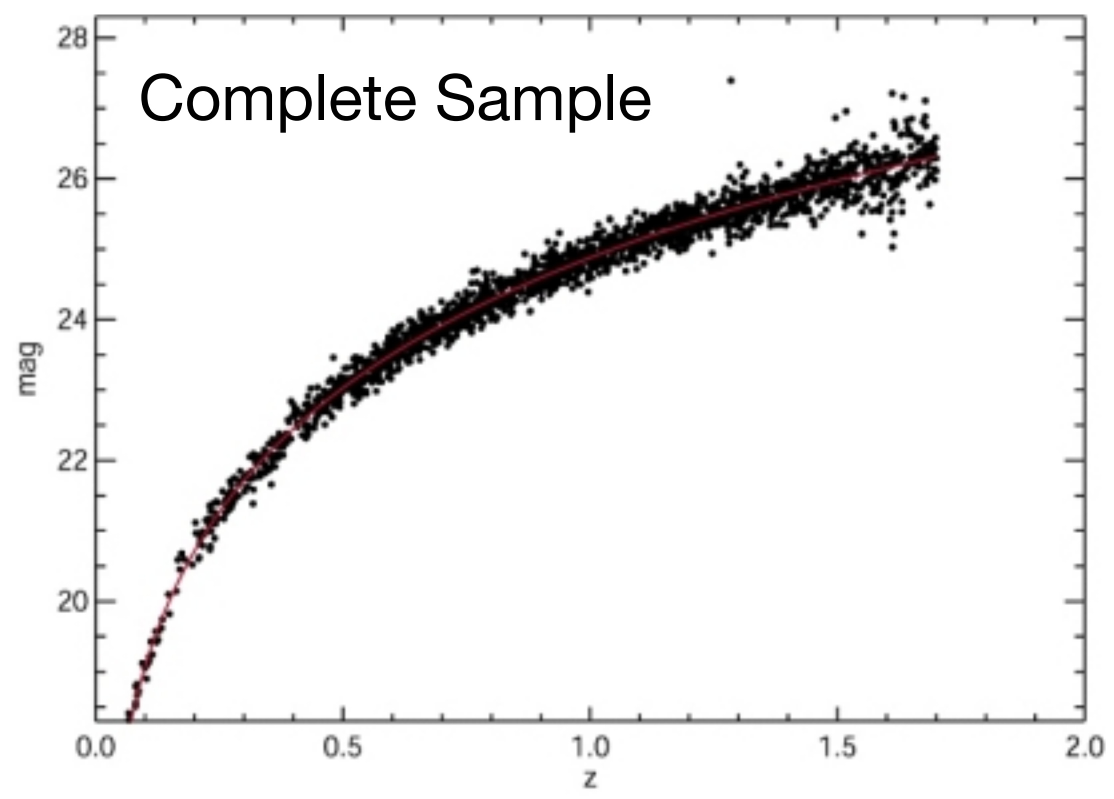
# Characterizing the Sample Selection for Supernova Cosmology

