Normal-Normal Inference

STAT 251, Unit 6

Overview

Review

Notation for the Normal Distribution

Inference with the Normal Distribution: μ unknown, σ^2 known

Review

Normal Distribution

The normal distribution is a "workhorse" distribution in statistics. It has two parameters, the mean μ and the variance σ^2.

Refer to Unit 2C for a review of concepts such as the empirical rule and the prevalence of the normal distribution in statistical inference.

Notation for the Normal Distribution

Notation

Technical Note: If we want to refer to more than one value, we can consider the vector that contains all values of interest. Conventional notation for a vector is a bold-faced symbol, such as $\mathbf{y}$ instead of y, or $\boldsymbol{\theta}$ instead of θ. For example, the vector of observed data would be written as

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

where the first element of the vector is the first observation (i.e., y_1), the second element is the second observation, and so forth.

Notation (cont.)

With our established pattern of representing a parameter of interest as θ, we could say that the Normal distribution is $f(y|\boldsymbol{\theta})$ where

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}.$$

That is, $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. But I will refer to the normal distribution's parameters specifically as μ and σ^2 rather than generically as θ_1 and θ_2.

Inference with the Normal Distribution: μ unknown, σ^2 known

Structure of Units 6–8

Suppose $y_i|(\mu, \sigma^2) \overset{iid}{\sim} N(\mu, \sigma^2)$.

There are two parameters to consider. We will work through the following progression:

- Unit 6: Inference with μ unknown but σ^2 known.
- Unit 7: Inference with σ^2 unknown but μ known.
- Unit 8: Inference with μ and σ^2 unknown

While Unit 8 is the only realistic scenario among the three, that unit will rely on results that we establish in Units 6 and 7.

Prior for μ (with known σ^2)

Consider the normal family of distributions—that is, the collection of all probability density functions $f(\cdot|\mu, \sigma^2)$ such that

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

and $\mu \in (-\infty, \infty), \sigma^2 > 0$. Note that any value for μ could be used with a normal distribution.

Question: To allow μ to potentially be any value (positive/negative/zero, arbitrarily small/moderate/large), which of the following distributions would work as a prior distribution for μ?

a. $\mu \sim \textit{Normal}$
b. $\mu \sim \textit{Beta}$
c. $\mu \sim \textit{Gamma}$
d. $\mu \sim \textit{Binomial}$
e. $\mu \sim \textit{Poisson}$

Question: To allow μ to potentially be any value (positive/negative/zero, arbitrarily small/moderate/large), which of the following distributions would work as a prior distribution for μ?

a. $\mu \sim \mathit{Normal}$
b. $\mu \sim \mathit{Beta}$
c. $\mu \sim \mathit{Gamma}$
d. $\mu \sim \mathit{Binomial}$
e. $\mu \sim \mathit{Poisson}$

Only (a.) gives the requested support of $(-\infty, \infty)$.

We'll consider $\mu \sim N(m, v)$ as the prior distribution, and then derive the posterior distribution of μ.

More Notation

There are two **distinct** normal distributions to keep track of:

$$Y|\{\mu, \sigma^2\} \overset{iid}{\sim} N(\mu, \sigma^2) \tag{1}$$

$$\mu \sim N(m, v) \tag{2}$$

Equation (1) is for the likelihood; Equation (2) is for the prior distribution of the parameter μ.

Notice I adopt the following notation:

- m and v are the mean and variance of the **prior distribution** of the parameter μ
- μ and σ^2 are the (conditional) mean and variance of the **response variable**, Y_i.

More Notation

Consider adult female heights, and suppose $Y_i | \{\mu, \sigma^2 = 9\} \stackrel{iid}{\sim} N(\mu, 9)$ and that $\mu \sim N(65, 1)$.

μ represents:

m=65 implies:

$\sigma^2 = 9$ implies:

v=1 implies:

More Notation

Consider adult female heights (measured in inches), and suppose $Y_i|\{\mu, \sigma^2 = 9\} \overset{iid}{\sim} N(\mu, 9)$ and that $\mu \sim N(65, 1)$.

μ represents: The average height (in inches) among the population of *all* adult females

m=65 implies: Our prior belief about the value of μ is centered at 65 inches. We expect μ to be about 65 inches.

More Notation

Consider adult female heights (measured in inches), and suppose $Y_i|\{\mu, \sigma^2 = 9\} \overset{iid}{\sim} N(\mu, 9)$ and that $\mu \sim N(65, 1)$.

$\sigma^2 = 9$ implies: Individual female heights have an average squared distance from μ of 9.

- Roughly speaking, if we selected one adult female at random and observed her height, it would tend to differ from the population average, μ, by a magnitude of *about* $\sigma = 3$.

v=1 implies: Our prior belief has a variance of 1 $inch^2$ (and thus a standard deviation of 1 inch).

- The smaller the value of v, the stronger our belief that μ is close to $m = 65$.

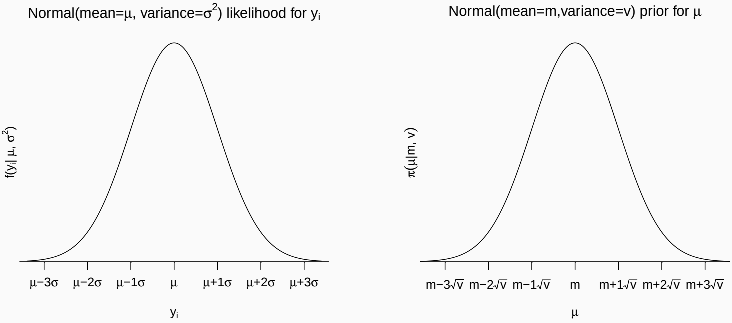
Comparison of Prior and Likelihood

Consider adult female heights (measured in inches), and suppose $Y_i|\{\mu, \sigma^2 = 9\} \overset{iid}{\sim} N(\mu, 9)$ and that $\mu \sim N(65, 1)$.

