Advanced Topics in Statistics

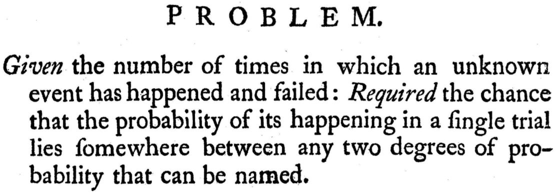
∼ Introduction to Bayesian Statistics ∼

1

UNIVERSITY OF EXETER

How did it all start?

In 1763, Reverend Thomas Bayes of Tunbridge Wells wrote



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 ſingle trial lies ſomewhere between any two degrees of probability that can be named.

What this means in modern language, is that given $r \sim Binomial(\theta, n)$, what is

$$P(\theta_1 < \theta < \theta_2 | r, n)?$$

Note that the above is a probability statement about the parameter, which is meaningless in 'classical' statistics that treats parameters as fixed quantities.

Bayes was the first statistician to use probability distributions to represent uncertainty about the model parameter, and thus think about the parameter as a random variable.

Two different interpretation of probability

Recall that there are two different approaches for defining the probability of an event:

	Relative frequency	**Subjective probability**
Probability	Limit of relative frequency	Degree of belief
Statistics	Frequentist approach	Bayesian approach
Unknown model parameters	Treated as fixed	Treated as random variables
Estimating parameters, expressing uncertainty	Point estimates, confidence intervals	Use distribution of RV to express uncertainty about unknown quantities

Note that the differences between the two statistical framework is a result of the two different ways the concept of probability can be interpreted.

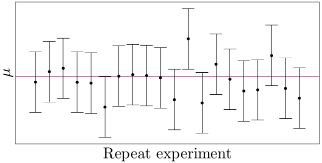
Estimating parameters - **Frequentist approach.**

One of the main goals of inferential statistics is estimating model parameters, e.g. mean, variance.

Frequentist approach:

- Find point estimates of the *fixed* parameter.
- To assess uncertainty we can construct (e.g. 95%) **confidence intervals**.

The confidence interval is not a probability statement about the model parameter (which is assumed to be fixed), but about the endpoints of the interval.



The endpoints of the CIs calculated from repeated sampling will vary, but 95% of these CIs will contain the true value of the parameter.

However, each calculated CI either contains the true model parameter or not (0/1 probability event).

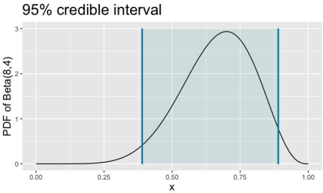
The frequentist approach relies on the notion of repeatability.

Estimating parameters - Bayesian approach I.

In **Bayesian statistics** we treat the model parameter as a random quantity, and as such it can be characterised by its probability distribution. Essentially it's the uncertainty in the inference that we quantify using probabilities.

Assigning probabilities to parameters, and thus using probability distributions has many advantages:

- Tells us what we want: what are the plausible values for the parameter of interest?
- Easy to get estimate of location (median/mean).
- No (difficult to interpret) confidence intervals: just report, say, central area that contains 95% of distribution. This is called a 95% **credible interval**: an interval within which the unobserved parameter value falls with 0.95 probability.



E.g. if $\theta \sim Beta(8,4)$, $P(0.39 < \theta < 0.89) = 0.95$.

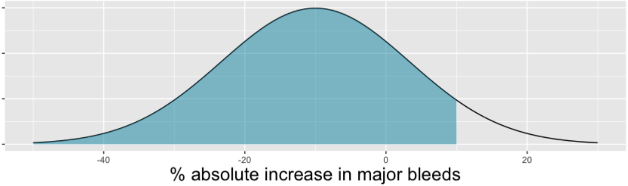
That is, we can give two fixed endpoints such that, the *subjective probability* that the unknown parameter lies between these two endpoints is 0.95.

Estimating parameters - Bayesian approach II.

Using direct probability distributions has many advantages:

- Relevant tail areas give probability of interest (no p-values).

