Statistical Methods in Geography - GEOG 2006
Review – April 2013

INTRO AND DATA HANDLING:

3 types of statistics:

1. Descriptive – describe variables – central tendency, dispersion, distribution shape
2. Inferential – infer characteristics of a population based on a sample statistic
3. Relational – explore the association between variables and possibly make predictions

Data types

Data/values – are numbers organized by variable and observation/case
Data – discrete/categorical vs. continuous
Level/Scale of measurement:

- Nominal – qualitative categories
- Ordinal – ranked/ordered categories
- Interval – continuous, intervals are meaningful, arbitrary zero
- Ratio – as above but with an absolute zero

Standardize data to aid in comparison:
Rates: per unit area, per capita, etc
Proportions/percentages: divide by total

Error/Uncertainty

- All measurements have some uncertainty/error
  - Accuracy:
    - Distance of the measurement to the true value. Inaccuracy from a bias/systematic error - calibrate
  - Precision:
    - Agreement of repeated measurements with each other. Imprecision caused by random error.
    - Numerical precision – how many significant figures are there?
    - Measurement resolution – how refined is the scale of measurement?
- Measurement uncertainty:
  - Last significant digit is uncertain (could be up to half way to the next highest/lowest digit)
- Significant digits/figures:
  - All digits that convey meaning (+ the last one that is uncertain) – excludes placeholder #s
  - Don't add spurious numerical precision - round *after* calc. to correct # of SDs
    - x & ÷ Can have no more sig. figs than carried by the least precise measured #
    - + & - Can have no more decimal places than are in the measured # with the fewest decimal places
    - Counts, constants and conversion factors are not measured #s

Data display

- Reporting values – always use the correct unit
- Caption with Fig/Table #, Title, explanation & source (as warranted) for tables (caption above) and graphs (caption below)
- *Tables* – no grid lines (simple horiz. line above & below header, plus line at bottom).
- *Graphs* – choose appropriate type, label axes (incl. units), no main title (caption instead), simple b&w symbols best, no grid lines (typically).

- scale – typically zoom in to data. Watch for scale tricks, which can mislead!
- *Histograms* – divide data range into equally sized classes/bins, count frequency and make bar plot, # of bins = $\sqrt{n}$ or use Sturges rule (default in R)
- *Boxplot* – Q1,Q2 and Q3 form box, whiskers max and min (or 1.5xIQR away from box, if outliers)

DESCRIPTIVE STATISTICS:

Describe/summarize variables: 1) Central Tendency, 2) Dispersion/Variation/Spread, 3) Shape of distribution

Central Tendency

- Mean – used in many tests, but not robust to outliers, not appropriate for categorical variables
- Median – Q2/50th percentile, more robust to outliers, works for ordinal, interval, ratio data
- Mode/Modal class – most commonly observed value/largest frequency bin

Dispersion

- Range – max-min
- Interquartile range (IQR) - dispersion of the middle 50% of the data, Q3-Q1
- Deviation – difference from the mean, +ve for obs. > mean, -ve for obs. < mean
- Sum of squares – the sum of all the squared deviations in a variable
- Variance – sum of squares, divided by degrees of freedom – large numbers, awkward units
- Standard deviation – the square root of variance, essentially the average of all deviations adjusted for d.f., in sensible units

Quantiles

- Quartiles – rank/sort variable and divide into 4 equal quarters, can break a tie by averaging (among other methods)
- Quantiles – quartile concept can be extended – divide in to 5 quintiles, 10 deciles, 100 percentiles, etc.

Shape of distribution

- Use histogram, boxplot or analyze moments
- Skewness – amt. of asymmetry of a distribution: +ve long tail right; 0 = symmetrical; -ve long tail left
- Kurtosis – how peaked or flat a distribution is: +ve leptokurtic; 0 = mesokurtic; -ve platykurtic

Summarizing a variable

- Always include measure of central tendency *and* dispersion (shape of dist. is impt. too)
  - visual summary – boxplot or histogram
  - table summary – 5 or 6 number summary (min, max, Q1,Q2,Q3, & mean)

PROBABILITY:

- Underpins inferential statistics and our understanding of probabilistic processes
- Theoretical (mathematical), empirical (based on outcome frequency), subjective (opinion)

