

Bayes: 45
analysis
of the
IID $N(\mu, \sigma^2)$
sampling
model

$$\underline{\theta} = (\mu, \sigma^2)$$

Ans 132

p = 2

28 Feb 17

①

$$(\mu | \mathcal{B}) \sim p(\mu | \mathcal{B})$$

$$(I_i | \mu | \mathcal{B}) \stackrel{\text{IID}}{\sim} N(\mu, \sigma^2)$$

$$(z^i = 1, \dots, n)$$

what should we take for the prior?

Does a conjugate prior exist?

Yes,
but

it's a
bit
complicated

let's sneak up on the full complexity
with $\underline{\theta} = (\mu, \sigma^2)$ by temporarily
pretending that σ is known.

The new
model is

$$(\mu | \mathcal{B}) \sim p(\mu | \mathcal{B})$$

$$(I_i | \mu | \mathcal{B}) \stackrel{\text{IID}}{\sim} N(\mu, \sigma^2)$$

$$(z^i = 1, \dots, n) \xrightarrow{\text{Known}} \text{function - it's}$$

we've already worked
out the likelihood

$$f(\mu | \gamma^B) = c\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (\gamma_i - \mu)^2\right]. \quad (2)$$

But σ is known now, so it can be absorbed

into the constant: $f(\mu | \gamma^B) = c \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (\gamma_i - \mu)^2\right]$

As Bayesians, we need to think of the right-hand-side as a density in μ to identify the conjugate prior (if any), so let's expand out the $(\gamma_i - \mu)^2$ term & see what happens:

$$-\frac{1}{2\sigma^2} \sum_{i=1}^n (\gamma_i - \mu)^2 = -\frac{1}{2\sigma^2} \sum_{i=1}^n (\gamma_i^2 - 2\gamma_i \mu + \mu^2)$$

$$= \boxed{-\frac{1}{2\sigma^2} \sum_{i=1}^n \gamma_i^2} + \frac{\mu}{\sigma^2} \left(\sum_{i=1}^n \gamma_i \right) - \frac{n\mu^2}{2\sigma^2} = (*)$$

constant in μ

Now you can readily identify $\bar{\gamma} = \frac{1}{n} \sum_{i=1}^n \gamma_i$ as the maximum-likelihood estimator of μ

in this model, which means that we would (3)
 be well advised to make \bar{y} appear in
 expression (8) above. This is easy to do:

$$\frac{\mu}{\sigma^2} \sum_{i=1}^n y_i = \frac{n\mu}{\sigma^2} \left(\frac{1}{n} \sum_{i=1}^n y_i \right) = \frac{n\bar{y}\mu}{\sigma^2}.$$

$$\int_0^\infty -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 = C + \frac{n\bar{y}\mu}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}$$

$$= C - \frac{n}{2\sigma^2} (\mu^2 - 2\bar{y}).$$

Remembering that

$$f(\mu | y, \beta) = c \exp[-\frac{1}{2\sigma^2} (\mu^2 - 2\bar{y})],$$

a familiar object
 is emerging: the Normal density for μ

looks like $c_1 \exp[-c_2(\mu - c_3)^2]$, and

that's just what we have here if we

complete the square in μ :]
$$-\frac{1}{2\sigma^2} \sum_{i=1}^n (\gamma_i - \mu)^2 \quad (4)$$

$$= C - \frac{h}{2\sigma^2} (\mu^2 - 2\bar{\gamma}) = C - \frac{h}{2\sigma^2} (\mu^2 - 2\bar{\gamma} + \bar{\gamma}^2 - \bar{\gamma}^2)$$

$$= C - \frac{h}{2\sigma^2} ((\mu - \bar{\gamma})^2 - \bar{\gamma}^2)$$

but this is
also constant
in μ , so
finally

$$= C - \frac{h}{2\sigma^2} (\mu - \bar{\gamma})^2 + \boxed{\frac{h\bar{\gamma}^2}{2\sigma^2}}$$

$$L(\mu | \gamma, B) = \exp \left[-\frac{h}{2\sigma^2} (\mu - \bar{\gamma})^2 \right]$$

$$= C \exp \left[-\frac{1}{2\left(\frac{\sigma^2}{n}\right)} (\mu - \bar{\gamma})^2 \right] = N \left(\bar{\gamma}, \frac{\sigma^2}{n} \right)$$

