PO11Q: Introduction to Quantitative Political Analysis I

Worksheet Week 8 - Solutions

Dr Florian Reiche

F.Reiche@warwick.ac.uk

The t-distribution

1. How df changes the shape and critical values
   Using two-tailed, $\alpha = 0.05$ (so $t_{\text{crit}} = t_{0.025,df}$ with $df = n - 1$):

   - $n = 3$ ($df = 2$): $t_{\text{crit}} \approx 4.303$
   - $n = 5$ ($df = 4$): $t_{\text{crit}} \approx 2.776$
   - $n = 10$ ($df = 9$): $t_{\text{crit}} \approx 2.262$
   - $n = 20$ ($df = 19$): $t_{\text{crit}} \approx 2.093$
   - $n = 50$ ($df = 49$): $t_{\text{crit}} \approx 2.009$

   The normal reference is $z_{0.025} \approx 1.960$.
   Within 0.05 of 1.96 occurs by $n \approx 50$ ($df = 49$, 2.009 is 0.049 above 1.96).

   Explanation: as $df \to \infty$, the t distribution $\to N(0, 1)$, so tails thin and t_{crit} decreases toward 1.96.

2. One-tailed vs two-tailed critical regions
   With $df = 14$ and $\alpha = 0.05$:

   - Two-tailed: $t_{0.025,14} \approx 2.145$, cut points at ± 2.145
   - One-tailed (upper): $t_{0.05,14} \approx 1.761$, single cut point at $+1.761$

   For $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$, the one-tailed threshold $t_{0.05,14}$ is relevant because only the upper tail provides evidence against H_0.

   In both cases, the shaded area equals α:

   - two-tailed shades two symmetric regions each of area $\alpha/2$;
   - one-tailed shades a single upper tail of area α.

3. Relating t statistics to sample size
   Observed $t = 2.10$, two-tailed $\alpha = 0.05$.

   a. $df = 9$: $t_{0.025,9} \approx 2.262$. Since $|2.10| < 2.262$, do not reject H_0.
   b. $df = 29$: $t_{0.025,29} \approx 2.045$. Since $|2.10| > 2.045$, reject H_0.
   c. As n increases ($df \uparrow$), t_{crit} decreases toward 1.96; with the same observed t, rejection becomes more likely.

Confidence Interval – Dynamics

1. Exploring the role of sample size With 95% level ($\alpha = 0.05$) and $s = 0.9$, the half-width is

$$\text{half-width} = t^* \frac{s}{\sqrt{n}}, \quad t^* = t_{0.025,\,df},\ df = n - 1.$$

   Using standard t–quantiles:

   - $n = 3$ ($df = 2$): $t^* \approx 4.303 \;\Rightarrow\;$ half-width ≈ 1.295
   - $n = 5$ ($df = 4$): $t^* \approx 2.776 \;\Rightarrow\;$ half-width ≈ 1.118
   - $n = 10$ ($df = 9$): $t^* \approx 2.262 \;\Rightarrow\;$ half-width ≈ 0.643
   - $n = 15$ ($df = 14$): $t^* \approx 2.145 \;\Rightarrow\;$ half-width ≈ 0.498
   - $n = 20$ ($df = 19$): $t^* \approx 2.093 \;\Rightarrow\;$ half-width ≈ 0.421
   - $n = 30$ ($df = 29$): $t^* \approx 2.045 \;\Rightarrow\;$ half-width ≈ 0.336

2. Confidence level and t^*

   a) Fix $n = 10$ ($df = 9$), $s = 0.9$; two-tailed t^*:

   - 80%: $t_{0.10,9} \approx 1.383$
   - 90%: $t_{0.05,9} \approx 1.833$
   - 95%: $t_{0.025,9} \approx 2.262$
   - 99%: $t_{0.005,9} \approx 3.250$

   b) Higher confidence $\Rightarrow$ smaller α $\Rightarrow$ fatter central region $\Rightarrow$ larger quantile t^*; hence $2\,t^* \frac{s}{\sqrt{n}}$ increases with the level.

   c) Width ratio (same n, s) is the ratio of the two values of t^*:

$$\frac{\text{width}_{99\%}}{\text{width}_{90\%}} = \frac{t_{0.005,9}}{t_{0.05,9}} \approx \frac{3.250}{1.833} \approx 1.77.$$

3. Trade-off between variability and sample size

   a) At 95%, half-width $= t^* \frac{s}{\sqrt{n}}$.

