Statistical Models and Inference - Part IV

Alberto Garfagnini

Università di Padova

AA 2020/2021 - Stat Lect. 9

Bayesian inference for Normal distribution

- many random variables seem to follow a normal distribution, at least approximately
- any random variable that is the sum of a large number of similar size random variables from independent causes ➜ approximately normal
- let's analyze a single observation from a conditional density $f(y \mid \mu)$ that is known to be normal with known variance σ^2
- we have a discrete set of possible k values for the mean, $\mu_1, \mu_2 \ldots \mu_k$
- thanks to Bayes' theorem

$$P\left(\mu \mid D, \sigma\right) = \frac{f\left(D \mid \mu, \sigma\right) g\left(\mu \mid \sigma\right)}{\int f\left(D \mid \mu, \sigma\right) g\left(\mu \mid \sigma\right) d\mu}$$

Single observation Likelihood

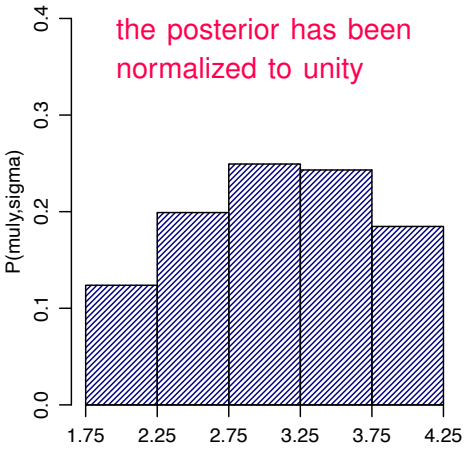
- the probability of the measurement of having a value y is

$$P\left(y \mid \mu, \sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right]$$

Example: single normal observation

The problem

- let's assume our variance is $\sigma^2 = 1$
- we know μ can have 5 possible values 2.0, 2.5, 3.0, 3.5 and 4.0
- a single observation is taken with $y = 3.2$



A Bayesian solution

- for the prior, we assume all values are equally possible
- we introduce a standardized variable $z = (y - \mu)/\sigma$
- let's report evaluated data in a table:

μ	$g(\mu\|\sigma)$ Prior	z	$f(y\|\mu,\sigma)$ Likelihood	$f \times g$	$P(\mu\|y,\sigma)$ Posterior
2.0	0.2	1.2	0.1942	0.03884	0.1238
2.5	0.2	0.7	0.3123	0.06245	0.1991
3.0	0.2	0.2	0.3910	0.07821	0.2493
3.5	0.2	-0.3	0.3814	0.07628	0.2431
4.0	0.2	-0.8	0.2897	0.05794	0.1847
				0.31372	1.0000

Estimating the mean of a Normal distribution

- given a set of N measurements, $D = \{y_j\}$, what is the best estimate of the parameter μ and how confident are we with the prediction ?
- let's assume that σ is known and is the same for all the measurements
- from Bayes' theorem

$$P\left(\mu \,\middle|\, D, \sigma\right) \propto P\left(D \,\middle|\, \mu, \sigma\right) \times P\left(\mu \,\middle|\, \sigma\right)$$

- we assume that data are independent, i.e. a measurement of one datum does not interfere on the outcome of another (given μ and σ)
- the Likelihood of the data is

$$P\left(D \,\middle|\, \mu, \sigma\right) = \prod_j P\left(y_j \,\middle|\, \mu, \sigma\right) = \prod_j \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(y_j - \mu)^2}{\sigma^2}\right]$$

- since the knowledge of the width of a Gaussian distribution does not tell us anything about the position of its centre, let us assume a uniform Prior pdf

$$P(\mu \,|\, \sigma) = P(\mu) = \begin{cases} \dfrac{1}{\mu_{max} - \mu_{min}} & \text{for } x \in [\mu_{min}, \mu_{max}] \\ 0 & \text{otherwise} \end{cases}$$

Estimating the mean of a Normal distribution

- let's combine Likelihood and Prior and write the natural logarithm of the posterior

$$L = \ln P\left(\mu \mid D, \sigma\right) = \text{const} - \sum_j \frac{(y_j - \mu)^2}{2\sigma^2}$$

