Introduction to **Bayesian Statistics**

for non-mathematicians

- **By:** Dr. J. Andrés Christen (Centro de Investigación en Matemáticas, CIMAT. Perteneciente a la red de centros CONACYT).
- **Prerequisits:** Elements of calulus and probobility (basic).
- **Lenght:** 5 hours (two sessions of 2:30 hours each).
- **e-mail:** jac@cimat.mx .

Introduction to **Bayesian Statistics** for non-mathematicians

- **By:** Dr. J. Andrés Christen (Centro de Investigación en Matemáticas, CIMAT. Perteneciente a la red de centros CONACYT).
- **Prerequisits:** Elements of calulus and probobility (basic).
- **Lenght:** 5 hours (two sessions of 2:30 hours each).
- **e-mail:** jac@cimat.mx .

Introduction to **Bayesian Statistics** for non-mathematicians

- **By:** Dr. J. Andrés Christen (Centro de Investigación en Matemáticas, CIMAT. Perteneciente a la red de centros CONACYT).
- **Prerequisits:** Elements of calulus and probobility (basic).
- **Lenght:** 5 hours (two sessions of 2:30 hours each).
- **e-mail:** jac@cimat.mx .

Introduction to **Bayesian Statistics** for non-mathematicians

- **By:** Dr. J. Andrés Christen (Centro de Investigación en Matemáticas, CIMAT. Perteneciente a la red de centros CONACYT).
- **Prerequisits:** Elements of calulus and probobility (basic).
- **Lenght:** 5 hours (two sessions of 2:30 hours each).
- **e-mail:** jac@cimat.mx .

Introduction to **Bayesian Statistics** for non-mathematicians

- **By:** Dr. J. Andrés Christen (Centro de Investigación en Matemáticas, CIMAT. Perteneciente a la red de centros CONACYT).
- **Prerequisits:** Elements of calulus and probobility (basic).
- **Lenght:** 5 hours (two sessions of 2:30 hours each).
- **e-mail:** jac@cimat.mx .

Texts:

1. Lee, P. (1994), *Bayesian Statyistics: An Introduction*, London: Edward Arnold.
2. J. O. Berger (1985), *Statistical Decision Theory: foundations, concepts and methods*, Second Edition, Springer-Verlag.
3. Bernardo, J. M. and Smith, A. F. M. (1994), *Bayesian Theory*, Wiley: Chichester, UK.
4. M. H. DeGroot (1970), *Optimal statistical decisions*, McGraw–Hill: NY.

Texts:

1. Lee, P. (1994), *Bayesian Statyistics: An Introduction*, London: Edward Arnold.
2. J. O. Berger (1985), *Statistical Decision Theory: foundations, concepts and methods*, Second Edition, Springer-Verlag.
3. Bernardo, J. M. and Smith, A. F. M. (1994), *Bayesian Theory*, Wiley: Chichester, UK.
4. M. H. DeGroot (1970), *Optimal statistical decisions*, McGraw–Hill: NY.

Texts:

1. Lee, P. (1994), *Bayesian Statyistics: An Introduction*, London: Edward Arnold.
2. J. O. Berger (1985), *Statistical Decision Theory: foundations, concepts and methods*, Second Edition, Springer-Verlag.
3. Bernardo, J. M. and Smith, A. F. M. (1994), *Bayesian Theory*, Wiley: Chichester, UK.
4. M. H. DeGroot (1970), *Optimal statistical decisions*, McGraw–Hill: NY.

Texts:

1. Lee, P. (1994), *Bayesian Statyistics: An Introduction*, London: Edward Arnold.
2. J. O. Berger (1985), *Statistical Decision Theory: foundations, concepts and methods*, Second Edition, Springer-Verlag.
3. Bernardo, J. M. and Smith, A. F. M. (1994), *Bayesian Theory*, Wiley: Chichester, UK.
4. M. H. DeGroot (1970), *Optimal statistical decisions*, McGraw–Hill: NY.

Texts:

Colin Howson and Peter Urbach (2006) **Scientific Reasoning: The Bayesian Approach**, (3rd Ed.), Open Court.

"Two English philosophers provocatively argue the case for Bayesian logic, with a minimum of complex math. They claim that Bayesian thinking is identical to the scientific method and give fascinating examples of how to analyze beliefs, such as Macbeth's doubting of the witches' prophecy, the discovery of Neptune on the strength of faith in Newton's laws but zero evidence, and why people get hooked on Dianetics.", – Discover.

"For the first time, we have a book that combines philosophical wisdom, mathematical skill, and statistical appreciation, to produce a coherent system." – Dennis V. Lindley, University College, London (ret.).

Conditional Probability

A cornerstone of Bayesian statistics is its (alternative) definition of probability, a definition sufficiently wide to cover many interesting cases. Let's start with some examples:

1. What is the probability that if I toss a coin it lands on "heads"?
2. What is the probability that your lecturer has more than the equivalent of 50 pesos in his pocket?
3. What is the probability that it rains tomorrow?
4. What is the probability that it rained yesterday in Washington?
5. What is the probability that our Galaxy has more than 10^9 stars?

Conditional Probability

A cornerstone of Bayesian statistics is its (alternative) definition of probability, a definition sufficiently wide to cover many interesting cases. Let's start with some examples:

1. What is the probability that if I toss a coin it lands on "heads"?
2. What is the probability that your lecturer has more than the equivalent of 50 pesos in his pocket?
3. What is the probability that it rains tomorrow?
4. What is the probability that it rained yesterday in Washington?
5. What is the probability that our Galaxy has more than 10^9 stars?

Conditional Probability

A cornerstone of Bayesian statistics is its (alternative) definition of probability, a definition sufficiently wide to cover many interesting cases. Let's start with some examples:

1. What is the probability that if I toss a coin it lands on "heads"?
2. What is the probability that your lecturer has more than the equivalent of 50 pesos in his pocket?
3. What is the probability that it rains tomorrow?
4. What is the probability that it rained yesterday in Washington?
5. What is the probability that our Galaxy has more than 10^9 stars?

Conditional Probability

A cornerstone of Bayesian statistics is its (alternative) definition of probability, a definition sufficiently wide to cover many interesting cases. Let's start with some examples:

1. What is the probability that if I toss a coin it lands on "heads"?
2. What is the probability that your lecturer has more than the equivalent of 50 pesos in his pocket?
3. What is the probability that it rains tomorrow?
4. What is the probability that it rained yesterday in Washington?
5. What is the probability that our Galaxy has more than 10^9 stars?