   - $(n, s) = (8, 0.6)$: $df = 7$, $t^* \approx 2.365$, so half-width $\approx 2.365 \cdot \frac{0.6}{\sqrt{8}} \approx 0.502$
   - $(n, s) = (20, 1.0)$: $df = 19$, $t^* \approx 2.093$, so half-width $\approx 2.093 \cdot \frac{1.0}{\sqrt{20}} \approx 0.468$

   Thus $(8, 0.6)$ is (slightly) wider—smaller n and larger t^* outweigh the smaller s.

   b) Keep $s = 0.9$, level 95%. Need $t^* \frac{0.9}{\sqrt{n}} \le 0.30$. Using $t_{0.025,\,n-1}$, the smallest n that satisfies this is about $n \approx 37$ (since $t_{0.025,36} \approx 2.03$ gives $2.03 \cdot 0.9/\sqrt{37} \approx 0.30$).

   c) Hold $n = 12$ ($df = 11$), level 90% so $t^* = t_{0.05,11} \approx 1.796$. Full width $= 0.9$ means

$$2\,t^* \frac{s}{\sqrt{12}} = 0.9 \quad\Rightarrow\quad s = \frac{0.9\sqrt{12}}{2\,t^*} \approx \frac{0.9 \cdot 3.464}{2 \cdot 1.796} \approx 0.868.$$

Confidence Interval – Calculations

1. The number of diners visiting a restaurant on a Thursday is normally distributed with a mean of 150 and standard deviation of 30. One Thursday only 100 people eat in the restaurant, and the manager says, “next week will be better”.

    a. What is the probability she is right?

    Finding a z-score allows claims to be made on how likely an event is to occur.

    $\mu = 150$
    $\sigma = 30$
    $x = 100$

    $Z = (x - \mu)/\sigma$
    $Z = (100 - 150)/30$
    $Z = -1.67$

    This z-score can be used with the one-sided z-tables to find a probability of a more extreme value being found. This value is 0.048.
    As the manager is referring to next week having MORE diners, this is the opposite of the probability found which is that there will be FEWER diners.
    Therefore, the probability she is right is, 1 - 0.048 = 0.95

    b. The number of diners on a Friday is also normally distributed with a mean of 200 and a standard deviation of 50. Which two values, symmetrical around the mean contain the number of Friday diners 80% of the time?

    Confidence interval. z-tables provide a z-score for a two tailed 80% probability.
    $z = 1.29$
    $\mu = 200$
    $\sigma = 50$

    $200 \pm (1.29 \times 50) = 135.5; 264.5$

2. A researcher is analysing individuals’ relative fear of being a victim of burglary on a 1-100 scale. A random sample of 9 individuals found a mean score of 47 on the scale with a sample variance of 158.76 for fear of being burgled.

    a. What distribution would be used to calculate an 80% confidence interval around this mean?
    A t-distribution as we don’t know the population standard deviation and n is small.

    b. Construct that interval.

$\bar{x} = 47$
$n = 9$
t from tables = 1.397

$s = \sqrt{158.76}$
$s = 12.6$

Confidence interval formula
$\bar{x} \pm t \times \frac{s}{\sqrt{n}}$
Lower bound
$47 - 1.397 \times \frac{12.6}{\sqrt{9}} = 47 - 5.867 = 41.13$

Upper bound
$47 + 1.397 \times \frac{12.6}{\sqrt{9}} = 47 + 5.867 = 52.87$

c. Is this a suitable sample size for seeing whether individuals are more nervous around burglary or murder, which is found to have an 80% confidence interval between 2.76 and 14.65?

Yes, as the bounds are very clearly different with the lower bound far from the upper bound for murder.

3. We are investigating the height of men in the UK. For this we have obtained a random sample of 100 UK men and found they had a mean height of 180cm with a standard deviation of 10cm.
   a. Construct a 95% confidence interval for the mean height of UK males.

$\bar{x} = 180$
$s = 10$
$n = 100$

As the population standard deviation is not known, the t distribution and t need to be used.

Find the t-score for a 95% confidence interval in the t-table with 99 df.
$t = 1.984$

Confidence interval:
$\bar{x} \pm t \times \frac{s}{\sqrt{n}}$

Lower bound:
$180 - 1.984 \times \frac{10}{\sqrt{100}} = 180 - 1.984 = 178.02$
Upper bound:
$180 + 1.984 \times \frac{10}{\sqrt{100}} = 181.98$

b. Select all true statements concerning the constructed confidence interval and justify your choice for each statement.
   i. The probability of the population mean being within the upper and lower bounds is 95%.
      FALSE - The population mean is fixed but unknown and therefore can either be inside the bounds or outside. The Probability is therefore 50%.

   ii. 95% of men's heights fall between the upper and lower bound.
      FALSE - The distribution calculated is not the distribution of men's height, but the sampling distribution of the mean male height.

   iii. 95% of the cases in the sample fall between the upper and lower bound.
      FALSE - The distribution calculated is not of men's height in this sample, but the sampling distribution of the mean male height.

   iv. On average 95% of confidence intervals constructed would contain the population mean.
      TRUE

   v. On average 95% of the means of samples with 100 respondents will fall within the upper and lower bands.
      FALSE - This confidence interval is not making statements about various sample means but rather about the population mean.

   vi. On average 95% of the sample means equal the population mean.
      FALSE - The confidence interval is a range and does not make claims about where the population mean is exactly.

R Exercises

see RScript in the Companion's Download Section