**Social Science Statistics Module I**

**Gwilym Pryce**

# Lecture 4
# Confidence Intervals for two means and for proportions

**Slides available from *Statistics & SPSS* page of** [www.gpryce.com](www.gpryce.com)

## Expectations & Support:

1. **Independent learning:**
   - this is a PG course and a degree of independent learning is assumed.
   - do the reading, attend labs, review the lectures, make use of the computer labs/online help in your own time.
2. **Lab Overview & Feedback:**
   - Please feedback to the tutors & Class Reps how you think that is going, how it could be improved.
   - Tutors and Class Reps will then report back to me how things are going each week.
3. **Talk to tutors if you are struggling:**
   - Let the tutors know if you are struggling (assuming you have done the reading, attended labs etc.)
   - Tutors cannot guarantee extra support, but it might be possible to arrange extra tutorials etc.

4. **Departmental Support:**
   - Struggling students should enquire whether their own dept has support to offer.
   - All the grad school courses are only intended to constitute a generic training component;
   - Individual depts & supervisors should supplement with additional training & support as necessary.
5. **Support from Maths Advisor Shazia Ahmed, University's Maths Adviser:**
   - If you have gone through steps 1 to 4, Shazia has agreed to run one-on-one sessions with students that are struggling with particular mathematical or statistical concepts (though she has made it clear that she cannot advise on SPSS problems, nor will she do the assignment for you).
   - Students who have particular problems in this regard can contact her directly: **Shazia Ahmed**, Maths Adviser, Student Learning Service, McMillan Reading Room, **Tel: 330 5631** Fax: 330 8063
6. **Tutor of Last Resort:**
   - Students who have gone through steps 1 to 5 above, and who still feel they are not receiving enough support, can email me directly
   - I will try to arrange individual or small group meetings for people who have tried all other avenues.
     - You will need to demonstrate that you have gone through steps 1 to 5.

# Aims & Objectives

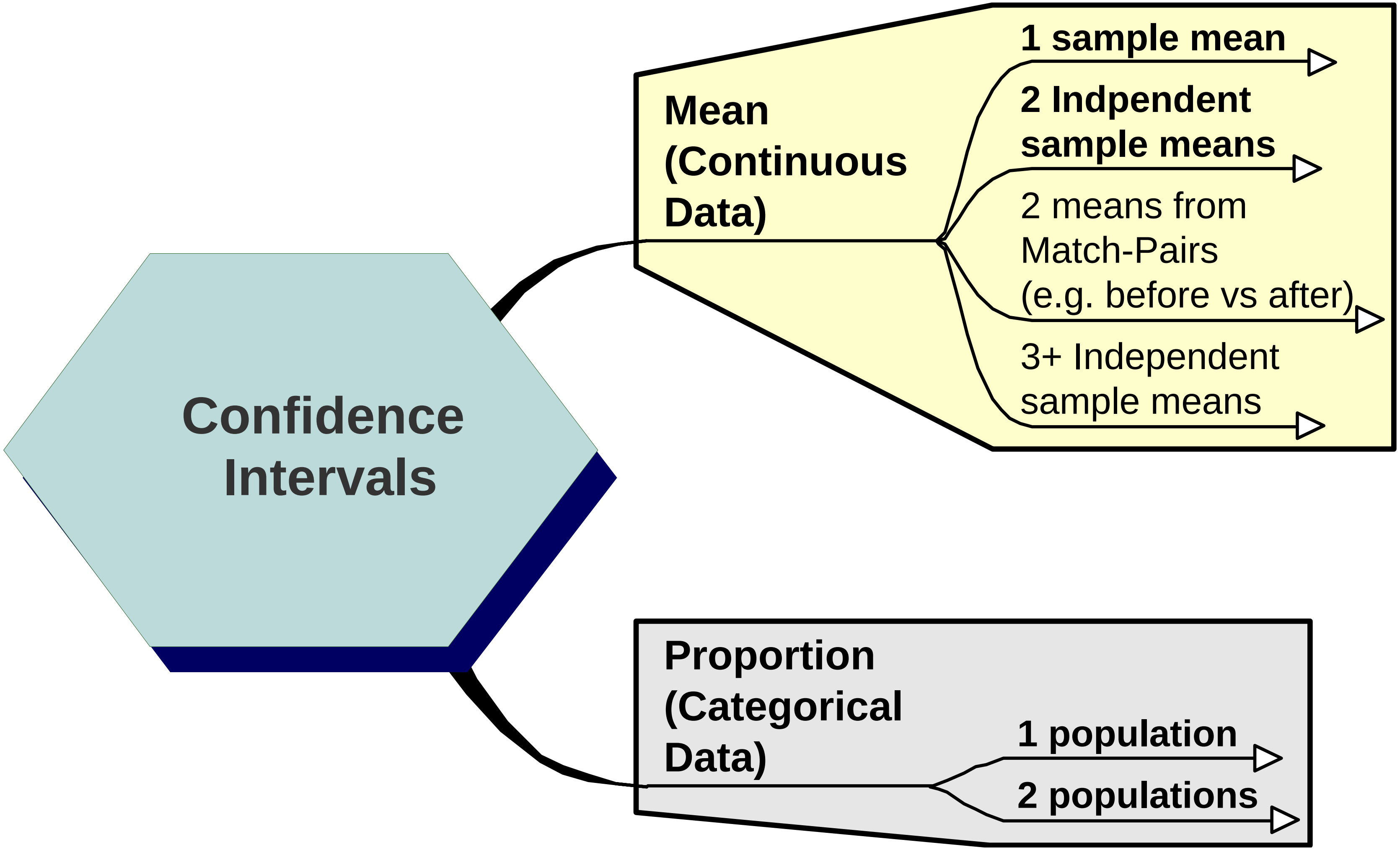
- Aim
  - To consider the appropriate confidence interval procedures for a range of situations.
- Objectives
  - By the end of this session, students should be able to:
    - Run confidence intervals on 2 means;
    - Run confidence intervals on proportions.

