Collocations

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

Tests 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

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.
    - H_a: $\mu_1 \neq \mu_2$
    - H_a: $\mu_1 - \mu_2 \neq 0$

Testing

- Null Hypothesis:
  - At IU linguistics students perform the same in statistics as computer science students.
    - $H_0: \mu_1 = \mu_2$
    - $H_0: \mu_1 - \mu_2 = 0$

Testing

- More specific: Research Hypothesis:
  - At IU linguistics students perform better in statistics than computer science students.
    - H_a: $\mu_1 > \mu_2$
    - H_a: $\mu_1 - \mu_2 > 0$

Testing

- More specific: Null Hypothesis
  - At IU linguistics students perform worse in statistics, or equal to computer science students.
    - H_0: $\mu_1 \leq \mu_2$
    - H_0: $\mu_1 - \mu_2 \leq 0$

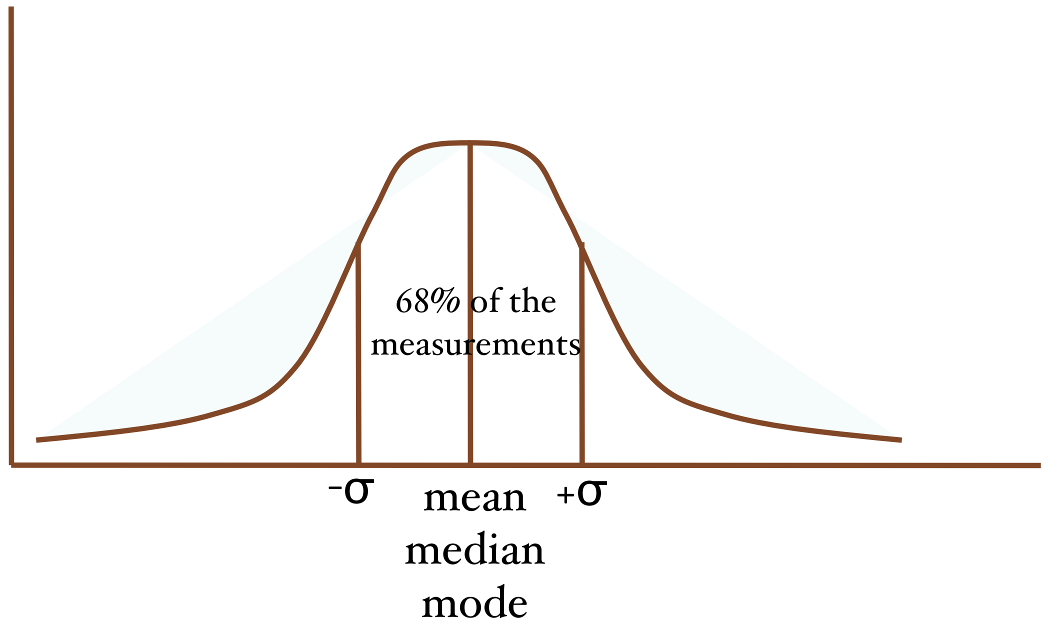
Testing

- 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.

