Vocabulary IQ test: research

Taras Svystun, Dmytro Kalitin, Yaroslav Tsymbalista

12/24/2021

Introduction

Topic and aim

In this mini-research project we aim to discover the dependencies between IQ and some other characteristics (age, education). Also we are looking for links between IQ level and the degree of approval of some informative statements (e.g. "I prefer to be barefoot" or "I had an imaginary friend as a child").

Libraries

```
library(ggplot2)
library(dplyr)
library(ggpubr)
library(e1071)
library(fitdistrplus)
```

Reading the data

```
df <- read.csv("cleaned.csv", sep = "\t")
# Remove outliers
df <- df[df$age < 94,]
df <- df[df$S27 > 0.5,]
```

```
IQ <- df$score_full
country <- df$country
education <- df$education
age <- df$age
s2 <- df$S2
s5 <- df$S5
s26 <- df$S26
s27 <- df$S27
```

Data overview

```
head(df, 10)
```

```
##    score_full country S2 S5 S26 S27 education age
## 1       25.25      US  2  4   2   5         3  54
## 2       20.80      US  4  4   3   1         2  16
## 3       32.55      US  5  1   4   5         4  32
## 4       16.65      IN  5  5   1   1         4  24
## 5       39.65      US  2  1   3   3         3  63
## 6       31.20      US  4  1   0   2         3  51
## 7       16.25      AU  4  1   4   4         2  40
## 8       19.45      CH  2  3   3   3         4  45
## 9       43.65      GB  5  1   3   1         3  71
## 10      32.25      US  4  4   4   2         1  14
```

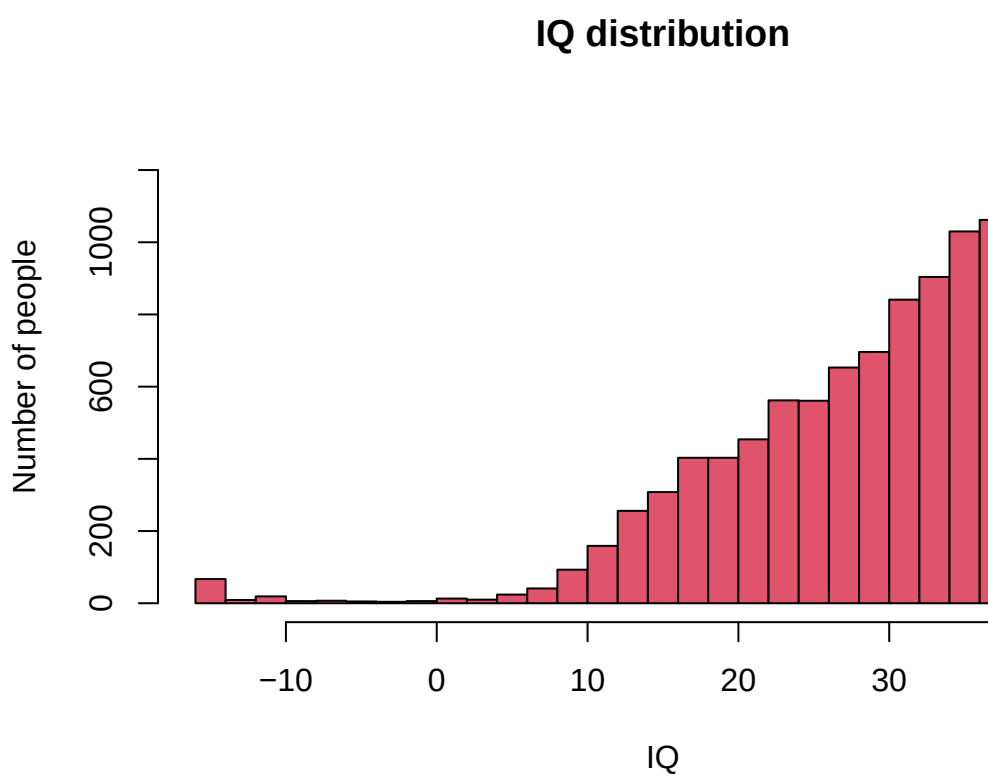
score_full - the total score of certain person for vocabulary IQ test. This test is not an ordianry IQ test. Here, people answer vocabular questions earning 1 point for right answer and loosing 0.35 for wrong.

What is s2, s5, s26, s27? These are following statements:

- S2 I avoid contacts with others.
- S5 I would be interested in getting my fortune told.
- S26 I always do the bare minimum I need to get by.
- S27 I like to play devils advocate.

Respondents answered to them in range (1=Disagree, 5=Agree). Based on them we will show you unobvious dependency.

```
set.seed(2021)
hist(IQ,  breaks = 24, main = "IQ distribution", col = "666", ylab = "Number of people")
```

