Norms and Basic Statistics

Cal State Northridge
Ψ427
Andrew Ainsworth PhD

Statistics AGAIN?

- What do we want to do with statistics?
  - Organize and Describe patterns in data
    - Taking incomprehensible data and converting it to:
      - Tables that summarize the data
      - Graphs
  - Extract (i.e. INFER) meaning from data
    - Infer POPULATION values from SAMPLES
    - Hypothesis Testing – Groups
    - Hypothesis Testing – Relation/Prediction

Descriptives

- Disorganized Data

Comedy	7	Suspense	8	Comedy	7	Suspense	7
Drama	8	Horror	7	Drama	5	Comedy	6
Horror	8	Comedy	5	Drama	3	Drama	3
Suspense	7	Horror	8	Comedy	6	Suspense	6
Horror	8	Comedy	6	Drama	7	Horror	9
Drama	5	Horror	9	Drama	6	Suspense	4
Drama	5	Horror	7	Suspense	3	Suspense	4
Horror	7	Suspense	5	Horror	10	Suspense	5
Horror	9	Suspense	6	Comedy	6	Drama	8
Comedy	7	Comedy	5	Comedy	4	Drama	4

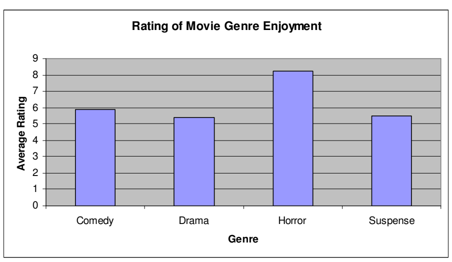
Descriptives

- Reducing and Describing Data

Genre	Average Rating
Comedy	5.9
Drama	5.4
Horror	8.2
Suspense	5.5

Descriptives

- Displaying Data

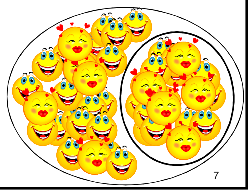


Inferential

- Inferential statistics:
  - Is a set of procedures to infer information about a population based upon characteristics from samples.
  - Samples are taken from Populations
  - Sample Statistics are used to infer population parameters

Inferential

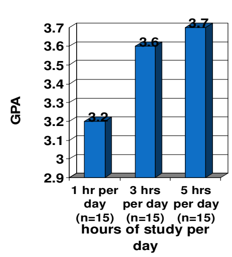
- **Population** is the *complete* set of people, animals, events or objects that share a common characteristic
- A **sample** is some subset or subsets, selected from the population.
  - **representative**
  - **simple random sample.**



	Population	Sample
Definition	The group (people, things, animals, etc.) you are intending to measure or study; they share some common characteristic	A subset of the population; used as a representative of the population
Size	Large to Theoretically Infinite	Substantially Smaller than the population (e.g. 1 to (population - 1))
Descriptive Characteristics	Parameters	Statistics
Symbols	Greek	Latin
Mean	μ	$\bar{X}$
Standard Deviation	σ	s or SD

Inferential

- Does the number of hours students study per day affect the grade they are likely to receive in statistics (Ψ320)?



Inferential

- Sometimes manipulation is not possible
- Is prediction possible?
- Can a relationship be established?
  - E.g., number of cigarettes smoked by per and the likelihood of getting lung cancer,
  - The level of child abuse in the home and the severity of later psychiatric problems.
  - Use of the death penalty and the level of crime.

Inferential

- Measured constructs can be assessed for co-relation (where the “coefficient of **correlation**” varies between -1 to +1)

-1 ———— 0 ———— 1

- “**Regression** analysis” can be used to assess whether a measured construct predicts the values on another measured construct (or multiple) (e.g., the level of crime given the level of death penalty usage).

