MATH9102

INSPECTING AND PREPARING DATA

What do I need to describe for numerical data?

Centre

- Discuss where the middle of the data falls
- Measures of central tendency
  - mean, median and mode

Spread

- Discuss how spread out the data is
- Refers to the variability in the data
  - Range, standard deviation, IQR

Shape

- Refers to the overall shape of the distribution
- Symmetrical, uniform, skewed, or bimodal

What do I need to describe for numerical data?

Unusual Occurrences

- Outliers (value that lies away from the rest of the data)
- Gaps
- Clusters

Context

- You must write your answer
  - With reference to the context in the problem you are investigating,
  - Using **correct statistical vocabulary** adhering to referencing scheme guidelines
  - Using complete sentences.

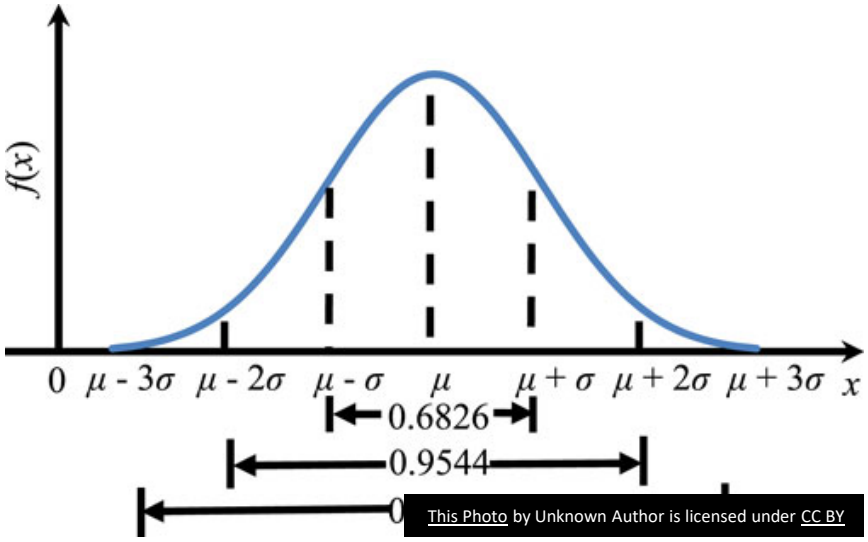
Parametric v Non-parametric

Parametric

- Make assumptions about the population from which the sample is taken
- Shape of the population (normally distributed)

Non-parametric

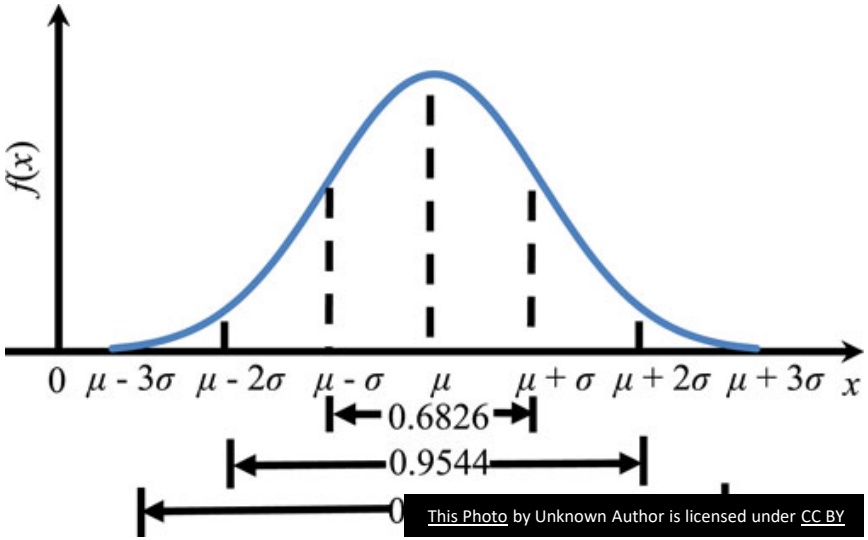
- Do not make assumptions about the population and its distribution
- Tolerant set of tests which don't expect your data to anything fancy
  - Not high-powered and don't promise more than they can deliver
  - May fail to detect differences that exist
- Use for nominal or ordinal data
- Use for small samples
- Use for skewed data



