Statistical Testing of Randomness

Masaryk University in Brno
Faculty of Informatics

Jan Krhovják

Basic Idea Behind the Statistical Tests

- Generated random sequences – properties as sample drawn from uniform/rectangular distribution
- Particular tests are based on test statistic
  - Expected value of some test statistic is known for the reference distribution
  - Generated random stream is subjected to the same test
  - Obtained value is compared against the expected value
- Boundless number of statistical test can be constructed
  - Some of them are accepted as the de facto standard
    - NIST battery consists of 15 such tests (e.g. frequency test)
  - Generators that pass such tests are considered “good”
    - Absolute majority of generated sequences must pass

Statistical Hypothesis Testing – Basics

- Null hypothesis (H_0) – denotes test hypothesis
  - H_0 = the sequence being tested is random
- Alternative hypothesis (H_A) – negates H_0
  - H_A = the sequence is not random
- Each test is based on some test statistic (TS)
  - TS is quantity calculated from our sample data
  - TS is random variable/vector obtained from transformation of random selection
  - TS have mostly standard normal or chi-square (χ^2) as reference distributions
- After each applied test must be derived conclusion that rejects or not rejects null hypothesis

Statistical Hypothesis Testing – Errors

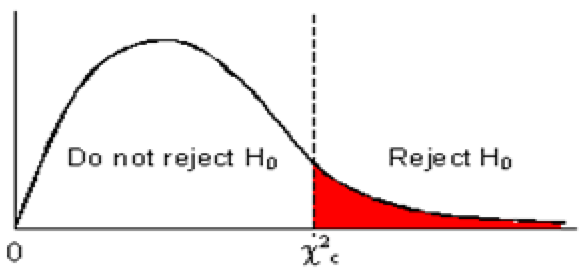
- Conclusion generation procedure and errors

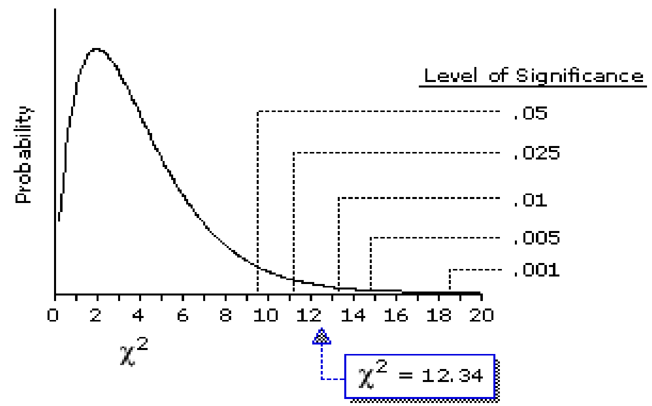
Real situation	Conclusion	
	H_0 is not rejected	**H_0 is rejected**
H_0 is true	good decision	type I error
H_0 is not true	type II error	good decision

- Probability of type I error (α) = level of significance
  - Set prior the test; typically between 0.0001 and 0.01
  - Nonrandom sequence, produced by “good” generator
- Probability of type II error (β)
  - Random sequence, produced by “bad” generator
- α and β are related to each other and to sample size

Statistical Hypothesis Testing – Core

- Critical value – threshold between rejection and non-rejection regions
- Two (quite similar) ways of testing
  - α => critical value; test statistic => value; compare
  - Test statistic => P-value; α; compare





Level of Significance (non-directional test)

df	.05	.025	.010	.005	.001
4	9.49	11.14	13.28	14.86	18.47

critical values of chi-square for **df** = 4

- NIST battery uses P-values
  - $P \leq \alpha$ => reject H_0
  - $P > \alpha$ => do not reject H_0

Frequency (Monobit) Test

- Basic idea
  - Number of zeros and ones expected in the truly random sequence should be the same
- Description
  - Length of the bit string: n
  - Sequence of bits: $\varepsilon = \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$
  - Test statistic: $s_{obs} = |S_n|/\sqrt{n}$
    - $S_n = X_1 + X_2 + \ldots + X_n$, where $X_i = 2\varepsilon_i - 1$ (conversion to ± 1)
    - The absolute value => half normal distribution
- Example (for n = 10)
  - $\varepsilon = 1011010101$; $S_n = 1-1+1+1-1+1-1+1-1+1 = 2$
  - $s_{obs} = |2|/\sqrt{10} = 0.632455532$; P-value = 0.5271
  - For $\alpha = 0.01$: $0.5271 > 0.01$ => ε "is random"

