

# **STA 291**

## **Lecture 5 Chap 4**

- **Graphical and Tabular Techniques for categorical data**
- **Graphical Techniques for numerical data**

# Review: Stratified Sampling

- Suppose the population can be divided into non-overlapping groups (“***strata***”) according to some criterion.

Example: All voters divided into male voters and female voters.

- Select a Simple Random Sample independently from each group.

- how it is different from SRS?
- (SRS) = any possible selection equally likely
- Any selection got discriminated/eliminated here in stratified sampling?

# Examples of Stratified Sampling

- The population is divided into male/female sub-populations (Two strata). Within each sub-population do an SRS.
- The population is divided into [Whites, Blacks, Hispanics, Asians, Others.] Five strata. Within each, do a SRS.  
Smaller groups may be **over-sampled**: For example: select from each group a SRS of same size  $n=500$ .

# How could stratification be useful?

- We may want to draw inference about population parameters for each subgroup
- When done right, estimators from stratified random samples are **more precise** than those from Simple Random Samples

# Important Sampling Plans: SRS and variations

- **Simple Random Sampling (SRS)**
  - Each possible sample has the same probability of being selected.
- **Stratified Random Sampling**
  - The population can be divided into a set of non-overlapping subgroups (the strata)
  - SRSs are drawn from each strata
- **Systematic Sampling (eg. Digital music)**

# Sampling Error

- Assume you take a SRS of 100 UK students and ask them about their political affiliation (Democrat, Republican, Independent)
- Now take another SRS of 100 UK students
- Will you get the same percentages?

- No, because of sampling variability.
- Also, the result will not be exactly the same as the population percentage, unless you take a “sample” consisting of the whole population of 30,000 students (this would be called a “census”)  
or if you are very lucky



# Sampling Error

- **Sampling Error** is the error that occurs when a *statistic* based on a sample estimates or predicts the value of a population *parameter*.
- In SRS, stratified RS, the sampling error can usually be quantified.
- In other sampling plans, there is also sampling variability, but its extent is not predictable.

# Nonsampling Error

- bias due to question wording, question order,
- nonresponse (people refuse to answer),

# Chapter 4 Display and Describe Categorical Data

- Summarize data using *graphs, tables, and numbers*.
- Condense the information from the dataset
- Bar chart, Pie chart, scatter plot

# Bar Graph

- features:
  - The bars are usually separated to emphasize that the variable is categorical rather than quantitative
  - For nominal variables (no natural ordering), order the bars by frequency, except possibly for a category “other” that is always last

# Pie Chart

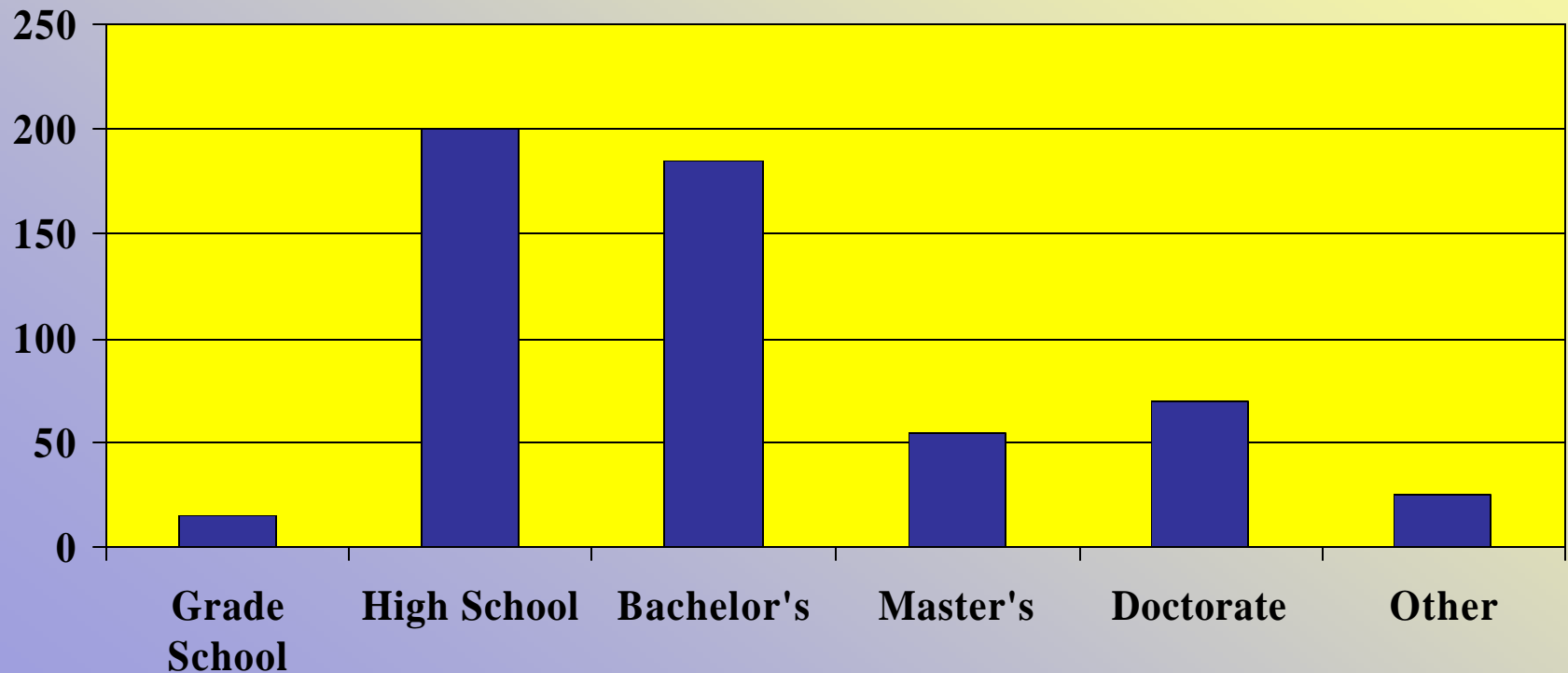
## (Nominal/Ordinal Data)

First Step: Create a Frequency Distribution

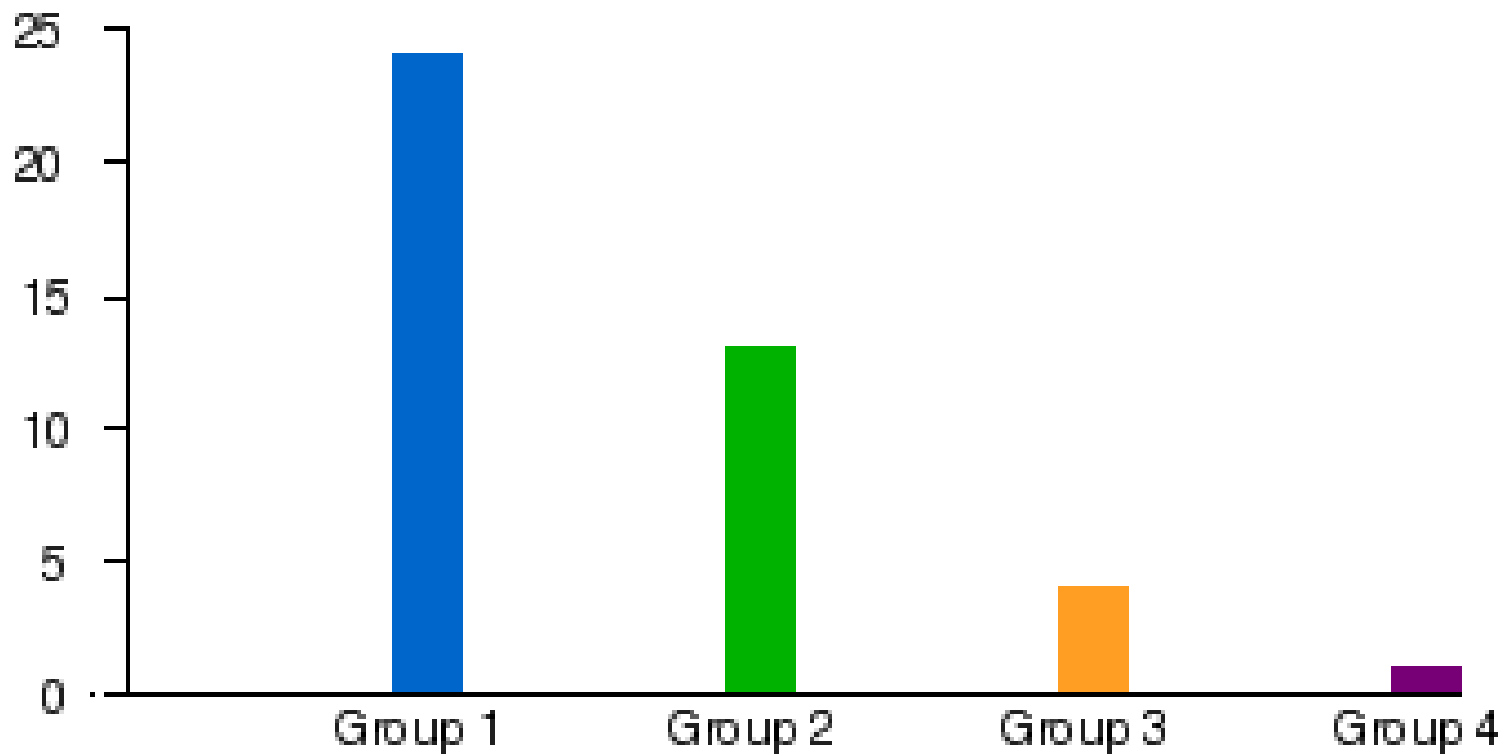
Highest Degree	Frequency (Number of Employees)	Relative Frequency
Grade School	15	
High School	200	
Bachelor's	185	
Master's	55	
Doctorate	70	
Other	25	
Total	550	

