

# Studying Correlations

## Linear correlations

- single dependent variable:  $y = mx + b$  (*fit a line*)
- multiple dependent variables:  $z = mx + ny + b$  (*fit a plane*)

**Nonlinear correlations:** *try to linearize them!*

**Example #1:** Exponential surface brightness of a disk galaxy

Raw form:  $I(r) = I_0 e^{-r/h}$

Linearized form:  $\ln I(r) = \ln I_0 - r/h$

In surface brightness (mags per sq arcsec):

( $\mu = -2.5 \log(I) + C$ , and remember  
 $\log(x) = \ln(x)/\log(10)$ )

$$\mu(r) = \mu_0 + \frac{2.5}{\ln 10} \frac{r}{h}$$

**Example #2:** Power law form of Tully-Fisher relationship

Raw form:  $L \sim V_{circ}^\alpha$

Linearized form:  $\log L = \alpha \log V_{circ} + C$

## Characterizing a linear (or linearized) relationship:

- Dataset of  $N$  points:  $(x_i, y_i)$
- Fit a line to data:  $y_{fit} = mx + b$
- Calculate **slope**, **intercept**, and their **uncertainties**:  $m \pm \sigma_m, b \pm \sigma_b$
- Calculate root-mean-square (RMS) **scatter** around the fit:  $\sigma_{RMS}^2 \equiv \frac{1}{N} \sum (y_i - y_{fit}(x_i))^2$

Five numbers to characterize a fit:  
 $m, \sigma_m, b, \sigma_b, \sigma_{RMS}$

## The importance of scatter

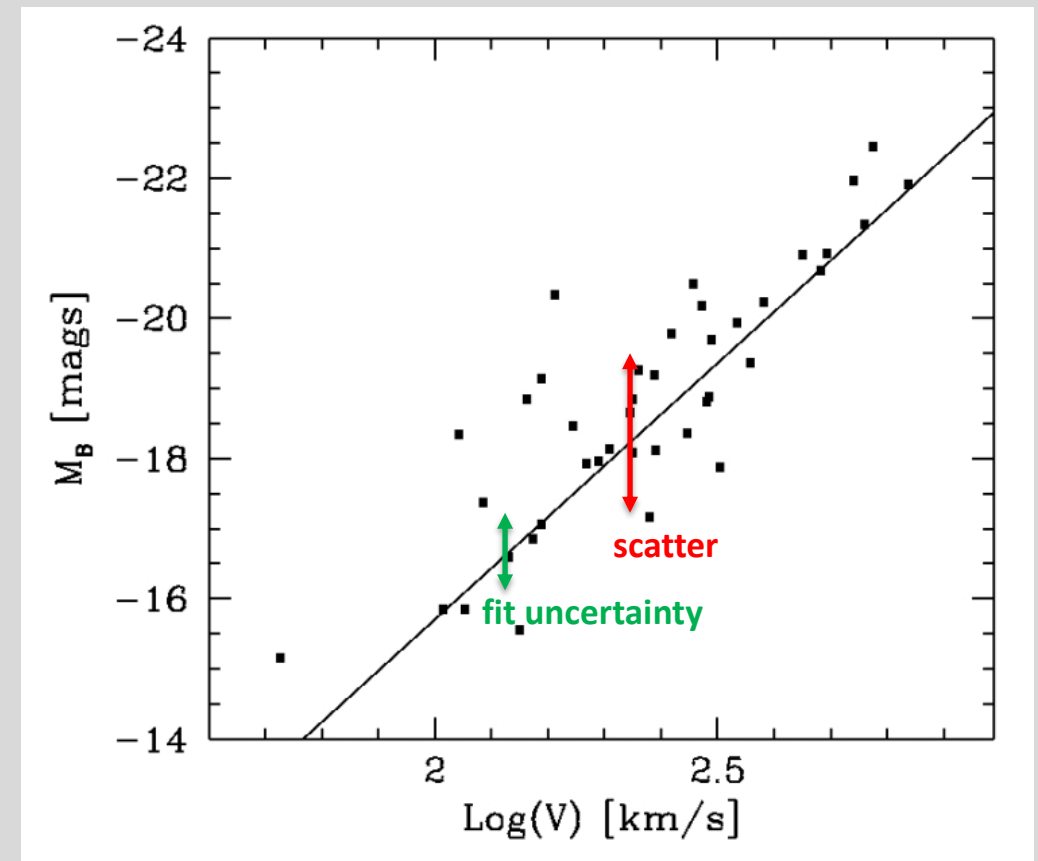
The uncertainties on the fit tell you how well-determined the fit parameters are.

The scatter of the fit tells you how well, on average, individual data points obey the relationship.

### Example: Tully Fisher Relationship $\Rightarrow$

Lower fit uncertainties ( $\sigma_m, \sigma_b$ ) mean that the overall TF relationship is better-determined.

Large scatter ( $\sigma_{RMS}$ ) means any one galaxy may not perfectly obey TF.



Characterizing a linear (or linearized) relationship (least squares fitting, assuming Gaussian statistics):

```
# make a linear fit, and calculate uncertainty and scatter

good = <some criterion> # dont want to include bad data

coeff, cov = np.polyfit(x[good],y[good],1,cov=True)

coeff_err = np.sqrt(np.diag(cov))

print(' slope = {:.3f} +/- {:.3f}'.format(coeff[0],coeff_err[0]))

print('intercept = {:.3f} +/- {:.3f}'.format(coeff[1],coeff_err[1]))

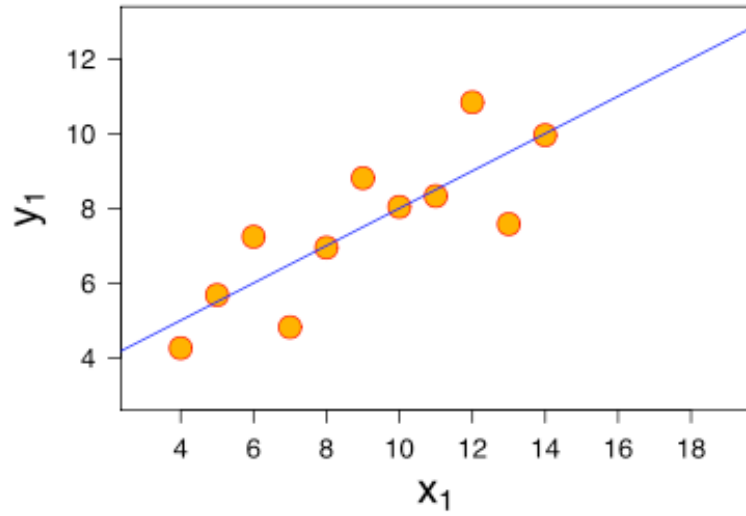
polynomial=np.poly1d(coeff)

xfit=np.linspace(x.min(),x.max())

plt.plot(xfit,polynomial(xfit),color='green',lw=3)

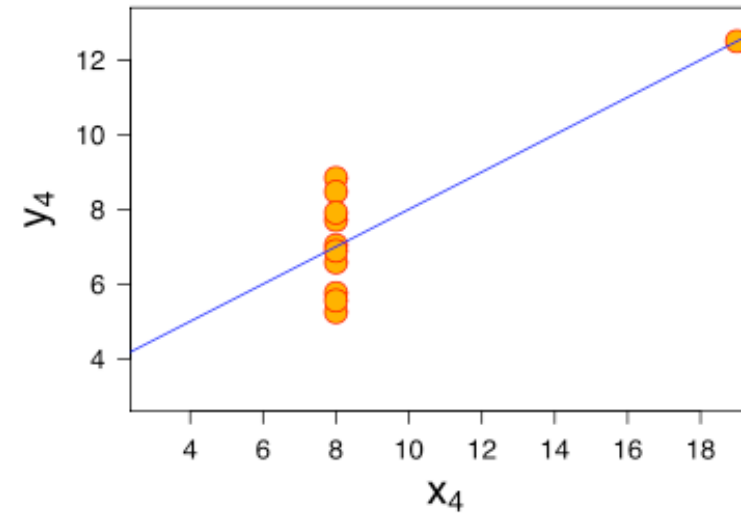
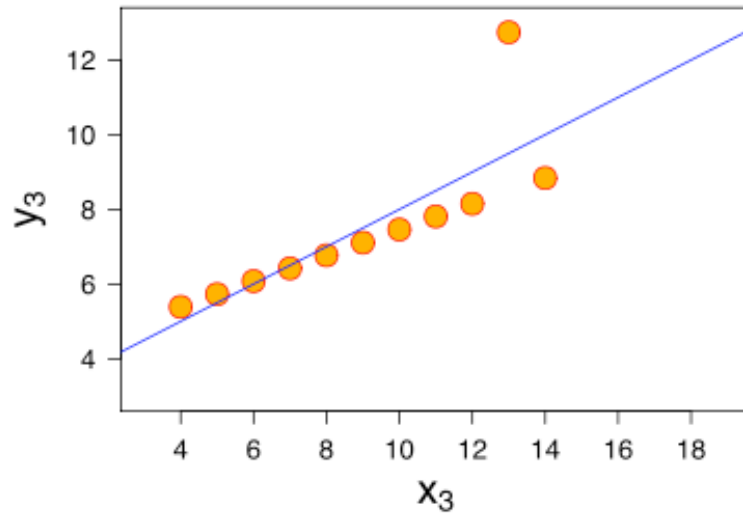
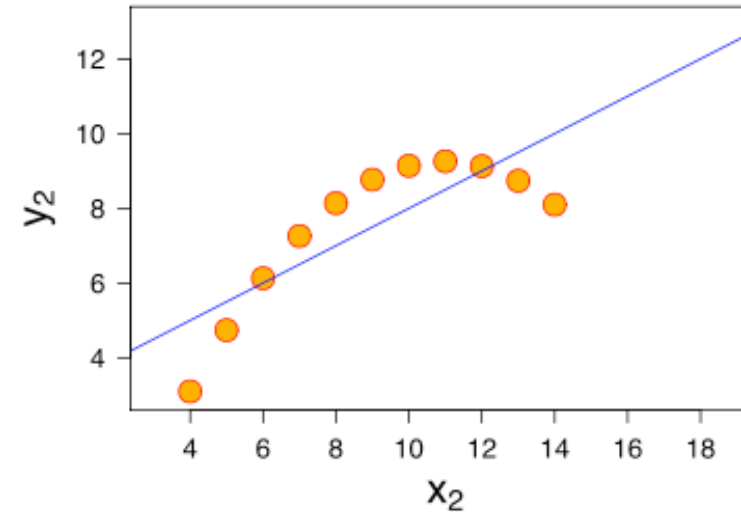
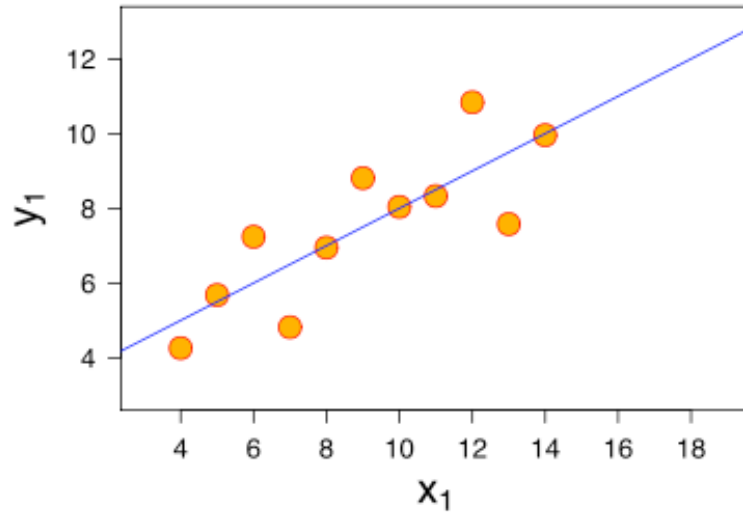
print(' scatter = {:.3f}'.format(np.std(y[good]-polynomial(x[good]))))
```