Conditional Probability

A cornerstone of Bayesian statistics is its (alternative) definition of probability, a definition sufficiently wide to cover many interesting cases. Let's start with some examples:

1. What is the probability that if I toss a coin it lands on "heads"?
2. What is the probability that your lecturer has more than the equivalent of 50 pesos in his pocket?
3. What is the probability that it rains tomorrow?
4. What is the probability that it rained yesterday in Washington?
5. What is the probability that our Galaxy has more than 10^9 stars?

More over...

A piece of maize with several kernels found in a clay pot believed to belong to the last days of the Mexica umpire are radiocarbon dated.

What is the age of the pot?

More over...

A piece of maize with several kernels found in a clay pot believed to belong to the last days of the Mexica umpire are radiocarbon dated.

What is the age of the pot?

More over...

A piece of maize with several kernels found in a clay pot believed to belong to the last days of the Mexica umpire are radiocarbon dated.

What is the age of the pot?

All probabilities are conditional

(on the person or *agent* speaking, assumptions made, data used, etc.).

Probability statements go beyond favorable/possible calculations.

In Bayesian statistics, **all** uncertainties about unknowns are measured with a probability distribution.

All probabilities are conditional

(on the person or *agent* speaking, assumptions made, data used, etc.).

Probability statements go beyond favorable/possible calculations.

In Bayesian statistics, **all** uncertainties about unknowns are measured with a probability distribution.

All probabilities are conditional

(on the person or *agent* speaking, assumptions made, data used, etc.).

Probability statements go beyond favorable/possible calculations.

In Bayesian statistics, **all** uncertainties about unknowns are measured with a probability distribution.

Informal Bayesian definition of Probability

Probability is an opinion hold by an *agent*, that may be turned into a bet under suitable circumstances.

If you say the probability of an event E is p, the you would take a bet of at most $a = \frac{1-p}{p}$ to 1 on E being true.

Informal Bayesian definition of Probability

Probability is an opinion hold by an *agent*, that may be turned into a bet under suitable circumstances.

If you say the probability of an event E is p, the you would take a bet of at most $a = \frac{1-p}{p}$ to 1 on E being true.

Preferences among events

Bayesian statistics, unlike other paradigms for inference, is based on a theory, that is, a set of axioms that creates a general procedure to make inferences. We briefly present the theory given in DeGroot (1970, cap. 6). We begin with a quote by DeGroot (1970, p. 70):

> *...suitable probabilities can often be assigned objectively and quickly because of wide agreement on the appropriateness of a specific distribution for a certain type of problem...On the other hand, there are some situations for which it would be very difficult to find even two people who would agree on the appropriateness of any specific distribution.*

Preferences among events

We have a total event Ω and a set o events @ ($(\Omega, @)$ is a *mesurable* set), we have:

$$A \succ B, \quad A \prec B, \quad A \sim B.$$

to mean that A is less (more, equal) likely that B. Also

$$A \preceq B$$

means that A is no more likely than B.

Preferences among events

We have a total event Ω and a set o events @ $((\Omega, @))$ is a *mesurable* set), we have:

$$A \succ B, \quad A \prec B, \quad A \sim B.$$

to mean that A is less (more, equal) likely that B. Also

$$A \preceq B$$

means that A is no more likely than B.

Axioms

A set of axioms are given for the preference relation $\preceq$, for a rational *agent*:

A complete ordering axiom:

Axiom

For any two events $A, B \in \copyright$, we have exactly one of the three following preference relations: $A \succ B, \quad A \prec B, \quad A \sim B.$

A transitivity axiom similar to this (a more general version is needed though):

Axiom

Si $A, B, C \in \copyright$, are three events $A \preceq B$ y $B \preceq C$, then $A \preceq C$.

Axioms

A set of axioms are given for the preference relation $\preceq$, for a rational *agent*:

A complete ordering axiom:

Axiom

For any two events $A, B \in \mathbb{Q}$, *we have exactly one of the three following preference relations:* $A \succ B, \quad A \prec B, \quad A \sim B$.

A transitivity axiom similar to this (a more general version is needed though):

Axiom

Si $A, B, C \in \mathbb{Q}$, *are three events* $A \preceq B$ *y* $B \preceq C$, *then* $A \preceq C$.

Axioms

A set of axioms are given for the preference relation $\preceq$, for a rational *agent*:

A complete ordering axiom:

Axiom

For any two events $A, B \in$ @, *we have exactly one of the three following preference relations:* $A \succ B, \quad A \prec B, \quad A \sim B.$

A transitivity axiom similar to this (a more general version is needed though):

Axiom

Si $A, B, C \in$ @, *are three events* $A \preceq B$ *y* $B \preceq C$, *then* $A \preceq C$.

A non triviality axiom

Axiom

For $A \in$ @ any event, then $\emptyset \preceq A$. Moreover, $\emptyset \preceq \Omega$.

And a continuity axiom, a technicality to be able to work with continuos distributions, like the gaussian:

Axiom

If $A_1 \supset A_2 \supset \cdots$ ia a decreasing sequence of events in @ and $B \in$ @ is another event such that $A_i \succeq B$ for all i, then $\cap_{i=1}^{\infty} A_i \succeq B$.

A non triviality axiom

Axiom

For $A \in @$ any event, then $\emptyset \preceq A$. Moreover, $\emptyset \preceq \Omega$.

And a continuity axiom, a technicality to be able to work with continuos distributions, like the gaussian:

Axiom

If $A_1 \supset A_2 \supset \cdots$ ia a decreasing sequence of events in $@$ and $B \in @$ is another event such that $A_i \succeq B$ for all i, then $\cap_{i=1}^{\infty} A_i \succeq B$.

The auxiliary experiment

One further axiom is needed. This axiom more or less says that some "standard" events are added to our sets of events, and this in turn are compared with the standard events.

Suppose for example, that we spin a roulette and all events regarding the final position of the roulette are compared with our "relevant" events.

Bayesian Inference

Uncertainty is quantified with a probability measure

Bayes' Theorem: Modify our probability measure with evidence

All probability is conditional (to assumptions made, *agent* speaking etc.)

$$P(\cdot \mid H), \quad \text{with } H = \text{particular context, } \textit{agent} \text{ speaking etc..}$$

Now, let $B \in$ @ and observable event What is the probability of $A \in$ @ given that we have observed B?

