

PHY 517 / AST 443: Observational Techniques in Astronomy

Lecture 5:
Statistics /
Spectroscopy

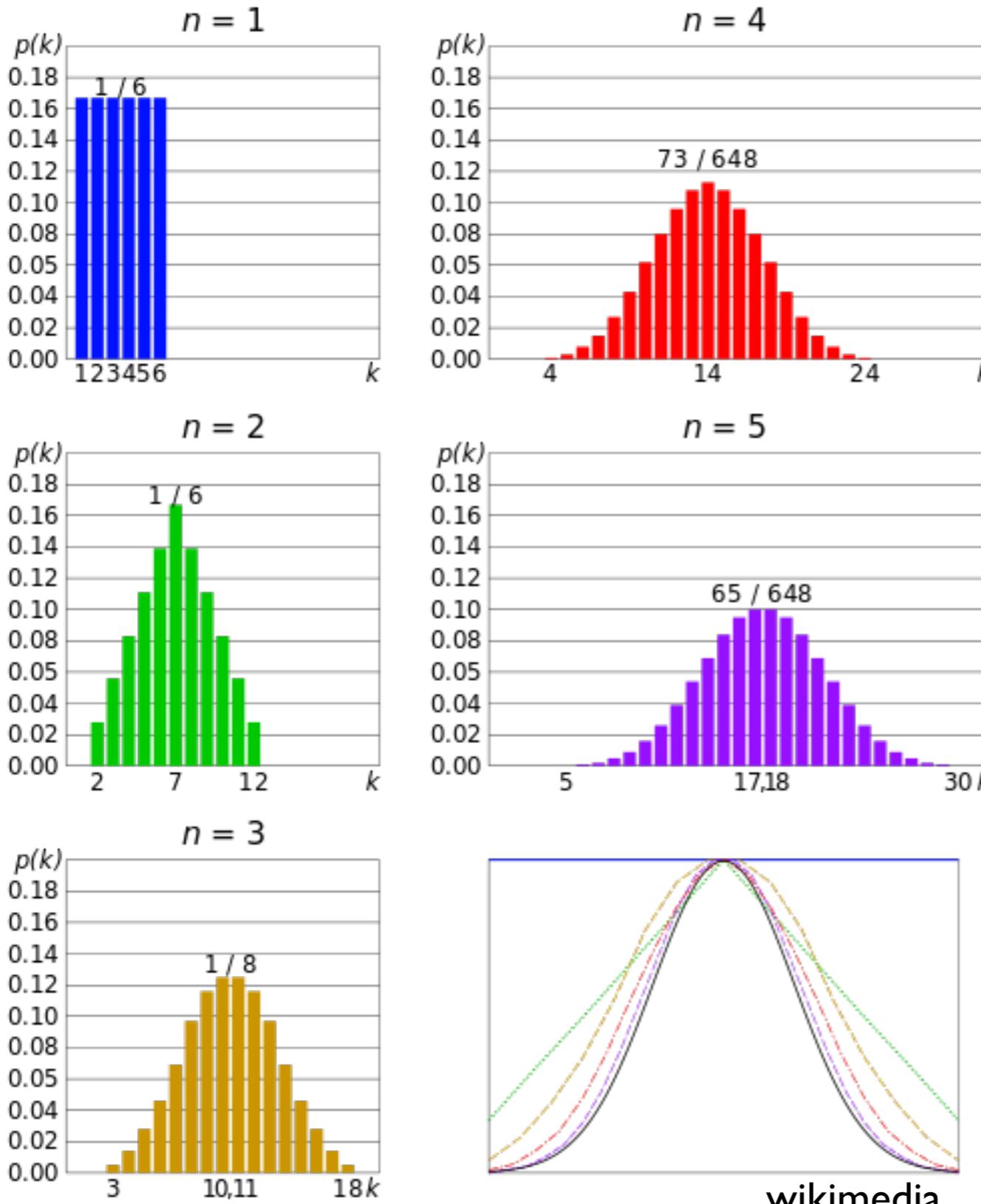
Last Time:

- sample distribution vs. parent population
- summary statistics
- uncertainty on the mean
- Binomial, Poisson, Gaussian 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 *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

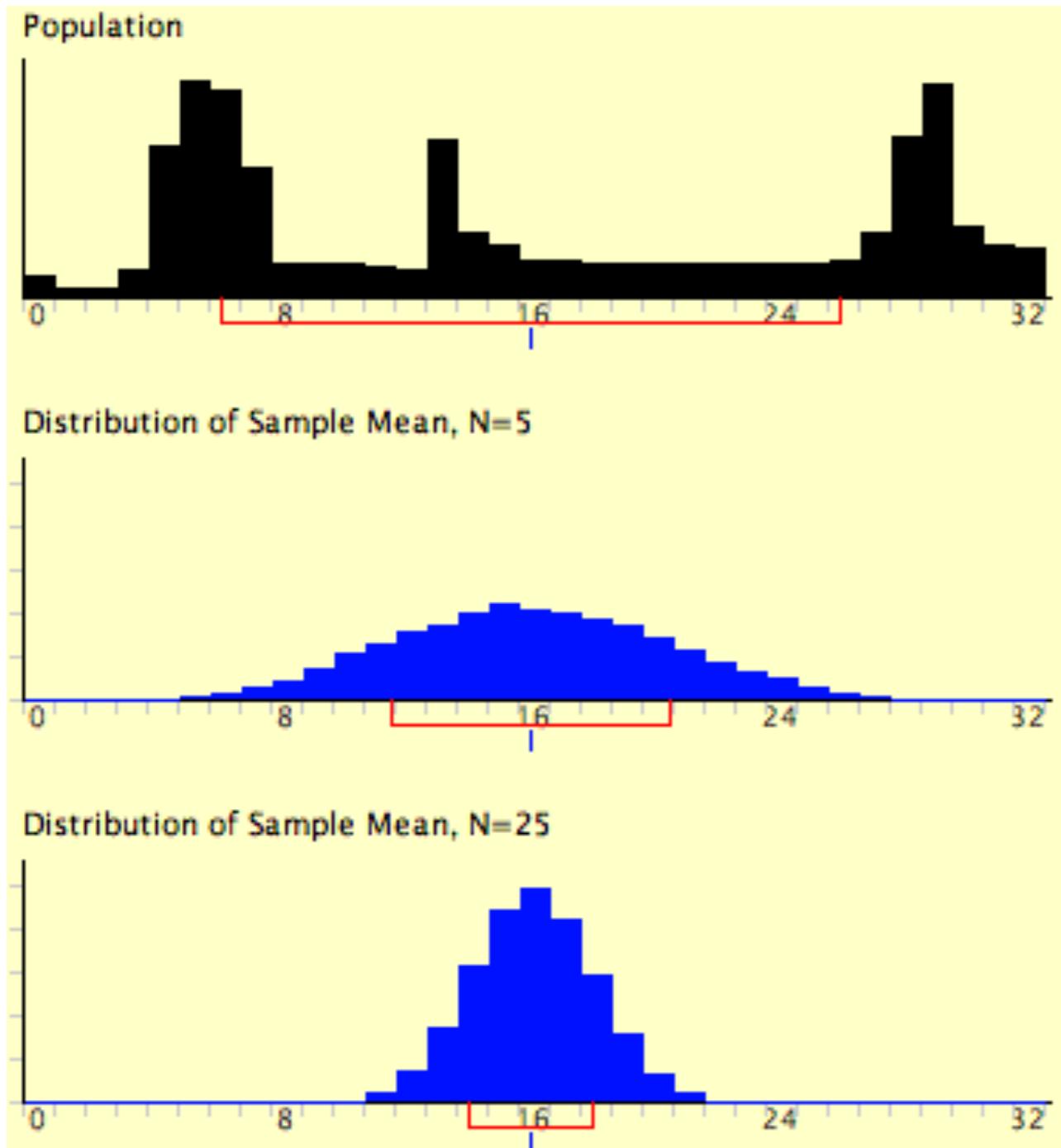
Central Limit Theorem



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



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:

