# Lab4

## Завантаження пакетів

```
if (!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(pacman, ggplot2,
   plotly, rio, rmarkdown, moments, agricolae, corrplot, tidyverse, corrr, model4you,
   Metrics, readxl, MASS, stats, dplyr)
```

```
library(pacman)
library(moments)
library(agricolae)
library(corrplot)
library(tidyverse)
library(corrr)
library(model4you)
library(Metrics)
library(readxl)
library(MASS)
library(stats)
library(dplyr)
library(ggplot2)
```

## Завантаження даних

```
# CSV
data_csv <- import("/Users/victoria/Downloads/healthcare-dataset-stroke-data.csv")
```

## 1. Опис даних

Цей набір даних містить дані про характеристики здоров'я людей та передбачення, чи був у них інфаркт.

Розглянемо перші рядки датасету:

```
head(data_csv)
```

```
##      id gender age hypertension heart_disease ever_married     work_type
## 1  9046   Male  67            0             1          Yes       Private
## 2 51676 Female  61            0             0          Yes Self-employed
## 3 31112   Male  80            0             1          Yes       Private
## 4 60182 Female  49            0             0          Yes       Private
## 5  1665 Female  79            1             0          Yes Self-employed
## 6 56669   Male  81            0             0          Yes       Private
##   Residence_type avg_glucose_level  bmi  smoking_status stroke
## 1          Urban            228.69 36.6 formerly smoked      1
## 2          Rural            202.21  N/A    never smoked      1
## 3          Rural            105.92 32.5    never smoked      1
## 4          Urban            171.23 34.4          smokes      1
## 5          Rural            174.12   24    never smoked      1
## 6          Urban            186.21   29 formerly smoked      1
```

```
data_csv$work_type = as.factor(data_csv$work_type)
data_csv<-group_by(data_csv, data_csv$work_type)
```

Для змінної work_type виведемо у вигляді лінійних діаграм значення середнього рівня глюкози в крові по кожній з градацій та значення середніх.

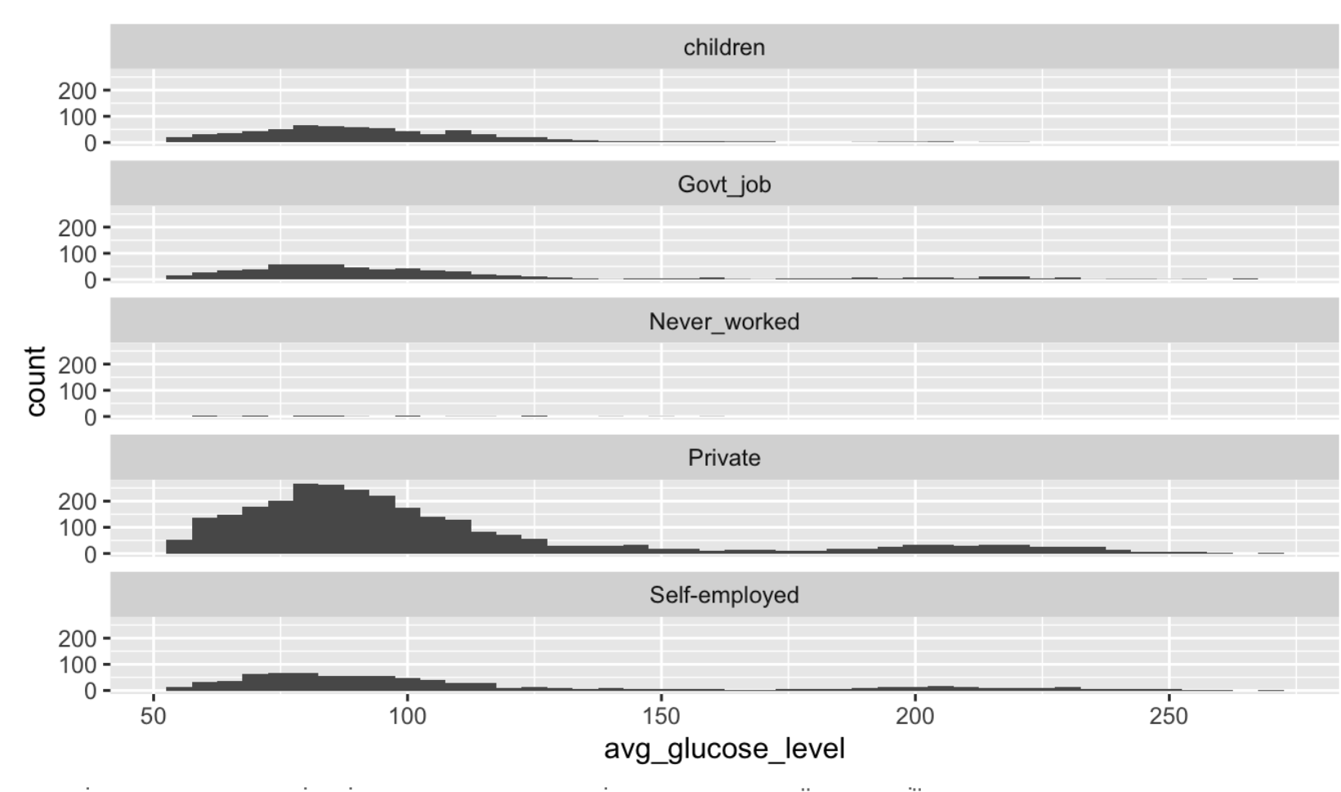
```
# stripchart with mean values
p <- ggplot(data_csv, aes(x=avg_glucose_level, y = work_type)) + geom_jitter(aes(color=work_type)) +
labs(title="Plot of work type by average glucose level") +
stat_summary(fun=mean, geom="point", shape = 18, size = 3, aes(color="mean glucose level")) +
scale_color_manual(values=c("tomato", "darkslategray3", "darkorchid4", "darkseagreen4", "grey", "orange"))
p
```