# We could display this data in a bar chart...

- **Bar Graph:** *If the data is ordinal, classes are presented in the natural ordering.*



- [http://en.wikipedia.org/wiki/Bar\\_chart](http://en.wikipedia.org/wiki/Bar_chart)



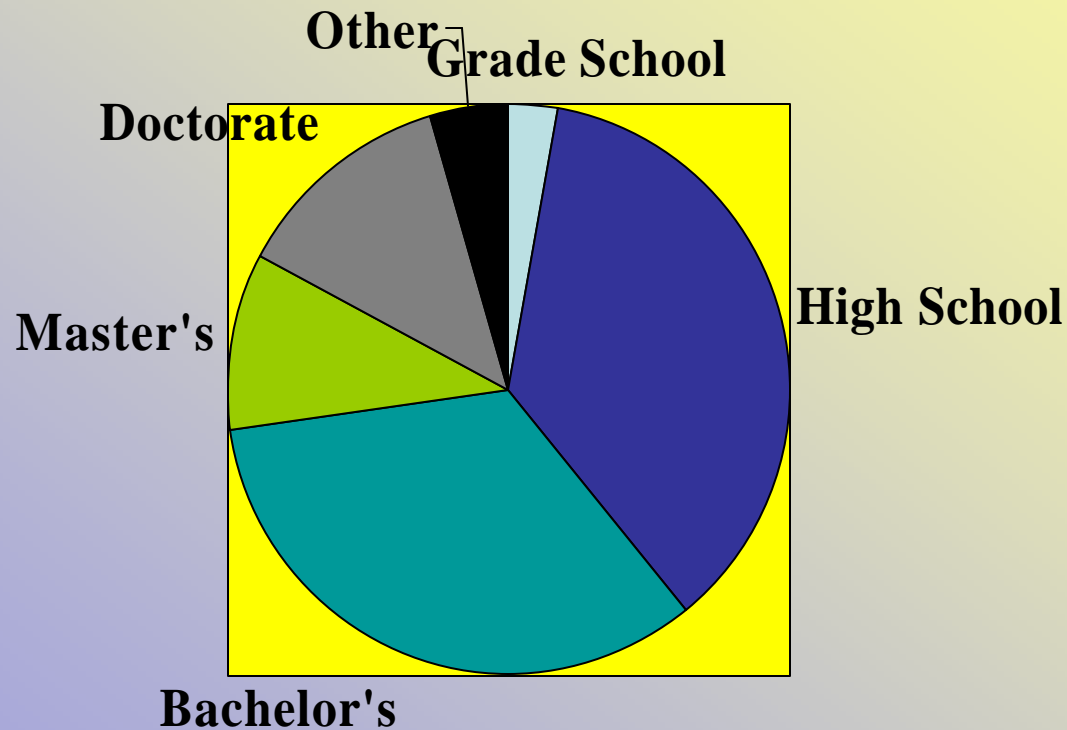
# Pie Chart

- Pie Chart: Pie is divided into slices; The area of each slice is proportional to the frequency of each class.

Highest Degree	Relative Frequency	Angle ( = Rel. Freq. x 360°)
Grade School	$15/550 = .027$	9.72
High School	$200/550 = .364$	131.04
Bachelor's	$185/550 = .336$	120.96
Master's	$55/550 = .1$	36.0
Doctorate	$70/550 = .127$	45.72
Other	$25/550 = .045$	16.2



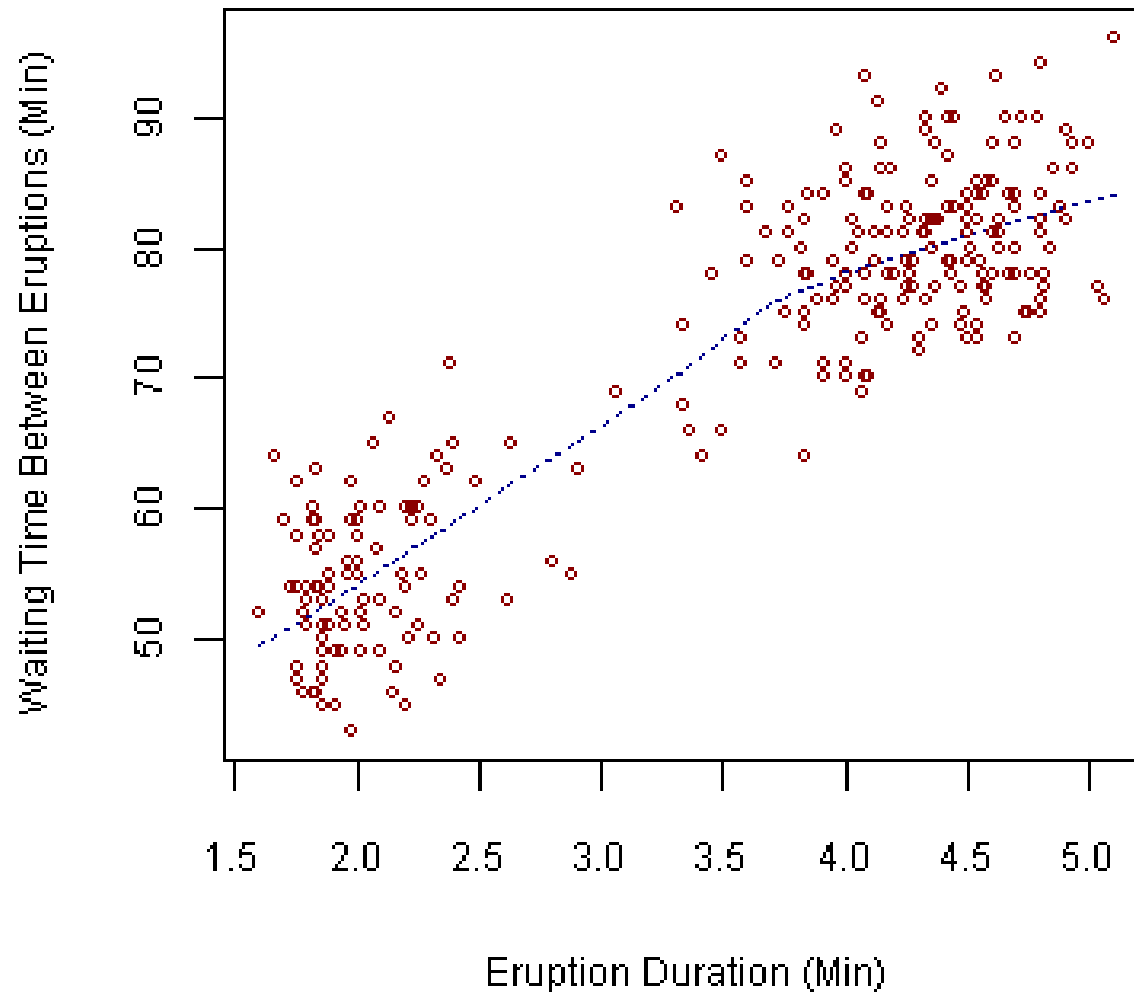
# Pie Chart for Highest Degree Achieved



# Scatter plot

- Plots with two variables  
(reveal the relationship between the two variables)