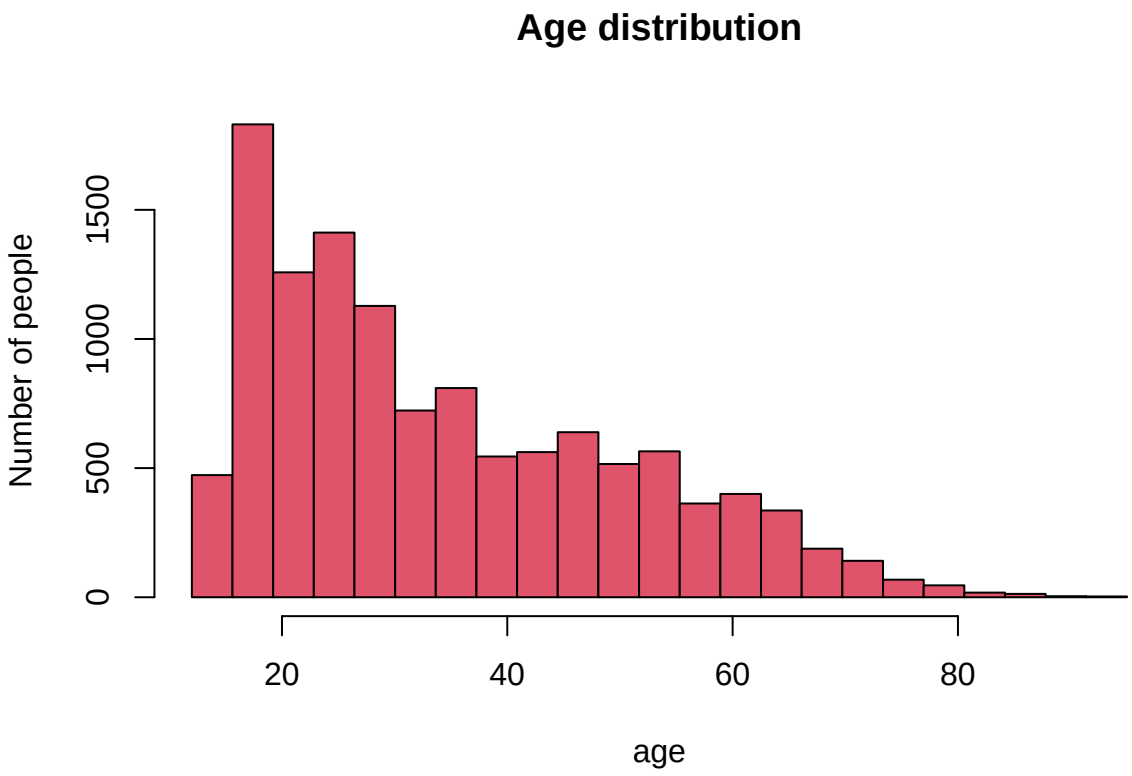


Graphical representation of data

```
summary(age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   21.00   30.00   34.52   46.00   93.00
```

```
hist(age, breaks = seq(12, 95, length = 24), main = "Age distribution", col = "666", ylab = "Number of p
```



8 most popular countries people from were surveyed in this sample

```
tail(sort(table(country)), 8)
```

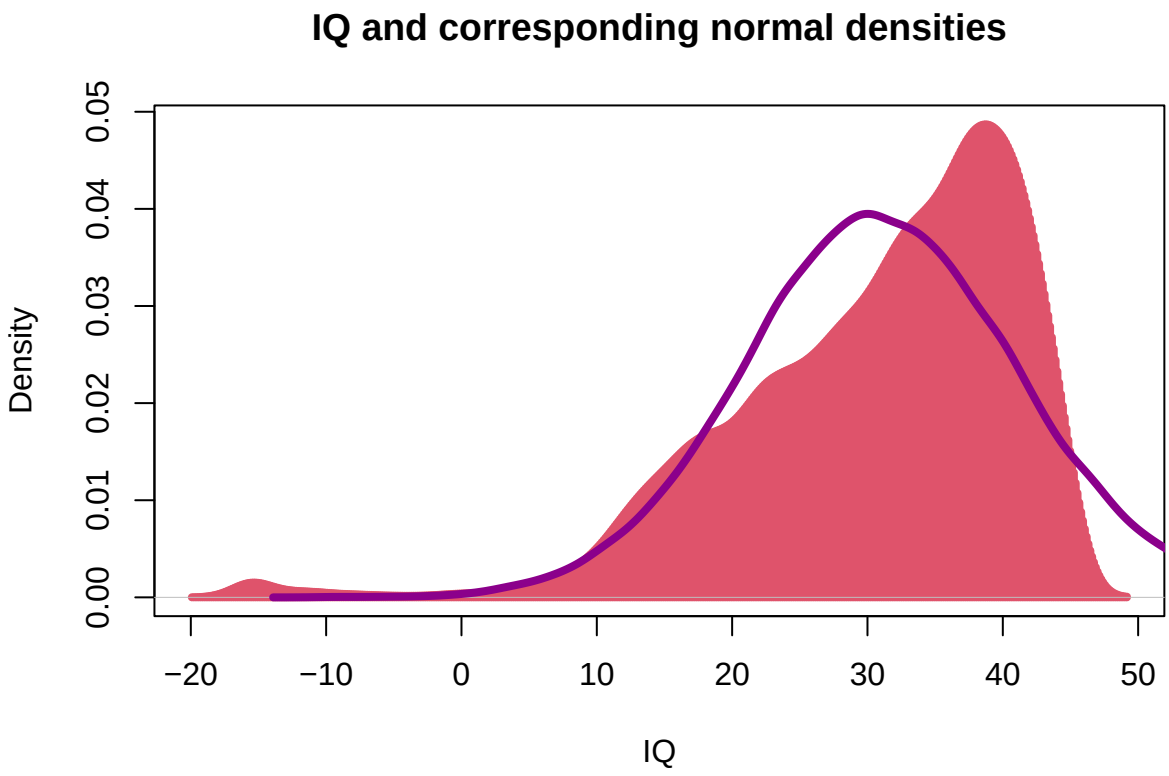
```
## country
##   NZ   DE   PH   IN   AU   CA   GB   US
##  124  184  202  232  534 1001 1186 6394
```

