Question 1

The probability density function f for the χ^2_ν distribution (the chi-squared distribution with ν degrees of freedom) is

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2},$$

where the support is $x \in \mathbb{R}$ and $x > 0$, and the degrees of freedom $\nu \in \{1, 2, \ldots\}$.

(a) Let $Y \sim \chi^2_\nu$. Assuming that we know $\mathrm{E}(Y) = \nu$ and $\mathrm{E}\left(Y^2\right) = \nu(\nu+2)$, find $\mathrm{Var}(Y)$.

(b) Assume that $X_1, X_2, \ldots, X_n \sim \mathrm{N}\left(\mu, \sigma^2\right)$. Use Part (a) to show that

$$\mathrm{Var}\left(S^2\right) = \frac{2\sigma^4}{n-1},$$

where S^2 is the sample variance, i.e. $S^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$, as usual.

Solution to Question 1

Part (a):

$$\mathrm{Var}(Y) = \mathrm{E}\left(Y^2\right) - (\mathrm{E}(X))^2 = \nu(\nu+2) - (\nu)^2 = 2\nu$$

Part (b):

Theorem 3.2.2 in the notes proves that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Therefore, using Part (a),

$$\begin{aligned}
\mathrm{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) &= 2(n-1) \\
\Rightarrow \left(\frac{n-1}{\sigma^2}\right)^2 \mathrm{Var}\left(S^2\right) &= 2(n-1) \\
\Rightarrow \mathrm{Var}\left(S^2\right) &= 2(n-1) \cdot \frac{\sigma^4}{(n-1)^2} \\
\Rightarrow \mathrm{Var}\left(S^2\right) &= \frac{2\sigma^4}{n-1}.
\end{aligned}$$

Question 2

Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables following a normal distribution with mean μ and variance σ^2. The value of μ is unknown, but σ^2 is known to be $\sigma^2 = 16$. Suppose we observe $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Given that $\overline{x} = 7$ and $n = 50$, construct a 99% confidence interval for μ.

Solution to Question 2

Since $\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, if we define

$$Z = \frac{\mu - \overline{X}}{\sigma/\sqrt{n}}$$

then $Z \sim \mathrm{N}(0,1)$. For any significance level α, if we define z_α to be the value such that $\mathrm{P}(Z < z_\alpha) = \alpha$, then

$$\begin{aligned}
&\mathrm{P}\left(Z < z_{1-\alpha/2}\right) = 1 - \alpha/2, \\
&\mathrm{P}\left(Z < z_{\alpha/2}\right) = \alpha/2, \\
\Rightarrow &\mathrm{P}\left(z_{\alpha/2} < Z < z_{1-\alpha/2}\right) = 1 - \alpha. \\
\Rightarrow &\mathrm{P}\left(z_{\alpha/2} < \frac{\mu - \overline{X}}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha \\
\Rightarrow &\mathrm{P}\left(\overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha
\end{aligned}$$

To construct a 99% confidence interval, $1 - \alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow \alpha/2 = 0.005$.

Using the table, we find $z_{0.995} = 2.576$, and therefore by symmetry of the normal distribution, $z_{0.005} = -2.576$. Since $\mathbf{X}$ is observed as $\mathbf{x}$ and $\overline{x} = 7$, a 99% confidence interval is therefore

$$\begin{aligned}
&\left(\overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \\
= &\left(7 - 2.576 \cdot \frac{4}{\sqrt{50}}, 7 + 2.576 \cdot \frac{4}{\sqrt{50}}\right).
\end{aligned}$$

Question 3

Suppose $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables following a normal distribution with mean μ and variance σ^2. The values of μ and σ^2 are both unknown. Suppose we observe $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ as $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. Given that the sample mean is $\overline{y} = 11$, the sample variance is $s^2 = 18$ and $n = 8$, construct a 90% confidence interval for μ.

Solution to Question 3

Here we use Student's t-test since

$$T = \frac{\mu - \overline{Y}}{s/\sqrt{n}} \sim t_{n-1},$$

where t_{n-1} denotes Student's t-distribution with $n-1$ degrees of freedom. **Note that the degrees of freedom is $n-1$ and not simply n.** Let $t_{n-1,\alpha}$ denote the value such that, if $T \sim t_{n-1}$, then

$$\mathrm{P}\left(T < t_{n-1,\alpha}\right) = \alpha.$$

Then

$$\begin{aligned}
&\mathrm{P}\left(t_{n-1,\alpha/2} < T < t_{n-1,1-\alpha/2}\right) = \alpha \\
\Rightarrow &\mathrm{P}\left(t_{n-1,\alpha/2} < \frac{\mu - \overline{X}}{s/\sqrt{n}} < t_{n-1,1-\alpha/2}\right) = \alpha \\
\Rightarrow &\mathrm{P}\left(\overline{X} + t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,1-\alpha/2} \cdot \frac{s}{\sqrt{n}}\right) = \alpha
\end{aligned}$$

