Basic Concepts of Statistics

Romain Madar (CNRS/IN2P3/LPC)

School Of Statistics

Carry-le-Rouet – 16/05/2022

School of Statistics

ATLAS EXPERIMENT

Run: 282712
Event: 1212489545
2015-10-21 09:39:30 CEST

General introduction

Statistics and probability are everywhere in science *and* in everyday life.

General introduction

Statistics and probability are everywhere in science *and* in everyday life.

Attempt to extract quantitative information from the "non fully certain"

General introduction

Statistics and probability are everywhere in science *and* in everyday life.

Attempt to extract quantitative information from the "non fully certain"

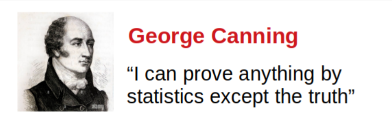
- single realisation of a measurement
- complex systems and/or dynamics (from the forecast, to a flipping coin)
- ...

General introduction

Statistics and probability are everywhere in science *and* in everyday life.

Attempt to extract quantitative information from the "non fully certain"

- single realisation of a measurement
- complex systems and/or dynamics (from the forecast, to a flipping coin)
- ...



General introduction

Statistics and probability are everywhere in science *and* in everyday life.

Attempt to extract quantitative information from the "non fully certain"

- single realisation of a measurement
- complex systems and/or dynamics (from the forecast, to a flipping coin)
- ...

George Canning

"I can prove anything by statistics except the truth"

Ernest Rutherford

"If your experiment needs a statistician, you need a better experiment"

General introduction

Statistics and probability are everywhere in science *and* in everyday life.

Attempt to extract quantitative information from the "non fully certain"

- single realisation of a measurement
- complex systems and/or dynamics (from the forecast, to a flipping coin)
- ...

George Canning

"I can prove anything by statistics except the truth"

Ernest Rutherford

"If your experiment needs a statistician, you need a better experiment"

Goals of the lecture

- recap the basics needed for the SOS
- learn how to be critical with statistics (in science, but not only)
- *focus* on meaning and (mis)intuition rather than mathematical rigour

Statistics versus probability (according to Persi Diaconis)

> *The problems considered by probability and statistics are inverse to each other. In probability theory we consider some underlying process which has some randomness [...] and we figure out what happens. In statistics we observe something that has happened, and try to figure out what underlying process would explain those observations.*

Statistics versus probability (according to Persi Diaconis)

> *The problems considered by probability and statistics are inverse to each other. In probability theory we consider some underlying process which has some randomness [...] and we figure out what happens. In statistics we observe something that has happened, and try to figure out what underlying process would explain those observations.*

Few personal tips for this lecture

- keywords/concepts will be listed at the end of each section
  $\rightarrow$ make sure you know the ideas behind them!

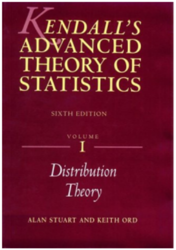
Statistics versus probability (according to Persi Diaconis)

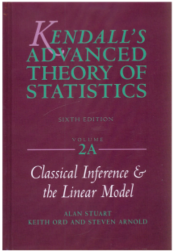
> *The problems considered by probability and statistics are inverse to each other. In probability theory we consider some underlying process which has some randomness [...] and we figure out what happens. In statistics we observe something that has happened, and try to figure out what underlying process would explain those observations.*

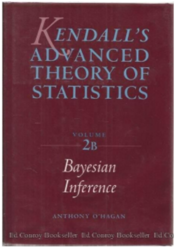
Few personal tips for this lecture

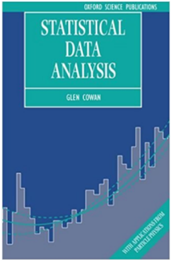
- keywords/concepts will be listed at the end of each section
  → make sure you know the ideas behind them!
- statistics is almost like a language: you need practice to learn it!
  → compute/code as much as simple examples as you can **by yourself!**

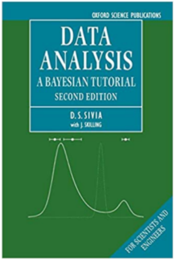
Some references

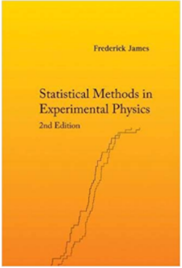


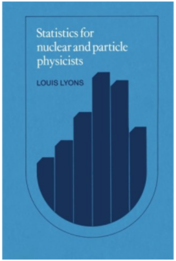


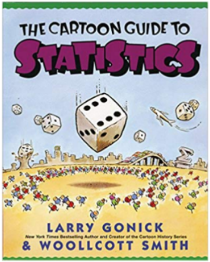












Content

1. Statistics
2. Probability
3. Statistical model
4. The two big schools
5. Parameter estimation and hypothesis testing

Statistics

Descriptive statistics

Definitions:

- Descriptive statistics $\sim$ "summarize" a sample
- sample = set of observations $\mathcal{S} \equiv \{x_1, x_2, ..., x_n\}$

Descriptive statistics

Definitions:

- Descriptive statistics $\sim$ "summarize" a sample
- sample = set of observations $\mathcal{S} \equiv \{x_1, x_2, ..., x_n\}$

Sample caracterisation:

- *What if the sample would be replaced by a single value?*
  - arithmetic mean: $\overline{x} = \frac{1}{n} \sum x_i$
  - median: value that separates the sample in half

Descriptive statistics

Definitions:

- Descriptive statistics $\sim$ "summarize" a sample
- sample = set of observations $\mathcal{S} \equiv \{x_1, x_2, ..., x_n\}$

Sample caracterisation:

- *What if the sample would be replaced by a single value?*
  - arithmetic mean: $\overline{x} = \frac{1}{n}\sum x_i$
  - median: value that separates the sample in half
- *How well this single value actually represents the sample?*
  - variance: $v_x = \overline{(x - \overline{x})^2}$; $\sigma_x \equiv \sqrt{v_x}$ - dispersion

Descriptive statistics

Definitions:

- Descriptive statistics $\sim$ "summarize" a sample
- sample = set of observations $\mathcal{S} \equiv \{x_1, x_2, ..., x_n\}$

Sample caracterisation:

- *What if the sample would be replaced by a single value?*
  - arithmetic mean: $\overline{x} = \frac{1}{n} \sum x_i$
  - median: value that separates the sample in half

- *How well this single value actually represents the sample?*
  - variance: $v_x = \overline{(x - \overline{x})^2}$; $\sigma_x \equiv \sqrt{v_x}$ - dispersion
  - Skewness: $\gamma_x = \overline{\left(\frac{x-\overline{x}}{\sigma_x}\right)^3}$ - asymmetry

Descriptive statistics

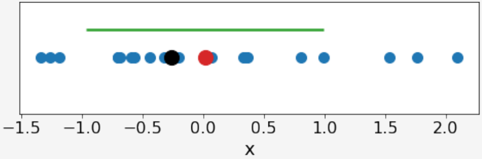
Definitions:

- Descriptive statistics $\sim$ "summarize" a sample
- sample = set of observations $\mathcal{S} \equiv \{x_1, x_2, ..., x_n\}$

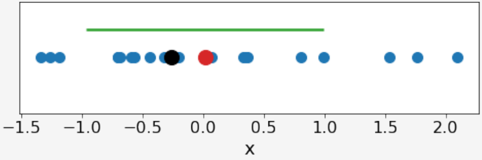
Sample caracterisation:

- *What if the sample would be replaced by a single value?*
  - arithmetic mean: $\overline{x} = \frac{1}{n}\sum x_i$
  - median: value that separates the sample in half
- *How well this single value actually represents the sample?*
  - variance: $v_x = \overline{(x-\overline{x})^2}$; $\sigma_x \equiv \sqrt{v_x}$ - dispersion
  - Skewness: $\gamma_x = \overline{\left(\frac{x-\overline{x}}{\sigma_x}\right)^3}$ - asymmetry
  - Kurtosis: $\beta_x = \overline{\left(\frac{x-\overline{x}}{\sigma_x}\right)^4}$ - importance of tails

Sample caracterisation - illustrations

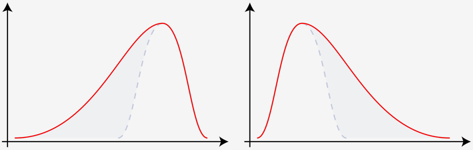


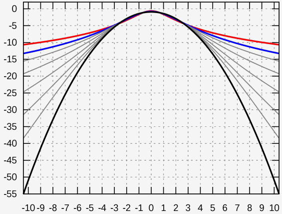
blue: x_i, red: mean. black: median, green: σ_x



blue: x_i, red: mean. black: median, green: σ_x

Skewness and Kurtosis (using probability functions)





Right plot: Kurtosis $\gamma = \infty$ (red), 2 (blue), $1, 1/2, 1/4, 1/8$, and $1/16$ (gray), 0 (black)

Sample caracterisation - comments

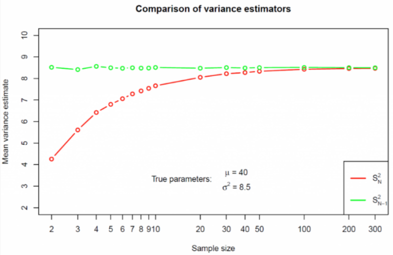
Notion of estimator (more on this later)

- e.g.: sample mean $\neq$ "true mean"
- sample mean $\equiv$ estimator of the true mean
- estimators can be biased - they don't converge to the true value

Sample caracterisation - comments

Notion of estimator (more on this later)

- e.g.: sample mean $\neq$ "true mean"
- sample mean $\equiv$ estimator of the true mean
- estimators can be biased - they don't converge to the true value



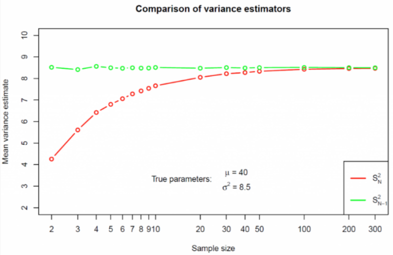
$\rightarrow$ sample variance v_x is a biased estimator of the true variance.

But $\frac{1}{n-\mathbf{1}} \sum (x_i - \overline{x})^2$ is unbiased.

Sample caracterisation - comments

Notion of estimator (more on this later)

- e.g.: sample mean $\neq$ "true mean"
- sample mean $\equiv$ estimator of the true mean
- estimators can be biased - they don't converge to the true value



$\rightarrow$ sample variance v_x is a biased estimator of the true variance.

But $\frac{1}{n-\mathbf{1}}\sum(x_i - \overline{x})^2$ is unbiased.

Statistical moments (more on this later)

- Order-r moment: $m_r = \overline{\left(\frac{x-\overline{x}}{\sigma_x}\right)^r}$ (relates directly to the mean of x^r)
- probability theory: all truth moments $\equiv$ exact underlying probability
- first moments $\equiv$ "main" features of the sample

Correlations

Multidimensional sample

- single observation i = several numbers: $x_i \rightarrow (x_i^{(1)}, x_i^{(2)}, ...x_i^{(p)})$
- *e.g.* biological dataset: person size, weight, age and genre

Previous description applies to each variable $x_i^{(j)}$ but one can now explore how variables behave wrt each other.

Correlations

Multidimensional sample

- single observation i = several numbers: $x_i \rightarrow (x_i^{(1)}, x_i^{(2)}, ...x_i^{(p)})$
- *e.g.* biological dataset: person size, weight, age and genre

Previous description applies to each variable $x_i^{(j)}$ but one can now explore how variables behave wrt each other.

Covariance and correlations between two variables a and b:

$$\mathrm{cov}_{ab} \equiv \overline{(a - \overline{a})(b - \overline{b})} \quad ; \quad \rho_{ab} \equiv \frac{\mathrm{cov}_{ab}}{\sigma_a \sigma_b}$$

Correlations

Multidimensional sample

- single observation i = several numbers: $x_i \to (x_i^{(1)}, x_i^{(2)}, ... x_i^{(p)})$
- *e.g.* biological dataset: person size, weight, age and genre

Previous description applies to each variable $x_i^{(j)}$ but one can now explore how variables behave wrt each other.

Covariance and correlations between two variables a and b:

$$\mathrm{cov}_{ab} \equiv \overline{(a - \overline{a})(b - \overline{b})} \quad ; \quad \rho_{ab} \equiv \frac{\mathrm{cov}_{ab}}{\sigma_a \sigma_b}$$

- probes if fluctuations around the mean are coherent for a and b

Multidimensional sample

- single observation i = several numbers: $x_i \to (x_i^{(1)}, x_i^{(2)}, ...x_i^{(p)})$
- *e.g.* biological dataset: person size, weight, age and genre

Previous description applies to each variable $x_i^{(j)}$ but one can now explore how variables behave wrt each other.

Covariance and correlations between two variables a and b:

$$\text{cov}_{ab} \equiv \overline{(a - \overline{a})(b - \overline{b})} \quad ; \quad \rho_{ab} \equiv \frac{\text{cov}_{ab}}{\sigma_a \sigma_b}$$

- probes if fluctuations around the mean are coherent for a and b
- covariance (and correlation) are symetric - fortunate
- covariance of x with itself is the variance
- $\rho_{a,b} \in [-1, 1]$; 0 = uncorrelated ($\neq$ indep!), (-)1 = (anti-)correlated

More on correlations

Covariance matrix or error matrix

- $C_{ij} = \rho_{ij} \times \sigma_i \sigma_j$ - real and symmetric.
- ρ_{ij} is the correlation matrix - symmetric with 1's on diagonal.

Why is this object so important?

- find pattern in a dataset (*e.g.* is age correlated to weight?)

More on correlations

Covariance matrix or error matrix

- $C_{ij} = \rho_{ij} \times \sigma_i \sigma_j$ - real and symmetric.
- ρ_{ij} is the correlation matrix - symmetric with 1's on diagonal.

Why is this object so important?

- find pattern in a dataset (*e.g.* is age correlated to weight?)
- encode the 'effective' amount of information in a dataset
  - having many correlated variables doesn't bring much information

More on correlations

Covariance matrix or error matrix

- $C_{ij} = \rho_{ij} \times \sigma_i \sigma_j$ - real and symmetric.
- ρ_{ij} is the correlation matrix - symmetric with 1's on diagonal.

Why is this object so important?

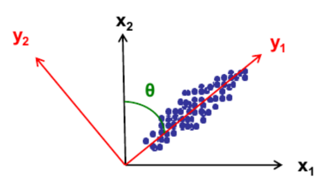
- find pattern in a dataset (*e.g.* is age correlated to weight?)
- encode the 'effective' amount of information in a dataset
  - having many correlated variables doesn't bring much information
  - error propagation (measuring two correlated variables $\sim$ measuring twice the *same* thing)
  - find directions which are uncorrelated (**P**rincipal **C**omponent **A**nalysis)

Covariance matrix or error matrix

- $C_{ij} = \rho_{ij} \times \sigma_i \sigma_j$ - real and symmetric.
- ρ_{ij} is the correlation matrix - symmetric with 1's on diagonal.

Why is this object so important?

