STAT5002 Weekly Independent Exercises - solution

Sheet 2 - Week 5

STAT5002

1 0-1 Box (specific example)

A box contains 10 tickets. 3 are $\boxed{1}$ and 7 are $\boxed{0}$. In the questions below, if necessary, round to 3 decimal places.

1.1

What is the mean μ and SD σ of the box (that is, of the list of numbers represented on the tickets in the box)?

`Answer:`

The sum of the numbers is 3, there are 10 numbers in all, so the mean is $\frac{3}{10} = 0.3$. This is also the *proportion of* $\boxed{1}$s.

The SD can be worked out in a few different ways. For any list of numbers $x_1, \dots, x_N$, the direct definition of the SD is

$$\text{SD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}\,,$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the mean.

Applying this to our list gives

$$
\begin{aligned}
\sigma &= \sqrt{\frac{1}{10}\left[(1-0.3)^2 + (1-0.3)^2 + (1-0.3)^2 + \underbrace{(0-0.3)^2 + \cdots + (0-0.3)^2}_{7\text{ terms}}\right]} \\
&= \sqrt{\frac{1}{10}\left[(3\times 0.7^2) + (7\times(-0.3)^2)\right]} \\
&= \sqrt{\frac{1}{10}\left[(3\times 0.49) + (7\times 0.09)\right]} \\
&= \sqrt{\frac{1}{10}\left[1.47 + 0.63\right]} \\
&= \sqrt{\frac{2.1}{10}} \\
&= \sqrt{0.21} \approx 0.458
\end{aligned}
$$

to 3 decimal places.

Another useful method of computing the SD is the use the formula

$$\text{SD} = \sqrt{\text{mean sq.} - (\text{mean})^2}$$

Since there are only $\boxed{0}$s and $\boxed{1}$s in the box, the mean and mean square are the same, i.e. both 0.3. So

$$\text{SD} = \sqrt{0.3 - (0.3^2} = \sqrt{0.3 - 0.09} = \sqrt{0.21}\,.$$

1.2

Suppose $n = 100$ tickets are drawn randomly, with replacement, yielding numbers $X_1, \dots, X_n$. Write $S = X_1 + \dots + X_n$ for the sum of the draws and $\bar{X} = S/n$ for the average of the draws.

1.2.1

What is $E(X_1)$?

`Answer:` For a single random draw X from a box with mean μ, $E(X) = \mu$. Here, X_1 behaves (individually) just like a single random draw from a box with mean $\mu = 0.3$, so $E(X) = 0.3$.

1.2.2

What is $SE(X_1)$?

Answer: For a single random draw X from a box with SD σ, $SE(X) = \sigma$. Here, X_1 behaves (individually) just like a single random draw from a box with SD $\sigma = \sqrt{0.21}$, so $SE(X) = \sqrt{0.21} \approx 0.46$.

1.2.3

What is $E(X_1 + X_2)$?

Answer: The sum $X_1 + X_2$ of two (independent) random draws from our box is like a *single* random draw from a bigger box: the box of all possible sums. This bigger box has mean $E(X_1) + E(X_2) = 2\mu = 0.6$.

1.2.4

What is $SE(X_1 + X_2)$?

Answer: For two independent random draws X_1 and X_2, since $SE(X_1) = SE(X_2) = \sigma = \sqrt{0.21}$,

$$SE(X_1 + X_2) = \sqrt{SE(X_1)^2 + SE(X_2)^2} = \sqrt{2}SE(X_1) = \sqrt{2}\,\sigma = \sqrt{0.42} \approx 0.648\,.$$

1.2.5

What is $E(S)$?

Answer: Extending the reasoning used to answer part (1.2.3) above,

$$E(S) = E(X_1 + \cdots + X_{100}) = E(X_1) + \cdots + E(X_{100}) = 100 \times 0.3 = 30\,.$$

1.2.6

What is $SE(S)$?

Answer: Extending the reasoning used to answer part (1.2.4) above,

$$SE(S)^2 = SE(X_1 + \cdots + X_{100})^2 = SE(X_1)^2 + \cdots + SE(X_{100})^2 = 100 \times 0.21 = 21\,.$$

Therefore $SE(S) = \sqrt{21} \approx 4.583$.

1.2.7

What is $E(\bar{X})$?

`Answer:` We know $E(S) = 30$ already, which is also the mean of the (very much) bigger box of all possible sample sums. $E(\bar{X})$ is, in turn the mean of the (also very big) box of all possible sample averages, each of which is a possible sum divided by 100. In other words, the box of all possible sample averages is obtained by dividing all possible sample sums by 100.

If we obtain a second list of numbers by multiplying/dividing a first list of numbers by a constant, the mean of the second list is also obtained by multiplying/dividing the mean of the first list by the same constant.

