Statistics for Engineering, 4C3/6C3

Assignment 3

Kevin Dunn, kevin.dunn@mcmaster.ca **Due date: 05 February 2014**

Assignment objectives: univariate data analysis

Question 1 [8]

In a question on the final exam in `4M3` there was an open-ended question. The data values are the grades achieved for the answer to that question, broken down by whether the student used a systematic method, or not. No grades were given for using a systematic method; grades were awarded only on answering the question.

Use a statistical test, at the 95% confidence level to check whether this difference is significant. Interpret your answer carefully and clearly.

Solution

Note: the purpose of the assignment is not to use the `t.test(...)` function, as you won't have access to that in the midterm and final exam. See the solution at the end of this PDF.

Question 2 [6]

Your company is creating a new product line that produces a plastic. A measure of the plastic's strength, q is possible. How many lab samples must you take to be sure the true strength is in a range of 5 units, centered about the *sample* average [use typical levels of confidence]? The device that takes the measurements has an error, characterized by a standard deviation of ± 3.3 units.

Solution

The objective is to calculate n, the number of samples. Let $\overline{x}_q$ be the average of these n samples, and this average will be distributed according to the normal distribution with mean and standard deviation as shown below, if the samples are taken independently (which may not be possible in practice!):

$$z = \frac{\overline{x}_q - \mu_q}{\sigma_q}$$

The value of z will lie within this confidence interval:

$$\begin{aligned} -c_n &\leq \frac{\overline{x}_q - \mu_q}{\sigma_q/\sqrt{n}} \leq +c_n \\ \overline{x}_q - c_n\frac{\sigma_q}{\sqrt{n}} &\leq \mu_q \leq \overline{x}_q + c_n\frac{\sigma_q}{\sqrt{n}} \\ \text{LB} &\leq \mu_q \leq \text{UB} \end{aligned}$$

As a start, we will assume we know the population standard deviation, so we use the normal distribution to calculate c_n as `qnorm(1-0.05/2)` = 1.96 (using software or tables), and we use σ_q

At this point all we know is that UB $-$ LB = 5 units = $\left(\overline{x}_q + c_n\frac{\sigma_q}{\sqrt{n}}\right) - \left(\overline{x}_q - c_n\frac{\sigma_q}{\sqrt{n}}\right) = 2 \cdot c_n\frac{\sigma_q}{\sqrt{n}}$.

These are the rest of the assumptions we have to make:

- Assume the error standard deviation, as given is σ_q = 3.3 units
- Use a 95% confidence interval

Solving the above equation for n at these values gives: $n = \left(\frac{2c_n\sigma_q}{5}\right)^2 = \left(\frac{2(1.96)(3.3)}{5}\right)^2 \sim 6.7$. This implies that 7 samples should be taken.

Next, we know we should have used the t-distribution, and that s_q = 3.3 units, so let's verify whether 7 samples from that distribution works:

$$\begin{array}{rcccl} \overline{x}_q - c_t\frac{s_q}{\sqrt{n}} & \leq & \mu_q & \leq & \overline{x}_q + c_t\frac{s_q}{\sqrt{n}} \\ \text{LB} & \leq & \mu_q & \leq & \text{UB} \end{array}$$

So test using 7 samples: The c_t = `qt(0.975,df=7-1) = 2.447`. UB - LB = $2c_t\frac{s_q}{\sqrt{n}} = 2 \times 2.447 \times \frac{3.3}{\sqrt{7}} = 5.6$ units, which is too large a range.

Try using 8 samples: c_t = `qt(0.975,df=8-1) = 2.364`. UB - LB = $2c_t\frac{s_q}{\sqrt{n}} = 2 \times 2.364 \times \frac{3.3}{\sqrt{8}} = 5.5$ units, which is too large a range.

Try using 9 samples: c_t = `qt(0.975,df=9-1) = 2.306`. UB - LB = $2c_t\frac{s_q}{\sqrt{n}} = 2 \times 2.306 \times \frac{3.3}{\sqrt{9}} = 5.07$ units, which is too large a range.

Try using 10 samples: c_t = `qt(0.975,df=10-1) = 2.262`. UB - LB = $2c_t\frac{s_q}{\sqrt{n}} = 2 \times 2.262 \times \frac{3.3}{\sqrt{9}} = 4.72$ units, which meets our requirement.

