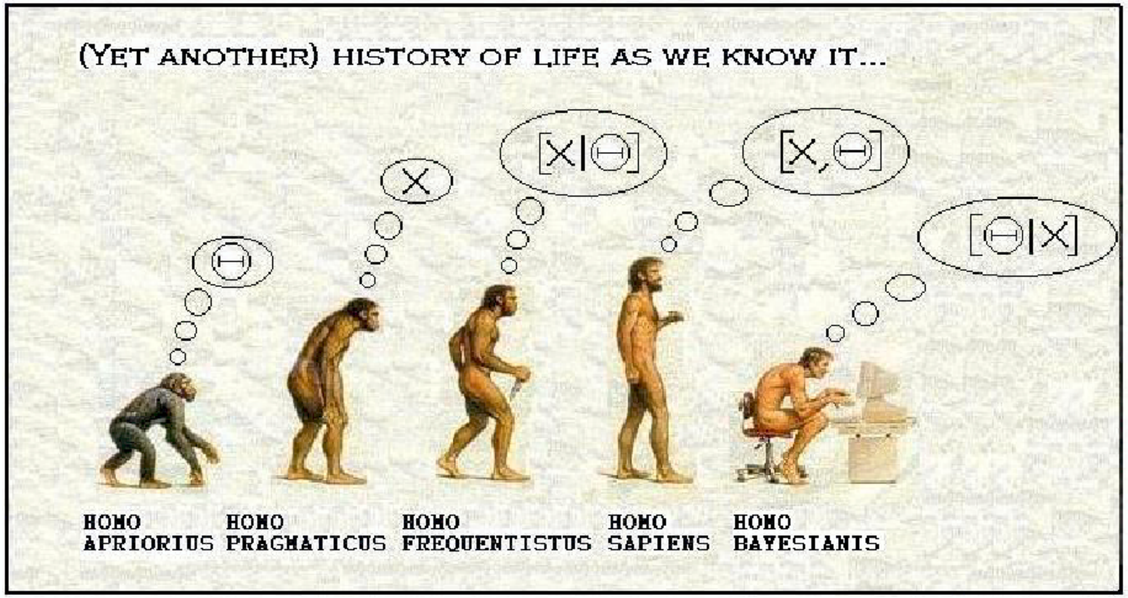
1. Introduction and non-Bayesian inference



Objective

Introduce the different *objective* and *subjective* interpretations of probability. Examine the various non-Bayesian treatments of statistical inference and comment on their associated problems.

Recommended reading

- Hájek, A. (2003). Interpretations of Probability. In *Stanford Encyclopedia of Philosophy*.

  http://plato.stanford.edu/entries/probability-interpret/

- The Wikipedia has a nice page on different interpretations of probability.

  http://en.wikipedia.org/wiki/Probability_interpretations

- Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*, Chapter 2.

  http://www.uv.es/bernardo/BT2.pdf

Probability

Kolmogorov

Probability theory developed from studies of games of chance by Fermat and Pascal and may be thought of as the study of randomness. It was put on a firm mathematical basis by Kolmogorov (1933).

The Kolmogorov axioms

For a *random experiment* with *sample space* $\mathbf{\Omega}$, then a probability measure P is a function such that

1. for any event $A \in \mathbf{\Omega}$, $P(A) \geq 0$.

2. $P(\mathbf{\Omega}) = 1$.

3. $P\left(\cup_{j \in J} A_j\right) = \sum_{j \in J} P(A_j)$ if $\{A_j : j \in J\}$ is a countable set of incompatible events.

The laws of probability can be derived as consequences of these axioms. However, this is a purely mathematical theory and does not provide a useful practical interpretation of probability.

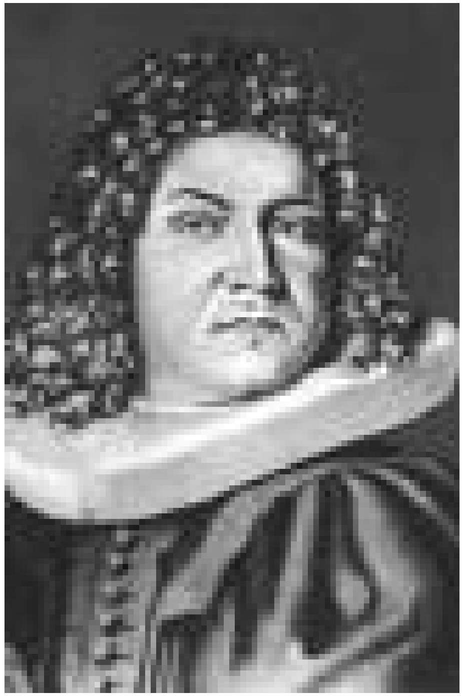
Interpretations of probability

There are various different ways of interpreting probability.

