MONASH University

Information Technology

FIT1006
Business Information Analysis

Lecture 15
Estimation

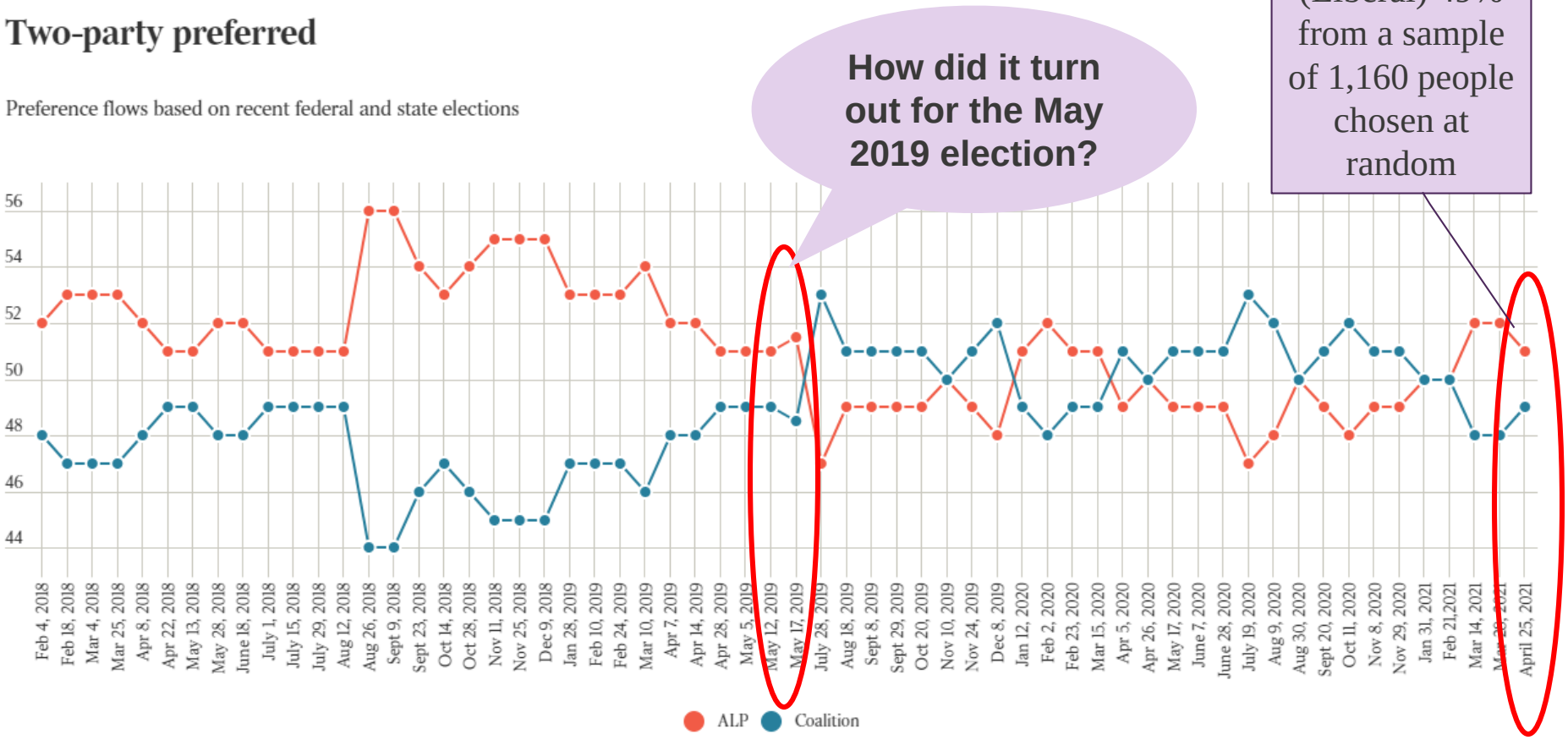
Topics covered:

- Estimation
  - Estimating population parameters using a sample
  - Creating a confidence interval for a population parameter
  - C.I. for the mean and proportion
  - Difference of means and proportions

Motivating Problem

- Would Labor win if we have an election today?
- The Australian Newspoll had the two-party preferred vote at: ALP 51% vs Coalition (Liberal) 49% from a sample of 1,160 people chosen at random (taken on 25 April 2021).
- Hint: Find a 95% CI for the expected Liberal-NP vote.
- Ref: http://www.theaustralian.com.au/national-affairs/newspoll

Who will win the election?



