Statistics and Modeling

Statistics is the grammar of science.
- Karl Pearson

There are three types of lies -- lies, damn lies, and statistics.
- Benjamin Disraeli? Mark Twain?

It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so.
- Mark Twain? Yogi Berra?

All models are wrong, but some models are useful.
- George E.P. Box

Data do not give up their secrets easily. They must be tortured to confess.
- Jeff Hopper, Bell Labs

Random vs Systematic Error

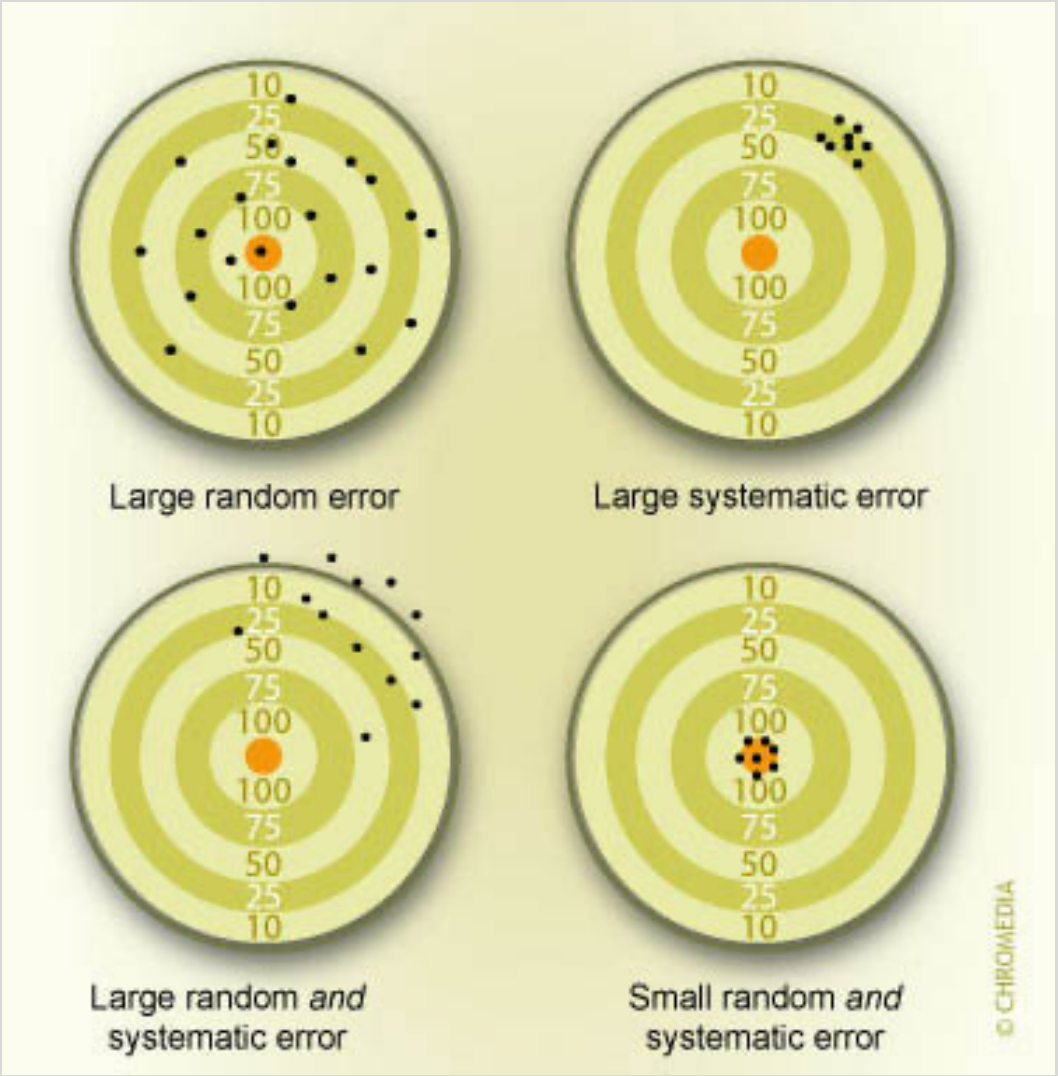
Precision: How well can you measure a quantity? How repeatable is your measurement? Usually captured by “random errors.”

Accuracy: How well does your measurement actually recover the value you are trying to measure? Source of “systematic errors.”

Random vs Systematic is critical to understand, extremely hard to quantify in practice.

If you measure a value and do not give some estimate of uncertainty or some discussion of systematic errors, your measurement is nearly useless.

The 10/90 rule: you spend 10% of your time getting “the answer”. You spend the other 90% understanding your uncertainties.




SYSTEMATIC ERROR
Random Error
Random Error
Random Error
Random Error
Random Error
Random Error
Random Error
daytonight

Characterizing distributions

- Moments
- 1st: Mean, $\bar{x}$ (location)
  - Other 1st-moment indicators:
    - *median* (robust estimator)
    - *mode*
- 2nd: Standard deviation, σ (width)
  - Other 2nd-moment indicators:
    - *Average deviation* (robust estimator): $AD = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$
    - *full-width half-maximum (FWHM)*
- 3rd: Skew, s (symmetry)
- 4th: Kurtosis, k (shape)

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

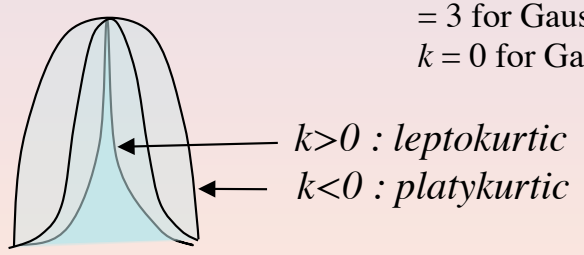
$$s = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3$$

$$k = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{\sigma}\right)^4 - 3$$

= 3 for Gaussian
k = 0 for Gaussian

1st & 2nd moments

skew



kurtosis

courtesy
Matt Bershady,
University of Wisconsin

Error Propagation

If errors are **gaussian** and **uncorrelated**, we can add each error source in quadrature.
(But uncorrelated gaussian errors are often a bad assumption!)

if each measurement of X has an uncertainty σ, the error in mean is given by $\sigma_{\bar{x}} = \sigma/\sqrt{N}$

For propagating small errors, we can use a **Taylor expansion**. If you are calculating some property C from measurements of x, y, and z:

$$C = f(x, y, z) \quad \Longrightarrow \quad \sigma_C^2 = \left[\left(\frac{\partial f}{dx}\right)\sigma_x\right]^2 + \left[\left(\frac{\partial f}{dy}\right)\sigma_y\right]^2 + \left[\left(\frac{\partial f}{dz}\right)\sigma_z\right]^2$$

Example, working out the absolute magnitude of M87: apparent mag (m) = 8.63 ± 0.04, distance (D)=16.0 ± 1.1 Mpc

- $M = -5\log d + 5 + m = -22.4$
- $\partial M/dm = 1$
- $\partial M/dD = -5/(D\ln 10) = -2.17/D$
- $\sigma_M^2 = [1\times 0.04]^2 + [(-2.17/16)\times 1.1]^2$
- so $\sigma_M = 0.15$ mag

but this characterizes random error, not systematic error!
it also assumes gaussian and uncorrelated errors!

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: $\sigma_{RMS}^2 \equiv \frac{1}{N}\sum\left(y_i - y_{fit}(x_i)\right)^2$

