STATS 2107
Statistical Modelling and Inference II
Tutorial 2
Solutions

Sharon Lee, Matt Ryan

Semester 2 2022

1. a. If $X \sim \chi^2_k$, show that $E(X) = k$ and $\text{Var}(X) = 2k$. **Hint: Use MGFs.**

Solutions:

$$\begin{aligned}
E[X] &= \frac{d}{dt} M_X(t)\Big|_{t=0} \\
&= \frac{d}{dt}(1-2t)^{-k/2}\Big|_{t=0} \\
&= k(1-2t)^{-k/2-1}\Big|_{t=0} \\
&= k. \\
E[X^2] &= \frac{d^2}{dt^2} M_X(t)\Big|_{t=0} \\
&= \frac{d}{dt} k(1-2t)^{-k/2-1}\Big|_{t=0} \\
&= k(k+2)(1-2t)^{-k/2-2}\Big|_{t=0} \\
&= k(k+2).
\end{aligned}$$

Now

$$\begin{aligned}
\text{Var}(X) &= E[X^2] - E[X]^2 \\
&= k(k+2) - k^2 \\
&= 2k.
\end{aligned}$$

---

b. Suppose $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ independently. Find the distribution of $X_1 + X_2$.

Solutions:
Use the fact that the MGF of a sum of independent random variables is equal to the product of the individual MGFs.

$$\begin{aligned}
M_{X_1+X_2}(t) &= M_{X_1}(t) \times M_{X_2}(t) \\
&= (1-2t)^{-k_1/2} \times (1-2t)^{-k_2/2} \\
&= (1-2t)^{-(k_1+k_2)/2}.
\end{aligned}$$

By observation and the uniqueness of MGFs, we recognise this as the MGF of a χ^2 distribution with $k_1 + k_2$ degrees of freedom

2. A random sample of 500 hospital records shows that the length of stay in one of South Australia's hospitals had a (sample) mean 5.4 days and (population) standard deviation 3.1 days.

   a. A health agency hypothesizes that the average length of stay is 5 days. Do the data support this hypothesis? You may use $\alpha = 0.05$.

Solutions:
Let μ be the true mean length of stay in the hospital. Then we are testing

$$H_0 : \mu = 5 \quad \text{against} \quad H_a : \mu \neq 5 \,.$$

The test statistic is

$$z = \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{5.4 - 5}{\frac{3.1}{\sqrt{500}}} \approx 2.89 \,.$$

The critical value is $z_{\alpha/2} = z_{0.025} \approx 1.96$. Hence there is sufficient evidence to reject H_0 since $z > 1.96$.

---

   b. For the hypothesis test in part a, and using the significance level $\alpha = 0.05$, find β for $\mu = 5.5$. **Hint: first calculate the power using the formula from the lectures.**

Solutions:
We want to use the formula for the two-sided hypothesis test that:

$$\begin{aligned} 1 - \beta &= \text{Power} \\ &= \Phi\left(-z_{\alpha/2}; \frac{\mu - \mu_0}{\sigma/\sqrt{n}}, 1\right) + 1 - \Phi\left(z_{\alpha/2}; \frac{\mu - \mu_0}{\sigma/\sqrt{n}}, 1\right) , \end{aligned}$$

Where Φ is the cdf of the normal distribution. In this situation, $\alpha = 0.05$, $\mu = 5.5$, $\mu_0 = 5$, $\sigma = 3.1$, and $n = 500$. Using this (and R) we get that

$$\begin{aligned} 1 - \beta &= 0.9501796 \\ \therefore \beta &\approx 0.05 \,. \end{aligned}$$

---

   c. How large should the sample size be if we require that $\alpha = 0.01$ and $\beta = 0.05$, assuming $\mu = 5.5$?

Solutions:
This is a two-sided z-test. Using the formula from lecture, we have

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{(\mu_a - \mu_0)^2} = 684.7765.$$

Hence, we need a sample of size 685 to provide the desired levels.

---

3. A study is to be conducted to investigate the amount of toxic chemicals in freshwater lakes. A common measure of toxicity for any pollutant is LC50 (lethal concentration killing 50% of test species), which is the concentration of the pollutant that will kill half of the test species in a given amount of time (usually 96 hours for fish species). In many studies, the natural logarithm of LC50 measurements, log(LC50), are normally distributed. For copper, the variance of log(LC50) measurements is around 0.4 mg/L (milligrams per litre) on fish species A and around 0.8 mg/L for fish species B.

a. Suppose 10 samples were collected for species A. Find the probability that the sample mean of log(LC50) will differ from the population mean by no more than 0.5.

