Probability and Statistics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 2: RANDOM VARIABLES

1 Introduction

2 Random variables

2.1 Introduction

2.2 Types of data

2.3 Looking at data

2.4 Formal definition of a random variable

3 Cumulative distribution functions

4 Density functions

4.1 Discrete random variables

4.2 Continuous random variables

5 A gentle introduction to moments

5.1 Mean of a random variable

5.2 Variance of a random variable

5.3 Rules for means and variances

5.4 Moments and moment generating functions

5.5 Useful results

5.5.1 Law of large numbers

5.5.2 Expected value of a function of a random variable

5.5.3 Chebyshev inequality

5.5.4 Jensen inequality

1 Introduction

Assessing probabilities of events

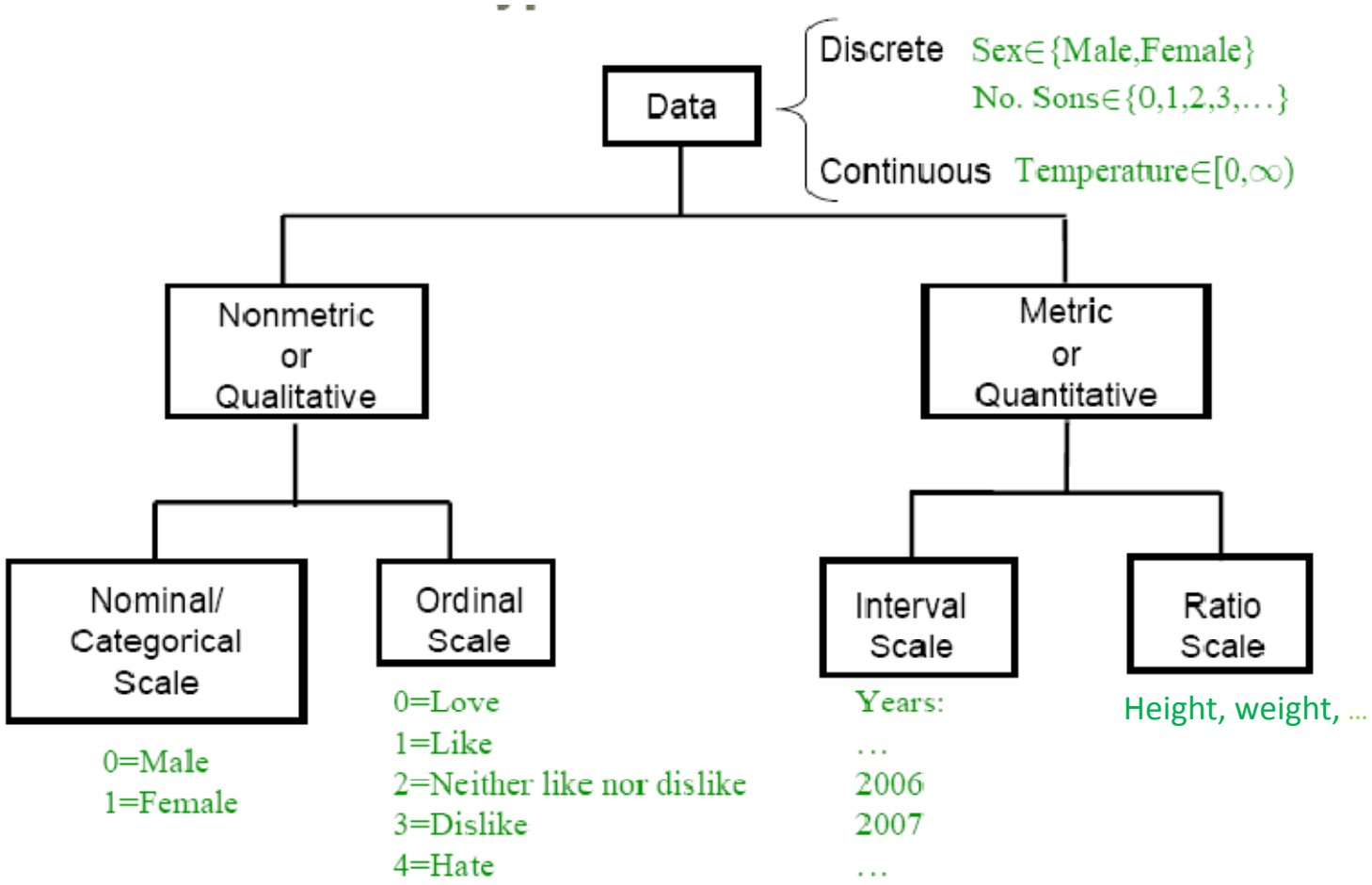
- As in Chapter 1, we would like to model our random experiment so as to be able to give values to the probabilities of events
- Recall that a probability space was defined by a triplet consisting of a
  - o sample space Ω
  - o collection of subsets $\mathscr{A}$ (of events)
  - o set function P[.] having domain and $\mathscr{A}$ counterdomain the interval [0,1]
- We need the notion of a *random variable* to describe events
- A *cumulative distribution function* will be used to give the probabilities of certain events defined in terms of random variables

2 Random variables

2.1 Introduction

- The outcome of an experiment need not be a number, for example, the outcome when a coin is tossed can be 'heads' or 'tails'. However, we often want to represent outcomes as numbers.
- A random variable is a function that associates a unique numerical value with every outcome of an experiment. The value of the random variable will vary from trial to trial as the experiment is repeated.
- Basically, there are two types of random variable - discrete and continuous.

2.2 Types of data



Scales of measurement

Data comes in various sizes and shapes and it is important to know about these so that the proper analysis can be used on the data. There are usually 4 scales of measurement that must be considered:

1. **Nominal Data**
   - classification data, e.g. m/f
   - no ordering, e.g. it makes no sense to state that M > F
   - arbitrary labels, e.g., m/f, 0/1, etc

2. **Ordinal Data**
   - ordered but differences between values are not important
   - e.g., political parties on left to right spectrum given labels 0, 1, 2
   - e.g., Likert scales, rank on a scale of 1..5 your degree of satisfaction
   - e.g., restaurant ratings

3. **Interval Data**
   - ordered, constant scale, but no natural zero
   - differences make sense, but ratios do not (e.g., 30°-20°=20°-10°, but 20°/10° is not twice as hot!
   - e.g., temperature (C,F), dates

4. **Ratio Data**
   - ordered, constant scale, natural zero
   - e.g., height, weight, age, length

Some computer packages (e.g. JMP) use these scales of measurement to make decisions about the type of analyses that should be performed. Also, some packages make no distinction between Interval or Ratio data calling them both *continuous*. However, this is, technically, not quite correct.

Only certain operations can be performed on certain scales of measurement. The following list summarizes which operations are legitimate for each scale. Note that you can always apply operations from a 'lesser scale' to any particular data, e.g. you may apply nominal, ordinal, or interval operations to an interval scaled datum.

- **Nominal Scale**. You are only allowed to examine if a nominal scale datum is equal to some particular value or to count the number of occurrences of each value. For example, gender is a nominal scale variable. You can examine if the gender of a person is F or to count the number of males in a sample.

- **Ordinal Scale**. You are also allowed to examine if an ordinal scale datum is less than or greater than another value. Hence, you can 'rank' ordinal data, but you cannot 'quantify' differences between two ordinal values. For example, political party is an ordinal datum with the NDP to left of Conservative Party, but you can't quantify the difference. Another example, are preference scores, e.g. ratings of eating establishments where 10=good, 1=poor, but the difference between an establishment with a 10 ranking and an 8 ranking can't be quantified.

- **Interval Scale**. You are also allowed to quantify the difference between two interval scale values but there is no natural zero. For example, temperature scales are interval data with 25C warmer than 20C and a 5C difference has some physical meaning. Note that 0C is arbitrary, so that it does not make sense to say that 20C is twice as hot as 10C.

