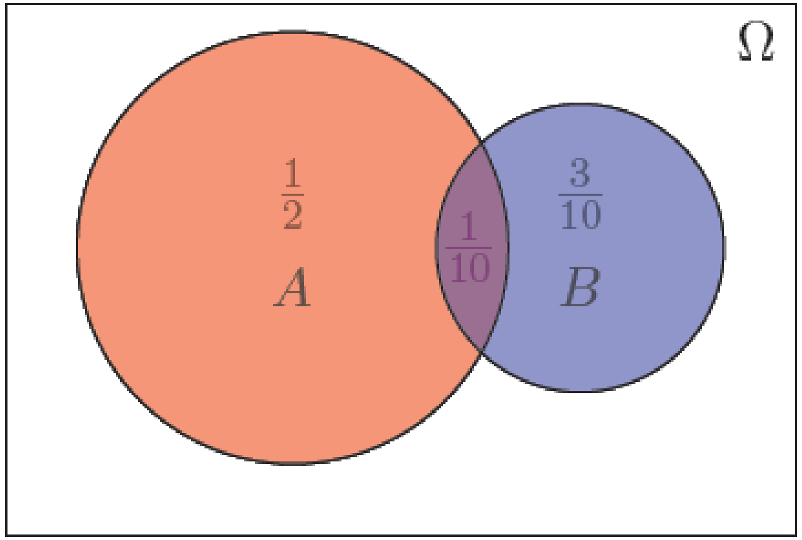
Week 12
Bayesian Inferences

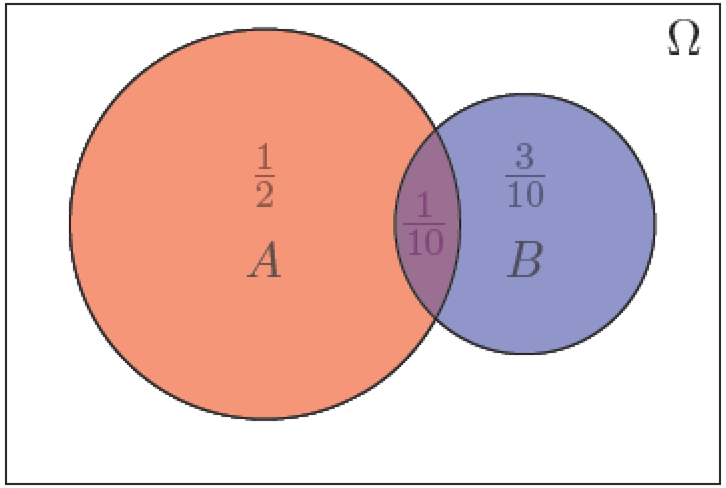
Revision - Probability



For two events A and B, calculate the following:

- $P(A) = \frac{3}{5}$
- $P(B) = \frac{2}{5}$
- $P(A|B) = \frac{1}{4}$
- $P(B|A) = \frac{1}{6}$

Revision – Bayes' Rule



$\text{P(A)} = \frac{1}{2} + \frac{1}{10} = \frac{3}{5}$

$\text{P(B)} = \frac{1}{10} + \frac{3}{10} = \frac{2}{5}$

$\text{P}(\text{A} \cap \text{B}) = \frac{1}{10}$

$\text{P(B|A)} = \frac{\text{P}(\text{A} \cap \text{B})}{\text{P(A)}} = \frac{\frac{1}{10}}{\frac{3}{5}} = \frac{1}{6}$

For 2 events A and B,

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

$= \frac{1}{6} * \frac{\frac{3}{5}}{\frac{2}{5}} = \frac{1}{4}$

Example – Medical Test

The probability of a certain medical test being positive (+) is 90% if a patient has a disease, D. 1% of the population have the disease, and the test records a false positive 5% of the time. If a patient has a positive test result, what is the probability the patient has the disease?

$P(+|D) = 0.9$

$P(D) = 0.01$

$P(+|D') = 0.05$

$P(D \cap +) = P(+|D)P(D) = 0.9(0.01) = 0.009$

$P(+) = P(D \cap +) + P(D' \cap +) = P(+|D)P(D) + P(+|D')P(D') = 0.009 + 0.05(1 - 0.01) = 0.0585$

$P(D|+) = \frac{P(D \cap +)}{P(+)} = \frac{0.009}{0.0585} = 0.1538$

Example – Further Medical Test

Suppose that the patient takes the same test again and the result is positive. What is the revised probability that the patient has the disease?

