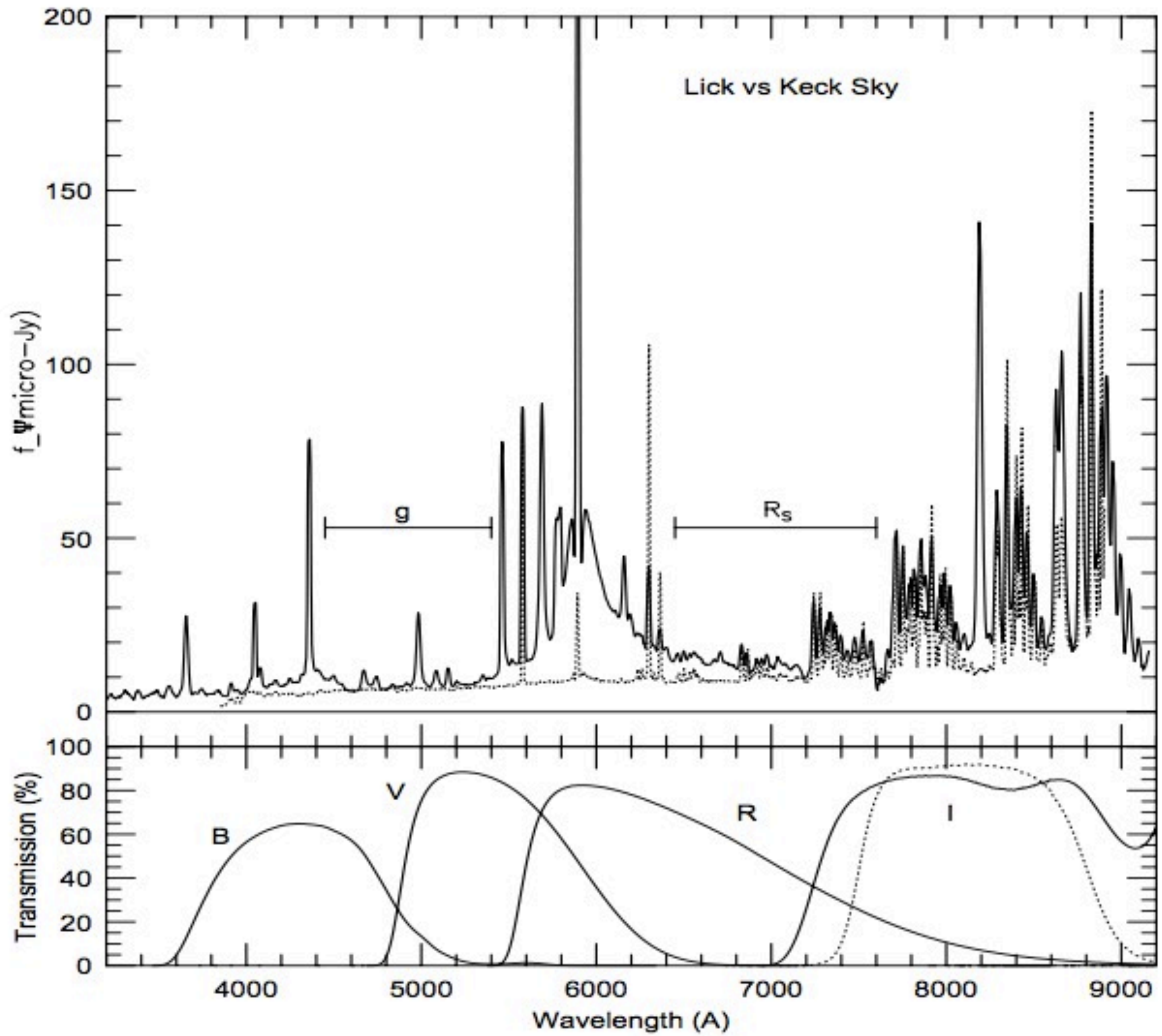


Sky Background

limits most astronomical observations!

always present:
emission from
atmosphere

(+city lights)



Schedule

- We avoid Full Moon ± 3 days for optical labs.
- Reminder: send in your exoplanet transit requests!
- After Exoplanet Lab observations are scheduled, schedule your Lab 2 observations
- (Radio Lab: have to wait for Sun to be further away from equator)
- Class-time from now will be mainly "data analysis help sessions" - still mandatory attendance, unless you have completed all your deadlines for the week