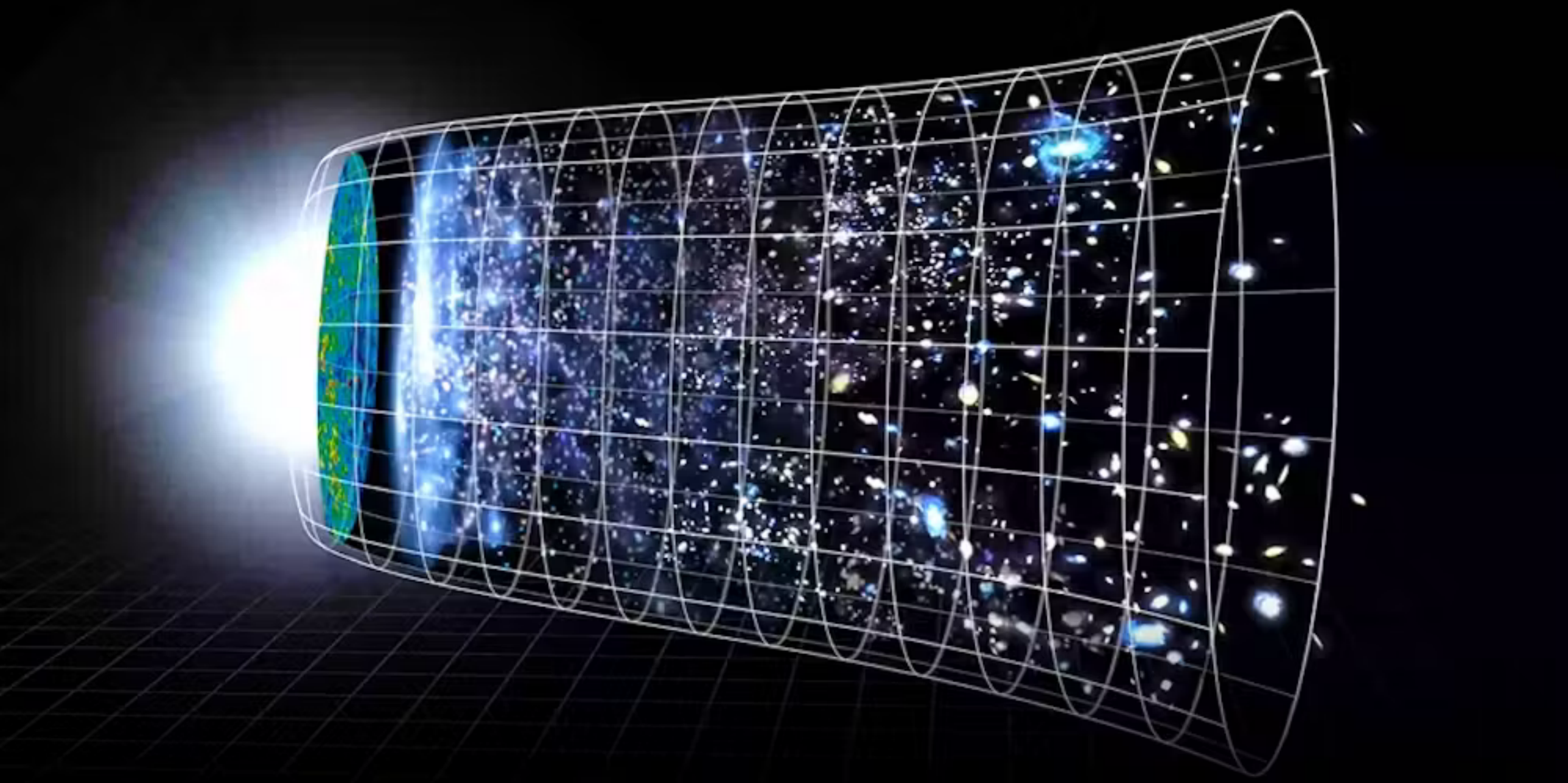
# Quantifying Inference Biases and Their Uncertainties in Distance Measurements



Hai Fu, University of Iowa

# Key Questions

- How do astronomers measure distances?
- What inference biases are important for distance measurements?
- How to estimate and correct these biases?
- How to incorporate the uncertainties of bias estimates?

# The Standard Candle Method

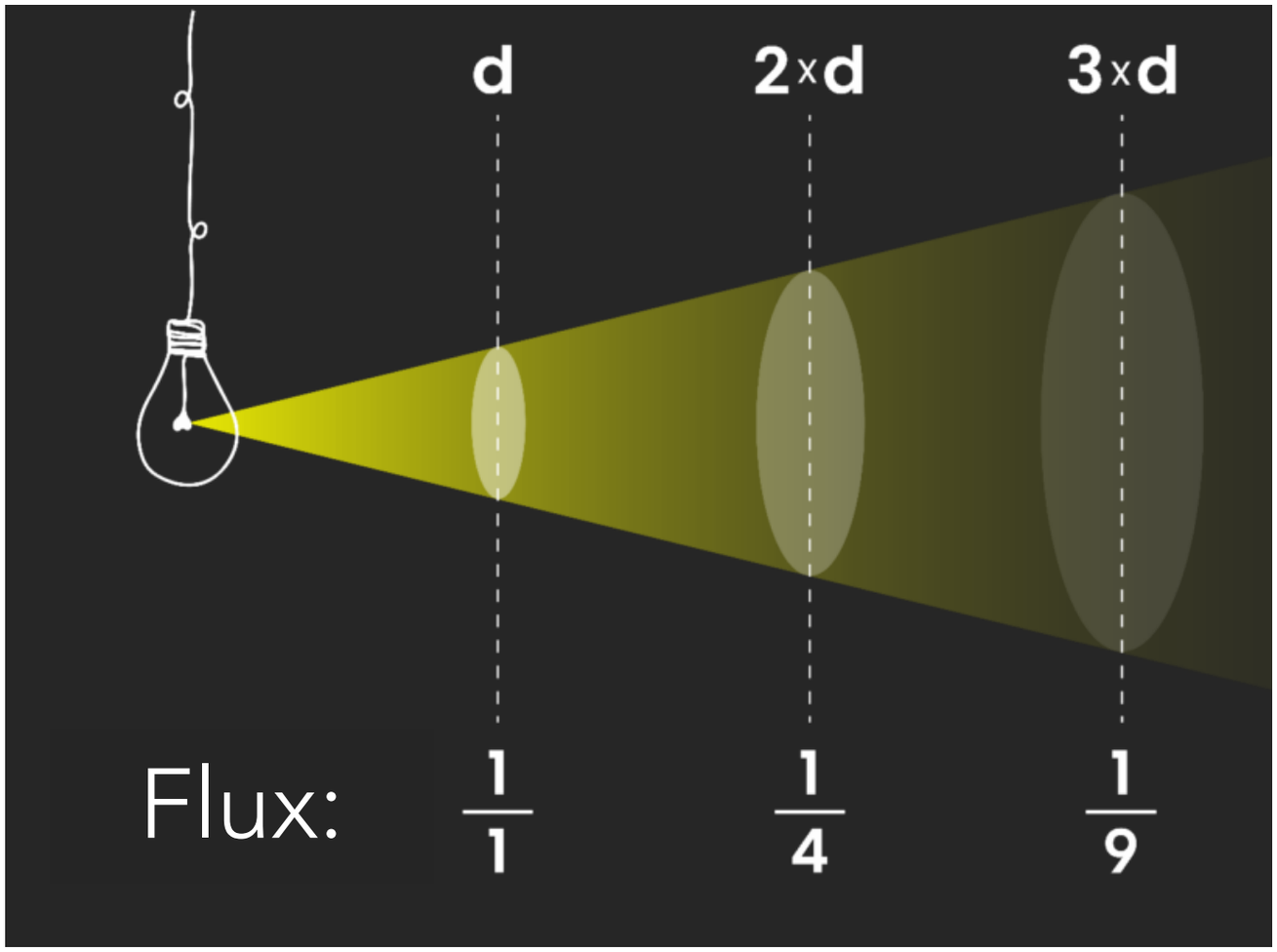
# Distance from Luminosity and Flux: the Inverse Square Law

- **Log Luminosity:** $y$
  - $y = \log L$

- **Log Flux:** $f$
  - $f = \log 4\pi F$

- **Log Distance:** $d$
  - $d = y - f = 2 \log D$

  given $D^2 = L/4\pi F$ (distance square law)




- ***Absolute & apparent magnitudes:*** $-2.5y$, $-2.5f$

  ***Distance modulus*** $\mu = m - M$: $2.5d$

# Major Standard Candles and Their Luminosity Relations

**Luminosity must be *predictable* from distance-independent observables:**

- **Pulsating variables** follow Leavitt (1912) law ($P$: pulsation period):

