# Statistical Literacy and Chance

Milo Schield
Augsburg College
Dept. of Business Administration. 2211 Riverside Drive. Minneapolis, MN 55454

**Abstract**

Statistical literacy studies statistics used as evidence in everyday arguments. This paper studies those aspects of chance that seem most relevant to statistical literacy. Web prevalences of chance-related terms are used to select chance topics relevant to statistical literacy such as 'better than chance', 'due to chance', chance in evolution, the Law of Very Large Numbers, Trojan numbers, informal statistical significance and a Bayesian view of confidence intervals and statistical significance.

## 1. Chance

Chance is an extremely basic concept in everyday usage as seen in the Yahoo prevalence (log10) of these terms in Appendix A: probability (7.6), possibility (8.1), accident (8.2), random (8.5), likely, risk and chance (8.6), certain (8.7), possible (8.9), and can (9.6).

### 1.1. Chance in Statistical Literacy

Consider the Yahoo prevalence of traditional and non-traditional statistical terms. Such rankings help in identifying topics for statistical literacy.

Prevalence of traditional statistical terms: "type 1 error," "type 2 error" (4.6), "prediction interval" biorhythm (4.7), Chi-squared test" (5.0), "hypothesis test," F-test (5.3), "significant result" "sampling distribution" "due to chance" (5.4), "binomial distribution", "random assignment" (5.6), "Central Limit theorem", "statistical tests" "significance level" (5.7), "statistical inference" (5.9), "null hypothesis", p-value, "random sampling" (6.0), "confidence level", "normal distribution" (6.2), "t-test", "statistical significance" "analysis of variance" (6.3), "random sample", "margin of error", "standard error", "confidence interval" (6.4), "random number" (6.7), "statistically significant", "standard deviation," "significant difference" (6.8) and "by chance" (6.9).

Prevalence of other terms with 'chance': "fortune telling" (5.3), "lie detector" (5.4), extraterrestrial (5.8), "bible codes" (5.9), "intelligent design" (6.1), paranormal, superstition (6.2), Tarot (6.3), astrology, UFO (6.4), prophecy, "global warming" ESP (6.5), psychic (6.6) coincidence (6.7) aliens (6.8), ruin (6.9), conspiracy, betting (7.0), gambling, lottery, prayer, bible (7.1), chaos, evolution, fortune disaster (7.2), accident (7.3), religion (7.4), insurance (7.5), risk, God (7.7), death (7.8), weather, sports (8.0), love (8.1) and life (8.2).

### 1.2. Chance in Traditional Statistics

Chance-related topics may be different in statistical literacy than in traditional statistics. McKenzie (2004) asked statistical educators in his session at the 2004 JSM to grade the following 30 statistical topics. The numbers shown in Table 1 are percentages: the count per 100 respondents in each category.

Six statistical inference concepts (*randomness*, *significance, sampling distributions*, *hypothesis tests, confidence intervals* and *random samples*) ranked among the top 9 core concepts.[1]

**Table 1: Statistical Topics Survey Results**

| | Percentage Of All Reponses[2] | Core Concept | TOP 3 Important | TOP 3 Difficult |
|---|---|---|---|---|
| 1 | Variability | 96 | 75 | 12 |
| 2 | Association vs. Causation | 82 | 31 | 6 |
| 3 | Randomness | 77 | 14 | 8 |
| 4 | Significance (Practical/Statistical) | 77 | 14 | 16 |
| 5 | Data Collect (Exp, Obs, surveys) | 75 | 24 | 4 |
| 6 | Sampling Dist (Law Lg. #, CLT) | 71 | 25 | 66 |
| 7 | Hyp. test (crit value, p-value, pwr) | 64 | 22 | 66 |
| 8 | Confidence Interval | 63 | 12 | 16 |
| 9 | Random Sample | 63 | 10 | 4 |
| 10 | Data types | 61 | 8 | 4 |
| 11 | Center | 59 | 6 | 0 |
| 12 | Assumptions | 55 | 8 | 20 |
| 13 | Graphing | 54 | 10 | 0 |
| 14 | Uncertainty | 54 | 10 | 2 |
| 15 | Distributions | 52 | 10 | 14 |
| 16 | Independence | 50 | 4 | 16 |
| 17 | Bias | 48 | 2 | 2 |
| 18 | Correlation | 48 | 2 | 6 |
| 19 | Shape | 45 | 0 | 0 |
| 20 | Data Exploration | 43 | 8 | 0 |
| 21 | Proportion | 41 | 0 | 0 |
| 22 | Least-squares Regression | 39 | 2 | 8 |
| 23 | Models | 38 | 4 | 12 |
| 24 | Comparisons | 38 | 2 | 2 |
| 25 | Prediction | 34 | 2 | 2 |
| 26 | Outliers (aspects of robustness) | 32 | 0 | 0 |
| 27 | Cross-sectional vs. longitudinal | 11 | 0 | 0 |
| 28 | Regression effect | 11 | 0 | 4 |
| 29 | Process | 7 | 0 | 0 |
| 30 | Transformations | 2 | 0 | 10 |

Some educators disagreed:[3] *randomness* (23%), *significance* (23%), *sampling distributions* (29%), *hypothesis*

[1] A core concept is "*a big idea or fundamental principle*."
[2] Maximum marks: Core concepts (**54**), Importance (38), Difficulty (38). Sum of marks: Core concepts (833), Importance (**154**), Difficulty (**150**). Respondents were not limited on core topics (the average respondent selected 17 items), but could only vote for three topics for the Top 3. The number of respondents inferred and used above: Core concepts (56 surveys), Top 3 Importance (51), Top 3 Difficulty (50)
[3] The concepts are not clearly exclusive so votes may have been split.

*tests* (36%), *confidence intervals* (37%), *random samples* (37%) and *bias* (52%).

Textbooks spend many more pages on hypothesis tests and sampling distributions than on association-causation. This may reflect difficulty (66% & 66% vs. 6%) more than importance (64% & 71% vs. 82%).

This list seems to omit some key topics in traditional statistics such as confounding, conditional probability, multivariate modeling/analysis and Bayesian statistics.

## 2. Key Ideas Involving Chance

The following 12 topics are candidates for key ideas involving chance in statistical literacy. While conditional probability and the grammar of rates and percentages are key elements of statistical literacy closely related to chance, they are covered in Schield (2004a).

### 2.1. Chance in Everyday Life

Chance (8.6) is extremely common in everyday usage. According to Wikipedia, chance has three distinct uses: Luck, Randomness and Probability. As luck, chance has a semi-causal status (by chance). As randomness, chance is coincidence. As a probability, chance measures uncertainty. Uncertainty and probability only refer to the last two forms – not to the causal form. Chance can also refer to different kinds of probabilities: analytical, empirical or subjective.

