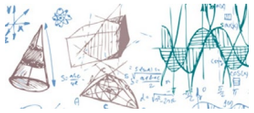
Introduction to probability, statistics and data handling

Agnieszka Obłąkowska-Mucha
Tomasz Szumlak

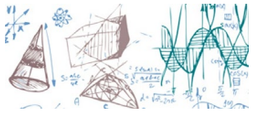
Faculty of Physics and Applied Computer Science
AGH University of Krakow



Last week recap

- **Probability** plays an important role in science, engineering and every day life!
- To answer a question we need to follow a specific pattern:
  - ✓ Define the topic of interest
  - ✓ Design an experiment
  - ✓ Collect data
  - ✓ Choose mathematical representation of the topic of interest (e.g., I want to measure a typical weight of people in large cities)
  - ✓ Analyse the data and make a statement

Random variables (II)

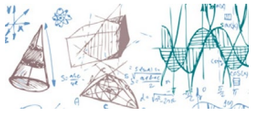


- There is no surprise, we can have either discrete or continuous RV
- Now, let's have a discrete RV X that can assume the following values: $X = \{x_1, x_2, \dots, x_n\}$. Suppose, these values are assumed with certain probabilities:

$$p(X = x_i) = f(x_i), i = 1, 2, \dots, n$$

- We can introduce **probability function**, that we call **probability distribution** for RV X
- In general, any function can be a probability function if:
  - Its values are always positive: $\boldsymbol{f(x) \geq 0 \vee x \in X \subset \Omega}$
  - The sum taken over all possible x_i is: $\boldsymbol{\sum_X f(x) = 1}$
- It is easy to extend all of this to RVs that are continuous, so we will not do that here (in principle we should remember that the sum changes into the integral)

Random variables (III)



Ex. 2 Again, let's look at the double coin toss. How do we define the probability distribution function (P.D.F.)?

The sample space $\Omega = \{HH, HT, TH, TT\}$, each of these events has the same probability $p(HH) = p(HT) = \cdots = 1/4$

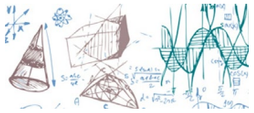
Using the Ex. 1 we can write:

$$p(X = 0) = p(TT) = 1/4$$

$$p(X = 1) = p(HT \cap TH) = 1/2$$

$$p(X = 2) = p(HH) = 1/4$$

x	0	1	2
$f(x)$	$1/4$	$1/2$	$1/4$



Distribution function (I)

- Closely related to P.D.F. is the cumulative distribution function or just distribution function (DF)
- We define it as follow:

$$F(x) = p(X \leq x)$$

- The DF has the following properties:
  - ✓ $F(x)$ must be non-decreasing

$$F(x_i) \leq F(x_j) \rightarrow \ x_i \leq x_j$$

  - ✓ Asymptotic behaviour

$$\lim_{x \to -\infty} F(x) = 0 \, , \lim_{x \to \infty} F(x) = 1$$

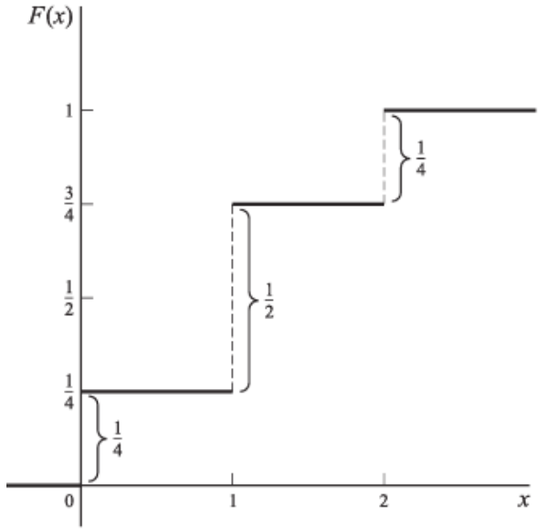
  - ✓ DF is continuous from the right

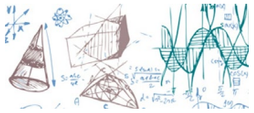
$$\lim_{h \to 0^+} F(x + h) = F(x), \ \vee \, x \in \mathrm{X}$$

Distribution function (II)

Ex. 4 Again, taking the two tosses example, we can work out the distribution function.

