Lecture 1:

Essential statistics

Shane Elipot

The Rosenstiel School of Marine and Atmospheric Science,
University of Miami

Created with {Remark.js} using {Markdown} + {MathJax}

Foreword

Statistics (*noun, plural in form but singular or plural in construction*)

1. *Merriam-Webster*: a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data

Foreword

Statistics (*noun, plural in form but singular or plural in construction*)

1. *Merriam-Webster*: a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data

2. *Oxford dictionnary of English*:

   a. the practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

   b. a collection of quantitative data (e.g. *The statistics of the data are unknown.*)

Foreword

Statistics (*noun, plural in form but singular or plural in construction*)

1. *Merriam-Webster*: a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data

2. *Oxford dictionnary of English*:

   a. the practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

   b. a collection of quantitative data (e.g. *The statistics of the data are unknown.*)

Statistic (*noun*) a single term or datum in a collection of statistics (e.g. *The mean of the data is zero.*)

References

[1] Bendat, J. S., & Piersol, A. G. (2011). *Random data: analysis and measurement procedures* (Vol. 729). John Wiley & Sons.

[2] Thomson, R. E., & Emery, W. J. (2014). *Data analysis methods in physical oceanography*. Newnes. dx.doi.org/10.1016/B978-0-12-387782-6.00003-X

[3] Taylor, J. (1997). *Introduction to error analysis, the study of uncertainties in physical measurements*.

[4] Press, W. H. et al. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

[5] Kanji, G. K. (2006). *100 statistical tests*. Sage.

[6] von Storch, H. and Zwiers, F. W. (1999). *Statistical Analysis in Climate Research*, Cambridge University Press

Foreword

A quote from *Data Analysis Methods in Physical Oceanography*, Thompson and Emery, Third Edition, 2014:

"Statistical methods are essential to determining the value of the data and to decide how much of it can be considered useful for the intended analysis. This statistical approach arises from the fundamental complexity of the ocean, a **multivariate** system with many **degrees of freedom** in which nonlinear dynamics and **sampling** limitations make it difficult to separate scales of **variability**."

What do **multivariate**, **degrees of freedom**, **sampling**, and **variability** mean in this context?

Lecture 1: Outline

1. Introduction
2. Estimation vs Truth
3. Fundamental Statistics
4. Common Distributions
5. Errors, Uncertainties, and hypothesis testing
6. Extra slides

1. Introduction

Introduction

In oceanography and atmospheric science, we have at our disposal a number of physical theories, that is equations, that describe the spatial and temporal variability of the climate system.

Introduction

In oceanography and atmospheric science, we have at our disposal a number of physical theories, that is equations, that describe the spatial and temporal variability of the climate system.

Specifically, in geophysical fluid dynamics, the Navier-Stokes equations describe the motion of a fluid in 2D or 3D. Yet, we do not know if these equations have reasonable physical solutions (If you figure this out, there's a $1M prize from the Clay Mathematics Institute). Assuming that they do, then the ocean, as an example, can be seen as being a *deterministic system*, which means that mathematical expressions can be used to describe completely the velocity and pressure field (e.g. $p(\mathbf{x}, t) = p_0 \cos(\omega t - \mathbf{k} \cdot \mathbf{x})$).

Introduction

In oceanography and atmospheric science, we have at our disposal a number of physical theories, that is equations, that describe the spatial and temporal variability of the climate system.

Specifically, in geophysical fluid dynamics, the Navier-Stokes equations describe the motion of a fluid in 2D or 3D. Yet, we do not know if these equations have reasonable physical solutions (If you figure this out, there's a $1M prize from the Clay Mathematics Institute). Assuming that they do, then the ocean, as an example, can be seen as being a *deterministic system*, which means that mathematical expressions can be used to describe completely the velocity and pressure field (e.g. $p(\mathbf{x}, t) = p_0 \cos(\omega t - \mathbf{k} \cdot \mathbf{x})$).

Yet, even if we do not know the analytical solutions, we can discretize the equations within a computer models to obtain approximate solutions describing the flow deterministically ... (in theory).

Introduction

A deterministic approach is not realistic because of three commonly acknowledged reasons. The ocean and the atmosphere are *complex*, *nonlinear*, and *unpredictable*.

Introduction

A deterministic approach is not realistic because of three commonly acknowledged reasons. The ocean and the atmosphere are *complex*, *nonlinear*, and *unpredictable*.

As an example, there is an estimated 4.7×10^{46} water molecules in the ocean (that's complexity) so that there are too many variables, and too many initial and boundary conditions to be specified, (jointly forming the number of *degrees of freedom* of the system) in order to solve all equations numerically in a computer model. Because many variables cannot be observed, or are unspecified at the start of a simulation, outcomes will appear *random* to the observer (that's *unpredictability*).

Introduction

A deterministic approach is not realistic because of three commonly acknowledged reasons. The ocean and the atmosphere are *complex*, *nonlinear*, and *unpredictable*.

As an example, there is an estimated 4.7×10^{46} water molecules in the ocean (that's complexity) so that there are too many variables, and too many initial and boundary conditions to be specified, (jointly forming the number of *degrees of freedom* of the system) in order to solve all equations numerically in a computer model. Because many variables cannot be observed, or are unspecified at the start of a simulation, outcomes will appear *random* to the observer (that's *unpredictability*).

Finally, ocean and atmosphere are *nonlinear* which means that you cannot really find a portion of the system (e.g. surface gravity waves) with a finite number of *degrees of freedom* whose evolution is isolated and can be made deterministic. Unknown perturbations will render the observations to contain randomness, or *noise*.

Introduction

Once we accept that the climate is not a purely deterministic system, but contain *randomness*, we can rely on a suite of tools especially applicable to *stochastic systems or processes*.

---

stochastic (*adjective, technical*)

having a random probability distribution or pattern that may be analysed statistically but may not be predicted precisely.

As a consequence, in the rest of this lecture, we will often discuss *random variables* (hereafter r.v.), that is variables to which statistical theory can be applied. In addition, we will take the approach that our system, or that our observational data, can be separated into *signal* plus *noise*. As an example, estimation of the seasonal cycle of ocean temperature (the sought after signal) is disturbed by changes due to ocean currents, but also changes due to the imperfections of your temperature sensors, etc.

2. Estimation vs truth

Estimation vs truth

We are in the business of *estimation*: we try to describe and/or understand the climate system by estimating the value of a *random variable*. This r.v. may be a physical variable such as air temperature, water vapor, sea ice area, sea surface temperature, sea level, snow cover, glacier volume, CO_2 concentration, etc.

Estimation vs truth

We are in the business of *estimation*: we try to describe and/or understand the climate system by estimating the value of a *random variable*. This r.v. may be a physical variable such as air temperature, water vapor, sea ice area, sea surface temperature, sea level, snow cover, glacier volume, CO_2 concentration, etc.

Or it may be a variable derived from one or several other r.v., in essence a *parameter*. This parameter is itself a r.v. As an example ocean heat content, the daily mean temperature, the decadal temperature trend, the acoustic travel time in water, the amplitude of the seasonal cyle of temperature, the adiabatic lapse rate, the power spectral density function of velocity, the precipitation rate, etc.

Estimation vs truth

We are in the business of *estimation*: we try to describe and/or understand the climate system by estimating the value of a *random variable*. This r.v. may be a physical variable such as air temperature, water vapor, sea ice area, sea surface temperature, sea level, snow cover, glacier volume, CO_2 concentration, etc.

Or it may be a variable derived from one or several other r.v., in essence a *parameter*. This parameter is itself a r.v. As an example ocean heat content, the daily mean temperature, the decadal temperature trend, the acoustic travel time in water, the amplitude of the seasonal cyle of temperature, the adiabatic lapse rate, the power spectral density function of velocity, the precipitation rate, etc.

The distinction between variable and parameter is maybe semantic. In any case, let's call ϕ a r.v. of interest.

Estimation vs truth

Unfortunatelly, it is likely that we will never know ϕ exactly, but only access an *estimate* that we will note $\widehat{\phi}$ (ϕ "hat"). This estimate means we use a given method or a given instrument to measure or calculate ϕ.

Estimation vs truth

Unfortunatelly, it is likely that we will never know ϕ exactly, but only access an *estimate* that we will note $\widehat{\phi}$ (ϕ "hat"). This estimate means we use a given method or a given instrument to measure or calculate ϕ.

As an example, we want to know the temperature of the room. We can:

1. use one temperature sensor at one fixed location (in the middle of the room), repeatedly through time, N times.
2. use N temperature sensors, once.
3. use one temperature sensor, used repeatedly N times, each time in a different corner of the room
4. etc.

Each one of these methods leads to one estimate of "the temperature of the room".

Expectation of an estimate

Let's assume that we design an experiment and obtain $\widehat{\phi}$, repeatedly N times. The *expectation* value of $\widehat{\phi}$, denoted $E[\widehat{\phi}]$, is

$$E[\widehat{\phi}] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} \widehat{\phi}_n$$

where $\widehat{\phi}_n$ is the estimate from the n-th experiment. Unfortunatelly, since we must have $N \to \infty$, the expectation cannot be known.

Expectation of an estimate

Let's assume that we design an experiment and obtain $\widehat{\phi}$, repeatedly N times. The *expectation* value of $\widehat{\phi}$, denoted $E[\widehat{\phi}]$, is

$$E[\widehat{\phi}] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} \widehat{\phi}_n$$

where $\widehat{\phi}_n$ is the estimate from the n-th experiment. Unfortunatelly, since we must have $N \to \infty$, the expectation cannot be known.

However, if we can reasonably make some assumptions about the statistics of the r.v. itself, and/or if we know the specifications of our instruments, then we can sometimes derive the expectation of our estimator.

Expectation of an estimate

Let's assume that we design an experiment and obtain $\widehat{\phi}$, repeatedly N times. The *expectation* value of $\widehat{\phi}$, denoted $E[\widehat{\phi}]$, is

$$E[\widehat{\phi}] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} \widehat{\phi}_n$$

where $\widehat{\phi}_n$ is the estimate from the n-th experiment. Unfortunatelly, since we must have $N \to \infty$, the expectation cannot be known.

However, if we can reasonably make some assumptions about the statistics of the r.v. itself, and/or if we know the specifications of our instruments, then we can sometimes derive the expectation of our estimator.

Why is this important?

Expectation of an estimate

Let's assume that we design an experiment and obtain $\widehat{\phi}$, repeatedly N times. The *expectation* value of $\widehat{\phi}$, denoted $E[\widehat{\phi}]$, is

$$E[\widehat{\phi}] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} \widehat{\phi}_n$$

where $\widehat{\phi}_n$ is the estimate from the n-th experiment. Unfortunatelly, since we must have $N \to \infty$, the expectation cannot be known.

However, if we can reasonably make some assumptions about the statistics of the r.v. itself, and/or if we know the specifications of our instruments, then we can sometimes derive the expectation of our estimator.

Why is this important? To assess how good our estimate is.

Expectation of an estimate

Let's assume that we have access to the expectation $E[\widehat{\phi}]$ of an estimator $\widehat{\phi}$ of the r.v. ϕ.

Expectation of an estimate

Let's assume that we have access to the expectation $E[\widehat{\phi}]$ of an estimator $\widehat{\phi}$ of the r.v. ϕ.

If we are lucky, $E[\widehat{\phi}] = \phi$, and the estimator $\widehat{\phi}$ is said to be *unbiased*.

Expectation of an estimate

Let's assume that we have access to the expectation $E[\widehat{\phi}]$ of an estimator $\widehat{\phi}$ of the r.v. ϕ.

If we are lucky, $E[\widehat{\phi}] = \phi$, and the estimator $\widehat{\phi}$ is said to be *unbiased*. Otherwise, it is said to be *biased* and we define *the bias of the estimator*:

$$b[\widehat{\phi}] = E[\widehat{\phi}] - \phi,$$

which constitutes a *systematic error* of the estimate.

Expectation of an estimate

Let's assume that we have access to the expectation $E[\widehat{\phi}]$ of an estimator $\widehat{\phi}$ of the r.v. ϕ.

If we are lucky, $E[\widehat{\phi}] = \phi$, and the estimator $\widehat{\phi}$ is said to be *unbiased*. Otherwise, it is said to be *biased* and we define *the bias of the estimator*:

$$b[\widehat{\phi}] = E[\widehat{\phi}] - \phi,$$

which constitutes a *systematic error* of the estimate.

In addition, the value of the estimator will change from one experiment to the next, so we define the *variance* of the estimator, denoted $\mathrm{Var}[\widehat{\phi}]$, as

$$\mathrm{Var}[\widehat{\phi}] = E[(\widehat{\phi} - E[\widehat{\phi}])^2],$$

which constitutes the *random error* of the estimate.

Expectation of an estimate

The bias and the variance of the estimator contributes both to its *total error*, which can be assessed by the *mean square error* (MSE):

$$MSE[\widehat{\phi}] = E[(\widehat{\phi} - \phi)^2] = \mathrm{Var}[\widehat{\phi}] + (b[\widehat{\phi}])^2,$$

usually reported as the *root mean square error*:

$$RMS = \sqrt{MSE[\widehat{\phi}]} = \sqrt{\mathrm{Var}[\widehat{\phi}] + (b[\widehat{\phi}])^2}$$

Another sometimes useful quantity is the *normalized rms error*:

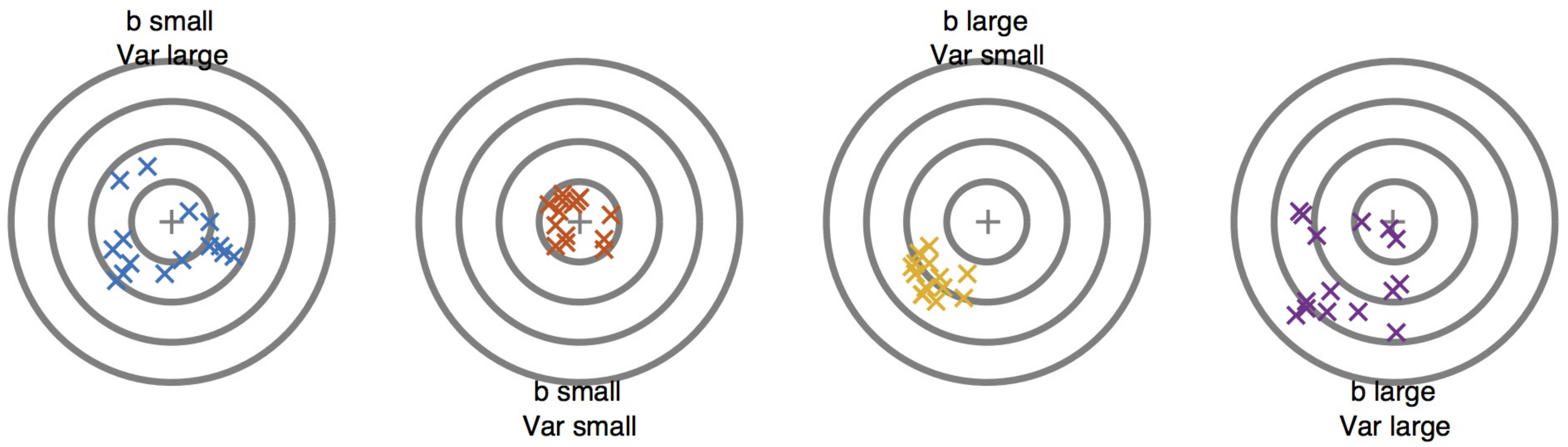
$$\varepsilon[\phi] = \frac{\sqrt{E[(\widehat{\phi} - \phi)^2]}}{\phi} \quad \text{for} \quad \phi \neq 0,$$

which is unitless and can be reported as a percentage.

Estimation vs truth

From the International Organization for Standardization (ISO) publication 5725-1:1994 *Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions*

Introduction 0.1 ISO 5725 uses two terms "trueness" and "precision" to describe the accuracy of a measurement method. "Trueness" refers to the closeness of agreement between the arithmetic mean of a large number of test results and the true or accepted reference value. "Precision" refers to the closeness of agreement between test results.



Example

Specification sheets of Seabird 911 CTD Plus sensors:

Manual version 018 | Section 2: Description of SBE 9*plus* | SBE 9*plus*

Specifications

	Temperature (°C)	Conductivity (S/m)	Pressure	A/D Inputs
Measurement Range	-5 to +35	0 to 7	0 to full scale range (in meters of deployment depth capability): 1400 / 2000 / 4200 / 6800 / 10500 meters	0 to +5 volts
Initial Accuracy	± 0.001	± 0.0003	± 0.015% of full scale range	± 0.005 volts
Typical Stability	0.0002/month	0.0003/month	0.02% of full scale range/year	0.001 volts/month
Resolution at 24 Hz	0.0002	0.00004	0.001% of full scale range	0.0012 volts

Example

Specification sheets of Seabird 911 CTD Plus sensors:

Manual version 018 Section 2: Description of SBE 9*plus* SBE 9*plus*

Specifications

	Temperature (°C)	Conductivity (S/m)	Pressure	A/D Inputs
Measurement Range	-5 to +35	0 to 7	0 to full scale range (in meters of deployment depth capability): 1400 / 2000 / 4200 / 6800 / 10500 meters	0 to +5 volts
Initial Accuracy	± 0.001	± 0.0003	± 0.015% of full scale range	± 0.005 volts
Typical Stability	0.0002/month	0.0003/month	0.02% of full scale range/year	0.001 volts/month
Resolution at 24 Hz	0.0002	0.00004	0.001% of full scale range	0.0012 volts

An interpretation is that the accuracy is the total error, or RMS error which is originally only the random error. The stability implies that the bias, originally zero, increases with time.

2. Fundamental statistics

Fundamental statistics

Let's consider a random variable x, for which we obtain (that is measure, calculate) values X_n, dependent on an index n along a given dimension, or scale (e.g. temperature as a function of time, sea level along a satellite track, oxygen concentration as a function of depth etc).

