STAT2371 Assignment

Ze Hong Zhou (46375058)

2022-08-25 (last edited 2022-08-27)

```
library(tidyverse)
```

The original source code for this assignment can be found here after I make the repo public.

Question 1

Suppose that two independent binomial random variables X_1 and X_2 are observed where X_1 has a Binomial(n, p) distribution and X_2 has a Binomial($2n, p$) distribution. You may assume that n is known, whereas p is an unknown parameter. Define two possible estimators of p

$$\hat{p}_1 = \frac{1}{3n}(X_1 + X_2) \quad \textbf{and} \quad \hat{p}_2 = \frac{1}{2n}(X_1 + 0.5X_2).$$

(a) Show that both of the estimators $\hat{p}_1$ and $\hat{p}_2$ are unbiased estimators of p.

$$\begin{aligned}
E(\hat{p}_1) &= \frac{1}{3n}(E(X_1) + E(X_2)) \quad \text{applying expected value linearity} \\
&= \frac{1}{3n}(n \cdot p + 2n \cdot p) \quad \text{applying the binomial expectation formula} \\
&= p \\
\text{bias}(\hat{p}_1, p) &= E(\hat{p}_1) - p = 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
E(\hat{p}_2) &= \frac{1}{2n}(E(X_1) + 0.5E(X_2)) \\
&= \frac{1}{2n}(n \cdot p + 0.5 \cdot 2n \cdot p) \\
&= p \\
\text{bias}(\hat{p}_2, p) &= E(\hat{p}_2) - p = 0.
\end{aligned}$$

Hence $\hat{p}_1$ and $\hat{p}_2$ are unbiased estimators of p.

(b) Find Var($\hat{p}_1$) and Var($\hat{p}_2$).

$$\begin{aligned}
\mathrm{Var}(\hat{p}_1) &= \frac{1}{9n^2}(\mathrm{Var}(X_1) + \mathrm{Var}(X_2)) \quad \text{applying the formula for independent case} \\
&= \frac{1}{9n^2}(np(1-p) + 2np(1-p)) \quad \text{applying the binomial variance formula} \\
&= \frac{p(1-p)}{3n}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{Var}(\hat{p}_2) &= \frac{1}{4n^2}(\mathrm{Var}(X_1) + 0.5^2\mathrm{Var}(X_2)) \\
&= \frac{1}{4n^2}(np(1-p) + 0.25 \cdot 2np(1-p)) \\
&= \frac{3p(1-p)}{8n}.
\end{aligned}$$

(c) Show that both estimators are consistent estimators of p.

Let $\varepsilon > 0$.

$$\begin{aligned}
\lim_{n\to\infty} P(|\hat{p}_1 - p| > \varepsilon) &\leq \lim_{n\to\infty} \frac{E\left((\hat{p}_1 - p)^2\right)}{\varepsilon^2} \quad \text{applying Markov's inequality} \\
&= \lim_{n\to\infty} \frac{\mathrm{Var}(\hat{p}_1)}{\varepsilon^2} \quad \text{since } \hat{p}_1 \text{ is unbiased} \\
&= \lim_{n\to\infty} \frac{p(1-p)}{3n\varepsilon^2} \\
&= 0 \quad \text{for all } p \in [0,1].
\end{aligned}$$

Applying the squeeze theorem, $\lim_{n\to\infty} P(|\hat{p}_1 - p| > \varepsilon) = 0$. Similarly,

$$\begin{aligned}
\lim_{n\to\infty} P(|\hat{p}_2 - p| > \varepsilon) &\leq \lim_{n\to\infty} \frac{E\left((\hat{p}_2 - p)^2\right)}{\varepsilon^2} \\
&= \lim_{n\to\infty} \frac{\mathrm{Var}(\hat{p}_2)}{\varepsilon^2} \\
&= \lim_{n\to\infty} \frac{3p(1-p)}{8n\varepsilon^2} \\
&= 0 \quad \text{for all } p \in [0,1].
\end{aligned}$$

So $\lim_{n\to\infty} P(|\hat{p}_2 - p| > \varepsilon) = 0$. Hence $\hat{p}_1$ and $\hat{p}_2$ are weakly consistent estimators of p.