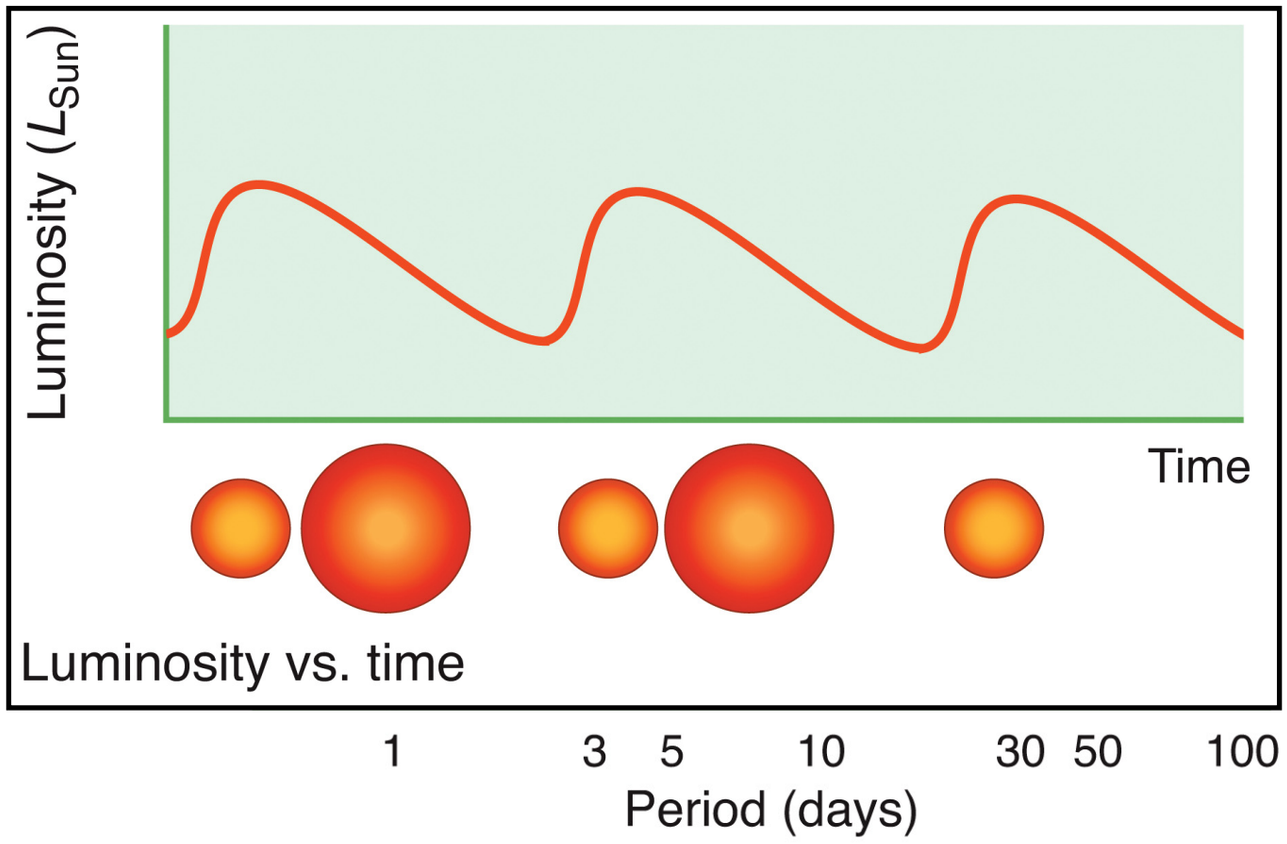
$$y = y_0 + \beta(\log P - 1)$$

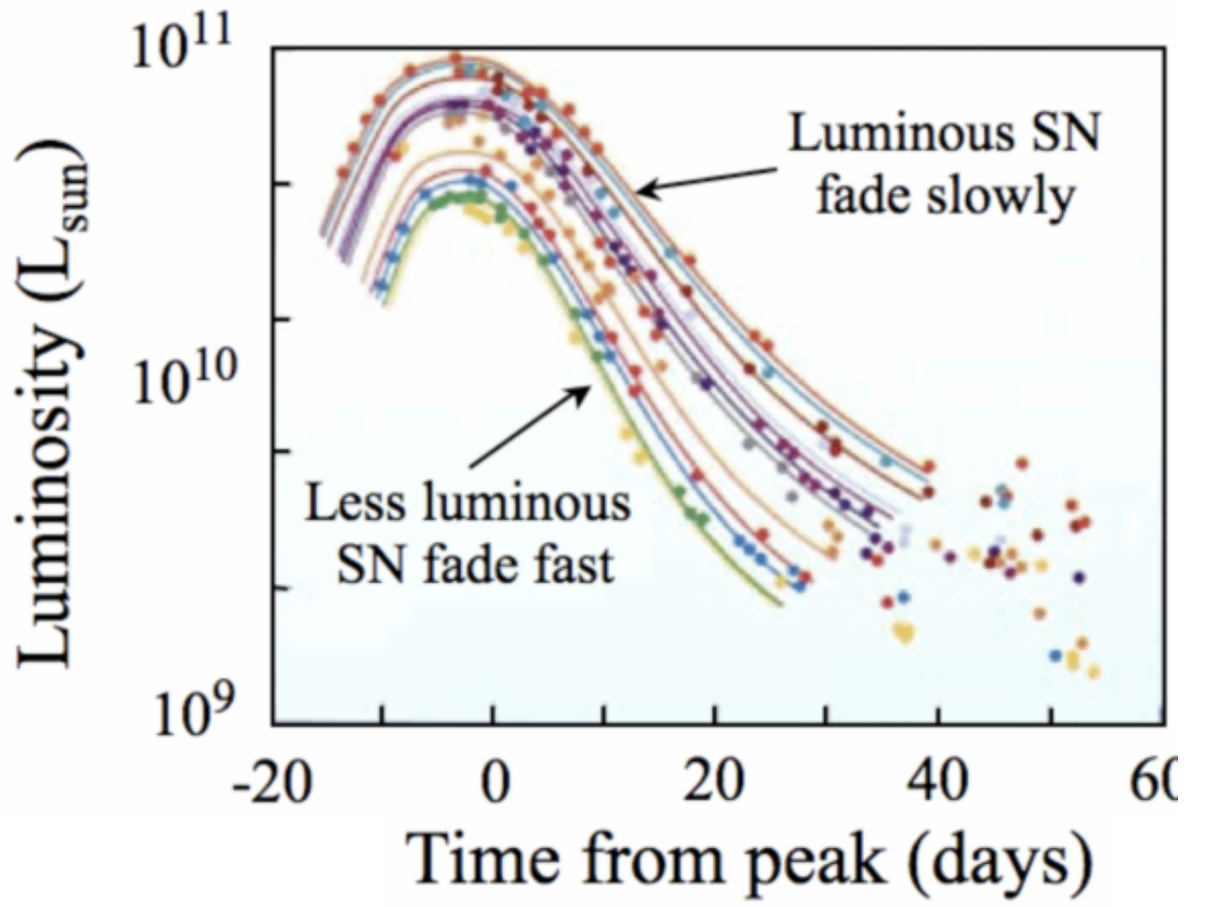
- **Type Ia supernovae** follow Phillips (1993) law ($\Delta m$: light decline rate):

$$y = y_0 + \beta \Delta m$$

- **Disk galaxies** follow the Tully-Fisher (1977) relation ($W_{\rm HI}$: rotation velocity, $\sin i$: inclination angle correction):

$$y = y_0 + \beta(\log W_{\rm HI}/\sin i - 2.5)$$

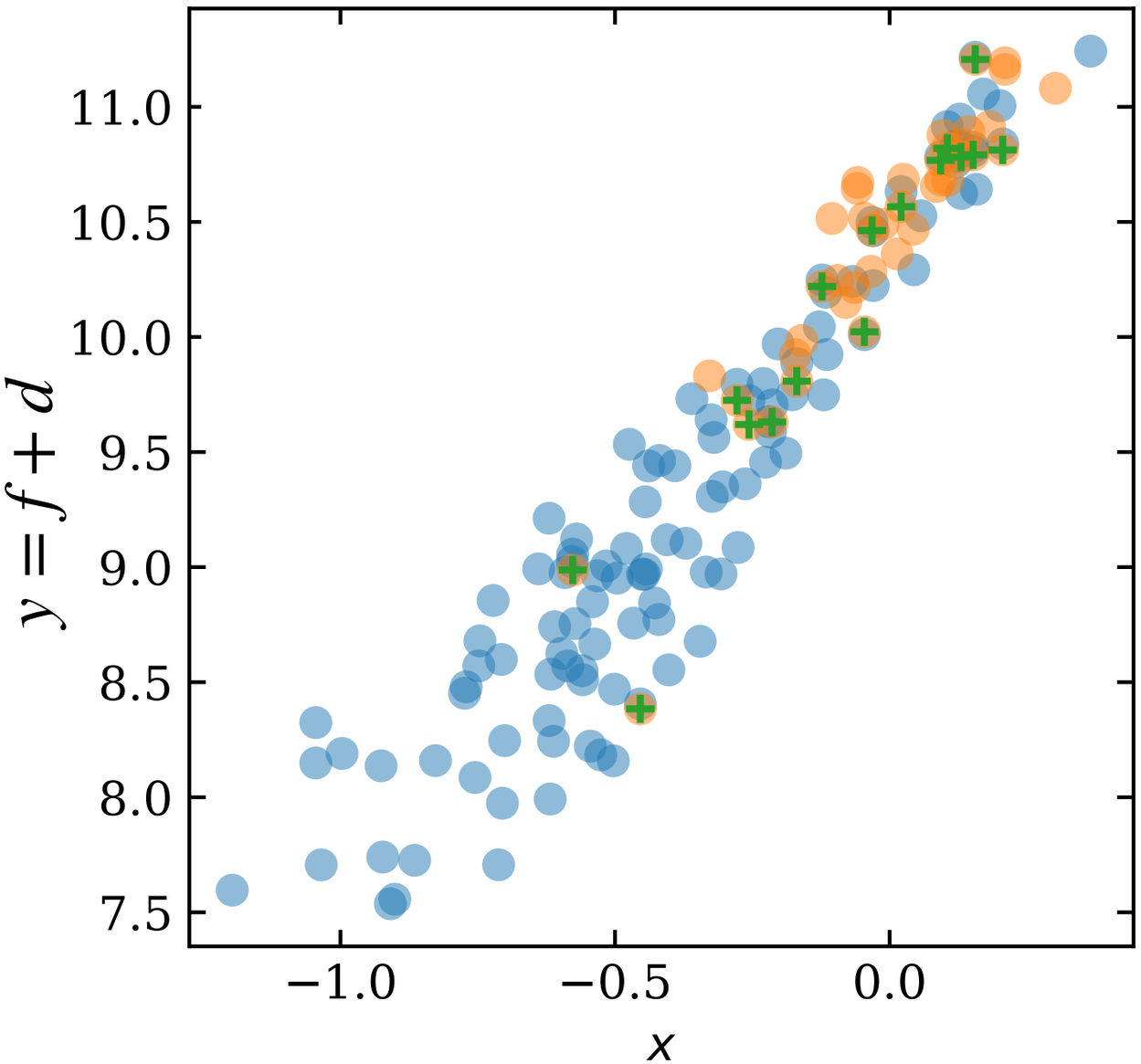


# Distance Measurement Requires Calibration of Luminosity Relations

**Luminosity Relation in a General Form:**

$$y = \beta x + \gamma$$




**Data in a calibration sample**:

$x_i$ (luminosity predictor)

$y_i = f_i + d_i$ (distances from priors)

**Inference (Regression):**

*Fit the data with model*

$y = \beta x + \gamma$ to infer $\beta, \gamma$

**Distance Measurements:**

$d_j = \gamma + \beta x_j - f_j$

for objects without distance priors

**Therefore, if $\beta, \gamma$ are biased, distances will be biased.**
**Consequently, any distance-dependent quantities will be biased.**

Fitting the **slope** & the **intercept** $(y = \beta x + \gamma)$ from data points $\{x_i, y_i = f_i + d_i, \sigma_{x,i}, \sigma_{y,i}\}$ is a **linear regression.**

It should be **pretty easy**, RIGHT?