But be careful  
with fits...



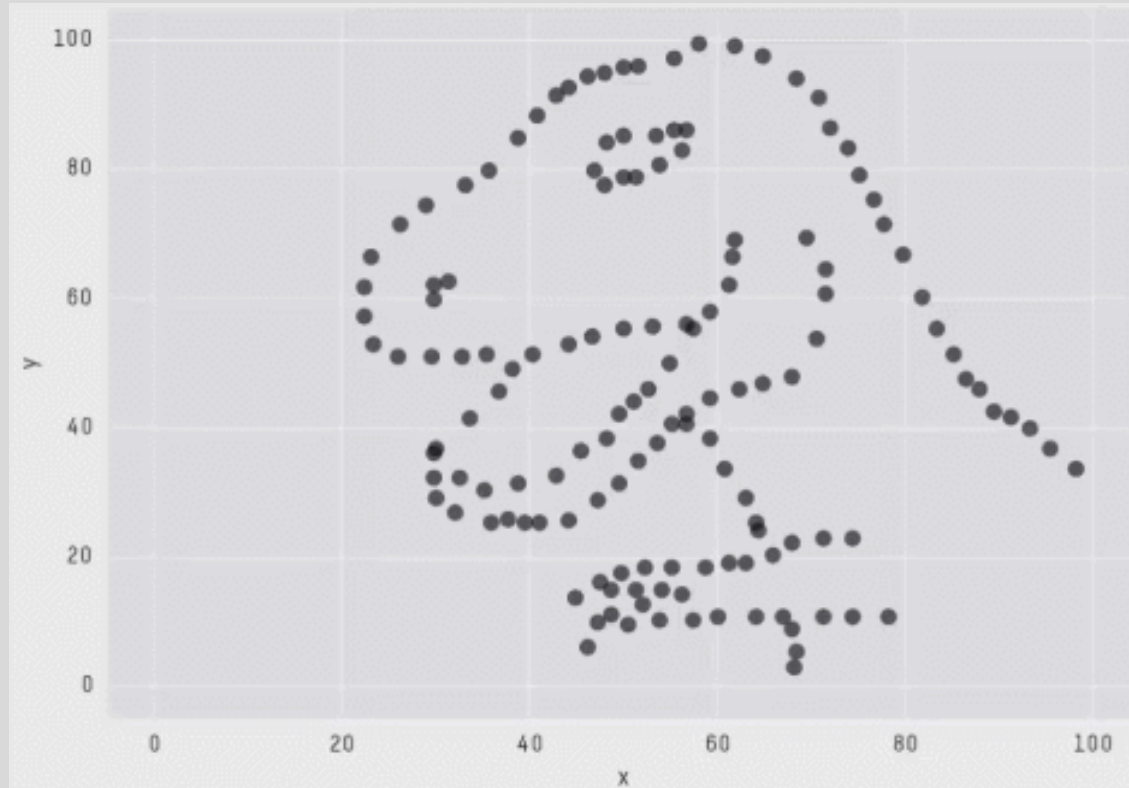
Anscombe's quartet: Fit  $y=mx+b$  and get the same  $r$  (correlation coefficient),  $m$ ,  $b$ ,  $\sigma_m$ ,  $\sigma_b$ ,  $\sigma_{RMS}$

But be careful  
with fits...



Anscombe's quartet: Fit  $y=mx+b$  and get the same  $r$  (correlation coefficient),  $m$ ,  $b$ ,  $\sigma_m$ ,  $\sigma_b$ ,  $\sigma_{RMS}$

## ***Beware the datasaurus!***



X Mean: 54.2659224  
Y Mean: 47.8313999  
X SD : 16.7649829  
Y SD : 26.9342120  
Corr. : -0.0642526

***Moral of the story: ALWAYS PLOT YOUR DATA!***

## Modeling Uncertainty

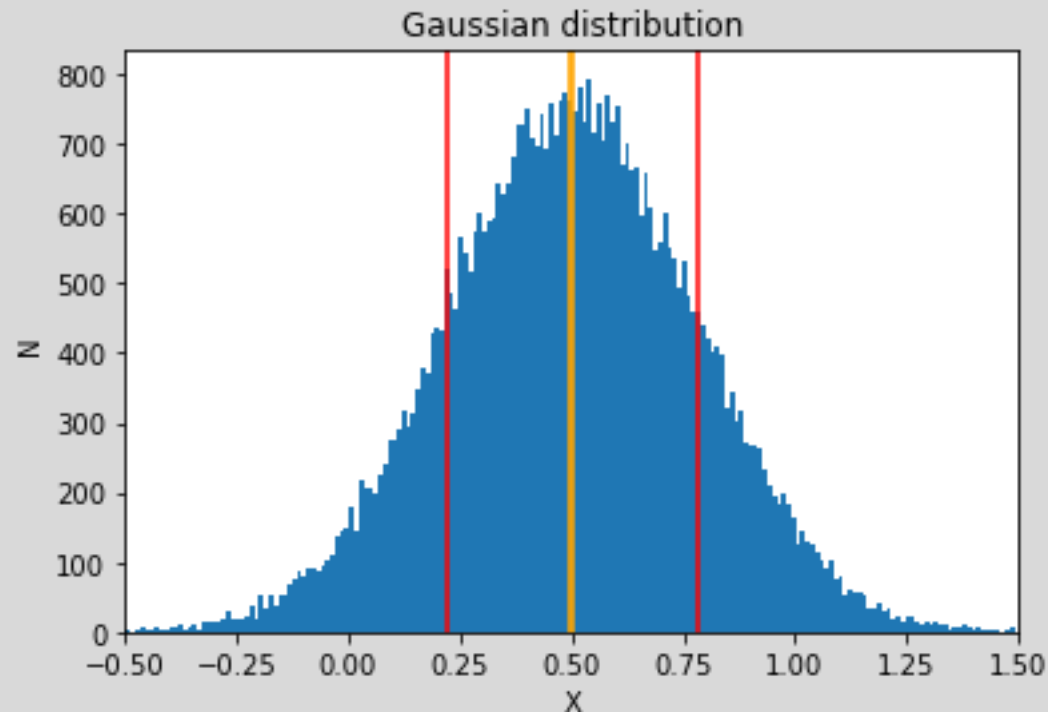
Let's say you have a repeated measurements of some value. How do we estimate the best value and uncertainty.

If your errors are independent and follow a Gaussian distribution:

- measure mean and standard deviation ( $\bar{x}, \sigma$ )
- “standard error in the mean” is given by  $\sigma/\sqrt{N}$

Is this a good assumption? Take a distribution of 50,000 measurements with  $\bar{x}, \sigma = 0.5, 0.28$ , look at distribution.