(d) Show that $\hat{p}_1$ is the most efficient estimator among all unbiased estimators.

$\underset{\sim}{X} \equiv \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, $\underset{\sim}{x} \equiv \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

$$\begin{aligned}
f_{\underset{\sim}{X}}(\underset{\sim}{x}, p) &= f_{X_1}(x_1, p) \cdot f_{X_2}(x_2, p) \quad \text{applying the r.v. independence definition} \\
&= \begin{cases} \binom{n}{x_1} p^{x_1}(1-p)^{n-x_1} \cdot \binom{2n}{x_2} p^{x_2}(1-p)^{2n-x_2} & \text{if } x_1 \in [0,n] \cap \mathbb{N} \text{ and } x_2 \in [0,2n] \cap \mathbb{N} \\ 0 & \text{if otherwise} \end{cases} \\
&= \begin{cases} \binom{n}{x_1}\binom{2n}{x_2} p^{x_1+x_2}(1-p)^{3n-(x_1+x_2)} & \text{if } x_1 \in [0,n] \cap \mathbb{N} \text{ and } x_2 \in [0,2n] \cap \mathbb{N} \\ 0 & \text{if otherwise} \end{cases} \\
&= \begin{cases} \binom{n}{x_1}\binom{2n}{x_2} exp\left(\begin{bmatrix} ln(p) \\ ln(1-p) \end{bmatrix}^T \begin{bmatrix} x_1 + x_2 \\ 3n - (x_1 + x_2) \end{bmatrix} \right) & \text{if } x_1 \in [0,n] \cap \mathbb{N} \text{ and } x_2 \in [0,2n] \cap \mathbb{N} \\ 0 & \text{if otherwise} \end{cases}
\end{aligned}$$

Since n is known and fixed, $\underset{\sim}{X}$ has a pdf in the exponential family, and any sufficient static is also complete. $X_1 + X_2$ is thus sufficient and complete by the sufficient statistic factorisation theorem. By the Lehmann-Scheffé theorem, $E(\hat{p}_1 \mid X_1 + X_2) = \hat{p}_1$ is the unique MVUE, i.e. it is the most efficient estimator among all unbiased estimators.

(e) Derive the efficiency of the estimator $\hat{p}_1$ relative to $\hat{p}_2$.

$$\begin{aligned}
\text{eff}(\hat{p}_2, \hat{p}_1, p) &= \frac{\text{MSE}(\hat{p}_2, p)}{\text{MSE}(\hat{p}_1, p)} = \frac{\text{Var}(\hat{p}_2) + (\text{bias}(\hat{p}_2, p))^2}{\text{Var}(\hat{p}_1) + (\text{bias}(\hat{p}_1, p))^2} = \frac{\frac{3p(1-p)}{8n}}{\frac{p(1-p)}{3n}} \\
&= \frac{9}{8}
\end{aligned}$$

Question 2

The random variables $X_1, X_2, \ldots, X_{2n}$ are independent and normally distributed with common variance σ^2. However, $X_1, X_2, \ldots, X_n$ have mean 0 while $X_{n+1}, X_{n+2}, \ldots, X_{2n}$ have mean μ.

