The University of Sydney
School of Mathematics and Statistics

Solutions to Tutorial Week 9

MATH1905: Statistics (Advanced) Semester 2, 2017

Web Page: `http://sydney.edu.au/science/maths/MATH1905`
Lecturer: Michael Stewart

Due to the public holiday, some material relevant to these exercises may not yet have been covered. Please take note of lecturer and tutor announcements relating to this.

1. A random sample of 80 observations from a population **known** to have a standard deviation of 12 gave a sample average of $\bar{x} = 6.55$.

(a) Taking the observed value as an *estimate* of the (unknown) population mean μ, compute a standard error to go with this estimate.

Solution: The standard error associated with an estimate is

- the standard deviation of the corresponding estimator *or*
- an estimate of this if it involves any unknown quantities.

In this case, the sample mean has a *known* standard deviation: $\frac{\sigma_0}{\sqrt{n}} = \frac{12}{\sqrt{80}} \approx 1.342$.

(b) Provide 3 (two-sided) confidence intervals for μ at confidence levels

Solution: A $100(1-\alpha)\%$ confidence interval is of the form

$$\bar{x} \pm c\,(\text{s.e.})$$

where the constant c satisfies the relation

$$P\left\{-c \leq \frac{\bar{X}-\mu}{\text{S.E.}} \leq c\right\} = 1 - \alpha\,.$$

(*take note of the distinction between capitals (random variables) and lower case (values taken by them)*). In this case, where the "true" standard error is known, the random variable

$$\frac{\bar{X}-\mu}{\text{S.E.}} \sim N(0,1)$$

and so we need "critical values" from the standard normal distribution; these can be obtained using the function `qnorm()` and noting the symmetry of the standard normal distribution the "two-sided" condition above is equivalent to the "one-sided" condition

$$P\left\{\frac{\bar{X}-\mu}{\text{S.E.}} \leq c\right\} = 1 - \frac{\alpha}{2}\,.$$

(i) 90%

Solution: This corresponds to $\alpha = 0.1$, so the correct value of the "mutliplier" c is given by the command

```
qnorm(0.95)
```

```
[1] 1.644854
```

(since here $1 - \frac{\alpha}{2} = 0.95$). Thus a 90% confidence interval is given by $\bar{x} \pm 1.645(\text{s.e.})$ which gives

```
6.55+c(-1,1)*1.645*1.342
```

```
[1] 4.34241 8.75759
```

(ii) 95%

Solution: This corresponds to $\alpha = 0.05$, so the correct value of the "mutliplier" c is given by the command

```
qnorm(0.975)
```

```
[1] 1.959964
```

(since here $1 - \frac{\alpha}{2} = 0.975$). Thus a 95% confidence interval is given by $\bar{x} \pm 1.96$(s.e.) which gives

```
6.55+c(-1,1)*1.96*1.342
```

```
[1] 3.91968 9.18032
```

(iii) 99%

Solution: This corresponds to $\alpha = 0.01$, so the correct value of the "mutliplier" c is given by the command

```
qnorm(0.995)
```

```
[1] 2.575829
```

(since here $1 - \frac{\alpha}{2} = 0.995$). Thus a 99% confidence interval is given by $\bar{x} \pm 2.576$(s.e.) which gives

```
6.55+c(-1,1)*2.576*1.342
```

```
[1]  3.093008 10.006992
```

(c) State what assumption(s) and/or theoretical result(s) one needs for these confidence intervals to be

(i) valid;

Solution: We need the sample mean to be *exactly* normally distributed, which follows if the original observations are also normally distributed.

(ii) *approximately* valid.

Solution: By the Central Limit Theorem if the sample size is "large" the even if the original observations are not normally distributed, the sample mean is *approximately* normally distributed. This sample size is "large" (so long as the "parent" population is not too non-normal) so in this sense the intervals are approximately valid.