Thus the likelihood density in this simpler sampling model - ^{IID} Gaussian, with known σ - is itself Gaussian with mean $\bar{\gamma}$ & s.d. $\frac{\sigma}{\sqrt{n}}$.

(This makes excellent sense from the frequentist point of view: $(\bar{Y} | \mu, \sigma^2) \sim N(\mu, \frac{\sigma^2}{n})$)

so it's no surprise that

$$(\bar{Y} | \bar{y}, \sigma^2) \sim N(\bar{y}, \frac{\sigma^2}{n})$$

Bayesian

frequentist
repeated
sampling

has incidentally

also shown that \bar{y} is sufficient for μ
in the IID $N(\mu, \sigma^2)$ sampling model
(again, no surprise).

ok, the likelihood

is Gaussian (as a density for μ); if
a conjugate prior exists & follows the
pattern of all of our other conjugate-
prior calculations, it would also have to
be Gaussian.

and this would work if the product of two Gaussian densities for μ is another Gaussian density for μ . we don't ^{yet} know if

this is true, but let's optimistically ^{that} pretend it is and see how it works out.

Suppose we take $\mu \sim N(\mu_0, \sigma_0^2)$

as our prior for μ in this simplified model.

Then, by Bayes's theorem,

$$\text{(posterior)} = \text{(posterior)} = c \cdot (\text{prior}) \cdot (\text{likelihood})$$

$$p(\mu | \bar{y}, \mathcal{B}) = p(\mu | \bar{y}, \mathcal{B}) = c p(\mu | \mathcal{B}) \cdot l(\mu | \bar{y}, \mathcal{B})$$

$$= c \exp\left[-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right] \cdot \exp\left[-\frac{1}{2\left(\frac{\sigma^2}{n}\right)}(\mu - \bar{y})^2\right]$$

$$= c \exp\left[-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 - \frac{1}{2\left(\frac{\sigma^2}{n}\right)}(\mu - \bar{y})^2\right].$$

Expand out both $(\mu - \mu_0)^2$ and $(\mu - \bar{\gamma})^2$ terms, collect everything into a quadratic in μ that looks like $A\mu^2 + B\mu + C$,

and complete the square in μ to get

an expression of the form $-\mathcal{D}(\mu - E)^2 + F$

(trust me, all of this works) the result
(it's just ugly algebra)

will be $p(\mu | \mathcal{B}) = c \exp[-\mathcal{D}(\mu - E)^2 + F]$

$$= c \exp[-\mathcal{D}(\mu - E)^2] e^F$$

F ← so e^F is a constant depending on μ

absorbed into c . Thus $p(\mu | \mathcal{B}) = c \exp[-\mathcal{D}(\mu - E)^2]$

i.e., in this simplified model in which,

$$(m | \mathcal{B}) \sim N(\mu_0, \sigma_0^2)$$

$$(I_i | \mu, \mathcal{B}) \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$$

$$(i=1, \dots, n) \quad \text{known}$$

if the prior for μ is Gaussian then the likelihood & posterior

are also Gaussian, i.e., there is indeed a conjugate prior for μ in this model & it's Gaussian.