- **Ratio Scale**. You are also allowed to take ratios among ratio scaled variables. Physical measurements of height, weight, length are typically ratio variables. It is now meaningful to say that 10 m is twice as long as 5 m. This ratio hold true regardless of which scale the object is being measured in (e.g. meters or yards). This is because there is a natural zero.

Coding … of categorical variables

3 "dummy variables are sufficient

Hair Colour
{Brown, Blond, Black, Red}

No order →

$$(x_{Brown}, x_{Blond}, x_{Black}, x_{Red}) \in \{0,1\}^4$$

Peter: Black	Peter: $\{0,0,1,0\}$
Molly: Blond	Molly: $\{0,1,0,0\}$
Charles: Brown	Charles: $\{1,0,0,0\}$

Company size
{Small, Medium, Big}

Implicit order →

$$x_{size} \in \{0,1,2\}$$

Company A: Big	Company A: 2
Company B: Small	Company B: 0
Company C: Medium	Company C: 1

2.3 Looking at data

How do we know whether a variable is quantitative or qualitative?

Ask:

- What are the *n* individuals/units in the sample (of size "*n*")?
- What is being recorded about those *n* individuals/units?
- Is that a number (→ quantitative) or a statement (→ categorical)?

Categorical
Each individual is assigned to one of several categories.

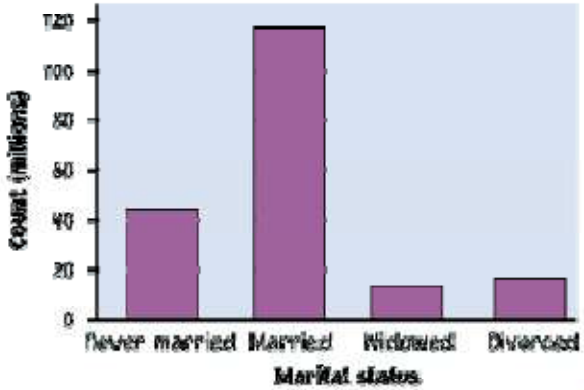
Quantitative
Each individual is attributed a numerical value.

Individuals in sample	DIAGNOSIS	AGE AT DEATH
Patient A	Heart disease	56
Patient B	Stroke	70
Patient C	Stroke	75
Patient D	Lung cancer	60
Patient E	Heart disease	80
Patient F	Accident	73
Patient G	Diabetes	69

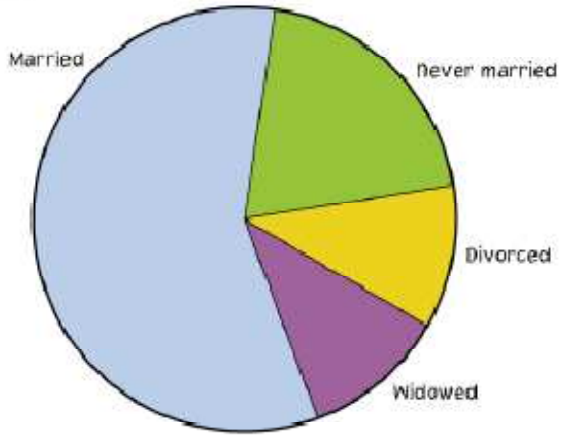
Ways to chart categorical data

Because the variable is categorical, the data in the graph can be ordered any way we want (alphabetical, by increasing value, by year, by personal preference, etc.)

- **Bar graphs**
  Each category is represented by a bar.



- **Pie charts**
  The slices must represent the parts of one whole.



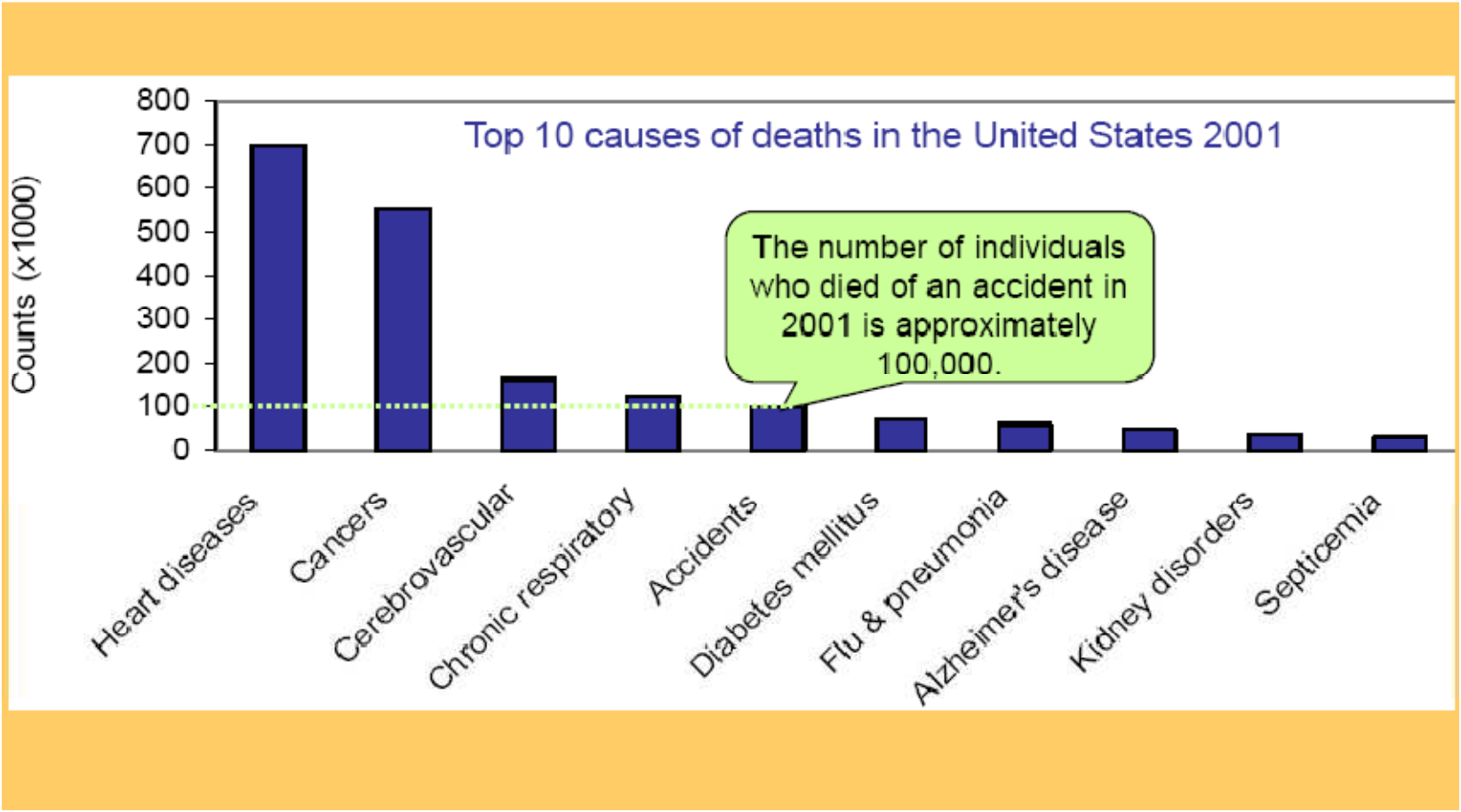
Example: Top 10 causes of death in the United States 2001

