# JOINT INFERENCE

BST228 Applied Bayesian Analysis



Speaker notes

*No notes on this slide.*

# RECAP

- Inference of location $\mu$ given data $\mathbf{y}$ and precision $\tau$.
- Inference of precision $\tau$ given data $\mathbf{y}$ and location $\mu$.



Speaker notes

- We considered univariate inferences, assuming one of the parameters was known.
- The natural next step is to infer both parameters together.
- In the instrument manufacturer example, we may want to jointly estimate the concentration of a marker in a sample and the measurement error by running replicates. This means we do not need to rely on the reported precision of the instrument.

# JOINT INFERENCE IN THEORY

$$p\left(\boldsymbol{\theta} \mid \mathbf{y}\right) = \frac{p\left(\mathbf{y} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{p\left(\mathbf{y}\right)}$$



Speaker notes

- We often do not know any of the parameters of our model and need to infer them jointly.
- This includes hierarchical models, regression, non-parametric models, etc.
- Bayes theorem remains unchanged for multivariate inference, interpretation of prior, likelihood, and posterior stay the same.
- However, priors and posteriors are now multivariate distributions which require careful analysis.

# JOINT INFERENCE IN PRACTICE

We have posterior

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = p(\theta_1, \ldots, \theta_q \mid \mathbf{y})$$

for $q$ parameters.



Speaker notes

- We need to handle high-distributions in `q` dimensions which can be both computationally and conceptually challenging.
- We often report summaries of the posterior, such as marginal distributions for each of he parameters (see next slide).

# JOINT AND MARGINAL DISTRIBUTIONS

For a two-parameter posterior, by the law of total probability,

$$
\begin{aligned}
p(\theta_1 \mid \mathbf{y}) &= \int d\theta_2 \, p(\theta_1, \theta_2 \mid \mathbf{y}) \\
&= \int d\theta_2 \, p(\theta_1 \mid \theta_2, \mathbf{y}) \, p(\theta_2 \mid \mathbf{y}) \, .
\end{aligned}
$$

The marginal posterior $p(\theta_1 \mid \mathbf{y})$ is an average of the conditional posterior $p(\theta_1 \mid \theta_2, \mathbf{y})$ weighted by the marginal posterior $p(\theta_2 \mid \mathbf{y})$. Here, $\theta_2$ is a nuisance variable that is not of primary concern. The integral is referred to as marginalization.



Speaker notes

*No notes on this slide.*

# MARGINAL DISTRIBUTION EXAMPLE

The marginal posterior chemical concentration is

$$p\left(\mu \mid \mathbf{y}\right) = \int d\tau \, p\left(\mu, \tau \mid \mathbf{y}\right).$$



Speaker notes

- The distribution of likely concentrations is what we are ultimately interested in.
- We thus marginalize with respect to the instrument precision, and the precision $\tau$ is a nuisance parameter.
- But before we can evaluate the marginal posterior, we need to obtain the joint posterior.

# JOINT POSTERIOR FOR NORMAL DATA

The joint posterior is

$$
\begin{aligned}
p\left(\mu,\tau\mid\mathbf{y}\right) &\propto p\left(\mu,\tau\right)p\left(\mathbf{y}\mid\mu,\tau\right)\\
&\propto p\left(\mu,\tau\right)\left(\frac{\tau}{2\pi}\right)^{n/2}\exp\left(-\frac{\tau\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{2}\right).
\end{aligned}
$$

