Welcome to MTH6102: Bayesian Statistical Methods

Eftychia Solea

Queen Mary University of London

2023

Outline of the course

Lectures

- Three hours of lectures each week:
  - Lecture A: Wednesday, 9:00-11:00, G.O.Jones:LT.
  - Lecture B: Friday, 11:00-12:00, Bancroft: 2.40.

- Lectures are both on-campus and Q-reviewed.

- Each lecture will contain one or more of the following:
  - lecture slides;
  - hand-written examples;
  - demonstration of R code.

IT classes

- IT classes:
  - Thursday, 9:00am, Bancroft:1.23 (71) PC Lab
  - Thursday, 10:00am, Bancroft:1.23 (71) PC Lab
- You're meant to attend one of these two IT classes, starting this week.
- You will use the statistical software R.
- There will be an R practical to work through each IT class available in QMPlus the previous day.
- These will be a mix of individual questions and answers, plus me or the TA going over examples when there is a desire for this.
- Plus you can ask about non-R material.

TA

- **TA:** Maria Pintado Serrano
- **E-mail:** m.f.pintadoserrano@qmul.ac.uk.

Outline of the course

Weekly sequence

- The sequence each week is:
  - Lecture A: Wednesday, 9:00-11:00, G.O.Jones:LT.
  - Lecture B: Friday, 11:00-12:00, Bancroft:2.40.
  - IT class: Thursday (starting in week 1).
  - Office hours: Wednesday, 14:45-15:45pm, room MB-324 (starting in week 2).

- Plus an exercise sheet to work through each week for practice (not to be handed in) and you can ask about it in IT sessions.

Lecture notes and course website

- All course material can be found on QMPlus.
- Slides will be put on QMPlus before each lecture.
- There is a more formal set of notes (single pdf file) on QMPlus.
- This sometimes has more formal proofs and definitions than the lectures.
- Please, be sure to visit QMPlus early and often.

Discussion forum

- I encourage you to use the discussion forum.
- A great way to interact online and work together on the assignments.
- You can post any question you have about module's material
- Another student might comment or respond.
- I will respond after one student has commented on the question.
- Please post the question to the forum before emailing me.

Outline of the course

Assessment

We have the following assessment pattern:

- 20% coursework.
- 80% final exam.

Coursework

- There will be 5 sets of exercise sheet questions to be handed in.
- Each one counts for 4% of module total.
- Submit on QMPlus, every two weeks.
- First one, starting in week 2, is due to be handed in by start of week 4, on **Monday, 16h Oct at 11:00**, based on the first two weeks' material.
- Assignments will not be accepted past the time they are due.
- Assignment must be done individually.
- See Assessment Information section on QMPlus page.

Exam

- Entirely written, not computer-based, in January 2024.
- You are allowed to bring 3 pages of A4 notes.
- You could also bring a non-programmable calculator.
- **Past exam papers:** There are 4 past exam papers available on QMPlus, as this is the fifth year the module has run.

Reading list

- Bayesian Statistics: An Introduction, 2014 (4th ed.) by P M Lee.
- All of statistics, 2011 by Larry Wasserman (Chapters 11 and 24).
- Computational Bayesian Statistics, 2019 by M. Antonia Amaral Turkman, Carlos Daniel Paulino and Peter Mulller.
- Bayesian Data Analysis, 2013 (3rd edition) by A Gelman, J B Carlin, H S Stern and D B Rubin.
- A First Course in Bayesian Statistical Methods, 2009 by Hoff, Peter D.

Module overview

- This course is an introduction to Bayesian statistics
- There are two main approaches to statistical learning: frequentist statistics, or classical and Bayesian statistics.
- So far at Queen Mary, the statistics for inference and estimation has been in the frequentist, or classical.
- Bayesian statistics is an alternative approach-attempts to treat all statistical inference as probabilistic inference
- It has some advantages that we will mention later.
- It is also becoming more commonly used, especially for more complex modelling work.

What is a probability?