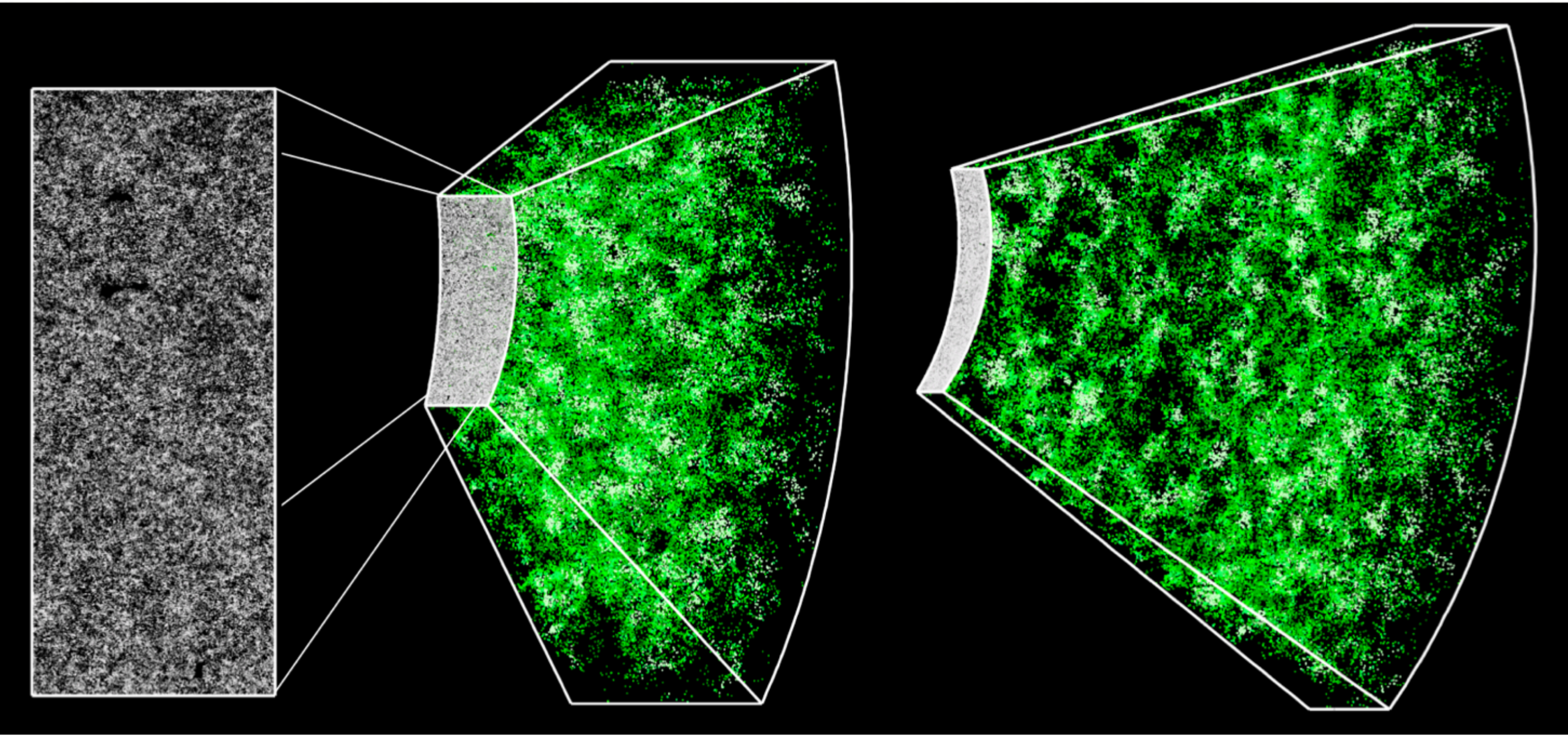
What **biases** could there be?

# Malmquist Bias
# (distance-dependent)

# Astronomical surveys cover larger volumes at greater distances

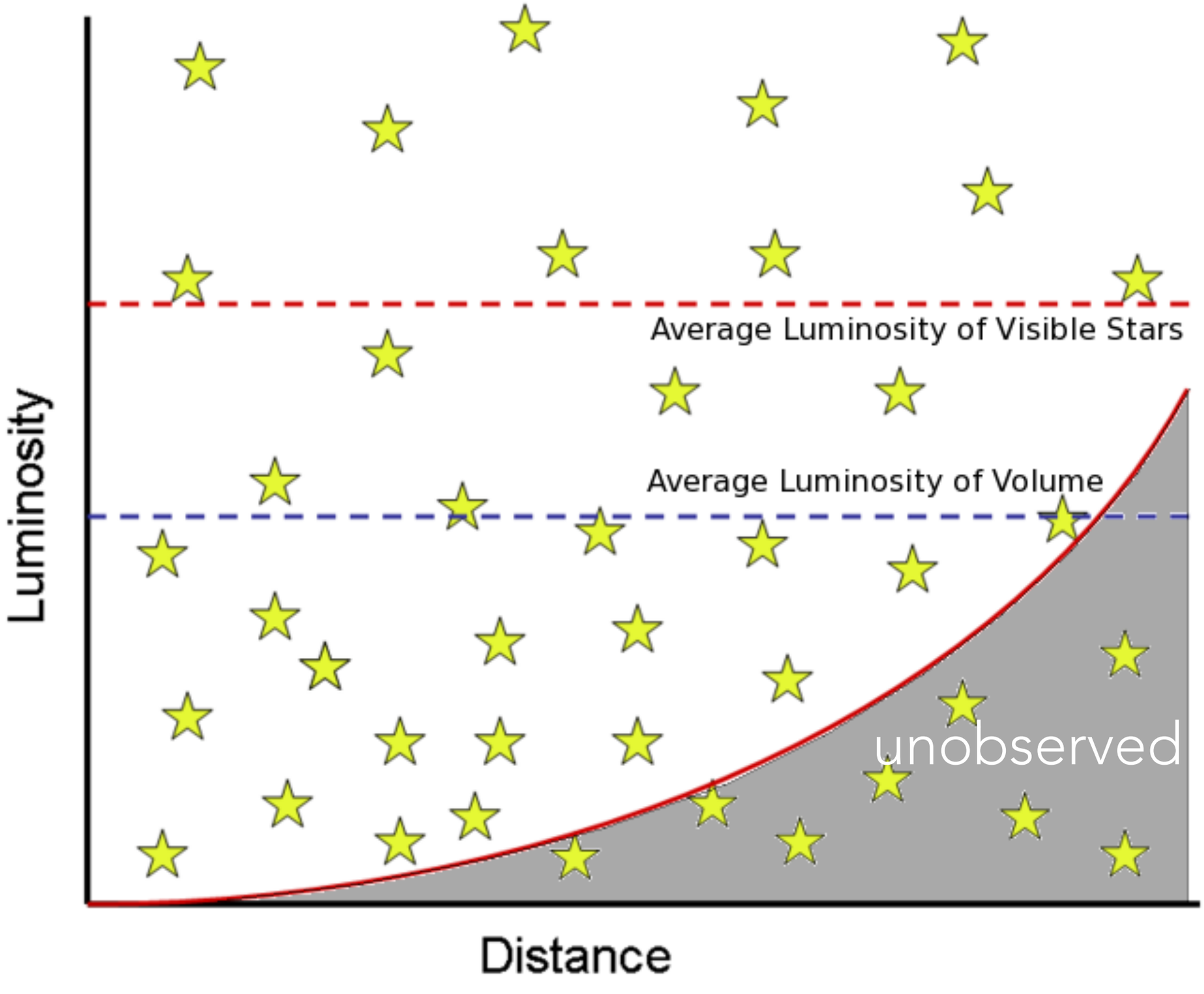
Each survey focuses on a fixed area on the sky, leading to the spatial **volume** to increase with **distance squared**:

$$dV = \Theta D^2 dD$$



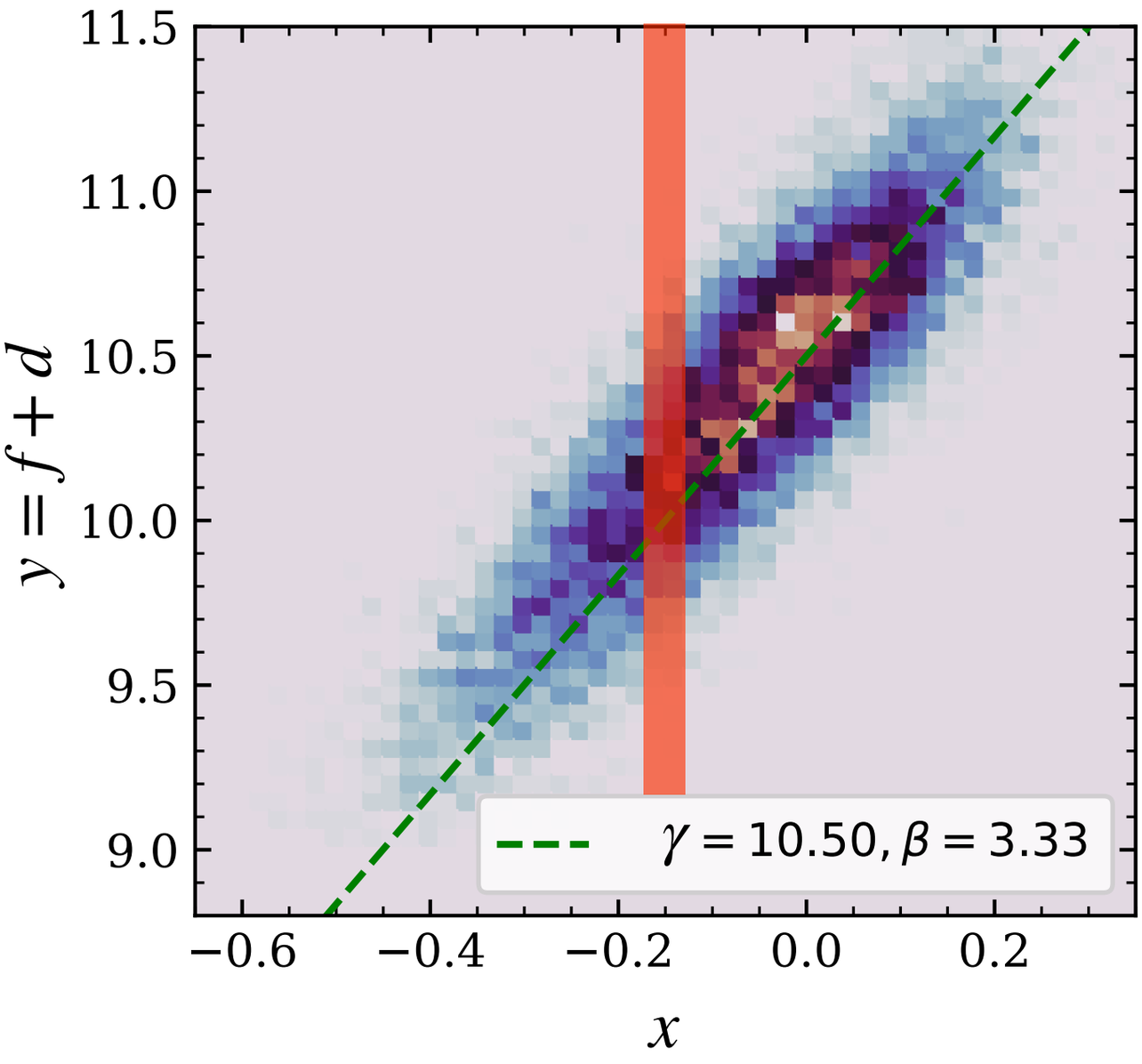
## But only more luminous objects can be detected at greater distances

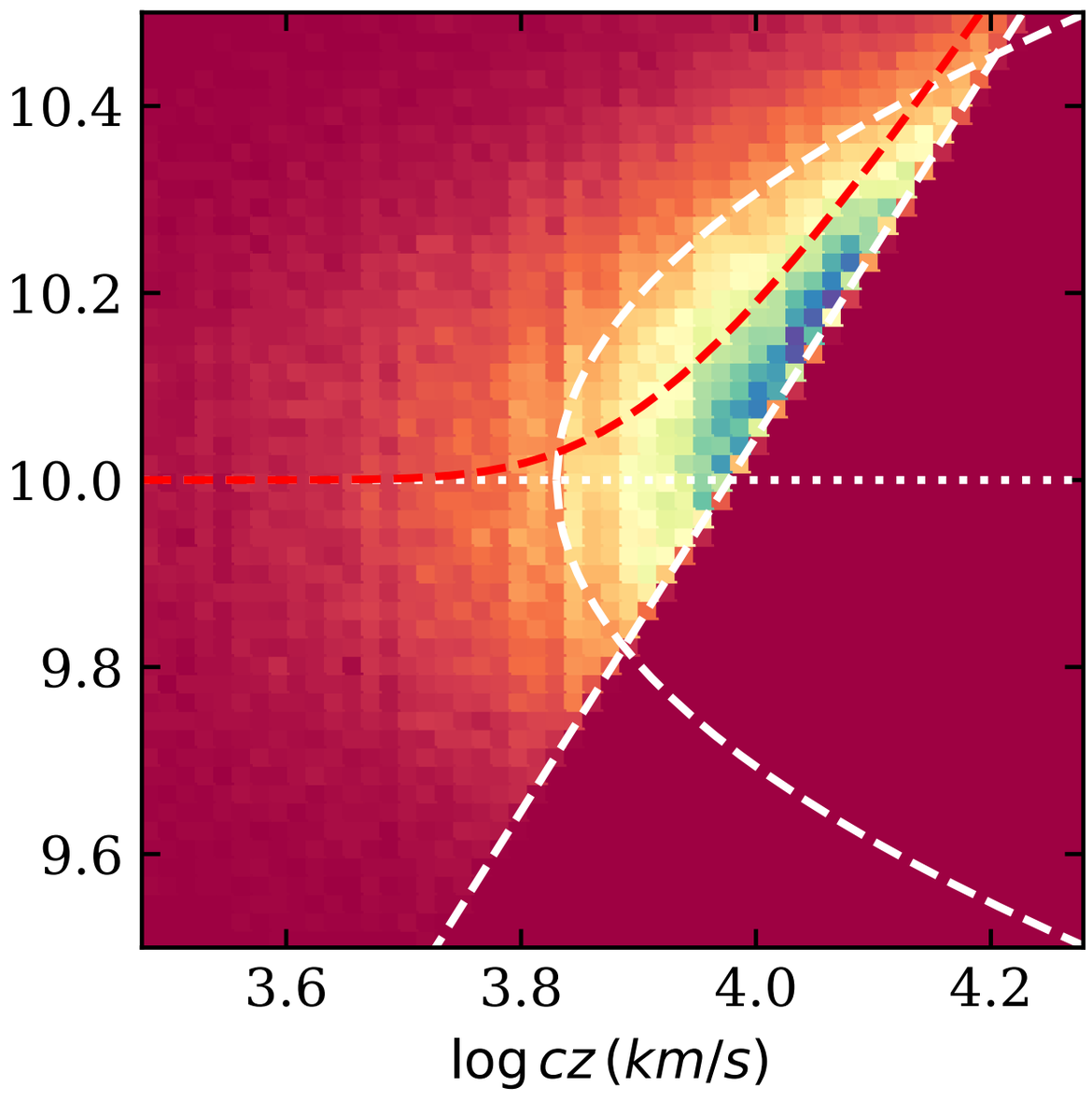
As distance increases, both **survey volume** and **luminosity-limit** increase



# Malmquist Bias: Sample Selection and Luminosity Dispersion

- All luminosity relations (e.g., Leavitt, Phillips, Tully-Fisher) have **intrinsic dispersions** and **measurement errors**. e.g., there are **scatters in luminosity** for Cepheids selected to have the **same *P***.
- As distance increases, surveys cover larger volume but can only detect the more luminous objects, causing the **distance-dependent Malmquist bias.**