- SPSS can produce confidence intervals for the mean when you have the original data
    - go to Analyze, Descriptive Statistics, Explore
- But its not so useful when you have only summary information.
    - I.e. when you are only given the mean, s.d. & n

… or when you want a CI for something other than the mean of one population

- E.g. if you want a CI for the difference between 2 means;
- E.g. if you want a CI for a proportion (particularly if you want to use the more robust Wilson method)

- In situations like these you either need to be familiar with the appropriate formulas or you need to know how to use the custom macros… this lecture introduces both.



Confidence Intervals

**Mean (Continuous Data)**

- **1 sample mean**
- **2 Indpendent sample means**
- 2 means from Match-Pairs (e.g. before vs after)
- 3+ Independent sample means

**Proportion (Categorical Data)**

- **1 population**
- **2 populations**

# Summary of all SSS1 macros:

| Confidence Intervals (CI) | | Hypothesis tests | |
|---|---|---|---|
| Macro command | Definition | Macro Command | Definition |
| **CI_L1M** | Large sample CI for one mean | **H_L1M** | Large sample significance test on one mean |
| **CI_S1M** | Small sample CI for one mean | **H_S1M** | Small sample significance test on one mean |
| **CI_S2MP** | Small independent samples CI for difference between 2 means (pooled variance) | **H_S2MP** | Small independent samples significance test for equality of 2 means (pooled variance) |
| **CI_S2MD** | Small independent samples CI for difference between 2 means (different variances) | **H_S2MD** | Small independent samples significance test for equality of 2 means (different variances) |
| **CI_L1P** | Large sample CI for one proportion (presents output for both Traditional and Wilson methods of calculation) | **H_L1P** | Large sample significance test on one proportion |
| **CI_L2P** | Large sample CI for comparing two proportions (presents output for both Traditional and Wilson methods of calculation) | **H_L2P** | Large samples significance test on two proportions |
| | | **H_S2VF** | Simple small sample F-test on equality of two variances (see also Levene's test in the SPSS help menu for more sophisticated test of homogenous variances). |
| **N_L1M** | Sample size for desired margin or error for the mean | | |

# Plan

- 1. CI for two independent means
  - 1.1 Pooled Variances
  - 1.2 Different Variances
- 2. CI for two paired means
- 3. CI for one proportion
- 4. Sample size determination

# 1. CI for two independent means

- Sometimes we want to compare the means of two independent populations.
  - E.g. sample mean height from a population of girls of a particular age vs sample mean height from a population of boys.
    - is the difference between the means a freak result arising from sampling variation?
    - or does it reflect true differences in height between the population of boys and the population of girls?
- One way of tackling this quandary is to estimate the confidence interval for the difference in the two means.
    - This will tell us the range of likely values for that difference the in the whole population.

- The following calculations assumes that the two populations (and hence the two samples) are ***<u>independent</u>***:
    - i.e. someone in the first population cannot occur in the second.
    - This is distinct from situations where the researcher observes the same person before and after a treatment (for such experiments we use a *Paired Samples Confidence Interval*).
- There are two formulas for calculating the confidence interval for comparing two population means:
    - one assumes equal (or *homogeneous*) variances across the two populations,
    - the other assumes unequal (or *heterogeneous*) variances across the two populations.
  - – Later on in the course we shall look at hypothesis tests that help us decide on whether or not the variances are the same (e.g. Levene’s test).