  E.g. “There is an 89% probability that the absolute increase in major bleeds is less than 10 percent with low-dose PLT transfusions” (Tinmouth et al, Transfusion, 2004)



  Note that the above statement doesn't make sense in the frequentist framework, since there are no probabilities assigned to parameters, they are treated as fixed.

- Easy to make predictions (see later).
- There is a procedure for adapting the distribution in the light of additional evidence: i.e. **Bayes theorem** allows us to learn from experience.

Bayes' Theorem I.

Previously we've seen that Bayes' Theorem allows us to flip conditioning around. For two events A and B we had

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

In **Bayesian statistics** this formula takes the form

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{p(\text{data})},$$

- $p(\theta)$ is the prior information,
- $p(\text{data}|\theta)$ is the likelihood,
- $p(\theta|\text{data})$ is the posterior,
- $p(\text{data})$ is often called the evidence.

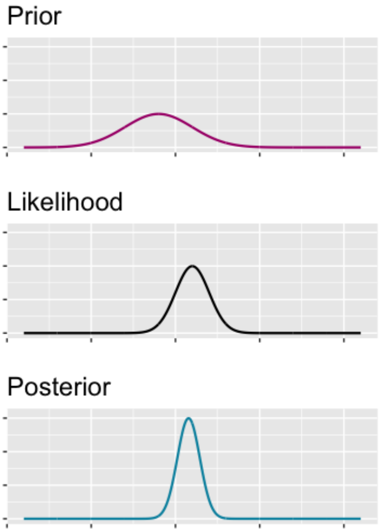
In this setting Bayes' Theorem gives us a tool to **update our beliefs in light of new evidence**.

Bayes' Theorem II.

The prior probability is our (subjective) beliefs about the parameter before the evidence (data) is taken into account.

Once data is collected, we use the likelihood to **update** our beliefs about the model parameter.

This updated belief is what we call the posterior distribution.



Prior distribution.

Reflects our uncertainty about the model parameter before seeing the data.

When choosing a prior we may consider:

- Information from previous studies.

  E.g. posterior from one problem (e.g. today's temperature) can become the prior for another problem (e.g. tomorrow's temperature).

- Expert opinion.

  This can take into account domain-specific knowledge, judgement, experience.

- Physical science.

Priors can also be used to impose constraints on variables (e.g. based on physical or assumed properties) and bound variables to plausible ranges.

The above are called 'informative' priors. They express specific, definite information about a variable.

But in certain cases we might want to minimise the introduced information. In these cases we can choose a so called vague prior.

PRIOR DISTRIBUTION.

When we want to minimise the introduced information we can use vague priors.

- Vague priors are vague with respect to the likelihood. It essentially means that prior mass is diffusively spread over the range of parameter values that are plausible, i.e. supported by the data (likelihood).
- Vague priors are often called 'non-informative' priors. However it's better to refer to them as 'vague', 'diffuse' or 'minimally informative' priors, since 'vague' never actually means completely uninformative.

  'There is no such thing as a 'noninformative' prior. Even improper priors give information: all possible values are equally likely' (Fisher, 1996)

- The simplest vague prior is one that assigns equal probabilities to all possible values. These are called flat priors.
- The uniform distribution (often with a wide range), or the normal distribution with large variance (say 10^6) are popular choices when a vague prior is required.

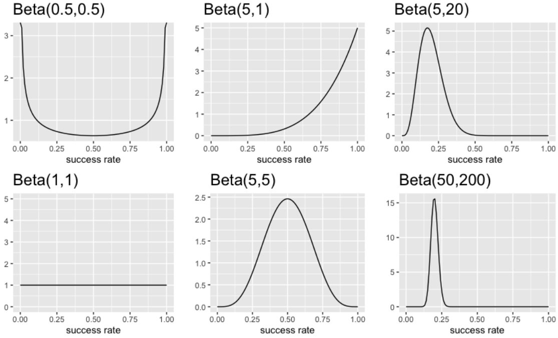
Overall, intuitively, the posterior is a compromise between the prior distribution and the likelihood. The smaller the sample size the larger weight the prior distribution gets.

