SUMMARY

- RECAP ON PROBABILITY: DISTRIBUTIONS, $E[\cdot]$, $V[\cdot]$, ecc → STATISTICAL METHODS

es.

RANDOM EXPERIMENT: "$n = 10$ RANDOM PERSON WITH REPLACEMENT AND COUNT HOW MANY BORN IN ITALY" = X

BORN	COUNT
ITALY	24
ABROAD	12
TOT	36

- $X \sim Bin\left(n = 10, p = \frac{24}{36} = \frac{2}{3}\right)$

$$\rightarrow P(X = 3) = \binom{10}{3} \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^{10-3}$$

- IN PYTHON:

```
# 10 random number in [1, 37)
np.random.choice(range(1,37), size=10)
```

es. MULTINOMIAL DISTRIBUTION

	BORN	COUNT
P	PIEDMONT	10
I	ITALY NOT PIEDMONT	14
E	EU NOT ITALY	5
A	OUT OF EU	7
	TOT	36

$X_1 = \#$ BORN IN PIEDMONT
$X_2 = \#$ BORN IN ITALY NOT PIEDMONT
$X_3 = \#$ BORN IN EU NOT ITALY
$X_4 = \#$ BORN OUT OF EU

$$\rightarrow P(X_1 = 2, X_2 = 3, X_3 = 0, X_4 = 5) = ?$$

- ... BUT FIRST, 1 POSSIBLE DISPOSITION:

$$P(\text{"SEQUENCE } P, I, I, A, A, A, I, P, A, A\text{"}) = \frac{10}{36} \cdot \frac{14}{36} \cdot \frac{14}{36} \cdot \frac{7}{36} \cdot \frac{7}{36} \cdot \frac{7}{36} \cdot \frac{14}{36} \cdot \frac{10}{36} \cdot \frac{7}{36} \cdot \frac{7}{36}$$

- BUT $\forall$ SYMBOL → MULTIPLE COMBINATIONS: 2P, 3I, 0E, 5A

$$\rightarrow P(X_1 = 2, X_2 = 3, X_3 = 0, X_4 = 5) = \binom{10}{2}\binom{8}{3}\binom{5}{0}\binom{5}{5} \cdot \left(\frac{10}{36}\right)^2 \left(\frac{14}{36}\right)^3 \left(\frac{5}{36}\right)^0 \left(\frac{7}{36}\right)^5 =$$

$$= \frac{10!}{2!\,3!\,0!\,5!} \cdot \left(\frac{10}{36}\right)^2 \left(\frac{14}{36}\right)^3 \left(\frac{5}{36}\right)^0 \left(\frac{7}{36}\right)^5$$

- **MULTINOMIAL DISTRIBUTION** : ~> GENERALIZATION OF BINOMIAL EXPERIMENT : 1 CLASS ⟶ K+1 CLASSES

GIVEN $K+1$ CLASSES AS POSSIBLE OUTCOMES, IF WE MAKE n CHOICES INDEPENDENTLY WITH THE SAME CLASS PROBABILITIES $\pi_1, \dots, \pi_{K+1}$ / $\pi_{K+1} = 1 - \sum_{j=1}^{K} \pi_j$

-> $P\left(\underbrace{X_1 = x_1, \dots, X_{K+1} = x_{K+1}}_{\text{MULTINOMIAL RANDOM VECTOR}}\right) = \underbrace{\frac{n!}{x_1! \cdot \dots \cdot x_{K+1}!}}_{\text{MULTINOMIAL COEFFICIENT}} \cdot \pi_1^{x_1} \cdot \dots \cdot \pi_{K+1}^{x_{K+1}}$ / $n = \sum_{j=1}^{K} x_j + x_{K+1}$

(R.V. ~> OUTCOME)

(RS. BINOMIAL : $K=1$ -> p, $(1-p)$ ~>)

- $\begin{cases} K+1: & n^\circ \text{ OF MUTUALLY EXCLUSIVE EVENTS } (P(A \cap B) = 0) \text{ / } K: n^\circ \text{ R.V.s} \\ n: & n^\circ \text{ OF TRIALS} \end{cases}$

RS.

WHEN 3 POSSIBLE DEFECT A, B, C / X_i = # OF i DEFECTIVE OUTCOME

-> WHEN RANDOM SAMPLING -> MULTINOMIAL (n, π_A, π_B)

- **RANDOM VECTORS** :

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} \text{ / } E(X) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_d) \end{pmatrix} \rightsquigarrow d \times 1 \text{ VECTOR}$$

- $V(X_i)$ AND $CoV(X_i, X_j)$ ARE COLLECTED IN A $d \times d$ MATRIX CALLED "VARIANCE - COVARIANCE MATRIX"

$$\underset{d \times d}{Var\,CoV(\bar{X})} = \begin{pmatrix} V(X_1) & CoV(X_1, X_2) & \dots & CoV(X_1, X_d) \\ CoV(X_1, X_2) & V(X_2) & & \\ \vdots & & \ddots & \\ CoV(X_1, X_d) & & & V(X_d) \end{pmatrix}$$

- $P(\bar{x} \in A) = \begin{cases} \sum f(x_1, \dots, x_d) \\ \int_A \dots \int f(x)\, dx \end{cases}$

es.

IN THE PLACE OF BIRTH EXAMPLE, P OF SEEING NO ITALIANS AT ALL ?

- $P(X_1 = 0,\ X_2 = 0,\ X_3 = ANY) = f(0,0,0) + f(0,0,1) + \dots + f(0,0,10)$

- PROPERTY: THE MARGINAL DISTRIBUTION OF A SUBSET OF COMPONENTS IS AGAIN MULTINOMIAL

$$\rightarrow \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} \sim \text{MULTINOMIAL}(n, \pi_1, \dots, \pi_d) \quad / \quad X_1 \sim Bin(n, \pi_1)$$

$$\rightarrow \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \text{MULTINOMIAL}(n, \pi_1, \pi_2)$$

- THE DISTRIBUTION OF PARTIAL SUMS OF COMPONENTS ARE MULTINOMIAL:

$$\rightarrow (X_1,\ X_2 + \dots + X_d) \sim \text{MULTINOMIAL}(n, \pi_1, \pi_2 + \dots + \pi_d)$$

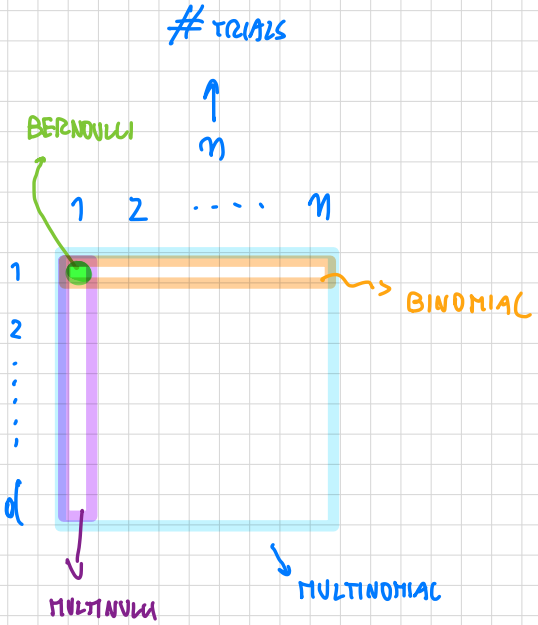
- IF $n = 1$ -> <u>MULTINULLI</u>:

- $X \sim \text{MULTINULLI}(\pi_1, \dots, \pi_d)$

- $E(\vec{X}) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_d) \end{pmatrix} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_d \end{pmatrix}$

- $V(X_i) = \pi_i(1 - \pi_i)$

RVs ← K



- **SUMMARY:**

  - $\bar{X} \sim$ MULTINOMIAL $(n, \tilde{\pi}_1, \dots, \tilde{\pi}_{K+1})$

    - $X_{K+1} = n - \sum_{i=1}^{K} X_i$
    - X_i : INTEGERS IN $0 \le X_i \le n$

    $\hookrightarrow f(x_1, \dots, x_K) = \binom{n}{x_1, \dots, x_{K+1}} \tilde{\pi}_1^{x_1} \cdot \tilde{\pi}_2^{x_2} \cdots \tilde{\pi}_K^{x_K} \cdot \left(1 - \sum_i \tilde{\pi}_i\right)^{x_{K+1}}$ (FROM MULTINOMIAL)

  ↑ —> 1 R.V.

  - $K = 1$: $f(x_1) = \binom{n}{x_1} \cdot \tilde{\pi}_1^{x_1} \cdot (1 - \tilde{\pi}_1)^{n - x_1}$ —> BINOMIAL DISTRIBUTION

    (# EXTRACTIONS IN EXPERIMENT = 1)

  - $n = 1$: $f(x_1, \dots, x_K) = \tilde{\pi}_1^{x_1} \cdot \dots \cdot \tilde{\pi}_K^{x_K} \cdot \left(1 - \sum_j \tilde{\pi}_j\right)^{1 - \sum x_j}$ —> MULTINULLI DISTRIBUTION

  - $K = 1, n = 1$: $f(x_1) = \begin{cases} 1, & p = \tilde{\pi}_1 \\ 0, & p = 1 - \tilde{\pi}_1 \end{cases}$ —> BERNOULLI DISTRIBUTION

- CONSIDER $\bar{X} \sim$ MULMNOULLI $(\tilde{\pi}_1, \dots, \tilde{\pi}_K)$: ~> OUTCOMES SHAPE : $\{0, \dots, 1_i, \dots 0\}$

  -> SINCE $X_i \sim$ Bernoulli $(\tilde{\pi}_i)$ —> $E[X_i] = \tilde{\pi}_i$, $V[X_i] = \tilde{\pi}_i (1 - \tilde{\pi}_i)$

  $\hookrightarrow$ $\begin{cases} 1, & p = \tilde{\pi}_i \\ 0, & p = 1 - \tilde{\pi}_i \end{cases}$

  $\hookrightarrow Cov(X_i, X_j)_{i \neq j} = E[X_i X_j] - E[X_i] E[X_j] =$ (0·0, 0·1, 1·0)

  $= 0 - \tilde{\pi}_i \cdot \tilde{\pi}_j = -\tilde{\pi}_i \tilde{\pi}_j$

- PROPERTIES:

  - IF $\bar{X} \sim$ MULMNOMIAL $(n, \tilde{\pi}_1, \dots, \tilde{\pi}_{K+1})$ -> $\bar{X} = \sum_{i=1}^{K} n_i$ / $n_i \sim$ MULMNULLI $(\tilde{\pi}_1, \dots, \tilde{\pi}_K)$

  - $E(\bar{X}) = \sum_{i=n}^{K} E(n_i) = n \begin{pmatrix} \tilde{\pi}_n \\ \vdots \\ \tilde{\pi}_K \end{pmatrix}$, $V(X_i) = n \tilde{\pi}_i (1 - \tilde{\pi}_i)$

  - $Cov(X_i, X_j) = -n \tilde{\pi}_i \tilde{\pi}_j$

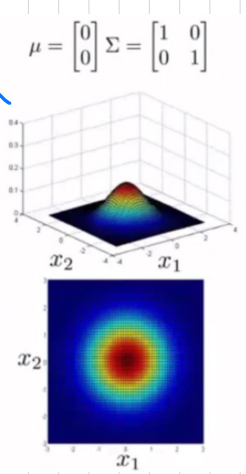
  - $E[\bar{a} \bar{X} + \bar{b}] = \bar{a} E[\bar{X}] + \bar{b}$

  - $Var\,Cov(\underset{l \times K}{A} \cdot \underset{K+1}{\bar{X}}) = A \cdot Var\,Cov(\bar{X}) \cdot A^T$

—>

SO IF $\bar{X} \sim$ MULMNOMIAL $(n, \tilde{\pi}_1, \dots, \tilde{\pi}_K)$ —> $Var\,Cov(\bar{X}) = n \cdot \begin{pmatrix} \tilde{\pi}_1 (1 - \tilde{\pi}_1) & -\tilde{\pi}_i \tilde{\pi}_j & \cdots \\ -\tilde{\pi}_i \tilde{\pi}_j & \tilde{\pi}_2 (1 - \tilde{\pi}_2) & \\ \vdots & & \ddots \end{pmatrix}$

• MULTIVARIATE NORMAL :



K = 2

1. DEFINE THE MULTIVARIATE STD NORMAL :

$$\bar{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_K \end{pmatrix} \Big/ Z_1, \dots, Z_K \overset{iid}{\sim} N(0,1)$$

$$\cdot f(z_1, \dots, z_K) = \prod_{i=1}^{K} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{z_i^2}{2}\right)} = (2\pi)^{-\frac{K}{2}} e^{-\frac{1}{2}\|z\|^2}$$

L_2 NORM

$$\cdot E(\bar{Z}) = \bar{0}^T \quad , \quad Var\,Cov(\bar{Z}) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \\ 0 & \dots & & 1 \end{pmatrix} = I_K$$

IDENTITY MATRIX $K \times K$

2. DEFINE ANY MULTIVARIATE NORMAL VIA LINEAR TRANSFORMATION

• STARTING WITH $\bar{Z} \sim N_K(\bar{0}, I_K)$, $\underset{m \times 1}{\bar{\mu}}$, $\underset{m \times K}{A}$:

→ MULTIVARIATE NORMAL : $\bar{X} = \underset{m \times K}{A} \cdot \underset{K \times 1}{\bar{Z}} + \underset{m \times 1}{\bar{\mu}}$

$$\cdot E(\bar{X}) = A \cdot \underbrace{E(\bar{Z})}_{= \bar{0}^T} + \bar{\mu} = \bar{\mu}$$

$$\cdot Var\,Cov(\bar{X}) = Var\,Cov(A \cdot \bar{Z}) = A \cdot \overbrace{Var\,Cov(\bar{Z})}^{I} \cdot A^T = A \cdot A^T$$

ex.

$$\mu = (0 \;\; 1)^T, \quad A = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \Big/ K = n = 2$$

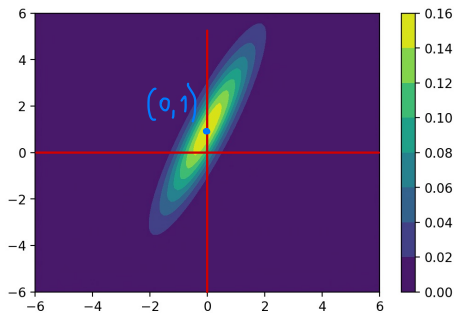
$$\rightarrow \begin{pmatrix} X \\ Y \end{pmatrix} = A \cdot \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} + \mu = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} z_1 \\ 2z_1 + z_2 + 1 \end{pmatrix}$$

• WHERE $z_1, z_2 \overset{iid}{\sim} N(0,1)$

$$\rightarrow E\begin{pmatrix} X \\ Y \end{pmatrix} = \bar{\mu} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\rightarrow Var\,Cov\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$$

$$\rho = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{2}{\sqrt{5}}$$



· PROOF THAT $Cov(X_l, X_m) = -n\,\tilde{\pi}_l\,\tilde{\pi}_m$, IF $m \neq l$:

/ $\bar{X} = \sum M_i$, $M_1, \dots, M_n \sim \text{MULTINOMIAL}(\tilde{\pi}_1, \dots, \tilde{\pi}_k)$

↳ $\sim \text{MULTINOMIAL}(n, \tilde{\pi}_1, \dots, \tilde{\pi}_k)$

$\sim$> l-TH COMPONENT OF i-TH VECTOR $\bar{M}_i$

→

$$Cov(X_l, X_m) = Cov\left(\sum_{i=1}^{n} M_{il}, \sum_{j=1}^{n} M_{jm}\right) =$$

IF $i \neq j$ -> $Cov = 0$, $\because \bar{M}_i$ INDEP.

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(M_{il}, M_{jm}) \overset{!}{=} \sum_{\text{IF } i=j} Cov(M_{il}, M_{im})$$

$$= n\left(-\tilde{\pi}_l\,\tilde{\pi}_m\right) = -n\,\tilde{\pi}_l\,\tilde{\pi}_m$$

· DENSITY OF A MULTIVARIATE NORMAL :

$$\underset{n\times 1}{\bar{X}} \sim \underset{n\times k}{A} \cdot \underset{k\times 1}{\bar{Z}} + \underset{n\times 1}{\bar{\mu}} \sim \mathcal{N}\left(\bar{\mu}, \underset{n\times n}{\Sigma = A \cdot A^T}\right)$$

IF Σ^{-1} EXISTS, WE CAN APPLY THE JACOBIAN METHOD TO PROVE THAT THE DENSITY OF $\bar{X}$ ALSO EXISTS, THAT IS :

ELLIPSOIDE IN $\mathbb{R}^n$

$$\mathbb{R}^n \sim> f(\bar{x}) = (2\tilde{\pi})^{-\frac{n}{2}} \cdot |\det(\Sigma)|^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2}(\bar{x}-\bar{\mu})^T \Sigma^{-1} \cdot (\bar{x}-\bar{\mu})\right\}$$

$\mathbb{R}^2$ ~> · IF $n = 2$:

$$\cdot\ \bar{X} = \begin{pmatrix} X \\ Y \end{pmatrix} \quad \text{->} \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

$$\text{->} \quad \Sigma^{-1} = \frac{1}{\det(\Sigma)}\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix} = \underbrace{\frac{1}{\sigma_X^2\sigma_Y^2 - \rho^2\sigma_X^2\sigma_Y^2}}_{\sigma_X^2\sigma_Y^2(1-\rho^2)} \begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix} =$$

$$= \frac{1}{1-\rho^2}\begin{pmatrix} 1/\sigma_X^2 & -\rho/\sigma_X\sigma_Y \\ -\rho/\sigma_X\sigma_Y & 1/\sigma_Y^2 \end{pmatrix}$$

(... AFTER SOME ALGEBRA), FOR $\rho \neq \pm 1$

→

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

ELLIPSE IN $\mathbb{R}^2$

es.