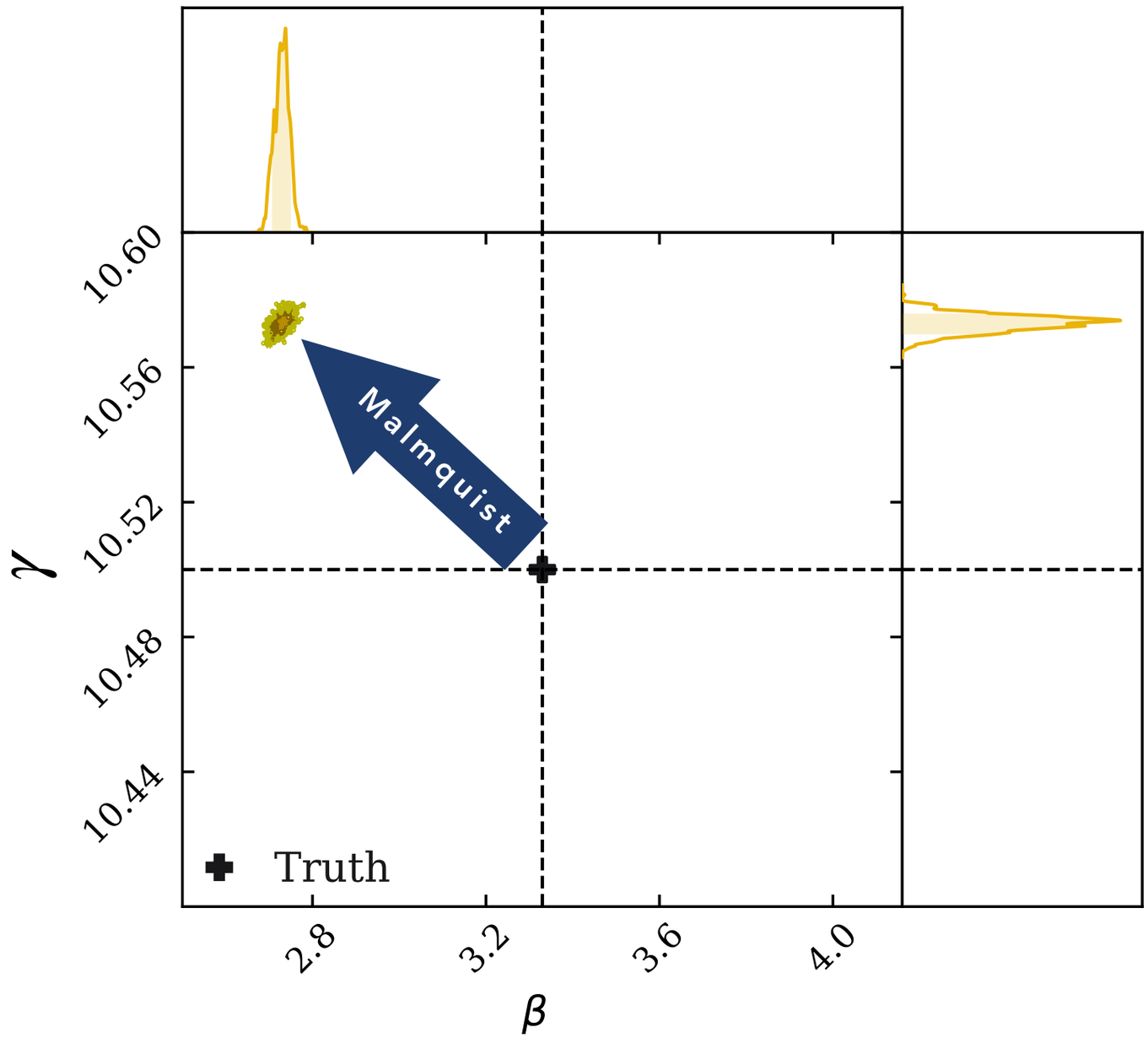



# *D*-dependent Malmquist Bias: Formula & Effects on Regression Coefficients

- As distance increases, the mean luminosity of standard candles increases due to observational selection (Willick 1994):

$$y(x) - \langle \tilde{y} \rangle_{x,d} = -\sigma_y \sqrt{\frac{2}{\pi}} \frac{\exp[-(f_l + d - y(x))^2/(2\sigma_y^2)]}{\mathrm{erfc}[(f_l + d - y(x))/(\sqrt{2}\sigma_y)]} < 0$$

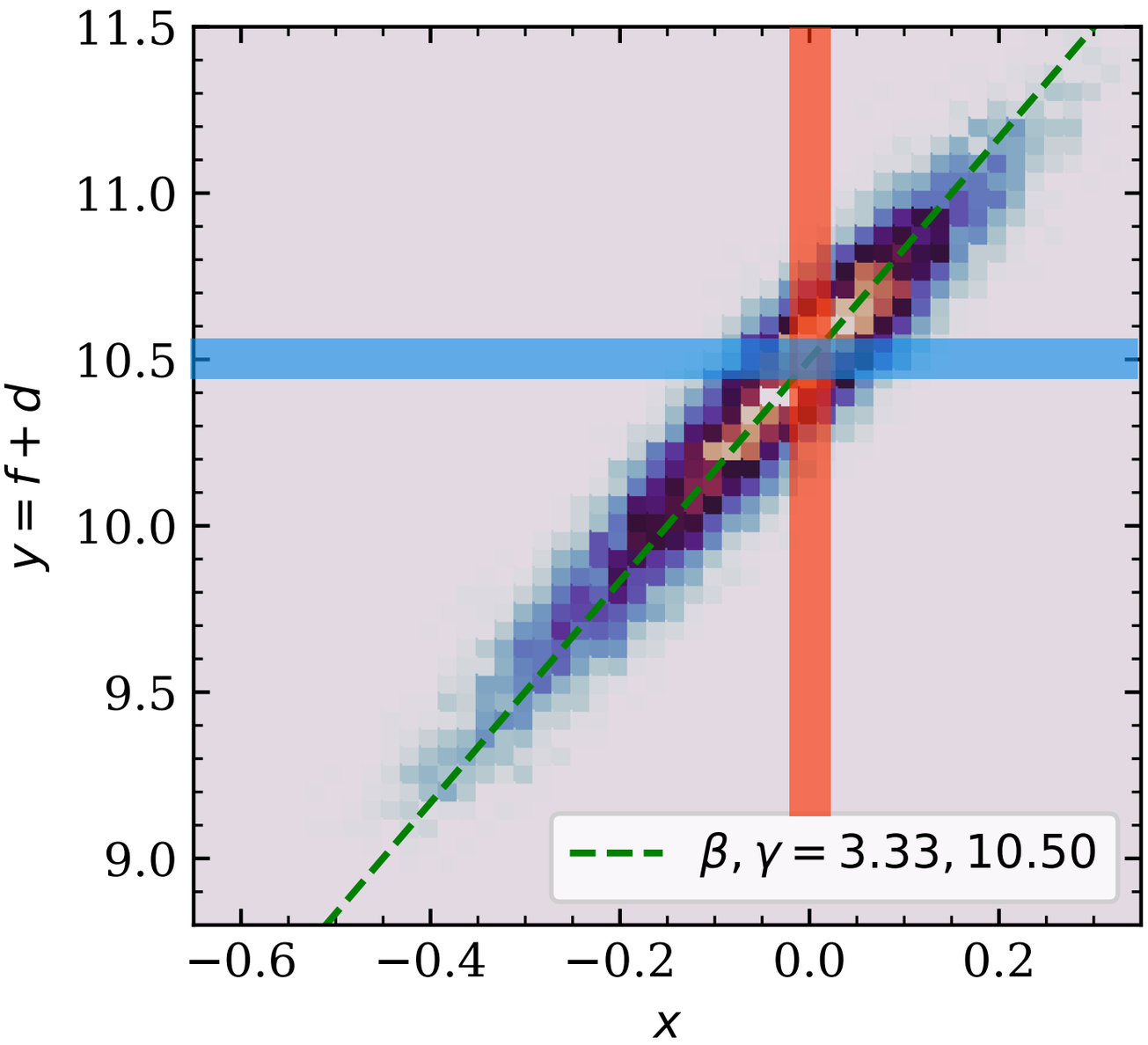
- The shift in the mean causes **underestimated slope** and **overestimated intercept.**

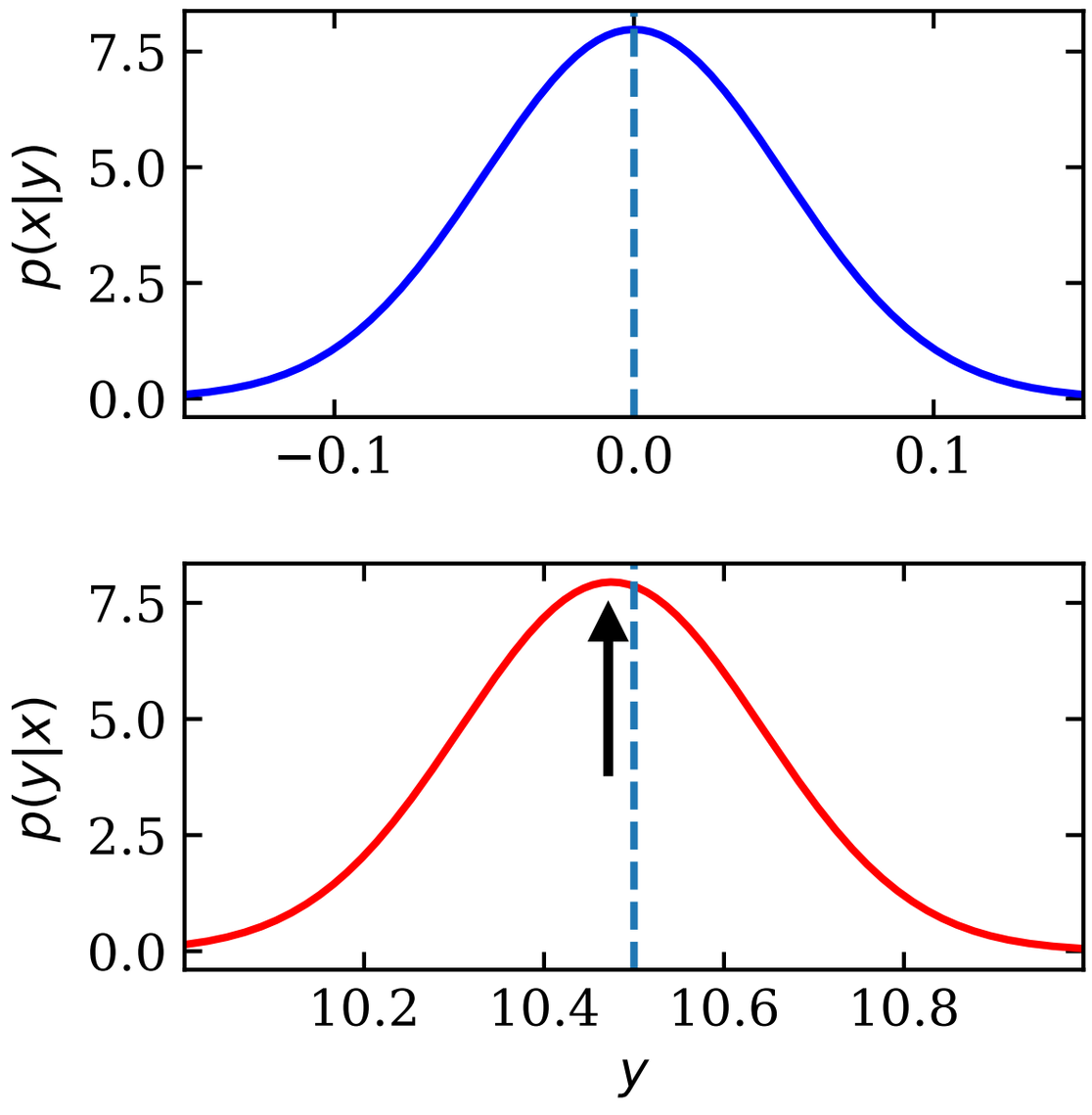


# Eddington Bias (Generalized)

# The General Eddington Bias in Luminosity Relations

- **Gaussian scatters in *x*** propagate to **skewed Gaussian scatters in *y*** when the distribution function in *y* is **non-uniform** (as is always the case).
- Given the probability identify, $p(y\,|\,x) = p(x\,|\,y)p(y)/p(x)$, we have $p(y\,|\,x) \propto p(x\,|\,y)p(y)$ at any given x. Thus if $p(x\,|\,y)$ is a Gaussian, $p(y\,|\,x)$ would be a skewed Gaussian with its mean shifted to **lower luminosity than predicted**.



