Introduction to statistics

Maria Süveges

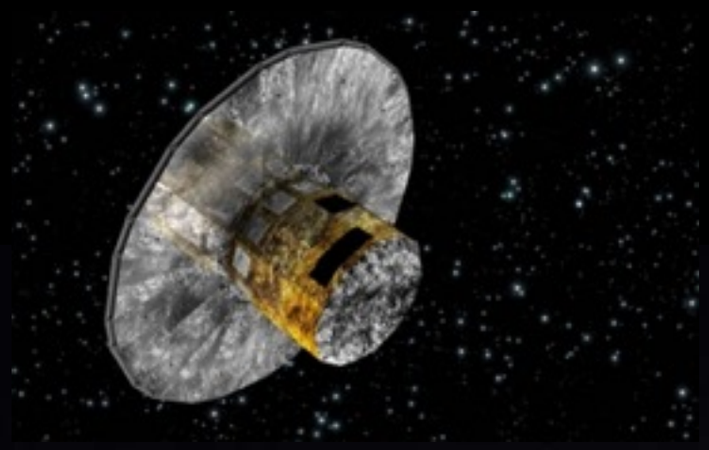
University of Geneva
Switzerland

LSSTC DSFP
Northwestern University, Evanston
August 1-5, 2016

Exercises:

https://github.com/LSSTC-DSFP/LSST-DSFP-Resources/tree/master/Session1/Tuesday

The need for statistics

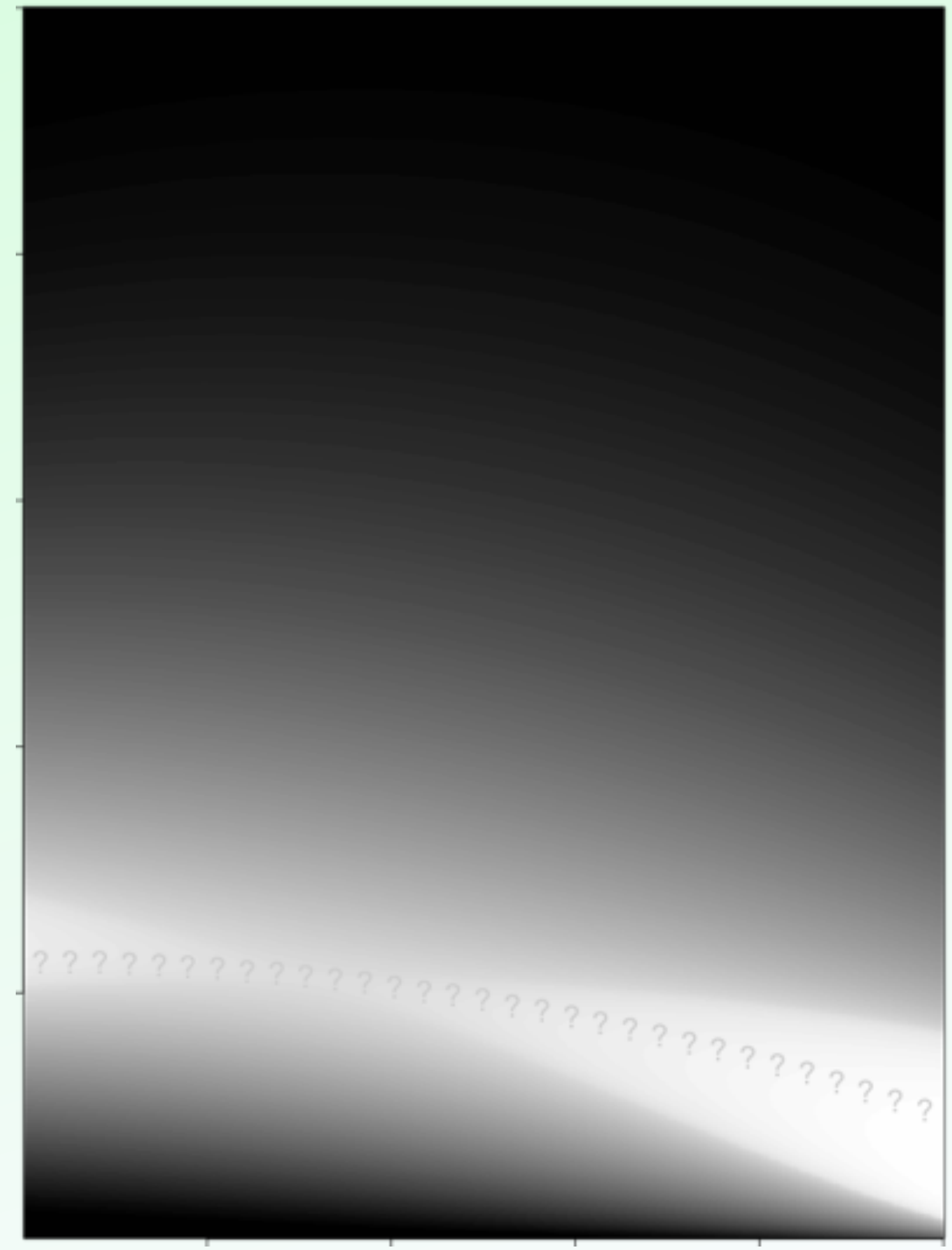


The need for statistics

The need for statistics

Astronomers' view:

Statisticians' view:



Outline

- Reminder: probabilities
- Random variables and distributions
- Exploratory techniques
  - Tools
  - Recognizing distributions and discrepancies
- Estimation
  - Method of moments
  - Robust methods
  - Likelihood and maximum likelihood estimation
- Classical hypothesis testing
- Model selection

Probabilities

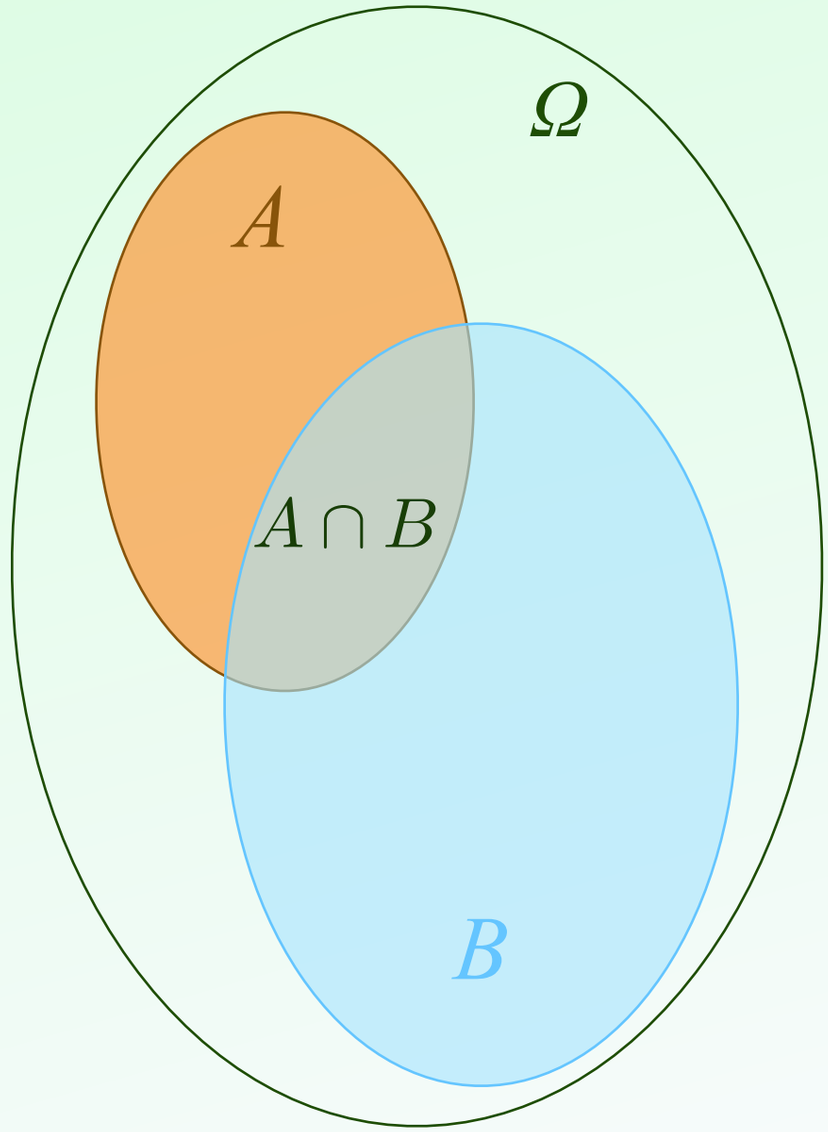
Probability: definition

Axioms (Kolmogorov):

Let Ω be a collection of possible elementary events, and A and B events such that $A, B \in \Omega$.

1. $P(A) \geq 0$ for all A;
2. $P(\Omega) = 1$;
3. $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

   for all countable disjoint sets $A_1, A_2, \ldots \in \Omega$



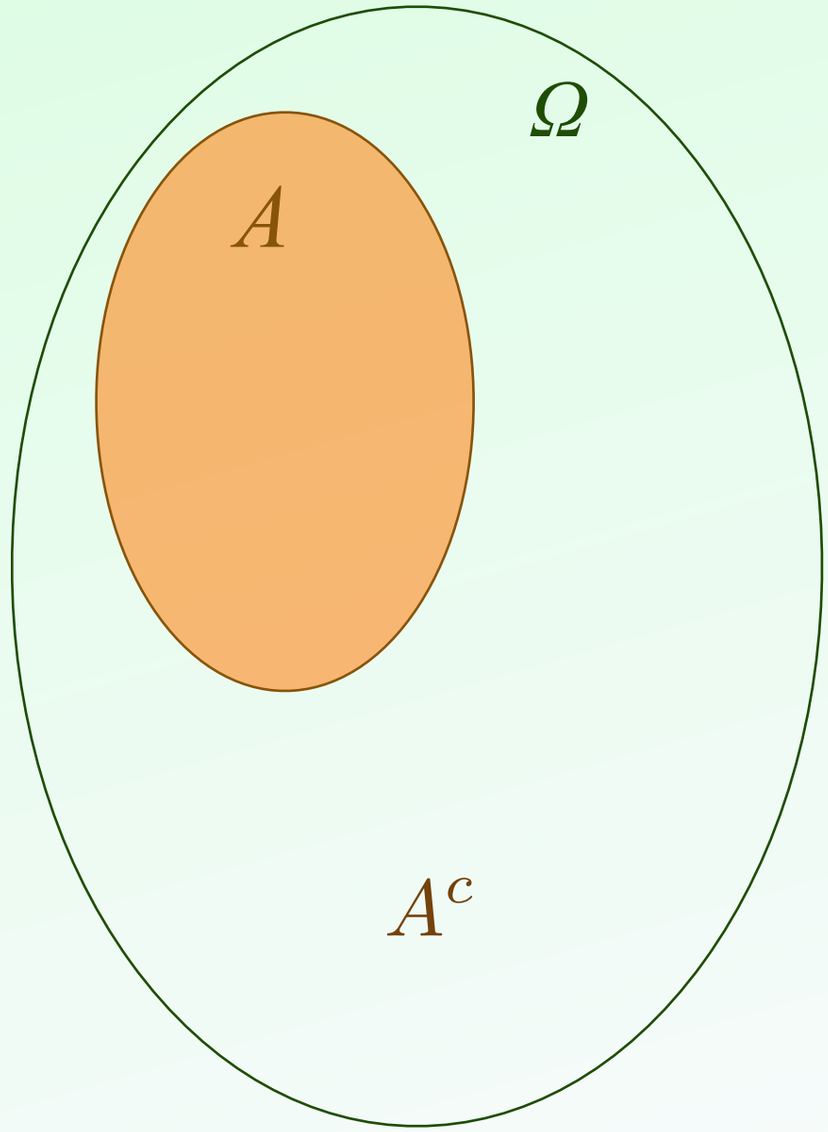
Probability: definition

Axioms (Kolmogorov):

Let Ω be a collection of possible elementary events, and events A and B such that $A, B \in \Omega$

1. $P(A) \geq 0$ for all A;
2. $P(\Omega) = 1$;
3. $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

   for all countable disjoint sets $A_1, A_2, \ldots \in \Omega$



Some consequences:

$$P(A) + P(A^c) = 1$$

Probability: definition

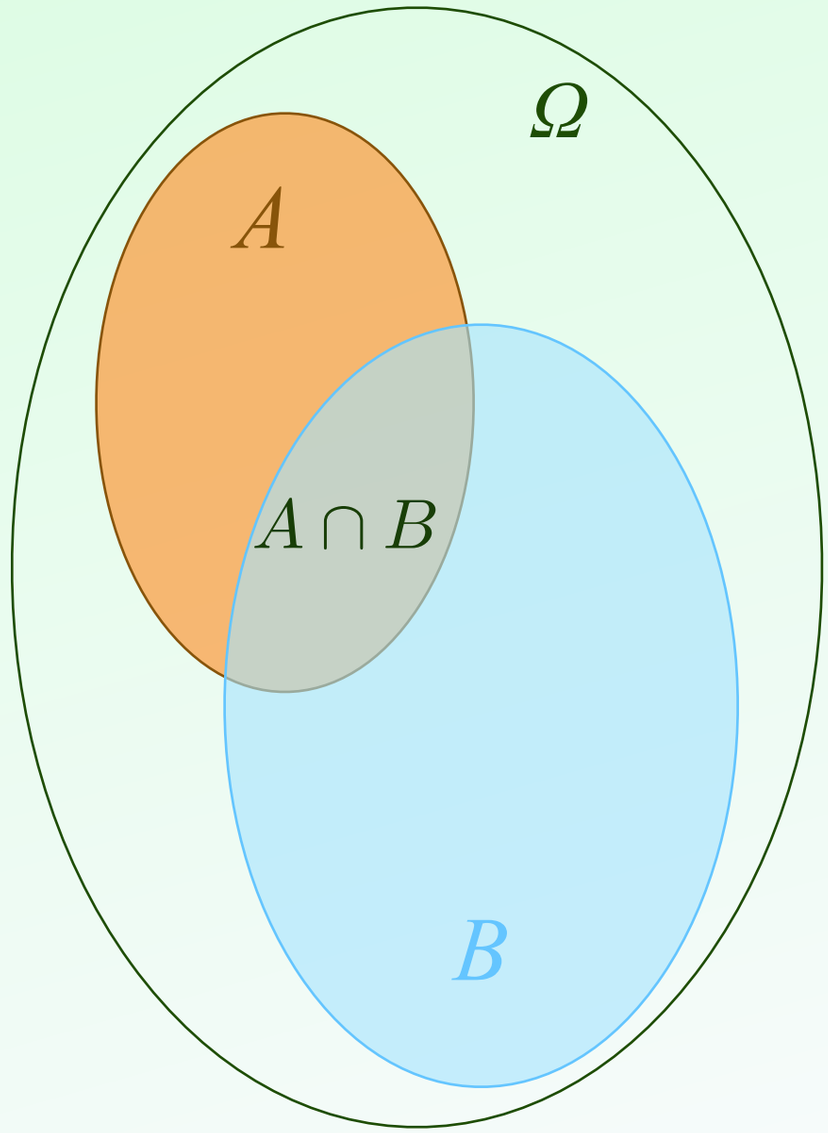
Axioms (Kolmogorov):

Let Ω be a collection of possible elementary events, and A and B events such that $A, B \in \Omega$.

1. $P(A) \geq 0$ for all A;
2. $P(\Omega) = 1$;
3. $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

   for all countable disjoint sets

   $A_1, A_2, \ldots \in \Omega$



Some consequences:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

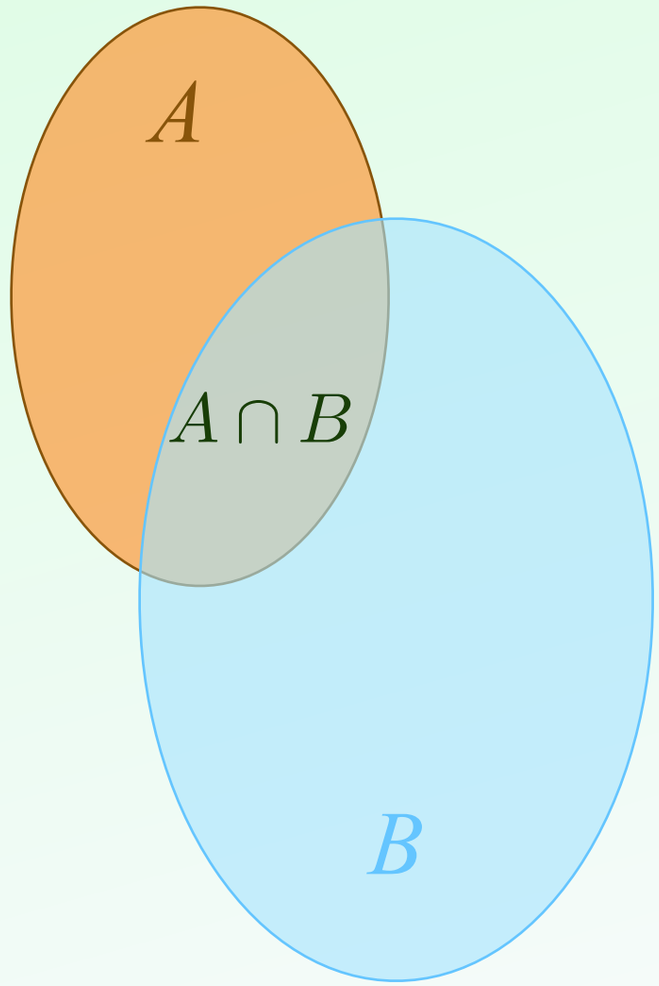
Probability: definition

Axioms (Kolmogorov):

Let Ω be a collection of possible elementary events, and A and B events such that $A, B \in \Omega$.

1. $P(A) \geq 0$ for all A;
2. $P(\Omega) = 1$;
3. $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

   for all countable disjoint sets $A_1, A_2, \ldots \in \Omega$



Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Probability: definition

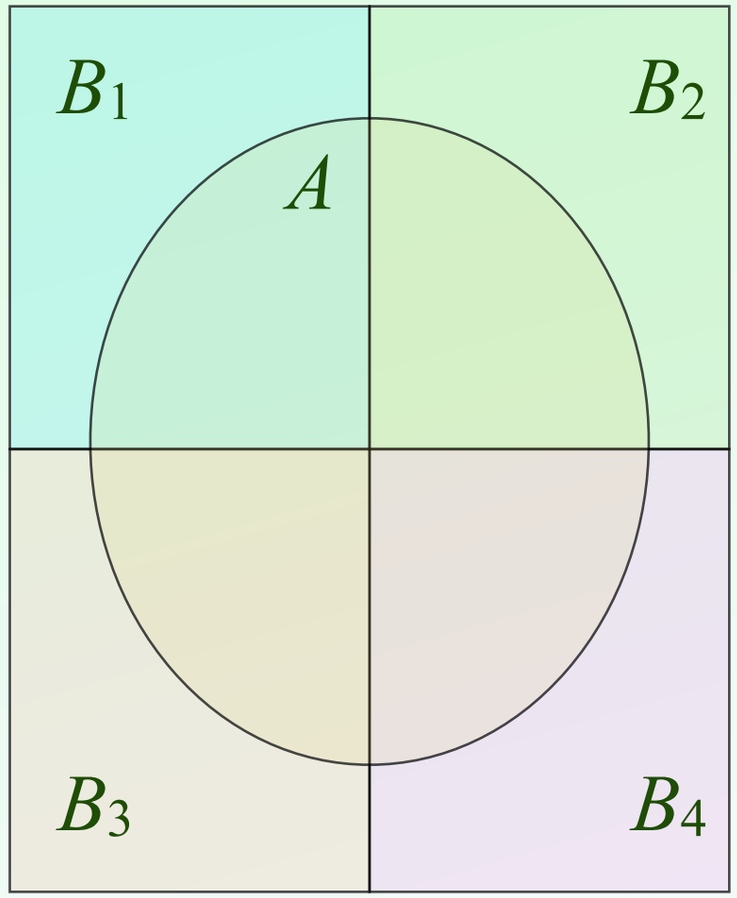
Axioms (Kolmogorov):

Let Ω be a collection of possible elementary events, and A and B events such that $A, B \in \Omega$.

1. $P(A) \geq 0$ for all A;
2. $P(\Omega) = 1$;
3. $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

   for all countable disjoint sets

   $A_1, A_2, \ldots \in \Omega$



Some consequences (law of total probability):

$$P(A) = \sum_{i=1}^{N} P(A|B_i)P(B_i)$$

Random variables and their distribution

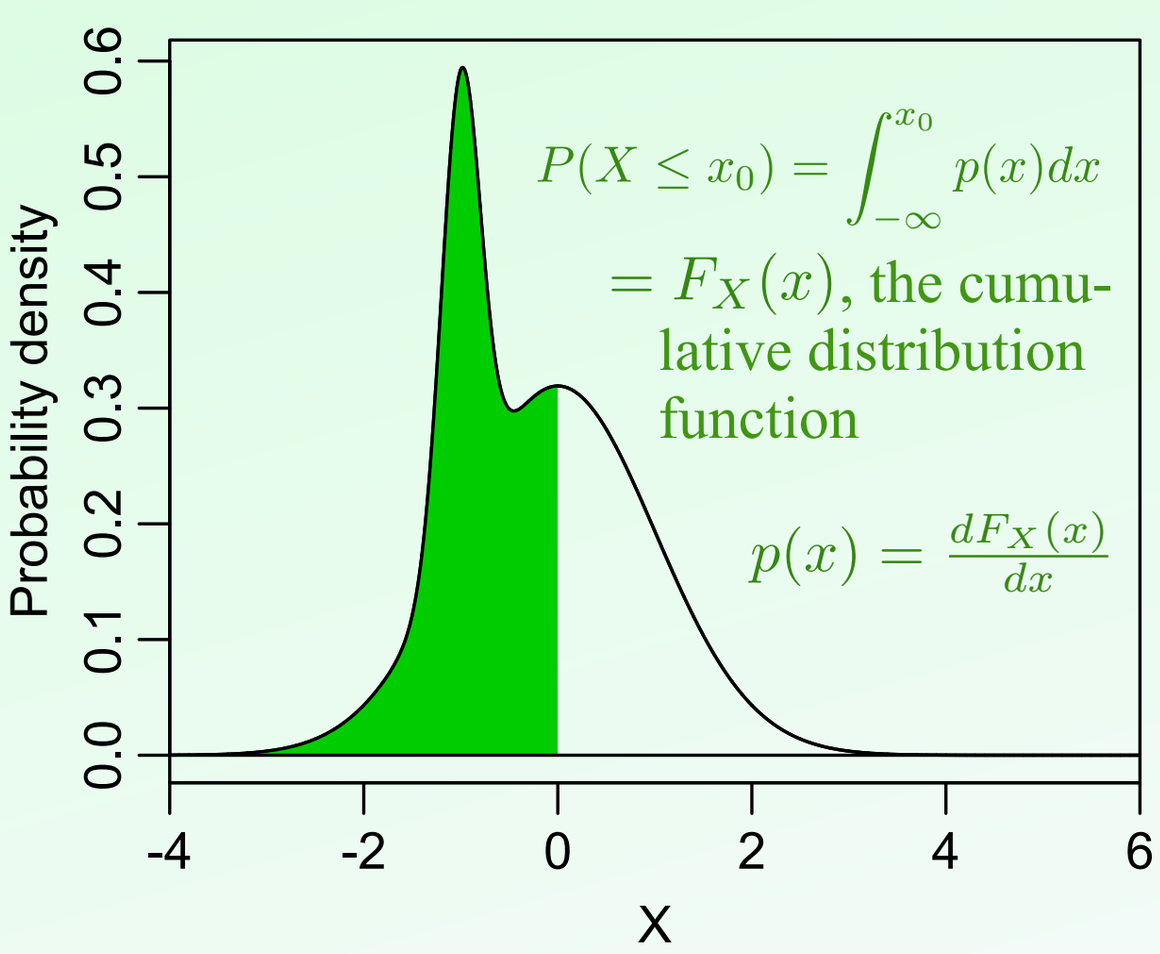
Probability: random variables

Random variable:

the outcome of an "experiment", with a probability for each outcome

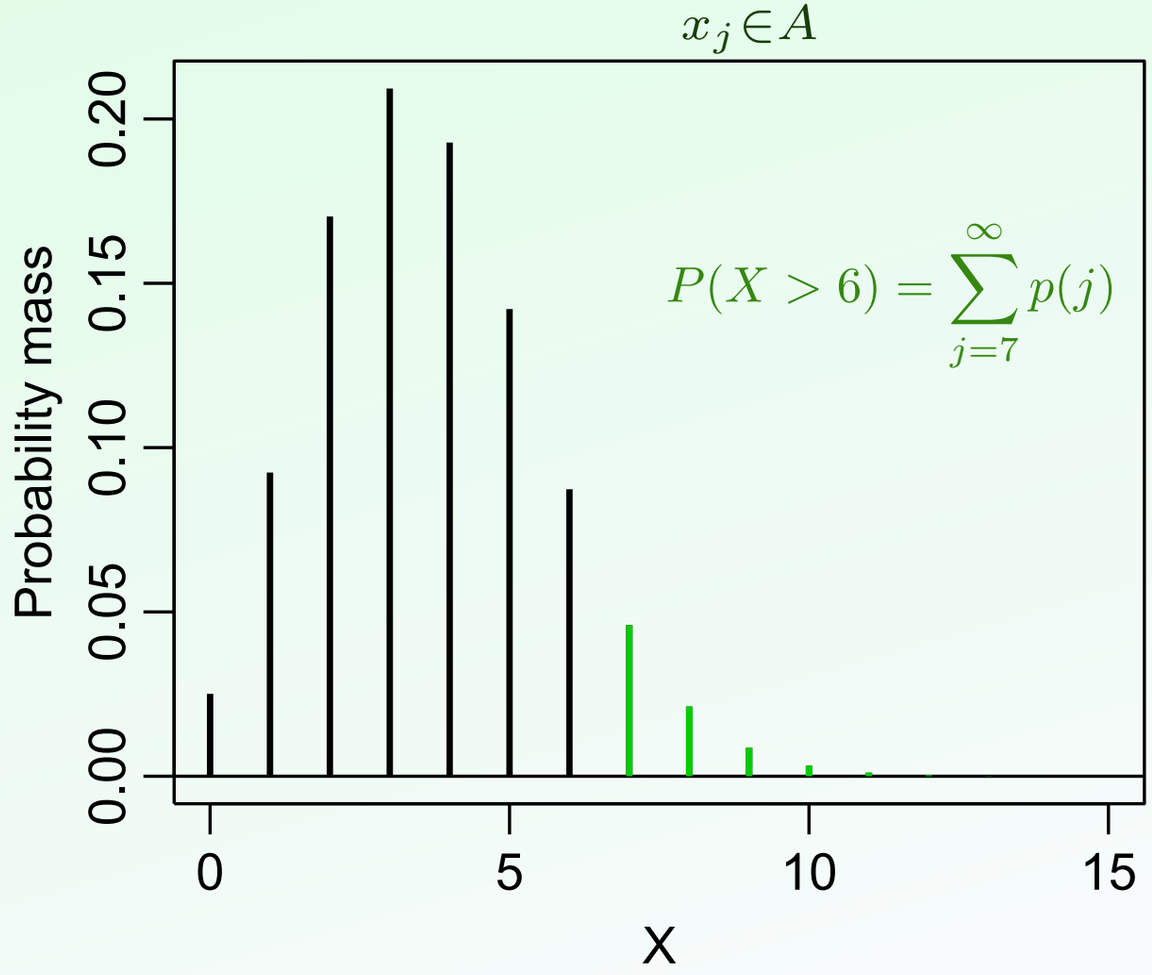
Continuous

$$P(X \in A) = \int_A p(x)dx$$



Discrete

$$P(X \in A) = \sum_{x_j \in A} p(x_j)$$



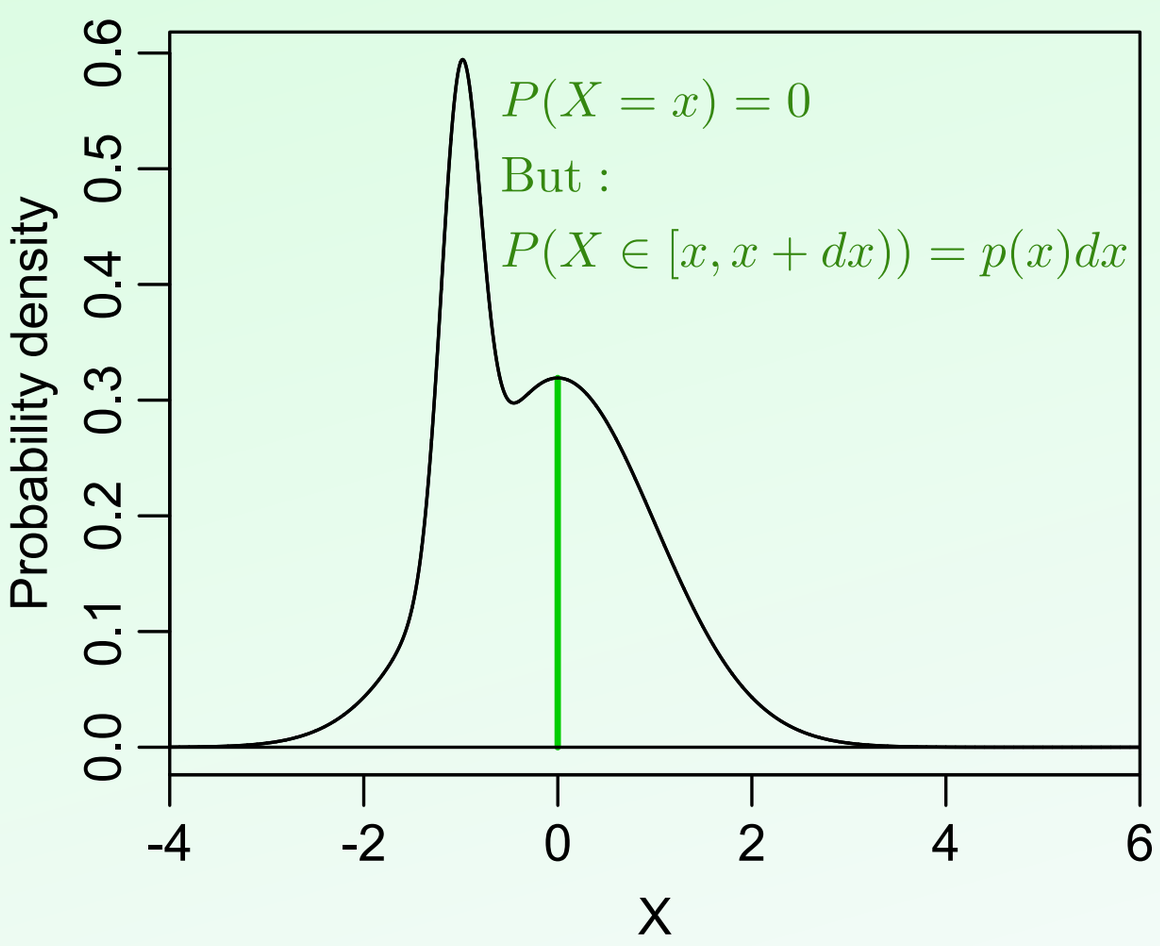
Probability: random variables

Random variable:

the outcome of an "experiment", with a probability for each outcome

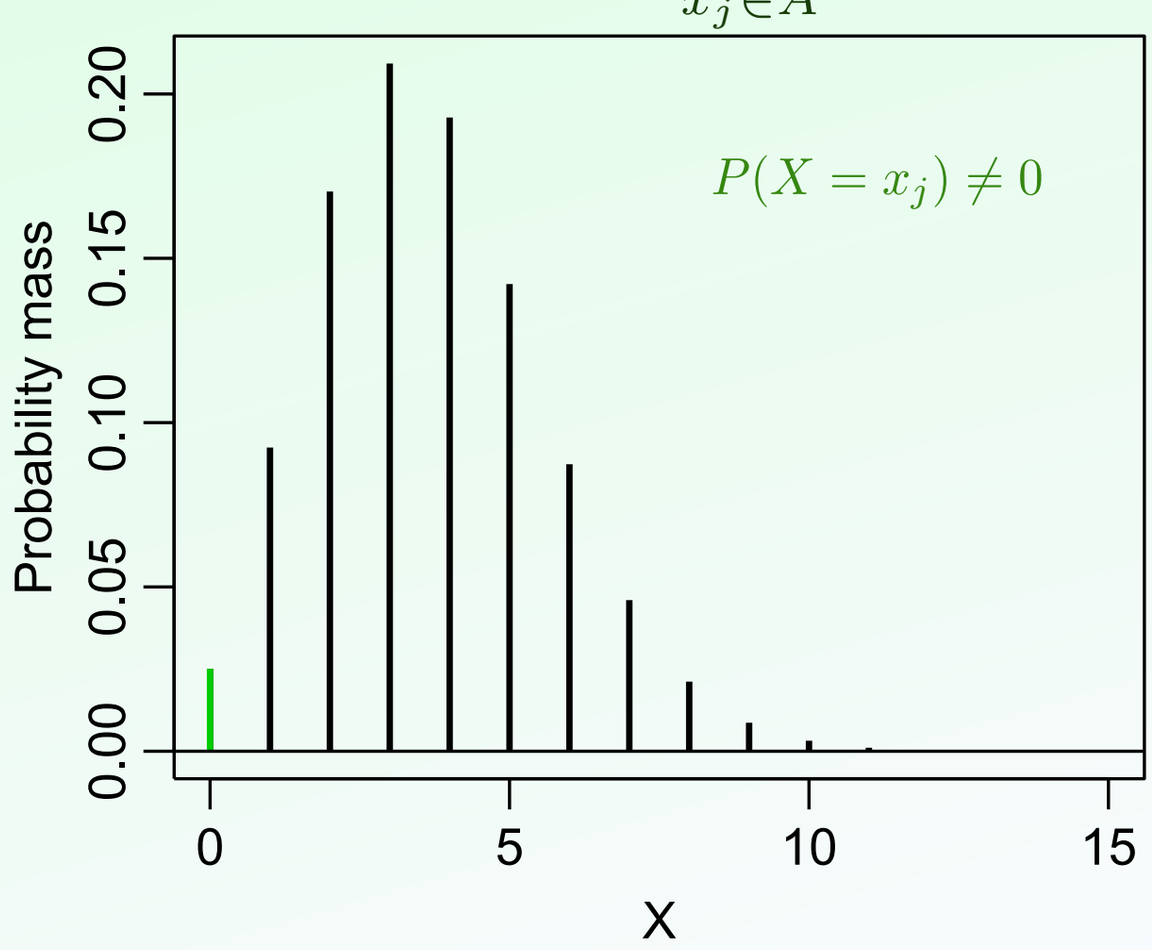
Continuous

$$P(X \in A) = \int_A p(x)dx$$



Discrete

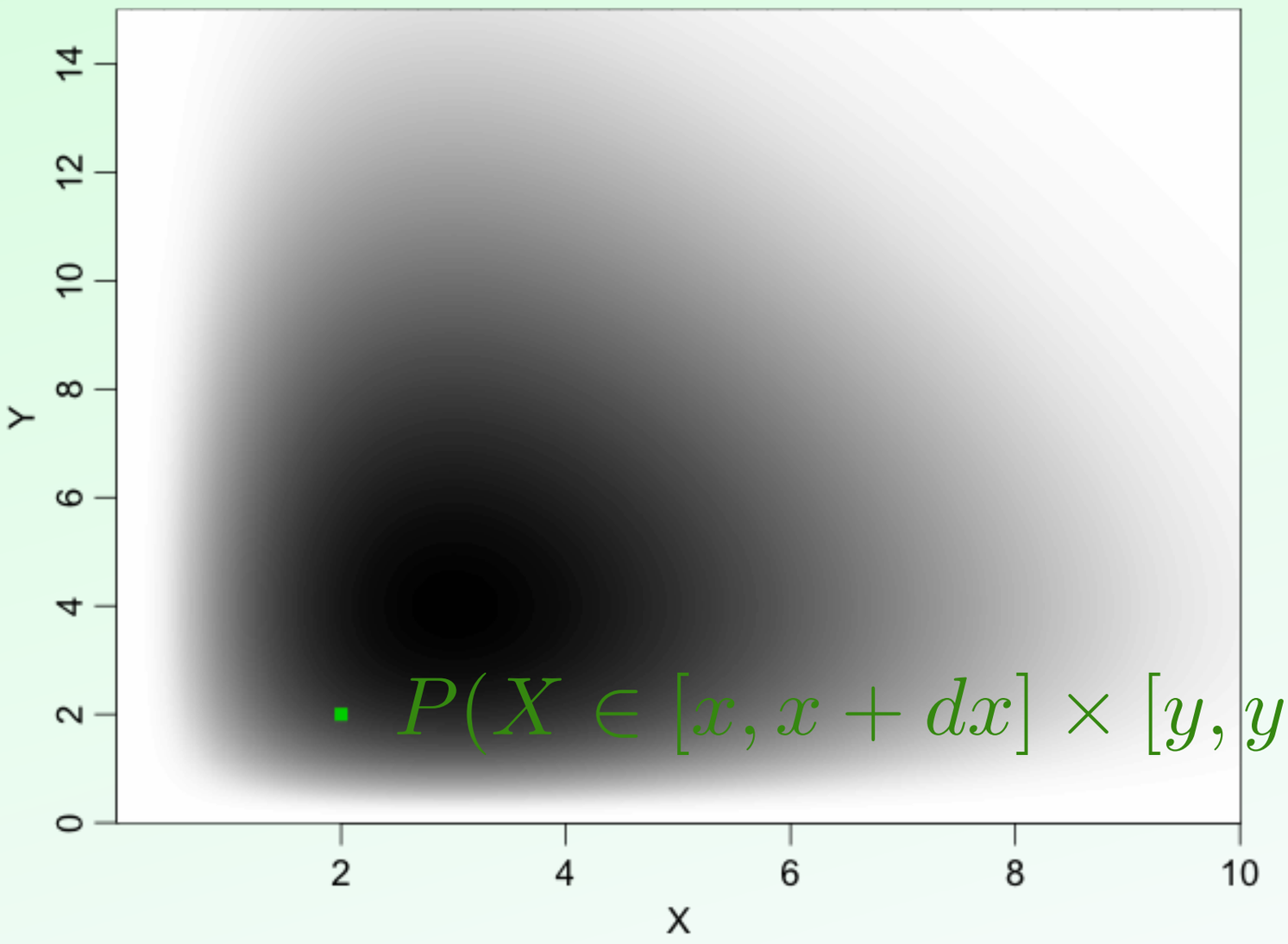
$$P(X \in A) = \sum_{x_j \in A} p(x_j)$$