Chance is commonly involved with sports, weather (8.0), gambling (7.1) and betting (7.0). Chance magazine[4] is a good source of articles on these topics.

Chance grammar (8.6) is common but often ambiguous. Chance grammar can blur the distinction between a past prevalence (the percentage of W who are P) and an indefinite event (People who are W are more likely to have P). In a Bernoulli model this difference is irrelevant since the subjects are identical and the outcome prevalence is constant over time. But in reality there is no such guarantee so the move from prevalence to chance may be disputable. Chance grammar can blur the distinction between process and outcome so the random element is unknown. C.f., "the chance that A contains B" (the chance that a 95% confidence interval contains the population parameter). Chance grammar can ignore the context (c.f., "50% chance of rain")

Chance can function as a premise or as a disputable conclusion. E.g., An outcome is unlikely "if due to chance" (1.3) or is unlikely "to be due to chance" (3.3). But "due to chance" (5.3) or "by chance" (6.9) are ambiguous. For example, Sir Ronald Fisher said, "First convince us that a finding is not *due to chance*, and only then, assess how impressive it is." Probability was defined as "the likelihood that results in a test were *due to chance*." "Significance refers to whether a result is extreme enough to be unlikely to have arisen *by chance*."

### 2.2. Better Than Chance

"Better than chance" (4.8) is the claim of various systems for gambling[5] (7.1) and betting[6] (7.0). Some say that lie detectors (5.4) and polygraphs (5.4) aren't any better than chance when used in field conditions.[7] Ekman & O'Sullivan (1991) tested whether we can do much better than chance[8] in deciding if someone is deliberately lying.[9]

"Better than chance" (4.8), "not due to chance" (4.5) or "just a coincidence"[10] (5.9) is commonly involved with psychic[11] phenomena (6.6), extra-sensory perception[12] (6.5), astrology (6.4), tarot (6.3), the paranormal (6.2), bible codes[13] (5.9) and fortune telling (5.3) and biorhythms[14] (4.7). Are their results better than chance? Studies (*A Double-blind Test of Astrology*) say No.[15] One test of psychic power is to see if a person can influence the output of a random number generator. See Jefferys (1990) and Dobyns (1992).

To interpret "better than chance", students must understand chance, the conditional probability involved in accuracy (confirmation vs. prediction), random numbers, chance of rare events, prediction interval versus confidence interval and "statistically significant."

### 2.3. Due to Chance: Evolution

"Due to chance" (5.4) is a key element in evolution (7.2). Some say, "Life is so complex that it is can't be due *just* to chance." A classic reply is "A thousand monkeys, typing on a thousand typewriters will eventually type the entire works of William Shakespeare."[16]

---

[4] The index for 1988-1997 contains entries on baseball, basketball, the bible, figure skating, football, gambling, games, golf, hockey, horseracing, lotteries, marathons, soccer, sports statistics and tennis. www.stat.duke.edu/~dalene/chance/chanceweb/index1to10.html

[5] *Win 86% of the time* www.pokerliving.net

[6] RacingPicks.Com - Horse Race Handicapping ... will predict race outcomes *better than chance*...www.racingpicks.com/basics.htm

[7] http://antipolygraph.org/cgi-bin/forums/YaBB.pl/YaBB.pl?board=Policy&action=display&num=86

[8] For n possible outcomes (n > 1), XP = [R-100%/n)]/(1- 1/n). For n = 5, see the Lawshe (1975) content-validity index (CVI).

[9] College students 52.8%, CIA, FBI & military 55.7%, police 55.8%, trial judges 56.7%, psychiatrists 57.6% and secret service 64.1%.

[10] http://cms.psychologytoday.com/articles/pto-20040715-000008.html, www-class.unl.edu/bios101e/News%20Stories/COINCIDENCE2.htm

[11] Mirror readers beat the odds in my weekly Psychic Challenge… Week after week, your correct responses are significantly *better than chance* would predict. www.digitalzodiac.com/urigeller/nov8.html

[12] E.S.P. Lottery Secrets "Learn to Consistently attain *better than chance* results!" www.mindovermatterovermoney.com/lottery.html

[13] Pro: www.biblemysteries.com/library/codes1.htm
Con: http://cs.anu.edu.au/~bdm/dilugim/torah.html

[14] www.skepdic.com/biorhyth.html

[15] http://skeptico.blogs.com/skeptico/2005/02/what_do_you_mea.html www.skepsis.nl/astrot.html, www.psychicinvestigator.com/demo/AstroSkc.htm

[16] "If 17 billion monkeys on each of 17 billion habitable planets in each of 17 billion galaxies in the universe would be typing away at

But the claim that chance *alone* could produce the evolution of life[17] is extremely weak.

**Model #1**: Pure chance without genetic inheritance. Suppose that an evolutionary process involves rolling a ten-sided die where the desired evolutionary outcome (a pair of songbirds?) is a run of all ones in a thousand tries. For this ten-sided die, we expect a one in the next 10 rolls. The chance that the next thousand rolls of this die will give all ones is 1 chance in $10^{1,000}$ tries. This outcome is *expected* in $10^{1,000}$ tries. If random tries occur at a rate of 317 times/second ($10^{10}$ times/year), then it would take $10^{990}$ years which is immense compared to the age of the earth ($\sim 5 \cdot 10^9$ years). If this model were at all reasonable, it is virtually impossible that life could have evolved "due just to chance alone." This is close to the chance that a tornado can turn a junkyard into a Boeing 747 jumbo jet.[18]

**Model #2**: Chance plus genetic inheritance. Now suppose that genetic inheritance preserves *every* success (one) while all non-successes (non-ones) are subject to random variation. To *expect* 1,000 successes at ten tries per success, we need 10,000 tries: $10^4$. Given the same rate as used above ($10^{10}$ times/year), we obtain the 1,000 ones in $10^{-6}$ years, 0.5 minutes or 30 seconds.

**Analysis:** Look at the change in getting a very unlikely outcome. Chance alone: $10^{990}$ years (much longer than the age of the universe). Chance with genetic inheritance: 30 seconds. Even if the particulars are unrealistic, the point is that a causal process (a non-random process) working together with chance can achieve an unlikely outcome much more quickly than chance alone could possibly do. Evolving apes from single-cell organisms is much more likely if done gradually with genetic inheritance and a small amount of chance than if done in a single step by pure chance alone.

Does saying this make one anti-evolution and pro-intelligent design? No! Evolution is more than *just pure* chance. "Not by chance *alone*" is a red herring in arguing about evolution and intelligent design.