- differentiating L and setting it to zero

$$\frac{dL}{d\mu} = \sum_j \frac{y_j - \mu}{\sigma^2} = 0 \Rightarrow \mu_\circ = \frac{1}{N}\sum_j y_j$$

- the reliability of the estimate is given by the second derivative

$$\left.\frac{d^2L}{d\mu^2}\right|_{\mu_\circ} = -\sum_j \frac{1}{\sigma^2} = \frac{N}{\sigma^2}$$

- therefore

$$\mu = \mu_\circ \pm \frac{\sigma}{\sqrt{N}}$$

Estimating the mean of a Normal distribution

- our estimate relies on the validity of a quadratic expansion of the natural logarithm of the Posterior around the maximum
- for the Gaussian distribution, this is an exact identity, because all higher derivatives of L are zero

- what happens when data have individual errors σ_j ?
- our single measurement Likelihood becomes

$$P\left(y_j \mid \mu, \sigma_j\right) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{1}{2}\frac{(y_j - \mu)^2}{\sigma_j^2}\right]$$

- and the logarithm of the Posterior

$$L = \ln P\left(\mu \mid \{y_j\}, \{\sigma_j\}\right) = \text{const} - \sum_j \frac{(y_j - \mu)^2}{2\sigma_j^2}$$

- taking the derivative of L and setting it to zero

$$\frac{dL}{d\mu} = \sum_j \frac{y_j - \mu}{\sigma_j^2} = 0 \Rightarrow \mu_\circ = \sum_j \frac{y_j}{\sigma_j^2} \Big/ \sum_j \frac{1}{\sigma_j^2}$$

- the reliability of the estimate is given by the second derivative

$$\left.\frac{d^2L}{d\mu^2}\right|_{\mu_\circ} = -\sum_j \frac{1}{\sigma_j^2} \quad \text{and, therefore} \quad \mu = \mu_\circ \pm \left(\sum_j 1/\sigma_j^2\right)^{-1/2}$$

Single observation with a Normal Prior

- let's assume our Prior has a Normal shape with mean m and variance s^2, $\texttt{Norm}(m, s^2)$

$$g(\mu \mid m, s) \propto \exp\left[-\frac{1}{2}\frac{(\mu - m)^2}{s^2}\right]$$

- the shape of the Likelihood is

$$f(y \mid \mu, \sigma) \propto \exp\left[-\frac{1}{2}\frac{(y - \mu)^2}{\sigma^2}\right]$$

- the product Likelihood $\times$ prior becomes

$$f(y \mid \mu, \sigma) \times g(\mu \mid m, s) \propto \exp -\frac{1}{2}\left[\frac{(y - \mu)^2}{\sigma^2} + \frac{(\mu - m)^2}{s^2}\right]$$

- with little algebra, it can be seen that the Posterior is a Normal distribution itself with mean and variance given by

$$m' = \frac{\sigma^2 m + s^2 y}{\sigma^2 + s^2} \quad \text{and} \quad (s')^2 = \frac{\sigma^2 s^2}{\sigma^2 + s^2}$$

➜ the $\texttt{Norm}(m, s^2)$ distribution is the conjugate family for the normal observation distribution (i.e. likelihood) with known variance

Updating rules for Normal inference with fixed variance

Single observation y

- the precision is the reciprocal of the variance, and we know form basics probability that precisions are additive. Therefore:

$$\frac{1}{(s')^2} = \frac{1}{s^2} + \frac{1}{\sigma^2} = \frac{\sigma^2 + s^2}{\sigma^2 s^2}$$

- the Posterior mean is given by

$$m' = \frac{\sigma^2 m + s^2 y}{\sigma^2 + s^2} = \frac{\sigma^2}{\sigma^2 + s^2} m + \frac{s^2}{\sigma^2 + s^2} y$$

- which can also be written as

$$m' = \frac{1/s^2}{1/\sigma^2 + 1/s^2} m + \frac{1/\sigma^2}{1/\sigma^2 + 1/s^2} y$$

Multiple observations $y_1, y_2 \ldots y_n$

- with the definition $\overline{y} = \frac{1}{N}\sum_j y_j$, it is possible to demonstrate that