# Old Faithful Eruptions

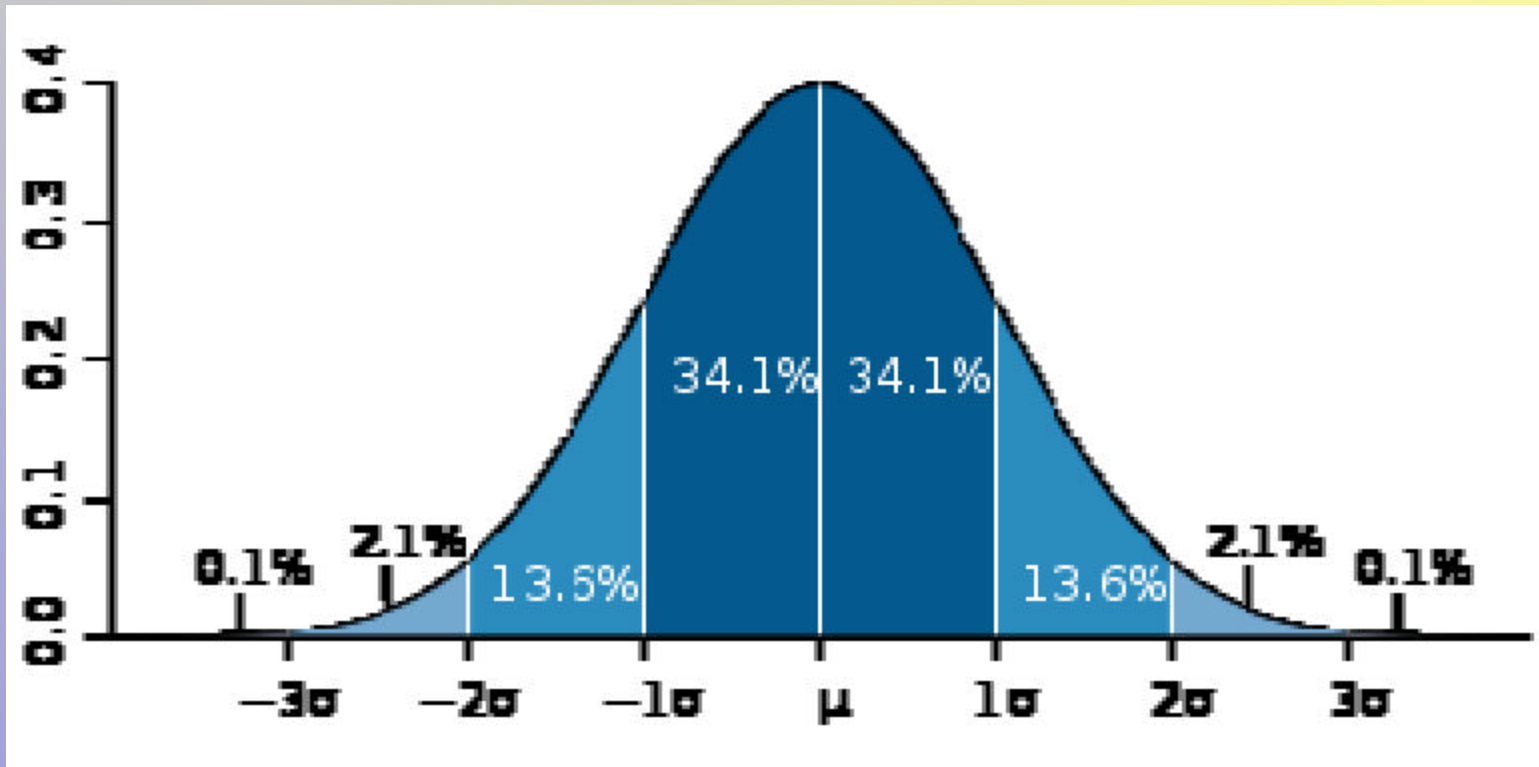


- Dynamic graph: graph change over time – movie or animation.
- Try watch more of those movies at <http://www.gapminder.org>

# Distribution of a (continuous, numerical) variable

- Histogram
- Smoothed histogram – distribution

# A distribution



# Frequency Tables

- Suppose the variable can only take one of 5 possible values.
- We can condense a large sample ( $n=2000$ ) to

value	1	2	3	4	5
frequency	365	471	968	134	62

# Contingency tables

- More complicated tables
- by rows and columns (cross tabulation)



# Homework 2

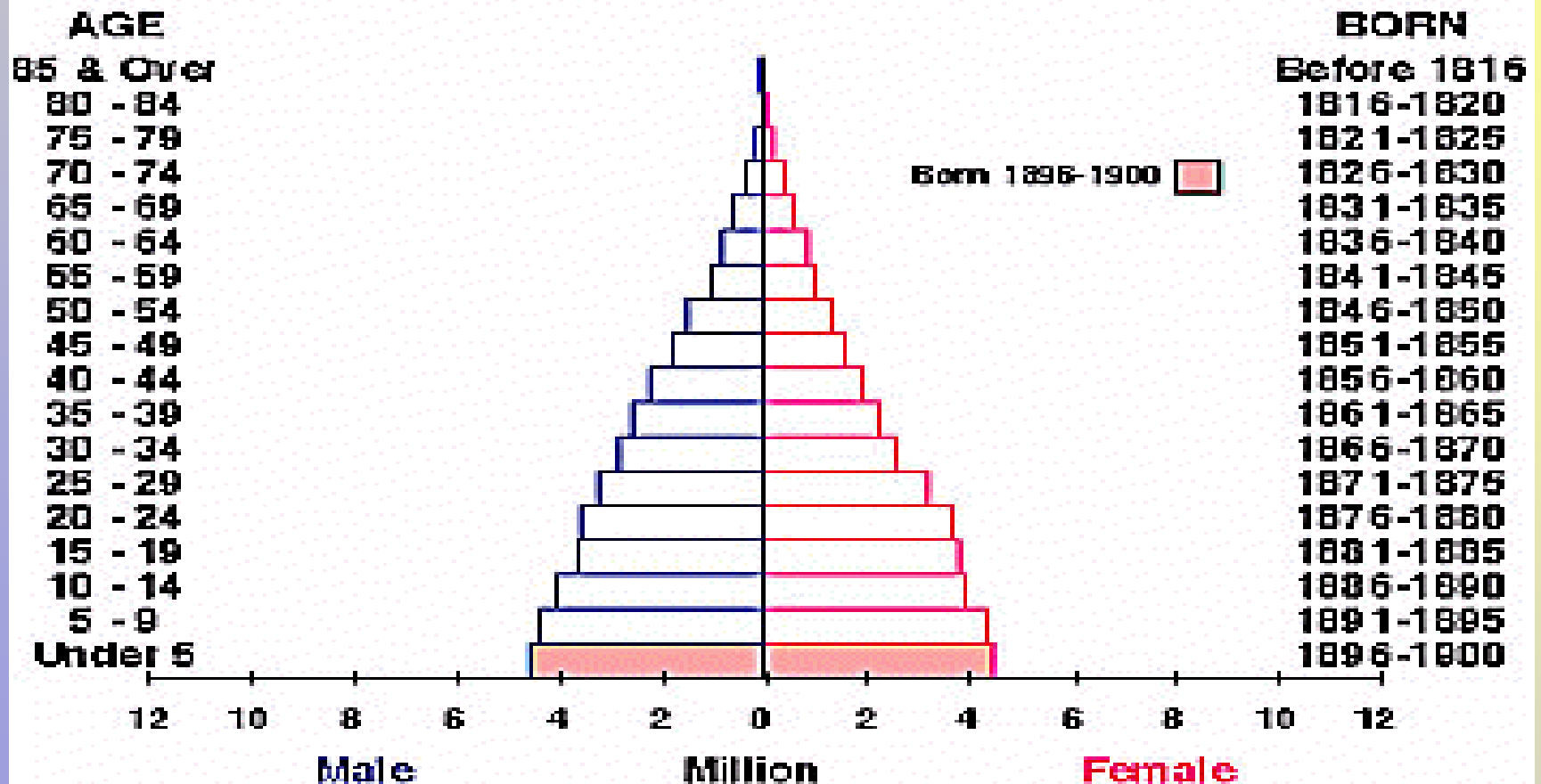
- Due Tuesday next week (Feb 5, 11 PM).
- Online homework assignment.

# Attendance Survey Question 5

- On a 4”x6” index card (or little piece of paper)
  - Please write down your *name* and *section number*.
  - Today’s Question: What is “SRS” stands for in statistical observational study?

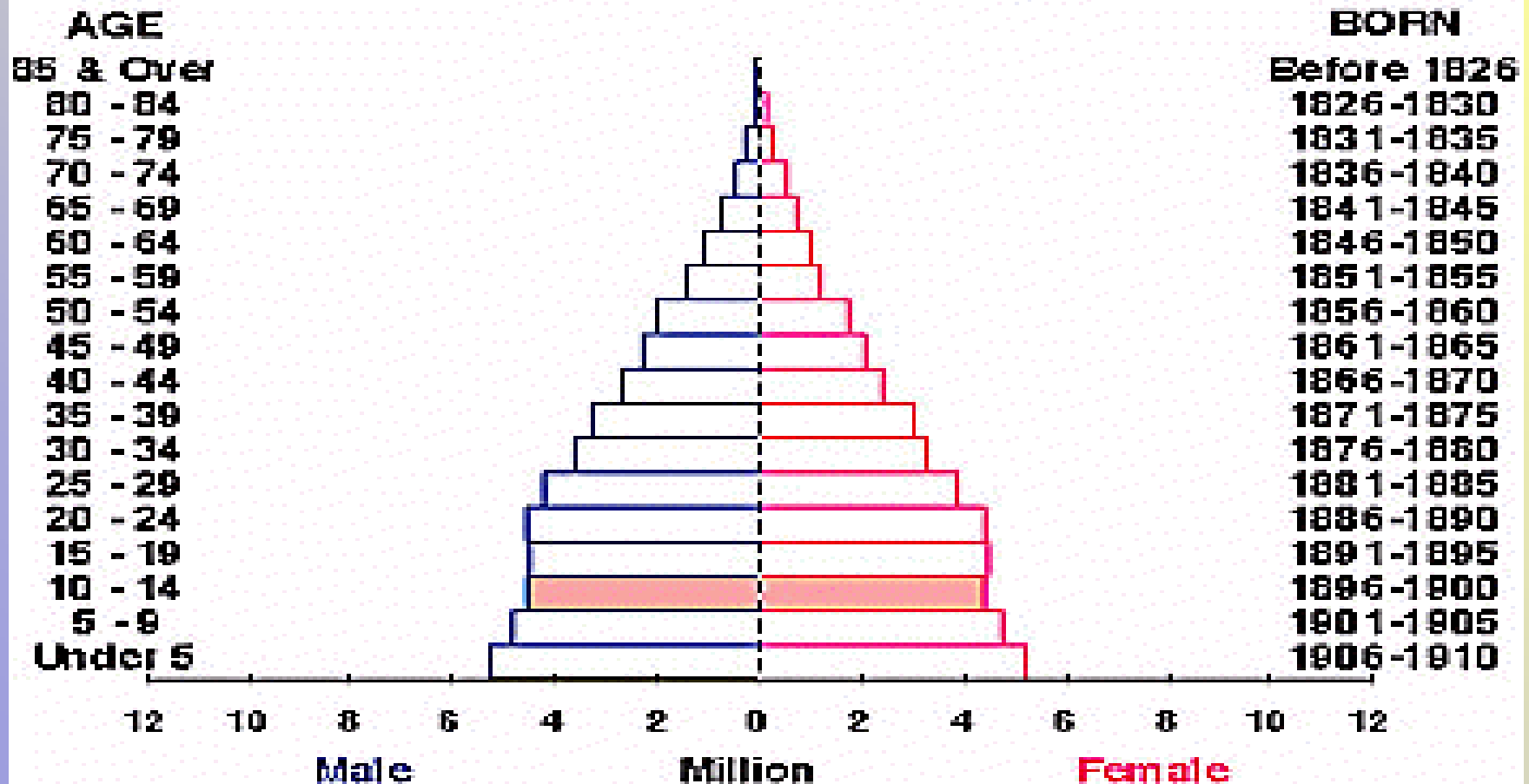
# Histogram of Numbers of Males vs. Females