(yellow: mean, red: mean  $\pm 1\sigma$ )



## Modeling Uncertainty

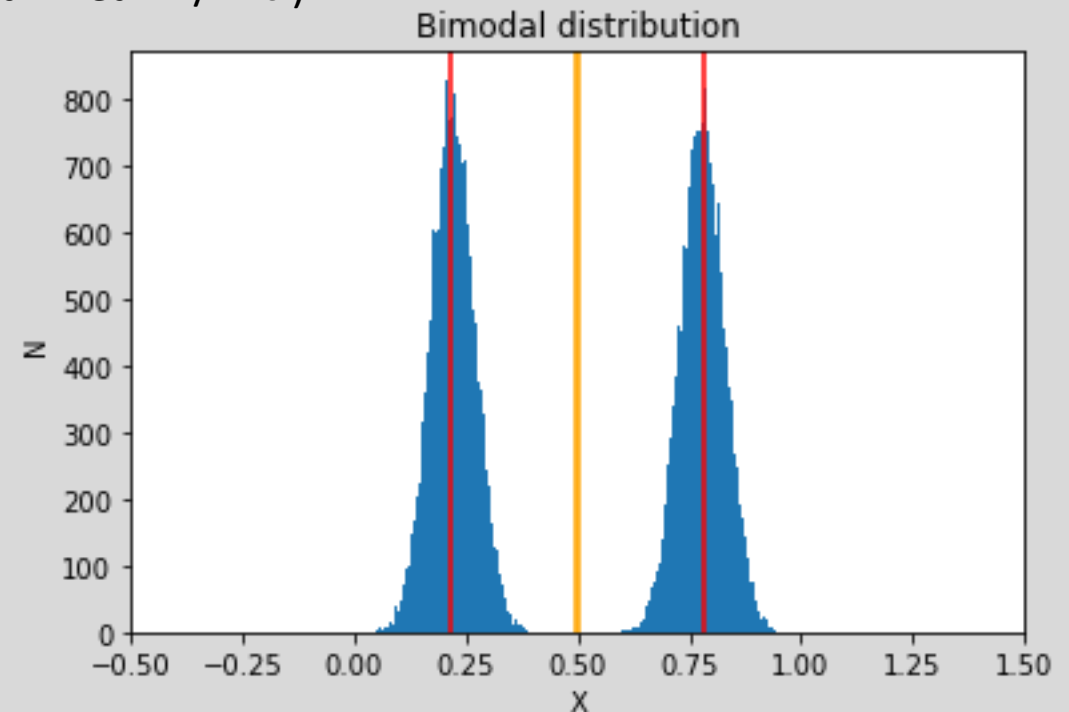
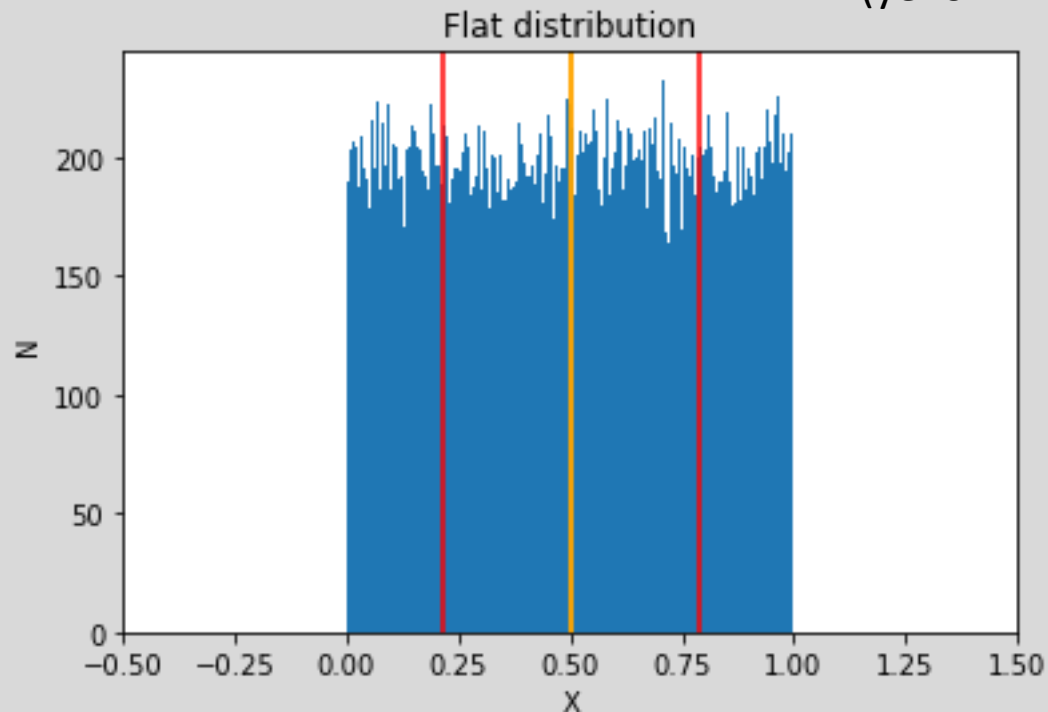
Let's say you have a repeated measurements of some value. How do we estimate the best value and uncertainty.

If your errors are independent and follow a Gaussian distribution:

- measure mean and standard deviation ( $\bar{x}, \sigma$ )
- “standard error in the mean” is given by  $\sigma/\sqrt{N}$

Is this a good assumption? Take a distribution of 50,000 measurements with  $\bar{x}, \sigma = 0.5, 0.28$ , look at distribution.

(yellow: mean, red: mean  $\pm 1\sigma$ )





## Modeling Uncertainty

Let's say you have a repeated measurements of some value. How do we estimate the best value and uncertainty.

If your errors are independent and follow a Gaussian distribution:

- measure mean and standard deviation ( $\bar{x}, \sigma$ )
- “standard error in the mean” is given by  $\sigma/\sqrt{N}$

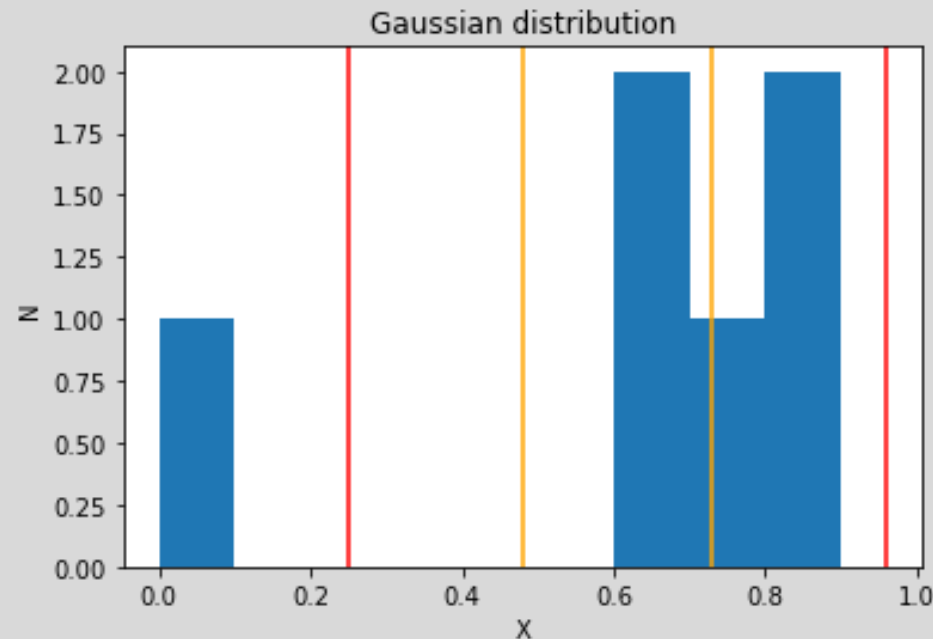
What about small N? Take a distribution of 8 measurements with  $\bar{x}, \sigma = 0.5, 0.28$ , look at distribution.

(yellow: mean  $\pm$  std err, red: mean  $\pm 1\sigma$ )

### Moral of the story:

*Assuming gaussian statistics is often a bad idea!*

*Gotta look at your data!*



## Advanced Parameter Estimation

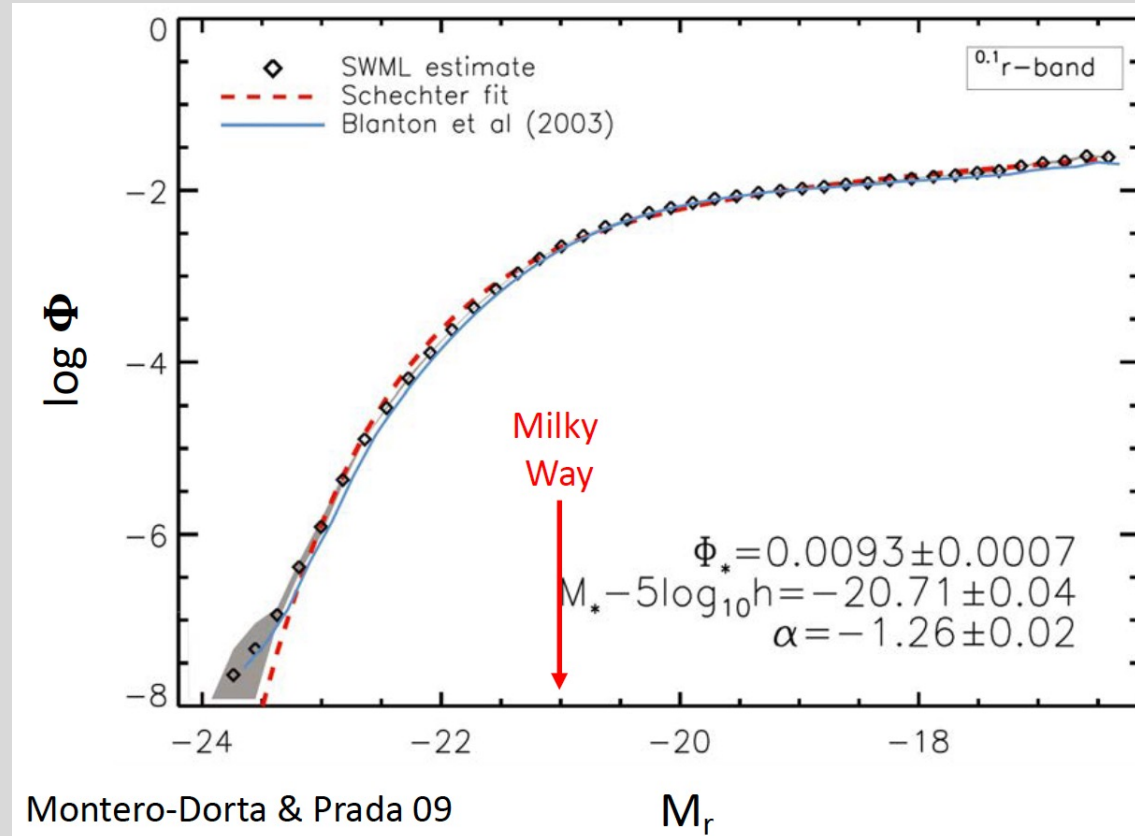
Let's say you have a sample of galaxies in a galaxy cluster, and you've measured all their luminosities. Now want to model the luminosity function of the cluster – the number of galaxies as a function of their luminosity,  $N(L)$ .