(d) Provide 3 lower confidence limits for μ at levels

Solution: A $100(1-\alpha)$% lower confidence limit is (loosely speaking) the "lowest value of μ consistent with the data" (in a particular sense). It is of the form

$$\bar{x} - c(\text{s.e.})$$

where the (positive) constant c is chosen so that

$$P\left\{\frac{\bar{X} - \mu}{\text{S.E.}} \leq c\right\} = 1 - \alpha\,.$$

(i) 90%

Solution: The required c is given by

```
qnorm(0.9)
```

```
[1] 1.281552
```

which gives a lower confidence limit of

```
6.55 - 1.282*1.342
```

```
[1] 4.829556
```

(ii) 95%

Solution: The required c is given by

```r
qnorm(0.95)
```

```
[1] 1.644854
```

which gives a lower confidence limit of

```r
6.55 - 1.645*1.342
```

```
[1] 4.34241
```

(iii) 99%

Solution: The required c is given by

```r
qnorm(0.99)
```

```
[1] 2.326348
```

which gives a lower confidence limit of

```r
6.55 - 2.326*1.342
```

```
[1] 3.428508
```

(e) Compute a value for the z-statistic for testing the hypothesis $H_0: \mu = 5$.

Solution: This is

$$\frac{\text{estimate} - \text{hypothesised value}}{\text{s.e.}} = \frac{\bar{x} - 5}{\sigma_0/\sqrt{n}}$$

which is

```r
(6.55-5)/1.342
```

```
[1] 1.154993
```

(f) Compute a p-value for a test of H_0 against the (one-sided) alternative $H_1: \mu > 5$.

Solution: This is a one-sided test, with larger values of the z-statistic constituting more evidence against H_0. Thus the event "at least as much evidence against H_0 as was observed" is simply that the z-statistic take a value *at least this large*. Under H_0 the z-statistic has a $N(0,1)$ distribution, so the p-value is $P(Z \geq 1.155)$ where $Z \sim N(0,1)$, which is given by

```r
1-pnorm(1.155)
```

```
[1] 0.1240452
```

(g) Write a sentence giving an interpretation of the p-value.

Solution: The value of the test statistic is not very large, indeed even if the true $\mu = 5$ it would take a larger value more than 10% of the time by chance alone. Thus this does **not** constitute very strong evidence against H_0.

(h) Is this significant at the

(i) 10% level;

(ii) 5% level;

(iii) 1% level?

Solution: For a result to be significant at level α the p-value must be *smaller* than α. So the answer is "no" to all 3.

(i) How do your answers change if the alternative is two-sided?

Solution: The only change is that the p-value is doubled (since the one-sided p-value is ≤ 0.5; otherwise the two-sided p-value would be 1). This constitutes *even less* evidence against H_0.

(j) How do your hypothesis test answers relate to your confidence interval/limit answers?

Solution: There is a strong connection between hypothesis tests and confidence intervals: both provide mechanisms for identifying parameter values that are "plausible" or "consistent with the data" in a particular sense. Remember, large p-value means "data consistent with H_0" (in this sense). Therefore in general we have the relationship:

$$\text{p-value of } H_0\colon \mu = \mu_0 \geq \alpha \;\Leftrightarrow\; \mu_0 \text{ is in } 100(1-\alpha)\% \text{ confidence interval.}$$

There are one-sided and two-sided versions of this relationship:

- if the alternative is two-sided i.e. $H_1\colon \mu \neq \mu_0$, the corresponding confidence interval is also two-sided;
- if the alternative is one-sided, the confidence interval is also one-sided, specifically
  - if the alternative is $H_1\colon \mu > \mu_0$, the corresponding confidence interval is of the form $[\ell, \infty)$;
  - if the alternative is $H_1\colon \mu < \mu_0$, the corresponding confidence interval is of the form $(-\infty, u]$;
  - the symbols ℓ and u here respectively represent lower and upper "confidence limits";
    - they may be interpreted as smallest (resp. largest) parameter values consistent with the data.