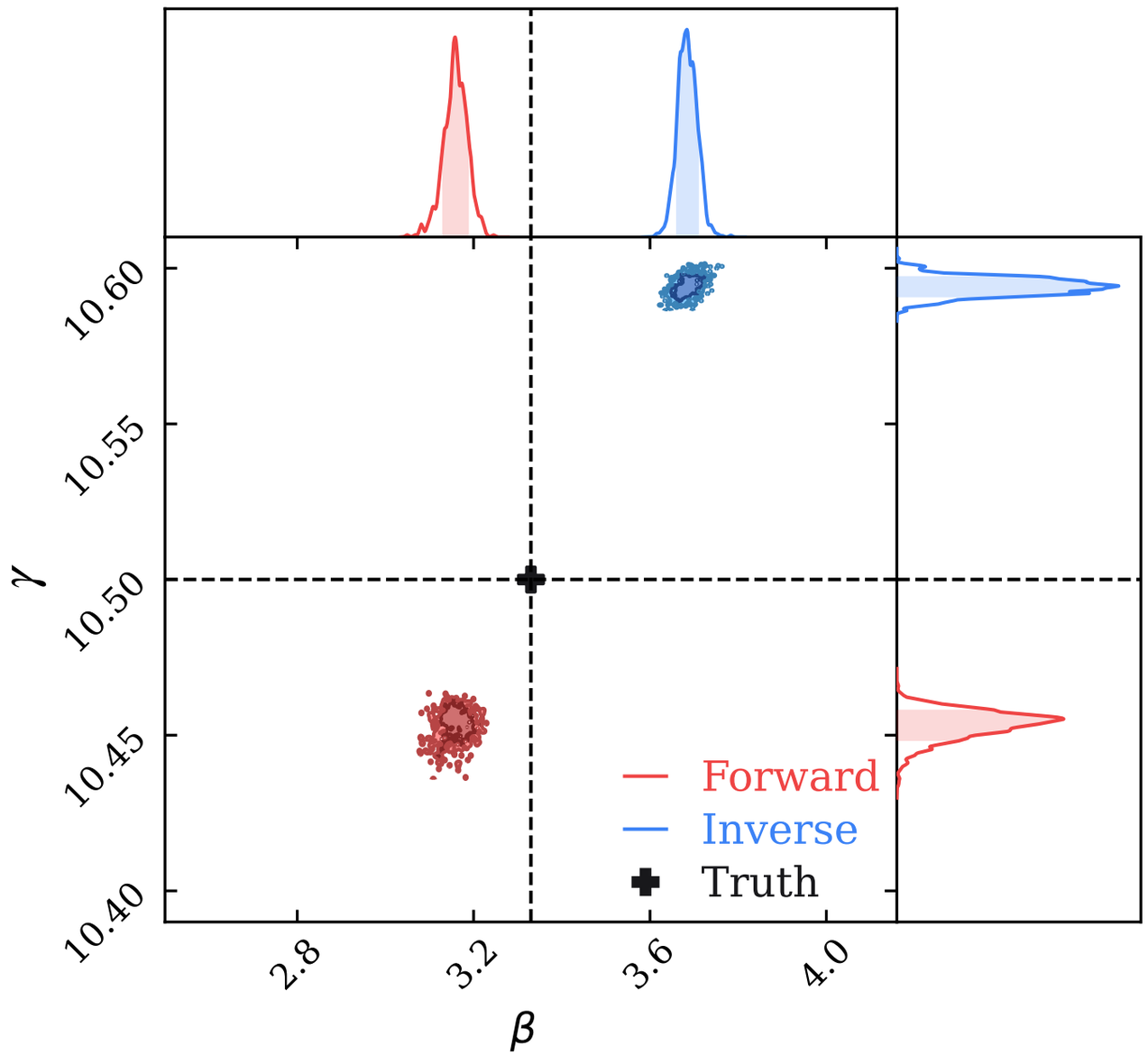



# General Eddington Bias: Formula & Effects on Regression Coefficients

Gaussian scatter in $\tilde{x}$ causes **a bias** of **the mean *y* at the given** $\tilde{x}$, $\langle y \rangle_{\tilde{x}}$ (Fu 2025):

$$y(\tilde{x}) - \langle y \rangle_{\tilde{x}} = -\beta\sigma_x^2 \frac{d \ln p(\tilde{x})}{d\tilde{x}} > 0$$

- The classic Eddington bias: $\tilde{x} - \langle x \rangle_{\tilde{x}} = -\sigma^2 d \ln p(\tilde{x})/d\tilde{x}$ is a special case of the above.
- Since dimmer objects are more numerous, $dp(\tilde{x})/d\tilde{x} < 0$, so that $\langle y \rangle_{\tilde{x}} < y(\tilde{x})$. This leads to **lower inferred intercept ($\gamma$)** for **forward model**, and the **opposite** for **inverse model**.



# How to mitigate the biases and incorporate bias uncertainties?

# Maximize Data Likelihood: $\chi^2$ as an example

- The preferred mitigation method should be able to infer $\beta, \gamma$ and constrain the bias-contributing parameters $\sigma_x, p(x|\boldsymbol{\eta}), \sigma_y, f_l$ at the same time .
- This leads us to the method of **maximizing the data likelihood**:
  - Compute the probability of each data point $\{x_i, f_i, d_i\}$, when accounting for the sample selection function
  - Multiply the probabilities to obtain the likelihood of the full dataset
  - Maximize the likelihood to infer all parameters $\beta, \gamma; \sigma_x, p(x|\boldsymbol{\eta}), \sigma_y$ and their statistical uncertainties through a Monte-Carlo Process.
- The simplest example is $\chi^2$ **minimization**:
  - **Probability of $y_i$ given $x_i$ and model $\theta$:**

$$P(y_i|x_i,\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{(y_i - y(x_i|\boldsymbol{\theta}))^2}{2\sigma_i^2}\right]$$

  - **Data likelihood is the product of $P(y_i|x_i,\boldsymbol{\theta})$ for all data points:**

$$\ln\mathscr{L} = \ln\prod P(y_i|x_i,\boldsymbol{\theta}) = -\sum\left[\ln(\sqrt{2\pi}\sigma_i) + \frac{(y_i - y(x_i|\boldsymbol{\theta}))^2}{2\sigma_i^2}\right]$$