Probability: multivariate

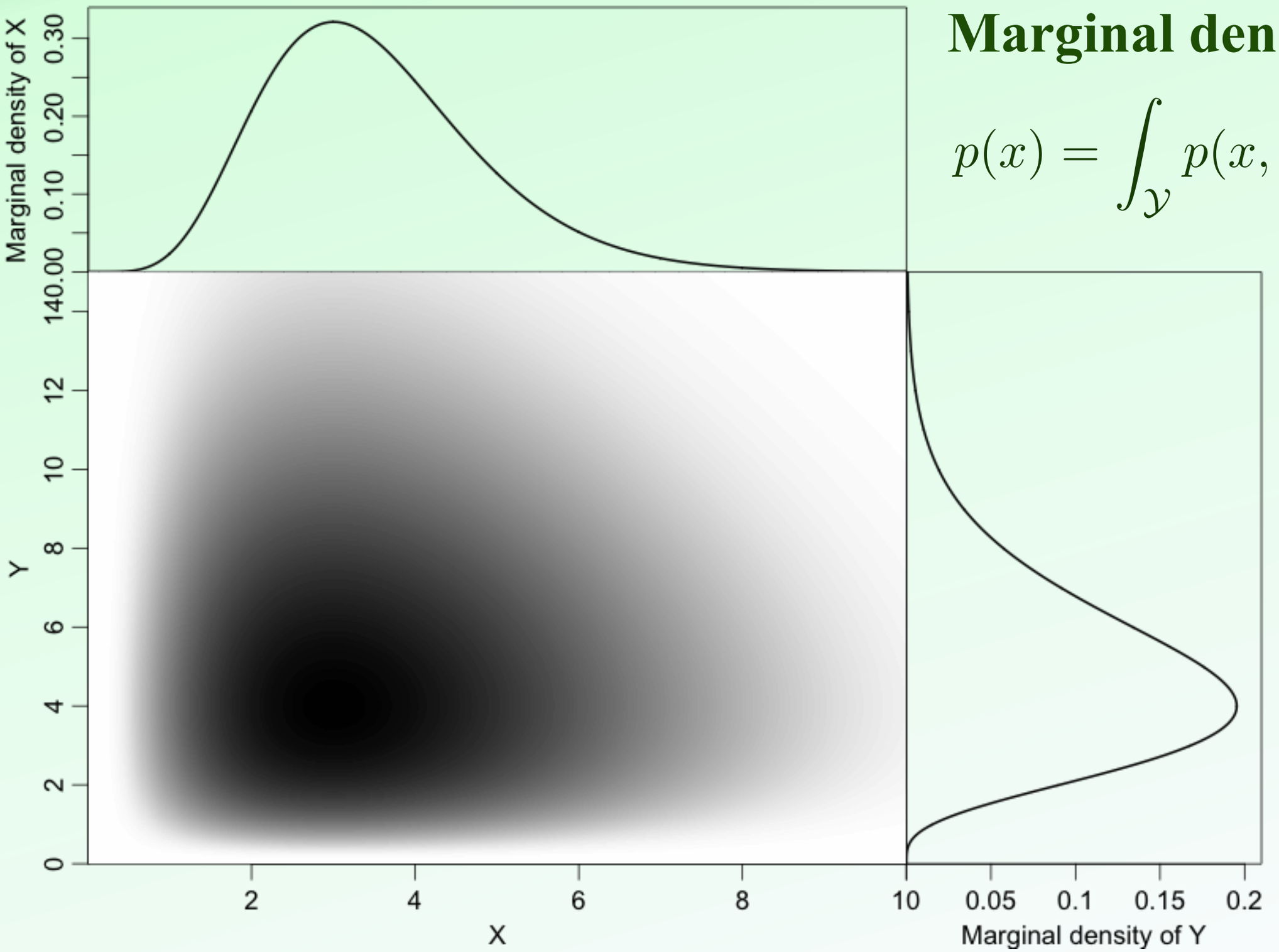
Probability of *X* falling in set *A*:

$$P(X \in A) = \int_A p(x_1, x_2, \ldots, x_D) dx_1 dx_2 \ldots dx_D$$



- $P(X \in [x, x + dx] \times [y, y + dy]) = p(x, y) dx dy$

Probability: multivariate



Marginal density:

$$p(x) = \int_{\mathcal{Y}} p(x, y) dy$$

Probability: conditional

Formal definition:

$$p(y \mid x) = \frac{p(x, y)}{p(x)}$$

Law of total probability:

$$p(x) = \int_{\mathcal{Y}} p(x \mid y) p(y) dy$$

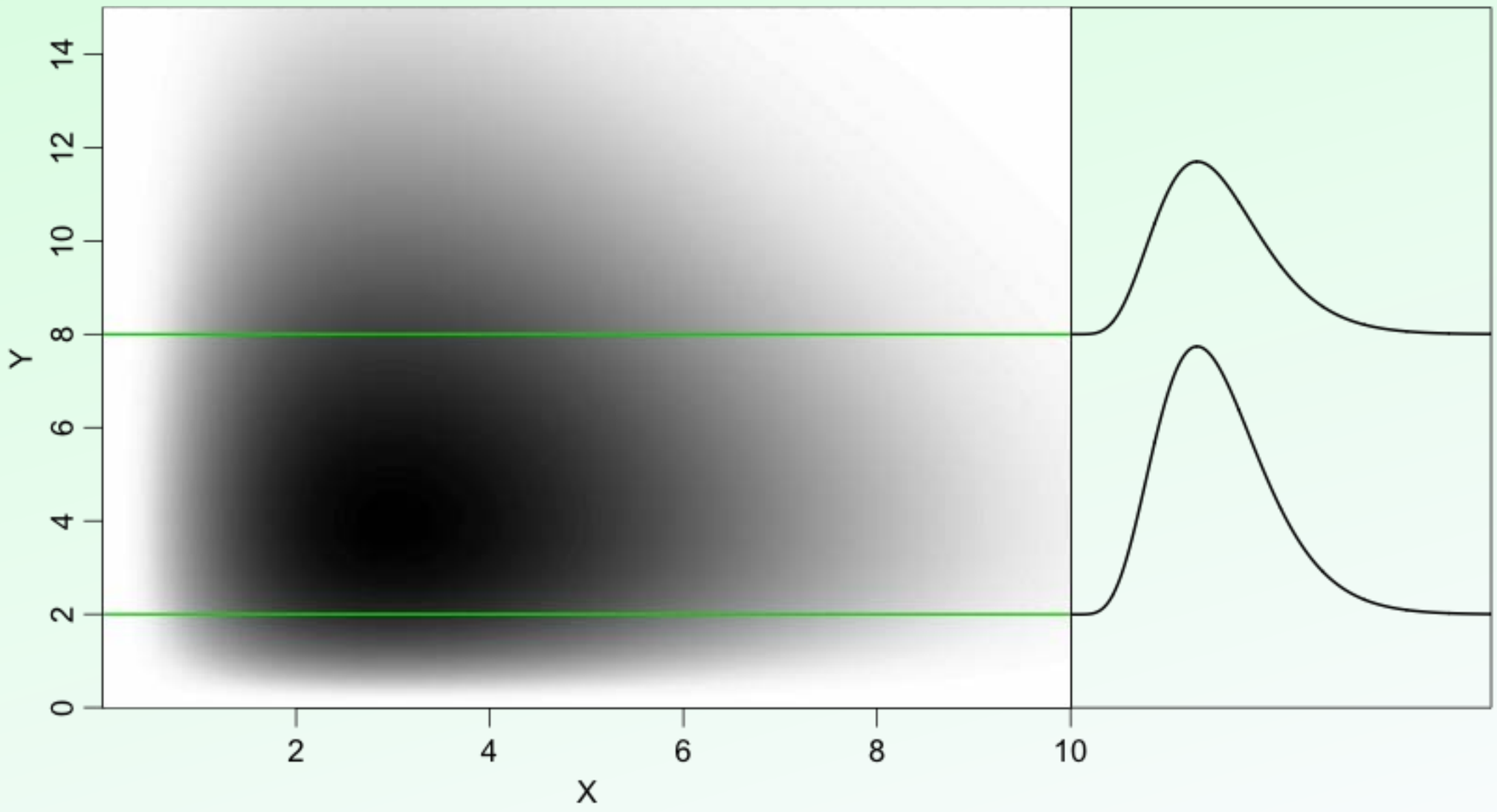
Bayes rule:

$$p(y \mid x) = \frac{p(x, y)}{p(x)} = \frac{p(x \mid y) p(y)}{p(x)} = \frac{p(x \mid y) p(y)}{\int_{\mathcal{Y}} p(x \mid y) p(y) dy}$$

Probability: conditional

Conditional probability (intuitively):

$$p(x \mid y = y_0) = p(x, y_0)$$



Moments of distributions

Definitions:

- **Expected value (mean)**

$$\mathrm{E}(X) = \int_{\mathcal{X}} xp(x)dx$$
$$\mathrm{E}(f(X)) = \int_{\mathcal{X}} f(x)p(x)dx$$

- **Variance**

$$\mathrm{Var}(X) = E\left([X - E(X)]^2\right)$$

- **Moments**
  - ⋆ **noncentral:**

$$\mu_r(X) = \mathrm{E}\left(X^r\right)$$

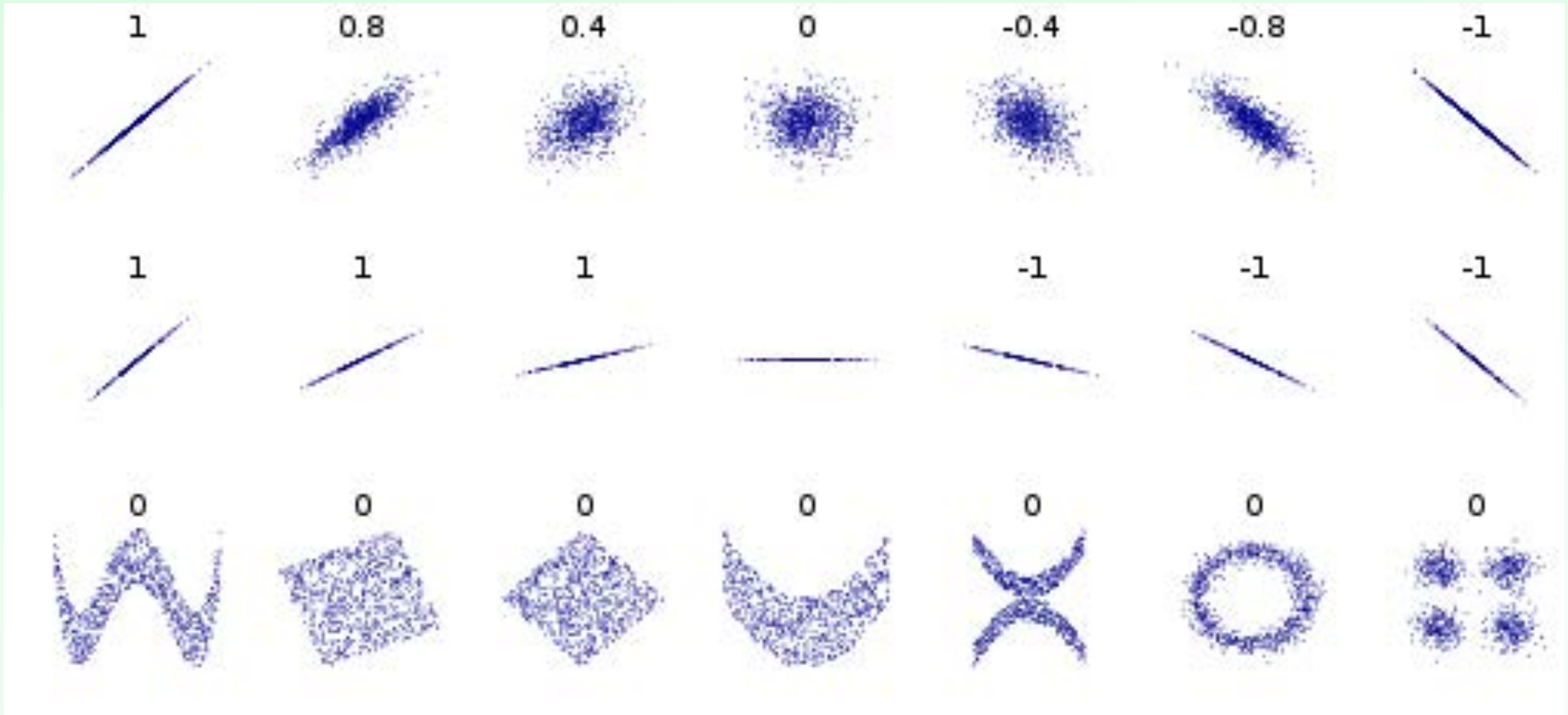
  - ⋆ **central:**

$$\tilde{\mu}_r(X) = \mathrm{E}\left([X - \mathrm{E}(X)]^r\right)$$

Covariance and correlation

Definitions:

- **Covariance:** $\mathrm{Cov}(X,Y) = \mathrm{E}[X - \mathrm{E}(X)]\mathrm{E}[Y - \mathrm{E}(X)]$
- **Correlation:** $\mathrm{Corr}(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}Y)}}$



Important distributions: Poisson

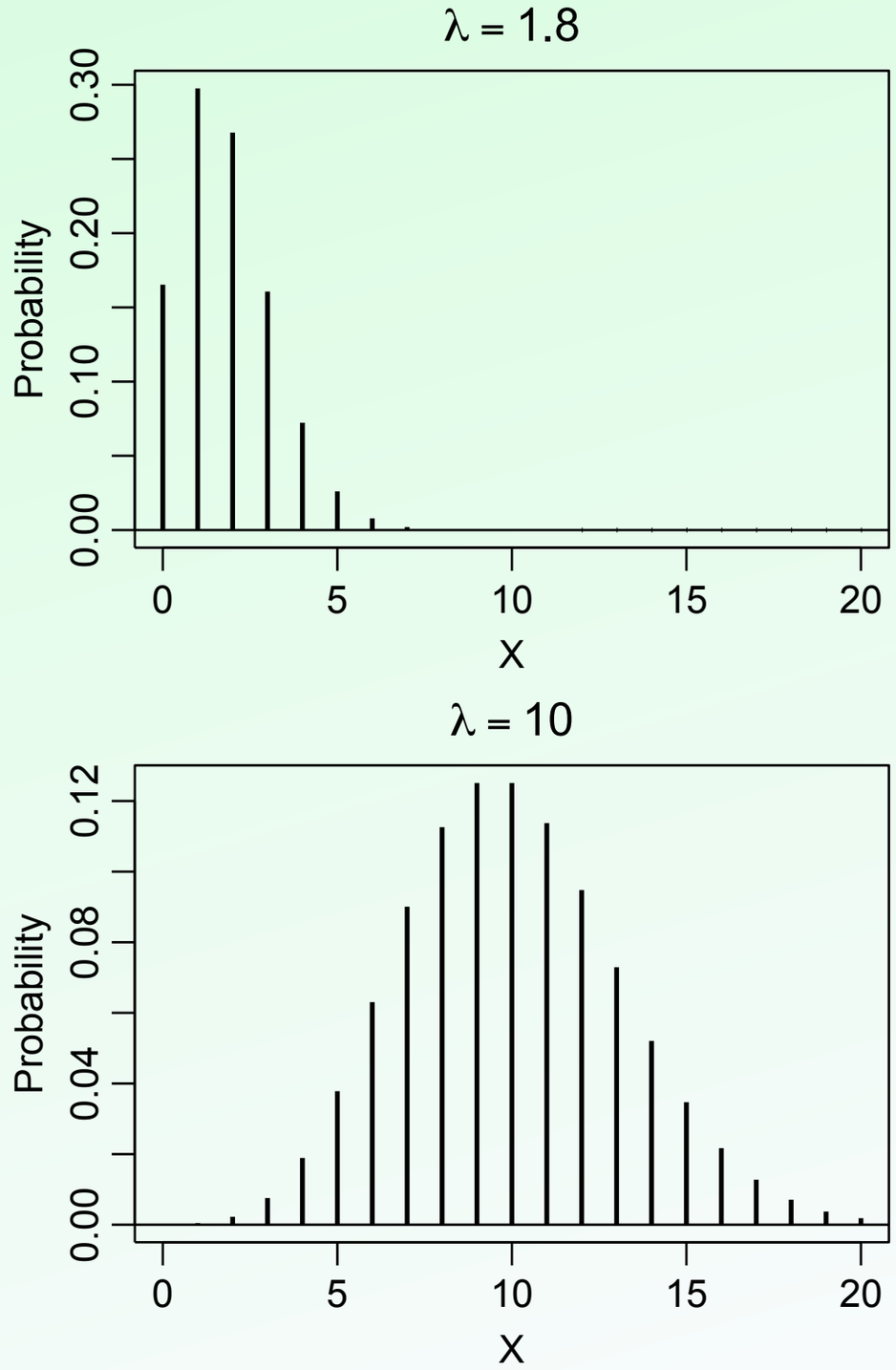
Probability distribution:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\lambda \in \mathbb{R}^+,\ k \in \mathbb{N}$$

The distribution for counts of events (satisfying some conditions…)

Mean	λ
Mode	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
Median	$\approx \left\lfloor \lambda + \frac{1}{3} - \frac{0.02}{\lambda} \right\rfloor$
Variance	λ



Important distributions: binomial

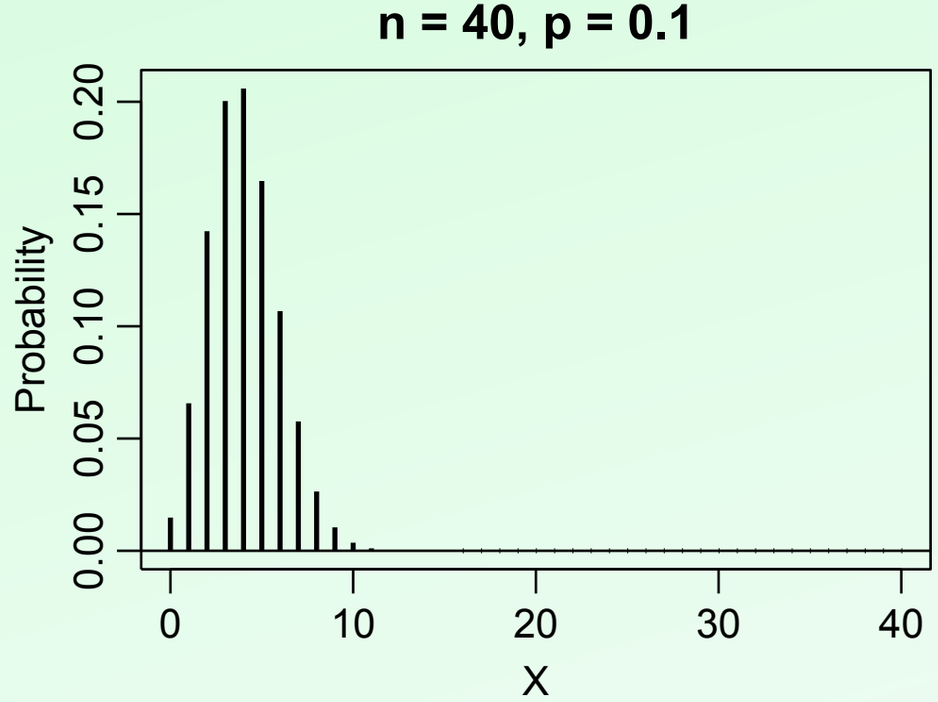
Probability distribution:

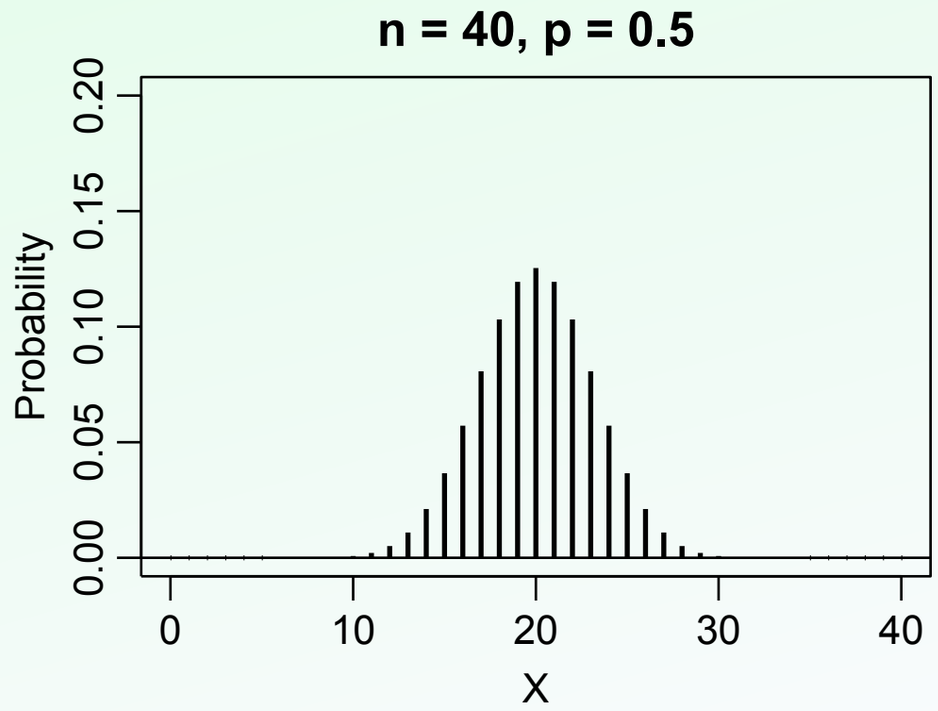
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$p \in [0,1],\ n \in \mathbb{N},\ k = 1, 2, \ldots, n$$

Example: tossing a dice and getting 6

Mean	np
Mode	$\lfloor np \rfloor$ or $\lceil np \rceil$
Median	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p - 1 \rceil$
Variance	$np(1-p)$





Important distributions: Gaussian (normal)

Density:

$$\phi(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

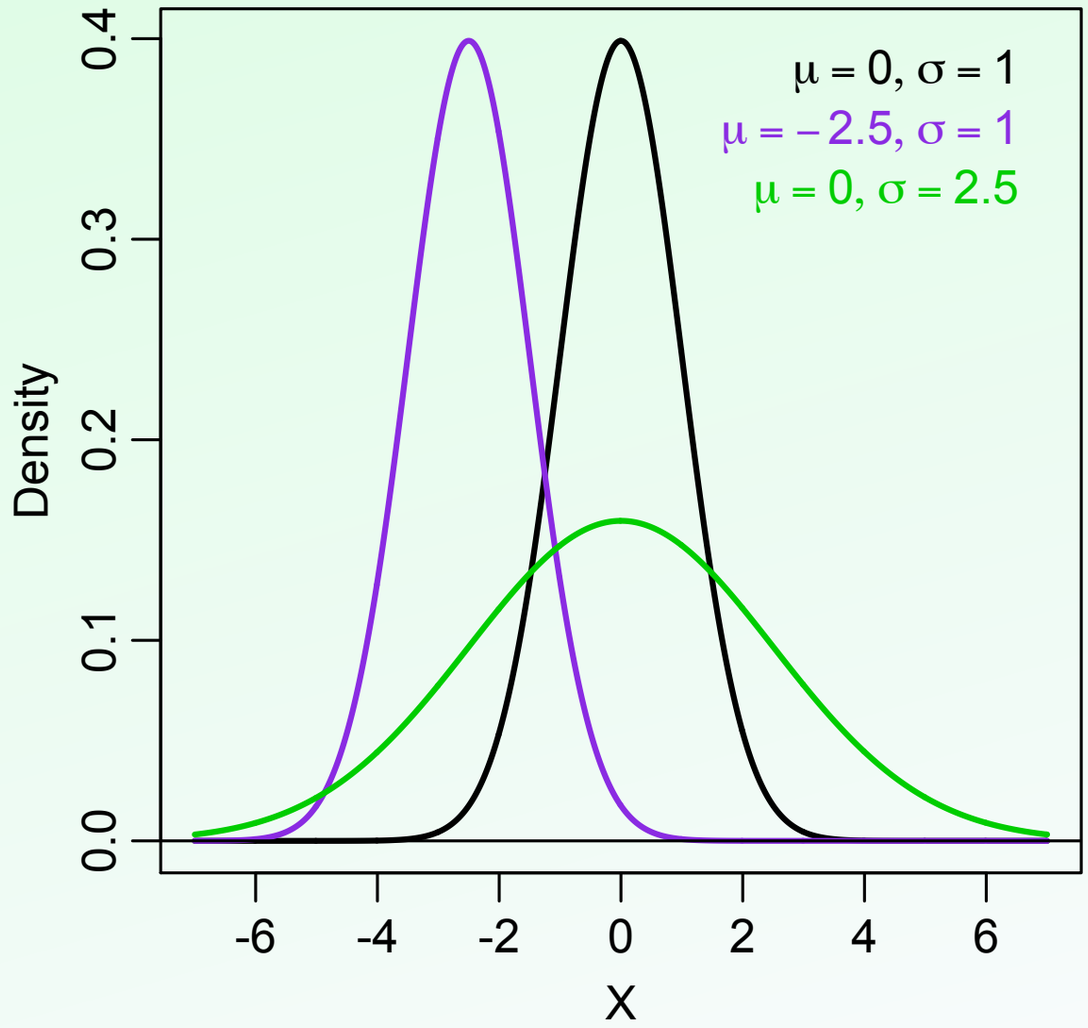
$$\mu \in \mathbb{R},\ \sigma > 0$$

Probably the most used distribution.

Standard normal variables:

$$\mu = 0,\ \sigma = 1$$

Mean	μ
Mode	μ
Median	μ
Variance	σ



Important distributions: χ^2_ν

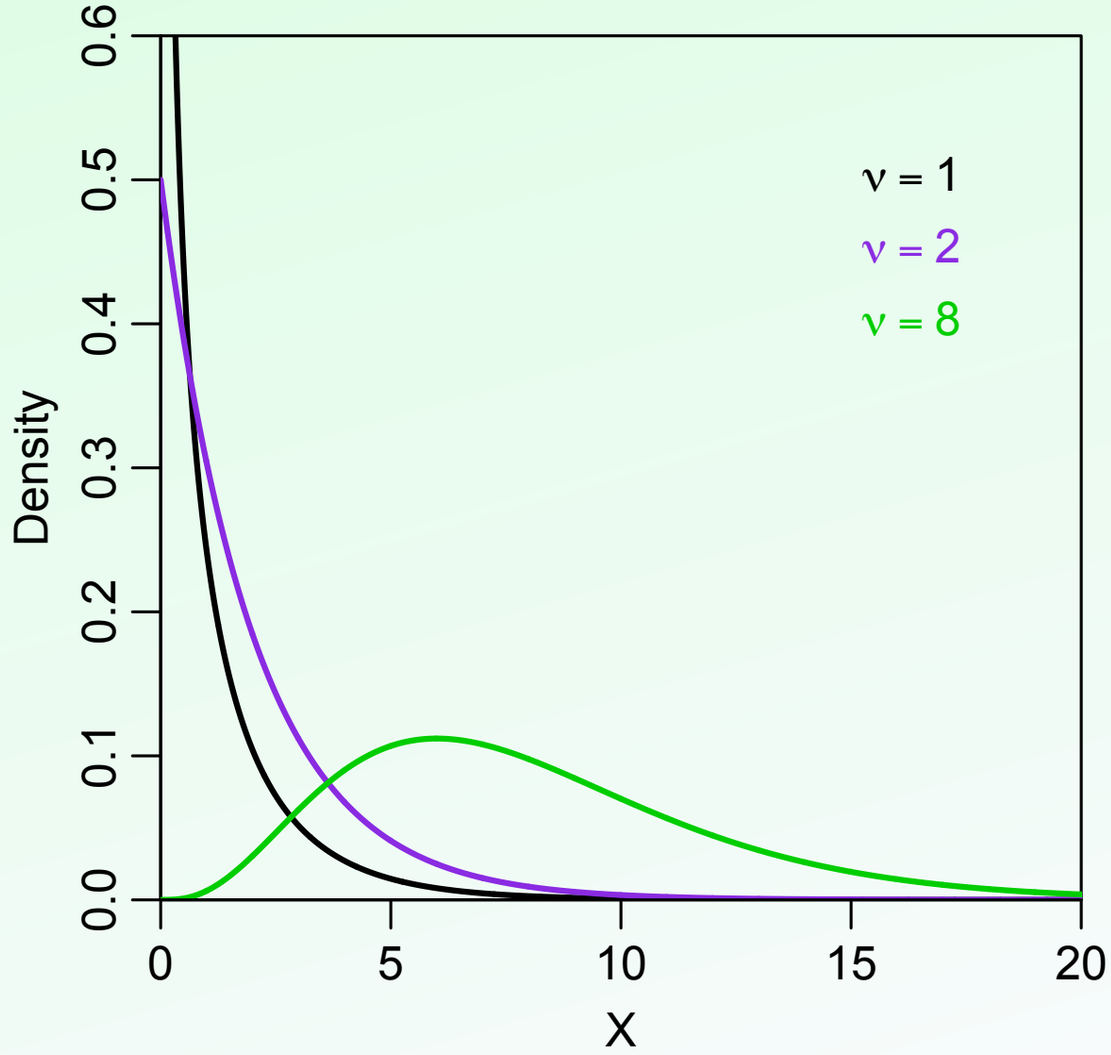
Density:

$$p(x;\nu) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}$$

$$\nu = 1, 2, \ldots$$

The distribution of the sum of the squares of ν independent standard normal variables

Mean	$\boldsymbol{\nu}$
Mode	$\max\{0, \nu - 2\}$
Median	$\approx \nu\left(1 - \frac{2}{9\nu}\right)$
Variance	2ν



Important distributions: t_ν

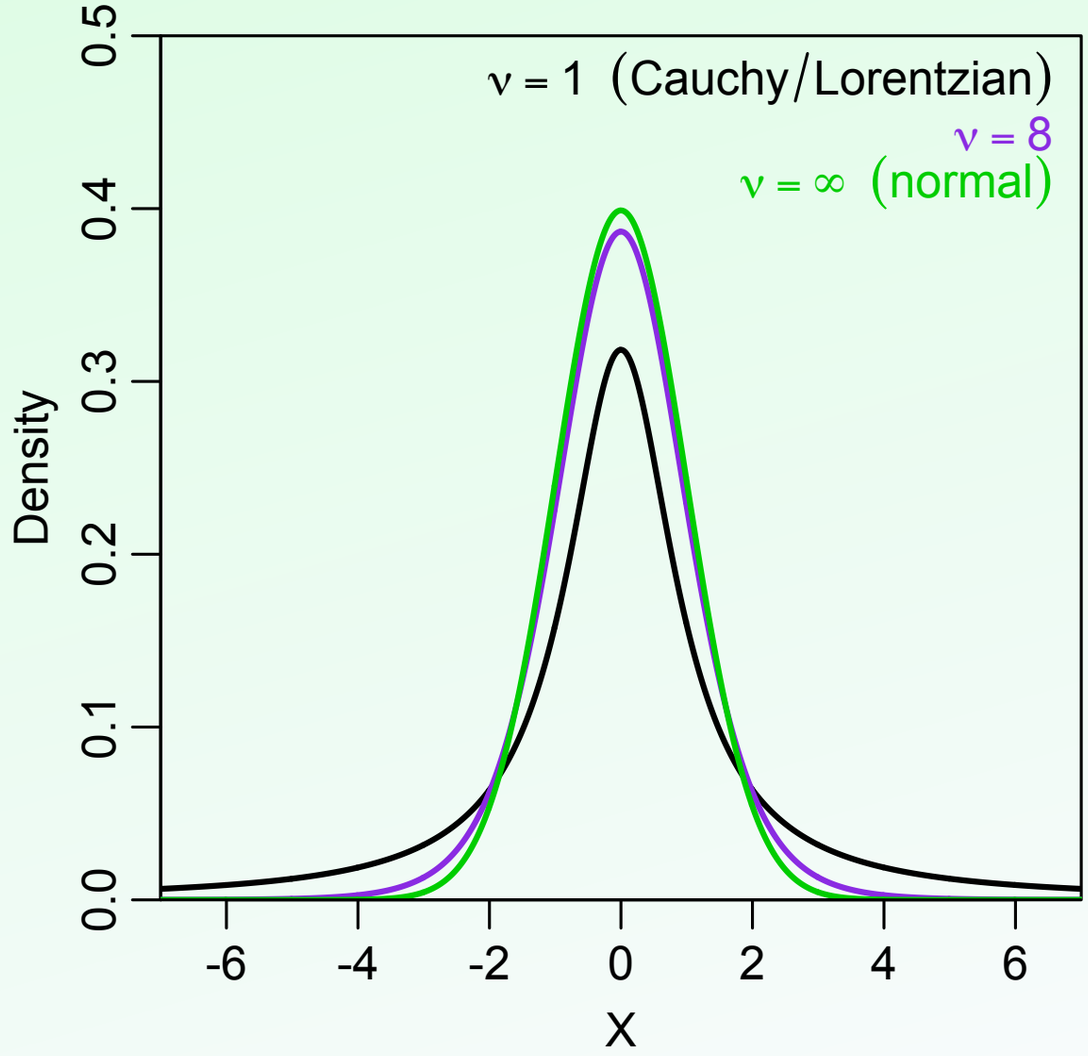
Density:

$$p(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$\nu \in \mathbb{R}^+,\ x \in \mathbb{R}$

The distribution of the estimated coefficients in linear regression (variance unknown)

Mean	0 if $\nu>1$
Mode	**0**
Median	**0**
Variance	$\frac{\nu}{\nu-2}$ if $\nu>2$ ∞ if $1<\nu\leq 2$



Important distributions: multivariate normal

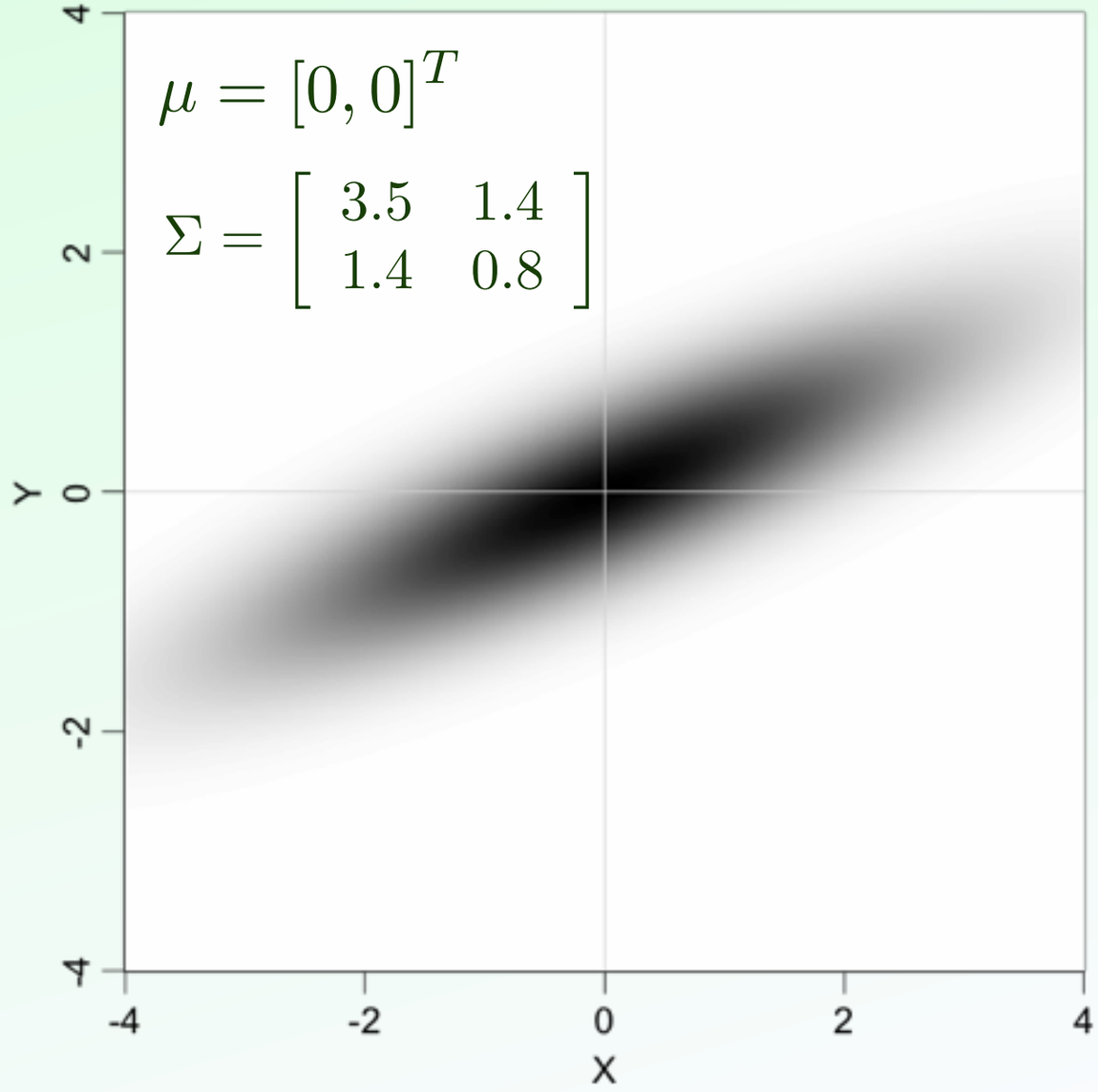
Density:

$$\phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi \det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$ pos. def.

Central role in classical statistics as limiting distribution of many estimators

Mean	$\boldsymbol{\mu}$
Mode	$\boldsymbol{\mu}$
Variance	$\boldsymbol{\Sigma}$



Important distributions: multivariate normal

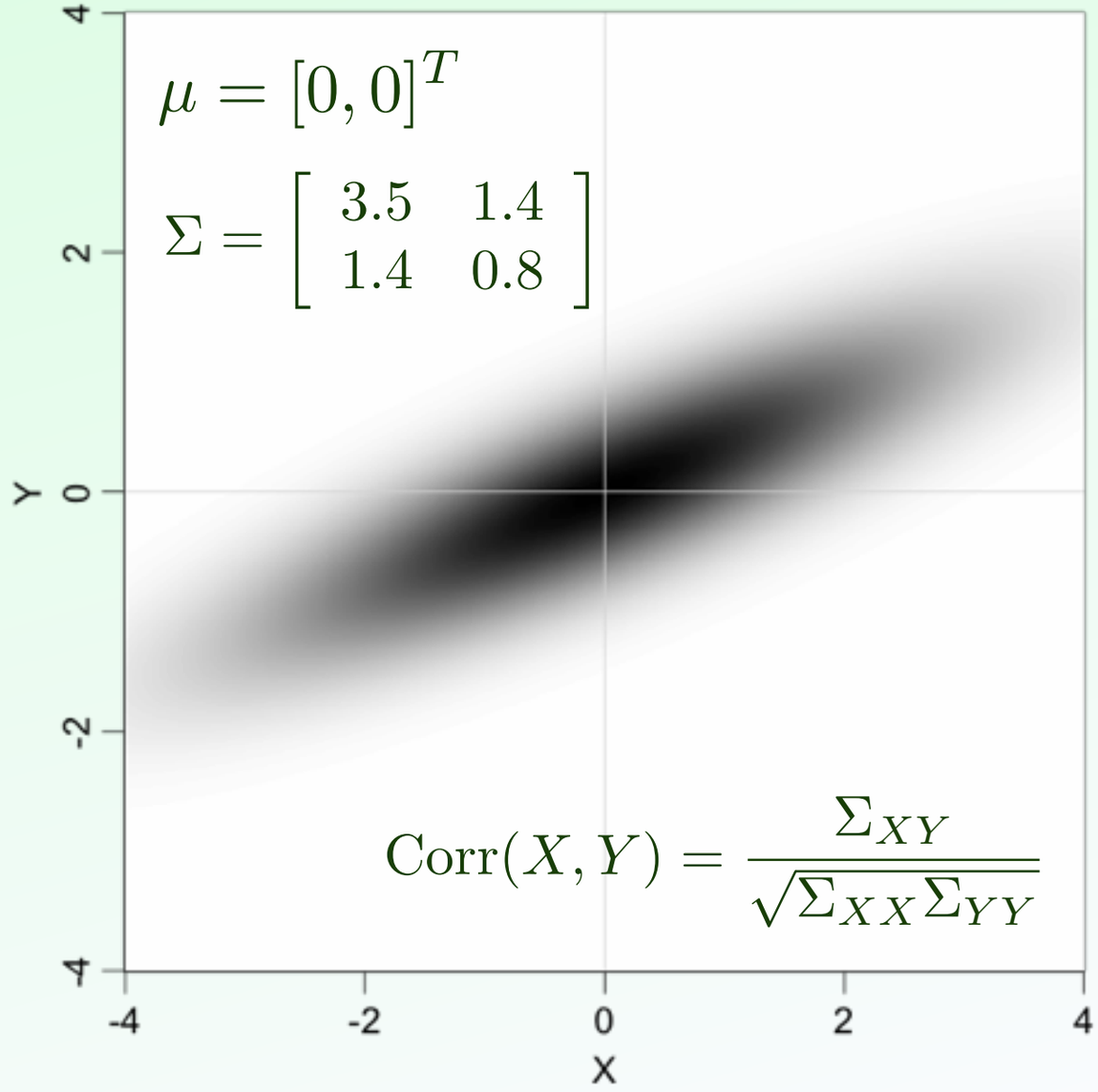
Density:

$$\phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{2\pi \det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$\mu \in \mathbb{R}^n,\ \Sigma \in \mathbb{R}^{n \times n}$ pos. def.

