02462 – Signals and data

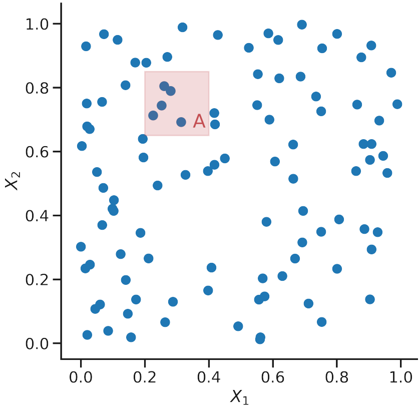
Technical University of Denmark,
DTU Compute, Institut for Matematik og Computer Science.

Overview

Continuous Random Variables

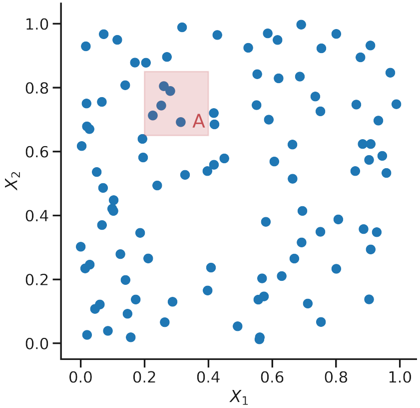
Probability in a Continuous World

- Random variables can be *continuous* instead of *discrete*
  - when discrete, it takes specific values (e.g. $\{0, 1\}$ or the integers)
  - continuous variables can range over whole intervals in $\mathbb{R}$.
- Our concept of probability $\mathbb{P}(X \in A)$ works in the continuous setting as well.
- Can we define something like the probability mass function?



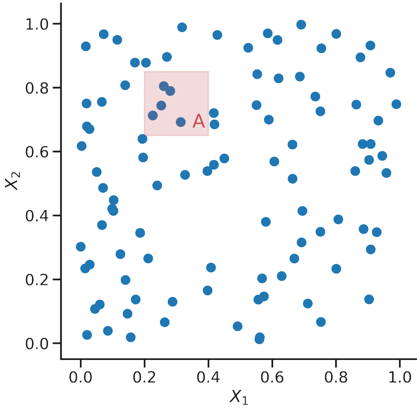
Probability in a Continuous World

- Random variables can be *continuous* instead of *discrete*
  - when discrete, it takes specific values (e.g. $\{0, 1\}$ or the integers)
  - continuous variables can range over whole intervals in $\mathbb{R}$.
- Our concept of probability $\mathbb{P}(X \in A)$ works in the continuous setting as well.
- Can we define something like the probability mass function?



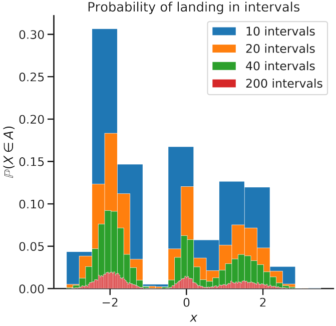
Probability in a Continuous World

- Random variables can be *continuous* instead of *discrete*
  - when discrete, it takes specific values (e.g. $\{0, 1\}$ or the integers)
  - continuous variables can range over whole intervals in $\mathbb{R}$.
- Our concept of probability $\mathbb{P}(X \in A)$ works in the continuous setting as well.
- Can we define something like the probability mass function?



Probability Density

- As the intervals grow smaller, the probability decreases towards 0.
- *Intuition*: probability of smaller intervals add up to larger intervals.
- We can define a *probability density* $p(x)$ at each point x that can be *integrated* to get probabilities $\mathbb{P}(X \in A)$.



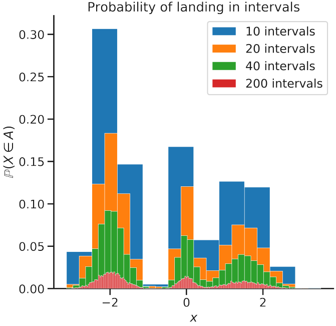
Probability Density

A continuous random variable X is defined in terms of its probability density function $p(x) \geq 0$ for which,

$$\mathbb{P}(X \in A) = \int_A p(x)\,\mathrm{d}x\,, \quad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1 \tag{1}$$

Probability Density

- As the intervals grow smaller, the probability decreases towards 0.
- *Intuition*: probability of smaller intervals add up to larger intervals.
- We can define a *probability density* $p(x)$ at each point x that can be *integrated* to get probabilities $\mathbb{P}(X \in A)$.