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ \frac{1}{4} & 0 \le x < 1 \\ \frac{3}{4} & 1 \le x < 2 \\ 1 & 2 \le x < \infty \end{cases}$$





Continuous RV

- There is a natural extension to continuous RV, however the exact definition is based on the properties of the distribution function
- **Def.1** We say, that a non-discrete random variable X is continous if its distribution function may be represented as:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du\,, (-\infty < x < \infty)$$

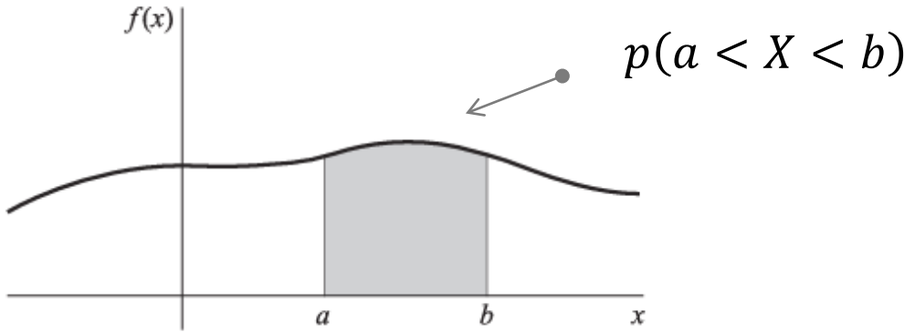
- We know already, that the function $f(x)$ should represent a P.D.F:

$$f(x) \geq 0 \qquad \int_{-\infty}^{\infty} f(x)dx = 1$$

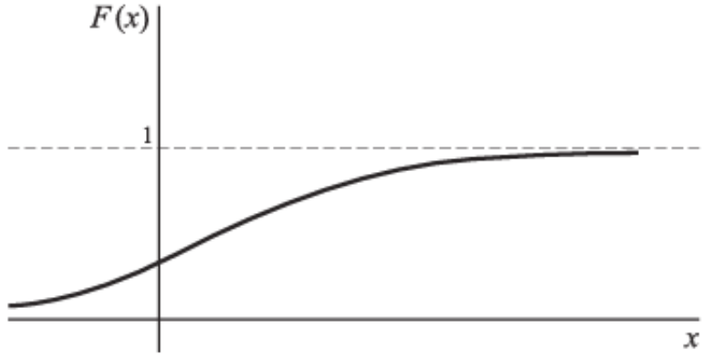
- There are some interesting properties related to the CRV
  - The probability that X takes on any one particular value is zero!
  - The interval probability can be estimated as: $p(a < X < b) = \int_{a}^{b} f(x)dx$

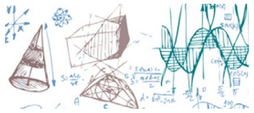
Graphical interpretation

- Let f(x) be the **density function** for a random variable X. This function can be represented as a graph by some curve



- The **distribution function** is, in turn, a monotonically increasing function which value goes from 0 to 1





Joint distributions (I)

- If X and Y are two discrete RVs, we can define P.D.F. of X and Y as follow:

$$p(X = x, Y = y) = f(x, y)$$

$$f(x, y) \geq 0 \qquad \sum_x \sum_y f(x, y) = 1$$

- If we assume that: $X = \{x_1, x_2, \dots, x_m\}$ and $Y = \{y_1, y_2, \dots, y_n\}$, then the probability of the event that $X = x_i$ and $Y = y_j$ is given by:

$$p(X = x_i, Y = y_j) = f(x_i, y_j)$$

- Respective probabilities for $X = x_i$ and $Y = y_j$ are given by:

$$p(X = x_i) = f_1(x_i) = f_x(x_i) = \sum_k f(x_i, y_k)$$

$$p(Y = y_j) = f_2(y_j) = f_y(y_j) = \sum_l f(x_l, y_j)$$

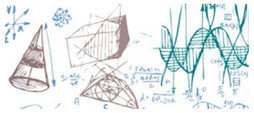
Joint distributions (II)

X \ Y	y_1	y_2	...	y_n	Totals ↓
x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	...	$f(x_1, y_n)$	$f_1(x_1)$
x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	...	$f(x_2, y_n)$	$f_1(x_2)$
⋮	⋮	⋮		⋮	⋮
x_m	$f(x_m, y_1)$	$f(x_m, y_2)$	...	$f(x_m, y_n)$	$f_1(x_m)$
Totals →	$f_2(y_1)$	$f_2(y_2)$	...	$f_2(y_n)$	1 ← Grand Total

