L665 ML for NLP

Damir Cavar
Indiana University
Jan. 2018

Agenda

- Syllabus
- Descriptive Statistics
- Normal distribution and Significance tests
- Assignment

Descriptive Statistics

- For different text types:
  - Does the choice of words say something specific about genre, century, the author etc.?
  - Which words are more frequent in what kind of texts?
  - Which syntactic constructions are more or less frequent in what kind of texts?
  - How many different words (not tokens) do we find in what kind of texts?

Descriptive Statistics

- Information theory (Shannon, 1948):
- Counting letters in English prose by one author:
  - Proportion of E = 13%
  - Proportion of W = 2%
- Same proportions in (long enough) text by any other author.

Inductive Statistics

- Given the descriptive statistic facts we can make predictions:
- The distribution of words: the most frequent
- The distribution of letters: E 13%, W 2%

Inductive Statistics

- We can use statistical properties to e.g. predict
- the type of text
- the century it is created
- the origin of the author
- the language of the document
- the part-of-speech of a word
- the argument structure of a verb
- ...

Word Counts

- Count words and build a decreasing frequency profile
- What kind of regularities do you see in the frequency profile of one document?
- How can we compare the frequency profile of two or more documents?

Results

- Open vs. Closed Class Lexicon
  - Function words belong to the closed class.
  - Verbs, nouns, adjectives etc. belong to the open class.
- Function words are most frequent.
- Function words make up the smallest part of a natural language lexicon.

FOXTROT


HOW'S YOUR 500-WORD HISTORY PAPER COMING?
WELL, BASED ON PAST PER-FORMANCE, I'M ALREADY A QUARTER DONE.
WHAT DO YOU MEAN?
I KNOW I'LL USE THE WORD "THE" ABOUT 25 TIMES, "AND" AT LEAST 15. "IN," "IF," "IT" AND "BUT" SHOULD GIVE ME ANOTHER 30-40. TOSS IN THE USUAL "IS," "WAS," "WILL BE" VERB ASSORTMENT AND I'M SITTING COMFORTABLY AT 120-PLUS WORDS BEFORE I EVEN START.
THE KEY TO WRITING A HISTORY ESSAY IS KNOWING YOUR ESSAY HISTORY.
YOU NEVER FAIL TO AMAZE ME.
THANKS. WHO'D WE FIGHT IN WORLD WAR I, BY THE WAY?

Probability Theory

- The chance of a particular outcome occurring is determined by the ratio of the number of favorable outcomes to the total number of outcomes.
-
- Approach: frequency based

$$\text{probability of favorable outcome} = \frac{\text{number of favorable outcomes}}{\text{total number of outcomes}}$$

Statistics

- Inductive Statistics
  - analytic: conclusions from data
    - including probabilities
- Descriptive Statistics
  - description of data
  - display or presentation of data

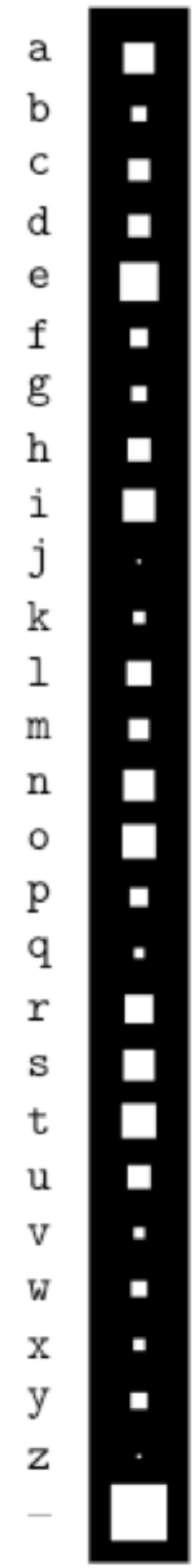
N-gram Models

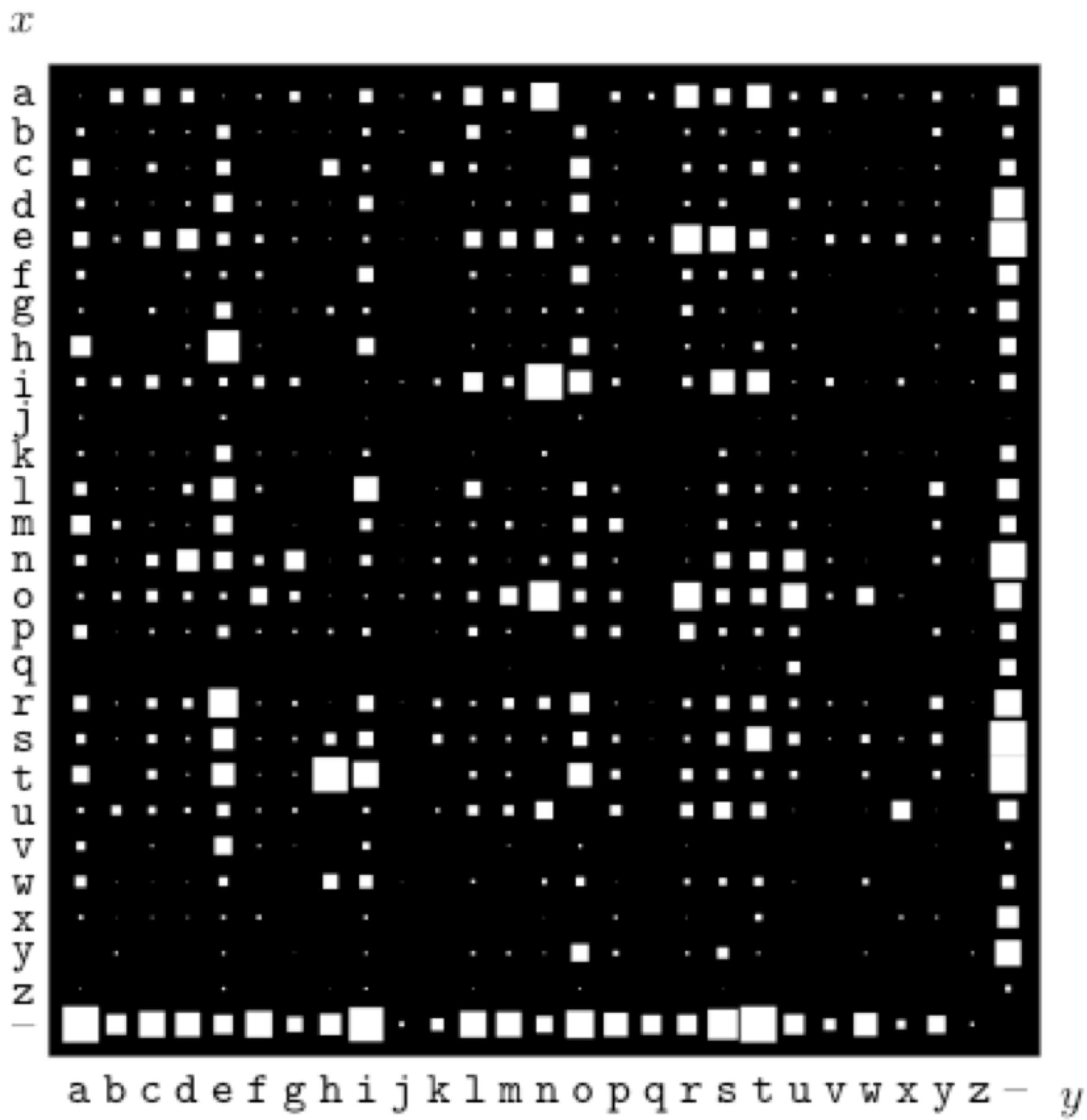
- List all possible symbol combinations of length *n* for a given corpus,
  - symbols: phones, phonemes, characters, morphemes, words (tokens or types), sentences, paragraphs etc.
- together with their frequencies (absolute + number of all elements/tokens; relative)

Frequency Profiles

- Unigram
- Bi-gram
  - Tables/graphics taken from MacKay (2003)

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928





x
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
u
v
w
x
y
z
–
a b c d e f g h i j k l m n o p q r s t u v w x y z –
y

Relative Frequency Theory

- If an experiment is repeated an extremely large number of times and a particular outcome occurs a percentage of the time, then the particular percentage is close to the probability of that outcome.

