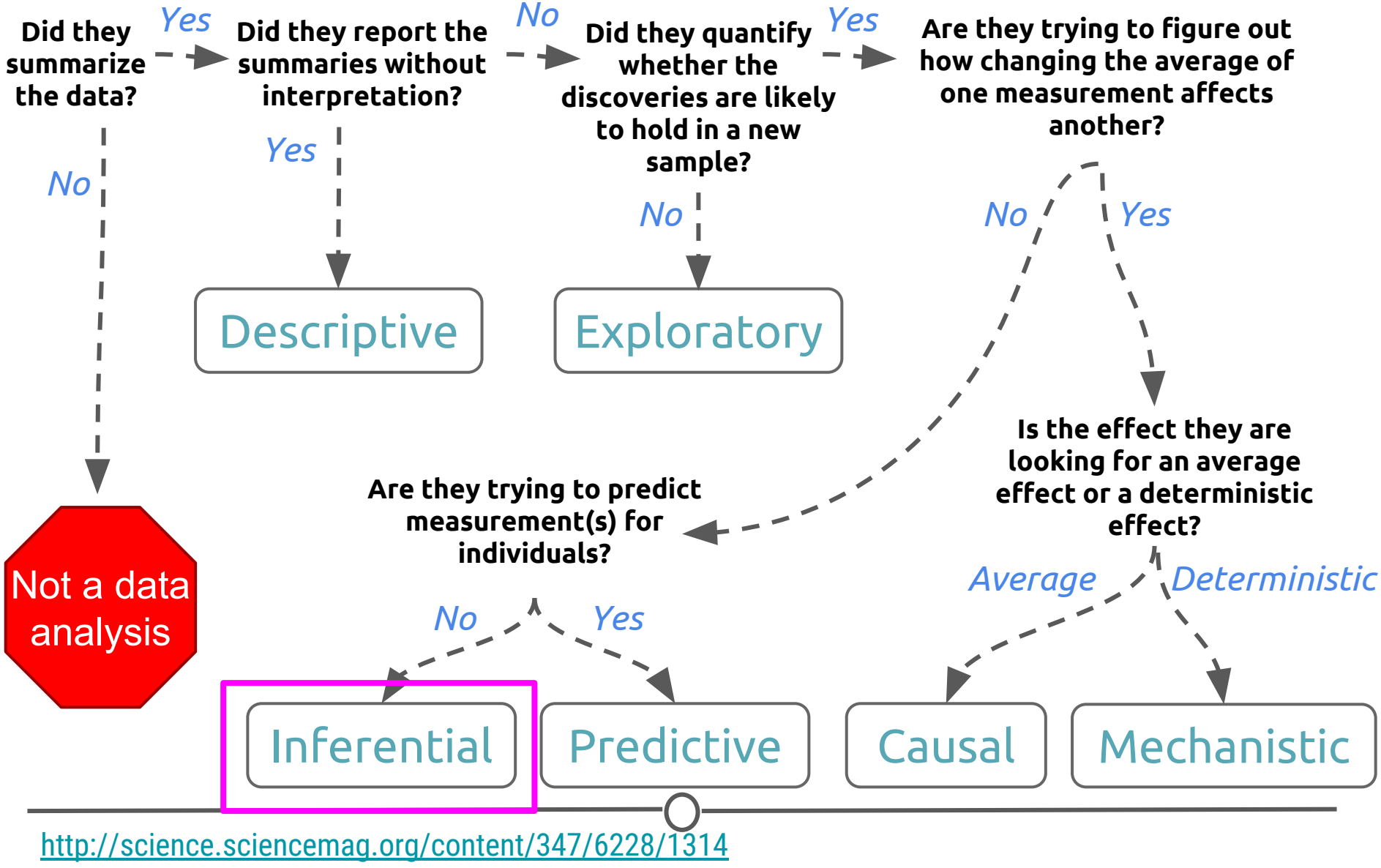
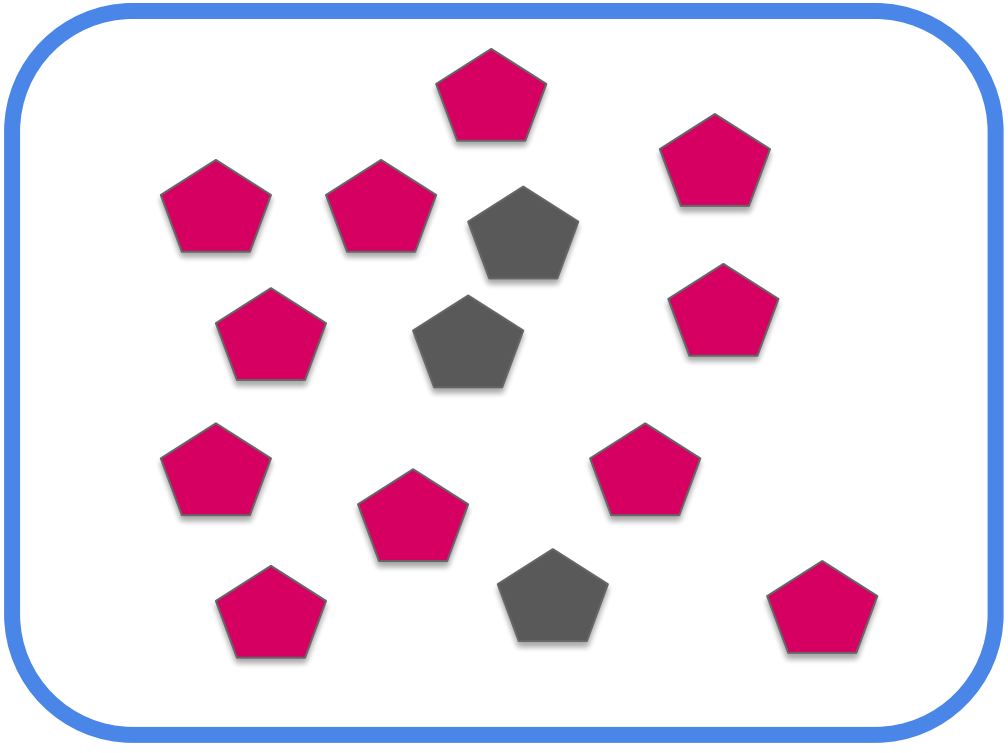
Inferential Analysis

