Revision

Probability and Statistical Inference

Bojan Božić

TU Dublin

Winter 2020

Student Feedback

- Open Brightspace and select the module.
- Go to Assessment $\rightarrow$ Surveys
- Complete **Q6A Survey**
- Time: 15mins

Basics of Statistics

- Science of collection, presentation, analysis, and reasonable interpretation of data.
- Statistics provides a rigorous scientific method for gaining insight into data.

Variables

- For our purposes a statistical experiment or observation is any process through which measurements are obtained.
- Common to use the letter x to represent the quantitative results of an experiment or observation.
- X is a variable, x is the value of the variable.

Variables

- Not only something we measure.
  - Things we can manipulate.
  - Compute.
  - Or control for.
- Others we measure indirectly.
  - There will sometimes be a difference between the numbers we use to represent a thing we are measuring and the actual value of the thing (if we were measuring it directly).
  - Measurement error.
  - E.g. psychological tests are approximate measures.

Study Design

- A careful advance plan of data collection and the analytic approach is needed to answer the question under investigation in a scientific way.
- The basic elements of a study design.
  - Selecting an appropriate sample size for a specified level of power and level of significance.
  - Select appropriate measures.
  - Selecting methods of sampling, data collection, and analysis appropriate to the study's objectives.

Guidelines for Presenting Descriptive Statistics

- Ensure that you are using the most appropriate way of summarizing and presenting your data.
- Be as efficient as possible when presenting your findings.
  - All charts and tables should, as far as possible, be self-explanatory.
  - Use appropriate visualisation for the variables of interest.
- Be consistent in the way you present your findings.
- Ensure that your data are not presented in a way that may be misleading and/or confusing.

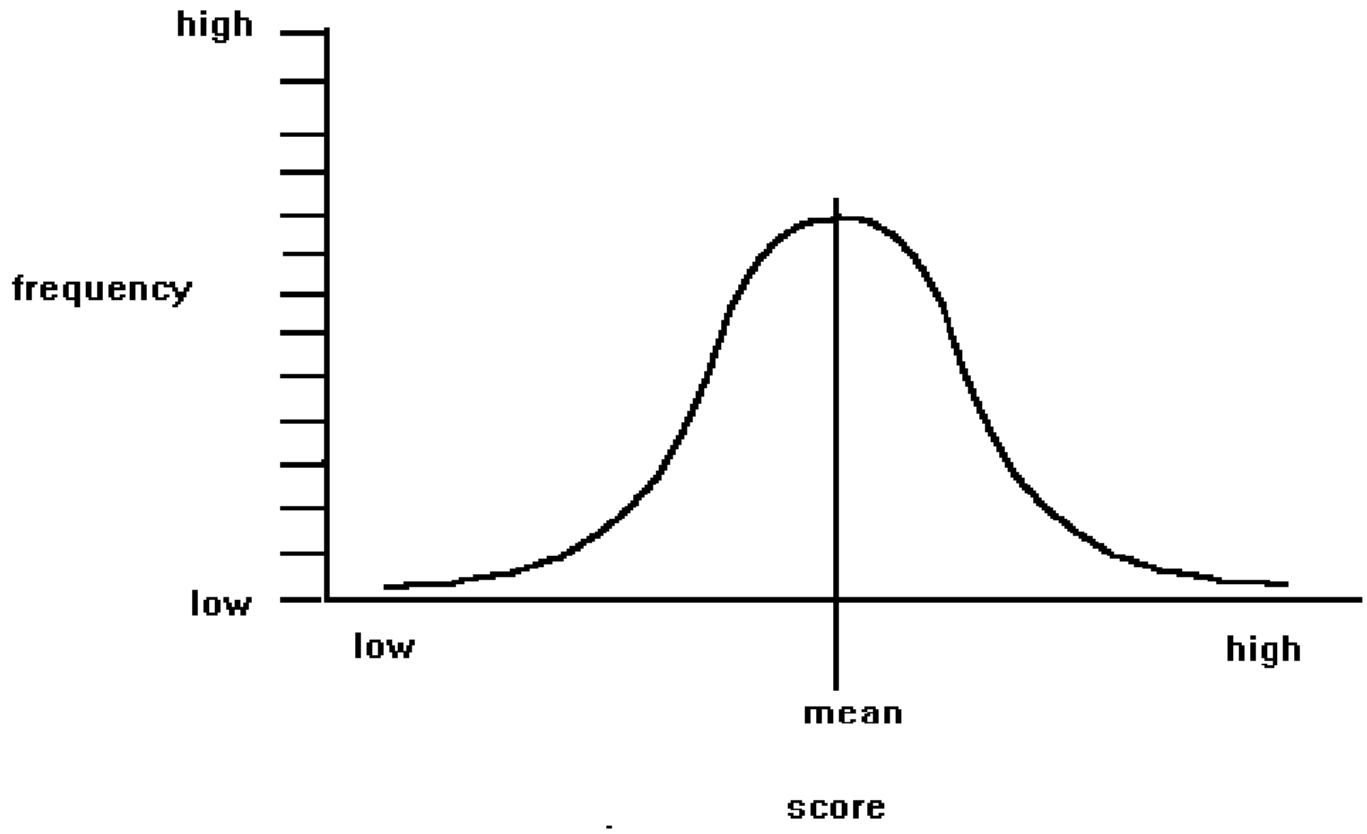
What do I need to describe for numerical data?