(a) Write down the joint pdf of $X_1, X_2, \ldots, X_{2n}$ and hence the likelihood function and log-likelihood of (μ, σ^2).

$$\begin{aligned}
f_{\underset{\sim}{X}}(\underset{\sim}{x}) &= \prod_{i=1}^{n}(f_{X_i}(x_i)) \prod_{i=n+1}^{2n}(f_{X_i}(x_i)) \quad \text{since each element of } \underset{\sim}{X} \text{ is independent} \\
&= \prod_{i=1}^{n}\left((2\pi\sigma^2)^{-1/2}\text{exp}\left[\frac{-1}{2}\left(\frac{x_i}{\sigma}\right)^2\right]\right) \prod_{i=n+1}^{2n}\left((2\pi\sigma^2)^{-1/2}\text{exp}\left[\frac{-1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]\right) \\
&= (2\pi\sigma^2)^{-n}\text{exp}\left(\frac{-1}{2}\left[\sum_{i=1}^{n}\left(\left[\frac{x_i}{\sigma}\right]^2\right) + \sum_{i=n+1}^{2n}\left(\left[\frac{x_i-\mu}{\sigma}\right]^2\right)\right]\right)
\end{aligned}$$

$$L\left(\underset{\sim}{\theta} \equiv \begin{bmatrix} m \\ s^2 \end{bmatrix} \mid \underset{\sim}{X}\right) = (2\pi s^2)^{-n}\exp\left(\frac{-1}{2}\left[\sum_{i=1}^{n}\left(\left[\frac{X_i}{s}\right]^2\right) + \sum_{i=n+1}^{2n}\left(\left[\frac{X_i - m}{s}\right]^2\right)\right]\right)$$

$$l\left(\underset{\sim}{\theta} \mid \underset{\sim}{X}\right) = -n\ln(2\pi s^2) - \frac{1}{2}\left[\sum_{i=1}^{n}\left(\left[\frac{X_i}{s}\right]^2\right) + \sum_{i=n+1}^{2n}\left(\left[\frac{X_i - m}{s}\right]^2\right)\right]$$

(b) Show that the maximum likelihood estimators of μ and σ^2 are

$$\hat{\mu} = \frac{1}{n}\sum_{j=n+1}^{2n} \quad \textbf{and} \quad \hat{\sigma}^2 = \frac{1}{2n}\left(\sum_{j=1}^{2n}(X_j^2) - n\hat{\mu}^2\right).$$

$$s\left(\underset{\sim}{\theta} \mid \underset{\sim}{X}\right) = \begin{bmatrix} \frac{\partial}{\partial m}\left[\frac{-1}{2}\sum_{i=n+1}^{2n}\left(\left[\frac{X_i - m}{s}\right]^2\right)\right] \\ \frac{\partial}{\partial s^2}\left[-n\ln(2\pi s^2) - \frac{1}{2}\left[\sum_{i=1}^{n}\left(\left[\frac{X_i}{s}\right]^2\right) + \sum_{i=n+1}^{2n}\left(\left[\frac{X_i - m}{s}\right]^2\right)\right]\right] \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=n+1}^{2n}\left(\frac{X_i - m}{s^2}\right) \\ \frac{-n}{s^2} + \frac{1}{2}\left[\sum_{i=1}^{n}\left(\left[\frac{X_i}{s^2}\right]^2\right) + \sum_{i=n+1}^{2n}\left(\left[\frac{X_i - m}{s^2}\right]^2\right)\right] \end{bmatrix}$$

$$\underset{\sim}{0} = s\left(\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} \mid \underset{\sim}{X}\right)$$

$$\underset{\sim}{0} = \begin{bmatrix} \sum_{i=n+1}^{2n}(X_i) - n\hat{\mu} \\ -n\hat{\sigma}^2 + \frac{1}{2}\left[\sum_{i=1}^{n}\left(X_i^2\right) + \sum_{i=n+1}^{2n}\left([X_i - \hat{\mu}]^2\right)\right] \end{bmatrix}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=n+1}^{2n}(X_i) \\ \frac{1}{2n}\left[\sum_{i=1}^{n}\left(X_i^2\right) + \sum_{i=n+1}^{2n}\left(X_i^2 - 2X_i\hat{\mu} + \hat{\mu}^2\right)\right] \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n}\sum_{i=n+1}^{2n}(X_i) \\ \frac{1}{2n}\left[\sum_{i=1}^{n}\left(X_i^2\right) + \sum_{i=n+1}^{2n}(X_i^2) - 2n\hat{\mu}^2 + n\hat{\mu}^2\right] \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n}\sum_{i=n+1}^{2n}(X_i) \\ \frac{1}{2n}\left[\sum_{i=1}^{2n}\left(X_i^2\right) - n\hat{\mu}^2\right] \end{bmatrix}$$

Assuming that this point gives a maximum of L, $\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=n+1}^{2n}(X_i) \\ \frac{1}{2n}\left[\sum_{i=1}^{2n}\left(X_i^2\right) - n\hat{\mu}^2\right] \end{bmatrix}$ is the MLE of $\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$.