Thus,

$$E(\bar{X}) = E\left(\frac{S}{100}\right) = \frac{1}{100}E(S) = \frac{30}{100} = 0.3\,.$$

But we knew this already! $E(\bar{X}) = \mu$

1.2.8

What is $SE(\bar{X})$?

`Answer:`

Note that $SE(\bar{X})$ is also the SD of the (very big) box of all possible sample averages. We already know $SE(S)$, the SD of the (very big) box of all possible sample sums.

Similarly to the previous question, if a second list of numbers is obtained by multiplying/dividing a first list by a *postive* constant, the SD of the second list is obtained by multiplying/dividing the SD of the first list by the same constant. This is easily checked in general: if the first list is $x_1, \ldots, x_N$ and the second list $y_1, \ldots, y_N$ is obtained via $y_1 = cx_1$,...,$y_N = cx_N$, then the second list has mean

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} cx_i = c\frac{1}{N}\sum_{i=1}^{N} x_i = c\bar{x}$$

and it has SD

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(cx_i - c\bar{x})^2} = \sqrt{c^2\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = c\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

(if $c > 0$) which is simply c times the SD of the first list.

Therefore, we get

$$SE(\bar{X}) = SE\left(\frac{S}{100}\right) = \frac{1}{100}SE(S) = \frac{\sqrt{21}}{100} \approx 4.583100 \approx 0.046\,.$$

1.3

By appealling to the Central Limit Theorem, determine a value v such that the interval $0.3 \pm v$, i.e. $[0.3 - v, 0.3 + v]$, serves as an (approximate) 98% prediction interval for $\bar{X}$. In other words, find v such that

$$P\left\{0.3 - v \leq \bar{X} \leq 0.3 + v\right\} \approx 0.98\,.$$

The R output below may be useful for this.

```r
qnorm(0.95)
```

```
[1] 1.644854
```

```r
qnorm(0.975)
```

```
[1] 1.959964
```

```r
qnorm(0.98)
```

