

PROBABILITY AND STATISTICS
BRIEFLY ANNOTATED DEFINITIONS AND CONCEPTS



MARLOS VIANA

1. INTRODUCTION

The following are simple annotated definitions, concepts and related topics for reference and discussion. They outline the basic reasoning underscoring the classical approach to statistical inference, summarized later on in Comment 2.38. Included in the objectives of these annotations is the introducing of statistics as a language of scientific inquiry. On the other hand, research questions in general are expressed in a variety of intrinsic languages (e.g., biological, psychometric) which, when confronted with a statistical language and reasoning, are subject to all levels of adaptation, translation and, not uncommonly, truncation. These notes are true to the common language of statistics and probability, and may serve to contrast it against one's own language and reasoning of scientific investigation. Comments indicated with a (*) may be discussed after completing a first reading of the basic concepts.

2. PROBABILITY

2.1. The motif of probability and statistics is *uncertainty*, which is present everywhere, independently of you and me. It is present in the back of one's mind, it is present in the very fabric of nature. One knows when it is present. The objects of uncertainty are propositions, expressed by language, about certain natural conditions of interest. Coins, dice and decks of cards do not *have* uncertainty. Single particles of light do. Heisenberg¹ uncertainty principle;

2.2. *Randomness* is a condition which often leads to uncertainty. However, uncertainty does not require its presence. A spinning top shows a random trajectory and one is uncertain of its final resting position. Whether or not a certain person that I know (and you don't) is alive is not a random condition, and nevertheless, you are uncertain on which condition (proposition) is true. Randomness, mechanics;

2.3. Propositions are often expressed in terms of *measurements*, which result from the common practice of associating categories, rankings or numbers with natural conditions perceived around us. All measurements are representations of certain conditions. History (e.g., psychometrics): The work of Charles E. Spearman on intelligence (1904), of Luis L. Thurstone on attitude (1928), of Clark L. Hull (1943) on learning; Edward Lee Thorndike, Lee J. Cronbach, Patrick Suppes. When a condition is represented by a number, say, then all similar conditions become subject to the algebraic properties of the adopted set of numbers. Time, length, area, volume, velocity, density...anxiety, depression, pain, wellness- in all cases the conditions become (adequately of not) subject to the underlying algebra (sum, multiplication, ordering, commutativity, associativity); measuring is giving numerical names to conditions- apparent sense of domination, understanding,

Date: January 15, 2003.

Lectures notes prepared for the GRCC's Fall 2002 Biostatistics Rotation. Copyright ©2002 by Marlos Viana. Comments, corrections, suggestions to viana@uic.edu. Selected biographic citations were abstracted from <http://www-gap.dcs.st-and.ac.uk/history/BiogIndex.html>.

¹Werner Heisenberg (1901,1976): the theoretical or actual position and velocity of an object cannot both be measured exactly.

synthesis (John Locke² on the role of words and language) ;

2.4. The results of measurements, or quantification mechanisms in general, are often represented as *variables*. *Discrete* variables take on possible values like the natural numbers (countably many 0,1,2,3,4,...), e.g., the number of fish caught at Lake Michigan at any given day (0,1,2,...); *continuous* variables take on all possible values like the real numbers, e.g., the average weight of fish caught at Lake Michigan at any given day (0.51723 Kg, say);

2.5. A *random variable* is a variable connected to a proposition (and subject to uncertainty). Statisticians write X, Y, Z,... to indicate random variables. A random variable X may represent the response to a given clinical intervention or the systolic blood pressure at a given time point. When the uncertainty of particular variables is dependent of time, we may write X(t), Y(t), Z(t) etc, to indicate such dependence (follow-up studies, circadian observations, periodograms, the state of a disease condition at a given time point, etc.);

2.6. The *unit of analysis* is the unit (subjects, specimens, organs, time-points) of comparative interest on a proposition. Comparing the pulse rate among subjects (the subject is the unit), comparing the visual acuity between fellow eyes (the eye is the unit of analysis), comparing electric potential between cell lines (the cell line is the unit);

2.7. A *binary* random variable is one related to propositions with only two possible outcomes, usually represented by X=1 (positive, yes, like, survived) and X=0 (negative, no, dislike, died)- It is a discrete random variable;

2.8. *Probability* is a numerical expression of one's current uncertainty about a given proposition (usually expressed in terms of random variables). It is therefore personal and temporal. There are many ways of eliciting and quantifying uncertainty from one's mind (e.g., betting challenges). Early history (probability): e.g., Abraham De Moivre (Doctrine of Chances, 1718, 1738, 1756), Jacob Bernoulli (Ars Conjectandi, 1713-The Art of Conjecturing), Pierre Simon (Marquis de) Laplace. Contemporary (subjective probability): The works of Amos Tversky, Patrick Suppes, Leonard J. Savage (1917-1971) and others- Probability is the square of something...for Richard Feynman³, it is God's debris for Scott Adams⁴. Nevertheless, concluded Persi Diaconis, it is something for which our brain is not well-wired for.... Contemporary views on subjective probability: John von Neumann and Oskar Morgenstern, Frank P. Ramsey, and Leonard J. Savage among others (utility functions, personal probabilities, decision making);

2.9. *Symmetry*, like uncertainty, is a perceived notion which is all around us- we will not defined it here, e.g., $\triangleleft | \triangleright$. Symmetry often means that certain "labels" are irrelevant to the one's perceived likelihood of a given proposition. Example: your (indifferently left or right) visual acuity is 20/20. Symmetry is often associated with the notion of invariance or *immunity to change*. Related: mirror symmetry, mosaics, wallpaper patterns. History: Herman Weyl (algebra, physics), Joe Rosen⁵ (symmetry principle). It is the building block of developmental human vision;

²English philosopher (1634-1704). In *Essay Concerning Human Understanding*, 1689: Frequently, the idea signified by the word is not clear, and sometimes words are used even when there are no ideas corresponding to them.

