# Bayesian Statistics: Thomas Bayes to David Blackwell

Kathryn Chaloner

Department of Biostatistics
Department of Statistics & Actuarial Science
University of Iowa
kathryn-chaloner@uiowa.edu, or biostatistics@uiowa.edu
Field of Dreams, Arizona State University, Phoenix AZ

November 2013

# Probability

1. What is the probability of “heads” on a toss of a fair coin?
2. What is the probability of “six” upermost on a roll of a fair die?
3. What is the probability that the 100th digit after the decimal point, of the decimal expression of $\pi$ equals 3?
4. What is the probability that Rome, Italy, is North of Washington DC USA?
5. What is the probability that the sun rises tomorrow? (Laplace)

My answers: (1) $\frac{1}{2}$ (2) $\frac{1}{6}$ (3) $\frac{1}{10}$ (4) 0.99
Laplace’s answer to (5) 0.9999995

# Interpretations of Probability

- There are several *interpretations* of probability. The interpretation leads to methods for inferences under uncertainty. Here are the 2 most common interpretations:
  1. as a long run frequency (often the only interpretation in an introductory statistics course)
  2. as a subjective degree of belief
- You cannot put a long run frequency on an event that cannot be repeated.
  1. The 100th digit of $\pi$ is or is not 3. The 100th digit is constant no matter how often you calculate it.
  2. Similarly, Rome is North or South of Washington DC.

# The Mathematical Concept of Probability

First the Sample Space

- Probability theory is derived from a set of rules and definitions.
- Define a sample space $S$, and $\mathcal{A}$ a set of subsets of $S$ (events) with specific properties.
  1. $S \in \mathcal{A}$.
  2. If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$.
  3. If $A_1, A_2 \ldots$ is an infinite sequence of sets in $\mathcal{A}$ then $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$.
- A set of subsets of $S$ with these properties is called a $\sigma$-algebra.

# What is Probability?

## Probability is a Function

- Now define a non-negative function, $P$, on each event in $\mathcal{A}$, also satisfying specific properties.
  1. For any $A \in \mathcal{A}$, $P(A) \geq 0$ and $P(S) = 1$.
  2. If $A$ and $B$ are in $\mathcal{A}$ and $AB = \emptyset$ then $P(A \cup B) = P(A) + P(B)$.
  3. If $A_1, A_2, A_3, \ldots$ is an infinite sequence of events such that $A_i A_j = \emptyset$ for all $i \neq j$ then $P(\cup_{i=1}^{\infty} A_i) = \Sigma_{i=1}^{\infty} P(A_i)$.
- There is nothing controversial about the mathematical concept of probability.

# Subjective Probability

- Should “degrees of belief” follow rules?
- Should an individual represent their beliefs by a probability distribution? (A probability distribution follows the rules of probability).
- Building on others (Bayes, 1763; Laplace, 1814; Ramsey, 1926) Savage, 1954, gave a set of assumptions for rational behavior that beliefs should obey. They lead to a probability distribution as a consequence.
- Several alternative frameworks also led to probability as a consequence, most of them involve either gambling or a concept of what it means to be “rational”.

# Assumptions of DeGroot, 1971

$A \prec B$ pronounced "$A$ is less likely than $B$", $A \sim B$ as "$A$ is equally likely as $B$", and $A \succ B$ as "$B$ is less likely than $A$"

- $SP_1$: For any two events $A$ and $B$, exactly one of the following three relations must hold: $A \prec B$, $A \sim B$ or $A \succ B$.
- $SP_2$: If $A_1$, $A_2$, $B_1$, and $B_2$ are four events such that $A_1 A_2 = B_1 B_2 = \emptyset$ and $A_i \preceq B_i$ for $i = 1, 2$, then $A_1 \cup A_2 \preceq B_1 \cup B_2$. If, in addition, either $A_1 \prec B_1$ or $A_2 \prec B_2$, then $A_1 \cup A_2 \prec B_1 \cup B_2$.
- $SP_3$: If $A$ is any event then $\emptyset \preceq A$. Furthermore $\emptyset \prec S$.
- $SP_4$: If $A_1 \supset A_2 \supset \dots$ is a decreasing sequence of events and B is some fixed event such that $A_i \succeq B$ for $i = 1, 2, \dots$, then $\cap_{i=1}^{\infty} A_i \succeq B$.
- $SP_5$: There exists a random variable which has a uniform distribution on the interval $[0, 1]$.

