Refresher Course in Calculus, Probability, and Statistics

Day 3: Probability

Instructor: Dr. Guillaume Wald

9 September 2026

Introduction

"Life's most important questions are, for the most part, nothing but probability problems."

Pierre Simon Laplace

"Doubt is not a pleasant condition, but certainty is absurd."

Voltaire

- Most aspects of the world around us have an element of uncertainty:
  - Will it rain tomorrow?
  - Will I win the lottery next week?
  - Will I earn more than 6,000 CHF/month some day?
- Theory of probability provides mathematical tools for quantifying and describing this randomness and dealing with uncertainty.
- Probability is needed to understand regression analysis and statistics.

Probability

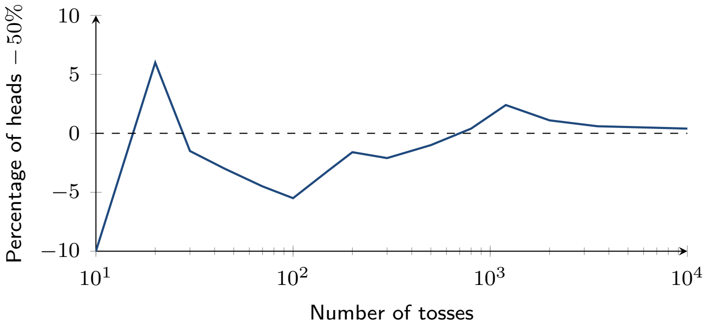
Basic definitions

- **Trial:** event whose outcome is unknown (also called: experiment, observation)
  - flipping a coin
  - rolling a die
- **Sample space:** specification of all possible outcomes of a trial
  - for flipping a coin: $S = \{\text{heads}, \text{tails}\}$
  - for rolling a die: $S = \{1, 2, 3, 4, 5, 6\}$
- **Outcome:** mutually exclusive potential result of a random trial
- **Event:** specification of the outcome of one trial (single outcome or set of outcomes)
  - event "heads in flipping a coin": $A = \{\text{heads}\}$
  - event "odd number in rolling a die": $B = \{1, 3, 5\}$
- **Mutual exclusivity:** events that cannot occur together are mutually exclusive (e.g. passing a test and not passing a test)
- **Independence:** the outcome of one trial has no relationship with the outcome of another trial

Probability

Basic definitions (continued)

- **Relative frequency definition of probability:**
  If an experiment is repeated n times under essentially identical conditions and the event A occurs m times, then as n gets large the ratio $\frac{m}{n}$ approaches the probability of A:
  $$P(A) = \lim_{n\to\infty} \frac{m}{n}.$$



Kerrich's experiment: tossing a coin 10,000 times.

Probability

Properties

Event	Probability
A	$P(A) \in [0,1]$
not A	$P(A^c) = P(\bar{A}) = 1 - P(A)$
A or B	$P(A \cup B) = P(A) + P(B)$ if A, B mutually exclusive
A or B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
A and B	$P(A \cap B) = P(A)\,P(B)$ if A, B independent
A and B	$P(A \cap B) = P(A \mid B)\,P(B) = P(B \mid A)\,P(A)$
A given B	$P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{P(B \mid A)\,P(A)}{P(B)}$ *Bayes*' theorem

Random Variables

- **Random variable** (or stochastic variable): variable whose value results from measurement of a random process
  - numerical summary of a random process
  - probabilities are associated with the possible values of the random variable
- Examples of a random variable:
  - number of times Word crashes while writing a term paper
  - number of times heads comes up when tossing a coin 10 times
- Random variables can be:
  - **Discrete:** can take only a limited number of values (gender, outcome of a die, ...)
  - **Continuous:** can take any value in an interval (time, earnings, ...)