This makes sense: compare the range (5 units) to the standard deviation of 3.3 units. This implies that most of the raw data comes from a range of $\pm 2 \times 3.3$ or ± 6.6 units, which is *wider* than the range. You have to take a large number of samples, and average them, to get your precision to the level required for the confidence interval.

Full grade for this question, especially for 600-level students, requires a discussion on using the t-distribution (and not just the normal distribution).

Question 3 [6]

Consider the BOD data set discussed in class, 27 January. We showed in class that there is no difference between the two methods. However, we felt uncertain about that result as it went against our expectations. Repeat the example, to discover any underlying problems with the data, and proceed to show a more careful analysis of the data.

Solution

As pointed out in class, once you detect and remove the unusual data point (outlier) which is biasing the confidence interval to span zero, the revised statistical test does not span zero any more.

Question 4 [12]

The ammonia concentration in your wastewater treatment plant is measured every 6 hours. The data for one year are available from the dataset website.

1. Use a visualization plot to hypothesize from which distribution the data might come. Which distribution do you think is most likely? Once you've decided on a distribution, use a q-q plot to test your decision.
2. Estimate location and spread statistics assuming the data are from a normal distribution. You can investigate using the `fitdistr` function in R, in the MASS package, or any other appropriate method of assessing the distribution's parameters.
3. What if you were told the measured values are not independent. How does it affect your answer?
4. What is the probability of having an ammonia concentration greater than 40 mg/L when:
    - you may use only the data (do not use *any* estimated statistics)

- you use the estimated statistics for the distribution?

Note: Answer this entire question using computer software to calculate values from the normal distribution. But also make sure you can answer the last part of the question by hand, (when given the mean and variance), and using a table of normal distributions.

Solution

```
nh4 <- read.csv('http://datasets.connectmv.com/file/ammonia.csv')
summary(nh4$Ammonia)              # just to check we've got the right data
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#    9.99   30.22   36.18   36.09   42.37   58.74

# Investigate the histogram or density plots first
hist(nh4$Ammonia)
plot(density(nh4$Ammonia))

# The qq-plot confirms it is normal, apart from the right-hand-side tail
library(car)
png(file='ammonia-qqplot.png')
qqPlot(nh4$Ammonia)
dev.off()

# Estimate the parameters of the distribution
nh4.mean = mean(nh4$Ammonia)  # 36.09499
nh4.sd = sd(nh4$Ammonia)      # 8.518928

# Advanced: use the MASS package in R to estimate
# (very similar results)
fitdistr(nh4$Ammonia, "normal")


level <- 40

# Using only the data to calculate p(Ammonia > level):
# calculate fraction of samples greater than ``level``
sum(nh4$Ammonia > level) / length(nh4$Ammonia)

# Using the normal distribution to estimate p(Ammonia > level):

# Calculate a z-value first, then the cumulative probability
z <- (level - nh4.mean)/nh4.sd
1 - pnorm(z)

# Or, you can get the answer more directly:
1 - pnorm(level, mean=nh4.mean, sd=nh4.sd)

# More correctly, we should have used the t-distribution,
# because we actually estimated the standard deviation
# We basically get the same answer
1 - pt(z, df=(length(nh4$Ammonia)-1))
```

1. When plotting a histogram, it seems that an appropriate distribution might be the normal distribution. A q-q plot shows it it mostly normal, apart from the right hand side tail (upper tail) which is slightly heavier, outside the given limits, than would be found on the normal distribution.
2. Assuming the data are normal, we can calculate the distribution's parameters as $\overline{x} = \hat{\mu} = 36.1$ and $s = \hat{\sigma} = 8.52$.
3. The fact that the data are not independent is not an issue. To calculate estimates of the parameter's distribution (the mean and standard deviation) we do not need to assume independence. One way to see this: if I randomly reorder the data, I will still get the same value for the mean and standard deviation. The assumption of

independence is required for the central limit theorem, but we have not used that theorem here.