- The classical interpretation.
- Logical probability.
- Frequentist probability.
- Propensities.
- Subjective probability.

For a full review, see e.g. Gillies (2000).

Classical probability



Bernoulli

This derives from the ideas of Jakob Bernoulli (1713) contained in the *principle of insufficient reason* (or *principle of indifference*) developed by Laplace (1814) which can be used to provide a way of assigning epistemic or subjective probabilities.

The principle of insufficient reason

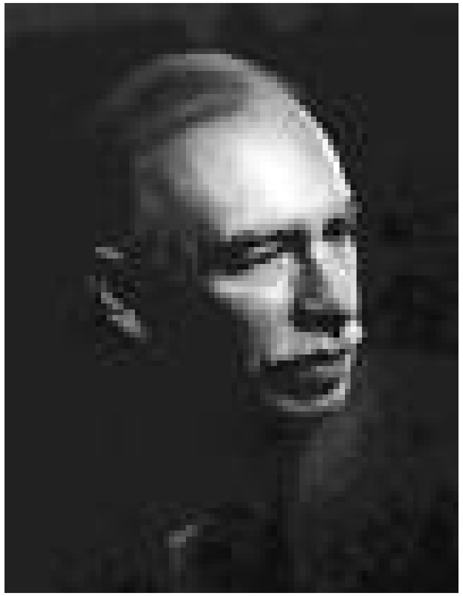
If we are ignorant of the ways an event can occur (and therefore have no reason to believe that one way will occur preferentially compared to another), the event will occur equally likely in any way.

Thus the probability of an event is the coefficient between the number of favourable cases and the total number of possible cases.

This is a very limited definition and cannot be easily applied in infinite dimensional or continuous sample spaces.

http://en.wikipedia.org/wiki/Principle_of_indifference

Logical probability



Keynes

Carnap

This approach, which extends the classical concept of probability, was developed by Keynes (1921) and Carnap (1950). The probability of a proposition H given evidence E is interpreted as the (unique) degree to which E logically entails H.

Logical probabilities are constructed using the idea of formal languages as follows:

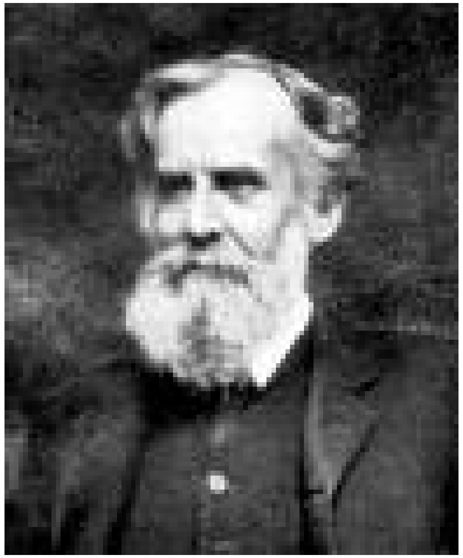
- Consider a language, L, with predicates $H_1, H_2, \ldots$ and a finite number of constants $e_1, e_2, \ldots, e_n$.

- Define a probability measure $P(\cdot)$ over sentences in L in a way that only takes into account their syntactic structure.

- Then use the standard probability ratio formula to create conditional probabilities over pairs of sentences in L.

Unfortunately, as noted by Bernardo and Smith (1994), "the logical view of probability is entirely lacking in operational content". Such probabilities are assumed to exist and depend on the formal language in which they are defined.

Frequentist probability



Venn

Von Mises

The idea comes from Venn (1876) and was expounded by von Mises (1919).

Given a repeatable experiment, the probability of an event is the limit of the proportion of times that the event will occur when the number of repetitions of the experiment tends to infinity.

This is a restricted definition of probability. It is impossible to assign probabilities in non repeatable experiments.

Propensities

Popper

This theory was developed by Popper (1957).