Adopt the Schechter function:  $N(L) = \Phi_0 L^\alpha e^{-L/L_*}$

## Advanced Parameter Estimation

Let's say you have a sample of galaxies in a galaxy cluster, and you've measured all their **absolute magnitudes**. Now want to model the luminosity function of the cluster – the number of galaxies as a function of their **absolute mag**,  $N(M)$ .

Adopt the Schechter function:  $N(M)dM = 0.4 \ln 10 \phi_* 10^{-0.4(\alpha+1)(M-M_*)} e^{-10^{-0.4(M-M_*)}} dM$



For a huge sample of galaxies (~ 100,000), it might look something like this.

## Advanced Parameter Estimation

Let's say you have a sample of galaxies in a galaxy cluster, and you've measured all their **absolute magnitudes**. Now want to model the luminosity function of the cluster – the number of galaxies as a function of their **absolute mag**,  $N(M)$ .

Adopt the Schechter function: 
$$N(M)dM = 0.4 \ln 10 \phi_* 10^{-0.4(\alpha+1)(M-M_*)} e^{-10^{-0.4(M-M_*)}} dM$$

But you don't have a huge sample, since you are looking at one specific cluster. How do you estimate  $\alpha$ , and  $L_*$ ?

### Standard approach:

Bin galaxies by magnitude to create  $N(M)$ , then do a (non-linear) chi-sq fit, and solve for the parameters.

### Problems:

- The errors in  $N(M)$  come from magnitude uncertainties, low  $N$  Poisson statistics, and binning decisions. They are complex and non-Gaussian!
- Our detection rate drops for fainter galaxies, so we systematically undercount them (“incompleteness”). We need to add corrections to the data to account for this before making the fit.

## Advanced Parameter Estimation

Let's say you have a sample of galaxies in a galaxy cluster, and you've measured all their **absolute magnitudes**. Now want to model the luminosity function of the cluster – the number of galaxies as a function of their **absolute mag**,  $N(M)$ .

Adopt the Schechter function: 
$$N(M)dM = 0.4 \ln 10 \phi_* 10^{-0.4(\alpha+1)(M-M_*)} e^{-10^{-0.4(M-M_*)}} dM$$

But you don't have a huge sample, since you are looking at one specific cluster. How do you estimate  $\alpha$ , and  $L_*$ ?

### Alternative approach:

Make a model of what the luminosity function should look like for a given set of LF parameters. Add the uncertainties and incompleteness to that model, then estimate the likelihood that you would measure the data you have, given that model.

You don't "correct" the data, you alter the model to account for uncertainty and systematic error.

This is an approach that uses **Bayesian statistics**

## Bayesian Estimation

We speak in terms of probabilities. What is the probability you'd get the data you measure given some underlying model?

$$\text{Bayes' theorem: } P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

A = your dataset

B = the parameters you're trying to measure

$P(B|A)$ : **The posterior probability**. What is the probability of B, given that you've measured A? Your best estimate is the B that is most-likely.

$P(A|B)$ : **The likelihood function**. What is the likelihood of measuring A, given that model B is true?

$P(B)$ : **The prior**. What is the probability of B?

$P(A)$ : Normalizing factor. What is the probability you could measure A to begin with? Usually we just set this = 1, since we have measured the dataset!

## Bayesian Estimation

We speak in terms of probabilities. What is the probability you'd get the data you measure given some underlying model?

$$\text{Bayes' theorem: } P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

A = your dataset

B = the parameters you're trying to measure

$P(B|A)$ : **The posterior probability**: the probability that some particular set of  $\alpha$  and  $M_*$  is true, given my dataset.

$P(A|B)$ : **The likelihood function**: the probability of measuring my dataset given some particular value of  $\alpha$  and  $M_*$

$P(B)$ : **The prior**. my prior beliefs about reasonable possibilities for  $\alpha$  and  $M_*$

$P(A)$ : Normalizing factor. What is the probability you could measure A to begin with? Usually we just set this = 1, since we have measured the dataset!

## Bayesian Estimation

Yes, but how do we actually do this? A cartoon sketch:

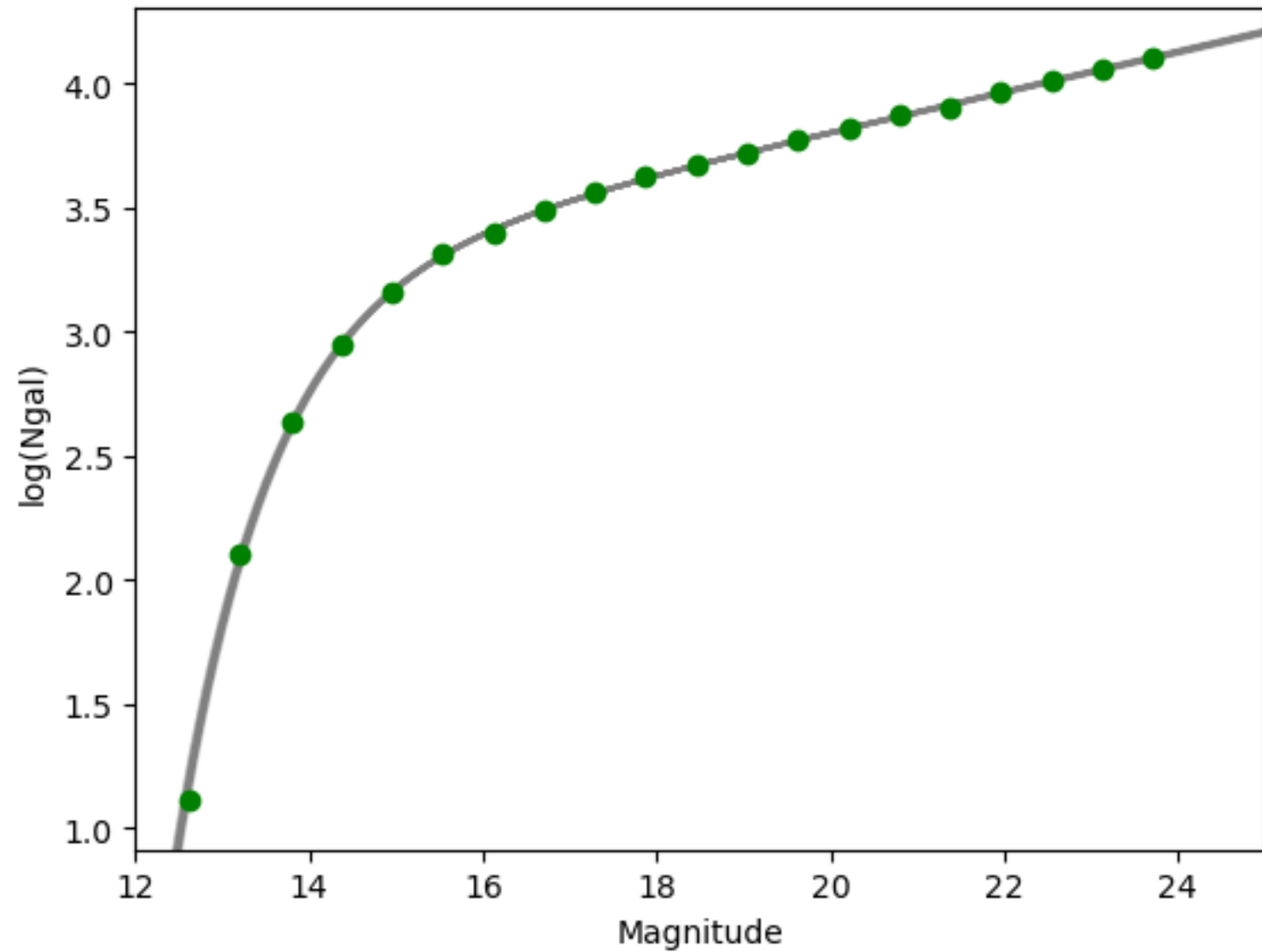
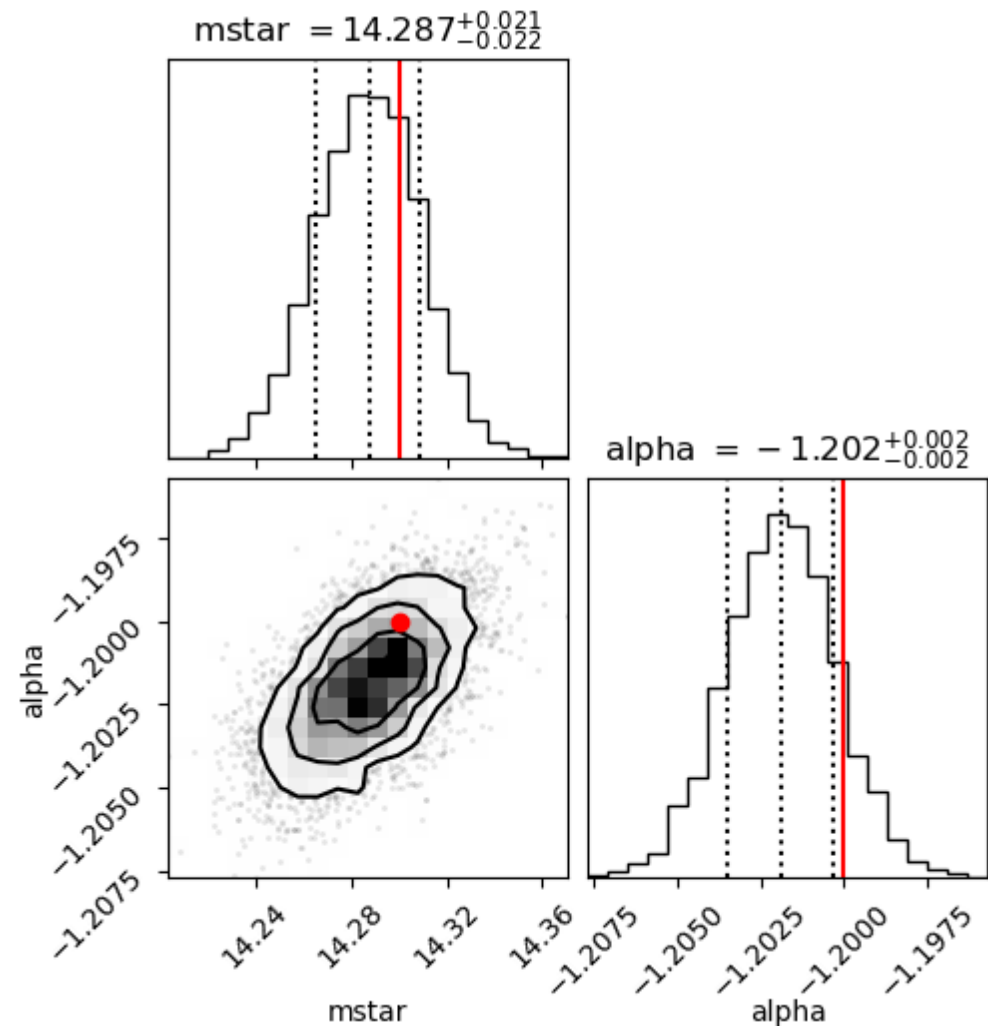
1. **You:** Decide your priors. What are reasonable value ranges (and probabilities) for the parameters of your model ( $\alpha$  and  $M_*$ )?
2. **You:** Start with an initial guess for  $(\alpha, M_*)$  at random, *given your priors*.
3. **Code:** Model the observed luminosity function for those values of  $(\alpha, M_*)$ , including observational uncertainties. Use this model LF to estimate the likelihood of finding galaxies with the absolute magnitudes in my sample.
4. **Code:** tweak the choices of  $(\alpha, M_*)$ , go back to step 3. Do this loop many times.
5. **Code:** Plot the likelihood histograms calculated from the many different parameter tweaks.