- find pattern in a dataset (*e.g.* is age correlated to weight?)
- encode the 'effective' amount of information in a dataset
  - having many correlated variables doesn't bring much information
  - error propagation (measuring two correlated variables $\sim$ measuring twice the *same* thing)
  - find directions which are uncorrelated (**P**rincipal **C**omponent **A**nalysis)



- x_1 and x_2 both have a large σ
- but, they are highly correlated
- most of the information is in y_1 (largest σ)
  $\rightarrow$ idea of dimension reduction
  $\rightarrow$ idea of pre-processing in ML

Correlation and dependence

Correlation $\equiv$ *linear* dependence $\Rightarrow$ dependence

BUT

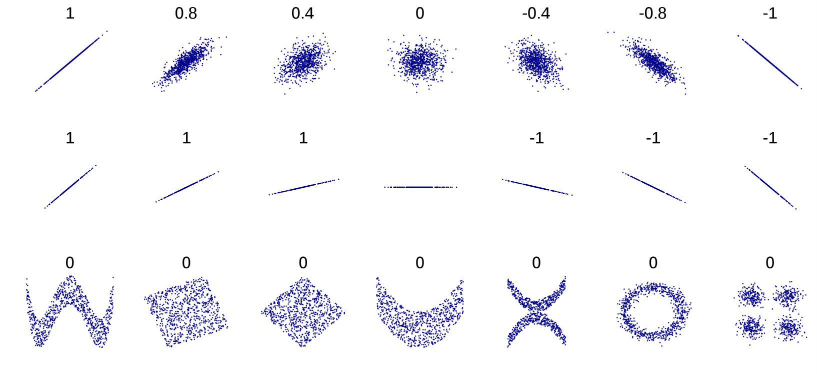
Non-correlation *dosen't* imply independence (matter of vocabulary)

Correlation and dependence

Correlation $\equiv$ *linear* dependence $\Rightarrow$ dependence

BUT

Non-correlation *dosen't* imply independence (matter of vocabulary)



NEVER confuse correlation and causality

Correlation between observations doesn't (necessarily) imply causality

NEVER confuse correlation and causality

Correlation between observations doesn't (necessarily) imply causality

Coluche

“N’allez jamais a l’hopital, on y meurt dix fois plus que chez soi”

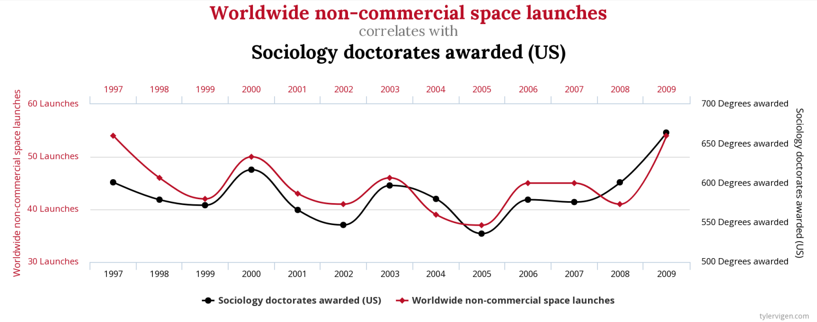
(Never go to the hospital, people there die 10 times more than at home)

Correlation between observations doesn't (necessarily) imply causality

Coluche

"N'allez jamais a l'hopital, on y meurt dix fois plus que chez soi"

(Never go to the hospital, people there die 10 times more than at home)



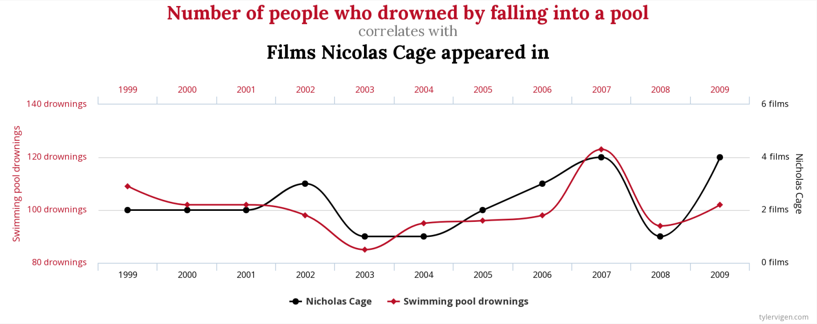
NEVER confuse correlation and causality

Correlation between observations doesn't (necessarily) imply causality

Coluche

"N'allez jamais a l'hopital, on y meurt dix fois plus que chez soi"

(Never go to the hospital, people there die 10 times more than at home)



Keywords and concepts

Part I

descriptive statistics – sample – mean – (co)variance – (de)correlation

Probability

Some definitions

Caution: what follows is *not* mathematically rigorous

Random variable and associated probability

- a random variable X describes an observable which is not certain
- all possible outcomes - realisations - of X form a set Ω
- a probability P_i is associated to each realisation i of Ω
- $\{P_i\}$ must satisfy $P_i \in [0, 1]$ and $\sum P_i = 1$

Some definitions

Caution: what follows is *not* mathematically rigorous

Random variable and associated probability

- a random variable X describes an observable which is not certain
- all possible outcomes - realisations - of X form a set Ω
- a probability P_i is associated to each realisation i of Ω
- $\{P_i\}$ must satisfy $P_i \in [0,1]$ and $\sum P_i = 1$

Simple concrete example: a flippin coin

- $X =$ result of tossing the coin
- $\Omega = \{\text{head}, \text{tail}\} \equiv \{0, 1\}$
- $P_0 = 1/2$ and $P_1 = 1/2$

Some definitions

Caution: what follows is *not* mathematically rigorous

Random variable and associated probability

- a random variable X describes an observable which is not certain
- all possible outcomes - realisations - of X form a set Ω
- a probability P_i is associated to each realisation i of Ω
- $\{P_i\}$ must satisfy $P_i \in [0,1]$ and $\sum P_i = 1$

Simple concrete example: a flippin coin

- $X =$ result of tossing the coin
- $\Omega = \{\text{head}, \text{tail}\} \equiv \{0, 1\}$
- $P_0 = 1/2$ and $P_1 = 1/2$

$\rightarrow$ these notions can be defined and manipulated without any sample

Coming back to estimators - I

Previously: sample mean $\neq$ "true mean". What is the true mean?

$$\mu = \sum_{\Omega} P_i x_i \quad ; \quad \sigma^2 = \sum_{\Omega} P_i \times (x_i - \mu)^2 \quad ; \quad m_r = \sum_{\Omega} P_i \times \left(\frac{x_i - \mu}{\sigma}\right)^r$$

Coming back to estimators - I

Previously: sample mean $\neq$ "true mean". What is the true mean?

$$\mu = \sum_{\Omega} P_i x_i \quad ; \quad \sigma^2 = \sum_{\Omega} P_i \times (x_i - \mu)^2 \quad ; \quad m_r = \sum_{\Omega} P_i \times \left(\frac{x_i - \mu}{\sigma}\right)^r$$

$\rightarrow$ These quantities can be computed without any sample.

Coming back to estimators - I

Previously: sample mean $\neq$ "true mean". What is the true mean?

$$\mu = \sum_{\Omega} P_i x_i \quad ; \quad \sigma^2 = \sum_{\Omega} P_i \times (x_i - \mu)^2 \quad ; \quad m_r = \sum_{\Omega} P_i \times \left(\frac{x_i - \mu}{\sigma}\right)^r$$

$\rightarrow$ These quantities can be computed without any sample.

$\rightarrow$ Estimators connect actual (finite) observations - a sample - and these true quantities, usually not known. **Ultimate goal: find $\mathbf{P_i}$**

Coming back to estimators - I

Previously: sample mean $\neq$ "true mean". What is the true mean?

$$\mu = \sum_{\Omega} P_i x_i \quad ; \quad \sigma^2 = \sum_{\Omega} P_i \times (x_i - \mu)^2 \quad ; \quad m_r = \sum_{\Omega} P_i \times \left(\frac{x_i - \mu}{\sigma}\right)^r$$

→ These quantities can be computed without any sample.

→ Estimators connect actual (finite) observations - a sample - and these true quantities, usually not known. **Ultimate goal: find $\mathbf{P_i}$**

→ This "connection" can more or less good (cf. later).

Coming back to estimators - I

Previously: sample mean $\neq$ "true mean". What is the true mean?

$$\mu = \sum_{\Omega} P_i x_i \quad ; \quad \sigma^2 = \sum_{\Omega} P_i \times (x_i - \mu)^2 \quad ; \quad m_r = \sum_{\Omega} P_i \times \left(\frac{x_i - \mu}{\sigma}\right)^r$$

$\rightarrow$ These quantities can be computed without any sample.

$\rightarrow$ Estimators connect actual (finite) observations - a sample - and these true quantities, usually not known. **Ultimate goal: find $\mathbf{P_i}$**

$\rightarrow$ This "connection" can more or less good (cf. later).

Note: the "true mean" is called expected value and noted $\mathbb{E}(x)$

Coming back to estimators - I

Previously: sample mean $\neq$ "true mean". What is the true mean?

$$\mu = \sum_{\Omega} P_i x_i \quad ; \quad \sigma^2 = \sum_{\Omega} P_i \times (x_i - \mu)^2 \quad ; \quad m_r = \sum_{\Omega} P_i \times \left(\frac{x_i - \mu}{\sigma}\right)^r$$

$\rightarrow$ These quantities can be computed without any sample.

$\rightarrow$ Estimators connect actual (finite) observations - a sample - and these true quantities, usually not known. **Ultimate goal: find $\mathbf{P_i}$**

$\rightarrow$ This "connection" can more or less good (cf. later).

Note: the "true mean" is called expected value and noted $\mathbb{E}(x)$

E.g. of the flipping coin

- $\mu = 1/2$, $\sigma = 1/2$, $m_r = 1$ if r is even and 0 if r is odd

Conditional probabilities and bias theorem

Bias theorem - math version

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

Conditional probabilities and bias theorem

Bias theorem - math version

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

Bias theorem - meaningful version (to me, at least)

$$P(hypothesis|evidence) = P(hypothesis) \times \frac{P(evidence|hypothesis)}{P(evidence)}$$

Conditional probabilities and bias theorem

Bias theorem - math version

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

Bias theorem - meaningful version (to me, at least)

$$P(hypothesis|evidence) = P(hypothesis) \times \frac{P(evidence|hypothesis)}{P(evidence)}$$

- *hypothesis*: the event we are interested in (*e.g.* theory)
- *evidence*: what we observed (*e.g.* measurement)

Conditional probabilities and bias theorem

Bias theorem - math version

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

Bias theorem - meaningful version (to me, at least)

$$P(hypothesis|evidence) = P(hypothesis) \times \frac{P(evidence|hypothesis)}{P(evidence)}$$

- *hypothesis*: the event we are interested in (*e.g.* theory)
- *evidence*: what we observed (*e.g.* measurement)

Comments

- many ways to *understand* this fundamental equation
- in some case, each of these term has a clear meaning
- these two posts are quit interesting post 1 and post 2

Understanding Bias theorem

Example: $hypothesis = fire$ and $evidence = smoke$

$$P(fire|smoke) = P(fire) \times \frac{P(smoke|fire)}{P(smoke)}$$

Understanding Bias theorem

Example: *hypothesis* = *fire* and *evidence* = *smoke*

$$P(fire|smoke) = P(fire) \times \frac{P(smoke|fire)}{P(smoke)}$$

- $P(hypothesis|evidence)$: proba that there is a fire if there is smoke
  $\rightarrow$ difficult to assess (many sources of smoke), that's **the posterior**

Understanding Bias theorem

Example: *hypothesis* = *fire* and *evidence* = *smoke*

$$P(fire|smoke) = P(fire) \times \frac{P(smoke|fire)}{P(smoke)}$$

- $P(hypothesis|evidence)$: proba that there is a fire if there is smoke
  $\rightarrow$ difficult to assess (many sources of smoke), that's **the posterior**
- $P(hypothesis)$: proba that there is a fire
  $\rightarrow$ this our **prior** knowledge about the hypothesis (often **arbitrary**)

Understanding Bias theorem

Example: *hypothesis* = *fire* and *evidence* = *smoke*

$$P(fire|smoke) = P(fire) \times \frac{P(smoke|fire)}{P(smoke)}$$

- $P(hypothesis|evidence)$: proba that there is a fire if there is smoke
  → difficult to assess (many sources of smoke), that's **the posterior**
- $P(hypothesis)$: proba that there is a fire
  → this our **prior** knowledge about the hypothesis (often **arbitrary**)
- $P(evidence|hypothesis)$: proba that there is smoke if there is fire
  → **easy to assess** (fire produces smoke)
  → That is **the** interst of bias theorem

Understanding Bias theorem

Example: $hypothesis = fire$ and $evidence = smoke$

$$P(fire|smoke) = P(fire) \times \frac{P(smoke|fire)}{P(smoke)}$$

- $P(hypothesis|evidence)$: proba that there is a fire if there is smoke
  $\rightarrow$ difficult to assess (many sources of smoke), that's **the posterior**
- $P(hypothesis)$: proba that there is a fire
  $\rightarrow$ this our **prior** knowledge about the hypothesis (often **arbitrary**)
- $P(evidence|hypothesis)$: proba that there is smoke if there is fire
  $\rightarrow$ **easy to assess** (fire produces smoke)
  $\rightarrow$ That is **the** interst of bias theorem
- $P(evidence)$: proba that there is smoke somewhere
  $\rightarrow$ the evidence is rare (valuable) to observe or not (indifferent)

Understanding Bias theorem

Example: *hypothesis* = *fire* and *evidence* = *smoke*

$$P(fire|smoke) = P(fire) \times \frac{P(smoke|fire)}{P(smoke)}$$

- $P(hypothesis|evidence)$: proba that there is a fire if there is smoke
  → difficult to assess (many sources of smoke), that's **the posterior**
- $P(hypothesis)$: proba that there is a fire
  → this our **prior** knowledge about the hypothesis (often **arbitrary**)
- $P(evidence|hypothesis)$: proba that there is smoke if there is fire
  → **easy to assess** (fire produces smoke)
  → That is **the** interst of bias theorem
- $P(evidence)$: proba that there is smoke somewhere
  → the evidence is rare (valuable) to observe or not (indifferent)

N.B.: $P(evidence)$ is independent from the hypothesis, and is sometime impossible to compute. It is often seen as a "normalization factor" and dropped while comparing different hypothesis.

Everyday life questions are often bayesian

Few examples:

- I'm not feeling so well $\rightarrow$ Am I sick ?
- There are clouds $\rightarrow$ will it rain?
- I go out in a bar $\rightarrow$ will I end up drunk?
- I attend to a school statistics $\rightarrow$ will I learn something?

Everyday life questions are often bayesian

Few examples:

- I'm not feeling so well $\rightarrow$ Am I sick ?
- There are clouds $\rightarrow$ will it rain?
- I go out in a bar $\rightarrow$ will I end up drunk?
- I attend to a school statistics $\rightarrow$ will I learn something?

Always the same thinking:

1. you observe a fact
2. you wonder the probability of something, given you this fact happened
3. you have (somtimes rough/wrong) prior, based on past knowledge
4. your brain applies Bias theorem, even you don't know it!

Continous random variables

Generalization to the continuous case

- There is a whole continuum of outcome (realization) for X
- Probability described by a density probability function (PDF), $f(x)$:

$$P(x \in [x_1, x_2]) = \int_{x_1}^{x_2} f(x)\mathrm{d}x \quad ; \quad \int_{\Omega} f(x)\mathrm{d}x = 1$$

Continous random variables

Generalization to the continuous case

- There is a whole continuum of outcome (realization) for X
- Probability described by a density probability function (PDF), $f(x)$:

$$P(x \in [x_1, x_2]) = \int_{x_1}^{x_2} f(x)\mathrm{d}x \quad ; \quad \int_{\Omega} f(x)\mathrm{d}x = 1$$

Moments definitions

$$\mu = \int_{\Omega} x\, f(x)\mathrm{d}x \ ; \ \sigma^2 = \int_{\Omega} (x-\mu)^2 f(x)\mathrm{d}x \ ; \ m_r = \int_{\Omega} \left(\frac{x-\mu}{\sigma}\right)^r f(x)\mathrm{d}x$$

Continous random variables

Generalization to the continuous case

- There is a whole continuum of outcome (realization) for X
- Probability described by a density probability function (PDF), $f(x)$:

$$P(x \in [x_1, x_2]) = \int_{x_1}^{x_2} f(x)\mathrm{d}x \quad ; \quad \int_\Omega f(x)\mathrm{d}x = 1$$

Moments definitions

$$\mu = \int_\Omega x\, f(x)\mathrm{d}x \;;\; \sigma^2 = \int_\Omega (x-\mu)^2 f(x)\mathrm{d}x \;;\; m_r = \int_\Omega \left(\frac{x-\mu}{\sigma}\right)^r f(x)\mathrm{d}x$$

Characteristic function of a PDF

- Fourier transform of the PDF: $\varphi_x(t) = \mathbb{E}(e^{itx}) = \int f(x)e^{itx}\mathrm{d}x$
- many manipulations easier in Fourier space - as in many other fields

Continous random variables

Generalization to the continuous case

- There is a whole continuum of outcome (realization) for X
- Probability described by a density probability function (PDF), $f(x)$:

$$P(x \in [x_1, x_2]) = \int_{x_1}^{x_2} f(x)\mathrm{d}x \quad ; \quad \int_\Omega f(x)\mathrm{d}x = 1$$

Moments definitions

$$\mu = \int_\Omega x\, f(x)\mathrm{d}x \;\; ; \;\; \sigma^2 = \int_\Omega (x-\mu)^2 f(x)\mathrm{d}x \;\; ; \;\; m_r = \int_\Omega \left(\frac{x-\mu}{\sigma}\right)^r f(x)\mathrm{d}x$$

Characteristic function of a PDF

- Fourier transform of the PDF: $\varphi_x(t) = \mathbb{E}(e^{itx}) = \int f(x)e^{itx}\mathrm{d}x$
- many manipulations easier in Fourier space - as in many other fields
- $e^{itx} = \sum \frac{(itx)^n}{n!} \Rightarrow \varphi_x(t) \sim$ linear combination of all moments

Continous random variables

Generalization to the continuous case

- There is a whole continuum of outcome (realization) for X
- Probability described by a density probability function (PDF), $f(x)$:

$$P(x \in [x_1, x_2]) = \int_{x_1}^{x_2} f(x)\mathrm{d}x \quad ; \quad \int_\Omega f(x)\mathrm{d}x = 1$$

Moments definitions

$$\mu = \int_\Omega x\, f(x)\mathrm{d}x \ ; \ \sigma^2 = \int_\Omega (x-\mu)^2 f(x)\mathrm{d}x \ ; \ m_r = \int_\Omega \left(\frac{x-\mu}{\sigma}\right)^r f(x)\mathrm{d}x$$