When all the algebraic dust settles, you get the following

Math Fact

$$\left. \begin{array}{l} (\mu | \mathcal{B}) \sim N(\mu_0, \sigma_0^2) \\ (\bar{\gamma}_i | \mu, \mathcal{B}) \stackrel{\text{IID}}{\sim} N(\mu, \sigma^2) \\ (i=1, \dots, n) \end{array} \right\} \quad \left. \begin{array}{l} \bar{\gamma} = \frac{1}{n} \sum_{i=1}^n \bar{\gamma}_i \text{ is sufficient} \\ \text{for } \mu \text{ and} \\ (\mu | \bar{\gamma}, \mathcal{B}) = (\mu | \bar{\gamma}, \mathcal{B}) \\ = N(\mu_*, \sigma_*^2) \end{array} \right.$$

in which $\mu_* = \frac{\left(\frac{1}{\sigma_0^2}\right)\mu_0 + \left(\frac{n}{\sigma^2}\right)\bar{\gamma}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$

$$= \frac{\left(\frac{\sigma^2}{\sigma_0^2}\right)\mu_0 + n\bar{\gamma}}{\sigma^2/\sigma_0^2 + n} = \frac{n\mu_0 + n\bar{\gamma}}{n + 1}$$

$$\mu_0 \triangleq \frac{\sigma^2}{\sigma_0^2}$$

(9)

and $\frac{1}{\sigma_*^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$, from which

$$\sigma_*^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} = \frac{\sigma^2}{\frac{\sigma^2}{\sigma_0^2} + n} = \frac{\sigma^2}{n_0 + n}.$$

This requires a fair amount of unpacking, which fortunately turns out to be highly intuitive, as follows. [① In

the frequentist approach we're accustomed to characterizing the variability of a Gaussian(Normal) distribution, in terms of its variance σ^2 or its SD σ ;

if we write $\theta_i \sim N(\mu, \sigma^2)$, σ^2 and σ ⁽¹⁰⁾ represent the uncertainty about θ_i .

before observing it. From an information

perspective this is like describing how little we know about θ_i (pessimistic).

Bayesians have found it useful to define an optimistic quantity related to σ

and σ^2 that describes how much we know

about θ_i : Definition If a random variable

θ has a distribution with variance σ^2 ,

people say that the precision of this

distribution is $\frac{1}{\sigma^2}$ (i.e. precision = reciprocal variance)

Let's apply this idea to the IID $\overset{\text{known}}{N(\mu, \sigma^2)}$
Sampling model.

Prior:

$$(\mu | \mathcal{B}) \sim N(\mu_0, \frac{1}{\sigma_0^2})$$

so the prior mean (of μ) is μ_0 and the
prior precision is $\frac{1}{\sigma_0^2}$

Likelihood:

$$\ell(\mu | \bar{y}, \mathcal{B}) = N(\bar{y}, \frac{\sigma^2}{n})$$

so the likelihood mean is \bar{y} and the
likelihood precision is $\frac{n}{\sigma^2}$.

Posterior:

$$p(\mu | \bar{y}, \mathcal{B}) = N(\mu_*, \sigma_*^2), \text{ so the posterior}$$

mean is μ_* (more about this below)

& the posterior precision is

$$\frac{1}{\sigma_*^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

i.e. important result

$$(\text{posterior precision}) = (\text{prior precision}) + (\text{likelihood precision})$$

Now earlier the result was

$$\hat{\mu}_* = \frac{\left(\frac{1}{\sigma_0^2}\right)\mu_0 + \left(\frac{n}{\sigma^2}\right)\bar{Y}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

$$= \frac{\underset{\text{prior precision}}{\left(\mu_0\right)} \cdot \underset{\text{prior mean}}{\left(\mu_0\right)} + \underset{\text{likelihood precision}}{\left(n\right)} \cdot \underset{\text{sample mean}}{\left(\bar{Y}\right)}}{\underset{\text{prior precision}}{\left(\mu_0\right)} + \underset{\text{likelihood precision}}{\left(n\right)}}$$

i.e.,

important result

(posterior mean) is a weighted average of the prior mean & the sample mean, with weights given by the (prior precision) & $(\text{likelihood precision})$.

