Statistical concepts in Earth data modeling

Joachim Vogt

Fall 2022

Modeling of Earth System Data

JACOBS
UNIVERSITY

Statistical concepts in Earth data modeling — Overview

1. Random variables and model distributions
   - Continuous random variables, CDF, PDF, quantiles
   - Expectation, moments, mean, variance, correlation
2. Empirical distributions and sample statistics
   - Histograms, kernel density estimators, sample statistics
   - Bivariate data: sample correlation, graphical representation
3. Data modeling and residual distributions
   - Motivation, terminology, parameter estimation strategies
   - Distribution of residuals, graphical normality tests
4. Confidence intervals and standard errors
   - Significance level, confidence intervals for the mean
   - Error propagation concepts and formulas
5. Hypothesis testing and binary classification
   - Binary classification, statistical hypothesis testing, error types
6. Bootstrap approach to error estimation
   - Monte Carlo error propagation, bootstrap principles, applications
7. Project: Statistical concepts

Random variables and model distributions

Cumulative distribution function and density

Measurements and their errors are modeled by *random processes*, and their outcomes by (continuous) random variables.

For *continuous random variables*, probabilities do not refer to single values but to ranges of values. The probability to find a continuous random variable U in the range $[a, b]$ is denoted as $\mathrm{P}(a \leq U \leq b)$.

Cumulative distribution function (CDF): $\Phi(u) \; = \; \mathrm{P}(U \leq u)$.

If $\Phi(u)$ is a smooth (differentiable) function, then the derivative

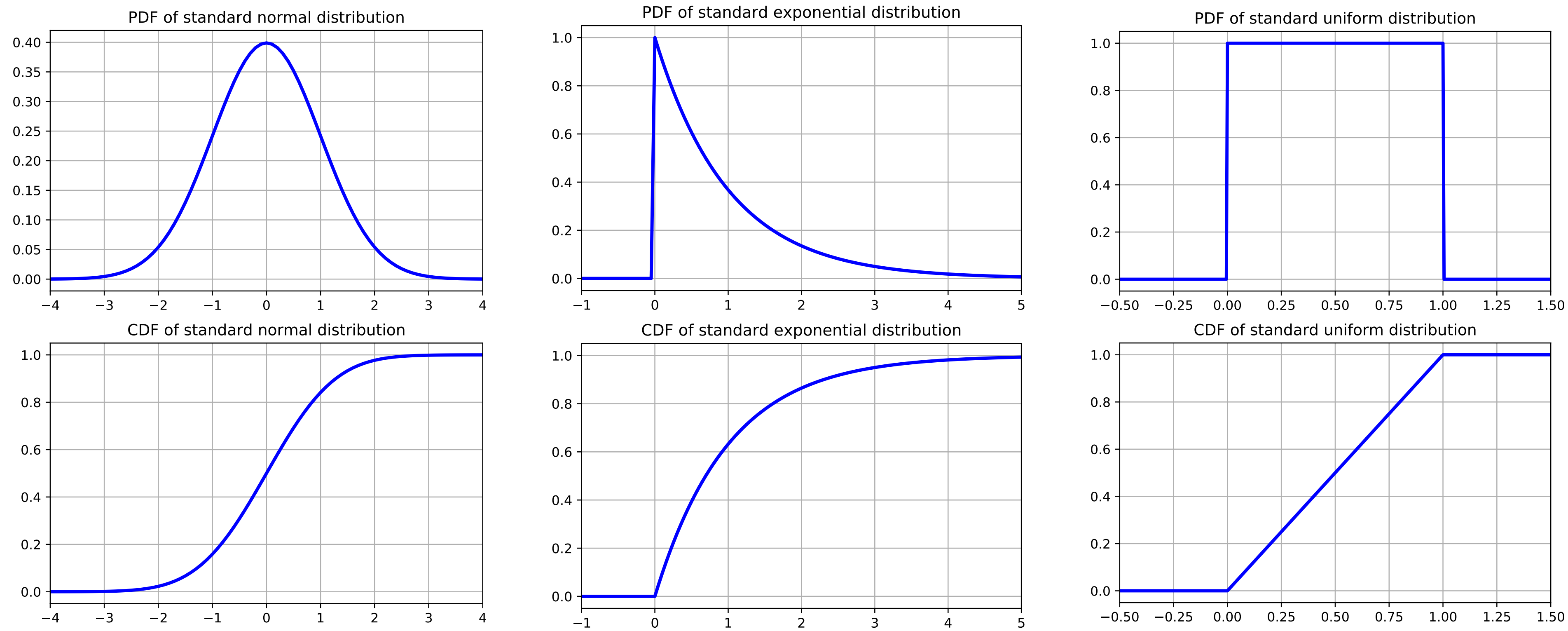
$$p(u) \; = \; \frac{\mathrm{d}\Phi}{\mathrm{d}u} \; = \; \Phi'(u)$$

is the *probability density function (PDF)*.

The PDF is also called density, distribution function, or simply distribution.

- PDF is *non-negative*: $p(u) \geq 0$.
- $\mathrm{P}(a \leq U \leq b) \; = \; \Phi(b) - \Phi(a) \; = \; \int_a^b p(u) \, \mathrm{d}u$.
- *Normalization*: $\int_{-\infty}^{\infty} p(u) \, \mathrm{d}u \; = \; 1$.

Important examples



Gaussian distribution = normal distribution:

$$p(u) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(u-\mu)^2/2\sigma^2} \;, \quad \Phi(u) \;=\; \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{u-\mu}{\sqrt{2\sigma^2}}\right)\right] \;.$$

Exponential distribution: $p(u) \;=\; \lambda\,\mathrm{e}^{-\lambda u} \;, \quad \Phi(u) \;=\; 1 \,-\, \mathrm{e}^{-\lambda u} \;.$

Uniform distribution: $p(u)$ is constant and $\Phi(u)$ is linear for $u \in [a, b]$.

Quantiles

Quantiles divide the domain of a distribution into intervals of equal probability. If L is the number of intervals and $\Phi = \Phi(u)$ the CDF, the quantiles $u_{L;k}$ are given through $\Phi(u_{L;k}) = \frac{k}{L}$, $k = 1, 2, \ldots, L-1$.

- $L = 2$: *median*, $\Phi(\text{median}) = \frac{1}{2}$. The areas under the PDF left and right of the median are equal.
- $L = 4$: *quartiles*, $\Phi(Q_1) = \frac{1}{4}$ and $\Phi(Q_3) = \frac{3}{4}$. The difference is called *interquartile range*: $\text{iqr} = Q_3 - Q_1$.
- $L = 10$: *deciles*.
- $L = 100$: *percentiles*, e.g., $\Phi(u) = 0.9$ defines the 90th percentile.

Quantile function or percent-point function (PPF): inverse CDF.

- $\Phi^{-1}\left(\frac{1}{4}\right) = Q_1$.
- $\Phi^{-1}\left(\frac{1}{2}\right) = Q_2 = \text{median}$.
- $\Phi^{-1}\left(\frac{3}{4}\right) = Q_3$.
- $\Phi^{-1}\left(\frac{k}{L}\right) = u_{L;k}$.

Expectation and moments

For a continuous random variable U, *expectation* E is defined as

$$\mathrm{E}\{f(U)\} \;=\; \int_{-\infty}^{\infty} f(u)\, p(u)\, \mathrm{d}u \;.$$

Moments of a random variable (or distribution): $\mathrm{E}\{U^n\},\; n = 1, 2, \ldots$

- *Mean*: $\mathrm{E}\{U\} = \bar{U}$.
- *Centered moments*: $\mathrm{E}\{(U - \bar{U})^n\},\; n = 2, 3, \ldots$
- *Variance (second centered moment)*: $\mathrm{E}\{(U - \bar{U})^2\} = (\Delta U)^2 = \sigma^2$.
- *Skewness (normalized centered third moment)*: $\mathrm{E}\left\{\left(\frac{U-\bar{U}}{\Delta U}\right)^3\right\}$.

  Skewness measures the symmetry of the distribution relative to the mean value. A symmetric distribution has zero skewness.
- *Kurtosis (normalized centered fourth moment)*: $\mathrm{E}\left\{\left(\frac{U-\bar{U}}{\Delta U}\right)^4\right\} - 3$.

  Kurtosis measures the flatness/peakedness of the distribution as compared to a Gaussian distribution (zero kurtosis).

Bivariate random processes: joint CDF and joint PDF

Consider two continuous random variables U and V, then

$$\mathrm{P}(a \le U \le b, c \le V \le d)$$

is the probability to find U in $[a, b]$ and (simultaneously) V in $[c, d]$.

- *Joint cumulative distribution function (joint CDF)*: $\Phi = \Phi(u, v)$.
- *Joint probability density function (joint PDF)*: $p = p(u, v)$.
- Relationship between $\Phi(u, v)$ and $p(u, v)$:

$$\Phi(u, v) \;=\; \mathrm{P}(U \le u, V \le v) \;=\; \int_{-\infty}^{u} \int_{-\infty}^{v} p(\tilde{u}, \tilde{v}) \, \mathrm{d}\tilde{u} \, \mathrm{d}\tilde{v} \; .$$

- *Independence*: $\Phi(u, v) = \Phi_u(u) \cdot \Phi_v(v)$, $p(u, v) = p_u(u) \cdot p_v(v)$.
- *Marginal densities*: $p_u(u) \;=\; \int p(u, v) \, \mathrm{d}v$ and $p_v(v) \;=\; \int p(u, v) \, \mathrm{d}u$.
- *Expectation*: $\mathrm{E}\{f(U, V)\} \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u, v) \, p(u, v) \, \mathrm{d}u \, \mathrm{d}v$.
- *Covariance*: $\mathrm{cov}(U, V) \;=\; \mathrm{E}\{(U - \bar{U})(V - \bar{V})\}$.

Association and correlation

In a statistical context, the term *association* refers to a general kind of relationship between two variables that gives rise to statistical dependence.

Correlation is sometimes used in the same general sense but usually refers more strictly to *linear relationships*.