We use a prior ansatz $\tau\sim\text{Gamma}\left(a_0,b_0\right)$ and $\mu\sim\text{Normal}\left(\nu_0,\frac{1}{\kappa_0\tau}\right)$ such that

$$
\begin{aligned}
p\left(\mu,\tau\mid\mathbf{y}\right) &\propto \tau^{1/2}\exp\left(-\frac{\kappa_0\tau}{2}\left(\mu-\nu_0\right)^2\right)\tau^{a_0-1}\exp\left(-b_0\tau\right)\tau^{n/2}\exp\left(-\frac{\tau n}{2}\sum_{i=1}^{n}\frac{y_i^2-2\mu y_i+y_i^2}{n}\right)\\
&\propto \tau^{a_0-1+\frac{n+1}{2}}\exp\left(-\tau\left[b_0+\frac{\kappa_0}{2}\left(\mu^2-2\mu\nu_0+\nu_0^2\right)+\frac{n}{2}\left(s-2\mu\bar{y}+\mu^2\right)\right]\right),
\end{aligned}
$$

where $s=\sum_{i=1}^{n}\frac{y_i^2}{n}$ is the second moment of the sample. We note $a_n=a_0+\frac{n}{2}$ and consider the term in brackets, which we call $L$.



Speaker notes

- Deriving the joint posterior is tedious but a worthwhile exercise.
- This derivation is the most fiddly algebra of the course, but I encourage you to verify the derivation in your own time.
- Using samplers to explore the posterior (see later lectures) allows us to side-step these derivations.
- Aside: Bayes had a bit of a revival starting in the late 90s because computational statistics became feasible.

# JOINT POSTERIOR FOR NORMAL DATA

$$
\begin{aligned}
L &= b_0 + \frac{\kappa_0}{2}\left(\mu^2 - 2\mu\nu_0 + \nu_0^2\right) + \frac{n}{2}\left(s - 2\mu\bar{y} + \mu^2\right) \\
&= b_0 + \frac{1}{2}\left(\kappa_0\mu^2 - 2\kappa_0\mu\nu_0 + \kappa_0\nu_0^2 + ns - 2n\mu\bar{y} + n\mu^2\right) \\
&= b_0 + \frac{\kappa_0 + n}{2}\left(\mu^2 - 2\mu\frac{\kappa_0\nu_0 + n\bar{y}}{\kappa_0 + n} + \frac{\kappa_0\nu_0^2 + ns}{\kappa_0 + n}\right) \\
&= b_0 + \frac{\kappa_0 + n}{2}\left(\mu^2 - 2\mu\frac{\kappa_0\nu_0 + n\bar{y}}{\kappa_0 + n} + \left(\frac{\kappa_0\nu_0 + n\bar{y}}{\kappa_0 + n}\right)^2 - \left(\frac{\kappa_0\nu_0 + n\bar{y}}{\kappa_0 + n}\right)^2 + \frac{\kappa_0\nu_0^2 + ns}{\kappa_0 + n}\right) \\
&= b_0 + \frac{\kappa_0 + n}{2}\left(\mu - \frac{\kappa_0\nu_0 + n\bar{y}}{\kappa_0 + n}\right)^2 + \frac{\kappa_0 + n}{2}\left(\frac{\kappa_0\nu_0^2 + ns}{\kappa_0 + n} - \left(\frac{\kappa_0\nu_0 + n\bar{y}}{\kappa_0 + n}\right)^2\right).
\end{aligned}
$$

We note $\kappa_n = \kappa_0 + n$ and $\nu_n = \frac{\kappa_0\nu_0 + n\bar{y}}{\kappa_0 + n}$. We further consider the first and last terms which is $b_n$.



Speaker notes

*No notes on this slide.*

# JOINT POSTERIOR FOR NORMAL DATA

$$
\begin{aligned}
b_n &= b_0 + \frac{\kappa_0 + n}{2} \left( \frac{\kappa_0 \nu_0^2 + ns}{\kappa_0 + n} - \left( \frac{\kappa_0 \nu_0 + n\bar{y}}{\kappa_0 + n} \right)^2 \right) \\
&= b_0 + \frac{1}{2} \left( \kappa_0 \nu_0^2 + ns - \frac{\kappa_0^2 \nu_0^2 + 2\kappa_0 \nu_0 n \bar{y} + n^2 \bar{y}^2}{\kappa_0 + n} \right) \\
&= b_0 + \frac{\left(\kappa_0 \nu_0^2 + ns\right)\left(\kappa_0 + n\right) - \kappa_0^2 \nu_0^2 - 2\kappa_0 \nu_0 n \bar{y} - n^2 \bar{y}^2}{2\left(\kappa_0 + n\right)} \\
&= b_0 + \frac{\kappa_0^2 \nu_0^2 + n\kappa_0 \nu_0^2 + \kappa_0 n s + n^2 s - \kappa_0^2 \nu_0^2 - 2\kappa_0 \nu_0 n \bar{y} - n^2 \bar{y}^2}{2\left(\kappa_0 + n\right)} \\
&= b_0 + \frac{n}{2\left(\kappa_0 + n\right)} \left( \kappa_0 \left( s - 2\nu_0 \bar{y} + \nu_0^2 \right) + n \left( s - \bar{y}^2 \right) \right).
\end{aligned}
$$