Central role in classical statistics as limiting distribution of many estimators

Mean	$\boldsymbol{\mu}$
Mode	$\boldsymbol{\mu}$
Var-covar.	$\boldsymbol{\Sigma}$



Important distributions: others

F_{ν_1,ν_2} (classical hypothesis testing in linear models)

Exponential (time between events of a Poisson process)

Gamma (survival time)

Laplace (double exponential) (prior for the lasso regression)

Uniform (probability integral transform)

Log-normal (multiplicative CLT)

Extreme-value distributions (distribution of maxima)

(Generalized) Pareto (distribution of extremes above threshold)

Inverse Wishart (prior on the covariance matrix of a normal)

…

Important distributio

F_{ν_1,ν_2} (classical hypo

Exponential (time bet

Gamma (survival time

Laplace (double expo

Uniform (probability

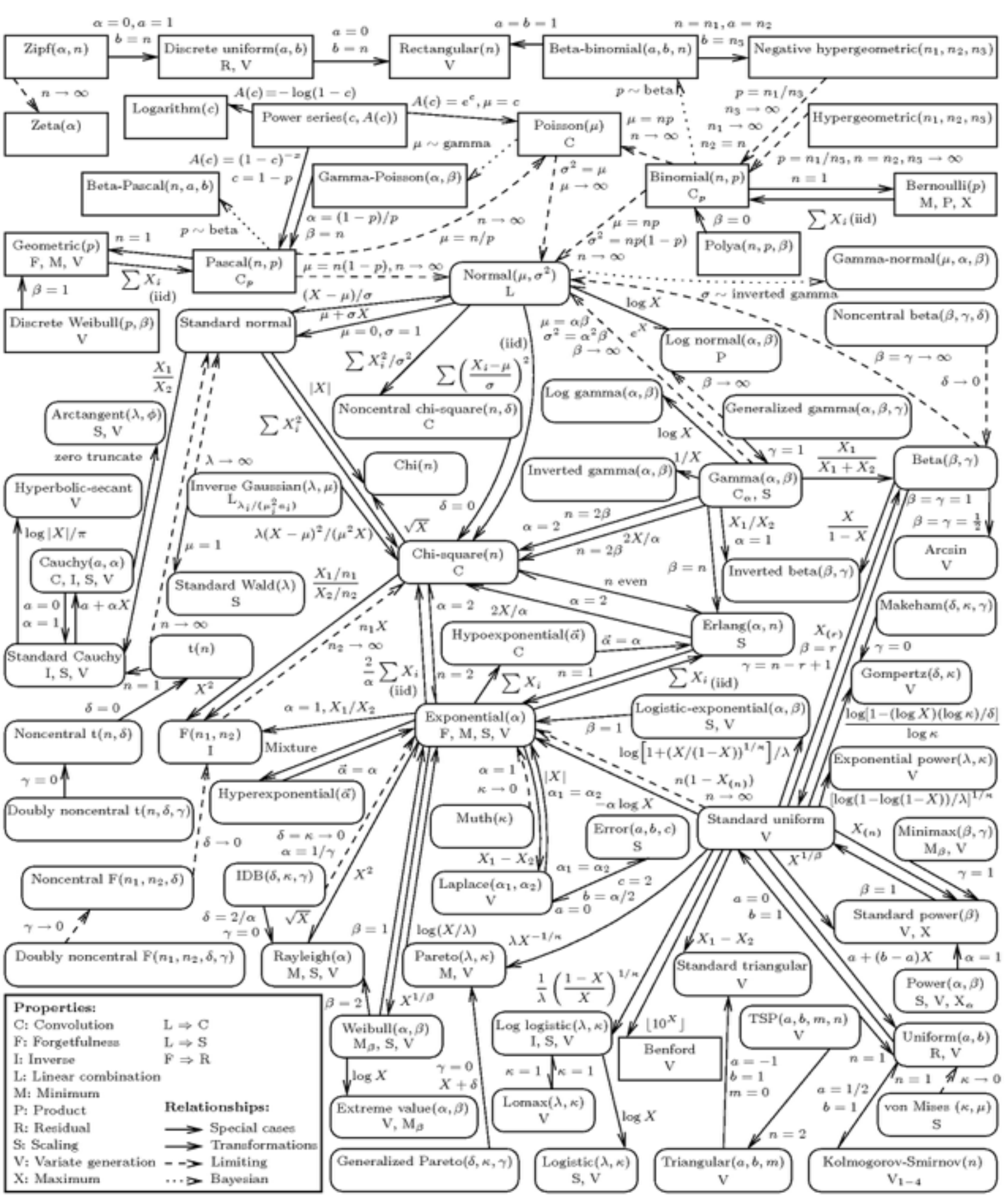
Log-normal (multipli

Extreme-value distrib

(Generalized) Pareto

Inverse Wishart (prio

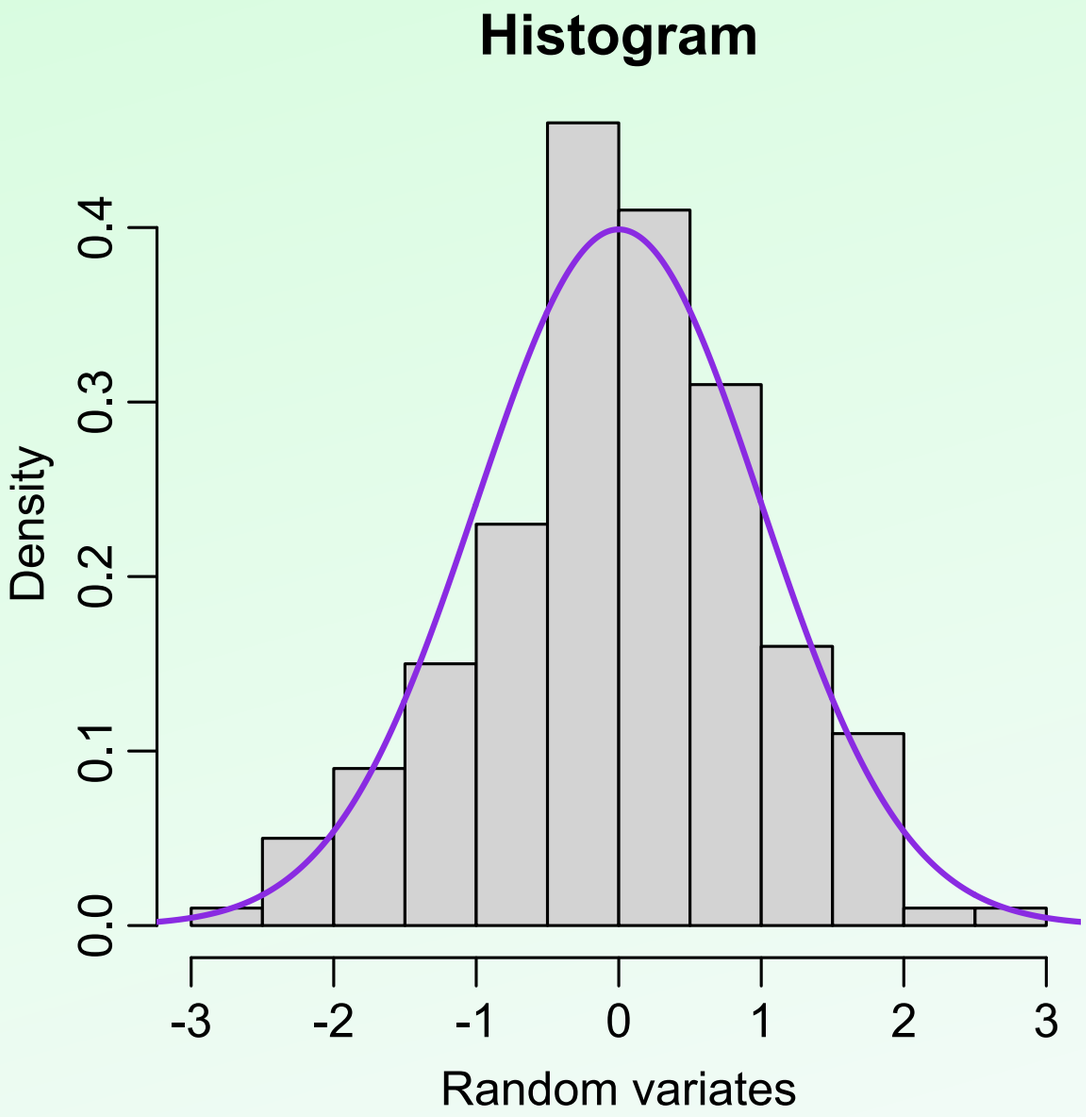
…



Finding the way on the map: techniques

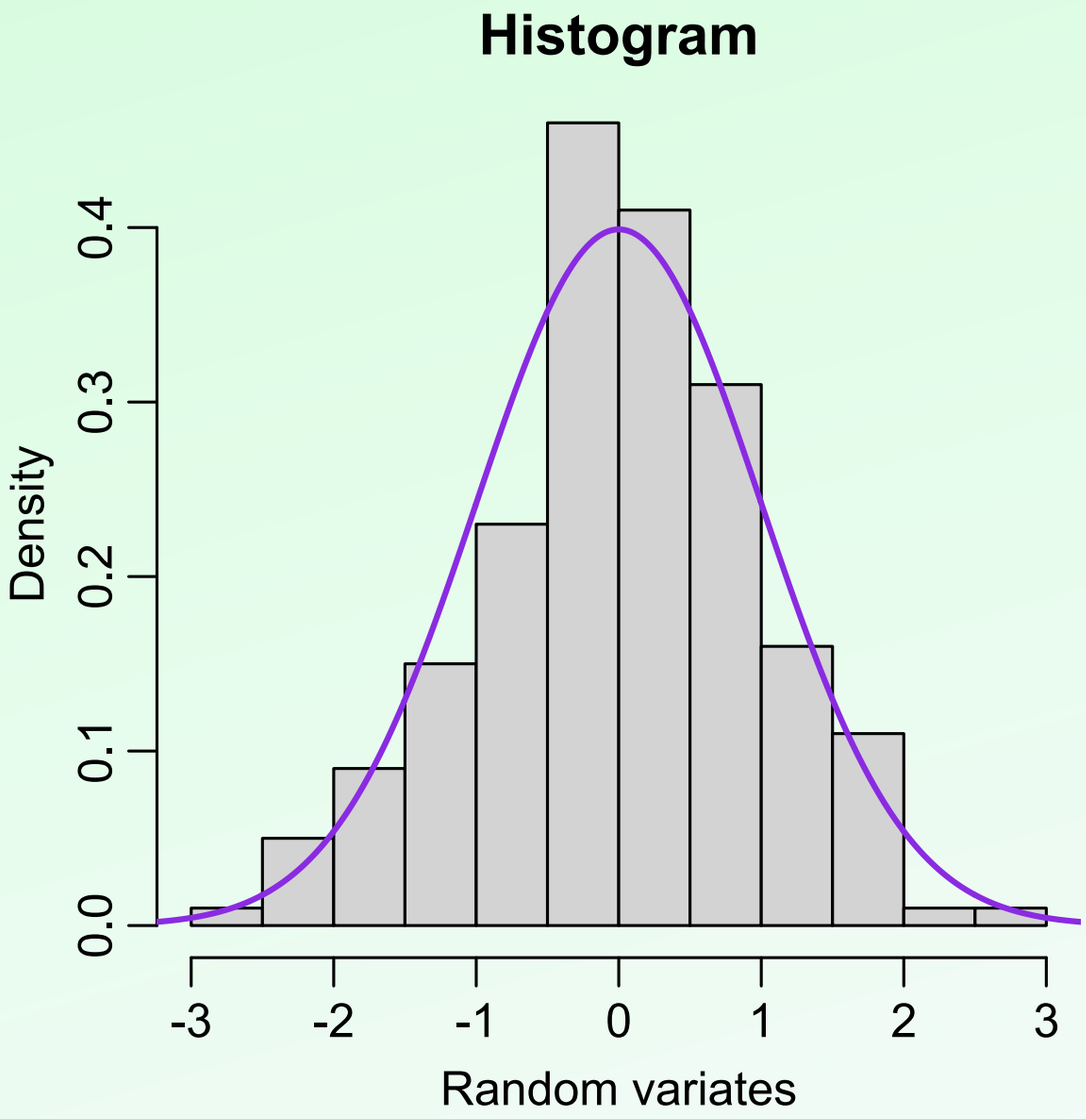
Diagnostics: the QQ plot

- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when $n \longrightarrow \infty$)



Diagnostics: the QQ plot

- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when $n \longrightarrow \infty$)



Statisticians' preference:
quantile-quantile (QQ) plot.

It consists of the pairs

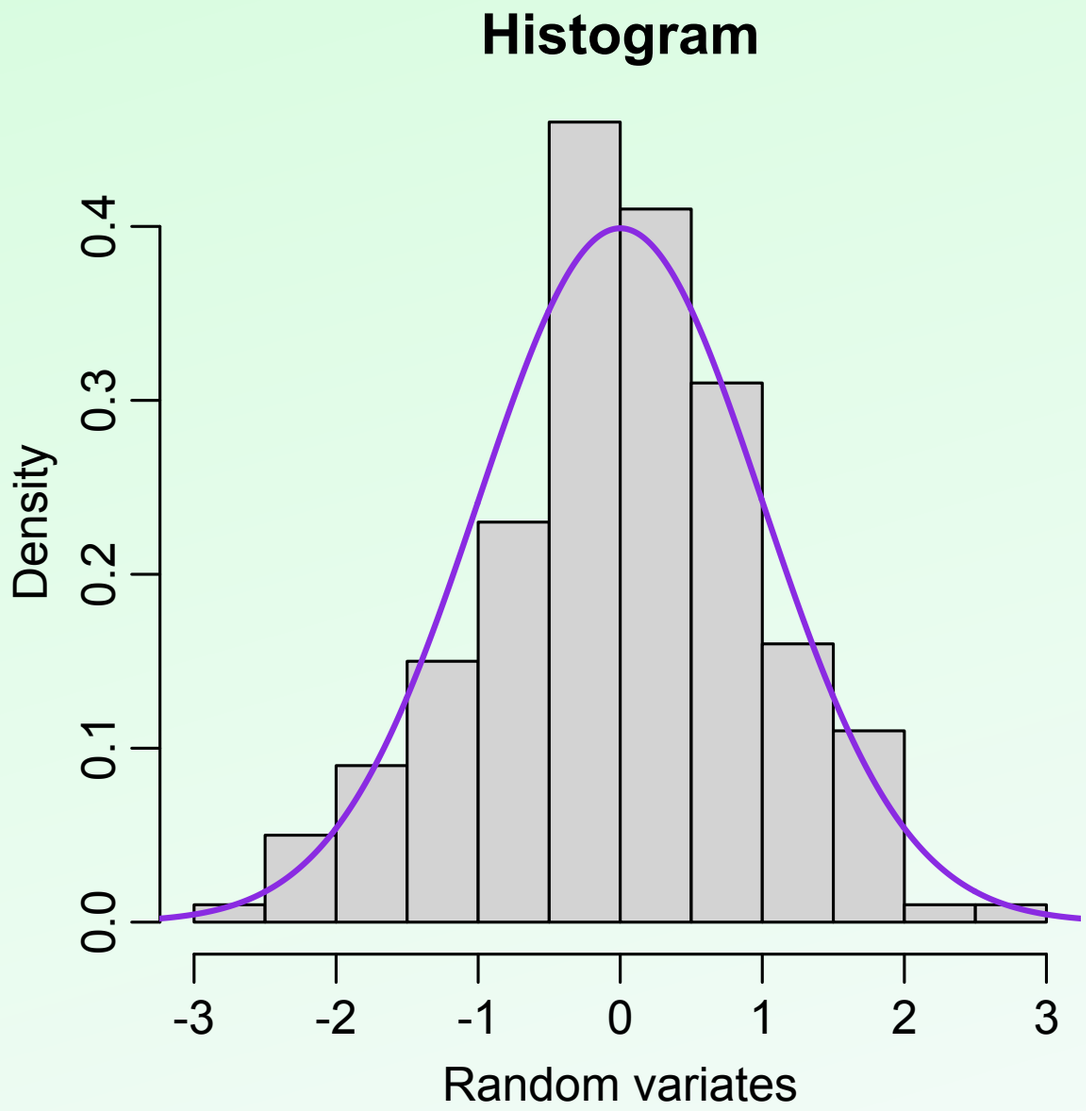
$$\left\{ \Phi^{-1}\left(\frac{j}{n+1}\right), Y_{(j)} \right\},$$

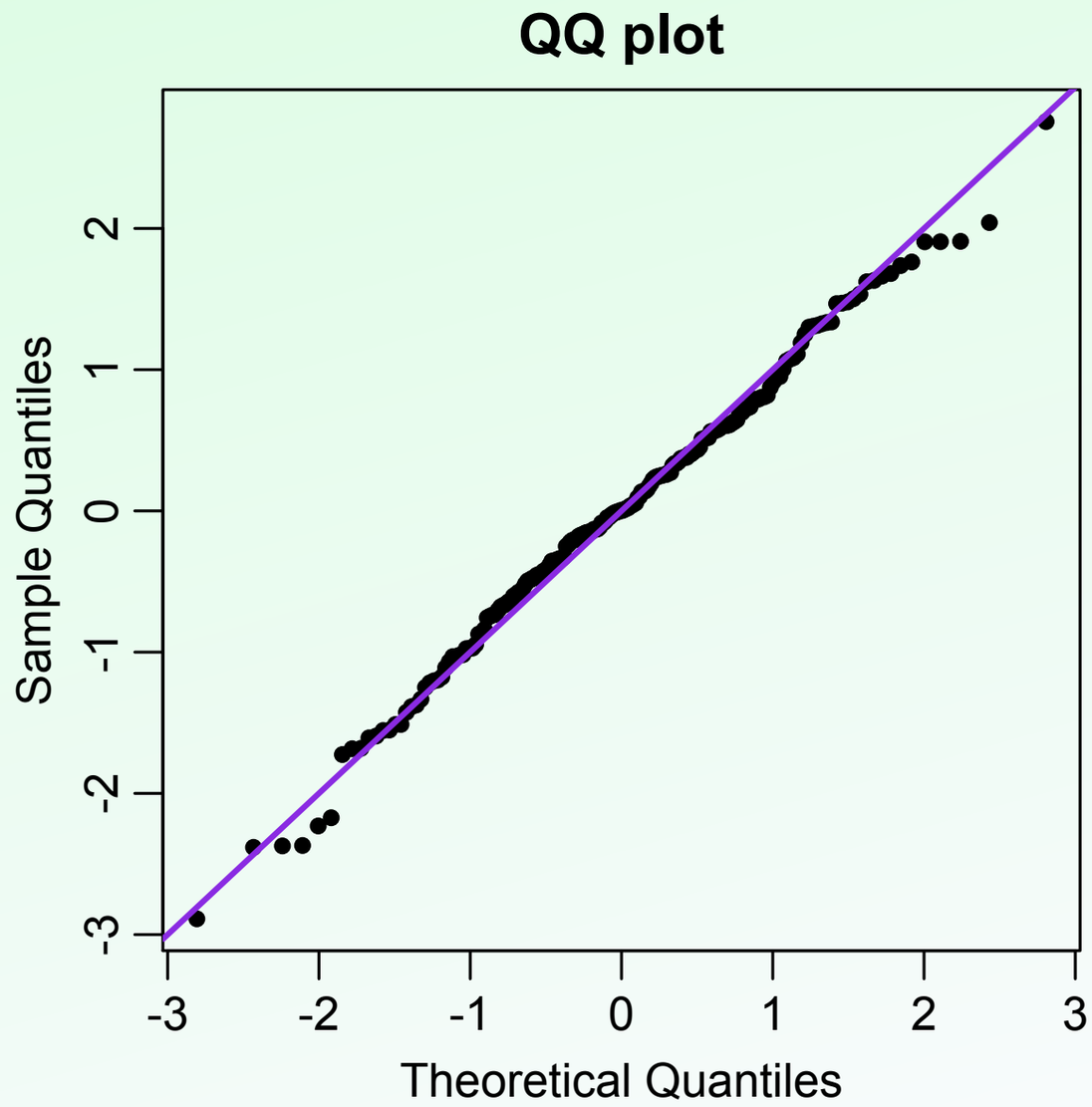
where
$\Phi^{-1}\left(\frac{j}{n+1}\right)$ is the inverse of the standard normal distribution, and
$Y_{(j)}$ is the ordered sample.

✳ Exercises 1-2

Diagnostics: the QQ plot

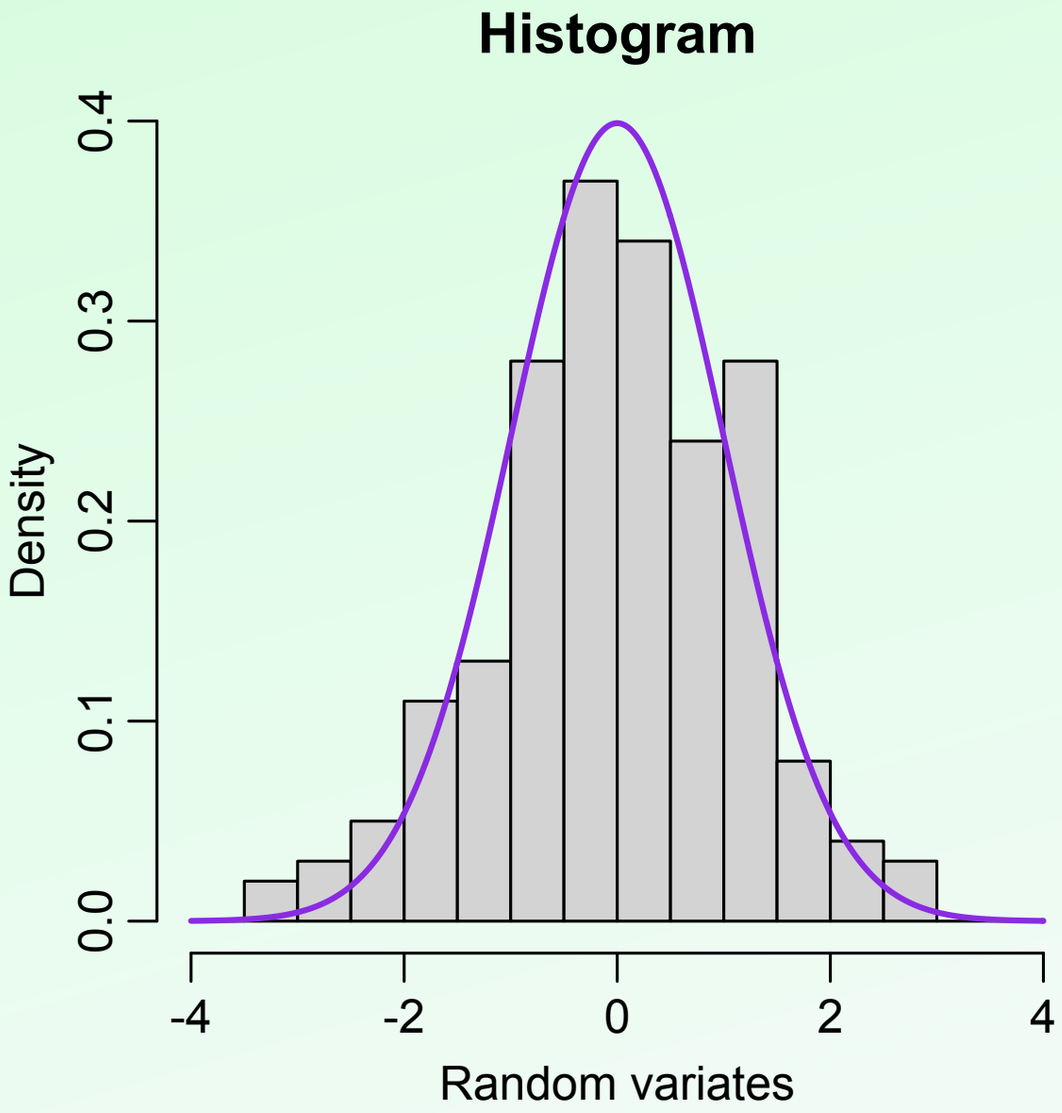
- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when $n \longrightarrow \infty$)





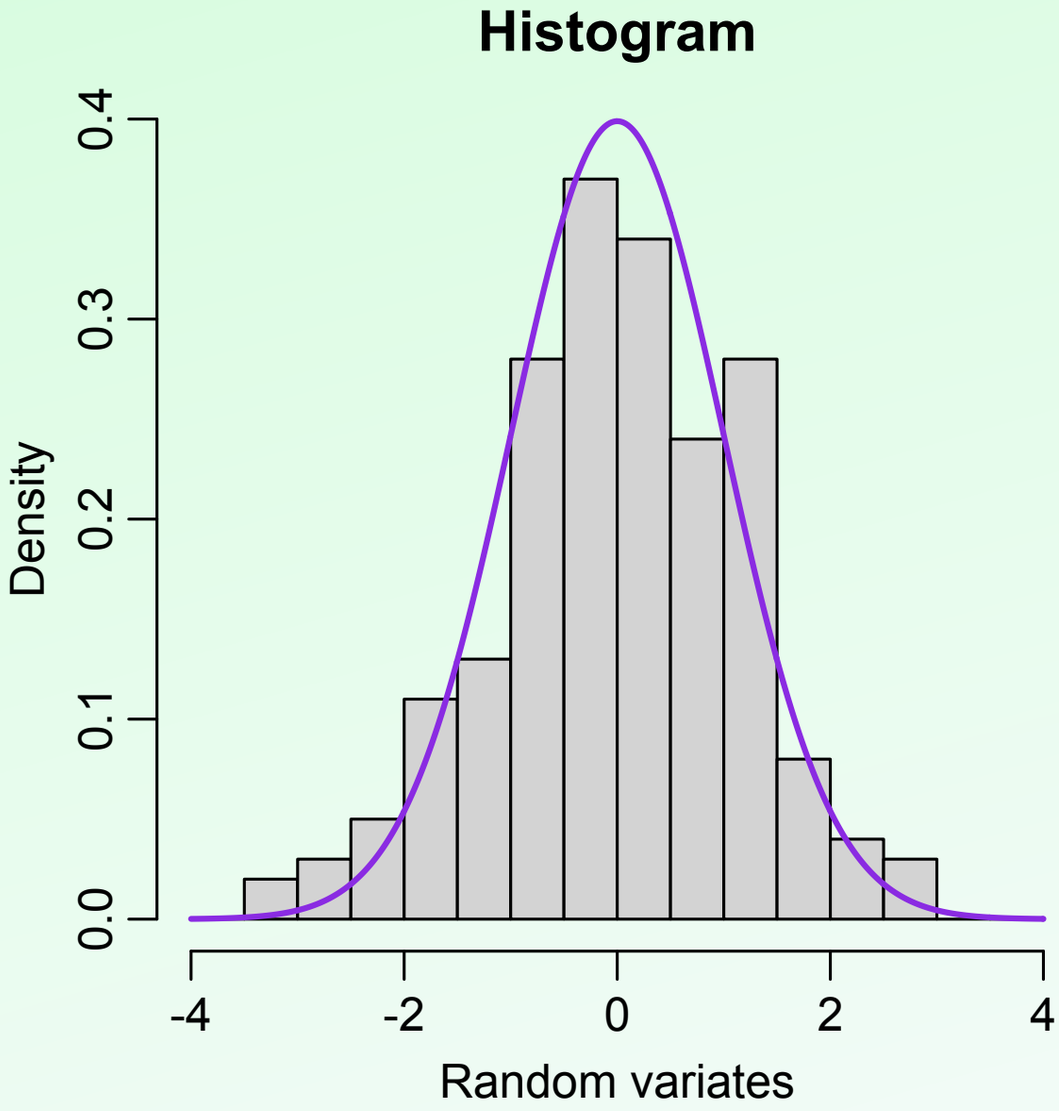
Diagnostics: the QQ plot

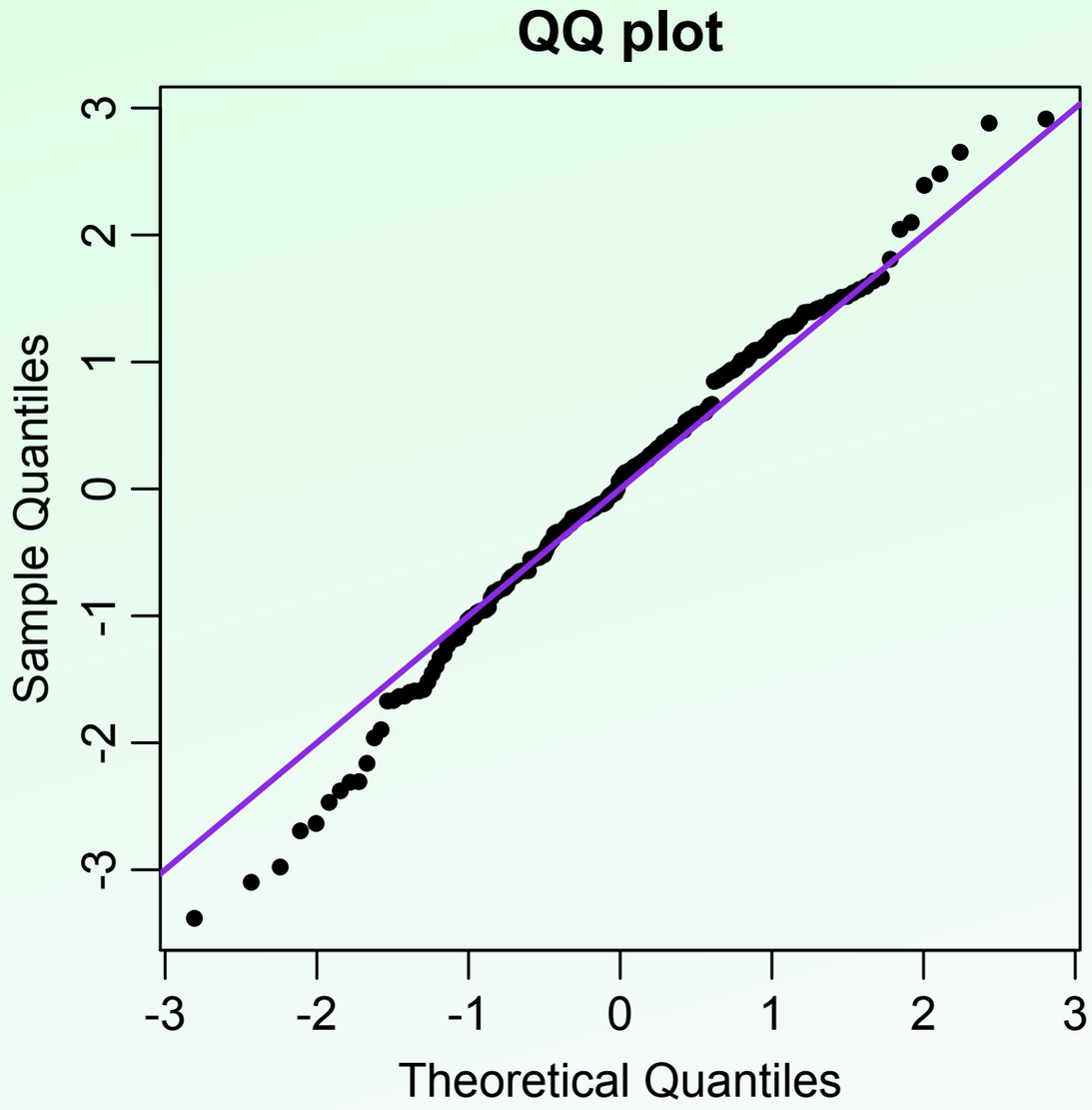
- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when $n \longrightarrow \infty$)



Diagnostics: the QQ plot

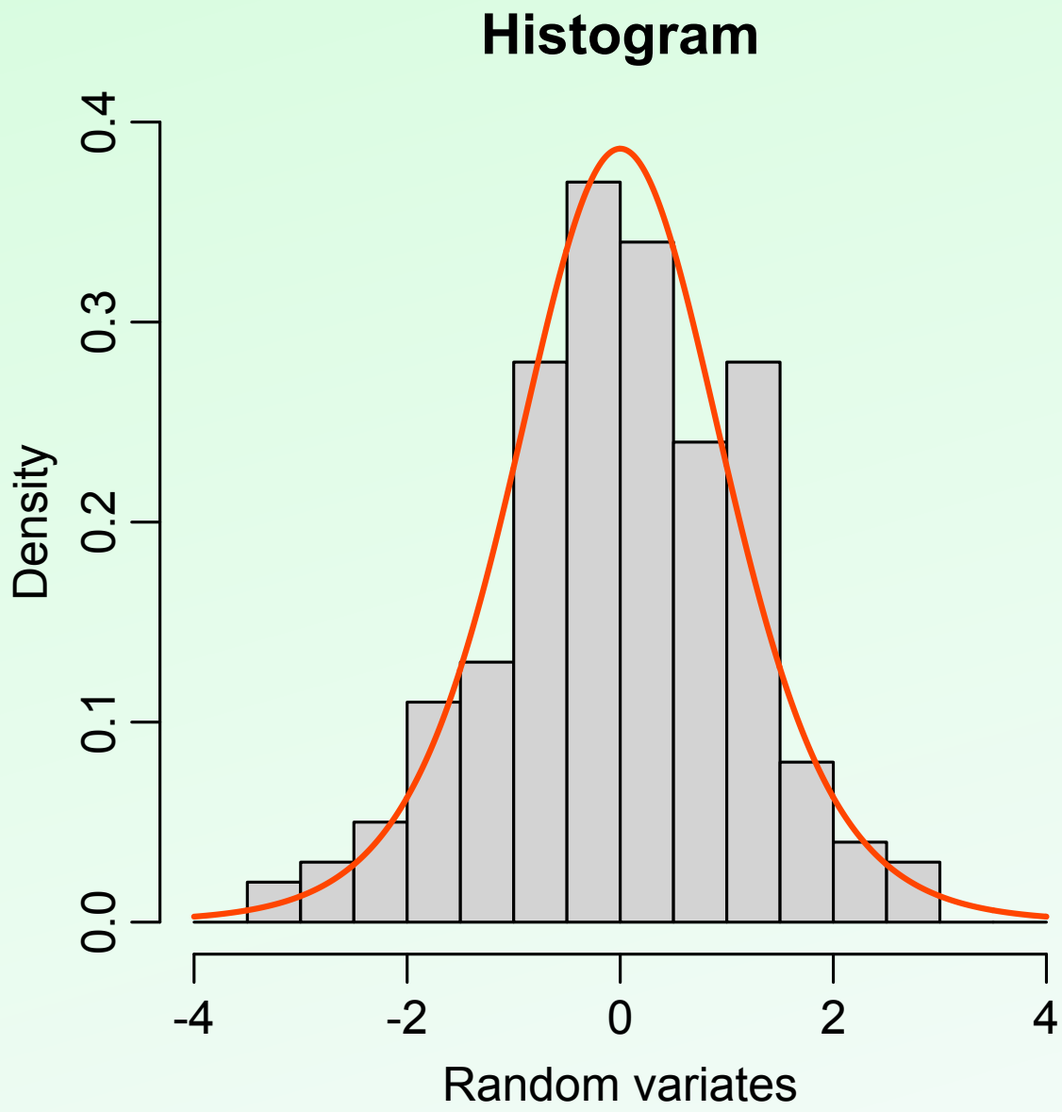
- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when $n \longrightarrow \infty$)

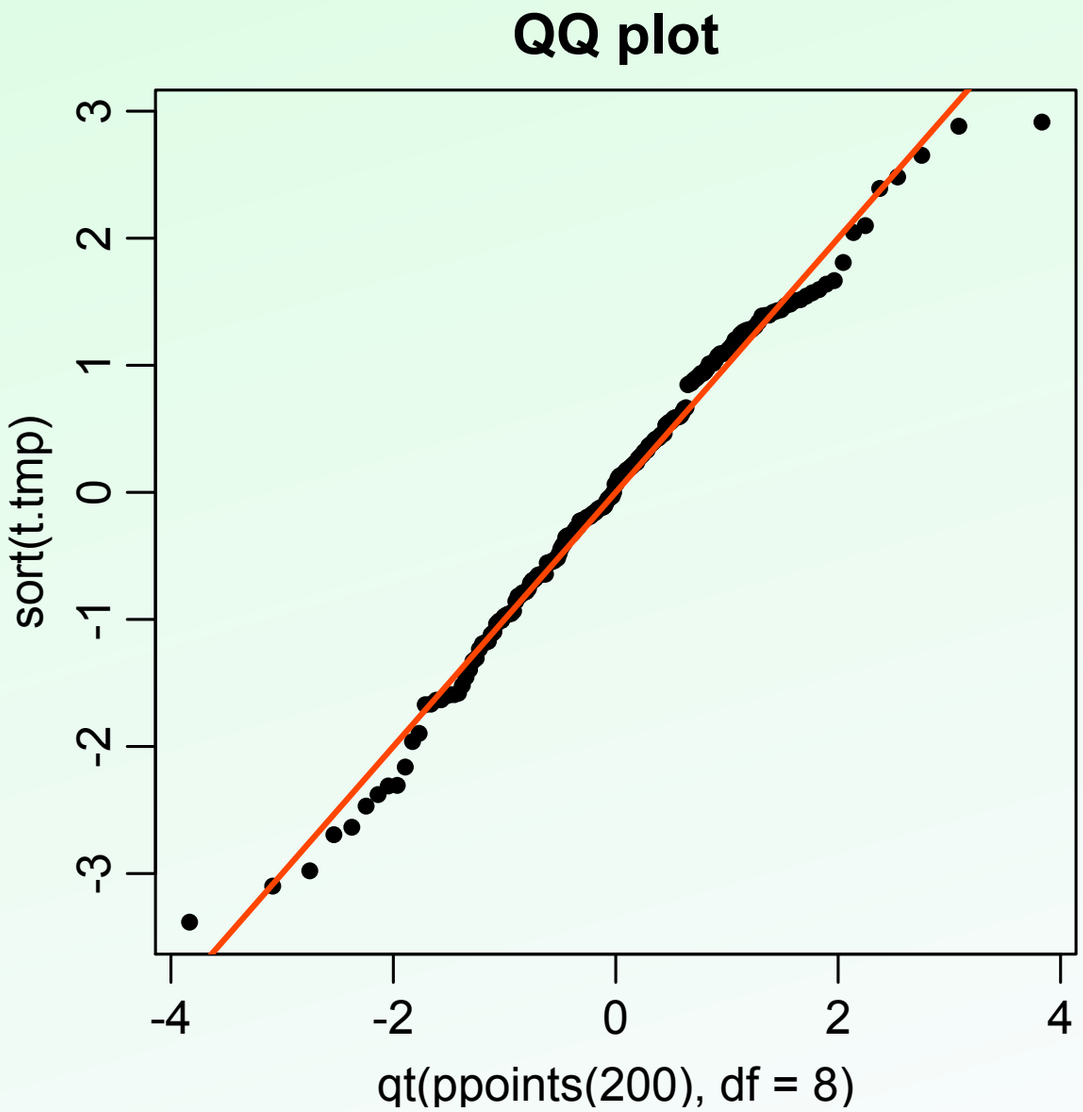




Diagnostics: the QQ plot

- Parametric models are nearly always only approximate
- In classical statistics, CIs are based on asymptotic behaviour (that is, when $n \longrightarrow \infty$)





Diagnostics: the QQ plot

- **"*x* is a detection at 3σ level":** this assertion relies on the unsaid assumption of normality ("Gaussianity").

Check the distributions!

If the distribution is indeed normal:

$$|x - \mu| > 3\sigma \quad \Longleftrightarrow \quad P(X \text{ is more extreme than } x) < 0.0027$$

But if the distribution is the t_8 as in the previous pages:

$$|x - \mu| > 3\sigma \quad \Longleftrightarrow \quad P(X \text{ is more extreme than } x) < 0.017$$

✳ Exercise 3

Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables, such that both $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ are finite.

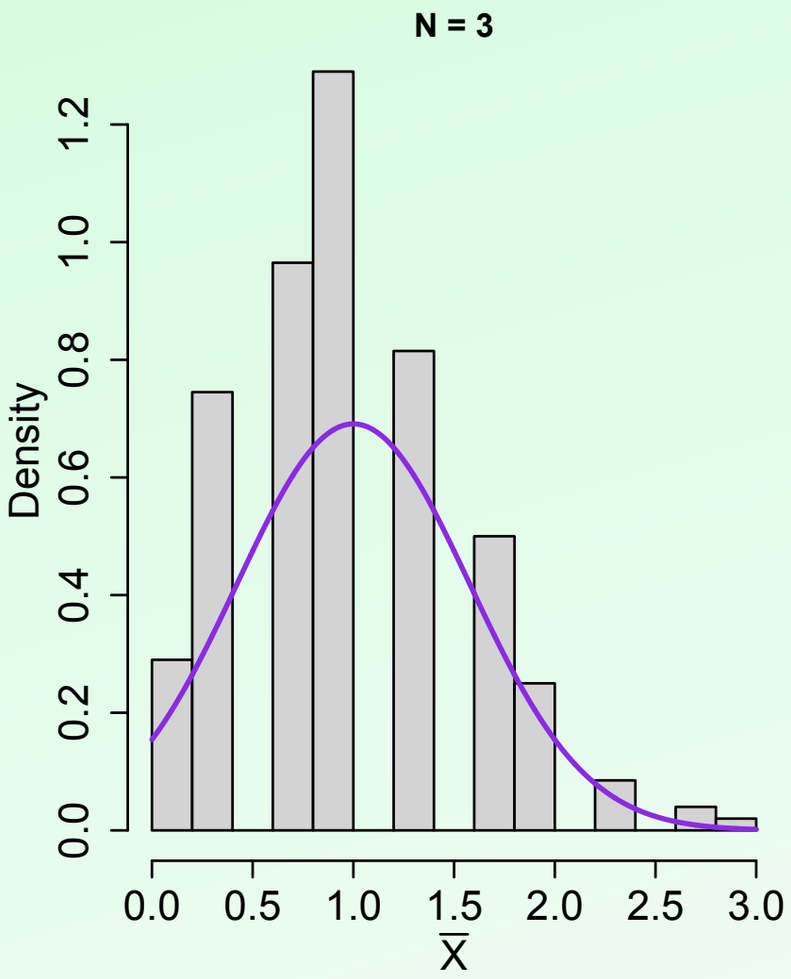
Then, defining $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ and $S_n = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$, we have

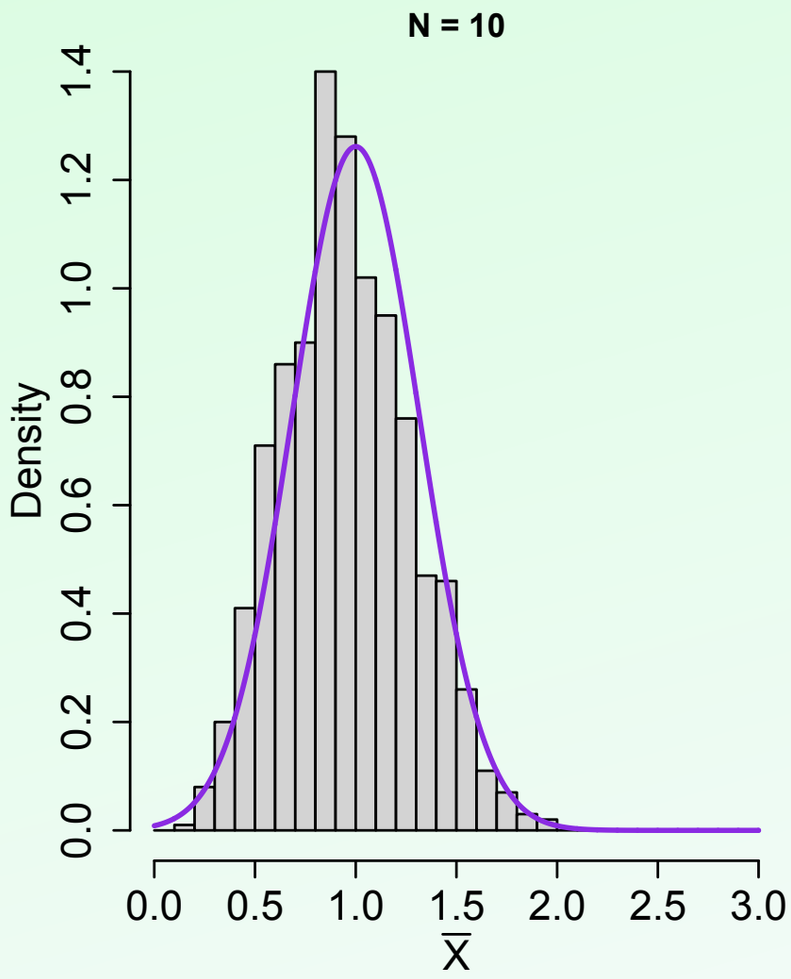
$$S_n \dot{\sim} \mathcal{N}(0,1),$$

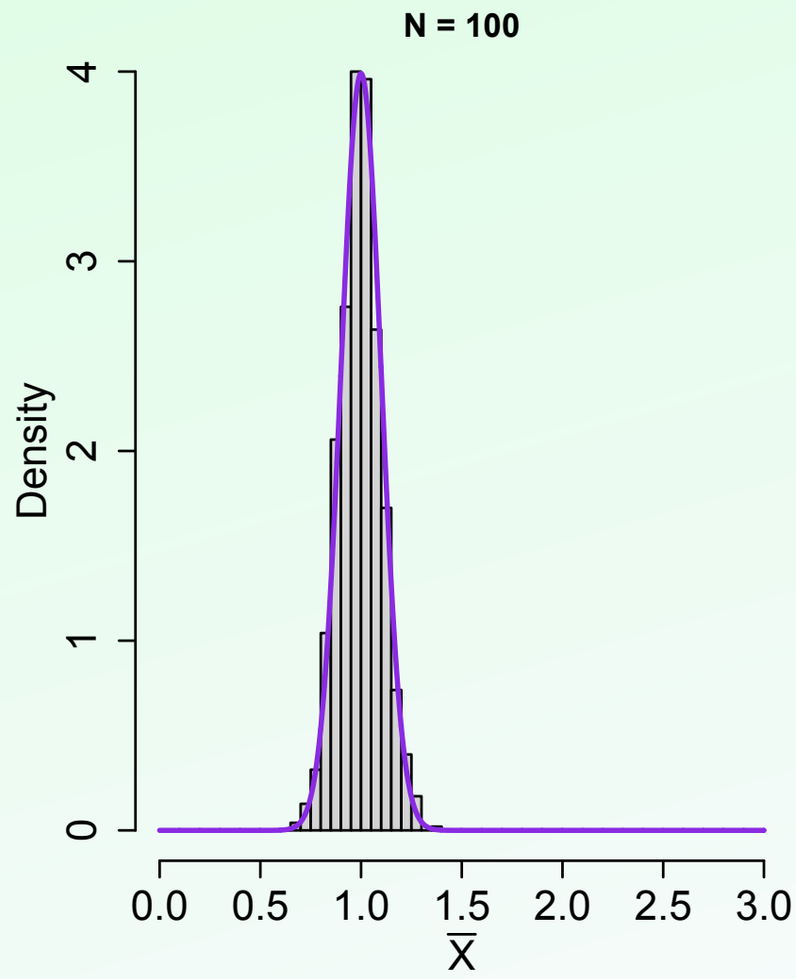
$$\bar{X} \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \longleftrightarrow \quad \sum_{i=1}^{N} X_i \dot{\sim} \mathcal{N}\left(\mu, n\sigma^2\right)$$

Illustration CLT

$X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Pois}(\lambda)$; distribution of $\frac{1}{N}\sum_{i=1}^{N} X_i$?





