Math 183
Statistical Methods

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Math 183
Statistical Methods

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today:

- Presentation of the course
- Chapter 1: Introduction to data

Course Home

Instructor's webpage:

`www.math.ucsd.edu/~eaamari/`

- Lecture slides
- Homework sets
- Course syllabus
- Provisional course calendar
- Links to R and RStudio
- Office hour info and location

How the Course is Graded

The one following formula giving you the better result will be used:

Formula 1		Formula 2	
20%	Homework	20%	Homework
20%	Midterm Exam 1	20%	Best Midterm Exam
20%	Midterm Exam 2	60%	Final Exam
40%	Final Exam		

- Your worst homework grade will be dropped for computing your final *Homework* score.
- No makeup exams.
- The grading scheme will be curved and scaled to the best student in class.

Homework

- Homework is due weekly on Friday's lecture.
- Late assignments will not be accepted.
- Your worst homework grade will be dropped.
- Randomly selected problems on the assignment will be graded.

Tacit homework: Read the textbook!

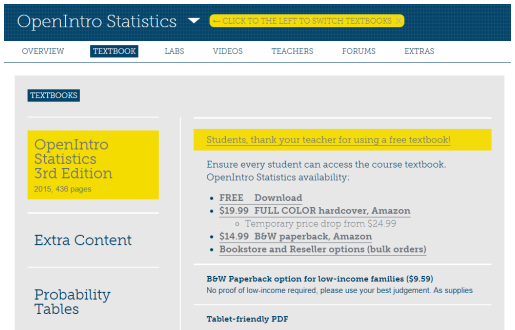
- Homework handed back on Discussion sections.
- No homework re-grading will be allowed after the section ends. This means that if you come back after you went out the room, your grade is fixed and your homework will not be regraded. Complaints/reclamation during the section will be considered with concern.

Class Calendar

	Monday	Tuesday	Wednesday	Thursday	Friday
Week 1	October 2 Chapter 1	3	4 Chapter 2	5 Discussion	6 Chapter 2 HW 1 due
Week 2	9 Chapter 2	10	11 Chapter 2	12 Discussion	13 Chapter 3 HW 2 due
Week 3	16 Chapter 3	17	18 Chapter 3	19 Discussion	20 Midterm Exam I
Week 4	23 Chapter 3	24	25 Chapter 3	26 Discussion	27 Chapter 4 HW 3 due
Week 5	30 Chapter 4	31	November 1 Chapter 4	2 Discussion	3 Chapter 4 HW 4 due
Week 6	6 **Review**	7	8 Midterm Exam II	9 Discussion	10 Veterans Day
Week 7	13 Chapter 5	14	15 Chapter 5	16 Discussion	17 Chapter 5 HW 5 due
Week 8	20 Chapter 5	21	22 Chapter 6	23 Thanksgiving	24 Thanksgiving
Week 9	27 Chapter 6	28	29 Chapter 6	30 Discussion	December 1 HW 6 due
Week 10	4 Chapter 7	5	6 Chapter 7	7 Discussion	8 **Review** HW 7 due
Final Week	11	12	13	14 Final Exam 11:30am-2:29pm	15

Components you Need

Textbook: *OpenIntro Statistics, Third edition*, by Diez, Barr, & Cetinkaya-Rundel
→ Free pdf available online.



Software: *R and RStudio*
→ Open-source statistical programming language and development environment used in data analysis.

Calculators:

- Used on exams and homework
- Need not be graphics, nor have statistical functions
- Cannot be your phone or computer (for exams)

Content of this Course

The lectures will cover Chapters 1 to 7 of the textbook.

- Introduction to data
- Introduction to probability:
  - Discrete and continuous random variables
  - Binomial, Poisson and Gaussian distributions
  - Central limit theorem
- Data analysis and inferential statistics:
  - Graphical techniques
  - Confidence intervals, hypothesis testing
  - Curve fitting (regression)

Before Carrying On...

Any questions so far?

What Does Data Look Like?

RStudio Source Editor

email50

Filter

	spam	to_multiple	from	cc	sent_email	time	image	attach	dollar	winner	inherit	viagra	password	num_char	line_breaks	format	r
1	0	0	1	0	1	2012-01-04 05:19:16	0	0	0	no	0	0	0	21.705	551	1	
2	0	0	1	0	0	2012-02-16 12:10:06	0	0	0	no	0	0	0	7.011	183	1	
3	1	0	1	4	0	2012-01-04 07:36:23	0	2	0	no	0	0	0	0.631	28	0	
4	0	0	1	0	0	2012-01-04 09:49:52	0	0	0	no	0	0	0	2.454	61	0	
5	0	0	1	0	0	2012-01-27 01:34:45	0	0	9	no	0	0	1	41.623	1088	1	
6	0	0	1	0	0	2012-01-17 09:31:57	0	0	0	no	0	0	0	0.057	5	0	
7	0	0	1	0	0	2012-03-17 21:18:55	0	0	0	no	0	0	0	0.809	17	0	
8	0	0	1	0	1	2012-03-31 06:58:56	0	0	0	no	0	0	0	5.229	88	1	
9	0	0	1	1	1	2012-01-10 17:57:54	0	0	0	no	0	0	0	9.277	242	1	
10	0	0	1	0	0	2012-01-07 11:29:16	0	0	23	no	0	0	0	17.170	578	1	
11	0	0	1	0	0	2012-02-22 16:57:02	0	0	4	no	0	0	0	64.401	1167	1	
12	0	0	1	0	0	2012-02-04 15:26:09	0	0	0	no	0	0	2	10.368	198	1	
13	1	0	1	0	0	2012-01-24 08:15:56	0	0	3	yes	0	0	0	42.793	712	1	
14	1	1	1	0	0	2012-02-08 18:22:46	0	2	2	no	0	0	0	0.451	24	0	

Showing 1 to 15 of 50 entries

What Does Data Look Like?

RStudio Source Editor

email50

Filter

	spam	to_multiple	from	cc	sent_email	time	image	attach	dollar	winner	inherit	viagra	password	num_char	line_breaks	format
1	0	0	1	0	1	2012-01-04 05:19:16	0	0	0	no	0	0	0	21.705	551	1
2	0	0	1	0	0	2012-02-16 12:10:06	0	0	0	no	0	0	0	7.011	183	1
3	1	0	1	4	0	2012-01-04 07:36:23	0	2	0	no	0	0	0	0.631	28	0
4	0	0	1	0	0	2012-01-04 09:49:52	0	0	0	no	0	0	0	2.454	61	0
5	0	0	1	0	0	2012-01-27 01:34:45	0	0	9	no	0	0	1	41.623	1088	1
6	0	0	1	0	0	2012-01-17 09:31:57	0	0	0	no	0	0	0	0.057	5	0
7	0	0	1	0	0	2012-03-17 21:18:55	0	0	0	no	0	0	0	0.809	17	0
8	0	0	1	0	1	2012-03-31 06:58:56	0	0	0	no	0	0	0	5.229	88	1
9	0	0	1	1	1	2012-01-10 17:57:54	0	0	0	no	0	0	0	9.277	242	1
10	0	0	1	0	0	2012-01-07 11:29:16	0	0	23	no	0	0	0	17.170	578	1
11	0	0	1	0	0	2012-02-22 16:57:02	0	0	4	no	0	0	0	64.401	1167	1
12	0	0	1	0	0	2012-02-04 15:26:09	0	0	0	no	0	0	2	10.368	198	1
13	1	0	1	0	0	2012-01-24 08:15:56	0	0	3	yes	0	0	0	42.793	712	1
14	1	1	1	0	0	2012-02-08 18:22:46	0	2	2	no	0	0	0	0.451	24	0

