Stat 306:
Finding Relationships in Data.
Lecture 4
2.2 (continued) + 2.5 Intervals for simple linear regression

t-test

Age vs. Money

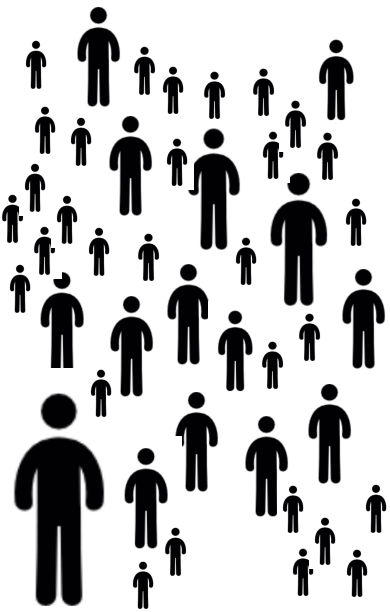
Dependent variable

χ → old (0)
χ → young (1)

Independent variable

γ → dollars ($)
In bank account

Population



Population parameters

μ_0 , μ_1 , σ^2

Hypothesis Test

$H_0 : \mu_0 = \mu_1$

$H_1 : \mu_0 \neq \mu_1$

Sample statistics

$\bar{y}_0 = 56$
$\bar{y}_1 = 27$
$\bar{y}_0 - \bar{y}_1 = 29$
$s_p = 10.81$

t = 2.68, *df* = 7
p-value = 0.03
95% C.I. = [3.4, 54.6]

Sample, n=9

χ	γ
old	71
old	54
old	43
young	45
young	21
young	11
young	30
young	45
young	10

linear regression

Age vs. Money

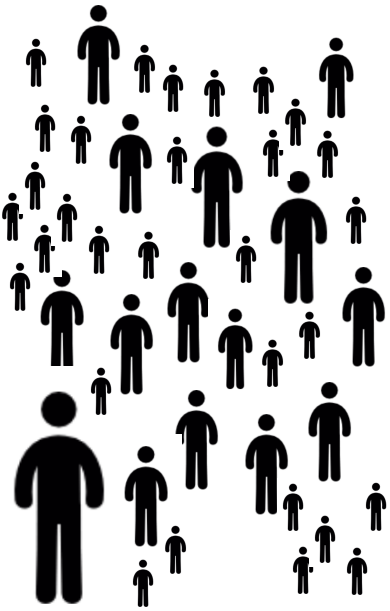
PREDICTOR *variable*

$X \longrightarrow$ Age in Years

RESPONSE *variable*

$Y \longrightarrow$ dollars ($) In bank account

Population



Population parameters

$\beta_0, \beta_1, \sigma^2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]
p-value = 0.036

Sample, n=9

x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

linear regression

Age vs. Money

fixed value

PREDICTOR *variable*

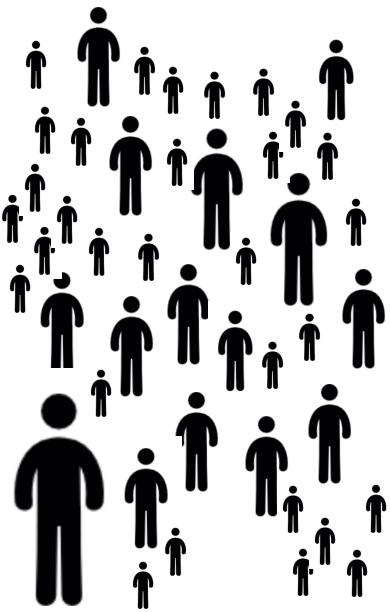
X ⟶ Age in Years

Random variable

RESPONSE *variable*

Y ⟶ dollars ($) In bank account

Population



Population parameters

$\beta_0, \beta_1, \sigma^2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

$b_0 = 17.7$

$b_1 = 0.55$

$s = 15.5$

$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]

p-value = 0.036

Sample, n=9

x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

Section 2.2 - Statistical linear regression model

What is a random variable?

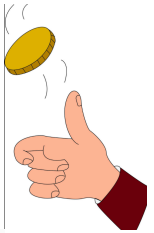
"A random variable, Y, is a variable whose possible values are numerical outcomes of a random phenomenon."

For a Random Variable, Y, we typically want to talk about the Expectation and Variance.

Example 1: $E[Y] = 3.5$ $\quad$ $Var(Y) = 2.92$

Example 2: $E[Y] = \theta$ $\quad$ $Var[Y] = \theta\,(1-\theta)$

Example 3: $E[Y] = \beta_0 + \beta_1 X$ $\quad$ $Var(Y) = \sigma^2$



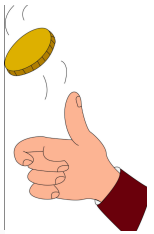
Questions?

Section 2.2 - Statistical linear regression model

What is a random variable?

"A random variable, Y, is a variable whose possible values are numerical outcomes of a random phenomenon."

We can ONLY talk about the Expectation and Variance of Y, if Y is a random variable.



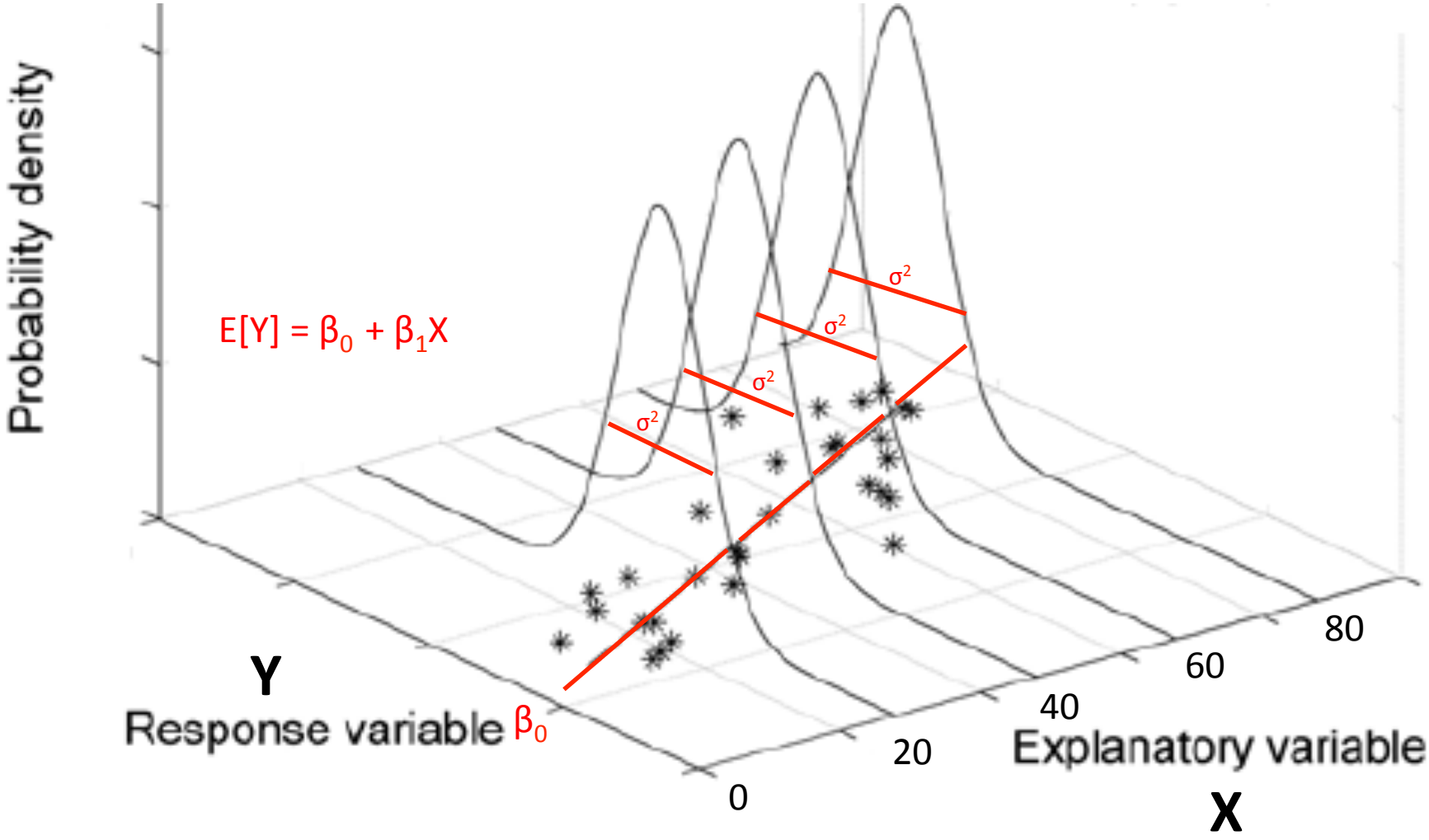
Example 1: $E[Y] = 3.5$ $Var(Y) = 2.92$

Example 2: $E[Y] = \theta$ $Var[Y] = \theta\,(1-\theta)$

Example 3: $E[Y] = \beta_0 + \beta_1 X$ $Var(Y) = \sigma^2$

Questions?