Source: https://www.theaustralian.com.au/nation/newspoll

Australia's conservative party retains power in shocking election result

The Labor Party has lost the "unloseable" election.

By Rachel Withers | Updated May 25, 2019, 2:19pm EDT

SHARE

Prime Minister Scott Morrison and his family celebrate his party's surprise win. | Tracey Nearmy/Getty Images

Labor has lost the "unloseable" election

The Labor Party was widely favored to win this election — so much so that popular gambling website Sportsbet opted to pay out to Labor-backers **two days early**, to the tune of $1.3 million (there was no such luck then or now for the man who placed a record-breaking **$1 million bet on Labor** on rival site Ladbrokes).

And while polls had narrowed in recent weeks, Labor remained clearly in front, with some pundits now blaming the "**shy Tory factor**" (essentially people telling pollsters they plan to vote for one before actually voting for another) for this surprise upset. On-air commenters in Australia are questioning whether they can ever really trust polling again…

Estimates

- Two types of estimates:
- Point Estimates, where we estimate the actual (exact) value of a population parameter.
- Because point estimates are rarely correct it is more usual to define an Interval Estimate. This the range over which we expect the value of the population parameter vary with a given level of confidence.

Estimation, main characters:

Parameter	Population	Sample
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	π	p

$\sigma_{\bar{x}}$ = standard error of the sample mean

σ_{p} = standard error of the sample proportion

The sample values are used to estimate the unknown population parameters, taking into account variability introduced by sampling.

Deriving a confidence interval

In the following slides a confidence interval is derived based on our understanding of the Normal distribution.

To simplify learning the basic technique, this lecture assumes that we know the population variance (*not true in practice*) or that the sample size is large enough that the Central Limit Theorem is true.

In the following lecture the model is adjusted for the case when sample sizes are small and the population variance is unknown.

https://flux.qa (Feed

Question 1:

What is the z value for a 95% CI?