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

The importance of scatter

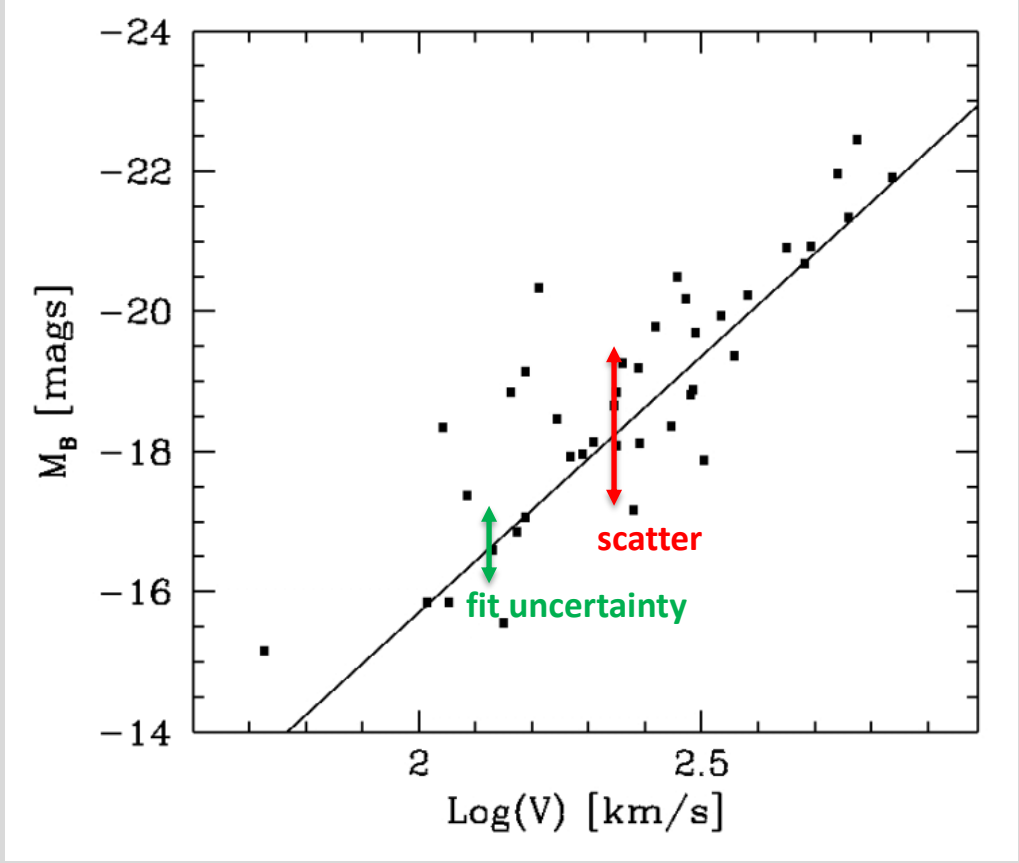
The uncertainties on the fit tell you how well-determined the fit parameters are.

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

Example: Tully Fisher Relationship ⇒

Lower fit uncertainties (σ_m, σ_b) mean that the overall TF relationship is better-determined.

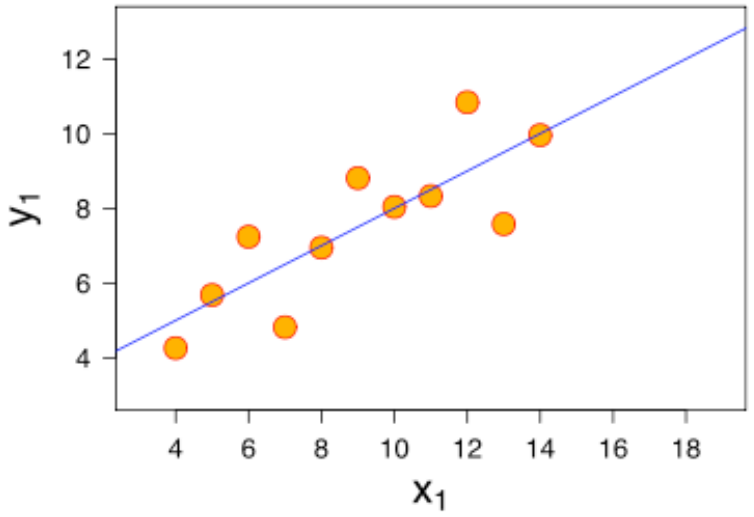
Large scatter (σ_{RMS}) means any one galaxy may not perfectly obey TF.



