Cardiovascular Diseases

Laura Ruffoni Gabriele Oliveto Riccardo De Sury Leonardo Nossa

Dataset Outlining And Index

Blood Composition: Cholesterol and Glucose

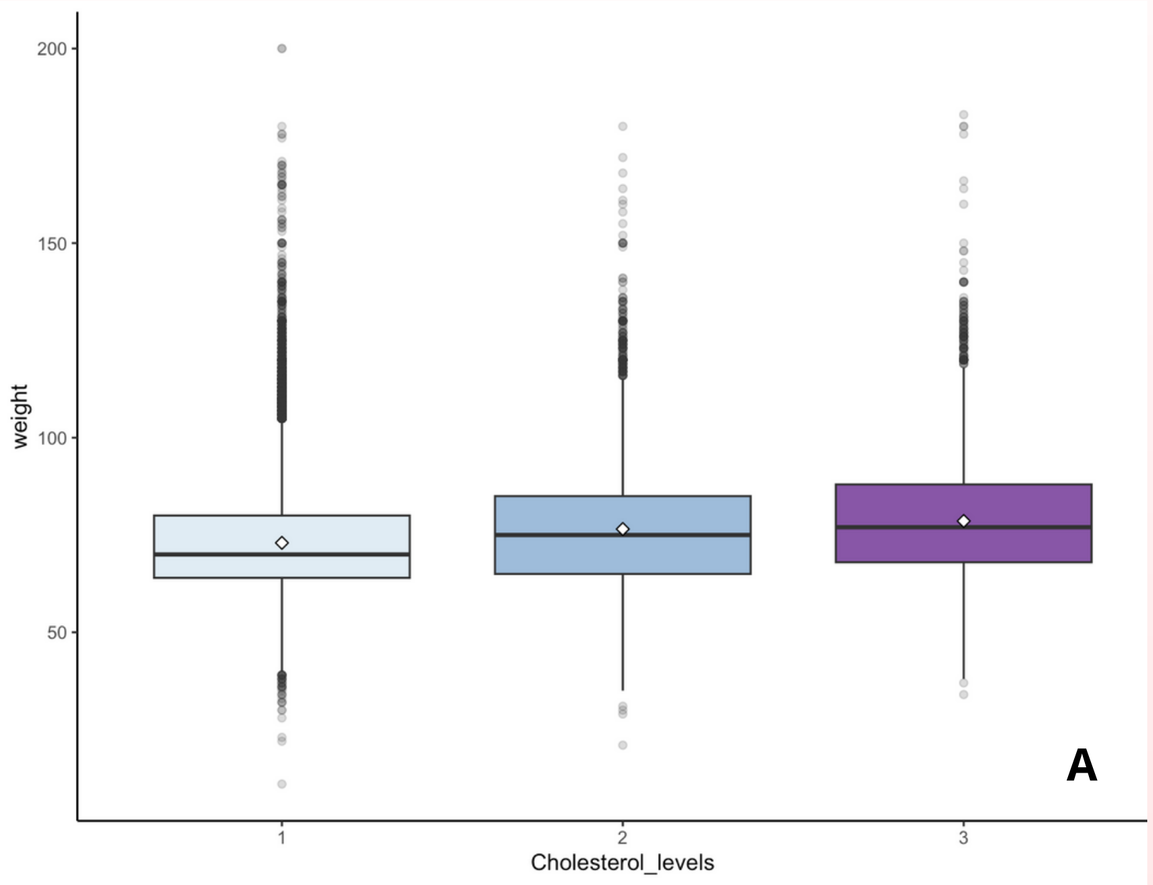
Age Influence

Habits: Sport, Smoking and Drinking

Predictive Model

Best vs Worst Habits

Cholesterol

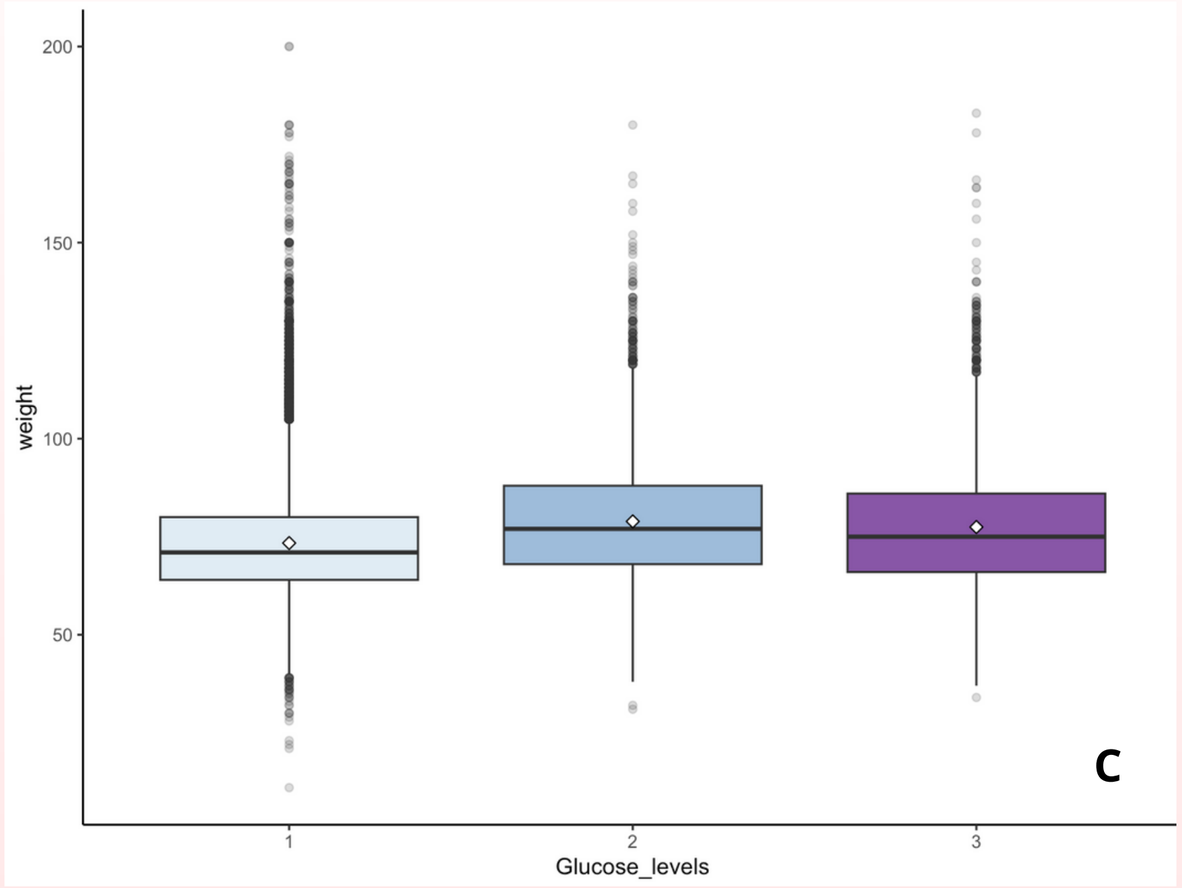


B	Perc_ill	Weight	BMI
Normal cholesterol level	43.475	72.981	27.023
Above normal cholesterol level	59.515	76.533	28.535
Well above normal cholesterol level	76.142	78.595	29.503

