

# UL Winter School Quantitative Stream

## Unit B0: Refresher Course

Brendan Halpin

Department of Sociology, University of Limerick – [brendan.halpin@ul.ie](mailto:brendan.halpin@ul.ie)

January 8-9, 2018

# What is Quantitative Method?

- Distinct in using number
- Large amounts of relatively shallow data
- Data may be shallow, but is strictly comparable: compare and contrast
- Tends to look to explicit causal explanations

# What is Quantitative Method?

- Clearly defined meanings allocated numerical representations
- Thus easily manipulated
- Descriptive statistics and graphics
- Analytical statistics and graphics

# Causal relationships from empirical data?

- QM often concerned with causal accounts, “low level” theories
- The experiment is probably the strongest way of arguing from data
  - “Experimental control” means everything is the same except the input of interest
  - A strong inference that differences in the result are caused by the difference in the input
- Experiments are rarely possible in social science: therefore we use “observational” data, and compare and contrast (“statistical control”)

# Numbers as information

- Designing – asking structured questions so answers can be mapped onto numbers
- Coding – turning answers to numbers and entering them on the computer
- Labelling – attaching meaning to the numbers (not essential but very helpful!)
- Reporting/analysing – very easy once the preceding steps are accomplished

# Univariate summaries of data

- Numerical summaries
  - Data type hierarchy: Nominal, Ordinal, Interval, Ratio
  - Measures of central tendency: means, modes and medians
  - For data with few distinct values: frequency distributions
  - Measures of spread: range, inter-quartile range, standard deviation
- Graphical summaries
  - Bar and pie charts
  - Histograms
  - Box plots

# Bivariate summaries of data

- Numerical
  - Crosstabulations
  - Comparing means across groups
  - Correlations
- Graphical
  - Side-by-side and stacked bar charts, histograms, box-plots
  - Scatter plots

# Sampling

- The use of sampling is another characteristic of QM
- Calculations based on representative samples *approximate* those of the reference population
- *Random* sampling is a powerful way of ensuring representativity
- What does *random* mean?



## Simple random sampling

- In “Simple Random Sampling” every element in the reference population has the same chance of being selected
- SRS needs a clear *sampling frame* (e.g., a list of everyone in the population) and a random selection process
- E.g., a list of all students in a university, “put the names in a hat”
- Often difficult to get a good sampling frame
- SRS more important as an ideal type for reasoning about statistics

## Varieties of sampling

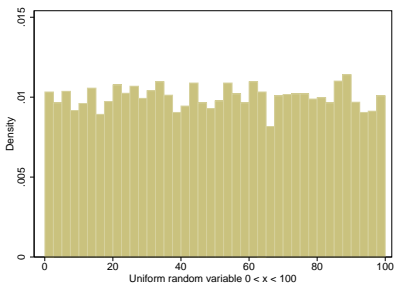
- Non-representative: accidental sampling, volunteer sampling
- Quasi-representative: quota sampling
- Representative: SRS, cluster sampling, stratified sampling
- What is representativity?

# Data distributions

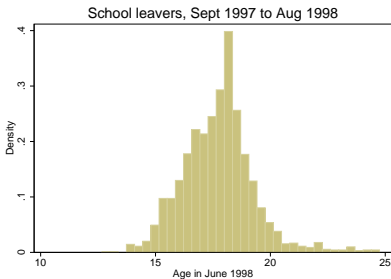
- We have seen how to display and summarise the distribution of variables:
  - Categorical: frequency distribution, percentage distribution, bar and pie charts
  - Continuous (interval/ratio): mean, median, IQR, standard deviation, histogram, box-plot

# Uniform distribution

- The shape of the histogram tells us about the distribution of the variable
- If a variable is “uniformly” distributed we see a flat distribution between the extremes:



- More often we see “heaped distributions” where more of the observations cluster around the centre, like this age example from the ESRI School-Leavers’ Survey:



# Distribution patterns

There are many patterns we might see in histograms and distributions:

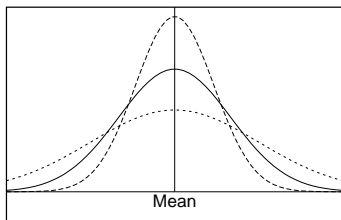
- Uniform
- Extremes
- Bimodal
- Uni-modal
  - Asymmetric
    - Positively skewed (to right)
    - Negatively skewed (to left)
  - Symmetric (with different levels of kurtosis)
    - platykurtic – flatter
    - mesokurtic – average
    - leptokurtic – very concentrated around centre