Cryptography

- What do counts tell us?
  - Using counts to match encrypted code with word types
- Language = cryptography
  - Navajo
- Linguistics = cryptoanalysis

Zipf's Law

- Principle of Least Effort
  - Minimize probable effort of work.
    - Not immediate, but considering future work.
- Reciprocal relationship between word frequencies and rank in a frequency table:

$$f \propto \frac{1}{r}$$

  - There is a constant *k* such that *f* multiplied with *r* equals *k*.

Zipf’s Law

- Explanation:
  - Count the words and sort them based on their frequency.
  - The 50th most common word should occur with three times the frequency of the 150th most common word.
- Problem:
- Slight inaccuracy for words in the rank 100.

Zipf's Law

- Zipf's explanation:
  - Speaker and hearer are trying to minimize their effort.
  - Speaker: tendency for a small vocabulary
  - Hearer: large vocabulary of individually rarer words
    - reduction of ambiguity
- Relevance:
  - For most words there is little data in the corpus.
    - Collocation, distribution, use

Zipf's Other Laws

- Number of meanings of a word obeys the law:

$$m \propto \sqrt{f} \qquad\qquad m \propto \frac{1}{\sqrt{r}}$$

- The lower the rank of a word in the frequency scale, the less meanings the word has.
  - High frequent words are more often ambiguous.
  - Low frequent words are less ambiguous.

Zipf's Other Laws

- Clumping of content words:
  - Measure the distance between words (the same words)
    - characters, lines, pages
  - Calculate the frequency F of the different intervals I
- Content words occur near another occurrence of the same word.

Statistics

- Types of statistics
  - Numerical statistics
    - Numbers, average, mean
  - Pictorial statistics
    - Presentation of numerical statistics
  - Inductive statistics
    - Process numbers and/or pictures

Numerical Statistics

- Measures of central tendencies of data
  - Mean
  - Median
  - Mode
- Measures of variation/variability
  - Spread in data
- Measurement scales

Numerical Statistics

- Arithmetic Mean
  - Data set:

File	**Count words**
Flo031201.txt	10346
Flo031202a.txt	5031
Flo031202b.txt	11876
Flo031203.txt	12175
Flo031204.txt	10943

Numerical Statistics

- Arithmetic Mean

$$arithmetic\ mean = \frac{sum\ of\ measures}{number\ of\ measures}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

  - example:

$$\frac{10346 + 5031 + 11876 + 12175 + 10943}{5} = 10074.2$$

Numerical Statistics

- Median
  - Middle value of ordered measure values

File	**Count words**
Flo031202a.txt	5031
Flo031201.txt	10346
Flo031204.txt	10943
Flo031202b.txt	11876
Flo031203.txt	12175

Numerical Statistics

- Median
  - Decrease relevance of outliers:

File	**Count words**
Flo031202a.txt	5031
Flo031201.txt	10346
Flo031204.txt	10943
Flo031202b.txt	11876
Flo031203.txt	12175

Numerical Statistics

- Median

  with even number of elements:

File	Count words
Flo031202a.txt	5031
Flo031201.txt	10346
Flo031204.txt	10943
Flo031202b.txt	11876

Arithmetic mean of the two middle values:

$$\frac{10346 + 10943}{2} = 10644.5$$

Median

- Measures need to be sorted (irrespective of whether decreasing or increasing):

$$median = \begin{cases} x_{k+1} & \text{if } n \text{ is odd}, n = 2k \\ \frac{x_k + x_{k+1}}{2} & \text{if } n \text{ is even}, n = 2k+1 \end{cases}$$

Numerical Statistics

- Mean: 10074.2
- Median: 10943
- Mean is reduced on the basis of the outlier:
  Flo031202a.txt 5031
- Median may be a better indicator of central tendency if outliers/extreme values are present.

Numerical Statistics

- Mode

  The measure value that occurs most often:

File	Count words
Flo031202a.txt	5031
Flo031201.txt	10943
Flo031204.txt	10943
Flo031202b.txt	6329
Flo031203.txt	12175

Mode = 10943

Numerical Statistics

- Approximation of
  - Mode
    - *mean* – 3 (*mean* – *median*)
  - Median
    - (2 *mean* + *mode*) / 3
  - Mean
    - (3 *median* – *mode*) / 2

Numerical Statistics

- Notation
  - Mean (x bar): $\overline{\mathrm{x}}$
  - Mean of a population:
  - Sum of values: μ

$$\sum$$

Numerical Statistics

- Notation example:
  - Arithmetic mean:

$$\overline{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Numerical Statistics

- Arithmetic mean for grouped data:

Files	Count words
35%	0-4999
30%	5000-9999
25%	10000-14999
10%	15000-19999

- With 100 sample documents what is the arithmetic mean?

Numerical Statistics

- Arithmetic mean for grouped data:

$$\bar{x} = \frac{\sum fx}{n}$$

  - f = frequency
  - x = midpoint

Numerical Statistics

- Arithmetic mean for grouped data:

Files	**Midpoint**	***fx***	**Count words**
35	2500	87500	0-4999
30	7500	225000	5000-9999
25	12500	312500	10000-14999
10	17500	175000	15000-19999

$$\overline{x} = \frac{\sum fx}{n} = \frac{87500 + 225000 + 312500 + 175000}{100} = \frac{800000}{100} = 8000$$

Numerical Statistics

- Median for grouped data:

$$median = L + \frac{w}{f_{med}}\left(.5n - \sum f_b\right)$$

- – L = lower class limit that contains the interval
- – n = total number of measurements
- – w = class width
- – f_{med} = frequency of the class containing the median
- – Σf_b = sum of the frequencies for all classes before the median class

Numerical Statistics

- Median for grouped data:

Files	Count words
35	0-4999
30	5000-9999
25	10000-14999
10	15000-19999

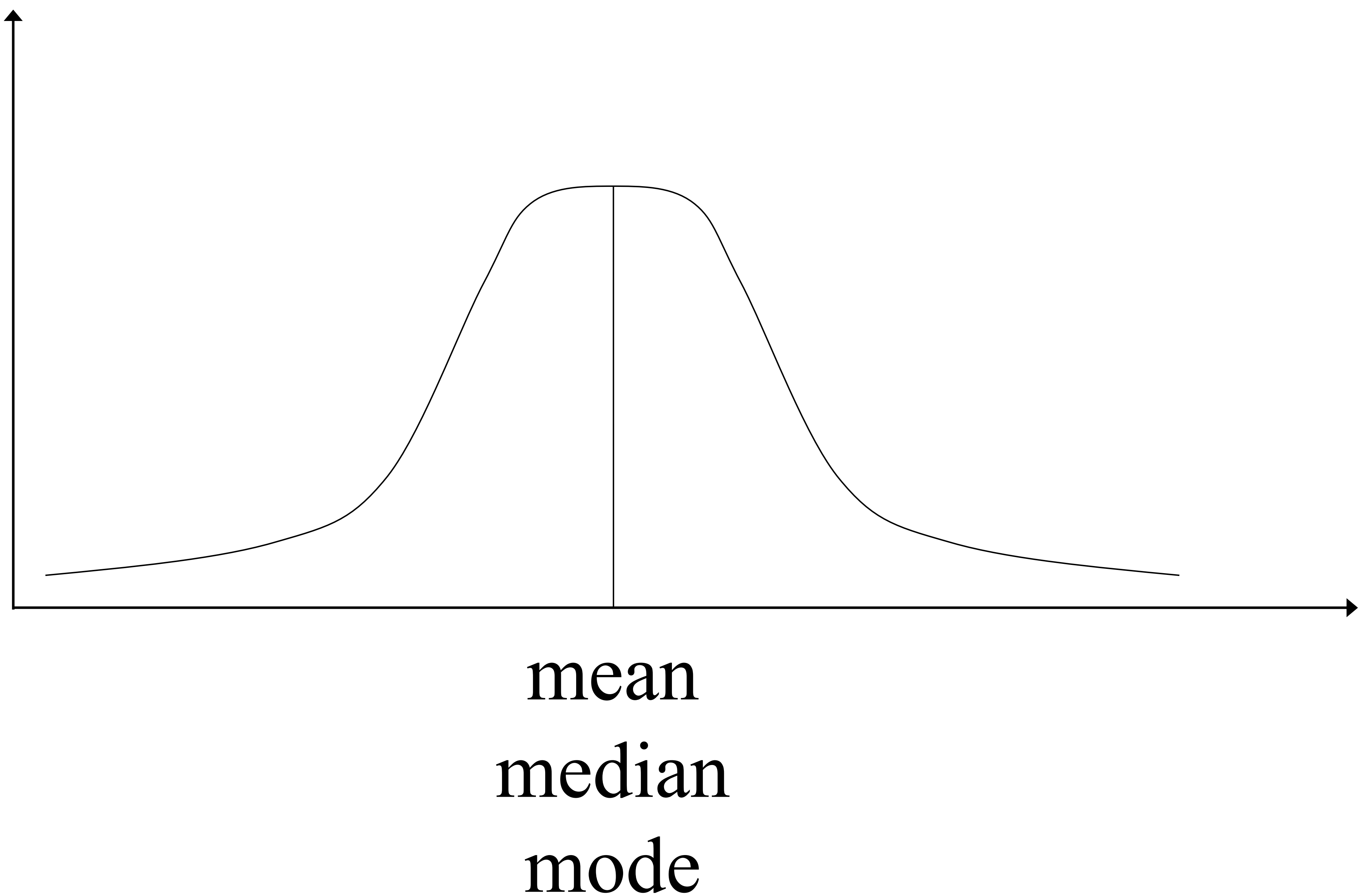
$$median = 5000 + \frac{4999}{30}(50 - 35) = 7499.5$$

Numerical Statistics

- Distribution
  - Symmetric distribution
  - Skewed curves
    - negatively skewed curves
    - positively skewed curves

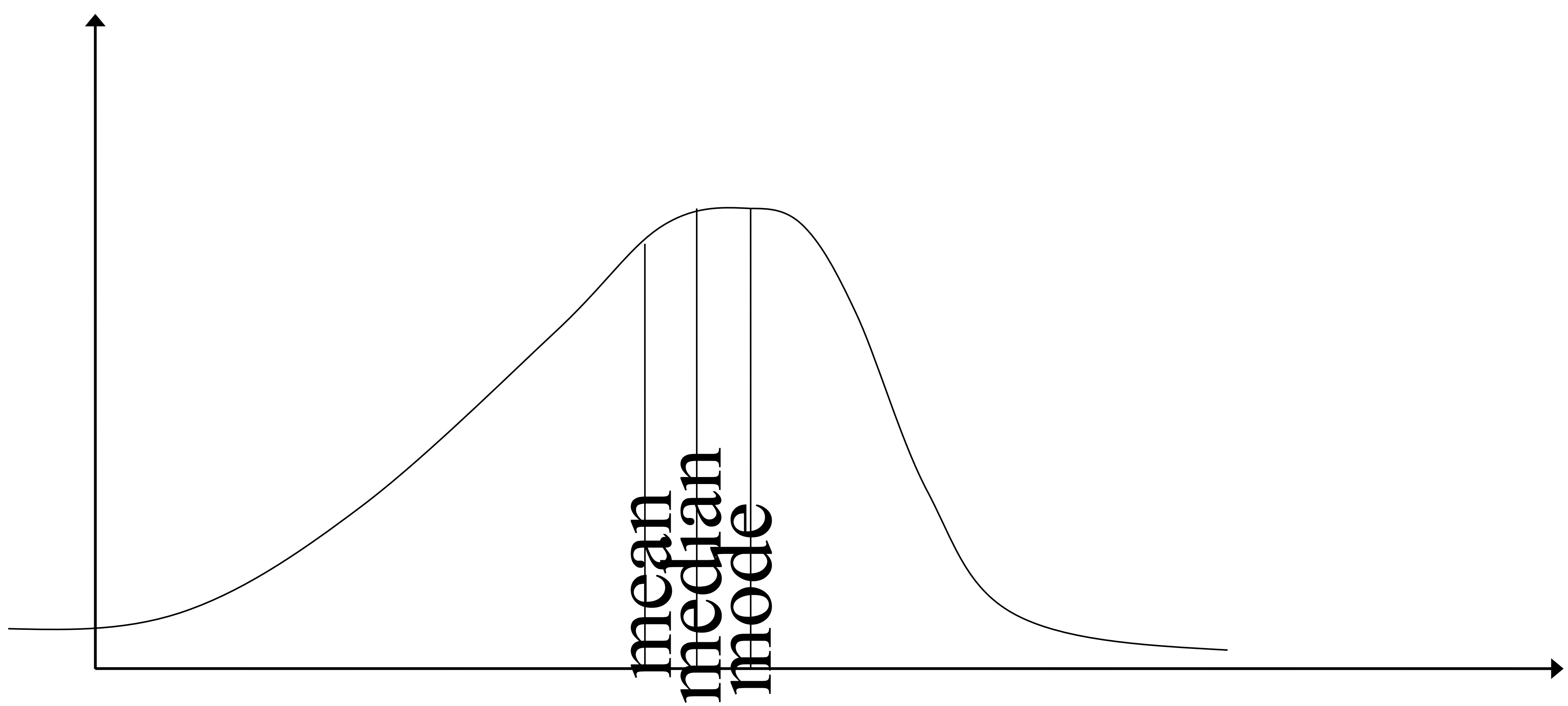
Numerical Statistics

- Symmetric distribution: Mean, median and mode are equal.



