M248 June 2018

2021-04-22

Updates

2021-05-01: Solution refactored for new reference style. Removed needless formula manipulations in solutions.

2021-04-27: Partial solution uploaded.

Setup

Path to script: **ljk233/AutomatingM248/src/r/calculators.r**

```
source("../src/r/calculators.r")
```

Section A

q.01

A. *Calculating probabilities using a probability function (HB p.7)* **:** ***p.d.f.***

If $f(x) = \frac{2}{9}(1+x), / > x \in (-1, 2)$ then the probability $P(0 \leq X < 1)$ is given by

$$\begin{aligned} P(0 \leq X < 1) &= \int_0^1 \frac{2}{9}(1+x)\,dx \\ &= \frac{2}{9}\left[x + \frac{1}{2}x^2\right]_0^1 \\ &= \frac{2}{9}\left(1 + \frac{1}{2} - (0+0)\right) = \frac{1}{3}. \end{aligned}$$

```
f <- function (x) {(2/9)*(1+x)}  # declare the fn
integrate(f, 0, 1)  # integrate over range
```

```
## 0.3333333 with absolute error < 3.7e-15
```

q.02

C. *Normalising constants (Book A, p.111)*

Normalising constant K is given by

$$1 = \int_0^1 \frac{1}{K}x^4\,dx = \frac{1}{K}\int_0^1 x^4\,dx.$$

Therefore, K is

$$K = \int_0^1 x^4\,dx = \left[\frac{1}{5}x^5\right]_0^1 = \frac{1}{5}(1-0) = \frac{1}{5}.$$

```
f <- function (x) {(5)*(x**4)}  # declare the fn
integrate(f, 0, 1)  # integrate over range (0,1)
```

```
## 1 with absolute error < 1.1e-14
```

q.03

F. *Calculating probabilities using the c.d.f. :* ***discrete***

The $P(X \leq 4)$ is given by

$$\begin{aligned} P(X \leq 4) = F(4) &= p(0) + \cdots + p(4) \\ &= 0.1 + \cdots + 0.1 \\ &= 0.6. \end{aligned}$$

q.04

D. *Calculating probabilities using the c.d.f. :* ***continuous***

The probability $P(1 \leq X \leq 2) = F(2) - F(1)$ is given by

$$\begin{aligned} F(2) - F(1) &= 1 - \frac{1}{3}\sqrt{9 - 2^2} - \left(1 - \frac{1}{3}\sqrt{9 - 1^2}\right) \\ &= 1 - \frac{1}{3}\sqrt{5} - 1 + \frac{1}{3}\sqrt{8} \\ &= \frac{1}{3}(\sqrt{8} - \sqrt{5}). \end{aligned}$$

q.05

E. *Choosing a model based on the range of the standard probability models (HB p.26, 27)*

The range of X is $\{1, 2, 3, 4, 5\}$.

- Cannot be a *Bernoulli*, as range of Bernoulli is $\{0, 1\}$.
- Cannot be a *binomial*, as range of Binomial includes 0.
- Cannot be a *geometric* or *Poisson*, as these have no upper boundaries (X has max value of 5).
- Cannot be a continuous, as X is discrete.

Therefore it must be a **discrete uniform distribution.**

q.06

E. *Mean of a rv (HB p.7) :* ***discrete***

The $E(X) = \sum_x x\, p(x)$, so

$$\sum_x x\, p(x) = 0(0.1) + 1(0.25) + \cdots + 4(0.2) = 2.15.$$

```
x <- seq.int(from = 0, to = 4)
f <- c(0.1, 0.25, 0.25, 0.2, 0.2)
sum(x*f)
```

```
## [1] 2.15
```

q.07

E. *Mean of a linear function of a rv (HB p.9)*

For a random variable $Y = 20 - 3X$, then

$$E(Y) = aE(X) + b,$$

where $E(X) = 1$, $a = -3$, and $b = 20$. Therefore,

$$E(Y) = -3(1) + 20 = 17.$$

q.08

B. *Variance of a linear function of rvs (HB p.9)*

For a random variable $Y = 20 - 3X$, then

$$S(Y) = \sqrt{V(Y)} = \sqrt{a^2 V(X)} = |a|\sqrt{V(X)},$$

where $V(X) = 4$ and $a = -3$. Therefore,

$$S(Y) = |-3|\sqrt{4} = 6.$$

q.09

B. *Poisson distribution (HB p.8) and Poisson process (HB p.10)*

Let X model the number of email arriving in the office per hour. Then $X \sim \text{Poisson}(\mu)$, where $\mu = 5$ is the average number of emails received in an hour.