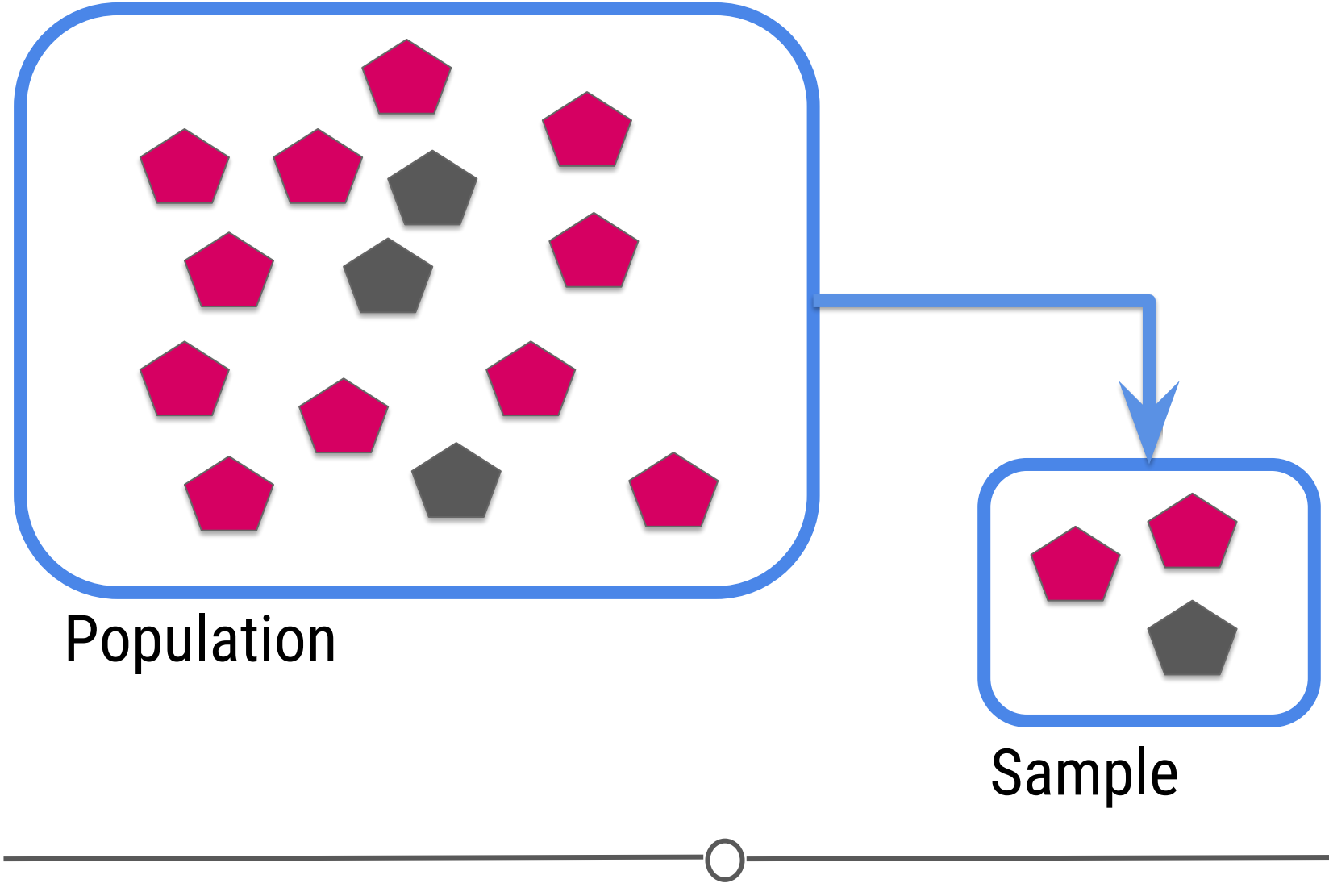
Data Analysis

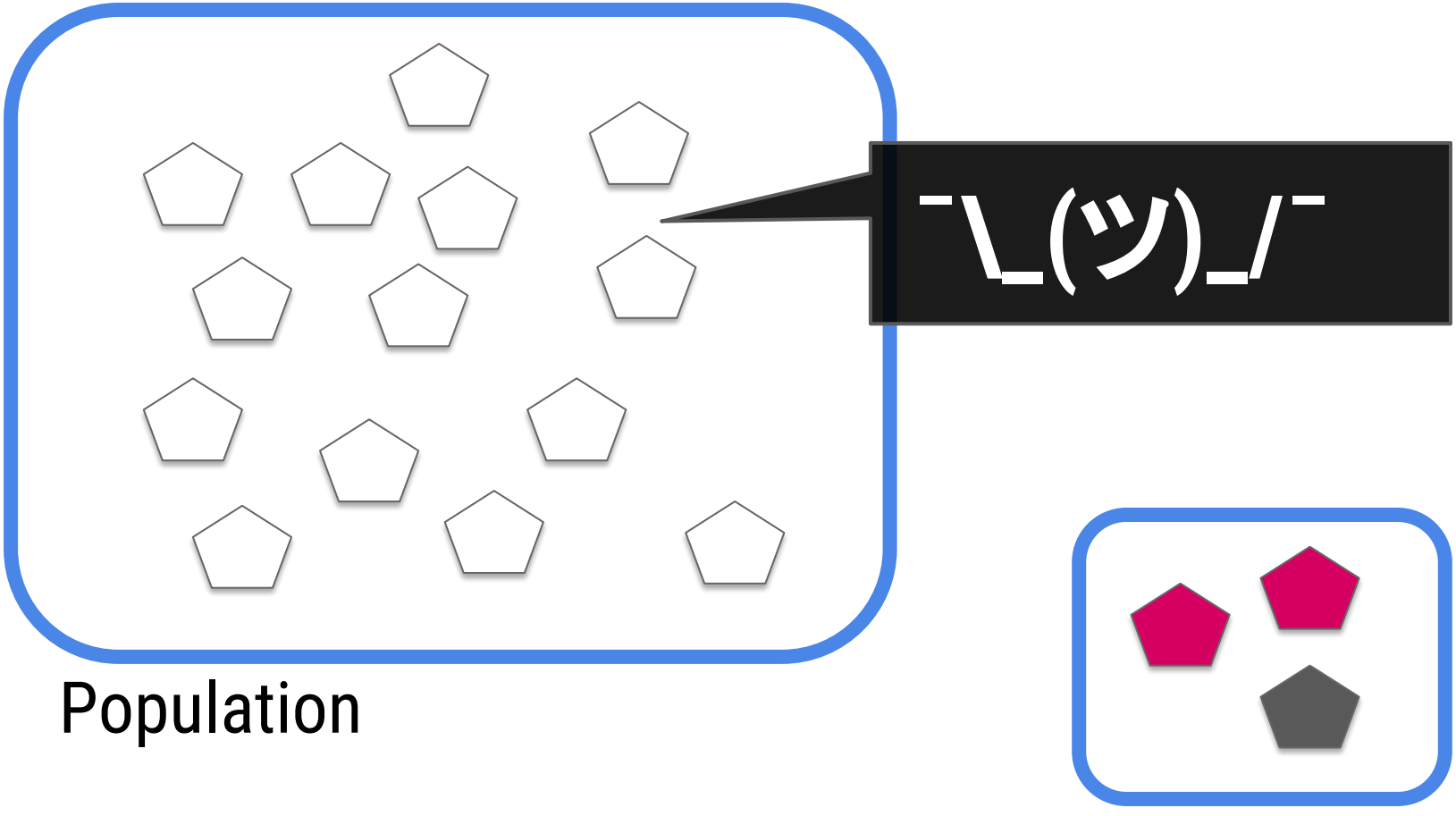


http://science.sciencemag.org/content/347/6228/1314



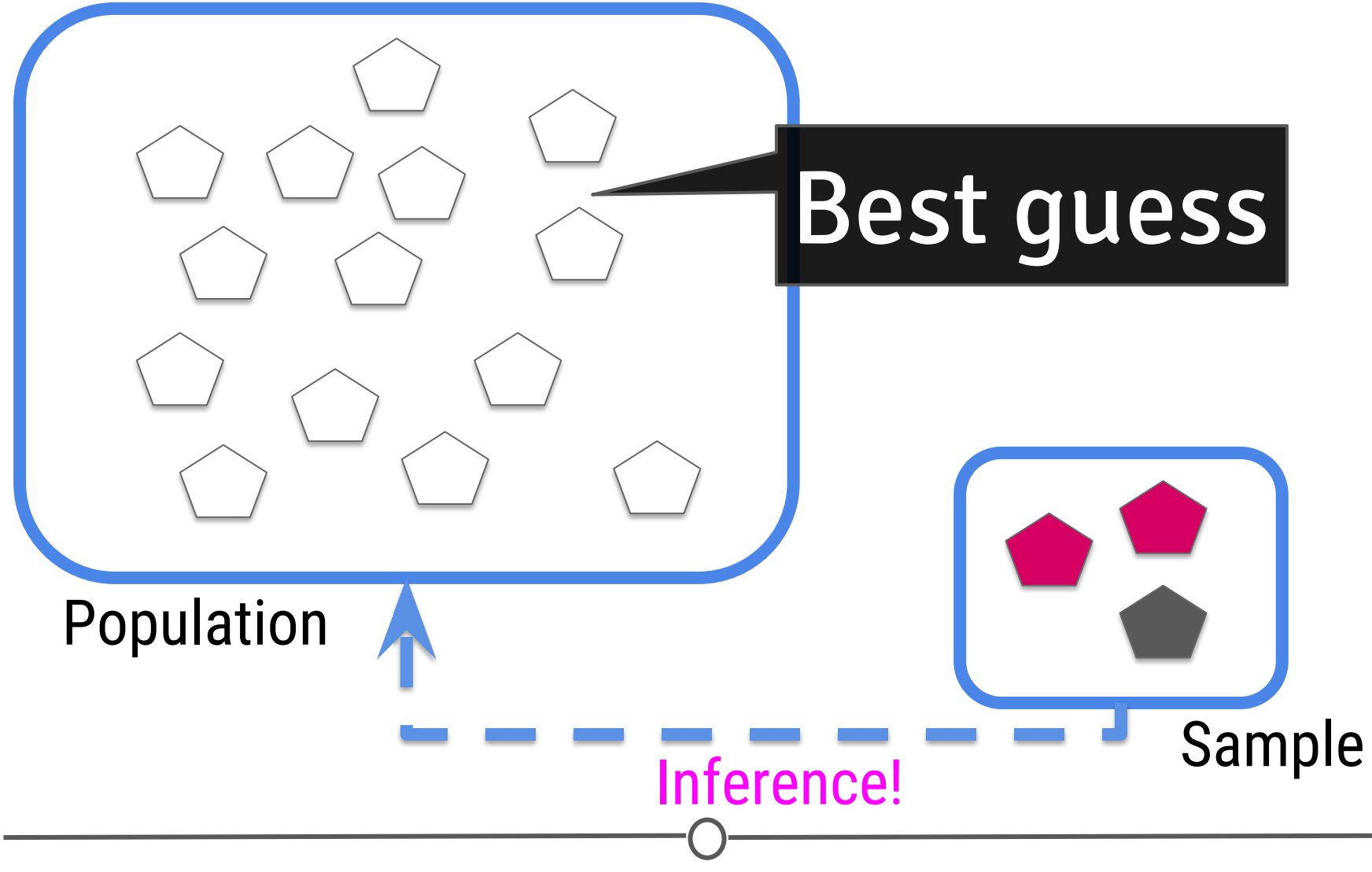
Population

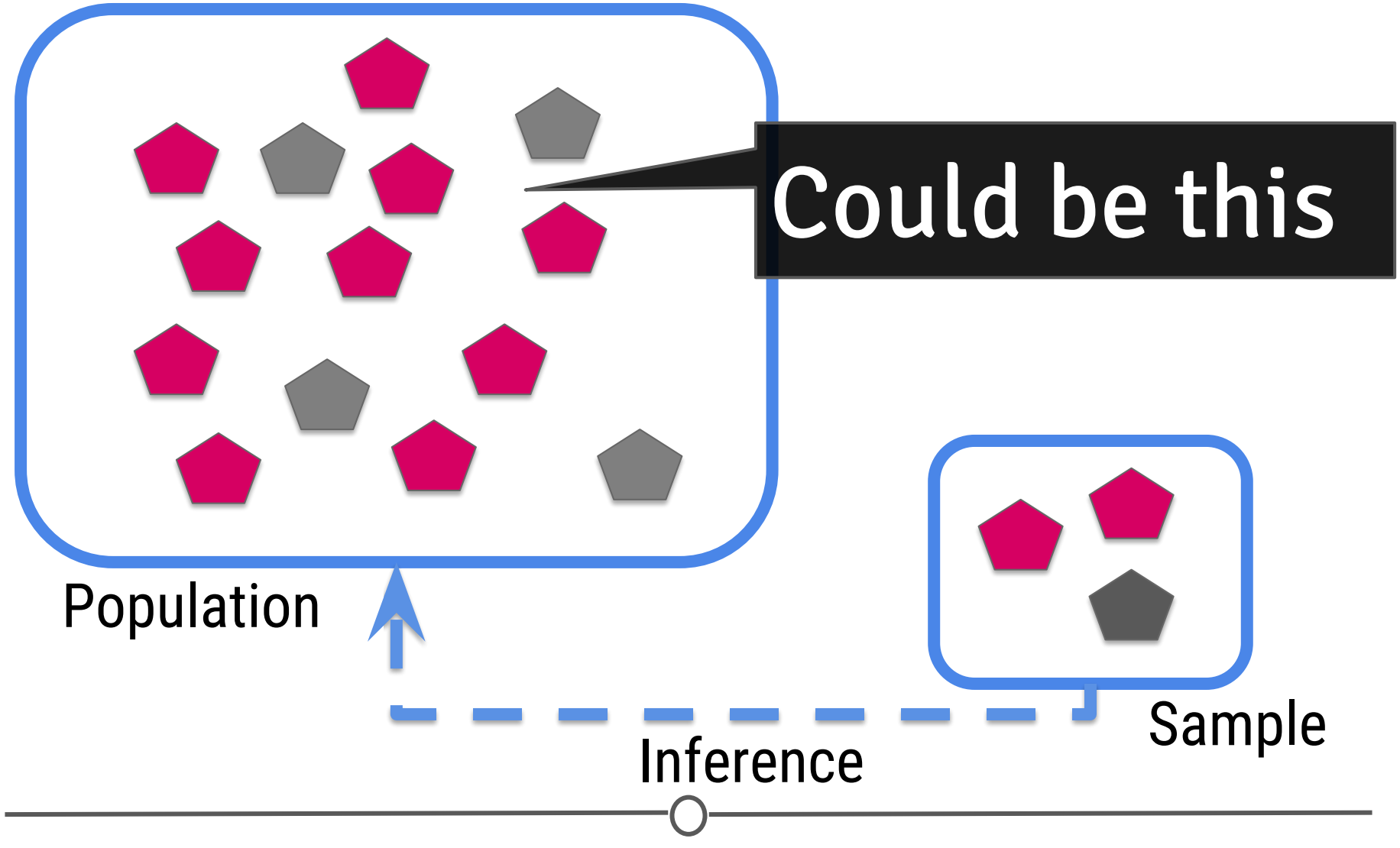

Population
Sample

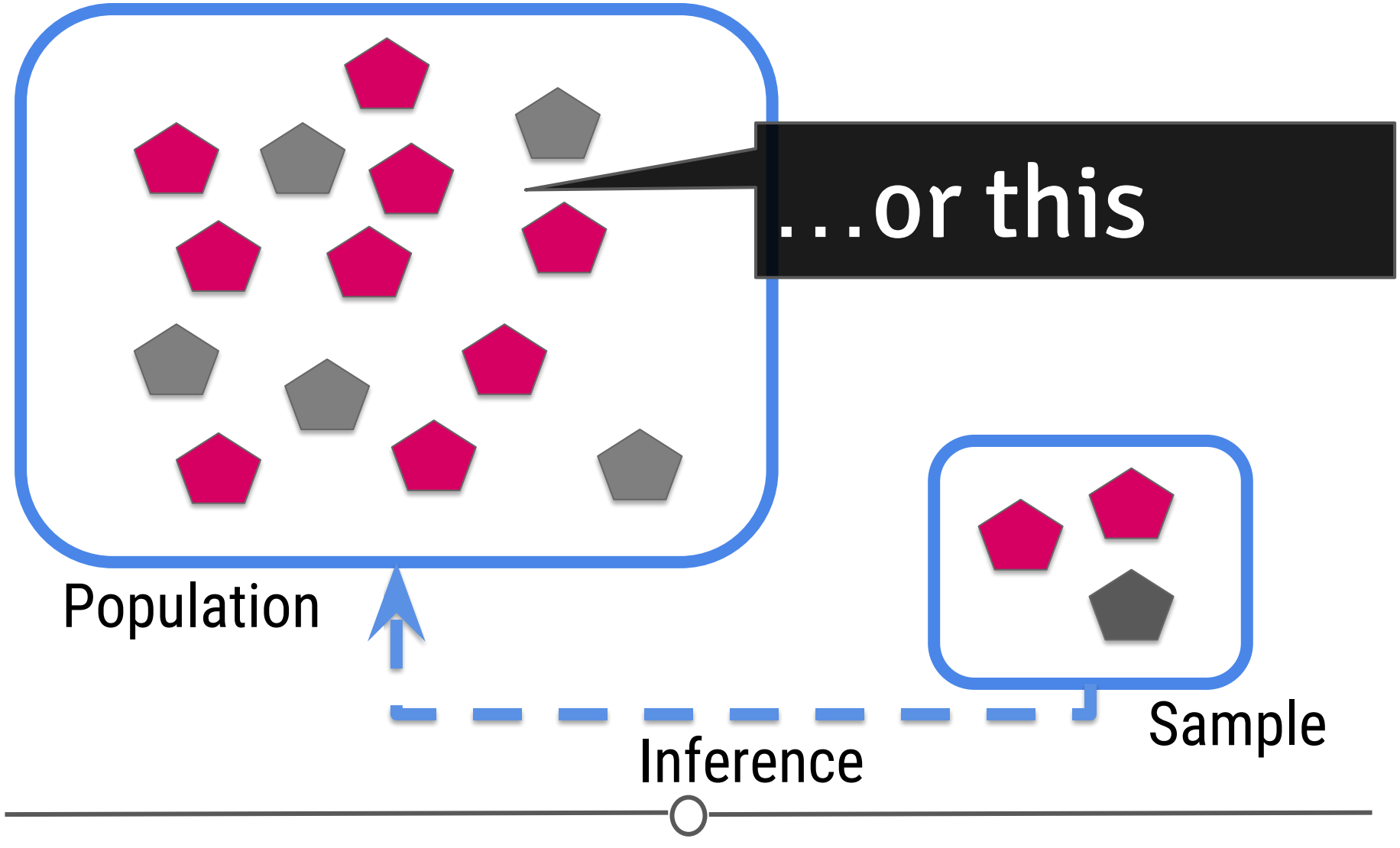


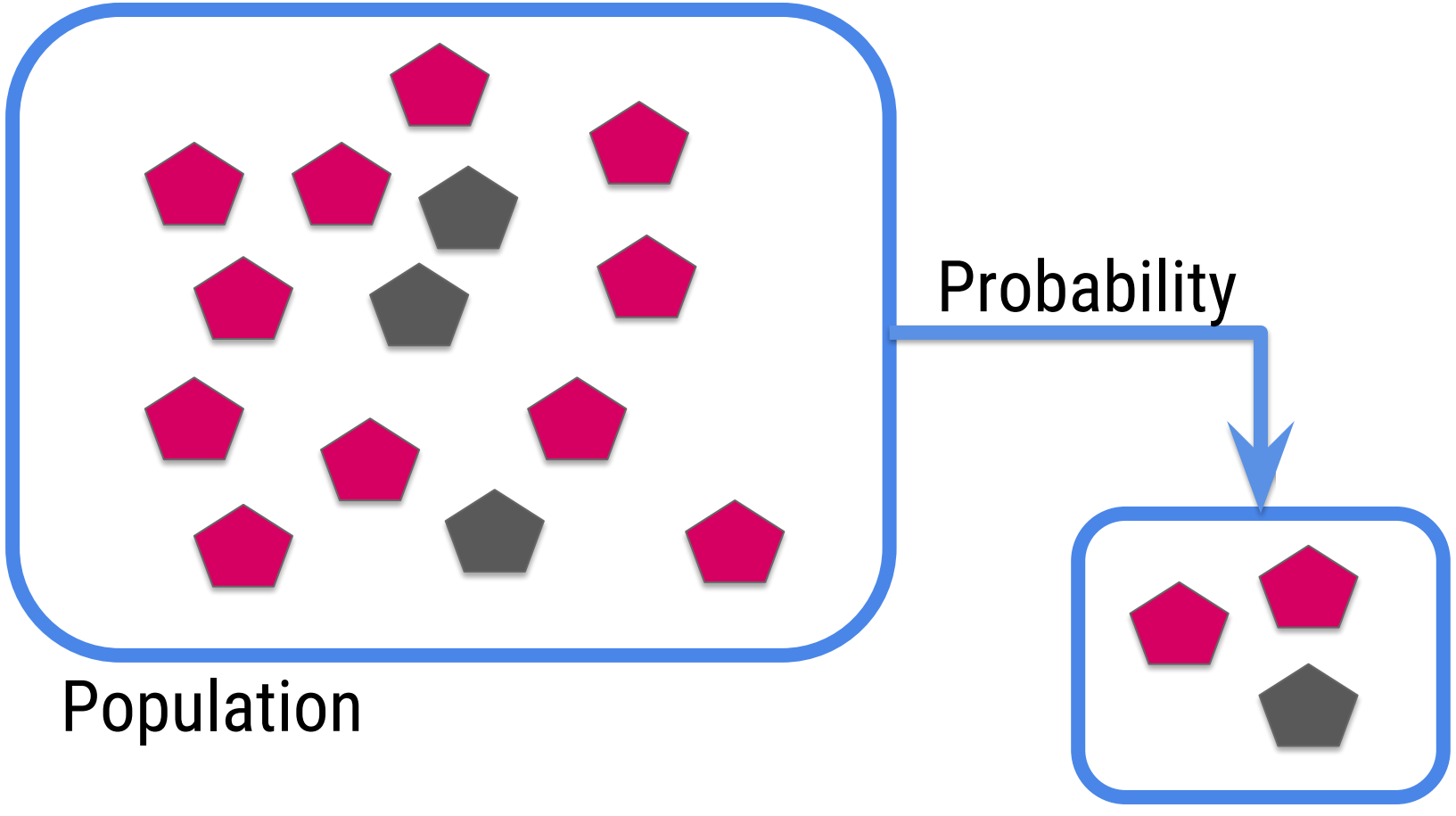
Population

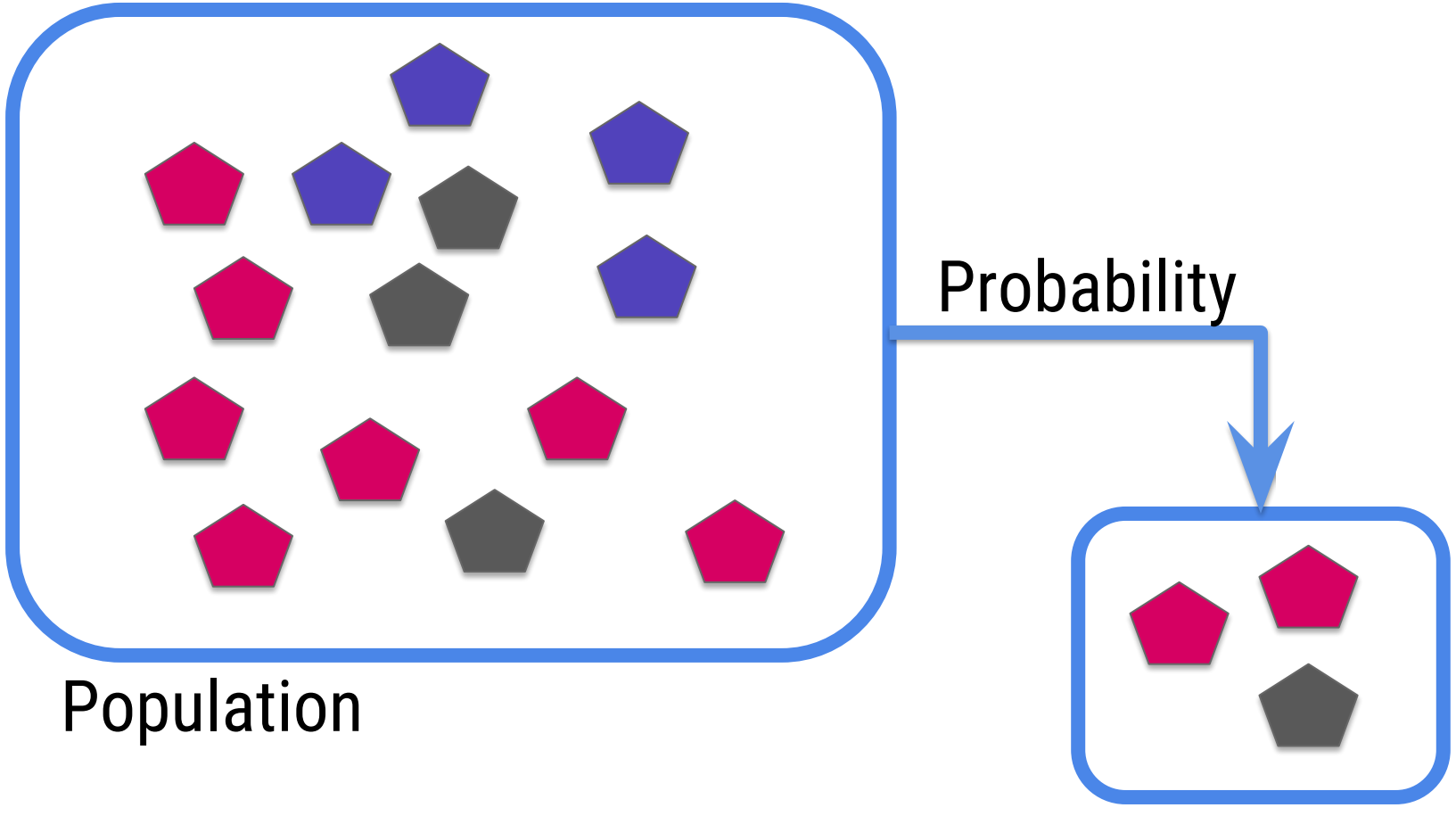
Sample

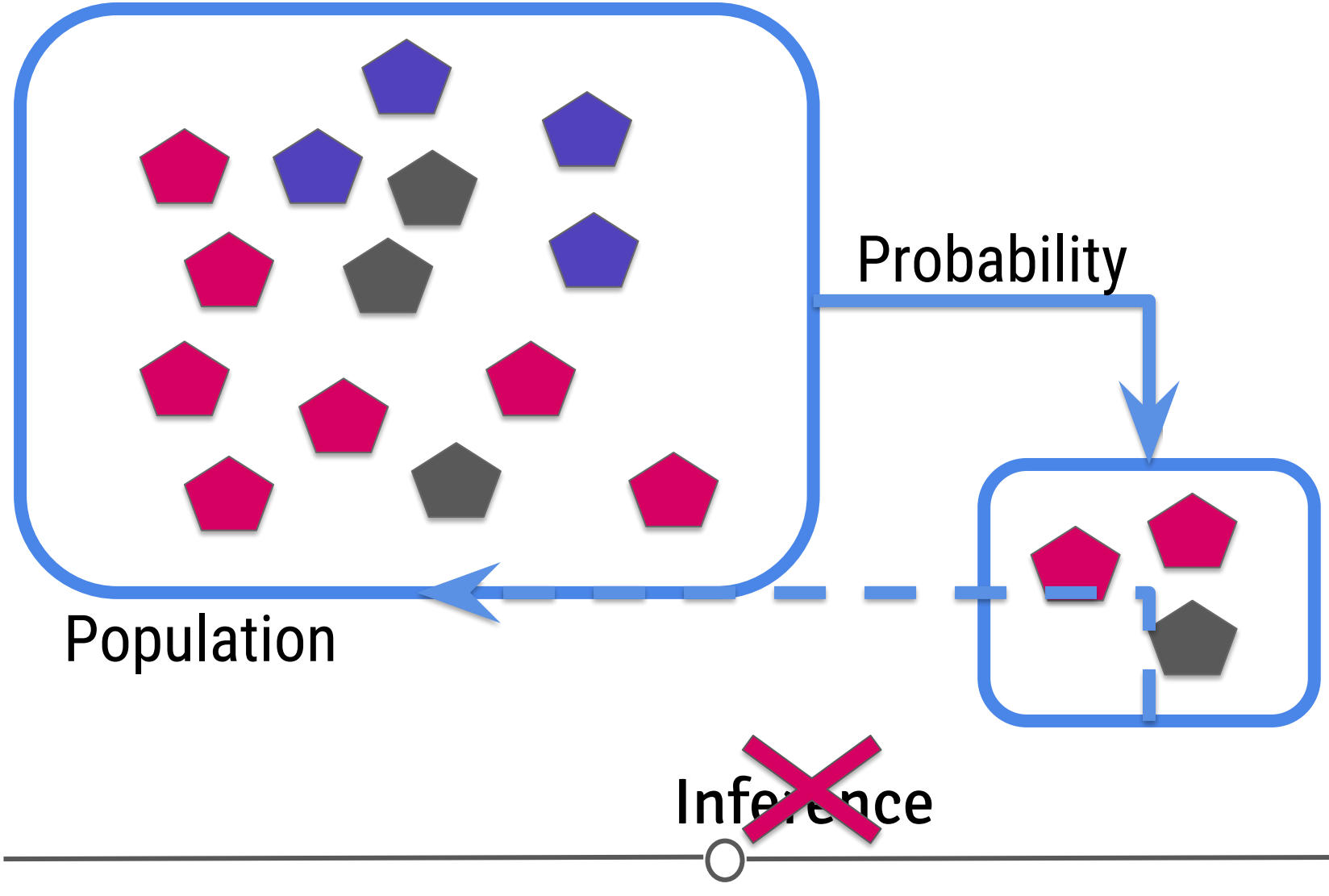

Best guess
Population
Sample
Inference!


Could be this
Population
Sample
Inference


...or this
Population
Sample
Inference


Probability
Population


Probability
Population


Probability
Population
Inference

Published in final edited form as:
Epidemiology. 2013 January ; 24(1): 23–31. doi:10.1097/EDE.0b013e3182770237.

The Effect of Air Pollution Control on Life Expectancy in the United States: An Analysis of 545 US counties for the period 2000 to 2007

Andrew W. Correia,
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 2, 4th Floor, Boston, MA 02115

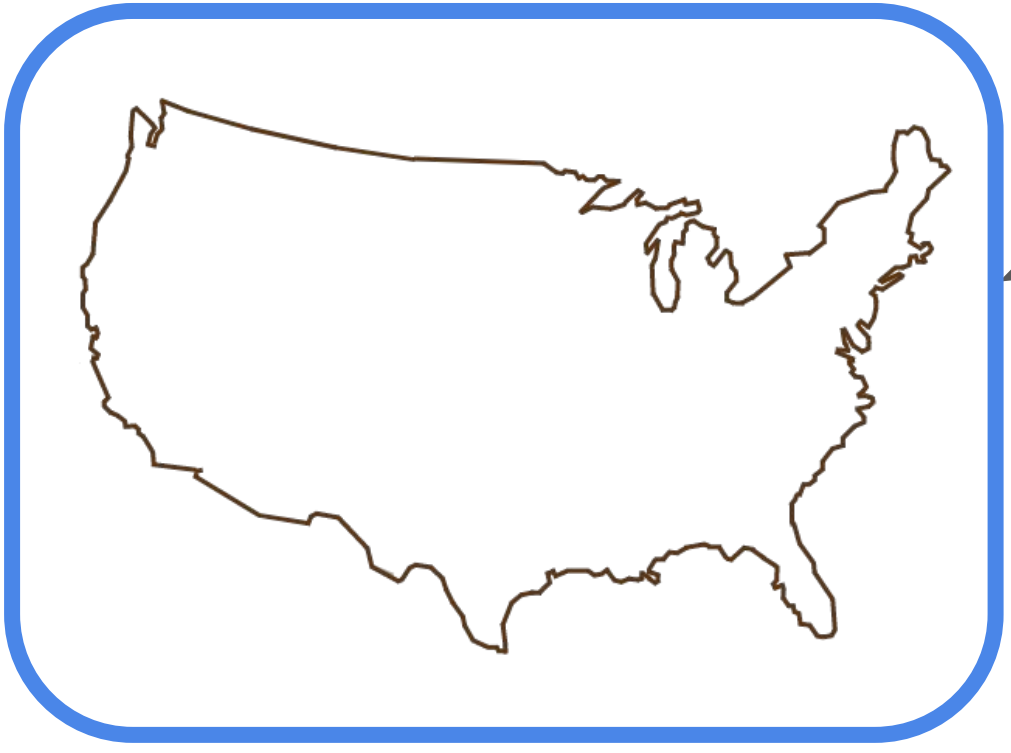
C. Arden Pope III,
Department of Economics, Brigham Young University, 142 Faculty Office Building, Provo, UT 84602

Douglas W. Dockery,
Departments of Environmental Health and Epidemiology, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 1, 1301B, Boston, MA 02115

Yun Wang,
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 2, 4th Floor, Boston, MA 02115

Majid Ezzati, and
MRC-HPA Centre for Environment and Health and Department of Epidemiology and Biostatistics, Imperial College London, Norfolk Place, St Mary's Campus, London W2 1PG

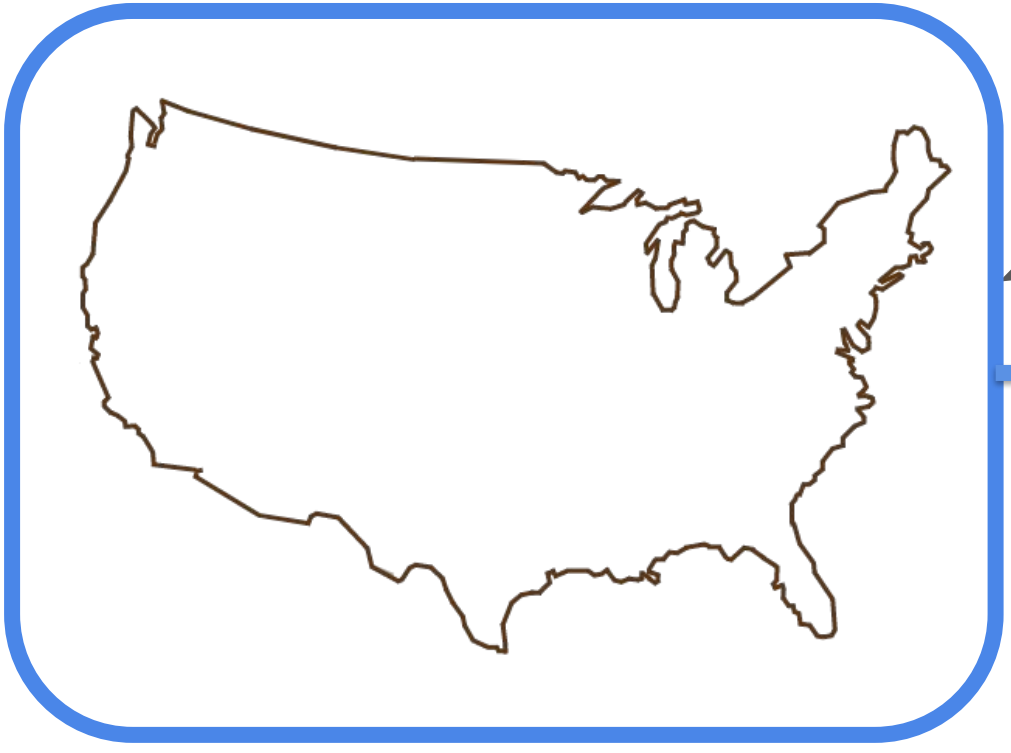
Francesca Dominici