$\mu = (0 \;\; 1)^T$, $A = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$ $\Big/ k = n = 2$

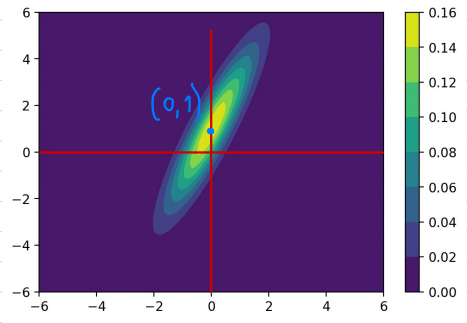
$\rightarrow \begin{pmatrix} X \\ Y \end{pmatrix} = A \cdot \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} + \mu = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} z_1 \\ 2z_1 + z_2 + 1 \end{pmatrix}$

- WHERE $z_1, z_2 \overset{iid}{\sim} N(0,1)$

$\rightarrow E\begin{pmatrix} X \\ Y \end{pmatrix} = \bar{\mu} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

$\rightarrow Var\,Cov \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$

- $\rho = \frac{2}{\sqrt{5}} \neq \pm 1$



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

$$f(X,Y) = (2\pi)^{-1} \cdot \left(1 - \left(\frac{2}{\sqrt{5}}\right)^2 \cdot 1 \cdot 5\right)^{-\frac{1}{2}} \cdot \exp\left\{ -\frac{1}{2\cdot\left(1-\frac{4}{5}\right)} \left[X^2 - 2\cdot\frac{2}{\sqrt{5}} \cdot X \cdot \left(\frac{Y-1}{\sqrt{5}}\right) + \left(\frac{Y-1}{\sqrt{5}}\right)^2 \right] \right\}$$

$$= (2\pi)^{-1} \cdot \exp\left\{ -\frac{5}{2} \left[X^2 - \frac{4}{5} X (Y-1) + \left(\frac{Y-1}{\sqrt{5}}\right)^2 \right] \right\}$$

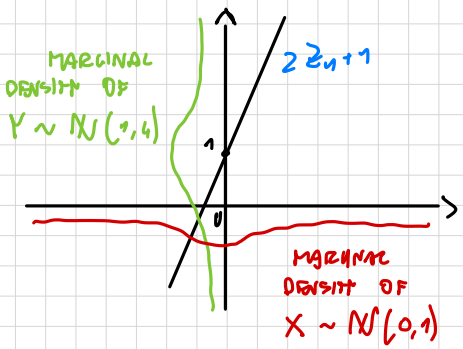
es. DEGENERATE CASE, $\rho = \pm 1$

$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} z_1 \\ 2z_1 + 1 \end{pmatrix}$ $\rightarrow$ LINEAR CORRELATION

$\hookrightarrow \rho = \frac{2}{\sqrt{1\cdot 4}} = 1$

- $E\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
- $Var\,Cov \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$

$\rightarrow$ SINCE $Y = 2X + 1$ $\rightarrow$ NO 2D DENSITY



- PROPERTY :

MARGINAL DISTRIBUTIONS OF SUBVECTORS OF MULTIVARIATE NORMAL VECTORS ARE MULTIVARIATE NORMAL THEMSELVES

-> $\bar{X} = \begin{pmatrix} \bar{X}_1 \\ n_1 \times 1 \\ \bar{X}_2 \\ n_2 \times 1 \end{pmatrix}$

- $E(\bar{X}) = \begin{pmatrix} E(\bar{X}_1) \\ E(\bar{X}_2) \end{pmatrix} = \begin{pmatrix} \bar{\mu}_1 \\ n_1 \times 1 \\ \bar{\mu}_2 \\ n_2 \times 1 \end{pmatrix} = \bar{\mu}$

- $Var\,Cov(\bar{X}) = \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ n_1 \times n_1 & n_1 \times n_2 \\ \Sigma_{12}^T & \Sigma_{22} \\ n_2 \times n_1 & n_2 \times n_2 \end{pmatrix}$

-> $\bar{X}_1 = N_{n_1}(\bar{\mu}_1, \Sigma_{11})$

-> $\bar{X}_2 = N_{n_2}(\bar{\mu}_2, \Sigma_{22})$

- <u>N.B.</u>: Σ IN GENERAL CAN HAVE $n_1 \neq n_2$

es. $n_1 = 1$, $n_2 = n-1$

$n_1 = 1$ { $n_2 = n-1$ {

$\sigma^2_{X_1}$	$Cov(X_1, X_2) \dots Cov(X_1, X_n)$
$Cov(X_1, X_2)$ $\vdots$ $Cov(X_1, X_n)$	Σ_{22}

• CONDITIONAL NORMAL DISTRIBUTIONS :

- IT CAN BE PROVED THAT CONDITIONAL DISTRIBUTIONS OF NORMAL ARE NORMAL

→ IF Σ_{11} IS NOT SINGULAR (Σ_{11}^{-1} EXISTS)

→ $X_2 \mid X_1 = x_1 \sim N_{n_2}\left(\bar{\mu}_2 + \Sigma_{21}\cdot\Sigma_{11}^{-1}\left(\bar{x}_1 - \bar{\mu}_1\right), \Sigma_{22} - \Sigma_{21}\cdot\Sigma_{11}^{-1}\Sigma_{12}\right)$

- IF $n = 2$, $n_1 = 1$, $n_2 = 1$:

$\Leftrightarrow \rho \neq \pm 1$

- $\begin{pmatrix} X \\ Y \end{pmatrix}$: $Y \mid X = x \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X), (1-\rho^2)\sigma_Y^2\right)$

- $P(\underbrace{X \le x_1 \cap Y \le y_1}_{A}) = \left/ \begin{pmatrix} X \\ Y \end{pmatrix} \sim \begin{pmatrix} z_1 \\ 2z_1 + z_2 + 1 \end{pmatrix}\right.$

$$= \iint_A f(x,y)\,dx\,dy = \int_{-\infty}^{x_1}\int_{-\infty}^{y_1} f(x,y)\,dx\,dy =$$

$P(x,y) = P(y)\,P(x \mid y)$

$$= \int_{-\infty}^{x_1} f_X(x)\underbrace{\int_{-\infty}^{y_1} f_{Y\mid X}(y \mid x)\,dy}_{\text{CDF OF A NORMAL}}\,dx$$

```
library(mvtnorm)

# Mean vector: (mu_1, ..., mu_n)
mean <- c(0, 1)
# Covariance matrix
cov <- matrix(c(1, 2,
                2, 4), nrow = 2, ncol = 2)

lower_bounds <- c(-Inf, -Inf)
# Upper bounds
upper_bounds <- c(2, 1)

# Compute the probability
probability <- pmvnorm(
    lower = lower_bounds,
    upper = upper_bounds,
    mean = mean,
    sigma = cov
)
# paste(): concatenate float and str
print(paste("Probability:", probability))
```

```
[1] "Probability: 0.5"
```

- IN GENERAL : $P(X_1 \le x_1, \dots, X_n \le x_n) = \int_{-\infty}^{x_1}\dots\int_{-\infty}^{x_n} f(t_1,\dots,t_n)\,dt_1\cdot\dots\cdot dt_n$ → SOLVABLE IN PYTHON, R

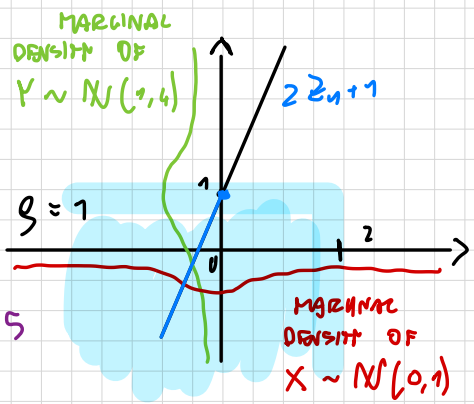
ex.

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} z_1 \\ 2z_1 + 1 \end{pmatrix} \rightarrow E\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\rightarrow \mathrm{Var\,Cov}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

- $P(X \le 2 \cap Y \le 1) =$

$$= P\left[\begin{pmatrix} X \\ Y \end{pmatrix} \in A\right] = P[X \le 0] = 0.5$$



SIMULATION :

- THE BASE IS TO EXTRACT x_0 FROM $X \sim U[0,1]$

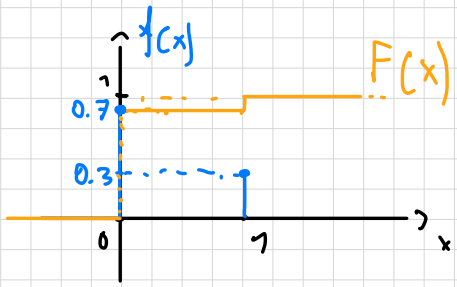
  ↳ `np.random.random(1)`

→ EVERY EXTRACTION OF $X \sim f(x)$ CAN BE DONE BASED ON THE EXTRACTION FROM THE UNIFORM DISTRIBUTION :

es.

$X \sim \text{Bernoulli}(p_0)$:

→ RETURN $\begin{cases} 0 & \text{IF } x_0 \leq p_0 \\ 1 & \text{IF } x_0 > p_0 \end{cases}$



es.

$X \sim \text{Binomial}(n = n_0, p = p_0)$

→ COMPUTE CDF AND APPLY SAME TRICK

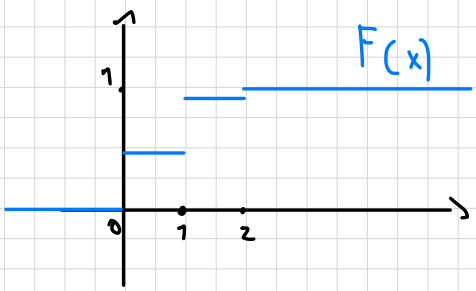
$$f_X(x_0) = \begin{cases} k=0 : \binom{n_0}{0} p_0^0 (1-p_0)^{n_0} \\ k=1 : \binom{n_0}{1} p_0^1 (1-p_0)^{n_0-1} \\ \dots \end{cases}$$

$n_0 = 2,\ p_0 = 0.3$

$= 0.49$

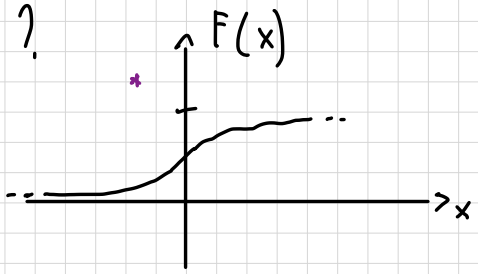
$= 0.42$

→ $F_X(x_0) = \begin{cases} 0 & , x \leq 0 \\ 0.49 & , 0 < x \leq 1 \\ 0.91 & , 1 < x \leq 2 \\ 1 & , x \geq 2 \end{cases}$



→ RETURN $\begin{cases} 0 & , p = 0.49 \\ 1 & , p = 0.42 \\ 2 & , p = 0.09 \end{cases}$

- WHAT IF X HAS DENSITY $f_X(x)$ CONTINOUS ?

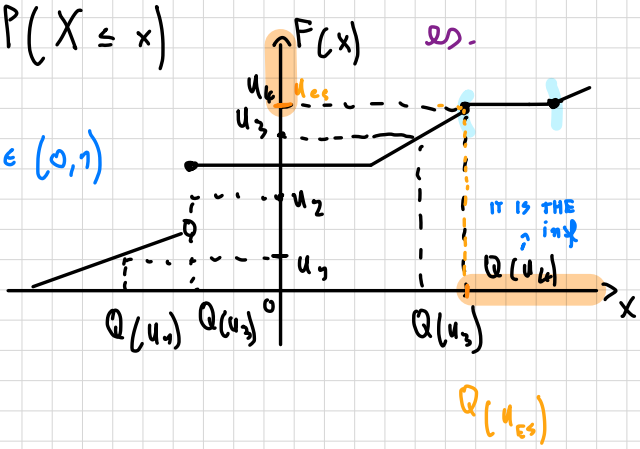


  - $F_X(x_0) = P(X \leq x_0)$

  ex. $Z \sim N(0,1)$ *

-> WE NEED TO DEFINE A GENERALIZED INVERSE OF F :

- IF X HAS CDF $F(x) = P(X \leq x)$



-> QUANTILE FUNCTION :

$$Q(u) = \inf\{x : F(x) \geq u\} \quad \forall u \in (0,1)$$

- THEOREM :

$$Q(u) \leq x \iff u \leq F(x)$$

$$\forall u \in (0,1), \ \forall x \in \mathbb{R}$$

- PROOF :

  - IF $u \leq F(x)$ -> $Q(u)$ IS AN INFIMUM : $Q(u) \leq x$
  - IF $Q(u) \leq x$ -> $F(Q(u)) \leq F(x)$

  -> $u \leq F(Q(u)) \leq F(x)$

- THEOREM :

IF X IS A R.V. WITH QUANTILE FUNCTION $Q(u)$, $u \in (0,1)$
AND IF $U \sim$ UNIFORM $(0,1)$

-> $Q(u)$ AND X HAVE SAME DISTRIBUTION

- PROOF :

  WHEN $F_X(x) = P(X \leq x)$, CDF OF X

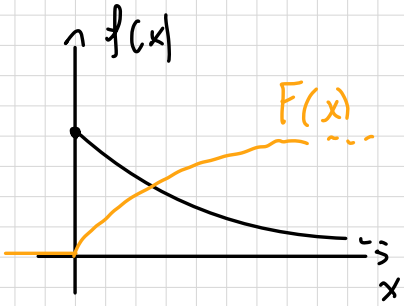
  -> CDF OF $Q(U)$: $F_{Q(U)}(x) = P(Q(U) \leq x) \overset{\text{THEOREM ABOVE}}{=} P(U \leq F(x)) \overset{\text{SINCE } U \text{ IS UNIFORM}}{=} F(x)$

-> THIS RESULTS SIMPLIFIES THE GENERAL ALGORITHM TO SIMULATE AN UNIVARIATE R.V. :

1. OBTAIN $U \sim \text{UNIFORM}(0,1)$
2. RETURN $X = Q(U)$

es.

$X \sim Exp(\lambda) \rightarrow f_X(x) = \lambda e^{-\lambda x}, \quad x > 0$



- $F_X(x) = \int_0^x \lambda e^{-\lambda x} = -e^{-\lambda t} \Big|_0^x = 1 - e^{-\lambda t}$

$\left(y = 1 - e^{-\lambda x} \Rightarrow 1 - y = e^{-\lambda x} \Rightarrow \ln(1-y) = -\lambda x \rightarrow x = -\frac{\ln(1-y)}{\lambda} \right)$

- $Q(u) = F_X^{-1}(x)$ -> (NOT SURE AS NOTATION)

-> SIMULATION :

1. GET $U \sim \text{UNIFORM}(0,1)$
2. RETURN $X = -\frac{\ln(1-U)}{\lambda}$ TO SIMULATE $X \sim Exp(\lambda)$

- HOW DO WE SIMULATE U EXTRACTION (STEP 1.)

↳ USE OF PSEUDO-RANDOM NUMBER GENERATION

- ONE CLASS OF ALGORITHM TO DO THAT : LINEAR CONGRUENTIAL GENERATORS
  -> THEY EXPLOIT THE UNPREDICTABILITY OF THE REMAINDER OF A DIVISION BETWEEN 2 INTEGERS

- ALGORITHM :
  - CHOOSE A POSSIBLY LARGE NUMBERS : M, a, c ~> USED VALUES US. : $a = 7^5$, $M = 2^{31} - 1$, $c = 0$
  - CHOOSE SEED L_0
  - DEFINE FOR $i = 1, 2, \ldots$: $L_i = (a L_{i-1} + c) \bmod M$

-> EXTRACT $u_i = \frac{L_i}{M}$: RATIONAL NUMBER IN $[0,1]$

• MULTIVARIATE SIMULATION

HOW DO WE SIMULATE FROM $\bar{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$, WITH A DENSITY $f(x_1, \dots, x_d)$?

