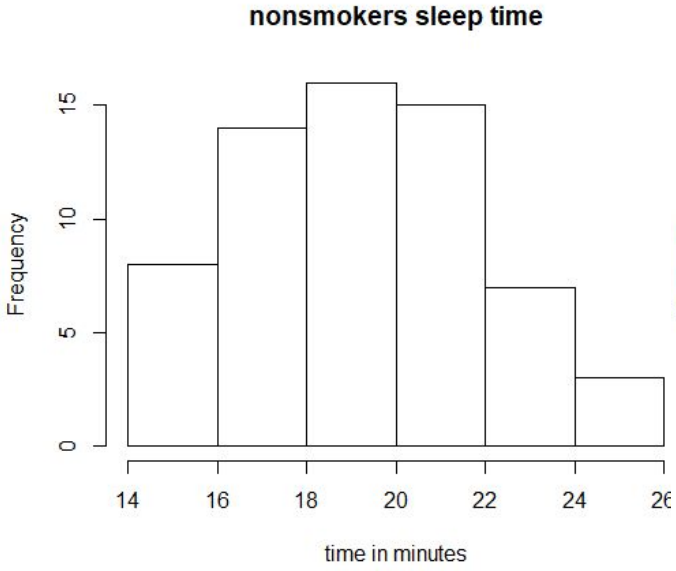
Topic: Effect of smoking on sleep pattern

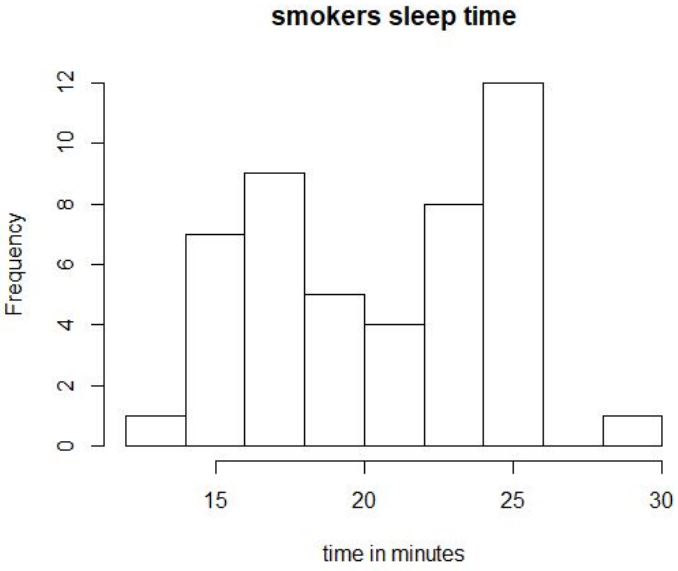
Conduct study to help understand the effect of smoking on sleep patterns. The random variable considered is T, the time in minutes that it takes to fall asleep. Samples of smokers and nonsmokers yield these observations on T. A total of 63 non-smokers and 47 smokers were surveyed.

Part 1) Is there any time difference between the 2 groups?

Part 2) Write down the report of your observation that you believe to be correct. Perform statistical analysis to support your Argument.

Part 1)
If we plot the histogram of sleeptime of smokers and histogram of nonsmokers, we will see these two graphs does not look similar at all. By observing the data, we can also see smokers' data varied a lot more.

