# 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

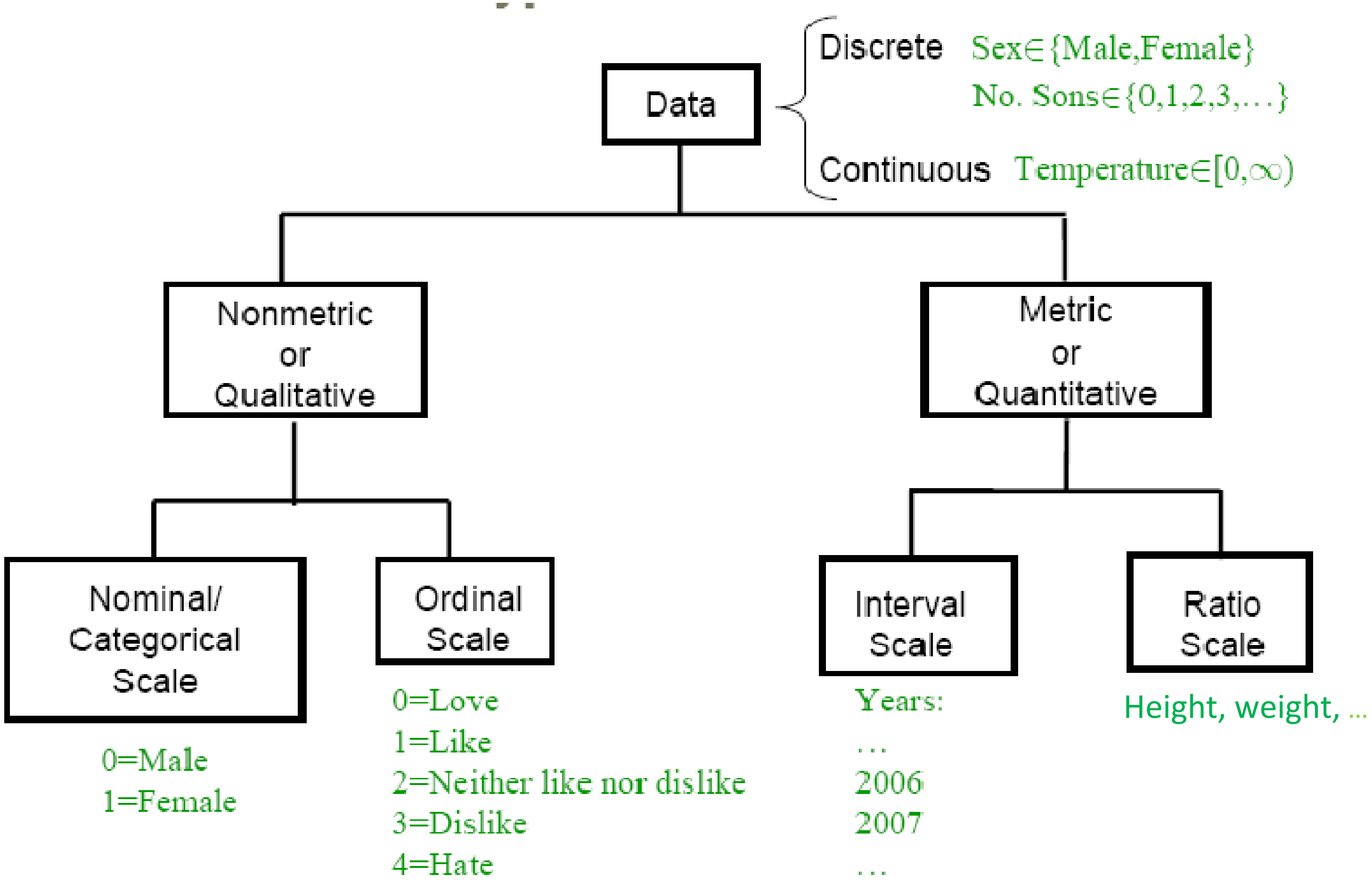
- As in Chapter 1, we would like to model our random experiment so as to be able to give values to the probabilities of events
- Recall that a probability space was defined by a triplet consisting of a
  - o sample space Ω
  - o collection of subsets $\mathscr{A}$ (of events)
  - o set function P[.] 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.

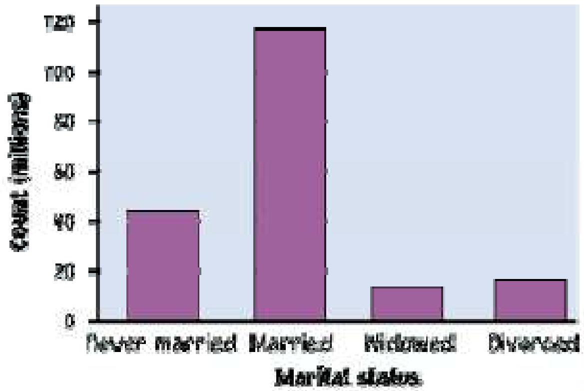
**Quantitative**
Each individual is attributed a numerical value.

| 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 |

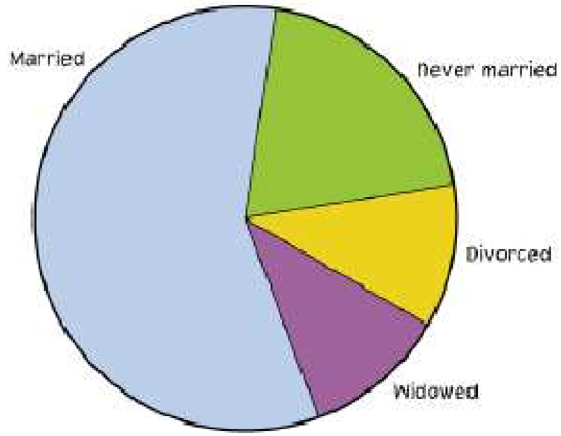
# Ways to chart categorical data

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.



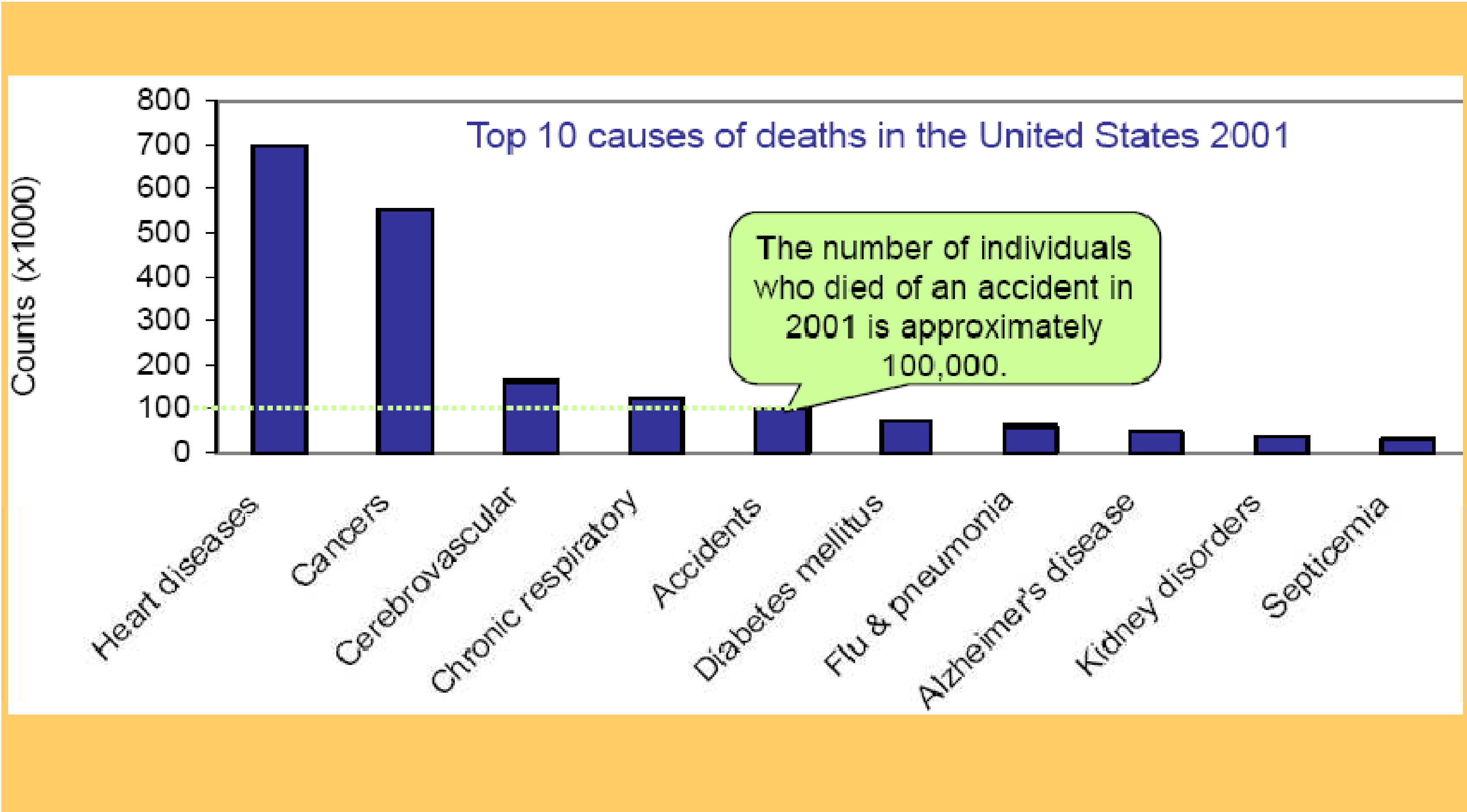
## Example: Top 10 causes of death in the United States 2001

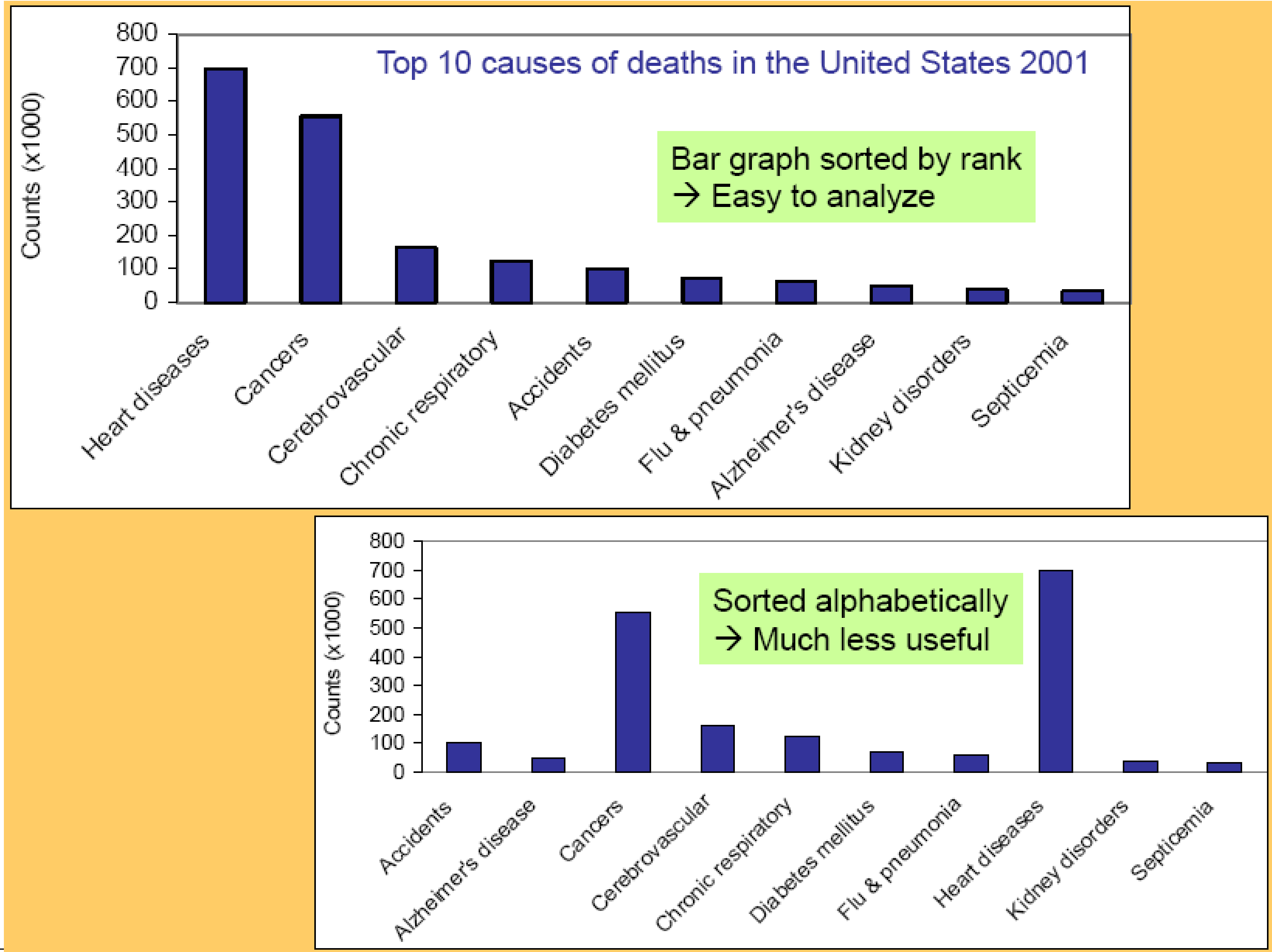
| Rank | Causes of death | Counts | % of top 10s | % of total deaths |
|---|---|---|---|---|
| 1 | Heart disease | 700,142 | 37% | *29%* |
| 2 | Cancer | 553,768 | 29% | *23%* |
| 3 | Cerebrovascular | 163,538 | 9% | *7%* |
| 4 | Chronic respiratory | 123,013 | 6% | *5%* |
| 5 | Accidents | 101,537 | 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