The probability $P(X = 3)$ is given by

$$p(3) = e^{-5}\left(\frac{5^3}{3!}\right) \simeq 0.140.$$

```r
dpois(3, 5)
```

```
## [1] 0.1403739
```

q.10

D. *Poisson distribution (HB p.8) and Poisson process (HB p.10)*

The $P(X < 3) = P(X \leq 2)$ is given by

$$\begin{aligned} P(X \leq 2) = \sum_{i=0}^{2} e^{-5}\left(\frac{5^i}{i!}\right) &= e^{-5}\sum_{i=0}^{2}\frac{5^i}{i!} \\ &= e^{-5}\left(\frac{5^0}{0!} + \frac{5^1}{1!} + \frac{5^2}{2!}\right) \\ &= e^{-5}(1 + 5 + 12.5) \\ &\simeq 0.125. \end{aligned}$$

```
ppois(2, 5)
```

```
## [1] 0.124652
```

q.11

D. *Exponential distribution (HB p.10) and Poisson process (HB p.10)*

Let T model the waiting time between emails arriving in the office. Then $T \sim M(\lambda)$, where $\lambda = 5$ is the average number of emails received in an hour.

Ten minutes is 1/6 hours, so $P(T < 1/6)$ is given by

$$P\left(T \leq \frac{1}{6}\right) = F\left(\frac{1}{6}\right) = 1 - e^{-5\left(\frac{1}{6}\right)} \simeq 0.565.$$

```
pexp(1/6, 5)
```

```
## [1] 0.5654018
```

q.12

F. *Transforming the parameter of a Poisson distribution (HB p.10)*

The number of emails that will arrive in 3 hours is distributed Poisson(λt), where $\lambda = 5$ emails per hour and $t = 3$. Therefore Poisson(15).

q.13

A. *Population quantiles of a rv (HB p.11) :* ***continuous***

The α-quantile of a continuous rv X with c.d.f. $F(X)$ is defined as $F(x) = \alpha$.

If $F(x) = 1 - \sqrt{1 - x}$ and $\alpha = q_L = 1/4$, then it must be that

$$\begin{aligned}
F(x) = \alpha \rightarrow 1 - \sqrt{1-x} &= \frac{1}{4} \\
1 - \frac{1}{4} &= \sqrt{1-x} \\
\left(\frac{3}{4}\right)^2 &= 1 - x \\
x &= 1 - \frac{9}{16} \\
&= \frac{7}{16}.
\end{aligned}$$

```
F <- function (x) {1 - sqrt(1 - x)};
F(x = 7/16)
```

```
## [1] 0.25
```

q.14

C. *Difference between two independent normal rvs (HB p.11)*

If $X \sim N(2, 4)$, $Y \sim N(1, 3)$, then $U = X - Y$ will also have a normal distribution with parameters

$$E(U) = E(X - Y) = E(X) - E(Y) = 2 - 1 = 1,$$

and

$$V(U) = V(X - Y) = V(X) + V(Y) = 4 + 3 = 7.$$

Hence $U = X - Y \sim N(1, 7)$.

