Mathematics in Machine Learning - DSE@polito

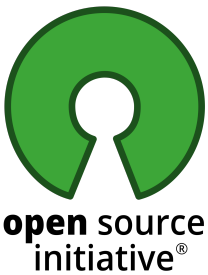
Gasparini's part

Giuseppe Concialdi
@concialdi_g

Christian Montecchiani
@Christian_Montecchiani

Michele Veronesi
@mveronesi

Spring 2022



`https://github.com/users/mveronesi/projects/4/views/1`

Buy us some alcohol plz `https://paypal.me/veronesimichele`

Contents

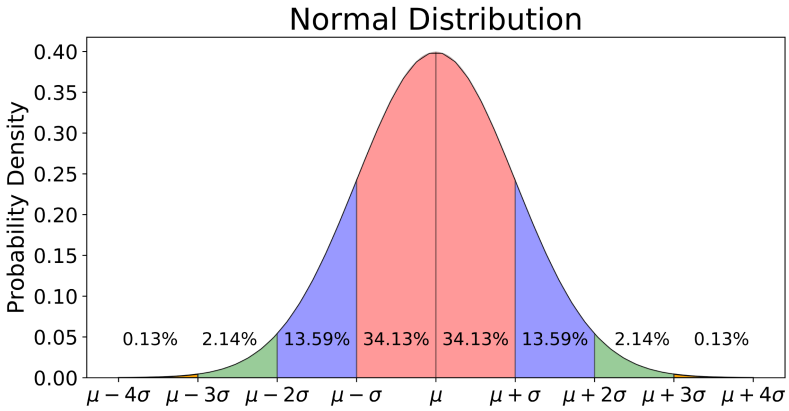


Figure 1: Normal distribution representation

1 Multivariate normal distribution

We start introducing the univariate normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$, which means that X has normal distribution with mean μ and variance σ^2. Then, X is a continuous random variable with probability density function (PDF) reported in equation 1 and represented in figure 1.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \tag{1}$$

The cumulative distribution function (CDF) is $P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$. Its mean is $\mu = \mathbb{E}[X]$ and the variance is

$$\sigma^2 = Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Now, we want to generalize the univariate normal to higher dimensions, i.e., a random vector.

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_d)^T \in \mathbb{R}^d$$

We will do that in two steps.

1. Define the standard multivariate normal random variable

$$\boldsymbol{Z} = (Z_1, \ldots, Z_k)^T \in \mathbb{R}^k$$

2. Define the general multivariate normal random variable.

Step 1 The univariate normal $Z \sim \mathcal{N}(0, 1)$ is well known. We define $\boldsymbol{Z} = (Z_1, \ldots, Z_k) \in \mathbb{R}^k$ as the vector having each component $Z_i \sim \mathcal{N}(0, 1)$ independent and identically distributed (i.i.d.). Its joint density is reported in equation 2.

$$f_{\boldsymbol{Z}}(z_1, \ldots, z_k) = \prod_{i=1}^{k} f_{Z_i}(z_i) \tag{2}$$

Step 2 Given a matrix $A \in \mathbb{R}^{d \times k}$ and a vector $\mu \in \mathbb{R}^d$, we define

$$\boldsymbol{X} = \mu + A \cdot Z$$

to be a normal random vector with mean μ and variance-covariance matrix $\Sigma = A \cdot A^T$.

In general, the expected value of a random vector $\boldsymbol{X}$ is the vector of expected values

$$\mu = \mathbb{E}[\boldsymbol{X}] = \mathbb{E}\begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix} = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix}$$

whereas the variance-covariance matrix at X has variances on the main diagonal and covariances in all the other entries.

$$\Sigma = VarCov(\boldsymbol{X}) = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \ldots & Cov(X_1, X_d) \\ Cov(X_2, X_1) & Var(X_2) & \ldots & Cov(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_d, X_1) & Cov(X_d, X_2) & \ldots & Var(X_d) \end{pmatrix}$$

Recap: moments and their algebra for one dimensional random variable
The expectation of a random variable X is

$$\mathbb{E}[x] = \begin{cases} \int_{\mathcal{X}} x f_X(x)\ dx \text{ if } X \text{ is continuous} \\ \sum_{x_i \in \mathcal{X}} x_i f_X(x_i) \text{ if } X \text{ is discrete} \\ \text{other cases including non existance} \end{cases}$$

Furthermore, the k-th moment of a random variable is $\mathbb{E}[X^k]$. Since $\mathbb{E}$ is a linear operator, we have that $\mathbb{E}[aX+b] = a\mathbb{E}[x] + b$ with a, b constants. Instead, $Var(aX+b) = a^2 Var(X)$ is not linear. When you study more than one random variable, you also have

- $\mathbb{E}[aX+bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
- $Var(aX+bY) = Var(aX) + Var(bY) + 2Cov(aX, bY) = a^2 Var(X) + b^2 Var(Y) + 2ab \cdot Cov(X,Y)$
- $Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[y])] = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$
- For random vectors: $VarCov(A\boldsymbol{X} + b) = A \cdot VarCov(\boldsymbol{X}) \cdot A^T$

Back now to the general multivariate normal random variable $\boldsymbol{X} = A\boldsymbol{Z} + \boldsymbol{\mu}$. We have that

$$\mathbb{E}[\boldsymbol{X}] = \mathbb{E}[A\boldsymbol{Z} + \boldsymbol{\mu}] = A \cdot \mathbb{E}[\boldsymbol{Z}] + \boldsymbol{\mu} = A \cdot \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} + \boldsymbol{\mu} = \boldsymbol{\mu}$$

and

$$VarCov(\boldsymbol{X}) = VarCov(A\boldsymbol{Z} + \boldsymbol{\mu}) = A \cdot VarCov(\boldsymbol{Z}) \cdot A^T = A \cdot I \cdot A^T = AA^T = \Sigma$$

We now define the Pearson correlation coefficient between two random variables (or vectors) as

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} \text{ with } -1 \leq \rho \leq 1$$

If $d = k$, then $X = AZ + \mu,\ A \in \mathbb{R}^{k \times k}$, and if $A^T A$ is invertible, so that $\Sigma^{-1} = (AA^T)^{-1}$ exists, then X has a density which can be easily retrieved from the standard normal density:

$$f_X(x) = (2\pi)^{-k/2} \cdot |det(\Sigma)|^{-1/2} \cdot \exp\{-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\}$$

Note that $det(\Sigma) \neq 0$ if and only if Σ is invertible.
Example in the bivariate case (k=2)
We have that our random vector is $(X,Y)^T$. The joint density is:

$$f_{(X,Y)}(x,y) = (2\pi)^{-1} \cdot |det \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}| \cdot \exp\{-\frac{1}{2} \cdot \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}^T \cdot \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}\}$$

where

$$det \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} = \sigma_X^2\sigma_Y^2 - \sigma_{XY}^2 = \frac{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}{\sigma_X^2\sigma_Y^2} \cdot \sigma_X^2\sigma_Y^2 = (1 - \frac{\sigma_{XY}^2}{\sigma_X^2\sigma_Y^2}) \cdot \sigma_X^2\sigma_Y^2 = (1-\rho_{XY}^2) \cdot \sigma_X^2\sigma_Y^2$$

We use this to compute

$$\Sigma^{-1} = \frac{1}{(1-\rho_{XY}^2)\sigma_X^2\sigma_Y^2} \cdot \begin{bmatrix} \sigma_Y^2 & -\sigma_{XY} \\ -\sigma_{XY} & \sigma_X^2 \end{bmatrix}$$

$$f(x,y) = \frac{1}{2\pi\sqrt{(1-\rho^2)\sigma_X^2\sigma_Y^2}} \cdot \exp\{-\frac{1}{2(1-\rho^2)} \cdot [(\frac{x-\mu_X}{\sigma_X}) - 2\rho(\frac{x-\mu_X}{\sigma_X}) \cdot (\frac{y-\mu_Y}{\sigma_Y}) + (\frac{y-\mu_Y}{\sigma_Y})^2]$$

In the extreme case, we have $\rho = 1$ or $\rho = -1$, then $(1-\rho^2) = 0$ and the density does not exist.

The multivariate normal has good properties: marginal and conditionals distributions are also normal, in addition to linear transformations (which are normal by the way we have defined $X = AZ + \mu$).
Consider $X \sim \mathcal{N}_d(\mu, \Sigma)$ and partition it into two sub-vectors

$$X = \begin{pmatrix} X_1 \in \mathbb{R}^{d_1} \\ X_2 \in \mathbb{R}^{d_2} \end{pmatrix} \text{ where } d_1 + d_2 = d$$

Then, accordingly,

$$\mathbb{E}[X] = \begin{pmatrix}\mathbb{E}[X_1]\\ \mathbb{E}[X_2]\end{pmatrix} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} = \mu$$

$$VarCov(X) = \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12} = \Sigma_{21}^T\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$$

Then, as said, the marginal distribution are normal

$$X_1 \sim \mathcal{N}_{d_1}(\mu_1, \Sigma_{11}) \ \ \& \ \ X_2 \sim \mathcal{N}_{d_2}(\mu_2, \Sigma_{22})$$

Also conditional distributions are normal (proof is omitted):
if Σ_{11}^{-1} exists, then
$X_2|X_1 = x_1 \sim$ The conditional distribution of the random vectorX_2 given that the random vector X_1 equals x_1 (a constant vector) i.e.,

$$X_2|X_1 = x_1 \sim \mathcal{N}_{d_2}(\underbrace{\mu_2 + \Sigma_{21}\cdot\Sigma_{11}^{-1}\cdot(x_1-\mu_1)}_{\text{conditional expected value}}, \quad \underbrace{\Sigma_{22} - \Sigma_{21}\cdot\Sigma_{11}^{-1}\cdot\Sigma_{12}}_{\text{conditional var-cov matrix}} \quad)$$

1.1 Exam exercises

1.1.1 Exercise 1

Given (X, Y), a random vector with joint density

$$f_{(X,Y)}(x,y) = \lambda x e^{-x(\lambda+y)}, \ (x>0), \ (y>0)$$

find marginal and conditional densities of X, Y, $X|Y$, $Y|X$.

Marginal of X.

$$\begin{aligned}
f_X(x) &= \int f_{(X,Y)}(x,y)dy \ (x>0)\\
&= \int_0^{+\infty} \lambda x e^{-x(\lambda+y)}dy \ (x>0)\\
&= \lambda x e^{-\lambda x}\cdot\int_0^{+\infty} e^{-xy}dy \ (x>0) (1)\\
&= \lambda \cancel{x} e^{-\lambda x}\cdot\left[-\frac{e^{-xy}}{\cancel{x}}\right]_0^{+\infty} \ (x>0)\\
&= \lambda e^{-\lambda x}\cdot[0+1] \ (x>0)\\
&= \lambda e^{-\lambda x} \ (x>0)
\end{aligned}$$

We expand (1) imposing $t = -xy$, then $dt = -xy\ dy \iff dt = -x\ dy$

$$\int e^{-xy} = -\frac{1}{x}\int e^{-xy}(-x)dy = -\frac{1}{x}\int e^t dt = -\frac{1}{x}e^{-xy} + c$$

Marginal of Y.

$$\begin{aligned}
f_Y(y) &= \int_0^{\infty} \lambda x e^{-x(\lambda+y)}dx \ (y>0)\\
&= \lambda\frac{1}{\lambda+y}\underbrace{\int_0^{\infty} x e^{-x(\lambda+y)}(\lambda+y)}_{\text{Expectation of } Exp(\lambda+y)} dx \ (y>0) \ (2)\\
&= \frac{\lambda}{\lambda+y}\cdot\frac{1}{\lambda+y} = \frac{\lambda}{(\lambda+y)^2} \ (y>0)
\end{aligned}$$

Conditional of $Y|X$.

$$f_{Y|X}(y|x) = \frac{f_{(X,Y)}(x,y)}{f_X(x)} = \frac{\cancel{\lambda x e^{-x\lambda}}\cdot e^{-xy}}{\cancel{\lambda e^{-\lambda x}}} = x e^{-xy}(y>0)$$

which is $\sim Exp(x)$.

Conditional of $X|Y$.

$$f_{X|Y}(x|y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)} = \frac{\lambda x e^{-x(\lambda+y)}}{\frac{\lambda}{(\lambda+y)^2}} = \lambda x e^{-x(\lambda+y)} \cdot \frac{(\lambda+y)^2}{\lambda}$$

1.1.2 Exercise 2

Let (X, Y) be a gaussian random vector with mean $\mu = (0, 0)$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 2 \\ 2 & 8 \end{pmatrix}$$

Define $W = X - 3Y$ and $Z = Y - \alpha X$ with $\alpha \in \mathbb{R}$.

Compute mean and variance of W.

$$\mu_W = \mathbb{E}[W] = \mathbb{E}[X - 3Y] = \mathbb{E}[X] - 3\mathbb{E}[Y] = 0$$

$$\sigma_W^2 = Var(W) = Var(X - 3Y) = Var(X) + 9 \cdot Var(Y) - 6 \cdot Cov(X, Y) = 1 + 9 \cdot 8 - 6 \cdot 2 = 61$$

Determine for which value of α W and Z are independent. We need the variance-covariance matrix of the vector (W, Z).

$$\begin{pmatrix} W \\ Z \end{pmatrix} = \begin{pmatrix} 1 & -3 \\ -\alpha & 1 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2(?, ?)$$

$$\mathbb{E}[(W, Z)] = \mathbb{E}\left[\begin{pmatrix} 1 & -3 \\ -\alpha & 1 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Because (X, Y) has 0 as expected value. In order to have independence,

$$VarCov((W, Z)) = VarCov\left(\begin{pmatrix} 1 & -3 \\ -\alpha & 1 \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \end{pmatrix} \right) = \begin{pmatrix} 1 & -3 \\ -\alpha & 1 \end{pmatrix} \cdot VarCov((X, Y)) \cdot \begin{pmatrix} 1 & -\alpha \\ -3 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & -3 \\ -\alpha & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 \\ 2 & 8 \end{pmatrix} \cdot \begin{pmatrix} 1 & -\alpha \\ -3 & 1 \end{pmatrix} = \cdots = \begin{pmatrix} 61 & 5\alpha - 22 \\ 5\alpha - 22 & \alpha^2 - 4\alpha + 8 \end{pmatrix}$$

Then, to have independence, we impose

$$Cov(W, Z) = 5\alpha - 22 = 0 \iff \alpha = \frac{22}{5}$$

Compute $E(Y|X)$ and $E(X|Y)$. We can interpret the expected value as a random variable itself. Then

$$\mathbb{E}[Y|X = x] = \mu_Y + Cov(X, Y) \cdot Var(X)^{-1} \cdot (x - \mu_X) = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2} \cdot (x - \mu_X) = 0 + \frac{2}{1} \cdot (x - 0) = 2x$$

$$\mathbb{E}[X|Y = y] = \mu_X + \frac{\sigma_{XY}}{\sigma_Y^2} \cdot (y - \mu_Y) = 0 + \frac{2}{8} \cdot (y - 0) = \frac{y}{4}$$

1.1.3 Exercise 3

Let Z_1, Z_2, Z_3 be i.i.d. standard normal and let

$$\begin{pmatrix} X \\ Y \\ S \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Compute mean and variance of (X, Y, S). For the properties of the multivariate random variables we have

$$\mathbb{E}[(X, Y, S)] = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\Sigma = VarCov((X, Y, S)) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

Write the probability that both X and S are negative as a double integral. Suggest possible computational strategies. We have that

$$\begin{pmatrix} X \\ S \end{pmatrix} \sim \mathcal{N}_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix})$$

Then

$$\begin{aligned}
&\mathbb{P}((X<0)\cap(S<0)) = \int_{-\infty}^{0}\int_{-\infty}^{0} f_{(X,S)}(x,s)\ dx\ ds \\
&= \int_{-\infty}^{0}\int_{-\infty}^{0} \frac{1}{2\pi}\cdot|\det\begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}|\cdot\exp\{-\frac{1}{2}\cdot\begin{pmatrix} x \\ s \end{pmatrix}^T\cdot\begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}\cdot\begin{pmatrix} x \\ s \end{pmatrix}\}\ dx\ ds \\
&= \int_{-\infty}^{0}\int_{-\infty}^{0} -\frac{1}{\pi}\cdot\exp\{\frac{x^2+sx+x+3s^2}{2}\}\ dx\ ds
\end{aligned}$$

We can solve it using Monte Carlo Crude, numerical integration (e.g. with the trapezoid method) or using some routines in Python or R.

Find the conditional distribution of S given $(X=x, Y=y)$.

$$\begin{aligned}
&f_{S|X,Y}(s|x,y) = \frac{f_{(X,Y,S)}(x,y,s)}{f_{(X,Y)}(x,y)} = \\
&= \frac{\frac{1}{2\pi}\cdot|\det\begin{pmatrix} 1 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}|\cdot\exp\{-\frac{1}{2}\cdot\begin{pmatrix} x \\ y-1 \\ s \end{pmatrix}^T\cdot\begin{pmatrix} 1 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}^{-1}\cdot\begin{pmatrix} x \\ y-1 \\ s \end{pmatrix}\}}{\frac{1}{2\pi}\cdot|\det\begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}|\cdot\exp\{-\frac{1}{2}\cdot\begin{pmatrix} x \\ y-1 \end{pmatrix}^T\cdot\begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}^{-1}\cdot\begin{pmatrix} x \\ y-1 \end{pmatrix}\}} \\
&= \frac{\frac{1}{2\pi}\cdot|\det\begin{pmatrix} 1 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}|\cdot\exp\{-\frac{1}{2}\cdot\begin{pmatrix} x \\ y-1 \\ s \end{pmatrix}^T\cdot\begin{pmatrix} 1 & 0 & 1 \\ 0 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}/4\cdot\begin{pmatrix} x \\ y-1 \\ s \end{pmatrix}\}}{\frac{1}{2\pi}\cdot|\det\begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}|\cdot\exp\{-\frac{1}{2}\cdot\begin{pmatrix} x \\ y-1 \end{pmatrix}^T\cdot\begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}/4\cdot\begin{pmatrix} x \\ y-1 \end{pmatrix}\}} \\
&= \frac{4\cdot\exp\{(x^2\cdot 2^{-1}+sx-3y+2sy+2y^2-s+3s^2\cdot 2^{-1}+2)/4\}}{4\cdot\exp\{(x^2\cdot 2^{-1}+2y^2-4y+2)/4\}} \\
&= \exp\{(sx+y+2sy-s+\frac{3}{2}s^2)/4\}
\end{aligned}$$

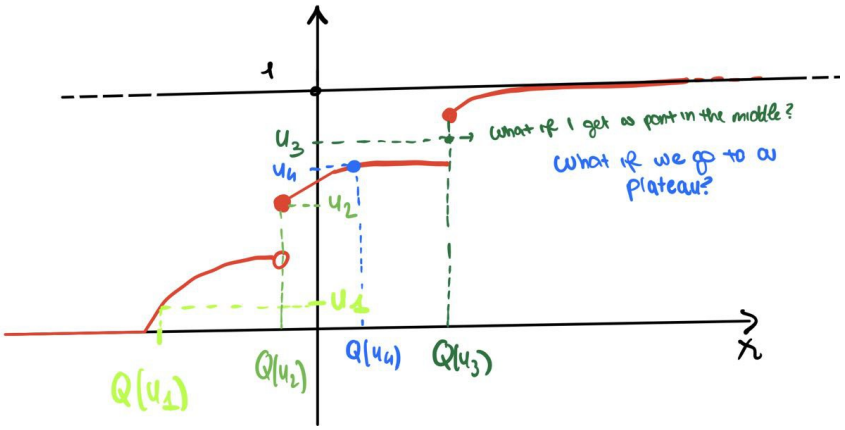


Figure 2: Example of a cumulative distribution of a Random Variable

2 Monte Carlo Methods

Many algorithms in machine learning and data science make use of **Monte Carlo techniques**. There are 3 main uses of these techniques:

1. **Simulate** random objects and processes in order to observe their behaviour.
2. **Estimate** numerical quantities by repeated sampling.
3. **Solve** complicated optimization problems through randomized algorithms.