The normal distribution

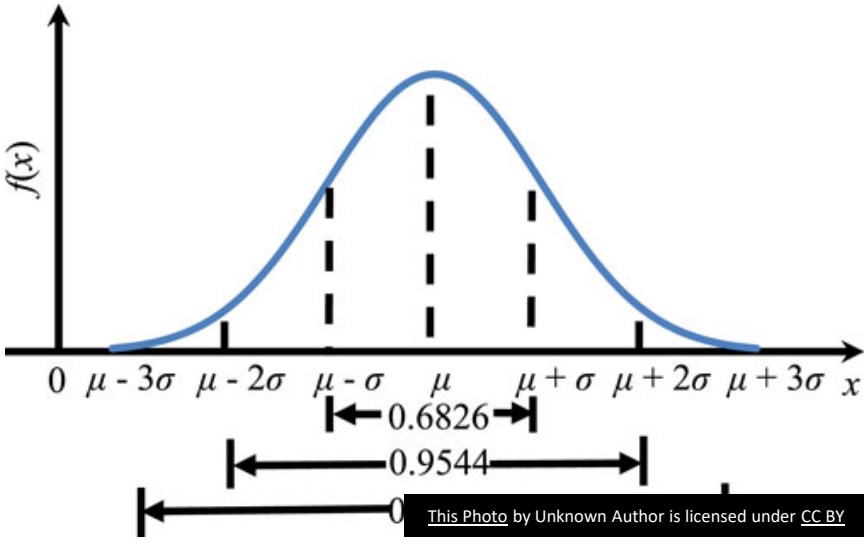
When we take a random sample from a population and compute a statistic (e.g., the sample mean), the Central Limit Theorem (CLT) tells us that this statistic will have an approximately normal distribution, even if the population itself is not perfectly normal (as long as the sample size is reasonably large).

The normal distribution



Parametric tests assume normality.

- The test statistics (like the t-statistic or F-statistic) are derived under the assumption that data, or at least the sampling distribution of the mean, follows a normal distribution.

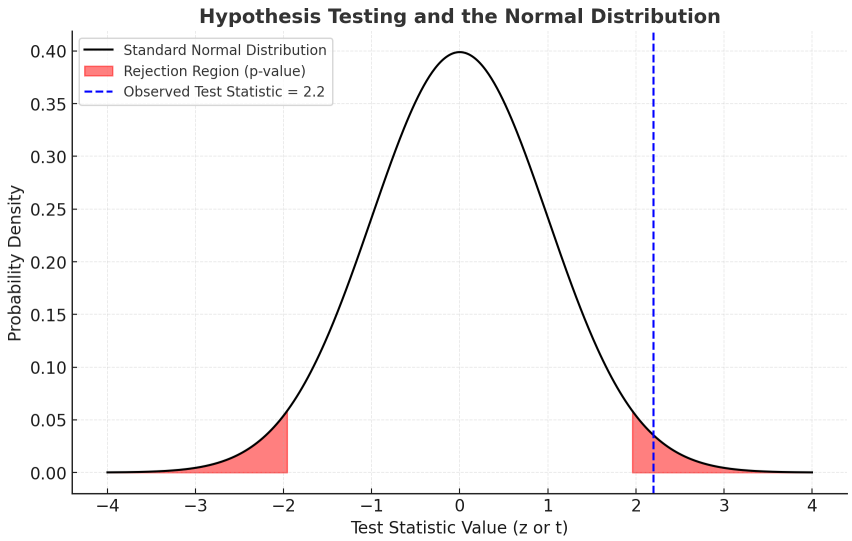


The normal distribution

If the assumption of normality is badly violated (especially in small samples), the test statistics may not follow their theoretical distributions, making p-values invalid.

For large samples, the Central Limit Theorem helps:

- A normal approximation works even when data are not strictly normal, which is why parametric tests are often considered “robust” to moderate deviations from normality.



Hypothesis Testing

Hypothesis testing depends on calculating a p-value:

- The probability of observing a test statistic as extreme as the one we obtained, assuming the null hypothesis is true.

These probabilities are computed using the normal distribution (or distributions derived from it: t, F, χ^2).

Without the normal distribution framework, we couldn't assign these probabilities reliably.