## 2. Дисперсійний аналіз

Проведемо дисперсійний аналіз для цих даних. Перевіримо виконання умов ДА: нормальність залишків та гомогенність дисперсій по групах.

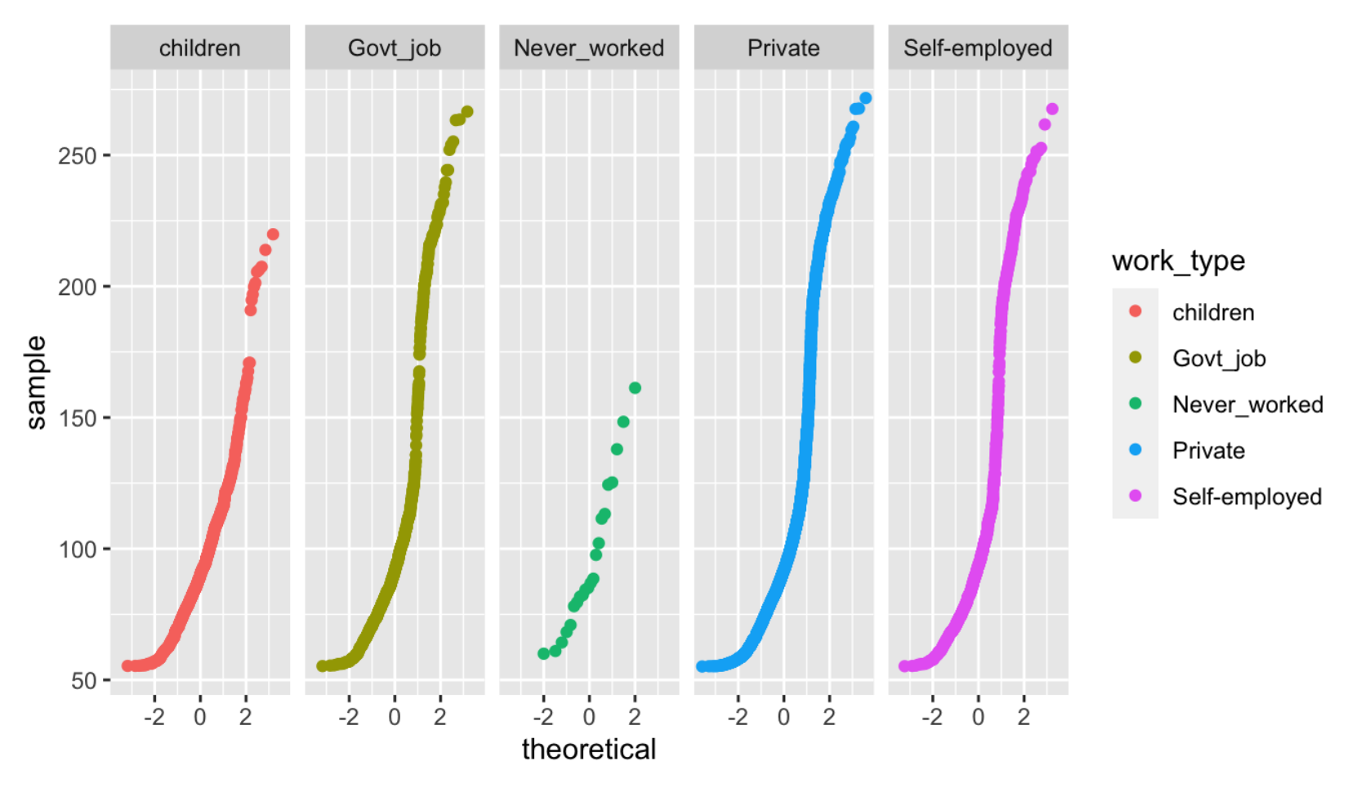
Візуальний тест на нормальність:

```
ggplot(data_csv, aes(x = avg_glucose_level)) + geom_histogram(binwidth=5) + facet_wrap(~ work_type, ncol = 1)
```