Characteristic function of a PDF

- Fourier transform of the PDF: $\varphi_x(t) = \mathbb{E}(e^{itx}) = \int f(x)e^{itx}\mathrm{d}x$
- many manipulations easier in Fourier space - as in many other fields
- $e^{itx} = \sum \frac{(itx)^n}{n!} \Rightarrow \varphi_x(t) \sim$ linear combination of all moments
- *knowing all moments* $\equiv$ *knowing the full PDF*

Continous random variables

Generalization to the continuous case

- There is a whole continuum of outcome (realization) for X
- Probability described by a density probability function (PDF), $f(x)$:

$$P(x \in [x_1, x_2]) = \int_{x_1}^{x_2} f(x)\mathrm{d}x \quad ; \quad \int_\Omega f(x)\mathrm{d}x = 1$$

Moments definitions

$$\mu = \int_\Omega x\, f(x)\mathrm{d}x \ ; \ \sigma^2 = \int_\Omega (x-\mu)^2 f(x)\mathrm{d}x \ ; \ m_r = \int_\Omega \left(\frac{x-\mu}{\sigma}\right)^r f(x)\mathrm{d}x$$

Characteristic function of a PDF

- Fourier transform of the PDF: $\varphi_x(t) = \mathbb{E}(e^{itx}) = \int f(x)e^{itx}\mathrm{d}x$
- many manipulations easier in Fourier space - as in many other fields
- $e^{itx} = \sum \frac{(itx)^n}{n!} \Rightarrow \varphi_x(t) \sim$ linear combination of all moments
- *knowing all moments* $\equiv$ *knowing the full PDF*
- moments are the Taylor expension coefficients: $m_r = (-i)^r \left.\frac{\mathrm{d}^r \varphi_X}{\mathrm{d}t^r}\right|_{t=0}$

Important PDF examples

Binomial law: efficiency, trigger rates, …

$$B(k;n,p) = C_k^n p^k (1-p)^{n-k}, \mu = np, \sigma = \sqrt{np(1-p)}$$

Poisson distribution: counting experiments, hypothesis testing

$$P(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}, \mu = \lambda, \sigma = \sqrt{\lambda}$$

Gauss distribution (aka Normal): many use-case (asymptotic convergence)

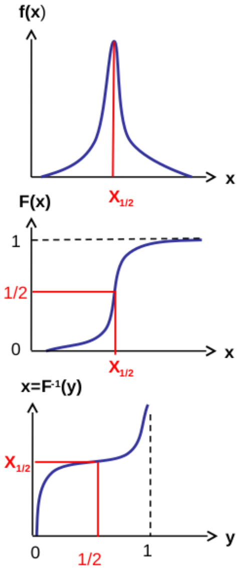
$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Cauchy distribution (aka Breit-Wigner): particle decay width, ...

$$f(x;x_0,\gamma) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$$

μ and σ not defined (divergent integral)

Cumulative distribution and quantiles



Probability density function: $f(x)$

Cumulative distribution: $F(x)=y$

Inverse cumulative distribution: $x=F^{-1}(y)$

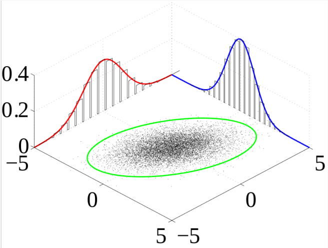
Median: x such that $F(x)=1/2 \rightarrow x_{1/2} = F^{-1}(1/2)$

Quantile of order α: $x_\alpha = F^{-1}(\alpha)$

Multidimensional PDF

How to describe several random variables simulataneously?

- X and Y are two random variables $\rightarrow$ PDF is f_{XY},
- several questions can be asked about X, Y or both.

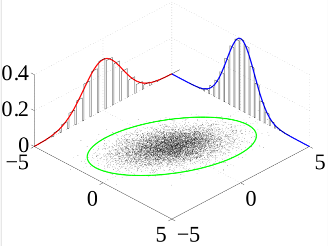


- Probability that $X \in [x, x + \mathrm{d}x]$ and $Y \in [y + \mathrm{d}y]$:
  $\mathrm{d}^2 P(x, y) = f_{XY}(x, y)\mathrm{d}x\mathrm{d}y$
- Probability that $X \in [x, x + \mathrm{d}x]$
  $\mathrm{d}P(x) = \left(\int_y f_{XY}(x, y)\mathrm{d}y\right)\mathrm{d}x$
  $\rightarrow$ this is the marginal PDF

Multidimensional PDF

How to describe several random variables simulataneously?

- X and Y are two random variables $\rightarrow$ PDF is f_{XY},
- several questions can be asked about X, Y or both.



- Probability that $X \in [x, x + \mathrm{d}x]$ and $Y \in [y + \mathrm{d}y]$:
  $\mathrm{d}^2P(x, y) = f_{XY}(x, y)\mathrm{d}x\mathrm{d}y$
- Probability that $X \in [x, x + \mathrm{d}x]$
  $\mathrm{d}P(x) = \left(\int_y f_{XY}(x, y)\mathrm{d}y\right)\mathrm{d}x$
  $\rightarrow$ this is the marginal PDF

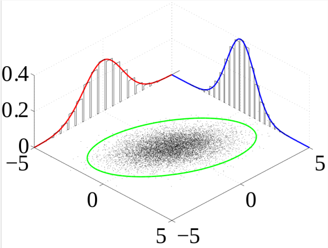
Independent variables $\rightarrow f_{XY}(x, y) = f_X(x) \times f_Y(y)$

- Why? Because marginal PDF is independent from Y behaviour

Multidimensional PDF

How to describe several random variables simulataneously?

- X and Y are two random variables → PDF is f_{XY},
- several questions can be asked about X, Y or both.



- Probability that $X \in [x, x+\mathrm{d}x]$ and $Y \in [y+\mathrm{d}y]$:
  $\mathrm{d}^2 P(x,y) = f_{XY}(x,y)\mathrm{d}x\mathrm{d}y$
- Probability that $X \in [x, x+\mathrm{d}x]$
  $\mathrm{d}P(x) = \left(\int_y f_{XY}(x,y)\mathrm{d}y\right)\mathrm{d}x$
  → this is the marginal PDF

Independent variables → $f_{XY}(x,y) = f_X(x) \times f_Y(y)$

- Why? Because marginal PDF is independent from Y behaviour

$$\rightarrow \mathrm{d}P(x) = \left(\int_y f_{XY}(x,y)\mathrm{d}y\right)\mathrm{d}x = \underbrace{\left(\int_y f_Y(y)\mathrm{d}y\right)}_{=1} f_X(x)\mathrm{d}x$$

Multidimensional normal distribution

$$f(\vec{x};\vec{\mu},\Sigma) \;=\; \frac{1}{\sqrt{(2\pi)^n \det\Sigma}}\,\exp\left(-\frac{1}{2}\,(\vec{x}-\vec{\mu})^T\,\Sigma^{-1}\,(\vec{x}-\vec{\mu})\right)$$

- $\vec{\mu}$ mean position of $\vec{x}$, Σ covariance matrix

Multidimensional normal distribution

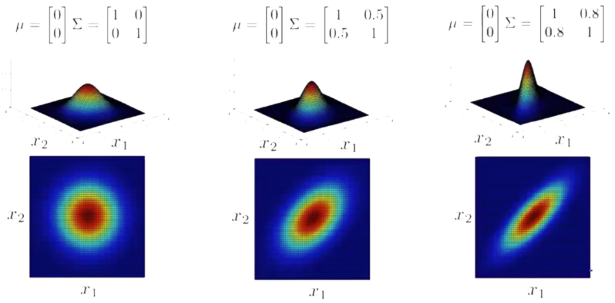
$$f(\vec{x};\vec{\mu},\Sigma) \;=\; \frac{1}{\sqrt{(2\pi)^n \det\Sigma}}\,\exp\left(-\frac{1}{2}\,(\vec{x}-\vec{\mu})^T\,\Sigma^{-1}\,(\vec{x}-\vec{\mu})\right)$$

- $\vec{\mu}$ mean position of $\vec{x}$, Σ covariance matrix
- Σ encodes correlations between x_i and x_j: *if* $\Sigma = diag(\sigma_i)$, *then* $f(\vec{x};\vec{\mu},\Sigma) = \prod_i \mathcal{N}(x_i;\mu_i,\sigma_i)$ - indep. x_i)

Multidimensional normal distribution

$$f(\vec{x};\vec{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^n \det\Sigma}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right)$$

- $\vec{\mu}$ mean position of $\vec{x}$, Σ covariance matrix
- Σ encodes correlations between x_i and x_j: *if* $\Sigma = diag(\sigma_i)$, *then* $f(\vec{x};\vec{\mu},\Sigma) = \prod_i \mathcal{N}(x_i;\mu_i,\sigma_i)$ - indep. x_i)



Central limit theorem

Caution: what follows is *not* mathematically rigorous

If n random variables $\{X_i\}$ are distributed according to the same PDF f_X with a defined mean μ_x and a std σ_x, **then** the random variable $Y = \frac{1}{n}(X_1 + ... + X_n)$ is following a normal distribution of mean μ_x and std $\sigma_x/\sqrt{n}$.

Central limit theorem

Caution: what follows is *not* mathematically rigorous

If n random variables $\{X_i\}$ are distributed according to the same PDF f_X with a defined mean μ_x and a std σ_x, **then** the random variable $Y = \frac{1}{n}(X_1 + ... + X_n)$ is following a normal distribution of mean μ_x and std $\sigma_x/\sqrt{n}$.

For 2 variables $Y = X_1 + X_2$

- The PDF of Y is $f_Y(y) = \int f_{X_1}(x_1) \times f_{X_2}(y - x_1)\mathrm{d}x_1 \to$ convolution!
- Caracteristic function: $\varphi_Y(t) = \varphi_{X_1}(t) \times \varphi_{X_2}(t) = \varphi_X(t)^2$ - same PDF!
- 1st and 2nd moments known : $\varphi_x(t) \sim$ 2nd order Taylor expansion

Central limit theorem

Caution: what follows is *not* mathematically rigorous

If n random variables $\{X_i\}$ are distributed according to the same PDF f_X with a defined mean μ_x and a std σ_x, **then** the random variable $Y = \frac{1}{n}(X_1 + ... + X_n)$ is following a normal distribution of mean μ_x and std $\sigma_x/\sqrt{n}$.

For 2 variables $Y = X_1 + X_2$

- The PDF of Y is $f_Y(y) = \int f_{X_1}(x_1) \times f_{X_2}(y - x_1)\mathrm{d}x_1 \rightarrow$ convolution!
- Caracteristic function: $\varphi_Y(t) = \varphi_{X_1}(t) \times \varphi_{X_2}(t) = \varphi_X(t)^2$ - same PDF!
- 1st and 2nd moments known : $\varphi_x(t) \sim$ 2nd order Taylor expansion

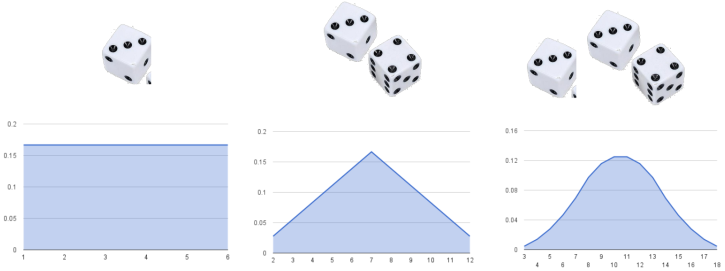
Generalizing for sum of n variables:

- $\varphi_Y(t) = \varphi_x(t)^n \sim \left(1 - \frac{t^2}{n}\right)^n \rightarrow e^{-t/2}$ for $n \rightarrow \infty$
- going back to real space, a normal distribution is obtained

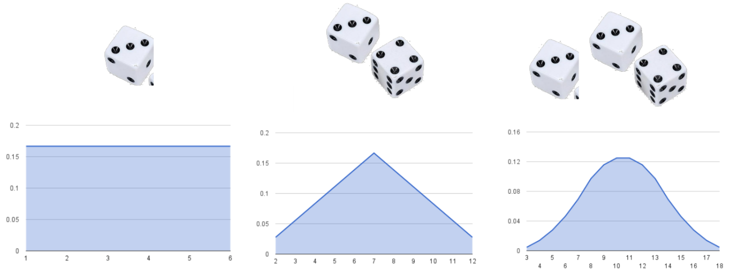
N.B. this reasonning doesn't explain why $\sigma_Y = \sigma_x/\sqrt{n}$, this needs to properly re-scale Y.

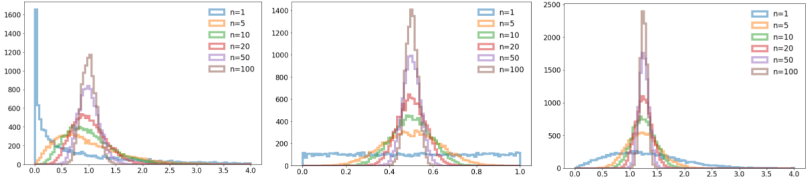
Central limit theorem – continued

One way to understand why it works



One way to understand why it works





Central limit theorem - homework

Proof

Proove that $\sigma_Y = \sigma_X/\sqrt{n}$ with the proper scalings to define Y.

Application

Proove, using the CLT, that a Poisson distribution $P(n;\lambda)$ tends to a normal distribution for large numbers.

Hint: $N = 1 + 1 + 1.... + 1$ N-times

Function of random variables

Final observable is very often a combination of (random) variable.

- $\mathcal{O} = g(X_1, X_2, ..., X_n) \equiv g(\vec{X})$. $\mathcal{O}$ is also a random variable
- what is the PDF of $\mathcal{O}$, knowing $f_{\vec{X}}$? *Not trival* (think about a sum)!

Function of random variables

Final observable is very often a combination of (random) variable.

- $\mathcal{O} = g(X_1, X_2, ..., X_n) \equiv g(\vec{X})$. $\mathcal{O}$ is also a random variable
- what is the PDF of $\mathcal{O}$, knowing $f_{\vec{X}}$? *Not trival* (think about a sum)!
- *What can we say about $\mathcal{O}$ then?* Do we need to know the full $f_{\vec{X}}$?

Function of random variables

Final observable is very often a combination of (random) variable.

- $\mathcal{O} = g(X_1, X_2, ..., X_n) \equiv g(\vec{X})$. $\mathcal{O}$ is also a random variable
- what is the PDF of $\mathcal{O}$, knowing $f_{\vec{X}}$? *Not trival* (think about a sum)!
- *What can we say about $\mathcal{O}$ then?* Do we need to know the full $f_{\vec{X}}$?

Taylor expension around the mean $\vec{\mu}$:

$$\mathcal{O} \approx g(\vec{\mu}) + \sum_i \left.\frac{\partial g}{\partial X_i}\right|_{\vec{X}=\vec{\mu}} (X_i - \mu_i)$$

Function of random variables

Final observable is very often a combination of (random) variable.

- $\mathcal{O} = g(X_1, X_2, ..., X_n) \equiv g(\vec{X})$. $\mathcal{O}$ is also a random variable
- what is the PDF of $\mathcal{O}$, knowing $f_{\vec{X}}$? *Not trival* (think about a sum)!
- *What can we say about $\mathcal{O}$ then?* Do we need to know the full $f_{\vec{X}}$?

Taylor expension around the mean $\vec{\mu}$:

$$\mathcal{O} \approx g(\vec{\mu}) + \sum_i \left.\frac{\partial g}{\partial X_i}\right|_{\vec{X}=\vec{\mu}} (X_i - \mu_i)$$

$\rightarrow \overline{\mathcal{O}} \approx g(\vec{\mu})$ since $\overline{X_i - \mu_i} = 0$

Function of random variables

Final observable is very often a combination of (random) variable.

- $\mathcal{O} = g(X_1, X_2, ..., X_n) \equiv g(\vec{X})$. $\mathcal{O}$ is also a random variable
- what is the PDF of $\mathcal{O}$, knowing $f_{\vec{X}}$? *Not trival* (think about a sum)!
- *What can we say about $\mathcal{O}$ then?* Do we need to know the full $f_{\vec{X}}$?

Taylor expension around the mean $\vec{\mu}$:

$$\mathcal{O} \approx g(\vec{\mu}) + \sum_i \left.\frac{\partial g}{\partial X_i}\right|_{\vec{X}=\vec{\mu}} (X_i - \mu_i)$$

$\rightarrow \overline{\mathcal{O}} \approx g(\vec{\mu})$ since $\overline{X_i - \mu_i} = 0$

$\rightarrow \sigma^2_{\mathcal{O}} \approx \sum_{i,j} \frac{\partial g}{\partial X_i} \frac{\partial g}{\partial X_j}(\vec{\mu}) \times \text{cov}(i,j)$ since $\overline{(X_i - \mu_i)(X_j - \mu_j)} = \text{cov}(i,j)$

Function of random variables

Final observable is very often a combination of (random) variable.