- **Probability density function (pdf):** summarizes the probabilities of outcomes
- For **discrete random variables** it is like a table contrasting the possible values of the variable (outcomes in the sample space of a trial) with the probability that each value occurs
  - indicates the probability of each possible value occurring
  - the value of the pdf of a discrete random variable X, denoted $f(x)$, is the probability that X takes the value x:

$$f(x) = P(X = x)$$

  - $0 \leq f(x) \leq 1$
  - if X takes n possible values $x_1, \ldots, x_n$:

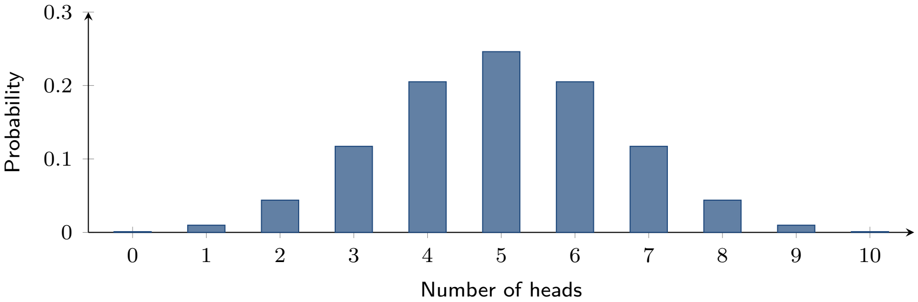
$$P(X = x_1) + \cdots + P(X = x_n) = f(x_1) + \cdots + f(x_n) = 1$$

Discrete Random Variables

Probability Density Function: Example

- pdf for the number of heads out of 10 coin tosses:

x	0	1	2	3	4	5	6	7	8	9	10
$f(x)$	.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010



- The probability of an event (or compound event) can be computed:
  - probability of event "exactly 7 heads":

$$P(X = 7) = 0.1172 = 11.72\%$$

  - probability of compound event "even number of heads":

$$\begin{aligned} &P(X{=}0) + P(X{=}2) + P(X{=}4) + P(X{=}6) + P(X{=}8) + P(X{=}10) \\ &\quad = 0.0010 + 0.0439 + 0.2051 + 0.2051 + 0.0439 + 0.0010 = 0.5000 = 50\% \end{aligned}$$

  - probability of compound event "no more than 3 heads":

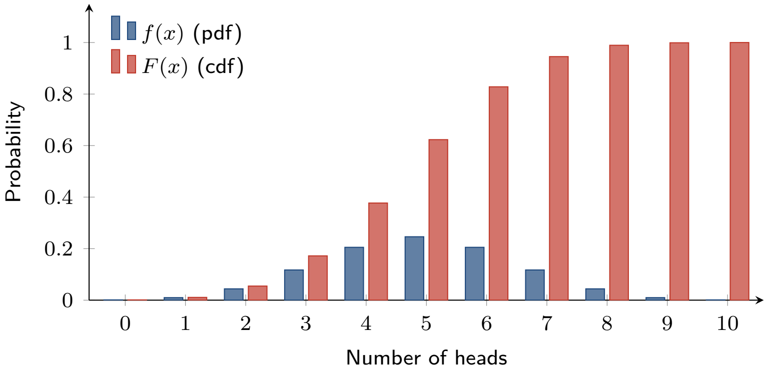
$$\begin{aligned} &P(X{=}0) + P(X{=}1) + P(X{=}2) + P(X{=}3) \\ &\quad = 0.0010 + 0.0098 + 0.0439 + 0.1172 = 0.1719 = 17.19\% \end{aligned}$$

- **Cumulative distribution function (cdf):** an alternative way to summarize probabilities
- the cdf of a random variable X, denoted $F(x)$, is the probability that X is less than or equal to x:
$$F(x) = P(X \leq x)$$
- cdf for the number of heads out of 10 tosses:

# heads	0	1	2	3	4	5	6	7	8	9	10
$f(x)$	.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010
$F(x)$	.0010	.0107	.0547	.1719	.3770	.6230	.8281	.9453	.9893	.9990	1.000

Discrete Random Variables

Cumulative Density Function (continued)