Bayesian Inference

Uncertainty is quantified with a probability measure

Bayes' Theorem: Modify our probability measure with evidence

All probability is conditional (to assumptions made, *agent* speaking etc.)

$$P(\cdot \mid H), \quad \text{with } H = \text{particular context, } \textit{agent} \text{ speaking etc..}$$

Now, let $B \in @$ and observable event What is the probability of $A \in @$ given that we have observed B?

Bayesian Inference

Uncertainty is quantified with a probability measure

Bayes' Theorem: Modify our probability measure with evidence

All probability is conditional (to assumptions made, *agent* speaking etc.)

$$P(\cdot \mid H), \quad \text{with } H = \text{particular context, } \textit{agent} \text{ speaking etc..}$$

Now, let $B \in$ @ and observable event What is the probability of $A \in$ @ given that we have observed B?

We are talking about the event $A \mid H, B$ and we may calculate its probability by means of

$$P(A \mid H, B) = \frac{P(A \cap B \mid H)}{P(B \mid H)},$$

or

$$P(A \mid H, B) = \frac{P(B \mid H, A)P(A \mid H)}{P(B \mid H)}.$$

Let's look closer at

$$P(A \mid H, B) = \frac{P(B \mid H, A)P(A \mid H)}{P(B \mid H)}.$$

- $P(A \mid H)$ we call it *a priori* probability or "prior", for A.
- $P(A \mid H, B)$ we call it *a posteriori* o posterior probability for A, given that we have observed B.
- $P(B \mid H, A)$ is our model...How the observables would be if we knew A? How the data B would be if we knew what we don't know A (unknown parameters, for example)?
- $P(B \mid H)$ is a normalization constant $P(\cdot \mid H, B)$ is a modified measure, then we may say that

$$P(\cdot \mid H, B) \propto P(B \mid H, \cdot)P(\cdot \mid H).$$

Commonly, conditioning on H is only done implicitly.

Let's look closer at

$$P(A \mid H, B) = \frac{P(B \mid H, A)P(A \mid H)}{P(B \mid H)}.$$

- $P(A \mid H)$ we call it *a priori* probability or "prior", for A.
- $P(A \mid H, B)$ we call it *a posteriori* o posterior probability for A, given that we have observed B.
- $P(B \mid H, A)$ is our model...How the observables would be if we knew A? How the data B would be if we knew what we don't know A (unknown parameters, for example)?
- $P(B \mid H)$ is a normalization constant $P(\cdot \mid H, B)$ is a modified measure, then we may say that

$$P(\cdot \mid H, B) \propto P(B \mid H, \cdot)P(\cdot \mid H).$$

Commonly, conditioning on H is only done implicitly.

Let's look closer at

$$P(A \mid H, B) = \frac{P(B \mid H, A) P(A \mid H)}{P(B \mid H)}.$$

- $P(A \mid H)$ we call it *a priori* probability or "prior", for A.
- $P(A \mid H, B)$ we call it *a posteriori* o posterior probability for A, given that we have observed B.
- $P(B \mid H, A)$ is our model...How the observables would be if we knew A? How the data B would be if we knew what we don't know A (unknown parameters, for example)?
- $P(B \mid H)$ is a normalization constant $P(\cdot \mid H, B)$ is a modified measure, then we may say that

$$P(\cdot \mid H, B) \propto P(B \mid H, \cdot) P(\cdot \mid H).$$

Commonly, conditioning on H is only done implicitly.

Let's look closer at

$$P(A \mid H, B) = \frac{P(B \mid H, A)P(A \mid H)}{P(B \mid H)}.$$

- $P(A \mid H)$ we call it *a priori* probability or "prior", for A.
- $P(A \mid H, B)$ we call it *a posteriori* o posterior probability for A, given that we have observed B.
- $P(B \mid H, A)$ is our model...How the observables would be if we knew A? How the data B would be if we knew what we don't know A (unknown parameters, for example)?
- $P(B \mid H)$ is a normalization constant $P(\cdot \mid H, B)$ is a modified measure, then we may say that

$$P(\cdot \mid H, B) \propto P(B \mid H, \cdot)P(\cdot \mid H).$$

Commonly, conditioning on H is only done implicitly.

Let's look closer at

$$P(A \mid H, B) = \frac{P(B \mid H, A) P(A \mid H)}{P(B \mid H)}.$$

- $P(A \mid H)$ we call it *a priori* probability or "prior", for A.
- $P(A \mid H, B)$ we call it *a posteriori* o posterior probability for A, given that we have observed B.
- $P(B \mid H, A)$ is our model...How the observables would be if we knew A? How the data B would be if we knew what we don't know A (unknown parameters, for example)?
- $P(B \mid H)$ is a normalization constant $P(\cdot \mid H, B)$ is a modified measure, then we may say that

$$P(\cdot \mid H, B) \propto P(B \mid H, \cdot) P(\cdot \mid H).$$

Commonly, conditioning on H is only done implicitly.

Let's look closer at

$$P(A \mid H, B) = \frac{P(B \mid H, A)P(A \mid H)}{P(B \mid H)}.$$

- $P(A \mid H)$ we call it *a priori* probability or "prior", for A.
- $P(A \mid H, B)$ we call it *a posteriori* o posterior probability for A, given that we have observed B.
- $P(B \mid H, A)$ is our model...How the observables would be if we knew A? How the data B would be if we knew what we don't know A (unknown parameters, for example)?
- $P(B \mid H)$ is a normalization constant $P(\cdot \mid H, B)$ is a modified measure, then we may say that

$$P(\cdot \mid H, B) \propto P(B \mid H, \cdot)P(\cdot \mid H).$$

Commonly, conditioning on H is only done implicitly.