2.1 Simulations

Simulation means using a computer to generate *realizations* or *observations* of a random variable with a given distribution. We can do it in 2 steps:

1. **Generate** a random number from a uniform distribution, $U(0,1)$.
2. Return the **inverse** of the Cumulative Distribution Function (CDF) of the distribution of the random variable that you want to simulate.

To do the **step 2** we must define the **quantile function**, which is a generalize inverse function of a CDF.

Definition If X has a CDF $F_X(x)$, where $x \in \mathbb{R}$ then its quantile function:

$$Q(u) = inf\{x : u \leq F_X(x)\}^1$$

Notice that u is a number between $(0,1)$, which is an open interval because the quantile function is not well defined in zero and one.

The quantile function has an important **property**:

$$Q(u) \leq x \iff u \leq F(x) \tag{3}$$

Proof

- $\Leftarrow$ if $u \leq F_X(x)$, then by definition of infimum we obtain $Q(u) \leq x$.
- $\Rightarrow$ suppose that $Q(u) \leq x$ then since CDF is a non-decreasing function we have that $F(Q(u)) \leq F_X(x)$ and by graphically visualize each case in the figure 2, we can say that $u \leq F(Q(u)$ therefore, $u \leq F(Q(u) \leq F(x)$ which implies $u \leq F(x)$.

Theorem 2.1 *If X is a random variable with quantile function $Q(u)$, $u \in (0,1)$, and if $U \sim Uniform(0,1)$ then $Q(U)$ and X have the same distribution.*

[1] We use the *infimum* $(\cdot)$ function.

Proof By definition we know that $F_X(x) = \mathbb{P}(X \leq x)$ and this implies that $F_{Q(U)}(x) = \mathbb{P}(Q(U) \leq x)$. Starting from this last definition:

$$\begin{aligned} F_{Q(u)}(x) &= \mathbb{P}(Q(U) \leq x) = (\ by\ definition\ ) \\ &= \mathbb{P}(U \leq F(x)) = (\ since\ U \sim U(0,1)\ ) \\ &= F(x) = \mathbb{P}(X \leq x) \end{aligned}$$

Therefore, $Q(U)$ and X have the same CDF $\Rightarrow$ they have the same distribution. This justify the general algorithm:

Algorithm 1 Simulate Random Variable

Require: CDF of the distribution
Draw $U \sim U(0,1)$
return $X = Q(U)$

Example Suppose you want to simulate the following distribution which is a Bernoulli:

$$f_X(X = x) \begin{cases} 0.3 & \text{if } x = 1 \\ 0.7 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

```
import numpy as np

u = np.random.random()  # Generate random number from U(0,1)

if u < 0.3:
    x = 1
else:
    x = 0

    print(f"The observation simulate is {x} ")
```

Example Suppose you want to simulate the following distribution which is a Exponential:

$$X \sim Exponential(\lambda),$$

$$f_X(X = x) = \lambda e^{-\lambda x}\ (\lambda > 0)$$

Firstly, let's calculate the cumulative distribution function of the exponential distribution:

$$F_X(x) = \int_{-\infty}^{\infty} \lambda e^{-\lambda x} = \begin{cases} 1 - e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Notice that the quantile function $Q(u)$ can be computed faster just by using the following result:

$$\begin{aligned} u &= 1 - e^{-\lambda x} \\ 1 - u &= e^{-\lambda x} \\ \log(1-u) &= -\lambda x \\ x &= -\frac{\log(1-u)}{\lambda} \end{aligned}$$

```
import numpy as np

u = np.random.random()  # Generate random number from U(0,1)

x = - np.log(1-u) / lambda

print(f"The observation simulate is {x} ")
```

How does the RND generator work? Generally speaking, they use an algorithm based usually on linear congruential generators, which are based on the remainder of an integer division (the most unpredictable programmable thing in Mathematics).

Choose **M**(large), **a**, and **c** integers and define for $i = 1, 2, \ldots$

$$L_i = (a\ L_{i-1} + c)\ mod\ M \tag{4}$$

Starting from a seed L_0. Then define $u_i = \frac{L_i}{M}$. For some large i, $u_i, u_{i+1}, \ldots$ behave as if they were I.I.D. $U(0,1)$.

We have seen how to generate a single observation x from a density $f_X(\cdot)$. Let's now simulate a random vector:

$$\mathbf{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$$

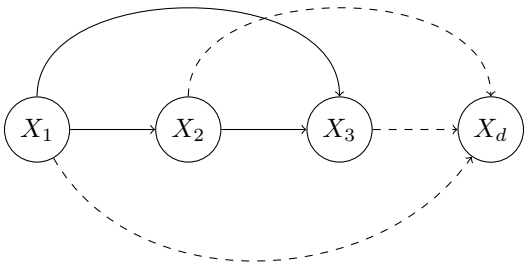
with density $f(x_1, \ldots, x_d)$.
We can see it as:

$$f((\underbrace{x_1, \ldots, x_{d-1}}_{x}, \underbrace{x_d}_{y})) = f_{X_1,\ldots,X_{d-1}}(x_1, \ldots, x_{d-1}) \times f_{X_d|X_1,\ldots,X_{d-1}}(x_d) \tag{5}$$

Generalizing equation 5:

$$f(x_1, \ldots, x_d) = \underbrace{f_{X_1}(x_1)}_{marginal\ of\ X_1} \times f_{X_2|X_1}(x_2) \times f_{X_3|X_1,X_2}(x_3) \times \ldots \underbrace{f_{X_d|X_1,\ldots,X_{d-1}}(x_d)}_{conditional\ of\ X_d\ given\ the\ previous\ RVs} \tag{6}$$

Which can be represented as a complete DAG as the figure below:

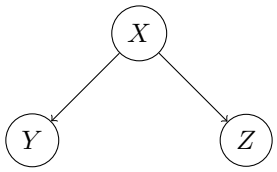


This suggest the following algorithm:

Algorithm 2 Simulate Random Vector

Require: Conditional probabilities density function
Generate X_1 from $f_{X_1}(x_1)$
Generate X_2 from $f_{X_2|X_1}(x_2|X_1 = x_1)$
...
Generate X_d from $f_{X_d|X_1,\ldots X_{d-1}}(x_d|X_1 = x_1, X_2 = x_2, \ldots, X_{d-1} = x_{d-1})$
return Vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)$

Notice that is a series of univariate simulation giving the multivariate simulation. Also, if not all the arrows are present the algorithm simplifies a bit. For example:



In this case the algorithm becomes:

1. Simulate X from f_X
2. Simulate Y from $f_{Y|X}$
3. Simulate Z from $f_{Z|X}$ (Notice that $f_{Z|X,Y}$ is not necessary)

There might be a special algorithm for particular multivariate distribution, which is more efficient than this general algorithm. For example, take

$$X \sim N_d(\mu, \Sigma) \tag{7}$$

suppose $\Sigma = AA'$ for some matrix A, e.g. $A = chol(\Sigma)$ the *Choleski decomposition* of a non-negative definite matrix. Then the algorithm becomes:

1. Simulate $\mathbf{Z} = (Z_1, \ldots, Z_d)^T$ **i.i.d** from $N(0,1)$
2. Set $\mathbf{X} = \mu + A\mathbf{Z}$, where $\mathbf{X} \sim N(\mu, \Sigma)$

Back to the general algorithm it can be change in the following way: sample from a conditional distribution instead marginalize by integrating over a joint distribution. This is a well know algorithm called **Gibbs sampler** and this is its pseudo-code:
The sequence of sampling can proved to be a **Markov Chain** with steady state the desired distribution. After

Algorithm 3 Gibbs sampler

Require: Conditional probabilities density function
  Initialize $X^{(0)}$
  for all $j \in [0, B]$ **do**
    Generate X_1^{j+1} from $f_{X_1|X_2,\ldots X_d}(x_1|x_2^j, x_3^j, \ldots, x_d^j)$
    Generate X_2^{j+1} from $f_{X_d|X_1,X_3,\ldots X_d}(x_d|x_1^{j+1}, x_3^j, \ldots, x_d^j)$
    ...
    Generate X_i^{j+1} from $f_{X_i|X_1,\ldots,X_{i-1},X_{i+1},\ldots X_d}(x_i|x_1^{j+1}, x_2^{j+1}, \ldots, x_{i-1}^{j+1}, x_{i+1}^j, \ldots, x_d^j)$
    ...
    Generate X_d^{j+1} from $f_{X_d|X_1,\ldots X_{d-1}}(x_d|x_1^{j+1}, x_2^{j+1}, \ldots, x_{d-1}^{j+1})$
  end for
  return Vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)$

a while (when B is large enough) the chain will forget the initial state $X^{(0)}$. If we keep on doing it, after a while when $C \gg B$, X^C will be independent realization of $\mathbf{X}$ with no memory of X^B.

Another Markov Chain Monte Carlo method for simulating a random vector $\mathbf{X}$ more general then Gibbs, useful when we do not have f_X, but we know f_X up to proportionality constant difficult to compute. This method is called **Metropolis-Hastings** algorithms. It can be proved that, due to the acceptance probability

Algorithm 4 Metropolis-Hastings

  Initialize $X^{(0)}$
  for all $j \in [0, B]$ **do**
    Sample $\mathbf{y}$ from a proposal density $y \sim q(\cdot|\mathbf{x}^j)$
    Sample u from a $U(0,1)$
    Accept $\mathbf{y}$ with probability $\alpha(\mathbf{x}, \mathbf{y}) = min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\}$
    if $\alpha > u$ **then**
      $\mathbf{x}^{j+1} = \mathbf{y}$
    else
      Keep on
    end if
  end for
  return Vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)$

α, the target distribution is actually a steady-state distribution for the Markov Chain: $\mathbf{X}^0, \mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^b, \ldots$, so that, for B large, approximately $\mathbf{X}^B \sim f_{\boldsymbol{X}}(x_1, \ldots, x_d)$

2.2 Estimates

Now, we describe how Monte Carlo techniques can be used to estimate complicated integrals, probabilities and expectation.

Crude Monte Carlo
Suppose we wish to compute the expectation $\mu = \mathbb{E}(\mathbf{Y})$ of some continuous RV $\mathbf{Y}$ with probability density function $f_Y(y)$, but the integral is too difficult.

The idea of **Crude Monte Carlo** (CMC) is to approximate μ by simulating many independent copies of $Y_1, \ldots, Y_n$ of $\mathbf{Y}$ and then take the sample mean $\bar{Y}$ as estimator of $\mu = \mathbb{E}(\mathbf{Y})$. Then by the **Law of Large Numbers** (LLN) $\bar{Y}$ converges in probability as $n \to \infty$. Recall LLN:

$$\bar{g} = \frac{1}{N}\sum_{i=1}^{N} g(x_i) \xrightarrow{P} \mathbb{E}(g(\mathbf{X})) \tag{8}$$

Important consequence: for large N, $\frac{1}{N}\sum_{i=1}^{N} g(x_i)$ can be taken as an approximate of μ. The higher N is the better it is, the error is controlled by the *Central Limit Theorem*, that give us the *confidence*.

Often, $g(\mathbf{X})$ is an indicator function $\mathbb{1}_{[\mathbf{X}\in A]}$:

$$g(\mathbf{X}=1)\begin{cases} 0 & \text{if } \mathbf{X}\notin A \\ 1 & \text{if } \mathbf{X}\in A \end{cases} \tag{9}$$

in which case $\mathbb{E}[g(\mathbf{X})] = \mathbb{P}(x \in A)$. Let's see an example of Monte Carlo estimation: suppose you want to know the $\mathbb{P}(S \leq 1)$ and you cannot compute it easily.

1. **Simulate** $(x_1^j, \ldots, x_d^j)$ from $(X_1, \ldots, X_d)$
2. **Compute** $s^j = g(x_1^j, \ldots, x_d^j)$
3. **Count** $\frac{1}{N}\sum_{j=1}^{N}(s^j \leq 1)$

3 General Linear Models

It is possible to model a dataset as a statistical model. The main idea is that there are some variable of interest called *responses* that have some dependencies with other variables called *predictors* and we want to exploit these correlation to predict the value of the response variable.

3.1 Definition

A general linear model is with n observations and p predictors is characterized as $\underset{n\times 1}{Y} = \underset{n\times p}{X} \cdot \underset{p\times 1}{B} + \underset{n\times 1}{E}$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & z_1 & \dots \\ 1 & x_2 & z_2 & \dots \\ \vdots & \vdots & \vdots & \\ 1 & x_n & z_n & \dots \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

These values can be seen as realization of random variables. In particular:

response — $\boldsymbol{Y}$ is an observable response random vector
design — $\boldsymbol{X}$ collects the observed levels of predictors controlled by the experiment
coefficients — $\boldsymbol{B}$ is a vector of unknown unobservable parameters
errors — $\boldsymbol{E}$ is an unobservable random vector of errors

It is possible to assume that the errors have a multivariate normal distribution, where each error is **independent** from the others follows a $N(0, \sigma^2)$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim N(\bar{0}, \sigma^2 I_n)$$

so that Y has expected value

$$\mathbb{E}[Y] = \mathbb{E}[X\beta + \varepsilon] = \mathbb{E}[X\beta] + \overset{\curvearrowright 0}{\mathbb{E}[\varepsilon]} = \mathbb{E}[X\beta] = X\beta$$

and variance-covariance matrix

$$VarCov[Y] = I_n VarCov[\varepsilon] I_n^T = {}^2 I_n \sigma^2 I_n I_n^T = \sigma^2 I_n$$

therefore, since Y is a linear transformation, it can be modeled as a Normal distribution

$$Y \sim \mathbb{N}_m(X\beta, \sigma^2 I_n)$$

3.1.1 Interpretation of the coefficients

In a liner model, the coefficients of the predictors have an interpretable meaning.

Example 1 Let's suppose we have a model employed to predict the temperature y in a particular room. In order to do this, we only have some predictors such as the humidity x_h and the pressure of the room x_p. The linear model associated can be written as:

$$Y^{(i)} = \beta_0 + \beta_h x_h^{(i)} + \beta_p x_p^{(i)} + \varepsilon^{(i)}$$

The coefficients β_j have the usual interpretation of slope, for example, the coefficient of the humidity predictor β_h can be seen as

> The average variation of temperature when the humidity is increased by one unit and pressure is held constant.

In general the coefficient β_j can be interpreted as

> The average variation of the response when the predictor x_j is increased by one unit and all the other variables are held constant.

[2] X and β are constant

Note that things may be more complicated, for example we could have a polynomial regressor model like

$$Y = \beta_0 + \beta_1 T + \beta_2 T^2 + \beta_3 P + \varepsilon$$

This is still a linear model (w.r.t. βs), since we can rewrite it in this way

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & T_1 & T_1^2 & P_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & T_n & T_n^2 & P_n \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Example 2 Suppose that we have the following linear model.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \\ Y_{m+1} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where we have $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ i.i.d. Then the following relations hold:

- $Y_i = \beta_0 + \varepsilon_i$ for $i = 1, \ldots, m$
- $Y_i = \beta_0 + \beta_1 + \varepsilon_i$ for $i = m+1, \ldots, n$
- $Y_i \sim \mathcal{N}(\beta_0, \sigma^2)$ for $i = 1, \ldots, m$
- $Y_i \sim \mathcal{N}(\beta_0 + \beta_1, \sigma^2)$ for $i = m+1, \ldots, n$

This is a comparison of two homoskedastic (same variance) normal random samples (e.g., t-test, confidence interval for the difference between the two normal means). In fact, the inference often is focused on the difference between the two means $\mu_2 - \mu_1 = \beta_0 + \beta_1 - \beta_0 = \beta_1$.

Back to the original expression $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ and

$$VarCov(\varepsilon) = \begin{pmatrix} \sigma^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma^2 \end{pmatrix}$$

Therefore, since Y is a linear transformation of ε, we have that

$$Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$$

β and σ^2 are both unknown parameters, we want to make inference about them since they are interesting features of our system. We start by seeking the *maximum likelihood estimation* of these parameters. Since $Y_1, \ldots, Y_n$ are independent normal, the likelihood of β and σ^2 is the joint density of the random vector $Y_1, \ldots, Y_n$ computed in the observations $y_1, \ldots, y_n$.

$$\begin{aligned} \mathcal{L}(\beta, \sigma^2 | y_1, \ldots, y_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\{-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (y_i - (X\beta)_i)^2\} \\ &= (2\pi\sigma^2)^{-n/2} \cdot \exp\{-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (y_i - (X\beta)_i)^2\} \\ &= (2\pi\sigma^2)^{-n/2} \cdot \exp\{-\frac{1}{2\sigma^2} \cdot (Y - X\beta)^T \cdot (Y - X\beta)\} \end{aligned}$$

To maximize this likelihood, we first maximize on β keeping σ^2 fixed, then we maximize over σ^2. So, for σ^2 fixed, we have

$$\max_{\beta} (2\pi\sigma^2)^{-n/2} \cdot \exp\{-\frac{1}{2\sigma^2} \cdot (Y - X\beta)^T \cdot (Y - X\beta)\}$$

which is equivalent to

$$\min_{\beta} (Y - X\beta)^T \cdot (Y - X\beta)$$

Therefore, performing the maximum likelihood estimation on a linear model with normal homoskedastic distribution on the error ε is equivalent to solving a famous least squares problem.