Probability is an innate disposition or propensity for things to happen. In particular, long run propensities seem to coincide with the frequentist definition of probability whereas it is not clear what individual propensities are, or whether they obey the probability calculus.

Subjective probability

Ramsey

The subjective concept of probability is as *degrees of belief*. A first attempt to formalize this was made by Ramsey (1926).

In reality, most people are irrational in that their own degrees of belief do not satisfy the probability axioms, see e.g. Kahneman et al (1982) and chapter 5 of this course. Thus, in order to formalize the definition of subjective probability, it is important to only consider *rational agents*, i.e. agents whose beliefs are logically consistent.

Consistent degrees of belief are probabilities

Cox (1946) formalized the conditions for consistent reasoning by assuming defining the necessary logical conditions. Firstly, relative beliefs in the truths of different propositions are assumed to be transitive, i.e. if we believe that $A \succeq B$ and $B \succeq C$, for three events A, B, C, then we must believe that $A \succeq C$, where $A \succeq B$ means A is at least as likely as B to occur. This assumption implies that we can represent degrees of belief by numbers, where the higher the value, the higher the degree of belief.

Secondly, it is assumed that if we specify how much we believe an event A to be true, then we also implicitly specify how much we believe it to be false.

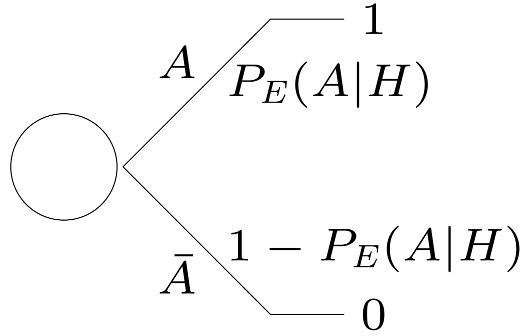
Finally it is assumed that if we specify how much we believe A to be true and how much we believe B to be true, given that A is true, then we are implicitly implying our degree of belief that both A and B are true.

Given these axioms, Cox was able to demonstrate that logical consistency could only be assured if the numbers used to represent degrees of belief are probabilities.

Defining subjective probabilities

Cox's method is non-constructive and does not give a method of defining a probability for a given event. There are various ways of doing this, usually based around the ideas of betting.

De Finetti (1937) defines la probability, $P_E(A|H)$, of an event A for an agent E with information H to be the maximum quantity, or fair price, that E would pay for a ticket in a lottery where they will obtain a (small) unit prize if and only if A occurs.



The expected gain in this lottery, is equal to

$$P_E(A|H) \times 1 + (1 - P_E(A|H)) \times 0 = P_E(A|H).$$

A *Dutch book* is defined to be a series of bets, each acceptable to the agent, which collectively imply that the agent is certain to lose money, however the world turns out.

It can now be shown that in order to avoid a Dutch book, then the agent's probabilities must satisfy the usual probability calculus.

Note that the definition proposed by De Finetti depends on the assumption that the *utility* of the agent for money is linear. A more general method of defining probabilities and utilities is provided by Savage (1954). See also O' Hagan (1988) chapters 1 to 3 for a definition based on betting and odds.

Statistical Inference

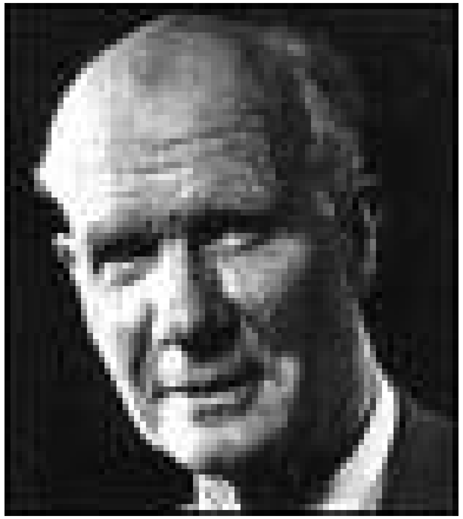
A number of different approaches to statistical inference have been developed based on both the frequentist and subjective concepts of probability.

- Classical (frequentist) inference.
- Likelihood based approaches.
- Fiducial statistics and related methods.
- Bayesian inference.

For full comparisons of the different approaches see e.g. Barnett (1999).

Classical inference

Neyman



Pearson

This approach was developed from the ideas of Neyman and Pearson (1933) and Fisher (1925).

