STAT 6200 — Introduction to Biostatistics
Lecture Notes

Introduction*

Statistics and Biostatistics:

The *field* of statistics: The study and use of theory and methods for the analysis of data arising from **random** processes or phenomena. The study of how we make sense of **data.**

- The field of statistics provides some of the most fundamental tools and techniques of the scientific method:
  - forming hypotheses,
  - designing experiments and observational studies,
  - gathering data,
  - summarizing data,
  - drawing inferences from data (e.g., testing hypotheses)
- A **statistic** (rather than the field of "statistics") also refers to a numerical quantity computed from sample data (e.g., the mean, the median, the maximum).

Roughly speaking, the field of statistics can be divided into

- *Mathematical Statistics:* the study and development of statistical theory and methods in the abstract; and
- *Applied Statistics:* the application of statistical methods to solve real problems involving randomly generated data, and the development of new statistical methodology motivated by real problems.

* Read Ch.1 of our text.

Biostatistics is the branch of applied statistics directed toward applications in the health sciences and biology.

- Biostatistics is sometimes distinguished from the field of biometry based upon whether applications are in the health sciences (biostatistics) or in broader biology (biometry; e.g., agriculture, ecology, wildlife biology).

- Other branches of (applied) statistics: psychometrics, econometrics, chemometrics, astrostatistics, environmetrics, etc.

Why biostatistics? What's the difference?

- Because some statistical methods are more heavily used in health applications than elsewhere (e.g., survival analysis, longitudinal data analysis).

- Because examples are drawn from health sciences.

  - Makes subject more appealing to those interested in health.

  - Illustrates how to apply methodology to similar problems encountered in real life.

We will emphasize the methods of data analysis, but some basic theory will also be necessary to enhance understanding of the methods and to allow further coursework.

- Mathematical notation and techniques are necessary! (No apologies.)

We will study what to do and how to do it, but also very important is why the methods are appropriate and what are the concepts justifying those methods.

- The latter (the why) will get you further than the former (the what).

Data*

Data Types:

Data are observations of random variables made on the elements of a population or sample.

- Data are the quantities (numbers) or qualities (attributes) measured or observed that are to be collected and/or analyzed.
- The word "data" is plural, "datum" is singular!
- A collection of data is often called a *data set* (singular).

Example — Low Birth Weight Infant Data

- Appendix B of our text contains a data set called lowbwt containing measurements and observed attributes on 100 low birth weight infants born in two teaching hospitals in Boston, MA.
- The variables measured here are

sbp = systolic blood pressure

sex = gender (1=male, 0=female)

tox = maternal diagnosis of toxemia (1=yes, 0=no)

grmhem = whether infant had a germinal matrix hemorrhage (1=yes, 0=no)

gestage = gestational age (weeks)

apgar5 = Apgar score (measures oxygen deprivation) at 5 minutes after birth

- Data are reproduced on the top of the following page.

* Read Ch.2 of our text.

- There are 6 variables here (sbp, sex, etc.) measured on 100 units/elements/subjects (the infants) of a random sample of size 100.
- An observation can refer to the value of a single variable for a particular subject, but more commonly it refers to the observed values of all variables measured on a particular subject.
    - There are 100 observations here.

Types of Variables:

Variable types can be distinguished based on their scale. Typically, different statistical methods are appropriate for variables of different scales.

Scale	Characteristic Question	Examples
Nominal	Is A different than B?	Marital status Eye color Gender Religious affiliation Race
Ordinal	Is A bigger than B?	Stage of disease Severity of pain Level of satisfaction
Interval	By how many units do A and B differ?	Temperature SAT score
Ratio	How many times bigger than B is A?	Distance Length Time until death Weight

Operations that make sense for variables of different scales:

	Operations that make sense			
Scale	Counting	Ranking	Addition/ Subtraction	Multiplication/ Division
Nominal	√			
Ordinal	√	√		
Interval	√	√	√	
Ratio	√	√	√	√

- Often, the distinction between interval and ratio scales can be ignored in statistical analyses. Distinction between these two types and ordinal and nominal are more important.

Another way to distinguish between types of variables is as **quantitative** or **qualitative**.

- Qualitative variables have values that are intrinsically nonnumeric (categorical).
  - E.g., Cause of death, nationality, race, gender, severity of pain (mild, moderate, severe).
  - Qualitative variables generally have either nominal or ordinal scales.
  - Qualitative variables can be reassigned numeric values (e.g., male=0, female=1), but they are still intrinsically qualitative.
- Quantitative variables have values that are intrinsically numeric.
  - E.g., survival time, systolic blood pressure, number of children in a family, height, age, body mass index.

Quantitative variables can be further subdivided into **discrete** and **continuous** variables.

- Discrete variables have a set of possible values that is either finite or countably infinite.
  - E.g., number of pregnancies, shoe size, number of missing teeth.
  - For a discrete variable there are gaps between its possible values. Discrete values often take integer (whole numbers) values (e.g., counts), but some discrete variables can take non-integer values.
- A continuous variable has a set of possible values including all values in an interval of the real line.
  - E.g., duration of a seizure, body mass index, height.
  - No gaps between possible values.

The distinction between discrete and continuous quantitative variables is typically clear theoretically, but can be fuzzy in practice.

- In practice the continuity of a variable is limited by the precision of the measurement. E.g., height is measured to the nearest centimeter, or perhaps millimeter, so in practice heights measured in millimeters only take integer values.
    - Another example: survival time is measured to the nearest day, but could, theoretically, be measured to any level of precision.
- On the other hand, the total annual attendance at UGA football games is a discrete (inherently integer-valued) variable, but, in practice, can be treated as continuous.
- In practice, all variables are discrete, but we treat some variables as continuous based upon whether their distribution can be "well approximated" by a continuous distribution.

Data Sources:

Data arise from experimental or observational studies, and it is important to distinguish the two.

- In an **experiment**, the researcher deliberately imposes a treatment on one or more subjects or experimental units (not necessarily human). The experimenter then measures or observes the subjects' response to the treatment.
    - Crucial element is that there is an intervention.

Example: To assess whether or not saccharine is carcinogenic, a researcher feeds 25 mice daily doses of saccharine. After 2 months, 10 of the 25 mice have developed tumors.

- By definition, this is an experiment, but not a very good one.

In the saccharine example, we don't know whether 10/25 with tumors is high because there is no **control group** to which comparison can be made.

Solution: Select 25 more mice and treat them exactly the same but give them daily doses of an inert substance (a **placebo**).

Suppose that in the control group only 1 mouse develops a tumor. Is this evidence of a carcinogenic effect?

Maybe, but there's still a problem:

- What if the mice in the 2 groups differ systematically? E.g., group 1 from genetic strain 1, group 2 from genetic strain 2.

Here, we don't know whether saccharine is carcinogenic, or if genetic strain 1 is simply more susceptible to tumors.

- We say that the effects of genetic strain and saccharine are **confounded** (mixed up).

Solution: Starting with 50 relatively homogeneous (similar) mice, randomly assign 25 to the saccharine treatment, and 25 to the control treatment.

- **Randomization** an extremely important aspect of experimental design.
  - In the saccharine example, we should start out with 50 homogeneous mice, but of course they will differ some. Randomization ensures that the two experimental groups will be **probabilistically alike** with respect to all **nuisance variables** (potential confounders). E.g., the distribution of body weights should be about the same in the two groups.

Another important concept, especially in human experimentation, is **blinding.**

- An experiment is blind if the subjects don't know which treatment they receive.
- E.g., suppose we randomize 25 of 50 migraine sufferers to an active drug and the remaining 25 to a placebo control treatment.
  - Experiment is blind if pills in the two treatment groups look and taste identical and subjects are not told which treatment they receive.
  - This guards against the **placebo effect.**
- An experiment is **double-blind** if the researcher who administers the treatments and measures the response does not know which treatment is assigned.
  - Guards against **experimenter effects**. (Experimenter may behave differently toward the subjects in the two groups, or measure the response differently in the two groups.)

Experiments are to be contrasted with **observational studies**.

- No intervention.
- Data collected on an existing system.
  - Less expensive.
  - Easier logistically.
  - More often ethically practical.
  - Interventions often not possible.
- Experiments have many advantages and are strongly preferred when possible. However, experiments are rarely feasible in public health/epidemiology.
  - In health sciences/medicine, experiments involving humans are called **clinical trials**.

Types of Observational Studies:

1. Case studies or case series.

   - A descriptive account of interesting characteristics (e.g., symptoms) observed in a single case (subject with disease) or in a sample of cases.

   - Typically are unplanned and don't involve any research hypotheses. No comparison group.

   - Poor design, but can generate research hypotheses for subsequent investigation.

2. Case-control study.

   - Conducted **retrospectively** (by looking into past).

   - Two types of subjects included:

     $$\text{cases} = \text{subjects with the disease/outcome of interest}$$
     $$\text{controls} = \text{subjects without the disease/outcome}$$

   - History of two groups is examined to determine which subjects were exposed to, or otherwise possessed, a prior characteristic. Association between exposure and disease then quantified.

   - Controls are often matched to cases based on similar characteristics.

- Advantages:

  - Useful for studying rare disease.
  - Useful for studying diseases with long latency periods.
  - Can explore several potential risk factors (exposures) for disease simultaneously.
  - Can use existing data sources - cheap, quick, easy to conduct.

- Disadvantages:

  - Prone to methodological errors and biases.
  - Dependent on high quality records.
  - Difficult to select an appropriate control group.
  - More difficult statistical methods required for proper analysis.

3. Cross-sectional Studies.

- Collect data from a group of subjects at one point in time.
- Sometimes called prevalence studies, due to their focus on a single point in time.
- Advantages:
  - Often based on a sample of the general population, not just people seeking medical care.
  - Can be carried out over a relatively short period of time.
- Disadvantages:
  - Difficult to separate cause and effect because measurement of exposure and disease are made at one point in time, so it may not be possible to determine which came first.
  - Are biased toward detecting cases with disease of long duration and can involve misclassifications of cases in remission or under effective medical treatment.
  - Snapshot in time can be misleading in a variety of other ways.

4. Cohort Studies.

   - Usually conducted **prospectively** (forward in time).

   - A **cohort** is a group of people who have something in common at a particular point in time and who remain part of the group through time.

   - A cohort of disease-free subjects are selected and their exposure status evaluated at the start of the study.

   - They are then followed through time in order to observe who develops disease. Association between exposures (risk factors) and disease are then quantified.

- Advantages:

  - Useful when exposure of interest is rare.
  - Can examine multiple effects (e.g., diseases) of a single exposure.
  - Can elucidate temporal relationship between exposure and disease, thereby getting closer to causation.
  - Allows direct measurement of incidence of disease.
  - Minimizes bias in ascertainment of exposure.

- Disadvantages:

  - Inefficient for studying rare diseases.
  - Generally requires a large number of subjects.
  - Expensive and time-consuming.
  - Subjects can be lost to follow-up (drop out of study) leading to bias.

- Cohort studies can also be conducted retrospectively by identifying a cohort in present, determining exposure status in past, and then determining subsequent disease occurrence between time of exposure and present through historical records.

Data Presentation:

Even quite small data sets are difficult to comprehend without some summarization. Statistical quantities such as the mean and variance can be extremely helpful in summarizing data, but first we discuss tabular and graphical summaries.

Tables:

One of the most important means of summarizing the data from a single variable is to tabulate the **frequency distribution** of the variable.

- A frequency distribution simply tells how often a variable takes on each of its possible values. For quantitative variables with many possible values, the possible values are typically binned or grouped into intervals.

Example - Gender in this Class (Nominal Variable):

Gender	Frequency	Relative Frequency (proportion)	Relative Frequency (percent)
Female			
Male			
Total			

- Here, the relative frequency as a proportion is just

$$\text{Relative frequency (proportion)} = \text{Frequency}/n$$

  where n =sample size.

- The relative frequency as a percent is

  $$\text{Relative Frequency (percent)} = \text{Relative frequency (proportion)} \times 100\%$$

- It is worth distinguishing between the *empirical* relative frequency distribution, which gives the proportion or percentage of observed values, and the *probability* distribution, which gives the probability that a random variable takes each of its possible values.
  - The latter can be thought of as the relative frequency distribution for an infinite sample size.

Example - Keypunching Errors (Discrete Quantitative Variable):

A typist entered 156 lines of data into a computer. The following table gives the number of errors made for each line.

Number of Errors	Frequency	Relative Frequency (%)
0	124	
1	27	
2	5	
3 or more	0	
Total		

- Here, it was not necessary to bin the data.

Example - Age at Death (in Days) for SIDS Cases:

The following table contains the age at death in days for 78 cases of sudden infant death syndrome (SIDS, or Crib Death) occuring in King County, WA, during 1976–1977.

Age Interval (Days)	Frequency	Relative Frequency (%)	Cumulative Frequency	Cumulative Relative Frequency (%)
1–30	6	7.69	6	7.69
31–60	13	16.67	19	24.36
61–90	23	29.49	42	53.85
91–120	18	23.08	60	76.92
121–150	6	7.69	66	84.62
151–180	5	6.41	71	91.03
181–210	3	3.85	74	94.87
211–240	2	2.56	76	97.44
241–270	0	0	76	97.44
271–300	1	1.28	77	98.72
301–330	1	1.28	78	100.00

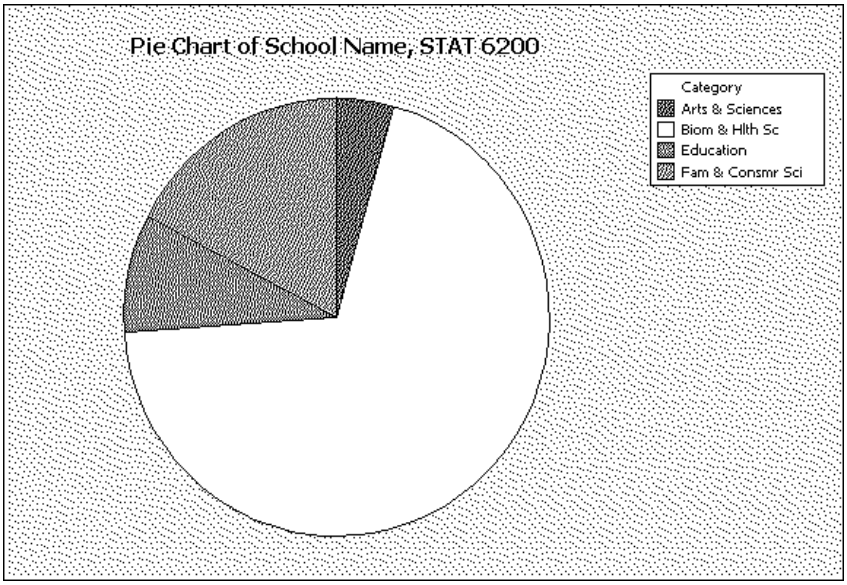
- Here it is necessary to bin the data. The bins should be
  - Mutually exclusive (non-overlapping).
  - Exhaustive (every observed value falls in a bin)
  - The handling of cutpoints between bins should be consistent and clearly defined.
  - (preferrably) The bins should be of equal width, although it can be better to violate this rule sometimes, especially for the smallest and largest bins.