- DeGroot proves that under the 5 assumptions about subjective beliefs, $SP_1$ to $SP_5$, for every event $A \in S$ there is a unique probability distribution $P$ that agrees with the subjective beliefs. He first constructs it and then he verifies that it is unique.

1. $SP_1$: For any two events $A$ and $B$, exactly one of the following three relations must hold: $A \prec B$, $A \sim B$ or $A \succ B$.
2. $SP_2$: If $A_1$, $A_2$, $B_1$, and $B_2$ are four events such that $A_1 A_2 = B_1 B_2 = \phi$ and $A_i \preceq B_i$ for $i = 1, 2$, then $A_1 \cup A_2 \preceq B_1 \cup B_2$. If, in addition, either $A_1 \prec B_1$ or $A_2 \prec B_2$, then $A_1 \cup A_2 \prec B_1 \cup B_2$.

- **Lemma 1:** Suppose that $A$, $B$, and $D$ are events such that $AD = BD = \emptyset$. Then $A \prec B$ if and only if $A \cup D \prec B \cup D$.
- **Proof:** Suppose $A \prec B$ then $SP_2$ gives us that $A \cup D \prec B \cup D$. Now suppose that $A \succ B$. Then, again by $SP_2$, $A \cup D \succ B \cup D$.

# Conditional Probability

For any three events $A$, $B$ and $D$, we need to extend the relation $\preceq$ to $(A|D) \preceq (B|D)$ to mean that: the event $B$ is at least as likely to occur as event $A$ given that event $D$ has occured.

- **Assumption CP**: For any three events $A$, $B$ and $D$, $(A|D) \preceq (B|D)$ if and only if $A \cap D \preceq B \cap D$.
- **Theorem**: If the relation $\preceq$ satisfies $SP_1$ to $SP_5$, and $CP$, then the function $P$ defined earlier is the unique probability distribuition which has the following property: for any three events $A$, $B$ and $D$ such that $P(D) > 0$, $(A|D) \preceq (B|D)$ if, and only if, $P(A|D) \leq P(B|D)$.

# Bayes Theorem

- Conditional probability of an event $A$ given event $B$, where $P(B) \neq 0$ is defined as

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)}$$

- More generally, if $A_i$, $i = 1, \ldots, k$ is a partition of $S$, that is $A_iA_j = \emptyset$, and $\Sigma_{i=1}^{i=k} P(A_i) = 1$. Then

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\Sigma_{i=1}^{i=k} P(B|A_i)P(A_i)}.$$

# Bayesian Statistics

- Using Bayesian statistics we can put a probability distribution on quantities that are fixed, but unknown.
- Specify a joint distribution for the data $y$ and the unknown parameters $\theta$: $p(y, \theta)$. Alternatively, specify a likelihood $p(y|\theta)$ and a prior distribution $p(\theta)$.
- After observing data $y = y_0$ update the prior distribution using Bayes theorem to $p(\theta|y = y_0)$.

# Outline

- Probability: Fermat (1601) & Pascal (1623), Jacob Bernoulli (1654), De Moivre (1718).
- **Bayes:** Subjective probability for inference about an unknown $\theta$ conditional on the data.
- Laplace: Proved Bayes' Theorem independently, and used it for inference.
- **Savage**
- **Blackwell**

Bayes, Savage and Blackwell all came from "non-traditional" backgrounds and faced challenges related to religion, or disability, or race.