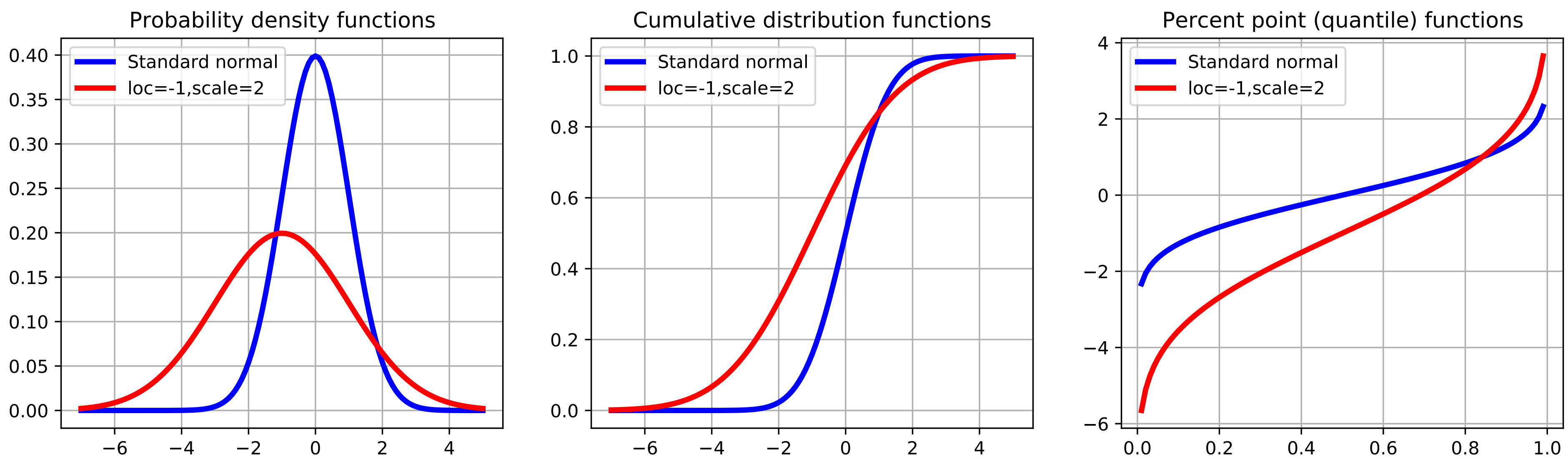
- *Pearson's correlation coefficient* $r = \frac{\text{cov}(U,V)}{\Delta U \cdot \Delta V}$ is the most popular measure of (linear) correlation.
- The correlation coefficient of independent random variables is zero.
- The square of r is called *coefficient of determination* r^2.

Further correlation/association measures

- *Spearman's rank correlation coefficient* is given by $\rho = \frac{\text{cov}(Ru,Rv)}{\Delta Ru \cdot \Delta Rv}$. Here Ru and Rv are arrays formed from the magnitude-based ranks.
- The rank correlation coefficient ρ is less affected by outliers, and yields high values also for monotonic relationships that are nonlinear.
- Association measures based on information theory: mutual information, Kullback-Leibner divergence, information entropy.

Introduction to the SciPy module `stats`

- `rv_continuous`: class for continuous random variables, e.g., normal (Gaussian), uniform, Pareto, t, chi-squared, Cauchy, ...
- Class methods:
  - moments, PDF, CDF, PPF, generic transformations (location and scale), distribution fitting, generation of random numbers (variates);
  - demonstrated for the uniform distribution;
  - to be applied to the normal distribution (exercise).



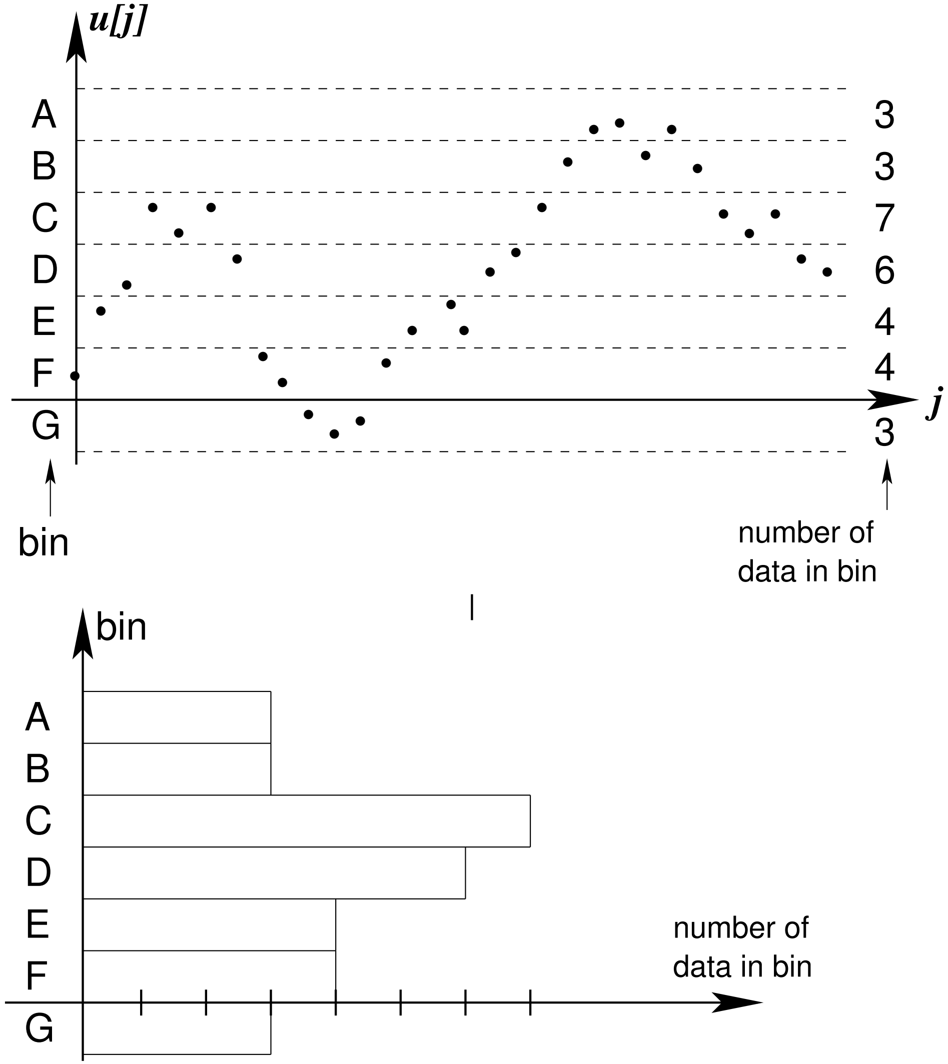
Empirical distributions and sample statistics

Empirical distributions, density estimation, histograms

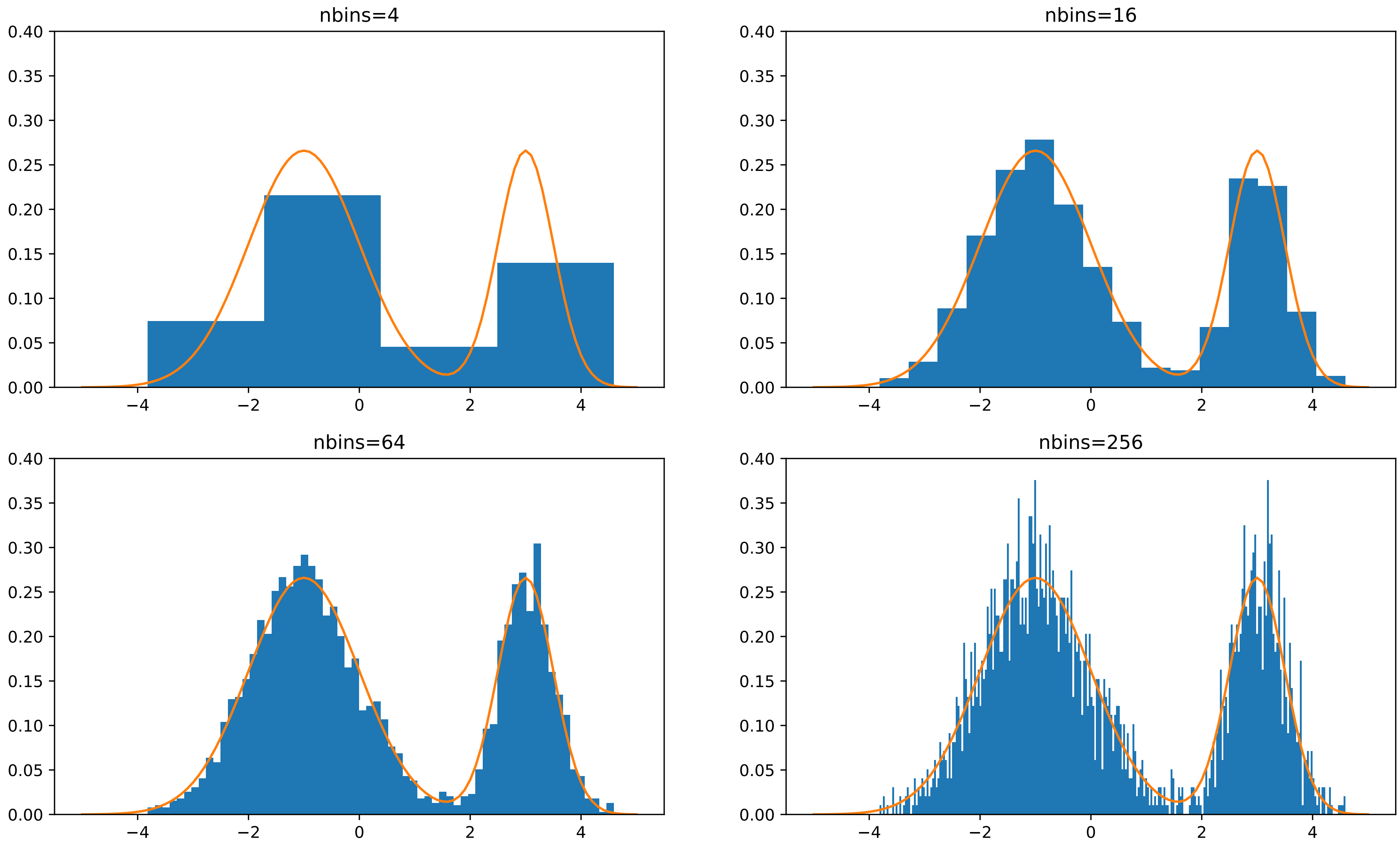
To connect probability theory with observations, a data set (sample) is understood as a *realization* of the underlying random process.

Inference of distribution parameters or more general aspects (*statistics*) of the underlying process from the observations is called *estimation*.