$$\ln\mathscr{L} = -\sum\left[\ln(\sqrt{2\pi}\sigma_i)\right] - \chi^2/2$$

# Simulated Data $\{x_i, f_i, d_i\}$ for Testing

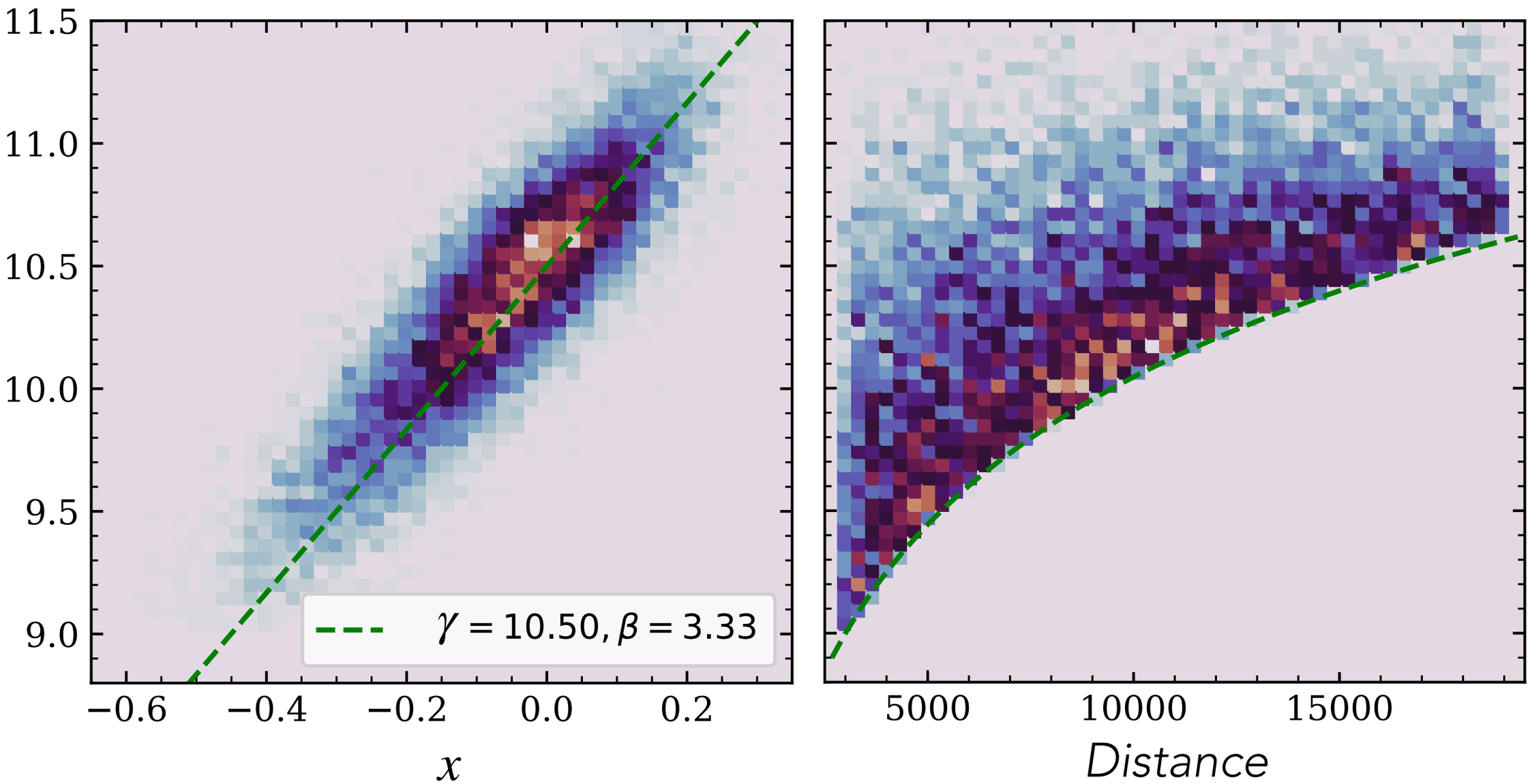
**Scatter in $y$:** 0.2 dex

**Log flux limit:** $f_l = 5.736$

**Scatter in $x$ :** 0.045 dex

**Distribution in $x$:** Powerlaw x Exponential

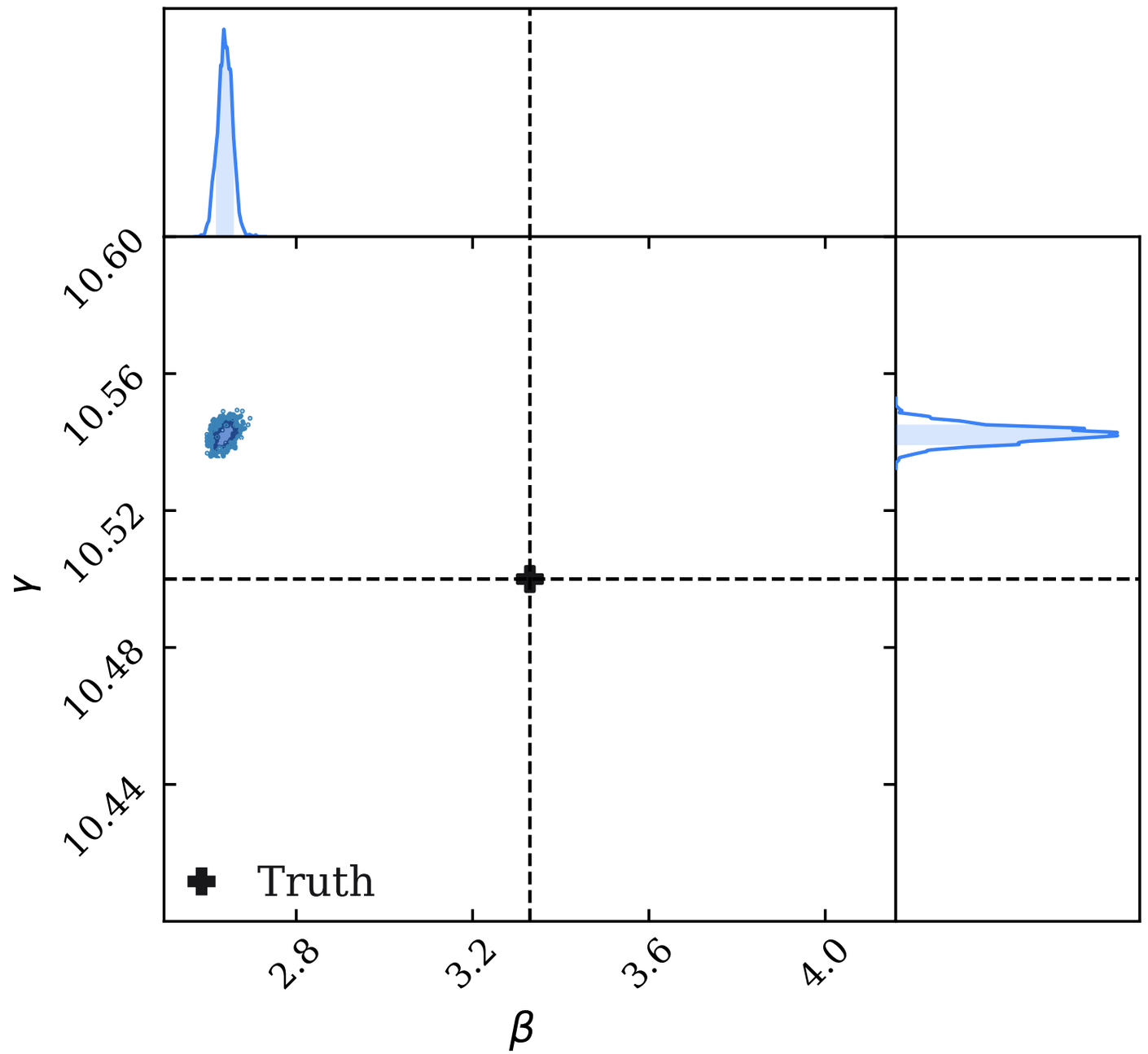
**Distance range:** 3,000 - 19,000



# No Bias Correction — Simple $\chi^2$ minimization

**Probability of $f_i$ with no selection function and no scatter in $x$ (Blue curves):**

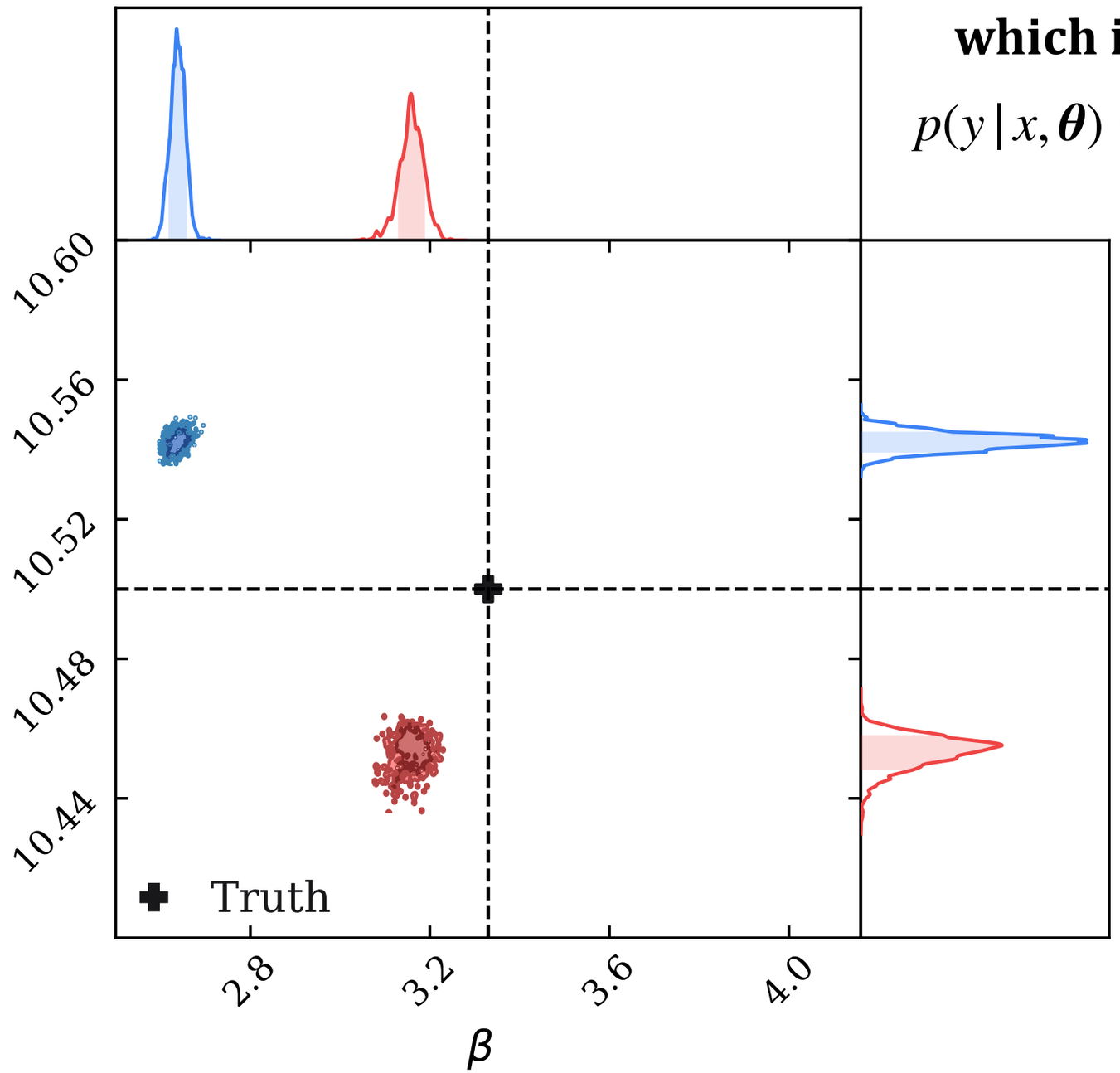
$$P(f_i \mid x_i, d_i, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left[-\frac{(f_i + d_i - y(x_i \mid \boldsymbol{\theta}))^2}{2\sigma_y^2}\right]$$



# Correcting Distance-Dependent Malmquist Bias

**Probability of $f_i$ with step selection function but no scatter in $x$ (Red curves):**

$$p(f_i \,|\, x_i, d_i, \boldsymbol{\theta}) = \frac{\sqrt{2}}{\sqrt{\pi}\sigma_y} \frac{\exp\left[-(f_i + d_i - y(x_i \,|\, \boldsymbol{\theta}))^2 / 2\sigma_y^2\right]}{\operatorname{erfc}\left[(f_l + d_i - y(x_i \,|\, \boldsymbol{\theta}))/\sqrt{2}\sigma_y\right]}$$