Since we have observed $\mathbf{Y}$ as $\mathbf{y}$, and $\overline{y} = 11$, $s^2 = 18$ and $n = 8$, and since we want a 90% confidence interval, which implies $\alpha = 0.1 \Rightarrow 1 - \alpha/2 = 0.95$, we find in the table that $t_{7,0.95} = 1.895$. By symmetry of the t-distribution around 0, $t_{7,0.05} = -1.895$. Therefore, our 90% confidence interval is

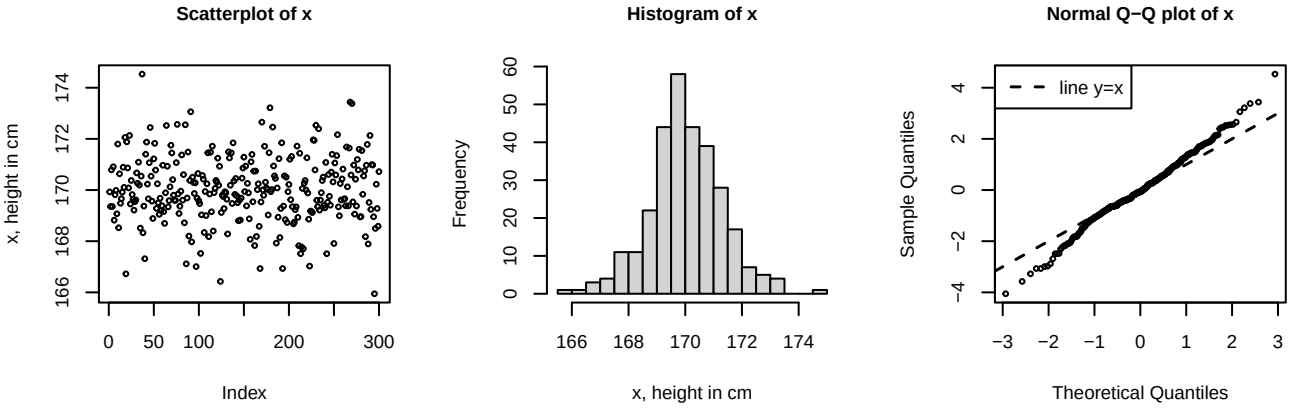
$$\left(11 - 1.895\frac{\sqrt{18}}{\sqrt{8}}, 11 + 1.895\frac{\sqrt{18}}{\sqrt{8}}\right)$$

$$= \left(11 - 1.895\frac{3\sqrt{2}}{2\sqrt{2}}, 11 + 1.895\frac{3\sqrt{2}}{2\sqrt{2}}\right)$$

$$= \left(11 - 1.895\left(\frac{3}{2}\right), 11 + 1.895\left(\frac{3}{2}\right)\right)$$

Question 4

Suppose $X_1, X_2, \ldots, X_n$ are the random variables representing the heights of the $n = 300$ students in a particular module, measured in cm. These random variables are observed as $x_1, x_2, \ldots, x_n$, which are plotted below in (a) a scatterplot of the data, (b) a histogram of the data, (c) a Q-Q plot of the data after being standardised by the sample mean and variance. Do these plots suggest that $X_1, X_2, \ldots, X_n$ follow a normal distribution? Justify your answer.



Solution to Question 4

Although the scatterplot and histogram may suggest the data is normal, it is not possible to tell conclusively from these plots.

The Q-Q plot shows that the sample quantiles do not agree with the theoretical quantiles (do not lie along the line $y = x$) for a large proportion of the quantiles. Therefore, this suggests that the random variables which have been observed do not follow a normal distribution.

Note: it is also possible to argue that most of the points in the Q-Q plot lie along the line $y = x$, and so the data seems to be normally distributed. However, if it were only the most extreme 5 or 6 observations at each end (in the intervals $[-3, -2]$ and $[2, 3]$) that did follow the line $y = x$, then that argument would be more acceptable. The problem is the large number of values in the intervals $[-1, -2]$ and $[1, 2]$ (note the overplotting) that do not follow the line $y = x$; it is these values which suggest that the data is not normally distributed.