Statistics for Engineering, 4C3/6C3

Assignment 3

Kevin Dunn, kevin.dunn@mcmaster.ca **Due date: 08 February 2013**

Note: Assignment objectives

- Demonstrate an understanding of tests for differences, both paired and unpaired.
- Correctly calculate and interpret confidence intervals.
- Find limits for monitoring charts, and understand their purpose.

Question 1 [12]

Similar, but different from the 2012 final exam.

A company has been producing a polymer for the past 5 years. There are plenty of historical data available, two values per day, that give the conversion of the raw material monomer to the final product, the polymer. An engineering team has just finished a sequence of experiments to test whether a cheaper catalyst, **B**, has any effect on product conversion when compared to the existing catalyst, **A**.

The engineers created a set of $N = 8$ experiments over the 4 days, where they tested the catalyst in an alternating manner: these are experiments 13 to 20 in the table below, together with the resulting conversion.

Experiment	Catalyst used	Conversion [%]	Experiment	Catalyst used	Conversion [%]
1	**A**	85	11	**A**	84
2	**A**	78	12	**A**	80
3	**A**	81	13	**B**	76
4	**A**	79	14	**A**	79
5	**A**	97	15	**B**	71
6	**A**	70	16	**A**	76
7	**A**	87	17	**B**	86
8	**A**	74	18	**A**	75
9	**A**	89	19	**B**	92
10	**A**	77	20	**A**	83

1. Describe why the 8 experiments, numbered 13 to 20, could show a misleading result when trying to test the difference between catalysts **A** and **B** using only those 8 data points.
2. Draw a table to show how *you* would have allocated the choice of catalyst **A** or **B** for runs 13 to 20 over those 4 days.
3. Calculate the confidence interval that engineering team would have calculated from only those 8 data points.
4. Given that this experimental work has already been completed, show the first steps of the calculations you could do to extract some additional value from these data. Use the other data in the table to demonstrate your method.

Solution

1. The 8 experiments are performed 2 per day (A and B run on each day). The experimental values will not be independent, as there might be something in the process that changes from morning to evening (e.g. ambient temperature; or shift/crew operating the process). This will lead to a misleading result. The experiments should have been carried out in random order.
2. One way to generate A/B random results in R, is to write `runif(11) > 0.5` and you then get: `FALSE TRUE TRUE TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE`. From this I would assign the experiments as `B, A, A, A, B, A, B, B`. Any other random allocation would work.

In certain circumstances it might be appropriate to pair the experiments on each day, if one can guarantee that all other conditions are kept constant. In this case, one would randomly pick A or B to be first, then run the other one afterwards.

3. From the data, and assuming the samples of A and B come from the normal distribution, and are independent of each other, and that the variances can be pooled:

$$\bar{x}_A = 78.25$$
$$\bar{x}_B = 81.25$$
$$s_A = 3.594$$
$$s_B = 9.5$$
$$s_P^2 = \frac{(4-1)(3.594)^2 + (4-1)(9.5)^2}{4-1+4-1} = 51.58$$
$$c_t = 2.447 \text{ (the 95\% confidence interval's critical value with 6 DOF)}$$
$$\text{LB} = (81.25 - 78.25) - 2.447 \cdot \sqrt{51.58\left(\frac{1}{4} + \frac{1}{4}\right)} = -9.43$$
$$\text{UB} = (81.25 - 78.25) + 2.447 \cdot \sqrt{51.58\left(\frac{1}{4} + \frac{1}{4}\right)} = 15.43$$

indicating there is no apparent difference between the two catalysts from these limited, and problematic data.

4. One thing to consider is building a dot plot for the differences from first to second run in the day, and check the difference of the 8 runs against historical data.

   For example, from historical data we can calculate:

   - diff1 = 85 - 78 = 7
   - diff2 = 81 - 79 = 2
   - diff3 = 97 - 70 = 27
   - diff4 = 87 - 74 = 13
   - diff5 = 89 - 77 = 12
   - diff6 = 84 - 80 = 4

   Notice these differences are all positive, indicating that even without changing catalyst there is a definite tendency to obtain higher conversion in the first run of the day.

   Now contrast these differences with the average difference from the actual tests:

   - actual diff 1 = 76 - 79 = -3
   - actual diff 2 = 71 - 76 = -5
   - actual diff 3 = 86 - 75 = +9
   - actual diff 4 = 92 - 83 = +9

   Here the average difference is (-3 -5 +9 +9)/4 = 2.5 percentage units.

   I would use the fact that the other runs show a high difference (on average 10.8 percentage units), with no change in catalyst. This actually leads one to conclude the new catalyst has *lowered* conversion. Certainly it would be better to obtain a longer sequence of data prior to the experiment, to verify the average within-day difference.