In our specific examples, we have $\alpha = 0.1, 0.05, 0.01$, both two-sided and one-sided versions. In all cases, the p-value is bigger than α, thus in all cases the hypothesised value 5 is in the corresponding confidence interval.

2. A p-value of 0.98 indicates that the null hypothesis is true. Comment.

Solution: A p-value of 0.98 indicates that the observed test statistic is consistent with the null hypothesis. A statistical test does not prove that H_0 is true or false.

3. A random sample of 30 households was selected as part of a study on electricity usage, and the number of kilowatt-hours (kWh) was recorded for each household in the sample for the March quarter of 2013. The average usage was found to be 375kWh. In a very large study in the March quarter of the previous year it was found that the standard deviation of the usage was 81kWh. Assuming the standard deviation is unchanged and that the usage is normally distributed, which of the following options is an appropriate expression for the 99% confidence interval for the mean usage in the March quarter of 2013.

(a) $375 \pm 2.756 \times \dfrac{81}{\sqrt{30}}$

(b) $375 \pm 2.576 \times \dfrac{9}{\sqrt{30}}$

(c) $375 \pm 2.326 \times \dfrac{81}{\sqrt{30}}$

(d) $375 \pm 2.576 \times \dfrac{81}{\sqrt{30}}$

Solution: The main point here is that the (population) standard deviation is assumed known. Thus we need to identify the value c such that

$$P(-c \leq Z \leq c) = 0.99$$

which also satisfies

$$P(Z \leq c) = 0.995$$

therefore c can be obtained using the command

```
qnorm(0.995)
```

```
[1] 2.575829
```

The assumed *standard deviation* is 81 and so the correct answer is (d).

4. A p-value of 0.2 means:

(a) there is 20% chance H_0 true,

(b) there is 20% chance H_1 true,

(c) there is strong evidence against H_0,

(d) the data are consistent with H_0.

Solution: The correct answer is (d).

5. The performance in kilometres/litre (km/l) of a particular model of car tested by machine is 7 km/l and the distribution appears normal. Company engineers have redesigned the carburettor in an effort to improve the performance and have equipped a sample of 36 cars with this new carburettor. When tested the average performance of the sample was 7.6 km/l, with a sample standard deviation of 1.5 km/l. The sample has no "outliers". What are the null (H_0) and alternative (H_1) hypotheses to be tested using the sample?

(a) $H_0 : \mu = 7,\ H_1 : \mu \neq 7$,

(b) $H_0 : \mu < 7,\ H_1 : \mu = 7$,

(c) $H_0 : \mu > 7,\ H_1 : \mu \leq 7$,

(d) $H_0 : \mu = 7,\ H_1 : \mu > 7$.

Solution: (d) $H_0 : \mu = 7$ against $H_1 : \mu > 7$.

6. In the preceding problem the appropriate p-value of the test is

(a) 0.6554,

(b) 0.0082,

(c) between 0.01 and 0.025,

(d) none of the these.

Solution: The correct answer is (c). The procedure is:

- assumptions: X_i's iid $N(\mu, \sigma^2)$;
- test statistic: $\dfrac{\bar{X} - 7}{S/\sqrt{n}} \sim t_{35}$ if the true population mean is indeed 7 (i.e. if H_0 true);
- observed value of test statistic: $\dfrac{7.6 - 7}{1.5/\sqrt{36}} = 2.4$;
- p-value $= P(t_{35} \geq 2.4) =$ `1-pt(2.4,35)` ≈ 0.011.

The p-value is smaller than $\alpha = 0.05$ therefore we reject the null hypothesis at the 5% level of significance. Thus, there is evidence to suggest that the redesigned carburettor improves fuel efficiency.