Beta distribution I.

What is a reasonable form for a prior distribution for a proportion?

We would need a distribution that takes values on the interval $(0, 1)$, yet flexible enough, so it can be appropriate for a number of scenarios.

The Beta distribution satisfies both conditions. It is defined on the unit interval, and its probability density function can take various forms depending on the chosen parameters.



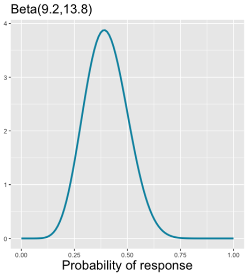
Beta distribution II.

For $\theta \sim Beta(a, b)$ we have

$$
\begin{aligned}
p(\theta|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1) \\
E(\theta|a,b) &= \frac{a}{a+b} \\
Var(\theta|a,b) &= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}
$$

Example - Drug.

- Consider a drug to be given for relief of chronic pain.
- Experience with similar compounds has suggested that annual response rates between 0.2 and 0.6 could be feasible.
- Interpret this as a distribution with mean = 0.4, standard deviation 0.1.
- A $Beta(9.2, 13.8)$ distribution has these properties.



Gamma distribution.

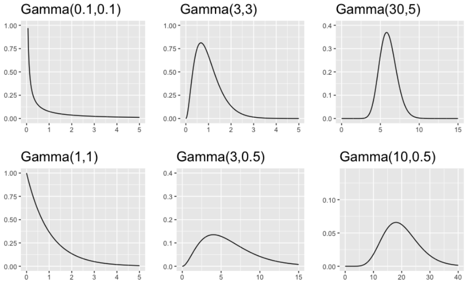
When the parameter can only take positive values, the gamma distribution is a popular choice as a prior due to its flexibility.

If $Y \sim Gamma(a, b)$ then

$$p(y|a,b) = \frac{b^a}{\Gamma(a)} y^{a-1} \mathrm{e}^{-by}$$

$$E(Y|a,b) = \frac{a}{b}$$

$$Var(Y|a,b) = \frac{a}{b^2}$$



- $Gamma(1, b)$ distribution is exponential with mean $1/b$.
- Used as conjugate prior distribution for inverse variances (precisions), the Poisson parameter, the exponential parameter, etc.
- Used as sampling distribution for skewed positive valued quantities (alternative to log normal likelihood).
- $Gamma(0.001, 0.001)$ is a vague prior for positive parameters (i.e. precisions).

PREDICTIVE DISTRIBUTION.

The next step would be to observe the data, and using the likelihood update our beliefs to get the posterior distribution.

However, there are many scenarios where it's useful to predict the data distribution before (or even after) seeing the data. These include:

- Design of studies.
- Assessing whether observed data is compatible with expectations.
- Can inform policy makers. E.g. cost-effectiveness models.

Before observing a quantity Y, we can provide its predictive distribution by integrating out the unknown parameter

$$p(Y) = \int p(Y|\theta)p(\theta)\mathrm{d}\theta.$$

Heuristically this provides a weighted average of the potential data models, where the weighting is given by the prior.

Predictive distribution - Example.

We have a coin that gives 'Head' with some unknown probability θ, and we want to predict the number of heads in n coin tosses, Y_n.

The number of heads follows a Binomial(θ, n) distribution.

The unknown parameter θ gives the proportion of heads in n tosses, therefore we can use the Beta distribution as a prior for θ.

Therefore, we have

$$\begin{aligned} \theta &\sim Beta(a, b), \\ Y_n &\sim Binomial(\theta, n). \end{aligned}$$

The exact predictive distribution in this case is known as the Beta-Binomial distribution, which has the form

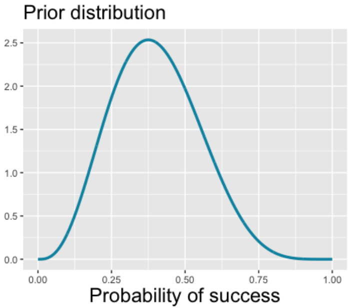
$$p(y_n) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \binom{n}{y_n} \frac{\Gamma(a+y_n)\Gamma(b+n-y_n)}{\Gamma(a+b+n)}, \quad y_n = 0, 1, \ldots, n.$$

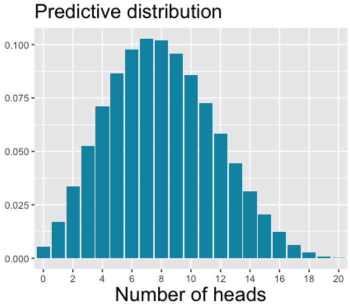
Predictive distribution - Example.