Showing 1 to 15 of 50 entries

observation (case)

variable

Describing a Data Set

spam	Indicator for whether the email was spam.
to_multiple	Indicator for whether the email was addressed to more than one recipient.
from	Whether the message was listed as from anyone (this is usually set by default for regular outgoing email).
cc	Indicator for whether anyone was CCed.
sent_email	Indicator for whether the sender had been sent an email in the last 30 days.
time	Time at which email was sent.
image	The number of images attached.
attach	The number of attached files.
dollar	The number of times a dollar sign or the word "dollar" appeared in the email.
winner	Indicates whether "winner" appeared in the email.
inherit	The number of times "inherit" (or an extension, such as "inheritance") appeared in the email.
viagra	The number of times "viagra" appeared in the email.
password	The number of times "password" appeared in the email.
num_char	The number of characters in the email, in thousands.
line_breaks	The number of line breaks in the email (does not count text wrapping).

Types of variables:

- Numeric (= quantitative)
  - Discrete (space between possible values)
  - Continuous (real numbers)
- Categorical (= qualitative)
  - Nominal (with no natural ordering)
  - Ordinal (with a natural ordering)

Describing a Data Set

spam	Indicator for whether the email was spam.
to_multiple	Indicator for whether the email was addressed to more than one recipient.
from	Whether the message was listed as from anyone (this is usually set by default for regular outgoing email).
cc	Indicator for whether anyone was CCed.
sent_email	Indicator for whether the sender had been sent an email in the last 30 days.
time	Time at which email was sent.
image	The number of images attached.
attach	The number of attached files.
dollar	The number of times a dollar sign or the word "dollar" appeared in the email.
winner	Indicates whether "winner" appeared in the email.
inherit	The number of times "inherit" (or an extension, such as "inheritance") appeared in the email.
viagra	The number of times "viagra" appeared in the email.
password	The number of times "password" appeared in the email.
num_char	The number of characters in the email, in thousands.
line_breaks	The number of line breaks in the email (does not count text wrapping).

Types of variables:

- Numeric (= quantitative)
  - Discrete (space between possible values)
  - Continuous (real numbers)
- Categorical (= qualitative)
  - Nominal (with no natural ordering)
  - Ordinal (with a natural ordering)

Describing a Data Set

spam	Indicator for whether the email was spam.
to_multiple	Indicator for whether the email was addressed to more than one recipient.
from	Whether the message was listed as from anyone (this is usually set by default for regular outgoing email).
cc	Indicator for whether anyone was CCed.
sent_email	Indicator for whether the sender had been sent an email in the last 30 days.
time	Time at which email was sent.
image	The number of images attached.
attach	The number of attached files.
dollar	The number of times a dollar sign or the word "dollar" appeared in the email.
winner	Indicates whether "winner" appeared in the email.
inherit	The number of times "inherit" (or an extension, such as "inheritance") appeared in the email.
viagra	The number of times "viagra" appeared in the email.
password	The number of times "password" appeared in the email.
num_char	The number of characters in the email, in thousands.
line_breaks	The number of line breaks in the email (does not count text wrapping).

Types of variables:

- Numeric (= quantitative)
  - Discrete (space between possible values)
  - Continuous (real numbers)
- Categorical (= qualitative)
  - Nominal (with no natural ordering)
  - Ordinal (with a natural ordering)

Describing a Data Set

spam	Indicator for whether the email was spam.
to_multiple	Indicator for whether the email was addressed to more than one recipient.
from	Whether the message was listed as from anyone (this is usually set by default for regular outgoing email).
cc	Indicator for whether anyone was CCed.
sent_email	Indicator for whether the sender had been sent an email in the last 30 days.
time	Time at which email was sent.
image	The number of images attached.
attach	The number of attached files.
dollar	The number of times a dollar sign or the word "dollar" appeared in the email.
winner	Indicates whether "winner" appeared in the email.
inherit	The number of times "inherit" (or an extension, such as "inheritance") appeared in the email.
viagra	The number of times "viagra" appeared in the email.
password	The number of times "password" appeared in the email.
num_char	The number of characters in the email, in thousands.
line_breaks	The number of line breaks in the email (does not count text wrapping).

Types of variables:

- Numeric (= quantitative)
  - Discrete (space between possible values)
  - Continuous (real numbers)
- Categorical (= qualitative)
  - Nominal (with no natural ordering)
  - Ordinal (with a natural ordering)

Be Careful About Data Types

Some variables encoded with numbers are not numeric.

- 0/1 for TRUE/FALSE
- ZIP codes (92093)

Not all numeric variables look like numbers.

- Dates (Friday the 13th, 2017)
- GPS coordinates (40°26’ 46” N 79°58’ 56” W)

Be Careful About Data Types

Some variables encoded with numbers are not numeric.

- 0/1 for TRUE/FALSE
- ZIP codes (92093)

Not all numeric variables look like numbers.

- Dates (Friday the 13th, 2017)
- GPS coordinates (40°26' 46" N 79°58' 56" W)

Data types determine how you analyze them. Tools are specifically suited to on of them.

	1 variable	2 variables
Categorical	Bar chart	Contigency table
Numeric	Histogram	Scatterplot

Various situations come various visualizations

One Categorical Variable

See the "attachment" variable as categorical.

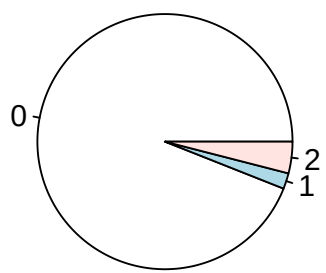
```
> email50$attach
 [1] 0 0 2 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 1 0 0 0 0 0 0 0 0 0 0
       0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> table(email50$attach)
 0  1  2
47  1  2
```

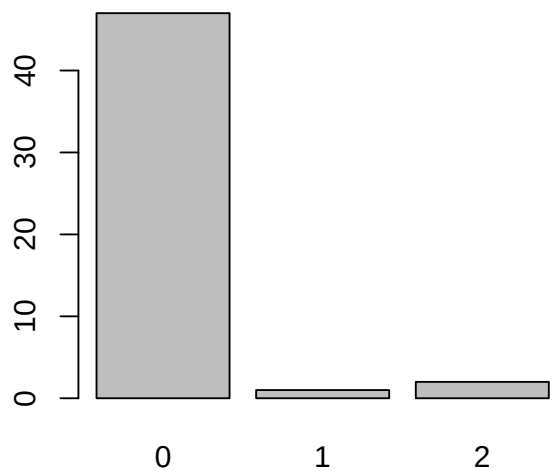
One Categorical Variable

See the "attachment" variable as categorical.

```
> email50$attach
  [1] 0 0 2 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 1 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> table(email50$attach)
 0  1  2
47  1  2
```

```
pie(table(email50$attach))
barplot(table(email50$attach))
```





Two Categorical Variables

Contingency table: correlation between frequencies of two categorical variables.

```
> addmargins(table(email50$attach, email50$to_multiple))
          0  1 Sum
  0      41  6  47
  1       1  0   1
  2       1  1   2
  Sum    43  7  50
```

- Rows: $\{0, 1, 2\}$ are the number of attachment
- Columns: $\{0, 1\}$ $(= \{No, Yes\})$ indicates if the email was sent to several addresses.