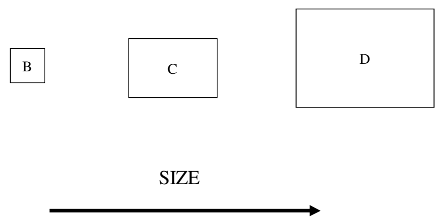
Measurement

- Statistical analyses depend upon the measurement characteristics of the data.
- **Measurement** is a process of assigning numbers to constructs following a set of rules.
- We normally measure variables into one of four different levels of measurement:
  - **Nominal**
  - **Ordinal**
  - **Interval**
  - **Ratio**

Ordinal Measurement

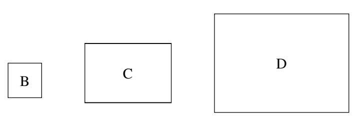
- Where Numbers Representative Relative Size Only

Contains 2 pieces of information



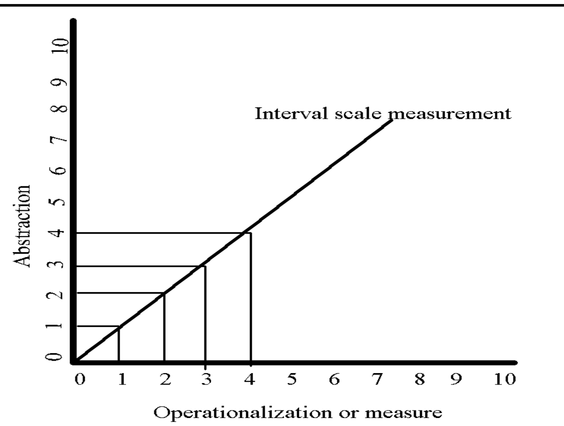
Interval Measurement:

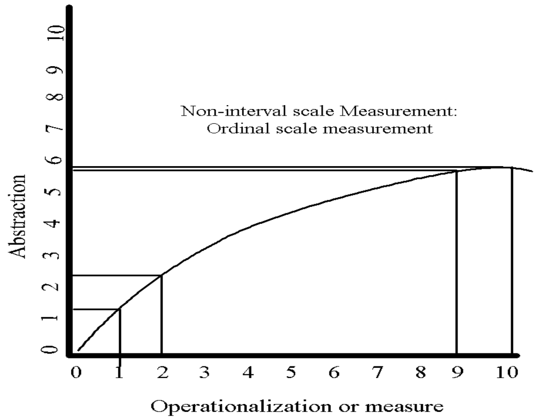
- Where Equal Differences Between Numbers Represent Equal Differences in Size



Numbers representing Size	**1**	**2**	**3**
Diff in numbers		**2-1=1**	**3-2=1**
Diff in size		**Size C – Size B =Size X**	**Size D – Size C = Size X**

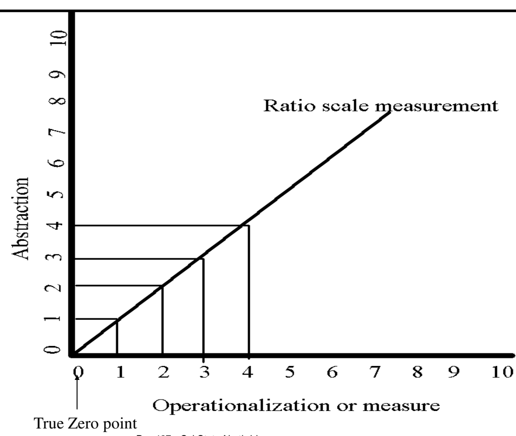
SIZE





Measurement

- **Ratio Scale Measurement**
  - In ratio scale measurement there are **four** kinds of information conveyed by the numbers assigned to represent a variable:
    - Everything Interval Measurement Contains Plus
    - A meaningful 0-point and therefore meaningful ratios among measurements.