The previously seen predictive distribution can be derived by integrating the product of the Binomial and the Beta distributions:

$$p(y_n) = \int_0^1 \binom{n}{y_n} \theta^{y_n} (1-\theta)^{n-y_n} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} \mathrm{d}\theta.$$

The Beta-Binomial distribution is discrete with $E(Y_n) = n\frac{a}{a+b}$.





Note that if $a = b = 1$ (which is $Unif(0,1)$), then $p(y_n)$ is uniform over $0, 1, \ldots, n$.

POSTERIOR DISTRIBUTION.

The posterior distribution represents are beliefs about the parameter after having observed the data. It is a combination of the prior distribution and the likelihood function.

The posterior distribution can be used to get

- Point estimates by calculating the mean or median of the posterior.
- Interval estimates by computing credible intervals.
- Prediction for future data.
- Probability of different hypotheses.

In simple cases, e.g. where both prior and likelihood are conjugate (see later), exact expression for the posterior distribution can be found.

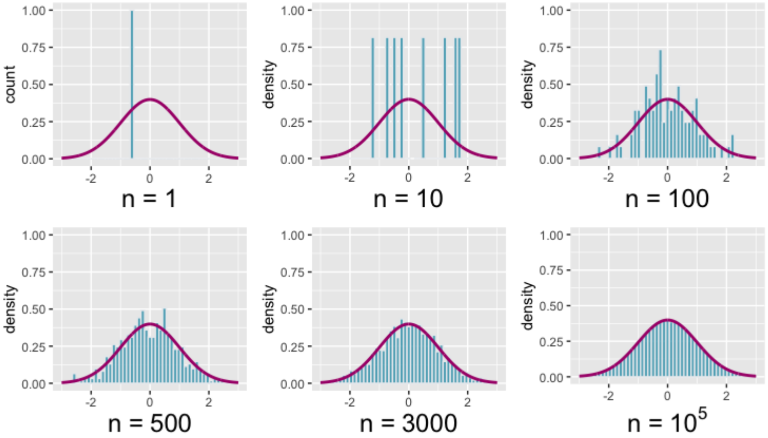
In more complex cases, the posterior might be intractable.

Can use simulation to build up posterior.

Simulation.

Since the posterior is often hard to compute, or even analytically intractable, we use simulation to approximate the posterior distribution.

We can use the simulated samples to build up the distribution:



n = 1

n = 10

n = 100

n = 500

n = 3000

n = 10^5

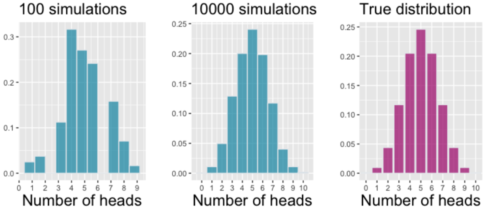
Forward sampling.

Suppose we want to know the probability of getting 8 or more heads when we toss a fair coin 10 times.

- Since this is a relatively simple question, we can use algebra to find the answer. For $X \sim Binom(0.5, 10)$ we have

$$
\begin{aligned}
P(\geq 8) = \sum_{k=8}^{10} p(x|\theta = 0.5, n = 10) =& \binom{10}{8}\left(\frac{1}{2}\right)^{8}\left(\frac{1}{2}\right)^{2} \\
&+ \binom{10}{9}\left(\frac{1}{2}\right)^{9}\left(\frac{1}{2}\right)^{1} + \binom{10}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^{0} \\
=& 0.0547
\end{aligned}
$$

- For more complicated problems, simulation can be the only feasible approach. Simulation uses the computer to repeatedly throw a set of 10 coins and count the proportion of throws with 8 or more heads.