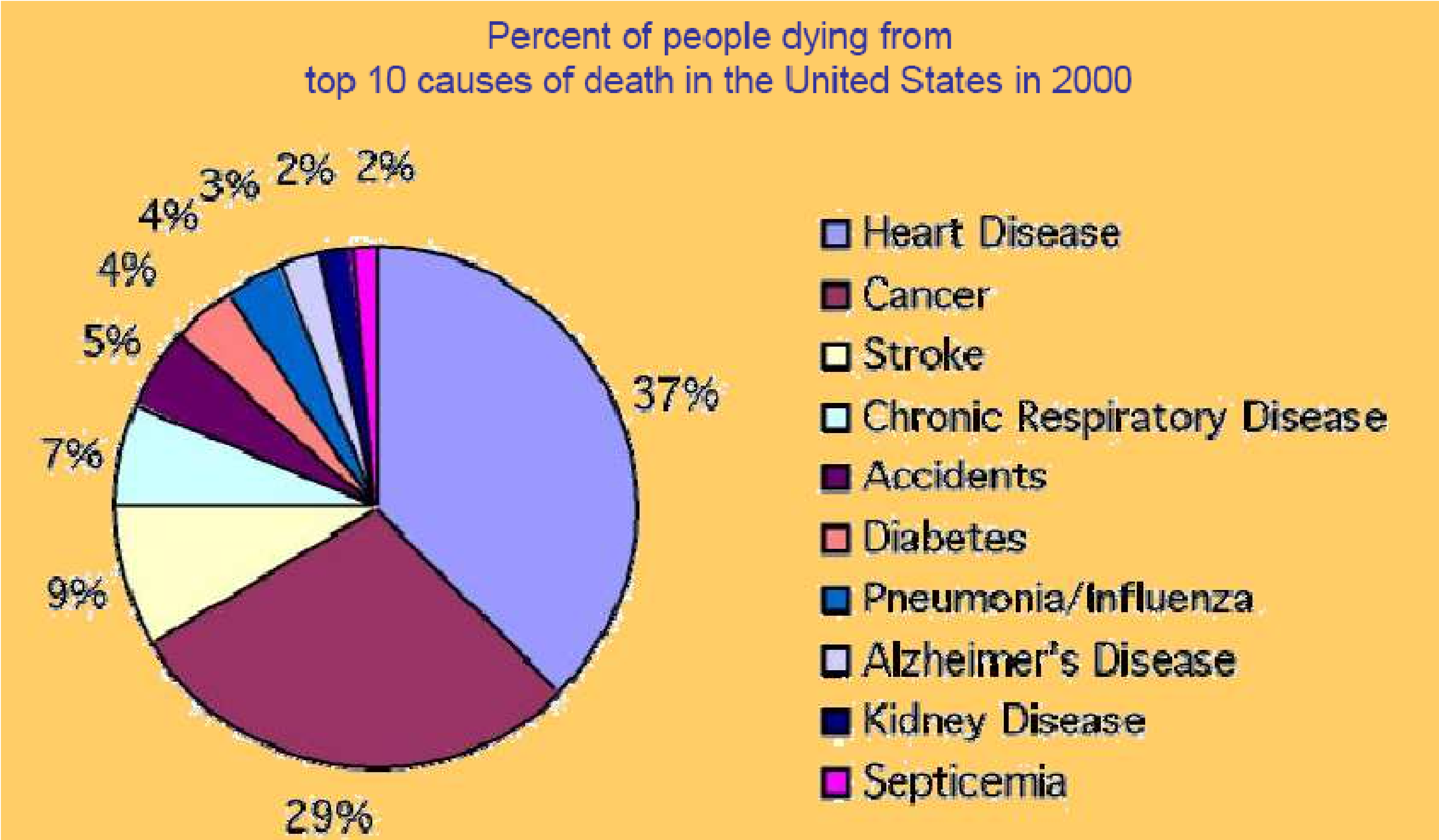
Each category is represented by one bar. The bar's height shows the count (or sometimes the percentage) for that particular category.

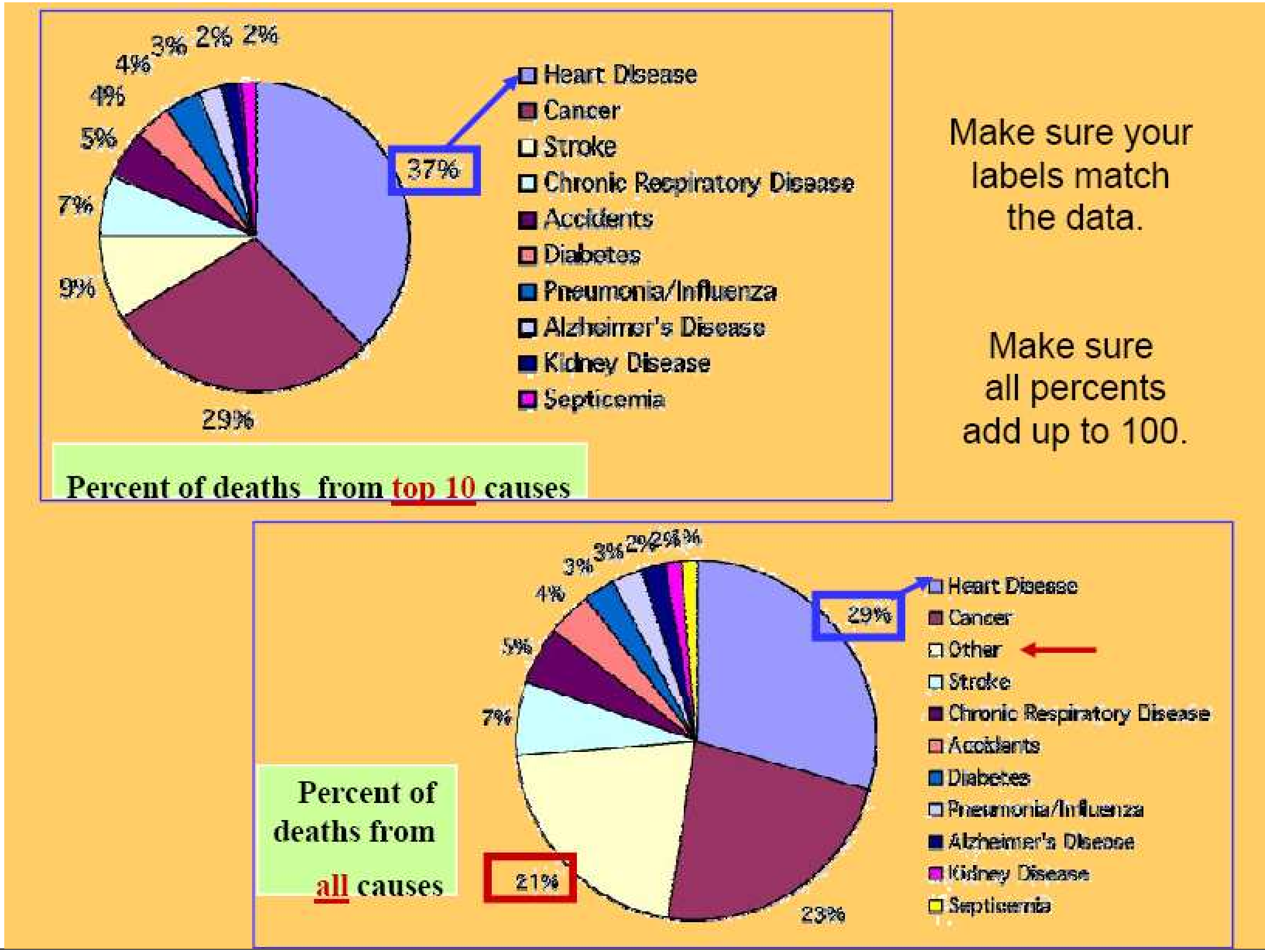



Top 10 causes of deaths in the United States 2001
Counts (x1000)
800
700
600
500
400
300
200
100
0
Heart diseases
Cancers
Cerebrovascular
Chronic respiratory
Accidents
Diabetes mellitus
Flu & pneumonia
Alzheimer's disease
Kidney disorders
Septicemia
Bar graph sorted by rank
→ Easy to analyze
Sorted alphabetically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

# 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.
3%
3%
4%
5%
7%
29%
21%
23%
Heart Disease
Cancer
Other
Stroke
Chronic Respiratory Disease
Accidents
Diabetes
Pneumonia/Influenza
Alzheimer's Disease
Kidney Disease
Septicemia
Percent of deaths from all causes

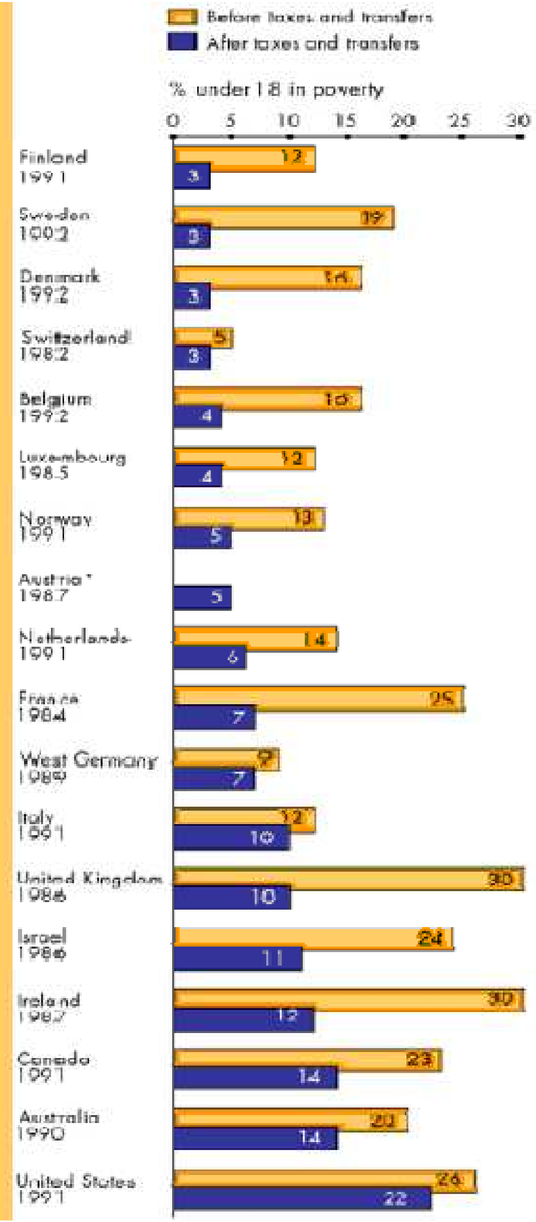
# Child poverty before and after government intervention—UNICEF, 1996

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.

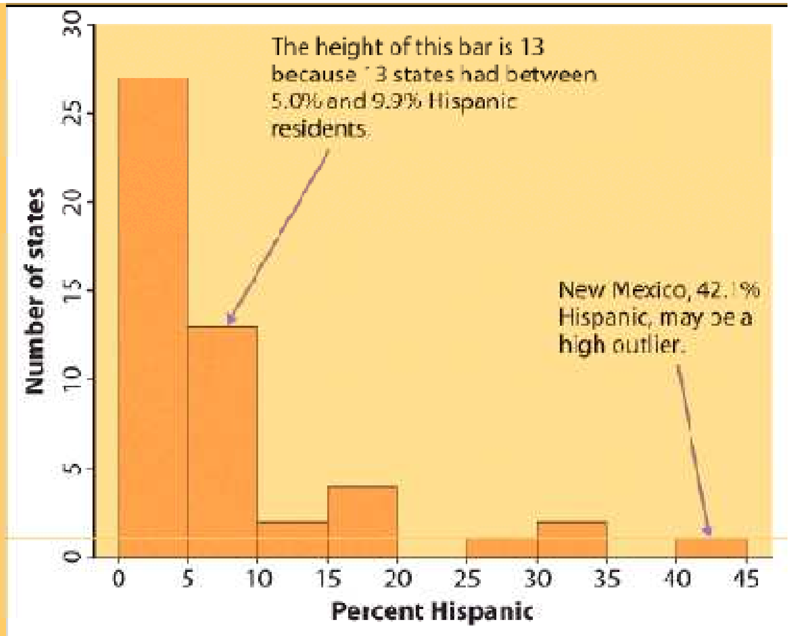
- Line graphs: time plots

  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 |

## Percent of Hispanic residents in each of the 50 states

| 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 |
| RhodeIsland | 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 |
| RhodeIsland | 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

| | |
|---|---|
| 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.

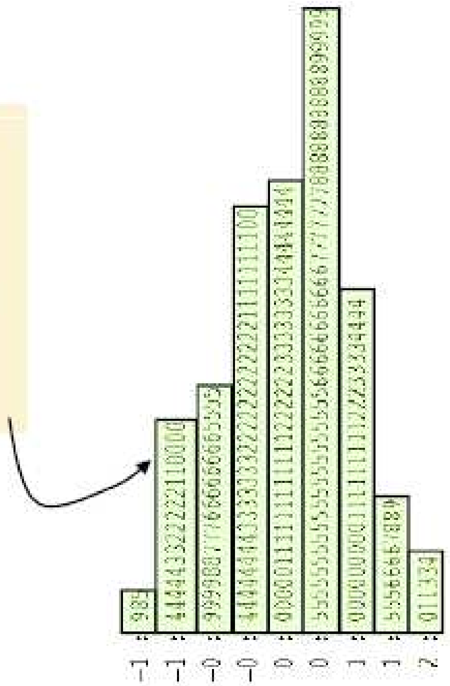
- When the observed values have too many digits, trim the numbers before making a stem plot.