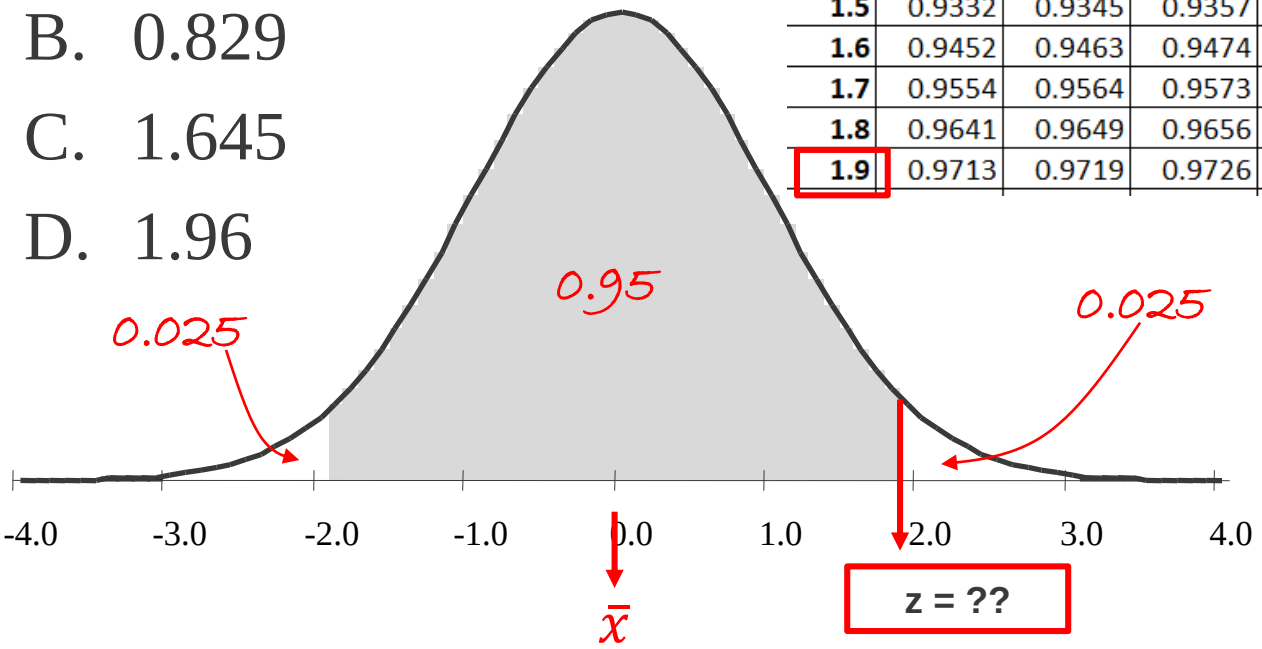
A. 0.510

B. 0.829

C. 1.645

✓ D. 1.96

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750

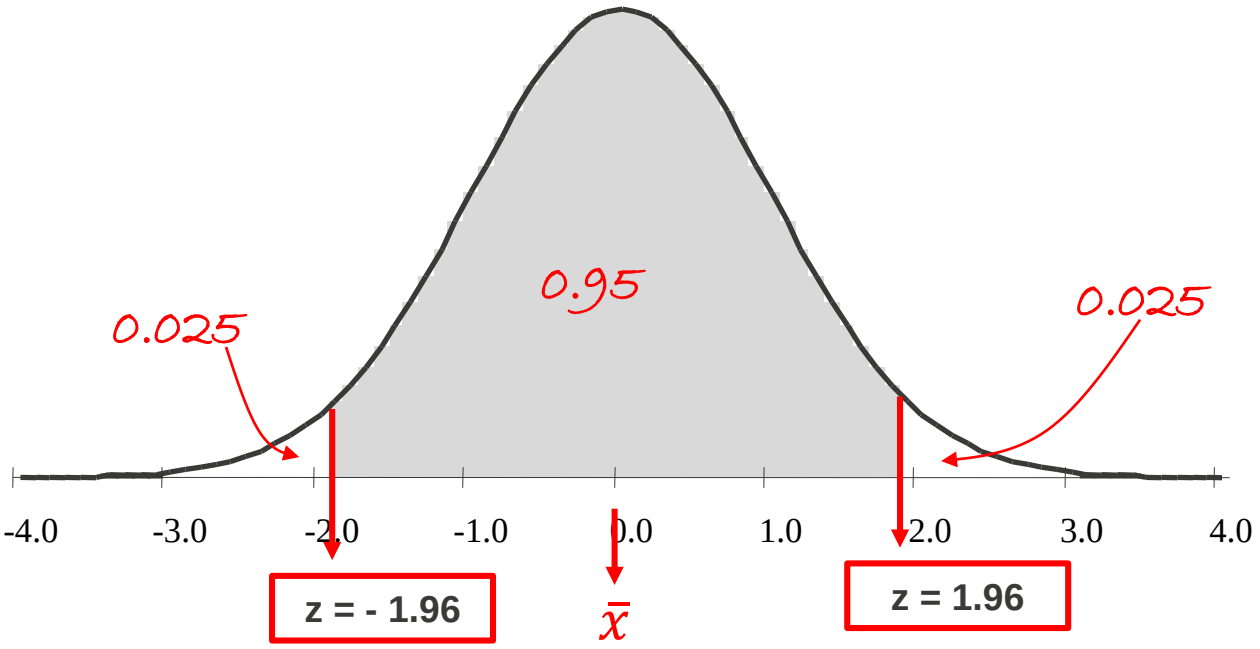


95% Confidence Interval

From the Standard Normal distribution:

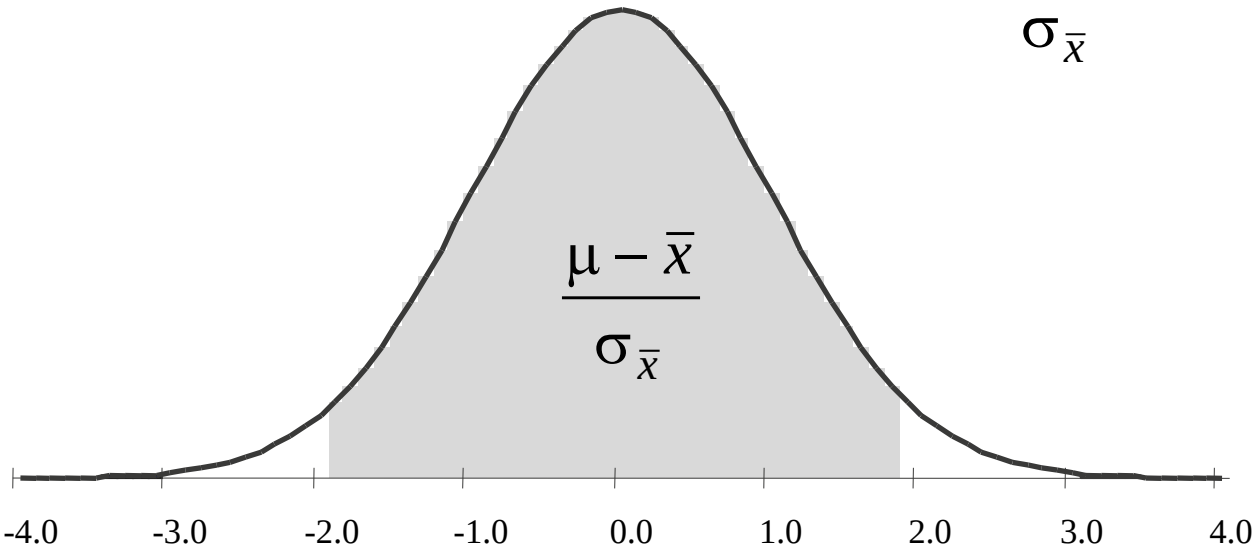
$P(-1.96 < z < 1.96) = 0.95.$

This is a 95% confidence interval for z.



95 % C.I. for μ

- Problem: using a sample taken from a population, estimate population mean, μ, using $\bar{x}$ and σ known.
- Construct a 95% C.I. for the population mean.
- Standardised error of the estimate is $\dfrac{\mu - \bar{x}}{\sigma_{\bar{x}}}$.