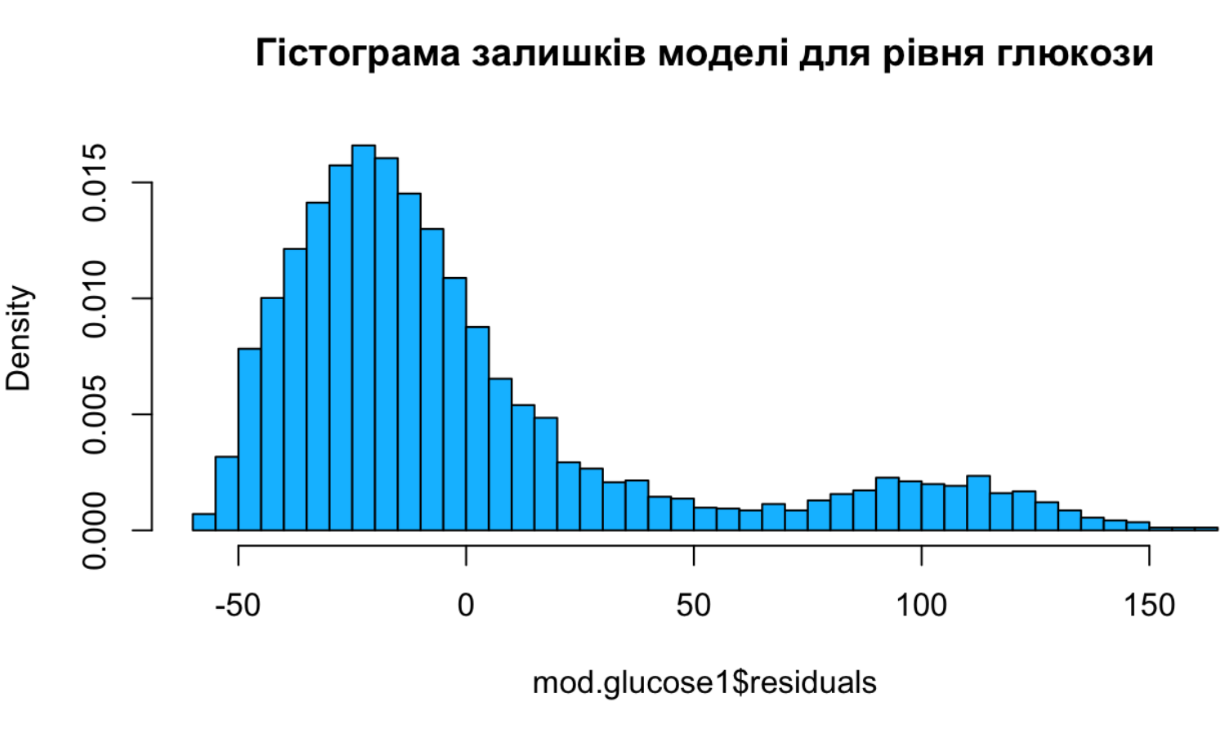


Зобразимо Q-Q діаграму для перевірки узгодженості з гіпотетичним розподілом для кожної градації.

```
library(ggplot2)
ggplot(data_csv, aes(sample = avg_glucose_level, col = work_type))+
  geom_qq()+
  facet_grid( ~ work_type)