- Probability is a way to describe the likelihood of an uncertain event in advance, before we observe whether it happens or not.
- Let A be an event, then the probability that A will occur is written as $P(A)$

$$P(A) = \text{probability that } A \text{ will occur.}$$

- $P(A) = 1$ means that the event will definitely happen. $P(A) = 0$ means that the event will definitely not happen.
- What is the meaning of probabilities between 0 and 1?

First interpretation: Relative Frequency

- If the experiment can be **repeated** potentially infinitely many times, then the probability of an event can be defined through relative frequencies.

$P(A)$=the proportion of the time that A occurs in the long run, if the experiment is repeated under identical conditions.

- For example, suppose the experiment is to toss a coin, and suppose the event A is head.
- Then, $P(A) = 0.5$ means that, if we were to toss the coin a very large number of times, the proportion of tosses we observe heads tends to 0.5 or 50% as the number of tosses increases.

First interpretation: Relative Frequency

- Example. Using computer, we can simulate tosses of a fair coin.
- We obtain the following frequency table:

tosses	head	proportion
10	3	0.3
100	61	0.61
1000	481	0.481
10,000	4966	0.4966
100,000	50,022	0.5002
1,000,000	500,456	0.500

- As the number of tosses increases, the proportion of the time that head occurs gets closer and closer to 0.5
- This verifies the claim that $P(A) = 0.5$.

First interpretation: Relative Frequency

Comments on the Relative Frequency interpretation

- This is the most widely accepted interpretation of probability.
- It is regarded as **objective**, because answers can be verified (e.g., by computer simulation).
- This interpretation makes sense for experiments that can be repeated under similar conditions.
- It does not make as much sense for special or one-time situations that cannot be repeated.

Classical statistics or frequentist statistics use the Relative Frequency interpretation of probability. Probabilities are viewed as limiting relative frequencies.

Second interpretation: Personal or subjective probability

- The personal or subjective probability interpretation is

$P(A) =$ degree to which an individual believes that A is going to happen

- This value will obviously differ from one person to another.
- Individual's probabilities may differ because
  - they have varying amounts and kinds of knowledge
  - people are not equally good at assessing uncertainty.

Second interpretation: Personal or subjective probability

Example

- A report by the Environmental Protection Agency: "Global warming is most likely to raise sea level 15 cm by 2050 and 34 cm by 2100...There is a 1% chance that global warming will raise sea level 1 meter in the next 100 years."

- Senator John Kerry (March 17, 2006): "I can say to absolute certainty that if things stay exactly as they are today...within the next thirty years, the Arctic ice sheet is gone...If that melts, you have a level of sea level increase that wipes out Boston harbor, New York harbor."
- This does not refer to any limiting frequency. It reflects different strengths of beliefs about global warming.

Second interpretation: Personal or subjective probability

Comments on the Personal or Subjective Probability interpretation

- For evaluating the risks of rare or one-time events, this may be the only way
- Many subjective probabilities are simply an individual's statement of personal beliefs and biases

Bayesian Statistics uses the Personal or Subjective Probability interpretation as a degree of belief.

- Bayesian methods combine expert opinion and evidence to update subjective probabilities.

Frequentist statistics vs Bayesian statistics

These two different interpretations of probability lead to two schools of statistical inference:

- The frequentist statistics, and the
- The Bayesian statistics.

Frequentist statistics and inference

- This has been the mainstream of statistics for the past century. The frequentist point of view is based on the following
  - F1. Probability refers to limiting relative frequencies. Probabilities are objective frequencies.
  - F2. Parameters are fixed but unknown constants. Because they are not random, no useful probability statements can be made about parameters.
  - F3. Statistical procedures should be designed to have well-defined long-run frequency properties. This is based on the principle of (hypothetical) repeated sampling.

Interval example: Frequentist statistics and inference

- Suppose $X_1, \ldots, X_n$ i.i.d from $\mathcal{N}(\theta, 1)$. We wish to provide some sort of interval estimate C of θ.
- **Frequentist approach.** We construct the confidence interval

$$C = [\bar{X} - \frac{1.96}{\sqrt{n}}, \bar{X} + \frac{1.96}{\sqrt{n}}],$$

where $\bar{X}$ is the average of $X_1, \ldots, X_n$.