- When plotting a moderate number of observations, you can split each stem.

# Stem plot or histogram?

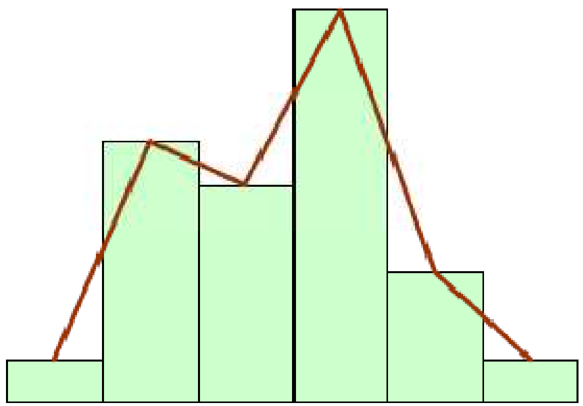
Stemplots are quick and dirty histograms that can easily be done by hand, and therefore are very convenient for back of the envelope calculations. However, they are rarely found in scientific or laymen publications.

```
−1 : 985
−1 : 4444332222211000
−0 : 99998877766666665555
−0 : 444444433333332222222222221111111100
 0 : 0000011111111111122222233333333344444444
 0 : 555555555555555555555556666666666677777778888888888899999
 1 : 000000000111111111222333334444
 1 : 55566667889
 2 : 011334
```

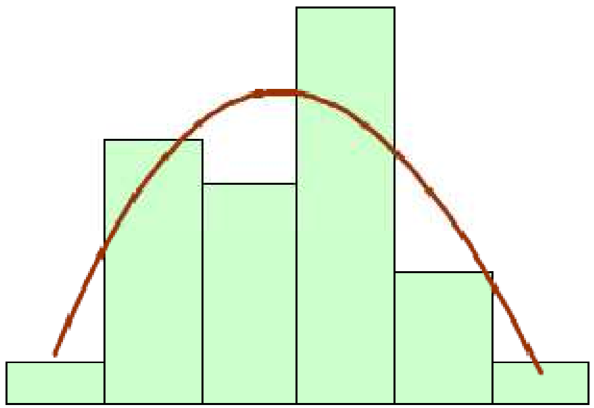


# Interpreting histograms

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.**

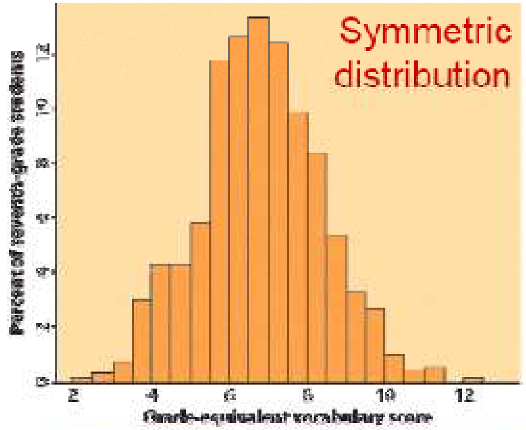


Histogram with a line connecting each column → too detailed



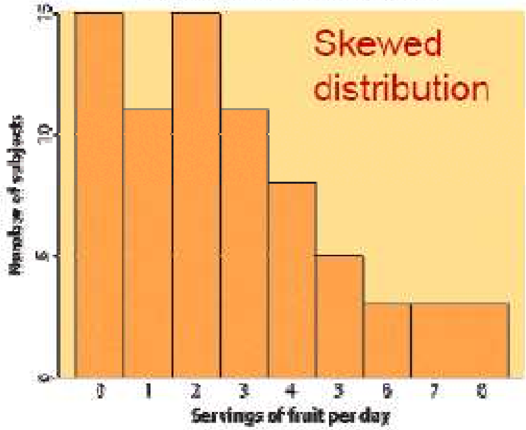
Histogram with a smoothed curve highlighting the overall pattern of the distribution

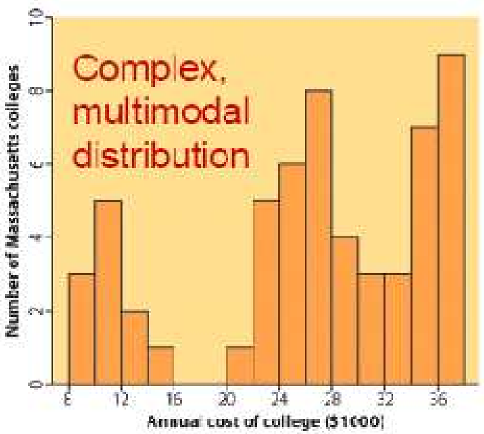
# Most common distribution shapes




- 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.



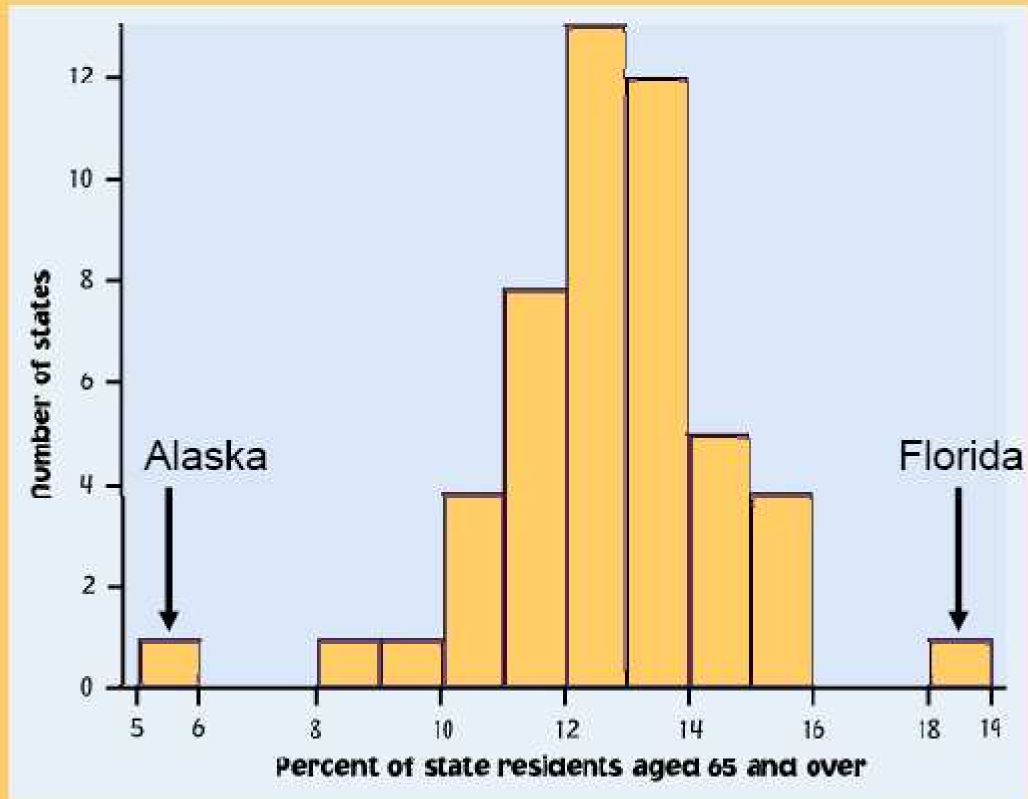

- Not all distributions have a simple overall shape, especially when there are few observations.

# 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.

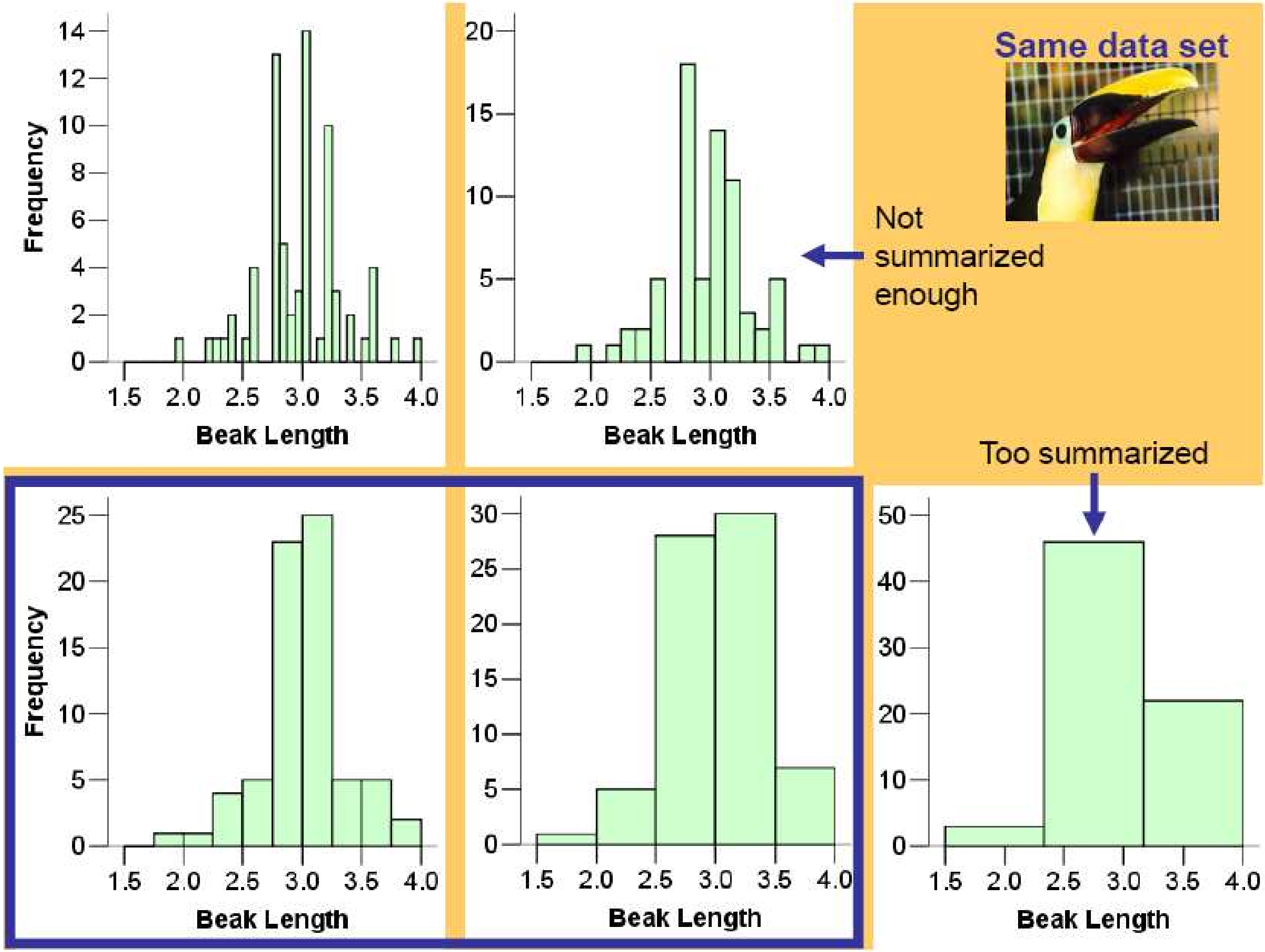
What bin size should you use?

- 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

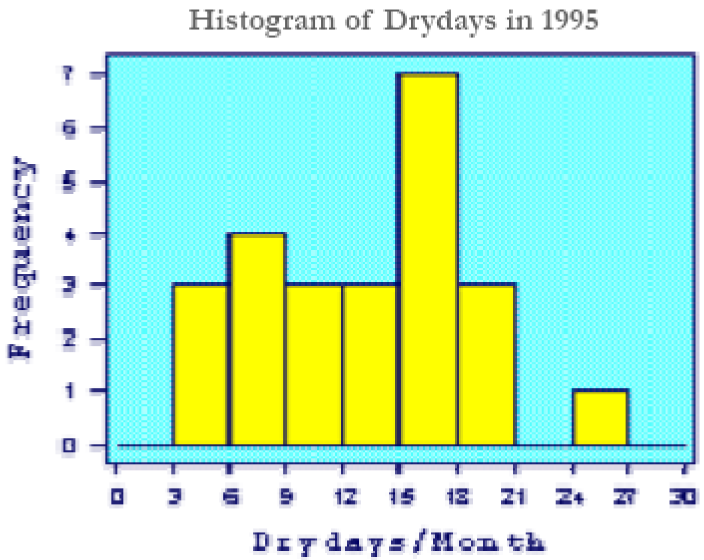
*(There isn't a unique or "perfect" solution)*

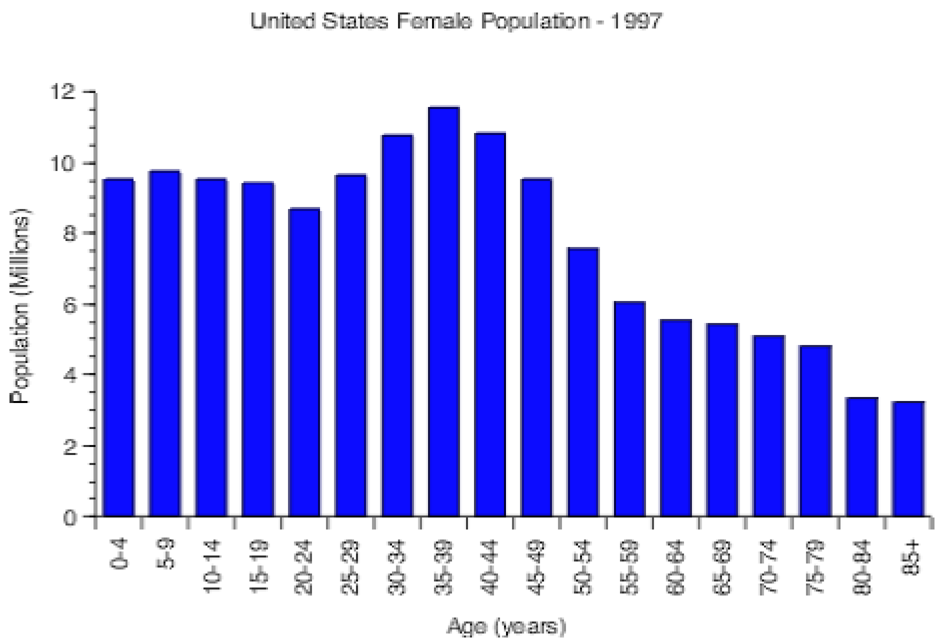

Same data set
Not summarized enough
Too summarized
Frequency
Beak Length
1.5 2.0 2.5 3.0 3.5 4.0

IMPORTANT NOTE:

Your data are the way they are.