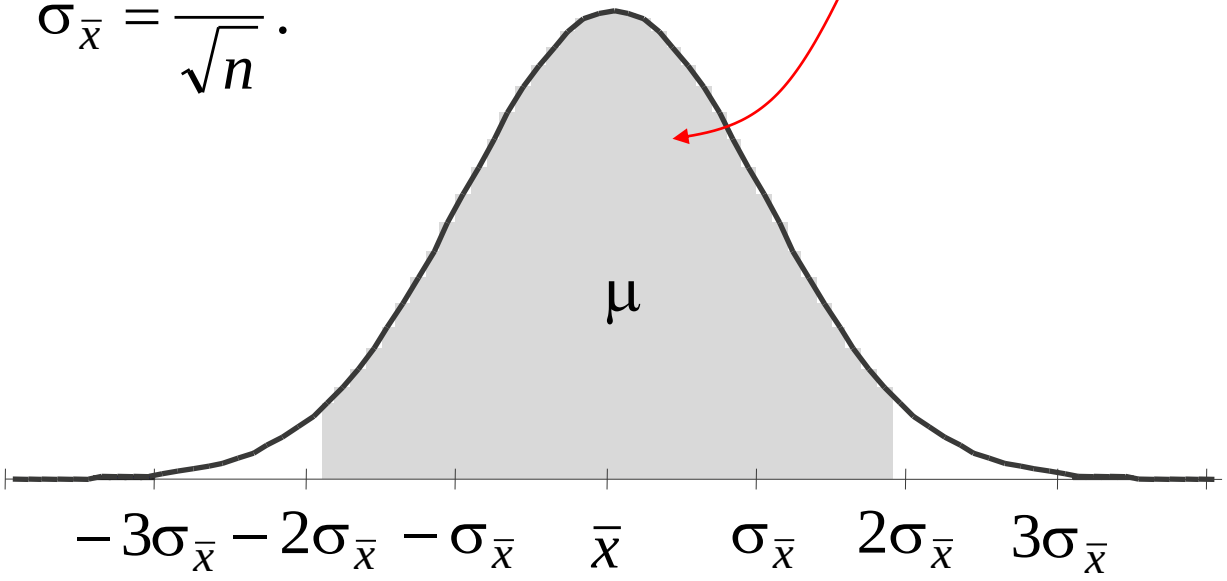
95 % C.I. for μ

- Rescale the distribution, by un-standardising so now:

$$P(\bar{x} - 1.96\sigma_{\bar{x}} < \mu < \bar{x} + 1.96\sigma_{\bar{x}}) = 0.95$$

$$\text{So a 95\% C.I. for } \mu \text{ is : } \mu = \bar{x} \pm 1.96\sigma_{\bar{x}}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$



95 % C.I. for μ

- Algebraic derivation (*you can skip if you want*)

The true value of the population mean is μ which is unknown.

Take a sample, calculate $\bar{x}$ and s (sample mean and st dev).

The standard error (deviation) of $\bar{x}$ is $\sigma_{\bar{x}}$, which is unknown.

The standardised error of the estimate of μ is $\dfrac{\bar{x}-\mu}{\sigma_{\bar{x}}}$,

which has a normal $N(0,1)$ distribution.

95 % C.I. for μ continued

- Algebraic derivation (*you can skip if you want*)

$$P(-1.96 < z < 1.96) = 0.95$$

$$\text{So } P(-1.96 < \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < 1.96) = 0.95, \text{ manipulating}$$

$$P(-1.96\sigma_{\bar{x}} < \bar{x} - \mu < 1.96\sigma_{\bar{x}}) = 0.95$$

$$P(\bar{x} - 1.96\sigma_{\bar{x}} < \mu < \bar{x} + 1.96\sigma_{\bar{x}}) = 0.95$$

$$\text{Thus a 95\% C.I. for } \mu \text{ based on the sample mean } \bar{x} \text{ is :}$$

$$\mu = \bar{x} \pm 1.96\sigma_{\bar{x}}$$

$$\text{Finally, } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ to calculate C.I.}$$

Example

- A sample of 100 students were sampled and their age recorded.

  Summary statistics: $\bar{x}$ = 20.1, σ = 1.2

- Calculate a 95% C.I. for μ, the average age of students at the university.

$$\mu = \bar{x} \pm 1.96\sigma_{\bar{x}} \text{ where } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