Solutions:
Let $\bar{X}_A$ denote the sample mean of the species A sample, and μ_A denote the population mean for this species. We have $\sigma_A^2 = 0.4$ and sample size $n = 10$.

$$\begin{aligned}
P(|\bar{X}_A - \mu_A| \leq 0.5) &= P\left(\left|\frac{\bar{X}_A - \mu_A}{\frac{\sigma_A}{\sqrt{n}}}\right| \leq \frac{0.5}{\frac{\sigma_A}{\sqrt{n}}}\right) \\
&= P(|Z| \leq 2.5) \\
&= P(-2.5 \leq Z \leq 2.5) \\
&= 1 - 2P(Z > 2.5) \\
&= 1 - 2(0.00621) \\
&= 0.9876
\end{aligned}$$

---

b. If we want the sample mean (for species A) to differ from the population by no more than 0.5 with probability 0.95, how many samples do we need to collect?

Solutions:
We want

$$\begin{aligned}
P(|\bar{X}_A - \mu_A| \leq 0.5) &= P\left(|Z| \leq \frac{0.5}{\sqrt{\frac{0.4}{n}}}\right) \\
&= 1 - P\left(|Z| > \frac{0.5}{\sqrt{\frac{0.4}{n}}}\right) \\
&= P\left(-\frac{0.5}{\sqrt{0.4}}\sqrt{n} \leq Z \leq \frac{0.5}{\sqrt{0.4}}\sqrt{n}\right) \\
&= 0.95
\end{aligned}$$

We know that $P(|Z| > 1.96) = 0.05$. Hence it follows that $\frac{0.5}{\sqrt{\frac{0.4}{n}}} = 1.96$, which implies $n = \frac{0.4}{\left(\frac{0.5}{1.96}\right)^2} = 6.15$. Hence, we need to collect 7 samples.

---

c. Assuming the population mean for both species is the same, what is the probability that the sample mean of species A will exceed the sample mean of species B by at least 1 mg/L, if we collected 10 measurements from each species?

Solutions:
Let $\bar{Y}_B$ denote the sample mean of the species B sample. We have $E(\bar{X}_A - \bar{Y}_B) = \mu_A - \mu_B = 0$. Also, since $\bar{X}_A$ and $\bar{Y}_B$ are independent, then $\text{Var}(\bar{X}_A - \bar{Y}_B) = \frac{\sigma_A^2 + \sigma_B^2}{n} = \frac{0.4+0.8}{10} = 0.12$. Hence,

$$\begin{aligned}
P(\bar{X}_A - \bar{Y}_B \geq 1) &= P\left(\frac{\bar{X}_A - \bar{Y}_B}{\sqrt{0.12}} \geq \frac{1}{\sqrt{0.12}}\right) \\
&= P\left(Z \geq \frac{1}{\sqrt{0.12}}\right) \\
&= P(Z \geq 2.89) \\
&= 0.0019
\end{aligned}$$

4. Suppose $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ are independent random samples, with $X_i \sim N(\mu_1, \sigma_1^2)$ and $Y_j \sim N(\mu_2, \sigma_2^2)$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.

   a. State $E(\bar{X} - \bar{Y})$.

Solutions:
$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2$

   b. State $\text{Var}(\bar{X} - \bar{Y})$.

Solutions:
As $\bar{X}$ and $\bar{Y}$ are independent,

$$\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\,.$$

   c. What is the sample size needed so that $(\bar{X} - \bar{Y})$ will be within k units of $(\mu_1 - \mu_2)$ with probability $1 - \alpha$? You may assume $m = n$.

Solutions:
It is required that $P(|(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)| \le k) = \alpha$. Observe that $\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$. Taking $n = m$, we have

$$\begin{aligned}
1 - \alpha &= P(|(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)| \le k) \\
&= P\left(\left|\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}}\right| \le \frac{k}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}}\right) \\
&= P\left(|Z| \le \frac{k}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}}\right)\,.
\end{aligned}$$

Equivalently, this is

$$P\left(|Z| \ge \frac{k}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}}\right) = \alpha\,.$$

Since $Z \sim N(0, 1)$, it follows that

$$\frac{k}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n}}} = z_{\alpha/2}\,,$$

and hence

$$n = \frac{(\sigma_1^2 + \sigma_2^2) z_{\alpha/2}^2}{k^2}\,.$$