Choices to be made before conducting a hypothesis test/building a model

Deciding if you have sufficient data and sufficient variability within that data

Correcting for non-response, design effect

- Weighting variable
- Be careful of scale up weighting

Missing data

- Decide what level of data is missing
- Decide what the pattern is
- Decide why it is missing
- Correct accordingly or ignore

Normality

- Inspect and test for normality
- Be aware of allowable limits
- If not normal
  - Could use non-parametric tests
  - OR
  - Apply a transformation to see if that results in a normal distribution

Linearity and homoscedasticity (scale variables)

- Inspect your scatterplot
- If assumptions are not addressed could consider transformation but also non-parametric test

Preparing your data

You may need to consider doing the following:

- Making a decision about missing data
- Making a decision about outliers
- Recoding your variables
  - E.g. to reduce the number of categories
  - Doing so will not be objective but working with categorization at all is highly contested and highly political
- Weighting your data to correct for bias, address design effects, make sample more representative of the population
- Selecting cases
  - To work only with particular sub-groups of data
- Splitting your file
  - Allows you to organise your output by category of variable you are interested in

Why?

Its all about the bias….

Bias

We want our sample to be representative of the population from which it is taken

A sample is biased if it is systematically different from the population

- Under representation
- Over representation

Bias

Selection Bias

- Choosing only to use certain cases or groups of cases in your analysis
  - Randomness is not achieved

Sampling Bias

- Caused by non-random sampling
- Is a fact of life in most cases as you mainly use convenience samples
- Need to recognise and try to minimise the impact

Time interval Bias

- By specifying a time bounded interval for collecting data

Confirmation Bias

- Choosing data which is more likely to support your belief which is determined in advance of analysis

Omitted Variable Bias

- Choosing not to collect data about certain concepts or not to include variables which represent certain concepts in predictive models

Missing Data

Missing data affects the validity of statistical analysis and can lead to invalid conclusions

- Reduces statistical power
  - the probability that the test will reject the null hypothesis when it is does not hold.
- Can reduce the representativeness of the samples.
  - Can cause bias in the estimation of parameters.

May complicate the analysis of the study.

Missing Data

What is certain in quantitative research?

- Measurement error
- Missing data

Missing data can be:

- Due to preventable errors, mistakes, or lack of foresight by the researcher
- Due to problems outside the control of the researcher
- Deliberate, intended, or planned by the researcher to reduce cost or respondent burden
- Due to differential applicability of some items to subsets of respondents
- Etc.

Why are the values missing

Automated Data Collection (e.g., web scraping, monitoring, sensors)

Non-response analogue: Some sources may be systematically missing. For example, scraping social media may exclude private accounts, deleted posts, or platforms that block crawlers.

Unobserved factors: Technical constraints (e.g., only collecting during certain times of day, only capturing certain device types) can bias the dataset.

Example: Traffic monitoring systems may miss cyclists or pedestrians if only cars are recorded.

Why are the values missing

Experimental Studies

Attrition (dropout bias): Participants may leave an experiment early, and dropout is often non-random (e.g., those not experiencing a treatment benefit may withdraw).

Non-compliance: Some participants may not follow protocols (e.g., skipping medication, ignoring instructions). If this is not random, it biases results.

Unobserved factors: Conditions like lab environment, researcher effects, or hidden participant traits may affect outcomes but aren’t measured.

Why are the values missing

Observational / Administrative Data

Coverage bias: Administrative records (e.g., hospital data, school registers) only capture people who interact with those systems, missing those outside.

Systematic exclusions: Some groups may be underrepresented (e.g., undocumented migrants missing from government datasets).

Mechanism knowledge: Sometimes we know who is excluded (e.g., lack of insurance → not in claims data), which allows adjustment with auxiliary variables.

Why are the values missing

Survey-Specific Issues (with generalisable analogues)

Survey refusals: Equivalent in non-survey contexts to “opt-out” of participation (e.g., patients refusing to be part of a clinical registry).

“Don’t know” responses: Analogous to uncertain or incomplete data logs (e.g., IoT devices transmitting null values).

Differential applicability: Just as some survey items apply only to subsets of respondents, certain measurements in experiments may apply only to subsets (e.g., biomarkers only measurable in those with sufficient sample material).