```

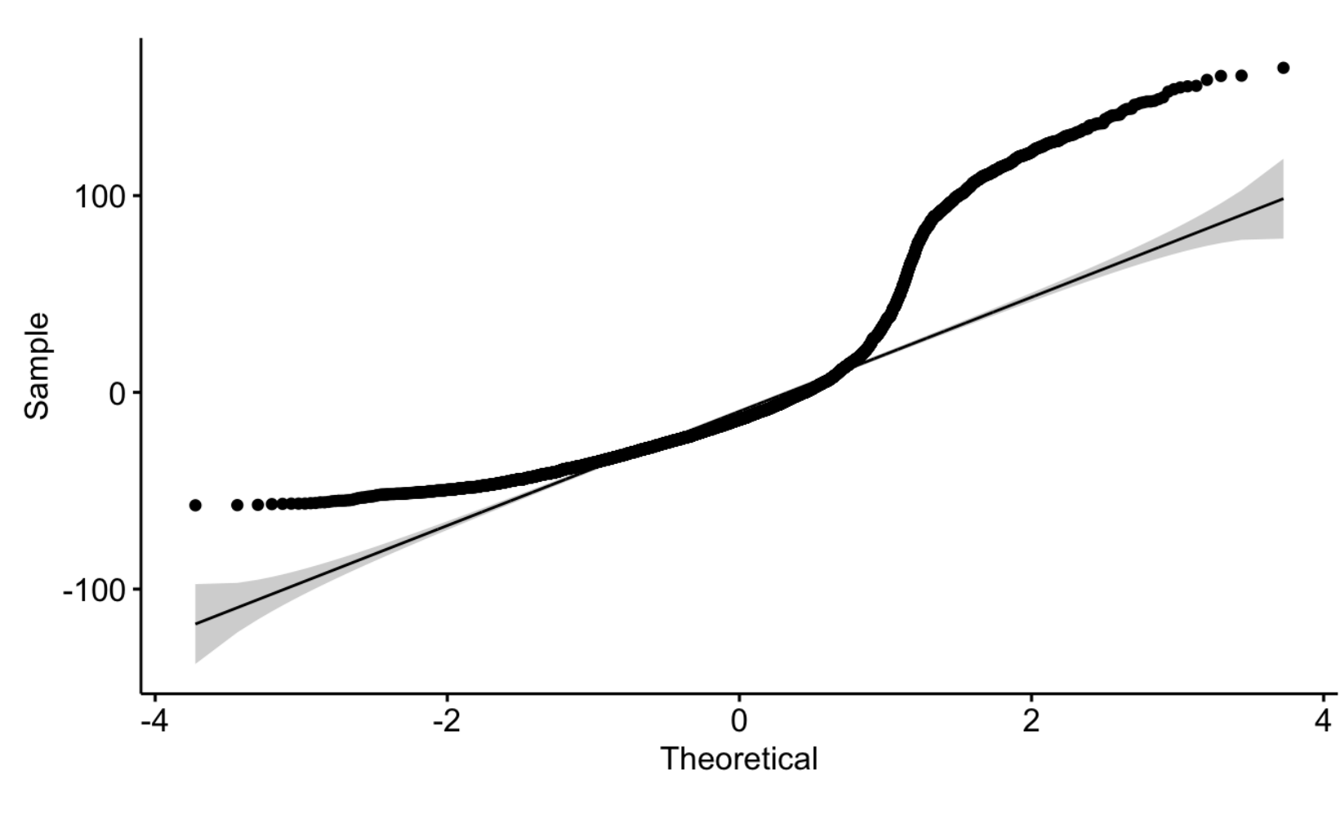


```
mod.glucose1 <- aov(avg_glucose_level ~ work_type, data = data_csv)
hist(mod.glucose1$residuals, breaks = "FD",
     main = "Гістограма залишків моделі для рівня глюкози", col = "deepskyblue", freq = F)
```