Posterior Distribution

Suppose a random variable with $X_i = 0, 1$, that is $X_i \mid p \sim Be(p)$ independent and uncertainty about $p \in [0, 1]$ is quantified with $f(p)$ and $p \sim Beta(\alpha, \beta)$ *a priori*. We obtain that $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and

$$P(p \leq p_0 \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid p \leq p_0) P(p \leq p_0)}{P(\mathbf{X})}.$$

But

$$P(\mathbf{X} \mid p \leq p_0) P(p \leq p_0) = P(\mathbf{X}, p \leq p_0) = \int_0^{p_0} f(\mathbf{X}, p) dp.$$

Now $f(\mathbf{X}, p) = f(\mathbf{X} \mid p) f(p)$ and then

Posterior Distribution

Suppose a random variable with $X_i = 0, 1$, that is $X_i \mid p \sim Be(p)$

independent and uncertainty about $p \in [0, 1]$ is quantified with $f(p)$ and $p \sim Beta(\alpha, \beta)$ *a priori*. We obtain that $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and

$$P(p \leq p_0 \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid p \leq p_0) P(p \leq p_0)}{P(\mathbf{X})}.$$

But

$$P(\mathbf{X} \mid p \leq p_0) P(p \leq p_0) = P(\mathbf{X}, p \leq p_0) = \int_0^{p_0} f(\mathbf{X}, p) dp.$$

Now $f(\mathbf{X}, p) = f(\mathbf{X} \mid p) f(p)$ and then

$$P(p \leq p_0 \mid \mathbf{X}) \propto \int_0^{p_0} f(\mathbf{X} \mid p) f(p) dp.$$

The left hand side of the above expression is the posterior cdf of p, and thus by definition its **posterior density** is

$$f(p \mid \mathbf{X}) \propto f(\mathbf{X} \mid p) f(p).$$

Moreover

$$f(\mathbf{X} \mid p) = \prod_{i=1}^{n} f(X_i|p) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}$$

and

$$f(p) = B(\alpha, \beta)^{-1} p^{\alpha-1}(1-p)^{\beta-1},$$

and then

$$f(p \mid \mathbf{X}) \propto p^{(\alpha+\sum_{i=1}^n X_i)-1}(1-p)^{(\beta+n-\sum_{i=1}^n X_i)-1}.$$

Therefore

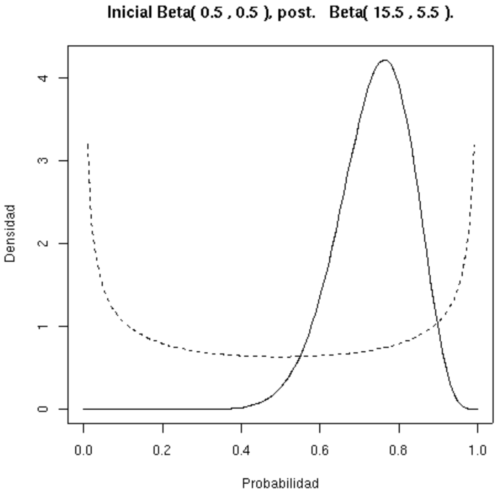
$$p \mid \mathbf{X} \sim Beta\left(\alpha + \sum_{i=1}^{n} X_i, \beta + n - \sum_{i=1}^{n} X_i\right).$$

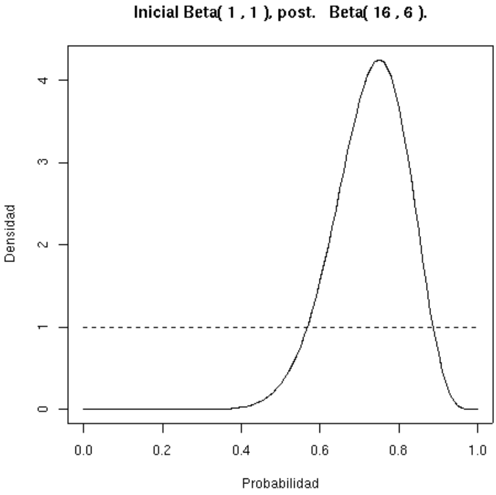
We present some priors and posterior (Beta) for p

Therefore

$$p \mid \mathbf{X} \sim Beta\left(\alpha + \sum_{i=1}^{n} X_i, \beta + n - \sum_{i=1}^{n} X_i\right).$$

We present some priors and posterior (Beta) for p


Inicial Beta(0.5 , 0.5), post.   Beta(15.5 , 5.5).
Densidad
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
Probabilidad


Inicial Beta(1 , 1), post.   Beta(16 , 6).
Densidad
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
Probabilidad

Example 2

We have a couple that has had 5 pregnancies and all 5 have been male, What is the probability that thire next pregnancy results is female?

Example 2

1. Are pregnancies independent with respect to the resulting gender?
2. Are there only two possible outputs?

Then the Bernoulli inference model explained above is valid and should be used.

Check possibilities in R.

Example 2

1. Are pregnancies independent with respect to the resulting gender?
2. Are there only two possible outputs?

Then the Bernoulli inference model explained above is valid and should be used.
Check possibilities in R.

Example 2

1. Are pregnancies independent with respect to the resulting gender?
2. Are there only two possible outputs?

Then the Bernoulli inference model explained above is valid and should be used.
Check possibilities in R.

Example 2

1. Are pregnancies independent with respect to the resulting gender?
2. Are there only two possible outputs?

Then the Bernoulli inference model explained above is valid and should be used.

Check possibilities in R.

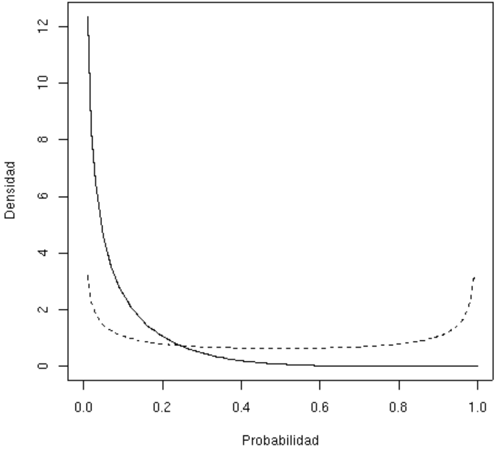
Example 2

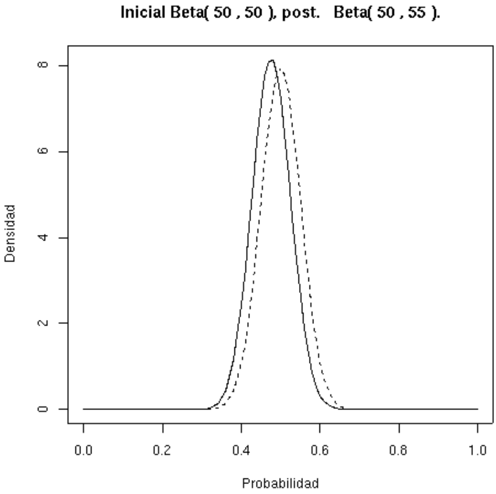
But the question is ...what prior would you use?

