Finding Outliers in Monte Carlo Computations

Prof. Michael Mascagni

Department of Computer Science
Department of Mathematics
Department of Scientific Computing
Graduate Program in Molecular Biophysics
Florida State University, Tallahassee, FL 32306 **USA**
AND
Applied and Computational Mathematics Division, Information Technology Laboratory
National Institute of Standards and Technology, Gaithersburg, MD 20899-8910 **USA**
E-mail: mascagni@fsu.edu or mascagni@nist.gov
URL: http://www.cs.fsu.edu/∼mascagni

Research supported by ARO, DOE, NASA, NATO, NIST, National Science Fund (Bulgaria) & NSF, with equipment donated by Intel and Nvidia

Outline of the Talk

Introduction and Motivation

- Monte Carlo methods compute quantities of interest by statistical sampling
  1. Partial results are combined to produce a mean and variance used to publish a confidence interval
  2. The partial results are independent, identically distributed (i.i.d.) random variables with finite mean and variance, if the Monte Carlo methods is constructed correctly
  3. These results are summed up in a sample mean and variance, and under the assumptions above, the mean should be normally distributed as per the De Finetti version the Central Limit Theorem (CLT)
- There are many methods that can be used to detect outliers when the underlying distribution is normal
  1. For sample size, $N \leq 30$, one can use Dixon's Q-test
  2. For larger sample sizes, $N > 30$, one can use the significance test of Pearson and Hartley
  3. The last will be a general technique for detecting and removing outliers based on Chauvenet's criterion

Introduction and Motivation

- The consideration of outlier identification and removal leads naturally to other topics
- Topics not considered here
  1. Construction of confidence intervals
  2. Using *p*-values based on the confidence interval parameters to identify outliers
- We will consider the related problem of goodness of fit, specifically whether observed data fit particular distributions
  1. Discrete Distributions: χ^2 test
  2. Continuous Distributions
     - 2.1 The Kolmogorov-Smirnov Test
     - 2.2 The Anderson-Darling Test
     - 2.3 The Shapiro-Wilk Test
  3. Combining different goodness of fitness tests

A Version of the Central Limit Theorem

- Let $X_1, X_2, \ldots, X_N$ be a sequence of i.i.d. random variables with
  1. $\mathrm{E}[X_i] = \mu$
  2. $\mathrm{Var}[X_i] = \sigma^2 < \infty$
- Consider the sample mean

$$S_N = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- Then as N approaches infinity

$$\lim_{n \to \infty} \sqrt{n}(S_N - \mu) \xrightarrow{\text{in distribution}} N(0, \sigma^2)$$

- This is the DeFinetti version of the CLT, and is sufficient for our purposes in Monte Carlo

Why Are Partial Results from Monte Carlo Approximately Normal?

- Given a quantity of interest, Z, one may be able to define a Monte Carlo method to compute it based on
  1. A random/stochastic process that can be realized, S
  2. An estimator of Z, $Est_Z(S)$ with the following properties
     - 2.1 $\mathrm{E}[Est_Z(S)] = Z + b$, where b is the known bias, when $b = 0$ we have an unbiased estimator
     - 2.2 $\mathrm{Var}[Est_Z(S)] < \infty$, the estimator has finite variance
- With such an estimator, the various samples, $Z_1, Z_2, \ldots$, satisfy the conditions for the CLT stated previously
- The sample mean, $\frac{1}{N}\sum_{i=1}^{N} Z_i$, will be approximately Gaussian in distribution
- Thus, to work with outliers in Monte Carlo, it suffices to use methods geared to the normal distribution