Do not try to force them into a particular shape.

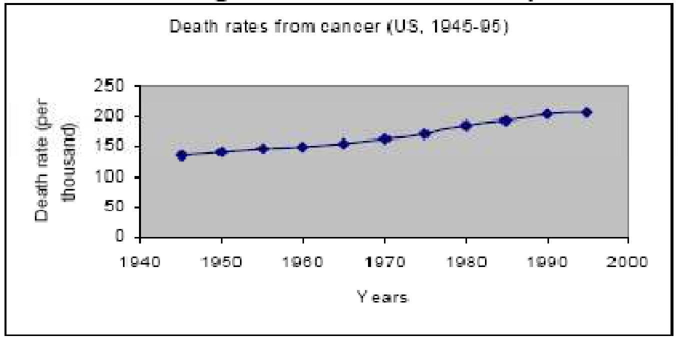


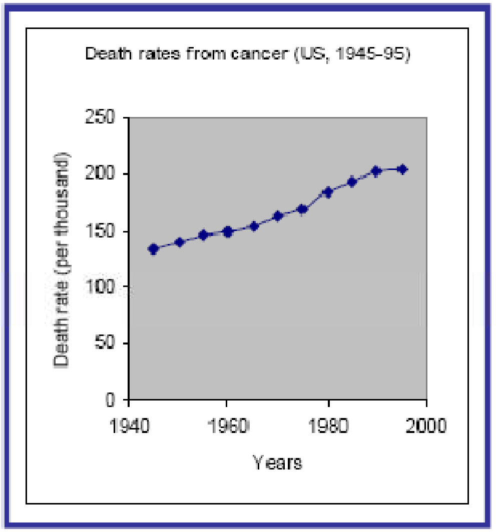



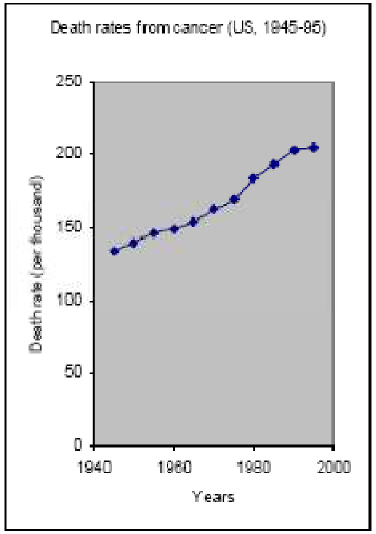

*It is a common misconception that if you have a large enough data set, the data will eventually turn out nice and symmetrical.*

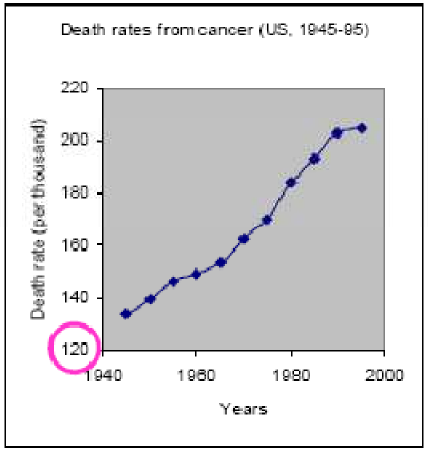
# Cautionary note : scale matters when visualizing data

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

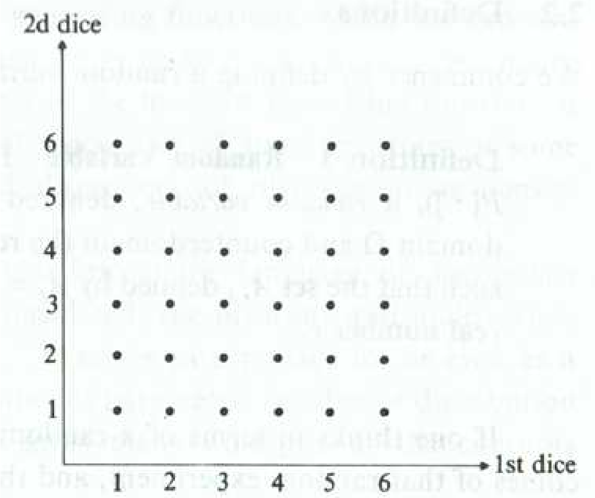
**Definition** **Random Variable** For a given probability space $(\Omega, \mathscr{A}, P[\cdot])$, a *random variable*, denoted by $X$ or $X(\cdot)$, is a function with domain $\Omega$ and counterdomain the real line. The function $X(\cdot)$ must be such that the set $A_r$, defined by $A_r = \{\omega: X(\omega) \leq r\}$, 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 $\{\omega\colon X(\omega) \le r\}$ belongs to $\mathscr{A}$ for every real number $r$. $\mathscr{A}$ consists of the four subsets: $\phi$, $\{\text{head}\}$, $\{\text{tail}\}$, and $\Omega$. Now, if $r < 0$, $\{\omega\colon X(\omega) \le r\} = \phi$; and if $0 \le r < 1$, $\{\omega\colon X(\omega) \le r\} = \{\text{tail}\}$; and if $r \ge 1$, $\{\omega\colon X(\omega) \le r\} = \Omega = \{\text{head, tail}\}$. Hence, for each $r$ the set $\{\omega\colon X(\omega) \le r\}$ belongs to $\mathscr{A}$; so $X(\cdot)$ is a random variable. ////

**EXAMPLE** Consider the experiment of tossing two dice. $\Omega$ can be described by the 36 points displayed in Fig. . $\Omega = \{(i, j): i = 1, \ldots, 6$ 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. ////



# 3 Cumulative distribution functions

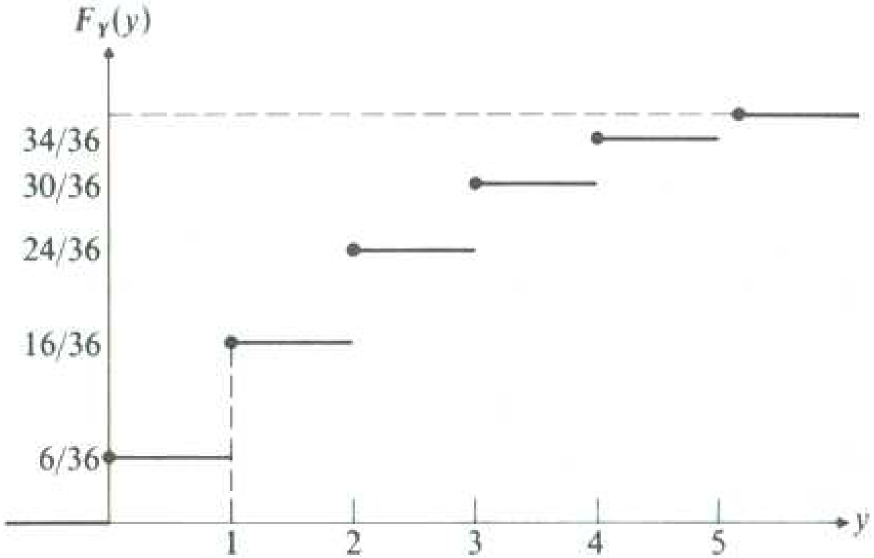
- In both examples above, we described random variables in terms of a random experiment rather than in specifying their functional form; such will usually be the case

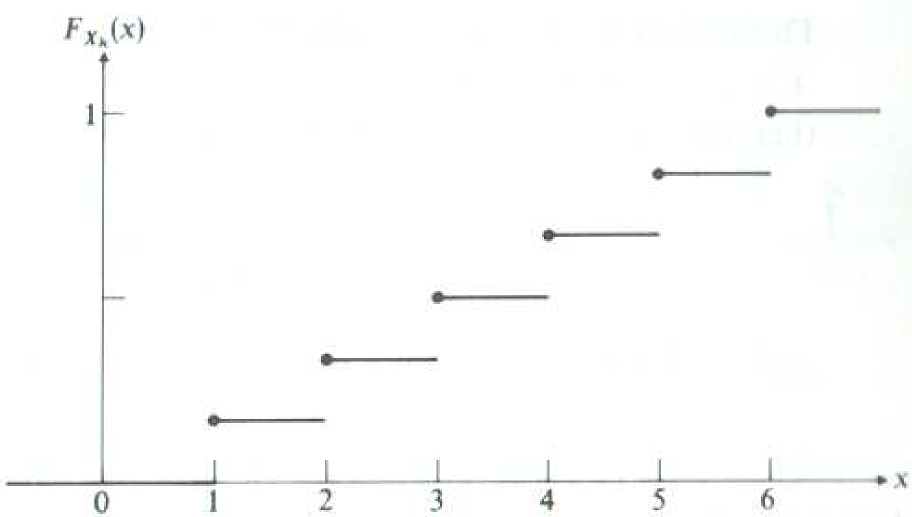
**Definition 2 Cumulative distribution function** The *cumulative distribution function* of a random variable $X$, denoted by $F_X(\cdot)$, is defined to be that function with domain the real line and counterdomain the interval

$[0, 1]$ which satisfies $F_X(x) = P[X \leq x] = P[\{\omega: X(\omega) \leq x\}]$ for every real number $x$. ////

- Note that a cumulative distribution function is uniquely defined for each random variable and that, when it is known, it can be used to find probabilities of events defined in terms of its corresponding random variable.

EXAMPLE 4 In the experiment of tossing two fair dice, let $Y$ denote the absolute difference. The cumulative distribution of $Y$, $F_Y(\cdot)$, is sketched in Fig. A Also, let $X_k$ denote the value on the upturned face of the $k$th die for $k = 1, 2$. $X_1$ and $X_2$ are different random variables, yet both have the same cumulative distribution function, which is $F_{X_k}(x) = \sum_{i=1}^{5} \frac{i}{6} I_{[i, i+1)}(x) + I_{[6, \infty)}(x)$ and is sketched in Fig. B ////



**Properties of a Cumulative Distribution Function $F_X(\cdot)$**

(i) $F_X(-\infty) \equiv \lim_{x \to -\infty} F_X(x) = 0$, and $F_X(+\infty) \equiv \lim_{x \to +\infty} F_X(x) = 1$.

(ii) $F_X(\cdot)$ is a monotone, nondecreasing function; that is, $F_X(a) \le F_X(b)$ for $a < b$.

(iii) $F_X(\cdot)$ is continuous from the right; that is,

$$\lim_{0 < h \to 0} F_X(x + h) = F_X(x).$$

Except for (ii), we will not prove these properties. Note that the event $\{\omega\colon X(\omega) \le b\} = \{X \le b\} = \{X \le a\} \cup \{a < X \le b\}$ and $\{X \le a\} \cap \{a < X \le b\} = \phi$; hence, $F_X(b) = P[X \le b] = P[X \le a] + P[a < X \le b] \ge P[X \le a] = F_X(a)$ which proves (ii). Property (iii), the continuity of $F_X(\cdot)$ from the right, results from our defining $F_X(x)$ to be $P[X \le x]$. If we had defined, as some authors do, $F_X(x)$ to be $P[X < x]$, then $F_X(\cdot)$ would have been continuous from the left.

## Simplified definition for cumulative distribution function

- The aforementioned properties of a cumulative distribution function are important in that they lead to a simplified definition of a cumulative distribution function, avoiding the mentioning of random variables

**Definition** **Cumulative distribution function** Any function $F(\cdot)$ with domain the real line and counterdomain the interval [0, 1] satisfying the above three properties is defined to be a *cumulative distribution function*. ////

# 4 Density functions

## 4.1 Discrete random variables

### Definitions

**Definition Discrete random variable** A random variable $X$ will be defined to be *discrete* if the range of $X$ is countable. If a random variable $X$ is discrete, then its corresponding cumulative distribution function $F_X(\cdot)$ will be defined to be *discrete*. ////

By the range of $X$ being countable we mean that there exists a finite or denumerable set of real numbers, say $x_1, x_2, x_3, \ldots$, such that $X$ takes on values only in that set. If $X$ is discrete with distinct values $x_1, x_2, \ldots, x_n, \ldots$, then $\Omega = \bigcup_n \{\omega\colon\ X(\omega) = x_n\} = \bigcup_n \{X = x_n\}$, and $\{X = x_i\} \cap \{X = x_j\} = \phi$ for $i \neq j$; hence $1 = P[\Omega] = \sum_n P[X = x_n]$ by the third axiom of probability.