Question 2 [16]

A major aim of many engineers is/will be to reduce the carbon footprint of their company's high-profile products. Next week your boss wants you to evaluate a new raw material that produces 2.6 $\frac{\text{kg } CO_2}{\text{kg product}}$ less than the current material,

but the final product's brittleness must be the same as achieved with the current raw material. This is a large reduction in CO_2, given your current production capacity of 51,700 kg of product per year. Manpower and physical constraints prevent you from running a randomized test; you don't have a suitable database of historical data either.

One idea you come up with is to use to your advantage the fact that your production line has three parallel reactors, TK104, TK105, and TK107. They were installed at the same time, they have the same geometry, the same instrumentation, *etc*; you have pretty much thought about every factor that might vary between them, and are confident the 3 reactors are identical. Typical production schedules split the raw material between the 3 reactors. Data on the website contain the brittleness values from the three reactors for the past few runs on the current raw material.

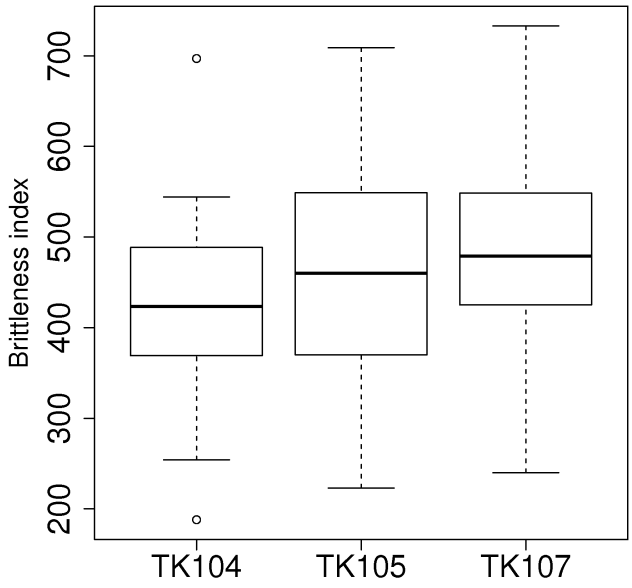
1. Which two reactors would you pick to run your comparative trial on next week?
2. 600-level students: see the question below, where you use pairing to answer the same question.

Solution

The purpose of this question is to compare two systems. There are two ways: either compare one group to another group, or to have paired tests. We could consider this a paired test, because the material is run in both reactors at the same conditions. In this answer we compare reactor I to reactor J as groups. Our answer will be to run experiments in the reactors that show the smallest difference.

Note: This question also has missing data, denote as `NA` in R. Most real data sets that you deal with will have missing data and the questions will expect to deal with them. For example, the degrees of freedom will be reduced because of the missing data. Use this solution to see how to write code in R that deals with missing values.

We can start by looking at the data. A box plot is a reasonable way to compare both the location and spread of the brittleness values from each reactor.



The standard way to test for differences between two groups of samples is given by the equation below - it is derived as coming from the normal distribution with mean of $\mu_A - \mu_B$ and the standard deviation as shown in the denominator.

$$z = \frac{(\overline{x}_B - \overline{x}_A) - (\mu_B - \mu_A)}{\sqrt{\sigma^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}}$$

Assuming the two *population* means are identical, the z-value is a direct estimate of the probability with which that assumption is wrong. A z-value around zero indicates that the assumption was true, a large or small z-value indicates

that the assumption was wrong.

So we can calculate the z-value, and the corresponding probability for each pair of reactor differences using the code below.

But the next problem we face is that we don't know the value of σ. We can estimate it however, by pooling the variances of the two groups. Strictly speaking we should do a check for comparable variances before pooling them, but you may assume that we can.

When we use the pooled variance now, then the assumption that the z-value follows the normal distribution is not correct anymore; it follows the t-distribution, with the pooled number of degrees of freedom. Once we have the z-value we can calculate the probability of finding a z-value of at least that big. Anything beyond that is the risk that we are wrong.

We can also expand the z value into a confidence interval at a given confidence level. We do this in the code at the 95% level (see `LB` and `UB` terms).

