MONASH University

Information Technology

FIT1006
Business Information Analysis

Lecture 16
Estimation

Topics covered:

- Small samples.
- The t-Distribution.
  - which adjusts the C.I. when s is estimated from the data by s and corrects for small samples.
- Setting the sample size for a required level of accuracy.

We did this yesterday

Motivating Problem

- Would Labor have won a Federal Election if an election is to be held today?
- The Australian Newspoll had the two-party preferred vote at: Labor 51% Liberal-NP 49% from a sample of 1,160 people chosen at random.
- Hint: Find a 95% CI for the expected Liberal-NP vote.
- Ref: http://www.theaustralian.com.au/national-affairs/newspoll

From yesterday …

Are you 95% confidence that Labor would win?

- Find a 95% CI for the expected Labor vote.
- p = 0.51, n = 1,160.
- The 95% CI is:

$$\pi = p \pm 1.96\, \sigma_p \; ; \sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

$$\pi = 0.51 \pm 1.96 \sqrt{\frac{0.51 * 0.49}{1160}} = 0.51 \pm 0.029$$

- LCL (Lower Confidence Limit) = 0.51 – 0.029 = 0.481
- UCL (Upper Confidence Limit) = 0.51 + 0.029 = 0.539

Small Samples

- One of the fundamental assumptions of the Central Limit Theorem is that of large sample sizes are used.
- ‘Large’ means at least 30 in practice.
- When sample sizes are small and the variance of the population unknown, the Normal distribution cannot be used as the basis of a confidence interval.
- Instead the t-Distribution is used.