Section 2.2 - Statistical linear regression model

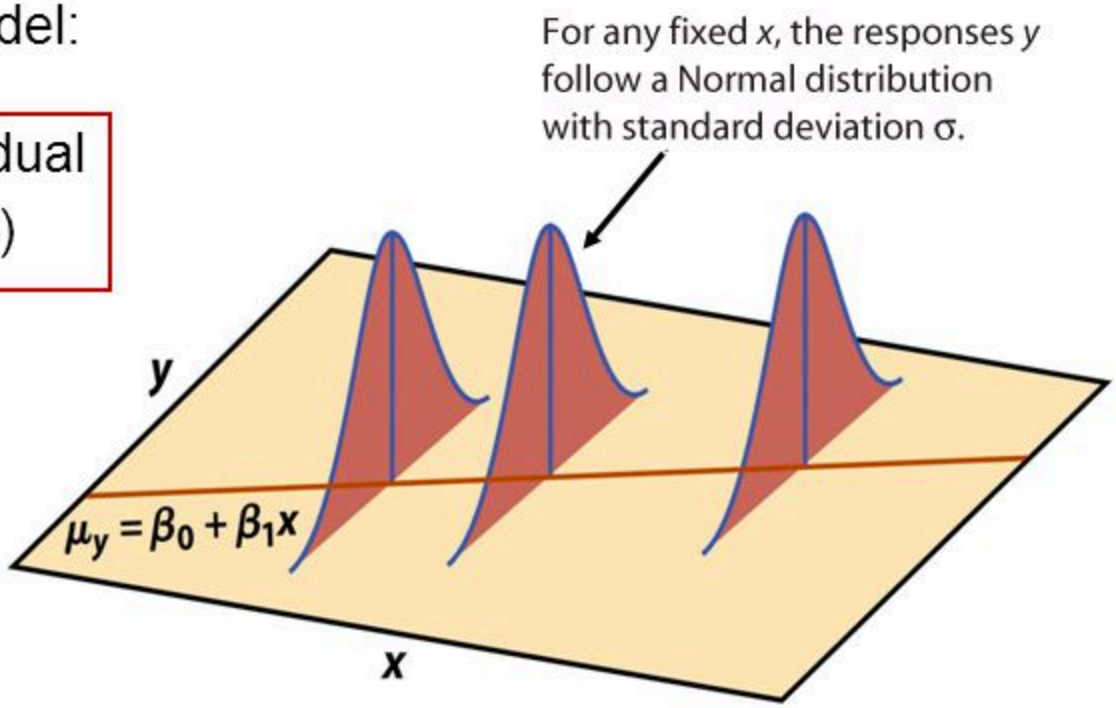


Simple linear regression model

In the population, the linear regression equation is $E(y) = \beta_0 + \beta_1 x$.

Sample data then fits the model:

Data =	fit	+	residual
y_i =	$(\beta_0 + \beta_1 x_i)$	+	(ε_i)



where the ε_i are **independent** and **Normally** distributed $N(0, \sigma)$.

Linear regression assumes **equal standard deviation of *y*** (σ is the same for all values of x).

Credit: Julien Stamp
http://slideplayer.com/slide/2342871/

Section 2.2 - Statistical linear regression model

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.
- $\hat{\theta}$ is a realization of a random variable $\hat{\Theta}$ assuming a probability model. Hence $\mathrm{E}(\hat{\Theta})$ and $\mathrm{Var}(\hat{\Theta})$ can be considered.
- A confidence interval for a parameter θ commonly has the form

$$\hat{\theta} \pm c \times se(\hat{\theta}),$$

where $se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$, and c depends on the confidence level. Typically $se(\hat{\theta}) = O(1/\sqrt{n})$ so that interval gets smaller as sample size n increases.

- Examples of 95% confidence intervals: with $\overline{y} = \hat{\mu}$,

$$\overline{y} \pm t_{n-1,0.975} \times se(\overline{y}); \quad se(\overline{y}) = s_y/\sqrt{n} \tag{2.34}$$

$$\hat{\pi} \pm z_{0.975} \times se(\hat{\pi}); \quad se(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})}/\sqrt{n}. \tag{2.35}$$

Here $z_{0.975}$ is the upper 0.975 quantile of the standard normal distribution and $t_{\nu,0.975}$ is the upper 0.975 quantile of the Student t distribution with degree of freedom parameter ν. (2.35) is based on an approximation that is valid for large n.

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.

$$\hat{\beta}_1 = b_1 = r_{xy} \frac{s_y}{s_x}$$

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.
- $\hat{\theta}$ is a realization of a random variable $\hat{\Theta}$ assuming a probability model. Hence $\mathrm{E}(\hat{\Theta})$ and $\mathrm{Var}(\hat{\Theta})$ can be considered.

"A random variable is a variable whose possible values are numerical outcomes of a random phenomenon."

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.
- $\hat{\theta}$ is a realization of a random variable $\hat{\Theta}$ assuming a probability model. Hence $\mathrm{E}(\hat{\Theta})$ and $\mathrm{Var}(\hat{\Theta})$ can be considered.

"A random variable is a variable whose possible values are numerical outcomes of a random phenomenon."

Consider the act of running an study as a "random phenomenon":

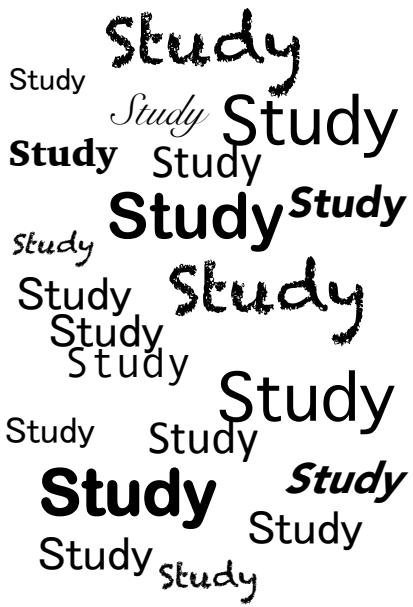
"A random variable is a variable whose possible values are numerical outcomes of a random phenomenon."

Consider the act of running an study as a "random phenomenon":

RANDOM variable

$\hat{\Theta}$ ⟶ Estimate from a study

Population



Sample

Study $\hat{\theta}_1$

Study $\hat{\theta}_2$

Study $\hat{\theta}_3$

Study $\hat{\theta}_4$

Study $\hat{\theta}_5$

Study $\hat{\theta}_6$

Hence we can consider:

$$\mathrm{E}[\hat{\Theta}]$$

$$\mathrm{Var}[\hat{\Theta}]$$

t-test

Age vs. Money

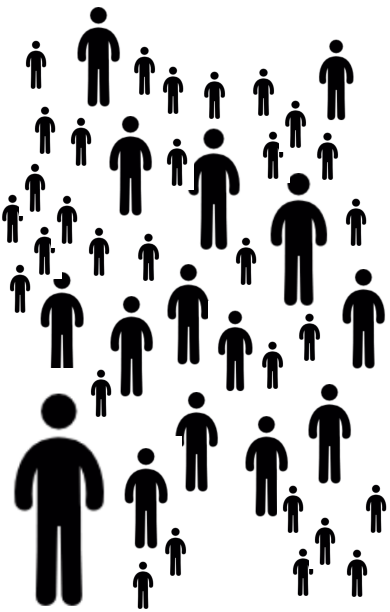
Dependent variable

χ → old (0)
χ → young (1)

Independent variable

γ → dollars ($)
In bank account

Population



Population parameters

μ_0 , μ_1 , σ^2

Hypothesis Test

$H_0 : \mu_0 = \mu_1$

$H_1 : \mu_0 \neq \mu_1$

Sample statistics

$\bar{y}_0 = 56$
$\bar{y}_1 = 27$
$\bar{y}_0 - \bar{y}_1 = 29$
$s_p = 10.81$

t = 2.68, *df* = 7
p-value = 0.03
95% C.I. = [3.4, 54.6]

Sample, n=9

χ	γ
old	71
old	54
old	43
young	45
young	21
young	11
young	30
young	45
young	10

t-test

Age vs. Money

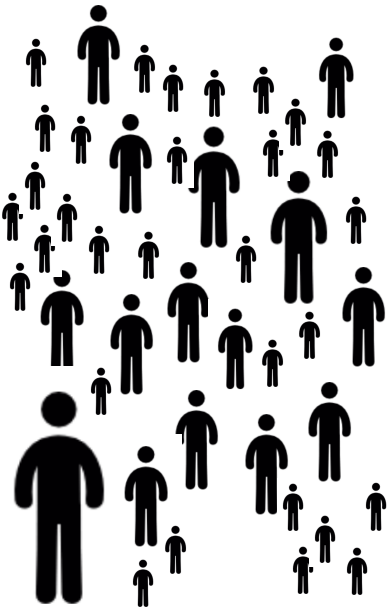
Dependent variable

χ → old (0)
χ → young (1)

Independent variable

γ → dollars ($)
In bank account

Population



Population parameters

μ_0 , μ_1 , σ^2

Hypothesis Test

$H_0 : \mu_0 = \mu_1$

$H_1 : \mu_0 \neq \mu_1$

Sample statistics

$\bar{y}_0 = 56$

Sample, n=9

χ	γ
old	71
old	54
old	43
young	45
young	21
young	11
young	30
young	45
young	10

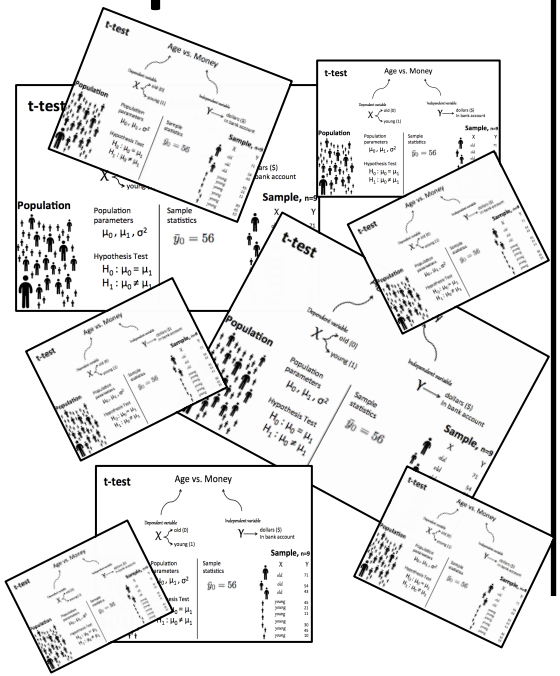
"A random variable is a variable whose possible values are numerical outcomes of a random phenomenon."