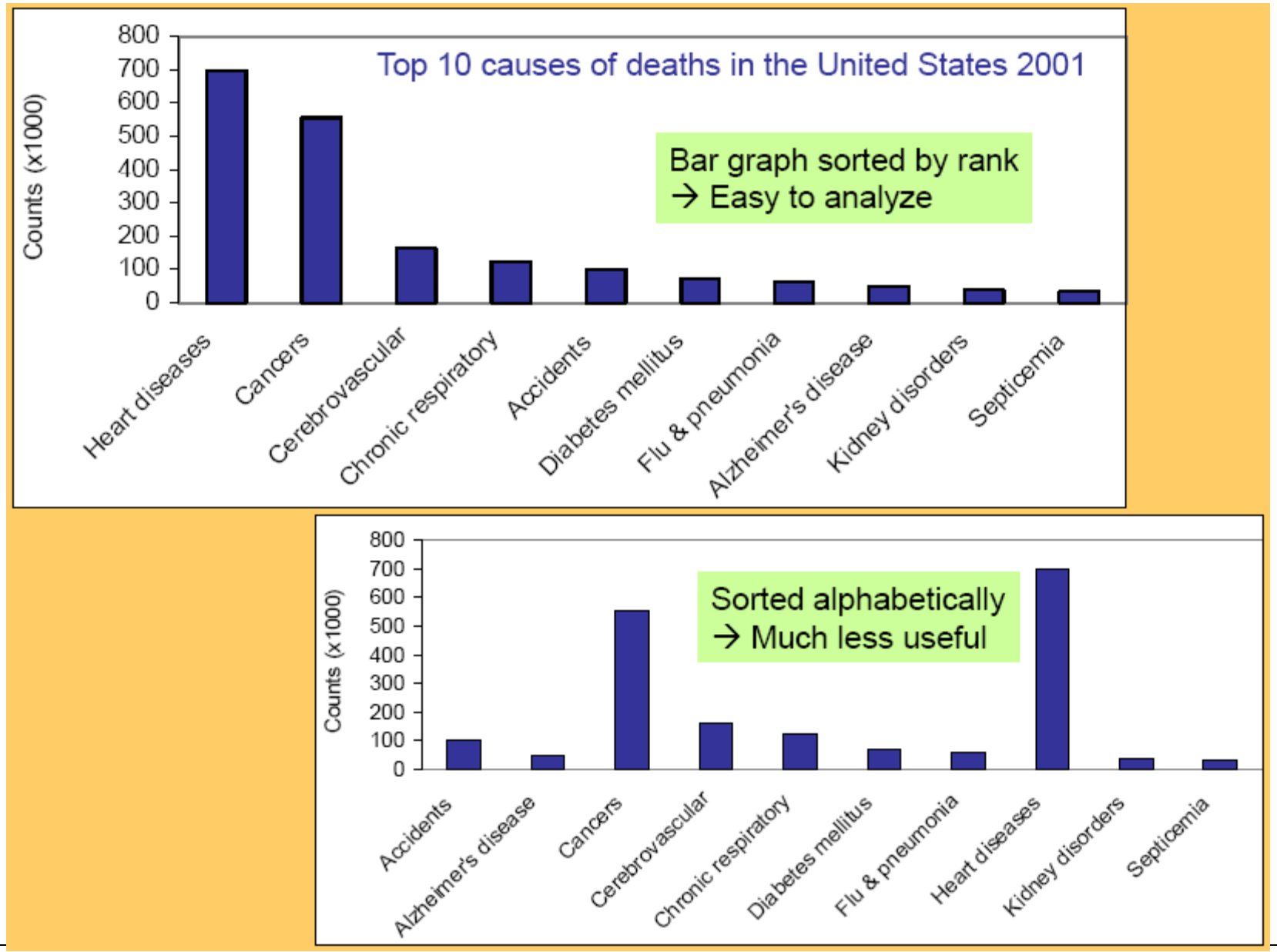
Rank	Causes of death	Counts	% of top 10s	% of total deaths
1	Heart disease	700,142	37%	*29%*
2	Cancer	553,768	29%	*23%*
3	Cerebrovascular	163,538	9%	*7%*
4	Chronic respiratory	123,013	6%	*5%*
5	Accidents	101,537	5%	*4%*
6	Diabetes mellitus	71,372	4%	*3%*
7	Flu and pneumonia	62,034	3%	*3%*
8	Alzheimer's disease	53,852	3%	*2%*
9	Kidney disorders	39,480	2%	*2%*
10	Septicemia	32,238	2%	*1%*
	All other causes	*629,967*		*26%*

For each individual who died in the United States in 2001, we record what was the cause of death. The table above is a summary of that information.

Bar graphs

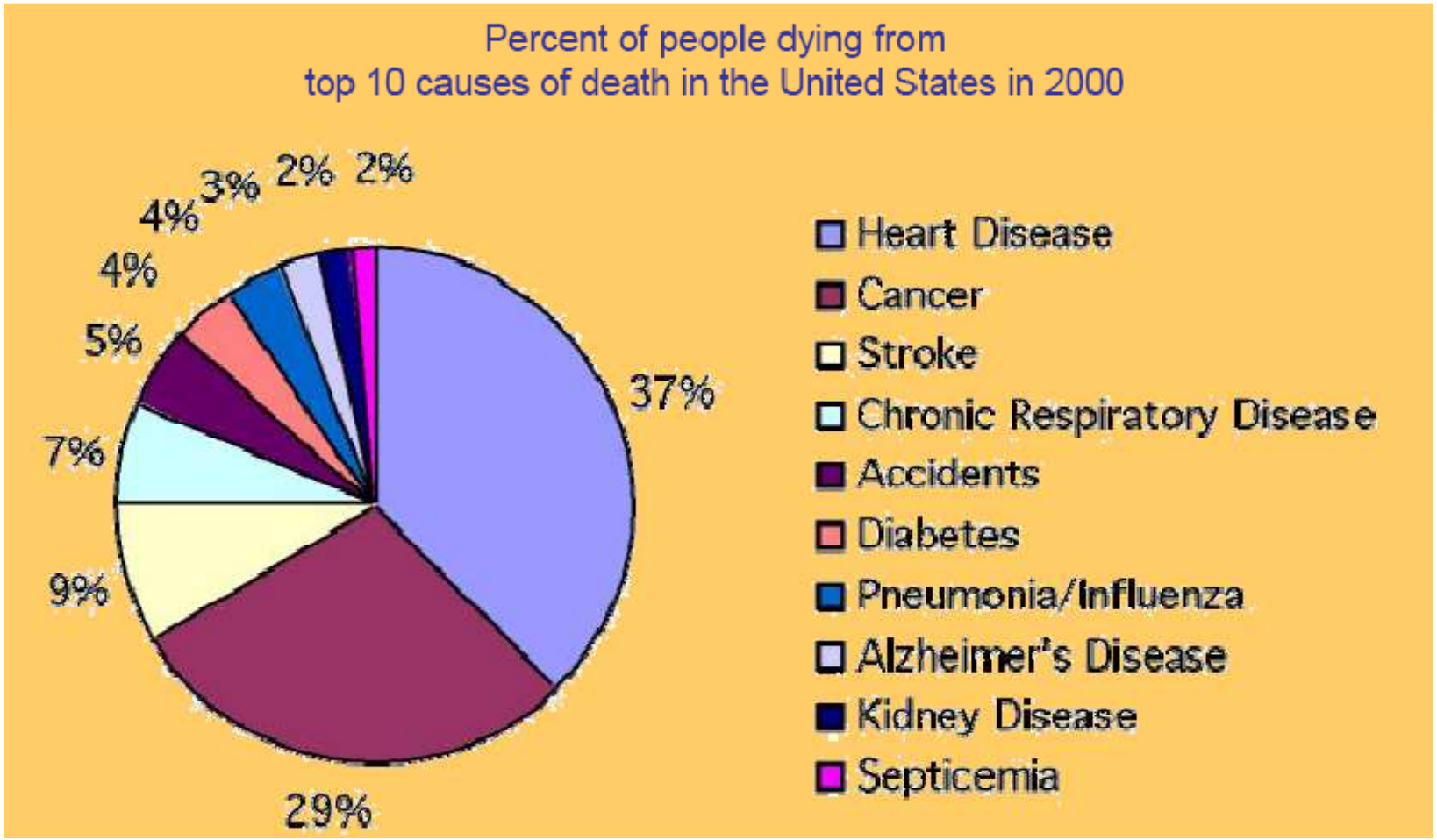
Each category is represented by one bar. The bar's height shows the count (or sometimes the percentage) for that particular category.