Isaac Newton, 1642–1727.
Jacob Bernoulli, 1654– 1705.
Abraham de Moivre, 1667 – 1754.
**Thomas Bayes, 1701 –1761.**
Jacob Emanuel Euler, 1707 – 1756.
Pierre-Simon Laplace, 1749 – 1847.
J Carl Friedrich Gauss, 1777 – 1855

**Thomas Bayes**

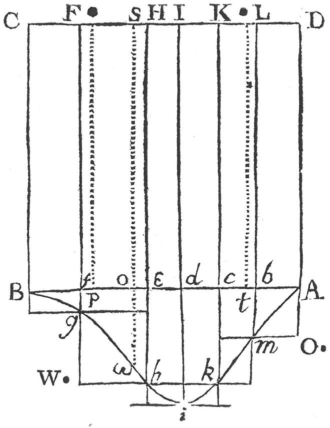
**He was born in 1701 or 1702 in Hertfordshire. His family then moved to a working class area of London where his father was a Presbyterian minister**

**He attended the University of Edinburgh 1719 until at least 1721.**

**He moved to Tunbridge Wells in ~1930 to take up a ministry.**

**Published a theological (1731) and mathematical (1736) treatise: his mathematical publication defends Newton's calculus against Berkley's philosophy. He died in 1761 and his statistical papers were published in 1763.**

# Bayes Example



# Bayes Example

PROBLEM.

*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a fingle trial lies fomewhere between any two degrees of probability that can be named.

# Bayes' Example:

- Roll $W$ once so that it is equally likely to land anywhere on the length of the table. Let $\theta$ represent where ball $W$ lands as a fraction of the length of the table $0 \leq \theta \leq 1$. Then $\theta$ is uniformly distributed on the interval $[0, 1]$.
- Let $X$ be the number of times that in $n$ independent rolls of the ball $O$, it lands to the left of the ball $W$.
- Note that $X \in \{0, 1, 2, \ldots, n-1, n\}$ and there are $n+1$ possible values for $Y$.

- The marginal distribution of $X$ is

$$\begin{aligned} & Pr(X = x) \\ = & \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} d\theta \text{ for } x \in \{0, 1, \ldots, n\} \\ = & \frac{1}{n+1}. \end{aligned}$$

# Bayes' Example Continued

- Bayes used Bayes Theorem to make *inference*.
- If we observe $X = x$ then inferences are through $p(\theta|X = x)$ the posterior distribution.

$$
\begin{aligned}
& p(\theta|X = x) \\
= \quad & \frac{p(X = x|\theta)p(\theta)}{Pr(X = x)} \\
= \quad & \frac{\binom{n}{x}\theta^{x}(1-\theta)^{n-x}}{(n+1)^{-1}} \quad \text{for } 0 \le \theta \le 1 \\
= \quad & \frac{(n+1)!}{(n-x)!x!}\theta^{x}(1-\theta)^{n-x} \quad \text{for } 0 \le \theta \le 1 \\
\text{and} \quad & 0 \quad \text{elsewhere}
\end{aligned}
$$

$$
\begin{aligned}
& p(\theta|X=x) \\
= \; & \frac{(n+1)!}{(n-x)!x!}\theta^{x}(1-\theta)^{n-x} \;\; \text{for } 0 \le \theta \le 1, \text{ and } 0 \text{ elsewhere} \\
= \; & \frac{\Gamma(n)}{\Gamma(n-x+1)\Gamma(x+1)}\theta^{x+1-1}(1-\theta)^{n-x+1-1} \;\; \text{for } 0 \le \theta \le 1 \\
= \; & \frac{1}{B(x+1,n-x+1)}\theta^{x+1-1}(1-\theta)^{n-x+1-1} \;\; 0 \le \theta \le 1
\end{aligned}
$$

Today we would recognize this as a Beta(x+1,n-x+1)distribution. The mean, the expectation of $\theta$ given $X=x$, is $\frac{x+1}{n+2}$ which Bayes also derived.