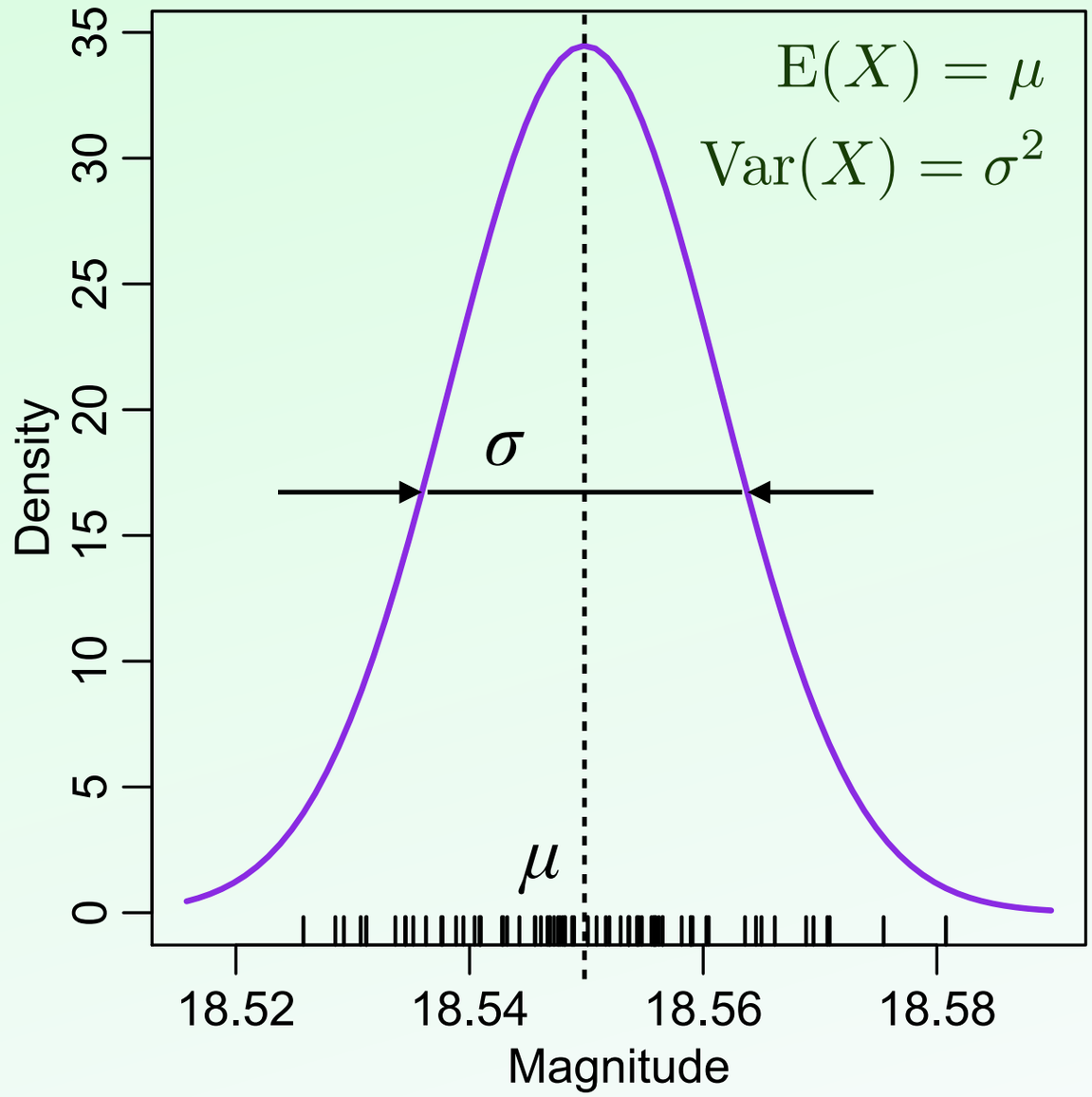


Learning the distribution: parameter estimation

Estimators: method of moments

Principle:

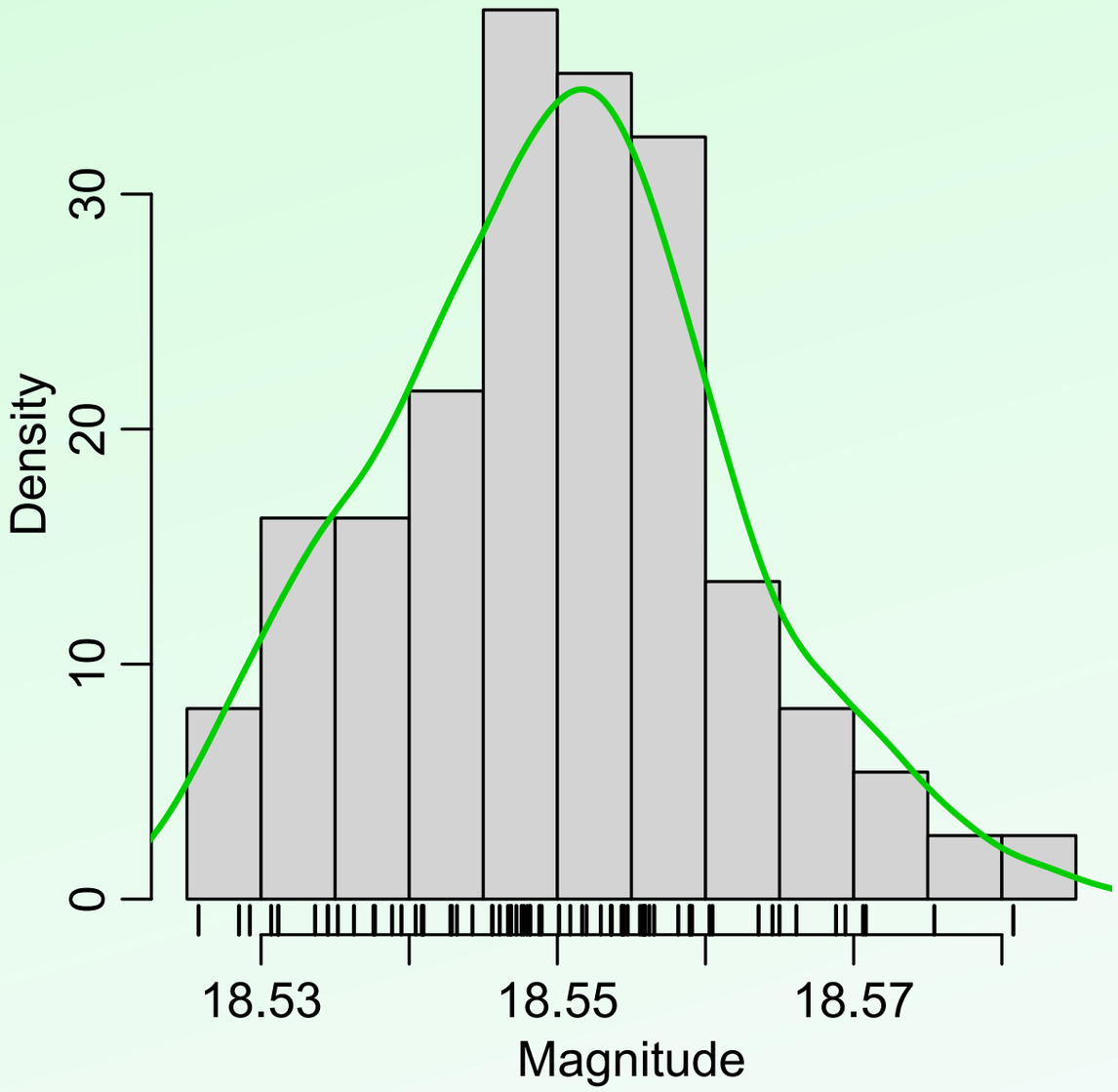
1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.



Estimators: method of moments

Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.



$$\mathrm{E}(X) = \mu$$
$$\mathrm{Var}(X) = \sigma^2$$

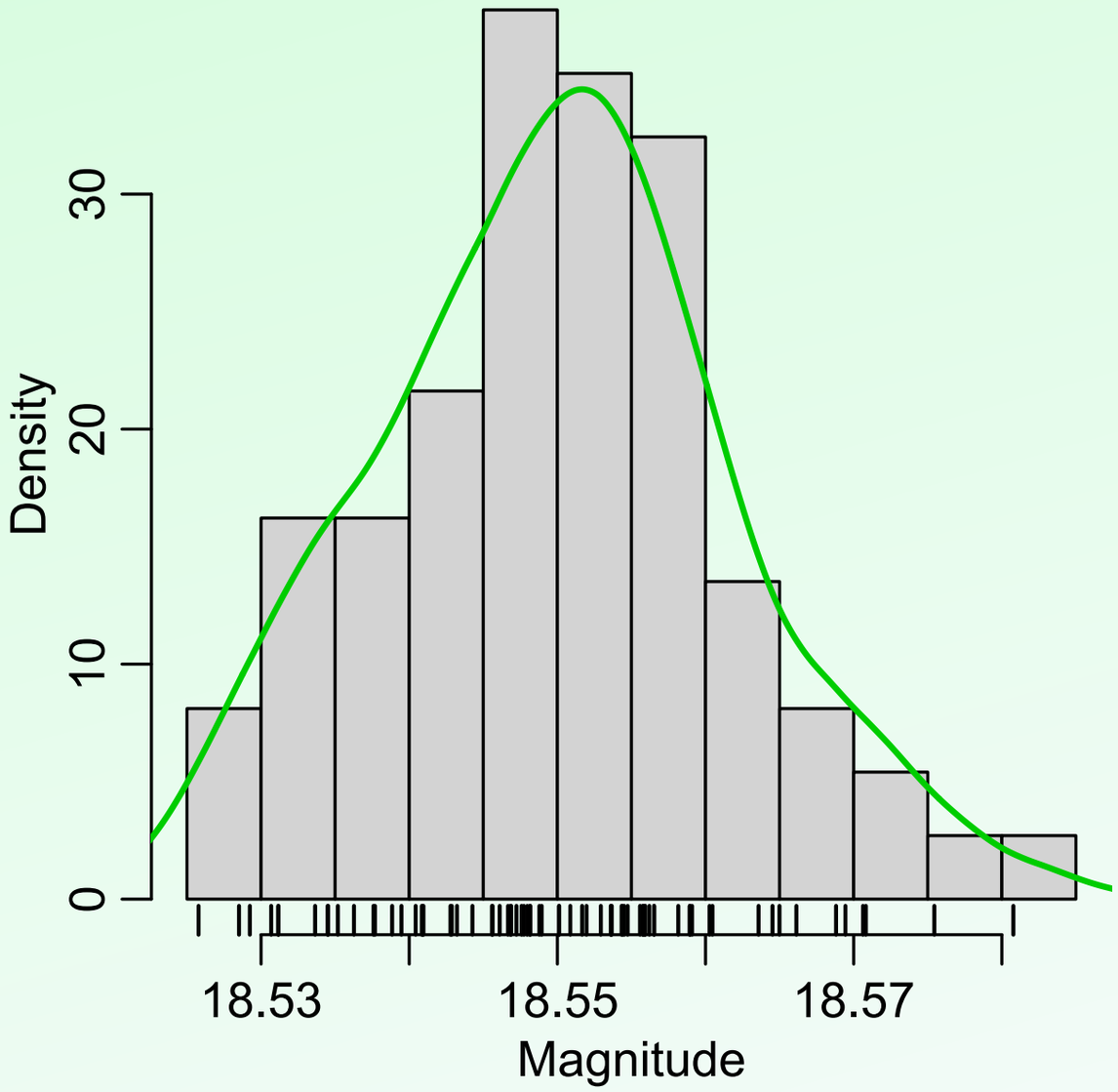
Can we assume
a normal distribution?

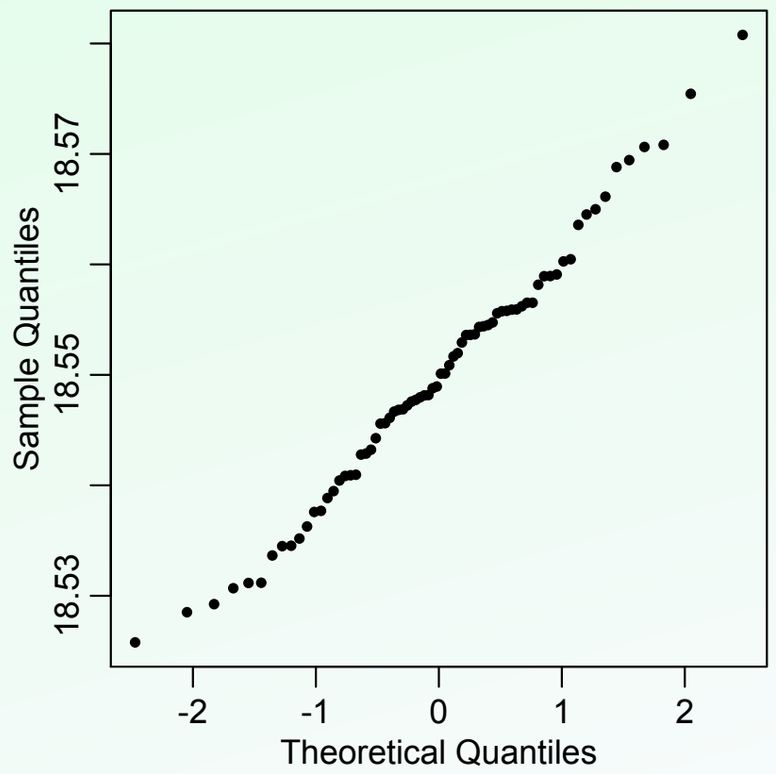
Estimators: method of moments

Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.

$$\mathrm{E}(X) = \mu$$
$$\mathrm{Var}(X) = \sigma^2$$

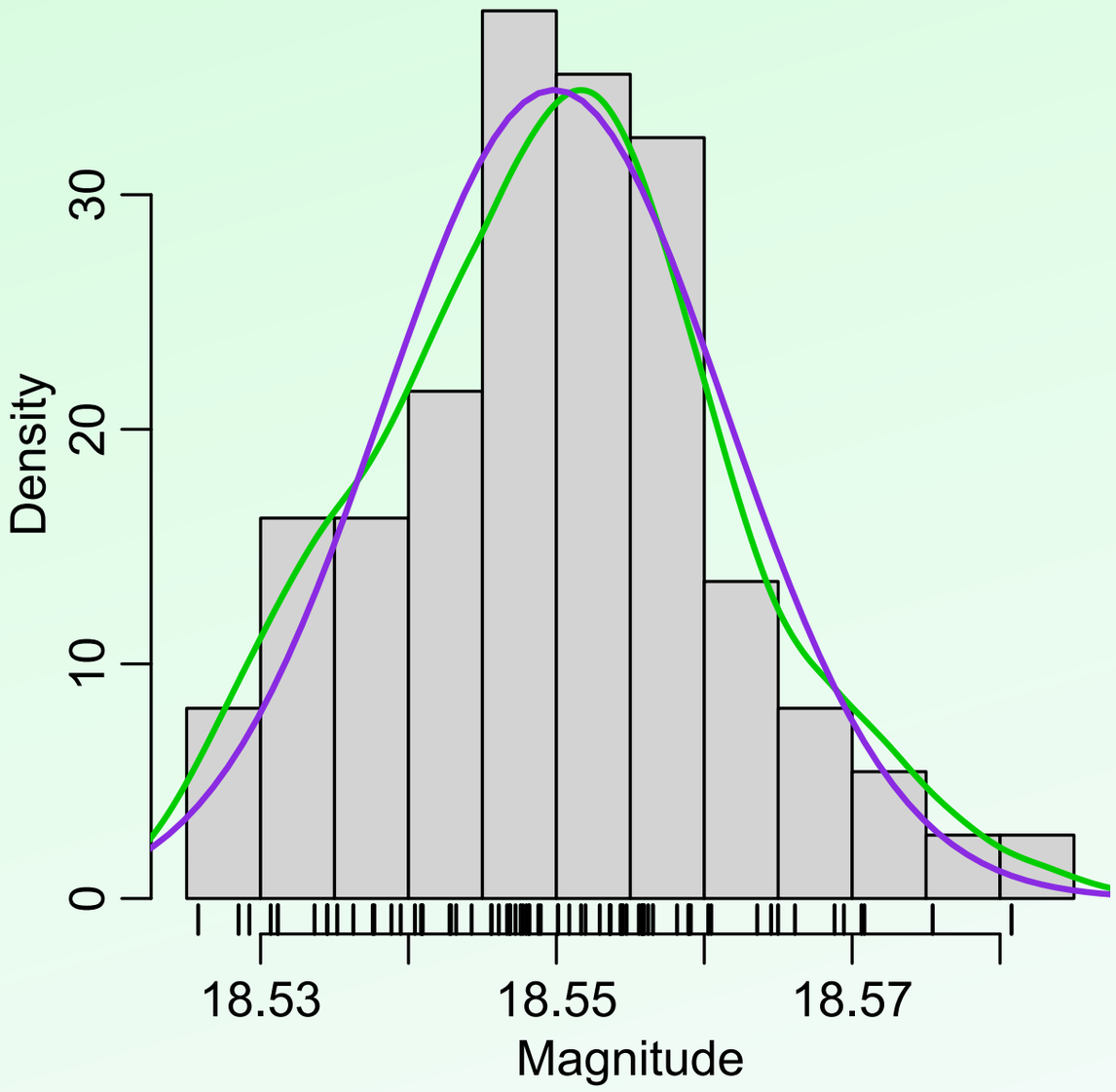




Estimators: method of moments

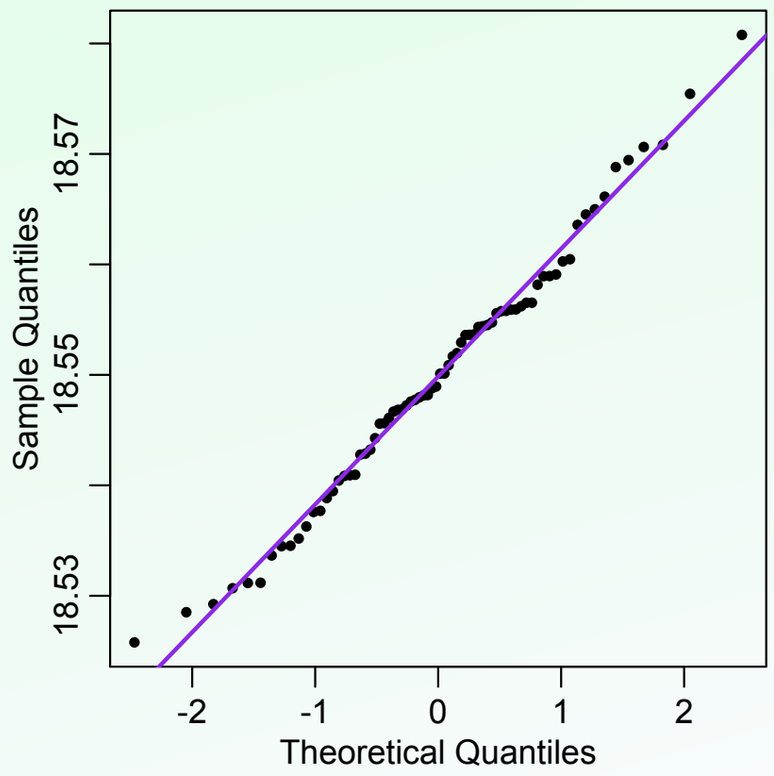
Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.



$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i = 18.5499$$

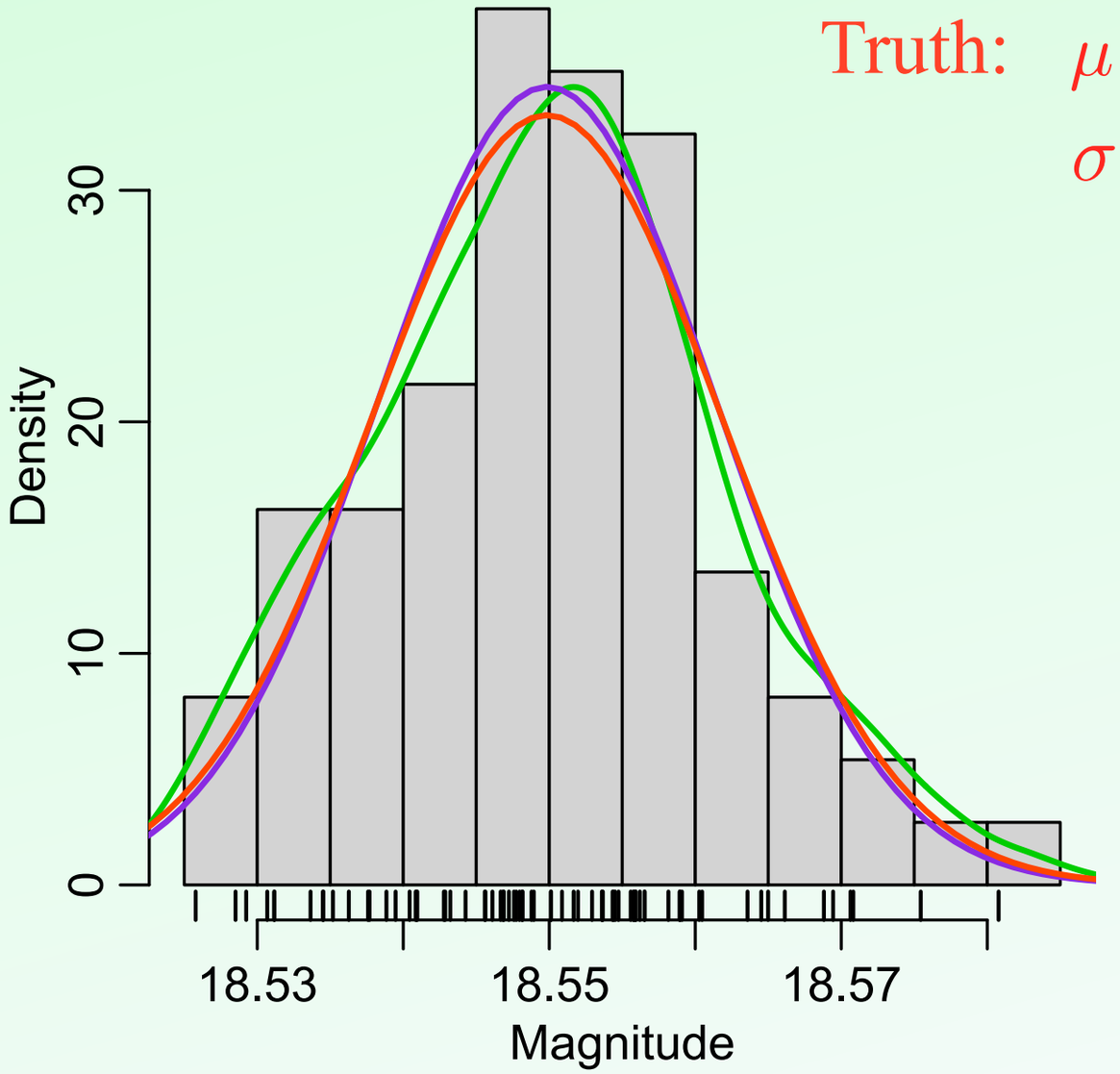
$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{\mu})^2 = 0.0116^2$$



Estimators: method of moments

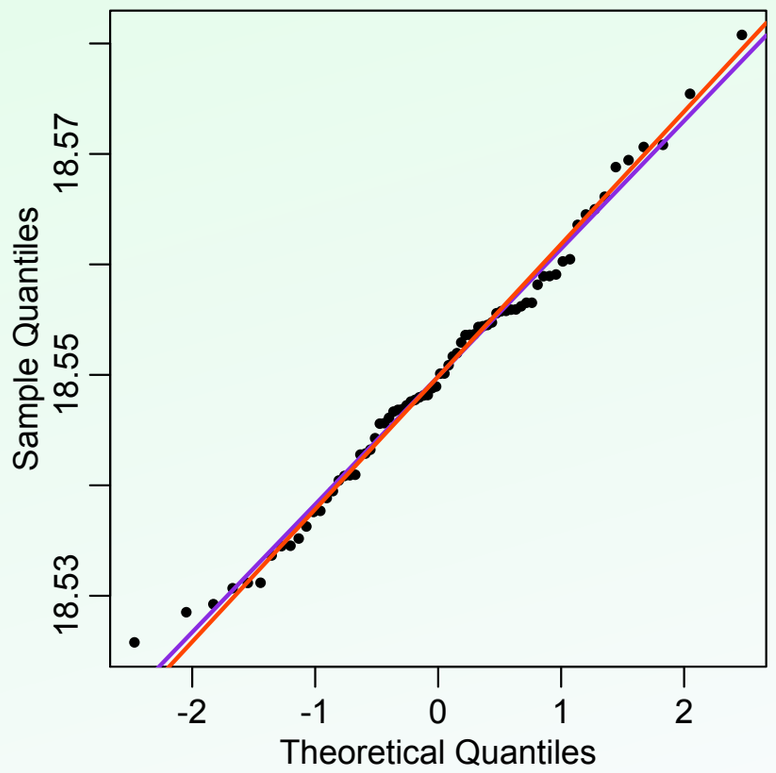
Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.



Truth: $\mu = 18.54984$ $\quad \hat{\mu} = 18.5499$

$\sigma = 0.012$ $\quad \hat{\sigma}^2 = 0.0116^2$

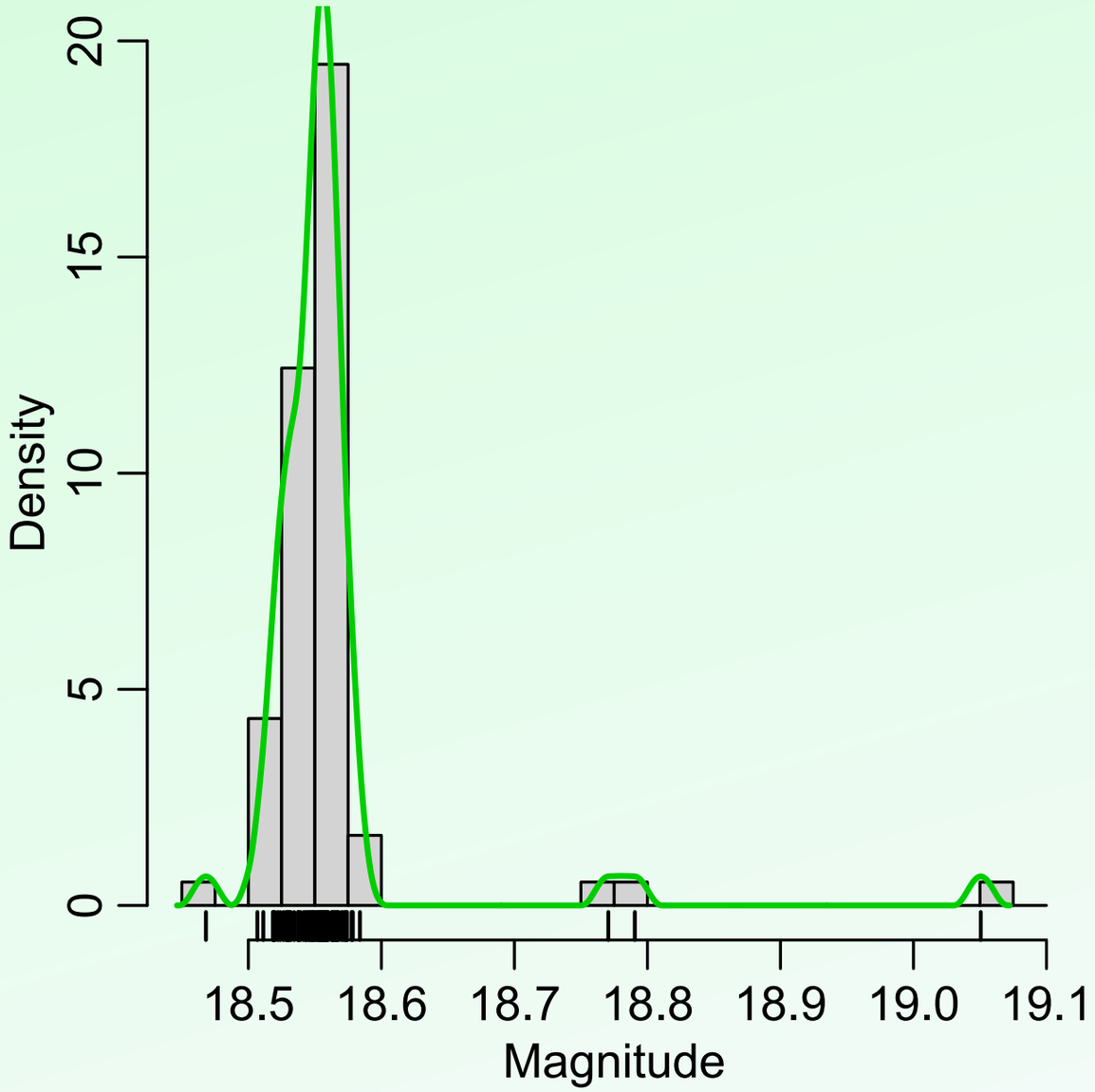


✳ Exercise 4

Estimators: method of moments (robustness)

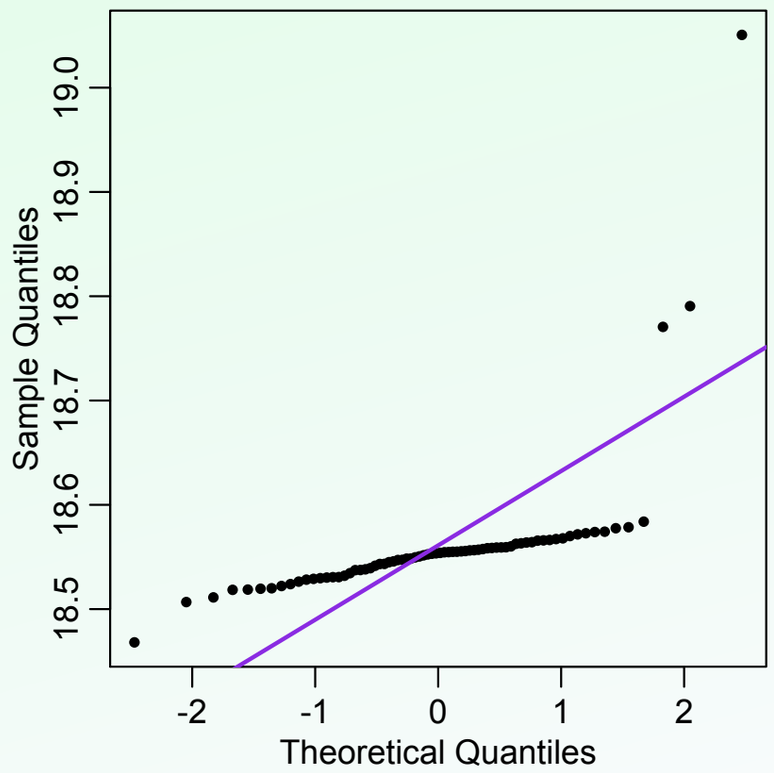
Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.



$$\mathrm{E}(X) = ?$$

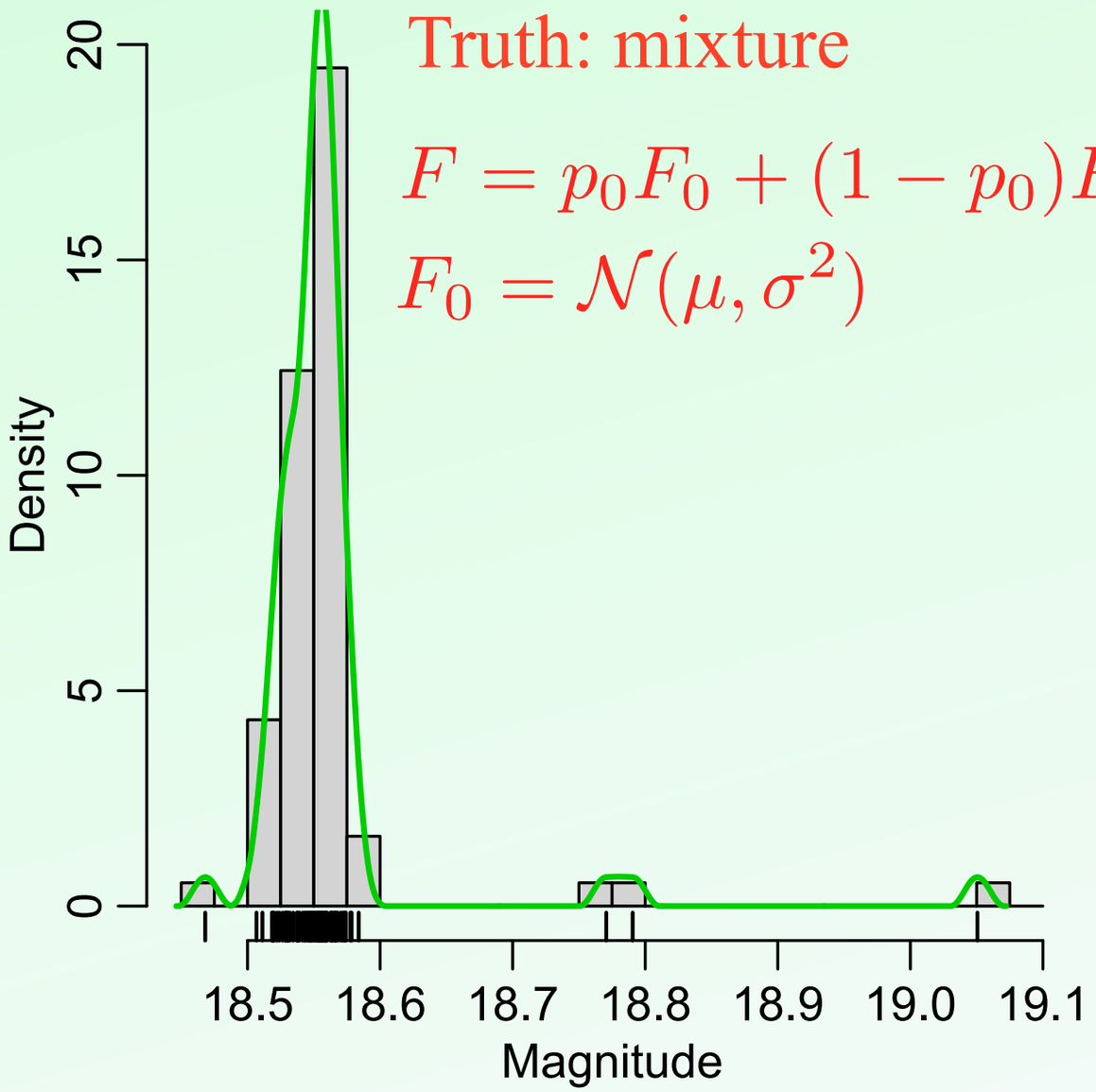
$$\mathrm{Var}(X) = ?$$



Estimators: method of moments (robustness)

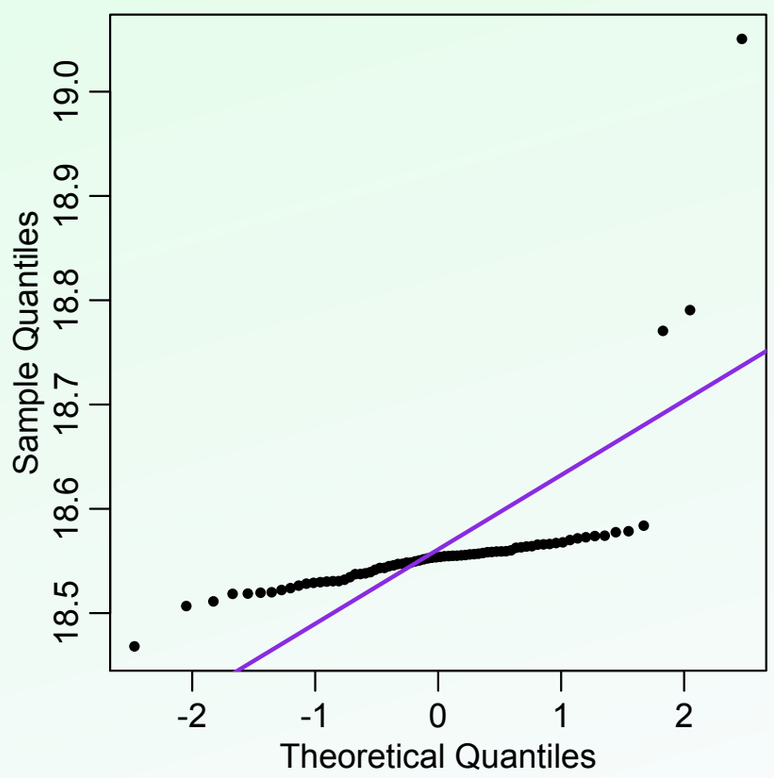
Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.



