MIE237 Term Test 1 Solutions

2016-02-09

Examination Type B; Calculator Type 2 Permitted

50 Minutes; 40 Marks Available

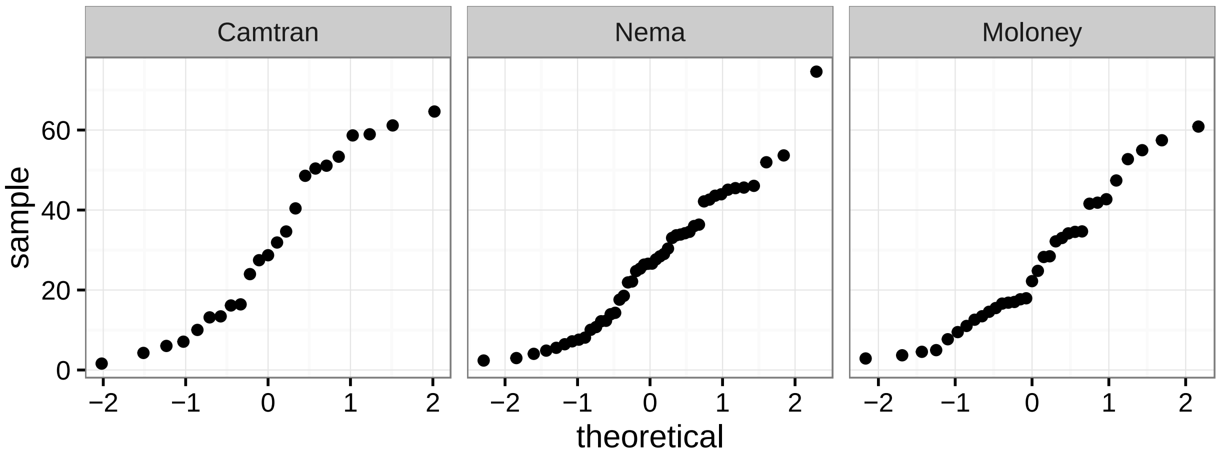
An electricity distribution company (a company that delivers electricity to homes and businesses) has accumulated a dataset related to 102 failed small transformers and wants to analyse some aspects of the data. Here are the first 10 rows of the dataset:

ID	Manufacturer	Size	Age
RY0303	Nema	100KVA	10.7
FD9446	Nema	100KVA	24.7
WZ4786	Moloney	75KVA	11.0
IW5825	Moloney	50KVA	15.5
FZ4835	Moloney	75KVA	41.6
TX9351	Nema	50KVA	36.0
JR0207	Camtran	50KVA	51.1
AB2067	Camtran	100KVA	16.4
BP3860	Moloney	75KVA	28.4
RW5898	Nema	50KVA	21.9

The dataset has 4 variables: `ID, Manufacturer, Size, Age`. The variable `ID` contains the serial number of the transformer. The variable `Manufacturer` contains the manufacturer name, one of: `Camtran, Nema, Moloney`. The variable `Size` contains a description of the transformer's power rating. The variable `Age` contains the age in years of the transformer at the time of its failure.

1. **(15 marks total)** Here is a table of summary statistics with the count, mean age, and standard deviation of age broken down by manufacturer, followed by a normal quantile plot of the ages for each manufacturer.

Manufacturer	Count	Mean Age	SD Age
Camtran	23	31.39	20.91
Nema	46	26.61	16.45
Moloney	33	26.00	16.82



Produce a 95% confidence interval for the difference in mean age at failure between `Camtran` and `Moloney` transformers, commenting on any relevant assumptions you might have needed to make.

```
##
##  Two Sample t-test
##
## data:  Age by Manufacturer
## t = 1.0661, df = 54, p-value = 0.2911
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.741064 15.509558
## sample estimates:
## mean in group Camtran mean in group Moloney
##              31.38601              26.00176
```

The pooled variance is 345.6957322.

Both normal quantile plots indicate non-normal data, but the sample sizes are large enough so that it shouldn't cause problems with the confidence interval. The standard deviations are well within the 3:1 guideline for the equal variance assumption.

2. **(10 marks total)** The company wants to look at the `Manufacturer` and `Size` variables. Here is a summary table with counts by these two variables, followed by `R` output for the χ^2 test of independence with some values removed (replaced with `MISSING`).

```
##
##  Pearson's Chi-squared test
##
## data:  tx$Size and tx$Manufacturer
## X-squared = 12.049, df = MISSING, p-value = MISSING
```

	100KVA	75KVA	50KVA	Sum
Camtran	9	3	11	23
Nema	9	17	20	46
Moloney	8	18	7	33
Sum	26	38	38	102

a. **(3 marks)** Produce a 95% confidence interval for the proportion of transformers that are manufactured by Nema, commenting on any relevant assumptions you might have needed to make.

```
##       method  x   n      mean     lower     upper
## 1 asymptotic 46 102 0.4509804 0.3544152 0.5475456
```

b. **(2 marks)** Compute the "expected cell count" for the top left cell (corresponding to `Camtran` and `100KVA`).

Here is the full table:

```
chisq.test(tx$Manufacturer, tx$Size)$expected
```

```
##                 tx$Size
## tx$Manufacturer    100KVA     75KVA     50KVA
##         Camtran  5.862745  8.568627  8.568627
##         Nema    11.725490 17.137255 17.137255
##         Moloney  8.411765 12.294118 12.294118
```

c. **(2 marks)** How many out of the 9 expected cell counts would you need to calculate using multiplication and division of marginal totals before you can simply use addition and subtraction to produce the rest?