- Then,

$$P_\theta(\theta \in C) = 0.95 \quad \forall \theta \in \mathbb{R} \tag{1}$$

Example continued

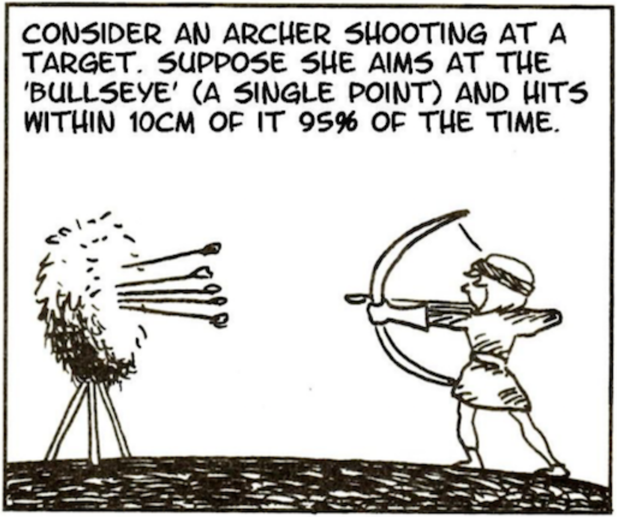
- In frequentist approach the statement (1) is about the random interval C which covers θ with probability 0.95, referred to as the confidence level.
- The interval is random because it is a function of the data.
- The parameter is a fixed, unknown quantity.
- The confidence level, 95%, is a property of the procedure over hypothetical repeated samples. It is not a property of any one specific confidence interval.
  - It is NOT correct to say: The true θ lies in C with probability 0.95.
  - A single 95% confidence interval does or does not cover the true value. You don't know whether it does or doesn't.

Example continued

- To make the meaning clearer, suppose we repeat the experiment many times
    - On day 1 you collect data from $\mathcal{N}(\theta, 1)$ and construct a [valid] 95% confidence interval C_1 for θ
    - On day 2 you collect data from $\mathcal{N}(\theta, 1)$ and construct a [valid] 95% confidence interval C_2 for θ

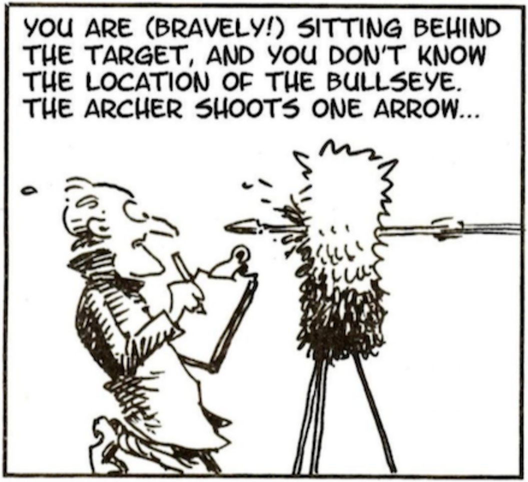
      $\vdots$
    - On day 100 you collect data from $\mathcal{N}(\theta, 1)$ and construct a [valid] 95% confidence interval C_{100} for θ
- In the long run, with repeated sampling, the intervals trap the parameter θ 95 percent of the time.

Archery example: Frequentist statistics and inference



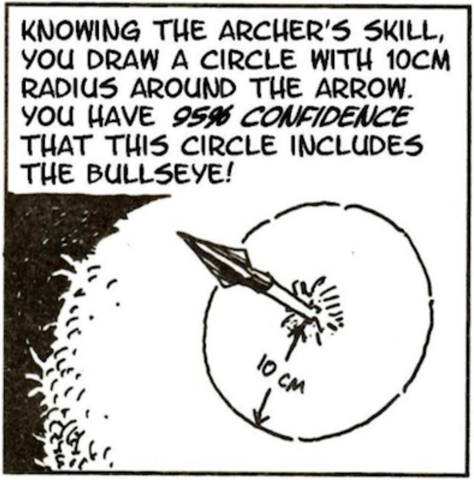
Adapted from Gonick & Smith, The Cartoon Guide to Statistics