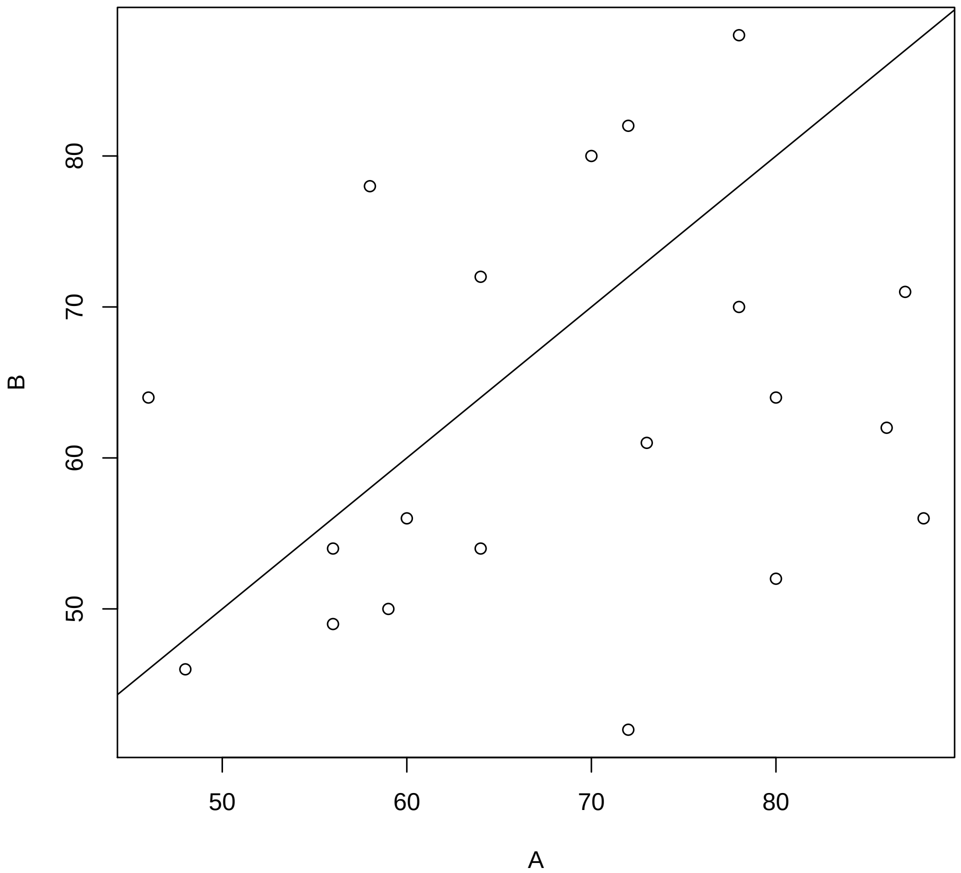
7. An insurance assessor has received estimates from two different repair garages (A and B) for minor repairs on 20 cars and wants to know if there strong evidence of a difference in estimates on average.

The estimates on each of the 20 cars, from each of the two garages are entered into the R objects `A` and `B` as below:

```
A=c(48,56,87,88,86,64,80,78,72,70,80,58,72,60,64,46,56,59,73,78)
B=c(46,49,71,56,62,54,52,88,82,80,64,78,42,56,72,64,54,50,61,70)
```

Based on the following output, answer the questions that follow:

```
plot(A,B)
abline(0,1)
```




```r
mean(A)
```

```
[1] 68.75
```

```r
mean(B)
```

```
[1] 62.55
```

```r
sd(A)
```

```
[1] 12.71499
```

```r
sd(B)
```

```
[1] 12.92275
```