Theoretical probability

- Sample space (S): all possible outcomes which are mutually exclusive, exhaustive, equally likely to occur
- Event (A): combination of outcomes or single outcomes
- Probability of event A: $P(A) = (A)/(S)$
- Probability is always a number between 0 and 1: $0 \leq P(A) \leq 1$
- Probability of getting any outcome in the sample space is 1: $P(S)=1$
- Therefore the probability of *not* getting A the complement of A: $P(A^c) = 1-P(A)$
- Probability of 2 *disjoint* (no outcomes in common) events (addition) P(A or B) = P(A) + P(B)
- Probability of 2 *independent* events occurring at the same time (multiplication) P(A and B) = P(A) x P(B)
- Statistical independence: outcome of one event has no bearing on the probability of another

Empirical probability

- *The law of large numbers*: Relative frequency will converge on the probability as *n* increases for independent events
- P(A) = Count of A / Count of trials

Probability distributions

- Area under the curve relates to probability; P(total area) = 1
- Tables and software determine the cumulative probability (area to the left of the value) of getting up to a given value by random chance
- With a symmetric probability distribution the area under the left tail at a given value is the same as the area under the right tail at that same value (probability of getting more than that value)
- Use probability rules to determine probabilities that are not in the tails:
  - $P(A^c) = 1-P(A)$
  - P(A or B) = P(A) + P(B)

INFERENTIAL STATISTICS:

Infer characteristics of a population from a sample. Evaluate hypotheses that this characteristic is probable given that it is impossible to know the true value of population characteristics

Samples vs. Populations

- Characteristics of a sample are statistics (depend on the sample/sample error), ...of a population are parameters (invariant and typically unknowable)

Sampling design

By taking a representative sample it is possible to infer characteristics of the population

- Ensure that the research question is: clear, focused, tractable, feasible, original/interesting
- Determine the scope of the study based on scale, controlling/influencing factors
- Define the target population (conceptual) & sampling frame (operational)
- Use a probabilistic sampling design: random, systematic, stratified random, hybrid of these.
- Ensure sampling design accounts for controlling factors (which may bias or confound your result)
- Ensure sample is representative/unbiased – reflecting the relevant characteristics of the pop.

- Ensure *n* is high enough – reduce uncertainty, raise the power of your test

Central Limit Theorem (CLT)

- The sampling distribution (values are a given statistic (e.g., the mean) from all possible samples of a population) will be more normal as n increases
- ($\bar{x}_s$) is centred on the population mean (μ)
- σ_s of the sampling dist. is $\sigma/\sqrt{n}$
- Assumptions: 1) sampled values are mostly independent, 2) sampling probabilistic & unbiased, 3) n is large (>30), 4) n < 10% of N if sampling w/o replacement 5) $\sigma > 0$

Standard error of the mean (SEM)

- The standard deviation of the sampling distribution
- A measure of uncertainty/unreliability of the sample mean

Normalizing data

- To compare data on difference scales
- Data are scaled so that the mean is 0 and the standard deviation is 1
- z-score: indicates how many standard deviations a value is from the mean

Confidence intervals

- The probability that a given interval surrounding the sample mean actually contains μ
- CI expressed as a percent, α is the corresponding probability of μ not being within the CI
  - $\alpha = 1 - (CI/100)$
- Use $t_{\alpha/2}$ if $n < 30$

Philosophy behind hypothesis testing

Deduction – specific conclusions from general principles (flaw: how to get general principle?)
Induction – general conclusions from observing specific examples (flaw: black swan issue)
Science – test falsifiable claims with observations (hypothetico-deductive method)

- Hypotheses – never proven, but they can be falsified

Hypothesis test – is the measured effect due to random chance?

