Chi-Square and t-distribution Vignette

Oscar A. Trevizo

2023-05-09

Contents

1 Build Chi-Squared and t-distribution dataset

Running x <- rchisq(n, 1) creates n random values from a chi-squared distribution and one degree of freedom.

Running x <- rt(n, 1) creates n random values from a t-distribution with one degree of freedom.

Make normal probability plots using 30 random values for various degrees of freedom from each of these distributions.

Calculate p-values using Shapiro-Wilk tests.

The H_O null hypothesis is that the data are normally distributed.

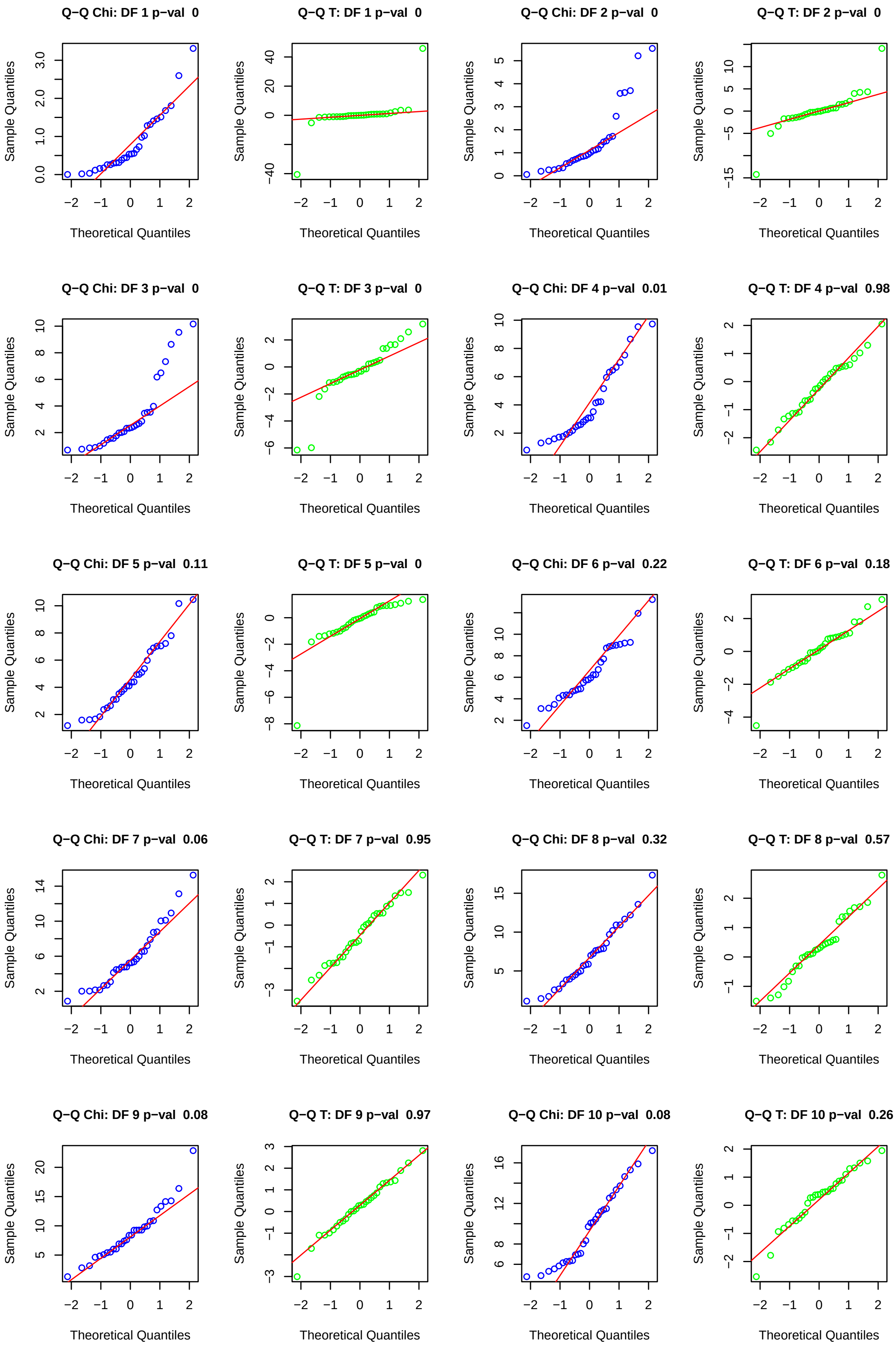
The H_A alternative hypothesis is that the data are not normally distributed.

