07 - Normal analysis

HCI/PSYCH 522
Iowa State University

February 10, 2022

Overview

- Inference for means
  - Estimating 1 mean
  - Comparing 2 means

Estimating 1 mean

Suppose we have

- n numerical observations,
- with the same population mean μ and
- population standard deviation σ, and
- observations are independent.

Let Y_i be the value for the ith observation and assume $Y_i \stackrel{ind}{\sim} N(\mu, \sigma^2)$.

The sample can be summarized by the sample mean

$$\overline{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$

and sample variance

$$S^2 = \frac{(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + \cdots (Y_n - \overline{Y})^2}{n - 1}$$

(or the sample standard deviation $S = \sqrt{S^2}$.)

Sample statistics in R

```
heights <- c(66.9, 63.2, 58.7, 64.2, 65.1)

length(heights) # number of observations

## [1] 5

mean(heights) # sample mean

## [1] 63.62

var(heights)  # sample variance

## [1] 9.417

sd(heights)   # sample standard deviation

## [1] 3.068713
```

Parameter estimation

If we assume $Y_i \stackrel{ind}{\sim} N(\mu, \sigma^2)$, then we can use these sample statistics to estimate population parameters:

- $\hat{\mu} = \overline{Y}$,
- $\hat{\sigma} = S$, and
- $\hat{\sigma}^2 = S^2$.

Please remember that sample statistics are only estimates (not the true values).

Posterior belief about population mean

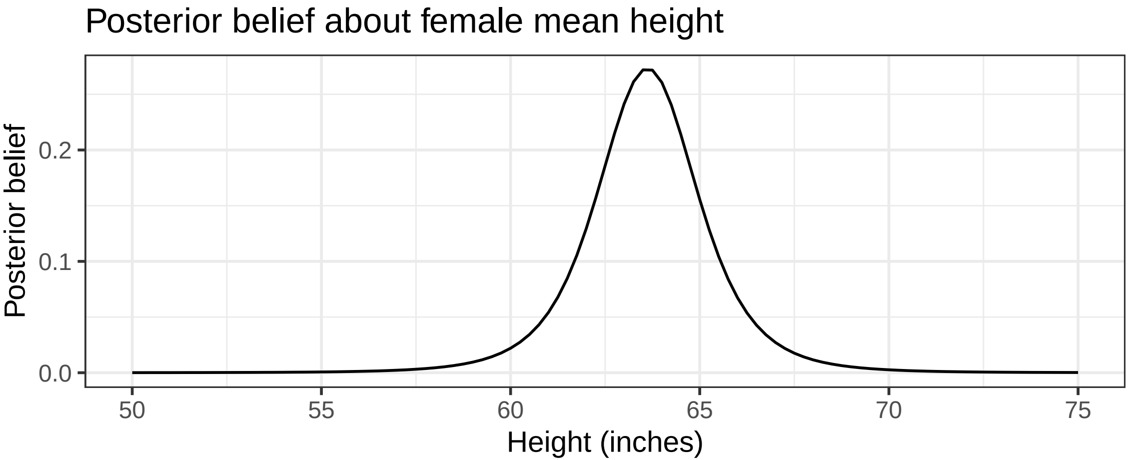
Our posterior belief about the population mean is

$$\mu|y \sim t_{n-1}(\overline{y}, s^2/n)$$

where

- $y = (y_1, \ldots, y_n)$ is the data,
- n is the sample size,
- $\overline{y}$ is the sample mean,
- s^2 is the sample variance, and
- $t_{n-1}(\overline{y}, s^2/n)$ is a T distribution with
  - $n-1$ degrees of freedom,
  - location $\overline{y}$, and
  - scale s.