Discrete Random Variables

Special Case: Binary Variables

- Important special case of a discrete random variable: one that can take only two possible values (0 or 1)
  - such variables are called *binary*, *indicator*, *dummy*, or *dichotomous* variables
- Probability distribution of a binary random variable is governed by a **Binomial** or **Bernoulli distribution**
  - a sequence of n independent Bernoulli trials with constant probability of success p at each trial; we are interested in the total number of successes x

Jakob Bernoulli
1655–1705

- **Example:** a student takes a 60-question multiple-choice exam, each question with 4 options, and guesses every answer at random.
  - what are the chances of getting exactly 30 questions right?
  - what are the chances of getting 30 or more right?
  - possible model: binomial with $p = 0.25$ and $n = 60$

Discrete Random Variables

Special Case: Binary Variables (continued)

- The **Binomial distribution**:

$$P(x) = \underbrace{\binom{n}{x}}_{\text{binomial coefficient}} p^x (1-p)^{n-x} = \underbrace{\frac{n!}{x!\,(n-x)!}}_{\text{binomial coefficient}} p^x (1-p)^{n-x}$$

  - x the number of successes
  - p the probability of success
  - n the number of trials

$\rightarrow$ A binomial coefficient (or a combination) represent the **number of ways to choose x items from a set of n items** where order does not matter.

Discrete Random Variables

Special Case: Binary Variables (continued)

- The **Binomial distribution**:

$$P(x) = \underbrace{\binom{n}{x}}_{\text{binomial coefficient}} p^x (1-p)^{n-x} = \underbrace{\frac{n!}{x!\,(n-x)!}}_{\text{binomial coefficient}} p^x (1-p)^{n-x}$$

  - x the number of successes
  - p the probability of success
  - n the number of trials

  → A binomial coefficient (or a combination) represent the **number of ways to choose x items from a set of n items** where order does not matter.

- Example (continued):
  - Probability of getting exactly 30 questions right by guessing:

$$P(30) = \binom{60}{30} 0.25^{30} (1-0.25)^{60-30} = \frac{60!}{30!\,30!} \cdot 0.25^{30} \cdot 0.75^{30} \approx 1.83 \times 10^{-5} = 0.0018\%$$

  - So the probability of getting 30 or more questions right by guessing:

$$P(x \geq 30) = 1 - P(x < 30) = 1 - \sum_{i=0}^{29} P(x=i) \approx 2.67 \times 10^{-5} = 0.0027\%$$

Discrete Random Variables

Expected Value

- **Mathematical expectation:** the mean of a random variable (its long-run average value over many repeated trials or occurrences)
- Expected value of a discrete random variable X taking values $x_1, \dots, x_n$:

$$E(X) = \mu_X = x_1 P(X = x_1) + \cdots + x_n P(X = x_n) = \sum_{i=1}^{n} x_i f(x_i) = \sum_x x\, f(x)$$

→ a weighted average of the possible outcomes (weights = probabilities)

- For a **binary variable**:

$$E(B) = 1 \times p + 0 \times (1 - p) = p$$

→ for the binomial distribution with n trials:

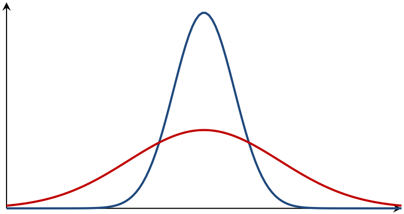
$$E(B) = \mu_B = np$$

- For **linear combinations**:

$$E(aX + bY + c) = a\,E(X) + b\,E(Y) + c$$

Discrete Random Variables

Variance and Standard Deviation



- **Variance** and **standard deviation** measure the dispersion or "spread" of a probability distribution
  - larger variance and standard deviation