Figure A is a boxplot showing the relationship between cholesterol (three levels) and body weight

Figure B is a table showing the percentage of ill patients for each cholesterol level and the average weight and average BMI values

Glucose



D	Perc_ill	Weight	BMI
Normal glucose level	47.478	73.387	27.225
Above normal glucose level	58.623	78.904	29.377
Well above normal glucose level	61.622	77.466	28.905

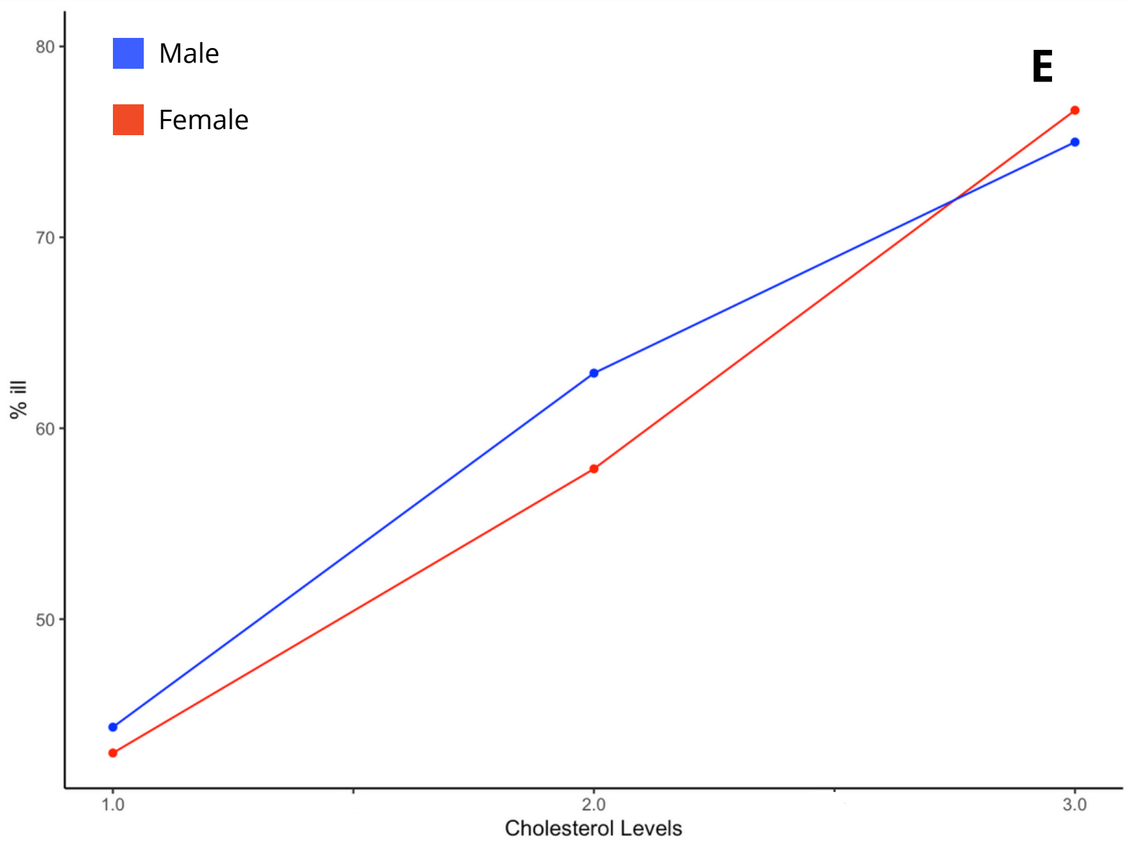
Figure C is a boxplot showing the relationship between glucose (for the three levels) and body weight

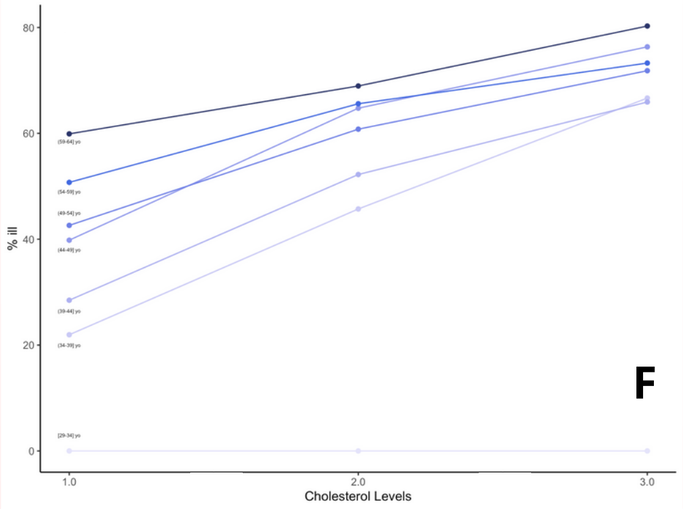
Figure D is a table showing the percentage of ill patients for each glucose level and the average weight and BMI values

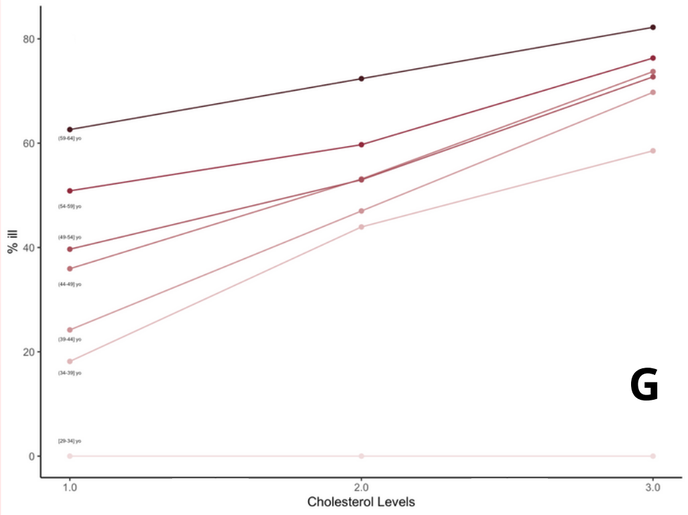
Cholesterol

Figure E is a lineplot showing the relationship between cholesterol (for the three levels) and percentage of people affected by cardiovascular disease

