HEADS

PhD PROGRAMME IN HEALTH DATA SCIENCE

Descriptive Statistics

Examples in R

Basics of Health Intelligent Data Analysis

PhD Programme in Health Data Science

Cláudia Camila Dias

Pedro Pereira Rodrigues

U.PORTO
FMUP FACULDADE DE MEDICINA UNIVERSIDADE DO PORTO

MEDCIDS
DEPARTAMENTO DE MEDICINA DA COMUNIDADE, INFORMAÇÃO E DECISÃO EM SAÚDE

CINTESIS
Health. Research.

Variables

Classification of variables

variable → what is observed or measured

Types of variables:

Categorical: variable can take on one of a limited, and usually fixed, possible values

- **Ordinal**: is a categorical variable where the variables have natural, ordered categories (e.g. severity scores)
- **Nominal**: a variable with values which have no ordering value (e.g. gender)

Continuous: is one which can take on infinitely many, uncountable values

Categorical Data

Categorical data

Frequencies tables

Frequency count (n): number of observed cases in each category

Relative frequency (%): proportion of cases in each category

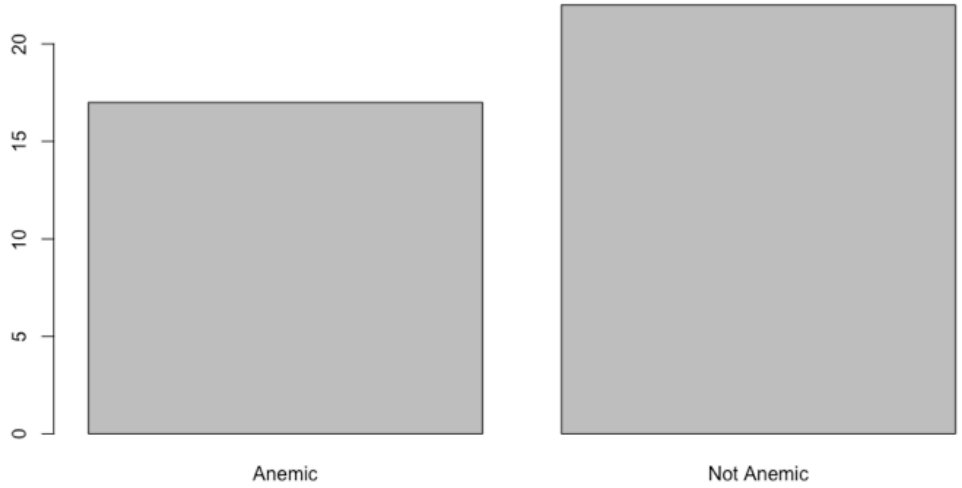
Table 2 - Socio-demographic characteristics of the participants in the study (n=300).

	Total	
	n	(%)
Gender		
Male	263	(87)
Female	37	(13)
Mother's ethnicity		
Caucasian	241	(82)
African American	19	(7)
Hispanic	13	(4)
Asian	15	(5)
Other	6	(2)
Father's ethnicity		
Caucasian	245	(83)
African American	20	(7)
Hispanic	13	(4)
Asian	11	(4)
Other	6	(2)
Mother's education		
High school	32	(11)
College	264	(89)
Father's education		
High school	52	(18)
College (Grad/Post Grad)	242	(82)

Categorical data – graphical representation

```
library(foreign)
develop.data <- as.data.frame(read.spss("develop.sav", use.missings=T))
attach(develop.data)
summary(group)
plot(group)
```

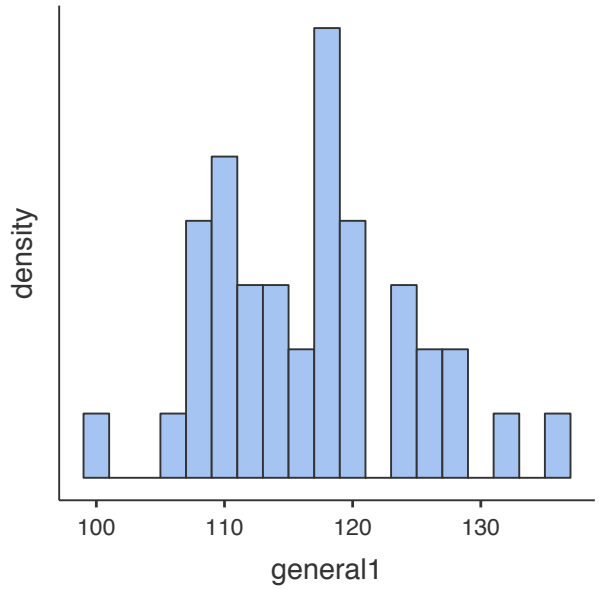
```
> summary(group)
   Anemic Not Anemic
       17         22
>
```



Continuous variables

Continuous data – graphical representation

Histogram – describe the distribution of continuous variables



Unlike the bar chart, the bars are adjacent.