  $\Rightarrow$ more "spread out" values
- **Variance** of a discrete random variable X:

$$\mathrm{var}(X) = \sigma_X^2 = E\big[(X - \mu_X)^2\big] = E(X^2) - \mu_X^2$$

  - For a **binary variable**:

$$\mathrm{var}(B) = p(1-p)$$

  - For the **binomial distribution** with n trials:

$$\mathrm{var}(B) = np(1-p)$$

  - For **linear combinations**:

$$\mathrm{var}(aX+bY+c) = a^2\,\mathrm{var}(X)+b^2\,\mathrm{var}(Y)+2ab\;\mathrm{cov}(X,Y) = a^2\sigma_X^2+b^2\sigma_Y^2+2ab\,\sigma_{XY}$$

- **Standard deviation:** $\sigma_X = \sqrt{\text{var}(X)} = \sqrt{\sigma_X^2}$
  - same units of measure as the random variable
- For our example with the exam:
  - how many questions should we expect to get right by guessing?

$$E(B) = \mu_B = np = 60 \times 0.25 = 15$$

  - apply an empirical rule that $\approx 2/3$ of a (symmetric) distribution lies in the range $\mu_B \pm \sigma_X$ and 95% is covered by $\mu_B \pm 2\sigma_X$

$$\sigma_B = \sqrt{\text{var}(B)} = \sqrt{np(1-p)} = \sqrt{60 \times 0.25 \times (1 - 0.25)} = 3.354$$

  - $15 \pm 3.3541 = [11.65,\ 18.35]$
  - $15 \pm 6.7082 = [8.29,\ 21.71]$

- **Skewness** measures lack of symmetry:

$$\text{Skewness} = \frac{E\left[(X-\mu_X)^3\right]}{\sigma_X^3}$$

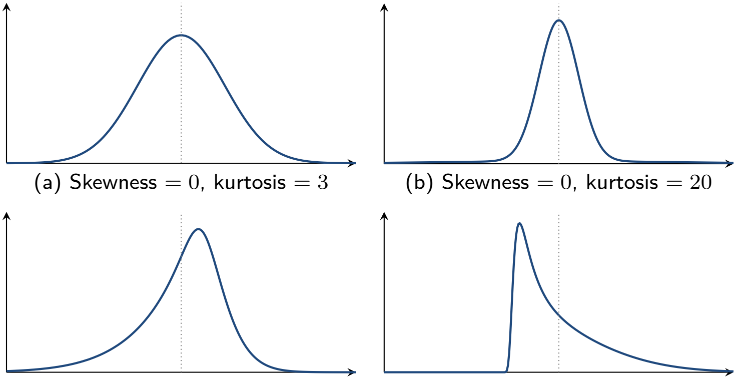
  - $S = 0$: symmetric; skewness is unitless
    $S < 0$: left-skewed (long left tail); $S > 0$: right-skewed (long right tail)

- **Kurtosis** measures how much mass is in the tails of a distribution:

$$\text{Kurtosis} = \frac{E\left[(X-\mu_X)^4\right]}{\sigma_X^4}$$

  - how much of the variance of X comes from extreme values
  - benchmark value is 3 (normal distribution); $K - 3$ is called "excess kurtosis"
  - $K = 3$: mesokurtic; $K < 3$: **platykurtic**; $K > 3$: **leptokurtic**
  - kurtosis is unitless and cannot be negative

Four Distributions with Different Skewness and Kurtosis



(a) Skewness $= 0$, kurtosis $= 3$

(b) Skewness $= 0$, kurtosis $= 20$

(c) Skewness $= -1.0$, kurtosis $= 5$

(d) Skewness $= 1.4$, kurtosis $= 5$

Note: (a),(b) are Gaussian, while (c),(d) are sinh–arcsinh (SHASH). A Gaussian forced to high levels of skewness would grow a non-convex bump in the tail, so SHASH is used for panels (c) and (d) to keep smooth, convex tails at the same kurtosis. The dotted line marks the mean ($x = 0$).