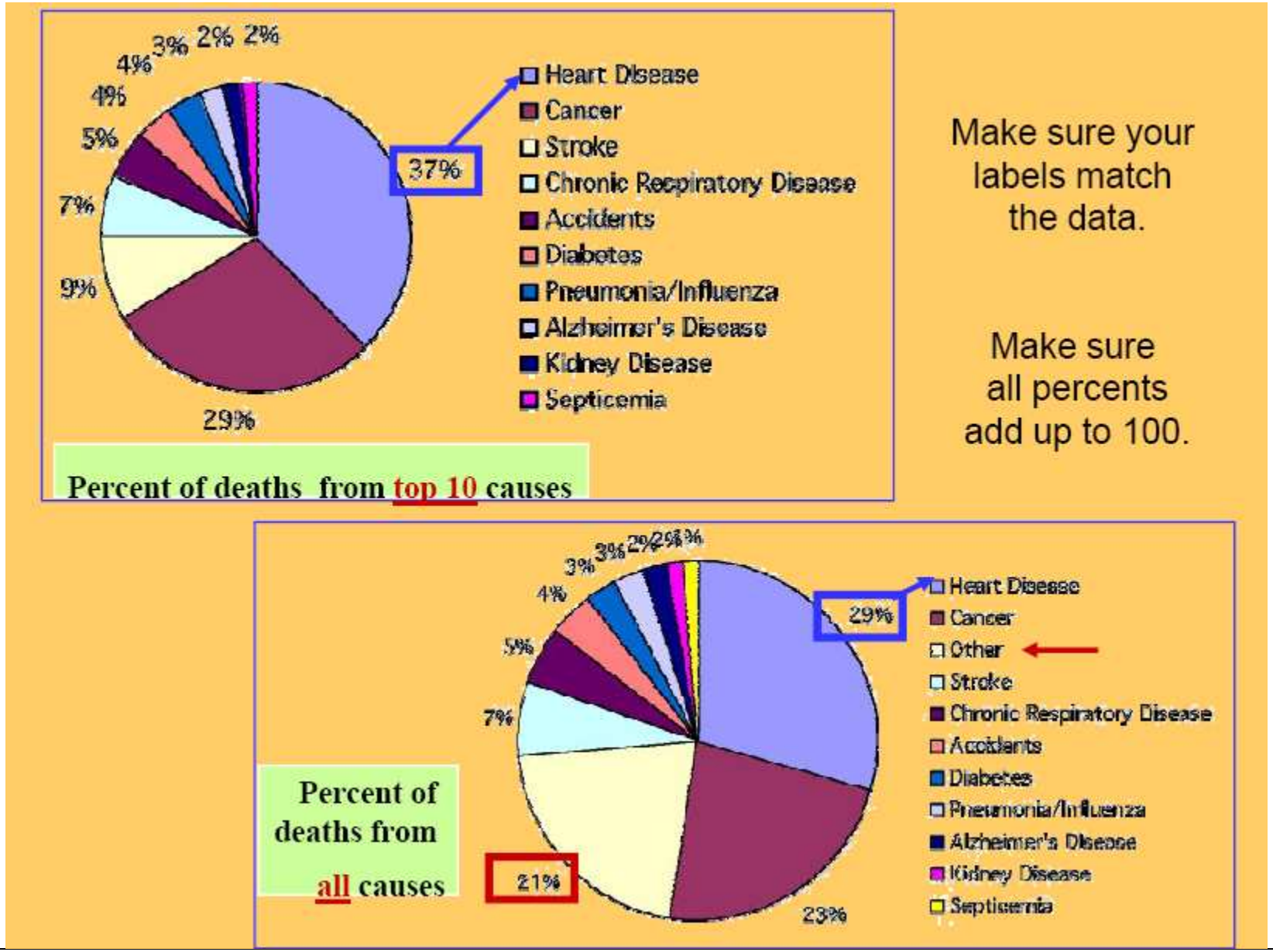



Top 10 causes of deaths in the United States 2001
Counts (x1000)
800
700
600
500
400
300
200
100
0
Heart diseases
Cancers
Cerebrovascular
Chronic respiratory
Accidents
Diabetes mellitus
Flu & pneumonia
Alzheimer's disease
Kidney disorders
Septicemia
Bar graph sorted by rank
→ Easy to analyze
Counts (x1000)
800
700
600
500
400
300
200
100
0
Accidents
Alzheimer's disease
Cancers
Cerebrovascular
Chronic respiratory
Diabetes mellitus
Flu & pneumonia
Heart diseases
Kidney disorders
Septicemia
Sorted alphabetically
→ Much less useful

Pie charts

Each slice represents a piece of one whole. The size of a slice depends on what percent of the whole this category represents.




3% 2% 2%
4%
4%
5%
7%
9%
29%
37%
Heart Disease
Cancer
Stroke
Chronic Respiratory Disease
Accidents
Diabetes
Pneumonia/Influenza
Alzheimer's Disease
Kidney Disease
Septicemia
Percent of deaths from top 10 causes
Make sure your labels match the data.
Make sure all percents add up to 100.
Percent of deaths from all causes
29%
21%
23%
Heart Disease
Cancer
Other
Stroke
Chronic Respiratory Disease
Accidents
Diabetes
Pneumonia/Influenza
Alzheimer's Disease
Kidney Disease
Septicemia

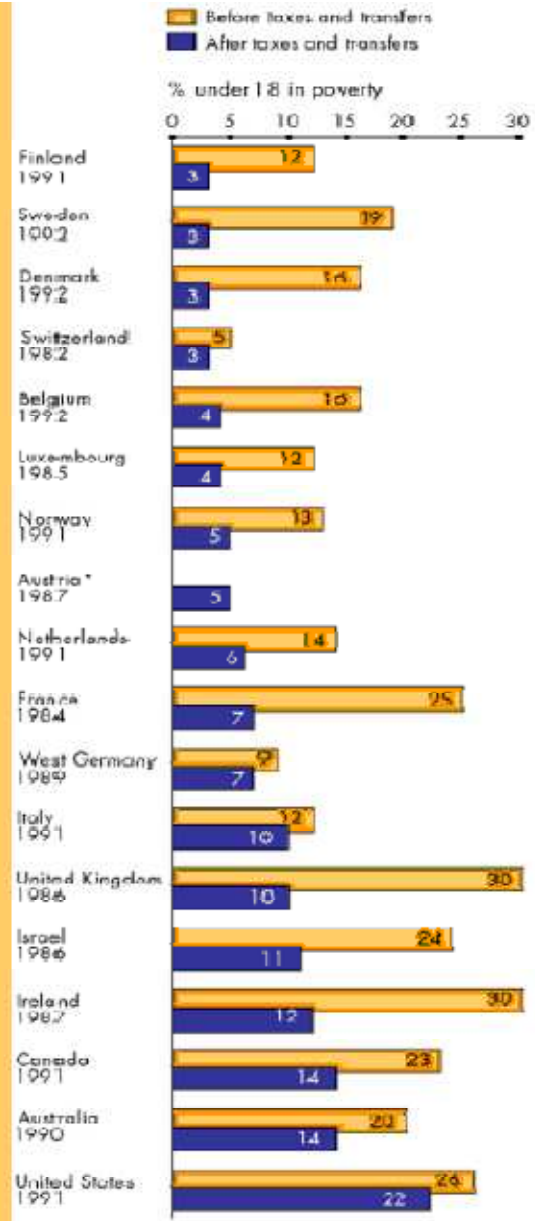
Child poverty before and after government intervention—UNICEF, 1996

What does this chart tell you?

- The United States has the highest rate of child poverty among developed nations (22% of under 18).
- Its government does the least—through taxes and subsidies—to remedy the problem (size of orange bars and percent difference between orange/blue bars).

Could you transform this bar graph to fit in 1 pie chart? In two pie charts? Why?

The poverty line is defined as 50% of national median income.



Ways to chart quantitative data

- Histograms and stemplots

  These are summary graphs for a single variable. They are very useful to understand the pattern of variability in the data.