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

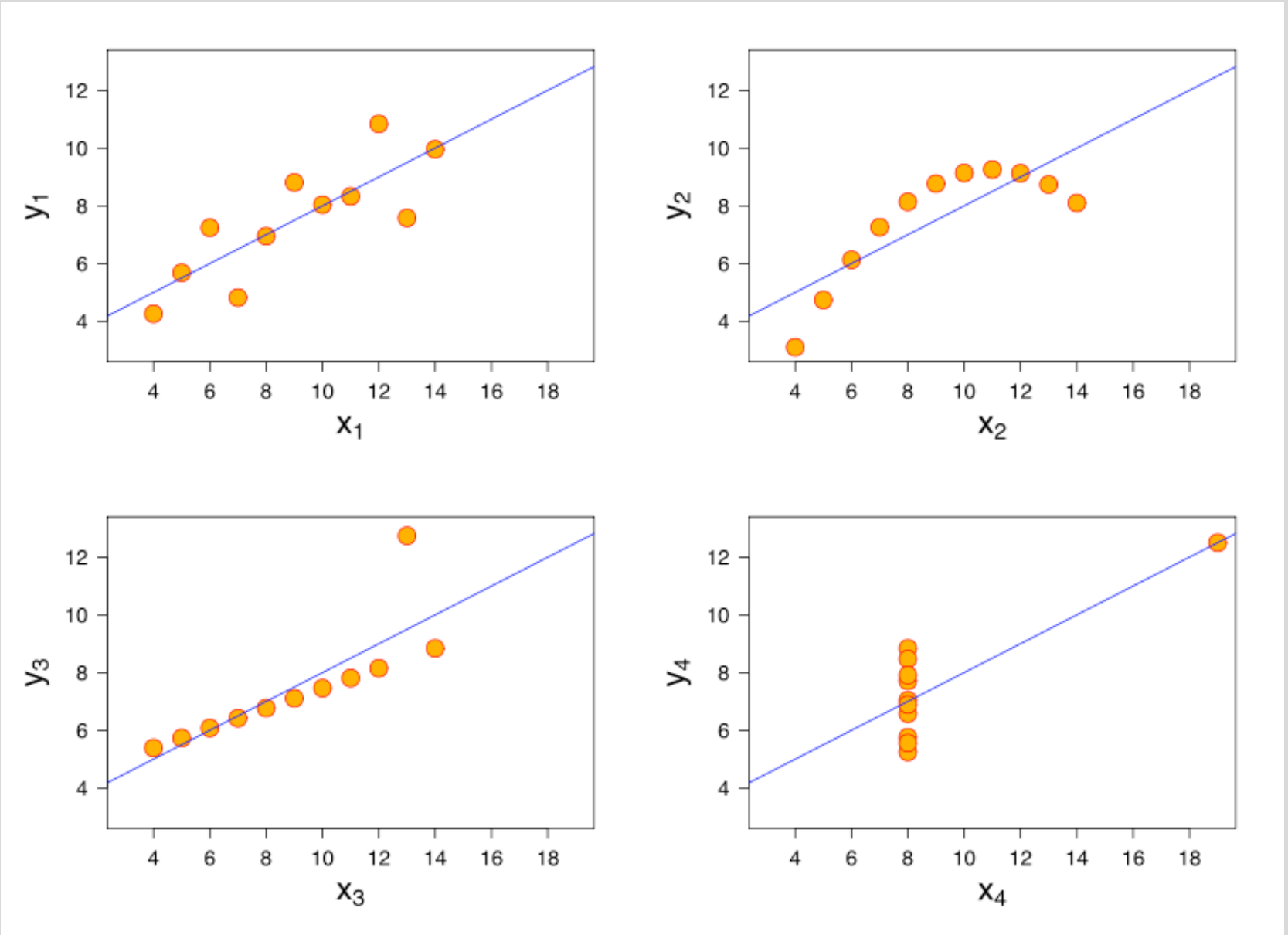
```
# make a linear fit, and calculate uncertainty and scatter

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...



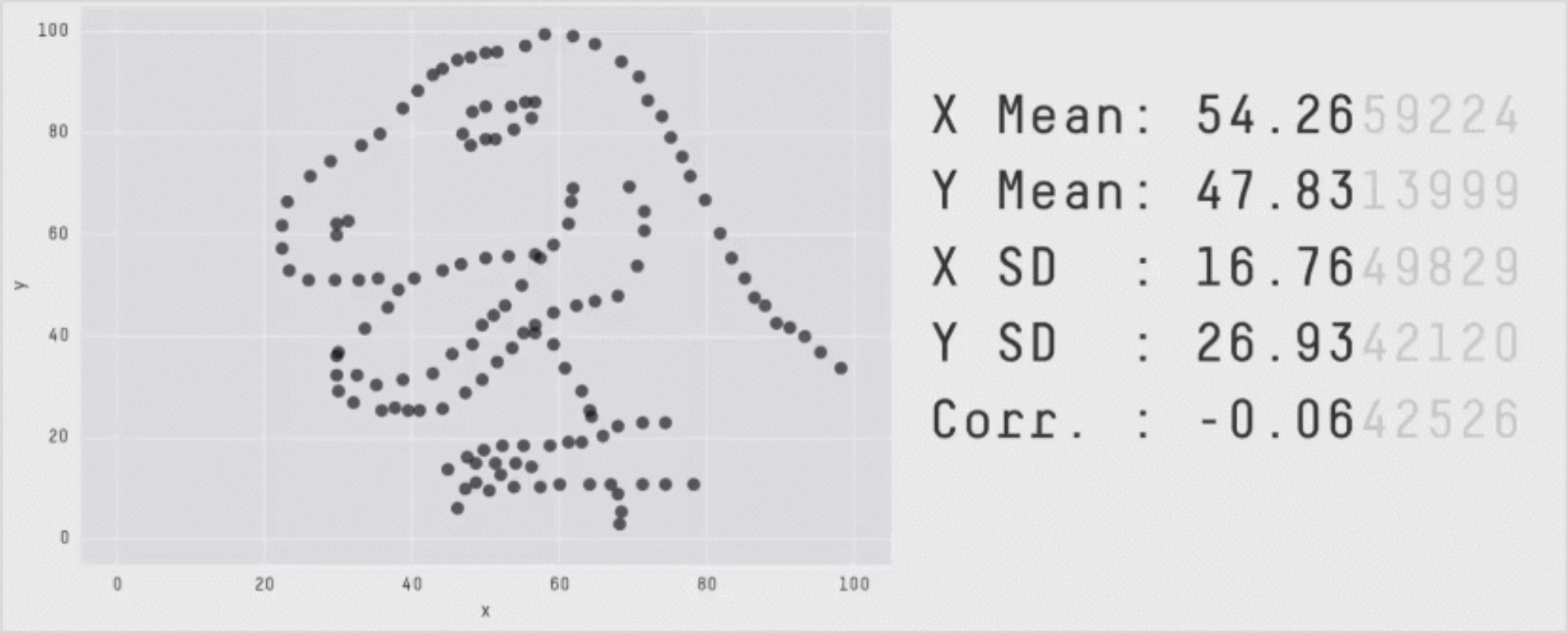
Anscombe's quartet: Fit y=mx+b and get the same r (correlation coefficient), m, b, σ_m, σ_b, σ_{RMS}

But be careful with fits...



Anscombe's quartet: Fit y=mx+b and get the same r (correlation coefficient), m, b, σ_m, σ_b, σ_{RMS}

Beware the datasaurus!



Moral of the story: ALWAYS PLOT YOUR DATA!

Modeling Uncertainty

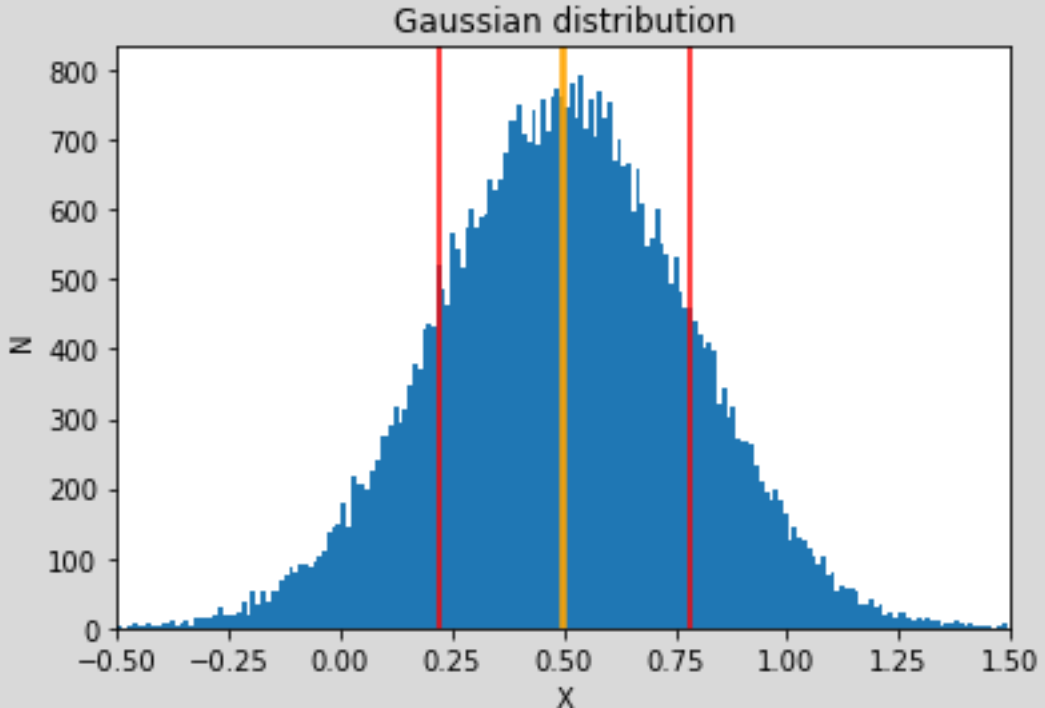