Each bar represents the number of cases in a range of values.

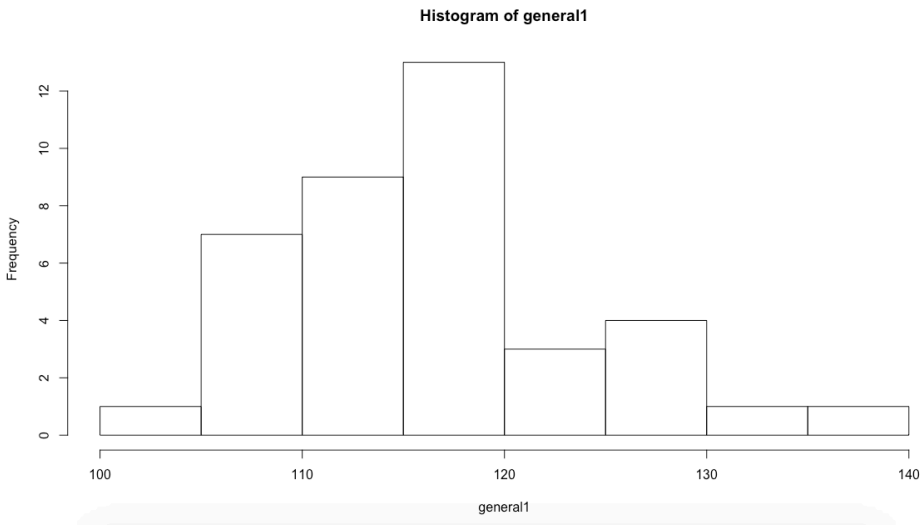
Descriptives

	general1
N	39
Missing	0
Mean	117
Median	118
Minimum	100
Maximum	136
25th percentile	111
50th percentile	118
75th percentile	120

Continuous data – graphical representation

```
library(foreign)
develop.data <- as.data.frame(read.spss("develop.sav", use.missings=T))
attach(develop.data)
summary(general1)
hist(general1)
```

```
> summary(general1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  100.0   111.0   118.0   117.2   120.0   136.0
>
```



Summary measures

Measures of central tendency

Frequency tables are not useful for describing continuous variables. We then use central tendency measures, which describe the center of the distribution.

Suppose:

X represents the variable " newborn weight"

Let x_i the weight of the newborn i and i=1....n (n newborn in the sample)

The average of X is defined by:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \mathrm{K} + x_n}{n}$$

Summary measures

Measures of central tendency

Suppose that

X represents the variable “newborn weight”

Let x_i the weight of the newborn i and i=1....n (n newborn in the sample)

If the values x_1, x_2, ..., x_n are sorted by ascending order, the **median** of a set of observations is just the middle value.

Summary measures

Measures of central tendency – Median (odd number of values)

Consider the following newborn weights in grams:

3434 2987 4189 2333 3761

Sorting the values, the median is

2333 2987 **3434** 3761 4189

Summary measures

Measures of central tendency – Median (even number of values)

Consider the following newborn weights in grams:

3434 2987 4189 3761

Sorting the values, the median is

2987 **3434 3761** 4189

(3434+3761)/2 = **3597.5**

Summary measures

Measures of central tendency – Median

If the median weight of newborns is 3300g, this means that

half of newborns in this sample are under 3300g and

half of newborns in this sample are above 3300g

Summary measures

So what measure of central tendency should I use?

Mean or median?

Both describe the center of the distribution…

Summary measures

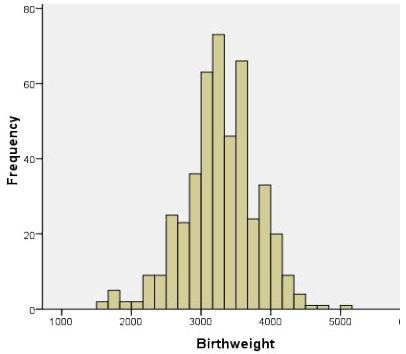
So what measure of central tendency should I use?