Consider the act of running an study as a "random phenomenon":

RANDOM *variable*

$\bar{Y}_0$ ⟶ Estimate from a study

Population



Sample

$$\bar{y}_0 = 46$$
$$\bar{y}_0 = 86$$
$$\bar{y}_0 = 36$$
$$\bar{y}_0 = 66$$
$$\bar{y}_0 = 55$$

Hence we can consider:

$$\mathrm{E}[\bar{Y}_0] = \mu_0$$

$$\mathrm{Var}[\bar{Y}_0] = \sigma^2/n$$

http://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.
- $\hat{\theta}$ is a realization of a random variable $\hat{\Theta}$ assuming a probability model. Hence $\mathrm{E}(\hat{\Theta})$ and $\mathrm{Var}(\hat{\Theta})$ can be considered.

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.
- $\hat{\theta}$ is a realization of a random variable $\hat{\Theta}$ assuming a probability model. Hence $\mathrm{E}(\hat{\Theta})$ and $\mathrm{Var}(\hat{\Theta})$ can be considered.
- A confidence interval for a parameter θ commonly has the form

$$\hat{\theta} \pm c \times se(\hat{\theta}),$$

where $se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$, and c depends on the confidence level. Typically $se(\hat{\theta}) = O(1/\sqrt{n})$ so that interval gets smaller as sample size n increases.

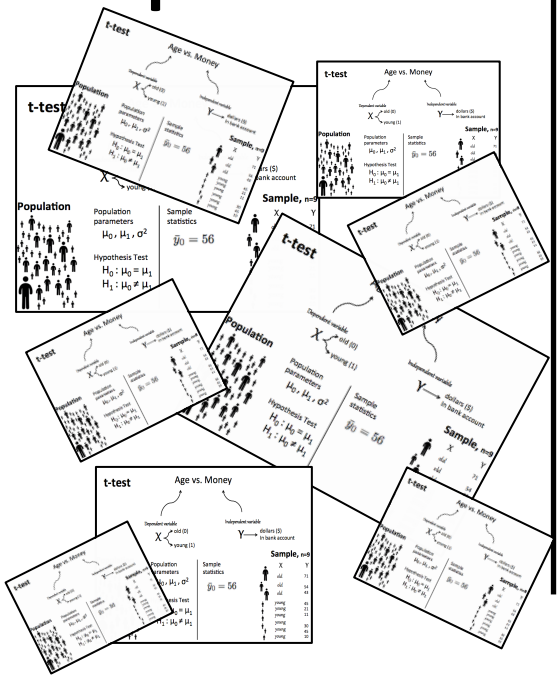
"A random variable is a variable whose possible values are numerical outcomes of a random phenomenon."

Consider the act of running an study as a "random phenomenon":

RANDOM *variable*

$\bar{Y}_0 \longrightarrow$ Estimate from a study

Population



Sample

$\bar{y}_0 = 46$

$\bar{y}_0 = 86$

$\bar{y}_0 = 36$

$\bar{y}_0 = 66$

$\bar{y}_0 = 55$

Hence we can consider:

$$\mathrm{E}[\bar{Y}_0] = \mu_0$$

$$\mathrm{Var}[\bar{Y}_0] = \sigma^2/n$$

$se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}\,(\hat{\Theta})}$

$$\mathrm{se}(\bar{y}_0) = s/\sqrt{n}$$

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.
- $\hat{\theta}$ is a realization of a random variable $\hat{\Theta}$ assuming a probability model. Hence $\mathrm{E}(\hat{\Theta})$ and $\mathrm{Var}(\hat{\Theta})$ can be considered.
- A confidence interval for a parameter θ commonly has the form

$$\hat{\theta} \pm c \times se(\hat{\theta}),$$

where $se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$, and c depends on the confidence level. Typically $se(\hat{\theta}) = O(1/\sqrt{n})$ so that interval gets smaller as sample size n increases.

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.
- $\hat{\theta}$ is a realization of a random variable $\hat{\Theta}$ assuming a probability model. Hence $\mathrm{E}(\hat{\Theta})$ and $\mathrm{Var}(\hat{\Theta})$ can be considered.
- A confidence interval for a parameter θ commonly has the form

$$\hat{\theta} \pm c \times se(\hat{\theta}),$$

where $se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$, and c depends on the confidence level. Typically $se(\hat{\theta}) = O(1/\sqrt{n})$ so that interval gets smaller as sample size n increases.

The "standard error" is the error of the sample statistic with respect to estimating the population parameter. If **n** is very large, then **se** is very small. As n increases linearly, **se** decreases by 1/sqrt(n).

https://en.wikipedia.org/wiki/Standard_error

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.
- $\hat{\theta}$ is a realization of a random variable $\hat{\Theta}$ assuming a probability model. Hence $\mathrm{E}(\hat{\Theta})$ and $\mathrm{Var}(\hat{\Theta})$ can be considered.
- A confidence interval for a parameter θ commonly has the form

$$\hat{\theta} \pm c \times se(\hat{\theta}),$$

where $se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$, and c depends on the confidence level. Typically $se(\hat{\theta}) = O(1/\sqrt{n})$ so that interval gets smaller as sample size n increases.

Typically the "confidence level" is 95% or 80% or 90%.

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.
- $\hat{\theta}$ is a realization of a random variable $\hat{\Theta}$ assuming a probability model. Hence $\mathrm{E}(\hat{\Theta})$ and $\mathrm{Var}(\hat{\Theta})$ can be considered.
- A confidence interval for a parameter θ commonly has the form

$$\hat{\theta} \pm c \times se(\hat{\theta}),$$

  where $se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$, and c depends on the confidence level. Typically $se(\hat{\theta}) = O(1/\sqrt{n})$ so that interval gets smaller as sample size n increases.
- Examples of 95% confidence intervals: with $\overline{y} = \hat{\mu}$,

$$\overline{y} \pm t_{n-1,0.975} \times se(\overline{y}); \quad se(\overline{y}) = s_y/\sqrt{n} \tag{2.34}$$

$$\hat{\pi} \pm z_{0.975} \times se(\hat{\pi}); \quad se(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})}/\sqrt{n}\,. \tag{2.35}$$

  Here $z_{0.975}$ is the upper 0.975 quantile of the standard normal distribution and $t_{\nu,0.975}$ is the upper 0.975 quantile of the Student t distribution with degree of freedom parameter ν. (2.35) is based on an approximation that is valid for large n.

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.
- $\hat{\theta}$ is a realization of a random variable $\hat{\Theta}$ assuming a probability model. Hence $\mathrm{E}(\hat{\Theta})$ and $\mathrm{Var}(\hat{\Theta})$ can be considered.
- A confidence interval for a parameter θ commonly has the form

$$\hat{\theta} \pm c \times se(\hat{\theta}),$$

  where $se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$, and c depends on the confidence level. Typically $se(\hat{\theta}) = O(1/\sqrt{n})$ so that interval gets smaller as sample size n increases.
- Examples of 95% confidence intervals: with $\overline{y} = \hat{\mu}$,

$$\overline{y} \pm t_{n-1,0.975} \times se(\overline{y}); \quad se(\overline{y}) = s_y/\sqrt{n} \tag{2.34}$$

$$\hat{\pi} \pm z_{0.975} \times se(\hat{\pi}); \quad se(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})}/\sqrt{n}. \tag{2.35}$$

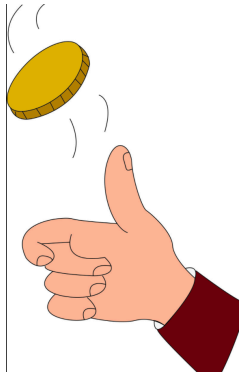
  Here $z_{0.975}$ is the upper 0.975 quantile of the standard normal distribution and $t_{\nu,0.975}$ is the upper 0.975 quantile of the Student t distribution with degree of freedom parameter ν. (2.35) is based on an approximation that is valid for large n.

