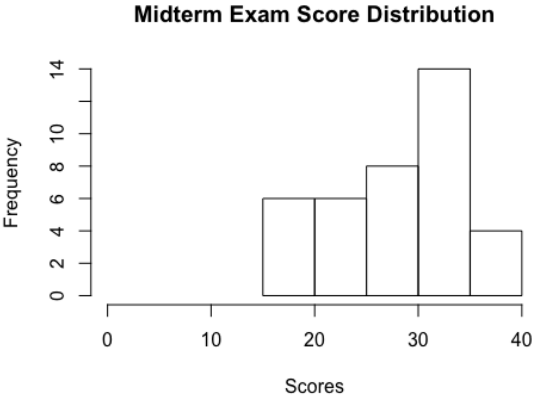
Statistics

Week 8 Recitation

ESD, SUTD

Term 5, 2017

Midterm Exam



Notes: *Regrade requests* must be made DURING this recitation.
Return the exams at the end of class.

Midterm Exam

Q2: The x and y values of each point on a Q-Q plot represent the values of the same quantile on two distributions. The quantile itself is not explicit in the plot.

Q3: The sample sizes for the high risk patients are very uneven. Simpson's paradox.

Q4: Do not exclude the outlier(63) when computing Q_1, the median $\tilde{x}$, and Q_3. (Because we need Q_1 and Q_3 to identify the outlier in the first place!)

Midterm Exam

Q5(b):

- The sum is not infinite, so cannot use the formula for an infinite geometric sum.
- An explicit formula for EMWA$_t$ is given in the Week 2 slides. Such formulas could go on the cheat sheet.
- You can show that the weights sum up to 1 using *induction*. When $t = 0$, EWMA$_0 = x_1$, so the weights sum to 1. Now suppose the weights sum to 1 for EWMA$_k$, then since EWMA$_{k+1} = \alpha\, x_{k+1} + (1 - \alpha)$ EWMA$_k$, the sum of the weights for EWMA$_{k+1}$ is $\alpha + (1 - \alpha)\times$ (sum of weights for EWMA$_k$) $= \alpha + 1 - \alpha = 1$. Then, letting $k = 0, 1, 2, \ldots$ and applying the previous argument, we see that the weights sum to 1 for all $t \in \mathbb{N}$.

Midterm Exam

Q7(a): The confidence interval start at 0, not $-\infty$, since variance and standard deviation can never be negative.

Use $\chi^2_{n-1,\alpha}$:

- Upper 99% CI and lower 99% CI.
- $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2$, σ^2 is the denominator: use α not $1-\alpha$
- One-sided: use α not $\alpha/2$.

Q8(a): Correct: $H_0 : \mu = 160,\ H_1 : \mu \neq 160.$
Incorrect: $H_0 : \mu \in [158, 162],\ H_1 : \mu \notin [158, 162]$

(c): Since the alternative hypothesis is two-sided, the p-value is also two-sided.

Midterm Exam

Q9: This is a *matched pairs* design. Since the sample size is small, the normality assumption is required.

What's wrong with the following hypothesis?

- $\mu_{before} = 125$

  $H_0 : \mu_{after} = 125,\ H_1 : \mu_{after} < 125$

Q10: Here, X_i follows a geometric distribution.

Exercise: Chi-squared Test of Independence with R

The results of a survey with 237 students in an Australian university are recoreded in a built-in data frame **survey** in R. The **Smoke** column records the students' smoking habit. The allowed values are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". The **Exer** column records their exercise level, which can be "Freq" (frequently), "Some" and "None".

Test the hypothesis whether the students' smoking habit is independent of their exercise level at 0.05 significance level.

Useful Codes

```
library(MASS)     # Load the MASS package

# Create contingency table
tbl = table(survey$Smoke, survey$Exer)

chisq.test(tbl)   # Chi-squared test
```

Exercise: Chi-squared Test of Independence with R

Questions: What's the problem with running chi-squared test directly with this contingency table? How to solve this issue?

Hint

```
# Combine categories:
ctbl = cbind(tbl[,"Freq"], tbl[,"None"]+tbl[,"Some"])
```