# The Normal Distribution

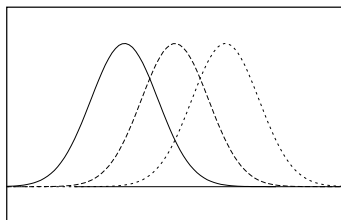
- One very important theoretically defined distribution is the “Normal Distribution”
- Occurs in the real world, e.g., where multiple measures of a fixed value have additive errors, or where there is a typical value around which cases are distributed
- The normal distribution is
  - symmetric (no skew)
  - mesokurtic (between flatter and pointier)
- The mean, mode and median are the same
- The farther you go from the mean, the lower the proportion of cases, in each direction symmetrically

## Defined by mean and std deviation

- What makes the normal distribution useful is that its form is well understood:
  - It is completely characterised by its mean and its standard deviation



Same mean, different SD



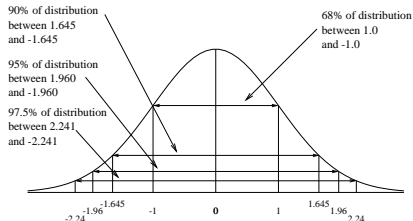
Same SD, different mean

- <http://teaching.sociology.ul.ie:3838/apps/normsd>



# Reading the Normal Distribution

- About 68% of the cases in a normal distribution will be within 1 standard deviation on either side of its mean
- 95% of cases will be within  $\pm 1.96$  std dev of the mean
- 97.5% of cases will be within  $\pm 2.24$  standard deviations of the mean



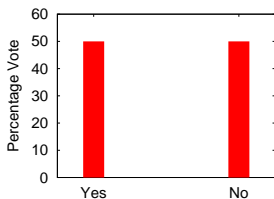
# The most important thing!

- The immediately most important thing about the normal distribution?
- Take a large sample from a population and calculate a statistic (e.g., a mean)
- Repeat a large number of times and make a histogram of your results
- These will cluster around the true population mean in a normal distribution, with
  - Mean:  $\mu$ , the true population value
  - Standard deviation:  $\frac{\sigma}{\sqrt{N}}$
- $\Rightarrow$  Sample statistics are normally distributed

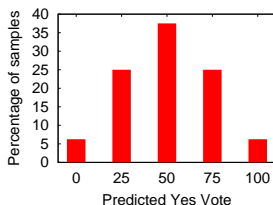
# How wrong are samples?

- A random sample gives an “approximately correct” result – how wrong is it likely to be?
- Large samples are more correct, measures of things with more variability are likely less correct
- Explore a simple case:
  - Binary outcome: yes or no (say 50:50 in population)
  - Sample size of 4 (very small, work through the details by hand)

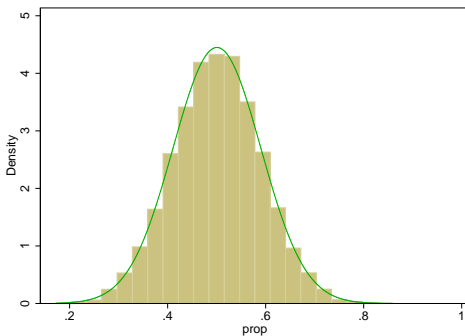
# Sampling distribution, $N=4$



## Sampling dist for $N=4$



# N=1,000; Replications: 10,000



# The Central Limit Theorem

- For a sufficiently large sample, sample estimates are distributed normally
  - Mean:  $\mu$ , the true population value
  - Standard deviation:  $\frac{\sigma}{\sqrt{N}}$
  - The “standard deviation of the sampling distribution” is called the “standard error”
- This holds no matter what the distribution of the original variable

# Sampling distributions

- Therefore, any statistic calculated on a single sample can be considered as being drawn from a “sampling distribution” with mean  $\mu$  and standard deviation  $\frac{\sigma}{\sqrt{N}}$
- This allows us to reason about how much sampling error we can expect
- For instance, 95% of the time our sample statistic will be in the range  $\mu \pm 1.96 \times \frac{\sigma}{\sqrt{N}}$

## “Confidence” intervals

- If we know that we have a 95% chance of falling in the range  $\mu \pm 1.96 \times SE$ , we can turn it around:
- There is a 95% chance that the true answer is in the range  $\bar{x} \pm 1.96 \times SE$
- Since we don't know  $\sigma$ , the population standard deviation, we estimate it using the sample standard deviation,  $s$ :
- Confidence interval:  $\bar{x} \pm 1.96 \times \frac{s}{\sqrt{N}}$
- Interpretation: in 95% of large simple random samples, the true value will fall within the CI



## Different confidence levels

- 95% implies a 1 in 20 chance of the true value not falling in the interval
- Confidence levels can be changed: 99% CI is  $\bar{x} \pm 2.58 \times SE$