Σ^{-1} EXISTS

->

- TAKE $\bar{X} \sim N_d(\bar{\mu}, \Sigma)$ AND FIND $A \,/\, \Sigma = A \cdot A^T$, FOR EX.: $A = \text{chol}(\Sigma)$ (CHOLESKY DECOMPOSITION)

```
A <- matrix(c(
    4, 12, -16,
    12, 37, -43,
    -16, -43, 98
), nrow = 3, ncol = 3)
L_T <- chol(A) # A = L * L^T
```

-> TRIANGULAR LOWER

L_T
[,1] [,2] [,3]
[1,] 2 6 -8
[2,] 0 1 5
[3,] 0 0 3

L
[,1] [,2] [,3]
[1,] 2 0 0
[2,] 6 1 0
[3,] -8 5 3

$A = LL^T$
[,1] [,2] [,3]
[1,] 4 12 -16
[2,] 12 37 -43
[3,] -16 -43 98

- THEN:

1) SIMULATE $\bar{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix}$, iid $N(0,1)$

2) SET $\bar{X} = \bar{\mu} + A \cdot \bar{Z}$, IN THIS WAY $\bar{X} \sim N_d(\bar{\mu}, AA^T)$

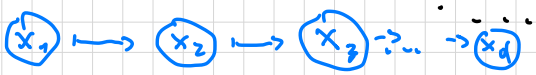
- THIS WORKS ONLY WITH MULTIVARIATE NORMAL
  -> PROBLEM CAN BE REDUCED TO A SERIES OF UNIVARIATE SIMULATION

- A GENERAL ALGORITHM TO APPLY FOR ANY DISTRIBUTION, TO GENERATE $\bar{X}$:

BACKWARDS $f(x,y) = f_X(x) \cdot f_{Y|X=x}(x)$

$$f(x_1, \dots, x_d) = f_{X_1}(x_1) \cdot f_{X_2|X_1}(x_2|x_1) \cdot f_{X_3|X_1,X_2}(x_3|x_1,x_2) \cdot \dots \cdot f_{X_d|X_1 \dots X_{d-1}}$$

COMPLETE DAG



- ALGORITHM:

1. GENERATE x_1 FROM THE MARGINAL f_{X_1}
2. " x_2 " $f_{X_2|X_1}$
   ⋮
d. " x_d " $f_{X_d|X_1,\dots,X_{d-1}}$

GIBBS SAMPLER

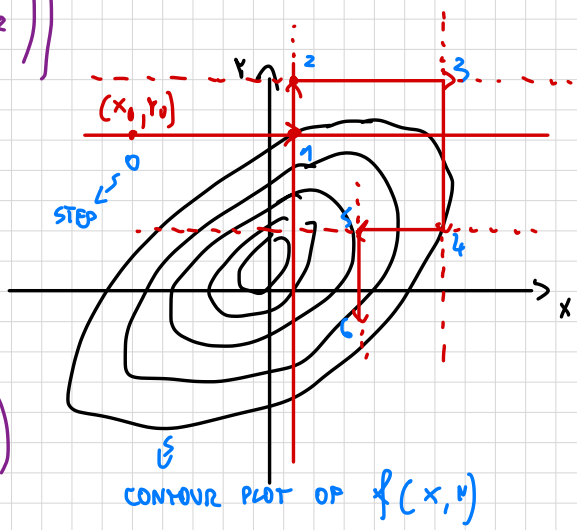
ex.

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \right)$$

0. START AT $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$

FOR $j = 1, \dots, B$:

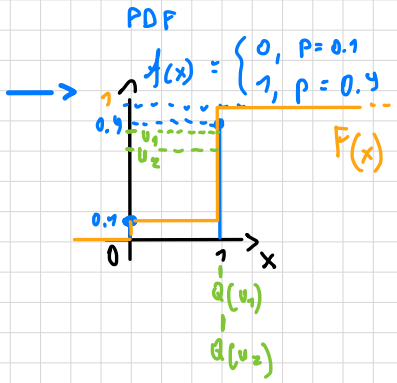
1. GENERATE x_{j+1} FROM $f_{X|Y}(x \mid y_j)$
2. GENERATE y_{j+1} FROM $f_{Y|X}(y \mid x_{j+1})$



-> AT ITERATION B : (x_B, y_B) WITH JOINT DENSITY $f(x, y)$

-> (x_B, y_B) WILL BE A GENUINE SIMULATED VALUE FROM $f(x, y)$

- UTILITY :



SIMULATION :

```
x <- round(runif(1), 2)
```

$u_1 \sim U[0,1]$: 0.86 -> 1

u_2 : 0.71 -> 1

u_3 : ....

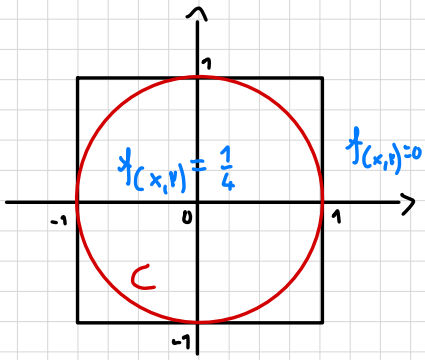
- CONTOUR PLOT OF A MULTIVARIATE NORMAL :

MONTE CARLO SIMULATION:

GIVEN X, Y iid $\sim U[-1,1]$

$$f_X(x) = f_Y(y) = \frac{1}{2} \quad (-1 < x < 1)$$

$$\rightarrow f_{X,Y}(x,y) = \frac{1}{4}$$



- WANT IS $P\left[\begin{pmatrix} X \\ Y \end{pmatrix} \in C\right]$?

$$\rightarrow P\left[\begin{pmatrix} X \\ Y \end{pmatrix} \in C\right] = \iint_C f_{(x,y)}\, dx\, dy = \frac{1}{4} \iint_C dx\, dy =$$

$$= \frac{1}{4} (\text{AREA OF } C) = \frac{1}{4} \cdot (\tilde{\pi} \cdot 1^2) = \frac{\tilde{\pi}}{4}$$

MONTE CARLO APPROXIMATION:

1. GENERATE $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \dots, \begin{pmatrix} X_B \\ Y_B \end{pmatrix}$ iid $\begin{pmatrix} X \\ Y \end{pmatrix}$ FOR LARGE B

2. SET $y(x,y) = \left(\begin{pmatrix} X \\ Y \end{pmatrix} \in C\right) = \begin{cases} 1 & \text{IF } \begin{pmatrix} X \\ Y \end{pmatrix} \in C \\ 0 & \text{IF } \begin{pmatrix} X \\ Y \end{pmatrix} \notin C \end{cases}$

- BY LLN: $\frac{1}{B} \sum_{i=1}^{B} y(x_i, y_i) = \frac{\#\text{ TIMES } \begin{pmatrix} x \\ y \end{pmatrix} \in C}{B} \approx P\left(\begin{pmatrix} X \\ Y \end{pmatrix} \in C\right)$

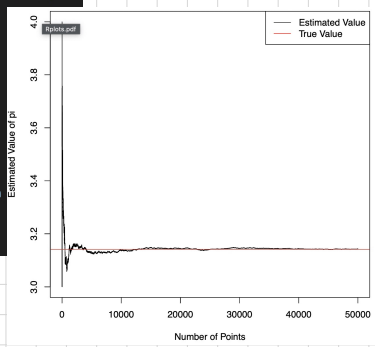
$$\rightarrow \frac{1}{B} \sum_{i=1}^{B} y(x_i, y_i) \overset{\text{APPROXIMATION}}{\approx} E[y(x,y)]$$

- TO APPROXIMATE $\tilde{\pi}$:

$$\rightarrow \frac{1}{B} \Sigma y(x_i, y_i) \approx \frac{\tilde{\pi}}{4}$$

$$\rightarrow \frac{4}{B} \Sigma y(x_i, y_i) \approx \tilde{\pi}$$

```r
# Number of random points
n <- 10000000

# Generate random x and y coordinates within the square
x <- runif(n, -1, 1)
y <- runif(n, -1, 1)

# Calculate distance from the origin
dist <- sqrt(x^2 + y^2)

# Check if points fall within the circle
points_in_circle <- dist <= 1

# Estimate pi
pi_estimate <- 4 * sum(points_in_circle) / n

# Print the estimated value of pi
print(pi_estimate)
```

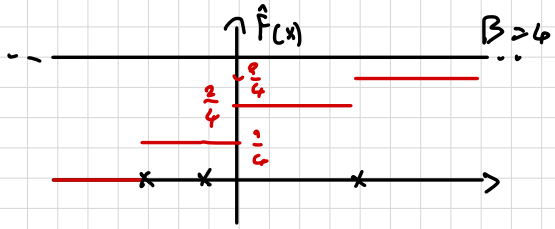


25. EMPIRICAL CDF

SUPPOSE X UNIVARIATE R.V. $\quad F(x) = \int_{-\infty}^{x} f(t)\,dt = P(X \le x)$

- HOW TO APPROX. $F(x_0)$:

1. GENERATE $x_1, \dots, x_B$ iid X
2. $\hat{F}(x_0) = \frac{1}{B} \sum_{i=1}^{B} (x_i \le x_0)$



- EMPIRICAL CDF DO NOT PROVIDE AN ESTIMATE OF $f(x)$
  - $\frac{d}{dx} \hat{F}(x)$ DO NOT MAKE A MEANING

-> AN HISTOGRAM IS A DENSITY ESTIMATE, BUT VERY PRIMITIVE

- HOW TO MEASURE VARIABILITY IN A MEASURE?
  - $\bar{y} = \frac{1}{B} \sum_{i=1}^{B} y(\tilde{x}_i^*)$: SAMPLE MEAN ESTIMATING $E[y(\tilde{x})]$
    - $\tilde{x}_1^*, \dots, \tilde{x}_B^*$ ARE SYNTHETIC OBSERVATIONS
  - IF $V(y(\tilde{x}))$ IS ASSUMED TO EXISTS -> CLT APPLIES:

$$\frac{\bar{y} - E[y(\tilde{x})]}{\sqrt{\frac{V(y(\tilde{x}))}{B}}} \xrightarrow[B \to \infty]{d} N(0,1)$$

-> SAMPLE VARIANCE: $s_y^2 = \frac{\sum_{i=1}^{B} (y(\tilde{x}_i^*) - \bar{y})^2}{B-1}$

-> APPROXIMATE ASYMPTOTIC C.I. OF LEVEL $1-\alpha$: $\bar{y} \pm z_\alpha \sqrt{\frac{s_y^2}{B}}$

- SOMETIMES WE CANNOT DO iid SIMULATION -> iid MONTE CARLO NOT POSSIBLE: MARKOV CHAIN MONTE CARLO

ESSENCE OF MCMC:

-> 2 WAYS:

1. STORE MANY MARKOV CHAINS AND USE LAST OBSERVATION $\tilde{x}_i^t$

OR

2. USE A SINGLE VERY LONG MARKOV CHAIN AND SPLIT IT

} THERE WILL BE SIMILAR LLN FOR MARKOV CHAINS ENSURING $\bar{y}$ IS A GOOD ESTIMATE OF $E(y(\tilde{x}))$

· LINEAR MODELS :

A QUANTITATIVE RESPONSE VARIABLE Y HAS TO BE EXPLAINED IN TERMS OF 1 OR MORE PREDICTORS (FEATURES)

$$\bar{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ \vdots & & & \\ 1 & x_{m,1} & \dots & x_{m,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_m \end{pmatrix}$$

RANDOM VECTOR OF OBSERVATIONS — INTERCEPT COLUMN — $p-1$ FEATURES — VECTOR OF COEFFICIENTS — ERROR TERM

$$\rightarrow Y_i = \beta_0 + \beta_1 \cdot x_{i,1} + \dots + \beta_{p-1} \cdot x_{i,p-1} + \varepsilon_i$$

- ONCE OBSERVED $Y_1, \dots, Y_m$ BECOMES NUMBERS $y_1, \dots, y_m$
- $\varepsilon_1, \dots, \varepsilon_m$ ARE ERRORS / iid NORMAL $(0, \sigma^2)$
- $\beta_0, \dots, \beta_{p-1}$ ARE UNKNOWN COEFFICIENTS
- $x_{i,1}, \dots, x_{i,p-1}$ WILL BE A VECTOR OF CONSTANT FEATURES ASSOCIATED TO THE i-TH RESPONSE

$$\rightarrow Y_i = N\left(\beta_0 + \beta_1 \cdot x_{i,1} + \dots + \beta_{p-1} \cdot x_{i,p-1},\ \sigma^2\right)$$

(ERROR VARIANCE)

es.

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{matrix} \} m_1 \\ \} m - m_1 \end{matrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{pmatrix}$$

1 BINARY FEATURE ONLY

- $Y_i = \beta_0 + \varepsilon_i \quad , i = 1, 2, \dots, m_1$
- $Y_i = \beta_0 + \beta_1 + \varepsilon_i \quad , i = m_1+1, \dots, m$

$\rightarrow$

$$Y_1, \dots, Y_{m_1} \text{ iid } N(\underbrace{\beta_0}_{\mu}, \sigma^2)$$

$$Y_{m_1+1}, \dots, Y_m \text{ iid } N(\underbrace{\beta_0 + \beta_1}_{\nu}, \sigma^2)$$

- THE PREVIOUS EXAMPLE IS THE TWO-NORMAL-SAMPLE PROBLEM IN TWO GROUP EXPERIMENT:

WE ARE COMPARING TWO MEANS μ AND ν IN LINEAR MODEL: $\begin{cases} \mu = \beta_0 \\ \nu = \beta_0 + \beta_1 \end{cases}$

-> $\nu - \mu = \beta_1$: DIFFERENCE OF 2 NORMAL MEANS, EACH GROUP SAME $V(\cdot)$

↳ t-TEST AND RELATED t C.I.

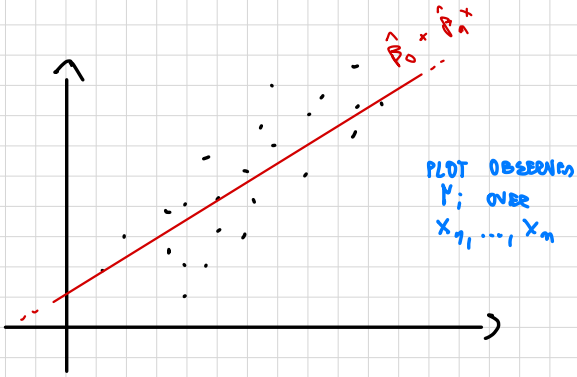
EX. 1 SAMPLE MODEL, NULL MODEL

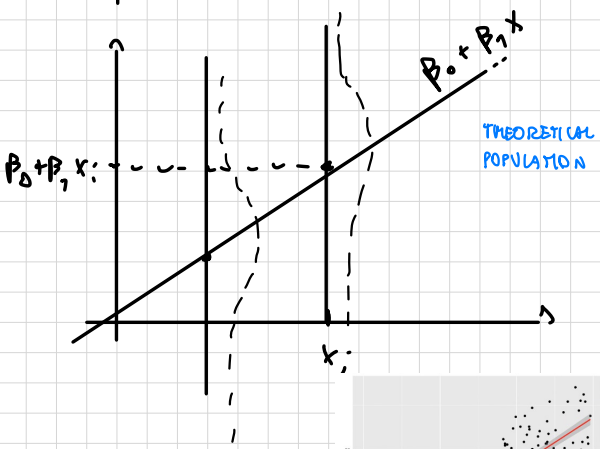
$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \beta_0 + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \rightarrow Y_1, \dots, Y_n \overset{iid}{\sim} N(\underbrace{\beta_0}_{\mu}, \sigma^2)$$

EX. SIMPLE LINEAR REGRESSION -> $p = 2$, 1 QUANTITATIVE FEATURE

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

-> $Y_i \sim N(\beta_0 + \beta x_i, \sigma^2)$





- WE DO NOT KNOW $\beta_0, \beta_1, \sigma^2$

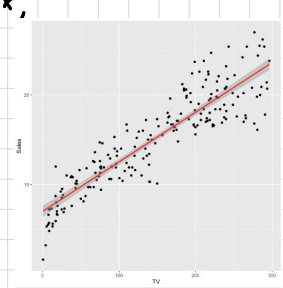
-> WE WILL ESTIMATE THEM BASED ON DATA

- IN R:

```
advertising <- read_csv("./datasets/advertising.csv")
```

$\bar{Y}$ $\bar{X}_1$

```
simple_reg <- lm(Sales ~ TV, data = advertising)
```



• ESTIMATION IN LINEAR MODELS:

LINEAR MODEL:

$$\underset{n\times 1}{\bar{Y}} = \underset{n\times p}{X} \cdot \underset{p\times 1}{\beta} + \underset{n\times 1}{\bar{\varepsilon}}$$

SINCE $\bar{\varepsilon} \sim N_n\left(\bar{0}, \sigma^2 \underset{n\times n}{I_n}\right)$:

- $\bar{Y} = X\bar{\beta} + \bar{\varepsilon}$ IS A LINEAR COMBINATION OF MULTIVARIATE NORMAL

$\rightarrow \bar{Y} \sim N\left(E(\bar{Y}), \text{VarCoV}(\bar{Y})\right)$

- $E(\bar{Y}) = X\bar{\beta} + \underbrace{E(\bar{\varepsilon})}_{=0} = X\bar{\beta}$

- $\text{VarCoV}(\bar{Y}) = \text{VarCoV}(X\bar{\beta} + \bar{\varepsilon}) = \text{VarCoV}(\bar{\varepsilon}) = \sigma^2 I$

$\rightarrow \bar{Y} \sim N(X\bar{\beta}, \sigma^2 I)$

$\hookrightarrow \bar{Y}$ ARE INDEPENDENT, BUT NOT iid

- MLE TO ESTIMATE PARAMETERS $\beta_0, \dots, \beta_{p-1}, \sigma^2$:

$$\mathcal{L}(\bar{\beta}, \sigma^2; Y_1, \dots, Y_n) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2\sigma}\left(Y_i - \sum_{j=1}^{B-1}\beta_j \cdot x_{i,j}\right)^2\right]\right] =$$

$$= (2\pi\sigma^2)^{-n/2} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\underbrace{\left(Y_i - \sum_{j=1}^{B-1}\beta_j \cdot x_{i,j}\right)^2}_{(\bar{Y}-X\bar{\beta})^T\cdot(\bar{Y}-X\bar{\beta}) \text{ IN MATRIX FORM}}\right\}$$

- MAXIMIZING $\log\mathcal{L}$:

$$\log\mathcal{L}(\beta, \sigma^2, Y_1, \dots, Y_B) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\bar{Y}-X\bar{\beta})^T(\bar{Y}-X\bar{\beta})$$