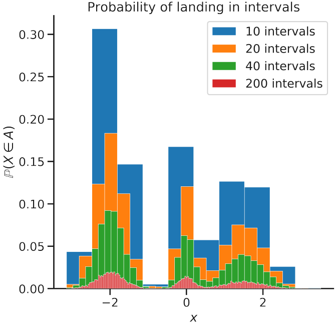
Probability Density

A continuous random variable X is defined in terms of its probability density function $p(x) \geq 0$ for which,

$$\mathbb{P}(X \in A) = \int_A p(x)\,\mathrm{d}x\,, \quad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1 \tag{1}$$

Probability Density

- As the intervals grow smaller, the probability decreases towards 0.
- *Intuition*: probability of smaller intervals add up to larger intervals.
- We can define a *probability density* $p(x)$ at each point x that can be *integrated* to get probabilities $\mathbb{P}(X \in A)$.



Probability Density

A continuous random variable X is defined in terms of its probability density function $p(x) \geq 0$ for which,

$$\mathbb{P}(X \in A) = \int_A p(x)\,\mathrm{d}x\,, \quad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1 \tag{1}$$

Example — the Uniform Density

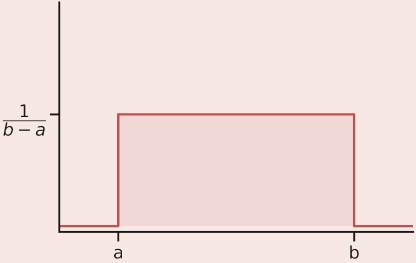
Problem *You are running a program in a loop, and you know that each iteration takes 1 hour to finish. You open your computer at some point during the day — how long until the next iteration finishes?*

If $U \in (0, 1)$ is a fraction of one hour, then since any waiting time between 0 and 1 is equally likely we can model it using a *continuous uniform distribution*.

Continuous Uniform Distribution

A random variable U follows a continuous uniform distribution on an interval (a, b) if it has density,

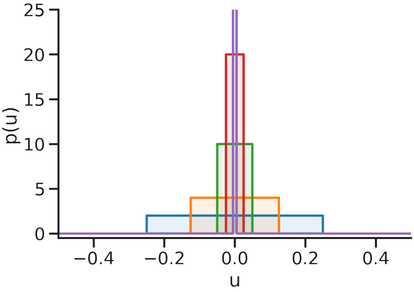
$$p(u) = \frac{1}{b-a}\,\mathbb{1}[u \in (a, b)]$$



Density is not Probability — Density is Unbounded

- Probability is bounded as $\mathbb{P}(X \in A) \leq 1$, but the density can be arbitrarily high.
- The *integral* of the density is always 1 — if limited to a small interval, that means high density.

Example The uniform density on $(-\epsilon/2, \epsilon/2)$ grows to $1/\epsilon$.



Density is not Probability — Density is on Different Scale

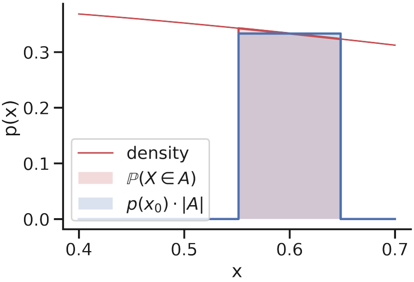
- In a small interval $A_h = (x_0 - {}^h\!/_2, x_0 + {}^h\!/_2)$ the density is almost constant, so

$$\mathbb{P}(X \in A_h) = \int_A p(x)\,\mathrm{d}x \approx h \cdot p(x_0)$$

$$\Rightarrow$$

$$p(x_0) \approx \frac{\mathbb{P}(X \in A_h)}{h}.$$

- Probability density is closer to *probability per interval unit*.

