Analysis of Scientific Data – Summer Semester, 2022

Week 3 Tutorial Solutions

Part A

A paper appearing in a 2015 issue of *Biological Conservation* used a randomised response technique to investigate rates of illegal fishing of red abalone in Northern California. The study involved interviews with people at various sites. Once verbal consent to participant was obtained, each respondent was given a coin and an envelope containing two cards. On one card was written the question

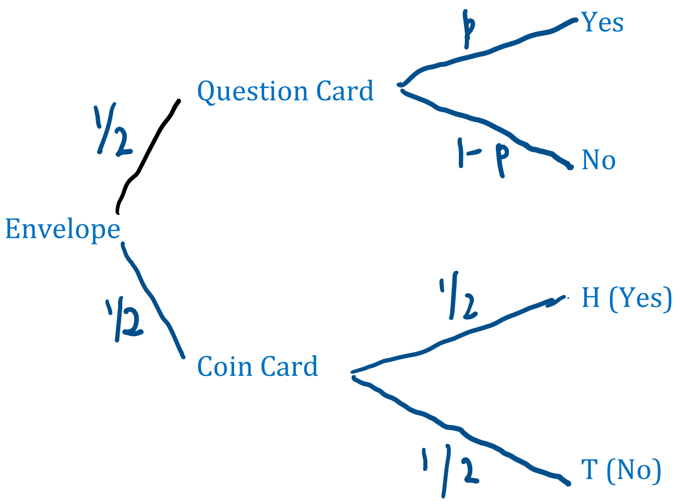
In the past year have you ever taken abalone under the minimum legal-size limit?

while on the other card was written the question

Did you get heads on the coin toss?

Without the interviewer watching, each respondent first tossed the coin, noting the outcome. They then chose a card at random and answered the question on it to the interviewer. Out of 279 respondents, 102 answered 'Yes'.

a) Denoting the unknown proportion by p, draw a tree diagram showing the conditional probabilities of answering 'Yes' following this procedure.



P(Question Card and Yes)= $p \times \frac{1}{2}$

P(Question Card and No) = $(1 - p) \times \frac{1}{2}$

P(Coin Card and Yes) = $\frac{1}{2} \times \frac{1}{2}$

P(Coin Card and No) = $\frac{1}{2} \times \frac{1}{2}$

P(Yes) = P(Question card and Yes) + P(Coin card and Yes)

P(Yes) = $\left(\frac{1}{2} \times p\right) + \left(\frac{1}{2} \times \frac{1}{2}\right)$

P(Yes) = $\frac{1}{2}p + \frac{1}{4}$

b) Based on your tree diagram in (a), what is the estimated proportion of people who have actually taken abalone under the minimum legal-size limit in the last year?

Estimate, P(Yes) = 102/279 = 0.366

That is; $\frac{1}{2}p + \frac{1}{4}$ = 0.366 Solving this, p = 0.232

Thus 23.2% of people have taken abalone under the minimum legal size limit.

Part B

An expensive piece of equipment in a laboratory is starting to show signs of age. Let X be the number of days in any week that the equipment is working and suppose that X has the following probability distribution:

x	0	1	2	3	4	5
$P(X = x)$	0.01	0.09	0.25	0.34	0.24	0.07

a) What is the expected number of days in a week that the equipment is working?

E(X) = $\sum_x x \times P(X = x)$

E(X) = $0 \times 0.01 + 1 \times 0.09 + \cdots + 5 \times 0.07$

E(X) = 2.92 days

b) What is the standard deviation of the number of days in a week that the equipment is working?

Var(X) = $\sum_x (x - E(X))^2 \times P(X = x)$

Var(X) = $(0 - 2.92)^2 \times 0.01 + (1 - 2.92)^2 \times 0.09 + \cdots + (5 - 2.92)^2 \times 0.07$

Var(X) = 1.2136

sd(X) = $\sqrt{1.2136}$
sd(X) = 1.102 days

c) What is the expected value of the total number of days that the equipment is working over a 40-week period?