³The Character of Physical Law, MIT Press; ISBN: 0262560038; (February 15, 1967)

⁴God's Debris, Andrews McMeel Pub (Cal); ISBN: 0740721909; (September 15, 2001)

⁵Symmetry Discovered: Concepts and Applications in Nature and Science, Dover Pubns; ISBN: 0486294331; (April 1998)

2.10. *Uniform* probabilities reflect particular symmetries in the object of a given proposition. If certain properties of the object are irrelevant and do not alter one's beliefs when moved around, then the proposition should be assigned with equally likely (or uniform) probability. Face six will show up when the die comes to rest. Replace "six" by one, or two, or... five- should your expectations on the proposition remain constant, then the probability assigned to the event must be 1/6. Immunity to change is (Joe Rosen's view of) symmetry. This leads to uniform probabilities. Probabilities given by synthetic formulas are often limited to propositions about symmetric objects: propositions about coins, dice⁶, deck of cards e.g., *number of favorable events divided by total number of events* and similar definitions of probability. Other names related to symmetry: exchangeable, permutable conditions. History: Alfréd Haar, Bruno De Finetti. Dice and historical decisions, miracles⁷;

2.11. A *probability model* for a discrete random variable X is a table describing, in one column, the possible outcomes of X, say, x_1, x_2, \dots, x_k , and, in the other column, the corresponding probabilities

$$P(x_1), P(x_2), \dots, P(x_k).$$

All probabilities are real numbers bounded in the interval [0,1] (this includes the extremes 0 and 1) and the sum of these probabilities is one. Symbolically, we write, $0 \leq P(x_i) \leq 1$, and $\sum_i P(x_i) = 1$;

2.12. *Risk*⁸

...is simply danger mathematized, often unfittingly, and the mathematized expectation of gain, falsified by replacing the individual with the average. Mathematical expectation does not capture the amusement or recreational value of gambling, the adrenalin of the win, the benefits of living with hope and dreams, the fact that despite the odds, individuals have won fortunes and have their lives changed for better and for worse. Indeed, the relief from the dull safety of averaging can be viewed as benefits of gambling.;

2.13. *Minimal risk*⁹

means that the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests.;

2.14. Weighted averages. The Oxford English Diccionary¹⁰ has many interesting usages of *weighting*, including:

- Statistics. To multiply the components of (an average) by compensating factors; to treat (the components of any numerical quantity) similarly.
- 1901 A. L. BOWLEY Elem. Statist. 111 The very important statistical method known as weighting the average. Ibid., Should we weight the numbers given by the total numbers of inhabitants of the contributing counties, or by their distance from London, or by some quantity derived from these? 1927 C. SPEARMAN Abilities of Man App. p. xviii, We urgently require to know how the single tests should be relatively weighted in their combination. 1971 I. G. GASS et al. Understanding Earth v. 82 The individual

⁶Take a short break and enjoy rolling virtual bones at <http://www.irony.com/igroll.html>

⁷Many fascinating parables on uncertainty, probability and statistics are accounted in *The Broken Dice, and Other Mathematical Tales of Chance* by Ivar Ekeland, Carol Volk (Translator) University of Chicago Press (Trd); ISBN: 0226199924; (May 1996)

⁸Legalized Gambling as a Strategy for Economic Development, by R. Goodman. The quote is from the book's review by Philip J. Davis, SIAM News, October, 1994

⁹Code of Federal Regulations- Title 45 Department of Health and Human Services, Part 46- Protection of Human Subjects (45 CFR 46)-

<http://www.med.umich.edu/irbmed/FederalDocuments/hhs/HHS45CFR46.html>

¹⁰<http://dictionary.oed.com/>

data were weighted according to quality, so that a poorly determined result makes a smaller contribution to the mean than a precisely determined value. 1976 Daily Record (Glasgow) 30 Nov., Replies were weighted by age and General Election voting to make sure they were representative of all Record readers. 1977 Whitaker's Almanack 1978 1219 In working out the [cost-of-living] index figure, the price changes are weighted that is, given different degrees of importance in accordance with the pattern of consumption of the average family.

2.15. The *mathematical expectation* or expected value of a discrete random variable X is the weighted average of its possible values with weights the corresponding probabilities; We write, symbolically, $E(X) = \sum_i x_i P(x_i)$ to indicate the weighted average. This is an example of a fundamental type of calculation with probability models;

2.16. The expected value of X is also known as the *model mean* for the probability model of X . This particular expected value is often represented by μ_x (the Greek lower-case m, mu), that is;

$$\mu_x = \sum_i x_i P(x_i).$$

Typically, the model mean is dominated by those outcomes X with highest probability. Center of mass, model location;

2.17. The *model variance* is also an expected value. Now we look at possible values of the squared distance $(x - \mu_x)^2$ between X and its mean μ_x , and determine the weighted average, with weights $P(x)$ as above. That is, the model variance of X is the expected value of $(X - \mu_x)^2$. Symbolically, we indicate it by σ_x^2 , so that

$$\sigma_x^2 = \sum_i (x_i - \mu_x)^2 P(x_i).$$

The model variance is dominated by those square distances $(X - \mu)^2$ that occur with highest probability. The more likely these distances (caused by observations dispersed away from the model mean), the higher the model variance. Variability, dispersion, spread; The *model standard deviation* of X is σ_x , the (positive) square root of the model variance;

2.18. Both the model mean and the model variance are expressions reflecting certain characteristics of the underlying probability model for X . The model mean reflects the model *center* or location, whereas the model variance reflects the overall distribution of probability *spread* among the possible values of X . Relative center, relative spread;

2.19. A *joint* probability model for a pair (X, Y) of discrete random variables is a table, such as in Table 2.1, describing the possible *joint* events and the corresponding probabilities (p_1, \dots, p_4) . The joint model

TABLE 2.1. Joint Probability Model.

x	y	P(x,y)
1	1	p ₁
0	1	p ₂
1	0	p ₃
0	0	p ₄

represents one's uncertainty about both outcomes, jointly;

2.20. The *model covariance* for (X,Y) is also a expected value. Now we look at possible values of $(x - \mu_x)(y - \mu_y)$ and determine the weighted average, with weights $P(x, y)$, just as with the model mean and the model variance. Symbolically, we indicate the model covariance by σ_{xy} , and write,

$$\sigma_{xy} = \sum_i (x_i - \mu_x)(y_i - \mu_y)P(x_i, y_i);$$

2.21. The *model correlation* for (X,Y) is the model covariance σ_{xy} scaled by the product of the corresponding standard deviations. We indicate the model correlation by ρ_{xy} , and write¹¹, following the definition,