$$\mu = 20.1 \pm 1.96\frac{1.2}{\sqrt{100}} = 20.1 \pm 1.96 \times 0.12$$

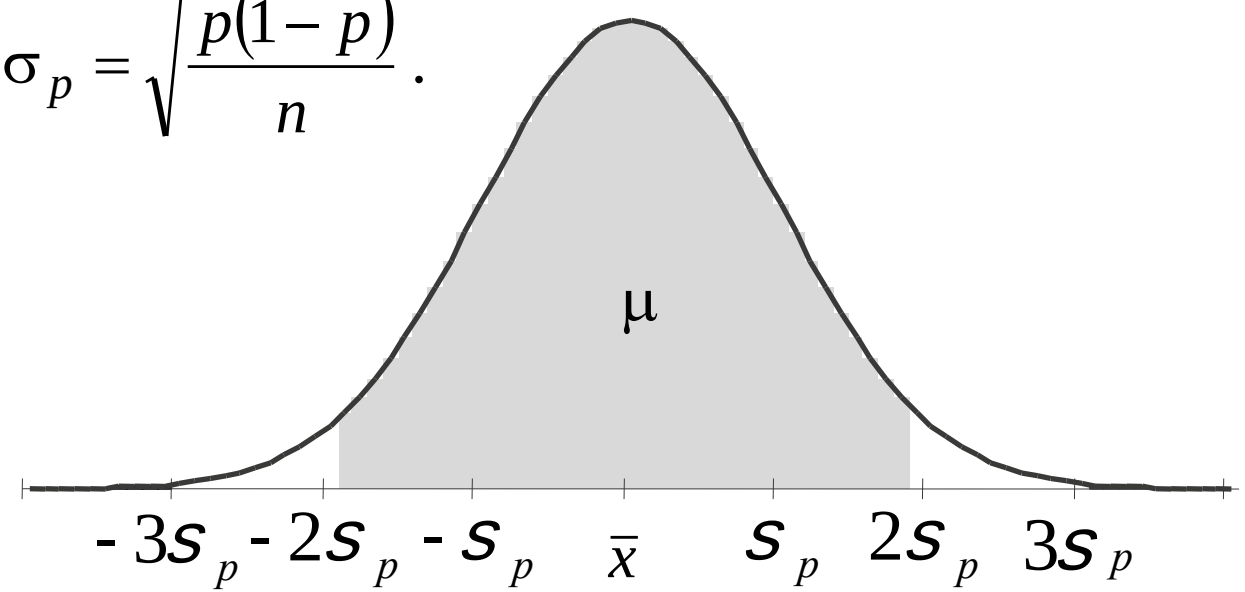
95 % C.I. for π

- In the same way that μ was estimated:

$$P(p - 1.96\sigma_p < \pi < p + 1.96\sigma_p) = 0.95$$

$$\text{So a 95\% C.I. for } \pi \text{ is : } \pi = p \pm 1.96\sigma_p$$

$$\text{Estimate } \sigma_p = \sqrt{\frac{p(1-p)}{n}}.$$



Example

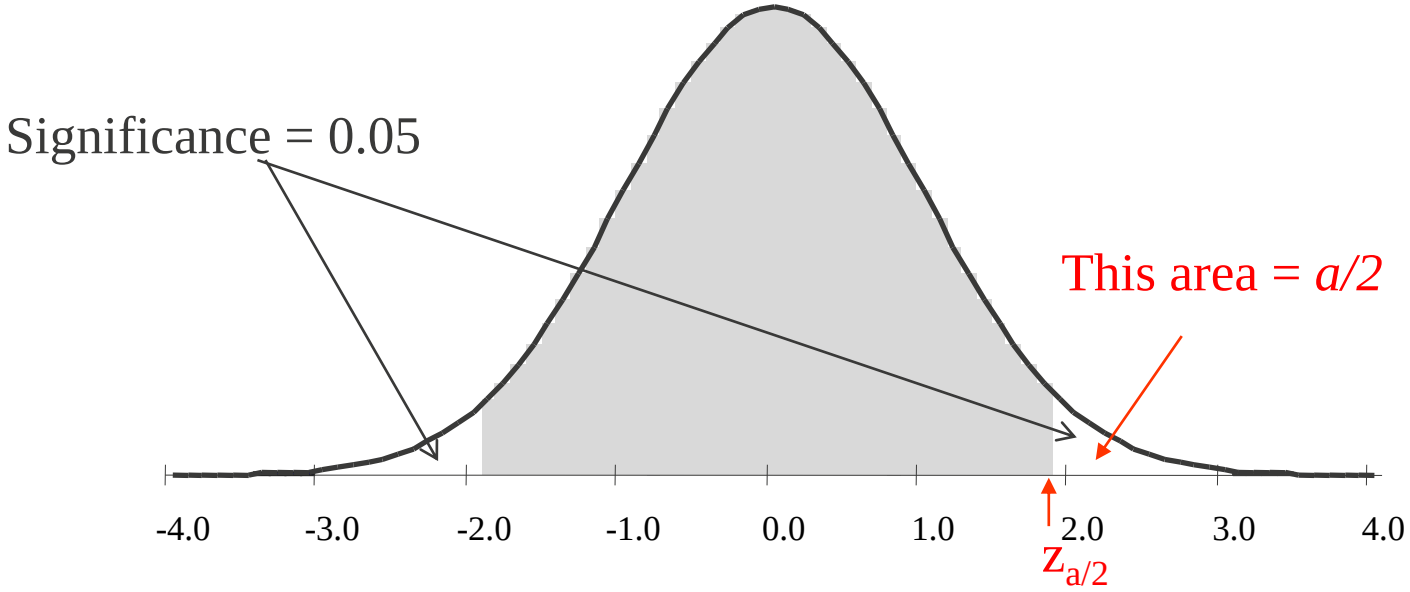
- A sample of 100 students were sampled and 12 left-handed students counted.
- Calculate a 95% C.I. for π, proportion of left-handed students at the university.

$$\pi = p \pm 1.96\sigma_p \text{ where } \sigma_p = \sqrt{\frac{p(1-p)}{n}}.$$

$$\pi = 0.12 \pm 1.96\sqrt{\frac{(0.12)(0.88)}{100}} = 0.12 \pm 1.96 \times 0.032$$

Significance

- Significance is the probability that the C.I. does not contain the population statistic.
- Significance = 1- Confidence. So a 95% confidence has a significance, *a* = 0.05.



90% Confidence Interval

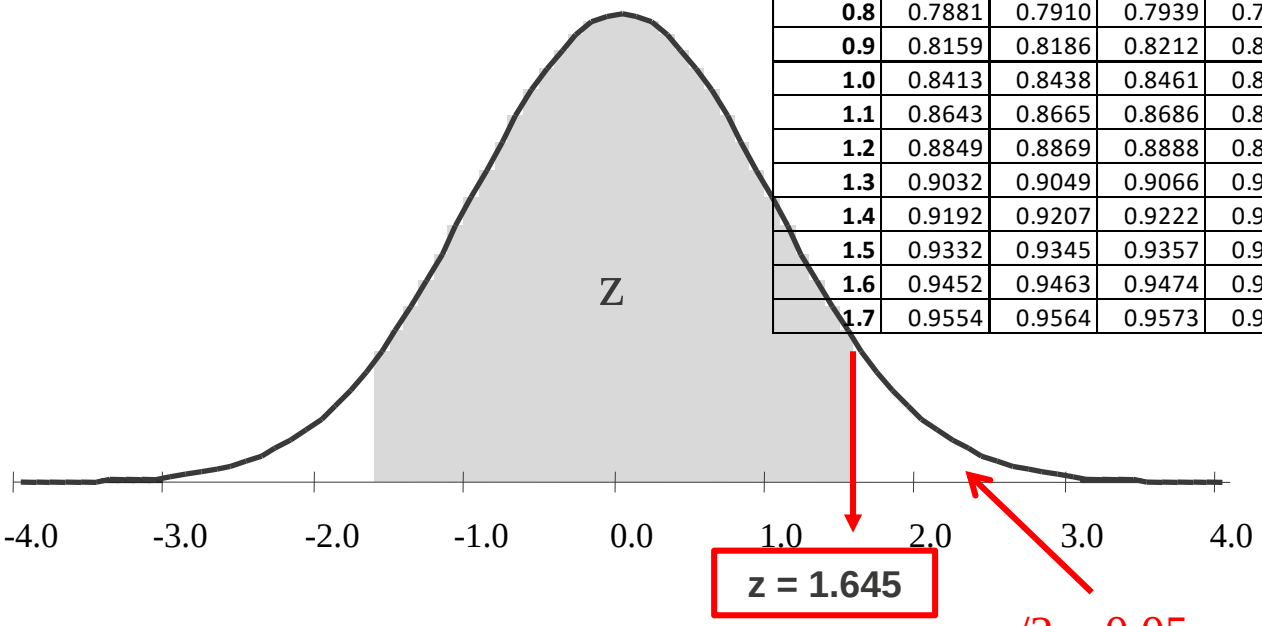
We want P(- ? < z < ?) = 0.90.