Let $Y = X_1 + X_2 + \cdots + X_{40}$

Then $E(Y) = E(X_1) + E(X_2) + \cdots + E(X_{40})$

$E(Y) = 40 \times 2.92 = 116.8\ days$

d) Suppose the number of days in a week that the equipment is working is independent from week to week. What is the standard deviation of the total number of days that the equipment is working over a 40-week period?

$Var(Y) = Var(X_1) + Var(X_2) + \cdots + Var(X_{40})$ since independent

$Var(Y) = 40 \times 1.2136 = 48.544$

sd(Y) = $\sqrt{48.544}$ = 6.97 days

Alternatively, sd(Y) = $\sqrt{40} \times sd(X)$

Part C – Random Variables

Suppose a random person has a 1 in 3 chance of passing their driving test on each attempt. Let X be the number of attempts that a random person to pass their driving test. The probability distribution of X can be found using a simulation process. You can roll a six-sided die again and again until you see the number 5 or 6 and record the number of attempts. If you don't have a die use the sample(1:6, 1) in R count the number of attempts that you had to use this function to get 5 or 6. Share with your tutor the number of attempts you had to make to see 5 or 6. The tutor will share a "csv" file with you (in online tutorials) or show the values of X on white board (in face-to-face tutorials, you then need to type them and create your "csv" file).

a) Is X a discrete or a continuous random variable?

Discrete

b) Read the "csv" file into RStudio.

c) Use min() and max() in RStudio to find the minimum and maximum number of attempts to see number 5 or 6?

d) What is the shape of the distribution of X. You can use barplot() in RStudio to create a bar chart of X to determine the shape of X.

e) Use the prop.table() in RStudio to obtain the probability distribution function of X.

f) What is probability that a randomly selected person makes at least 2 attempts to pass the driving test?

g) Using the probability distribution of X from part c), what is the expected number of attempts taken, E(X)?

Part D – Binomial Distribution

a) Suppose 56% of STAT1201 students have brown eyes and we take a random sample of 4 students. Let X be the number of students with brown eyes in the sample. What is the distribution of X?

$$X \sim \text{Binomial}(4, 0.56)$$

This assumes independent samples, which is reasonable for random samples from a large population.

b) Let X be the number of towns in which it will rain tomorrow among five neighbouring towns. Is X a Binomial random variable?

No – rainfall in neighbouring towns is unlikely to be independent.

Many processes that evolve spatially or temporally violate the independence assumption. Can you think of any others?

c) Suppose 10% of people are left-handed and let X be the number of left-handed people in sample of 20 individuals. What is the probability of at least one left-handed person in the sample?

$$X \sim \text{Binomial}(20, 0.1)$$

We want $P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{20}{0} 0.1^0 (1 - 0.1)^{20-0} = 0.9^{20} = 0.878$

Using dbinom() in R, first find P(X=0):

```
dbinom(x=0, size=20, prob=0.1)
[1] 0.1215767
```

Then $P(X \geq 1) = 1$ −0.1215767 = 0.8784233

d) Suppose a drug has a 20% chance of making a person drowsy. Out of a sample of 80 people who each take the drug, what is the probability that no more than 10 of them experience drowsiness?

Assume individuals in the sample are independent. (Why might this not be the case? We might question the assumption of independence if all individuals in the sample were related.) Let X be the random variable representing the number of people that are made drowsy by the drug in a sample of 80 people. Then

$$X \sim \text{Binomial}(80, 0.2)$$

We want $P(X \leq 10)$ which can be evaluated in R as

```
pbinom(q=10,size=80,prob=0.2)
[1] 0.0564609
```

Alternatively,

```
sum(dbinom(x=0:10,size=80,prob=0.2))
[1] 0.0564609
```

e) In Part B, we computed the expected value and standard deviation of X, the number of days ina week (Mon – Fri) that a piece of equipment is working, to be 2.92 and 1.102, respectively. The probabilities given in the table actually came from a Binomial distribution. What are the parameters of the Binomial distribution?