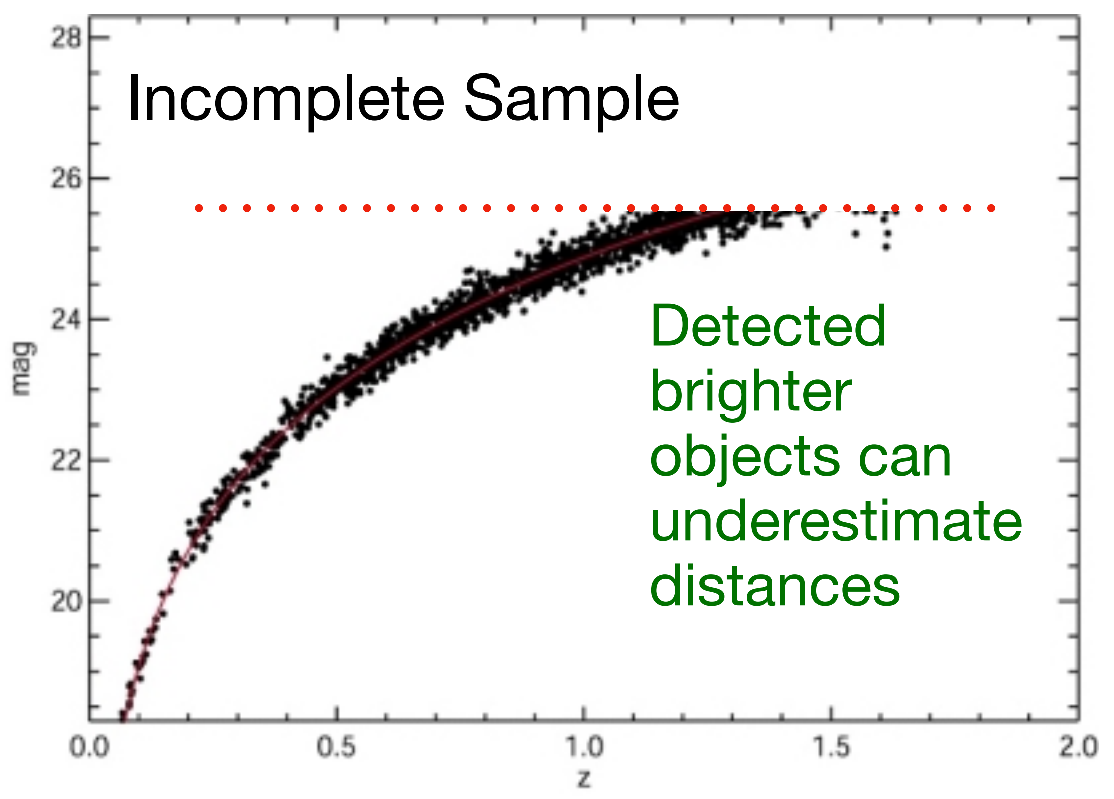
Author: Alex Kim (Lawrence Berkeley National Laboratory)
Internal Reviewers: Rahul Biswas, Phil Marshall, Gautham Narayan
DESC Paper Tacking Link

# Motivation

## Sample Selection An Important Ingredient for Cosmology Analysis

- Incomplete sample selection must be considered in SN Ia cosmology analysis
    - Canonical example is a magnitude-limited sample, which can lead to Malmquist bias in distances estimated from standard candles with an intrinsic magnitude dispersion




- DESC planned sample selection is more complicated than just a magnitude limit
    - Detection based on multiple color/phase measurements
    - Typing based on machine-learned photometric SN Ia classification

# Motivation

## Alert Brokers a Component of DESC Transient Pipelines

- Third-party brokers provide customized subsets of near real-time LSST-discovered alerts/objects
- DESC plans to use brokers to identify targets for real-time spectroscopic follow-up and likely Type Ia supernovae to include in the cosmology analysis
    - Brokers would then in part determine sample selection
- To evaluate brokers for use in our transient pipelines DESC needs to determine the computational requirements for calculating the “sample selection”
- After saying this for 1+ year, I was asked by Rahul for quantitative requirements

# Approach

- Selection function is not analytic or depends on many measurements
- Quantities needed for cosmology analysis that depend on the selection function are calculated using Monte Carlo
    - Uncertainties in these quantities depend on the number of Monte Carlo samples
    - Each Monte Carlo sample is processed through the transient pipeline (broker) to get its selection probability
- Uncertainties propagate into errors in the
    - Position of the maximum likelihood, i.e. parameter estimators
    - Hessian at maximum likelihood (proxy for parameter uncertainty)
- Requirements on the precisions of parameters and their uncertainties translate into a minimum **number of simulated MC samples to process through the transient pipeline (broker)**

# Model and Data

## Illustrative Toy Example

- Article presents an illustrative toy example to present the procedure; requirements are dependent on model and selection, which will not be specified for a few years
- Model
  - Perfect standard candle with intrinsic magnitude dispersion
  - Second background population with a different mean magnitude, broad intrinsic magnitude dispersion
  - Model parameters: distance modulus at the redshift; intrinsic fraction of the candle relative to background population; true type of each object

# Model and Selection Function

## Illustrative Toy Example

- Data
  - N candles at a fixed redshift that pass sample selection
  - Measured magnitude for each candle (independent)
- Sample Selection — Two criteria
  - Magnitude-limit cutoff ($S=1$)
  - Standard candles selection with false-positives, false negatives ($\tau=1$)