P(+|D) = 0.9

P(D) = 0.01

P(+|D') = 0.05

$P(D|+) = 0.153846 \overset{\text{update}}{\Rightarrow} P(D) = 0.153846$ and $P(D') = 1 - 0.153846 = 0.846154$

$P(D \cap +) = P(+|D)P(D) = 0.9(0.153846) = 0.1384614$

$P(+) = P(D \cap +) + P(D' \cap +) = P(+|D)P(D) + P(+|D')P(D') = 0.1385 + 0.05(0.846) = 0.1807691$

$P(D|+) = \frac{P(D \cap +)}{P(+)} = \frac{0.1384614}{0.1807691} = 0.76596$. This is the updated posterior probability of D.

Prior Distributions

The prior distribution is a description of the knowledge about the parameter in question prior to observation of the data.

There are different types of prior distributions which include:

- uninformed prior – you have no prior knowledge
- subjective or informed prior – incorporates information from an expert's opinion or your level of knowledge
- conjugate prior – the same family as the posterior
- improper prior – does not normalise to unity

Theory – Posterior Distribution

Assume that parameter(s) θ describe (part of) the distribution of the data . Then by Bayes' rule:

$$\boxed{P(\theta|) = \frac{P(|\theta)P(\theta)}{P()}}$$

where

- $P(\theta)$ is called the prior probability;
- $P(|\theta)$ is called the likelihood (or sampling distribution);
- $P(\theta|)$ is called the posterior probability;
- $P()$ is the normalising constant.

Note that $P()$ is a constant in the equation above so as a consequence, we may only need the prior distribution and the likelihood and use

$$P(\theta|) \propto P(|\theta)P(\theta).$$

The Bayesian Way

There are four steps involved:

1. assume a prior distribution of the parameter θ before analysing the new data set;
2. find an appropriate likelihood function for the observed data – $P(|\theta)$;
3. get the posterior distribution – $P(\theta|)$; and
4. make inference based on the posterior distribution.

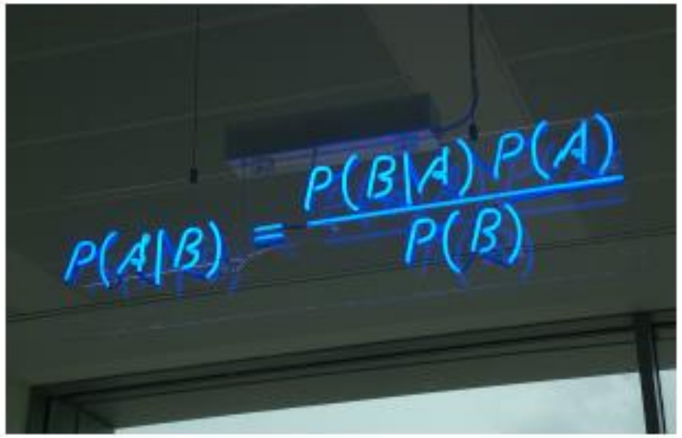


Image: Wikimedia Commons CC BY-SA 3.0

A bit of history

Ronald Fisher's Image: Wikimedia Commons CC BY 2.0

- So far the majority of what you have been learning in this course is the frequentist approach to statistical inference.
- Frequentist approach was popularised by Fisher, Neyman and Pearson in the early 20th century.
- Bayesian approach uses the rule/theorem by Thomas Bayes at its foundation.
- *Bayesian inference*, unlike Frequentist inference, makes distributional assumptions on the parameters.

Frequentist vs. Bayesian

Frequentist:

- In the frequentist (or classical) approach to statistics, probability is interpreted as long run frequencies (Lecture 2).
- The goal of frequentist inference is to create procedures (for example, methods of estimation) with long run guarantees.
- In frequentist inference, sampling processes are random while parameters are fixed, unknown quantities.