Let's say you have a repeated measurements of some value. How do we estimate the best value and uncertainty.

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 +/- 1σ)



Modeling Uncertainty

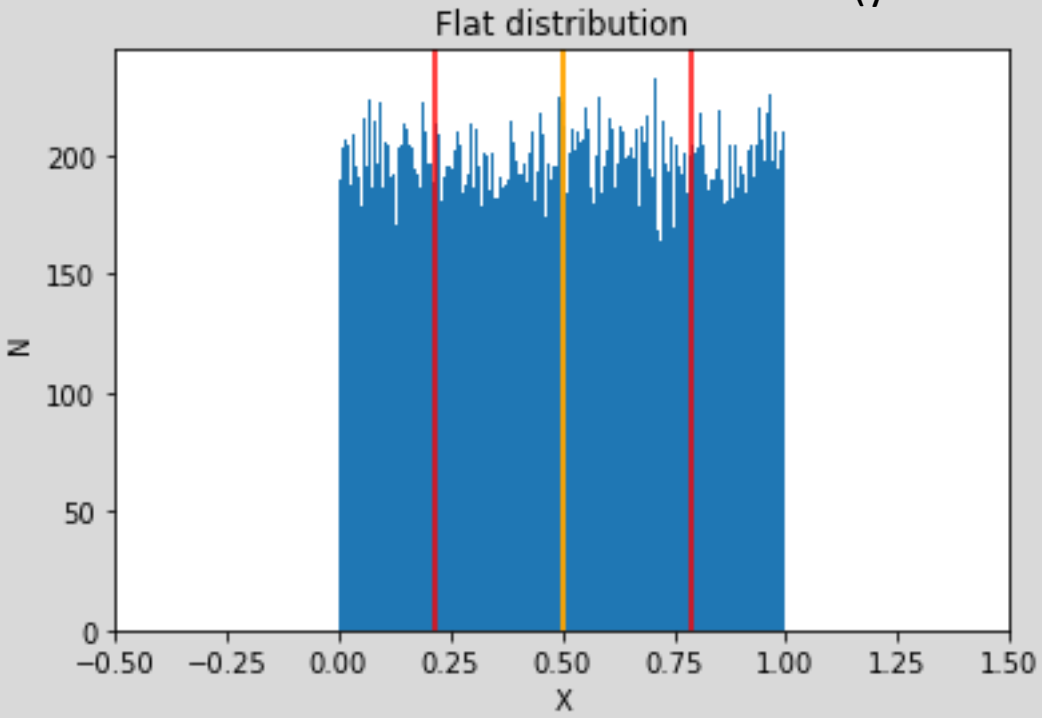
Let's say you have a repeated measurements of some value. How do we estimate the best value and uncertainty.

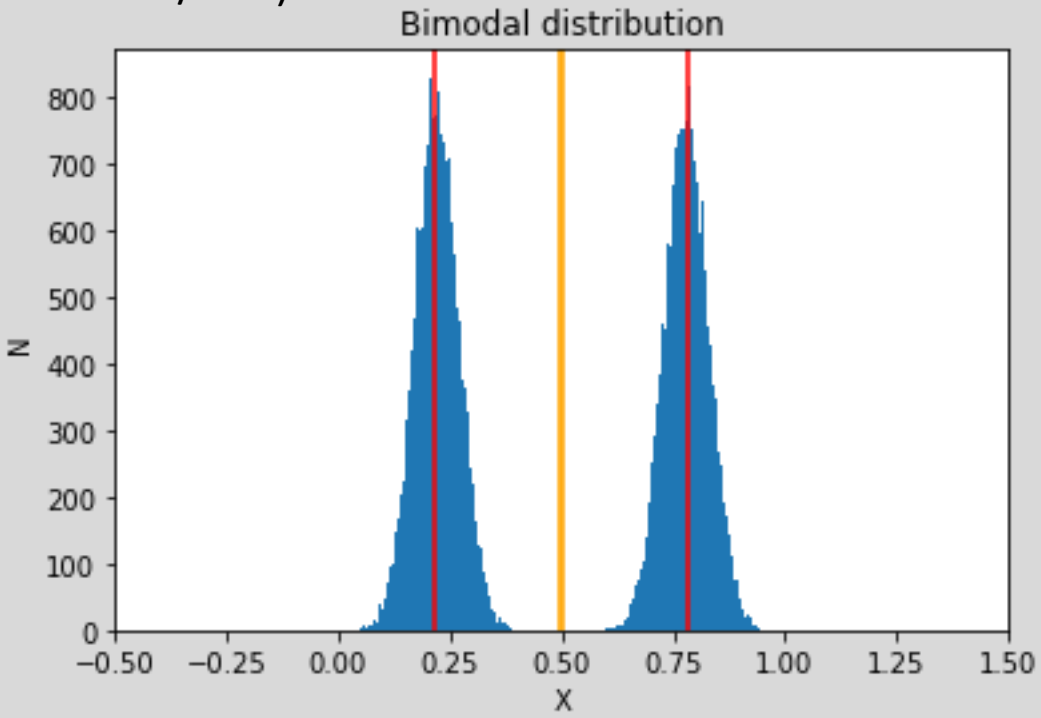
If your errors are independent and follow a Gaussian distribution:

- 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 +/- 1σ)





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 $\sigma/\sqrt{N}$

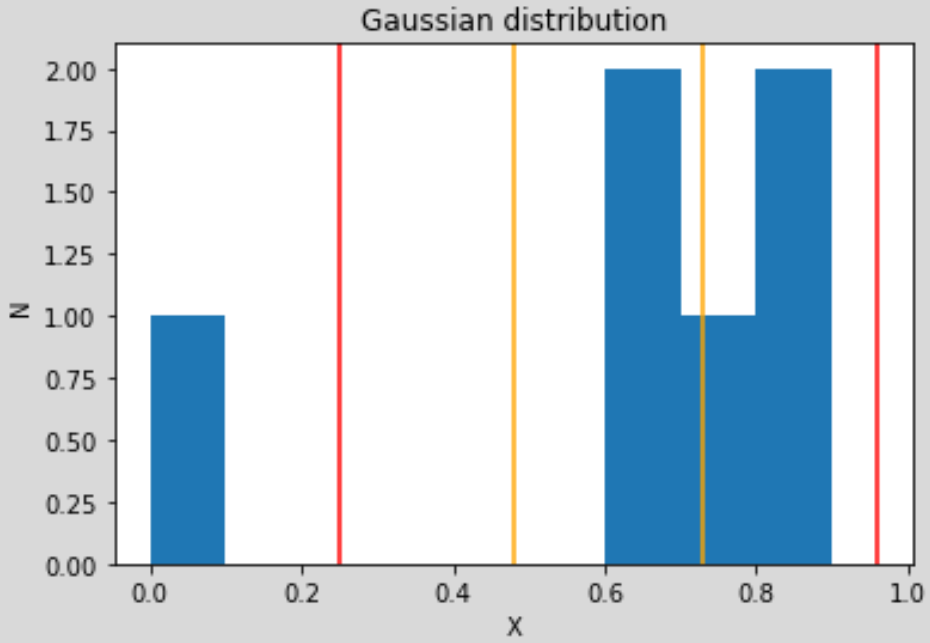