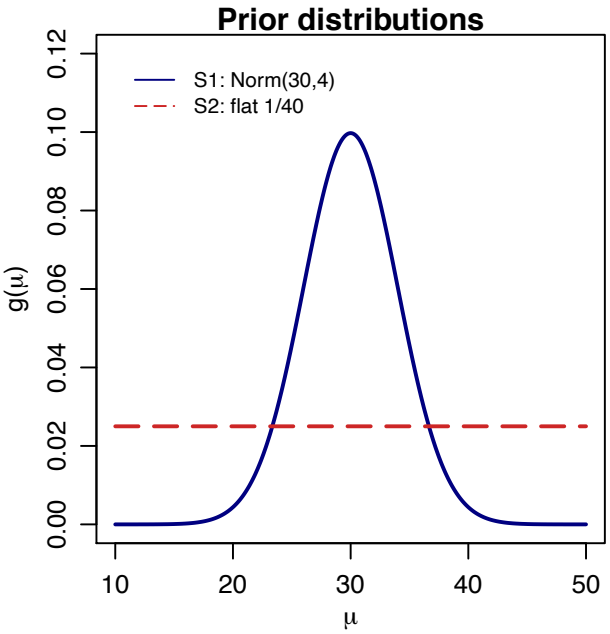
$$\frac{1}{(s')^2} = \frac{\sigma^2 + n s^2}{\sigma^2 s^2} \quad \text{and} \quad m' = \frac{1/s^2}{n/\sigma^2 + 1/s^2} m + \frac{n/\sigma^2}{n/\sigma^2 + 1/s^2} \overline{y}$$

Example : aquaculture

- two students have been asked to estimate the average length of a special fish in its first year of age leaving in a mountain lake
- previous studies in other lakes have shown that the length of the fish has a Normal distribution with known standard deviation $\sigma = 2$ cm

Assigning the Priors

1. Student 1 decides that her prior mean is $m = 30$ cm. Moreover she thinks that for such kind of fish in its first year it is not possible to have length below 18 cm or above 42 cm. Therefore her standard deviation is $s = 4$ cm
   ➜ her Prior is $\texttt{Norm}(30, 4^2)$
2. Student 2 does not know anything about this kind of fish, therefore he decides to assume a Uniform Prior



Example : aquaculture

- they take a random sample of 12 fish in their first year and they find that the sample mean is

$$\overline{y} = 32 \text{ cm}$$

Evaluating the Posteriors

1. Student 1, using the simple rule for the conjugate prior, gets:

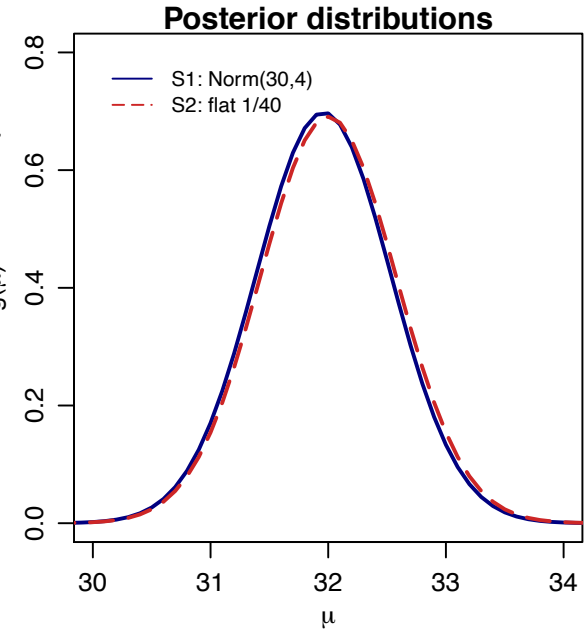
$$\frac{1}{(s')^2} = \frac{1}{4^2} + \frac{12}{2^2} \quad \text{which gives:} \quad s' = 0.5714$$

The Posterior mean is

$$m' = \frac{1/4^2}{1/0.3265} \times 30 + \frac{12/2^2}{1/0.3265} \times 32 = 31.96$$