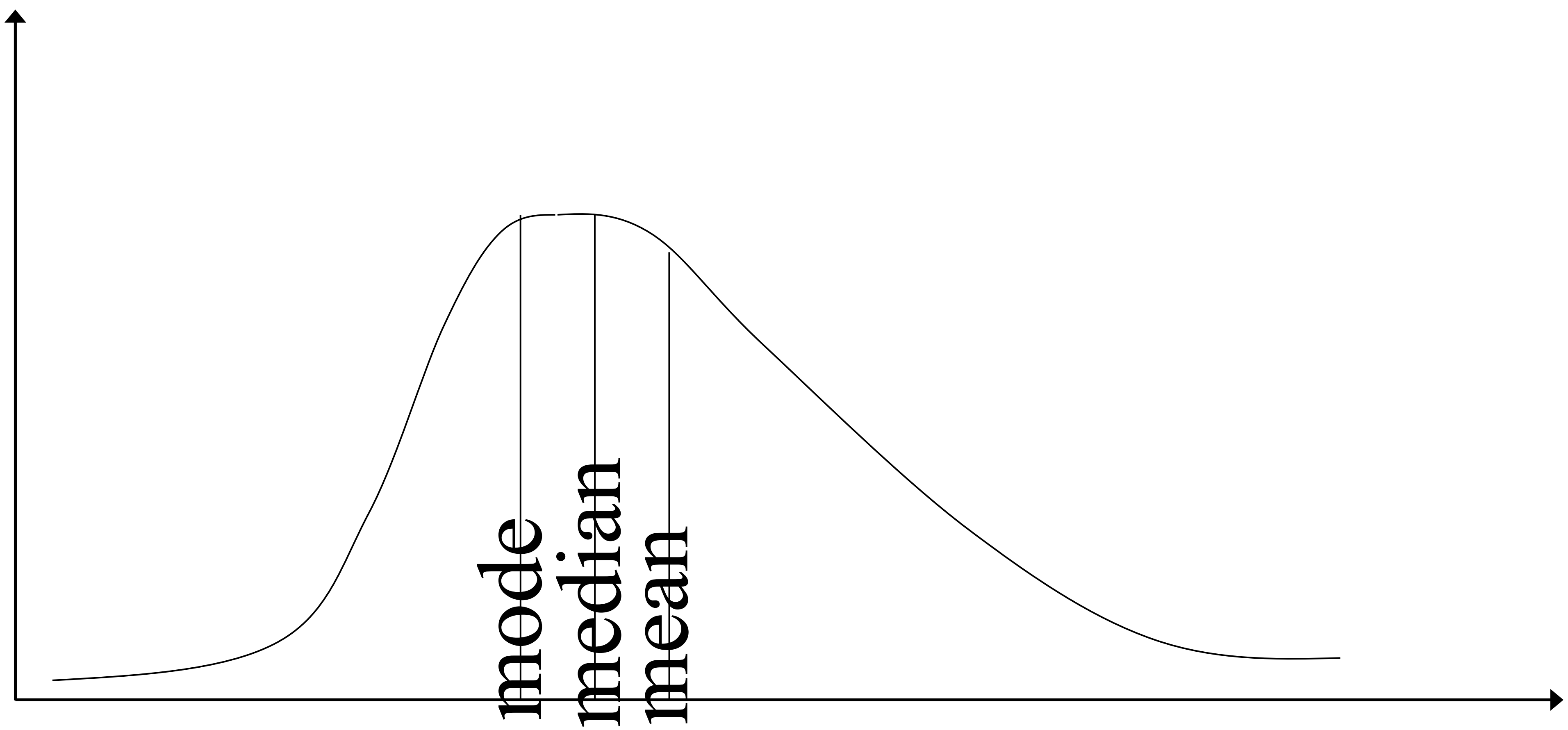
Numerical Statistics

- Skewed curves: Negatively skewed distribution: mean < median < mode



Numerical Statistics

- Skewed curves
  - Positively skewed distribution: mode < median < mean



Numerical Statistics

- Variability

Experiment 1	Experiment 2
195	10
210	0
199	400
200	20
205	380
190	200
200	390
201	200

Numerical Statistics

- Variability
  - For both experiments:
    - mean: 200
    - mode: 200
    - median: 200
  - Experiment 2 has greater variation.
- Measure of variation:
  - Range
  - Deviation
  - Variance

Numerical Statistics

- Range
  - Difference between largest and smallest value:
    - Experiment 1: 210 – 190 = 20
    - Experiment 2: 400 – 0 = 400
- Deviation
  - Distance of the measurements away from the mean:
    - Experiment 1: less
    - Experiment 2: more

Numerical Statistics

- Deviation
  - Distance of the measurements away from the mean:
    - Experiment 1: less
    - Experiment 2: more

$$sd = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$$

Numerical Statistics

- Variance
  - Sum of squared deviations of n measurements from their mean divided by (n – 1):
    - Experiment 1: ?
    - Experiment 2: ?

$$s_N^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad s_{N-1}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Numerical Statistics

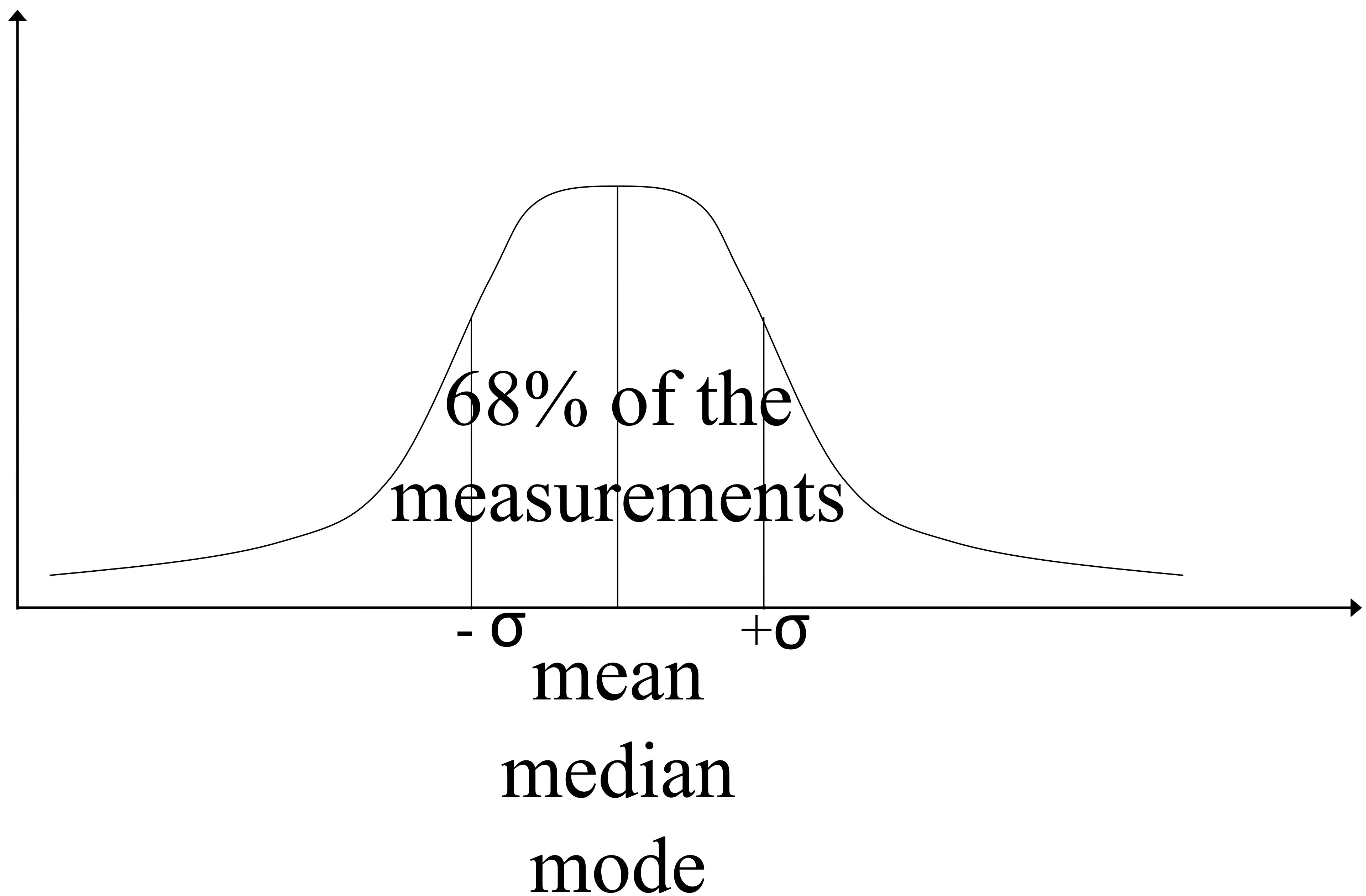
- Standard deviation
  - Positive square root of the variance.
    - Experiment 1: ?
    - Experiment 2: ?