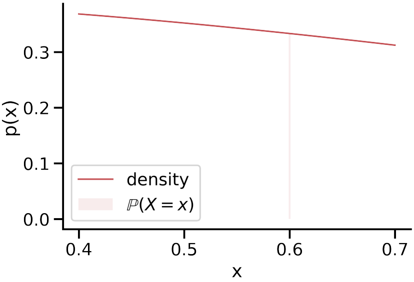


Density is not Probability — Density Does Not Vanish

- No matter what value $p(x)$ has, $\mathbb{P}(X = x)$ *is always* 0.

$$\mathbb{P}(X = x) = \int_x^x p(x')\,\mathrm{d}x' = 0$$

- *Intuition*: Even with an infinite number of samples, you would not necessarily ever see a particular value a.



Rules of Probability — Revisited

Conveniently, the probability rules you have learned extend to probability densities painlessly.

$$p(y|x) = \frac{p(x,y)}{p(x)} \qquad \text{(conditional)}$$

$$p(x,y) = p(y|x)p(x) = p(x|y)p(y) \qquad \text{(product rule)}$$

$$p(y) = \int_{-\infty}^{\infty} p(x,y)\,\mathrm{d}x \qquad \text{(marginals)}$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int_{-\infty}^{\infty} p(y|x)p(x)\,\mathrm{d}x} \qquad \text{(Bayes)}$$

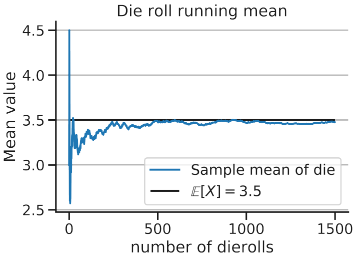
$$X, Y \text{ independent} \Leftrightarrow p(x,y) = p(x)p(y) \qquad \text{(independence)}$$

with $p(x,y)$ being the natural joint density such that

$$\mathbb{P}(X \in A \cap Y \in B) = \int_B \int_A p(x,y)\,\mathrm{d}x\,\mathrm{d}y\,. \qquad \text{(joint density)}$$

Expectation

- The long-run frequency of an event was related to the probability $\mathbb{P}(A) \approx {}^{N_A}/_{N}$. What is the long-run frequency of a random variable X?



If x_n is the n'th sample of a random variable X, the *law of large numbers* tells us that it will converge towards the *expectation* $\mathbb{E}[X]$,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} x_n = \mathbb{E}[X] \qquad \text{(law of large numbers)}$$

Expectation

The expectation $\mathbb{E}[X]$ for a random variable X is

$$\mathbb{E}[X] = \sum_x P(x)x \ \text{ (if discrete)}, \quad \mathbb{E}[X] = \int_{-\infty}^{\infty} p(x)x\,dx \ \text{ (if continuous)}$$

Expectation of Functions and Losses

Problem *For random data X, your model receives a loss $L = g(X)$. What is the expected loss $\mathbb{E}[L]$? What if you know the density of X but not the density of L?*

A helpful theorem in this case is so simple that most people use it subconsciously,

The Law of the Unconscious Statistician

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} p(x) g(x) \, \mathrm{d}x \,, \quad \text{when } Y = g(x). \tag{LOTUS}$$

Warning be careful not to think that $\mathbb{E}[g(X)]$ is equal to $g(\mathbb{E}[X])$. This is *not* true.

A related property of $\mathbb{E}$ is that of *linearity*,

$$\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y].$$

Warning Products are different. $\mathbb{E}[XY]$ is *not* equal to $\mathbb{E}[X]\,\mathbb{E}[Y]$ unless X and Y are independent.

Mean, Variance, and Standard Deviation

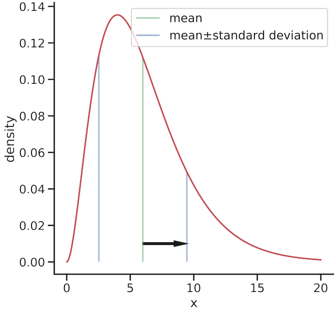
Different expectations tell us different things about X.