Two Categorical Variables

Contingency table: correlation between frequencies of two categorical variables.

```
> addmargins(table(email50$attach,email50$to_multiple))
          0  1 Sum
  0      41  6  47
  1       1  0   1
  2       1  1   2
  Sum    43  7  50
```

- Rows: $\{0, 1, 2\}$ are the number of attachment
- Columns: $\{0, 1\}$ $(= \{No, Yes\})$ indicates if the email was sent to several addresses.

Use contingency tables to apprehend 2 categorical variables:

$$\%\{\ 1 \text{ - sended email with attachment}\} = \frac{1+1}{43} \simeq 4.65\%.$$

$$\%\{\ 1 > \text{ - sended email with attachment}\} = \frac{0+1}{7} \simeq 14.28\%.$$

Two Categorical Variables

Contingency table: correlation between frequencies of two categorical variables.

```
> addmargins(table(email50$attach,email50$to_multiple))
          0  1 Sum
  0      41  6  47
  1       1  0   1
  2       1  1   2
  Sum    43  7  50
```

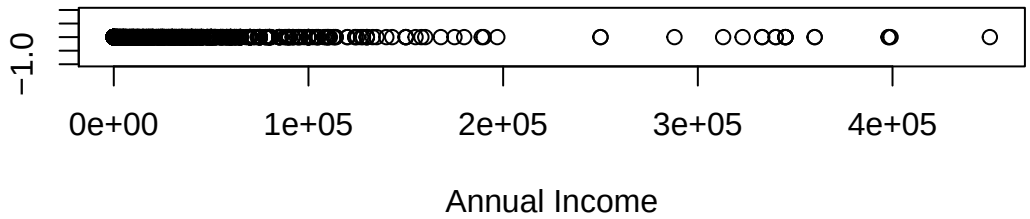
- Rows: $\{0, 1, 2\}$ are the number of attachment
- Columns: $\{0, 1\}$ $(= \{No, Yes\})$ indicates if the email was sent to several addresses.

Use contingency tables to apprehend 2 categorical variables:

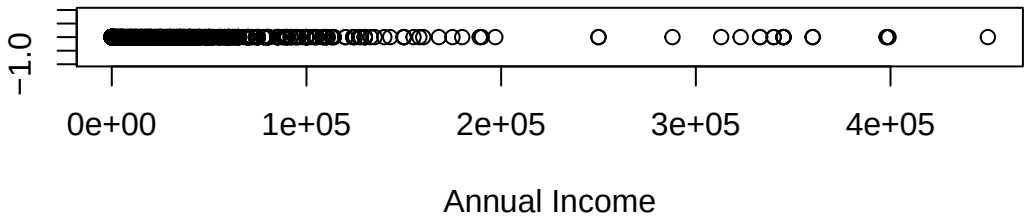
$$\%\{\ 1 \text{ - sended email with attachment}\} = \frac{1+1}{43} \simeq 4.65\%.$$
$$\%\{\ 1 > \text{ - sended email with attachment}\} = \frac{0+1}{7} \simeq 14.28\%.$$

Be careful with sample size! (We only have **one** email user.)