Archery example: Frequentist statistics and inference

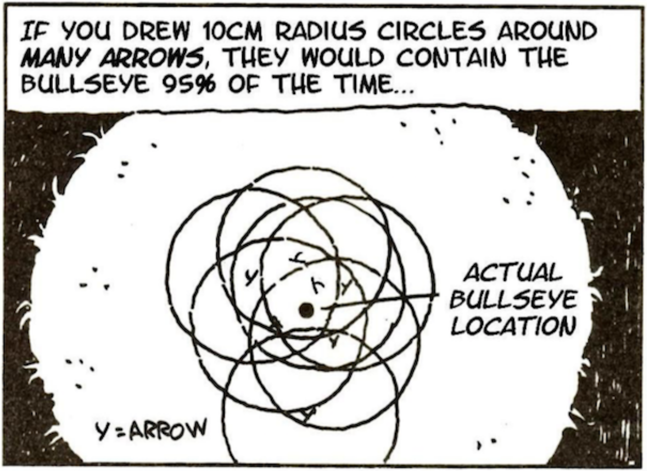


You want to learn where the bullseye is

Frequentist statistics and inference



Archery example: Frequentist statistics and inference



The circle traps the truth bullseye's location in 95% of shootings. To define anything frequentist, you have to imagine repeated experiments.

Classical or frequentist statistics: Long-run frequency interpretation

- Statistical procedures should be designed to have well-defined long-run frequency properties. This is based on the principle of (hypothetical) repeated sampling.

- **Long-run frequency interpretation.** If we could repeat the sampling procedure many times, we would get many intervals, and 95% of them would cover the true value.

- **Question:** How does Bayesian inference differ?

Bayesian approach

B1. Probability describes degree of belief, not limiting frequency. As such, we can make probability statements about lots of things, not just data which are subject to random variation.

B2. Parameters are viewed as unobserved random variables, for which we can make probability statements.

B3. We make inferences about a parameter θ by producing a probability distribution for θ. Inferences, such as intervals may be extracted from this distribution.

Bayesian approach

- **Bayesian approach.** In Bayesian statistics we express our beliefs and uncertainty about the unknown parameter θ using a probability distribution, $p(\theta)$, called the prior distribution.
- In frequentist statistics, we do not have a probability distribution for the parameters
  - only for data we observed, or might have observed.
- This is a characteristic of the Bayesian approach-all unknown parameters are treated as random variables and they are given a prior distribution.

Bayesian approach

- We then combine this with the observed data $X = x \in \mathbb{R}^n$, $X = (X_1, \ldots, X_n)$ to update our beliefs about the parameters.
- Using Baye's theorem

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}, \tag{2}$$

  where $p(x|\theta)$ is the likelihood viewed as conditional probability (conditional on the event $\{X = x\}$)
- The Baye's Theorem yields a distribution over θ-the posterior probability of θ given x, $p(\theta|x)$