3.2 Maximum Likelihood Estimate

The likelihood is, again,

$$\mathcal{L}(\beta, \sigma^2; y_1, \dots, y_n) = (2\pi\sigma^2)^{-n/2} \cdot \exp\{-\frac{1}{2\sigma^2} \cdot (Y - X\beta)^T \cdot (Y - X\beta)\}$$

We take the logarithm in natural basis in order to simplify the calculation.

$$\log \mathcal{L}(\beta, \sigma^2; y_1, \dots, y_n) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \cdot (Y - X\beta)^T \cdot (Y - X\beta)$$

As stated before, we can see that maximizing the (log) likelihood is equivalent to minimizing

$$\begin{aligned}
&\min_\beta (y - X\beta)^T \cdot (y - X\beta) \iff \\
&\min_\beta (y^T - \beta^T X^T) \cdot (y - X\beta) \iff \\
&\min_\beta (y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta) \iff \\
&\min_\beta (y^T y - 2y^T X\beta + \beta^T X^T X\beta)
\end{aligned}$$

We can look for stationary points by differentiating with respect to the vector β: since it is a convex function (a sum of squares) they will be all minima.

$$\begin{aligned}
&\frac{\partial}{\partial\beta}(y^T y - 2y^T X\beta + \beta^T X^T X\beta) = 0 \\
&\iff \\
&-2X^T y + 2X^T X\beta = 0 \\
&\iff \\
&X^T X\beta = X^T y
\end{aligned}$$

If $X^T X$ is invertible, then we can write explicetely

$$\hat{\beta} = (X^T X)^{-1} \cdot X^T y$$

3.3 Estimates of Variance

We have found an estimator for β, but we still have to discuss how to estimate σ^2.
Remember the MLE, $\mathcal{L}(\hat{\beta}, \sigma^2; y_1, \dots, y_n)$, now we take the log of the likelihood and we differentiate with respect to σ^2.

$$\begin{aligned}
&\frac{\partial}{\partial\sigma^2} \log \mathcal{L}(\hat{\beta}, \sigma^2; y_1, \dots, y_n) = 0 \\
&\frac{\partial}{\partial\sigma^2} \left(-\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} ||Y - X\hat{\beta}||^2 \right) = 0 \\
&-\frac{n}{2} \cdot \frac{1}{2\pi\sigma^2} \cdot 2\pi - \frac{1}{2} ||Y - X\hat{\beta}||^2 \cdot \frac{1}{(\sigma^2)^2} = 0 \\
&\Rightarrow \hat{\sigma}^2 = \frac{||Y - X\hat{\beta}||^2}{n}
\end{aligned}$$

Notice that $Y - X\beta$ is an important vector, so important that we give it a name, since it appears in both the estimators:

$$e = Y - X\beta, \text{ called } \textbf{vector of residuals} \tag{10}$$

To recap, we have obtain that the 2 *inference* of the two *unknown* parameters are:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \text{ and } \quad \hat{\sigma}^2 = \frac{e^T e}{n} = \frac{||Y - X\hat{\beta}||^2}{n} \tag{11}$$

Now we are asking, which are their sampling distributions? Notice that in both estimators the **response vector** Y compare. And we know its distribution which is a multivariate normal distribution of $Y \sim N_m(X\beta, \sigma^2 I)$.

Distribution of β we know that $\hat{\beta} \sim N_p(?, ?)$.

- $\mathbb{E}[\hat{\beta}] = \mathbb{E}[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^T\mathbb{E}[Y] = (X^TX)^{-1}(X^TX)\beta = \beta$ This means that our predictor is **unbiased** (i.e., a good estimator).
- $VarCov[\hat{\beta}] = VarCov[\underbrace{(X^TX)^{-1}X^T}_{\text{A}}Y] = \underbrace{(X^TX)^{-1}X^T}_{\text{A}} \cdot \sigma^2 I \cdot \underbrace{X(X^TX)^{-1}}_{\text{A}^T} =$
  $= (X^TX)^{-1} \cdot (X^TX) \cdot \sigma^2 \cdot (X^TX)^{-1} = \sigma^2(X^TX)^{-1}$

So (components of *multivariate normal* are still *normal*)

$$\Rightarrow \hat{\beta}_j \sim N\left(\beta_j, \sigma^2(X^TX)^{-1}_{(j,j)}\right)$$

Regarding $\hat{\sigma}^2$ it can be proved that:

$$\frac{e^Te}{n} \sim \chi^2(n-p)$$ A *chi squared* distribution of (n-p) degrees of freedom.

Example with the Null Model In this case we have that the number of parameters p is equal to 1 (just β_0). So, we have that the sample covariance ($\hat{\sigma}^2$) is:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n} = \frac{n-1}{n} \cdot S^2$$

Where $S^2 = \frac{\sum e_i^2}{n}$ and we prefer it to estimate σ^2 since it is **unbiased**. Therefore, back to $\hat{\sigma}^2$, we have that:

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\left[\frac{n-1}{n} \cdot S^2\right] = \frac{(n-1)\sigma^2}{n}$$

So the MLE of $\hat{\sigma}^2$ tends to underestimate σ^2, that is why we prefer the **unbiased** estimate of σ^2, that in general it is (ere p indicates the number of predictors)

$$\hat{\sigma^2} = \frac{e^Te}{n-p} = \frac{\sum(Y_i - \hat{Y})^2}{n-p}$$

3.4 Exercise in R

We start analyzing some linear models using R as programming language.

Description The data set contains the waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. It is composed by 272 observations on 2 variables:

1. `eruption`: eruption time in minutes
2. `waiting`: waiting time to next eruption in minutes

Firstly, we build the linear model:

$$\texttt{eruption\_lm} \leftarrow \texttt{lm (eruptions} \sim \texttt{waiting)}$$

We apply the `lm(·)` function to a formula (`(eruptions` $\sim$ `waiting)`) that describes the variable `eruptions` by the variable `waiting` and save the linear regression model in a new variable `eruption_lm`.
Then we extract the parameters of the estimated regression equation with the coefficient function (`coefficients(·)`):

$$\texttt{coeffs} \leftarrow \texttt{coefficients (eruption\_lm)}$$

That has the output in Table 3.4

$Intercept\ (\hat{\beta}_0)$	$Waiting\ (\hat{\beta}_1)$
-1.874016	0.075628

Table 1: Output of `coefficients(·)`

Exam Question: *How to interpret* $\hat{\beta}_1$?
On average for each further minute of waiting the average eruption lasts 0.075 more.
You can use another method called: `summary(eruption_lm)`, which gives two important outputs:

1. A table that describes the distribution of the **residuals**.
2. A table that describes the description of the distribution of the **coefficients**.

Min	$1Q$	$Median$	$3Q$	Max
-1.29917	-0.37689	-0.03508	0.34909	1.19329

Table 2: Residuals

$Coeff$	$Estimate$	$Std.Error$	$t-value$	$Pr(>\|t\|)$
(Intercept)	-1.874016	0.160143	-11.70	$< 2e^{-16}$***
Waiting	0.075628	0.002219	34.09	$< 2e^{-16}$***

Table 3: Coefficients

Recall that each β_i has its distribution: $\hat{\beta}_i \sim N(\beta^3, \sigma^2(X^TX)^{-1}_{(i+1,i+1)})$.
Notice that the estimated **standard error** of each $\hat{\beta}_i$ is therefore the Mean Squared Residual (MSR): $\sqrt{MSR(X^TX)^{-1}}$.
Notice that we can standardize each component of β:

$$\frac{\hat{\beta}_i - 0}{\sqrt{MSR(X^TX)^{-1}}} \sim t(n-p) \text{ [t-value]}$$

We can estimate the $(1-\alpha)$-level **confidence interval** for β_i is:

$$\hat{\beta}_i \pm t_{\frac{\alpha}{2}}(n-p) \cdot \underbrace{\sqrt{MSR(X^TX)^{-1}_{(i+1,i+1)}}}_{\text{std. error of } \beta_i}$$

We used the t-statistic to compute the confidence of the test:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

where the T-statistics is: $T = \frac{\hat{\beta}_i - 0}{\sqrt{MSR(X^TX)^{-1}_{(i+1,i+1)}}}$ and then we can compute the p-value.

[3] Square root of the variance is the standard deviation of the sampling distribution of $\hat{\beta}_i$ which is often called **standard error**.

Recap on t-distribution It is the distribution of the difference between two Gaussian random variable (then t-distribution is Gaussian itself). Note that in the t-testing above, we subtract 0 to the coefficient because we are testing the null hypothesis H_0. When we obtain a t-value far from expectation of the difference of the two Gaussian, then the p-value is small (and the null hypothesis is rejected because the variable is significant). If the t-value is near to zero, then the p-value is high and the null hypothesis is verified (the variable is not significant).

3.5 Exercise Insulate

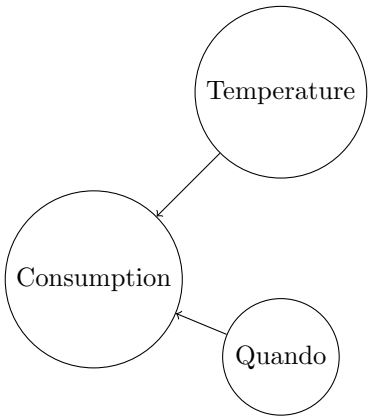
Description The data set insulate is one person's record of weekly gas consumption. It is composed by 56 observations on the following three variables:

- **Consumption**: is the fuel consumption of a certain house in several days of the year, measured in 1000 cubic feet. It represents the *response* of the linear model. (In the matrix model is y_i)
- **Temperature**: it represents the outside temperature measured in Celsius degrees. It represents a *quantitative predictor* of the linear model. (In the matrix model is x_i)
- **Quando**: it is a binary *qualitative predictor*:
  1. "prima": means the measure was taken before the house was better insulate. (In the matrix model is codified with 1)
  2. "dopo": means the measure was taken after the house was better insulate. (In the matrix model is codified with 0)

We can build the linear model in the matrix form:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\text{Observed consumption}} = \begin{bmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_n & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

We can also represent the linear model with an informal dependence graph:



We expect that the sign of β_1 to be negative (since consumption and temperature have a negative relationship), instead we expect the sign of β_2 to be (depending on our coding) positive since before insulation consumption is on average higher than after insulation.
We can plot the data to better understand the key point of the exercise that is the use of an interaction model with respect to only one additive model, Fig. 3.

We can use the additive model and the result that we obtain is in Fig. 4. As you can see the additive model have a limit which is that the two lines must be **parallel**. Now we are asking: *Can we explore more complex models?* For example, we may try a polynomial regression, that is represented by the following graph Fig. 3.5

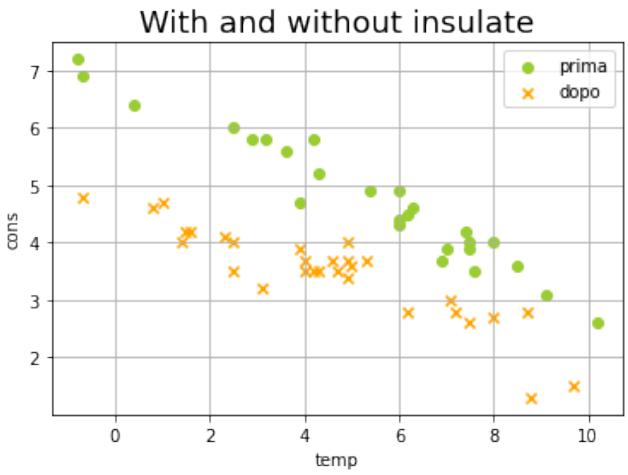


Figure 3: Insulate data points

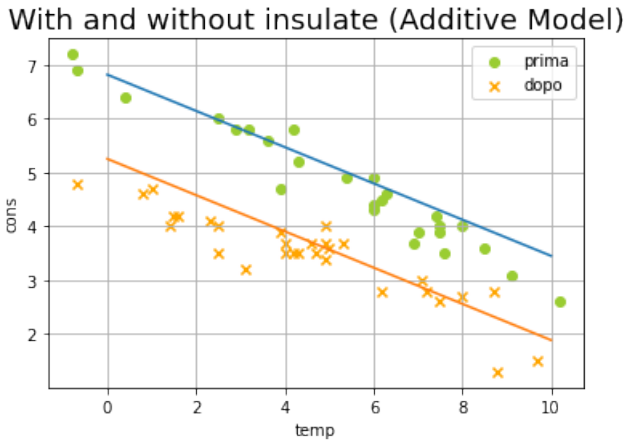
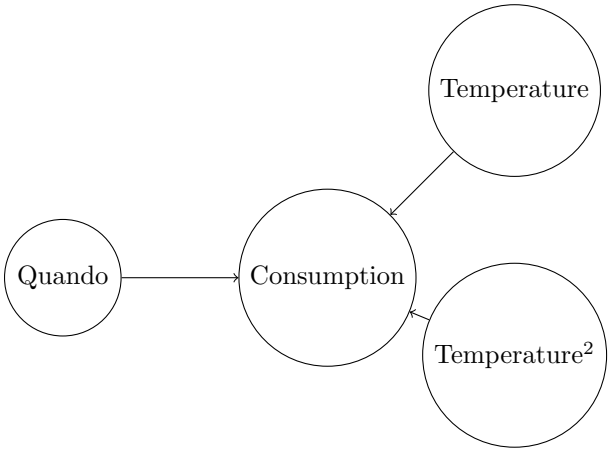


Figure 4: Additive model

Figure 5: Graph Insulate model



Fitting this model we have found that the *p-value* for `temp`^{2} is $\simeq 0.07$[4] (fairly high), which suggest that the `temp`^{2} is not very significant for our analysis. More interestingly, we could explore the possibility of an *interaction* between temperature and quando. In our case the interaction could be:

$$temperature \times \mathbb{1}_{[\text{quando=prima}]} \tag{12}$$

[4]This value was given by the R software, in particular with the following command `summary(lm(cons ~ temp+quando)`

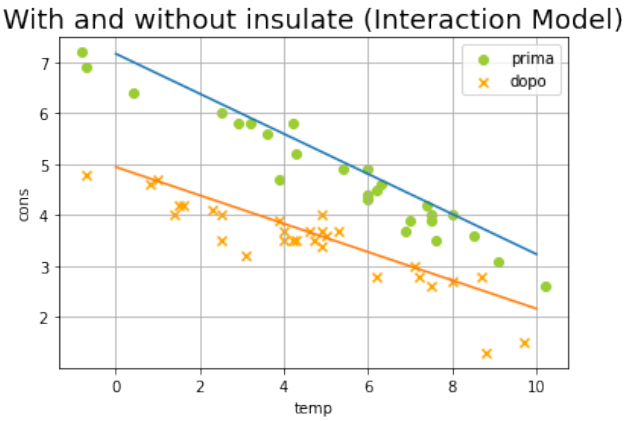


Figure 6: Interaction model

In terms of the design matrix X we obtain:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ 1 & x_2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n-1} & 1 & x_{n-1} \\ 1 & x_n & 1 & x_n \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

which means:

- $\mathbb{E}(Y) = \beta_0 + \beta_1 x$ if quando = "dopo"
- $\mathbb{E}(Y) = \beta_0 + \beta_1 x + \beta_2 + \beta_3 x = \underbrace{(\beta_0 + \beta_2)}_{\substack{\text{different intercepts, w.r.t.} \\ \text{the first additive model}}} + \underbrace{(\beta_1 + \beta_3)x}_{\substack{\text{different} \\ \text{slope}}}$ if quando = "prima"

As you can see from the Fig. 6 the interception is very significant since it better fit the data.

4 F-Test

Until now, we have used tests for individual coefficients, for a fixed i of interest:

$$\begin{aligned} H_0 &: \beta_i = 0 \\ H_a &: \beta_i \neq 0 \end{aligned} \tag{13}$$

And we reject the *null hypothesis* if the t-statistic:

$$\left| \frac{\hat{\beta}}{\sqrt{\frac{e'e}{n-p}^5 (X'X)^{-1}_{(i+1,i+1)}}} \right| > t_{1-\frac{\alpha}{2}}(n-p) \tag{14}$$

is larger than $t_{\frac{\alpha}{2}}(n-p)$. This is exactly the test that is done when using in the command `summary` of a linear model. We have used it in the example before "insulate" case study to show that:

- a $\texttt{temp}^2$ term is not necessary since the test was not significant (p-*value* $\simeq 7$ %)
- an interaction term `temp` $\times$ `quando` was instead significant, since p-*value* $\simeq 0.000731$

Notice that these are **marginal** tests, since they test for significance of the i-th predictor when **all** other predictors are kept in the model. The question that we are trying to answer with the previous test is: *Is the i-th predictor necessary or can we eliminate it, and keep the other ones?*
Often, we need to test a whole data set of predictors, and ask ourselves whether they are really necessary. This cannot be done by testing individually each predictor.

4.1 All-Or-Nothing Test

As an important example, ask if all predictors are significant:

$$\begin{gathered} H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \\ H_a : not\ H_0 \end{gathered} \tag{15}$$

Basically with this test we are asking if just β_0 (intercept) is significant to obtain the response. We can see this problem in matrix form as the following:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Where $y_1, \ldots, y_n \sim N(\beta_0, \sigma^2)$ and they are **I.I.D**. The parameter β_0 can be estimated using the MLE and it concides with the sample mean:

$$\beta_0 = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{Y} \tag{16}$$

Recap: we have seen a test for all the parameters (All-Or-Nothing) and a test for a single predictor. *Why do not try to test a subset of predictors?*

4.2 Nested Model

Let's take X as a matrix (n×p), which is the form of:

$$X = [X_1, X_2] \tag{17}$$

where X_1 is a model matrix (n×k) and X_2 is a model matrix (n×(p-k)). So the linear models become:

$$\begin{aligned} Y_1 &= X_1\beta_1 + \varepsilon_1 \\ Y_2 &= X_2\beta_2 + \varepsilon_2 \end{aligned}$$

These models are called nested models within the linear model $Y = X\beta + \varepsilon$.

Suppose we wish to use whether the full model matrix X or reduced matrix X_1. Moreover, suppose that $\underbrace{\hat{\beta} \text{ is the estimate of } \beta}_{\text{Full Model}}$ and $\underbrace{\hat{\beta_1} \text{ the estimate of } \beta_1}_{\text{Reduced Model}}$.