- In this example we have also tabulated the cumulative frequency and the cumulative relative frequency. The cumulative frequency simply counts the number of observations $\leq$ the current value (or current bin if the data are binned).
  - The cumulative relative frequency expresses the same information as a percent by multiplying by $\frac{100\%}{n}$.

Graphs:

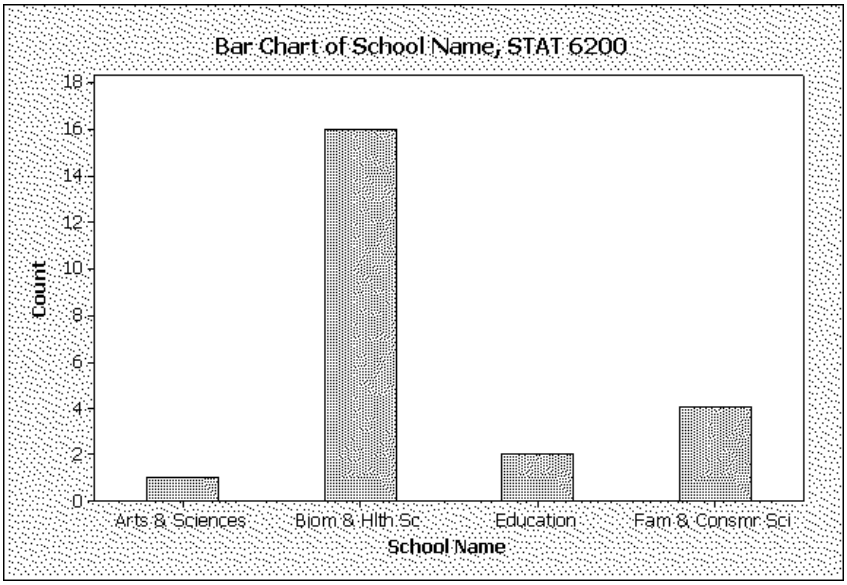
Frequency distributions can often be displayed effectively using graphical means such as the **bar chart**, **pie chart**, or **histogram**.

- Pie charts are useful for displaying the relative frequency distribution of a nomianl variable. Here is an example created in Minitab of the relative frequency distribution of the school affiliation of students in this class.



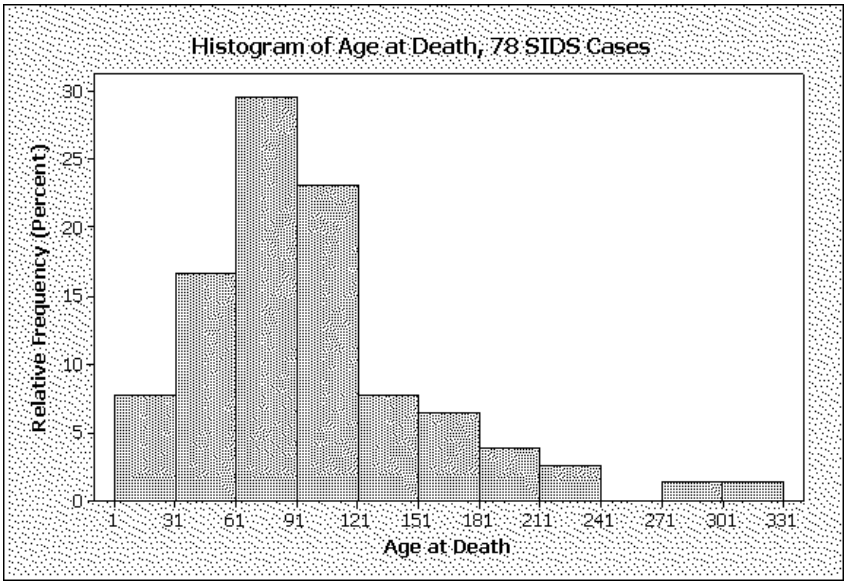
- A **legend**, or **key** is important in many different graph types, but is especially crucial in a pie chart.

- Bar charts display absolute or relative frequency distributions for categorical variables (ordinal or nominal). Here is a Minitab bar chart of the school affiliations of students in this class.

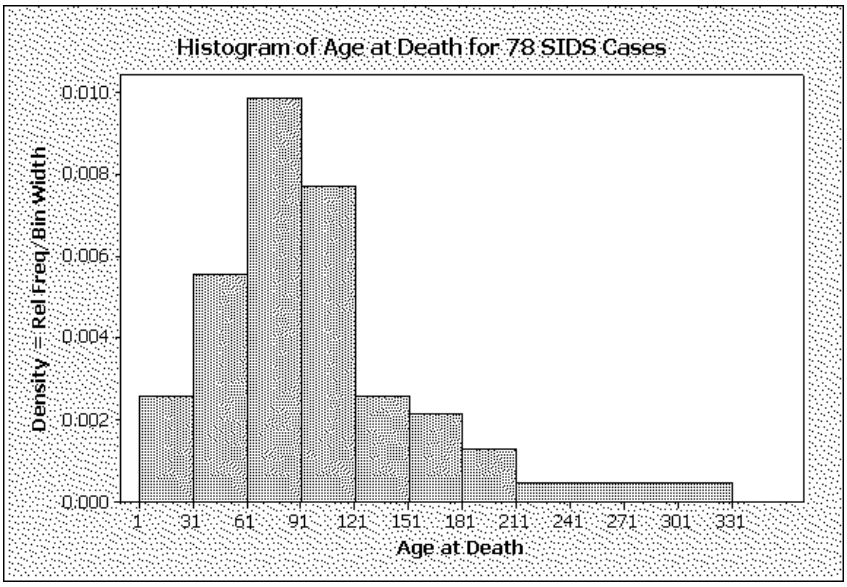


- Note that the horizontal axis in a bar chart has no scale. The categories can be re-ordered arbitrarily without affecting the information contained in the plot.

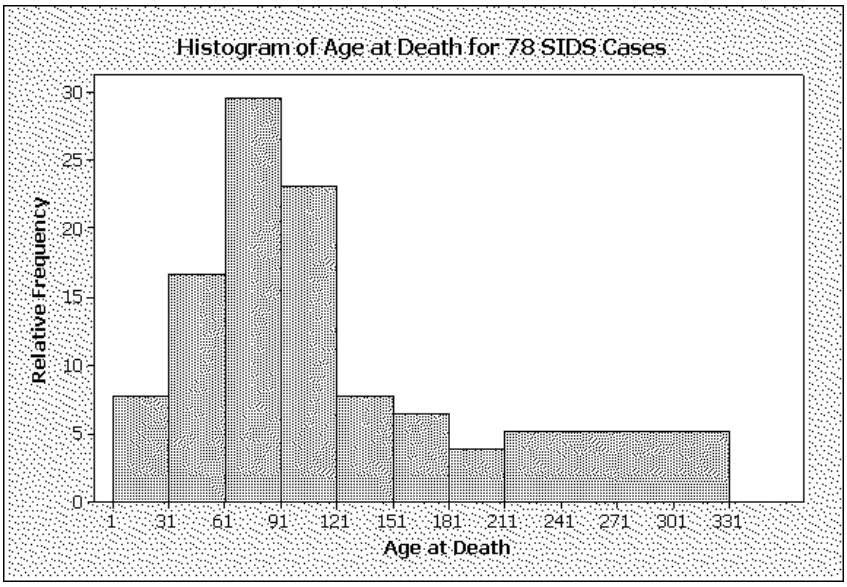
A **histogram** depicts the frequency distribution of a quantitative random variable. Below is a histogram of the Age at Death data for the 78 SIDS cases in King Co., WA.



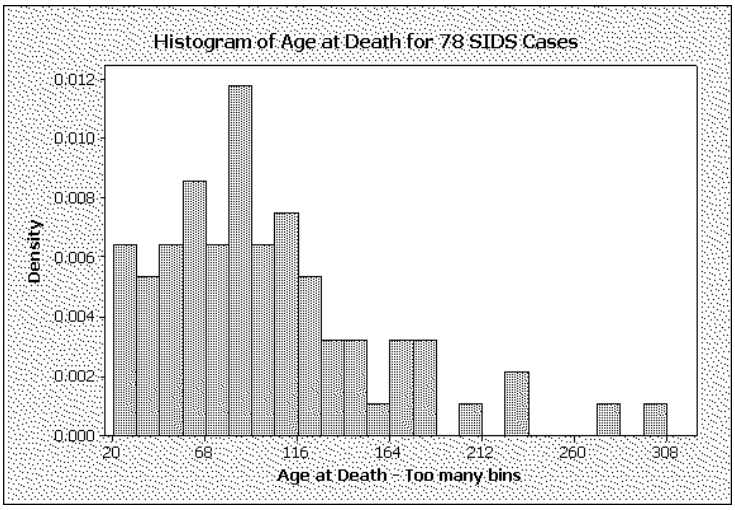
- Histograms are sometimes constructed so that the height of each bar gives the frequency (or relative frequence) in each interval. This is ok if the intervals all have the same width, but can be misleading otherwise.
  - Here's an example of what can go wrong with unequal bins when frequency or relative frequency is plotted.



- The above example can be fixed by making the relative frequency in each interval equal to the area in each bar, not the height. That is, the height of each bar should be equal to Rel Freq/Bin Width. Here's a fixed version of the histogram given above.



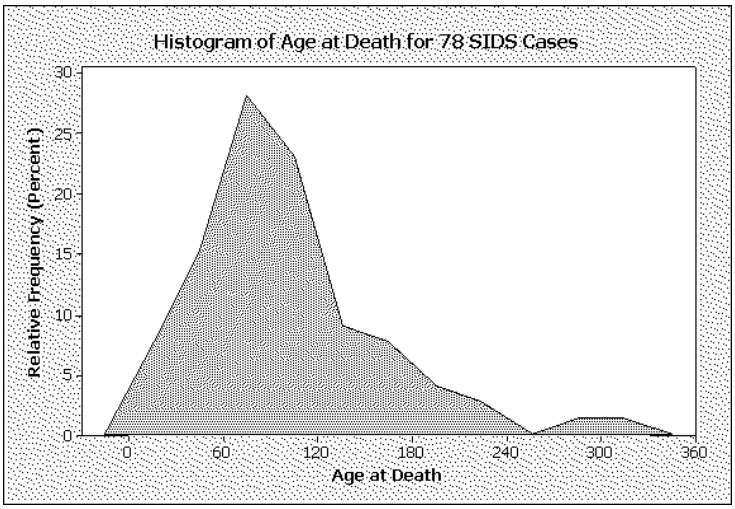
Note that the choice of number of bins and bin width can affect histograms dramatically. Here's a different choice of bins for the SIDS data (a bad choice).



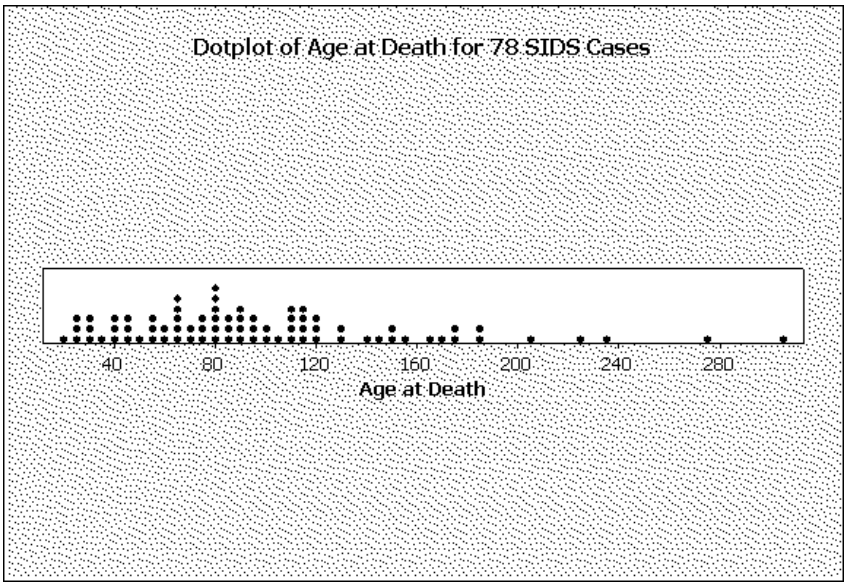
- It is not easy to give a general rule on how many bins should be used in a histogram, but somewhere between 5 and 20 bins is typically avisable.

Frequency polygons are formed by plotting lines between the midpoints of the tops of the bars of a histogram. The histogram should have equal bin widths and the lines should extend down to 0 at the right and left extremes of the data.

- Here is a frequency polygon for the SIDS data. Its principle advantages are that (i) it is continuous, and (ii) multiple frequency polygons can be displayed on the same plot.



A **one-way scatter plot** is just a plot of the real line with tick marks, or sometimes dots, at each observed value of the variable. Here is a one-way scatter plot, or dotplot, for the SIDS data

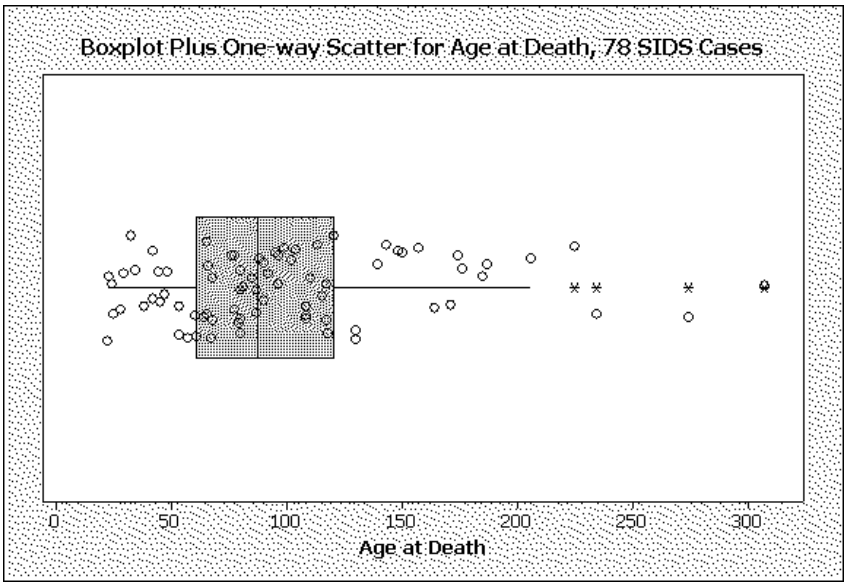


Another plot useful for summarizing the distribution of a single variable is the **boxplot**.