q.15

C. *Probabilities for a normal distribution (HB p.12)*

We know that $X \sim N(100, 15^2)$. Then the probability $P(90 < X < 125)$ is given by

$$\begin{aligned} P(X < 125) - P(X < 90) &= P\left(Z < \frac{125 - 100}{15}\right) - P\left(Z < \frac{90 - 100}{15}\right) \\ &\simeq \Phi(1.67) - \Phi(-0.67) \\ &= \Phi(1.67) - (1 - \Phi(0.67)) \\ &= 0.9525 + 0.7486 - 1 \simeq 0.701. \end{aligned}$$

```
mean_ <- 100; std <- 15;
pnorm(125, mean_, std) - pnorm(90, mean_, std)
```

```
## [1] 0.6997171
```

```
## Small error due to rounding when transforming X -> Z
```

q.16

B. *Distribution of the sample mean (HB p.12)*

We know $\mu_X = 69.1$, $\sigma_X^2 = 9.4$ and $n = 40$. Then $\overline{X}_{40}$ will have the distribution $\overline{X}_{40} \sim N(\mu_X, \sigma_X^2/n) = N(69.1, 0.235)$.

And so $P(X < 68)$ is given by

$$\begin{aligned} P(X < 68) &= P\left(Z < \frac{68 - 69.1}{\sqrt{0.235}}\right) \\ &\simeq \Phi(-2.27) \\ &= 1 - \Phi(2.27) \\ &= 1 - 0.9884 \simeq 0.012. \end{aligned}$$

```
pnorm(68, 69.1, sqrt(9.4/40))
```

```
## [1] 0.01163031
```

q.17

C. *Distribution of the sample total (HB p.13) and quantiles of any non-standard normal (HB p.12)*

The sample total T_n is distributed $T_n \sim N\big(n\mu, (\sqrt{n}\sigma)^2\big)$.

If $X \sim N(\mu, \sigma^2)$, then the α-quantile x is given by $x = \sigma q_\alpha + \mu$.

Therefore the α-quantile of s will be given by

$$s = q_\alpha \sqrt{n\sigma^2} + n\mu = q_\alpha \sigma \sqrt{n} + n\mu.$$

q.18

C. *Variance of an unbiased estimator (U7.1, Ex.6)*

If $X_1 \sim N(\mu, 1)$, $X_2 \sim N(\mu, 4)$, $X_3 \sim N(\mu, 4)$ and $\widehat{\mu} = \frac{1}{6}(4X_1 + X_2 + X_3)$, then

$$\begin{aligned} V(\widehat{\mu}) = V\left(\frac{1}{6}(4X_1 + X_2 + X_3)\right) &= \left(\frac{1}{6}\right)^2 V(4X_1 + X_2 + X_3) \\ &= \frac{1}{36}\Big\{V(4X_1) + V(X_2) + V(X_3)\Big\} \\ &= \frac{1}{36}\{4^2\, V(X_1) + V(X_2) + V(X_3)\} \\ &= \frac{1}{36}\{16\,(1) + 4 + 4\} \\ &= \frac{1}{36}\{24\} \\ &= \frac{2}{3}. \end{aligned}$$

q.19

A. *Transforming a confidence interval (HB p.14)*

The transformation $X = \frac{5}{9}(Y - 32)$ is both linear and increasing.

When $Y = 245.3$,

$$X = \frac{5}{9}(245.3 - 32) = 118.5,$$

and $Y = 249.8$,

$$X = \frac{5}{9}(249.8 - 32) = 121.0.$$

Therefore, a new 95% confidence interval in degrees Celsius is $(118.5, 121.0)$.

q.20

E. *Critical values (HB p.16) of a t-test (HB p.17)*

The critical value of a one-tail test t-test is the $(1 - \alpha)$-quantile of $t(\nu)$, where ν is the degrees of freedom. The test is at a 1% significance level so $\alpha = 0.01$, and has a sample size $n = 101$, meaning $q_{1-\alpha} = 0.99$ and $\nu = n - 1 = 101 - 1 = 100$.

