2023 S1 DATA1001/1901 Main Exam

Di Warren

April 20, 2023

1 DATA1001/1901 Main Exam

Spacer (Multiple Choice Section)

Instructions for Multiple Choice Section

The Multiple Choice Section is worth 50% of the total examination. There are **20** multiple choice questions, and each question is of equal value.

Answers to the Multiple Choice questions must be entered on the **Multiple Choice Answer Sheet** before the end of the examination. For each question, choose at most one option.

Question 1 (LO3 / T2 Graphical Summaries: ggplot) **Points: 1**

Consider the following data.

```
str(iris)

 'data.frame': 150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1  ...
```

What output would be produced from the following R code?

```
boxplot(iris$Sepal.Length~iris$Species)
```

(A) An error message

(B) 1 boxplot

(C) 2 boxplots

→ (D) 3 boxplots

(E) 4 boxplots

Spacer (Extended Answer Questions)

Instructions for Extended Answer Questions

The Extended Answer Section is worth 50% of the total examination. There are **4** extended answer questions, with total marks as indicated.

Answers to the Extended Answer questions must be in the spaces provided.
Give concise, precise answers, with evidence, in context.

Spacer (Exam Concept Sheet)

Exam Concept Sheet

This unit is focused on words, not formulae. The following sheet is given for your reference.

Numerical Summaries

SD = RMS of gaps from the mean = $\sqrt{\text{mean of (gaps from the mean)}^2}$
IQR = 75% percentile - 25% percentile = $Q_3 - Q_1$
Identifing outliers: $LT = Q_1 - 1.5 * IQR$; $UT = Q_3 + 1.5 * IQR$

Models

Normal: $X \sim N(mean, SD^2)$; thresholds ($\pm 1/2/3$ SD : 68%/95%/99.7%)
Linear: $\hat{y} = a + bx$, where $b = r\frac{SD_y}{SD_x}$ and $a = \bar{y} - b\bar{x}$.
Linear strip at x^*: $y^* \sim N(\bar{y} + rz_{x^*}SD_y, RMSError)$, where $RMSError = \sqrt{1-r^2}SD_y$.
Binomial: $X \sim Bin(n, p)$, then $P(X = x \text{ successes}) = \binom{n}{x}p^x(1-p)^{n-x}$, for $0 \leq x \leq n$.
Box Model: Given a population with mean M and standard deviation SD, and a sample taken with replacement of size n, the Sample Sum has EV = nM and SE = $\sqrt{n}$SD, and the Sample Mean has EV = M and SE = SD/$\sqrt{n}$.

Hypothesis Testing (HATPC)

Test	Null Hypothesis	Assumptions
1 Sample Proportion	Ho: proportion = constant	independent; constant P(success)
1 Sample T	Ho: mean = constant	independent; population Normal (if small n)
2 Sample T	Ho: difference in 2 means = constant	independent, Normal populations
Chi-squared (model)	Ho: model holds	Cochran's Rule
Chi-squared (independence)	Ho: 2 variables are independent	Cochran's Rule
Regression	Ho: slope = 0	looks linear; homoscedastic residuals

R Code

```
# IDA
str(iris)
library(tidyverse)
ggplot(iris, aes(x=Sepal.Length)) + geom_histogram()

# Modelling
pnorm(5,4,3)   # Given X ~ N(4,9), find the lower tail area from 5 down.
qnorm(0.4,4,3)  # Given X ~ N(4,9), find the 40th percentile
pnorm(r*qnorm(x)) # Estimate y percentile from x percentile, in linear model
sample(c(1:6),3,replace = T)  # 3 rolls of a fair die
```