Monte Carlo Errors and Elementary Statistics

- Given that we want to compute Z, and the we have a stochastic process and estimator that produces our Monte Carlo estimates: $Z_1, Z_2, \ldots$, we compute via the following three running sums $N, Z^{'}$ and $Z^{''}$:
  1. $N = N + 1$: the number of samples
  2. $Z^{'} = Z^{'} + Z_i$: the running sum
  3. $Z^{''} = Z^{''} + (Z_i * Z_i)$: the running sum of squares
- Then we compute the sample mean and variance as
  1. $\bar{Z} = \frac{1}{N} \sum_{i=1}^{N} Z_i = \frac{Z^{'}}{N}$
  2. $\mathrm{Var}[Z] = \frac{1}{N-1} \sum_{i=1}^{N} \left(Z_i - \bar{Z}\right)^2 = \frac{1}{N-1} \sum_{i=1}^{N} Z_i^2 - 2 Z_i \bar{Z} + \bar{Z}^2 = \frac{1}{N-1} \left(Z^{''} - \bar{Z}^2\right)$
- From above we know that $\bar{Z}$ should be approximately normal with mean and variance given by their estimates
- It is customary to publish Monte Carlo errors as the sample mean plus or minus the square root of the variance of the sample mean: $\bar{Z} \pm \sqrt{\mathrm{Var}[Z]}/N^{1/2}$
- This last value is called the standard error, and provides the variance to construct a confidence interval bases on normal theory for the Monte Carlo estimate

Dixon's Q-test: Detection of a Single Outlier

Theory

- In a set of i.i.d. computations, one or more of the values may differ considerably from the majority of the rest
- In this case there is always a strong motivation to eliminate those deviant values and not to include them in any subsequent calculation
- This is permitted only if the suspect values can be "legitimately" characterized as outliers
- Usually, an outlier is defined as an observation that is generated from a different model or a different distribution than was the main "body" of data
- Although this definition implies that an outlier may be found anywhere within the range of observations, it is natural to suspect and examine as possible outliers only the extreme values.

Dixon's *Q*-test: Detection of a Single Outlier

Theory (Cont.)

- The Dixon's *Q*-test is the simpler test of this type
- This test allows us to examine if one (and only one) observation from a small set of replicate observations (typically 3 to 30) can be "legitimately" rejected or not
- The *Q*-test is based on the statistical distribution of "sub-range ratios" of ordered data samples, drawn from the same normal population
- A normal distribution of data is assumed whenever this test is applied. In case of the detection and rejection of an outlier, *Q*-test cannot be reapplied on the set of the remaining observations

Dixon's Q-test: Detection of a Single Outlier

Practice

The test is very simple and it is applied as follows:

1. The N values comprising the set of observations under examination are arranged in ascending order: $x_1 < x_2 < \cdots < x_N$
2. The statistic experimental Q-value (Q_{exp}) is calculated. This is a ratio defined as the difference of the suspect value from its nearest one divided by the range of the values (Q: rejection quotient). Thus, for testing x_1 or x_N (as possible outliers) we use the following Q_{exp} values:

$$Q_{exp} = \frac{x_2 - x_1}{x_N - x_1} \quad \text{or} \quad Q_{exp} = \frac{x_N - x_{N-1}}{x_N - x_1} \tag{1}$$

Dixon's Q-test: Detection of a Single Outlier

Practice (Cont.)

3. The obtained Q_{exp} value is compared to a critical Q-value (Q_{crit}) found in tables. This critical value should correspond to the confidence level (α) we have decided to run the test (usually: α =95%) Note: Q-test is a significance test.
4. If $Q_{exp} > Q_{crit}$, then the suspect value can be characterized as an outlier and it can be rejected, if not, the suspect value must be retained and used in all subsequent calculations
5. The null hypothesis associated to Q-test is as follows:"There is no a significant difference between the suspect value and the rest of them, any differences must be exclusively attributed to random errors"

- A table containing the critical Q values for different N and α follows

A Table of Critical Values of Q Depending on N and α

N	α=0.001	α=0.002	α=0.005	α=0.01	α=0.02	α=0.05	α=0.1	α=0.2
3	0.999	0.998	0.994	0.988	0.976	0.941	0.886	0.782
4	0.964	0.949	0.921	0.889	0.847	0.766	0.679	0.561
5	0.895	0.869	0.824	0.782	0.729	0.643	0.559	0.452
6	0.822	0.792	0.744	0.698	0.646	0.563	0.484	0.387
7	0.763	0.731	0.681	0.636	0.587	0.507	0.433	0.344
8	0.716	0.682	0.633	0.591	0.542	0.467	0.398	0.314
9	0.675	0.644	0.596	0.555	0.508	0.436	0.370	0.291
10	0.647	0.614	0.568	0.527	0.482	0.412	0.349	0.274
15	0.544	0.515	0.473	0.438	0.398	0.338	0.284	0.220
20	0.491	0.464	0.426	0.393	0.356	0.300	0.251	0.193
25	0.455	0.430	0.395	0.364	0.329	0.277	0.230	0.176
30	0.430	0.407	0.371	0.342	0.310	0.260	0.216	0.165