What about small N? Take a distribution of 8 measurements with $\bar{x}, \sigma$ = 0.5, 0.28, look at distribution.

(yellow: mean +/- std err, red: mean +/- 1σ)

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).

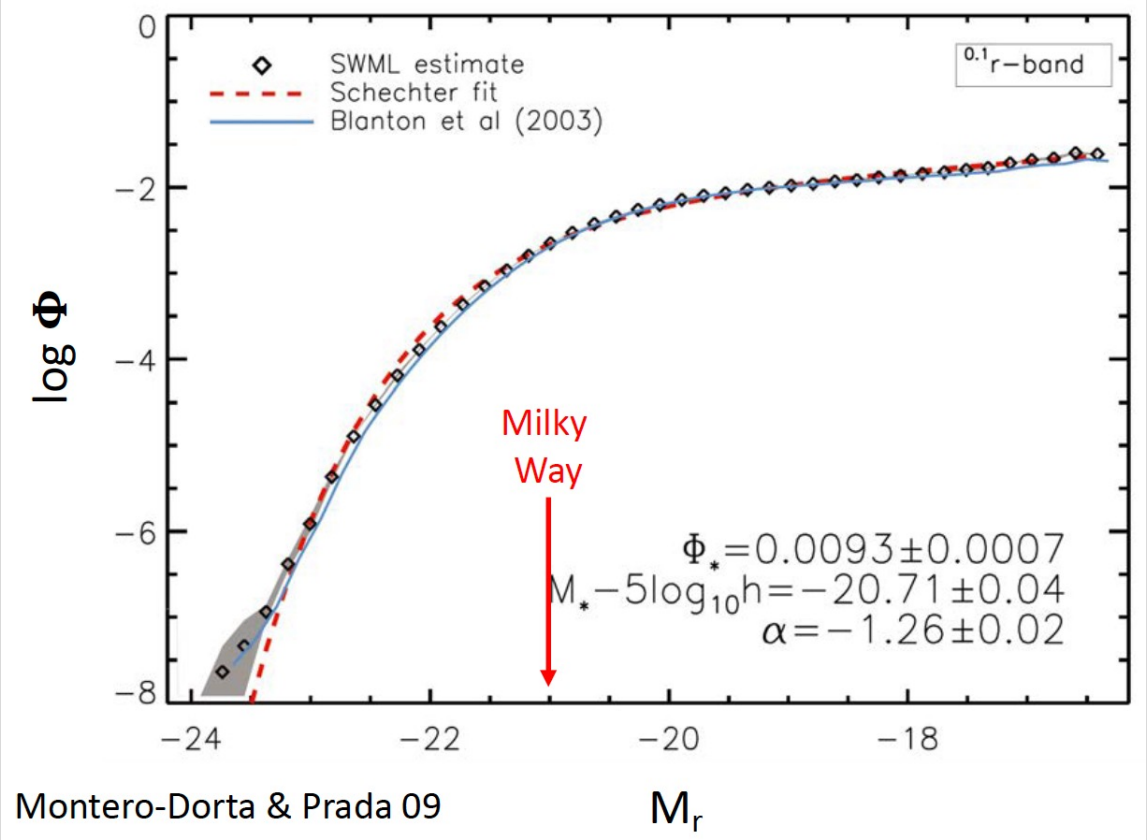
Adopt the Schecter function: $N(L) = \Phi_0 L^{\alpha} e^{-L/L_*}$

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 Schecter 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 Schecter 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 α, 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 Schecter 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 α, 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 dont "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 α and M_* is true, given my dataset.