Right Margin

Bottom Margin

Because the respective probabilities: $p(X = x_i)$ and $p(Y = y_j)$ are found on the margins of the joint probability table, we call both functions $f_1(x_i)$ and $f_2(y_j)$ the **marginal probability functions** of X and Y



Joint distributions (III)

- It is essential to note that for both marginal density functions we have:

$$\sum_i f_1(x_i) = 1, \sum_j f_2(y_j) = 1$$

- The two above are equivalent of:

$$\sum_i \sum_j f(x_i, y_j) = 1$$

- Since all of these functions represent P.D.F. they must be normalised, or in other words the probability of all entries is 1
- The **joint distribution function** of RVs X and Y is given by

$$F(x, y) = p(X \leq x, Y \leq y) = \sum_{u \leq x} \sum_{v \leq y} f(u, v)$$

- So, to get a value of $F(x, y)$ for a given pair (x, y) we need to sum-up all the entries for which $x_i \leq x$ and $y_j \leq y$

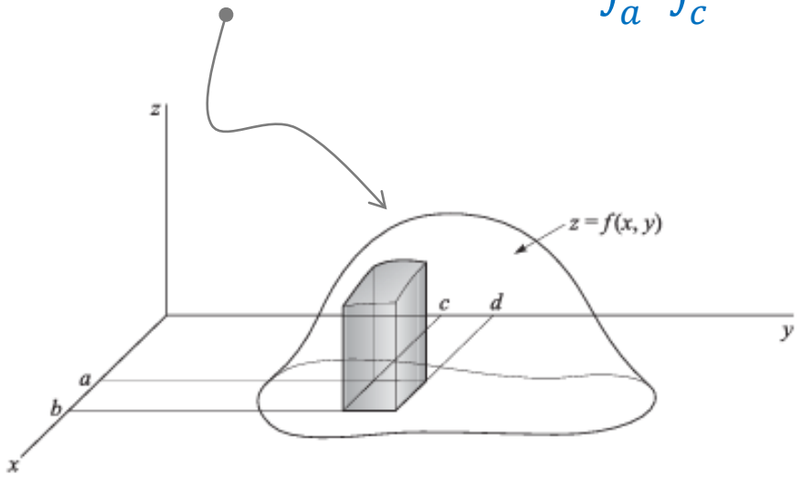
Joint distributions (IV)

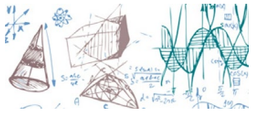
- Again, by analogy we can easily obtain the joint probability function for continuous RVs X and Y:

$$f(x,y) \geq 0, \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

- The probability can be estimated using the joint P.D.F. as follow:

$$p(a < X < b, c < Y < d) = \int_{a}^{b}\int_{c}^{d} f(x,y)dxdy$$





Joint distributions (V)

- Now, we can define the joint distribution function:

$$F(x,y) = p(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv$$

- We also have the following: $\dfrac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y)$

- By analogy, the respective marginal functions can be defined for both density, $f(x,y)$, and distribution, $F(x,y)$ functions

$$f_1(x) = \int_{-\infty}^{\infty} f(x,v)dv\,, f_2(y) = \int_{-\infty}^{\infty} f(u,y)du$$

$$F_1(x) = p(X \leq x) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(u,v)dudv \qquad F_2(y) = p(Y \leq y) = \int_{-\infty}^{\infty}\int_{-\infty}^{y} f(u,v)dudv$$

Joint distributions (VI)

Example: Students in a class of 100 were classified according to gender (G) and smoking (S) as follows:

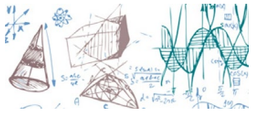
		S			
		s	q	n	
G	male	20	32	8	60
	female	10	5	25	40
		30	37	33	100

s: “now smokes”,
q: “did smoke but quit”
n: “never smoked”.

Identify:
- joint distribution function
- marginal distributions

Find the probability that a randomly selected student
1. is a male;
2. is a male smoker;
3. is either a smoker or did smoke but quit;
4. is a female who is a smoker or did smoke but quit.