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 2, 4th Floor, Boston, MA 02115, fdominic@hsph.harvard.edu, P: (617) 432-1056; F: (617)-739-1781



Population

The **population**: every individual in the USA

What if we want to know the effect of air pollution on **everyone in the United States**?



Population

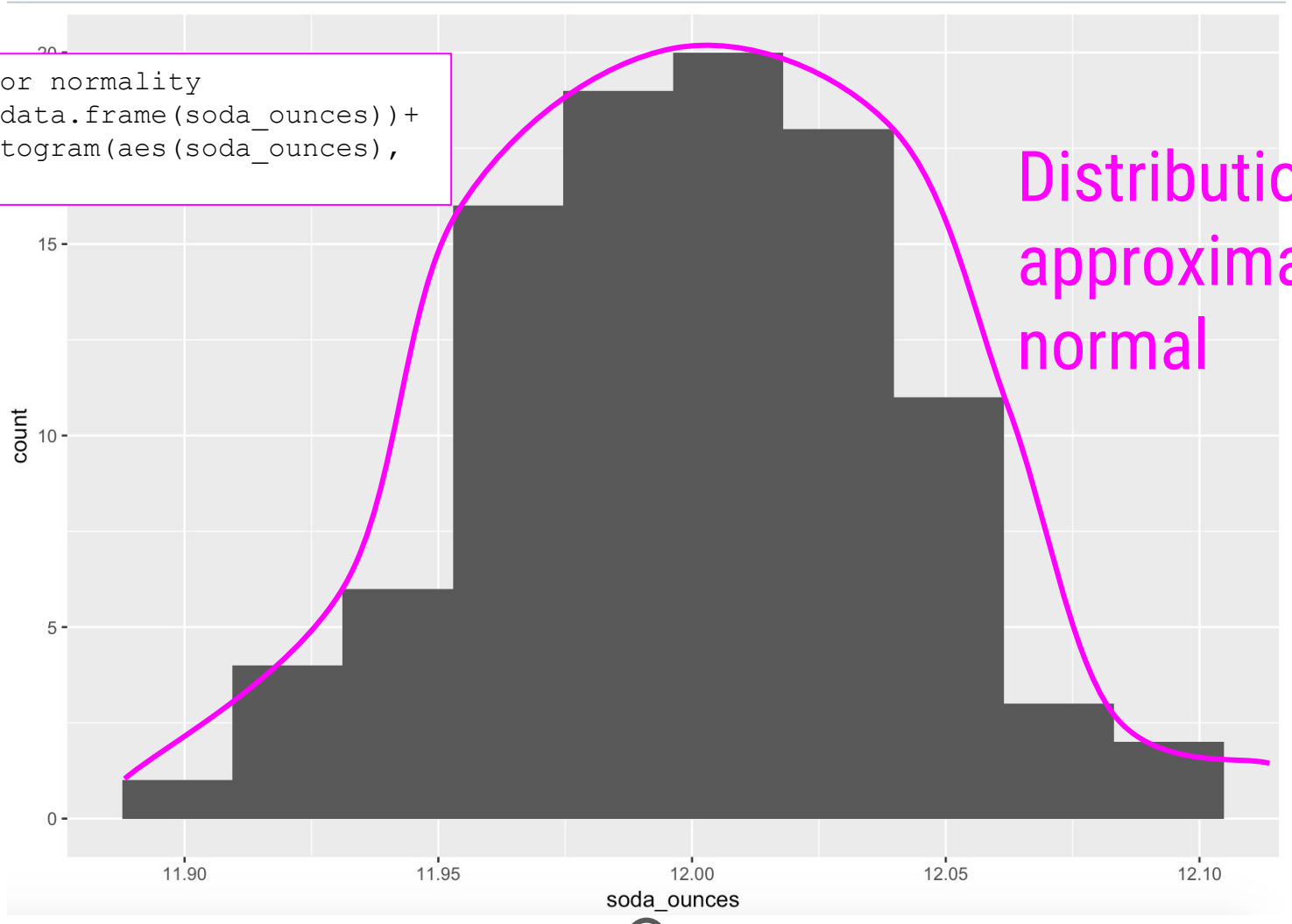
The **population**: every individual in the USA

555 US counties

Sample

```
> ## generate the dataset
> set.seed(34)
> soda_ounces <- rnorm(100, mean = 12, sd = 0.04)
> head(soda_ounces)
[1] 11.99444 12.04799 11.97009 11.97699 11.98946 11.98178
```

```
## check for normality
ggplot(as.data.frame(soda_ounces))+
  geom_histogram(aes(soda_ounces),
bins = 10)
```



Distribution is approximately normal

```
> ## carry out t-test
> t.test(soda_ounces, mu = 12)

        One Sample t-test

data:  soda_ounces
t = -0.074999, df = 99, p-value = 0.9404
alternative hypothesis: true mean is not equal to 12
95 percent confidence interval:
 11.99187 12.00754
sample estimates:
mean of x
  11.9997
```

95% CI contains the value 12, the expected mean!

Absenteeism from School in Rural New South Wales

Description

The `quine` data frame has 146 rows and 5 columns. Children from Walgett, New South Wales, Australia, were classified by Culture, Age, Sex and Learner status and the number of days absent from school in a particular school year was recorded.

Usage

```
quine
```

Format

This data frame contains the following columns:

`Eth`

: ethnic background: Aboriginal or Not, (`"A"` or `"N"`).

`Sex`

: sex: factor with levels (`"F"` or `"M"`).

`Age`

: age group: Primary (`"F0"`), or forms `"F1,"` `"F2"` or `"F3"`.

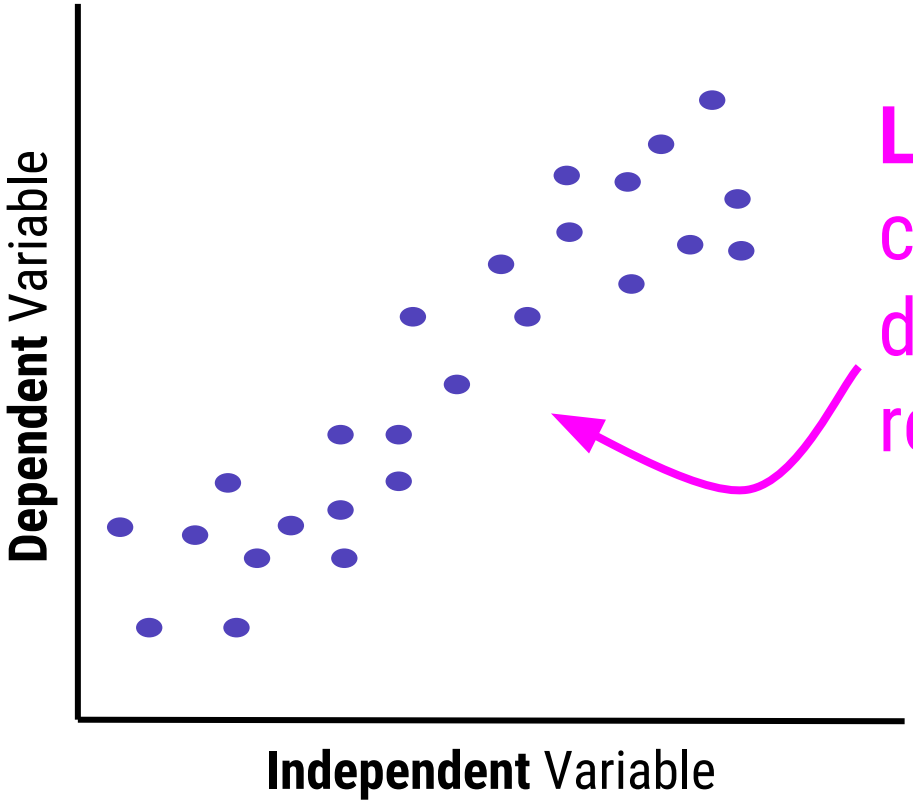
`Lrn`

: learner status: factor with levels Average or Slow learner, (`"AL"` or `"SL"`).