Characteristics of classical inference

- Frequentist interpretation of probability.

- Inference is based on the likelihood function $l(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})$.

- We can only quantify (a priori) the uncertainty about $\mathbf{X}$. $\boldsymbol{\theta}$ is fixed.

- Inferential procedures are based on asymptotic performance:

  ◇ An estimator $\mathbf{t} = \mathbf{t}(\mathbf{X})$ is defined.
  ◇ The plausibility of a given value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is measured by the density

$$l(\boldsymbol{\theta}_0|\mathbf{x}) \propto f(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}_0),$$

  (assuming that $t(\cdot)$ is a sufficient statistic).
  ◇ If $\mathbf{t}$ does not lie in the tail of $f(\mathbf{t}|\boldsymbol{\theta}_0)$ then $\boldsymbol{\theta}_0$ is a plausible value of $\boldsymbol{\theta}$.

Classical estimation and hypothesis tests

Classical point estimation is based on choosing estimators with good asymptotic properties (unbiasedness, minimum variance, efficiency etc.)

There are a number of possible methods of choosing an estimator, e.g. method of moments, maximum likelihood, etc. Note that in particular, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is defined so that $l(\hat{\boldsymbol{\theta}}|\mathbf{x}) > l(\boldsymbol{\theta}|\mathbf{x})$ for all $\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}$. The MLE is asymptotically unbiased, efficient etc.

For interval estimation, an interval $(l(\mathbf{x}),\ u(\mathbf{x}))$ is chosen so that

$$P(l(\mathbf{X}) < \boldsymbol{\theta} < u(\mathbf{X})|\boldsymbol{\theta}) = 1 - \alpha$$

for some fixed probability level α.

Finally, hypothesis testing is based on rejecting $\boldsymbol{\theta}_0$ at level α if

$$P(t(\mathbf{X}) > t(\mathbf{x})|\boldsymbol{\theta}_0) < \alpha.$$

The principle of sufficiency

Definition 1
A statistic $\mathbf{t} = \mathbf{t}(\mathbf{x})$ is said to be sufficient for $\boldsymbol{\theta}$ if

$$f(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta}) = f(\mathbf{x}|\mathbf{t}).$$

The *sufficiency principle* (Fisher 1922) is as follows.

If a sufficient statistic, $\mathbf{t}$, exists, then for any two samples $\mathbf{x}_1$, $\mathbf{x}_2$ of the same size such that $\mathbf{t}(\mathbf{x}_1) = \mathbf{t}(\mathbf{x}_2)$ then the conclusions given $\mathbf{x}_1$ and $\mathbf{x}_2$ should be the same.

All standard methods of inference satisfy this principle.

The Fisher-Neyman factorization theorem

This gives a useful characterization of a sufficient statistic which we shall use in Chapter 3.

Theorem 1
A statistic $\mathbf{t}$ is sufficient for $\boldsymbol{\theta}$ if and only if there exist functions g and h such that

$$l(\boldsymbol{\theta}|\mathbf{x}) = g(\mathbf{t}, \boldsymbol{\theta})h(\mathbf{x}).$$

Proof See e.g. Bernardo and Smith (1994). ∎

The principle of repeated sampling

The inference that we draw from $\mathbf{x}$ should be based on an analysis of how the conclusions change with variations in the data samples, which would be obtained through hypothetical repetitions, under exactly the same conditions, of the experiment which generated the data x in the first place.

This principle follows directly from the frequentist definition of probability and is much more controversial. It implies that measures of uncertainty are just hypothetical asymptotic frequencies. Thus, it is impossible to measure the uncertainty about $\boldsymbol{\theta}$ *a posteriori*, given $\mathbf{x}$.

Criticisms of classical inference

Global criticisms

As noted earlier, the frequentist definition of probability restricts the number of problems which can be reasonably analyzed. Furthermore, classical inference appears to be a form of *cookbook* containing many seemingly *ad-hoc* procedures.

Specific criticisms

We can also criticize many specific aspects of the classical approach.

Firstly, there are often problems with estimation. An optimal method of choosing an estimator is not generally available and some typically used approaches may provide very bad estimators.

One possibility is to attempt to use an unbiased estimator.