Bayesian:

- In the Bayesian approach, probability is regarded as a measure of subjective degree of belief.
- Bayesian statements are probability statements about the uncertainty of the parameters.
- The data are a given, the uncertainty on parameters can vary.

Maximum likelihood estimation

Suppose that the likelihood function depends on k parameters $\theta_1, \theta_2, \ldots, \theta_k$. Choose as estimates those values of the parameters that maximize the likelihood

$L(x_1, x_2, \ldots, x_n | \theta_1, \theta_2, \ldots, \theta_k)$.

Formally, $$\boxed{\hat{\theta}_{\text{ML}} = \arg\max_{\theta} \mathcal{L}(\theta|) = \arg\max_{\theta} f(|\theta).}$$

where θ is the parameter(s) and X is the data.

The maximum likelihood (ML) is a frequentist approach.

Example – maximum likelihood estimator

Given X “successes” in n trials, find the maximum likelihood estimate of the parameter p of the corresponding binomial distribution.

To find the value of p which maximizes L(p) = $\binom{n}{x} p^x (1-p)^{n-x}$, it will be convenient to make use of the fact that the value of p which maximizes L(p) will also maximize

lnL(p) = ln $\binom{n}{x}$ + xln(p) + (n – x)ln(1 – p)

Thus, we get

$$\frac{d[\ln L(p)]}{dp} = \frac{x}{p} - \frac{n - x}{1 - p}$$

And equating this derivative to 0 and solving for p, we find that the likelihood function has a maximum at p = $\frac{x}{n}$. This is the maximum likelihood estimate of the binomial parameter p, and we refer to $\hat{p} = \frac{x}{n}$ as the corresponding maximum likelihood estimator.

Example – maximum likelihood estimator

If $X_1, X_2, \ldots, X_n$ are the values of a random sample from an exponential population, find the maximum likelihood estimator of its parameter λ.

Since the likelihood function is given by

$L(\lambda) = f(X_1, X_2, \ldots, X_n; \lambda) = \prod_{i=1}^{n} \lambda\, e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum X_i}$,

Differentiation of $\ln L(\lambda) = n\ln(\lambda) - \lambda \sum X_i$ with respect to λ yields

$$\frac{d[\ln L(\lambda)]}{d\lambda} = \frac{n}{\lambda} - \sum X_i$$

Equating this derivative to 0 and solving for λ, we get the maximum likelihood estimate $\hat{\lambda} = \frac{n}{\sum X_i} = \frac{1}{\bar{X}}$. Hence, the maximum likelihood estimator is $\hat{\lambda} = \frac{1}{\bar{X}}$.

Special probability densities

Uniform distribution

If X ~ Unif(a, b), then $f(x) = \frac{1}{b-a}$; $a < X < b$

$E(X) = \frac{a+b}{2}$ and $Var(X) = \frac{(b-a)^2}{12}$

Beta distribution

If X ~ Beta(α, β), then $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} X^{\alpha-1}(1 - X)^{\beta-1}$; $0 < X < 1$, $\alpha > 0$ and $\beta > 0$.

$E(X) = \frac{\alpha}{\alpha+\beta}$ and $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Note that Beta(1, 1) = Unif(0, 1)

Maximum A Posteriori Estimate (MAP)

- Maximum a posteriori (MAP) estimate is the mode of the posterior distribution or more formally,

$$\boxed{\hat{\theta}_{\mathsf{MAP}} = \arg\max_{\theta} f(\theta|) = \arg\max_{\theta} f(|\theta) f(\theta).}$$

  where θ is the parameter(s) and X is the data, once again.
- Prior to sampling four fruits from the bag, I have no information so I may assume an uninformative prior, say $p \sim U(0,1)$ and so $f(p) = 1$ for $0 < p < 1$.
- Then the posterior density is given as

$$\begin{aligned} f(p|X) &= \frac{f(X|p)f(p)}{f(X)} = \frac{f(X|p)\cdot 1}{f(X)} \\ &= f(X|p) \end{aligned}$$

- In this case, the likelihood is equal to posterior density and so the MAP estimate is equal to the ML estimate.

Theorem 1

If X is a binomial random variable and the prior distribution of p is a beta distribution with the parameters α and β, then the posterior distribution of p|X=x is a beta distribution with the parameters x + α and n – x + β.