Figure F and ***Figure G*** are lineplots (as above) where male and female are divided in different age ranges

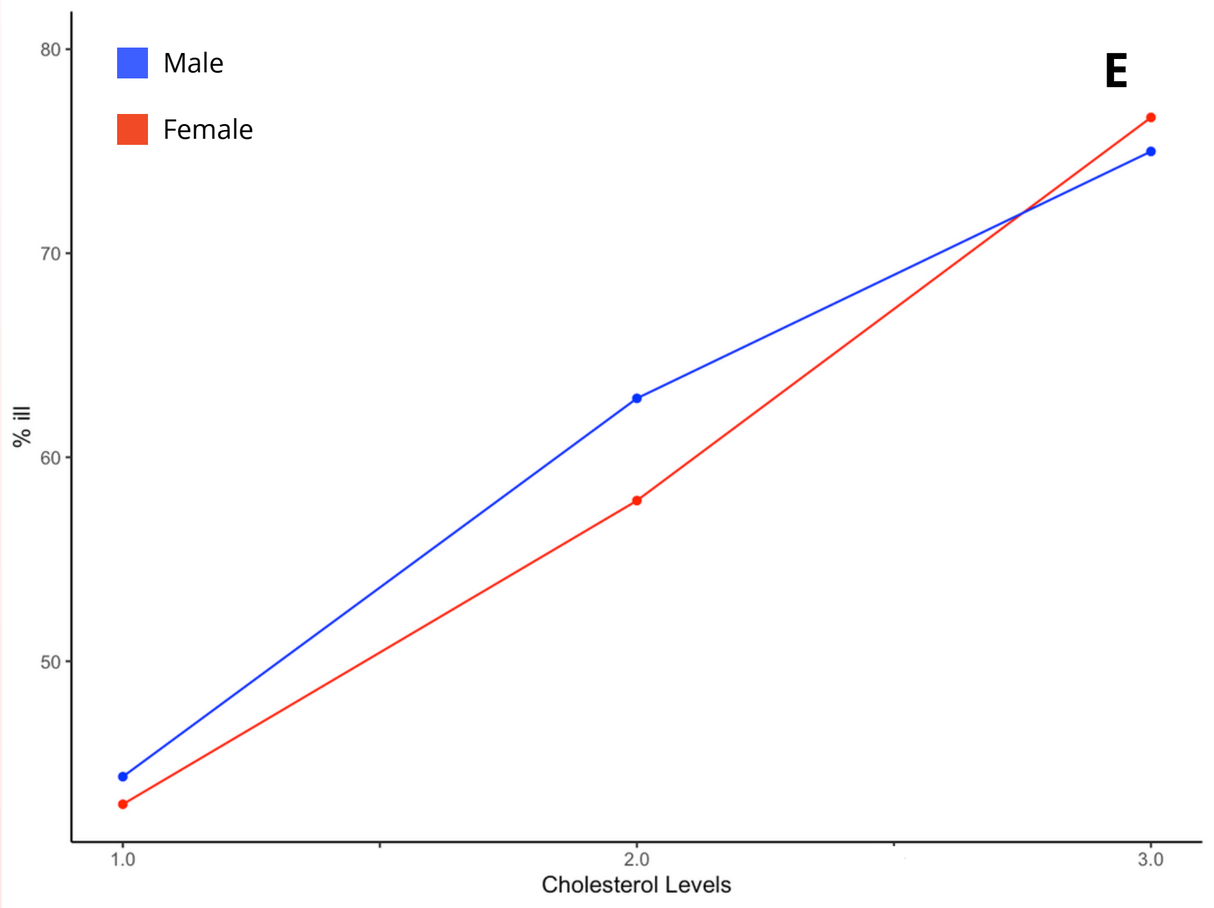






Cholesterol

Hypothesis test on level 3 of cholesterol:



$$X_{female.1}, \dots, X_{female.N} \quad iid \quad X_{female}i \sim Be(p)$$

$$X_{male.1}, \dots, X_{male.N} \quad iid \quad X_{male}i \sim Be(p) \qquad E[X] = p$$

$$Var(X) = p * (1 - p)$$

$$H_0: D = 0 \quad H_1: D \neq 0$$

$$D = p_{female} - p_{male}$$

$$\widehat{D} = \widehat{p}_{female} - \widehat{p}_{male} \sim N(p_{female} - p_{male}, \frac{Var(X_{female})}{N_{female}} + \frac{Var(X_{male})}{N_{male}})$$