To interpret "due to chance" students needed to understand chance, conditional probability and why something that has one chance in N of occurring is to be *expected* in N tries (Law of Very Large Numbers) even though that outcome may be more unlikely than not.[19]

the rate of one 41 character line per second for 17 billion years, the odds that they would have come up with "To be or not to be, that is the question" would still be only around: 0.000000000005%." http://wetware.hjalli.com/000067.html

[17] Gene mutation, propagation and recombination may be by chance. See www.talkorigins.org/faqs/chance/chance.html

[18] Sir Fred Hoyle: "evolution is as likely as a tornado blowing through a junkyard and assembling a Boeing 747 jumbo jet."

[19] =BinomDist(1,1000,1/1000,0) = 36.81%.

### 2.4. Law of Very Large Numbers

Coincidences (6.7), small chance (6.0) and random events (5.8) are common in everyday usage. The Law of Very Large Numbers says "The unlikely is almost certain given enough tries." (Brignell, 2004)[20]

The chance of K successes in N tries is given by the binomial distribution (where the probability of success is P). Consider success as a run of micro-events each of which is very unlikely. To simplify the math, consider the chance of at least 1 run (K > 0). Let p be the probability of a success per micro-event. Let L be the length of a run. The chance of a run of L successes is P where $P = p^L$. Let N be the number of random samples of size L. Let C be the Chance of at least 1 run of L successes in N tries. So, C equals 1 - Chance of NO run of L successes in N tries of size L.

Eq. 1 Chance (no run L successes| N tries, size L) $= (1-p^L)^N$ for all N >= L.

Eq. 2 C = Chance(≥ 1 run L successes| N tries size L) $= 1 - (1-p^L)^N$ for all N >= L. [21]

The results for a fair coin are shown in Table 2.

**Table 2: Chance(>=1 run L heads| N tries size L)**

| L | L=1 | L=2 | L=3 | L=4 | L=8 |
|---|---|---|---|---|---|
| N = 1 | 0.500 | | | | |
| 2 | **0.750** | 0.438 | | | |
| 4 | 0.938 | **0.684** | 0.414 | 0.228 | |
| 16 | 1.000 | 0.990 | 0.882 | **0.644** | 0.061 |
| 256 | 1.000 | 1.000 | 1.000 | 1.000 | **0.633** |

One expects a run of 8 heads in 8 flips of a fair coin in 256 tries of size 8. As shown, there is a 63.3% chance of *at least one* run of 8 heads in 256 tries. This particular result is an instance of the Law of Very Large Numbers: As the number of tries increases, the unlikely becomes almost certain.

One can solve for the number of trials, N, needed to achieve a certain level of confidence (C) in obtaining at least one run of length L successes in N tries of size L.

Eq. 3 $\ln(1-C) = \ln[(1-p^L)^N] = N \cdot \ln(1-p^L)$

Eq. 4 $N = \ln(1-C) / \ln(1-p^L)$

Table 3 shows the number of sets (N) of length L needed to achieve a level of confidence, C, in obtaining *at least one* run of L heads in flipping a fair coin.

**Table 3: N sets of L for Chance at least 1 set Heads**

| Chance | 50% | 75% | 90% | 99% |
|---|---|---|---|---|
| L | | | | |
| 1 | 1.0 | 2.0 | 3.3 | 6.6 |
| 2 | 2.4 | 4.8 | 8.0 | 16.0 |
| 4 | 10.7 | 21.5 | 35.7 | 71.4 |
| 8 | 177.1 | 354.2 | 588.3 | 1,176.6 |

[20] This is the related to (but not exactly the opposite of) the Law of Small Probability: specified events of small probability do not occur by chance. www.leaderu.com/offices/billcraig/docs/design.html

[21] $C(K|N, P) = \{N!/[K!(N-K)!]\}P^K(1-P)^{(N-K)}$. $C(K=0|N, P) = (1-P)^N$ so $C(K>0|N,P) = 1-(1-P)^N$. If $P = p^L$ then $C(K>0|N, p, L) = 1-[1-(p^L)]^N$.

For a given L, to go from 25.0% to 43.8% to 68.4% to 90% to 99% involves a doubling in N for each step as does going from 29.3% to 50% to 75% to 93.7%.

Now consider the chance of at least 1 run of L successes in N tries where N is the number of tries of size L *in which one desirable event is expected*: $N = (1/p)^L$

Eq. 5 $\quad C = 1 - [1-(p^L)]\text{\^{}}[(1/p)^L]$

If $N = (1/p)^L$ then in the limit of large N the chance of *at least 1 run* of L successes in N tries[22] is [1-(1/e)] or 63.2121%. The chance of at least one run[24] for a given N is never less than this limit.

Suppose that $N = M(1/p)^L$. In the limit of large N, this is the chance of at least one run of L successes.

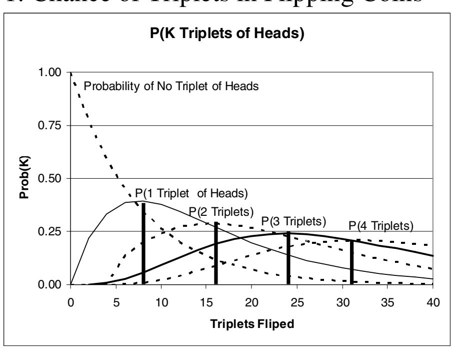
Eq. 6 $\quad C = 1 - [1-(p^L)]\text{\^{}}[M(1/p)^L]$

If M = 1, C = 63.21% in the limit of large N; if M = 2, C = 86.47%; and if M = 3, C = 95.02%. In the limit of large N where $N = M(1/P)^L$, in N tries of size L the chance[23] of at least one run of L successes is $1-(1/e)^M$. This chance for a given N is never less than this limit.[24]

*2.4A Large Number Law; Binomial Distribution*

The binomial distribution gives the chance of *exactly* K successes. Figure 1 illustrates K runs of triplets:

Figure 1: Chance of Triplets in Flipping Coins




If the chance of *at least one* (K > 0) desirable outcome is at least 1-1/e when N = (1/P) where $P = p^L$, then what is the chance of *exactly one* desirable outcome? As shown in Figure 1, this is a maximum when N = 1/P.

One may be tempted to say that getting *exactly* 1 run of L successes in N tries of probability P is *more likely than not* when N = 1/P, but that is false.[25] One can say, "Getting *exactly* one run of L successes in N tries of size L is *more likely than getting no such runs* when N = 1/P. Can we say anything else about *exactly* 1 run?

The chance of exactly 1 success with probability P for N tries where N = 1/P is greater[26] than that for N = 1/P+1 and greater[27] than that for N = 1/P-1 for K > 1. If K = 1 the distribution has two equal modes so the chance of *exactly* 1 run of successes (K = 1) with probability P in N tries is *most likely* when N = 1/P.[28] If K = 0, the distribution has 1 mode

The quantitative form of the Law of Very Large Numbers says, *If the chance of a success is P and if N = 1/P, then there is at least a 63% chance (it is more likely than not) that there will be at least 1 success* and *it is "most likely" that there will be exactly one success.*

*2.4B More Combinations using Smaller Samples*

If N is the number of macro-tires (the number of sets of size L) and if n is the number of micro-tries, then in the aforementioned, n = N·L: 256·8 = 2,048.