*Many different Bayesian codes available online, the one I use is [emcee](#) ([Foreman-Mackey+12](#))*



Example 1: mock sample of  $N=100,000$  galaxies with no incompleteness or uncertainty.

Constructed using  $m_* = 14.3, \alpha = -1.2$

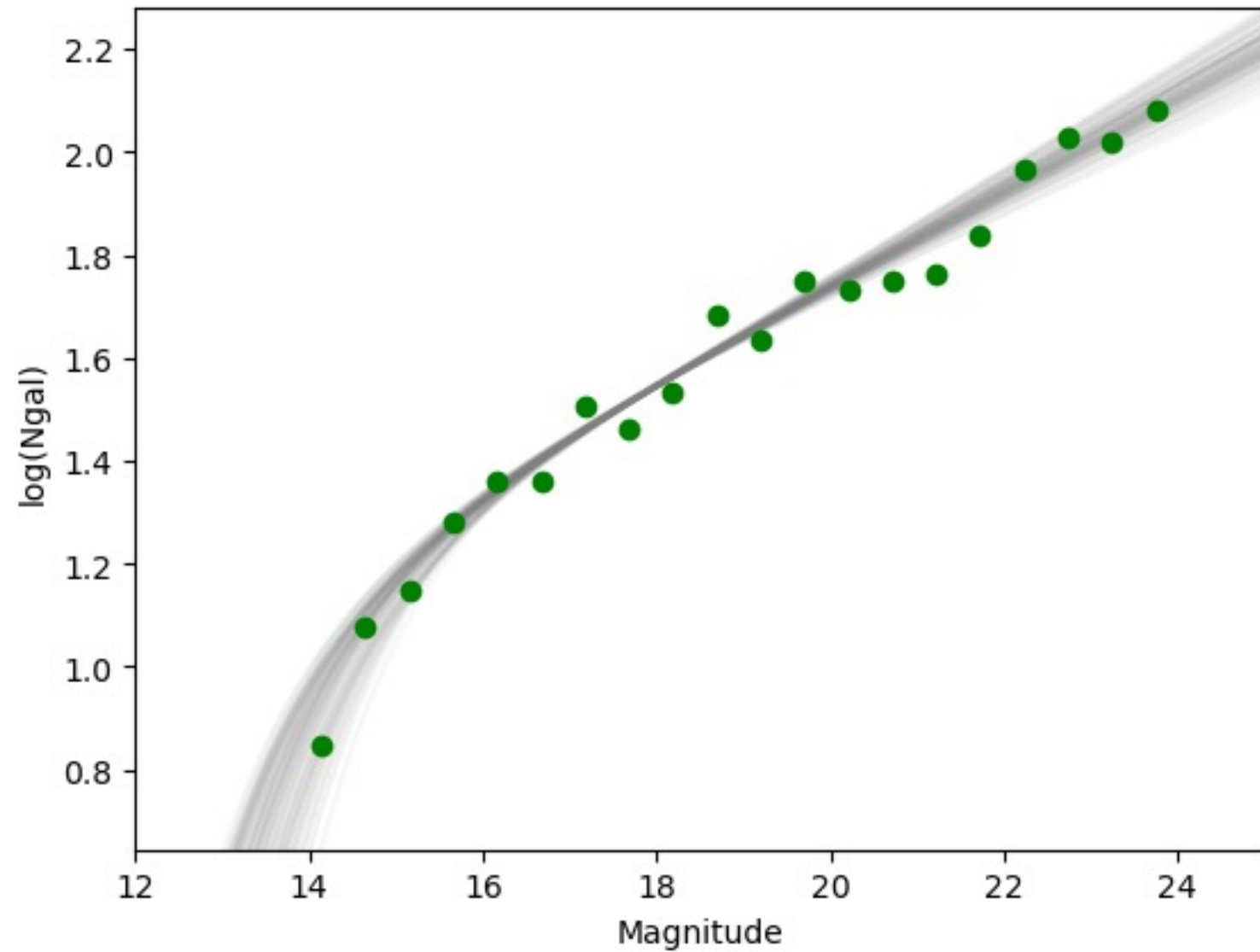
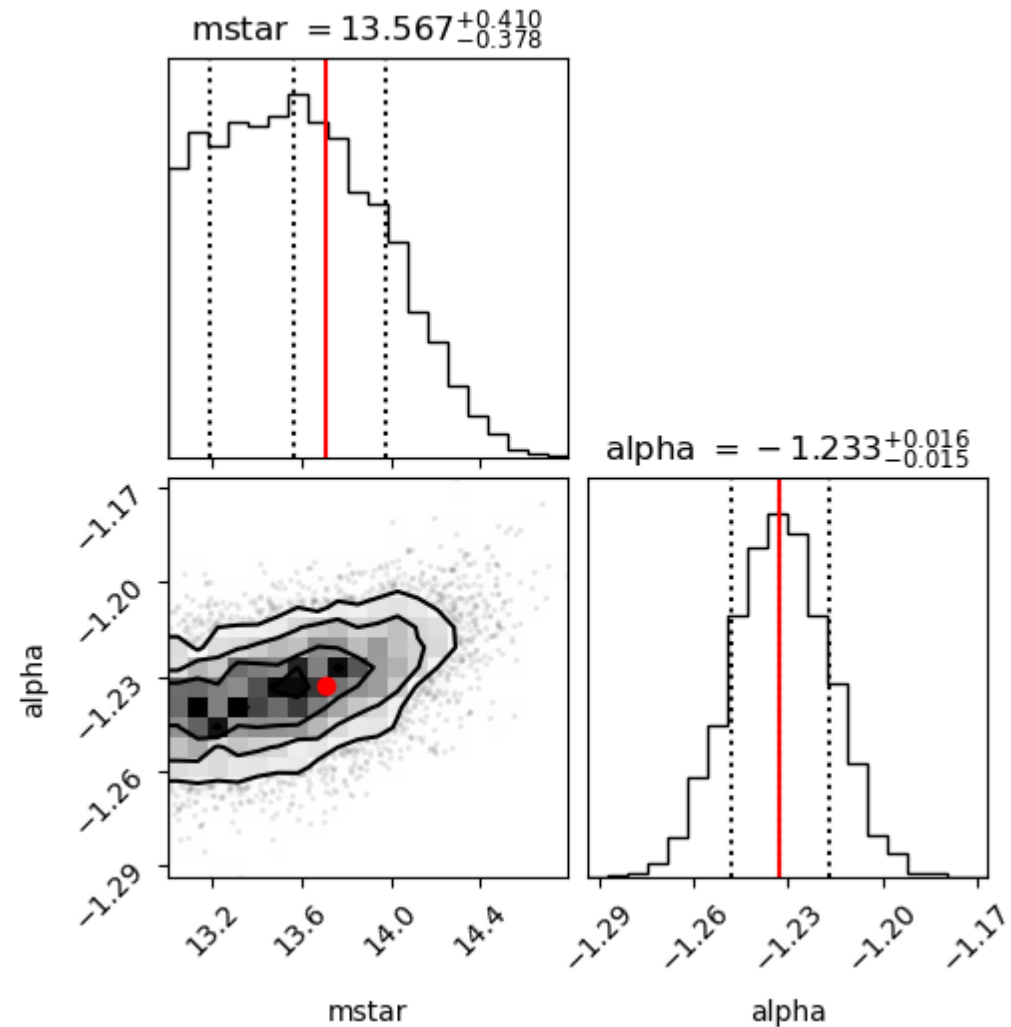