A boxplot summarizes the distribution of a variable by locating the 25th, 50th and 75th **percentiles** of the data, plus two **adjacent values.**

- A p^{th} percentile of a data set is a number such that at least $p\%$ of the data are $\leq$ this value and at least $100 - p\%$ of the values are $\geq$ this value.
  - The **median** is the 50^{th} percentile.
  - The 25^{th}, 50^{th} and 75^{th} percentiles are sometimes called the first, second, and third **quartiles** of the data.
- The box in a boxplot extends from the 25th to the 75th percentiles of the data. The line in the box locates the median.

Here is a boxplot for the SIDS data. This boxplot also includes a one-way scatterplot of the data.



- The lines extending on either side of the box are called *whiskers*. They indicate roughly the extent of the data.
  - The whiskers sometimes extend to the 10th and 90th percentiles.
  - In Minitab's implementation of a boxplot, however, the whiskers extend to the adjactent values, which are defined to be the most extreme values in the data set that are not more than 1.5 times the width of the box beyond either quartile.
  - The width of the box is the distance between the first and third quartile. This distance is called the **interquartile range**.
- The term **outlier** is used to refer to data points that are not typical of the rest of the values. Exactly what constitutes "not typical" is somewhat controversial, but one way to define an outlier is as a point beyond the adjacent values.
  - Based on this definition there are four large outliers in the SIDS data (marked by *'s) and no small outliers.

There are a variety of tabular and graphical methods to summarize the **joint frequency distribution** of two variables.

For two qualitative variables, a **contingency table** or **cross-tabulation** is useful.

- This is just a table where the rows represent the values of one variable, the columns the values of the other variable, and the cells give the frequency with which each combination of values is observed.

Here is a contingency table giving the joint frequency distribution of grmhem (germinal matrix hemorrhage) and tox (diagnosis of toxemia for mother) for the low birth weight data:

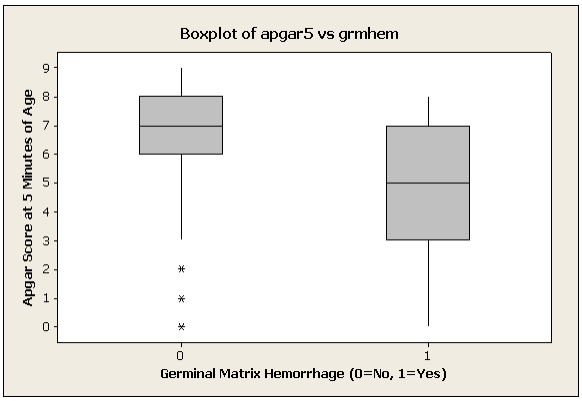
		Germinal Hemorrhage		
		No	Yes	
Toxemia	No	65	14	79
	Yes	20	1	21
		85	15	100

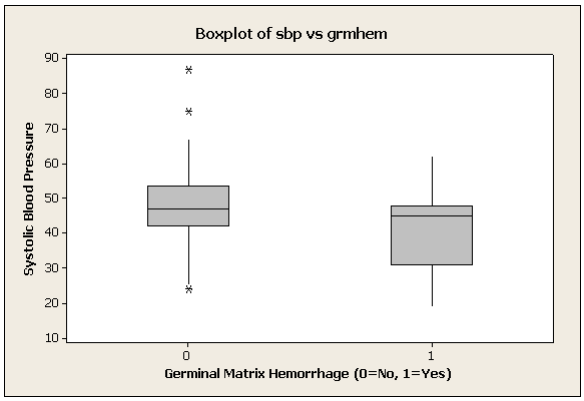
- Notice that the margins of the table give the univariate frequency distributions of the two variables.

Cross-tabulations can be constructed when one or more of the variables are quantitative. In this case, it may be necessary to bin the quantitative variable(s). E.g., here is a cross-tab of gestage (gestastional age) and grmhem:

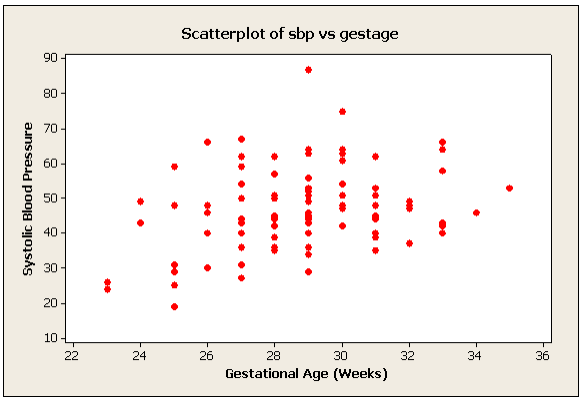
		Germinal Hemorrhage		
		No	Yes	
Gestational Age	23–25	7	4	11
	26–28	26	4	30
	29–31	38	6	44
	32–34	13	1	14
	≥ 35	1	0	1
		85	15	100

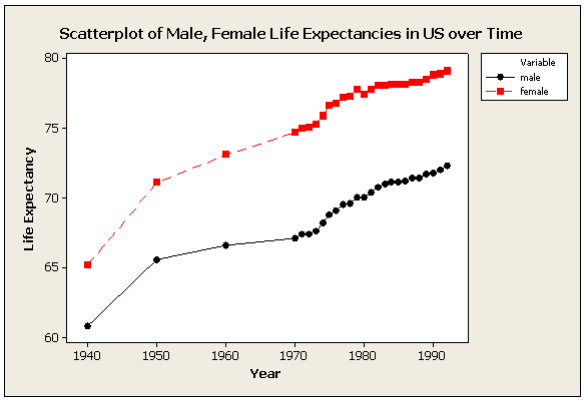
For displaying the relationship between a quantitative variable and a qualitative variable, side-by-side boxplots or superimposed frequency polygons can be useful.





The most useful graphical tool for displaying the relationship between two quantitative variables is a **two-way scatterplot.** Here is one that displays systolic blood pressure vs. gestational age for the low birth weight data.





Line graphs are useful when a variable is measured at each of many consecutive points in time (or some other dimension like depth of the ocean). In such a situation it is useful to construct a scatterplot with the measured variable on the vertical axis and time on the horizontal. Connecting the points gives a sense of the time trend and any other temporal pattern (e.g., seasonality).

- Above is a line graph displaying US Life Expectancies over time. Male and female life expectancies are plotted on the same graph here.

Numerical Summary Measures*

In talking about numerical summary measures, it is useful to distinguish between whether a particular quantity (such as the mean) is computed on a sample or on an entire population.

Sometimes, we have data on all units (e.g., subjects) in which we have interest. That is, we have observations on the entire population.

- E.g., the diameters of the nine planets of the solar system are

Planet	Diameter (miles)
Mercury	3030
Venus	7520
Earth	7926
Mars	4217
Jupiter	88838
Saturn	74896
Uranus	31762
Neptune	30774
Pluto	1428

- The mean diameter of the 9 planets is

$$(3030 + 7520 + \cdots + 1428)/9 = 27,821.22 \quad \text{miles.}$$

Summary measures such as the mean and variance are certainly useful for such data, but it is important to realize that there is no need to estimate anything here or to perform statistical inference.

- The mean diameter of the 9 planets in our solar system *is* 27,821.22 miles. This is a population quantity or **parameter** that can be computed from direct measurements on all population elements.

* Read Ch.3 of our text.

In contrast, the more common situation is one in which we can't observe the entire population. Instead we select a subset of the population called a **sample**, chosen to be representative of the population of interest.

Summary measures computed on the sample are used to make **statistical inference** on the corresponding population quantities.

- That is, we don't know the parameter, so we estimate its value from a sample, quantify the uncertainty in that estimate, test hypotheses about the parameter value, etc.

- E.g., we don't know the proportion of US registered voters who approve of President Bush's job performance, so we take a representative sample of the population of size 1,000, say, and ask each sample member whether they approve. The proportion of these 1,000 sample members who approve (a sample statistic) is used to estimate the corresponding proportion of the total US population (the parameter).

    - Note that this estimate will almost certainly be wrong. One of the major tasks of statistical inference is in determining how wrong it is likely to be.

Notation:

Random variables will be denoted by Roman letters (e.g., x, y).

Sample quantities: Roman letters (e.g., mean=$\bar{x}$, variance=s^2)
Population quantities: Greek letters (e.g., mean= μ, variance= σ^2).

Suppose we have a sample on which we measure a random variable that we'll call x (e.g., age at death for 78 SIDS cases):

$$225, 174, 274, 164, \ldots, 32, 44$$

A convenient way to refer to these numbers is as $x_1, x_2, \ldots, x_n$ where n is the sample size. Here,

$$x_1 = 225, x_2 = 174, x_3 = 274, \ldots, x_{77} = 32, x_{78} = 44.$$

Summation notation: many statistical formulas involve summing a series of number like this, so it is convenient to have a shorthand notation for $x_1 + x_2 + \cdots + x_n$. Such a sum is denoted by $\sum_{i=1}^{n} x_i$. That is,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

Similarly,

$$\sum_{i=1}^{3} 4(x_i - y_i)^2 = 4(x_1 - y_1)^2 + 4(x_2 - y_2)^2 + 4(x_3 - y_3)^2.$$

Measures of Location:

Mean: The sample mean measures the location or **central tendency** of the observations in the sample. For a sample $x_1, \ldots, x_n$, the mean is denoted by $\bar{x}$ and is computed via the formula

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- The mean gives the point of balance for a histogram of the sample values and is affected by every value in the sample.
- Sample mean for age at death, SIDS cases:

$$\bar{x} = \frac{1}{78}\sum_{i=1}^{78} x_i = \frac{1}{78}(225 + 174 + \cdots + 44) = 99.29$$

- The population mean is the same quantity computed on all the elements in the population.
  - In the SIDS example, the population is not clearly defined. We may think of the SIDS cases in 1976–77 in King Co. Washington as representative of the entire US or of similar metropolitan areas in the US at that point in time, or as representative of King County at points in time other than 1976–77.
- The mean is not an appropriate measure for ordinal or nomianl variables.

Median: One feature of the mean that is sometimes undesirable is that it is affected by every value in the data set. In particular, this means that it is sensitive to extreme values, which at times may not be typical of the data set as a whole.

- The median does not have this feature, and is therefore sometimes more appropriate for conveying the "typical" value in a data set.

The median is defined as the 50^{th} percentile or middle value of a data set. That is, the median is a value such that at least half of the data are greater than or equal to it and at least half are less than or equal to it.

- If n is odd, this definition leads to a unique median which is an observed value in the data set.

  E.g., 9 health insurance claims (dollar amounts):

$$\begin{aligned} \text{data:} \quad & 1100, 1900, 600, 890, 690, 890000, 380, 1200, 1050 \\ \text{sorted data:} \quad & 380, 600, 690, 890, 1050, 1100, 1200, 1900, 890000 \\ \Rightarrow \quad & \text{median} = 1050, \quad \text{whereas the mean} = 99,756.67 \end{aligned}$$

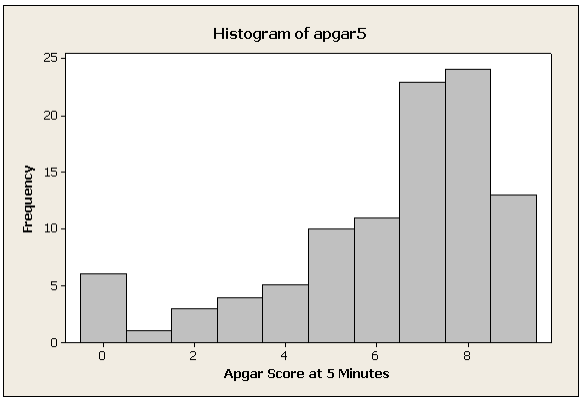
- If n is even, there are two "middle values" and either middle value or any number in between would satisfy the definition. By convention we take the average of the two middle values.

$$\begin{aligned} \text{sorted data:} \quad & 600, 690, 890, 1050, 1100, 1200, 1900, 890000 \\ \Rightarrow \quad & \text{median} = (1050 + 1100)/2 = 1075 \end{aligned}$$

- Notice that the median is unaffected by the size of the largest claim.
- The median is appropriate for ordinal qualitative data as well as quantitative data.

Mode: The mode is simply the most commonly occurring value. This quantity is not unique; there may be multiple modes.

- In the insurance claims data, all values were distinct, so all values were modes.
- A histogram of the apgar5 data is given below. From this plot it is easy to see that the mode is 8. The mean is 6.25 and the median is 7.



- The mode is especially useful for describing qualitative variables or quantitative variables that take on a small number of possible values.
  - The modal gender in this class is female. The modal academic program afiliation is BHSI.
- If two values occur more than others but equally frequently, we say the data are **bimodal**, or more generally multimodal.
  - The term bimodal is also sometimes used to describe distributions in which there are two peaks, not necessarily of the same height. E.g.:

Percentiles: Earlier, we noted that the median is the 50th percentile. We gave the definition of a percentile on p.20. A procedure for obtaining the p^{th} percentile of a data set of size n is as follows:

Step 1: Arrange the data in ascending (increasing) order.

Step 2: Compute an index i as follows: $i = \frac{p}{100}n$.

Step 3:

- If i is an integer, the p^{th} percentile is the average of the i^{th} and $(i+1)^{\text{th}}$ smallest data values.
- If i is not an integer then round i up to the nearest integer and take the value at that postion.

- For example, consider the 9 insurance claims again:

  sorted data: $380, 600, 690, 890, 1050, 1100, 1200, 1900, 890000$

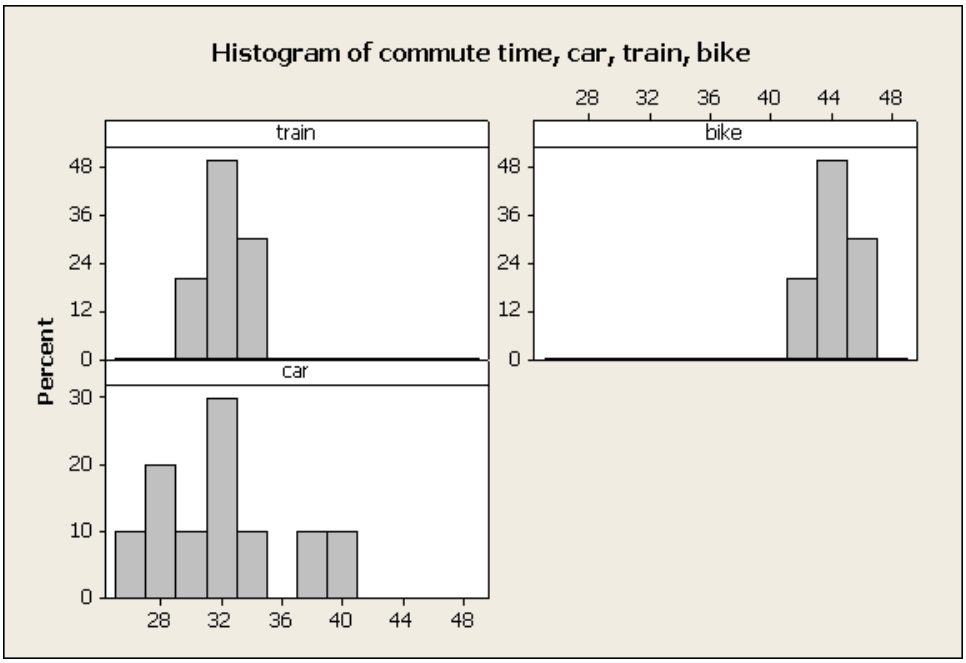
  - For the $p = 10$th percentile, $i = pn/100 = 10(9)/100 = .9$. Round up to 1, so that the 10th percentile is the first sorted value, or 380.
  - For the $p = 75$th percentile, $i = pn/100 = 75(9)/100 = 6.75$. Round up to 7, so that the 75th percentile is the seventh sorted value, or 1200.

