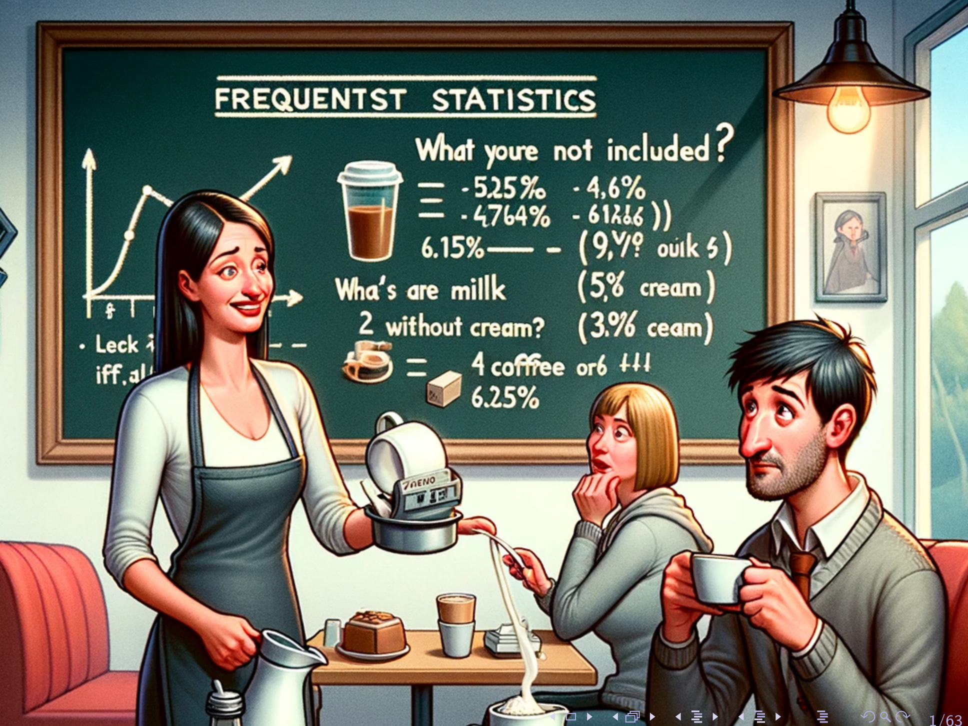

FREQUENTST STATISTICS
What youre not included?
-5.25% -4,6%
-4764%
6.15%
Wha's are milk
(5% cream)
2 without cream?
(3.% ceam)
4 coffee or6
6.25%

CHACAL 2024

Statistics for HEP

Lecture 1: intro & basic concepts

James Keaveney[1]

[1]james.keaveney@uct.ac.za

Jan 2024

UNIVERSITY OF CAPE TOWN

IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

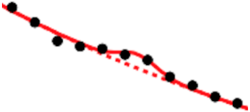
Who am I?

Details of these "lectures"

- 2×90 minute sessions - today and tomorrow
  - I will mainly talk from slides **but** I welcome interruptions/questions/discussion
  - pdf slides followed up with python notebook demos
  - accompanying *take-home* tutorial based on $H \to ZZ$ analysis with ATLAS OpenData
- Two guiding principles:
  - statistics is vast...far too much for 180 minutes.
  - there is always a wide range of prior knowledge experience
- Hence I concentrate on **fundamentals** will set you up well to understand domain specific techniques in your analysis later...
- I make no reference to experiment-specific tools or conventions...

Intro

- Statistics is
  - peculiar, counter-intuitive, often seems easier than it is
  - elusive: (you think you understand it, you realise you don't)^N
  - **fundamental to modern experimental particle physics**
- Incorrect statistical analysis can mean the difference between a discovery and not a discovery



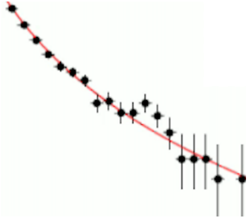


Figure: One of these *bumps* is a real discovery, the other is not...

Intro

- Statistics is
  - peculiar, counter-intuitive, often seems easier than it is
  - elusive: (you think you understand it, you realise you don't)^N
  - **fundamental to modern experimental particle physics**
- Incorrect statistical analysis can mean the difference between a discovery and not a discovery

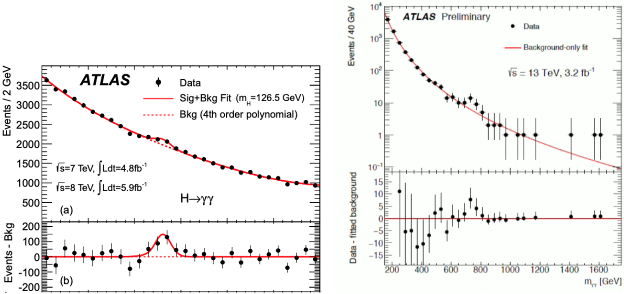
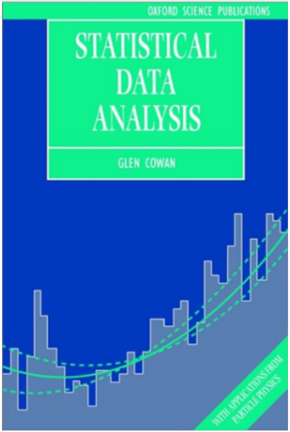


Figure: Discovery (left), Not a Discovery (right)

Intro

- Statistics is
  - elusive: (you think you understand it, you realise you don't)^N
- often need to refer back to textbooks...