Flat Priors:

- $13 < m_* < 16$
- $-1.5 < \alpha < -0.5$

Example 2: mock sample of  $N=1,000$  galaxies with no observational uncertainty.

Constructed using  $m_* = 14.3, \alpha = -1.2$

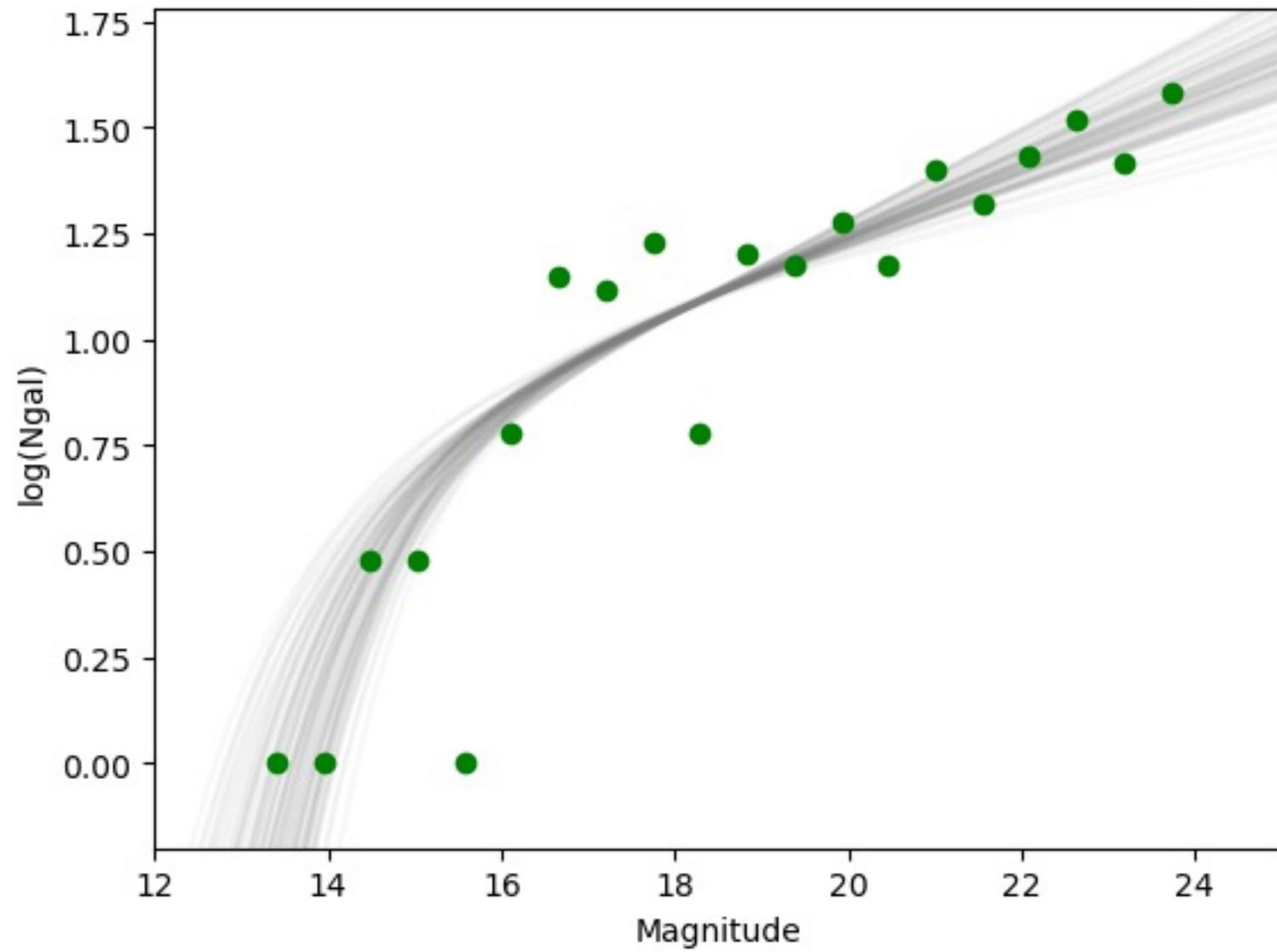
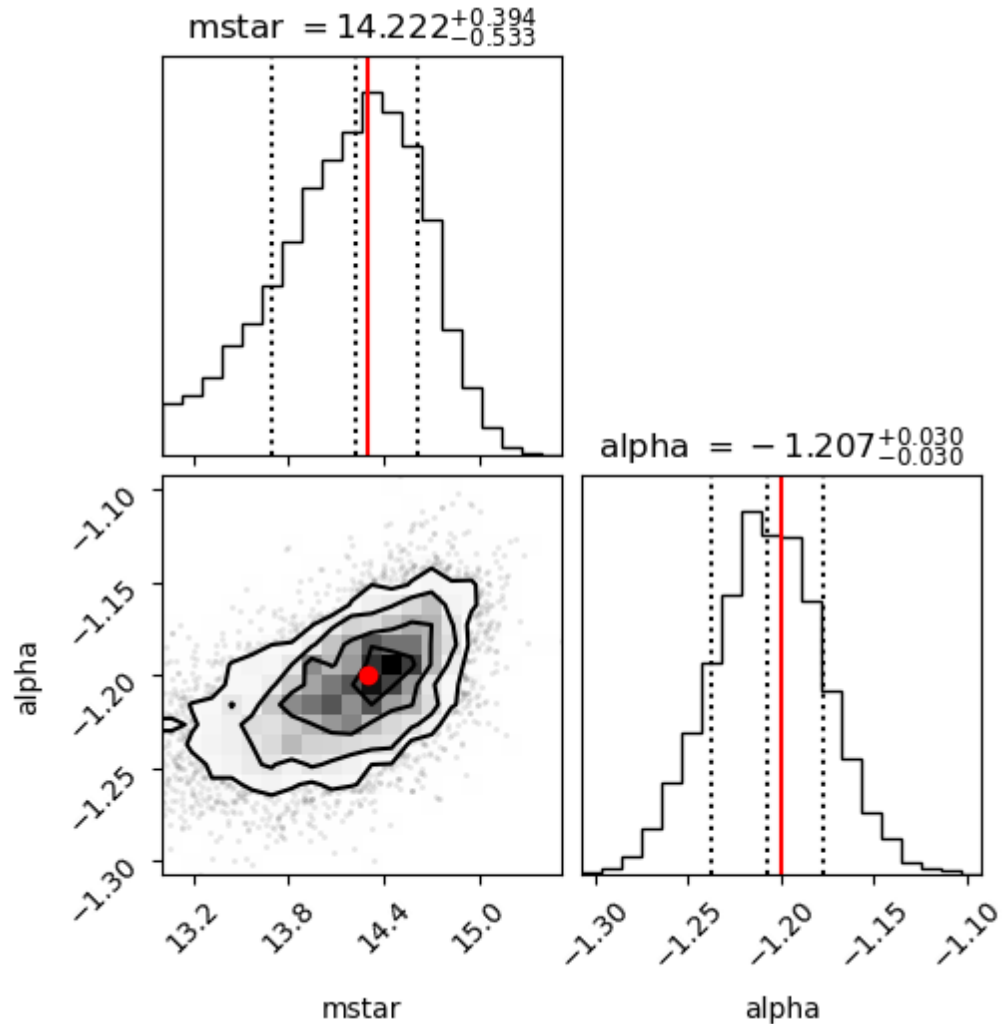


Flat Priors:

- $13 < m_* < 16$
- $-1.5 < \alpha < -0.5$

Example 3: mock sample of  $N=300$  galaxies with no observational uncertainty.