- $\mathcal{O} = g(X_1, X_2, ..., X_n) \equiv g(\vec{X})$. $\mathcal{O}$ is also a random variable
- what is the PDF of $\mathcal{O}$, knowing $f_{\vec{X}}$? *Not trival* (think about a sum)!
- *What can we say about $\mathcal{O}$ then?* Do we need to know the full $f_{\vec{X}}$?

Taylor expension around the mean $\vec{\mu}$:

$$\mathcal{O} \approx g(\vec{\mu}) + \sum_i \left.\frac{\partial g}{\partial X_i}\right|_{\vec{X}=\vec{\mu}} (X_i - \mu_i)$$

$\rightarrow \overline{\mathcal{O}} \approx g(\vec{\mu})$ since $\overline{X_i - \mu_i} = 0$

$\rightarrow \sigma^2_{\mathcal{O}} \approx \sum_{i,j} \frac{\partial g}{\partial X_i}\frac{\partial g}{\partial X_j}(\vec{\mu}) \times \mathrm{cov}(i,j)$ since $\overline{(X_i - \mu_i)(X_j - \mu_j)} = \mathrm{cov}(i,j)$

Comments:

- these equations are known as error propagation
- **this procedure is not exact** and relies on Taylor expansion
- only 1st and 2nd moments of $\vec{X}$ are needed (or their estimators)

Error propagation formula is not exact

(Counter) example with one variable

- X follows a normal distribution ($\sigma_X = 1, \mu_X = 0$), $Y = e^X$

Error propagation formula is not exact

(Counter) example with one variable

- X follows a normal distribution ($\sigma_X = 1, \mu_X = 0$), $Y = e^X$
- approximate formula gives: $\overline{Y} = e^{\mu_X} = 1$ and $\sigma_Y = e^{\mu_X}\sigma_X = 1$

Error propagation formula is not exact

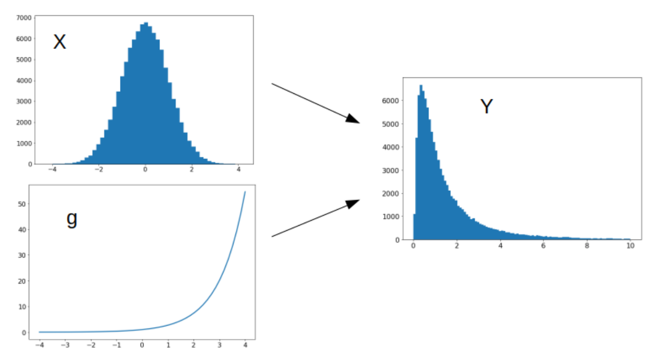
(Counter) example with one variable

- X follows a normal distribution ($\sigma_X = 1, \mu_X = 0$), $Y = e^X$
- approximate formula gives: $\overline{Y} = e^{\mu_X} = 1$ and $\sigma_Y = e^{\mu_X}\sigma_X = 1$
- correct result (from estimator) is $\overline{Y} = 1.6$ and $\sigma_Y = 2.2$

Error propagation formula is not exact

(Counter) example with one variable

- X follows a normal distribution ($\sigma_X = 1, \mu_X = 0$), $Y = e^X$
- approximate formula gives: $\overline{Y} = e^{\mu_X} = 1$ and $\sigma_Y = e^{\mu_X}\sigma_X = 1$
- correct result (from estimator) is $\overline{Y} = 1.6$ and $\sigma_Y = 2.2$



Keywords and concepts

Part I: statistics

descriptive statistics – sample – mean – (co)variance – (de)correlation

Part II: probability

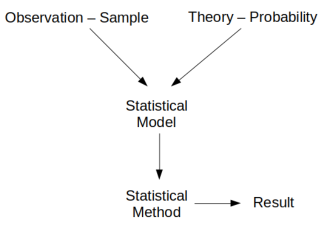
Bias theorem – prior – posterior – random variable – (marginal) PDF –
moments – caracteristic function – (in)dependent variables –
CLT – error propagation

Content

1. Statistics
2. Probability
3. Statistical model
4. The two big schools
5. Parameter estimation and hypothesis testing

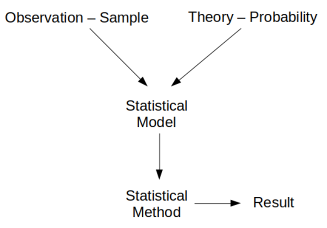
Statistical model

Statistical model: what, why, how



What? missing piece between the "sample" and "probablity"

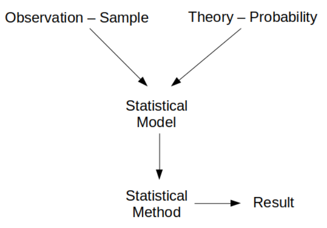
Statistical model: what, why, how



What? missing piece between the "sample" and "probablity"

Why? because a measurement is always one realization of a random variable.

Statistical model: what, why, how

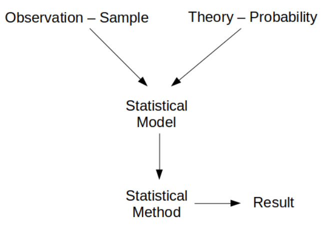


What? missing piece between the "sample" and "probablity"

Why? because a measurement is always one realization of a random variable.

N.B. Statistical methods will be introduced in the next sections

Statistical model: what, why, how



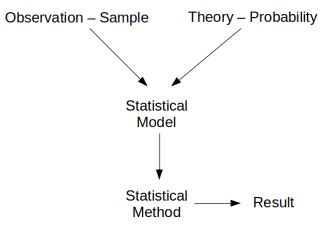
What? missing piece between the "sample" and "probablity"

Why? because a measurement is always one realization of a random variable.

N.B. Statistical methods will be introduced in the next sections

How? physical model + fluctuation model = statistical model

Statistical model: what, why, how



What? missing piece between the "sample" and "probablity"

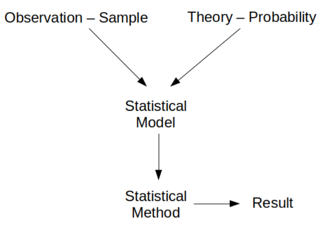
Why? because a measurement is always one realization of a random variable.

N.B. Statistical methods will be introduced in the next sections

How? physical model + fluctuation model = statistical model

Statistical model ingredients:

- (pseudo-)observations, written $\vec{x}$ (or x)
- parameters we want: parameter(s) of interest, written $\vec{\mu}$ or μ (POI)
- parameters we don't care about: nuisance parameters, written $\vec{\theta}$ or θ



What? missing piece between the "sample" and "probablity"

Why? because a measurement is always one realization of a random variable.

N.B. Statistical methods will be introduced in the next sections

How? physical model + fluctuation model = statistical model

Statistical model ingredients:

- (pseudo-)observations, written $\vec{x}$ (or x)
- parameters we want: parameter(s) of interest, written $\vec{\mu}$ or μ (POI)
- parameters we don't care about: nuisance parameters, written $\vec{\theta}$ or θ

A statistical model is also called likelihood function $\mathcal{L}(\vec{\mu}, \vec{\theta}; \vec{x})$. It can be seen as the probability that the physical model predicts the observable $\vec{x}$, given the parameters $(\vec{\mu}, \vec{\theta})$.

Statistical model: particle physics experiment - I

Model ingredients:

- collisions are performed and detected with an efficiency $\epsilon = 0.10$.

Statistical model: particle physics experiment - I

Model ingredients:

- collisions are performed and detected with an efficiency $\epsilon = 0.10$.
- what is measured is a number of events N for a given final state

Statistical model: particle physics experiment - I

Model ingredients:

- collisions are performed and detected with an efficiency $\epsilon = 0.10$.
- what is measured is a number of events N for a given final state
- the physics model tells us $N_{\text{exp}}(\sigma) = \sigma \times L \times \epsilon$
  - σ: cross-section of the studied final state, parameter of interest
  - L: integrated luminosity ($\sim$ amount of collisions)
  - ϵ: detection efficiency

Statistical model: particle physics experiment - I

Model ingredients:

- collisions are performed and detected with an efficiency $\epsilon = 0.10$.
- what is measured is a number of events N for a given final state
- the physics model tells us $N_{\text{exp}}(\sigma) = \sigma \times L \times \epsilon$
  - σ: cross-section of the studied final state, parameter of interest
  - L: integrated luminosity ($\sim$ amount of collisions)
  - ϵ: detection efficiency
- the fluctuation model tells us $P(N; N_{\text{exp}})$ is a Poisson distribution.

Statistical model: particle physics experiment - I

Model ingredients:

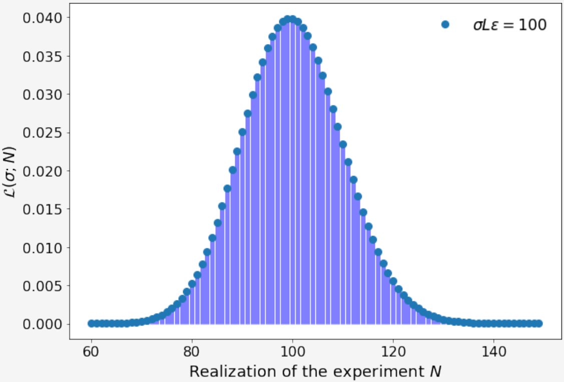
- collisions are performed and detected with an efficiency $\epsilon = 0.10$.
- what is measured is a number of events N for a given final state
- the physics model tells us $N_{\text{exp}}(\sigma) = \sigma \times L \times \epsilon$
    - σ: cross-section of the studied final state, parameter of interest
    - L: integrated luminosity ($\sim$ amount of collisions)
    - ϵ: detection efficiency
- the fluctuation model tells us $P(N; N_{\text{exp}})$ is a Poisson distribution.

Statistical model

$$\mathcal{L}(\sigma; N) = e^{-\sigma L \epsilon} \frac{(\sigma L \epsilon)^N}{N!}$$

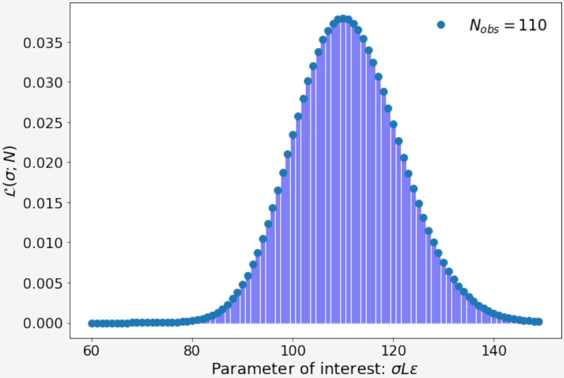
Illustration of the Likelihood

Given a value of σ, what's the "probability" to observe N ?



Anticipation: frequentist "usage" of the likelihood

If we observed a value for N, what's the "probability" that $\sigma = X$?



Anticipation: bayesian "usage" of the likelihood

Statistical model: particle physics experiment - II

Model ingredients:

- the physics model tells us $N_{exp}(\sigma) = \sigma \times L \times \epsilon$
  - σ: cross-section of the studied final state, parameter of interest
  - L: integrated luminosity, with a known uncertainty δL
  - ϵ: detection efficiency, with a uncertainty $\delta\epsilon$

Statistical model: particle physics experiment - II

Model ingredients:

- the physics model tells us $N_{\mathrm{exp}}(\sigma) = \sigma \times L \times \epsilon$
  - σ: cross-section of the studied final state, parameter of interest
  - L: integrated luminosity, with a known uncertainty δL
  - ϵ: detection efficiency, with a uncertainty $\delta \epsilon$
- the fluctuation model tells us $P(N; N_{\mathrm{exp}})$ is a Poisson distribution.

Statistical model: particle physics experiment - II

Model ingredients:

- the physics model tells us $N_{\text{exp}}(\sigma) = \sigma \times L \times \epsilon$
  - σ: cross-section of the studied final state, parameter of interest
  - L: integrated luminosity, with a known uncertainty δL
  - ϵ: detection efficiency, with a uncertainty $\delta\epsilon$
- the fluctuation model tells us $P(N; N_{\text{exp}})$ is a Poisson distribution.

Systematic uncertainties turn numbers into new random variables.
They PDFs depends on parameters, we don't really care about: nuisances parameters. *Example* of systematic parametrization:

$$P(L; L_{\text{truth}}) = \mathcal{N}(L; \mu = L_{\text{truth}}, \sigma = \delta L)$$

Statistical model: particle physics experiment - II

Model ingredients:

- the physics model tells us $N_{\text{exp}}(\sigma) = \sigma \times L \times \epsilon$
  - σ: cross-section of the studied final state, parameter of interest
  - L: integrated luminosity, with a known uncertainty δL
  - ϵ: detection efficiency, with a uncertainty $\delta\epsilon$
- the fluctuation model tells us $P(N; N_{\text{exp}})$ is a Poisson distribution.

Systematic uncertainties turn numbers into new random variables.
They PDFs depends on parameters, we don't really care about: nuisances parameters. *Example* of systematic parametrization:

$$P(L; L_{\text{truth}}) = \mathcal{N}(L; \mu = L_{\text{truth}}, \sigma = \delta L)$$

Statistical model

$$\mathcal{L}(\sigma, L_{\text{truth}}, \epsilon_{\text{truth}}; N) = e^{-\sigma L \epsilon} \frac{(\sigma L \epsilon)^N}{N!} \times P(L; L_{\text{truth}}) \times P(\epsilon; \epsilon_{\text{truth}})$$

More realistic statistical model

In realistic experiment:

- histograms are used - not only event counts
- several samples can be considered simultaneously
- Many processes are usually needed to describe data
- Some are known (backgrounds), others are to be measured (signals)

More realistic statistical model

In realistic experiment:

- histograms are used - not only event counts
- several samples can be considered simultaneously
- Many processes are usually needed to describe data
- Some are known (backgrounds), others are to be measured (signals)

Statistical model (without systematics)

$$\mathcal{L}(\vec{\mu};\vec{x}) = \prod_{\text{bin } i \ \text{ region } j} P_{\text{Poisson}}(x_{i,j} \mid \sum_{bkg} N_{i,j}^{\text{bkg}} + \sum_{sig} N_{i,j}^{\text{sig}}(\mu_{sig}))$$

- $\vec{\mu} = (\sigma_{sig_1}, .., \sigma_{sig_n})$: signal x-sec to be measured (*e.g.* several Higgs prod.)
- $x_{i,j}$: observed number of events in the bin i of the region j

More realistic statistical model

In realistic experiment:

- histograms are used - not only event counts
- several samples can be considered simultaneously
- Many processes are usually needed to describe data
- Some are known (backgrounds), others are to be measured (signals)

Statistical model (without systematics)

$$\mathcal{L}(\vec{\mu};\vec{x}) = \prod_{\text{bin } i \ \text{ region } j} P_{\text{Poisson}}(x_{i,j} \mid \sum_{bkg} N_{i,j}^{\text{bkg}} + \sum_{sig} N_{i,j}^{\text{sig}}(\mu_{sig}))$$

- $\vec{\mu} = (\sigma_{sig_1}, .., \sigma_{sig_n})$: signal x-sec to be measured (*e.g.* several Higgs prod.)
- $x_{i,j}$: observed number of events in the bin i of the region j

Questions for the audience. From a statistical point of view:

- What is more relvant: more regions or more bins?

More realistic statistical model

In realistic experiment:

- histograms are used - not only event counts
- several samples can be considered simultaneously
- Many processes are usually needed to describe data
- Some are known (backgrounds), others are to be measured (signals)

Statistical model (without systematics)

$$\mathcal{L}(\vec{\mu};\vec{x}) = \prod_{\text{bin } i \ \text{ region } j} P_{\text{Poisson}}(x_{i,j} \mid \sum_{bkg} N_{i,j}^{\text{bkg}} + \sum_{sig} N_{i,j}^{\text{sig}}(\mu_{sig}))$$

- $\vec{\mu} = (\sigma_{sig_1}, .., \sigma_{sig_n})$: signal x-sec to be measured (*e.g.* several Higgs prod.)
- $x_{i,j}$: observed number of events in the bin i of the region j

Questions for the audience. From a statistical point of view:

- What is more relvant: more regions or more bins?
- Does the order of bins in histograms matters for the result?

More realistic statistical model

In realistic experiment:

- histograms are used - not only event counts
- several samples can be considered simultaneously
- Many processes are usually needed to describe data
- Some are known (backgrounds), others are to be measured (signals)

Statistical model (without systematics)

$$\mathcal{L}(\vec{\mu};\vec{x}) = \prod_{\text{bin } i \ \text{ region } j} P_{\text{Poisson}}(x_{i,j} \mid \sum_{bkg} N_{i,j}^{\text{bkg}} + \sum_{sig} N_{i,j}^{\text{sig}}(\mu_{sig}))$$

- $\vec{\mu} = (\sigma_{sig_1}, .., \sigma_{sig_n})$: signal x-sec to be measured (*e.g.* several Higgs prod.)
- $x_{i,j}$: observed number of events in the bin i of the region j

Questions for the audience. From a statistical point of view:

- What is more relvant: more regions or more bins?
- Does the order of bins in histograms matters for the result?
- Why do we multiply terms?