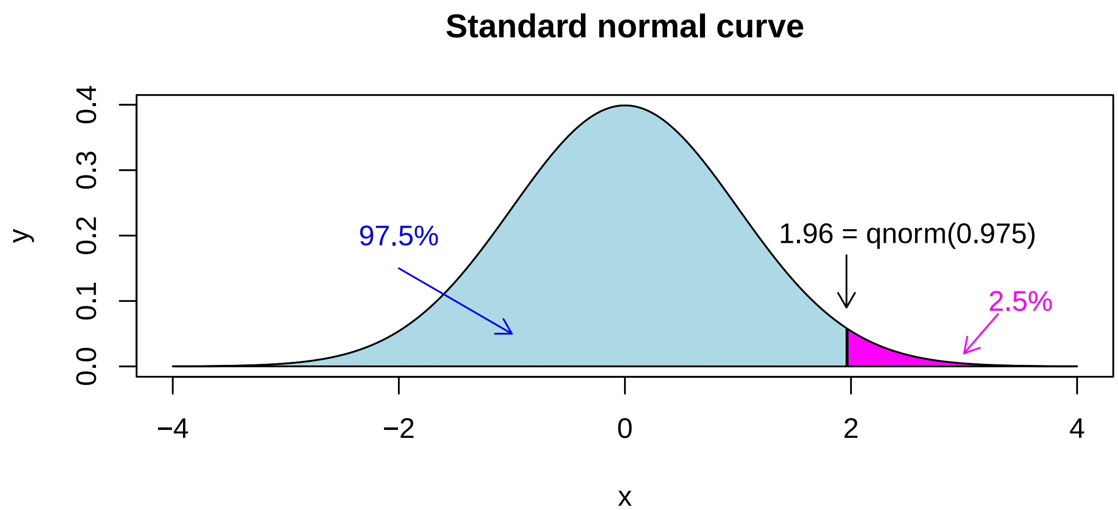
```
[1] 2.053749
```

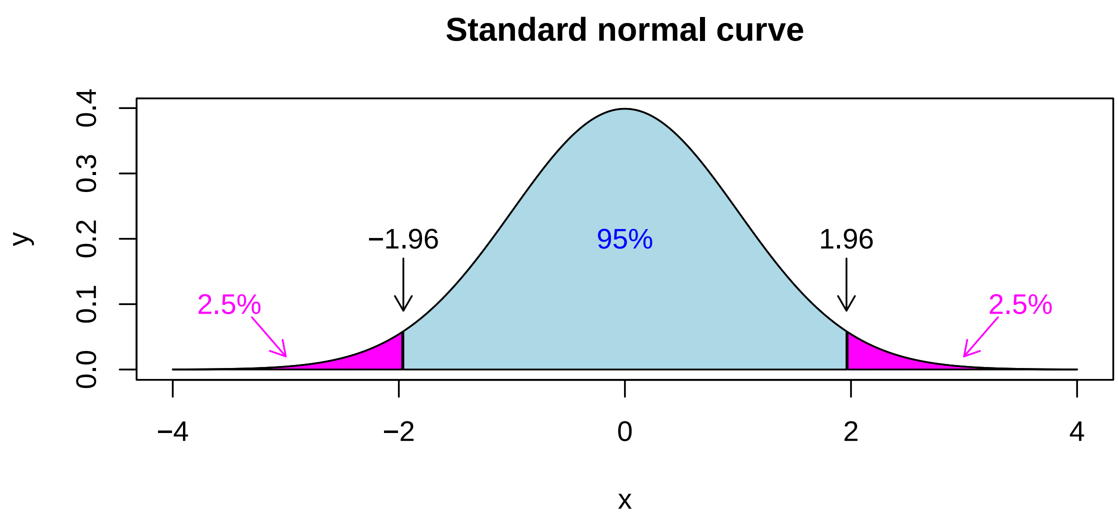
```r
qnorm(0.99)
```

```
[1] 2.326348
```

`Answer:`

Using the R output above, we have that for the **standard normal curve**, 99% of the area under the curve is to the left of 2.326, and 1% is to the right. By symmetry, another 1% is to the left of -2.326. Therefore there is 98% of the area under the standard normal curve between -2.326 and $+2.326$.



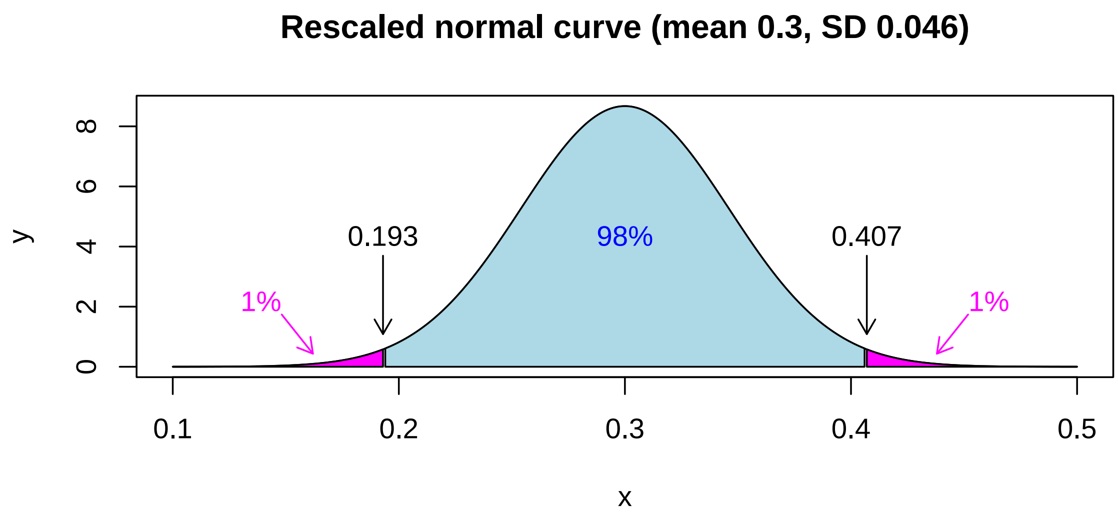


If a histogram follows the **standard** normal curve, the list of numbers represented has mean ≈ 0 and SD ≈ 1, **and** approximately 98% of the values lie between ± 2.326. But if a list of numbers has mean μ and SD σ, **and** has a normal shape, then approximately 98% of the values lie between $\mu \pm 2.326\sigma$.

We have shown above that the list of all possible sample averages has mean 0.3 and SD ≈ 0.046. Furthermore, since $n = 100$ is "quite large", the Central Limit Theorem tells us the histogram of these is approximately normally-shaped. Therefore approximately 98% of all possible sample averages lie in the interval

$$0.3 \pm 2.326 \times 0.046\,, \quad \text{i.e.} \quad 0.3 \pm 0.107\,,$$

giving the interval $[0.193, 0.407]$.



Finally, the "random experiment'' of drawing a random sample of size 100 and taking the sample average $\bar{X}$ is equivalent to a *single* draw from the box of all possible sample averages (described by the rescaled normal curve above), and so

$$P\left\{0.3 - 0.107 \le \bar{X} \le 0.3 + 0.107\right\}\,.$$

We should thus take $w = 0.107$.

2 0-1 Box (general case)

Repeat question 1, but for a box with N tickets: pN are $\boxed{1}$ and $(1-p)N$ are $\boxed{0}$. Write out answers to the questions below in terms of general sample size n and proportion of $\boxed{1}$s p.

2.1

What is the mean μ and SD σ of the box (that is, of the list of numbers represented on the tickets in the box)?