(c) Derive the Cramér-Rao lower bounds for the variances of unbiased estimators of $\tau(\mu, \sigma^2)$.

$$
\begin{aligned}
I\left(\begin{bmatrix}\mu\\ \sigma^2\end{bmatrix}\right) &= -E\left(\frac{\partial s}{\partial \underset{\sim}{\theta}^T}\left(\underset{\sim}{\theta} = \begin{bmatrix}\mu\\ \sigma^2\end{bmatrix} \mid \underset{\sim}{X}\right)\right)\\
&= -E\left(\begin{bmatrix} \sum_{i=n+1}^{2n}\left(\frac{-1}{\sigma^2}\right) & -\sum_{i=n+1}^{2n}\left(\frac{X_i-\mu}{\sigma^4}\right) \\ -\sum_{i=n+1}^{2n}\left(\frac{X_i-\mu}{\sigma^4}\right) & \frac{n}{\sigma^4} - \left[\sum_{i=1}^{n}\left(\left[\frac{X_i}{\sigma^3}\right]^2\right) + \sum_{i=n+1}^{2n}\left(\left[\frac{X_i-\mu}{\sigma^3}\right]^2\right)\right]\end{bmatrix}\right)\\
&= \begin{bmatrix}\frac{n}{\sigma^2} & 0 \\ 0 & \frac{-n}{\sigma^4} + \sigma^{-6}\left[\sum_{i=1}^{n}\left(E\left(X_i^2\right)\right) + \sum_{i=n+1}^{2n}\left(E\left(\left[X_i-\mu\right]^2\right)\right)\right]\end{bmatrix}\\
&= \begin{bmatrix}\frac{n}{\sigma^2} & 0 \\ 0 & \frac{-n}{\sigma^4} + \sigma^{-6}\left[\sum_{i=1}^{n}\left(\sigma^2+0^2\right) + \sum_{i=n+1}^{2n}\left(\sigma^2\right)\right]\end{bmatrix}\\
&= \begin{bmatrix}\frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{\sigma^4}\end{bmatrix}\\
\left(I\left(\begin{bmatrix}\mu\\ \sigma^2\end{bmatrix}\right)\right)^{-1} &= \begin{bmatrix}\frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^4}{n}\end{bmatrix}
\end{aligned}
$$

The CRLB for the variances of unbiased $\tau(\mu, \sigma^2)$ estimators is

$$
\begin{bmatrix}\frac{\partial \tau}{\partial \mu} & \frac{\partial \tau}{\partial \sigma^2}\end{bmatrix}\begin{bmatrix}\frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^4}{n}\end{bmatrix}\begin{bmatrix}\frac{\partial \tau}{\partial \mu} \\ \frac{\partial \tau}{\partial \sigma^2}\end{bmatrix} = \left(\frac{\partial \tau}{\partial \mu}\right)^2\frac{\sigma^2}{n} + \left(\frac{\partial \tau}{\partial \sigma^2}\right)^2\frac{\sigma^4}{n}.
$$

(d) Briefly explain whether μ or σ^2 have unbiased estimators that attain the relevant Cramér-Rao lower bounds.

