Analyzing the Dynamics of Bone Mineral Index

[*Notebook*]

Factors impacting BMI undergo a statistical spotlight, revealing their significant effects succinctly.

Group Number 7

Ashith Shetty(21bcs019)
Jaishana Bindhu Priya(21bcs045)
Zaid Ragib(21bcs069)

Table of Content

Bone Mineral Density

01 Osteoporosis, a silent disease

A condition characterized by weak and brittle bones prone to fractures. BMD measurements help in identifying individuals at higher risk for osteoporosis-related fractures

02 Smoking and BMD

Smoking has a detrimental effect on bone health and can lead to decreased bone mineral density. It interferes with the absorption of calcium.

03 Diet and Bone Health

A diet rich in calcium and vitamin D can help promote bone health and reduce the risk of developing osteoporosis.

Data

Harvard Dataverse

Link

Lifestyle changes

This data set is the original data of the original paper "Nonlinear Association between Serum Uric Acid levels and risk of Osteoporosis: A Retrospective Study", including clinical baseline data, and dual-energy X-ray measurement results

The relationship between serum uric acid levels and osteoporosis is a topic of ongoing research. Osteoporosis is a bone disease that develops when bone mineral density and bone mass decreases, or when the structure and strength of bone changes. This can lead to a decrease in bone strength that can increase the risk of fractures

Data Analysis

Dataset had many Null values

```
print(df.isnull().sum())
```

```
Gender      0
Age        36
Height     34
Weight     34
BMI        34
L1-4        0
L1.4T       0
FN          0
FNT         0
TL          0
TLT         0
ALT         2
AST         2
BUN         1
CREA        3
URIC        0
FBG        16
HDL-C      17
LDL-C      14
Ca          2
P           5
Mg          3
Calsium     0
```

Data Visualisation and Data Cleaning

Visualisation
Check the difference before and after handling null using mean and median respectively

Handling Null Values
Use mean and median

Population Parameters

Calculate Mean

Calculate Median

Check Distribution

Unit 2: Parametric Test

Large Sample Z Test

The z-test for the mean is a statistical test for a population mean. The z-test can be used when the population is normal and is known, or for any population when the sample size n is at least 30.

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \qquad \frac{\sigma}{\sqrt{n}} = \text{standard error} = \sigma_{\bar{x}}$$

When $n \geq 30$, the sample standard deviation s can be substituted for σ.

Unit 2: Parametric Test

Small Sample T Test

The t-test for the mean is a statistical test for a population mean. The t-test can be used when the population is normal or nearly normal, □ is unknown, and n < 30.

The **test statistic** is the sample mean $\bar{x}$ and the **standardized test statistic** is *t*.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

The degrees of freedom are d.f. = $n - 1$.

Unit 2: Parametric Test

Proportion T Test

The z-test for a population is a statistical test for a population proportion. The z-test can be used when a binomial distribution is given such that np 5 and nq 5.

The **test statistic** is the sample proportion $\hat{p}$ and the **standardized test statistic** is *z*.

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

Unit 3: NonParametric Test

Runs Test

Rather than try to guess whether the data of a sample have been selected at random, statisticians have devised a nonparametric test to determine randomness. This test is called the runs test.

a. There are four runs, as shown.

MM FFF M FF

1 2 3 4

b. There are three runs, as shown.

H T HHH

1 2 3

c. There are six runs, as shown.

A B AAA BB A BBB

1 2 3 4 5 6

Unit 3: NonParametric Test

Sign Test

When using the sign test, the researcher hypothesizes the specific value for the median of a population; then he or she selects a sample of data and compares each value with the conjectured median.

If it is below the conjectured median, it is assigned a minus sign.
And if it is exactly the same as the conjectured median, it is assigned a 0.

Unit 3: NonParametric Test

Wilcoxon Rank Sum Test

The Wilcoxon tests consider differences in magnitudes by using ranks.

$$Z = \frac{R - \mu_R}{\sigma_R} \quad \text{where} \quad \mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} \quad \sigma_R = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$$

R sum of ranks for smaller sample size (n_1 or n_2)
$n1$ First sample size
$n2$ Second sample size
$n1 \geq 10$ and $n2 \geq 10$
Note that if both samples are the same size, either size can be used as $n1$.

Unit 3: NonParametric Test

Wilcoxon Sign Sum Test

When the samples are dependent, as they would be in a before- and-after test using the same subjects, the Wilcoxon signed-rank test can be used in place of the t test for dependent samples. Again, this test does not require the condition of normality. Wilcoxon Signed-rank Table is used to find the critical values.

Step 4 Make the decision. Reject the null hypothesis if the test value is less than or equal to the critical value.

Unit 4: Goodness of fit Test

Independence test

Homogeneity test

2x2 contigency test

Yates correction

Unit 4: Goodness of fit Test

Chi- square Goodness of Fit test

Goodness of fit refers to how close the observed data are to those predicted from a hypothesis

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei}$$

Unit 4: Goodness of fit Test

Test for Independence

The absence of association between two cross-tabulated variables. The percentage distributions of the dependent variable within each category of the independent variable are identical.

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei}$$

Unit 4: Goodness of fit Test

Homogeneity Test

To test whether the proportions of elements that have a common characteristic are the same for each population

$$\chi^2 = \sum \frac{(Oi - Ei)^2}{Ei}$$

Unit 4: Goodness of fit Test

2X2 Contingency Table

To test whether the proportions of elements that have a common characteristic are the same for each population

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi^2(1)df$$

Unit 4: Goodness of fit Test

Yates Correction Test in 2x2 contigency

To test whether the proportions of elements that have a common characteristic are the same for each population

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi^2(1)df$$

Unit 5: Sample Estimation

Errors in sample

Type 1 error:

- The probability of finding a difference when compared our sample with population, and in reality there is no difference
- Known as the α (or "type 1 error")

Type 2 error:

- The probability of not finding a difference thatactually exists between two groups (or betweensample and population).
- Known as the β (or "type 2 error")

Estimation of Sample Size

(1) One sample (single mean):

The one-sample estimation of sample size involves determining the number of observations needed to estimate a population mean with a specific level of confidence and precision, considering factors like standard deviation, margin of error, and confidence level.

$$n = \left(\frac{zs}{d} \right)^2$$

where:
n = sample size
z = confidence interval in standard error units
s = standard error of the mean
d= acceptable magnitude of error

(2) Sample Size for a Proportion:

The sample size for a proportion refers to the number of observations required to estimate a population proportion with a specified level of confidence and margin of error.

$$n = \frac{Z^2 pq}{E^2}$$

Where:
n = number of items in samples
Z2 = square of confidence interval in standard error units
p = estimated proportion of success
q = (1-p) or estimated the proportion of failures
E2 = square of maximum allowance for error between true

Two Sample test:

For two means :

A two-sample test for two means is a statistical analysis comparing the means of two independent groups to determine if there is a significant difference between them.

For the hypothesis: $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

$$N = n_1 + n_2 = \frac{4\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{(d = \mu_1 - \mu_2)^2}$$

Two Sample test:

Two-sample proportions:

A two-sample test for two proportions is a statistical method used to compare the proportions of two independent groups and determine if there is a significant difference between them.

For the hypothesis: $H_0 : P_1 = P_2 \quad \text{vs.} \quad H_1 : P_1 \neq P_2$

$$N = n_1 + n_2 = \frac{4(z_{1-\alpha/2} + z_{1-\beta})^2 \left[\left(\frac{P_1 + P_2}{2} \right) \left(1 - \frac{P_1 + P_2}{2} \right) \right]}{\left(d = P_1 - P_2 \right)^2}$$