Section 2.2 - Statistical linear regression model

Example 2: All we can know about Y is that:

Pr(Y=0) = 1-θ

Pr(Y=1) = θ



From the notes:

$\hat{\pi} \pm z_{0.975} \times se(\hat{\pi}); \quad se(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})}/\sqrt{n}, \ \mathrm{Var}(\hat{\Pi}) = \pi(1-\pi)/n.$

Rcode:

http://www.r-tutor.com/elementary-statistics/interval-estimation/interval-estimate-population-proportion

Population

Population parameters

θ

Hypothesis Test

$H_0 : \theta = 0.5$

$H_1 : \theta \neq 0.5$

Is the coin fair?

Sample statistics

$\hat{\theta} = 60/90 = 0.667$

We can also think of $\hat{\Theta}$ as a Random variable…. so:

$$\mathrm{Var}[\hat{\Theta}] = \theta(1-\theta)/n$$

Sample, n=90

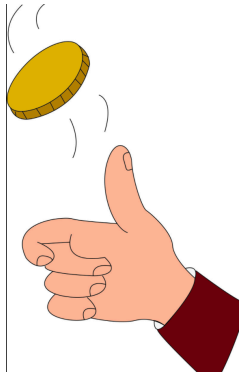
	y
heads	0
tails	1
heads	0
tails	1
tails	1
heads	0
tails	1
…	
tails	1
tails	1

Section 2.2 - Statistical linear regression model

Example 2: All we can know about Y is that:

Pr(Y=0) = 1-θ

Pr(Y=1) = θ



From the notes:

$\hat{\pi} \pm z_{0.975} \times se(\hat{\pi}); \quad se(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})}/\sqrt{n}, \ \mathrm{Var}(\hat{\Pi}) = \pi(1-\pi)/n.$

Rcode:

http://www.r-tutor.com/elementary-statistics/interval-estimation/interval-estimate-population-proportion

Population

Population parameters

θ

Hypothesis Test

$H_0 : \theta = 0.5$

$H_1 : \theta \neq 0.5$

Is the coin fair?

Sample, n=90

Sample statistics

$\hat{\theta} = 60/90 = 0.667$

We can also think of $\hat{\Theta}$ as a Random variable…. so:

$\mathrm{Var}[\hat{\Theta}] = \theta(1-\theta)/n$

$se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$

$se(\hat{\theta}) = \sqrt{\hat{\theta}(1-\hat{\theta})}/\sqrt{n}$

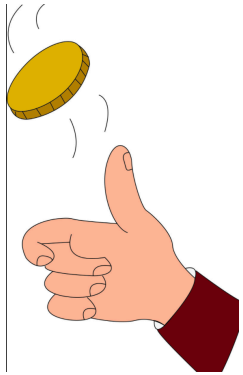
	y
heads	0
tails	1
heads	0
tails	1
tails	1
heads	0
tails	1
…	
tails	1
tails	1

Section 2.2 - Statistical linear regression model

Example 2: All we can know about Y is that:

Pr(Y=0) = 1-θ

Pr(Y=1) = θ



From the notes:

$\hat{\pi} \pm z_{0.975} \times se(\hat{\pi})$; $se(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})}/\sqrt{n}$, $\mathrm{Var}(\hat{\Pi}) = \pi(1-\pi)/n$.

Rcode:

http://www.r-tutor.com/elementary-statistics/interval-estimation/interval-estimate-population-proportion

Population

Population parameters

θ

Hypothesis Test

$H_0 : \theta = 0.5$

$H_1 : \theta \neq 0.5$

Is the coin fair?

Sample, n=90

Sample statistics

$$\hat{\theta} = 60/90 = 0.667$$

$$se(\hat{\theta}) = \sqrt{\hat{\theta}(1-\hat{\theta})}/\sqrt{n}$$

= 0.05

Using the standard error, we can construct a 95% confidence interval:

$$[\hat{\theta} - z_{0.975}se(\hat{\theta}), \hat{\theta} + z_{0.975}se(\hat{\theta})]$$

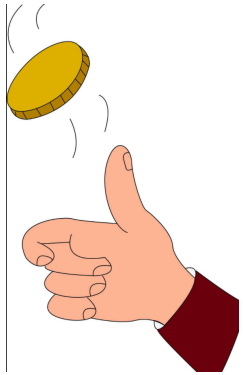
	y
heads	0
tails	1
heads	0
tails	1
tails	1
heads	0
tails	1
…	
tails	1
tails	1

Section 2.2 - Statistical linear regression model

Example 2: All we can know about Y is that:

Pr(Y=0) = 1-θ

Pr(Y=1) = θ



From the notes:

$\hat{\pi} \pm z_{0.975} \times se(\hat{\pi})$; $se(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})}/\sqrt{n}$, $\mathrm{Var}(\hat{\Pi}) = \pi(1-\pi)/n$.

Rcode:

http://www.r-tutor.com/elementary-statistics/interval-estimation/interval-estimate-population-proportion

Population

Population parameters

θ

Hypothesis Test

$H_0 : \theta = 0.5$

$H_1 : \theta \neq 0.5$

Is the coin fair?

Sample, n=90

Sample statistics

$\hat{\theta} = 60/90 = 0.667$

$se(\hat{\theta}) = \sqrt{\hat{\theta}(1-\hat{\theta})}/\sqrt{n}$

= 0.05

For 95% C.I. :

$[\hat{\theta} - z_{0.975} se(\hat{\theta}), \hat{\theta} + z_{0.975} se(\hat{\theta})]$

= [0.569, 0.764]

	y
heads	0
tails	1
heads	0
tails	1
tails	1
heads	0
tails	1
…	
tails	1
tails	1

- Questions?

Background on interval estimation

An interval estimate is an interval of plausible values for unknown population quantity, computed based on the observed data. A longer interval has a large confidence level.

- θ is generic (scalar) parameter to be estimated; for example, population mean μ, proportion π, regression coefficient β.
- $\hat{\theta}$ is an estimator of θ based on the data; for example, $\hat{\mu} = \overline{y}$.
- $\hat{\theta}$ is a realization of a random variable $\hat{\Theta}$ assuming a probability model. Hence $\mathrm{E}(\hat{\Theta})$ and $\mathrm{Var}(\hat{\Theta})$ can be considered.
- A confidence interval for a parameter θ commonly has the form

$$\hat{\theta} \pm c \times se(\hat{\theta}),$$

  where $se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$, and c depends on the confidence level. Typically $se(\hat{\theta}) = O(1/\sqrt{n})$ so that interval gets smaller as sample size n increases.
- Examples of 95% confidence intervals: with $\overline{y} = \hat{\mu}$,

$$\overline{y} \pm t_{n-1,0.975} \times se(\overline{y}); \quad se(\overline{y}) = s_y/\sqrt{n} \tag{2.34}$$

$$\hat{\pi} \pm z_{0.975} \times se(\hat{\pi}); \quad se(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})}/\sqrt{n}. \tag{2.35}$$

  Here $z_{0.975}$ is the upper 0.975 quantile of the standard normal distribution and $t_{\nu,0.975}$ is the upper 0.975 quantile of the Student t distribution with degree of freedom parameter ν. (2.35) is based on an approximation that is valid for large n.

linear regression

Age vs. Money

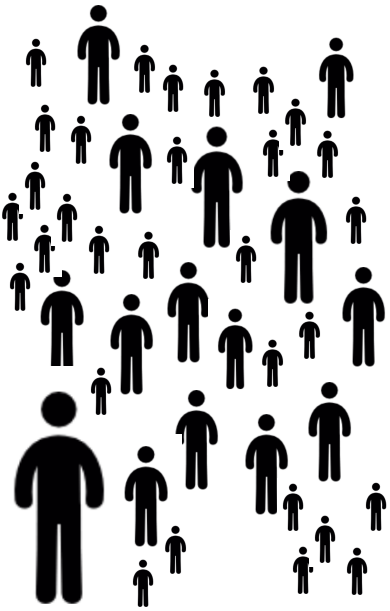
PREDICTOR *variable*

χ ⟶ Age in Years

RESPONSE *variable*

γ ⟶ dollars ($) In bank account

Population



Population parameters

β_0 , β_1 , σ^2

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

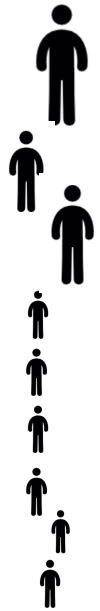
Sample statistics

b_0 = 17.7
b_1 = 0.55
s = 15.5
R^2 = 0.49

For parameter β_1 :