Furthermore, another expression for

μ_* (p. 8) is

$$\mu_* = \left(\frac{\sigma^2}{\sigma_0^2} \right) \mu_0 + \bar{y}$$

$$\left(\frac{\sigma^2}{\sigma_0^2} \right) + n$$

so in the

weighted average, the sample mean gets

n votes & the prior mean gets $(\frac{\sigma^2}{\sigma_0^2})$

votes, so $\mu_0 = \frac{\sigma^2}{\sigma_0^2}$ must be the

$$\mu_0 = \frac{\sigma^2}{\sigma_0^2}$$

prior sample size

This makes excellent sense: to get a diffuse (low-information-content)

prior, choose μ_0 close to 0, which is

equivalent to making σ_0^2 large

~~so σ_0 (large)~~

$p(\mu|y)$

(that's exactly what
"diffuse" means)

In the limit as $\sigma_0^2 \rightarrow \infty$ you get (14)

what's called an improper prior:

$(\mu | \mathcal{B}) = N(\mu_0, \infty)$ is equivalent to

$$p(\mu | \mathcal{B}) \propto C$$

\uparrow
constant



$$\begin{aligned} &= \text{Uniform}(-\infty, \infty) \\ &= U(-\infty, \infty) \end{aligned}$$

This is called
an improper

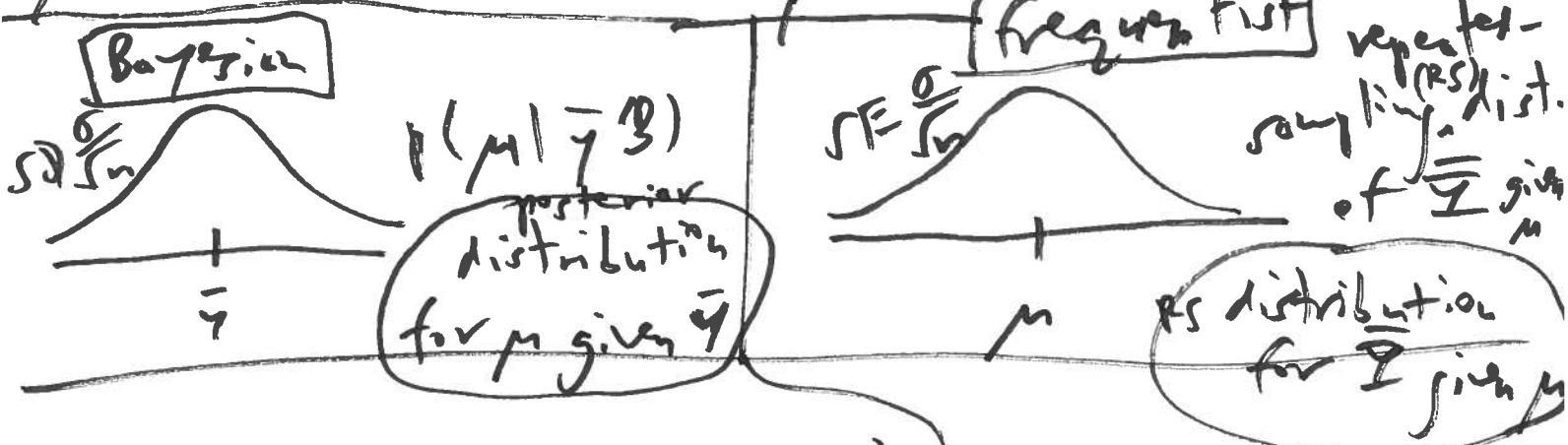
prior because it integrates to infinity,
whereas proper prior distributions integrate

to 1.