$\mathrm{E}(X) = ?$

$\mathrm{Var}(X) = ?$



Estimators: method of moments (robustness)

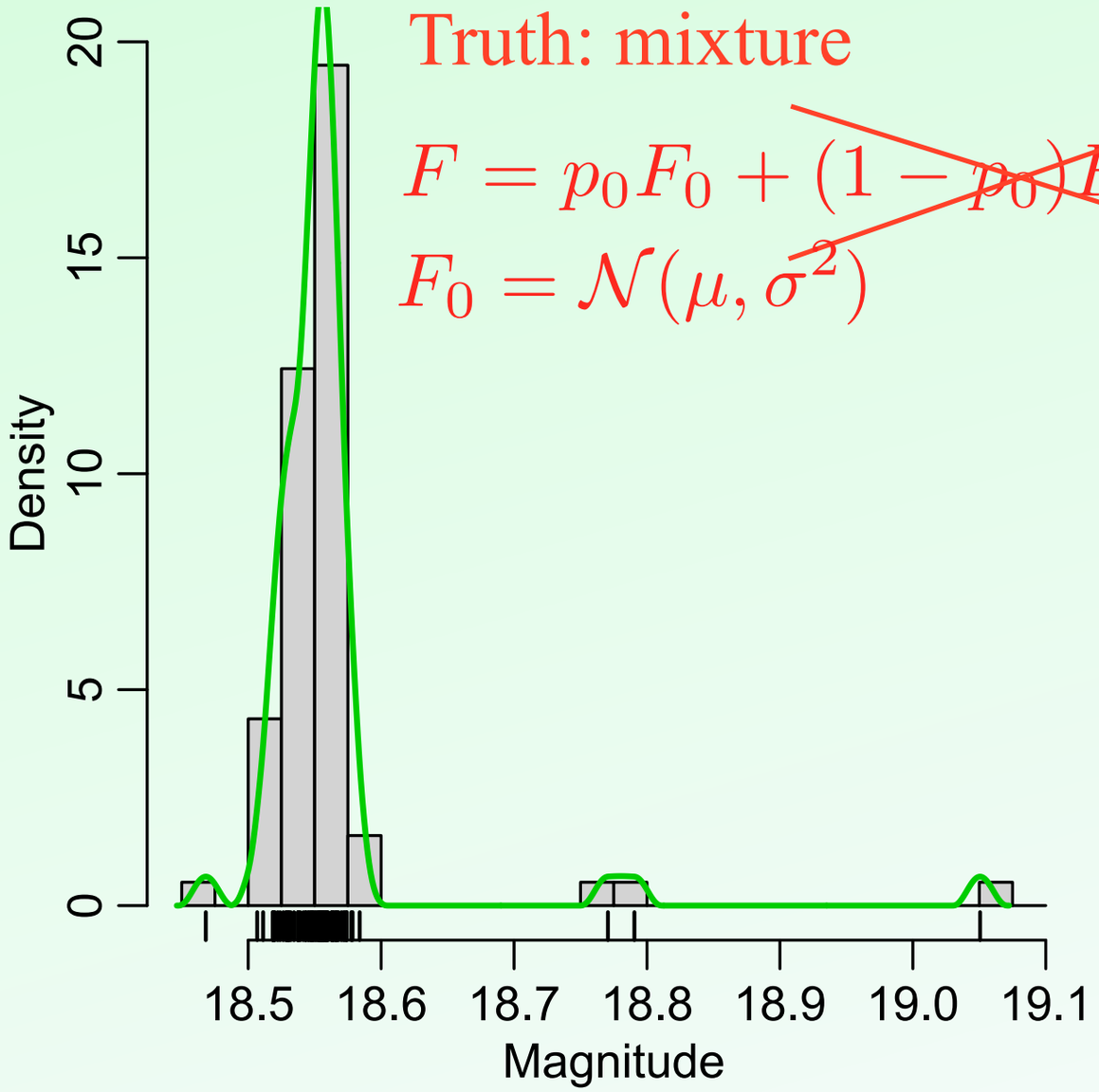
Principle:

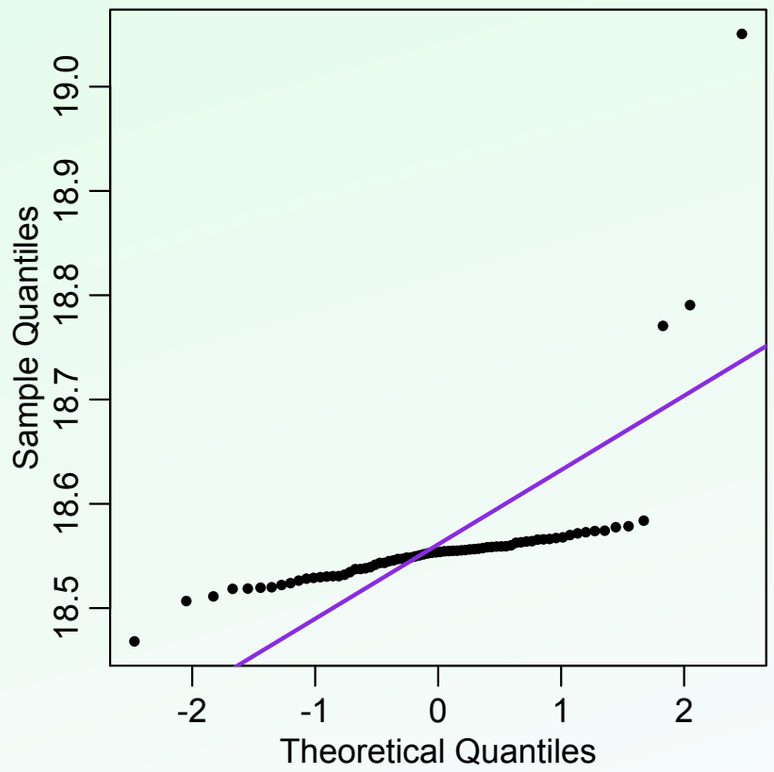
1. Assume a distribution $F(x; \theta)$.
2. Alternative 1: identify and cut the outliers (often iteratively)

Truth: mixture

$$F = p_0 F_0 + (1 - p_0) F_{\text{cont}}$$

$$F_0 = \mathcal{N}(\mu, \sigma^2)$$

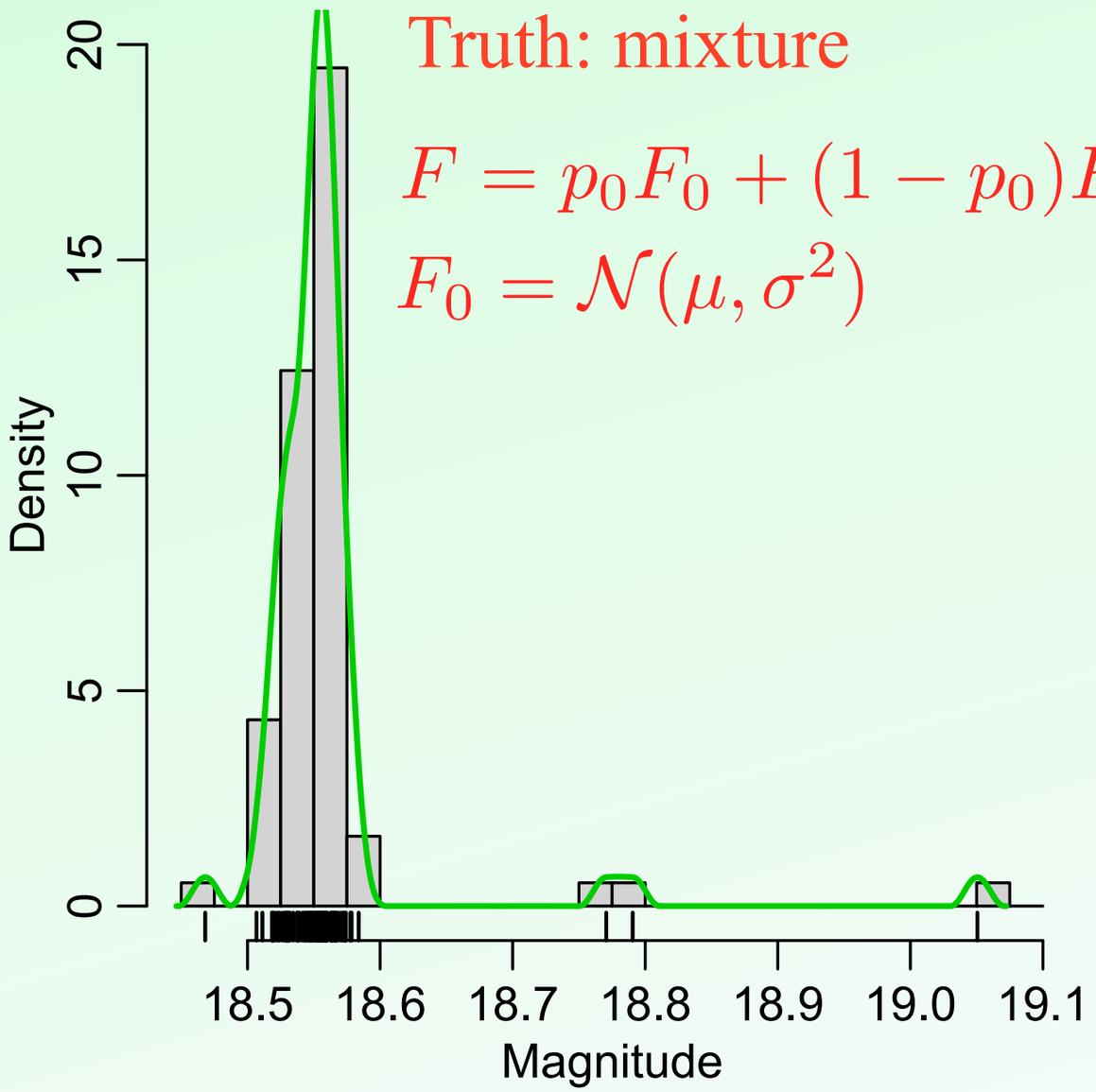




Estimators: method of moments (robustness)

Principle:

1. Assume a distribution $F(x; \theta)$.
2. Alternative 2: use quantile estimators as estimators of the moments.



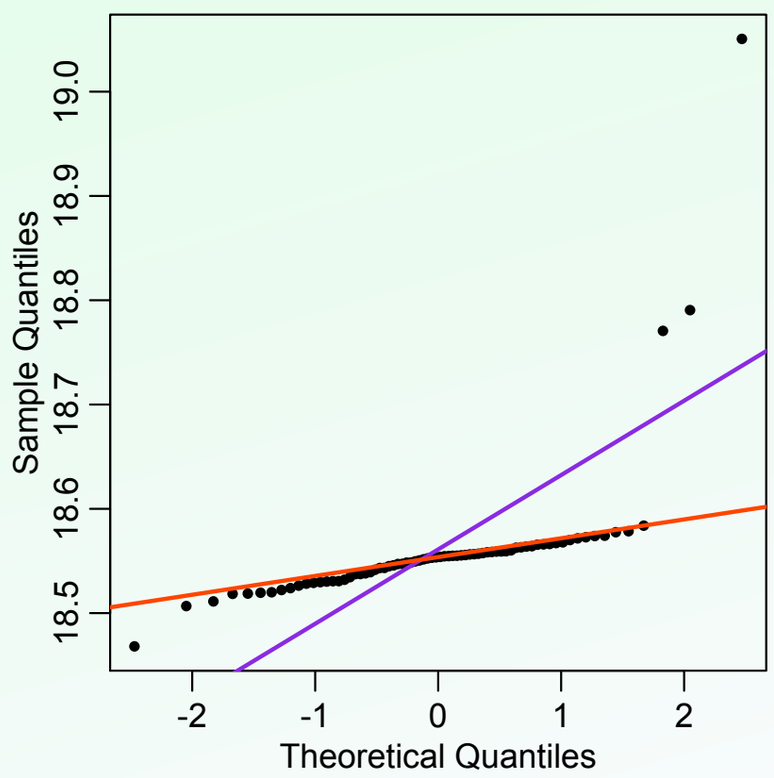
Truth: mixture

$$F = p_0 F_0 + (1 - p_0) F_{\text{cont}}$$

$$F_0 = \mathcal{N}(\mu, \sigma^2)$$

$$\hat{\mu} = \text{median}(X)$$

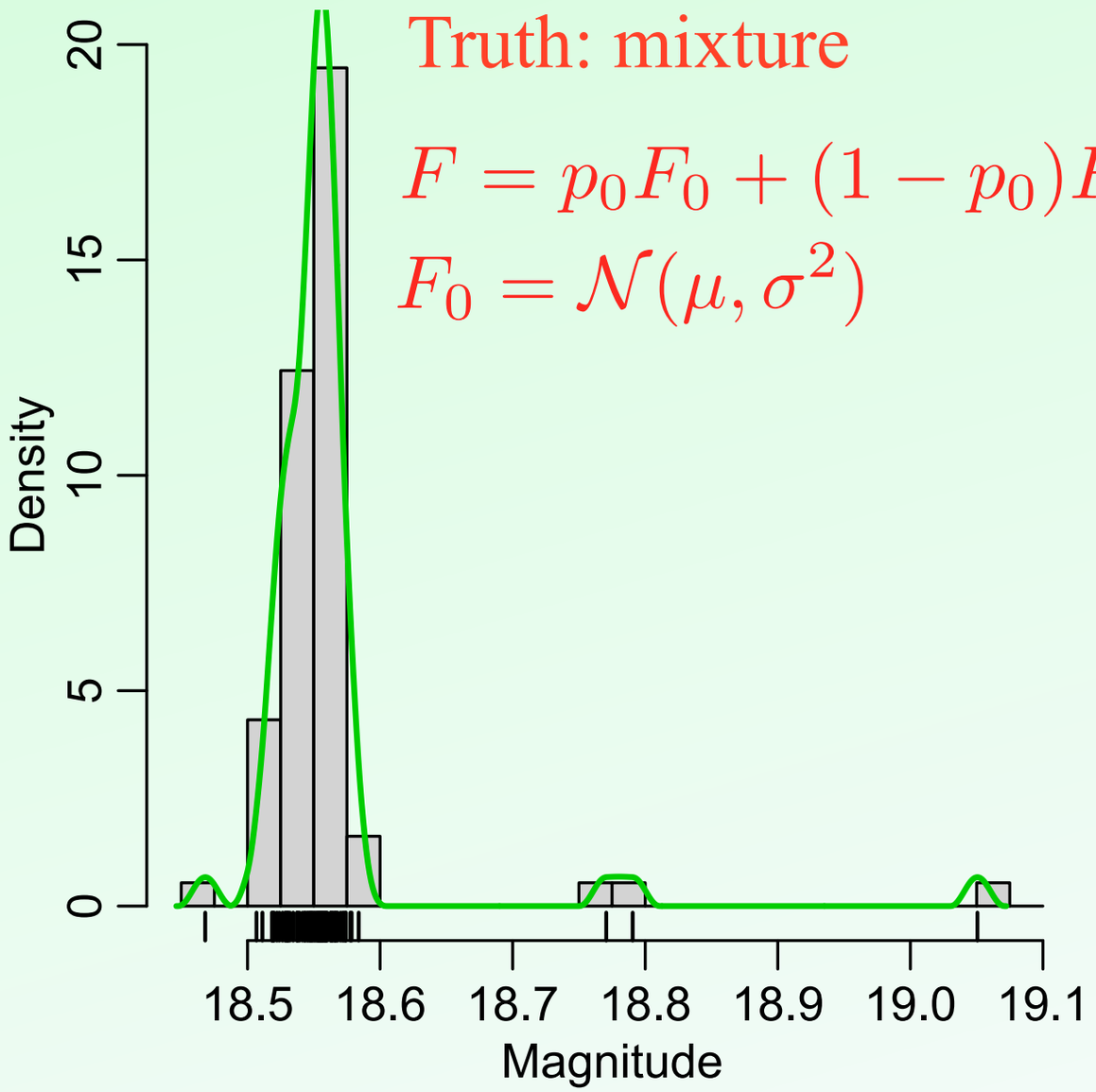
$$\hat{\sigma} = 1.4826\,\text{MAD}(X)$$



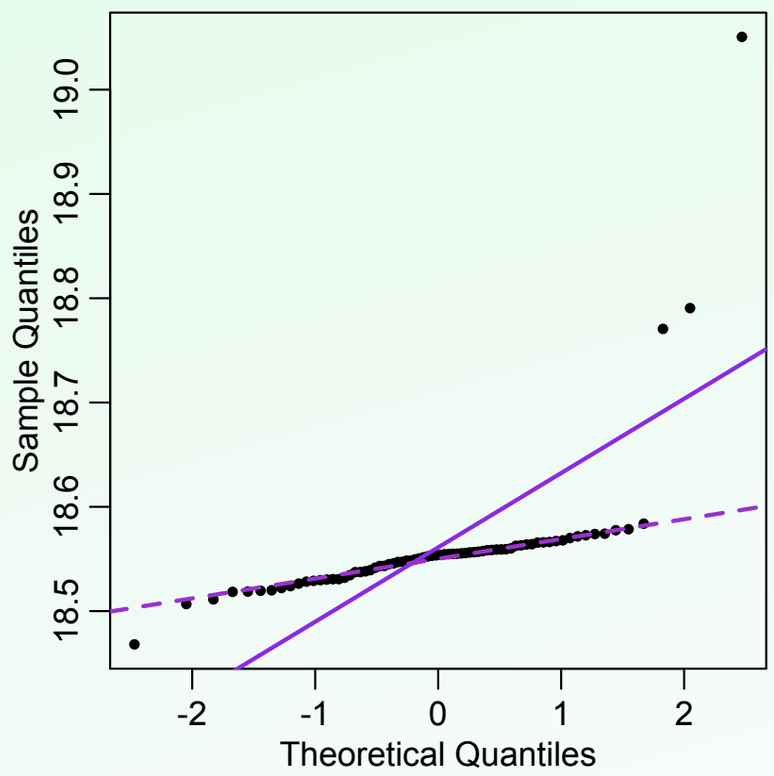
Estimators: method of moments (robustness)

Principle:

1. Assume a distribution $F(x; \theta)$.
2. Alternative 2: use quantile estimators as estimators of the moments.



$$\hat{\mu} = \text{median}(X)$$
$$\hat{\sigma} = \text{IQR}(X)/1.349$$



Estimators: method of moments (robustness)

Principle:

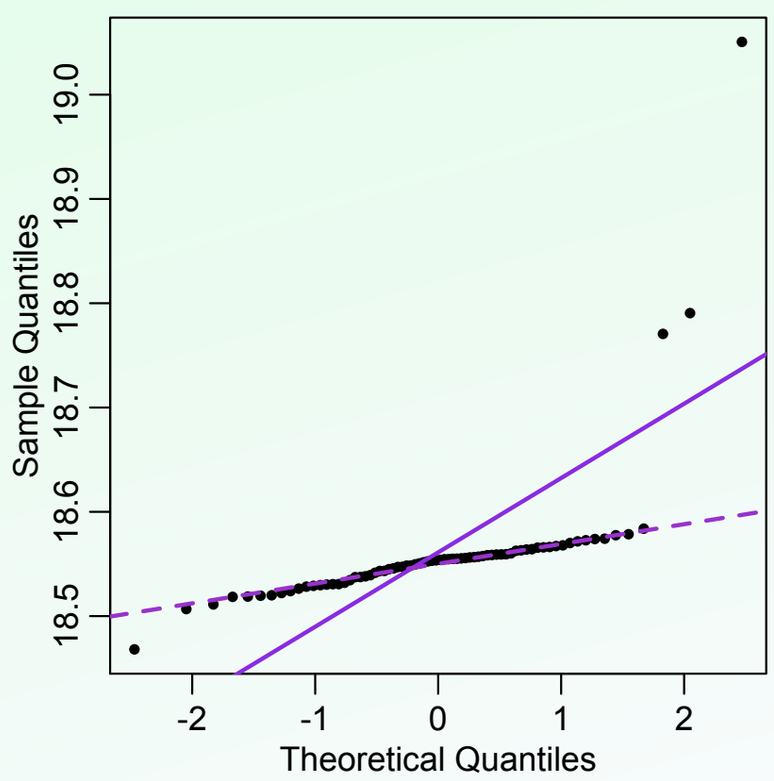
1. Assume a distribution $F(x; \theta)$.
2. Alternative 2: use quantile estimators as estimators of the moments.

$$\hat{\mu} = \text{median}(X)$$
$$\hat{\sigma} = \text{IQR}(X)/1.349$$

Generalization:

L-estimators

`https://en.wikipedia.org/wiki/L-estimator`



Estimators: method of moments (robustness)

Principle:

1. Assume a distribution $F(x; \theta)$.
2. Alternative 3: define loss functions and minimize this loss over the sample.

   Loss function: $\rho(u)$

   We minimize $\sum_{i=1}^{N} \rho(u_i)$, where for example $u_i = x_i - \mu$.

Estimator:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{N} \rho(u_i; \theta)$$

Generalization:

M-estimators

`https://en.wikipedia.org/wiki/M-estimator`

Estimators: method of moments (robustness)

Examples of loss functions:

- sum of squared residuals

  $\rho(u) = u^2$

- sum of absolute residuals

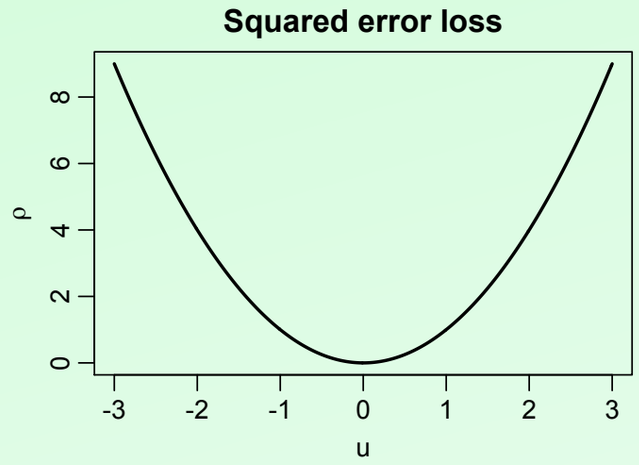
  $\rho(u) = |u|$

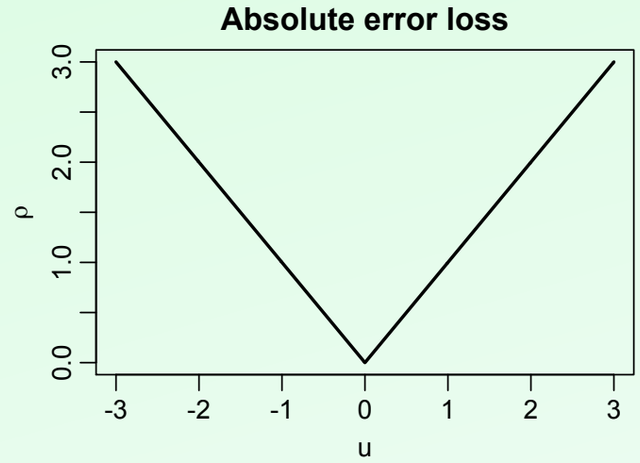
- Huber's loss function

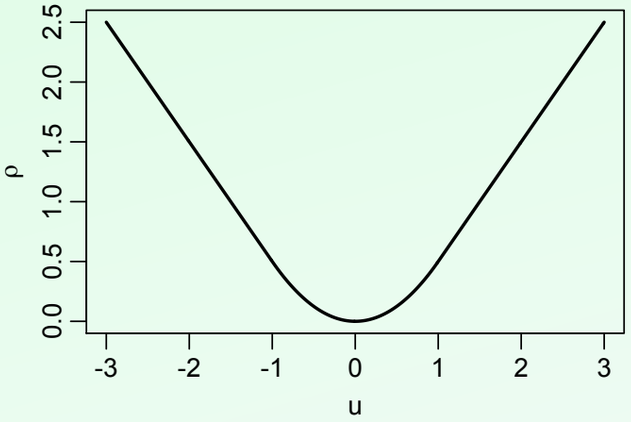
$$\rho(u) = \begin{cases} \frac{1}{2}u^2) & \text{if } |u| \leq \delta \\ \delta(|u| - \frac{1}{2}\delta) & \text{if } |u| > \delta \end{cases}$$

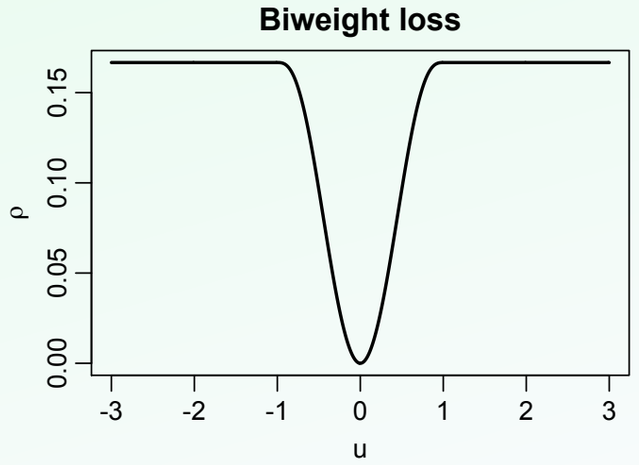
- Tukey's biweight

$$\rho(u) = \begin{cases} \frac{1}{6}[1 - (1 - u^2)^3] & \text{if } |u| \leq 1 \\ \frac{1}{6} & \text{if } |u| > 1 \end{cases}$$









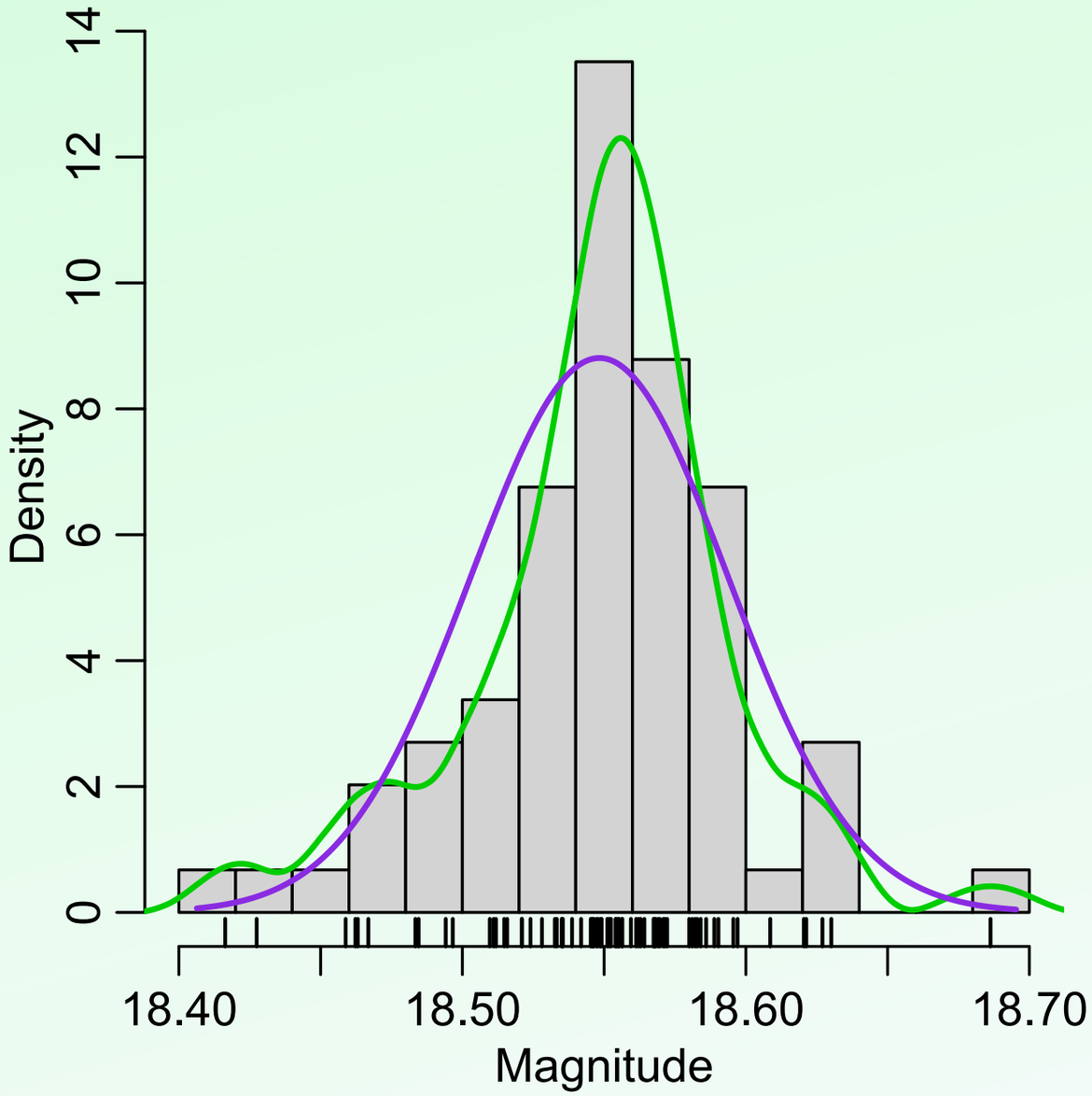
Estimators: method of moments

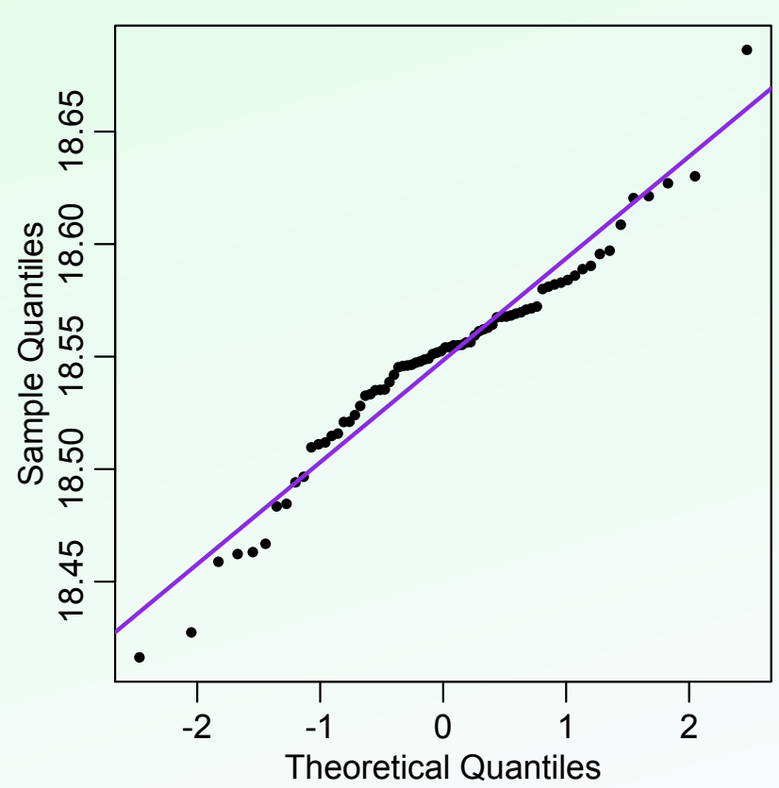
Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.

$$\mathrm{E}(X) = \mu$$

$$\mathrm{Var}(X) = \sigma^2$$

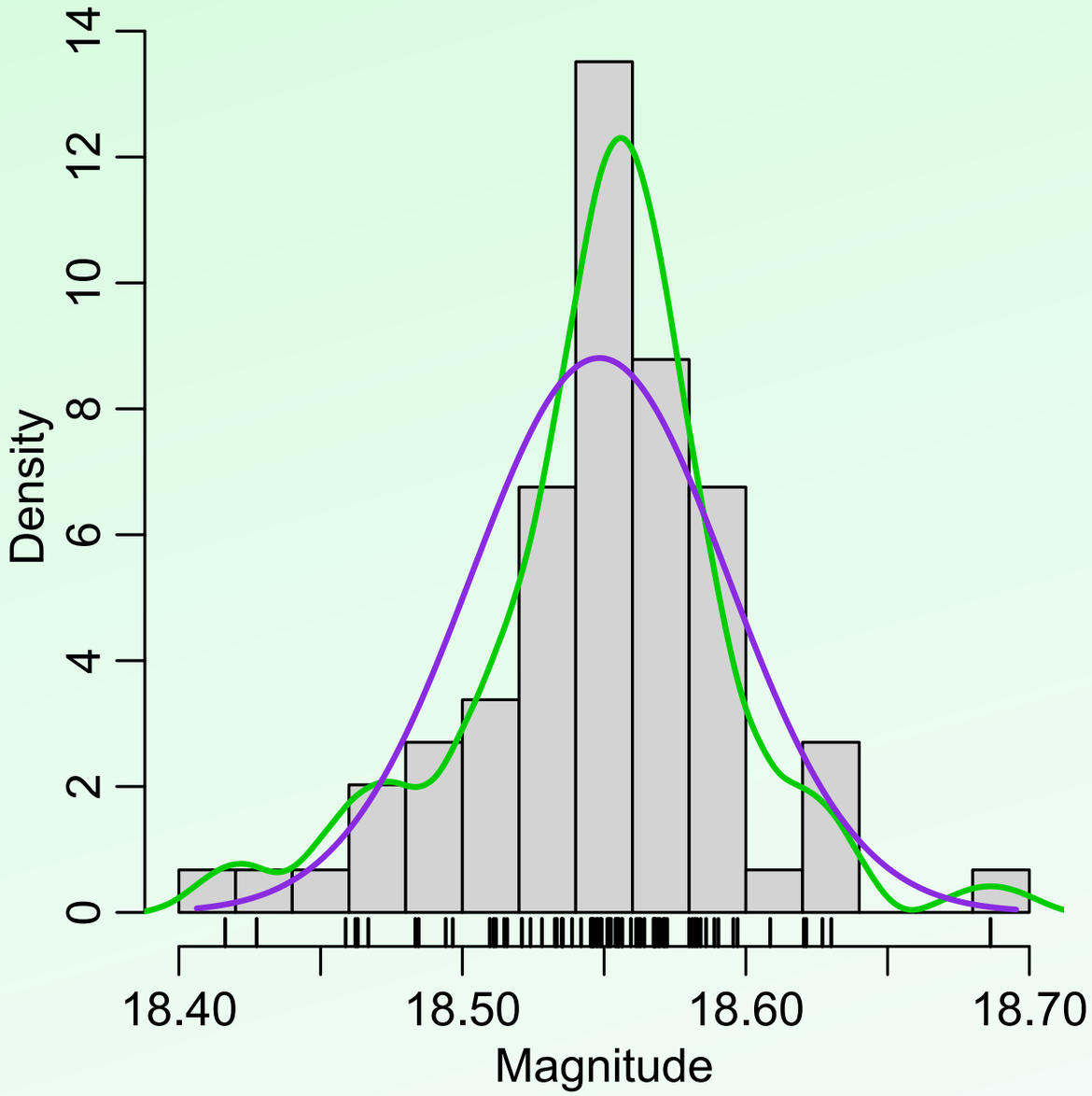




Estimators: method of moments

Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.



$$\mathrm{E}(X) = \mu$$

$$\mathrm{Var}(X) = \sigma^2$$

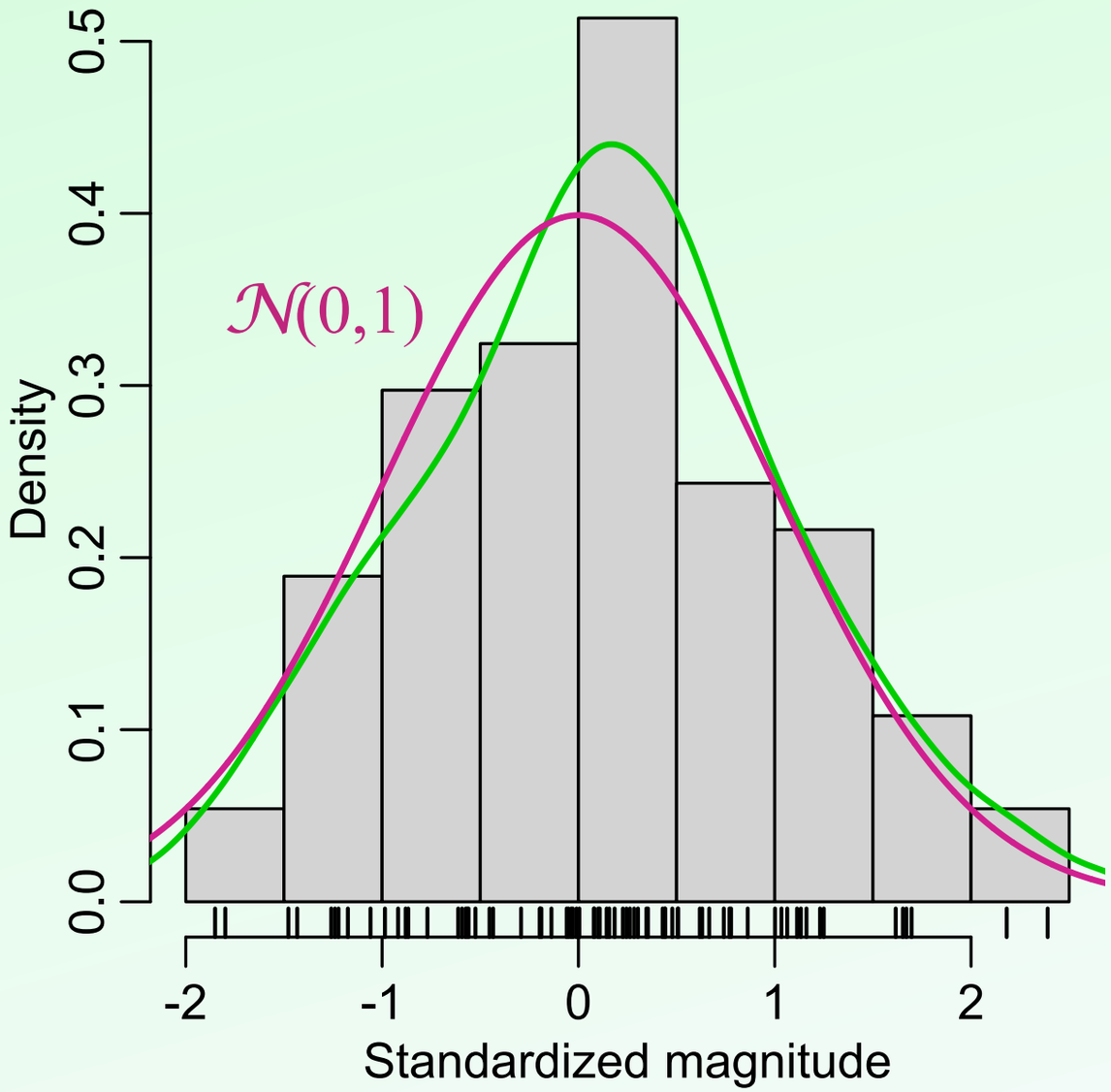
Truth: heteroscedastic

Pairs of
(X_1, σ_1),
$\ldots$,
(X_N, σ_N)
are given

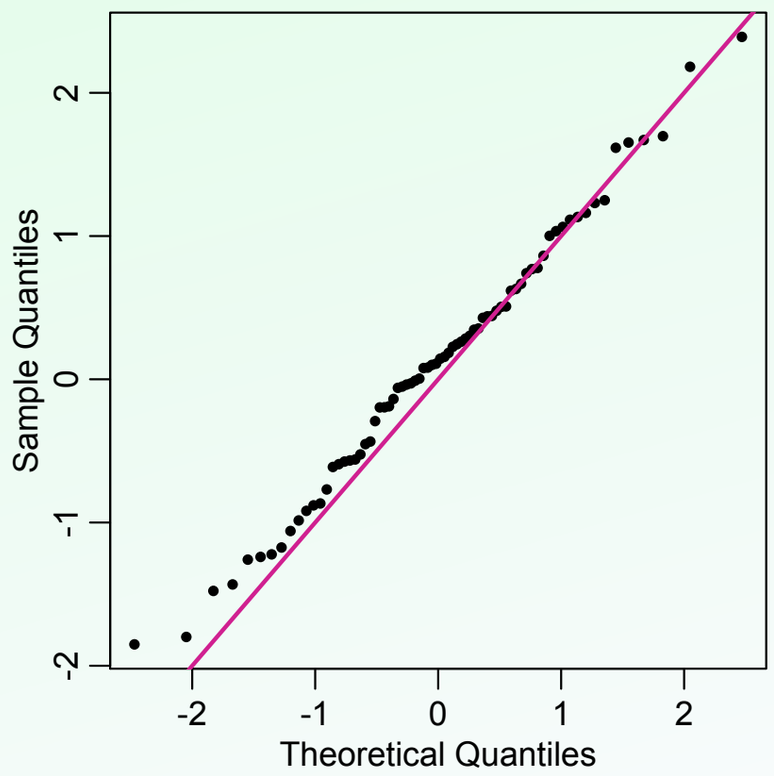
Estimators: method of moments

Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.



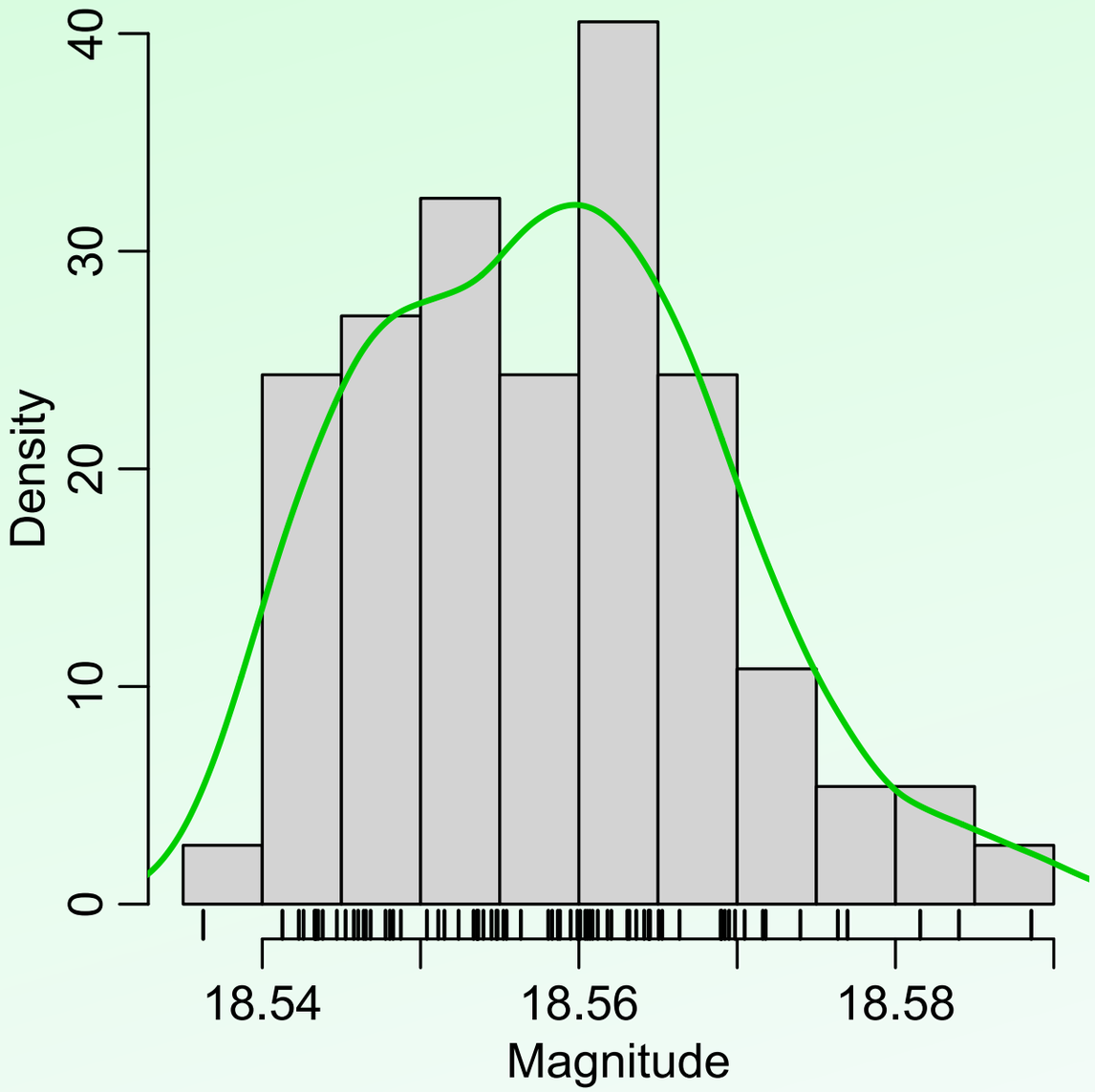
After standardization,
should be $\mathcal{N}(0,1)$



Estimators: method of moments

Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.



$$\mathrm{E}(X) = \mu$$

$$\mathrm{Var}(X) = \sigma^2$$

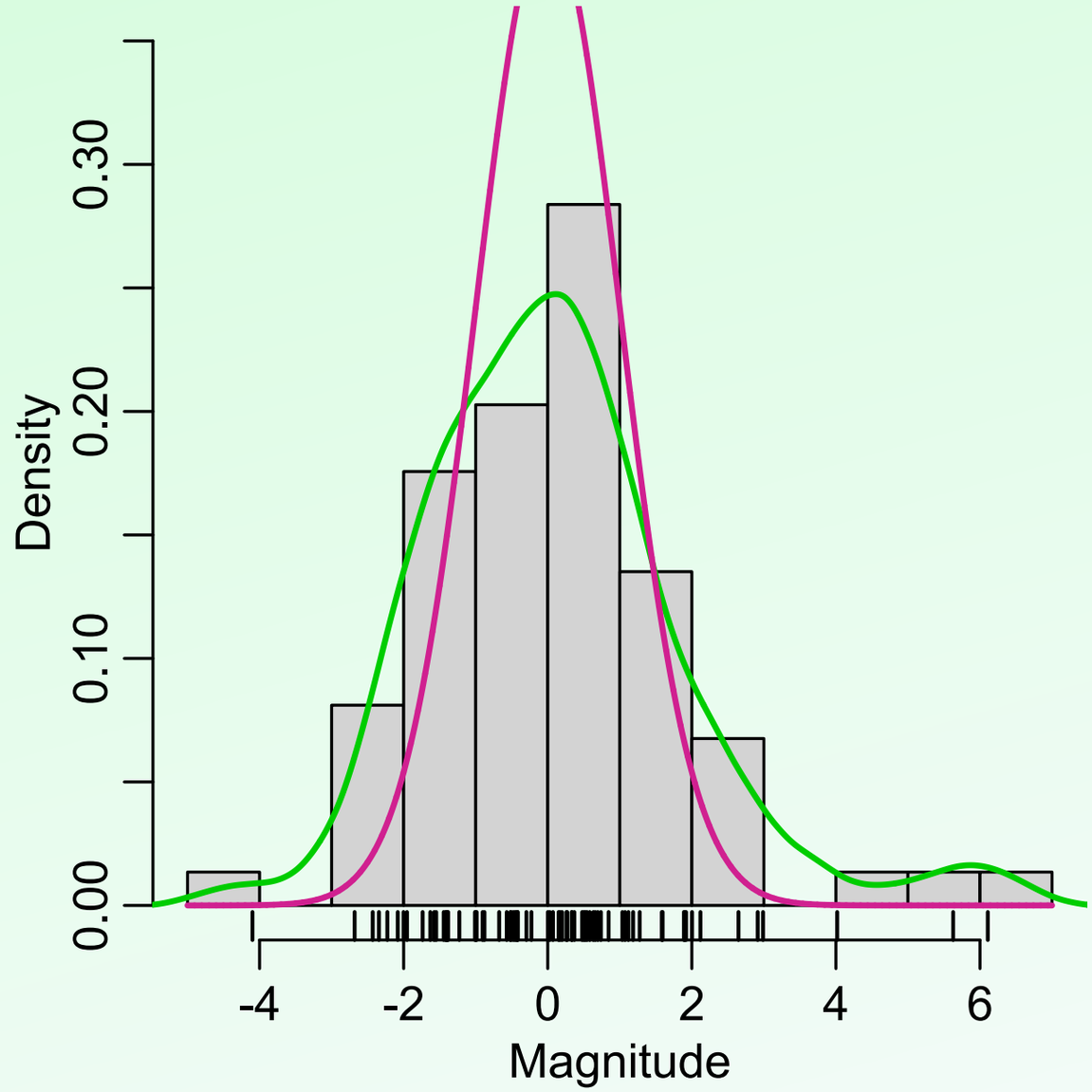
Can we assume
a normal distribution?