Example 1
Let X be an observation from a Poisson distribution: $\mathcal{P}(\lambda)$ and assume that we wish to estimate $\phi = e^{-2\lambda}$. Then only one unbiased estimator of ϕ exists and takes the value $\tilde{\phi} = (-1)^X = \pm 1$. However $0 < \phi \leq 1$ for all λ.

An alternative is to use the method of moments. However, this approach is not always available.

Example 2
Suppose that we wish to estimate the location parameter θ of a Cauchy distribution. Then no simple method of moments estimator exists.

Also, method of moments of estimators, when they do exist do not have the optimality properties of e.g. maximum likelihood estimators.

It is more common to use maximum likelihood estimators. These are justified asymptotically but can have very bad properties in small samples.

Example 3

Let $X \sim \mathcal{DU}[1, \theta]$. Then given a sample of size 1, the MLE of θ is $\hat{\theta} = X$. The bias of the MLE is

$$E[X] - \theta = \frac{\theta + 1}{2} - \theta = \frac{1 - \theta}{2},$$

which can be enormous if θ is large.

Whatever the sample size, the MLE is always underestimates the true value in this experiment.

Example 4

Suppose that $\mathbf{Y} \sim \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2 \mathbf{I}\right)$ where $\dim(\boldsymbol{\theta}) = n$, and that we wish to estimate $\boldsymbol{\theta}$ given $\mathbf{Y} = \mathbf{y}$. Then clearly, the maximum likelihood estimator is $\hat{\boldsymbol{\theta}} = \mathbf{y}$. ▮

However, if $n \geq 3$ then this estimator is inadmissible as the James-Stein estimator $\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|\mathbf{y}\|^2}\right) \mathbf{y}$ has lower mean squared error.

We will return to this example in Chapter 10.

A more important problem is interpretation.

Example 5

Suppose that we carry out an experiment and find that a 95% confidence interval for θ based on the sample data is equal to $(1, 3)$. How do we interpret this?

This means that if we repeated the same experiment and procedure by which we constructed the interval many times, 95% of the constructed intervals would contain the true value of θ. ▮

It does not mean that the probability that θ lies in the interval $(1, 3)$ is 95%.

There are often problems in dealing with nuisance parameters in classical inference.

Example 6

Let $Y_{i,j} \sim \mathcal{N}\left(\phi_i, \sigma^2\right)$ for $i = 1, \ldots, n$ and $j = 1, 2$.

Suppose that the parameter of interest is the variance, σ^2 and that $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ are nuisance parameters.

The likelihood is $l\left(\sigma^2, \boldsymbol{\phi}|\mathbf{y}\right) \propto$

$$\sigma^{-2n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_{i,1} - \phi_i)^2 + (y_{i,2} - \phi_i)^2\right).$$

and now, the most natural way of estimating σ^2 is to maximize the *profile likelihood*. This is calculated as follows.

Supposing σ^2 known, we first maximize the likelihood with respect to ϕ.

$$
\begin{aligned}
l_P(\sigma^2|\mathbf{y}) &= \sup_{\boldsymbol{\phi}} l\left(\sigma^2, \boldsymbol{\phi}|\mathbf{y}\right) \\
&= \sigma^{-2n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_{i,1}-\bar{y}_i\right)^2 + \left(y_{i,2}-\bar{y}_i\right)^2\right) \\
&= \sigma^{-2n} \exp\left(-\frac{1}{4\sigma^2}\sum_{i=1}^{n}\left(y_{i,1}-y_{i,2}\right)^2\right)
\end{aligned}
$$

Secondly, we maximize the profile likelihood with respect to σ^2, which implies that

$$
\hat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n}(y_{i,1}-y_{i,2})^2.
$$

However,

$$
E\left[\hat{\sigma}^2\right] = \frac{\sigma^2}{2}
$$

for any value of n and this estimator is inconsistent.

It is also not entirely clear how predictions should be carried out.

Example 7

Let $X|\theta \sim f(\cdot|\theta)$. Then typically, the predictive distribution is estimated, given the sample data, by substituting θ by its MLE, leading to the estimated predictive density $f(x|\hat{\theta})$. However this procedure clearly underestimates the predictive uncertainty due to the fact that θ is unknown.

Likelihood based inference

Barnard

This approach is based totally on the likelihood function and derives from Barnard (1949) and Barnard et al (1962). Firstly, the likelihood principle is assumed.

The likelihood principle

This principle is originally due to Barnard (1949). Edwards (1992) defines the likelihood principle as follows.