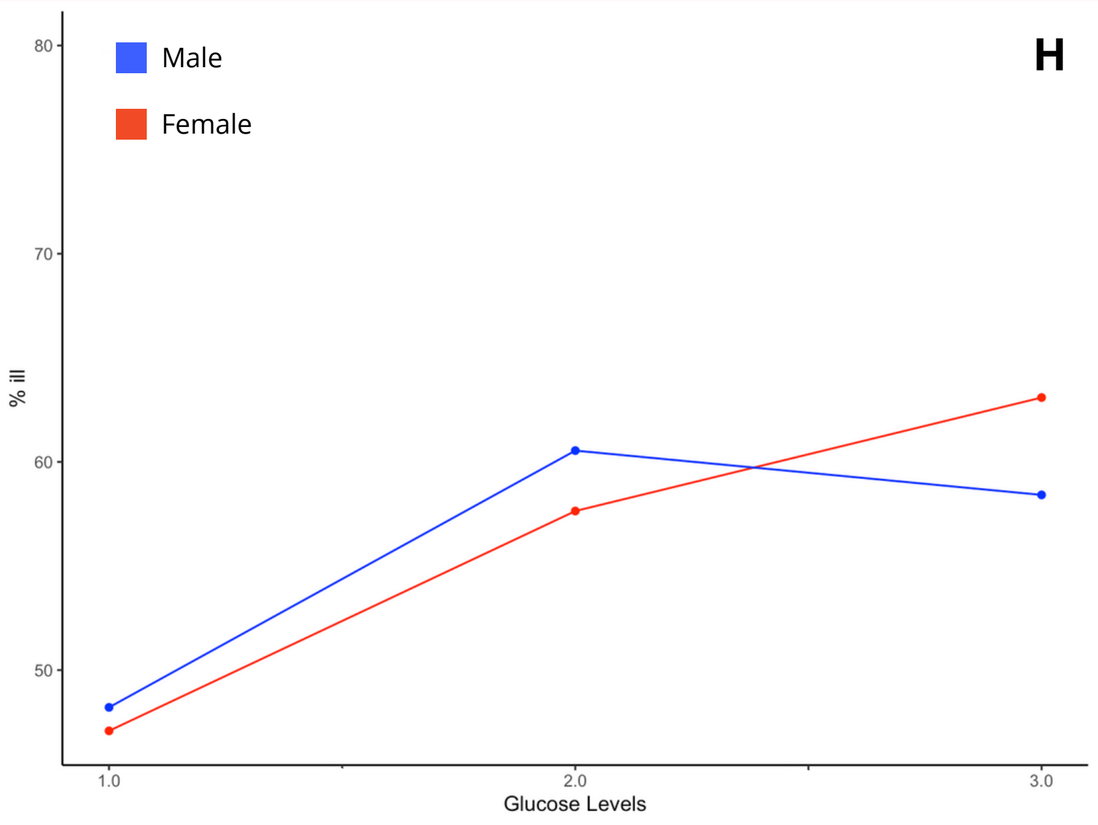
$$Under\ H_0: \widehat{D} \sim N(0, sd^2_D) \qquad N >> 1 \qquad Z = \frac{D-0}{\sqrt{sd^2_D}}$$

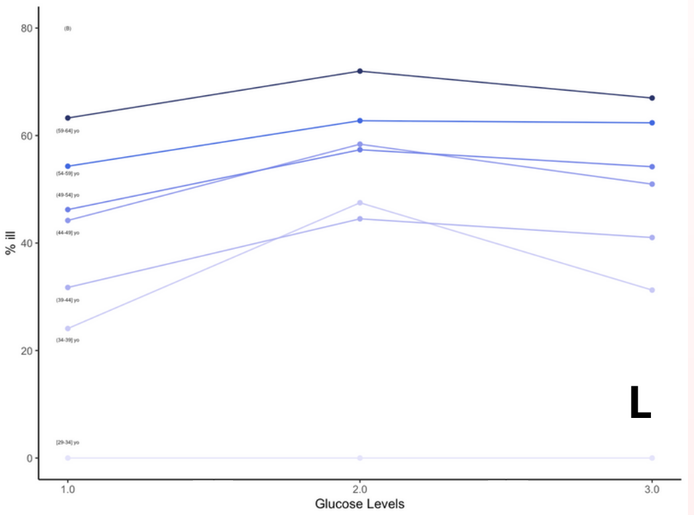
$$p - value = 0.114104$$

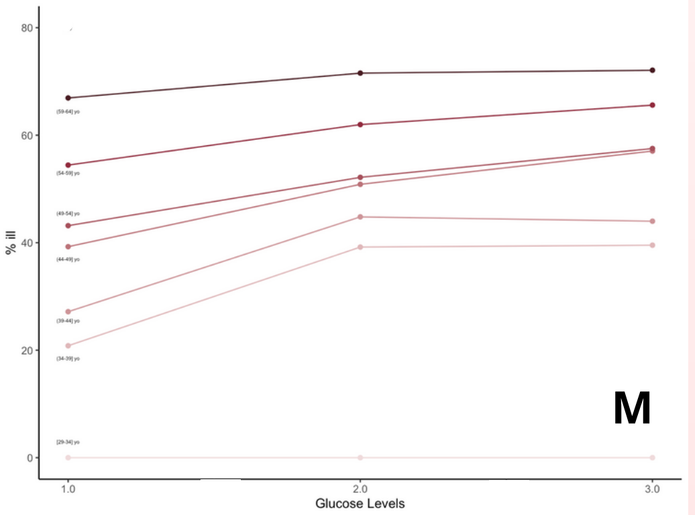
Glucose

Figure H is a lineplot showing the relationship between cholesterol (for the three levels) and percentage of people affected by cardiovascular disease

Figure L and ***Figure M*** are lineplots (as above) where male and female are divided in different age ranges