Fundamental statistics

Let's consider a random variable x, for which we obtain (that is measure, calculate) values X_n, dependent on an index n along a given dimension, or scale (e.g. temperature as a function of time, sea level along a satellite track, oxygen concentration as a function of depth etc).

Statistical theory often considers r.v. that are continuous, as in $X(t)$. Since we are typically dealing with digital or numerical data that are discrete, we will take a discrete approach in this course, as in $X_n = X(t_n) = X(n\Delta t)$ where Δt is the step of the record.

Sometimes, we will need to revert to continuous notations when needed; as an example:

$$\sum_{n=0}^{N} X_n \Delta t \Longleftrightarrow \int_0^T X(t)dt, \quad T = N\Delta t$$

Mean

The first statistical quantity to consider is the *true mean, population mean, or expectation* of x:

$$\mu_x \equiv E[X] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} X_n$$

Mean

The first statistical quantity to consider is the *true mean, population mean, or expectation* of x:

$$\mu_x \equiv E[X] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} X_n$$

Unfortunatelly, we only have a finite numbers of estimates $X_n,\ n = 1, \ldots, N$ so we compute the *sample mean*

$$\overline{X} \equiv \frac{1}{N} \sum_{n=1}^{N} X_n = \widehat{\mu}_x$$

where the last equality means that the **sample mean is an estimator of the true mean of** x.

Variance

The next statistical quantity of interest is the *variance* of x:

$$\sigma_x^2 \equiv E[(X - \mu_x)^2].$$

$\sigma_x = \sqrt{\sigma_x^2}$ is called the *standard deviation*.

Variance

The next statistical quantity of interest is the *variance* of x:

$$\sigma_x^2 \equiv E[(X - \mu_x)^2].$$

$\sigma_x = \sqrt{\sigma_x^2}$ is called the *standard deviation*.

An estimator of σ_x^2 is the *sample variance* s_x^2:

$$s_x^2 = \widehat{\sigma}_x^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \overline{X})^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(X_n - \frac{1}{N}\sum_{n=1}^{N} X_n\right)^2$$

Note the factor $\frac{1}{N-1}$ instead of $\frac{1}{N}$ in the previous expression; see section 4.1 of reference [1] for an explanation.

Mean

Let's go back to the estimator $\overline{X}$ of μ_x. It is an estimator, so is it biased? How much does it vary?

Mean

Let's go back to the estimator $\overline{X}$ of μ_x. It is an estimator, so is it biased? How much does it vary?

The expectation is an mathematical operator which is *linear*:

$$E[aX + bY] = aE[X] + bE[Y]$$

Mean

Let's go back to the estimator $\overline{X}$ of μ_x. It is an estimator, so is it biased? How much does it vary?

The expectation is an mathematical operator which is *linear*:

$$E[aX + bY] = aE[X] + bE[Y]$$

Let's use this rule for $\overline{X}$:

$$E[\overline{X}] = E\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] = \frac{1}{N}\sum_{n=1}^{N} E[X_n] = \frac{1}{N}(N\mu_x) = \mu_x$$

Note that $E[X_n] = \mu_x$ is a definition, valid for all n!

Mean

Let's go back to the estimator $\overline{X}$ of μ_x. It is an estimator, so is it biased? How much does it vary?

The expectation is an mathematical operator which is *linear*:

$$E[aX + bY] = aE[X] + bE[Y]$$

Let's use this rule for $\overline{X}$:

$$E[\overline{X}] = E\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] = \frac{1}{N}\sum_{n=1}^{N} E[X_n] = \frac{1}{N}(N\mu_x) = \mu_x$$

Note that $E[X_n] = \mu_x$ is a definition, valid for all n!

Since $E[\overline{X}] = \mu_x$ then it is said that $\overline{X}$ *is an unbiased estimator of* μ_x (see slide on bias). This means that the more observations of x we obtain, the more accurate the estimation of the mean will be.

Mean

Let's go back to the estimator $\overline{X}$ of μ_x. It is an estimator, so is it biased? How much does it vary?

The expectation is an mathematical operator which is *linear*:

$$E[aX + bY] = aE[X] + bE[Y]$$

Let's use this rule for $\overline{X}$:

$$E[\overline{X}] = E\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] = \frac{1}{N}\sum_{n=1}^{N} E[X_n] = \frac{1}{N}(N\mu_x) = \mu_x$$

Note that $E[X_n] = \mu_x$ is a definition, valid for all n!

Since $E[\overline{X}] = \mu_x$ then it is said that $\overline{X}$ *is an unbiased estimator of* μ_x (see slide on bias). This means that the more observations of x we obtain, the more accurate the estimation of the mean will be. That does not mean that $\overline{X}$ is free of errors ...

Mean

How much does the sample mean estimator vary? Recall the definition of the MSE of an estimator; for $\overline{X}$ it is

$$E[(\overline{X} - \mu_x)^2] = \mathrm{Var}[\overline{X}] + (b[\overline{X}])^2 = \mathrm{Var}[\overline{X}] + 0$$

Mean

How much does the sample mean estimator vary? Recall the definition of the MSE of an estimator; for $\overline{X}$ it is

$$E[(\overline{X} - \mu_x)^2] = \mathrm{Var}[\overline{X}] + (b[\overline{X}])^2 = \mathrm{Var}[\overline{X}] + 0$$

Under some assumption (that the X_n are *independent*), it can be shown (your homework, or section 4.1 of reference [1]) that

$$\mathrm{Var}[\overline{X}] = \frac{\sigma_x^2}{N} = \frac{\mathrm{Var}[X]}{N}$$

The **variance of the mean estimator** is the **true variance** of the data (σ_x^2) divided by the number of observation (N).

Mean

How much does the sample mean estimator vary? Recall the definition of the MSE of an estimator; for $\overline{X}$ it is

$$E[(\overline{X} - \mu_x)^2] = \mathrm{Var}[\overline{X}] + (b[\overline{X}])^2 = \mathrm{Var}[\overline{X}] + 0$$

Under some assumption (that the X_n are *independent*), it can be shown (your homework, or section 4.1 of reference [1]) that

$$\mathrm{Var}[\overline{X}] = \frac{\sigma_x^2}{N} = \frac{\mathrm{Var}[X]}{N}$$

The **variance of the mean estimator** is the **true variance** of the data (σ_x^2) divided by the number of observation (N). Since we typically do not know the true variance, we substitute for the sample variance to obtain the *standard error of the mean*, or *random error for the mean*:

$$\mathrm{s.e.}[\overline{X}] \equiv \sqrt{\mathrm{Var}[\overline{X}]} = \sqrt{\frac{\widehat{\sigma}_x^2}{N}} = \frac{s_x}{\sqrt{N}}$$

Mean

s.e.$[\overline{X}]$ is a measure of the uncertainty, or of our capability of estimating the mean value of x.

Mean

s.e.$[\overline{X}]$ is a measure of the uncertainty, or of our capability of estimating the mean value of x.

Example from *Beal, L. M. et al. (2015), Capturing the Transport Variability of a Western Boundary Jet: Results from the Agulhas Current Time-series experiment (ACT), J. Phys. Oceanogr., 45, 1302-1324,* *doi:10.1175/JPO-D-14-0119.1*

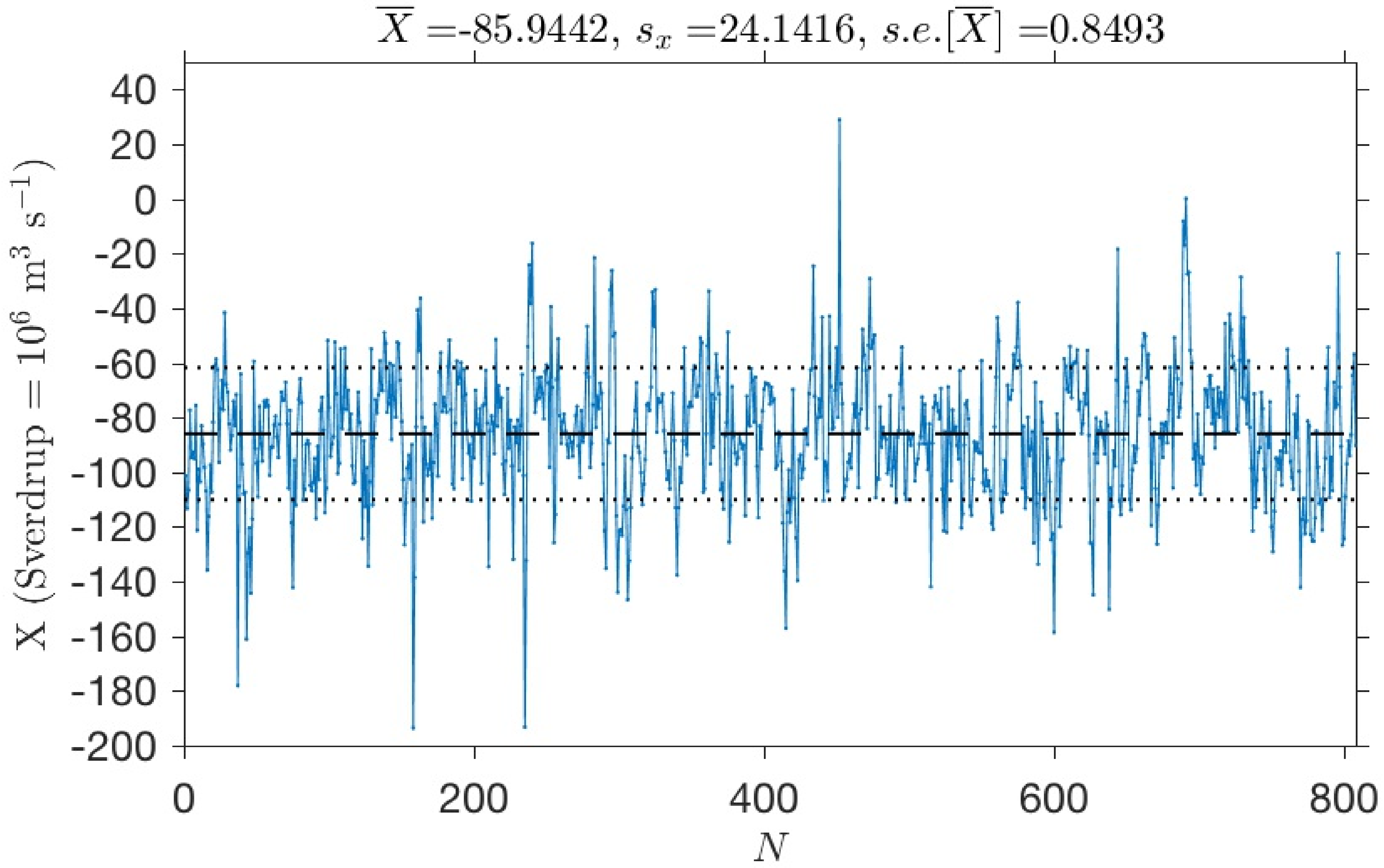
TABLE 3. Statistics (Sv) of time series for the western boundary jet transport T and the boundary layer transport T_{box}. Negative values are transport to the southwest with the Agulhas Current. Estimated errors are an upper bound and propagate from the derivation of geostrophic velocity from CPIES and from CM instrumental and sampling errors, as described in the appendix.

	T	T_{box}
Mean	−84	−77
Median	−79	−76
Standard deviation	24	32
Decorrelation time scale	7	17
Standard error of the mean	2	4
Estimated error (20 h)	14.8	6
Estimated error (mean)	9.0	0.5

errors, while for CPIES data these errors are similar (see the appendix). As seen above, observed differences between overlapping CPIES- and CM-derived transports (Table 2) show actual CPIES errors are likely 30% smaller than these estimates. Nevertheless, we combine these independent CM and CPIES errors and sum with the standard error (Kanzow et al. 2010) to estimate a mean and total error for the western boundary jet transport of −84 ± 11 Sv. This total error is an upper

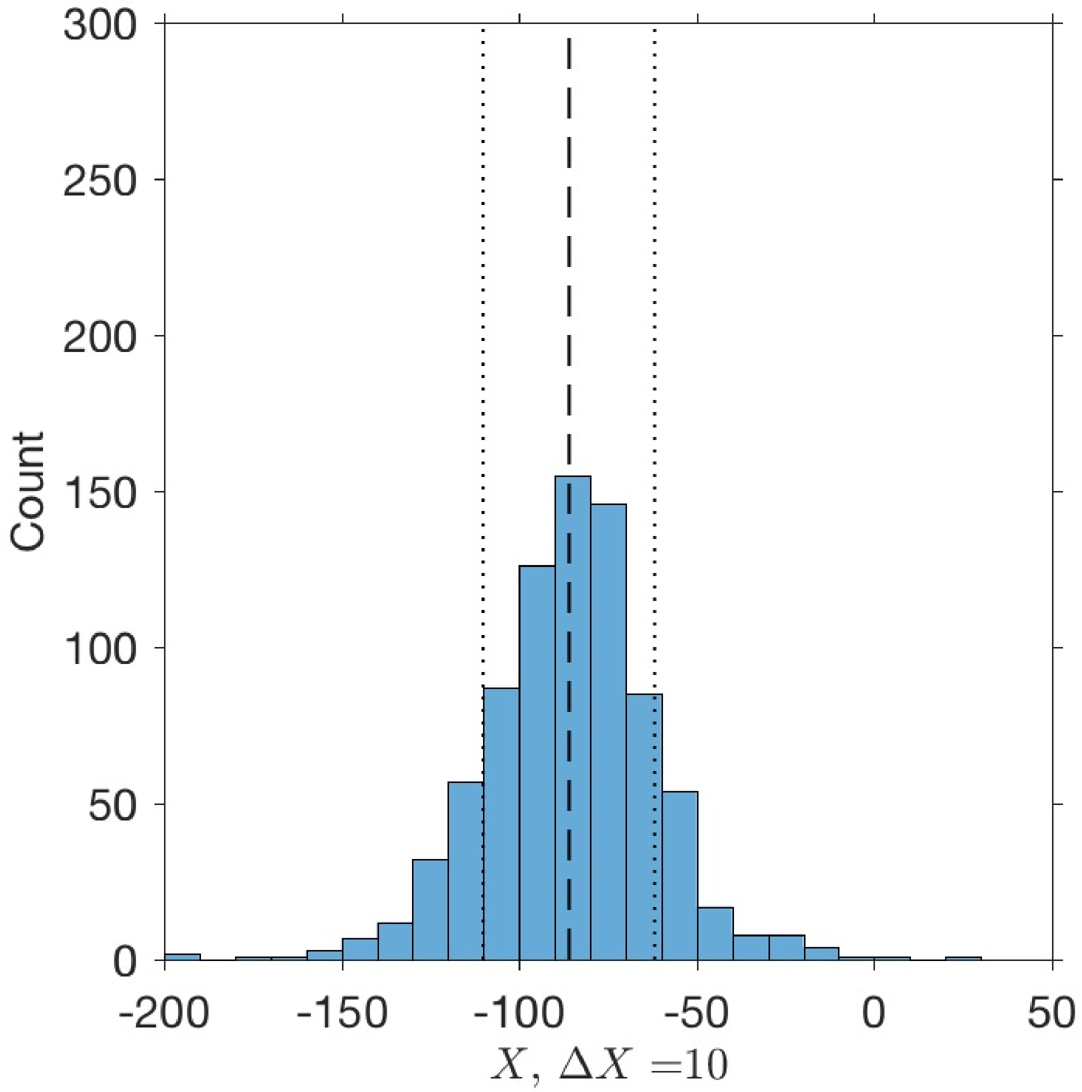
Example

Agulhas current boundary transport from *Beal, L. M. and S. Elipot, Broadening not strengthening of the Agulhas Current since the early 1990s, Nature, 540, 570573,* *doi:10.1038/nature19853*



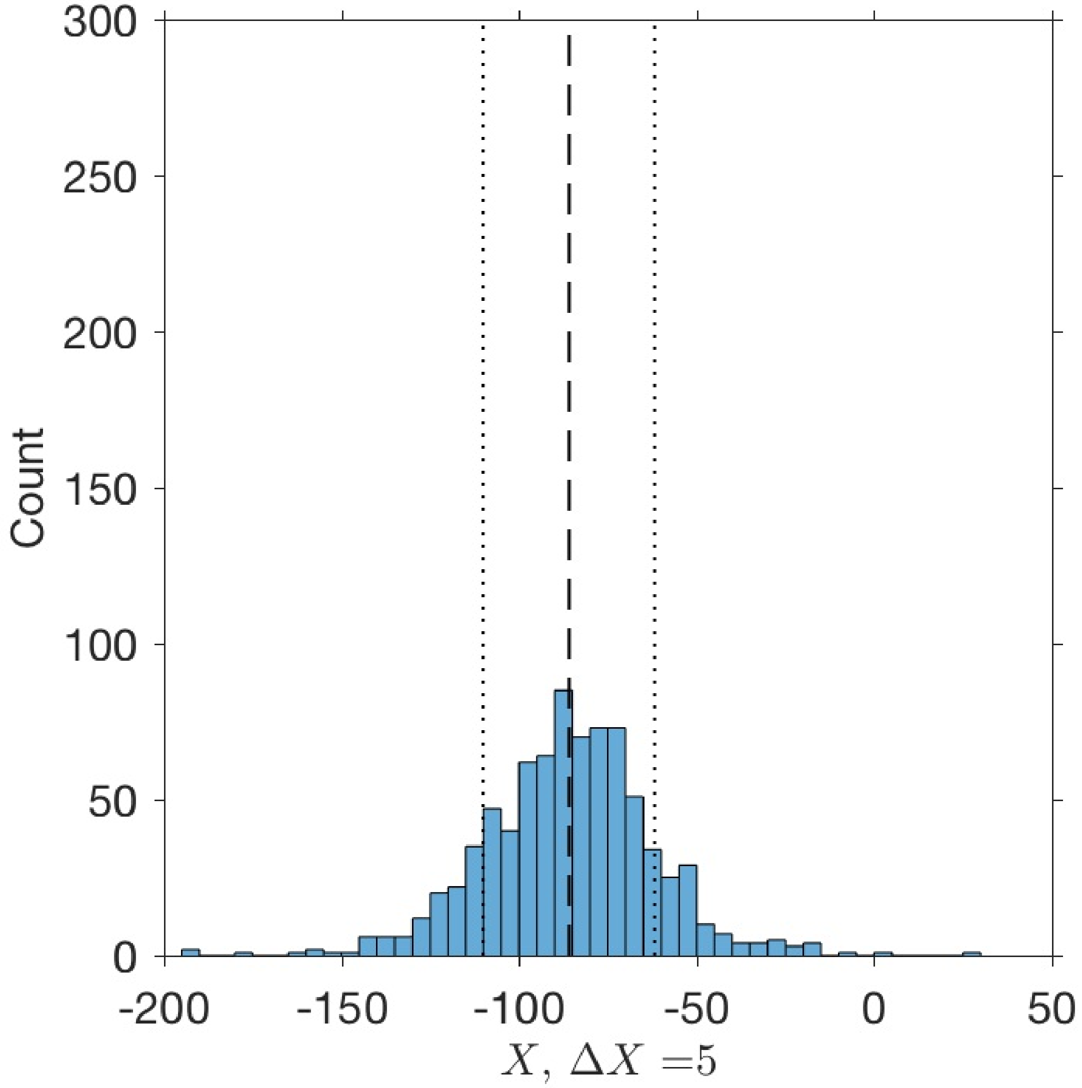
Depending on your data, your point of view, and your interests, the mean and the variance may tell you a lot, or little, about your data.