- The cumulative distribution function of $\theta$ given $X = x$ is

$$
\begin{aligned}
& Pr(\theta \leq \omega) \\
= \; & \frac{1}{B(x+1, n-x+1)} \int_{\theta=0}^{\theta=\omega} \theta^{x+1-1}(1-\theta)^{n-x+1-1} d\theta \\
= \; & \frac{B(\omega; x+1, n-x+1)}{B(x+1, n-x+1)}
\end{aligned}
$$

  where $B(\omega; x+1, n-x+1)$ is the incomplete Beta function and $B(x+1, n-x+1)$ is the usual Beta function.

- Beta functions and incomplete Beta functions can now be calculated numerically.

# Thomas Bayes' Results

Stigler, 1989

- Bayes noted that evaluation of $\int_0^f \theta^x (1-\theta)^{n-x} d\theta$ would complete the solution. The first extensive tables of this Incomplete Beta function were not compiled until the 20th century.
- Bayes sought to bound the Incomplete Beta function above and below, but his bounds were not very close.
- It is thought (Stigler, 1989) that Bayes was reluctant to publish his work without better bounds. It was read to the Royal Society after his death by Richard Price.

# Pierre-Simone Laplace

**Pierre–Simon Laplace, 1749--1827.**

**He is thought to be the son of a small cottager or a cider merchant. He attended a Benedictine priory school. At 16 he went to the University of Caen, and then to the Ecole Militaire.**

**He made important contributions to Mathematics, physics, astronomy and statistics, especially to Bayesian statistics.**

- Laplace also approached the problem of Binomial sampling for making inferences about a proportion $\theta$. He used De Moivre's urn models rather than rolling balls on a table.
- He derived Bayes' theorem without being aware of Bayes' work.
- He used Stirling's formula to derive a large sample normal approximation to the Beta posterior.
- In 1785 he published a long memoir that was later used to derive approximations to incomplete Beta functions.
- Laplace's 1774 analysis was directed towards showing what is now called posterior consistency.

- Bayes found the mean of $\frac{x+1}{n+2}$ when we observe $x$ successes out of $n$ trials.
- Laplace derived the same expression of $\frac{x+1}{n+2}$.
- Laplace used as an illustration the calculation of the probability of the event that the sun will rise tomorrow.
- Suppose the sun has risen every day for 5000 years. Laplace substituted $n = 5000 \times 365 = 1,825,000$ and $x = 5000 \times 365 = 1,825,000$ and so

$$\begin{aligned} Pr(\text{The sun will rise tomorrow}) &= \frac{1825001}{1825002} \\ &= 0.9999995 \end{aligned}$$

# Frank Ramsey

Frank Ramsey (1903 to 1930)

Born in Cambridge where his father was a mathematician.

He studied mathematics at Cambridge where he became a student of John Maynard Keynes. Despite his atheism he was appointed as a fellow at Cambridge in 1924 (non-Christians were allowed to become faculty in 1913).

Ramsey advised Wittgenstein on his PhD thesis.

Keynes argued against subjective probability, but Ramsey disagreed. His work was ignored until 1944.

- Ramsey developed a theory of rational behavior using betting. He developed a framework for rational behavior by avoiding bad bets. He developed the basis of Bayesian decision theory.
- His mentor was Keynes.
- He died when he was 26.
- It has been speculated that had Ramsey not died at 26, subjective probabilities and Bayesian statistics would have become standard. As it is, the frequentist interpretation became the norm.