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y};$$

2.22. Both the model covariance and the model correlation reflect the amount of statistical *association* between the two random outcomes X, Y. This is evident by considering the relative positions of observations x_i and y_i relative to the corresponding means, so that each component $(x_i - \mu_x)(y_i - \mu_y)$ will be positive when both outcomes are at the same side of their corresponding means and will be negative otherwise. Therefore, relatively large (positive or negative) values of the model covariance will reflect a higher (positive or negative) relative association between the two random variables. The model correlation picks up on the same interpretation, with the added fact that its range is now bounded by the interval $(-1, 1)$. Values toward either extremes reflect tighter associations relative to correlation values around zero;

2.23. The *Bernoulli* probability model is useful for representing one's uncertainty about a 0-1 (discrete, binary) random variable. The Bernoulli model is described in Table 2.2. Under the Bernoulli model, the

TABLE 2.2. Bernoulli Probability Model.

x	P(x)
1	p
0	1-p

calculus with expected values leads to the model mean $\mu_x = p$, and to the model variance $\sigma_x^2 = p(1 - p)$. Interpretations of center and spread reflect different values of p: Largest variance arises when $p=0.5$ (most uncertain), the least when $p=1$, or with $p=0$ (most certain);

2.24. We say that the Bernoulli model has one single *parameter*, namely, p. Both the model mean and the model variance depend on the same parameter p. In general, one can perform similar calculations with all probability models, obtain the model mean, the model variance, and derive interpretations. However, you cannot decide (in the Bernoulli model, for example) which outcome, $X=0$ or $X=1$, is the most likely one for as long as the actual value of the parameter p remains unknown to you. You cannot decide. Synthetic, analytic forms of a model;

2.25. The *binomial* model is another example of a discrete probability model. It represents one's uncertainty on the total number

$$X = X_1 + \dots + X_n$$

of *positive* results when observing n *independent* and *identically distributed* Bernoulli random variables with (common) parameter p. This refers to the assumption of identically distributed random variables. The other assumption, that these random variables are statistically independent, is a rather subtle notion (we will refer to it later on in Comment 2.32). For the moment, think of statistical independence as a condition under

¹¹ ρ indicates the Greek letter rho, for lower-case Roman r.

which propositions verified from previous outcomes (of X_1 , say) do not affect one's *betting expectations* on future propositions (X_2). The point here is introducing the binomial model as an example of a simple model in which its construction clearly depends on (two) assumptions, and such that when these assumptions are ignored the model may be improperly applied. Models require judgments, often subjective, on their proper adequacy and usage. This is complementary to simple deductive reasoning (e.g., simple calculus following known formulas);

2.26. As a simple example of *calculations with probability models*, you may want to verify that the binomial probability model for $X = X_1 + X_2$ ($n = 2$) can be summarized as in Table 2.3. Again, x represents the

TABLE 2.3. Binomial Probability Model, $n=2$.

x	$P(x)$
0	$(1 - p)^2$
1	$2p(1-p)$
2	p^2

number of positive counts, in this case, $x=0,1,2$, and the parameter p stands for the probability of a positive observation in any one of the two (statistically independent) trials X_1, X_2 . You may calculate, similarly to the Bernoulli model, the model mean and the model variance for the binomial model, to find $\mu_x = 2p$ and $\sigma_x^2 = 2p(1 - p)$;

2.27. The same reasoning and assumptions applied to $n = 2$ Bernoulli variables leads to the binomial model with an arbitrary (however fixed) number, (n), of independent and identical Bernoulli random variables. The resulting binomial model is given¹² by

$$P(x | p) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, 2, \dots, n;$$

2.28. We may distinguish between *analytic* forms of the model, as when the model parameter is known, and *synthetic* forms of the model, as when the parameter of the (Bernoulli) model is unknown. Synthetic forms are useful, nevertheless, to obtain and describe properties of all models within the same family (say, Bernoulli) of models. Simple analogy: analytic geometry- drawing a specific line, studying properties of families of line e.g., slope, intercept;

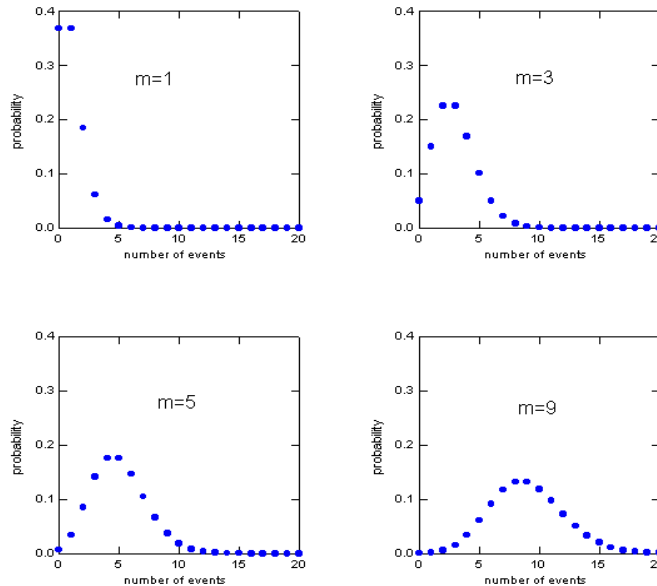
2.29. The binomial probability model may be extended to count any possible number $0, 1, 2, \dots$ of events resulting from adding a very large number of independent Bernoulli events that occur with very small probability. Think of counting the random number of radioactive emissions during a time interval of t units of time (or linear space) assuming that the probability (p) of an emission in any smaller segment of length d units is proportional to the length of these segments. If these events occur at a rate of $f = p/d$, then, approximately, we have a binomial probability model, with mean ft and number of trials equal to the integer number closest to t/d , describing the number $(0, 1, 2, 3, \dots)$ of emission events during the period of t units of time. For (infinitely) small values of d , we obtain the probability model describing the exact number of radioactive counts in the interval of duration t . Using the convention $m = ft$, the resulting probabilities for the number x of emissions are given by

$$P(x | m) = \frac{m^x}{x!} e^{-m}, \quad x = 0, 1, 2, 3, \dots,$$

¹²The binomial coefficients $\binom{n}{x}$ are given by $\frac{n!}{x!(n-x)!}$.

This is the *Poisson*¹³ probability model. Its mean and variance are both equal to m . Figure 2.1 illustrates the Poisson probabilities for selected values of m ;

FIGURE 2.1. Poisson probabilities for $m = 1, 3, 5$ and 9 .