- Let $Y^{(2)} = X\hat{\beta}$ be the projection of Y onto the space of SPAN(X) spanned by the columns of X.
- Let $Y^{(1)} = X_1\hat{\beta_1}$ be the projection of Y onto the space of SPAN(X_1) spanned by the columns of X_1.

In order to decide whether the features of X_2 are needed, we may compare the estimated error[6] terms of the 2 models (the full rank and the reduced one).

If the outcome of this comparison is that: there is a little difference between the two models, then it is appropriate to use the *reduced model*, because it has fewer parameters than the full rank model, while explaining the data just as well. The comparison is done between the squared norms (which correspond to the *Euclidean distance*), so between $\underbrace{\|Y - Y^{(2)}\|^2}_{full\ model\ X}$ and $\underbrace{\|Y - Y^{(1)}\|^2}_{reduced\ model\ X_1}$.

Because of the *nested nature* of the Linear Models $SPAN(X_1)$ is a sub-space of $SPAN(X)$, consequently the projection of $Y^{(2)}$ onto the $SPAN(X_1)$ is the same as the orthogonal projection of Y onto $SPAN(X_1)$. By Pitagoras' theorem:

$$\|Y^{(2)} - Y^{(1)}\|^2 + \|Y - Y^{(2)}\|^2 = \|Y - Y^{(1)}\|^2 \tag{18}$$

The concept can be visualized graphically in the figure 7:

We reject the hypothesis of using only a subset, H_0, when the distance between the full model and the reduced model (composed by the excluded terms in H_0) is too large.

[6]The error is computed by the Sum of Residuals Squares

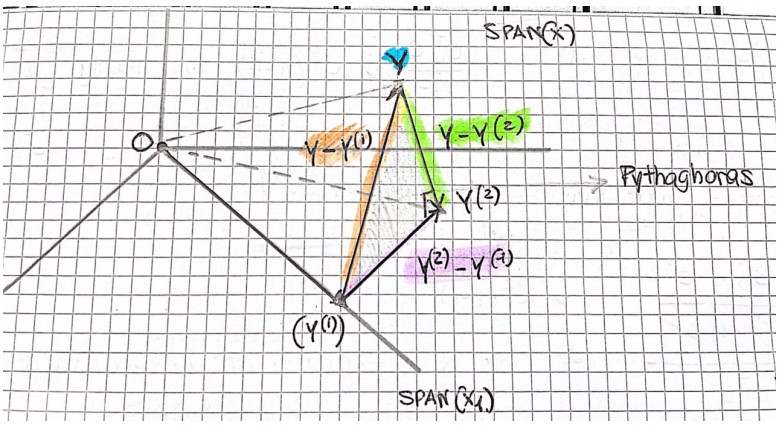


Figure 7: Geometric meaning of F-test

4.3 Example

Take again the *All-Or-Nothing* test, where the hypothesis to be checked is:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$
$$H_a : \ not\ H_0$$

In this case we are comparing the full model with respect to the model with only the intercept and we reject when: ($\hat{Y_0}$ is the estimate of null model and $\hat{Y}$ is the estimate of the full model)

$$\underbrace{e_0^T e_0}_{\|Y_0 - \hat{Y_0}\|^2} - \underbrace{e^T e}_{\|Y - \hat{Y}\|^2} = \|\hat{Y} - \hat{Y_0}\|^2$$

is too large. How do we define the size to be rejected? We use a statistic called F-statistic. We already know that $\frac{e^T e}{\sigma^2} \sim \chi^2(n-p)$ and similarly $\frac{e_0^T e_0}{\sigma^2} \sim \chi^2(p-1)$ and if the null hypothesis is true:

$$\frac{\frac{e_0^T e_0 - e^T e}{(p-1)}}{\frac{e^T e}{(n-p)}}$$

This is a F-distribution, because it is division of two χ^2 distributions, in particular is and F-distribution $\sim F(p-1, n-p)$.

Otherwise, if the null hypothesis is not true, that ratio tends to be larger, since the two projections $\hat{Y}$ and $\hat{Y_0}$ are too different. Therefore, it makes sense to reject the null hypothesis if:

$$F = \frac{e_0^T e_0 - e^T e}{e^T e} \cdot \frac{n-p}{p-1} > F_\alpha(p-1, n-p)$$

Notice that we can also test different sub-spaces, which can be done by using the ANOVA (Analysis of Variance) which takes into account a small and large model, respectively the reduced and full model. Moreover, you can notice that in the case we are testing a reduced model with just one predictor it can be shown that the F-test is equivalent to a marginal T-test for that predictor.

Now, we start analyzing applications of F-test with **binary predictor**, such as `sex`.

4.4 Example Qualitative binary predictor

Suppose that we have just one predictor: `sex` which is a qualitative and binary. we can represent the model in the matrix form in the following way:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Here, β_1 expresses the extra variation when the feature (`sex`) is present, with respect when it is not present. We can see β_1 as the difference in means with respect to females and males: $\beta_1 = \mu_{\text{male}} - \mu_{\text{female}}$. In fact the confidence interval given by the (F-test or T-test)[7] is equivalent to the confidence interval for testing the difference between two means.

4.5 Example 2 binary predictors

Suppose now that we have the following data set in Table 4, where:

- `male`: indicates the sex of the person and if he is a male it is 1.
- `education`: indicates the level of education of the person and it can *low* or *high*.
- `salary`: this is the response and it indicates the salary of the person.

$Salary$	$Male$	$Education$
y_1	0	L
y_2	0	H
$\vdots$	$\vdots$	$\vdots$
y_{n-1}	1	H
y_n	1	L

Table 4: Salary table

We can build the linear model in the matrix form as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Notice that it is a **full rank matrix** and **additive model**. Let's analyze the meaning of each coefficient:

- β_0 it is the mean response for **females** with **low** education level.
- $\beta_0 + \beta_1$ it is the mean response for **males** with **low** education level.
- $\beta_0 + \beta_2$ it is the mean response for **females** with **high** education level.
- $\beta_0 + \beta_1 + \beta_2$ it is the mean response for **males** with **high** education level.

Since the model is *additive* it must respect the following equation:

$$\beta_2 = \mu_{MH} - \mu_{ML} = \mu_{FH} - \mu_{FL} \tag{19}$$

Where β_2 represents the effect of high with respect to low education level, no matter the value of the sex feature. Similarly β_1 is the main effect of male with respect to female no matter what education level is. This is **additive specification** of the model since we do not have interactions.
We can build the interaction model:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The third column of the X matrix represent the interaction, which as we have already seen can be represented as a product: `male` $\times$ `education`.

[7] Since we are testing just one predictor the two tests are equivalent

4.6 More than binary qualitative predictors

Remember that in linear models, having a single binary predictor corresponds to have a 2 sample normal problem, which can be done by confidence interval for the difference of means of $\mu_1 - \mu_2$.
For example, look at the matrix form below:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Therefore, having a single *more-than binary predictors* corresponds to a comparison of a p-normal group instead of just 2.

Notice that, in coding, a p-levels predictor one level is absorbed into the intercept then we have p-1 columns.

Example 4.1 *Suppose we have 4 brands (A, B, C, D) for golf's balls and we want to compare them with respect to the distance that they are able to achieve. We can expect that the distributions of all balls are normal with a different mean for each ball brand, but the same variance.*

Let's analyze it with R. We start by building the linear model: `golf = lm(distance ~ brand)`*, let's use the following command:* `summary(golf)` *and let's analyze the output, which is composed of 3 different parts:*

1. ***Residuals table*** *in table 5*
2. ***Coefficients table****: in table 6*
3. ***F-statistic:*** `0.2219 on 3 and 28 DF, p-value: 0.8804`

Min	$1Q$	$Median$	$3Q$	Max
-48.638	-31.703	-0.481	32.947	42.475

Table 5: Residuals

$Coeff$	$Estimate$	$Std.Error$	$t-value$	$Pr(>\|t\|)$
(Intercept)	199.862	12.262	16.299	8.04e-16 ***
brandB	8.333	17.341	0.481	0.634
brandC	5.275	17.341	0.304	0.763
brandD	-4.737	17.341	-0.273	0.78

Table 6: Coefficients

An alternative way to do the *All-Or-Nothing* global F test is doing the ANOVA, which can be done in R with the following easy command `anova(lm(distance ~ brand))`.
As you can see from the results obtained in the table 6 the ball brand alone are not significant predictors, so we can study how the additive model given by brand and club works.
To analyze that we need to use one more binary predictor to individuate the club, so the new data set becomes 7: We can build the linear model in R by using the following command `lm(distance ~ brand + club)`. The

$(Intercept)$	B	C	D	$clubIRON$
1	0	0	0	0
1	1	0	0	0
1	0	1	0	1
⋮	⋮	⋮	⋮	⋮
1	1	0	0	0
1	0	1	0	0
1	0	0	1	1
1	0	0	1	1

Table 7: Golf data set

'+' indicates an *additive* model, where the effects of one factors simply add to the effects of the other level (no

	Df Sum	Sq mean	F value	Pr(>F)
brand	3	801	267	0.01061 *
club	1	32093	32093	$< 2e^{-16}$ ***

Table 8: Golf data set

interaction are allowed).
We now have two interesting F-tests: **N.B** the first line of the table 8is used testing brands group : $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. The second line is used to test the differences between club groups $H_0 : \beta_4 = 0$ ANOVA test is a way to show two tests together.

We can go further, allow of interactions and ask ourselves whether the model can be reduced to additive or not. We can do this in R with the following line of code: `golf_interaction = lm(distance ~ brand × club)`.
The matrix of predictors become:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{bmatrix}$$

Where the last three $(\beta_5, \beta_6, \beta_7)$ are the coefficients of the interactions. Now, by doing the following command `summary(golf_interaction)` which give us the following table: It is first most interesting to test whether any

	Estimate	Std. Error	t-value	$Pr(> \|t\|)$
(Intercept)	228.425	2.927	78.051	$< 2e^{-16}$***
brandB	5.300	4.139	267	0.21259
brandC	14.675	4.139	267	0.00165 **
brandD	1.325	4.139	267	0.75163
clubIRON	-57.125	4.139	-13.502	$6.552e^{-13}$ ***
brandB: clubIRON	6.075	5.853	267	0.30966
brandC: clubIRON	-18.800	5.853	267	0.00373 **
brandD: clubIRON	-12.125	5.853	267	0.04923*

Table 9: Golf data set

interaction at all is present: $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$. It can be tested in two ways:

- Using `anova(large)`
- Using `anova(small, large)`

This problem can be represented in the following way:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Let's see the meaning of the predictors:

- β_0 is the average response when factor is set at the *reference* level.
- $\beta_i, \forall i \in [0, p-1]$ are the **differential mean** responses of other levels with respect to the reference level.

Take home message: when analyzing a full factorial design with two factors A and G:

1. Fit first the `lm (Y ~ A × G)` and tests for interactions.
2. If interactions are not significant, switch to an additive model `lm (Y ~ A + G)`

3. Next, if you suspect G (or A) may not be influential, test for main effects of G (or A, respectively): $H_0 : \beta_G = 0$, where β_G refers to all beta's related to G. (or $H_0 : \beta_A = 0$) if one (or both) of the, are not significant, you can switch to a one-factor (or null) model.

In the golf example, interactions appear to be all significant, with the combination BRAND C and CLUB DRIVER appearing to give the highest mean distance.
To conclude, testing:

1. interactions between pairs of factors.

2. main effects of single factors.

Are important contributions to the *interpretability* of your methods.

In the full factorial approach so far A and G were symmetric (in the golf example,we are formally interested in both BRAND and CLUB). In another famous design, one factor is of interest, whereas the other factor, the blocking factor is a necessary nuisance. In the golf example we may be interesetd in BRANDS and have them tested not by a robot, but by players (people). Players here are a blocking factor, meaning each player plays with A,B,C,D (randomixed order).

Finally, let's add that sometimes a full factorial design is not possible, in which case we choose only a subset of combinations of levels and do a fractionaò factorial design. ($\rightarrow$ Latin Square Design)

Example 4.2 *Suppose you have the data set represented in table 10 we want to analyze with R:*

club	*A*	*B*	*C*	*D*
DRIVER	226.4	238.3	240.5	219.8
DRIVER	232.6	231.7	246.9	228.7
DRIVER	234	227.7	240.3	232.9
DRIVER	220.7	237.2	244.7	237.6
IRON	163.8	184.4	179	157.8
IRON	179.4	180.6	168	161.8
IRON	168.6	179.5	165.2	162.1
IRON	173.4	186.2	156.5	160.3

Table 10: Golf data set

1. First thing to do is to build the linear model: `lm(distance ~ brand + club)`

2. Now let's estimate the values of the coefficients, we can easily do that by using: `summary(lm(distance ~ brand)`, which output is the following:
Residuals Coefficients

Min	1*Q*	*Median*	3*Q*	*Max*
-48.638	-31.703	-0.481	32.947	42.475

Coeff	*Estimate*	*Std.Error*	$t - value$	$Pr(>\|t\|)$
(Intercept)	199.862	12.262	16.299	8.04e-16 ***
brandB	8.333	17.341	0.481	0.634
brandC	5.275	17.341	0.304	0.763
brandD	-4.737	17.341	-0.273	0.78

Exam Question: What does -4.737 represents?
It is the estimate of the difference between mean response corresponding to BRAND C and the reference which is BRAND A.

The following is the result of the *The All-Or-Nothing* F-test, studying the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \tag{20}$$

`F-statistic: 0.2219 on 3 and 28 DF, p-value: 0.8804` The *The All-Or-Nothing* test is not significant, which mean that BRAND A alone is not a significant predictor, on the other hand CLUB is significant.

5 Confidence Intervals and Prediction Intervals

In many situations, linear regression is most useful when we wish to predict how a new response variable will behave on the basis of a new explanatory vector $\boldsymbol{x}$. For example, it may be difficult to measure the response variable, but by knowing the estimated regression line and the value for $\boldsymbol{x}$, we will have a reasonable good idea what Y or the expected value of Y is going to be.
Seeing the problem in another way, we have the following training data set:

$$\begin{bmatrix} y_1 & 1 & x_{1,1} & \dots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_n & 1 & x_{n,1} & \dots & x_{n,p-1} \end{bmatrix}$$

and we have a new sample:

$$\begin{bmatrix} 1 & x_{n+1,1} & \dots & x_{n+1,p-1} \end{bmatrix} \tag{21}$$

We are interested to the corresponding response, Y_{n+1}, in particular we may be interested in two values:

1. The mean $\mathbb{E}(Y_{n+1})$ (*inference problem on a function*)
2. The actual value of Y_{n+1} (*prediction problem*)

5.1 Inference Problem

Consider the new sample $\mathbf{x}$ in equation 21 and let its response $Y \sim N(\mathbf{x}^T\beta, \sigma^2)$ with β and σ^2 unknown. First we are looking at the *expected value* of Y, that is:

$$\mathbb{E}(Y_{n+1}) = \mathbf{x}^T\beta \tag{22}$$

We know the value of $\mathbf{x}$, but we do not know the value of β and for this reason we also do not know the expected value of Y.
Of course, the **point estimate** of $\mathbb{E}(Y_{n+1})$ and the **point prediction** of Y_{n+1} will be the same:

$$\hat{Y}_{n+1} = \mathbf{x}^T\hat{\beta} \tag{23}$$

But the **confidence interval** for $\mathbb{E}(Y_{n+1})$ will be different from the **prediction interval** for Y_{n+1}, but both centered in the estimate of Y_{n+1}.

Confidence Interval Let's start with confidence interval for $\mathbb{E}(Y_{n+1})$. We know that $\mathbf{x}^T\hat{\beta}$ is a random variable (an estimator). What is its sampling distribution?

$$\hat{\beta} \sim N_p(\beta, \sigma(X^TX)^{-1}) \tag{24}$$

Therefore,

$$\mathbf{x}^T\hat{\beta} \sim N_p(\mathbf{x}\beta, \mathbf{x}VarCov(\hat{\beta})\mathbf{x}^T) = N_p(\mathbf{x}\beta, \mathbf{x}\sigma(X^TX)^{-1}\mathbf{x}^T)$$

As usual:

$$\frac{\mathbf{x}\hat{\beta} - \mathbf{x}\beta}{\sqrt{(\mathbf{x}\sigma^2(X^TX)^{-1}\mathbf{x}^T)\underbrace{\frac{e^Te}{n-p}}_{\text{MSR, estimate of } \sigma^2}}} \sim t(n-p) \tag{25}$$

And we can build the **confidence interval** in the following way:

$$\mathbf{x}\hat{\beta} \pm t_{\frac{\alpha}{2}(n-p)} \cdot \sqrt{\frac{e^Te}{n-p}} \cdot \sqrt{\mathbf{x}(X^TX)^{-1}\mathbf{x}^T}$$

5.2 Inference Problem

What about predictions?

$$\underbrace{Y_{n+1} - \mathbf{x}\hat{\beta}}_{\text{two random variables}} \qquad \sim N(?, ?)$$

What distributions does it have? Notice that they are difference of random variables, so we can find out mean and variance.

- **Mean:**

$$\mathbb{E}(Y_{n+1} - \mathbf{x}\hat{\beta}) = \mathbb{E}(Y_{n+1}) - \mathbb{E}(\mathbf{x}\hat{\beta}) = \mathbf{x}\beta - \mathbf{x}\beta = 0$$

- **Variance**

$$Var(Y_{n+1} - \mathbf{x}\hat{\beta}) = Var(Y_{n+1}) - Var(\mathbf{x}\hat{\beta}) - 2Cov(Y_{n+1} - \mathbf{x}\hat{\beta})^{8} = \sigma^2 + \sigma^2\mathbf{x}(X^TX)^{-1})\mathbf{x}^T = \sigma^2(1 + \mathbf{x}(X^TX)^{-1}\mathbf{x}^T)$$

Given those we obtain that:

$$Y_{n+1} - \mathbf{x}\hat{\beta} \sim N(0,\ \sigma^2(1 + \mathbf{x}(X^TX)^{-1}\mathbf{x}^T))$$

And based on the probabilistic property:

$$\mathbb{P}\left(-t_{\frac{\alpha}{2}(n-p)} \leq \frac{Y_{n+1} - \mathbf{x}\hat{\beta}}{\sqrt{\mathbf{x}\sigma^2(X^TX)^{-1})\mathbf{x}^T)\frac{e^Te}{n-p}}} \leq t_{\frac{\alpha}{2}(n-p)}\right) = 1 - \alpha$$

I can isolate the **prediction interval** which is:

$$\mathbf{x}\hat{\beta} \pm t_{\frac{\alpha}{2}(n-p)} \cdot \sqrt{\frac{e^Te}{n-p}} \cdot \sqrt{1 + \mathbf{x}(X^TX)^{-1}\mathbf{x}^T}$$

The 1+ under the square root makes the prediction interval wider.