4. The probability of having an ammonia concentration greater than 40 mg/L:

   - When counting the fraction of the samples greater than 40 mg/L (i.e. we only use the data themselves): **34.4%** (see code)
   - When using the estimated values of the mean and standard deviation from the normal distribution, we can calculate a z-value, then find the area under the normal distribution corresponding to this z: **32.3%** (see code)

     Note: We should use actually be using the t-distribution, since we used *an estimate* of the population variance and not the true population variance to calculate z. However, since the degrees of freedom, $n - 1 = 1439$, are so large, there is no practical difference in our answer.

Question 5 [5 (600-level students; extra credit for 400-levels students)]

The confidence interval for the population mean takes one of two forms below, depending on whether we know the variance or not. At the 90% confidence level, for a sample size of 13, compare and comment on the upper and lower bounds for the two cases. Assume that $s = \sigma = 3.72$.

$$-c_n \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq c_n$$

$$-c_t \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq c_t$$

Solution

This question aims for you to prove to yourself that the t-distribution is **wider (more broad)** than the normal distribution. The 90% region spanned by the t-distribution with 12 degrees of freedom has upper and lower limits at `qt((1-0.9)/2, df=12)`, i.e. from **-1.782** to **1.782**. The equivalent 90% region spanned by the normal distribution is `qnorm((1-0.9)/2)`, spanning from **z=-1.64** to **z=1.64**. Everything else in the center of the 2 inequalities is the same, so we only need to compare c_t and c_n.

Question 6 [0 (for practice)]

From the 2011 midterm

Sulphur dioxide is a byproduct from ore smelting, coal-fired power stations, and other sources.

These 11 samples of sulphur dioxide, SO_2, measured in parts per billion [ppb], were taken from our plant. Environmental regulations require us to report the 90% confidence interval for the mean SO_2 value.

$$180,\ 340,\ 220,\ 410,\ 101,\ 89,\ 210,\ 99,\ 128,\ 113,\ 111$$

1. What is the confidence interval that must be reported, given that the sample average of these 11 points is 181.9 ppb and the sample standard deviation is 106.8 ppb?
2. Why might Environment Canada require you to report the confidence interval instead of the mean?

Solution

1. From the central limit theorem, assuming the 11 values are independent, the mean SO_2 value, $\overline{x} \sim \mathcal{N}\left\{\mu, \sigma^2/n\right\}$, where μ and σ are the distribution from which the raw values come.

   Using an estimate for $\sigma = \hat{s} = 106.8$ we can construct the z-value and confidence interval. z will be t-distributed with $n - 1 = 10$ degrees of freedom, so $c_t = 1.81$ At the 90% confidence level we can then write:

$$\begin{aligned} -c_t &\leq \frac{\overline{x} - \mu}{s/\sqrt{n}} \leq +c_t \\ \overline{x} - c_t \frac{s}{\sqrt{n}} &\leq \mu \leq \overline{x} + c_t \frac{s}{\sqrt{n}} \\ 181.9 - 1.81 \times \frac{106.8}{\sqrt{11}} &\leq \mu \leq 181.9 + 1.81 \times \frac{106.8}{\sqrt{11}} \\ 123.6 \text{ ppb} &\leq \mu \leq 240.2 \text{ ppb} \end{aligned}$$

2. Environment Canada may require the confidence interval since in addition to providing an estimate of the mean (just the midpoint of the CI), it also provides an *estimate of the spread* – variability in your process – if n is known, without requiring access to the raw data.

   A wide CI gives an indication that you might in fact be polluting too much on some days, and compensating on others, which is not desirable. The confidence interval's width can also be compared between plants to find the most variable polluters.

Question 7 [0 (for practice)]

From the 2011 midterm

A concrete slump test is used to test for the fluidity, or workability, of concrete. It's a crude, but quick test often used to measure the effect of polymer additives that are mixed with the concrete to improve workability.

The concrete mixture is prepared with a polymer additive. The mixture is placed in a mold and filled to the top. The mold is inverted and removed. The height of the mold minus the height of the remaining concrete pile is called the "slump".

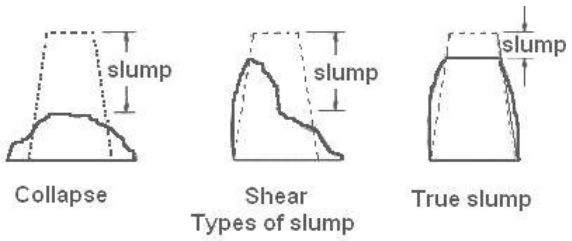


Illustration from Wikipedia

Your company provides the polymer additive, and you are developing an improved polymer formulation, call it B, that hopefully provides the same slump values as your existing polymer, call it A. Formulation B costs less money than A, but you don't want to upset, or loose, customers by varying the slump value too much.