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

P(B): ***The prior***. my prior beliefs about reasonable possibilities for α 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 (α and M_*)?

2. **You**: Start with an initial guess for (α, M_*) at random, *given your priors*.

3. **Code**: Model the observed luminosity function for those values of (α, M_*), including observational uncertainties. Use this model LF to estimate the likelihood of finding galaxies with the absolute magnitudes in my sample.

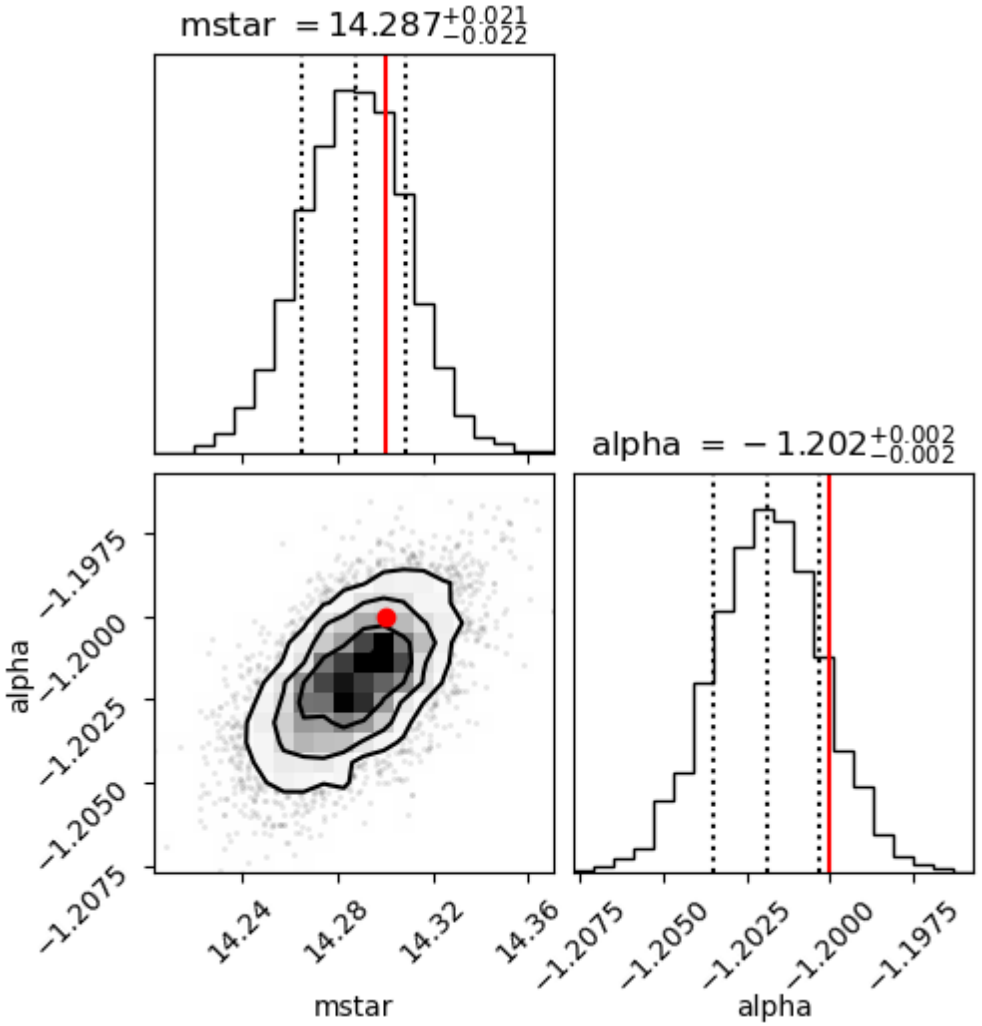