Summary measures	Continuous variables	Advantages	Disadvantages
Central tendency	Mean	Uses all data Algebraically defined	Distorted by extreme values
	Median	Not distorted by extreme values	Ignores too much information; Not algebraically defined

Summary measures

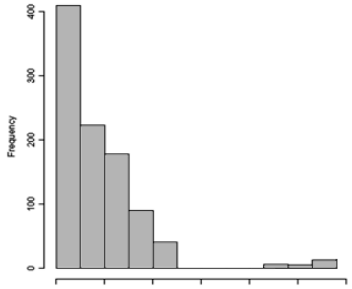
Symmetric distribution:

- mean = median;
- use the **mean** for its advantages.



Asymmetric distribution:

- the mean is affected by extreme values
- so, does not describe the centre of the distribution
- use the median



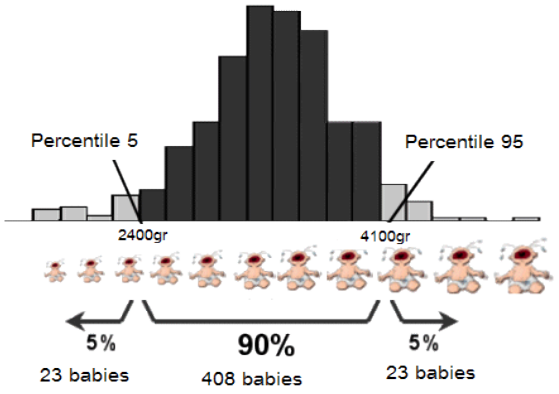
Summary measures - Percentile

The median is thus the cutoff point in the middle (50%) of the distribution.

- We can generalize this idea to other cutoff points (the percentiles)
- The 5th percentile is the value below which are 5% of cases (or above which there are 95% of cases) .
    - Between the 5th and 95th percentiles there are 90% of cases

Quartiles

- Percentile 25 : 1º quartile
- Percentile 50 : 2º quartile or median
- Percentile 75 : 3º quartile



Summary measures - Percentiles

But to describe the center of the distribution ...

is it enough to use measures of central tendency?

Two students had an equal average (15) in the masters course (10 units).

Grades of each of the 10 subjects:

15, 15, 15, 15, 15, 15, 15, 15, 15, 15

10, 10, 10, 10, 10, 20, 20, 20, 20, 20

Are both students' grades equally distributed?

Summary measures

We need measures that describe the dispersion of values:

Range : difference between the highest and the lowest value

Interquartile range: difference between percentiles 25 and 75

Variance: mean squared deviations from the mean $VAR(X) = S^2 = \frac{\sum (x_i - \overline{X})^2}{n}$

Standard deviation: squared root of variance $S = \sqrt{\frac{\sum (x_i - \overline{X})^2}{n}}$

Summary measures - variability

	Advantages	Disadvantages
Standard deviation	Uses all the data; Algebraically defined; Same units of the data;	Sensitive to extreme values;
Percentile	Not distorted by extreme values	Not to be calculated on small samples
Range	Easy to compute	Uses only 2 values; Distorted by extreme values

Summary measures

min(general1)
quantile(general1,.25)
median(general1)
mean(general1)
quantile(general1,.75)
max(general1)

sd(general1)
range(general1)

```
> min(general1)
[1] 100
> quantile(general1,.25)
25%
111
> median(general1)
[1] 118
> mean(general1)
[1] 117.1538
> quantile(general1,.75)
75%
120
> max(general1)
[1] 136
>
```

```
> sd(general1)
[1] 7.548226
> range(general1)
[1] 100 136
```

Relate Two Variables

Categorical X Categorical

Contingency tables

- are used to study the relationship between two categorical variables

- describe the frequency of the categories of one of the variables relatively to the categories of the other

			smoked in early pregnancy		
			No	Yes	Total
Mother's age	13-20 Years	Count	13	7	20
		% within Mother's age	65,0%	35,0%	100,0%
		% within smoked in early pregnancy	3,4%	12,1%	4,5%
	21-30 Years	Count	237	42	279
		% within Mother's age	84,9%	15,1%	100,0%
		% within smoked in early pregnancy	61,7%	72,4%	63,1%
	31-35 Years	Count	82	6	88
		% within Mother's age	93,2%	6,8%	100,0%
		% within smoked in early pregnancy	21,4%	10,3%	19,9%
	36-55 Years	Count	52	3	55
		% within Mother's age	94,5%	5,5%	100,0%
		% within smoked in early pregnancy	13,5%	5,2%	12,4%
Total		Count	384	58	442
		% within Mother's age	86,9%	13,1%	100,0%
		% within smoked in early pregnancy	100,0%	100,0%	100,0%