1. You have a single day to run your tests (experiments). Preparation, mixing times, measurement and clean up take 1 hour, only allowing you to run 10 experiments. Describe all precautions, and why you take these precautions, when planning and executing your experiment. Be very specific in your answer (use bullet points).

2. The following slump values were recorded over the course of the day:

Additive	Slump value [cm]
A	5.2
A	3.3
B	5.8
A	4.6
B	6.3
A	5.8
A	4.1
B	6.0
B	5.5
B	4.5

What is your conclusion on the performance of the new polymer formulation (system B)? Your conclusion must either be "send the polymer engineers back to the lab" or "let's start making formulation B for our customers". Explain your choice clearly.

To help you, $\overline{x}_A = 4.6$ and $s_A = 0.97$. For system B: $\overline{x}_B = 5.62$ and $s_B = 0.69$.

Note: In your answer you must be clear on which assumptions you are using and, where necessary, why you need to make those assumptions.

3. Describe the circumstances under which you would rather use a paired test for differences between polymer A and B.

4. What are the advantage(s) of the paired test over the unpaired test?

Solution

1. The basic rule is to control what you can and randomize against what you cannot. You should have mentioned some of these items:

   - Control: clean equipment thoroughly between runs.
   - Control: other factors that might affect the slump: temperature, humidity.
   - Control: ensure the same person prepares all mixtures, or randomize the allocation of people if you have to use more than 1 person. Don't let person 1 prepare all the A mixtures and person 2 the B mixtures.
   - Control: mixing times and how the mixture is created could have an effect. This should ideally be done by the same person.
   - Randomize the order of all the A and B experiments: don't run all the A's, then all the B's, as that will confound with other factors. For example, even though temperature might vary during the day, if we randomize the run order, then we prevent temperature from affecting the results.
   - Use raw materials (cement, binder, other ingredients) from all possible suppliers. And the supplier raw materials should be representative.

2. We will initially assume that $\mu_A = \mu_B$, in other words, the outcome is "let's start making formulation B for our customers". We will construct a confidence interval for the difference, $\mu_B - \mu_A$ and interpret that CI.

   - Assume the slump values within each group are independent, which will be true if we take the precautions above. We do this because then we can use the central limit theorem (CLT) to state $\overline{x}_A \sim \mathcal{N}\left(\mu_A, \sigma_A^2/n_A\right)$ and that $\overline{x}_B \sim \mathcal{N}\left(\mu_B, \sigma_B^2/n_B\right)$.
   - Note: we don't require the samples within each group to be normally distributed.
   - Assume the variances are the same: $\sigma_A^2 = \sigma_B^2 = \sigma^2$: this is required to simplify the next step.
   - Assume the $\overline{x}_A$ and $\overline{x}_B$ means are independent. This allows us to calculate a variance value, $\mathcal{V}\{\overline{x}_B - \overline{x}_A\}$ from which we can create a z-value for $\mu_B - \mu_A$:

$$z = \frac{(\overline{x}_B - \overline{x}_A) - (\mu_B - \mu_A)}{\sqrt{\mathcal{V}\{\overline{x}_B - \overline{x}_A\}}}$$

That denominator variance can be written as:

$$\begin{aligned}\mathcal{V}\left\{\overline{x}_B - \overline{x}_A\right\} &= \mathcal{V}\left\{\overline{x}_B\right\} + \mathcal{V}\left\{\overline{x}_A\right\} \\ &= \sigma^2\left(\frac{1}{n_B} + \frac{1}{n_A}\right)\end{aligned}$$

using our previous assumption that the variances are equal. We can verify this with an F-test, but won't do it here.

Because we do not have an external estimate of the variance, σ^2, available, we must assume a good estimate for it can be found by pooling the estimated variances of the group A and B samples (which requires our equal variance assumption from earlier).

$$\begin{aligned}s_P^2 &= \frac{4s_A^2 + 4s_B^2}{4+4} \\ s_P^2 &= \frac{4(0.97)^2 + 4(0.69)^2}{4+4} = 0.709\end{aligned}$$

This pooling also gives us 8 degrees of freedom for the t-distribution, which is how the z-value is distributed.