Heteroscedastic again:
pairs of
(X_1, σ_1),
$\ldots$,
(X_N, σ_N)
are given

Estimators: method of moments

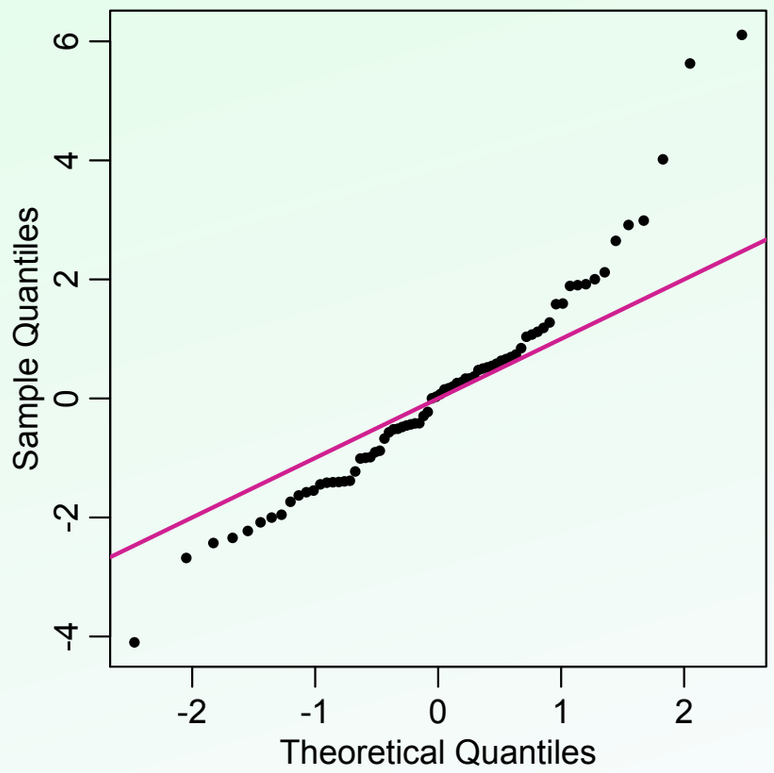
Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.



$$\mathrm{E}(X) = \mu$$

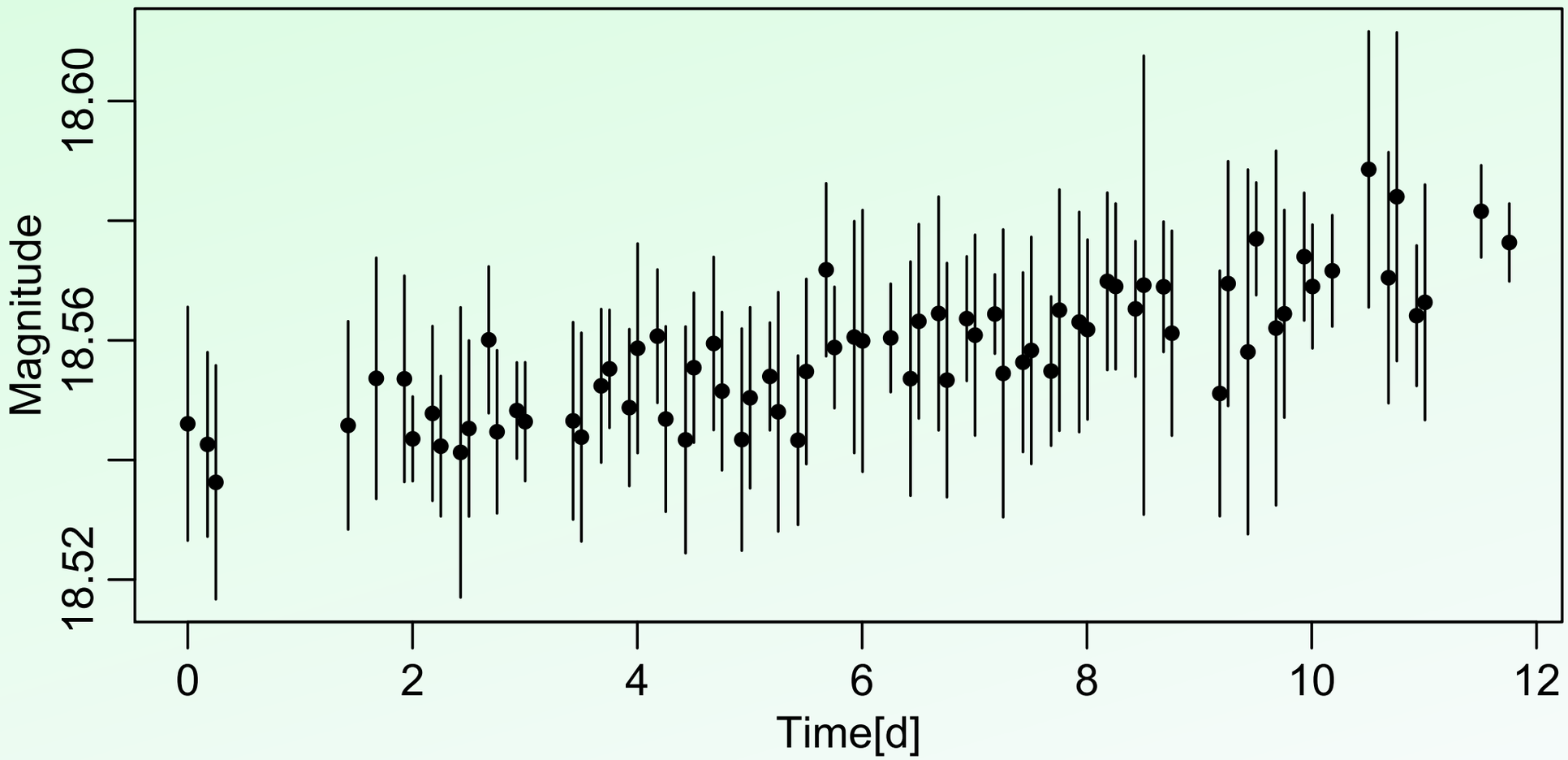
$$\mathrm{Var}(X) = \sigma^2$$



Estimators: method of moments

Principle:

1. Assume a distribution $F(x; \theta)$.
2. Equate the moments of the distribution with the empirical moments.



Modelling:
likelihood methods

Estimators: maximum likelihood

Principle:

1. Assume a (joint!) distribution for the sample: $\boldsymbol{X} \sim F(\boldsymbol{x}, \boldsymbol{\theta})$
2. Find the parameters that maximize the probability to obtain the sample

Log-likelihood: the probability density of the sample, with the observed values taken as fixed, and regarding the parameters as its arguments:

$$\ell(\boldsymbol{\theta}; x_1, \ldots, x_N) = \log f(x_1, \ldots, x_N; \boldsymbol{\theta})$$

Estimating equation:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; x_1, \ldots, x_N)$$

Maximize the log-likelihood with an appropriate method!

Estimators: maximum likelihood

Uncertainty of $\hat{\boldsymbol{\theta}}$: the estimators above are themselves random variables (since the X_i were so).

Theorem: Under some conditions on the likelihood, the maximum likelihood estimator is asymptotically multivariate normal:

if $\boldsymbol{\theta}_0$ is the true value, then

$$\lim_{n\to\infty} \mathrm{E}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) = \boldsymbol{\theta}_0$$

$$\lim_{n\to\infty} \mathrm{Var}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) = I^{-1}$$

where $\quad I = \mathrm{E}_{f(x|\theta_0)}\left(-\dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}}\right)$ (Fisher information).

In practice, the expected value is replaced by the sample average at $\hat{\boldsymbol{\theta}}$.

Standard deviation of the ML estimator $\hat{\theta}_i$: $\quad s_i = \sqrt{([I^{-1}]_{ii})}$

Confidence intervals for $\hat{\theta}_i$: $\quad [\hat{\theta}_i - z_{1-\alpha/2} s_i,\ \hat{\theta}_i + z_{1-\alpha/2} s_i]$

Estimators: maximum likelihood

Cramér-Rao bound:

Let $T(\boldsymbol{\theta})$ be an unbiased estimator of a parameter vector of a model. Then

$$\mathrm{Cov}[T(\boldsymbol{\theta})] \geq I(\boldsymbol{\theta})^{-1}$$

If an estimator attains this limit, it is said to be *efficient*.

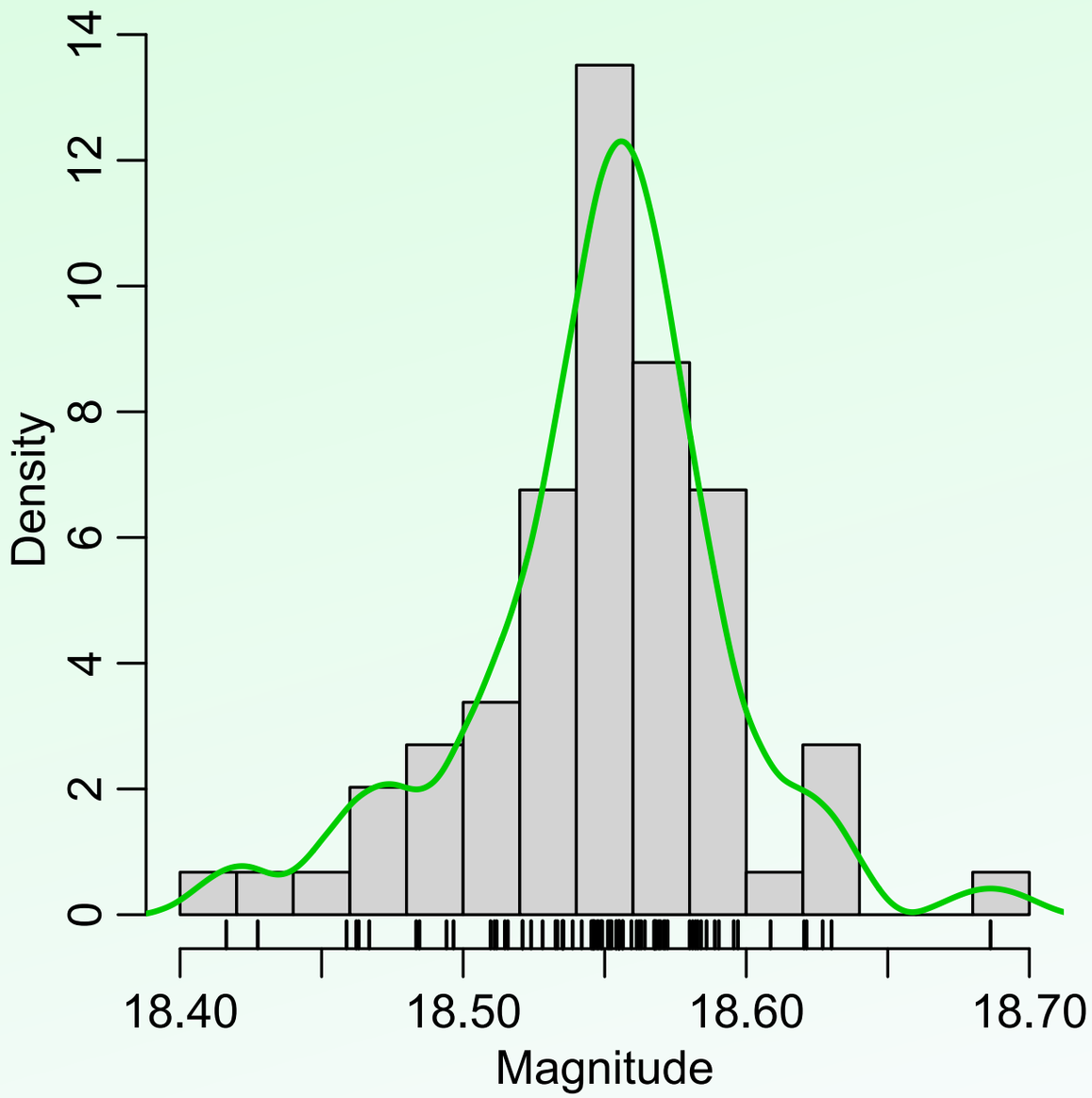
Since the maximum likelihood is asymptotically unbiased, it is efficient when the sample size is large (tends to infinity). Thus, maximum likelihood estimators have an optimal character among the unbiased estimators.

Biased estimators can improve on the Cramér-Rao bound; you need to find the bias correction, though.

Estimators: maximum likelihood

Examples: $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$ (our heteroscedastic sample)

$$\ell(\boldsymbol{\theta}; x_1, \ldots, x_N) = -\sum_{i=1}^{N} \log \sigma_i - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma_i^2} + \text{cst.}$$



Estimators: maximum likelihood

Examples: $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$ (our heteroscedastic sample)

$$\ell(\boldsymbol{\theta}; x_1, \ldots, x_N) = -\sum_{i=1}^{N} \log \sigma_i - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma_i^2} + \text{cst.}$$

known

In this case, we can easily give an analytic solution:

1. derive the log-likelihood w.r.t. the parameter of interest (μ)
2. equate it with zero
3. solve the equation.

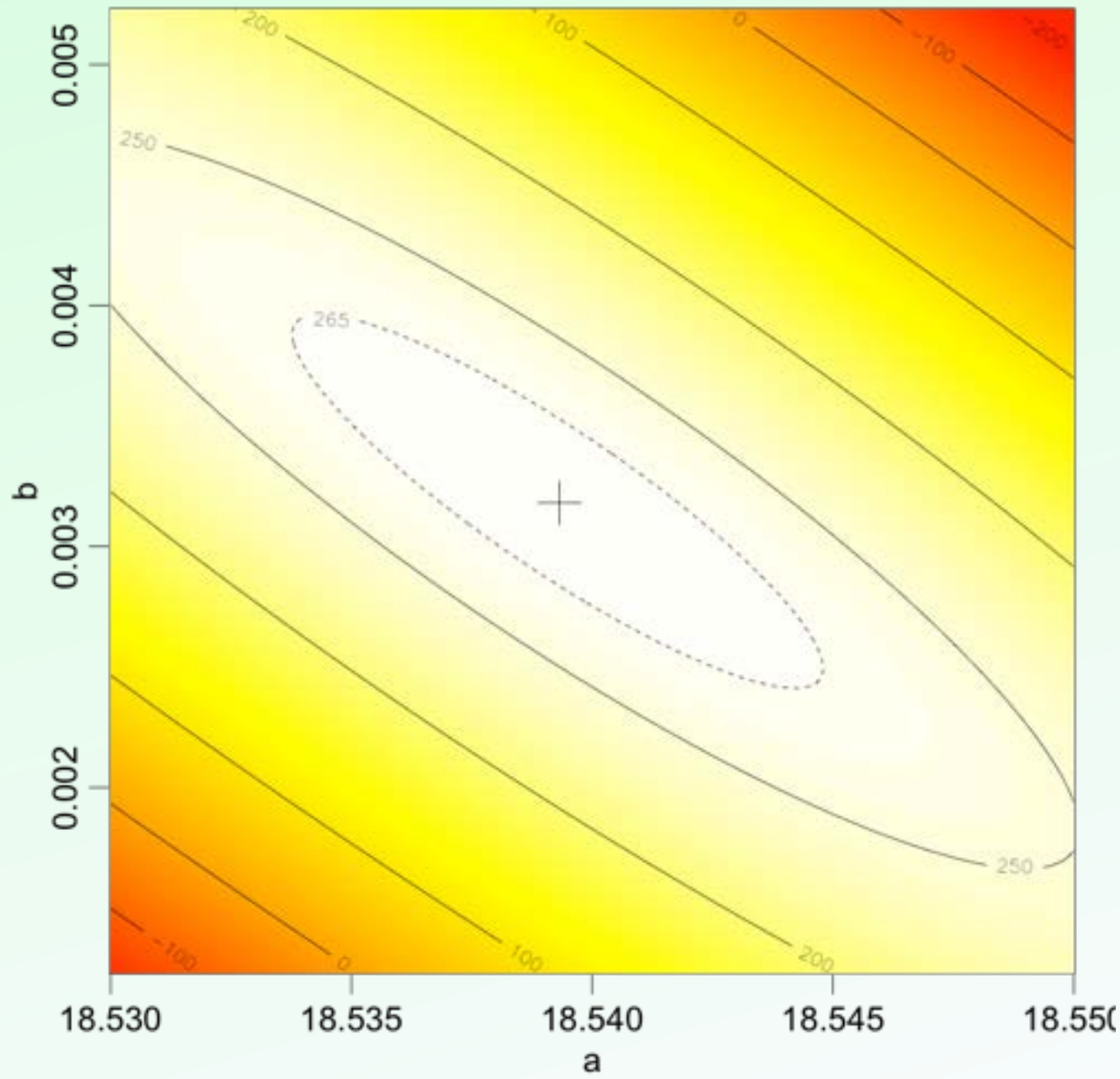
Estimator: the weighted sample mean

$$\hat{\mu} = \frac{1}{W} \sum_{i=1}^{N} w_i x_i, \quad \text{where} \quad w_i = \frac{1}{\sigma_i^2} \quad \text{and} \quad W = \sum_{i=1}^{N} w_i.$$

Estimators: maximum likelihood

Examples: $X_i \sim \mathcal{N}(a + b\,t_i, \sigma_i^2)$ (our time-dependent sample)

$$\ell(a, b; x_1, \ldots, x_N) = -\sum_{i=1}^{N} \frac{[x_i - (a + b\,t_i)]^2}{2\sigma_i^2} + \text{cst.}$$



Estimators: maximum likelihood

Examples: $X_i \sim \mathcal{N}(a + b\,t_i, \sigma_i^2)$ (our time-dependent sample)

$$\ell(a, b; x_1, \ldots, x_N) = -\sum_{i=1}^{N} \frac{[x_i - (a + b\,t_i)]^2}{2\sigma_i^2} + \text{cst.}$$

Solution (in this case, closed formula):

$$\hat{b} = \frac{\sum_{i=1}^{N} w_i (t_i - \bar{t}) x_i}{\sum_{i=1}^{N} w_i (x_i - \bar{x}) x_i}$$

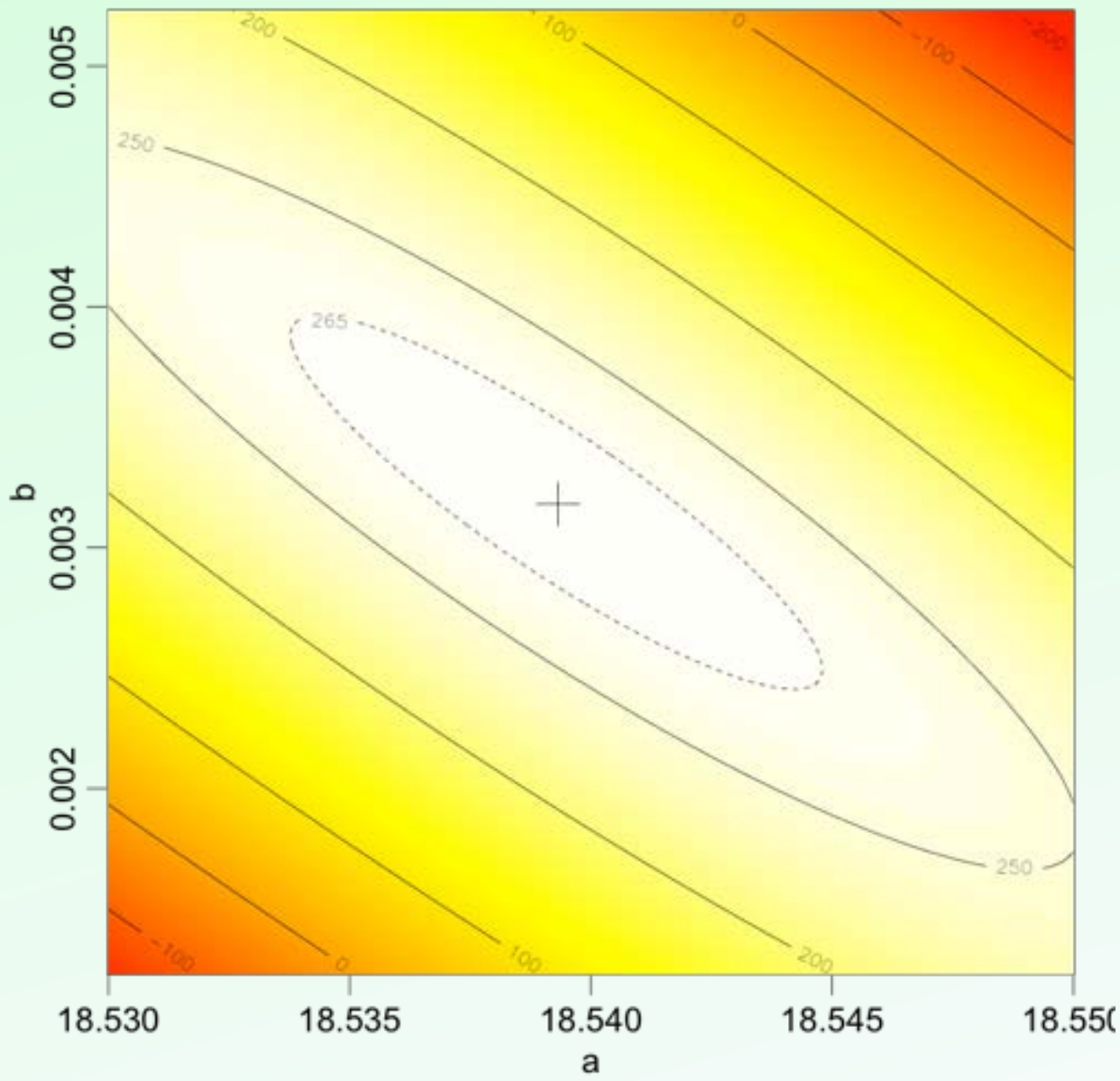
$$\hat{a} = \bar{x} - \hat{b}\bar{t}$$

where $\bar{x} = \sum_{i=1}^{N} w_i x_i$ and $\bar{t} = \sum_{i=1}^{N} w_i t_i$

with $w_i = \frac{1}{\sigma_i^2}$, $W = \sum_{i=1}^{N} w_i$.

Estimators: maximum likelihood

Uncertainty of $\hat{b}$ and $\hat{a}$: In this case, with normally distributed errors, theory is exact!



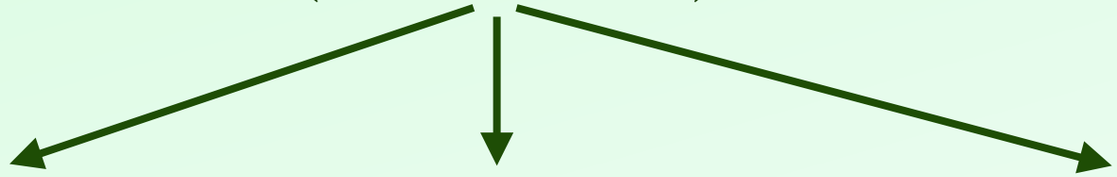
Estimators: maximum likelihood

Generalized: $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \Sigma)$

$$\ell(\boldsymbol{\beta}; \boldsymbol{X}) = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\log\det\boldsymbol{\Sigma} + \text{cst.}$$

Solution:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}X^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$



$\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$ $\quad$ $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ $\quad$ $\boldsymbol{\Sigma}$ general

σ unknown, must also be estimated

ordinary least squares

$\boldsymbol{\Sigma}$ known: the above formula (weighted least squares)

$\boldsymbol{\Sigma}$ unknown: must also be estimated

$\boldsymbol{\Sigma}$, if non-constrained, contains $N(N+1)/2$ unknowns: must be constrained

Estimators: maximum likelihood

Examples: t_ν-distribution with $\boldsymbol{\theta} = \nu$ and

$$f(x \mid \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Log-likelihood:

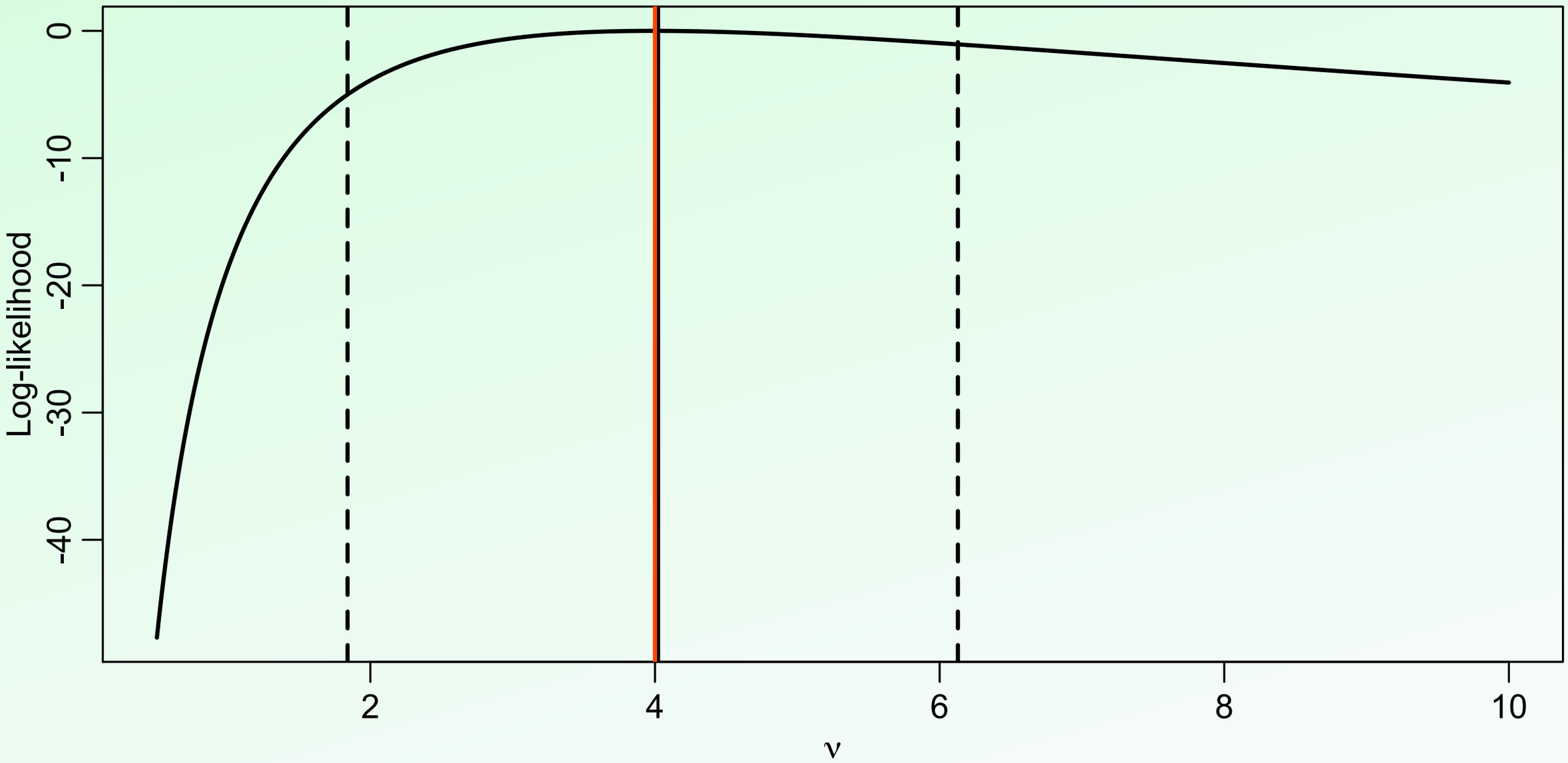
$$\ell(\theta; x_1, \ldots, x_N) = N \log \Gamma(\frac{\nu+1}{2}) - N \log \Gamma(\frac{\nu}{2}) - \frac{N}{2} \log \nu - \frac{\nu+1}{2} \sum_{i=1}^{N} \log\left(1 + \frac{x_i^2}{\nu}\right) + \text{ cst.}$$

Estimate by numerical optimization

Variance of $\hat{\nu}$, according to theory: compute the second derivative of the log-likelihood, take its negative, and invert it

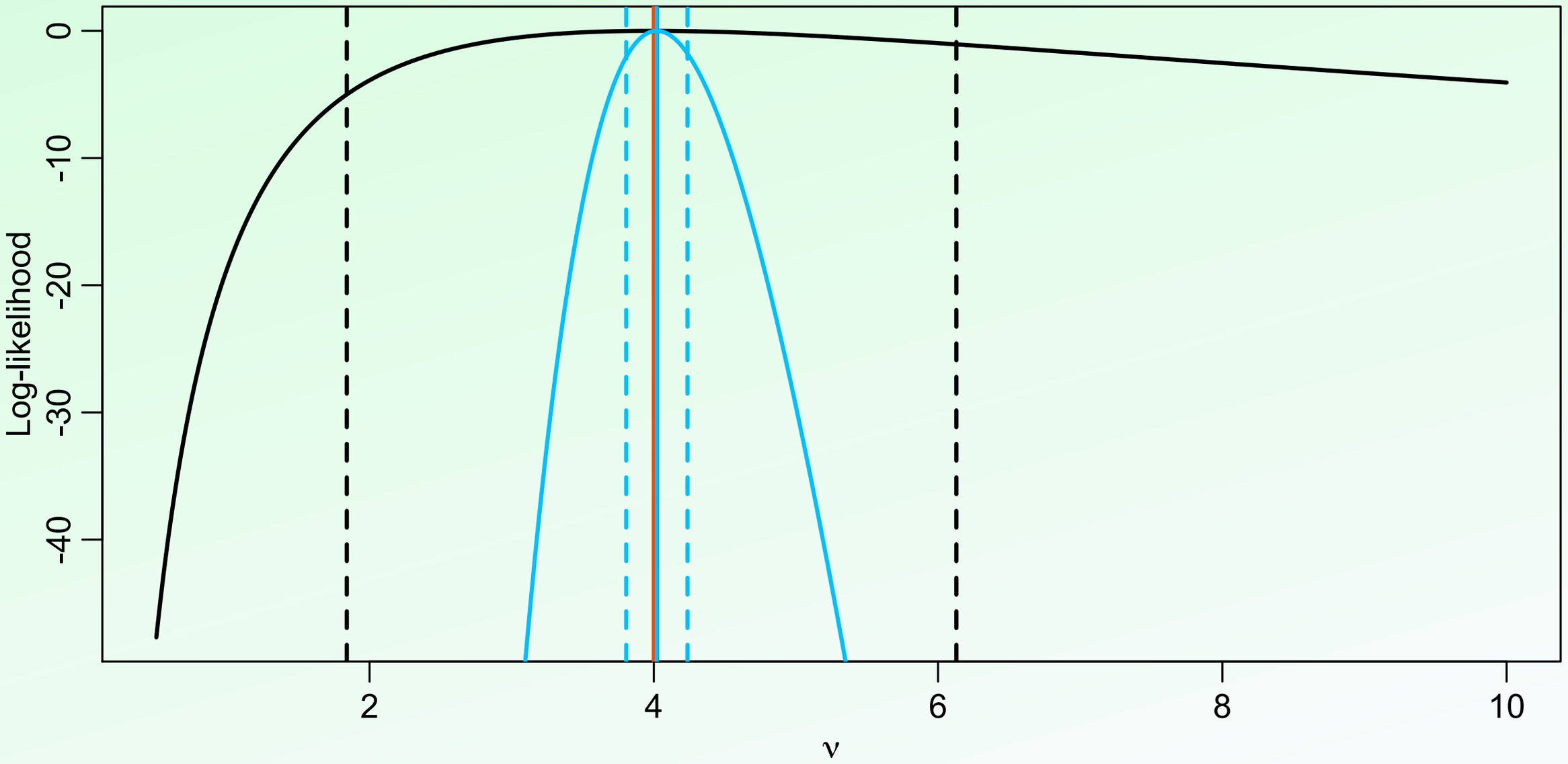
Estimators: maximum likelihood

Examples: t_ν-distribution; simulate a sample of 130 t_4 variates



Estimators: maximum likelihood

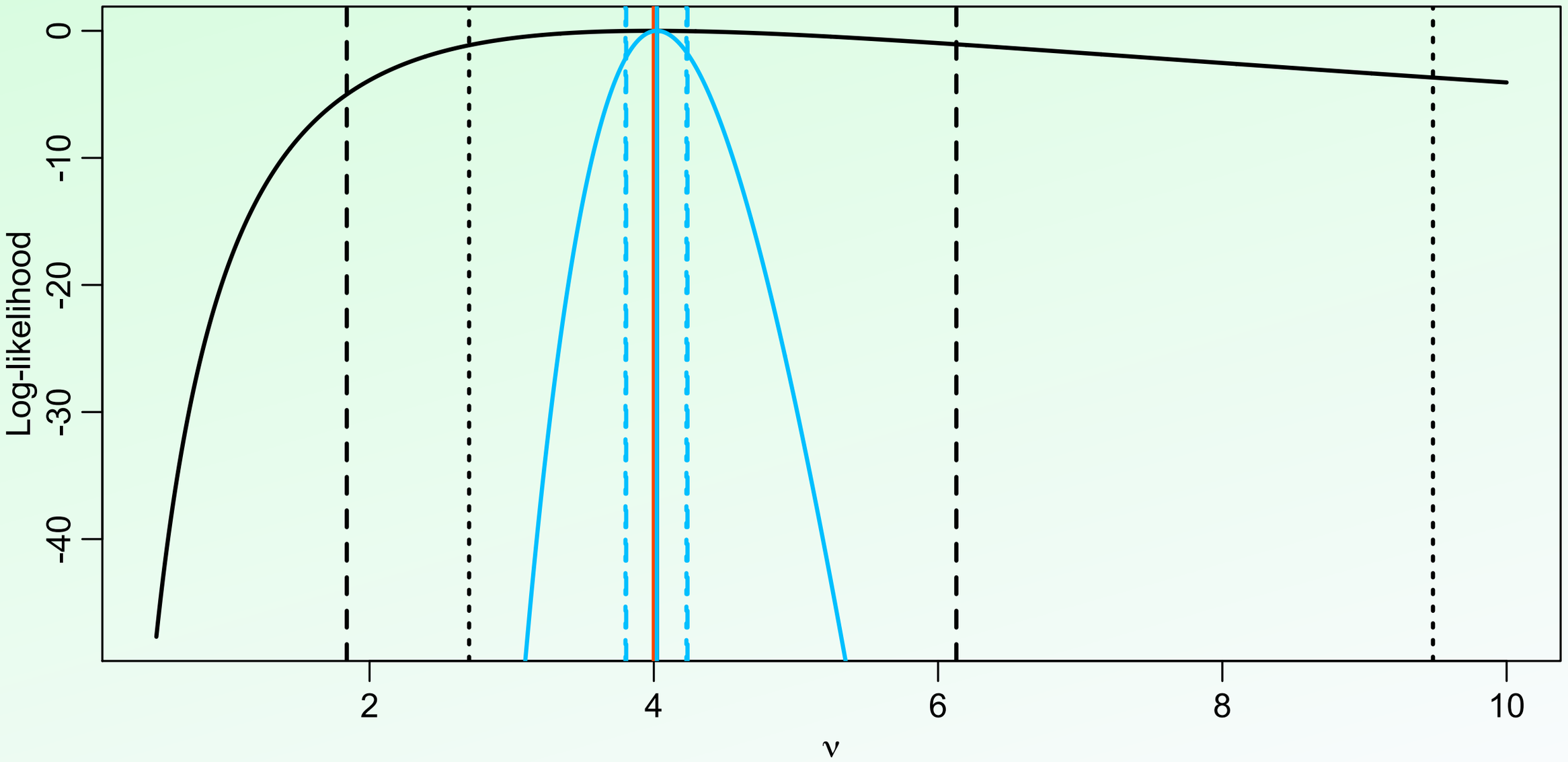
Examples: t_ν-distribution; simulate a sample of 13000 t_4 variates



Estimators: maximum likelihood

Examples: t_ν-distribution; simulate a sample of 13000 t_4 variates

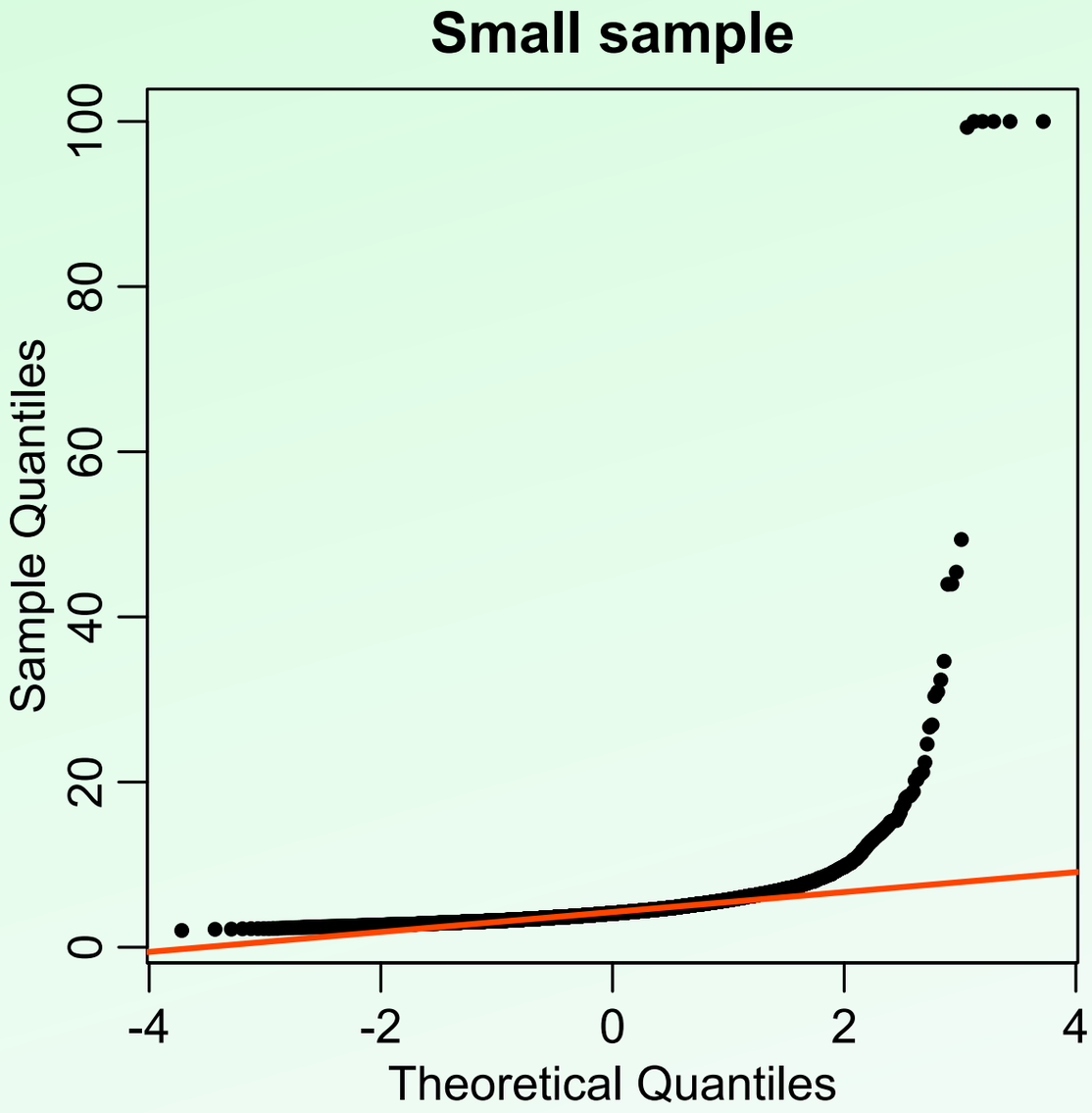
Repeat now both simulations 5000 times, and take the (0.025,0.975) quantiles of the 5000 estimates!

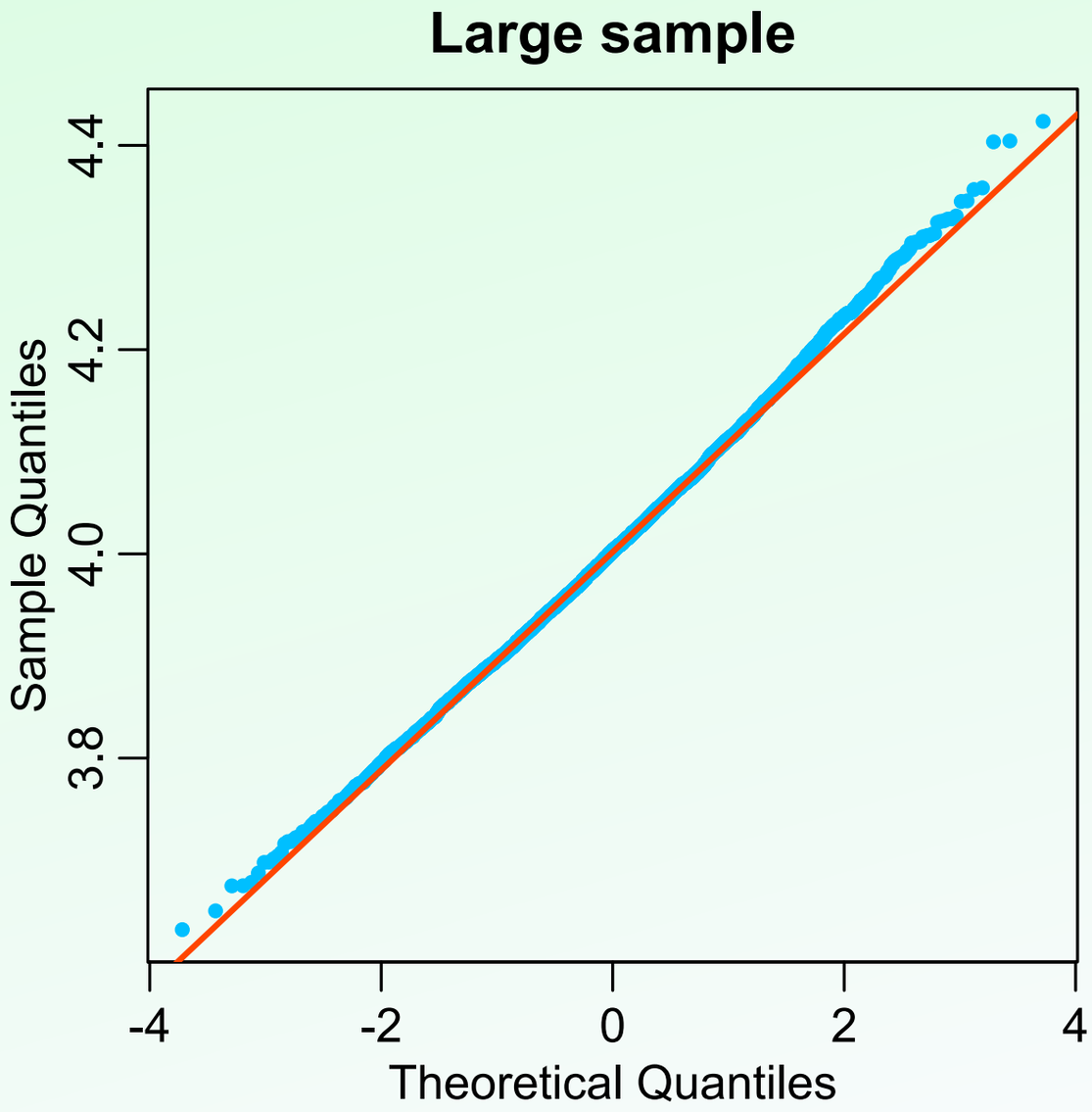


Estimators: maximum likelihood

Examples: t_ν-distribution; simulate a sample of 13000 t_4 variates

Repeat now both simulations 5000 times, and take the (0.025,0.975) quantiles of the 5000 estimates!