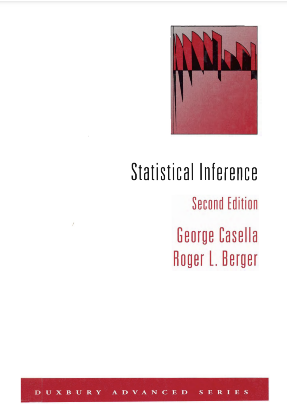
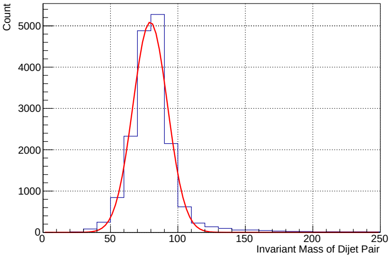


Figure: Books I love

Basic concepts: random variables

- **Results may vary** - the reason why the field of statistics exists.
- Results of repeated "identical", experiments may vary.
  - Instability in apparatus/environment/experimenter
  - Fundamental QM unpredictability of the system
- A variable is **random** when it cannot be predicted with absolute certainty

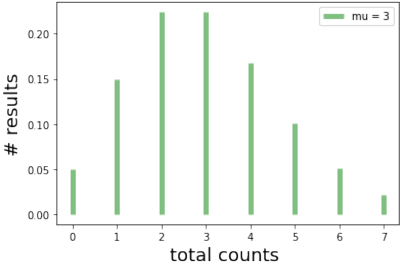


Basic concepts: statistical hypothesis

- A statistical hypothesis is a **formal** claim about a state of nature structured within the framework of a statistical model.
- To be useful and scientific, it should come with a prediction for an experiment so that the hypothesis can be tested.
- As **results may vary** the prediction should be **probabilistic**
- **Simple hypothesis** - everything about the prediction is specified

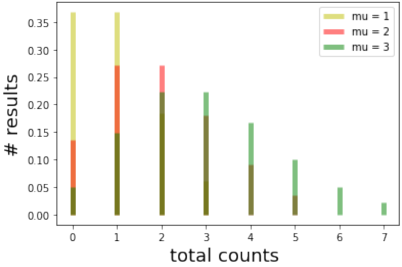
Basic concepts: simple statistical hypotheses

- **Simple hypothesis** everything about the prediction is specified
- Poisson hypothesis with fixed mean for a counting experiment



Basic concepts: composite statistical hypotheses

- **Composite hypothesis** not everything about the prediction is specified
- Poisson hypothesis with unknown mean for a counting experiment



Basic concepts: probability

- Statistics and Probability: two *definitions*
  - **Bayesian**: Given some data/evidence, we assign probability to some *hypothesis*, e.g. given this LHC data, how sure are we the Higgs boson exists?
  - **Frequentist**: Given some *hypothesis*, how likely is the data we observe, e.g. assuming the Higgs boson exists, how likely is the data that we observe?
- Frequentist approaches are more popular in particle physics
- I will mainly discuss frequentist ideas

Basic concepts: random variables and probability

- Frequentist Probability
  - interpreted as a **limiting frequency**??
- Imagine a *repeatable* experiment repeated n times, with S the set of all possible results
- A is a subset of possible results

$$P(A) = \lim_{n \to +\infty} \frac{N_{\text{result in A}}}{n}$$

- This definition satisfies the **3 axioms of probability**:
  1. $P(A) > 0$ for all A - probabilities can't be negative
  2. $\int_S P(A) = 1$ - *something* must happen
  3. For two mutually exclusive sets A and B, $(A \cap B = 0)$, $P(A \cup B) = P(A) + P(B)$.

Ice-breaker 1

- What does the *mean* mean?

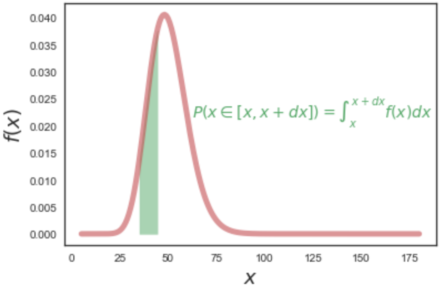
Python code / notebooks

- Code used to make the following plots (unless stated otherwise) available at link to github

Basic concepts: probability density functions (pdf)

- Imagine an experiment with all possible results characterised by a single continuous variable x
- S corresponds to the $(1D)$ space of all possible results
- What is the probability of observing a result in the interval $[x, x + dx]$?
    - given by $f(x)$ (pdf)

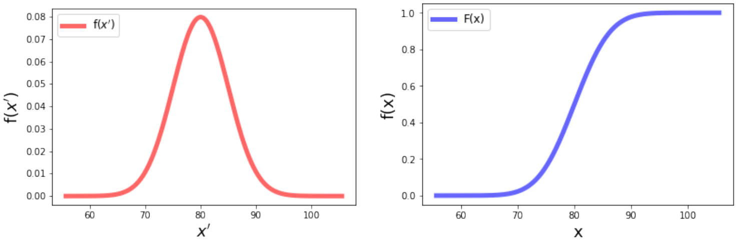
$$P(x \in [x, x + dx]) = \int_{x}^{x+dx} f(x)dx$$