Table: A Table of Q_{crit} for Dixon's Q-Test

Grubb's Significance Test

- Grubb's test detects one outlier at a time from $X_1, X_2, \ldots, X_N$ assumed normal
- It is based on distinguishing between the following hypotheses:
  1. H_0: There are no outliers in $X_1, X2, \ldots, X_N$ (null hypothesis)
  2. H_1: The is at least one outlier in $X_1, X_2, \ldots, X_N$
- With knowledge of $\overline{\mu}$ and $\overline{\sigma^2}$ from the N data points, the test statistic is:

$$G = \frac{\max\limits_{i=1,\ldots,N} |X_i - \overline{\mu}|}{\overline{\sigma}}$$

- For the two-sided test, the null hypothesis is rejected at significance level α if

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\frac{\alpha}{2N},N-2}}{N-2+t^2_{\frac{\alpha}{2N},N-2}}}$$

where $t_{\frac{\alpha}{2N},N-2}$ denotes the critical value of the t-distribution with $N-2$ degrees of freedom at a significance level of $\frac{\alpha}{2N}$

Pearson and Hartley's Significance Test

- For random samples larger than 30 objects, possible outliers may be identified by using the significance thresholds of Pearson and Hartley
- The test statistic q has to be calculated as follows:

$$q = \left| \frac{X_1 - \bar{\mu}}{\bar{\sigma}} \right|, \text{ where}$$

1. X_1 is the object to be tested
2. $X_1, X_2, \ldots, X_N$ are the data
3. $\bar{\mu}$ is the computed mean of all objects (including the value of X_1)
4. $\bar{\sigma}^2$ is the computed variance of all the objects

- X_1 is regarded to be an outlier if the q exceeds the critical threshold q_{crit} for a given level of significance α and a sample size N

Pearson and Hartley's Significance Test

N	$q_{crit}, \alpha = 0.05$	$q_{crit}, \alpha = 0.01$	N	$q_{crit}, \alpha = 0.05$	$q_{crit}, \alpha = 0.01$
1	1.645	2.326	55	3.111	3.564
2	1.955	2.575	60	3.137	3.587
3	2.121	2.712	65	3.160	3.607
4	2.234	2.806	70	3.182	3.627
5	2.319	2.877	80	3.220	3.661
6	2.386	2.934	90	3.254	3.691
8	2.490	3.022	100	3.283	3.718
10	2.568	3.089	200	3.474	3.889
15	2.705	3.207	300	3.581	3.987
20	2.799	3.289	400	3.656	4.054
25	2.870	3.351	500	3.713	4.106
30	2.928	3.402	600	3.758	4.148
35	2.975	3.444	700	3.797	4.183
40	3.016	3.479	800	3.830	4.214
45	3.051	3.511	900	3.859	4.240
50	3.083	3.539	1000	3.884	4.264

Table: q_{crit} for Various N's and Significance $\alpha = 0.05$ and $\alpha = 0.01$

Chauvenet's Criterion

Procedure

- You have a data set: $X_1, X_2, \ldots X_N$ that is assumed to be $N(\mu, \sigma^2)$
- You want to throw away all observations which are outliers
- This is how you do it using Chauvenet's criterion:
  1. Calculate $\overline{\mu}$ and $\overline{\sigma^2}$ from the $\overline{N}$ data points, and let $Z = \frac{|X_i - \overline{\mu}|}{\overline{\sigma}}$
  2. If $\overline{N} \times \text{erfc}(Z/\sqrt{2}) < \frac{1}{2}$, then we reject x_i
  3. Repeat the previous until all points pass the rejection criterion, using only the non-rejected points
  4. Report the final $\overline{\mu}, \overline{\sigma^2}$, and $\overline{N}$
- When the dust settles, you have two data sets: The set of all good data points, and the set of "bad" points