2.30. The *Gaussian*¹⁴ curve, also called *normal curve*, is an example of a *density function*. Functions like this one are utilized to calculate probabilities when the underlying random variable is represented on a continuous domain, like that of the real numbers. The graph, $y = f(x)$, of these densities is obtained by the expression

$$f(x | \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < +\infty,$$

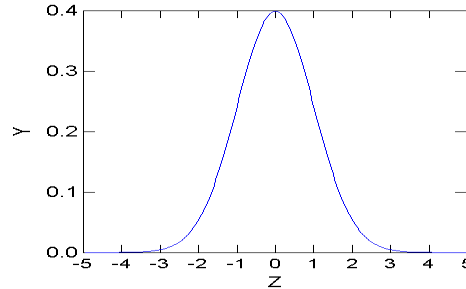
where μ is the model mean and σ^2 the model variance. Figure 2.2 illustrates the graph of the standard normal curve ($\mu = 0, \sigma = 1$). Note that this is a two-parameter probability model. In analogy with the mathematical expectations derived for discrete models, as in Comments 2.16 and 2.17, here, integral calculus leads us to define and verify that the model mean and model variance for the normal probability model are, respectively,

$$\mu = \int xf(x)dx, \quad \sigma^2 = \int (x - \mu)^2f(x)dx;$$

¹³French mathematician Siméon-Denis Poisson, 1781-1840.

¹⁴Johann Carl Friedrich Gauss, German mathematician, 1777-1855

FIGURE 2.2. The (standard) normal density function- The Gauss curve.



2.31. The probability of a random variable X exceeding a given value, say c , is symbolically expressed as $P(X \geq c)$. When the domain of X is continuous, its value is the area $\int_c^\infty f(x)dx$ under the density function bounded by the graph of the density, the vertical line at $x=c$, and the horizontal segment $x \geq c$ in the x -axis. Classically, probabilities of that nature may be used to characterize the *statistical magnitude* of the value c , under the assumption (or hypothesis) that the probability model of X is given by the present density. It can be argued that c is *statistically large*, given the present probability density, when the (one-tail) probability $P(X \geq c)$ is small. We will return to this notion later on in Comment 3.15;

2.32. *. *Statistical independence* is often understood to mean something like *mechanical* independence. Clearly, when sequentially sampling two marbles from a fixed-composition urn, we are safe to say that the marble selected in the first trial has no physical effect on the marble to be selected in the subsequent trial. They are, should we say, *mechanically* independent. However, the outcome of these two trials are statistically dependent unless we have *sharp* knowledge about the composition of the urn. On the other hand, we can find statistical independence where there is mechanically perfect association. The classical example is Kepler's¹⁵ laws of planetary motions, describing the motions of the planets in the solar system. As originally formulated, the laws do not take into account the random, seemingly independent, gravitational interactions of the various planets on each other;

2.33. *. *Marginal* and *conditional* probabilities can be derived from joint probability models, as described above in Comment 2.19. Here is a similar table, with a particular interpretation. Think of D as disease

TABLE 2.4. Joint Probability Model.

d	t	P(d,t)
1	1	p_1
0	1	p_2
1	0	p_3
0	0	p_4

condition (0= condition absent, 1= condition present) and of T as a screening test result (0= suggestive of

¹⁵German astronomer Johannes Kepler, whose analysis of the observations of the 16th-century Danish astronomer Tycho Brahe enabled him to announce his first two laws in the year 1609 and a third law in 1618, the law of planetary motions.

condition present, 1= suggestive of condition absent). The marginal probability model, P_D , for the random outcome D is obtained as $P_D(d) = P(d, 0) + P(d, 1)$, so that

$$P_D(1) = P(1, 0) + P(1, 1) = p_3 + p_1, \quad P_D(0) = P(0, 0) + P(0, 1) = p_4 + p_2.$$

The marginal model for the other random outcome, (T), is obtained similarly. Marginal probability models represent one's uncertainty about propositions related to each outcome alone. The conditional probability model refers to the uncertainty about one of the events when (given that) the other event is now known to you. For example, we indicate the conditional probability of $T = t$ given $D = d$ by $P(t | d)$ and calculate it as the ratio

$$P(t | d) = \frac{P(t, d)}{P_D(d)}$$

of joint and marginal probabilities. In particular, the test *sensitivity* is the conditional probability

$$P(1 | 1) = \frac{P(1, 1)}{P_D(1)} = \frac{p_1}{p_3 + p_1}.$$

All usual descriptors of clinical and lab-related tests, such as test sensitivity, test specificity and predictive values (positive and negative) are conditional probabilities. Disease prevalence and apparent disease prevalence are marginal probabilities;

2.34. *. Reversing Probabilities: Bayes¹⁶ Formula. The connection between $P(T | D)$ and $P(D | T)$ can be understood as follows: From Comment 2.33, we obtain,

$$P(d | t) = \frac{P(d, t)}{P_T(t)} = \frac{P(t | d)P_D(d)}{P_T(t)}.$$

In addition, because

$$P_T(t) = P(0, t) + P(1, t) = P(t | 0)P_D(0) + P(t | 1)P_D(1),$$

it follows, making the substitution, that

$$(2.1) \quad P(d | t) = \frac{P(t | d)P_D(d)}{P(t | 0)P_D(0) + P(t | 1)P_D(1)},$$

which is known as Bayes Formula. In the language of clinical tests, it connects the test sensitivity or specificity, $P(T | D)$, with the test predictive values, $P(D | T)$, in such way that one can related sensitivities and predictive values by properly adjoining the marginal information. The consequences of this rather simple argument, however, have proved to be far-reaching and lead to important interpretations in the core of the Bayesian, subjectivist, school¹⁷ of statistical inference and probability¹⁸.

2.35. *. Bayes formula is often expressed as the equality between the ratios

$$\boxed{\frac{P(1 | t)}{P(0 | t)} = \frac{P(t | 1)}{P(t | 0)} \times \frac{P_D(1)}{P_D(0)}};$$

in which the LHS ratio is interpreted as the *posterior odds* on the presence of the disease condition given the test result T, and the RHS is the product of the *Bayes factor* (or *likelihood ratio*), and the *prior odds* on the disease condition. Learning, as an interpretation of Bayes formula, is following the path

$$\frac{P_D(1)}{P_D(0)} \rightarrow \text{observe data } T = t_1 \rightarrow \text{updated prior odds: } \frac{P(1 | t_1)}{P(0 | t_1)} \rightarrow \text{observe data } T = t_2 \dots$$

¹⁶Reverend Thomas Bayes, 1702-1761, England. His theory of probability appears in *Essay towards solving a problem in the doctrine of chances* published in the Philosophical Transactions of the Royal Society of London in 1764.