Basic concepts: cumulative density functions (cdf)

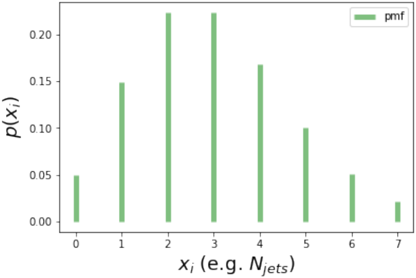
- cdf: $F(x)$
  - probability for x' to have a value $\leq x$

$$F(x) = \int_{-\infty}^{x} f(x) dx'$$



Basic concepts: probability mass function (pmf)

- If x can only assume discrete values (x_i), we use a *pmf* to describe its distribution
- pmf: $p(x_i) = P(x = x_i)$ where P is a probability.



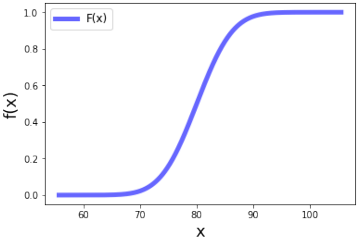
$$\sum_{x_i} p(x_i) = 1$$

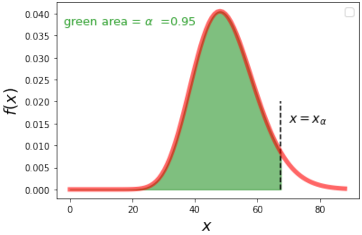
- Many examples of discrete observables in particle physics!

Basic concepts: quantiles

- the quantile x_α is the value of x such that $F(x_\alpha) = \alpha$
- simply the inverse of the cdf

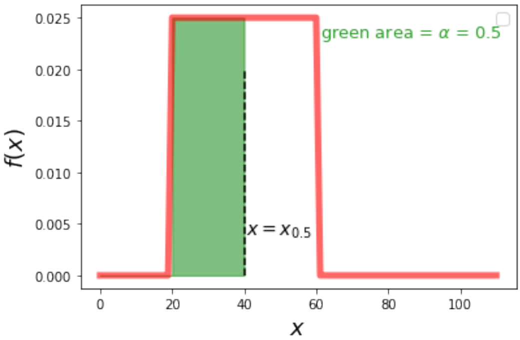
$$x_\alpha = F^{-1}(\alpha)$$





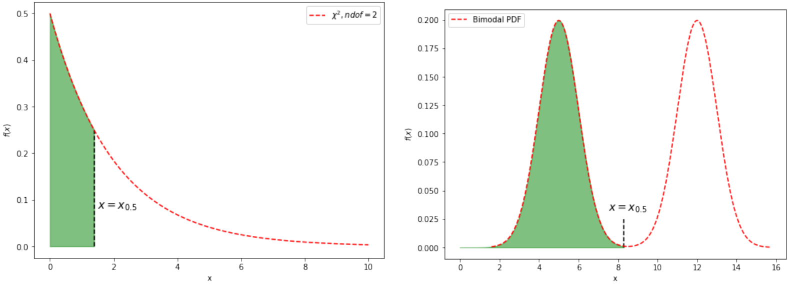
Basic concepts: median

- $x_{0.5}$ is a special case known as the **median**
- median often interpreted as the *typical location of x*
- when can this interpretation break down?



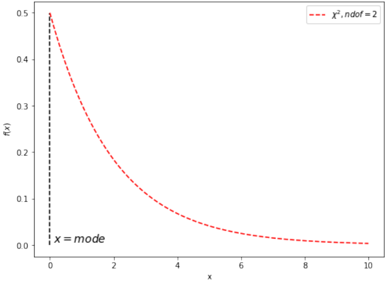
Basic concepts: median

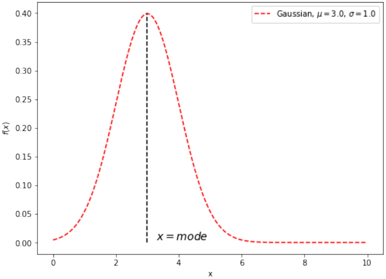
- The median often interpreted as the *typical location of x*
- when can this interpretation break down?



Basic concepts: mode

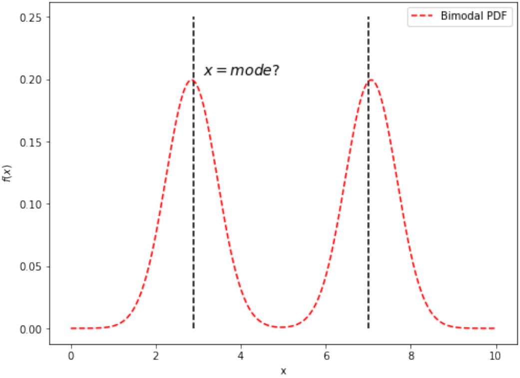
- The **mode** is the value of x for which $pdf(x)$ is maximal
  - The *typical location of the variable* is often better captured by the mode





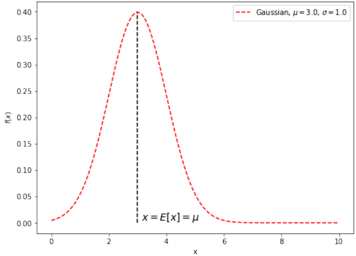
Basic concepts: mode

- mode is the value of x for which $pdf(x)$ is maximal
- when can this breakdown?