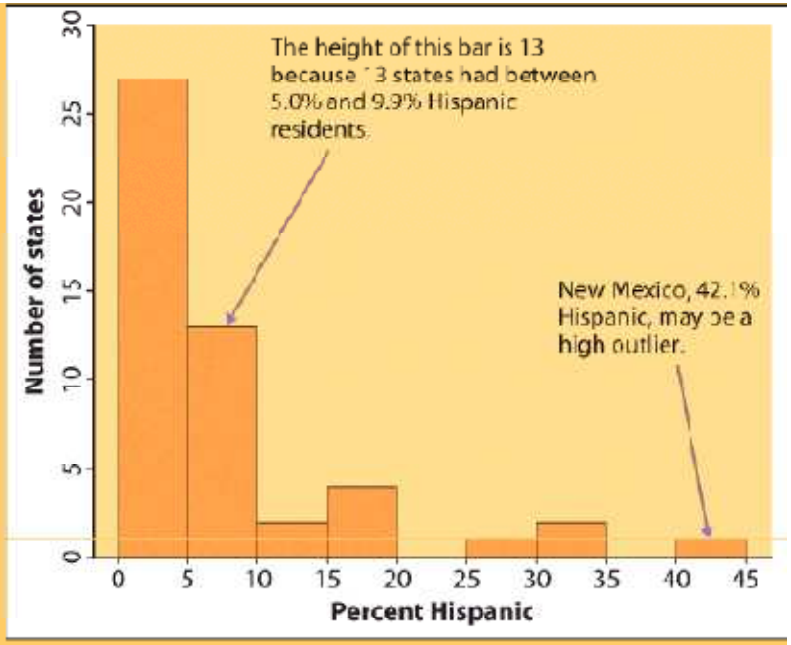
- Line graphs: time plots

  Use when there is a meaningful sequence, like time. The line connecting the points helps emphasize any change over time.

Histograms

The range of values that a variable can take is divided into equal size intervals.

The histogram shows the number of individual data points that fall in each interval.



The first column represents all states with a Hispanic percent in their population between 0% and 4.99%. The height of the column shows how many states (27) have a percent in this range.

The last column represents all states with a Hispanic percent in their population between 40% and 44.99%. There is only one such state: New Mexico, at 42.1% Hispanics.

Stem plots

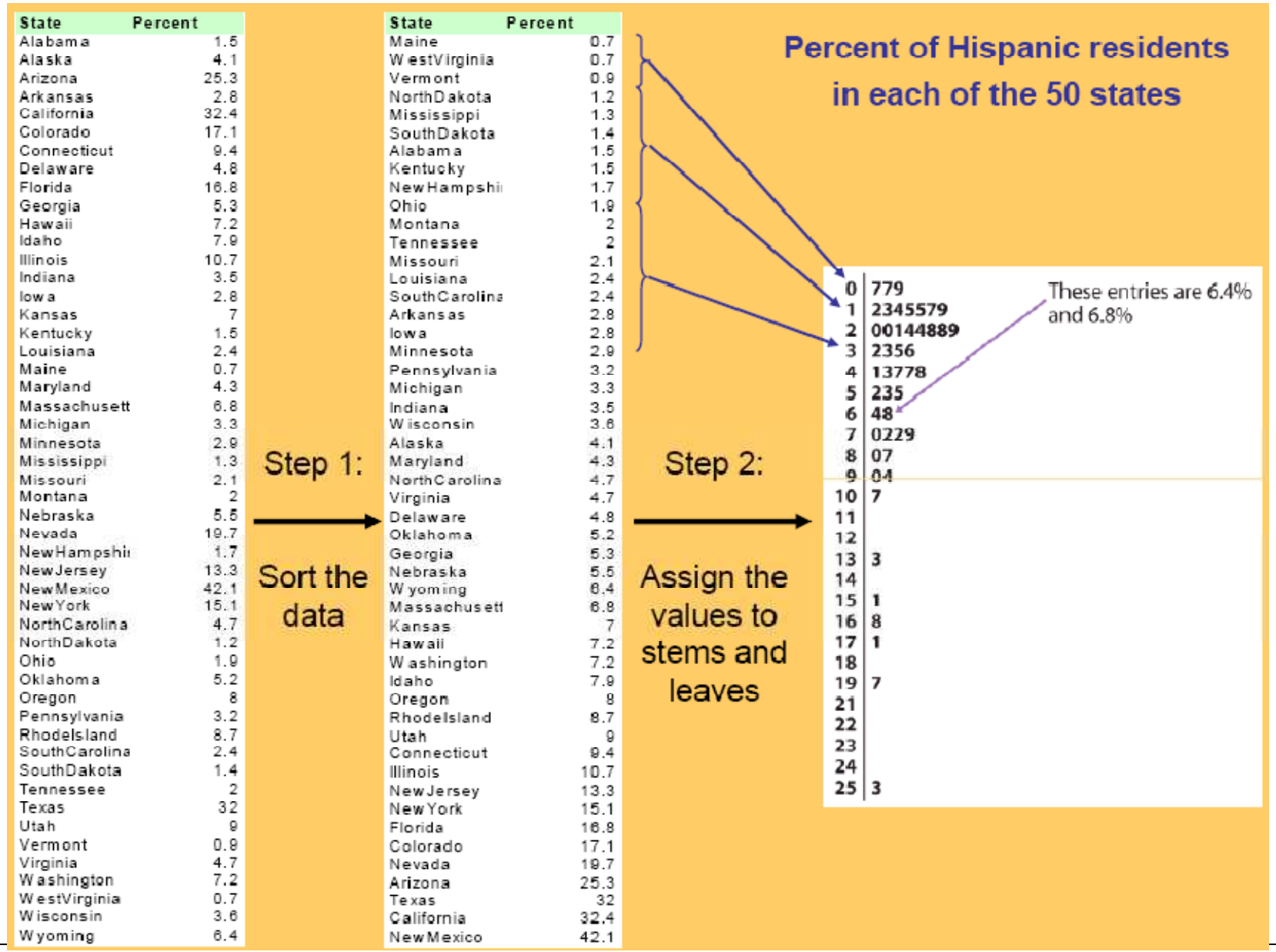
How to make a **stemplot:**

1) Separate each observation into a **stem**, consisting of all but the final (rightmost) digit, and a **leaf**, which is that remaining final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.

2) Write the stems in a vertical column with the smallest value at the top, and draw a vertical line at the right of this column.

3) Write each leaf in the row to the right of its stem, in increasing order out from the stem.