- **N.B.:** $\max \log \mathcal{L} \iff \min \left(\bar{Y} - X\bar{\beta}\right)^T \left(\bar{Y} - X\bar{\beta}\right)$

$$\rightarrow \min \left(\bar{Y} - X\bar{\beta}\right)^T \left(\bar{Y} - X\bar{\beta}\right) = \min_{\beta} \left(Y^T Y - 2 \cdot \underbrace{\beta^T X^T Y}_{1\times p \;\; p\times n \;\; n\times 1 \;\rightarrow\; 1\times 1} + \beta^T X^T X \beta\right)$$

- **N.B.:** IMPORTANT RULES FOR MATRIX DIFFERENTIATION

$$\cdot \frac{\partial}{\partial x} A\bar{x} = A^T \qquad , \qquad \cdot \frac{\partial}{\partial x} x^T A x = 2Ax$$

$$\rightarrow \frac{\partial}{\partial \beta} \left(Y^T Y - 2 \cdot \underbrace{\beta^T X^T Y}_{1\times 1} + \beta^T X^T X \beta\right) = -2X^T Y + 2X^T X \bar{\beta} = 0$$

$$\rightarrow \bar{X}^T \cdot \bar{X} \cdot \bar{\beta} = \bar{X}^T \cdot \bar{Y} \qquad \text{NORMAL EQUATION}$$

- PROVIDED THAT $\left(\bar{X}^T \bar{X}\right)^{-1}$ EXISTS:

$$\rightarrow \beta = \left(\bar{X}^T \bar{X}\right)^{-1} \cdot \bar{X}^T \cdot \bar{Y}$$

$\rightarrow$

ESTIMATED COEFFICIENTS

- $\hat{\beta} = \left(X^T \cdot X\right)^T \cdot X \cdot \bar{Y}$ — LS ESTIMATE OF β, CORRESPONDING TO OUR DATA
- $\hat{\beta} = \left(X^T \cdot X\right)^T \cdot X \cdot \bar{Y}$ (R.VEC.) — LS ESTIMATE OF β, A RANDOM VECTOR

- $\hat{Y} = X \cdot \hat{\beta}$ — PROJECTION OF OUR DATA ONTO SPAN (X)
- $\bar{Y} = X \cdot \bar{\beta}$ — CORRESPONDING RANDOM VECTOR

RESIDUALS

- $\bar{e} = \bar{Y} - \hat{Y}$ — DIFFERENCE BETWEEN OBSERVED DATA AND THEIR PROJECTION
- $\bar{E} = \bar{Y} - \hat{Y}$ — RESIDUALS

ex. NULL MODEL $\rightsquigarrow Y_i = \beta_0 + \varepsilon_i$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{X} \cdot \beta_0 + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\rightarrow \quad \beta_0 = (X^T X)^{-1} \cdot X^T Y = \begin{pmatrix} (1 \dots 1) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \end{pmatrix} \cdot (1 \dots 1) \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} =$$

$$= \dots = \frac{1}{n} \sum Y_i$$

$\rightarrow$ LSE OF β_0 IS THE SAMPLE MEAN

ex. LINEAR REGRESSION $\rightsquigarrow Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

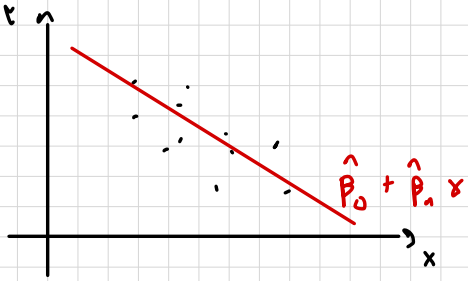
$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\cdot \hat{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = (X^T \cdot X)^{-1} X^T \cdot Y = \left(\begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \right)^{-1} \cdot \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} =$$

$$= \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} =$$

$$= \dots = \begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 = \dfrac{\sum (x_i - \bar{x}) Y_i}{\sum (x_i - \bar{x})^2} \end{cases}$$

$\bar{x}$ MEAN



- WHAT ABOUT σ^2 ?

$$\rightarrow \frac{\partial}{\partial \sigma^2} \log \mathcal{L}\left(\hat{\beta}, \sigma^2\right) = 0 \quad \text{(log LIKELIHOOD)}$$

$$\rightarrow \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(Y - X\hat{\beta}\right)^T\left(Y - X\hat{\beta}\right)\right) = 0$$

$$\rightarrow -\frac{n}{2}\frac{1}{2\pi\sigma^2}2\pi - \frac{1}{2}\left(Y - X\hat{\beta}\right)^T\left(Y - X\hat{\beta}\right)\frac{1}{(\sigma^2)^2} = 0$$

- IF WE CAN EXCLUDE THE SOLUTION $\sigma^2 = 0$:

$$\rightarrow \hat{\sigma}^2 = \frac{\left(Y - X\hat{\beta}\right)^T\left(Y - X\hat{\beta}\right)}{n} = \frac{e^T \cdot e}{n}$$

(e: RESIDUAL $\bar{e}$; $\left(Y - X\hat{\beta}\right)^2 = \left(Y - \hat{Y}\right)^2$; $e^T \cdot e$: SUM OF SQUARED RESIDUALS)

↳ IT IS ALSO THE MLE OF $\hat{\sigma}^2$

2D. NULL MODEL

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{X} \cdot \beta_0 + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \rightsquigarrow Y_i = \beta_0 + \varepsilon_i$$

$$\rightarrow \hat{\sigma}^2 = \frac{e^T \cdot e}{n} = \frac{\sum_i \left(Y_i - \hat{\beta}_0\right)^2}{n} = \frac{\sum\left(Y_i - \bar{Y}\right)^2}{n}$$

- ACTUALLY A BETTER ESTIMATE OF σ^2 : $S^2 = \frac{\sum\left(Y_i - \bar{Y}\right)^2}{n-1}$

... OR : $MSR = \frac{e^T \cdot e}{n-p}$ (p: DEGREE OF FREEDOMS)

- $\hat{\beta}$ IS A R.V. ITSELF : IT'S ALSO NORMAL SINCE IS A LINEAR TRANS. OF Y ($\rightsquigarrow \mathcal{N}_m\left(E(\beta), \mathrm{VarCoV}(\hat{\beta})\right)$)

$$\cdot E\left(\hat{\beta}\right) = E\left(\left(X^TX\right)^{-1}X^TY\right) = \left(X^TX\right)^{-1}X^T \cdot E(Y) = \left(X^TX\right)^{-1}X^T \cdot X\beta = \beta \quad (E(Y))$$

$$\cdot \mathrm{VarCoV}\left(\hat{\beta}\right) = \mathrm{VarCoV}\left(\left(X^TX\right)^{-1}X^TY\right) = \left(X^TX\right)^{-1}X^T \cdot \mathrm{VarCoV}(Y) \cdot \left(\left(X^TX\right)^{-1}X^T\right)^T =$$

($\mathrm{VarCoV}(AX) = A \cdot \mathrm{VarCoV}(X) \cdot A^T$)

$$= \left(X^TX\right)^{-1}X^T \cdot \sigma^2 I \cdot \left(X^TX\right)^{-1} \cdot X = \left(X^TX\right)^{-1} \cdot \sigma^2 I$$

→ A SINGLE COMPONENT : $\hat{\beta}_i \sim \mathcal{N}\left(\beta_i, \sigma^2\left(X^TX\right)^{-1}_{i+1,i+1}\right)$, $i = 0, \dots, p-1$

($\left(X^TX\right)^{-1}_{i+1,i+1}$: SQUARED OF STANDARD ERROR OF $\hat{\beta}_0$; σ^2 IT WILL BE ESTIMATED WITH MSR)

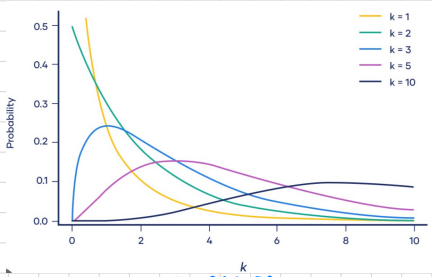
CONFIDENCE INTERVALS FOR β.

- RECALL: $\bar{Y} = \underset{n \times p}{X} \cdot \beta + \varepsilon$ (X: GIVEN) $\varepsilon \sim N(\bar{0}, \sigma^2 I)$ → LINEAR MODEL
  - ε: R.V., OBSERVED INTO VALUES $\bar{Y}$
  - n: # SAMPLES, p: # FEATURES
- AN ESTIMATE (MLE) IS: → ESTIMATOR: $\hat{\sigma}^2 = \frac{E^T E}{n}$

$$\hat{\sigma}^2 = \frac{e^T e}{n} = \text{SUM OF SQUARED RESIDUALS} = \frac{\sum_i e_i^2}{n}$$

e_i (OBSERVED RESIDUALS): $e_i = Y_i - \hat{Y}_i$

- IT CAN BE PROVEN THAT: $\frac{E^T E}{\sigma^2} \sim \chi^2_{n-p}$ (DOF)
  - IF R.V. $X \sim \chi^2_k \rightarrow E[X] = k$



$$\rightarrow E\left[\frac{E^T E}{\sigma^2}\right] = n - p \rightarrow E\left[\frac{E^T E}{n-p}\right] = E[MSR] = \sigma^2$$

(UNBIASED)

↳ MSR IS A BETTER ESTIMATOR THAN $\frac{E^T E}{n}$, WHICH IS BIASED

- C.I.:
  - $\hat{\beta}_i \sim N\left(\beta_i, \sigma^2 (X^T X)^{-1}_{i+1, i+1}\right)$
  - THEOREM: $\frac{N(0,1)}{\sqrt{\frac{\chi^2_k}{k}}} \sim t_k$

→ WE CAN STANDARDIZE IT:

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 (X^T X)^{-1}_{i+1, i+1}}} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\frac{E^T E}{n-p} \cdot \frac{\sigma^2}{\sigma^2} (X^T X)^{-1}_{i+1, i+1}}} = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-p}}{n-p}}} \sim t_{n-p}$$

- $\sqrt{\sigma^2 (X^T X)^{-1}_{i+1, i+1}}$ = SE$(\hat{\beta}_i)$
- $\frac{E^T E}{n-p} \cdot \frac{1}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p}$
- $\hat{\beta}_i - \beta_i \sim N(0,1)$
- THEOREM

- **C.I. FOR β:**

  - t - STUDENT C.I. FROM ELEMENTARY STATISTICS: $SE(\hat{\beta}_i) = \sqrt{\sigma^2 (X^T X)^{-1}_{i+1,i+1}} = \sqrt{\frac{S^2}{n}}$

    - SINCE: $P\left[\beta_i - t_{\frac{\alpha}{2}, n-p} \cdot SE(\hat{\beta}_i) \le \hat{\beta}_i \le \beta_i + t_{\frac{\alpha}{2}, n-p} \cdot SE(\hat{\beta}_i)\right] = 1-\alpha$

    $\rightarrow$ $C.I._{1-\alpha} = \left[\beta_i - t_{\frac{\alpha}{2}, n-p} \cdot SE(\hat{\beta}_i) \; ; \; \beta_i + t_{\frac{\alpha}{2}, n-p} \cdot SE(\hat{\beta}_i)\right]$

- **C.I. FOR $\hat{\sigma}^2$:**

$$C.I._{1-\alpha}: \quad P\left[\chi^2_{1-\frac{\alpha}{2}, n-p} < \frac{E^T E}{\sigma^2} < \chi^2_{\frac{\alpha}{2}, n-p}\right] = 1-\alpha$$

ex. TWO SAMPLE PROBLEM

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{matrix} \} n_1 \\ \} n - n_1 \end{matrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{pmatrix}$$

1 BINARY FEATURE ONLY

- $Y_i = \beta_0 + \varepsilon_i \quad , i = 1, 2, \dots, n_1$
- $Y_i = \beta_0 + \beta_1 + \varepsilon_i \; , i = n_1+1, \dots, n$

$\rightarrow$ $Y_1, \dots, Y_{n_1}$ iid $N(\underbrace{\beta_0}_{\mu}, \sigma^2)$

$Y_{n_1+1}, \dots, Y_n$ iid $N(\underbrace{\beta_0 + \beta_1}_{\nu}, \sigma^2)$

$\rightarrow$

DIFFERENCE OF MEANS: $\beta_1 = \mu - \nu$

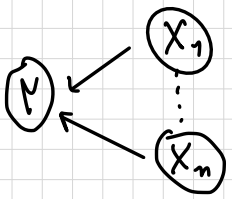
$\rightarrow$ $p = 2$.·. 2 FEATURES

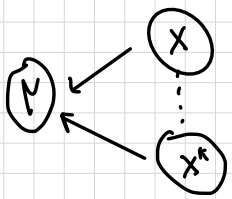
$\rightarrow$ $C.I._{1-\alpha} = \left[\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} \cdot SE(\hat{\beta}_1) \; ; \; \hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2} \cdot SE(\hat{\beta}_1)\right]$

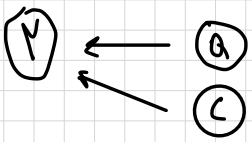
- **INTUITIVE GRAPHS :**

  - Y : RESPONSE
  - $X_1, X_2, \dots$; $Z_1, Z_2, \dots$: QUANTITATIVE FEATURES
  - $b_1, b_2, \dots$: BINARY FEATURES
  - $a_1, a_2, \dots$: CATEGORICAL FEATURES

->

- (Y) NULL MODEL
- (Y) ← (b) . 2 NORMAL SAMPLE PROBLEM
- (Y) ← (X) : SIMPLE LINEAR REGRESSION
- 

. MULTIPLE LINEAR REGRESSION
- 

: POLYNOMIAL REGRESSION ~> BUT LINEARITY IS IN β
- (Y) ← (a) : SINGLE CATEGORICAL FEATURE (ONEWAY ANOVA)
- 

: TWO-WAY ANOVA

TESTS ON β_i :

- NULL HYPOTHESIS -> $H_0 : \beta_i = \beta_i^0$

- IF WE TEST $H_0 : \beta_i = \beta_i^0 = 0$ -> WE ARE TESTING IF FEATURE X_i HAS INFERENCE ON Y :

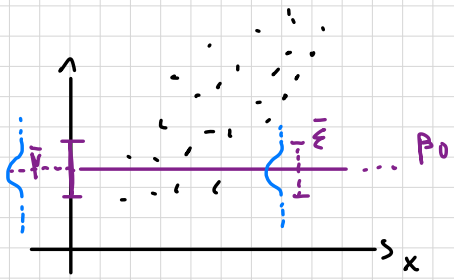
$$\tilde{Y} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ \vdots & \vdots & & \\ 1 & x_{m1} & \cdots & x_{m,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_i = 0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{pmatrix}$$

$X_{(\cdot,i)} = 0$

ex. SIMPLE REGRESSION

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 = 0 \end{pmatrix} + \tilde{\varepsilon}$$

$= 0$



-> IF $\beta_1 = 0$ -> $\tilde{Y} = \beta_0 + \tilde{\varepsilon}$ -> $\tilde{X}$ NO INFERENCE ON Y

ex. 2 NORMAL SAMPLES

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{matrix} m_1 \\ m - m_1 \end{matrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{pmatrix}$$

-> EQUIVALENT TO COMPARING 2 NORMAL MEANS : $\beta_1 = \mu_X - \mu_Y$

-> $H_0 : \beta_1 = 0 \equiv H_0 : \mu_X - \mu_Y = 0 \equiv H_0 : \mu_X = \mu_Y$

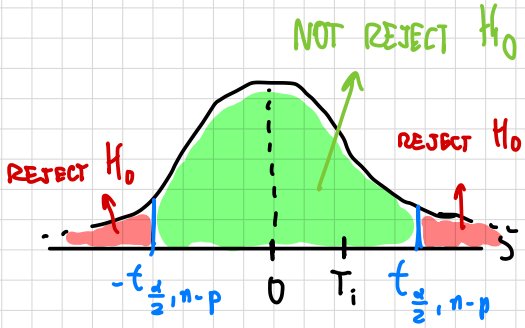
- IN GENERAL : IF $H_0 : \beta_i = 0$ IS TRUE -> X_i IS USELESS IN EXPLAINING Y

- **HOW TO REJECT / NOT REJECT H_0 ?**

- REJECT H_0 IF:

$$T_i = \left| \frac{\hat{\beta}_i - \beta_i^{0}}{SE(\hat{\beta}_i)} \right| > t_{\frac{\alpha}{2}, n-p}$$

(β_i^0: H_0)



ex. ADVERTISING DATASET

```
Residuals:
     Min       1Q   Median       3Q      Max
-15.5632  -3.5293   0.6714   4.2504   8.6796

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.2357 β0  0.6535  18.724  < 2e-16 ***
Radio        0.1244 β1  0.0237   5.251 3.88e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.963 on 198 degrees of freedom
Multiple R-squared:  0.1222,    Adjusted R-squared:  0.1178
F-statistic: 27.57 on 1 and 198 DF,  p-value: 3.883e-07
```

- FIT: Y: SALES, X: RADIO

- $Y = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{RADIO} =$

$= 12.2357 + 0.1264 \cdot \text{RADIO}$

- $T_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = 5.251$

- $SE(\hat{\beta}_1) = \sqrt{\frac{e^T e}{n-p} (X^T X)^{-1}_{i+1, i+1}} = 0.0237$

$n = 200$ ~> 200 SAMPLES

$p = 2$ ~> β_0, β_1

-> SUPPOSE WE CHOOSE $\alpha = 0.05$:

$$t_{\frac{\alpha}{2}=0.025,\ n-p=198} = 1.972 \left(\text{LARGE DATASET} \uparrow \approx z_{0.025} = 1.96 \right)$$

-> $T_1 = 5.251 > t_{\frac{\alpha}{2}, n-p} = 1.972$ -> H_0 REJECTED

- **p-VALUE:**

- ↓ p-VALUE -> WE BELIEVE LESS IN H_0: OUR OBSERVED VALUE IS EXTREME

->
- p-VALUE $\leq \alpha$ -> H_0 REJECTED
- p-VALUE $> \alpha$ -> H_0 NOT REJECTED

• ALL OR NOTHING TEST:

- $H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$

  ↳ OPPOSITE H_p : $H_1 : \exists \beta_i \neq 0$

-> H_0 REFERS TO NULL MODEL ~> $\beta_0 = k \neq 0$, $\beta_i = 0$, $i \neq 0$, $k \neq 0$

- IN THIS CASE : $E(\bar{Y}) = \beta_0$, $\hat{\beta}_0 = \bar{Y}$ -> $\begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \beta_0 + \bar{\varepsilon}$, $\bar{\varepsilon} \sim N(0, \sigma^2)$

- VECTOR OF FITTED VALUES:

$$\hat{\bar{Y}} = \begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_m \end{pmatrix} = X \cdot \hat{\beta} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \hat{\beta}_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \bar{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix}$$

-> NULL MODEL ⊂ FULL MODEL :

- FULL MODEL : β_i MAY NOT BE 0
- NULL MODEL : $\beta_1 = \dots = \beta_{p-1} = 0$, $\beta_0 = k$

- TESTING H_0 ≡ NULL MODEL IS ENOUGH TO EXPLAIN Y

- IF NULL MODEL IS ENOUGH (H_0) -> $E_0^T E_0 - E^T E \longmapsto$ SMALL NUMBER ?