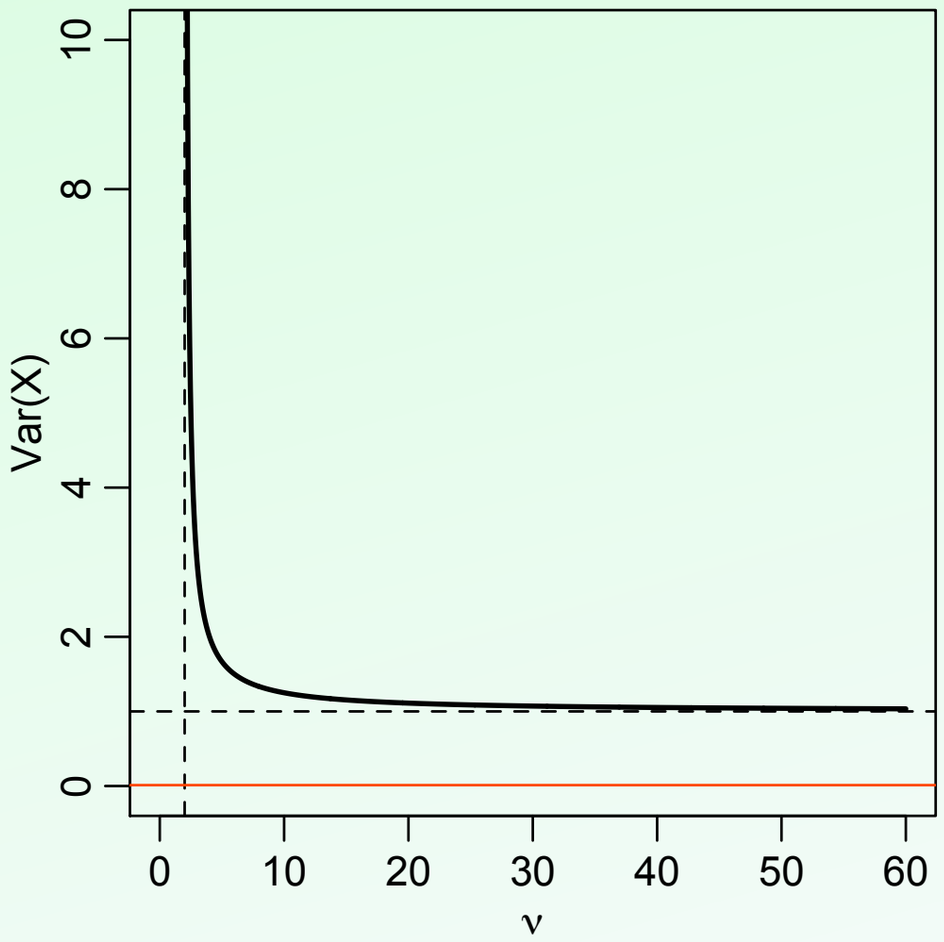


Estimators: maximum likelihood

Examples: our scaled-shifted t_ν-distributed sample from earlier

In this case, we should assume that the errors around the true magnitude m are rescaled (by a factor a)

$$\ell(\theta; x_1, \ldots, x_N) = N \log \Gamma(\frac{\nu+1}{2}) - N \log \Gamma(\frac{\nu}{2}) - \frac{N}{2} \log \nu$$
$$- \frac{\nu+1}{2} \sum_{i=1}^{N} \log\left(1 + \frac{(x_i - m)^2}{a^2 \nu}\right) + \text{cst.}$$

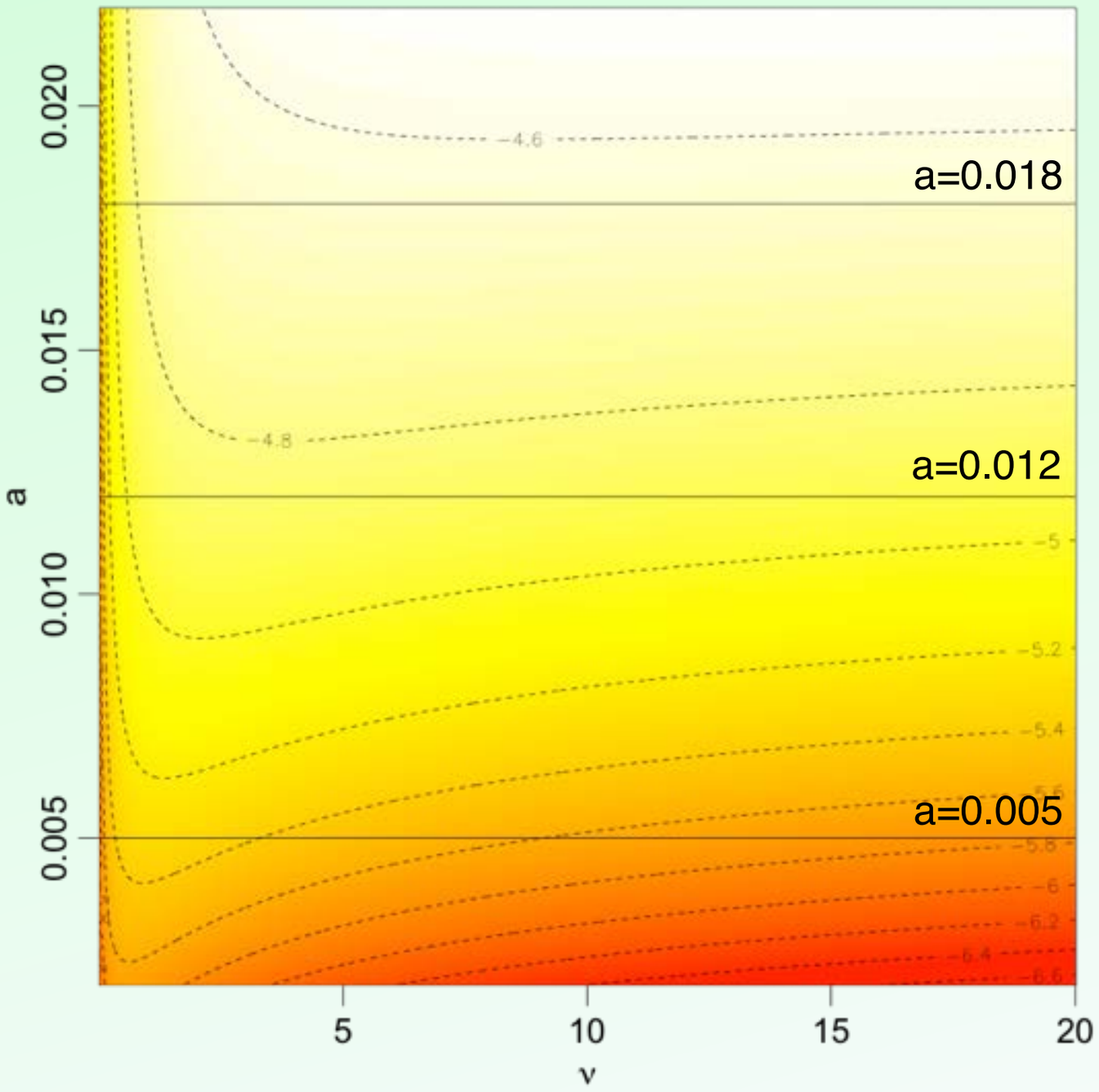


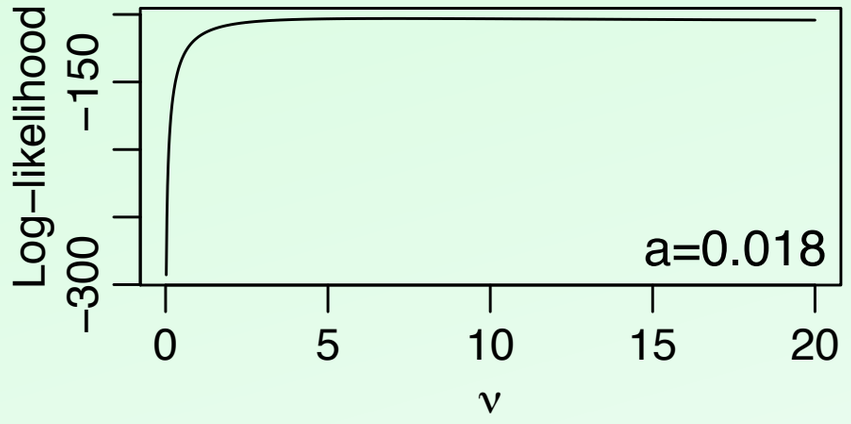
Var(X) as a function of ν if $X \sim t_\nu$

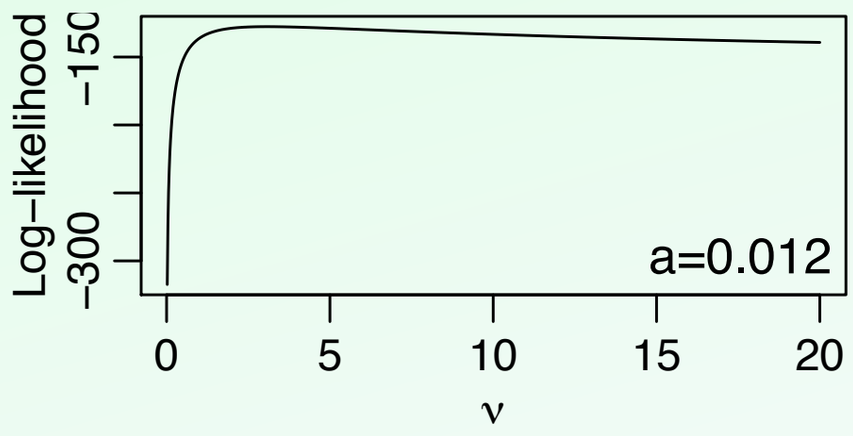
Empirical variance in the sample

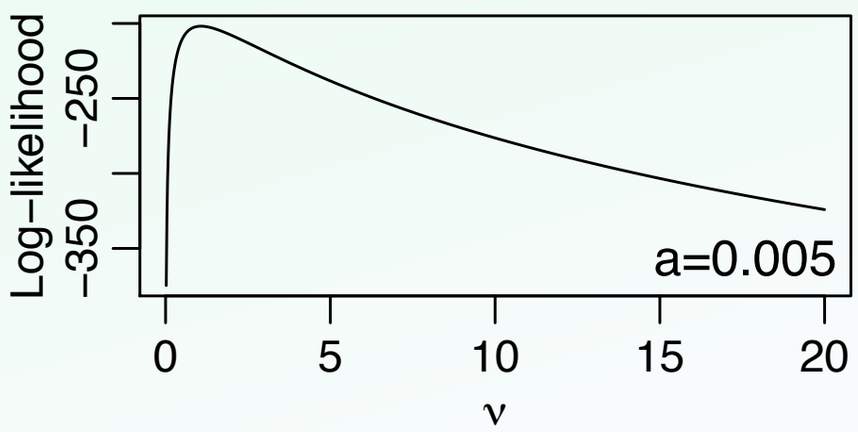
Estimators: maximum likelihood

Examples: our scaled-shifted t_ν-distributed sample from earlier



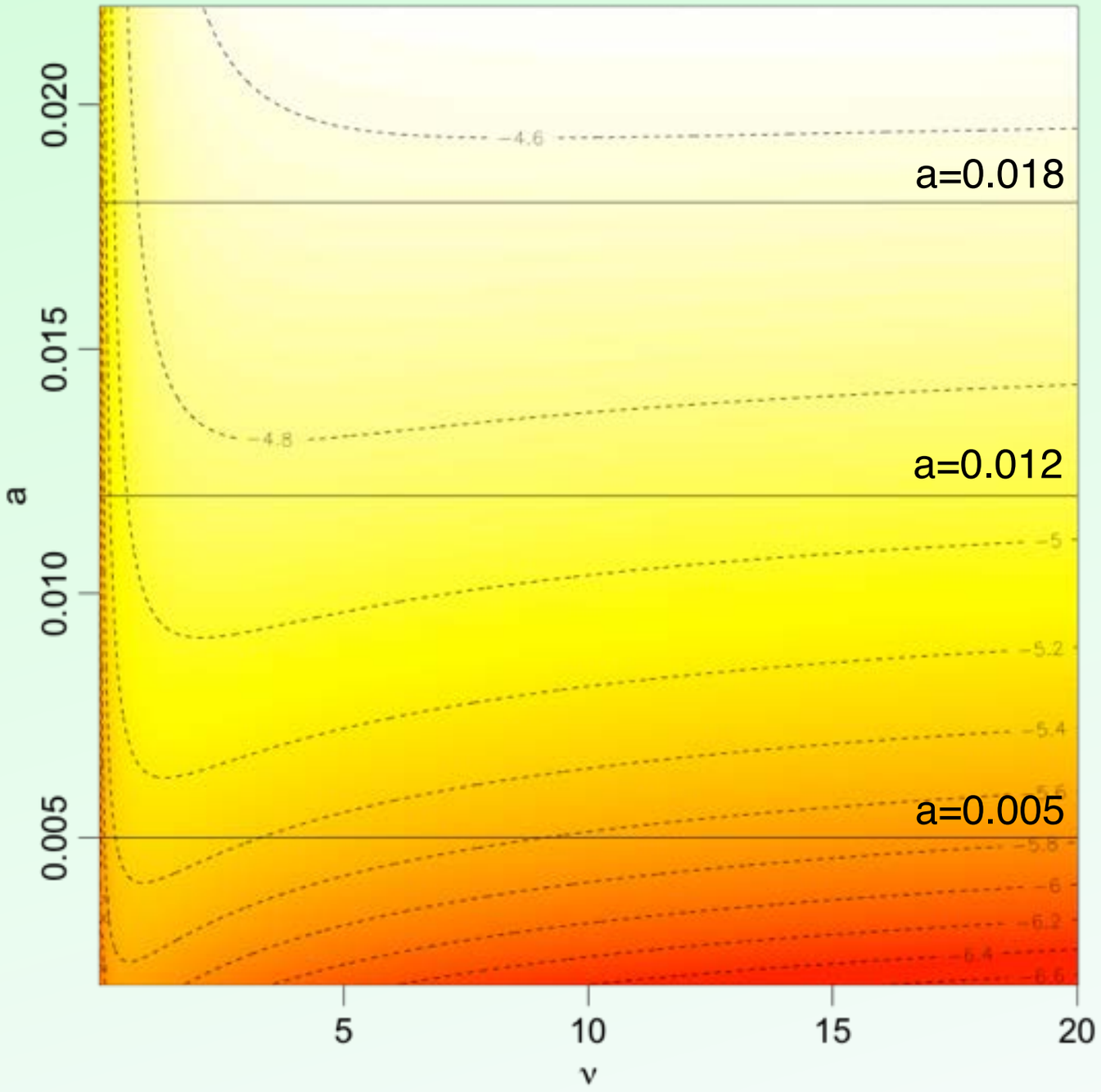






Estimators: maximum likelihood

Examples: our scaled-shifted t_ν-distributed sample from earlier



The first derivative wrt a:

$$\frac{\partial \ell}{\partial a} = \sum_{i=1}^{N} \frac{1}{a} \frac{\frac{(x_i-\mu)^2}{\nu}}{a^2 + \frac{(x_i-\mu)^2}{\nu}}$$

$$\frac{\partial \ell}{\partial a} = 0 \quad \text{iff} \quad x_i = \mu \ \forall \ i$$

$$\lim_{a \to \infty} \frac{\partial \ell}{\partial a} \longrightarrow 0$$

Estimators: maximum likelihood

Generalized:

- the distribution is non-Gaussian
  - e.g. waiting times between events (exponential or gamma distr.,…)
  - e.g. counting processes in $\mathbb{R}^d$ (Poisson, …)
  - e.g. fractions (binomial, multinomial, …)
  - e.g. mixtures (distributions of galaxies or stars in colour space)
- the model is non-linear
- need to estimate the basis functions at the same time
- high-dimensional problems: $N \ll$ number of parameters
- hierarchical models: even the parameters are random
- estimate of incomplete or missing data
- semi- and non-parametric methods
- etc.

Classical hypothesis testing

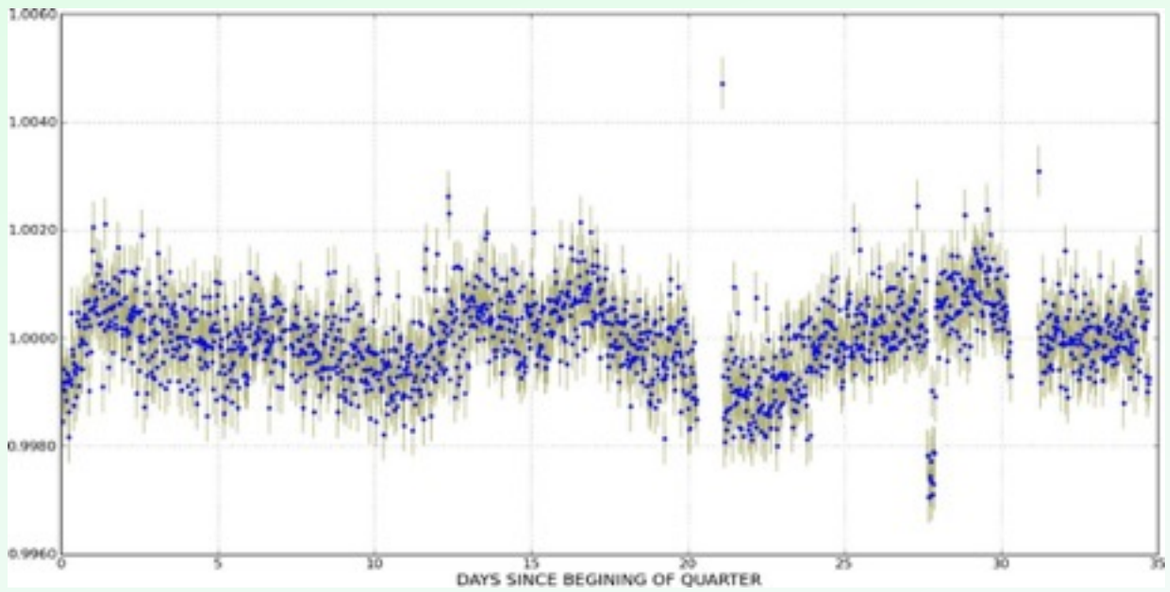
Suppose we have a theory predicting an interesting effect:

- the expansion of the universe is accelerating
- B-mode polarization in the cosmic microwave background
- an Earth-like planet in a star's habitable zone
- the Cepheids are not perfectly stable "cosmic clocks"
- …

Its effects on observables can be formulated in a model where its presence or absence can be detected, and we also have some appropriate data to decide.

Null hypothesis:

No planet here, only the star



Alternative hypothesis:

There is a planet

Needs to be put in a precise form: white noise, correlated white noise…

Classical hypothesis testing

Null hypothesis
$\mathbf{H_0}$

Alternative hypothesis
$\mathbf{H_1}$

Select test statistic T
and significance level α

the distribution F_0 of T under $\mathbf{H_0}$ should be known!

no

the distribution of T under $\mathbf{H_1}$ not necessary to know

yes

Check model assumptions

- independence of observations?
- distributional assumption (misspecified models must be treated in a different way)

no

yes

Model assessment

1. Compute the observed value t_{sample} of T
2. Compute p-value $\mathrm{P}(T \geq t_{\text{sample}}) = 1 - F_0(t_{\text{sample}})$
3. Decision: if p-value $> \alpha$: **cannot reject** $\mathrm{H_0}$
   if p-value $\leq \alpha$: **reject** $\mathrm{H_0}$

Classical hypothesis testing

Meaning of a "test statistic value t_{obs} significant at the 95% confidence level:

The assumed null model, $\mathbf{H_0}$ would produce a value such as t_{obs} or more extreme with a probability equal to or less than 0.05.

But, though it is supportive:

- it does not prove $\mathbf{H_1}$: maybe another alternative could have produced the effect (and the extreme test statistic)
- how likely was $\mathbf{H_1}$ originally? Conditional arguments: if prior data did not support it, then taking into account this low prior probability, the P-value weakens a lot

		Decision	
		$\mathbf{H_0}$ not rejected	**$\mathbf{H_0}$ rejected**
Truth	**$\mathbf{H_0}$ true**	$P_{H_0}(\mathbf{H_0}$ is not rejected$) = 1-\alpha$	Type I error: $P_{H_0}(\mathbf{H_0}$ is rejected$) = \alpha$
	$\mathbf{H_1}$ true	Type II error: $P_{H_1}(\mathbf{H_0}$ is not rejected$) = 1-\beta$	$P_{H_1}(\mathbf{H_0}$ is rejected$) = \beta$

Likelihood ratio test (LR)

Suppose we have two nested models:

- M with parameters (ψ, λ);
- M_0 with parameters (ψ_0, λ),

where ψ_0 is some specified, fixed value of ψ (often 0). The parameter vector ψ is of dimension p.

Let the corresponding **log-likelihoods** be $\ell(\psi, \lambda)$ and $\ell(\psi_0, \lambda)$.

Maximize both, to obtain $\ell(\hat{\psi}, \hat{\lambda})$ and $\ell(\psi_0, \hat{\lambda}_{\psi_0})$.

Then the **likelihood ratio statistic**

$$W(\psi_0) = 2[\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi_0, \hat{\lambda}_{\psi_0})]$$

is distributed approximately as

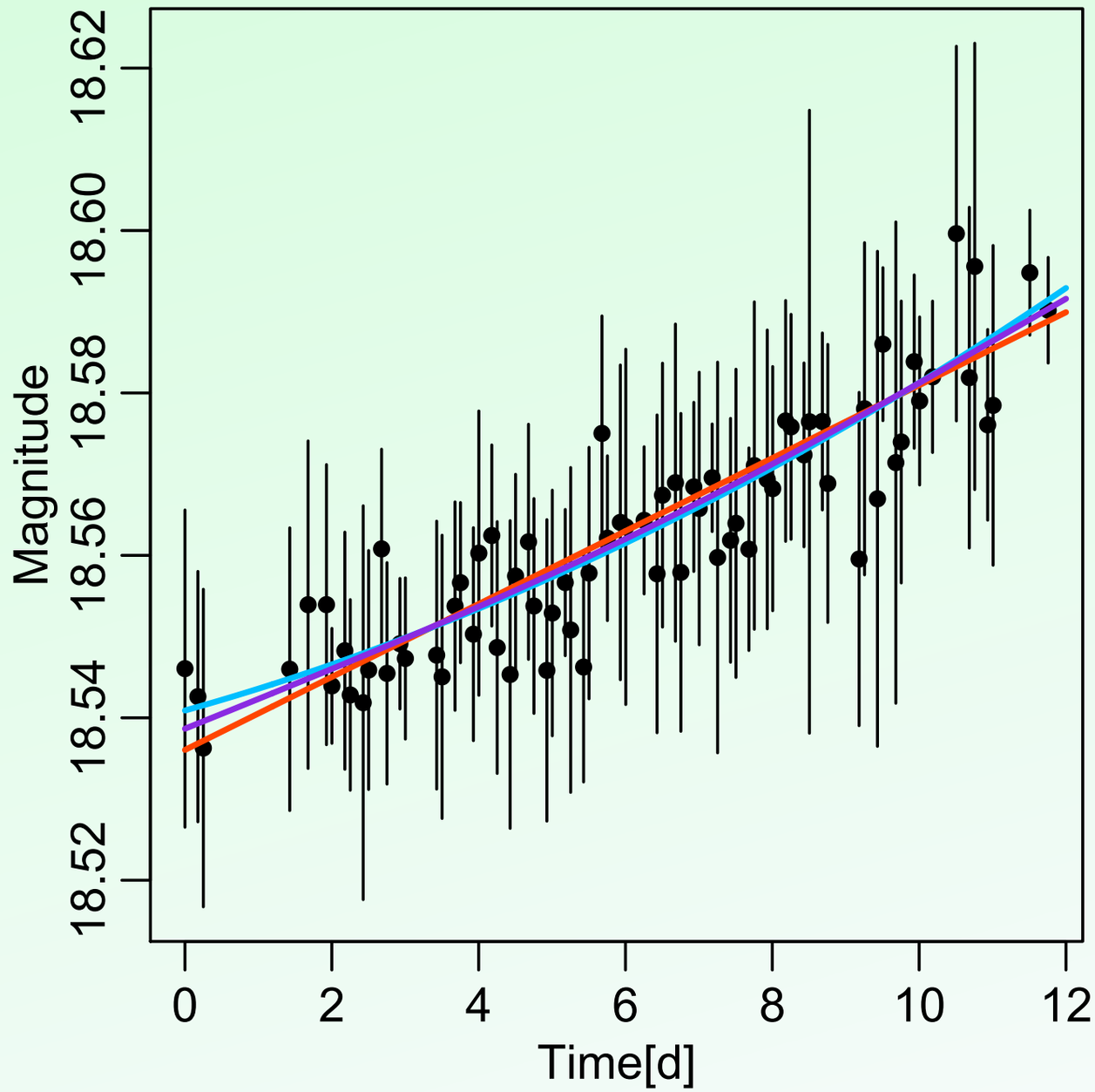
$$W(\psi_0) \dot{\sim} \chi^2_p$$

Some classical tests

- **Tests for the mean**
  - *z*-test (variance is known)
  - *t*-test (variance is estimated)
  - two-sample *t*-test, paired difference test
- **Tests for the variance**
- **Correlation tests:**
  - Fisher
  - Ljung-Box, Durbin-Watson
- **Adequacy tests**
- **Distribution tests:**
  - Kolmogorov-Smirnov, Anderson-Darling
  - Wilcoxon
  - Tests for normality: Jarque-Bera, omnibus, Shapiro-Wilks
- **Tests for regression parameters**
- **Likelihood ratio tests**

Model selection

Which model to choose? Linear, quadratic, or smooth spline?



Model selection

Which model to choose? Linear, quadratic, or smooth spline?

Most frequently used methods:

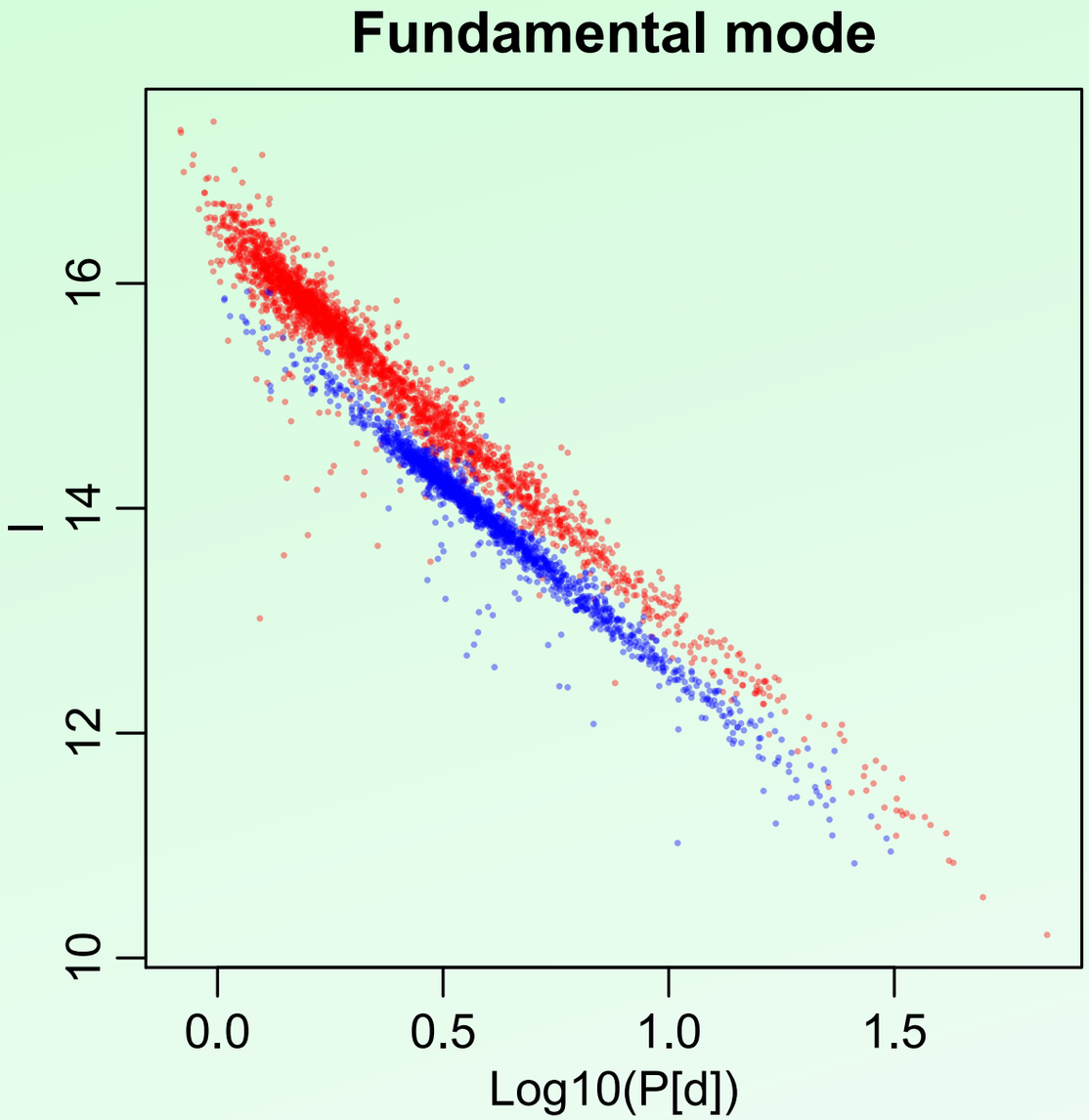
- Likelihood Ratio (LR)
  - only for nested models
  - it is permissive, and does not take into account the fact that a larger model will inevitably improve the fit
- Akaike Information Criterion $\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}) + 2p$
  - also for non-nested models
  - not consistent: if n goes to infinity, it selects a larger model than the true
  - more conservative than LR
- Bayes Information Criterion $\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}}) + p \log N$
  - also for non-nested models
  - consistent: if n goes to infinity, it selects the correct model
  - more conservative than LR or AIC, but can be too conservative for small samples
- Many modified versions

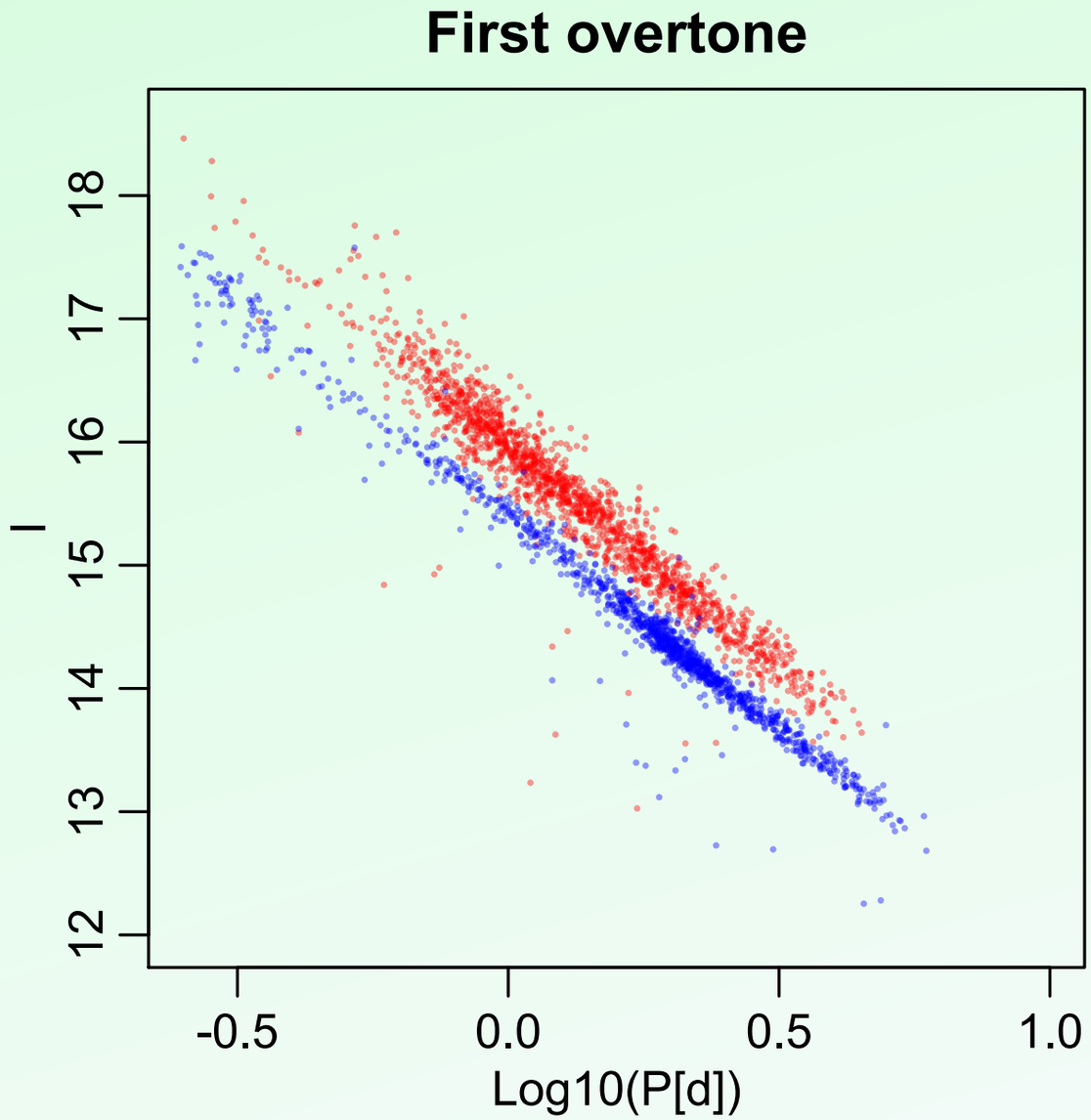
Good luck!

Exercise:

Period-Luminosity relationship of Cepheids

Data: OGLE-III classical Cepheids





Data and the fit

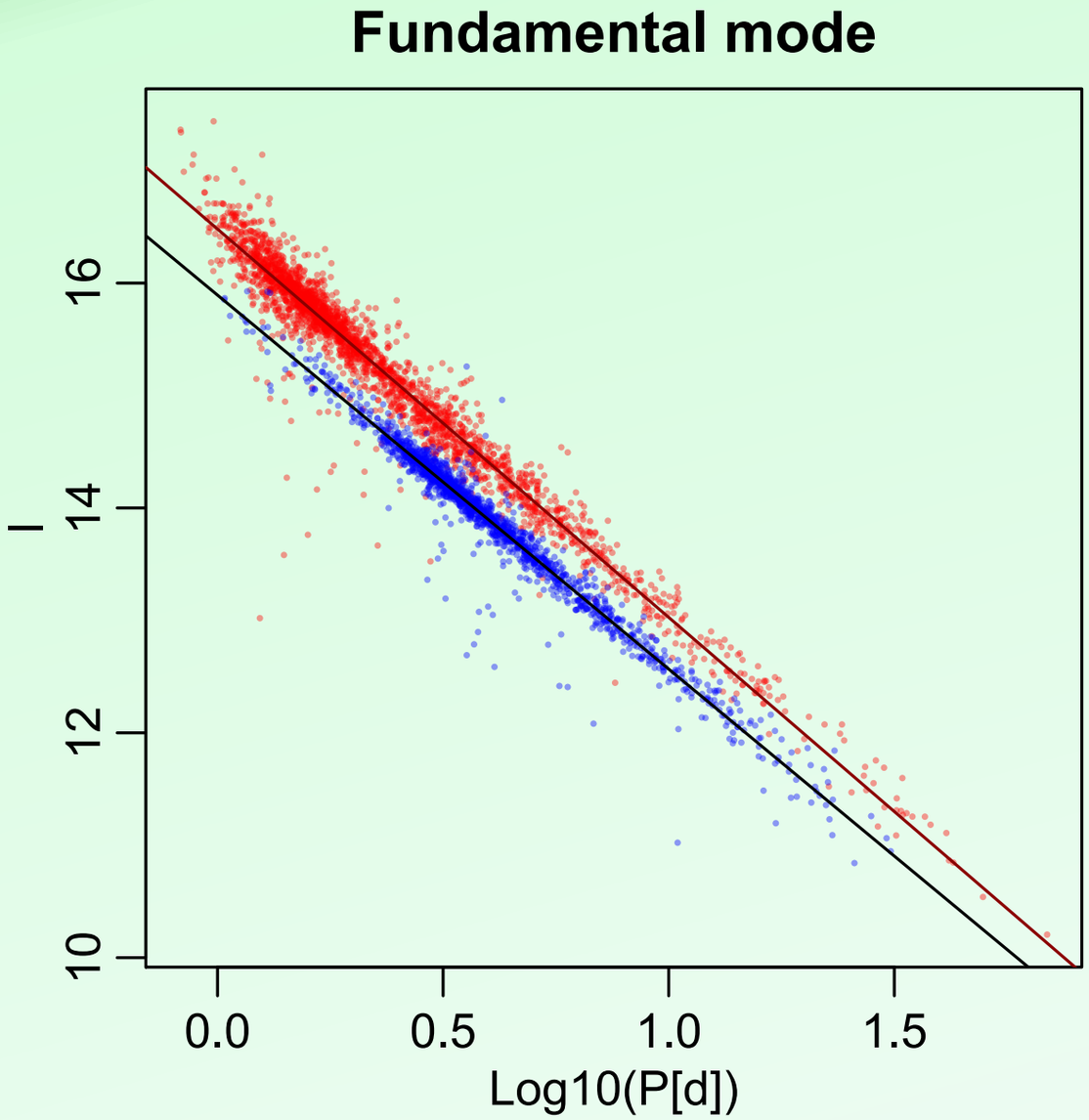
OLS Regression Results

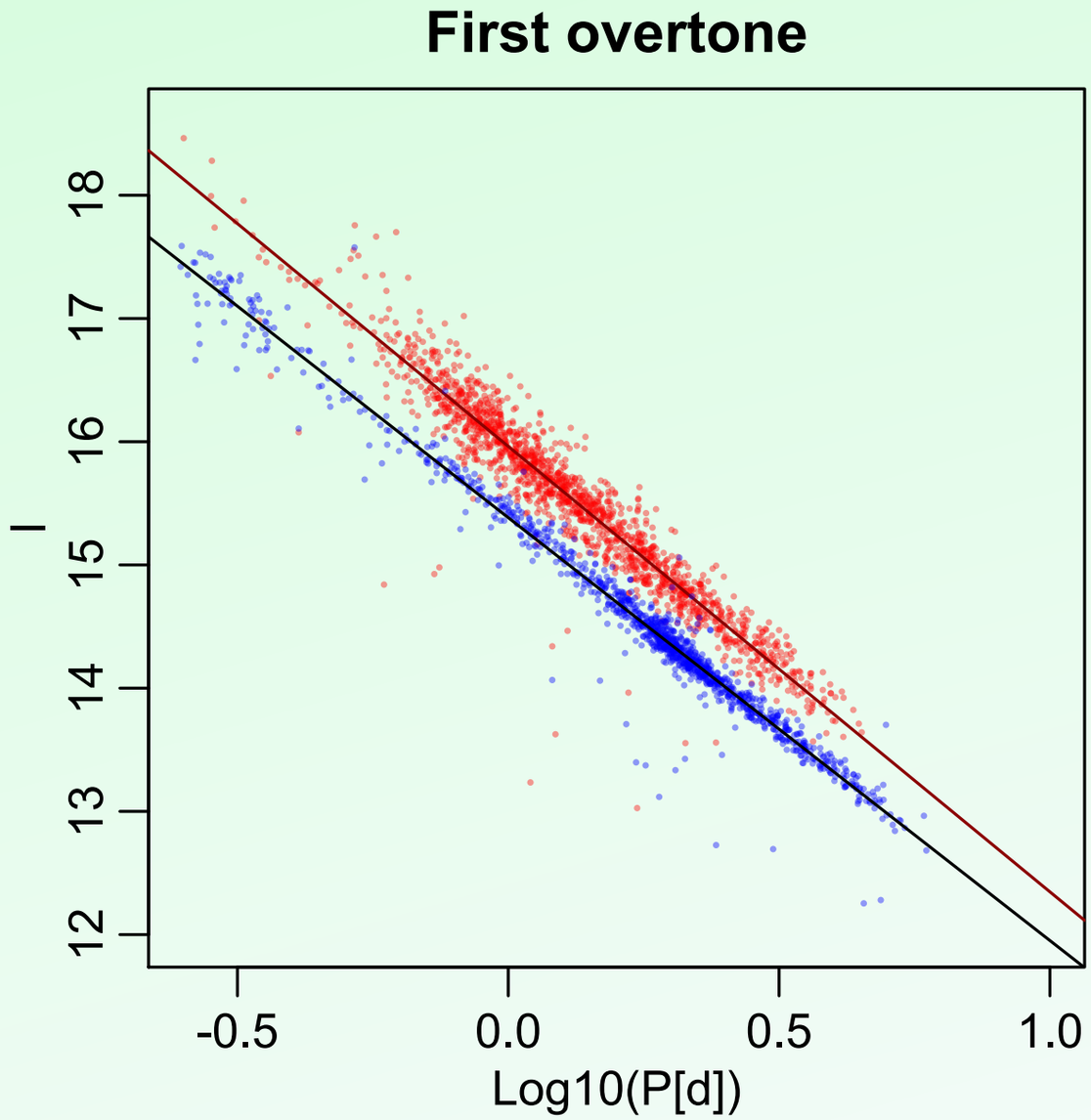
Dep. Variable:	W	R-squared:	0.918
Model:	OLS	Adj. R-squared:	0.918
Method:	Least Squares	F-statistic:	1.835e+04
Date:	Mon, 01 Aug 2016	Prob (F-statistic):	0.00
Time:	16:01:03	Log-Likelihood:	108.98
No. Observations:	1632	AIC:	-214.0
Df Residuals:	1630	BIC:	-203.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	15.9624	0.007	2428.764	0.000	15.949	15.975
logP1	-3.6155	0.027	-135.469	0.000	-3.668	-3.563

Omnibus:	1179.983	Durbin-Watson:	1.781
Prob(Omnibus):	0.000	Jarque-Bera (JB):	49075.172
Skew:	-2.890	Prob(JB):	0.00
Kurtosis:	29.235	Cond. No.	4.84