Using that z-value and filling our assumed difference of zero for the true means, we can construct a 95% confidence interval:

$$\begin{aligned} -c_t &\leq z \leq +c_t \\ (\overline{x}_B - \overline{x}_A) - c_t\sqrt{s_P^2\left(\tfrac{1}{n_B} + \tfrac{1}{n_A}\right)} &\leq \mu_B - \mu_A \leq (\overline{x}_B - \overline{x}_A) + c_t\sqrt{s_P^2\left(\tfrac{1}{n_B} + \tfrac{1}{n_A}\right)} \\ 1.02 - 2.3\sqrt{0.709\left(\tfrac{1}{5} + \tfrac{1}{5}\right)} &\leq \mu_B - \mu_A \leq 1.02 + 2.3\sqrt{0.709\left(\tfrac{1}{5} + \tfrac{1}{5}\right)} \\ -0.21 &\leq \mu_B - \mu_A \leq 2.2 \end{aligned}$$

The statistical conclusion is that there is **no difference between formulation A and B**, since the CI spans zero. However, the practical interpretation is that the CI only just contains zero, and this should cause us to stop, and really consider the risk of the statistical conclusion.

If one of the data points were in error just slightly, or if we ran a single additional experiment, it is quite possible the CI will *not span zero* anymore. In my mind, this risk is too great, and we risk upsetting the customers.

So my conclusion would be to "send the polymer engineers back to the lab" and have them improve their formulation until that CI spans zero more symmetrically.

3. A paired test should be used when there is something is common *within* pairs of samples in group A and B, but that commonality does not extend between the pairs. Some examples though you could have mentioned:

   Pairing is appropriate: person 1 mixes polymer for test A and B; person 2 mixes polymer for test A and B (but with different time and agitation level that person 2); person 3 mixes ... *etc* Pairing *not* appropriate: person 1 mixes all the polymer A samples; person 2 mixes all the polymer B samples (pairing won't fix this, and even the unpaired results will be inaccurate - see precautions mentioned above). Pairing appropriate: you only have enough cement and raw materials to create the concrete mixture for 2 samples: one for A and one for B. You repeat this 5 times, each time using a different supplier's raw materials.

   In other words, pairing is appropriate when there is something the prevents the $\overline{x}_A$ and $\overline{x}_B$ quantities from being independent.

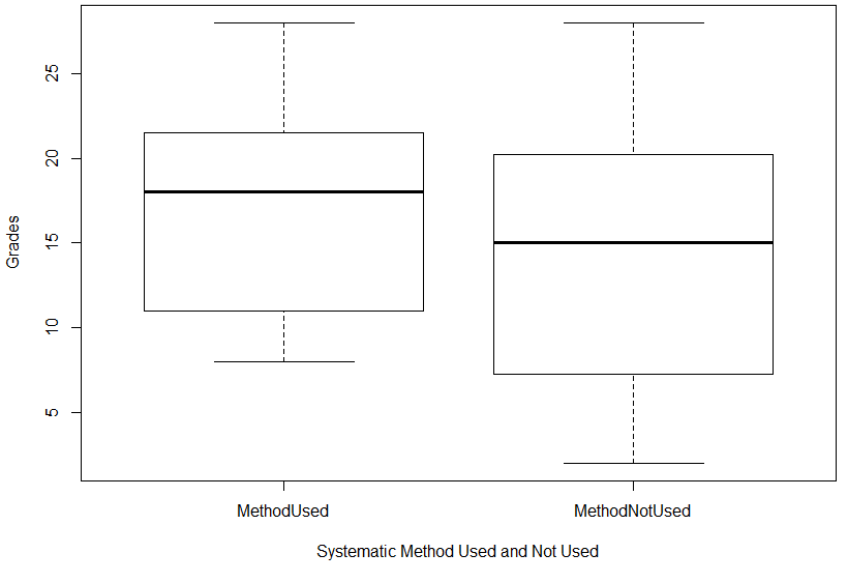
4. The one advantage of the paired test is that it will cancel out any effect that is common between the pairs (whether that effect actually affects the slump value or not). Pairing is a way to guard against *potential effect*.