- 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.

The Normal Distribution



The Normal curve is a mathematical abstraction which conveniently describes or "models" many frequency distributions of scores in real-life.

Properties of z-scores

- 1.96 cuts off the top 2.5% of the distribution.
- -1.96 cuts off the bottom 2.5% of the distribution.
- As such, 95% of z-scores lie between -1.96 and 1.96.
  - This should ring some bells for you now when thinking about statistical significance at level 0.05.
- 99% of z-scores lie between -2.58 and 2.58,
- 99.9% of them lie between -3.29 and 3.29.

Normal Distribution in summary

- Many psychological/biological properties are normally distributed.
- This is very important for statistical inference (extrapolating from samples to populations).
- z-scores provide a way of:
  1. comparing scores on different raw-score scales;
  2. showing how a given score stands in relation to the overall set of scores.

Normal distribution in summary

- The logic of z-scores underlies many statistical tests:
  1. Scores are normally distributed around their mean.
  2. Sample means are normally distributed around the population mean.
  3. Differences between sample means are normally distributed around zero ("no difference").
- We can exploit these phenomena in devising tests to help us decide whether or not an observed difference between sample means is due to chance.

Z test

$$z = \frac{(\textit{the statistic of interest} - \textit{the expected value of the statistic under the null hypothesis})}{\textit{the standard error of the statistic}}$$

- A Z-Test is any statistical test for which the distribution of the test statistic under the null hypothesis is normal.
- For large samples Z-Test = T-Test = Chi-Square Test due to the Central Limit Theorem.

Central Limit Theorem

- Many real-world observations can be approximated by, and tested against, the same expected pattern: the normal distribution. In this familiar symmetric bell-shaped pattern, most observations are close to average, and there are fewer observations further from the average. The size of flowers, the physiological response to a drug, the breaking force in a batch of steel cables - these and other observations often fit a normal distribution.
- There are, however, many important things we would like to measure and test that do not follow a normal distribution. Household income doesn't – high values are much further from the average than low values are.
- But even when raw data does not fit a normal distribution, there is often a normal distribution 'lurking' within it. This makes it possible to still use the normal distribution to test ideas about non-normal data.

Central Limit Theorem

- For both normal and non-normal data we can take many independent random samples of size n from the population and if n is large enough then the distribution of the sample means will approach a normal distribution.
  - We can extrapolate from this to other statistics based on the mean in the linear model.
- How large is large enough?
  - The closer the population distribution is to a normal distribution, the fewer samples you need to take to demonstrate the theorem.
  - Populations that are heavily skewed or have several modes may require more and larger sample sizes.

Central Limit Theorem

- States that the sampling distribution of any statistic will be normal or nearly normal, if the sample size (no. of samples in the sampling distribution) is large enough.
- How large is "large enough"?
- The answer depends on two factors:
  - Requirements for accuracy.
    - The more closely the sampling distribution needs to resemble a normal distribution, the more sample points will be required.
  - The shape of the underlying population.
    - The more closely the original population resembles a normal distribution, the fewer sample points will be required.

