Statistics for Engineers, 4C3/6C3

Assignment 3

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 27 January 2011**

Assignment objectives

- Interpreting and using confidence intervals.
- Using tests of differences between two population samples.
- Simulating (creating) data and verifying whether theory matches the simulation.

Question 1 [1]

Your manager is asking for the average viscosity of a product that you produce in a batch process. Recorded below are the 12 most recent values, taken from consecutive batches. State any assumptions, and clearly show the calculations which are required to estimate a 95% confidence interval for the mean. Interpret that confidence interval for your manager, who is not sure what a confidence interval is.

Raw data:	[13.7, 14.9, 15.7, 16.1, 14.7, 15.2, 13.9, 13.9, 15.0, 13.0, 16.7, 13.2]
Mean:	14.67
Standard deviation:	1.16

You should the course statistical tables, rather than computer software, to calculate any limits.

Solution

The confidence interval for a mean requires the assumption that the individual numbers are taken from a normal distribution, and they are sampled independently (no sample has an effect on the others). Under these assumptions we can calculate a z-value for the sampled mean, $\overline{x}$, and construct upper and lower bounds reflecting the probability of sampling that z-value.

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

Since we don't know the value of σ, we use the sampled value, $s = 1.16$. But this means our z-value is no longer normally distributed, rather it is t-distributed. The limits, $\pm c_t$ that contain 95% of the area under the t-distribution, with 11 degrees of freedom, are 2.20 (or any close approximation from the tables provided). From this we get the confidence interval:

$$\begin{aligned} -c_t \quad &\leq \quad \frac{\overline{x} - \mu}{s/\sqrt{n}} \quad \leq \quad c_t \\ 14.67 - \frac{2.20 \times 1.16}{\sqrt{12}} \quad &\leq \quad \mu \quad \leq \quad 14.67 + \frac{2.20 \times 1.16}{\sqrt{12}} \\ 13.93 \quad &\leq \quad \mu \quad \leq \quad 15.41 \end{aligned}$$

This confidence interval means that we have 95% confidence that the true average viscosity lies within these bounds. If we took 100 groups of 12 samples, then the limits calculated from 95 of those groups are expected to contain the true mean. It is **incorrect** to say that there is 95% probability the true mean lies within these bounds; the true mean is fixed, there is no probability associated with it.

Question 2 [1.5]

1. At the 95% confidence level, for a sample size of 7, compare and comment on the upper and lower bounds of the confidence interval that you would calculate if:

   (a) you know the population standard deviation

   (b) you have to estimate it for the sample.

   Assume that the calculated standard deviation from the sample, s matches the population $\sigma = 4.19$.

2. As a follow up, overlay the probability distribution curves for the normal and t-distribution that you would use for a sample of data of size $n = 7$.

3. Repeat part of this question, using larger sample sizes. At which point does the difference between the t- and normal distributions become *practically* indistinguishable?

4. What is the implication of this?

Solution

1. This question aims for you to prove to yourself that the t-distribution is **wider (more broad)** than the normal distribution, and as a result, the confidence interval is wider as well. This is because we are less certain of the data's spread when using the estimated variance.

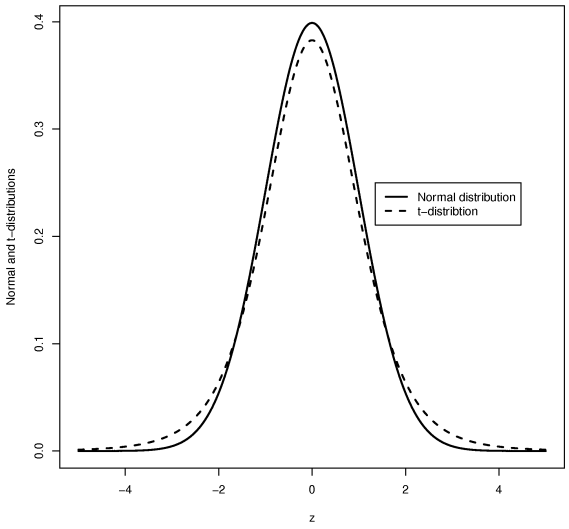
   The confidence intervals are:

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

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

   The 95% region spanned by the t-distribution with 6 degrees of freedom has upper and lower limits at $c_t = \pm$ `qt((1-0.95)/2, df=6)`, i.e. from **-2.45** to **2.45**. The equivalent 95% region spanned by the normal distribution is $c_n = \pm$ `qnorm((1-0.95)/2)`, spanning from **z=-1.96** to **z=1.96**. Everything else in the center of the 2 inequalities is the same, so we only need to compare c_t and c_n.