- **Joint probability** that $X = x$ and $Y = y$ (joint pdf of X and Y):

$$f(x, y) = P(X = x,\ Y = y)$$

  - Sum of all joint probabilities: $\displaystyle\sum_x \sum_y f(x, y) = 1$

- **Marginal probability distribution** of random variable X:

$$f_X(x) = \sum_y f(x, y) \qquad \text{for each value of } X$$

  - just another name for its probability distribution
  - computed from the joint distribution of X and Y by adding up the probabilities of all outcomes for which X takes on a specific value

- **Conditional pdf:** the probability that random variable Y takes the value y given that $X = x$:

$$P(Y = y \mid X = x) = \frac{P(X = x,\, Y = y)}{P(X = x)} \qquad \Longrightarrow \qquad f(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

- **Statistical independence:**

$$P(Y = y \mid X = x) = P(Y = y) \qquad \Longrightarrow \qquad f(y \mid x) = f(y) = f_Y(y)$$

Several Discrete Random Variables

Example

- Joint distribution of computer crashes (C) and operating system (S):

Operating system	$C{=}0$	$C{=}1$	$C{=}2$	$C{=}3$	$C{=}4$	Total
Windows ($S{=}0$)	0.531	0.08	0.05	0.029	0.01	0.70
Linux ($S{=}1$)	0.269	0.02	0.01	0.001	0.00	0.30
Total	0.80	0.10	0.06	0.03	0.01	1.00

- Probability the computer does not crash at all *and* runs Linux:
  Joint probability: $P(C{=}0, S{=}1) = 0.269 = 26.9\%$
- Probability a random computer runs Windows:
  Marginal probability: $P(S{=}0) = 0.531 + 0.08 + 0.05 + 0.029 + 0.01 = 0.70 = 70\%$
- Probability a Linux computer does not crash:
  Conditional probability: $P(C{=}0 \mid S{=}1) = \dfrac{0.269}{0.30} = 0.8967 = 89.67\%$

→ Under statistical independence this would be $P(C{=}0 \mid S{=}1) = P(C{=}0) = 0.80$;
since $0.8967 \neq 0.80$, crashes and S are *not* independent.

- **Conditional expectation** (conditional mean) of Y given $X = x$, if Y can take k values $y_1, \ldots, y_k$:

$$E(Y \mid X = x) = \sum_{i=1}^{k} y_i \, P(Y = y_i \mid X = x) = \sum_{y} y \, f(y \mid x)$$

- Example: how many crashes do I expect when working with each OS?
  - Linux computer:

$$E(C \mid S{=}1) = 0 \cdot \frac{0.269}{0.30} + 1 \cdot \frac{0.02}{0.30} + 2 \cdot \frac{0.01}{0.30} + 3 \cdot \frac{0.001}{0.30} + 4 \cdot \frac{0.00}{0.30} \approx 0.143$$

  - Windows computer:

$$E(C \mid S{=}0) = 0 \cdot \frac{0.531}{0.70} + 1 \cdot \frac{0.08}{0.70} + 2 \cdot \frac{0.05}{0.70} + 3 \cdot \frac{0.029}{0.70} + 4 \cdot \frac{0.01}{0.70} \approx 0.439$$

Several Discrete Random Variables

Conditional Variance

- **Conditional variance** (variance of Y given a value for X):

$$\operatorname{var}(Y \mid X = x) = \sum_{i=1}^{k} \left[\left(y_i - E(Y \mid X = x) \right)^2 P(Y = y_i \mid X = x) \right]$$

- Example: conditional variance of crashes for each OS
  - Linux computer:

$$\begin{aligned} \operatorname{var}(C \mid S{=}1) &= \sum_{i=1}^{4} \left[(c_i - E\,(C \mid S{=}1))^2 P(C = c_i \mid S{=}1) \right] \\ &= (0 - 0.143)^2 \times \tfrac{0.269}{0.30} + (1 - 0.143)^2 \times \tfrac{0.02}{0.30} + (2 - 0.143)^2 \times \tfrac{0.01}{0.30} \\ &\quad + (3 - 0.143)^2 \times \tfrac{0.001}{0.30} + (4 - 0.143)^2 \times \tfrac{0.00}{0.30} \approx 0.209 \end{aligned}$$