**Definition** **Discrete density function of a discrete random variable** If $X$ is a discrete random variable with distinct values $x_1, x_2, \ldots, x_n, \ldots$, then the function, denoted by $f_X(\cdot)$ and defined by

$$f_X(x) = \begin{cases} P[X = x_j] & \text{if } x = x_j, j = 1, 2, \ldots, n, \ldots \\ 0 & \text{if } x \neq x_j \end{cases} \tag{1}$$

is defined to be the *discrete density function* of $X$. ////

## Synonyms for discrete density function

The values of a discrete random variable are often called *mass points*; and, $f_X(x_j)$ denotes the *mass* associated with the *mass point* $x_j$. *Probability mass function*, *discrete frequency function*, and *probability function* are other terms used in place of *discrete density function*. Also, the notation $p_X(\cdot)$ is sometimes used intead of $f_X(\cdot)$ for discrete density functions. $f_X(\cdot)$ is a function with domain the real line and counterdomain the interval [0, 1]. If we use the indicator function,

$$f_X(x) = \sum_{n=1}^{\infty} P[X = x_n] I_{\{x_n\}}(x),$$

where $I_{\{x_n\}}(x) = 1$ if $x = x_n$ and $I_{\{x_n\}}(x) = 0$ if $x \neq x_n$.

# Relation between density and cumulative distribution functions

**Theorem 1** Let $X$ be a discrete random variable. $F_X(\cdot)$ can be obtained from $f_X(\cdot)$, and vice versa.

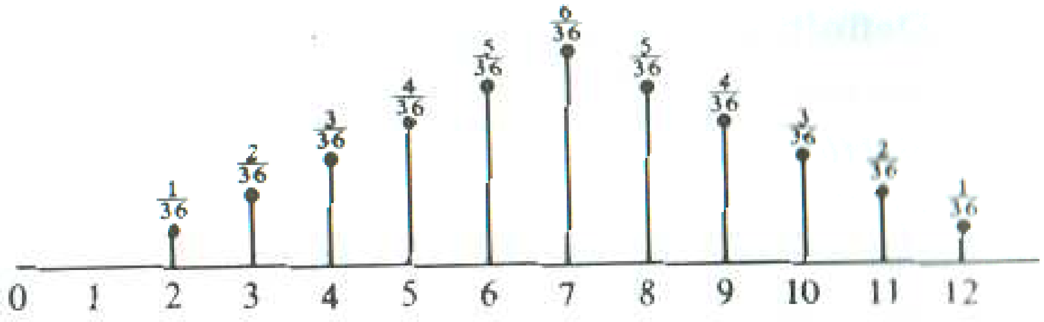
PROOF Denote the mass points of $X$ by $x_1, x_2, \ldots$. Suppose $f_X(\cdot)$ is given; then $F_X(x) = \sum_{\{j: x_j \leq x\}} f_X(x_j)$. Conversely, suppose $F_X(\cdot)$ is given; then $f_X(x_j) = F_X(x_j) - \lim_{0<h\to 0} F_X(x_j - h)$; hence $f_X(x_j)$ can be found for each mass point $x_j$; however, $f_X(x) = 0$ for $x \neq x_j$, $j = 1, 2, \ldots$, so $f_X(x)$ is determined for all real numbers. ////

EXAMPLE To illustrate what is meant in Theorem 1, consider the experiment of tossing a single die. Let $X$ denote the number of spots on the upper face:

$$f_X(x) = \left(\frac{1}{6}\right) I_{\{1, 2, \dots, 6\}}(x),$$

and

$$F_X(x) = \sum_{i=1}^{5} (i/6) I_{[i, i+1)}(x) + I_{[6, \infty)}(x).$$




According to Theorem 1, for given $f_X(\cdot)$, $F_X(x)$ can be found for any $x$; for instance, if $x = 2.5$,

$$F_X(2.5) = \sum_{\{j:\, x_j \le 2.5\}} f_X(x_j) = f_X(1) + f_X(2) = \frac{2}{6}.$$

And, if $F_X(\cdot)$ is given, $f_X(x)$ can be found for any $x$. For example, for $x = 3$,

$$f_X(3) = F_X(3) - \lim_{0<h\to 0} F_X(3-h) = \left(\frac{3}{6}\right) - \left(\frac{2}{6}\right) = \frac{1}{6}. \qquad ////$$

## Simplified definition for discrete density function

**Definition Discrete density function** Any function $f(\cdot)$ with domain the real line and counterdomain [0, 1] is defined to be a *discrete density function* if for some countable set $x_1, x_2, \ldots, x_n, \ldots,$

(i) $f(x_j) > 0$ for $j = 1, 2, \ldots.$

(ii) $f(x) = 0$ for $x \neq x_j; j = 1, 2, \ldots.$

(iii) $\sum f(x_j) = 1$, where the summation is over the points $x_1, x_2, \ldots, x_n, \ldots.$ ////

- This definition allows us to speak about discrete density functions without reference to some random variable.
- We can therefore talk about properties of discrete density functions without referring to a random variable

# 4.2 Continuous

## Definitions

**Definition** **Continuous random variable** A random variable $X$ is called *continuous* if there exists a function $f_X(\cdot)$ such that $F_X(x) = \int_{-\infty}^{x} f_X(u)du$ for every real number $x$. The cumulative distribution function $F_X(\cdot)$ of a continuous random variable $X$ is called *absolutely continuous*. ////

**Definition** **Probability density function of a continuous random variable** If $X$ is a continuous random variable, the function $f_X(\cdot)$ in $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$ is called the *probability density function* of $X$. ////

## Synonyms for probability density function

- Other names instead of probability density function include density function, continuous density function, integrating density function, …
- Strictly speaking, we should speak of **A** probability density function instead of **THE** probability density function
- Indeed, all the definition requires is that the integral of $f$ gives $F$ for every $x$, and more than one function $f$ may satisfy such a requirement

# Relation between density and cumulative distribution functions

**Theorem 2** Let $X$ be a continuous random variable. Then $F_X(\cdot)$ can be obtained from an $f_X(\cdot)$, and vice versa.

PROOF If $X$ is a continuous random variable and an $f_X(\cdot)$ is given, then $F_X(x)$ is obtained by integrating $f_X(\cdot)$; that is, $F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$. On the other hand, if $F_X(\cdot)$ is given, then an $f_X(x)$ can be obtained by differentiation; that is, $f_X(x) = dF_X(x)/dx$ for those points $x$ for which $F_X(x)$ is differentiable. ////

# Simplified definition for continuous density function

**Definition** **Probability density function** Any function $f(\cdot)$ with domain the real line and counterdomain $[0, \infty)$ is defined to be a *probability density function* if and only if

(i) $f(x) \geq 0$ for all $x$.

(ii) $\int_{-\infty}^{\infty} f(x)\, dx = 1$. ////

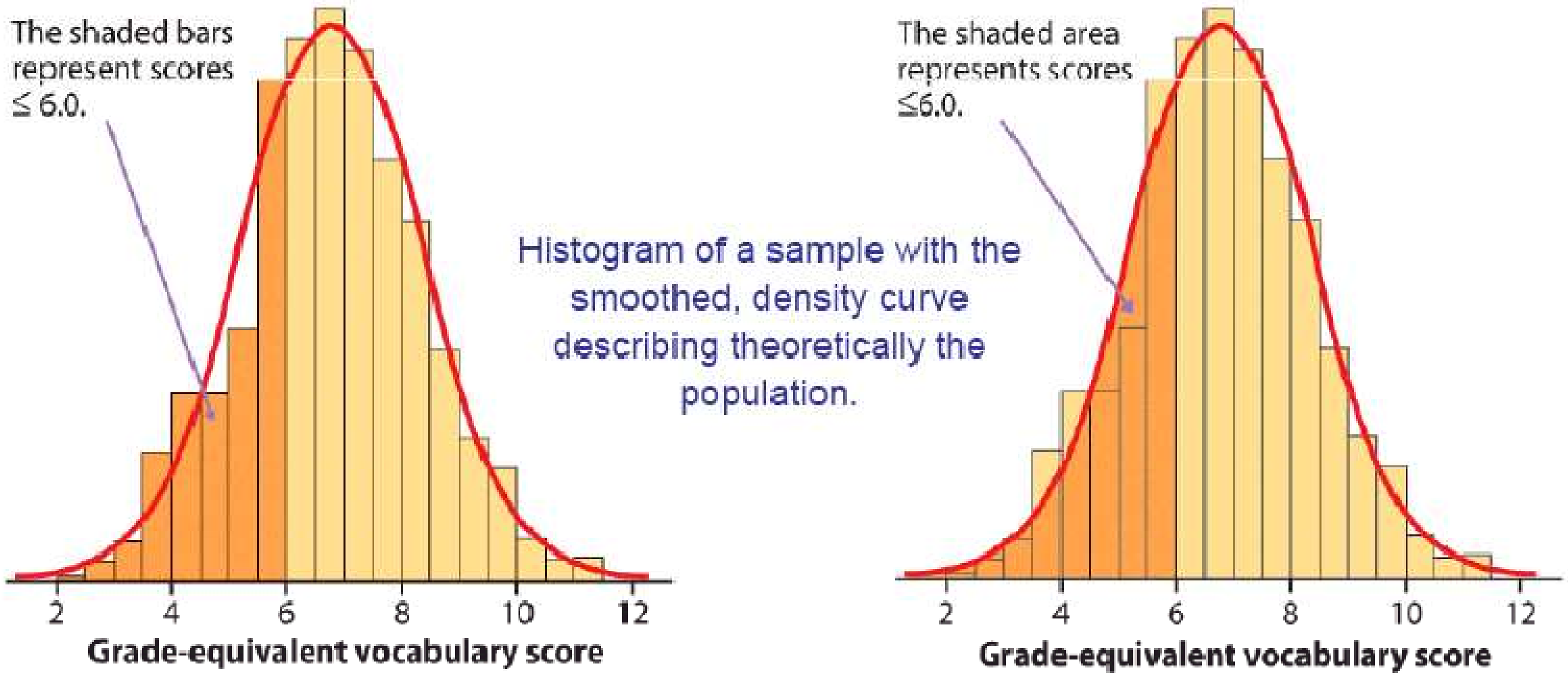
- With this definition , we can speak of probability density functions without reference to random variables

# Density curves

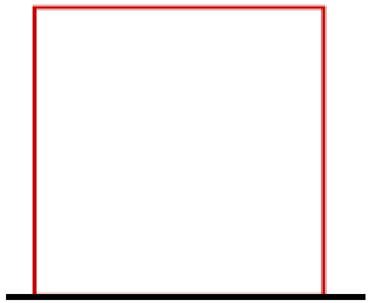
A **density curve** is a mathematical model of a distribution.

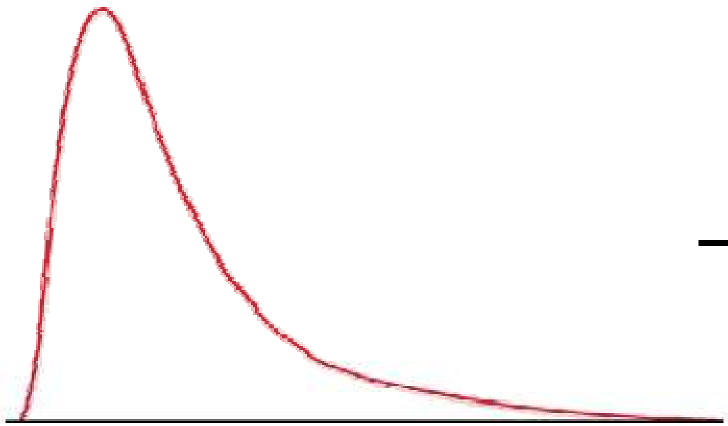
The total area under the curve, by definition, is equal to 1, or 100%.

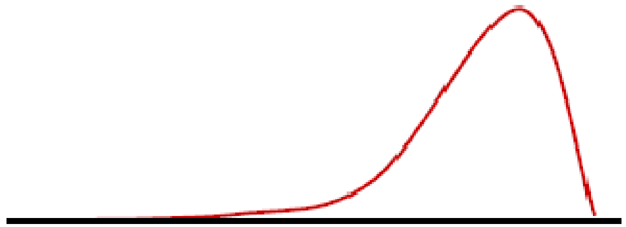
The area under the curve for a range of values is the proportion of all observations for that range.



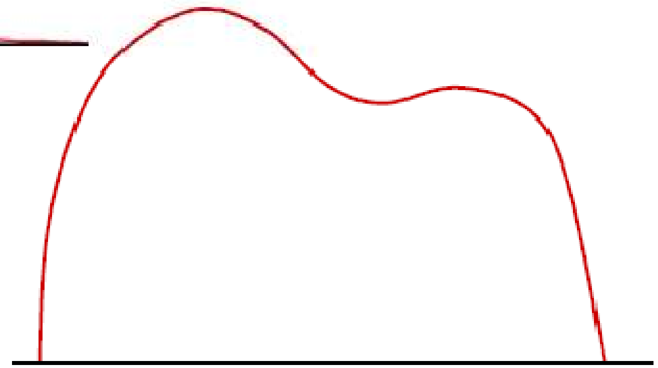
Density curves come in any imaginable shape.