2. The question asked to overlay the probability distributions (not cumulative probability distributions):



where the above figure was generated with the R-code:

```
n <- 7
z <- seq(-5, 5, 0.01)
prob.norm <- dnorm(z)
prob.t <- dt(z, df=n-1)

bitmap('../images/overlaid-distributions-assign2.jpg', res=300)
plot(z, prob.norm, type="l", ylab="Normal and t-distributions", lwd=2)
lines(z, prob.t, lty=8, lwd=2)  # dashed line
legend(x=1.35, y=0.25, legend=c("Normal distribution", "t-distribtion"),
       lty=c(1,8), lwd=c(2,2))
dev.off()
```

3. Repeated use of the above code but changing n shows that little *practical* difference between the distributions with as little as $n = 20$ samples. After $n = 40$ and especially $n = 60$, there is almost no *theoretical* difference between them.

4. This means that when we do any analysis of large samples of data, say $n > 50$, and if those data are independently sampled, then we can just use the normal distribution's critical value (e.g. the ± 1.96 value for 95% confidence, which you now know from memory), instead of looking up the t-distribution's values.

   Since the wider values from the t-distribution reflect our uncertainty in using an *estimate of the variance*, rather than the population variance, this result indicates that our estimated variances are a good estimate of the population variance for largish sample sizes.

Question 3 [1]

You plan to run a series of 22 experiments to measure the economic advantage, if any, of switching to a corn-based raw material, rather than using your current sugar-based material. You can only run one experiment per day, and there is a high cost to change between raw material dispensing systems. Describe two important precautions you would implement when running these experiments, so you can be certain your results will be accurate.

Solution

Some important precautions one has to take are:

1. Keep all disturbance factors as constant as possible: e.g. use the same staff for all experiments (*Corn* and *Sugar*), keep other variables on the process as constant as possible.

2. Randomize the **order** of the experiments, despite the cost, to obtain independent experimental measurements. For example, if you cannot use the same staff for all experiments, then the experiment order must be randomization. Do not, for example, use group A staff to run the *Corn* experiments and group B staff to run the *Sugar* experiments.

   Randomization is expensive and inconvenient, but is the insurance we pay to ensure the results are not confounded by unmeasured disturbances.

3. Use representative lots of corn- and sugar-based materials. You don't want to run all your experiments on one batch of corn or sugar. What if the batch was bad/good? Then your results will be biased to show no difference if there really was a difference, or shows a significant difference, when perhaps it was only a really good batch of raw materials.

Question 4 [1.5]

We have emphasized several times in class this week that engineering data often violate the assumption of independence. In this question you will create sequences of autocorrelated data, i.e. data that lack independence. The simplest form of autocorrelation is what is called lag-1 autocorrelation:

$$x_k = \phi x_{k-1} + a_k$$

For this question let $a_k \sim \mathcal{N}\left(\mu = 0, \sigma^2 = 25.0\right)$ and consider these 3 cases:

A: $\phi = +0.7$

B: $\phi = 0.0$

C: $\phi = -0.6$

For each case above perform the following analysis (if you normally submit code with your assignment, then only provide the code for one of the above cases):

1. Simulate the following $i = 1, 2, \ldots, 1000$ times:
   - Create a vector of 100 autocorrelated x_k values using the above formula, using the current level of ϕ
   - Calculate the mean of these 100 values, call it $\overline{x}_i$ and store it in a vector
2. Use this vector of 1000 means and answer:
   - Assuming independence, which is obviously not correct for 2 of the 3 cases, nevertheless, from which population should $\overline{x}$ be from, and what are the 2 parameters of that population?
   - Now, using your 1000 simulated means, estimate those two population parameters.
   - Compare your estimates to the theoretical values.