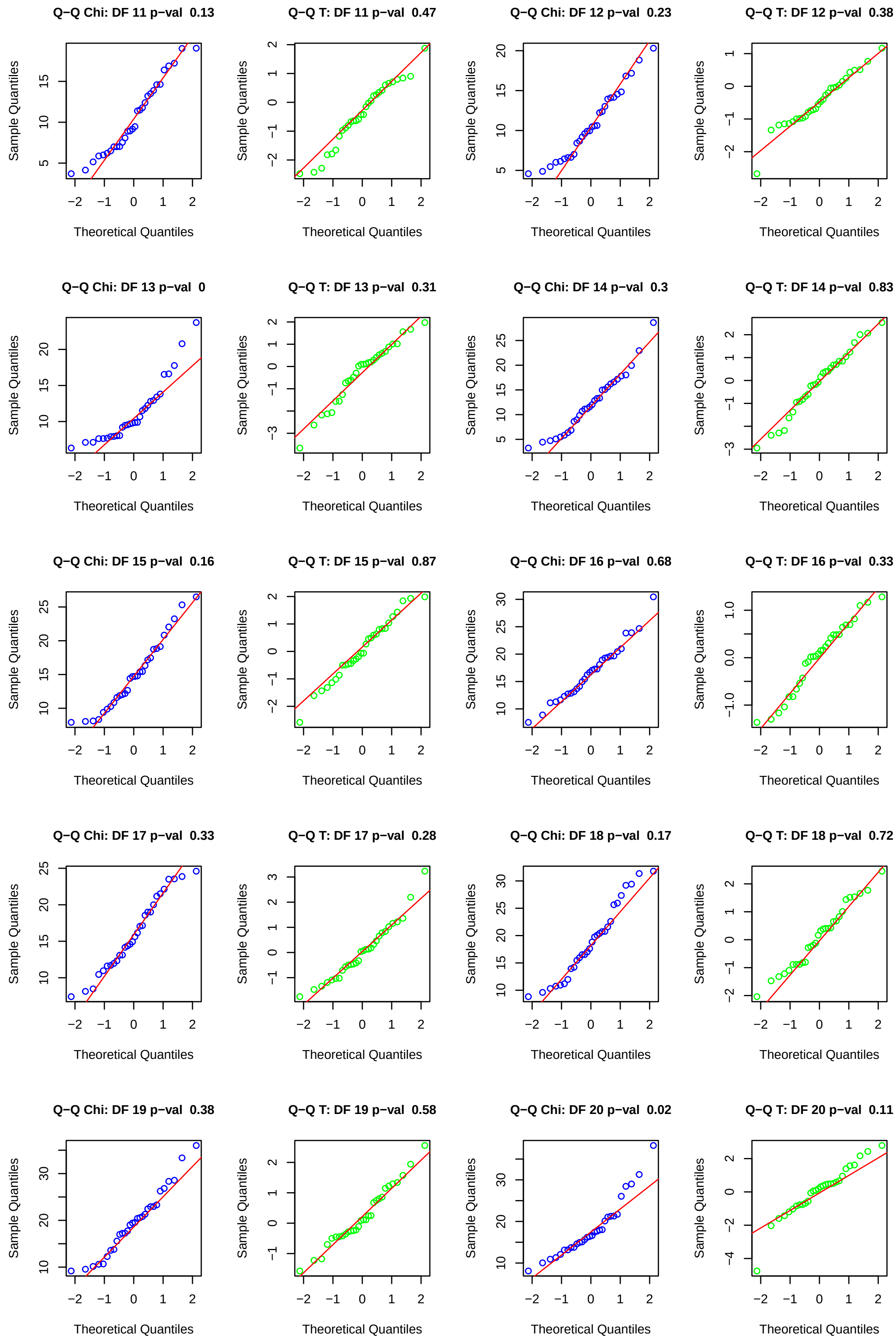
∴ Therefore, if *p-value* less than 0.05, we reject the null hypothesis and state that the data are plausibly not normally distributed.