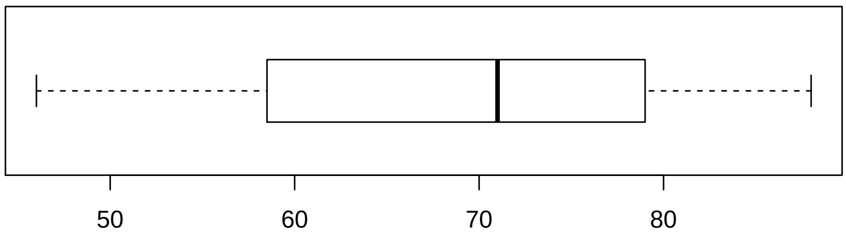
```r
sd(A-B)
```

```
[1] 15.52112
```

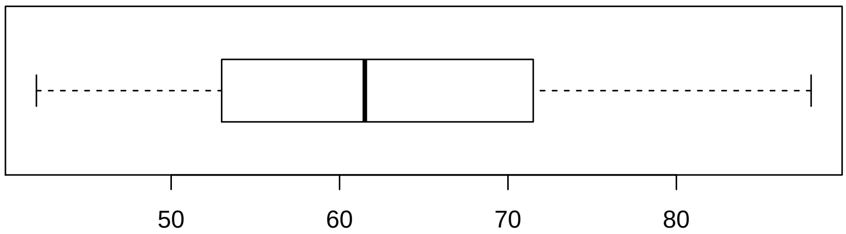
```
qt(c(0.9,0.95,0.975,0.99,0.995),df=19)
```

```
[1] 1.327728 1.729133 2.093024 2.539483 2.860935
```

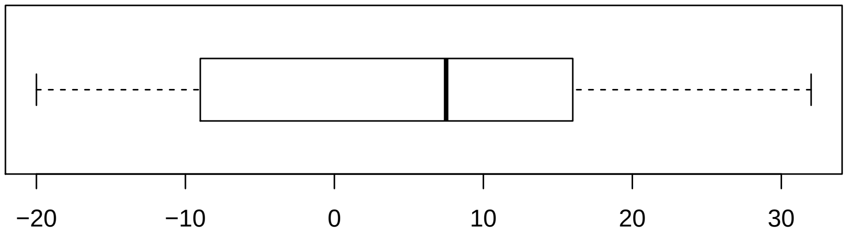
```
boxplot(A,horizontal=T)
```




```
boxplot(B,horizontal=T)
```




```
boxplot(A-B,horizontal=T)
```



(a) Let μ_d represent the true mean difference for minor repairs of this type. Explain why a paired test is appropriate.

Solution: Although there are two samples, corresponding observations in each are paired so a paired t-test is appropriate (as opposed to us having two *independent* samples).

(b) Making an appropriate normality assumption, set up the appropriate hypotheses and perform a paired t-test (i.e. compute observed value of test statistic and obtain p-value).

Solution: The "appropriate" normality assumption is that the *differences* are normally distributed with some mean μ. The hypotheses are then $H_0\colon \mu = 0$ versus the alternative $H_1\colon \mu \neq 0$ (there was no indication – before the fact – which garage might give the higher quotes, so we should use a two-sided alternative here).

Writing $\bar{d}$ for the observed mean difference and s_d for the observed sample standard deviation of the differences, the observed value of the test statistic is

$$\frac{\bar{d}}{s_d/\sqrt{n}}$$

whose numerical value is given by the commands below:

```
dbar=mean(A)-mean(B)
dbar
```

```
[1] 6.2
```

```
sdiff=sd(A-B)
sdiff
```

```
[1] 15.52112
```

```
stat=dbar/(sdiff/sqrt(20))
stat
```

```
[1] 1.78642
```

Under the null hypothesis of no actual mean difference, the statistic would have a t_{19} distribution. The p-value is thus given by the probability

$$2P(t_{19} > 1.786)$$

Using the output above, note that $P(t_{19} > 1.786)$ is between $P(t_{19} > 2.093)$ and $P(t_{19} > 1.729)$, that is between 0.025 and 0.05. Thus the *two-sided* p-value is between 0.05 and 0.1.

We can of course compute this directly using the R command

```
2*(1-pt(1.786,df=19))
```

```
[1] 0.09006958
```

This is not small, thus there is not really enough evidence to suggest a difference.

(c) Comment on the assumption of normality by referring to the appropriate boxplot(s) above.

Solution: The appropriate boxplot is the last one, the boxplot of *differences*. It is reasonably symmetric with no outliers, so the normality assumption *for the differences* seems reasonable.