- Sampling replacement
  - When a case is drawn it is put back again
  - Eligible for future draws

- Sampling without replacement
  - When a case is drawn it is not put back again
  - Not eligible for future draws

16. Types of sampling

(1) Simple random sampling
- Every element has an equal chance of being selected
- Standard to which all other sampling techniques are compared
- Give each person/observation a unique id
- Use random number tables or computer generated random numbers.
- Basis for all calculations in CI's

(2) Symmetric Sampling
- Take every $k^{th}$ element after a random start eg N=125
  - K=125/5=25
  - Suppose it was 7 from 125, K=125/7=17.8 — Call it 17
- Generate a random start between 1 and 125
- Treat list as circular
- Randomly sort the list

17. Advantages and Disadvantages on Symmetric Sampling
- Advantages
  - Easy to Execute and Understand (budget/time constraints)
  - Clustered Selection Eliminated
    - Where SRS samples are uncommonly close together in a population
    - Simple Random Sample (SRS) must increase the number of samples or running more than one survey
- Disadvantages
  - Assumes Size of Population Can be Determined
  - Need for Natural Degree of Randomness
  - Great Risk of Data Manipulation

(3) Stratified Sampling
- Might want to ensure there is representation from predefined groups
  - eg age groups in political polling
- Predefined number of samples from each group
- Can then use Simple Random Sampling or Symmetric Sampling

18. Central Limit Theorem:

the sample distribution of the mean approximates a normal distribution as sample size increases, regardless of the distribution of the underlying population

19. Standard Error
- The standard error of a statistic is a measure of its accuracy
- The standard error of the sample is equal to $\frac{\sigma}{\sqrt{n}}$
  - n = sample size
  - $\sigma$ = population standard deviation
- We will use SE to indicate our estimate of standard error of the sample means

- We can calculate this estimate by using the sample standard deviation $SE = \frac{SD}{\sqrt{n}}$

20. Confidence Intervals
- A confidence interval is another estimate
- A 95% level of confidence means we expect that 95% of confidence intervals calculated from these random samples will contain the true population mean
- In other words, if you conducted your study your study 100 times you would 100 different confidence intervals
- We would expect that 95 out of those 100 confidence intervals will contain the true population mean

21. How do we tell if our data is normal
- Skewness (Symmetry)
  - A positively skewed distribution has scores clustered to the left, with the tail extending to right
  - A negatively skewed distribution has scores clustered to the right
  - Skewness is 0 in a normal distribution, so the farther away from 0, the more non-normal the distribution
- Kurtosis: how heavy the tails of the distribution are
  - Kurtosis in a normal distribution is close to 3
  - Positive excess kurtosis indicate too many observations are close to the mean
  - Negative excess kurtosis indicates too many observations are in the tails
  - Significance ($p < 0.05$) indicates data is non-normal

22. Basic Idea of Hypothesis Testing
- Have Data
- Construct a Hypothesis: A statement concerning the parameters or form of the probability distribution for a designated population or populations
- Decide on a level of significance
- Calculate a test statistic
  - A statistic is a quantity derived (得出) from sample
- Evaluate our evidence against our hypothesis based on the test statistic and the level of significance

23. Significance Level
- The significance level is the probability of rejecting the Null Hypothesis when the Null Hypothesis is in fact true (Type I error)
- The significance level is a threshold
- We choose this ourselves before we analyse our data
- We will compare p-values to this pre-defined significance level

- Significance is a yes/no question based on the significance level. A p-value can therefore not be more significant than another one, since they are either smaller or larger than the chosen significance level

24. Test Statistics

Depends on:
- Form of the null hypothesis
- Type of data
- Story behind data

25. Relationship between significance level and CI
- A 95% level of confidence for the mean: we expect that 95% of confidence intervals calculated from these random samples will contain the true population mean

- In 5% of samples, the confidence level will not contain the true population mean
- The significance level is the probability of rejecting the Null Hypothesis is in fact true
  - In 5%, we will say that the true population mean does not equal the location, even when it does
- If the p value is less than your significance (alpha) level, the hypothesis test is statistically significant
- If the confidence interval does not contain the null hypothesis value, the results are statistically significant
- If the p value is less than alpha, the confidence interval will not contain the null hypothesis value

26. Type 1 Error
- Type 1 Error: the probability of rejecting the Null Hypothesis when the Null Hypothesis is in fact true
- In a hypothesis test such as we have looked at so far we know this already, because we chose it

27. Type 2 Error
- Type 2 Error: the probability of failing to reject the null hypothesis when the null hypothesis is in fact false

28. Confidence level
- the probability of failing to reject the null hypothesis when the null hypothesis is in fact true

29. Power
- the probability of rejecting the null hypothesis when the null Hypothesis is in fact false