From the expression of $s\left(\underset{\sim}{\theta} \mid \underset{\sim}{X}\right)$, it depends linearly on

$$
\sum_{i=1}^{2n}\left(X_i^2\right), \quad \text{and} \quad \sum_{i=n+1}^{2n}\left(X_i\right).
$$

$$
\begin{aligned}
E\left(\sum_{i=1}^{2n}\left(X_i^2\right)\right) &= 2n\sigma^2 + n\mu^2\\
E\left(\sum_{i=n+1}^{2n}\left(X_i\right)\right) &= n\mu
\end{aligned}
$$

So only unbiased estimators of $a(2\sigma^2 + \mu^2) + b\mu + c$, where a, b, and c are real numbers, can attain the CRLB. Setting $(a = 0, b = 1, c = 0)$, there exist unbiased estimators for μ that attain the CRLB variance, but there is no such combination for σ^2.

(e) For either maximum likelihood estimator in (b) above that is biased, find the unique minimum variance unbiased estimator, and explain why it is so.

$$\begin{aligned} E\left(\hat{\sigma}^2\right) &= \frac{1}{2n}\left(2n\sigma^2 + n\mu^2 - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) \\ &= \frac{2n-1}{2n}\sigma^2 \end{aligned}$$

Hence the estimator for σ^2 is biased. From the pdf of $\underset{\sim}{X}$, the exponent of e can be split into the product of a function of $\begin{bmatrix}\mu \\ \sigma^2\end{bmatrix}$ alone and another function of $\underset{\sim}{X}$ alone. Hence the pdf is in the exponential family and any sufficient static is also complete. Via the sufficient statistic factorisation theorem, it is also apparent from the exponent that a sufficient and complete statistic is

$$S \equiv \begin{bmatrix} \sum_{i=1}^{2n}\left(X_i^2\right) \\ \sum_{i=n+1}^{2n}\left(X_i\right) \end{bmatrix}.$$

By the Lehmann-Scheffé theorem, the unique MVUE is

$$\begin{aligned} E\left(\frac{2n}{2n-1}\hat{\sigma}^2 \mid S\right) &= \frac{1}{2n-1}E\left(\sum_{j=1}^{2n}(X_j^2) - n\hat{\mu}^2 \mid S\right) \\ &= \frac{1}{2n-1}\left(\sum_{j=1}^{2n}(X_j^2) - n\hat{\mu}^2\right) \quad \text{(conditioning on } S \text{ gives no variability).} \end{aligned}$$