## 1.1 Pooled variance (see M&M p.537)

- The confidence interval for the difference between two population means is given by:

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where,

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

## Alternatively, we can use the macro command:

- CI_S2Mp
  - Small Independent Samples CI
  - for difference between 2 means
  - (pooled variance M&M p.538)
- The syntax for the command is entered as follows:

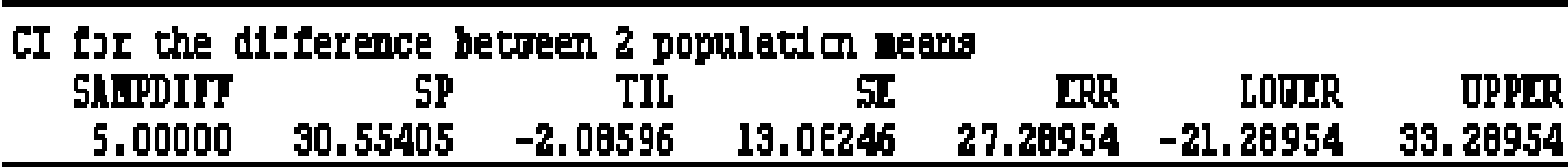
**CI_S2Mp** n1=(?)  n2=(?)  x_bar1=(?)  x_bar2=(?)  s1=(?)  s2=(?)  c=(?).

**E.g. mean height of girls in our sample of 10 = 100 cm (s.d. = 30cm), and the mean height of 12 boys is 94cm (sd = 31cm). All are the same age.**

- To find 95% confidence interval for the difference in population means we would enter the following:

  **CI_S2Mp n1=(10) n2=(12) x_bar1=(100) x_bar2=(94) s1=(30) s2=(31) c=(.95).**

- which results in a v. wide interval:



```
CI for the difference between 2 population means
   SAMPDIFF          SP        TVAL          SE         ERR       LOWER       UPPER
    6.00000    30.55405    -2.08596    13.06246    27.28954   -21.28954    33.28954
```

i.e.

$$\mu_1 - \mu_2 = 6 \pm 27.28954$$

# 1.2 Different Variance (see M&M p.532)

- The confidence interval for the difference between two population means is given by:

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}$$

where, $df$ = $\min[n_1\text{-}1, n_2\text{-}1]$

## Alternatively, we can use the macro command:

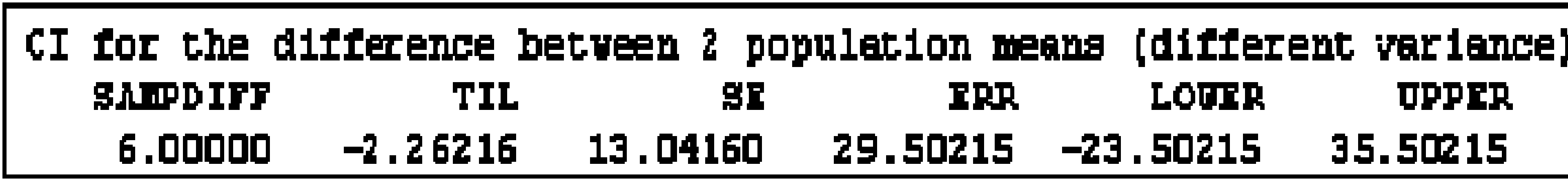
- CI_S2Md
  - Small Independent Samples CI for differences between 2 means
  - different variances (M&M p.532).
  - Arguments[1] are entered in the same way as for CI_S2Mp:

**CI_S2Md** n1=(?) n2=(?) x_bar1=(?) x_bar2=(?) s1=(?) s2=(?) c=(?).

[1] Argument = “*Independent variable determining the value of function*” (OED)

## Applying CI_S2Md to our girl/boy heights “difference in means” example:

CI_S2Md n1=(10) n2=(12) x_bar1=(100) x_bar2=(94) s1=(30) s2=(31) c=(.95).



```
CI for the difference between 2 population means (different variance)
   SAMPDIFF          TIL          SE          ERR       LOWER       UPPER
    6.00000     -2.26216    13.04160     29.50215   -23.50215    35.50215
```

i.e.

$$\mu_1 - \mu_2 = 6 \pm 29.50215$$

# 2. CI for two paired means (see M&M p.501-503)

- Suppose we have two sets of observations on the same individuals:
  - as in a "before and after" trial,
  - our two samples are said to be "paired"
- We can:
  - compute the mean & s.d. of the difference between the two sets of results
    - e.g. average "improvement" & s.d of "improvement"
  - apply the one sample confidence interval for the mean procedure.
    - If large sample use: CI_L1M n=(?) x_bar=(?) s=(?) c=(?).
    - If small sample use: CI_S1M n=(?) x_bar=(?) s=(?) c=(?).