2. Student 2 uses a flat prior and gets

$$(s')^2 = \frac{2^2}{12} = 0.3333 \quad \text{and} \quad m' = 32$$



Example: K/Ar rock dating methods

The problem

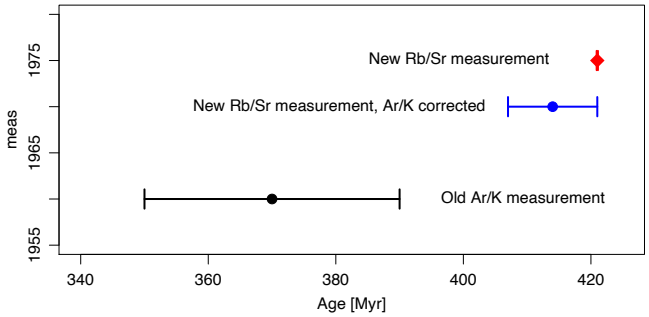
- K/Ar dating methods have been developed in the 60s for geochronology and archaeology research
- a measurement in the 60s of some rock samples gave an age $T_1 = 370 \pm 20$ Myr
- in the 70s, new methods based on the Rb/Sr method allowed to reach more precise measurements with a precision of $\sigma_2 = 8$ Myr and with a measurement result $T_2 = 421$ Myr

How to combine the measurements

- we assume that the measurements of the rocks with the K/Ar method gave $t_1 \sim \text{Norm}(\mu = 370, \sigma^2 = 20^2)$
- investigations with Rb/Sr will produce results of the type $t \sim \text{Norm}(\mu, 8^2)$ with well established precision
- the new prior is $\text{Norm}(m = 370, s = 8^2)$:

$$(s')^2 = \frac{\sigma^2 s^2}{\sigma^2 + s^2} = \frac{8^2 \cdot 20^2}{8^2 + 20^2} \sim 55 \sim 7^2$$

$$m' = \frac{\sigma^2 m + s^2 t_2}{\sigma^2 + s^2} = \frac{8^2 \cdot 370 + 20^2 \cdot 421}{8^2 + 20^2} = 414$$



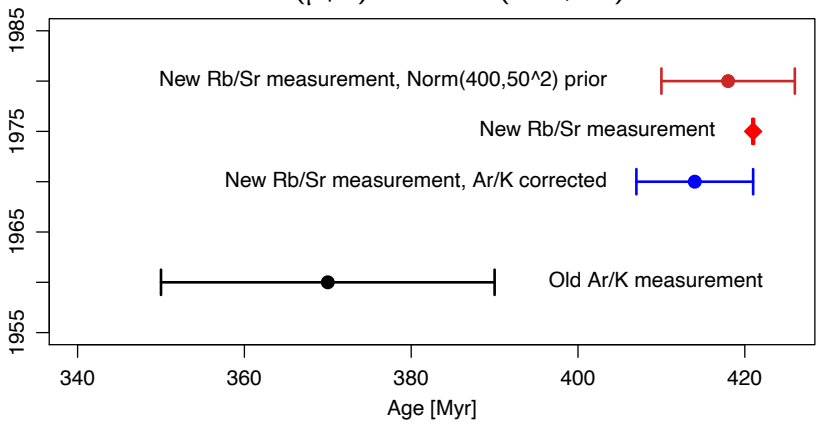
- the posterior for the age of the rock is $\text{Norm}(414, 7^2)$

Example: K/Ar rock dating methods (2)

Performing new measurements

- another scientist performs the same measurements but he is not aware of previous K/Ar dating results
- he considers a Normal prior with the assumption that the rock age is 400 ± 50 Myr
- the posterior variance is $(s')^2 = \frac{\sigma^2 s^2}{\sigma^2 + s^2} = \frac{1}{50^{-2} + 8^{-2}} \sim 62 \sim 8^2$
- and the posterior mean is $m' = \frac{\sigma^2 m + s^2 t_2}{\sigma^2 + s^2} = 62 \cdot \left(\frac{400}{50^2} + \frac{421}{8^2} \right) = 418$
- therefore the posterior is