Independent RVs



- We learned how to calculate probability of independent events:

$$p(\mathbb{A} \cap \mathbb{B}) = p(\mathbb{B}|\mathbb{A})p(\mathbb{A}) = p(\mathbb{B})p(\mathbb{A})$$

- This definition can also be used for probability functions. Say, X and Y are RVs. If the events $X = x$ and $Y = y$ are independent for all x and y, then we say that X and Y are independent RVs. We also have:

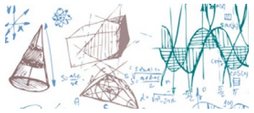
$$p(X = x, Y = y) = p(X = x) \cdot p(Y = y)$$

$$f(x, y) = f_1(x) f_2(y)$$

- Similarly, we say that X and Y are independent RVs if the events $X \leq x$ and $Y \leq y$ are independent for all x and y. We can write:

$$p(X \leq x, Y \leq y) = p(X \leq x)p(Y \leq y) \rightarrow \qquad F(x, y) = F_1(x)F_2(y)$$

Mathematical Expectation



- The **mathematical expectation** (or expected value) is one of the most important notions in statistics. Let's start (as usual…) from a DRV
- **Definition 1**. Assume that X is a DRV having the possible values as follow $\{x_1, x_2, \dots, x_n\}$, the **expectation** of X is defined as:

$$E[X] = x_1 \cdot p(X = x_1) + \cdots + x_n \cdot p(X = x_n) = \sum_{i/1}^{i/n} x_i \cdot p(X = x_i)$$

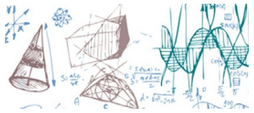
$$E[X] = x_1 \cdot f(x_1) + \cdots + x_n \cdot f(x_n) = \sum_{i/1}^{i/n} x_i \cdot f(x_i)$$

where: $f(x_i)$ is the DRV's P.D.F.

- NOTE. When the respective probabilities for events x_i are all equal, we have (**arithmetic mean**):

$$E[X] = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_i x_i$$

Mathematical Expectation

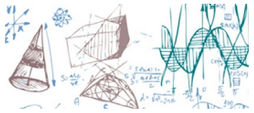


- For a CRV X having P.D.F. $f(x)$, the expectation of X is defined as follow (we always silently assume that the integral converges absolutely):

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

- If we know completely the P.D.F. of RV X then we call $E[X]$ the **mean value** of **X** and denote it by μ_X
- Since, the mean gives a single value that represents the values of RV X we call it a **measure of central tendency** (remember we loose something here – data reduction)
- NOTE something quite here – the mean value is a **single number** – it is not a RV! We call it a parameter of a RV X's P.D.F.
- The crucial point is that we **assumed that we know** and **understand** the P.D.F. of a RV X – we are going to learn soon that this is usually **not the case**! All we can know is a **sample** drawn from a **population** that is described by an **unknown** P.D.F. **This is the core of statistical reasoning**!

Mathematical Expectation



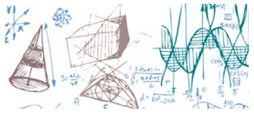
- **Example 1**. Say, that we play a game where we toss a single die (assumed fair). A player wins if she/he has 2 (20$) or 4 (40$), looses if a 6 turns up. Find the expected amount of money to be won:

x_j	0	+20	0	+40	0	−30
$f(x_j)$	1/6	1/6	1/6	1/6	1/6	1/6

$$E[X] = (0\$) \cdot \left(\frac{1}{6}\right)_1 + (20\$) \cdot \left(\frac{1}{6}\right)_2 + (0\$) \cdot \left(\frac{1}{6}\right)_3 + (40\$) \cdot \left(\frac{1}{6}\right)_4$$
$$+(0\$) \cdot \left(\frac{1}{6}\right)_5 + (-30\$) \cdot \left(\frac{1}{6}\right)_6 = 5$$

- A player is expected to win 5$. So, for the game to be fair she/he is expected to pay 5$ in order to play the game…
- For fun – you can check if „Euro Millions” is a fair game…