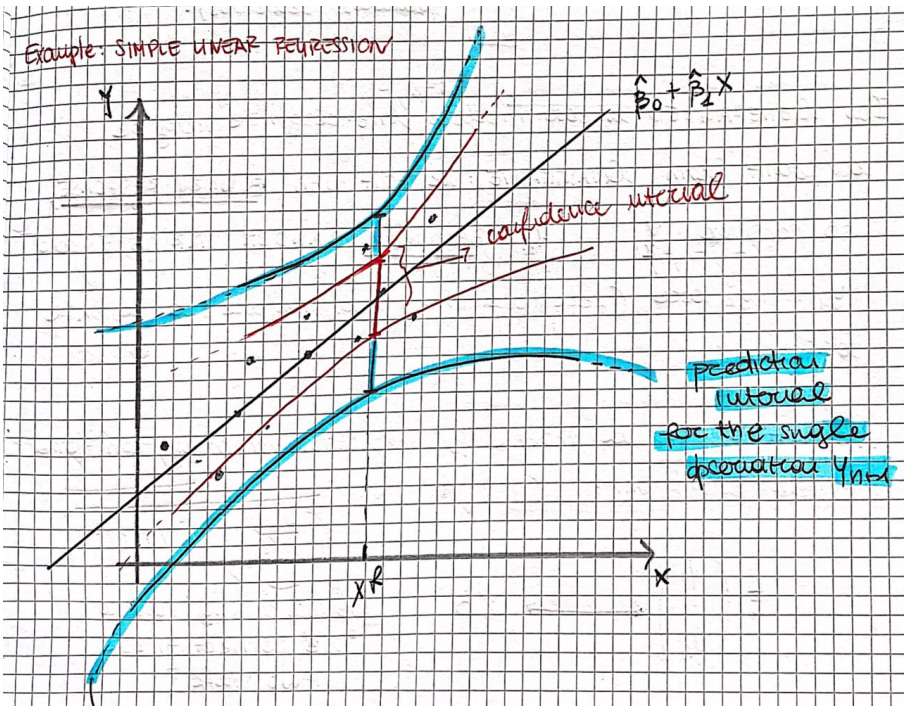


Figure 8: Simple Linear Regression

5.3 Exercises on linear models

Exercise 1. Replicate in Python the results of this short version of the R analysis of the insulate data.

```
# reading the data
insulate=read.table("220330insulate.txt", col.names=c("quando","temp","cons"))
attach(insulate) # attach/detach dynamics
# fitting a regression model for consumption with interaction
summary(regr2)
# obtain 90% confidence intervals for the regression coefficients
confint(regr2, level=0.9)
detach(insulate) # attach/detach dynamics
```

The following python codes reproduces the results of the R analysis:

[8] If Y_{n+1} is *independent* from $Y_1, \ldots, Y_n$ the covariance = 0

```
# import useful libraries
import pandas as pd
from statsmodels.formula.api import ols
# reading the data
insulate = pd.read_table("220330insulate.txt", sep="\textbackslash_s+", \
           header=None, names=["quando", "temp", "cons"])
# create the interaction column
insulate["int"] = insulate["temp"] * insulate["quando"].map({"dopo": 1., "prima": 0.})
# fitting a regression model for consumption with interaction
regr2 = ols("cons~quando+temp+int", data=insulate).fit()
print(regr2.summary())
# obtain 90% confidence intervals for the regression coefficients
regr2.conf_int(.1).rename(columns={0: "[0.05", 1: "0.95]"})
```

Exercise 2. The distances obtained having 10 different golf players hit four different brands of golf balls are collected. We are interested in comparing brands, not players. The experiment is balanced and randomized, since each player hits a ball of each of the four different brands in a randomized order. Such an experimental plan is called a randomized block design. The example taken from McClave JT., Benson PG. e Sincich T. (2014). Statistics for Business and Economics. Pearson Education.

a) Read in the data and transform to long format.

```
golf = pd.read_table('golf.csv', sep='\s+')
```

b) Fit an additive model containing both BRAND and GOLFER as qualitative predictors (factors) for DISTANCE.

```
model = ols("DISTANCE~BRAND+GOLFER", golf).fit()
```

c) It would not be possible to introduce all brand by player interaction terms with this data. Why?
- The additive model features 1 intercept and a total of 12 parameters:

- 4 brands are encoded with 3 binary variables
- 10 golfers are encoded with 9 binary variable

Therefore, there are 27 possible interaction terms of BRAND*GOLFER for a total of $p = 12 + 27 = 39$ parameters to be estimated. However, there are only $m = 40$ observation, thus $p < m$ and it is not possible to fit this kind of model.

d) Test for BRAND differences; are they significant? Are they important? Which is the best brand?
- Yes, those differences are important because we are interested in BRAND effects. By performing ANOVA we get significant p-values ≈ 0

e) Test for GOLFER differences; are they significant? Are they important?
- By perfoming ANOVA of GOLFER differences we also get significant p-values but we are not interested in measuring the performances of the different golfers, therefore it is not meaningful for the analysis.

f) Calculate a 95% confidence interval for the mean distance difference between BRAND C and BRAND A.

```
model.conf_int(.05).rename(columns={0: "[0.025", 1:"0.975]"})
```

It is possible to see that the estimated difference BRANDC - BRANDA lies in the range $[9.70, 26.86]$ with 95% confidence.

g) Calculate a 95% confidence interval for the mean distance difference between BRAND C and BRAND A without accounting for GOLFER and comment on the difference with the previous one.
- We fit a new model without accounting for GOLFER

```
model2 = ols("DISTANCE~BRAND", golf).fit()
model2.conf_int(.05).rename(columns={0: "[0.025", 1:"0.975]"})
```

Instead, if we do not account for GOLFER but for BRAND effect alone the estimated difference between brand A and C is in the range $[1.65, 34.81]$ The confidence interval for the difference between brandC and brandA is much smaller when accounting for GOLFER.

Therefore, having GOLFER as a blocking factor, improves precision, diminishes variability of our estimates.

5.4 Binary Regression and Generalized Linear Models

So far, we have dealt with a quantitative response

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np-1} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

And the predictors can be both quantitative and (coded) qualitative features.
In **Generalized** Linear Models, the response may be different from quantitative: we start with a binary response.
Consider $Y_1, \dots, Y_n$ Bernoulli random variable and their observed values can be either 0 or 1. We want to keep a matrix of predictors linearly transformed through β coefficients.

$$\mathbb{E} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \overset{?}{\Longleftrightarrow} X\beta$$

We know that for binary y_i, $\mathbb{E}[y_i] = p$ (probability of $y_i = 1$). If we suppose to have only one quantitative predictor x we get a situation like this: Here, the lines not getting out of $[0, 1]$ are useless. So, we cannot directly connect p_i and x through $\beta_0 + \beta_1 x$. We need a transformation of p_i that maps $[0, 1] \mapsto \mathbb{R}$. A very popular mapping is the logit transformation

$$\text{logit}(p) = \log \frac{p}{1-p}$$

Now, it is possible to connect $\text{logit}(p) = X\beta$ because the output of the logit is a number in $[-\infty, +\infty]$. This is called **logistic regression** a special Generalized Linear Model. GLM are characterized by:

1. A response which is *not* necessarily quantitative observed in n independent units, therefore $Y_1, \dots, Y_n$ are independent but not necessarily normal, as in the linear models seen so far
2. p predictors in the design matrix X are linearly combined via $\beta_0, \dots, \beta_{p-1}$: $\underset{n \times p}{X} \quad \underset{p \times 1}{\beta}$
3. A link function (e.g., the logit) that transforms the expectations of Y into real numbers

$$\text{link}(\mathbb{E}[Y_i]) = X_{(i)} \cdot \beta$$

The inverse of the link is called activation function, which is the sigmoid function $S(x) = \frac{1}{1+e^{-x}}$.

The resulting distributions and estimators are more complex than the normal linear models.

6 Introduction to Bayesian Statistics

6.1 Bayes theorem for elementary probability

- Ω is our sample space, or outcome space, i.e., the set of all possible outcomes of a random experiment.
- $\mathbb{P}$ is a probability or certain subsets of Ω, e.g., $\mathbb{P}(A)$ is the probability of event $A \subseteq \Omega$

Example $\Omega = \{(i,j) : i,j = 1,\dots 6\}$ represents the possible outcomes of tossing two dices. A possible probability is the uniform probability in case the two dices are fair and independent, i.e.,

$$\mathbb{P}(\{(i,j)\}) = \frac{1}{36} \ \forall i,j$$

Consider a partition of Ω $H_1, H_2, \dots$ (i.e., a family of events, subsets) such that:

- $H_i \cap H_j = \emptyset$ if $i \neq j$ (mutually exclusive)
- $\bigcup_i H_i = \Omega$ (exhaustive, i.e., they cover the entire set Ω)

For a generic event E, we can write $E = \bigcup_i (E \cap H_i)$ thanks to the mutually exclusiveness, then

$$\mathbb{P}(E) = \mathbb{P}(\bigcup_i (E \cap H_i)) = \sum_i \mathbb{P}(E \cap H_i) = \sum_i \mathbb{P}(H_i) \cdot \mathbb{P}(E|H_i)$$

thanks to the total probability law and the definition of conditional probability. Now, suppose that we are interested in a specific H_k, then

$$\mathbb{P}(H_k|E) = \frac{\mathbb{P}(H_k \cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|H_k) \cdot \mathbb{P}(H_k)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|H_k) \cdot \mathbb{P}(H_k)}{\sum_i \mathbb{P}(H_i) \cdot \mathbb{P}(E|H_i)}$$

The last is the formulation of the Bayes theorem. In the mind of Bayes, Pierce and Laplace, at the end of '700, $H_1, H_2, \dots$ are possible states of the world, whereas E is *evidence*, retrieved from data. Roughly speaking, we do not know what the state of the world is, but we have some prior ideas about it expressed with the prior probabilities $\mathbb{P}(H_1), \mathbb{P}(H_2), \dots$ We then perform an experiment and collect the evidence E occurred. So, we update accordingly to E our prior probabilities into posterior using Bayes theorem. Thus, we are going from $\mathbb{P}(H_i)$ to $\mathbb{P}(H_i|E)$, which is called *Bayesian learning* and it is the basic operation of *Bayesian statistics*.

6.1.1 Example: diagnostic tests

We want to test some subjects, picked at random from a population, in order to find a specific disease. In this case, our sample space Ω is divided in two partitions:

- D if the subject has the disease
- $\overline{D}$ if the subject does not have the disease.

Our prior knowledge is expressed in the following probabilities:

- $\mathbb{P}(T_+|D) = 0.7$ is the *sensitivity*, i.e., how much is good the test to reveal a diseased subject.
- $\mathbb{P}(T_-|\overline{D}) = 0.95$ is the *specificity*, i.e., how much is good the test to reveal a not diseased subject.
- $\mathbb{P}(D) = 0.01$ is the *prevalence*, i.e., how diffused is the disease in the population, fixed the time. Note that it may change over the time.

Now we perform the test on a subject and this result positive. We can now update $\mathbb{P}(D)$ as follow.

$$P(D|T_+) = \frac{P(D) \cdot P(T_+|D)}{P(D) \cdot P(T_+|D) + P(\overline{D}) \cdot P(T_+|\overline{D})} = 0.124$$

A very low number. This is due to a very low prevalence and to the fact that we are doing screening (and not, for example, testing people with symptoms or contacts). A possible solution to this unfortunate situation is to take another test on the same subject, independently w.r.t. the first test. Then we have a new posterior probability which is the following.

$$P(D|T_{1+}, T_{2+}) = \frac{P(D) \cdot P(T_{1+} \cap T_{2+}|D)}{P(D) \cdot P(T_{1+} \cap T_{2+}|D) + P(\overline{D}) \cdot P(T_{1+} \cap T_{2+}|\overline{D})} \approx 0.66$$

where $P(T_{1+} \cap T_{2+}|D) = P(T_{1+}|D) \cdot P(T_{2+}|D)$ thanks to the independence of the two tests. We will rewrite these results (expressed in terms of events) using binary random variables called indicators.

$$D? = \begin{cases} 0 \text{ if the subject is not diseased} \\ 1 \text{ if the subject is diseased} \end{cases} \quad T_1? = \begin{cases} 0 \text{ if test 1 is negative} \\ 1 \text{ if test 1 is positive} \end{cases} \quad T_2? = \begin{cases} 0 \text{ if test 2 is negative} \\ 1 \text{ if test 2 is positive} \end{cases}$$

They have discrete densities, also called probability mass functions.

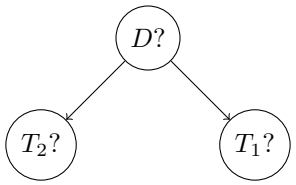
6.2 Elementary Bayesian Networks

We define the random vector $(D?, T?)$ with joint density

$$f_{D?,T?}(x, y) = f_{D?}(x) \cdot f_{T?|D?}(y|x)$$

This formulation suggests the following Directed Acyclic Graph (DAG).

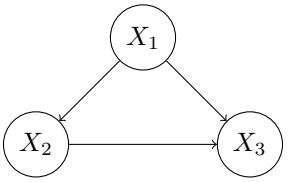
This is a natural representation of the way we were given the diagnostic problem. It is also a way to represent the joint density. You have the prior probability of $T?|D?$, and you update your knowledge going back from $T?$ (the evidence) to $D?$, obtaining $P(D?|T?)$. This operation is also called *bayesian updating* or *evidence propagation*. The DAG in case of two independent tests $T_1?$ and $T_2?$ is the following.



Note that $T_1?$ and $T_2?$ are on different paths since they are conditionally independent given $D?$. So, a Bayesian network is a live representation of a joint density of a random vector where:

1. Vertices are the components of the vector
2. Edges represent various dependencies / conditional independencies between nodes.

In case of fully connected DAG, we show how to compute the joint density for a random vector with 3 components.



We have that

$$f_{X_1,X_2,X_3}(x_1, x_2, x_3) = f_{(X_1,X_2)}(x_1, x_2) \cdot f_{(X_3|X_1,X_2)}(x_3|x_1, x_2) = f_{X_1}(x_1) \cdot f_{(X_2|X_1)}(x_2|x_1) \cdot f_{(X_3|X_1,X_2)}(x_3|x_1, x_2)$$

Note that if we can cancel some of the arrows (to express conditional independence), then calculations are simplified. In general, we have that the joint density can be expressed as

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \prod_{X_i \in nodes} f_{(X_i|\text{Parents}(X_i))}(x_i|\text{parents}(x_i))$$

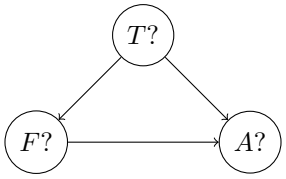
6.3 The fraud detection example

Text Suppose you are working for a financial institution and you are asked to build a fraud detection system. You plan to use the following information. When the card holder is traveling abroad, fraudulent transaction are more likely since tourists are prime targets for thieves. More precisely, 2% of transactions are fraudulent when the card holder is traveling, whereas only 1% of the transactions are fraudulent when he is not traveling. On average, 5% of all transactions happen while card holder is traveling. If a transaction is fraudulent, then the likelihood of a purchase abroad increases, unless the card holder happens to be traveling. More precisely, when the card holder is not traveling, 10% of the fraudulent transactions are abroad purchases, whereas only 1% of the legitimate transactions are abroad purchases. On the other hand, when the card holder is traveling, 90% of the transactions are abroad purchases regardless of the legitimacy of the transactions.

Solution We deal with the following binary random variables.

$$T? = \begin{cases} 1 \text{ if traveling} \\ 0 \text{ otherwise} \end{cases} \quad F? = \begin{cases} 1 \text{ if transaction is fraudolent} \\ 0 \text{ otherwise} \end{cases} \quad A? = \begin{cases} 1 \text{ if transaction was done abroad} \\ 0 \text{ otherwise} \end{cases}$$

The DAG is the following.



Note that $T?$ is not dependent by any other random variable, and its probability distribution is:

$$f_{T?}(t) = \begin{cases} 0.05 \text{ if } t = 1 \\ 0.95 \text{ if } t = 0 \end{cases}$$

(5% of all transactions happen while card holder is traveling) Therefore, it is a Bernoulli random variable, i.e., $T? \sim Bernoulli(0.05)$. Furthermore, from the text, we know that:

- $P(F = 1|T = 1) = 0.02$ and $P(F = 0|T = 1) = 0.98$ (2% of transactions are fraudulent when the card holder is traveling)
- $P(F = 1|T = 0) = 0.01$ and $P(F = 0|T = 0) = 0.99$ (1% of the transactions are fraudulent when he is not traveling)
- $P(A = 1|T = 0, F = 1) = 0.1$ and $P(A = 0|T = 0, F = 1) = 0.9$ (when the card holder is not traveling, 10% of the fraudulent transactions are abroad purchases)
- $P(A = 1|F = 0) = 0.01$ and $P(A = 0|F = 0) = 0.99$ (1% of the legitimate transactions are abroad purchase)
- $P(A = 1|T = 1 \cap F = 0, 1) = 0.9$ and $P(A = 0|T = 1 \cap F = 0, 1) = 0.1$ (when the card holder is traveling, 90% of the transactions are abroad purchases regardless of the legitimacy of the transactions.)

We now apply the formula needed to find the joint density of the random vector $(T?, F?, A?)$.

$$\begin{aligned} \pi_{(T?,F?,A?)}(t, f, a) &= \prod_{I \in \{T?,F?,A?\}} \pi_{(I|\text{Parents}(I))}(i|\text{parents}(i)) \\ &= \pi_T(t) \cdot \pi_{F|T}(f|t) \cdot \pi_{A|T,F}(a|t, f) \end{aligned}$$

We have that the distribution $\pi_{(A|F,T)}(a|f, t)$ is

t	f	$\pi_{(A\|F,T)}(0\|f,t)$	$\pi_{(A\|F,T)}(1\|f,t)$
0	0	0.99	0.01
0	1	0.9	0.1
1	0	0.1	0.9
1	1	0.1	0.9

Thus we have the following joint distribution.

t	f	a	$\pi(t, f, a)$
0	0	0	0.931095
0	0	1	0.009405
0	1	0	0.00855
0	1	1	0.00095
1	0	0	0.0049
1	0	1	0.0441
1	1	0	0.0001
1	1	1	0.0009

Once you have the joint distribution, you can answer to any question about conditional probabilities, given any conditioning (evidence) configuration.