IQ distribution

Now it's time to analyze the IQ distribution. The first natural assumption is that IQ is normally distributed. Let us look at the densities.

```
set.seed(2021)
plot(density(IQ), main = "IQ and corresponding normal densities", col="666", lwd=4, type="h", xlab="IQ")

normal <- rnorm(12000, mean = mean(IQ), sd=sd(IQ))
lines(density(normal),col="darkmagenta", lwd=4)
```

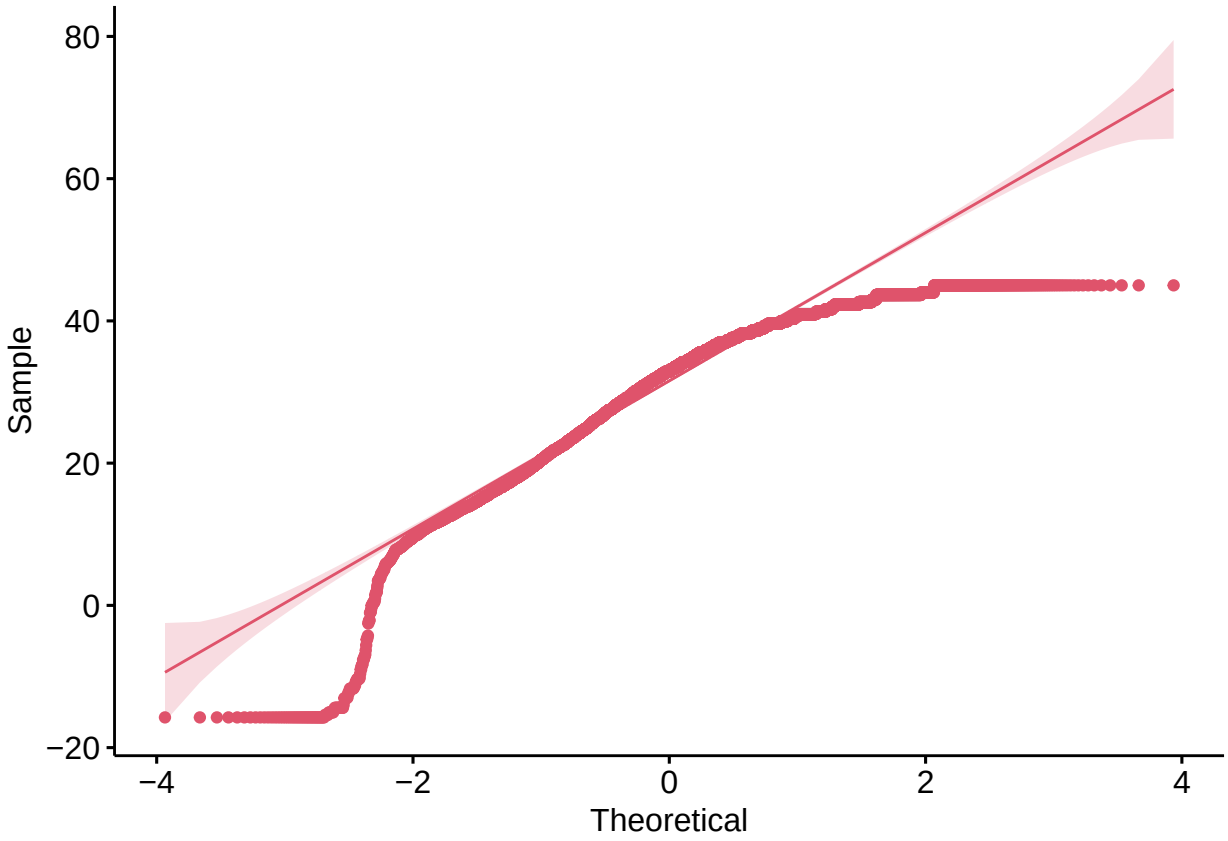



```
skewness(IQ)
```

```
## [1] -1.165102
```

It doesn't seem like normal distribution. Let's deep into details and look at the quantile-quantile plot (the straight line is corresponding to quantiles of the normal distribution).

```
ggqqplot(IQ, color = "666")
```



As we can see from the graphs, IQ fluctuates from the normal below -10 and after 40. It could be seen in 2 different ways. First one is the plot of densities. We see that IQ density has heavy left tail, which means negative skewness. We proved that by calculating skew. So it drops quickly to the right hand side, while normal distribution is symmetric (i.e. skewness is zero).

Second one is by using qq plot, which shows the correlation between quantiles of given distribution.