Constructed using  $m_* = 14.3, \alpha = -1.2$

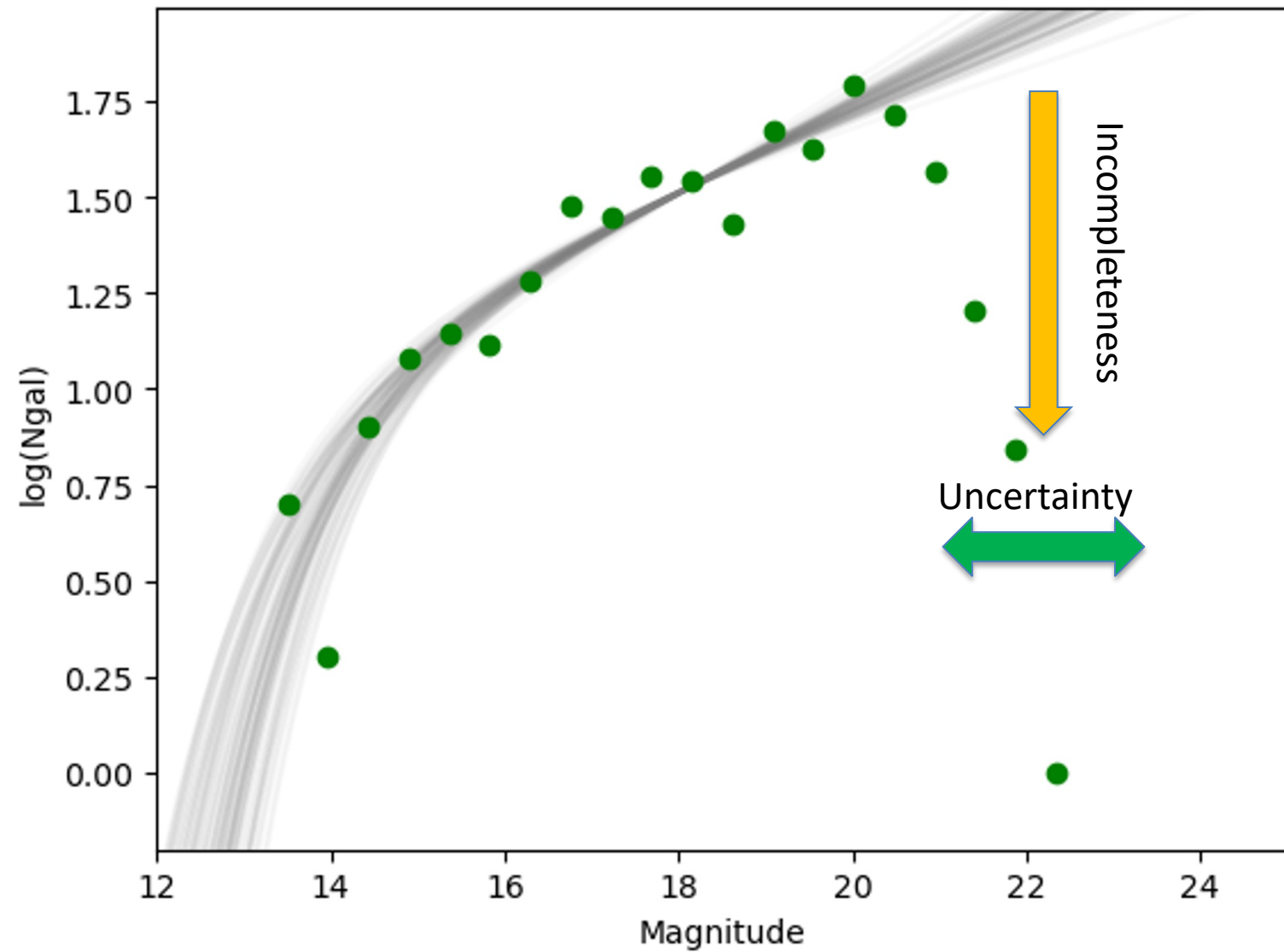
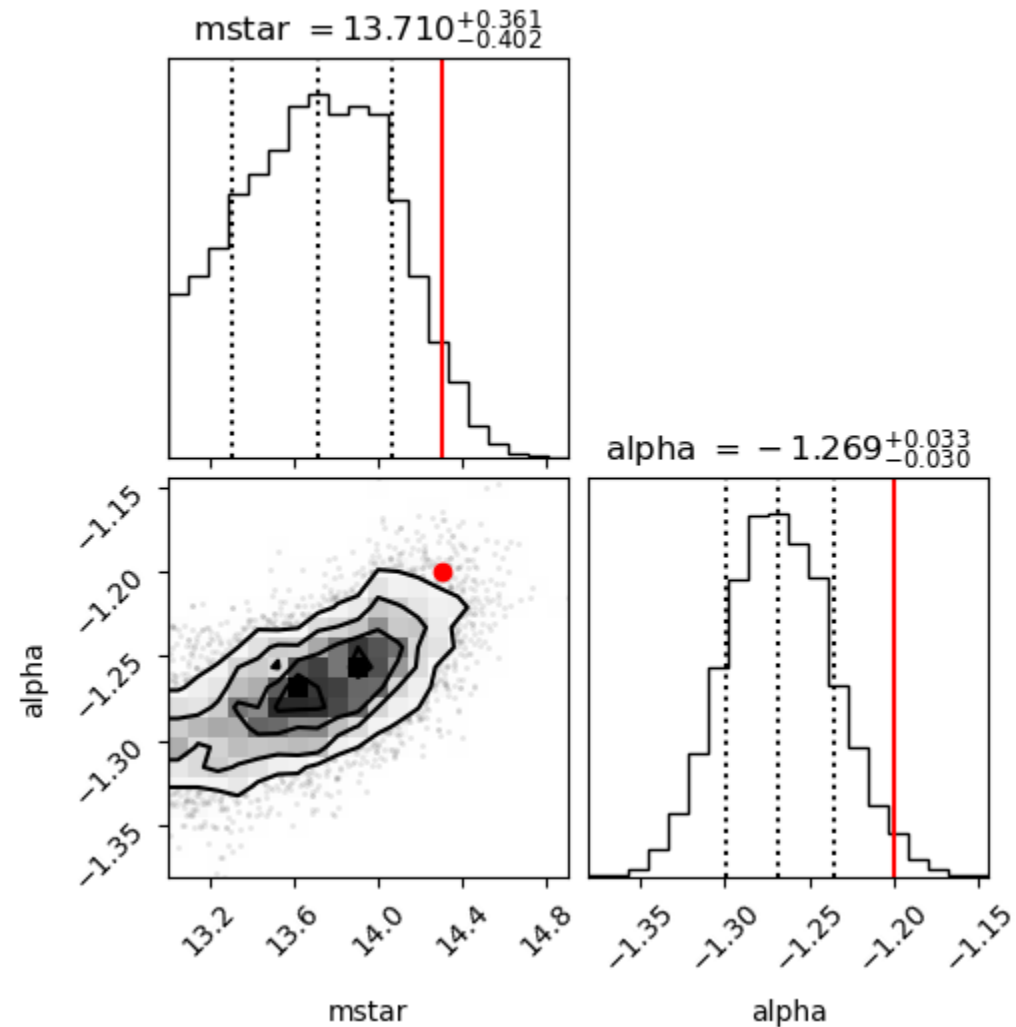


Flat Priors:

- $13 < m_* < 16$
- $-1.5 < \alpha < -0.5$

Example 4: mock sample of  $N=500$  galaxies with incompleteness and photometric uncertainty.