Basic concepts: expectation value

- The **expectation value** $E[x]$ of a variable x distributed according to $f(x)$ is often referred to as the **mean** μ.
- $E[x]$ is **not** a function of x, rather depends on form of $f(x)$.

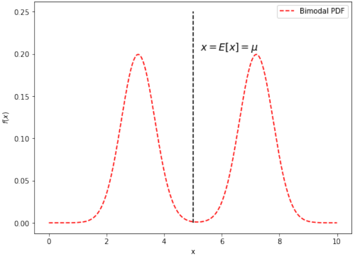


$$E[x] = \int_{-\infty}^{\infty} x.f(x)dx = \mu$$

- If the $f(x)$ is *concentrated in one region*, $E[x]$ represents a measure of where values of x are likely to be observed.
- **When can this interpretation break down?**

Basic concepts: expectation value

- What if $f(x)$ is *multimodal*?, e.g, two gaussian peaks



$$E[x] = \int_{-\infty}^{\infty} x.f(x)dx = \mu$$

- x is never equal to μ!

Basic concepts: variance

- Functions of x also have expectation values
  - e.g. squared difference between x and μ.
- $E[(x-\mu)^2]$ is called the **variance** V
  - V measures how *spread out* $f(x)$ is
  - Note $E[(x-\mu)^2] = E[x^2] - \mu^2$
- usually use the **standard deviation** σ instead
  - $\sigma = \sqrt{V}$

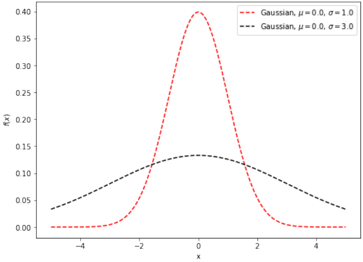


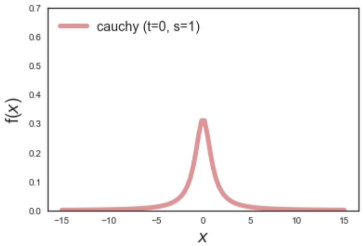
Figure: The two pdfs have the same μ but different σ

Basic concepts: mean & standard deviation limitations

- When the pdf has *fat tails*, μ and σ stop being useful
  - e.g. the Cauchy pdf

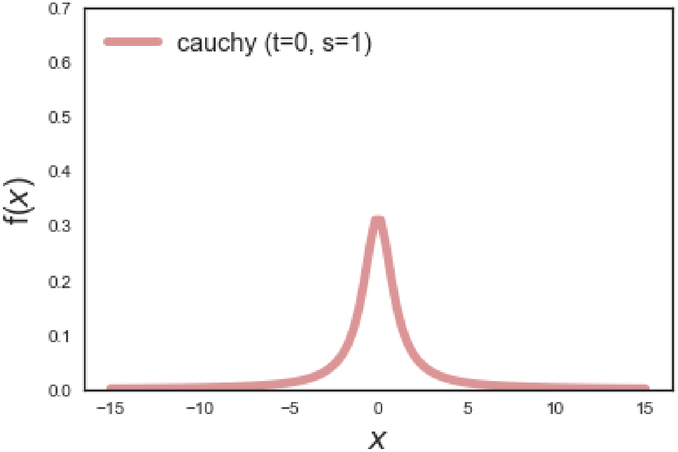
$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma}\left[\frac{\gamma^2}{(x - x_0)^2 + \gamma^2}\right],$$

- This pdf comes up a lot in physics
- $E[x]$ is undefined!
- $E[(x - \mu)^2]$ is undefined!



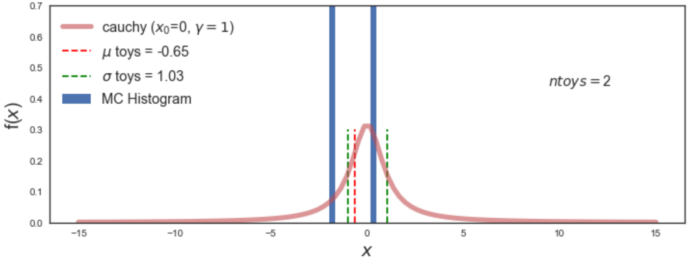
Basic concepts: mean & standard deviation limitations

- $E[x]$ is undefined!
- $E[(x - \mu)^2]$ is undefined!
- Taking the μ and σ of random numbers distributed according to a Cauchy does not work



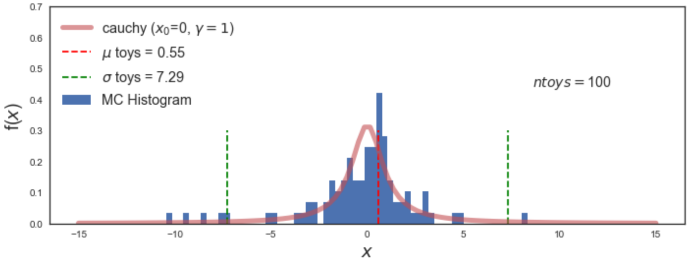
Basic concepts: mean & standard deviation limitations