100 simulations

10000 simulations

True distribution

Proportion with 8 or more 'heads':

After 100 tosses: 0.07

After 10000 tosses: 0.0543

True probability: 0.053

Forward sampling - General Monte Carlo analysis.

The previous method can be extended to answer a wide range of inferential questions.

In general:

- Suppose we have a logical function f containing uncertain parameters.
- We can express our uncertainty as a prior distribution.
- Simulate many values from this prior distribution.
- Calculate f at the simulated values ('iterations').
- Obtain an empirical predictive distribution for f.
- To get a point estimate for f, we can take the mean or median of this predictive distribution.
- Sometimes termed *probabilistic sensitivity analysis*.

Bayesian inference using Gibbs sampling.

The prior distribution is known. But how can we generate samples from a distribution we don't explicitly know?

- Samples from the posterior can be generated in several ways, without exact knowledge of $p(\theta|y)$.
- One method we can use is the so-called Gibbs sampling.
- A Gibbs sampler uses the conditional distribution of each parameter (conditional on all other parameters) to generate values from the posterior distribution.
- *The details of the Gibbs sampling algorithm will be discussed later.*

The Gibbs sampler provides the basis of the BUGS program, which is a language for specifying complex Bayesian models.

- BUGS stands for 'Bayesian inference Using Gibbs Sampling'.
- It's a software package for performing Bayesian inference (based on Gibbs sampling).
- It constructs an object-oriented internal representation of the model.
- BUGS is used in the following software: WinBUGS, OpenBUGS, JAGS.

- JAGS (Just another Gibbs Sampler) is a popular program for Bayesian inference, that can be called from R. (It can be installed like an `R` package).

Some aspects of the BUGS language.

Some of the most widely used commands/syntax of JAGS include

- `<-` represents logical dependence, e.g. `m <- a + b*x`
- `~` represents stochastic dependence, e.g. `r ~ dunif(a,b)`
- We can use arrays and loops

```
for(i in 1:n){
r[i] ~ dbin(p[i],n[i])
p[i] ~ dunif(0,1)
}
```

- Some functions can appear on the left-hand-side of an expression, e.g.

```
logit(p[i]) <- a + b*x[i]
log(m[i]) <- c + d*y[i]
```

- We can use the command `mean(p[])` to take mean of a whole array, `mean(p[m:n])` to take mean of elements `m` to `n`. Similarly for `sum(p[])`.
- We can restrict priors to the positive range: `dnorm(0,1)I(0,)`
- A wide range of R functions can also be used with JAGS.

When finding probabilities we make use of the command `ifelse(statement,1,0)`, which returns 1 if the statement is true, and returns 0 if it is false. An example of a tail probability: `ifelse(x>5,1,0)`.

Some common Distributions in JAGS.

The following table shows how the most common probability distributions are defined in JAGS.

Expression	Distribution	Usage
`dbin`	binomial	`r ~ dbin(p,n)`
`dnorm`	normal	`x ~ dnorm(mu,tau)`
`dpois`	Poisson	`r ~ dpois(lambda)`
`dunif`	uniform	`x ~ dunif(a,b)`
`dgamma`	gamma	`x ~ dgamma(a,b)`
`dbeta`	beta	`x ~ dbeta(a,b)`

The syntax is similar to R's own syntax, but there are some differences:

- The normal distribution is parameterised in terms of its mean and precision = 1/variance = $1/\text{sd}^2$.
- The order in which the parameters of the binomial distribution has to be listed is the opposite what R uses.
- Functions cannot be used as arguments in distributions (you need to create new nodes). E.g. $N(\alpha + \beta \cdot x, \tau)$ would have to be defined as

```
y ~ dnorm(mu,tau)
mu <- alpha + beta * x
```

Coin example with $P(\geq 8)$.

The model for this example is $Y \sim Binomial(0.5, 10)$, and we want to know $P(Y \geq 8)$.