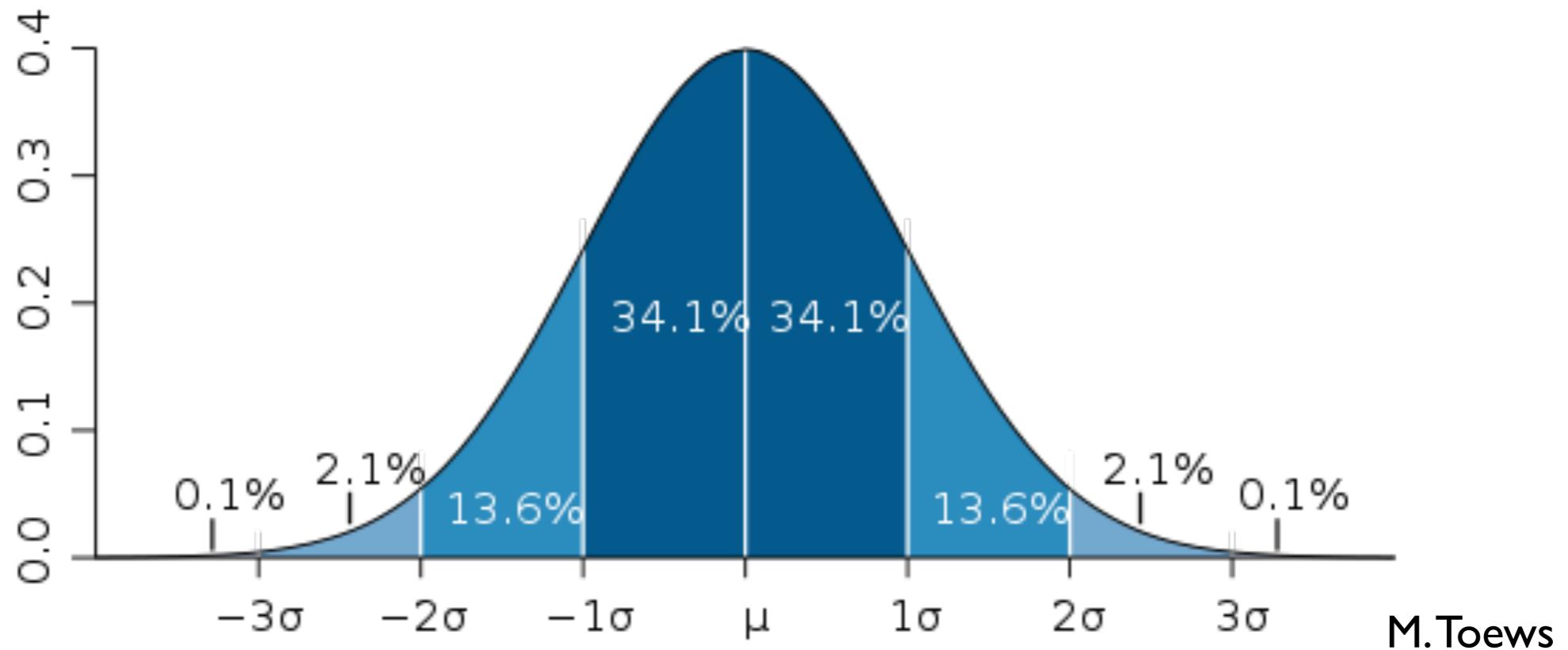
- Mean Flux = $N_{\text{photons}} / \Delta t$ from star - sum of many “draws” of single photons hitting CCD, “success” = detection
- Poisson uncertainty: $\sqrt{N_{\text{photons}}}$

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:



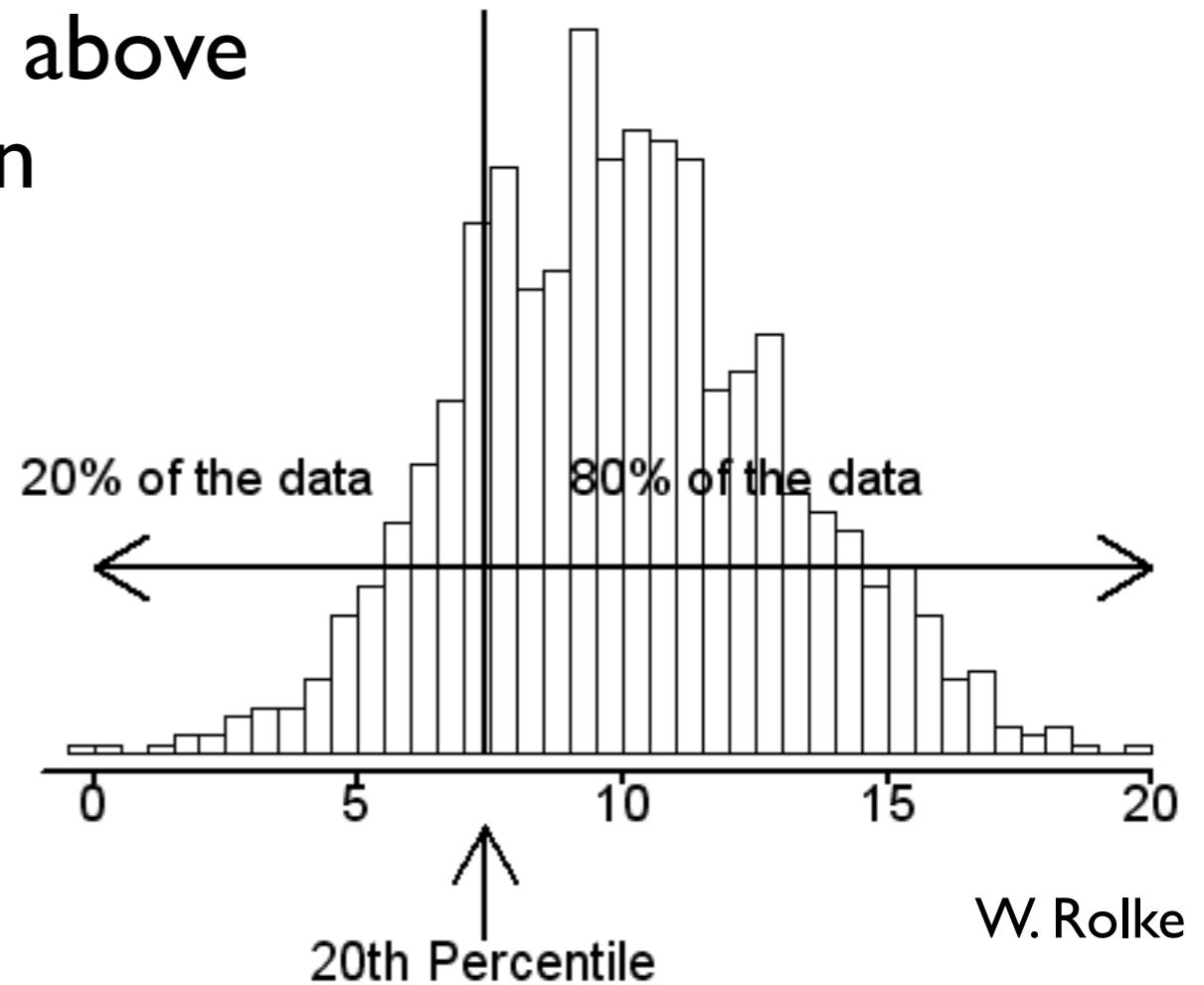
measurements should fall:

- 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

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.:

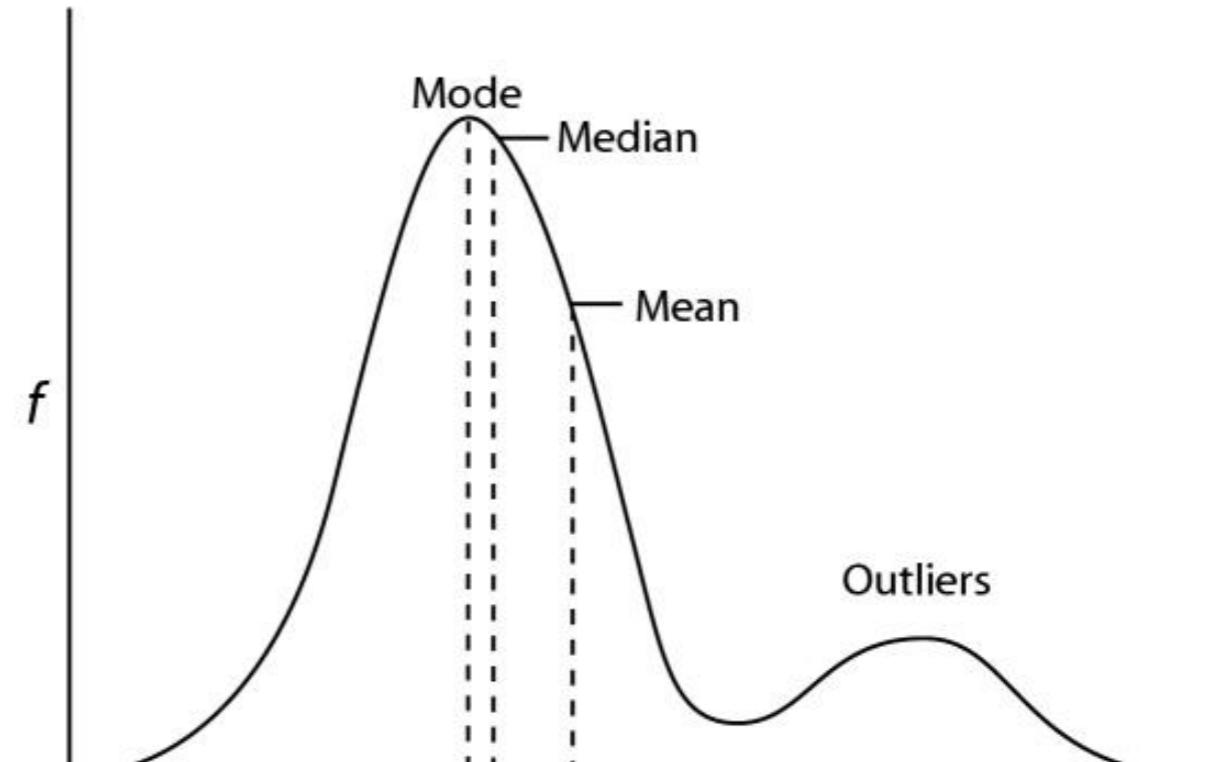
$$99.123^{+0.005}_{-0.004}$$



W. Rolke

Outliers

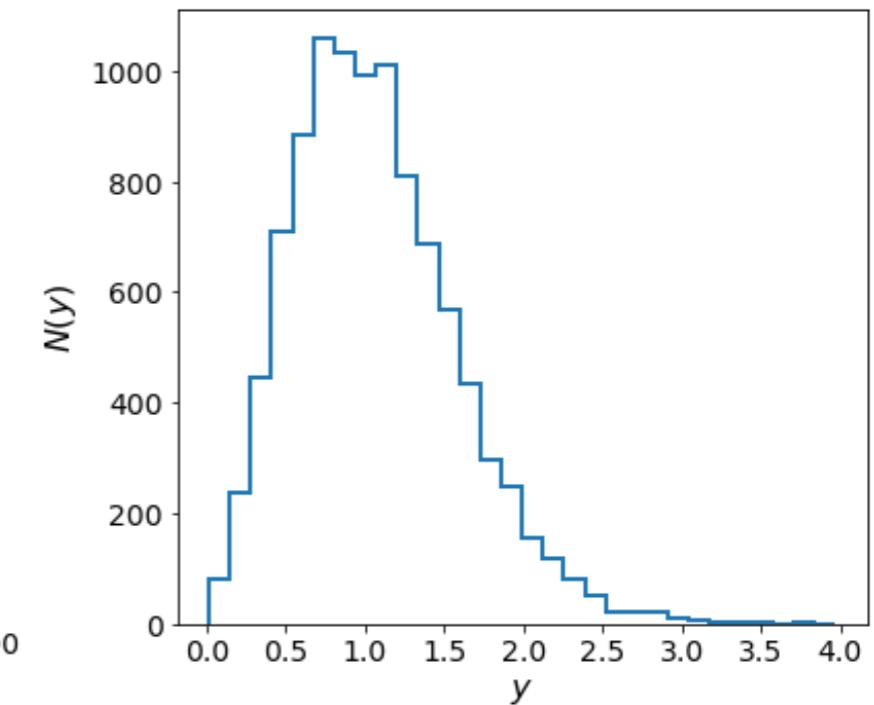
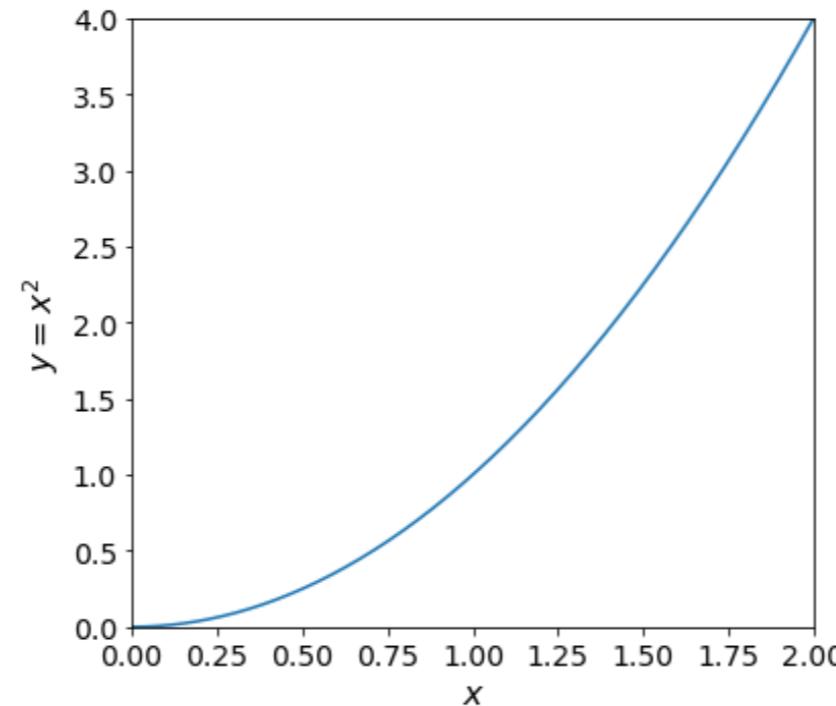
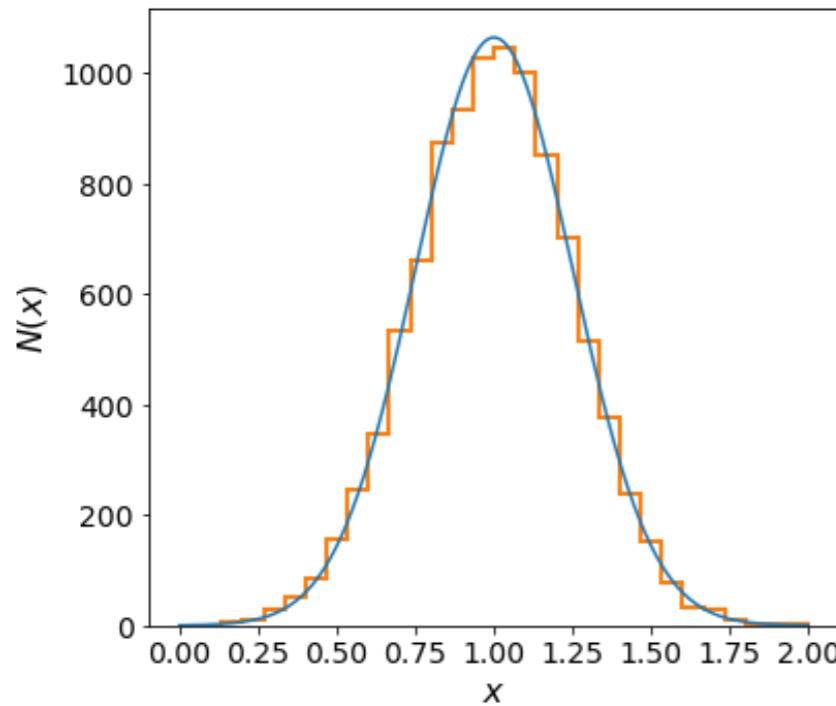
- for normal distribution, median = mean
- what if distribution is “almost” normal, but has a few outliers? e.g. *cosmic rays in dark frame*
- mean: significantly affected by outliers
- median: robust against (small number of) outliers
- **sometimes**, it’s ok to remove gross outliers (“sigma-clipping”),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