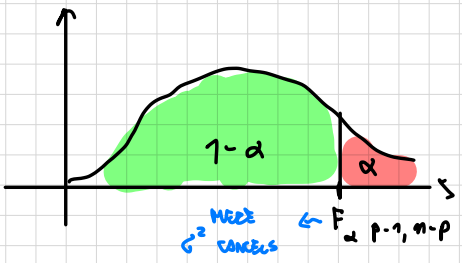
$$/ \quad E_0^T E_0 - E^T E = \| \hat{Y} - \hat{Y}_0 \|$$ EXTENSION OF PYTHAGORA'S THEOREM

-> WE REJECT H_0 IF $E_0^T E_0 - E^T E$ IS TOO LARGE ?

- **F TEST :**

  - IT CAN BE PROVEN THAT :

$$\frac{E_0^T E_0 - E^T E}{\sigma^2} \sim \chi^2_{p-1}$$



AND IF H_0 IS TRUE :

$$F_{\text{STATISTIC}} = \frac{\frac{E_0^T E_0 - E^T E}{\sigma^2 (p-1)}}{\frac{E^T E}{\sigma^2 (n-p)}} \sim F_{p-1,n-p}$$

F DISTRIBUTION

- BY PYTHAGORA'S THEOREM : $F = \frac{\frac{\|\hat{Y} - \hat{Y}_0\|^2}{p-1}}{\frac{\|Y - \hat{Y}\|^2}{n-p}} = \frac{\frac{\sum(\hat{Y}_i - \bar{Y})^2}{p-1}}{\frac{\sum(Y_i - \hat{Y})^2}{n-p}}$

  EXPLAINED VARIANCE

  UNEXPLAINED VARIANCE

- H_0 REJECTED $\Leftrightarrow F > F_{\alpha, p-1, n-p}$ (DOF$_1$, DOF$_2$, UNDER H_0)

  ↳ FROM COMPUTERS : REJECT IF p-VALUE $< \alpha$

20. ONE WAY ANOVA (ANALYSIS OF VARIANCE)

ONE WAY ANOVA : LINEAR MODEL WITH CATEGORICAL FEATURE WITH MORE THAN 2 CLASSES, ONE-HOT ENCODING USED

- USING $p = 4$ CLASSES :

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ \vdots & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

(A | B | C | D)

→ A COLUMN NEEDS TO BE ERASED TO HAVE FULL RANK ()

- INTERPRETATION :

β_0 : AVG PRICE FOR A,

β_1 : AVG DIFF. IN PRICE B - A

β_2 : AVG DIFF. IN PRICE C - A

β_3 : AVG DIFF. IN PRICE D - A

-> TESTING $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ MEANS ALL GROUPS HAVE SAME EXPECTATION

-> WE REJECT IF $F = \frac{\text{VARIANCE BETWEEN GROUPS}}{\text{VARIANCE WITHIN GROUPS}}$ IS LARGE.

- AN F TEST ON A SMALL MODEL VS LARGE MODEL CAN BE DONE IN R USING : anova(small, large) (NESTED LINEAR MODELS)

eg.

Y : FINAL LAUREA SCORE (/110)

PREDICTORS:

X_1 : MATURITÀ, X_2 : ANALYSIS I SCORE

X_3 : CHEMISTRY SCORE, X_4 : PHYSICS SCORE

->

SMALL MODEL : $m_1 \leftarrow lm(Y \sim X_1 + X_2)$

LARGE MODEL : $m_2 \leftarrow lm(Y \sim X_1 + X_2 + X_3 + X_4)$ (CONTAINS ALL PREDICTORS)

-> anova(m_1, m_2)

- IN R:

- $DOF_1 = n - p_1 - 1 = 200 - 2 - 1 = 197$
- $DOF_2 = n - p_2 - 1 = 200 - 3 - 1 = 196$
- $RSS = \sum (Y_i - \hat{Y}_i)^2 = e^T e$
- $Df = DOF_2 - DOF_1$
- SUM OF Sq = $RSS_2 - RSS_1$
- $F = \frac{\frac{\sum (\hat{Y}_i - \bar{Y})^2}{p-1}}{\frac{\sum (Y_i - \hat{Y})^2}{n-p}} = 0.0034$

```
# Fit the full model
full_model <- lm(Sales ~ TV + Radio + Newspaper, data = advertising)
# Fit the reduced model (without the Newspaper predictor)
reduced_model <- lm(Sales ~ TV + Radio, data = advertising)
# Perform the F-test
f_test <- anova(reduced_model, full_model)
# Print the F-test results
print(f_test)
```

```
Analysis of Variance Table

Model 1: Sales ~ TV + Radio
Model 2: Sales ~ TV + Radio + Newspaper
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    197 541.21
2    196 541.20  1 0.0092861 0.0034 0.9538
```

-> $P[X > 0.0034] \approx 0.95$ = p-value > α -> H_0 NOT REJ

- **CONFIDENCE AND PREDICTION INTERVALS:**

SUPPOSE $[\ X_f\ ]$: NEW COMBINATION OF PREDICTORS

eg. LINEAR REGRESSION $X_f = [1\ \ X_f^\circ]$

- FITTING VALUE OF X_f IS : $\hat{Y}_f = X_f \hat{\beta}$

WHERE: $E[\overbrace{X_f \cdot \hat{\beta}}^{\text{R.V.}}] = X_f \cdot \beta$

- C.I FOR. $\hat{Y}_f = X_f \hat{\beta}$ / $X_f \hat{\beta} = X_f (X^T X)^{-1} X^T Y$;

  - $E[\hat{Y}_f] = E[X_f (X^T X)^{-1} X^T Y] = X_f \cdot (X^T X)^{-1} X^T \cdot X\beta = X_f \beta$

  - $V[\hat{Y}_f] = X_f (X^T X)^{-1} X^T \cdot Var Cov(Y) \cdot X (X^T X)^{-1} X_f^T =$

    $= \sigma^2 \cdot X_f (X^T X)^{-1} (X^T X) (X^T X)^{-1} X_f^T$

$\rightarrow X_f \hat{\beta} \sim N(X_f \beta,\ \sigma^2 \cdot X_f (X^T X)^{-1} X_f^T)$

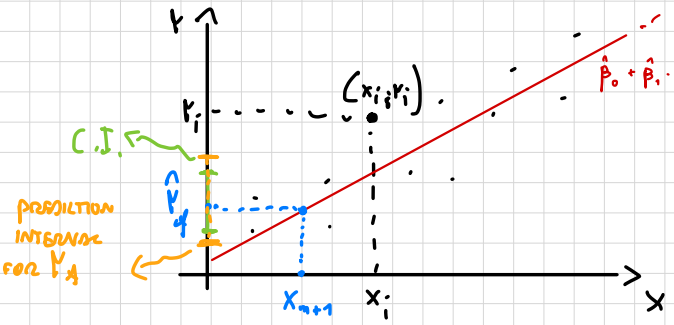
$\rightarrow$

$$C.I_{1-\alpha} : X_f \cdot \hat{\beta} \pm t_{\frac{\alpha}{2}, n-p} \cdot \sqrt{\frac{e^T e}{n-p} \cdot X_f (X^T X)^{-1} X_f^T}$$

eg. SIMPLE LINEAR REGRESSION

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\tilde{x}_{n+1} = \begin{bmatrix} 1 & x_{n+1} \end{bmatrix}$$



- WHAT WILL BE THE NEW VALUE OF Y_{n+1} CORRESPONDING TO x_{n+1} ?

→ POINT PREDICTION : $\hat{Y}_{n+1} = \tilde{x}_{n+1} \cdot \hat{\beta}$

- PREDICTION INTERVAL :

IN GENERAL $Y_{n+1} \sim N(x_{n+1}\beta, \sigma^2)$

-> PREDICTION ERROR . $Y_{n+1} - \hat{Y}_{n+1} = Y_{n+1} - \tilde{x}_{n+1}\hat{\beta} \sim N$

$$\cdot E(Y_{n+1} - \hat{Y}_{n+1}) = E(Y_{n+1}) - E(\hat{Y}_{n+1}) = \tilde{x}_{n+1}\beta - \tilde{x}_{n+1}\beta = 0$$

$$\cdot V(Y_{n+1} - \hat{Y}_{n+1}) = V(Y_{n+1}) + V(\hat{Y}_{n+1}) + 2\underbrace{Cov}^{0}(Y_{n+1}, x_{n+1}\hat{\beta}) =$$

$$= \sigma^2 + \sigma^2 x_{n+1}(X^T X)^{-1} x_{n+1}^T = \sigma^2\left(1 + x_{n+1}(X^T X)^{-1} x_{n+1}^T\right)$$

PROPERTY ↑

$$\rightarrow \frac{Y_{n+1} - x_{n+1}\hat{\beta}}{\sqrt{\frac{E^T E}{n-p}\left(1 + x_{n+1}(X^T X)^{-1} x_{n+1}^T\right)}} \sim t_{n-p}$$

-> PREDICTION INTERVAL :

$$x_{n+1}\hat{\beta} \pm t_{\frac{\alpha}{2}, n-p} \cdot \sqrt{\frac{E^T E}{n-p}\left(1 + x_{n+1}(X^T X) x_{n+1}^T\right)}$$

→ A BIT LARGER THEN C.I.

- **2 BINARY PREDICTORS:**

eg.

SEX: M → 0, F → 1 ; EDUCATION: L → 0, H → 1 ⇝ 2 LEVELS OF FACTOR "EDUCATION"

- IN R CATHEGORICAL FEATURES ARE CALLED "FACTORS", AND THEIR VALUES ARE CALLED "LEVELS"

- ~~lm(INCOME ~ SEX)~~ → NO MEANING

↳ lm(INCOME ~ AS.FACTORS(SEX)) → $\begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}$ (M/F → 0/1)

- GENERIC ADDITIVE MODEL WITH 2 BINARY FEATURES:

(2 BINARY FEATURES)

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \bar{\varepsilon}$$

→

0 0 $\mu_{M,L} = \beta_0$: MEAN RESPONSE IF SEX = M, EDUC = L

1 0 $\mu_{F,L} = \beta_0 + \beta_1$: MEAN RESPONSE IF SEX = F, EDUC = L

0 1 $\mu_{M,H} = \beta_0 + \beta_2$: MEAN RESPONSE IF SEX = M, EDUC = H

1 1 $\mu_{F,H} = \beta_0 + \beta_1 + \beta_2$: MEAN RESPONSE IF SEX = F, EDUC = H

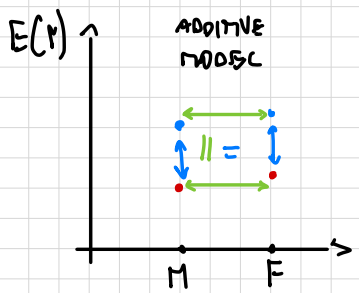
- IN R, lm($Y \sim \overset{\text{SEX}}{X_1} + \overset{\text{EDUC}}{X_2}$) : ADDITIVE MODEL (NO INTERACTION)

↳ NO INTERACTION: DIFFERENCE BETWEEN 2 LEVELS IT'S THE SAME, NO MATTER WHAT OTHER FACTOR IS

→

$\mu_{F,L} - \mu_{M,L} = \mu_{F,H} - \mu_{M,H} = \beta_1$

$\mu_{F,H} - \mu_{F,L} = \mu_{M,H} - \mu_{M,L} = \beta_2$

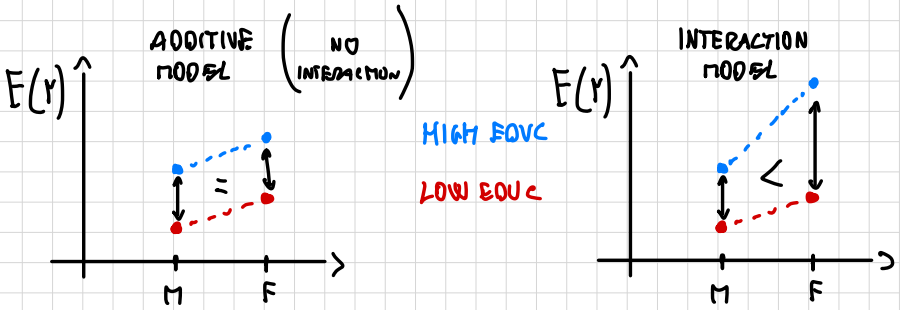


- **GENERAL INTERACTION MODEL:**

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

INTERACTION MODEL

- GRAPHICALLY:



es.

SEX: M / F , EDUCATION: L / H

→

$$\mu_{M,L} = \beta_0 = E(Y) \text{ IF M AND L}$$

$$\mu_{F,L} = \beta_0 + \beta_1 = E(Y) \text{ IF F AND L}$$

$$\mu_{M,H} = \beta_0 + \beta_2 = E(Y) \text{ IF M AND H}$$

$$\mu_{F,H} = \beta_0 + \beta_1 + \beta_2 + \beta_3 = E(Y) \text{ IF F AND H}$$

- IN THIS CASE:

$$\mu_{F,L} - \mu_{M,L} = \beta_1 \neq \mu_{F,H} - \mu_{M,H} = \beta_1 + \beta_3$$

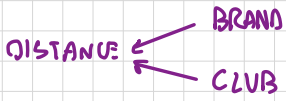
(≠ → DIFFERENT FROM ADDITIVE MODEL)

↳ WHEN INTERACTION IS PRESENT, EACH COMBINATION OF FACTOR 1 AND A LEVEL OF FACTOR 2 HAS NO (INTERACTION)

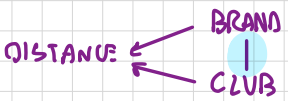
ex. GOLF. R IN VS CODE

IN R:

ADDITIVE MODEL (NO INTERACTION)



INTERACTION MODEL



- ANOVA (SMALL, LARGE) GIVES A TEST FOR NULL H_p, H_0: ALL INTERACTIONS = 0

-> IN THIS EXAMPLE, p_VALUE = 0.00179 -> SMALL -> H_0 REJECTED

L> $\exists$ EVIDENCE THAT SOME INTERACTIONS $\neq 0$

- DIAGNOSTIC PLOTS:

  - THEY ARE USED TO BE SURE WE ARE USING THE "RIGHT" MODEL

  -> ONE WAY IS TO LOOK AT RESIDUAL PLOTS: $E_i = Y_i - \hat{Y}_i$

  L> E_i SHOULD BEHAVE AS NOISE (IF RIGHT MODEL), SIMILAR TO ε

  -> ALSO A PLOT E_i VS NEW FEATURE W_i CAN TELL IF W_i SHOULD BE INCLUDED IN MODEL: IF PATTERNS ARE EVIDENT (ex. LINEARITY)

  -> W_i SHOULD BE INCLUDED IN THE MODEL

- **MEASURES OF FIT:**

GIVEN $\bar{Y}$: VECTOR OF TRUE VALUES, $\hat{Y}_0$: VECTOR FROM NULL MODEL ($\beta_i = 0,\ \beta_0 \neq 0$)

$\hat{Y}$: FITTED VALUES BY THE MODEL

→ THE FURTHER AWAY IS $\hat{Y}$ FROM $\hat{Y}_0$ → BETTER EXPLANATION OF DATA BY THE MODEL

- R^2:

$$R^2 = \frac{NUM}{DEN} = \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\|Y - \hat{Y}_0\|^2} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

(PRED. $\hat{Y}_i$; $\bar{Y}$: MEAN OF DATA)

PITHAGORA EXTENDED TO $\mathbb{R}^n$

- NUM: $\|\hat{Y} - \hat{Y}_0\|^2 = \|Y - \hat{Y}_0\|^2 - \|Y - \hat{Y}\|^2 = TSS - E^T E$