$$P(\mu|T) = \text{Norm}(418, 8^2)$$



Bayesian Credible Interval for Normal μ

The variance is known

- $\{y_1 \dots y_n\}$ follows Norm(μ, σ^2)
- $\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$
- using either a flat prior or Norm(m, s^2) prior, the $(1-\alpha)\times 100\%$ credible interval for μ is:

$$m' \pm z_{\alpha/2} \times s'$$

with $z_{\alpha/2}$ the quantiles for a standardized normal distribution

The variance is unknown

- we evaluate the sample variance $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \mu)^2$
- the $(1-\alpha)\times 100\%$ credible interval for μ is:

$$m' \pm t_{\alpha/2} \times s'$$

with $t_{\alpha/2}$ the quantiles for a Student's t distribution with $n-1$ degrees of freedom

Predictive density for the next observation

- $y_1, \dots, y_n, y_{n+1}$ is a random sample from a Normal distribution with mean μ and known variance σ^2
- with bayesian statistics it is possible to write a conditional probability for the next random observation, given the actual random sample:

$$f\left(y_{n+1} | y_1 \dots y_n\right)$$

- the question is how to combine the uncertainty from the measured sample with that in the observation distribution

➜ by writing Bayes theorem and using the likelihood and prior distribution

- a Theorem of probability theory says that if

$$X \sim \text{Norm}(\mu_X, \sigma_X^2) \quad \text{and} \quad Y \sim \text{Norm}(\mu_Y, \sigma_Y^2)$$

➜

$$Z = X + Y \sim \text{Norm}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

- therefore, writing
  $y_{n+1} = y_{n+1} - \mu + \mu$
  $y_{n+1} - \mu \sim \text{Norm}(0, \sigma^2)$
  $\mu \sim \text{Norm}(m, s^2)$
- we get:

$$y_{n+1} \sim \text{Norm}(m,\ \sigma^2 + s^2)$$

Confidence Interval versus Credibility Interval

- we perform inference about the population mean when we have a random sample from a normally distributed population

Frequentist Confidence Interval

$$P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{y} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$\overline{y}$ is the sample mean and follows a Norm$(\mu, \sigma^2/n)$ distribution

we can re-write it as

$$P\left(\overline{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

- μ is a fixed but unknown parameter
- $(1-\alpha)\times 100\%$ of the intervals so computed will contain the true value
- by taking the random sample, and computing $\overline{y}$ there is nothing random left to attach a probability
- the computed interval either contains the true value or it does not

Confidence Interval versus Credibility Interval

Bayesian Credibility Interval

- using a flat prior for μ, the posterior mean is $m' = \overline{y}$
- the posterior variance is $(s')^2 = \sigma^2/n$

the interval is

$$P\left(\overline{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

the same form of the frequentist C.I., but with different interpretation:

- μ is a random variable ➜ probability statements are allowed
- the Credibility Interval is computed from the posterior distribution, given the observed sample
- the Credibility Interval contains a conditional probability of containing μ, given the data
- we are not concerned with a repetition of the experiment giving all possible data sets ➜ the only data set that matters is the one that occurred

Frequentist one-side HT for Normal μ

1 - setup the null hypothesis and alternative hypotheses

$$H_\circ : \mu \leq \mu_\circ \quad \text{versus} \quad H_1 : \mu > \mu_\circ$$

2 - the null distribution for $\overline{y}$ is Norm$(\mu_\circ, \sigma^2/n)$
The null distribution of the standardized variable $z \sim$ Norm$(0,1)$:

$$z = \frac{\overline{y} - \mu_\circ}{\sigma/\sqrt{n}}$$

3 - choose a level of significance α

4 - determine the rejection region. This is the region that has probability α when the NULL hypothesis is true. For $\alpha = 0.05$ ➜ the rejection region is $z > 1.645$

5 - take the sample and compute $\overline{y}$. If the value falls in the rejection region, we reject the hypothesis at level of significance α, otherwise we do not reject the NULL hypothesis

6 - or we can compute the P-value, which is the probability of observing what we observed, or something more extreme, given the NULL hypothesis:

$$P_{value} = P\left(z \geq \frac{\overline{y} - \mu_\circ}{\sigma/\sqrt{n}}\right)$$

if $P_{value} \leq \alpha$, we reject the NULL hypothesis, otherwise we cannot reject it