# Model and Data

## Likelihood

Fraction of model objects selected and classified as a candle

$$\bar{S}(\mu, p_0) = p(S = 1, \tau = 1 | \mu, p_0)$$

Probability of a model object being selected and classified as candle and having some magnitude

$$R(m, \mu, p_0) = p(m, S = 1, \tau = 1 | \mu, p_0)$$

Likelihood can be expressed as

$$\begin{aligned} L(\mu, p_0; \{m\}) &\equiv \prod_{i=1}^{N} p(m_i | S_i = 1, \tau_i = 1, \mu, p_0) \\ &= \prod_{i=1}^{N} \frac{p(m_i, S_i = 1, \tau_i = 1 | \mu, p_0)}{p(S_i = 1, \tau_i = 1 | \mu, p_0)} \\ &= \bar{S}(\mu, p_0)^{-N} \prod_{i=1}^{N} R(m_i, \mu, p_0). \end{aligned}$$

# Model and Data

## Maximum Likelihood and Hessian

Maximum likelihood and

$$0 = -\frac{1}{\bar{S}}\frac{\partial \bar{S}}{\partial \theta} + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R(m_i)}\frac{\partial R(m_i)}{\partial \theta}$$

Hessian

$$\begin{aligned} H_{ij} &= -\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \\ &= -\sum_{k=1}^{N}\left(\frac{1}{R(m_k)}\frac{\partial^2 R(m_k)}{\partial \theta_i \partial \theta_j} - \frac{1}{R^2(m_k)}\frac{\partial R(m_k)}{\partial \theta_i}\frac{\partial R(m_k)}{\partial \theta_j}\right) \\ &+ N\left(\frac{1}{\bar{S}}\frac{\partial^2 \bar{S}}{\partial \theta_i \partial \theta_j} - \frac{1}{\bar{S}^2}\frac{\partial \bar{S}}{\partial \theta_i}\frac{\partial \bar{S}}{\partial \theta_j}\right) \end{aligned}$$

- Depend on partial derivatives of S̄ and R with respect to model parameters
- Give familiar results when S̄ is parameter-independent

# S̄ and Its Partials

## Where Numerical Errors Enter

S̄ involves an integral/sum

$$\bar{S}(\mu, p_0) = p(S = 1, \tau = 1 | \mu, p_0)$$
$$= \int \sum_{T=0}^{1} p(S = 1, \tau = 1, m, T | \mu, p_0) dm$$

which is estimated using Monte Carlo integration

$$\bar{S}(\mu, p_0) \approx \frac{p(S = 1 | \mu, p_0)}{N_{\mathrm{MC}}} \sum_{i=1}^{N_{\mathrm{MC}}} \sum_{T=0}^{1} \underline{p(\tau = 1 | S_i = 1, m_i, T, \mu, p_0)} p(T | S_i = 1, m_i, \mu, p_0)$$
$$\times \frac{p(m_i | S_i = 1, \mu, p_0)}{p(m_i | S_i = 1, \mu_0, p_{00})}$$