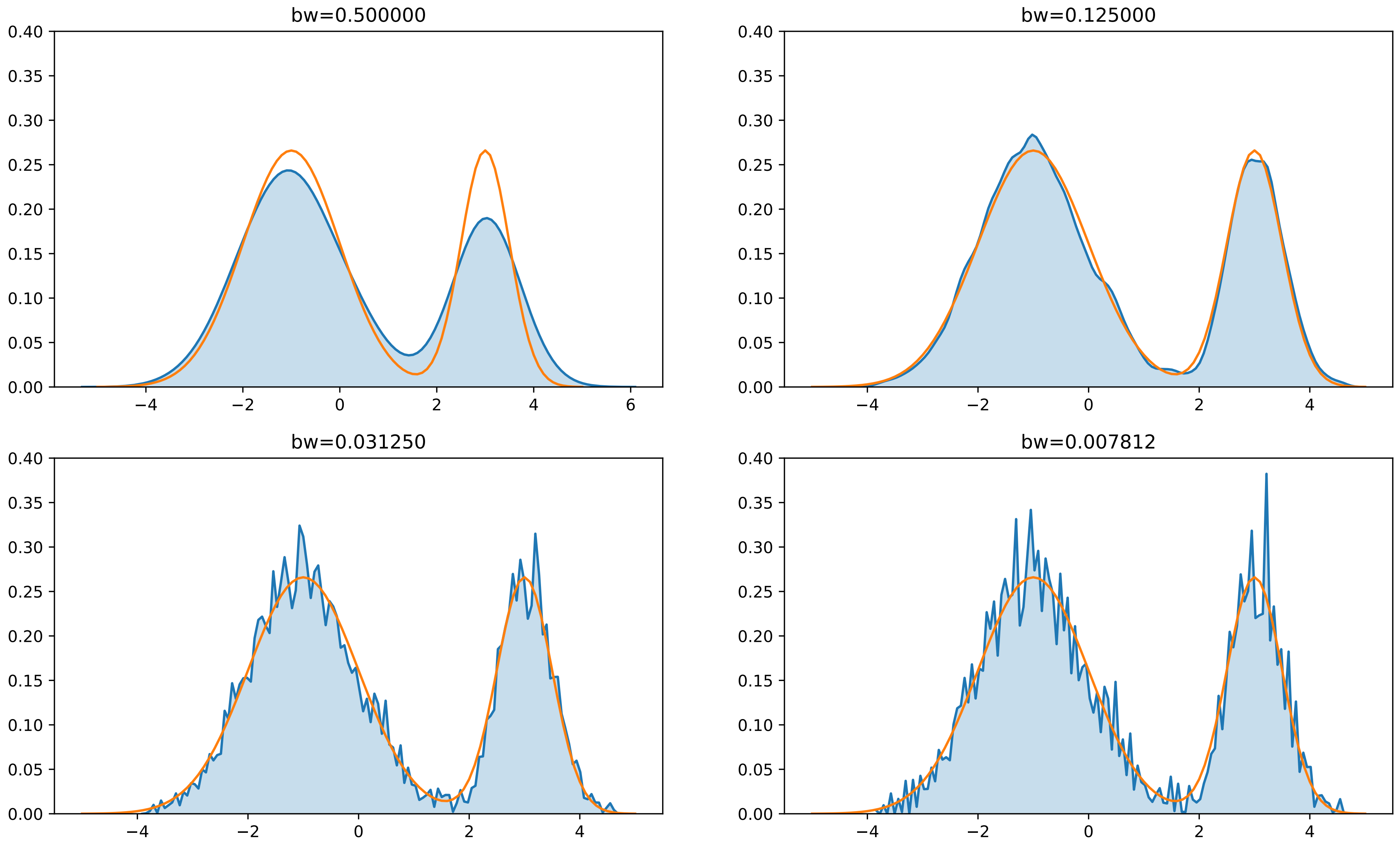
Density estimation: construction of an estimator for the PDF (also called empirical distribution). A simple and popular density estimator is the *histogram*. More refined are *kernel density estimators* which acculumate contributions from predefined functions (kernels) around the data.



Histogram dependence on bin width



Kernel density estimation



Sample statistics

Consider a series of measurements (of equal quality) constituting a *data set* $u = \{u_1, u_2, \ldots, u_N\}$. To obtain estimates of expectation values, denoted by $\hat{\mathrm{E}}\{\cdots\}$ or $\langle\cdots\rangle$, carry out arithmetic *averaging* as defined through

$$\langle f(u) \rangle \;=\; \frac{1}{N} \sum_j f(u_j) \;.$$

Such estimators are called *sample statistics*.

Sample moments (empirical moments)

- *Mean*: $\langle u \rangle \;=\; \bar{u} \;=\; u_{\text{mean}}$.
- *Variance*: $\left\langle (u - \bar{u})^2 \right\rangle \;=\; (\Delta u)^2$.
- *Moments*: $\langle u^n \rangle\,,\; n = 1, 2, \ldots$
- *Centered moments*: $\langle (u - \bar{u})^n \rangle\,,\; n = 2, 3, \ldots$

Sample (empirical) quantiles: sort the observations into ascending order and determine the values that separate the sorted data set into L sections.

Bivariate data: sample statistics, covariance, correlation

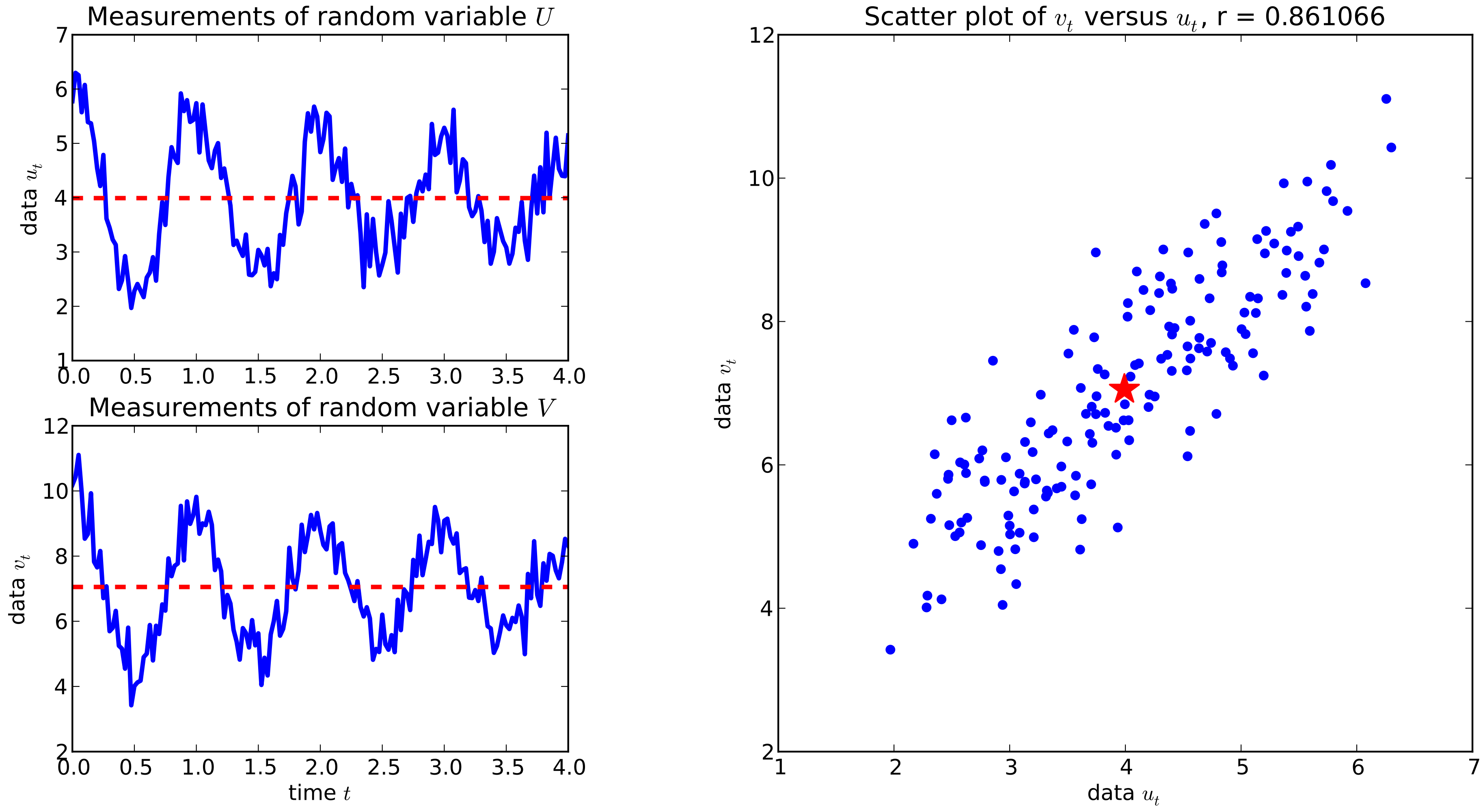
Two data sets u and v, sample statistics: $\langle f(u,v)\rangle \;=\; \frac{1}{N}\sum_j f(u_j,v_j)$.

Covariance: $\mathrm{cov}(u,v) \;=\; \langle (u-\bar{u})(v-\bar{v})\rangle \;=\; \frac{1}{N}\sum_j (u_j-\bar{u})(v_j-\bar{v})$.

Pearson's linear correlation coefficient: $r \;=\; \dfrac{\mathrm{cov}(u,v)}{\Delta u \cdot \Delta v} \in [-1,1]$.

- Values close to $\left\{\begin{array}{r} 1 \\ 0 \\ -1 \end{array}\right\}$ suggest that u and v are $\left\{\begin{array}{c} \text{correlated} \\ \text{uncorrelated} \\ \text{anti-correlated} \end{array}\right\}$.
- r measures the *goodness-of-fit to a linear model*.
- Relative variance captured by the linear model is given by r^2.
- Large linear correlation ($|r| \approx 1$) does not imply a causal relationship between the variables.
- Zero linear correlation does not imply statistical independence.
- *Outliers* in data sets can affect r significantly.