95% C.I. = [0.05, 1.05]
p-value = 0.036

Sample, n=9



χ	γ
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

linear regression

Age vs. Money

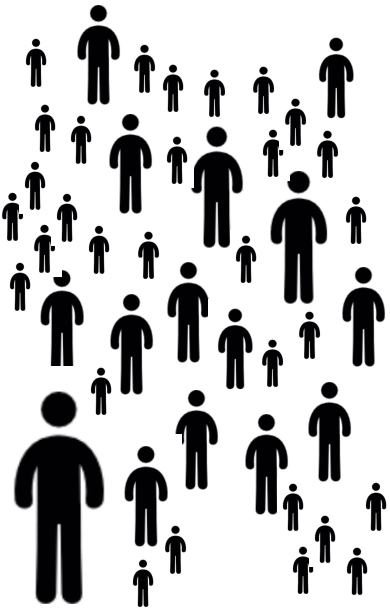
PREDICTOR *variable*

X ⟶ Age in Years

RESPONSE *variable*

Y ⟶ dollars ($) In bank account

Population



Population parameters

$\beta_0, \beta_1, \sigma^2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]

p-value = 0.036

Sample, n=9

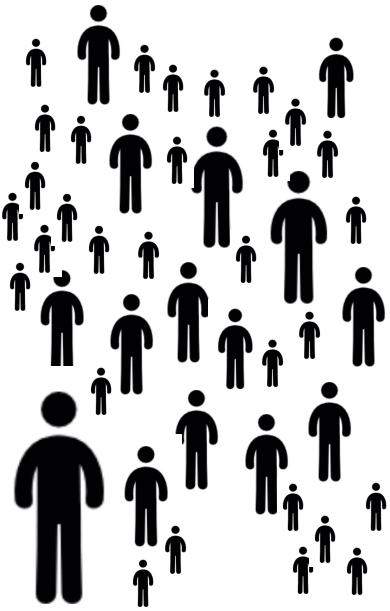
X	Y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

linear regression

Steps to get 95% C.I. for b_1

1. Consider the sample statistic b_1 as the random variable B_1
2. Determine $Var[B_1]$
3. Define $se(b_1)$ as an estimate of $sqrt(Var(B_1))$
4. 95% C.I. = $[\, b_1 - c^*se(b_1) \,,\, b_1 + c^*se(b_1) \,]$

Population



Population parameters

$\beta_0 , \beta_1 , \sigma^2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]

p-value = 0.036

Sample, n=9

X	Y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

2.5 Intervals for simple linear regression

The following is an inference result from introductory statistics (but the proof is given in a mathematical statistics course). If the values $y_1, \ldots, y_n$ are realizations from a random sample from a normal population with mean μ and variance σ^2, then the 95% confidence interval for μ is:

(2.39) $$\overline{y} \pm t_{n-1,0.975} \times se(\overline{y}); \quad se(\overline{y}) = s/\sqrt{n},$$

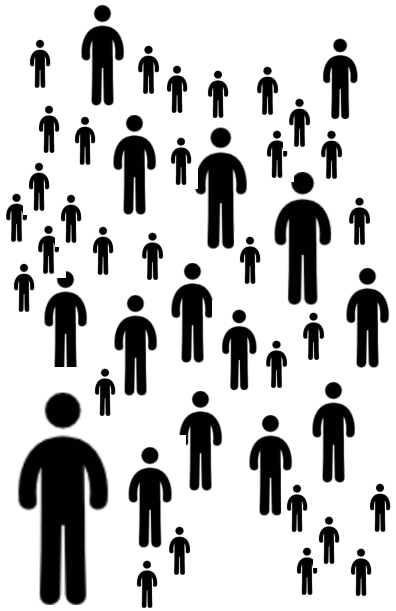
where $t_{\nu,0.975}$ is the 0.975 quantile of the Student t distribution with degree of freedom parameter ν. This means area to the left of $t_{\nu,0.975}$ under the t_ν density curve is 0.975. *Some statistics books denote this critical value as $t_{0.025,\nu}$, to indicate the area to the right is 0.025.*

“from introductory statistics…”

RANDOM *variable*

$$Y \sim Normal(\mu, \sigma^2)$$

Population



Population parameters

μ , σ^2

Hypothesis Test

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

Sample statistics

$$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$$

$$s = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)}.$$

We can also think of $\bar{Y}$ as a Random variable…. so:

$$\mathrm{Var}[\bar{Y}] = \sigma^2 / n$$

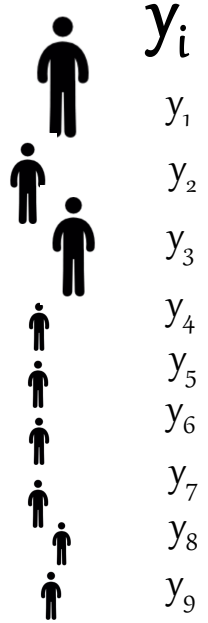
$se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$

$$se(\bar{y}) = s/\sqrt{n},$$

the 95% confidence interval for μ is:

$$\bar{y} \pm t_{n-1,0.975} \times se(\bar{y});$$

Sample, n=9

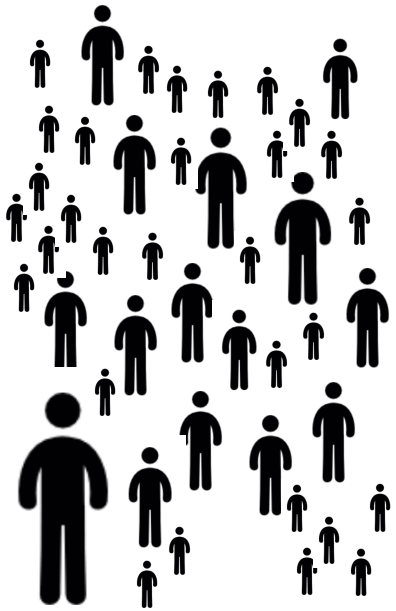


"from introductory statistics…"

RANDOM *variable*

$$Y \sim Normal(\mu, \sigma^2)$$

Population



Population parameters

μ , σ^2

Hypothesis Test

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

Sample statistics

$$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$$

$$s = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)}.$$

Step1

We can also think of $\bar{Y}$ as a Random variable…. so:

Step2

$$\mathrm{Var}[\bar{Y}] = \sigma^2 / n$$

Step3

$se(\hat{\theta})$ is an estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$

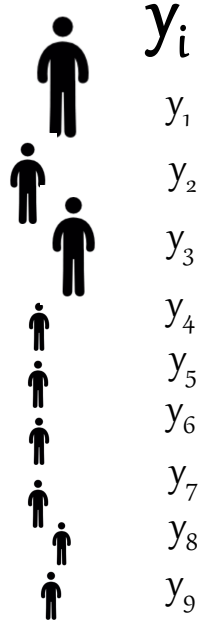
$$se(\bar{y}) = s/\sqrt{n},$$

Step4

the 95% confidence interval for μ is:

$$\bar{y} \pm t_{n-1,0.975} \times se(\bar{y});$$

Sample, n=9



2.5 Intervals for simple linear regression

The following is an inference result from introductory statistics (but the proof is given in a mathematical statistics course). If the values $y_1, \ldots, y_n$ are realizations from a random sample from a normal population with mean μ and variance σ^2, then the 95% confidence interval for μ is:

(2.39) $$\overline{y} \pm t_{n-1,0.975} \times se(\overline{y}); \quad se(\overline{y}) = s/\sqrt{n},$$

where $t_{\nu,0.975}$ is the 0.975 quantile of the Student t distribution with degree of freedom parameter ν. This means area to the left of $t_{\nu,0.975}$ under the t_ν density curve is 0.975. *Some statistics books denote this critical value as $t_{0.025,\nu}$, to indicate the area to the right is 0.025.*

Rcode

2.5.2 Derivations

linear regression

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1

2. Determine Var[B_1]

3. Define se(b_1) as an estimate of sqrt(Var(B_1))

4. 95% C.I. for parameter β_1 = [b_1 - c*se(b_1) , b_1 + c*se(b_1)]

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1
2. Determine Var[B_1]
3. Define se(b_1) as an estimate of sqrt(Var(B_1))
4. 95% C.I. = [b_1 - c*se(b_1) , b_1 + c*se(b_1)]

2.5.2 Derivations

The standard errors come from the variances when the estimators are considered as random variables. $\hat{\beta}_1$ as a random variable $\hat{B}_1$ is:

(2.50)
$$\hat{B}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{(n-1)s_x^2} = \sum_{i=1}^{n} a_i Y_i,$$

where

(2.51)
$$a_i = (x_i - \overline{x})/[(n-1)s_x^2].$$

2.5.2 Derivations

$$b_1 = r_{xy}\frac{s_y}{s_x}$$

Recall: $r_{xy} = \frac{s_{xy}}{s_x s_y}$

$$= \frac{s_{xy}}{s_x s_y}\frac{s_y}{s_x}$$

Recall: $s_{xy} = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x^2}$$

$$= \frac{\sum_{i=1}^{n}(y_i)(x_i - \bar{x})}{(n-1)s_x^2} - \frac{\sum_{i=1}^{n}(\bar{y})(x_i - \bar{x})}{(n-1)s_x^2}$$

$$= \frac{\sum_{i=1}^{n}(y_i)(x_i - \bar{x})}{(n-1)s_x^2} - \bar{y}\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{(n-1)s_x^2}$$

2.5.2 Derivations

$$b_1 = r_{xy}\frac{s_y}{s_x}$$

Recall: $r_{xy} = \frac{s_{xy}}{s_x s_y}$

$$= \frac{s_{xy}}{s_x s_y}\frac{s_y}{s_x}$$

Recall: $s_{xy} = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x^2}$$

$$= \frac{\sum_{i=1}^{n}(y_i)(x_i-\bar{x})}{(n-1)s_x^2} - \frac{\sum_{i=1}^{n}(\bar{y})(x_i-\bar{x})}{(n-1)s_x^2}$$

$$= \frac{\sum_{i=1}^{n}(y_i)(x_i-\bar{x})}{(n-1)s_x^2} - \bar{y}\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{(n-1)s_x^2}$$