Determined by running
MC realizations through
the pipeline

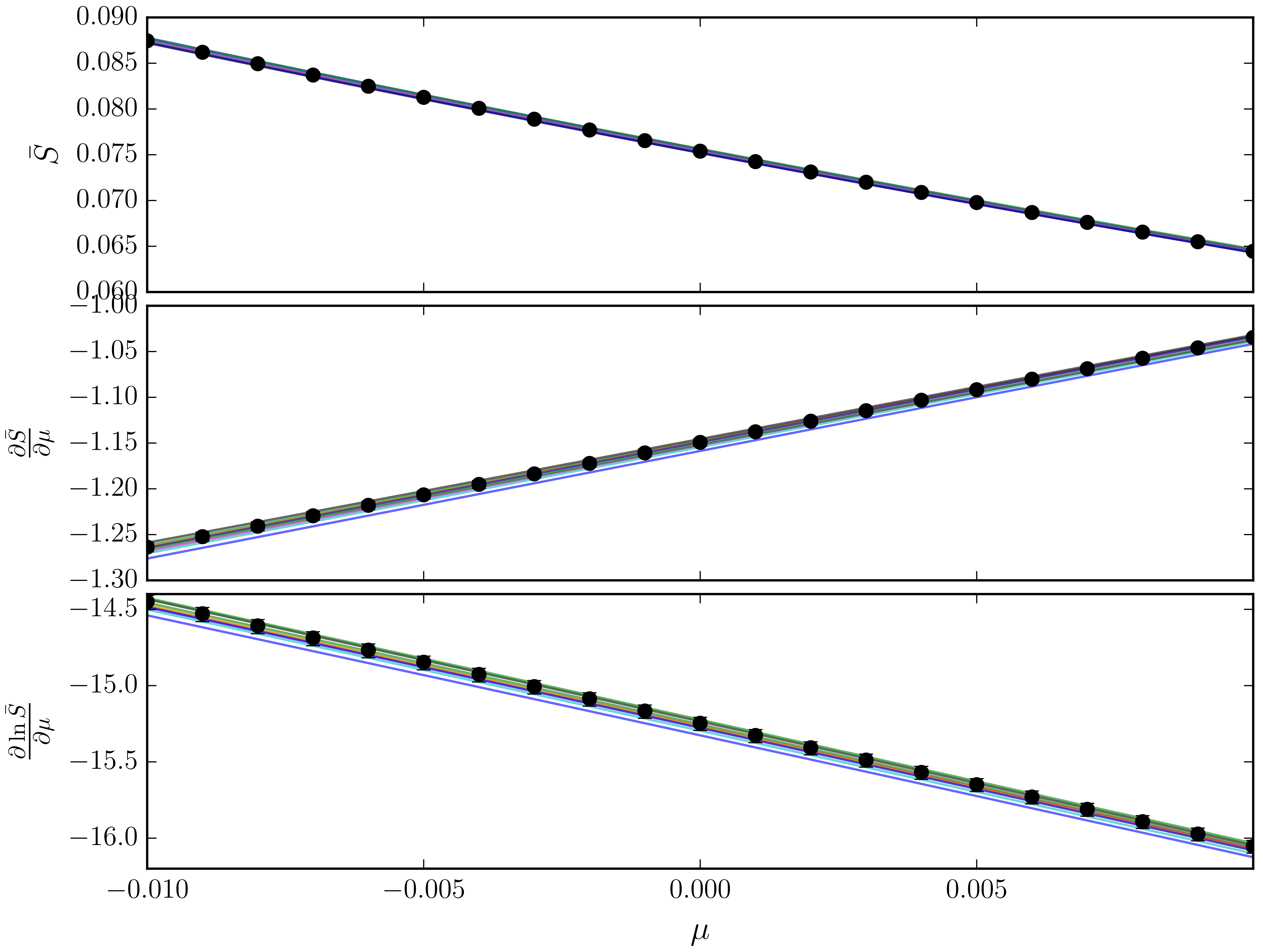
with N draws from $p(m_i | S_i = 1, \mu_0, p_{00})$