Example 2

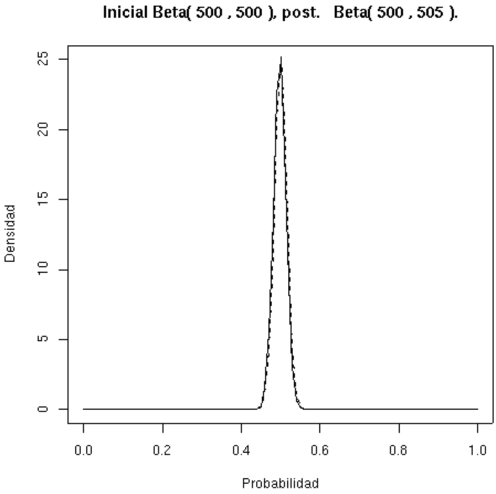
But the question is **...what prior would you use?**

Inicial Beta(0.5 , 0.5), post. Beta(0.5 , 5.5).





Inicial Beta(50 , 50), post.   Beta(50 , 55).
Densidad
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Probabilidad


Inicial Beta(500 , 500), post.   Beta(500 , 505).
Densidad
0
5
10
15
20
25
0.0
0.2
0.4
0.6
0.8
1.0
Probabilidad

Example: Normal sampling

In this case $X_i \sim N(\theta, \sigma^2)$, $i = 1, 2, \ldots, n$ (independent) with σ known and $\theta \sim N(\theta_0, \sigma_0^2)$ *a priori*:

$$f(\theta \mid \mathbf{X}) \propto \exp - \left\{ \frac{(\theta - \theta_0)^2}{2\sigma_0^2} + \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2} \right\}.$$

We see that the posterior is of the form $\exp h(\theta)$, where $h(\cdot)$ is a quadratic function of θ. Then $\theta \mid \mathbf{X}$ has a Normal distribution. Compleating the squares we obtain

$$f(\theta \mid \mathbf{X}) \propto \exp\left\{ -\frac{(\theta - \theta_p)^2}{2\sigma_p^2} + C \right\},$$

where $\sigma_p^2 = 1/(\sigma_0^{-2} + n\sigma^{-2})$, $\theta_p = \sigma_p^2(\mu_0/\sigma_0^2 + nm/\sigma^2)$, $m = 1/n \sum_{i=1}^{n} x_i$ and C does not depend on θ. Then

$$\theta \mid \mathbf{X} \sim N(\theta_p, \sigma_p^2).$$

Point and interval estimation

The main objective of any Bayesian analysis is finding the posterior distribution of interest. A secondary (although very important issue) is making proper outlines of this posterior distribution. For example, if we have

$$f(\theta_1, \theta_2 \mid \mathbf{X})$$

(a bivariate distribution), what would you do if only θ_1 is of interest?

We need the posterior of θ_1, and this may be obtained my marginalization, that is

$$f(\theta_1 \mid \mathbf{X}) = \int f(\theta_1, \theta_2 \mid \mathbf{X}) d\theta_2.$$

This is the so called marginal posterior density of θ_1 and etc.

Assuming we have the posterior $f(\theta \mid \mathbf{X})$, we only need to report it somehow: How would you report the following distributions (see figura 2).

We need the posterior of θ_1, and this may be obtained my marginalization, that is

$$f(\theta_1 \mid \mathbf{X}) = \int f(\theta_1, \theta_2 \mid \mathbf{X}) d\theta_2.$$

This is the so called marginal posterior density of θ_1 and etc.

Assuming we have the posterior $f(\theta \mid \mathbf{X})$, we only need to report it somehow: How would you report the following distributions (see figura 2).

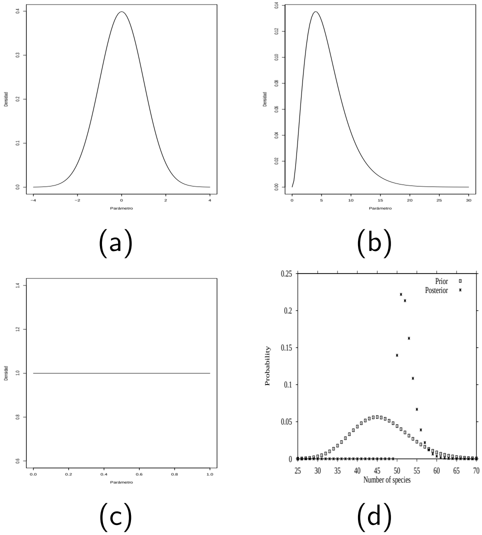


Figure: How would you report the following posterior distributions?

The concept of (point or interval or else) “estimation” in Bayesian statistics is only understood as outlines of the relevant posterior distribution (of course, there are good and bad outlines). Therefore, for example, point estimation may be understood as making an outline of a complete probability distribution with a single point, as absurd as this may be.

We could use the expected value of the posterior distribution.
Or we could use the maximum of the posterior distribution, this is the so called the MAP (*Maximum a posteriori*).

The concept of (point or interval or else) “estimation” in Bayesian statistics is only understood as outlines of the relevant posterior distribution (of course, there are good and bad outlines). Therefore, for example, point estimation may be understood as making an outline of a complete probability distribution with a single point, as absurd as this may be.

We could use the expected value of the posterior distribution.
Or we could use the maximum of the posterior distribution, this is the so called the MAP (*Maximum a posteriori*).

The concept of (point or interval or else) “estimation” in Bayesian statistics is only understood as outlines of the relevant posterior distribution (of course, there are good and bad outlines). Therefore, for example, point estimation may be understood as making an outline of a complete probability distribution with a single point, as absurd as this may be.

We could use the expected value of the posterior distribution.
Or we could use the maximum of the posterior distribution, this is the so called the MAP (*Maximum a posteriori*).

Calculus

At the end of the day, we will need

1. $f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n)$, a model.
2. $f(\theta_1, \theta_2, \ldots, \theta_n)$ a prior distribution for parameters.
3. The normalization constant

$$f(\mathbf{X}) = \int\int\cdots\int f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n) f(\theta_1, \theta_2, \ldots, \theta_n) d\theta_1 d\theta_2 \cdots d\theta_n$$

4. To obtain the posterior

$$f(\theta_1, \theta_2, \ldots, \theta_n \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n) f(\theta_1, \theta_2, \ldots, \theta_n)}{f(\mathbf{X})}.$$

5. And outlines of these posteriors, like marginal distributions etc. $f(\theta_1 \mid \mathbf{X}) = \int\int\cdots\int f(\theta_1, \theta_2, \ldots, \theta_n \mid \mathbf{X}) d\theta_2 d\theta_3 \cdots d\theta_n$.