   This makes the test more sensitive to the difference actually being tested for (formulation A vs B) and prevents confounding from the effect we are not testing for (suppliers' raw material).

   Unpaired tests, but with randomization will only prevent us from being misled, however that supplier effect is still present in the 10 experimental values. The 5 difference values used in the paired tests will be free from that effect.

Solution by *Ghassan Marjaba*

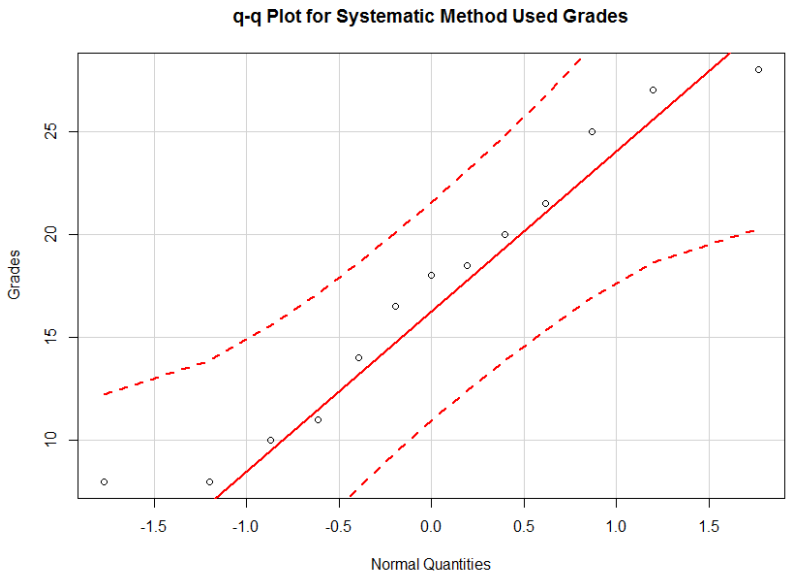
The first step would be to visualize the data. Below I plotted a box plot and a scatter plot. This would help me possibly identify outliers or any trends that may be useful. This step is not strictly required from a statistical point of view, that is we can just calculate the confidence interval and interpret the results, however plotting the data when possible is useful.



The box plot, which we have seen in a previous assignment, shows that there is significant overlap. This means that a statistical test would be needed to make judgement on whether there is a difference or not.

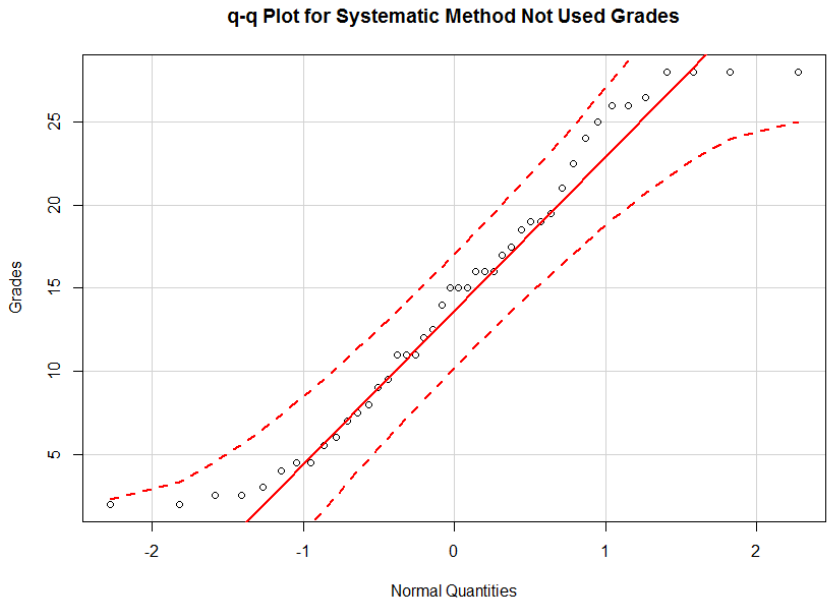
The histograms (not shown) are interesting. At first glance we may think that the grades are not normally distributed, which we know they should be. If they were not, we have to look closely at the reason. So to check, I used the q-q-plots shown below.