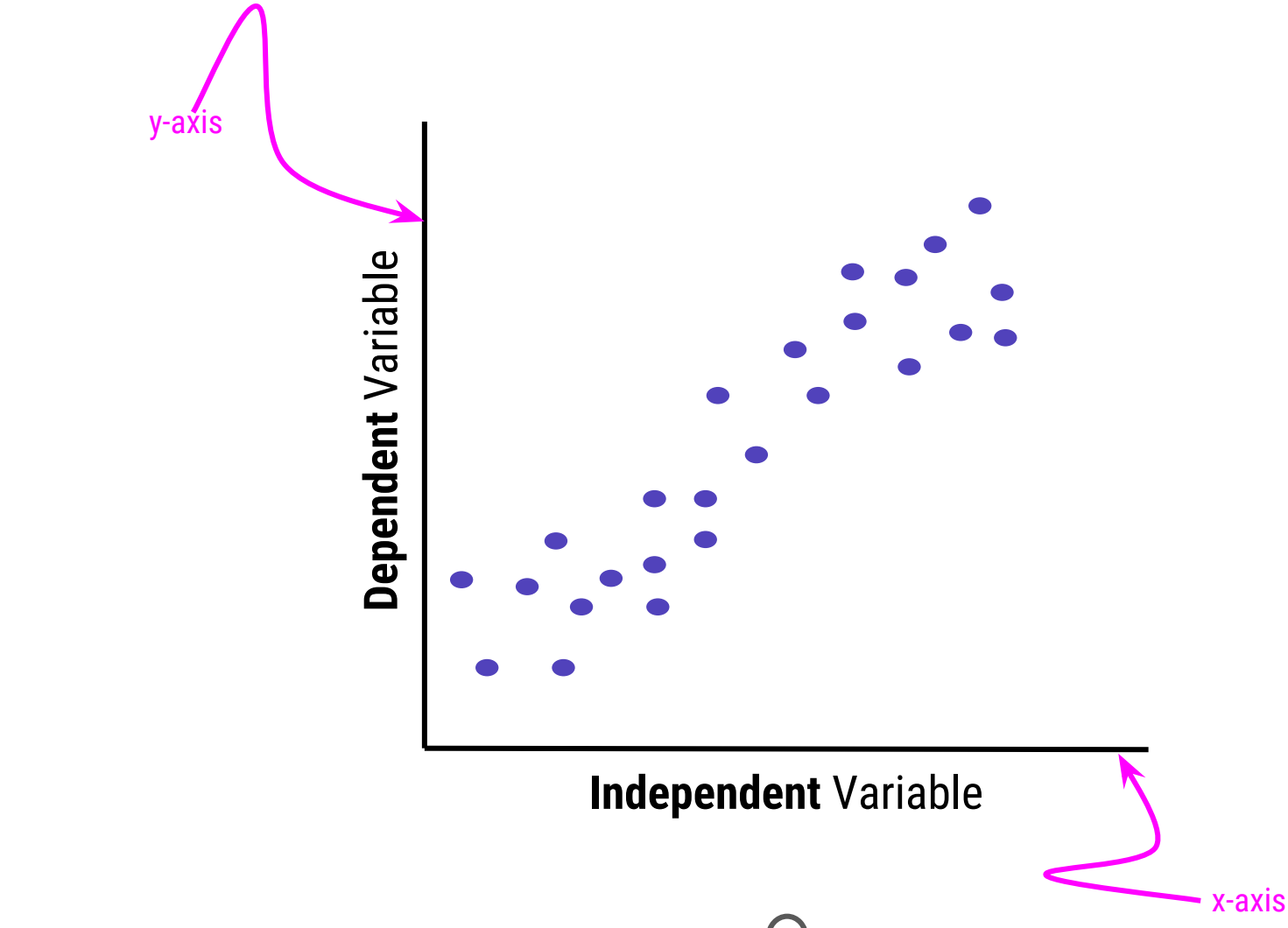
```
> library(MASS)
>
> ## take a look at the raw values
> table(quine$Eth, quine$Sex)

    F  M
A 38 31
N 42 35
```

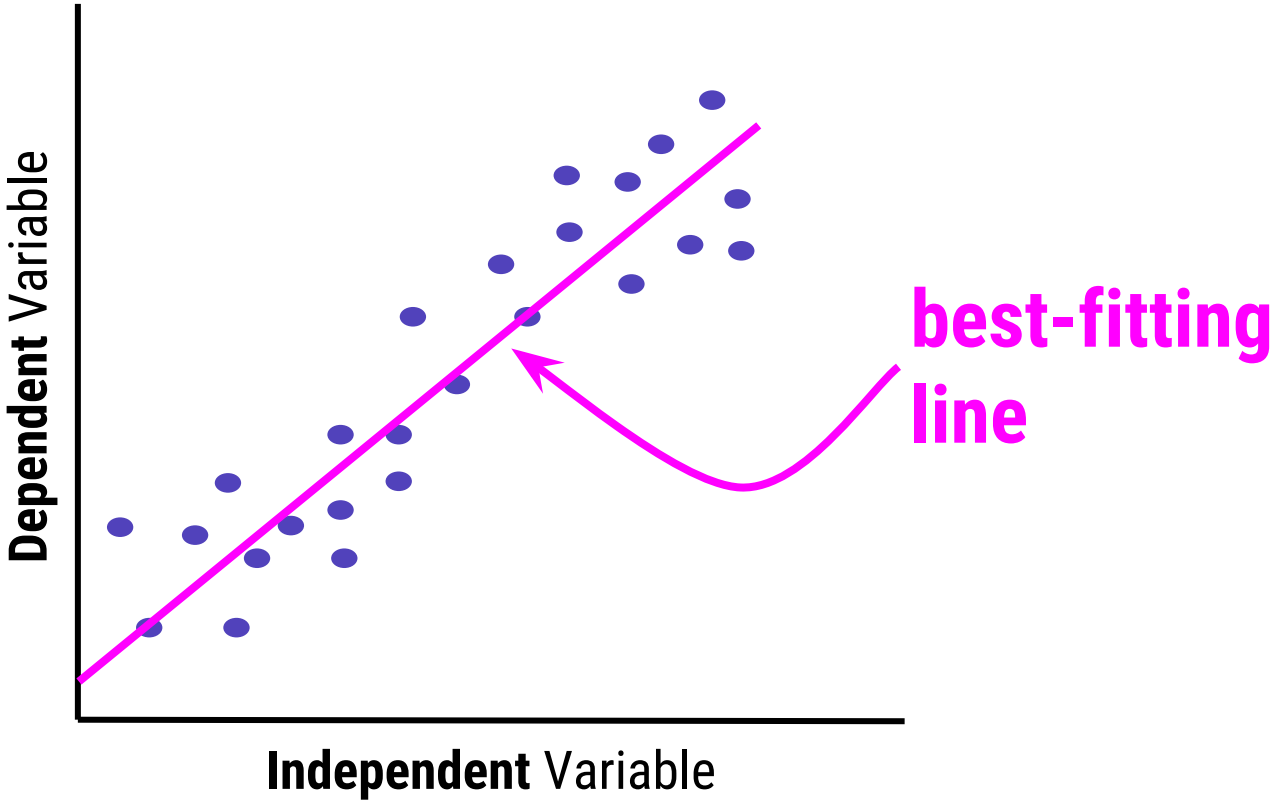
```
> ## test for differences in proportions between groups
> prop.test(table(quine$Eth, quine$Sex))

        2-sample test for equality of proportions with continuity correction

data:  table(quine$Eth, quine$Sex)
X-squared = 2.606e-30, df = 1, p-value = 1
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1616919  0.1722321
sample estimates:
   prop 1    prop 2
0.5507246 0.5454545
```