Within the framework of a statistical model, all the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data. ...For a continuum of hypotheses, this principle asserts that the likelihood function contains all the necessary information.

This is equivalent to supposing that if we have experiments, E_i of the same size with sample data $\mathbf{x}_i$ and sampling distributions $f_i(\cdot|\boldsymbol{\theta})$ for $i = 1, 2$, then the two experiments provide the same evidence (EV) and hence the same inference about $\boldsymbol{\theta}$ if $f_1(\mathbf{x}_1|\boldsymbol{\theta}, E_1) \propto f_2(\mathbf{x}_2|\boldsymbol{\theta}, E_2)$, that is

$$l(\boldsymbol{\theta}|E_2, \mathbf{x}_2) = c\,l(\boldsymbol{\theta}|E_1, \mathbf{x}_1) \quad \text{for some } c \;\Rightarrow EV[E_1, \mathbf{x}_1] = EV[E_2, \mathbf{x}_2].$$

A second assumption of likelihood inference is that the likelihood ratio for $l(\boldsymbol{\theta}_1|\mathbf{x})/l(\boldsymbol{\theta}_2|\mathbf{x})$, for any two values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is a measure of the evidence supplied by the data in favour of $\boldsymbol{\theta}_1$ relative to $\boldsymbol{\theta}_2$.

This is much harder to apply in practice, in particular in the presence of nuisance parameters. For example, how can we marginalize the likelihood function?

Stopping rules

Stopping rules are often used in classical statistics (e.g. clinical trials) to make it possible to stop early if the results are sufficiently favourable or unfavourable.

Frequentist statisticians must choose the stopping rule before the experiment begins and must stick to it exactly (otherwise the good frequentist properties of the test are lost).

Example 8

Assume we are interested in testing whether a coin is biased in favour of heads. Then we may consider various different stopping rules, e.g.

- toss the coin a fixed number, n times.
- toss the coin until the first tail is seen.
- toss the coin until the r'th tail is seen.

The stopping rule principle

The stopping rule principle is the following.

In a sequential experiment, the evidence provided by the experiment about the value of the unknown parameters $\boldsymbol{\theta}$ should not depend on the stopping rule.

This is clearly a consequence of the likelihood principle as the likelihood function is independent of the stopping rule. However ...

Classical hypothesis tests do not satisfy the stopping rule principle

Example 9

Suppose that $\theta = P(\text{head})$. We wish to test $H_0 : \theta = 1/2$ against the alternative $H_1 : \theta > 1/2$ at a 5% significance level.

Suppose that we observe 9 heads and 3 tails. This information is not sufficient for us to write down the likelihood function. We need to know the sampling scheme or stopping rule.

Suppose that we fix the number of tosses of the coin to be 12. Then

$$X = \#\ \text{heads}|\theta \sim \mathcal{BI}(12, \theta) \quad \text{and} \quad l(\theta|x = 9) = \left(\begin{array}{c} 12 \\ 9 \end{array} \right) \theta^9 (1 - \theta)^3.$$

Thus, the p-value is

$$p_1 = \frac{1}{2}^{12} \left[\left(\begin{array}{c} 12 \\ 9 \end{array} \right) + \cdots + \left(\begin{array}{c} 12 \\ 12 \end{array} \right) \right] \approx .075$$

and we do not reject the null hypothesis.

Suppose now that we decided to calculate the number of heads X until the third tail occurs. Thus, $X|\theta \sim \mathcal{NB}(3, \theta)$ and

$$l(\theta|x) = \left(\begin{array}{c} 11 \\ 9 \end{array} \right) \theta^9 (1-\theta)^3 \quad \text{and the p-value is}$$

$$\begin{aligned} p_2 &= \left(\begin{array}{c} 11 \\ 9 \end{array} \right) \theta^9 (1-\theta)^3 + \left(\begin{array}{c} 12 \\ 10 \end{array} \right) \theta^{10} (1-\theta)^3 + \cdots \\ &= .0325 \quad \text{and we reject the null hypothesis.} \end{aligned}$$

The reason is that in order to carry out a hypothesis test, then we must specify the sample space or stopping rule. This is different in the two cases that we have seen:

1. $\Omega = \{(u, d) : u + d = 12\}$

2. $\Omega = \{(u, d) : d = 3\}$

The conditionality principle

Suppose that we have the possibility of carrying out two experiments E_1 and E_2 in order to make inference about $\boldsymbol{\theta}$ and that we choose the experiment to carry out by tossing an unbiased coin. Then our inference for $\boldsymbol{\theta}$ should only depend on the selected experiment.