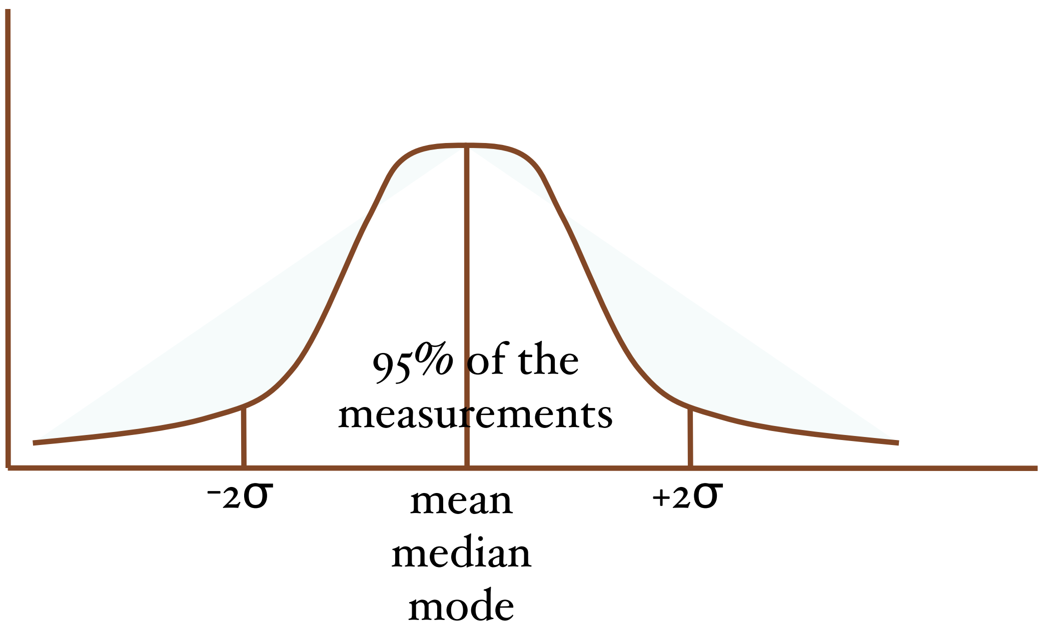
Testing

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

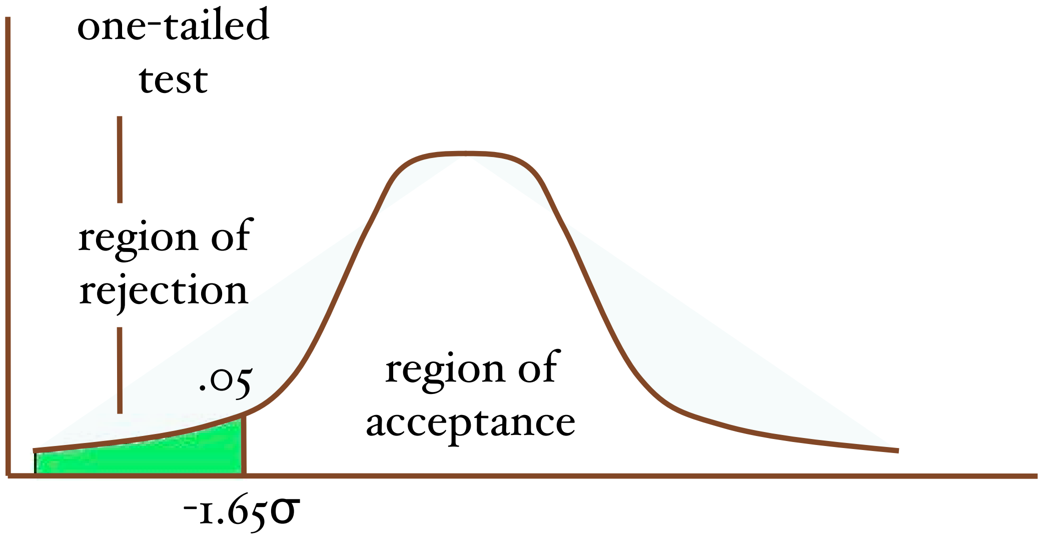
Numerical Statistics



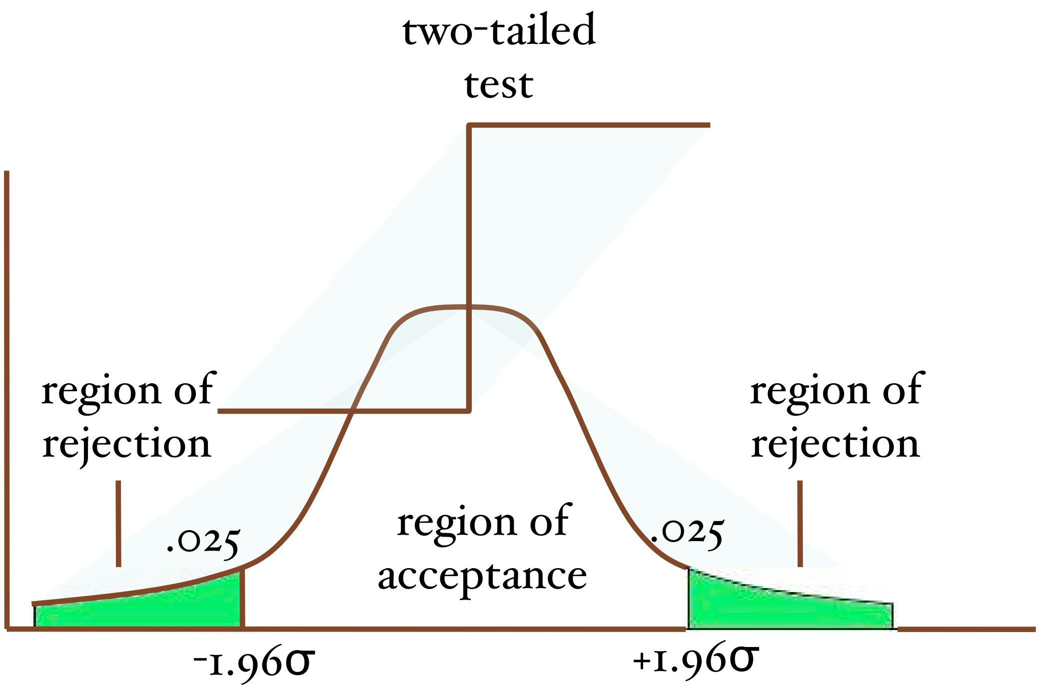
Numerical Statistics



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

- Probability as significance level
- Example: Collocations
  - Null Hypothesis: independence of two words
  - $P(w_1 w_2) = P(w_1)\ P(w_2)$

chi-square (X2) 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 (χ2) 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 (χ2) 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 (χ2) 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 (χ2) 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 (χ^2) 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 (χ^2) test

- The larger χ^2, the more likely the variables are related.
- Square effect of cells with large differences.

chi-square (χ^2) test

- Probability distribution of χ^2:
  - 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 (χ^2) test

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

chi-square (χ^2) test

- Collocations
  - new, companies

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