Frequency Test within a Block

- Basic idea
  - Number of zeros and ones expected in a M-bit block of truly random sequence should be the same
  - M = 1 => Frequency (Monobit) Test.
- Description
  - Number of non-overlapping blocks: $N = \lfloor n/M \rfloor$
  - Proportion of ones in each M-bit block: π
  - Test statistic: $\chi^2_{obs} = 4M\Sigma(\pi_i - 1/2)^2$, where $1 \leq i \leq N$
- Example (for n = 10 and M = 3)
  - ε = 0110011010; N_1 = 011, N_2 = 001, N_3 = 101
  - $\pi_1 = 2/3$, $\pi_2 = 1/3$, $\pi_3 = 1/3$; $\chi^2_{obs} = 1$; P-value = 0.8012
  - For $\alpha = 0.01$: $0.8012 > 0.01$ => ε "is random"

Runs Test

- Basic idea
  - A run is the uninterrupted sequence of identical bits
  - Number of runs determines the speed of oscillation
- Description
  - Proportion of ones: $\pi = (\Sigma\varepsilon_i)/n$
  - Test statistic: $\chi^2_{obs} = \Sigma r(k) + 1$, where $1 \leq k \leq n-1$
    - If $\varepsilon_k = \varepsilon_{k+1}$, then $r(k) = 0$, otherwise $r(k) = 1$
- Example (for n = 10)
  - $\varepsilon = 1001101011$; $\pi = 6/10 = 3/5$
  - $\chi^2_{obs} = (1+0+1+0+1+1+1+1+0)+1 = 7$
  - P-value = 0.1472
  - For $\alpha = 0.01$: $0.1472 > 0.01 \Rightarrow \varepsilon$ "is random"

Cumulative Sums Test

- Basic idea
  - A cumulative sums of the adjusted (–1, +1) digits in the sequence should be near zero
- Description
  - Normalizing: $X_i = 2\varepsilon_i - 1$ (conversion to ± 1)
  - Partial sums of successively larger subsequences
    - Forward: $S_1 = X_1$; $S_2 = X_1 + X_2$; ... $S_n = X_1 + X_2 + \ldots + X_n$
    - Backward: $S_1 = X_n$; $S_2 = X_n + X_{n-1}$; ... $S_n = X_n + X_{n-1} + \ldots + X_1$
  - Test statistic (normal distribution): $s_{obs} = \max_{1 \le k \le n} |S_k|$
- Example (for n = 10)
  - ε = 1011010111; X = 1,–1,1,1,–1,1,–1,1,1,1
  - S(F) = 1,0,1,2,1,2,1,2,3,4; s_{obs} = 4; P-value = 0.4116
  - For α = 0.01: 0.4116 $>$ 0.01 => ε “is random”

NIST Testing Strategy

1. Select (pseudo) random number generator
2. Generate sequences
   a) Generate set of sequences or one long sequence
      i. Divide the long sequence to set of subsequences
3. Execute statistical tests
   a) Select the statistical tests
   b) Select the relevant input parameters
4. Examine (and analyse) the P-values
   a) For fixed α a certain percentage are expected to failure
5. Assign Pass/Fail

Interpretation of Empirical Results

- Three scenarios may occur when analysing P-values
  - The analysis indicate a deviation from randomness
  - The analysis indicate no deviation from randomness
  - The analysis is inconclusive
- NIST has adopted two approaches
  - Examination of the proportion of sequences that pass the statistical test
  - Check for uniformity of the distribution of P-values
- If either of these approaches fails => new experiments with different sequences
  - Statistical anomaly? Clear evidence of non-randomness?

Proportion of Sequences Passing a Test

- Example
  - 1000 binary sequences; $\alpha = 0.01$
  - 996 sequences with P-values > 0.01
  - Proportion is 996/1000 = 0.9960
- The range of acceptable proportions
  - Determined by confidence interval
  - $p' \pm 3 \cdot \sqrt{(p' \cdot (1-p')/n)}$, where $p' = 1 - \alpha$; n is sample size
  - If proportion falls outside => data are non-random
- Threshold is the lower bound
  - For n=100 and $\alpha = 0.01$ it is 0.96015
  - For n=1000 and $\alpha = 0.01$ it is 0.98056

Uniform Distribution of P-values

- Interval [0,1] divided to 10 subintervals
- Visually may be illustrated by using histogram
  - P-values within each subinterval are counted
- Chi-square (χ^2) goodness-of-fit test
  - Level of significance $\alpha = 0.0001$
  - Test statistic: $\chi^2 = \Sigma(o_i - e_i)^2/e_i$
    - o_i is observed number of P-values in i^{th} subinterval
    - e_i is expected number of P-values in i^{th} subinterval
      - Sample size multiplied by probability of occurrence in each subinterval (i.e. for sample size n it is n/10)
  - P_T-value is calculated and compared to α
    - P_T-value > 0.0001 => sequence is uniformly distributed

Conclusion

- Randomness testing is based on statistical hypothesis testing
- Each statistical test is based on some function of data (called the test statistic)
- There exists many statistical tests
  - No set of such tests can be considered as complete
  - New testable statistical anomaly can be ever found
- Correct interpretation of empirical results should be very tricky