MSc - Data Mining

Topic 01 : Module Overview

Part 04 : A Review of Statistical Concepts

Dr Bernard Butler and Dr Kieran Murphy

Department of Computing and Mathematics, WIT.
(bbutler@tssg.org and kmurphy@wit.ie)

Spring Semester, 2021

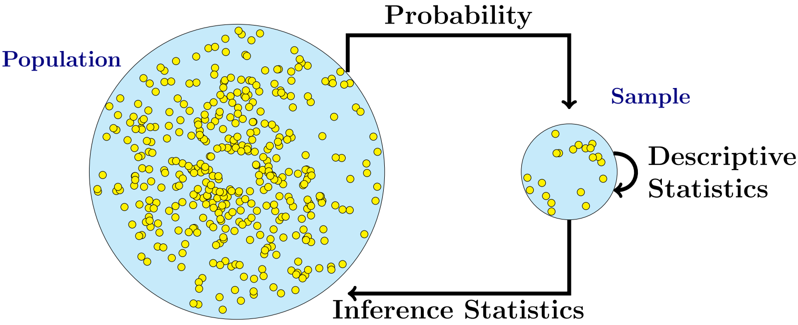
Outline

- Probability in five minutes (probably)
- Type of Data
- Descriptive statistics

Part I

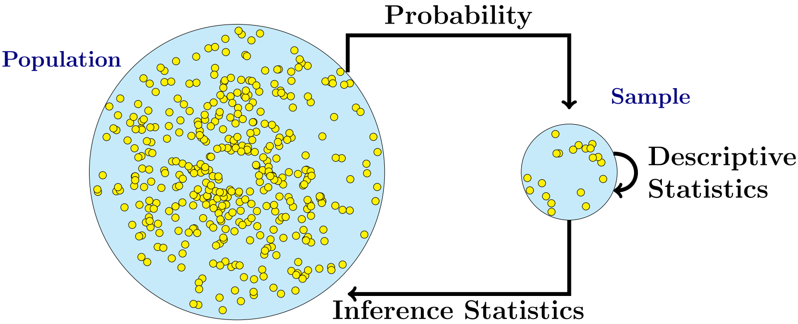
Fundamental Concepts

“Central Dogma” of Statistics I



- The population is the group of all items of interest.
  - Frequently very large or even sometimes infinite.
  - Population attributes are called population parameters.
- A sample is a subset of the population drawn from the population.
  - Potentially very large, but usually of orders smaller than the population.
  - Sample attributes are called sample statistics.

“Central Dogma” of Statistics II



- Descriptive Statistics is a set of methods of organising, summarising, and presenting data in a convenient and informative way.
- Inferential statistics is a set of methods, to draw conclusions or inferences about characteristics of populations based on data from a sample.

> So we use the sample statistics to make inferences about the population parameters

Descriptive vs Inference Statistics

We use sample statistics to make inferences about population parameters.

- Therefore, we can make an estimate, prediction, or decision about a (unknown) population based on (known) sample data.

Rationale

- Large populations make investigating each member impractical and expensive.
- Easier and cheaper to take a sample and make estimates about the population from the sample.

Warning

Such conclusions and estimates are not always going to be correct. For this reason, we build into the statistical inference “measures of reliability” namely confidence level and significance level.

— **even when all of the assumptions made in constructing the sample are true!.**

How good is your data? I

Validity

A valid measurement is one which actually measures what it claims to measure.

- Unemployment figures are validly measured using the Labour Force Survey not the Live Register

Reliability

A reliable measurement is one which will give approximately the same result time after time, when taken on the same individual or object.

- Most physical measurements are reliable, for example measuring your weight using a bathroom scales.
- Some measurements may be reliable but not necessarily valid.
  - Are exams reliable measuring devices?
  - Are exams results valid measurements of intelligence?

How good is your data? I

Validity

A valid measurement is one which actually measures what it claims to measure.

- Unemployment figures are validly measured using the Labour Force Survey not the Live Register

Reliability

A reliable measurement is one which will give approximately the same result time after time, when taken on the same individual or object.

- Most physical measurements are reliable, for example measuring your weight using a bathroom scales.
- Some measurements may be reliable but not necessarily valid.
  - Are exams reliable measuring devices?
  - Are exams results valid measurements of intelligence?