4: the degrees of freedom $(r-1)(c-1)$.

d. **(3 marks)** Perform the test of independence with null hypothesis (informally) expressed as: H_0 : `Manufacturer` and `Size` are independent, commenting on any relevant assumptions you might have needed to make.

```
chisq.test(tx$Size, tx$Manufacturer)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tx$Size and tx$Manufacturer
## X-squared = 12.049, df = 4, p-value = 0.01699
```

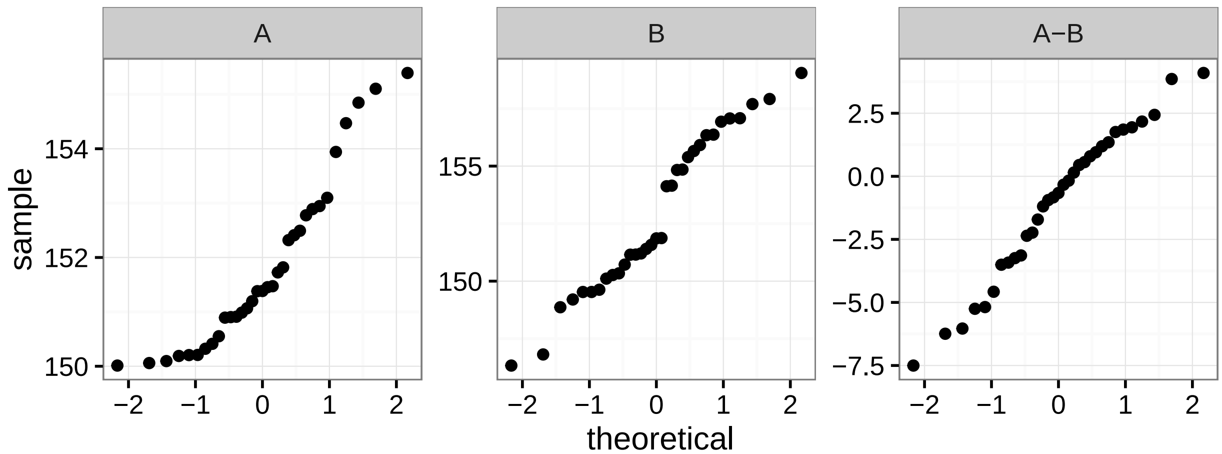
The expected cell counts all exceed 5, so the p-value is accurate.

3. **(10 marks total)** The company happens to still have all the `Moloney` transformers in storage and decides to do some electrical testing on two of the "windings" (essentially, a wire wound around a metal core—the details don't matter) in each of these transformers. Let's call the windings `A` and `B` within each unit. A current is passed through each winding and the amount of heat generated is measured. (If you are a transformer expert and this makes no sense, this is all made up, and please forgive me.)

A summer student working at the company produces the following summaries of the data gathered, consisting of: mean and standard deviation for each of the `A` and `B` winding experiments, and the standard deviation of the unit-by-unit differences between `A` and `B` experiments.

Count	A Temp Mean	A Temp SD	B Temp Mean	B Temp SD	A–B Diff Temp SD
33	151.82	1.56	152.88	1.56	3.01

Here are the normal quantile plots for the `A` and `B` winding experiments and also for the `A-B` differences.



Perform the appropriate hypothesis test to evaluate if there is a difference in temperature between `A` and `B` winding experiments.

```
t.test(tx_AB$`A-B`)
```

```
##
##  One Sample t-test
##
## data:  tx_AB$`A-B`
## t = -2.0253, df = 32, p-value = 0.05124
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.126291769  0.006077164
## sample estimates:
## mean of x
## -1.060107
```

4. **(5 marks total)** Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$. The least squares estimators for β_0 and β_1 are on the aid sheet—you'll need them here. (In this question for economy of notation I've used lowercase y to refer to "data" and "random variable" interchangeably.)

a. **(1 mark)** Show that the fitted regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ always passes through the point $(\overline{x}, \overline{y})$ for any dataset $\{(y_1, x_1), \ldots, (y_n, x_n)\}$.

Plug $\overline{x}$ into the equation to get: $\hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 \overline{x} = \overline{y}$

b. **(2 marks)** Show that $E(\overline{y}) = \beta_0 + \beta_1 \overline{x}$.

$\overline{y} = \sum(\beta_0 + \beta_1 x_i + \varepsilon_i)/n$ so

$$
\begin{aligned}
E(\overline{y}) &= E\left(\sum(\beta_0 + \beta_1 x_i + \varepsilon_i)/n\right) \\
&= \sum(\beta_0 + \beta_1 x_i + E(\varepsilon_i))/n \\
&= \sum(\beta_0 + \beta_1 x_i)/n \\
&= \beta_0 + \beta_1 \overline{x}
\end{aligned}
$$

c. **(2 marks)** Show that $E(\hat{\beta}_1) = \beta_1$.

$$
\begin{aligned}
E\left(\hat{\beta}_1\right) &= E\left(\frac{S_{xy}}{S_{xx}}\right) \\
&= E\left(\frac{\sum(y_i - \overline{y})(x_i - \overline{x})}{S_{xx}}\right) \\
&= \frac{\sum(E(y_i) - E(\overline{y}))(x_i - \overline{x})}{S_{xx}} \\
&= \frac{\sum((\beta_0 + \beta_1 x_i)) - (\beta_0 + \beta_1 \overline{x})(x_i - \overline{x})}{S_{xx}} \\
&= \beta_1 \frac{\sum(x_i - \overline{x})(x_i - \overline{x})}{S_{xx}} \\
&= \beta_1
\end{aligned}
$$