Speaker notes

*No notes on this slide.*

# UPDATE RULES

$$\kappa_n = \kappa_0 + n,$$

$$\nu_n = \frac{\kappa_0 \nu_0 + n\bar{y}}{\kappa_0 + n},$$

$$a_n = a_0 + \frac{n}{2},$$

$$b_n = b_0 + \frac{n}{2\left(\kappa_0 + n\right)} \left(\kappa_0 \left(s - 2\nu_0 \bar{y} + \nu_0^2\right) + n \operatorname{var} \mathbf{y}\right).$$



Speaker notes

- We derived four update rules for the posterior parameters, but the approach quickly becomes infeasible for more complex models.
- We have the joint posterior, and we can evaluate the marginal posterior distribution.
- In your own time, consider the limiting cases of large observation precision $\tau \to \infty$, large sample size $n \to \infty$, and large prior precision $\kappa_0 \to \infty$. Do the limiting cases agree with your intuition?

```r
> # Define some random variables with correct support.
> n <- 7
> y <- rnorm(n)
> tau <- rgamma(1, 5, 5)
> mu <- rnorm(1)
> nu_0 <- rnorm(1)
> kappa_0 <- rgamma(1, 5, 5)
> a_0 <- rgamma(1, 5, 5)
> b_0 <- rgamma(1, 5, 5)
>
> reference <- dnorm(mu, mean = nu_0, sd = 1 / sqrt(kappa_0 * tau),
+   log = TRUE) + dgamma(tau, shape = a_0, rate = b_0, log = TRUE) +
+   sum(dnorm(y, mean = mu, sd = 1 / sqrt(tau), log = TRUE))
> print(paste("reference", reference))
[1] "reference -12.7749474363713"
>
> # Replace distributions by explicit values.
> test_value <- log(kappa_0 * tau / (2 * pi)) / 2 - kappa_0 * tau / 2 *
+   (mu - nu_0)^2 + a_0 * log(b_0) - lgamma(a_0) + (a_0 - 1) * log(tau) -
+   b_0 * tau + n * log(tau / (2 * pi)) / 2 - tau / 2 * sum((y - mu)^2)
>
> stopifnot(all.equal(test_value, reference))
>
> # Group terms and introduce a normalization constant to absorb terms.
> evaluate_log_norm <- function(n, kappa, a, b) {
+   return(
+     (log(kappa) - (n + 1) * log(2 * pi)) / 2 +
+     a * log(b) - lgamma(a)
+   )
+ }
```

Speaker notes

- Using R or another programming language can be a convenient way to verify algebraich manipulation by evaluating the manipulated expressions at some arbitrary values.

# MARGINAL POSTERIOR FOR $\mu$

Recall

$$
\begin{aligned}
p\left(\mu \mid \mathbf{y}, a_n, b_n\right) &= \int d\tau \sqrt{\frac{\kappa_n \tau}{2\pi}} \exp\left(-\frac{\kappa_n \tau}{2}\left(\mu - \nu_n\right)^2\right) \frac{b_n^{a_n}}{\Gamma(a_n)} \tau^{a_n - 1} \exp\left(-b_n \tau\right) \\
&\propto \int d\tau\, \tau^{a_n + 1/2 - 1} \exp\left(-\left(b_n + \frac{\kappa_n \left(\mu - \nu_n\right)^2}{2}\right)\tau\right),
\end{aligned}
$$

where $\{\nu_n, \kappa_n, a_n, b_n\}$ are posterior parameters. The integrand is the kernel of a gamma distribution with effective parameters.