Interval example: Bayesian approach

- Next, using the posterior one finds an interval C such that

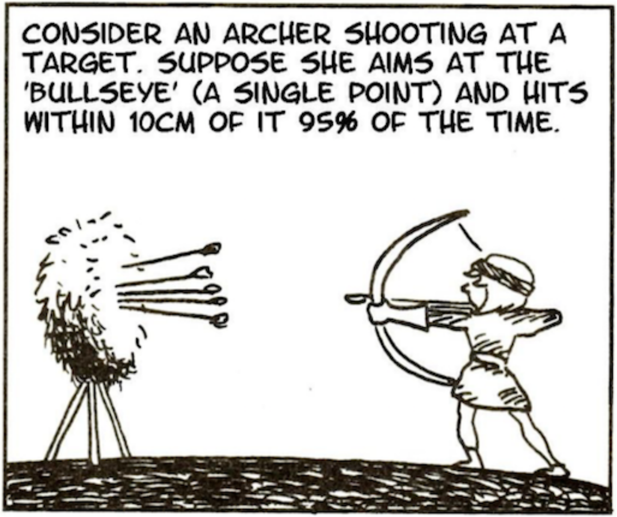
$$\int_C p(\theta|x)\, d\theta = 0.95$$

- We can report that

$$P(\theta \in C|x) = 0.95 \tag{3}$$

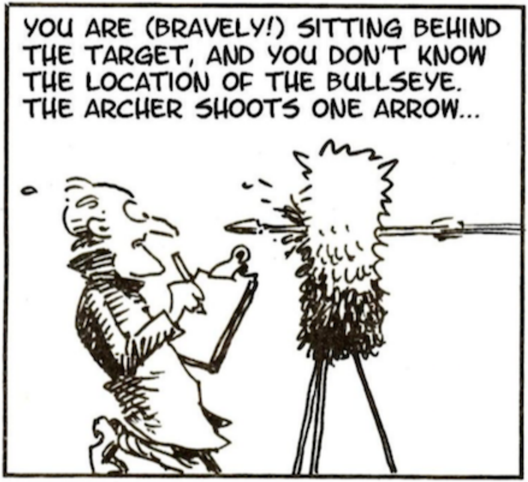
- This a degree-of-belief probability statement about θ given the data. It is not the same as (1).
- Bayesian methods allow to say the true θ lies in C with probability 0.95
- C is called a credible interval and has a direct interpretation in terms of probability.

Archery example: Bayesian approach



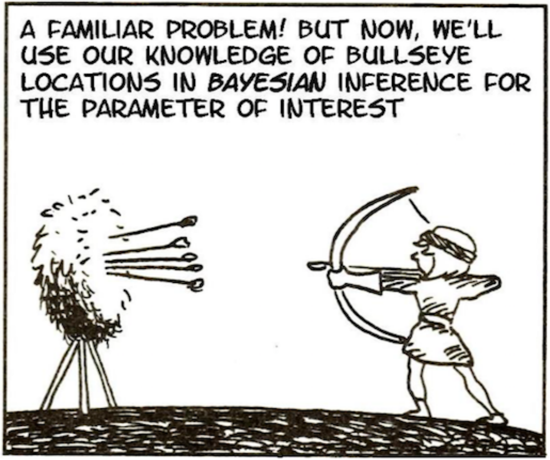
Adapted from Gonick & Smith, The Cartoon Guide to Statistics

Archery example: Frequentist statistics and inference



You want to learn where the bullseye is

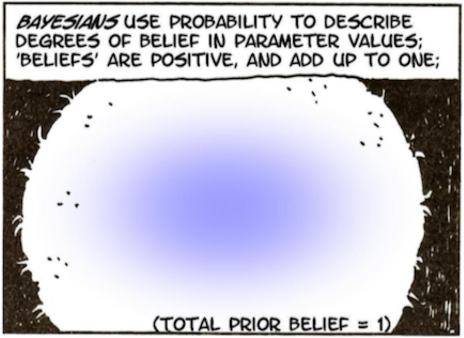
Archery example: Bayesian approach



The parameter of interest θ is the bullseye location.

Archery example: Bayesian approach

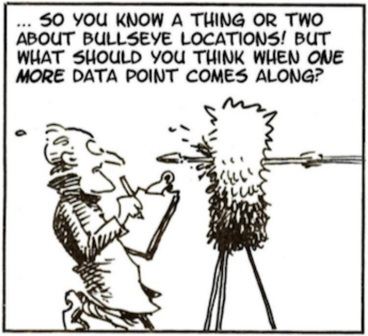
You don't know the location exactly, but do have some ideas.



blue shaded region is your prior $p(\theta)$ that describes your beliefs about the plausible values of θ

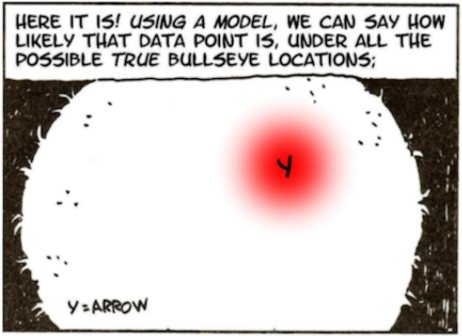
Archery example: Bayesian approach

You don't know the location exactly, but do have some ideas.