- $E[x]$ is undefined!
- $E[(x-\mu)^2]$ is undefined!
- Taking the μ and σ of random numbers distributed according to a Cauchy does not work



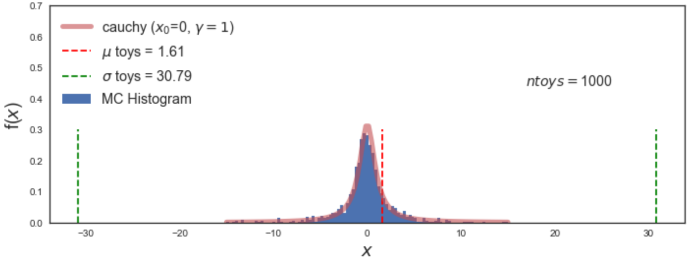
Basic concepts: mean & standard deviation limitations

- $E[x]$ is undefined!
- $E[(x - \mu)^2]$ is undefined!
- Taking the μ and σ of random numbers distributed according to a Cauchy does not work



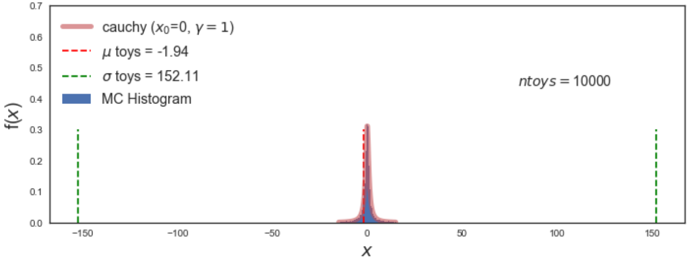
Basic concepts: mean & standard deviation limitations

- $E[x]$ is undefined!
- $E[(x-\mu)^2]$ is undefined!
- Taking the μ and σ of random numbers distributed according to a Cauchy does not work



Basic concepts: mean & standard deviation limitations

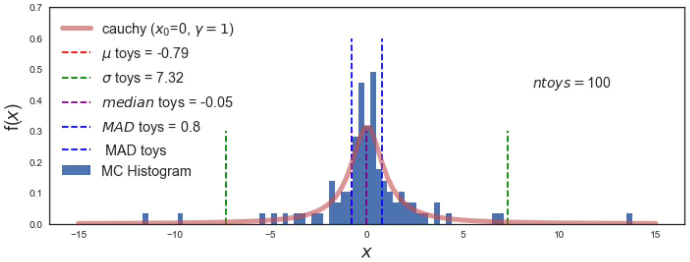
- $E[x]$ is undefined!
- $E[(x-\mu)^2]$ is undefined!
- Taking the μ and σ of random numbers distributed according to a Cauchy does not work



Basic concepts: alternatives: median and MAD

- If you suspect your data has fat tails, it's can be better to avoid the μ and σ
- Instead of μ how about the median?
- Instead of σ how about something MAD? (Mean Absolute Deviation)

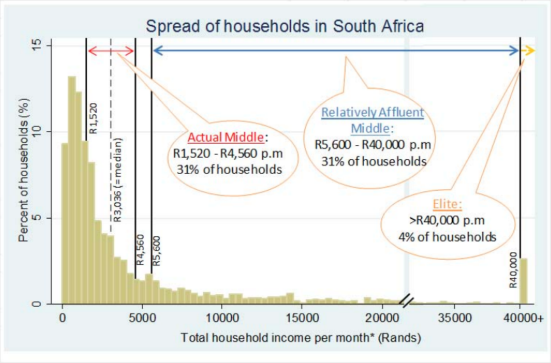
$$MAD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \mu(x)|$$



Basic concepts: alternatives: mode

- When does the median fail?

Figure 1: The spread of households within the income distribution in South Africa, 2008

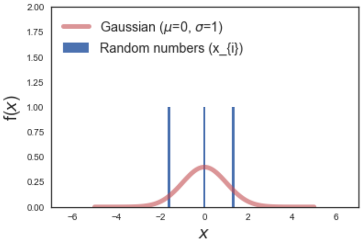


Source: NIDS 2008, own estimates

Figure: Source: *Who are the middle class in South Africa? Does it matter for policy? Visagie 2013*

Basic concepts: *random* numbers

- We have been talking about abstract notions of probability
  - but what about real data?
  - imagine some data x_i: n observations of some quantity x
  - what then is the μ and σ of x_i?

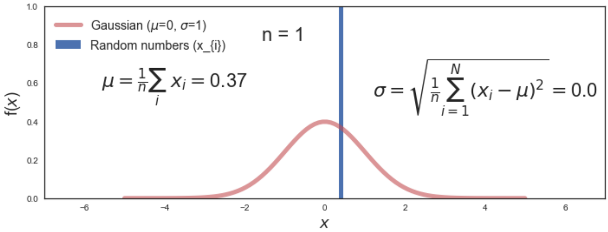


$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i,$ $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i - \mu)^2}$

- Let's think about how these definitions correspond to the defns. for pdfs

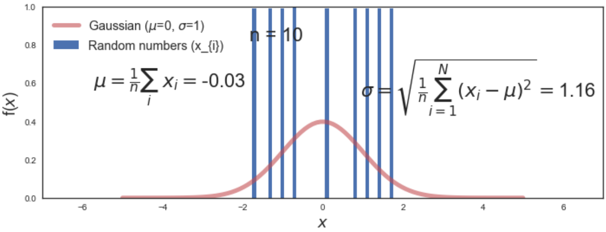
Basic concepts: *random* numbers

- *Random* numbers are useful in simulating data that is governed by a pdf
- Software tools can generate random numbers that are governed by any pdf...