95% CI contains zero : no statistical difference in proportions between groups

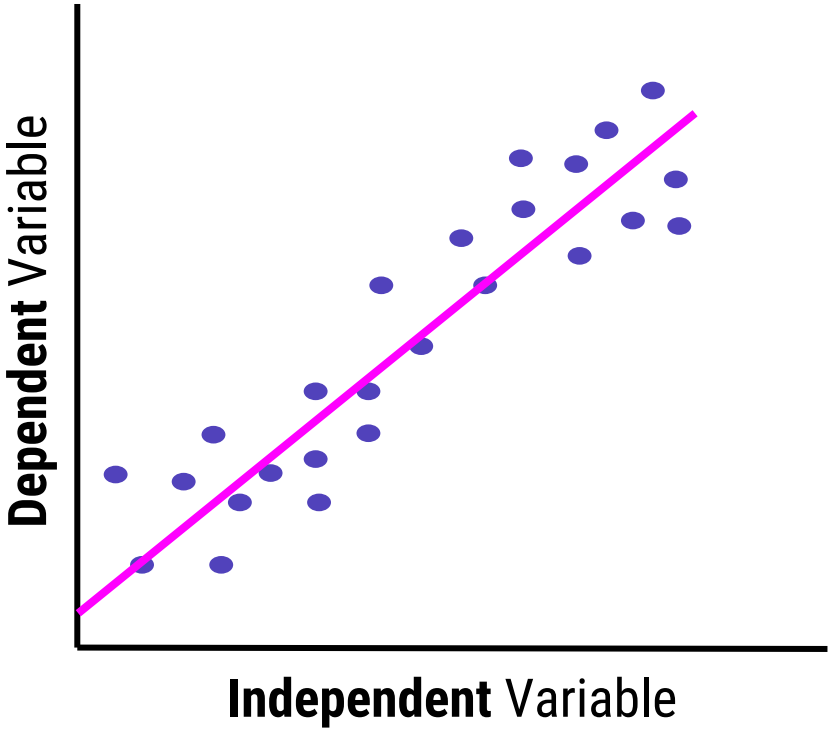


Linear regression can be used to describe this relationship

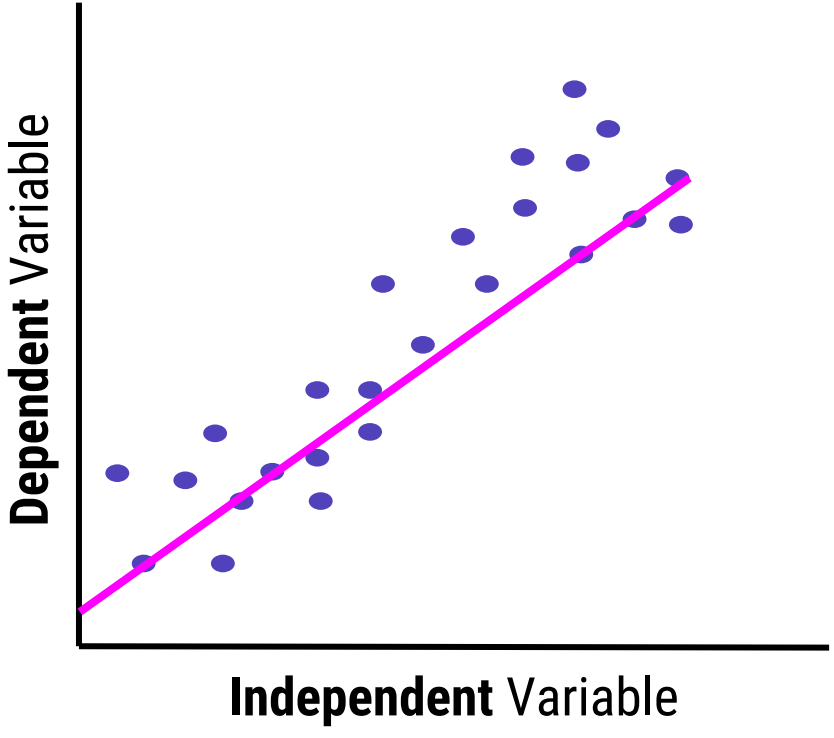

y-axis
Dependent Variable
Independent Variable
x-axis

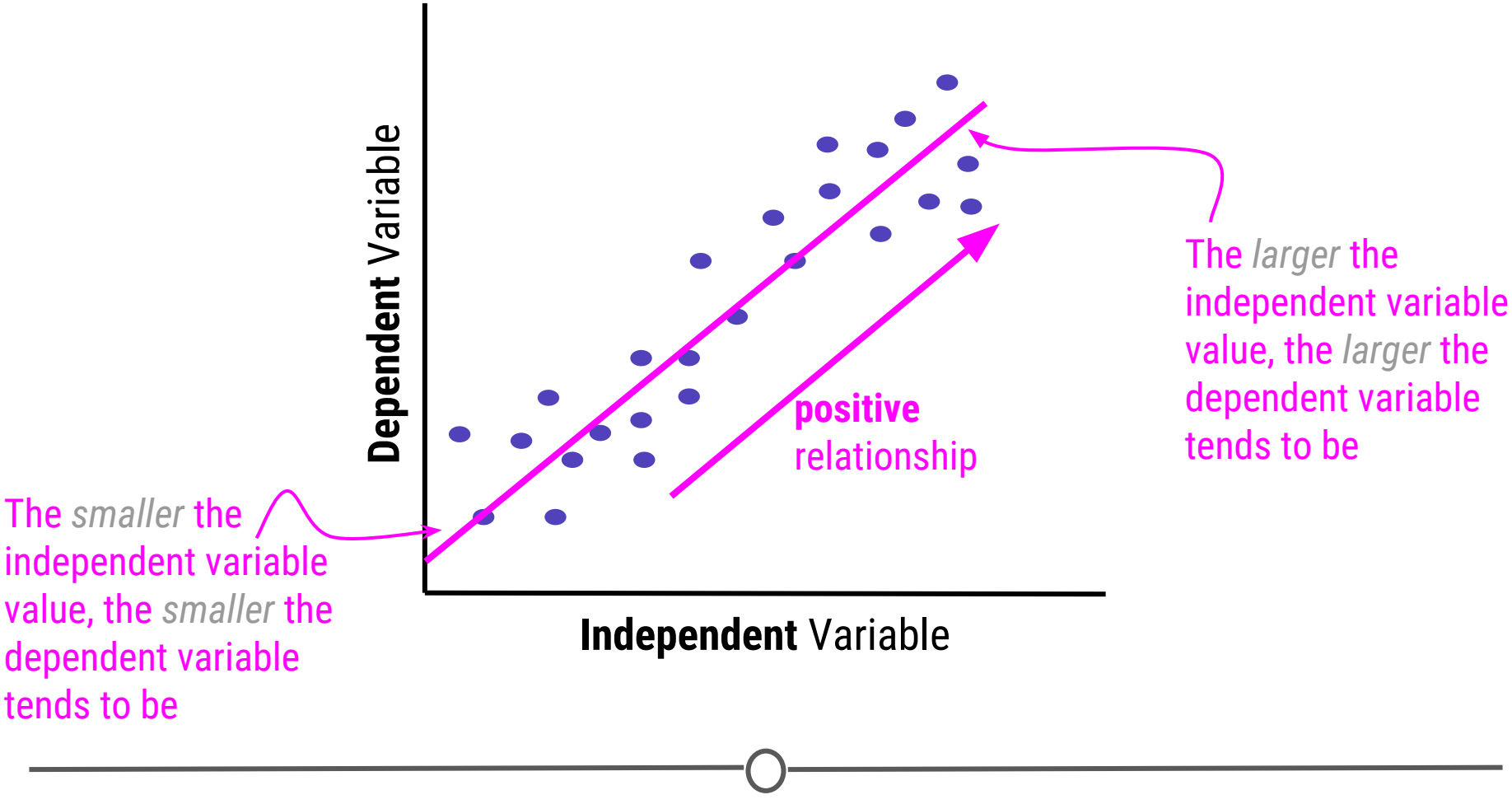

Dependent Variable
Independent Variable
best-fitting line

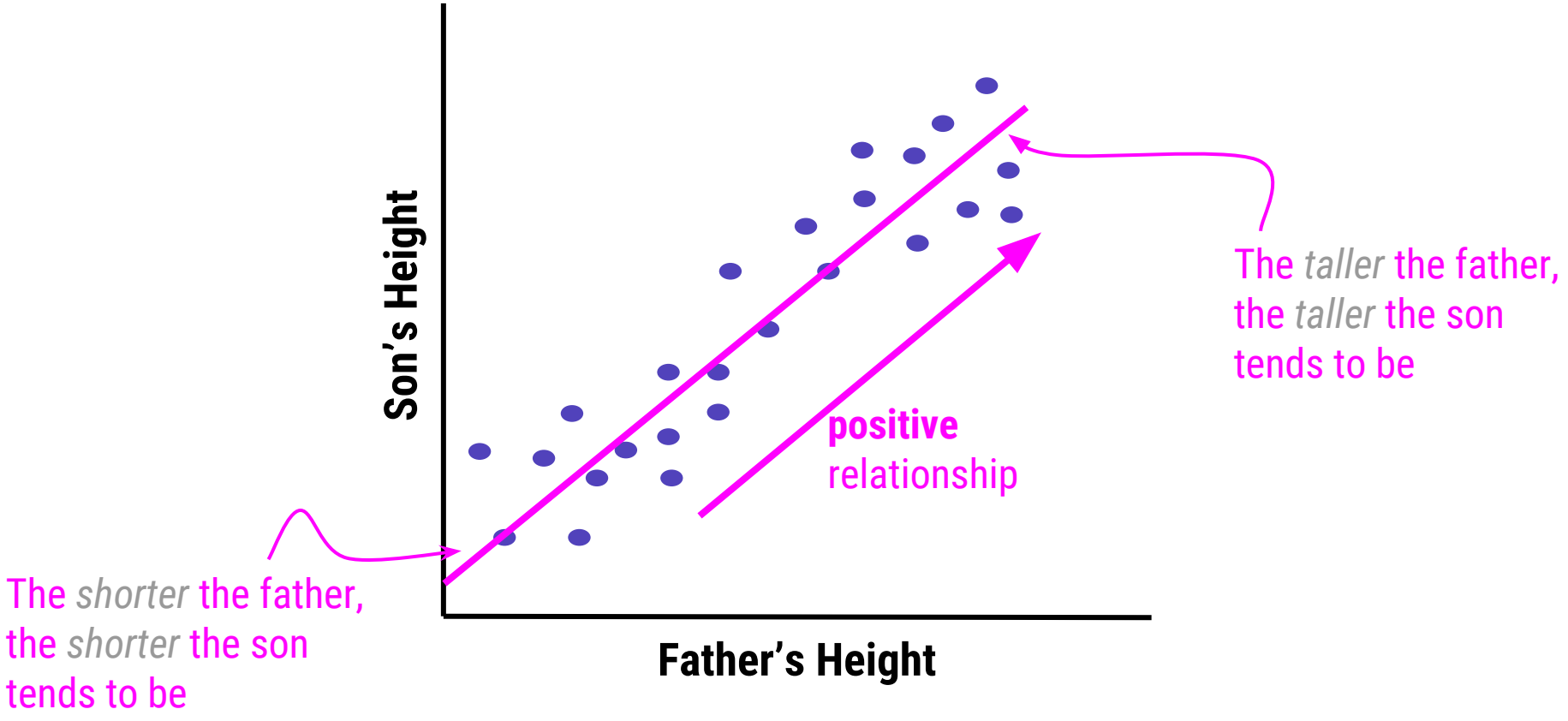
Best-fitting line

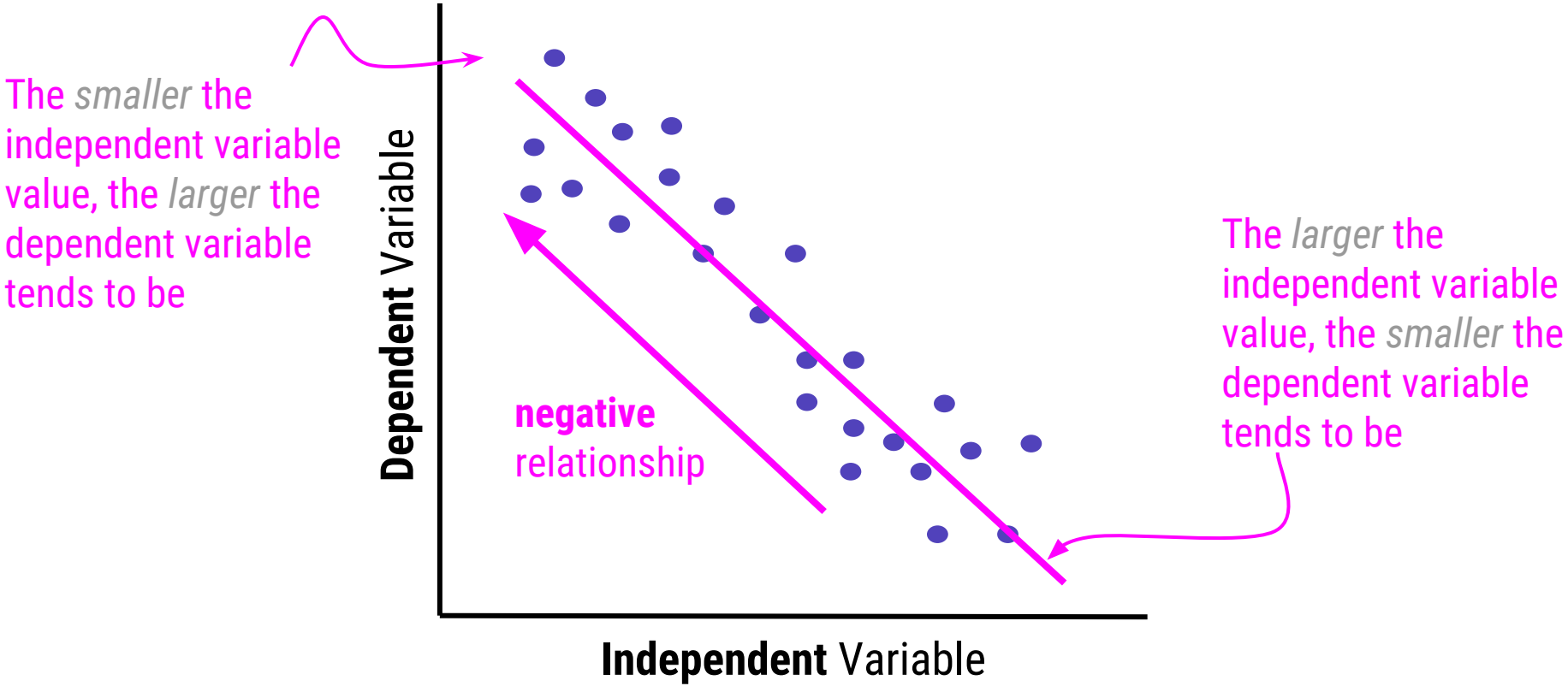


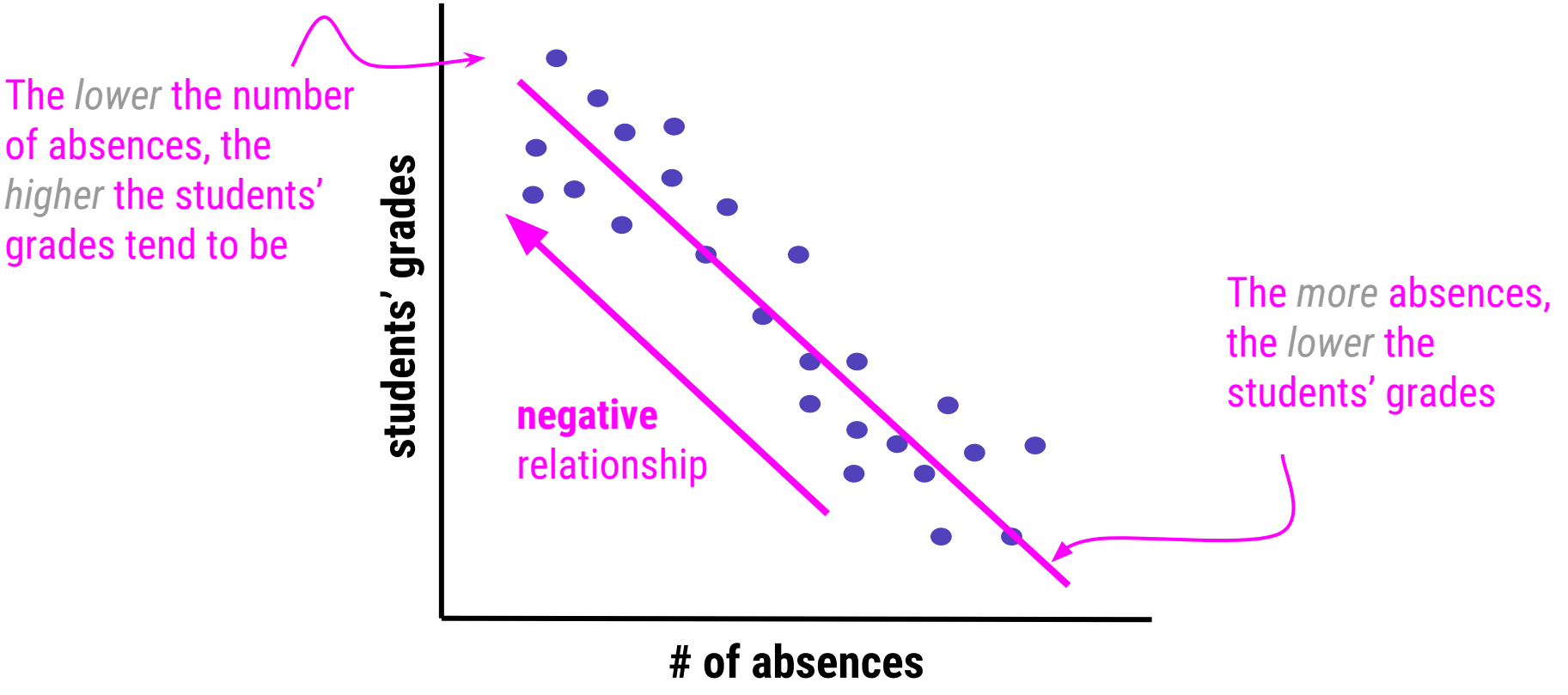
NOT a best-fitting line

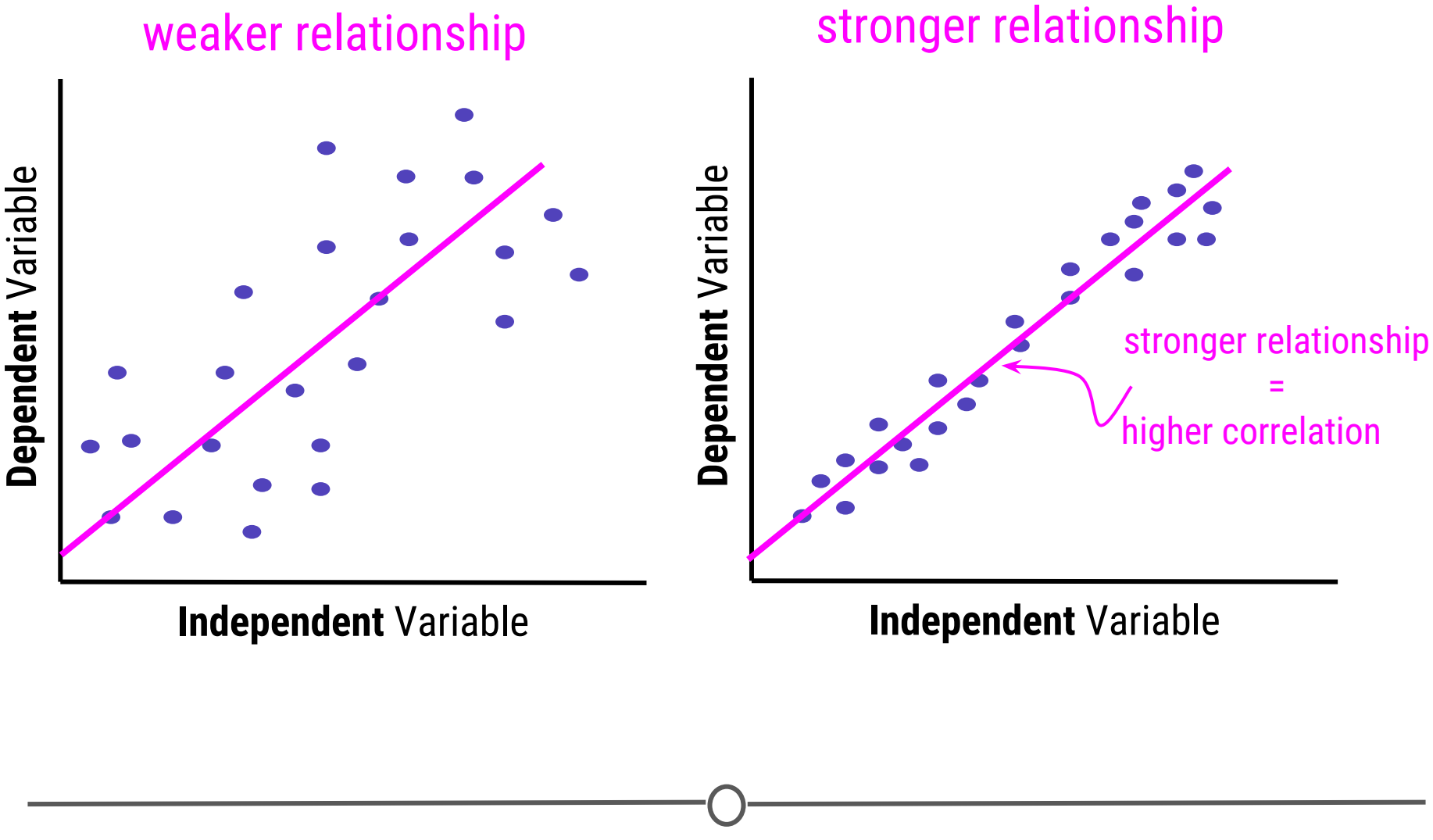


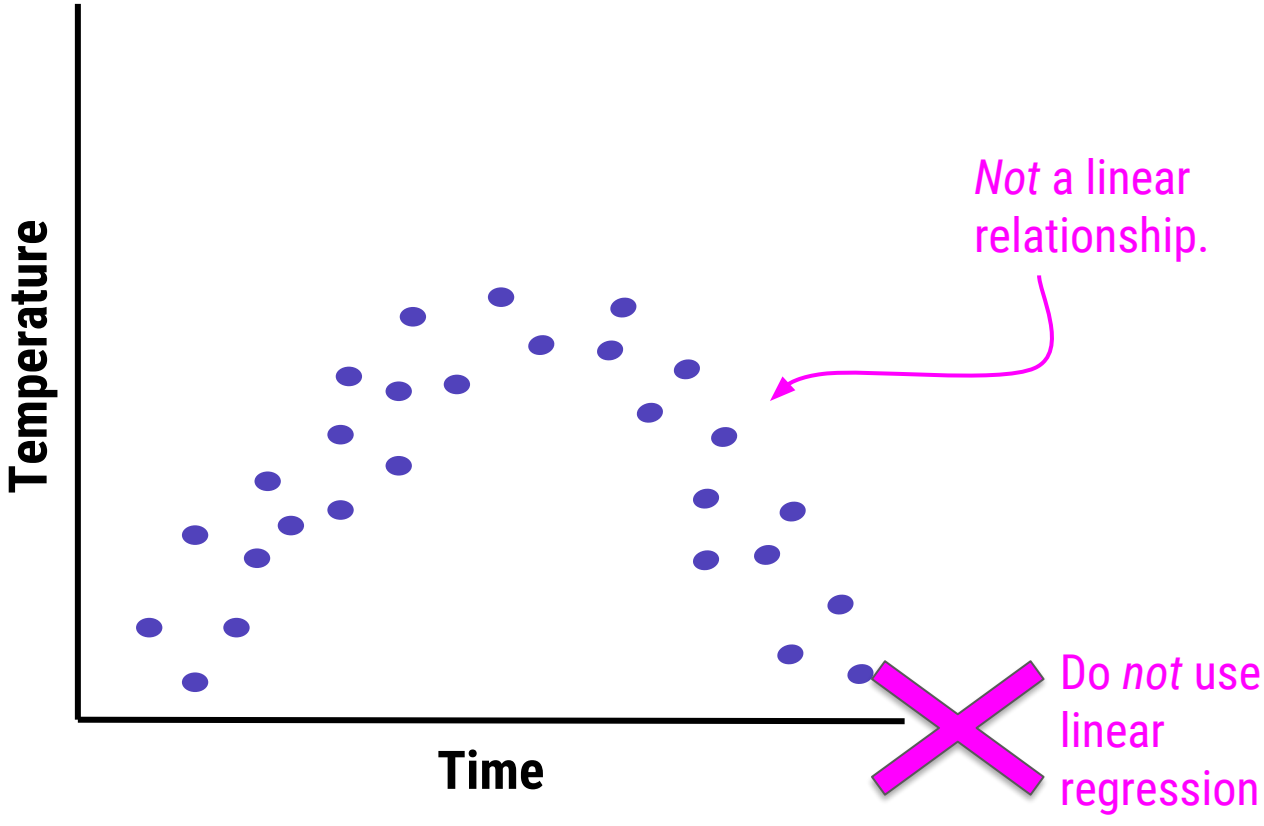

Dependent Variable
Independent Variable
positive relationship
The larger the independent variable value, the larger the dependent variable tends to be
The smaller the independent variable value, the smaller the dependent variable tends to be