# 1900

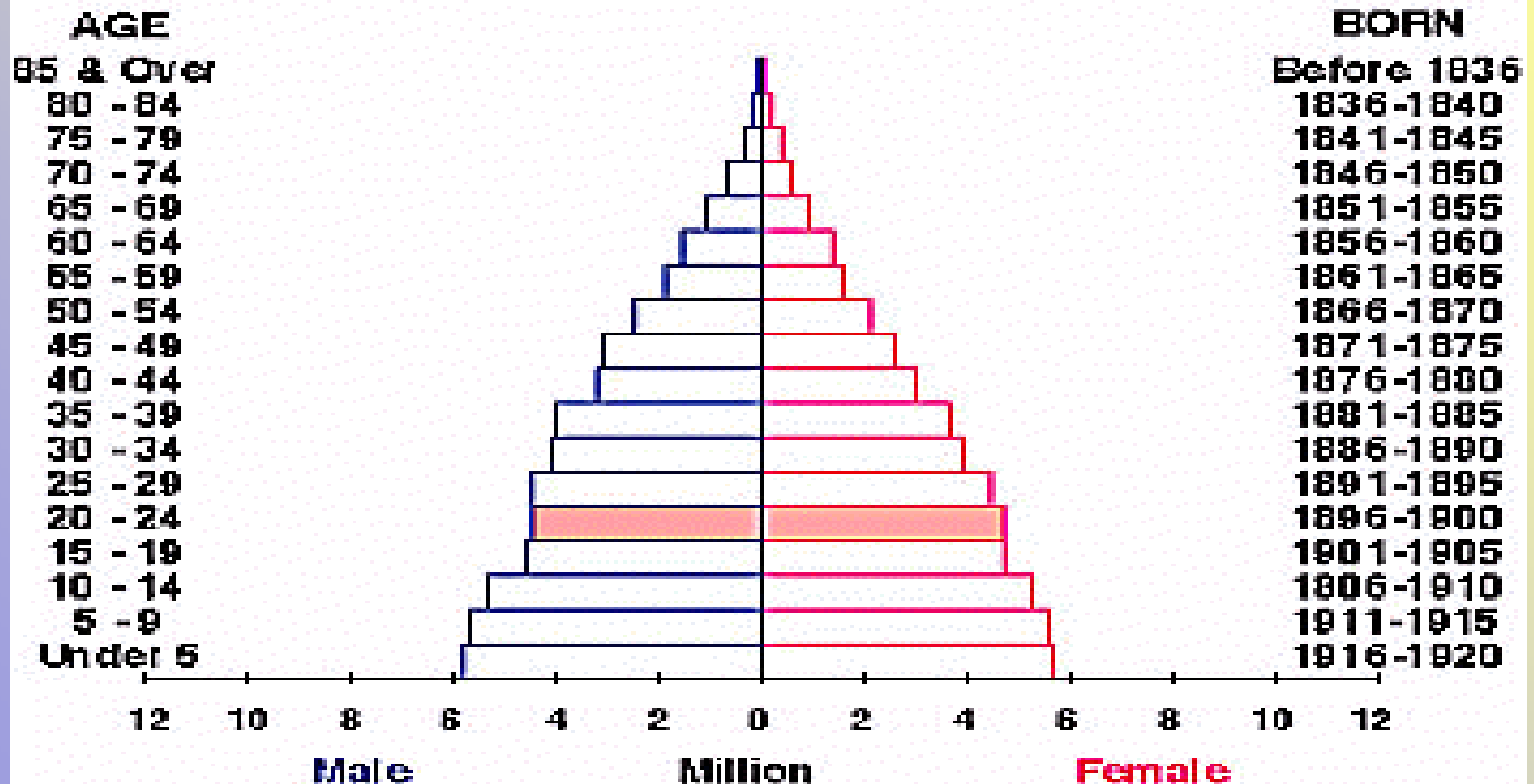


# Histogram of Numbers of Males vs Females

# 1910

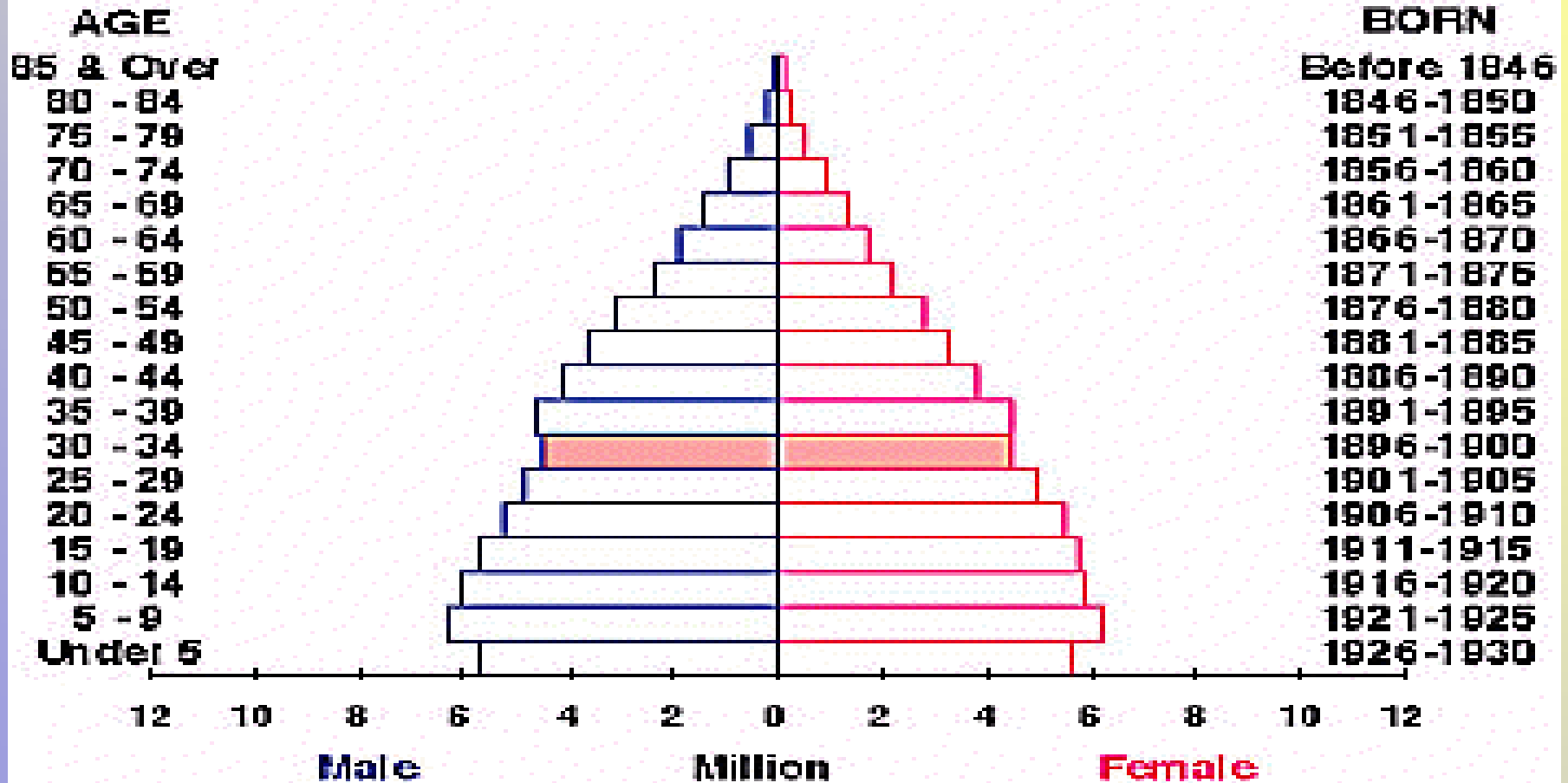


# Histogram of Numbers of Males vs. Females 