Зобразимо Q-Q діаграму для перевірки узгодженості з гіпотетичним розподілом.

```
library(ggpubr)
ggqqplot(mod.glucose1$residuals)
```



За допомогою статитичного критерію перевіряємо згоду з гіпотетичним розподілом.

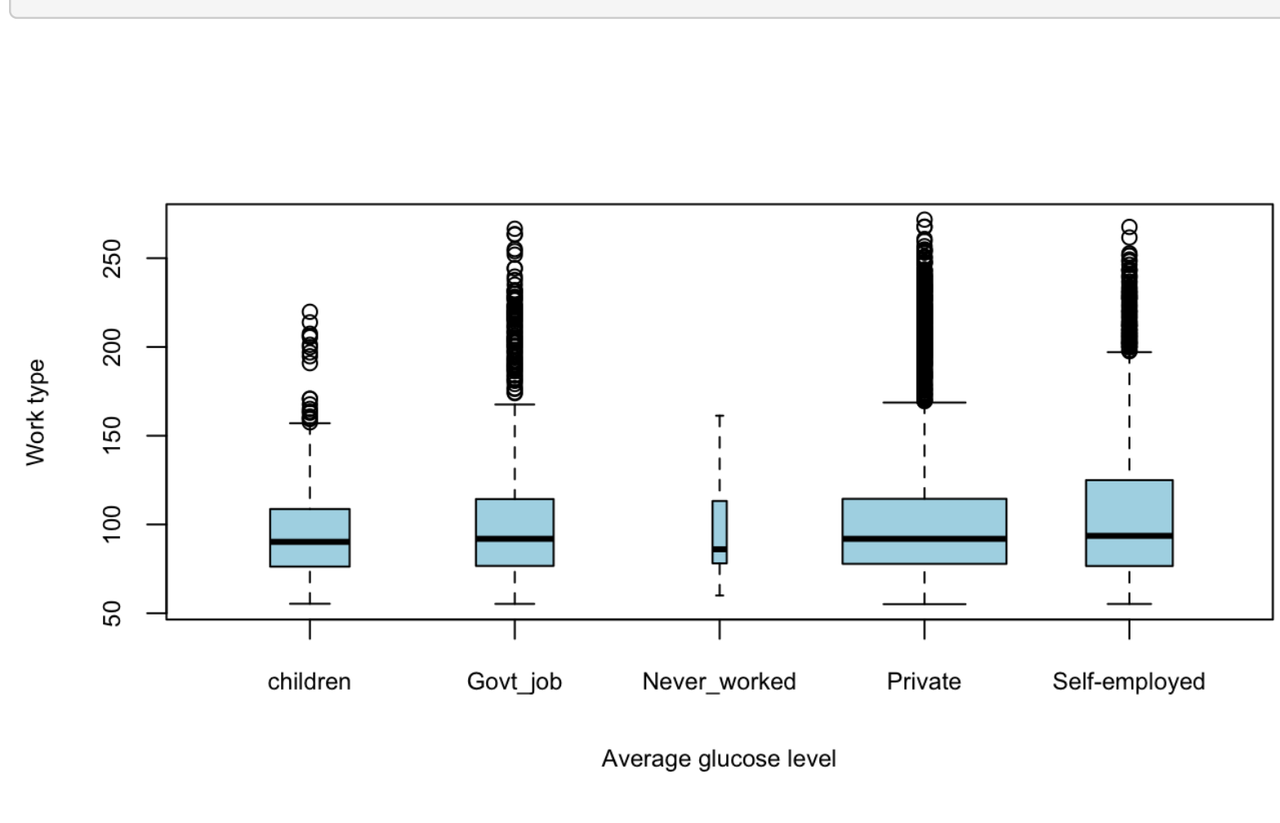
```
subsample <- sample(mod.glucose1$residuals, size=500, replace=TRUE)
shapiro.test(subsample)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subsample
## W = 0.81948, p-value < 2.2e-16
```

Бачимо, що порушена умова нормальності залишків.