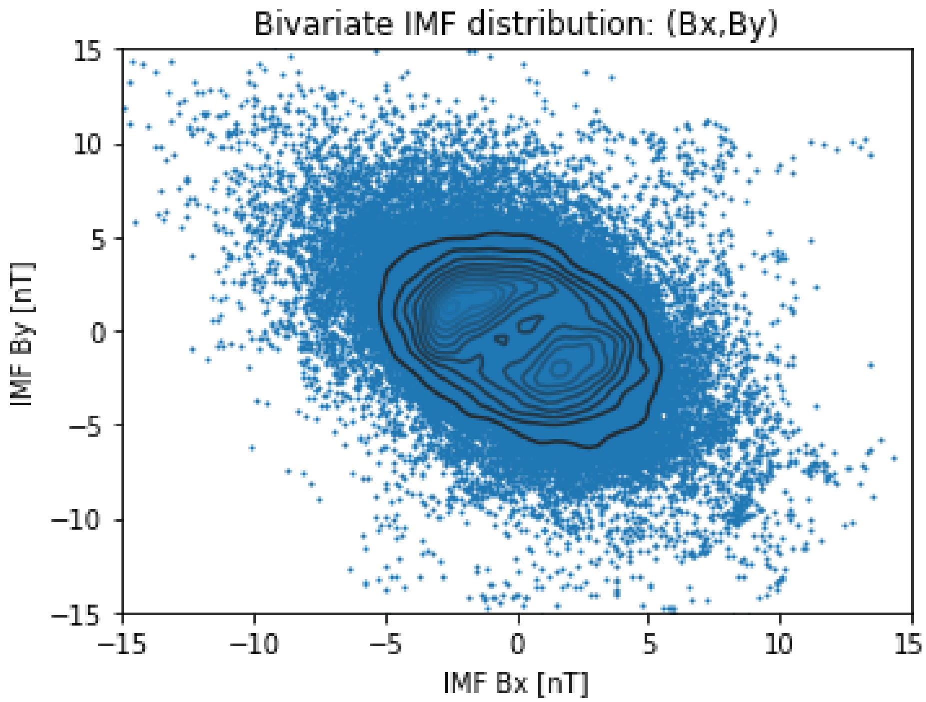
Example: correlated time series

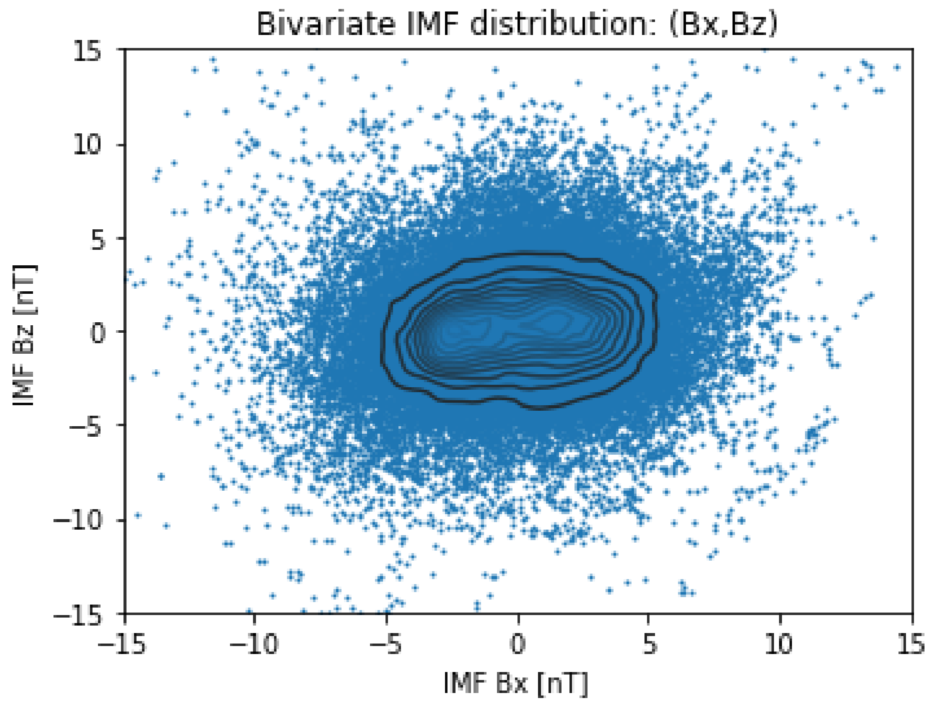


Numerical Software Lab

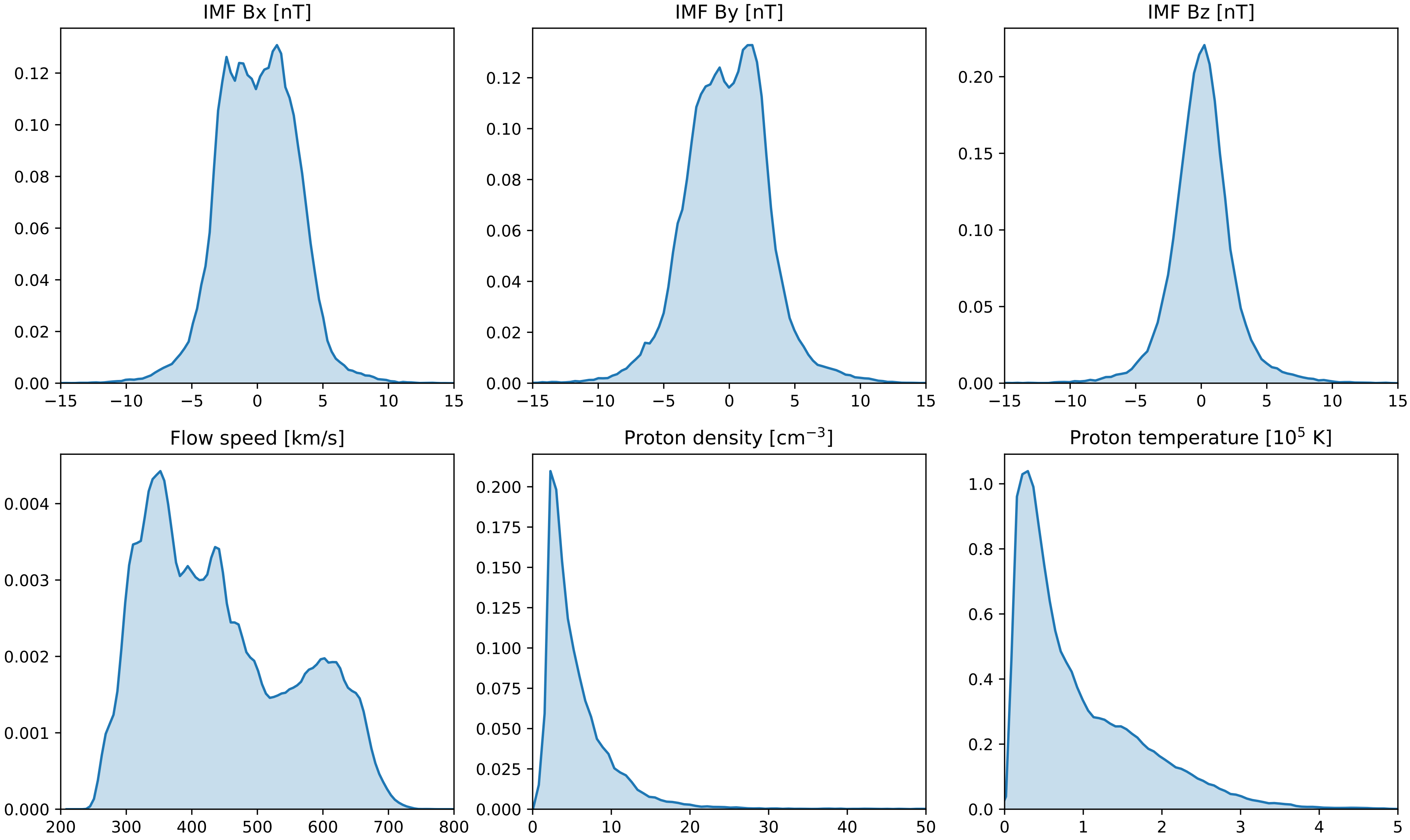
Histograms, kernel density estimators, sample statistics

- Pseudo-random number generators (`numpy.random`, `scipy.stats`).
- Histogram dependence on bin width.
- Kernel density estimation dependence on bandwidth and kernel type.
- Statistical accuracy of sample moments for varying sample size.
- Graphical representation of bivariate data.





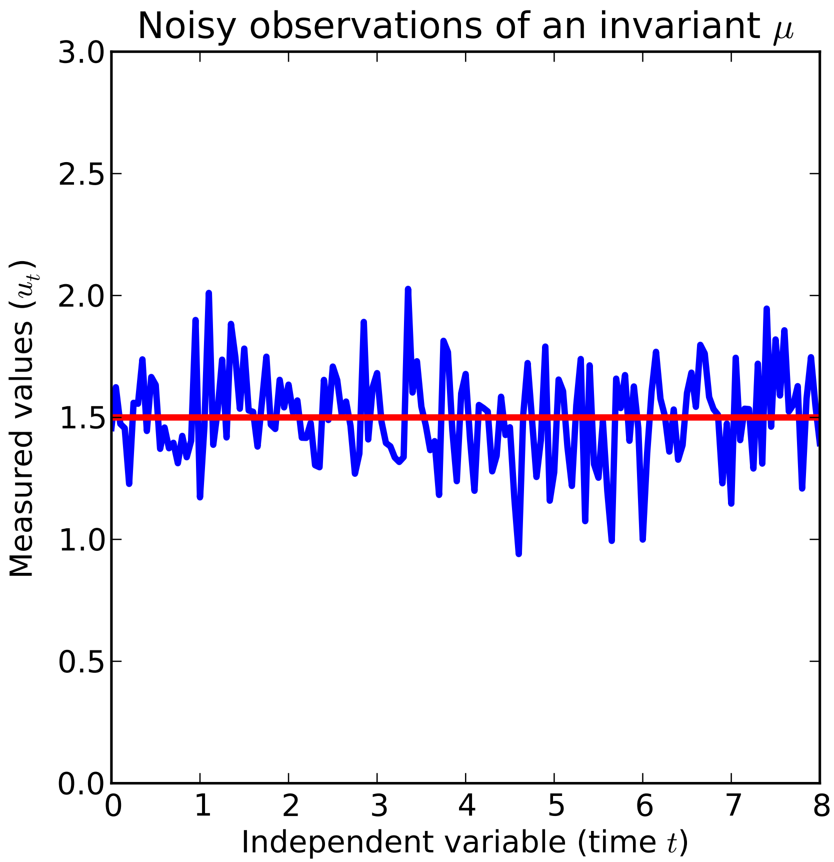
Empirical distributions of solar wind data



Data modeling and residual distributions

Random measurement errors

The term *noise* refers to *random* components of *imperfect measurements*. Random errors can be reduced by suitable averaging operations, and must be distinguished from systematic errors that may yield a bias.



Example: repeated measurements d_t of an invariant parameter μ in the presence of (additive) noise r_t:

$$d_t = \mu + r_t \; .$$

Questions:

- How to *estimate* μ from d_t ?
- *Uncertainty* of estimate ?
- Noise characteristics ?
- *Validity* of assumptions ?