A quicker way is to make N+L-1 micro tries and treat adjacent tries as overlapping series: 1 to L, 2 to L+1, etc. so n = N+L-1 = 263. See Schilling (1990).

A still quicker way is to make less than N micro-tries[29] by using the combinatorial technique featured in the birthday problem to obtain N different paths of length L where $N = (1/p)^L$. While this approach requires fewer micro-tries, it is more difficult to explain.

## 2.5. Regression to the Mean

'Regression to the mean' occurs when the extremes on a test tend to move closer to the mean on a retest. The **regression fallacy** is to claim there is always an external determinate cause for the regression effect. An alternate explanation is that some of the extremes (good or bad) on a test are due to chance and cannot be replicated. Evidence for this is that some of the extremes on the *retest* were further from the mean than on their original test so we can't expect them to replicate.

## 2.6. Trojan Numbers

In Greek mythology, a large wooden horse was used as a ruse to enter Troy. Thus, calling something a 'Trojan' means, "this is not what it seems." In statistics, there

---

[22] If $Y = P^{-L}$, $[1-(P^L)]\text{\^{}}P^{-L} = [1-1/Y]^Y$. This well-known limit is 1-1/e.
[23] If $Y = M/(P^L)$, $[1-(P^L)]\text{\^{}}[M(1/P)^L] = [1-(M/Y)]^Y$. Invert & expand:
$[1-(M/Y)]^{-Y} = (1^{-Y}) + (-Y)(1^{-Y-1})(-M/Y)/1! +$
$(-Y)(- Y-1)(1^{-Y-2})[(-M/Y)^2]/2! + (-Y)(-Y-1)(- Y-2)(1^{-Y-3})[(-M/Y)^3]/3!...$
Limit for large Y of $[1-(M/Y)]^{-Y} = 1 +M/1! +M^2/2! + M^3/3! \ldots$
Expand: $e^M = 1 + M/1! + M^2/2! + M^3/3! + M^4/4! \ldots$
Thus in the limit for large Y, $[1-(M/Y)]^{-Y} = e^M$
So for large $M/(P^L)$, $[1-(P^L)]\text{\^{}}[M(1/P)^L] = [1-(M/Y)]^Y = (1/e)^M$.
[24] $dC/dY = d\{1-[1-(M/Y)]^Y\}/dY = -Y[1-(M/Y)]^{Y-1} [-M(-1/Y^2)] < 0$
[25] =BinomDist(K, 256, 1/256, 0) gives 36.72% (K=0) so the chance of at least 1 success is 63.28%: 36.86% (K=1), 18.43% (K=2), 6.12% (K=3), 1.52% (K=4), 0.30% (K=5) and 0.05% (K=6).

[26] $C(K|N=(1/P)+1, P) = \{[(1/P)+1]!/[K![(1/P)+1-K]!]\}P^K(1-P)^{[(1/P)+1-K]}$
$C(K|N=1/P, P) = \{(1/P)!/[K![(1/P)-K]!]\}P^K(1-P)^{[(1/P)-K]}$.
$C[K|N=1/P+1]/C[K|N=1/P]=[\{[(1/P)+1]!/[(1/P)+1-K]!]\}(1-P)^{[(1/P)+1-K]}]$
$/ [\{(1/P)!/[(1/P)-K]!]\}(1-P)^{[(1/P)-K]}] = \{[(1/P)+1]/[(1/P)+1-K]\}(1-P)$
$= (1+P)(1-P)/\{1 + P + P{\cdot}K) < 1$ for all K > 0.
[27] $C(K|N=1/P, P) = \{(1/P)!/[K![(1/P)-K]!]\}P^K(1-P)^{[(1/P)-K]}$.
$C(K|N=(1/P)-1, P) = \{[(1/P)-1]!/[K![(1/P)-1-K]!]\}P^K(1-P)^{[(1/P)-1-K]}$
$C(K|N=1/P,P) / C[K|N=(1/P)-1, P) = [\{(1/P)!/[(1/P)-K]!]\}(1-P)^{[(1/P)-K]}]$
$/ [\{[(1/P)-1]!/ [(1/P)-1-K]!]\}(1-P)^{[(1/P)-1-K]}]= \{(1/P)/[(1/P)-K]\} / (1-P)^{-1}$
$=(1-P)/\{P[(1/P)-K]\}=(1-P)/(1-KP)$. GT/EQ/LT 1 for K GT/EQ/LT 1.
[28] This is true in the strongest sense for K > 1 and true in a weaker sense (no other value of N has a higher probability) for K = 1.
[29] $C = n!/[L!(n-L)!\} < n^L$. If $C = (1/p)^L$ then $n < 1/p$.

are two kinds of Trojan Numbers.[30] One is where the small margin of error for a large sample is taken to be the margin of error for a small subset. The other is where a large size sample with a rare outcome is presumed to have a small margin of error.

The formula for standard error involving proportions is $\sqrt{[P(1-P)/N]}$. For small P, this is $\sqrt{[P/N]}$. The number of expected successes, K, is given by $K = P \cdot N$, so this standard error is $P/\sqrt{K}$. For a rare event, the standard error is determined entirely by K (the number of successes) – not by N (the size of the sample). This is why sample size can be considered a Trojan Number when dealing with percentages and rates. If K is 36, the standard error (P/6) is 1/6[th] the expected value (P). Thus, K > 30 is a good rule of thumb for rare events.

## 2.7. Long-run, Large-sample Bias

In the long run or in larger samples, the sample mean is expected to equal the population mean. But what is expected is not necessarily most likely.[31] When the population mode is different from the mean, then long-run, large-sample bias is expecting that a small sample mean is most likely to equal the population mean when in fact it is more likely to equal the population mode and to move toward the population mean as more data is collected. (Brooks, 2004, Phone conversation).

## 2.8. Informal Statistical Inference

Being "statistically significant" (6.8) is a most common sign of statistical inference. But many news stories don't say if the data is from a sample and if so don't use this phrase or give the sample size or p-value. To evaluate statistical significance, students need simple sufficient conditions. See Harradine (2004) and Pfannkuch & Horring (2004). Here are some rules:

The difference in sample means is statistically significant if the data from two random samples has no overlap, if the boxes (25[th] to 75[th] percentiles) in two box plots[32] do not overlap and the sample size > 9, if the boxes overlap but n > 9 [(IQR of larger box) / (difference in means)]$^2$ or if n > 9 [(larger range) / (difference in means)]$^2$ since Range > IQR.

The difference in sample means is statistically significant[33] if $n > 16/ES^2$ where Effect Size (ES) is the difference in means / pooled standard deviation.