Please note that the erfc$(\cdot)$ is the complimentary error function:
$\text{erfc}(y) = \frac{2}{\sqrt{\pi}} \int_y^\infty e^{-t^2}\,\mathrm{d}t = 1 - \text{erf}(y)$, where $\text{erf}(y) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^y e^{-t^2}\,\mathrm{d}t$, if $Z \sim N(0,1)$, then
$P[\alpha < Z < \beta] = \frac{1}{2}\left[\text{erf}\left(\frac{\alpha}{\sqrt{2}}\right) - \text{erf}\left(\frac{\beta}{\sqrt{2}}\right)\right]$

What are Goodness of Fit Tests?

- Outliers are problematic items in a data set, and the removal of outliers has many applications
  1. Cleaning up data so that statistical tests can compute representative values
  2. Detecting stochastic computations that are either very unusual or have been corrupted by faults, and hence making them more fault tolerant
- Goodness of fit tests are similar to tests for detecting outliers, but here the assumption is that the entire data set are i.i.d. random variables from a certain distribution
- Goodness of fit tests exist for both discrete and continuous probability distributions
  1. Discrete: Chi-squared test
  2. Continuous: Kolmogorov-Smirnov, Anderson-Darling, and Shapiro-Wilk (note all are expected distribution function (EDF) tests)

The Chi-Squared Goodness of Fit Test

The χ^2 test is for discrete probability distributions, consider the problem of rolling two six-sided dice, which has 36 possible outcomes of which the die totals are the integers 2-12

s : Value of the total of the 2 dice
p_s : Probability of a certain total occurring

s	2	3	4	5	6	7	8	9	10	11	12
p_s	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

If we throw dice $N = 144$ times, here is a possible outcome:

s	2	3	4	5	6	7	8	9	10	11	12
Observed: Y_s	2	4	10	12	22	29	21	15	14	9	6
Expected: Np_s	4	8	12	16	20	24	20	16	12	8	4

The Chi-Squared Goodness of Fit Test

Is a pair of dice loaded? Is this result consistent with the standard discrete distribution?

We can't make a definite yes/no statement, but we can give a probabilistic answer using the χ^2 statistic:

$$\chi^2 = \sum_{1 \le s \le k} \frac{(Y_s - Np_s)^2}{Np_s} = \frac{1}{N} \sum_{1 \le s \le k} \left(\frac{Y_s^2}{p_s} \right) - N$$

Recall that: $\sum_{1 \le s \le k} Y_s = N$ and $\sum_{1 \le s \le k} p_s = 1$

k	: Number of bins
N	: Number of observations
$\nu = k - 1$	: degrees of freedom (one less than the number of bins)

The Chi-Squared Goodness of Fit Test

Consider the following sets of data from 144 rolls:

Value of s	2	3	4	5	6	7	8	9	10	11	12
Experiment 1 Y_s	4	10	10	13	20	18	18	11	13	14	13
Experiment 2 Y_s	3	7	11	15	19	24	21	17	13	9	5

$$\chi_1^2 = 29\frac{59}{120} \qquad \chi_2^2 = 1\frac{11}{120}$$

We look these values up in χ^2 table with $\nu = 10$ degrees of freedom:

χ_1^2 is too high, this value occurs by chance 0.1% of the time

χ_2^2 is too low, this value occurs by chance 0.01% of the time

Both indicate a significant departure from randomness

Rule of thumb: To use the χ^2 test N should be large enough to make each $Np_s \geq 5$

The Chi-Squared Goodness of Fit Test

The general recipe for the Chi-Squared Goodness of fit test

1. *N* independent observations
2. Count the number of observations in in the *k* categories (bins)
3. Compute χ^2
4. Look up χ^2 Chi-Squared distribution table with $\nu = k - 1$ d.o.f.