Depending on your data, your point of view, and your interests, the mean and the variance may tell you a lot, or little, about your data. Thus, you may want to look at the *frequency distribution* plot, or *histogram*, which is a count of your data values in a number of discrete intervals.



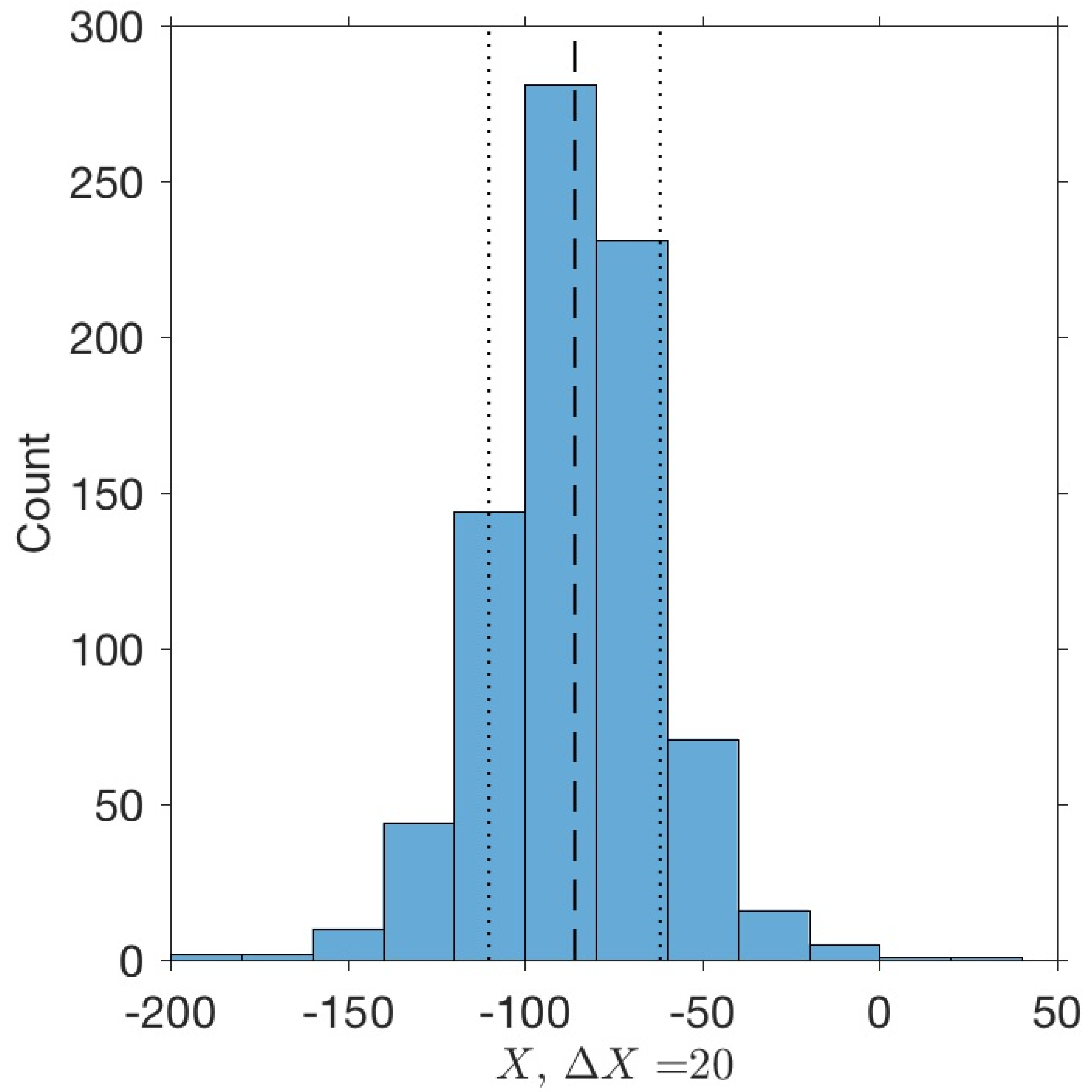
Histogram

There is no general rule (only recommendations) on how to choose the size of the bins or the number of bins to be used.



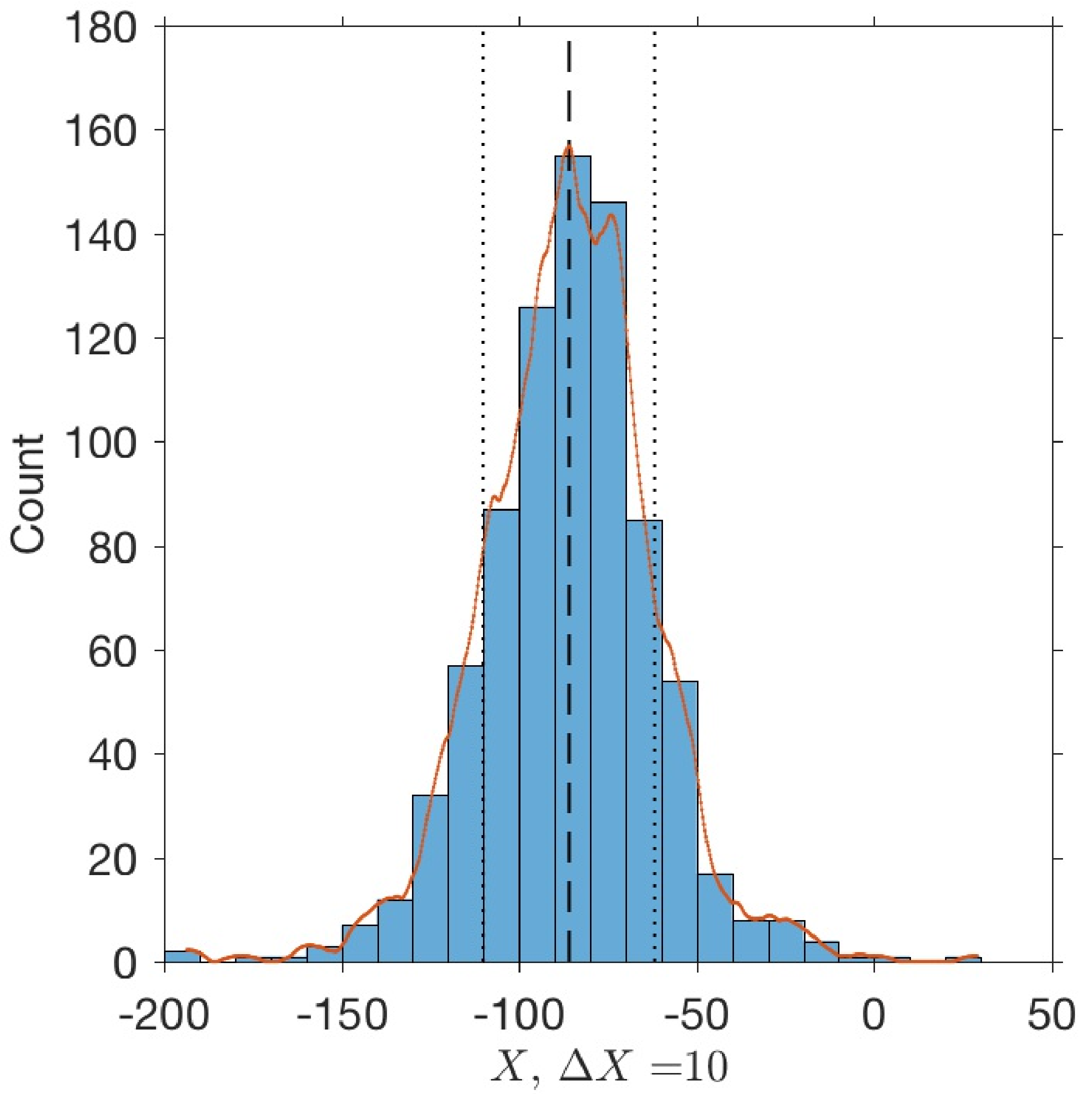
Histogram

There is no general rule (only recommendations) on how to choose the size of the bins, or the number of bins to be used.



Histogram

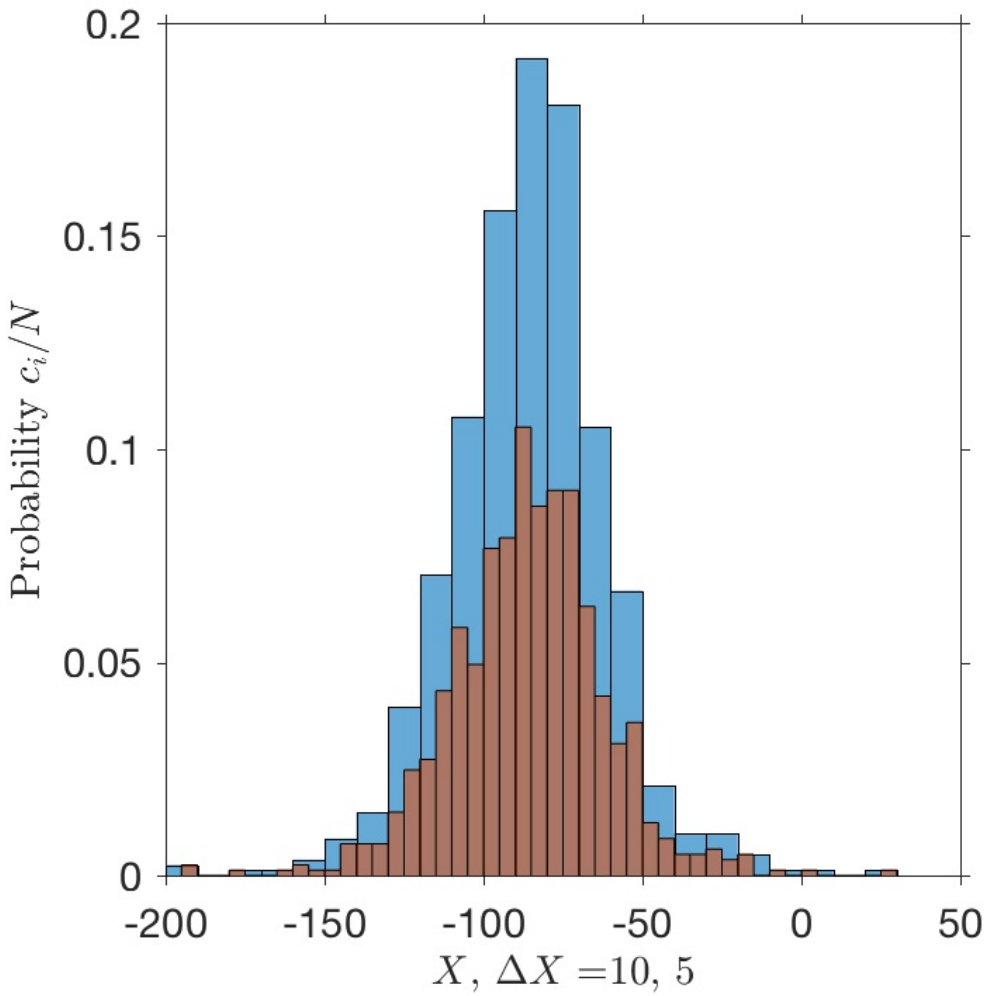
The frequency of occurences of a given value $x = a$ is quantity derived from x and can be itself estimated. The red line in this plot shows a *kernel estimate* of the histogram (see practical session this afternoon).



Probability function

Let's consider the probability, or relative count of occurences, to obtain a value X in the interval $[x_{k-1}, x_k]$. This defines a *probability function*:

$$P^*(x_{k-1} \leq X \leq x_k) = P_k^* = \frac{c_k}{N}, \quad c_k \text{ count in } [x_{k-1}, x_k]$$



$$\sum_k P_k^* = 1, \quad k\text{: interval index}$$

But the overall values of P_k^* still depend on the width ΔX of the bins. See this example with $\Delta X = 10$ and $\Delta X = 5$.

PDF and CFD

Let's consider instead the discrete *probability density function*, or *PDF*:

$$P(x_{k-1} \leq X \leq x_k) = P_k = \frac{c_k}{N\Delta X}, \quad \Delta X = x_k - x_{k-1}$$

PDF and CFD

Let's consider instead the discrete *probability density function*, or *PDF*:

$$P(x_{k-1} \le X \le x_k) = P_k = \frac{c_k}{N\Delta X}, \quad \Delta X = x_k - x_{k-1}$$

and the discrete *cumulative (probability) distribution function*, or *CDF*:

$$F(x_k) = P^*(x_0 \le X \le x_k) = P^*(X \le x_k) = \sum_{i \le k} P_i^*$$

PDF and CFD

Let's consider instead the discrete *probability density function*, or *PDF*:

$$P(x_{k-1} \leq X \leq x_k) = P_k = \frac{c_k}{N\Delta X}, \quad \Delta X = x_k - x_{k-1}$$

and the discrete *cumulative (probability) distribution function*, or *CDF*:

$$F(x_k) = P^*(x_0 \leq X \leq x_k) = P^*(X \leq x_k) = \sum_{i \leq k} P_i^*$$

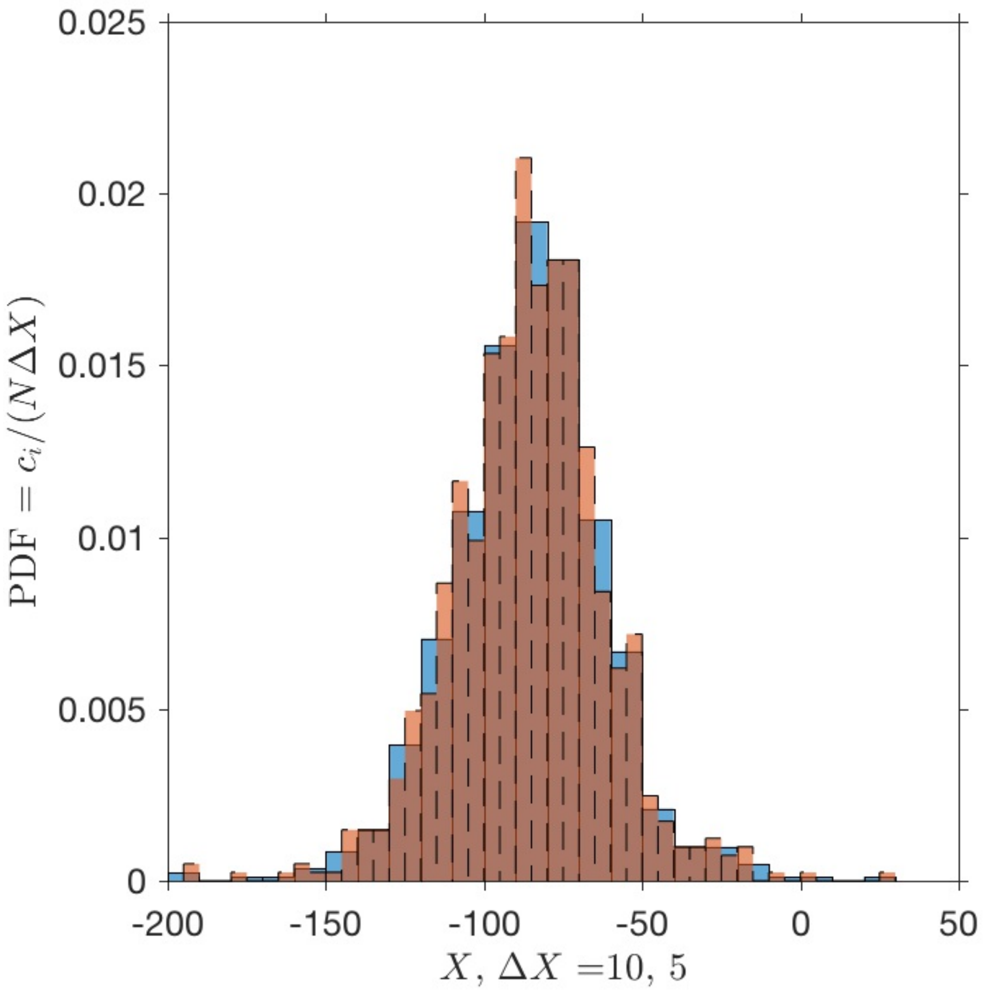
Since all the values X are contained between $\min[X]$ and $\max[X]$:

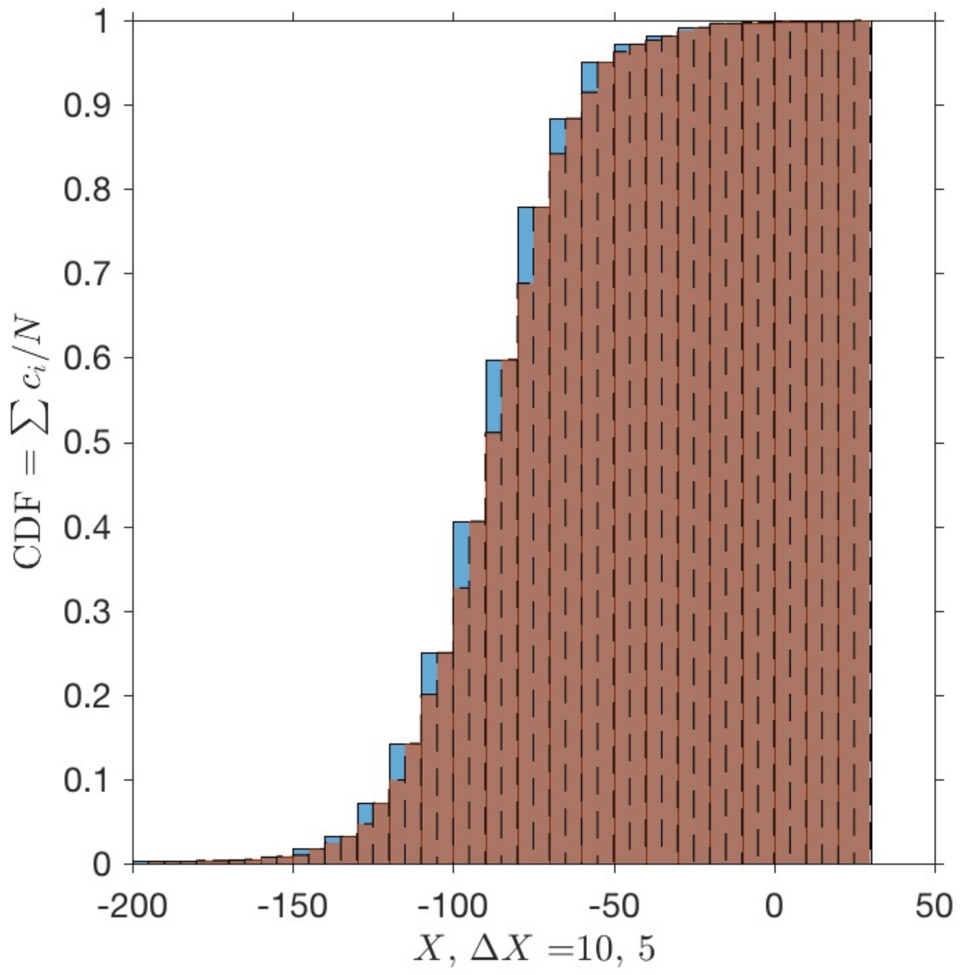
$$P(\min[X] \leq X \leq \max[X]) = 1 \quad \text{or} \quad \sum_k P_k \Delta X = \left(\frac{N}{N\Delta X}\right)\Delta X = 1$$

$$F(\max[X]) = 1$$

PDF and CFD

The overall values of the PDF and CDF do not depend on the bin width. However, their resolution (or detailed shapes) do.





PDF and CDF

As one reduces the size of the bins, $\Delta X \to 0$, the *continuous PDF* and *CDF* are approximated:

$$P(x \le X \le x + \Delta X) \longrightarrow p(x) = \frac{df(x)}{dx}$$

$$F(x) = P(X \le x)\Delta X \longrightarrow f(x) = \int_{-\infty}^{x} p(x')\, dx'$$

with the property

$$\sum_k P_k \Delta X = 1 \longrightarrow \int_{-\infty}^{+\infty} p(x)\, dx = f(+\infty) - f(-\infty) = 1 - 0 = 1$$

PDF and statistics

We can now give some formal definitions:

$$\mu_x \equiv \int_{-\infty}^{+\infty} xp(x)\,dx, \quad \sigma_x^2 \equiv \int_{-\infty}^{+\infty} (x-\mu_x)^2 p(x)\,dx$$
$$\mu_n \equiv \int_{-\infty}^{+\infty} (x-\mu_x)^n p(x)\,dx$$

The last expression defines $\mu_n(x)$ the n-th *central moment* of x.

PDF and statistics

We can now give some formal definitions:

$$\mu_x \equiv \int_{-\infty}^{+\infty} xp(x)\,dx, \quad \sigma_x^2 \equiv \int_{-\infty}^{+\infty} (x-\mu_x)^2 p(x)\,dx$$

$$\mu_n \equiv \int_{-\infty}^{+\infty} (x-\mu_x)^n p(x)\,dx$$

The last expression defines $\mu_n(x)$ the n-th *central moment* of x.

The discrete equivalent is

$$\mu_n \equiv E[(X-\mu_x)^n]$$

PDF and statistics

We can now give some formal definitions:

$$\mu_x \equiv \int_{-\infty}^{+\infty} x p(x)\, dx, \quad \sigma_x^2 \equiv \int_{-\infty}^{+\infty} (x-\mu_x)^2 p(x)\, dx$$
$$\mu_n \equiv \int_{-\infty}^{+\infty} (x-\mu_x)^n p(x)\, dx$$

The last expression defines $\mu_n(x)$ the n-th *central moment* of x.

The discrete equivalent is

$$\mu_n \equiv E[(X-\mu_x)^n]$$

The second central moment μ_2 is the variance, by definition. The mean is the first moment about the origin (0), that is $\mu_x = \int_{-\infty}^{+\infty} (x-0) p(x)\, dx$

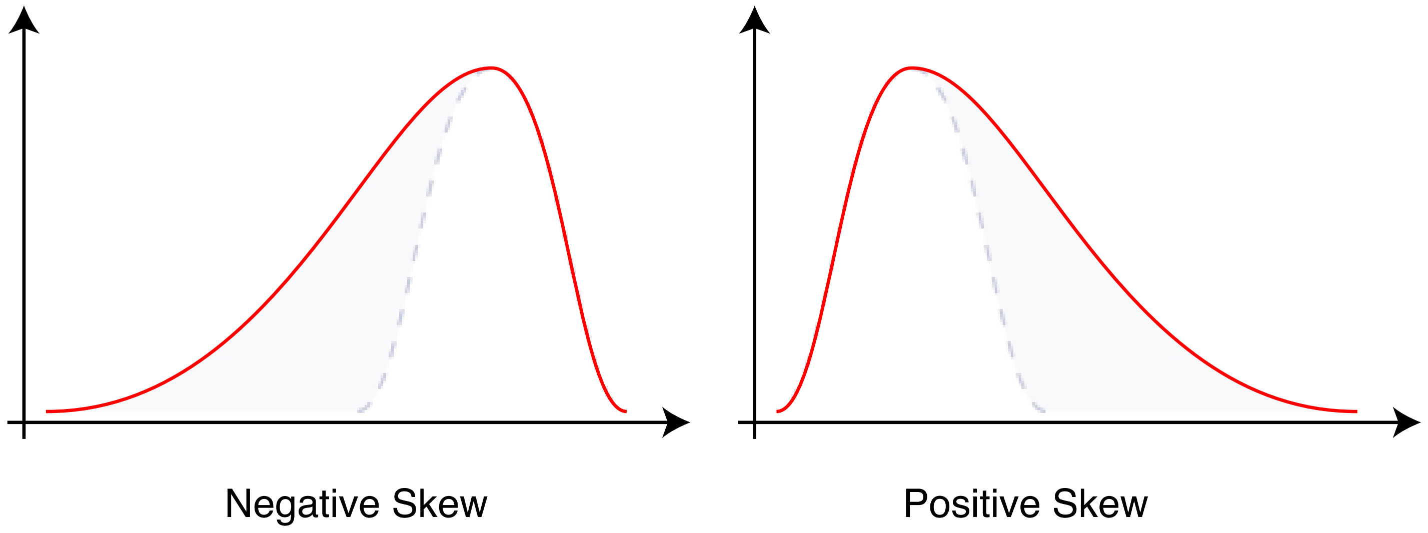
3rd moment: skewness

The third normalized central moment is called the *skewness*. It describes the tendency for an *asymmetry* between positive excursions and negative excursions of the PDF:

$$\gamma_x \equiv \frac{\mu_3}{(\mu_2)^{3/2}} = \frac{\mu_3}{(\sigma_x^2)^{3/2}}$$

One (biased) estimator is

$$\widehat{\gamma}_x \equiv \frac{\frac{1}{N}\sum_{n=1}^{N}(X_n - \overline{X})^3}{\left[\frac{1}{N}\sum_{n=1}^{N}(X_n - \overline{X})^2\right]^{3/2}}$$



4th moment: kurtosis

The fourth normalized central moment is called the *kurtosis*. It describes the *peakedness* (concentration near μ_x), or a tendency for *long tails* (concentration far from μ_x):

$$\kappa_x \equiv \frac{\mu_4}{(\mu_2)^2} = \frac{\mu_4}{(\sigma_x^2)^2}$$

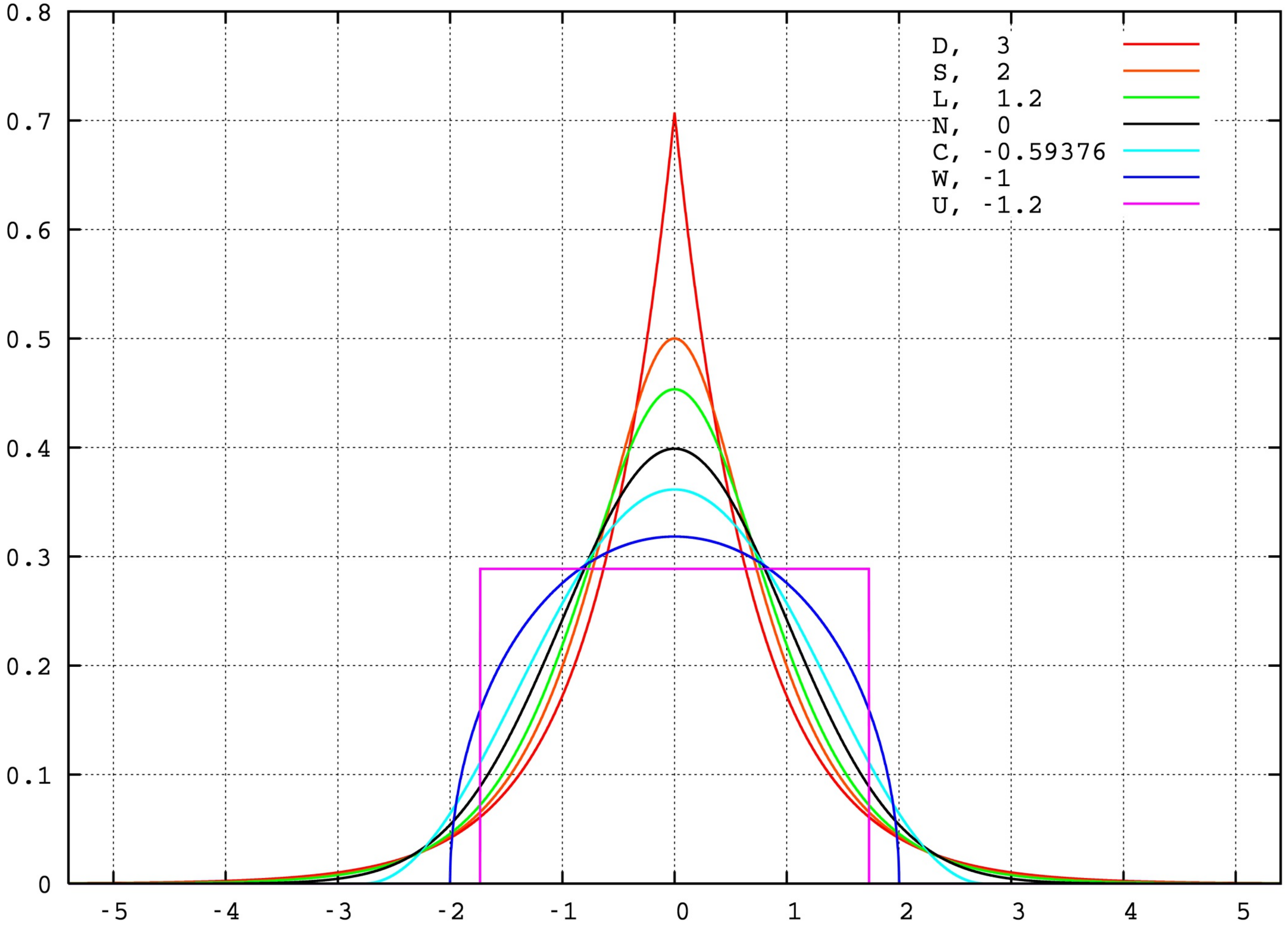
One (biased) estimator is

$$\widehat{\kappa}_x \equiv \frac{\frac{1}{N}\sum_{n=1}^{N}(X_n - \overline{X})^4}{\left[\frac{1}{N}\sum_{n=1}^{N}(X_n - \overline{X})^2\right]^{4/2}}$$

Because the kurtosis of a normal, or Gaussian, distribution is equal to 3, often the *excess kurtosis* $\kappa_x - 3$ is considered. *See Moors (1986), "The Meaning of Kurtosis: Darlington Reexamined".*

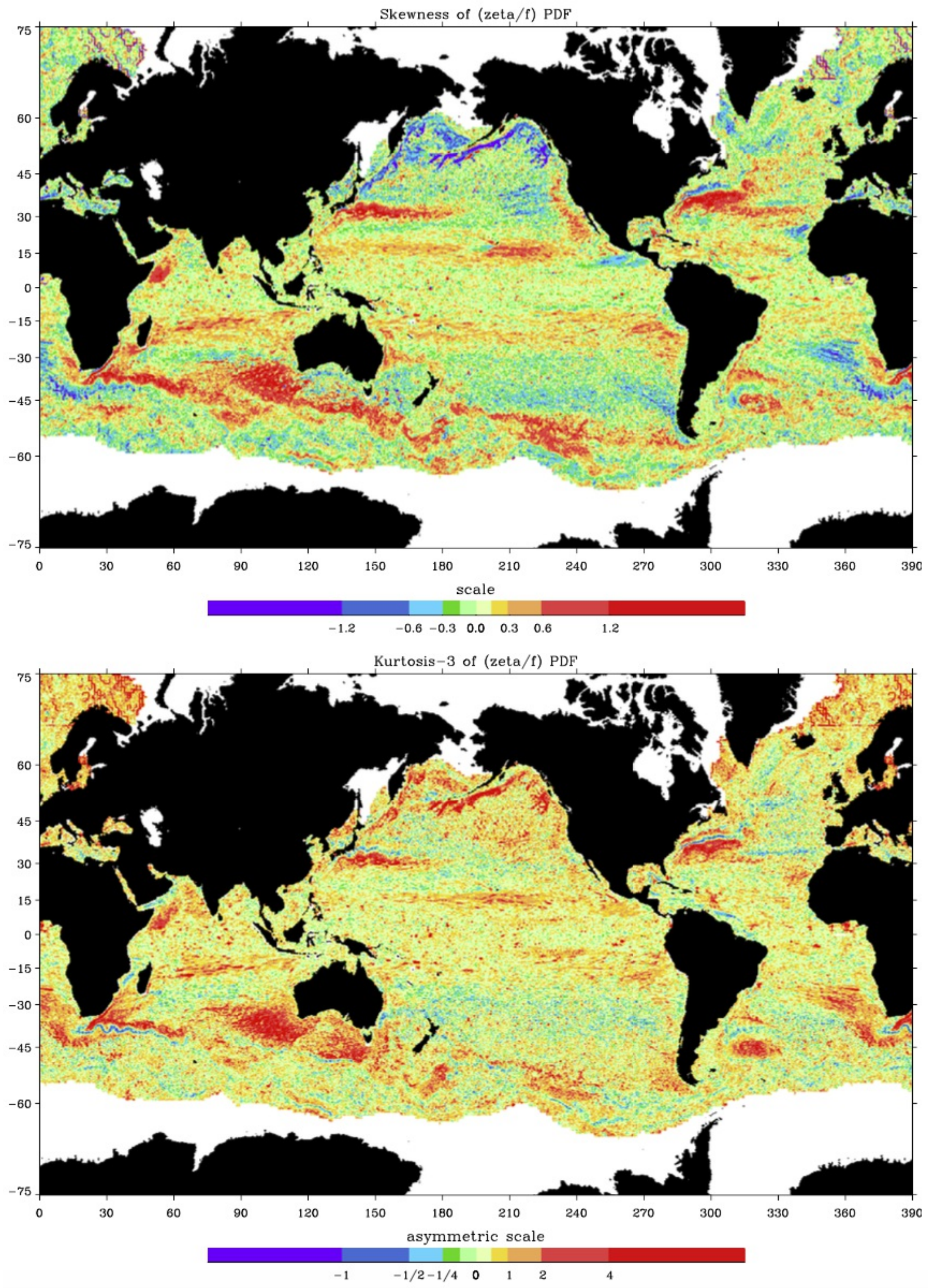
Illustration of Kurtosis

Distributions corresponding to different values of excess kurtosis.



Positive excess kurtosis corresponds to long tails and peakedness.

Kurtosis: example



Hughes et al. (2010) Identification of jets and mixing barriers from sea level and vorticity measurements using simple statistics

They used the statistics and PDF of sea level anomalies and derived relative vorticity to show that strong oceanic jets tend to be identified by a zero contour in skewness coinciding with a low value of kurtosis.

Why is this useful?

I think that plotting the histogram of your data, and further estimating in detail its PDF, gives you a holistic, or global view, of the *data population* from which your sample is drawn. Maybe you will find that the estimated PDF of a sample on a given day is different from the estimated PDF on another day ...

Why is this useful?

I think that plotting the histogram of your data, and further estimating in detail its PDF, gives you a holistic, or global view, of the *data population* from which your sample is drawn. Maybe you will find that the estimated PDF of a sample on a given day is different from the estimated PDF on another day ...

It also allows you to answer questions such as: what is the most probable value of the data? How often do we observe extreme values? etc.