Therefore the critical value is the **0.99**-quantile of $t(100)$, which is **2.364.**

```
qt(0.99, 100)
```

```
## [1] 2.364217
```

q.21

Significance level of a hypothesis test (HB p.17) and Interpreting a p-values (HB p.18)

B. False. p-value is the *significance probability*, not the probability H_0 is correct.

E. False. The significance level is $100\alpha = 1\%$, so $\alpha = 0.01$. We can see that $p = 0.02 > 0.01 = \alpha$, so H_0 is not rejected.

q.22

A. *Calculating the power of a hypothesis test (HB p.18)*

We know from the following from question that the:

- test is **one-sided**
- sample size $n = 20$;
- standard deviation $\sigma = 0.25$
- test is at a **5%** significance level
  - so $q_{1-\alpha} = q_{0.95} = 1.645$
- true value of the mean $\mu = 0.2$, and the hypothesised mean is $\mu_0 = 0$
  - so $d = \mu_0 - \mu = 0.2$.

The power of the test will therefore be

$$\begin{aligned}1 - \Phi\left(q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right) &= 1 - \Phi\left(1.645 - \frac{0.2}{0.25/\sqrt{20}}\right)\\ &\simeq 1 - \Phi(-1.93)\\ &= 1 - \{1 - \Phi(1.93)\}\\ &= \Phi(1.93)\\ &= 0.9732.\end{aligned}$$

```
calc_test_power(
  a = 0.05, d = 0.2, sd = 0.25, n = 20, one_sided = TRUE)
```

```
## [1] 0.973373
```

```
## Error due to rounding
```

q.23

A. *Choosing the sample size of a hypothesis test (HB p.19)*

We know the following from the question that the:

- test is **two-sided**
- standard deviation $\sigma = 1$
- test is at a 5% significance level
  - so $q_{1-\alpha/2} = q_{0.975} = 1.960$
- discrepancy $d = 0.25$
- power is $85\% \equiv 0.85$
  - so $q_{1-\gamma} = q_{1-0.85} = q_{0.15} = -q_{0.85} = -1.036$

The required sample size n will therefore be

$$\begin{aligned}n = \frac{\sigma}{d^2}(q_{1-\alpha/2} - q_{1-\gamma})^2 &= \frac{1}{0.25^2}(1.960 - -1.036)^2\\ &= 16(1.960 + 1.036)^2\\ &= 143.6 \simeq 144.\end{aligned}$$

```
calc_sample_size(
  a = 0.05, d = 0.25, sd = 1, pow_ = 85, one_sided = FALSE)
```

```
## [1] 143.6544
```

q.24

C. *Degrees of freedom of a **Chi-squared** goodness-of-fit test (HB p.21)*

The degrees of freedom of a **chi-squared** goodness-of-fit test is $k - p - 1$, where

- $k = 4$ is the number of categories
- $p = 1$ is the number of estimated parameters
  - p was estimated

Therefore the degrees of freedom is $\nu = 4 - 1 - 1 = 2$.

q.25

F. *Calculating the test statistic of **Chi-squared** goodness-of-fit test (HB p.21)*

The **chi-squared** test statistic is

$$\chi^2 = \sum \frac{(O_i - Ei)^2}{E_i},$$

so in this case, the **chi-squared** test statistic will be

$$\chi^2 = \frac{(44 - 55.4)^2}{55.4} + \cdots + \frac{(10 - 6.2)^2}{6.2} \simeq 6.94.$$

```
obs <- c(44, 27, 19, 10)
exp <- c(55.4, 24.7, 13.7, 6.2)

calc_chi_sq(obs, exp)
```

```
## [1] 6.939416
```

q.26

A. *Calculating S_{xx}, S_{yy} and S_{xy} (HB p.22)*

The value of S_{yy} is given by

$$S_{yy} = \sum y_i^2 - \frac{\{\sum y_i\}^2}{n}.$$

We know that $n = 14$, $\sum y_i^2 = 34692$, and $\sum y_i = 600$. Therefore,

$$S_{yy} = 34692 - \frac{600^2}{14} = 8977.7142\ldots \simeq 8977.7.$$

```
calc_slr_std(sum_y = 600, sum_sq = 34692, n = 14)
```

```
## [1] 8977.714
```