χ^2<1% or χ^2>99%	reject
1%<χ^2<5% or 95%<χ^2<99%	suspect
5%<χ^2<10% or 90%<χ^2<95%	almost suspect
otherwise	accept

The Chi-Squared Goodness of Fit Test

- Why is the χ^2 distributed as a Chi-squared with $\nu = k - 1$ degrees of freedom? This is Pearson's argument
  1. If $\eta_1, \eta_2, \ldots, \eta_N \sim N(0,1)$ then $Q = \sum_{i=1}^{N} \eta_i^2$ is Chi-squared distributed with N degrees of freedom
  2. Consider $f_i, i = 1, \ldots N$ to be i.i.d. Binomial random variables, and consider $m = \sum_{i=1}^{N} f_i$
  3. $E[m] = Np$, and $var[m] = Np(1-p) = Npq$, so that $\chi = \frac{m - Np}{\sqrt{Npq}}$ will have mean zero and unit variance
     - 3.1 Laplace and DeMoivre, in an early version of the Central Limit Theorem, showed that $\chi \to N(0,1)$
     - 3.2 Thus $\chi^2 = \frac{(m-Np)^2}{(Npq)} = \frac{(m-Np)^2}{(Np)} + \frac{(N-m-Nq)^2}{(Nq)} \to \chi^2(2)$
     - 3.3 Since $\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$ is the sum of squares of approximate normals from the multinomial extension from binomial distribution, it is approximately $\chi^2(N-1)$ since $\sum_{i=1}^{N} O_i = \sum_{i=1}^{N} E_i = N$ removes one d.o.f.

A Table of Chi-Squared Percentages

$$P[X \leq x] = \int_0^x \frac{1}{\Gamma(\nu/2)2^{\nu/2}}\, y^{\nu/2-1} e^{-\nu/2}\, dx$$

ν	0.01	0.025	0.05	0.25	0.50	0.75	0.95	0.975	0.99
1	0.000	0.001	0.004	0.102	0.455	1.323	3.841	5.024	6.635
2	0.020	0.051	0.103	0.575	1.386	2.773	5.991	7.378	9.210
3	0.115	0.216	0.352	1.213	2.366	4.108	7.815	9.348	11.345
4	0.297	0.484	0.711	1.923	3.357	5.385	9.488	11.143	13.277
5	0.554	0.831	1.145	2.675	4.351	6.626	11.070	12.833	15.086
6	0.872	1.237	1.635	3.455	5.348	7.841	12.592	14.449	16.812
7	1.239	1.690	2.167	4.255	6.346	9.037	14.067	16.013	18.475
8	1.646	2.180	2.733	5.071	7.344	10.219	15.507	17.535	20.090
9	2.088	2.700	3.325	5.899	8.343	11.389	16.919	19.023	21.666
10	2.558	3.247	3.940	6.737	9.342	12.549	18.307	20.483	23.209
11	3.053	3.816	4.575	7.584	10.341	13.701	19.675	21.920	24.725
12	3.571	4.404	5.226	8.438	11.340	14.845	21.026	23.337	26.217
15	5.229	6.262	7.261	11.037	14.339	18.245	24.996	27.488	30.578
20	8.260	9.591	10.851	15.445	18.338	23.828	31.410	34.170	37.566
30	14.953	16.791	18.493	24.478	29.336	34.800	43.773	46.979	50.892
50	29.707	32.357	34.764	42.942	49.335	56.334	67.505	71.420	76.154
$n > 30$	$\nu + \sqrt{2\nu}x_p + \frac{2}{3}x_p^2 - \frac{2}{3} + O(\nu^{-1/2})$								
x_p	-2.326	-1.960	-1.645	-0.675	0.00	0.6745	1.6449	1.9600	2.3263

The Kolmogorov-Smirnov Test

Chi-Squared (χ^2): testing data from discrete distributions
Kolmogorov-Smirnov (K-S): testing data from continuous distributions

- N i.i.d. observations of the random variable $X : X_1, X_2, \ldots, X_N$
- X has as cumulative density function (CDF): $F(x) = P(X \leq x)$
- $X_1, X_2, \ldots, X_N$ define the empirical CDF (ECDF) $F_N(x)$:

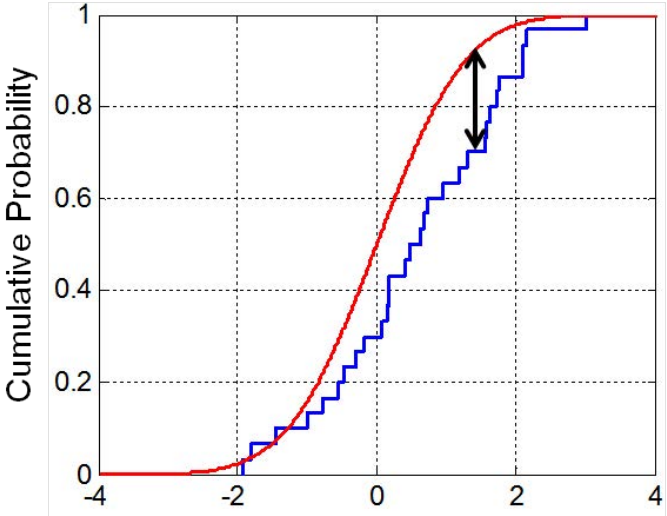
$$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} \chi_{(-\infty, x]}(X_i), \text{ where}$$

$$\chi_I(y) = \begin{cases} 1, & y \in I \\ 0, & y \notin I \end{cases} \text{ is the indicator function of the interval } I \subset \mathbb{R}$$

- Then we define the following measure of deviation of the ECDF from the exact (hypothesized) CDF:

$$D_N^* = \sup_x |F_N(x) - F(x)|$$

A Graph Illustrating How the K-S Statistic is Computed



The Kolmogorov-Smirnov Test

- When $X \sim U[0,1)$ we have

$$F(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases} \quad \text{and so} \quad D_N^* = \sup_x |F_N(x) - F(x)|$$

  is the star-discrepancy, which is used in quasirandom number generation
- The use of the ∞-norm (sup-norm) is the standard topology used in probability for studying distributions via their CDFs
- The standard, two-sided K-S statistic is $K_N = \sqrt{N} D_N^*$
  1. As $N \to \infty$ we get the asymptotic distribution K_∞
  2. Kolmogorov showed that

$$K_\infty = \sup_{t \in [0,1]} |B(t)|, \text{where } B_t := (W_t \mid W_1 = 0),\ t \in [0,1]$$

     is the Brownian Bridge process, using Donsker's theorem
  3. More generally $\sqrt{n} D_n \xrightarrow{n \to \infty} \sup_t |B(F(t))|$

The Kolmogorov-Smirnov Test

Non-central variants of the K-S test based on differences between $F(x)$ and $F_N(x)$

$$K_N^+ = \sqrt{N} \max_{-\infty<x<+\infty} (F_N(x) - F(x))$$

maximum deviation when F_N is greater than $F(\cdot)$

$$K_N^- = \sqrt{N} \max_{-\infty<x<+\infty} (F(x) - F_N(x))$$

maximum deviation when F_n is less than $F(\cdot)$
We get a table similar to χ^2 to find the percentile, but unlike χ^2, the table fits any size of N

Note that $K_N = \sqrt{N} D_N^* = \max(K_N^-, K_N^+)$

The Kolmogorov-Smirnov Test

Simple procedure to obtain K_N^+, K_N^- used the test the null hypothesis

$$H_0\text{: The } X_1, X_2, \ldots, X_N \text{ are drawn from the CDF } F(\cdot)$$

1. Obtain observations $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N$
2. Sort them into $X_1 \leq X_2 \leq \ldots \leq X_N$
3. We use the fact that $F_N(X_j) = \frac{j}{N}$ to calculate K_N^+, K_N^- as follows:

$$K_N^+ = \sqrt{N} \max_{1 \leq j \leq N} \left(\frac{j}{N} - F(X_j) \right)$$

$$K_N^- = \sqrt{N} \max_{1 \leq j \leq N} \left(F(X_j) - \frac{j-1}{N} \right)$$

4. The asymptotic distribution of $\max(K_N^+, K_N^-)$ is given by

$$F_\infty(x) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2} = \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} e^{-(2k-1)^2\pi^2/(8x^2)}$$