- $\mu_{104} - \mu_{105}$: z = 1.25; risk we are wrong: 89.1%; CI: $-31.4 \leq \mu_{104} - \mu_{105} \leq 134$
- $\mu_{104} - \mu_{107}$: z = 1.41; risk we are wrong: 91.6%; CI $-21.4 \leq \mu_{104} - \mu_{107} \leq 120$
- $\mu_{105} - \mu_{107}$: z = -0.0532; risk we are wrong: 52.1% and $-81.8 \leq \mu_{105} - \mu_{107} \leq 77.6$ (note that the minimum risk is 50%; the risk is not 47.8%)

While all three reactors have confidence intervals that span zero at the 95% level, notice how the interval gives us a feel for the degree of difference. Clearly **reactors TK105 and TK107 are the most similar**, however all 3 are statistically equivalent from a confidence interval point of view. Contrast this to using a hypothesis test, which you may have encountered in other statistical courses. A hypothesis test just tells you "yes" or "no"; a confidence interval gives a much better engineering feel for the degree of difference.

A full solution to this question requires you report the z-values and its corresponding risk.

```
# We are going to be doing this 3 times. Rather write a function to
# do the work for general "groupA" and "groupB" vectors.
group_difference <- function(groupA, groupB)
{
    # This function assumes either group has missing data.  Calculate
    # the mean and variance omitting the missing values

    A.mean <- mean(groupA[!is.na(groupA)])
    A.var <- var(groupA[!is.na(groupA)])
    A.N <- length(groupA[!is.na(groupA)])

    B.mean <- mean(groupB[!is.na(groupB)])
    B.var <- var(groupB[!is.na(groupB)])
    B.N <- length(groupB[!is.na(groupB)])

    difference <- B.mean - A.mean
    var.DOF <- (A.N - 1 + B.N - 1)
    var.pooled <- ((A.N - 1) * A.var + (B.N - 1) * B.var) / var.DOF

    sd.denom <- sqrt(var.pooled *(1/A.N + 1/B.N))
    z <- (difference - 0) / sd.denom
    t.critical <- pt(z, var.DOF)

    LB <- difference - qt(0.975, df=var.DOF)*sd.denom
    UB <- difference + qt(0.975, df=var.DOF)*sd.denom
return(list(z, t.critical, LB, UB))
}

brittle <- read.csv('http://datasets.connectmv.com/file/brittleness-index.csv')
attach(brittle)  # Now we can access the variables directly, without $ symbols
```

```
# Let's start though by plotting boxplots of the data
bitmap('../images/brittleness-boxplot.png', type="png256", width=7, height=7, res=250, pointsize=14)
par(mar=c(4.2, 4.2, 0.2, 0.2))  # (bottom, left, top, right);
boxplot(brittle, ylab="Brittleness index", cex.lab=1.5, cex.main=1.8, cex.sub=1.8, cex.axis=1.8)
dev.off()

# 104 vs 105
group_difference(TK104, TK105)
# z = 1.253729
# t.critical = 0.891298 (1-0.1087021)
# -31.4 < mu.diff < 134

# 104 vs 107
group_difference(TK104, TK107)
# z = 1.405639
# t.critical = 0.9163178 (1-0.0836822)
# -21.4 < mu.diff < 120

# 105 vs 107
group_difference(TK105, TK107)
# z = -0.05326222
# t.critical = 0.4788878 (1-0.5211122)
# -81.8 < mu.diff < 77.6
```

Question 3 [600 level students: 8]

Repeat the above question, but assume samples of raw material were split in thirds and each third was run in one of the reactors.

Use a paired test and calculate the confidence interval for the reactor combinations to answer the question: which two reactors would you pick to run your comparative trial on next week?

Solution

Pairing assumes that each reactor was run with the same material, except that the material was split into thirds: one third for each reactor. As described in the section on paired tests we rely on calculating the difference in brittleness, then calculating the z-value of the average difference. Contrast this to the unpaired tests, where we calculated the difference of the averages.

The code below shows how the paired differences are evaluated for each of the 3 combinations. The paired test highlights the similarity between TK105 and TK107, the same as the unpaired test. However the paired test shows much more clearly how different tanks TK104 and TK105 are, and especially TK104 and TK107.