Similar expressions for the partial derivatives of S̄

# Numerical Errors in S̄ and Its Partial Derivatives

## Illustrative example case using 10,000 MC Samples

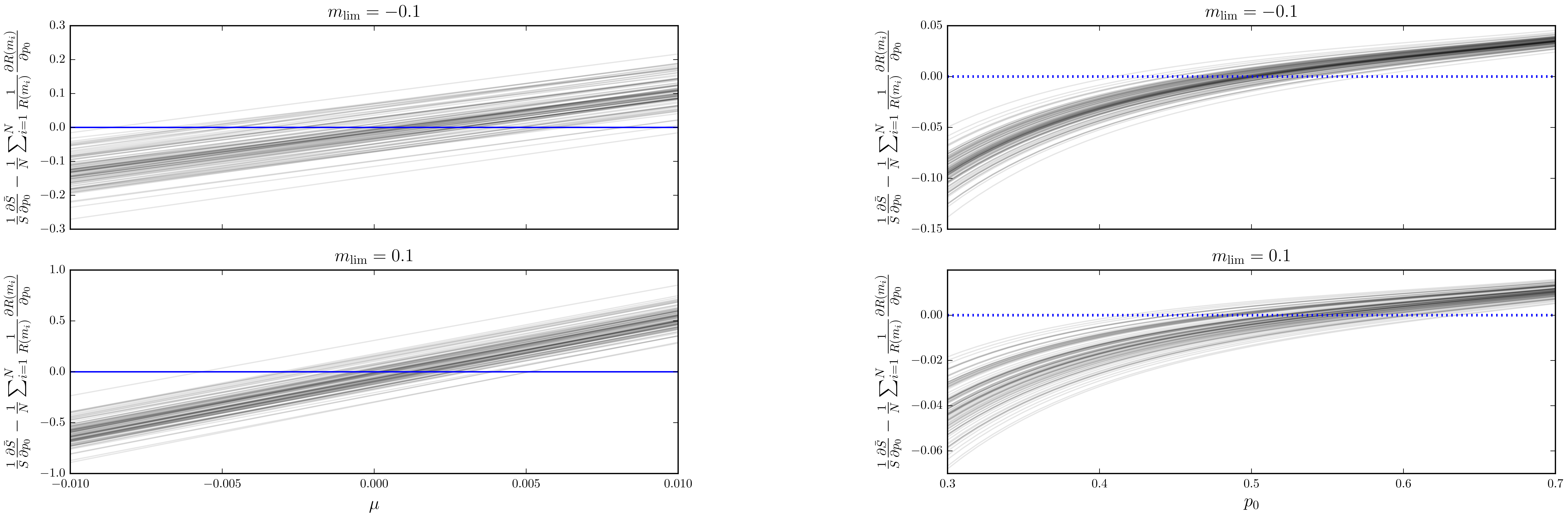
- Calculate S̄ as a function of μ for one MC realization
- Repeat for many MC realizations
- Calculate mean and dispersion or those many realizations
- Plotted are
  - Mean and dispersion as points
  - Several realizations as lines
- To note
  - For one MC realization the errors at different μ are correlated
  - Increasing/decreasing the number of MC samples decrease/increase the dispersion
  - Details of model used in this calculation in article



# Best-Fit Estimators

## Finding the Zero

$$0 = -\frac{1}{\bar{S}}\frac{\partial \bar{S}}{\partial \theta} + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{R(m_i)}\frac{\partial R(m_i)}{\partial \theta}$$