0

2.5.2 Derivations

$$b_1 = r_{xy}\frac{s_y}{s_x}$$

$$= \frac{\sum_{i=1}^{n}(y_i)(x_i - \bar{x})}{(n-1)s_x^2}$$

$$= \sum_{i=1}^{n}\frac{(x_i - \bar{x})}{(n-1)s_x^2}(y_i)$$

2.5.2 Derivations

$$b_1 = r_{xy}\frac{s_y}{s_x}$$

$$= \frac{\sum_{i=1}^{n}(y_i)(x_i-\bar{x})}{(n-1)s_x^2}$$

$$= \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{(n-1)s_x^2}(y_i)$$

Step 1. Consider the sample statistic b_1 as the random variable B_1 :

$$B_1 = \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{(n-1)s_x^2}(Y_i)$$

$$= \sum_{i=1}^{n} a_i Y_i \quad \text{, where:} \quad a_i = \frac{(x_i-\bar{x})}{(n-1)s_x^2}$$

2.5.2 Derivations

2.5.2 Derivations

The standard errors come from the variances when the estimators are considered as random variables. $\hat{\beta}_1$ as a random variable $\hat{B}_1$ is:

(2.50) $$\hat{B}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{(n-1)s_x^2} = \sum_{i=1}^{n} a_i Y_i,$$

where

(2.51) $$a_i = (x_i - \overline{x})/[(n-1)s_x^2].$$

Step 1. Consider the sample statistic b_1 as the random variable B_1 :

$$B_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{(n-1)s_x^2}(Y_i)$$

$$= \sum_{i=1}^{n} a_i Y_i \quad \text{, where:} \quad a_i = \frac{(x_i - \bar{x})}{(n-1)s_x^2}$$

2.5.2 Derivations

Step 1. Consider the sample statistic b_1 as the random variable B_1 :

$$B_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{(n-1)s_x^2}(Y_i)$$

$$= \sum_{i=1}^{n} a_i Y_i \quad \text{, where:} \quad a_i = \frac{(x_i - \bar{x})}{(n-1)s_x^2}$$

Step 2. Determine Var[B_1]

First, recall that for random variable Y_i, we have:

$$(2.33) \qquad Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

2.5.2 Derivations

Step 2. Determine Var[B_1]

First, recall that for random variable Y_i, we have:

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2). \tag{2.33}$$

2.5.2 Derivations

Step 2. Determine Var[B_1]

First, recall that for random variable Y_i, we have:

$$Y_i \sim N(\beta_0 + \beta_1 x_i,\, \sigma^2). \tag{2.33}$$

Then from page 28 of the course notes, we have solutions for Var[B_1] and E[B_1] :

$\hat{B}_1 = \sum_{i=1}^{n} a_i Y_i$ as a random variable using $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ with a_i as in (2.51).

$$\mathrm{E}(\hat{B}_1) = \sum_{i=1}^{n} a_i \mathrm{E}(Y_i) = \sum_{i=1}^{n} a_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} a_i + \beta_1 \sum_{i=1}^{n} a_i x_i \tag{2.56}$$

$$= 0 + \beta_1 \sum_{i=1}^{n} \frac{(x_i - \overline{x})x_i}{[(n-1)s_x^2]} = \beta_1, \tag{2.57}$$

$$\mathrm{Var}(\hat{B}_1) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} a_i^2 = \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{[(n-1)s_x^2]^2} \tag{2.58}$$

$$= \frac{\sigma^2}{(n-1)s_x^2}. \tag{2.59}$$

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1

2. **Determine Var[B_1]** $= \frac{\sigma^2}{(n-1)s_x^2}.$

3. Define se(b_1) as an estimate of sqrt(Var(B_1))

4. 95% C.I. = [b_1 - c*se(b_1) , b_1 + c*se(b_1)]

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic b_1 as the random variable B_1

2. Determine $\text{Var}[B_1] = \frac{\sigma^2}{(n-1)s_x^2}$.

3. **Define $se(b_1)$ as an estimate of $\text{sqrt}(\text{Var}(B_1))$**

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}\, s_x}$$

where:

$$\hat{\sigma} = \text{residual SD} = \left\{(n-2)^{-1}\sum_{i=1}^{n} e_i^2\right\}^{1/2}$$

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1
2. Determine Var[B_1] $= \frac{\sigma^2}{(n-1)s_x^2}$.
3. Define se(b_1) as an estimate of sqrt(Var(B_1)) : $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}\, s_x}$
4. **95% C.I. = [b_1 - c*se(b_1) , b_1 + c*se(b_1)]**

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1
2. Determine Var[B_1] $= \frac{\sigma^2}{(n-1)s_x^2}$.
3. Define se(b_1) as an estimate of sqrt(Var(B_1)) : $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}\, s_x}$
4. **95% C.I. = [b_1 - c*se(b_1) , b_1 + c*se(b_1)]**

we take c = $t_{n-2,0.975}$

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1
2. Determine Var[B_1] $= \frac{\sigma^2}{(n-1)s_x^2}$.
3. Define se(b_1) as an estimate of sqrt(Var(B_1)) : $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}\, s_x}$
4. **95% C.I. = [b_1 - c*se(b_1) , b_1 + c*se(b_1)]**

we take c = $t_{n-2,0.975}$

Then we have :

95% C.I. for β_1 : $[b_1 - t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}, \quad b_1 + t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}]$

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1
2. Determine Var[B_1] $= \frac{\sigma^2}{(n-1)s_x^2}$.
3. Define se(b_1) as an estimate of sqrt(Var(B_1)) : $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}\, s_x}$
4. **95% C.I. = [b_1 - c*se(b_1) , b_1 + c*se(b_1)]**

we take c = $t_{n-2,0.975}$

Then we have :

95% C.I. for β_1 : $[b_1 - t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}, \quad b_1 + t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}]$

where: $\hat{\sigma} = \text{residual SD} = \left\{(n-2)^{-1}\sum_{i=1}^{n} e_i^2\right\}^{1/2}$ (also known as "s")

linear regression

Age vs. Money

fixed value

PREDICTOR *variable*

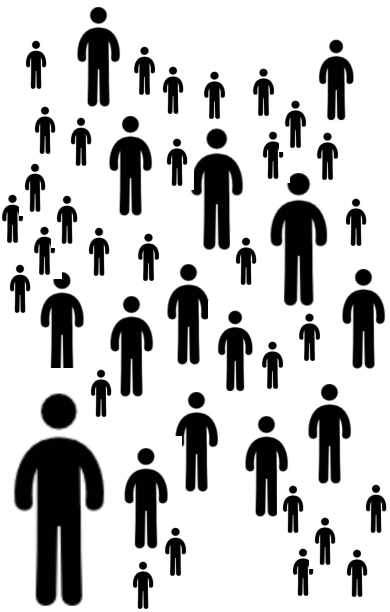
X ⟶ Age in Years

Random variable

RESPONSE *variable*

Y ⟶ dollars ($) In bank account

Population



Population parameters

$\beta_0, \beta_1, \sigma^2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

$b_0 = 17.7$

$b_1 = 0.55$

$s = 15.5$

$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]

p-value = 0.036

Sample, n=9

x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

linear regression

Age vs. Money

fixed value

PREDICTOR *variable*

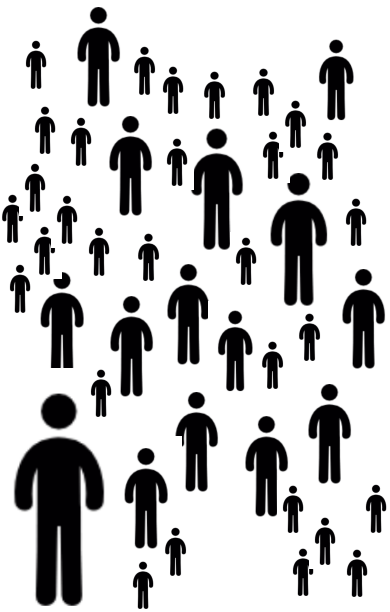
X ⟶ Age in Years

Random variable

RESPONSE *variable*

Y ⟶ dollars ($) In bank account

Population



Population parameters

$\beta_0, \beta_1, \sigma^2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

$b_0 = 17.7$

$b_1 = 0.55$

$s = 15.5$

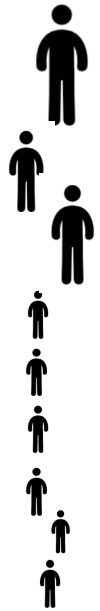
$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]

p-value = 0.036

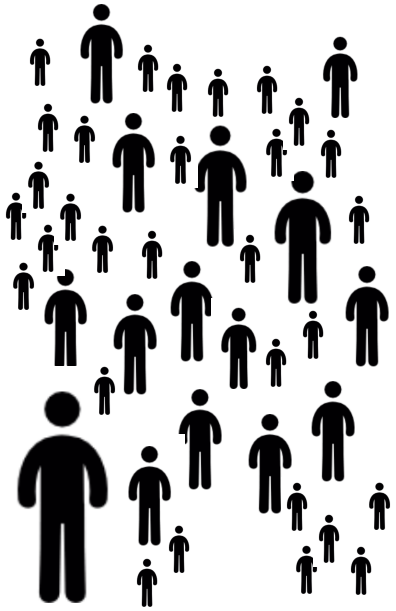
Sample, n=9



x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

Age vs. Money

Population



Population parameters

$\beta_0, \beta_1, \sigma^2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

$b_0 = 17.7$

$b_1 = 0.55$

$s = 15.5$

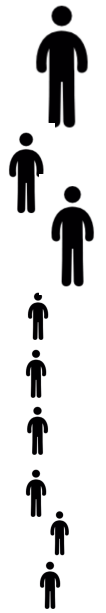
$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]

p-value = 0.036

Sample, n=9



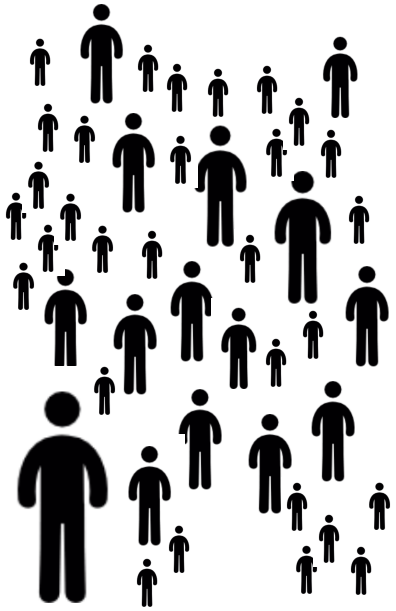
x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

s = $\hat{\sigma}$ = residual SD = $\left\{(n-2)^{-1}\sum_{i=1}^{n} e_i^2\right\}^{1/2}$

```
> residuals <- y - b0_hat - b1_hat*x
> s <- sqrt( (1/(n-2))*sum(residuals^2) )
> s
[1] 15.5308
```

