

ISDS Individual Assignment

Chen-Wei Huang

2025-10-26

Q1.(a) For each of the three variables, size, time, and treats, state whether each one is categorical, ordinal, discrete or continuous.

Variable Type Summary

Variable	Type	Explanation
size	Ordinal (categorical)	Represents ordered size categories (Small < Medium < Large < Very large).
time	Continuous (numerical)	Measured in minutes; can take any real numeric value.
treats	Discrete (numerical)	Integer count of treats given.

```
dog_1 <- read.csv("Dog_Data_1.csv")
dog_2 <- read.csv("Dog_Data_2.csv")

dog_1$size <- factor(dog_1$size,
levels = c("Small", "Medium", "Large", "Very large"))
dog_2$size <- factor(dog_2$size,
levels = c("Small", "Medium", "Large", "Very Large"))
```

Q1.(b) Draw a bar chart showing the proportions of each size of dog.

```
size_summary <- aggregate(frequency ~ size, data = dog_2, sum)

size_summary$proportion <- size_summary$frequency / sum(size_summary$frequency)

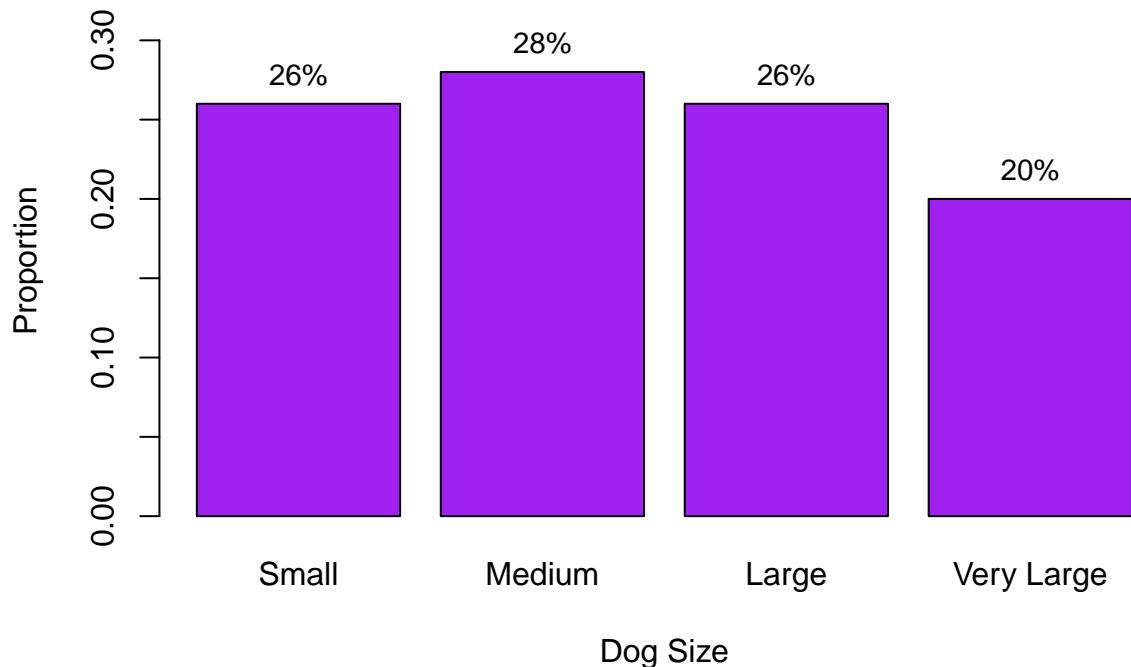
size_summary

##          size frequency proportion
## 1      Small        13     0.26
## 2    Medium        14     0.28
## 3    Large         13     0.26
## 4 Very Large       10     0.20

mp <- barplot(size_summary$proportion,
               names.arg = size_summary$size,
               col = "purple",
               ylim = c(0, max(size_summary$proportion) * 1.15),
               main = "Proportion of Dog Sizes",
               xlab = "Dog Size",
               ylab = "Proportion")

text(mp, size_summary$proportion,
      labels = paste0(round(100 * size_summary$proportion, 1), "%"),
      pos = 3, cex = 0.9)
```