- $\mathbb{E}[X]$ is also the *mean* of a distribution and gives the location of X.
- The *variance* describes the "spread" of the distribution,

$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \qquad \text{(variance)}$$

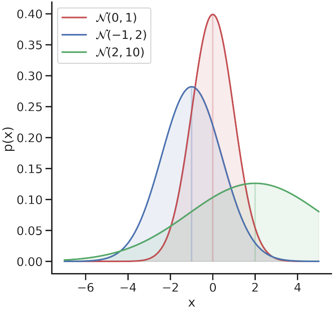
- The variance is hard to interpret as it measures a squared distance, so we often compute the *standard deviation* instead,

$$\text{sd}(X) = \sqrt{\text{Var}(X)} \qquad \text{(standard deviation)}$$



Normal Distribution

- The Normal (or Gaussian) distribution is the most common distribution.
  - It occurs frequently in nature.
  - It plays an important role in statistics.
  - It is mathematically convenient[1].



The Normal Distribution

A random variable X follows a normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ if it has the density function,

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \tag{2}$$

where μ is the mean parameter and σ^2 is the variance parameter.

[1] allowing its application in complicated models without too many headaches.

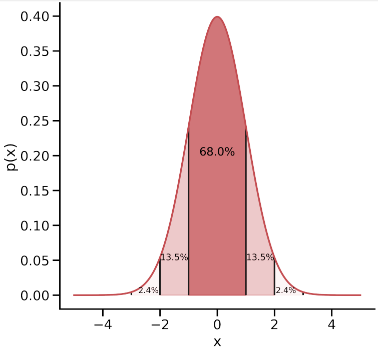
Normal Properties — Concentration around the Mean

Mean, Variance, and the 68-95-99.7 Rule

The parameters of $\mathcal{N}(\mu, \sigma^2)$ correspond to *mean* and *variance*,

$$\mu = \mathbb{E}[X], \quad \sigma^2 = \mathrm{Var}(X).$$

Almost all of the probability density is within 3 standard deviations: 68% is in $[-\sigma, \sigma]$, 95% in $[-2\sigma, 2\sigma]$ and 99.7% in $[-3\sigma, 3\sigma]$.

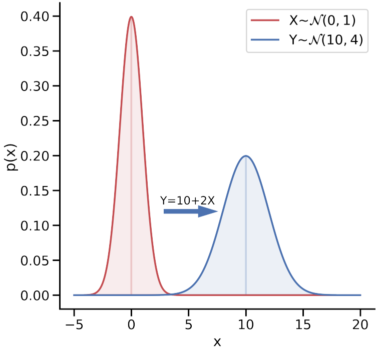


Normal Properties — Location-Scale Family

Scaling and Translating Normal Variables

The normal distribution is in the *location-scale family*, so if $X \sim \mathcal{N}(\mu, \sigma^2)$ then scaling/translating it results in another normal variable,

$$Y = aX + b \Rightarrow Y \sim \mathcal{N}(\mu + b, a^2\sigma^2).$$



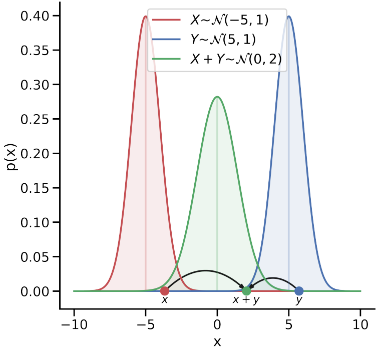
Normal Properties — Normal plus Normal is Normal

Linear Combinations of Normal Variables

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ then $Z = X + Y$ is *normal* and distributed as,

$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2). \tag{3}$$

This is a *unique property* of the normal.



Digression: (Almost) Everything is (Almost) Normal

- If X_n is normal, $Y = \frac{1}{N}\sum_{n=1}^{N} X_n$ is also normal.
- No matter how X_n is distributed, Y is *close to normal*, as long as N is large enough.
- Adding many small effects washes everything out except for mean and variance,

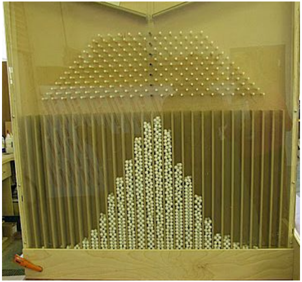
  Genetics Interacting genes coding for height or IQ.