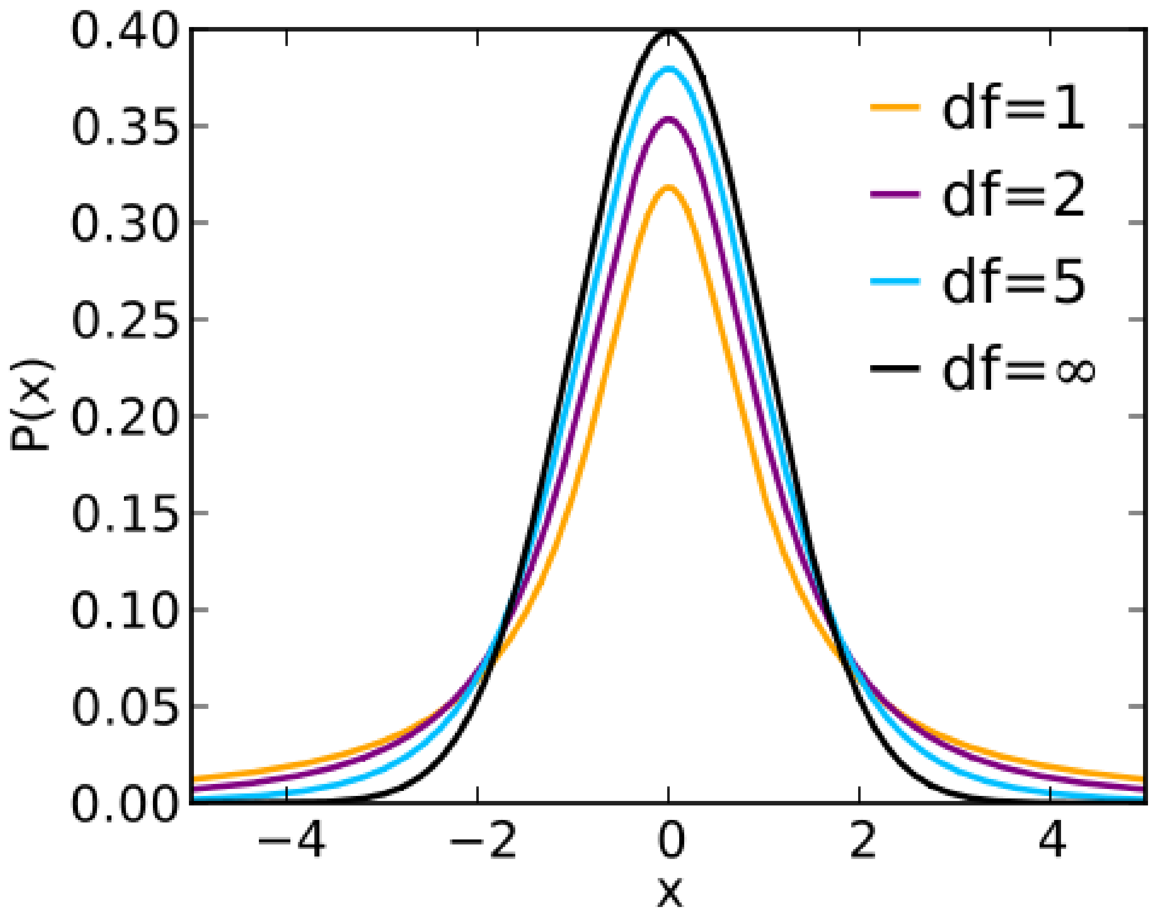
Student’s t-Distribution

- The t-Distribution was derived by W. S. Gosset, a scientist working for the Guinness brewery. He published under the pseudonym ‘student.’ As a consequence the distribution is commonly known as student’s t distribution.
- The t-Distribution has three parameters, μ, σ and ‘degrees of freedom’, ν.
- The t distribution is (heavy-tailed) for small values of n. As n increases, the shape of the t-Distribution becomes closer to the Normal distribution.

Degrees of Freedom

- The number of degrees of freedom or ν, refers to the number of observations that are free to vary when determining the variance or standard error of a sample.
- The general rule for calculating the number of degrees of freedom is to count the number of observations and subtract 1 for each statistic that is derived from the sample.
- In practice, for one-sample problems, ν equals the number of observations less 1 (because we use the *derived* sample mean).

Comparison of *t* and *z*



Source: http://en.wikipedia.org/wiki/Student's_t-distribution

Tables for the t-Distribution.

- On Excel file PROBDIST.XLS

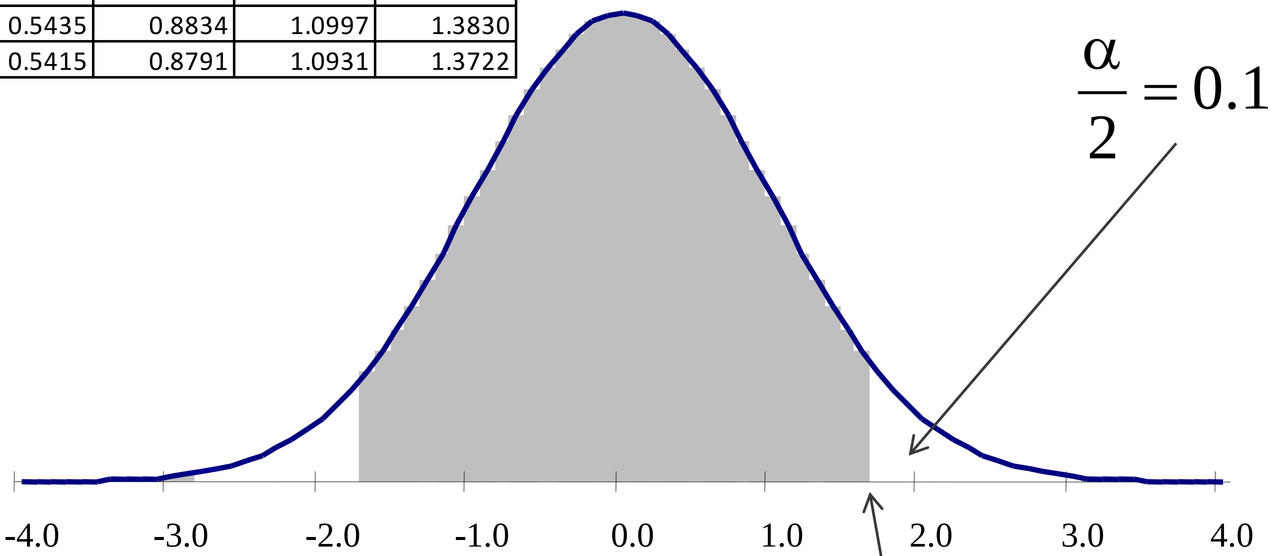
Critical Values of the t Distribution								
Table gives upper critical values only								
				a				
n	**0.300**	**0.200**	**0.150**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**
1	0.7265	1.3764	1.9626	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.6172	1.0607	1.3862	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.5844	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.5686	0.9410	1.1896	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.5594	0.9195	1.1558	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.5534	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.5491	0.8960	1.1192	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.5459	0.8889	1.1081	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.5435	0.8834	1.0997	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.5415	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.5399	0.8755	1.0877	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.5386	0.8726	1.0832	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.5375	0.8702	1.0795	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.5366	0.8681	1.0763	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.5357	0.8662	1.0735	1.3406	1.7531	2.1314	2.6025	2.9467