q.27

C. *Calculating S_{xx}, S_{yy}, and S_{xy} (HB p.22)*

Value of S_{xy} is given by

$$S_{xy} = \sum x_i\, y_i - \frac{\sum x_i \sum y_i}{n}.$$

We know that $n = 14$, $\sum x_i\, y_i = 6912$, $\sum x_i = 118$, and $\sum y_i = 600$. Therefore,

$$S_{xy} = 6912 - \frac{(118)(600)}{14} = 1854.8571\ldots \simeq 1854.9.$$

```
calc_slr_std(sum_x = 118, sum_y = 600, sum_sq = 6912, n = 14)
```

```
## [1] 1854.857
```

q.28

B. *Confidence intervals for $\widehat{\beta}$ (HB p.22)*

A $100(1-\alpha)\%$ confidence interval for $\widehat{\beta}$ is given by

$$\widehat{\beta} \pm t\frac{s}{\sqrt{S_{xx}}}.$$

We know that $\widehat{\beta} = 0.76\; s^2 = 91.5$, $S_{xx} = 574.29$, and $t \simeq 2.086$ is the 0.975-quantile of $t(20)$. So,

$$\beta^- = 0.76 - 2.086\left(\sqrt{\frac{91.5}{574.24}}\right) \simeq -0.072,$$

and

$$\beta^+ = 0.76 + 2.086\left(\sqrt{\frac{91.5}{574.24}}\right) \simeq 1.593.$$

Therefore $(-0.072, 1.593) \approx (-0.08, 1.60)$.

```
calc_conf_int_beta(
  ci = 0.95, b = 0.76, n = 21, Sxx = 574.24, var = 91.5)
```

```
## [1] -0.0726654  1.5926654
```

```
## Small error due to rounding
```

q.29

F. *Elements of a statistical report (HB p.24)*

> The **Discussion** should contain your assessment of the statistical evidence relating to the original question or hypothesis.

q.30

A. *Elements of a statistical report (HB p.24)*

> The **Method** should include [...] the statistical test used to check the model...

Section B

q.31

(a) *Binomial distribution (HB p.8)*

Let X be a discrete rv that represents the number of apples that passes Ping's acceptability criteria. Then X is modelled by the binomial distribution, $X \sim B(n, p)$, where $n = 5$ is the sample size and $p = 0.4$ is the probability that an individual apple passes the acceptability criteria, so $X \sim B(5, 0.4)$.

The probability $P(X = 2)$ is given by

$$P(X = 2) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{5}{2} 0.4^2 (1-0.4)^3 = 0.3456.$$

```r
dbinom(2, 5, 0.4)
```

```
## [1] 0.3456
```

(b) *Geometric distribution (HB p.8)*

Let Y be a discrete rv that represents the sample number of the apple that passes Ping's acceptability criteria in a sample. Then Y is modelled by the geometric distribution, $Y \sim G(p)$, where $p = 0.4$ is the probability that an individual apple passes the acceptability criteria, so $Y \sim G(0.4)$.

The probability $P(Y = 3)$ is given by

$$P(Y = 3) = (1-p)^{y-1} p = 0.6^2 0.4 = 0.144.$$

```r
dgeom(2, 0.4)  # note, @param x -> (x-1) as defined by M248
```

```
## [1] 0.144
```

(c) *Expected value of a standard probability model (HB p.26, 27)*

The expected number of apples, $E(Y)$, Ping is needed to examine before finding one that meets her acceptability criteria is given by

$$E(Y) = \frac{1}{p} = \frac{1}{0.4} = 2.5.$$

q.32

Variance of a rv (HB p.9) : ***continuous***

The variance $V(X)$ of a continuous random variable X is defined as

$$V(X) = E\{(X-\mu)^2\} = E(X^2) - E(X)^2$$

We know that $E(X) = 3/5$, so let us calculate $E(X^2)$,

$$
\begin{aligned}
E(X^2) &= \int_0^1 x^2 \left\{12x^2(1-x)\right\} dx \\
&= 12 \int_0^1 x^4 - x^5 \, dx \\
&= 12 \left[\frac{1}{5}x^5 - \frac{1}{6}x^6\right]_0^1 \\
&= 12 \left(\frac{1}{5}(1)^5 - \frac{1}{6}(1)^6 - (0-0)\right) \\
&= 12 \left(\frac{1}{30}\right) \\
&= \frac{2}{5}.
\end{aligned}
$$

