Lecture 4 - Levels of Measurement, Types of Measurement and Scale Scores

Tony Tan

University of Oslo

Thursday, 27 October 2022

Table of Contents

Last time

- Review of some essential probability theory required for the course
- Concepts of expected value, variance, covariance
- Statistical reasoning and the concepts of parameters, estimators and estimates

Today

- Levels of measurement – nominal, ordinal, interval, ratio
- Mean, median, mode
- Criterion-referenced and norm-referenced tests
- Different ways to define test scores
- Some exercises

Table of Contents

Measurement

- The purpose of measurement is to quantify an attribute
- We assign a number to an item response
- Based on item responses we assign a test score
- The test score is determined by the item responses but the reverse is typically not true – the same test score can be obtained from different item responses

Ceiling and floor of a test

- Any test has a floor and ceiling in that there is a range of levels of the attribute that the test can actually measure
- With a mathematics test containing 10 binary items, some test-takers may score 0 and some may score 10
- The underlying construct can be too low or too high for the test to be able to measure it
- Example: giving a university calculus exam to a class of second-graders
- Example: giving a diagnostic test for dementia to a group of university students

Measurement scales – levels of measurement

- Certain physical measurements have specific properties
- We can think of a measurement of length as being twice as large as another measurement of length
- We can think of the interval between 5 °C and 6 °C as being the same as that between 25 °C and 26 °C
- To what extent to item scores and test scores fulfill these properties?

Ratio scale

- Some measurements have the property that the doubling of the measurement can be be interpreted as being twice as large
- For example, 20 cm is twice as long as 10 cm
- We of course also have that the difference between 30 cm and 20 cm is the same as the difference between 20 cm and 10 cm

Interval scale

- The difference between two observations is interpreted in the same way
- Consider temperature as measured by celsius degrees (°C)
  - The difference between 3 °C and 2 °C and the difference between 1003 °C and 1002 °C has the same interpretation
  - We can of course also say that 3 °C is higher than 2 °C
  - However, we can't say that 10 °C is twice as warm as 5 °C

Ordinal scale

- Some measurements have an ordering, but do not have the property of equal intervals
- Mohs scale for hardness of minerals
  - The hardness is determined by which mineral scratches another mineral
  - A mineral A is harder than a mineral B if A scratches B
  - However, if A scratches B and B scratches C, we wouldn't be able to say that THE difference in hardness between A and B is the same as the difference in hardness between B and C
  - We could of course also not say that A is twice as hard as B or as C

Nominal scale

- If there is no ordering of the measurements, we have a nominal scale
- It won't make sense to speak of a measurement being twice as large, having equal distances or having any order

The scale of item scores?

- Consider a Likert scale
  - Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree
  - If we assign integer scores from 1 to 5 to the categories, we are imposing a scale on the items
    - Is Agree twice as large as Disagree?
    - Is the difference between Strongly agree and Agree the same as that between Agree and Neither agree nor disagree?
  - It seems that the scale is actually ordinal

The scale of test scores?

- Consider a test score defined by the number of items correct on a 20 item test with binary item scores
- Does this test score have the property of equal intervals?
- Is a score of 10 twice as good as a score of 5?

Different test scores

- Consider again a 20-item test with binary item scores
- We are not forced to define a test score which takes integer values from 0 to 20
- We can apply a transformation, such as the test score to the power of 2 or another transformation
- The choice of metric is the choice of how numbers are assigned to observations

Table of Contents

The expected value (revision)

Let X be a discrete R.V. that can take k different values with probabilities $p_1, \ldots, p_k$. Then

$$\mathbb{E}(X) = \sum_{i=1}^{k} x_i \, p_i.$$

The median

- The median denotes the value c such that

$$\mathbb{P}(X \leqslant c) \geqslant 0.5 \text{ and } \mathbb{P}(X \geqslant c) \geqslant 0.5.$$

- We can think of the median as the value of the R.V. X which lies in the middle of the probability mass function or the probability density function .
- Such a measure will give a better idea of the typical value of X when the density of X is skewed .

The mode

- For a discrete R.V. X, the mode denotes the value of X which has the highest probability mass associated with it.
- For a continuous R.V., the mode denotes the value of X for which the density $f(x)$ reaches its highest value.
- Note that the mode may not be unique.

Example

Let X be a discrete R.V. defined by

X	1	2	3	4
$\mathbb{P}(X)$	0.1	0.5	0.2	0.2

What is a) the expected value, b) the median and c) the mode of X?

a $\mathbb{E}(X) = 1 \times 0.1 + 2 \times 0.5 + 3 \times 0.2 + 4 \times 0.2 = 2.5$.

b Since $\mathbb{P}(X \leqslant 2) = 0.6$ and $\mathbb{P}(X \geqslant 2) = 0.9$, the median of X is 2.

c Since $\mathbb{P}(X = 2)$ reaches the higest value 0.5, the mode of X is 2.