- DEN: $\|Y - \hat{Y}_0\|^2 = E_0^T E_0 = TSS$

TOTAL SUM OF SQUARES

RSS: $E^T E$ — RESIDUAL SUM OF SQUARES

- $0 \leq R^2 \leq 1$

- PROBLEM: IF A PREDICTOR IS ADDED → $\uparrow R^2$, SHOULDN'T HAPPEN
  - SOL: R^2_{ADJ}

- R^2_{ADJ}:

$$R^2_{ADJ} = 1 - \frac{n-1}{n-p} \cdot \frac{RSS}{TSS}$$

n: # SAMPLES; p: # PREDICTORS

- AKAIKE INFORMATION CRITERION:

IT'S A GENERALIZATION OF R^2 AND R^2_{ADJ} TO NON-LINEAR MODELS (BUT ALSO FOR LINEAR WORKS)

$$AIC = 2k - 2 \cdot \ln(\hat{L})$$

k: # ESTIMATED PARAMETER IN THE MODEL (→ $p+1$ IN LINEAR MODELS $(\beta_0, \dots, \beta_{p-1}, \sigma^2)$)

$\hat{L}$: MAXIMUM VALUE OF LIKELIHOOD FUNCTION

- BAYESIAN INFORMATION CRITERION:

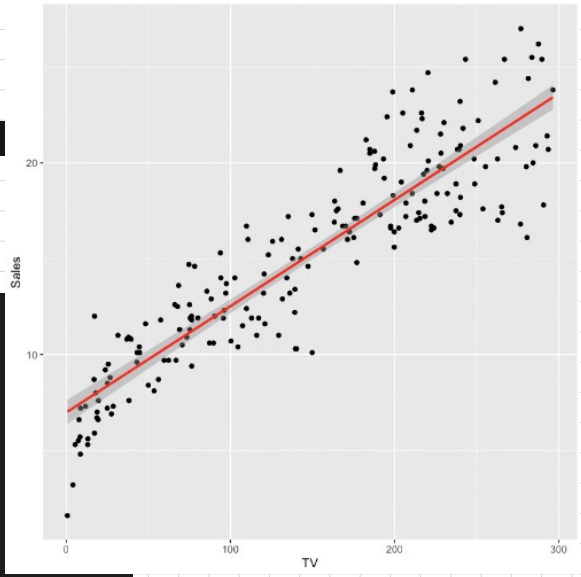
$$BIC = k \cdot \ln(n) - 2 \ln(\hat{L})$$

n: # DATA POINTS

→ ↓ AIC, BIC → BETTER MODEL ~> UNLIKE R^2, R^2_{ADJ}

R ANALYSIS :

```
> advertising <- read_csv("./datasets/advertising.csv")
> simple_reg <- lm(Sales ~ TV, data = advertising)
> summary(simple_reg)
```




```
Residuals:
    Min      1Q  Median      3Q     Max 
-6.4438 -1.4857  0.0218  1.5042  5.6932 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.974821   0.322553   21.62   <2e-16 ***
TV          0.055465   0.001896   29.26   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.296 on 198 degrees of freedom
Multiple R-squared:  0.8122,    Adjusted R-squared:  0.8112 
F-statistic: 856.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

1 2 3

- SIMPLE REG: $\underset{n\times 1}{\bar{Y}} = X \underset{2\times 1}{\bar{\beta}} + \underset{n\times 1}{\bar{\varepsilon}}$ / $\bar{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, $\underset{n\times 2}{X} = \begin{pmatrix} 1 & x_{1,1} \\ \vdots & \vdots \\ 1 & x_{1,n} \end{pmatrix}$ (TV)

- $\hat{\beta}_0 = 6.974827$ → INTERCEPT OF LINEAR REGRESSION
- $\hat{\beta}_1 = 0.055465$ → COEFFICIENT RELATED TO "TV" PREDICTOR, x_1

→ $Y_i = \begin{pmatrix} 1 & x_{1,i} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \varepsilon \overset{\varepsilon \sim N(0, \sigma^2)}{=} \beta_0 + \beta_1 x_{1,i} + \varepsilon$ (TV predictor)

- RESIDUALS :
  - RESIDUAL ESTIMATOR : $Y - X\hat{\beta}$
    - → TO COMPUTE THEM : $e_i = Y - \underset{2\times 1}{X_i}\hat{\beta}$, $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$

```
library(tibble)
y <- advertising$Sales
x <- cbind(1, advertising$TV)
e <- y - x %*% simple_reg$coefficients
head(tibble(
  lm_res = simple_reg$residuals,
  manual_res = as.vector(e)
))
```

→ ACTUALLY, $\exists$ 200 RESIDUALS

A tibble: 6 x 2

lm_res	manual_res
<dbl>	<dbl>
2.3627348	2.3627348
0.9569962	0.9569962
4.0711845	4.0711845
1.1222658	1.1222658
0.8971480	0.8971480
-0.2573650	-0.2573650

- 3 :

$\min\{e_i\} = -6.4438$, $\max\{e_i\} = 5.6932$

- $\begin{cases} 1Q = -1.4857 : & P[e_i < -1.4857] = 25\% \\ 3Q = 1.5042 : & P[e_i < 1.5042] = 75\% \end{cases}$

MEDIAN = 0.0218 : $P[e_i < 0.0218] = 50\%$

- **COEFFICIENTS :**

- $\hat{\beta}_0 \overset{1}{=} 6.974827$, $\hat{\beta}_1 \overset{2}{=} 0.055465$

```
Residuals:
    Min      1Q  Median      3Q     Max 
-6.4438 -1.4857  0.0218  1.5042  5.6932 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.974821   0.322553  21.62   <2e-16 ***
TV          0.055465   0.001896  29.26   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.296 on 198 degrees of freedom
Multiple R-squared:  0.8122,    Adjusted R-squared:  0.8112 
F-statistic: 856.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

3

```
beta_hat <- solve(t(x) %*% x) %*% t(x) %*% y
head(tibble(
  lm_coeff = simple_reg$coefficients,
  manual_coeff = as.vector(beta_hat)
))
```

lm_coeff	manual_coeff
<dbl>	<dbl>
6.97482149	6.97482149
0.05546477	0.05546477

- **COEFFICIENTS ESTIMATOR :** $\hat{\beta} = (X^T X)^{-1} X^T Y$

- $\forall \hat{\beta}_i$: $SE_i = \sqrt{\hat{\sigma}_m^2 (X^T X)^{-1}_{i+1, i+1}}$

```
n <- nrow(x)
p <- ncol(x)
rms <- t(e) %*% e / (n - p)

# SE for TV
tv_se <- sqrt(rms * solve(t(x) %*% x)[2, 2])
intercept_se <- sqrt(rms * solve(t(x) %*% x)[1, 1])
tv_se
intercept_se
```

5

A matrix: 1 x 1 of type dbl
0.001895551

4

A matrix: 1 x 1 of type dbl
0.3225535

→ $\downarrow SE_i \rightarrow \uparrow \hat{\beta}_i$ ACCURACY

- $\hat{\sigma}_m^2 = \frac{e^T e}{n-p} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-p} = RMS = \text{RESIDUAL MEAN SQUARES} = S^2$

- $SE_0 \overset{4}{=} 0.322553$

- $SE_1 \overset{5}{=} 0.001896$ an. → $SE_1 = \sqrt{\frac{e^T e}{n-p} \cdot \sqrt{\left(\begin{bmatrix} 1 & \cdots & 1 \\ x_{1,1} & \cdots & x_{1,m} \end{bmatrix} \begin{bmatrix} 1 & x_{1,1} \\ \vdots & \vdots \\ 1 & x_{1,1} \end{bmatrix}^{-1} \right)_{2,2}}}$

- **T - VALUE :**

GIVEN $H_0 : \hat{\beta}_i^0 = 0$:

- $T_i = \left| \frac{\hat{\beta}_i - \hat{\beta}_i^0}{SE(\hat{\beta}_i)} \right| \rightarrow \begin{cases} T_0 \overset{6}{=} \frac{\hat{\beta}_0 - 0}{SE(\hat{\beta}_0)} = \frac{①}{④} = 21.62 \\ T_1 = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{②}{⑤} = 29.26 \end{cases}$

```
# for TV
t_val <- simple_reg$coefficients[2] / tv_se
t_val
```

A matrix: 1 x 1 of type dbl
6 29.2605

→ IF $T_i > t_{\frac{\alpha}{2}, n-p}$ (200 - 2) → H_0 REJECTED

- **p - VALUE :**

- FOR 2 - SIDED TEST :

$p\text{-VALUE}_i = 2 \cdot P[\hat{\beta}_i \geq T_i] \rightarrow \begin{cases} \text{IF } p\text{-value}_i < \alpha \rightarrow H_0 \text{ REJECT} \\ \text{IF } p\text{-value}_i > \alpha \rightarrow H_0 \text{ NOT REJECT} \end{cases}$

→ $p\text{-VALUE}_1 \overset{8}{=} 7.9279 \cdot 10^{-74} < 2.2 \cdot 10^{-16}$ → H_0 REJECT

```
# multiply by two because the alternative hyp is two-sided
p_val <- 2 * pt(t_val, n - 2, lower.tail = FALSE)
p_val
```

8

A matrix: 1 x 1 of type dbl
7.927912e-74

- **SIGNIF. CODES:**

THEY INDICATES THE STRONGNESS TO REJECT H_0, BASED ON p-VALUE; :

- '***' : STRONGLY REJECT H_0
- ⋮
- ' ' (NULL) : NOT REJECT H_0

↓p-value

```
Residuals:
    Min      1Q  Median      3Q     Max 
-6.4438 -1.4857  0.0218  1.5042  5.6932 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.974821   0.322553  21.62   <2e-16 ***
TV          0.055465   0.001896  29.26   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.296 on 198 degrees of freedom
Multiple R-squared:  0.8122,    Adjusted R-squared:  0.8112 
F-statistic: 856.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

STRONGNESS

- **LOGISTIC REGRESSION AS GLM**: ~> GENERALIZED LINEAR MODEL

SO FAR WE ASSUMED: $\bar{Y} \sim N(X\bar{\beta}, \sigma^2 I)$

, BUT SOME ASSUMPTIONS MAY NOT BE REALISTIC:

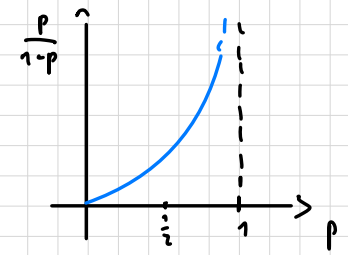
1. IF $\sigma^2 \neq$ CONST. $\longrightarrow$ TRY $\log(\bar{Y})$ OR OTHER TRANSFORMATIONS
2. IF RESIDUALS DO NOT LOOK $\sim N(\dots,\dots)$ ~> ∃ PATTERNS $\longrightarrow$ TRY SIMPLE TRANSFORMATIONS OR ENRICH MODEL WITH NEW PREDICTORS $(\log(\bar{Y}), \log(K), K^2, \dots)$
3. IF OBSERVATIONS CLUSTER INTO 2 GROUPS $\longrightarrow$ INTRODUCE GROUP LABEL AS A PREDICTOR
4. RESIDUALS OR Y_i's LOOK CORRELATED $\longrightarrow$ USE SERIES MODEL (eg. MARKOVIAN) — NEXT COURSE
5. Y CANNOT BE ASSUMED TO BE NORMAL $\longrightarrow$ GLMs

(5) ↑

- ASSUME: $Y_i \sim \text{Bernoulli}(p_i)$, $Y_1, \dots, Y_n$ INDEPENDENT

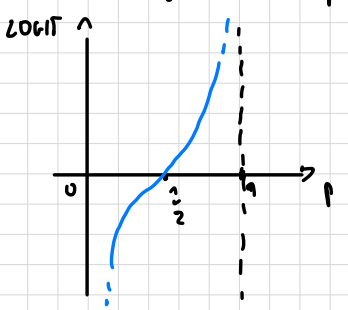
$$\rightarrow \underset{n\times p}{X} \cdot \underset{p\times 1}{\beta} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad E(\bar{Y}) = \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}$$

- PROBLEM: $p_i \in [0,1]$ -> WE SELDOM CAN USE $\bar{p} = X\bar{\beta}$
  - SOL: IDEA OF GLM's -> TRANSFORM $p: [0,1] \longmapsto [-\infty, +\infty]$ — LINK FUNCTION
    - ODDS:
      - $\frac{p}{1-p}: [0,1] \longmapsto [0, +\infty]$



ex. $p = 0.05$ -> ODDS: $\frac{1}{20}$: "ONE TO TWENTY"

    - LOGIT (LOG ODDS):
      - $\log\left(\frac{p}{1-p}\right): [0,1] \longmapsto [-\infty, +\infty]$



- GLM FOR p:

$$\begin{pmatrix} \text{logit}(p_1) \\ \vdots \\ \text{logit}(p_n) \end{pmatrix} = X \cdot \tilde{\beta}$$

~> NO "SIGNAL + ERROR" STRUCTURE

$\because Y_i \sim \text{Bernulli}(p_i)$

- $\text{logit}(p) \longmapsto p$: $[-\infty; +\infty] \longmapsto [0, 1]$

$$l = \log\left(\frac{p}{1-p}\right) \rightarrow e^l = \frac{p}{1-p} \rightarrow (1-p)e^l = p$$

$$\rightarrow p = \frac{e^l}{1+e^l}$$ INVERSE LOGIT TRANSFORM.

$$\rightarrow p_i = \frac{e^{X\beta_{(i)}}}{1+e^{X\beta_{(i)}}}$$

- IN R: https://stats.oarc.ucla.edu/r/dae/logit-regression/