P(–1.645 < z < 1.645) = 0.90.

Cumulative Probabilities for the Standard Normal Distribution

Table gives P(Z<z) for Z = N(0,1)

z	0.00	0.01	0.02	0.03	0.04	0.05
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599



A General C.I. for μ and π

- Based on the normal distribution a confidence interval at the *a* significance is.

$$\mu = \bar{x} \pm z_{a/2}\sigma_{\bar{x}} \quad \text{where} \quad \sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- The sample standard deviation *s* is used as an estimate of the population standard deviation, σ.
- The C.I. for p is created the same way.

$$\pi = p \pm z_{a/2}\sigma_p \quad \text{where} \quad \sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

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

Question 2

If the sample size increases, the width of the corresponding confidence interval will:

- ✓ A. decrease,
- B. be unaffected,
- C. increase,
- D. varies, depending on the data.

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

Question 3

If the confidence level increases, the width of the confidence interval will:

A. decrease,
B. be unaffected,
✓ C. increase,
D. varies, depending on the data.

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

Motivating Problem: Group Activity

- Find a 95% CI for the expected Labor vote.
- p = 0.51, n = 1,160.
- The 95% C$\pi = p \pm 1.96\,\sigma_p$; $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$I is:

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

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

Sums and Differences of Variables

- Consider two independent random variables, X and Y:
  - $E(X + Y) = E(X) + E(Y)$
  - $E(X - Y) = E(X) - E(Y)$
  - $Var(X + Y) = Var(X) + Var(Y)$
  - $Var(X - Y) = Var(X) + Var(Y)$
- Rule: always add variances.

The Difference of Means

Consider two populations 1 and 2 with means μ_1 and μ_2.

Let σ_1^2 and σ_2^2 be the population variances.

We take samples of size n_1 and n_2.

Let $\overline{X}_1$ and $\overline{X}_2$ be the sample means, Then :

$$E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$$

$$Var(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2_{\overline{X}_1 - \overline{X}_2}$$

The Difference of Proportions

Consider two populations 1 and 2 with population proportions π_1 and π_2.

We take samples of size n_1 and n_2

Let P_1 and P_2 be the sample proportions

Then :

$$E(P_1 - P_2) = \pi_1 - \pi_2$$

$$Var(P_1 - P_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

Difference of means and proportions

- When finding the confidence interval for the difference of means or proportions use the following to calculate standard error.

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Reading/Questions (Selvanathan)

- Sampling inference and sampling distributions.
  - Reading: 7th Ed. Chapter 9.
  - Questions: 7th Ed. 9.4, 9.12, 9.13, 9.18, 9.24, 9.25

- Estimation
  - Reading: 7th Ed. Chapter 9 + Sections 10.1, 10.2, 10.3.
  - Questions: 7th Ed. 9.4, 9.12, 9.13, 9.18, 9.24, 9.25, 10.1, 10.2, 10.6, 10.9, 10.26, 10.36, 10.58, 10.60, 10.61, 10.71.

Next lecture

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