In this simplified model, in which $(I_i | \mu | \mathcal{B}) \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$, with the improper $U(-\infty, \infty)$ prior the posterior & likelihood coincide: $(\mu | \bar{Y} | \mathcal{B}) \sim N(\bar{Y}, \frac{\sigma^2}{n})$

and the Bayesian & frequentist approaches (IS)

yield identical answers:



$$p(\mu | \bar{y}, \mathcal{B}) = c_1 e^{-c_2(\mu - \bar{y})^2}$$

Bayesian posterior

$$p(\bar{y} | \mu, \mathcal{B}) = c_2 e^{-c_3(\bar{y} - \mu)^2}$$

Note: The constants c_1 and c_2 are identical
in both cases

frequentist
repeated-sampling
distribution

Interesting thing

about the Normal distribution: if I
give you the density $c_1 e^{-c_2(\mu - \bar{y})^2}$, you
can't tell if it's a Normal distribution

for μ centred at \bar{Y} or a Normal^⑯
distribution for \bar{I} centered at μ . (!)

This is a special case of the Bernstein-von Mises

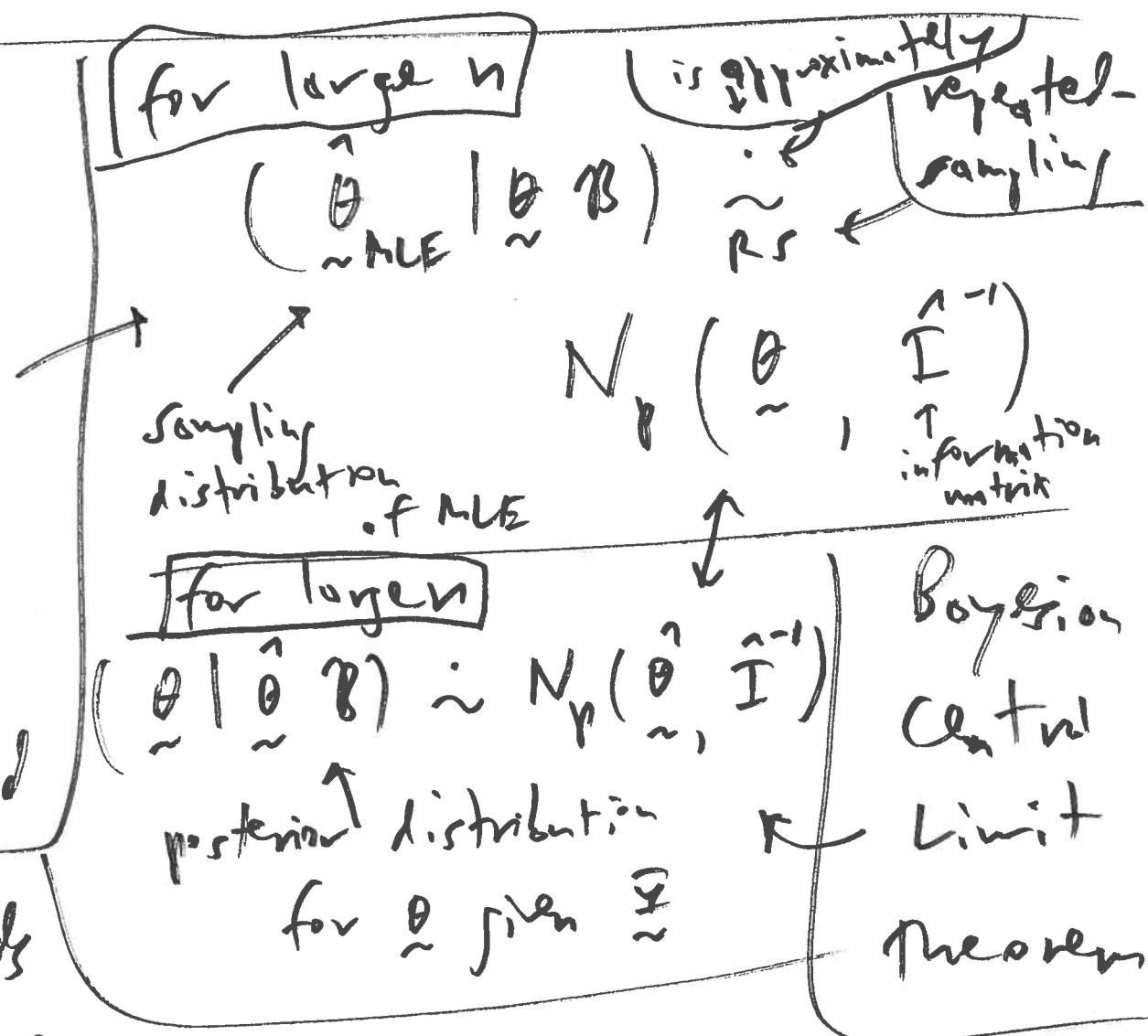
Informal
statement
of the
theorem

Frequentist (maximum-likelihood) & Bayesian inferential conclusions will be similar when (a) n is large (in that case both the posterior & likelihood will be close to Normal) and (b) the prior has relatively low information content (ie., the prior is diffuse).

more technical
way of
putting this | $\theta = (\theta_1, \dots, \theta_p)$ $p \geq 1$
 $(\bar{I}_i | \underline{\theta}, \mathcal{B}) \stackrel{i.i.d.}{\sim} p(y_i | \underline{\theta}, \mathcal{B})$

Bayesian analysis says $(\hat{\theta} | \mathcal{B}) \sim p(\theta | \mathcal{B})$ (17)

Frequentist
Central
Limit
Theorem
for
Maximum
Likelihood



This holds under some technical regularity conditions, including the stipulation that the inference problem is regular (i.e., the set of possible values of \hat{I}_i does not depend on $\hat{\theta}$)