## e.g. Mean Quality of Life score for 100 amputees: sample average improvement since amputation = 5.3, with s.d. of improvement = 4.2. What is the 99% confidence interval for the population mean improvement?

- **CI_S1M** n=(100) x_bar=(5.3) s=(4.2) c=(0.99).

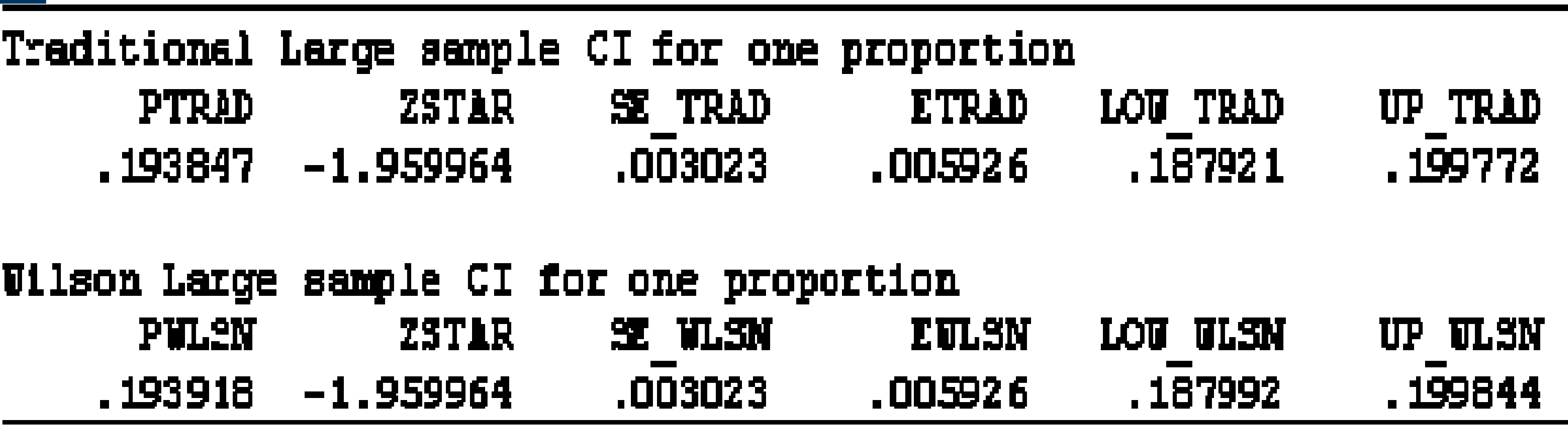
```
Small sample confidence interval for the population mean
    N          X_BAR        TIL          SE         ERR       LOWER      UPPER
100.00000    5.30000   -2.62641     .42000    1.10309    4.19691    6.40309
```

- The experiment (!) has produced a fairly narrow interval for the improvement score, even at the 99% confidence level
  - NB lower bound is positive, so amputation likely to beneficial on average in population.

## 3. CI for one proportion

e.g. Suppose 3,314 out of a sample of 17,096 students reveal that they are binge drinkers (M&M p. 572ff), find the 95% confidence interval for the proportion of binge drinkers.

**CI_L1P** n=(17096) x=(3314) c=(.95).



```
Traditional Large sample CI for one proportion
     PTRAD       ZSTAR     SE_TRAD       ETRAD    LOW_TRAD     UP_TRAD
   .193847   -1.959964     .003023     .005926     .187921     .199772

Wilson Large sample CI for one proportion
     PWLSN       ZSTAR     SE_WLSN       EWLSN    LOW_WLSN     UP_WLSN
   .193918   -1.959964     .003023     .005926     .187992     .199844
```

- As it happens, there is very little difference between the Traditional and Wilson methods in this particular example.
  - Using the latter method, we estimate with 95% confidence that between 18.799% and 19.984% of college students are frequent binge drinkers.