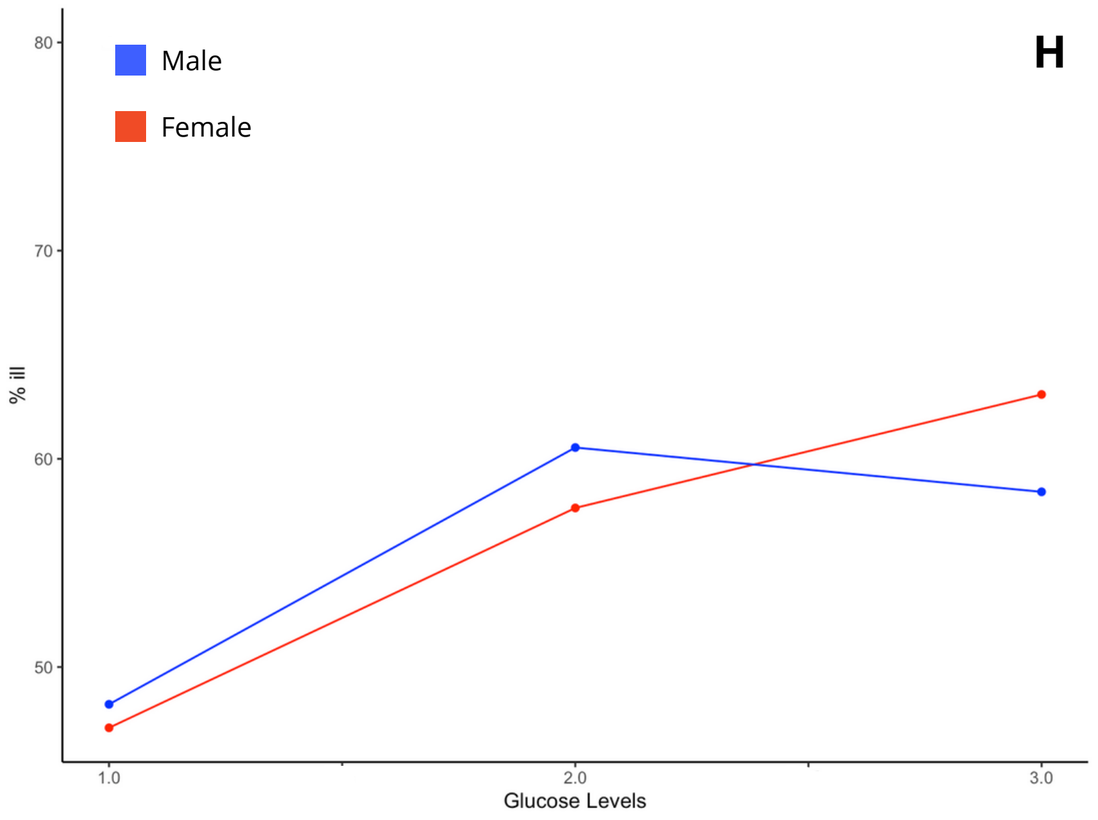






Glucose

Hypothesis Test on level 3 and 2 of glucose:



$\rightarrow$ *Level 3*

$$H_0: D = 0 \quad H_1: D > 0$$

$$p - value = 0.00070965459$$

$\rightarrow$ *Level 2*

$$H_0: D = 0 \quad H_1: D < 0$$

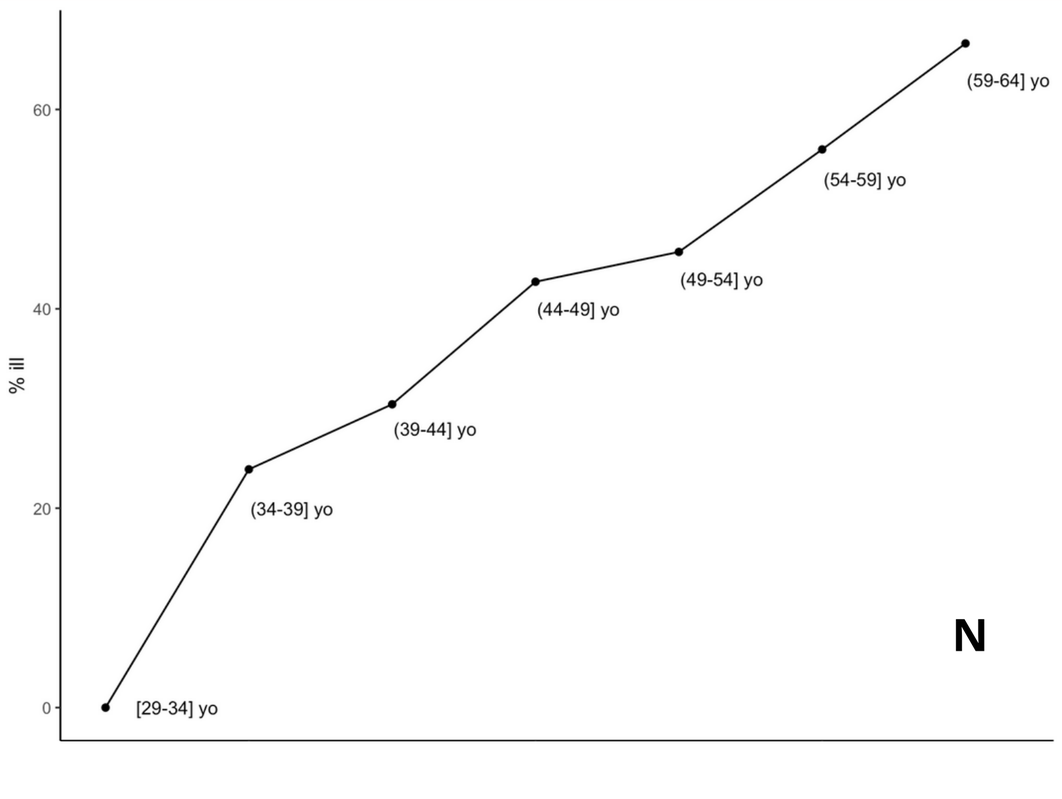
$$p - value = 0.02383739$$

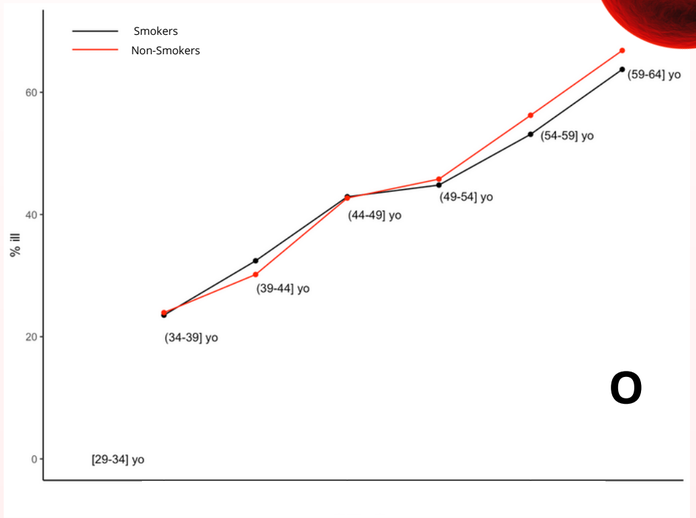
Age

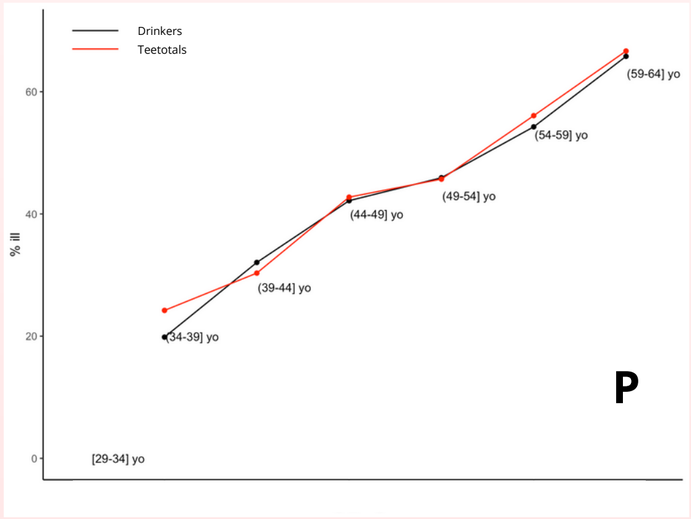
Figure N is a lineplot showing the relationship between age and percentage of people affected by cardiovascular disease