$$
\begin{aligned}
a' &= a_n + \frac{1}{2} \\
b' &= b_n + \frac{\kappa_n \left(\mu - \nu_n\right)^2}{2}.
\end{aligned}
$$



Speaker notes

*No notes on this slide.*

# MARGINAL POSTERIOR FOR $\mu$

The integral thus evaluates to the inverse normalization constant $\Gamma(a')b'^{-a'}$, and

$$p\left(\mu \mid \mathbf{y}\right) \propto \left(1 + \frac{a_n \kappa_n \left(\mu - \nu_n\right)^2}{2 a_n b_n}\right)^{-\frac{2a_n+1}{2}},$$

where we have absorbed a factor of $b_n$ in the normalization constant and added a factor $a_n$ to nominator and denominator. We compare the expression with the kernel of a non-centered, scaled Student's t-distribution

$$\left(1 + \frac{\kappa \left(\mu - \nu\right)^2}{q}\right)^{-\frac{q+1}{2}}$$

with $q$ degrees of freedom, location $\nu$, and precision $\kappa$.



Speaker notes

*No notes on this slide.*

## MARGINAL POSTERIOR FOR $\mu$

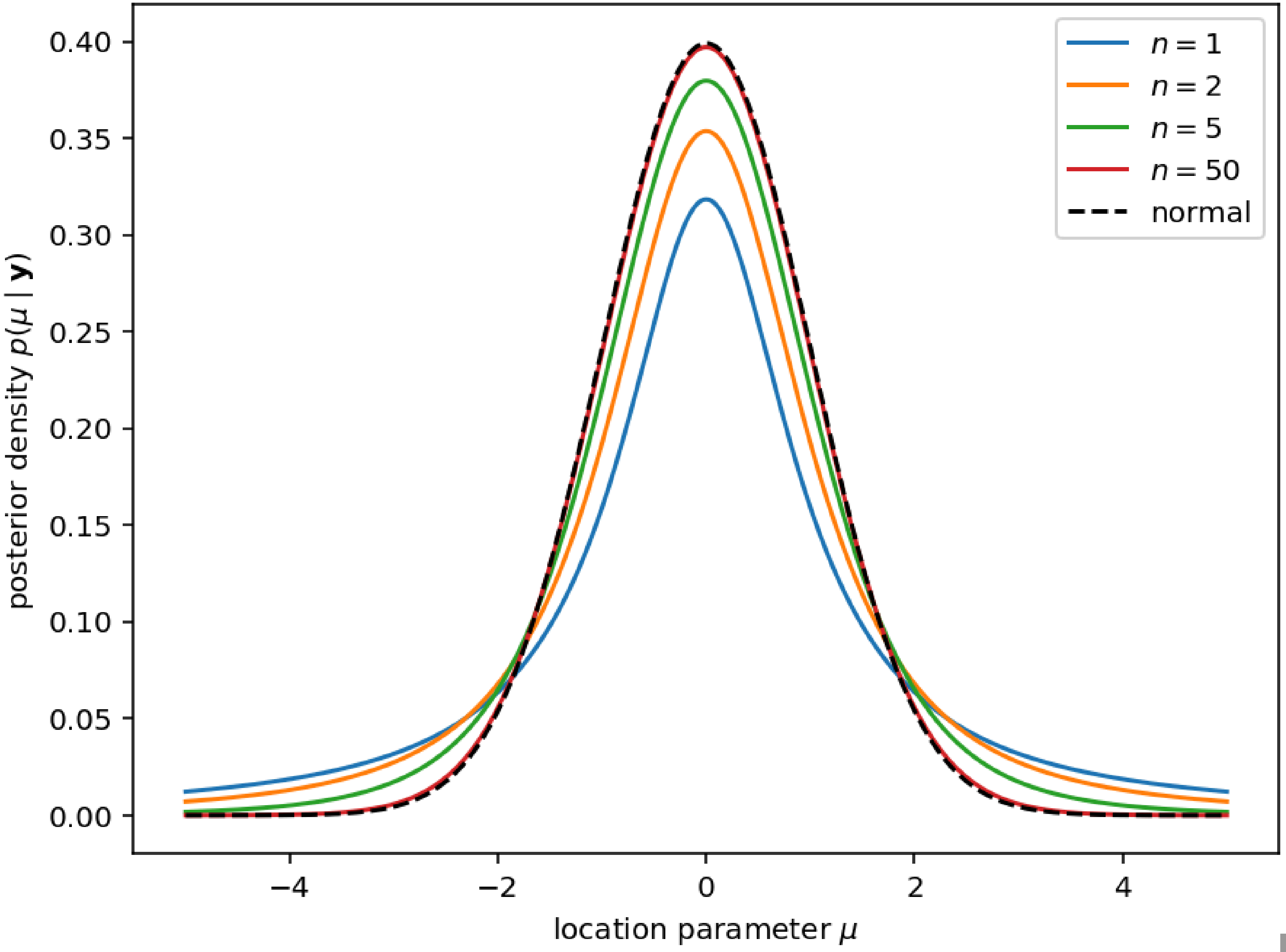
Matching terms, we have a non-centered, scaled Student's t-distribution with $2a_n$ degrees of freedom. The marginal posterior is