Terminology

In model equations of the type $d_t = m_t + r_t$ or $d(t) = m(t) + r(t)$,

- the model function $m_t = m(t)$ may be referred to as *prediction*, and
- the noise term $r_t = d_t - m_t$ may also be called *residual* or error.

The residual is (a realization of) a random variable to be characterized through the underlying distribution. Assuming a normal distribution with zero mean, we need to specify only the variance σ^2.

The prediction $m(t)$ may depend on one or more *model parameters* $\boldsymbol{\omega}$. We write $m(t|\boldsymbol{\omega})$ to make this dependence explicit, e.g., $m(t|\mu) = \mu$.

Important aspects of modeling

- *Model selection*: based on what you know (or presume) about the underlying process(es), choose an appropriate model function m.
- *Parameter fitting*: estimate the model parameters from the data.
- Compute *parameter uncertainties* from measurement errors.
- Check the *validity of the chosen model* using statistical tests.

Parameter estimation strategies

Parameter estimation can be based on different optimality criteria.

- Maximize the (data) likelihood $P(d|m,\boldsymbol{\omega})$: probability distribution needs to be known for *maximum-likelihood (ML) estimation*.
- Minimize the *absolute deviation* $\sum_t |d(t) - m(t|\boldsymbol{\omega})|$: robust class of estimators but only few analytical results.
- Minimize $\sum_t [d(t) - m(t|\boldsymbol{\omega})]^2$ to yield a *least squares estimator*: well understood and many analytical results but less robust.

Minimum absolute deviation example

For $m(t|\mu) = \mu$, show that the median minimizes absolute deviation.

We write $A(\mu) = \sum_t |d_t - \mu| = \sum_t |\mu - d_t| = \sum_t s_t \cdot (\mu - d_t)$ with

$$s_t \;=\; \mathrm{sgn}(\mu - d_t) \;=\; \begin{cases} +1 & , \quad d_t < \mu \;, \\ -1 & , \quad d_t > \mu \;. \end{cases}$$

The derivative $A'(\mu) = \sum_t s_t$ is zero when the values d_t occur in equal numbers above and below μ, hence the estimator for μ is the median.

Maximum-likelihood (ML) estimation

When data form an independent and identically distributed (i.i.d.) sample drawn from a random variable with *density* $p(d|m,\boldsymbol{\omega}) = p(d|\boldsymbol{\omega})$, the *likelihood function* is $\prod_t p(d_t|\boldsymbol{\omega})$. ML estimators for $\omega_1, \omega_2, \ldots$ satisfy:

$$0 \;=\; \sum_t \frac{1}{p(d_t|\boldsymbol{\omega})} \frac{\partial p(d_t|\boldsymbol{\omega})}{\partial \omega_n} \;=\; \sum_t \frac{\partial \ln p(d_t|\boldsymbol{\omega})}{\partial \omega_n} \;, \;\; n = 1, 2, \ldots$$

Maximum-likelihood estimation example

Construct the ML estimator for λ in $p(d) \;=\; \lambda\, \mathrm{e}^{-\lambda d}\; (d \geq 0)$.

The derivative of $\ln p = \ln \lambda - \lambda d$ is $\frac{\partial \ln p}{\partial \lambda} = \frac{1}{\lambda} - d$. Using the sample mean $\bar{d} = \frac{1}{N} \sum_t d_t$ we find

$$0 \;=\; \sum_t \left(\frac{1}{\lambda} - d_t \right) \;=\; \frac{N}{\lambda} - \sum_t d_t \;=\; N \left(\frac{1}{\lambda} - \bar{d} \right) \;.$$

The ML estimator of λ is thus $\hat{\lambda} = 1/\bar{d}$.

Least squares approach to parameter estimation

When the measurement errors $r_t = d_t - m_t$ form an i.i.d. sample drawn from a *normal distribution* $p(r_t|\boldsymbol{\omega}) \propto \exp\{-r_t^2/2\sigma^2\}$, maximizing the likelihood is equivalent to minimizing $\sum_t r_t^2/2\sigma^2 \propto \sum_t (d_t - m_t)^2$, i.e., the least squares approach.

When the uncertainty of individual measurements is not independent of t, then $\sigma \to \sigma_t$ and the *least squares condition* becomes

$$\chi^2 = \sum_t \left(\frac{d(t) - m(t|\boldsymbol{\omega})}{\sigma_t} \right)^2 \stackrel{!}{=} \text{Min} .$$

Least squares estimation example

Consider again the model $m(t|\mu) = \mu$, and assume that σ does not change with t. Compute the least squares estimator for μ.

Result: $\hat{\mu} = \dfrac{1}{N} \sum\limits_t d_t = \bar{d}$.

Checking the normality assumption

A large class of statistical techniques rely on the assumption that residuals or other data are normally distributed. Methods to check this assumption are referred to as *normality tests*.

Graphical methods

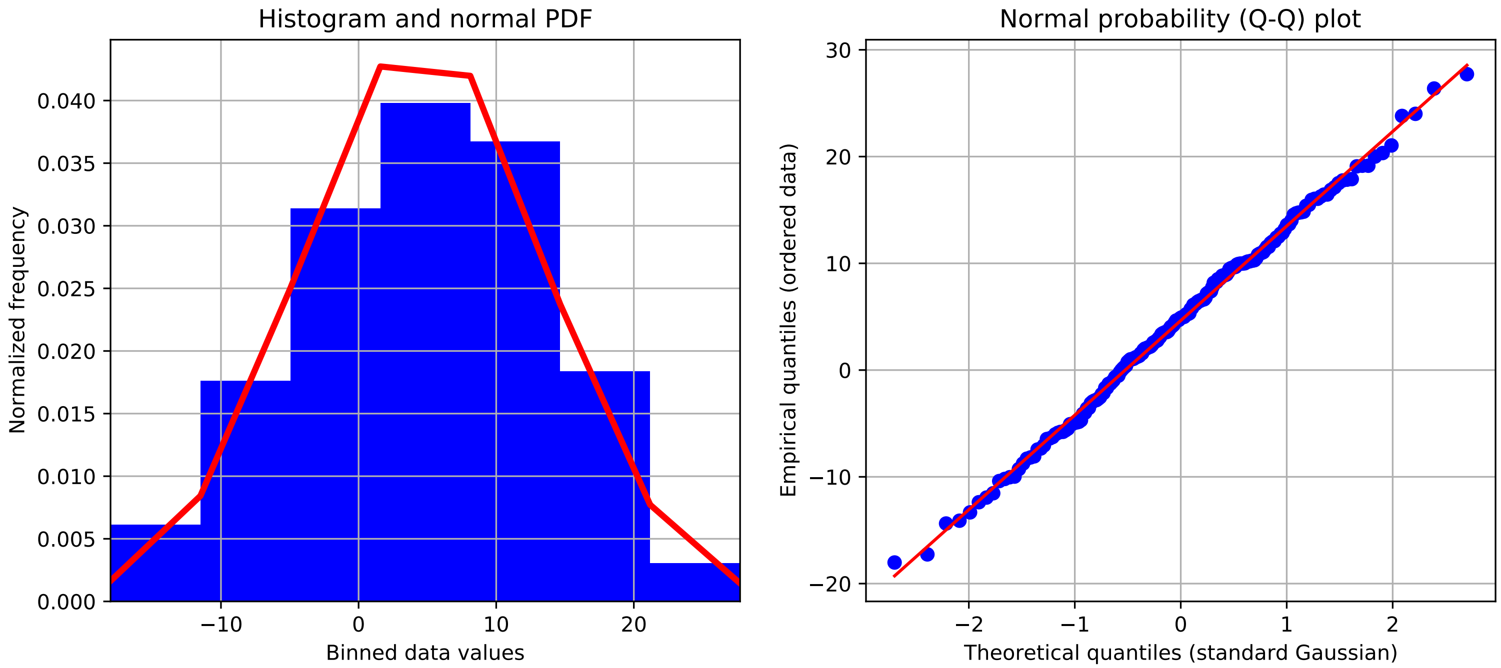
- Visual inspection of the histogram versus the model PDF curve.
- Normal probability plots: draw empirical probabilities or quantiles versus the corresponding values predicted by a normal distribution.
  - Normal P-P plot: empirical CDF (ordered data) vs normal CDF.
  - Normal Q-Q plot: empirical PPF (quantiles) vs normal PPF.

Numerical methods

- Check sample statistics and compare with theoretical values as predicted by the normal distribution (moments, extreme values).
- Hypothesis testing, with normality as the null hypothesis: e.g., Anderson-Darling test, D'Agostino's K^2 test, Shapiro-Wilk test.