Categorical X Categorical

Contingency Table

			smoked in early pregnancy		
			No	Yes	Total
Mother's age	13-20 Years	Count	13	7	20
		% within Mother's age	65,0%	35,0%	100,0%
		% within smoked in early pregnancy	3,4%	12,1%	4,5%
	21-30 Years	Count	237	42	279
		% within Mother's age	84,9%	15,1%	100,0%
		% within smoked in early pregnancy	61,7%	72,4%	63,1%
	31-35 Years	Count	82	6	88
		% within Mother's age	93,2%	6,8%	100,0%
		% within smoked in early pregnancy	21,4%	10,3%	19,9%
	36-55 Years	Count	52	3	55
		% within Mother's age	94,5%	5,5%	100,0%
		% within smoked in early pregnancy	13,5%	5,2%	12,4%
Total		Count	384	58	442
		% within Mother's age	86,9%	13,1%	100,0%
		% within smoked in early pregnancy	100,0%	100,0%	100,0%

Categorical X Categorical

table(group,soclass)

```
> table(soclass, group)
        group
soclass  Anemic Not Anemic
  High       3          6
  Medium     7          8
  Low        7          8
>
```

library(gmodels)
CrossTable(soclass, group, format="SPSS")

```
> CrossTable(soclass, group, format="SPSS")

   Cell Contents
|-------------------------|
|                   Count |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  39

             | group
     soclass |     Anemic  | Not Anemic  |  Row Total |
-------------|-------------|-------------|-------------|
        High |           3 |           6 |           9 |
             |       0.217 |       0.168 |             |
             |      33.333% |      66.667% |      23.077% |
             |      17.647% |      27.273% |             |
             |       7.692% |      15.385% |             |
-------------|-------------|-------------|-------------|
      Medium |           7 |           8 |          15 |
             |       0.033 |       0.025 |             |
             |      46.667% |      53.333% |      38.462% |
             |      41.176% |      36.364% |             |
             |      17.949% |      20.513% |             |
-------------|-------------|-------------|-------------|
         Low |           7 |           8 |          15 |
             |       0.033 |       0.025 |             |
             |      46.667% |      53.333% |      38.462% |
             |      41.176% |      36.364% |             |
             |      17.949% |      20.513% |             |
-------------|-------------|-------------|-------------|
Column Total |          17 |          22 |          39 |
             |      43.590% |      56.410% |             |
-------------|-------------|-------------|-------------|
```

Categorical X Categorical

Table 3
Reoperation in patients with abdominal surgery after diagnosis (n = 767).

	Reoperation					
	No (n = 306,61%)			Yes (n = 192,39%)		
	n	Col%	Row%	n	Col%	Row%
Gender						
Male	132	43%	58%	95	49%	42%
Female	174	57%	64%	97	51%	36%
Smoking habits						
Never smoke	157	57%	67%	79	47%	33%
Ex-smoker	59	21%	60%	40	24%	40%
Smoker	59	21%	55%	49	29%	45%
Age at diagnosis						
A1	31	10%	62%	19	10%	38%
A2	209	68%	57%	155	81%	43%
A3	66	22%	79%	18	9%	21%
Location						
L1	155	56%	63%	92	52%	37%
L2	22	8%	71%	9	5%	29%
L3	102	37%	58%	75	43%	42%
L4						
No	269	91%	63%	157	87%	37%
Yes	25	9%	52%	23	13%	48%
Behaviour						
B1	56	19%	81%	13	7%	19%
B2	106	36%	57%	79	44%	43%
B3	132	45%	60%	88	49%	40%
Perianal disease						
No	250	82%	65%	135	70%	35%
Yes	56	18%	50%	57	30%	50%
Time between diagnosis and surgery						
0–6 months	129	43%	62%	80	42%	38%
7–13 months	28	9%	58%	20	11%	42%
13–36 months	46	15%	64%	26	14%	36%
>36 months	100	33%	61%	64	34%	39%