# Leonard Jimmie Savage

**Leonard "Jimmie" Savage (Born Leonard Ogashevitz 1917 in Detroit, died aged 53 in 1971). His grandfather was an immigrant and his father a realtor. He had poor eyesight and he was initially educated at home, then at Central High School in Detroit. *"He was a brilliant child, but he paid no attention to what was going on in school because he couldn't see what was going on in school. The teachers thought he was more or less feebleminded."* His teachers refused to recommend him to the University of Michigan.**

**His father persuaded a friend to recommend him to Wayne University where he studied Engineering. He did well enough to be admitted later to UM to study Chemical Engineering. He was expelled from UM. He was allowed back to take math courses. His grades improved to C in analytic geometry; B in calculus; B in differential equations. Inspired by a faculty member, Raymond Wilder, he received all A's from then on.**

# Jimmie Savage

- Went on to a PhD in Math (1941) at UM, Institute for Advanced Study (1941-2). He joined the war effort and worked with von Neumann and with Courant. In 1946 he began at the University of Chigago, followed by a Guggenheim, and a Fulbright award.
- His 1954 book, *the Foundations of Statistics* built on Ramsey's and von Neumann's work. He developed the theory of utility and optimal decision makeing. The proofs were very rigorous.
- In the late 50's to the late 70's the debate between Bayesians and non-Bayesians became acrimonious.
- Savage persevered.

# Jimmie Savage

- Savage developed much of the Bayesian statistics in use today.
- His PhD students were Morris DeGroot and Don Berry who both continued his work on making Bayesian inference the foundations of modern statistics. Bayesian inference is now not controversial, in either mathematical statistics or its many applications (economics, biostatistics, F.D.A.).
- His many contributions to statistics, mathematics, economics (Milton Friedman, M.A. Gershick), biology and medicine have stood the test of time.

http://www-history.mcs.st-and.ac.uk/Biographies/Savage.html

# David Blackwell

David Blackwell, was born in Centralia IL in 1919 and died in 2010. He went to elementary school and high school in Centralia. He entered the University of Illinois in Urbana-Champagne and earned a bachelors degree in math in 1938, a masters in 1939, and a PhD in 1941 at the age of 22. Like Savage, he did a post-doc at the Institute for Advanced Study in 1941-2 and was a Fellow at Princeton. After his post doc he applied to all 105 HBCU's for a position. He also applied to Berkeley where Neyman supported him but others had race-based objections. He had appointments at Southern University (Baton Rouge), Clarke College and Howard in 1944. He visited Berkeley for one year and was then hired there as Professor, in the newly formed Statistics Department.

- He is known for the Rao-Blackwell theorem. If $X$ denoted data, and $g(X)$ is any estimator of a parameter $\theta$, then the conditional expectation of $g(X)|T(X)$ is also an estimator and is typically a better estimator. It cannot be a worse estimator – "Rao-Blackwellization".
- His collaboration with Girshick led to the classic book *Theory of Games and Statistical Decisions* by David Blackwell and M.A. Girshick.
- He worked on dynamic programming for use in optimal sequential statistial decision making and design.
- His book *Basic Statistics 1969* was the first fully Bayesian introductory textbook.

# Preface

*This book indicates the content of a lower-division basic statistics course I have taught several times at Berkeley. The students come from all departments of the university, and many of them have forgotten high-school algebra. The mathematical level of the course is modest: Any student who can do arithmetic, substitute in simple formulas, plot points, and draw a smooth curve through plotted points is ready for the course. But he must be prepared to think seriously about frivolous examples, as balls in urns are used to illustrate practically every idea introduced. The approach is intuitive, informal, concrete, decision-theoretic, and Bayesian.*

# David Blackwell's Conversation

Jimmie Savage "had just one influence, but it was a big one".