In the case of TK104 and TK105 the difference might seem surprising - take a look back at the box plots and how much they overlap. However a paired test cannot be judged by a box plot, because it looks at the case-by-case difference, not the overall between group difference. A better plot with which to confirm the really large z-value for the TK105 and TK107 difference is the plot of the differences.

```
brittle <- read.csv('http://datasets.connectmv.com/file/brittleness-index.csv')
attach(brittle)

# Calculates the paired difference
paired_difference <- function(groupA, groupB, alpha=0.95)
{
    # This function assumes either group has missing data.
    # Find the subset of observations in common.

    groupA.sub <- groupA[!is.na(groupA) & !is.na(groupB)]
    groupB.sub <- groupB[!is.na(groupA) & !is.na(groupB)]
```

```
    diffs <- groupB.sub - groupA.sub
    diffs.mean <- mean(diffs)
    diffs.sd <- sd(diffs)
    diffs.N <- length(diffs)

    plot(groupB.sub-groupA.sub, type="b")

    z <- (diffs.mean - 0) / (diffs.sd/sqrt(diffs.N))
    t.critical <- pt(z, df=(diffs.N-1))
    c.t <- qt(1-(1-alpha)/2, df=(diffs.N-1))
    LB <- diffs.mean  - c.t * diffs.sd / sqrt(diffs.N)
    UB <- diffs.mean  + c.t * diffs.sd / sqrt(diffs.N)

return(list(z, t.critical, diffs.N-1, LB, UB))
}

paired_difference(TK104, TK105, alpha=0.95)
# (z=2.64, t.critical=0.991, DOF=17, LB=9.81, UB=88.4)

paired_difference(TK104, TK107, alpha=0.95)
# (z=12, t.critical=1, DOF=19, LB=48.3, UB=68.7)

paired_difference(TK105, TK107, alpha=0.95)
# (z=-0.33, t.critical=0.37, DOF=20, LB=-46.1, UB=33.5)
```

The confidence intervals are:

$$\begin{aligned} 9.81 &\leq \mu_{105-104} \leq 88.4 \\ 48.3 &\leq \mu_{107-104} \leq 68.7 \\ -46.1 &\leq \mu_{107-105} \leq 33.5 \end{aligned}$$

You should also consider how the reduction in degrees of freedom affects this test; contrast the results to those when using an unpaired test.

Question 4 [3]

Final exam, 2012: Describe why *and* how real-time control charts* can save companies money in processes where there is a long time delay from the point of production until getting the lab results back from quality control testing.

* We've been calling them monitoring charts, but "control charts" is also a term that is used.

Solution

The earlier the problem is detected, the less off-spec product will have to be reworked, or in the worst case, scraped. Also, the sooner the problem is detected, the more likely the operators and engineers on the line can solve it correctly. If left for later, the true cause of the problem might never be correctly uncovered.

If there is monitoring at frequent points, we are able to pin point where the problem occurred.

The real-time data should be plotted as soon as possible on the control chart so that the off-spec production can be detected. Don't wait for the much slower off-line lab results; plot the real-time data. The real-time data is more likely to contain an accurate fingerprint of the process problem than the laboratory data.

Question 5 [9 = 2 + 2 + 1 + 1 + 3]

Final exam, 2012: A filling line in our company fills bottles by weight. The main quality criterion is to ensure the fill weight is stable over time, around the target of 300 mg. Using data from many days of operation, the standard deviation of fill weights was determined to be 21.7 mg and it remains constant, since the equipment has weekly maintenance.

1. Calculate the Shewhart chart control limits that could be used to monitor the fill weights.
2. A basic Shewhart chart is really not the best choice to monitor the fill weights. Explain why not, and what can be used instead.
3. Calculate the lower and upper control limit for an EWMA chart using $\lambda = 0.1$.
4. Calculate the lower and upper control limit for an EWMA chart using $\lambda = 0.9$.
5. Comment on the Shewhart and two EWMA limits you have calculated in this question, and explain why their numeric values make sense in the context of this monitoring problem.

Solution

1. A Shewhart chart limits could be set at UCL $= 300 + 3 \times 21.7 = 365.1$ mg and LCL $= 234.9$ mg assuming we are monitoring the raw data directly. If we monitor with subgroups, then divide the standard deviation by the square root of the subgroup size, ie $\pm 3 \cdot \frac{\sigma}{\sqrt{n}}$. There is no need to use the a_n correction factor if we know the standard deviation.
2. Shewhart charts do not rapidly catch drift from target; a CUSUM does, or an EWMA chart tuned to be close to CUSUM behaviour (i.e. with small λ). Keeping fill weights on target is critical to avoid overfilling and losing money, or under-filling and short-changing customers.
3. The EWMA control limits are: LCL $= 300 - 3 \times 21.7 \times \sqrt{\frac{0.1}{2 - 0.1}} = 285$ mg and UCL $= 315$ mg. Again, divide the standard deviation by $\sqrt{n}$ if you have assumed a subgroup size.
4. The limits with $\lambda = 0.9$ assuming monitoring of the raw data are: LCL $= 300 - 3 \times 21.7 \times \sqrt{\frac{0.9}{2 - 0.9}} = 241.1$ mg and UCL $= 358.8$ mg.
5. $\lambda = 0.9$ approximates a Shewhart chart's limits, while $\lambda = 0.1$ is approximating a CUSUM chart and using more past data. Note how tight the limits are with $\lambda = 0.1$ it will quickly pick up drift and raise an alarm.

END