Univariate Parametric & Nonparametric Statistics & Statistical Tests

- Kinds of variables & why we care
- Univariate stats
  - qualitative variables
  - parametric stats for ND/Int variables
  - nonparametric stats for ~ND/~Int variables
- Univariate statistical tests
  - tests qualitative variables
  - parametric tests for ND/Int variables
  - nonparametric tests for ~ND/~Int variables
  - of normal distribution shape for quantitative variables

Kinds of variable → The "classics" & some others …

Labels

- aka → identifiers
- values may be alphabetic, numeric or symbolic
- different data values represent unique vs. duplicate cases, trials, or events
- e.g., UNL ID#

Nominal

- aka → categorical, qualitative
- values may be alphabetic, numeric or symbolic
- different data values represent different "kinds"
- e.g., species

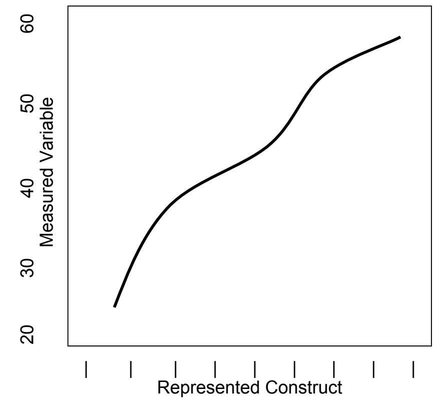
Ordinal

- aka → rank order data, ordered, seriated data
- values may be alphabetic or numeric
- different data values represent different "amounts"
  - only "trust" the ordinal information in the value
  - don't "trust" the spacing or relative difference information
- has no meaningful "0"
  - don't "trust" ratio or proportional information
- e.g., 10 best cities to live in
  - has ordinal info → 1^{st} is better than 3^{rd}
  - no interval info → 1^{st} & 3^{rd} not "as different" as 5^{th} & 7^{th}
  - no ratio info → no "0^{th} place"
  - no prop info → 2^{nd} not "twice as good" as 4^{th}
  - no prop dif info→ 1^{st} & 5^{th} not "twice as different" as 1^{st} & 3^{rd}

Interval

- aka → numerical, equidistant points
- values are numeric
- different data values represent different "amounts"
  - all intervals of a given extent represent the same difference anywhere along the continuum
  - "trust" the ordinal information in the value
  - "trust" the spacing or relative difference information
- has no meaningful "0" (0 value is arbitrary)
  - don't "trust" ratio or proportional information
- e.g., # correct on a 10-item spelling test of 20 study words
  - has ordinal info → 8 is better than 6
  - has interval info → 8 & 6 are "as different" as 5 & 3
  - has prop dif info→ 2 & 8 are "twice as different" as 3 & 5
  - no ratio info → 0 not mean "can't spell any of 20 words"
  - no proportional info → 8 not "twice as good" as 4

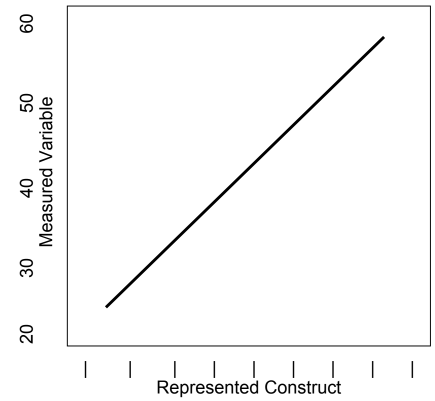
Ordinal Measure



Positive monotonic trace

"more means more but doesn't tell how much more"

Interval Measure



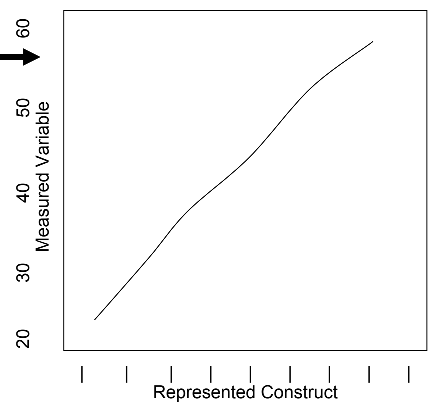
Linear trace

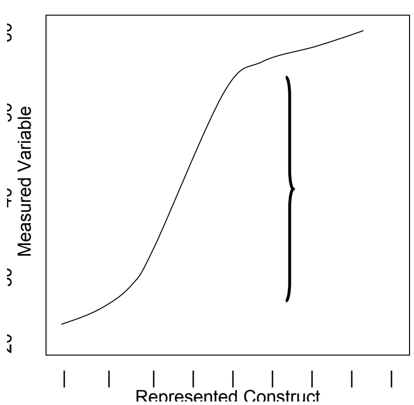
"more how much more"

y = mx + c

"Nearly" Interval Scale

- "good" summative scales
- how close is "close enough"