A first discussion on uncertainties

Caution

Systematic uncertainty estimation *and* treatment is not an exact science.

(While statistics deals with the non-certain, systematic uncertainties says we don't exactly know the PDF quantifying the non-certain)

A first discussion on uncertainties

Caution

Systematic uncertainty estimation *and* treatment is not an exact science.

(While statistics deals with the non-certain, systematic uncertainties says we don't exactly know the PDF quantifying the non-certain)

Two big classes of uncertainties

- with a statistical nature (typically coming from a *measurement*)

A first discussion on uncertainties

Caution

Systematic uncertainty estimation *and* treatment is not an exact science.

(While statistics deals with the non-certain, systematic uncertainties says we don't exactly know the PDF quantifying the non-certain)

Two big classes of uncertainties

- with a statistical nature (typically coming from a *measurement*)
- without a statistical nature (typically coming from *calculation*)

A first discussion on uncertainties

Caution

Systematic uncertainty estimation *and* treatment is not an exact science.

(While statistics deals with the non-certain, systematic uncertainties says we don't exactly know the PDF quantifying the non-certain)

Two big classes of uncertainties

- with a statistical nature (typically coming from a *measurement*)
- without a statistical nature (typically coming from *calculation*)
- in general: both are present at the same time
- difficult to statistically treat/interpret in the same way

A first discussion on uncertainties

Caution

Systematic uncertainty estimation *and* treatment is not an exact science.

(While statistics deals with the non-certain, systematic uncertainties says we don't exactly know the PDF quantifying the non-certain)

Two big classes of uncertainties

- with a statistical nature (typically coming from a *measurement*)
- without a statistical nature (typically coming from *calculation*)
- in general: both are present at the same time
- difficult to statistically treat/interpret in the same way

Implications:

- arbitrariness (and a loooot of discussion that go with it)
- always check the robustness of the conclusion wrt to those
- that's the way it is, no choice! $\rightarrow$ be *smartly* practical!

Keywords and concepts

Part I: statistics

descriptive statistics – sample – mean – (co)variance – (de)correlation

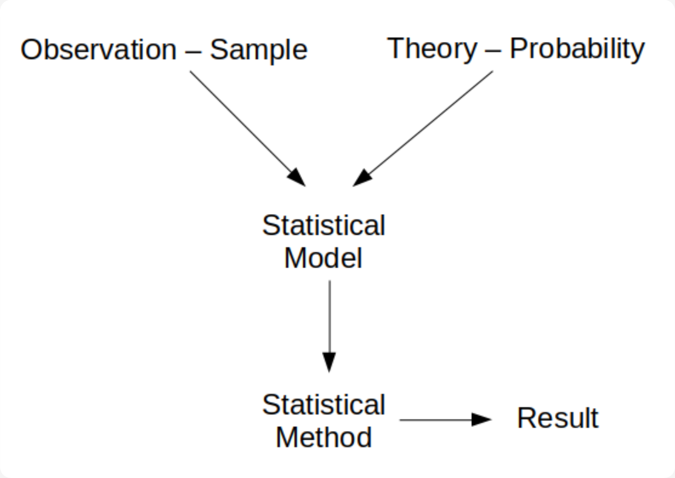
Part II: probability

Bias theorem – prior – posterior – random variable – (marginal) PDF – moments – caracteristic function – (in)dependent variables – CLT – error propagation

Part III: statistical model

Likelihood – nuisance parameter – parameter of interest – systematic uncertainties

Overview



Observation – Sample
Theory – Probability
Statistical Model
Statistical Method
Result

Content

1. Statistics
2. Probability
3. Statistical model
4. The two big schools
5. Parameter estimation and hypothesis testing

The two big schools

Fequentist versus bayesian

	Frequentist	**Bayesian**
probability	frequency of occurence	degree of belief

Fequentist versus bayesian

	Frequentist	**Bayesian**
probability	frequency of occurence	degree of belief
parameters	fixed (once chosen)	uncertain

Fequentist versus bayesian

	Frequentist	**Bayesian**
probability	frequency of occurence	degree of belief
parameters	fixed (once chosen)	uncertain
observation	fluctuates	certain (once observed)

Fequentist versus bayesian

	Frequentist	**Bayesian**
probability	frequency of occurence	degree of belief
parameters	fixed (once chosen)	uncertain
observation	fluctuates	certain (once observed)

The two approaches in a nutshell:

- frequenstist $\rightarrow$ probability of observation, given a model
- bayesian $\rightarrow$ probability of a model, given an observation

Fequentist versus bayesian

	Frequentist	**Bayesian**
probability	frequency of occurence	degree of belief
parameters	fixed (once chosen)	uncertain
observation	fluctuates	certain (once observed)

The two approaches in a nutshell:

- frequenstist → probability of observation, given a model
- bayesian → probability of a model, given an observation

Methodologies

- *frequenstist:* estimates frequencies, by emulating repetitions of the experiment (toys) for a given parameter, using the **likelihood** as PDF

Fequentist versus bayesian

	Frequentist	Bayesian
probability	frequency of occurence	degree of belief
parameters	fixed (once chosen)	uncertain
observation	fluctuates	certain (once observed)

The two approaches in a nutshell:

- frequenstist $\rightarrow$ probability of observation, given a model
- bayesian $\rightarrow$ probability of a model, given an observation

Methodologies

- *frequenstist:* estimates frequencies, by emulating repetitions of the experiment (toys) for a given parameter, using the **likelihood** as PDF
- *bayesian:* exploits the Bayes theorem to compute the posterior $P(para|obs)$, using the prior $P(para)$ and $P(obs|para)$ - the **likelihood**

Is a flipping coin tricked?

The experiment:

We toss a coin 113 times and we got 'tail' 68 times. Is the coin tricked?

Is a flipping coin tricked?

The experiment:

We toss a coin 113 times and we got 'tail' 68 times. Is the coin tricked?

Statistical Model assuming $N = 113$ is large enough to apply CLT

$$\mathcal{L}(p; N_{tail}) = \frac{1}{\sqrt{2\pi N}}\, e^{-\frac{1}{2}\left(\frac{pN - N_{tail}}{\sqrt{N}}\right)^2}$$

- N (known parameter): number of tosses
- N_{tail} (observation): number of time tail is obtained
- p (parameter): balance between the two sides (tricked $p \neq 1/2$).

Is a flipping coin tricked?

The experiment:

We toss a coin 113 times and we got 'tail' 68 times. Is the coin tricked?

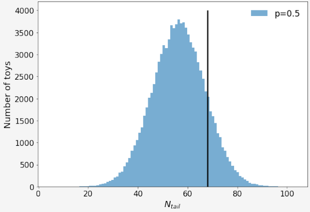
Statistical Model assuming $N = 113$ is large enough to apply CLT

$$\mathcal{L}(p; N_{tail}) = \frac{1}{\sqrt{2\pi N}}\, e^{-\frac{1}{2}\left(\frac{pN - N_{tail}}{\sqrt{N}}\right)^2}$$

- N (known parameter): number of tosses
- N_{tail} (observation): number of time tail is obtained
- p (parameter): balance between the two sides (tricked $p \neq 1/2$).

Let's try to analyze the same experiment with both frequentist and bayesian approaches

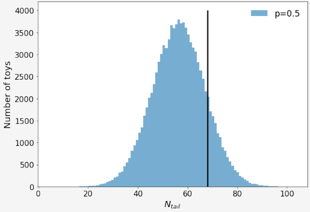
Is a flipping coin tricked? Frequentist approach



Toys with a normal coin

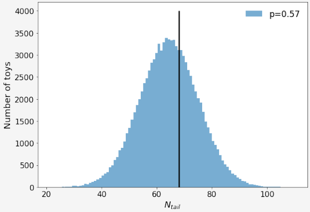
14.1% of pseudo-experiments using an normal coin would lead to $N_{tail} \geq 68$

Is a flipping coin tricked? Frequentist approach



Toys with a normal coin

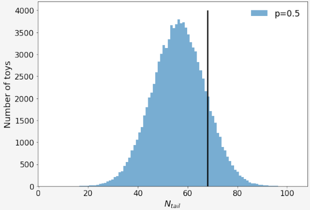
14.1% of pseudo-experiments using an normal coin would lead to $N_{tail} \geq 68$



Toys with a tricked coin

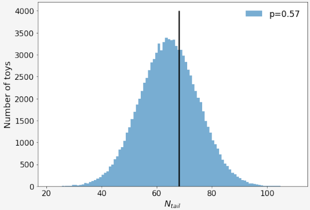
36.8% of pseudo-experiments using an tricked coin with $p = 0.57$ would lead to $N_{tail} \geq 68$

Is a flipping coin tricked? Frequentist approach



Toys with a normal coin

14.1% of pseudo-experiments using an normal coin would lead to $N_{tail} \geq 68$

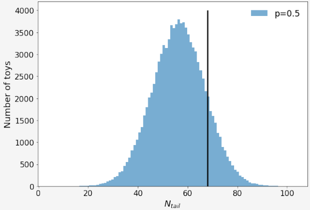


Toys with a tricked coin

36.8% of pseudo-experiments using an tricked coin with $p = 0.57$ would lead to $N_{tail} \geq 68$

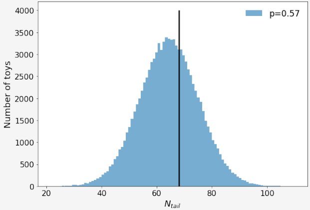
In the end, is the coin tricked?

Is a flipping coin tricked? Frequentist approach



Toys with a normal coin

14.1% of pseudo-experiments using an normal coin would lead to $N_{tail} \geq 68$



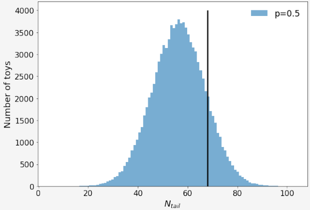
Toys with a tricked coin

36.8% of pseudo-experiments using an tricked coin with $p = 0.57$ would lead to $N_{tail} \geq 68$

In the end, is the coin tricked?

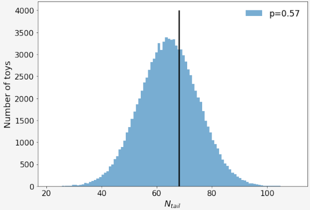
- we can only state confidence levels for each scenario

Is a flipping coin tricked? Frequentist approach



Toys with a normal coin

14.1% of pseudo-experiments using an normal coin would lead to $N_{tail} \geq 68$



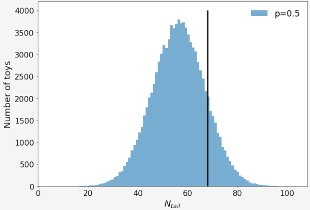
Toys with a tricked coin

36.8% of pseudo-experiments using an tricked coin with $p = 0.57$ would lead to $N_{tail} \geq 68$

In the end, is the coin tricked?

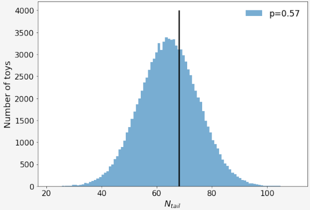
- we can only state confidence levels for each scenario
- according to you, is $p = 0.57$ more probable than $p = 0.50$?

Is a flipping coin tricked? Frequentist approach



Toys with a normal coin

14.1% of pseudo-experiments using an normal coin would lead to $N_{tail} \geq 68$



Toys with a tricked coin

36.8% of pseudo-experiments using an tricked coin with $p = 0.57$ would lead to $N_{tail} \geq 68$

In the end, is the coin tricked?

- we can only state confidence levels for each scenario
- according to you, is $p = 0.57$ more probable than $p = 0.50$?
  $\rightarrow$ this question has no sense in frequentist

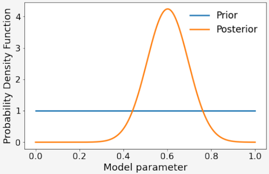
Is a flipping coin tricked? Bayesian approach

$$P(p|N_{tail}) = Prior(p) \times \frac{P(N_{tail}|p)}{P(N_{tail})}$$

Is a flipping coin tricked? Bayesian approach

$$P(p|N_{tail}) = Prior(p) \times \frac{P(N_{tail}|p)}{P(N_{tail})}$$

Flat prior



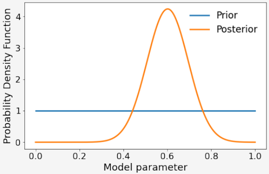
Most probable value is

$p = 0.60$

Is a flipping coin tricked? Bayesian approach

$$P(p|N_{tail}) = Prior(p) \times \frac{P(N_{tail}|p)}{P(N_{tail})}$$

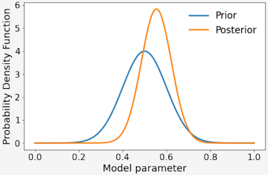
Flat prior



Most probable value is

$p = 0.60$

Wide center prior



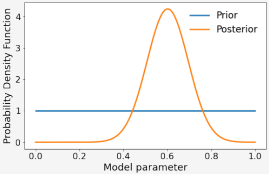
Most probable value is

$p = 0.55$

Is a flipping coin tricked? Bayesian approach

$$P(p|N_{tail}) = Prior(p) \times \frac{P(N_{tail}|p)}{P(N_{tail})}$$

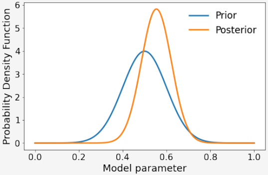
Flat prior



Most probable value is

$p = 0.60$

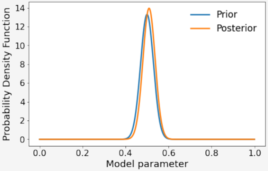
Wide center prior



Most probable value is

$p = 0.55$

Narrow centerd prior



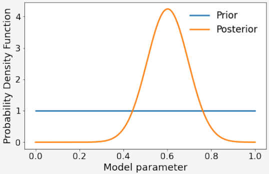
Most probable value is

$p = 0.51$

Is a flipping coin tricked? Bayesian approach

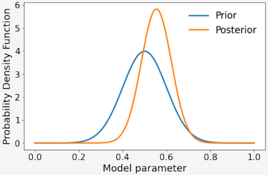
$$P(p|N_{tail}) = Prior(p) \times \frac{P(N_{tail}|p)}{P(N_{tail})}$$

Flat prior



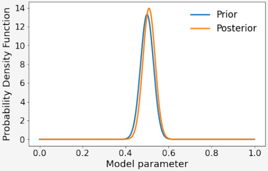
Most probable value is $p = 0.60$

Wide center prior



Most probable value is $p = 0.55$

Narrow centerd prior



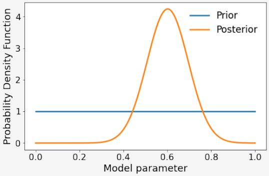
Most probable value is $p = 0.51$

In the end, is the coin tricked?

Is a flipping coin tricked? Bayesian approach

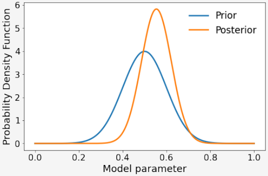
$$P(p|N_{tail}) = Prior(p) \times \frac{P(N_{tail}|p)}{P(N_{tail})}$$

Flat prior



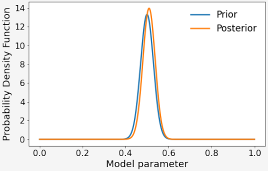
Most probable value is $p = 0.60$

Wide center prior



Most probable value is $p = 0.55$

Narrow centerd prior



Most probable value is $p = 0.51$

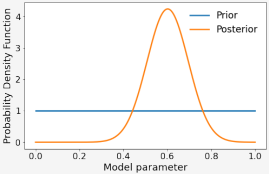
In the end, is the coin tricked?

- we can only state credibility interval for p, which is prior-dependent

Is a flipping coin tricked? Bayesian approach

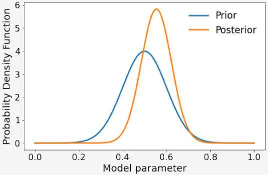
$$P(p|N_{tail}) = Prior(p) \times \frac{P(N_{tail}|p)}{P(N_{tail})}$$

Flat prior



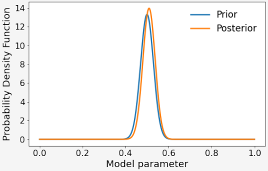
Most probable value is $p = 0.60$

Wide center prior



Most probable value is $p = 0.55$

Narrow centerd prior



Most probable value is $p = 0.51$

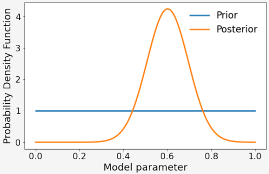
In the end, is the coin tricked?

