05 - Normal distribution

HCI/PSYCH 522
Iowa State University

February 8, 2022

Overview

- Normal distribution
  - Numerical data
- Inference for means
  - Estimating 1 mean
  - Comparing 2 means
  - Comparing 3+ means

Normal

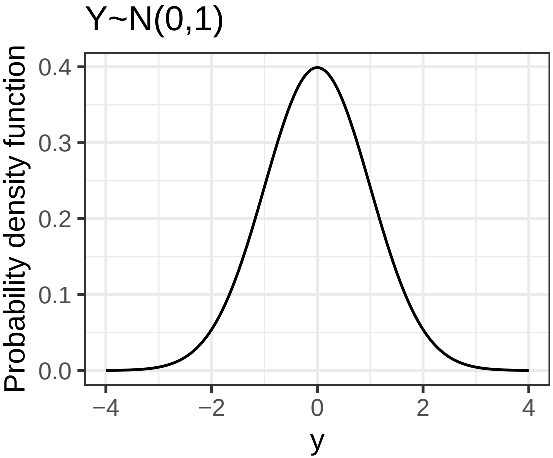
We typically model numerical data with a normal distribution. If $Y \sim N(\mu, \sigma^2)$, then

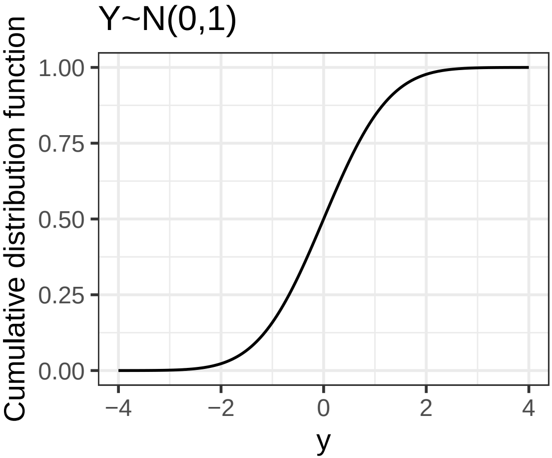
- the expected value $E[Y] = \mu$,
- variance $Var[Y] = \sigma^2$,
- standard deviation $SD[Y] = \sigma$,
- probability density function (bell-shaped curve)

$$f(y) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right),$$

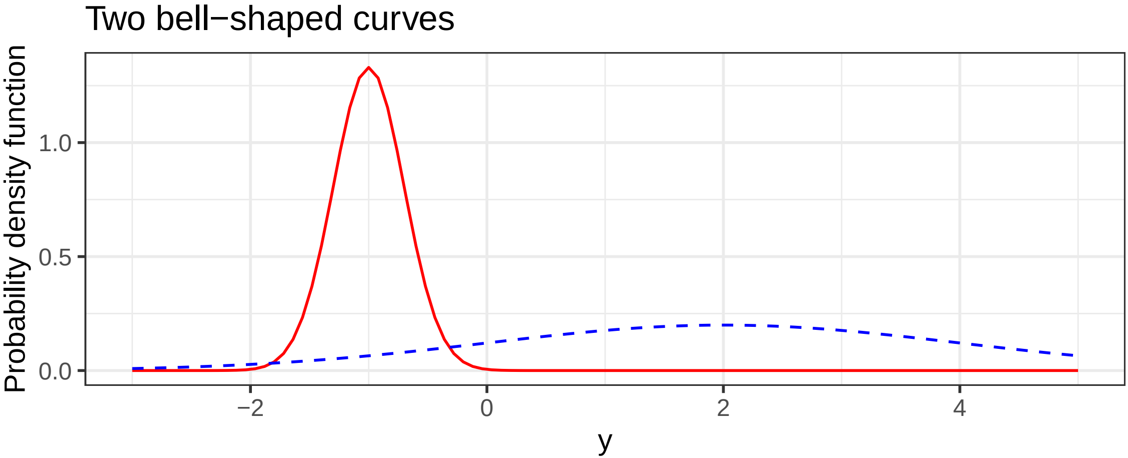
- cumulative distribution function $P(Y \leq y)$.