How good is your data? I

Validity

A valid measurement is one which actually measures what it claims to measure.

- Unemployment figures are validly measured using the Labour Force Survey not the Live Register

Reliability

A reliable measurement is one which will give approximately the same result time after time, when taken on the same individual or object.

- Most physical measurements are reliable, for example measuring your weight using a bathroom scales.
- Some measurements may be reliable but not necessarily valid.
  - Are exams reliable measuring devices?
  - Are exams results valid measurements of intelligence?

How good is your data? II

Bias

Sometimes when measurements are made a systematic error is made which underestimates or overestimates the true value. Such a measurement is called a biased measurement.

- Suppose your bathroom scales always overestimated your weight.
- Car speedometers are deliberately biased to overestimate a car's real speed.

Variability

Some datasets are more variable than others.

- A dataset consisting of the ages of 100 students in WIT will be less variable than a dataset consisting of the ages 100 randomly chosen Irish people.

How good is your data? II

Bias

Sometimes when measurements are made a systematic error is made which underestimates or overestimates the true value. Such a measurement is called a biased measurement.

- Suppose your bathroom scales always overestimated your weight.
- Car speedometers are deliberately biased to overestimate a car's real speed.

Variability

Some datasets are more variable than others.

- A dataset consisting of the ages of 100 students in WIT will be less variable than a dataset consisting of the ages 100 randomly chosen Irish people.

How good is your data? II

Bias

Sometimes when measurements are made a systematic error is made which underestimates or overestimates the true value. Such a measurement is called a biased measurement.

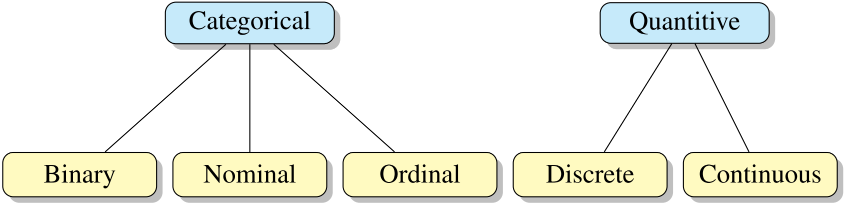
- Suppose your bathroom scales always overestimated your weight.
- Car speedometers are deliberately biased to overestimate a car's real speed.

Variability

Some datasets are more variable than others.

- A dataset consisting of the ages of 100 students in WIT will be less variable than a dataset consisting of the ages 100 randomly chosen Irish people.

Types of Data



2 Categories	More Categories	Order matters	Numerical	+ Uninterrupted
TRUE/FALSE YES/NO, etc	Gender, Religion, Favourite Team, etc		**CAN** be represented by integers on the number line	Takes values from intervals on the number line
			Number of students in this class	Duration of this class

Cannot order

Cannot add / subtract / multiple or divide

Dimensionality of Data Sets

We will frequently use a table to represent our data:

- Each row represents a observation / subject / case.
    - Univariate — single measurement per subject.
    - Bivariate — two measurements per subject.
    - Multivariate — Multiple measurements per subject.
- Each column represents a variable / attribute.
    - Each variable is then Binomial/Nominal/Ordinal/Discrete/Continuous

Tips dataset

```
df.head(10)
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
5	25.29	4.71	Male	No	Sun	Dinner	4
6	8.77	2.00	Male	No	Sun	Dinner	2
7	26.88	3.12	Male	No	Sun	Dinner	4
8	15.04	1.96	Male	No	Sun	Dinner	2
9	14.78	3.23	Male	No	Sun	Dinner	2