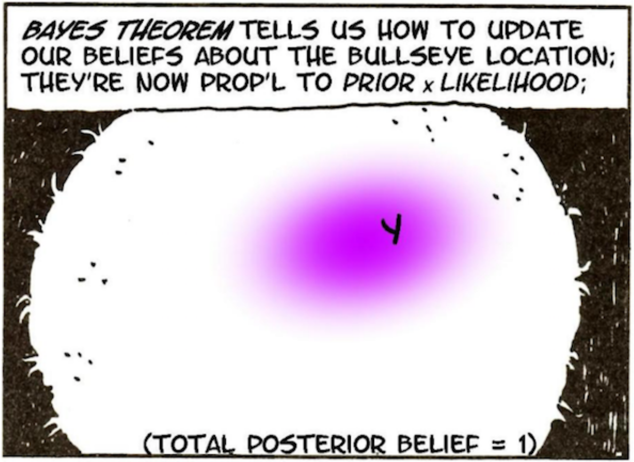
Archery example: Bayesian approach

What to do when the data comes along?



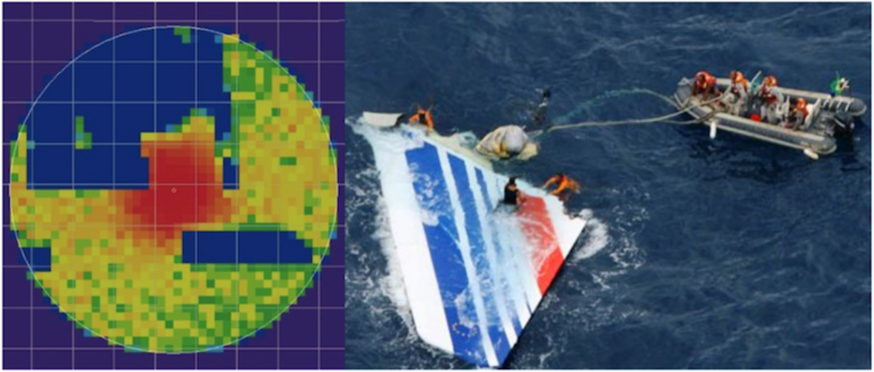
red shaded region is your likelihood, $p(\text{data}|\theta)$, where data is the arrow hit.

Archery example: Bayesian approach



purple region represents the posterior $p(\theta|\text{data})$: your updated beliefs about bullseye location, θ.

Bayesian approach in practice



- During the search for Air France 447, from 2009-2011, knowledge about the black box location was described using Bayesian inference.
- Eventually, the black box was found in the red area.
- For further, see Search for the Wreckage of Air France Flight AF 447

Bayesian and frequentist statistics

To summarise,

- Bayesian treats probability as beliefs, not frequencies.
- Bayesian inference is a method for stating and updating beliefs.
- A Bayesian perspective requires us to assign a prior probability to θ.
- Bayesian is subjective: Two different Bayesian statisticians may assign different priors to θ, and thus obtain different conclusions.

Bayesian and frequentist statistics

- Frequentist wishes to avoid such subjectivity.
- The goal of a frequentist approach is to develop an objective statistical theory, in which two statisticians employing the methodology must necessarily draw the same conclusions from a particular data set.
- The idea is to create procedures with long-run frequency guarantees.
- Important features of frequentist approach is the principle of repeated sampling and the frequentist interpretation of probabilities.

Bayesian and frequentist statistics

	Frequentist	Bayesian
Probability is:	limiting relative frequency	degree of belief, subjective
Parameter θ is	fixed constant	random variable
Probability statements are about:	procedures	parameters
Frequency guarantees	yes	no

Overview of this course

- This course will expose you to Bayesian statistical methods for inference and prediction. The emphasis is on methodologies with some theory and applications.