To formalize this, note that the experiment E_i may be characterized as $E_i = (\mathbf{X}_i, \boldsymbol{\theta}, f_i)$ which means that in this experiment, the variable $\mathbf{X}_i$ is generated from $f_i(\mathbf{X}_i|\boldsymbol{\theta})$.

Now define the composite experiment $E^\star$ which consists of generating a random variable K where $P(K = 1) = P(K = 2) = \frac{1}{2}$ and then performing experiment K so that $E^\star = (\underbrace{(K, \mathbf{X}_K)}_{\mathbf{X}}, \boldsymbol{\theta}, \frac{1}{2} f_K(\mathbf{X}_K))$. Then conditionality implies that

$$EV[E^\star, \mathbf{x}] = \begin{cases} EV[E_1, \mathbf{x}_1] & \text{if } K = 1 \text{ so } \mathbf{x} = (1, \mathbf{x}_1) \\ EV[E_2, \mathbf{x}_2] & \text{if } K = 2 \text{ so } \mathbf{x} = (1, \mathbf{x}_2) \end{cases}$$

Birnbaum (1962) demonstrates the following theorem which relates the sufficiency, likelihood and conditionality principles.

Theorem 2
The likelihood principle is equivalent to the sufficiency principle and the conditionality principle.

We shall demonstrate that sufficiency plus conditionality $\Rightarrow$ likelihood. See Birnbaum (1962) for a full proof.

Proof of the theorem

Proof Let E_1, E_2 and $\mathbf{x}_1$ and $\mathbf{x}_2$ be the two experiments and data samples figuring in the statement of the likelihood principle so that

$$l(\boldsymbol{\theta}|E_1, \mathbf{x}_1) = cl(\boldsymbol{\theta}|E_2, \mathbf{x}_2)$$

for some constant c and define the composite experiment $E^\star = E_1$ as in the conditionality principle.

Define the statistic $\mathbf{t}$ by

$$\mathbf{t} = \mathbf{t}(\mathbf{x}) = \begin{cases} (1, \mathbf{x}_1) & \text{if } K = 2 \\ \mathbf{x} & \text{otherwise} \end{cases}$$

Now note that if $\mathbf{t} \neq (1, \mathbf{x}_1)$ then

$$f(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta}) = \begin{cases} 1 & \text{if } \mathbf{t} = \mathbf{t}(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases}$$

whereas if $\mathbf{t} = \mathbf{x} = (1, \mathbf{x}_1)$, then

$$f(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta}) = \frac{\frac{1}{2} f_1(\mathbf{x}_1|\boldsymbol{\theta})}{\frac{1}{2} f_1(\mathbf{x}_1|\boldsymbol{\theta}) + \frac{1}{2} f_2(\mathbf{x}_2|\boldsymbol{\theta})} = \frac{c}{1+c}$$

and if $\mathbf{t} = (1, \mathbf{x}_1)$ but $\mathbf{x} = (2, \mathbf{x}_2)$ then

$$f(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta}) = \frac{\frac{1}{2} f_2(\mathbf{x}_2|\boldsymbol{\theta})}{\frac{1}{2} f_1(\mathbf{x}_1|\boldsymbol{\theta}) + \frac{1}{2} f_2(\mathbf{x}_2|\boldsymbol{\theta})} = \frac{1}{1+c}$$

so $f(\mathbf{x}|\mathbf{t}, \boldsymbol{\theta}) = f(\mathbf{x}|\mathbf{t})$ and $\mathbf{t}$ is a sufficient statistic.

It follows from the sufficiency principle that

$$EV[E^\star, (1, \mathbf{x}_1)] = EV[E^\star, (2, \mathbf{x}_2)]$$

and then the conditionality principle ensures that

$$EV[E^\star, \mathbf{x}_1] = EV[E_1, (1, \mathbf{x}_i)] = EV[E^\star, (2, \mathbf{x}_2)] = EV[E_2, \mathbf{x}_2]$$

which is the likelihood principle. ■

In the same way, it can be demonstrated that likelihood + sufficiency $\Rightarrow$ conditionality or that likelihood + conditionality $\Rightarrow$ sufficiency.