Comment on the results, and the implication of this regarding tests of significance (i.e. statistical tests to see if a significant change occurred or not).

Solution

We expect that case B should match the theoretical case the closest, since data from case B are truly independent, since the autocorrelation parameter is zero. We expect case A and C datasets, which violate that assumption of independence, to be biased one way or another. This question aims to see **how** they are biased.

```
# Number of simulations
nsim <- 1000
x.mean <- numeric(nsim)

set.seed(37)             # so that you can reproduce these results
for (i in 1:nsim)
{
    N <- 100             # number of points in the autocorrelated sequence
    phi <- +0.7          # change this line for case A, B and C
    spread <- 5.0        # standard deviation of random variables
    x <- numeric(N)
    x[1] = rnorm(1, mean=0, sd=spread)
    for (k in 2:N){
       x[k] <- phi*x[k-1] + rnorm(1, mean=0, sd=spread)
    }
    x.mean[i] <- mean(x)
}
theoretical <- sqrt(spread^2/N)
```

```
# Show some output to the user
c(theoretical, mean(x.mean), sd(x.mean))
```

You should be able to reproduce the results I have below, because the above code uses the `set.seed(...)` function, which forces R to generate random numbers in the same order on my computer as yours (as long as we all use the same version of R).

- Case A: `0.50000000, 0.00428291, 1.65963302`
- Case B: `0.50000000, 0.001565456, 0.509676562`
- Case C: `0.50000000, 0.0004381761, 0.3217627596`

The first output is the same for all 3 cases: this is the theoretical standard deviation of the distribution from which the $\overline{x}_i$ values come: $\overline{x}_i \sim \mathcal{N}\left(\mu, \sigma^2/N\right)$, where $N = 100$, the number of points in the autocorrelated sequence. This result comes from the central limit theorem, which tells us that $\overline{x}_i$ should be normally distributed, with the same mean as our individual x-values, but have smaller variance. That variance is σ^2/N, where σ is the variance of the distribution from which we took the raw x values. That theoretical variance value is $25/100$, or theoretical standard deviation of $\sqrt{(25/100)} = \mathbf{0.5}$.

But, the central limit theorem only has one *crucial* assumption: that those raw x values are independent. We intentionally violated this assumption for case A and C.

We use the 1000 simulated values of $\overline{x}_i$ and calculate the average of the 1000 $\overline{x}_i$ values and the standard deviation of the 1000 $\overline{x}_i$ values. Those are the second and third values reported above.

We see in all cases that the mean of the 1000 values nearly matches 0.0. If you run the simulations again, with a different seed, you will see it above zero, and sometimes below zero for all 3 cases. So we can conclude that lack of independence *does not* affect the estimated mean.

The major disagreement is in the variance though. Case B matches the theoretical variance; data that are positively correlated have an inflated standard deviation, 1.66; data that are negatively correlated have a deflated standard deviation, 0.32 when $\phi = -0.6$.

This is problematic for the following reason. When doing a test of significance, we construct a confidence interval:

$$\begin{aligned} -c_t &\leq \frac{\overline{x} - \mu}{s/\sqrt{n}} \leq +c_t \\ \overline{x} - c_t\frac{s}{\sqrt{n}} &\leq \mu \leq \overline{x} + c_t\frac{s}{\sqrt{n}} \\ \text{LB} &\leq \mu \leq \text{UB} \end{aligned}$$

We use an estimated standard deviation, s, whether that is found from pooling the variances or found separately (it doesn't really matter), but the main problem is that s is not accurate when the data are not independent:

- For positive correlations (quite common in industrial data): our confidence interval will be too wide, likely spanning zero, indicating no statistical difference, when in fact there might be one.
- For negative correlations (less common, but still seen in practice): our confidence interval will be too narrow, more likely to indicate there is a difference.

For some reason, this question was poorly answered in almost all assignments. The main purpose was for you to see how modern statistical work is done: we repeat many simulations, calculate an average value and compare it to theoretically expected values.