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad s_{N-1} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

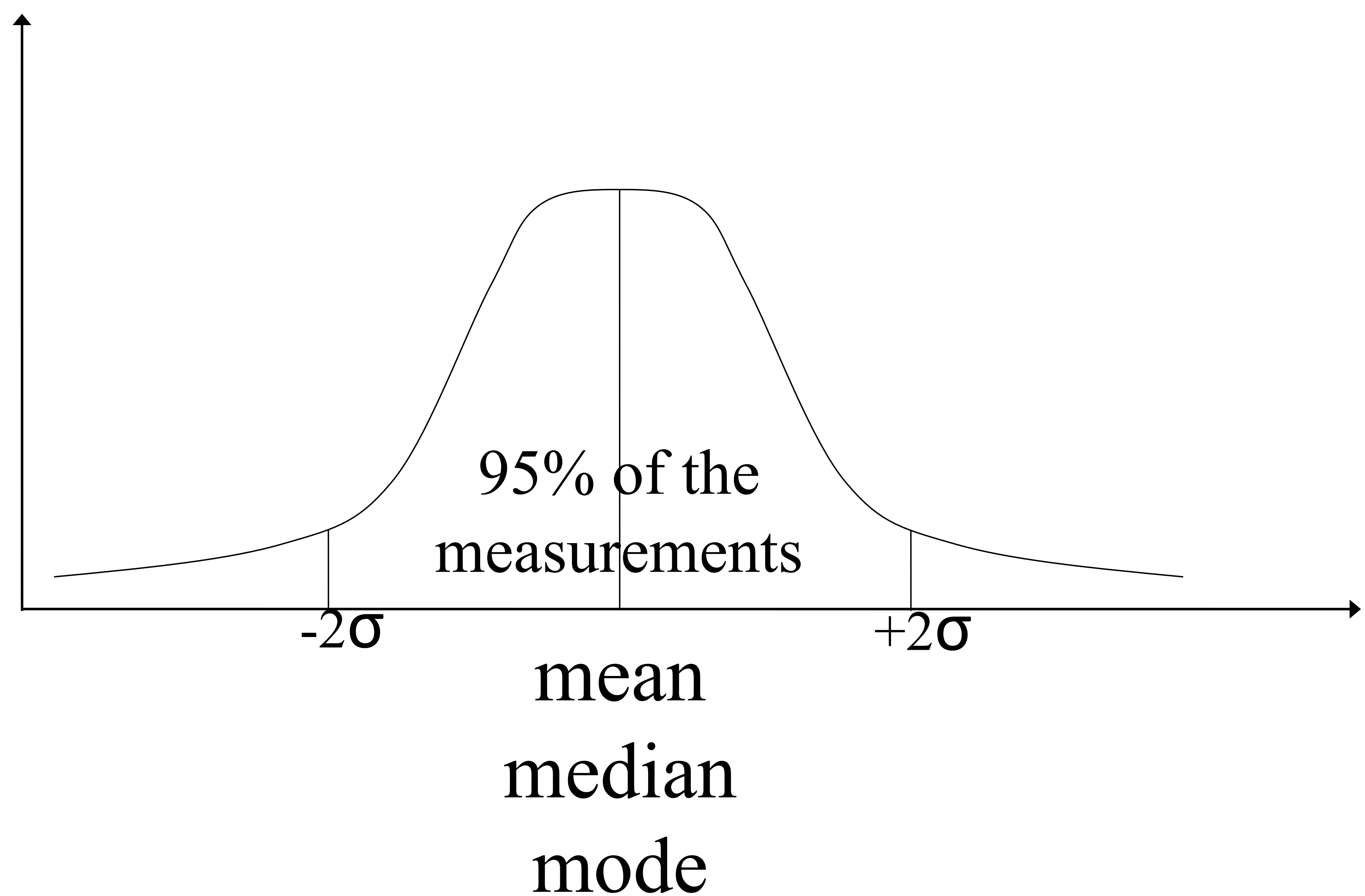
Numerical Statistics

- Notation
  - s^2 = variance of a sample
  - σ^2 = variance of a population
  - s = standard deviation of a sample
  - σ = standard deviation of a population

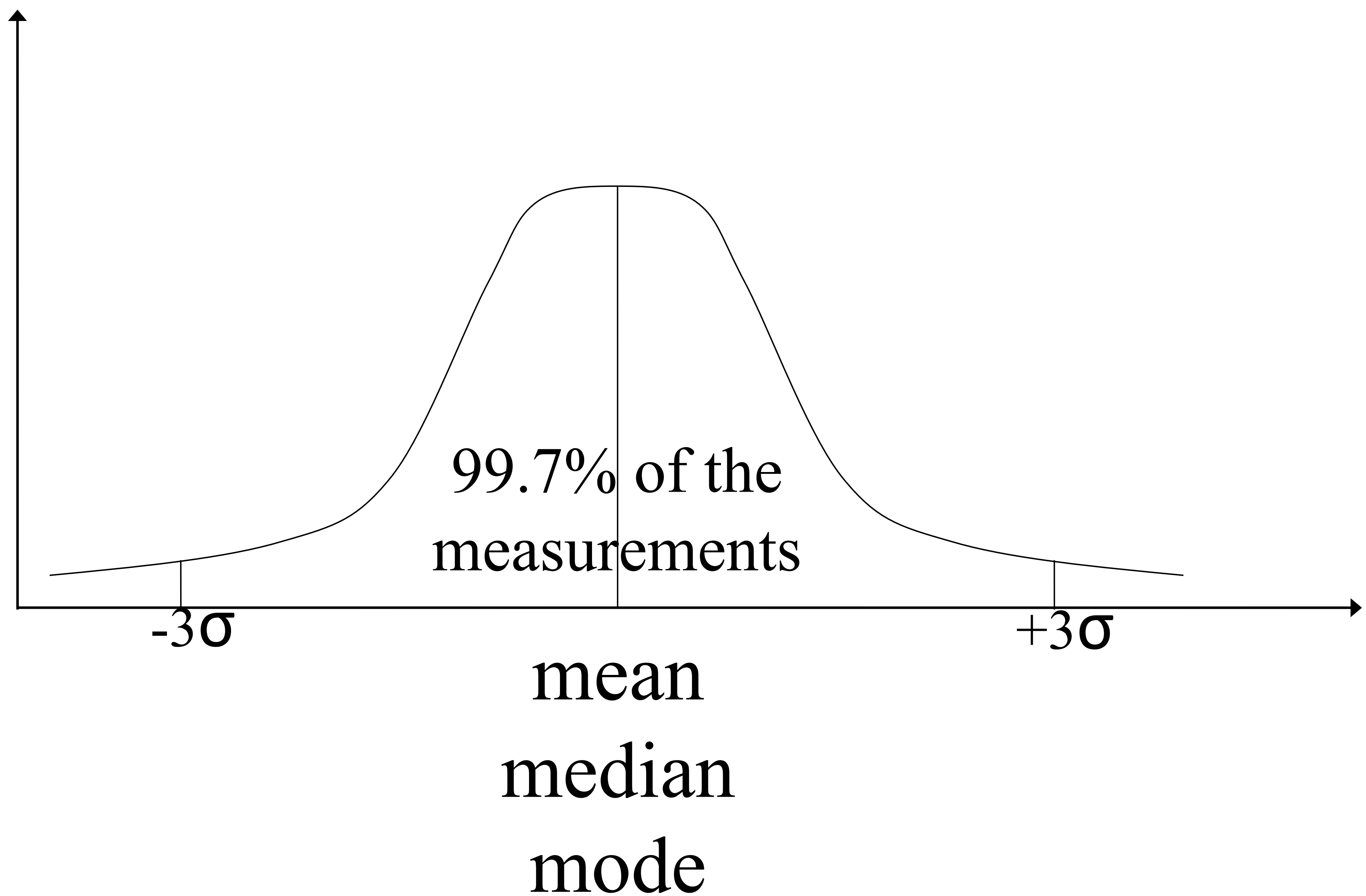
Numerical Statistics



Numerical Statistics



Numerical Statistics



Collocations

- Words in context
  - distribution
  - fixed expressions
  - collocations
    - statistical properties
    - function words

How to test for collocations?