Normal

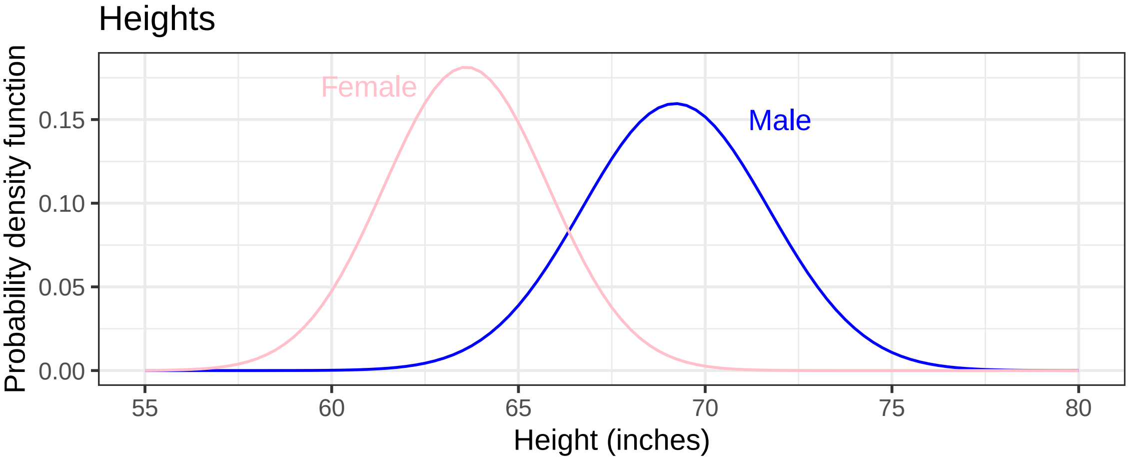

Y~N(0,1)
Probability density function
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
y


Y~N(0,1)
Cumulative distribution function
1.00
0.75
0.50
0.25
0.00
−4
−2
0
2
4
y

Normal

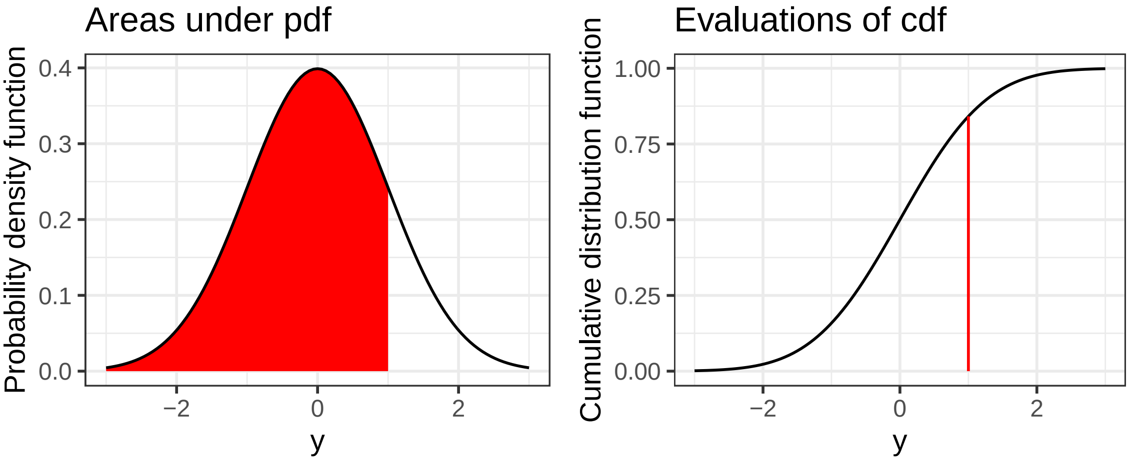

Two bell−shaped curves
Probability density function
1.0
0.5
0.0
−2
0
2
4
y

Heights


Heights
Female
Male
Probability density function
0.15
0.10
0.05
0.00
55
60
65
70
75
80
Height (inches)

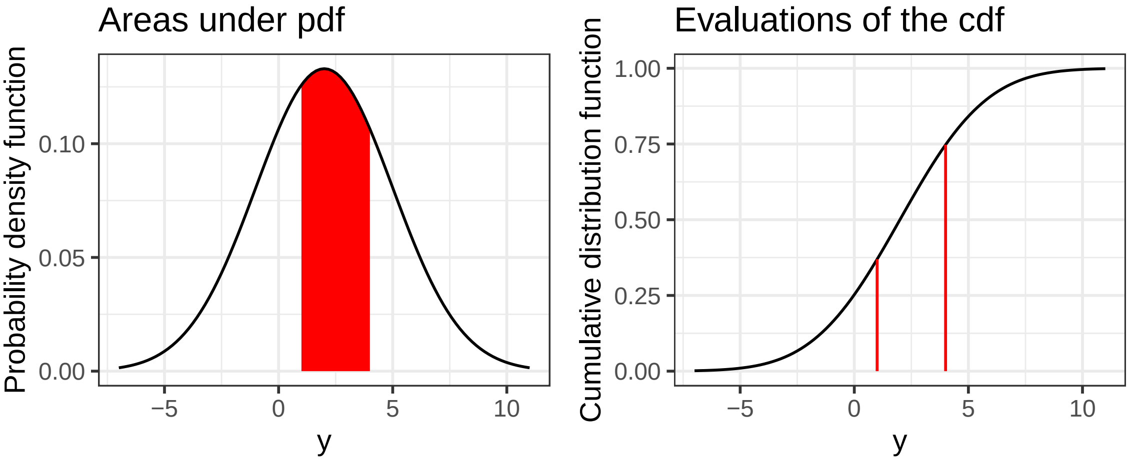
Probabilities

Let $Y \sim N(0,1)$ and calculate $P(Y < 1)$.



Probabilities

Let $Y \sim N(2, 3^2)$ and calculate $P(1 < Y < 4) = P(Y < 4) - P(Y < 1)$.



