The Data Science of Particle Physics

basic concepts II

James Keaveney[1]

[1]james.keaveney@uct.ac.za
Room 5.05, RW James

July 2024

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)

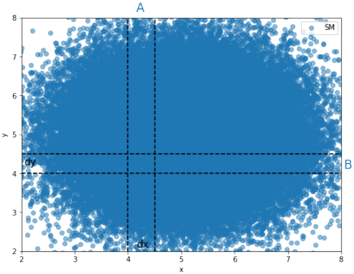


Figure: 5000 toy experiments with results (x_i, y_i) distributed as a 2-d Gaussian

- 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

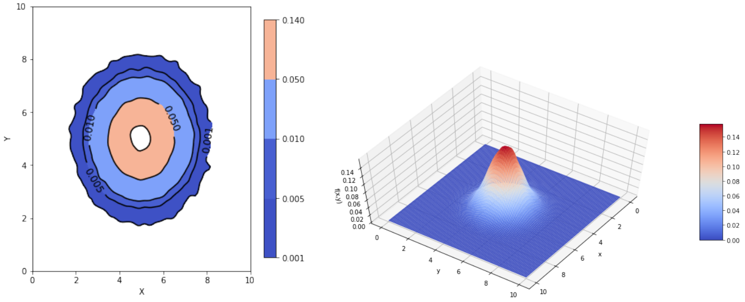
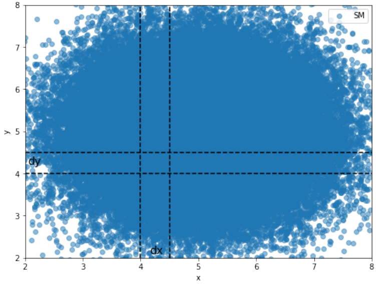


Figure: underlying pdf $f(x, y)$ of (x_i, y_i) dataset in 2- and 3-D

- $f(x, y)$ 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 $f_x(x)$



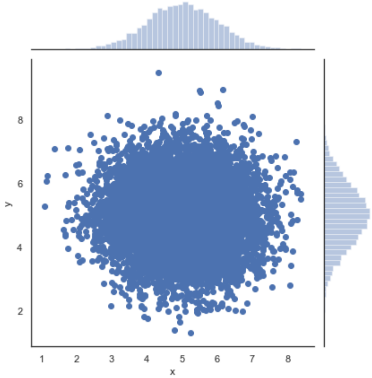
$$f_x(x) = \int_{-\infty}^{\infty} f(x, y)dy$$

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"

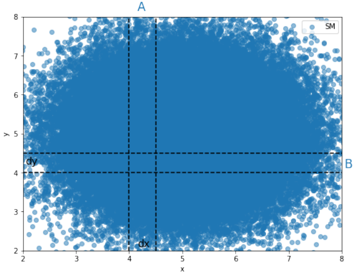


Figure: 5000 toy experiments with results (x_i, y_i) distributed as a 2-d Gaussian

- 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"

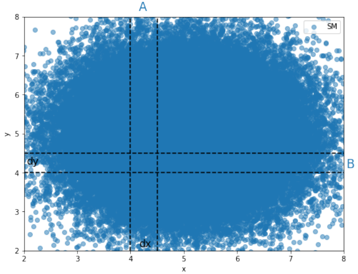


Figure: 5000 toy experiments with results (x_i, y_i) distributed as a 2-d Gaussian

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

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

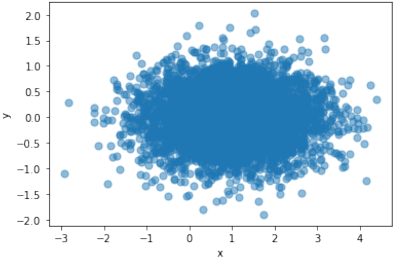
- often don't know the pdf of (x, y) but instead have a sample of N measurements
- we define r as the **sample correlation coefficient** by inserting estimates of V_x, V_y and V_{xy} into the formula for ρ_{xy}
- Recall: $V_{xy} = E[xy] - E[x]E[y]$