X ~ Bin(n, p)

$f(x|p) = \binom{n}{x} p^{x}(1-p)^{n-x}$; x = 0, 1, 2, …, n

$h(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$; $0 < p < 1$

$g(p|x) = \frac{f(x|p)h(p)}{f(x)} = \frac{\Gamma(\alpha+n+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} p^{x+\alpha-1}(1-p)^{n-x+\beta-1}$; $0 < p < 1$

Theorem 2

If $\overline{X}$ is the mean of a random sample of size n from a normal population with the known variance σ^2 and the prior distribution of μ is a normal distribution with the mean μ_0 and the variance σ_0^2, then the posterior distribution $\mu|\overline{X}=\bar{x}$ is a normal distribution with mean μ_1 and the variance σ_1^2, where

$$\mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} \text{ and } \sigma_1^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$f(\bar{x}|\mu) = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2\right]; \ -\infty < \bar{x} < \infty$$

$$h(\mu) = \frac{1}{\sigma_0\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]; \ -\infty < \mu < \infty$$

$$g(\mu|\bar{x}) = \frac{f(\bar{x}|\mu)h(\mu)}{f(\bar{x})} = \frac{1}{\sigma_1\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_1}{\sigma_1}\right)^2\right]; \ -\infty < \mu < \infty$$

Binomial likelihood with Beta prior

- Suppose the likelihood is modelled by $Bin(n, p)$, your prior distribution is $p \sim Beta(\alpha, \beta)$ and your observation is x successes out of n.
- Then the posterior distribution is a Beta distribution with parameters $\alpha + x$ and $\beta + n - x$.
- We call Beta distribution a conjugate prior for the Binomial likelihood function.
- **Formal definition**: if the posterior distribution is in the same family as the prior distribution then the prior and posterior are conjugate distributions and the prior is called a conjugate prior for the likelihood function.
- Here $\hat{p}_{\text{ML}} = \frac{x}{n}$ and $\hat{p}_{\text{MAP}} = \frac{x + \alpha - 1}{n + \alpha + \beta - 2}$.

Normal likelihood with normal prior

Normal likelihood with Normal prior

- Suppose that the parameter of your interest is μ which is estimated by the mean of the data $\bar{X}$ sampled n times from $N(\mu, \sigma^2)$ with known σ^2. Note: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.
- You assume a prior: $\mu \sim N(\mu_0, \sigma_0^2)$.
- Then the posterior distribution is given as

$$
\begin{aligned}
f(\mu|\bar{X} = \bar{x}) &\propto f(\bar{X}|\mu)f(\mu) \\
&\propto \exp\left(-\frac{1}{2\sigma^2/n}(\bar{x} - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \\
&\propto \exp\left(-\frac{1}{2 \cdot \kappa^2}(\mu - \tau)^2\right)
\end{aligned}
$$

where $\kappa^2 = \dfrac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}$ and $\tau = \dfrac{n\sigma_0^2\bar{x} + \sigma^2\mu_0}{\sigma^2 + n\sigma_0^2}$.

Summary

Parameter	Likelihood	Prior	ML Est.	MAP Estimate	Credible Interval
μ	$N\left(\mu, \frac{\sigma^2}{n}\right)$	$N(\mu_0, \sigma_0^2)$	$\bar{x}$	μ_1	$\mu_1 \pm Z_{1\text{-}\alpha/2}\sigma_1$
p	Bin(n, p)	Beta(α, β)	$\frac{x}{n}$	$\frac{x + \alpha - 1}{n + \alpha + \beta - 2}$	Out of scope

$$\mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} \text{ and } \sigma_1^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

Example

Kevin, a biology student, poses a statistical model for his scores on standard IQ tests. He thinks that, in general, his scores are normally distributed with unknown mean μ and variance of 80.

Expert opinion is that the IQ of biology students, μ, is a normal random variable, with mean 110 and variance 120.

Kevin took the test and scored 98.

$X \sim N(\mu, \sigma^2=80)$ and $\mu \sim N(\mu_0=110, \sigma_0^2=120)$

What is an estimate of Kevin's IQ?

$\hat{\mu} = \bar{X} = 98$ since $n = 1$ (The classical estimate of μ is given by the sample mean, which happens to be the ML estimate).