- Statistics
- Significance tests

Significance

- Notations:
  - Type I error rate of .05
  - Alpha level of .05 or $\alpha = .05$
  - Finding is significant at the .05 level
  - Confidence level is 95%
  - 95% certainty that a result is not due to chance
  - A 1 in 20 chance of obtaining the result
  - Are of the region of rejection is .05
  - p-value is .05 or $p = .05$

Testing

- Statistics as testing of scientific hypotheses
- Strategies:
  - Formulating a Research Hypothesis or Alternative Hypothesis (Ha)
    - Statement of the expectation to be tested

Testing

- Strategies:
  - Derivation of a statement that is the opposite of the research hypothesis: Null Hypothesis (H0)
    - Testing the null hypothesis

Testing

- Statistics as testing of scientific hypotheses
- Strategies:
  - If the null hypothesis can be rejected, this is evidence in favor of the research hypothesis.

Testing

- Strategies:
  - Usually:
    - No prove for research hypothesis, just support for it.

Testing

- Research Hypothesis:
  - At IU linguistics students perform differently in statistics than computer science students.
    - Ha: $\mu 1 \neq \mu 2$
    - Ha: $\mu 1 - \mu 2 \neq \odot$

Testing

- Null Hypothesis:
  - At IU linguistics students perform the same in statistics as computer science students.
    - H0: $\mu1 = \mu2$
    - H0: $\mu1 - \mu2 = 0$

Testing

- More specific: Research Hypothesis:
  - At IU linguistics students perform better in statistics than computer science students.
    - Ha: $\mu 1 > \mu 2$
    - Ha: $\mu 1 - \mu 2 > 0$

Testing

- More specific: Null Hypothesis
  - At IU linguistics students perform worse in statistics, or equal to computer science students.
    - H0: $\mu 1 \leq \mu 2$
    - H0: $\mu 1 - \mu 2 \leq 0$

Testing

- Given the distribution of a known area
  - e.g. normal distribution
- estimate the probability of obtaining a certain value as a result of chance.
- If the probability is low, the likelihood for a mere coincidence is low, i.e. a certain theory is correct.

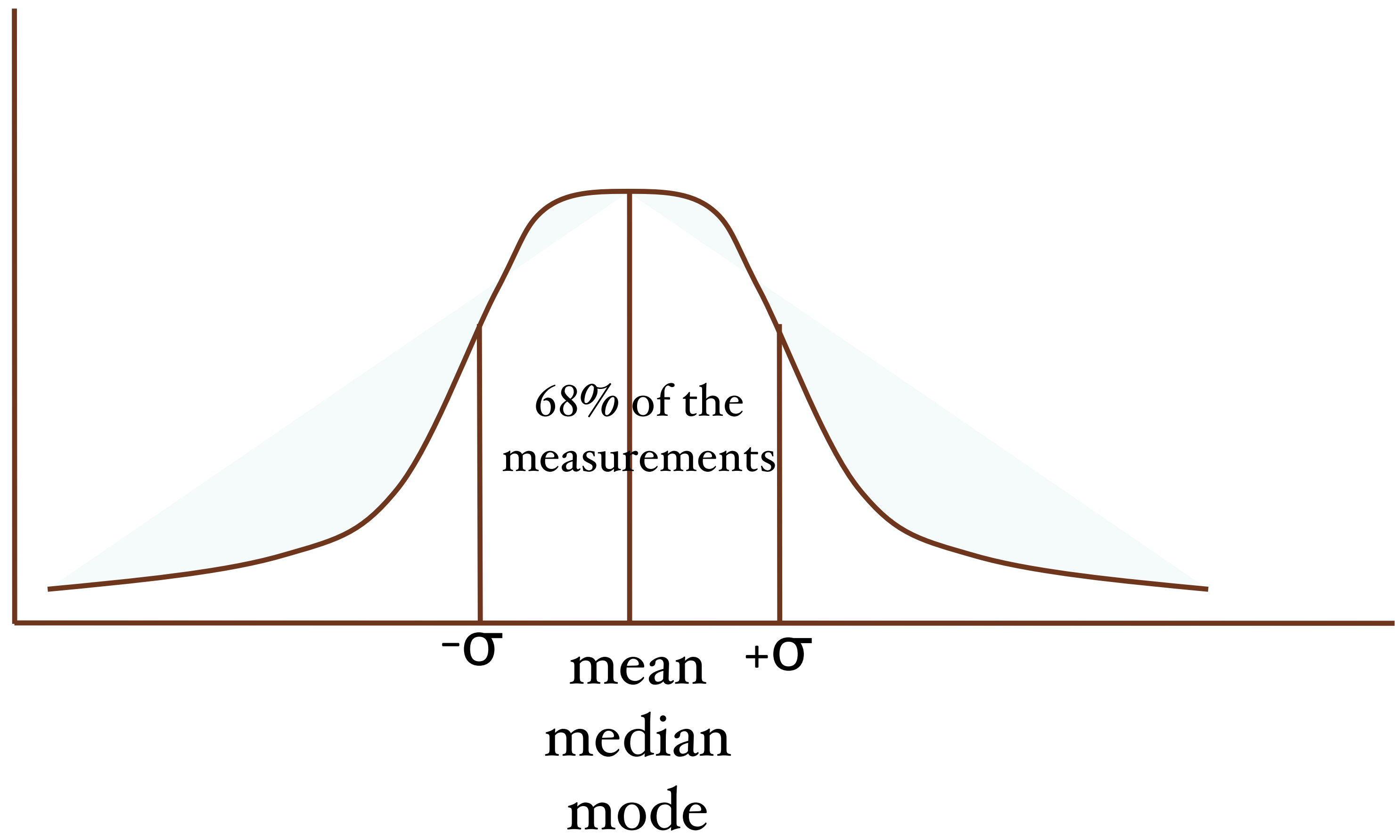
Testing

- Two possible outcomes of test:
  - Rejection of null hypothesis
  - Acceptance of null hypothesis