Why to assess normality?

- For continuous data in a sample its all about bias.
- The closer our data is to normal the better.
  - Need to assess how far away from normal our data is.
  - Decide if the risk of bias is at a level we are happy to accept.
- For a single sample if 95% of our data falls within accepted conventions then we expect our test statistics for that sample to be within the 95% confidence interval we are working to under the central limit theorem.

Hypothesis Testing

Hypothesis Testing

- Hypothesis may concern an effect (e.g. correlation) in the population or a difference between groups in a population.
- The general goal of a hypothesis test is to rule out chance (sampling error) as a plausible explanation for the results from a research study.
- All hypothesis testing starts with the null hypothesis : that there is no effect or difference in the population.

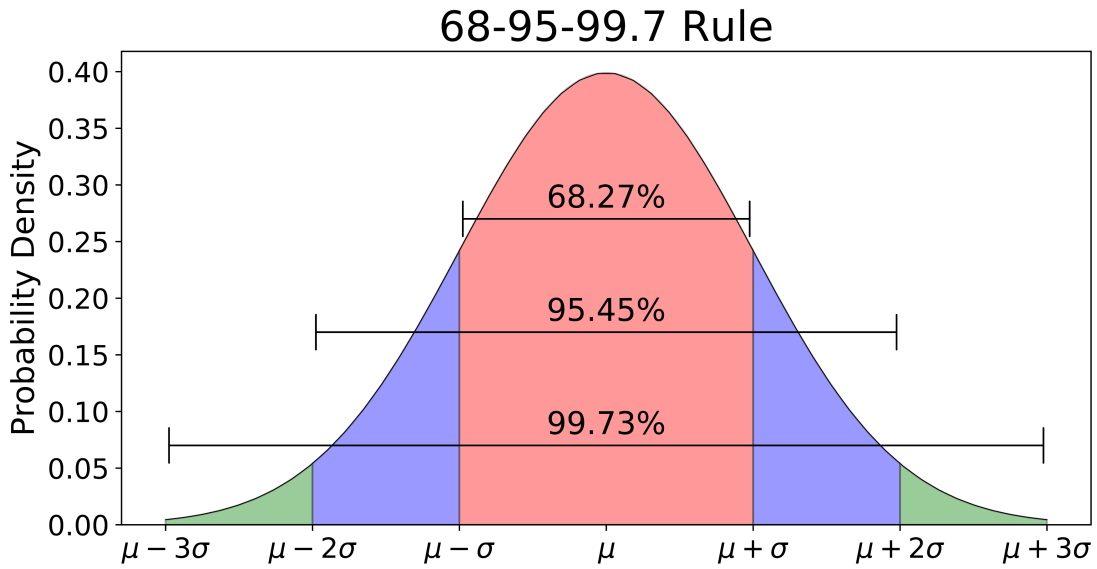
The Null Hypothesis, the Alpha Level, the Critical Region, and the Test Statistic

- State the hypotheses and select an α level or level of significance.
- The null hypothesis, H_0, always states there is no effect or no difference.
- The α level establishes a criterion, or "cut-off", for making a decision about the null hypothesis e.g. 95%, 99%.

The Null Hypothesis, the Alpha Level, the Critical Region, and the Test Statistic

- Compute the test statistic.
- The test statistic (e.g. a z-score) forms a ratio comparing the obtained difference between the sample and the hypothesized population versus the amount of effect or difference we would expect.
- This can be correlation, explanation of variance etc.

Critical Values



The critical values are set by moving toward the tails of the distribution. The higher the significance threshold, the more space under the tail.

Also, hypothesis testing can entail a one or two-tailed test, depending on if a hypothesis is directional (increase/decrease) in nature.

Steps of Testing and Significance

- The curve represents all of the possible outcomes for a given hypothesis.
- In this manner we move from talking about a distribution of data to a distribution of potential values for a sample of data.

The Null Hypothesis, the Alpha Level and the Test Statistic

- Generally a large value for the test statistic shows that the obtained mean difference (or variance difference or frequency difference) is more than would be expected if there is no effect/difference between groups.
- If it is large enough to be in the critical region (as computed by consulting relevant statistical tables) we conclude that the difference is significant or that there is a significant effect.
  - In this case we have evidence to support rejecting the null hypothesis.
- If it is relatively small, then the test statistic will have a low value.
  - In this case, we conclude that the evidence from the sample is not sufficient, and the decision is we do not have sufficient evidence to reject the null hypothesis or we retain it.