```
df.shape
```

```
(244, 7)
```

```
df.dtypes
```

```
total_bill    float64
tip           float64
sex          category
smoker       category
day          category
time         category
size            int64
dtype: object
```

Terminology to Describe Variables

Central Tendency

- Where, “generally” are the scores?

mean, median

- Is there a “meaningful” (subjective) characterisation of where “most” scores are situated?

mode

Dispersion

- How “spread out” are the scores?

standard deviation, variance, range, IQR

- Is it not meaningful to talk about a “typical” observation?

Shape

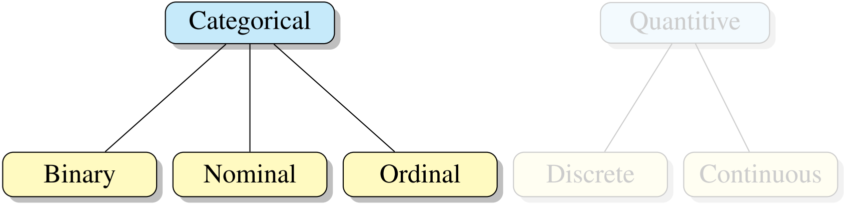
Do the observations appear to be

- Unimodal (one most-likely score, others less likely)?
- Symmetric or Skewed?
- Are there outliers —- atypical/extreme values?
- Are there missing values?

Part II

Analysing Categorical Data

Types of Data — Categorical Data



Categorical

Quantitive

Binary

Nominal

Ordinal

Discrete

Continuous

Binary	Nominal	Ordinal	Discrete	Continuous
2 Categories	More Categories	Order matters	Numerical	+ Uninterrupted
TRUE/FALSE YES/NO, etc	Gender, Religion, Favourite Team, etc		**CAN** be represented by integers on the number line	Takes values from intervals on the number line
			Number of students in this class	Duration of this class

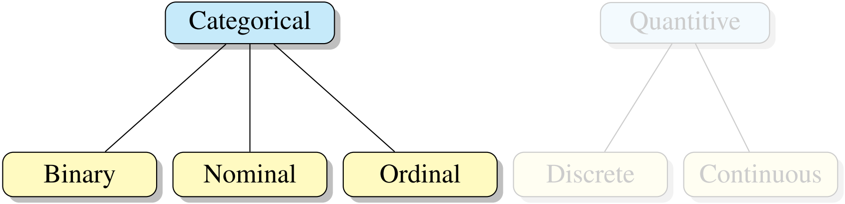
Cannot order

⇑

Cannot add / subtract / multiple or divide

⟹ Limited to methods that only deal with counts/frequencies

Types of Data — Categorical Data



Categorical

Quantitive

Binary

Nominal

Ordinal

Discrete

Continuous

2 Categories

TRUE/FALSE
YES/NO,
etc

More Categories

Gender,
Religion,
Favourite Team, etc

Order matters

Nun

CA
sent
on t

Nun
in th

Graphical Methods

- Bar chart
- Pie chart

Numerical Methods

- Frequency counts
- Number of unique values

⇑

Cannot order

Cannot add / subtract / multiple or divide

⟹ Limited to methods that only deal with counts/frequencies

Example: tips.day

```
df.day.nunique()
```

```
4
```

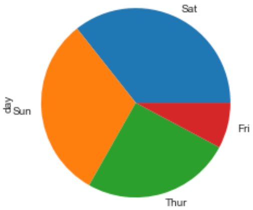
```
df.day.value_counts()
```

```
Sat     87
Sun     76
Thur    62
Fri     19
Name: day, dtype: int64
```

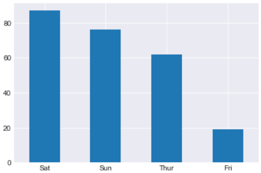
```
df.day.describe()
```

```
count     244
unique      4
top       Sat
freq       87
Name: day, dtype: object
```