Upper critical value

n	0.300	0.200	0.150	0.100
1	0.7265	1.3764	1.9626	3.0777
2	0.6172	1.0607	1.3862	1.8856
3	0.5844	0.9785	1.2498	1.6377
4	0.5686	0.9410	1.1896	1.5332
5	0.5594	0.9195	1.1558	1.4759
6	0.5534	0.9057	1.1342	1.4398
7	0.5491	0.8960	1.1192	1.4149
8	0.5459	0.8889	1.1081	1.3968
9	0.5435	0.8834	1.0997	1.3830
10	0.5415	0.8791	1.0931	1.3722

Upper critical value is based on upper region.

If you're looking for a 80% confidence level, then $\alpha = 1 - 0.8 = 0.2$



$$\frac{\alpha}{2} = 0.1$$

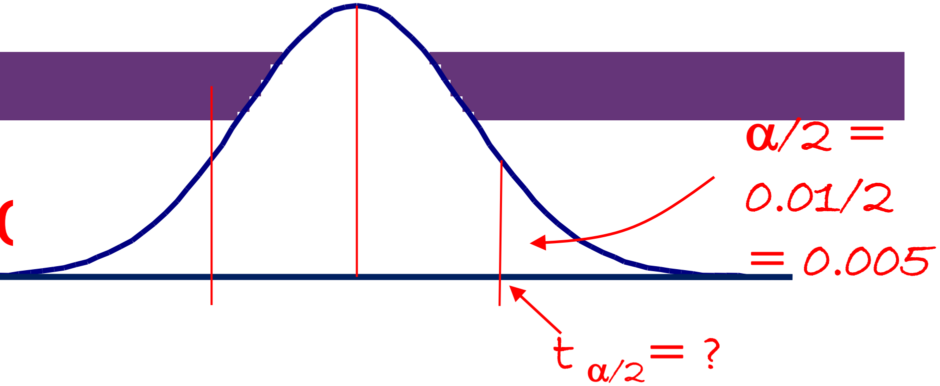
$$t_{(4,0.1)} =$$

If your sample size is 5, then $v = 5 - 1 = 4$

Example 1

- Five experiments were conducted to determine the amount of silica in water, measured in parts per million (ppm).
- Data: 229, 255, 280, 203, 229.
- Estimate the mean amount of silica using a 99% confidence interval.

https://flux.qa (Feed cc



Question 1

For a sample size of 5, and a 99% confidence interval, the corresponding t statistic is: $v = 5-1 = 4$

A. 3.7469

✓ B. 4.6041

C. 3.3649

D. 4.0321

n	0.300	0.200	0.150	0.100	0.050	0.025	0.010	0.005
1	0.7265	1.3764	1.9626	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.6172	1.0607	1.3862	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.5844	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.5686	0.9410	1.1896	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.5594	0.9195	1.1558	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.5534	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.5491	0.8960	1.1192	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.5459	0.8889	1.1081	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.5435	0.8834	1.0997	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.5415	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693

Solution

Given
Data: 229, 255, 280, 203, 229

$\bar{x} = 239.2$ and $s = 29.3$

Find the mean and std dev. from these values.

$\alpha = (1 - \text{confidence level}) = (1 - 0.99) = 0.01$, Thus $\dfrac{\alpha}{2} = 0.005.$

The sample size is 5, hence DOF (υ) is 4.

From tables of the t-distribution $t_{(4,0.005)} = 4.604.$

A 99% CI for μ is $\mu = \bar{x} \pm t_{\frac{\alpha}{2}}\left(\dfrac{s}{\sqrt{n}}\right).$

Thus a 99% CI is $\mu = 239.2 \pm 4.604\left(\dfrac{29.3}{\sqrt{5}}\right).$

i.e. $\mu = 239.2 \pm 60.3$ ppm at the 99% confidence level.

$178.9 < \mu < 299.5$

Confidence Intervals in SYSTAT

- The descriptive statistics menu in SYSTAT determines 95% confidence intervals by default, but can be set to any value. Using the data from the previous question.