Mathematical Expectation

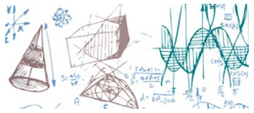


Example 2. Let's have a look how this works for a CRV. Say, the density function of a CRV X is given by:

$$f(x) = \begin{cases} \frac{1}{2}x & 0 < x < 2 \\ 0 & otherwise \end{cases}$$

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{2} x\left(\frac{1}{2}x\right)dx = \int_{0}^{2} \frac{x^2}{2}dx =$$

$$= \frac{x^3}{6}\Big|_0^2 = \frac{4}{3}$$



Moments

- The **rth moment of a RV X about the mean μ**, also called the **rth central moment**, is defined as follow:

$$\mu_r = E[(X-\mu)^r], r = 0, 1, 2, \ldots$$

$$\mu_0 = 1, \mu_1 = 0, \mu_2 = \sigma^2, \ldots$$

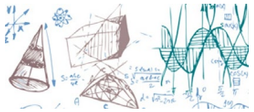
- Assuming absolute convergence we write explicitly for both DRV and CRV:

$$\mu_r = \sum (x-\mu)^r f(x)\,, \mu_r = \int_{-\infty}^{\infty} (x-\mu)^r f(x) dx$$

- The **rth moment of a RV X about the origin**, also called the **rth raw moment**, is defined as:

$$\mu_r' = E[X^r]$$

$$\mu_0' = \mu_0 = 1, \mu_1' = \mu$$



Moments

Let's consider the case of a continuous variable:

$$\mu_0 = \int_{-\infty}^{\infty} (x-\hat{x})^0 f(x)dx = 1$$

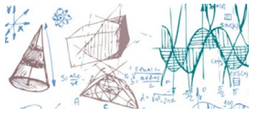
$$\mu_1 = \int_{-\infty}^{\infty} (x-\hat{x})^1 f(x)dx = 0$$

$$\mu_2 = \int_{-\infty}^{\infty} (x-\hat{x})^2 f(x)dx \stackrel{\text{def}}{=} VAR(X) = \sigma^2(X) = \text{VARIANCE}$$

$$\mu_3 = \int_{-\infty}^{\infty} (x-\hat{x})^3 f(x)dx = \text{SKEWNESS}$$

$$\mu_4 = \int_{-\infty}^{\infty} (x-\hat{x})^4 f(x)dx = \text{KURTOSIS}$$

- VARIANCE — a measure of the spread (dispersion) (always > 0)
- SKEWNESS — a measure of asymmetry
- KURTOSIS— a measure of the spread as compared with a special type of distribution – normal distribution



Moments

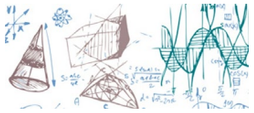
- A general formula that relates the both types of moments can be written as follow:

$$\mu_r = \mu'_r - \binom{r}{1}\mu'_{r-1}\mu + \cdots + (-1)^j \binom{r}{j}\mu'_{r-j}\mu^j + \cdots + (-1)^r \mu'_0 \mu^r$$

$$\mu_2 = \mu'_2 - \mu^2$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu + 2\mu^3$$

- Note, by using moments we can describe any probability distribution function. This is not trivial! Sometimes we do not know (even in principle) what is the P.D.F. of a RV X
- Assuming some concrete function may lead to completely wrong results of statistical analysis. However, we still can calculate the moments using a sample data taken experimentally
- We need, in principle, infinite number of moments to describe a given P.D.F.



Variance

- Another important metric used in statistics is **variance**

$$V[X] = E[(X - \mu)^2]$$

- The variance **cannot** be a **negative number**, the positive square root of the variance is called the **standard deviation**

- Writing explicitly we have:

$$\sigma_X = \sqrt{V[X]} = \sqrt{E[(X - \mu)^2]}$$

- If the probabilities are all equal, we have

$$V[X] = \sigma_X^2 = \sum_{i/1}^{i/n} (x_i - \mu)^2 f(x_i)$$

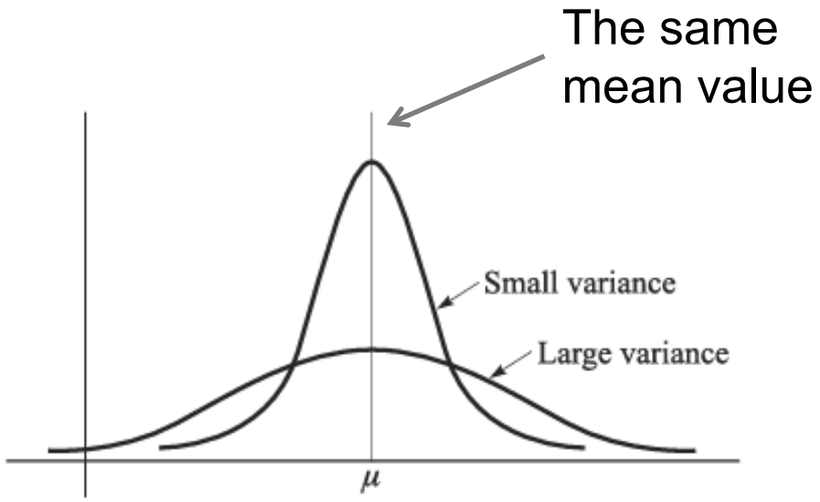
$$\sigma^2 = \frac{1}{n}[(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2]$$