- we can only state credibility interval for p, which is prior-dependent
- according to you, is $p = 0.57$ more probable than $p = 0.50$?

Is a flipping coin tricked? Bayesian approach

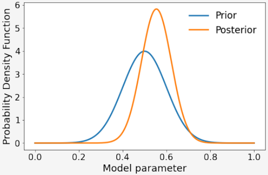
$$P(p|N_{tail}) = Prior(p) \times \frac{P(N_{tail}|p)}{P(N_{tail})}$$

Flat prior



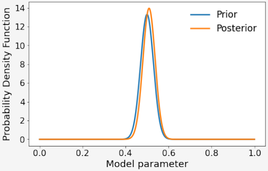
Most probable value is $p = 0.60$

Wide center prior



Most probable value is $p = 0.55$

Narrow centerd prior



Most probable value is $p = 0.51$

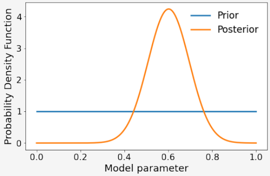
In the end, is the coin tricked?

- we can only state credibility interval for p, which is prior-dependent
- according to you, is $p = 0.57$ more probable than $p = 0.50$?
  $\rightarrow$ this question has now a clear answer in bayesian!

Is a flipping coin tricked? Bayesian approach

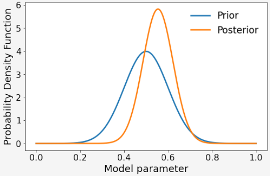
$$P(p|N_{tail}) = Prior(p) \times \frac{P(N_{tail}|p)}{P(N_{tail})}$$

Flat prior



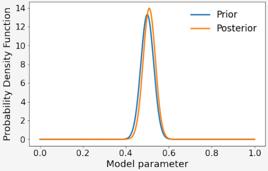
Most probable value is $p = 0.60$

Wide center prior



Most probable value is $p = 0.55$

Narrow centerd prior

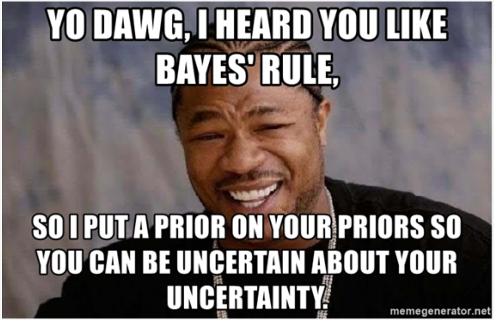


Most probable value is $p = 0.51$

In the end, is the coin tricked?

- we can only state credibility interval for p, which is prior-dependent
- according to you, is $p = 0.57$ more probable than $p = 0.50$?
  $\rightarrow$ this question has now a clear answer in bayesian!
  $\rightarrow$ expect it depends on the choice of the prior ...

Is a flipping coin tricked? Bayesian approach



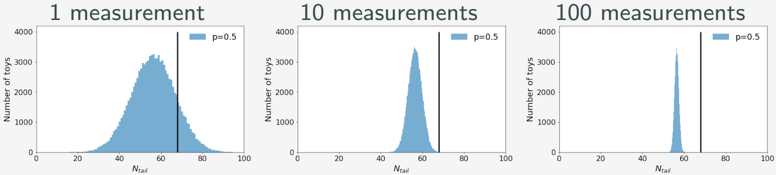
So … Is this coin tricked or not?

Well … statistics can't say for sure (science of handling the "not fully certain"). The unambiguous answer exists only in the limit of infinite number of measurements. **What both methods say in that case?**

So ... Is this coin tricked or not?

Well ... statistics can't say for sure (science of handling the "not fully certain"). The unambiguous answer exists only in the limit of infinite number of measurements. **What both methods say in that case?**

Frequentist



Frequentists say "Yes, the coin is tricked!"

Certainty comes from the extremely low fraction of pseudo-experiments of a normal coin, that would lead the observed result.

So ... Is this coin tricked or not?

Well ... statistics can't say for sure (science of handling the "not fully certain"). The unambiguous answer exists only in the limit of infinite number of measurements. **What both methods say in that case?**

Handling many measurements in Bayesian

- prior is build while accumlating knowledge, supressing the arbitrariness

So ... Is this coin tricked or not?

Well ... statistics can't say for sure (science of handling the "not fully certain"). The unambiguous answer exists only in the limit of infinite number of measurements. **What both methods say in that case?**

Handling many measurements in Bayesian

- prior is build while accumlating knowledge, supressing the arbitrariness
- Prior of i^{th} measurement = posterior of $(i-1)^{th}$ measurement

So ... Is this coin tricked or not?

Well ... statistics can't say for sure (science of handling the "not fully certain"). The unambiguous answer exists only in the limit of infinite number of measurements. **What both methods say in that case?**

Handling many measurements in Bayesian

- prior is build while accumlating knowledge, supressing the arbitrariness
- Prior of i^{th} measurement = posterior of $(i-1)^{th}$ measurement
- $P(p|N_{tail})_{N_{meas}} \propto \mathcal{L}^{N_{meas}-1} \times P(p)$

So ... Is this coin tricked or not?

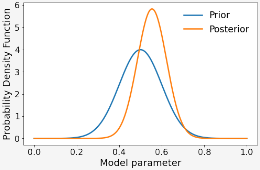
Well ... statistics can't say for sure (science of handling the "not fully certain"). The unambiguous answer exists only in the limit of infinite number of measurements. **What both methods say in that case?**

Handling many measurements in Bayesian

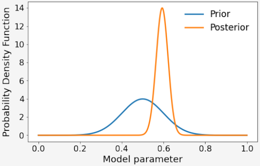
- prior is build while accumlating knowledge, supressing the arbitrariness
- Prior of i^{th} measurement = posterior of $(i-1)^{th}$ measurement
- $P(p|N_{tail})_{N_{meas}} \propto \mathcal{L}^{N_{meas}-1} \times P(p)$

Bayesian, wide prior

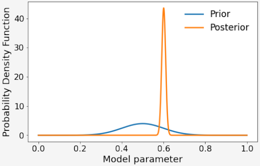
$N_{meas} = 1$



$N_{meas} = 10$



$N_{meas} = 100$



So ... Is this coin tricked or not?

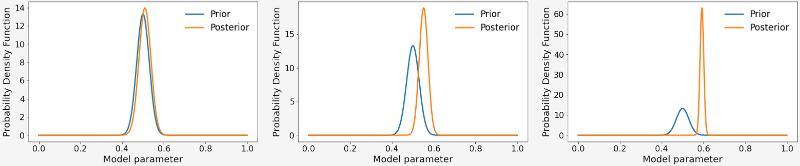
Well ... statistics can't say for sure (science of handling the "not fully certain"). The unambiguous answer exists only in the limit of infinite number of measurements. **What both methods say in that case?**

Handling many measurements in Bayesian

- prior is build while accumlating knowledge, supressing the arbitrariness
- Prior of i^{th} measurement = posterior of $(i-1)^{th}$ measurement
- $P(p|N_{tail})_{N_{meas}} \propto \mathcal{L}^{N_{meas}-1} \times P(p)$

Bayesian, narrow prior

$N_{meas} = 1$ $N_{meas} = 10$ $N_{meas} = 100$



So … Is this coin tricked or not?

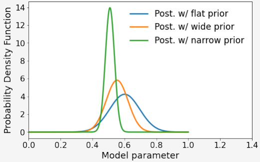
Well … statistics can't say for sure (science of handling the "not fully certain"). The unambiguous answer exists only in the limit of infinite number of measurements. **What both methods say in that case?**

Handling many measurements in Bayesian

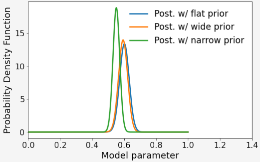
- prior is build while accumlating knowledge, supressing the arbitrariness
- Prior of i^{th} measurement = posterior of $(i-1)^{th}$ measurement
- $P(p|N_{tail})_{N_{meas}} \propto \mathcal{L}^{N_{meas}-1} \times P(p)$

Bayesian, posterior for various priors

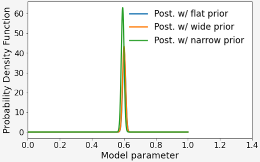
$N_{meas} = 1$



$N_{meas} = 10$



$N_{meas} = 100$



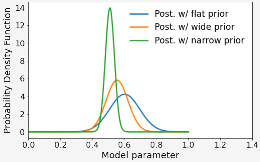
Well ... statistics can't say for sure (science of handling the "not fully certain"). The unambiguous answer exists only in the limit of infinite number of measurements. **What both methods say in that case?**

Handling many measurements in Bayesian

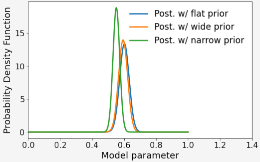
- prior is build while accumlating knowledge, supressing the arbitrariness
- Prior of i^{th} measurement = posterior of $(i-1)^{th}$ measurement
- $P(p|N_{tail})_{N_{meas}} \propto \mathcal{L}^{N_{meas}-1} \times P(p)$

Bayesian, posterior for various priors

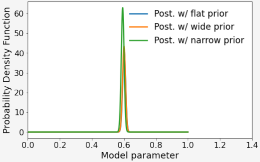
$N_{meas} = 1$



$N_{meas} = 10$



$N_{meas} = 100$



Bayesians also say "Yes, the coin is tricked!"

Frequentist v.s. Bayesian: what to take away

1. Both approaches handle differently the "non fully certain"

2. Final conlusions should be compatible, even if the question they adress are not exaclty the same.

3. Both approaches get unifed when

- there is an infinite number of measurements

Frequentist v.s. Bayesian: what to take away

1. Both approaches handle differently the "non fully certain"

2. Final conlusions should be compatible, even if the question they adress are not exaclty the same.

3. Both approaches get unifed when

- there is an infinite number of measurements
- the prior is uniform: $P(par|obs) = A \times \mathcal{L}(par; obs)$

(same equation, but its meaning and the question it addresses *are* different)

Frequentist v.s. Bayesian: what to take away

1. Both approaches handle differently the "non fully certain"

2. Final conlusions should be compatible, even if the question they adress are not exaclty the same.

3. Both approaches get unifed when

- there is an infinite number of measurements
- the prior is uniform: $P(par|obs) = A \times \mathcal{L}(par;obs)$

  (same equation, but its meaning and the question it addresses *are* different)

You cannot be wrong or right choosing one or the other approach. It's matter of taste (and history)

Frequentist v.s. Bayesian: what to take away

1. Both approaches handle differently the "non fully certain"

2. Final conlusions should be compatible, even if the question they adress are not exaclty the same.

3. Both approaches get unifed when

- there is an infinite number of measurements
- the prior is uniform: $P(par|obs) = A \times \mathcal{L}(par; obs)$

  (same equation, but its meaning and the question it addresses *are* different)

You cannot be wrong or right choosing one or the other approach. It's matter of taste (and history)

One thing I like from the two approaches

- probability intepretation from the frequentist
- ranking two theories using their probability, called Bias factors

Keywords and concepts

Part I: statistics

descriptive statistics – sample – mean – (co)variance – (de)correlation

Part II: probability

Bias theorem – prior – posterior – random variable – (marginal) PDF – moments – caracteristic function – (in)dependent variables – CLT – error propagation

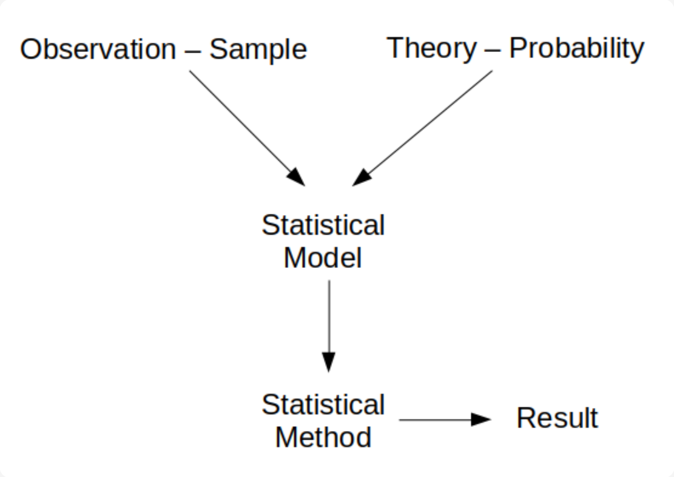
Part III: statistical model

Likelihood – nuisance parameter – parameter of interest – systematic uncertainties

Part IV: The two big school

Frequentist – occurence frequency – pseudo-data (toys) – bayesian – degree of belief

Overview


Observation – Sample
Theory – Probability
Statistical
Model
Statistical
Method
Result

Content

1. Statistics
2. Probability
3. Statistical model
4. The two big schools
5. Parameter estimation and hypothesis testing

Parameter estimation and hypothesis testing

Program of this section

Baics of parameter estimation in both frequentist and bayesian,
explained on a simple linear fit.

Program of this section

Baics of parameter estimation in both frequentist and bayesian, explained on a simple linear fit.

The last *fundamental aspect* of this lecture is the notion of uncertainty of the parameter of interest.

Program of this section

Baics of parameter estimation in both frequentist and bayesian, explained on a simple linear fit.

The last *fundamental aspect* of this lecture is the notion of uncertainty of the parameter of interest.

1. **Frequentist**
   - coming back on the notion of estimator, again
   - Maximum likelihood (ML) and χ^2 estimators
   - uncertainty: *confidence interval*, notion of coverage

Program of this section

Baics of parameter estimation in both frequentist and bayesian,
explained on a simple linear fit.

The last *fundamental aspect* of this lecture is the notion of
uncertainty of the parameter of interest.

1. **Frequentist**
    - coming back on the notion of estimator, again
    - Maximum likelihood (ML) and χ^2 estimators
    - uncertainty: *confidence interval*, notion of coverage
2. **Bayesian**
    - from the posterior to the parameter of interest
    - uncertainty: *credibility interval*
    - impact of priors of parmater

Program of this section

Baics of parameter estimation in both frequentist and bayesian, explained on a simple linear fit.

The last *fundamental aspect* of this lecture is the notion of uncertainty of the parameter of interest.

1. **Frequentist**
   - coming back on the notion of estimator, again
   - Maximum likelihood (ML) and χ^2 estimators
   - uncertainty: *confidence interval*, notion of coverage
2. **Bayesian**
   - from the posterior to the parameter of interest
   - uncertainty: *credibility interval*
   - impact of priors of parmater
3. **Coming back on nuisance parameters** (*i.e.* uncertainties on the model)

Frequentist approach: estimators

Definition: random variable which gives a 'good' estimate of your parameter of interest ($\hat{\mu} = \frac{1}{N} \sum_i x_i$ as estimator of $\mathbb{E}[X]$). Estimator depends on observation $\hat{\mu}(x_1, ..., x_n)$ and is *not* constant. N_{meas} needed to assess its quality.

Frequentist approach: estimators

Definition: random variable which gives a 'good' estimate of your parameter of interest ($\hat{\mu} = \frac{1}{N}\sum_i x_i$ as estimator of $\mathbb{E}[X]$). Estimator depends on observation $\hat{\mu}(x_1, ..., x_n)$ and is *not* constant. N_{meas} needed to assess its quality.

Properties: when $N_{meas} \to \infty$

1. consistency: "$P(\hat{\mu} \neq \mu_{truth}) \to 0$" (rigorously: $P(|\hat{\mu} - \mu_{truth}| > \epsilon) \to 0, \forall \epsilon > 0$)

Frequentist approach: estimators

Definition: random variable which gives a 'good' estimate of your parameter of interest ($\hat{\mu} = \frac{1}{N}\sum_i x_i$ as estimator of $\mathbb{E}[X]$). Estimator depends on observation $\hat{\mu}(x_1, ..., x_n)$ and is *not* constant. N_{meas} needed to assess its quality.

Properties: when $N_{meas} \to \infty$

1. consistency: "$P(\hat{\mu} \neq \mu_{truth}) \to 0$" (rigorously: $P(|\hat{\mu} - \mu_{truth}| > \epsilon) \to 0, \forall \epsilon > 0$)
2. bias: $b \equiv \mathbb{E}[\hat{\mu}] - \mu_{truth} = 0$

Frequentist approach: estimators

Definition: random variable which gives a 'good' estimate of your parameter of interest ($\hat{\mu} = \frac{1}{N}\sum_i x_i$ as estimator of $\mathbb{E}[X]$). Estimator depends on observation $\hat{\mu}(x_1, ..., x_n)$ and is *not* constant. N_{meas} needed to assess its quality.

Properties: when $N_{meas} \to \infty$

1. consistency: "$P(\hat{\mu} \neq \mu_{truth}) \to 0$" (rigorously: $P(|\hat{\mu} - \mu_{truth}| > \epsilon) \to 0, \forall \epsilon > 0$)
2. bias: $b \equiv \mathbb{E}[\hat{\mu}] - \mu_{truth} = 0$
3. efficiency: smallest variance $v_{\hat{\mu}}$

Frequentist approach: estimators

Definition: random variable which gives a 'good' estimate of your parameter of interest ($\hat{\mu} = \frac{1}{N}\sum_i x_i$ as estimator of $\mathbb{E}[X]$). Estimator depends on observation $\hat{\mu}(x_1, ..., x_n)$ and is *not* constant. N_{meas} needed to assess its quality.