Finally, let us use Kolmogorov-Smirnov test and check the p-value (Note: here D is the max distance between the two CDFs; H_0 - cdf's are equal).

```
ks.test(IQ, normal, alternative = "t")
```

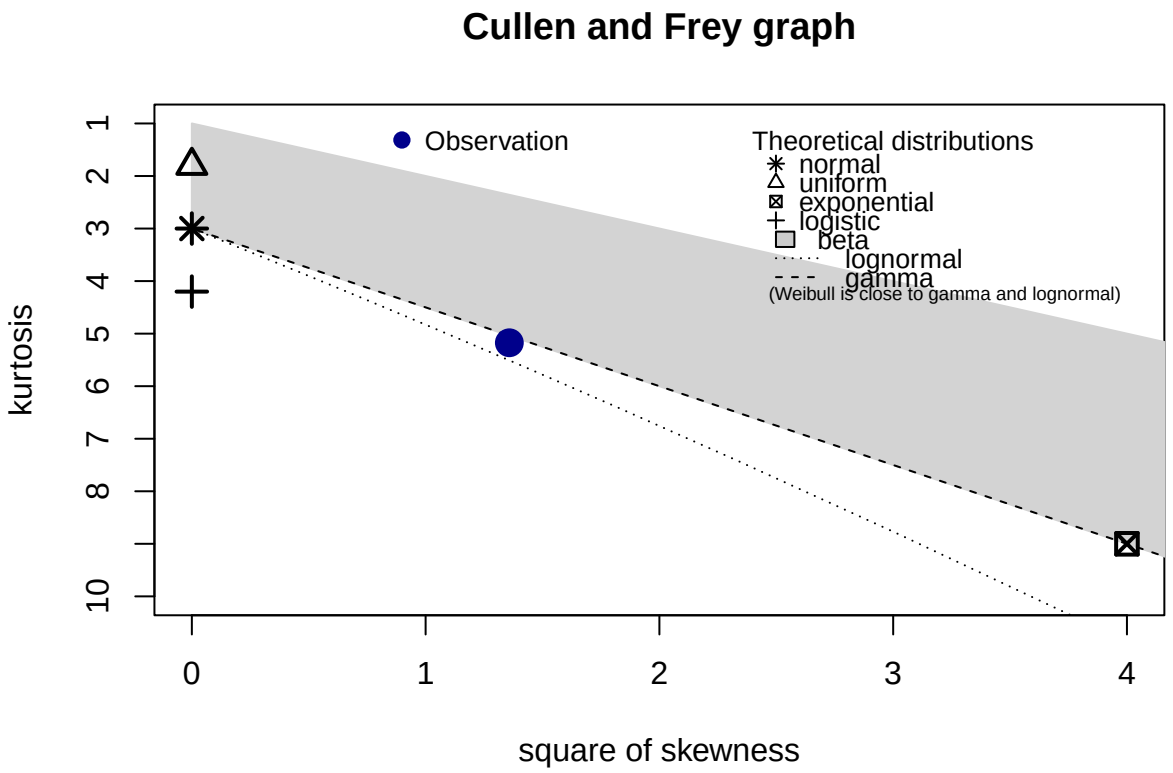
```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  IQ and normal
## D = 0.087307, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

p-value is very small, so we sure we can reject the hypothesis that IQ distribution is the normal one.

Skewness-kurtosis graph to fit the distribution

The function descdist() provides a skewness-kurtosis graph to help to choose the best candidate(s) to fit a given dataset. If we want to use it for discrete distributions we may use argument discrete=TRUE. We perform the fit on positive values of IQ score, we'll explain why latter.

```
IQ.positive <- IQ + 15.76
descdist(IQ.positive)
```




```
## summary statistics
## ------
## min:  0.01   max:  60.76
## median:  48.66
## mean:  46.6237
## estimated sd:  10.12304
## estimated skewness:  -1.165392
## estimated kurtosis:  5.173669
```

A dark-blue point on the graph represent the square of skewness and kurtosis of our observation.

We see that it is close to that ones of Gamma and lognormal distributions. Let us check whether IQ distributed as Gamma or lognormal.

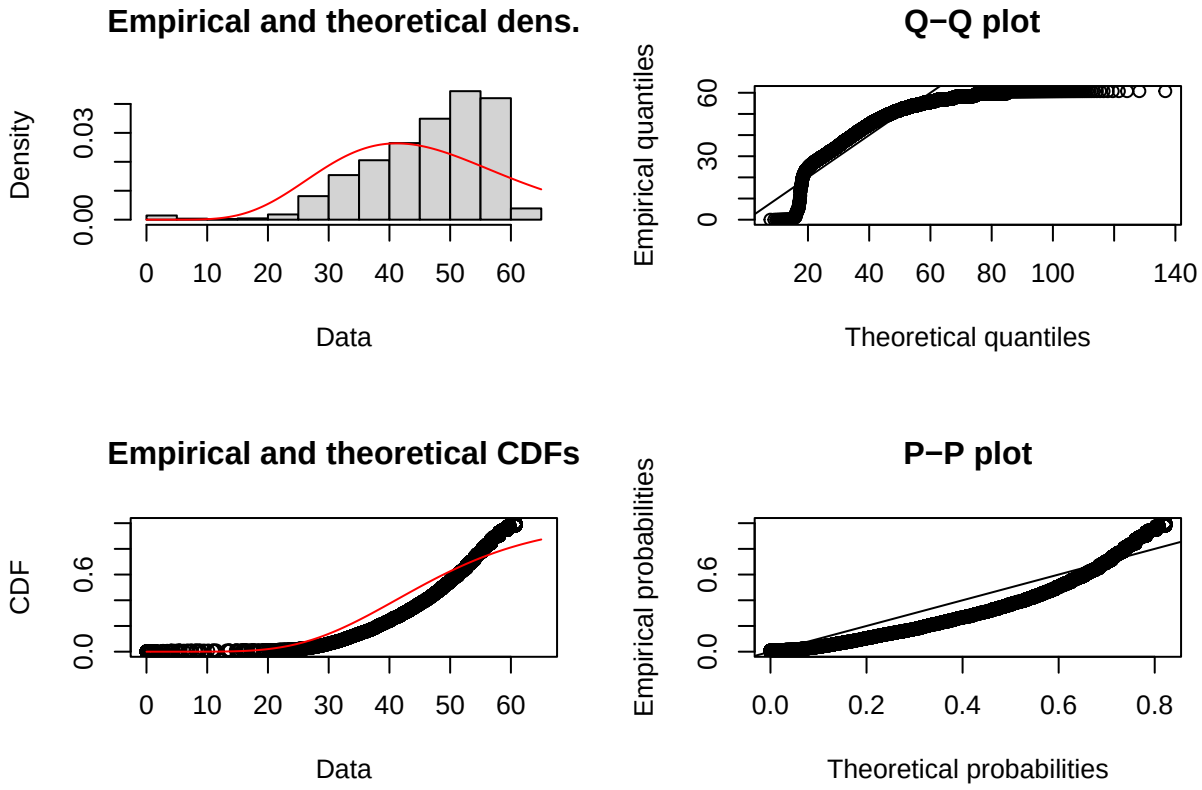
However, both distributions assume only positive values, that's why we added the minimum value of our sample to each of the values.

Function fitdist() will find for us maximum likelihood estimators for parameters to define distributions. There is the dataframe presented with estimates and standard errors of corresponding estimators. Note: PP (probability-probability) plot shows correlation between the values of theoretical (Gamma / lognormal) and emperical (IQ score) CDFs.