SILICA_PPM	
N of cases	5
Minimum	203.000
Maximum	280.000
Mean	239.200
95% CI Upper	275.575
95% CI Lower	202.825
Standard Dev	29.295

SILICA_PPM	
N of cases	5
Minimum	203.000
Maximum	280.000
Mean	239.200
99% CI Upper	299.519
99% CI Lower	178.881
Standard Dev	29.295

Example 2

- A shop reported the following numbers of shoppers over two weeks. Calculate a 95% confidence interval for the average number of customers.
- Data: 99 179 126 156 132 31 122 126 123 150 158 160 67 111

Descriptive statistics are:

SHOPPERS	
N of cases	14
Minimum	31.000
Maximum	179.000
Mean	124.286
Standard Dev	39.169

https://flux.qa (Feed code: SJ6KGV)

Question 2

For a sample size of 14, and a 95% confidence interval, the corresponding t statistic is:

A. 1.7709

✓ B. 2.1604

C. 1.7613

D. 2.1448

$\alpha/2 = 0.05/2 = 0.025$

n	0.300	0.200	0.150	0.100	0.050	0.025	0.010	0.005
8	0.5459	0.8889	1.1081	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.5435	0.8834	1.0997	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.5415	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.5399	0.8755	1.0877	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.5386	0.8726	1.0832	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.5375	0.8702	1.0795	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.5366	0.8681	1.0763	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.5357	0.8662	1.0735	1.3406	1.7531	2.1314	2.6025	2.9467
16	0.5350	0.8647	1.0711	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.5344	0.8633	1.0690	1.3334	1.7396	2.1098	2.5669	2.8982

Solution

Given:

SHOPPERS	
N of cases	14
Minimum	31.000
Maximum	179.000
Mean	124.286
Standard Dev	39.169

From the data: $\bar{x} = 124.3, \quad \sigma_{\bar{x}} = 10.5, \quad \underline{t_{0.025(13)} = 2.160}$

$\sigma / \sqrt{n} = 39.169/3.742 = 10.47$

$$95\% C.I. = 124.3 \pm 2.160 \times 10.5$$

$$= (101.7,\ 146.9)$$

SHOPPERS	
N of cases	14
95% CI Upper	146.901
95% CI Lower	101.670

Pooled Samples – Diff. of means

- The usual way to calculate the standard error
- For the difference of means is: $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$
- However, when we have two small samples of similar variance it is possible to calculate the variance of the 'pooled' sample which gives a smaller standard error.
- See following slide.

Pooled Samples – C.I. Calculations

We can determine a confidence interval for the difference of population means for small samples using the variance of the pooled sample.

Suppose we have $\bar{x}_1$ and $\bar{x}_2, s_1^2$ and s_2^2 we wish to find a C.I. for $\mu_1 - \mu_2$.

We assume both populations have the same variance and make an estimate of the population standard deviation with the formula

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \text{ and standard error } s_{\bar{x}_1 - \bar{x}_2} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

We use the t distribution with degrees of freedom $\nu = n_1 + n_2 - 2$.

Our $(1 - \alpha)$ confidence interval is given by $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2}$

Pooled Samples – Example

- The number of claims processed by two workers is measured over a period of (different) days.
- Worker A: 23, 45, 21, 22, 17, 42, 45, 41, 49, 19.
- Worker B: 33, 23, 19, 51, 32, 15.
- Calculate a 95% C.I. For the difference in the average number of claims (A-B) processed by the workers.

Pooled Samples – Summary Stats

	Worker A	Worker B
	23	33
	45	23
	21	19
	22	51
	17	32
	42	15
	45	
	41	
	49	
	19	
N	10.00	6.00
Mean	32.40	28.83
St Dev	12.92	12.97

Pooled Samples

Population standard deviation, $s = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ and

Standard error $s_{\bar{x}_1-\bar{x}_2} = s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$.

We use the t distribution with degrees of freedom $\nu = n_1 + n_2 - 2$.