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

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

Basic concepts: correlation coefficient examples

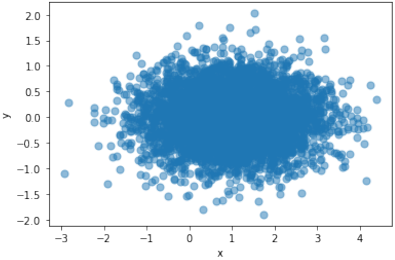
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} = ?$

Basic concepts: correlation coefficient examples

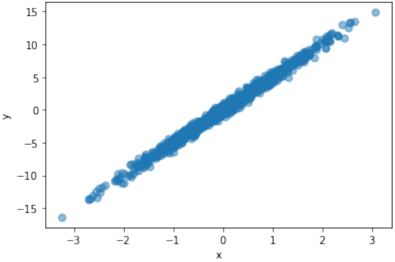
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} \approx 0.0$

Basic concepts: correlation coefficient examples

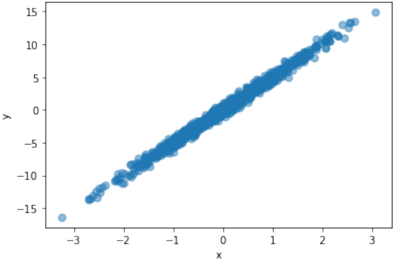
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} = ?$

Basic concepts: correlation coefficient examples

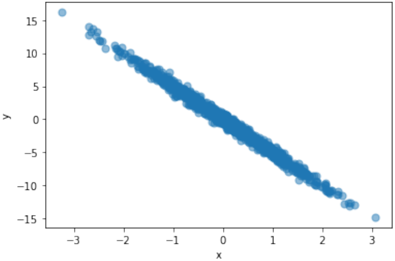
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} \approx 1.0$

Basic concepts: correlation coefficient examples

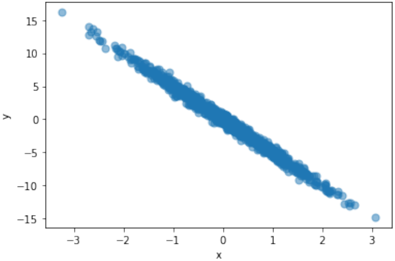
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} = ?$

Basic concepts: correlation coefficient examples

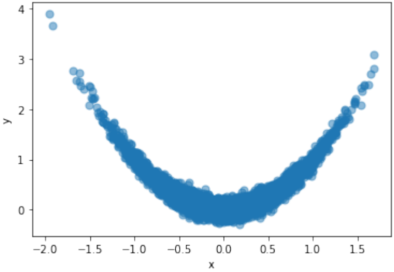
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} \approx -1.0$

Basic concepts: correlation coefficient examples

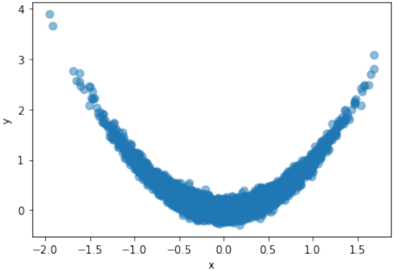
- Testing our intuition about r_{xy}
  - Generate N random (x, y) points according to some $pdf(x, y)$
  - We can calculate r_{xy} and compare to expectation from scatter plot of x and y



- $r_{xy} = ???$

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

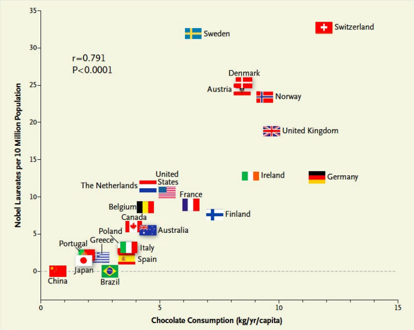
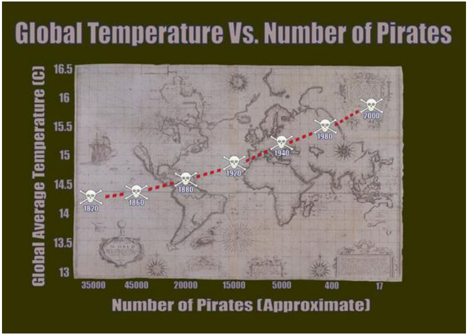


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

- 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.