## 4. Sample size determination

- Suppose you want to estimate the average weight of 5 year olds with a margin of error ***e*** of 2 pounds when you apply a 95% confidence interval.
- Sample size necessary for estimating the population mean with the desired accuracy will be given by:

$$\boxed{n = \frac{z^{*2}\sigma^2}{e^2}}$$

- Sample size necessary for estimating the population proportion with a desired level of accuracy would be:

$$\boxed{n = \frac{z^{*2}\pi^*(1-\pi^*)}{e^2}}$$

Where $\pi^*$ is your guesstimate of the population proportion

## Example:

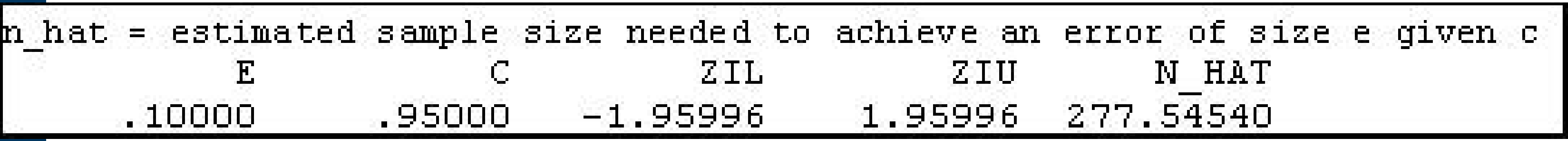
- For your PhD, you want to estimate the mean hourly wage rate of unskilled labour in Easterhouse, Glasgow. You would like your estimate to lie within ±£0.10 at the 95% confidence level. A 1987 study (large sample size) by the Department of Employment resulted in a standard deviation of £0.85. Using this as an approximation for $\sigma$, compute the necessary sample size to arrive at the desired level of accuracy.

- The maximum allowable error *e* = 0.1
- The z* value for 95% confidence interval = 1.96
- Our best estimate of the population s.d. $\sigma$ = 0.85
- Entering these values in the formula gives:
- round up to 278 to ensure our sample size is large enough.

$$\boxed{n = \frac{(1.96)^2(.85)^2}{(0.1)^2} = 277.556}$$

# Using the N1_L1M syntax:

- **N1_L1M**
  - Sample size for desired margin or error for the mean (M&M p.425).
- N1_L1M e=(0.1) c=(0.95) s=(0.85) .



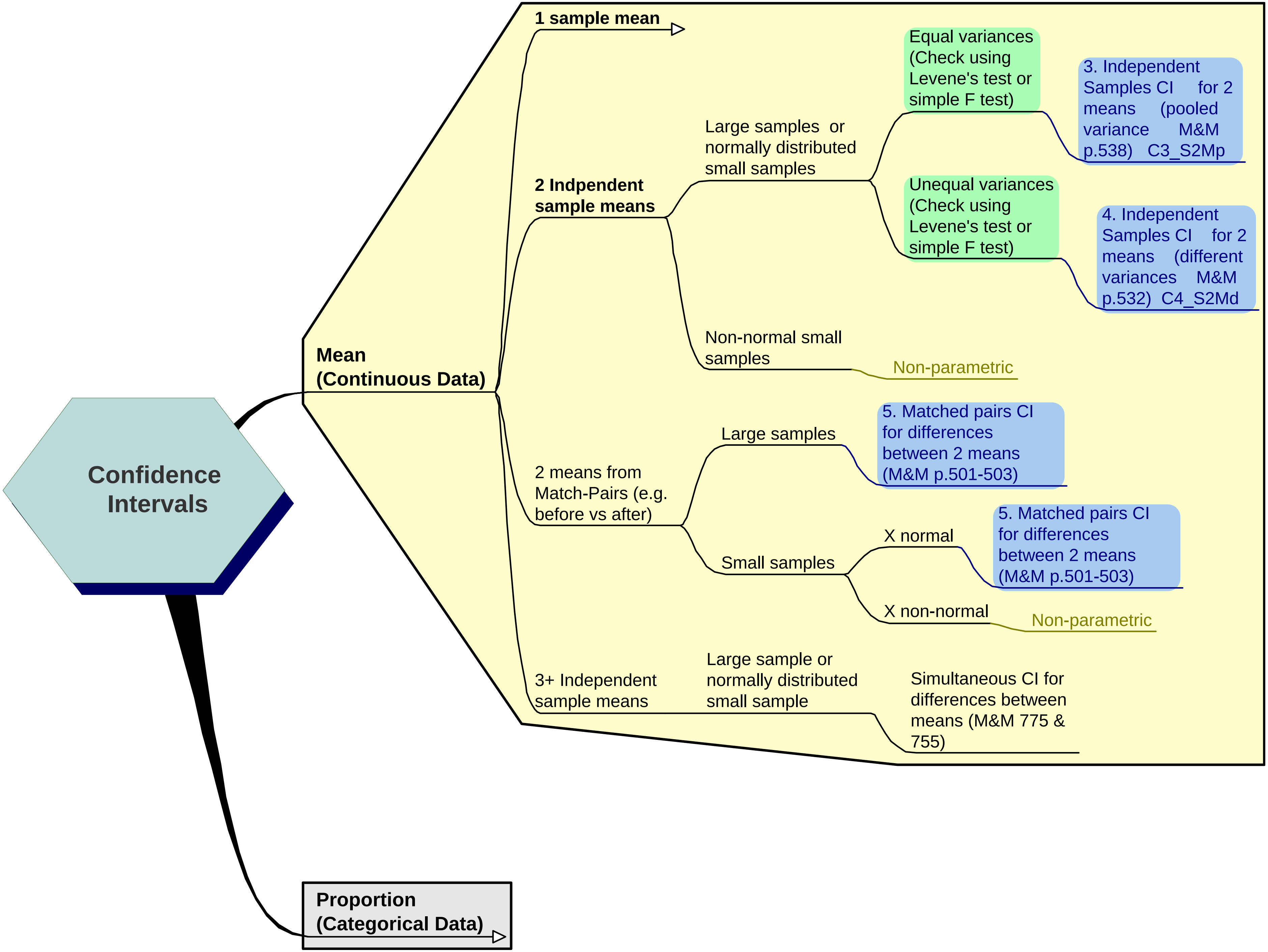
```
n_hat = estimated sample size needed to achieve an error of size e given c
          E           C         ZIL         ZIU       N_HAT
     .10000      .95000    -1.95996     1.95996   277.54540
```