```
my_fit <- glm(y~predictors, family=binomial)
```

Family	Default link function
binomial	(link = "logit") → logistic regression
gaussian	(link = "identity") → LINEAR MODELS
gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log") → for count variables, like number of accidents
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

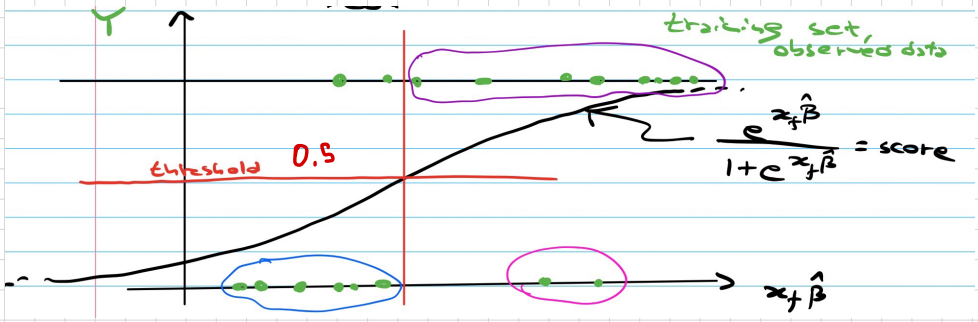
- METHODS FROM lm() CLASS: ~> es. SUMMARY (my_fit)

Function	Description
summary()	Displays detailed results for the fitted model
coefficients(), coef()	Lists the model parameters (intercept and slopes) for the fitted model
confint()	Provides confidence intervals for the model parameters (95 percent by default)
residuals()	Lists the residual values for a fitted model
anova()	Generates an ANOVA table comparing two fitted models
plot()	Generates diagnostic plots for evaluating the fit of a model
predict()	Uses a fitted model to predict response values for a new dataset

- $\hat{p}_f = \frac{e^{\bar{x}_f \cdot \hat{\beta}}}{1 + e^{\bar{x}_f \cdot \hat{\beta}}}$: ESTIMATES PROBABILITY OF $Y_f = 1$ CORRESPONDING TO A NEW CONFIGURATION OF PREDICTORS $\bar{x}_f$

→ IN ORDER TO ASSIGN CLASS $0/1$ TO $\bar{x}_f$ WE CAN SET A THRESHOLD : $\hat{p}_f > \text{THRESHOLD}$

ex. t = THRESHOLD $= 0.5$



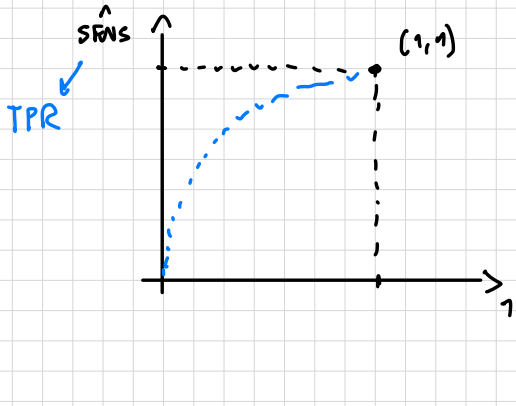
- CONFUSION MATRIX :

TRUE \ PREDICTED	1	0
1	TP	FN
0	FP	TN

- SENSITIVITY = RECALL $(c=1)$ = $P(\hat{Y}=1 \mid Y=1)$
  - ↳ $\hat{\text{SENSITIVITY}} = \frac{TP}{TP+FN} = TPR$
- SPECIFICITY = RECALL $(c=0)$ = $P(\hat{Y}=0 \mid Y=0)$
  - ↳ $\hat{\text{SPECIFICITY}} = \frac{TN}{TN+FP} = 1 - FPR$

- IF WE VARY t → CHANGE OF SENSITIVITY AND SPECIFICITY

- ROC CURVE :



ROC CURVE :

$\hat{SENS} = f(1 - \hat{SPEC})$

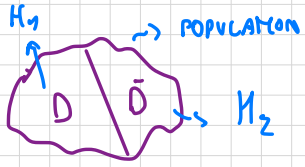
- **INTRODUCTION TO BAYESIAN STATISTICS:**

- BAYES THEOREM: $$P(B_n \mid A) = \frac{P(B_n)\,P(A \mid B_n)}{\sum_i P(B_i)\,P(A \mid B_i)}$$

$$\rightarrow P[H_p : H_k \text{, GIVEN EVIDENCE } E] : P(H_k \mid E) = \frac{P(H_n)\,P(E \mid H_k)}{\sum_i P(H_i)\,P(E \mid H_i)}$$

ex. DIAGNOSTIC TEST

SUPPOSE 2 H_ps: $\begin{cases} D: \text{DISEASED} \\ \bar{D}: \text{NOT DISEASED} \end{cases}$



- EVIDENCE WILL BE A TEST: $\begin{cases} T+: \text{SAYS YOU HAVE } D \\ T-: \text{SAYS YOU HAVE NOT } D \end{cases}$

→ PROBS = 1, GT = 1

- $P(T+ \mid D)$ = SENSITIVITY = TPR
- $P(T+ \mid \bar{D})$ = 1 - SPECIFICITY = $1 - P(T- \mid \bar{D})$ = FPR

$$\rightarrow P(D \mid T+) = \frac{P(D) \cdot P(T+ \mid D)}{P(T+)} = \frac{P(D) \cdot P(T+ \mid D)}{P(D)\,P(T+ \mid D) + P(\bar{D})\,P(T+ \mid \bar{D})}$$

- **BAYES THEOREM FOR R.V.:**

LET X, Y BE RANDOM VECTORS (~> R.VEC.) WITH JOINT DENSITY $f(x,y)$

- CONDITIONAL DENSITY OF X GIVEN $Y=y$:

$$f_{X|Y}(X=x \mid Y=y) := f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

$f_Y(y)$ -> MARGINAL DENSITY OF Y; $f_Y(y) = \int f(x,y)\,dx$

- CONDITIONAL DENSITY OF Y GIVEN $X=x$:

$$f_{Y|X}(Y=y \mid X=x) := f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

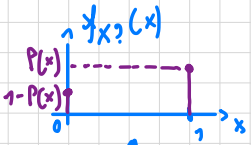
$f_X(x)$ -> MARGINAL DENSITY OF X; $f_X(x) = \int f(x,y)\,dy$

- APPLYING BAYES THEOREM: (IF DISCRETE: $\int \dots dt \longmapsto \sum_i$)

$$f_{X|Y=y}(x|y) \overset{\text{LAW OF TOTAL PROB.}}{:=} \frac{f_X(x)\cdot f_{Y|X=x}(y|x)}{f_Y(y)} \overset{\sum}{=} \frac{f_X(x)\cdot f_{Y|X=x}(y|x)}{\int f_X(t)\cdot f_{Y|X=t}(y|t)\,dt}$$

$f_Y(y) = \int f(x,y)\,dx$

- **INDICATOR OF EVENTS:**



X: EVENT -> $X? = \begin{cases} 1, & \text{IF } X \text{ IS TRUE} \\ 0, & \text{IF } X \text{ IS FALSE} \end{cases}$, $f_{X?}(x) = \begin{cases} P(X), & \text{IF } x=1 \\ 1-P(X), & \text{IF } x=0 \end{cases}$

ex. DIAGNOSTIC EXAMPLE, SUPPOSE: SENS = 0.9, SPEC = 0.95, PREVALENCE = $P(D) = 0.01$

$$f_{T?}(x) = \begin{cases} 1, & \text{IF } T+ \\ 0, & \text{IF } T- \end{cases}$$

- JOINT DENSITY OF $T?$ AND $D?$ (0/1): $f(x,y) := f_{T?|D?}(y|x)\cdot f_{D?}(x)$
- EXAMPLE: $f(\underset{T}{0}, \underset{D}{1}) = f_{D?}(1)\cdot f_{T?|D?}(0|1) = 0.01 \cdot \underbrace{(1-\text{SENS})}_{0.1} = 0.001$

ex. 2a

$P(D? = 1 \mid T? = 1)$:

P OF HAVING DISEASE GIVEN A POSITIVE RESULT ON A TEST

2. Consider a binary screening situation formulated in terms of the following random variables:

- $D?$ = indicator whether a randomly sampled person is diseased
- $T?$ = indicator whether the person has a single positive test
- $T_1?, T_2?$ = similar indicators for two conditionally independent tests

and compute the following probabilities.

(a) Consider a single test. Assuming $P(D? = 1) = 0.01$ (prevalence), $P(T? = 1|D? = 1) = 0.90$ (sensitivity) and $P(T? = 0|D? = 0) = 0.95$ (specificity), compute

$P(D? = 1|T? = 1)$ (positive predictive value)

PREV = 0.01

SENS = 0.9

SPEC = 0.95

$$\rightsquigarrow P(D? = 1 \mid T? = 1) = \frac{P(D? = 1, T? = 1)}{P(T? = 1)}$$

- $P(D? = 1, T? = 1) = \overbrace{P(T? = 1 \mid D? = 1)}^{\text{SENSITIVITY}} \overbrace{P(D? = 1)}^{\text{PREVALENCE}} = 0.90 \cdot 0.01 = 9 \cdot 10^{-3}$

- $P(T? = 1) = \overbrace{P(T? = 1 \mid D? = 0)}^{1-\text{SPECIFICITY}} \overbrace{P(D? = 0)}^{1-\text{PREVALENCE}} + \overbrace{P(T? = 1 \mid D? = 1)}^{\text{SENSITIVITY}} \overbrace{P(D? = 1)}^{\text{PREVALENCE}} =$

$= (1 - 0.95)(1 - 0.01) + 0.9 \cdot 0.01 = 0.0585$

$$\rightarrow P(D? = 1 \mid T? = 1) = \frac{9 \cdot 10^{-3}}{0.0585} = 0.1538 \rightsquigarrow \text{RELATIVELY LOW}$$

ex. 2b

(b) Consider repeating two conditionally independent test on the same person. Assume sensitivity and specificity of $T_1?$ and $T_2?$ are the same as $T?$. Compute

$P(D? = 1|T_1? = 1, T_2? = 1)$

$$\rightsquigarrow P(D? = 1 \mid T_1? = 1, T_2? = 1) = \frac{P(D? = 1, T_1? = 1, T_2? = 1)}{P(T_1? = 1, T_2? = 1)}$$

- $P(D? = 1, T_1? = 1, T_2? = 1) = \overbrace{P(T_1? = 1, T_2? = 1 \mid D? = 1)}^{\text{INDIP.}} P(D? = 1) =$

$= P(T_1? = 1 \mid D? = 1) P(T_2? = 1 \mid D? = 1) P(D? = 1) =$

$= 0.9 \cdot 0.9 \cdot 0.01$

- $P(T_1? = 1, T_2? = 1) = P(T_1? = 1, T_2? = 1 \mid D? = 1) P(D? = 1) +$

$+ P(T_1? = 1, T_2? = 1 \mid D? = 0) P(D? = 0) = \ldots =$ (INDIP. AS BEFORE)

$= 0.9 \cdot 0.9 \cdot 0.1 + (1 - 0.95)(1 - 0.95) \cdot (1 - 0.01)$

$$\rightarrow P(D? = 1 \mid T_1? = 1, T_2? = 1) = \frac{0.9 \cdot 0.9 \cdot 0.01}{0.9^2 \cdot 0.1 + 0.05^2 \cdot 0.99} = 0.766$$

- **BAYESIAN UPDATE :**

IT IS POSSIBLE TO UPDATE $P(D)$ GIVEN NEW EVIDENCE OF $T?$:

→ SAMPLE A RANDOM PERSON → (BAYESIAN UPDATE) → TEST IS POSITIVE → (BAYESIAN UPDATE) → AN OTHER TEST IS POSITIVE

$P(D?=1) = 0.01$ $\quad P(D?=1 \mid T?=1) = 0.154$ $\quad P(D?=1 \mid T_1?=1, T_2?=1) = 0.766$

es. 3

3. The following example is adapted from the *Encyclopedia of Biostatistics*, Wiley (2005). Suppose 100 non-diseased (D−) patients and 50 diseased (D+) patients have also been classified on the basis of radiological exams over five ordered levels

−− = very mild
− = mild
+− = neutral
+ = serious
++ = very serious

Here are the results:

	t_0 −−	t_1 −	t_2 +−	t_3 +	t_4 ++	total
D−	35	45	10	5	5	100
D+	5	10	10	10	15	50

t_x: $\tilde{D}$ | D ; 0 | 1

Using the only four possible cut-off choices and the default points (0,0) and (1,1), draw an empirical ROC curve.

- $(1,1)$: ALWAYS SAYS "PERSON HAS DISEASE"
- $(0,0)$: NEVER SAYS "PERSON HAS DISEASE"

→

TPR : SENSITIVITY

FPR : 1 − SPECIFICITY

- t_1 → $P1 = (0.05, 0.9)$:

$$TPR = \frac{TP}{TP + FN} = \frac{45}{45 + 5} = 0.9$$

$$FPR = \frac{FP}{FP + TN} = \frac{65}{65 + 35} = 0.65$$

- t_2 → $P2 = (0.2, 0.7)$:

$$TPR = \frac{35}{35 + 15} = 0.7$$

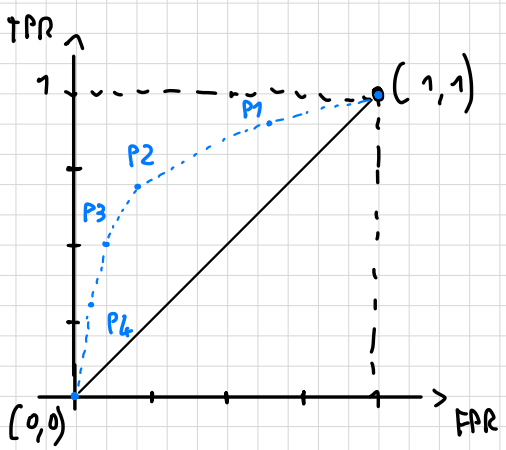
$$FPR = \frac{20}{20 + 80} = 0.2$$

⋮ → $P3 = (0.1, 0.5)$

- t_4 → $P4 = (0.05, 0.3)$:

$$TPR = \frac{15}{50} = 0.3$$

$$FPR = \frac{5}{100} = 0.05$$



· MULTIVARIATE DISTRIBUTIONS AND DAGs :

CONSIDER A R.VEC $\in \mathbb{R}^d$ $(x_1, \dots, x_d)$, WITH JOINT DENSITY $f(x_1, \dots, x_d)$

USING THE DEFINITION OF CONDITIONAL DENSITY:

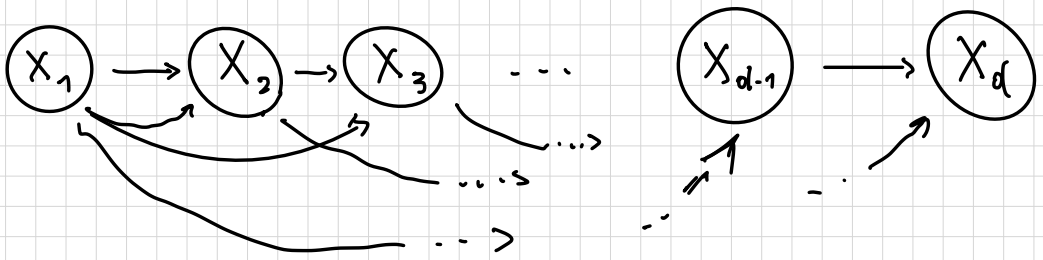
$$\cdot\ f(x_1, \dots, x_d) =$$

$$= f_{x_1, \dots, x_{d-1}}(x_1, \dots, x_{d-1}) \cdot f_{x_d \mid x_1, \dots, x_{d-1}}(x_d \mid x_1, \dots, x_{d-1}) =$$

$$= f_{x_1, \dots, x_{d-2}}(x_1, \dots, x_{d-2}) \cdot f_{x_{d-2} \mid x_1, \dots, x_{d-1}}(x_{d-2} \mid x_1, \dots, x_{d-1}) \cdot f_{x_{d-1} \mid x_1, \dots, x_d}(x_{d-1} \mid x_1, \dots, x_d)$$

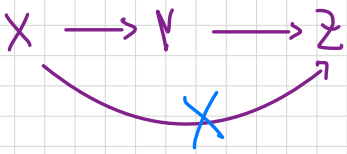
$$= \dots = f_{x_1}(x_1)\, f_{x_2 \mid x_1}(x_2 \mid x_1) \cdot f_{x_3 \mid x_2, x_1}(x_3 \mid x_2, x_1) \cdot \dots \cdot f_{x_d \mid x_1, \dots, x_{d-1}}(x_d \mid x_1, \dots, x_{d-1})$$

-> DAG:



· FOR SOME DISTRIBUTIONS -> ∄ SOME ARROWS -> JOINT DENSITY SIMPLIFIES

es.



$\rightarrow f(x, y, z) = f_X(x) \cdot f_{Y \mid X}(y \mid x) \cdot f_{Z \mid Y, X}(z \mid y, x)$

· NOW SUPPOSE $f_{Z \mid X, Y}(z \mid y, x) \overset{*}{=} f_{Z \mid Y}(z \mid y)$ $\rightsquigarrow$ DO NOT DEPEND ON X

· $f_{X, Y \mid Z}(x, y \mid z) = f_{X \mid Y}(x \mid y)\, f_{Z \mid X, Y}(z \mid x, y) \overset{*}{=} f_{X \mid Y}(x \mid y) \cdot f_{Z \mid Y}(z \mid y)$

$\rightarrow f(x, y, z) = f_X(x) \cdot f_{X \mid Y}(x \mid y) \cdot f_{Z \mid Y}(z \mid y)$

- MARKOVIAN CHAINS :

$X_1 \to X_2 \to \dots \to X_d$:

$$f(x_1, \dots, x_d) = f_{X_1}(x_1) \cdot f_{X_2|X_1}(x_2|x_1) \cdot f_{X_3|X_2}(x_3|x_2) \cdot \dots \cdot f_{X_d|X_{d-1}}(x_d|x_{d-1})$$

- INDEPENDENT RVs:

X_1 X_2 X_d : $f(x_1, \dots, x_d) = \prod_i f_{X_i}(x_i)$

- FACTORIZATION THEOREM FOR DAGs :

$$f(x_1, \dots, x_d) = \prod_i f_{X_i|\text{PARENTS}(X_i)}(x_i \mid \text{PARENTS}(x_i))$$

es.

$\forall i$: $X \to Y \leftarrow W$, Z $\quad Y = \text{PARENTS}(W, Z)$ $\quad X = \text{PARENTS}(Y)$

$f_{\dots}(\cdot) = 0$ $\to$ $f(x, y, w, z) = f_{Z|Y}(z|y) \cdot f_{W|Y}(w|y) \cdot f_{Y|X}(y|x) \cdot f_X(x)$

es. MARKOVIAN CHAIN : $X_1 \to X_2 \to \dots \to X_d$

$\to$ $f(x_1, \dots, x_d) = f_{X_1}(x_1) \cdot f_{X_2|X_1}(x_2|x_1) \cdot \dots \cdot f_{X_d|X_{d-1}}(x_d|x_{d-1})$

DAG :

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.

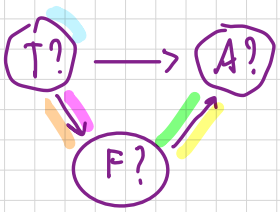
- $T? = \begin{cases} 1, & \text{CARD HOLDER IS TRAVELING} \\ 0, & \text{" IS NOT "} \end{cases}$

- $F? = \begin{cases} 1, & \text{TRANSACTION IS FRAUDULENT} \\ 0, & \text{" IS NOT "} \end{cases}$

- $A? = \begin{cases} 1, & \text{PURCHASE IS ABROAD} \\ 0, & \text{" IS NOT "} \end{cases}$

->

DAG:



BINARY BAYESIAN NETWORK.

- $f(x, y, z) = f_{T?}(x) \cdot f_{F?|T?}(y|x) \cdot f_{A?|T?,F?}(z|y,x)$ (T? F? A?)

- T? :

	T? = 0	T? = 1
p :	0.95	0.05

- F? | T? :

	T? = 0	T? = 1
F? = 0	0.99	0.98
F? = 1	0.01	0.02

- A? | T?, F?

	T? = 0		T? = 1	
	F? = 0	F? = 1	F? = 0	F? = 1
A? = 0	0.99	0.3	0.1	0.1
A? = 1	0.01	0.7	0.9	0.9

-> $f(0,0,0) = 0.95 \cdot 0.99 \cdot 0.99$