Symmetric distributions

- Let X be a R.V. taking values 1, 2 and 3 such that

$$\mathbb{P}(X = 1) = 0.2, \ \mathbb{P}(X = 2) = 0.6, \ \text{and} \ \mathbb{P}(X = 3) = 0.2.$$

  This is a symmetric distribution – whose mean and median coincide.
- Consider a $\mathcal{N}(0, 1)$ distribution. This is also a symmetric distribution, whose mean, median, and mode are all equal.

The sample mean

Let $\{x_i\}_{i=1}^n$ denote a sample. The sample mean is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In R, we can calculate the sample mean of a vector $\boldsymbol{x}$ by typing

```
x <- c(12, 8, 19, 13, 8)
mean(x)

## [1] 12
```

The sample median

Let $\{x_i\}_{i=1}^{n}$ denote a sample. The sample median is the middle value of the observations.
The median may be a more suitable measure of central tendency when the distribution is skewed.
In R, we can find the sample median by typing

```
median(x)

## [1] 12
```

The sample mode

The sample mode is equal to the most common occurence from a set of observations. It is meaningful for categorical data as an indicator of the most frequent observation.
We can find the sample mode by using the R function table():

```
table(x)

## x
##  8 12 13 19
##  2  1  1  1
```

The mode is 8 since it has the highest count.

Table of Contents

Criterion-referenced measurement

- Test scores are interpreted as an absolute measure of an underlying construct
- If an individual reaches a particular score, the individual is seen as having mastered this construct or having fulfilled this level of proficiency
- A test designed in this way is said to be a criterion-reference measurement
  - Driver licence test
  - Registered Nurse test
  - Citizenship tests (*Prøve i samfunnskunnskap*)

Norm-referenced measurement

- Test scores are seen as indicative of the level of proficiency with reference to a particular population
- The norms have been established from previous research with individuals from the same population
  - College entrance exams, SAT, ACT
  - Graduate Record Examinations (GRE)
  - Wechsler Intelligence Scale for Children

The duality of reference

- The same test can be interpreted in a criterion-referenced manner and in a norm-referenced manner
- Original motive for the Binet and Simon intelligence test was to identify children with intellectual disabilities – having nothing to do with national norms for intelligence
- Later the intelligence test has been used to refer to a population norm

Table of Contents

Standardised scores (Z-scores)

For ease of interpretation, we can standardise raw scores Y into $\mathcal{N}(0, 1)$:

$$Z = \frac{Y - \mu_Y}{\sigma_Y}.$$

If the resulting standardised scores follow a distribution similar to the standard normal distribution , we can interpret individual scores as percentiles of this distribution.

Linearly transformed scores

Wechsler Intelligence Test has mean 100 and standard deviation 15. This distribution can be obtained from the standard normal by a linear transformation:

$$S = cZ + a,$$

where $c = 15$ and $a = 100$.

Normalised scores and other non-linearly transformed scores

- If the raw scores are not normally distributed, it is often desirable to rescale the scores to approximately normal
- We can apply square-root or other non-linear transformations, or link the percentiles of the raw-score distribution to those from the normal distributions
- Criterion-referenced tests often have a threshold for certain levels of skill

Permissible operations

- Certain statistical operations are inappropriate for certain types of item or test scores
- If we have an ordinal level of measurement we should compute its median rather than the mean
- See Chapter 4 and 18 of McDonald (1999) for more details

Statistical tests and levels of measurement

- A hypothesis test of equal means requires approximately normally distributed scale scores (t-test)
  - Does not require the original scale to have interval level properties
- Hypothesis tests can be affected by a change of scale (see p. 60–61, McDonald (1999))

Review

- Levels of measurement
- Measures of central tendency
- Criterion-referenced and norm-referenced testing
- Linear and non-linear transformations of test scores

Table of Contents

L4 Task 1

Consider a R.V. X that takes values 1, 2, 3 and 4 with corresponding probabilities 0.1, 0.2, 0.4 and 0.3.

a What is its median?

b What is its mode?

L4 Task 2

Which of the following are symmetric distributions?

a A t_{49}-distribution.

b A χ^2_1-distribution.

c For the R.V. X with PMF

X	1	2	3	4	5
$\mathbb{P}(X)$	0.3	0.15	0.1	0.15	0.3