Перевірка умови гомогенності дисперсій по групах

```
boxplot(avg_glucose_level~work_type, horizontal = FALSE, data = data_csv, xlab = "Average glucose level", ylab = "Wo
```



Дисперсії в усіх групах не однакові. Перевірка по Бартлету це підтверджує:

```
bartlett.test(avg_glucose_level~work_type, data = data_csv)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  avg_glucose_level by work_type
## Bartlett's K-squared = 319.82, df = 4, p-value < 2.2e-16
```

Якщо кожна не розподілена нормально та дисперсії не однакові, можемо використати тест Крускала-Валліса. https://www.statology.org/anova-unequal-sample-size/

```
kruskal.test(avg_glucose_level ~ work_type, data = data_csv)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  avg_glucose_level by work_type
## Kruskal-Wallis chi-squared = 17.336, df = 4, p-value = 0.001663
```

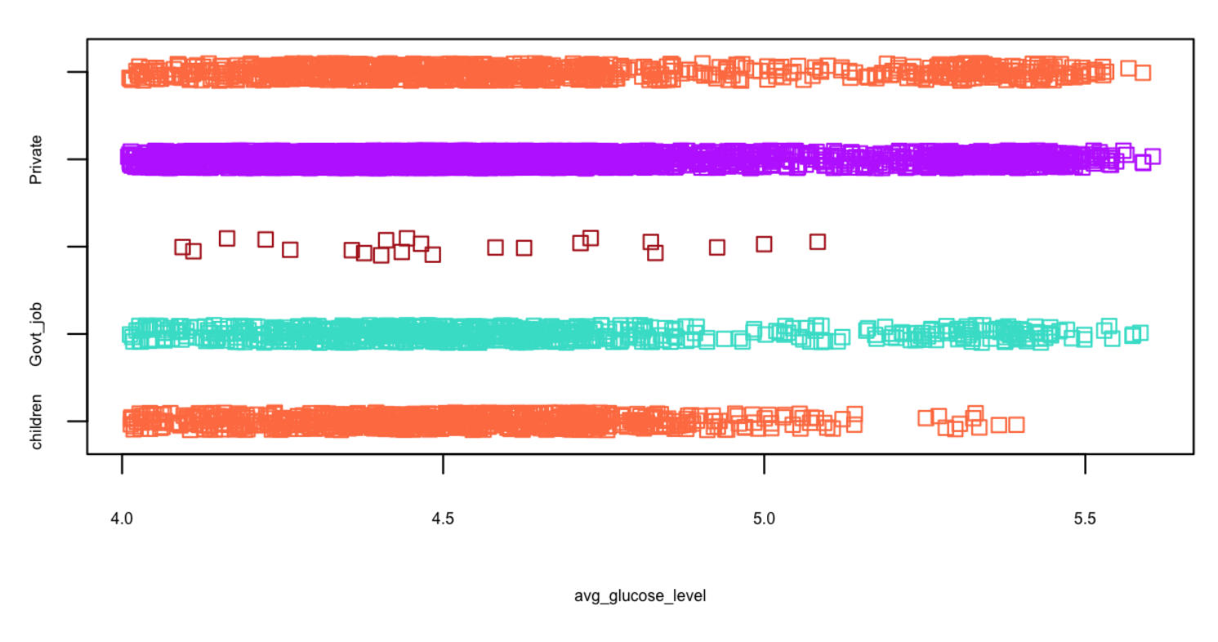
Інтерпретація: Оскільки p-значення менше рівня значущості 0.05, можна зробити висновок, що між групами типу роботи існують значні відмінності.

```
# Дисперсійний аналіз з поправкою Велча для випадку порушення умови гомогенності дисперсій (не працює для ненормальн

oneway.test(avg_glucose_level ~ work_type, data=data_csv)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  avg_glucose_level and work_type
## F = 31.478, num df = 4.00, denom df = 160.78, p-value < 2.2e-16
```

```
# Прологарифмуємо дані та побудуємо нову модель.
glucose.log <- data_csv
glucose.log$avg_glucose_level <- log(glucose.log$avg_glucose_level)
stripchart(avg_glucose_level ~ work_type, data=glucose.log, method = "jitter",
           col = c("coral", "turquoise", "firebrick", "darkorchid1"),
           cex.lab=0.5, cex.axis=0.5, cex.main=0.5, cex.sub=0.5)
```