- In particular, we will study
  - Prior distributions; conjugate priors; non-informative priors.
  - Point estimates, credible intervals.
  - Markov chain Monte Carlo.
  - Model selection.
  - Predictive distributions.
  - Missing data; hierarchical models.

This week lectures

This week lectures will also cover:

- Likelihood.
- Maximum likelihood estimator (MLE).
- We start Bayesian inference next week.

MLE and likelihood

- Let Y be a random variable (discrete/continuous) with probability distribution $p(y|\theta)$.
- Let $Y_1, \ldots, Y_n$ be a sample from the population $p(y|\theta)$
- In frequentist statistic, the idea is to construct various estimators of θ, and choose the best estimator according to some criteria (bias, variance).
- A point estimator is any function $W(Y_1, \ldots, Y_n)$ of the sample.
- **Important:** An estimator is itself a random variable since a new experiment will produce new data to compute it.

MLE and likelihood

- There is one particular estimator that is widely used in frequentist statistics, namely the maximum likelihood estimator (MLE).
- This estimator is popular because if often yields natural estimators (sample mean and sample proportion) and has favourable asymptotic properties.
- To understand the MLE, we must understand the notion of likelihood from which it derives.
- The concept of likelihood is also needed for Bayesian statistics.

Formal definition

Definition 1 (The Likelihood function)

Let $p(y|\theta)$ denote the joint probability density (pdf) or probability mass function (pmf) of the sample $Y = (Y_1, \ldots, Y_n)$. Then, given that $Y = y$ is observed, the function of θ defined by

$$\mathcal{L}(\theta|y) = p(y|\theta)$$

is called the likelihood function

- If $Y = y$ is discrete, then $\mathcal{L}(y|\theta) = P_\theta(Y = y)$.
- We treat $p(y \mid \theta)$ as function of θ for fixed y, and provides the basis for maximum likelihood estimation.

Binomial model

- Suppose we toss a (biased) coin that has probability q of showing heads.
- We toss it n times.
- Then the number of heads X is binomially distributed

$$X \sim Bin(n, q)$$

- Suppose we observe k heads (i.e. $X = k$).
- What is the likelihood function?

Binomial likelihood

- The data is discrete - the number of heads.
- So the likelihood is the Binomial probability mass function.
- For a given value of q, the probability that $X = k$ is

$$p(k \mid q) = P(X = k) = \binom{n}{k} q^k (1-q)^{n-k}$$

- So if we observe k heads, the likelihood is

$$\mathcal{L}(q|k) = p(k \mid q) = \binom{n}{k} q^k (1-q)^{n-k}$$

Binomial likelihood

Likelihood is

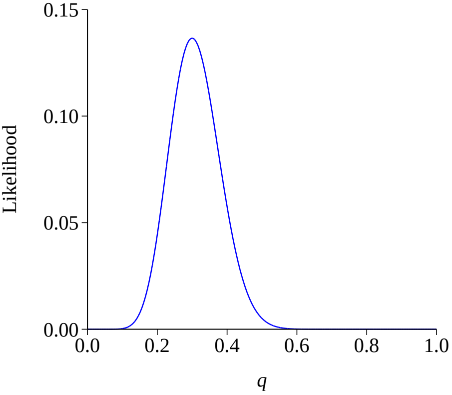
$$p(k \mid q) = \binom{n}{k} q^k (1-q)^{n-k}$$

Plotted as a function of q.

Here,

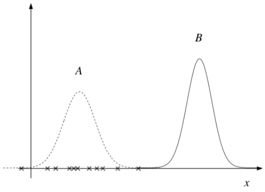
$$n = 40$$

$$k = 12$$



Maximum likelihood

- Suppose we want to estimate parameters θ.
- As suggested in Figure, the likelihood can be used to evaluate choices of θ.



- Density A assigns higher probability to the observed data x than density B, and thus would be preferred according to the principle of maximum likelihood.