And so,

$$
V(X) = E(X^2) - E(X)^2 = \frac{2}{5} - \left(\frac{3}{5}\right)^2 = \frac{1}{25}.
$$

Therefore, $V(X) = 1/25$.

q.33

(a) *Finding the likelihood function of a sample (HB p.13)*

If $f(x;\theta) = \theta e^{-\theta x}$, then the likelihood of θ, $L(\theta)$, will be

$$
\begin{aligned}
L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) &= \theta e^{-\theta x_1} \times \theta e^{-\theta x_2} \times \theta e^{-\theta x_3} \times \theta e^{-\theta x_4} \\
&= \theta e^{-\theta(4.5)} \times \theta e^{-\theta(1.5)} \times \theta e^{-\theta(6)} \times \theta e^{-\theta(4.4)} \\
&= \theta \times \theta \times \theta \times \theta \times e^{-\theta(4.5)} \times e^{-\theta(1.5)} \times e^{-\theta(6)} \times e^{-\theta(4.4)} \\
&= \theta^4 \, e^{-16.4\theta}.
\end{aligned}
$$

Hence, we can see that $L(\theta) = \theta^4 \, e^{-16.4\theta}$.

(b) *Finding the MLE of an estimator (HB p.13)*

If $L(\theta) = \theta^4 \, e^{-16.4\theta}$ then, by the **product rule**, $L'(\theta)$ will be given by

$$
\begin{aligned}
L'(\theta) &= \theta^4(-16.4 \, e^{-16.4\theta}) + 4\,\theta^3 \, e^{-16.4\theta} \\
&= 4\,\theta^3 \, e^{-16.4\theta} - 16.4\,\theta^4 \, e^{-16.4\theta} \\
&= \theta^3 e^{-16.4\theta}\{4 - 16.4\,\theta\}.
\end{aligned}
$$

Comparing this with the form of the equation given in the question, $L'(\theta) = \theta^a e^{-b\theta} L_1(\theta)$, we can see that

- $a = 3$
- $b = 16.4$
- $L_1(\theta) = 4 - 16.4\,\theta$

(c) *Finding the MLE of an estimator (HB p.13)*

The MLE of θ, $\widehat{\theta}$, is the solution to

$$L'(\theta) = \theta^3 e^{-16.4\theta}\{4 - 16.4\,\theta\} = 0.$$

Given the range of $\theta > 0$, then $\theta^3, e^{-16.4\theta} > 0$, so this reduces to solving

$$\begin{aligned} 0 &= 4 - 16.4\,\theta \\ 16.4\,\theta &= 4 \\ \theta &= \frac{4}{16.4} \simeq 0.244. \end{aligned}$$

(d) *MLE of a standard probability distribution (HB p.26m 27)*

The MLE of an exponential distribution is $\widehat{\lambda} = 1/\overline{X}$.

We can see from the data that $\overline{x} = \frac{1}{4}(4.5 + \cdots + 4.4) = 16.4/4$.

And so, $1/\overline{x} = \frac{1}{16.4/4} = \frac{4}{16.4} \simeq 0.244$.

This value matches that given in q.33(c).

q.34

(a) *Assumptions for two-sample t-intervals (HB p.16)*

The *assumption of an equal variance* is valid if the larger of the sample variances divided by the smaller is less than **3.**

We have been told $s_C^2 = 103.84$ and $s_N^2 = 115.99$, so

$$\frac{s_N^2}{s_C^2} = \frac{115.99}{103.84} \simeq 1.117 < 3.$$

Therefore, the assumption of equal variances is valid.