  - Windows computer:

$$\operatorname{var}(C \mid S{=}0) = \sum_{i=1}^{4} \left[(c_i - E\,(C \mid S{=}0))^2 P(C = c_i \mid S{=}1) \right] \approx 0.809$$

- **Covariance** between X and Y:

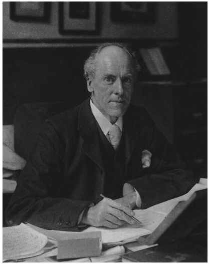
$$\operatorname{cov}(X,Y) = \sigma_{XY} = E\left[(X-\mu_X)(Y-\mu_Y)\right] = E\left(XY\right) - \mu_X\mu_Y$$

  - $\sigma_{XY} > 0$ indicates a positive association: $X > \mu_X \;\Rightarrow\; Y > \mu_Y$, and $X < \mu_X \;\Rightarrow\; Y < \mu_Y$.
  - $\sigma_{XY} < 0$: negative association.
  - $\sigma_{XY} = 0$: neither negative nor positive association.

- **Correlation** between X and Y:

$$\operatorname{corr}(X,Y) = \rho = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

  - Pearson product-moment correlation coefficient
  - unitless measure
  - $-1 \leq \rho \leq 1$



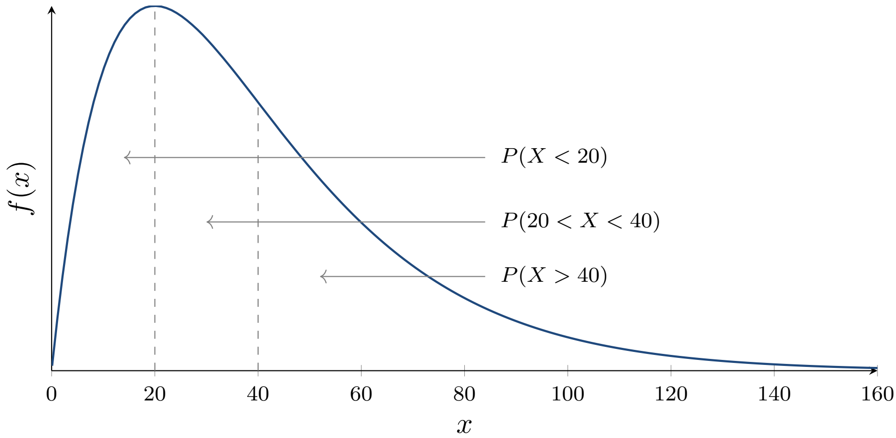
Karl Pearson
1857–1936

Continuous Random Variables

- Continuous random variables can take any value in an interval
  - GDP, interest rates, income, stock-market indices, exchange rates are all (treated as) continuous variables
- Because they can take uncountably many values, the probability that any single value occurs is zero
  - e.g. if X is tomorrow's EUR|CHF exchange rate, $P(X = 0.9300) = 0$
- For continuous variables, probability statements are meaningful only when we consider outcomes within intervals
  - e.g. $P(0.9295 \leq X \leq 0.9305)$ is meaningful ("0.9300" is really an interval)

- The **probability density function (pdf)** summarizes the probabilities of a continuous variable
- Let X be a continuous random variable with pdf $f(x)$:
  - $f(x) \geq 0$
  - $\displaystyle\int_{-\infty}^{\infty} f(x)\, dx = 1$
  - $P(a \leq X \leq b) = \displaystyle\int_{a}^{b} f(x)\, dx$
- The probability that X falls in the interval $[a, b]$ equals the area under the pdf between these points.

pdf of a Continuous Random Variable



Hill, Griffiths, & Lim (2008)

- The **cumulative distribution function (cdf)** is defined as in the discrete case: the probability that the random variable is less than or equal to a value:

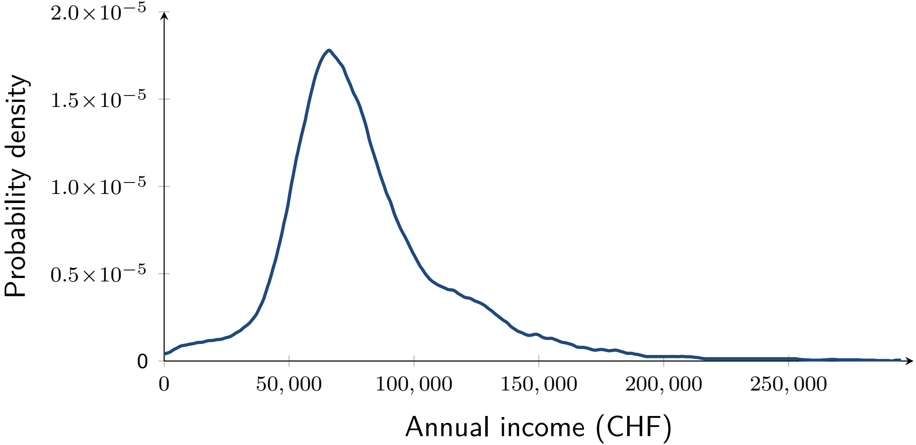
$$P(X \leq a) = \int_{-\infty}^{a} f(x)\, dx = F(a)$$

- the cdf is obtained by *integrating* the pdf
- hence the pdf can be obtained by *differentiating* the cdf:

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

An Empirical *pdf*

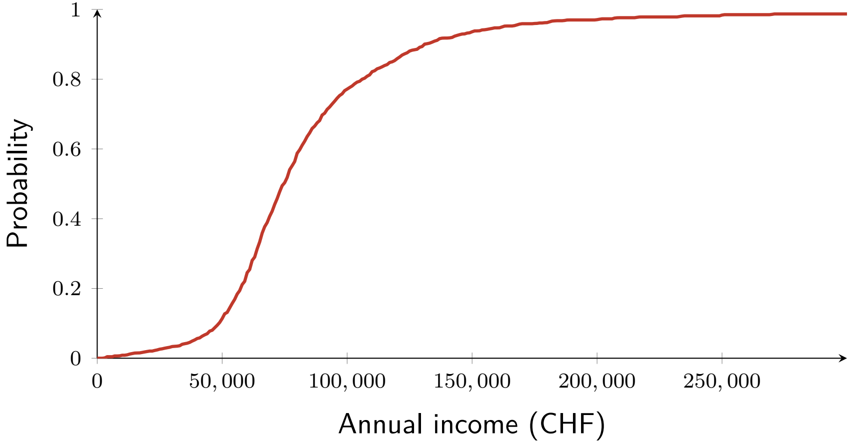
Male Workers' Income in Switzerland



Data: Swiss Labour Force Survey 2008

An Empirical *cdf*

Male Workers' Income in Switzerland



Data: Swiss Labour Force Survey 2008

- **Expectation:**

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

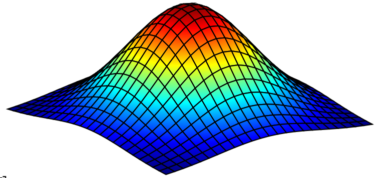
  - Compared to the discrete case, the integral simply replaces the sum
  - Interpretation: the average value of X over an infinite number of draws

- **Variance:**

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

Several Continuous Random Variables

Joint, Marginal, and Conditional Probability Distributions



- The joint probability density function $f(x, y)$ is a *surface*; probabilities are *volumes* under the surface
- Probability that $X \in [a, b]$ **and at the same time** $Y \in [c, d]$:

$$P(a \leq X \leq b,\, c \leq Y \leq d) = \int_{x=a}^{b} \int_{y=c}^{d} f(x, y)\, dx\, dy$$

- Marginal probability density function for Y:

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

- Conditional probability density function:

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

$\rightarrow$ unlike the discrete case, $f(y \mid x)$ is not a probability but a density used to compute probabilities:

$$E(Y \mid X = x) = \int_{-\infty}^{\infty} y\, f(y \mid x)\, dy$$

Normal Distribution

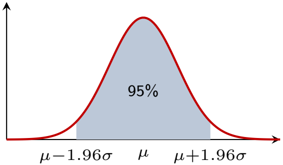
- **Normal (Gaussian) distribution:** continuous probability distribution with a bell-shaped pdf
- A normally distributed random variable X with mean μ and variance σ^2 is written

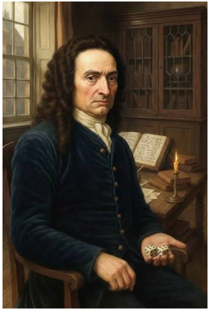
$$X \sim N(\mu, \sigma^2)$$

- pdf of X:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- **Standard normal distribution:** $Z \sim N(0, 1)$
  - tables/software give the standard normal cdf, usually denoted $\Phi(z) = P(Z \leq z)$
  - **standardization** to compute probabilities for $X \sim N(\mu, \sigma^2)$: $\quad z = \dfrac{x - \mu}{\sigma}$





Abraham de Moivre
1667–1754

Normal Distribution (continued)

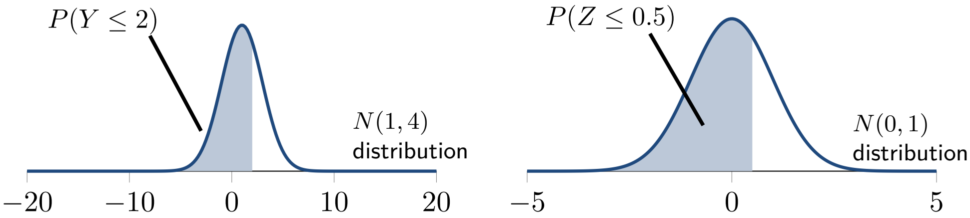
Calculating probabilities

- What is the probability that $Y \leq 2$ if $Y \sim N(1,4)$?
  - Step 1: z-standardize to get the z-score

$$z = \frac{x-\mu}{\sigma} = \frac{2-1}{2} = 0.5$$

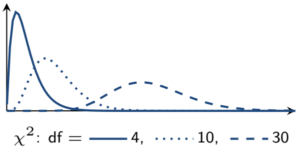
  - Step 2: compare the z-score to the distribution table

$$P\left(\frac{Y-1}{2} \leq 0.5\right) = \Phi(0.5) = 0.691$$

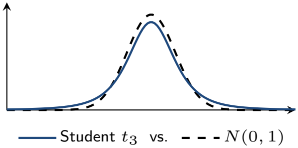


Other important distributions

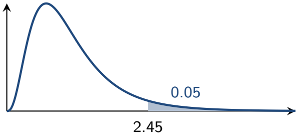
- **Chi-square distribution**
  - $V = Z_1^2 + Z_2^2 + \cdots + Z_m^2 \sim \chi^2_{(m)}$
    with m degrees of freedom
  - where $Z_i \sim N(0,1)$, $i = 1, \ldots, m$, and the Z_i are independent
  - as $m \to \infty$, $\chi^2_{(m)} \to Z \sim N(0,1)$, and $\chi^2_{(1)} = Z_1^2$.



- **Student t-distribution**
  - $t = \dfrac{Z}{\sqrt{V/m}} \sim t_{(m)}$ with m degrees of freedom
  - where $Z \sim N(0,1)$ and $V \sim \chi^2_{(m)}$
  - as $m \to \infty$, $t_{(m)} \to Z \sim N(0,1)$



- **F-distribution**
  - $F = \dfrac{V_m/m}{V_n/n} = \dfrac{\chi^2_m}{m} \cdot \dfrac{n}{\chi^2_n} \sim F_{(m,n)}$
    with m, n degrees of freedom
  - relationship to χ^2: $\lim\limits_{n\to\infty} F = \dfrac{\chi^2_m}{m}$



F-distribution