Hedges & Shah 2003

Error Propagation

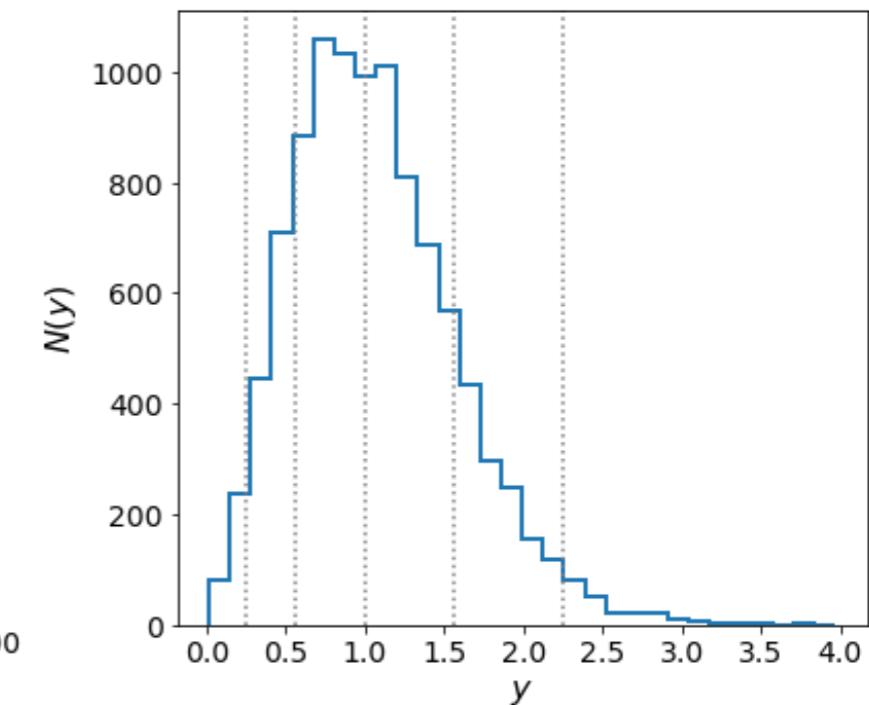
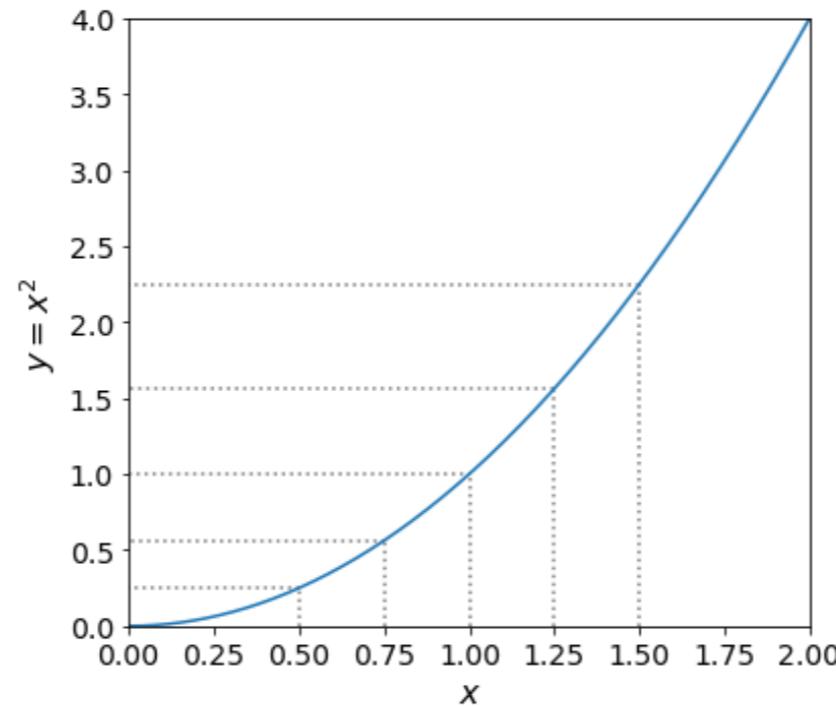
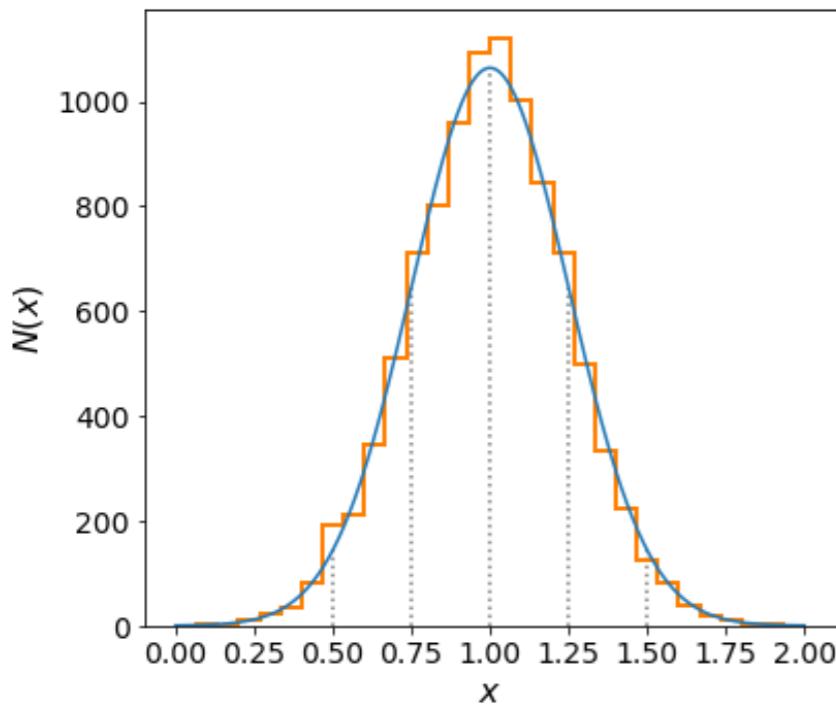
- suppose you have measured ($x \pm \sigma_x$), but you are really interested in $y=f(x)$? How do you determine ($y \pm \sigma_y$)?



- in general, if $f(x)$ is non-linear, $N(y)$ is not Gaussian

Error Propagation

- suppose you have measured ($x \pm \sigma_x$), but you are really interested in $y=f(x)$? How do you determine ($y \pm \sigma_y$)?



- percentiles transform easily (as long as $f[x]$ is monotonous)

Gaussian Error Propagation

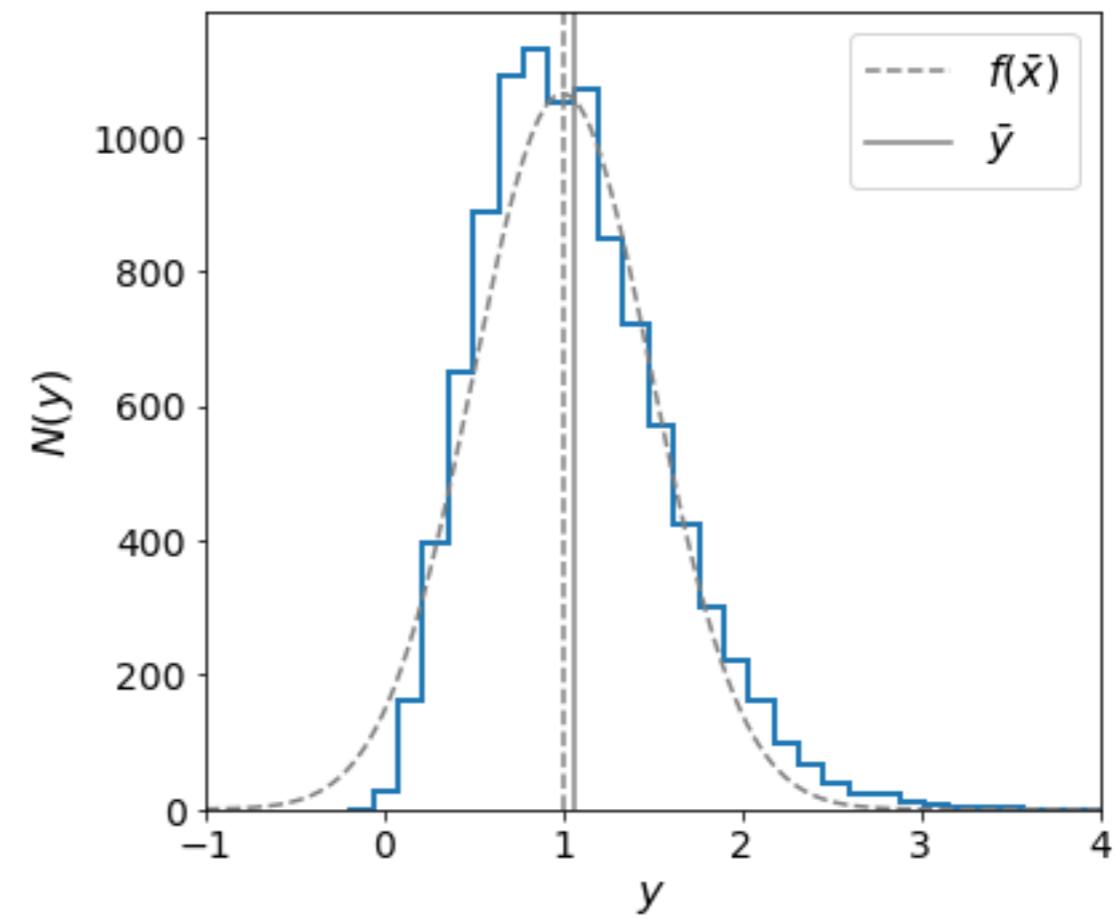
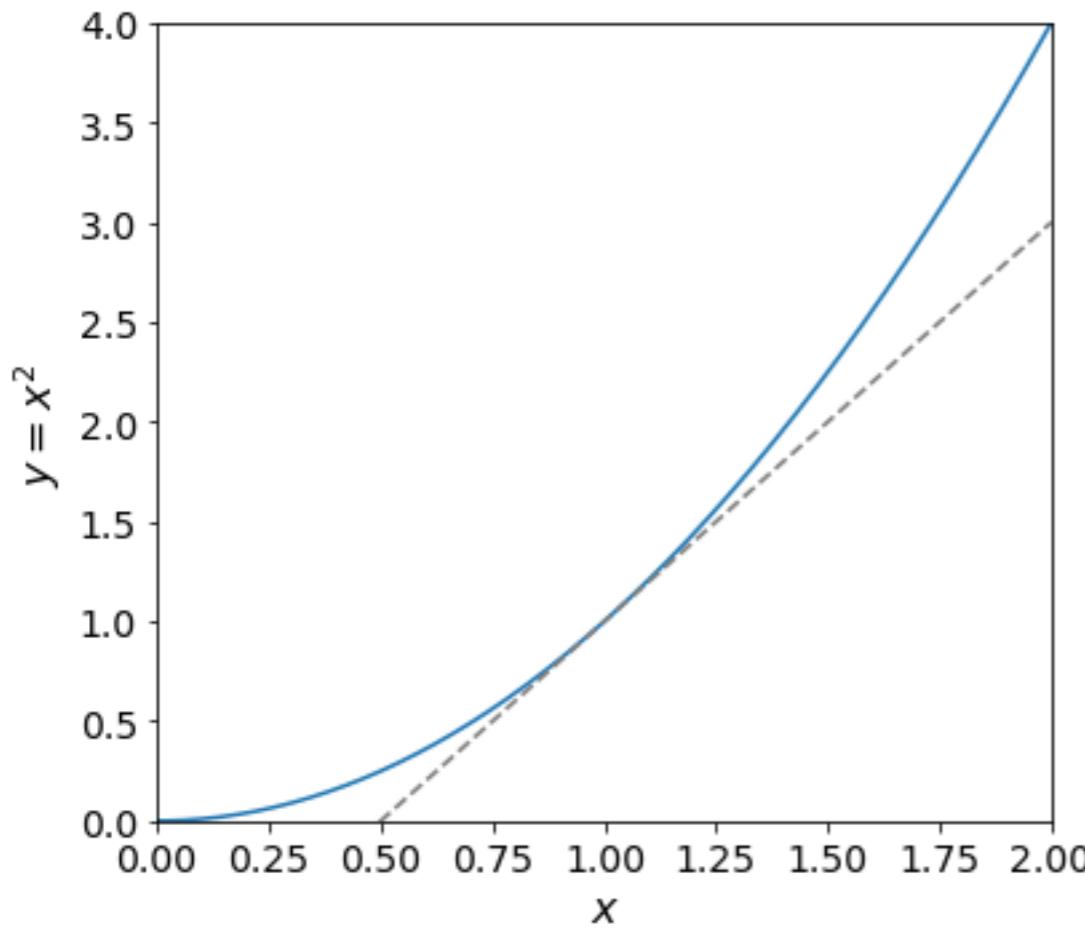
- suppose you have measured $(x_1 \pm \sigma_{x_1})$ and $(x_2 \pm \sigma_{x_2})$, but you are really interested in $y = f(x_1, x_2)$? How do you determine $(y \pm \sigma_y)$?
- Gaussian Error Propagation: approximate $p(y)$ as a Gaussian of mean $\bar{y} \simeq y(\bar{x}_1, \bar{x}_2)$ and width:

$$\sigma_y^2 = \left(\frac{\partial y}{\partial x_1} \right)^2 \sigma_{x_1}^2 + \left(\frac{\partial y}{\partial x_2} \right)^2 \sigma_{x_2}^2$$

Gaussian Error Propagation

- Gaussian Error Propagation: approximate $p(y)$ as a Gaussian of mean $\bar{y} \simeq y(\bar{x}_1, \bar{x}_2)$ and width:

$$\sigma_y^2 = \left(\frac{\partial y}{\partial x_1} \right)^2 \sigma_{x_1}^2 + \left(\frac{\partial y}{\partial x_2} \right)^2 \sigma_{x_2}^2$$



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

Significant detections?

- *The “known” flux of a star is $F = 1.00$ (arbitrary units). You measure a flux of $F = 0.99$. Did you detect an exoplanet transit?*
- A: Yes! This is a large enough difference to claim a detection.
- B: No! The difference is too small to claim a detection.
- C: Can't tell, because ...
- no uncertainties have been provided

Significant detections?

- *The “known” flux of a star is $F = (1.00 \pm 0.002)$ (arbitrary units). You measure a flux of $F = 0.99$. Did you detect an exoplanet transit?*

A: Yes! This is a 5σ detection.

B: No! The difference is too small to claim a detection.

C: Can't tell, because ...

no uncertainties have been provided

Significant detections?

- The “known” flux of a star is $F = (1.00 \pm 0.002)$ (arbitrary units). You measure a flux of $F = (0.99 \pm 0.002)$. Did you detect an exoplanet transit?

A: Yes! This is a 5σ detection.

B: Yes! This is a $\sim 3\sigma$ detection.

C: Can’t tell, because ...

Significance of differences

significance of a deviation from the “expected” value:

$$\frac{|x - \mu|}{\sigma_x}$$

in units of 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:

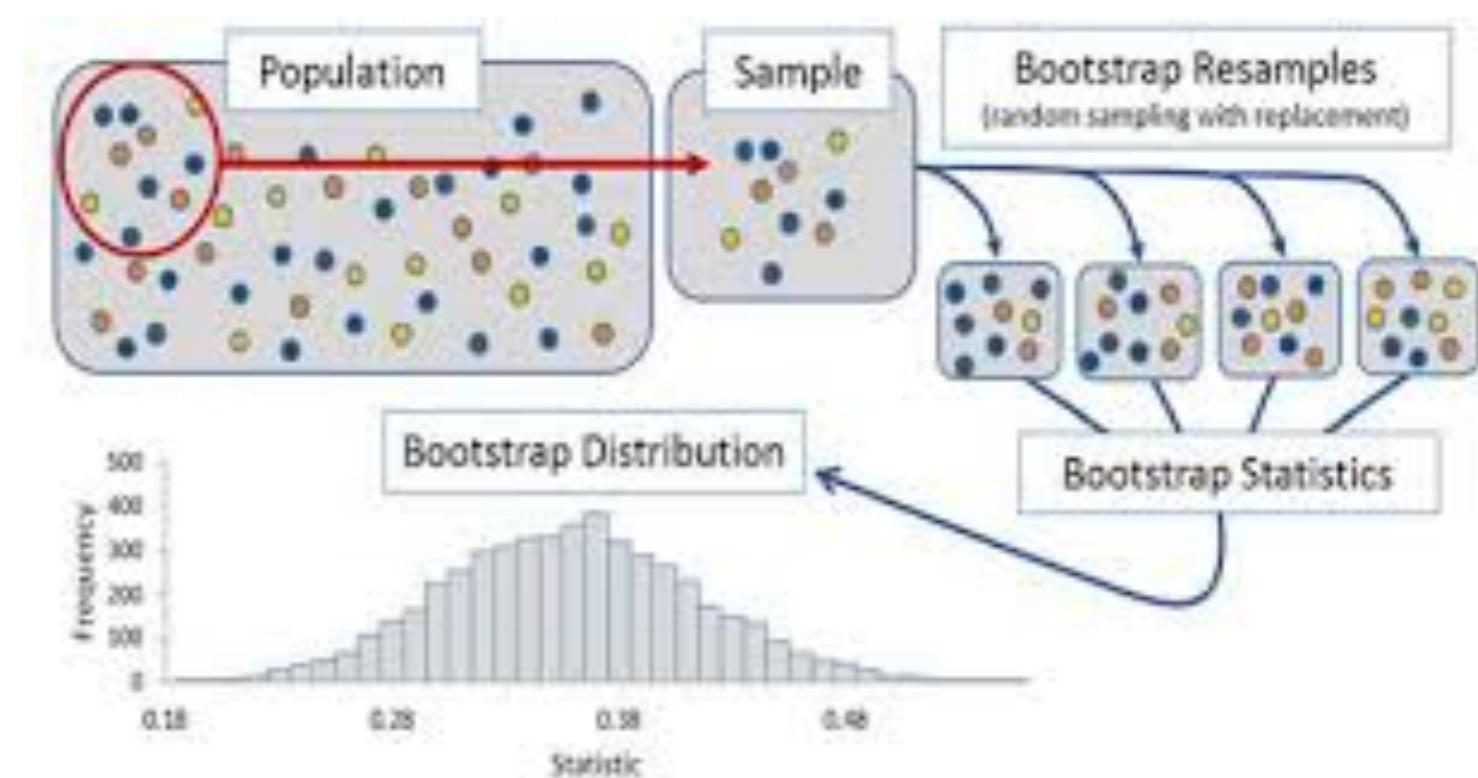
$$\frac{|x_1 - x_2|}{\sigma_{\text{tot}}} = \frac{|x_1 - x_2|}{\sqrt{\sigma_{x_1}^2 + \sigma_{x_2}^2}}$$