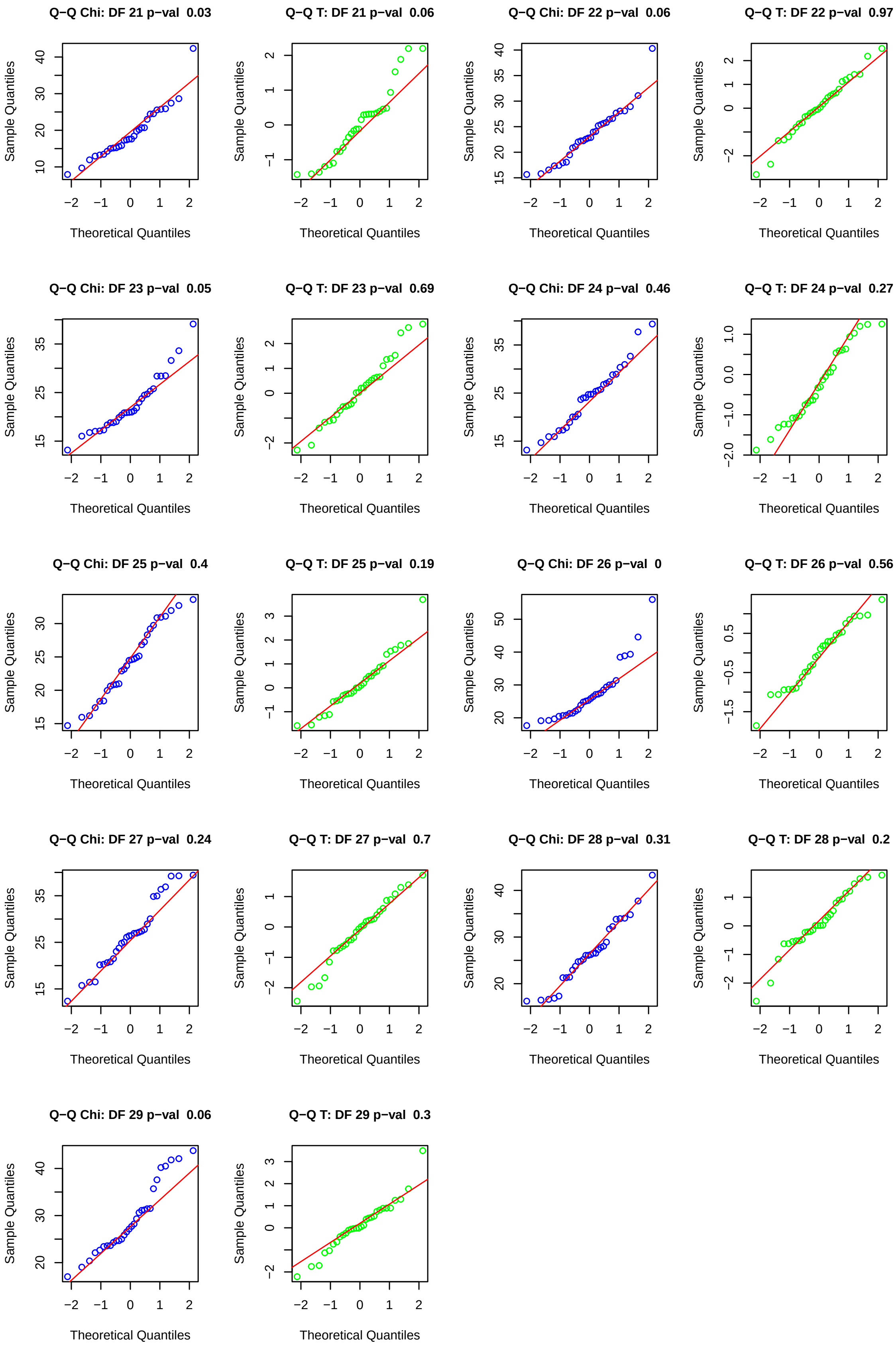
And if the *p-value* is greater than 0.05, we keep the null hypothesis and state that it is plausible that the data are normally distributed.

```
#R Code Here

# Sample size
N <- 30

# Let's bound our story to a maximum degrees of freedom
Max_D <- N-1

TRIALS <- 100
# Initialize two dataframes: one for chisq and one for t distribution
chi_p.value <- as.data.frame(matrix(NA, nrow = TRIALS*Max_D, ncol=3))
t_p.value <- as.data.frame(matrix(NA, nrow = TRIALS*Max_D, ncol=3))

par(mfrow = c(5, 4))

# Initialize observation number
```

```r
obs <- 0

# Loop from 1 to S-1 degree of freedom experiments each.
# for (i in 1:(N-1)) {
for (i in 1:Max_D) {

  # Now I am going to run 100 trials for each experiment
  # To do multiple tests and plot a boxplot of p-values
  # Trials within an experiment
  for (j in 1:TRIALS) {

    obs <- obs + 1
    # Run Chi-Square distribution trials