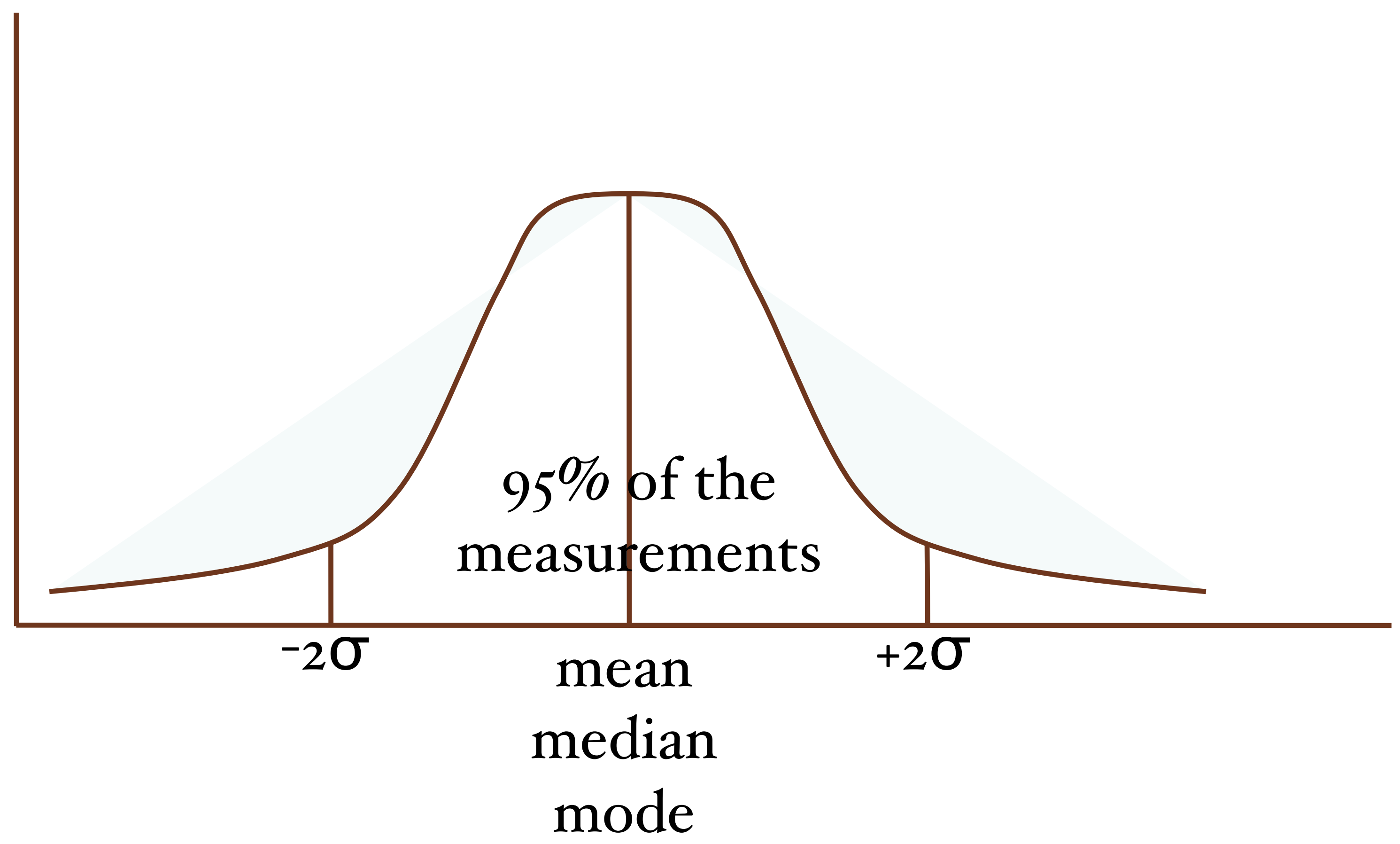
Testing

- z-score: determines the probability of obtaining a given value: How many standard deviations is a value above or below the mean?
  - With: x = value, μ = population mean, σ = population standard deviation