The Null Hypothesis, the Alpha Level, and the Test Statistic

- Can map to an equivalent probability.
- The phrase "statistical significance" means that these samples have a probability (p) that is less than the alpha level (so we report p value < 0.05, or p $=$ value achieved) if not less than critical value.

Errors in Hypothesis Tests

- Just because we find a statistically significant difference/effect does not necessarily indicate there is a causal relationship.
- Because the hypothesis test relies on sample data, and because sample data are not completely reliable, there is always the risk that misleading data will cause the hypothesis test to reach a wrong conclusion.
- Two types of error are possible.

Type I Errors

- Occur when the sample data appear to show an effect/difference when, in fact, there is none in the population.
- In this case the researcher will reject the null hypothesis and falsely conclude that there is an effect/difference.
- Type I errors are caused by unusual, unrepresentative samples.
- Just by chance the researcher selects an extreme sample with the result that the sample falls in the critical region even though there is no effect.
- The hypothesis test is structured so that Type I errors are very unlikely
- Specifically, the probability of a Type I error is equal to the alpha level.

Type II Errors

- Occurs when the sample does not appear to have an effect/difference when in fact this exists in the population.
- In this case, the researcher will fail to reject the null hypothesis.
- Type II errors are commonly the result of a very small effects/differences (not large enough to show up in the research study).

Power of a Hypothesis Test

- The **power** β of a hypothesis test is defined is the probability that the test will reject the null hypothesis when there is no effect.
- The power of a test depends on a variety of factors including the size of the effect and the size of the sample.

Possible Error ... ?

- Compare Type I and Type II error like this:
  - The only concern when you find statistical significance ($p < 0.05$) is Type I Error.
  - Is the difference between groups REAL or due to Random Sampling Error.
  - Thankfully, the p-value tells you exactly what the probability of that random sampling error is.
  - In other words, the p-value tells you how likely Type I error is.
- But, does the p-value tell you how likely Type II error is?
  - The probability of Type II error is better provided by Power.

Possible Error ... ?

- Probability of Type II error is provided by Power:
  - Statistical Power, also known as β (actually 1 – β).
- Power (Beta) is related to Alpha, but:
  - Alpha is the probability of having Type I error:
    - Lower number is better (i.e., 0.05 vs 0.01 vs 0.001).
    - Power is the probability of NOT having Type II error.
  - The probability of being right (correctly rejecting the null hypothesis):
    - Higher number is better (typical goal is 0.80).

Should it be statistically significant?

The most obvious thing you need to consider is if you REALLY should have found a statistically significant result?

- Just because you wanted your test to be significant doesn't mean it should be.
- This wouldn't be Type II error - it would just be the correct decision.

The Catch-22 of Power and P-values

- The larger your sample, the more likely you'll find statistically significant results.
  - Sometimes miniscule differences between groups or tiny correlations are 'significant'.
  - This becomes relevant once sample size grows to 100-150 subjects per group.
  - Once you approach 1000 subjects, it's hard not to find $p < 0.05$.
- Check previous lectures about what you can do to limit potential of type I or type II.

Measuring Effect Size

- A hypothesis test evaluates the statistical significance of the results from a research study.
- The hypothesis test is influenced not only by the size of the effect/difference but also by the size of the sample.
- Thus, even a very small effect/difference can be significant if it is observed in a very large sample.
- Finding a statistically significant result does not necessarily mean a large effect.
- It is recommended that the hypothesis test be accompanied by a measure of the **effect size**.
- Cohen's measures of effect size are used as standard.

Effect Size

- To get an idea of how 'important' a difference or association is, we can use Effect Size.
  - There are over 40 different types of effect size.
  - Depends on statistical test used.
- Effect size is like a 'descriptive' statistic that tells you about the magnitude of the association or group difference.
  - Not impacted by statistical significance.
  - Effect size can stay the same even if p-value changes.
  - Present the two together when possible.

Data Inspection and Preparation

Parametric vs 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.