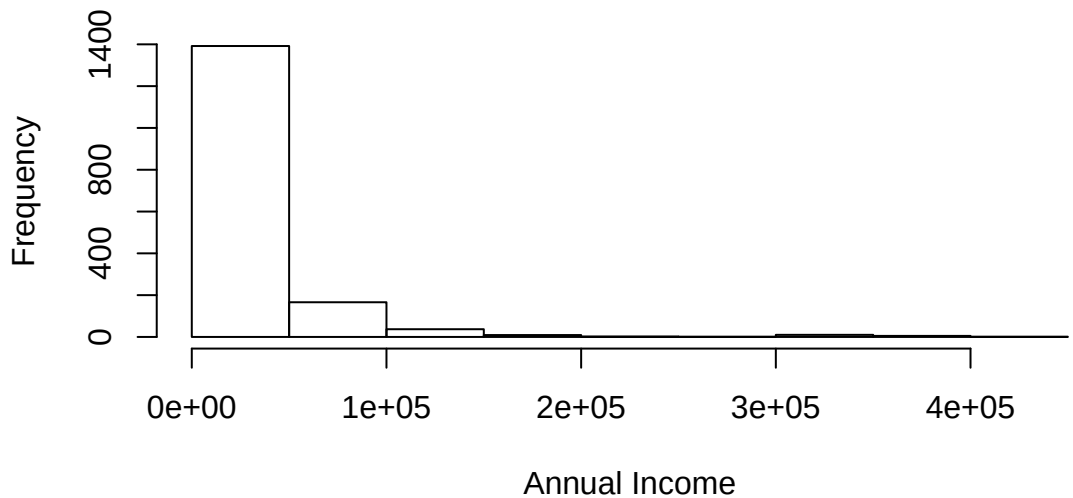
Numerical Data: Histograms



Numerical Data: Histograms

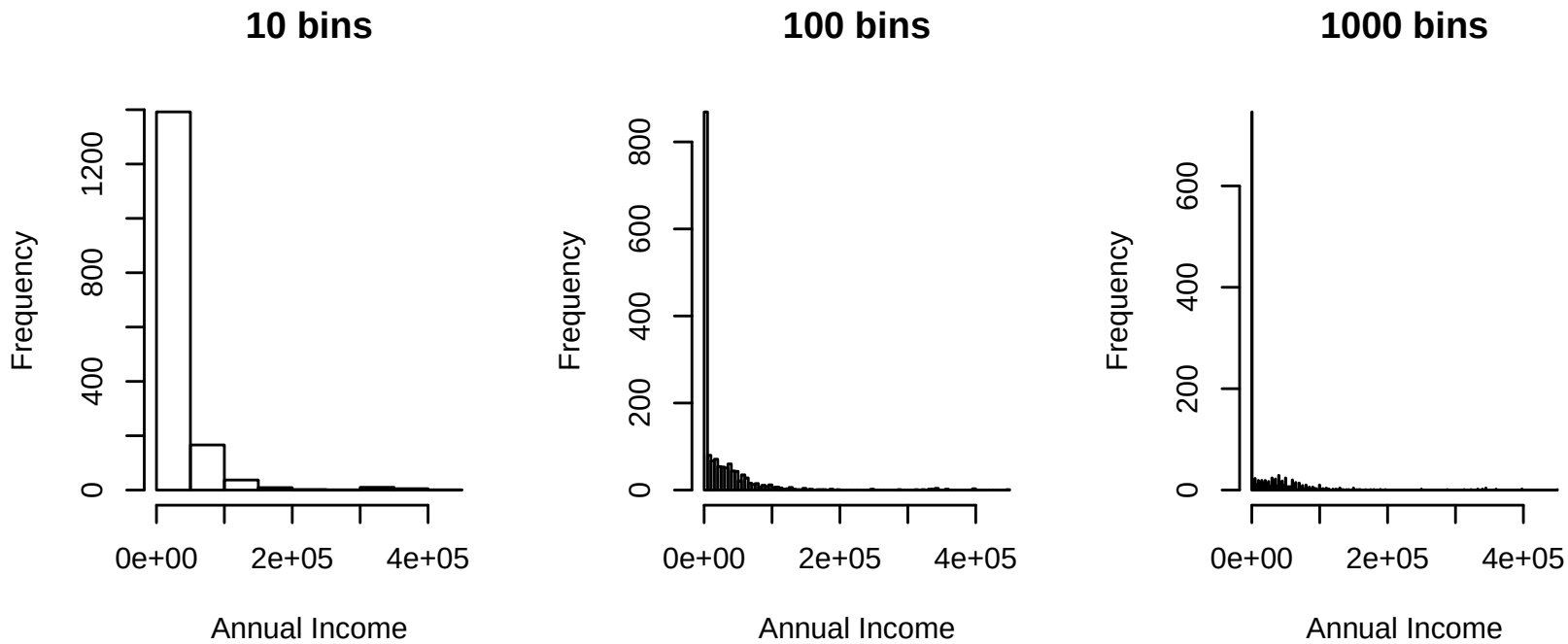


Group points into bins to get an **histogram** (R function `hist`).



Numerical Data: Histograms

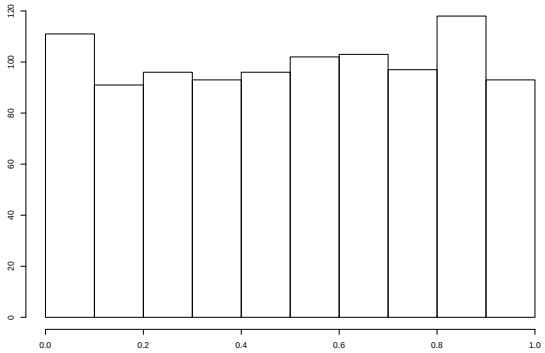
The choice of bin size influences crudely the histogram plot.



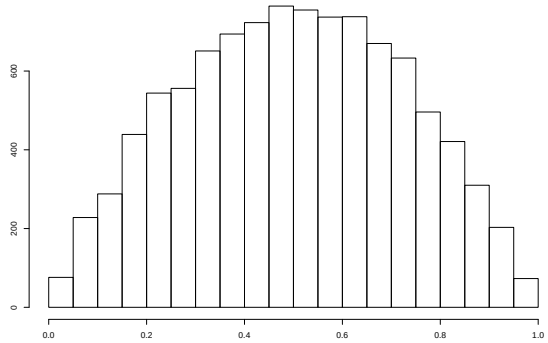
By default, R tries its best to display an informative histogram.

Numerical Data Vocabulary: Modes

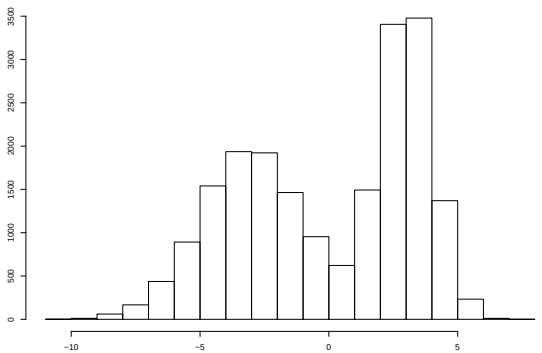
We say an histogram has a **mode** when it is peaked somewhere.



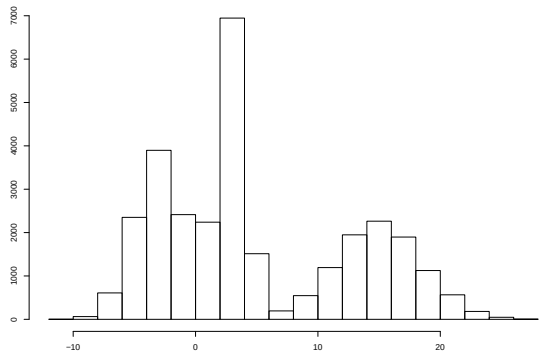
(a) No mode



(b) Unimodal (1 mode)



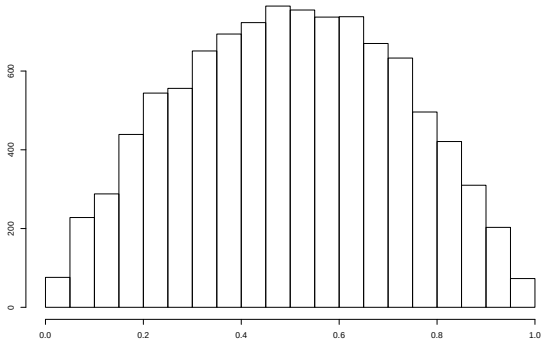
(c) Bimodal (2 modes)



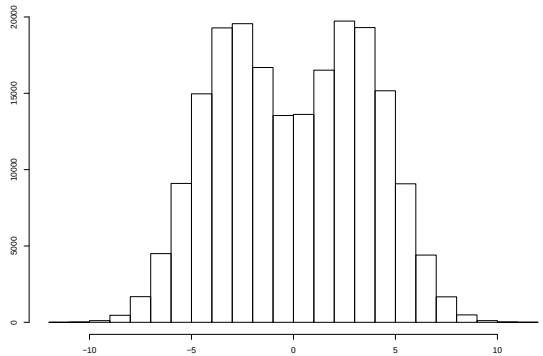
(d) Multimodal (> 2)

Numerical Data Vocabulary: Symmetry

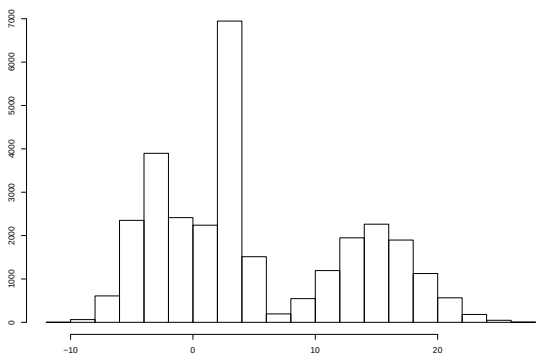
An histogram is **symmetric** if both sides of mode look the same.