# Summary:

- 1. CI for two independent means
- 2. CI for two paired means
- 3. CI for one proportion
- 4.  Sample size determination

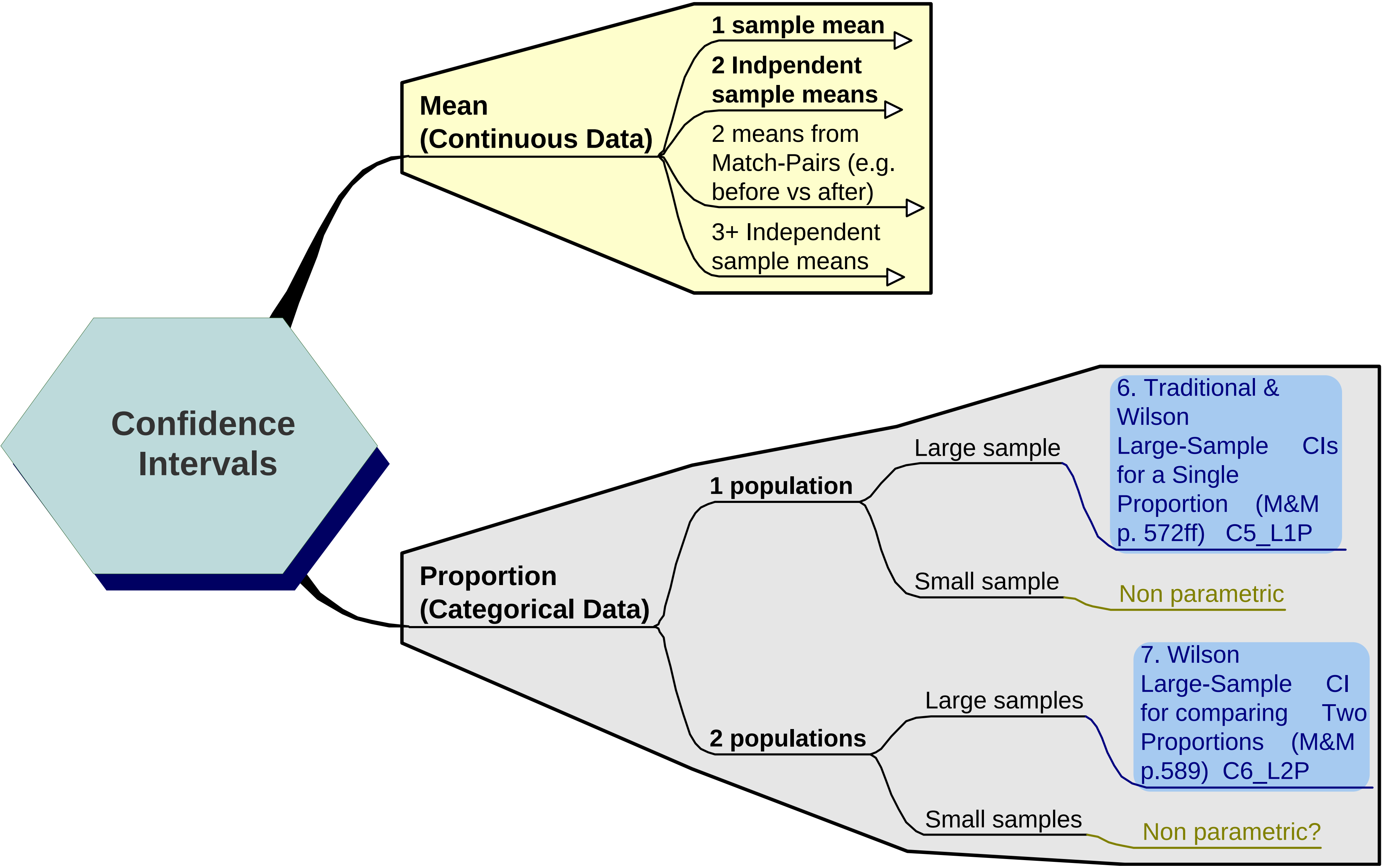
# Reading:

- Chapter 4 of Pryce (2005) Inference and Statistics in SPSS
- M&M 4th Ed.
  - section 6.3 and exercises for 6.3
  - Sections 6.1 (p. 415-429); 7.1 and 7.2. Chapter 8.



Confidence Intervals

- **Mean (Continuous Data)**
  - **1 sample mean**
  - **2 Indpendent sample means**
    - Large samples or normally distributed small samples
      - Equal variances (Check using Levene's test or simple F test)
        - 3. Independent Samples CI for 2 means (pooled variance M&M p.538) C3_S2Mp
      - Unequal variances (Check using Levene's test or simple F test)
        - 4. Independent Samples CI for 2 means (different variances M&M p.532) C4_S2Md
    - Non-normal small samples
      - Non-parametric
  - 2 means from Match-Pairs (e.g. before vs after)
    - Large samples
      - 5. Matched pairs CI for differences between 2 means (M&M p.501-503)
    - Small samples
      - X normal
        - 5. Matched pairs CI for differences between 2 means (M&M p.501-503)
      - X non-normal
        - Non-parametric
  - 3+ Independent sample means
    - Large sample or normally distributed small sample
      - Simultaneous CI for differences between means (M&M 775 & 755)
- **Proportion (Categorical Data)**



Confidence Intervals

Mean (Continuous Data)

**1 sample mean**

**2 Indpendent sample means**

2 means from Match-Pairs (e.g. before vs after)

3+ Independent sample means

**Proportion (Categorical Data)**

**1 population**

Large sample

6. Traditional & Wilson Large-Sample CIs for a Single Proportion (M&M p. 572ff) C5_L1P

Small sample

Non parametric

**2 populations**

Large samples

7. Wilson Large-Sample CI for comparing Two Proportions (M&M p.589) C6_L2P

Small samples

Non parametric?

# FAQ on SE & CIs:

- **Q1/ Is the "standard error" the same as the "margin of error"?**
- A/ No.  The "Standard Error" has a very precise statistical meaning:
  - SE is "the standard deviation of the sampling distribution of the mean (or proportion)".
  - That is, it is the name we give to the amount sample means will vary from sample to sample.
    - If sample means don't vary much from sample to sample (i.e. the "sampling distribution of the mean" is fairly peaked), then the standard deviation of means (i.e. the "SE of the mean") will be small.
    - If, on the other hand, sample means do vary considerably from sample to sample (i.e. the "sampling distribution of the mean" is well spread -- fairly flat) then we will find that the SE of the mean will be large.

- Note that when we refer to a "sampling distribution" we refer to the distribution of means from repeated samples OF THE SAME SIZE.
    - I.e. each sample we take has the same number of observations.
    - In other words, there will be a different sampling distribution for each sample size.
    - Hence, for each sampling distribution there will be a different standard deviation ("standard error").
- As you might expect, the larger the sample size, the more *peaked* the sampling distribution, and the smaller the standard error.
    - The sampling distribution we are interested in for a particular problem will of course be the one defined by the size of the sample we are dealing with at the time.