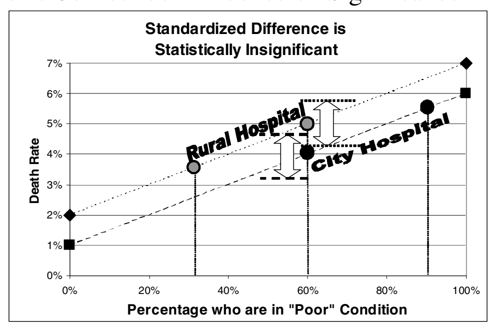
The difference in sample means is statistically significant if n > 9 and the median overlap (percentage of one distribution beyond mean of other) > 25%.[34] See Herrnstein and Murray 1994.

Replacing 9 and 16 with 30 would handle most sample variation in IQR, StdDev and tail probability for non-normal populations and 30 is more memorable.

## 2.9. Formal Statistical Inference

Schield (2004b) argued that statistical significance can be taught very quickly by focusing on the lack of overlap for conservative confidence intervals for proportions where $SE = 1/\sqrt{(2n)}$.[35] Confounder influence on statistical significance (4.5) is important. Schield (2004b) presented a graph (Figure 2) that showed how a confounder could make a statistically-significant association become insignificant (or vice versa).

Figure 2: Confounder Influence on Significance




## 2.10. Bayesian View of Confidence

Confidence intervals (6.4), margin of error (6.4) and confidence level (6.2) are common in everyday usage.[36] But some students think that a 95% confidence interval has a 95% chance of including the population parameter. A Frequentist would say this is true for a future randomly-selected 95% confidence interval but not for an already-selected 95% confidence interval.

Schield (1997) argued that from a decision-making or Bayesian perspective one should regard a Frequentist confidence as the chance of winning a bet about

[30] John Brignell (2004, p. 54) introduced "Trojan Numbers".

[31] Consider 90% group 1 (value = 100) and 10% group 2 (value = 200) so population mean is 110. For n < 10, we don't expect anything from group 2, so a sample mean of 100 (pop. mode & median) is most likely. For n > 10, a sample mean of 110 is most likely.

[32] Std. normal: z(75[th] percentile) = 0.67 so StdDev = 1.33·IQR/2 and StdErr = 1.33(IQR/2)/$\sqrt{n}$. If DM = difference in means then difference is statistically significant (2·SE < DM/2), if $\sqrt{n}$ > 3·IQR/dM or n > [3·IQR/dM]$^2$. If just touching, then DM = IQR so true for n > 9.

[33] Statistically significant if 2SE < DM/2. SE = DM/(ES$\sqrt{n}$).

[34] InvNorm(-0.67) = 25% so DM = 1.34 = IQR and n = 9.

[35] While touching 95% confidence intervals have p-value < 5%, the difference in means is still statistically significant at the 5% level.

[36] Confidence intervals must be distinguished from prediction intervals while standard error and margin or error must be distinguished from standard deviation. Consider this misuse. "from sixty eight different "scientific" uniformitarian measurements…" "we come to a mean age for the earth of about 30 (32.7) million years with a standard deviation of approximately 100 (99.0) million years. … we can be *confident* that the true age of the earth will lie within approximately two ***standard deviations*** on either side of our statistical mean.... we can be *virtually sure* the true age of the earth should be somewhere between **zero** and 230 million years." www.createleaders.org/chronicles/chronicles1999/Earthsage.html Comment: The standard error is really 12 M years so the 95% confidence interval for this non-random data is 30 ± 24 million years.

whether a fixed confidence interval contains a fixed population parameter give no prior knowledge.

It seems professionally negligent not to tell students how they should act when given a 95% confidence interval. Although true, it is not helpful to say, "This confidence interval either does or does not contain the fixed population parameter" while simultaneously saying in a three-door problem "The fixed prize has one chance in three of being behind any of three doors."

## 2.11. Bayesian View of Significance

Statistically significant" and "significant difference" (6.8) are more common then p-value (6.0), "reject the null" (5.3) or "accept the alternate" (4.5). "Statistically significant" means the data is unlikely to occur if due just to chance. "Rejecting the null" means the null is unlikely to be true given unlikely data.[37] Is being statistically significant logically sufficient to reject the null? If not[38], then determining a criterion for rejecting the null may involve Bayesian thinking.

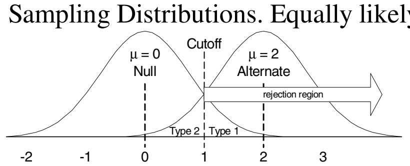
### *2.11.1 Sampling from One of Two Distributions*

Schield (1996) investigated sampling from overlapping distributions named Null and Alternate. The null hypothesis is that a given sample came from the Null distribution. The cutoff value used to reject the null hypothesis was chosen so alpha = beta.[39] If the null and alternate have equal standard errors then the cutoff is midway between the means.[40]

P(Null) = percentage of all samples that are from the Null distribution. P(Null|reject) = percentage of rejected samples that are Null. P(reject|Null) = percentage of Null samples that are rejected = Alpha.

Suppose the Null and Alternate are equally likely, P(null) = 50% in Figure 3, and the null is rejected.

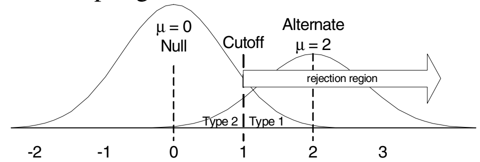
Figure 3: Sampling Distributions. Equally likely




Then the percentage of rejected samples that are null, P(Null|reject), will equal the percentage of null samples that are rejected, P(reject|Null) or alpha (Type 1 error as a percentage of all samples from the null distribution).

If the Alternate is less likely than the Null, P(Null) > 50% (Figure 4), then the percentage of rejected samples that are Null, P(Null|reject), will be larger than alpha.[41]

Figure 4: Sampling Distributions. Alternate Less likely




These results can be summarized in 2x2 tables.

**Table 4: Sampling from One of Two Distributions**

| | Alt | Null | ALL |
|---|---|---|---|
| Reject | | P(R\|N) | |
| Not | Beta | | |
| ALL | 100% | 100% | 100% |

| | Alt | Null | ALL |
|---|---|---|---|
| Reject | | P(N\|R) | 100% |
| Not | | | 100% |
| ALL | P(alt) | P(null) | 100% |

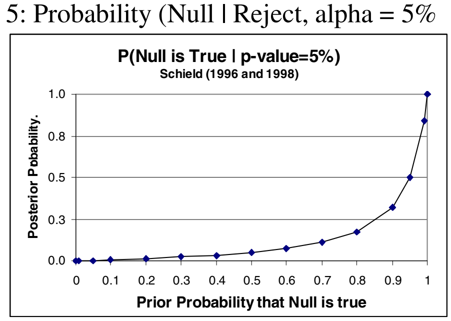
Although the names are suggestive, this is still Frequentist reasoning. Using Bayes rule, it follows that:

Eq. 7: P(Null | reject) = Alpha {P(null) / {[Alpha·P(Null)]+[1-P(Null)][1-Alpha]}

If reject and P(null) < ½, P(null | reject) < alpha.
If reject and P(null) = ½, P(null | reject) = alpha.
If reject and P(null) > ½, P(null | reject) > alpha.