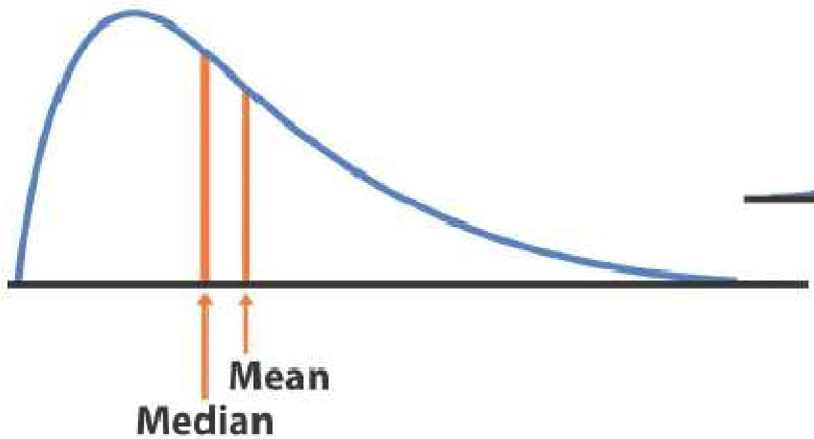


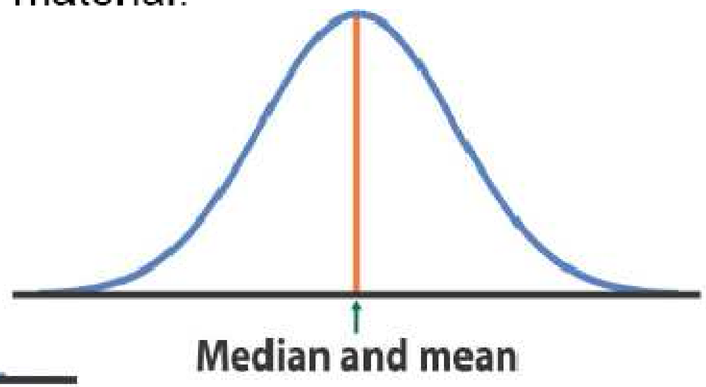
Some are well known mathematically and others aren't.



The **median** of a density curve is the equal-areas point: the point that divides the area under the curve in half.

The **mean** of a density curve is the balance point, at which the curve would balance if it were made of solid material.



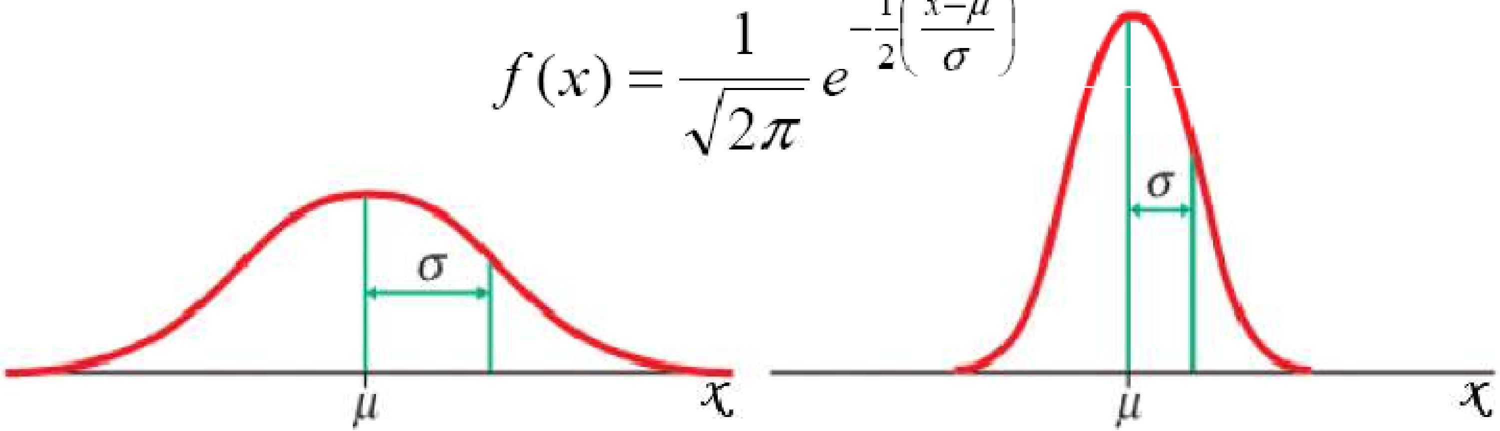

The median and mean are the same for a symmetric density curve.

The mean of a skewed curve is pulled in the direction of the long tail.

# Example: normal distribution

Normal – or Gaussian – distributions are a family of symmetrical, bell-shaped density curves defined by a mean μ (*mu*) and a standard deviation σ (*sigma*) : N(μ,σ).

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$




*e = 2.71828… The base of the natural logarithm*

*π = pi = 3.14159…*

# General remark

The notations for discrete density function and probability density function are the same, yet they have quite different interpretations. For discrete random variables $f_X(x) = P[X = x]$, which is not true for continuous random variables. For continuous random variables,

$$f_X(x) = \frac{dF_X(x)}{dx} = \lim_{\Delta x \to 0} \frac{F_X(x + \Delta x) - F_X(x - \Delta x)}{2\Delta x};$$

hence $f_X(x)2\Delta x \approx F_X(x + \Delta x) - F_X(x - \Delta x) = P[x - \Delta x < X \leq x + \Delta x]$; that is, the probability that $X$ is in a *small* interval containing the value $x$ is approximately equal to $f_X(x)$ times the width of the interval. For discrete random variables $f_X(\cdot)$ is a function with domain the real line and counterdomain the interval $[0, 1]$; whereas, for continuous random variables $f_X(\cdot)$ is a function with domain the real line and counterdomain the infinite interval $[0, \infty)$.

# 5 A gentle introduction to "moments"

## 5.1 Mean of a random variable

The mean $\mu$ of a random variable $X$ is a weighted average of the possible values of $X$, reflecting the fact that all outcomes might not be equally likely.

A basketball player shoots three free throws. The random variable $X$ is the number of baskets successfully made ("H").

| | | | |
|---|---|---|---|
| | HMM | HHM | |
| | MHM | HMH | |
| MMM | MMH | MHH | HHH |

| Value of $X$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Probability | 1/8 | 3/8 | 3/8 | 1/8 |

The mean of a random variable $X$ is also called **expected value** of $X$.

**Definition** **Mean** Let $X$ be a random variable. The *mean* of $X$, denoted by $\mu_X$ or $\mathscr{E}[X]$, is defined by:

(i)
$$\mathscr{E}[X] = \sum x_j f_X(x_j)$$

if $X$ is discrete with mass points $x_1, x_2, \ldots, x_j, \ldots$.

(ii)
$$\mathscr{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

if $X$ is continuous with probability density function $f_X(x)$.

(iii)
$$\mathscr{E}[X] = \int_0^{\infty} [1 - F_X(x)]\, dx - \int_{-\infty}^{0} F_X(x)\, dx$$

for an arbitrary random variable $X$. ////

# Explanation

Note what the definition says: In $\sum_j x_j f_X(x_j)$, the summand is the $j$th value of the random variable $X$ multiplied by the probability that $X$ equals that $j$th value, and then the summation is over all values. So $\mathscr{E}[X]$ is an "average" of the values that the random variable takes on, where each value is weighted by the probability that the random variable is equal to that value. Values that are more probable receive more weight. The same is true in integral form in (ii). There the value $x$ is multiplied by the approximate probability that $X$ equals the value $x$, namely $f_X(x)\,dx$, and then integrated over all values.

## Remarks

- When the indicated series do not converge, or the integrals do not exist, then the mean does not exist either
- The mean of a random variable is a **measure of central location** of the density of X

> $\mathscr{E}[X]$ is the center of gravity (or *centroid*) of the unit mass that is determined by the density function of $X$. So the mean of $X$ is a measure of where the values of the random variable $X$ are "centered."

- It can be shown that (i) and (ii) follow from (iii) in the discrete case and continuous case resp. The main use of (iii) is in those settings where we would like to find the mean of a random variable that is neither discrete nor continuous.

# Examples

For a discrete random variable $X$ with probability distribution →

| Value of $X$ | $x_1$ | $x_2$ | $x_3$ | $\cdots$ | $x_k$ |
|---|---|---|---|---|---|
| Probability | $p_1$ | $p_2$ | $p_3$ | $\cdots$ | $p_k$ |

the mean $\mu$ of $X$ is found by multiplying each possible value of $X$ by its probability, and then adding the products.

$$\mu_X = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$
$$= \sum x_i p_i$$

A basketball player shoots three free throws. The random variable $X$ is the number of baskets successfully made.

| Value of $X$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Probability | 1/8 | 3/8 | 3/8 | 1/8 |

The mean $\mu$ of $X$ is

$\mu$ = (0*1/8) + (1*3/8) + (2*3/8) + (3*1/8)

= 12/8 = 3/2 = 1.5

EXAMPLE 10 Consider the experiment of tossing two dice. Let $X$ denote the total of the two dice and $Y$ their absolute difference. The discrete density functions for $X$ and $Y$ are given in Example 6.

$$\mathscr{E}[Y] = \sum y_j f_Y(y_j) = \sum_{i=0}^{5} i f_Y(i) = 0 \cdot \tfrac{6}{36} + 1 \cdot \tfrac{10}{36}$$
$$+ 2 \cdot \tfrac{8}{36} + 3 \cdot \tfrac{6}{36} + 4 \cdot \tfrac{4}{36} + 5 \cdot \tfrac{2}{36} = \tfrac{70}{36}.$$
$$\mathscr{E}[X] = \sum_{i=2}^{12} i f_X(i) = 7.$$

Note that $\mathscr{E}[Y]$ is not one of the possible values of $Y$. ////

EXAMPLE __ Let $X$ be a continuous random variable with probability density function $f_X(x) = \lambda e^{-\lambda x} I_{[0, \infty)}(x)$.

$$\mathscr{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_0^{\infty} x \lambda e^{-\lambda x}\, dx = \frac{1}{\lambda}.$$

The corresponding cumulative distribution function is

$$F_X(x) = (1 - e^{-\lambda x}) I_{[0, \infty)}(x); \text{ so } \mathscr{E}[X] = \int_0^{\infty} [1 - F_X(x)]\, dx$$
$$- \int_{-\infty}^{0} F_X(x)\, dx = \int_0^{\infty} (1 - 1 + e^{-\lambda x})\, dx = 1/\lambda.$$ ////

## 5.2 Variance of a random variable

**Definition** **Variance** Let $X$ be a random variable, and let $\mu_X$ be $\mathscr{E}[X]$. The *variance of* $X$, denoted by $\sigma_X^2$ or var $[X]$, is defined by

(i) $$\text{var}\,[X] = \sum_j (x_j - \mu_X)^2 f_X(x_j)$$

if $X$ is discrete with mass points $x_1, x_2, \ldots, x_j, \ldots$.

(ii) $$\text{var}\,[X] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\,dx$$

if $X$ is continuous with probability density function $f_X(x)$.

(iii) $$\text{var}\,[X] = \int_0^{\infty} 2x[1 - F_X(x) + F_X(-x)]\,dx - \mu_X^2$$

for an arbitrary random variable $X$. ////

# Explanation

Note what the definition says: In (i), the square of the difference between the $j$th value of the random variable $X$ and the mean of $X$ is multiplied by the probability that $X$ equals the $j$th value, and then these terms are summed. More weight is assigned to the more probable squared differences. A similar comment applies for (ii). Variance is a *measure of spread* since if the values of a random variable $X$ tend to be far from their mean, the variance of $X$ will be larger than the variance of a comparable random variable $Y$ whose values tend to be near their mean.

## Remarks

- When the indicated series do not converge, or the integrals do not exist, then the mean does not exist either
- The variance of a random variable is a **measure of spread** or **dispersion** of the density of X

Whereas the mean can be interpreted as the center of gravity of a density, the variance can be seen as the moment of inertia of the same density with respect to a perpendicular axis through the center of gravity (elementary physics, mechanics)

- It can be shown that (i) and (ii) follow from (iii) in the discrete case and continuous case resp. The main use of (iii) is in those settings where we would like to find the variance of a random variable that is neither discrete nor continuous.

## Standard deviation

**Definition** **Standard deviation** If $X$ is a random variable, the *standard deviation* of $X$, denoted by $\sigma_X$, is defined as $+\sqrt{\text{var}[X]}$. ////

- The standard deviation of a random variable, like the variance, is a **measure of the spread or dispersion** of the values of a random variable.
- In many applications it is preferable to the variance as such a measure since it will have the same measurement units as the random variable itself.

# Examples

For a discrete random variable $X$ with probability distribution →

| Value of $X$ | $x_1$ | $x_2$ | $x_3$ | $\cdots$ | $x_k$ |
|---|---|---|---|---|---|
| Probability | $p_1$ | $p_2$ | $p_3$ | $\cdots$ | $p_k$ |

and mean $\mu_X$, the variance $\sigma^2$ of $X$ is found by multiplying each squared deviation of $X$ by its probability and then adding all the products.

$$\sigma_X^2 = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \cdots + (x_k - \mu_X)^2 p_k$$
$$= \sum (x_i - \mu_X)^2 p_i$$

A basketball player shoots three free throws. The random variable $X$ is the number of baskets successfully made.

| Value of X | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Probability | 1/8 | 3/8 | 3/8 | 1/8 |

$\mu_X = 1.5$.