- Percentiles not only give locate the center of a distribution (e.g., the median), but also other locations in a distribution.

Measures of Dispersion:

The two most important aspects of a unimodal distribution are the location (or central tendency) and the spread (or dispersion).

- E.g., consider the time it takes to commute to work by car, train, and bike. Suppose these are the distributions of commute time by these modes of transportation.



Comparison	Location	Spread
Train & Car	Same	Different
Train & Bike	Different	Same
Car & Bike	Different	Different

Measures of Dispersion: Range, Interquartile Range, Variance and Standard Deviation, Coefficient of Variation.

Range: The range is simple the maximum value minus the minimum value in the data set:

$$\text{range} = \max - \min.$$

- The range of the 9 insurance claims was $890,000 - 380 = 889,620$.

Inter-quartile Range: The range only depends upon the minimum and maximum, so it is heavily influenced by the extremes.

- That is, the range may not reflect the spread in most of the data.

The inter-quartile range is the difference between the third quartile (75th %'ile) and the first quartile (25th %'ile). That is,

$$\text{IQR} = Q_3 - Q_1.$$

- For the insurance claim data, we computed the 75th %'ile as $Q_3 = 1200$. To get the 25th percentile, $i = pn/100 = 25 * 9/100 = 2.25$. Rounding up, we take the third smallest value, or $Q_1 = 690$. Thus

$$\text{IQR} = \$1200 - \$690 = \$510.$$

Variance and Standard Deviation: The most important measures of dispersion are the variance and its square root, the standard deviation.

- Since the variance is just the square of the standard deviation, these quantities contain essentially the same information, just on different scales.

The range and IQR each take only two data points into account.

> *How might we measure the spread in the data accounting for the value of every observation?*

Consider the insurance claim data again:

Observation Number (i)	x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	1100	99756.67	-98656.7	9733137878
2	1900	99756.67	-97856.7	9575927211
3	600	99756.67	-99156.7	9832044544
4	890	99756.67	-98866.7	9774617778
5	690	99756.67	-99066.7	9814204444
6	890000	99756.67	790243.3	6.24485E+11
7	380	99756.67	-99376.7	9875721878
8	1200	99756.67	-98556.7	9713416544
9	1050	99756.67	-98706.7	9743006044
Sum=	897810		0	7.02547E+11
Mean=	99756.67		0	78060733578

One way to measure spread is to calculate the mean and then determine how far each observation is from the mean.

$$\text{Mean:} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{9} x_i = \frac{1}{9}(1100 + 1900 + \cdots + 1050) = 99756.67.$$

How far an observation is from the mean is quantified by the difference between that observation and the mean: $x_i - \bar{x}$. In the entire data set, we have 9 of these:

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_9 - \bar{x}.$$

One idea is to compute the average of these **deviations from the mean.** That is, compute

$$\frac{1}{9}\{(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_9 - \bar{x})\} = \frac{1}{9}\sum_{i=1}^{9}(x_i - \bar{x}).$$

Problem: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ (always!).

- Deviations from the mean always necessarily sum to zero. The positive and negative values cancel each other out.

Solution: Make all of the deviations from the mean positive by squaring them before averaging.

That is, compute

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_9 - \bar{x})^2$$

and then average. This gives the quantity

$$\frac{1}{9}\sum_{i=1}^{9}(x_i - \bar{x})^2 = 78060733578.$$

If $x_1, x_2, \ldots, x_9$ is the entire population, then $\bar{x} = \mu$, the population mean, and the population size is $N = 9$. In this case, our formula becomes

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2,$$

which is called the **population variance** of $x_1, \ldots, x_N$, and is usually denoted by σ^2.

Why is this called σ^2 rather than σ? Why the 2 exponent?

Because this is the average *squared* deviation from the mean (in the claims data example, the units of this quantity are squared dollars).

- To put the variance on the same scale as the original data, we sometimes prefer to work with the **population standard deviation** which is denoted as σ and is just the square root of the population variance σ^2:

$$\text{population standard deviation:} \quad \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}.$$

Suppose now that $x_1, \ldots, x_n$ are sample values.

How do we compute a sample variance to estimate the population variance, σ^2?

We could simply use $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. However, for reasons we'll discuss later, it turns out that it is better to define the **sample variance** as

$$\text{sample variance:} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- The sample standard deviation is simply the square root of this quantity:

$$\text{sample standard deviation:} \quad s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- The sample variance of the 9 insurance claims is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{9-1}\{(x_1 - \bar{x})^2 + \cdots (x_9 - \bar{x})^2\}$$
$$= 87818325275 \quad \text{(squared dollars)}$$

  and the sample standard deviation is

$$s = \sqrt{87818325275} = \$296,341.57$$

A Note on Computation:

The formula for s^2 that we just presented,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

conveys clearly the logic of the standard deviation: it is an average (in some sense) of the squared deviations from the mean. However, it is not a good formula to use for computing the SD because it

- is hard to use; and
- it tends to lead to round-off errors.

For computing, an equivalent but better formula is

$$s^2 = \frac{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2}{n-1}.$$

- When using this formula or any other that requires a series of calculations, keep all intermediate steps in the memory of your calculator until the end to avoid round-off errors.

Coefficient of Variation: In some cases the variance of a variable changes with its mean.

- For example, suppose we are measuring the weights of children of various ages.

  5 year old children (relatively light, on average)
  15 year old children (much heaver, on average)

  Clearly, there's much more variability in the weights of 15 year olds, but a valid question to ask is "Do 15 year old children's weights have more variabilty relative to their average?"

The coefficient of variation allows such comparisons to be made:

$$\text{population CV} = \frac{\sigma}{\mu} \times 100\%,$$
$$\text{sample CV} = \frac{s}{\bar{x}} \times 100\%.$$

- From current CDC data available on the web, I obtained standard deviations and means for the weights (in kg) of 5 and 15 year old male children as follows:

Age	s	$\bar{x}$	CV
5	2.74	18.39	0.15
15	12.05	56.28	0.21

- Thus, 15 year olds' weights are more variable relative to their average weight than 5 year olds.

- Note that the CV is a unitless quantity.

Mean and Variance (or SD) for Grouped Data:

- **Example – Lead Content in Boston Drinking Water**

  Consider the following data on the lead content (mg/liter) in 12 samples of drinking water in the city of Boston, MA.

$$\begin{aligned} \text{data:} \quad & .035, .060, .055, .035, .031, .039, .038, .049, .073, .047, .031, .016 \\ \text{sorted data:} \quad & .016, .031, .031, .035, .035, .038, .039, .047, .049, .055, .060, .073 \end{aligned}$$

- Notice that there are some values here that occur more than once.

Consider how the mean is calculated in such a situation:

$$\begin{aligned} \bar{x} &= \frac{.016 + .031 + .031 + .035 + .035 + .038 + \cdots + .073}{12} = .042 \\ &= \frac{.016(1) + .031(2) + .035(2) + .038(1) + \cdots + .073(1)}{(1) + (2) + (2) + (1) + \cdots + (1)} \\ &= \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i} \end{aligned}$$

where

$$\begin{aligned} k &= \text{the number of distinct values in the data} \\ m_i &= \text{the } i^{\text{th}} \text{ distinct value} \\ f_i &= \text{the frequency with which } m_i \text{ occurs} \end{aligned}$$

Similarly, consider the sample variance:

$$\begin{aligned} s^2 &= \Big\{(.016 - .042)^2 + (.031 - .042)^2 + (.031 - .042)^2 + (.035 - .042)^2 + (.035 - .042)^2 + \\ &\quad \cdots + (.073 - .042)^2\Big\}/(12 - 1) = .015 \\ &= \frac{(.016 - .042)^2(1) + (.031 - .042)^2(2) + (.035 - .042)^2(2) + \cdots + (.073 - .042)^2(1)}{[(1) + (2) + (2) + \cdots + (1)] - 1} \\ &= \frac{\sum_{i=1}^{k}(m_i - \bar{x})^2 f_i}{[\sum_{i=1}^{k} f_i] - 1} \end{aligned}$$

- **Another Example: Apgar Scores of Low Birthweight Infants**

  Here is a frequency distribution of the Apgar scores for 100 low birthweight infants in data set lowbwt.

Apgar Score	Frequency
0	6
1	1
2	3
3	4
4	5
5	10
6	11
7	23
8	24
9	13
Total=	100

  Using the formula for the mean for grouped data we have

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i} \\ &= \frac{0(6) + 1(1) + 2(3) + \cdots + 9(13)}{100} = 6.25\end{aligned}$$

  which agrees with the value we reported previously for these data.

  Similarly, the sample SD is

$$\begin{aligned}s &= \sqrt{\frac{\sum_{i=1}^{k}(m_i - \bar{x})^2 f_i}{[\sum_{i=1}^{k} f_i] - 1}} \\ &= \sqrt{\frac{(0-6.25)^2(6) + (1-6.26)^2(1) + (2-6.25)^2(3) + \cdots + (9-6.25)^2(13)}{100-1}} \\ &= 2.43\end{aligned}$$

z-Scores and Chebychev's Inequality:

The National Center for Health Statistics at the CDC gives the following estimate of the body mass index ($\frac{\text{weight}}{\text{height}^2}$) for 15 year old boys:

$$\bar{x} = 19.83$$

Suppose that a particular 15 year old boy, Fred, has a BMI equal to 25.

How overweight is Fred?

We know he is heavier than average for his age/gender group, but how much heavier?

- Relative to the variability in BMI for 15 year old boys in general, Fred's BMI may be close to the mean or far away.

Case 1: Suppose $s = 10$.

- This implies that the typical deviation from the mean is about 10. Fred's deviation from the mean is 5.17, so Fred doesn't seem to be unusually heavy.

Case 2: Suppose $s = 2$.

- This implies that the typical deviation from the mean is about 2. Fred's deviation from the mean is 5.17, so Fred does seems to be unusually heavy.

- Thus, the extremeness of Fred's BMI is quantified by its distance from the mean BMI *relative to the SD* of BMI.

The z-score gives us this kind of information.

$$z_i = \frac{x_i - \bar{x}}{s}$$

where

x_i = value of the variable of interest for subject i,

$\bar{x}$ = sample mean

s = sample standard deviation

Case 1: $z = \frac{25-19.83}{10} = .517$. Fred's BMI is .517 SD's above the mean.

Case 2: $z = \frac{25-19.83}{2} = 2.585$. Fred's BMI is 2.585 SD's above the mean.

- From NCHS data, the true SD for 15 year old boys is $s = 3.43$. So, Fred's BMI is $z = \frac{25-19.83}{3.43} = 1.51$ SD's above the mean.

How extreme is a z score of 2? 3? -1.5?

An exact answer to this question depends upon the distribution of the variable you are interested in.

However, a partial answer that applies to any variable is provided by Chebychev's inequality.

Chebychev's Theorem: At least $\left(1 - \frac{1}{k^2}\right) \times 100\%$ of the values of any variable must be within k SDs of the mean, for any $k > 1$.

This results implies (for example):

- At least 75% of the observations must be within 2 SDs, since for $k = 2$

$$\left(1 - \frac{1}{k^2}\right) \times 100\% = \left(1 - \frac{1}{2^2}\right) \times 100\% = \left(1 - \frac{1}{4}\right) \times 100\% = 75\%.$$

  - For the BMI example, we'd expect at least 75% of 15 year old males to have BMIs between $\bar{x} - 2s = 19.83 - 2(3.43) = 12.97$ and $\bar{x} + 2s = 19.83 + 2(3.43) = 26.69$.

- At least 89% of the observations must be within 3 SDs, since for $k = 3$

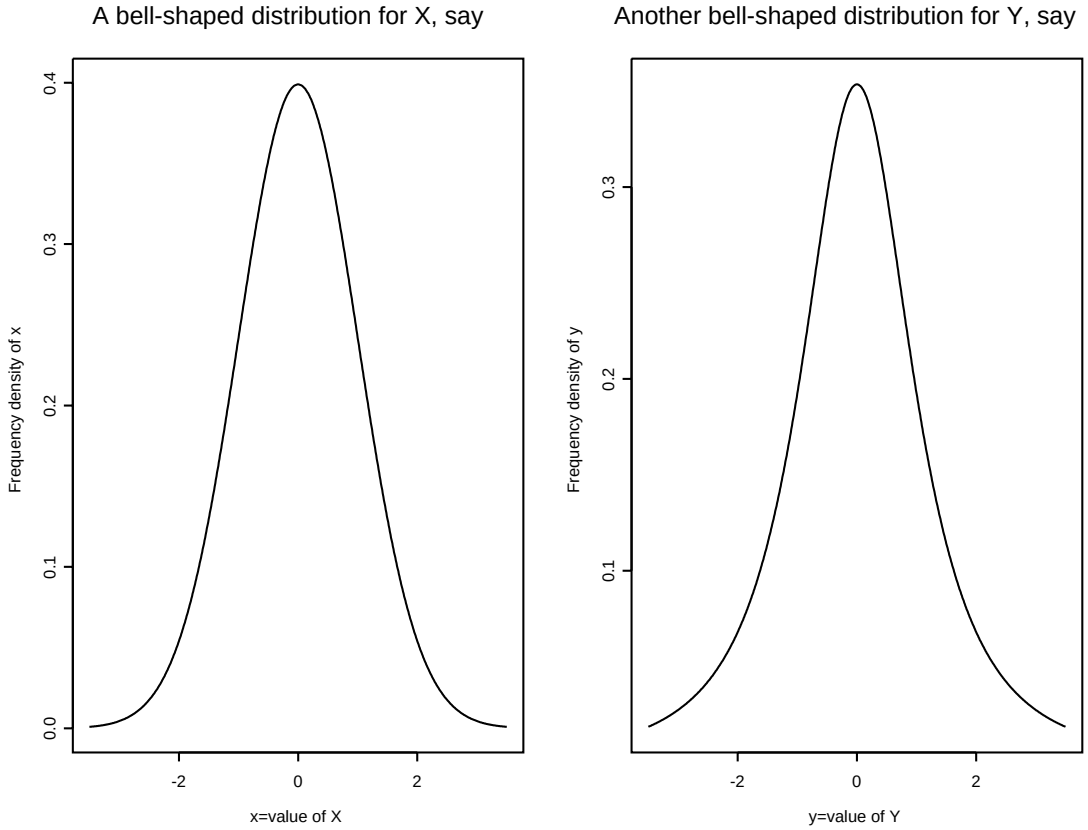
$$\left(1 - \frac{1}{k^2}\right) \times 100\% = \left(1 - \frac{1}{3^2}\right) \times 100\% = \left(1 - \frac{1}{9}\right) \times 100\% = 89\%.$$

  - For the BMI example, we'd expect at least 89% of 15 year old males to have BMIs between $\bar{x} - 3s = 19.83 - 3(3.43) = 9.54$ and $\bar{x} + 3s = 19.83 + 3(3.43) = 30.12$.