Table of Selected Percentiles of the Distributions of K_n^+ and K_n^-

	$p = 1\%$	$p = 5\%$	$p = 25\%$	p = 50%	$p = 75\%$	$p = 95\%$	$p = 99\%$
$n = 1$	0.01000	0.05000	0.2500	0.5000	0.7500	0.9500	0.9900
$n = 2$	0.01400	0.06749	0.2929	0.5176	0.7071	1.0980	1.2728
$n = 3$	0.01699	0.07919	0.3112	0.5147	0.7539	1.1017	1.3589
$n = 4$	0.01943	0.08789	0.3202	0.5110	0.7642	1.1304	1.3777
$n = 5$	0.02152	0.09471	0.3249	0.5245	0.7674	1.1392	1.4024
$n = 6$	0.02336	0.1002	0.3272	0.5319	0.7703	1.1463	1.4144
$n = 7$	0.02501	0.1048	0.3280	0.5364	0.7755	1.1537	1.4246
$n = 8$	0.02650	0.1086	0.3280	0.5392	0.7797	1.1586	1.4327
$n = 9$	0.02786	0.1119	0.3274	0.5411	0.7825	1.1624	1.4388
$n = 10$	0.02912	0.1147	0.3297	0.5426	0.7845	1.1658	1.4440
$n = 11$	0.03028	0.1172	0.3330	0.5439	0.7863	1.1688	1.4484
$n = 12$	0.03137	0.1193	0.3357	0.5453	0.7880	1.1714	1.4521
$n = 15$	0.03424	0.1244	0.3412	0.5500	0.7926	1.1773	1.4606
$n = 20$	0.03807	0.1298	0.3461	0.5547	0.7975	1.1839	1.4698
$n = 30$	0.04354	0.1351	0.3509	0.5605	0.8036	1.1916	1.4801
$n > 30$	$y_p - \frac{1}{6}n^{-1/2} + O(1/n)$, where $y_p^2 = \frac{1}{2}\ln(1/(1-p))$						
y_p	0.07089	0.1601	0.3793	0.5887	0.8326	1.2239	1.5174

The Anderson-Darling Goodness of Fit Test

- The Anderson-Darling (A-D) test is a goodness of fit test, like the better known Kolmogorov-Smirnov test, but it has many of the same building blocks
  1. Data, $X_1, X_2, \ldots X_N$, that has F_N as it's ECDF: $F_N(x) = \frac{1}{N} \sum_{i=1}^{N} \chi_{(-\infty,x]}(X_i)$
  2. The CDF we hypothesize fits the data, $F(x)$
- A-D is a quadratic empirical distribution function statistic of the form:

$$N \int_{-\infty}^{\infty} (F_N(x) - F(x))^2 \, w(x) \, dF(x),$$

  where the weight function for A-D is $w(x) = [F(x)\ (1 - F(x))]^{-1}$, which places extra emphasis on the tails of the distribution
- Thus the A-D statistic, A is given by:

$$A = N \int_{-\infty}^{\infty} \frac{(F_N(x) - F(x))^2}{F(x)\ (1 - F(x))} \, dF(x)$$

The Anderson-Darling Goodness of Fit Test

- A-D test assesses whether a sample comes from a specified distribution
- A-D makes use of the fact that, when given that the data comes from the hypothesized underlying distribution the frequency of the data can be assumed to follow a uniform distribution
- The data can be then tested for uniformity with a distance test (Shapiro 1980)
- The formula for the A-D test statistic A to assess if data $\{X_1 < \cdots < X_N\}$ comes from a distribution with CDF of F

$$A^2 = -N - S, \text{ where, } S = \sum_{i=1}^{N} \frac{2i-1}{N} \left[\ln(F(X_i)) + \ln\left(1 - F(X_{n+1-i})\right)\right]$$

- The test statistic can then be compared against the critical values of the known theoretical distribution
- Note that no parameters are estimated in relation to the CDF with this version of A

The Shapiro-Wilk Goodness of Fit Test

- Shapiro-Wilk (S-W) test checks whether $X_1 \leq X_2 \leq \ldots \leq X_N$ comes from a normally distributed population
- We then form the the following statistic

$$W = \frac{\left(\sum_{i=1}^{N} a_i X_i\right)^2}{\sum_{i=1}^{N}(X_i - \overline{X})^2}, \text{ using}$$

$$(a_1, \ldots, a_N) = \frac{m^{\mathsf{T}} V^{-1}}{(m^{\mathsf{T}} V^{-1} V^{-1} m)^{1/2}}$$