Maximum likelihood Estimators

- **Idea:** Pick that value of θ that makes the observed sample x most probable

Definition 2: Maximum likelihood Estimators

For each sample point $y = (y_1, \ldots, y_n)$, find the value of θ which maximizes the likelihood function as a function of θ (with y held fixed)

$$\hat{\theta}_{\mathrm{ML}}(y) = \mathrm{argmax}_{\theta \in \Theta} \mathcal{L}(\theta | y),$$

where Θ is the range of the parameter θ. The estimator $\hat{\theta}_{\mathrm{ML}}(Y_1, \ldots, Y_n)$ based on the sample is known as the maximum likelihood estimator, or MLE.

Log-likelihood

- When finding the MLE, it is easier to work with the log of the likelihood.
- The log function is monotonically increasing.
- So the same θ will maximize $\mathcal{L}(\theta|y)$ and $\log \mathcal{L}(\theta|y)$.
- The log-likelihood is denoted by

$$\ell(\theta; y) = \log \mathcal{L}(\theta|y).$$

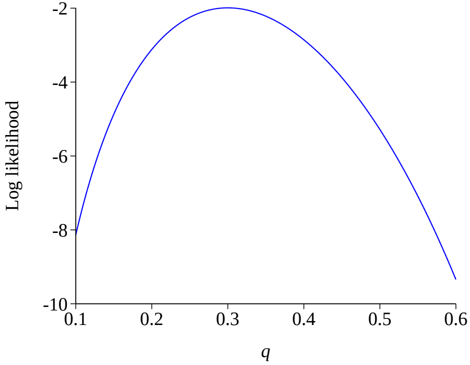
Binomial log-likelihood

$$\ell(q;k) = \log\binom{n}{k} + k\log(q) + (n-k)\log(1-q)$$

Plotted as a function of q.
Here,

$$n = 40$$

$$k = 12$$

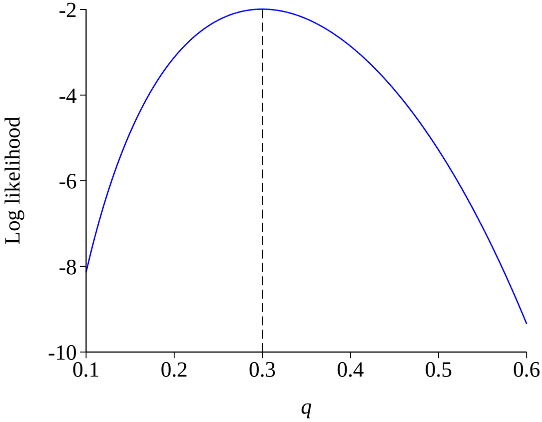


Binomial MLE

$n = 40$

$k = 12$

- If $0 < k < n$, differentiating $\ell(q; k)$ with respect to q and setting the result equal to 0, gives $\hat{q} = \frac{k}{n} = 0.3$
- Maximum likelihood estimator (MLE) is $\hat{q} = \frac{X}{n}$



Checking that it's a global maximum

- To check that we have found a maximum, we can calculate the second derivatives.
- For the binomial example, as $q \to 0$ or 1, $\ell \to -\infty$.
- So the stationary point must be a global maximum, and hence the MLE is $\hat{q} = \frac{X}{n}$.

Board example: Coins

A coin is taken from a box containing three coins, which gives heads with probability $q = 1/3$, $q = 1/2$, and $q = 2/3$. The mystery coin is tossed 80 times, resulting in 49 heads and 31 tails.

- What is the likelihood of this data for each type of coin? Which coin gives the maximum likelihood?
- Now suppose that we have a single coin with unknown probability q of landing heads. Find the likelihood and log likelihood functions given the same data. What is the MLE for q?

Work from scratch. Set the problem by defining random variables and pmf.

An example with continuous data

Example: Light bulbs

- The time until failure for a type of light bulb is exponentially distributed with parameter λ.
- We tested n bulbs and observe independently failure times $t = (t_1, \ldots, t_n)$.
- The unknown parameter is λ.
- Find the likelihood function and the log likelihood function
- Find the MLE for λ

Board example: Light bulbs

Suppose 5 bulbs are tested and have lifetimes of $2, 3, 1, 3, 4$ years, respectively.

- Find the MLE of λ.

Work from scratch. Set the problem by defining random variables and pmf.