Measurement

- Ratio Scale Measurement
  - If we have a true ratio scale, where 0 represents an a complete absence of the variable in question, then we form a meaningful ratio among the scale values such as:

$$4/2 = 2$$

  - However, if 0 is not a true absence of the variable, then the ratio 4/2 = 2 is not meaningful.

Percentiles and Percentile Ranks

- A **percentile** is the score at which a specified percentage of scores in a distribution fall below
  - To say a score 53 is in the 75th percentile is to say that 75% of all scores are less than 53
- The **percentile rank** of a score indicates the percentage of scores in the distribution that fall at or below that score.
  - Thus, for example, to say that the percentile rank of 53 is 75, is to say that 75% of the scores on the exam are less than 53.

Percentile

- **Scores** which divide distributions into specific proportions
  - Percentiles = hundredths

    $P_1, P_2, P_3, \ldots P_{97}, P_{98}, P_{99}$
  - Quartiles = quarters

    Q_1, Q_2, Q_3
  - Deciles = tenths

    $D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9$
- **Percentiles are the SCORES**

Percentile Rank

- What percent of the scores fall below a particular score?

$$PR = \frac{(Rank - .5)}{\text{N}} \times 100$$

- **Percentile Ranks are the Ranks not the scores**

Example: Percentile Rank

- Ranking no ties – just number them

Score:	1	3	4	5	6	7	8	10
Rank:	1	2	3	4	5	6	7	8

- Ranking with ties - assign midpoint to ties

Score:	1	3	4	6	6	8	8	8
Rank:	1	2	3	4.5	4.5	7	7	7

	Step 1	Step 2	Step 3	Step 4
Data	Order	Number	Assign Midpoint to Ties	Percentile Rank (Apply Formula)
9	1	1	1	2.381
5	2	2	2	7.143
2	3	3	4	16.667
3	3	4	4	16.667
3	3	5	4	16.667
4	4	6	7	30.952
8	4	7	7	30.952
9	4	8	7	30.952
1	5	9	10	45.238
7	5	10	10	45.238
4	5	11	10	45.238
8	6	12	12	54.762
3	7	13	14	64.286
7	7	14	14	64.286
6	7	15	14	64.286
5	8	16	17.5	80.952
7	8	17	17.5	80.952
4	8	18	17.5	80.952
5	8	19	17.5	80.952
8	9	20	20.5	95.238
8	9	21	20.5	95.238

- Steps to Calculating Percentile Ranks

- Example:

$$PR_3 = \frac{(Rank_3 - .5)}{\text{N}} \times 100 =$$

$$\frac{(4 - .5)}{21} \times 100 = 16.667$$

Percentile

$$X_P = (p)(n+1)$$

- Where X_P is the score at the desired percentile, p is the desired percentile (a number between 0 and 1) and n is the number of scores)
- If the number is an integer, than the desired percentile is that number
- If the number is not an integer than you can either round or interpolate; for this class we'll just round (round up when p is below .50 and down when p is above .50)

Percentile

- Apply the formula $X_P = (p)(n+1)$
  1. You'll get a number like 7.5 (think of it as *place1.proportion)*
  2. Start with the value indicated by *place1* (e.g. 7.5, start with the value in the 7th place)
  3. Find *place2* which is the next highest place number (e.g. the 8th place) and subtract the value in *place1* from the *value* in *place2*, this *distance1*
  4. Multiple the *proportion* number by the *distance1* value, this is *distance2*
  5. Add *distance2* to the value in *place1* and that is the *interpolated value*

Example: Percentile

- Example 1: 25th percentile:

  {1, 4, 9, 16, 25, 36, 49, 64, 81}
- X_{25} = (.25)(9+1) = 2.5
  - *place1* = 2, *proportion* = .5
  - Value in *place1* = 4
  - Value in *place2* = 9
  - *distance1* = 9 – 4 = 5
  - *distance2* = 5 * .5 = 2.5
  - *Interpolated value* = 4 + 2.5 = 6.5
  - 6.5 is the 25th percentile