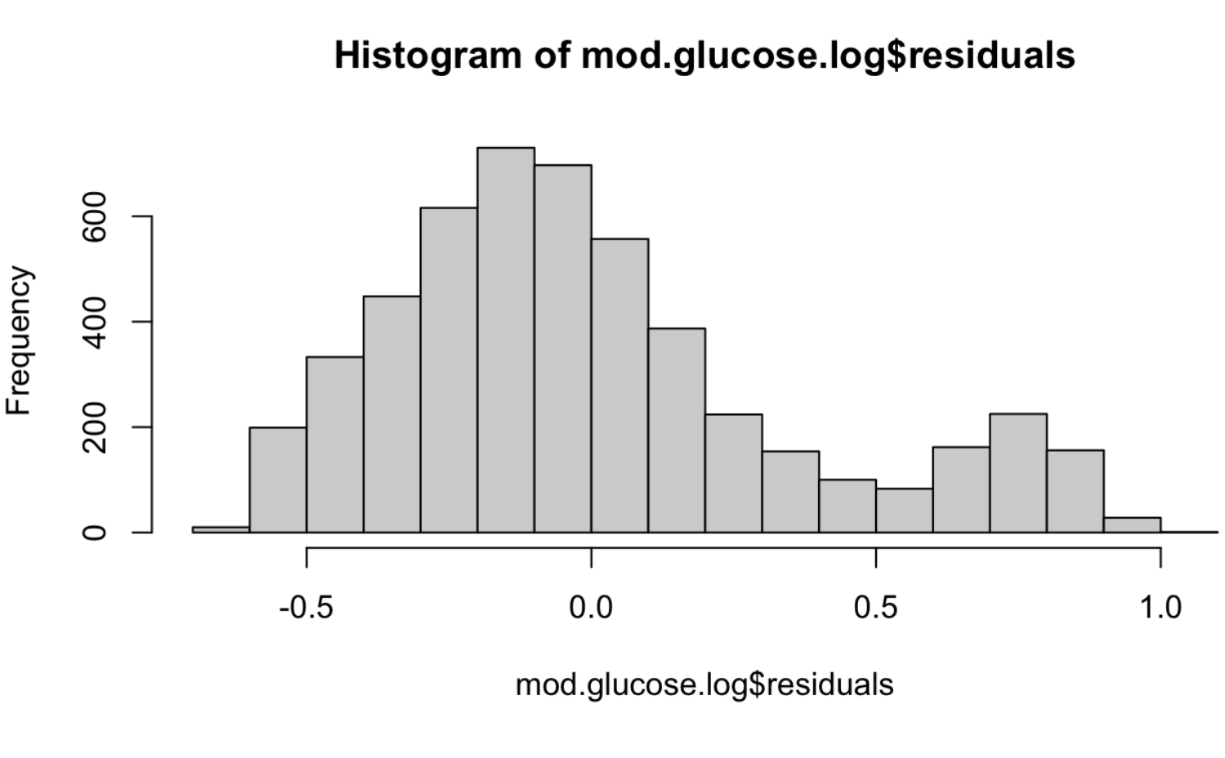
```
mod.glucose.log <- aov(avg_glucose_level ~ work_type, data=glucose.log)

#aov(Butterfat ~ Age, data=butter.log)

#summary(mod.butter.log)

#summary(lm(Butterfat ~ Breed, data=butter.log))

hist(mod.glucose.log$residuals)
```



```
subsample2 <- sample(mod.glucose.log$residuals, size=500, replace=TRUE)
shapiro.test(subsample2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subsample2
## W = 0.93286, p-value = 3.352e-14
```

```
bartlett.test(avg_glucose_level ~ work_type, data=glucose.log)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  avg_glucose_level by work_type
## Bartlett's K-squared = 140.42, df = 4, p-value < 2.2e-16
```