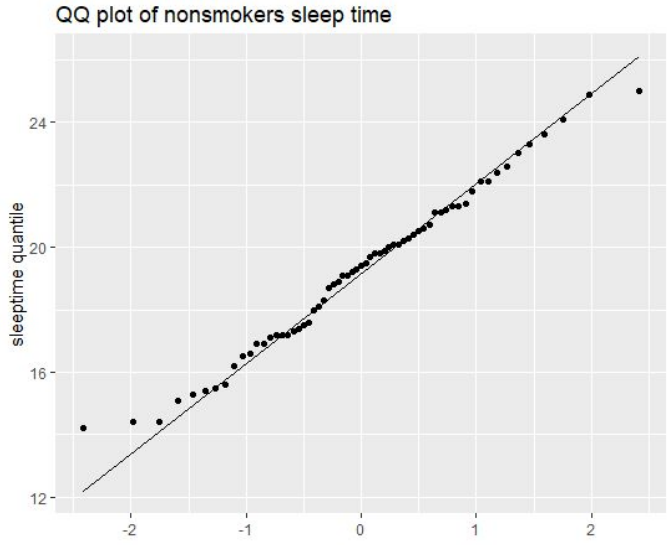


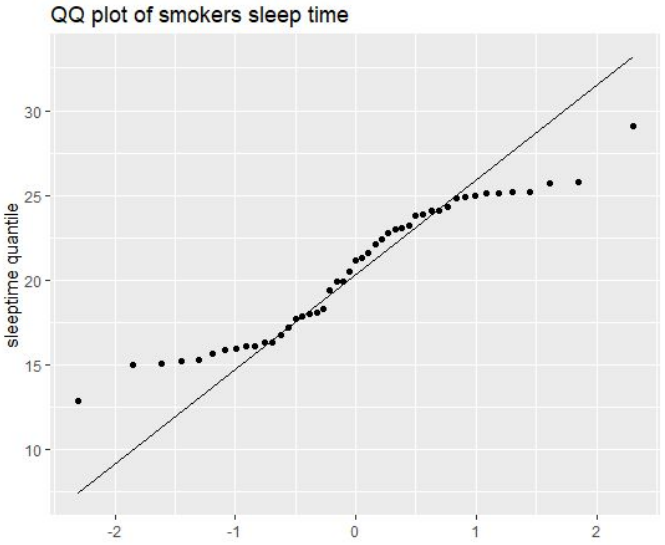


Part 2)
Let $\overline{T_n}$ be the random variable that represents the sample mean of nonsmokers' sleep time.
Let $\overline{T_s}$ be the random variable that represents the sample mean of smokers' sleep time.

I suspect there's a difference between $\overline{T_n}$ *and* $\overline{T_s}$, so I will compare their population means.

First we check if $\overline{T_n}$ *and* $\overline{T_s}$ are drawn from a normal distribution, i.e. check if smokers' sleeptime and nonsmokers' sleeptime are roughly normally distributed. We do this by plotting out the QQ plot of both, which results as follows:





The nonsmoker's QQ plot shows a very strong evidence that nonsmoker's sleeptime is normally distributed.
The smoker's QQ plot on the other hand, is very rough. It only roughly follows the normal distribution.

$$\overline{T_n} \sim N(\mu_n , \frac{\sigma_n^2}{n_n})$$
$$\overline{T_s} \sim N(\mu_s , \frac{\sigma_s^2}{n_s})$$

Then we set up the Hypothesis:
$H_0 : \mu_n - \mu_s = 0$
$H_A : \mu_n - \mu_s \neq 0$

So under the assumption of H_0, we get that

$$\overline{T_n} - \overline{T_s} \sim N(0, \frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s})$$

Next, I run R scripts to obtain the necessary information to perform statistical tests.
We get the following results:

Sample stat: $\overline{T_n} = 19.24$

Sample stat: $\overline{T_s} = 20.56$

Sample variance: $S_n^2 = 7.01$

Sample variance: $S_s^2 = 16.07$

$n_n = 63$

$n_s = 47$

With these data, we are able to perform the statistical tests.(Hand Written on next page)

① First perform F test to determine if $\sigma_n^2 = \sigma_s^2$

$H_0: \sigma_n^2 = \sigma_s^2$

$H_A: \sigma_n^2 \neq \sigma_s^2$

Test stat: $\frac{S_n^2}{\sigma_n^2} \times \frac{\sigma_s^2}{S_s^2} \sim F_{(n_n-1, n_s-1)}$

under H_0, $\frac{S_n^2}{S_s^2} \sim F_{(62,46)}$

$F = \frac{7.01}{16.07} \sim F_{(62,46)}$

$F = \frac{16.07}{7.01} \sim F_{(46,62)}$

$F = 2.29 \sim F_{(46,62)}$

Pvalue : $P(F_{46,62} \geq 2.29 \text{ or } F_{46,62} \leq \frac{1}{2.29})$

$= P(F_{46,62} \geq 2.29 \text{ or } F_{62,46} \geq 2.29)$

$P(F_{46,62} \geq 2.29)$ is between 0.1% and 1%

$P(F_{62,46} \geq 2.29)$ is between 0.1% and 1%

Therefore

$P(F_{46,62} \geq 2.29 \text{ or } F_{62,46} \geq 2.29)$ is between 0.2% and 2%

P-value is significance at 5% level, therefore we have enough evidence to reject null hypothesis, so the variance of non-smoker's sleeptime and smoker's sleeptime are different.