Why are the values missing

- Data Entry Errors
  - Human Error: Mistakes made during data entry can lead to missing values, such as typing errors or omitting values accidentally.
  - System Limitations: Certain data management systems may not enforce mandatory fields, leading to incomplete records.
- Data Integration Issues
  - Mismatched Sources: When merging data from different sources, discrepancies in data formats or schemas can lead to missing values for some records.
  - Duplicate Entries: During data consolidation, duplicate records may be eliminated, which could inadvertently remove data fields from some records.
- Data Corruption
  - File Corruption: Data files can become corrupted due to system crashes, hardware failures, or software bugs, resulting in missing or unusable data.
  - Incompatible Formats: Data transferred between incompatible systems may result in loss or corruption of certain fields.

Why are the values missing

- Incomplete Data Collection
  - Limited Data Collection Period: Data may be collected over a limited time, leaving gaps for periods not covered by the data collection process.
  - Insufficient Coverage: If the sample or data collection process does not encompass all relevant categories or time frames, some data points will be missing.
- Data Privacy and Security
  - Anonymization: In some cases, data may be intentionally withheld or anonymized to protect individuals' privacy, leading to missing values for certain sensitive fields.
  - Compliance Regulations: Data retention policies driven by privacy regulations (like GDPR) may lead to the deletion of certain records.
- Data Filtering or Preprocessing
  - Outlier Removal: During data cleaning, outliers may be removed, leading to missing data for certain records.
  - Thresholds: Data might be filtered out based on specific thresholds (e.g., only including records with a certain level of detail), resulting in missing entries for others.

Why are values missing?

Missing by researcher error

- May be missing completely at random
- May reflect researcher bias
- Due to preventable errors, mistakes, or lack of foresight by the researcher
- Due to problems outside the control of the researcher
  - E.g.
  - Perceived risk to researcher

Deliberate, intended, or planned by the researcher to reduce cost or respondent burden

Why are the values missing?

Code reason value is missing

- Depends on your domain but standards will apply
  - e.g. -99 social sciences data
  - NA by default in R
- Treat each reason differently

Can we impute a better value?

- Should we?

Why do we need to care about missing data?

Source of bias

- Introduces the possibility of making inferences on the basis of sample data that are inadvertently biased in unknown directions

Choice of treatment (e.g. deletion) can lead to loss of information and loss of statistical power through reduced sample size

Makes some common tests inappropriate or difficult to use

Why do we need to care about missing data?

The mechanism and the pattern of missing data have greater impact on results than does the amount of data missing

The logic of statistical inference presumes that the sample is randomly drawn from the population.

- Thus whether the missing data within a sample are random is important.

Why do we need to care about missing data?

When data are missing in a random fashion, there is no systematic difference between the available data and the missing data;

They are both random subsets of the data composing the entire sample.

Why are the Values Missing: The reason instructs the solution

1. Data missing at random (MAR).
   - The distribution of the missing data is related to the distribution of the observed data but not to the values in the variable which has missing values.

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	
29	
30	
30	
31	
44	118
46	93
48	141
51	104
51	116
54	97

Here the missing values for IQ score are related to the value of the Age variable.

Why are the Values Missing: The reason instructs the solution

2. Data missing completely at random (MCAR).
   - When the probability of missing data on a variable is unrelated to any other measured variable and is unrelated to the variable with missing values itself.
   - i.e. the missingness on the variable is completely unsystematic.

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

Here there does not appear to be a pattern to the missingness of IQ Score.

Why are the Values Missing: The reason instructs the solution

3. Data that are not missing at random (MNAR)
   - When the missing values on a variable are related to the values of that variable itself, even after controlling for other variables.

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	133
26	121
29	
30	
30	110
31	
44	118
46	
48	141
51	
51	116
54	

Data are missing on IQ but only the people with low IQ values have missing observations for this variable.

Evaluating Missing Data

Missing data mechanism

- Missing completely at random (MCAR)
  - Can be ignored
- Missing at random (MAR)
  - May possibly be ignored - need to identify the reason and then decide if it can be ignored
- Missing not at random (MNAR)
  - Not ignorable

Missing Data Mechanisms

The appropriateness of different missing data treatments depends (among other things) on the underlying missing data mechanism

"Real" missing data can seldom be classified into just one of the three (MCAR, MAR, MNAR)

Because we don't have access to the missing data (Y_{miss}), we can not empirically test whether or not the data is MNAR

If we know (or can convincingly argue) that the data is *not* MNAR, a test of whether the data is MCAR is available (in various packages in R).

Evaluating Missing Data – What should you do?