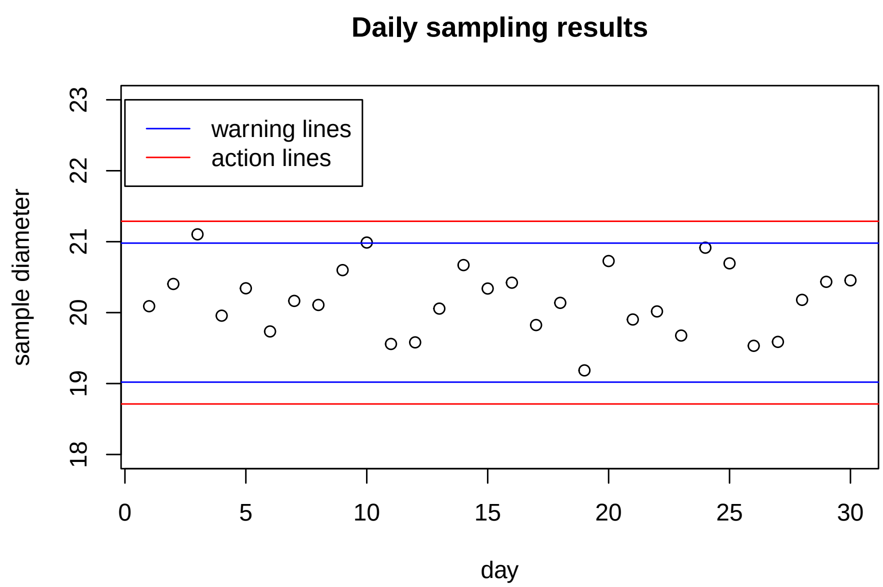
Question 3

(a) Suppose components are manufactured to a target diameter 20 cm. An earlier process capability assessment has shown that the standard deviation of the diameter is 0.5 cm. At the end of each day, a component is chosen at random from the day's production, and the diameter is measured to ensure that the process has not moved off target. In R, generate a data set of size 30 from Normal($\mu = 20, \sigma^2 = 0.25$), set the observations into a vector called data1. Plot the data and the appropriate target, warning and action lines using R function abline; label the lines. Comment on the results.

```r
set.seed(46375058)
#generate random samples
data1 <- rnorm(30, 20, sqrt(0.25))
#warning lines
warn <- qnorm(c(0.025, 0.975), 20, sqrt(0.25))
#action lines
act <- qnorm(c(0.005, 0.995), 20, sqrt(0.25))

#plot data points
plot(x = 1:30, y = data1, main = "Daily sampling results",
     xlab = "day", ylab = "sample diameter", ylim = c(18, 23))

#plot warning and action lines
abline(h = warn, col = "blue")
abline(h = act, col = "red")
#create legend
legend(x = 0, y = 23, legend = c("warning lines", "action lines"),
       col = c("blue", "red"), lty = 1)
```



Two data points (6.67% of data points) lie outside the warning lines, and the rest fall inside. This is to be expected with random variation.

(b) In R, generate data sets of size 30 from

(i) Normal($\mu = 20.2$, $\sigma^2 = 0.25$),

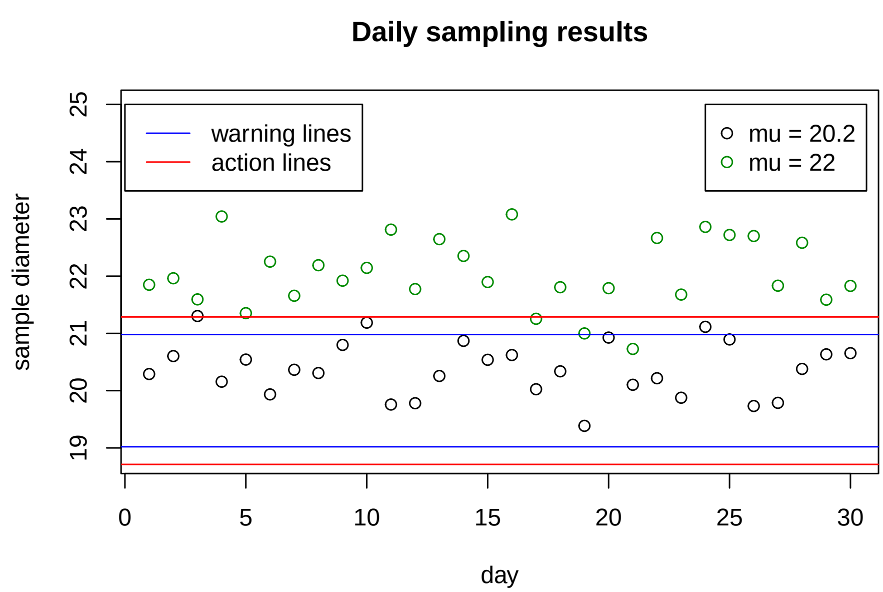
(ii) Normal($\mu = 22$, $\sigma^2 = 0.25$),

set the observations into vectors called data2.i and data2.ii, respectively. Plot the data and the appropriate target, warning and action lines; label the lines. Comment on the results.

```
set.seed(46375058)
#generate random samples
data2.i <- rnorm(30, 20.2, sqrt(0.25))
data2.ii <- rnorm(30, 22, sqrt(0.25))

#plot data points
plot(x = 1:30, y = data2.i, main = "Daily sampling results",
     xlab = "day", ylab = "sample diameter", ylim = c(18.8, 25))
points(x = 1:30, y = data2.ii, col = "green4")

#plot warning and action lines
abline(h = warn, col = "blue")
abline(h = act, col = "red")
#create legend
legend(x = 0, y = 25, legend = c("warning lines", "action lines"),
       col = c("blue", "red"), lty = 1)
legend(x = 24, y = 25, legend = c("mu = 20.2", "mu = 22"),
       col = c("black", "green4"), pch = 1)
```



As μ increases, there are a greater number of points above both the warning and action lines.