"Limited" Interval Scale

- provided interval data only over part of the possible range of the scale values / construct
- summative/aggregated scales

Binary Items

Nominal

- for some constructs different values mean different kinds
- e.g., male = 1   famale = 2

Ordinal

- for some constructs can rank-order the categories
- e.g., fail = 0   pass = 1

Interval

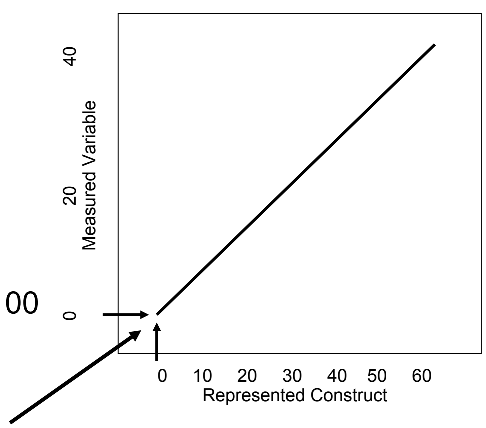
- only one interval, so “all intervals of a given extent represent the same difference anywhere along the continuum”

So, you will see binary variables treated as categorical or numeric, depending on the research question and statistical model.

Ratio

- aka → numerical, “real numbers”
- values are numeric
- different data values represent different “amounts”
  - “trust” the ordinal information in the value
  - “trust” the spacing or relative difference information
- has a meaningful “0”
  - “trust” ratio or proportional information
- e.g., number of treatment visits
  - has ordinal info → 9 is better than 7
  - has interval info → 9 & 6 are “as different” as 5 & 2
  - has prop dif info→ 2 & 8 are “twice as different” as 3 & 5
  - has ratio info → 0 does mean “didn’t visit”
  - has proportional info → 8 is “twice as many” as 4

Ratio Measure



Pretty uncommon in Psyc & social sciences

- tend to use arbitrary scales
- usually without a zero
  - 20 5-point items → 20-100

Linear scale & “0 means none”

Linear trace w/ 0

“more how much more”

$y = mx + c$

Kinds of variables → Why we care …

Reasonable mathematical operations

Nominal → ≠ =

Ordinal → ≠ < = >

Interval → ≠ < = > + - (see note below about * /)

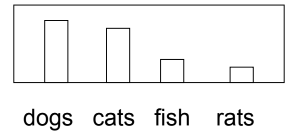
Ratio → ≠ < = > + - * /

Note: For interval data we cannot * or / numbers, but can do so with differences. E.g., while 4 can not be said to be twice 2, 8 & 4 are twice as different as are 5 & 3.

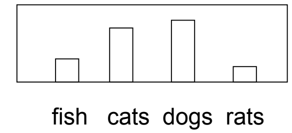
Data Distributions

We often want to know the "shape" of a data distribution.

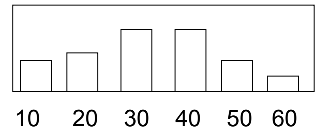
Nominal → can't do → no prescribed value order

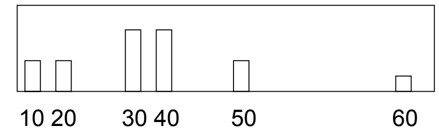


vs.

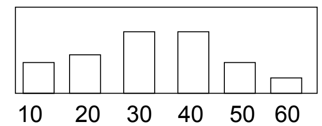


Ordinal → can't do well → prescribed order but not spacing





Interval & Ratio → prescribed order and spacing



Univariate Statistics for qualitative variables

Central Tendency – "best guess of next case's value"
- Mode -- the most common score(s)
- uni-, bi, multi-modal distributions are all possible

Variability – "index of accuracy of next guess"
- # categories
- modal gender is more likely to be correct guess of next person than is modal type of pet – more categories of the latter

Shape – symmetry & proportional distribution
- doesn't make sense for qualitative variables
- no prescribed value order

Parametric Univariate Statistics for ND/Int variables

Central Tendency – "best guess of next case's value"

- mean or arithmetic average → $M = \Sigma X / N$
- 1^{st} moment of the normal distribution formula
- since ND unimodal & symetrical → mode = mean = mdn

Variability – "index of accuracy of next guess"

- sum of squares → $SS = \Sigma(X - M)^2$
- variance → $s^2 = SS / (N-1)$
- standard deviation → $s = \sqrt{s^2}$
- std preferred because is on same scale as the mean
- 2^{nd} moment of the normal distribution formula
- average extent of deviation of each score from the mean

Parametric Univariate Statistics for ND/Int variables, cont.