Constructed using  $m_* = 14.3, \alpha = -1.2$

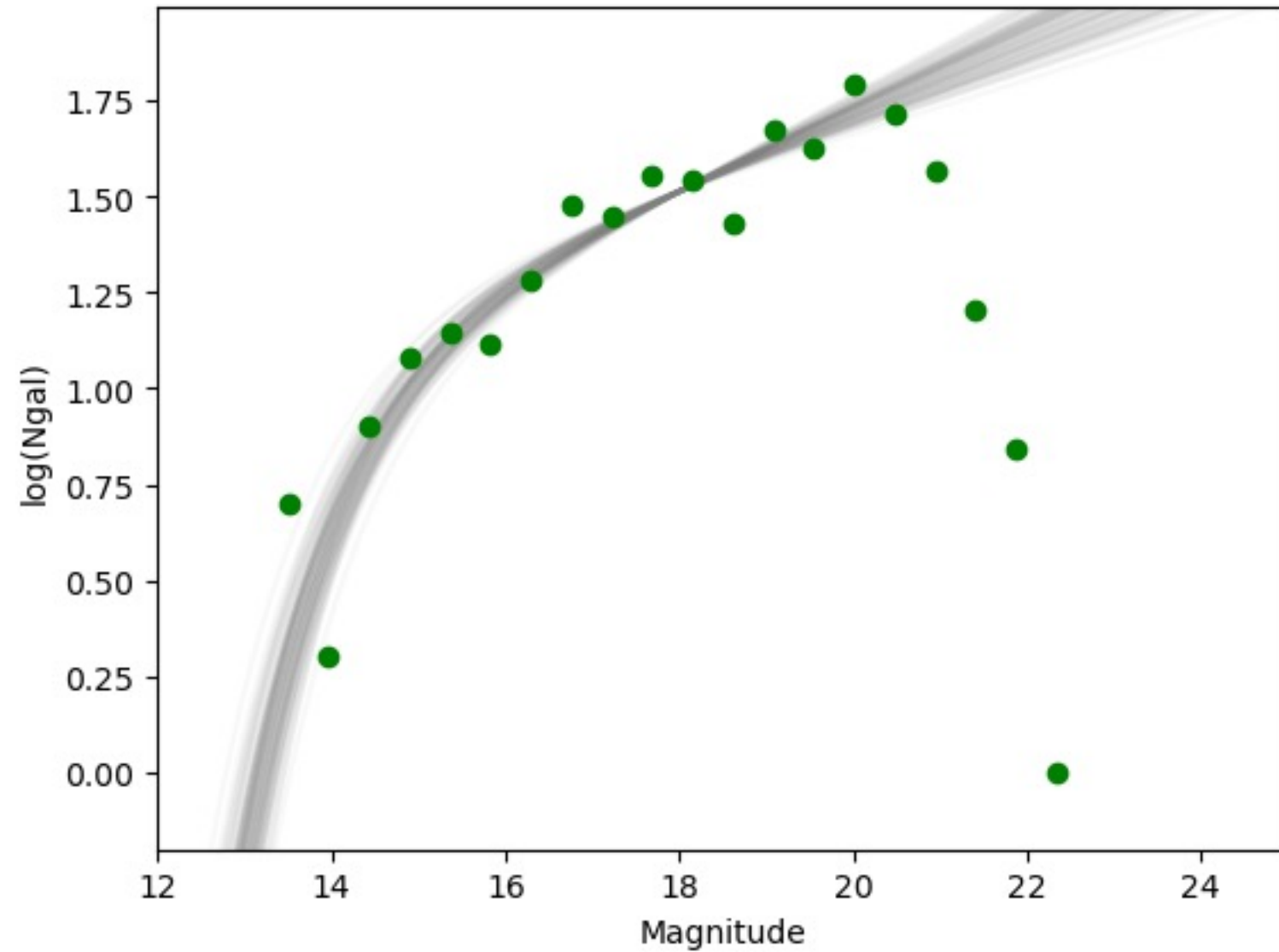
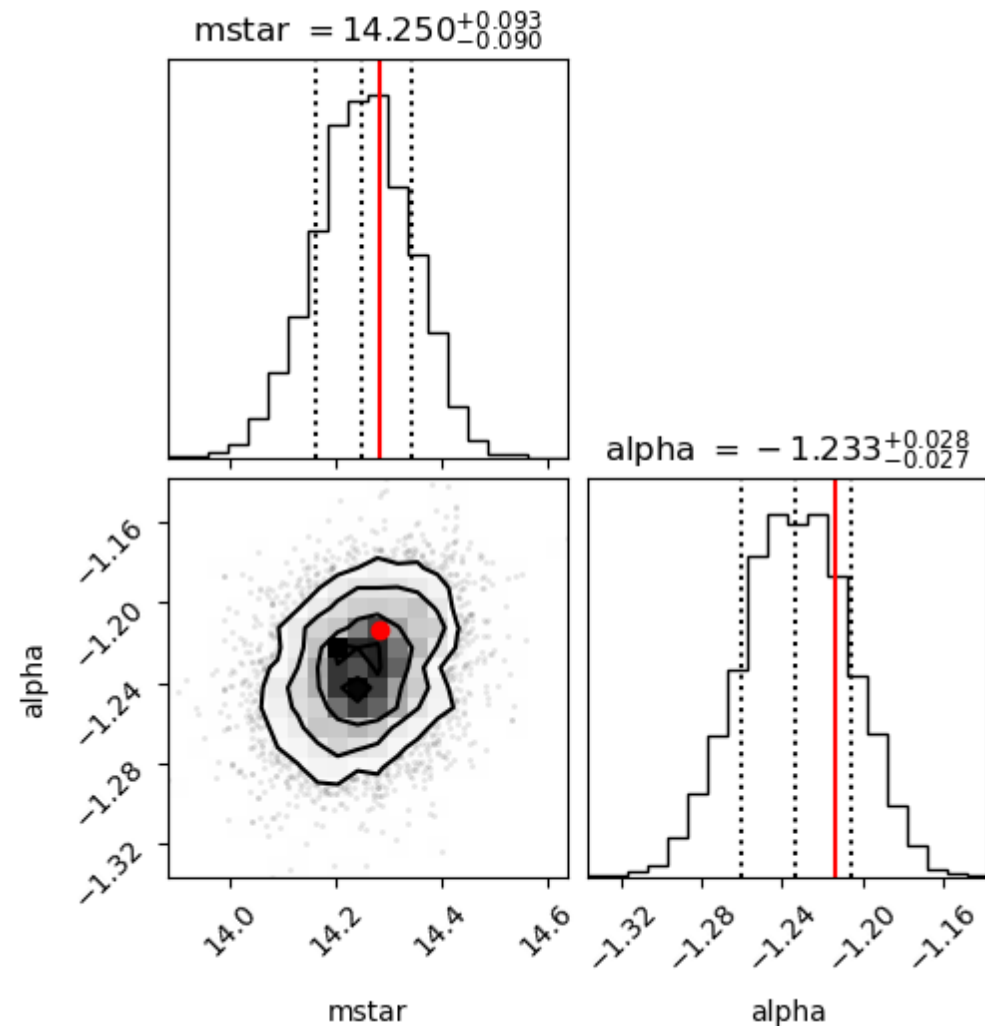


Flat Priors:

- $13 < m_* < 16$
- $-1.5 < \alpha < -0.5$

Example 4: mock sample of  $N=500$  galaxies with incompleteness and photometric uncertainty.

Constructed using  $m_* = 14.3, \alpha = -1.2$



Priors:

- $m_*$ : Gaussian peaked at  $m=14.3$  with  $\sigma=0.1$
- $\alpha$ :  $-1.5 < \alpha < -0.5$

## Citations: Why? When? How?

- To give proper credit to others' work
- To support your arguments and factual claims
- To ground your work in the broader scientific context
- To give a trail for readers to learn more about the topic.
- To demonstrate your awareness and understanding of other work in the field and establish your credentials as a researcher.

failed massive objects and lower-mass dwarf-sized galaxies (e.g., Lim et al. [2018](#), [2020](#); Doppel et al. [2021](#)).

The cluster environment offers additional evolutionary pathways to explain cluster UDGs. One possibility for cluster UDGs is that they started as otherwise normal dwarf galaxies but have been dynamically heated and “puffed up” by interactions within the cluster (Moore et al. [1996](#); Carleton et al. [2019](#); Liao et al. [2019](#); Tremmel et al. [2020](#)) or after gas is ram-pressure-stripped by the intracluster medium (Safarzadeh & Scannapieco [2017](#)). UDGs that are satellites in groups and clusters have also been shown to form by tidal stripping of otherwise normal galaxies either with (Carleton et al. [2019](#)) or without (Sales et al. [2020](#)) cored dark matter halos. However, not all cluster UDGs may have formed in response to the cluster environment; simulations show that a significant fraction of UDGs found in clusters may have been an object “born” as UDGs in the field environment and later accreted into the cluster (Sales et al. [2020](#)).



## Citations: Why? When? How?

- When making general statements about past work on a topic: “Previous studies have shown that galaxies in clusters are preferentially red [citations].”
- When giving factual data that someone else worked out: “The distance to the Virgo Cluster is 16.5 Mpc [citation].”
- When supporting an important claim that you are making: “Tidal stripping should affect low mass galaxies more than massive galaxies [citation(s)].”
- When planning to test a particular model, theory, or result: “Our observations will test tidal stripping models such as those of [citation].”

## Citations: Why? When? How?

Astronomical Literature (*Please use this format in this class!*)

Parenthetical (Author Year) in text, full citation in list at end of document.

- Single author: “Tidal tails are visible for 1-2 Gyr (Mihos 1995)”
- Two authors: “Galaxy mergers trigger starbursts (Mihos & Hernquist 1996)”
- Three or more authors: “The distance to VCC615 is 17.7 Mpc (Mihos et al 2015)”

Citations at end listed in alphabetical order.

- Format: Author(s) Year, Journal, Volume, Page or Article Number
- Example: Mihos et al 2015, ApJ, 638, 17

List multiple citations in chronological order: “The Virgo Cluster contains many diffuse galaxies (Binggeli et al 1996, Mihos et al 2015, Ferrarese et al 2019).”

If your sentence references the work as a grammatical part of the sentence, only the date is parenthetical: “The models of Mihos et al (1995) have shown that....”



## Citations: Why? When? How?

Physics Literature (*Please **do not** use this format in this class!*)

Bracket [Number in end list] in text, full citation in list at end.

- Single author: “Tidal tails are visible for 1-2 Gyr [3]”
- Two authors: “Galaxy mergers trigger starbursts [12]”
- Three or more authors: “The distance to VCC615 is 17.7 Mpc [25]”

Citations at end listed in order of appearance in text.

- Format: Number, Author(s) Year, Journal, Volume, Page or Article Number
- Example: [15] Mihos et al 2015, ApJ, 638, 17

List multiple citations in order of appearance in end list: “The Virgo Cluster contains many diffuse galaxies [12, 17, 32].”

If your sentence references the work specifically, just cite via citation number: “The models of [17] have shown that....”

## Citations: Why? When? How?

Other notes:

- Direct quoting is extraordinarily rare in the sciences. Do not quote your sources, just summarize their findings using your own words, and give citation to their work.
- Citations should be to journal articles, technical documents, preprints, or works published in scientific monographs (books).
- Data aggregator websites (including NED or Simbad) are typically not appropriate as cited source. If you get data using NED, look at the citation NED gives for the data, and cite that source directly.
- For examples of how citations are used, look at the (good) proposal example and the journal articles you find.