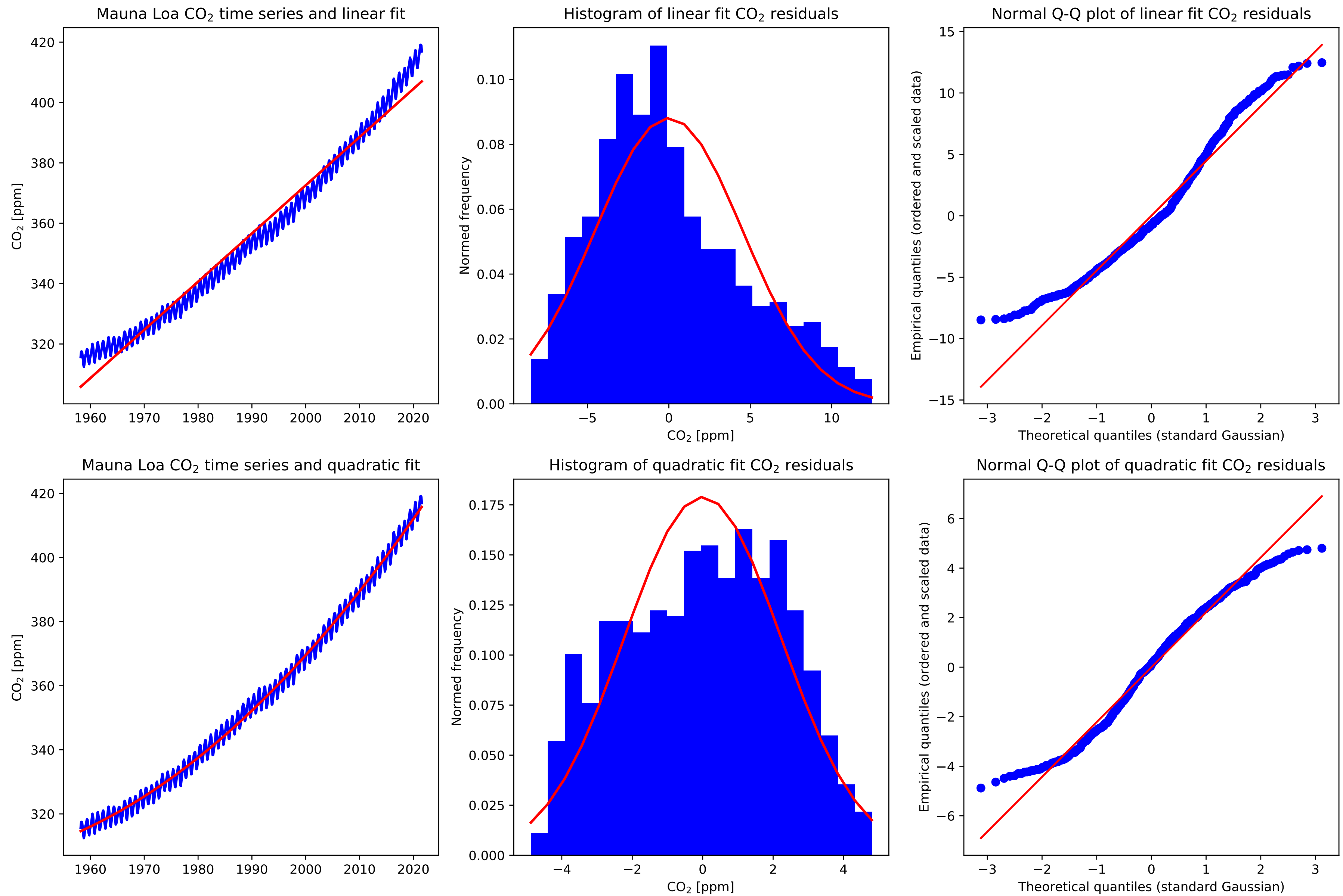
Histograms and normal probability (Q-Q) plots

- Graphical normality checks using random variates from
  - a Gaussian distribution (symmetric and normal tails),

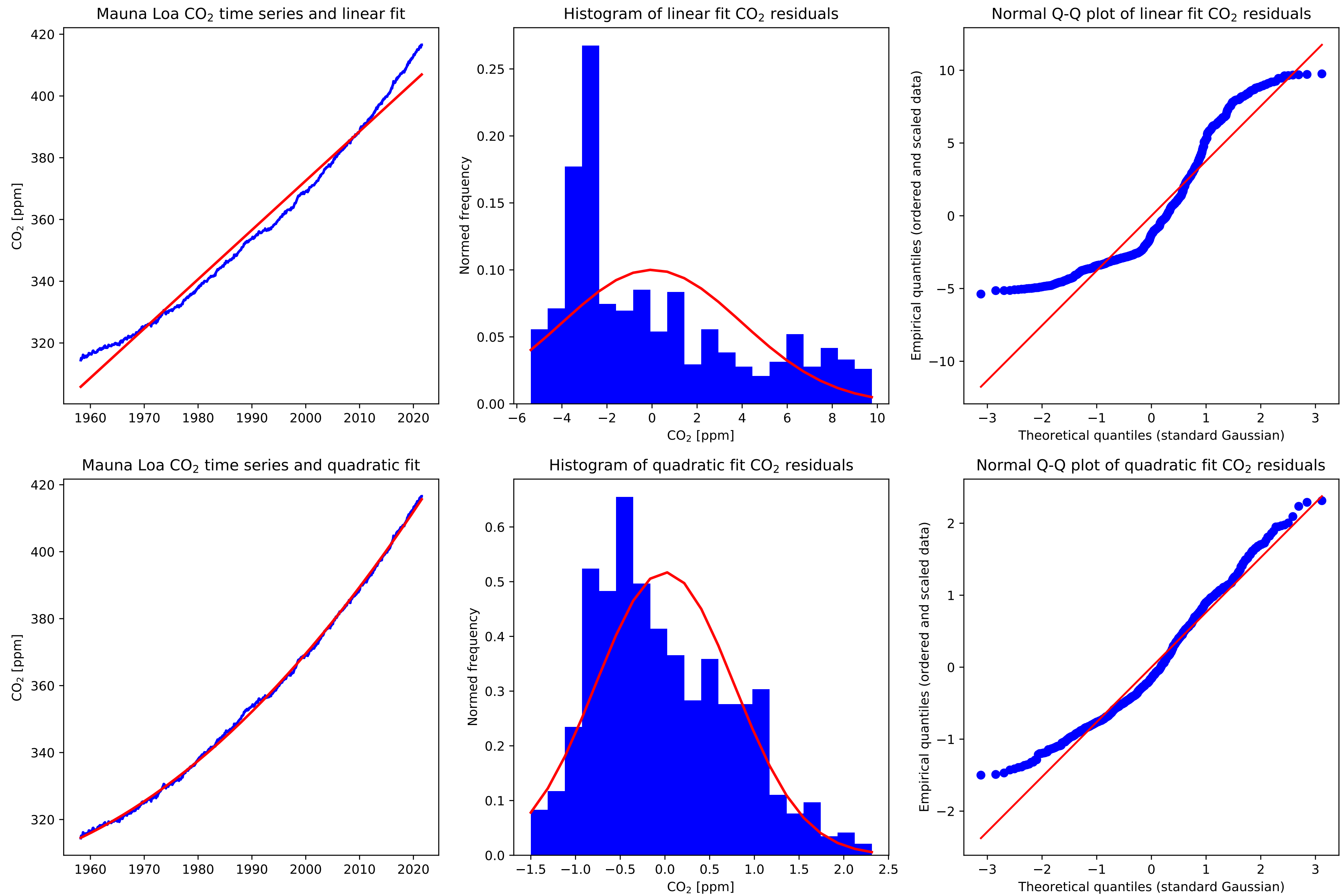


  - a symmetric distribution with short tails,
  - a symmetric distribution with long tails,
  - a non-symmetric (skewed) distribution.
- Exercise: Residuals of CO_2 Mauna Loa time series.

Residuals of CO_2 Mauna Loa time series (1)



Residuals of CO_2 Mauna Loa time series (2)



Confidence intervals and standard errors

Confidence intervals

Statistical significance: How reliable/robust are statistical estimates?

Confidence interval: estimated range of values containing an unknown (distribution, population) parameter for a given probability threshold.

- *Significance level* α: probability that the confidence interval *does not contain* the true value of the parameter. Commonly used values: $\alpha = 0.1, 0.05, 0.01$ (corresponding to percentages $10\%, 5\%, 1\%$).
- *Confidence level* $\gamma = 1 - \alpha$: probability that the true value of the parameter lies within the confidence interval.

Confidence interval for the mean of normally distributed measurements

- Consider data $\{u_1, u_2, \ldots, u_N\}$ distributed according to a normal distribution with (true) mean μ and (true) standard deviation σ.
- Estimates of μ and σ are the sample mean $\bar{u} = \frac{1}{N}\sum_{j=1}^{n} u_j$ and the sample standard deviation $\Delta u = \sqrt{\frac{1}{N-1}\sum_{j=1}^{n}(u_j - \bar{u})^2}$, respectively.
- A confidence interval for μ is of the form $[\bar{u} - h_\alpha, \bar{u} + h_\alpha]$ where $h_\alpha = h_\alpha(u)$ is the interval half-width for a prescribed significance α.

Confidence interval for the mean

Unknown standard deviation σ and small sample size N

In general, the confidence interval for the true mean μ must be constructed using *Student's t distribution* with $N-1$ degrees of freedom:

$$h_\alpha \; = \; t_{\alpha,N-1} \, \frac{\Delta u}{\sqrt{N}} \; .$$

Here $t_{\alpha,N-1}$ is the value of the t distribution quantile function at $1-\frac{\alpha}{2}$.

Unknown standard deviation σ and large sample size N

For large sample sizes ($N > 30$ is often recommended), the t distribution is well approximated by a normal distribution so that the half-width can be written as

$$h_\alpha \; = \; z_\alpha \, \frac{\Delta u}{\sqrt{N}}$$

where z_α is the score of the normal distribution quantile function at $1-\frac{\alpha}{2}$.

Known standard deviation σ

If σ is known, and z_α is the normal quantile function at $1-\frac{\alpha}{2}$, then $h_\alpha \; = \; z_\alpha \, \frac{\sigma}{\sqrt{N}}$.

Error propagation concepts

A scientific measurement needs to be furnished with an indication of its uncertainty (error, standard deviation, confidence interval).

Example: Numerical value and error of the atomic mass constant*

- $m_\mathsf{u} \;=\; (1.660\,539\,066\,60 \pm 0.000\,000\,000\,50) \cdot 10^{-27}$ kg ,
- or, in a more concise notation: $m_\mathsf{u} = 1.660\,539\,066\,60(50) \cdot 10^{-27}$ kg.

Error propagation: when erroneous variables $(U_1, U_2, \ldots)$ are transformed or combined, how large is the uncertainty of the resulting quantity (V) ?

Measure of uncertainty considered here: standard error, standard deviation

$$\Delta U \;=\; \sqrt{(\Delta U)^2} \;=\; \sqrt{\mathrm{E}\{(U - \bar{U})^2\}}$$

Error propagation example: scaling operation

Suppose $V \;=\; aU$ for an exact number $a \in \mathbb{R}$. Compute ΔV.

$$\Delta V \;=\; \sqrt{\mathrm{E}\{(aU - \overline{aU})^2\}} = \sqrt{\mathrm{E}\{a^2(U - \overline{U})^2\}} = \sqrt{a^2\mathrm{E}\{(U - \overline{U})^2\}} = |a|\sqrt{\mathrm{E}\{(U - \overline{U})^2\}} = |a|\Delta U.$$

*According to NIST, https://physics.nist.gov/, 23 July 2020.

Error propagation formulas

Power law: $V = U^q$ for an exact number $q \in \mathbb{R}$, then $\frac{\Delta V}{|V|} = |q| \frac{\Delta U}{|U|}$.
Note $\frac{\Delta V}{|V|} = \frac{\Delta U}{|U|}$ for $q = -1$, i.e. $V = U^{-1} = 1/U$.

In the following, the random variables $U_1, U_2, \ldots$ are assumed to be mutually uncorrelated: $\mathrm{cov}(U_j, U_k) = 0$ for $j \neq k$.