- no = probability of this result being due to random chance is high
- yes = result is significant (significantly higher than what would occur by random chance)

Hypotheses in inferential statistics

- H_0: Null hypothesis – status quo, no change, nothing special, random chance...
- H_1: Alternate hypothesis – there is a difference, change, effect, etc.
- H_0 & H_1 must be mutually exclusive and exhaustive
- Test H_0 – if this can be falsified, you can accept H_1
- Burden of proof is on H_1 – if the evidence is insufficient, then the fall back is H_0

Hypothesis test – decisions

- *Method 1* (takes a computer):
  - Decide on level of significance (α)
  - Calculate the test statistic
  - Determine the probability of obtaining this value (or one that is greater) due to random chance (*p*-value) – this is the area in the tail(s) of the distribution
  - If obtaining this test statistic is very improbable (*p*-value is < α) then reject H_0
- *Method 2* (use a table or computer):
  - Decide on level of significance (α)
  - Calculate the test statistic

- Determine the critical value of the test statistic based on α and the probability distribution – the probability of obtaining a value greater than the critical value due to random chance is α – this is the area in the tail(s) of the distribution (the rejection region)
- If the test statistic is greater (more extreme) than the critical value, then this is more improbable than (α); therefore *p*-value is < α and H_0 can be rejected

1-tailed vs. 2-tailed test?

- Language cues dictate which one to use (e.g., more than, fewer, vs. different from, not the same)
- A 1-tailed test has a single rejection region in one tail only (α)
- A 2-tailed test has two rejection regions (left tail $p=\alpha/2$ and right tail $p=\alpha/2$)
  - delimited by 2 test statistics (e.g., $|z_{\alpha/2}|$)
  - software will determine P(>|test statistic|), which accounts for both tails so p-value can be compared to α

Hypothesis test errors

Type I error

- Rejecting H_0 when it is actually true [an innocent person is wrongly convicted]
- Probability of this occurring is set by α although it may be less than α (if *p*-value is < α)

Type II error

- Not rejecting H_0 when H_1 is actually true [a guilty person is acquitted, due to insufficient evidence]

Degrees of freedom

How many values in a calculation are free to vary?

Probability distributions

- Theoretical distribution described by a *probability density function* (PDF) that takes parameters
- The area under the curve is the probability of finding a value within specific values on the x-axis The area under the entire curve = 1.

Normal distribution – parameters: mean and variance

- a typical distribution for variables with many independent samples, influenced by many random factors

t-distribution – parameter: degrees of freedom (d.f.)

- fatter than a normal distribution at small *n* so this offsets any issues with small sample size

F-distribution – parameters: d.f. numerator and d.f. denominator

- non-symetrical, >0, for testing ratios

χ^2*-distribution* – parameter: d.f.

- non-symetrical, >0, for testing difference between observed and expected

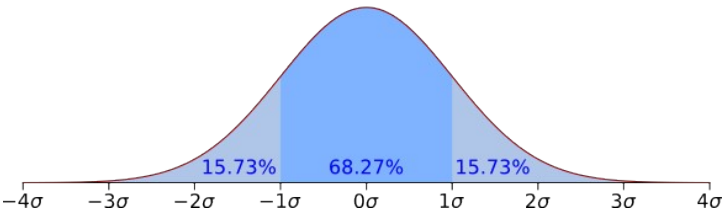


Figure 1, The normal distribution showing probabilities in the shaded areas and z-scores along the x axis. Source: wikimedia

Hypothesis Tests

Single sample difference of means

Test if the mean of a population is the hypothesized value (eg. 50). Can be one or two-tailed.

- H_0: $\mu = 50$, H_1: $\mu \neq 50$
- *Z* test, Test statistic: *Z,* Assumptions: 1,2,3,4
- *t*-test, Test statistic: *t,* Assumptions: 1,2,3

Two sample difference of means

Test if the means (medians for *W*) of two populations are different. Can be one or two-tailed.

H_0: $\mu_A = \mu_B$, H_1: $\mu_A \neq \mu_B$ OR H_0: $\mu_A - \mu_B = 0$, H_1: $\mu_A - \mu_B \neq 0$

- *t*-test, Test statistic: *t* with pooled standard deviation*,* Assumptions: 1,2,3,5
- *t*-test, Test statistic: *t* for unequal variance (Welch's *t*)*,* Assumptions: 1,2,3,6
- Wilcoxon rank sum, Test statistic: *W*, Assumptions: 1,
- paired *t*-test, Test statistic: *t* (paired)*,* Assumptions: 1,2,3,9

Equal variance

Test if the variances from two populations are different. Two-tailed.

H_0: $\sigma_A = \sigma_{B,,}$ H_1: $\sigma_A \neq \sigma_B$

- *F*-test, Test statistic: *F*, Assumptions : 1,2,3

Normal distribution

Test if the population is normally distributed. Two-tailed.