Why is this useful?

I think that plotting the histogram of your data, and further estimating in detail its PDF, gives you a holistic, or global view, of the *data population* from which your sample is drawn. Maybe you will find that the estimated PDF of a sample on a given day is different from the estimated PDF on another day ...

It also allows you to answer questions such as: what is the most probable value of the data? How often do we observe extreme values? etc.

In addition, the knowledge of your data distribution, and/or a choice of a *model* for your data distribution will allow you to define *confidence intervals* for your estimated parameters, and to proceed to conduct *hypothesis testing* in your research.

Why is this useful?

I think that plotting the histogram of your data, and further estimating in detail its PDF, gives you a holistic, or global view, of the *data population* from which your sample is drawn. Maybe you will find that the estimated PDF of a sample on a given day is different from the estimated PDF on another day ...

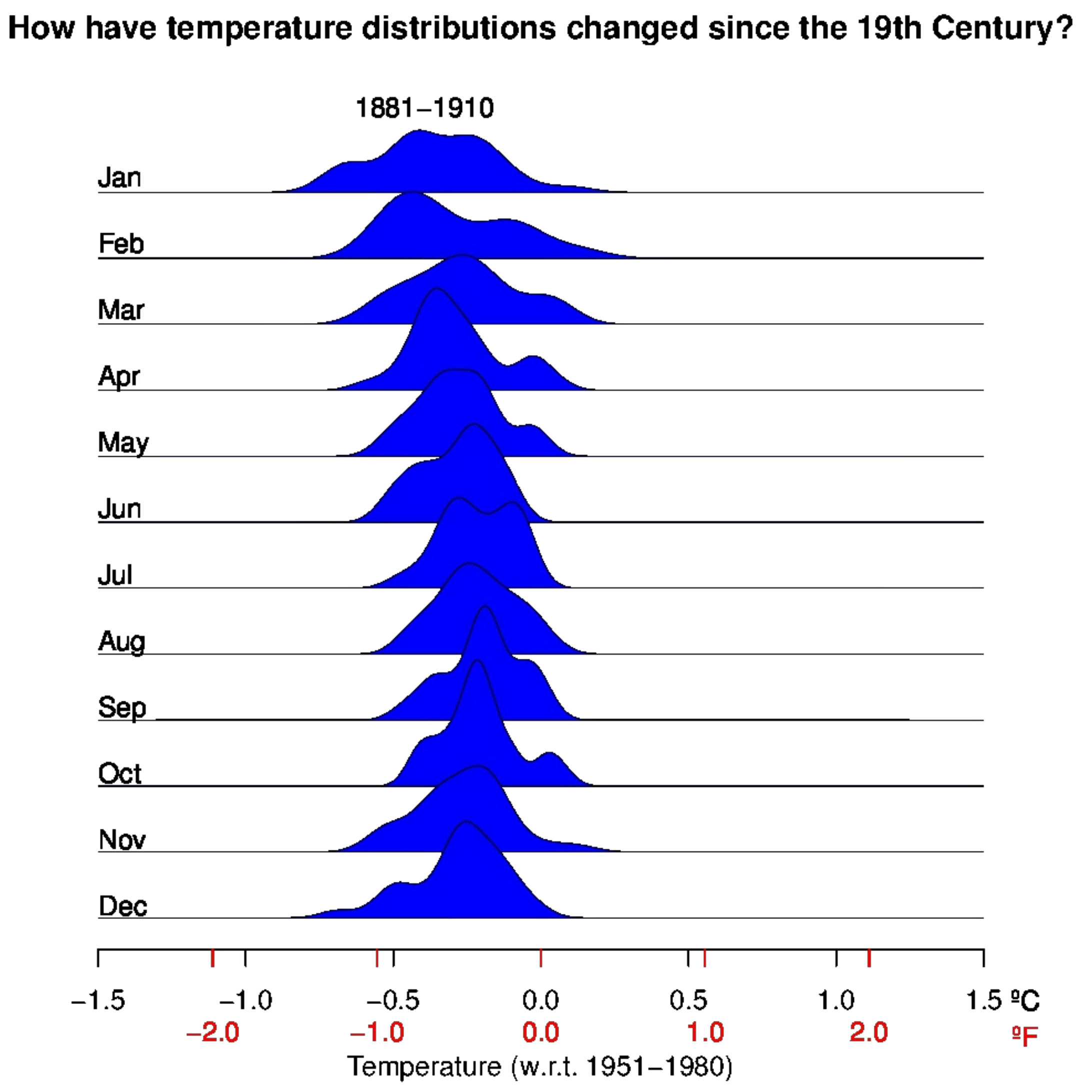
It also allows you to answer questions such as: what is the most probable value of the data? How often do we observe extreme values? etc.

In addition, the knowledge of your data distribution, and/or a choice of a *model* for your data distribution will allow you to define *confidence intervals* for your estimated parameters, and to proceed to conduct *hypothesis testing* in your research.

Before looking at this, we need to review the theory of various probability distribution functions.

But first an example

Animation (#joyplot) by Gavin Schmidt showing global temperature distribution in 10-yr windows, see his blog post on realclimate.org. Data from the NASA GISS Surface Temperature Analysis (GISTEMP) dataset.



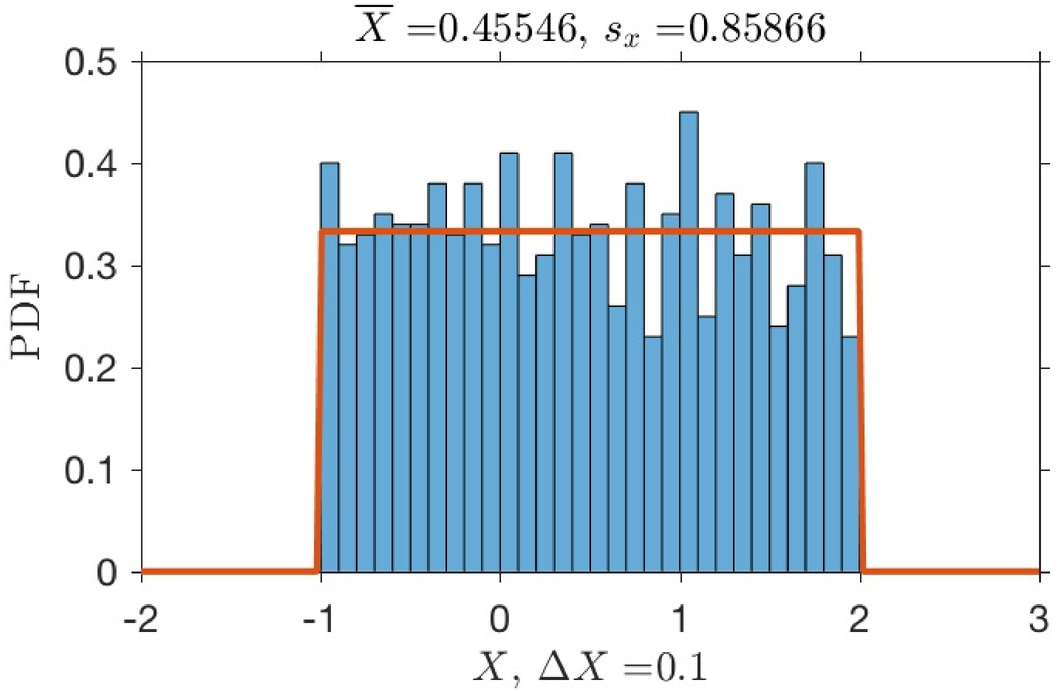
3. Common Distributions

The uniform distribution

A random variable x that is *uniformly distributed* between x_1 and x_2 has for PDF:

$$p(x) = \frac{1}{x_2 - x_1}, \quad x_1 \leq x \leq x_2$$
$$= 0, \quad \text{otherwise}$$

The mean of this distribution is $(x_1 + x_2)/2$ and its standard deviation is $(x_2 - x_1)/(2\sqrt{3})$



$$\mu_x = 0.5, \quad \sigma_x = \frac{2 - -1}{2\sqrt{(3)}} = 0.8660$$

```
z = rand(1000,1);x1 = -1; x2 = 2;
x = (x2-x1)*z + x1;
```

The normal distribution (or Gaussian)

A random variable x that is normally distributed with mean μ_x and standard deviation σ_x has for PDF:

$$p(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right] \equiv \mathcal{N}(\mu_x, \sigma_x)$$

The normal distribution (or Gaussian)

A random variable x that is normally distributed with mean μ_x and standard deviation σ_x has for PDF:

$$p(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right] \equiv \mathcal{N}(\mu_x, \sigma_x)$$

If $x \sim \mathcal{N}(\mu_x, \sigma_x)$ then the variable $z = \dfrac{x-\mu_x}{\sigma_x} \sim \mathcal{N}(0,1)$

Hereafter, $\sim$ will mean "distributed like".

The normal distribution (or Gaussian)

A random variable x that is normally distributed with mean μ_x and standard deviation σ_x has for PDF:

$$p(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right] \equiv \mathcal{N}(\mu_x, \sigma_x)$$

If $x \sim \mathcal{N}(\mu_x, \sigma_x)$ then the variable $z = \dfrac{x-\mu_x}{\sigma_x} \sim \mathcal{N}(0,1)$

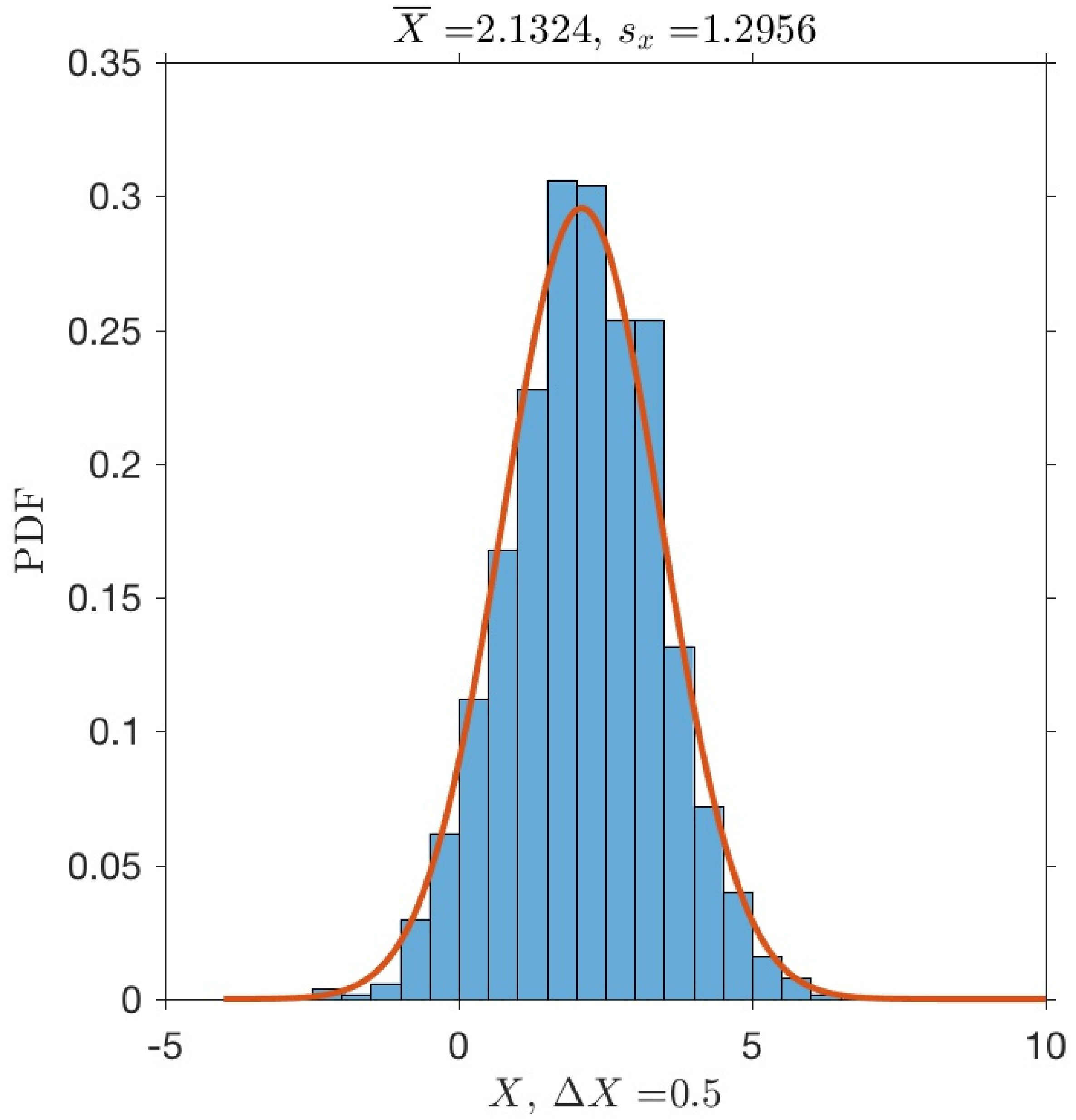
Hereafter, $\sim$ will mean "distributed like".

In Matlab, the following generates a data vector **z** containing 1000 samples from a r.v. $\sim \mathcal{N}(0,1)$, and a data vector **x** from a r.v. $\sim \mathcal{N}(2.1, 1.35)$

```
z = randn(1000,1);
x = 1.35*z + 2.1;
```

The normal distribution

The red curve is the theoretical normal PDF $\mathcal{N}(2.1, 1.35)$. The histogram is computed from a sample of size $N = 1000$.



The normal distribution

The normal distribution is of particular importance because of the *central limit theorem* which asserts roughly that the normal distribution is the result of the sum of a large number of independent random variable acting together.

The normal distribution

The normal distribution is of particular importance because of the *central limit theorem* which asserts roughly that the normal distribution is the result of the sum of a large number of independent random variable acting together.

To be more specific, let $x_1, x_2, \ldots, x_i, \ldots, x_N$ be N **independent** r.v. with individual means μ_i and variances σ_i^2. Now consider the new r.v.

$$x = a_1 x_1 + a_2 x_2 + \ldots + a_N x_N.$$

The central limit theorem states that, as $N \to +\infty$, x will be **normally** distributed with mean $\sum_k a_k \mu_k$ and variance $\sum_k a_k^2 \sigma_k^2$. In practice, the CLT is used for N "large".

The normal distribution

The normal distribution is of particular importance because of the *central limit theorem* which asserts roughly that the normal distribution is the result of the sum of a large number of independent random variable acting together.

To be more specific, let $x_1, x_2, \ldots, x_i, \ldots, x_N$ be N **independent** r.v. with individual means μ_i and variances σ_i^2. Now consider the new r.v.

$$x = a_1 x_1 + a_2 x_2 + \ldots + a_N x_N.$$

The central limit theorem states that, as $N \to +\infty$, x will be **normally** distributed with mean $\sum_k a_k \mu_k$ and variance $\sum_k a_k^2 \sigma_k^2$. In practice, the CLT is used for N "large".

In fact, we have already used the central limit theorem ...

The normal distribution and the CLT

Recall that the sample mean of the record $X_1, X_2, \ldots, X_N$ is defined as

$$\overline{X} = \frac{1}{N}\sum_{n=1}^{N} X_n = \left(\frac{1}{N}\right) X_1 + \left(\frac{1}{N}\right) X_2 + \ldots + \left(\frac{1}{N}\right) X_N$$

Here, $\overline{X}$ can be seen as a new r.v. which is the sum of individual r.v. (for which we have only one value) with the same population mean μ_x, and same population variance σ_x^2.

The normal distribution and the CLT

Recall that the sample mean of the record $X_1, X_2, \ldots, X_N$ is defined as

$$\overline{X} = \frac{1}{N}\sum_{n=1}^{N} X_n = \left(\frac{1}{N}\right) X_1 + \left(\frac{1}{N}\right) X_2 + \ldots + \left(\frac{1}{N}\right) X_N$$

Here, $\overline{X}$ can be seen as a new r.v. which is the sum of individual r.v. (for which we have only one value) with the same population mean μ_x, and same population variance σ_x^2.

Hence, the CLT states that, for N large enough, $\overline{X}$ is normally distributed with mean $\sum_{n=1}^{N}(1/N)\mu_x = (N/N)\mu_x = \mu_x$ and variance $\sum_{n=1}^{N}(1/N)^2\sigma_x^2 = (N/N^2)\sigma_x^2 = (1/N)\sigma_x^2$, where this last result was given previously without explanation.

The χ_n^2 distribution

Let $z_1, z_2, \ldots, z_n$ be n independent r.v. $\sim \mathcal{N}(0, 1)$. Let a new r.v. defined as

$$\chi^2 = z_1^2 + z_2^2 + \ldots + z_n^2$$

The χ_n^2 distribution

Let $z_1, z_2, \ldots, z_n$ be n independent r.v. $\sim \mathcal{N}(0,1)$. Let a new r.v. defined as

$$\chi^2 = z_1^2 + z_2^2 + \ldots + z_n^2$$

It is said that this r.v. is a "chi-squared" variable with n degrees of freedom (DOF).

The χ^2_n distribution

Let $z_1, z_2, \ldots, z_n$ be n independent r.v. $\sim \mathcal{N}(0,1)$. Let a new r.v. defined as

$$\chi^2 = z_1^2 + z_2^2 + \ldots + z_n^2$$

It is said that this r.v. is a "chi-squared" variable with n degrees of freedom (DOF). Such r.v. has for PDF:

$$p(x) = \frac{x^{\frac{n}{2}-1} \exp(-\frac{x}{2})}{2^{\frac{n}{2}} \Gamma(n/2)} \equiv \chi^2(n) \text{ or } \chi^2_n$$

The mean of this distribution is n and its variance is $2n$.