Shape – "index of symmetry"

- skewness → $$\frac{\Sigma (X - M)^3}{(N - 1) * s^3}$$
- 3^{rd} moment of the normal distribution formula
- 0 = symmetrical, + = right-tailed, - = left-tailed
- can't be skewed & ND

Shape –"index of proportional distribution"

- kurtosis → $M = \Sigma X / N$ $$\frac{\Sigma (X - M)^4}{(N - 1) * s^4} - 3$$
- 4th moment of the normal distribution formula
- 0 = prop dist as ND, + = leptokurtic, - = platakurtic

The four "moments" are all independent – all combos possible

- mean & std "make most sense" as indices of central tendency & spread if skewness = 0 and kurtosic = 0

Nonparametric Univariate Statistics for ~ND/~Int variables

Central Tendency – "best guess of next case's value"

- median → middle-most value, 50^{th} percentile, 2^{nd} quartile

How to calculate the Mdn

1. Order data values — 11 13 16 18 18 21 22

2. Assign depth to each value, starting at each end

11	13	16	18	18	21	22
1	2	3	4	3	2	1

3. Calculate median depth $D_{mdn} = (N+1) / 2$ — $(7 + 1) / 2 = 4$

4. Median = value at D_{mdn} (or average of 2 values @ D_{mdn}, if odd number of values) — 18

Nonparametric Univariate Statistics for ~ND/~Int variables

Variability – "index of accuracy of next guess"

• Inter-quartile range (IQR)→ range of middle 50%, 3^{rd}-1^{st} quartile

How to calculate the IQR

1. Order & assign depth to each value

11	13	16	18	18	21	22
1	2	3	4	3	2	1

2. Calculate median depth $D_{Mdn} = (N+1) / 2$ — $(7 + 1) / 2 = 4$

3. Calculate quartile depth $D_Q = (D_{Mdn} + 1) / 2$ — $(4 + 1) / 2 = 2.5$

4. 1^{st} Quartile value — Ave of 13 & 16 = 14.5

5. 3^{rd} Quartile value — Ave of 18 & 21 = 19.5

6. IQR – 3^{rd} - 1^{st} Q values — 19.5 – 14.5 = 5

Univariate Parametric Statistical Tests for qualitative variables

Goodness-of-fit X^2 test

• Tests hypothesis about the distribution of category values of the population represented by the sample

• H0: is the hypothesized pop. distribution, based on either ...
  • theoretically hypothesized distribution
  • population distribution the sample is intended to represent
  • E.g., 65% females & 35% males or 30% Frosh, 45% Soph & 25% Juniors

• RH: & H0: often the same !

• binary and ordered category variables usually tested this way

• gof X^2 compares hypothesized distribution & sample dist.

• Retaining H0: -- sample dist. "equivalent to" population dist.

• Rejecting H0: -- sample dist. "is different from" population dist.

Data & formula for the gof X^2

Frequency of different class ranks in sample

Frosh	Soph	Junior
25	55	42

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Observed frequency – actual sample values (25, 55 & 42)

Expected frequency – based on a priori hypothesis

• however expressed (absolute or relative proportions, %s, etc)

• must be converted to expected frequencies

Example of a gof X^2

RH: "about ½ are sophomores and the rest are divided between frosh & juniors

Frosh	Soph	Junior
25	55	54

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

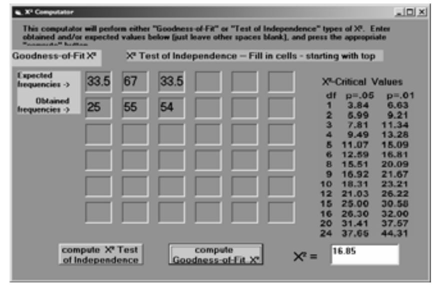
1. Obtain expected frequencies
- determine category proportions frosh .25 soph .5 junior .25
- determine category freq as proportion of total (N=134)
  - Frosh .25*122 = 33.5 Soph 67 Junior 33.5

2. Compute X^2
- $(25 – 33.5)^2/33.5 + (55-67)^2/67 + (54 – 33.5)^2/33.5 = 16.85$

3. Determine df & critical X^2
- df = k – 1 = 3 – 1 = 2
- $X^2_{2,.05} = 5.99$ $x^2_{2,.01} = 9.21$

4. NHST & such
- $X^2 > X^2_{2,.01}$, so reject H0: at p = .01
- Looks like fewer Frosh – Soph & more Juniors than expected