Figure O is a lineplot showing the relationship between smoke and percentage of people affected by cardiovascular disease

Figure P is a lineplot showing the relationship between alcohol and percentage of people affected by cardiovascular disease

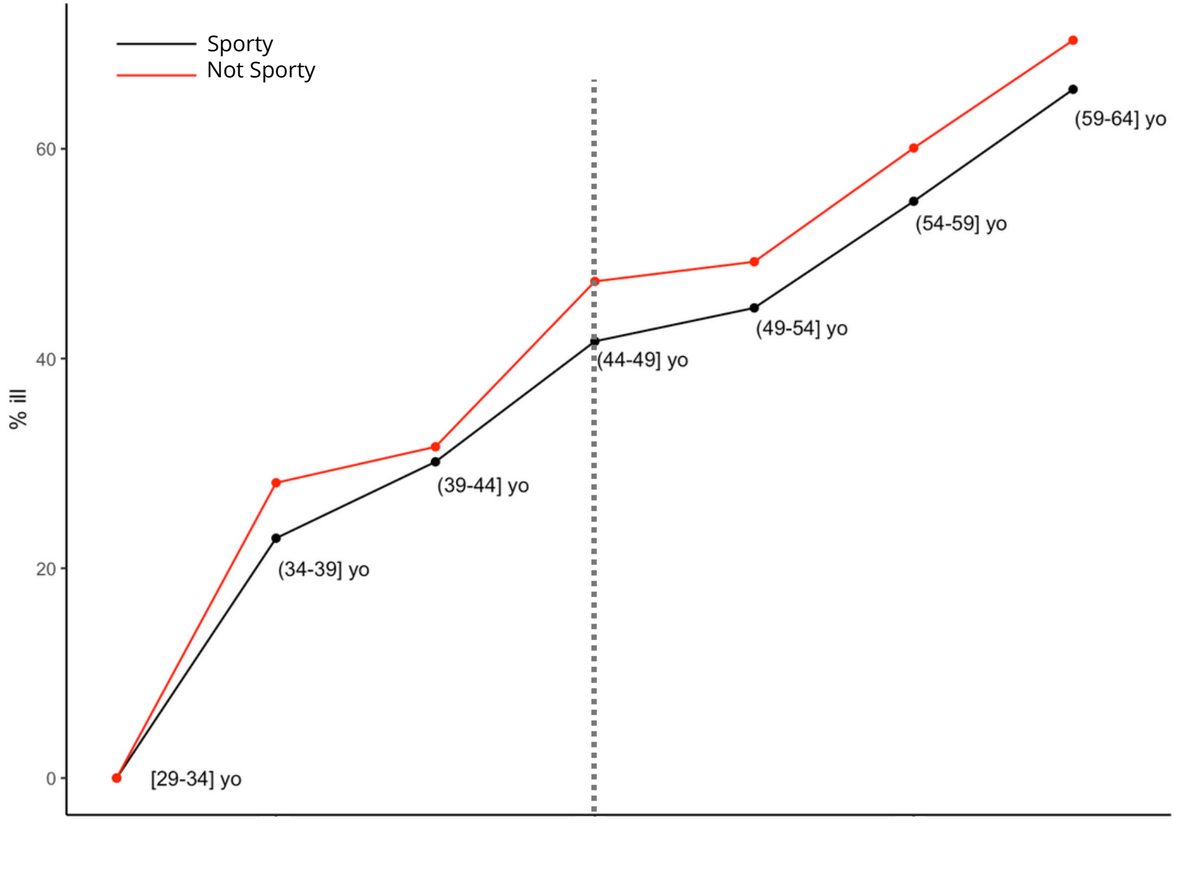






Sport

Hypothesis Test between sporty or non-sporty patients:



$$X_{sport.1}, \dots, X_{sport.N} \quad iid \quad X_{sport}i \sim Be(p)$$

$$X_{no\ sport.1}, \dots, X_{no\ sport.N} \quad iid \quad X_{no\ sport}i \sim Be(p)$$

$$H_0: D = 0 \quad H_1: D \neq 0$$

$$D = p_{no\ sport} - p_{sport}$$

$$\widehat{D} = \widehat{p}_{no\ sport} - \widehat{p}_{sport} \sim N(p_{no\ sport} - p_{sport}, \frac{Var(X_{no\ sport})}{N_{no\ sport}} + \frac{Var(X_{sport})}{N_{sport}})$$

$$Under\ H_0: \widehat{D} \sim N(0, sd^2_D) \quad N >> 1: \quad Z = \frac{D-0}{\sqrt{sd^2_D}}$$

$$p - value_{29-44\ yo} = 0.07199555 \qquad p - value_{45-64\ yo} = 0$$

Predictive Model

Model's Equation:

$$\log[p(cardio)/(1-p(cardio))] = \beta0 + \beta1*active + \beta2*weight + \beta3*age + \beta4*ap\_hi + \beta5*ap\_lo + \beta6*cholesterol$$