> *he explained to me that the Bayes approach was the right way to do inference. Let me tell you how that happened... an economist came in one day to talk to me. He said that he had a problem. They were preparing a recommendation to the Air Force on how to divide their research budget over the next 5 years and, in particular, they had to decide what fraction of it should be devoted to long-range research and what fraction of it should be devoted to developmental research.*

*"Now," he said, "one of the things that this depends upon is the probability of a major war in the next five years. If it's large then, of course, that would shift the emphasis towards developing what we already know how to do, and if it's small then there would be more emphasis on long-range research. . . . "if you could give me any guide as to how I could go about finding such a number I would be grateful"*

*"'Oh, I said to him, that question just does not make sense. Probability applies to a long sequence of repeatable events, and this is clearly a unique situation. The probability is either 0 or 1, but we won;t know for 5 years, I pontificated"' [Laughs]*

*‘So the economist looked at me and nodded and said ‘I was afraid you were going to say that. I have spoken to several other statisticians and they have all told me the same thing. Thank you very much”’.*

*A couple of weeks later Jimmie Savage came to visit . . .. I happened to mention this conversation that I had had, and then he started telling me about deFinetti and personal probability. Anyway, I walked out of his office half an hour later with a completely different view on things. I now understood what was the right way to do statistical inference*

*‘Looking back on it, I can see that I was emotionally and intellectually prepared for Jimmie’s message because I had been thinking in a Bayesian way about sequential analysis, hypothesis testing, and other statistical problems for some years.’*

# A Conversation with David Blackwell

**DeGroot**: Let's talk a little bit about the current state of statistics. What areas do you think are particularly important these days? Where do you see the field going?
**Blackwell**: I can tell you what I'd like to see happen. First, of course, I would like to see more emphasis on Bayesian statistics. . . ..

# Finally

**DeGroot:** You have a reputation as one of the finest lecturers in the field. Is that your style of lecturing?

**Blackwell:** I guess it is. I try to emphasize that with students. I notice that when students are talking about their theses or about their work, they want to tell you everything they know. So I say to them: You know much more about this topic than anybody else. We'll never understand it if you tell it all to us. Pick just one interesting thing. Maybe two.

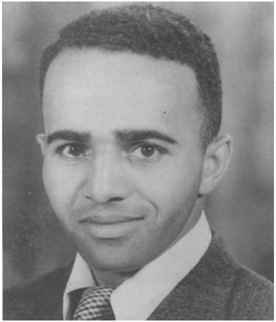
**DeGroot:** Thank you, David.

# David Blackwell

Kenneth Arrow, David Blackwell and M.A. Girshick,
Santa Monica September 1948

http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.ss/1177013814

**From "A Conversation with David Blackwell" by Morris DeGroot in *Statistical Science* vol. 1, 1989**



# References


- Steven Stigler, it A History of Statistics: The Measurement of Uncertainty Until 1900. Harvard University Press.
- J.J. O'Connor and E F Robertson, *Savage Biography*. http://www-history.mcs.st-and.ac.uk/Biographies/Savage.html
- Leonard J. Savage, 1954, The Foundations of Statistics, Dover Press.
- M.H. DeGroot, *Optimal Statistical Decisions*, Wiley Classics Library, 2004.
- J.J. O'Connor and E.F Robertson, *Biography of Frank Plumpton Ramsey*
- F.P. Ramsey, *Truth and Probability* Chapter VII in *The Foundations of mathematics and other Logical Essays* , 1931, reprinted in 2001, Routledge, ISBM 0-415-22546-9. http://www-history.mcs.st-and.ac.uk/Biographies/Ramsey.html
- American Statistical Association, *Statisticians in History* Biography of David Blackwell, http://www.amstat.org/about/statisticiansinhistory/index.cfm?fuseaction=main
- M.H. DeGroot, 1986, A Conversation with David Blackwell, Statistical Science, vol. 1, no. 1, p 40-53.
- David Blackwell, *Basic Statistics* McGraw-Hill, 1969. David Blackwell, M.A. Girshick, *Theory of Games and Statistical Decisions*, Dover Press, 1954.
- Wikipedia, entries for Jakob Bernoulli, Abraham de Moivre, Thomas Bayes, Pier-Simone Laplace, Frank P. Ramsey, Leonard Jimmie Savage, David Harold Blackwell. Accessed November 2013.