## 3. Аналіз контрастів

```
model<-lm(avg_glucose_level~work_type, data=data_csv)
summary(model)
```

```
##
## Call:
## lm(formula = avg_glucose_level ~ work_type, data = data_csv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.415 -29.242 -13.850   9.908 164.943
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              94.400      1.717  54.973  < 2e-16 ***
## work_typeGovt_job        13.379      2.456   5.448 5.35e-08 ***
## work_typeNever_worked     1.642      9.748   0.168    0.866
## work_typePrivate         12.397      1.908   6.496 9.01e-11 ***
## work_typeSelf-employed   18.245      2.329   7.835 5.65e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.01 on 5105 degrees of freedom
## Multiple R-squared:  0.01285,    Adjusted R-squared:  0.01208
## F-statistic: 16.61 on 4 and 5105 DF,  p-value: 1.559e-13
```

В якості базового рівня автоматично обирається група спостережень для типу роботи children (за алфавітом) – вона відповідає рядку (Intercept) в таблиці результатів аналізу. В цій групі рівень глюкози в крові дорівнює 94.400.

В другому рядку наведено інформацію про різницю між базовим рівнем (children) та типом роботи Govt_job: рівень глюкози суттєво вищий (Pr(>|t|), ніж в типі роботи Children (в середньому на 13.379, ніж в групі children).

В групі Never_worked між базовим рівнем різниця на 1.642, але це збільшення не було статистично значущим (Pr(>|t|) = 0.866).

Побудуємо матрицю контрастів для факторів:

```
contrasts(data_csv$work_type)
```

```
##               Govt_job Never_worked Private Self-employed
## children             0            0       0             0
## Govt_job             1            0       0             0
## Never_worked         0            1       0             0
## Private              0            0       1             0
## Self-employed        0            0       0             1
```

Це приклад матриці, що містить вагові коефіцієнти контрастів комбінацій умов (відносно базового рівня; середнєзначення n-го рівня мінус середнє значення базового рівня).

Контрасти сум:

```
contrasts(data_csv$work_type)<-contr.sum(n=5)
contrasts(data_csv$work_type)
```

```
##               [,1] [,2] [,3] [,4]
## children         1    0    0    0
## Govt_job         0    1    0    0
## Never_worked     0    0    1    0
## Private          0    0    0    1
## Self-employed   -1   -1   -1   -1
```

Базовий рівень, з яким порівнюються інші рівні, представляє собою середнє значення з середніх по кожній групі.

Знову побудуємо лінійну модель:

```
model2<-lm(avg_glucose_level~work_type, data=data_csv)
summary(model2)
```

```
##
## Call:
## lm(formula = avg_glucose_level ~ work_type, data = data_csv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.415 -29.242 -13.850   9.908 164.943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  103.533      2.013  51.438  < 2e-16 ***
## work_type1    -9.133      2.413  -3.785 0.000155 ***
## work_type2     4.247      2.429   1.748 0.080496 .
## work_type3    -7.490      7.701  -0.973 0.330763
## work_type4     3.264      2.113   1.544 0.122581
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.01 on 5105 degrees of freedom
## Multiple R-squared:  0.01285,    Adjusted R-squared:  0.01208
## F-statistic: 16.61 on 4 and 5105 DF,  p-value: 1.559e-13
```

Тепер перший рядок в таблиці з результатами аналізу (Intercept) містить середнє значення рівня глюкози, підраховане по середніх значеннях кожного типу роботи (загальне середнє). Далі - наскільки середні значення кожної групи відрізняються від загального середнього.

ANOVA застосувати не можемо, оскільки маємо суттєві відхилення від нормальності.

```
#mod.glucose1 <- aov(avg_glucose_level ~ work_type, data = data_csv)
#summary(mod.glucose1)

#model<-lm(avg_glucose_level~work_type, data=data_csv)
#summary(model)
```