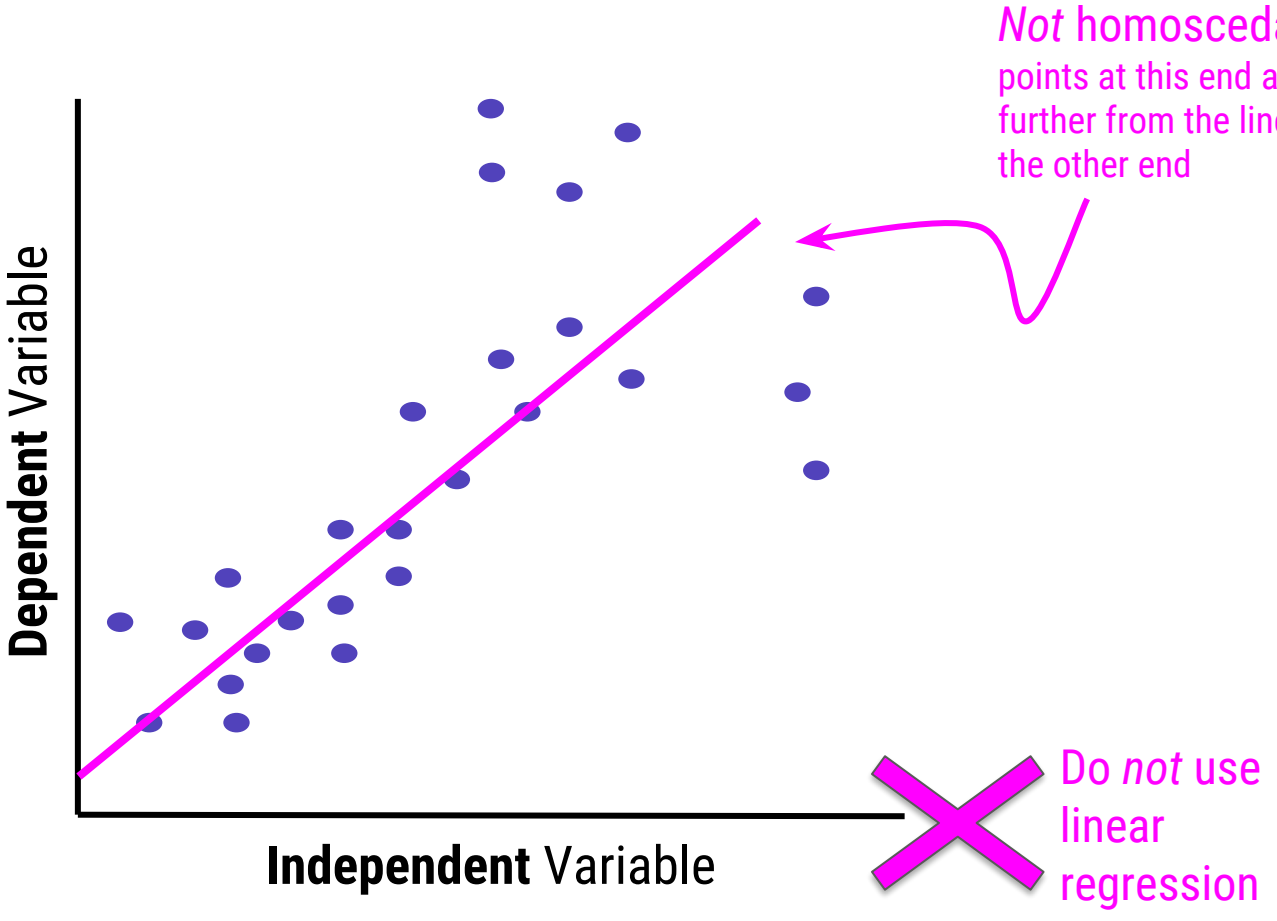

Son's Height
Father's Height
positive
relationship
The taller the father,
the taller the son
tends to be
The shorter the father,
the shorter the son
tends to be

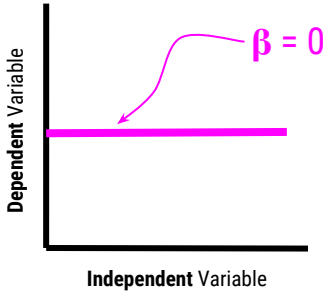

The *smaller* the independent variable value, the *larger* the dependent variable tends to be
Dependent Variable
negative relationship
The *larger* the independent variable value, the *smaller* the dependent variable tends to be
Independent Variable

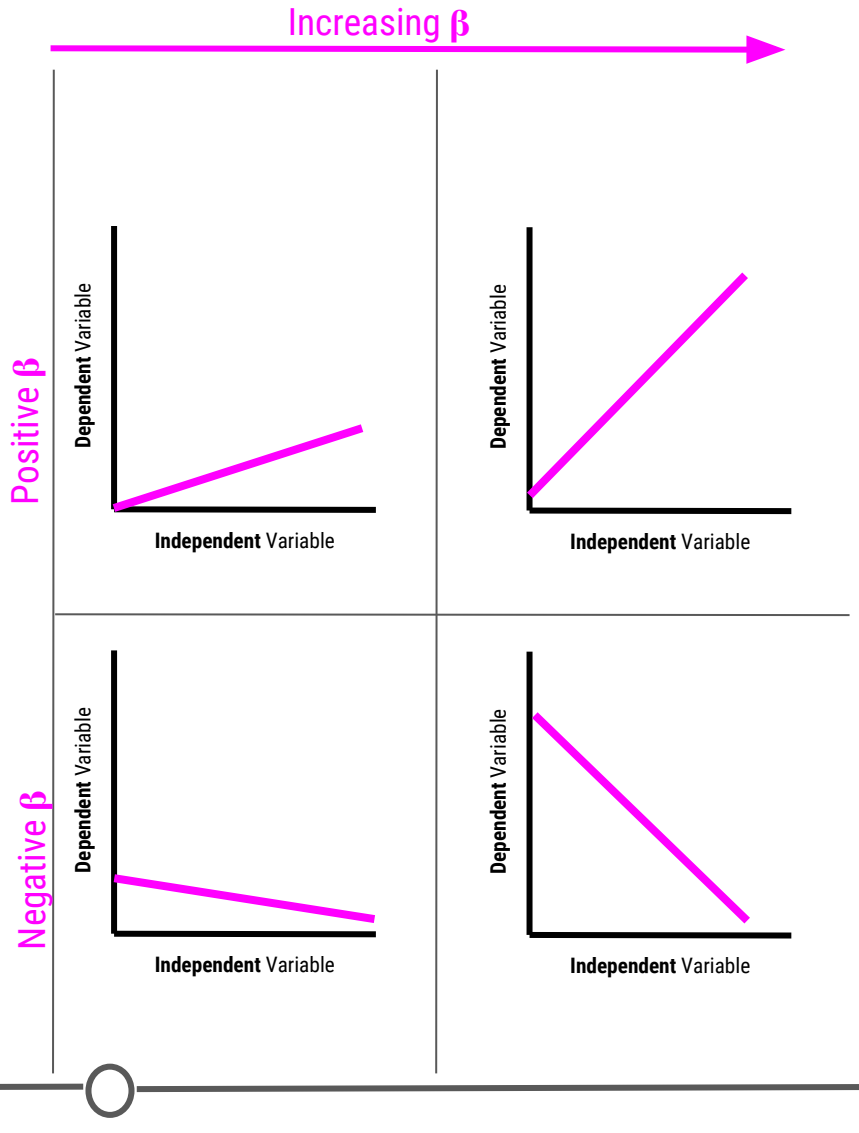

The *lower* the number of absences, the *higher* the students' grades tend to be
students' grades
negative relationship
The *more* absences, the *lower* the students' grades
# of absences

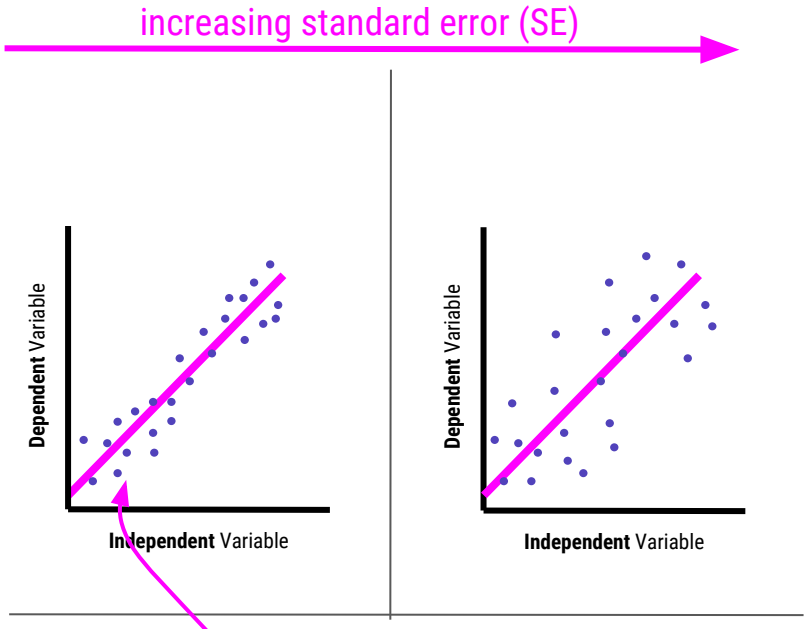

weaker relationship
Dependent Variable
Independent Variable
stronger relationship
Dependent Variable
stronger relationship
=
higher correlation
Independent Variable


Temperature
Time
Not a linear relationship.
Do *not* use linear regression


Not homoscedastic:
points at this end are much further from the line than at the other end
Dependent Variable
Independent Variable
Do *not* use linear regression


β = 0
Dependent Variable
Independent Variable


Increasing β
Positive β
Negative β
Dependent Variable
Independent Variable
Dependent Variable
Independent Variable
Dependent Variable
Independent Variable
Dependent Variable
Independent Variable



The *closer* the points are to the regression line, the *less uncertain* we are in our estimate

p-values : the probability of getting the observed results (or results more extreme) by chance alone

Girth, Height and Volume for Black Cherry Trees

Description

This data set provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. Note that girth is the diameter of the tree (in inches) measured at 4 ft 6 in above the ground.

Usage

```
trees
```

Format

A data frame with 31 observations on 3 variables.

`[,1]` `Girth` numeric Tree diameter in inches
`[,2]` `Height` numeric Height in ft
`[,3]` `Volume` numeric Volume of timber in cubic ft

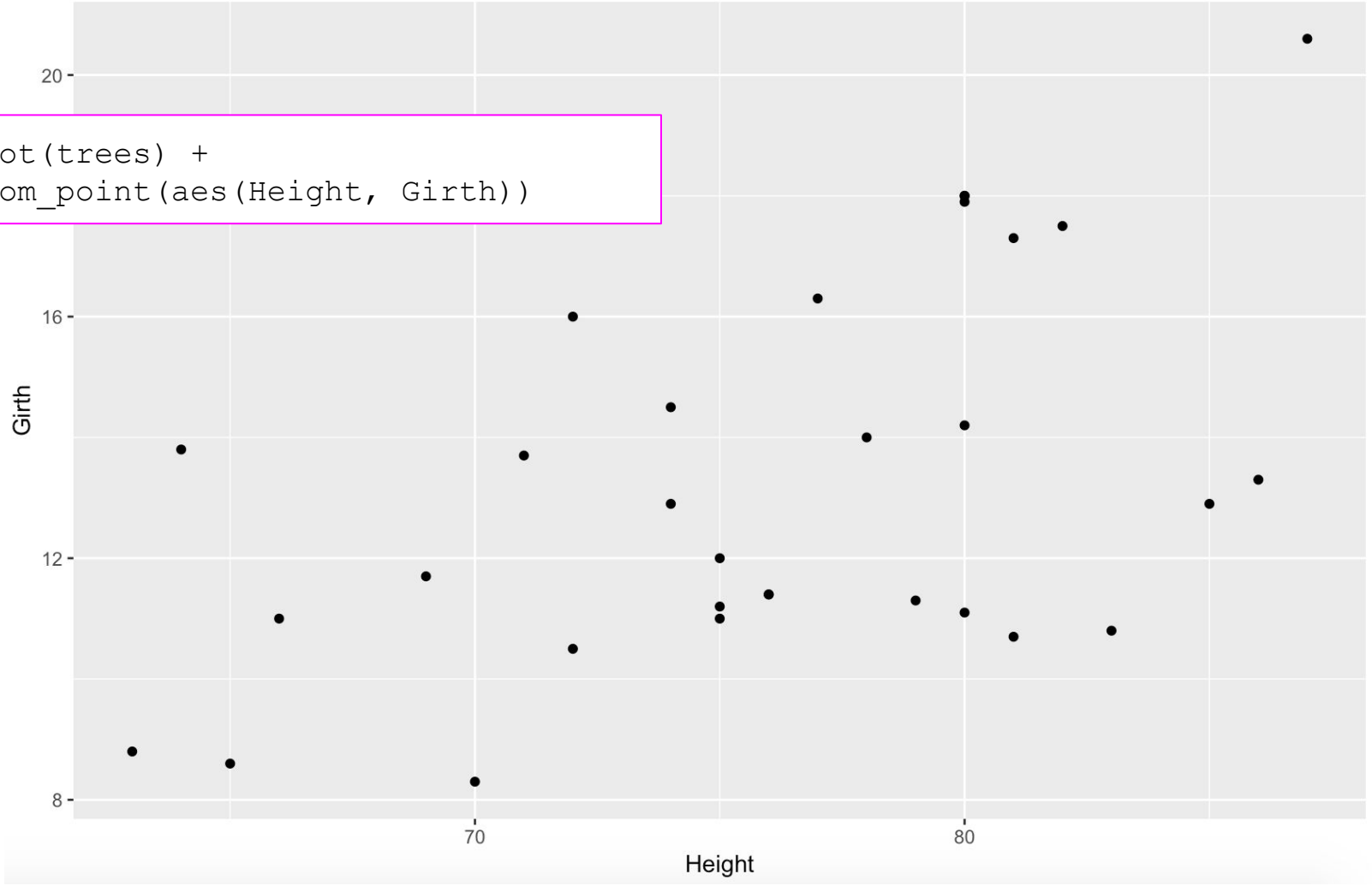
Source

Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1976) *The Minitab Student Handbook*. Duxbury Press.

