

Lung Capacity and Smoking in Youth

The health and smoking habits of 725 youth

Details

The data give information on the health and smoking habits of a sample of 725 youths, aged 3 to 19, in East Boston during middle to late 1970s.

A data frame with 726 observations on the following 5 variables. (The data frame `lungcap` contains the data only for smokers, and hence does not contain the variable `Smoke`.)

Age(years): the age of the subject in completed years; a numeric vector

LungCap(cc): the forced expiratory volume in litres, a measure of lung capacity; a numeric vector

Height(Inches): the height in inches; a numeric vector

Gender: the gender of the subjects: a numeric vector with females coded as 0 and males as 1

Smoke: the smoking status of the subject: a numeric vector with non-smokers coded as 0 and smokers as 1

Caesarean: Yes or No

Data could be easy accessed within R

```
>data(lungcap)
```

Minimum Steps for exploration:

1. Importing the dataset into R
2. Understanding the structure of dataset
3. Graphical exploration
4. Descriptive statistics
5. Insights from the dataset

Graphics:

`hist(x)` – histogram of data in `x`.

`stem(x)` – stem and leaf plot of data in `x`.

`plot(x,y)` – scatter plot of `y` against `x`.

`lines(supsmu(x,y))` – add smoother to existing scatter plot.

`boxplot(list(x1,x2,...))` – side-by-side boxplots of variables `x1`, `x2`, etc.

`boxplot(y ~ x)` – alternative method for boxplots if `y` is quantitative and `x` is categorical.

`barplot(x)` – barplot of `x` (where `x` contains the heights of the bars).

`abline(a,b)` – add the line $y = a + bx$ to an existing plot.

`abline(h=a)` – add a horizontal line at $y = a$ to an existing plot.

`abline(v=a)` – add a vertical line at $x = a$ to an existing plot.

`abline(model.fit)` – add a regression line based on the model `model.fit` to an existing plot.

`qqnorm(x)` – normal probability plot of data in `x`.

`qqline(x)` – adds a line to a normal probability plot passing through 1Q and 3Q

Probability distribution computations:

`dbinom(x, n, p)` – $P(X = x)$ where $X \sim B(n, p)$

`pnorm(x, mean, sd)` – $P(X < x)$ where $X \sim N(\text{mean}, \text{sd})$

`qnorm(p, mean, sd)` – the value of x in $p = P(X < x)$, where $X \sim N(\text{mean}, \text{sd})$

`pt(x, df)` – $P(X < x)$ where $X \sim t(\text{df})$

`qt(p, df)` – the value of x in $p = P(T < x)$, where $T \sim t(\text{df})$

`pchisq(x, df)` – $P(X^2 < x)$ where $X^2 \sim \chi^2(\text{df})$

Random sampling (without replacement):

`sample(n)` – a random arrangement of the first `n` positive integers.

`sample(n, size)` – a random sample of `size` values from among the first `n` positive integers.

Graphics:

`hist(x)` – histogram of data in `x`.

`stem(x)` – stem and leaf plot of data in `x`.

`plot(x,y)` – scatter plot of `y` against `x`.

`lines(supsmu(x,y))` – add smoother to existing scatter plot.

`boxplot(list(x1,x2,...))` – side-by-side boxplots of variables `x1`, `x2`, etc.

`boxplot(y ~ x)` – alternative method for boxplots if `y` is quantitative and `x` is categorical.

`barplot(x)` – barplot of `x` (where `x` contains the heights of the bars).

`abline(a,b)` – add the line $y = a + bx$ to an existing plot.

`abline(h=a)` – add a horizontal line at $y = a$ to an existing plot.

`abline(v=a)` – add a vertical line at $x = a$ to an existing plot.

`abline(model.fit)` – add a regression line based on the model `model.fit` to an existing plot.

`qqnorm(x)` – normal probability plot of data in `x`.

`qqline(x)` – adds a line to a normal probability plot passing through 1Q and 3Q

Probability distribution computations:

`dbinom(x, n, p)` – $P(X = x)$ where $X \sim B(n, p)$

`pnorm(x, mean, sd)` – $P(X < x)$ where $X \sim N(\text{mean}, \text{sd})$

`qnorm(p, mean, sd)` – the value of x in $p = P(X < x)$, where $X \sim N(\text{mean}, \text{sd})$

`pt(x, df)` – $P(X < x)$ where $X \sim t(\text{df})$

`qt(p, df)` – the value of x in $p = P(T < x)$, where $T \sim t(\text{df})$

`pchisq(x, df)` – $P(X^2 < x)$ where $X^2 \sim \chi^2(\text{df})$

Random sampling (without replacement):

`sample(n)` – a random arrangement of the first `n` positive integers.

`sample(n, size)` – a random sample of `size` values from among the first `n` positive integers.