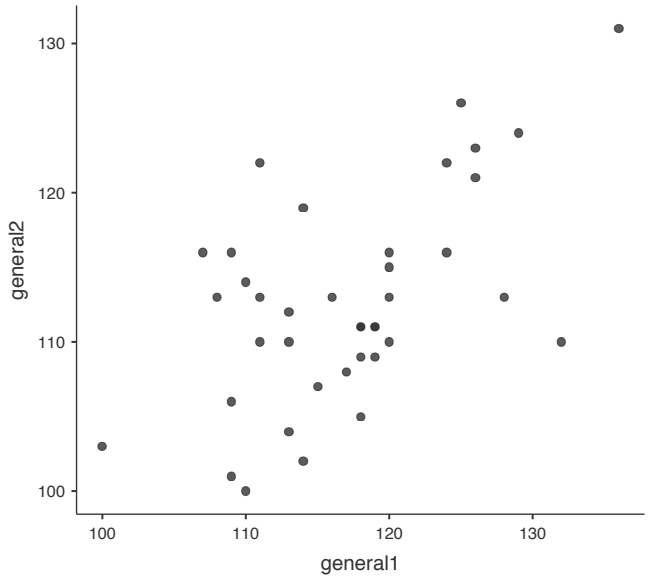
What % of women are reoperated?

Of the reoperated patients, what % has no perianal disease ?

Continuous x Continuous

Scatter plots

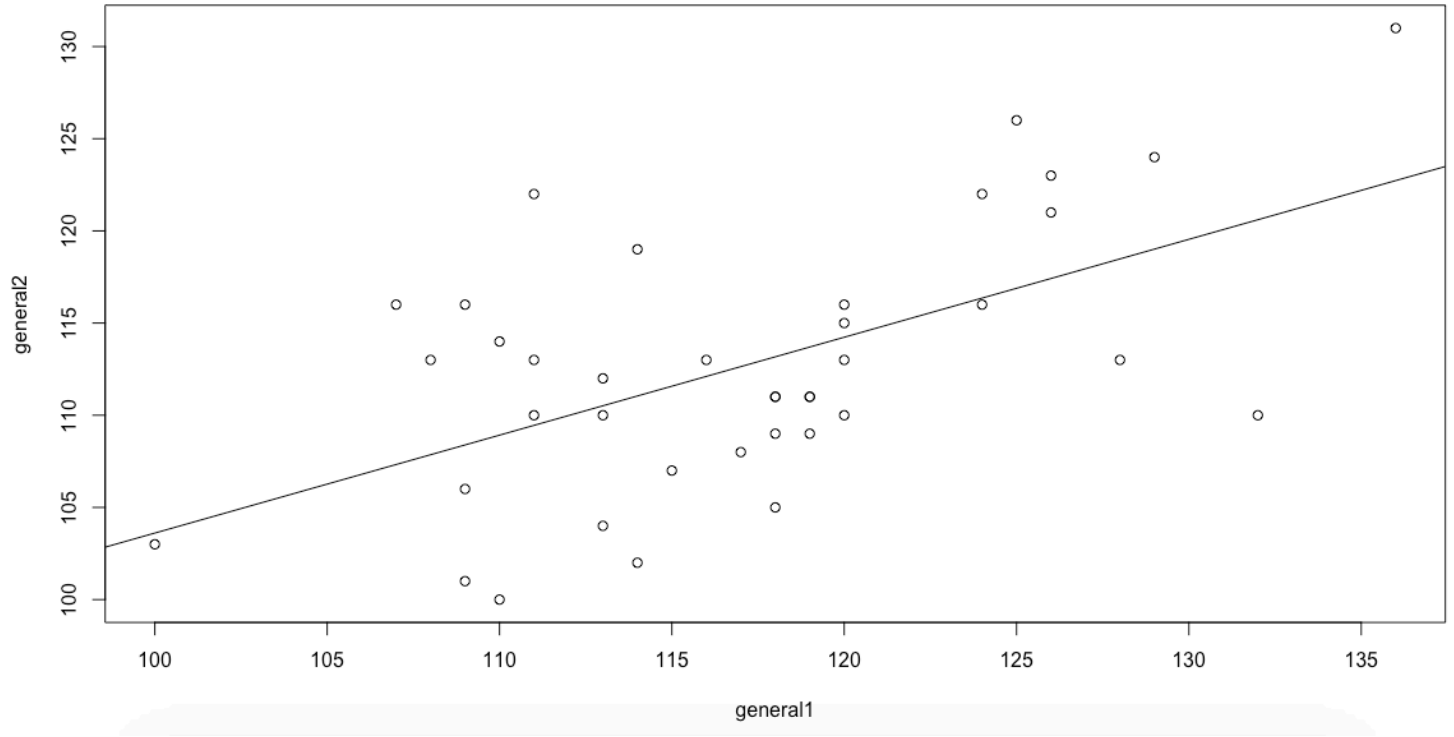
- are used to study the relationship between two continuous variables
- describe the position of each case in a frame where the coordinates are defined by the values of each variable