Addition (and difference): $V = a_1 U_1 + a_2 U_2 + a_3 U_3 \cdots = \sum_j a_j U_j$, then

$$(\Delta V)^2 = a_1^2 (\Delta U_1)^2 + a_2^2 (\Delta U_2)^2 + \cdots = \sum_j a_j^2 (\Delta U_j)^2 \ .$$

Multiplication (and division): $V = U_1^{q_1} \cdot U_2^{q_2} \cdot U_3^{q_3} \cdots = \prod_j U_j^{q_j}$, then

$$\frac{(\Delta V)^2}{V^2} = q_1^2 \frac{(\Delta U_1)^2}{U_1^2} + q_2^2 \frac{(\Delta U_2)^2}{U_2^2} \cdots = \sum_j q_j^2 \frac{(\Delta U_j)^2}{U_j^2} \ .$$

General *nonlinear estimation* rules $V = V(U_1, U_2, \ldots)$, small errors ΔU_j:

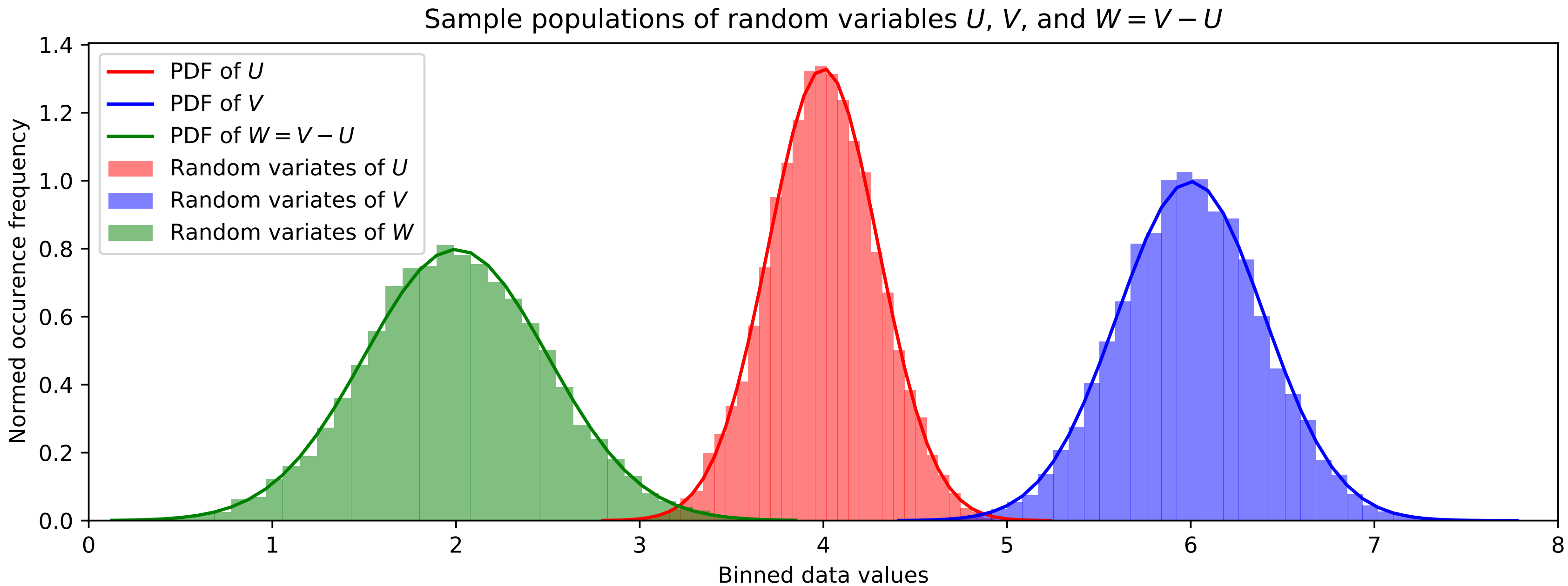
$$(\Delta V)^2 = \sum_j \left(\frac{\partial V}{\partial U_j} \right)^2 (\Delta U_j)^2 \ .$$

Numerical Software Lab

Confidence intervals for the mean of normally distributed data

- Large sample sizes (and/or known standard deviation): confidence interval are constructed from normal distribution parameters.
- Small sample sizes (and unknown standard deviation): confidence interval are constructed from Student's t distribution.

Standard errors for arithmetic operations is illustrated using normally distributed random variates.



Hypothesis testing and binary classification

Binary classification tests

Binary classification tests: terminology in biostatistics

- Condition: present or absent. Test result: positive or negative.
- True/False positive: predicted condition (test result) is positive and correctly/incorrectly classified by the test (type I error).
- True/False negative: predicted condition (test result) is negative and correctly/incorrectly classified by the test (type II error).
- *Sensitivity (true positive rate, TPR)*: proportion of correctly classified objects among those where the condition is present (sick persons).
- *Specificity (true negative rate, TNR)*: proportion of correctly classified objects among those where the condition is absent (healthy persons).
- *Prevalence*: overall proportion of objects where the condition is present (proportion of sick persons in the total population).

Outside biostatistics, sensitivity is usually referred to as the *power of a test* and written in the form $(1 - \beta)$ where β is the false negative rate.

Binary classification tests (continued)

Sensitivity, specificity, prevalence

Express sensitivity, specificity, prevalence using (conditional) probabilities.

Sensitivity: $\mathrm{P}(\text{pos}|\text{pre})$. Specificity: $\mathrm{P}(\text{neg}|\text{abs}) = \mathrm{P}(\overline{\text{pos}}|\overline{\text{pre}})$. Prevalence: $\mathrm{P}(\text{pre})$.

Further terminology in binary classification tests

- False positive rate (FPR, probability of false alarm): $\mathrm{P}(\text{pos}|\text{abs})$.
- False negative rate (FNR, miss rate): $\mathrm{P}(\text{neg}|\text{pre})$.
- Positive predictive value (PPV): $\mathrm{P}(\text{pre}|\text{pos})$.
- False discovery rate (FDR): $\mathrm{P}(\text{abs}|\text{pos})$.
- Negative predictive value (NPV): $\mathrm{P}(\text{abs}|\text{neg})$.
- False omission rate (FOR): $\mathrm{P}(\text{pre}|\text{neg})$.

False positives and false negatives

Relate FPR and FNR to sensitivity (TPR) and specificity (TNR).

$\text{FPR} = 1 - \mathrm{P}(\overline{\text{pos}}|\text{abs}) = 1 - \text{TNR}$. $\text{FNR} = 1 - \mathrm{P}(\overline{\text{neg}}|\text{pre}) = 1 - \text{TPR}$.

Statistical hypothesis testing

Statistical hypothesis: statement that can be tested by statistical means, typically a statement about the underlying distribution of a data set.

Statistical hypothesis testing: method in inferential statistics.

- Formulate a *null hypothesis* H_0 and (at least implicitly) an alternative hypothesis H_1 referring to a testable property of the distribution.
- Define a *test statistic* T sensitive to differences between H_0 and H_1.
- Choose a *significance level* α (maximum tolerated false positive rate).
- For the chosen test statistic T, find the value t from the sample.
- Compute the *p value* (probability value): probability that an empirical estimate of T is at least as extreme as the observed value t.
- *The null hypothesis H_0 is rejected (in favor of H_1) if $p < \alpha$.*
- *Type I error*: a true null hypothesis is rejected (false positive).
- *Type II error*: a false null hypothesis is not rejected (false negative).

Null hypothesis in medical tests: condition is absent (healthy person).

Examples of statistical tests

Z tests

- Under H_0, the test statistic is normally distributed: $T \sim \mathcal{N}(\mu, \sigma^2)$.
- *Location tests*: Suppose μ and σ are known, does the mean $\bar{u}$ of a particular sample differ significantly from μ?

Student's t test

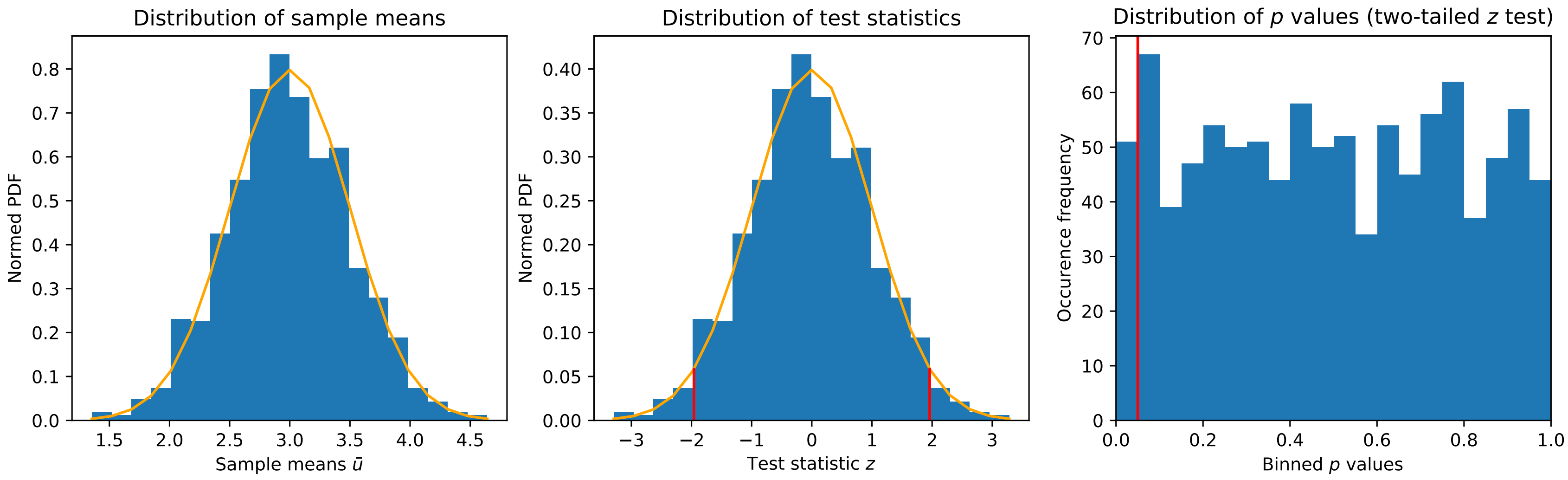
- Under H_0, the test statistic T follows a Student's t distribution.
- *One-sample location tests*: Assuming the mean μ of a normally distributed sample is known but its standard deviation σ is unknown, test if the sample mean $\bar{u}$ differs significantly from μ.
- *Two-sample location tests*: Compare the means of two samples.