STEM	LEAVES
0	9 9
1	
2	2
3	2 3 9 9
4	2 9
5	2 8
6	
7	0

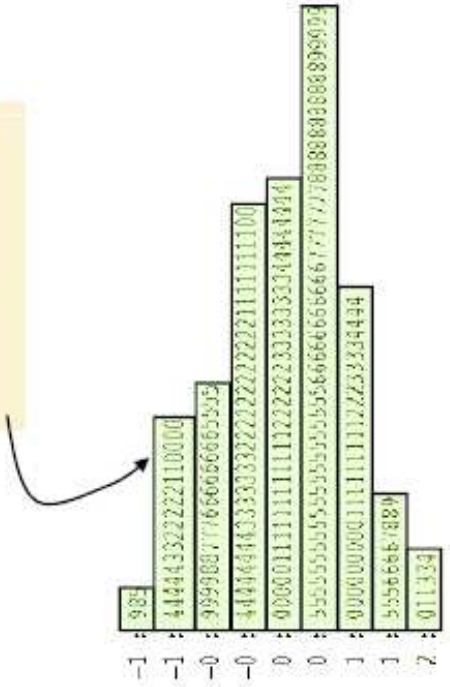

State Percent
Alabama 1.5
Alaska 4.1
Arizona 25.3
Arkansas 2.8
California 32.4
Colorado 17.1
Connecticut 9.4
Delaware 4.8
Florida 16.8
Georgia 5.3
Hawaii 7.2
Idaho 7.9
Illinois 10.7
Indiana 3.5
Iowa 2.8
Kansas 7
Kentucky 1.5
Louisiana 2.4
Maine 0.7
Maryland 4.3
Massachusett 6.8
Michigan 3.3
Minnesota 2.9
Mississippi 1.3
Missouri 2.1
Montana 2
Nebraska 5.5
Nevada 19.7
NewHampshi 1.7
NewJersey 13.3
NewMexico 42.1
NewYork 15.1
NorthCarolina 4.7
NorthDakota 1.2
Ohio 1.9
Oklahoma 5.2
Oregon 8
Pennsylvania 3.2
Rhodelsland 8.7
SouthCarolina 2.4
SouthDakota 1.4
Tennessee 2
Texas 32
Utah 9
Vermont 0.9
Virginia 4.7
Washington 7.2
WestVirginia 0.7
Wisconsin 3.6
Wyoming 6.4
Step 1:
Sort the data
State Percent
Maine 0.7
WestVirginia 0.7
Vermont 0.9
NorthDakota 1.2
Mississippi 1.3
SouthDakota 1.4
Alabama 1.5
Kentucky 1.5
NewHampshi 1.7
Ohio 1.9
Montana 2
Tennessee 2
Missouri 2.1
Louisiana 2.4
SouthCarolina 2.4
Arkansas 2.8
Iowa 2.8
Minnesota 2.9
Pennsylvania 3.2
Michigan 3.3
Indiana 3.5
Wisconsin 3.6
Alaska 4.1
Maryland 4.3
NorthCarolina 4.7
Virginia 4.7
Delaware 4.8
Oklahoma 5.2
Georgia 5.3
Nebraska 5.5
Wyoming 6.4
Massachusett 6.8
Kansas 7
Hawaii 7.2
Washington 7.2
Idaho 7.9
Oregon 8
Rhodelsland 8.7
Utah 9
Connecticut 9.4
Illinois 10.7
NewJersey 13.3
NewYork 15.1
Florida 16.8
Colorado 17.1
Nevada 19.7
Arizona 25.3
Texas 32
California 32.4
NewMexico 42.1
Step 2:
Assign the values to stems and leaves
Percent of Hispanic residents in each of the 50 states
0 | 779
1 | 2345579
2 | 00144889
3 | 2356
4 | 13778
5 | 235
6 | 48
7 | 0229
8 | 07
9 | 04
10 | 7
11 |
12 |
13 | 3
14 |
15 | 1
16 | 8
17 | 1
18 |
19 | 7
21 |
22 |
23 |
24 |
25 | 3
These entries are 6.4% and 6.8%

- To compare two related distributions, a back-to-back stem plot with common stems is useful.
- Stem plots do not work well for large datasets.
- When the observed values have too many digits, trim the numbers before making a stem plot.
- When plotting a moderate number of observations, you can split each stem.

Stem plot or histogram?

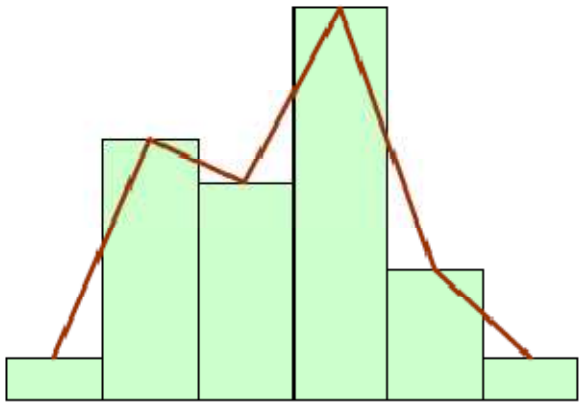
Stemplots are quick and dirty histograms that can easily be done by hand, and therefore are very convenient for back of the envelope calculations. However, they are rarely found in scientific or laymen publications.

```
-1 : 985
-1 : 444443322222110000
-0 : 999988777666666665555
-0 : 444444443333333222222222221111111100
 0 : 000001111111111112222223333333344444444
 0 : 555555555555555555556666666666677777778888888888899999
 1 : 000000000111111111222333344444
 1 : 55566667889
 2 : 011334
```

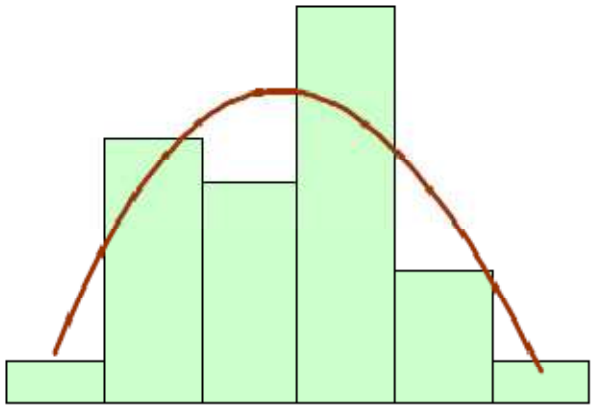


Interpreting histograms

When describing the distribution of a quantitative variable, we look for the overall pattern and for striking deviations from that pattern. We can describe the *overall* pattern of a histogram by its **shape, center**, and **spread.**



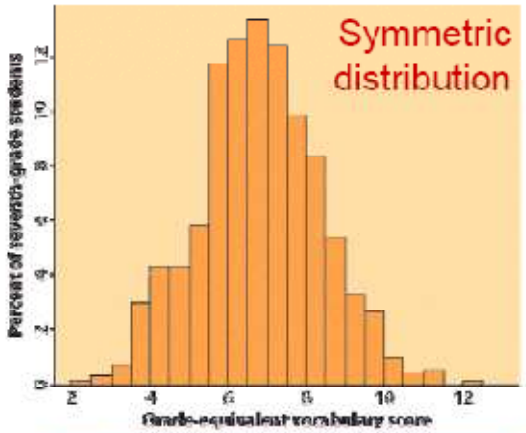
Histogram with a line connecting each column → too detailed

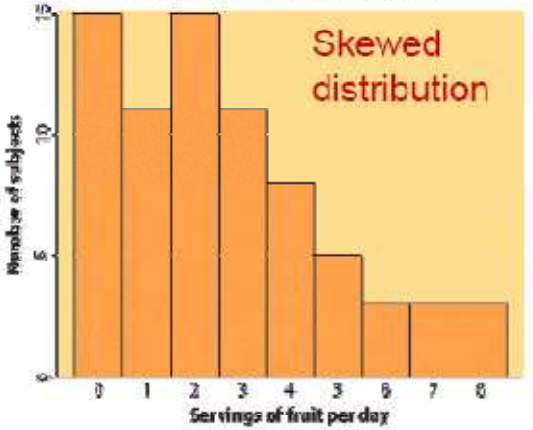


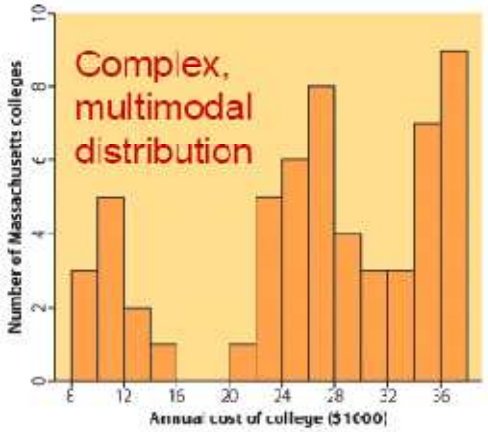
Histogram with a smoothed curve highlighting the overall pattern of the distribution

Most common distribution shapes

- A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.
- A distribution is **skewed to the right** if the right side of the histogram (side with larger values) extends much farther out than the left side. It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.