Calculus

At the end of the day, we will need

1. $f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n)$, a model.
2. $f(\theta_1, \theta_2, \ldots, \theta_n)$ a prior distribution for parameters.
3. The normalization constant
$$f(\mathbf{X}) = \int \int \cdots \int f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n) f(\theta_1, \theta_2, \ldots, \theta_n) d\theta_1 d\theta_2 \cdots d\theta_n$$
4. To obtain the posterior
$$f(\theta_1, \theta_2, \ldots, \theta_n \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n) f(\theta_1, \theta_2, \ldots, \theta_n)}{f(\mathbf{X})}.$$
5. And outlines of these posteriors, like marginal distributions etc. $f(\theta_1 \mid \mathbf{X}) = \int \int \cdots \int f(\theta_1, \theta_2, \ldots, \theta_n \mid \mathbf{X}) d\theta_2 d\theta_3 \cdots d\theta_n$.

Calculus

At the end of the day, we will need

1. $f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n)$, a model.
2. $f(\theta_1, \theta_2, \ldots, \theta_n)$ a prior distribution for parameters.
3. The normalization constant

$$f(\mathbf{X}) = \int \int \cdots \int f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n) f(\theta_1, \theta_2, \ldots, \theta_n) d\theta_1 d\theta_2 \cdots d\theta_n$$

4. To obtain the posterior

$$f(\theta_1, \theta_2, \ldots, \theta_n \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n) f(\theta_1, \theta_2, \ldots, \theta_n)}{f(\mathbf{X})}.$$

5. And outlines of these posteriors, like marginal distributions etc. $f(\theta_1 \mid \mathbf{X}) = \int \int \cdots \int f(\theta_1, \theta_2, \ldots, \theta_n \mid \mathbf{X}) d\theta_2 d\theta_3 \cdots d\theta_n$.

Calculus

At the end of the day, we will need

1. $f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n)$, a model.
2. $f(\theta_1, \theta_2, \ldots, \theta_n)$ a prior distribution for parameters.
3. The normalization constant

$$f(\mathbf{X}) = \int \int \cdots \int f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n) f(\theta_1, \theta_2, \ldots, \theta_n) d\theta_1 d\theta_2 \cdots d\theta_n$$

4. To obtain the posterior

$$f(\theta_1, \theta_2, \ldots, \theta_n \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n) f(\theta_1, \theta_2, \ldots, \theta_n)}{f(\mathbf{X})}.$$

5. And outlines of these posteriors, like marginal distributions etc. $f(\theta_1 \mid \mathbf{X}) = \int \int \cdots \int f(\theta_1, \theta_2, \ldots, \theta_n \mid \mathbf{X}) d\theta_2 d\theta_3 \cdots d\theta_n$.

Calculus

At the end of the day, we will need

1. $f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n)$, a model.
2. $f(\theta_1, \theta_2, \ldots, \theta_n)$ a prior distribution for parameters.
3. The normalization constant
$$f(\mathbf{X}) = \int\int\cdots\int f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n) f(\theta_1, \theta_2, \ldots, \theta_n) d\theta_1 d\theta_2 \cdots d\theta_n$$
4. To obtain the posterior
$$f(\theta_1, \theta_2, \ldots, \theta_n \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid \theta_1, \theta_2, \ldots, \theta_n) f(\theta_1, \theta_2, \ldots, \theta_n)}{f(\mathbf{X})}.$$
5. And outlines of these posteriors, like marginal distributions etc. $f(\theta_1 \mid \mathbf{X}) = \int\int\cdots\int f(\theta_1, \theta_2, \ldots, \theta_n \mid \mathbf{X}) d\theta_2 d\theta_3 \cdots d\theta_n$.

Hypotheses testing as an inference problem

If $\theta \in \Theta$ our interest is

$$H_1 : \theta \in \Theta_1, \quad H_2 : \theta \in \Theta_2$$

these hypotheses, where Θ_1 y Θ_2 form a *partition* of Θ, that is, $\Theta_1 \cap \Theta_2 = \emptyset$ y $\Theta_1 \cup \Theta_2 = \Omega$. In Bayesian statistics terms, given a model $f(X \mid \theta)$, a *a priori* $f(\theta)$ and observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, What could it mean to "test" the above hypotheses?

Remmember:

Uncertanty is quantified with a probability measure

Hypotheses testing as an inference problem

If $\theta \in \Theta$ our interest is

$$H_1 : \theta \in \Theta_1, \quad H_2 : \theta \in \Theta_2$$

these hypotheses, where Θ_1 y Θ_2 form a *partition* of Θ, that is, $\Theta_1 \cap \Theta_2 = \emptyset$ y $\Theta_1 \cup \Theta_2 = \Omega$. In Bayesian statistics terms, given a model $f(X \mid \theta)$, a *a priori* $f(\theta)$ and observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, What could it mean to "test" the above hypotheses?
Remmember:
Uncertanty is quantified with a probability measure

Let $f(\theta)$ an *a priori* for θ. We calculate

$$P(H_i) = \int_{\Theta_i} f(\theta \mid \mathbf{X}) d\theta$$

and "prefer" or "data support" H_1 if $P(H_1) > P(H_2)$ (equivalently for H_2). Moreover, we could have more than two hypotheses

$$H_i : \quad \theta \in \Theta_i,$$

and we would only require the the corresponding posterior probability for each of them.

Let $f(\theta)$ an *a priori* for θ. We calculate

$$P(H_i) = \int_{\Theta_i} f(\theta \mid \mathbf{X}) d\theta$$

and "prefer" or "data support" H_1 if $P(H_1) > P(H_2)$ (equivalently for H_2). Moreover, we could have more than two hypotheses

$$H_i : \quad \theta \in \Theta_i,$$

and we would only require the the corresponding posterior probability for each of them.

Example: Hypothese testing

We have an experimental treatment for a condition which is used in 20 patients with similar cohort characteristics, from which 15 have recover from the condition (success). The standard treatment has a probability of success of 50%. The following hypotheses is stated *The experimental treatment is superior to the standard treatment.*

The hypotheses can be translated as

$$H_1 : \theta > 0.5, \quad H_2 : \theta \leq 0.5$$

where θ is the probability of success of the experimental treatment. Not much is known about the experimental treatment and a uniform (flat; Beta(1, 1)) prior is used. The corresponding posterior is Beta(16, 6), see figure.