Example: Percentile

- Example 2: 75th percentile
  {1, 4, 9, 16, 25, 36, 49, 64, 81}
- $X_{75} = (.75)(9+1) = 7.5$
  - *place1* = 7, *proportion* = .5
  - Value in *place1* = 49
  - Value in *place2* = 64
  - *distance1* = 64 – 49 = 15
  - *distance2* = 15 * .5 = 7.5
  - *Interpolated value* = 49 + 7.5 = 56.5
  - 56.5 is the 75th percentile

Quartiles

- To calculate Quartiles you simply find the scores the correspond to the 25, 50 and 75 percentiles.
- $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$

Reducing Distributions

- Regardless of numbers of scores, distributions can be described with three pieces of info:
  - **Central Tendency**
  - Variability
  - Shape (Normal, Skewed, etc.)

Measures of Central Tendency

Measure	Definition	Level of Measurement	Disadvantage
Mode	Most frequent value	nom., ord., int./rat.	Crude
Median	Middle value	ord., int./rat.	Only two points contribute
Mean	Arithmetic average	int./rat.	Affected by skew

The Mean

- Only used for interval & ratio data.

$$Mean = M_X = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Major advantages:
  - The sample value is a very good estimate of the population value.

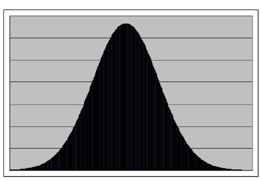
Reducing Distributions

- Regardless of numbers of scores, distributions can be described with three pieces of info:
  - Central Tendency
  - **Variability**
  - Shape (Normal, Skewed, etc.)

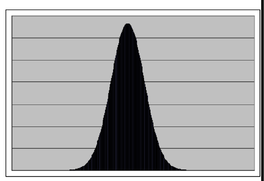
How do scores spread out?

- Variability
  - Tell us how far scores spread out
  - Tells us how the degree to which scores deviate from the central tendency

How are these different?



Mean = 10



Mean = 10

Measure of Variability

Measure	Definition	Related to:
Range	Largest - Smallest	Mode
Interquartile Range	X_{75} - X_{25}	Median
Semi-Interquartile Range	(X_{75} - X_{25})/2	
Average Absolute Deviation	$\frac{\sum \lvert X_i - \bar{X} \rvert}{N}$	Mean
Variance	$\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}$	
Standard Deviation	$\sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}}$	

The Range

- The simplest measure of variability
  - **Range (R)** = $X_{highest} - X_{lowest}$
  - Advantage – Easy to Calculate
  - Disadvantages
    - Like Median, only dependent on two scores → unstable

      {0, 8, 9, 9, 11, 53} Range = 53

      {0, 8, 9, 9, 11, 11} Range = 11
    - Does not reflect all scores

Variability: IQR

- Interquartile Range
  - = $P_{75} - P_{25}$ or $Q_3 - Q_1$
  - This helps to get a range that is not influenced by the extreme high and low scores
  - Where the range is the spread across 100% of the scores, the IQR is the spread across the middle 50%

Variability: SIQR

- Semi-interquartile range
  - $=(P_{75} - P_{25})/2$ or $(Q_3 - Q_1)/2$
  - IQR/2
  - This is the spread of the middle 25% of the data
  - The average distance of Q1 and Q3 from the median
  - Better for skewed data

Variability: SIQR

- Semi-Interquartile range

Variance

- The average squared distance of each score from the mean
- Also known as the **mean square**
- Variance of a sample: s^2
- Variance of a population: σ^2

Variance

- When calculated for a sample

$$s^2 = \frac{\sum \left(X_i - \bar{X} \right)^2}{N - 1}$$

- When calculated for the entire population

$$\sigma^2 = \frac{\sum \left(X_i - \bar{X} \right)^2}{N}$$

Standard Deviation

- Variance is in squared units
- What about regular old units
- Standard Deviation = Square root of the variance

$$s = \sqrt{\frac{\sum\left(X_i - \bar{X}\right)^2}{N-1}}$$

Standard Deviation

- Uses measure of central tendency (i.e. mean)
- Uses all data points
- **Has a special relationship with the normal curve**
- Can be used in further calculations
- Standard Deviation of Sample = *SD* or *s*
- Standard Deviation of Population = σ

Why N-1?