¹⁷An interesting summary of many more Schools of Thought are outlined at <http://cepa.newschool.edu/het/thought.htm>.

¹⁸Among their founders and contemporary names: Harold Jeffreys (1891-1989), Frank P. Ramsey (1903-1930), Bruno de Finetti (1906-1985), Dennis V. Lindley, L. J. Savage (1917-1971).

many times over;

2.36. *. *Markov Models*¹⁹ describe the uncertainty associated with the states which a (e.g., biological, mechanical, molecular) system may occupy progressively with time. Consider an enzyme level which may vary into one of three states (1=below minimum level, 2=within normal range, 3=above maximum level). Write $X_1, X_2, X_3, X_4, \dots$ to indicate these random states at sequentially observed time-points. Here is a possible sequence of enzyme levels:

3, 3, 3, 3, 2, 3, 2, 2, 2, 3, 2, 2, 2, 2, 1, 1, 1, 2

The *Markov Property* states that these events vary in such a way that the state of the system at the next time-point (X_{t+1}) is defined, in probability terms, by knowing the state of the system in the previous time-point (X_t). All past history beyond the immediate past is irrelevant for predicting the future state of the system. This is summarized by the conditional probability statement

$$P(X_{t+1} | X_t, X_{t-1}, \dots, X_1) = P(X_{t+1} | X_t);$$

2.37. *. When the same Markov property applies for all time-points the system is said to be *homogeneous*, in which case it is characterized by the *transition probabilities* among the states of interest. In the example above (Comment 2.36), we need to specify the transition probabilities among the states $\mathcal{A} = \{1, 2, 3\}$: this leads to the *transition probability matrix*

$$A = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix},$$

which describes all transition probabilities $p_{ij} = P(X_{t+1} = i | X_t = j)$. Note that all entries in the transition probability matrix are probabilities and that the sum of the entries in each column is equal to 1. If P_0 is the probability model for X at baseline ($t = 0$), then the model P_1 describing the system one time step into the future is calculated by the product $P_1 = AP_0$, whereas model P_k describing the system k time steps into the future is, by iteration, derived as

$$P_k = A^k P_0,$$

and eventually reaches a *stationary distribution*;

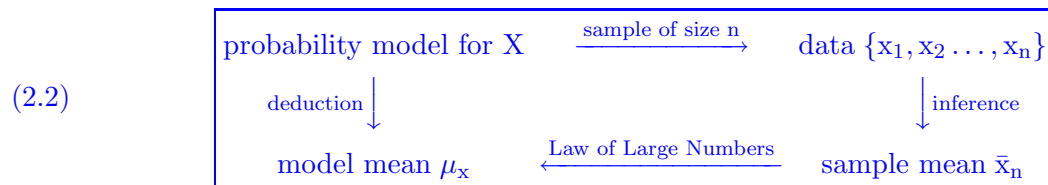
2.38. *Summary.* We have discussed probability as a measure of one's uncertainty about certain well-defined²⁰ events or propositions. Coins or dice do not *have* probabilities. They stand quiet and there remain so, according to Newton's laws. The object of a probabilistic statement is a given proposition- which is true or is false, always. However, at a given time, one is uncertain about whether or not a given proposition obtains. The uncertainty on the color of the marble is just the same whether the marble is yet to be sampled from the urn or has already been drawn and is resting outside of our view - it does not matter, to resolve the uncertainty about the face showing up, whether a coin is at that very moment spinning around the table or at rest inside a book. The flipping of the coin or the shaking of the urn are just transient events which do not *add* uncertainty to a proposition about the face showing up when the coin is at rest or to a proposition about the color of the marble already out of the urn.

We have indicated that expected values can be mathematically derived (deductive calculus) from a given probability model (such as the Bernoulli, the binomial, the Poisson or the normal models). Expected values trace and describe properties of these models, and expected values of different relationships among the random variables lead to different parametric interpretations (e.g., center, variability, correlation). The study of these parametric properties obtained from deductive calculations is at the core of probability theory. These parametric entities (e.g., mean, variance, correlation) become the object of interest: these are the entities

¹⁹Russian mathematician, Andrey Andreyevich Markov, 1856, 1922

²⁰Fuzzy-logic arguments to inference are based on the view that this is a heavy-handed assumption.

which statisticians would make statements about - We may know that the probability model is Bernoulli and yet be uncertain as to which parameter (p) determines it. With complete knowledge (say: model is binomial, n is 10 and p is 0.7) at hand the task is complete: all propositions have a probabilistic answer. Analogy: we know that the curve describing the moving particle is a line, and yet we do not know its intercept nor its slope. We may only know that the density function of a probability model is symmetric around zero, and yet be uncertain to which model this is. Analogy: we know that the moving particle is described by a polynomial curve and yet we do not know its degree. At that point, when models are (partially or fully) unknown, we will have to change our language (deduction vs. inference) and obtain adequate ways of reasoning about these unknown objects. The classical view of parametric statistical inference is illustrated in the diagram below. We have discussed the two blocks of the deductive side (probability) of the diagram. In the sequence we will comment on the basic components of its inferential (statistics) side.



3. STATISTICAL INFERENCE

3.1. Under the classical paradigm for inference, when a probability model is unknown, inferential arguments may lead to *estimates* and other inferences about its parameters (expected values). Statistical inference starts with data and ends with reasonable estimates, such as the sample mean \bar{x}_n , of one or more parameters descriptives of the model for X ;

3.2. Fully-specified probability models can be programmed into a computer so that the machine *simulates* a (nearly) random sample of a determined size (say, n) from the given model. Here is an example of a random sample of size 20 from a Bernoulli model²¹ with $p = 0.5$:

0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1;

3.3. A *statistic* is a *meaningful* summary of data, often with the objective of learning about one or more model characteristics, such as the model mean or the model variance. Those are the primary and simplest objects of *statistical inference* statements. When the model and its parameters are known²² then all questions have (probabilistic) answers. You do not have to sample, to experiment, to simulate. Only when model parameters are unknown, one then has to extend the language of calculus of probability with the wording and reasoning of statistical induction or inference- this is the right-hand side of Diagram (2.2). History: A. M. Legendre²³, F. Galton²⁴, R. Fisher²⁵ and others;

²¹Here is another: 2, 2, 2, 1, 1, 2, 1, 2, 1, 1, 2, 2, 1, 1, 1, 2, 1, 1, 2, 1 obtained from the virtual dice roller mentioned earlier on in Comment 2.10- note its usefulness in generating random assignments of subjects to treatments

²²Classical examples are found in genetics where the Hardy-Weinberg equilibrium condition is perhaps the best example: If we mate two individuals that are heterozygous (e.g., Aa) for a trait, we should obtain 25% of their offspring as homozygous (AA), 50% heterozygous (Aa) and 25% homozygous (aa).