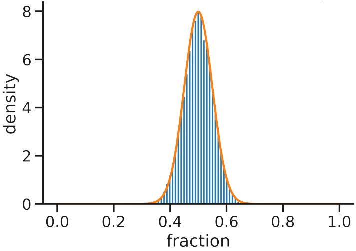
  Finance Thousands of traders influencing fluctuating stock prices.

  Physics Particle interactions producing Brownian motion.

- In statistics, this is formalized in the *central limit theorem*[2].

Galton board





[2] which is a bit beyond the scope of this course — you should just be familiar with the principle.

Making Decisions

Probability becomes useful once we start using it to make *decisions and predictions*.
Decisions can be framed as *optimization*, where we try to pick the best option under some measure. To do this we need to...

1. Determine what criterion we want to optimize (classification accuracy, a patient's health, our income).
2. Determine what *actions* we have available (choice of image class, medical treatment, stock to buy).
3. Evaluate the probability of each possible outcome if we take a specific action.
4. Pick the option that has the best *expected* outcome.

Example: Classification

- We want to find a *decision rule* $D(x)$ that maps x to the its class C.
- We choose a loss function L that penalizes wrong classifications

$$L(C, D(x)) = \mathbb{1}[C \neq D(x)] = \begin{cases} 1 & \text{incur a loss if } D(x) \text{ does not match } C \\ 0 & \text{no loss if } C = D(x) \end{cases}$$

- If $(x, C) \sim p(x, C)$ is the probability of drawing a particular combination of observation and class the *expected loss* is,

$$\rho(D) = \mathbb{E}_{p(x,C)}[L(C, D(x))] \, . \tag{4}$$

- If all we have are samples $(x_n, c_n) \sim p(x, C)$ we can approximate the expected loss as,

$$\tilde{\rho}(D) = \frac{1}{N} \sum_{n=1}^{N} L(c_n, D(x_n)) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}[c_n \neq D(x_n)] \, . \tag{5}$$

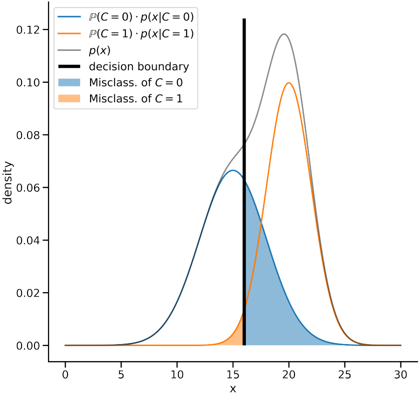
which simply counts the *fraction of classification errors on the dataset*.

Example Continued

The joint data generating density $p(x, C)$ might look like,

$$p(x|C=k) = \mathcal{N}(x; \mu_k, \sigma_k^2), \quad \mathbb{P}(C=1) = \alpha, \quad \mathbb{P}(C=0) = 1-\alpha.$$

A simple decision rule sends everything $x < d$ smaller than *decision boundary d* to class $C = 0$ and the rest to $C = 1$.



Accuracy is Not Everything

- accuracy might not always be the correct thing to optimize for.

		Predicted: Healthy	Predicted: Sick
True	Healthy	True Negative (TN)	False Positive (FP)
True	Sick	False Negative (FN)	True Positive (TP)

Accuracy just measures how often the test is right $\mathrm{Acc} = (TN + TP)/N$.

Recall measures the number of ill people who are caught by the test,

$$\mathrm{recall} = \frac{TP}{FN + TP}$$

Precision measures the risk of misdiagnosis,

$$\mathrm{prec} = \frac{TP}{FP + TP}$$

F1-score tries to balance the two and is in common use,

$$F_1 = 2\frac{\mathrm{prec} \cdot \mathrm{recall}}{\mathrm{prec} + \mathrm{recall}}.$$

- losses can be weighted using financial loss of outcome or another metric.