Question 1

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 1

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 2

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 2

This is similar to Question 1, but here we use Student's t-test since

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

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 3

Suppose $Z_1, Z_2, \ldots, Z_n$ are independent and identically distributed random variables following an unknown distribution F_Z. The mean μ of the distribution F_Z is unknown, but the variance of F_Z is known to be $\sigma^2 = 7$. Suppose we observe $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ as $\mathbf{z} = (z_1, z_2, \ldots, z_n)$. Given that the sample mean is $\overline{z} = 6$ and $n = 12$, construct a 95% confidence interval for μ.

Solution to Question 3

If the distribution is unknown, but the variance is known, we can use Chebyshev's inequality. For any X and any $k > 0$,

$$\mathrm{P}\left(|X - \mathrm{E}(X)| < k\sqrt{\mathrm{Var}(X)}\right) \geq 1 - \frac{1}{k^2}.$$

We know, by linearity of expectation and properties of the variance and since the Z_i are independent (Proposition 1.2.6):

$$\begin{aligned}\mathrm{E}\left(\overline{Z}\right) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(Z_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \\ \mathrm{Var}\left(\overline{Z}\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Z_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}\end{aligned}$$

Then,

$$\begin{aligned}&\mathrm{P}\left(\left|\overline{Z} - \mu\right| < k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}. \\ \Rightarrow &\mathrm{P}\left(\left|\mu - \overline{Z}\right| < k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}. \\ \Rightarrow &\mathrm{P}\left(-k\frac{\sigma}{\sqrt{n}} < \mu - \overline{Z} < k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}. \\ \Rightarrow &\mathrm{P}\left(\overline{Z} - k\frac{\sigma}{\sqrt{n}} < \mu < \overline{Z} + k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}.\end{aligned}$$

To find the value of k,

$$\begin{aligned}&1 - \frac{1}{k^2} = 0.95 \\ \Rightarrow &\frac{1}{k^2} = 0.05 = \frac{1}{20} \\ \Rightarrow &k^2 = 20 \\ \Rightarrow &k = \sqrt{20} = 2\sqrt{5}\end{aligned}$$

If $1 - \frac{1}{k^2} = 0.95$, then $k = \sqrt{0.05} = \frac{1}{\sqrt{20}} = \frac{1}{2\sqrt{5}}$. Since $\overline{z} = 6$ and $n = 12$ and $\sigma^2 = 7$, the 95% confidence interval is

$$\left(\overline{z} - k\frac{\sigma}{\sqrt{n}} < \mu < \overline{z} + k\frac{\sigma}{\sqrt{n}}\right)$$

$$= \left(6 - 2\sqrt{5}\cdot\frac{\sqrt{7}}{\sqrt{12}}, 6 + 2\sqrt{5}\cdot\frac{\sqrt{7}}{\sqrt{12}}\right)$$

$$= \left(6 - 2\sqrt{5}\cdot\frac{\sqrt{7}}{2\sqrt{3}}, 6 + 2\sqrt{5}\cdot\frac{\sqrt{7}}{2\sqrt{3}}\right)$$

$$= \left(6 - \frac{\sqrt{5}\sqrt{7}}{\sqrt{3}}, 6 + \frac{\sqrt{5}\sqrt{7}}{\sqrt{3}}\right).$$

Question 4

Suppose the heights of two groups of people are recorded. Group A consists of n people and their heights are recorded (in cm) as $x_1, x_2, \dots, x_n$ with $n = 10$, sample mean $\overline{x} = 171.5$ and sample variance $s_x^2 = 2$. Group B consists of m people and their heights are recorded as $y_1, y_2, \dots, y_m$, with $m = 12$, $\overline{y} = 170$ and sample variance $s_y^2 = 3$. We wish to test if the average heights of the two groups are significantly different or not. We start by assuming that the measurements $x_1, x_2, \dots, x_n$ are observations of the independent random variables $X_1, X_2, \dots, X_n$, respectively, which follow a normal distribution with unknown mean μ_1 and unknown variance σ_1^2. We also assume that the $y_1, y_2, \dots, y_m$ are observations of the independent random variables $Y_1, Y_2, \dots, Y_m$, respectively, following a normal distribution with unknown mean μ_2 and unknown variance σ_2^2. We also assume that although the variances are unknown, they are equal i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

(a) What is the null hypothesis for this test?