```
fit.gamma <- fitdist(IQ.positive, distr="gamma", method="mle")
fit.gamma
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters:
##        estimate  Std. Error
## shape 8.5819373 0.108481836
## rate  0.1840534 0.002395863
```

```
plot(fit.gamma)
```

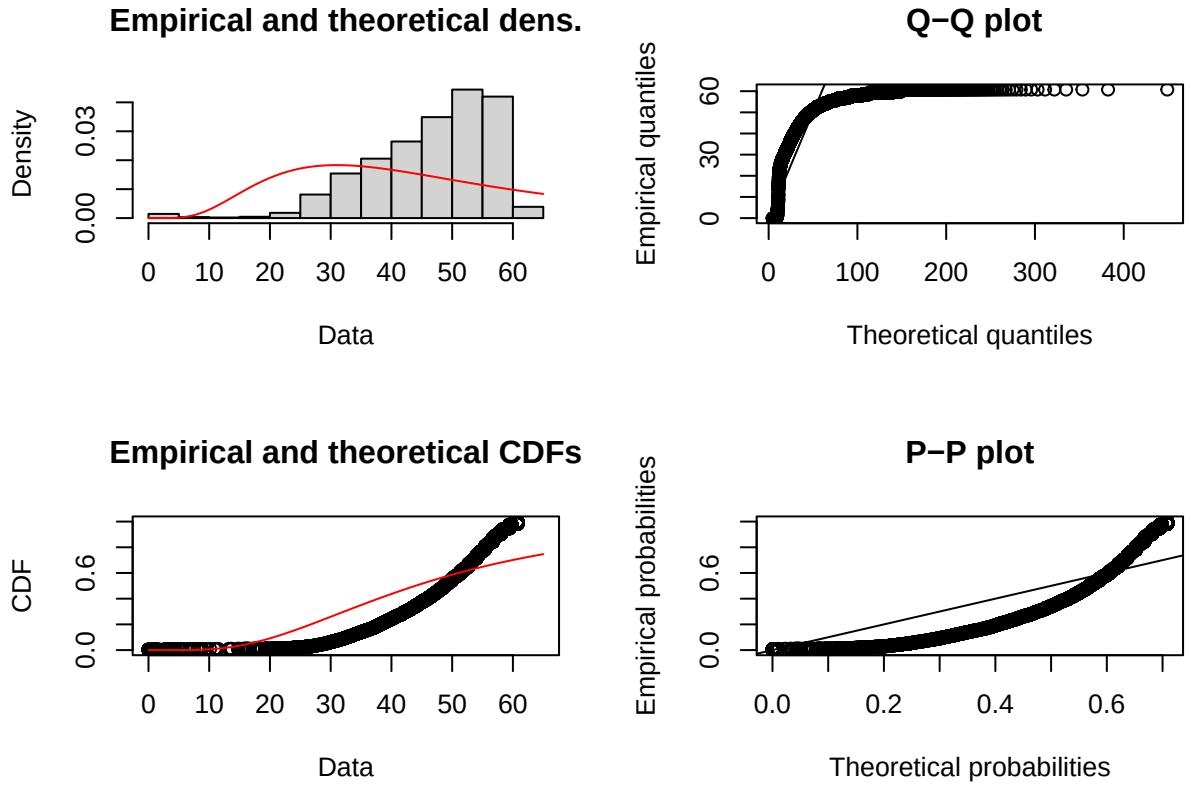


We see that our distribution of IQ test score doesn't fit quite well Gamma distribution, still it has some similarities.