Doing gof X^2 "by hand" – Computators & p-value calculators



The top 2 rows of the X^2 Computator will compute a gof X^2

If you want to know the p-value with greater precision, use one of the online p-value calculators

Univariate Parametric Statistical Tests for ND/Int

1-sample t-test

Tests hypothesis about the mean of the population represented by the sample (μ -- "mu")

- H0: value is the hypothesized pop. mean, based on either ...
  - theoretically hypothesized mean
  - population mean the sample is intended to represent
  - e.g., pop mean age = 19
- RH: & H0: often the same !
- 1-sample t-test compares hypothesized μ & x
- Retaining H0: -- sample mean "is equivalent to" population μ
- Rejecting H0: -- sample mean "is different from" population μ

Example of a 1-sample t-test

$$t = \frac{\overline{X} - \mu}{SEM} \qquad SEM = (s^2 / n)$$

The sample of 22 has a mean of 21.3 and std of 4.3

1. Determine the H0: μ value
- We expect that the sample comes from a population with an average age of 19   μ = 19

2. Compute SEM & t
   - SEM = 4.3^2 / 22 = .84
   - t = (21.3 – 19) / .84 = 2.74

3. Determine df & t-critical or p-value
   - df = N-1 = 22 – 1 = 21
   - Using t-table $t_{21,.05}$ = 2.08 $t_{21,.01}$ = 2.83
   - Using p-value calculator p = .0123

4. NHST & such
- t > $t_{2,.05}$ but not t2,.05 so reject H0: at p = .05 or p = .0123
- Looks like sample comes from population older than 19

Univariate Parametric Statistical Tests for ~ND/~In

1-sample median test

Tests hypothesis about the median of the population represented by the sample H0: value is the hypothesized pop. median, based on either ...

- theoretically hypothesized mean
- population mean the sample is intended to represent
- e.g., pop median age = 19

- RH: & H0: often the same !
- 1-sample median test compares hypothesized & sample mdns
- Retaining H0: -- sample mdn "is equivalent to" population mdn
- Rejecting H0: -- sample mdn "is different from" population mdn

Example of a 1-sample median test

age data → 11 12 13 13 14 16 17 17 18 18 18 20 20 21 22 22

1. Obtain obtained & expected frequencies
   - determine hypothesized median value → 19
   - sort cases in to above vs. below H0: median value
   - Expected freq for each cell = ½ of sample → 8

<19	>19
11	5

2. Compute X^2
- $(11 – 8)^2/8 + (5 – 8)^2/8 = 2.25$

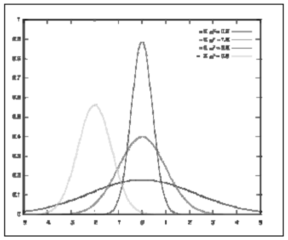
3. Determine df & X^2-critical or p-value
   - df = k-1 = 2 – 1 = 1
   - Using X^2-table $X^2_{1,.05}$ = 3.84 $X^2_{1,.05}$ = 6.63
   - Using p-value calculator p = .1336

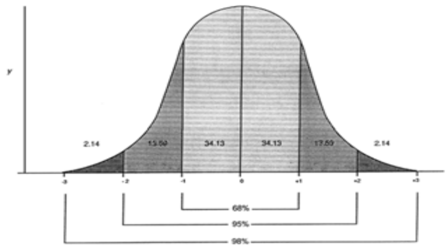
4. NHST & such
- X2 < X2 1, .05 & p > .05 so retain H0:
- Looks like sample comes from population with median not different from 19

Tests of Univariate ND

One use of gof X^2 and related univariate tests is to determine if data are distributed as a specific distribution, most often ND.

No matter what mean and std, a ND is defined by symmetry & proportional distribution





Using this latter idea, we can use a gof X^2 to test if the frequencies in segments of the distribution have the right proportions

• here we might use a k=6 gof X2 with expected frequencies based on % of 2.14, 13.59, 34.13, 34.13, 13.59 & 2.14

Tests of Univariate ND

One use of t-tests is to determine if data are distributed as a specific distribution, most often ND.

ND have skewness = 0 and kurtosis = 0

Testing Skewness

t = skewness / SES

Standard Error of Skewness

$SES \approx \sqrt{(6 / N)}$

Testing Kurtosis

t = kurtosis / SEK

Standard Error of Kurtosis

$SES \approx \sqrt{(24 / N)}$

Both of these are "more likely to find a significant divergence from ND, than that divergence is likely to distort the use of parametric statistics – especially with large N."