6.4 Review of conditional expectation

Given two random variables (X, Y) (but the extension to random vectors is trivial) with joint density $f_{(X,Y)}(x, y)$, the conditional expectation of Y given $X = x$ is

$$\mathbb{E}(Y|X = x) = \begin{cases} \sum_j y_j f_{(Y|X)}(y_j|x) \text{ if (X,Y) discrete random vector} \\ \int y \cdot f_{(Y|X)}\ dy \text{ if (X,Y) continuous random vector} \end{cases}$$

where $f_{(Y|X)} = \frac{f_{(X,Y)}(x,y)}{f_X(x)}$. Note that, for fixed x, $\mathbb{E}(Y|X = x)$ is a number.

Example

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right)$$

where $\rho = \frac{Cov(X,Y)}{\sqrt{\sigma_X^2\sigma_Y^2}} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$. Recall that $\mathbb{E}(Y|X = x_0) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x_0 - \mu_X)$, which is a line as a function of x_0. This can be rewritten as

$$\frac{\mathbb{E}(Y|X = x_0) - \mu_Y}{\sigma_Y} = \rho\frac{x_0 - \mu_X}{\sigma_X} \tag{26}$$

Galton and Pearson considered X as the height of fathers, Y as the height of sons and fitted a bivariate normal to (X, Y). They noticed that, given a father x_0 tall, his son was predicted by $\mathbb{E}(Y|X = x_0)$. Since this prediction was satisfying equation 26, they called this a "regression effect", meaning regression towards the mean, since usually $|\rho| < 1$. Now, $\mathbb{E}(Y|X = x_0)$ is a number, but since X is a random variable itself, we can also consider $\mathbb{E}(Y|X)$ as a function of X, i.e., a random variable called "conditional expectation of Y given X". Thus, we have that

$$\mathbb{E}(Y|X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)$$
$$\mathbb{E}(X|Y) = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y)$$

We now study some properties of conditional expectations (which are random variables). In the normal case we have:

$$\begin{aligned} \mathbb{E}(\mathbb{E}(Y|X)) &= \mathbb{E}(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)) \\ &= \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(\underbrace{\mathbb{E}(X) - \mu_X}_{\text{same values}}) \\ &= \mu_Y = \mathbb{E}(Y)\ \square \end{aligned}$$

While in the continuous case:

$$\begin{aligned} \mathbb{E}(\mathbb{E}(Y|X)) &= \mathbb{E}(\int_Y y \cdot f_{Y|X}(y|x)\ dy) \\ &= \int_X \int_Y y \cdot f_{Y|X}(y|x)\ dy\ f_X(x)\ dx \\ &= \int_Y y \int_X f_{X,Y}(x, y)\ dx\ dy \\ &= \int_Y y \cdot f_Y(y)\ dy = \mathbb{E}(Y)\ \square \end{aligned}$$

Other properties:

- $\mathbb{E}(a|X) = a$ if a is a constant
- $\mathbb{E}(aX + bY|Z) = a\mathbb{E}(X|Z) + b\mathbb{E}(Y|Z)$ (linearity of the expectation)
- $\mathbb{E}(XY|X) = X\mathbb{E}(Y|X)$ or more generally $\mathbb{E}(g(x) \cdot Y|X) = g(x) \cdot \mathbb{E}(Y|X)$

Finally, we also need a concept of conditional variance:

$$Var(Y|X) = \mathbb{E}((Y - \mathbb{E}(Y|X))^2|X)$$

The last, is a random variable itself. Let's compute its expectation.

$$\begin{aligned}
\mathbb{E}(Var(Y|X)) &= \mathbb{E}(\mathbb{E}[(Y-\mathbb{E}(Y|X))^2|X)] \\
&= \mathbb{E}(\mathbb{E}[Y^2+(\mathbb{E}(Y|X))^2-2Y\mathbb{E}(Y|X)|X]) \\
&= \mathbb{E}(\mathbb{E}(Y^2|X)+\mathbb{E}(\mathbb{E}(Y|X)^2|X)-2\mathbb{E}(Y\mathbb{E}(Y|X)|X)) \\
&= \mathbb{E}(Y^2)+\mathbb{E}(\mathbb{E}(Y|X)^2)-2\mathbb{E}(\mathbb{E}(Y|X)^2) \\
&= \mathbb{E}(Y^2)-\mathbb{E}(\mathbb{E}(Y|X)^2)+\mathbb{E}(Y)^2-\mathbb{E}(Y)^2 \\
&= \underbrace{\mathbb{E}(Y^2)-\mathbb{E}(Y^2)}_{Var(Y)}+(\mathbb{E}(Y)^2-\mathbb{E}(\mathbb{E}(Y|X)^2)) \\
&= Var(Y)-(\mathbb{E}(\mathbb{E}(Y|X)^2)-\mathbb{E}(Y)^2) \\
&= Var(Y)-Var(\mathbb{E}(Y|X))
\end{aligned}$$

So finally we have that

$$Var(Y) = \mathbb{E}(Var(Y|X)) + Var(\mathbb{E}(Y|X))$$

In the normal example it becomes

$$\begin{aligned}
Var(Y) &= \mathbb{E}(\sigma_Y^2(1-\rho^2)) + Var(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)) \\
&= \sigma_Y^2(1-\rho^2) + \rho^2\frac{\sigma_Y^2}{\sigma_X^2}\sigma_X^2 \\
&= \sigma_Y^2 - \rho^2\cdot\sigma_Y^2 + \rho^2\cdot\sigma_Y^2 \\
&= \sigma_Y^2 = Var(Y)\ \square
\end{aligned}$$

6.5 Bayesian statistics

We have reviewed Bayes theorem in elementary form (for events).

$$P(H_j|E) = \frac{P(H_j)\cdot P(E|H_j)}{\sum_{i\in[n]} P(H_i)\cdot P(E|H_i)}\ \forall j\in[n]$$

where $H_1,\dots H_n$ are the partitions of the sample space Ω. While for random vectors we have

$$f_{X|Y}(x|y) = \frac{f_X(x)\cdot f_{Y|X}(y|x)}{\int f_X(t)\cdot f_{Y|X}(y|t)\ dt}$$

and we studied elementary (binary or discrete) bayesian networks). Now we rethink our approach to statistics in a Bayesian way. So far, we have seen frequentist (or classical) statistics: the model for data is

$$f(\underbrace{x_1,\dots,x_n}_{data};\theta)$$

where θ is the unknown parameter. This last was considered a fixed constant, we can try to:

1. estimate it with a point estimate
2. estimate it with an interval (or region)
3. test hypotesis about it

Example 1

$$X_1,\dots,X_n \sim i.i.d.\ Bernoulli(\theta)$$

MLE of $\theta = \hat{\theta} = \frac{\sum X_i}{n}$ = relative frequency. We have seen confidence intervals and tests about θ.

Example 2

$$X_1,\dots,X_n \sim i.i.d.\ Normal(\mu,\sigma^2)$$

then $\theta = (\mu,\sigma^2)^T$. We estimate the parameter with

$$\hat{\theta} = \begin{pmatrix}\overline{X}\\ \frac{n-1}{n}S^2\end{pmatrix}\ \text{where}\ \frac{n-1}{n}S^2 = \frac{n-1}{n}\cdot\frac{\sum_i(x_i-\overline{X})^2}{n-1} = \frac{\sum_i(x_i-\overline{X})^2}{n}$$

Example 3

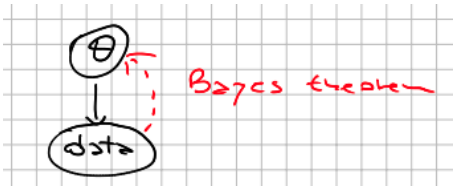
$$Y = (Y_1, \ldots, Y_n) \text{ independents with } Y \sim \mathcal{N}_n(X\beta,\ \sigma^2 I_n)$$

where X contains known numbers x_{ij}. We have $\theta = (\beta,\ \sigma^2)^T$, so this is the linear model. We have seen confidence intervals for β, for some linear combinations of β and several tests.

In Bayesian statistics, θ is considered a random variable (random vector, random element) itself. We do not observe it, but we can make inferences about it based on data. In particular, before seing the data, θ has a distribution called *prior*, expressing our information about it (it may be a subjective component). After seeing the data, we update (in a Bayesian sense) our prior distribution about θ into a posterior distribution using Bayes theorem. If θ is a vector with a prior density $\pi(\theta)$, then applying Bayes theorem:

$$\underbrace{\pi(\theta|data)}_{\text{posterior}} = \frac{\pi(\theta) \cdot f(data|\theta)}{\int \pi(t) \cdot f(data|t)\ dt} \propto \underbrace{\pi(\theta)}_{\text{prior}} \cdot \underbrace{f(data|\theta)}_{\text{likelihood}}$$

where $f(data|\theta)$ is the density of the data conditional on θ. Note that the joint density of θ and the data is given by $\pi(\theta) \cdot f(data|\theta)$. From the point of view of Bayesian networks, we have

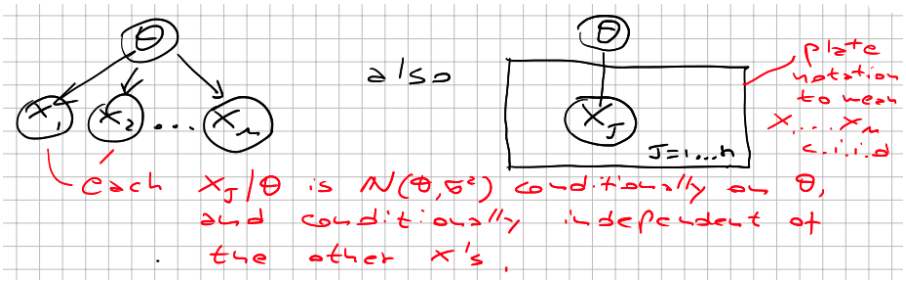


where θ is a particular node of interest. Of course, we will have to extend our Bayesian networks beyond discrete (binary) nodes. Data will be observations (realizations, instances) of random variables $X_1, \ldots, X_n, Y_1, \ldots, Y_n$.

Back to example 2

$$X_1, \ldots, X_n|\theta \sim \mathcal{N}_n(\theta,\ \sigma^2)$$

Since now θ is a random variable, we better say c.i.i.d. (conditional i.i.d. on θ). From a Bayesian Network point of view, this "single gaussian random sample" standard case is



What about the prior on θ? What is an appropriate distribution on it? If we assume θ is also normal, we have computational advantages since the computation of the posterior can be done by paper and pencil.

$$\theta \sim \pi(\theta) \sim \mathcal{N}(\mu_0,\ \sigma_0^2)$$

where μ_0, σ_0^2 are known numbers expressing our prior knowledge about θ.

$$f(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \cdot \exp\{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\}$$

$$\begin{aligned} f(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\{-\frac{1}{2\sigma^2}(x_i-\theta)^2\} \\ &= (2\pi\sigma^2)^{-n/2} \cdot \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\} \\ &= (2\pi\sigma^2)^{-n/2} \cdot \exp\{-\frac{1}{2\sigma^2}(\sum_{i=1}^{n} x_i^2 - 2\cdot\theta\sum_{i=1}^{n} x_i + n\cdot\theta^2)\} \\ &= (2\pi\sigma^2)^{-n/2} \cdot \exp\{-\frac{1}{2\sigma^2}(\sum_{i=1}^{n} x_i^2 - 2n\theta\overline{x} + n\theta^2)\} \end{aligned}$$

Therefore, applying Bayes theorem, we have

$$
\begin{aligned}
f(\theta|x_1,\dots,x_n) &= \frac{f(\theta)\cdot f(x_1,\dots,x_n|\theta)}{\int f_X(t)f(t|\theta)\ dt} \\
&= \frac{(2\pi\sigma_0^2)^{-1/2}\cdot\exp\{-(2\sigma_0^2)^{-1}\cdot(\theta-\theta_0)^2\}\cdot(2\pi\sigma^2)^{-n/2}\cdot\exp\{-(2\sigma^2)^{-1}(\sum_{i=1}^n x_i^2 - 2n\theta\overline{x}+n\theta^2)\}}{\int(2\pi\sigma_0^2)^{-1/2}\cdot\exp\{-(2\sigma_0^2)^{-1}(t-\theta_0)^2\}\cdot(2\pi\sigma^2)^{-n/2}\cdot\exp-(2\sigma^2)^{-1}(\sum_{i=1}^n x_i - 2t\overline{x}+n\theta^2)\}\ dt} \\
&\propto \exp\{-\frac{1}{2\sigma^2}(\theta^2 - 2\theta_0\theta + \cancel{\theta_0^2})\}\cdot\exp\{-\frac{1}{2\sigma^2}(\sum_{i=1}^n x_i^2 - 2n\theta\overline{x}+n\theta^2)\}
\end{aligned}
$$

Now we set $\frac{1}{\sigma^2}=\tau$, the precision, and $\frac{1}{\sigma_0^2}=\tau_0$, the prior precision.

$$
\begin{aligned}
&\propto\ \exp\{\frac{1}{2}(\tau_0\theta^2 - 2\tau_0\theta_0\theta + n\tau\theta^2 - 2n\tau\overline{x}\theta)\} \\
&= \exp\{\frac{1}{2}(\theta^2(\tau_0+n\tau) - 2\theta(\tau_0\theta_0+n\tau\overline{x}))\} \\
&= \exp\{-\frac{1}{2}(\tau_0+n\tau)\cdot(\theta^2 - 2\theta\frac{\tau_0\theta_0+n\tau\overline{x}}{\tau_0+n\tau}) \pm (\frac{\tau_0\theta_0+n\tau\overline{x}}{\tau_0+n\tau})\} \\
&= \exp\{-\frac{1}{2}(\tau_0+n\tau)\cdot(\theta - \frac{\tau_0\theta_0+n\tau\overline{x}}{\tau_0+n\tau})^2 - (\frac{\tau_0\theta_0+n\tau\overline{x}}{\tau_0+n\tau})^2\cdot(-\frac{1}{2}(\tau_0+n\tau))\} \\
&\propto \exp\{-\frac{1}{2}(\tau_0+n\tau)\cdot(\theta - \underbrace{\frac{\tau_0\theta_0+n\tau\overline{x}}{\tau_0+n\tau})^2}_{\text{kernel of a normal density}}\ \}
\end{aligned}
$$

$$
\propto f(\theta|x_1,\dots,x_n)
$$

Finally, we have the posterior density (che fatica)

$$
\theta|x_1\dots,x_n \sim \mathcal{N}\left(\frac{\tau_0\theta_0+n\tau\overline{x}}{\tau_0+n\tau},\ \frac{1}{\tau_0+n\tau}\right)
$$

Remarks

1. $E(\theta|x_1,\dots,x_n)$ is the posterior mean of θ given $x_1,\dots,x_n$. A Bayesian estimator is

$$
\frac{\tau_0\theta_0+n\tau\overline{x}}{\tau_0+n\tau} = \underbrace{\frac{\tau_0}{\tau_0+n\tau}\cdot\theta_0 + (1-\frac{\tau_0}{\tau_0+n\tau})\overline{x}}_{\text{weighted average of prior guess and MLE}}
$$

   where:

   - $\frac{\tau_0}{\tau_0+n\tau}$ is the prior weight
   - θ_0 is the prior guess, our estimate without seeing any data
   - $\overline{x}$ is the maximum likelihood estimate (MLE) of θ, the bast we can do about estimating θ without prior knowledge

2. As $n\to+\infty$, prior weight goes to 0: data speak more and more as they accumulate.

3. The prior weight depends on τ_0 (the prior precision) and $n\tau$ (the data precision). Notice that the posterior precision is $\tau_0+n\tau$, which is the prior precision plus the data precision. Therefore, precision is additive in the presence of random sampling (i.e., i.i.d. sampling from a frequentist point of view, of c.i.i.d. sampling from a Bayesian point of view).

4. More about prior precision: the higher is τ_0, the more informative is our prior distribution. Conversely, the lower is τ_0, the less informative is our prior distribution. In the limit, $\tau_0\to 0$, the prior becomes degenerate, i.e., $\mathcal{N}(\theta_0,\infty)$ does not exist. Nonetheless, the posterior distribution becomes

$$
\theta|x_1,\dots,x_n \sim \mathcal{N}(\overline{x},\ \frac{1}{n\tau})
$$

   This happens to the normal and to some other cases, just because normal is well behaved. $\tau_0=0$ is often used as a totally non informative prior distribution (prior ignorance), also called flat prior or uniform prior. The problems with flat priors are the following.

(a) They are sometimes not proper distributions, like in the normal case.

(b) They are not invariant under reparametrization; if instead of θ, you use $\varphi(\theta)$ as parameter, a flat prior on θ does not necessarily induce a flat prior on $\varphi(\theta)$.

5. Comments 1-3 apply not only to the mean, byt to the whole prior distribution.

6. The posterior on θ is normal, like the prior was. This is called conjugacy: if we choose a prior on a certain conjugate class, the posterior will belong to the same class.