1920



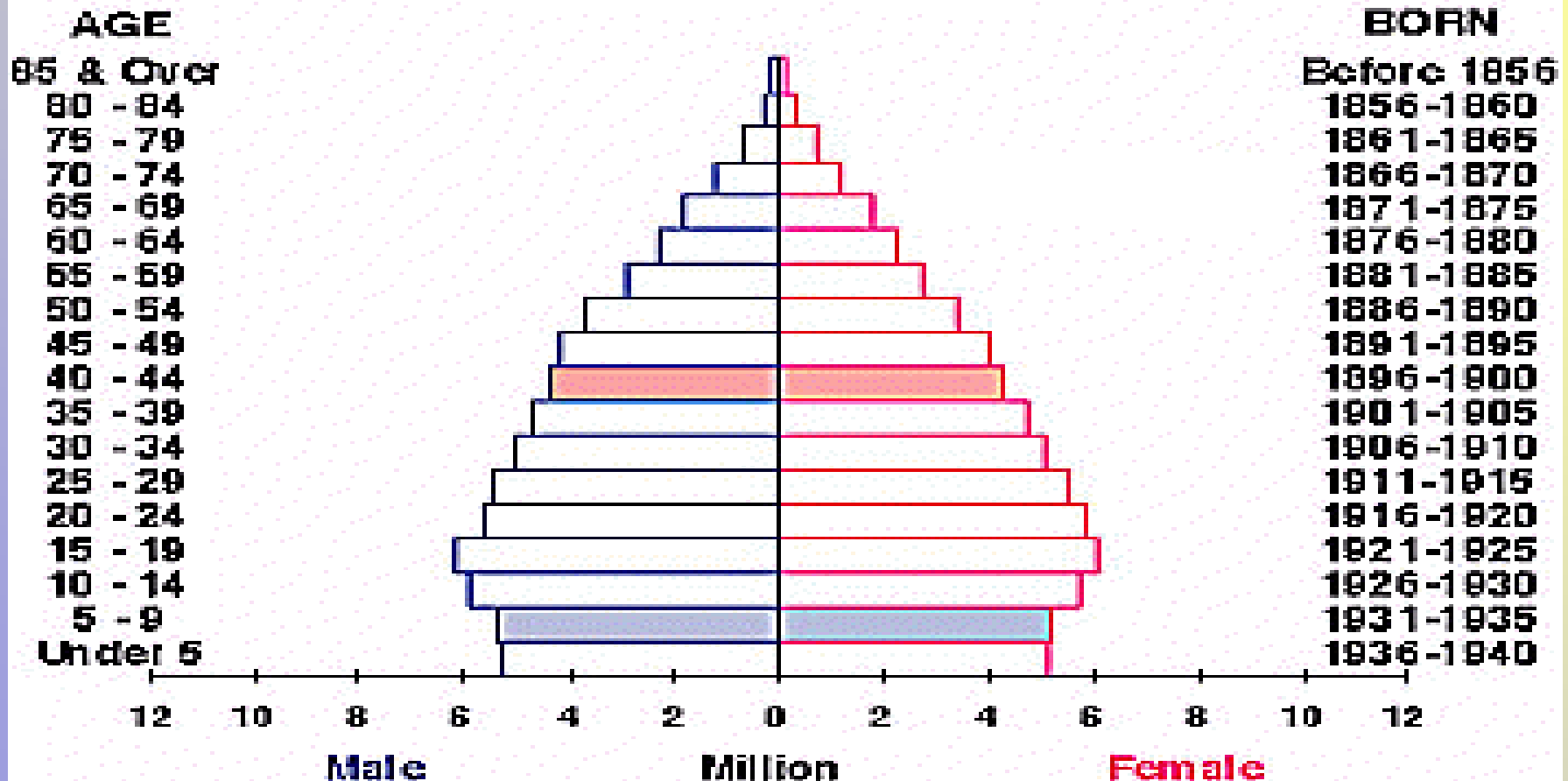
# Histogram of Numbers of Males vs. Females

# 1930



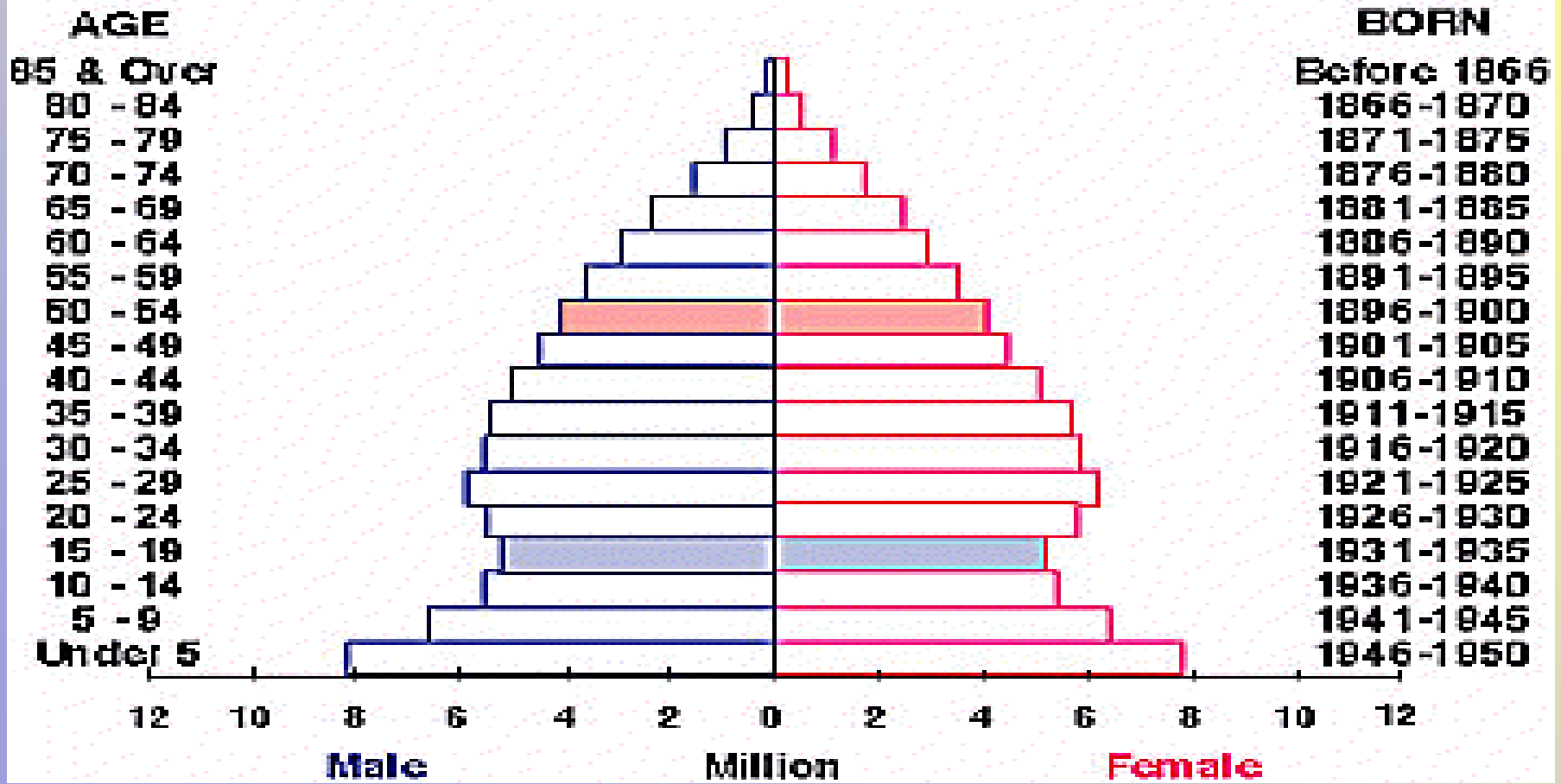
# Histogram of Numbers of Males vs. Females