- Note that Chebychev's Thm just gives a lower bound on the percentage falling within k SDs of the mean. At least 75% should fall within 2 SDs, but perhaps more.

  - Since it only gives a bound and not a more exact statement about a distribution, Chebychev's Thm is of limited practical value.

We can make a much more precise statement if we know that the distribution of the variable in which we're interest is bell-shaped. That is, shaped roughly like this:



- Think of the above pictures as idealized histograms as the sample size grows toward infinity.
- One particular bell-shaped distribution is the **normal distribution**, which is also known as the **Gaussian distribution**.
  - The normal distribution is particularly important in statistics, but it is not the only possible bell-shaped distribution. The distribution above left is normal, the one above right is similar, but not exactly normal (notice difference in the tails).

For data that follow the normal distribution, the following precise statements can be made:

- Excatly 68% of the observations lie within 1 SD of the mean.
- Exactly 95% of the observations lie within 2 SDs of the mean.
- Exactly 99.7% of the observations lie within 3 SDs of the mean.

In fact, for normally distributed data we can calculate the percentage of the observations that fall in any range whatsoever.

This is very helpful if we know our data are normally distributed.

However, even if the data aren't known to be exactly normal, but are known to be bell-shaped, then the exact results stated above will be approximately true. This is known as the *empirical rule.*

Empirical rule: for data following a bell-shaped distribution:

- Approximately 68% of the observations will fall with 1 SD of the mean.
- Approximately 95% of the observations will fall with 2 SDs of the mean.
- Nearly all of the observations will fall with 3 SDs of the mean.

BMIs of 15 Year-old Boys:

- At age 15, suppose that BMI follows an approximately bell-shaped distribution.
  - Then we would expect approximately 68% of 15 year old boys to have BMIs falling in the interval $(16.40, 23.26) = \bar{x} \pm 1s$. Fred's BMI was 25, so his BMI is more extreme than two-thirds of boys his age.
  - We would expect 95% of 15 year-old boys to have BMIs falling in the interval $(12.97, 26.69) = \bar{x} \pm 2s$ and nearly all to fall in the interval $(9.54, 30.12) = \bar{x} \pm 3s$. (Compare these results with the Chebychev bounds).
- In fact, BMI is probably not quite bell-shaped for 15 year olds. It may be for 5 year olds, but by age 15, there are many obese children who probably skew the distribution to the right (lots of large values in the right tail). Therefore, the empirical rule may be somewhat inaccurate for this variable.

Introduction to Probability*

- Note: we're going to skip ch.s 4 & 5 for now, but we'll come back to them later.

We all have an intuitive notion of probability.

- "There's a 75% chance of rain today."
- "The odds of Smarty Jones winning the Kentucky Derby are 2 to 1."
- "The chances of winning the Pick-5 Lottery game are 1 in 2.3 million."
- "The probability of being dealt four of a kind in a 5 card poker hand is 1/4164."

All of these statements are examples of quantifying the uncertainty in a random phenomenon. We'll refer to the random phenomenon of interest as the *experiment*, but don't confuse this use with an experiment as a type of research study.

- The experiments in the examples above are
  - An observation of today's weather
  - The results of the Kentucky Derby
  - A single play of the Pick-5 Lottery game
  - The rank of a 5-card poker hand dealt from a shuffled deck of cards

* Read Ch.6 of our text.

An experiment generates an outcome through some random process.

Experiment	Outcome
Weather	Rains, Does not rain
Kentucky Derby	Smarty Jones wins, places, shows,..., does not finish
Lottery	Win, Lose
Poker Hand	Royal Flush, Straight Flush, Four-of-a-kind, ...

- Set of outcomes is called the **sample space** and should consist of mutually exclusive, exhaustive set of outcomes.

An **event** is some description of the outcome of an experiment whose probability is of interest.

- A variety of events can be defined based on the outcome of a given experiment:
- E.g., Events that could be defined regarding the outcome of the Kentucky Derby:
  - Smarty Jones finishes
  - Smarty Jones finishes third or better (wins, places, or shows)
  - Smarty Jones wins.
- Events of interest need not be mutually exclusive or exhaustive.
- The terms "chance(s)", "likelihood", and "probability" are basically synonymous ways to describe the probability of an event. We denote the probability of an event A by

$$P(A)$$

  - The odds of an event describes probability too, but is a bit different. The odds of an event A are defined as

$$\text{odds}(A) = \frac{P(A)}{P(A^c)},$$

    where A^c denotes the event that A does not occur, which is known as the **complement** of A.

A number of different operations can be defined on events.

- One is the complement: A^c denotes the event that A does not occur.
- The **union** of events A and B is denoted

$$A \cup B.$$

  The union of A and B is the event that A occurs or B occurs (or both).

  - $\cup$ can be read as "or" (inclusive).

- The **intersection** of events A and B is denoted

$$A \cap B.$$

  The intersection of A and B is the event that A occurs and B occurs.

  - $\cap$ can be read as "and".

The following *Venn diagrams* describe these operations pictorially.

There are a number of legitimate ways to assign probabilities to events:

- the classical method
- the relative frequency method
- the subjective method

Whatever method we use, we require

1. The probability assigned to each experimental outcome must be between 0 and 1 (inclusive). That is, if O_i represents the i^{th} possible outcome, we must have

$$0 \leq P(O_i) \leq 1, \quad \text{for all } i.$$

   – Probabilities are between 0 and 1, but they are often expressed as percentages by multiplying by 100%. That is, to say there is a .75 chance of rain is the same as saying the chance of rain is 75%.

2. The sum of the probabilities for all experimental outcomes must equal 1. That is, for n mutually exclusive, exhaustive outcomes $O_1, \ldots, O_n$, we must have

$$P(O_1) + P(O_2) + \cdots + P(O_n) = 1.$$

Classical Method: When all n experimental outcomes are equally likely, we assign eqach outcome a probability of $1/n$.

- E.g., when tossing a fair coin, there are $n = 2$ equally likely outcomes, each with probability $1/n = 1/2$.

- E.g., If we pick a card from a well-shuffled deck and observe its suit, then there are $n = 4$ possible outcomes, so

$$P(\heartsuit) = P(\spadesuit) = P(\diamondsuit) = P(\clubsuit) = 1/n = 1/4.$$

The classical method is really a special case of the more general **Relative Frequency Method**. The probability of an event is the relative frequency with which that event occurs if we were to repeat the experiment a very large number of times under identical circumstances.

- I.e., if the event A occurs m times in n identical replications of an experiment, then

$$P(A) = \frac{m}{n} \text{when } n \to \infty.$$

- Suppose that the gender ratio at birth is 50:50. That is, suppose that giving birth to a boy and giving birth to a girl are equaly likely events. Then by the clasical method

$$P(\text{Girl}) = \frac{1}{2}.$$

  This is also the long run relative frequency. As $n \to \infty$ we should expect that

$$\frac{\text{number of girls}}{\text{number of births}} \to \frac{1}{2}.$$

There are several rules of probability associated with the union, intersection, and complement operations on events.

Addition Rule: For two events A and B

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Venn Diagram:

Example Consider the experiment of having two children and let

$$A = \text{event that first child is a girl}$$

$$B = \text{event that second child is a girl}$$

- Assume $P(A) = P(B) = 1/2$ (doesn't depend on birth order and gender of second child not influenced by gender of first child).

Then the probability of having at least one girl is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{2} - P(A \cap B)$$

But what's $P(A \cap B)$ here?

One way to determine this is by enumerating the set of equally likely outcomes of the underlying experiment here.

The experiment is observing the genders of two children. It has sample space (set of possible outcomes):

$$\{(M, M), (M, F), (F, M), (F, F)\}$$

which are all equally likely (have probability 1/4 each).

- The probability of an event is the sum of the probabilities of the outcomes satisfying the event's definition.

Here, the event $A \cap B$ corresponds to the outcome (F, F) so

$$P(A \cap B) = \frac{1}{4}$$

and

$$P(A \cup B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$$

- Notice that this agrees with the answer we would have obtained by summing the probabilities of the outcomes corresponding to at least one girl:

$$P(A \cup B) = P\{(M, F)\} + P\{(F, M)\} + P\{(F, F)\} = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}.$$

Complement Rule: For an event A and its complement A^c

$$P(A^c) = 1 - P(A).$$

- This is simply a consequence of the addition rule and the facts that

$$\begin{aligned} & P(A \cup A^c) = P(\text{entire sample space}) = 1, \\ \text{and} \quad & P(A \cap A^c) = P(A \text{ and not } A \text{ occur}) = 0 \end{aligned}$$

  Thus, by the addition rule

$$1 = P(A \cup A^c) = P(A) + P(A^c) - \underbrace{P(A \cap A^c)}_{=0} = P(A) + P(A^c)$$

$$\Rightarrow \qquad P(A) = 1 - P(A^c)$$

A third rule is the multiplication rule, but for that we need the definitions of conditional probability and statistical independence.

Conditional Probability:

For some events, whether or not one event has occurred is clearly relevant to the probability that a second event will occur.

- We just computed that the probability of having at least one girl in two births as $\frac{3}{4}$.

  Now suppose I know that my first child was a boy.

- Clearly, knowing that I've had a boy affects the chances of having at least one girl (it decreases them). Such a probability is known as a conditional probability.

The conditional proabability of an event A given that another event B has occurred is denoted $P(A|B)$ where $|$ is read as "given".

Independence of Events Two events A and B are independent if knowing that B has occurred gives no information relevant to whether or not A will occur (and vice versa). In symbols A and B are independent if

$$P(A|B) = P(A).$$

Multiplication Rule: The *joint probability* of two events $P(A \cap B)$ is given by

$$P(A \cap B) = P(A|B)P(B),$$

or since the A and B can switch places

$$P(A \cap B) = P(B \cap A) = P(B|A)P(A)$$

- Note that this relationship can also be written as

  $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  as long as $P(B) \neq 0$, or

  $$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

  as long as $P(A) \neq 0$.

Example: Again, the probability of having at least one girl in two births is $\frac{3}{4}$. Now suppose the first child is known to be a boy, and the second child's gender is unknown.

What is the conditional probability of at least one girl given that the first child is a boy?

Again, let

$$A = \text{event that first child is a girl}$$

$$B = \text{event that second child is a girl}$$

Then we are interested here in $P(A \cup B|A^c)$.

By the multiplication rule

$$P(A \cup B|A^c) = \frac{P\{(A \cup B) \cap A^c\}}{P(A^c)}.$$

We known that the probability that the first child is a girl is $P(A) = \frac{1}{2}$, so

$$P(A^c) = 1 - P(A) = 1 - \frac{1}{2} = \frac{1}{2}.$$

In addition, the probability in the numerator, $P\{(A \cup B) \cap A^c\}$, is the probability that at least one child is a girl (the event $A \cup B$) *and* the first child is a boy (the event A^c).

Both of these events can happen simultaneously only if the first child is a boy and the second child is a girl. That is, only if the outcome of the experiment is $\{(M, F)\}$. Thus,

$$P\{(A \cup B) \cap A^c\} = P(A^c \cap B) = P\{(M, F)\} = \frac{1}{4}.$$

Therefore, the conditional probability of at least one girl given that the first child is a boy is

$$P(A \cup B|A^c) = \frac{P\{(A \cup B) \cap A^c\}}{P(A^c)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

Another Example:

Suppose that among US adults, 1 in 3 obese individuals has high blood pressure, while 1 in 7 normal weight individuals has high blood pressure.

Suppose also that the percentage of US adults who are obese, or **prevalence** of obesity, is 20%.

What is the probability that a randomly selected US adult is obese and has high blood pressure?

Let

A = event that a randomly selected US adult is obese

B = event that a randomly selected US adult has high b.p.

Then the information given above is

$$P(A) = \frac{1}{5} \qquad P(B|A) = \frac{1}{3} \qquad P(B|A^c) = \frac{1}{7}.$$

By the multiplication rule, the probability that a randomly selected US adult is obese and has high blood pressure is

$$P(A \cap B) = P(B \cap A) = P(B|A)P(A) = \left(\frac{1}{3}\right)\left(\frac{1}{5}\right) = \frac{1}{15}.$$

Note that in general given that an event A has occurred, either B occurs, or B^c must occur, so the complement rule applies to conditional probabilities too:

$$P(B^c|A) = 1 - P(B|A).$$

With this insight in hand, we can compute all other joint probabilities relating to obesity and high blood pressure:

- The probability that a randomly selected US adult is obese and does not have high blood pressure is

$$P(A \cap B^c) = P(B^c \cap A) = P(B^c|A)P(A) = [1 - P(B|A)]P(A) = \left(\frac{2}{3}\right)\left(\frac{1}{5}\right) = \frac{2}{15}.$$

- The probability that a randomly selected US adult is not obese and does have high blood pressure is

$$P(A^c \cap B) = P(B \cap A^c) = P(B|A^c)P(A^c) = P(B|A^c)[1 - P(A)] = \left(\frac{1}{7}\right)\left(\frac{4}{5}\right) = \frac{4}{35}.$$

- The probability that a randomly selected US adult is not obese and does not have high blood pressure is

$$P(A^c \cap B^c) = P(B^c \cap A^c) = P(B^c|A^c)P(A^c) = [1 - P(B|A^c)][1 - P(A)] = \left(\frac{6}{7}\right)\left(\frac{4}{5}\right) = \frac{24}{35}.$$

These results can be summarized in a table of joint probabilities:

		Obese: Yes (event A)	Obese: No (event (A^c))	
High B.P.	Yes (event B)	$\frac{1}{15}$	$\frac{4}{35}$	$\frac{19}{105}$
	No (event B^c)	$\frac{2}{15}$	$\frac{24}{35}$	$\frac{86}{105}$
		$\frac{3}{15} = \frac{1}{5}$	$\frac{28}{35} = \frac{4}{5}$	1

Independence: Two events A and B are said to be **independent** if knowing whether or not A has occurred tells us nothing about whether or not B has or will occur and vice versa.

In symbols, A and B are independent if

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B).$$

- Note that under independence of A and B, the multiplication rule becomes
$$P(A \cap B) = P(A|B)P(B) = P(A)P(B)$$
and the addition rule becomes
$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A)P(B).$$

- Note also that the terms *mutually exclusive* and *independent* are often confused, but they mean different things.
  - Mutually exclusive events A and B are events that can't happen simultaneously. Therefore, if I know A has occurred, that tells me something about B; namely, that B can't have occurred. So mutually exclusive events are necessarily **dependent** (not independent).