²³French mathematician Adrien-Marie Legendre, 1752-1833.

²⁴British anthropologist Francis Galton, 1822-1911.

²⁵Sir Ronald Aylmer Fisher, statistician and geneticist, 1890-1962.

3.4. The *Law of Large Numbers*²⁶ is a criteria of meaningfulness of a given summary or statistic. For example, it says that (for most random variables) the *sample mean*

$$\bar{X} = \frac{1}{n} \sum_i X_i$$

approaches (in a mathematical limit sense) the model mean μ_x , as the sample size, n , increases to infinity. In that sense we say that the sample mean is a meaningful (consistent) *estimate* of the model mean;

3.5. Similarly, the sample variance for X ,

$$S^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2,$$

is a meaningful estimate of the model variance σ^2 . In fact, as the sample size increases to infinity, S^2 converges to the underlying model's variance, σ^2 ;

3.6. Similarly, the sample correlation for the joint outcomes (X, Y) ,

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})/n}{S_x S_y},$$

is an estimate of the model correlation, ρ , and, as the sample size increases to infinity, r converges to the model correlation (see Comment 2.21). In the above expression, S_x indicates the sample standard deviation for X , and S_y the corresponding standard deviation for Y ;

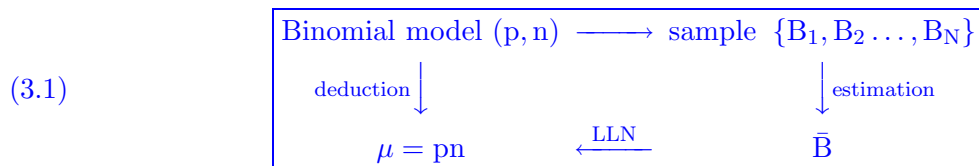
3.7. *Simulation and sampling.* Computer programs can simulate the tossing of a (fair or biased) coin or rolling of a die. To generate a sample of size $N = 5$ from a Binomial model with probability of success $p = 1/6$ and number of trials $n = 5$ we roll our (fair) die $n = 5$ times, count and record the number B of trials in which a six-dot face shows up. We then repeat this experiment four additional times (because $N=5$) and take note of all outcomes of B . Table 3.1 summarizes the results from this experiment, in which a computer program worked out the remaining experiments, sampling $N = 10, 50, 100, 500, 1000$ Binomial events. Because $p = 1/6$ and $n = 5$ are fixed, the expected value of B under these conditions is

$$\mu = 5 \times \frac{1}{6} = 0.83, \quad \text{when } p = 1/6, \quad n = 5.$$

Starting with sample data $\{B_1, \dots, B_N\}$ from B and evaluating the *sample mean*, \bar{B} ,

$$\bar{B} = \frac{B_1 + \dots + B_N}{N},$$

we can observe the Law of Large Numbers (LLN) in action (Comment 3.4): the sample mean *statistically approaches* the model mean μ , as the sample size N increases. In this context, Diagram (2.2) can be illustrated as



²⁶First established by Jacob Bernoulli (1654-1705), published in 1713.

The same argument would apply to the model's variance, which in this case, is $\sigma^2 = np(1 - p) = 25/36 = 0.694$. Table 3.2 summarizes the simulated results for the Binomial model with $n = 5$ and $p = 1/2$. Note that then,

$$\mu = 5 \times \frac{1}{2} = 2.5, \quad \text{when } p = 1/2, \quad n = 5,$$

and the sample mean \bar{B} is expected to statistically converge to the model mean $\mu = 2.5$. Also note that the model mean when $p = \frac{1}{6}$ ($\mu = 0.83$) is shifted to the left relative to the model mean with $p = \frac{1}{2}$ ($\mu = 2.5$). This reflects the interpretation given to the model mean, namely that of a location parameter. Moreover, we are expected to observe a *statistical* shift between the two corresponding samples. This is apparent from the data and corresponding sample averages shown in the two tables. The variance under $p = \frac{1}{2}$ is $\sigma^2 = np(1 - p) = 5/4 = 1.25$. This is larger than the model variance for the previous model. As a consequence, we should expect to observe more variability or spread, relative to the model mean, in samples from a fair coin compared with samples from a fair die. In the two tables, this difference in variability is reflected in the way the data are distributed among the possible values of B.

TABLE 3.1. Frequency counts of simulated N outcomes from a Binomial model with $n = 5, p = 1/6$

B=0	1	2	3	4	5	N	average
1	2	2	0	0	0	5	1.20
5	3	2	0	0	0	10	0.70
20	17	10	1	2	0	50	0.96
38	40	16	6	0	0	100	0.90
197	208	83	11	1	0	500	0.82
393	409	167	27	3	1	1000	0.84

TABLE 3.2. Frequency counts of simulated N outcomes from a binomial model with $n = 5, p = 1/2$

B=0	1	2	3	4	5	N	average
0	0	2	2	1	0	5	2.80
0	2	4	2	2	0	10	2.40
2	2	16	19	8	3	50	2.76
1	18	31	37	11	2	100	2.45
16	74	159	153	82	16	500	2.52
32	151	336	282	164	35	1000	2.50

3.8. The importance of the normal probability model (introduced earlier on in Comment 2.30) comes from the fact that, for most random variables, the probability model which best describes the average of statistically independent random copies of those variables is a normal probability model. This is the essence of the *Central Limit Theorem*, which, similarly to the Law of the Large Numbers is a convergence result from probability theory. In this case, the statistical distance between the frequency distribution resulting from averages of samples of size n , and a normal curve, decreases

as the sample size, n , increases²⁷. In 1920, G. Pólya gave this theorem the name “the central limit theorem of probability theory.” This name continues to be used today²⁸;