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.208e+01  1.485e-01 -81.363  < 2e-16 ***
active      -2.278e-01  2.623e-02  -8.687  < 2e-16 ***
weight       9.365e-03  7.852e-04  11.927  < 2e-16 ***
age          1.414e-04  4.424e-06  31.952  < 2e-16 ***
ap_hi        5.858e-02  1.135e-03  51.633  < 2e-16 ***
ap_lo        1.002e-02  1.792e-03   5.590 2.27e-08 ***
cholesterol  4.512e-01  1.658e-02  27.219  < 2e-16 ***
```

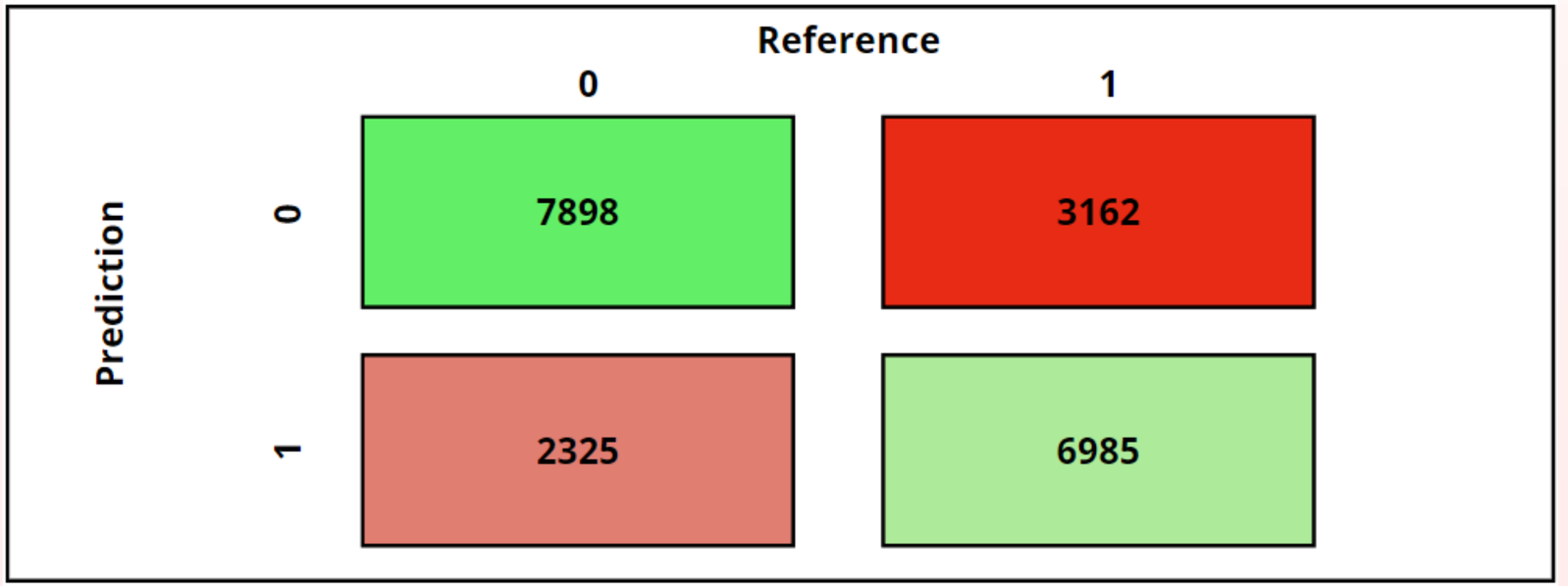
VIF

Active	Weight	Age	Ap_hi	Ap_lo	Cholesterol
1.001071	1.045474	1.011153	1.803825	1.791676	1.016573

McFadden
0.1913449

Predictive Model

The optimized treshold chosen to maximize **accuracy** is **0.486**



DETAILS

Accuracy	95% CI	NIR	P-value [Acc>NIR]	Kappa
0.7306	(0.7245, 0.7367)	0.5019	< 2.2e-16	0.4611

McNemar's Test P-Value	Sensitivity	Specificity
< 2.2e-16	0.7726	0.6884

Best vs Worst Habits

Figure Q shows percentage of sick and healthy patients having favorable or negative habits

	Patient A	Patient B
Active	1	0
Weight	65	71
Age	49	41
Ap_hi	110	160
Ap_lo	70	120
Cholesterol	1	3
Probability	0.1737812	0.9325049
Output	**NOT SICK**	**SICK**

Thanks For Listening!

Additional Information A

This is the header of the dataset used to for our project:

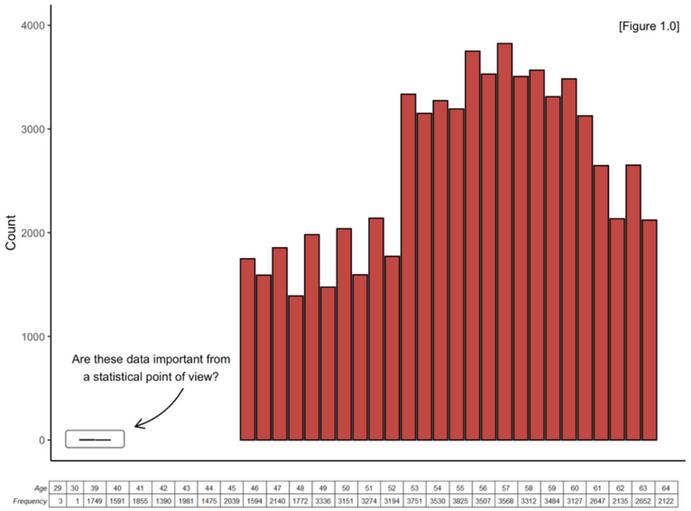
id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	age_years	bmi	bp_category
0	18393	2	168	62	110	80	1	1	0	0	1	0	50	21.96712	Hypertension Stage 1
1	20228	1	156	85	140	90	3	1	0	0	1	1	55	34.92768	Hypertension Stage 2

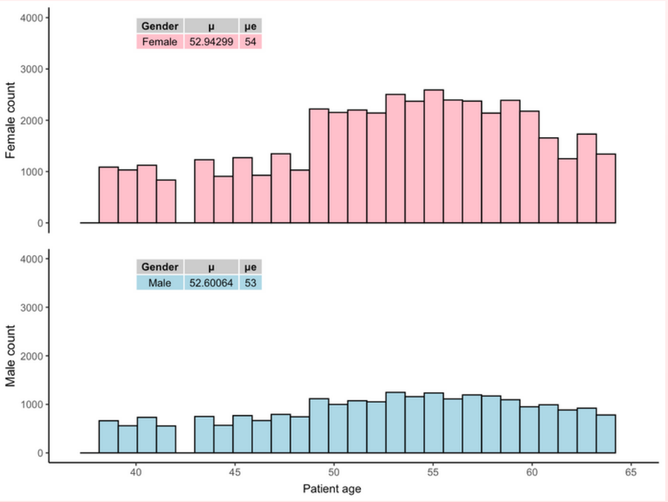
It’s downloadable from kaggle at the following link : https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease

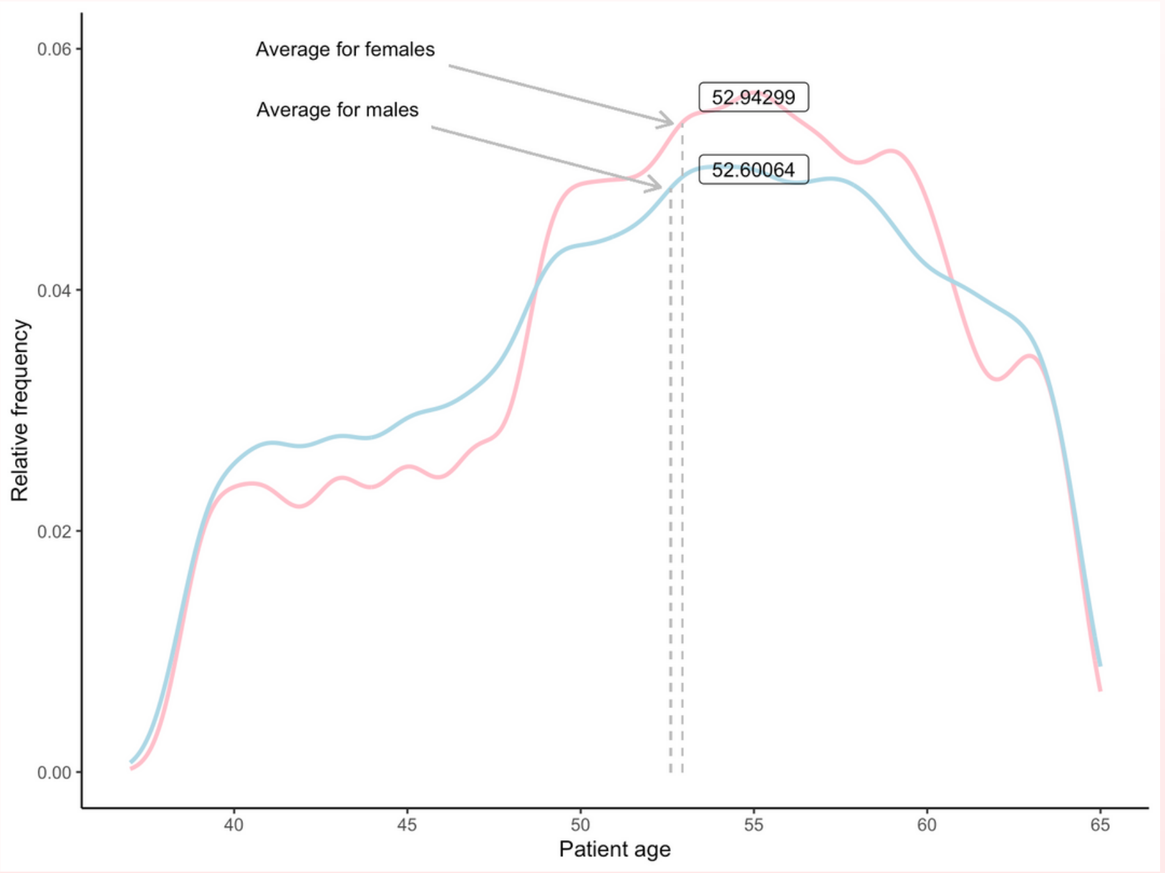
Bibliography:

- Life Science Journal 2013; Dukjae Lee; **A Comparison of Choice-based Landscape Preference Models between British and Korean Visitors to National Parks**
- Maturitas 168 (2023) 49–52; Hira Shakoor et al; **The benefits of physical activity in middle-aged individuals for cardiovascular disease outcomes**
- Preventive Medicine 27, 1–9 (1998); Hussain R. Yusuf et al ; **Impact of Multiple Risk Factor Profiles on Determining Cardiovascular Disease Risk**

Additional Information B







Additional information C

$X_{sport.1}, \dots, X_{sport.N}$ iid $X_{sport}i \sim Be(p)$ $\quad f(x) = \begin{cases} p & if\ x = 1\ \ cardiovascular\ disease \\ 1-p & if\ x = 0 \end{cases}$ $\quad E[x] = p$

$Var(x) = p * (1-p)$

$X_{no\ sport.1}, \dots, X_{no\ sport.N}$ iid $X_{no\ sport}i \sim Be(p)$

$Xi_{sport} \perp Xi_{no\ sport}$ *and unknown and uncommon variances*

→ *Point Estimation:* $\hat{p}$ *and* sd^2 *both unbiased*

$H_0: D = 0 \quad H_1: D \neq 0$

$p = \hat{p} = \frac{\#\ patients\ with\ disease}{N} \qquad N >> 1\ for\ CLT \rightarrow \hat{p} \sim N(p, \frac{p*(1-p)}{N})$

$Under\ H_0: \hat{D} \sim N(0, sd^2_D)$

$sd^2 = p * (1-p) = \hat{p} * (1-\hat{p})$

$N >> 1: Z = \frac{D-0}{\sqrt{sd^2_D}}$

$\hat{p}_{no\ sport} = 0.315856 \qquad \hat{p}_{sport} = 0.3015353$

$p - value = 2 * (1 - \Phi_{(|z|)})$

$sd^2_{no\ sport} = 0.216091 \qquad sd^2_{sport} = 0.2106117$

$D = p_{no\ sport} - p_{sport} \qquad \hat{D} = \hat{p}_{no\ sport} - \hat{p}_{sport} \sim N(p_{no\ sport} - p_{sport}, \frac{Var(X_{no\ sport})}{N_{no\ sport}} + \frac{Var(X_{sport})}{N_{sport}})$

$sd^2_D = \frac{sd^2_{no\ sport}}{N_{no\ sport}} + \frac{sd^2_{sport}}{N_{sport}}$

Additional information D