- When using a sample (which we always do) we want a statistic that is the best estimate of the parameter

$$E\left(\frac{\sum\left(X_i - \bar{X}\right)^2}{N-1}\right) = \sigma^2 \qquad E\left(\sqrt{\frac{\sum\left(X_i - \bar{X}\right)^2}{N-1}}\right) = \sigma$$

Degrees of Freedom

- Usually referred to as *df*
- Number of observations minus the number of restrictions

__+__+__+__=10 - 4 free spaces

2 +__+__+__=10 - 3 free spaces

2 + 4 +__+__=10 - 2 free spaces

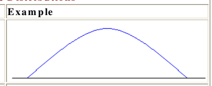
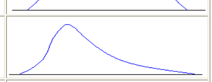
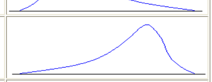
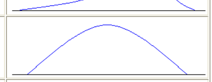
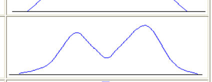
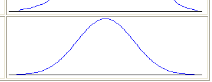
2 + 4 + 3 +__=10

Last space is not free!! Only 3 dfs.

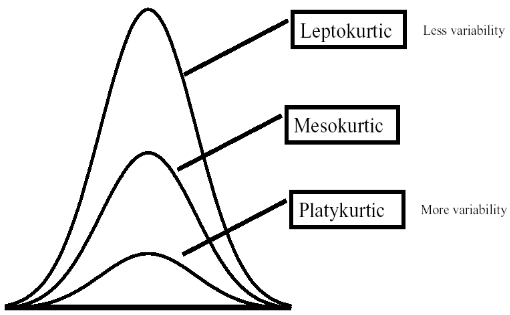
Reducing Distributions

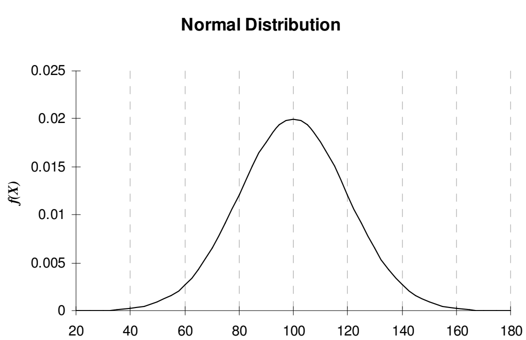
- Regardless of numbers of scores, distributions can be described with three pieces of info:
  - Central Tendency
  - Variability
  - **Shape (Normal, Skewed, etc.)**

Terms that Describe Distributions

Term	Features	Example
"Symmetric"	left side is mirror image of right side	
"Positively skewed"	right tail is longer then the left	
"Negatively skewed"	left tail is longer than the right	
"Unimodal"	one highest point	
"Bimodal"	two high points	
"Normal"	unimodal, symmetric, asymptotic	

Symmetrical, Bell-Shaped





Example: The Mean = 100 and the Standard Deviation = 20

Normal Distribution (Characteristics)

- Horizontal Axis = possible X values
- Vertical Axis = density (i.e. *f*(X) related to probability or proportion)
- Defined as

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}}(e)^{-(X-\mu)^2/2\sigma^2}$$