(c) This quality control procedure can be treated as a sequence of independent hypothesis testing $H_{0,i}$: the process is under control ($\mu = 20$) versus

(i) $H_{1,i}$: the process is out of control ($\mu = 20.2$),

(ii) $H_{1,i}$: the process is out of control ($\mu = 22$),

for $i = 1, \ldots, 30$. Suppose that the quality control procedure rejects $H_{0,i}$ if the observation d_i is outside the action lines. For both (i) and (ii), find α, probability of Type I error, and β, probability of Type II error.

Let X represent the data from a single sample.

$$\begin{aligned}\alpha &= P\left(\left|\frac{X-20}{\sigma}\right| > z_{0.995} \mid H_0\right)\\ &= P(|Z| > z_{0.995})\\ &= 2*(1-0.995) = 0.01\end{aligned}$$

When $\mu = 20.2$,

$$\begin{aligned}\beta &= P\left(\left|\frac{X-20}{\sigma}\right| < z_{0.995} \mid \mu = 20.2\right)\\ &= P\left(20 - z_{0.995}\sigma < X < 20 + z_{0.995}\sigma \mid \mu = 20.2\right)\\ &= 0.9838.\end{aligned}$$

The inline code used is `(pnorm(20 + qnorm(0.995)*sqrt(0.25), 20.2, sqrt(0.25)) - pnorm(20 - qnorm(0.995)*sqrt(0.25), 20.2, sqrt(0.25))) %>% round(4)`.

Similarly, when $\mu = 22$,

$$\begin{aligned}\beta &= P\left(\left|\frac{X-20}{\sigma}\right| < z_{0.995} \mid \mu = 22\right)\\ &= P\left(20 - z_{0.995}\sigma < X < 20 + z_{0.995}\sigma \mid \mu = 22\right)\\ &= 0.0772.\end{aligned}$$

The inline code used is `(pnorm(20 + qnorm(0.995)*sqrt(0.25), 22, sqrt(0.25)) - pnorm(20 - qnorm(0.995)*sqrt(0.25), 22, sqrt(0.25))) %>% round(4)`.

(d) Let us consider this problem within the Bayesian framework. Suppose that, after analysing a relevant archive of data, we know that P(the process is under control) = 0.1 and P(the process is out of control) = 0.9. Using Bayes' formula, for both (i) and (ii), calculate

$$\begin{aligned}R_0 &= P(\textbf{the process is under control} \mid H_1 \textbf{ is accepted}),\\ R_1 &= P(\textbf{the process is out of control} \mid H_0 \textbf{ is accepted}).\end{aligned}$$

Comment on which pair, (α,β) or (R_0,R_1), would better reflect producer's and customer's risks.

$$\begin{aligned}R_0 &= \frac{P(\text{the process is under control and } H_1 \text{ is accepted})}{P(H_1 \text{ is accepted})} \quad \text{applying Bayes' formula}\\ &= \frac{P(\text{the process is under control and } H_1 \text{ is accepted})}{P(\text{the process is under control and } H_1 \text{ is accepted}) + P(\text{the process is out of control and } H_1 \text{ is accepted})}\\ &= \frac{0.1\cdot 0.01}{0.1\cdot 0.01 + 0.9\cdot(1-0.9838)} = 0.001128\end{aligned}$$

Similarly,

$$
\begin{aligned}
R_1 &= \frac{P(\text{the process is out of control and } H_0 \text{ is accepted})}{P(H_0 \text{ is accepted})} \\
&= \frac{P(\text{the process is out of control and } H_0 \text{ is accepted})}{P(\text{the process is out of control and } H_0 \text{ is accepted}) + P(\text{the process is under control and } H_0 \text{ is accepted})} \\
&= \frac{0.9 \cdot 0.0772}{0.9 \cdot 0.0772 + 0.1 \cdot 0.99} = 0.412389
\end{aligned}
$$