Resampling Methods

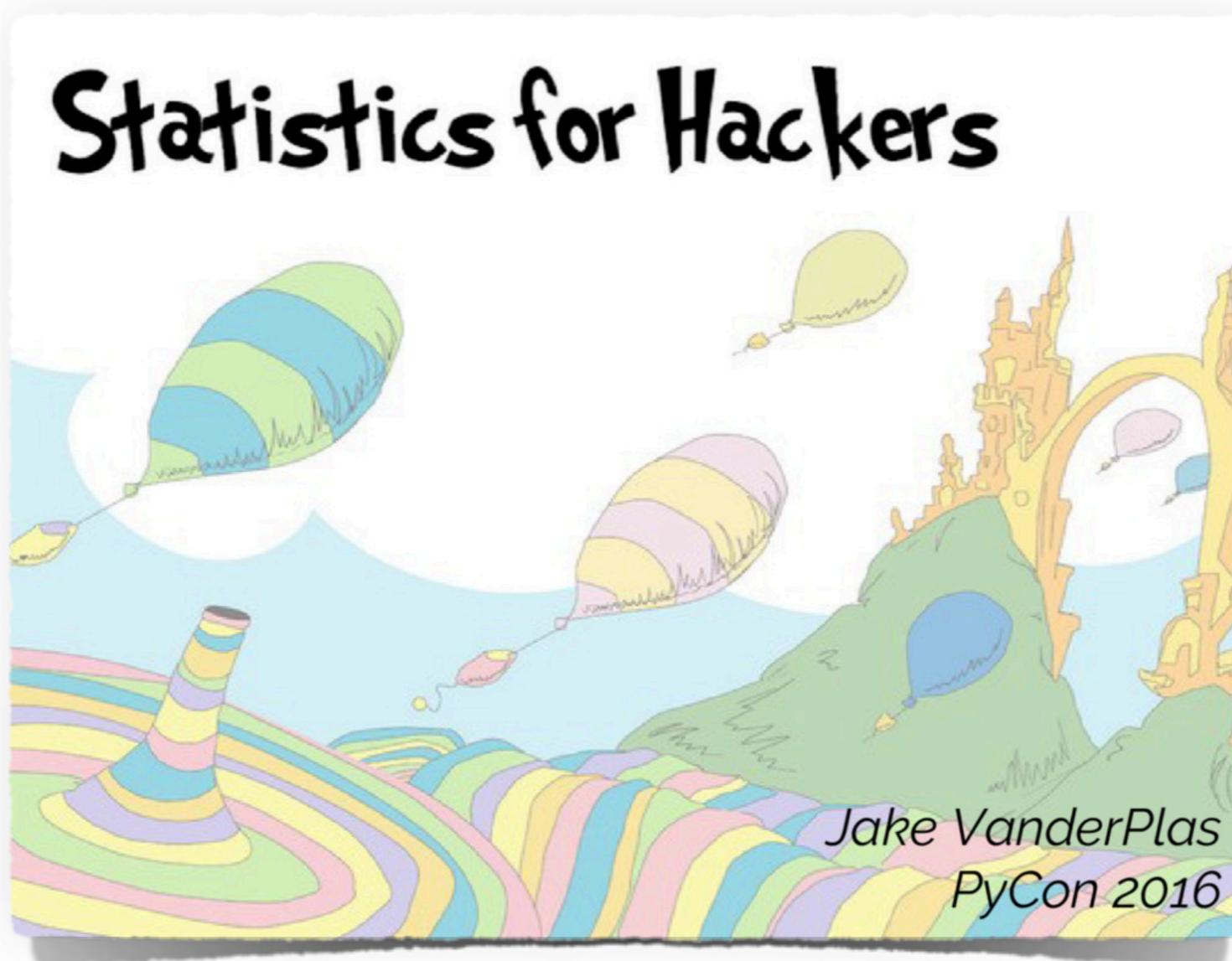
- suppose you have a sample distribution, and do not know the shape of the parent population
- ... and you want to measure *something* from your data, along with an uncertainty on *something*
- (*and the CLT might not hold, e.g. n too small and/or distribution is skewed*)
- can almost always use a resampling method (bootstrap, jackknife, ...)
 - use subsets of the data
 - draw randomly from the data with replacement
 - swap labels on data points

Bootstrapping

- 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 bootstrap distribution



More on resampling techniques



<https://speakerdeck.com/jakevdp/statistics-for-hackers>

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

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

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?

$$\chi_{\nu}^2 = \frac{\chi_{\min}^2}{\nu}$$

- 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 χ_{\min}^2/ν 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

$\chi_{\min}^2/\nu \approx 1$ model is a good fit

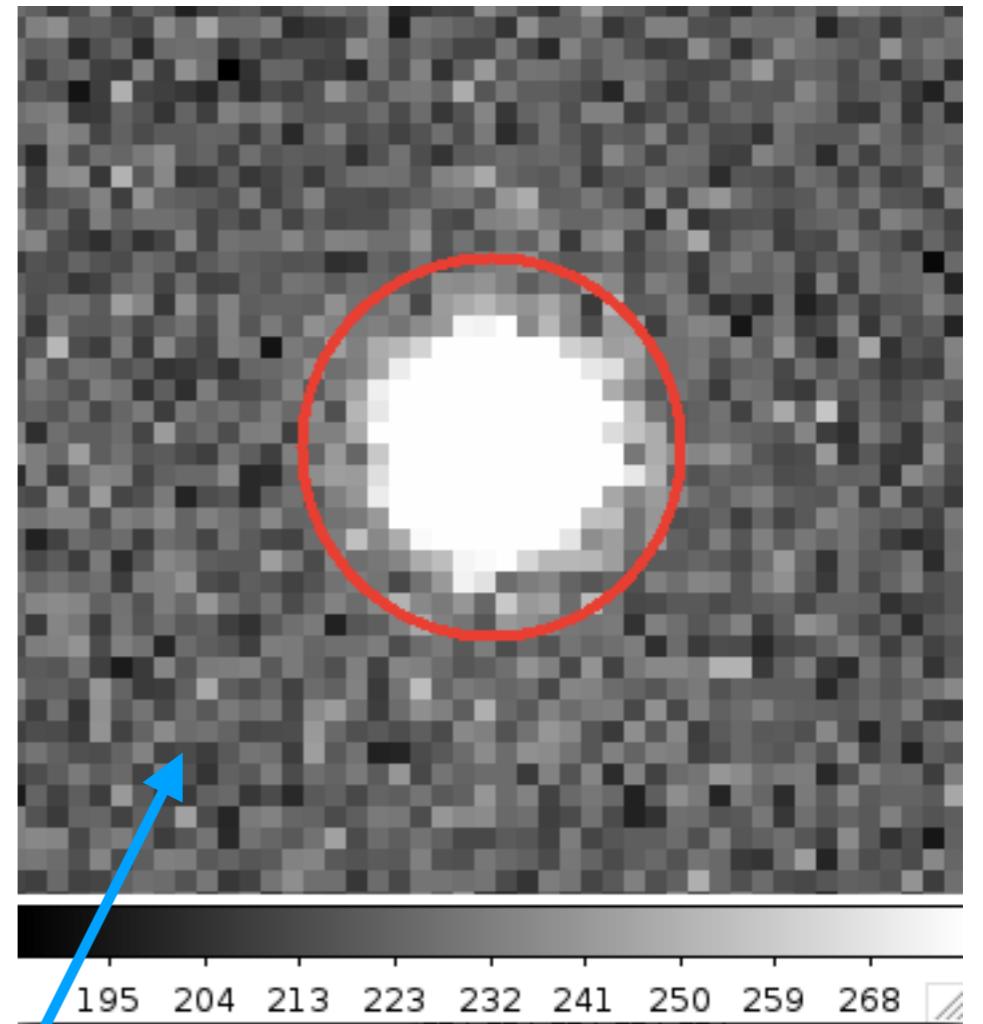
$\chi_{\min}^2/\nu \gg 1$ model is a bad fit

$\chi_{\min}^2/\nu \ll 1$ model is overfitting the data

Signal-to-Noise in CCD images

Signal-to-Noise

- there is always “noise” in a CCD image
- even in bias exposure: read-out noise
- specifically, we refer to the pixel-to-pixel variation in counts as noise
- we can only detect a signal (i.e. an object) when it is larger than the noise

