R Dataframe Build Vignette: WORK IN PROGRESS

Oscar A. Trevizo

2023-05-03

Contents

WORK IN PROGRESS

Load libraries

```
# library(DAAG)
# library(ggplot2)
# library(dplyr)
# library(caret)
```

Build a dataframe

```
#R code here

# Let's assign our 30 experiments to a variable "N".
N <- 30
S <- 100

# Initialize vector AVG of size 30
AVG <- numeric(N)

# Run a loop that gets a new data sample, calculates the mean and sd, and stores the result in AVG.
for (i in 1:N) {
  # AVG is an indexed vector that will take values from A[1] to A[30] as we loop through.
  # Store the mean of a Normally distributed sample of size S into AVG[i]. It will happen N times.
  AVG[i] <- mean(rnorm(S))
}
```

```
# Coming out of the loop, we have a vector AVG with 30 values.
# Calculate the standard deviation of AVG.
SIGMA <- sd(AVG)

# Display SD below
paste('The standard deviation of AGG is ', round(SIGMA, 2))
```

[1] "The standard deviation of AGG is 0.08"

```
#R code here

####
#
# 1C Strategy
#
# 1. Build our dataset.
# 2. Plot the data.
#
# To build our dataset:
# Create a dataframe with four columns, one for each sample.
# We already obtained the first AVG experiment from above. That will be my first dataframe column.
# Next we will bind 3 more experiments by column to create the 4 column dataframe:
# The dataframe will have dimensions 30 observations by 4 experiment.

# Assign the first experiment (the first 30 means of 100 Normally distributed values).

par(mfrow = c(4, 2))

df <- as.data.frame(AVG)


# Now we loop through the next 3 columns (i.e., the next 3 experiments).
for (i in 1:3) {
  # Run a loop that gets a new data sample, calculates the mean and sd, and stores the result in AVG.
  # First get an N vector AVG with the means of random numbers
  for (j in 1:N) {
    # AVG is an indexed vector that will take values from A[1] to A[30] as we loop through.
    # Store the mean of a Normally distributed sample of size S into AVG[i]. It will happen N times.
    AVG[j] <- mean(rnorm(S))
  }
  TRIAL <-as.data.frame(AVG)
  df <- cbind(df, TRIAL)
}

colnames(df) <- c('Trial_1', 'Trial_2', 'Trial_3', 'Trial_4')

# Normal probability plot is also known as a QQ plot.
qqnorm(df$Trial_1, ylim=c(-0.2, 0.2), col='blue')
qqline(df$Trial_1, ylim=c(-0.2, 0.2), col='red')
boxplot(df$Trial_1)

qqnorm(df$Trial_2, ylim=c(-0.2, 0.2), col='blue')
qqline(df$Trial_2, ylim=c(-0.2, 0.2), col='red')
```

```r
boxplot(df$Trial_2)

qqnorm(df$Trial_3, ylim=c(-0.2, 0.2), col='blue')
qqline(df$Trial_3, ylim=c(-0.2, 0.2), col='red')
boxplot(df$Trial_3)

qqnorm(df$Trial_4, ylim=c(-0.2, 0.2), col='blue')
qqline(df$Trial_4, ylim=c(-0.2, 0.2), col='red')
boxplot(df$Trial_4)
```

Normal Q–Q Plot

Normal Q–Q Plot

Normal Q–Q Plot

Normal Q–Q Plot

4