The possible values for X were {0,1,2,…,5}. If $X \sim \text{Binomial}(n, p)$, then we must have *n* = 5. For a $\text{Binomial}(n, p)$ distribution we know $E(X) = np$. So

$$5p = 2.92$$

Solving for *p* gives

$$p = \frac{2.92}{5} = 0.584$$

So $X \sim \text{Binomial}(5{,}0.584)$. We could compare the probabilities from this binomial distribution with those reported in the table in **Part B**:

```
dbinom(x=0:5,size=5,prob=0.584)
[1] 0.01245 0.08744 0.24553 0.34468 0.24194 0.06793
```

Part E – Normal Distribution

Based on the student survey data, suppose that pulse rates while completing the survey come from a Normal distribution with mean 71.7 bpm and standard deviation 11.7 bpm.

It is possible to answer these questions without using the standardization procedure (calculating z-scores) if R is available. Since we will be using this procedure a lot when we come to hypothesis testing, it is important that you are comfortable doing these calculations.

a) What is the probability that a random student has a pulse rate of above 90 bpm?

Let X be the pulse rate of the students. Then X follows a Normal distribution with mean = 71.7 bpm and sd = 11.7 bpm.

$$P(X \geq 90) = P\left(\frac{X - 71.7}{11.7} \geq \frac{90 - 71.7}{11.7}\right) = P(Z \geq 1.564) = 1 - P(Z \leq 1.564)$$

In R we can calculate $P(Z \leq 1.564)$ as

```
pnorm(1.564)
[1] 0.9410912
```

So $P(Z \geq 1.564) = 1 - 0.9411 = 0.0589.$

b) What is the probability that a random student has pulse rate between 60 and 80 bpm?

$$P(60 \leq X \leq 80) = P\left(\frac{60 - 71.7}{11.7} \leq \frac{X - 71.7}{11.7} \leq \frac{80 - 71.7}{11.7}\right) = P(-1 \leq Z \leq 0.7094)$$

$$= P(Z \leq 0.7094) - P(Z \leq -1) = 0.7610 - 0.1587 = 0.6023$$

c) What is the highest pulse rate that of a student would be in the bottom 10% of pulse rates?

In this question we want to find the value q (quantile) such that

$$0.10 = P(X \leq q).$$

Again, we standardize X

$$0.1 = P\left(\frac{X - 71.7}{11.7} \leq \frac{q - 71.7}{11.7}\right) = P\left(Z \leq \frac{q - 71.7}{11.7}\right)$$

Using the qnorm function in R

```
qnorm(0.1)
[1] -1.281552
```

That is, $P(Z \leq -1.2816) = 0.1$. So we need to solve the equation

$$-1.281552 = \frac{q - 71.7}{11.7}$$

$$-1.281552 \times 11.7 = q - 71.7$$

$$q = 71.7 - 1.281552 \times 11.7 = 57$$

So the highest pulse rate for a student in the bottom 10% is 57 bpm.

d) Suppose that you are taking a random sample of 4 students from this survey. What is the probability that average pulse rate these 4 students is less than 80.5 bpm?

Note that we need to find $P(\bar{X} \leq 80.5)$. Thus, we first need the distribution of $\bar{X}$.

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

$$\bar{X} \sim N(71.7, \frac{11.7}{\sqrt{4}})$$
$$\bar{X} \sim N(71.7, 5.85)$$
$$P(\bar{X} \leq 80.5) = P(Z \leq \frac{\bar{X} - 71.1}{5.85})$$
$$P(\bar{X} \leq 80.5) = P(Z \leq 1.5)$$

Using pnorm(1.5) in R; the answer is 0.93.

e) In a random sample of 5 students, what is the probability that at least 3 of them have a pulse rate over 90 bpm while completing the survey?

From part (a) we know the probability that a student has a pulse rate greater than 90 bpm is 0.0589. Let X be the random variable representing the number of students having a pulse rate over 90 bpm in a sample of 5 students. Assuming students are selected independently at random from the student population, X has a Binomial(5, 0.0589) distribution.

We want $P(X \geq 3)$ which can be evaluated in R as

```
sum(dbinom(x=3:5,size=5,prob=0.0589)
[1] 0.001867087
```

Reference

Several of these questions are adapted from Section 6.6 of *Mind on Statistics*, a useful source of other practice questions. See the ECP for details.