Properties: when $N_{meas} \to \infty$

1. consistency: "$P(\hat{\mu} \neq \mu_{truth}) \to 0$" (rigorously: $P(|\hat{\mu} - \mu_{truth}| > \epsilon) \to 0, \forall \epsilon > 0$)
2. bias: $b \equiv \mathbb{E}[\hat{\mu}] - \mu_{truth} = 0$
3. efficiency: smallest variance $v_{\hat{\mu}} \equiv$ Rao-Cramér-Fréchet (RCF) limit

$$v_{\hat{\mu}} \geq -\frac{\left(1 + \frac{\partial b}{\partial \mu}\right)^2}{\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \mu^2}\right]}$$

Frequentist approach: estimators

Definition: random variable which gives a 'good' estimate of your parameter of interest ($\hat{\mu} = \frac{1}{N}\sum_i x_i$ as estimator of $\mathbb{E}[X]$). Estimator depends on observation $\hat{\mu}(x_1, ..., x_n)$ and is *not* constant. N_{meas} needed to assess its quality.

Properties: when $N_{meas} \to \infty$

1. consistency: "$P(\hat{\mu} \neq \mu_{truth}) \to 0$" (rigorously: $P(|\hat{\mu} - \mu_{truth}| > \epsilon) \to 0, \forall \epsilon > 0$)
2. bias: $b \equiv \mathbb{E}[\hat{\mu}] - \mu_{truth} = 0$
3. efficiency: smallest variance $v_{\hat{\mu}} \equiv$ Rao-Cramér-Fréchet (RCF) limit

$$v_{\hat{\mu}} \geq -\frac{\left(1 + \frac{\partial b}{\partial \mu}\right)^2}{\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \mu^2}\right]}$$

Two important examples of estimators

1. Maximum likelihood estimator (MLE): $\hat{\mu}$ which maximizes $\mathcal{L}(\mu; x)$
   $\to$ numerically easier to minimze $-2\ln \mathcal{L}(\mu; x)$ - negative log likelihood (NLL)
2. χ^2 estimator: $\hat{\mu}$ which minimizes $\chi^2(\mu) \equiv \sum_i w_i \left(X_i^{pred}(\mu) - x_i\right)^2$

Frequentist approach: estimators

Definition: random variable which gives a 'good' estimate of your parameter of interest ($\hat{\mu} = \frac{1}{N} \sum_i x_i$ as estimator of $\mathbb{E}[X]$). Estimator depends on observation $\hat{\mu}(x_1, ..., x_n)$ and is *not* constant. N_{meas} needed to assess its quality.

Question 1 for the audience:

In frequentist, we sayed that the parameters are fixed (once chosen), while here were are talking about $P(\hat{\mu})$ or $\mathbb{E}[\hat{\mu}]$ ... So in the end, is there in frequentist a probability associated to the parameter or not?

Frequentist approach: estimators

Definition: random variable which gives a 'good' estimate of your parameter of interest ($\hat{\mu} = \frac{1}{N} \sum_i x_i$ as estimator of $\mathbb{E}[X]$). Estimator depends on observation $\hat{\mu}(x_1, ..., x_n)$ and is *not* constant. N_{meas} needed to assess its quality.

Question 1 for the audience:
In frequentist, we sayed that the parameters are fixed (once chosen), while here were are talking about $P(\hat{\mu})$ or $\mathbb{E}[\hat{\mu}]$ ... So in the end, is there in frequentist a probability associated to the parameter or not?

Question 2 for the audience:
Why consistency and bias of an estimator are different?

Example: linear fit

Model $N^{pred}(p_0, p_1; t) = p_0 + p_1 t$

4 estimators (or "cost function") are used:

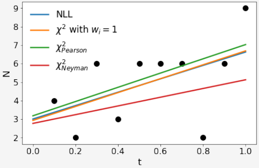
$$-2 \log \mathcal{L}_{poisson}$$

$$\chi^2(p_0, p_1) = \sum_i (N_i^{pred}(p_0, p_1) - N_i)^2$$

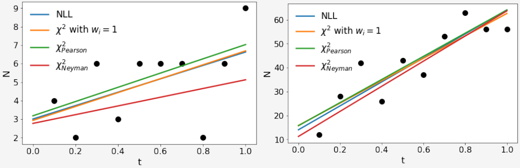
$$\chi^2_{Pearson}(p_0, p_1) = \sum_i \left(\frac{N_i^{pred}(p_0, p_1) - N_i}{\sqrt{N_i^{pred}(p_0, p_1)}} \right)^2$$

$$\chi^2_{Neyman}(p_0, p_1) = \sum_i \left(\frac{(N_i^{pred}(p_0, p_1) - N_i)^2}{\sqrt{N_i}} \right)^2$$

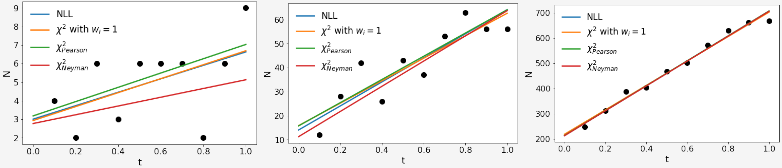
Example: linear fit



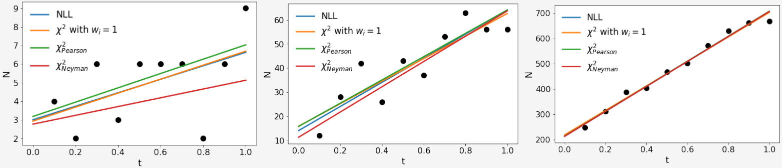
Example: linear fit



Example: linear fit



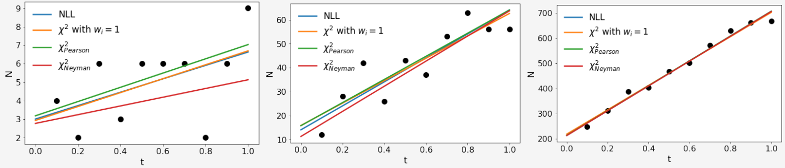
Example: linear fit



Comments:

- $\chi^2_{pearson} \equiv -2\log\mathcal{L}_{Gauss} \approx -2\log\mathcal{L}_{Poiss}$ for large numbers

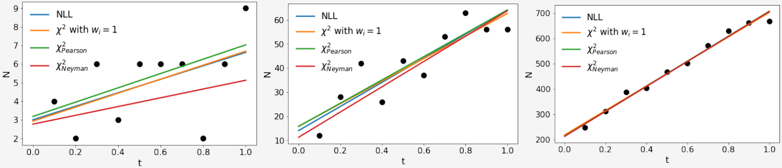
Example: linear fit



Comments:

- $\chi^2_{pearson} \equiv -2\log\mathcal{L}_{Gauss} \approx -2\log\mathcal{L}_{Poiss}$ for large numbers
- $\sqrt{N_i} \approx \sqrt{N_i^{pred}}$, justifing Neyman's approx (simpler to compute)

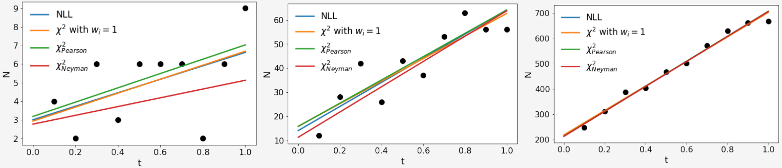
Example: linear fit



Comments:

- $\chi^2_{pearson} \equiv -2\log\mathcal{L}_{Gauss} \approx -2\log\mathcal{L}_{Poiss}$ for large numbers
- $\sqrt{N_i} \approx \sqrt{N_i^{pred}}$, justifing Neyman's approx (simpler to compute)
- Interpreting χ^2: distance, in unit of error, between data and model

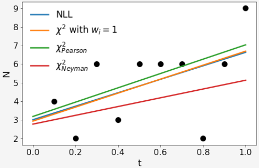
Example: linear fit

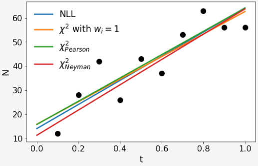


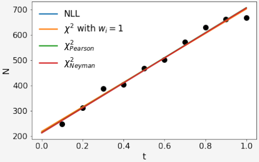
Comments:

- $\chi^2_{pearson} \equiv -2 \log \mathcal{L}_{Gauss} \approx -2 \log \mathcal{L}_{Poiss}$ for large numbers
- $\sqrt{N_i} \approx \sqrt{N_i^{pred}}$, justifing Neyman's approx (simpler to compute)
- Interpreting χ^2: distance, in unit of error, between data and model
- Doing a fit is always possible. Is the result statisfying?

Example: linear fit



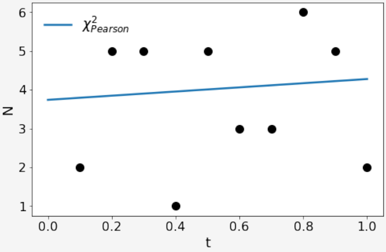


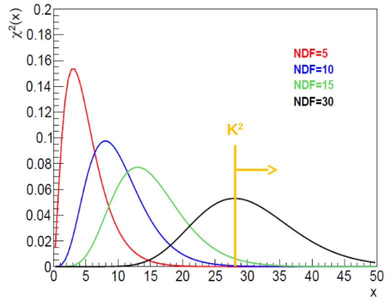


Comments:

- $\chi^2_{pearson} \equiv -2 \log \mathcal{L}_{Gauss} \approx -2 \log \mathcal{L}_{Poiss}$ for large numbers
- $\sqrt{N_i} \approx \sqrt{N_i^{pred}}$, justifing Neyman's approx (simpler to compute)
- Interpreting χ^2: distance, in unit of error, between data and model
- Doing a fit is always possible. Is the result statisfying?
  $\rightarrow$ goodness-of-fit is possible to evaluate since χ^2 PDF is known

The basics of goodness-of-fit





$\chi^2_{min} = 6.7$ with 10 data points ($nDoF = 10$) $\rightarrow$ blue PDF tells us this is a good fit, *even if* not a point is on the line.

We can actually compute the fraction of pseudo-data that would lead to a higher χ^2 (p-value), to quantify this statement.

Food for thought

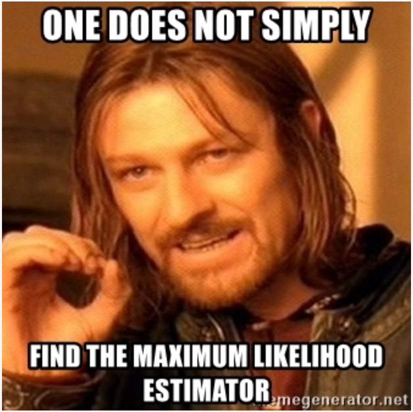
1. Perform a fit of an histogram in ROOT, with quite wide binning. Do you recover the true value? Does the result depends on the number of bins? How to solve it?

Food for thought

1. Perform a fit of an histogram in ROOT, with quite wide binning. Do you recover the true value? Does the result depends on the number of bins? How to solve it?

2. Imagine you have one dataset, but you want to fit simultaneously two distributions of these events. How to write the χ^2?

Frequentist parameter uncertainty



Frequentist parameter uncertainty

Confidence interval and level $\mu \in [\mu_{min}, \mu_{max}]$ @ α CL

- $\equiv$ the true value is in $[\mu_{min}, \mu_{max}]$ in α% of all possible realisations
- μ_{min} (μ_{max}) is the lower (upper) bound
- α is the confidence level
- μ_{min} and μ_{max} are random variables (as μ_{hat}): fluctuate with data

Frequentist parameter uncertainty

Confidence interval and level $\mu \in [\mu_{min}, \mu_{max}]$ @ α CL

- $\equiv$ the true value is in $[\mu_{min}, \mu_{max}]$ in α% of all possible realisations
- μ_{min} (μ_{max}) is the lower (upper) bound
- α is the confidence level
- μ_{min} and μ_{max} are random variables (as μ_{hat}): fluctuate with data

How to get confidence interval? Not trivial in general! Need approx

Frequentist parameter uncertainty

Confidence interval and level $\mu \in [\mu_{min}, \mu_{max}]$ @ α CL

- $\equiv$ the true value is in $[\mu_{min}, \mu_{max}]$ in α% of all possible realisations
- μ_{min} (μ_{max}) is the lower (upper) bound
- α is the confidence level
- μ_{min} and μ_{max} are random variables (as μ_{hat}): fluctuate with data

How to get confidence interval? Not trivial in general! Need approx

- simplest approx $\rightarrow$ use the variance of μ estimator:

$$\mu_{min/max} = \hat{\mu} \pm n\sqrt{v_{\hat{\mu}}}$$

Frequentist parameter uncertainty

Confidence interval and level $\mu \in [\mu_{min}, \mu_{max}]$ @ α CL

- $\equiv$ the true value is in $[\mu_{min}, \mu_{max}]$ in α% of all possible realisations
- μ_{min} (μ_{max}) is the lower (upper) bound
- α is the confidence level
- μ_{min} and μ_{max} are random variables (as μ_{hat}): fluctuate with data

How to get confidence interval? Not trivial in general! Need approx

- simplest approx $\rightarrow$ use the variance of μ estimator:

$$\mu_{min/max} = \hat{\mu} \pm n\sqrt{v_{\hat{\mu}}}$$

n is called "number of σ" and $\alpha(n)$ is known for a normal PDF:

- $\alpha(1) = 68\%$
- $\alpha(1.64) = 90\%$
- $\alpha(1.95) = 95\%$
- $\alpha(2) = 95.4\%$
- $\alpha(3) = 99.7\%$
- $\alpha(5) = 99.99994\%$

Frequentist parameter uncertainty

Quality of a given confidence interval

- CI $\equiv$ random variable: consider the limit of ∞ number of meas.
- Coverage $\equiv$ probability P that the true parameter *actually is* in C
- "Confidence level = what we target" while "coverage = what we get"

The 3 cases

1. $P = \alpha$: perfect coverage $\rightarrow$ ideal
2. $P > \alpha$: over-coverage $\rightarrow$ acceptable (conservative conclusions)
3. $P < \alpha$: under-coverage $\rightarrow$ dangerous (agressive conclusions)

Frequentist parameter uncertainty

Quality of a given confidence interval

- CI $\equiv$ random variable: consider the limit of ∞ number of meas.
- Coverage $\equiv$ probability P that the true parameter *actually is* in C
- "Confidence level = what we target" while "coverage = what we get"

The 3 cases

1. $P = \alpha$: perfect coverage $\rightarrow$ ideal
2. $P > \alpha$: over-coverage $\rightarrow$ acceptable (conservative conclusions)
3. $P < \alpha$: under-coverage $\rightarrow$ dangerous (agressive conclusions)

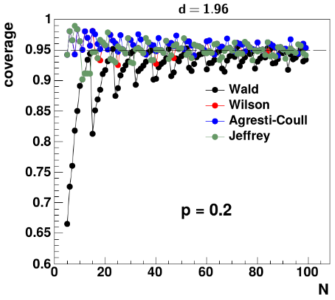
In practice: estimating coverage can be done using toys experiment (CPU-intensive for realistic models).