Consider the amount of missing data

- Percent of cases with missing data
- Percent of variables having missing data
- Percent of data values that are missing

Consider the pattern of missing data

- Missing by design
- Missing data patterns

Why are the Values Missing

Understand why each value is missing

Delete observations or variables where you do not intend to impute a value

- Drop variable
- Drop observation

Must report that you have done this and why.

Missing Data

Useful Paper

Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May;64(5):402-6. doi: 10.4097/kjae.2013.64.5.402. Epub 2013 May 24. PMID: 23741561; PMCID: PMC3668100.

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/#:~:text=Missing%20data%20can%20reduce%20the,estimates%2C%20leading%20to%20invalid%20conclusions.&text=Missing%20data%20(or%20missing%20values,in%20the%20observation%20of%20interest.

What usually happens?

If our sample is large, we may be able to allow cases to be excluded.

If our sample is small, we will try to use a substitution method so that we can retain enough cases to have sufficient power to detect effects.

In either case, we need to make certain that we understand the potential impact that missing data may have on our analysis.

Goals of a Missing Data Treatment

Preserve the essential characteristics of the data

- Distributions of the variables
- Relationships among the variables

Maintain the representativeness of the analyzed data

Provide valid statistical inference (control Type I error)

Maximize the statistical power of the study and its statistical analyses (minimize Type II error)

Avoid bias and instability in the parameter estimates and standard errors for statistical models

Missing data

If missing data represent less than 5% of the total and is missing in a random pattern from a large data set, almost any procedure for handling missing values yields similar results

Tabachnik and Fidell, Using Multivariate Statistics, 6th Edition, Pearson

Missing data


```
# install.packages("naniar")  # if not already installed

library(naniar)

# Example: running Little's MCAR test on a dataset called 'survey'

result <- mcar_test(survey)

print(result)
```

This test evaluates whether the missing data pattern is consistent with MCAR.

A **non-significant p-value (p > 0.05)** suggests that the data could plausibly be MCAR.

A **significant p-value (p < 0.05)** suggests the data is *not* MCAR (likely MAR or MNAR).

Missing Data Treatments – Deletion Methods

Listwise deletion (complete case analysis)

- Deletes the case if any variable is missing data
- All cases with missing scores on one or more variables are excluded from the analysis.
- The advantage of this method is that the remaining dataset is complete.
  - But this complete dataset has a reduced sample size and power, caused by the loss of the incomplete cases.
- The chance of having a biased dataset is substantial if data is not MCAR.
- In most situations, the disadvantages of listwise deletion far outweigh its advantages

Missing Data Treatments – Deletion Methods

Pairwise deletion (available case analysis)

- Deletes case only when considering the variable for which data is missing, can still use it for other variables.
- Any given case may contribute to some analyses but not to others.
  - So the sample size varies from test to test/model to model - the sample size will be the same for some analyses but will be reduced for others.
- Using this method the assumption of the MCAR mechanism can produce unbiased estimates
- But also the inconsistency of the sample size can lead to problems in computing standard errors.

Typical Missing Data Treatments - Imputation

Replace missing values with a substitution (uses historical data)

Cold deck (uses data from previous study or historical study)

Hot deck (donor case) imputation

- Various forms : mean, nearest neighbor, random

Mean substitution

- (Variable) mean substitution
- Mean substitution with added random error

Regression imputation

- Regression predicted value imputation
- Regression imputation with added random error

Modern Missing Data Treatment

Maximum likelihood (ML)

- Estimates summary statistics or statistical models using all available data
- Uses each cases available data to compute maximum likelihood estimates.
- The maximum likelihood estimate of a parameter is the value of the parameter that is most likely to have resulted in the observed data.
- The likelihood is computed separately for those cases with complete data on some variables and those with complete data on all variables.
- These two likelihoods are then maximized together to find the estimates.
- Gives unbiased parameter estimates and standard errors.
- Limited to linear models.

Modern Missing Data Treatment

Multiple imputation

- Imputes individual data values in multiple complete datasets, averaging the results of the statistical analyses across these datasets
- Because it uses an imputation method with error built in, the multiple estimates should be similar, but not identical.
- The result is multiple data sets with identical values for all of the non-missing values and slightly different values for the imputed values in each data set.
- The statistical analysis of interest, such as ANOVA or logistic regression, is performed separately on each data set, and the results are then combined.
- Because of the variation in the imputed values, there should also be variation in the parameter estimates, leading to appropriate estimates of standard errors and appropriate p-values.