```
fit.lognorm <- fitdist(IQ.positive, distr="lnorm", method="mle")
fit.lognorm
```

```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters:
##          estimate  Std. Error
## meanlog 3.7827114 0.005382280
## sdlog   0.5906301 0.003805798
```

```
plot(fit.lognorm)
```



And lognormal distribution doesn't fit event more to the observation.

As a result, we can conclude that IQ test score distribution doesn't fit any known distribution, however, it seems to be a mixture of some distributions. One of them might be Gamma or lognormal distribution.

Finding the correlations

In order to search for any dependency we used the following correlation tests: Pearson and Kendall rank correlation tests.

Pearson product-moment correlation test

The Pearson correlation coefficient is calculated by the known formula: $Corr(X,Y) = \frac{Cov(X,Y)}{(\sigma_X \sigma_Y)}$. Denote this coefficient as r, then the following t-statistic $t = r\sqrt{\frac{n-2}{1-r^2}}$ has Student distribution with $n-2$ degrees of freedom. This t-value is used to test the hypothesis whether correlation coefficient is equal to zero (Note: H_0 - correlation equal 0).

```
s2.cor1 <- cor.test(IQ, s2, method = "pearson")
s5.cor1 <- cor.test(IQ, s5, method = "pearson")
s26.cor1 <- cor.test(IQ, s26, method = "pearson")
s27.cor1 <- cor.test(IQ, s27, method = "pearson")
```

```r
age.cor1 <- cor.test(IQ, age, method = "pearson")
education.cor1 <- cor.test(IQ, education, method = "pearson")
s2.cor1
```

```
##
##  Pearson's product-moment correlation
##
## data:  IQ and s2
## t = -2.0575, df = 12040, p-value = 0.03966
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03659701 -0.00088749
## sample estimates:
##         cor
## -0.01874823
```

```r
s5.cor1
```

```
##
##  Pearson's product-moment correlation
##
## data:  IQ and s5
## t = -19.805, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1948678 -0.1602724
## sample estimates:
##       cor
## -0.177625
```

```r
s26.cor1
```

```
##
##  Pearson's product-moment correlation
##
## data:  IQ and s26
## t = -19.64, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1934444 -0.1588310
## sample estimates:
##        cor
## -0.1761922
```

```r
s27.cor1
```

```
##
##  Pearson's product-moment correlation
##
## data:  IQ and s27
## t = 24.307, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.1991891 0.2332407
## sample estimates:
##       cor
## 0.2162806
```

```
age.cor1
```

```
##
##  Pearson's product-moment correlation
##
## data:  IQ and age
## t = 48.333, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3880441 0.4179629
## sample estimates:
##       cor
## 0.4031112
```

```
education.cor1
```

```
##
##  Pearson's product-moment correlation
##
## data:  IQ and education
## t = 36.024, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2957084 0.3279559
## sample estimates:
##      cor
## 0.311922
```

As we can see, Statement2 "I avoid contacts with others" has no correlation with IQ - p-value is large enough and the coefficient r is close to zero.

But also some unexpected results were obtained.

- Statement5 "I would be interested in getting my fortune told."
- Statement26 "I always do the bare minimum I need to get by."
- Statement27 "I like to play devil's advocate."

The answers to these 3 statements correlate with IQ in some way (because p-value is very small) S27 has positive correlation: the more I think over statements I hear from various sources from different angles (even if I agree with to begin with) - the higher my IQ is. Well, pretty expected result. S5 and S26 have negative correlation: the less I am interested in getting my fortune told and the more I do my best - the higher my IQ is. And 2 expected positive correlations: age and education.

Kendall rank correlation tau

Kendall rank correlation coefficient, just as Person one, assumes values from -1 to 1, interprets in the same way and has the following formula: $\tau = \frac{(number of concordant pairs) - (number of discordant pairs)}{C_n^2}$. Where concordant pairs are such (x_i, y_j) and (y_i, y_j), so that either $x_i > y_i$ and $x_j > y_j$ or $x_i < y_i$ and $x_j < y_j$; otherwise they are said to be discordant.

Under the null hypothesis of independence of X and Y, the sampling distribution of τ has an expected value of zero. The precise distribution cannot be characterized in terms of common distributions, but may be calculated exactly for small samples; for larger samples, it is common to use an approximation to the normal distribution, with mean zero and variance $\frac{2(2n+5)}{9n(n-1)}$ (Note: H_0 - correlation equal 0).

```
s27.cor2 <- cor.test(IQ, s27, method = "kendall")
age.cor2 <- cor.test(IQ, age, method = "kendall")
education.cor2 <- cor.test(IQ, education, method = "kendall")
s27.cor2
```

```
##
##  Kendall's rank correlation tau
##
## data:  IQ and s27
## z = 22.046, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.1482618
```

```
age.cor2
```

```
##
##  Kendall's rank correlation tau
##
## data:  IQ and age
## z = 54.34, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.3350525
```

```
education.cor2
```

```
##
##  Kendall's rank correlation tau
##
## data:  IQ and education
## z = 39.887, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2756653
```

As we can see, similar results to Pearson tests.

Additional Pearson and Kendall tests

As we know, IQ correlates with education, S26 and S5. Its natural to assume that S26 and S5 correlates with level of education, lets check it. (Note: H_0 - correlation equal 0)

```
s26.cor2ed_1 <- cor.test(education, s26, method = "pearson")
s26.cor2ed_2 <- cor.test(education, s26, method = "kendall")

s26.cor2ed_1
```

```
##
##  Pearson's product-moment correlation
##
## data:  education and s26
## t = -19.903, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1957082 -0.1611236
## sample estimates:
##        cor
## -0.1784711
```

```
s26.cor2ed_2
```

```
##
##  Kendall's rank correlation tau
##
## data:  education and s26
## z = -20.111, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.1536352
```

Both methods show similar results. We can reject zero hypothesis (correlation = 0) as p-value is very small and from confidence interval we see that correlation is less than 0 in Pearson and Kendall methods.