In this particular example there is a known theoretical relationship between ϕ and the inflated/deflated variance. But in most situations the affect of violating assumptions is too difficult to derive mathematically, so we use computer power to do the work for us: but then we still have to spend time thinking and interpreting the results.

Question 5 [2]

We emphasized in class that the best method of testing for a significant difference is to use an external reference data set. The data I used for the example in class are available on the dataset website, including the 10 new data points from feedback system B.

1. Use these data and repeat for yourself (in R, MATLAB, or Python) the calculations described in class. Reproduce the dot plot, but particularly, the risk value of 11%, from the above data. Note the last 10 values in the set of 300 values are the same as "group A" used in the course slides. The 10 yields from group B are: $[83.5, 78.9, 82.7, 93.2, 86.3, 74.7, 81.6, 92.4, 83.6, 72.4]$.

2. The risk factor of 11% seemed too high to reliably recommend system B to your manager. The vendor of the new feedback has given you an opportunity to run 5 more tests, and now you have 15 values in group B:

$$[83.5, 78.9, 82.7, 93.2, 86.3, 74.7, 81.6, 92.4, 83.6, 72.4, \mathbf{79.1}, \mathbf{84.6}, \mathbf{86.9}, \mathbf{78.6}, \mathbf{77.1}]$$

   Recalculate the average difference between 2 groups of 15 samples, redraw the dot plot and calculate the new risk factor. Comment on these values and *make a recommendation to your manager*. Use bullet points to describe the factors you take into account in your recommendation.

Note: You can construct a dot plot by installing the `BHH2` package in R and using its `dotPlot` function. The `BHH2` name comes from Box, Hunter and Hunter, 2nd edition, and you can read about their case study with the dot plot on page 68 to 72 of their book. The case study in class was based on their example.

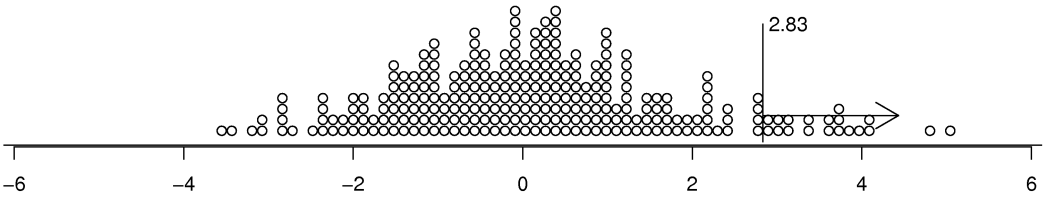
Solution

1. Most people constructed the box-plot without a problem and showed the average difference between group A and B to be 3.04. The number of sequential historical runs that had a larger difference than 3.04 was 31 out of 281, or a risk level of 11%.

2. The boxplot for the additional data points will be different, since we use the average difference between groups of 15 contiguous samples. We can form 271 such groups from the 300 historical data points.

   The average of the last 15 group A points was 79.54%, while the average of the 15 new, group B, data points was 82.37%, so an average difference of 2.83%. This is actually a smaller difference than when we used two groups of 10 samples!

   But, from the 271 historical differences between the means, only 19 had a value greater than 2.83, so the risk factor is 19/271 = 7.0%. The dotplot is shown below



   This is a greatly reduced risk factor, and was found using only 5 more experimental runs. Let's say that this risk is still too high, and you'd like 5% risk. You can recalculate the dotplot with groups of 20 experiments and find the point at which you will have 5% risk and set that as your minimum bound, and let your supplier know that you won't be purchasing the system unless you can see an improvement at that level.

For many companies, this 7% would present an acceptable level of risk, especially if the payoff is high. Other factors you would have to take into account are:

- The size of your budget: what portion of your annual budget will this new change cost?
- What are the annual maintenance fees to keep this improvement in place?
- What is the rate of return on this investment (ROI); is it above your company's internal ROI level? Most companies have an established ROI level that they need to see before approving new capital expenses.

There are obviously other factors, depending on the unique case at each company, but the point is that statistical tools here have allowed you translate the data into dollar values and risk values, which are more interpretable to your manager and colleagues.