Obesity and High Blood Pressure Example: The fact that obesity and high b.p. are not independent can be verified by checking that

$$\frac{1}{3} = P(B|A) \neq P(B) = \frac{19}{105}.$$

Alternatively, we can check independence by checking whether $P(A \cap B) = P(A)P(B)$. In this example,

$$0.0667 = \frac{1}{15} = P(A \cap B) \neq P(A)P(B) = \left(\frac{1}{5}\right)\left(\frac{19}{105}\right) = 0.0362$$

Bayes' Theorem:

We have seen that when two events A and B are dependent, then $P(A|B) \neq P(A)$.

- That is, the information that B has occurred affects the probability that A will occur.

Bayes' Theorem provides a way to use new information (event B has occurred) to go from our probability before the new information was available ($P(A)$, which is called the **prior probability**) to a probability that takes the new information into account ($P(A|B)$, which is called the **posterior probability**).

- Bayes' Theorem allows us to take the information about $P(A)$ and $P(B|A)$ and compute $P(A|B)$.

Obesity and High B.P. Example:

Recall

A = event that a randomly selected US adult is obese

B = event that a randomly selected US adult has high b.p.

and

$$P(A) = \frac{1}{5} \qquad P(B|A) = \frac{1}{3} \qquad P(B|A^c) = \frac{1}{7}.$$

Suppose that I am a doctor seeing the chart of a patient, and the only information contained there is that the patient has high b.p.

Assuming this patient is randomly selected from the US adult population, what is the probability that the patient is obese?

That is, what is $P(A|B)$?

By the multiplication rule, we know that

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \qquad (*)$$

Let's examine the numerator and denominator of this expression and see if we can use the information available to compute these quantities.

First, notice that the denominator is $P(B)$, the probability of high blood pressure. If a random subject has high b.p., then the subject either

a. has high b.p. and is obese, or
b. has high b.p. and is not obese.

That is,

$$B = (B \cap A) \cup (B \cap A^c)$$

so, by the addition rule

$$P(B) = P(B \cap A) + P(B \cap A^c) - \underbrace{P\{(B \cap A) \cap (B \cap A^c)\}}_{=0}$$

Therefore,

$$P(B) = P(B \cap A) + P(B \cap A^c).$$

- This relationship is sometimes called the *law of total probability*, and is just based on the idea that if B occurs, it has to occur with either A or A^c.

So, now (*) becomes

$$P(A|B) = \frac{P(A \cap B)}{P(B \cap A) + P(B \cap A^c)}. \qquad (**)$$

Now consider the numerator, $P(A \cap B)$. By the multiplication rule and using the fact that $(A \cap B) = (B \cap A)$, we have

$$P(A \cap B) = P(B \cap A) = P(B|A)P(A),$$

which is useul because we know these quantities.

Applying the same logic to the two joint probabilities in the denominator of (**), we have that

$$P(B \cap A) = P(B|A)P(A) \quad \text{and} \quad P(B \cap A^c) = P(B|A^c)P(A^c).$$

Therefore, (**) becomes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}. \qquad (\dagger)$$

- Equation (†) is known as Bayes' Theorem.

In the example, Bayes' Theorem tells us that the probability that the high b.p. patient is obese is

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{(1/3)(1/5)}{(1/3)(1/5) + (1/7)(4/5)} = \frac{1/15}{19/105} = 0.368 \end{aligned}$$

In the example above, we used the law of total probability to compute $P(B)$ as

$$P(B) = P(B \cap A) + P(B \cap A^c) = P(B|A)P(A) + P(B|A^c)P(A^c)$$

where A and A^c were mutually exclusive, exhaustive events.

- Bayes' Theorem generalizes to apply to the situation in which we have several mutually exclusive, exhaustive events.

Let $A_1, A_2, \ldots, A_k$ be k mutually exclusive, exhaustive events. Then for any of the events A_i, $i = 1, \ldots, k$, Bayes' Theorem becomes:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots P(B|A_k)P(A_k)}.$$

Another Example — Obesity and Smoking Status:

Let

B = event that a randomly selected US adult is obese

A_1 = event that a randomly selected US adult has never smoked

A_2 = event that a randomly selected US adult is an ex-smoker

A_3 = event that a randomly selected US adult is a current smoker

and suppose

$$P(B) = .209, \qquad P(B|A_1) = .208, \qquad P(B|A_2) = .239, \qquad P(B|A_3) = .178$$
$$P(A_1) = 0.520, \qquad P(A_2) = 0.250, \qquad P(A_3) = 0.230.$$

Given that a randomly selected US adult is obese, what's the probability that he/she is a former smoker?

By Bayes' Theorem

$$P(A_2|B) = \frac{P(B|A_2)P(A_2)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)}$$
$$= \frac{(.239)(.250)}{(.208)(.520) + (.239)(.250) + (.178)(.230)} = .286$$

- Note that the denominator is just $P(B)$, so since we happen to know $P(B)$ here, we could have simplified our calculations as

$$P(A_2|B) = \frac{P(B|A_2)P(A_2)}{P(B)}$$
$$= \frac{(.239)(.250)}{.209} = .286$$

Diagnostic Tests

One important application of Bayes' Theorem is to diagnostic or screening tests.

- **Screening** is the application of a test to individuals who have not yet exhibited any clinical symptoms in order to classify them with respect to their probability of having a particular disease.
  - Examples: Mammograms for breast cancer, Pap smears for cervical cancer, Prostate-Specific Antigen (PSA) Test for prostate cancer, exercise stress test for coronary heart disease, etc.

Consider the problem of detecting the presence or absence of a particular disease or condition.

Suppose there is a "gold standard" method that is always correct.

- E.g., surgery, biopsy, autopsy, or other expensive, time-consuming and/or unpleasant method.

Suppose there is also a quick, inexpensive screening test.

- Ideally, the test should correctly classify individuals as positive or negative for the disease. In practice, however, tests are subject to misclassification errors.

Definitions:

- A test result is a **true positive** if it is positive and the individual has the disease.
- A test result is a **true negative** if it is negative and the individual does not have the disease.
- A test result is a **false positive** if it is positive and the individual does not have the disease.
- A test result is a **false negative** if it is negative and the individual does have the disease.
- The **sensitivity** of a test is the conditional probability that the test is positive, given that the individual has the disease.
- The **specificity** of a test is the conditional probability that the test is negative, given that the individual does not have the disease.
- The **predictive value of a positive test** is the conditional probability that an individual has the disease, given that the test is positive.
- The **predictive value of a negative test** is the conditional probability that an individual does not have the disease, given that the test is negative.

Notation: Let

A = event that a random individual's test is positive

B = event that a random individual has the disease

Then

sensitivity =	predictive value positive =
specificity =	predicitve value negative =

Estimating the Properties of a Screening Test:

Suppose data are obtained to evaluate a screening test where the true disease status of each patient is known. Such data may be displayed as follows:

		Truth: Diseased (event B)	Truth: Not Diseased (event B^c)	
Test Result	+ (event A)	a	b	$n_{1.}$
	− (event A^c)	c	d	$n_{2.}$
		$n_{.1}$	$n_{.2}$	n

What properties of the screening test can be estimated if the data are obtained:

1. *from a random sample of n subjects?*
2. *from random samples of $n_{.1}$ diseased and $n_{.2}$ nondiseased subjects?*
3. *from random samples of $n_{1.}$ subjects with positive test results and $n_{2.}$ subjects with negative results?*

1. Suppose a random sample of n subjects is obtained, and each subject is tested via both the screening test and the gold standard.

   In this case,

   $$\text{estimated sensitivity} =$$
   $$\text{estimated specificity} =$$
   $$\text{estimated predictive value positive} =$$
   $$\text{estimated predictive value negative} =$$

2. Suppose that random samples of $n_{.1}$ diseased and $n_{.2}$ nondiseased subjects are obtained, and each subject is tested with the screening test.

   In this case,

   $$\text{estimated sensitivity} =$$
   $$\text{estimated specificity} =$$

   but predictive value positive and negative cannot be estimated directly without additional information about the probability (prevalence) of disease.

3. Suppose now that random samples of $n_{1.}$ subjects with positive screening test results and $n_{2.}$ subjects with negative screening test results are obtained. Each subject is then tested with the gold standard approach.

   In this case,

   $$\text{estimated predictive value positive} =$$
   $$\text{estimated predictive value negative} =$$

   but sensitivity and specificity cannot be estimated directly without additional information about the probability of a positive test result.

Notice that only in case 1 is it possible to obtain estimates of all four quantities from simple proportions in the contingency table.

- However, this approach is not particularly quick, easy or efficient because, for a rare disease, it will require a large n to obtain a sufficient sample of truly diseased subjects.

- Approach 2 is generally easiest, and predictive values can be computed from this approach using Bayes' Theorem if the prevalence of the disease is known as well.

Suppose we take approach 2. As before, let

$$A = \text{event that a random individual's test is positive}$$

$$B = \text{event that a random individual has the disease}$$

Suppose $P(B)$, the prevalence of disease, is known. In addition, suppose the sensitivity $P(A|B)$ and specificity $P(A^c|B^c)$ are known (or have been estimated as on the previous page).

Then, according to Bayes' Theorem, $P(B|A)$, the predictive value of a positive test result, is given by

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

Similarly, $P(B^c|A^c)$, the predictive value of a negative test result, is given by

$$P(B^c|A^c) = \frac{P(A^c|B^c)P(B^c)}{P(A^c|B^c)P(B^c) + P(A^c|B)P(B)}$$

Suppose that a new screening test for diabetes has been developed. To establish its properties, $n_{\cdot 1} = 100$ known diabetics and $n_{\cdot 2} = 100$ known non-diabetics were tested with the screening test. The following data were obtained:

		Truth	
		Diabetic (event B)	Nondiabetic (event B^c)
Test Result	+ (event A)	80	10
	− (event A^c)	20	90
		100	100

- Suppose that it is known that the prevalence of diabetes is $P(B) = .07$ (7%).

- The sensitivity $P(A|B)$ here is estimated to be $80/100 = .8$.
- The specificity $P(A^c|B^c)$ here is estimated to be $90/100 = .9$.

From the previous page, the predictive value positive is

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

Therefore, the estimated predictive value positive is

$$\text{estimated} \quad P(B|A) = \frac{(.8)(.07)}{(.8)(.07) + (1-.9)(1-.07)} = .376$$

- This result says that if you've tested positive with this test, then there's an estimated chance of 37.6% that you have diabetes.

ROC Curves:

There is an inherent trade-off between sensitivity and specificity.

Example — CT Scans The following data are ratings of computed tomography (CT) scans by a single radiologist in a sample of 109 subjects with possible neurological problems. The true status of these patients is also known.

		True Disease Status		
		Normal	Abnormal	
Radiologist's Rating	1	33	3	36
	2	6	2	8
	3	6	2	8
	4	11	11	22
	5	2	33	35
		58	51	109

Here, the radiologist's rating is an ordered categorical variable where

1 = definitely normal

2 = probably normal

3 = questionable

4 = probably abnormal

5 = definitely abnormal

If the CT scan is to be used as a screening device for detecting neurological abnormalities, where should the cut-off be set for the diagnosis of abnormality?

Suppose we diagnose every patient with a rating ≥ 1 as abnormal.

- Obviously, we will catch all true abnormals this way — the sensitivity of this test will be 1.
- However, we'll also categorize all normals as abnormal — the specificity will be 0.

Suppose we diagnose every patient with a rating ≤ 5 as normal.

- Obviously, we won't incorrectly diagnose any normals as abnormal — the specificity will be 1.
- However, we won't detect any true abnormalities — the sensitivity of this test will be 0.

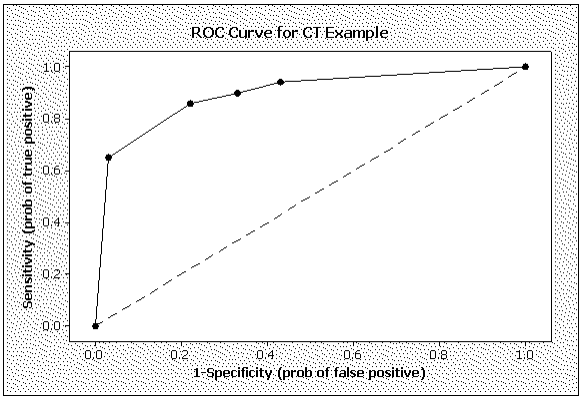
Clearly, we'd prefer to use some threshold between 1 and 5 to diagnose abnormality.

- We can always increase the sensitivity by setting the threshold high, but this will decrease the specificity.
- Similarly, a low threshold will increase the specificity at the cost of sensitivity.

For each possible threshold value, we can compute the sensitivity and specificity as follows:

Test Positive Criterion	Sensitivity	Specificity
≥ 1	1.00	0.00
≥ 2	0.94	0.57
≥ 3	0.90	0.67
≥ 4	0.86	0.78
≥ 5	0.65	0.97
> 5	0.00	1.00

A plot of the sensitivity versus (1 – specificity) is called a **receiver operating characteristic curve**, or ROC curve. The ROC curve for this example is as follows:



- An ROC curve is often used to help determine an appropriate threshold for a screening test. The point closest to the upper-left corner of the plot has the highest combination of sensitivity and specificity.
  - In this example, the ROC curve suggests that we use a rating ≥ 4 to classify patients as abnormal.
- The dashed line in this plot shows where there are equal probabilities of a false positive and a false negative. smallskip
  - A test falling on this line misclassifies normal subjects with the same frequency with which it misclassifies abnormal subjects.
  - Such a test classifies no better than chance, and thus has no predicitve value.

Estimation of Prevalence from a Screening Test:

Suppose we apply a screening test with known sensitivity and specificity to a new population for which the prevalence of the disease is unknown.

Without applying the gold standard test, can we estimate the prevalence?

Let's reconsider the diabetes example. Recall how we defined events:

A = event that a random individual's test is positive

B = event that a random individual has the disease

Previously, we obtained

$$\text{estimated sensitivity} = 0.8 = \widehat{P(A|B)}$$

$$\text{estimated specificity} = 0.9 = \widehat{P(A^c|B^c)}.$$

(hats indicate that these are estimated quantities).

Recall also that we knew the prevalence of diabetes to be .07.

However, now suppose that this prevalence value was for the US population and we decide now to apply the screening value in Canada.

Suppose that we screen $n = 580$ Canadians with our screening test and we obtain $n_{1\cdot} = 105$ positive test results:

		Truth: Diseased (B)	Truth: Not Diseased (B^c)	
Test Result	$+$ (A)	?	?	$n_{1\cdot} = 105$
	$-$ (A^c)	?	?	$n_{2\cdot} = 475$
		?	?	$n = 580$

What is the prevalence of diabetes among Canadians?