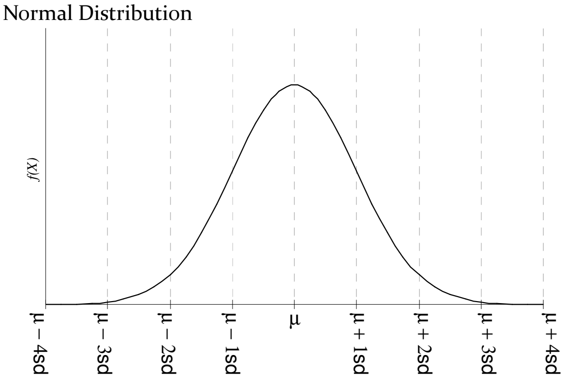
$$f(X_i) = \frac{1}{(s)\sqrt{2*(3.14159265)}} * (2.71828183)^{-(X_i-\bar{X})^2/2s^2}$$

- The distribution relies on only the **mean** and **s**

Normal Distribution (Characteristics)

- Bell shaped, symmetrical, unimodal
- Mean, median, mode all equal
- No real distribution is perfectly normal
- But, many distributions are approximately normal, so normal curve statistics apply
- Normal curve statistics underlie procedures in most inferential statistics.

Normal Distribution



The standard normal distribution

- A normal distribution with the added properties that the mean = 0 and the s = 1
- Converting a distribution into a standard normal means converting raw scores into Z-scores

Z-Score Formula

- Raw score $\rightarrow$ Z-score

$$Z_i = \frac{X_i - \bar{X}}{s} = \frac{\text{score - mean}}{\text{standard deviation}}$$

- Z-score $\rightarrow$ Raw score

$$X_i = Z_i(s) + \bar{X}$$

Properties of Z-Scores

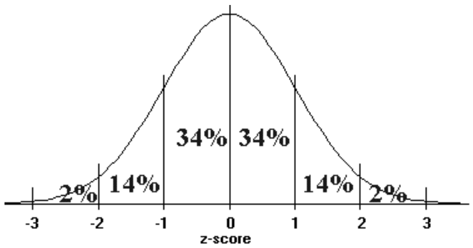
- Z-score indicates how many SD's a score falls above or below the mean.
- Positive z-scores are above the mean.
- Negative z-scores are below the mean.
- Area under curve → probability
- Z is continuous so can only compute probability for range of values

Properties of Z-Scores

- Most z-scores fall between -3 and +3 because scores beyond 3sd from the mean
- Z-scores are standardized scores → allows for easy comparison of distributions

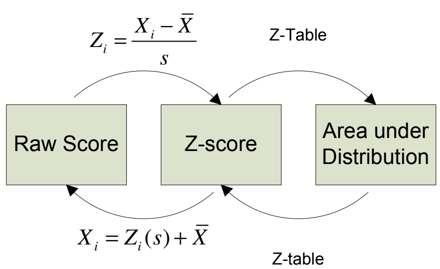
The standard normal distribution

- Rough estimates of the SND (i.e. Z-scores):



Have→Need Chart

When rough estimating isn't enough



What about negative Z values?

- Since the normal curve is symmetric, areas beyond, between, and below positive z scores are identical to areas beyond, between, and below negative z scores.
- There is no such thing as negative area!

Norms and Norm-Referenced Tests

- Norm - statistical representations of a population (e.g. mean, median).
- Norm-referenced test (NRT) – Compares an individual's results on the test with the pre-established norm
- Made to compare test-takers to each other
- I.E. - The Normal Curve

Norms and Norm-Referenced Tests

- Normally rather than testing an entire population, the norms are inferred from a representative sample or group (inferential stats revisited).
- Norms allow for a better understanding of how an individual's scores compare with the group with which they are being compared
- Examples: WAIS, SAT, MMPI, Graduate Record Examination (GRE)

Criterion-Referenced Tests

- Criterion-referenced tests (CRTs) - intended to measure how well a person has mastered a specific knowledge set or skill
- Cutscore – point at which an examinee passes if their score exceeds that point; can be decided by a panel or by a single instructor
- Criterion – the domain in which the test is designed to assess