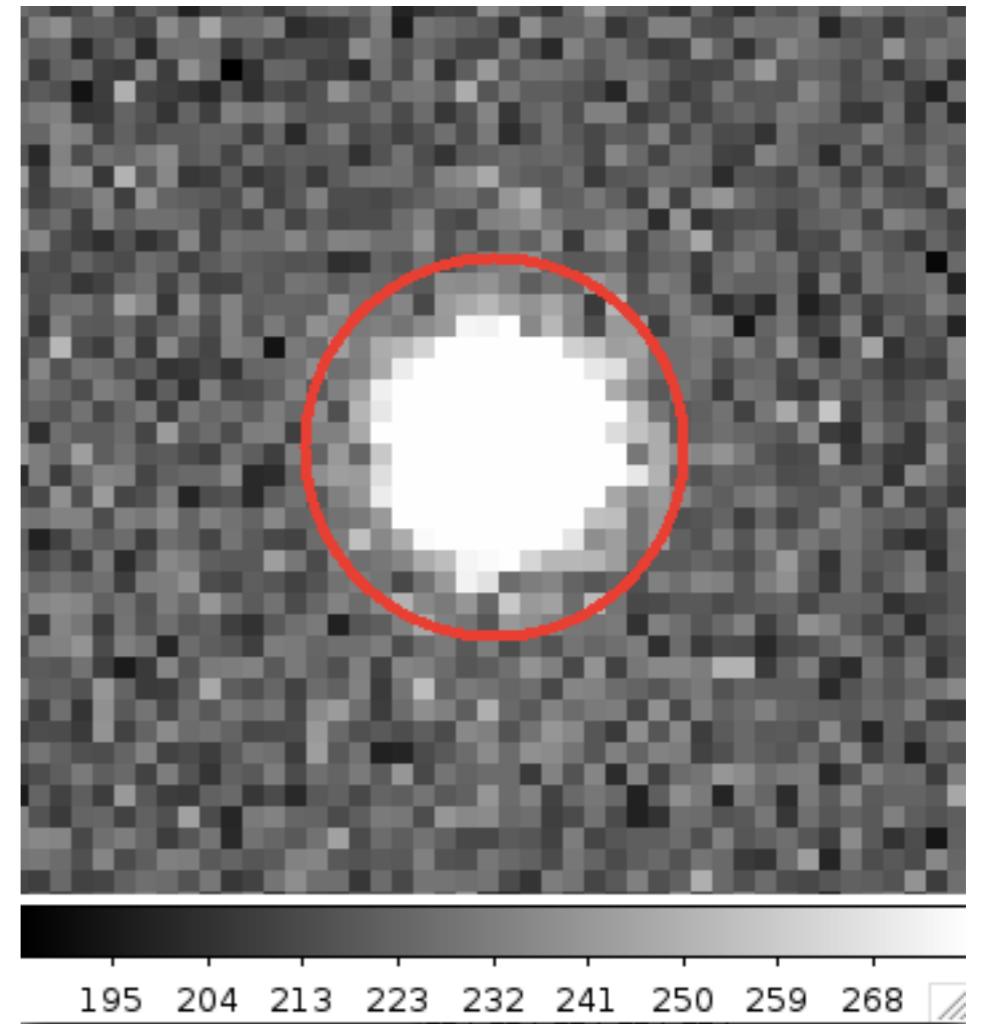


for example: noise from sky background is the standard deviation of pixels without objects

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:

$$N_{\text{object}} = N_{\text{total}} - N_{\text{background}}$$

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

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

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

Signal-to-Noise

noise contributions:

- shot noise from source

$$\begin{aligned}\sigma_{\text{object}} &= \sqrt{N_{\text{object}}} \\ &= \sqrt{S_{\text{object}} \times t}\end{aligned}$$

- sky noise

$$\begin{aligned}\sigma_{\text{sky}} &= \sqrt{N_{\text{sky}}} \\ &= \sqrt{s_{\text{sky}} \times n_{\text{pix}} \times t}\end{aligned}$$

- dark current noise

$$\begin{aligned}\sigma_{\text{dk}} &= \sqrt{N_{\text{dk}}} \\ &= \sqrt{s_{\text{dk}} \times n_{\text{pix}} \times t}\end{aligned}$$

- read-out noise

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

already a std. dev.

Signal-to-Noise

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

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

“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 case 1: very bright object $N_{\text{object}} \gg N_{\text{other}}$

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

Signal-to-Noise

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

limiting case 2: faint objects

$$N_{\text{sky}} \gg N_{\text{other}}$$

$$\begin{aligned} SNR &= \frac{N_{\text{object}}}{\sqrt{N_{\text{sky}}}} = \frac{s_{\text{object}} \times t}{\sqrt{s_{\text{sky}} \times n_{\text{pix}} \times t}} \\ &\propto \sqrt{t} \end{aligned}$$

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

Patat 2004

moonlight:

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

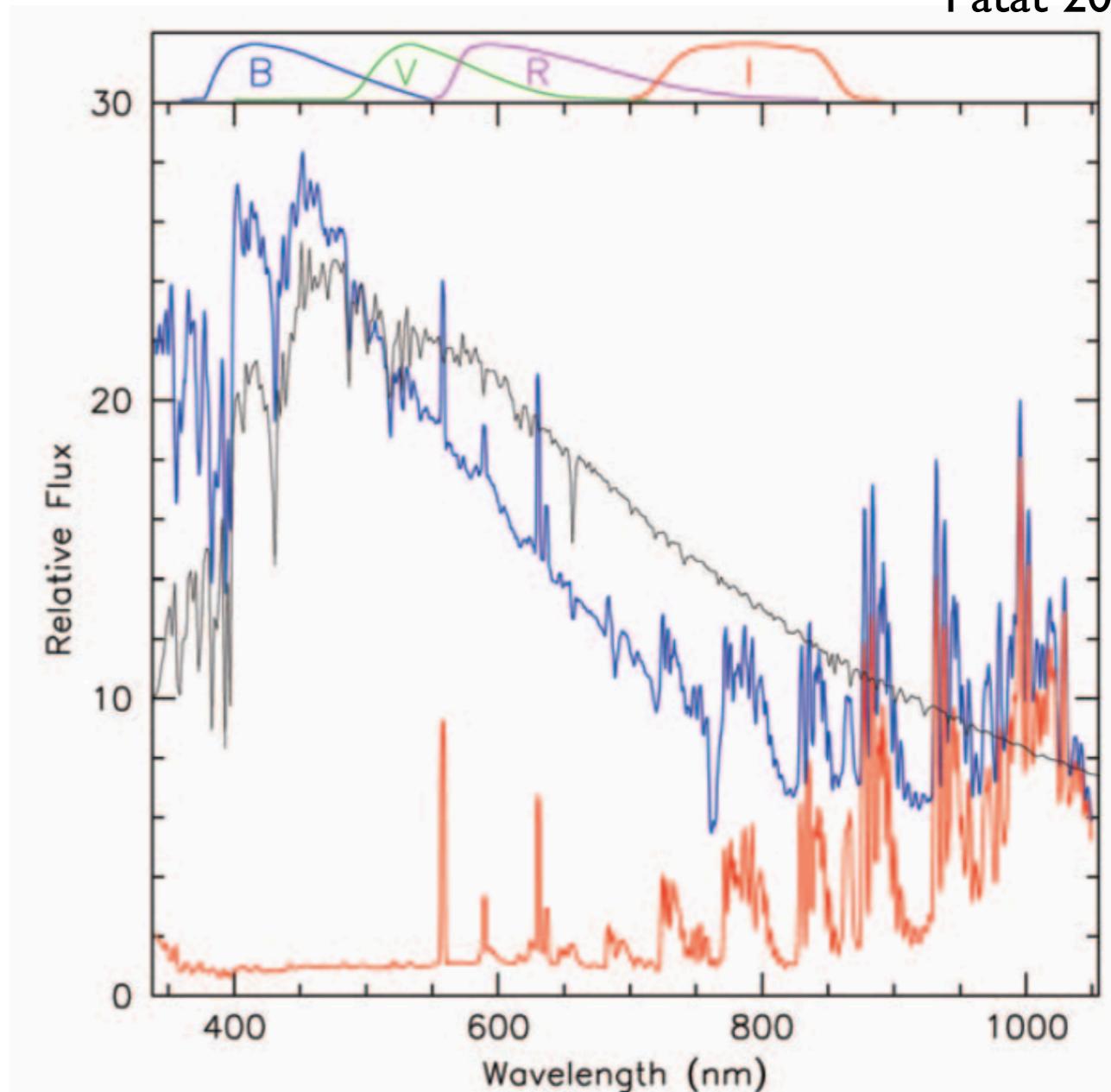
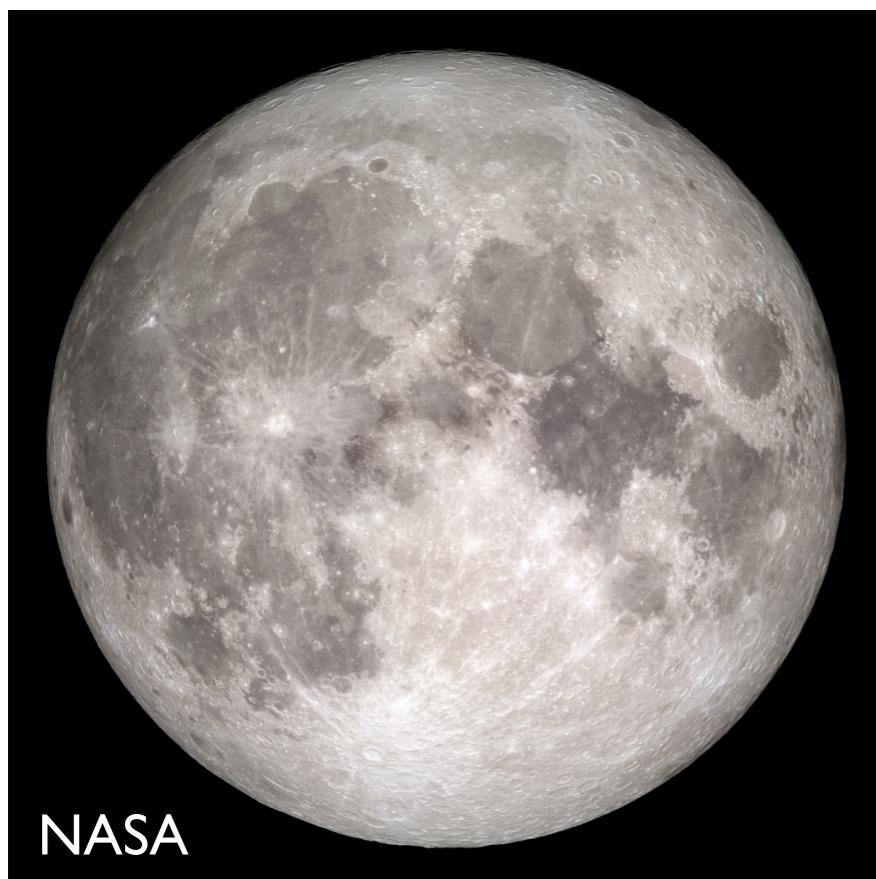
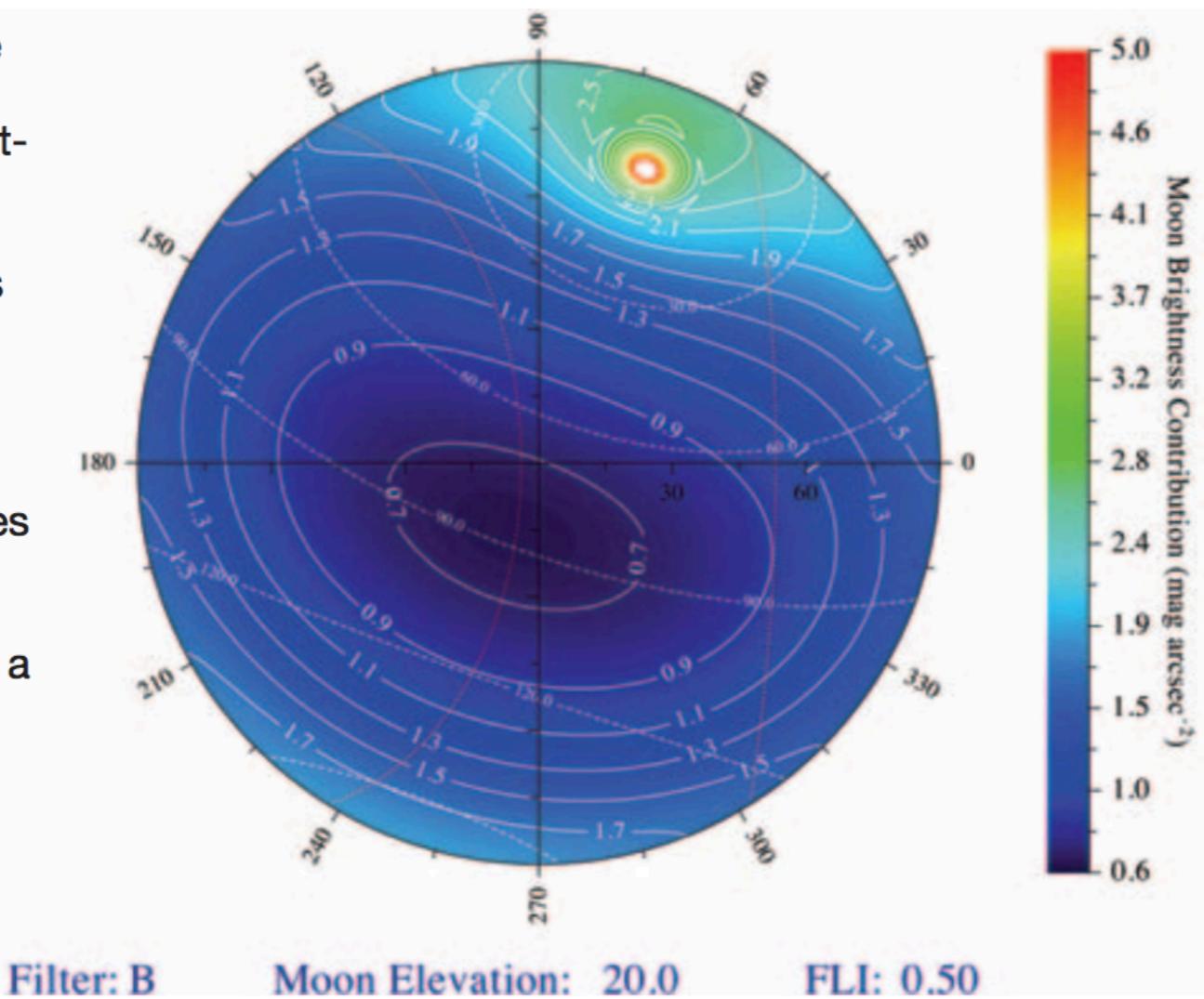