Now we're finally ready for the full model: ⑧

$$\theta = (\mu, \sigma^2) \\ (p=2)$$

$$(\mu, \sigma^2 | \mathcal{B}) \sim p(\mu, \sigma^2 | \mathcal{B}) \\ (I_i | \mu, \sigma^2, \mathcal{B}) \stackrel{\text{IID}}{\sim} N(\mu, \sigma^2) \\ (i=1, \dots, n) \quad \Sigma = (\Sigma_{ij}, \dots, \Sigma_{nn}) \quad \tilde{y} = (y_1, \dots, y_n)$$

It turns out that there is a conjugate prior when both μ & σ^2 are unknown.

The simplest way

to specify it is to remember from AMS 131 that you can write $p(\mu, \sigma^2 | \mathcal{B})$ as a product of a marginal distribution & a conditional distribution: $p(\mu, \sigma^2 | \mathcal{B}) = p(\mu | \mathcal{B}) \cdot p(\sigma^2 | \mu, \mathcal{B})$

and it's also true that $p(\mu, \sigma^2 | \mathcal{B}) = p(\sigma^2 | \mathcal{B}) \cdot p(\mu | \sigma^2, \mathcal{B})$

The second of these options makes more sense scientifically, because you can't think about μ without knowing something about the scale (symbol) of the distribution, and moreover

(19)

it's easier to work with because we already know the conjugate prior for μ when σ^2 is known, namely $p(\mu | \sigma^2, \mathcal{B})$ is Normal (see pp. ⑦-⑧ above in today's notes). [So, what

should we take for $p(\sigma^2 | \mathcal{B})$? A hint can be found in the frequentist fact that, in this model, the repeated-sampling distribution of the sample variance s^2 satisfies $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$ (see p. ⑫ of the 23 Feb notes)]

Switch this around to get $\sigma^2 \sim \frac{(n-1)s^2}{\chi_{n-1}^2} = (n-1)s^2 (\chi_{n-1}^2)^{-1}$.

It looks like we need a distribution that behaves like the reciprocal (inverse) of χ^2 .

There is such a distribution: it's called the scaled inverse χ^2 distribution.