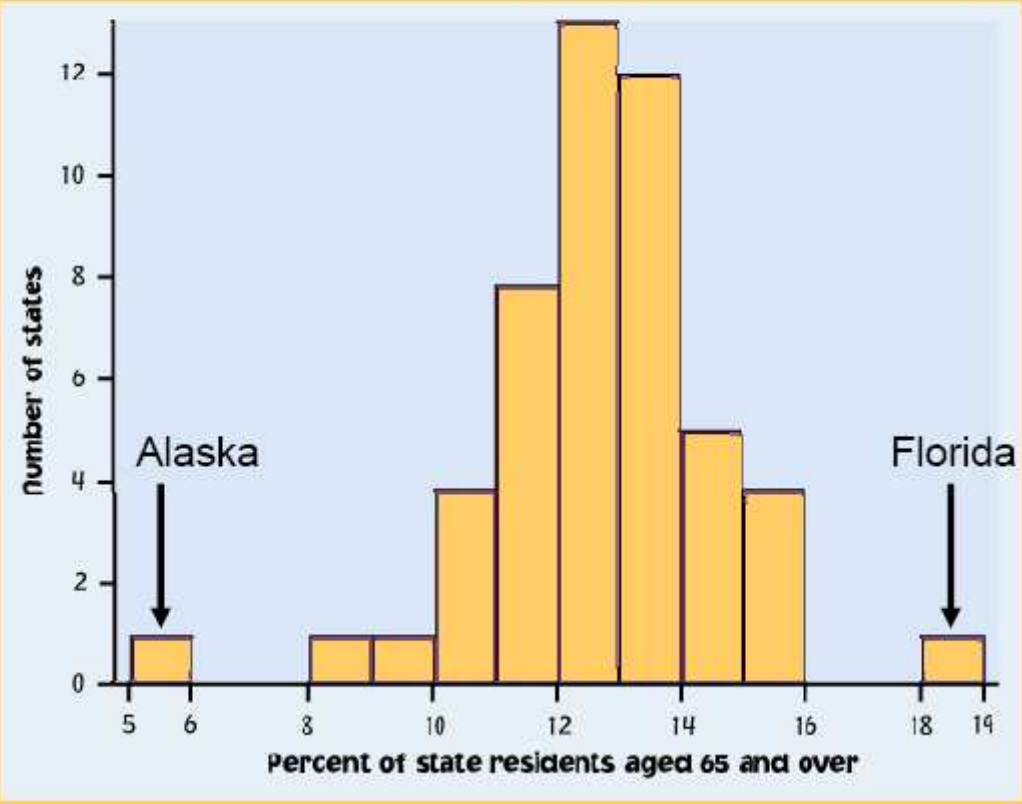
- Not all distributions have a simple overall shape, especially when there are few observations.

Outliers

An important kind of deviation is an **outlier.** Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

The overall pattern is fairly symmetrical except for 2 states that clearly do not belong to the main trend. Alaska and Florida have unusual representation of the elderly in their population.

A large gap in the distribution is typically a sign of an outlier.



How to create a histogram?

It is an iterative process – try and try again.

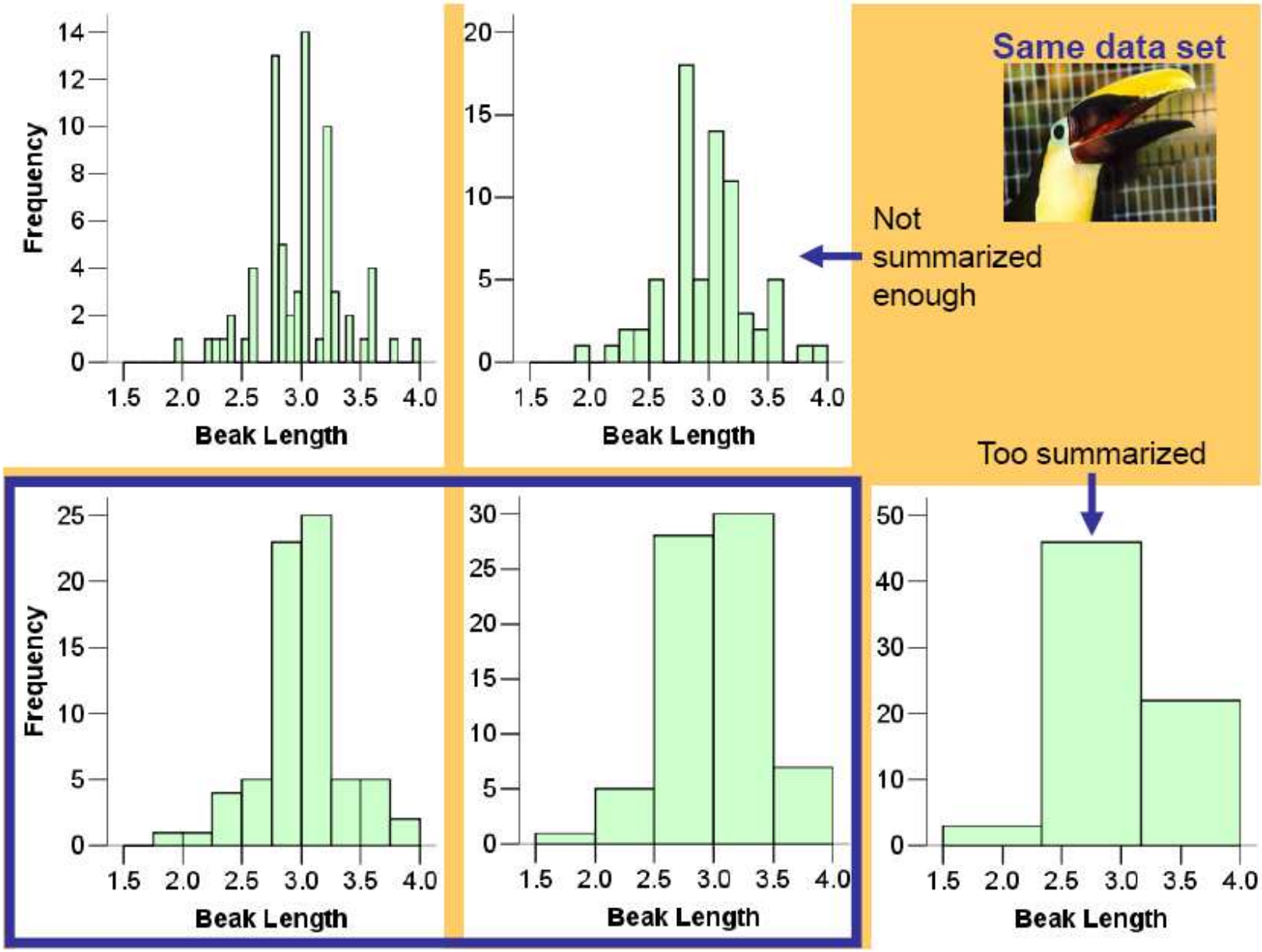
What bin size should you use?

- Not too many bins with either 0 or 1 counts
- Not overly summarized that you loose all the information
- Not so detailed that it is no longer summary

➔ rule of thumb: start with 5 to 10 bins

Look at the distribution and refine your bins

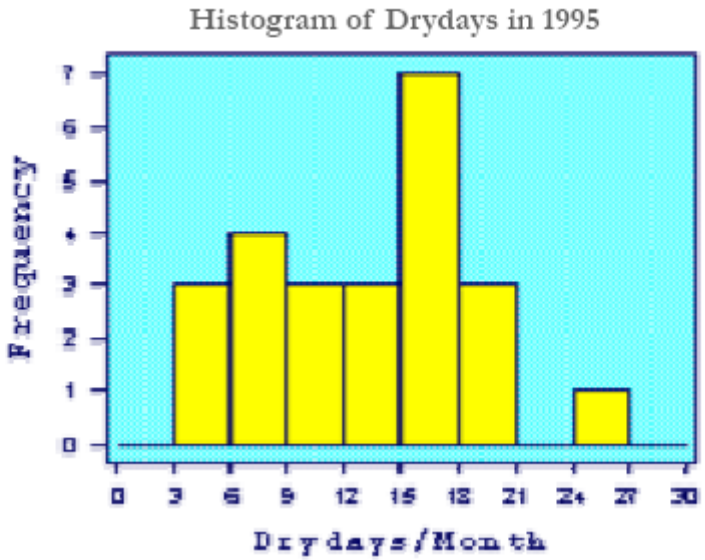
(There isn't a unique or "perfect" solution)