Figure 2: Comparison between the night sky spectrum during dark time (red line, Patat 2003) and bright time (blue line). The latter was obtained with FORS1 on September 1, 2004 using the low dispersion grism 150I and no order sorter filter. Due to the very blue continuum, the spectral region at wavelengths redder than 650 nm is probably contaminated by the grism second order. Both spectra have been normalized to the continuum of the first one at 500 nm. For comparison, the model spectrum of a solar-type star is also plotted (black line). For presentation, this has been normalized to the moonlit night sky spectrum at 500 nm. The upper plot shows the standard *BVRI* Johnson-Cousins passbands.

Sky Background

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

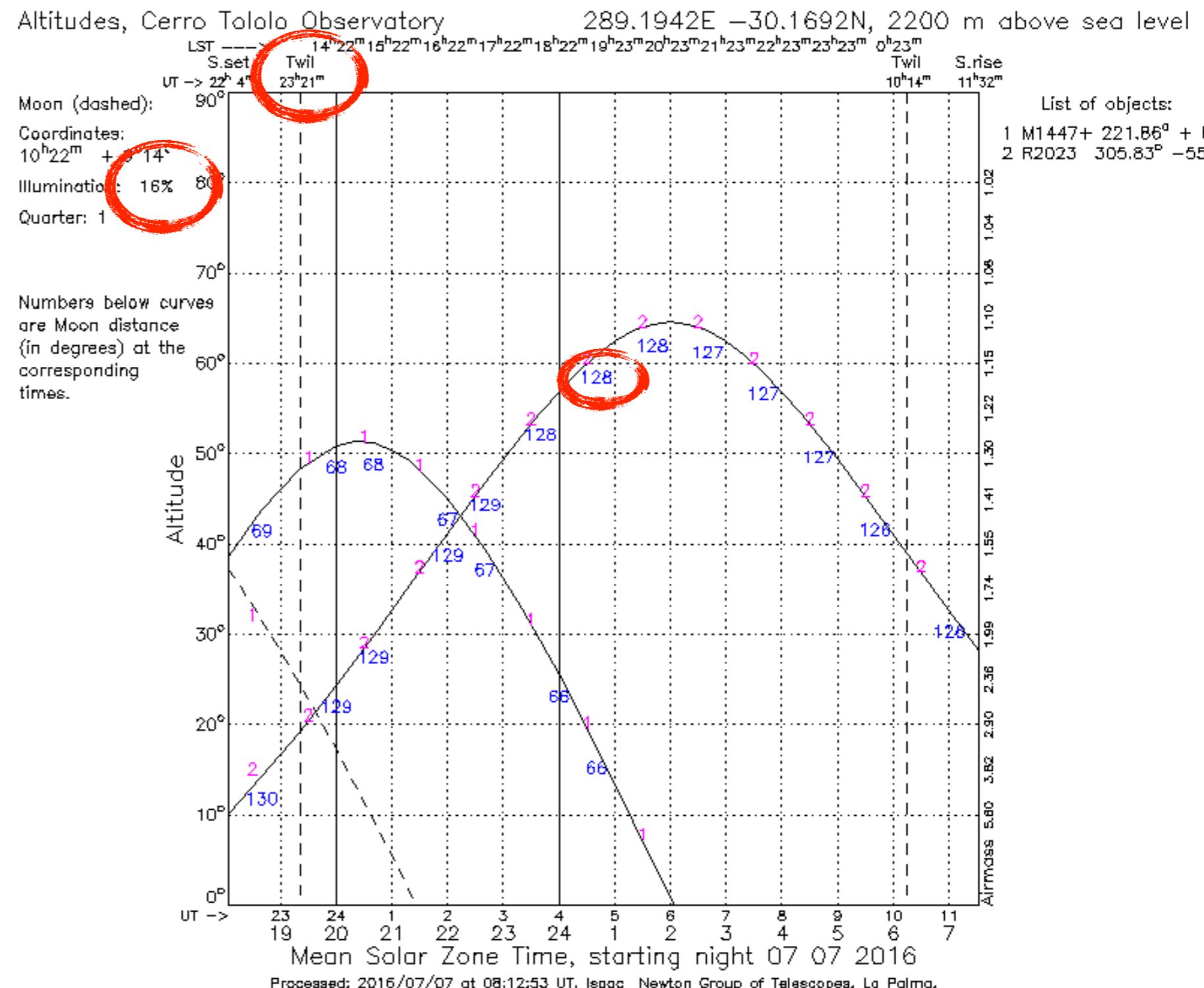
Figure 5: Example isophotal *alt-az* map for the expected moonlight contribution. The dashed white lines trace the loci at constant angular distance from the moon, while the two dotted red lines indicate the extreme apparent lunar paths during a full Saros cycle.



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)