- **CONDITIONAL EXPECTATIONS:**

GIVEN A PAIR OF R.V. X, Y:

$$E[Y|X=x] = \begin{cases} \sum_y y\, f_{Y|X}(y|x) & \text{DISCRETE} \\ \int_y y\, f_{Y|X}(y|x)\, dy & \text{CONTINUOUS} \end{cases} \Big/ \quad f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

ex. BIVARIATE NORMAL

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right) \Big/ \rho = \frac{Cov(X,Y)}{\sigma_x\,\sigma_y}$$

$$\rightarrow E[Y|X=x] = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x) \quad *$$

↳ IT'S A NUMBER NOT A R.V

↳ LINEAR FUNCTION OF x

ex. X = HEIGHT OF A FATHER, Y = HEIGHT OF HIS SON

SUPPOSE $x = 1.95$, $\mu_x = 1.75 = \mu_y$, $\rho = 0.9$, $\sigma_x = \sigma_y = 0.1$

$\rightarrow$

$$E[Y|X=1.95] = 1.75 + 0.9\frac{0.1}{0.1}(1.95 - 1.75) = 1.93$$

↳ NOT AS TALL AS HIS FATHER

↓

REGRESSION TO THE MEAN EFFECT (GALTON, 1885)

* $\rightarrow$

$$\frac{E[Y|X=x] - \mu_y}{\sigma_y} = \rho\frac{x-\mu_x}{\sigma_x}$$

HOW MANY STD THE SON IS DEVIATING FROM MEAN

HOW MANY STD THE FATHER IS DEVIATING FROM MEAN

- **PROPERTIES:**

1. $E[E[Y|X]] = E[Y]$

MAIN PROPERTY

PROOF:

$$E\left[\int_Y y\, f_{Y|X}(y|x)\, dy\right] = \int_X \int_Y y\, f_{Y|X}(y|x)\, f_X(x)\, dy\, dx =$$

$$= \int_X \int_Y y \cdot \frac{f(x,y)}{f_X(x)}\, f_X(x)\, dx = \int_Y y \cdot \overbrace{\int_X f(x,y)\, dx}^{f_Y(y)}\, dy = E[Y]$$

ex. BIVARIATE NORMAL

$$E[E[Y|X]] = E\left[\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(X - \mu_X)\right] =$$

$$= \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} \underbrace{E[X - \mu_X]}_{=0} = \mu_Y = E[Y]$$

2. $E(a|X) = a$ / a = CONST.
3. $E(aX + bY|Z) = a E(X|Z) + b E(Y|Z)$
4. $E(g(X) \cdot Y|X) = g(X) \cdot E(Y|X)$

- **CONDITIONAL VARIANCE:**

$$V(Y|X) = E\left[(Y - E(Y|X))^2 \,|\, X\right]$$

5. $E[V(Y|X)] = V(Y) - V(E(Y|X))$

↳ $V(Y) = E[V(Y|X)] + V(E(Y|X))$ OTHER WAY TO COMPUTE VAR

PROOF:

$$E(V(Y|X)) = E\left(E\left((Y - E(Y|X))^2 \,|\, X\right)\right) = E\left(E(Y^2 + E^2(Y|X) - 2Y \cdot E(Y|X)\,|\,X)\right) \overset{\text{LINEARITY}}{=} E\left(E(Y^2|X) + E(E(E^2(Y|X)|X)) - 2E(E(Y E(Y|X)|X)\right) =$$

$$= E(Y^2) + E(E^2(Y|X)) - 2E(E(Y|X) \cdot \overbrace{E(1|X)}^{=1}) = E(Y^2) - E(E^2(Y|X)) \quad (+E^2(Y) - E^2(Y)) =$$

(THEN PROPERTY 1)

$$= E(Y^2) - E^2(Y) + E^2(Y) - E(E^2(Y|X)) = V(Y) - (E(E^2(Y|X)) - E^2(Y)) = V(Y) - V(E(Y|X))$$

BAYESIAN STATISTICS

- θ IS A PARAMETER ITSELF, A R.V. (OR R.VEC.)

  e.g. $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$

- BEFORE SEEING ANY DATA, θ HAS A PRIOR DISTRIBUTION

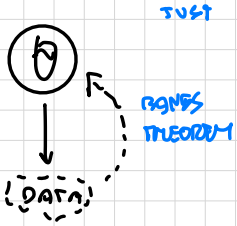
- AFTER SEEING DATA, WE UPDATE OUR PRIOR DISTRIBUTION ON θ INTO A POSTERIOR DISTRIBUTION, USING BAYES THEOREM

- GIVEN θ A VECTOR WITH PRIOR DENSITY $\tilde{\pi}(\theta)$:

$$\underbrace{\tilde{\pi}_{\theta|DATA}(\theta \mid DATA)}_{\text{POSTERIOR}} = \frac{f(DATA \mid \theta) \cdot \underbrace{\tilde{\pi}(\theta)}_{\text{PRIOR}}}{\int \tilde{\pi}(t) \cdot f(DATA \mid t)\, dt}$$

↳ OFTEN JUST $\tilde{\pi}(\theta \mid DATA)$

→



- SUPPOSE:

$$X_1, \dots, X_n \mid \theta \overset{iid}{\sim} N(\theta, \sigma^2)$$

R.V. ↑ KNOWN ↑

- WE ASSUME $\theta \sim N(\theta_0, \sigma_0^2) = \underbrace{\tilde{\pi}(\theta)}_{\text{PRIOR}}$ / μ_0, σ_0^2 ARE KNOWN NUMBERS EXPRESSING OUR KNOWLEDGE

→

$$f(\underbrace{x_1, \dots, x_n}_{DATA} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \prod_{i=1}^{n} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2}}_{\text{NORMAL DENSITY}} =$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i-\theta)^2\right\} =$$

$$\sum_i (x_i-\theta)^2 = \sum_i (x_i^2 - 2\theta x_i + \theta^2) = \sum_i x_i^2 - 2\theta \sum_i x_i + n\theta^2 =$$

$$= \dots = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}\left(\sum x_i^2 - 2\theta n\bar{x} + n\theta^2\right)\right\}$$

$$n\bar{x} = n \cdot \frac{\sum x_i}{n}$$

- APPLYING BAYES THEOREM:

$$\tilde{\Pi}(\theta \mid \text{DATA}) = \frac{\mathcal{L}(\text{DATA} \mid \theta) \cdot \hat{\Pi}(\theta)}{\underbrace{\int \hat{\Pi}(t) \cdot \mathcal{L}(\text{DATA} \mid t)\, dt}_{\text{CONST.}}} \propto \mathcal{L}(\text{DATA} \mid \theta) \cdot \hat{\Pi}(\theta)$$

$$\cdot \mathcal{L}(\text{DATA} \mid \theta) \cdot \hat{\Pi}(\theta) =$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma_0^2}}}_{\text{CONST.}} \cdot \exp\left\{ -\frac{1}{2\sigma_0^2}(\theta - \theta_0)^2 \right\} \cdot \underbrace{\left(2\pi\sigma^2\right)^{-\frac{n}{2}}}_{\text{CONST.}} \cdot \exp\left\{ -\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - 2\theta n\bar{x} + n\theta^2 \right) \right\}$$

$$\propto \exp\left\{ -\frac{1}{2\sigma_0^2} \underbrace{(\theta - \theta_0)^2}_{\theta^2 - 2\theta\theta_0 + \theta_0^2} \right\} \cdot \exp\left\{ -\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - 2\theta n\bar{x} + n\theta^2 \right) \right\}$$

- GIVEN $\begin{cases} \tau = \frac{1}{\sigma^2} & : \text{PRECISION} \\ \tau_0 = \frac{1}{\sigma_0^2} & : \text{PRIOR PRECISION} \end{cases}$

CONST. → REMOVED FOR PROPORTIONALITY

$$\to \exp\{\dots\} \cdot \exp\{\dots\} = \exp\left\{ -\frac{1}{2}\tau_0\left(\theta^2 - 2\theta\theta_0 + \theta_0^2\right) - \frac{1}{2}\tau\left(\sum_i x_i^2 - 2\theta n\bar{x} + n\theta^2 \right) \right\} =$$

$$= \exp\left\{ -\frac{1}{2}\left(\tau_0\theta^2 - 2\tau_0\theta\theta_0 - 2\tau\theta n\bar{x} + \tau n\theta^2 \right) \right\} =$$

$$= \exp\left\{ -\frac{1}{2}\left((\tau_0 + n\tau)\theta^2 - 2(\tau_0\theta_0 + n\tau\bar{x})\theta \right) \right\} =$$

TRICK TO COMPLETE THE SQUARE: $\pm \left(\frac{\tau_0\theta_0 + n\tau\bar{x}}{\tau_0 + n\tau} \right)^2$

$$= \exp\left\{ -\frac{1}{2}(\tau_0 + n\tau)\left(\theta^2 - 2\frac{\tau_0\theta_0 + n\tau\bar{x}}{\tau_0 + n\tau}\theta \right) \right\} =$$

$$\propto \exp\left\{ -\frac{1}{2}(\tau_0 + n\tau)\left(\theta - \frac{\tau_0\theta_0 + n\tau\bar{x}}{\tau_0 + n\tau} \right)^2 - \left(\frac{\tau_0\theta_0 + n\tau\bar{x}}{\tau_0 + n\tau} \right)^2 \cdot \left(-\frac{1}{2}(\tau_0 + n\tau) \right) \right\}$$

$$\propto \exp\left\{ -\frac{1}{2}(\tau_0 + n\tau)\left(\theta - \underbrace{\frac{\tau_0\theta_0 + n\tau\bar{x}}{\tau_0 + n\tau}} \right)^2 \right\} \propto \hat{\Pi}(\theta \mid x_1, \dots, x_n)$$

KERNEL OF A NORMAL DENSITY

$$\to \theta \mid x_1, \dots, x_n \sim N\left(\frac{\tau_0\theta_0 + n\tau\bar{x}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau} \right)$$

BAYESIAN ANALYSIS OF BINARY RANDOM SAMPLE:

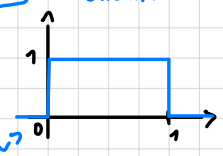
- RANDOM SAMPLES MAY BE BINARY, eg. BERNOULLI:

$$X_1, \dots, X_n \mid p \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$$

- PRIOR $\tilde{\Pi}(p)$:
  - WE COULD THINK IT AS A DISCRETE R.V. → DISCRETE BNs
  - ...OR AS A CONTINUOS R.V., eg WITH A BETA DISTRIBUTION:

$$p \sim \text{Beta}(a,b) \rightarrow \tilde{\Pi}(p) = \frac{p^{a-1}(1-p)^{b-1}}{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt}, \quad 0 < p < 1$$

(PRIOR DENSITY; the integral in the denominator is the NORMALIZING CONSTANT $B(a,b)$)



- SPECIAL CASE, $a = b = 1$:

$$\tilde{\Pi}(p) = \frac{p^{1-1}(1-p)^{1-1}}{\int_0^1 dt} = \frac{1}{\int_0^1 dt} = \mathbb{1}_{[0,1]}(p)$$

- $E[p] = \int_0^1 p \cdot \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}\,dp = \frac{B(a+1,b)}{B(a,b)} = \frac{a}{a+b}$
- $V[p] = \frac{ab}{(a+b)^2(a+b+1)}$

- N.B.: EULER'S BETA FUNCTION: $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$
  - $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ / $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$ (GAMMA FUNCTION)
  - PROPERTY: $\Gamma(x) = (x-1)\Gamma(x-1)$ / IF $x = n \in \mathbb{Z}$: $\Gamma(n) = (n-1)!$

- $X_1, \dots, X_n \mid p \overset{\text{iid}}{\sim} \text{Bernoulli}(p) \quad / \quad p \sim \text{Beta}(a, b)$

- CONDITIONAL DENSITY:

$$f(x_1, \dots, x_n \mid p) = \prod_{i=1}^{n} f_{X_i \mid p}(x_i \mid p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

↳ $\mathcal{L}(p, x_1, \dots, x_n)$ ↳ LIKELIHOOD FUNCTION $= p^{\sum x_i} \cdot (1-p)^{n - \sum x_i}$

- BAYES POSTERIOR:

$$\Pi(p \mid x_1, \dots, x_n) = \frac{\tilde{\Pi}(p) \cdot f(x_1, \dots, x_n \mid p)}{\int_0^1 \tilde{\Pi}(q) \cdot f(x_1, \dots, x_n \mid q)\, dq} =$$

$$= \frac{p^{a-1}(1-p)^{b-1}}{\underbrace{\left[\int_0^1 t^{a-1}(1-t)^{b-1}\, dt\right]}_{\text{CONST. PROP.}}} \cdot p^{\sum x_i} \cdot (1-p)^{n - \sum x_i} \propto p^{a-1}(1-p)^{b-1} \cdot p^{\sum x_i} (1-p)^{n - \sum x_i}$$

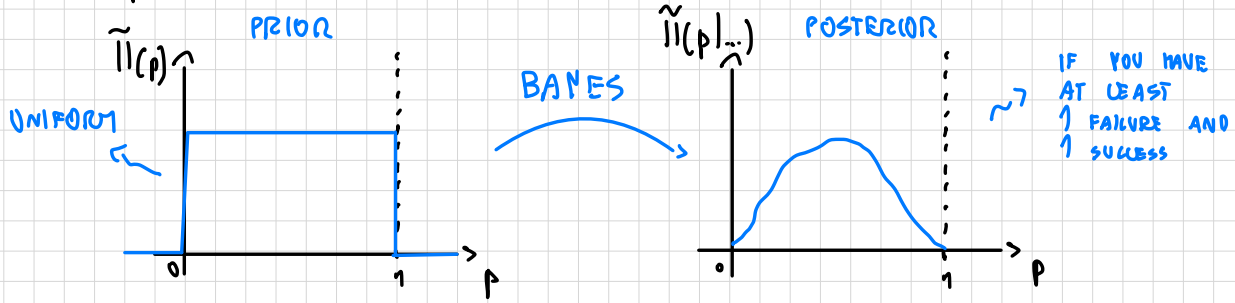
$\left[\int_0^1 \tilde{\Pi}(q) \cdot f(x_1, \dots, x_n \mid q)\, dq\right]$ CONST. PROP.

IT IS A BETA AGAIN

$$\cdot\ p^{a-1}(1-p)^{b-1} \cdot p^{\sum x_i} (1-p)^{n - \sum x_i} = p^{a + \sum x_i - 1} (1-p)^{b + n - \sum x_i - 1}$$

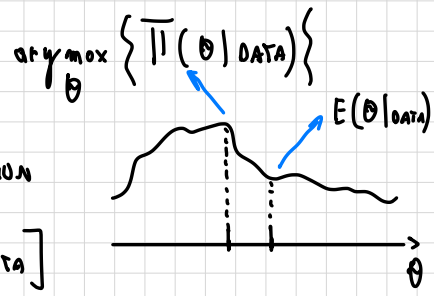
$$\rightarrow \Pi(p \mid x_1, \dots, x_n) = \frac{p^{a + \sum x_i - 1} (1-p)^{b + n - \sum x_i - 1}}{\int_0^1 t^{a + \sum x_i - 1} (1-t)^{b + n - \sum x_i - 1}\, dt}$$

$$\rightarrow p \mid x_1, \dots, x_n \sim \text{Beta}\left(a + \sum_i x_i,\ b + n - \sum x_i\right)$$



BAYESIAN STATISTICS CORE:

· POINT ESTIMATION:



IT IS AN APPROXIMATE SUMMARY OF THE POSTERIOR DISTRIBUTION

↳ ex. POSTERIOR EXPECTED VALUE: $E[\theta \mid \text{DATA}]$

↳ POSTERIOR MODE, MAP: MAXIMUM A POSTERIORI PROBABILITY

· NORMAL RANDOM SAMPLE:

$$E(\mu \mid x_1,\dots,x_n) = \frac{\tilde{\tau}_0\mu_0 + n\tilde{\tau}\bar{x}}{\tilde{\tau}_0 + n\tilde{\tau}} = \frac{\tilde{\tau}_0}{\tilde{\tau}_0+n\tilde{\tau}}\mu_0 + \frac{n\tilde{\tau}}{\tilde{\tau}_0+n\tilde{\tau}}\bar{x} =$$

WEIGHTED AVERAGE OF: · μ_0: PRIOR MEAN · $\bar{x}$: SAMPLE MEAN

$$= \left(1 - \frac{n\tilde{\tau}}{\tilde{\tau}_0+n\tilde{\tau}}\right)\mu_0 + \frac{n\tilde{\tau}}{\tilde{\tau}_0+n\tilde{\tau}}\bar{x} \quad / \quad w = \frac{n\tilde{\tau}}{\tilde{\tau}_0+n\tilde{\tau}}$$

· $\downarrow \tilde{\tau}_0$ ($\to$ HIGHER σ_0^2: HIGHER UNCERTAINTY) $\longrightarrow$ $\uparrow$ DATA IMPORTANCE FOR ESTIMATION

· IF $n \mapsto \infty$: $\lim_{n\to\infty} \frac{n\tilde{\tau}}{\tilde{\tau}_0+n\tilde{\tau}} = 1$ $\longrightarrow$ ONLY DATA COUNTS

↳ $n \mapsto \infty$: $(1-w)\mu_0 + w\bar{x}$ ($1-w \to 0$, $w \to 1$)

· BINARY DATA CASE, Beta(a,b):

$$E(p \mid x_1,\dots,x_n) = \frac{a + \sum x_i}{a+\sum x_i + b + n - \sum x_i} = \frac{a+\sum x_i}{a+b+n} =$$

$$= \frac{a+b}{a+b+n}\cdot\frac{a}{a+b} + \frac{n}{a+b+n}\cdot\frac{\sum x_i}{n} =$$

SAME PROPERTIES AS ABOVE

$$= \left(1 - \frac{n}{a+b+n}\right)\frac{a}{a+b} + \frac{n}{a+b+n}\bar{x}$$

↳ $\hat{p}$: MLE OF p

· IF $a=b=1$: $E(p \mid x_1,\dots,x_n) = \frac{1+\sum x_i}{2+n} = \tilde{p}$ BAYES ESTIMATOR

↳ IT GIVES UNIFORM ON p: GOOD PRIOR IGNORANCE

EX. $x_1 = \dots = x_n = 1 \longrightarrow \tilde{p} = \frac{1+n}{2+n}$ LAPLACE SUCCESSION RULE