H_0: sample came from a normally distributed population, H_1: sample did not come from a norm pop

- Shapiro-Wilk (W), Assumptions: 1,3

Association between variables

Test if variables are associated with each other

H_0: Variables are statistically independent, H_1: Variables are statistically dependent (two-tailed only)

- Contingency analysis, Test statistic: χ^2, Assumptions: 10,11,12

H_0: $\rho = 0$, H_1: $\rho \neq 0$ (can also be one-tailed)

- Pearson correlation analysis, Test statistic: *t,* Assumptions: 10, 13, 2

H_0: $\rho_s = 0$, H_1: $\rho_s \neq 0$ (can also be one-tailed)

- Spearman rank correlation analysis, Test statistic: Z_{rs}, Assumptions: 10, 14, 8

Linear regression tests

Test that the slope coefficient is significant (can be one or two-tailed)

H_0: $\beta_1 = 0$, H_1: $\beta_1 \neq 0$

- *t*-test, Test statistic: *t*, Assumptions: 10, 13, 2

Goodness of fit (R^2) -test that the model predicts a significant amount of the variance in y (one-tailed)

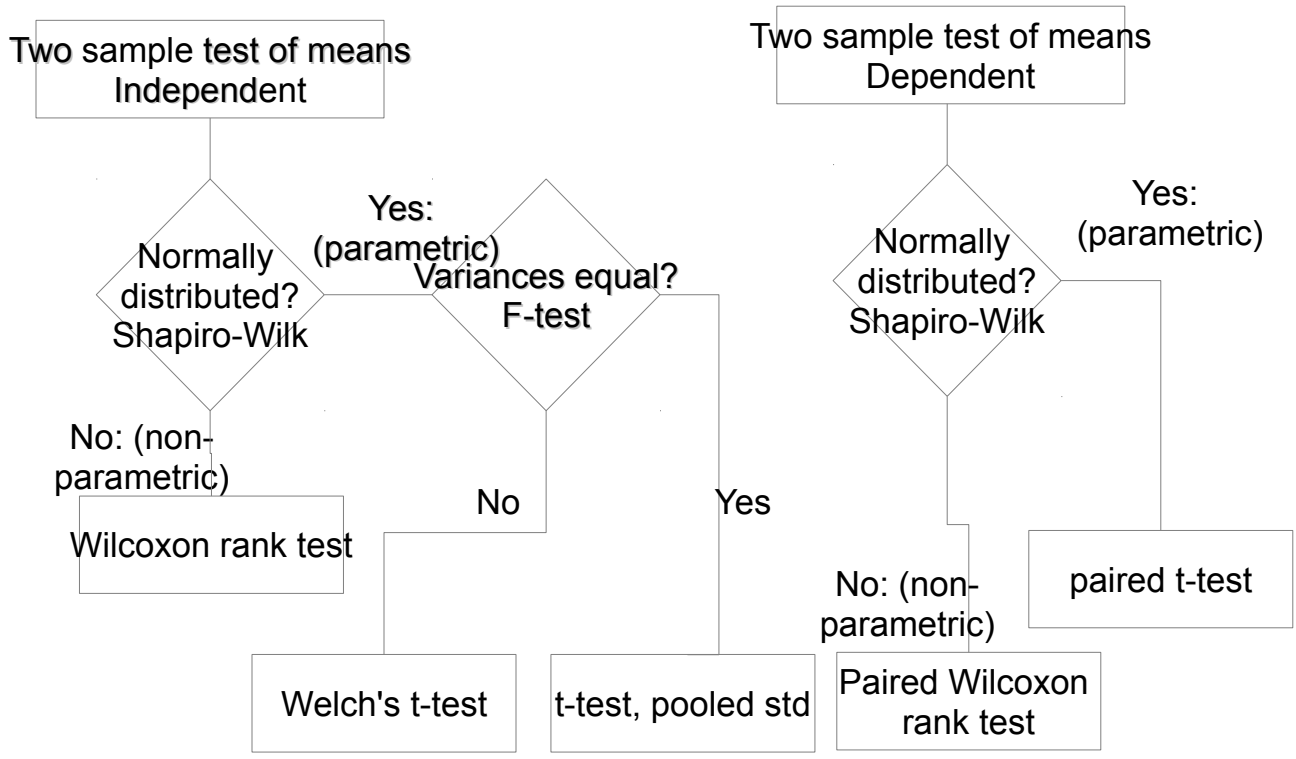
H_0: model is not significant, H_1: model is significant