Posterior belief about female mean height



Credible interval in R

```
t.test(heights, conf.level = 0.95)

##
##  One Sample t-test
##
## data:  heights
## t = 46.358, df = 4, p-value = 1.295e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  59.80969 67.43031
## sample estimates:
## mean of x
##     63.62
```

Calculating posterior probabilities

What is our belief that mean female height is greater than 60 inches?

$$P(\mu > 60|y)$$

```
1-pt((60-mean(heights))/(sd(heights)/sqrt(length(heights))), df = length(heights)-1)

## [1] 0.9711426
```

or

```
plst <- function(q, df, location, scale) { # location-scale t distribution
  pt( (q-location)/scale, df = df)
}
1-plst(60, df = length(heights)-1, location = mean(heights), scale = sd(heights)/sqrt(length(heights)))

## [1] 0.9711426
```

Comparing 2 means

Suppose we have groups indexed by $g = 1, \ldots, G$

- n_g numerical observations in group g,
- the same population mean μ_g within a group and
- same population standard deviation σ_g within a group,
- all observations are independent.

Let Y_{ig} be the value for the ith observation in the gth group and assume $Y_{ig} \stackrel{ind}{\sim} N(\mu_g, \sigma_g^2)$.

When we collect data, we will have a sample mean and sample standard deviation for each group.

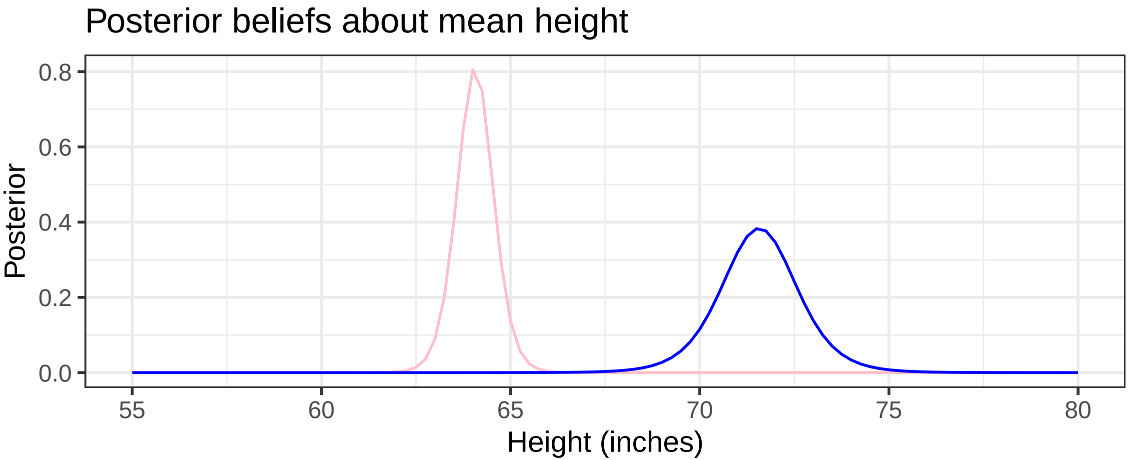
Sample statistics in R

```
d <- read_csv("heights.csv")

d %>%
  group_by(sex) %>%
  summarize(n = n(),
            mean = mean(height),
            sd = sd(height))

## # A tibble: 2 x 4
##   sex        n  mean    sd
##   <chr>  <int> <dbl> <dbl>
## 1 female    11  64.1  1.59
## 2 male       7  71.6  2.66
```

Posterior beliefs



Posterior probabilities

What is the probability that males are, on average, taller than females?

$$P(\mu_{\text{male}} > \mu_{\text{female}}|y)$$

We use a Monte Carlo approach

```
rlst <- function(n, df, location, scale) {
  location+scale*rt(n, df = df)
}
n_reps <- 100000
mu_female <- rlst(n_reps, df = 11-1, location = 64.1, scale = 1.59/sqrt(11))
mu_male   <- rlst(n_reps, df =  7-1, location = 71.6, scale = 2.66/sqrt(7))
mean(mu_male > mu_female)

## [1] 0.99981
```

Credible interval for the difference

```
a <- 1 - 0.95
quantile(mu_male - mu_female, prob = c(a/2, 1-a/2))

##      2.5%     97.5%
##  4.822371 10.161489
```

Using built in R functions

```
d <- read_csv("heights.csv")
t.test(height ~ sex, data = d)

##
##  Welch Two Sample t-test
##
## data:  height by sex
## t = -6.7492, df = 8.7839, p-value = 9.392e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.033670  -4.981915
## sample estimates:
## mean in group female   mean in group male
##             64.06364             71.57143
```