(b) *Exact confidence intervals for the difference between two normal means (HB p.16)*

Given two independent samples from distributions with a common variance, the pooled estimate of the common variance is given by

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

We have been given sample variances $s_C^2 = 103.84$ and $s_N^2 = 115.99$, and sample sizes $n_C = 7$ and $n_N = 13$, so

$$\begin{aligned} s_P^2 &= \frac{(7-1)103.84 + (13-1)115.99}{7 + 13 - 2} \\ &= \frac{(6)103.84 + (12)115.99}{18} \\ &= \frac{2014.92}{18} \\ &= 111.94. \end{aligned}$$

And so the pooled estimate of the common population standard deviation is $\sqrt{11.94} \simeq 10.58$ years.

```r
calc_pooled_sample_var(
  v1 = 103.84, n1 = 7, v2 = 115.99, n2 = 13)**(0.5)
```

```
## [1] 10.58017
```

(c) *Exact confidence intervals for the difference between two normal means (HB p.16)*

To calculate a $90\% = 100(1-\alpha)\%$ exact confidence interval, we require the $(1-(\alpha/2))$-quantile of $t(\nu)$, where ν is the degrees of freedom.

Let us calculate α,

$$\begin{aligned}90 &= 100(1-\alpha)\\ \frac{90}{100} &= 1-\alpha\\ \alpha &= 1-0.9 = 0.1.\end{aligned}$$

There will be $n_C + n_P - 2 = 7 + 13 - 2 = 18$ degrees of freedom for the t-distribution.

Therefore, we require the 0.95-quantile of $t(18)$, which is $q_{0.95} = 1.734$.

```r
qt(0.95, 18)
```

```
## [1] 1.734064
```

(d) *Exact confidence intervals for the difference between two normal means (HB p.16)*

We know that:

- $d = \overline{x}_C - \overline{x}_N = 7.21$
- $n_C = 7$, $n_N = 13$
- $t \simeq 1.734$
- $s_P = \sqrt{111.94}$,

so an exact confidence interval for the difference between two normal means, $d = \mu_1 - \mu_2$, is given by

$$\begin{aligned}d^- &= d - t\,s_P\sqrt{\frac{1}{n_C}+\frac{1}{n_N}} = 7.21 - \left\{1.734(111.94)^{\frac{1}{2}}\left(\frac{1}{7}+\frac{1}{13}\right)^{\frac{1}{2}}\right\} \simeq -1.39,\\ d^+ &= d + t\,s_P\sqrt{\frac{1}{n_C}+\frac{1}{n_N}} \simeq 15.81.\end{aligned}$$

Hence, a 90% two-sample t-interval for the difference between the population mean age of patients of the type in the study who had had a coronary event and those who have not is approximately $(-1.39, 15.81)$.

It can be seen the realised 90% confidence interval for d contains 0, so it does not support a claim that the population ages differ.

```r
tconfint_diff_means (
  a = 0.9, x1 = 49.29, n1 = 7, x2 = 42.08, n2 = 13, sp = sqrt(111.94))
```

```
## [1] -1.39106 15.81106
```

(e) *Repeated experiments interpretation of confidence intervals (HB p.14)*

If a large number of samples of size 7 and size 13 were drawn independently from the populations of patients in Groups C and N, respectively, and the mean difference and a 90% confidence interval for the mean difference was found, then approximately 90% of these intervals would contain the true mean difference in patient's age.

The 90% confidence interval actually observed, (-1.39, 15.81), is just one observation on a random interval, and may or may not contain the population mean.

q.35

1. It is a **one-sided test** as $H_1 : p > 0.5$.
2. A p-value of 0.055 corresponds to "weak or little evidence against the null hypothesis", not "little to no evidence against the null hypothesis".
3. A hypothesis test does not *prove* the null hypothesis to be true or not; it is instead a test as to whether to reject or not reject H_0.

q.36

(a) *The Mann–Whitney test (HB p.20)*

Let the hypotheses be

$$H_0 : \ell = 0, \; H_1 : \ell \neq 0$$

where ℓ is the underlying difference in location between the populations from which the samples were drawn.