Continuous x Continuous

plot(x=general1, y=general2)

abline(lm(general2 ~ general1))

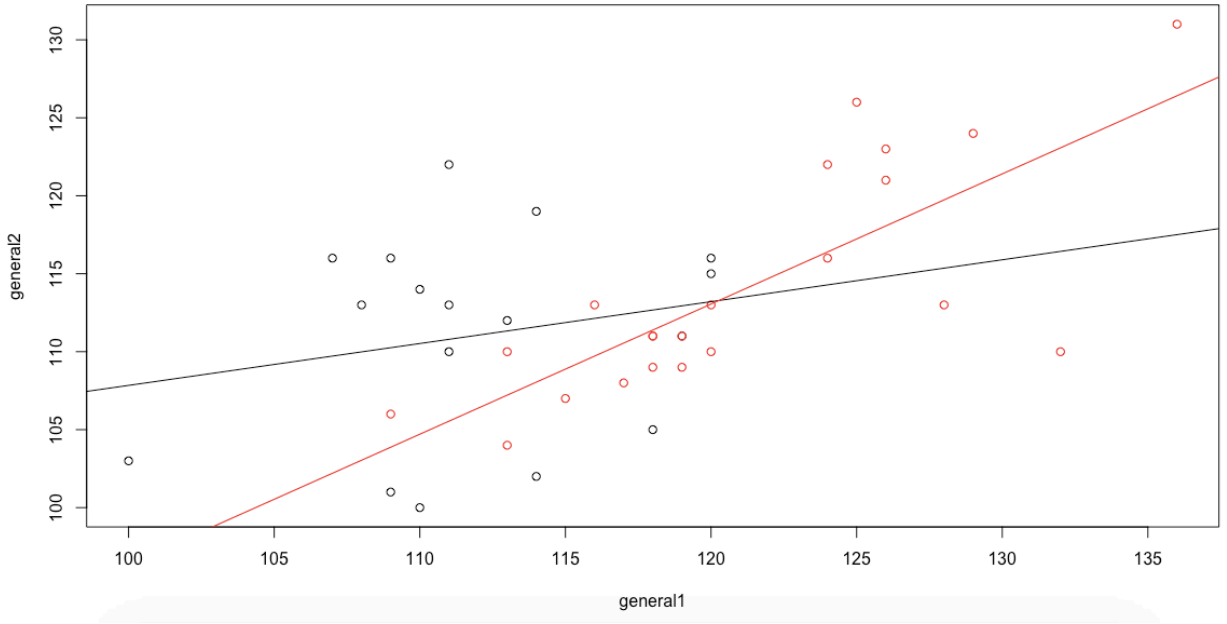


Continuous x Continuous

plot(general1, general2,col=group)

abline(lm(general2 ~ general1, subset=group==“Anemic”), col=1)

abline(lm(general2 ~ general1, subset=group==“Not Anemic”), col=“red”)