Example: binomial distribution, with parameter of interest p

$$P(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k}$$

$$\hat{p} = \frac{k}{N}$$

$$p \in \left[\hat{p} - d\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}; \hat{p} + d\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\right] \quad \text{(Wald interval)}$$



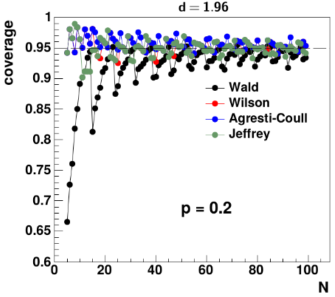
Frequentist parameter uncertainty

Example: binomial distribution, with parameter of interest p

$$P(k;N,p) = \binom{N}{k} p^k (1-p)^{N-k}$$

$$\hat{p} = \frac{k}{N}$$

$$p \in \left[\hat{p} - d\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}; \hat{p} + d\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\right] \quad \text{(Wald interval)}$$



Take away messages:

- notation $\mu = X^{+Y}_{-Z}$ (assuming 68% C.L.) is sometimes only indicative
- only object which contains the full information is likelihood
- OK to manipulate these approximate quanties - just know what they are(n't)

Bayesian parameter estimation

From the posterieur to the final value: given $f(\mu) \equiv P(\mu|data)$

Bayesian parameter estimation

From the posterieur to the final value: given $f(\mu) \equiv P(\mu|data)$

- few options for the central value
  - most probable value (MPV) or *mode*: $\hat{\mu}$ for which $f(\mu)$ is max

Bayesian parameter estimation

From the posterieur to the final value: given $f(\mu) \equiv P(\mu|data)$

- few options for the central value
    - most probable value (MPV) or *mode*: $\hat{\mu}$ for which $f(\mu)$ is max
    - mean: $\hat{\mu} = \int \mu f(\mu) \mathrm{d}\mu$

Bayesian parameter estimation

From the posterieur to the final value: given $f(\mu) \equiv P(\mu|data)$

- few options for the central value
  - most probable value (MPV) or *mode*: $\hat{\mu}$ for which $f(\mu)$ is max
  - mean: $\hat{\mu} = \int \mu f(\mu) \mathrm{d}\mu$
  - median: $\hat{\mu}$ such as $P(\mu > \hat{\mu}) = P(\mu < \hat{\mu}) = 1/2$

Bayesian parameter estimation

From the posterieur to the final value: given $f(\mu) \equiv P(\mu|data)$

- few options for the central value
  - most probable value (MPV) or *mode*: $\hat{\mu}$ for which $f(\mu)$ is max
  - mean: $\hat{\mu} = \int \mu f(\mu) \mathrm{d}\mu$
  - median: $\hat{\mu}$ such as $P(\mu > \hat{\mu}) = P(\mu < \hat{\mu}) = 1/2$
- few options for the **credibility** interval of **credibility** degree α
  - symetric around the mean: $[\mathbb{E}[\mu] - a, \mathbb{E}[\mu] + a]$, with

$$\int_{\mathbb{E}[\mu]-a}^{\mathbb{E}[\mu]+a} \mu f(\mu) \mathrm{d}\mu = \alpha$$

Bayesian parameter estimation

From the posterieur to the final value: given $f(\mu) \equiv P(\mu|data)$

- few options for the central value
  - most probable value (MPV) or *mode*: $\hat{\mu}$ for which $f(\mu)$ is max
  - mean: $\hat{\mu} = \int \mu f(\mu) \mathrm{d}\mu$
  - median: $\hat{\mu}$ such as $P(\mu > \hat{\mu}) = P(\mu < \hat{\mu}) = 1/2$

- few options for the **credibility** interval of **credibility** degree α
  - symetric around the mean: $[\mathbb{E}[\mu] - a, \mathbb{E}[\mu] + a]$, with

$$\int_{\mathbb{E}[\mu]-a}^{\mathbb{E}[\mu]+a} \mu f(\mu) \mathrm{d}\mu = \alpha$$

  - probability symetric around the mean $[a, b]$ such as

$$\int_{a}^{\mathbb{E}[\mu]} \mu f(\mu) \mathrm{d}\mu = \int_{\mathbb{E}[\mu]}^{b} \mu f(\mu) \mathrm{d}\mu = \alpha/2$$

Bayesian parameter estimation

From the posterieur to the final value: given $f(\mu) \equiv P(\mu|data)$

- few options for the central value
  - most probable value (MPV) or *mode*: $\hat{\mu}$ for which $f(\mu)$ is max
  - mean: $\hat{\mu} = \int \mu f(\mu)\mathrm{d}\mu$
  - median: $\hat{\mu}$ such as $P(\mu > \hat{\mu}) = P(\mu < \hat{\mu}) = 1/2$

- few options for the **credibility** interval of **credibility** degree α
  - symetric around the mean: $[\mathbb{E}[\mu] - a, \mathbb{E}[\mu] + a]$, with

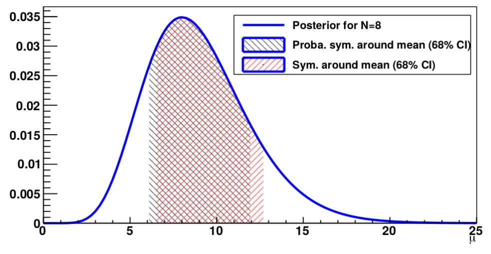
$$\int_{\mathbb{E}[\mu]-a}^{\mathbb{E}[\mu]+a} \mu f(\mu)\mathrm{d}\mu = \alpha$$

  - probability symetric around the mean $[a, b]$ such as

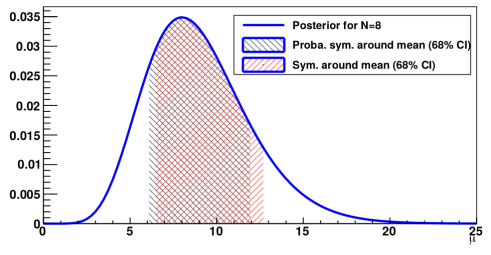
$$\int_{a}^{\mathbb{E}[\mu]} \mu f(\mu)\mathrm{d}\mu = \int_{\mathbb{E}[\mu]}^{b} \mu f(\mu)\mathrm{d}\mu = \alpha/2$$

  - Replace $\mathbb{E}[\mu]$ by the mode, or the median ...

Bayesian parameter estimation



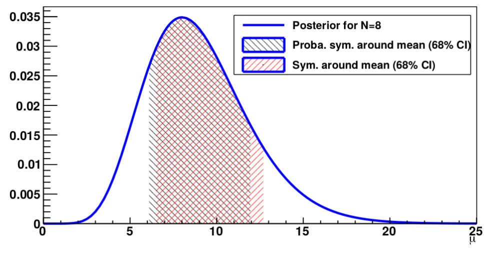
Bayesian parameter estimation



Take away messages:

- as in frequentist, the notation $\mu = X^{+Y}_{-Z}$ is sometimes only indicative
- the only object which contains the full information is the posterior

Bayesian parameter estimation



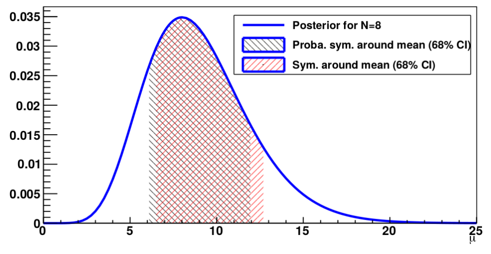
Take away messages:

- as in frequentist, the notation $\mu = X^{+Y}_{-Z}$ is sometimes only indicative
- the only object which contains the full information is the posterior

Few reminders

- impact of the prior decreases with the number of measurements
- frequentist $\approx$ bayesien with flat prior (numbers are $=$ but meaning is $\neq$)

Bayesian parameter estimation



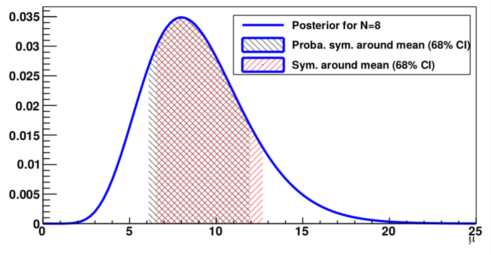
Take away messages:

- as in frequentist, the notation $\mu = X^{+Y}_{-Z}$ is sometimes only indicative
- the only object which contains the full information is the posterior

Few reminders

- impact of the prior decreases with the number of measurements
- frequentist $\approx$ bayesien with flat prior (numbers are $=$ but meaning is $\neq$)
- questions: (1) why there is no coverage in bayesian?

Bayesian parameter estimation



Take away messages:

- as in frequentist, the notation $\mu = X^{+Y}_{-Z}$ is sometimes only indicative
- the only object which contains the full information is the posterior

Few reminders

- impact of the prior decreases with the number of measurements
- frequentist $\approx$ bayesien with flat prior (numbers are $=$ but meaning is $\neq$)
- questions: (1) why there is no coverage in bayesian?
  (2) Why the 3 properties of frequentist estimator are defined in baysien?

Coming back to model uncertainties - I

Frequentist approach imagine you measure energy response r_E of a detector using a dedicated data d_E

- this measure is described by a likelihood $\mathcal{L}_{energy}(r_E, d_E)$
- the parameter of interest will be better known with more data
- this unknown can be added to the stat model using the full likelihood

$$\mathcal{L}(\mu, r_E; data, d_E) = \mathcal{L}(\mu, ; data)\mathcal{L}_{energy}(r_E, d_E)$$

- this is notion of auxiliary measurement.
- $\mathcal{L}_{energy}(r_E, d_E)$ is usally too complex to be implemented.
- One uses its approximation (Taylor Expension of order 2 of NLL around the min, leading to a gaussian likelihood)

Coming back to model uncertainties - II

Bayesian approach imagine you have a calculation with some approximations, to which an uncertainty is associated.

- this uncertainty is closer to a degree of beleif
- a prior $\pi(\theta)$ is required to quantify, were the true value of θ is more likely to be
- this unknown can be added to the stat model using the full likelihood

$$\mathcal{L}(\mu, \theta; \mathit{data}) = \mathcal{L}(\mu, ; \mathit{data})\, \pi(\theta)$$

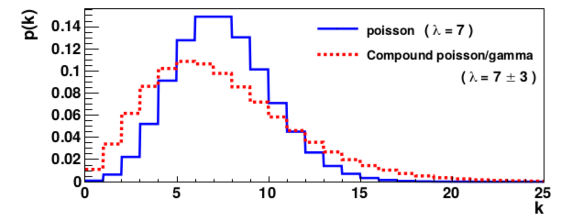
- this final likelihood is marginalized over θ:

$$\mathcal{L}_m(\mu; \mathit{data}) = \int \mathcal{L}(\mu, \theta; \mathit{data})\, \pi(\theta) \mathrm{d}\theta$$

- Interpretation: average all possible situations (defined by a θ value), accounting for the probability to actually have this value

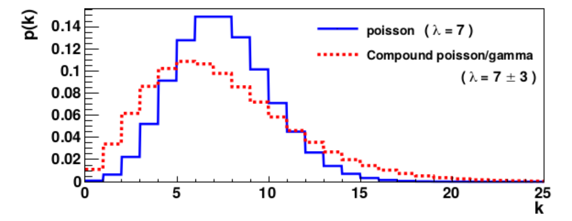
Coming back to model uncertainties - III

Example of marginalization



Coming back to model uncertainties - III

Example of marginalization



What's the proper way to implement uncertainties?

- no absolute answer to this question → arbitrariness
- make your choice depending on the context (ease interpretation or calculation, or ...?)
- always check the robustness of your conclusion wrt these choices

Test of Hypothesis

Why it is relevant

Most emblematic question: is there a signal in my data?

Test of Hypothesis

Why it is relevant

Most emblematic question: is there a signal in my data?

Formalism

- 2 hypothesis: H_1 =there is signal and H_0: there is no signal

$\rightarrow$ test statistics $t \equiv$ random variable, discrimating H_1 from H_0

Test of Hypothesis

Why it is relevant

Most emblematic question: is there a signal in my data?

Formalism

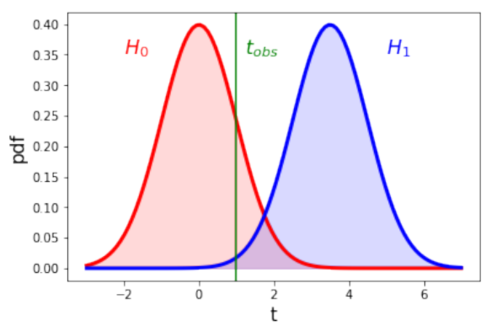
- 2 hypothesis: H_1 =there is signal and H_0: there is no signal

$\rightarrow$ test statistics $t \equiv$ random variable, discrimating H_1 from H_0

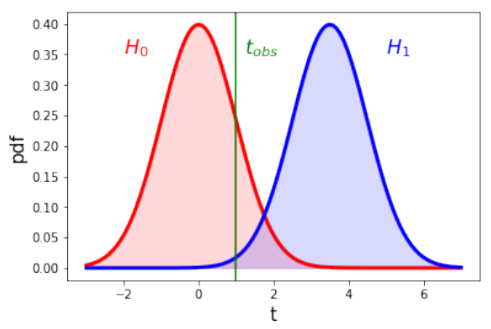
Most naive approch: event count as test statistics $t = N$

- *e.g.* H_1 predicts $N_1 = 110$, while H_0 predicts $N_1 = 100$
- observation $N_{obs} = 112$: do I reject the signal hypothesis?
- Steps of test hypothesis
  - find distribution of t in both hypothesis $f(t|H_0)$ and $f(t|H_1)$
  - check where t_{obs} fall wrt to $f(t|H_0)$ and $f(t|H_1)$
  - conclude with a confidence level ($p-$value)

Test of Hypothesis



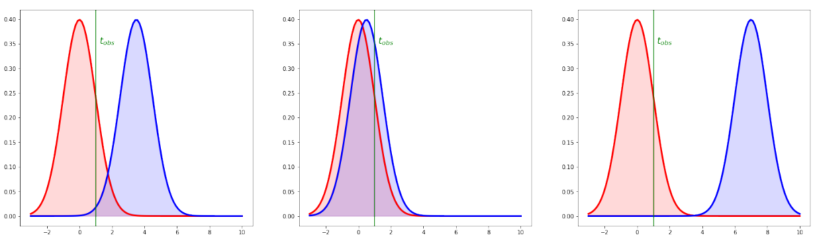
Test of Hypothesis



Quantitative agreement with an hypothsis: *p*-value

p-value = probability to observe what you observed in measurement or "more extreme" values

How to find exclusion limit



→ Increase the signal until the signal hypothesis get rejected (at a given confidence level).

Test of Hypothesis


Egon Pearson


Jerzy Neyman

Pearson-Neyman Lemma (1933)

- the most powerful statistical test is **N**egative **L**og **L**ikelihood ratio

$$NLL \equiv -2 \log \frac{\mathcal{L}(H_1|data)}{\mathcal{L}(H_0|data)}$$

Test of Hypothesis

Egon Pearson

Jerzy Neyman

Pearson-Neyman Lemma (1933)

- the most powerful statistical test is **N**egative **L**og **L**ikelihood ratio

$$NLL \equiv -2 \log \frac{\mathcal{L}(H_1|data)}{\mathcal{L}(H_0|data)}$$

→ an otpimal test statistics exists and we know it.

→ this always turns any n-dim problem into a 1-dim problem

e.g. imagine you have two event counts (N_1, N_2), instead of one N

Test of Hypothesis

Egon Pearson

Jerzy Neyman

Pearson-Neyman Lemma (1933)

- the most powerful statistical test is **N**egative **L**og **L**ikelihood ratio

$$NLL \equiv -2 \log \frac{\mathcal{L}(H_1|data)}{\mathcal{L}(H_0|data)}$$

$\rightarrow$ an otpimal test statistics exists and we know it.

$\rightarrow$ this always turns any n-dim problem into a 1-dim problem

e.g. imagine you have two event counts (N_1, N_2), instead of one N

In practice: hunders or thousands of event counts!

Keywords and concepts

Part I: statistics

descriptive statistics – sample – mean – (co)variance – (de)correlation

Part II: probability

Bias theorem – prior – posterior – random variable – (marginal) PDF – moments – caracteristic function – (in)dependent variables – CLT – error propagation

Part III: statistical model

Likelihood – nuisance parameter – parameter of interest – systematic uncertainties

Part IV: The two big school

Frequentist – occurence frequency – pseudo-data (toys) – bayesian – degree of belief

Part VI: Parameter estimation & hypothesis testing

estimator and its properties – χ^2 – confidence/credibility level/interval – coverage – p-value – LLR

Concluding remarks

Statistics deals with the 'not fully known'
$\rightarrow$ not a single way $\rightarrow$ some arbitrariness

1. Statistics $\equiv$ link between measurement and conclusion

Concluding remarks

Statistics deals with the 'not fully known'
$\rightarrow$ not a single way $\rightarrow$ some arbitrariness

1. Statistics $\equiv$ link between measurement and conclusion

2. Want to understand a method? Make sure to properly identify the question it addresses!

Concluding remarks

Statistics deals with the 'not fully known'
$\rightarrow$ not a single way $\rightarrow$ some arbitrariness

1. Statistics $\equiv$ link between measurement and conclusion

2. Want to understand a method? Make sure to properly identify the question it addresses!

3. Don't restrict yourself to one method/approach

Concluding remarks

Statistics deals with the 'not fully known'
$\rightarrow$ not a single way $\rightarrow$ some arbitrariness

1. Statistics $\equiv$ link between measurement and conclusion

2. Want to understand a method? Make sure to properly identify the question it addresses!

3. Don't restrict yourself to one method/approach

4. All these warnings, subtelties and arbitrariness don't matter any more when 'the peak is clear'

Concluding remarks

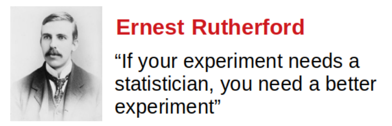
Statistics deals with the 'not fully known'
$\rightarrow$ not a single way $\rightarrow$ some arbitrariness

1. Statistics $\equiv$ link between measurement and conclusion

2. Want to understand a method? Make sure to properly identify the question it addresses!

3. Don't restrict yourself to one method/approach

4. All these warnings, subtelties and arbitrariness don't matter any more when 'the peak is clear'



Thanks for you attention !