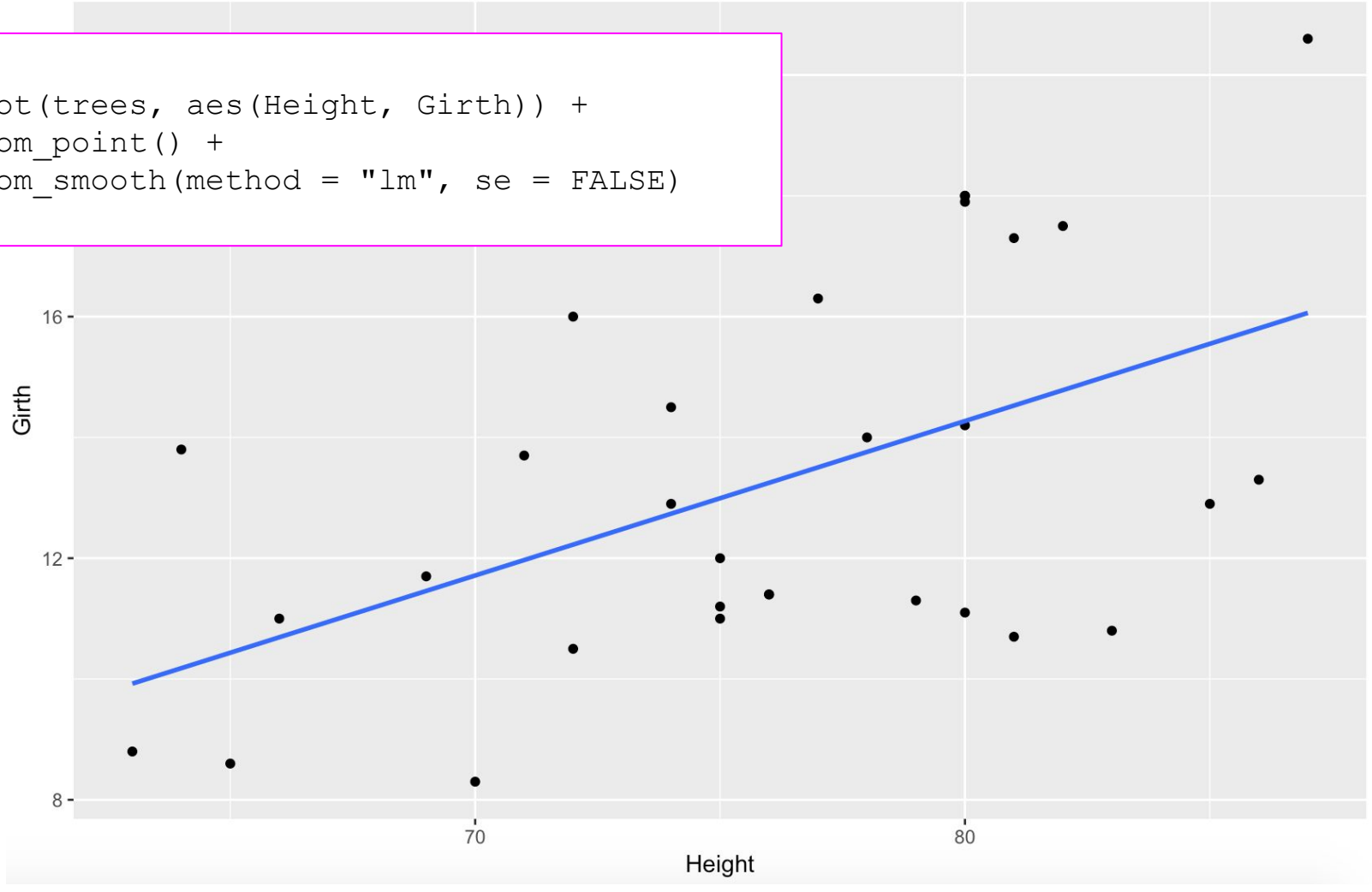
References

Atkinson, A. C. (1985) *Plots, Transformations and Regression*. Oxford University Press.

```
ggplot(trees) +
  geom_point(aes(Height, Girth))
```



```
ggplot(trees, aes(Height, Girth)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



```
> fit <- lm(Girth ~ Height , data = trees)
>
> summary(fit)

Call:
lm(formula = Girth ~ Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2386 -1.9205 -0.0714  2.7450  4.5384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.18839    5.96020  -1.038  0.30772
Height       0.25575    0.07816   3.272  0.00276 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.728 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

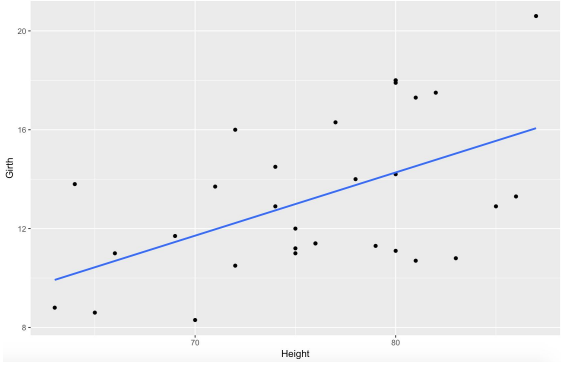
For every one inch increase in height, the girth will increase by 0.255 inches

β estimate

SE

p-value

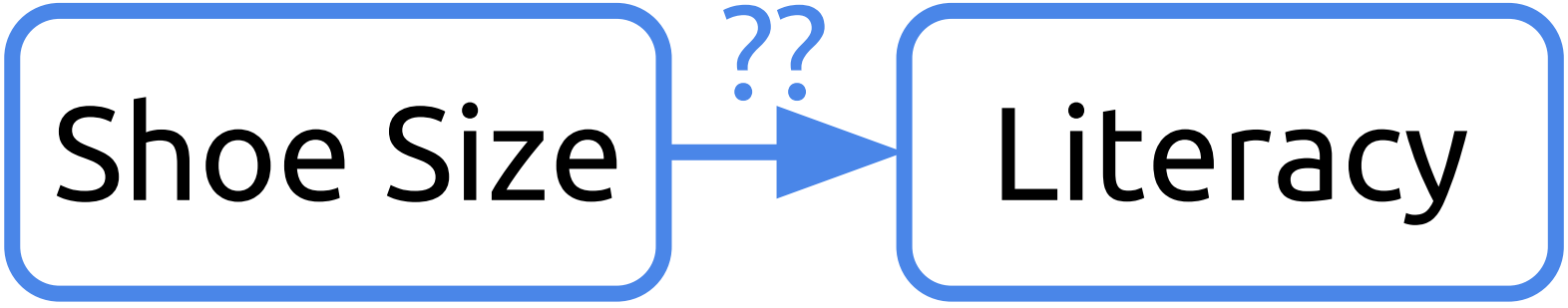
```
> fit <- lm(Girth ~ Height , data = trees)
>
> summary(fit)

Call:
lm(formula = Girth ~ Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2386 -1.9205 -0.0714  2.7450  4.5384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.18839    5.96020  -1.038  0.30772
Height       0.25575    0.07816   3.272  0.00276 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.728 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```



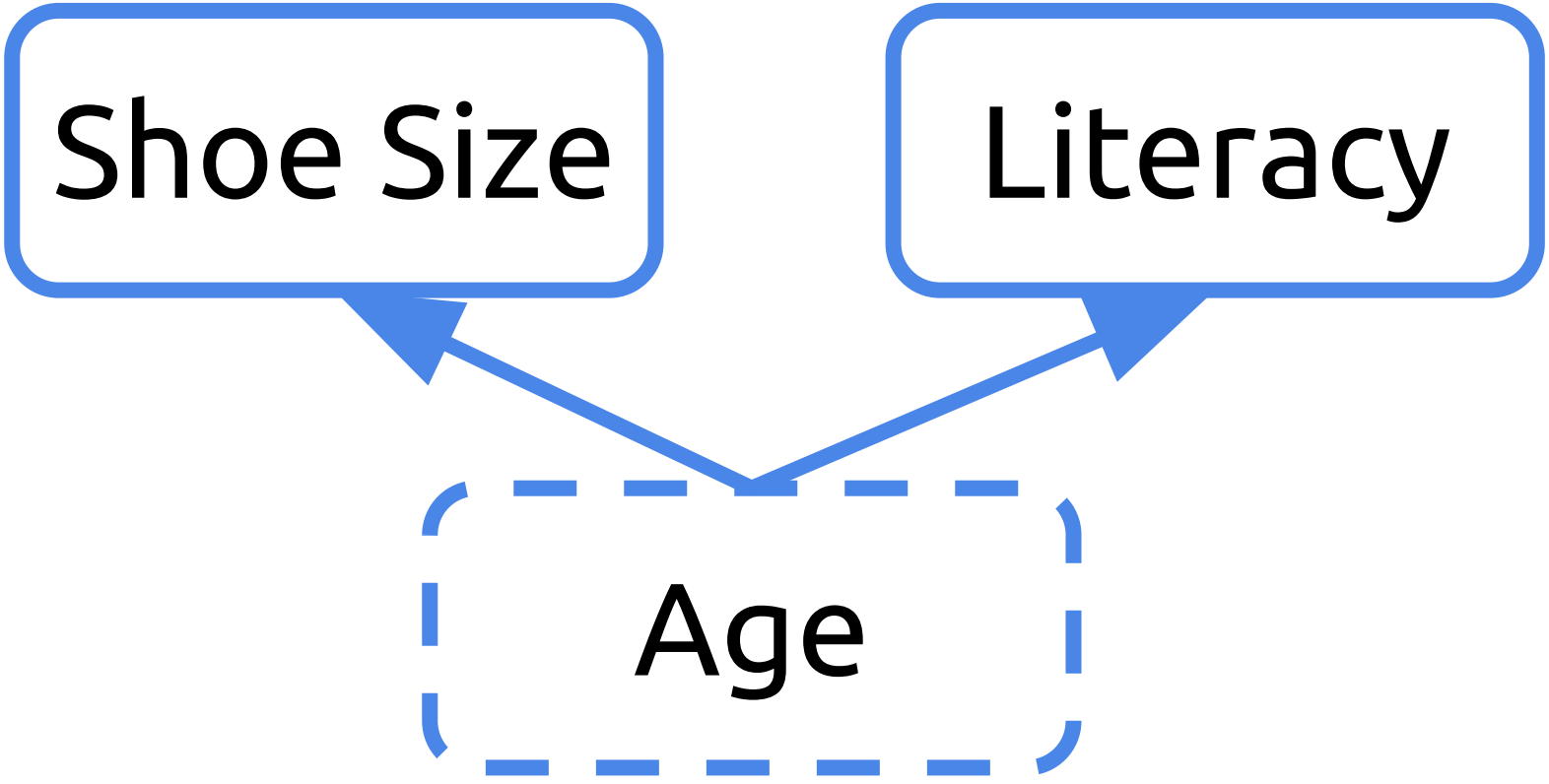
Describes the strength of the correlation

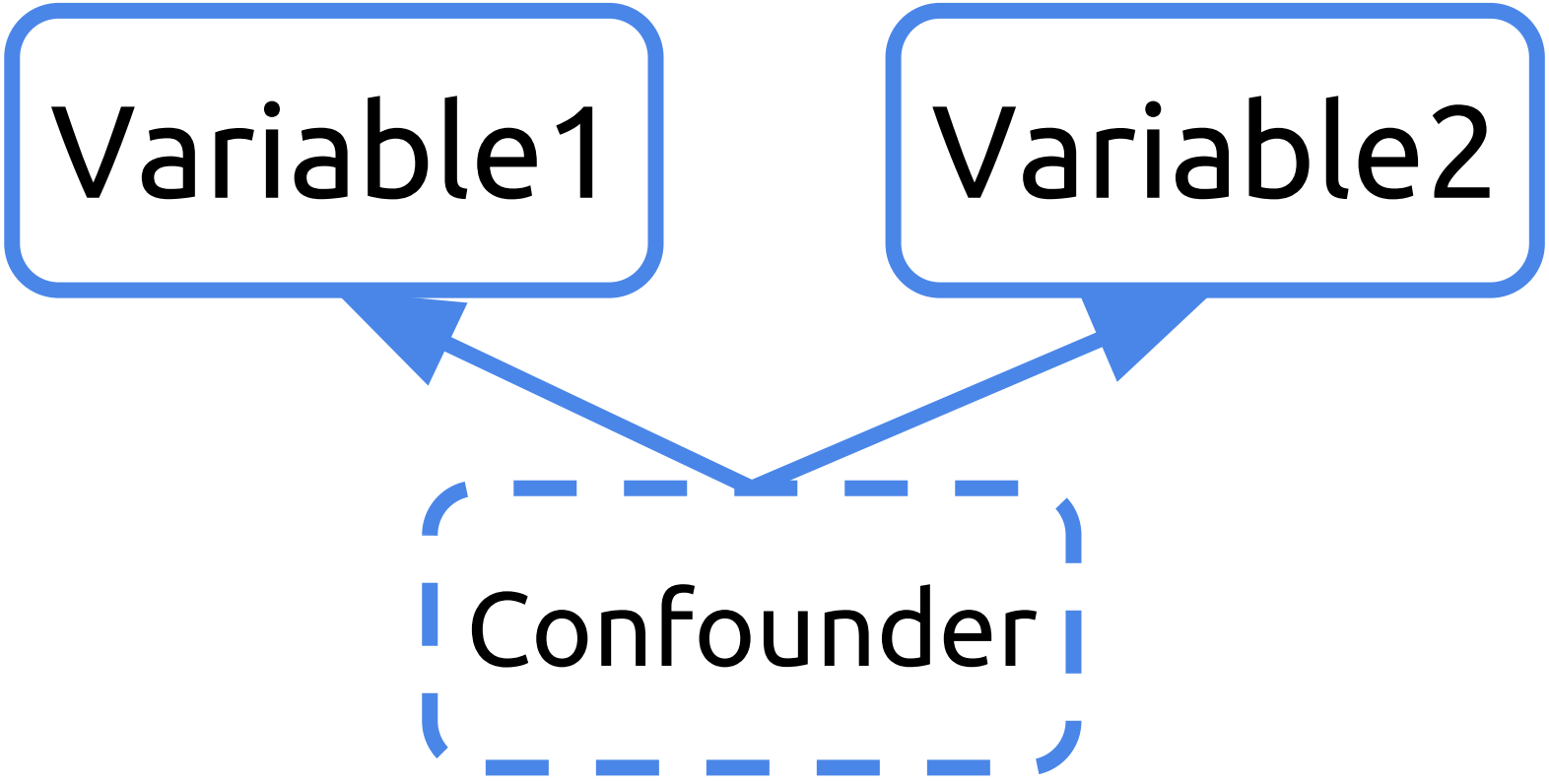
```
> #install.packages("broom")
> library(broom)
>
> tidy(fit)
         term   estimate std.error statistic     p.value
1 (Intercept) -6.1883945 5.9601994 -1.038286 0.307716768
2      Height  0.2557471 0.0781583  3.272169 0.002757815
```

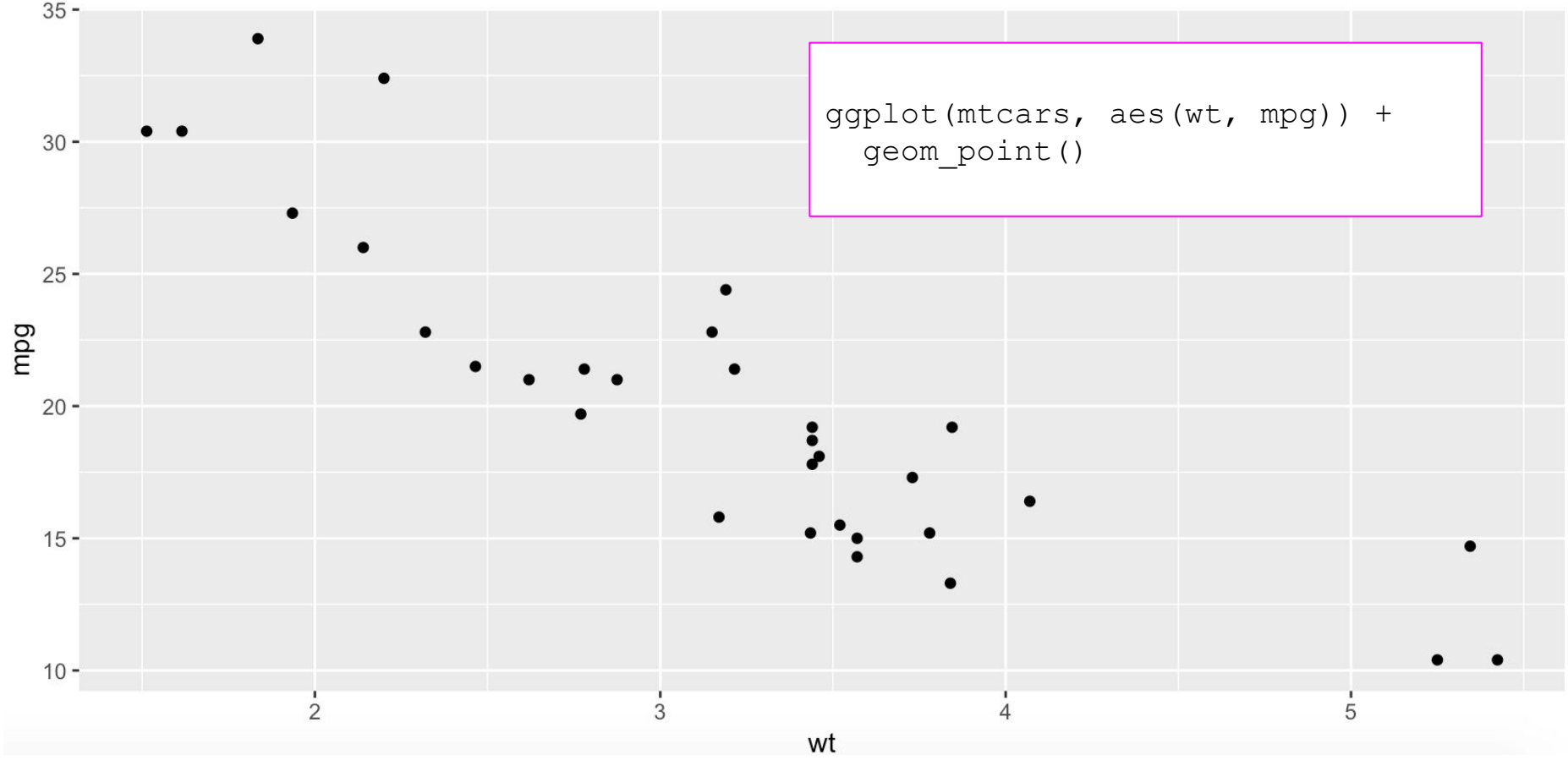

Small shoes
Not literate
Big shoes
Somewhat literate


Shoe Size
??
Literacy


Small shoes
Not literate
Young
Big shoes
Somewhat literate
Middle aged

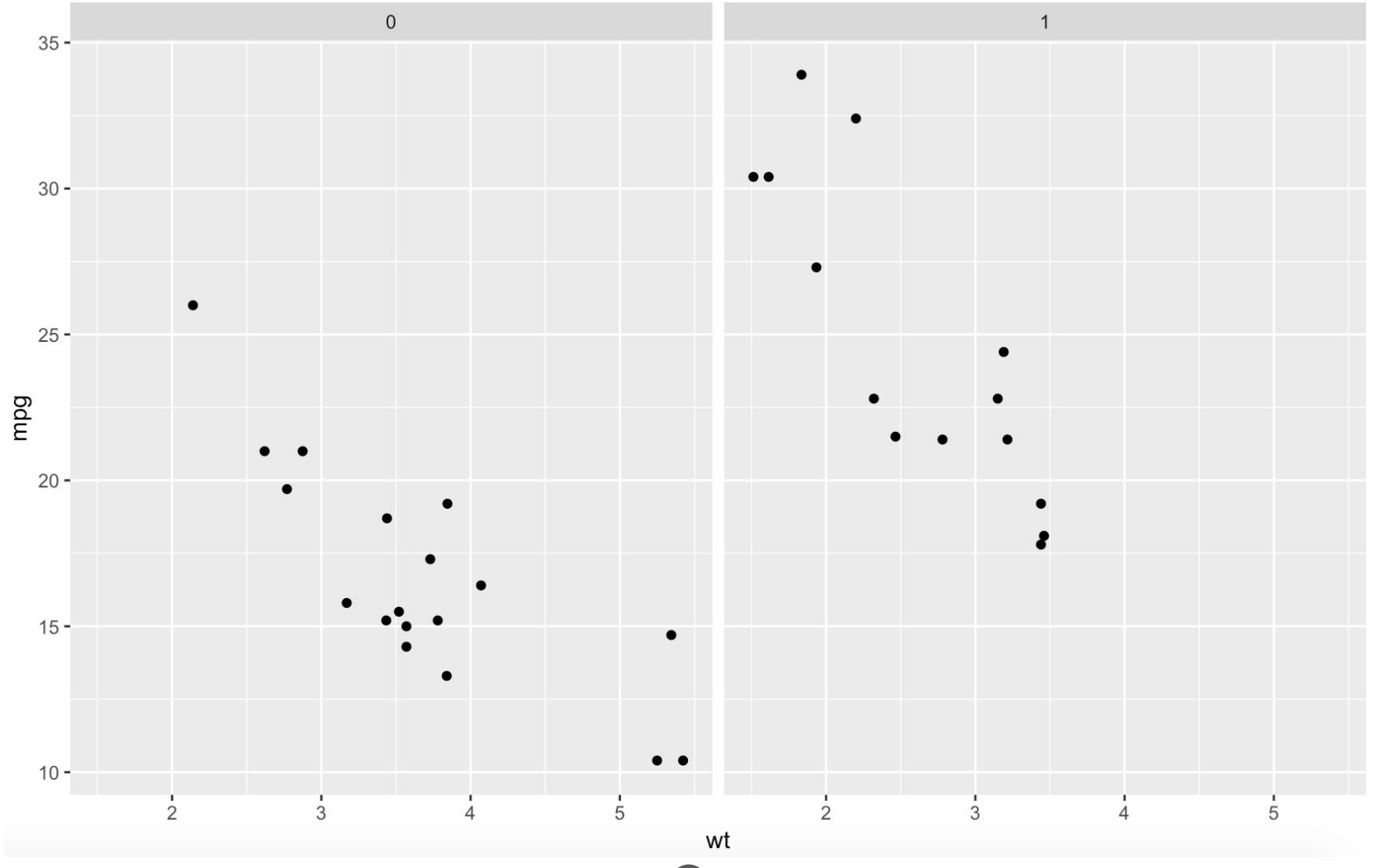

Shoe Size
Literacy
Age

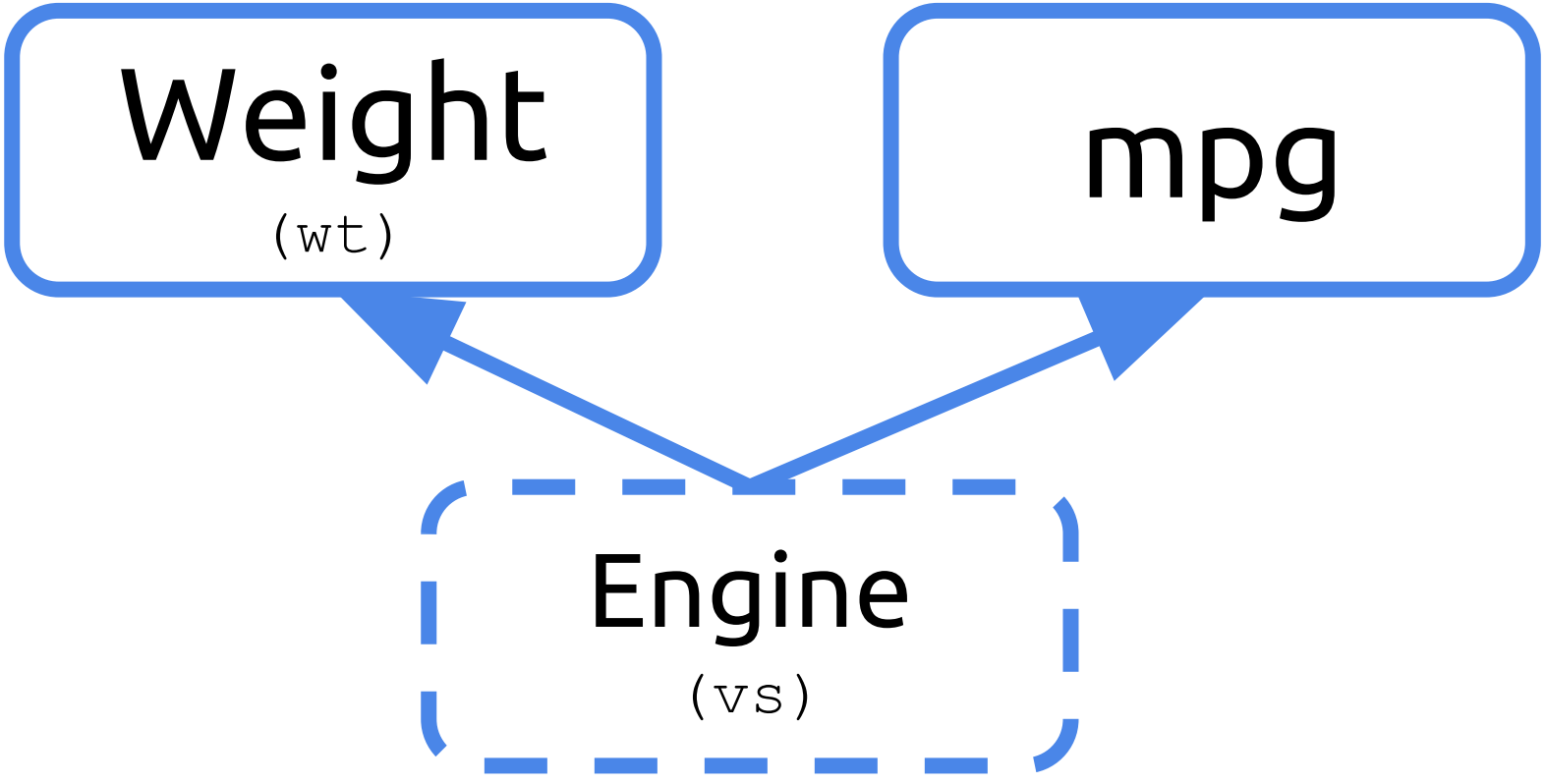

Variable1
Variable2
Confounder


ggplot(mtcars, aes(wt, mpg)) +
geom_point()
mpg
wt
35
30
25
20
15
10
2
3
4
5