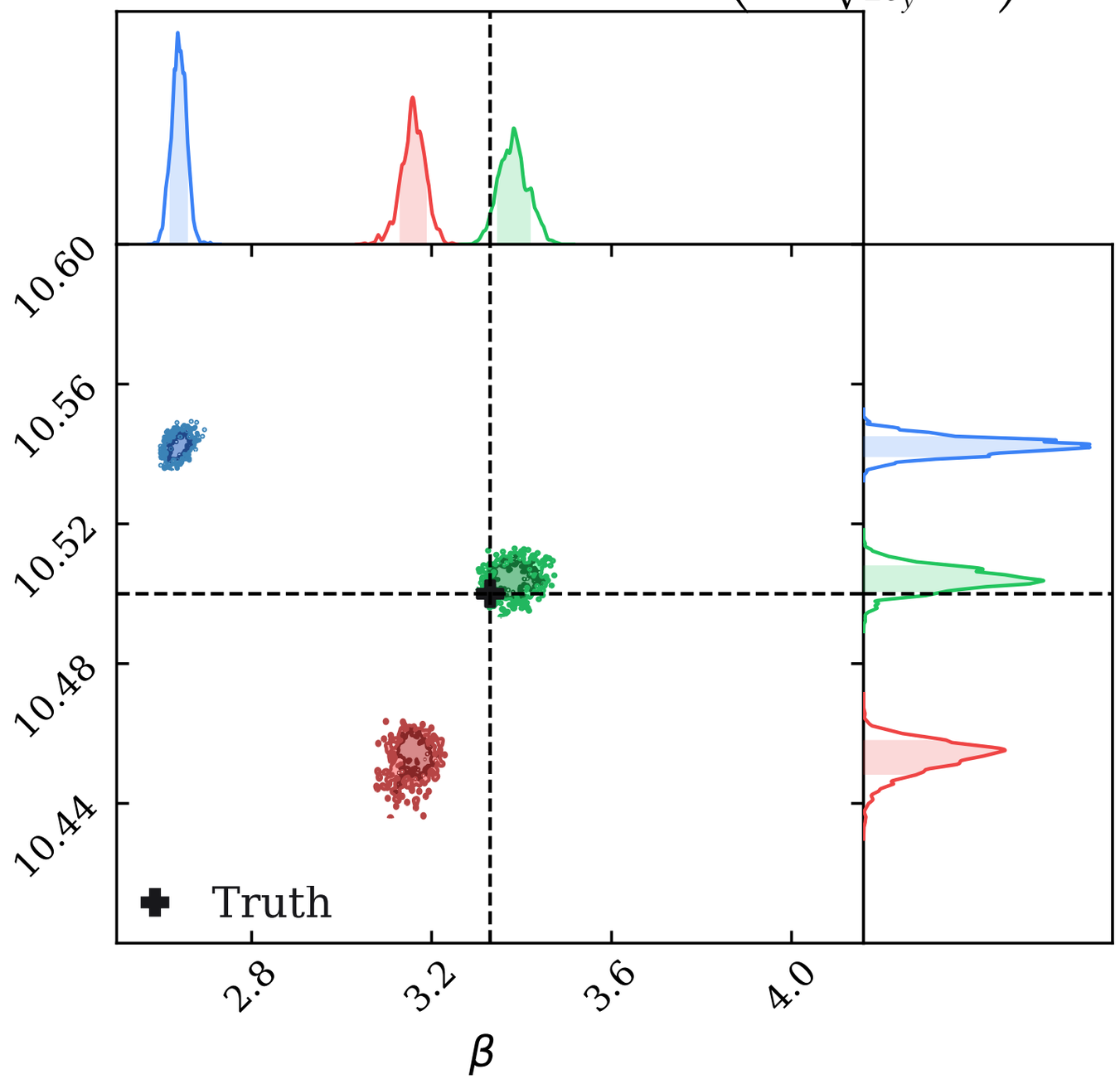

**which is derived from:**

$$p(y \,|\, x, \boldsymbol{\theta}) = \frac{p(y, x \,|\, \boldsymbol{\theta})}{\int_{f_l + d}^{\infty} p(y, x \,|\, \boldsymbol{\theta}) dy}$$

# Correcting Both Eddington and Malmquist Biases

**Probability of $f_i$ with step selection selection, $\sigma_x$, and $p(x)$ (Green curves):**

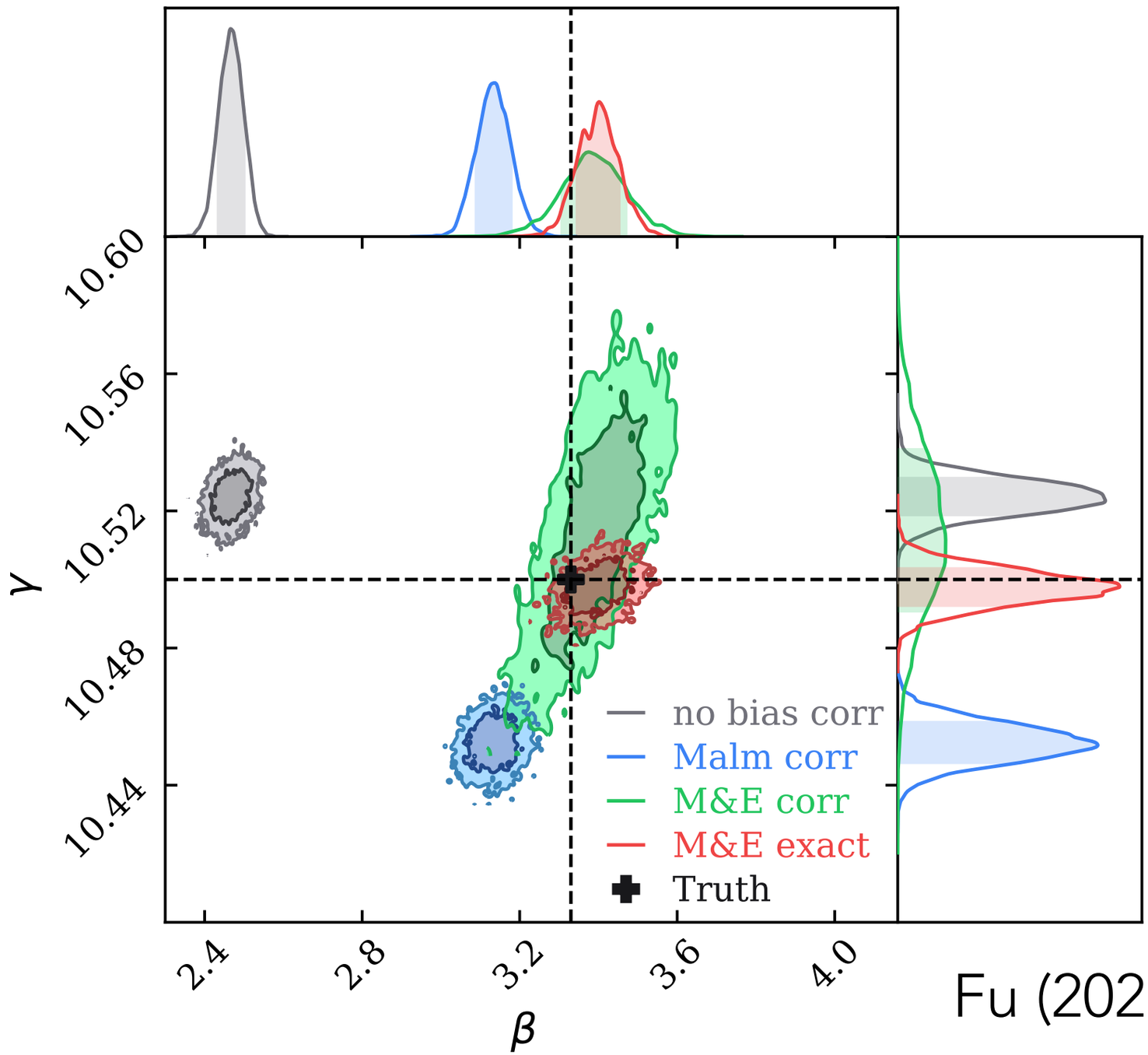
$$p(f_i \mid x_i, d_i, \boldsymbol{\theta}) = \frac{\sqrt{2}}{\sqrt{\pi}\sigma_y} \frac{\int_{-\infty}^{\infty} \exp\left[-\frac{(x-x_i)^2}{2\sigma_x^2}\right] \exp\left[-\frac{(f_i + d_i - y(x_i \mid \boldsymbol{\theta}))^2}{2\sigma_y^2}\right] p(x \mid \boldsymbol{\eta}) dx}{\int_{-\infty}^{\infty} \exp\left[-\frac{(x-x_i)^2}{2\sigma_x^2}\right] \operatorname{erfc}\left(\frac{f_l + d_i - y(x_i \mid \boldsymbol{\theta})}{\sqrt{2}\sigma_y}\right) p(x \mid \boldsymbol{\eta}) dx}$$



# Finally, Propagating the Uncertainties of Bias Estimates

Bias correction can not be exact, so the **uncertainties of the bias estimates** must be combined with **standard errors** to give the **total statistical uncertainty**.

By allowing bias-contributing parameters ($\sigma_y, f_l, \sigma_x, \boldsymbol{\eta}$ in $p(x\,|\,\boldsymbol{\eta})$) to vary freely, their uncertainties are constrained by the data and propagated to the posterior of $(\beta, \gamma)$ (**Green vs Red contours**)




Fu (2025) **ApJ** 992 214

# Summary

- Two important inference biases have been discussed:
  - Distance-dependent Malmquist bias due to observational selection, luminosity dispersion, and volume increase w/ distance.
  - Generalized Eddington bias due to scatter of the independent variable and its non-uniform distribution.
- Both biases can be mitigated by properly accounting for the above factors in the data likelihood function.
- When the bias-related parameters are allowed to vary freely, the uncertainties of bias estimates are constrained by the data and are propagated into the total statistical uncertainties.