Another prototype: binary data

$$X_1, \ldots, X_n \sim \text{ c.i.i.d. } Bernoulli(\theta)$$

$$f_{X_i|\theta}(x|\theta) = \theta^x \cdot (1-\theta)^{1-x} \; x \in \{0,1\}$$

The likelihood is:

$$\prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i} \cdot (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} \cdot (1-\theta)^{n-\sum_{i=1}^n x_i} \; x_i \in \{0,1\}$$

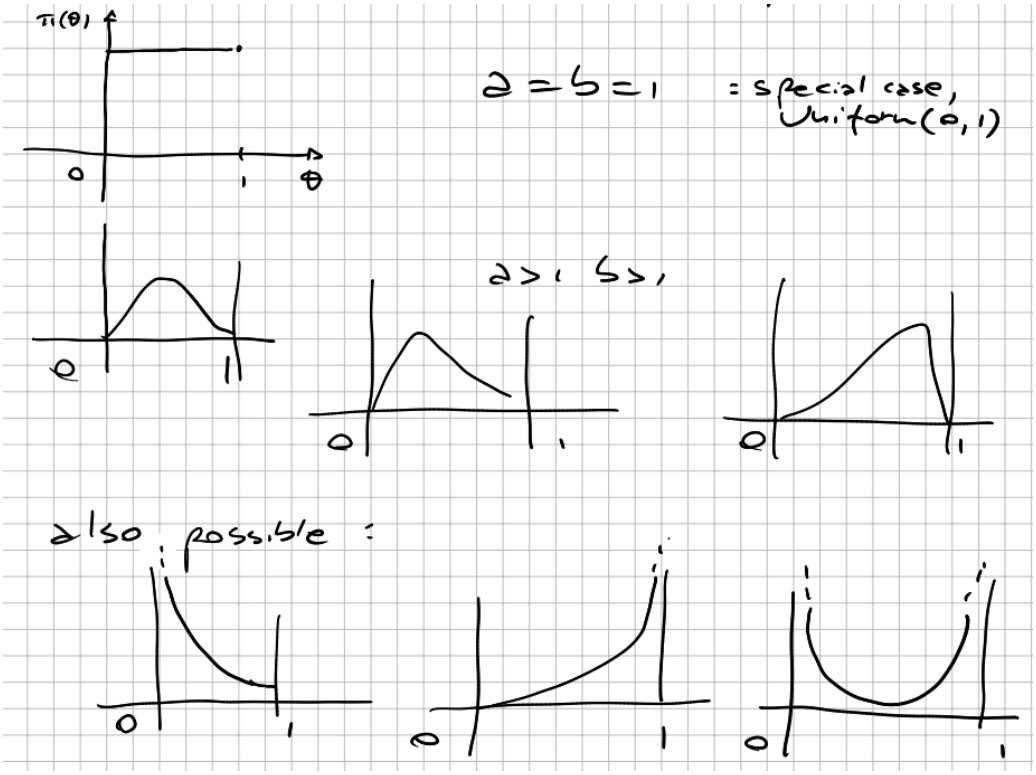
The conjugate prior follows a Beta distribution.
Recall: θ is said to have a Beta distribution with parameters a and b $\theta \sim Beta(a,b)$ if

$$\pi(\theta) = \frac{\theta^{a-1} \cdot (1-\theta)^{b-1}}{B(a,b)}$$

where $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\, dt$ is the appropriate normalization constant, a special function invented by Euler and called Euler's Beta function. A property of this function is the following.

$$B(a,b) = \frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)}$$

where $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}\, dx$ is another special function called Gamma. Beta densities can have different shapes:



Another property is

$$E(\theta) = \frac{a}{a+b} \text{ and } Var(\theta) = \frac{a \cdot b}{(a+b)^2 \cdot (a+b+1)}$$

Let's use $\pi(\theta)$ of the beta kind in the binary example.

$$\begin{aligned}
\text{posterior} &\propto \text{prior} \times \text{likelihood} \\
\text{posterior} &\propto \theta^{a-1} \cdot (1-\theta)^{b-1} \cdot \theta^{\sum_{i=1}^n x_i} \cdot (1-\theta)^{n-\sum_{i=1}^n x_i} \\
&= \underbrace{\theta^{a+\sum_{i=1}^n x_i - 1} \cdot (1-\theta)^{b+n-\sum_{i=1}^n x_i - 1}}_{\text{kernel of a Beta}}
\end{aligned}$$

Therefore we recognize that this is a beta distribution as well (another example of proportionally calculation as in the normal example). The posterior distribution is the following.

$$\theta|x_1,\ldots,x_n \sim Beta\left(a+\sum_{i=1}^n x_i,\ b+n-\sum_{i=1}^n x_i\right)$$

$$\begin{aligned}
E(\theta|x_1,\ldots,x_n) &= \frac{a+\sum_{i=1}^n x_i}{a+\sum_{i=1}^n + b + n - \sum_{i=1}^n x_i} \\
&= \frac{a+\sum_{i=1}^n x_i}{a+b+n} \\
&= \underbrace{\frac{a+b}{a+b+n}}_{\text{prior weight}} \cdot \underbrace{\frac{a}{a+b}}_{\text{prior mean}} + (1-\frac{a+b}{a+b+n}) \cdot \underbrace{\frac{\sum_{i=1}^n x_i}{n}}_{\text{MLE of } \theta}
\end{aligned}$$

One final remark.

- In example 1 (section 6.5), $\pi(\theta|x_1,\ldots,x_n)$ depends on the data only through $\overline{x}$.
- In example 2 (section 6.5), $\pi(\theta|x_1,\ldots,x_n)$ depends on the data only through $\sum_{i=1}^n x_i$.
- $\overline{x}$ and $\sum_{i=1}^n x_i$ respectively are said to be *sufficient statistics*.

6.6 Exercises on Bayesian Networks

6.6.1 Exercise 1

Text The two chefs Mark and Louise may cook, alone or together, a certain cake; Mark cooks the cake with probability 0.5, and Louise does it with probability 0.6, independently of one another. When he cooks alone, Mark uses the egg yolks with probability 0.4, while if he cooks with Louise they use them with probability 0.3; when she cooks alone, Louise uses them with probability 0.1, while if neither of them cooks the cake, egg yolks are anyway used for the cake with probability 0.2. Similar probabilities for vanilla extract are: Mark alone 0.2, Louise alone 0.4, together 0.3, neither of them 0.5.

Formalization We have the following binary random variables involved in the problem.

$$M? = \begin{cases} 1 \text{ if Mark cooks} \\ 0 \text{ otherwise} \end{cases} \qquad L? = \begin{cases} 1 \text{ if Louise cooks} \\ 0 \text{ otherwise} \end{cases}$$

$$E? = \begin{cases} 1 \text{ if egg yolks are used} \\ 0 \text{ otherwise} \end{cases} \qquad V? = \begin{cases} 1 \text{ if vanilla is used} \\ 0 \text{ otherwise} \end{cases}$$

From the text, we can infer the following probabilities

- $P(M=1) = P(M=0) = 0.5$
- $P(L=1) = 0.6$ and $P(L=0) = 0.4$

$M?$	L?	$P(E=1\|M,L)$	$P(E=0\|M,L)$	$P(V=1\|M,L)$	$P(V=0\|M,L)$
0	0	0.2	0.8	0.5	0.5
0	1	0.1	0.9	0.4	0.6
1	0	0.4	0.6	0.2	0.8
1	1	0.3	0.7	0.3	0.7

Table 11: Conditional probabilities inferred from the text

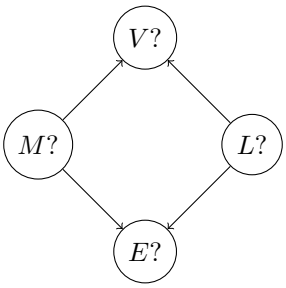
Joint density Now we compute the joint density of the random vector $(M?, L?, E?, V?)^T$. To do so, we apply the formula

$$\pi_{(M?,L?,E?,V?)}(m,l,e,v) = \prod_{I \in \{M?,L?,E?,V?\}} \pi_{(I|Parents(I))}(i|parents(i))$$

$M?$	$L?$	$E?$	$V?$	$\pi(m,l,e,v)$
0	0	0	0	$0.5 \times 0.4 \times 0.8 \times 0.5 = 0.08$
0	0	0	1	$0.5 \times 0.4 \times 0.8 \times 0.5 = 0.08$
0	0	1	0	$0.5 \times 0.4 \times 0.2 \times 0.5 = 0.02$
0	0	1	1	$0.5 \times 0.4 \times 0.2 \times 0.5 = 0.02$
0	1	0	0	$0.5 \times 0.6 \times 0.9 \times 0.6 = 0.162$
0	1	0	1	$0.5 \times 0.6 \times 0.9 \times 0.4 = 0.108$
0	1	1	0	$0.5 \times 0.6 \times 0.1 \times 0.6 = 0.018$
0	1	1	1	$0.5 \times 0.6 \times 0.1 \times 0.4 = 0.012$
1	0	0	0	$0.5 \times 0.4 \times 0.6 \times 0.8 = 0.096$
1	0	0	1	$0.5 \times 0.4 \times 0.6 \times 0.2 = 0.024$
1	0	1	0	$0.5 \times 0.4 \times 0.4 \times 0.8 = 0.064$
1	0	1	1	$0.5 \times 0.4 \times 0.4 \times 0.2 = 0.016$
1	1	0	0	$0.5 \times 0.6 \times 0.7 \times 0.7 = 0.147$
1	1	0	1	$0.5 \times 0.6 \times 0.7 \times 0.3 = 0.063$
1	1	1	0	$0.5 \times 0.6 \times 0.3 \times 0.7 = 0.063$
1	1	1	1	$0.5 \times 0.6 \times 0.3 \times 0.3 = 0.027$

Table 12: Joint density of the random vector

Point A Draw the appropriate DAG for the problem.



Point B Observe vanilla has been used but not egg yolks. Compute the probability that Mark has cooked the cake. We have that:

- $P(E = 0 \cap V = 1|M = 1) = \frac{P(E=0 \cap V=1 \cap M=1)}{P(M=1)} = \frac{0.024+0.063}{0.5} = 0.174$
- $P(E = 0 \cap V = 1|M = 0) = \frac{P(E=0 \cap V=1 \cap M=0)}{P(M=0)} = \frac{0.08+0.108}{0.5} = 0.376$

Thus the required probability is the following.

$$\begin{aligned} P(M = 1|E = 0 \cap V = 1) &= \frac{P(M = 1) \cdot P(E = 0 \cap V = 1|M = 1)}{P(M = 1) \cdot P(E = 0 \cap V = 1|M = 1) + P(M = 0) \cdot P(E = 0 \cap V = 1|M = 0)} \\ &= \frac{0.5 \times 0.174}{0.5 \times 0.174 + 0.5 \times 0.376} = \frac{0.087}{0.275} = 0.3163636 \end{aligned}$$

Point C Observe that vanilla has been used but not egg yolks. Compute the probability that Mark has cooked the cake alone. We have that:

- $P(M = 1 \cap L = 0) = 0.5 \times 0.4 = 0.2$
- $P(M = 0 \cup L = 1) = P(M = 0) + P(L = 1) - P(M = 0 \cap L = 1) = 0.5 + 0.6 - 0.5 \times 0.6 = 0.8$
- $P(E = 0 \cap V = 1|M = 1 \cap L = 0) = \frac{P(E=0,V=1,M=1,L=0)}{P(M=1 \cap L=0)} = \frac{0.024}{0.2} = 0.12$
- $P(E = 0 \cap V = 1|M = 0 \cup L = 1) = \frac{P(E=0 \cap V=1 \cap M=0 \cup L=1)}{P(M=0 \cup L=1)} = \frac{0.08+0.108+0.063}{0.8} = 0.31375$

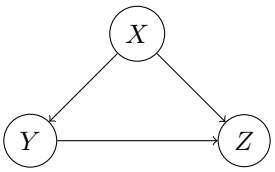
$$P(M = 1 \cap L = 0 | E = 0 \cap V = 1) =$$
$$= \frac{P(E = 0 \cap V = 1 | M = 1 \cap L = 0) \cdot P(M = 1 \cap L = 0)}{P(E = 0 \cap V = 1 | M = 1 \cap L = 0) \cdot P(M = 1 \cap L = 0) + P(M = 0 \cup L = 1) \cdot P(E = 0 \cap V = 1 | M = 0 \cup L = 1)}$$
$$= \frac{0.12 \times 0.2}{0.12 \times 0.2 + 0.8 \times 0.31375} = \frac{0.024}{0.275} = 0.0872727273$$

6.6.2 Exercise 2

Text A normal vector $(X, Y, Z)^T$ is given such that:

- $X \sim \mathcal{N}(0, 1)$;
- $Y|X = x \sim \mathcal{N}(x, 1)$;
- $Z|X = x \cap Y = y \sim \mathcal{N}(x + y, 1)$ [i.e., $Z|X = x, Y = y \sim \mathcal{N}(x + y, 1)$]

Formalization The DAG is the following.



From the text, we infer the following probability mass functions.

$$\pi_X(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\{-\frac{1}{2}x^2\}$$
$$\pi_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\{-\frac{1}{2}(y - x)^2\}$$
$$\pi_{Z|X,Y}(z|x, y) = \frac{1}{\sqrt{2\pi}} \cdot \exp\{-\frac{1}{2}(z - x - y)^2\}$$

Point A Compute the joint density of $(X, Y, Z)^T$.

$$\pi_{X,Y,Z}(x, y, z) = (2\pi)^{-3/2} \cdot \exp\{-\frac{1}{2}x^2 - \frac{1}{2} \cdot (x^2 + y^2 - 2xy) - \frac{1}{2} \cdot (x^2 + y^2 + z^2 - 2xz - 2yz + 2xy)\}$$
$$= (2\pi)^{-3/2} \cdot \exp\{-\frac{1}{2}y^2 + xy - \frac{1}{2}x^2 - \frac{1}{2}y^2 - \frac{1}{2}z^2 + xz - yz + xy\}$$
$$= (2\pi)^{-3/2} \cdot \exp\{-y^2 + 2xy - \frac{1}{2}x^2 - \frac{1}{2}z^2 + xz - yz\}$$

Point B Use the properties of conditional expectations to find the mean vector and the variance-covariance matrix of $(X, Y, Z)^T$. We have that:

- $E(X) = 0$
- $E(Y) = E(E(Y|X)) = E(X) = 0$
- $E(Z) = E(E(Z|X, Y)) = E(X + Y) = E(X) + E(Y) = 0$
- $Var(X) = 1$
- $Var(Y) = E(Var(Y|X)) + Var(E(Y|X)) = E(1) + Var(X) = 1 + 1 = 2$
- $Var(Z) = E(Var(Z|X, Y)) + Var(E(Z|X, Y)) = E(1) + Var(X + Y) = 1 + Var(X) + Var(Y) + 2 \cdot Cov(X, Y) = 1 + 1 + 2 + 2 = 6$
- $Cov(X, Y) = E(XY) - E(X) \cdot E(Y) = E(XY) = E(E(XY|X)) = E(X \cdot E(Y|X)) = E(XX) = E(X^2) = 1$
- $Cov(X, Z) = E(XZ) - E(X) \cdot E(Z) = E(XZ) = E(E(XZ|X, Y)) = E(X \cdot E(Z|X, Y)) = E(X \cdot (X + Y)) = E(X^2 + XY) = E(X^2) + E(XY) = E(X^2) + Cov(X, Y) = 1 + 1 = 2$
- $Cov(Y, Z) = E(YZ) - E(Y) \cdot E(Z) = E(YZ) = E(E(YZ|X, Y)) = E(Y \cdot E(Z|X, Y)) = E(Y \cdot (X + Y)) = E(XY) + E(Y^2) = 1 + 2 = 3$

Note that $Var(X) = E(X^2) - E(X)^2 = E(X^2) = 1$ (this is how we concluded the computation of $Cov(X, Y)$) and $Var(Y) = E(Y^2) - E(Y)^2 = E(Y^2) = 2$. Furthermore, when a variable is given, you can threat it as a constant and bring it out the expected value (e.g., $E(XZ|X, Y) = x \cdot E(Z|X, Y)$). Then the random vector has the following distribution.

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}\Big(\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},\ \Sigma = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 6 \end{pmatrix}\Big)$$

6.7 Bayesian estimation

Once we have computed (or approximated) a posterior distribution, expressing our updated knowledge about something uncertain and interesting, what do we do with it? We can reinterpret concepts and methods from classical statistics using Bayes terminology.

6.7.1 Point estimation

A Bayesian estimate is a summary of the posterior distribution.

1. A posterior mean $E(\theta|data)$.
2. A posterior mode (MAP = Maximum A Posteriori).

$$MAP = \arg\max \underbrace{\pi(\theta|data)}_{\text{posterior density}} = \arg\max \underbrace{\pi(\theta)}_{\text{prior density}} \cdot \underbrace{\mathcal{L}(\theta|data)}_{\text{likelihood}}$$

   So, a MAP is similar to a Maximum Likelihood Estimate, except there is an extra $\pi(\theta)$ term, which plays a role similar to penalizing terms in Ridge or Lasso (penalized likelihood approaches).

3. Other choices minimizing some specific loss functions.

Example: binary random sample + conjugate prior

$$\theta \sim Beta(a, b)$$

$$X_1, \ldots, X_n|\theta \sim \text{ c.i.i.d. } Bernoulli(\theta)$$

We showed that

$$\theta|x_1, \ldots, x_n \sim Beta(a + S, b + n - S) \text{ where } S = \sum_{i=1}^{n} x_i$$

$$E(\theta|x_1, \ldots, x_n) = \frac{a + S}{a + b + n}$$

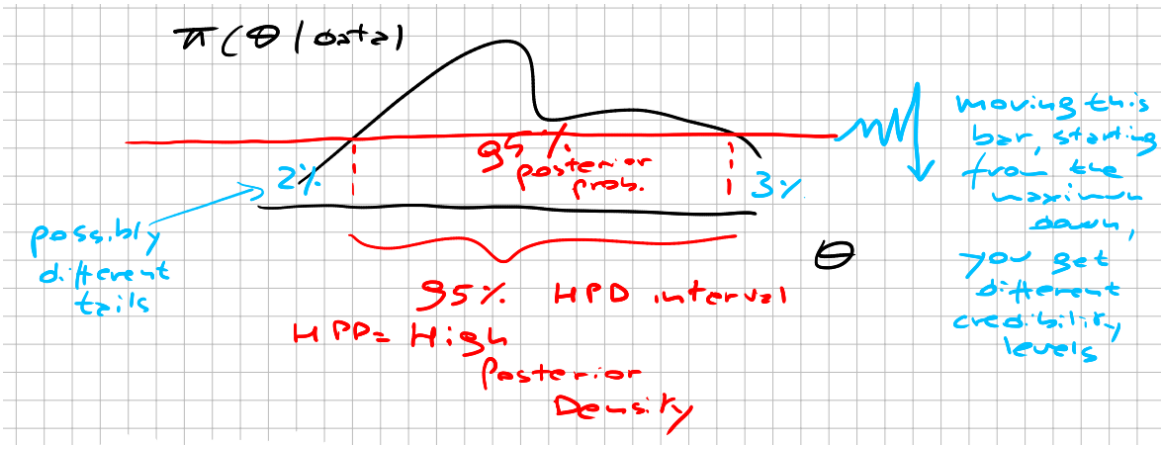
Notice: suppose $a = b = 1$ (uniform distribution), and suppose you see all success, i.e., $S = n$. Then

$$E(\theta|x_1, \ldots, x_n) = \frac{1 + n}{1 + 1 + n} = \frac{n + 1}{n + 2}$$

called Laplace succession rule.
Example: Suppose men have seen the sun raising 50000 times. Then, the Bayes estimate of the probability that the sun will raise tomorrow is $\frac{50001}{50002}$.