- *F*-test, Test statistic: *F*, Assumptions: 10, 13, 2

Test assumptions:

1) Random independent sample(s), 2) Population(s) from which the sample(s) is/are drawn is/are normally distributed, 3) Variable measured at the interval/ratio scale, 4) n>30, 5) Population variances are equal, 6) Population variances unequal, 7) Populations have a similar shape, 8) Variables at the ordinal scale (or downgraded), 9) Random samples dependent (paired/matched in time), 10) Random sample of paired variables, 11) 2 categorical variables, 12) No more than 20% of expected frequencies <5, no expected frequencies < 2, 13) Variables have a linear association, 14) Variables have a monotonic association

Difference of means flowchart



RELATIONAL STATISTICS :
Explore relationships between variables by examining scatterplots and testing for associations. Use models to predict y (dependent variable) from x (an independent variable).

Contingency analysis
Determine if two categorical variables are associated with each other. See hypothesis tests.

Scatterplots
Plot continuous variables to examine the direction, form and strength of the association between them

Correlation
Covariance - The amount that the variance in x is related to the variance in y
Correlation - A 'standardized' covariance

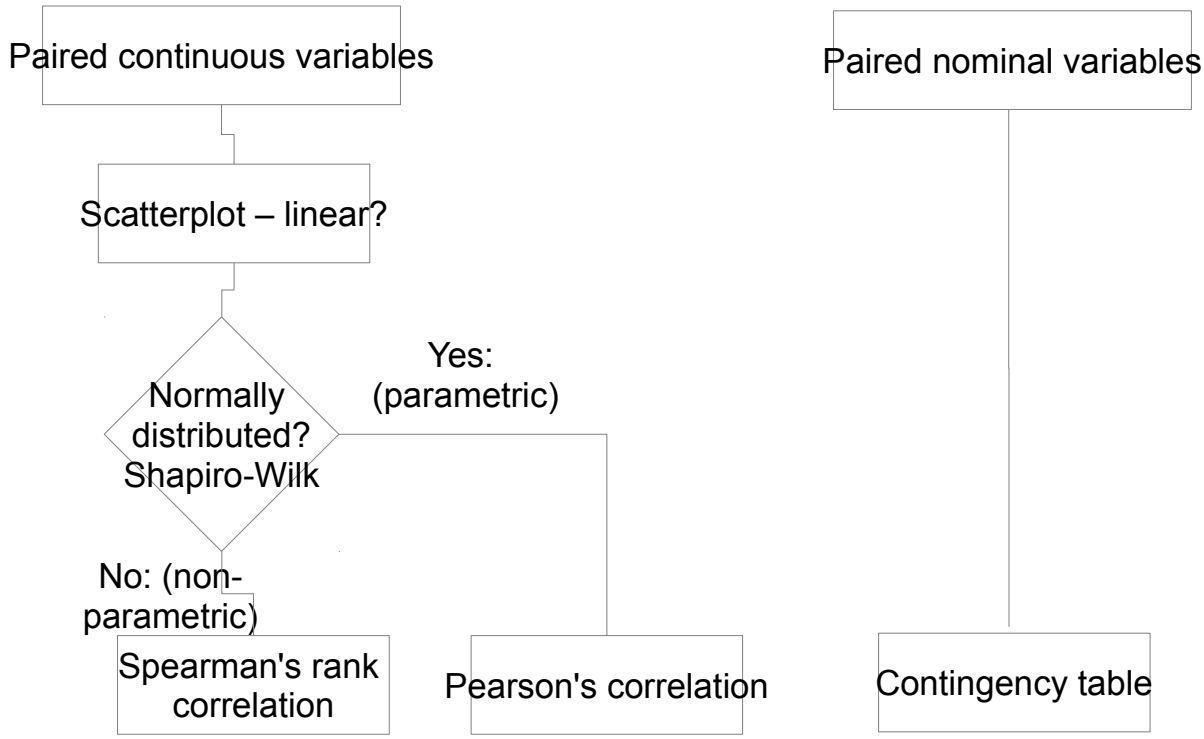
- Pearson's correlation coefficient (r): -1 to +1, 0= no correlation
- Spearman's rank correlation (r_s): -1 to +1, 0=no correlation (non-parametric alternative)
- Can test significance