# 1940



# Histogram of Numbers of Males vs. Females

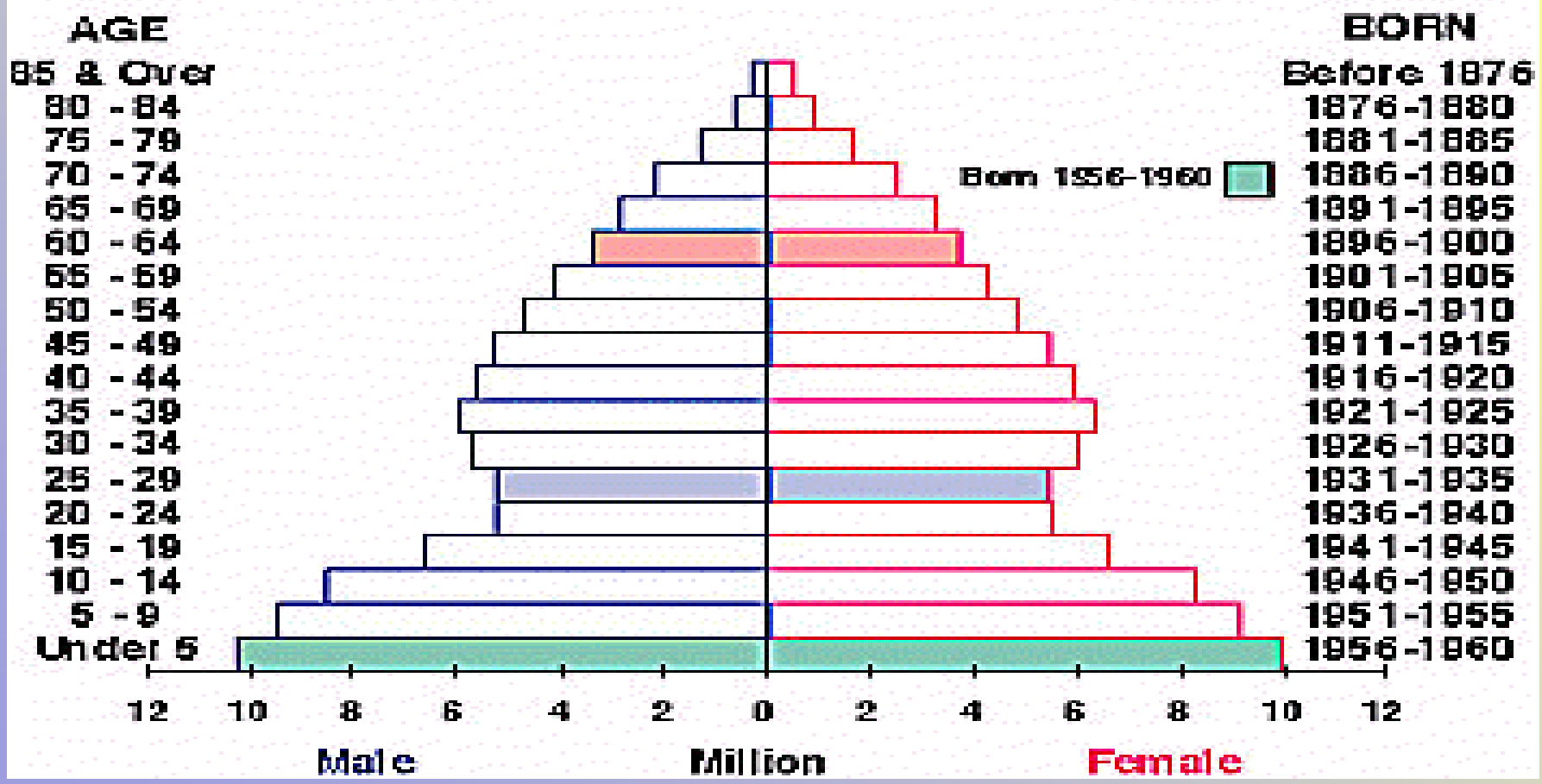
# 1950





# Histogram of Numbers of Males vs. Females

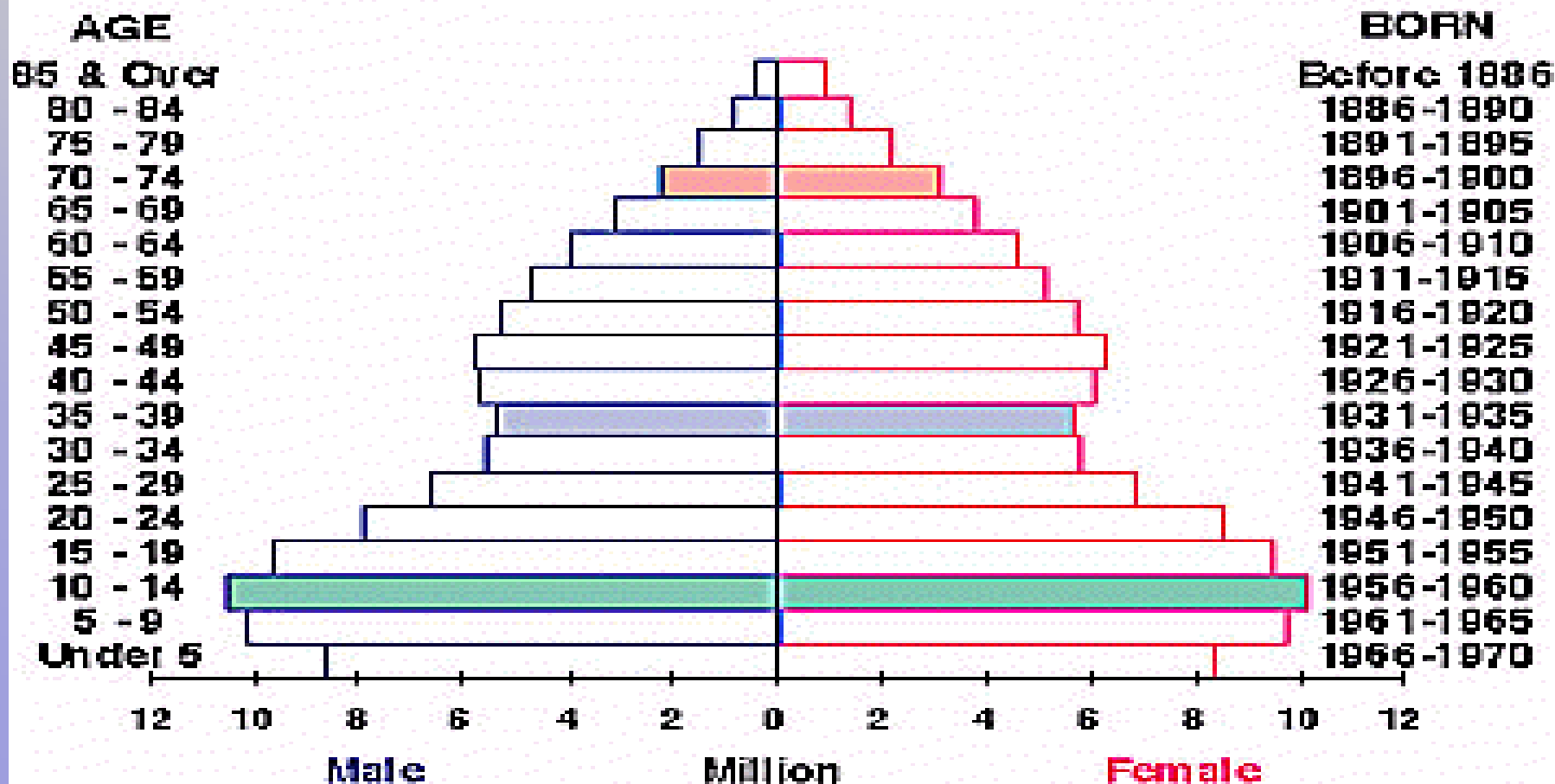
# 1960



# Histogram of Numbers of Males

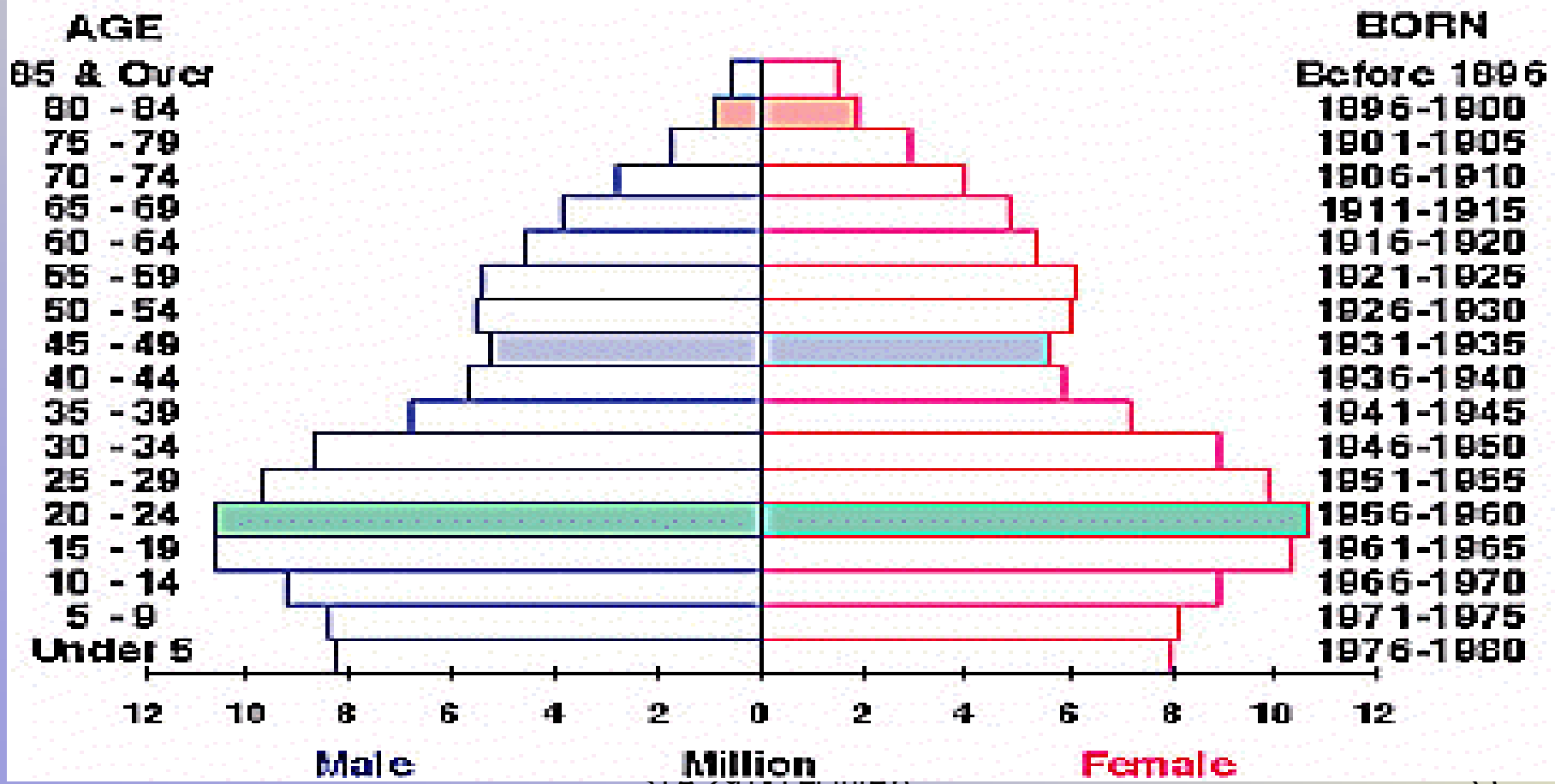
vs Females

# 1970



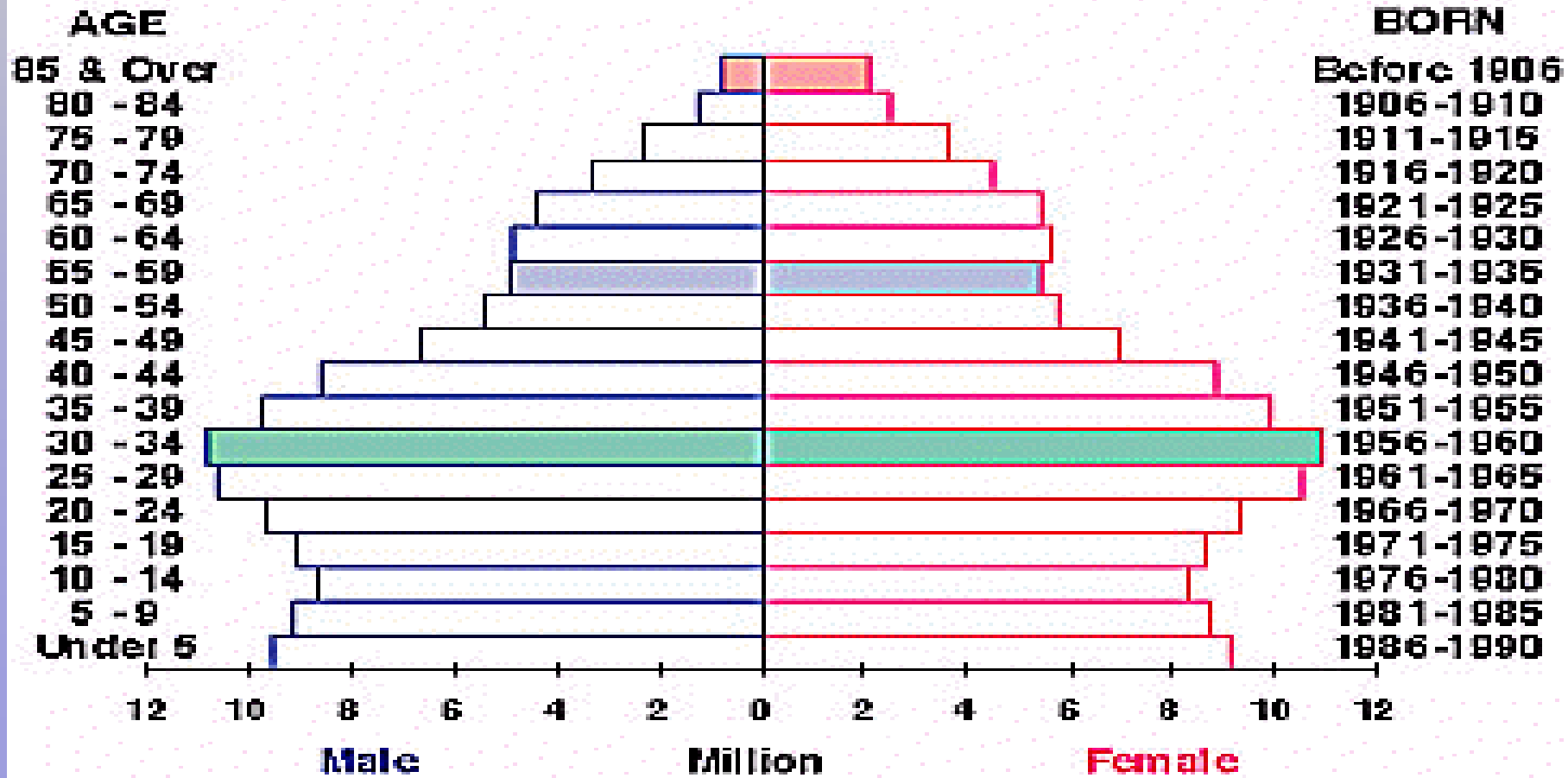
# Histogram of Numbers of Males vs. Females

# 1980



# Histogram of Numbers of Males vs. Females

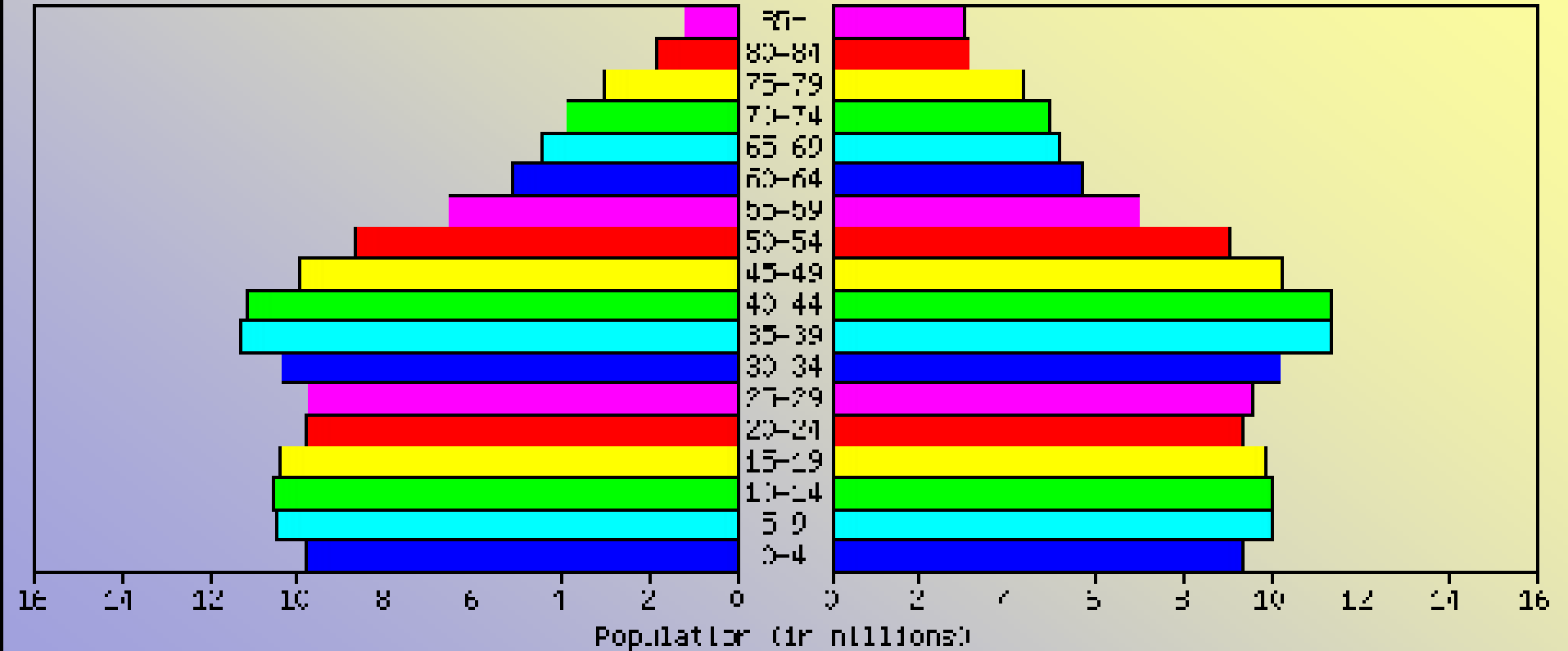
# 1990



### United States: 2000

YFLC

FCYFLC



Source: U.S. Census Bureau, International Data Base.

- Dynamic graph: graph changes over time
- [www.gapminder.org](http://www.gapminder.org)
- <http://www.gapminder.org/videos/ted-talks/hans-rosling-ted-talk-2007-seemingly-impossible-is-possible/>

# Histogram (for continuous numerical type data)

- Divide the range of possible values into many (contiguous, non-overlap) intervals, then count how many times data falls into each interval.
- Plot based on this table is called histogram.

# Data Table: Murder Rates per1000

Alabama	11.6	Alaska	9.0
Arizona	8.6	Arkansas	10.2
California	13.1	Colorado	5.8
Connecticut	6.3	Delaware	5.0
D C	78.5	Florida	8.9
Georgia	11.4	Hawaii	3.8
.....		.....	

- Difficult to see the “big picture” from these numbers
- Try to condense the data...