(e) Consider the cusum procedure that rejects H_0: $\mu = 20$, if $|S_t| > c_t$, where c_t is a threshold. Under H_0, find a sequence of the thresholds c_t such that P(Type I error) = 0.01 for $t = 1, \ldots, 30$. In R, plot the thresholds c_t and the cusum charts for data1 and data2.i. Provide comments.

Let X_i represent the data from the sample on day i.

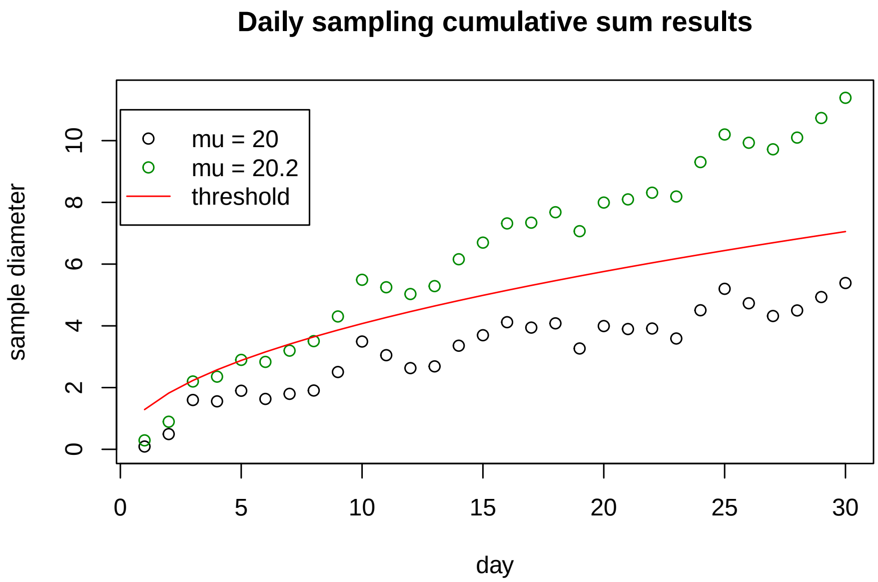
$$
\begin{aligned}
0.01 &= P(|S_t| > c_t \mid \mu = 20) \\
P(|Z| > z_{0.995}) &= P\left(\left|\frac{\sum_{i=1}^{t}(X_i) - t \cdot 20}{\sqrt{t}\sigma}\right| > \frac{c_t}{\sqrt{t}\sigma} \mid \mu = 20\right) = P\left(|Z| > \frac{c_t}{\sqrt{t}\sigma}\right) \\
c_t &= \sqrt{t}\sigma z_{0.995}
\end{aligned}
$$

```
#compute thresholds
thr <- sqrt(1:30) * sqrt(0.25) * qnorm(0.995)

#compute cumulative sums
cs_d1 <- cumsum(data1 - 20)
cs_d2i <- cumsum(data2.i - 20)

#plot points
plot(1:30, cs_d1, main = "Daily sampling cumulative sum results",
     xlab = "day", ylab = "sample diameter", ylim = c(0, 11.5))
points(x = 1:30, y = cs_d2i, col = "green4")
#plot threshold
lines(1:30, thr, col = "red")
lines(1:30, -thr, col = "red")

#create legend
legend(x = 0, y = 11, legend = c("mu = 20", "mu = 20.2", "threshold"),
       col = c("black", "green4", "red"), lty = c(0, 0, 1), pch = c(1, 1, NA))
```



The data points for $\mu = 20$ all lie below the threshold line while the data for $\mu = 20.2$ lies above the threshold line when $t = 5$ and when $t \geq 9$. The accumulated error when $\mu \neq 20$ is obvious as the days progress.