The variance $\sigma^2$ of $X$ is

$\sigma^2$ = 1/8*(0−1.5)$^2$ + 3/8*(1−1.5)$^2$ + 3/8*(2−1.5)$^2$ + 1/8*(3−1.5)$^2$

= 2*(1/8*9/4) + 2*(3/8*1/4) = 24/32 = 3/4 = .75

**EXAMPLE** Let $X$ be the total of the two dice in the experiment of tossing two dice.

$$\text{var}[X] = \sum (x_j - \mu_X)^2 f_X(x_j)$$
$$= (2-7)^2 \tfrac{1}{36} + (3-7)^2 \tfrac{2}{36} + (4-7)^2 \tfrac{3}{36} + (5-7)^2 \tfrac{4}{36}$$
$$+ (6-7)^2 \tfrac{5}{36} + (7-7)^2 \tfrac{6}{36} + (8-7)^2 \tfrac{5}{36} + (9-7)^2 \tfrac{4}{36}$$
$$+ (10-7)^2 \tfrac{3}{36} + (11-7)^2 \tfrac{2}{36} + (12-7)^2 \tfrac{1}{36} = \tfrac{210}{36}. \qquad ////$$

**EXAMPLE** Let $X$ be a random variable with probability density given by $f_X(x) = \lambda e^{-\lambda x} I_{[0,\infty)}(x)$; then

$$\text{Var}[X] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\,dx$$
$$= \int_0^{\infty} \left(x - \frac{1}{\lambda}\right)^2 \lambda e^{-\lambda x}\,dx$$
$$= \frac{1}{\lambda^2}. \qquad ////$$

# 5.3 Rules for means and variances

If $X$ is a random variable and $a$ and $b$ are fixed numbers, then

$$\mu_{a+bX} = a + b\mu_X$$

$$\sigma^2_{a+bX} = b^2\sigma^2_X$$

If $X$ and $Y$ are two independent random variables, then

$$\mu_{X+Y} = \mu_X + \mu_Y$$

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$

If $X$ and $Y$ are NOT independent but have correlation $\rho$, then

$$\mu_{X+Y} = \mu_X + \mu_Y$$

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\rho\sigma_X\sigma_Y$$

# Expected value of a function of a random variable

**Definition** **Expectation** Let $X$ be a random variable and $g(\cdot)$ be a function with both domain and counterdomain the real line. The *expectation* or *expected value* of the function $g(\cdot)$ of the random variable $X$, denoted by $\mathscr{E}[g(X)]$, is defined by:

(i) $$\mathscr{E}[g(X)] = \sum_j g(x_j) f_X(x_j)$$

if $X$ is discrete with mass points $x_1, x_2, \ldots, x_j, \ldots$ (provided this series is absolutely convergent).

(ii) $$\mathscr{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

if $X$ is continuous with probability density function $f_X(x)$ (provided $\int_{-\infty}^{\infty} |g(x)| f_X(x)\, dx < \infty$). ////

# General rules for means

**Theorem** Below are properties of expected value:

(i) $\mathscr{E}[c] = c$ for a constant $c$.

(ii) $\mathscr{E}[cg(X)] = c\mathscr{E}[g(X)]$ for a constant $c$.

(iii) $\mathscr{E}[c_1 g_1(X) + c_2 g_2(X)] = c_1 \mathscr{E}[g_1(X)] + c_2 \mathscr{E}[g_2(X)]$.

(iv) $\mathscr{E}[g_1(X)] \leq \mathscr{E}[g_2(X)]$ if $g_1(x) \leq g_2(x)$ for all $x$.

(similarly for variances)

# Example

You invest 20% of your funds in Treasury bills and 80% in an "index fund" that represents all U.S. common stocks. Your rate of return over time is proportional to that of the T-bills ($X$) and of the index fund ($Y$), such that $R = 0.2X + 0.8Y$.

Based on annual returns between 1950 and 2003:

- Annual return on T-bills $\mu X = 5.0\%$ $\sigma X = 2.9\%$
- Annual return on stocks $\mu Y = 13.2\%$ $\sigma Y = 17.6\%$
- Correlation between $X$ and $Y \rho = -0.11$

$\mu_R = 0.2\mu_X + 0.8\mu_Y = (0.2*5) + (0.8*13.2) = 11.56\%$

$\sigma^2{}_R = \sigma^2{}_{0.2X} + \sigma^2{}_{0.8Y} + 2\rho\sigma_{0.2X}\sigma_{0.8Y}$

$= 0.2*2\sigma^2{}_X + 0.8*2\sigma^2{}_Y + 2\rho*0.2*\sigma_X*0.8*\sigma_Y$

$= (0.2)2(2.9)2 + (0.8)2(17.6)2 + (2)(-0.11)(0.2*2.9)(0.8*17.6) = 196.786$

$\sigma_R = \sqrt{196.786} = 14.03\%$

The portfolio has a smaller mean return than an all-stock portfolio, but it is also less risky.

# Remark 1: means and variance are expectations

Expectation or expected value is not really a very good name since it is not necessarily what you "expect." For example, the expected value of a discrete random variable is not necessarily one of the possible values of the discrete random variable, in which case, you would not "expect" to get the expected value. A better name might be "average value" rather than "expected value."

# Remark 2: There are two ways of computing variances

**Theorem** If $X$ is a random variable, $\text{var}[X] = \mathscr{E}[(X - \mathscr{E}[X])^2] = \mathscr{E}[X^2] - (\mathscr{E}[X])^2$ provided $\mathscr{E}[X^2]$ exists.

PROOF (We first note that if $\mathscr{E}[X^2]$ exists, then $\mathscr{E}[X]$ exists.)* By our definitions of variance and $\mathscr{E}[g(X)]$, it follows that $\text{var}[X] = \mathscr{E}[(X - \mathscr{E}[X])^2]$. Now $\mathscr{E}[(X - \mathscr{E}[X])^2] = \mathscr{E}[X^2 - 2X\mathscr{E}[X] + (\mathscr{E}[X])^2] = \mathscr{E}[X^2] - 2(\mathscr{E}[X])^2 + (\mathscr{E}[X])^2 = \mathscr{E}[X^2] - (\mathscr{E}[X])^2$. ////

# 5.4 Moments and moment generating functions

## First and second moments

**Definition** **Moments** If $X$ is a random variable, the *rth moment* of $X$, usually denoted by $\mu'_r$, is defined as

$$\mu'_r = \mathscr{E}[X^r]$$

if the expectation exists. ////

Note that $\mu'_1 = \mathscr{E}[X] = \mu_X$, the mean of $X$.

**Definition** **Central moments** If $X$ is a random variable, the *rth central moment* of $X$ about $a$ is defined as $\mathscr{E}[(X-a)^r]$. If $a = \mu_X$, we have the *rth central moment* of $X$ about $\mu_X$, denoted by $\mu_r$, which is

$$\mu_r = \mathscr{E}[(X-\mu_X)^r].$$

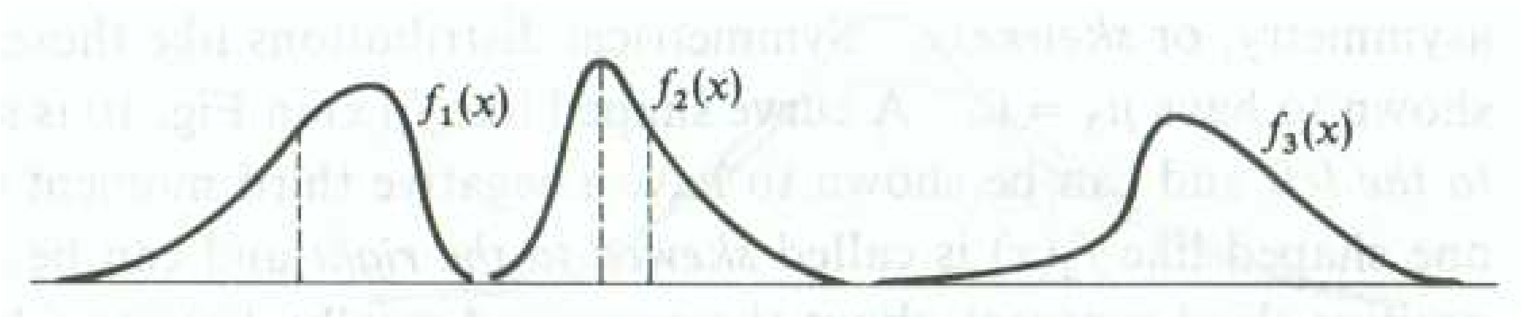
////

Note that $\mu_1 = \mathscr{E}[(X-\mu_X)] = 0$ and $\mu_2 = \mathscr{E}[(X-\mu_X)^2]$, the variance of $X$.

## Third central moment

- The third moment about the mean is sometimes called a measure of asymmetry, or **skewness**
- The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero.
- Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right.
- By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail.
- Some measurements have a lower bound and are skewed right. For example, in reliability studies, failure times cannot be negative.

- Knowledge of the third moment almost gives no clue as to the shape of the distribution … (e.g., $f_3(x)$ is far from symmetrical but its third moment is zero )




- The ratio

$$\mu_3/\sigma^3$$

is called the **coefficient of skewness** and is unitless.

The quantity $\mathscr{s}$ = (mean − median)/(standard deviation) provides an alternative measure of skewness. It can be proved that $-1 \leq \mathscr{s} \leq 1$.

## Fourth central moment

- The fourth moment about the mean is sometimes called a measure of excess, or **kurtosis**
- It refers to the degree of flatness of a density near its center
- Positive values of

$$\mu_4/\sigma^4 - 3,$$

(the **coefficient of excess** or kurtosis) are sometimes used to indicate that a density is more flat around its center than the density of a normal curve. This measure however suffers from the same failing as does the measure of skewness. It does not always measure what it is supposed to.

## The importance of moments … or not?

- In applied statistics, the first two moments are of great importance. It is usually necessary to know at least the location of the distribution and to have some idea about its dispersion or spread
- These characteristics can be estimated by examining a sample drawn from a set of objects known to have the distribution in question (see future chapters)
- In some cases, if all the moments are known, then the density can be determined.
- Hence, it would be useful if a function could be found that would give a representation of all the moments. Such a function is called a **moment generating function**.

# Moment generating functions

**Definition** **Moment generating function** Let $X$ be a random variable with density $f_X(\cdot)$. The expected value of $e^{tX}$ is defined to be the *moment generating function* of $X$ if the expected value exists for every value of $t$ in some interval $-h < t < h$; $h > 0$. The moment generating function, denoted by $m_X(t)$ or $m(t)$, is

$$m(t) = \mathscr{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx$$

if the random variable $X$ is continuous and is

$$m(t) = \mathscr{E}[e^{tX}] = \sum_{x} e^{tx} f_X(x)$$

if the random variable is discrete. ////

# Origin of the name "moment generating function"

One might note that a moment generating function is defined in terms of a density function, and since density functions were defined without reference to random variables (see Definitions 6 and 9), a moment generating function can be discussed without reference to random variables.

If a moment generating function exists, then $m(t)$ is continuously differentiable in some neighborhood of the origin. If we differentiate the moment generating function $r$ times with respect to $t$, we have

$$\frac{d^r}{dt^r} m(t) = \int_{-\infty}^{\infty} x^r e^{xt} f_X(x)\, dx,$$

and letting $t \to 0$, we find

$$\frac{d^r}{dt^r} m(0) = \mathscr{E}[X^r] = \mu'_r,$$

where the symbol on the left is to be interpreted to mean the $r$th derivative of $m(t)$ evaluated as $t \to 0$. Thus the moments of a distribution may be obtained from the moment generating function by differentiation, hence its name.

# Example

EXAMPLE .. Let $X$ be a random variable with probability density function given by $f_X(x) = \lambda e^{-\lambda x} I_{[0, \infty)}(x)$.

$$m_X(t) = \mathscr{E}[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx = \frac{\lambda}{\lambda - t} \qquad \text{for } t < \lambda.$$

$$m'(t) = \frac{dm(t)}{dt} = \frac{\lambda}{(\lambda - t)^2} \qquad \text{hence} \quad m'(0) = \mathscr{E}[X] = \frac{1}{\lambda}.$$

And $$m''(t) = \frac{2\lambda}{(\lambda - t)^3}, \qquad \text{so } m''(0) = \mathscr{E}[X^2] = \frac{2}{\lambda^2}.$$ ////

## Importance of moment generating functions

**Theorem** Let $X$ and $Y$ be two random variables with densities $f_X(\cdot)$ and $f_Y(\cdot)$, respectively. Suppose that $m_X(t)$ and $m_Y(t)$ both exist and are equal for all $t$ in the interval $-h < t < h$ for some $h > 0$. Then the two cumulative distribution functions $F_X(\cdot)$ and $F_Y(\cdot)$ are equal. ////

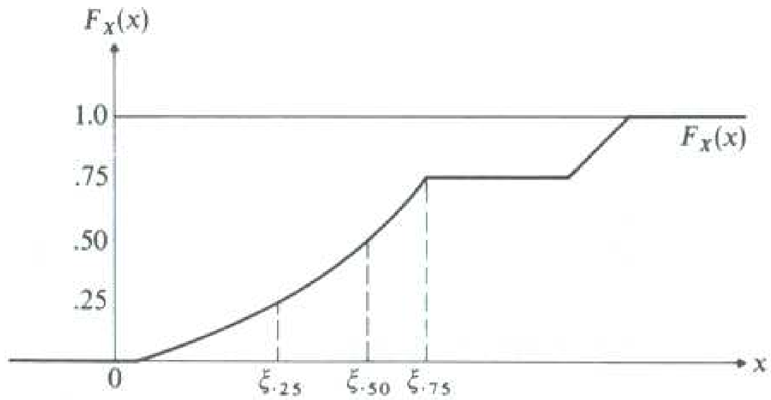
- In principle it is possible that there exists a sequence of moments for which there is a large collection of different distributions functions having these same moments → so a sequence of moments does not determine uniquely the corresponding distribution function
- However, the above theorem indicates that if a moment generating function of a random variable exists, it does uniquely determine the corresponding distribution function

# Further characterizing densities: about quantiles and medians

**Definition** **Quantile** The *qth quantile* of a random variable $X$ or of its corresponding distribution is denoted by $\xi_q$ and is defined as the smallest number $\xi$ satisfying $F_X(\xi) \geq q$. ////

**Note**: equality is attained in the continuous case

**Definition** **Median** The *median* of a random variable $X$, denoted by $\text{med}_X$, med $(X)$, or $\xi_{.50}$, is the .5th quantile. ////