Statistical Analysis with Missing Data

What do you get when you don't specify what you want?

What choices do you have within a given analysis procedure?

- Listwise
  - Complete-case analysis removes all data for a case that has one or more missing values for variables of interest
- Pairwise
  - Attempts to minimize the loss that occurs in listwise deletion.
  - Will conduct relevant analysis for variables if data exists in one or more of them and it makes sense to do so.

Missing data treatments carried out prior to analysis

- Ad hoc methods (Listwise, pairwise, single imputation, etc.)
- Modern methods(Maximum Likelihood, Multiple Imputation)

Missing Data

If you find a variable with a large amount of missing data you need to find out why it is missing

If it is not missing at random you need to deal with it in your tests

- In R most functions have a series of na parameters you can set to indicate what you want to do e.g. na.omit=true

General Steps for Dealing with Missing Data

Identify patterns/reasons for missing and recode correctly

Understand distribution of missing data

- Consider the probability of missingness
- Are certain groups more likely to have missing values?
- Example: Respondents in service occupations less likely to report income
- Are certain responses more likely to be missing?
- Example: Respondents with high income less likely to report income

General Steps for Dealing with Missing Data

Decide on best method of analysis

- Use what you know about
- Why data is missing
  - Distribution of missing data

Decide on the best analysis strategy to yield the least biased estimates

- Deletion Methods
  - Listwise deletion, pairwise deletion
- Single Imputation Methods
  - Mean/mode substitution, dummy variable method, single regression
- Model-Based Methods
  - Maximum Likelihood, Multiple imputation

Handling Missing Data in R

See the Quarto notebook accompanying the lecture

Calculate the % for the variables of interest

Visualise the patterns

Make a judgement on whether it is MAR, MCAR or MNAR and what you will do about it

Handling Missing Data in R

You can choose to eliminate all the data

- `ydata <- na.omit(survey)`

You can filter those that are na for all relevant variables

Most modelling functions offer you an option for handling missing data

- E.g. na.rm=True

Imputing missing data

- `Simputation package`
- `Hmisc` package contains several functions that are helpful for missing value imputation (agreImpute(), impute() and transcan())
  - https://cran.r-project.org/web/packages/Hmisc/index.html#:~:text=Contains%20many%20functions%20useful%20for,of%20R%20objects%20to%20LaTeX
- `mitools` package
  - https://cran.r-project.org/web/packages/mitools/index.html

Survey Data – Sources of Bias

Oversampling

- Some groups are purposefully sampled more frequently than others to produce a large enough sample size for analysis.

Undersampling

- Sample size is intentionally reduced to ensure equal representation of different groups.

Non-response

- When individuals who choose not to participate in a survey have different characteristics or opinions than those who participate
- This can result in under- or overestimating certain aspects of a population, leading to a skewed representation of the data.

Weighting

Use to

- Correct for any known bias that may exist in the final sample
- Scale-up frequencies so that frequencies calculated from the sample represent estimates for the population as a whole
- To address ‘design effects’ arising from the sampling methods used.

Example

Suppose we undertake a survey

We have selected a random sample for this purpose

Suppose we collect data on gender and marital status.

Example

Work out your actual proportions surveyed:

Demographic group	Percentage
Unmarried male	18.5%
Unmarried female	18.7%
Married male	21.2%
Married female	20.9%
Widowed man	2.3%
Widowed woman	3.1%
Divorced man	4.7%
Divorced woman	6.6%

Work out your expected proportions based on population

Demographic group	Percentage
Unmarried male	20%
Unmarried female	16.9%
Married male	23.8%
Married female	23.7%
Widowed man	1.4%
Widowed woman	4.7%
Divorced man	4.1%
Divorced woman	5.4%

Weighting – Correcting for Non-Response

We could conduct our analysis doing nothing to correct for this disparity, but we would be running the risk of creating biased estimates (perhaps males and females prefer different facilities)

By not correcting for bias in our sample we are giving more weight to some categories than others.