Choices to be made before testing for statistical significance

- 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
  - 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:

- Weighting your data to correct for bias, address design effects, make sample more representative of the population.
- 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.
- 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.

Weighting

Use to:

- Correct for any known bias that may exist in the final sample.
  - E.g. due to non-response or sampling method chosen.
- 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.

Weighting - Correcting for Non-Response

- Suppose we undertake a survey within a university to see what proportions of students use the different facilities available in the library.
- We have selected a random sample for this purpose.
- We know that within the university the population is split 50:50 male: female.
- But in our sample we have 40:60 male: female split in respondents.

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 female students than male.

Weighting - Correcting for Non-Response

- If we want to have equal representation we can multiply all responses by female students by 50/60 (0.8333) and all responses by male students by 50/40 (1.25).
- This **weighting variable** will have two values to use: 0.8333 for females, 1.25 for males.
- By applying this weighting our proportions would be adjusted to 50:50 male: female.

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

Weighting - Correcting for Non-Response

- In reality weighting variables are used to correct for known biases in a number of 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 - Scaling up frequencies

- Suppose in our previous example we have 500 students in our sample and the total population was 10,000.
- To scale up the sample we would create a weighting variable that had the same value for all respondents $10{,}000/500 = 20$.
- This would increase all responses by a factor of 20.
- So if in our sample 15 students said they used a particular facility, with our weighting applied our frequency would show 300.

Weighting - Scaling up frequencies

- We can create a weighting variable to both correct for bias and scale up frequency to generate population estimates.
- So in our example we have a weighting variable of 0.833 or 1.25 for male and female.
- If we want to scale this up for our population we multiply these by 20 giving us 16.6666 for female and 25 for male.
- In the datasets earlychildhood.sav and afterschools.sav (Paul Connolly) this is what the variables FEWT and FWST are designed to do.
    - They correct for non-response and scale up to provide estimates for the national population as a whole.

Weighting - Correcting for design effects

- Due to cluster sampling:
  - User 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 actually doing interviews, but 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 similar to 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 effects

- Use multi-level modelling (examine each cluster).
  - Beyond the scope of this module.
- A compromise might be to 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.

Dealing with Weight Variables

- Check if a weighting variable is included in your dataset.
- If not:
  - Run your analysis but check what sampling method was used.
  - This may impact the way you report your findings (e.g. if clustering was used be careful when reporting results that are only just statistically significant or employ a stricter level of statistical significance).

Dealing with Weight Variables

- 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.
  - Scale up frequency – check if it is also for non-response.
    - If just used to scale up frequencies then run analysis on unweighted data unless your hypothesis concerns estimates for the population.
    - If used for both then you need to recompute to remove the scaling element.
  - NOTE: never conduct tests for statistical significance on scaled up frequencies.
    - It is likely you will always get a statistically significant result.

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 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.
- 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 similar to the distribution of the observed data.
2. Data missing completely at random (MCAR).
   - The distribution of the missing data does not depend on the distribution of the observed data either.
3. Data that are not missing at random (MNAR) .

Why are the Values Missing: The reason instructs the solution

- By Design—Completely Random
  - Missing Completely at Random (MCAR)
  - If the probability of a response depends on neither the observed nor the missing value that could have been collected or recorded, the missing data are missing completely at random.
- Intentionally Missing—Researcher controlled
  - Missing At Random (MAR)
  - Certain questions not asked of certain respondents/available for certain cases.
  - Some data dropped from analysis.

Why are the Values Missing

- Refusals - We may know mechanism
  - Adjusted for gender, race, education.
  - May be missing at random.
  - Otherwise, bias is likely w/o Auxiliary Variables.
- Missing because of "don't know" responses.
  - Between agree and disagree?
  - Can we impute a better value?
  - Should we?
- Missing by researcher error
  - May be missing completely at random.
  - May reflect researcher bias.
  - Perceived risk to researcher.
  - Missing observation worse than missing value.

Why are the Values Missing

Code reason value is missing.

- Depends on your domain but standards will apply e.g. -99.
- Treat each reason differently.

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 (Ymiss), 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 (e. g. in SPSS Missing Values Analysis).