From these q-q plots, we can confirm that the data is normally distributed and proceed.



There are many advantages to using paired tests, however the experiment, or in this case the data collection should have been set up as a paired test. This is not likely to be possible for this situation since we cannot have the same student solve the problem systematically and then non-systematically with the results being independent.

Therefore, we will use the unpaired test.



Note that I will use the mean and standard deviation (instead of robust statistics) since there appears to be no outliers.

- Let A be systematic method not used $\Rightarrow \overline{x_A} = 14.44$ and $s_A = 8.34$
- Let B be systematic method used $\Rightarrow \overline{x_B} = 17.35$ and $s_B = 6.905$

Note that the median for A is 15 and for B is 18. They are relatively very close to the means, which gives us some level of comfort that the data is not being skewed by any outliers.

Note: this part to check for pooling the variances is not required for 400-level students (or for 600-level students, but you should understand what is being checked here). To calculate the confidence interval, we need to pool the variances, but we can only do so if they are similar enough. To check for that, we need to calculate the confidence interval for the ratio of the two variances. Assuming the degrees of freedom used for estimating s_A is 13 and for estimating s_B is 44, we can use R to calculate the confidence interval for s_A^2 / s_B^2. The point along the cumulative F-distribution which has an area of 0.025 (95% confidence), is:

$$LB = qf(0.025,\ 44,\ 13) \times s_A^2 / s_B^2 = 0.451 \times \left(\frac{8.3^2}{6.9^2}\right) = 0.653$$

$$UB = qf(0.975,\ 44,\ 13) \times s_A^2 / s_B^2 = 2.764 \times \left(\frac{8.3^2}{6.9^2}\right) = 4.00$$

The 95% confidence interval does span 1 which means that we can pool the variance.

Calculating the pooled variance:

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{(13-1)(8.3)^2 + (44-1)(6.9)^2}{13+44-2} = 64.85 \Rightarrow s_p = 8.05$$

Calculating the confidence interval for $\mu_B - \mu_A$

$$LB = (\overline{x_B} - \overline{x_A}) - c_t \sqrt{s_p^2 \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

$$UB = (\overline{x_B} - \overline{x_A}) + c_t\sqrt{s_p^2(\frac{1}{n_A}+\frac{1}{n_B})}$$

where $c_t = qt(0.025,\ df = 55) = -2.004$

therefore

$$LB = (\overline{x_B} - \overline{x_A}) - c_t\sqrt{s_p^2(\frac{1}{n_A}+\frac{1}{n_B})} = -2.19$$
$$UB = (\overline{x_B} - \overline{x_A}) + c_t\sqrt{s_p^2(\frac{1}{n_A}+\frac{1}{n_B})} = 8.0$$

The confidence interval does span zero, which means that there NO statistical difference in the grades of using the systematic method versus not using the systematic method. Having said that, the lower bound is close to zero, i.e. it does not symmetrically span zero. I would still be inclined to say that there is a difference, even if very slight. Consider the smaller sample size of students using the systematic method (13 vs the 44 from the other group).

600-level students: consider writing a code where you randomly select N samples from the one group and N samples from the other group. Construct a confidence interval and see whether it spans zero. Then repeat the code, each time selecting a different set of N samples.

```
grades <-
read.csv('http://datasets.connectmv.com/file/systematic-method.csv')
boxplot(grades, xlab="Systematic Method Used and Not Used",
ylab="Grades")

hist(grades$MethodUsed, xlab="Systematic Method Used",
main="Histogram of Grades")
hist(grades$MethodNotUsed, xlab="Systematic Method Not Used",
main="Histogram of Grades")

library(car)
qqPlot(grades$MethodUsed, xlab="Normal Quantities", ylab="Grades",
main="q-q Plot for Systematic Method Used Grades")
qqPlot(grades$MethodNotUsed, xlab="Normal Quantities", ylab="Grades",
main="q-q Plot for Systematic Method Not Used Grades")

xbarA = mean(grades$MethodNotUsed, na.rm = TRUE)
xbarB = mean(grades$MethodUsed, na.rm = TRUE)
sigmaA = sd(grades$MethodNotUsed, na.rm = TRUE)
sigmaB = sd(grades$MethodUsed, na.rm = TRUE)
```