Weighting Factor

Demographic Group	Expected Distribution	Actual Distribution	Weighting Factor
Unmarried Male	20%	18.5%	1.08
Unmarried Female	16.9%	18.7%	0.90
Married Male	23.8%	21.2%	1.12
Married Female	23.7%	20.9%	1.14
Widowed Man	1.4%	2.3%	0.61
Widowed Woman	4.7%	3.1%	1.52
Divorced Man	4.1%	4.7%	0.87
Divorced Woman	5.4%	6.6%	0.82

Calculate the weights for each demographic group by dividing the expected percentage by the actual percentage

Weighting – Correcting for Non-Response

How do we use the weighting factor?

We multiply all responses for that category by the corresponding weighting factor.

We apply the weight in the calculation of all relevant statistics (e.g. weighted mean).

In R: simple multiplier applied to all variables of interest, can include in your dataset

Weighting – Correcting for Non-Response

Weighting variables are normally used to correct for known biases in several factors.

In the youthcohort.sav (Paul Connolly) the weighting variable is used to correct for four factors related to non-response (gender, qualifications achieved, region in which they live and type of school)

Principle for calculating weighting variable remains the same.

Weighting – Correcting for design effects

Due to cluster sampling

- Used in large-scale surveys where it would be too costly to select a simple random sample

Suppose we need to conduct a survey of 10-11 yr. old pupils in school

Assuming it is possible to create a full-list of all eligible pupils in the country, then the selection of a simple random sample of 2000 pupils could mean having to interview children scattered across 800 schools throughout the country

- Lots of cost in doing interviews
- Huge amount of time in negotiating access, seeking permission etc.

Weighting – Correcting for design effects

An alternative is to use some form of cluster sampling

- Select 100 schools at random
- Survey all children in those schools
- If on average 20 pupils per school takes part this gives us our sample of 2000 pupils

But this may underestimate the amount of variation within the population as a whole

- Pupils within each cluster (each school) are likely to be more like each other than to those outside the school
- Thus, we are likely to underestimate variation unless we correct for this 'design effect'
- Consequence: increased risk of Type I error

Weighting – correcting for design effect

Correct Approach: Use multi-level modelling (examine each cluster)

- Beyond the scope of this module

Compromise Approach: Employ a weighting variable that attempts to reduce the overall size of the sample so that standard errors are increased to account for this

In earlychildhood.sav and afterschools.sav WTCORRCT corrects for known biases by non-response and provides a correction due to sampling.

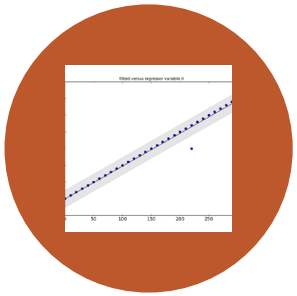
What do I need to do?

Check if a weighting variable is included in your dataset.

If yes

- Find out what type of weighting variable it is
- Design effect – apply for all analysis
- Non-response – apply for all analysis – but check the sampling method and use caution in reporting as above

Outliers



CASES THAT HAVE DATA VALUES THAT ARE VERY DIFFERENT FROM THE DATA VALUES FOR THE MAJORITY OF CASES IN THE DATA SET.

IMPORTANT BECAUSE THEY CAN CHANGE THE RESULTS OF OUR DATA ANALYSIS.

WHETHER WE INCLUDE OR EXCLUDE OUTLIERS FROM A DATA ANALYSIS DEPENDS ON THE REASON WHY THE CASE IS AN OUTLIER AND THE PURPOSE OF THE ANALYSIS.

Univariate and Multivariate Outliers

Univariate outliers

- Cases that have an unusual value for a single variable

Multivariate outliers

- Cases that have an unusual combination of values for several of the variables.
- The value for any of the individual variables may not be a univariate outlier, but, in combination with other variables, is a case that occurs very rarely.

Outliers

Reasons for outliers

- Data entry error
- Failure to specify a particular value for missing data
- Outlier not a true member of population of interest
- Outlier is a true member of population of interest with an extreme score

What to do?

- Transform to standardized variables
  - Look at histogram
  - Sometimes transforming data can “pull in” the outlier
- Censoring outliers
  - May need to delete case/s and run with and without outlier

Standard Scores Detect Univariate Outliers

One way to identify univariate outliers is to convert all values for a variable to standard scores.

If the sample size is small (80 or fewer cases)

- a case is an outlier if its standard score is ±2.5 or beyond.

If the sample size is larger than 80 cases

- a case is an outlier if its standard score is ±3.29 or beyond

This method applies to interval/ratio level variables, and to ordinal level variables that are treated as metric.

It does not apply to nominal level variables.