Bayesian one-side HT for Normal μ

1 - the posterior distribution

$$g(\mu | y_1 \ldots y_n)$$

summarizes our entire belief about the parameter, after having seen the data

2 - we setup the two hypotheses:

$$H_\circ : \mu \leq \mu_\circ \quad \text{versus} \quad H_1 : \mu > \mu_\circ$$

3 - we choose a level of significance α

4 - testing a one-side hypothesis is done by computing the following:

$$P(H_\circ : \mu \leq \mu_\circ | y_1 \ldots y_n) = \int_{-\infty}^{\mu_\circ} g\left(\mu \,\middle|\, y_1 \ldots y_n\right)\, d\mu$$

when the Posterior is Norm$(m', (s')^2)$ the computation is straightforward:

$$P(H_\circ : \mu \leq \mu_\circ | y_1 \ldots y_n) = P\left(Z \leq \frac{\mu_\circ - m'}{s'}\right)$$

5 - if the probability is less than α, we reject the NULL hypothesis and we can conclude that $\mu > \mu_\circ$

Example: One-Side Hypothesis Test (F)

The problem

- we wish to estimate the length of one-year old mountain trouts in a mountain river
- from measurements performed in the previous years, we know that the average length is $\mu_\circ = 31$ cm
- we want to test if the mean length is greater than that value, i.e.

$$H_\circ : \mu \leq 31 \text{ cm} \quad \text{versus} \quad H_1 : \mu > 31 \text{ cm}$$

with $\alpha = 0.05$

Frequentist approach

- the researchers measure $n = 12$ fish samples and measure $\overline{y} = 32$ cm
  - we build the normalized variable

$$z = \frac{\overline{y} - 31}{\sigma/\sqrt{n}} = \frac{32 - 31}{2/\sqrt{12}} = 1.732$$

  - we compute the P_{value}:

$$P_{value} = P\left(z > \frac{32 - 31}{2/\sqrt{12}}\right) = P(z > 1/732) = 0.04163678$$

  - the value is less that the level of significance, so the NULL hypothesis is rejected

Example: One-Side Hypothesis Test (B)

Bayesian approach

- the researchers measure $n = 12$ fish samples and compute $\overline{y} = 32$ cm
  - building the normalized variable

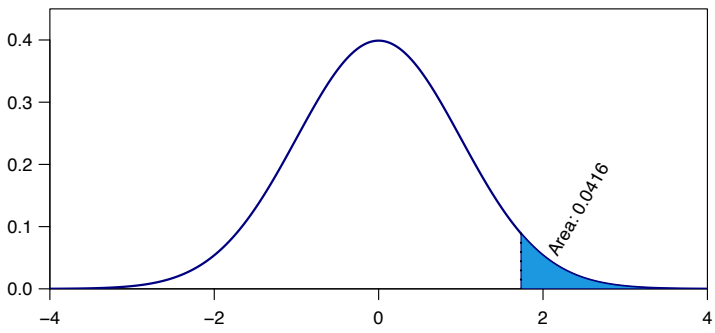
$$z = \frac{\overline{y} - 31}{\sigma/\sqrt{n}} = \frac{32 - 31}{2/\sqrt{12}} = 1.732$$

  - we compute the P_{value}:

$$P_{value} = P\left(z > \frac{32 - 31}{2/\sqrt{12}}\right) = P(z > 1/732) = 0.0416$$

  - the value is less that the level of significance, so the NULL hypothesis is rejected

```
y_ave <- 32
sigma <- 2
n <- 12
z <- (y_ave - 31)/(sigma/sqrt(n))
pnorm(z, lower.tail=FALSE)
[1] 0.04163226
```