Using the law of total probability followed by the multiplication rule, we have

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^c) \\ &= P(A|B)P(B) + P(A|B^c)P(B^c) \\ &= P(A|B)P(B) + P(A|B^c)[1 - P(B)] \end{aligned}$$

With a little algebra, we can solve for $P(B)$, the prevalence of diabetes as follows:

$$P(B) = \frac{P(A) - P(A|B^c)}{P(A|B) - P(A|B^c)} = \frac{P(A) - [1 - P(A^c|B^c)]}{P(A|B) - [1 - P(A^c|B^c)]}$$

$P(A)$, the probability of a positive test result can be estimated as

$$\widehat{P(A)} = \frac{n_{1\cdot}}{n} = \frac{105}{580}$$

and the other quantities in this expression, $P(A|B)$ and $P(A^c|B^c)$, are the sensitivity and specificity of our test.

Therefore, we can estimate the prevalence of diabetes in Canada as

$$\begin{aligned} \widehat{P(B)} &= \frac{\widehat{P(A)} - [1 - P(\widehat{A^c|B^c})]}{P(\widehat{A|B}) - [1 - P(\widehat{A^c|B^c})]} \\ &= \frac{\frac{105}{580} - [1 - .9]}{.8 - [1 - .9]} = .116 \end{aligned}$$

Risk Difference, Relative Risk and Odds Ratio:

Three quantities that are often used to describe the difference between the probability (or risk) of disease between two populations are the risk difference, risk ratio, and odds ratio.

- We will call the two populations the exposed and unexposed populations, but they could be whites and non-whites, males and females, or any two populations (i.e., the "exposure" could be being male).

1. **Risk difference:** One simple way to quantify the difference between two probabilities (risks) is to take their difference.

$$\text{Risk difference} = P(\text{disease}|\text{exposed}) - P(\text{disease}|\text{unexposed}).$$

- Independence between exposure status and disease status corresponds to a risk difference of 0.

- Risk difference ignores the magnitude of risk. E.g., suppose that among males, the exposed and unexposed groups have disease risks of .51 and .50, but among females, the exposed and unexposed groups have risks of .02 and .01.

  - Risk difference is .01 for males and for females. Risk difference does not convey the information that being exposed doubles the risk for females.

2. **Relative risk:** (also known as risk ratio).

$$\text{RR} = \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})}$$

- Independence between exposure status and disease status corresponds to a relative risk of 1.
- Relative risk especially useful for quantifying exposure effect for rare diseases.
  - E.g., the probability that a man over the age of 35 dies of cancer is 0.002679 for current smokers, and .000154 for nonsmokers.

$$\text{RR} = \frac{.002679}{.000154} = 17.4 \quad \text{risk difference} = .002679 - .000154 = .002525.$$

- Risk difference and RR convey different types of information - both useful.

3. **Odds ratio:** RR takes ratio of probabilities. As an alternative to using probability of disease, can compare odds of disease in exposed and unexposed group. This leads to the odds ratio (OR):

$$\text{OR} = \frac{\text{odds}(\text{disease}|\text{exposed})}{\text{odds}(\text{disease}|\text{unexposed})}.$$

- Recall that the odds of an event A are given by

$$\text{odds}(A) = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)},$$

so the OR is

$$\text{OR} = \frac{P(\text{disease}|\text{exposed})/[1 - P(\text{disease}|\text{exposed})]}{P(\text{disease}|\text{unexposed})/[1 - P(\text{disease}|\text{unexposed})]} \qquad (*)$$

- Independence between exposure status and disease status corresponds to an odds ratio of 1.

- The OR conveys similar information to that of the RR. The main advantages of the OR are that

    a. It has better statistical properties. We'll explain this later, but for now take my word for it.

    b. It can be calculated in cases when the RR cannot.

  The latter advantage comes from the fact that using Bayes' Theorem, it can be shown that

$$\text{OR} = \frac{P(\text{exposure}|\text{diseased})/[1 - P(\text{exposure}|\text{diseased})]}{P(\text{exposure}|\text{nondiseased})/[1 - P(\text{exposure}|\text{nondiseased})]} \qquad (**)$$

  – I.e.. (*) and (**) are mathematically equivalent formulas.

  – This equivalence is useful because in some contexts, the probability of exposure can be estimated among diseased and nondiseased but the probability of disease given exposure status cannot. This occurs in case-control studies.

Example — Contraceptive Use and Heart Attack

A case-control study of oral contraceptive use and heart attack. 58 female heart attack victims were identified and each of these "cases" was matched to one "control" subject of similar age, etc. who had not suffered a heart attack.

		Heart Attack: Yes	Heart Attack: No
Contraceptive Use	Yes	23	11
	No	35	47
		58	58

In this case, the column totals are fixed by the study design. Therefore, the probability of heart attack given whether or not oral contraceptives have been used cannot be estimated.

Why?

- Thus, we cannot estimate the risk of disease in either the exposed or unexposed group, and therefore cannot estimate the RR or risk difference.

However, we can estimate probabilities of contraceptive use given presence or absence of heart attack:

$$\hat{P}(\text{contraceptive use}|\text{heart attack}) = 23/58 = .397,$$
$$\hat{P}(\text{contraceptive use}|\text{no heart attack}) = 11/58 = .190.$$

And from these quantities we can estimate the odds ratio:

$$\widehat{\text{OR}} = \frac{\frac{23}{58}\left(1-\frac{11}{58}\right)}{\frac{11}{58}\left(1-\frac{23}{58}\right)} = \frac{\left(\frac{23}{58}\right)\left(\frac{47}{58}\right)}{\left(\frac{11}{58}\right)\left(\frac{35}{58}\right)} = \frac{23(47)}{11(35)} = 2.808.$$

- Interpretation: The odds of heart attack are 2.8 times higher for women who took oral contraceptives than for women who did not.

Theoretical Probability Distributions*

Probability Distributions:

Some definitions:

- A **variable** is any characteristic that can be measured or observed and which may vary (or differ) among the units measured or observed.

- A **random variable** is a variable that takes on different numerical values according to a chance mechanism

  - E.g., any variable measured on the elements of a randomly selected sample.

  - Discrete random variables are random variables that can take on a finite or countable number of possible outcomes (e.g., number of pregnancies).

  - A continuous random variable can (theoretically, at least) take on any value in a continuum or interval (BMI).

- A **probability function** is a function which assigns a probability to each possible value that can be assumed by a discrete random variable.

  The probability function of a discrete random variable (r.v.):

  - defines all possible values of the r.v.

  - gives the probabilities with which the r.v. takes on each of those values.

* Read Ch.7 of our text.

Example Let X =the number of ears affected by one or more episodes of otitis media (ear infection) during the first two years of life. Suppose the probability distribution function for this random variable is

x	$P(X = x)$
0	.13
1	.48
2	.39

- The notation used above is typical. Here, big X is the random variable, little x is a particular value of the random variable, and we are giving the probability that the random variable X takes on the value x for each possible x.
- Note that values of x that are not listed are assumed to have probability 0.
- Of course, the probability function must assign valid probabilities. In particular,
  - when summed over all possible values, the probabilities must sum to 1:
$$\sum_{\text{all } x} P(X = x) = 1.$$
  - and each probability must be between 0 and 1:
$$0 \leq P(X = x) \leq 1 \qquad \text{for all } x.$$
- Probability functions can be given in tables as above, or graphs (e.g., a bar graph), or as a mathematical formula.

The probability function allows computation of probabilities for events defined in terms of the random variable.

- E.g., by the addition rule, the probability of having at least one ear infection during the first two years of life is
$$\begin{aligned} P(X > 0) &= P(X = 1 \cup X = 2) \\ &= P(X = 1) + P(X = 2) - \underbrace{P(X = 1 \cap X = 2)}_{=0} = .48 + .39 = .87 \end{aligned}$$

Expected Value, Variance

The **expected value** of a random variable is the mean, or average value of the r.v. over the population of units on which the r.v. is defined.

- For a random variable X, its expected value is usually denoted $\mathrm{E}(X)$, or μ_X, or simply μ.

The expected value for a discrete r.v. can be computed from its probability distribution as follows:

$$E(X) = \sum_{\text{all } x} xP(X = x),$$

where this sum is taken over all possible values x of the r.v. X.

- E.g., the expected number of ears affected by ear infection during the first two years of life is computed as follows:

x	$P(X = x)$	$xP(X = x)$
0	.13	0(.13)
1	.48	1(.48)
2	.39	2(.39)
		$E(X) = 1.26$

  - Interpretation: the mean number of ears affected by otitis media during the first two years of life is 1.26.

The **variance** of a random variable is the population variance of the r.v. over the population of units on which the r.v. is defined.

- The variance of X is usually denoted $\text{var}(X)$, or σ_X^2, or simply σ^2.
- The formula for the variance of a random variable involves taking expectations:

$$\text{var}(X) = \text{E}\{(X - \mu_X)^2\},$$

which, for a discrete r.v. simplifies to

$$\text{var}(X) = \sum_{\text{all } x} (x - \mu_X)^2 P(X = x),$$

where again this sum is taken over all possible values x of the r.v. X.

- E.g., the variance of the number of ears affected by ear infection during the first two years of life is computed as follows:

x	$P(X = x)$	μ_X	$(x - \mu_X)^2 P(X = x)$
0	.13	1.26	$(0 - 1.26)^2(.13)$
1	.48	1.26	$(1 - 1.26)^2(.48)$
2	.39	1.26	$(2 - 1.26)^2(.39)$
			$\text{var}(X) = .452$

- The population standard deviation of X is $\sigma_x = \sqrt{\sigma_X^2}$ or $\sqrt{.452} = .673$ in our example.

The Binomial Probability Distribution

Many random variables that can be described as event counts where there is a max number of events that can occur, can be thought of as arising from a **binomial experiment**.

A binomial experiment has the following properties:

1. The experiment consists of a sequence of n identical trials.

2. Two outcomes are possible on each trial, one a "success" and the other a "failure".

3. The probability of success, denoted by p, is the same for each trial.

   - Since the probability of a failure is just $1 - p$, this means that the failure probability is the same for each trial as well.

4. The trials are independent (what happens on one trial doesn't affect what happens on any other trial).

In a binomial experiment we are interested in X, the r.v. defined to be the total number of successes that occur over the n trials.

- Note that "success" and "failure" are just convenient labels. A success could be identified as the birth of a girl, and failure as the birth of a boy, or vice versa. That is, "success" simply denotes the event of interest that is being counted.

- X in a binomial trial is a discrete random variable with possible values $0, 1, 2, \ldots, n$.

For any experiment with the above properties, X will necessarily have a particular distribution, the **binomial probability distribution** that is completely determined by n and p.

Examples:

A. The number of heads that occur in 4 coin flips

   1. Each coin flip is an identical trial.
   2. Two outcomes (Heads,Tails) are possible, where "success"= Heads.
   3. Probability of success= $P(\text{Heads}) = 1/2$ on each trial.
   4. Coin flips are independent.

B. The number of obese subjects out of 3 randomly selected US adults.

1. Observing obesity status of each randomly selected US adult is an identical trial.
2. Two outcomes are possible (obese, not obese) where "success" = subject is obese.
3. Probability of success= $P(\text{obese}) = .209$ on each trial.
4. Because selection of subjects is at random, obesity status is independent from subject to subject.

Counter Examples:

C. The number of lifetime miscarriages experienced by a randomly selected woman over the age of 50. Suppose the woman had had 5 lifetime pregnancies.

1. The $n = 5$ pregnancies are the trials, but they are not identical. They occur at different ages, under different circumstances (woman's health status differs, environmental exposures differ, fathers may differ, etc.).
2. Two outcomes are possible (miscarriage, not miscarriage) where "success" = miscarriage.
3. Probability of success not constant on each trial. Probability of miscarriage may be higher when woman is older, may depend on birth order, etc.
4. Pregnancy outcome may not be independent from one pregnancy to the next (if previous pregnancy was a miscarriage, that may increase the probability that next pregnancy will be miscarriage).

D. Out of the n hurricanes that will form in the Atlantic next year, how many will make landfall in the state of Florida?

1. Each hurricane represents a trial. Not identical.
2. Two outcomes possible (hit FL, not hit FL). "Success" = hit FL.
3. Probabilities of hitting Florida may not be constant from hurricane to hurricane depending upon when and where they form, but *a priori*, it may be reasonable to assume that these probabilities are equal from one hurricane to the next.
4. Hurricane paths are probably not independent. If the previous hurricane hit FL, that may increase the chances that the next hurricane will follow the same path and hit FL as well.

For any binomial experiment, the probability of any given number of "successes" out of n trials is given by the **binomial probability function**.

Let the random variable X = the number of successes out of n trials, where p is the success probability on each trial. Then the probability of x successes is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

- Here, $\binom{n}{x}$ (read "n choose x) is shorthand notation for $\frac{n!}{x!(n-x)!}$ where $a!$ ("a factorial") is given by

  $a! = a(a-1)(a-2)\cdots(2)(1)$, and, by convention we define$0! = 1$.

For example, to compute the probability of 3 heads out of 4 coin flips, $n = 4$, $p = \frac{1}{2}$, X = number of heads, where we are interested in $X = x$ where $x = 3$.

Then the binomial probability function says that

$$\begin{aligned} P(X = 3) &= \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \frac{4!}{3!(4-3)!} \left(\frac{1}{2}\right)^3 \left(1 - \frac{1}{2}\right)^{4-3} \\ &= \frac{4(3)(2)(1)}{\{(3)(2)(1)\}\{(1)\}} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 = 4\left(\frac{1}{2}\right)^4 = 0.25 \end{aligned}$$

Where does this formula come from?

Let's consider example B.

Let X =number of obese subjects out of $n = 3$ randomly chosen US adults where $p = .209$.

Forgetting the formula for a minute, how could we compute $P(X = 2)$, say?

One way is to list all of the possible outcomes of the experiment of observing 3 subjects and add up the probabilities for the outcomes that correspond to 2 obese subjects.

Possible outcomes:

Outcome Number	First Subject	Second Subject	Third Subject	Probability of Outcome
1	O	O	O	
2	O	O	N	
3	O	N	O	
4	O	N	N	
5	N	O	O	
6	N	O	N	
7	N	N	O	
8	N	N	N	

- Outcomes 2, 3, and 5 corresponse to getting a total of $X = 2$ obese subjects out of $n = 3$. What are the probabilitites of these three outcomes?

Probability of (O, O, N):

- Recall that for independent events, the joint probability of the events is the product of the individual probabilities of each event. Here, whether the subject is obese is independent from subject to subject.

So, the probability of observing (O, O, N) is

$$p \times p \times (1 - p) = p^2(1 - p)^1 = p^x(1 - p)^{n-x}$$

where $n = 3$, $x = 2$.

Probability of (O, N, O):

$$p \times (1-p) \times p = p^2(1-p)^1 = p^x(1-p)^{n-x}$$

where $n = 3$, $x = 2$.

Probability of (N, O, O):

$$(1-p) \times p \times p = p^2(1-p)^1 = p^x(1-p)^{n-x}$$

where $n = 3$, $x = 2$.