(a) Symmetric



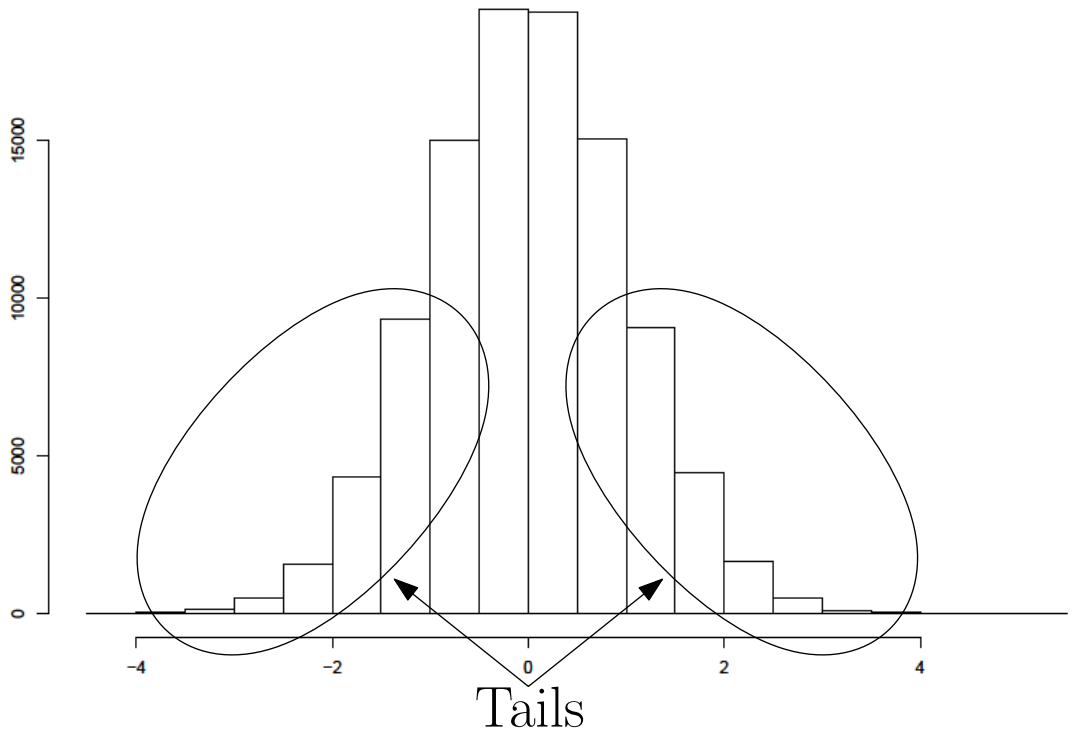
(b) Symmetric



(c) Not symmetric

Numerical Data Vocabulary: Tails

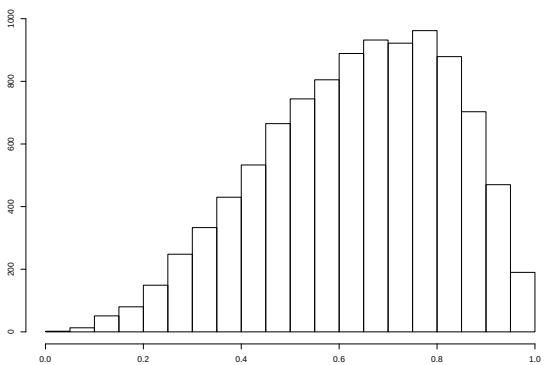
The **tails** of an histogram are the parts away from the center.



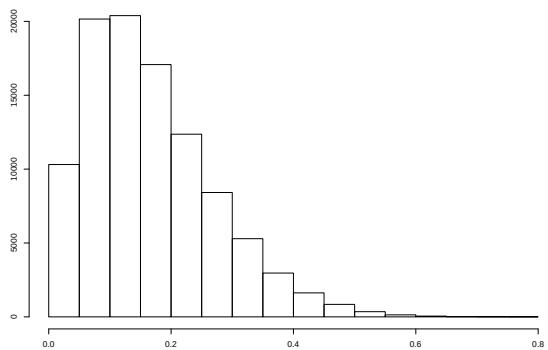
Numerical Data Vocabulary: Skewness

When an histogram is not symmetric, we can describe further its asymmetry by saying it is

- **Skewed left:** if the left tail is longer than the right tail.
- **Skewed right:** if the right tail is longer than the left tail.



(a) Skewed left

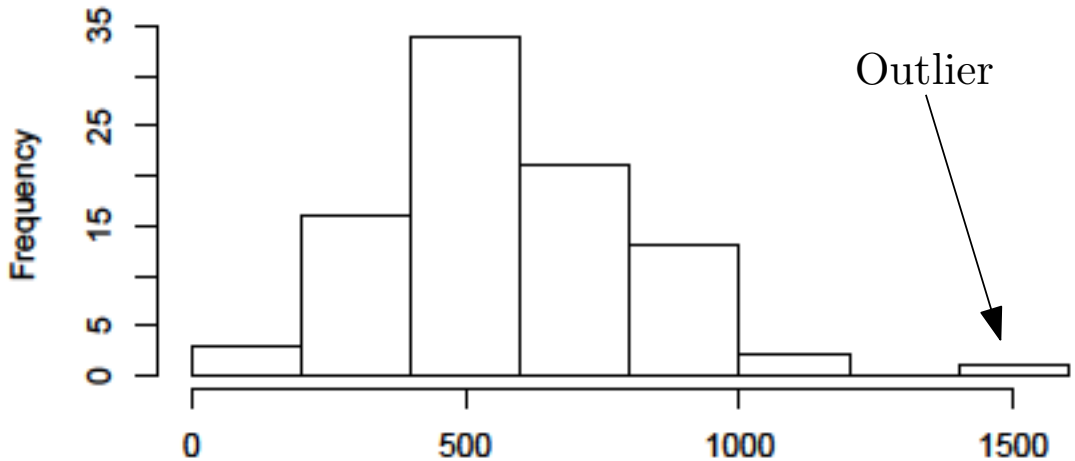


(b) Skewed right

Skewed left/right = the left/right tail stretches out longer.

Numerical Data Vocabulary: Outlier

An **outlier** is an observation that appears extreme relative to the rest of the data. (= Not conventional)