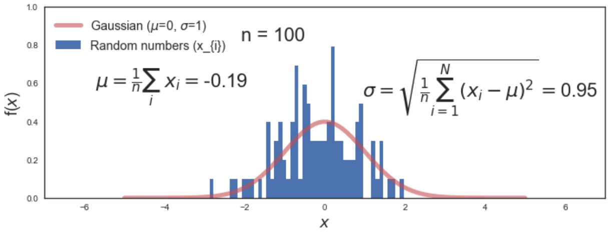
Basic concepts: *random* numbers

- *Random* numbers are useful in simulating data that is governed by a pdf



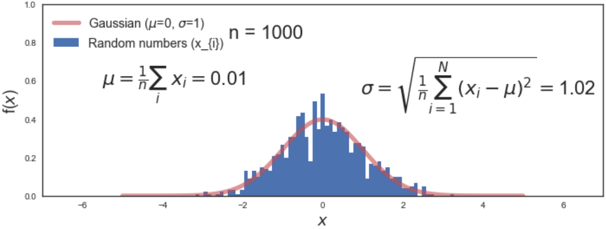
Basic concepts: *random* numbers

- *Random* numbers are useful in simulating data that is governed by a pdf



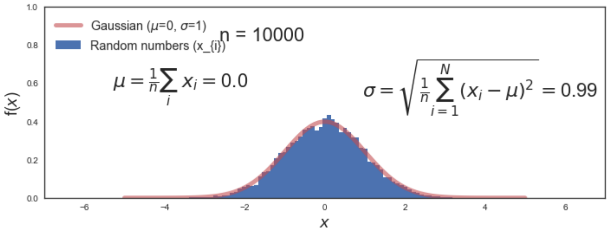
Basic concepts: *random* numbers

- *Random* numbers are useful in simulating data that is governed by a pdf



Basic concepts: *random* numbers

- *Random* numbers are useful in simulating data that is governed by a pdf



Basic concepts: *joint* pdf

- A *result* can correspond to more than one quantity, e.g, (x, y)
- **toy example**:
  - x and y both obey Gaussian pdfs
  - imagine each result as a point (x_i, y_i)

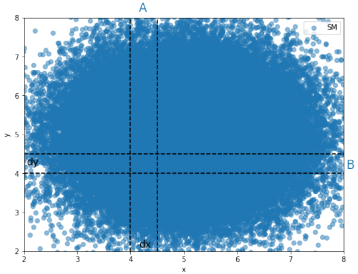


Figure: 5000 toy experiments with results (x_i, y_i) distributed as a 2-d Gaussian

- A = x observed in $[x, x + dx]$
- B = y observed in $[y, y + dy]$

$$P(A \cap B) = f(x, y)dxdy$$

Basic concepts: *joint* pdf

- pdf of multiple observables (x, y) is known as a **joint** pdf

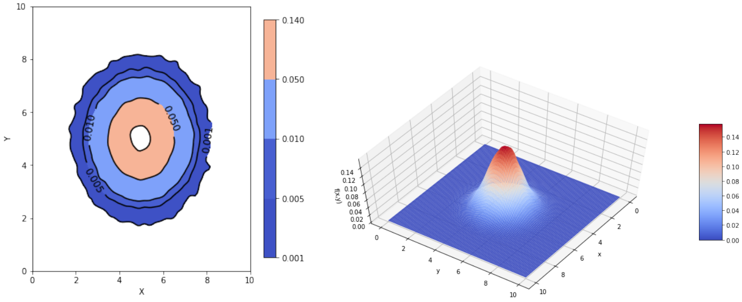
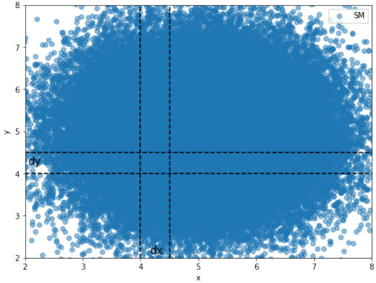


Figure: underlying pdf $f(x, y)$ of (x_i, y_i) dataset in 2- and 3-D

- $f(x, y)$ corresponds to the **density** of points **in the limit of infinite points**
- any experiment (x_i, y_i) must assume some value, one has the condition $\int \int f(x, y) dxdy = 1$

Basic concepts: *marginal* pdf

- If you know the joint pdf $f(x, y)$, you might want to know the pdf of x **regardless** of the value of y
  - this is given by the **marginal** pdf $f_x(x)$