    CHI_SAMPLE <- rchisq(N, i)
    CHI_SHAPIRO <- shapiro.test(CHI_SAMPLE)

    # Store it in the dataframe
    chi_p.value[obs, 1] <- i
    chi_p.value[obs, 2] <- CHI_SHAPIRO$p.value
    chi_p.value[obs, 3] <- 'chisq'

    # Run t-distribution trials.
    T_SAMPLE <- rt(N, i)
    T_SHAPIRO <- shapiro.test(T_SAMPLE)

    # Store it in the dataframe
    t_p.value[obs, 1] <- i
    t_p.value[obs, 2] <- T_SHAPIRO$p.value
    t_p.value[obs, 3] <- 't'

  }

  # Q-Q plot the last trial
  qqnorm(CHI_SAMPLE,
        #ylim=c(-2, 2),
        col='blue',
        main=paste('Q-Q Chi: DF', i, 'p-val ', round(CHI_SHAPIRO$p.value,2)),
        cex.main=1,)
  qqline(CHI_SAMPLE,
        #ylim=c(-2, 2),
        col='red')

  qqnorm(T_SAMPLE,
        #ylim=c(-2, 2),
        col='green',
        main=paste('Q-Q T: DF', i, 'p-val ', round(T_SHAPIRO$p.value,2)),
        cex.main=1,)
  qqline(T_SAMPLE,
        #ylim=c(-2, 2),
        col='red')

}
```