```
df.day.value_counts().plot(kind='pie');
```



```
df.day.value_counts().plot(kind='bar')
plt.xticks(rotation=0);
```



These are horrible visualisations — pie chart does not show counts, and what about the category order in the bar charts

Comments

- Pie charts are often over used (especially by newspapers) but they have significant limitations:
    - Comparing two pie charts is problematic since people find it harder to compare angles to lengths.
    - If one uses relative frequencies (or percentages) then the number of observations needs to be clearly stated.
    - Should not be used if the number of categories is large.

> *To my mind, the best use of a pie chart is when you have one value that is overwhelmingly larger than the rest and you don't want the audience to focus on the actual values, but just bamboozle them with the overwhelming size of the leading segment. Of course, this seems to come close to embracing the old adage, "There are lies, damn lies and statistics."*
>
> *— www.juiceanalytics.com/writing/writing/the-problem-with-pie-charts*

See also

- www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=00018S

  (Edward Tufte is a statistician and author of 4 books on data visualisation.
- "Save the Pies for Dessert" article at www.perceptualedge.com/articles/08-21-07.pdf

Part III

Analysing Quantitative Data

Types of Data — Quantitative Data

Categorical

Graphical Methods

- Histogram — for shape, spread
- Boxplot — for shape, location, outliers
- …

Numerical Methods

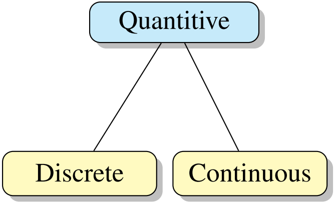
- mean, median — centre
- standard deviation, range — spread
- z-scores, quantiles — location
- …

YES/NO, etc

Religion, Favourite Team, etc

Cannot order

Cannot add / subtract / multiple or divide



Numerical	Numerical+Uninterrupted
CAN be represented by integers on the number line	Takes values from intervals on the number line
Number of students in this class	Duration of this class

Can add / subtract / multiple and divide (even for discrete data)

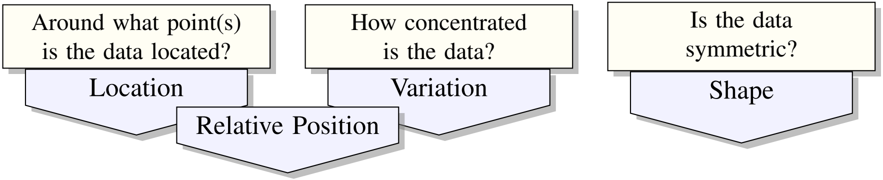
Numerical Methods for Quantitive Data

Around what point(s) is the data located?