The χ_n^2 distribution

Let $z_1, z_2, \ldots, z_n$ be n independent r.v. $\sim \mathcal{N}(0,1)$. Let a new r.v. defined as

$$\chi^2 = z_1^2 + z_2^2 + \ldots + z_n^2$$

It is said that this r.v. is a "chi-squared" variable with n degrees of freedom (DOF). Such r.v. has for PDF:

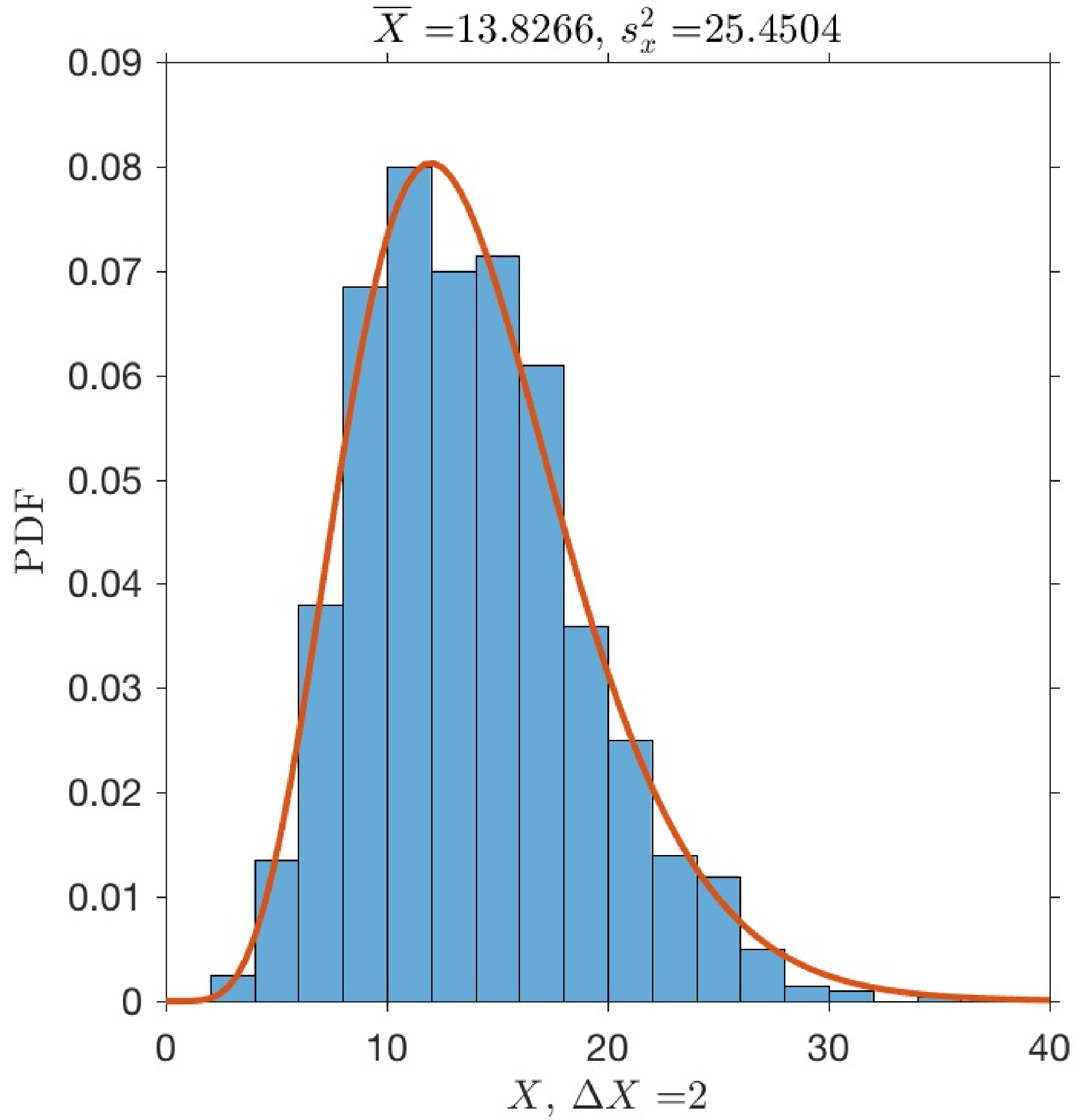
$$p(x) = \frac{x^{\frac{n}{2}-1}\exp(-\frac{x}{2})}{2^{\frac{n}{2}}\Gamma(n/2)} \equiv \chi^2(n) \text{ or } \chi_n^2$$

The mean of this distribution is n and its variance is $2n$.

Examples of χ^2 variables are power spectral density function estimates (see Lecture 4 on time series analysis) or variance estimates. The sample variance of $x \sim \mathcal{N}(\mu_x, \sigma_x)$ is $s_x^2 \sim \frac{\sigma_x^2}{N-1}\chi^2(N-1)$

The χ_n^2 distribution

In this example, the red curve is the theoretical χ_n^2 PDF for $n = 14$. The histogram is computed from a sample of size $N = 1000$.



The F distribution

If $x \sim \chi^2_n$ and $y \sim \chi^2_m$ then the following variable

$$\frac{x/n}{y/m} \sim F(n, m)$$

is F distributed with n and m degrees of freedom. The mathematical expression for this distribution is very complicated and not very useful here, see reference [1].

The mean value of the F distribution is $m/(m-2)$ for $m > 2$ and its variance is

$$\frac{2m^2(n+m-2)}{n(m-2)^2(m-4)} \text{ for } m > 4$$

The F distribution arises as an example when testing for the equality of two population variances (see practical this afternoon).

The Student's t distribution

Let y and z be two independent r.v. with $y \sim \chi^2_n$ and $z \sim \mathcal{N}(0,1)$.
Let be a new r.v. defined as

$$t = \frac{z}{\sqrt{\frac{y}{n}}}$$

It is said that this r.v. is a Student's t variable with n degrees of freedom (DOF).

The Student's t distribution

Let y and z be two independent r.v. with $y \sim \chi^2_n$ and $z \sim \mathcal{N}(0,1)$.
Let be a new r.v. defined as

$$t = \frac{z}{\sqrt{\frac{y}{n}}}$$

It is said that this r.v. is a Student's t variable with n degrees of freedom (DOF). Such r.v. has for PDF:

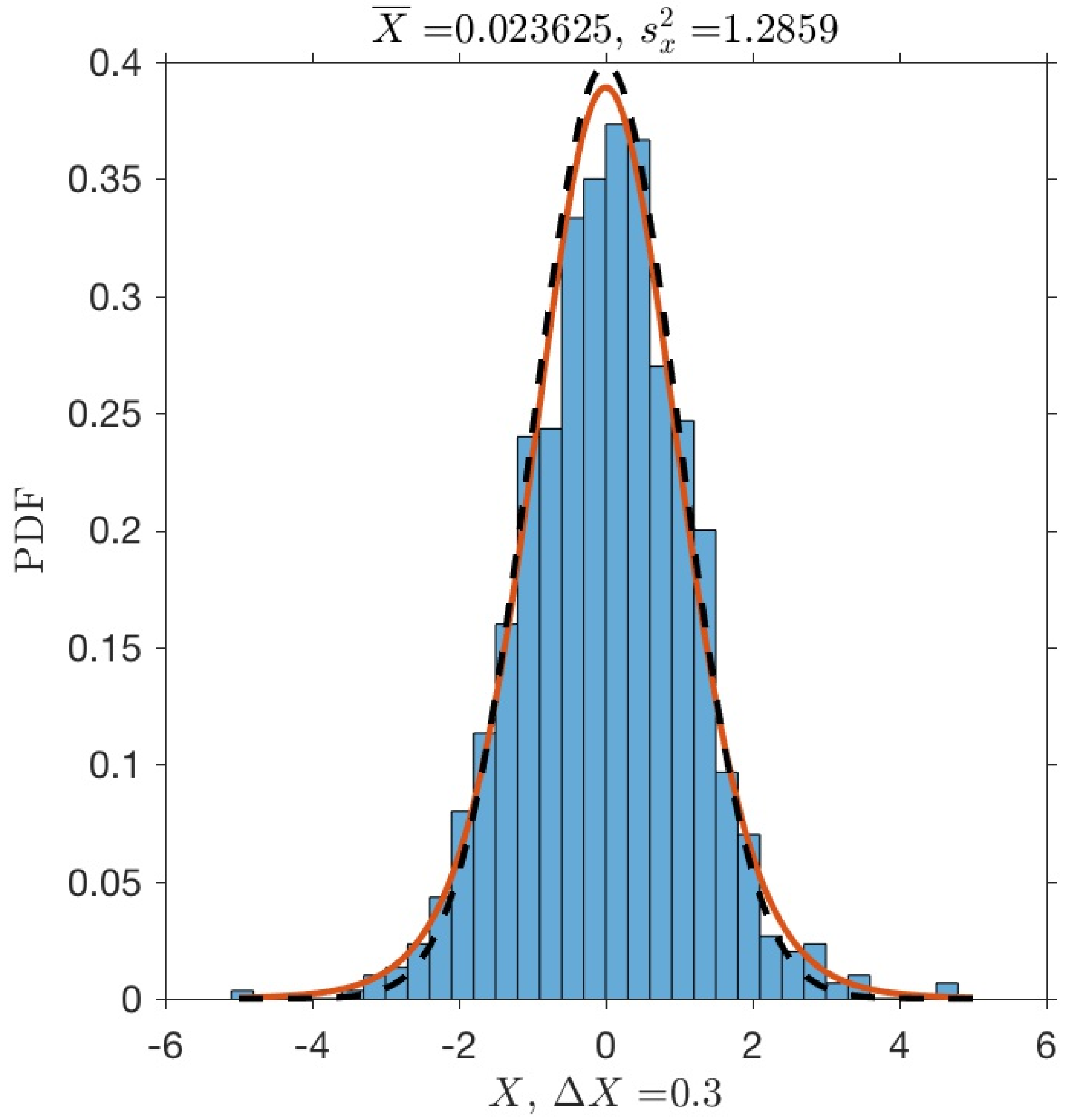
$$p(x) = \frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\Gamma(n/2)} \left[1 + \frac{x^2}{n}\right]^{\frac{n+1}{2}} \equiv t(n) \text{ or } t_n$$

The mean is 0 for $n > 0$ and the variance is $\frac{n}{n-2}$ for $n > 2$.

An example of t distributed r.v. is the estimate of the mean of a population with unknown variance, as we will see later.

The Student's t distribution

In this example, the red curve is the theoretical t_n PDF for $n = 10$. The histogram is computed from a sample of size $N = 1000$. The black curve is the theoretical PDF $\mathcal{N}(0, 1)$.



The Gamma (Γ) distribution family

In fact, the χ^2_n distribution and the exponential distribution are two particular cases of the Gamma (Γ) distribution family. A random variable x that is *Gamma distributed* with parameters α and β has for PDF:

$$p(x) = \frac{x^{\alpha-1}\exp[-x/\beta]}{\beta^\alpha \Gamma(\alpha)}, \quad \beta > 0;\, 0 \leq x \leq +\infty$$

with the Γ function defined as

$$\begin{aligned}
\Gamma(\alpha) &= \int_0^{+\infty} x'^{\alpha-1}\exp(-x')\,dx' \\
\Gamma(n) &= (n-1)! \quad \text{for } n \text{ integer} \\
\Gamma(\alpha) &= (\alpha-1)\Gamma(\alpha-1) \text{ for } \alpha \text{ continuous with } \Gamma(1) = 1
\end{aligned}$$

$\alpha = 1 \longrightarrow$ exponential distribution

$\alpha = \frac{n}{2}, \quad \beta = 2 \longrightarrow \chi^2_n$ distribution

4. Uncertainties, errors and hypothesis testing

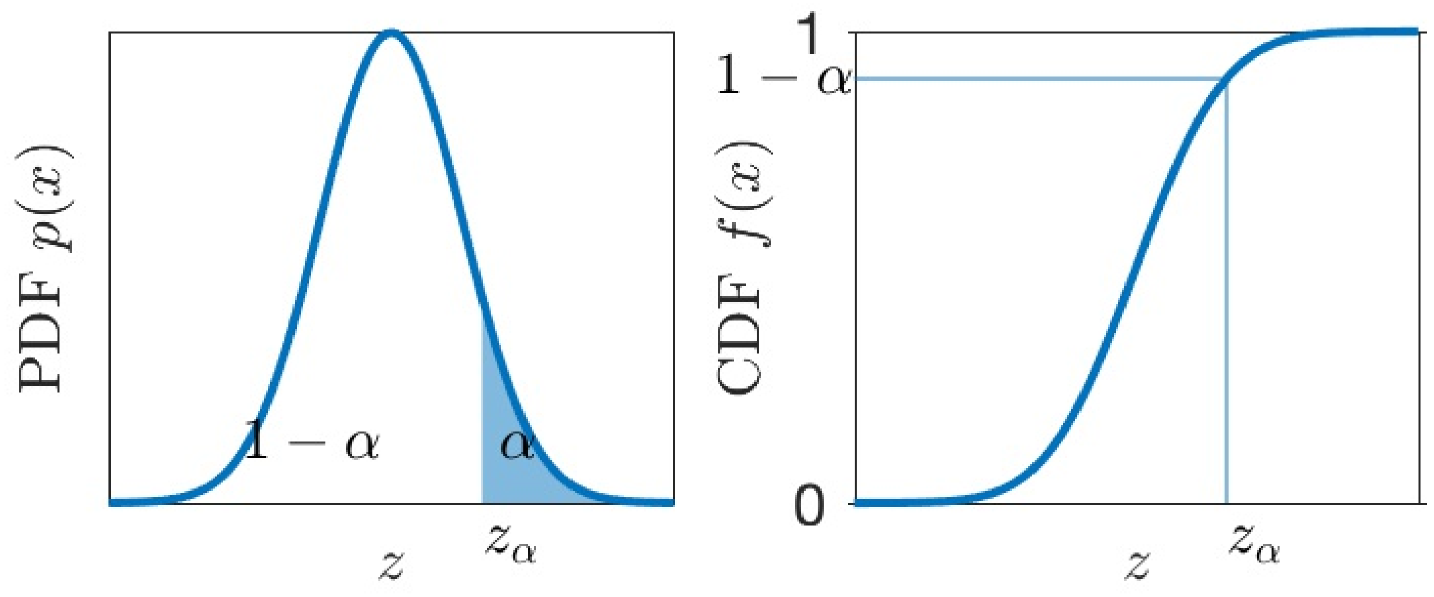
Probability statements

We use distributions to make probability statements about our r.v. estimates.

Probability statements

We use distributions to make probability statements about our r.v. estimates. It is useful to consider the following. For any given PDF $p(z)$, and associated CDF $f(z)$, of the variable z, let's denote z_α the value that corresponds to $f(z) = 1 - \alpha$, that is

$$f(z_\alpha) = \int_{-\infty}^{z_\alpha} p(z)\, dz = \text{Prob}[z \leq z_\alpha] = 1 - \alpha$$



BEWARE that this the convention used here. It is notably different in Matlab where icdf('normal',1-α,0,1) returns the value z_α

Confidence intervals

PDFs are used to derive confidence intervals (CI), the interpretation of which is subtle. What is a CI for you?

Confidence intervals

PDFs are used to derive confidence intervals (CI), the interpretation of which is subtle. What is a CI for you?

Given an estimate $\widehat{\phi}$ of a quantity ϕ, and a chosen significance level α, we construct an interval with lower bound ϕ_L and upper bound ϕ_U so that this interval is expected to cover the true, unknown, but fixed value of ϕ, with probability $1 - \alpha$.

Confidence intervals

PDFs are used to derive confidence intervals (CI), the interpretation of which is subtle. What is a CI for you?

Given an estimate $\widehat{\phi}$ of a quantity ϕ, and a chosen significance level α, we construct an interval with lower bound ϕ_L and upper bound ϕ_U so that this interval is expected to cover the true, unknown, but fixed value of ϕ, with probability $1 - \alpha$.

In other words, if we could repeat the estimation and calculation of the CI many times, we can expect that the true unknown parameter ϕ is covered by the calculated CI, 95 out of 100 times.

Confidence intervals

PDFs are used to derive confidence intervals (CI), the interpretation of which is subtle. What is a CI for you?

Given an estimate $\widehat{\phi}$ of a quantity ϕ, and a chosen significance level α, we construct an interval with lower bound ϕ_L and upper bound ϕ_U so that this interval is expected to cover the true, unknown, but fixed value of ϕ, with probability $1-\alpha$.

In other words, if we could repeat the estimation and calculation of the CI many times, we can expect that the true unknown parameter ϕ is covered by the calculated CI, 95 out of 100 times.

There is no probability statement about ϕ, only about $\widehat{\phi}$ and $[\phi_L, \phi_U]$.

Confidence intervals

As a concrete example, consider the sample mean $\overline{X}$ of $x \sim \mathcal{N}(\mu_x, \sigma_x)$.