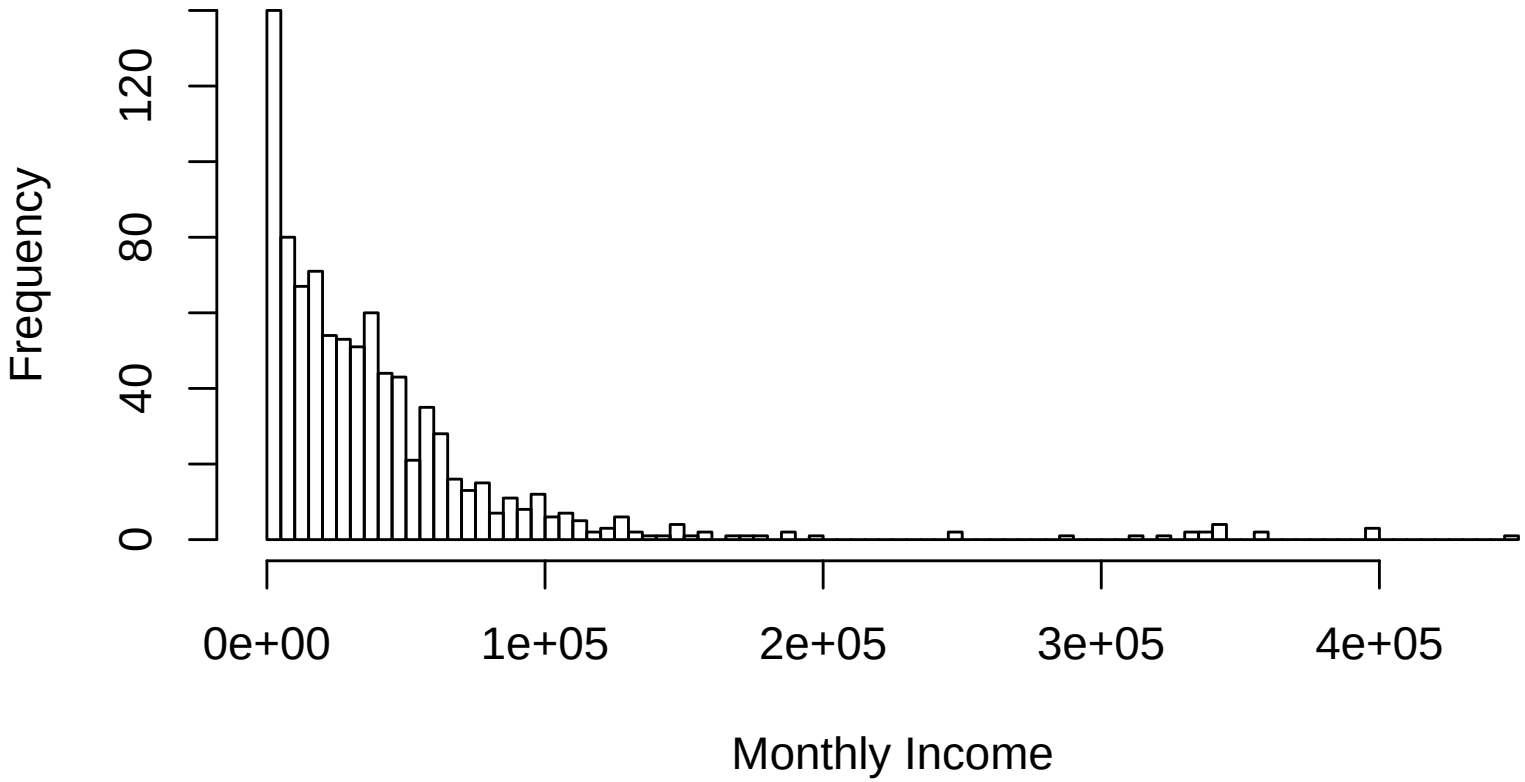
Examples:

- Extreme values in precision measurements for astrophysics
- Trolls' answers in online questionnaires

Sometimes outliers are informative, sometimes just annoying.

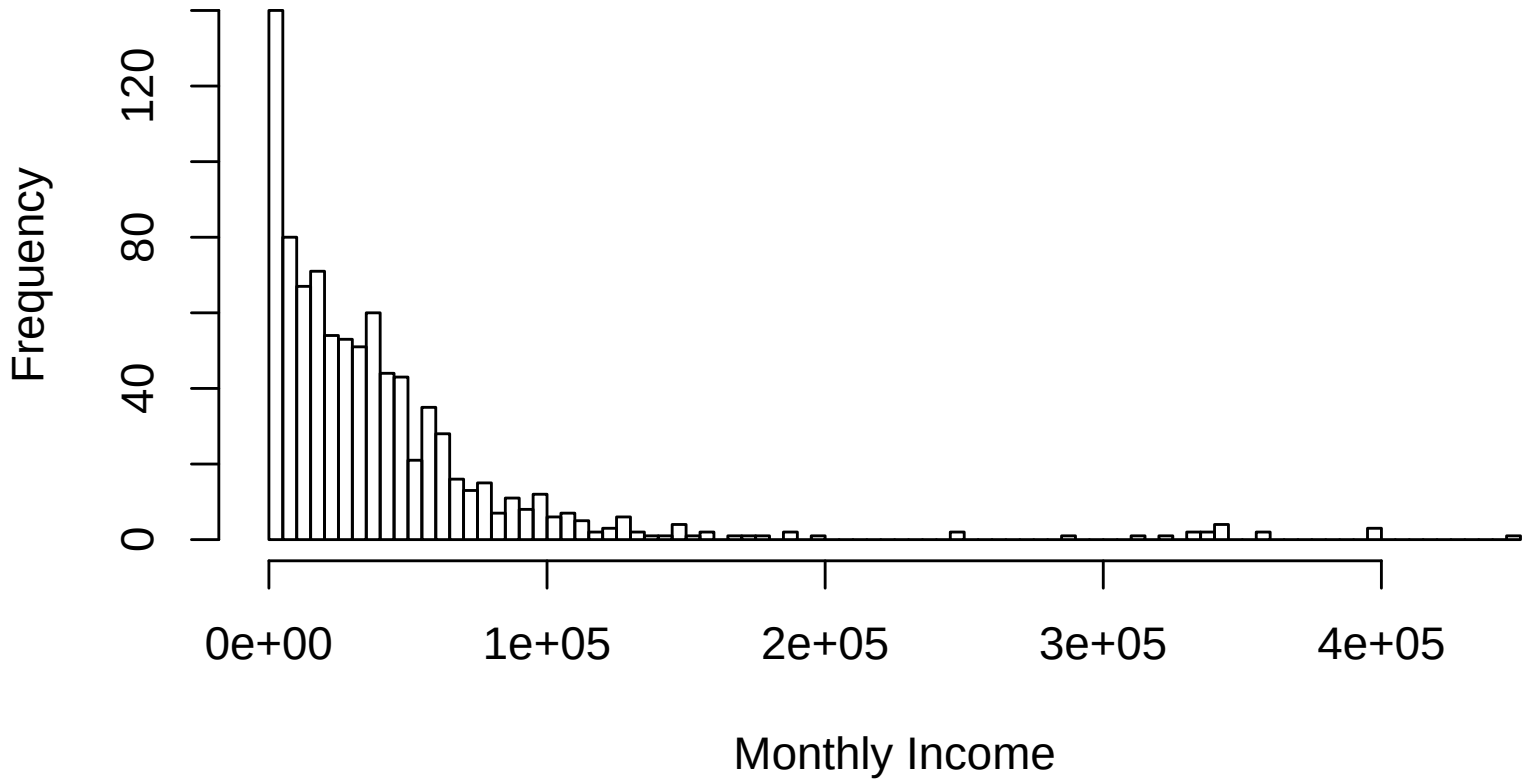
Practice

Describe this histogram.