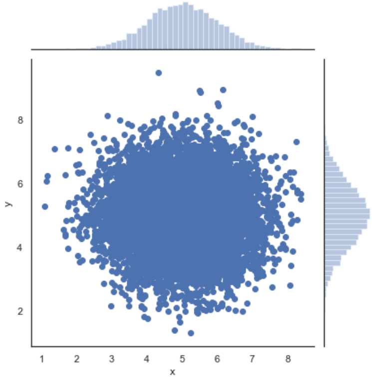
$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

similarly-

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

Basic concepts: *marginal* pdf

- If you know the joint pdf $f(x, y)$, you might want to know the pdf of x **regardless** of the value of y
  - this is given by the **marginal** pdf $f_x(x)$



$$\int_{-\infty}^{\infty} f_x(x)dx = 1$$

similarly-

$$\int_{-\infty}^{\infty} f_y(y)dy = 1$$

Basic concepts: *conditional* probability I

- What if you want to know the pdf of x **but you do care** about the value of y?
- **conditional probability:**
  - probability for y to be in $[y, y + dy]$ (B) with any x given that x is in $[x, x + dx]$ with any y (A)
  - usually referred to as $P(B|A)$, "probability of B given "A"

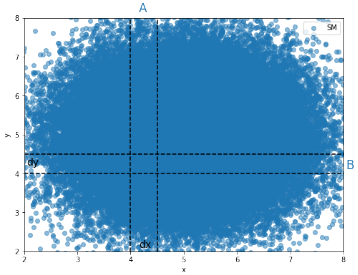


Figure: 5000 toy experiments with results (x_i, y_i) distributed as a 2-d Gaussian

- A = x observed in $[x, x + dx]$
- B = y observed in $[y, y + dy]$

$$P(A \cap B) = f(x, y)dxdy$$

Basic concepts: *conditional* probability II

- What if you want to know the pdf of x **but you do care** about the value of y?
- **conditional probability:**
  - probability for y to be in $[y, y+dy]$ (B) with any x given that x is in $[x, x+dx]$ with any y (A)
  - usually referred to as $P(B|A)$, "probability of B given "A"

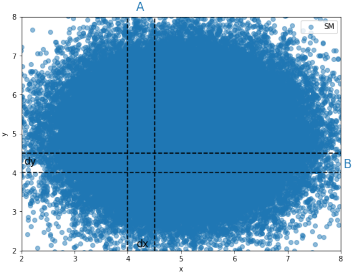


Figure: 5000 toy experiments with results (x_i, y_i) distributed as a 2-d Gaussian

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{f(x,y)dxdy}{f_x(x)dx}$$

Basic concepts: covariance

- Often a result corresponds to multiple quantities, e.g., x and y
- The **covariance** of x and y (V_{xy}) is defined as

$$V_{xy} = E[(x - \mu_x)(y - \mu_y)] = E[xy] - E[x]E[y]$$

- Suppose
  - x being **greater** than μ_x increases the probability to find y **greater** than μ_y
  - x being **less** than μ_x increases the probability to have y **less** than μ_y.
- Then $V_{xy} > 0$, and the variables are said to be **positively correlated** or just "correlated".

Basic concepts: covariance

- Often a result corresponds to multiple quantities, e.g., x and y
- The **covariance** of x and y (V_{xy}) is defined as

$$V_{xy} = E[(x-\mu_x)(y-\mu_y)] = E[xy] - E[x]E[y]$$

- Suppose
  - x being **greater** than μ_x increases the probability to find y **less** than μ_y
  - x being **less** than μ_x increases the probability to have y **greater** than μ_y.
- Then $V_{xy} < 0$, and the variables are said to be **negatively correlated** or anti-correlated.

Basic concepts: linear correlation coefficient

- One often thinks of the dimensionless **correlation coefficient** or "correlation"

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}$$

- correlation coefficient is covariance divided by the product of the standard deviations ($-1.0 < \rho_{xy} < 1.0$)

Basic concepts: linear correlation coefficient

- often don't know the pdf of (x, y) but instead have a sample of N measurements
- we define r as the **sample correlation coefficient** by inserting estimates of V_x, V_y and V_{xy} into the formula for ρ_{xy}
- Recall: $V_{xy} = E[xy] - E[x]E[y]$

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}$$

$$r_{xy} = \frac{(1/n)\sum_n x_i y_i - (\mu_x \mu_y)}{\sqrt{(1/n)\sum (x_i - \mu_x)^2}\ \sqrt{(1/n)\sum (y_i - \mu_y)^2}}$$

Basic concepts: linear correlation coefficient

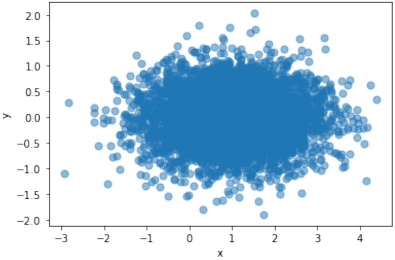
- often don't know the pdf of (x, y) but instead have a sample of N measurements
- we define r as the **sample correlation coefficient** by inserting estimates of V_x, V_y and V_{xy} into the formula for ρ_{xy}
- Recall: $V_{xy} = E[xy] - E[x]E[y]$

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}$$

$$r_{xy} = \frac{\sum_n x_i y_i - (\mu_x \mu_y)}{\sqrt{\sum(x_i - \mu_x)^2} \sqrt{\sum(y_i - \mu_y)^2}}$$