Our (1-α) confidence interval is given by $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_{\bar{x}_1-\bar{x}_2}$

From the data :

$\bar{x}_1 = 32.40, \quad s_1 = 12.92, \quad n_1 = 10, \bar{x}_2 = 28.83, \quad s_2 = 12.97, \quad n_2 = 6$

$$s = \sqrt{\frac{9\times 12.92^2 + 5\times 12.97^2}{14}} = 12.94$$

$\nu = n_1 + n_2 - 2 =$
$10 + 6 - 2 = 14$

$$s_{\bar{x}_1-\bar{x}_2} = 12.94\sqrt{\frac{1}{10} + \frac{1}{6}} = 6.68$$

$$t_{(0.025,14)} = 2.147$$

$$95\%C.I. = (32.40 - 28.83) \pm 2.147\times 6.68 \quad = 3.57 \pm 14.34$$

$$= (-10.78, 17.91)$$

Pooled Samples – SYSTAT Output

Variable	N	Mean	Standard Deviation
WORKERA	10.000	32.400	12.920
WORKERB	6.000	28.833	12.968

Separate Variance

Variable	Mean Difference	95.00% Confidence Interval Lower Limit	Upper Limit	t	df	p-Value
WORKERA	3.567	-11.214	18.348	0.533	10.634	0.605
WORKERB						

Pooled Variance

Variable	Mean Difference	95.00% Confidence Interval Lower Limit	Upper Limit	t	df	p-Value
WORKERA	3.567	-10.762	17.896	0.534	14.000	0.602
WORKERB						

https://flux.qa (Feed code: SJ6KGV)

Question 3

To reduce the width of a confidence interval of a population mean. It is necessary to: (*best answer*)

A. Increase sample size

B. Decrease sample size

C. Increase confidence level

D. Increase significance

✓ E. (A or D)

F. (B or C)

Factors Affecting Sample Size

- Factors affecting width of confidence interval:
  - The degree of confidence required, 99, 95, 90% etc.
  - The number of degrees of freedom for small samples.
  - The standard error of the estimate.
- Degrees of Freedom increases and Standard Error diminish as sample size increases.
- For n > 30, the values of the t-Distribution are close enough to the Normal distribution and so we must adjust sample size to further reduce standard error.

Choosing a Sample Size

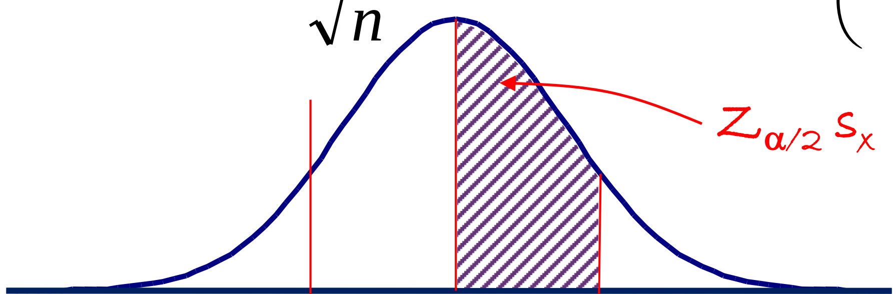
The confidence level for estimating the population mean is

$\mu = \bar{x} \pm Z_{\alpha/2} s_{\bar{x}}$

Thus, $Z_{\alpha/2} s_{\bar{x}}$ is half the width of the confidence interval.

Suppose we want to ensure that the half width is less than a desired value, E. We want $Z_{\alpha/2} s_{\bar{x}} < E$. But $s_{\bar{x}} = s / \sqrt{n}$.

We want a value of n such that $\dfrac{Z_{\alpha/2} s}{\sqrt{n}} \le E$, that is, $n \ge \left(\dfrac{Z_{\alpha/2} s}{E} \right)^2$.



Example 4

- A bank is interested in determining the average disposable income of its customers. From a pilot study they estimate the standard deviation of average disposable income to be $90. How many customers should they sample if they want to obtain an accuracy of $5 at the 95% level?

Solution:

Using a one sided calculation :

$$n \geq \left(\frac{z_{\alpha/2} s}{E} \right)^2$$

At 95% CI,
$z_{\alpha/2} = 1.96$
$s = \$90$
$E = \$5$

$$\geq \left(\frac{1.96 \times 90}{5} \right)^2$$

$$n \geq 1244.6 \ \ or \ \ 1245$$

$$\left(\frac{\bar{x}}{\$5 \quad | \quad \$5} \right)$$

What You Should Know

- You should have some idea of degrees of freedom and be able to read the table for the t distribution.
- You should be able to calculate a confidence interval for the population mean based on a small sample.
- You should be able to calculate the required sample size for a given confidence interval.

Reading/Questions (Selvanathan)

- Reading: Estimation
  - 7th Ed. Sections 10.3, 10.5, 11.1, 11.2.
- Questions: Estimation
  - 7th Ed. Questions and Data 10.40, 10.46, 10.53, 10.56, 10.72, 10.76, 10.77.