If alpha is 5%, the following pairs give the prior chance the Null is true and the posterior chance the Null is true when the Null is rejected: (1%, 0.1%), (50%, 5%) and (99%, 84%). Only if P(null) ≤ 50% can we reject the Null when P(reject|null) ≤ 5% (Frequentist) and P(Null|reject) ≤ 5% (Bayesian). Figure 5 shows this relationship for an alpha of 5%.

Figure 5: Probability (Null | Reject, alpha = 5%




From Eq. 7, one can determine the alpha needed to obtain a given posterior probability for the truth of the null given the data.[42] Eq. 8 identifies the alpha needed to obtain a 5% posterior probability: P(null|reject)=5%.

Eq. 8: Alpha = 5%·[1-P(Null)] / {5%·[1-P(Null)] + 95%·P(Null)]}

---

[37] Students must be familiar with "*reject the null,*" distinguish "*fail to reject the null*" and "*accept the null,*" and distinguish "*statistically-significant risk factor*" from "*significant risk factor*."

[38] Let P = Null is true. Let Q = some outcomes (z < |2|) are likely. This argument seems of the form, "If P then Q. -Q. Therefore -P." But this form wouldn't fit if Q and -Q are not logical opposites or if the alternate is not the logical opposite of the null.

[39] Given a cutoff, alpha is the chance of Type 1 error given the null is true, beta is the chance of Type 2 error given the alternate is true.

[40] C = Cutoff. Assume SE are equal. Alpha = P(z>C|μ=0), Beta = P(z<$\mu_A$-$C_A$ | μ=$\mu_A$). If Alpha = Beta, $C_A$ = C = $\mu_A$/2. P(Null|data) = P(null)·P(z<$\mu_A$/2|μ=0)/[P(null)·P(z<$\mu_A$/2|μ=0)+P(alt)·P(z<$\mu_A$/2|μ=$\mu_A$)]

[41] The height of the alternate is less than that of the null to illustrate that P(Alternate) < P(null).

[42] Eq 3b, Schield (1996): P(data | Null) = {P(Null | data)·[1-P(Null)]} / {P(Null | data)·[1-P(Null)] + [1-P(Null | data)]·P(Null)}

Alpha decreases very rapidly as the chance that the alternate is true approaches zero.[43] If the alternate is unlikely and if rejection requires the null has no more than a 5% chance of being true (Bayesian), then the alpha required is approximately 5% times the prior chance the alternate is true.[44]

*2.11.2 Application to Statistical Significance*

Schield (1996, 1998) argued that this Frequentist model serves as a suitable Bayesian view of statistical tests. Gonen et al. (2005) argued similarly for a Bayesian Two-Sample t Test using a more complex model. From a Frequentist perspective, this move is unjustified. If the truth or falsity of the null is a fact then there is no Frequentist probability that the null is true. But given the increasing use of significance tests involving alternates whose efficacy is unknown or unlikely, Frequentists may be motivated to present a Bayesian view hypothetically saying "If hypothesis testing were similar to identifying which of two distributions was the source for a random sample…" Consider the benefits.

First, this model is consistent with the Principal Principle: that Frequentist and Bayesian analyses will give similar answers when one has no other information on the matter – when the prior is uninformative (the null and alternate are equally likely). See Schield (1997) and Aitkin (1991, 1998). In this model if the prior chance of the null being true is 50%, then the posterior chance of the null being true on rejection will be alpha.

Second, this model provides qualified support for *Fisher's Rejection Rule*: reject the null when the p-value is less than 5%. When R. A. Fisher formulated hypothesis testing, horticulturalists had good reasons to think the effects were more likely due to the treatment than to chance (Pnull < 50%) so Fisher's rule works.

Third, this model identifies conditions where using Fisher's Rejection rule is inappropriate. Today hypothesis testing evaluates the influence of treatments or exposures that seem less likely to be true than is the null: Pnull > 50%. In these cases, rejecting the null using Fisher's guideline means doing so when the posterior chance the null is true given data to reject is greater than 5%.[45] This is Lindley's paradox: when Bayesian and Frequentist tests result in contradictive evidence. (See Shafer, 1982 and Jefferys, 1995.)

A Bayesian rejection rule would be to reject the null only when the sample statistic is '95% Bayes' significant': P(Alternate | Data) ≥ 95%. If rejection is justified only when the result is 'Bayesian significant', then this simple model shows that being 'statistically significant' is not sufficient to be 'Bayes significant.'

Schield (1998) recommended presenting the prior needed to reject the null with a 95% Bayes' confidence in the same way Frequentists present a p-value.[46]

If Frequentist hypothesis testing is to be influenced by Bayesian reasoning, then a first step might be to restrict Fisher's rule for rejecting the null hypothesis to those cases where the alternate is more likely than the null.[47] This might prevent some of the results in section 2.2 from being touted as being "statistically significant."

### 2.12. Significance for Rare Events

John Brignell (2000) claimed a relative risk must be at least 2 to be statistically significant for rare outcomes. Assume that RR = 1 in the population so that the chance of the desired outcome is the same in both exposure and control groups. Assume that we randomly sample for just the exposure group so the mean of the control group is the same as that in the population. Assume the outcome of interest is rare (P<< 1%) and that the sample sizes (N) are quite large, so the number of outcomes expected (K) is greater than 1 since K=N•P. In this case, the frequency of rare events is Poisson. The variance of the Poisson equals the expected value (K). The standard deviation (SE of the distribution) is the square root of the variance: √K. The upper limit of a 95% ***confidence*** interval is the mean (the # expected) plus 2 standard deviations (SE). This is [K + 2√K]. In risk, the upper limit is P + 2√(P/N)

Assuming N1 in the test group and N0 in the control (where P1=P0=P), the resulting relative risk due to chance is [P + 2√(P/N1)]/P = 1+2√[1/(P·N1)] = 1+2/√K. For K=4, the ratio is 2. This rule may apply for larger K since we excluded variation in the control group.[48] The argument is reversible. If a relative risk of 1.2 is to be statistically significant then at least 100 events of interest are needed in the test group. If the outcome prevalence is 1%, this requires 10,000 subjects.

[43] If P(Alternate|reject) = 95%, then these pairs relate P(alternate) and alpha: (50%, 5%), (10%, 0.58%), (1%, 0.053%), (0.1%, 0.0053%).