Basic concepts: correlation coefficient examples

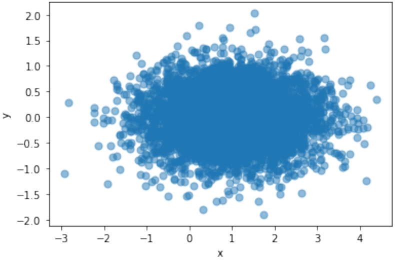
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} = ?$

Basic concepts: correlation coefficient examples

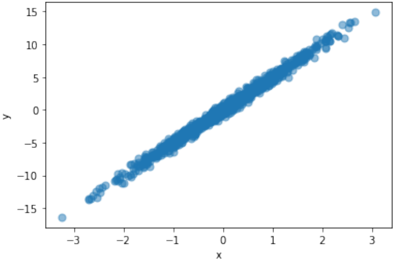
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} \approx 0.0$

Basic concepts: correlation coefficient examples

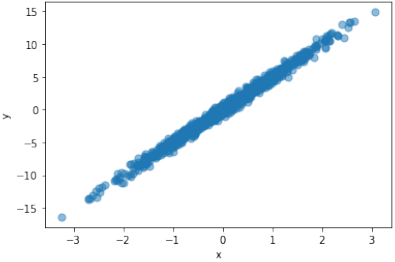
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} = ?$

Basic concepts: correlation coefficient examples

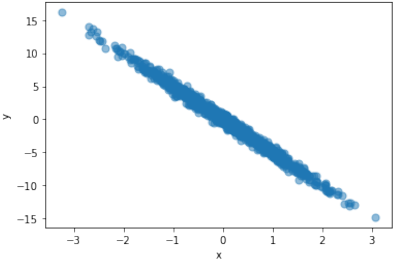
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} \approx 1.0$

Basic concepts: correlation coefficient examples

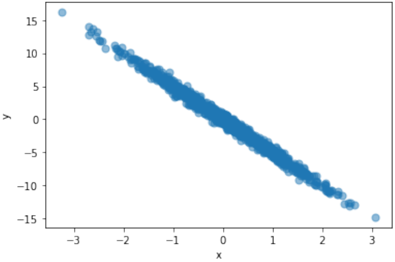
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} = ?$

Basic concepts: correlation coefficient examples

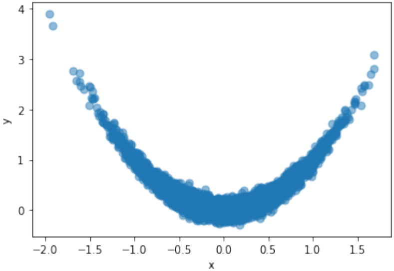
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} \approx -1.0$

Basic concepts: correlation coefficient examples

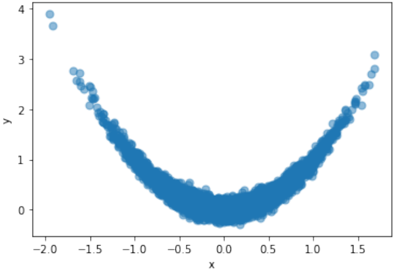
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} = ???$

Basic concepts: correlation coefficient examples

- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} \approx 0.0$!!!
- x and y are clearly related, but have r_{xy} vanishes due to the symmetry of $f(x, y)$ about 0
- shows the limitation of considering r_{xy} only

Basic concepts: mutual information

- The *mutual information*, $I(x; y)$, captures the inter-dependence of variables much better

$$\mathrm{I}(x; y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P(x, y) \log \left(\frac{P(x, y)}{P(x)\, P(y)} \right)$$

Basic concepts: mutual information

Basic concepts: correlation $!=$ causation

- Just because x and y have $r_{xy} > 0$, it doesn't guarantee that changes in x **cause** changes in y

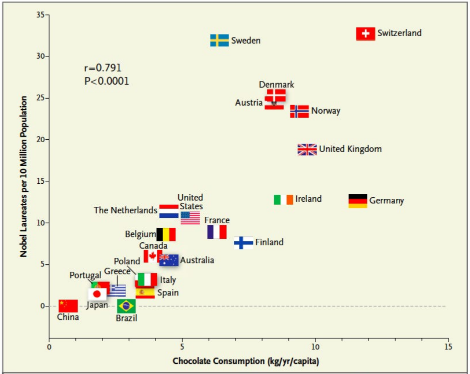
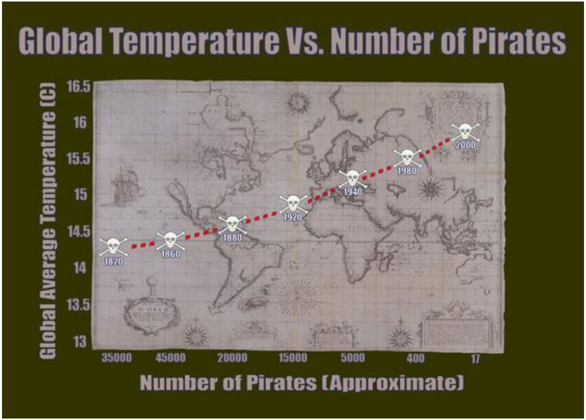


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

- Should we eat more chocolate?
- Unfortunately (probably) not.

Basic concepts: correlation $!=$ causation

- Just because x and y have $r_{xy} > 0$, it doesn't guarantee that changes in x **cause** changes in y



- Should we bring back pirates?
- Unfortunately (probably) not.