We stated earlier that $\overline{X} \sim \mathcal{N}(\mu_x, \sigma_x/\sqrt{N})$. As such, we can state that the new "transformed" variable

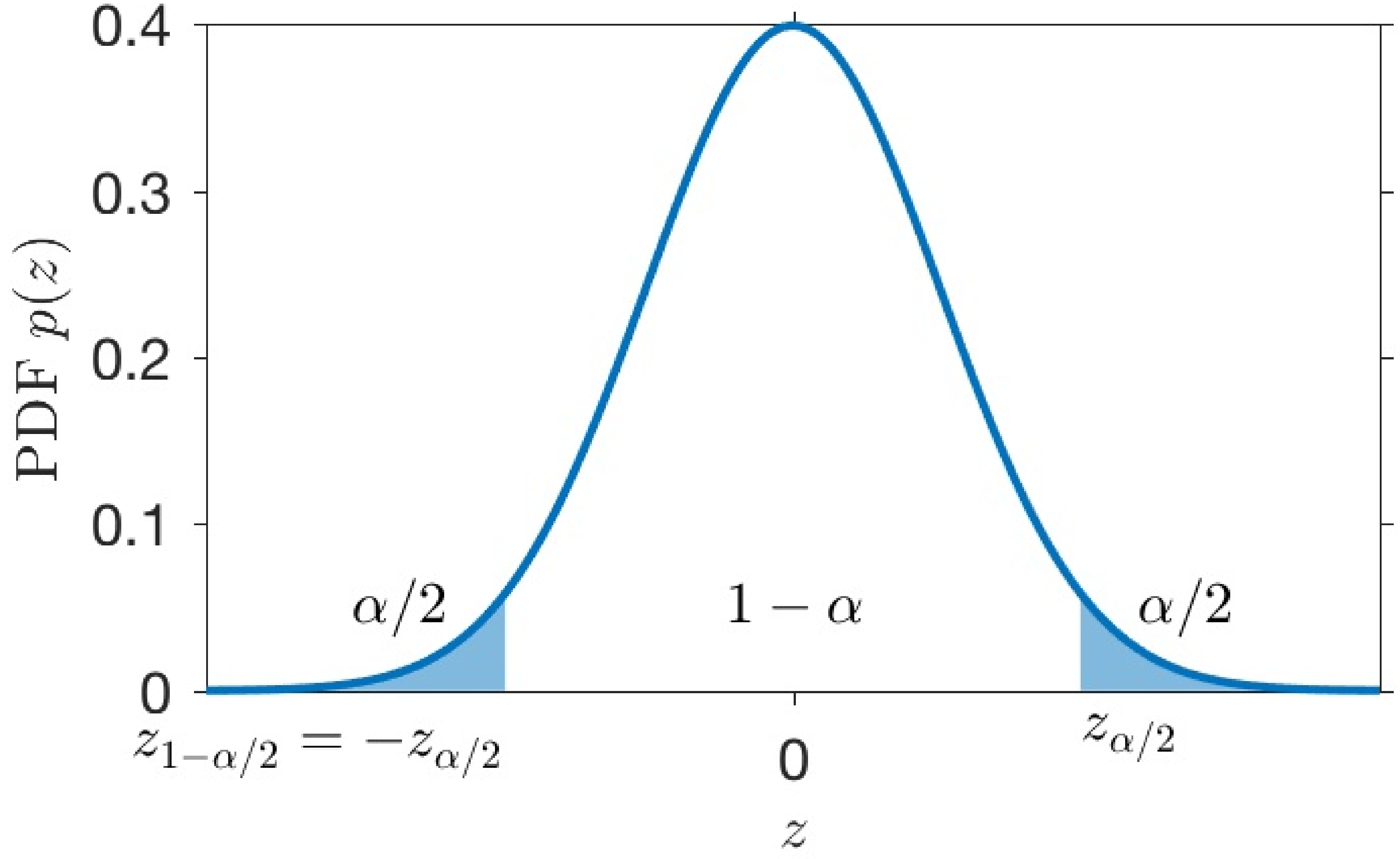
$$z = \frac{\overline{X} - \mu_x}{\sigma_x/\sqrt{N}} \sim \mathcal{N}(0,1)$$

and that we can find two z values such that

$$\text{Prob}\left[z_{1-\alpha/2} < \frac{\overline{X} - \mu_x}{\sigma_x/\sqrt{N}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

Confidence intervals: normal case

$$\text{Prob}\left[z_{1-\alpha/2} < \frac{\overline{X} - \mu_x}{\sigma_x/\sqrt{N}} \leq z_{\alpha/2}\right] = 1 - \alpha$$



Since, the normal distribution is symmetric around zero,

$$z_{1-\alpha/2} = -z_{\alpha/2}$$

Confidence intervals: normal case

As a result, the normalized calculated variable z is such that

$$-z_{\alpha/2} < z = \frac{\overline{X} - \mu_x}{\sigma_x/\sqrt{N}} \leq z_{\alpha/2}$$

with $1 - \alpha$ probability. After rearranging, we can state that the true mean μ_x of the r.v. x is such that

$$\overline{X} - \frac{\sigma_x z_{\alpha/2}}{\sqrt{N}} \leq \mu_x < \overline{X} + \frac{\sigma_x z_{\alpha/2}}{\sqrt{N}}$$

with a *confidence of* $100(1 - \alpha)\%$.

In common parlance, a 95% CI for μ_x is

$$\left[\overline{X} - \frac{\sigma_x z_{\alpha/2}}{\sqrt{N}}, \overline{X} + \frac{\sigma_x z_{\alpha/2}}{\sqrt{N}}\right]$$

Confidence intervals: normal case

Typical intervals used are:

$$90\%\ \text{CI}: \quad \alpha = 0.1 \longrightarrow z_{\alpha/2} = 1.6449$$

$$95\%\ \text{CI}: \quad \alpha = 0.05 \longrightarrow z_{\alpha/2} = 1.9600$$

$$99\%\ \text{CI}: \quad \alpha = 0.01 \longrightarrow z_{\alpha/2} = 2.5758$$

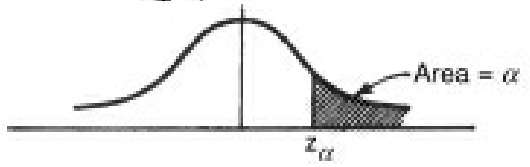
Before the advent of advanced softwares, people relied on statistical tables for the values of z, such as the ones found in the Appendices of references [1],[2],[3],[6].

Confidence intervals: normal case

Example from Bendat and Piersol (2011):

$$95\%\ \mathrm{CI}: \quad \alpha = 0.05 \longrightarrow \alpha/2 = 0.0250$$

$$\text{Value of } \alpha = \int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \mathrm{Prob}[z > z_\alpha]$$



z_α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0539
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Confidence intervals: normal case; example:

Using a CTD record of temperature at 24 Hz, we estimate the mean temperature near 70 db pressure level by averaging data points within .5 db of 70 db (falling rate is 1-2 m/s), $N = 11$. We find $\overline{T} = 19.21359$. From the specification sheet of the CTD 911 Plus, the accuracy of the temperature sensor is 0.001, which we interpret as being the random error or std of T, i.e. σ_T, ignoring a possible bias.

Confidence intervals: normal case; example:

Using a CTD record of temperature at 24 Hz, we estimate the mean temperature near 70 db pressure level by averaging data points within .5 db of 70 db (falling rate is 1-2 m/s), $N = 11$. We find $\overline{T} = 19.21359$. From the specification sheet of the CTD 911 Plus, the accuracy of the temperature sensor is 0.001, which we interpret as being the random error or std of T, i.e. σ_T, ignoring a possible bias. We use $(\overline{T} - \mu_T)/(\sigma_T/\sqrt{N}) \sim \mathcal{N}(0,1)$, and based on the previous formula, the 95% CI for the true mean μ_T is:

$$19.21359 - \frac{0.001 \times 1.96}{\sqrt{11}} \leq \mu_T < 19.21359 + \frac{0.001 \times 1.96}{\sqrt{11}}$$

$$\Rightarrow 19.21300 \leq \mu_T < 19.21418$$

Alternatively, once can state that the estimate of the mean with 95% uncertainty is

$$\mu_T = 19.21359 \pm 0.00059$$

Confidence intervals: t case

Now imagine that you obtain the data from the previous example but do not know the specification of the sensor used, that is σ_T. Instead, you can consider the sample standard deviation s_T as an estimate of the unknown σ_T.

Confidence intervals: t case

Now imagine that you obtain the data from the previous example but do not know the specification of the sensor used, that is σ_T. Instead, you can consider the sample standard deviation s_T as an estimate of the unknown σ_T. It can be shown (not obvious), that

$$\frac{\overline{T} - \mu_T}{s_T/\sqrt{N}} \sim t(N-1)$$

Confidence intervals: t case

Now imagine that you obtain the data from the previous example but do not know the specification of the sensor used, that is σ_T. Instead, you can consider the sample standard deviation s_T as an estimate of the unknown σ_T. It can be shown (not obvious), that

$$\frac{\overline{T} - \mu_T}{s_T/\sqrt{N}} \sim t(N-1)$$

Thus, a $100(1-\alpha)\%$ CI for the true mean μ_T is:

$$\left[\overline{T} - \frac{s_T t_{N-1;\alpha/2}}{\sqrt{N}} \leq \mu_T < \overline{T} + \frac{s_T t_{N-1;\alpha/2}}{\sqrt{N}}\right]$$

where $t_{N-1;\alpha/2}$ is the value of the t_{N-1} variable such that

$$\text{Prob}\left[t \leq t_{N-1;\alpha/2}\right] = 1 - \frac{\alpha}{2} \quad \text{or} \quad \text{Prob}\left[t > t_{N-1;\alpha/2}\right] = \frac{\alpha}{2}$$

The t distribution is symmetric like the normal distribution so that $t_{N;1-\beta} = -t_{N;\beta}$

Confidence intervals: t case

Going back to our CTD example, we still have $\overline{T} = 19.21359$ but now calculate $s_T = 0.02277$. Since $t_{40-1;0.05/2} = 2.0227$, the 95% CI for μ_T becomes:

$$19.21359 - \frac{0.02277 \times 2.0227}{\sqrt{1}} \leq \mu_T < 19.21359 + \frac{0.02277 \times 2.0227}{\sqrt{11}}$$

$$\Rightarrow 19.19829 \leq \mu_T < 19.22889$$

Alternatively once can state that the estimate of the mean with 95% uncertainty is

$$\mu_T = 19.21359 \pm 0.01530$$

Confidence intervals: t case

Going back to our CTD example, we still have $\overline{T} = 19.21359$ but now calculate $s_T = 0.02277$. Since $t_{40-1;0.05/2} = 2.0227$, the 95% CI for μ_T becomes:

$$19.21359 - \frac{0.02277 \times 2.0227}{\sqrt{1}} \leq \mu_T < 19.21359 + \frac{0.02277 \times 2.0227}{\sqrt{11}}$$

$$\Rightarrow 19.19829 \leq \mu_T < 19.22889$$

Alternatively once can state that the estimate of the mean with 95% uncertainty is

$$\mu_T = 19.21359 \pm 0.01530$$

Previously, we stated that

$$\mu_T = 19.21359 \pm 0.00059$$

So what is the right answer?

Confidence intervals: t case

Going back to our CTD example, we still have $\overline{T} = 19.21359$ but now calculate $s_T = 0.02277$. Since $t_{40-1;0.05/2} = 2.0227$, the 95% CI for μ_T becomes:

$$19.21359 - \frac{0.02277 \times 2.0227}{\sqrt{1}} \leq \mu_T < 19.21359 + \frac{0.02277 \times 2.0227}{\sqrt{11}}$$

$$\Rightarrow 19.19829 \leq \mu_T < 19.22889$$

Alternatively once can state that the estimate of the mean with 95% uncertainty is

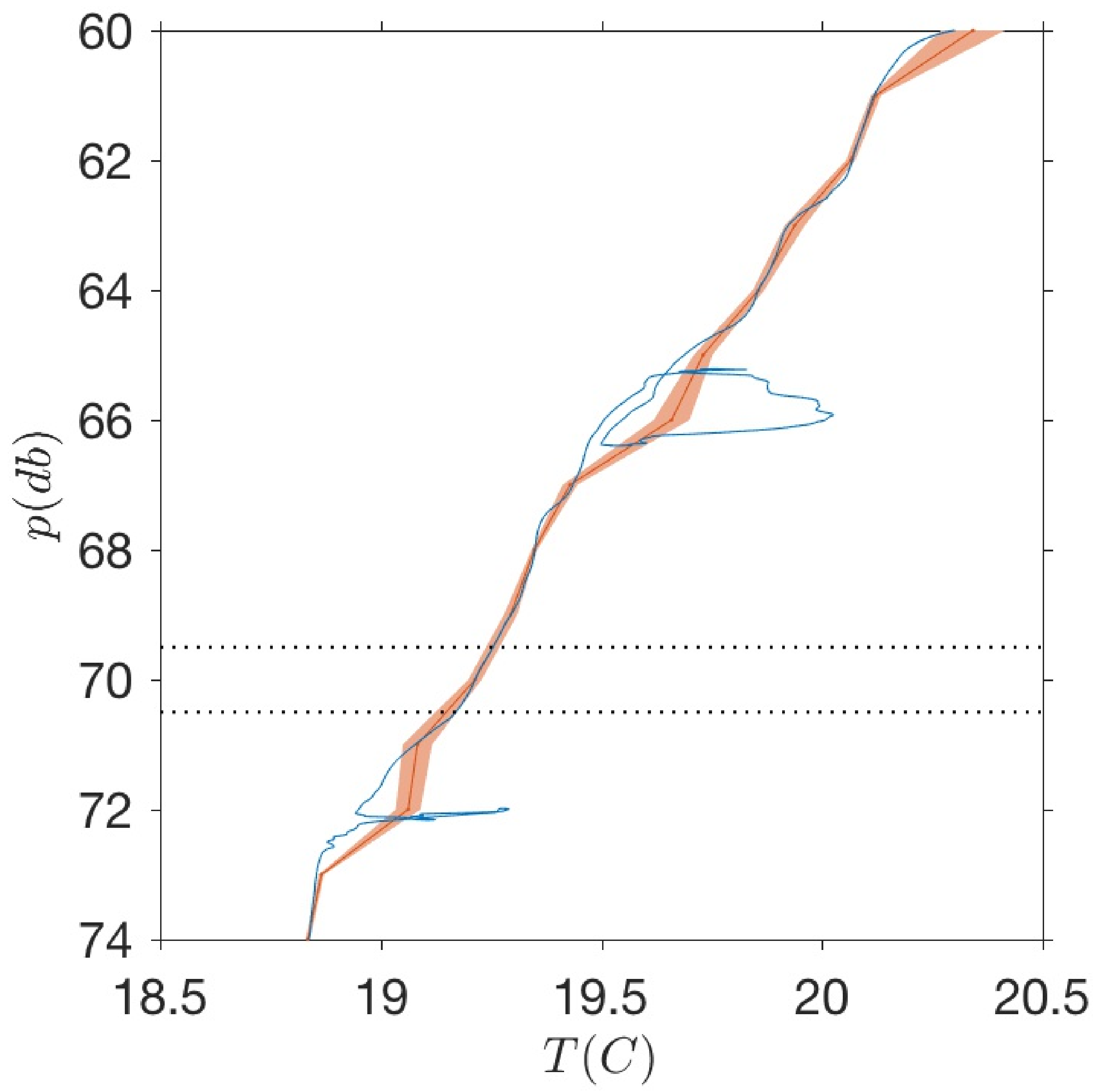
$$\mu_T = 19.21359 \pm 0.01530$$

Previously, we stated that

$$\mu_T = 19.21359 \pm 0.00059$$

So what is the right answer? It is arguable ...

To attempt to answer this question, let's look at the data. The blue curve is the 24 Hz temperature data, and the red curve and shading show the 1 db bin averages and 95% CI using the t distribution.



Instrumental vs sampling/model errors

When we use $\overline{T}$ as an estimate of the temperature at 70 db, we assumed that the $N = 11$ records of temperature at 24 Hz had the same expectation and variance.

Instrumental vs sampling/model errors

When we use $\overline{T}$ as an estimate of the temperature at 70 db, we assumed that the $N = 11$ records of temperature at 24 Hz had the same expectation and variance.

However, we know based on our oceanographic knowledge that the temperature is not necessarily a constant with depth. Here, the obvious temperature gradient with depth makes us think that the expected temperature at 69.5 db is not the same as at 70.5 db.

Instrumental vs sampling/model errors

When we use $\overline{T}$ as an estimate of the temperature at 70 db, we assumed that the $N = 11$ records of temperature at 24 Hz had the same expectation and variance.

However, we know based on our oceanographic knowledge that the temperature is not necessarily a constant with depth. Here, the obvious temperature gradient with depth makes us think that the expected temperature at 69.5 db is not the same as at 70.5 db.

Thus, when we calculate an arithmetic average ($\overline{T}$) of temperature that is a function depth, it is an estimate of a temperature quantity that is variable because of 1) the accuracy of the sensor, and 2) the varying expectation value.

Instrumental vs sampling/model errors

When we use $\overline{T}$ as an estimate of the temperature at 70 db, we assumed that the $N = 11$ records of temperature at 24 Hz had the same expectation and variance.

However, we know based on our oceanographic knowledge that the temperature is not necessarily a constant with depth. Here, the obvious temperature gradient with depth makes us think that the expected temperature at 69.5 db is not the same as at 70.5 db.

Thus, when we calculate an arithmetic average ($\overline{T}$) of temperature that is a function depth, it is an estimate of a temperature quantity that is variable because of 1) the accuracy of the sensor, and 2) the varying expectation value.

Traditionally, this second type of error is called a *sampling error* or a *model error*, which adds to the first type of error called *instrumental error*.

Instrumental vs sampling/model errors

When we use $\overline{T}$ as an estimate of the temperature at 70 db, we assumed that the $N = 11$ records of temperature at 24 Hz had the same expectation and variance.

However, we know based on our oceanographic knowledge that the temperature is not necessarily a constant with depth. Here, the obvious temperature gradient with depth makes us think that the expected temperature at 69.5 db is not the same as at 70.5 db.

Thus, when we calculate an arithmetic average ($\overline{T}$) of temperature that is a function depth, it is an estimate of a temperature quantity that is variable because of 1) the accuracy of the sensor, and 2) the varying expectation value.

Traditionally, this second type of error is called a *sampling error* or a *model error*, which adds to the first type of error called *instrumental error*.

Part of the analysis of your data is to understand, or model, the sources or variance and hence of errors when calculating derived quantities such as mean, variance etc.

Interlude: modeling signal and noise

Part of the problem of choosing the appropriate variance to estimate errors is choosing a model for the observations. As an example, the measured "process" x may be the sum of a given signal y plus instrumental noise ε.

$$x = y + \varepsilon$$

Interlude: modeling signal and noise

Part of the problem of choosing the appropriate variance to estimate errors is choosing a model for the observations. As an example, the measured "process" x may be the sum of a given signal y plus instrumental noise ε.

$$x = y + \varepsilon$$

If the signal and noise independent, then the total variance of the process is

$$\begin{aligned} \mathrm{Var}[x] &= \mathrm{Var}[y] + \mathrm{Var}[\varepsilon] \\ \sigma_x^2 &= \sigma_y^2 \qquad\quad + \sigma_\varepsilon^2 \end{aligned}$$

Interlude: modeling signal and noise

Part of the problem of choosing the appropriate variance to estimate errors is choosing a model for the observations. As an example, the measured "process" x may be the sum of a given signal y plus instrumental noise ε.