Variance

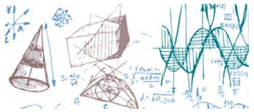
- In case when X is a CRV, we can write the variance as:

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

- We say that the variance is a measure of **the dispersion** (scatter) of the values of the RV **about the mean value** μ. For instance, if the values tend to be concentrated close to the mean, the variance is small



Theorems regarding variance



- **Theorem 4**. Let X be any RV:

$$\sigma^2 = E[(X-\mu)^2] = E[X^2] - \mu^2 = E[X^2] - E^2[X]$$

- **Theorem 5**. If c is any constant, we have:

$$V[cX] = c^2 V[X]$$

- **Theorem 6**. The quantity $E[(X-a)^2]$ is a minimum when $a = E[X]$

- **Theorem 7**. If X and Y are independent RVs,

$$V[X+Y] = V[X] + V[Y], \sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$
$$V[X-Y] = V[X] + V[Y], \sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

Functions of RV

- Last time we already indicated that functions of RV are of great importance for statistics
- Interestingly such function is also a RV itself! For instance:

$$Y = K(X)$$

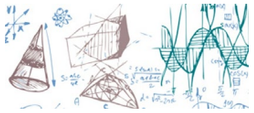
- Now, we can write formulas for the expectation value in a similar manner to defined in previous slides:

$$E[K(X)] = \sum_{i/1}^{i/n} K(x_i) \cdot f(x_i) = \sum_{i/1}^{i/n} K(x_i) \cdot p(X = x_i)$$

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

- In particular we can pick a special function: $K(X) = \boldsymbol{(X - \alpha)^l}$, and its expectation values are called l-th moments about point α (constant)

$$\boxed{m_l = E[(X - \alpha)^l]}$$



Covariance

- Next step, as usual, lead to more RVs. Let's see what's new if we consider two RVs X and Y with joint density function $f(x,y)$:

$$\mu_X = E[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)dxdy \qquad \mu_Y = E[Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x,y)dxdy$$

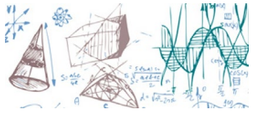
$$\sigma_X^2 = E[(X-\mu_X)^2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)^2 f(x,y)dxdy$$

$$\sigma_Y^2 = E[(Y-\mu_Y)^2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y-\mu_Y)^2 f(x,y)dxdy$$

- And what about the mixed terms? Analysis leads to the **covariance**

$$\sigma_{XY} = Cov[X,Y] = E[(X-\mu_X)(Y-\mu_Y)]$$

$$\sigma_{XY} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y)dxdy$$



Theorems regarding covariance

- **Theorem 1**. For any RVs the following is true:

$$\sigma_{XY} = Cov[X,Y] = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - E[X]E[Y]$$

- **Theorem 2**. In case the RVs X and Y are independent:

$$Cov[X,Y] = 0$$

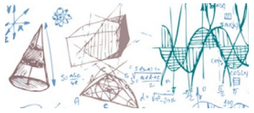
- **Theorem 3**. For any two RVs we have:

$$V[X \pm Y] = V[X] + V[Y] \pm 2Cov[X,Y]$$

- **Theorem 4**. For any two RVs we have:

$$|\sigma_{XY}| \leq \sigma_X \sigma_Y$$

- NOTE, that the converse of Theorem 9 is not necessarily true!



Correlation coefficient

- The covariance gives us a strong hint on how to measure **the dependence** of RVs. If X and Y are independent, then:

$$Cov[X, Y] = \sigma_{XY} = 0$$

- On the other hand, if they are completely dependent (e.g., $X = Y$), then:

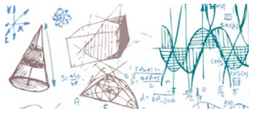
$$Cov[X, Y] = \sigma_{XY} = \sigma_X \sigma_Y$$

- So, one can use the following to measure the dependence of RVs:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- We call it the correlation coefficient and it is easy to note that its values vary between $[-1,1]$
- In case where the CC is equal zero, we call the RVs linearly uncorrelated. In general, however, the variables may or may not be independent.