To avoid notational confusion, let $\eta = \sigma^2$. (20)

Then $(\eta | B) \sim \chi^{-2}(r_0, \sigma_0^{-2})$ - read out loud

" η follows the scaled inverse χ^2 distribution with parameters r_0 and σ_0^{-2} " - if

$$p(\eta | B) = \begin{cases} \frac{\left(\frac{r_0}{2}\right)^{\frac{r_0}{2}}}{\Gamma\left(\frac{r_0}{2}\right)} (\sigma_0^{-2})^{\frac{r_0}{2}} \eta^{\frac{r_0}{2}-\left(\frac{r_0}{2}+1\right)} e^{-\frac{r_0 \sigma_0^2}{2\eta}} & \text{for } \eta > 0 \\ 0 & \text{otherwise} \end{cases}$$

see Gelman et al. (2014) p. 576

This is the same as the Inverse Gamma distribution

$\Gamma^{-1}\left(\frac{r_0}{2}, \frac{r_0 \sigma_0^2}{2}\right)$, which we met in an earlier discussion section.

It can be

shown that if $(\eta | B) \sim \tilde{\chi}^2(r_0, \sigma_0^{-2})$ then

$$E(\eta | B) = \frac{r_0}{r_0-2} \sigma_0^{-2} \quad (\text{or long as } r_0 > 2)$$

$$\text{and } V(\gamma | \mathcal{B}) = \frac{2r_0^2 \sigma_0^{-4}}{(r_0 - 2)^2 (r_0 - 4)} \quad (\text{as long as } r_0 > 4). \quad (2)$$

In Homework 2 you get a chance to work with this distribution a bit, to get used to it - it's not as bad as it looks. The

nice thing about $\chi^2(r_0, \sigma_0^2)$ as a prior for $\gamma = \sigma^2$ in the IID $N(\mu, \sigma^2)$ sampling model is that r_0 and σ_0^2 have directly interpretable meanings: r_0 is the prior sample size and σ_0^2 is the prior estimate of σ^2 . So to create a diffuse prior for σ^2 you can take r_0 small (close to 0). To summarize, it turns out that the conjugate prior for (μ, σ^2)

can be expressed hierarchically as follows (22):

$$(\sigma^2 | \mathcal{B}) \sim \chi^{-2}(n_0, \sigma_0^2)$$

$$(\mu | \sigma^2, \mathcal{B}) \sim N(\mu_0, \frac{\sigma^2}{k_0})$$

A note on Bayesian hierarchical models:

We've actually been ~~using~~ working with them all quarter without using that terminology.

A hierarchical statistical model

is just a way of expressing a model

in sequential layers.

In the fairer ICU care study

The sampling model was IID Bernoulli(θ) and the conjugate prior for θ was Beta(α, β):

We wrote this as $\left\{ \begin{array}{l} (\theta | \mathcal{B}) \sim \text{Beta}(\alpha, \beta) \\ ((I_i | \theta, \mathcal{B}) \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta)) \end{array} \right. \quad (i=1, \dots, n)$

This is a Bayesian hierarchical model with two layers \therefore no branches at the top

is the prior layer, and below that is (23)
the sampling distribution layer.

$$(\sigma^2 | \beta) \sim \chi^{-2}(\nu_0, \sigma_0^2)$$

$$(\mu | \sigma^2, \beta) \sim N(\mu_0, \frac{\sigma^2}{K_0})$$

This hierarchical

prior has 4 inputs,
referred to as
hyperparameters

(because they're parameters of the prior
distribution for the sampling model parameters)

I already gave the meaning of ν_0 and σ_0^2 .

K_0 and μ_0 play similar roles for μ :

\uparrow
Kappa K_0 is the prior sample size for μ
and μ_0 is the prior estimate of μ .

To create a diffuse prior for (μ, σ^2)
you can take both ν_0 and K_0 small (close
to 0), & it will then not matter much what you
take for $\sigma_0^2 + \mu_0$.