Practice

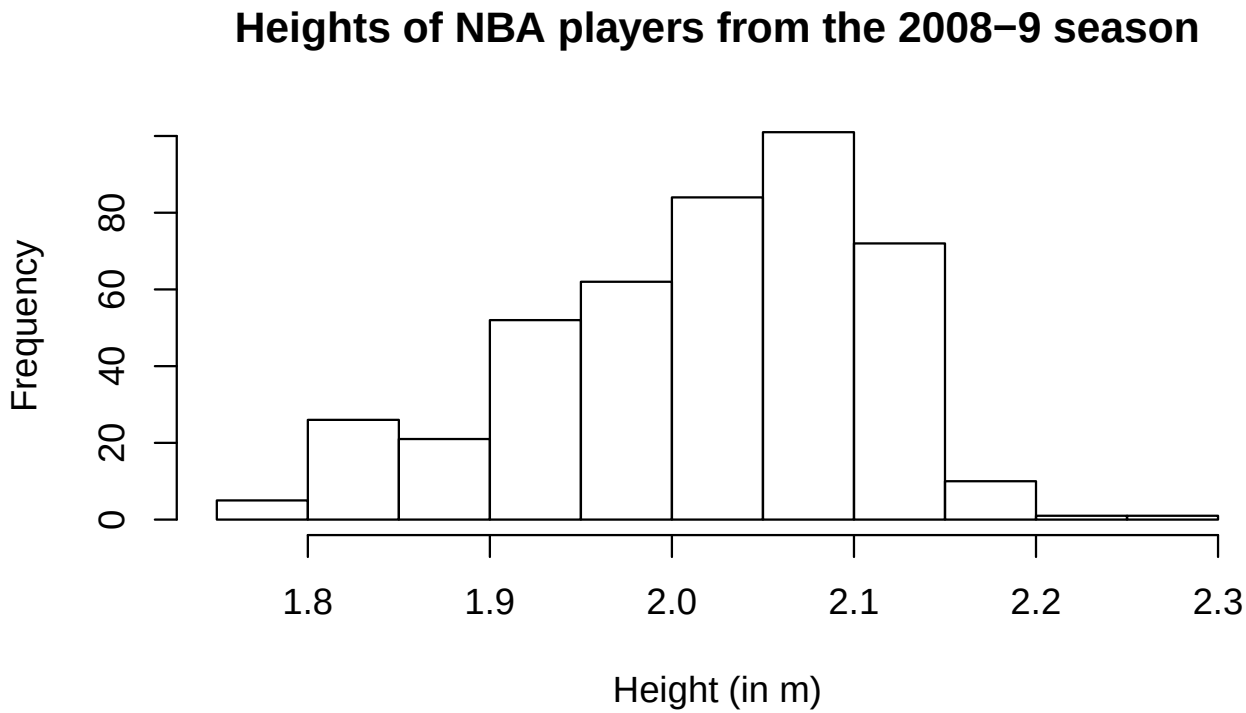
Describe this histogram.



Unimodal, (strongly) skewed right, some outliers.

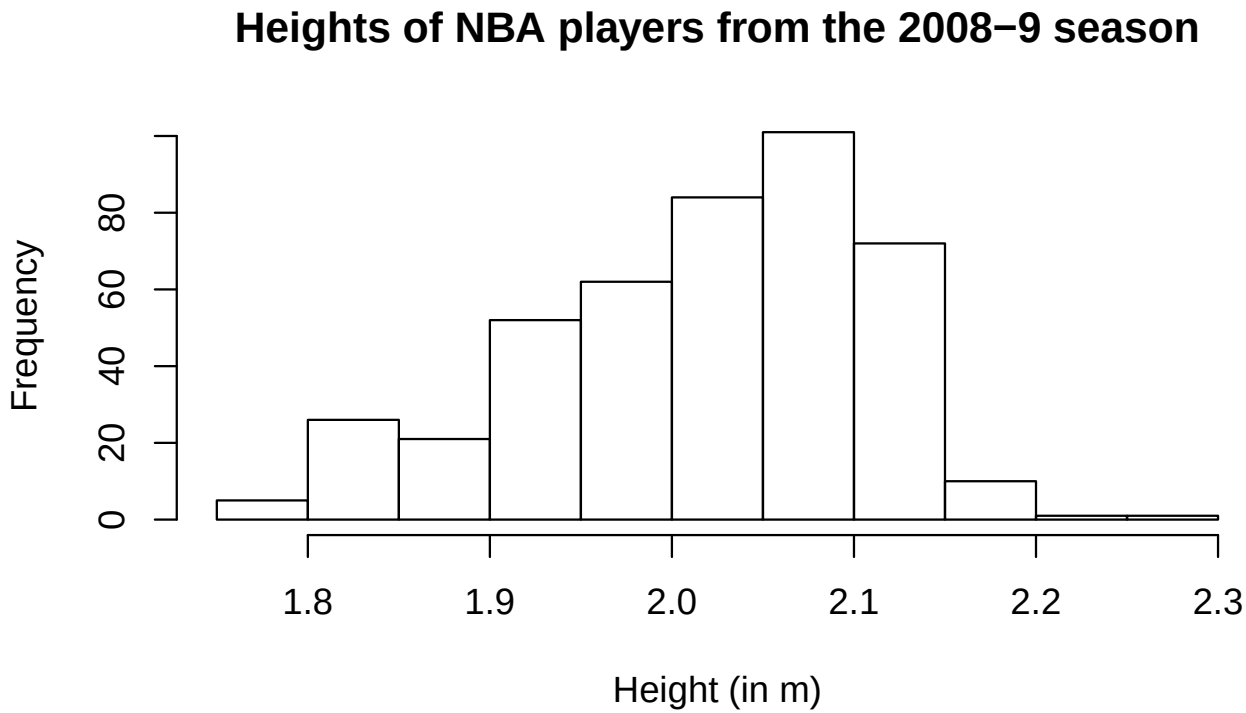
Practice

Describe this histogram.