Data and the fit



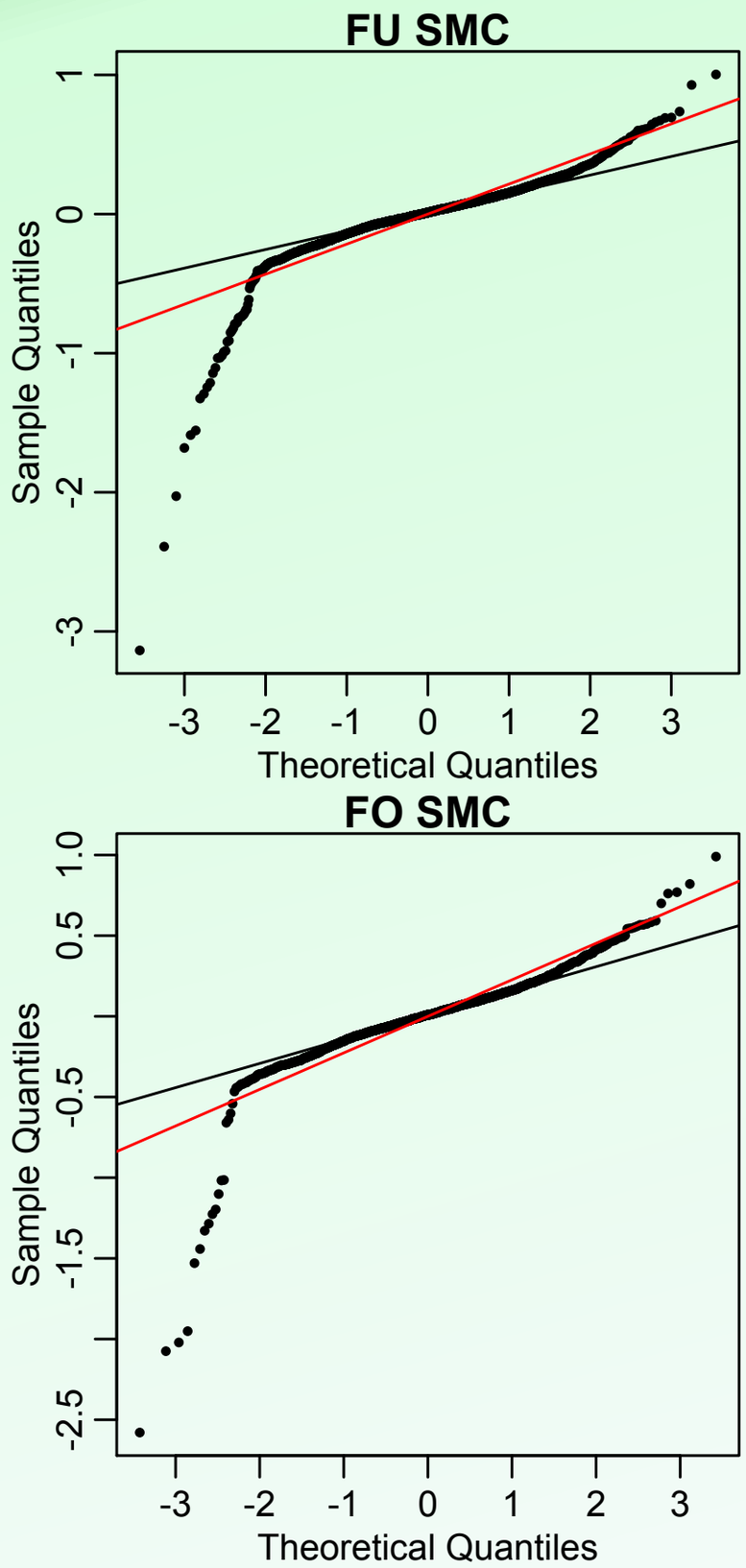


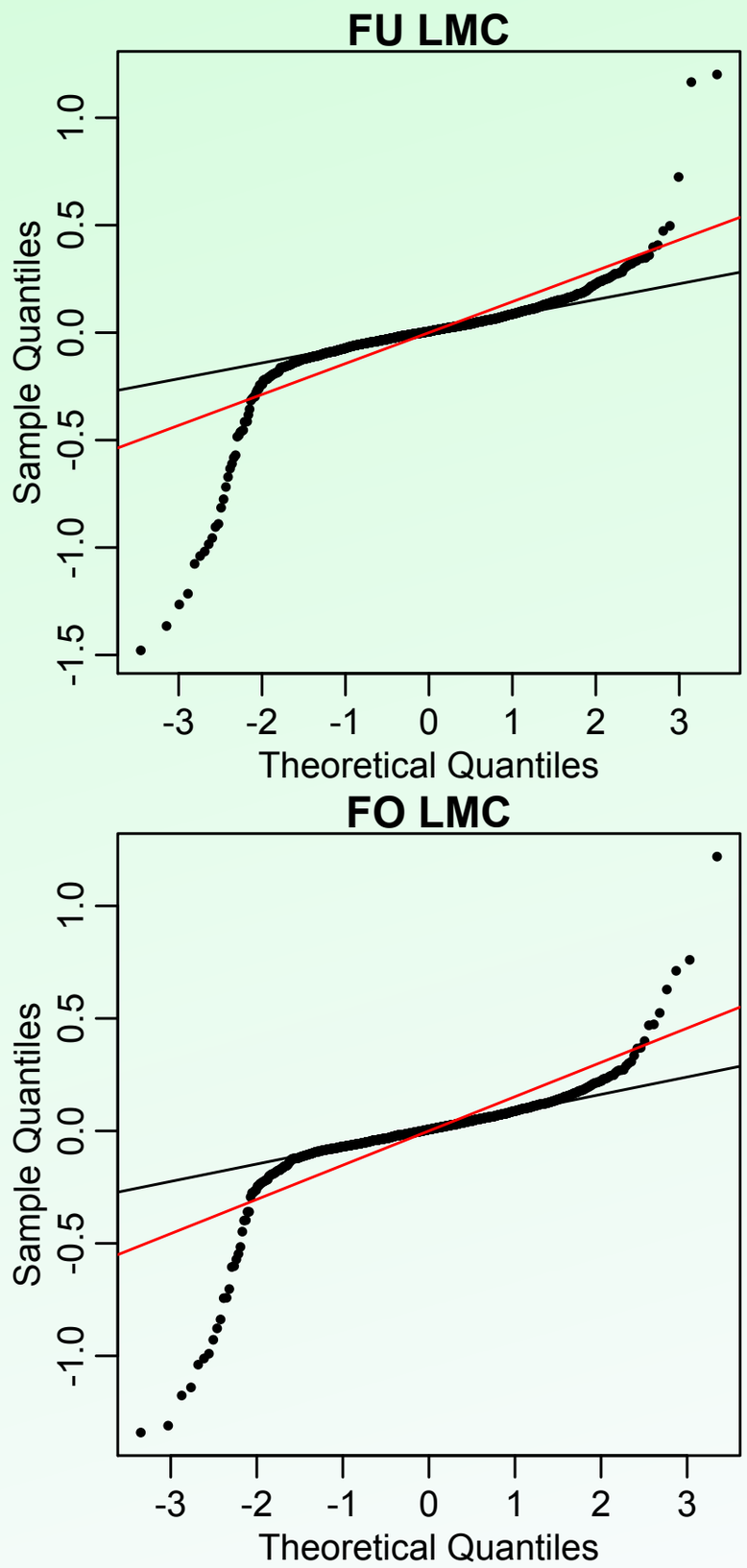
OGLE-III results for comparison:

LMC	coef	se	SMC	coef	se
a =	15.875	0.007	b =	16.387	0.016
b =	-3.309	0.011	a =	-3.310	0.020

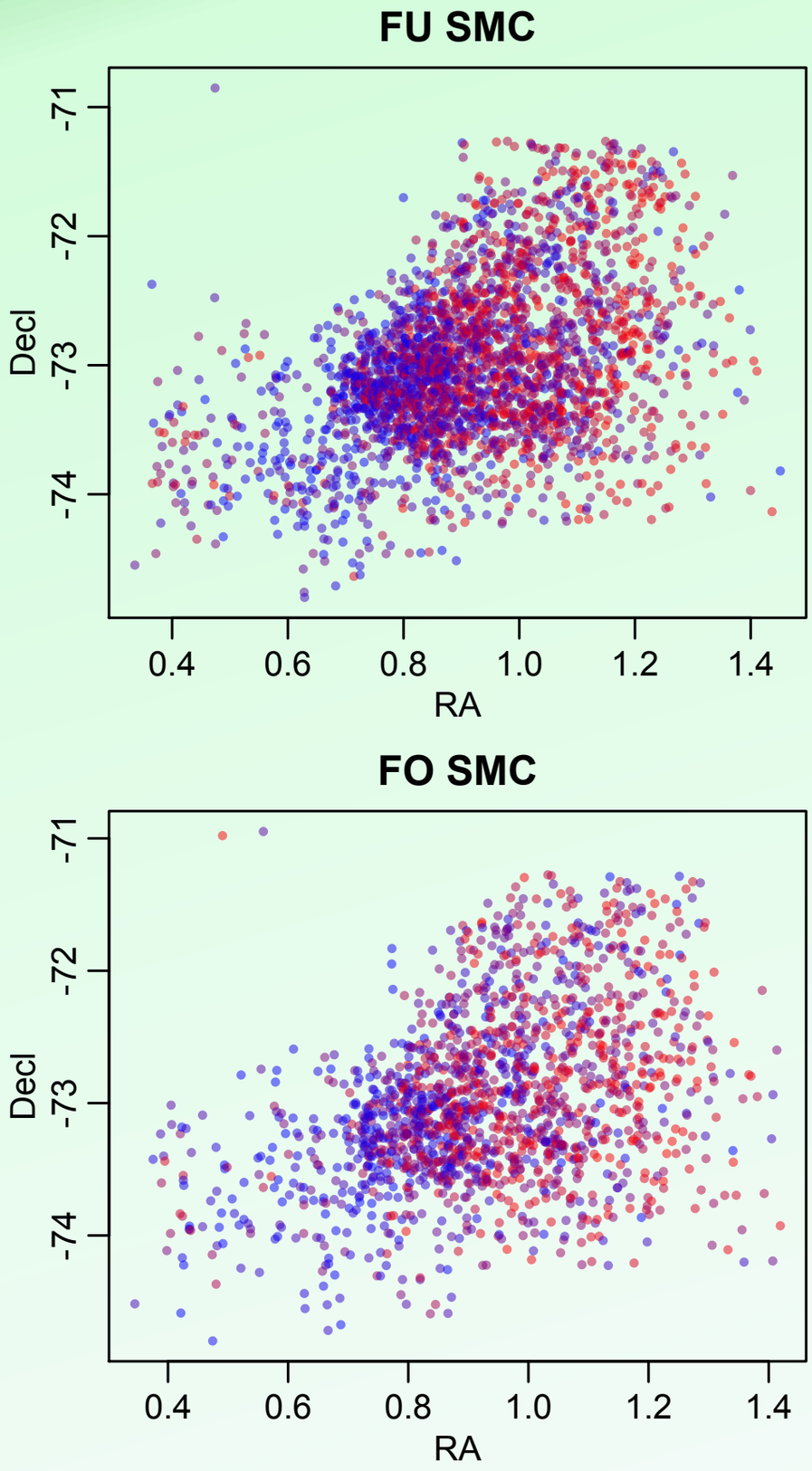
LMC	coef	se	SMC	coef	se
b =	15.375	0.006	b =	15.981	0.006
a =	-3.413	0.017	a =	-3.567	0.023

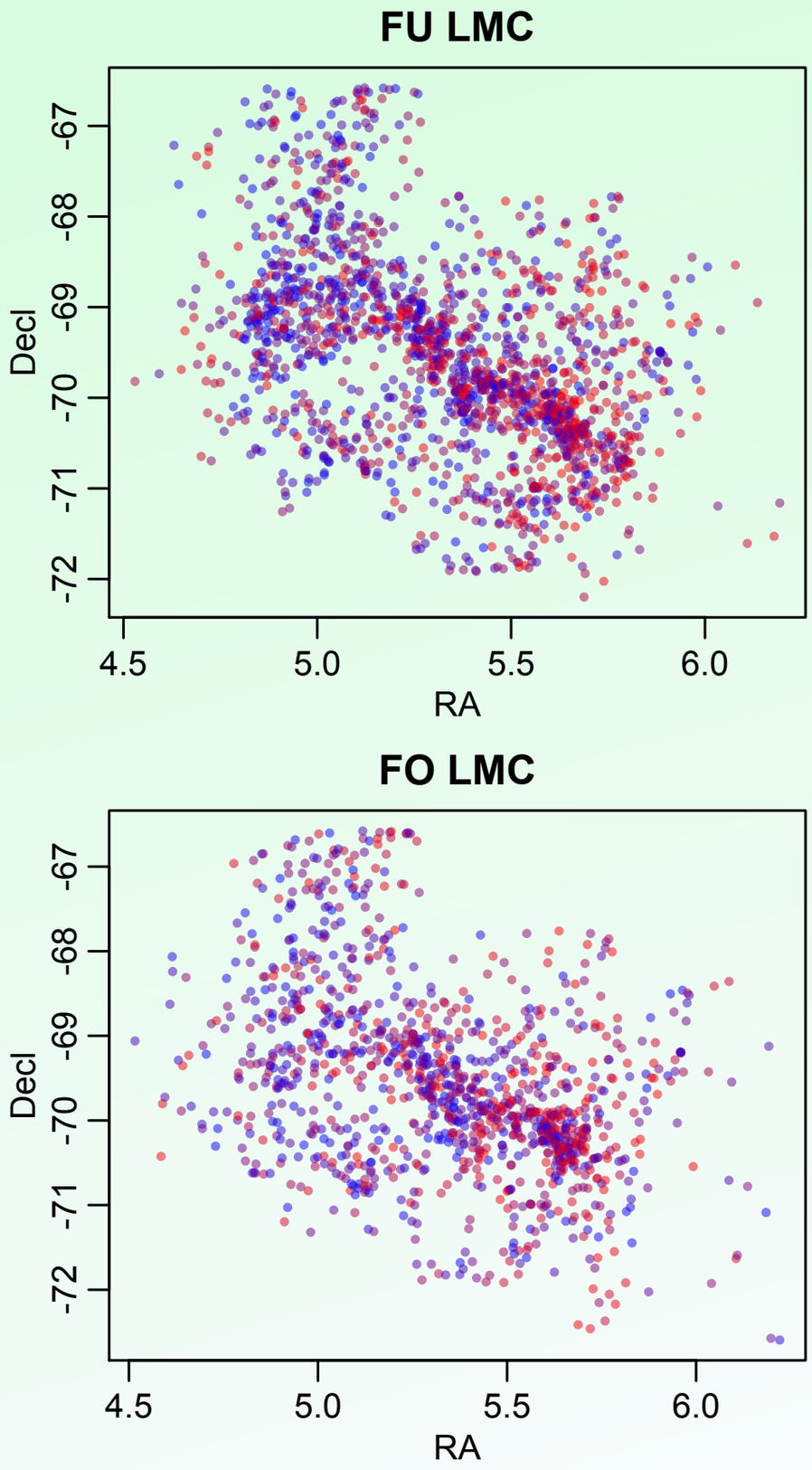
QQ-plot of residuals



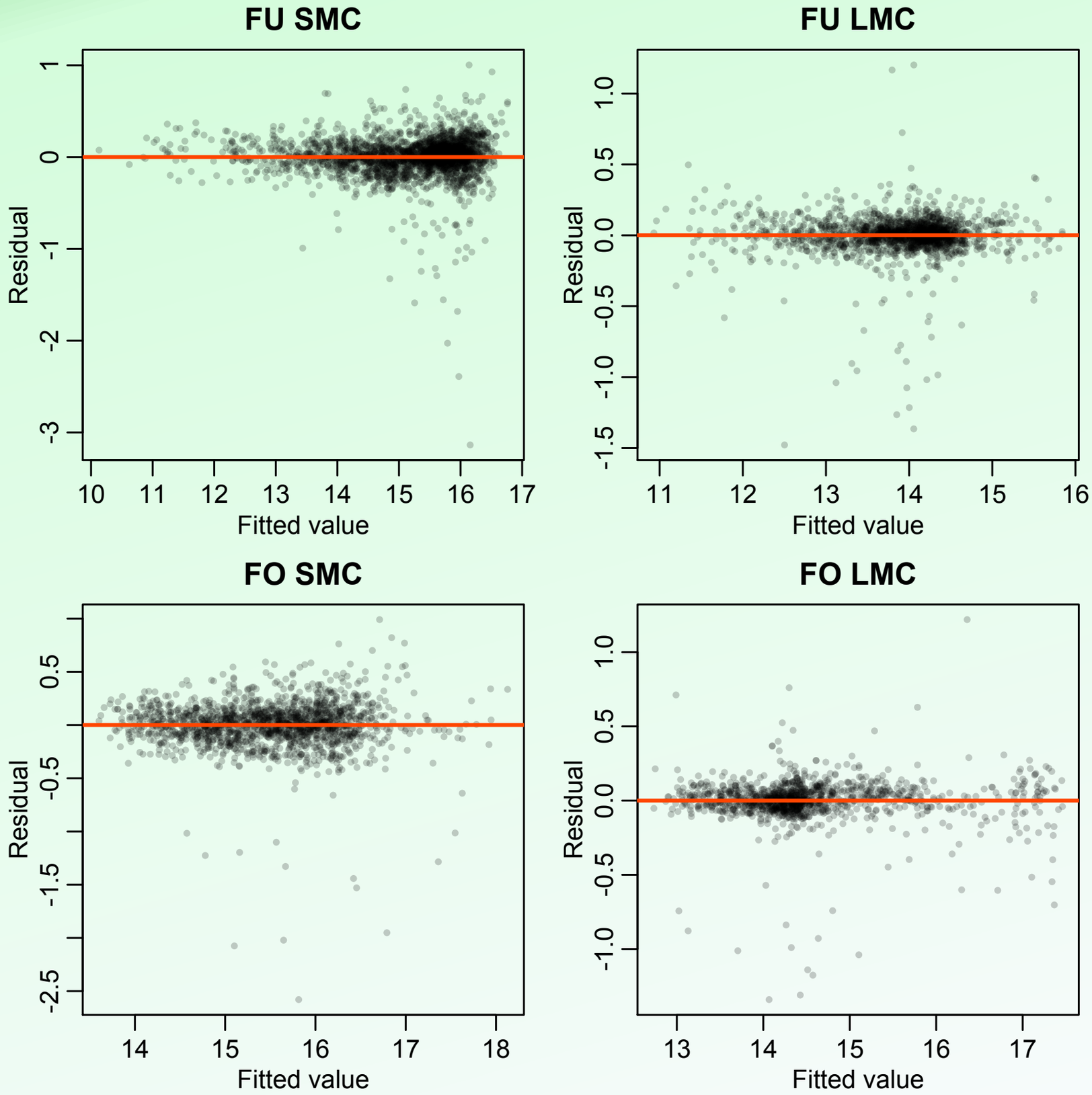


Skymap of the residuals

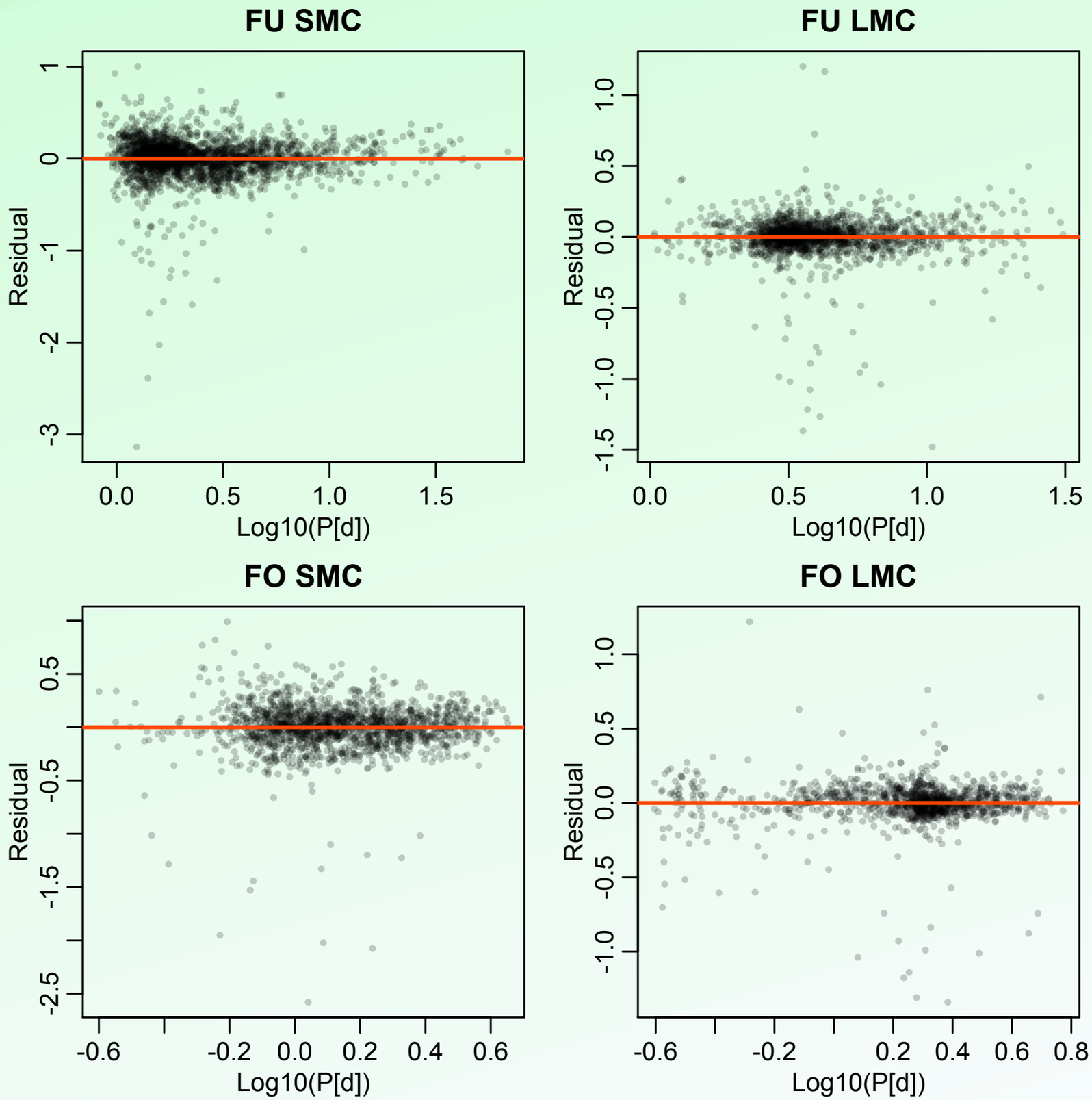




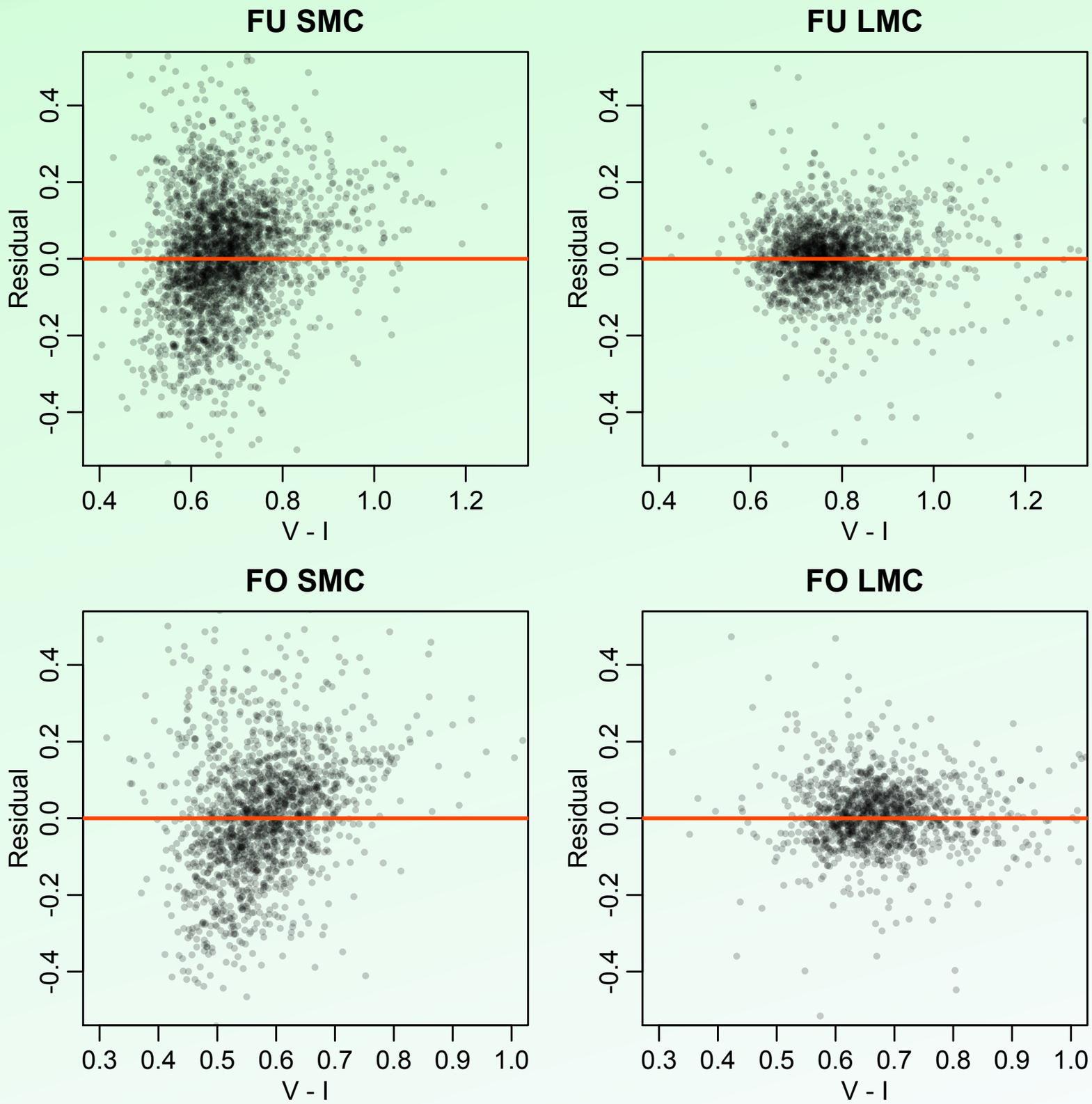
Residuals versus fitted value



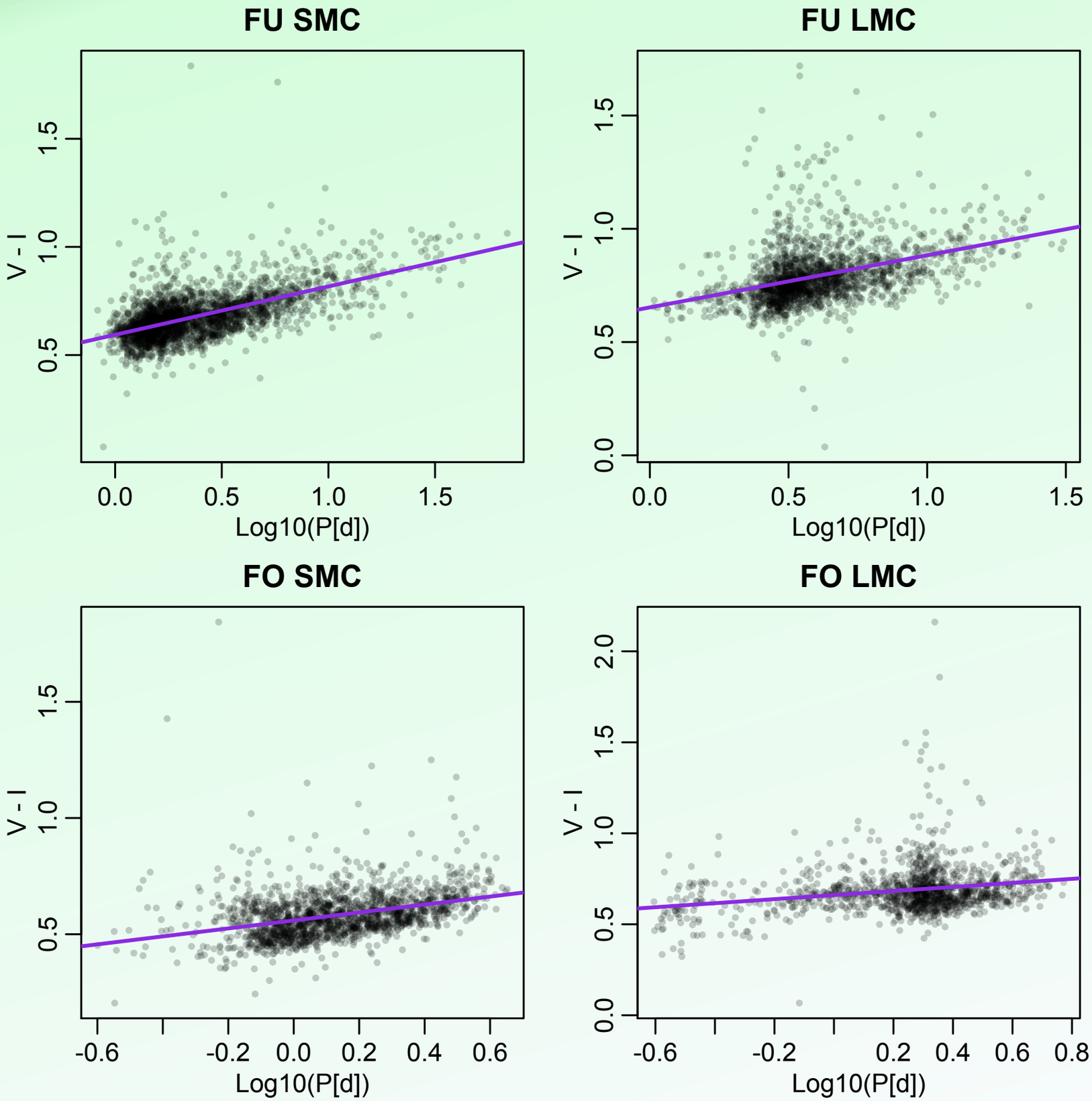
Residuals versus the covariate



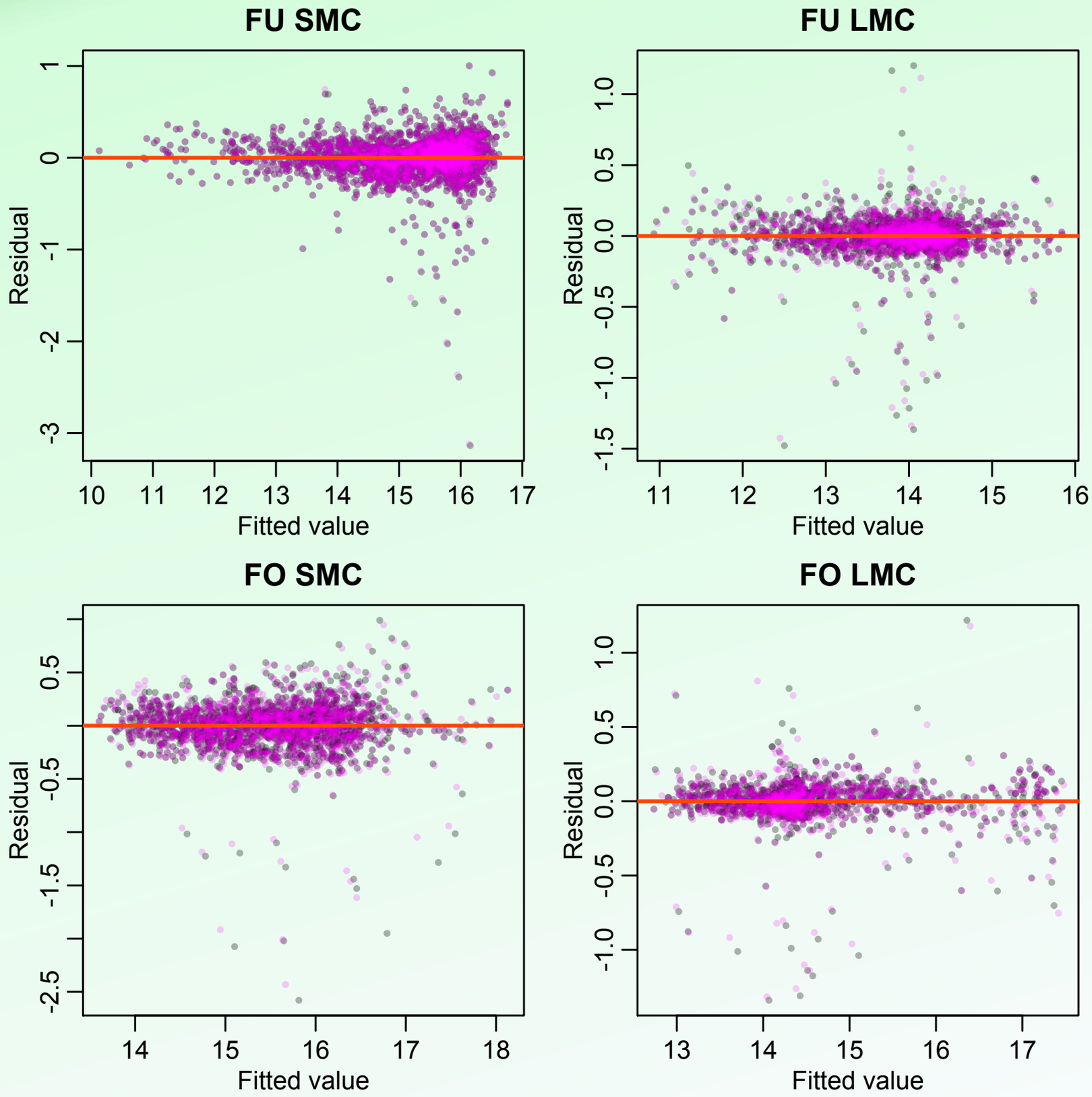
Residuals versus a new variable: V - I



Relationship between log(P) and V - I



Residuals in W ~ (log(P), r_{V-I})



Collinear vs. orthogonal variables: model comparison

Orthogonalized model

Collinear model

Model parameters

	Estimate	Std. Error	t-value	pval
const	15.3840	0.0058	2651.9087	0
logP1	-3.4322	0.0160	-214.9399	0
vi.resid0	-0.1954	0.0329	-5.9410	0

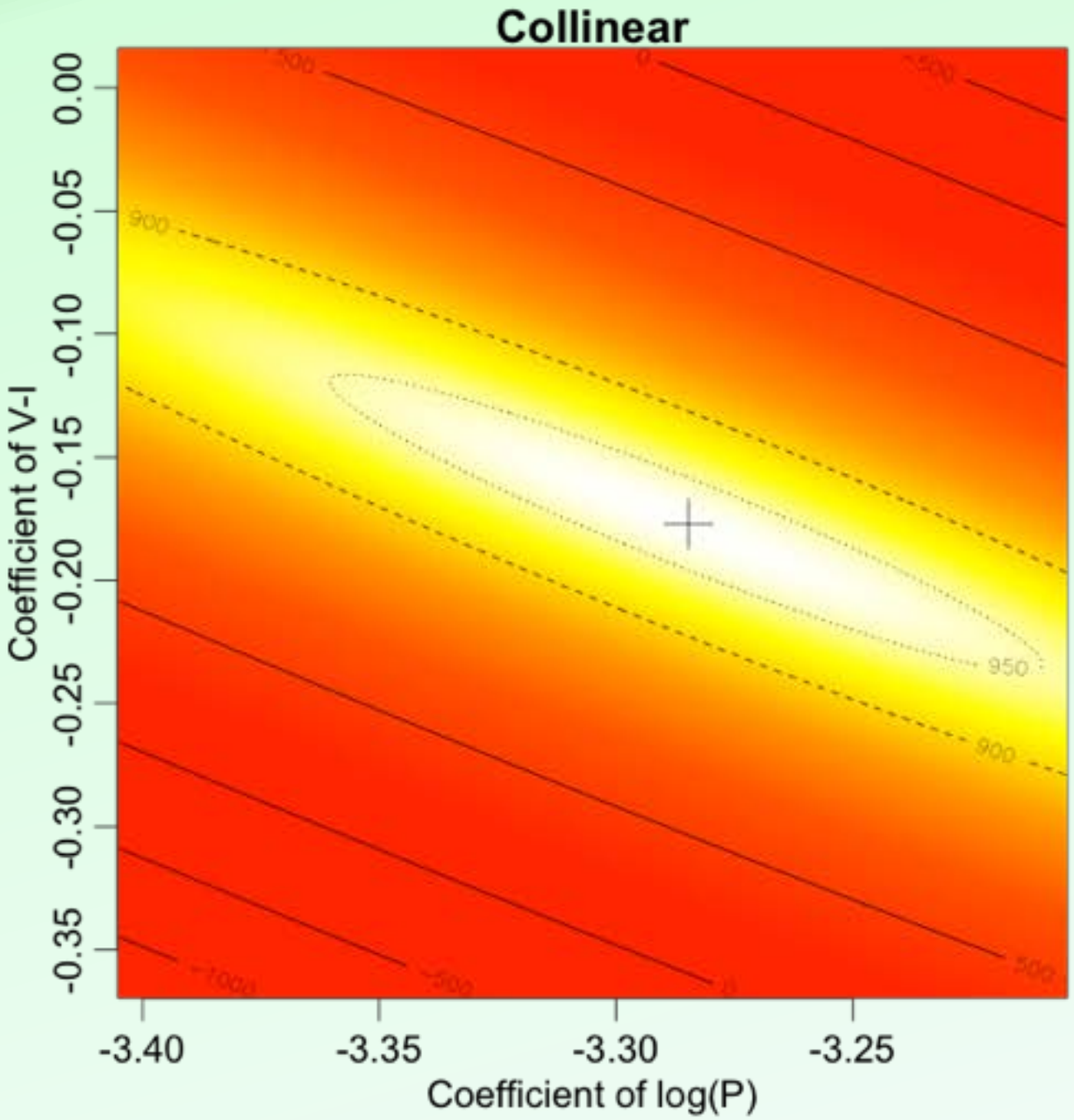
	Estimate	Std. Error	t-value	pval
const	16.0083	0.0206	778.7185	0
logP1	-3.2852	0.0168	-195.3723	0
VI	-0.1768	0.0275	-6.4323	0

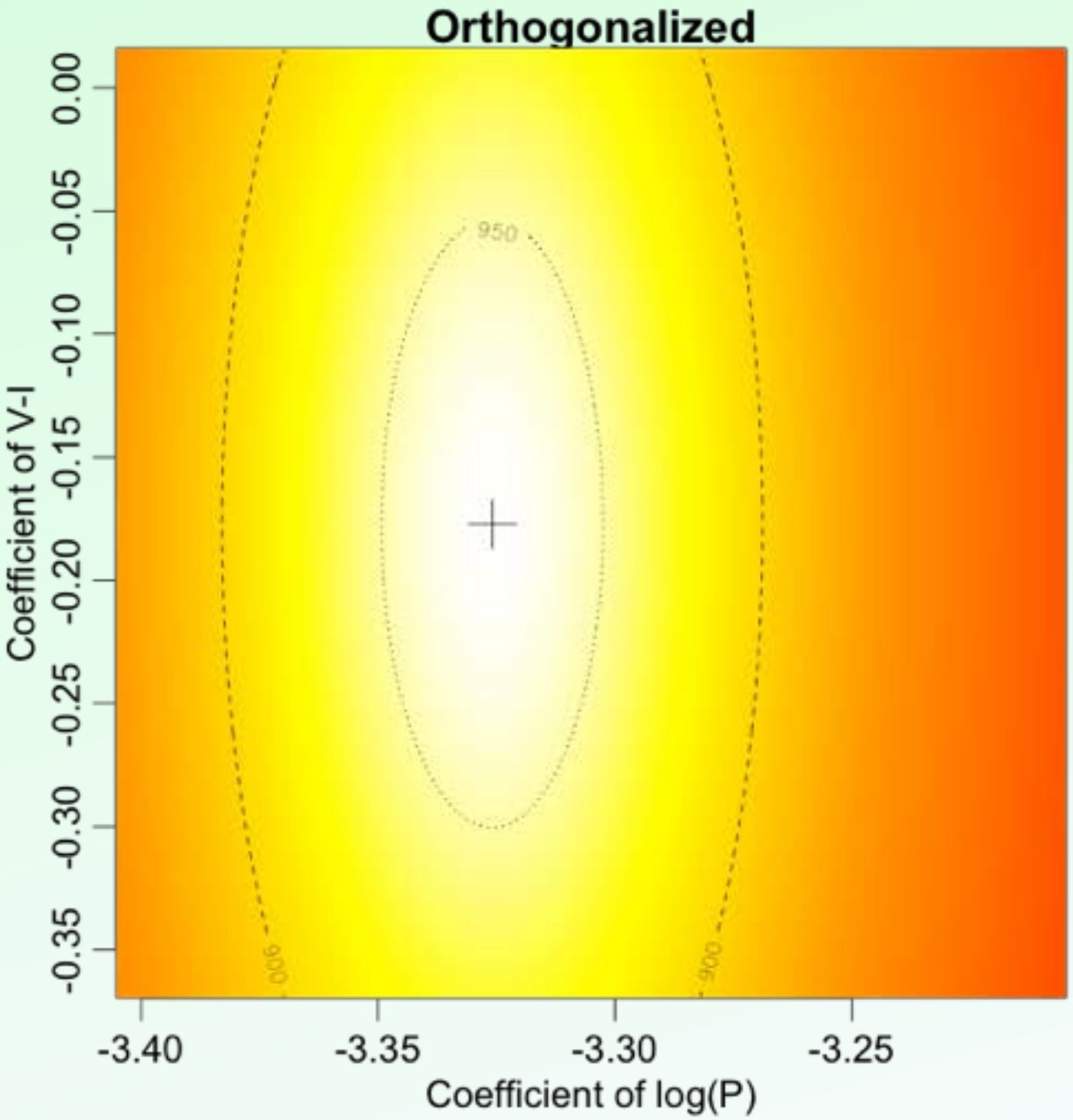
Correlation matrix of the parameters

	const	logP1	vi.resid0
const	1.0000	-0.6743	0
logP1	-0.6743	1.0000	0
vi.resid0	0.0000	0.0000	1

	const	logP1	VI
const	1.0000	-0.0982	-0.8728
logP1	-0.0982	1.0000	-0.3759
VI	-0.8728	-0.3759	1.0000

Collinear vs. orthogonal variables: likelihood surfaces





Take-home message

- **The probabilistic structure** of your problem is **just as important** as the effect or the physical law you want to detect/check

- **Look at your data** as many ways as possible: do extensive exploratory analysis, as many plots as you can imagine

- **Build up your likelihood very carefully**, aware of any problems your data might present: the likelihood will be the center and the machinery to obtain your results

Thanks!