(b) Assuming the null hypothesis is true, use Student's two-sample t-test to compute a p-value and decide whether or not the average heights of the two groups are signficantly different or not.

Solution to Question 4

Part (a):

The null hypothesis is that the two means are equal, i.e.

$$H_0 : \mu_1 = \mu_2$$

Part (b):

Using the hint,

$$s_p^2 = \frac{1}{10+12-2}\left((9)2 + (11)3\right) = \frac{51}{20}$$

Furthermore,

$$\sqrt{\frac{1}{10} + \frac{1}{12}} = \sqrt{\frac{22}{120}} = \sqrt{\frac{11}{60}}$$

Under the null hypothesis $\mu_1 - \mu_2 = 0$. Then, the observed value of the statistic is

$$t = \frac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{171.5 - 170}{\sqrt{\frac{51}{20}}\sqrt{\frac{11}{60}}}$$

$$= \frac{1.5}{\sqrt{\frac{17\cdot 3}{20}}\sqrt{\frac{11}{3\cdot 20}}} = \frac{1.5}{\frac{1}{20}\sqrt{17\cdot 11}}$$

$$= \frac{30}{\sqrt{17\cdot 11}} = 2.193817 \qquad \text{(using a calculator)}$$

If we look at the table for the cumulative distribution function of the t-distribution, in the row for $n+m-2 = 20$, we see that

$$\mathrm{P}(T < 2.086) = 0.975, \qquad \text{and} \qquad \mathrm{P}(T < 2.528) = 0.99.$$

Since $t = 2.193817$ falls between these two values, we can say that $p < 0.05$.

(Remember, $1 - \alpha/2 = 0.975 \Rightarrow \alpha = 0.05$; see Question 2.)

Therefore, we can reject the null hypothesis at significance level $\alpha = 0.05$, but not at the level $\alpha = 0.02$.

Question 5

A pharmaceutical company conducts a number of clinical trials simultaneously to test the effectiveness of different drug treatments for a particular disease. In each clinical trial $i \in \{1, 2, \ldots, n\}$, a group of patients is randomly divided into two subgroups, one of which is given drug treatment i while the other is given a placebo (a substance that has no effect on the disease, such as a sugar pill). After a period of time, the patients are examined and declared either to be cured or not to be cured. For each clinical trial, a statistical analysis is performed on the resulting data from the two subgroups.

(a) If the goal is to determine if a drug treatment is effective, what should the null hypothesis be for each statistical test?

(b) The results of the $n = 15$ statistical tests were the following p-values (in increasing order):

$$\begin{array}{cccccccc} 0.0001, & 0.0004, & 0.0019, & 0.0095, & 0.0201, & 0.0278, & 0.0298, & 0.0344, \\ 0.0459, & 0.3240, & 0.4262, & 0.5719, & 0.6528, & 0.7590, & 1.000. & \end{array}$$

If the pharmaceutical company declared in advance that a significance level of $\alpha = 0.05$ would be used, which of the p-values should be considered as significant (and therefore, which corresponding hypotheses should be rejected)?

Solution to Question 5

Part (a):

The null hypothesis for each statistical test $i \in \{1, 2, \ldots, n\}$ is:

$$H_0: \text{ drug treatment } i \text{ has no effect}$$

Part (b):

Although several p-values are less than $\alpha = 0.05$, we need to take into account the multiple hypothesis testing and include a correction for the multiple hypothesis testing.

Since there are $n = 15$ tests, if we use the Bonferroni correction, the adjusted significance level will be $\alpha' = \alpha/15 = 0.0033$.

If we compare the p-values to this adjusted threshold $\alpha' = 0.0033$, we see that only the three smallest p-values are below this threshold. These significant p-values are:

$$0.0001, \qquad 0.0004, \qquad 0.0019$$

The corresponding hypotheses will therefore be rejected and the corresponding drug treatments may be considered as effective.

Note that although the p-values

$$0.0095, \qquad 0.0201, \qquad 0.0278, \qquad 0.0298, \qquad 0.0344, \qquad 0.0459$$

are below the threshold $\alpha = 0.05$, they are not less than the adjusted threshold $\alpha' = 0.0033$, and therefore are not considered to be significant.