This prior, $\mu \sim N(m = 65, v = 1)$, represents being very sure (95%) that the **average** adult female height is between 63 and 67 inches, and being almost certain (99.7%) that the **average** adult female height is between 62 and 68 inches.

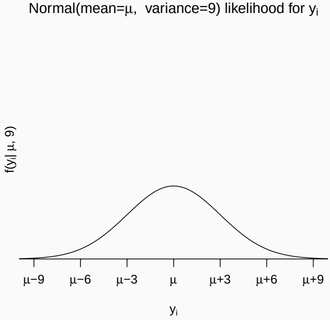
But even if μ and σ^2 were known, there is still a distribution for **individual** adult female heights. That distribution is assumed to be the $Normal(\mu, \sigma^2)$ distribution (our assumed distribution for the likelihood).

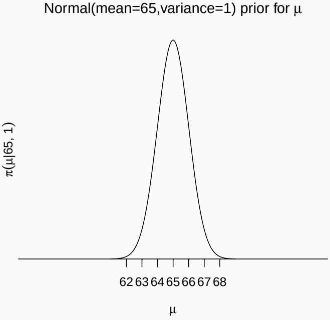
Comparison of Prior and Likelihood



Comparison of Likelihood and Prior

Consider adult female heights (measured in inches), and suppose $Y_i|\{\mu, \sigma^2 = 9\} \overset{iid}{\sim} N(\mu, \sigma^2 = 9)$ and that $\mu \sim N(m = 65, v = 1)$.





Posterior for μ (σ^2 known)

If $y_i|\{\mu, \sigma^2\} \overset{iid}{\sim} N(\mu, \sigma^2)$, and $\mu \sim N(m, v)$, and σ^2 is known,

then

$$\mu|\{y_1, \ldots, y_n, \sigma^2\} \sim N(m^\star, v^\star)$$

with

$$m^\star = \frac{nv\bar{y} + \sigma^2 m}{nv + \sigma^2}$$

and

$$v^\star = \left(\frac{n}{\sigma^2} + \frac{1}{v}\right)^{-1} = \frac{v\sigma^2}{nv + \sigma^2}$$

The derivation of this result will be shown on the board.

Example

Standardized tests are often designed so that they will be (approximately) normally distributed, and with a given mean and standard deviation. However, such tests are periodically modified in such a way that the distribution of the scores might be changed.

Scenario: Suppose that the current version of an IQ test is normally distributed with mean 100 and standard devation 15. A proposed modification of the test is anticipated to still have scores that are normally distributed with standard deviation 15, but the mean might be a little different. The new version is administered to 20 adults in a pilot study, and their scores are determined.

Setup

Likelihood: $y_i|\{\mu, \sigma^2 = 15^2\} \overset{iid}{\sim} N(\mu, \sigma^2 = 15^2)$, $i = 1, \ldots, n = 20$.

Prior Distribution for μ?

- Mean for current version equals 100.
- It is believed mean for modified version "might be a little different."
- I'll choose the $N(m = 100, v = 5^2)$ prior distribution for μ.

Consider some implications of this prior, $\mu \sim N(100, 5^2)$?

- $E(\mu) =$
- $Pr(95 \leq \mu \leq 105) \approx$
- $Pr(90 \leq \mu \leq 110) \approx$

Data

In the pilot study, the 20 scores are given by the table below.

Individual	1	2	3	4	5	6	7	8	9	10
Score	108	88	97	129	91	80	114	118	87	89
Individual	11	12	13	14	15	16	17	18	19	20
Score	106	106	102	112	90	119	100	106	93	97

Table 1: Scores for 20 adults in modified IQ test (fictional data)

```
scores <- c(108, 88, 97, 129, 91, 80, 114, 118,
            87, 89, 106, 106, 102, 112, 90, 119,
            100, 106, 93, 97)
```

In-class practice

For the IQ example,

1. Determine the posterior distribution of μ.
2. What is the mean of the posterior distribution?
3. What is the mode?
4. What is the central 95% posterior credible interval?
5. Does it seem reasonable to believe the new version has an average of 100, just as the current version does? Explain.
6. Plot the prior and posterior distribution on the same graph.

A few comments:

Conjugacy: If we use a normal distribution for the likelihood, and the variance is known, then: Normal(m,v) prior for $\mu \longrightarrow$ Normal$(m^\star, v^\star)$ posterior for μ.

Interpretation of Posterior Mean, $m^\star$: The posterior mean is a weighted average of the prior mean and sample mean of the data. For example, we can write $m^\star = w\bar{y} + (1 - w)m$, where $w = nv/(nv + \sigma^2)$ (and note that $0 < w < 1$).

The weights placed on each of these means depend on two factors: (1) the magnitudes of the variance of the data, σ^2, and the variance of the prior distribution, v; (2) the number of observations, n.

More on the Interpretation of Posterior Mean

Again, $m^\star = w\bar{y} + (1-w)m$, where $w = nv/(nv + \sigma^2)$.

Some intuition: As n increases, we get more and more influence from the likelihood on the posterior. Thus, the sample average, $\bar{y}$ becomes increasingly influential on the posterior mean of μ, $m^\star$.

If v is smaller than σ^2, then there is less variability in the prior belief than there would be in a random observation from the likelihood. Smaller variability implies it should be trusted more, so the prior mean can be relatively influential. Conversely, if v is very large, the prior belief is not very precise, so we don't expect μ to be that close to m, and thus the posterior mean shifts even more towards what the data suggest about μ.