② Testing for $\mu_n - \mu_s$

$H_0: \mu_n - \mu_s = 0$

$H_A: \mu_n - \mu_s \neq 0$

$\bar{T}_n - \bar{T}_s \sim N(0, \frac{v_n^2}{n_n} + \frac{v_s^2}{n_s})$

Sample stat: $20.56 - 19.24 = 1.32$

Data: $\bar{T}_n = 19.24$

$\bar{T}_s = 20.56$

$s_n^2 = 7.01$

$s_s^2 = 16.07$

Estimate $V(\bar{T}_n - \bar{T}_s)$ with sample variance

$\frac{7.01}{63} + \frac{16.07}{47}$

② Test stat:

$$T = \frac{1.32 - 0}{\sqrt{\frac{7.01}{63} + \frac{16.07}{47}}} = \frac{1.32}{0.673} = 1.96$$

degrees of freedom

$$\gamma = \frac{\left[\frac{7.01}{63} + \frac{16.07}{47}\right]^2}{\frac{(\frac{7.01}{63})^2}{62} + \frac{(\frac{16.07}{47})^2}{46}} = 75$$

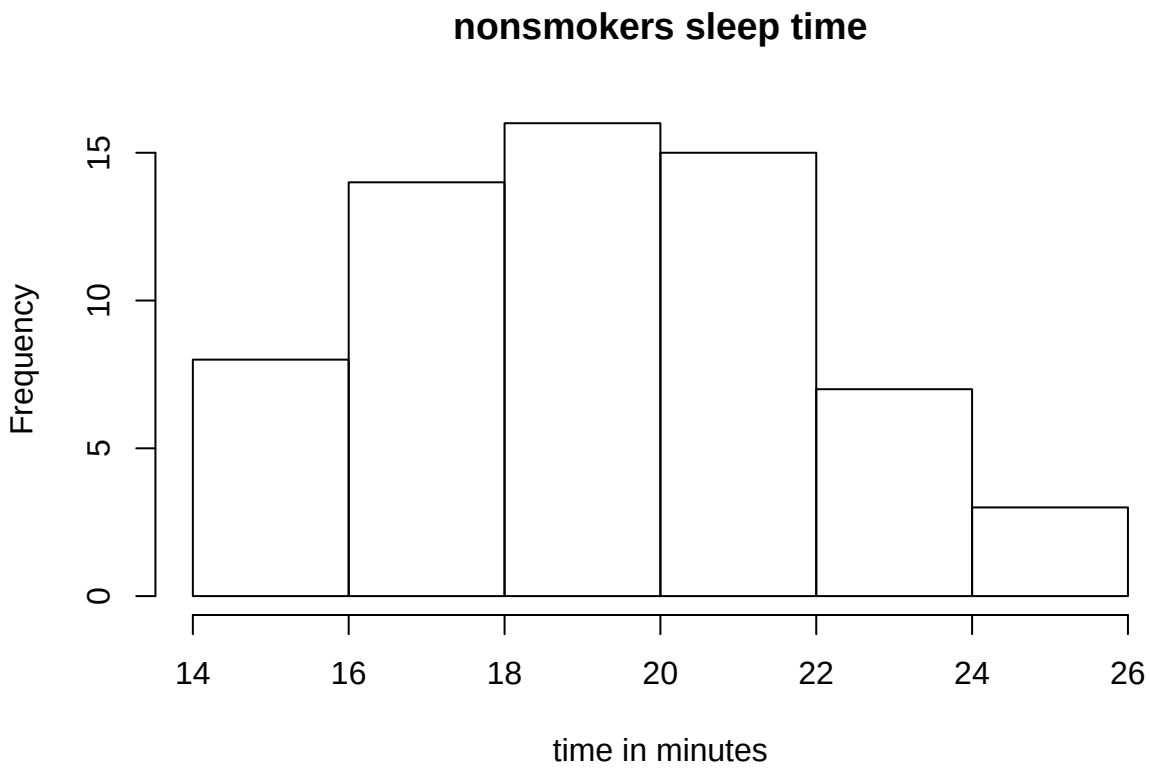
③ P-value:

$P(|T| \geq 1.96)$ is between 5% and 10%.

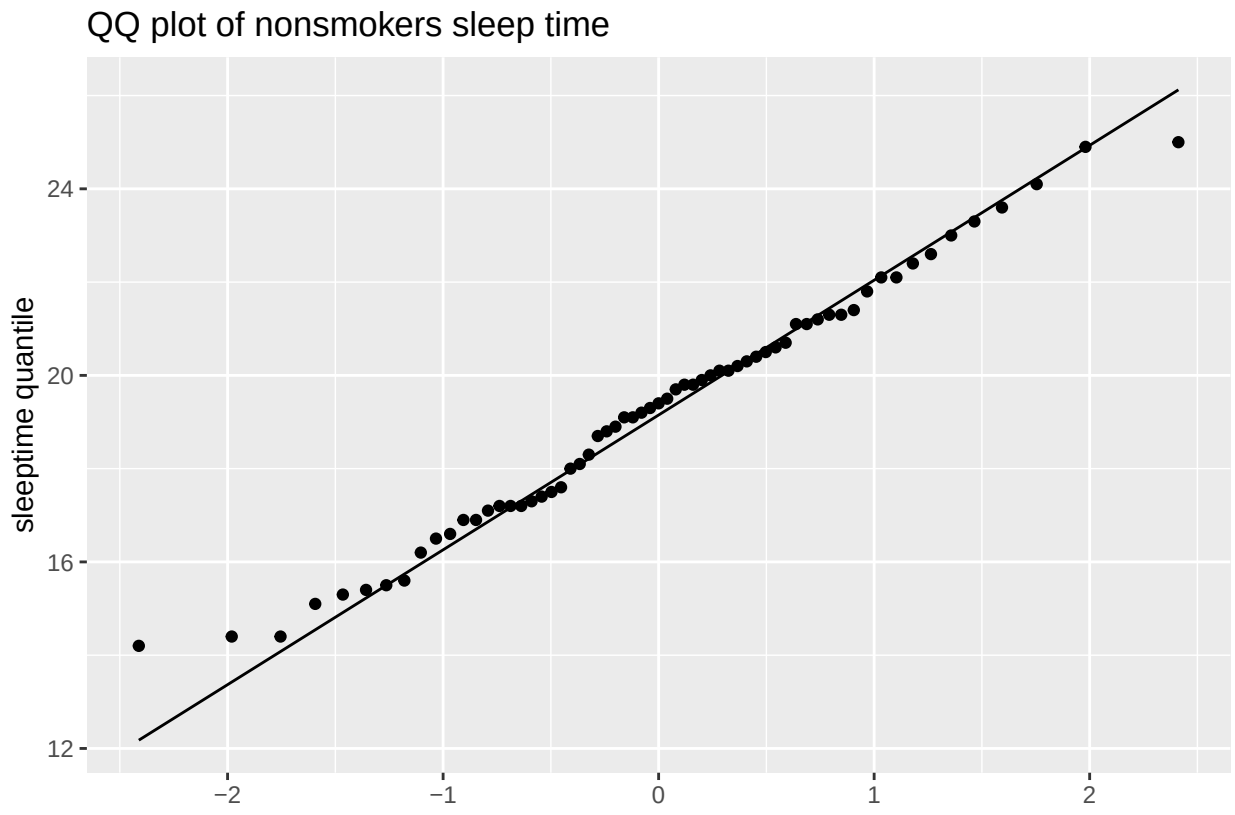
Therefore we fail to reject the null hypothesis at 5% significance level, i.e. there is not enough evidence to show that smoking makes a difference to time in minutes it takes to fall asleep.