```
ed.cor2s5_1 <- cor.test(s5, education, method = "pearson")
ed.cor2s5_2 <- cor.test(s5, education, method = "kendall")

ed.cor2s5_1
```

```
##
##  Pearson's product-moment correlation
##
## data:  s5 and education
## t = -12.345, df = 12040, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1294076 -0.0941319
## sample estimates:
##       cor
## -0.111805
```

```
ed.cor2s5_2
```

```
##
##  Kendall's rank correlation tau
##
## data:  s5 and education
## z = -11.946, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##         tau
## -0.09074052
```

Here we have the same situation as in previous case.

Important note: both Pearson and Kendall correlation coefficients test linear correlation!

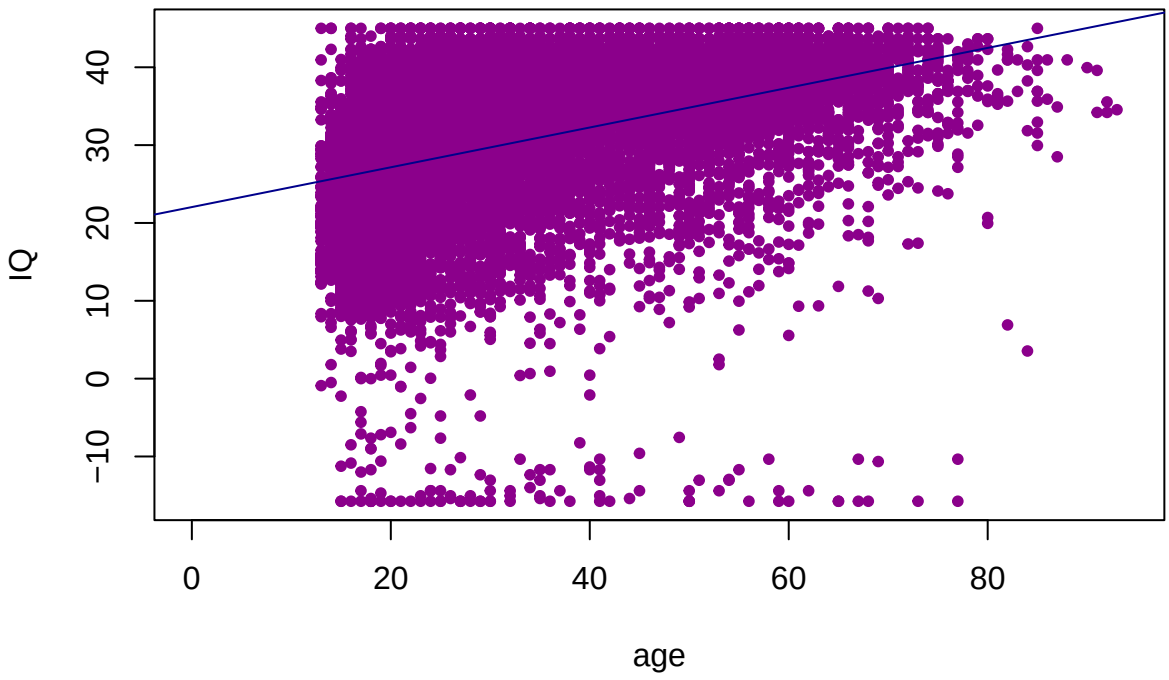
Regression analysis

Since we have found parameters which have largest correlation coefficients, we may build linear regression models to obtain more clear picture.

One essential notion is how to measure the goodness of our model. Usually, for this purpose determination coefficient r^2 is used. This measure is defined by the proportion of the total variability explained by the regression model. r^2 = Explained Variation of the model / Total variation of the model.

Let us start with age parameter.

```
age.linear <- lm(IQ~age)
plot(age, IQ, pch=20, xlim=c(0, 94), col="darkmagenta")
abline(age.linear, col="darkblue")
```

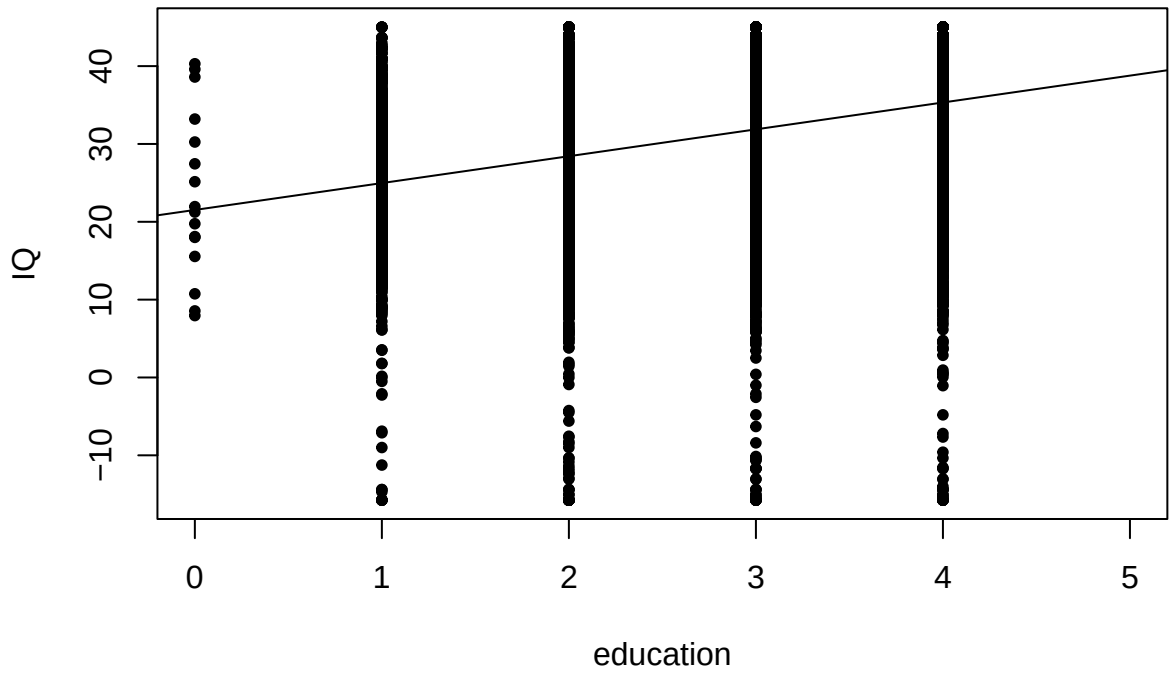