Age vs. Money

Population



Population parameters

$\beta_0, \beta_1, \sigma^2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

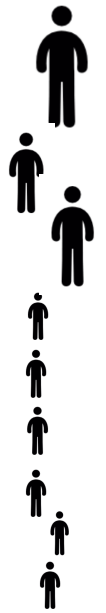
Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]
p-value = 0.036

Sample, n=9



x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

95% C.I. for b_1 = $[b_1 - t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}, \quad b_1 + t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}]$

```
> b1_CI95 <- c(b1 - qt(0.975,n-2)*(s/(sqrt(n-1)*sx)) , b1 + qt(0.975,n-2)*(s/(sqrt(n-1)*sx)))
> b1_CI95
[1] 0.04946904 1.04676990
```

linear regression

Age vs. Money

fixed value

PREDICTOR *variable*

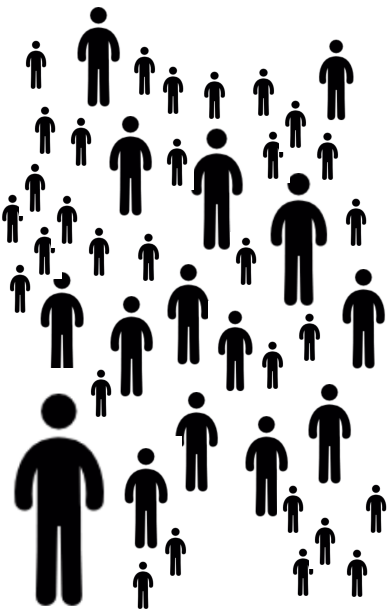
X ⟶ Age in Years

Random variable

RESPONSE *variable*

Y ⟶ dollars ($) In bank account

Population



Population parameters

$\beta_0, \beta_1, \sigma^2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

$b_0 = 17.7$

$b_1 = 0.55$

$s = 15.5$

$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]

p-value = 0.036

Sample, n=9

x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

- Questions?

2.5.2 Derivations

Question:

What is the $\text{subpopulation mean } \mu_Y(x) = \beta_0 + \beta_1 x$?

2.5.2 Derivations

Question:

What is the $\text{subpopulation mean } \mu_Y(x) = \beta_0 + \beta_1 x$?

Answer:

The average amount of money (i.e. the expectation of random variable Y), among people aged “x” years old.

2.5.2 Derivations

Question:

What is the subpopulation mean $\mu_Y(x) = \beta_0 + \beta_1 x$?

Answer:

The average amount of money (i.e. the expectation of random variable Y), among people aged "x" years old.

Steps to get 95% C.I. for the subpopulation mean:

1. Consider the subpopulation mean, $\mu_Y(x)$, as a random variable.
2. Determine Var[$\mu_Y(x)$]
3. Define the standard error, $se(\hat{\mu}_Y(x))$, as an estimate of sqrt(Var($\mu_Y(x)$)) :
4. **95% C.I. = [** $\hat{\mu}_Y(x)$ **– c*** $se(\hat{\mu}_Y(x))$**,** $\hat{\mu}_Y(x)$ **+ c*** $se(\hat{\mu}_Y(x))$ **]**

2.5.2 Derivations

Steps to get 95% C.I. for the subpopulation mean:

1. Consider the subpopulation mean, $\mu_Y(x)$, as a random variable.

$\hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ as a random variable is:

$$(2.52) \quad \hat{\mu}_Y(x) = (\overline{Y} - \hat{B}_1\overline{x}) + \hat{B}_1 x = \overline{Y} + \hat{B}_1(x - \overline{x}) = n^{-1}\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} a_i(x - \overline{x})Y_i = \sum_{i=1}^{n} c_i Y_i,$$

where

$$(2.53) \qquad c_i = n^{-1} + a_i(x - \overline{x}) = n^{-1} + \frac{(x - \overline{x})(x_i - \overline{x})}{(n-1)s_x^2}.$$

2.5.2 Derivations

Steps to get 95% C.I. for the subpopulation mean:

1. Consider the subpopulation mean, $\mu_Y(x)$, as a random variable.

$\hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ as a random variable is:

$$(2.52) \quad \hat{\mu}_Y(x) = (\overline{Y} - \hat{B}_1\overline{x}) + \hat{B}_1 x = \overline{Y} + \hat{B}_1(x - \overline{x}) = n^{-1}\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} a_i(x - \overline{x})Y_i = \sum_{i=1}^{n} c_i Y_i,$$

where

$$(2.53) \qquad c_i = n^{-1} + a_i(x - \overline{x}) = n^{-1} + \frac{(x - \overline{x})(x_i - \overline{x})}{(n-1)s_x^2}.$$

2. Determine Var[$\mu_Y(x)$]

2.5.2 Derivations

Steps to get 95% C.I. for the subpopulation mean:

1. Consider the subpopulation mean, $\mu_Y(x)$, as a random variable.

2. **Determine Var[$\mu_Y(x)$]**

$$
\begin{aligned}
(2.63)\quad \mathrm{Var}\left[\hat{\mu}_Y(x)\right] &= \sum_{i=1}^{n} c_i^2 \mathrm{Var}\left(Y_i\right) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \sum_{i=1}^{n} \left\{ n^{-1} + (x-\overline{x})(x_i-\overline{x})/[(n-1)s_x^2] \right\}^2 \\
(2.64)\quad &= \sigma^2 \sum_{i=1}^{n} \left\{ n^{-2} + \frac{(x-\overline{x})^2 (x_i-\overline{x})^2}{[(n-1)s_x^2]^2} + 2n^{-1} \frac{(x-\overline{x})(x_i-\overline{x})}{[(n-1)s_x^2]} \right\} \\
(2.65)\quad &= \sigma^2 \left\{ n^{-1} + \frac{(x-\overline{x})^2 \sum_i (x_i-\overline{x})^2}{[(n-1)s_x^2]^2} + 0 \right\} \\
(2.66)\quad &= \sigma^2 \left\{ n^{-1} + \frac{(x-\overline{x})^2}{[(n-1)s_x^2]} \right\}.
\end{aligned}
$$

2.5.2 Derivations

Steps to get 95% C.I. for the subpopulation mean:

1. Consider the subpopulation mean, $\mu_Y(x)$, as a random variable.

2. Determine Var[$\mu_Y(x)$] $= \sigma^2\left\{n^{-1} + \frac{(x-\overline{x})^2}{[(n-1)s_x^2]}\right\}$.

$$(2.63)\quad \mathrm{Var}\left[\hat{\mu}_Y(x)\right] = \sum_{i=1}^{n} c_i^2 \mathrm{Var}\left(Y_i\right) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \sum_{i=1}^{n} \left\{n^{-1} + (x-\overline{x})(x_i-\overline{x})/[(n-1)s_x^2]\right\}^2$$

$$(2.64)\quad = \sigma^2 \sum_{i=1}^{n} \left\{n^{-2} + \frac{(x-\overline{x})^2(x_i-\overline{x})^2}{[(n-1)s_x^2]^2} + 2n^{-1}\frac{(x-\overline{x})(x_i-\overline{x})}{[(n-1)s_x^2]}\right\}$$

$$(2.65)\quad = \sigma^2\left\{n^{-1} + \frac{(x-\overline{x})^2 \sum_i (x_i-\overline{x})^2}{[(n-1)s_x^2]^2} + 0\right\}$$

$$(2.66)\quad = \sigma^2\left\{n^{-1} + \frac{(x-\overline{x})^2}{[(n-1)s_x^2]}\right\}.$$

3. Define the standard error $se(\hat{\mu}_Y(x))$ **, as an estimate of sqrt(Var(** $\mu_Y(x)$ **)) :**

2.5.2 Derivations

Steps to get 95% C.I. for the subpopulation mean:

1. Consider the subpopulation mean, $\mu_Y(x)$, as a random variable.

2. Determine Var[$\mu_Y(x)$] $= \sigma^2\left\{n^{-1} + \frac{(x-\overline{x})^2}{[(n-1)s_x^2]}\right\}$.

3. Define the standard error $se(\hat{\mu}_Y(x))$, as an estimate of sqrt(Var($\mu_Y(x)$)) :

$$se(\hat{\mu}_Y(x)) = \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}}.$$

2.5.2 Derivations

Steps to get 95% C.I. for the subpopulation mean:

1. Consider the subpopulation mean, $\mu_Y(x)$, as a random variable.

2. Determine Var[$\mu_Y(x)$] $= \sigma^2\left\{n^{-1} + \frac{(x-\overline{x})^2}{[(n-1)s_x^2]}\right\}$.

3. Define the standard error $se(\hat{\mu}_Y(x))$, as an estimate of sqrt(Var($\mu_Y(x)$)) :

$$se(\hat{\mu}_Y(x)) = \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}}.$$

2.5.2 Derivations

Steps to get 95% C.I. for the subpopulation mean:

1. Consider the subpopulation mean, $\mu_Y(x)$, as a random variable.

2. Determine Var[$\mu_Y(x)$] $= \sigma^2\left\{n^{-1} + \frac{(x-\overline{x})^2}{[(n-1)s_x^2]}\right\}$.

3. Define the standard error $se(\hat{\mu}_Y(x))$, as an estimate of sqrt(Var($\mu_Y(x)$)) :

$$se(\hat{\mu}_Y(x)) = \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}}.$$

4. **95% C.I. = [$\hat{\mu}_Y(x)$ – c* $se(\hat{\mu}_Y(x))$, $\hat{\mu}_Y(x)$ + c* $se(\hat{\mu}_Y(x))$]**

The 95% confidence interval for subpopulation mean $\mu_Y(x) = \beta_0 + \beta_1 x$ is

(2.43) $$\hat{\mu}_Y(x) \pm t_{n-2,0.975} \times se(\hat{\mu}_Y(x)), \quad \hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

2.5.2 Derivations

The 95% confidence interval for subpopulation mean $\mu_Y(x) = \beta_0 + \beta_1 x$ is

$$\hat{\mu}_Y(x) \pm t_{n-2,0.975} \times se(\hat{\mu}_Y(x)), \quad \hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{2.43}$$

$$se(\hat{\mu}_Y(x)) = \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}}.$$

```
> myx <- 20
>  muhat_x <- b0+b1*myx
>  muhat_x
[1] 28.62758
> lowerCI <- muhat_x - qt(0.975,n-2) * s * sqrt(1/n + ((myx-xbar)^2)/((n-1)*sx^2))
> upperCI <-  muhat_x + qt(0.975,n-2) * s * sqrt(1/n + ((myx-xbar)^2)/((n-1)*sx^2))
> lowerCI
[1] 14.36777
> upperCI
[1] 42.88739
```

2.5.2 Derivations

The 95% confidence interval for subpopulation mean $\mu_Y(x) = \beta_0 + \beta_1 x$ is

$$\hat{\mu}_Y(x) \pm t_{n-2,0.975} \times se(\hat{\mu}_Y(x)), \quad \hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{2.43}$$

$$se(\hat{\mu}_Y(x)) = \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}}.$$

```
> myx <- 40
>  muhat_x <- b0+b1*myx
>  muhat_x
[1] 39.58997
> lowerCI <- muhat_x - qt(0.975,n-2) * s * sqrt(1/n + ((myx-xbar)^2)/((n-1)*sx^2))
> upperCI <-  muhat_x + qt(0.975,n-2) * s * sqrt(1/n + ((myx-xbar)^2)/((n-1)*sx^2))
> lowerCI
[1] 27.06291
> upperCI
[1] 52.11703
```

2.5.2 Derivations

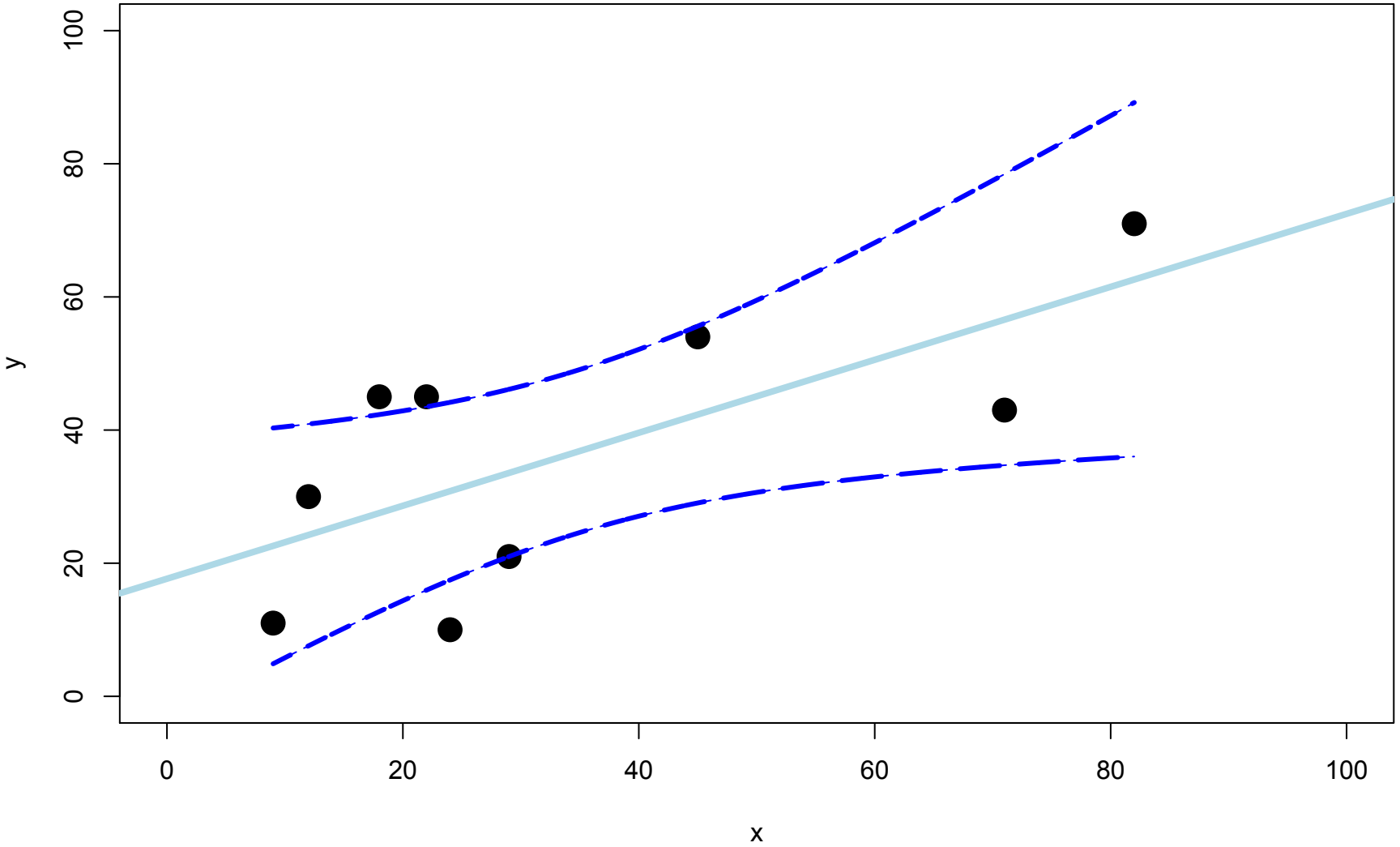
The 95% confidence interval for subpopulation mean $\mu_Y(x) = \beta_0 + \beta_1 x$ is

$$\hat{\mu}_Y(x) \pm t_{n-2,0.975} \times se(\hat{\mu}_Y(x)), \quad \hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{2.43}$$

$$se(\hat{\mu}_Y(x)) = \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}}.$$

```
> myx <- 80
>  muhat_x <- b0+b1*myx
>  muhat_x
[1] 61.51475
> lowerCI <- muhat_x - qt(0.975,n-2) * s * sqrt(1/n + ((myx-xbar)^2)/((n-1)*sx^2))
> upperCI <-  muhat_x + qt(0.975,n-2) * s * sqrt(1/n + ((myx-xbar)^2)/((n-1)*sx^2))
> lowerCI
[1] 35.8075
> upperCI
[1] 87.22199
```

Age vs. Money



Solid Blue Line is the "subpopulation mean" for each value of x:
The average amount of money (i.e. the expectation of random variable Y), among people aged "x" years old.

2.5.2 Derivations

The 95% confidence interval for subpopulation mean $\mu_Y(x) = \beta_0 + \beta_1 x$ is

$$\hat{\mu}_Y(x) \pm t_{n-2,0.975} \times se(\hat{\mu}_Y(x)), \quad \hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{2.43}$$

$$se(\hat{\mu}_Y(x)) = \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}}.$$

```
> myx <- 100
>  muhat_x <- b0+b1*myx
>  muhat_x
[1] 72.47714
> lowerCI <- muhat_x - qt(0.975,n-2) * s * sqrt(1/n + ((myx-xbar)^2)/((n-1)*sx^2))
> upperCI <-  muhat_x + qt(0.975,n-2) * s * sqrt(1/n + ((myx-xbar)^2)/((n-1)*sx^2))
> lowerCI
[1] 37.67466
> upperCI
[1] 107.2796
```