(b) *The Mann–Whitney test (HB p.20)*

The test statistic U_A is the sum of the ranks in sample A, which has been defined as the horses of heavy weights.

Therefore,

$$u_A = 2 + 4 + \cdots + 19 = 96.$$

Therefore the test statistic is $u_A = 96$.

(c) *Normal approximation to the null distribution of the Mann–Whitney test statistic (HB p.20)*

The mean and variance of the null distribution of the test statistic, $E(U_A)$ and $V(U_A)$ respectively, are given by

$$\begin{aligned} E(U_A) = \frac{1}{2}(n_A)(n_A + n_B + 1) &= \frac{1}{2}(8)(8 + 11 + 1) \\ &= \frac{1}{2}(160) \\ &= 80, \end{aligned}$$

and

$$\begin{aligned} V(U_A) = \frac{1}{12}(n_A)(n_B)(n_A + n_B + 1) &= \frac{1}{12}(8)(11)(8 + 11 + 1) \\ &= \frac{1}{12}(1760) \\ &= 146.666\ldots \\ &\simeq 146.67. \end{aligned}$$

```
# return (z, Eu, Vu)
calc_norm_approx_mann_whit(u = 96, nA = 8, nB = 11)
```

```
## [1]   1.321157  80.000000 146.666667
```

(d) *Normal approximation to the null distribution of the Mann–Whitney test statistic (HB p.20)*

The z-value corresponding to the approximate standard normal null distribution for this test is given by

$$\begin{aligned} Z = \frac{U_A - E(U_A)}{\sqrt{V(U_A)}} &\simeq \frac{96 - 80}{\sqrt{146.67)}} \\ &= \frac{96 - 80}{\sqrt{146.67)}} \\ &= 1.32114\ldots \\ &\simeq 1.32. \end{aligned}$$

Therefore $Z \simeq 1.32$.

```
# return (z, Eu, Vu)
calc_norm_approx_mann_whit(u = 96, nA = 8, nB = 11)
```

```
## [1]   1.321157  80.000000 146.666667
```

(e) *Calculating p-values (HB p.18)*

The test is two-sided, and so, using the table of probabilities of the standard normal distribution in the Handbook, the p-value is therefore

$$\begin{aligned} p &= 2P(\ |Z| \geq 1.32\ ) \\ &= 2(1 - \Phi(1.32)) \\ &= 2(1 - 0.9066) \\ &= 0.1868. \end{aligned}$$

Given that $p = 0.1868$, there is little to no evidence against the null hypothesis that the location of the differences between light and heavy weighted horses is zero.

q.37

Multiple linear regression (HB p.23)

(a) *Interpreting the p-values in multiple regression (HB p.23)*

For the two-sided test of the null hypothesis $H_0 : \beta_1 = 0$, since $p < 0.001 < 0.01$, there is strong evidence to suggest that β_1 is not equal to zero. Likewise, for the two-sided test of the null hypothesis $H_0 : \beta_2 = 0$, since $p < 0.001 < 0.01$, there is strong evidence to suggest that β_2 is not equal to zero. Therefore, there is strong evidence that both explanatory variables(x_1 and x_1 together influence the annual profit of a small campany.

(b) There is no particular pattern in the residual plot, so it seems the assumption that the residuals come from a distribution with constant, zero mean an constant variance is a reasonable one.

The points in the normal probability lot roughly follows a straight line, so the assumption that the residuals are normally distributed is also a reasonable one.

(c) *Using a fitted multiple regression (HB p.23) model*

Using the fitted multiple regression line, a small company with $x_1 = 8$ and $x_2 = 64$ is predicted to have annual profits (in £100,000s)

$$y = 0.930 - 0.3662(8) + 0.03543(64) \simeq 0.268.$$

Given the actual value was $y = 0.292$, this gives a residual value of

$$w = 0.292 - 0.268 = 0.024.$$

(d) The new fitted model $y = \alpha + \beta e^{\lambda x}$ is not linear. Hence, the new model could not be used to fit the data using the method of least squares.