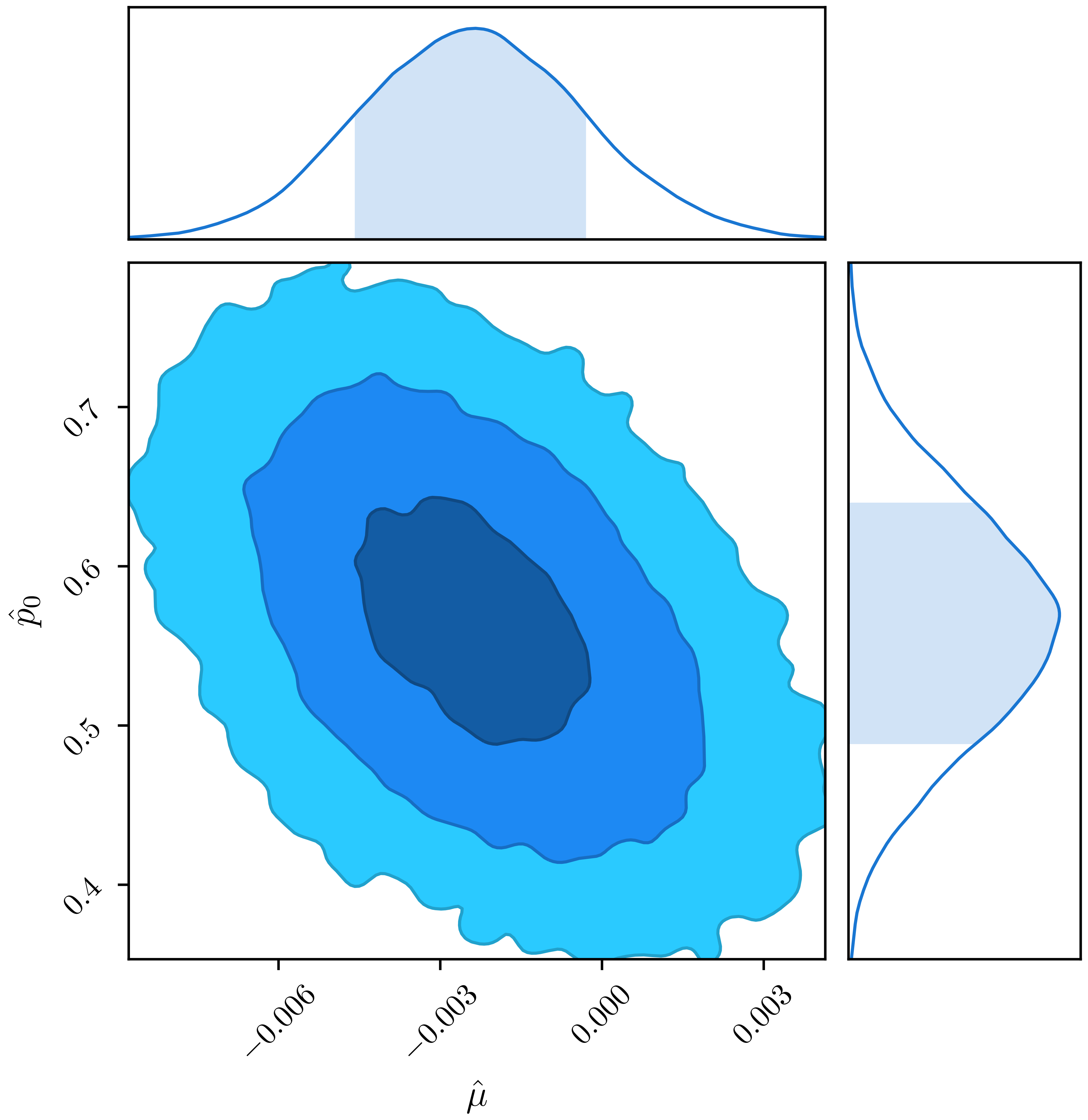



- Each line is the one MC calculation of the function whose zero is the coordinate of maximum likelihood
- Spread in zeros represent estimator errors
  - Spread larger for brighter limiting magnitude
  - Spread is to first order independent of the number of candles and the data that enter the analysis