[44] If P(Alternate) << 1, then per Eq. 8 alpha ~ 5%.P(Alternate).

[45] It would seem that the smaller the p-value, then the greater the evidence for rejecting the null. This is true if the sample size is increased but not necessarily true if the null and alternate vary.

[46] Mathews (1999) used a similar criterion, a 95% confidence that null is false, but added sampling from a distribution. . Assuming a 50% chance the alternate is a fluke (where the alternate is distributed uniformly), gave a two-tailed probability of 17% that the null is true given the data for a p-value of 5%. Mathews proposed a new criteria for significance: "no suggestion of significance should be made unless Pr(Fluke | data) < 0.05." Mathews Home Page: http://ourworld.compuserve.com/homepages/rajm/

[47] This seems less adventuresome than banning "statistically significant" (Mathews) or introducing another term ('Bayes significant').

[48] "It can be roughly reckoned that statistical variation will dominate for small samples, while confounding factors dominate for larger samples. In either case the minimum risk ratio that can be remotely acceptable is 2.0." P. 103. "…real scientists never accept an RR < 2." P. 102. See www.numberwatch.co.uk/2002%20July.htm

## 3. Work Needed

More work is needed to identify the role of chance in everyday arguments: in betting and insurance (c.f., the chance of ruin), in decision making (c.f., the value of perfect information) and in describing the chance an outcome is "clinically significant."[49] Work is needed on the possible influence of chance on the Lhotka curve (Murray, 2003), on Benford's law (Brooks, 2002) and on the extreme values of a distribution such as floods, peak temperatures and winners (Brignell, 2004).

## Acknowledgments

This work was performed under a grant from the W. M. Keck Foundation "to support the development of statistical literacy as an interdisciplinary curriculum in the liberal arts." Dr. Joel Best, Dr. Paul Alper, Dr. Larry Lesser and Dr. Robert Raymond provided articles and gave feedback on this paper. This is version 5b of *2005SchieldASA.pdf*.

## References

Aitkin, Murray (1991). *Posterior Bayes factors (with discussion)*. Journal of Royal Statistical Society B, 53(1):111-142.

Aitkin, Murray (1998). *Simpson's paradox and the Bayes factor*. Journal of Royal Statistical Society B, 60(1):269-270.

Brignell, John (2000). *Sorry, Wrong Number! The Abuse of Measurement.* Brignell Associates. P. 102-3.

Brignell, John (2004). *The epidemiologists.* Printing Data.com

Brooks, Dean (2002). *Practical Applications of Benford's Law for Integer Quantities.* www.StatLit.org/pdf/2002BrooksSpecial1.pdf

Dobyns, York H., (1992). *On the Bayesian Analysis of REG Data*. Jrnl of Scientific Exploration, Volume 6, Number 1, pp 23-46.

Ekman, P., & O'Sullivan, M. (1991). *Who can catch a liar?* American Psychologist, 46, 913-920.

Gonen, Mithat, Wesley Johnson, Yonggang Lu and Peter Westfall. (2005). *The Bayesian Two-Sample t Test* The American Statistician, Vol 59, No. 3. P. 252-257.

Harradine, Anthony (2004). *Within, Between and Beyond.* IASE Curriculum Design Round Table.

Jefferys, William H. (1990). *Bayesian Analysis of Random Event Generator Data.* Journal of Scientific Exploration, Volume 4, Number 2. pp.153-170.

Jefferys, William H. (1995). On p-values and chance. Journal of Scientific Exploration, Volume 9, Number 1. pp.121-122.

Lawshe, C. H. (1975). *A quantitative approach to content validity*. Personnel Psychology, 28: 563-575.

Lindley, D. V.(1957). *A statistical paradox*. Biometrika, 44:187-192.

Mathews, Robert, (1999). *Significance Levels for the Assessment of Anomalous Phenomena.* Journal of Scientific Exploration, Volume 13, Number 1, pp. 1-9.

McKenzie, John D., Jr. (2004). *Conveying the Core Concepts.* Proceedings of ASA Statistical Education Section, pp. 2755-2757.

Herrnstein, Richard and Charles Murray (1994). *The Bell Curve.* The Free Press.

Murray, Charles (2003). *Human Accomplishment.* Harper Collins.

Pfannkuch, Maxine and Julia Horring (2004). *Developing Statistical Thinking in Secondary School.* IASE Curriculum Roundtable.

Schield, Milo (1996). *Using Bayesian Inference in Classical Hypothesis Testing*. Proceedings of the ASA Section on Statistical Education, pp. 11-16.

Schield, Milo (1997). *Interpreting Statistical Confidence.* Proceedings of the ASA Section on Statistical Education.

Schield, Milo (1998a). *Using Bayesian Strength of Belief to Teach Classical Statistics.* Intrntnl Conference on Teaching Statistics.

Schield, Milo (1998b). *Statistical Literacy and Evidential Statistics*. Proceedings of the ASA Section on Statistical Education.

Schield, Milo (1999). *Difficulties in Describing and Comparing Rates and Percentages.* ASA Section on Statistical Education.

Schield, Milo (2004a). *Statistical Literacy Curriculum Design.* IASE Curriculum Design Roundtable.

Schield, Milo (2004b). *Three Graphs for Statistical Literacy.* International Conference on Mathematics Education (ICME-10).

Schilling, Mark (May 1990). *The Longest Run of Heads***.** College Mathematics Journal, Volume 21, No. 3, pp. 196-207.

Shafer, Glenn (1982). *Lindley's paradox (with discussion)*. Journal of American Statistical Association*,* 77(378):325-351.