3.9. Classical estimates such as the sample mean or the sample variance are called *point estimates*. The object of estimation is often a model parameter. This is what is estimated. They represent a single value which estimates the corresponding unknown parametric characteristic of the model (μ, σ^2, ρ , etc);

3.10. *. Given a sample x_1, x_2, \dots, x_n from a probability model, $P(x | \mu)$, where μ represents a parameter of interest (e.g., the model mean), the *likelihood function* associates to each value of μ the numerical product

$$\mathcal{L}(\mu) = P(x_1 | \mu)P(x_2 | \mu) \cdots P(x_n | \mu).$$

The value $\hat{\mu}$ of μ which maximizes the likelihood function is an estimate of μ . This principle of estimation is called *maximum likelihood estimation*;

3.11. *. In certain cases the theory is developed to estimate the whole density- this is called *density estimation*;

3.12. A *confidence interval* is an estimate which includes the point estimate and, in addition, reflects the amount of variability to which the estimate is subject to. The confidence interval is an interval, centered at the point estimate with range proportional to the standard deviation of the point estimate. For the model mean, these intervals take the generic form

$$\left(\bar{X} - t_\alpha \frac{S}{\sqrt{n}}, \bar{X} + t_\alpha \frac{S}{\sqrt{n}}\right),$$

where S is the sample standard deviation (see Comment 3.5), n is the sample size, and t_α is the $(1 - \alpha/2) \times 100$ -th percentile point obtained from the probability model for the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Under certain assumptions, this probability model follows the Student’s t distribution²⁹. The value of t_α modulates the amount of certainty, $(1 - \alpha)$, with which one wishes to estimate the unknown value of the model mean, such as 90%, 95%, 99%;

3.13. An *interpretation* for a confidence interval estimate is that, with probability $(1 - \alpha)$, one should expect *the statement*: the confidence interval $\bar{x} \pm t_\alpha \frac{s}{\sqrt{n}}$ includes the true value of μ_x , to be a correct one. Again, similarly to point estimates, the object of an interval estimate, or confidence interval, is a model parameter- this is what classical confidence intervals estimate (more on Comment A.1 in the Appendix);

²⁷More precisely, the sum of a large number n of independent identically distributed random variables with finite means and variances, normalized to have mean zero and variance one, is approximately normally distributed.

²⁸George Pólya, 1887-1985. A more apt description would be “the normal convergence theorem”. The (Bernoulli form of the) central limit theorem was first proved by De Moivre in 1733

²⁹First published in 1908 by William Sealey Gossett (1876-1937) under the pen name - imposed by his publisher - of “Student” .

3.14. The object of a *statistical hypothesis testing* includes the evaluation of a *null hypothesis* expressed in terms of one or more model parameters (Again, model parameters are the primary objects of classical inference statements). These hypotheses have the generic form $H : \mu = \mu_*$. Confidence intervals may be utilized to assess the hypothesis H by arguing that H should be rejected whenever the observed CI does not include the value, μ_* , proposed by the *null* hypothesis. Otherwise, the hypothesis is accepted³⁰. The Type I *error rate* associate with the decision rule for the proposed hypothesis is the value of α indicated above;

3.15. In addition, the hypothesis $H : \mu = \mu_*$ may be evaluated with reference to a *p-value*, such as

$$p = P(T \geq t \mid \mu_*),$$

which is the probability that the test statistic, T , (in itself a random variable) exceeds the particular observed value, t , of this statistic³¹. Here, as indicated by the above expression, the evaluation of the probability is under the assumption that μ_* is the correct model mean. A *small* p-value is interpreted as an indication that the observed value of the test statistic is statistically *large* enough to contest the proposed value of μ , and therefore, the classical argument goes, such proposed value should be rejected- that is, the null hypothesis should be rejected. This argument is at the center of the Neyman and Pearson³² school of parametric inference;

3.16. *. The notion of p-value is based on events that *could have happened*- but in fact did not. This is what the probability expressed by $P(T \geq t \mid \mu_*)$ means. It clearly depends on the probability mass lying to the right of the observed value t . Here is an adaptation of the classical example illustrating one of the conceptual difficulties intrinsic to the argument:³³ A binary data were left to a statistician, with a note saying: test the hypothesis that these data come from a Bernoulli model with $p = 0.5$ against the alternative hypothesis that $p > 0.5$. Here are the data:

0 1 1 0 1 1 1 1 1 1 0.

The statistician understood it as a binomial experiment, calculated and found the p-value to be

$$P(X \geq 9 \mid p = 0.5) = 0.03;$$

The other investigator who had conducted the experiment returns and adds that, in fact, the experiment was conducted with the purpose of stopping it after three negative results had been observed. Now, information that was only in the mind of one of the investigators becomes relevant, and the statistician has to recalculate the p-value according to a *negative binomial* model. With reference to this new set of possible outcomes, the new p-value

$$P(X \geq 9 \mid p = 0.5, \text{investigator's revelation}) = 0.07,$$

is qualitatively different from the first (if a hard line is drawn at the 0.05 criteria), and shows the subjectiveness of the classical argument, as well as its dependence on events that *could have*

³⁰Edwards, Lindman & Savage (1963, p.492) point to the fact that “If the null hypothesis is classically rejected, the alternative hypothesis is willingly embraced, but if the null hypothesis is not rejected, it remains in a kind of suspended disbelief.”

³¹It is common practice to indicate the p-value by the letter p- it has no connection with the same notation used elsewhere here, for example, to indicate the Bernoulli parameter (p)

³²Jerzy Neyman (1984-1981), Egon S. Pearson (1895-1980), son of Karl Pearson (1857-1936).

³³As discussed in Lindley & Phillips (1976).

happened and yet are not part of the observed data. Likelihood Principle;

3.17. Alternatively, null hypotheses may be evaluated by comparing the observed test statistic (e.g., the T test outlined in Comment 3.12) against a critical value c_α , obtained by solving an equation (in c_α , given μ_* and α) of the form

$$\alpha = P(T \geq c_\alpha \mid \mu_*);$$

3.18. The *statistical power* of a test statistic, indicated here by $\pi(\mu)$, is the probability with which the test statistic exceeds the critical value c_α (Comment 3.17), when the probability is evaluated under any given alternative hypothesis μ . Symbolically, we write

$$\pi(\mu) = P(T \geq c_\alpha \mid \mu).$$

Therefore, the power of a test is always a function of the parametric alternatives dictated by the probability model under consideration;

3.19. *Sample size and statistical power analysis.* In the above equation (Comment 3.18), the test statistics depends on the sample size (e.g., Comment 3.12). Therefore, one can set the value of α (the type I error), set the value of $(1 - \pi)$ (the type II error), set an *alternative of interest*³⁴, and then find out which value of the sample size, n , is a solution³⁵ of the resulting equation in the unknown n .