Proportion of Dog Sizes



Q1.(c) Draw a stem-and-leaf diagram of the time taken to groom the 50 dogs, and use it to find the modal grooming time.

```
stem(dog_1$time, scale = 2)
```

```
##  
##      The decimal point is at the |  
##  
##      11 | 00  
##      12 | 0  
##      13 | 000000  
##      14 | 0000000000  
##      15 | 000000  
##      16 | 000000  
##      17 | 00000  
##      18 | 000000
```

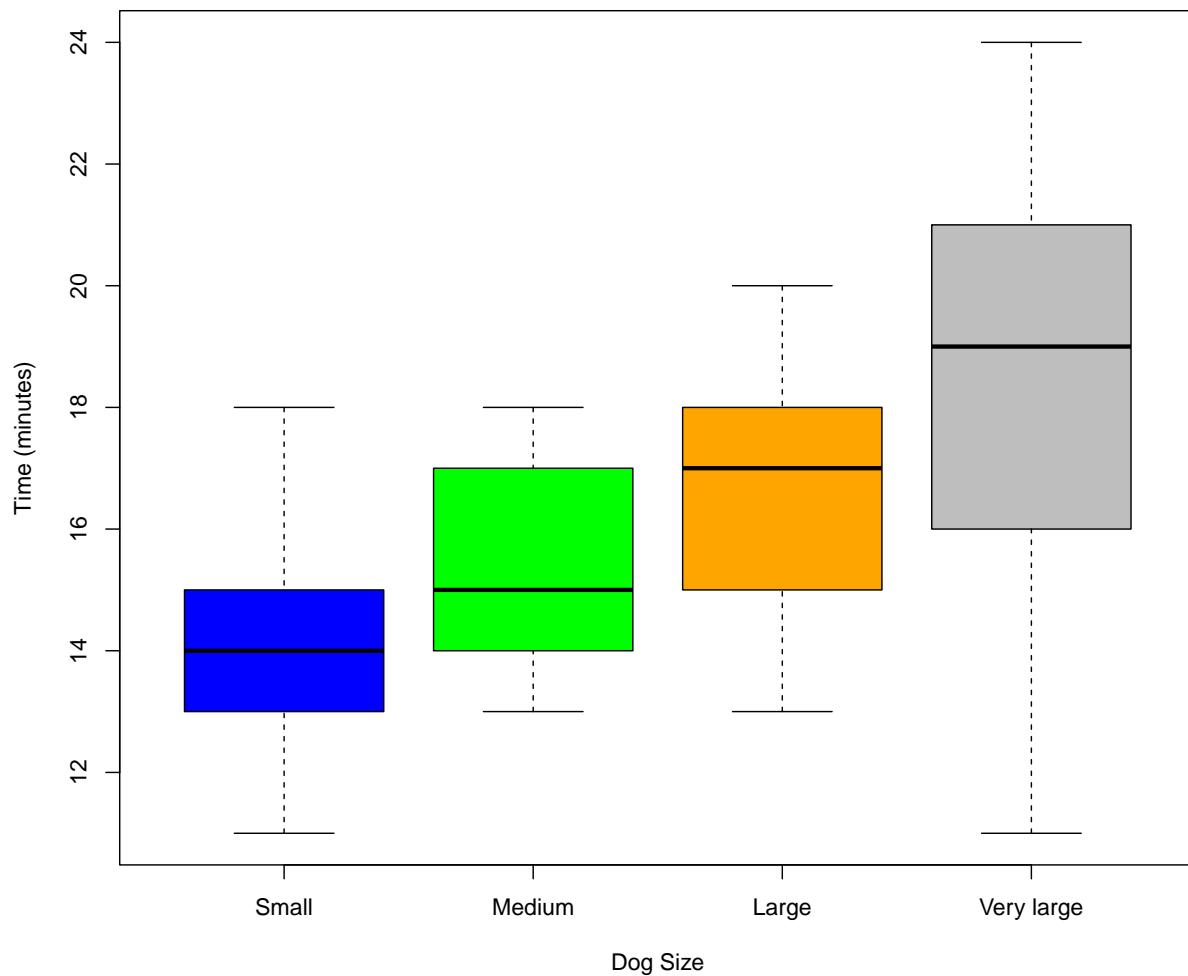
```
##   19 | 00
##   20 | 000
##   21 | 0
##   22 | 0
##   23 |
##   24 | 0
```

Therefore, The modal grooving time is 14 minutes.