$$\mu \mid \mathbf{y} \sim \text{StudentT}_{2a_n}\left(\nu_n, \frac{b_n}{\kappa_n a_n}\right).$$



Speaker notes

*No notes on this slide.*






Speaker notes

- The distribution, here shown for $\nu_n = 0$ has heaver tails than a normal distribution with the same parameters. However, even for relatively small sample size of $n = 50$, the Student's t-distribution closely approximates a normal distribution.
- This extra variance in the posterior for the chemical concentration is expected because we must also infer the observation precision given replicate measurements.

# PAIRED EXERCISE

Consider again the example data $\mathbf{y} = (2.1, 2.5, 1.6, 1.7)$.

- Using the derived update rules, what are the posterior parameters?
- Draw posterior samples of $\mu$? Can you think of two ways to obtain the samples?
- How do summary statistics of the posterior for $\mu$ and $\tau$ compare with inference assuming one known parameter?



Speaker notes

*No notes on this slide.*

```r
1  > # Declare data and known noise level.
2  > y <- c(2.1, 2.5, 1.6, 1.7)
3  > n <- length(y)
4  > # Define hyperparameters.
5  > nu_0 <- 0
6  > kappa_0 <- 1e-4
7  > a_0 <- 1e-3
8  > b_0 <- 1e-3
9  > # Update parameters and sample.
10 > nu_n <- (nu_0 * kappa_0 + n * mean(y)) / (kappa_0 + n)
11 > kappa_n <- kappa_0 + n
12 > a_n <- a_0 + n / 2
13 > b_n <- b_0 + n / 2 * (var(y) * n / (n - 1) + kappa_0 / (kappa_0 + n)
14 +                     * (mean(y) - nu_0) ** 2)
15 > tau_samples <- rgamma(1000, a_n, b_n)
16 > sigma_samples <- 1 / sqrt(tau_samples)
17 > mu_samples <- rnorm(1000, nu_n, 1 / sqrt(kappa_n * tau_samples))
18 > c(mean(mu_samples), sd(mu_samples),
19 +   mean(sigma_samples), sd(sigma_samples))
20 [1] 1.9702649 0.3226011 0.5906731 0.2788981
21 >
```

Speaker notes

- Lines #2-3 declare data and #5-8 declare hyperparameters.
- #10-14 evaluate the posterior parameters.
- #15-17 sample from the posterior in two steps. Alternatively, we could have directly sampled from a Student’s t-distribution.