## CIs and proportions

- We can also calculate CIs around proportions:  $p \pm 1.96 \times SE_p$ 
  - The SD of a proportion  $p$  is  $\sqrt{p \times (1 - p)}$
  - The SE of a proportion is therefore  $\sqrt{\frac{p \times (1 - p)}{N}}$
- For  $N = 1,000$  the CI is  $\pm 3\%$ :

$$1.96 \times SE = 1.96 \times \sqrt{\frac{0.5 \times 0.5}{1000}} = 0.031$$

# Student's $t$ -distribution

- The CLT speaks of large samples
- For smaller samples it is only approximately true, and the CI as defined above will be too narrow
- For means, we can use “Student's  $t$  distribution” in place of the normal distribution
- For large samples, we use  $z = 1.96$  for 95% CIs; for smaller samples we use  $t > 1.96$
- This makes for wider, correct CIs

# CIs – how fuzzy is the answer?

- Confidence intervals allow us to present sample information appropriately
  - Point estimate, *e.g.*, mean or other sample statistic: our “best guess” of the true value
  - Confidence interval: range in which we are “confident” true value (the population parameter) lies
- In combination, the point estimate and CI give us the answer with a measure of its precision

# Hypothesis testing

- Hypothesis testing is a way of using the reasoning behind CIs to make specific claims about the population
- Say we want to know if there is a relationship between two variables, e.g., whether taking a particular qualification has an effect on wages (positive or negative)
  - We begin with a hypothesis:  $W_a \neq W_b$  or  $W_a > W_b$
  - We negate the hypothesis to form a “{null hypothesis}”, called  $H_0$ :

$$H_0 : W_a = W_b$$

$$\Rightarrow H_0 : D_w = W_a - W_b = 0$$

- That is, on average, the wage difference is zero

# Testing the null hypothesis

- First we calculate a sample mean wage difference,  $\hat{D}_w$
- Then we construct a confidence interval at our chosen level of confidence (e.g., 95%):  $\hat{D}_w \pm z_{0.95} \times SE$
- If zero lies outside the interval, we are at least 95% sure the true population value is not zero, and we can reject the null hypothesis
- If zero lies within the interval, then zero is in the range of plausible values, so we **cannot reject the null hypothesis**
- We don't say we "**accept the null hypothesis**" because zero is just in the range of plausible values, and other values in this range are approximately as likely

# Rejecting the null hypothesis?

- Rejecting the null hypothesis constitutes support for the initial or “alternative” hypothesis
- Failing to reject the null hypothesis means the data fail to support the initial hypothesis: “there is no evidence that the course affects wage”
- Failure to support the initial hypothesis may be because
  - It is actually false, *i.e.*,  $D_w = 0$
  - The effect is small and/or very variable, and thus the sample is too small to detect it

# Significance

- Let's say we do a hypothesis test with a 95% confidence level, and we find the zero is way outside the CI
- We can try again with a 99% confidence level:
  - If it is still outside the interval we are not "at least 95%" but "at least 99%" sure that zero is not the true value
- If we keep trying with CIs with higher confidence levels we will eventually find one where zero is just outside the interval
- If that is at confidence level  $C$  we can say that we are  $C\%$  sure (not "at least" any more) that zero is not the true value



# $p$ -values

- This  $p = 100\% - C\%$  value is the probability that we get a sample statistic as different from zero as we did, even though the true value was zero
- This is known as the **significance** of the sample estimate, or its  $p$ -value
- We want it to be as small as possible, typically under 5% (0.05)
- $p$ -values are widely used – stats programmes report them in many places
- In general the interpretation is “what’s the probability of getting this result by chance if the null hypothesis was true?”

## $t$ -test shortcuts

- Rather than try repeatedly to get a CI just short of touching zero, we can calculate a  $t$ -statistic:

$$t = \frac{\bar{x}}{SE}$$

- If this  $t$  is greater than the “critical value”, e.g., 1.96 for large samples at 95% confidence, we can reject the null hypothesis
- If the CI doesn't include zero,  $t$  will be greater than the critical value
- We can also calculate the exact  $p$ -value for the  $t$ -statistic

# Error in hypothesis testing

- Another way of looking at significance is “the chance we would be wrong if we believe the initial hypothesis”
- For instance, if there is one chance in twenty ( $p = 0.05$ ) that the true value is outside the CI, then by basing our decision on the CI we will be wrong one time in twenty
- This is known as **Type I Error**: rejecting the null hypothesis when it is true
  - e.g., the true value might be zero but a small number of possible samples generate CIs that don't include zero
- If it is very important to avoid Type I error, we use high confidence levels (e.g., 99.5% instead of 95%) or insist on low  $p$ -values (e.g., 0.005 instead of 0.05)