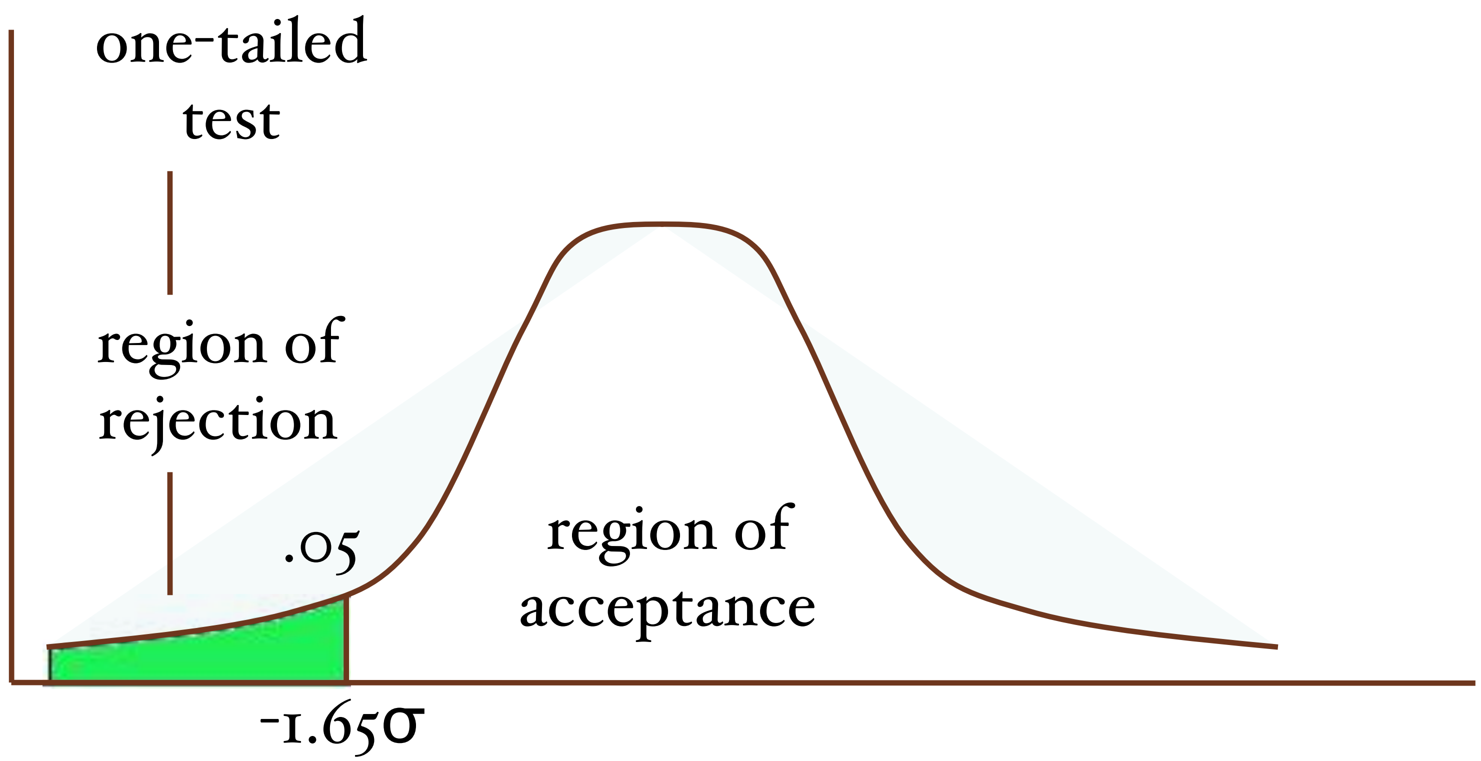
Numerical Statistics



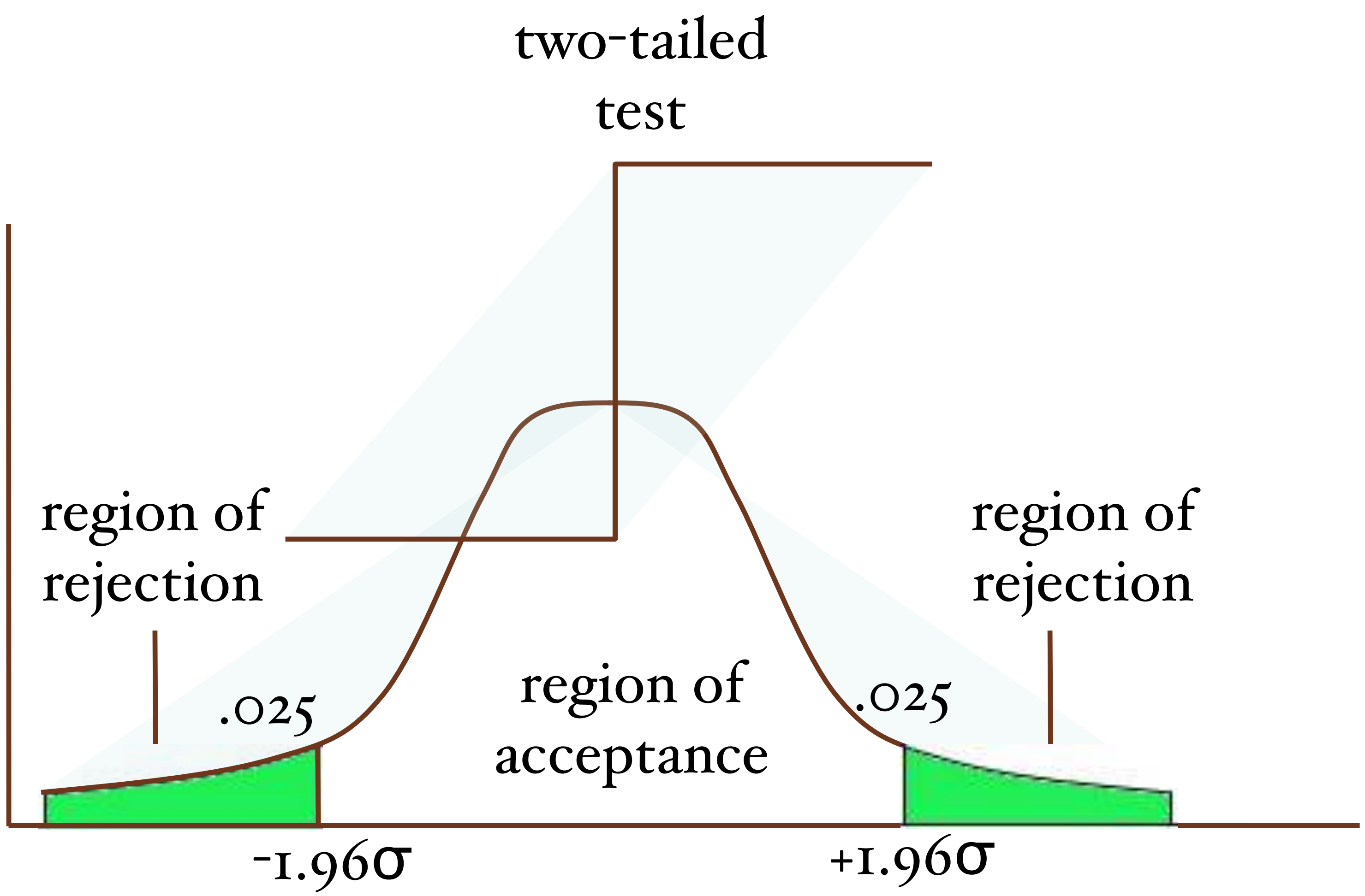
Numerical Statistics



Numerical Statistics



Numerical Statistics



Significance Table

	P	0.99	0.95	0.10	0.05	0.01	0.005	0.001
d.f.	1	0.00016	0.0039	2.71	3.84	6.63	7.88	10.83
	2	0.020	0.10	4.60	5.99	9.21	10.60	13.82
	3	0.115	0.35	6.25	7.81	11.34	12.84	16.27
	4	0.297	0.71	7.78	9.49	13.28	14.86	18.47
	100	70.06	77.93	118.5	124.3	135.8	140.2	149.4

Testing

- Critical value (tabled value):
  - z-value table for probability
  - Cutoff for acceptance or rejection of null hypothesis
    - weakest level: 95% / 5% $p = 0.05$
  - Decision made in advance

Testing

- Probability as significance level
- Example: Collocations
  - Null Hypothesis: independence of two words
  - $P(w1w2) = P(w1)\ P(w2)$

chi-square ($\chi2$) test

- Preferred activities over a population sample of 125 people:

	bowling	dancing	computer	total
male	30	29	16	75
female	12	33	5	50
total	42	62	21	125

chi-square ($\chi 2$) test

- Is the choice of activities related to the gender?
  - If the two variables are independent, we can use these probabilities to predict how many people should be in each cell.
  - If the actual number is different from the expectation for independence, the two variables must be related.

chi-square (χ2) test

- Research Hypothesis:
  - The variables are dependent.
- Null Hypothesis:
  - The variables are independent.

chi-square ($\chi 2$) test

- Overall probability of a person in the sample being:
  - male: 75/125 = .6
  - female: 50/125 = .4

chi-square (χ2) test

- Overall probability of each preference:
  - bowling: 42/125 = .336
  - dancing: 62/125 = .496
  - computer games: 21/125 = .168

chi-square (χ2) test

- Independent events: multiplication rule
  - The probability of two events occurring is the product of their two probabilities.

chi-square ($\chi 2$) test

- Probability of a person in the sample being male and preferring bowling:
  - P(male & bowling): .6 x .336 = .202
  - Expectation: .202 x 125 = 25.2

chi-square ($\chi 2$) test

- Multiplication of row total with column total and division by total number in sample:
- (75 x 42) / 125 = 25.2

	bowling	**dancing**	**computer**	**total**
male	30 (25.2)	29 (37.2)	16 (12.6)	75
female	12 (16.8)	33 (24.8)	5 (8.4)	50
total	42	62	21	125

chi-square (χ2) test

- Formula:

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

$$\chi^2 = \frac{(30-25.2)^2}{25.2} + \frac{(29-37.2)^2}{37.2} + \frac{(16-12.6)^2}{12.6} + \frac{(12-16.8)^2}{16.8} + \frac{(33-24.8)^2}{24.8} + \frac{(5-8.4)^2}{8.4} = 9.097$$

chi-square ($\chi2$) test

- The larger $\chi2$, the more likely the variables are related.
- Square effect of cells with large differences.

chi-square (χ2) test

- Probability distribution of χ2:
  - Critical values in table
  - Degree-of-freedom:
    - df = (number-of-rows - 1) x (number-of-columns – 1)
    - Example: (2 – 1) x (3 – 1) = 2
  - Example: 9.097 (< .025; > .01)

chi-square (χ2) test

- Example: 9.097 (< .025; > .01)
  - Significance (at levels: .05, .01)!
  - Rejection of Null Hypotheses (independence of variables)

chi-square (χ2) test

- Collocations
  - new, companies

	w1=new	**w1¬new**	**total**
w2=companies	8	4667	4675
w2¬companies	15820	14287181	14303001
total	15828	14291848	14307676

chi^2 (χ2) test

- Collocations
  - –ban, derenčin = 267771.9929697935

	r1=ban	r1¬ban	total
r2=derenčin	31	69	100
r2¬derenčin	3019	84972930	84975949
total	3050	84972999	84976049

chi^2 (χ2) test

- Collocations
  - ban, derenčin = ?

	r1=ban	r1¬ban	total
r2=derenčin	31 (0)	69 (99)	100
r2¬derenčin	3019 (3049)	84972930 (84972899)	84975949
total	3050	84972999	84976049

chi^2 (χ2) test

- Collocations
  - ban, derenčin = ?

	r1=ban	r1¬ban	total
r2=derenčin	31 (0)	69 (99)	100
r2¬derenčin	3019 (3049)	84972930 (84972899)	84975949
total	3050	84972999	84976049

- Is this significant?

Assignment

- Read:
  - Shravan & Broe (2010) The Foundations of Statistics: Chapter 1
  - Manning & Schuetze (1999) Foundations of Statistical Natural Language Processing: Chapter 1