Example

Kevin, a biology student, poses a statistical model for his scores on standard IQ tests. He thinks that, in general, his scores are normally distributed with unknown mean μ and variance of 80.

Expert opinion is that the IQ of biology students, μ, is a normal random variable, with mean 110 and variance 120.

Kevin took the test and scored 98.

$X \sim N(\mu, \sigma^2=80)$ and $\mu \sim N(\mu_0=110, \sigma_0^2=120)$

What is an 95% interval estimate of μ?

$\bar{X} \pm Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 98 \pm 1.96\frac{\sqrt{80}}{\sqrt{1}} = 98 \pm 17.53 = [80.47, 115.53]$

$W = 115.53 - 80.47 = 35.06$

Example

Kevin, a biology student, poses a statistical model for his scores on standard IQ tests. He thinks that, in general, his scores are normally distributed with unknown mean μ and variance of 80.

Expert opinion is that the IQ of biology students, μ, is a normal random variable, with mean 110 and variance 120.

Kevin took the test and scored 98.

$X \sim N(\mu, \sigma^2=80)$ and $\mu \sim N(\mu_0=110, \sigma_0^2=120)$

What is the posterior distribution of $\mu|\bar{x}$?

$$\mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{1(98)(120) + 110(80)}{1(120) + 80} = 102.8 \text{ and } \sigma_1^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} = \frac{80(120)}{1(120) + 80} = 48$$

$\mu|\bar{x} \sim N(\mu_1=102.8, \sigma_1^2=48)$

Example

Kevin, a biology student, poses a statistical model for his scores on standard IQ tests. He thinks that, in general, his scores are normally distributed with unknown mean μ and variance of 80.

Expert opinion is that the IQ of biology students, μ, is a normal random variable, with mean 110 and variance 120.

Kevin took the test and scored 98.

$X \sim N(\mu, \sigma^2=80)$ and $\mu \sim N(\mu_0=110, \sigma_0^2=120)$

What is the MAP estimate of μ?

$$\hat{\mu}_{\mathrm{MAP}} = \mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{1(98)(120) + 110(80)}{1(120) + 80} = 102.8$$

Example

Kevin, a biology student, poses a statistical model for his scores on standard IQ tests. He thinks that, in general, his scores are normally distributed with unknown mean μ and variance of 80.

Expert opinion is that the IQ of biology students, μ, is a normal random variable, with mean 110 and variance 120.

Kevin took the test and scored 98.

$X \sim N(\mu, \sigma^2=80)$ and $\mu \sim N(\mu_0=110, \sigma_0^2=120)$

What is an 95% credible interval of μ?

$$\mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{1(98)(120) + 110(80)}{1(120) + 80} = 102.8 \text{ and } \sigma_1^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} = \frac{80(120)}{1(120) + 80} = 48$$

$\mu_1 \pm Z_{1-\alpha/2}\sigma_1 = 102.8 \pm 1.96\sqrt{48} = 102.8 \pm 13.579 = [89.221, 116.379]$

$W = 116.379 – 89.221 = 27.158$

The credible interval is shorter than the Z-confidence interval because the posterior variance is smaller than the likelihood variance; this is a consequence of the incorporation of information from the prior distribution.

Example

Kevin, a biology student, poses a statistical model for his scores on standard IQ tests. He thinks that, in general, his scores are normally distributed with unknown mean μ and variance of 80.

Expert opinion is that the IQ of biology students, μ, is a normal random variable, with mean 110 and variance 120.

Kevin took the test and scored 98.

$X \sim N(\mu, \sigma^2{=}80)$ and $\mu \sim N(\mu_0{=}110, \sigma_0^2{=}120)$

What happens to the $\hat{\mu}_{MAP}$ estimate when the prior variance increases indefinitely?

$$\hat{\mu}_{MAP} = \mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{\frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2}}{\frac{n\sigma_0^2 + \sigma^2}{n\sigma_0^2}} = \frac{\bar{x} + \frac{\mu_0\sigma^2}{n\sigma_0^2}}{1+\frac{\sigma^2}{n\sigma_0^2}} \to \bar{x} = \hat{\mu}_{ML} \text{ as } \sigma_0^2 \to \infty$$