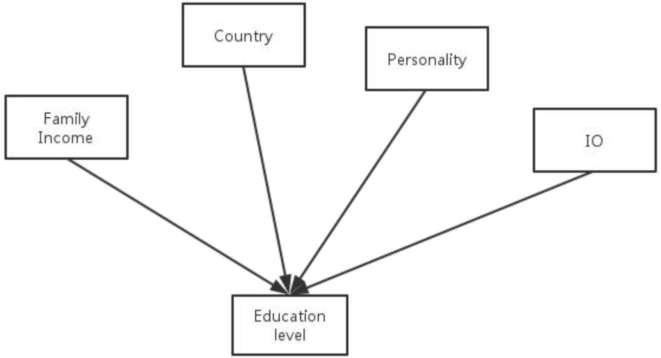
Multiple Testing & Bonferroni Correction

Multiple Testing

In last recitation, permutation test and z-test was done based on one hypothesis.

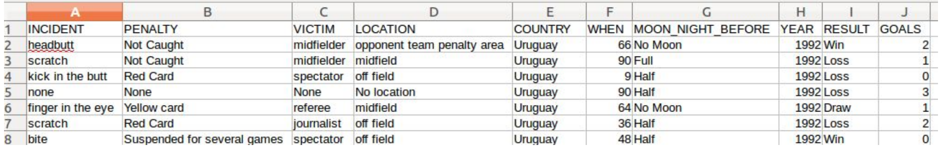
In real word problem, on attribute is affected by several factors.



Import Data

Luis Suarez is a famous Uruguayan soccer player who, among other things, is famous for his on-field “incident”.

This is dataset contains info on 1890 games that Suarez has played in over the course of career. We are interested in finding features which make it more likely that there is some sort of "incident" during the game

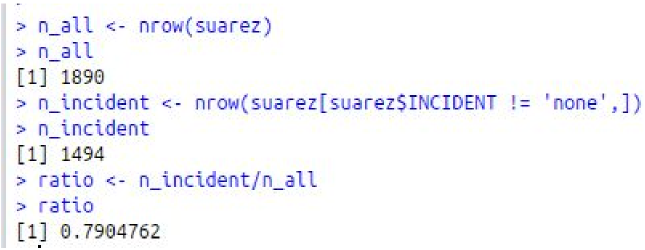


	A	B	C	D	E	F	G	H	I	J
1	INCIDENT	PENALTY	VICTIM	LOCATION	COUNTRY	WHEN	MOON_NIGHT_BEFORE	YEAR	RESULT	GOALS
2	headbutt	Not Caught	midfielder	opponent team penalty area	Uruguay	66	No Moon	1992	Win	2
3	scratch	Not Caught	midfielder	midfield	Uruguay	90	Full	1992	Loss	1
4	kick in the butt	Red Card	spectator	off field	Uruguay	9	Half	1992	Loss	0
5	none	None	None	No location	Uruguay	90	Half	1992	Loss	3
6	finger in the eye	Yellow card	referee	midfield	Uruguay	64	No Moon	1992	Draw	1
7	scratch	Red Card	journalist	off field	Uruguay	36	Half	1992	Loss	2
8	bite	Suspended for several games	spectator	off field	Uruguay	48	Half	1992	Win	0

General view about incidence

```
# all game records
n_all <- nrow(suarez)
# incident games
n_incident <- nrow(suarez[suarez$INCIDENT != 'none',])
# incident ratio
ratio <- n_incident/n_all

# group incident and non-incident to 1 and 0
incident <- rep(n_all,1)
incident[suarez$INCIDENT=='none'] = 0
```

Suarez had incidents in 79% games!!!



```
> n_all <- nrow(suarez)
> n_all
[1] 1890
> n_incident <- nrow(suarez[suarez$INCIDENT != 'none',])
> n_incident
[1] 1494
> ratio <- n_incident/n_all
> ratio
[1] 0.7904762
```

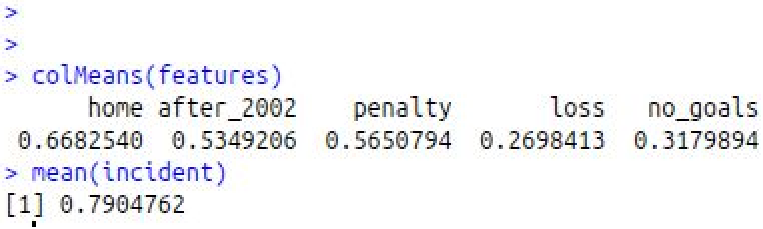
Multiple Hypothesis

We'll consider the following factors and whether or not they increase the likelihood of an incident:

1. Home or away match
2. After 2002 or before
3. Get cards for penalty
4. Lost the game?
5. Did Suarez scores 0 goals in the game?

```
#create a matrix to contain the 5 features, initialize all value as 0
features <- matrix(0,nrow = n_all,ncol = 5)
colnames(features) <- c('home','after_2002','penalty','loss','no_goals')

#if country is Uruguay, it is in home country, 0 means in others coutries
features[suarez$COUNTRY=='Uruguay','home'] <- 1

#the game happened  after 2002, set to 1.
features[suarez$YEAR >2002,'after_2002'] <- 1

#get card penalty
features[suarez$PENALTY %in% c('Yellow card', 'yellow card', 'Red card',
'red card'),'penalty'] <- 1

#game result is loss, set to 1
features[suarez$RESULT == "Loss","loss"] <- 1
#score no goal
features[suarez$GOALS == 0,"no_goals"] <- 1
```

General view each feature



```
>
>
> colMeans(features)
      home after_2002    penalty       loss   no_goals
 0.6682540  0.5349206  0.5650794  0.2698413  0.3179894
> mean(incident)
[1] 0.7904762
```

1. 66.8% games were in his home country.

2. 53.49% games in the dataset happened after 2002

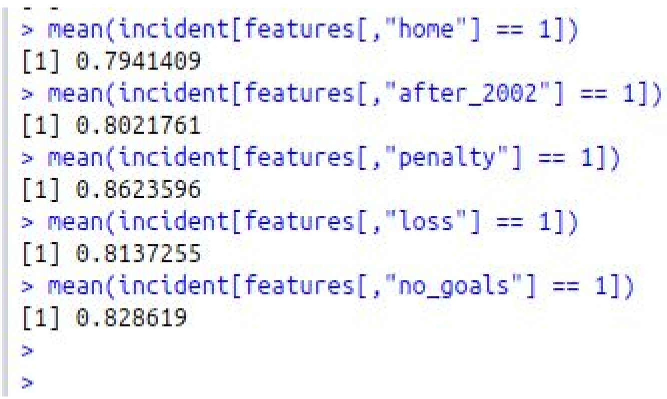
3. Suarez got card penalty in 56.50% soccer games.

4. Suarez lost 26.98% games

5. In 31.7% games, Suarez scored no goal

Suarez had incidents in 79.04% of all games.

General view each feature



```
> mean(incident[features[,"home"] == 1])
[1] 0.7941409
> mean(incident[features[,"after_2002"] == 1])
[1] 0.8021761
> mean(incident[features[,"penalty"] == 1])
[1] 0.8623596
> mean(incident[features[,"loss"] == 1])
[1] 0.8137255
> mean(incident[features[,"no_goals"] == 1])
[1] 0.828619
>
>
```

1. 79.42% incident rate for ‘home’ games, higher than 79.04% in general case.

2. 80.21% incident rate after 2002

3. 86.23% incident rate if he get card penalty

4. 81.37% incident rate in lost games.

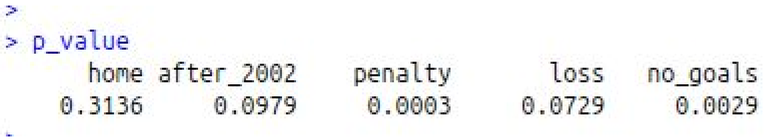
5. 82.86% incident rate without goal.

Suarez had incidents in 79.04% of all games.

Permutation test for each feature

Do permutation tests for other four features.

After 2002 or before?
card penalty or not?
Lost game or not?
No goals for Suarez？



```
>
> p_value
     home after_2002    penalty       loss   no_goals
   0.3136     0.0979     0.0003     0.0729     0.0029
```

Can we use critical value 5% here to verify the null hypothesis?

Bonferroni Correction

Bonferroni is one method for adjusting for multiple comparisons in hypothesis testing problems (probably the simplest). Other important methods involve controlling the "false discovery rate”.

$$P(T_i \text{ passes } \mid H_0) \leq \frac{\alpha}{n}$$

Can we use critical value 5% here to verify the null hypothesis?