Evaluating Missing Data

Missing data mechanism:

- Missing completely at random (MCAR)-Ignorable.
- Missing at random (MAR)-Conditionally ignorable.
- Missing not at random (MNAR)-Not ignorable.

Evaluating Missing Data

- 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.

Why is it a problem?

- In multivariate data a case will be excluded from the analysis if it is missing data for any variable included in the analysis.
- 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.

Typical Missing Data Treatments

- Deletion methods
  - Listwise deletion (complete case analysis)
    - Deletes the case if any variable is missing data.
  - Pairwise deletion (available case analysis)
    - Deletes case only when considering the variable for which data is missing, can still use it for other variables.
- 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.

Typical Missing Data Treatments

- 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 Treatments

- Maximum likelihood (ML)
  - Estimates summary statistics or statistical models using all available data.
- Multiple imputation
  - Imputes individual data values in multiple complete datasets, averaging the results of the statistical analyses across these datasets.

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).

Why do you see missing data treatments omitted from reports so frequently?

- Lack of awareness or familiarity.
- They are not convinced of the problems with older methods.
- The statistical literature on missing data is technically daunting.
- The techniques aren't incorporated into the standard statistical analysis procedures used by social scientists.
- Journal reviewers and editors have not required it.

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

Finding % missing

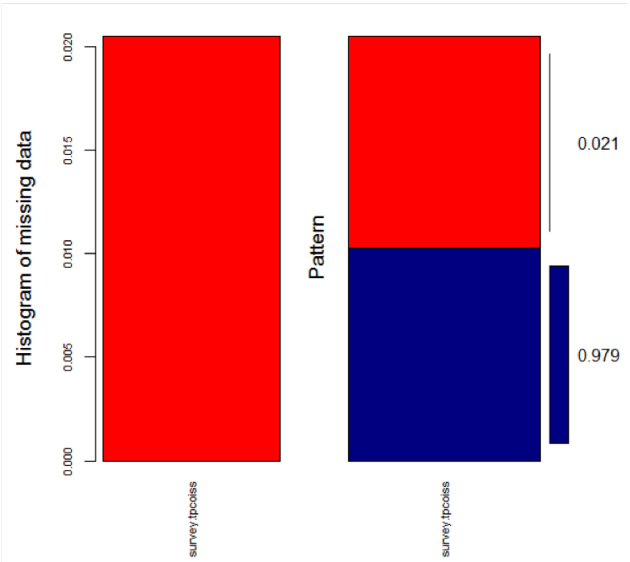
```
pMiss <- function(x){sum(is.na(x))/length(x)*100}
pMiss(survey$tpcois)
```

Output:

```
[1] 2.050114
```

=2.05% of survey$tpcois is missing

Handling Missing Data in R

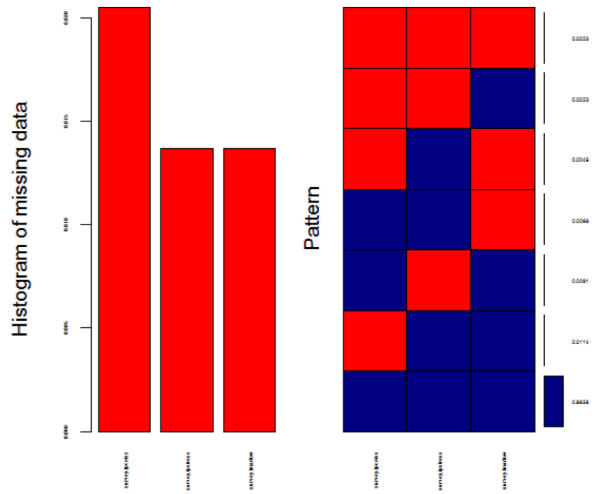


Looking at pattern

```
library(VIM)
aggr_plot <- aggr(ydata$tpcoiss,
col=c('navyblue','red'),
numbers=TRUE, sortVars=TRUE,
labels=names(ydata), cex.axis=.7,
gap=3, ylab=c("Histogram of
missing data","Pattern"))
```

```
Variable Count
survey.tpcoiss 0.02050114
```

Handling Missing Data in R



Variable	Count
survey.tpcoiss	0.02050114
survey.tpstress	0.01366743
survey.tmarlow	0.01366743

Looking at pattern

```
library(VIM)
aggr_plot <- aggr(ydata,
col=c('navyblue','red'),
numbers=TRUE, sortVars=TRUE,
labels=names(ydata), cex.axis=.3,
gap=3, cex.numbers=0.3,
ylab=c("Histogram of missing data",
"Pattern"))
```

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 relvant variables.
  - Using filter from the dplyr library.
- Most modelling functions offer you an option for handling missing data.
  - E.g. na.rm=True
- Imputing missing data.
  - Hmisc package contains several functions that are helpful for missing value imputation (agreImpute(), impute() and transcan())
  - mitools package

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.