Probabilities in R

Let $Y \sim N(-3, 4^2)$.

```
mn <- -3
s  <- 4
```

Calculate $P(Y < 0)$.

```
pnorm(0, mean = -3, sd = 4)

## [1] 0.7733726
```

Calculate $P(Y > 1)$.

```
1-pnorm(1, mean = -3, sd = 4)

## [1] 0.1586553
```

Probabilities in R

Let $Y \sim N(-3, 4^2)$.

```
mn <- -3
s  <- 4
```

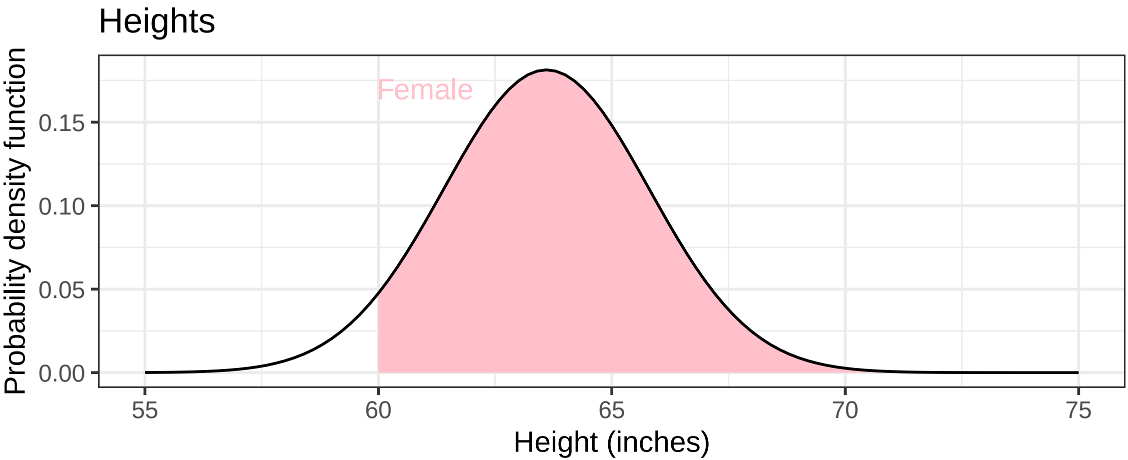
Calculate $P(0 < Y < 1) = P(Y < 1) - P(Y < 0)$.

```
pnorm(1, mean = -3, sd = 4) - pnorm(0, mean = -3, sd = 4)

## [1] 0.0679721
```

For continuous random variables, e.g. normal, $P(Y = y) = 0$ for any value y. This is NOT true for discrete random variables, e.g. binomial.

Probability female height is above 60 inches?



Probability female height is above 60 inches?

Let $Y \sim N(63.6, 2.2^2)$. Calculate $P(Y > 60)$.

```
1-pnorm(60, mean = 63.6, sd = 2.2)

## [1] 0.9491182
```

Estimating 1 mean

Suppose we have

- n numerical observations,
- with the same mean μ and
- and standard deviation σ,
- that 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

```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

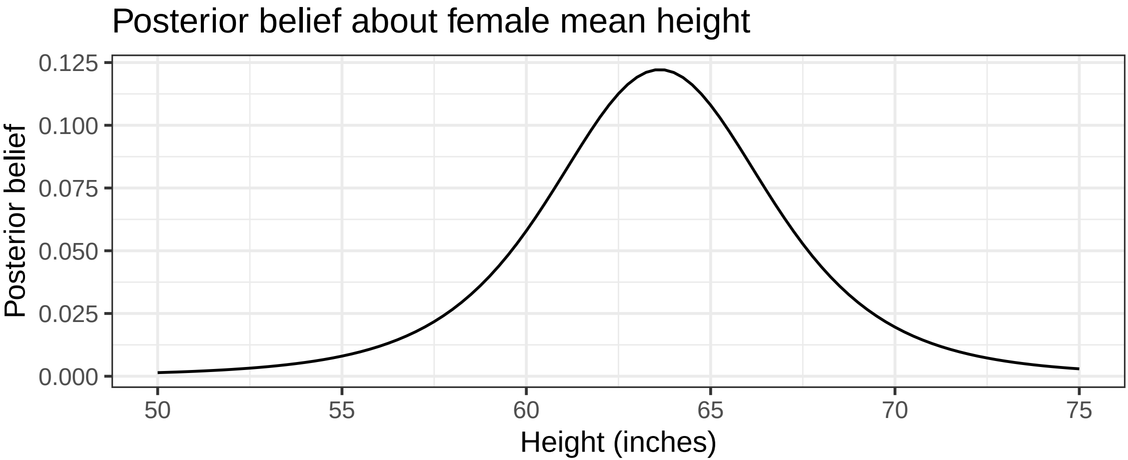
Our posterior belief about the population mean is

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

where

- $y = (y_1, \dots, y_n)$ is the data,
- n is the sample size,
- $\overline{y}$ is the sample mean,
- s^2 is the sample variance, and
- $t_{n-1}(\overline{y}, s^2)$ 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), df = length(heights)-1)

## [1] 0.8482461
```

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))

## [1] 0.8482461
```