- "Margin of error", on the other hand, is a much looser term.
    - It is usually how much our estimate (e.g. of the population mean) differs from the true value.
    - If we want our margin of error to be small, we have to use a large sample.
- The two concepts are not unrelated, however:
    - How close our sample estimate of the population mean will be to its true value will be determined by how much variation there is in sample means between samples.
    - So if the SE is small, the more accurate will be our estimate, and the smaller our margin of error will be.

- **Q2/ What scale is the SE measured in? Is it possible to read the standard error as an individual figure by itself e.g 5.3 without having the sample details? Compared to 1.9, which one would you say is a higher standard error?**
- Suppose we are looking at the height of girls and boys in cm.
    - Let's also assume that the samples we have for boys and girls are the same size.
    - If for boys, the SE = 5.3cm, then we are saying that, on average, sample means vary by 5.3cm from the true population mean (which happens to equal the mean of all sample means).
    - If for girls, on the other hand, sample means tend to vary only by 1.9 from the population mean, then we know that the sampling distribution of mean height is much flatter for boys than for girls.

- I.e. Mean height varies from sample to sample a lot more for boys than for girls.
  - This suggests that, for a given sample size, we shall be able to make a more accurate prediction of the population mean height of girls than of the population mean height of boys.

- **Q3/ It bothers me that an “error” can be inaccurate given a small sample size. Errors ARE inaccurate, how can it NOT be inaccurate? Only in statistics, right?**

- The problem is that we rarely know what the standard error of the mean is.
- Think for a moment why this might be.
  - If the SE of the mean is the "standard deviation of means across repeated samples" then you'd think that the only way we can calculate it is by taking repeated samples.
  - Strictly speaking, the only way to arrive at the true value of the SE is in fact to take an infinite number of samples!
  - So even if we could afford to take 100 samples, the standard deviation of all the means we have calculated would still only be an ESTIMATE of the true value of the standard error.

- In practice we usually only have enough time and money to take a single sample.
  - Our dilemma is that we somehow have to estimate from a single sample what the variation might be of means from repeated samples!
- All is not lost, however, because it turns out that the standard deviation of our single sample is related to the SE of the mean.
    - That is, the variation of the actual values of our variable within a particular sample is related to the variation of the mean of that variable from sample to sample.

- E.g. Average grade received SSS1.
  - If you had access to data on all previous classes, you could calculate the average grade for each class.
  - The sampling distribution of the mean would simply be the histogram of the means you have calculated for each class.
  - Now, what we are saying is that if you don't in fact have access to data on all previous classes, but only the current class, then the variation in marks amongst your colleagues in your year (the standard deviation of individual grades) will tell you something about how much the average grade is likely to vary from year to year (the standard error of the mean).
  - It won't be a perfect predictor but it’s the best we can do.

- What we do know is that the amount by which the average grade varies from year to year will depend on the size of class in each year (which we assume constant across all years).
  - If the size of the class in each year is 500, then the average grade will be pretty similar across years. If the class size in each year is only 10, then the average grade will vary considerably from year to year.
  - So, to account for the effect of sample size, our estimate of the standard error of the mean would be equal to the standard deviation of grades amongst your colleagues, divided by the *square root* of the number of students in the class.
  - For example, if the standard deviation of grades is 15 marks, and the size of the class is 50, then your estimate of the standard error of the mean would be 15/7.07 = 2.12. That is, you reckon that the mean grade in each year typically varies by 2.12 marks or so around the mean of all grades from all years (the "population mean").

- This statement is still rather vague, however, since we have said "typically".
    - It would be nice if we could give a probability to this.
    - That is, we'd like to say something like, that we are 95% sure that the average grade across all years lies between a and b.
  - But how can we work out where 95% of sample means lie?
    - To do this, we make use of the fact that the sampling distribution is normal (Central Limit Theorem) and that this means we can translate our knowledge of the sampling distribution (i.e. our estimate of how flat it is, the SE, which we have estimated to be 2.12), into finding the appropriate "margin of error".
    - This margin of error is found by multiplying our estimated standard error by the z score associated with the central 95% of z values, which turns out to be 1.96. So, 1.96 multiplied by 2.12, gives you a margin of error of 4.15 marks.

- We haven't said yet what the average grade in your year is. Lets say its 68 (you're a bright bunch!).
    - Therefore, we can be 95% sure that the average grade across all years is 68 plus or minus a margin of error of 2.12. I.e. we can be 95% sure that the population mean grade lies between 66 and 70, or thereabouts.
- The important assumption here, of course, is that the current class of students constitutes a simple random sample of all students in all years.
    - This would not be the case if, as some claim, students are gradually getting more intelligent (due to improvements in diet, pre-school education, and, apparently, computer games and TV!).