Linear regression
Model the relationship between x and y to predict y

- Residuals are the difference between the observed and the predicted y
- Minimize the residuals (sum of squared errors: SSE) to fit the best slope for the regression line (b_1)
  - $b_1 = SSXY/SSX$
- The intercept (b_0) completes the model: $y = b_0 + b_1x$ (Population: $y = \beta_0 + \beta_1x + \varepsilon$)
- Model goodness of fit – Coefficient of determination (R^2)

- R^2 = SSR/SSY
- Can test significance of the slope, intercept and the entire model (is R^2 significant)
- Assumptions: 1) relationship is linear between paired continuous variables, 2) Errors/residuals are independent, 3) Equal variance (plot does not thicken), 4) Residuals normally distributed
- Outliers – can change the slope of the regression line
- Causality – good models and high correlation do not guarantee that x causes y!

Testing associations flowchart



SPATIAL CONSIDERATIONS:

- Scale of measurement/study influences interpretation
- Ecological fallacy – do not use aggregate data to draw conclusions on an individual
- Modifiable Areal Unit Problem – spatial aggregations are typically arbitrary
- Autocorrelation – if present, violates the assumption of statistical independence
  - +ve clustered pattern
  - -ve dispersed pattern

Table 1. List of symbols

Symbol	Pop. equiv.	Meaning	Notes
x		a variable	
x_i		the i^{th} value of x	
$\bar{x}$	μ	the mean of x	x bar
$\tilde{x}$		the median of x	tilde
n	N	sample size	
Σ		the sum of	Greek sigma
P		the probability of	
p		the significance of the test statistic P(Type I error)	
α		significance level	Greek alpha
β		statistical power	Greek beta
b_1	β_1	slope	Greek beta
b_0	β_0	y-intercept	Greek beta
δ		difference	Greek delta
r	ρ	Pearson's correlation coefficient	Greek rho, say row
s	σ	standard deviation	Greek sigma
s^2	σ^2	variance	Greek sigma
χ^2		chi squared statistic	Greek chi, say k-eye
ε		model error	Greek epsilon
$\hat{y}$		predicted y	y hat

Key formulas

Mean

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

Median

$$\tilde{x} = ((n+1) \div 2)^{th} \text{ sorted value}$$

Deviation

$$d_i = x_i - \bar{x}$$

Sum of squares

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Skewness

$$Sk = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)s^3}$$

Kurtosis

$$Ku = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{(n-1)s^4} - 3$$

z-score

$$z = \frac{(x_i - \bar{x})}{s}$$

Confidence Interval

$$P(\bar{x} - |z_{\alpha/2}| \times SEM < \mu < \bar{x} + |z_{\alpha/2}| \times SEM) = 1 - \alpha$$

Standard Error of the Mean (SEM)

$$SEM = \frac{s}{\sqrt{n}}$$

Z statistic

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{(n)}}$$

Student's t

$$t = \frac{\bar{x} - \mu}{s/\sqrt{(n)}}$$

Student's t unequal variance (Welch's t-test)

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Student's t equal variance

$$t = \frac{\bar{x}_A - \bar{x}_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

F statistic for equal variance test

$$F=\frac{s_A^2}{s_B^2}(if\ s_A^2>s_B^2) \quad \text{or} \quad F=\frac{s_B^2}{s_A^2}(if\ s_B^2>s_A^2)$$

Contingency analysis – expected

$$E_{ij}=\frac{R_i\times C_j}{N}$$

Chi-squared statistic

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Covariance

$$cov(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{n-1}$$

Correlation

$$r=\frac{\sum_{i=1}^{n}z_{xi}\cdot z_{yi}}{n-1}$$

$$t=\frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Regression

$$SSY=\sum_{i=1}^{n}(y_i-\bar{y})^2 \qquad SSX=\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \qquad SSXY=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \qquad SSR=\sum_{i=1}^{n}(\hat{y}_i-y_i)^2$$

$$b_1=\frac{SSXY}{SSX} \qquad t=\frac{b_1}{se_{b_1}} \qquad se_{b_1}=\sqrt{\frac{SSE/(n-2)}{SSX}}$$