Practice

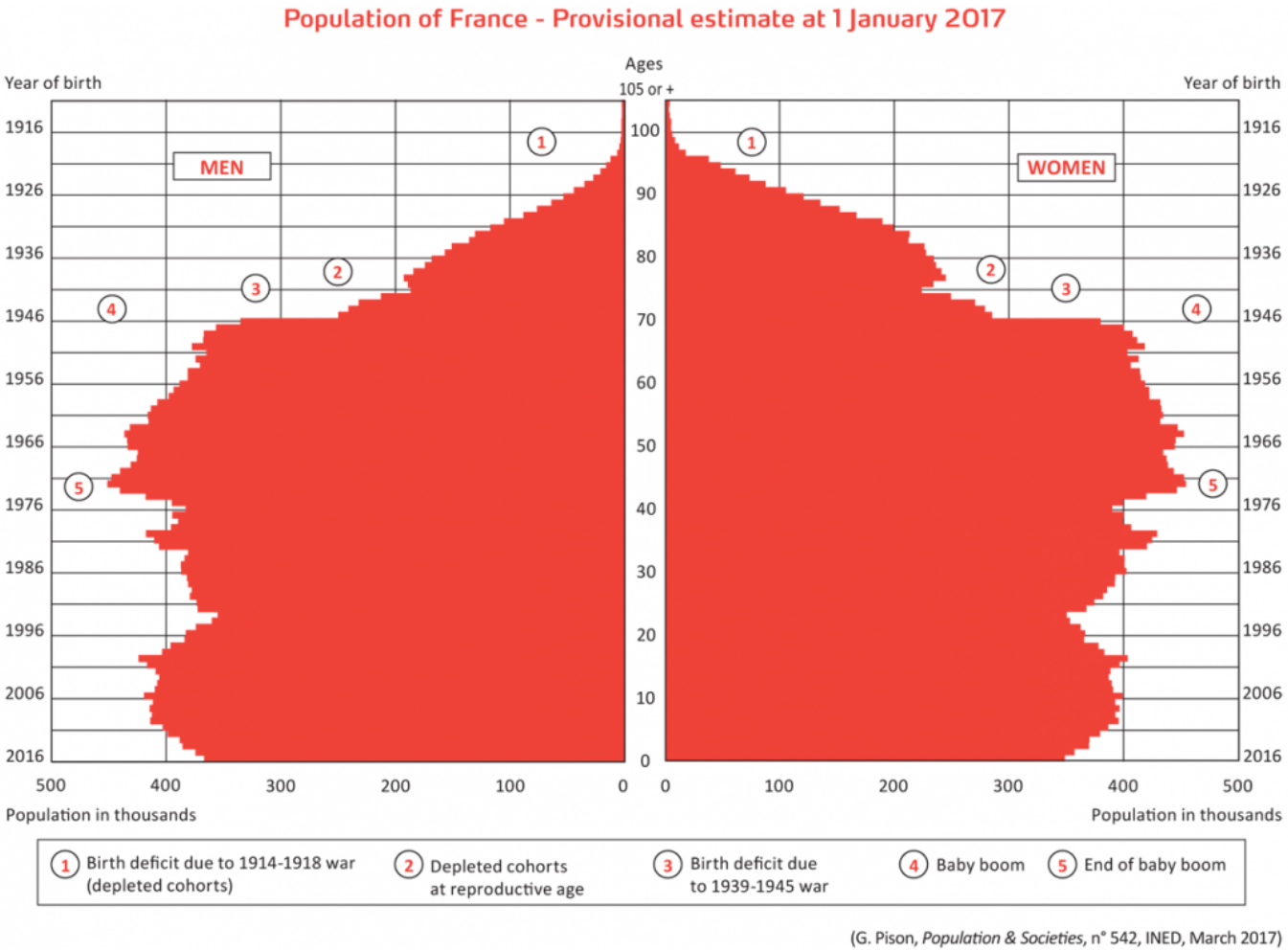
Describe this histogram.



Unimodal, skewed left, no outlier.

Practice

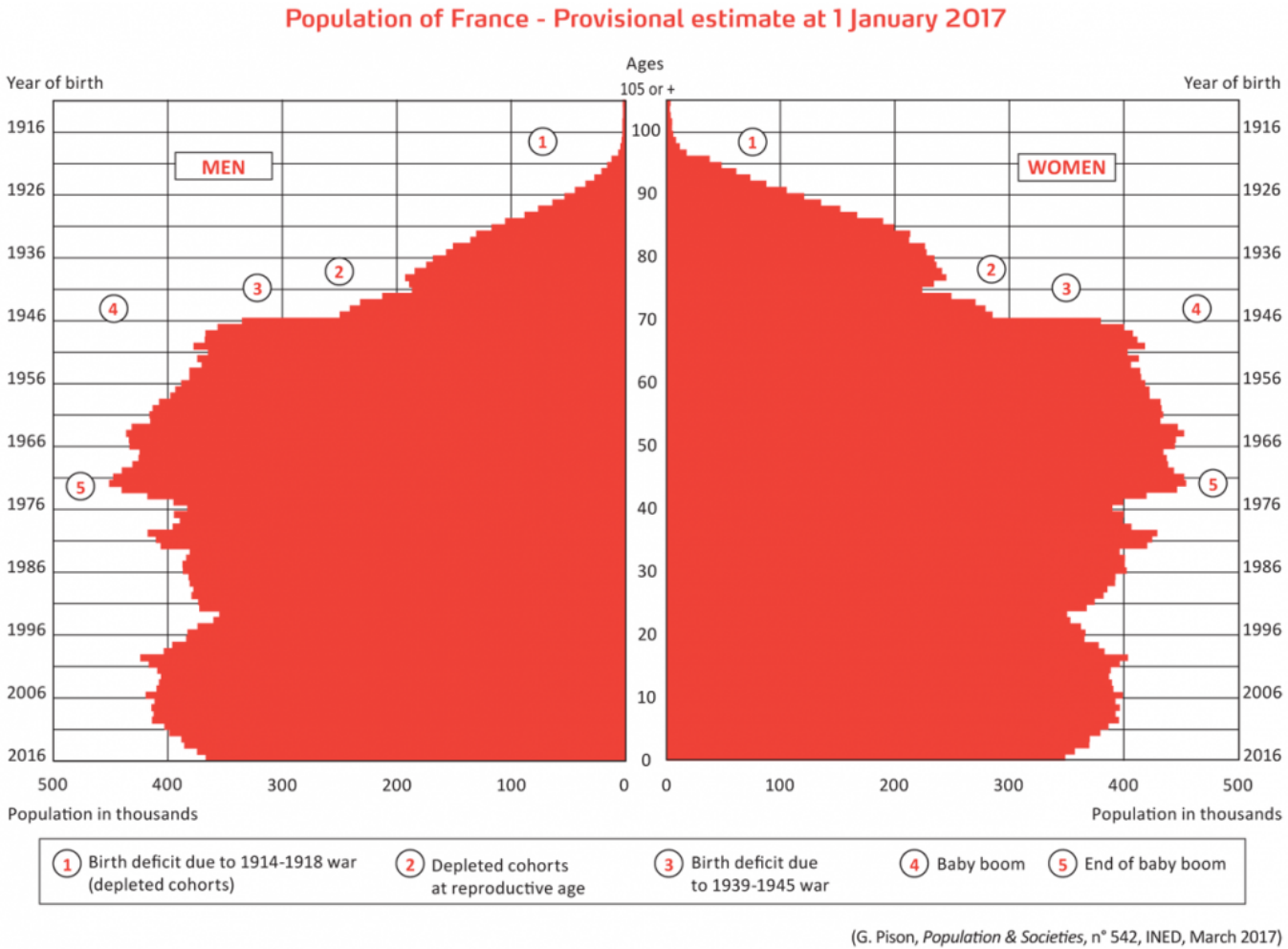
Describe these histograms.



(G. Pison, *Population & Societies*, n° 542, INED, March 2017)

Practice

Describe these histograms.



Multimodal, skewed right (up), no outlier.

What you Should Do After the Lecture

- Read Chapter 1
- Start Homework 1!
  Turn in the following exercises of Chapter 1 in the Textbook:
  - 1.6 (Stealers, study components)
  - 1.14 (Cats on Youtube)
  - 1.38 (Mammal life span)
  - 1.50 (Mix-and-match)
  - 1.58 (Exam scores)
  - 1.66 (Views on immigration)

  Due date is Friday 6th, October on lecture.

  Out of the 6 exercises here, 4 will be randomly chosen to be graded.
  Make sure you have written down your full name, the PID and the TA's name of your section.