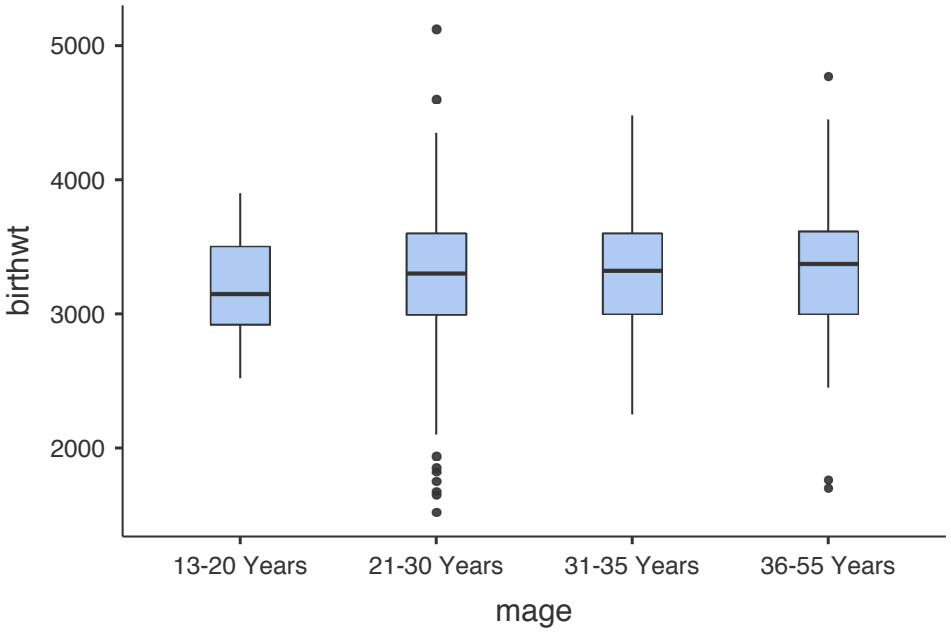


Continuous x Categorical

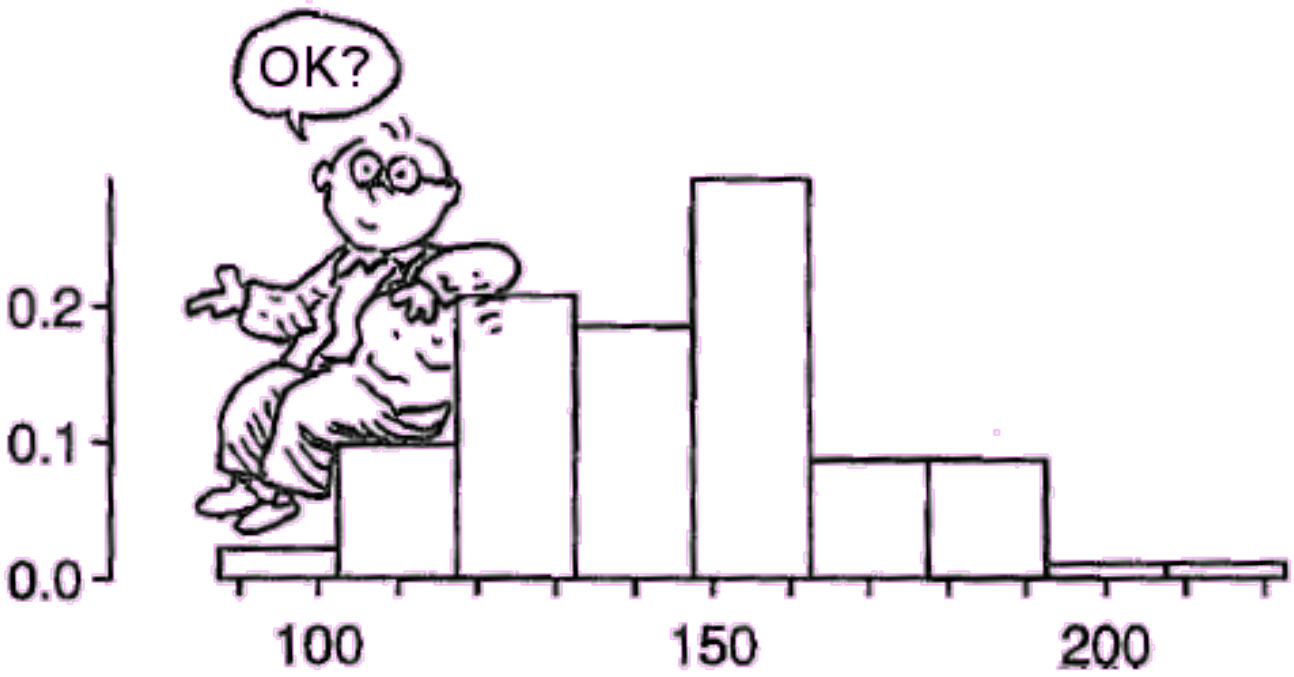
Boxplots

- are used to present the distribution of a continuous variable

- can easily be compared for different categories of a categorical variable



Boxplot (birthwt ~mage)


OK?
0.2
0.1
0.0
100
150
200