6.7.2 Interval (region) estimation

Suppose θ is one dimensional and continuous.



No symmetry is guaranteed. Some people prefer Equally Tailed Intervals (ETI): cut $\alpha/2$ from both the left and the right tails of the posterior distribution. They may be easier to compute. But, conceptually, HPD regions are similarly defined when θ is more-than-1 dimensional, while EQI are not generalizable to higher dimensions.

Example: binary random sample + conjugate prior

$$\theta|x_1,\dots,x_n \sim Beta(a+S, b+n-S)$$

We have that

$$\text{HPD intervals: solve in } (\theta_L,\theta_M) \begin{cases} \pi(\theta_L|x_1,\dots,x_n) = \pi(\theta_H|x_1,\dots,x_n) \\ \int_{\theta_L}^{\theta^H} \pi(t|x_1,\dots,x_n)\ dt = 0.95 \end{cases}$$

ETI: `qbeta(0.25, a+S, b+n-S), qbeta(0.975, a+S, b+n-S)` [R syntax]

6.7.3 Hypotesis testing

In the standard setup we have:

$$H_0 : \theta \in \Theta_0$$

where H_0 is the null hypothesis, θ is an unknown parameter of interest, Θ_0 is a specific subset of the parameters space. An alternative hypothesis is $\overline{H}_0 : \theta \notin \Theta_0$. From a Bayesian point of view, usually the important quantity is the posterior probability of the hypothesis:

$$P(\theta \in \Theta_0|data)$$

Example: binary random sample + conjugate prior

$$\begin{aligned} &\theta \sim Beta(a,b) \\ &X_1,\dots,X_n|\theta \sim \text{ c.i.i.d. } Bernoulli(\theta) \\ &\implies \theta|x_1,\dots,x_n \sim Beta(a+S, b+n-S)\ S = \sum_{i=1}^{n} x_i \end{aligned}$$

Suppose $H_0 : \theta > 1/2$ (so that $\Theta_0 = (\frac{1}{2}, 1]$) and you choose Laplace's uniform prior $\pi(\theta) = 1\ \theta \in [0,1]$. Then

$$P(H_0) = \frac{1}{2} \text{ (prior probability of } H_0)$$

$$P(H_0|data) = \frac{\int_{1/2}^{1} t^{1+s-1} \cdot (1-t)^{1+n-s-1}\ dt}{\underbrace{\int_0^1 v^s \cdot (1-v)^{n-s}\ dv}_{B(s,n-s) \text{ Euler's Beta function}}}$$

There may be problems with point hypothesis like $H_0 : \theta = 1/2$ since continuous priors on θ always give problems. In fact

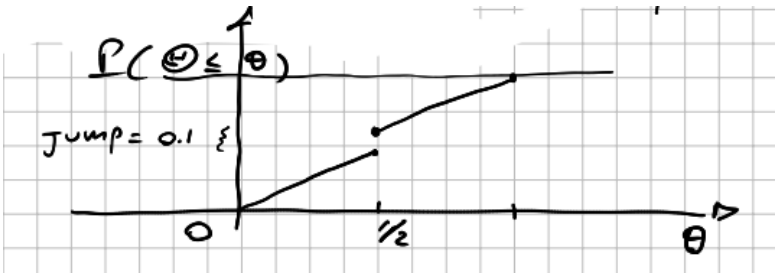
$$P(H_0) = \int_{1/2}^{1/2} 1\ dt = 0$$

With zero prior probabilities, posterior will also be zero.

$$P(H_0|data) \propto P(H_0) \cdot P(data|H_0) = 0$$

So we have few choices.

1. Give a positive mass to Θ_0 (but in the binary case you cannot use a Beta anymore, you may use $P(\theta = 1/2) = 0.1$) and mix it with a beta, to obtain



2. Relax the point null to an interval

$$H_0 : \theta \in [\frac{1}{2} - \varepsilon,\ \frac{1}{2} + \varepsilon]$$

   for an appropriate ε which is of practical relevance.

3. Use Bayes factor as follows.

Instead of looking at $P(H_0|data)$, look at the *posterior odds*.

$$\frac{P(H_0|data)}{P(\overline{H}_0|data)} = \frac{P(H_0) \cdot P(data|H_0)}{P(\overline{H}_0) \cdot P(data|\overline{H}_0)} = \text{prior odds } \times \text{ likelihood ratio}$$

Since the prior only affects the priors odds, we can separate it from the likelihood ratio and consider the likelihood ratio as a *weight of evidence*. The likelihood ratio is also called *Bayes factor* and it is an important concept when you want to quantify the sole contribution of evidence (e.g. forensic statistics). Finally, notice that from a Bayesian networks point of view, a hypothesis corresponds to a binary node which is updated when data are collected.

Example: binary random sample + conjugate prior (but reducing θ possible values to $\frac{1}{2}$ and $\frac{1}{3}$)

$$H_0 : \theta = \frac{1}{2} \left(\pi = P(\theta = \frac{1}{2})\right)$$
$$\overline{H}_0 : \theta = \frac{1}{3} \left(1 - \pi = P(\theta = \frac{1}{3})\right)$$

Bayes posterior odds:

$$\frac{P(H_0|data)}{P(\overline{H}_0|data)} = \frac{\pi}{1-\pi} \cdot \frac{\binom{n}{s} \cdot (\frac{1}{2})^s \cdot (1 - \frac{1}{2})^{n-s})}{\binom{n}{s} \cdot (\frac{1}{3})^s \cdot (1 - \frac{1}{3})^{n-s}}$$

Even using beta conjugate prior, Bayes factor can be defined

$$H_0 : \theta = \frac{1}{2}$$
$$\pi(\theta) = 1 \ \text{ conjugate uniform}$$
$$H_a : \theta \neq \frac{1}{2}$$

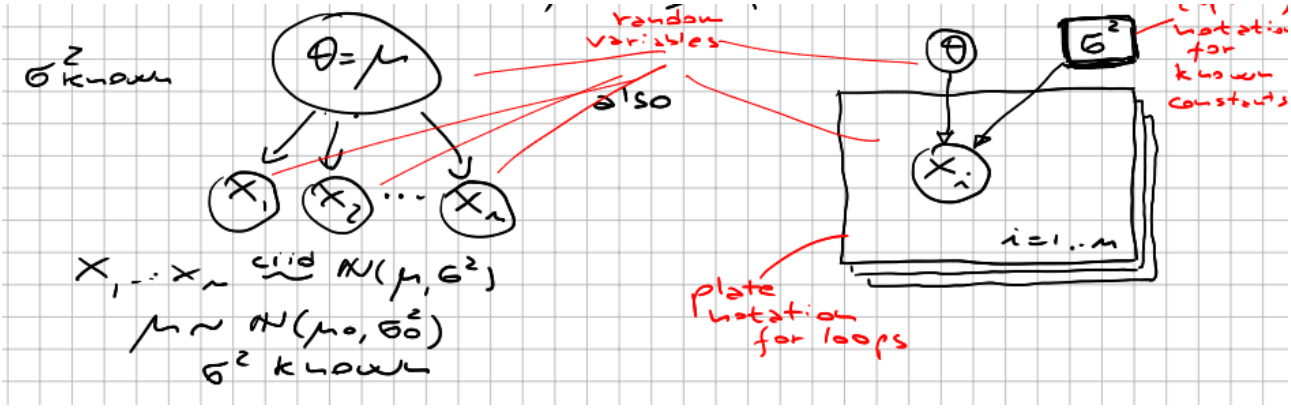
Bayes factor:

$$\frac{(1/2)^n}{\int t^s \cdot (1-t)^{n-s} \cdot \underbrace{\pi(\theta)}_{1}\ dt}$$

Althouth we can not say any more that the Bayes factor is not related to the uniform prior.

6.8 Hierarchical Bayes Models

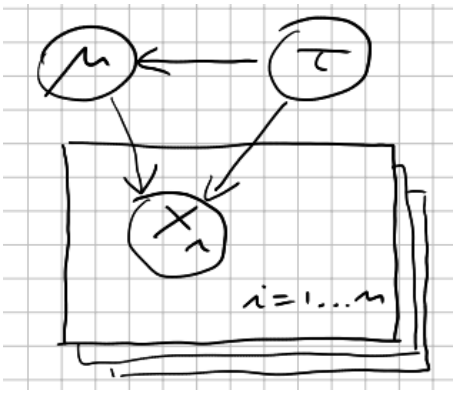
We did Bayes for a single random sample with known variance. In graphical terms:



Now we take into account the more difficult problem of considering σ^2 unknown, so that $\theta = (\mu, \sigma^2)^T$, i.e., the parameter of interest is bi-dimensional. There exists a conjugate prior as follows.

$$
\begin{aligned}
X_1, \ldots, X_n \sim \text{ c.i.i.d. } \mathcal{N}(\mu, \sigma^2) = \mathcal{N}(\mu, \frac{1}{\tau}) \\
\tau \sim Gamma(a, \lambda) \\
\frac{\mu}{\tau} \sim \mathcal{N}(\mu_0, \frac{1}{\tau \cdot \tau_0})
\end{aligned}
$$

where $\tau = 1/\sigma^2$ is the precision, $a, \lambda, \mu_0, \tau_0$ are hyperparameters to be chosen to express our level of prior uncertainty. Graphically:



The full prior on (μ, τ) is

$$
\begin{aligned}
\pi(\mu, \tau) &= \pi_\tau(\tau) \cdot \pi_{\mu|\tau}(\mu|\tau) \\
&\propto \frac{\lambda^a \cdot \tau^{a-1} \cdot e^{-\lambda \cdot \tau}}{\Gamma(\lambda)} \cdot \frac{1}{\sqrt{2 \cdot \frac{1}{\tau \cdot \tau_0}}} \cdot \exp\{-\frac{\tau \cdot \tau_0}{2} \cdot (\mu - \mu_0)^2\}
\end{aligned}
$$

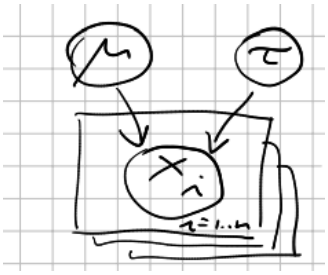
So that the posterior becomes

$$
\pi(\mu, \tau | x_1, \ldots, x_n) \propto \prod_{i=1}^{n} \underbrace{\sqrt{\tau} \exp\{-\frac{\tau \cdot \tau_0}{2} \cdot (\mu - \mu_0)^2\}}_{\text{likelihood}}
$$

and after some computations, we find out that

$$
\begin{aligned}
\tau | x_1, \ldots, x_n &\sim Gamma\left(a + \frac{n}{2}, \ \lambda + \frac{\sum_{i=1}^n (x_i - x)^2}{2} + \frac{n \cdot \tau_0 \cdot (\overline{x} - \mu_0)^2}{2 \cdot (\tau_0 + n)}\right) \\
\mu | \tau, x_1, \ldots, x_n &\sim \mathcal{N}\left(\frac{\tau_0 \cdot \mu_0 + n \cdot \overline{x}}{\tau_0 + n}, \ (\tau \cdot \tau_0 + n)^{-1}\right)
\end{aligned}
$$

So, the structure of the posterior is the same normal - inverse gamma as the prior. this is then another example of conjugacy, handy for calculations when it applies. But, what if we feel an independent prior is more appropriate? Graphically:



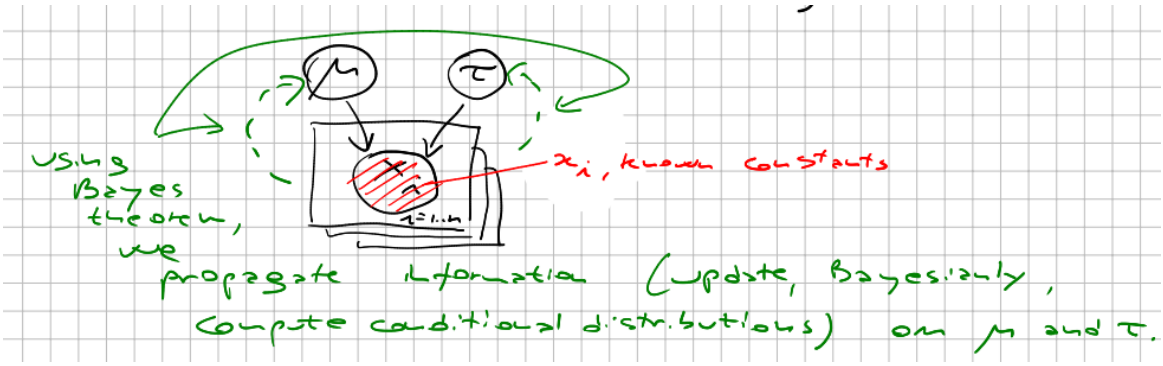
$$X_1, \ldots, X_n \sim \text{ c.i.i.d. } \mathcal{N}(\mu, \frac{1}{\tau})$$
$$\mu \sim \mathcal{N}(\mu_0, \frac{1}{\tau_0})$$
$$\tau \sim Gamma(a, \lambda)$$

Then the posterior is

$$\pi(\mu, \tau | x_1, \ldots, x_n) \propto \exp\{-\frac{\tau_0}{2}(\mu - \mu_0)^2\} \cdot \tau^{a-1} \cdot \exp\{-\frac{\tau}{2} \cdot \sum_{i=1}^{n}(x_i - \mu)^2\}$$

which is not a product of a normal and a gamma. We have a computational problem: that is our motivating example to introduce approximate methods to compute posteriors.

A possibility is Markov Chain Monte Carlo (MCMC). Suppose we observe $X_1 = x_1, \ldots X_n = x_n$ (i.e., we instantiate the X nodes to the observed values).



To do this, observe that

$$\pi(\mu | \tau, x_1, \ldots, x_n) = \mathcal{N}\left(\frac{\tau_0 \cdot \mu_0 + n \cdot \tau \cdot \overline{x}}{\tau_0 + n \cdot \tau}, \frac{1}{\tau_0 + n \cdot \tau}\right)$$

but also

$$\pi(\tau | \mu, x_1, \ldots, x_n) = \frac{\pi(\tau, \mu | x_1, \ldots, x_n)}{\pi(\mu | x_1, \ldots, x_n)}$$
$$\propto \tau^{a+\mu/2-1} \cdot \exp\{-(\lambda + \frac{1}{2} \cdot \sum_{i=1}^{n}(x_i - \mu)^2) \cdot \tau\}$$

which is exactly a Gamma density

$$\tau | \mu, x_1, \ldots, x_n \sim Gamma\left(a + \frac{\mu}{2}, \ \lambda + \frac{1}{2} \cdot \sum_{i=1}^{n}(x_i - \mu)^2\right)$$

This means that we can implement the Gibbs sampler, since all single conditional distributions of each parameter component (i.e., μ and τ) given all other parameters (and given data) are available. This motivating example sprung a lot of software activities in the 90s on computational Bayes, in particular the BUGS (Bayesian Using the Gibbs Sampling) for Unix. Today we use JAGS (Just Another Gibbs Sampler).

6.9 JAGS: The rats example

```
import numpy as np
import pyjags as jags
import matplotlib.pyplot as plt

n_rats = 30
n_measures_per_rat = 5
days = np.array([8, 15, 22, 29, 36])
measures = np.array(
[151, 145, 147, 155, 135, 159, 141, 159, 177, 134,
160, 143, 154, 171, 163, 160, 142, 156, 157, 152, 154, 139, 146,
157, 132, 160, 169, 157, 137, 153, 199, 199, 214, 200, 188, 210,
189, 201, 236, 182, 208, 188, 200, 221, 216, 207, 187, 203, 212,
203, 205, 190, 191, 211, 185, 207, 216, 205, 180, 200, 246, 249,
263, 237, 230, 252, 231, 248, 285, 220, 261, 220, 244, 270, 242,
248, 234, 243, 259, 246, 253, 225, 229, 250, 237, 257, 261, 248,
219, 244, 283, 293, 312, 272, 280, 298, 275, 297, 350, 260, 313,
273, 289, 326, 281, 288, 280, 283, 307, 286, 298, 267, 272, 285,
286, 303, 295, 289, 258, 286, 320, 354, 328, 297, 323, 331, 305,
338, 376, 296, 352, 314, 325, 358, 312, 324, 316, 317, 336, 321,
334, 302, 302, 323, 331, 345, 333, 316, 291, 324]).reshape(n_measures_per_rat, n_rats).transpose()

rats_jags_data = {
    'N': n_rats,
    'T': n_measures_per_rat,
    'Y': measures,
    'x': days
}

# alpha0 expected weight of rat at birth
# alpha.c intercept w.r.t. the average day
# beta.c slope w.r.t. the average day
# alpha[i] intercept of rat i
# beta[i] slope of rat i
# tau.c precision at avarage day
# tau.alpha precision of alpha
# tau.beta precision of beta
rats_jags_initialization = {
 'alpha': np.array([250] * n_rats),
 'beta': np.array([6] * n_rats),
 'alpha.c': 0,
 'beta.c': 2,
 'tau.c': 1,
 'tau.alpha': 1,
 'tau.beta': 1,
}

rats_jags_model_code = '''
model {
    for (i in 1:N) {
       for (j in 1:T) {
          mu[i,j] <- alpha[i] + beta[i]*(x[j] - x.bar);
          Y[i,j]   ~ dnorm(mu[i,j],tau.c)
       }
       alpha[i] ~ dnorm(alpha.c,tau.alpha);
       beta[i]  ~ dnorm(beta.c,tau.beta);
    }
    alpha.c    ~ dnorm(0,1.0E-4);
    beta.c     ~ dnorm(0,1.0E-4);
    tau.c      ~ dgamma(1.0E-3,1.0E-3);
    tau.alpha  ~ dgamma(1.0E-3,1.0E-3);
    tau.beta   ~ dgamma(1.0E-3,1.0E-3);
    sigma    <- 1.0/sqrt(tau.c);
    x.bar    <- mean(x[]);
    alpha0   <- alpha.c - beta.c*x.bar;
}
'''

model = jags.Model(
    code = rats_jags_model_code,
    data = rats_jags_data,
    init = rats_jags_initialization
)

# pyjags simulates parallely 4 chains
iterations = 10000
results = model.sample(iterations, vars=['alpha0', 'beta.c'])
```

```
def plot_hist(results):
    results = results.squeeze()
    plt.hist(results, bins=100, histtype='step')
    plt.show()

plot_hist(results['alpha0'])
plot_hist(results['beta.c'])
```