```
sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.7,0.3))
train <- data[sample, ]
test <- data[!sample, ]

model <- glm(cardio  ~ active+weight+age+ap_hi+ap_lo+cholesterol,  family="binomial", data=train)

summary(model)
pscl::pR2(model)["McFadden"]
caret::varImp(model)
car::vif(model)
```

```
validazione_best = data.frame(active = 1, weight = 67, age = 17885, ap_hi = 110, ap_lo = 70, cholesterol = 1)
validazione_worst = data.frame(active = 0, weight = 71, age = 14965, ap_hi = 160, ap_lo = 120, cholesterol = 3)

ext1 = predict(model, validazione_best, type="response")
ext2 = predict(model, validazione_worst, type="response")
ext1 = ext1 > 0.486
ext2 = ext2 > 0.486

predicted <- predict(model, test, type="response")
predicted = as.array(predicted)
res = as.data.frame(predicted >= 0.486)
colnames(res)[colnames(res) == 'predicted >= 0.486'] <- "Malato/NonMalato"
res$`Malato/NonMalato` <- ifelse(res$`Malato/NonMalato`, 1, 0)

res$`Malato/NonMalato` = factor(res$`Malato/NonMalato`, levels = c(0,1))
test$cardio = factor(test$cardio, levels = c(0,1))

confusion_matrix = confusionMatrix(
                   res$`Malato/NonMalato`,
                   test$cardio,
                   positive = NULL,
                   dnn = c("Prediction", "Reference"),
                   prevalence = NULL,
                   mode = "sens_spec")
```