1. Here $m = (m_1, \ldots, m_N)$ are the expected values of the order statistics of N i.i.d. $N(0,1)$ random variables: $X_1 \leq X_2 \leq \ldots \leq X_N$
2. V is the covariance matrix of those order statistics

Combining Goodness of Fitness Tests

- Given that we have a powerful tool to check for goodness of fit of continuous distributions
  1. The Kolmogorov-Smirnov test
  2. The Anderson-Darling test
- And we know the distribution of the K-S and χ^2 statistic, and many others
- One can apply several different goodness of fit tests to improve the ability to evaluate large samples

Combining Goodness of Fitness Tests

- Consider large amounts of data available for the χ^2 test
  1. Assume we have N samples for the χ^2 test with ν degrees of freedom, and N is appropriate based on the rule of thumb, etc.
  2. Also assume that we have r batches of N samples, and they create the following set of statistics: $\chi^2(1), \chi^2(2), \ldots, \chi^2(r)$
- One can now use the K-S test to see if $\chi^2(1), \chi^2(2), \ldots, \chi^2(r)$ with the null hypothesis that they are from $F(\cdot)$ that is χ^2 distributed with ν degrees of freedom

Combining Goodness of Fitness Tests

- Dilemma: We need a large N to tell F_N from F when they differ, but a large N in K-S will average out local random behavior
- Compromise: Consider a moderate size for N, say 1000, appropriate for a single K-S test
  1. Compute a fairly large number of K^+_{1000} on r different parts of the random sequence $K^+_{1000}(1), K^+_{1000}(2), \ldots, K^+_{1000}(r)$
  2. Apply the K-S test to he distribution of K^+_N, which is approximated by

$$F_{1000}(x) \approx F_\infty(x) = P(K \leq x) \approx 1 - e^{-2x^2}$$

- Significance: Detects both local and global random behavior

References I

- T. W. Anderson and D. A. Darling. Asymptotic theory of certain. *Ann. Math. Statist.*, (2):193–212, 06.
- T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.
- W. Chauvenet. *A Manual of Spherical and Practical Astronomy V*. Dover, N. Y., 1960.
- R. B. Dean and W. Dixon. Simplified statistics for small numbers of observations. *Analytical Chemistry*, 23(4):636–638, 1951.
- W. Dixon. Processing data for outliers. *Biometrics*, 9(1):74–89, 1953.
- B. A. Gould. On Peirce's criterion for the rejection of doubtful observations, with tables for facilitating its application. *Astronomical Journal*, 4(83), 1855.

References II

- F. E. Grubbs. Sample criteria for testing outlying observations. *Annals of Mathematical Statistics*, 21(1):27–58, 1950.
- F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.
- D. E. Knuth. *The Art of Computer Programming, Volume 2 (3rd Ed.): Seminumerical Algorithms*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1997.
- A. Kolmogorov. *Sulla determinazione empirica di una legge di distribuzione*, volume 4. 1933.
- Y. Li and M. Mascagni. Analysis of large-scale grid-based monte carlo applications. *International Journal of High Performance Computing Applications*, 17(4):369–382, 2003.

References III

- K. Pearson. On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it can be Reasonably Supposed to Have Arisen from Random Sampling. *Philosophical Magazine, 5th Series*, 50:157–175, 1900.
- B. Peirce. Criterion for the rejection of doubtful observations. *The Astronomical Journal*, 2:161–163, 1852.
- S. Ross. Peirce's criterion for the elimination of suspect experimental data. *Journal of Engineering Technology*, 20(2):1–12, 2003.
- S. S. Shapiro. *How to test normality and other distributional assumptions*. In: The ASQC basic references in quality control: statistical techniques 3, 1980.
- S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

References IV

- N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.*, 19(2):279–281, 06 1948.
- W. Stefansky. Rejecting outliers in factorial designs. *Technometrics*, 14(2):469–479, 1972.
- M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974.

Copyright Notice