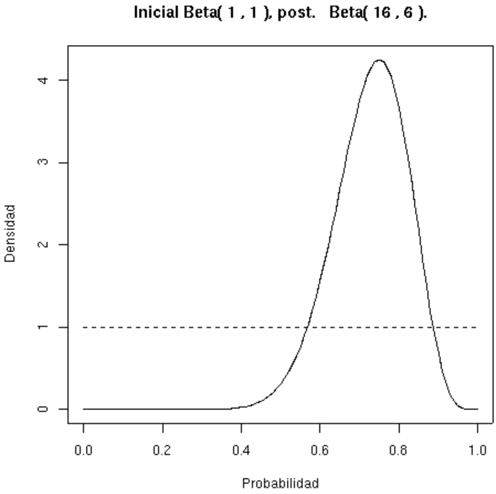
Example: Hypothese testing

We have an experimental treatment for a condition which is used in 20 patients with similar cohort characteristics, from which 15 have recover from the condition (success). The standard treatment has a probability of success of 50%. The following hypotheses is stated *The experimental treatment is superior to the standard treatment.*

The hypotheses can be translated as

$$H_1 : \theta > 0.5, \quad H_2 : \theta \leq 0.5$$

where θ is the probability of success of the experimental treatment. Not much is known about the experimental treatment and a uniform (flat; Beta(1, 1)) prior is used. The corresponding posterior is Beta(16, 6), see figure.


Inicial Beta(1 , 1), post.   Beta(16 , 6).
Densidad
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
Probabilidad

We have that *a priori* $P(H_2) = 0.5$ and *a posteriori* $P(H_2 \mid \mathbf{X}) = 0.01330185$.

$$H_1 : \theta > 0.5, \quad H_2 : \theta \leq 0.5$$

We may *we* (or *your*) conclude?

We have that *a priori* $P(H_2) = 0.5$ and *a posteriori* $P(H_2 \mid \mathbf{X}) = 0.01330185$.

$$H_1 : \theta > 0.5, \quad H_2 : \theta \leq 0.5$$

We may *we* (or *your*) conclude?

Radiocarbon Calibration, one det.

We have that $y \sim N(\mu(\theta), \sigma)$. Considering the errors in the calibration curve the model should be $y_j \sim N\left(\mu(\theta), \sqrt{\sigma(\theta)^2 + \sigma_j^2}\right)$.

Therefore the likelihood is

$$f(Data|\theta) = f(y|\theta)$$

And the posterior is $f(\theta|y) \propto f(\theta) f(y|\theta)$

$$Kf(\theta) \frac{1}{\sqrt{\sigma(\theta)^2 + \sigma^2}} \exp\left\{ \frac{(y - \mu(\theta))^2}{2\sigma^2} \right\},$$

Radiocarbon Calibration, one det.

We have that $y \sim N(\mu(\theta), \sigma)$. Considering the errors in the calibration curve the model should be $y_j \sim N\left(\mu(\theta), \sqrt{\sigma(\theta)^2 + \sigma_j^2}\right)$.
Therefore the likelihood is

$$f(Data|\theta) = f(y|\theta)$$

And the posterior is $f(\theta|y) \propto f(\theta)f(y|\theta)$

$$Kf(\theta)\frac{1}{\sqrt{\sigma(\theta)^2+\sigma^2}}\exp\left\{\frac{(y-\mu(\theta))^2}{2\sigma^2}\right\},$$

Radiocarbon Calibration, one det.

We have that $y \sim N(\mu(\theta), \sigma)$. Considering the errors in the calibration curve the model should be $y_j \sim N\left(\mu(\theta), \sqrt{\sigma(\theta)^2 + \sigma_j^2}\right)$.

Therefore the likelihood is

$$f(Data|\theta) = f(y|\theta)$$

And the posterior is $f(\theta|y) \propto f(\theta)f(y|\theta)$

$$Kf(\theta)\frac{1}{\sqrt{\sigma(\theta)^2 + \sigma^2}} \exp\left\{\frac{(y - \mu(\theta))^2}{2\sigma^2}\right\},$$

Solution to the Mexica pot problem

1. All radiocarbon dated corn kernels are associated to the same calendar date θ.
2. It is assumed that the pot was made "around" the same time as the corn was harvested.
3. Prior information on θ is provided by $f(\theta)$.

We have a series of radiocarbon determinations $y_1, y_2, \ldots, y_m$ with their standard errors $\sigma_1, \sigma_2, \ldots, \sigma_m$ corresponding to m corn kernels.

From point 2 above we have that $y_j \sim N(\mu(\theta), \sigma_j)$, and we also assume that these are independent, conditional on θ (and the standard errors). Considering the errors in the calibration curve the model should be $y_j \sim N\left(\mu(\theta), \sqrt{\sigma(\theta)^2 + \sigma_j^2}\right)$.

Solution to the Mexica pot problem

1. All radiocarbon dated corn kernels are associated to the same calendar date θ.
2. It is assumed that the pot was made "around" the same time as the corn was harvested.
3. Prior information on θ is provided by $f(\theta)$.

We have a series of radiocarbon determinations $y_1, y_2, \ldots, y_m$ with their standard errors $\sigma_1, \sigma_2, \ldots, \sigma_m$ corresponding to m corn kernels.

From point 2 above we have that $y_j \sim N(\mu(\theta), \sigma_j)$, and we also assume that these are independent, conditional on θ (and the standard errors). Considering the errors in the calibration curve the model should be $y_j \sim N\left(\mu(\theta), \sqrt{\sigma(\theta)^2 + \sigma_j^2}\right)$.

Solution to the Mexica pot problem

1. All radiocarbon dated corn kernels are associated to the same calendar date θ.
2. It is assumed that the pot was made “around” the same time as the corn was harvested.
3. Prior information on θ is provided by $f(\theta)$.

We have a series of radiocarbon determinations $y_1, y_2, \ldots, y_m$ with their standard errors $\sigma_1, \sigma_2, \ldots, \sigma_m$ corresponding to m corn kernels.

From point 2 above we have that $y_j \sim N(\mu(\theta), \sigma_j)$, and we also assume that these are independent, conditional on θ (and the standard errors). Considering the errors in the calibration curve the model should be $y_j \sim N\left(\mu(\theta), \sqrt{\sigma(\theta)^2 + \sigma_j^2}\right)$.

Solution to the Mexica pot problem

1. All radiocarbon dated corn kernels are associated to the same calendar date θ.
2. It is assumed that the pot was made “around” the same time as the corn was harvested.
3. Prior information on θ is provided by $f(\theta)$.

We have a series of radiocarbon determinations $y_1, y_2, \ldots, y_m$ with their standard errors $\sigma_1, \sigma_2, \ldots, \sigma_m$ corresponding to m corn kernels.