Fiducial inference and related methods

Fisher

Fiducial inference has the objective of defining a posterior measure of uncertainty for $\boldsymbol{\theta}$ without the necessity of defining a prior measure. This approach was introduced by Fisher (1930).

Example 10

Let $X|\mu,\sigma^2 \sim \mathcal{N}(\mu,\sigma^2)$. Suppose that we wish to carry out inference for μ.

We know that $T = \frac{(\bar{X}-\mu)\sqrt{n}}{S} \sim t_{n-1}$. Then for any t, $P(T > t) = p(t)$ where $p(t)$ is known. Fisher's idea is to write

$$\begin{aligned} p(t) &= P(T > t) \\ &= P\left(\frac{(\bar{X}-\mu)\sqrt{n}}{S} > t\right) \\ &= P\left(\mu < \bar{X} - \frac{St}{\sqrt{n}}\right) \end{aligned}$$

and then define $p(t) = P\left(\mu < \bar{x} - \frac{st}{\sqrt{n}}\right)$ to be the fiducial probability that μ is less than $\bar{x} - \frac{st}{\sqrt{n}}$.

Problems with the fiducial approach

- The probability measure is transferred from the sample space to the parameter space. What is the justification for this?
- What happens if no pivotal statistic exists?
- It is unclear how to apply the fiducial approach in multidimensional problems.

In many cases, fiducial probability intervals coincide with Bayesian credible intervals given specific non informative prior distributions. In particular, structural inference, see Fraser (1968) corresponds to Bayesian inference using so called Haar prior distributions. However we will see that the Bayesian justification for such intervals is more coherent.

References

Barnard, G.A. (1949). Statistical Inference, *Journal of the Royal Statistical Society. Series B*, *11*, 115–149.

Barnard, G.A., Jenkins, G.M. and Winsten, C.B. (1962) Likelihood Inference and Time Series. *Journal of the Royal Statistical Society, Series A*, **125**, 321–372.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. New York: Wiley.

Barnett, V. (1999). *Comparative Statistical Inference* (3rd ed.). Chichester: Wiley

Bernoulli, J. (1713). *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis*. Basel: Thurneysen Brothers.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, **57**, 269–306.

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press.

Cox, R. (1946). Probability, Frequency, and Reasonable Expectation, *Am. Jour. Phys.*, **14**, 1–13.

Edwards, A.W.F. (1992). *Likelihood* (2nd ed.) Cambridge: University Press.

Finetti, R. de (1937). La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de l'Institut Henri Poincaré*, **7**, 1–68.

Fisher, R.A. (1922). On the Mathematical Foundations Of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A*, **222**, 309-68.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fisher, R.A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, **26**, 528–535.

Fraser, D.A.S. (1968). *The Structure of Inference*. New York: Wiley.

Gillies, D. (2000). *Philosophical theories of probability*. Routledge.

Hájek, A. (2003). Interpretations of Probability. In *Stanford Encyclopedia of Philosophy*.

Kahneman, D., Slovic, P. and Tversky, A. (1982). *Judgement under Uncertainty: Heuristics and Biases*. Cambridge: University Press.

Keynes, J.M. (1921). *A Treatise on Probability*. London: Macmillan.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.

Laplace, P.S. (1814). *Théorie Analytique des probabilités*. Paris: Courcier Imprimeur.

Mises, R. von (1919). *Wahrscheinlichkeit, Statistik und Wahrheit*. Vienna, Springer.

Neyman, J. and Pearson, E. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses, *Philosophical Transactions of the Royal Society of London. Series A*, **231**, 289–337.

O' Hagan, A. (1988). *Probability: Methods and Measurement*. London: Chapman and Hall.

Pederson, J.G. (1978). Fiducial Inference. *International Statistical Review*, **46**, 147-170.

Popper, K. (1957). The Propensity Interpretation of the Calculus of Probability and of the Quantum Theory. In *Observation and Interpretation*. Buttersworth Scientific Publications.

Ramsey, F.P. (1926). Truth and Probability. In *Foundations of Mathematics and other Essays*. Routledge and Kegan.

Savage, L.J. (1954). *The Foundations of Statistics*. New York: Wiley.

Venn, J. (1876). *The Logic of Chance: An Essay on the Foundations and Province of the Theory of Probability*.