Presentation of Non-IID Tests

Patrick Hagerty, NSA

December 5, 2012

Partial Credit

- Raw Entropy can be a scarce resource.
- Many sources are not full-entropy.
- Convert pass/fail test to partial credit tests
- Not perfect.

How many outputs are needed to ensure the level of security desired?

Outline

- Entropy
- Entropic Statistics
  - Collision
  - Compression
  - Partial Collection
- 5 Non-IID Tests
  - Frequency
  - Collision
  - Compression
  - Partial Collection
  - Markov
- Why tests apply to non-IID sources

Entropy

- Measure of disorder
- Renyi Entropy

$$H_\alpha(\mathbf{p}) = \frac{1}{1-\alpha}\log_2\left(\sum_i p_i^\alpha\right)$$

- Min-Entropy

$$H_\infty(\mathbf{p}) = \min_i\left(-\log_2 p_i\right)$$

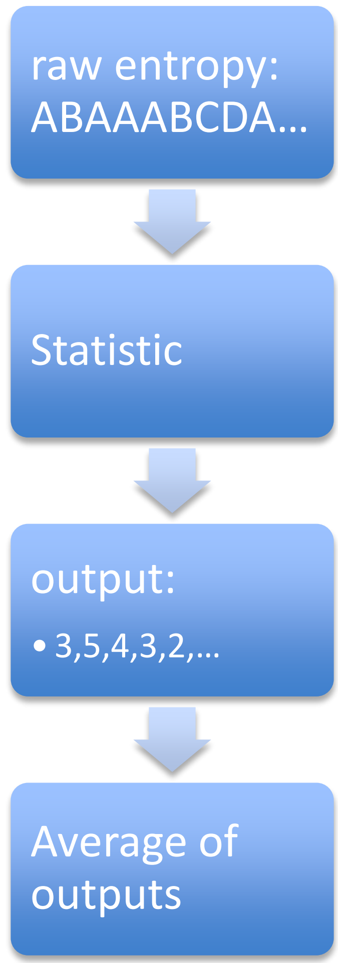
- Shannon Entropy

$$H_1(\mathbf{p}) = -\sum_i p_i \log_2 p_i$$

Min-Entropy

- Most conservative estimate $H_\infty \leq H_\alpha$
- Assumes time independent outputs
- Some dependencies can be accounted
- Efficient on IID sources

General Form Statistic



Ex. 1: Collision Statistic

- Repeat rate without distinguishing states (local)
- data set: ABCAACBCABBCABACBCBCA …
- collision blocks:
  - ABCA ACBC ABB CABA CBC …
- collision repeat rate
  - 4 4 3 4 3 …

Ex. 2: Compression Statistic

- weighted average of repeat rates of every state (global)
- data set: ABCAACBCABBCABACBCBCA …
- compression blocks
  - A: A..AA…A…A.A…..A
  - B: .B….B..BB..B..B.B..
  - C: ..C..C.C…C…C.C.C.
- Compression Repeat Rate
  - A: 1..31…4…4.2…..6
  - B: .2….5..31..3..3.2..
  - C: ..3..3.2…4…4.2.2.
- Compression Repeat Rate
  - 123313524314432432226

Ex. 3: Partial Collection Statistic

- Average repeat rate of every state in a block of output without distinguishing the states within the block (hybrid)
- data set: ABCAACBCABBCABACBCBCA …
- partial collection blocks:
  - ABC AAC BCA BBC ABA CBC BCA
- partial collection repeat rate
  - 0 1 0 1 1 1 0
- partial collection number of distinct elements
  - 3 2 3 2 2 2 3

Conservative Entropy Estimation

$$\mathbb{P} \Longrightarrow S_{\mathbb{P}} \Longleftarrow \mathbb{P}_1 \ \mathbb{P}_2 \ \cdots \ \mathbb{P}_n$$

$$\Downarrow$$

$$H(\mathbb{P}_1) \ H(\mathbb{P}_2) \ \cdots \ H(\mathbb{P}_n)$$

$$\Downarrow$$

$$\min_i H(\mathbb{P}_i)$$

Entropic Statistics

- A real-valued statistic, *S*, is **entropic** with respect to a function *H* if

$$\min_{\{\mathbf{p}:\mathbb{E}_{\mathbf{p}}[S]=m\}} H(\mathbf{p})$$

is monotonic in *m*.

In this case, calculus of variations is easy!