```
> ## model the data without confounder
> fit <- lm(mpg ~ wt, data = mtcars)
> tidy(fit)
         term  estimate std.error statistic      p.value
1 (Intercept) 37.285126  1.877627 19.857575 8.241799e-19
2          wt -5.344472  0.559101 -9.559044 1.293959e-10
```

For every 1000 lb increase in weight, there is a 5.34 mpg decrease in gas mileage


0
1
35
30
25
20
15
10
mpg
2
3
4
5
wt

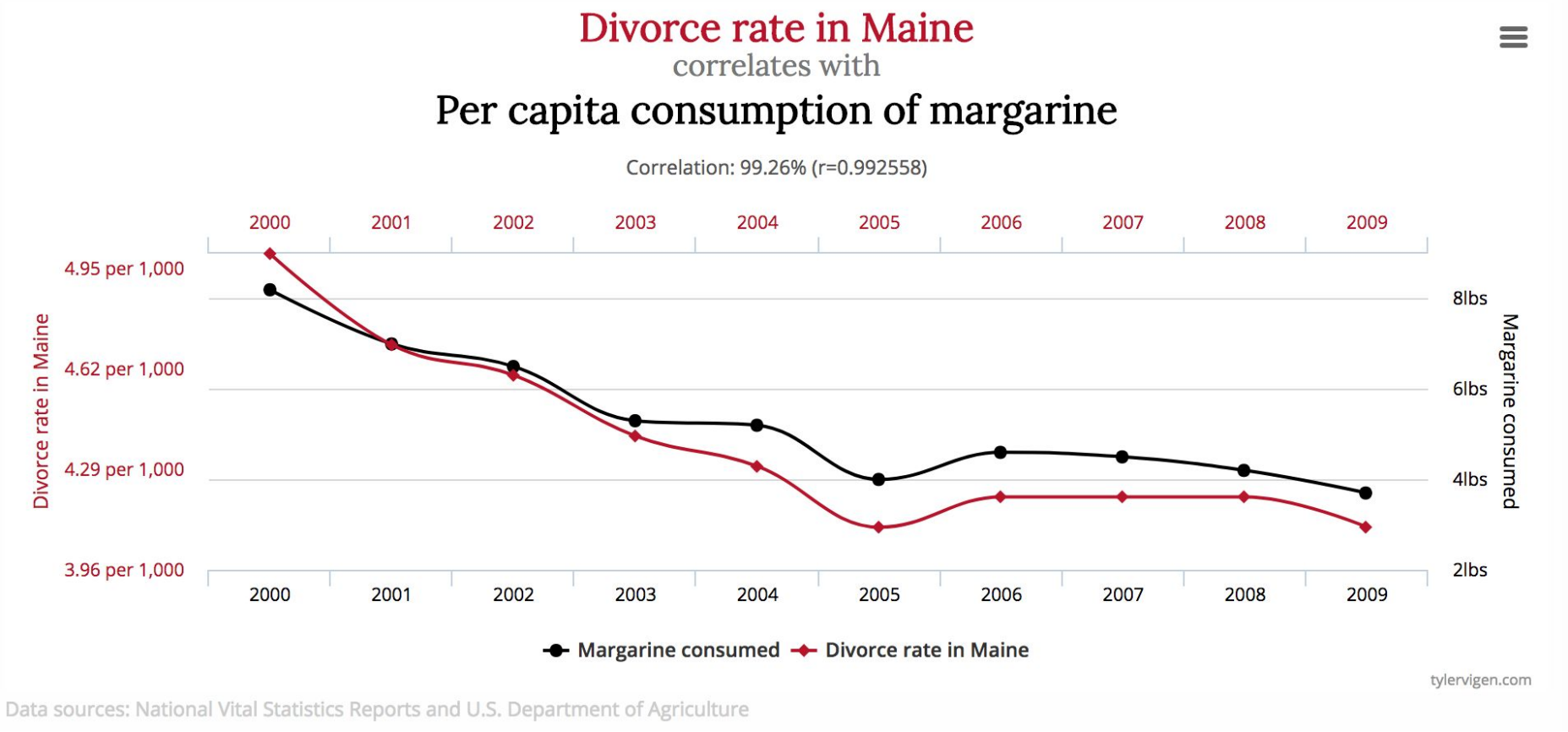

Weight
(wt)
mpg
Engine
(vs)

```
> ## include engine (vs) as a confounder
> fit <- lm(mpg ~ wt + vs, data = mtcars)
> tidy(fit)
         term  estimate std.error statistic      p.value
1 (Intercept) 33.004233 2.3553946 14.012188 1.920621e-14
2          wt -4.442814 0.6133645 -7.243350 5.632548e-08
3          vs  3.154367 1.1907378  2.649086 1.292580e-02
```

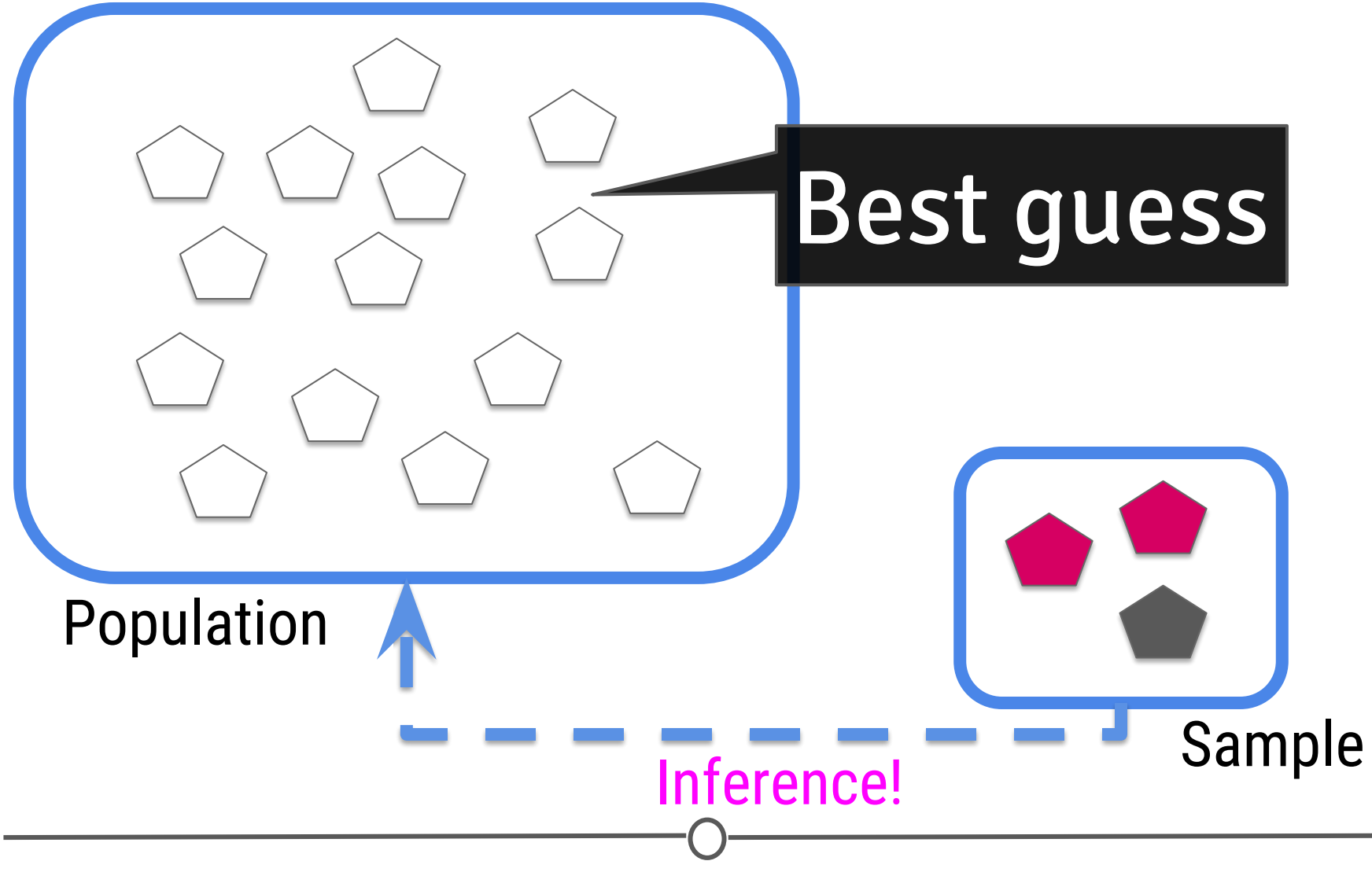
For a V-Shaped engine, for every 1000 lb increase in weight, there is a 4.44 mpg decrease in gas mileage

```
> ## include engine (vs) as a confounder
> fit <- lm(mpg ~ wt + vs, data = mtcars)
> tidy(fit)
         term  estimate std.error statistic      p.value
1 (Intercept) 33.004233 2.3553946 14.012188 1.920621e-14
2          wt -4.442814 0.6133645 -7.243350 5.632548e-08
3          vs  3.154367 1.1907378  2.649086 1.292580e-02
```

For two vehicles of the same weight, a straight engine will get 3.15 more mpg (on average) than a V-Shaped engine



http://www.tylervigen.com/spurious-correlations


Best guess
Population
Sample
Inference!