Q1.(d) Draw four box plots on the same axis, each showing the grooming times for a different size of dog. Discuss the extent to which the plots show evidence for or against Andy's position and Kim's position.

```
boxplot(time ~ size, data = dog_1,
         main = "Grooming Time by Dog Size",
         xlab = "Dog Size",
         ylab = "Time (minutes)",
         col = c("blue", "green", "orange", "gray"))
```

Grooming Time by Dog Size



Through the box plots, we can observe that as the size of the dog increases, the median grooming time also increases; therefore, Andy's statement is the correct one.

Q1.(e).(i) Based on the data, what percentage of dog owners do you predict will get their money back?

```
num_exceed <- sum(dog_1$time > 20)
```

```
total_dogs <- nrow(dog_1)

refund_percent <- (num_exceed / total_dogs) * 100
paste0(round(refund_percent, 1), "%")

## [1] "6%"
```

Therefore, 6% of dog owners will get their money back.

Q1.(e).(ii) Based on the data, what value of time should the boss set if they want only 1% of dog owners to get their money back? Justify your choice of value.

```
refund_cutoff <- quantile(dog_1$time, 0.99)

refund_cutoff

##    99%
## 23.02
```

Therefore, the boss should set the time at approximately 23 minutes if they want only 1% of dog owners to get their money back

Q2.

A local garage performs MOT tests for cars. The garage is large, and performs 100 MOT tests each day. The probability of a car failing its MOT is very small, as the garage offers a pre-test service which will generally catch any problems before the test is run. Nevertheless, failures do sometimes occur. The probability of a car failing its MOT is 0.005.

Q2.(a) Let X be the binomial random variable expressing the number of cars tested by the local garage in a given day that fail their MOT.

Q2.(a).(i) Give the parameter values for X .

$$X \sim \text{Binomial}(n = 100, p = 0.005)$$

$n = 100$ is the number of independent MOT tests per day

$p = 0.005$ is the probability that a single car fails the MOT test

X is the number of cars that fail their MOT in a day

Q2.(a).(ii). List each of the assumptions required for this distribution, and justify why they apply in this case.

Assumption	Description
Fixed number of trials	The number of trials n is known and constant (100 MOT tests daily)
Two possible outcomes	Each trial has only two possible outcomes: pass or fail.
Constant probability	The probability of failure $p = 0.005$ is the same for each trial.
Independent trials	The result of one MOT test does not affect the others.

**Q2.(b) Find $P(X > 0)$, the probability that at least one car fails their MOT in a given day.
Include your working in your submission.**

Q2(b) Find $P(X > 0)$

We have a binomial random variable $X \sim \text{Bin}(n = 100, p = 0.005)$.

The probability that at least one car fails the MOT test is:

$$P(X > 0) = 1 - P(X = 0)$$

Since

$$P(X = 0) = \binom{100}{0} (0.005)^0 (0.995)^{100} = (0.995)^{100},$$

we get

$$P(X > 0) = 1 - (0.995)^{100} \approx 1 - 0.60577 = 0.39423.$$

Therefore,

$$P(X > 0) \approx 0.394 \text{ (or } 39.4\%).$$

Q2.(c) The first car that fails an MOT test on any given day results in a 10% discount for the car's owner, as an apology for not performing better pre-test checks. If more than one car fails an MOT test on a given day, the garage is required to inform the DVLA that there may be an issue with their testing procedure.

**Find the probability that on any given day, the garage will need to inform the DVLA there may be an issue with their testing procedure given they have given a 10% discount that day.
Include your working in your submission.**

We are asked to find the conditional probability that the garage must report to the DVLA, given that a 10% discount has been given.