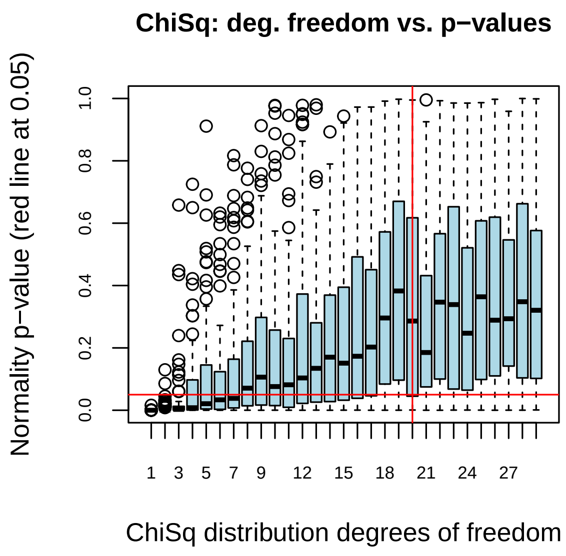

Q−Q Chi: DF 1 p−val 0
Q−Q T: DF 1 p−val 0
Q−Q Chi: DF 2 p−val 0
Q−Q T: DF 2 p−val 0
Q−Q Chi: DF 3 p−val 0
Q−Q T: DF 3 p−val 0
Q−Q Chi: DF 4 p−val 0.01
Q−Q T: DF 4 p−val 0.98
Q−Q Chi: DF 5 p−val 0.11
Q−Q T: DF 5 p−val 0
Q−Q Chi: DF 6 p−val 0.22
Q−Q T: DF 6 p−val 0.18
Q−Q Chi: DF 7 p−val 0.06
Q−Q T: DF 7 p−val 0.95
Q−Q Chi: DF 8 p−val 0.32
Q−Q T: DF 8 p−val 0.57
Q−Q Chi: DF 9 p−val 0.08
Q−Q T: DF 9 p−val 0.97
Q−Q Chi: DF 10 p−val 0.08
Q−Q T: DF 10 p−val 0.26
Sample Quantiles
Theoretical Quantiles

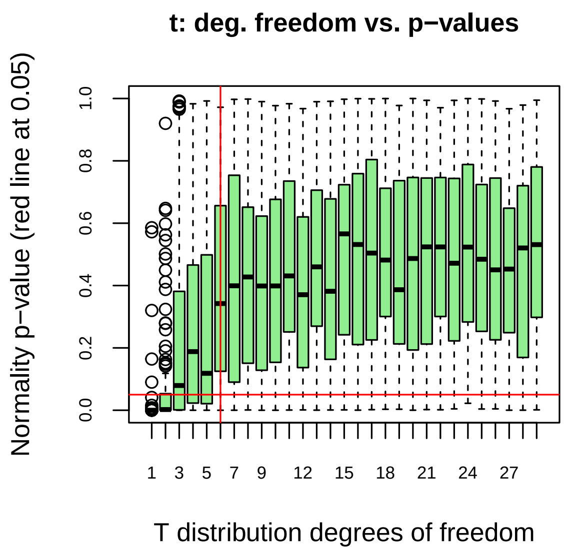

Q−Q Chi: DF 11 p−val 0.13
Q−Q T: DF 11 p−val 0.47
Q−Q Chi: DF 12 p−val 0.23
Q−Q T: DF 12 p−val 0.38
Q−Q Chi: DF 13 p−val 0
Q−Q T: DF 13 p−val 0.31
Q−Q Chi: DF 14 p−val 0.3
Q−Q T: DF 14 p−val 0.83
Q−Q Chi: DF 15 p−val 0.16
Q−Q T: DF 15 p−val 0.87
Q−Q Chi: DF 16 p−val 0.68
Q−Q T: DF 16 p−val 0.33
Q−Q Chi: DF 17 p−val 0.33
Q−Q T: DF 17 p−val 0.28
Q−Q Chi: DF 18 p−val 0.17
Q−Q T: DF 18 p−val 0.72
Q−Q Chi: DF 19 p−val 0.38
Q−Q T: DF 19 p−val 0.58
Q−Q Chi: DF 20 p−val 0.02
Q−Q T: DF 20 p−val 0.11
Sample Quantiles
Theoretical Quantiles


Q–Q Chi: DF 21 p–val 0.03
Q–Q T: DF 21 p–val 0.06
Q–Q Chi: DF 22 p–val 0.06
Q–Q T: DF 22 p–val 0.97
Q–Q Chi: DF 23 p–val 0.05
Q–Q T: DF 23 p–val 0.69
Q–Q Chi: DF 24 p–val 0.46
Q–Q T: DF 24 p–val 0.27
Q–Q Chi: DF 25 p–val 0.4
Q–Q T: DF 25 p–val 0.19
Q–Q Chi: DF 26 p–val 0
Q–Q T: DF 26 p–val 0.56
Q–Q Chi: DF 27 p–val 0.24
Q–Q T: DF 27 p–val 0.7
Q–Q Chi: DF 28 p–val 0.31
Q–Q T: DF 28 p–val 0.2
Q–Q Chi: DF 29 p–val 0.06
Q–Q T: DF 29 p–val 0.3
Sample Quantiles
Theoretical Quantiles