Change of variables (I)



- Let's assume we know distribution functions of one or more RVs. In practice, we are often interested in finding distributions of other RVs that depend on them (here we focus on CRV)
- **Theorem 1.** Let X be a CRV with P.D.F. given by $f(x)$. Next, define RV $U = \varphi(X)$, where $X = \omega(U)$. The P.D.F. of U is given by $g(u)$ where:

$$g(u)|du| = f(x)|dx|$$

$$g(u) = f(x)\left|\frac{dx}{du}\right| = f\big(\omega(u)\big)\omega'(u)$$

- For more than one variable things getting a bit more difficult…
- **Theorem 2.** Let X and Y be CRVs having joint P.D.F. $f(x, y)$. Let's define new variables $U = \varphi_1(X, Y)$ and $V = \varphi_2(X, Y)$, where $X = \omega_1(U, V)$ and $Y = \omega_2(U, V)$. Then the joint density function of U and V is given as:

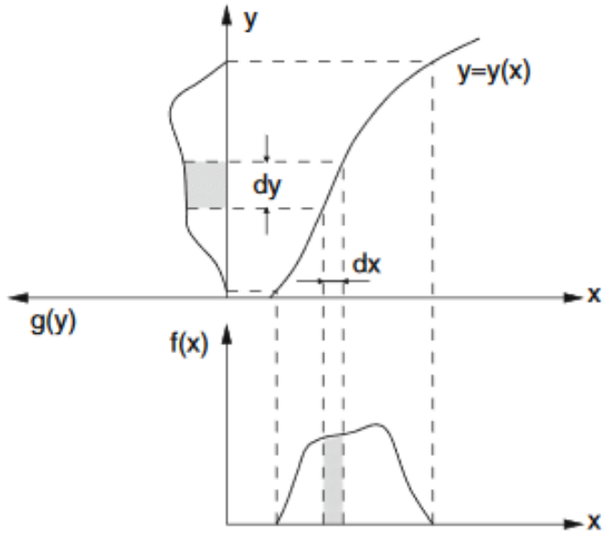
$$g(u, v)|dudv| = f(x, y)|dxdy|$$

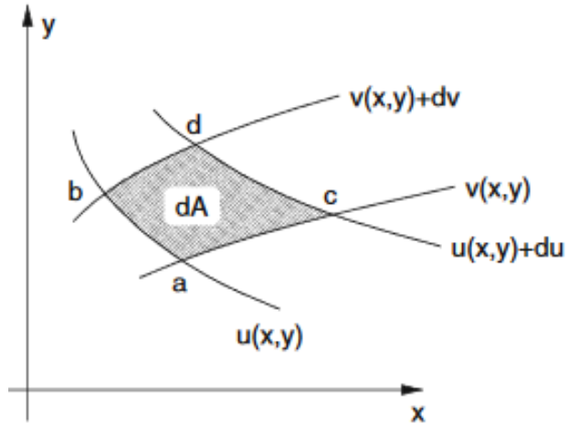
Change of variables (II)

$$g(u,v) = f(x,y)\left|\frac{\partial(x,y)}{\partial(u,v)}\right| = f(\omega_1(u,v),\omega_1(u,v))|J|$$

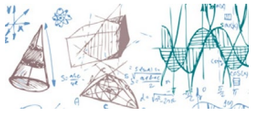
- For multi-dimensional case we have something brand new – Jacobian determinant or Jacobian

$$J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$$





Springer



Conditional P.D.F.s

- Let's assume that X and Y are CRVs. We define the conditional density function of Y given X, as:

$$f(y|x) = \frac{f(x,y)}{f_1(x)}$$

$$f(x|y) = \frac{f(x,y)}{f_2(y)}$$

- So, to define the conditional P.D.F. we need a joint P.D.F. and a marginal one to calculate an appropriate probability we do:

$$p(c < Y < d | x < X < x + dx) = \int_c^d f(y|x)dy$$

Conditional P.D.F.s

Example: Students in a class of 100 were classified according to gender (G) and smoking (S) as follows:

		S			
		s	q	n	
G	male	20	32	8	60
	female	10	5	25	40
		30	37	33	100

Calculate the probability that a randomly selected student is
1. a smoker given that he is a male;
2. female, given that the student smokes.