# Type II error

- However, there is a second type of error, **Type II**
- **Type II Error** is failing to reject the null hypothesis when it is false
- That is, failing to support the initial hypothesis even though it is true
- If we raise the confidence level we reduce the risk of Type I error but raise the risk of Type II error
- That is, if we make a special effort not to accept an initial hypothesis unless there is very clear evidence, we necessarily fail to accept it where there is only fairly clear evidence
- For a given  $p$ -value, we can only reduce the Type II error by increasing the sample size

# Association in tables

- We detect association in tables many ways
  - Comparing row percentages up and down columns
  - Column percentages across rows
  - Comparing observed with *expected* values
  - “Expected”  $\Rightarrow$  the concept of “independence”

# The $\chi^2$ test for association in tables

- Independence: no association between two variables
  - pattern of row percentages the same in all rows
  - pattern of column percentages the same in all columns
- But even if independence holds in the population, sampling variability leads to differences in percentages
- How big can the differences be before we can be convinced that there is really association in the population?

# Observed and Expected

- Method: compare the real table (“observed”) with hypothetical table under independence (“expected”)
- Summarise the difference into a single figure ( $\chi^2$  statistic, chi-sq)
- Compare  $\chi^2$  statistic with known distribution
- ... What is the probability of getting a sample statistic “at least this big” by simple sampling variability *if independence holds in the population?*

## Calculating the $\chi^2$ statistic

- The “expected” table has the same row and column totals, but the cell values are such that the percentages are the same as in the total row and column:

$$n_{ij} = \frac{R_i C_j}{T}$$

- For each cell we summarise the difference between observed ( $O$ ) and expected ( $E$ ) values as

$$\frac{(O - E)^2}{E}$$

- The summary for the table as a whole is the sum of this quantity across all cells:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$



# $\chi^2$ distribution

- This statistic is known to have a predictable distribution, the  $\chi^2$  distribution
- That is, if we take a large number of samples from a population where there is no association, and calculate the statistic, they will have a distribution in a known form, and we can calculate the probability of finding a value “at least as large as” any given number
- The distribution depends only on the “degrees of freedom” which is the number of rows minus one times the number of columns minus one:

$$df = (r - 1)(c - 1)$$

## Critical values and hypothesis testing

- Reading the table, we go to the row corresponding to the degrees of freedom, and read across until we get to the column with our chosen probability level (say 0.05) – this gives us the appropriate “critical value”
- If our  $\chi^2$  is bigger than the critical value, then there is at most one chance in 20 (*i.e.*, 0.05) that it has arisen by sampling variability and a 95% chance (*i.e.*,  $1 - 0.05$ ) that it is due to real association in the population
- When using computers, the exact  $p$ -value of the calculated  $\chi^2$  statistic is reported: if  $p \leq 0.05$  then we can reject the null hypothesis of no association with 95% confidence

# Critical values for $t$

- With both the  $\chi^2$  test and the  $t$  test, we
  - calculate a “test statistic”
  - compare it with a “critical value”, and/or
  - calculate its exact  $p$  value
- When testing for association in tables,  $\chi^2 = \sum \frac{(O-E)^2}{E}$
- When comparing a mean to zero,  $t = \frac{\bar{x}}{SE}$

# Multiple $t$ -tests

- We have looked at the “paired sample”  $t$ -test, where we compare a difference (between paired observations) to zero:

$$t = \frac{\bar{x}}{SE}$$

- This is a special case of the one-sample  $t$ -test, where we compare a sample statistic to a fixed reference value,  $r$ :

$$t = \frac{\bar{x} - r}{SE}$$

- This also applies to proportions, comparing to a reference value such as 50%:

$$t = (p - r) / \sqrt{p(1 - p) / N}$$

(note: use normal distribution, not  $t$ -distribution, as long as sample is large enough)

# The “independent-sample t-test”

- A third case is the “independent-sample t-test”, where we compare means across different (sub-)samples
- If we wish to test for differences across groups (e.g., differences in income between men and women) we are comparing one sample mean with another, not a sample mean with a fixed value
- We can consider the sample difference ( $\bar{x}_m - \bar{x}_w$ ) to be a point estimate of the population difference
- The null hypothesis is that  $\mu_m = \mu_w$ , or  $\mu_m - \mu_w = 0$

# Testing for differences in means

- To construct a CI or calculate the test statistic, we need the SE
- Where the groups have different variances this is

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- If we can assume the sub-population variances are the same it simplifies to

$$\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

# Test statistic

- The test statistic is standard, using the appropriate SE:

$$t = \frac{x_m - x_w}{SE}$$

- The degrees of freedom are complicated to calculate in the general case
- In the equal-variance case they are  $n - 2$  as two sample statistics are calculated