3.20. *. Important in the execution of any quantification plan is its *experimental design*. In statistics, experimental designs are obtained with the objective of making the most effective utilization of the experimental resources, and at the same time, providing a realistic connection between research question and the experimental data obtained (History: R. A. Fisher, F. Yates, G. W. Snedecor, H. Scheffé);

3.21. *. Also related to the planning of any study is the notion of *feasibility* of the proposed experimental trial, in which case a *pilot study* should be considered. In these studies, a small number of subjects is evaluated before proceeding with a full study. The pilot study can be used to³⁶ validate analytical methodology, assess variability, optimize sample collection time intervals, and provide other information;

3.22. *. *Bioequivalence*³⁷ means the absence of a significant difference in the rate and extent to which the active ingredient or active moiety in pharmaceutical equivalents or pharmaceutical alternatives becomes available at the site of drug action when administered at the same molar dose under similar conditions in an appropriately designed study. Where there is an intentional difference in rate (e.g., in certain controlled release dosage forms), certain pharmaceutical equivalents or alternatives may be considered bioequivalent if there is no significant difference in the extent to

³⁴This is a central issue in the calculation of the sample size. The alternative of interest is determined by the context within which the methodology is applied, and is not a statistical commodity. In particular, power analysis would be justified only when a clinically meaningful interpretation of the underlying measurement units did exist.

³⁵An interactive sample size calculator can be viewed online at UCLA's site <http://calculators.stat.ucla.edu/powercalc/>

³⁶e.g., <http://www.fda.gov/cder/guidance/>

³⁷Abstracted from Code of Federal Regulations 21CFR320.23 found at <http://www.access.gpo.gov/nara/cfr/> -note the carefully defined terms language.

which the active ingredient or moiety from each product becomes available at the site of drug action. This applies only if the difference in the rate at which the active ingredient or moiety becomes available at the site of drug action is intentional and is reflected in the proposed labeling, is not essential to the attainment of effective body drug concentrations on chronic use, and is considered medically insignificant for the drug.

APPENDIX A. APPENDED ANNOTATIONS

A.1. On Comment 3.13. Here is an experiment which illustrates an interpretation of interval estimates. The experiment consists of a sequence of sample averages based on binary (yes/no) outcomes considered earlier on in Comment 2.25. Figures A.1, A.2, A.3 and A.4 show 50 simulated (approximately) 95% confidence intervals for the binomial parameter p (the probability of success in a Bernoulli trial), based on different total number of trials (5,15,25 and 50). We note, in each case, that the statement *the interval contains the correct model mean*, that is, $p = 0.5$, would have been true in approximately 95% of the 50 simulated cases (there are about 2 or 3 total false statements in each case).

FIGURE A.1. 50 simulated 95% confidence intervals for $p=0.5$ based on $n=5$ binomial samples.

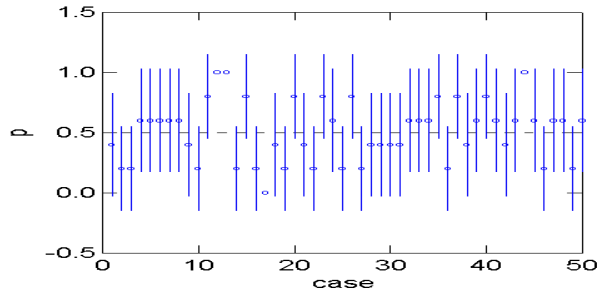


FIGURE A.2. 50 simulated 95% confidence intervals for $p=0.5$ based on $n=15$ binomial samples.

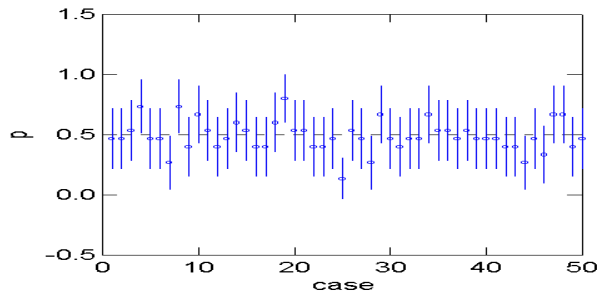
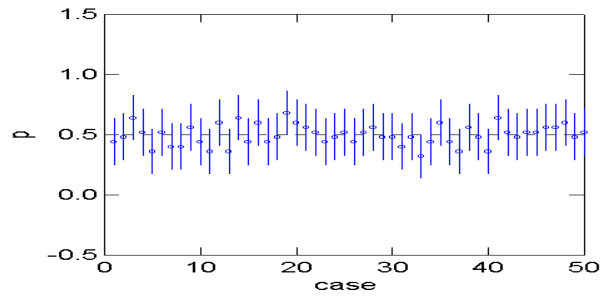
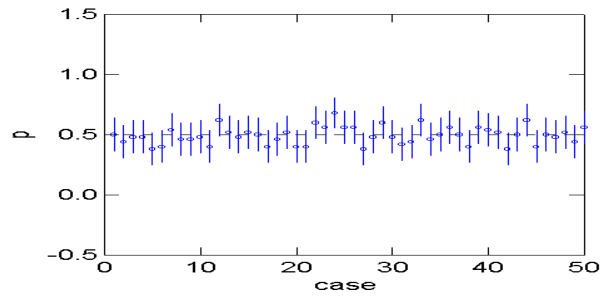


FIGURE A.3. 50 simulated 95% confidence intervals for $p=0.5$ based on $n=25$ binomial samplesFIGURE A.4. 50 simulated 95% confidence intervals for $p=0.5$ based on $n=50$ binomial samples

REFERENCES

- Edwards, W., Lindman, H. & Savage, J. (1963), 'Statistical inference for psychological research', *Psychological Review* **70**, 193–242.
- Lindley, D. & Phillips, L. (1976), 'Inference for a Bernoulli process (a Bayesian view)', *The American Statistician*.

Staff Statistician, General Clinical Research Center
Associate Professor, Department of Ophthalmology and Visual Sciences
viana@uic.edu