# Frequency Distribution

- A listing of intervals of possible values for a variable
- Together with a tabulation of the number of observations in each interval.

# Frequency Distribution

Murder Rate	Frequency
0-2.9	5
3-5.9	16
6-8.9	12
9-11.9	12
12-14.9	4
15-17.9	0
18-20.9	1
>21	1
Total	51

# Frequency Distribution

- Use intervals of same length (wherever possible)
- Intervals must be mutually exclusive: Any observation must fall into one and only one interval

# Relative Frequencies

- Relative frequency for an interval: The proportion of sample observations that fall in that interval
- Sometimes, percentages are preferred to relative frequencies

# Frequency and Relative Frequency and Percentage Distribution

Murder Rate	Frequency	Relative Frequency	Percentage
0-2.9	5	.10	10
3-5.9	16	.31	31
6-8.9	12	.24	24
9-11.9	12	.24	24
12-14.9	4	.08	8
15-17.9	0	0	0
18-20.9	1	.02	2
>21	1	.02	2
Total	51	1	100

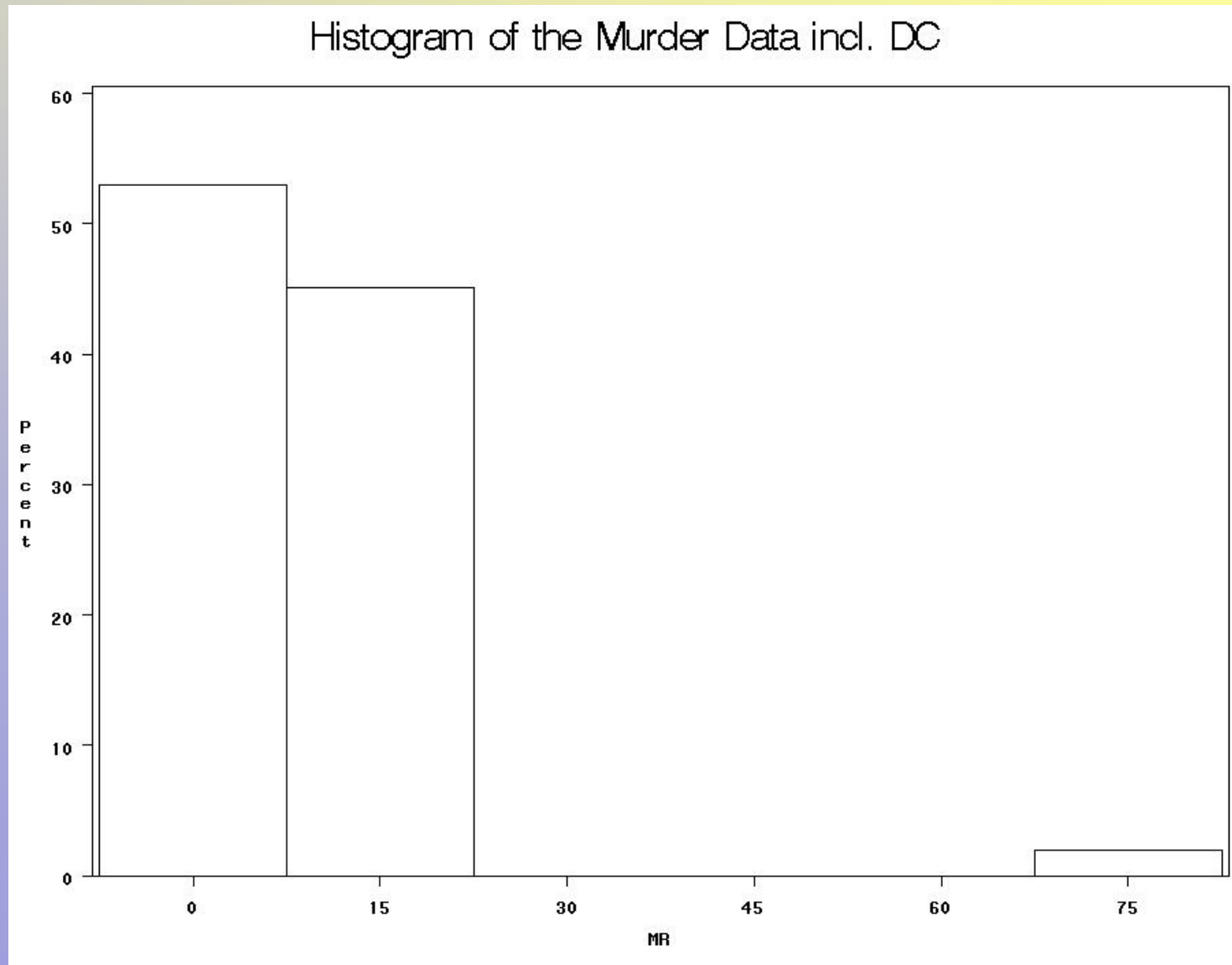
# Frequency Distributions

- Notice that we had to group the observations into intervals because the variable is measured on a continuous scale
- For discrete data, grouping may not be necessary (except when there are many categories)

# Histogram (for continuous numerical Data)

- Use the numbers from the frequency distribution to create a graph
- Draw a bar over each interval, the height of the bar represents the (relative) frequency for that interval
- Bars should be touching; I.e., equally extend the width of the bar at the upper and lower limits so that the bars are touching.

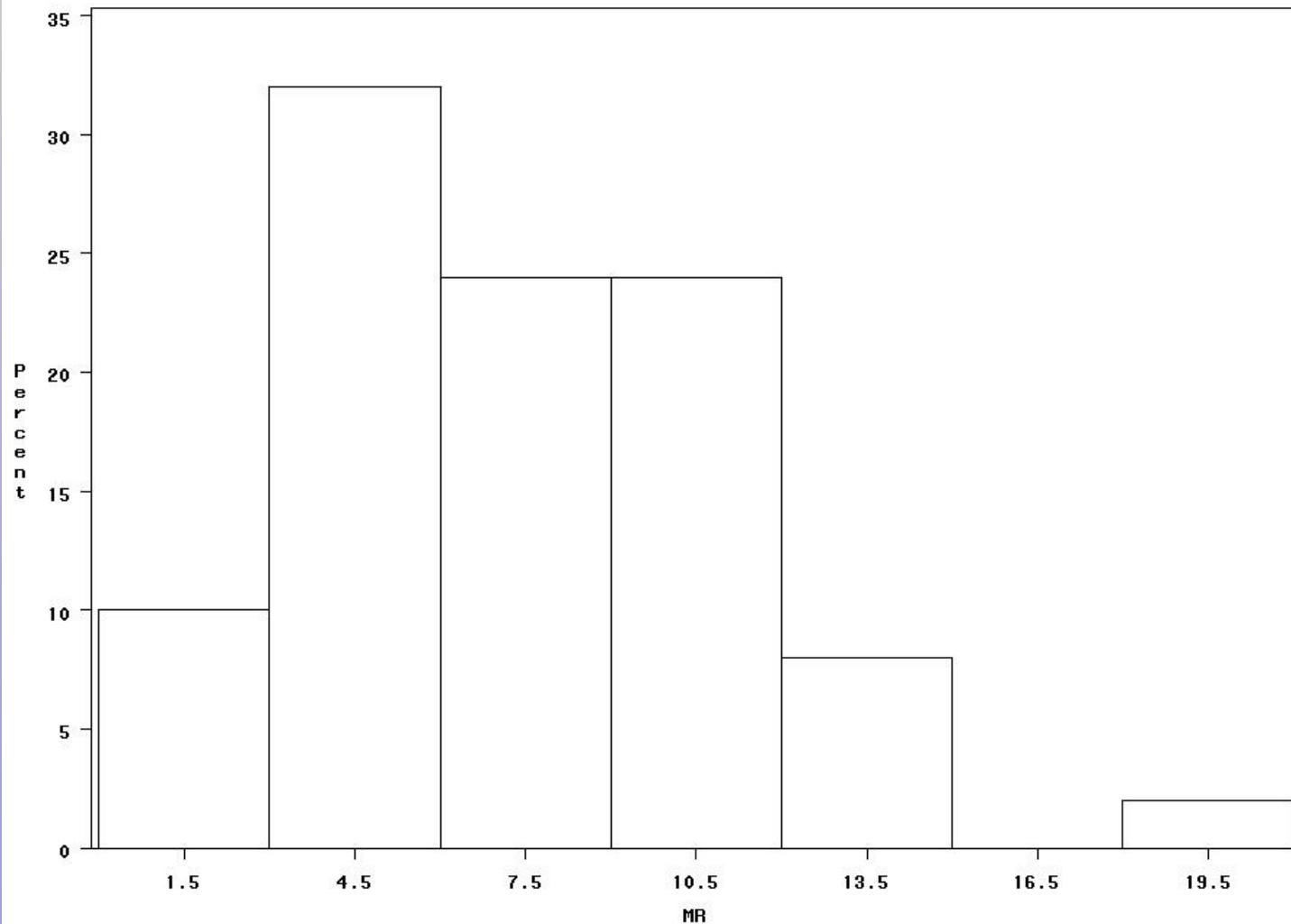
# Histogram





# Histogram w/o DC

Histogram of the Murder Data w/o DC



# Histogram

- Usually produced by software. We need to understand what they try to say.
- <http://www.shodor.org/interactivate/activities/histogram/>