Answer:

There are N numbers in the box, their sum is $(pN \times 1) + ((1-p)N \times 0) = pN$ so the mean is $\mu = \frac{pN}{N} = p$.

The definition of the SD yields

$$\begin{aligned}
\sigma &= \sqrt{\frac{1}{N}\left[\underbrace{(1-p)^2 + \cdots + (1-p)^2}_{pN \text{ terms}} + \underbrace{(0-p)^2 + \cdots + (0-p)^2}_{(1-p)N \text{ terms}}\right]} \\
&= \sqrt{\frac{1}{N}\left[pN(1-p)^2 + (1-p)Np^2\right]} \\
&= \sqrt{\frac{p(1-p)N}{N}\left[(1-p)+p\right]} \qquad \text{(taking out } p(1-p)N \text{ as a common factor)} \\
&= \sqrt{p(1-p)}\,.
\end{aligned}$$

Alternatively, we can use the formula

$$\text{SD} = \sqrt{\text{mean sq.} - (\text{mean})^2}$$

Since we only have $\boxed{1}$ and $\boxed{0}$ in the box (both of which are unchanged by squaring) the mean and mean-square are the same, i.e. p, which gives

$$\sigma = \sqrt{p - p^2} = \sqrt{p(1-p)}\,.$$

2.2

Suppose n tickets are drawn randomly, with replacement, yielding numbers $X_1, \ldots, X_n$. Write $S = X_1 + \ldots + X_n$ for the sum of the draws and $\bar{X} = S/n$ for the average of the draws. You may assume that n is large enough that the Central Limit Theorem applies.

2.2.1

What is $E(X_1)$?

Answer:

For a single random draw X from a box with mean μ, $E(X) = \mu$. Here, X_1 behaves (individually) just like a single random draw from a box with mean $\mu = p$, so $E(X) = p$.

2.2.2

What is $SE(X_1)$?

Answer:

For a single random draw X from a box with SD σ, $SE(X) = \sigma$. Here, X_1 behaves (individually) just like a single random draw from a box with SD $\sigma = \sqrt{p(1-p)}$, so $SE(X) = \sqrt{p(1-p)}$.

2.2.3

What is $E(X_1 + X_2)$?

Answer:

$E(X_1 + X_2) = E(X_1) + E(X_2) = 2p$.

2.2.4

What is $SE(X_1 + X_2)$?

Answer:

$SE(X_1 + X_2) = \sqrt{SE(X_1)^2 + SE(X_2)^2} = \sqrt{2p(1-p)}$.

2.2.5

What is $E(S)$?

Answer:

$$
\begin{aligned}
E(S) &= E\left(X_1 + \cdots X_n\right) \\
) &= E(X_1) + \cdots E(X_n) \\
&= \underbrace{p + \cdots + p}_{n \text{ terms}} \\
&= np\,.
\end{aligned}
$$

2.2.6

What is $SE(S)$?

Answer:

$$\begin{aligned} SE(S) &= SE\left(X_1 + \cdots + X_n\right) \\ &= \sqrt{SE(X_1)^2 + \cdots SE(X_n)^2} \\ &= \sqrt{\underbrace{p(1-p) + \cdots + p(1-p)}_{n \text{ terms}}} \\ &= \sqrt{np(1-p)}\,. \end{aligned}$$

2.2.7

What is $E(\bar{X})$?

Answer:

$$E(\bar{X}) = E\left(\frac{1}{n}S\right) = \frac{1}{n}E(S) = \frac{np}{n} = p\,.$$

2.2.8

What is $SE(\bar{X})$?

Answer:

$$SE(\bar{X}) = SE\left(\frac{1}{n}S\right) = \frac{1}{n}SE(S) = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}\,.$$

2.3

By appealling to the Central Limit Theorem, determine a value v such that the interval $p \pm v$, i.e. $[p - v, p + v]$, serves as an (approximate) 98% prediction interval for $\bar{X}$. In other words, find v such that

$$P\left\{p - v \leq \bar{X} \leq p + v\right\} \approx 0.98\,.$$

The R output below question 1.3 may be useful for this.

Answer:

Drawing a random sample (with replacement) of size n and then obtaining the sample average is equivalent to taking a single random draw from the box of all possible sample averages, which (as we have shown above) has mean equal to $E(\bar{X}) = p$ and SD equal to $SE(\bar{X}) = \sqrt{\frac{p(1-p)}{n}}$. It is also "normal-shaped'' (since the Central Limit Theorem applies). Thus using the reasoning of question (1.3), roughly 98% of the values lie between $p \pm 2.326\sqrt{\frac{p(1-p)}{n}}$. Therefore, the probability $\bar{X}$ takes a value in this range is (approximately) 0.98. Thus we should take $w = 2.326\sqrt{\frac{p(1-p)}{n}}$.