4. **Code**: tweak the choices of (α, M_*), go back to step 3. Do this loop many times.

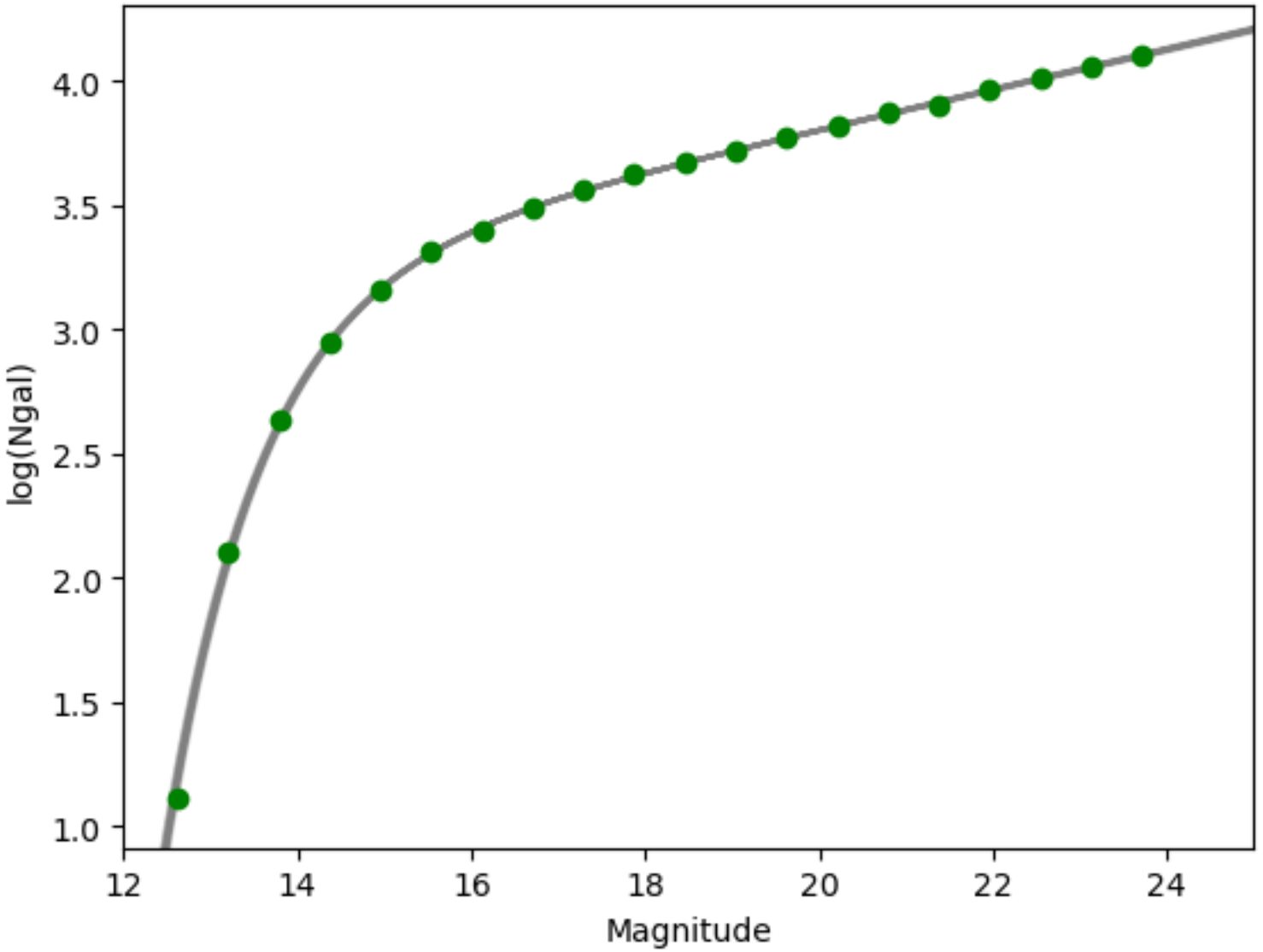
5. **Code**: Plot the likelihood histograms calculated from the many different parameter tweaks.

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



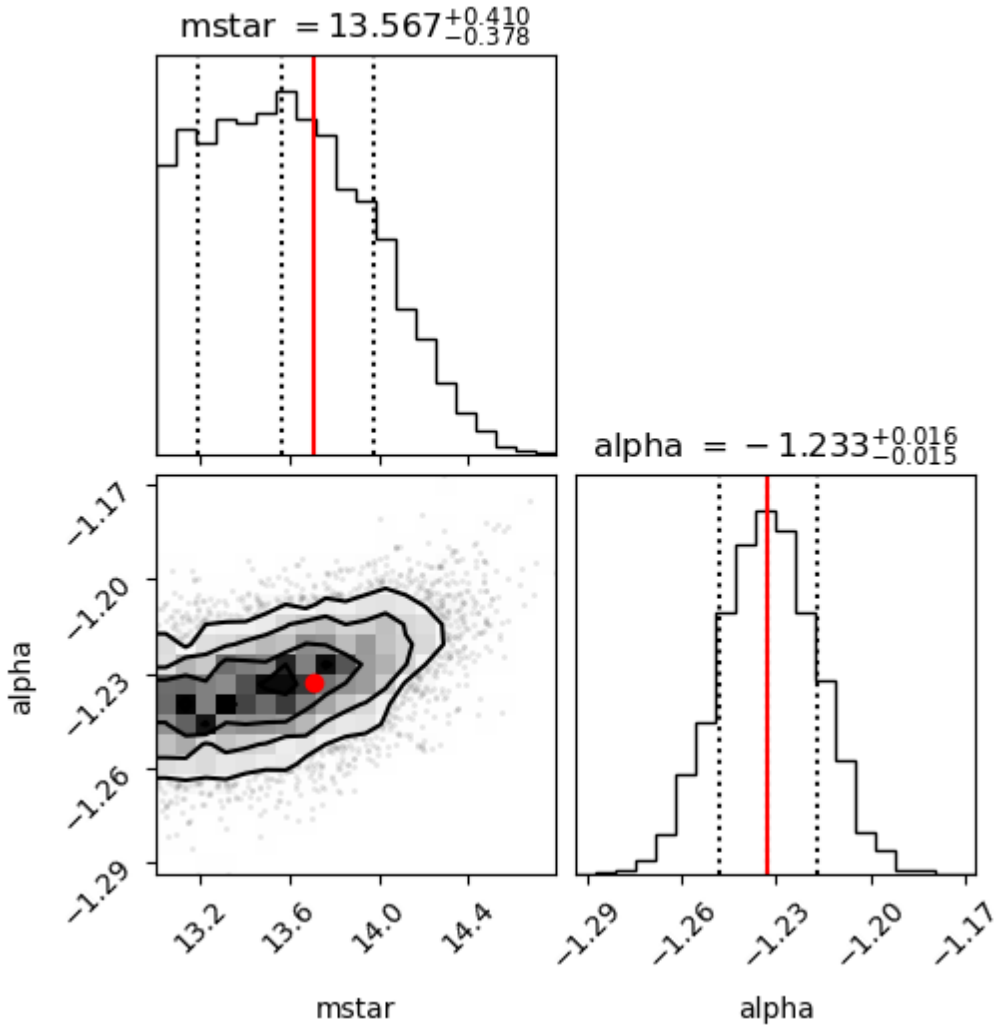


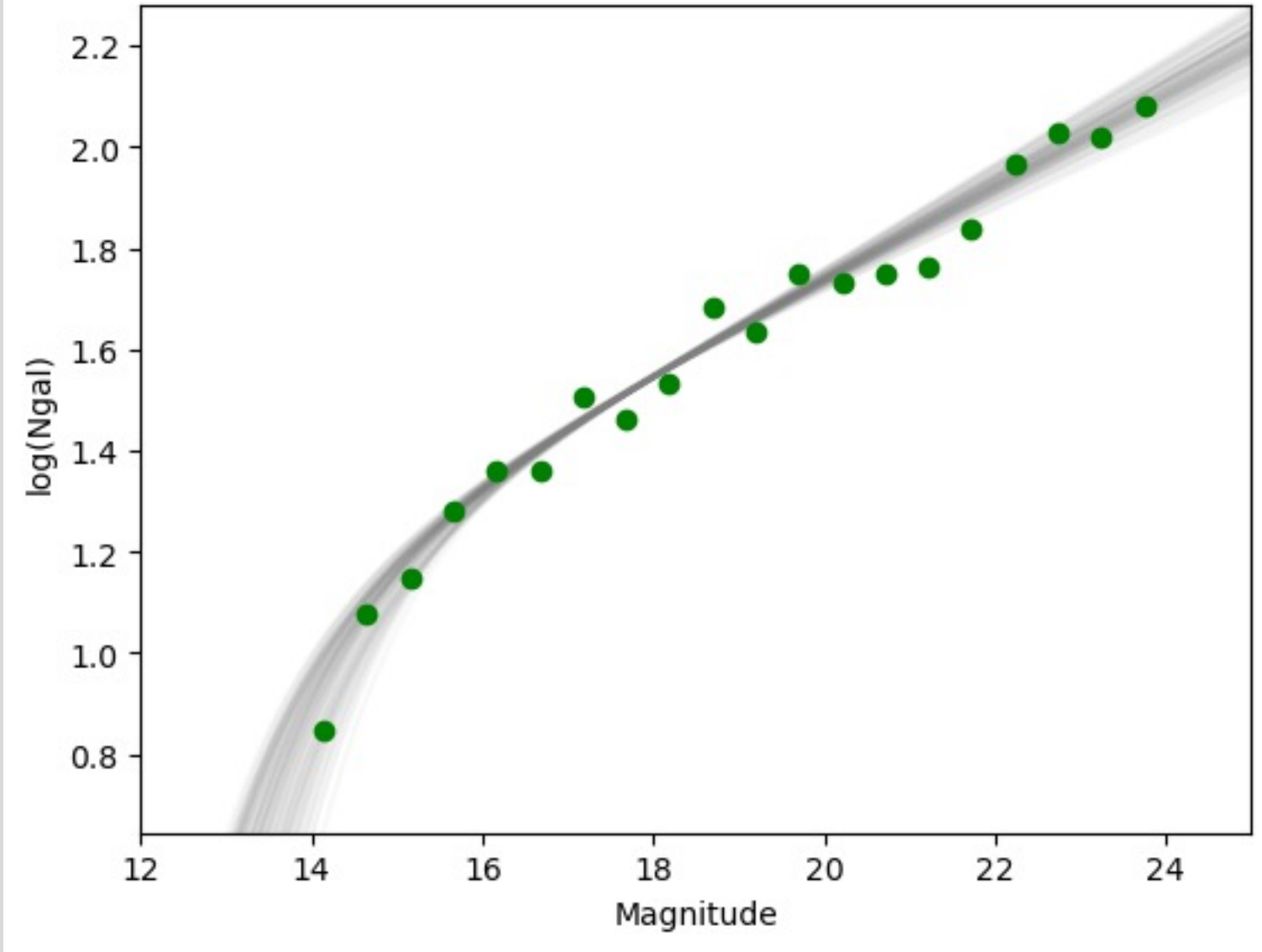
Flat Priors:

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

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

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



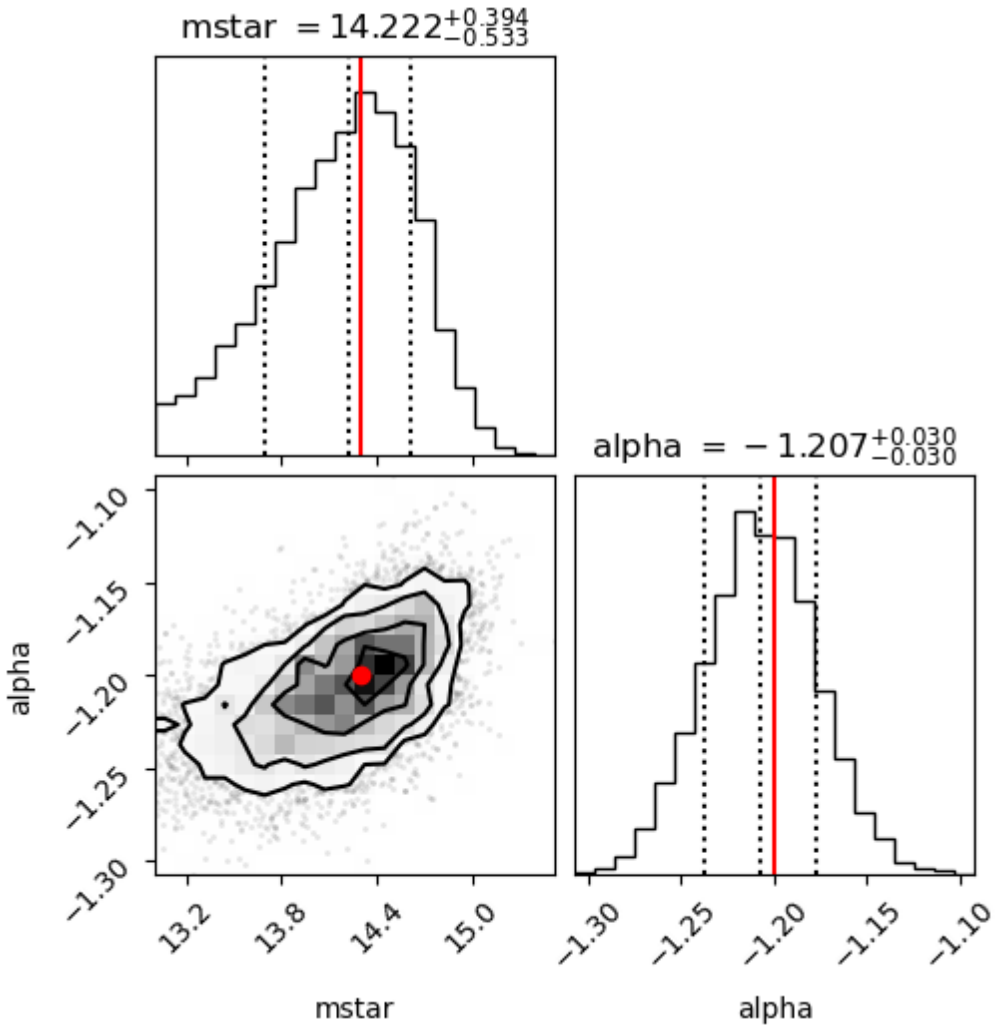


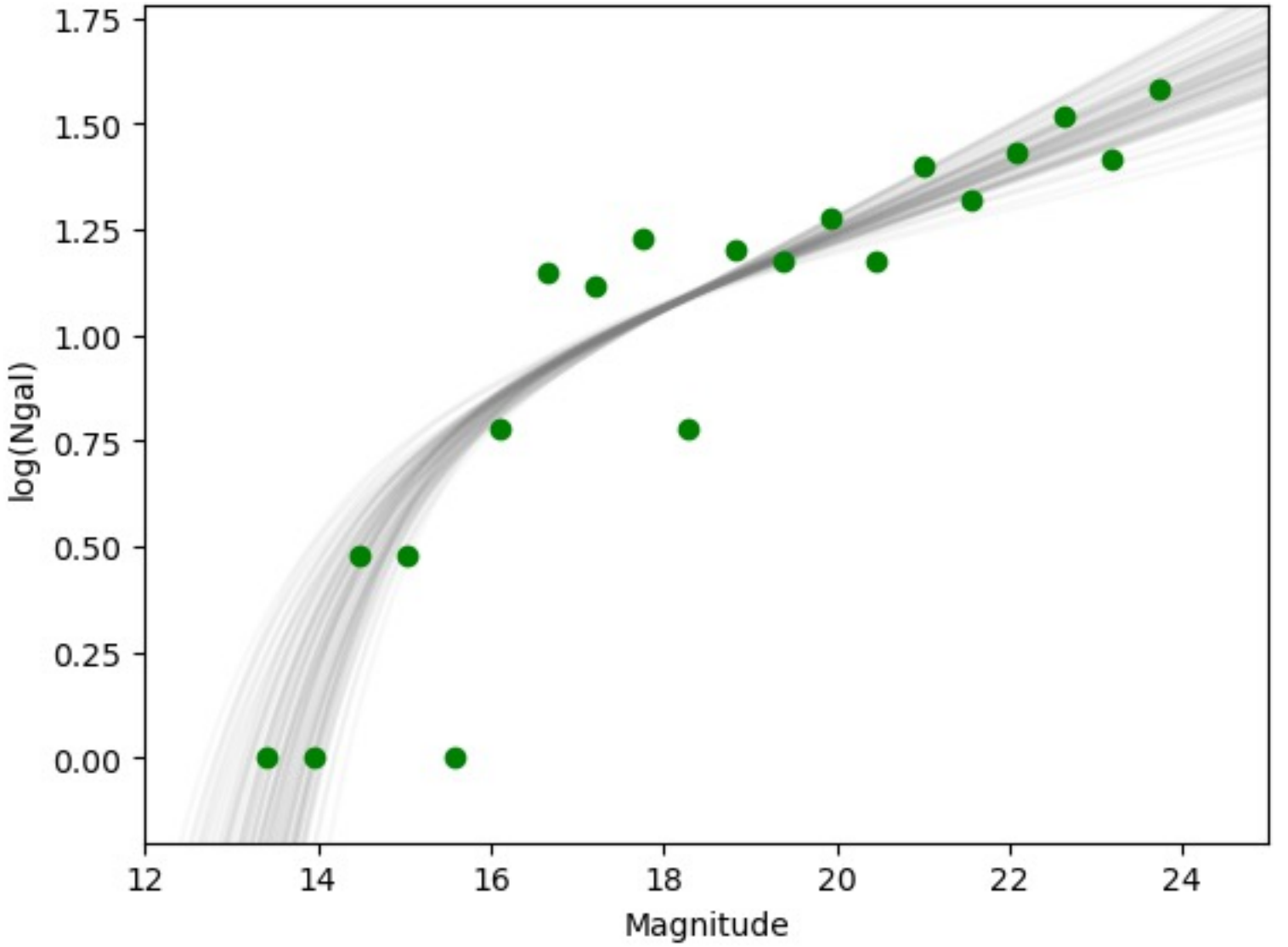
Flat Priors:

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

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

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



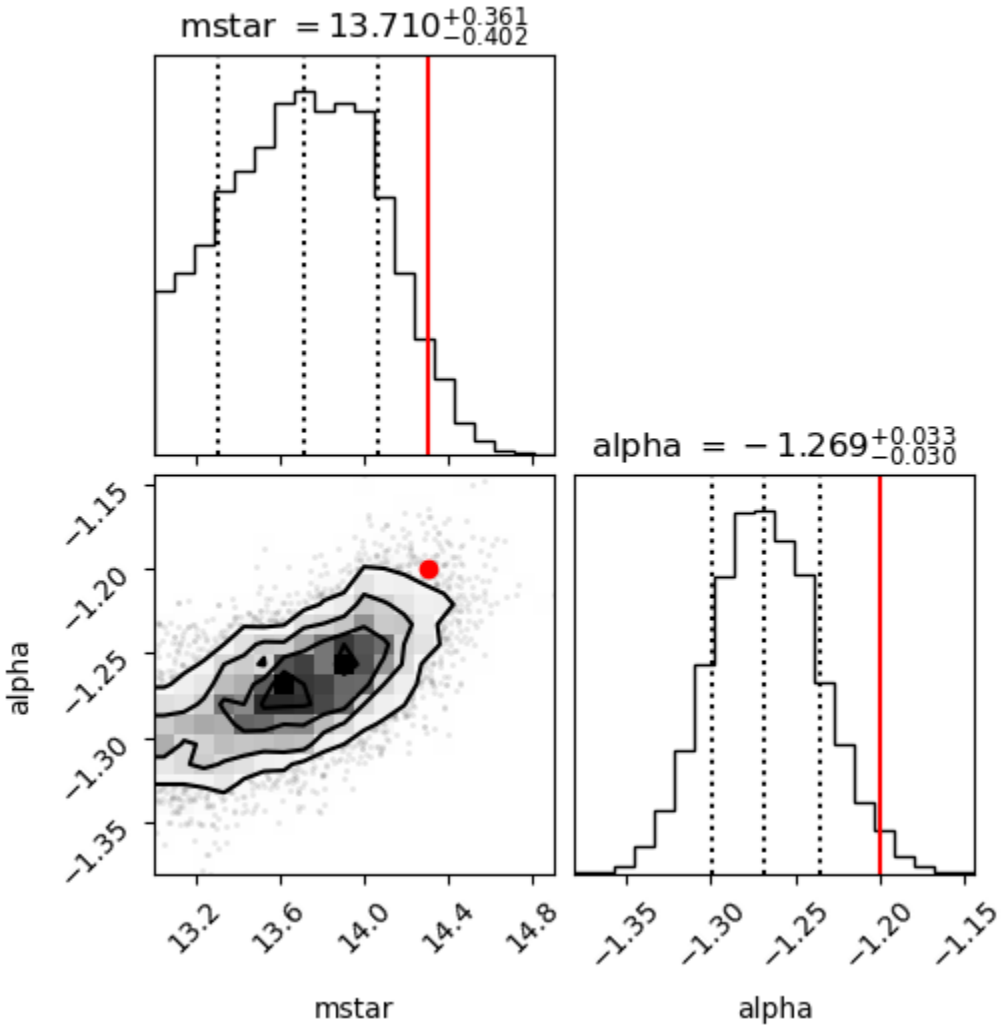


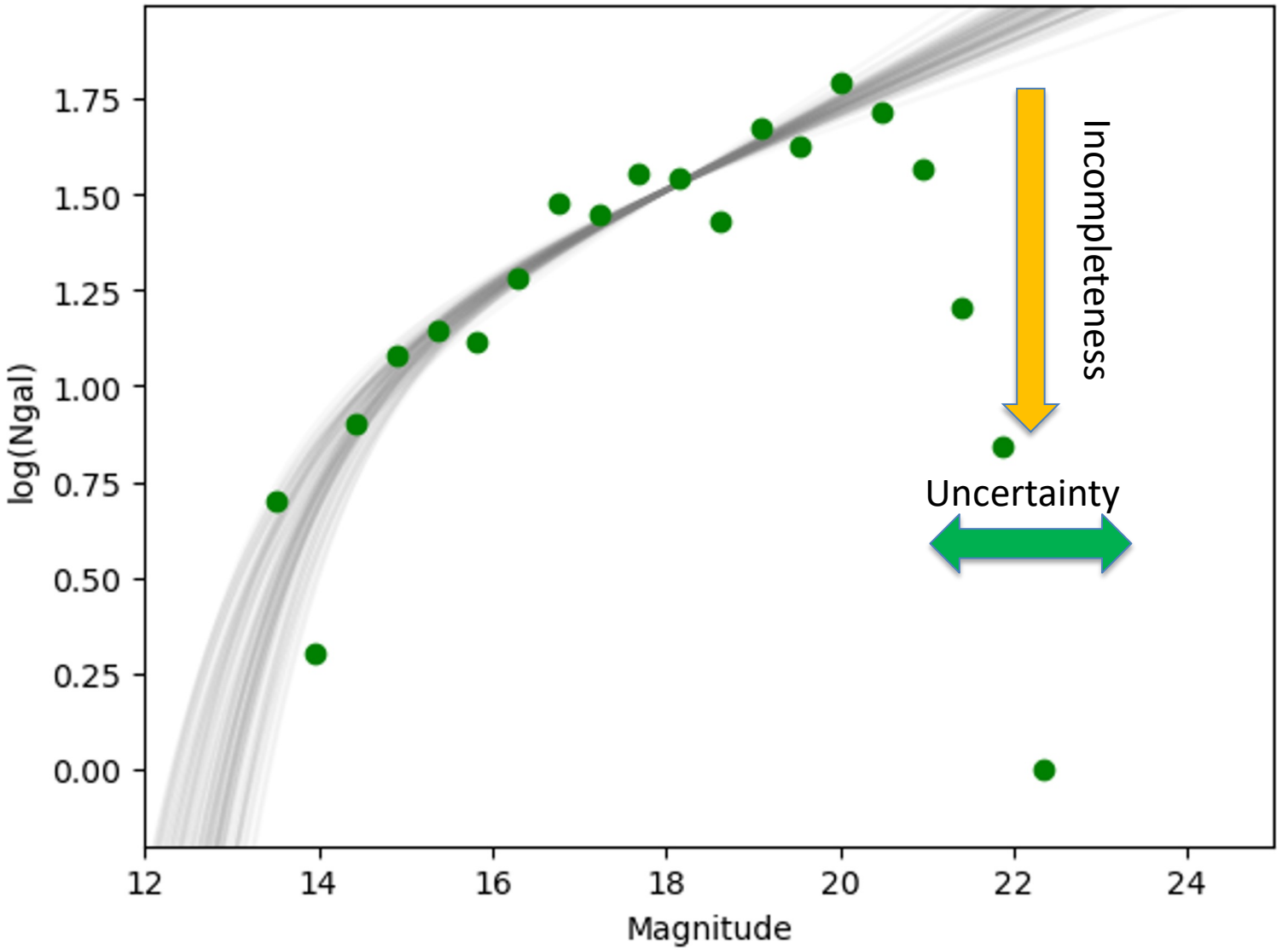
Flat Priors:

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

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

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



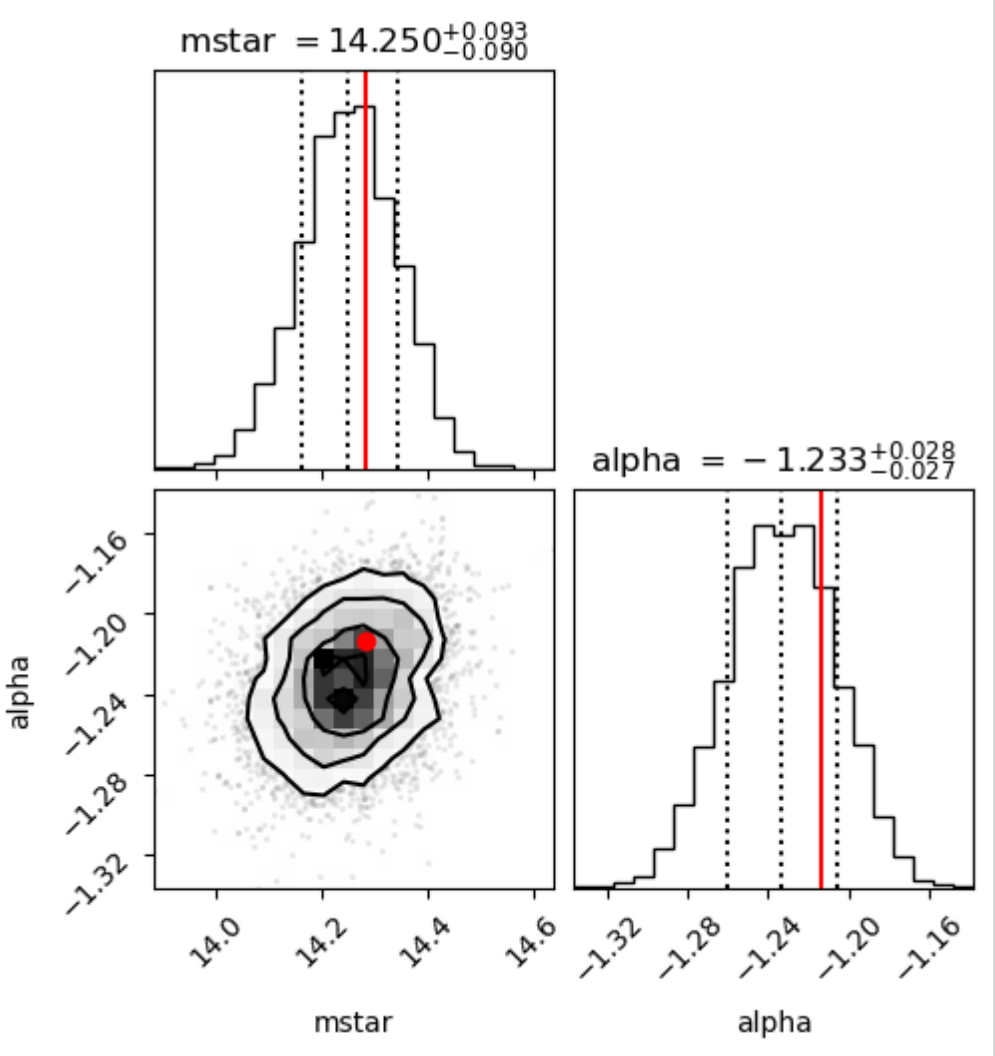


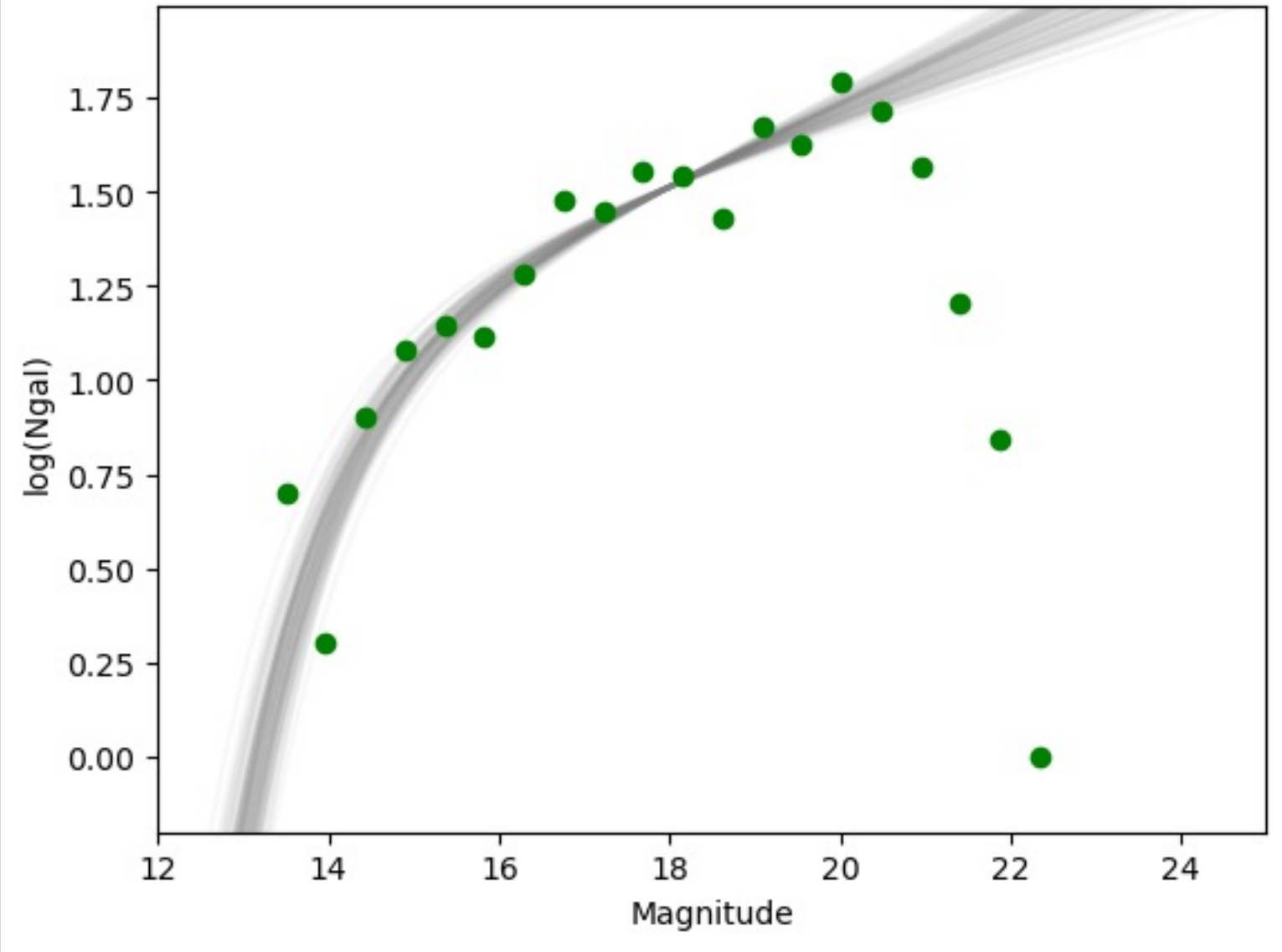
Flat Priors:

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

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 σ=0.1
- α: $-1.5 < \alpha < -0.5$