[49] www.acponline.org/journals/ecp/sepoct99/froehlich.htm

# Appendix A: Chance-Related Web Hits (Yahoo)

Log10 of 300 Yahoo hits (9/2005): the (10.0), can (9.6), may (9.5), could (9.2), possible (8.9), might (8.9), "due to" (8.7), certain (8.7), statistics (8.7), chance (8.6), risk (8.6), likely (8.6), random (8.5), life chance (8.2), luck (8.2), impossible (8.2), confidence (8.2), accident (8.2), possibility (8.1), percentage (8.1), love chance (8.1), gambling (8.1), lottery (8.0), odds (8.0), games chance (8.0), sports chance (8.0), statistical (7.9), astrology (7.9), betting (7.9), weather chance (7.9), unlikely (7.9), death chance (7.8), "likely to be" (7.8), variation (7.8), God chance (7.7), risk chance (7.7), ESP (7.7), psychic (7.7), probability (7.6), Tarot (7.6), probable (7.6), uncertainty (7.6), uncertain (7.5), insurance chance (7.5), incidence (7.5), likelihood (7.5), accidentally (7.5), prophecy (7.5), accidental (7.4), UFO (7.4), religion chance (7.4), random chance (7.4), hypothesis (7.4), statistics chance (7.4), forecasting (7.3), certainty (7.3), accident chance (7.3), coincidence (7.3), prevalent (7.3), variance (7.2), paranormal (7.2), disaster chance (7.2), fortune chance (7.2), evolution chance (7.2), research hypothesis (7.2), chaos chance (7.2), bible chance (7.1), gambling chance (7.1), prayer chance (7.1), lottery chance (7.1), betting chance (7.0), risk probability (7.0), due random chance (7.0), "by accident" (7.0), conspiracy chance (7.0), improbable (7.0), ESP tests (6.9), ruin chance (6.9), conceivable (6.9), chance probability (6.9), "intelligent design" (6.9), "by chance" (6.9), "unlikely to be" (6.9), "risk factor" (6.8), bible statistics (6.8), "chaos theory" (6.8), death probability (6.8), "significant difference" (6.8), "standard deviation" (6.8), "statistically significant" (6.8), aliens chance (6.8), prediction chance (6.8), randomness (6.7), coincidence chance (6.7), insurance probability (6.7), inconceivable (6.7), sports probability (6.7), weather probability (6.7), love probability (6.7), extraterrestrial (6.7), impossibility (6.7), "random number" (6.7), horoscope chance (6.7), psychic test (6.7), religion probability (6.6), truth probability (6.6), astrology test (6.6), psychic chance (6.6), evolution probability (6.6), chaos accident (6.6), God probability (6.6), ESP chance (6.5), prophecy chance (6.5), confounding (6.5), "to chance" (6.5), prediction probability (6.5), religion coincidence (6.5), "global warming" chance (6.5), "confidence interval" (6.4), astrology chance (6.4), "standard error" (6.4), UFO chance (6.4), "margin of error" (6.4), "random sample" (6.4), accident probability (6.4), gambling probability (6.3), disaster probability (6.3), betting probability (6.3), synchronicity (6.3), polygraph (6.3), "random number generator" (6.3), "analysis of variance" (6.3), "relative risk' (6.3), "statistical significance" (6.3), "T test" (6.3), evolution coincidence (6.2), "normal distribution" (6.2), paranormal chance (6.2), paranormal test (6.2), bible probability (6.2), forecasting chance (6.2), fortune probability (6.2), "sampling error" (6.2), "random selection" (6.2), "fat chance" (6.2), superstition chance (6.2), chaos probability (6.2), "lie detector" (6.2), "confidence level" (6.2), "fortune telling" (6.2), chaos coincidence (6.1), "intelligent design" chance (6.1), forecasting probability (6.1), pseudoscience (6.1), "random variables" (6.1), "random walk" (6.1), biorhythm (6.0), "statistically improbable" (6.0), "random sampling" (6.0), "p-value" (6.0), conspiracy probability (6.0), "null hypothesis" (6.0), "by coincidence" (6.0), "statistical association" (6.0), "small chance' (6.0), prayer probability (6.0), "if by chance" (6.0), polygraph test (6.0), "student test" (6.0), "intelligent design" test (5.9), ruin probability (5.9), "just a coincidence" (5.9), "statistical inference" (5.9), coincidence probability (5.9), "bible codes" chance (5.9), "lie detector" test (5.9), chaos coincidence chance (5.9), extraterrestrial chance (5.8), "random events" (5.8), "by luck" (5.8), "conditional probability" (5.7), "non-random" (5.7), "significance level" (5.7), aliens probability (5.7), "significant association" (5.7), "statistical tests" (5.7), prophecy probability (5.7), "mere coincidence" (5.6), "random chance" (5.6), "working hypothesis" (5.6), reject null hypothesis (5.6), psychic probability (5.6), "binomial distribution" (5.6), synchronicity chance (5.6), "random assignment" (5.6), "alternative hypothesis" (5.6), "just coincidence" (5.5), "strange coincidence" (5.5), superstition probability (5.5), "statistically insignificant" (5.5), "bible codes" (5.5), "random error" (5.5), "variance analysis" (5.5), "random sequence" (5.5), astrology probability (5.4), "lie detector" chance (5.4), "sampling distribution" (5.4), "due to chance" (5.4), polygraph chance (5.4), "significant result" (5.4), "extrasensory perception" (5.3), "reject the null" (5.3), "significant risk factor" (5.3), F-test (5.3), "fortune telling" chance (5.3), "than chance" (5.3), UFO probability (5.3), pseudoscience chance (5.3), "intelligent design" probability (5.3), extraterrestrial probability (5.3), "likely to be due to" (5.3), paranormal probability (5.2), "weird coincidence" (5.2), "due to accident" (5.2), "happen by chance" (5.1), "research hypothesis" (5.1), "statistical control" (5.1), "chi-squared test" (5.0), "occur by chance" (5.0), "result of chance" (4.9), "two sample test" (4.9), pseudoscience probability (4.9), "chance event" (4.9), "plausible hypothesis" (4.8), "better than chance" (4.8), "unlikely to be due to" (4.8), synchronicity probability (4.8), "due to luck" (4.8), "prediction interval" (4.7), biorhythm chance (4.7), "birthday problem" (4.7), "freak occurrence" (4.7), "statistically impossible" (4.7), "sample statistic" (4.7), "type 1 error" (4.6), "type 2 error" (4.6), "not just a coincidence" (4.5), "lie detector" probability (4.5), "not due to chance" (4.5), "accept the alternative" (4.5), "fail to reject the null" (4.5), "statistically unlikely" (4.5), "doesn't happen by chance" (4.4), "fortune telling" probability (4.3), "freak event" (4.2), "accept the null" (4.2), "due to chance alone" (4.1), "because of chance" (4.0), "due to sampling error" (4.0), "due to random chance" (4.0), "due to coincidence" (3.8), "non-random sample" (3.7), "due to randomness" (3.6), "likely due to chance" (3.6), "reject the alternative" (3.6), "simply due to chance" (3.6), "just due to chance" (3.5), "due to random error" (3.5), "not likely due to chance" (3.4), "test of randomness" (3.4), "to be due to chance" (3.3), "bible codes" probability (3.2), "to happen by chance" (3.2), "by means of chance" (3.2), "due merely to chance" (3.1), "accept the alternate" (3.1), "purely due to chance" (3.0), "unlikely to be due to chance" (3.0), "merely due to chance" (3.0), "due only to chance" (2.9), "accounted for by chance" (2.8), "unlikely to happen by chance" (2.8), "due purely to chance" (2.8), "reject the alternate" (2.8), "unlikely due to chance" (2.7), "due simply to chance" (2.7), "only due to chance" (2.7), "due just to chance" (2.6), "unlikely by chance" (2.1), the (10.0), "if due to chance" (1.3), "due purely to random chance" (1.1), "unlikely if due to chance" (0.8), "due just to luck" (0.3), "due only to randomness" (0.3), "due simply to randomness" (0.0)