How concentrated is the data?

Is the data symmetric?

Numerical Methods for Quantitive Data



Around what point(s) is the data located?

Location

How concentrated is the data?

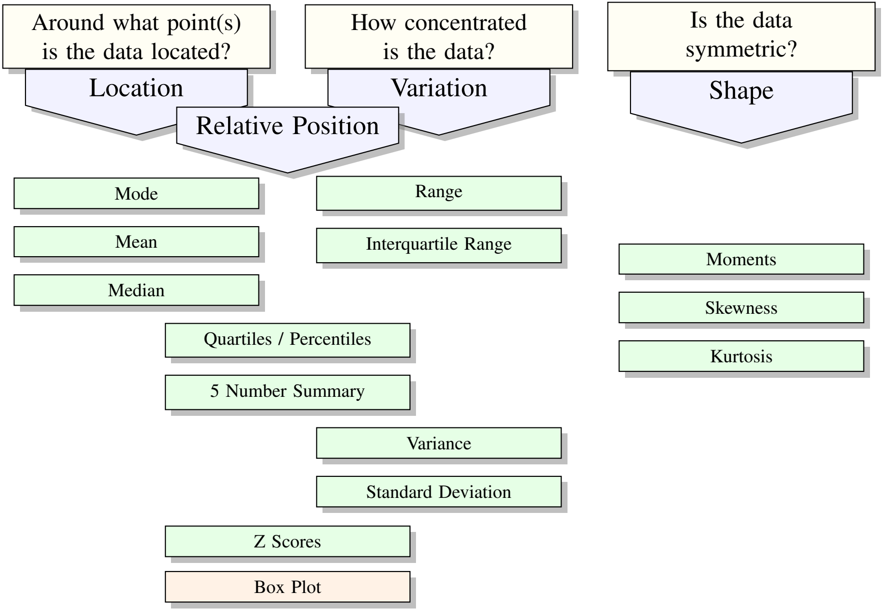
Variation

Relative Position

Is the data symmetric?

Shape

Numerical Methods for Quantitive Data



Around what point(s) is the data located?

Location

How concentrated is the data?

Variation

Is the data symmetric?

Shape

Relative Position

Mode

Mean

Median

Range

Interquartile Range

Quartiles / Percentiles

5 Number Summary

Variance

Standard Deviation

Z Scores

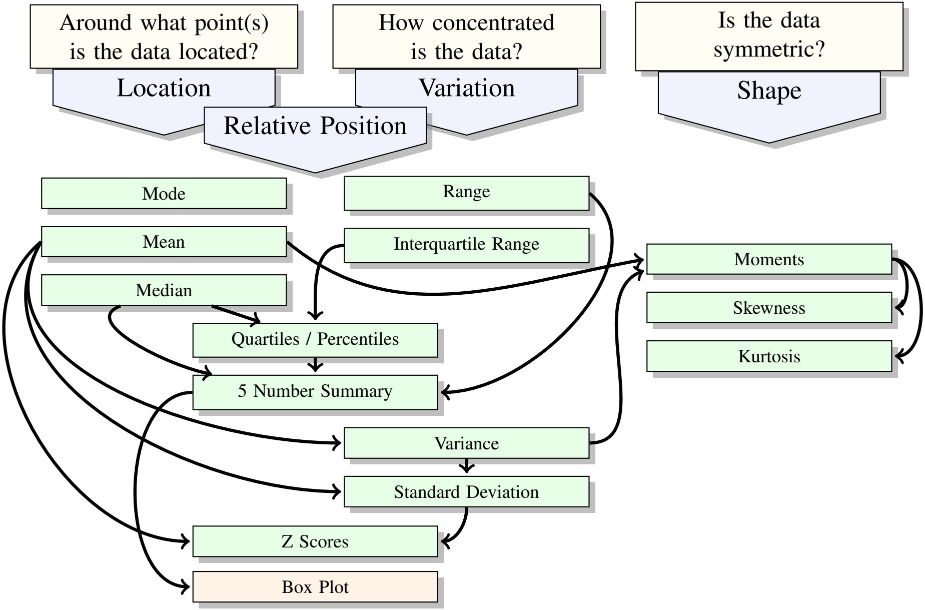
Box Plot

Moments

Skewness

Kurtosis

Numerical Methods for Quantitive Data



Around what point(s) is the data located?

Location

How concentrated is the data?

Variation

Is the data symmetric?

Shape

Relative Position

Mode

Mean

Median

Range

Interquartile Range

Quartiles / Percentiles

5 Number Summary

Variance

Standard Deviation

Z Scores

Box Plot

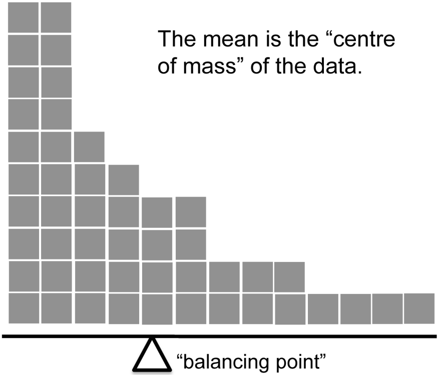
Moments

Skewness

Kurtosis

Mean, Mode

Properties of Arithmetic Mean

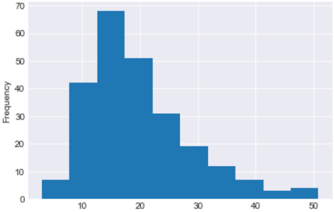


- ✘ The mean is sensitive to presence of outliers — so is not a robust metric.
- ✔ The trimmed mean addressed the outlier issue but now need to decide on appropriate trim value.
- ✘ The resulting value may be meaningless "average family has 2.3 children".
- ✔ Uses all of the available information in computing the metric.
- ✔ Has nice mathematical properties.*

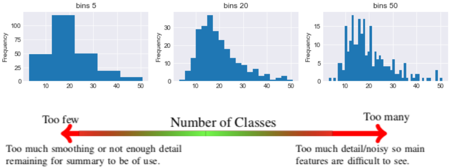
*Not an issue for you but it keeps mathematicians up at night.

Histogram and Box-Plot

```
df.total_bill.plot(kind='hist');
```



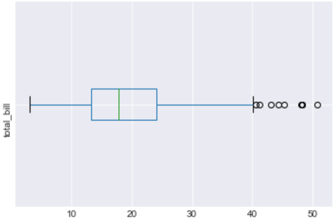
- Parameter `bin` controls level of granularity.



Too much smoothing or not enough detail remaining for summary to be of use.

Too much detail/noisy so main features are difficult to see.

- Distribution is uni-modal
- and right skewed (longer tail to the right)
- Boxplot clearly show outliers (dots), skewness



```
df.total_bill.plot(kind='box', vert=False)
plt.yticks(rotation=90);
```