Sleep time report anaylsis

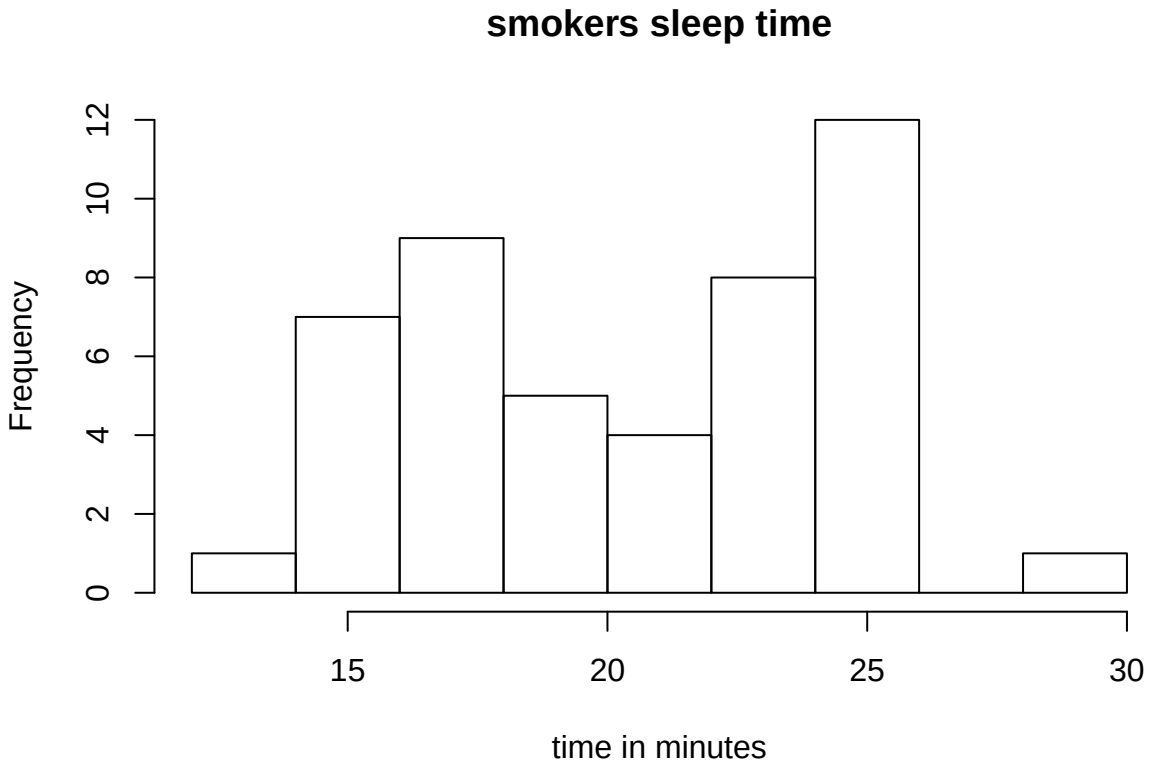
```
#Check for normality
nonsmoke_st = hist(sleep_time$nonsmokers, main = "nonsmokers sleep time", xlab = "time in minutes")
```



```r
nonsmoke_qplot = qplot(sample = sleep_time$nonsmokers, data = sleep_time) + geom_qq_line()
nonsmoke_qplot = nonsmoke_qplot + ggtitle("QQ plot of nonsmokers sleep time")
nonsmoke_qplot = nonsmoke_qplot  + ylab("sleeptime quantile")
nonsmoke_qplot
```