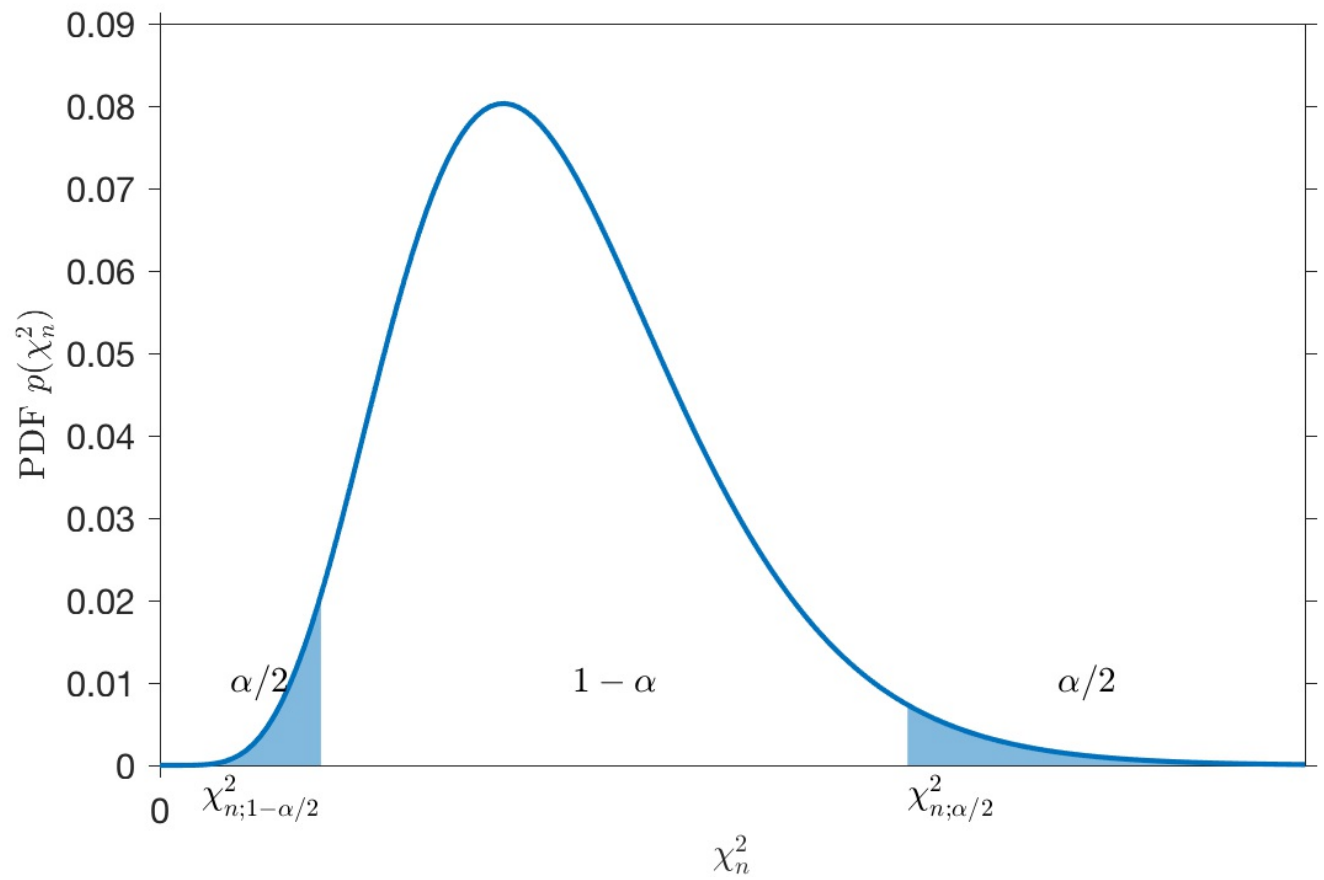
$$x = y + \varepsilon$$

If the signal and noise independent, then the total variance of the process is

$$\begin{aligned} \mathrm{Var}[x] &= \mathrm{Var}[y] + \mathrm{Var}[\varepsilon] \\ \sigma_x^2 &= \sigma_y^2 \qquad\; + \sigma_\varepsilon^2 \end{aligned}$$

In our previous example of the CTD profile, the sample variance of the measurements was likely the sum of the instrumental error and of the "error" from the background shear of temperature. I would tend to choose the second case (t distribution with unknown variance) to derive CIs.

Confidence intervals: χ^2_n case



$$\text{Prob}\left[\chi^2_{n;1-\alpha/2} < \chi^2_n \leq \chi^2_{n;\alpha/2}\right] = 1-\alpha$$

The χ^2 distribution is defined for positive values only, and is not symmetric: $\chi^2_{n;1-\beta} \neq -\chi^2_{n;\beta}$

Confidence intervals: χ^2_n case example

The χ^2 distribution can be used to derive CIs for variance estimates. It can be shown that for N samples drawn from a normally distributed r.v. x with variance σ_x^2, we have

$$\frac{(N-1)s_x^2}{\sigma_x^2} \sim \chi^2_{N-1}$$

which can be used to derive $100(1-\alpha)\%$ CI for variance estimates s_x^2 as

$$\frac{(N-1)s_x^2}{\chi^2_{N-1;\alpha/2}} \leq \sigma^2 < \frac{(N-1)s_x^2}{\chi^2_{N-1;1-\alpha/2}}$$

Hypothesis testing

Confidence intervals are particular cases of *hypothesis testing*, a case of data analysis that occurs frequently. See the introduction of *100 statistical tests* by G. K. Kanji (see reference [5]).

Hypothesis testing

Confidence intervals are particular cases of *hypothesis testing*, a case of data analysis that occurs frequently. See the introduction of *100 statistical tests* by G. K. Kanji (see reference [5]).

Hypothesis testing does **not** consist in proving or disproving hypotheses. Just like we will never know the true value of a r.v., we will never prove in an indeniable fashion that an hypothesis is true.

Hypothesis testing

Confidence intervals are particular cases of *hypothesis testing*, a case of data analysis that occurs frequently. See the introduction of *100 statistical tests* by G. K. Kanji (see reference [5]).

Hypothesis testing does **not** consist in proving or disproving hypotheses. Just like we will never know the true value of a r.v., we will never prove in an indeniable fashion that an hypothesis is true.

Hypothesis testing consists in showing that an hypothesis cannot be supported given its small probability. How small is your own choice, or the difference between getting published or not published, or a stakeholder taking a decision or action or inaction, etc.

Hypothesis testing

Confidence intervals are particular cases of *hypothesis testing*, a case of data analysis that occurs frequently. See the introduction of *100 statistical tests* by G. K. Kanji (see reference [5]).

Hypothesis testing does **not** consist in proving or disproving hypotheses. Just like we will never know the true value of a r.v., we will never prove in an indeniable fashion that an hypothesis is true.

Hypothesis testing consists in showing that an hypothesis cannot be supported given its small probability. How small is your own choice, or the difference between getting published or not published, or a stakeholder taking a decision or action or inaction, etc.

In general, the hypothesis we are trying to denouce, decry, etc, is one with no change (i.e. $a = b$, "the mean temperature today is the same as yesterday"), so that it is typically called the *null hypothesis*, H_0. When H_0 is rejected because of insufficient probability, we accept the *alternatice hypothesis* H_1 (i.e. $a \neq b$, "the mean temperature today is different from yesterday").

Hypothesis testing

Step 1

Define your practical problem in terms of simple hypotheses, a *null hypothesis* and an *alternate hypothesis* that typically leads to action. Decide if you are likely to conduct a *one-tailed* or *two-tailed* test.

Hypothesis testing

Step 1

Define your practical problem in terms of simple hypotheses, a *null hypothesis* and an *alternate hypothesis* that typically leads to action. Decide if you are likely to conduct a *one-tailed* or *two-tailed* test.

As an example, a null hypothesis is that the population mean μ_x of a r.v. x is equal to a given value μ_0 (maybe 0). Alternative hypotheses may be that μ_x is not equal to μ_0 (case 1, two-tailed test), or that μ_x is greater or smaller than μ_0 (cases $2, 3$, one-tailed tests).

1. $H_0 : \mu_x = \mu_0$
   $H_1 : \mu_x \neq \mu_0$
2. $H_0 : \mu_x = \mu_0$
   $H_1 : \mu_x > \mu_0$
3. $H_0 : \mu_x = \mu_0$
   $H_1 : \mu_x < \mu_0$

Hypothesis testing

Step 2

Derive a statistic, that is a number, that can be calculated from your data and your assumptions, typically under your null hypothesis H_0. Make sure that this number is going to be different when H_0 is true or when H_1 is true.

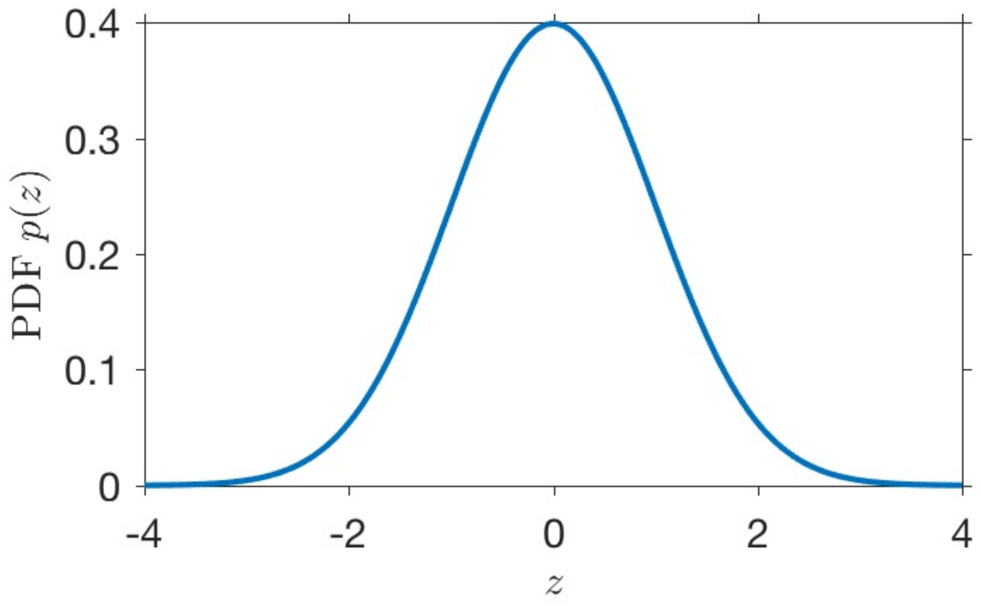
Hypothesis testing

Step 2

Derive a statistic, that is a number, that can be calculated from your data and your assumptions, typically under your null hypothesis H_0. Make sure that this number is going to be different when H_0 is true or when H_1 is true.

Following the previous example, we saw that if x is normally distributed with known variance σ_x, then the statistic

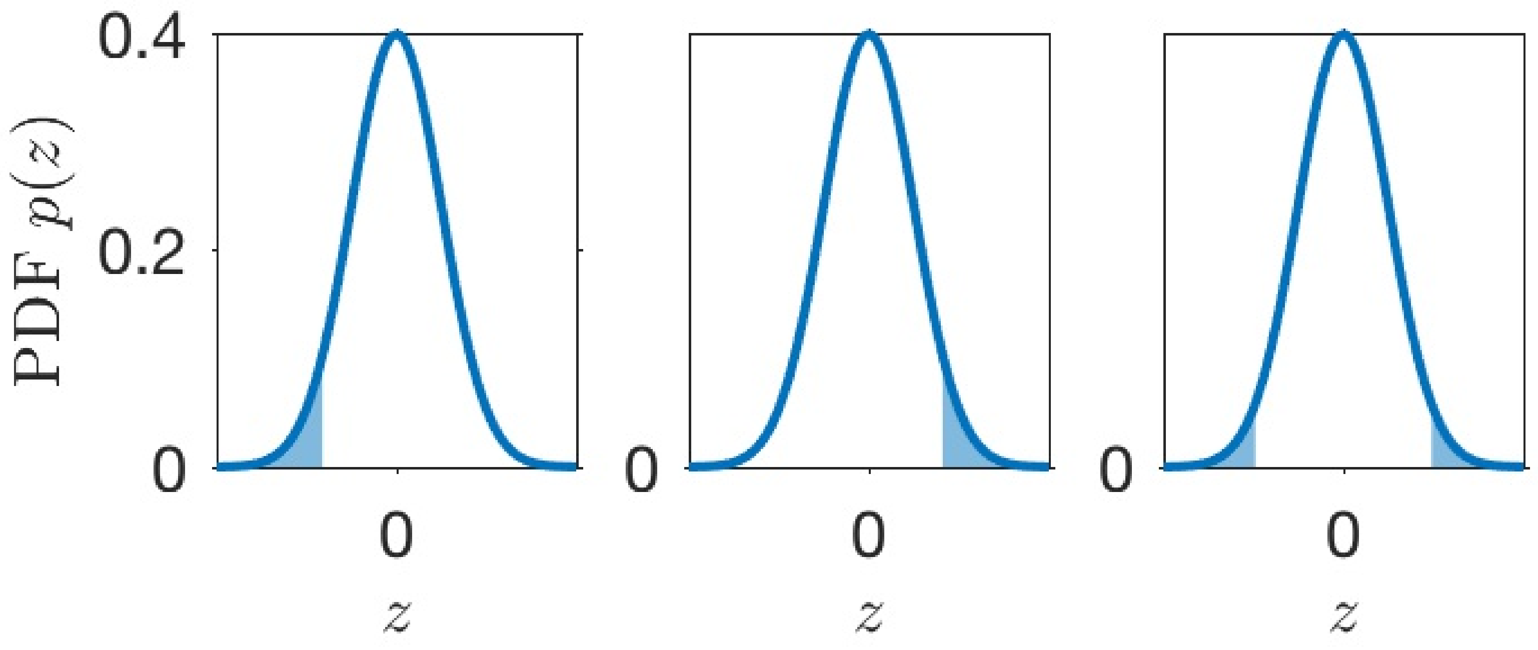
$$z = \frac{\overline{X} - \mu_x}{\sigma_x / \sqrt{N}} \sim \mathcal{N}(0, 1)$$



Hypothesis testing

Steps 3 & 4

Choose a *critical region* for your test statistic and a significance level α that determine the size of your critical region. Critical regions can be of three types; *right-sided* means that you reject H_0 if your test statistic is greater than or equal to some right critical value; *left-sided* you get it; or *both-sided* so that you reject H_0 if your test statistic is either greater than or equal to the right critical value or less than or equal to the left critical value.



Hypothesis testing

Steps 3 & 4: example

$$\alpha = 0.05, H_0 : \mu_0 = 4, N = 9, \overline{X} = 4.6, \sigma_x = 1.0 \rightarrow z = \frac{4.6 - 4}{1/\sqrt{9}} = 1.8$$

Case 1: $z_{1-0.05/2} = -1.96 < z = 1.8 < z_{0.05/2} = 1.96$

z is outside of the critical region! No reason to reject H_0 (i.e. we accept that the mean is not different from μ_0)

Case 2: $z = 1.8 > z_{1-0.05} = 1.64$

z is in the critical region for a right-sided test! We can reject H_0 (in the sense that the mean appears larger than μ_0)

Case 3: $z_{0.05} = -1.64 \leq z = 1.8$

z is outside the critical region for a left-sided test! No reason to reject H_0 (in the sense that it mean does not appear to be less than μ_0).

Hypothesis testing

Please see the book by G. K. Kanji, *100 Statistical Tests*, (2006)! It is very handy ...

One last thing: Error propagation

We saw common cases where the statistics were $x \sim z$, t, or χ^2 r.v. What if we are trying to assess the error or uncertainty for a r.v. y that is arbitrarily function of N variables x_n with *independent* random errors ε_{x_n} (maybe the RMS error)?

$$y = y(x_1, x_2, \ldots, x_N)$$

One last thing: Error propagation

We saw common cases where the statistics were $x \sim z, t$, or χ^2 r.v. What if we are trying to assess the error or uncertainty for a r.v. y that is arbitrarily function of N variables x_n with *independent* random errors ε_{x_n} (maybe the RMS error)?

$$y = y(x_1, x_2, \ldots, x_N)$$

An approximate formula for "small" errors is

$$\varepsilon_y^2 \approx \left(\frac{\partial y}{\partial x_1}\right)^2 \varepsilon_{x_1}^2 + \left(\frac{\partial y}{\partial x_2}\right)^2 \varepsilon_{x_2}^2 + \ldots + \left(\frac{\partial y}{\partial x_N}\right)^2 \varepsilon_{x_N}^2$$

See reference [3].

Practical session

Please download data at the following link:

Please download the Matlab code at the following link:

Make sure you have installed and tested the free jLab Matlab toolbox from Jonathan Lilly at www.jmlilly.net/jmlsoft.html

Extra slides

t-test for two population means (variances unknown and unequal)

Following test #9 of Kanji (2006), reference [5]

The test statistic is

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{\frac{1}{2}}}$$

which is used to test $\mu_1 = \mu_2$, so that $t \sim t(0, \nu)$ with

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}}$$

Kolmogorov-Smirnov test for distribution

The *Kolmogorov-Smirnov* test compares an empirical distribution function $\widehat{F}$ to a prescribed normal distribution function F with mean μ and standard deviation σ. It considers the statistic

$$D = \max_{X_i} |\widehat{F}(X_i) - F(\mu, \sigma)|$$

which measures the maximum distance between the two distribution (as seen on a Q-Q plot).

The issue is that this test is too conservative when the mean and std of F are calculated from the data. An alternative test is called the *Lilliefors test*, which is more stringent. See also test 20 of Kanji (2006), reference [5] for another test.

In Matlab:

```
h = kstest(x); h = lillietest(x);
```