Normality tests

- Check if a given sample follows a normal distribution.
- Anderson-Darling test: T is constructed using the empirical CDF.
- D'Agostino's K^2 test: T is based on skewness and kurtosis.
- Shapiro-Wilk test: related to the normal Q-Q plot.

Statistical hypothesis testing, implementations in `scipy.stats`

- *Z tests*: implementation of two-tailed test using the normal CDF, random number experiments to illustrate the significance level α.



- *t tests*: implementation of one-sample test using the normal CDF, test of function `ttest_1samp()`, exercise using `ttest_ind()`.
- *Normality tests*: illustration of Shapiro-Wilk test, additional exercise.

Bootstrap approach to error estimation

Parameter errors and model accuracy

How can we assess the accuracy of model parameters and predictions?

Analytical error propagation formulas: relationships between the standard deviations of input errors (data residuals) and output errors (parameter uncertainties), implicitly assuming that all distributions are Gaussian (normal theory statistics). For models that are nonlinear in the parameters, the expressions are valid only in the small-error limit.

Monte Carlo error estimates: can be generated using random variates from normal or non-normal distributions by generating surrogate data sets and studying ensembles of estimated parameters. The Monte Carlo approach works for linear and nonlinear parametric models.

Bootstrap approach to error estimation: special Monte Carlo method where the underlying distribution does not need to be known a priori. Surrogate measurements are constructed from a single data set through resampling with replacement.

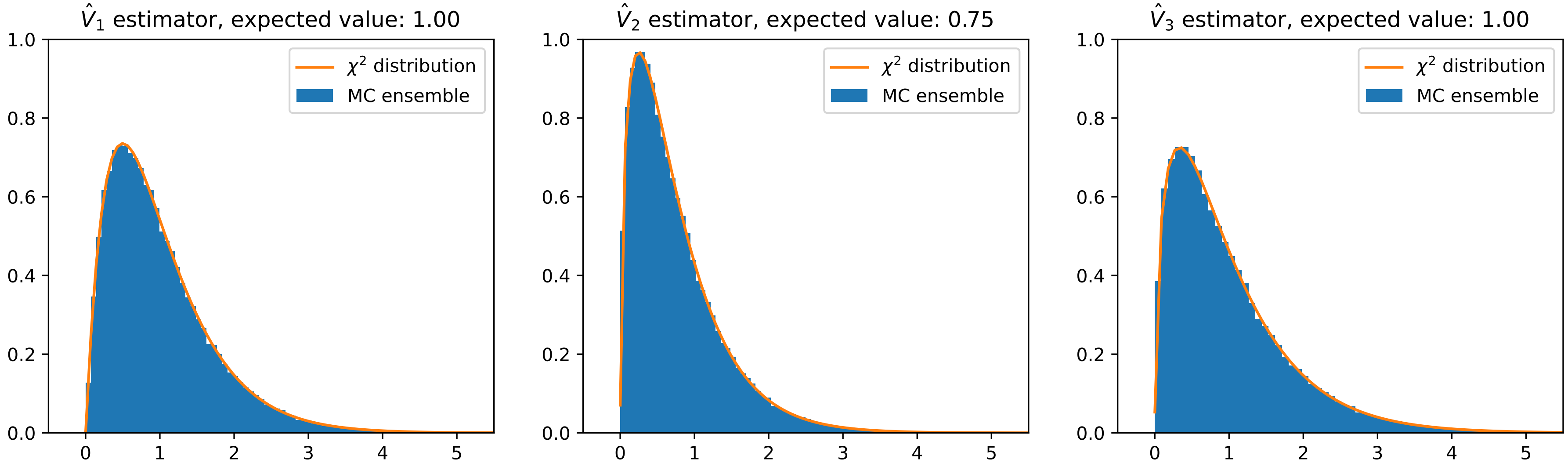
Illustration of the Monte Carlo method

Biased and unbiased estimators for the variance

Consider a sample $\{u_1, u_2, \ldots, u_N\}$ drawn from a standard normal distribution with mean $\mu = 0$ and variance $V = \sigma^2 = 1$. Using random number (Monte Carlo) experiments, test which of the following estimators for the variance is biased:

$$\hat{V}_1 = \frac{1}{N} \sum_k (u_k - \mu)^2 \ , \ \hat{V}_2 = \frac{1}{N} \sum_k (u_k - \bar{u})^2 \ , \ \hat{V}_3 = \frac{1}{N-1} \sum_k (u_k - \bar{u})^2 \ .$$

Here $\bar{u} = \frac{1}{N} \sum_k u_k$ denotes the sample mean.



Bootstrap approach to error estimation

The *bootstrap method* provides Monte Carlo estimates of errors and confidence intervals without reference to a model distribution.

- A *single data set* $\{u_1, u_2, \ldots, u_N\}$ is available.
- A statistic of interest v is computed from the data set.
- Based on $\{u_1, u_2, \ldots, u_N\}$, an *ensemble of surrogate data sets* $\{u_1^*, u_2^*, \ldots, u_N^*\}$ is constructed through *resampling with replacement*.
- Each data set in the resampling ensemble yields an estimate v^* of v.
- Standard errors and/or confidence intervals for the statistic v are obtained from the ensemble of bootstrap realizations v^*.

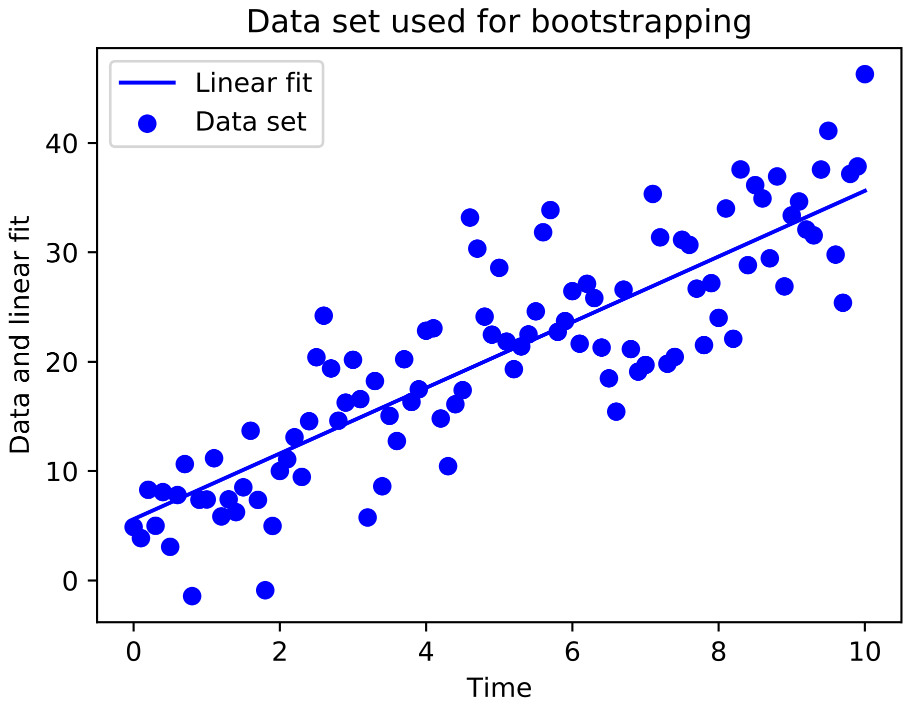
Bootstrap confidence interval for the mean (significance α, confidence $1-\alpha$)

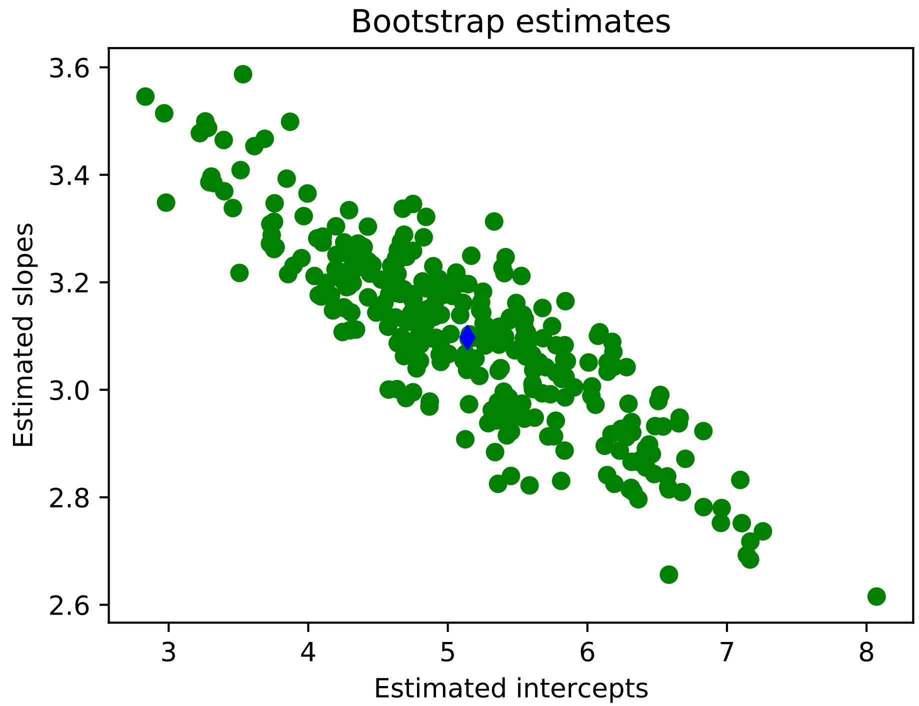
- Consider distribution of $D = \bar{u} - \mu$. Critical values: $D_- = D_{1-\alpha/2}$, $D_+ = D_{\alpha/2}$.
- $\mathrm{P}(D_- \leq D \leq D_+) = 1-\alpha \iff \mathrm{P}(\bar{u} - D_+ \leq \mu \leq \bar{u} - D_-) = 1 - \alpha$.
- *Confidence interval for the mean*: $[\bar{u} - D_+, \bar{u} - D_-]$.
- Consider $D^* = \bar{u}^* - \bar{u}$ as an approximation of $D = \bar{u} - \mu$, and sort the D^* ensemble to obtain D_-^* (quantile at $\frac{\alpha}{2}$) and D_+^* (quantile at $1 - \frac{\alpha}{2}$).
- *Boostrap confidence interval*: $[\bar{u} - D_+^*, \bar{u} - D_-^*]$.

Numerical Software Lab

Monte Carlo simulations and bootstrap method

- Monte Carlo distributions of variance estimators
- Monte Carlo simulations of straight line fits
- Bootstrap estimation of linear fit parameter errors
- Bootstrap confidence interval for the mean





Project: Statistical concepts