1.1 Boxplot: Visualize DF impact

```
# BOxplot of p-values

#chi_p.value[complete.cases(chi_p.value,)]
colnames(chi_p.value) <- c('deg_of_freedom', 'p_value', 'distribution')
colnames(t_p.value) <- c('deg_of_freedom', 'p_value', 'distribution')

par(mfrow = c(1, 2))

boxplot(p_value~deg_of_freedom, chi_p.value,
        col = 'lightblue',
        main='ChiSq: deg. freedom vs. p-values',
        ylab = 'Normality p-value (red line at 0.05)',
        xlab = 'ChiSq distribution degrees of freedom',
        cex.main=1,
        cex.axis=0.7)
abline(v=20, h=0.05, col='red')

boxplot(p_value~deg_of_freedom, t_p.value,
        col = 'lightgreen',
        main = 't: deg. freedom vs. p-values',
        ylab = 'Normality p-value (red line at 0.05)',
        xlab = 'T distribution degrees of freedom',
        cex.main=1,
        cex.axis=0.7)
abline(v=6, h=0.05, col='red')
```





- In both cases, *ChiSquare distribution and t-distribtution*, increasing the *number* of *degrees of freedom* makes the drawn distribution closer to a *Normally distributed* distribution.
- The sample size is fixed at 30 as stated in the problem.
- We can visually perceive *Q-Q plots* with data points aligning closer to the diagonal as we increase the *number* of *degrees of freedom*.

- Evidently, these are random samples. And as such, the random sample will varie from experiment to experiment.
- Therefore, I calculated the *Shapiro-Wilk Tests* for multiple experiment to assess an acceptable value of *degrees of freedom* for each distribution case.
- I store the *p-values* of the *Shapiro-Wilk Tests* in a *dataframe*. I ran 100 trials for each *degree of freedom* and for wach type of distribution.
- The H_o *null hypothesis* states that the sample likely came from a *Normal* distribution. If the *p-value* is higher than 0.05 we keep the *null hypothesis* stating that the same is *Normally distributed*.
- TheH_A *alternative hypothesis* states the opporsite, the sample des not come from a *Normal* distribution, and we reject and *null hypothesis*.
- Then I created *boxplots* for each type of distribution, *ChiSquare and t distribution.*
- The chart shows a *boxplot* for each *degrees of freedom* value.
- We can observe an upward or positive trend as we increase the *degrees of freedom*, the *p-value* increases.
- We also observe that the upward trend is more pronounced for the *t-distribution* than for the *ChiSquare distribution*.
- Furthermore, I included a *red* line on the key 0.05 *p-value* mark.
- And I included a vertical *red* line where the *interquartile range* of the *boxplots* clear, and are above the 0.05 *p-value* mark.
- For a *ChiSquare distribution*, when we have *degrees of freedom* greater than 20 we have a *p-values* above the 0.05 for the *interquartile range* (the boxes from the boxplot).
- For a *t distribution*, when we have *degrees of freedom* at least 6 we have a *p-values* above the 0.05 for the *interquartile range* (the boxes from the boxplot).
- Therefore, to answer the question, for 30 observations: *'Approximately how large degrees of freedom is necessary, in each instance, to obtain a consistent normal distribution shape?*
- The answer depends on how strict is the use case, the impact to people.
- For $ChiSquare$, I would use at least 20 *degrees of freedom* to have consistent draws of *Normally distributed* data.
- For $t - distribution$, I would use at least 6 *degrees of freedom* to have consistent draws of *Normally distributed* data.