Comparing μ of two Normal distributions

equal known variance

- samples and priors are independent ➜ the posteriors are independent

$$P\left(\mu_1 \,\middle|\, y_1 \dots y_{n1}\right) = \text{Norm}(m_1', (s_1')^2)$$

$$P\left(\mu_2 \,\middle|\, z_1 \dots z_{n2}\right) = \text{Norm}(m_2', (s_2')^2)$$

  with m_1', s_1' and m_2', s_2' determined by Normal μ inferences

- since both samples are independent, we can easily build the posterior distribution for $\mu_d = \mu_1 - \mu_2$:

$$P\left(\mu_d \,\middle|\, y_1 \dots y_{n1},\ z_1 \dots z_{n2}\right) = \text{Norm}(m_d', (s_d')^2)$$

- where $m_d' = m_1' - m_2'$ and $(s_d')^2 = (s_1')^2 + (s_2')^2$
- the $(1-\alpha)\times 100\%$ bayesian credible interval for $\mu_d = \mu_1 - \mu_2$ is:

$$m_d' \pm z_{\alpha/2} \times s_d'$$

  and can be written as:

$$m_1' - m_2' \pm z_{\alpha/2} \times \sqrt{(s_1')^2 + (s_2')^2}$$

Example: measuring the speed of light

The problem

- Michelson made two series of measurements in 1879 and 1882, respectively

Michelson (1879)				Michelson (1882)			
850	740	900	1070	883	816	778	796
930	850	950	980	682	711	611	599
980	880	1000	980	1051	781	578	796
930	650	760	810	774	820	772	696
1000	1000	960	960	573	748	748	797
				851	809	723	

- let's suppose the measurements are normally distributed with a known standard deviation, $\sigma = 100$
- we use independent priors, Norm$(m = 3 \cdot 10^5, s^2 = 500^2)$
- we compute the posteriors for μ_{1879} and μ_{1882} using the conjugate prior formulas
- we get:

$$\text{for} \quad \mu_{1879} : m_{1879}' = 299909 \text{ and } \quad (s1879')^2 = 499$$

$$\text{for} \quad \mu_{1882} : m_{1882}' = 299757 \text{ and } \quad (s1882')^2 = 434$$

Example: measuring the speed of light (2)

- the posterior distribution for $\mu_d = \mu_{1879} - \mu_{1882}$ will be Normal$(m'_d, (s'_d)^2)$, with

$$m'_d = 2999909 - 299757 = 152$$

and

$$(s'_d)^2 = 499 + 434 = 933 \sim 30.5^2$$

- the 95% Bayesian credible interval for $\mu_d = \mu_{1879} - \mu_{1882}$ is:

$$152 \pm 1.96 \times 30.5 = (92.1, 211.9)$$

- we perform an hypothesis test on the difference:

$$H_\circ : \mu_d \leq 0 \quad \text{versus} \quad H_1 : \mu_d > 0$$

- we compute the posterior probability of the null hypothesis $P(\mu_d < 0 \mid data)$:

$$\begin{aligned} P\left(\mu_d < 0 \mid data\right) &= P\left(\frac{\mu_d - m'_d}{s'_d} \leq \frac{0 - m'_d}{s'_d}\right) \\ &= P\left(z \leq \frac{0 - m'_d}{s'_d}\right) \end{aligned}$$

- since 0 lies outside the Bayesian credible interval, we reject the null hypothesis
  ➜ we conclude that the two sets of measurements are different

Comparing μ of two Normal distributions

variance unknown and flat priors are used

- we use independent flat priors for μ_1 and μ_2
- we get $m'_1 = \overline{y}$, $s'_1 = \sigma/\sqrt{n_1}$ and $m'_2 = \overline{z}$, $s'_2 = \sigma/\sqrt{n_2}$ and

- since we do not know the variance, we have to estimate it from the data:

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{n_1}(y_j - \overline{y})^2 + \sum_{k=1}^{n_2}(z_k - \overline{z})^2}{n_1 + n_2 - 2}$$

- since we used an estimate of the unknown true variance, the credible interval should be widened to allow for the additional uncertainty
- the $(1-\alpha) \times 100\%$ Bayesian credible interval for $\mu_1 - \mu_2$ is:

$$\overline{y} - \overline{z} \pm t_{\alpha/2} \times \hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $t_{\alpha/2}$ comes from a Sudent's t distribution with $n_1 + n_2 - 1$ degrees of freedom