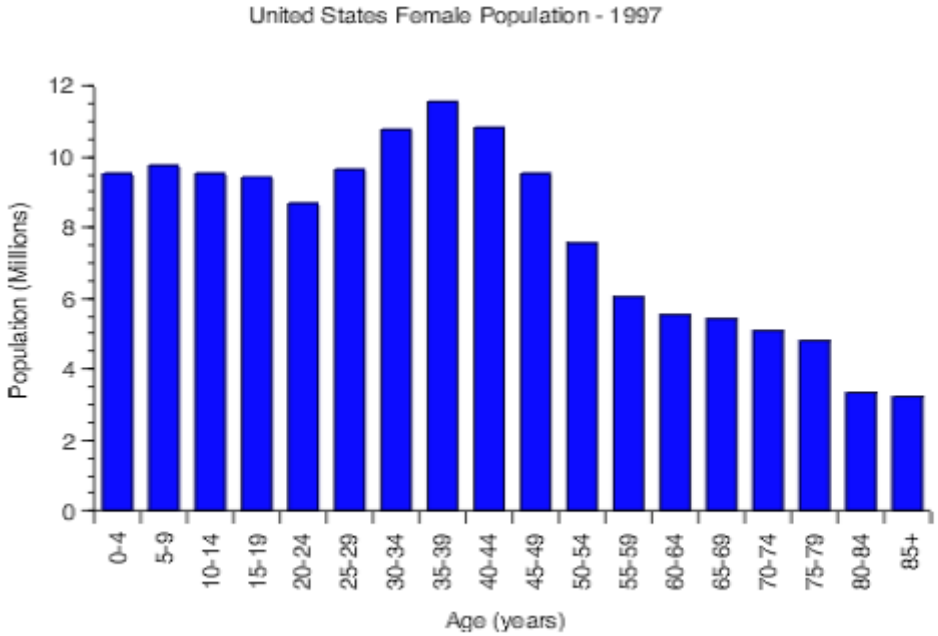

Frequency
Beak Length
Same data set
Not summarized enough
Too summarized
1.5 2.0 2.5 3.0 3.5 4.0

IMPORTANT NOTE:

Your data are the way they are.

Do not try to force them into a particular shape.

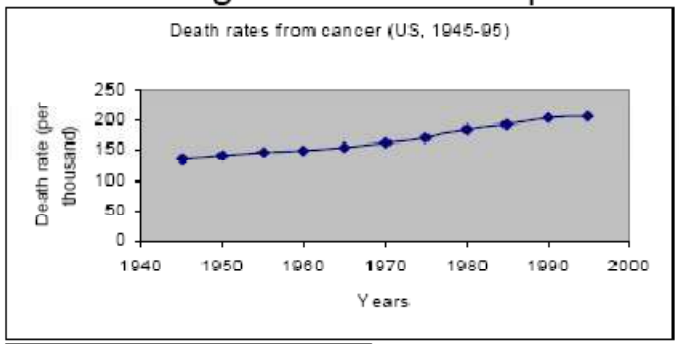


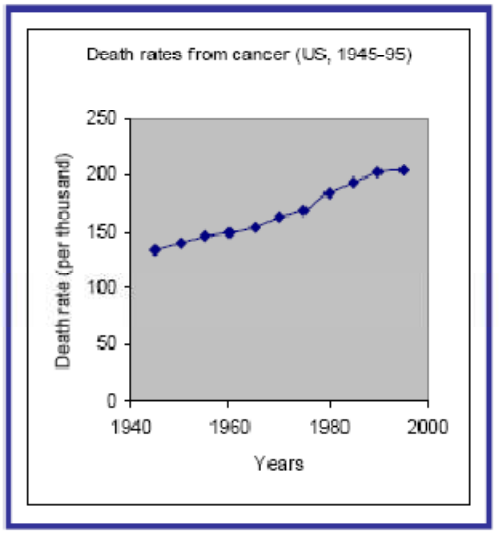


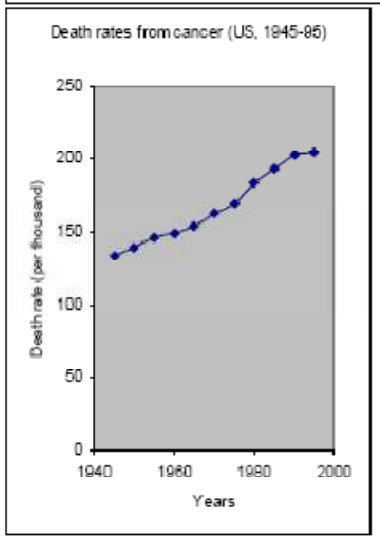
It is a common misconception that if you have a large enough data set, the data will eventually turn out nice and symmetrical.

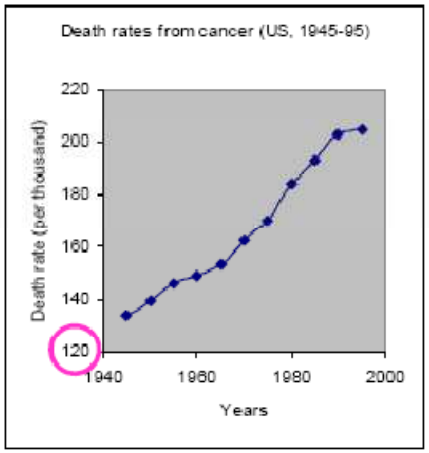
Cautionary note : scale matters when visualizing data

How you stretch the axes and choose your scales can give a different impression.









A picture is worth a thousand words,

BUT

There is nothing like hard numbers.
→ **Look at the scales.**

2.4 Formal definition of a random variable

Definition **Random Variable** For a given probability space $(\Omega, \mathscr{A}, P[\cdot])$, a *random variable*, denoted by X or $X(\cdot)$, is a function with domain Ω and counterdomain the real line. The function $X(\cdot)$ must be such that the set A_r, defined by $A_r = \{\omega: X(\omega) \leq r\}$, belongs to $\mathscr{A}$ for every real number r. ////

The use of words "random" and "variable" in the above definition is unfortunate since their use cannot be convincingly justified. The expression "random variable" is a misnomer that has gained such widespread use that it would be foolish for us to try to rename it.

EXAMPLE Consider the experiment of tossing a single coin. Let the random variable X denote the number of heads. $\Omega = \{\text{head, tail}\}$, and $X(\omega) = 1$ if $\omega = \text{head}$, and $X(\omega) = 0$ if $\omega = \text{tail}$; so, the random variable X associates a real number with each outcome of the experiment. We called X a random variable so mathematically speaking we should show that it satisfies the definition; that is, we should show that $\{\omega: X(\omega) \leq r\}$ belongs to $\mathscr{A}$ for every real number r. $\mathscr{A}$ consists of the four subsets: ϕ, {head}, {tail}, and Ω. Now, if $r < 0$, $\{\omega: X(\omega) \leq r\} = \phi$; and if $0 \leq r < 1$, $\{\omega: X(\omega) \leq r\} = \{\text{tail}\}$; and if $r \geq 1$, $\{\omega: X(\omega) \leq r\} = \Omega = \{\text{head, tail}\}$. Hence, for each r the set $\{\omega: X(\omega) \leq r\}$ belongs to $\mathscr{A}$; so $X(\cdot)$ is a random variable. ////

EXAMPLE Consider the experiment of tossing two dice. Ω can be described by the 36 points displayed in Fig. . $\Omega = \{(i, j): i = 1, \ldots, 6 \text{ and } j = 1, \ldots, 6\}$. Several random variables can be defined; for instance, let X denote the sum of the upturned faces; so $X(\omega) = i + j$ if $\omega = (i, j)$. Also, let Y denote the absolute difference between the upturned faces; then $Y(\omega) = |i - j|$ if $\omega = (i, j)$. It can be shown that both X and Y are random variables. We see that X can take on the values 2, 3, ..., 12 and Y can take on the values 0, 1, ..., 5. ////