# Errors in Best-Fit Estimators Due to MC Integration

**Illustrative example case using 10,000 MC Samples**

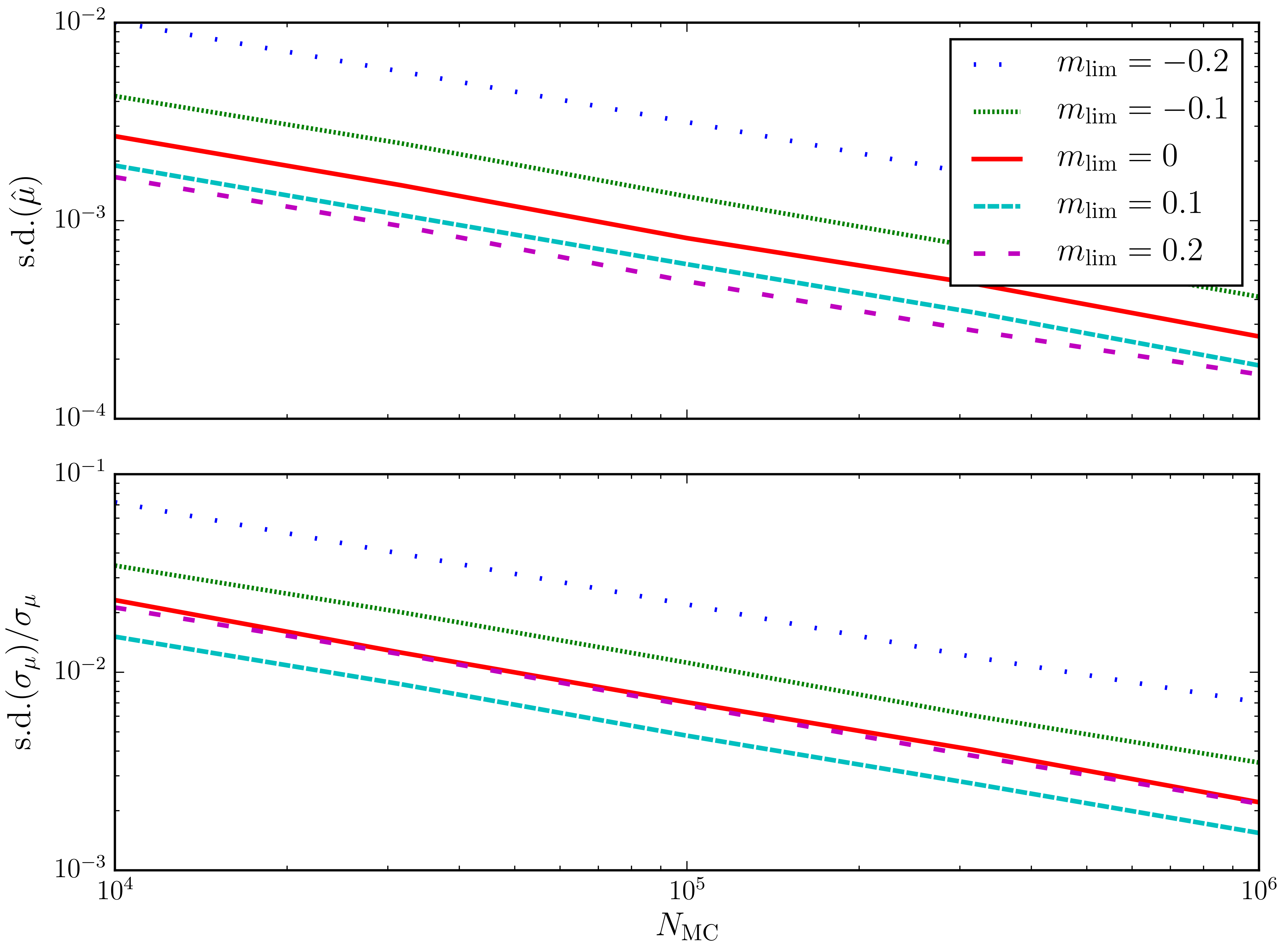
- Distribution of best-fit estimators for many MC realizations of $\bar{S}$
- Errors in the distance modulus and population fraction are correlated



# Dependence of Errors on the Number of MC Samples

## Requirements on Determining the Sample Selection

- Errors in the distance modulus estimator, fractional error in the estimator uncertainty (Hessian) as a function of the number of MC Samples
- Dependent on the magnitude limit
  - The more complete the sample the fewer Monte Carlo samples needed to achieve the same errors
- Scales as $N^{-1/2}$
- Desired precision translates into a computational requirement



# Scaling to Other Models

- Models whose sources have broad magnitude distributions require more MC samples
  - Monte Carlo integration variance goes as the variance of the sampling distribution
- Need only be concerned about sources that can change $\partial\bar{S}/\partial\theta$
  - Integrand also matters
- Models with low selection fraction require more integration precision
  - $\bar{S}$ enters equations as $\bar{S}^{-1}\partial\bar{S}/\partial\theta$

# Further Work

## Gaps to Fill for a Real SN Ia Analysis

- SNe Ia are standardizable candles with intrinsic subparameters
- Classifiers that use magnitude information should cause bias
- Combining spectroscopically confirmed and unconfirmed SNe Ia in one sample with different sample selections
- Examples in the article model intrinsic magnitude distributions as normally distributed
- Real sample selection will be date-dependent
- Sample selection may be stochastic, i.e. $0<S<1$
- Redshift-dependent rates in the model add a level of complexity in the likelihood
- Calculate partial derivatives of $\bar{S}$ for existing classifiers
- $w_o - w_a$ instead of $\mu$

# Conclusions

- For the range of toy models explored in the article and reasonable error targets, <~$10^6$ MC samples are required
- This is the number of real alerts a broker processes in one night
- No alarm bells yet for being able to process simulated data
- Cosmology analysis can occur 10+ years after sample selection
- Concluding Requirement: Containerize the pipeline state for future analysis on DESC computers