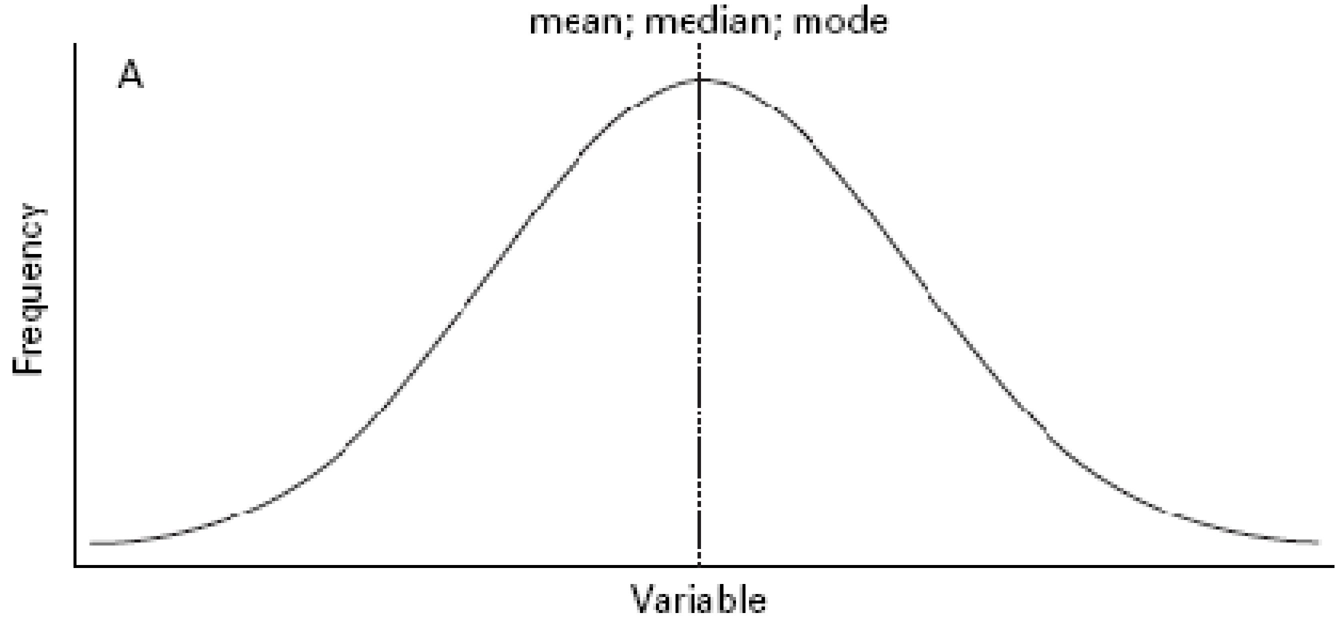


## Further characterizing densities: about modes and means

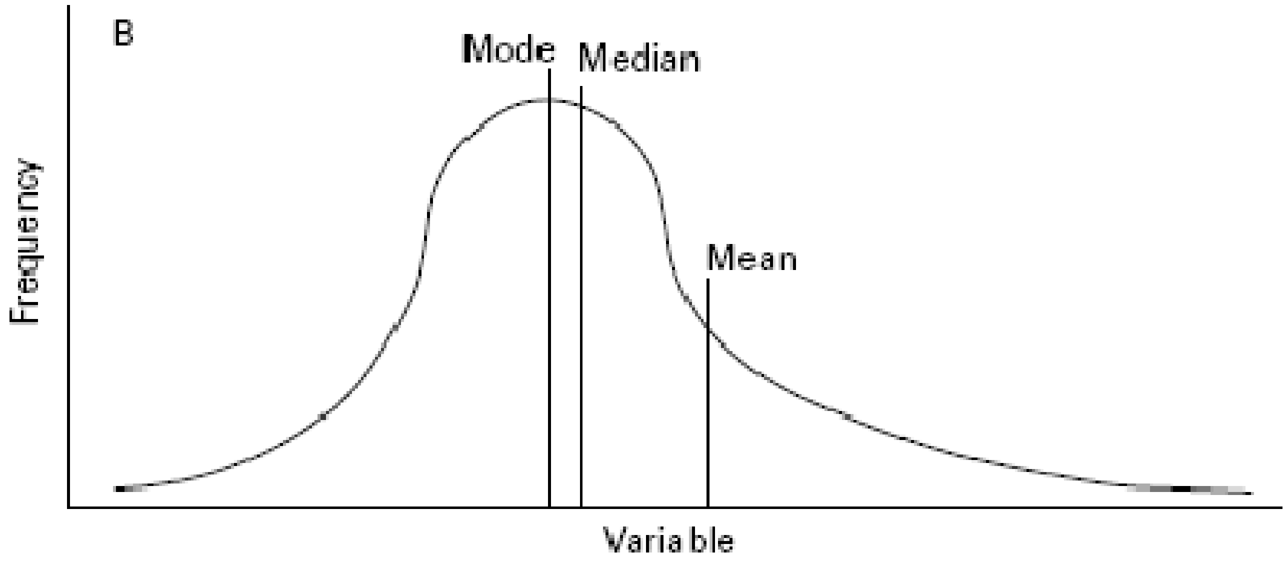
- So far we have seen two measures of location of the density of a random variable X:
  - o The **first moment** of X
  - o The **median** of X
- A third measure of central location is the **mode of X**, which is defined as that point (if that point exists) at which the density function *f* attains its maximum.

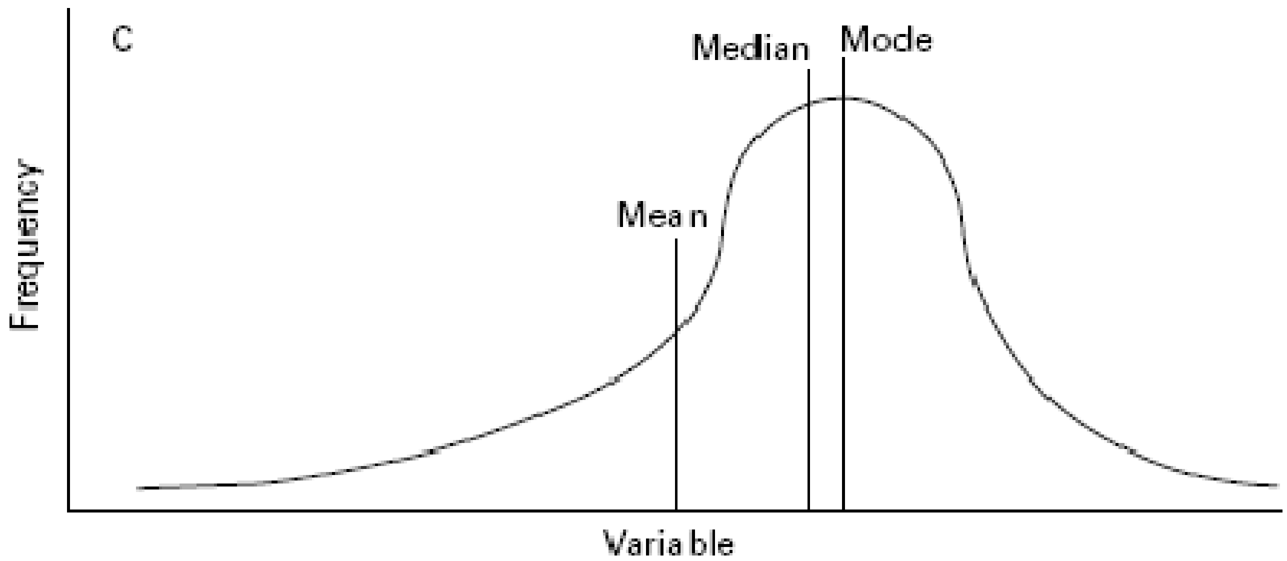
Applet Mean and Median

(http://bcs.whfreeman.com/ips6e)




- Location of mean, median and mode with different distributions: A = normal distribution; B (next slide) = Right (positive) skewed distribution; C (next slide) = Left (negative) skewed distribution.


B
Mode
Median
Mean
Frequency
Variable

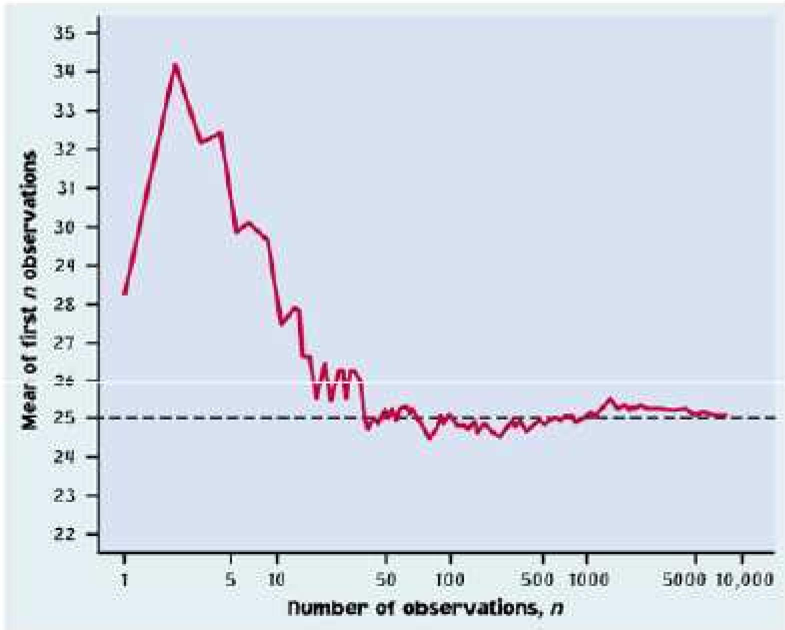

C
Median
Mode
Mean
Frequency
Variable

# 5.5 Useful results

## Law of large numbers

As the number of randomly drawn observations ($n$) in a sample increases, the mean of the sample ($x$ bar) gets closer and closer to the population mean $\mu$.

This is the **law of large numbers.** It is valid for <u>any</u> population.




*Note: We often intuitively expect predictability over a few random observations, but it is wrong. The law of large numbers only applies to really large numbers.*

# Applet Law of Large Numbers

(http://bcs.whfreeman.com/ips6e)

# Chebyshev inequality

**Theorem** Let $X$ be a random variable and $g(\cdot)$ a nonnegative function with domain the real line; then

$$P[g(X) \geq k] \leq \frac{\mathscr{E}[g(X)]}{k} \qquad \text{for every } k > 0.$$

PROOF Assume that $X$ is a continuous random variable with probability density function $f_X(\cdot)$; then

$$\begin{aligned}\mathscr{E}[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x)\,dx = \int_{\{x:\, g(x) \geq k\}} g(x) f_X(x)\,dx \\ &+ \int_{\{x:\, g(x) < k\}} g(x) f_X(x)\,dx \geq \int_{\{x:\, g(x) \geq k\}} g(x) f_X(x)\,dx \\ &\geq \int_{\{x:\, g(x) \geq k\}} k f_X(x)\,dx = kP[g(X) \geq k].\end{aligned}$$

Divide by $k$, and the result follows. A similar proof holds for $X$ discrete.

////

**Corollary Chebyshev inequality** If $X$ is a random variable with finite variance,

$$P[|X - \mu_X| \geq r\sigma_X] = P[(X - \mu_X)^2 \geq r^2\sigma_X^2] \leq \frac{1}{r^2} \qquad \text{for every } r > 0.$$

PROOF Take $g(x) = (x - \mu_X)^2$ and $k = r^2\sigma_X^2$

////

**Remark** If $X$ is a random variable with finite variance,

$$P[|X - \mu_X| < r\sigma_X] \geq 1 - \frac{1}{r^2},$$

which is just a rewriting ////

- **Chebyshev inequality** is used in various ways, to prove the law of large numbers or to establish $P[\mu_X - r\sigma_X < X < \mu_X + r\sigma_X] \geq 1 - \frac{1}{r^2}$

# Jensen inequality

**Definition** **Convex function** A continuous function $g(\cdot)$ with domain and counterdomain the real line is called *convex* if for every $x_0$ on the real line, there exists a line which goes through the point $(x_0, g(x_0))$ and lies on or under the graph of the function $g(\cdot)$. ////

**Theorem** **Jensen inequality** Let $X$ be a random variable with mean $\mathscr{E}[X]$, and let $g(\cdot)$ be a convex function; then $\mathscr{E}[g(X)] \geq g(\mathscr{E}[X])$.

PROOF Since $g(x)$ is continuous and convex, there exists a line, say $l(x) = a + bx$, satisfying $l(x) = a + bx \leq g(x)$ and $l(\mathscr{E}[X]) = g(\mathscr{E}[X])$. $l(x)$ is a line given by the definition of continuous and convex that goes through the point $(\mathscr{E}[X], g(\mathscr{E}[X]))$. Note that $\mathscr{E}[l(X)] = \mathscr{E}[(a + bX)] = a + b\mathscr{E}[X] = l(\mathscr{E}[X])$; hence $g(\mathscr{E}[X]) = l(\mathscr{E}[X]) = \mathscr{E}[l(X)] \leq \mathscr{E}[g(X)]$

////

- Note that in general,

$$\mathscr{E}[g(X)] \neq g(\mathscr{E}[X])$$

- **Jensen inequality** can be used to prove the Rao-Blackwell theorem
- The latter provides a method for improving the performance of an unbiased estimator of a parameter (i.e. reduce its variance) provided that a "sufficient" statistic for this estimator is available.
- With g(x)=$x^2$ (hence g is a convex function), Jensen inequality says

$$\mathscr{E}[X^2] \geq (\mathscr{E}[X])^2$$

and therefore that the variance of X is always non-negative

# Background reading:

- P Driscoll, F Lecky,M (2000) CrosbyAn intro to everyday statistics – 1. *J Accid Emerg Med* 17:205–211
- P Driscoll, F Lecky,M (2000) CrosbyAn intro to everyday statistics – 2. *J Accid Emerg Med* 17:274–281