```
summary(age.linear)
```

```
##
## Call:
## lm(formula = IQ ~ age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.471  -5.059   1.397   6.470  19.637
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.040747   0.201122  109.59   <2e-16 ***
## age          0.255589   0.005288   48.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.264 on 12040 degrees of freedom
## Multiple R-squared:  0.1625, Adjusted R-squared:  0.1624
## F-statistic:  2336 on 1 and 12040 DF,  p-value: < 2.2e-16
```

The same for education.

```
edu.linear <- lm(IQ~education)
plot(education, IQ, pch=20, xlim=c(0,5))
abline(edu.linear)
```




```
summary(edu.linear)
```

```
##
## Call:
## lm(formula = IQ ~ education)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.077  -5.377   2.024   6.973  20.027
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.5222     0.2737   78.63   <2e-16 ***
## education     3.4512     0.0958   36.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.618 on 12040 degrees of freedom
## Multiple R-squared:  0.0973, Adjusted R-squared:  0.09722
## F-statistic:  1298 on 1 and 12040 DF,  p-value: < 2.2e-16
```

Obviously, such small amount of x's doesn't give us clear picture, so there is no reason to build graphs for S5, S26 and S27. In general, these two models have poor graphical representation and are very bad for IQ prediction - both determination coefficients are near 0.1 which is very small. Moreover, the estimated variance of residuals is very large in both cases which is seen in the graphs. However, this models do state that age and education positively correlate with IQ - lines have positive slopes.

Let us look at determination coefficients of s5, s26 and s27.

```
s5.linear <- lm(IQ~s5)
summary(s5.linear)
```

```
##
## Call:
## lm(formula = IQ ~ s5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.631  -5.986   2.069   7.616  16.906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.07736    0.18593   183.3   <2e-16 ***
## s5          -1.19670    0.06042   -19.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.962 on 12040 degrees of freedom
## Multiple R-squared:  0.03155,    Adjusted R-squared:  0.03147
## F-statistic: 392.2 on 1 and 12040 DF,  p-value: < 2.2e-16
```

```
s26.linear <- lm(IQ~s26)
summary(s26.linear)
```

```
##
## Call:
## lm(formula = IQ ~ s26)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.183  -6.034   2.000   7.616  17.982
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.43334    0.20317  169.48   <2e-16 ***
## s26         -1.48313    0.07551  -19.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.965 on 12040 degrees of freedom
## Multiple R-squared:  0.03104,    Adjusted R-squared:  0.03096
## F-statistic: 385.7 on 1 and 12040 DF,  p-value: < 2.2e-16
```

```r
s27.linear <- lm(IQ~s27)
summary(s27.linear)
```

```
##
## Call:
## lm(formula = IQ ~ s27)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.364  -6.124   1.921   7.376  18.066
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.26357    0.24737  102.13   <2e-16 ***
## s27          1.67014    0.06871   24.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.884 on 12040 degrees of freedom
## Multiple R-squared:  0.04678,    Adjusted R-squared:  0.0467
## F-statistic: 590.8 on 1 and 12040 DF,  p-value: < 2.2e-16
```

Determination coefficients are even smaller but we indeed see some linear correlation of this parameters with IQ - p-values of slope equal zero are very small in all cases.

Conclusion

By doing this research, our team discovered that the most challenging part of work is data collection and decision-making when there is no exact answer for the question. One of our problems was that data was not representative as participants were not well distributed by age. Another problem was that not every answer was valid. Moreover, we noticed that used questions could not represent actual IQ as many people took maximum scores. Therefore this score does not show a difference between them. This could be a reason why we had so big residuals in linear regression models. Summing up, we cannot say that our results are pretty accurate. However, still, we learned a lot and made quite essential conclusions.

After analyzing the data of IQ results, we can say that there is no evidence that the type of personality (extrovert or introvert) has any influence on IQ score. However, we found that the desire to do the bare minimum and rely on fortune is negatively correlated with education. Moreover, the level of education and the desire to think over the facts is positively correlated with IQ level. Therefore, we can conclude that IQ mostly depends on actions, not properties.