Let $X \sim \text{Bin}(n = 100, p = 0.005)$.

The event of giving a discount corresponds to $X \geq 1$,

and the event of reporting to DVLA corresponds to $X > 1$.

Thus, we need to find:

$$P(X > 1 \mid X \geq 1) = \frac{P(X > 1)}{P(X \geq 1)}$$

We know that:

$$P(X \geq 1) = 1 - P(X = 0)$$

and

$$P(X > 1) = 1 - P(X = 0) - P(X = 1)$$

Substituting these into the conditional probability expression gives:

$$P(X > 1 \mid X \geq 1) = \frac{1 - P(X = 0) - P(X = 1)}{1 - P(X = 0)}$$

Now, using the binomial probabilities:

$$\begin{aligned} P(X = 0) &= \binom{100}{0} (0.005)^0 (0.995)^{100} = (0.995)^{100} \\ P(X = 1) &= \binom{100}{1} (0.005)^1 (0.995)^{99} = 100(0.005)(0.995)^{99} \end{aligned}$$

Hence,

$$P(X > 1 \mid X \geq 1) = \frac{1 - (0.995)^{100} - 100(0.005)(0.995)^{99}}{1 - (0.995)^{100}}$$

,

$$P(X > 1 \mid X \geq 1) = \frac{1 - 0.6058 - 0.304}{1 - 0.6058} = \frac{0.0902}{0.3942} \approx 0.229$$

Therefore,

$$P(X > 1 \mid X \geq 1) \approx 0.229$$

This means that, given a 10% discount has been issued, there is approximately a 22.9% chance that the garage will need to report to the DVLA.

Q2.(d) For sufficiently large values of n, and for p close to 0, the binomial random variable X~Bin(n,p) can be approximated by a Poisson random variable, expressed here as V~Pois(λ). X and V have the same expected value.

Give the value of λ for V, where V is approximating the random variable X defined by this question.

Given that $X \sim \text{Bin}(n = 100, p = 0.005)$, the Poisson approximation uses the same expected value:

$$\lambda = \mathbb{E}[X] = n \cdot p = 100 \times 0.005 = 0.5$$

Therefore, the approximating Poisson distribution is:

$$V \sim \text{Pois}(\lambda = 0.5)$$

Q2.(e) Calculate the value of P(V>0), and comment on this value in comparison to your answer for part b).

Given $V \sim \text{Pois}(\lambda = 0.5)$, the probability that at least one car fails the MOT is:

$$P(V > 0) = 1 - P(V = 0) = 1 - \frac{e^{-\lambda} \lambda^0}{0!} = 1 - e^{-0.5}$$

Using the exponential value:

$$P(V > 0) \approx 1 - 0.6065 = 0.3935$$

So, the approximated probability using the Poisson model is approximately:

$$P(V > 0) \approx 0.394 \text{ (or } 39.4\%)$$

This value is very close to the exact binomial calculation in part (b), which was:

$$P(X > 0) = 1 - (0.995)^{100} \approx 0.39423$$

Therefore, the Poisson approximation is highly accurate in this case, because $n = 100$ is large and $p = 0.005$ is small, satisfying the usual conditions for the Poisson approximation to the binomial distribution.

Q2.(f) X can take a maximum value of n. There is no maximum value V can take. What property or properties of X and/or V mean that this difference in maximum values does not prevent V from operating as an approximation to X?

Although the binomial variable $X \sim \text{Bin}(n = 100, p = 0.005)$ has a maximum possible value of $n = 100$, the Poisson approximation $V \sim \text{Pois}(\lambda = 0.5)$ has no upper limit.

However, the probability that X or V takes a value far from the mean is very small. Because $\lambda = np = 0.5$, both distributions are concentrated around small numbers, and the probability of getting values near or above 100 is almost zero.

Therefore, the extra values in the Poisson distribution beyond $n = 100$ do not make a real difference. Even though X has a maximum value and V does not, the Poisson model still gives a good approximation when n is large and p is small.