From point 2 above we have that $y_j \sim N(\mu(\theta), \sigma_j)$, and we also assume that these are independent, conditional on θ (and the standard errors). Considering the errors in the calibration curve the model should be $y_j \sim N\left(\mu(\theta), \sqrt{\sigma(\theta)^2 + \sigma_j^2}\right)$.

Solution to the Mexica pot problem

1. All radiocarbon dated corn kernels are associated to the same calendar date θ.
2. It is assumed that the pot was made "around" the same time as the corn was harvested.
3. Prior information on θ is provided by $f(\theta)$.

We have a series of radiocarbon determinations $y_1, y_2, \ldots, y_m$ with their standard errors $\sigma_1, \sigma_2, \ldots, \sigma_m$ corresponding to m corn kernels.

From point 2 above we have that $y_j \sim N(\mu(\theta), \sigma_j)$, and we also assume that these are independent, conditional on θ (and the standard errors). Considering the errors in the calibration curve the model should be $y_j \sim N\left(\mu(\theta), \sqrt{\sigma(\theta)^2 + \sigma_j^2}\right)$.

Solution to the Mexica pot problem

1. All radiocarbon dated corn kernels are associated to the same calendar date θ.
2. It is assumed that the pot was made "around" the same time as the corn was harvested.
3. Prior information on θ is provided by $f(\theta)$.

We have a series of radiocarbon determinations $y_1, y_2, \ldots, y_m$ with their standard errors $\sigma_1, \sigma_2, \ldots, \sigma_m$ corresponding to m corn kernels.

From point 2 above we have that $y_j \sim N(\mu(\theta), \sigma_j)$, and we also assume that these are independent, conditional on θ (and the standard errors). Considering the errors in the calibration curve the model should be $y_j \sim N\left(\mu(\theta), \sqrt{\sigma(\theta)^2 + \sigma_j^2}\right)$.

Solution to the Mexica pot problem

1. All radiocarbon dated corn kernels are associated to the same calendar date θ.
2. It is assumed that the pot was made "around" the same time as the corn was harvested.
3. Prior information on θ is provided by $f(\theta)$.

We have a series of radiocarbon determinations $y_1, y_2, \ldots, y_m$ with their standard errors $\sigma_1, \sigma_2, \ldots, \sigma_m$ corresponding to m corn kernels.

From point 2 above we have that $y_j \sim N(\mu(\theta), \sigma_j)$, and we also assume that these are independent, conditional on θ (and the standard errors). Considering the errors in the calibration curve the model should be $y_j \sim N\left(\mu(\theta), \sqrt{\sigma(\theta)^2 + \sigma_j^2}\right)$.

Solution to the Mexica pot problem

Therefore the likelihood is

$$f(Data|\theta) = f(y_1, y_2, \ldots, y_m|\theta) = \prod_{j=1}^{m} f(y_j|\theta)$$

And the posterior is $f(\theta|y_1, \ldots, y_m) \propto f(\theta) \prod_{j=1}^{m} f(y_j|\theta)$, or

$$f(\theta|y_1, \ldots, y_m) = Kf(\theta) \prod_{j=1}^{m} \frac{1}{\sqrt{\sigma(\theta)^2 + \sigma_j^2}} \exp\left\{ \frac{(y_j - \mu(\theta))^2}{2\sigma_j^2} \right\},$$

where K is a normalizing constant.

Solution to the Mexica pot problem

Therefore the likelihood is

$$f(Data|\theta) = f(y_1, y_2, \ldots, y_m|\theta) = \prod_{j=1}^{m} f(y_j|\theta)$$

And the posterior is $f(\theta|y_1, \ldots, y_m) \propto f(\theta) \prod_{j=1}^{m} f(y_j|\theta)$, or

$$f(\theta|y_1, \ldots, y_m) = K f(\theta) \prod_{j=1}^{m} \frac{1}{\sqrt{\sigma(\theta)^2 + \sigma_j^2}} \exp\left\{ \frac{(y_j - \mu(\theta))^2}{2\sigma_j^2} \right\},$$

where K is a normalizing constant.

Solution to the Mexica pot problem

Four radiocarbon dates are taken from 4 of the maize kernels. The obtained dates are:

sim1	340	20
sim2	370	20
sim3	355	20
sim4	360	20

The posterior distribution is calculated as above, see next slide, Figure (a).

However, knowledge of basic Mexican history tells us that the Mexica umpire fell to Conquistador Hernan Cortez in 1521 AD. Including such prior information we obtain the next slide, Figure (b).

Solution to the Mexica pot problem

Four radiocarbon dates are taken from 4 of the maize kernels. The obtained dates are:

sim1	340	20
sim2	370	20
sim3	355	20
sim4	360	20

The posterior distribution is calculated as above, see next slide, Figure (a).

However, knowledge of basic Mexican history tells us that the Mexica umpire fell to Conquistador Hernan Cortez in 1521 AD. Including such prior information we obtain the next slide, Figure (b).

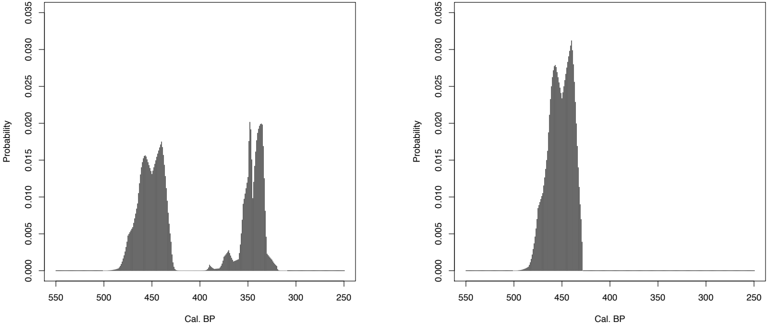
Solution to the Mexica pot problem

Four radiocarbon dates are taken from 4 of the maize kernels. The obtained dates are:

sim1	340	20
sim2	370	20
sim3	355	20
sim4	360	20

The posterior distribution is calculated as above, see next slide, Figure (a).

However, knowledge of basic Mexican history tells us that the Mexica umpire fell to Conquistador Hernan Cortez in 1521 AD. Including such prior information we obtain the next slide, Figure (b).



(a) (b)

Figure: Posterior distribution for the age of the maize kernels, (1) no prior (constant), (b) *a priori* distribution indicating $\theta \geq 429$ BP ($=$ 1521 AD).