Common Extremal Distributions

- near-uniform

$$\mathbf{p}_\theta[Z=i]=\begin{cases}\theta & i=i_1,\\ \frac{1-\theta}{n-1} & \text{otherwise}\end{cases}$$

- inverted near-uniform

$$\mathbf{p}_\theta[Z=i]=\begin{cases}\theta & i\in\{i_1,\cdots,i_{\lfloor\frac{1}{\theta}\rfloor}\},\\ 1-\lfloor\frac{1}{\theta}\rfloor\theta & i=i_{\lfloor\frac{1}{\theta}\rfloor+1},\\ 0 & \text{otherwise}\end{cases}$$

Non-IID Tests

- Based on Collision, Compression, and Partial Collection.
- Bins and Markov.
- Above tests are conservative.
- Underestimates entropy except in most pathological cases (or non-raw entropy).
- Theoretic bounds on IID, but are relatively robust for non-IID
- Simple compared to complete characterization of sources (years?)

General Setup

- Produce raw output from source
- Truncate (but keep embedded in larger space)
- Convert output into string of statistics
- Average statistics.
- Reduce average by # of standard deviations
- Compute lowest min-entropy distribution that attains the computed value
  - Partial Credit

Conservative Entropy Estimation

$$
\begin{tikzcd}
\mathbb{P} \arrow[r] & S \arrow[r] & B_\epsilon(S_{\mathbb{P}}) \\
 & & \{\mathbb{P}_\alpha \;:\; E[S_{\mathbb{P}_\alpha}] \in B_\epsilon(S)\} \arrow[u] \arrow[d] \\
\min_\alpha H(\mathbb{P}_\alpha) & & \{H(\mathbb{P}_\alpha)\} \arrow[ll]
\end{tikzcd}
$$

Theorems exist for Entropic Statistics: bins, collision, compression, and partial collection.

Collision Test

- Produce raw output until 1000 collisions
  - depends on sample size
- Calculate sample mean collision time, μ
- Calculate sample variance collision time σ^2
- Conservative estimate: $$\bar{\mu} = \mu - \frac{1.96\sigma}{\sqrt{v}}$$
- Find near-uniform distribution such that $$E_{\mathbb{P}_\theta}(S) = \mu$$
- Entropy is: $-\log_2 \theta$

Inversion by Quadrature

$$E_{\mathbb{P}_\theta}(S) = \theta\phi^{-2}\left(1 + \frac{1}{n}\left(\theta^{-1} - \phi^{-1}\right)\right)F(\phi) - \theta\phi^{-1}\frac{1}{n}\left(\theta^{-1} - \phi^{-1}\right)$$

$$F(1/z) = \Gamma(n+1, z)z^{-n-1}e^{-z}$$

Partial Collection Test

- Produce output data set (each output is from a space of *n* elements).
- Partition output into non-overlapping sets of size *n*
- Compute the sample mean and variance of the number of distinct elements in each set μ, σ^2
- Conservatively, account for error

$$\bar{\mu} = \mu - \frac{1.96\sigma}{\sqrt{v}}$$

- Minimize min-entropy of IID probability distributions on the *n* elements that has expected value of the statistic equal to $\bar{\mu}$

Inversion by Quadrature

$$E_{\mathbb{P}_\theta}[S] = 1 - (1-\theta)^n + (n-1)(1-(1-\phi)^n)$$

Compression Test

- Produce a sequence of output data
- Partition into two groups: a dictionary group of the first 1000 outputs and test group.
- Calculate sequence of distance between the index of each element in the test group and the index of the last time the element had appeared in the group.
- Calculate the mean and variance of the number of bits required to record each of these elements. μ, σ^2
- Conservatively, account for error

$$\bar{\mu} = \mu - \frac{1.96\sigma}{\sqrt{v}}$$

- Minimize min-entropy of IID probability distributions on the *n* elements that has expected value of the statistic equal to $\bar{\mu}$

Inversion by Quadrature

Really messy! See pub or reference

Frequency Test

- Bin data.
- Estimate $p_{\max}$
- Use Hoeffding's Inequality to bound actual $p_{\max}$ with specified confidence α

$$\epsilon = \sqrt{\frac{-\log(1-\alpha)}{2N}}$$

- Compute $-\log_2(p_{\max} + \epsilon)$

Markov Test

- Use data to populate transition probability matrix.
- Use Hoeffding Inequality to overestimate transition probabilities
- Write dynamic program to maximize "probability" chain of specified length.
- Entropy of chain is $-\log_2(p_{\max})$.
- Min-entropy plays nicely with Markov models.

Why apply to non-IID

- For IID distributions some tests are more accurate than others.
- For non-IID distribution we are conservative and take the lowest entropy estimate.
- Each test addresses one particular pathology.
- Union of tests gives confidence to the designer and user the entropy estimate.