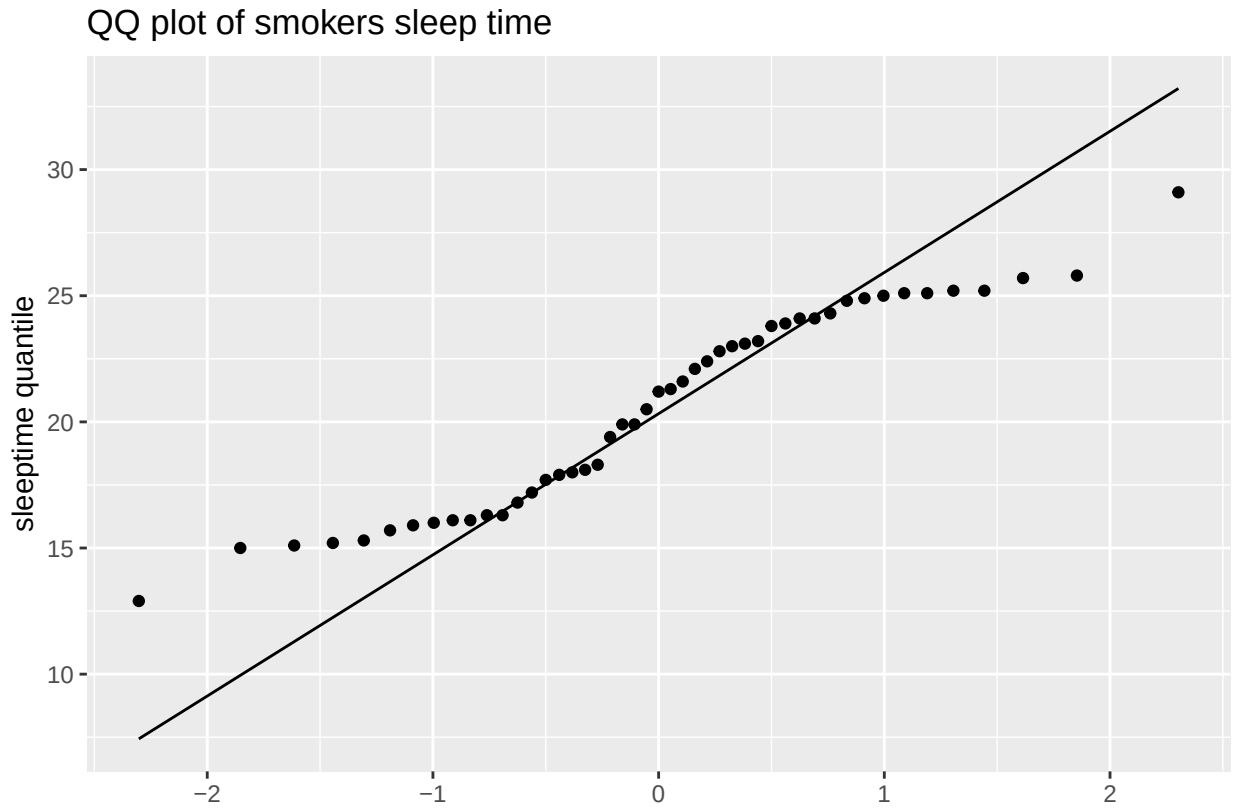
```r
smoke_st = hist(sleep_time$smokers, main = "smokers sleep time", xlab = "time in minutes")
```



```
smoke_qplot = qplot(sample = sleep_time$smokers, data = sleep_time) + geom_qq_line()
smoke_qplot = smoke_qplot + ggtitle("QQ plot of smokers sleep time") + ylab("sleeptime quantile")
smoke_qplot
```

```
## Warning: Removed 16 rows containing non-finite values (stat_qq).
```

```
## Warning: Removed 16 rows containing non-finite values (stat_qq_line).
```



Gather the necessary data to perform f test and t test.

```
num_nonsmoker = 63
num_smoker = 47
remove_na = which(complete.cases((sleep_time$smokers)))

mean_nonsmoker = mean(sleep_time$nonsmokers)
mean_nonsmoker
```

```
## [1] 19.24286
```

```
mean_smoker = mean(sleep_time$smokers[remove_na])
mean_smoker
```

```
## [1] 20.5617
```

```
var_nonsmoker = var(sleep_time$nonsmokers)
var_nonsmoker
```

```
## [1] 7.013134
```

```
var_smoker = var(sleep_time$smokers[remove_na])
var_smoker
```

```
## [1] 16.06807
```