This model is represented in JAGS as

```
jags.mod <- function(){
  Y ~ dbin(0.5,10)
  P8 <- ifelse(Y>7,1,0)
}
```

Which can be run using the following code:

```
jags.mod.fit <- jags(data = list(), model.file = jags.mod,
                     parameters.to.save = c("Y","P8"),
                     DIC=FALSE, n.iter = 100)
```

To get point estimates for $P(Y \geq 8)$ we have to look at the 'mu.vect' column of the outcome of `print(jags.mod.fit)`:

```
   mu.vect sd.vect 2.5% 25% 50% 75% 97.5%
P8   0.070   0.256    0   0   0   0     1
```

Running this simulation for 100 and 10000 iterations provided the previous estimated probabilities that Y will be 8 or more.

Note that the only reason we need 'DIC=FALSE' in the above code is that our model here does not contain a likelihood.

DRUG EXAMPLE.

Recall our drug example where a *Beta*(9.2, 13.8) distribution was suggested as a prior for the probability θ that a patient will respond to the drug.

Consider situation before giving 20 patients the treatment. What is the chance of getting 15 or more responders?

Here we have the following model

$$\begin{aligned} \theta &\sim Beta(9.2, 13.8) && \text{prior distribution} \\ y &\sim Binomial(\theta, 20) && \text{sampling distribution} \\ P_{crit} &= P(y \geq 15) && \text{Probability of exceeding critical threshold} \end{aligned}$$

In JAGS this translates to

```
jags.mod <- function(){
  theta ~ dbeta(9.2,13.8)
  y ~ dbin(theta,20)
  P.crit <- ifelse(y>=15,1,0)
}
```

Which we can run using

```
jags.mod.fit <- jags(data = list(), n.iter = 10000, DIC=FALSE,
                     parameters.to.save = c("theta","y","P.crit"),
                     model.file = jags.mod)
```

Drugs example - Output.

The `print(jags.mod.fit)` command gives the following output

```
       mu.vect sd.vect  2.5%  25%   50%    75%  97.5%
P.crit   0.015   0.122 0.000 0.00 0.000  0.000  0.000
theta    0.401   0.097 0.225 0.33 0.397  0.463  0.592
y        8.090   2.852 3.000 6.00 8.000 10.000 14.000
```

Note that the mean of the 0-1 indicator `P.crit` provides the estimated tail-area probability.

For comparison we can give the exact answers from closed-form analysis:

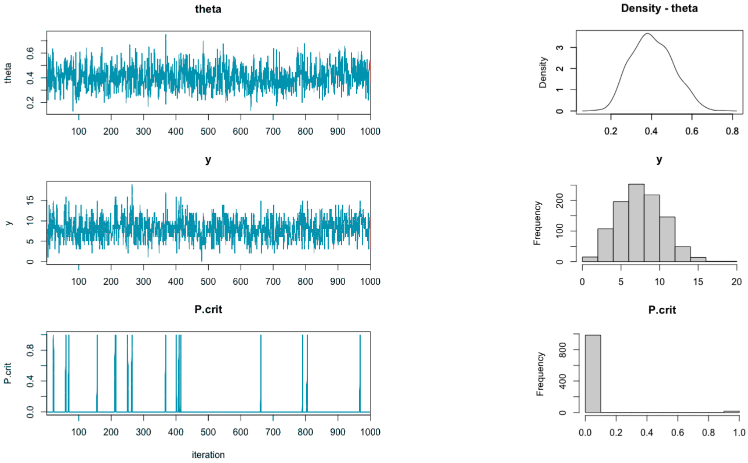
- θ: mean 0.4 and standard deviation 0.1
- y: mean 8 and standard deviation 2.93.
- Probability of at least 15: 0.015.

Can achieve arbitrary accuracy by running the simulation for longer.

These are independent samples, and so MC error $= SD/\sqrt{\text{Number of iterations}}$.

Drugs example - Output.

We can plot the simulated samples and the estimated density functions.



These are independent samples, and so there's auto-correlation and no concern with convergence (see details later).