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 are cases that have an unusual value for a single variable.
- Multivariate outliers are cases that have an unusual combination of values for a number of variables.
  - The value for any of the indvidual 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 of the scores 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 level variables, and to ordinal level variables that are treated as metric.
- It does not apply to nominal level variables.

Tools for Assessing Normality

- Histogram and Boxplot
- Normal Quantile Plot (also called Normal Probability Plot)
- Goodness of Fit Tests
  - Shapiro-Wilk Test (small dataset).
  - Kolmogorov-Smirnov Test (larger dataset).
  - These are unreliable if your dataset is large.
- Recognized heuristics
  - Standardized skewness and kurtosis.
  - Percentage of standardized scores falling within ranges (see previous slides and previous lectures for bounds).

Transformations to Improve Normality (removing skewness)

- Many statistical methods require that the numeric variables you are working with have an approximately normal distribution.
- Reality is that this is often times not the case.
- One of the most common departures from normality is skewness.
- There are many different types of transformation available.

Transforming Data

- Many of the statistical tests (parametric tests) are based on the assumption that the data are normally distributed.
- However, if we actually plot the data from a study, we rarely see perfectly normal distributions.
- Most often, the data will be skewed to some degree or show some deviation from mesokurtosis.
- Two questions immediately arise:
  - A) Can we analyze these data with parametric tests and. If not,
  - B) Is there something we can do to the data to make them more normal?

Transforming Data

- According to some researchers, sometimes violations of normality are not problematic for running parametric tests.
- When a variable is not normally distributed (a distributional requirement for many different analyses), we can create a transformed variable and test it for normality. If the transformed variable is normally distributed, we can substitute it in our analysis.

Transforming Data

Perform a mathematical operation on each of the scores in a set of data, and thereby converting the data into a new set of scores which are then employed to analyze the results of an experiment.

Tukey's Ladder of Powers

- To remove **right skewness** we typically take the square root, cube root, logarithm, or reciprocal of a the variable etc., i.e. V .5, V .333, $log_10(V)$ (think of V_0) , V -1, etc.
- To remove **left skewness** we raise the variable to a power greater than 1, such as squaring or cubing the values, i.e. V 2, V 3, etc.

Transforming for Postivie Skew

- Square roots, logarithmic, and inverse $(1/X)$ transforms "pull in" the right side of the distribution in toward the middle and normalize right (positive) skew.
- Inverse transforms are stronger than logarithmic, which are stronger than roots.
- A logarithmic transformation may be useful in normalizing distributions that have more severe positive skew than a square-root transformation.

Transforming for Negative Skew

- If skewness is actually negative, "flip" the curve over, so the skew left curves become skewed right, allowing us to use the transformation procedures of positively skewed distributions.
- Power transformation $Y = (X)p$
- Arcsine transformation - The arcsine of a number is the angle whose sine is that number.

Transforming Data

If transformation does not bring data to a normal distribution, the investigators might well choose a **nonparametric** procedure that does not make any assumptions about the shape of the distribution.

Transforming Data

- Transformations are obtained by computing a new variable.
  - R, either transform and save variable to dataset or apply transformation within whatever statistical formula.
- Functions
  - R functions
    - Logarithmic transformation log_10
    - Square Root sqrt
- For each of these calculations, there may be data values which are not mathematically permissible.
  - For example, the log of zero is not defined mathematically, division by zero is not permitted, and the square root of a negative number results in an "imaginary" value.
  - You need to decide what to do about this.

Good Information Source

- Andy Field's The Beast of Bias.
- `https://www.discoveringstatistics.com/repository/exploringdata.pdf`