Adding the probabilities of these mutually exclusive events together (addition rule) we get

$$P(X = 2) = p^2(1-p)^1 + p^2(1-p)^1 + p^2(1-p)^1 = 3p^2(1-p)^1$$

where for $n = 3$, $x = 2$

$$\binom{n}{x} = \binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3(2)(1)}{\{(2)(1)\}\{(1)\}} = 3.$$

- $\binom{3}{2}$ is the number of ways to arrange a sequence with 2 'O's and 1 'N'.
- More generally, $\binom{n}{x}$ gives the number of ways to choose x objects out of n to be of one type and $n - x$ to be of the other type.
- So, the probability of 2 obese subjects out of 3 randomly selected subjects is

$$\binom{n}{x} p^x(1-p)^{n-x} = \binom{3}{2}(.209)^2(1-.209)^{3-2} = 3(.209)^2(1-.209)^{3-2} = .1037.$$

The binomial formula can be used to compute the probability of x successes out of n trials where the success probability on each trial is p for any value of n and p.

However, it is convenient to have a table to give the answer for any given value of n and p, or, even better, a computer function that allows us to input n and p and outputs the answer.

- Table A.1 in Appendix A of our book gives binomial probabilities for selected values of n and p.

E.g., we computed the probability of $x = 3$ heads out of $n = 4$ coin flips to be .25. Table A.1 uses k instead of x, so we look up $n = 4$ and $k = 3$ on the left side of the table, $p = .5$ on the top and find the probability equals .2500 just as we computed.

- Note that the table only gives selected values of p where $p \leq .5$.

  What if we are interested in $p = .75$, say?

We can handle such a case by considering the number of failures rather than the number of successes.

That is, if X equals the number of successes out of n trials with success probability p, then

$$Y = n - X = \text{number of failures},$$

where the failure probability is $q = 1 - p$. We observe $X = x$ successes out of n trials if and only if we observe $Y = n - x$ failures. So,

$$\begin{aligned} P(X = x) = P(Y = n - x) &= \binom{n}{n-x} q^{n-x}(1-q)^{n-(n-x)} \\ &= \binom{n}{n-x} q^{n-x}(1-q)^{x}. \end{aligned}$$

Example Suppose that 55% of UGA undergraduates are women. In a random sample of 7 UGA undergraduates, what's the probability that 3 of them are women?

Here X = number of women (success) out of $n = 7$ "trials" where probability of woman on each trial is $p = .55$. If $x = 3$ women are observed, then we necessarily have observed $Y = n - x = 7 - 3 = 4$ men where the probability of observing a man is

$$q = 1 - p = 1 - .55 = .45.$$

So, the desired probability can be computed based on X:

$$\begin{aligned} P(X = 3) &= \binom{n}{x} p^x (1-p)^{n-x} = \binom{7}{3} (.55)^3 (1-.55)^{7-3} \\ &= \frac{(7)(6)(5)(4)(3)(2)(1)}{\{(3)(2)(1)\}\{(4)(3)(2)(1)\}} (.55)^3 (1-.55)^4 = .2388 \end{aligned}$$

or, equivalently, based on Y:

$$\begin{aligned} P(Y = 4) &= \binom{7}{4} (.45)^4 (1-.45)^{7-4} \\ &= \frac{(7)(6)(5)(4)(3)(2)(1)}{\{(4)(3)(2)(1)\}\{(3)(2)(1)\}} (.45)^4 (1-.45)^3 = .2388 \end{aligned}$$

- The latter probability is tabulated in Appendix A.1, but the former is not.

In addition, computer programs give binomial probabilities too. These have the advantage that they give the result for any value of n and p.

- In Minitab, select

  Calc > Probability Distributions > Binomial...

  Then select "Probability", and enter values for "Number of trials" and "Probability of success". The value of x desired can be input under "Input constant" or can be selected from data in a worksheet.

The binomial probability function gives the $P(X = x)$ for all possible values of x: $0, 1, 2, \ldots, n$. So, the probability function gives the entire probability distribution of X.

Once we know the probability distribution of a discrete r.v., we can compute its expected value and variance.

Recall:

$$\begin{aligned} E(X) &= \sum_{\text{all } x} xP(X = x) \\ &= 0P(X = 0) + 1P(X = 1) + \cdots + nP(X = n) = \mu_X \end{aligned}$$

and

$$\begin{aligned} \text{var}(X) &= \sum_{\text{all } x} (x - \mu_X)^2 P(X = x) \\ &= (0 - \mu_X)^2 P(X = 0) + \cdots + (n - \mu_X)^2 P(X = n) = \sigma_X^2 \end{aligned}$$

Fortunately, these formulas simplify for the binomial distribution so that we don't have to compute $P(X = 0), \ldots, P(X = n)$.

In a binomial experiment with n trials, each with success probability p, the number of successes X has the following expected value and variance:

$$\begin{aligned} E(X) &= np \\ \text{var}(X) &= np(1 - p) \end{aligned}$$

Example — Obesity Again

Suppose I take a random sample of $n = 4$ US adults. How many obese subjects should I expect to observe on average?

Here $n = 4$, $p = .209$, so I expect to observe

$$E(X) = np = 4(.209) = 0.836$$

obese adults out of a sample of $n = 4$.

- In a sample of $n = 1000$, I'd expect to observe $np = 1000(.209) = 209$ obese adults. (Make sense?)

The variance of the number of obese adults observed out of $n = 1000$ would be

$$\text{var}(X) = np(1-p) = 1000(.209)(1-.209) = 165.319$$

That is, the standard deviation is $\sqrt{165.319} = 12.9$.

- The interpretation here is that I could select $n = 1000$ US adults and count the number of obese subjects over and over again. Over the long run, the standard deviation of the number of obese subjects observed when repeating this binomial experiment again and again is 12.9.
    - That is, I expect to get about 209 out of 1000 obese subjects, but the actual number obtained is going to vary around 209, with typical deviation from 209 equal to 12.9.

The Poisson Probability Distribution

Another important discrete probability distribution that arisesoften in practice is the **Poisson probability distribution.**

- The binomial probability function gave the probability for the number of successes out of n trials.
  - Pertains to counts (of the number of successes) that are subject to an upper bound n.
- The Poisson probability function gives the probability for the number of events that occur in a given interval (often a period of time) assuming that events occur at a constant rate during that interval.
  - pertains to counts that are unbounded. Any number of events could, theoretically occur during the period of interest.
- In the binomial case, we know p = probability of the event (success) in each trial.
- In the Poisson case, we know λ = the mean (or expected) number of events that occur in the interval.
  - Or, equivalently, we could know the rate of events per unit of time. Then λ, the mean number of events during an interval of length t would just be $t \times$ rate.

Example — Traffic Accidents:

Based on long-run traffic history, suppose that we know that an average of 7 traffic accidents per month occur at Broad and Lumpkin. That is, $\lambda = 7$ per month. We assume this value is constant throughout the year.

What's the probability that in a given month we observe exactly 8 accidents?

Such probabilities can be computed by the Poisson probability function. If X = the number of events that occur according to a Poisson experiment with mean λ, then

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

- Here, e denotes the base of the natural logarithm function. This is a constant (like π) equal to $2.71828\ldots$.

In the example, the probability of getting exactly 8 accidents in a month is

$$P(X = 8) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-7}7^8}{8!} = .130.$$

- Poisson probabilities are tabulated in Table A.2 of Appendix A of our text. They also may be computed in computer programs like Minitab.

Often we are interested in cumulative probabilities.

- For example, we may be interested in the probability that we have no more than 8 accidents in a given month.

A probability $P(X \leq x)$ like this can be computed simply by summing up $P(X = 0), P(X = 1), \ldots, P(X = x)$.

- In this example, the probability of no more than 8 accidents in a month is given by

$$\begin{aligned} P(X \leq 8) &= P(X = 0) + \cdots + P(X = 8) \\ &= \frac{e^{-7}7^0}{0!} + \cdots + \frac{e^{-7}7^8}{8!} = .000912 + \cdots + .130 = .729. \end{aligned}$$

  Fortunately, computer programs like Minitab usually have functions for cumulative probabilities like this so that the individual probabilities need not be computed separately and then summed.

- Of course, if I were interested in knowing the probability of having more than x accidents in a month I could get that via

$$P(X > x) = 1 - P(X \leq x)$$

  So, for example, the probability of having 9 or more accidents in a month is $1 - .729 = .271$.

- Cumulative binomial probabilities can be computed in the same way as Poisson cumulative probabilities. That is, the formulas

$$P(X \leq x) = P(X = 0) + \cdots + P(X = x) \quad \text{and} \quad P(X > x) = 1 - P(X \leq x)$$

  hold for X a binomial outcome as well.

- The Poisson distribution has the remarkable property that its expected value (mean) and variance are the same. That is, for X following a Poisson distribution with mean λ,

$$\mathrm{E}(X) = \mathrm{var}(X) = \lambda$$

- For binomial experiments involving rare events (small p) and large values of n, the distribution of X = the number of success out of n trials is binomial, but it is also well approximated by the Poisson distribution with mean $\lambda = np$.

E.g., Suppose that the prevalence of testicular cancer among US males is .000113. Suppose we take a random sample of $n = 10,000$ male subjects. Then the probability of observing 2 or fewer males with a lifetime diagnosis of testicular cancer is given by the binomial cumulative probability:

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \binom{10000}{0} .000113^0 (1 - .000113)^{10000-0} + \cdots + \binom{10000}{2} .000113^2 (1 - .000113)^{10000-2} \\ &= .894312 \end{aligned}$$

This is the exact answer, but it is pretty well approximated by a Poisson probability with mean $\lambda = np = 10000(.000113) = 1.13$. Using the Poisson probability function

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &\approx \frac{e^{-1.13} 1.13^0}{0!} + \frac{e^{-1.13} 1.13^1}{1!} + \frac{e^{-1.13} 1.13^2}{2!} = .894301 \end{aligned}$$

Continuous Probability Distributions

Recall that for a discrete r.v. X, the probability function of X gave $P(X = x)$ for all possible x, thus describing the entire distribution of the r.v. X.

We'd like to do the same for a continuous r.v.

How do we calculate probabilities for continuous random variables?

- For a continuous r.v., the probability that it takes on any particular value is 0! Therefore, we can't use a probability function to describe it!
  - E.g., the probability that a randomly selected subject from this class weighs 146.923578234785079074... lbs is 0.

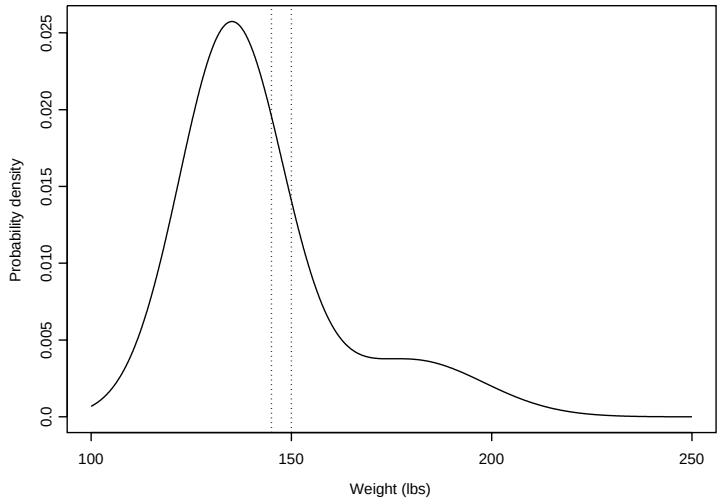
Instead of a probability function that gives the probability for each particular value of X, we quantify the probability that X falls in some interval or region of all possible values of X.

- This works because while the probability that a random student weighs 146.923578234785079074... lbs is 0, the probability that he/she weighs between 145 and 150 lbs, say, is not 0.

So, instead of describing the distribution of a continuous r.v. with a probability function, we use what is called the **probability density function.**

- The probability density function for a continuous r.v. X gives a curve such that the area under the curve corresponding to some interval on the horizontal axis gives the probability that X takes a value in that interval.

E.g., suppose the probability density function for X =body weight for a randomly selected student in this class looks like this:



- The dashed vertical lines are at weight=145 lbs and weight=150 lbs. The area under the curve between these lines gives the probability that a randomly selected student weighs between 145 and 150 lbs.
- In general, the are under the curve between x_1 and x_2 where $x_1 < x_2$ gives
$$P(x_1 < X < x_2)$$
- Note that the curve extends to the left and right, getting closer and closer to zero.
  - That is, weights greater than x lbs, say, are possible (have nonzero probability) no matter how big x is, but they are increasingly unlikely as x gets bigger.
  - Similarly, smaller and smaller weights are decreasingly probable.
- The entire area under the probability density function is 1, representing the fact that
$$P(-\infty < X < \infty) = 1$$

- Note that for a continuous r.v. X, $P(X = x) = 0$ for all x. Therefore,

$$P(X \leq x) = P(X = x) + P(X < x) = 0 + P(X < x) = P(X < x).$$

Similarly,

$$P(X \geq x) = P(X = x) + P(X > x) = 0 + P(X > x) = P(X > x).$$

  - That is, for X continuous, there's no difference between $<$ and $\leq$ probability statements, and also no difference between $>$ and $\geq$ probability statements. Not true in the discrete case.

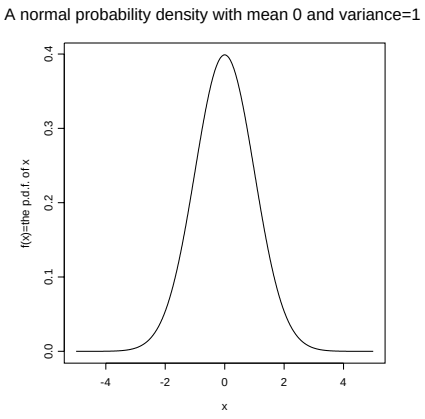
The Normal Distribution

Many continuous random variables have distributions such that

— values close to the mean are most probable, and values further away from the mean are decreasingly probable (unimodal)

— values c units larger than the mean are just as probable as values that are c units smaller than the mean (symmetry).

That is, many continuous random variables have probability distributions that look like this:



The probability density function or p.d.f. given above is the p.d.f. of the **normal probability distribution** (sometimes called the **Gaussian probability distribution**).

- The normal distribution is not the only distribution whose p.d.f. looks bell-shaped, but it is the most important one, and many real world random variables follow the normal distribution, at least approximately.

- The normal distribution, like the binomial and Poisson, is an example of a **parametric** probability distribution. It is completely described by a small number of **parameters**.

  - In the case of the binomial, there were two parameters, n and p.

  - In the case of the Poisson, there was just one parameter, λ, the mean of the distribution.

  - In the case of the normal, there are two parameters:

$$\mu = \text{the mean of the distribution, and}$$
$$\sigma^2 = \text{the variance of the distribution.}$$

- That is, if X is a r.v. that follows the normal distribution, then that means that we know exactly the shape of the p.d.f. of X except for $\mu = \mathrm{E}(X)$, the mean of X, and $\sigma^2 = \mathrm{var}(X)$, the variance of X.

  - We will use the notation

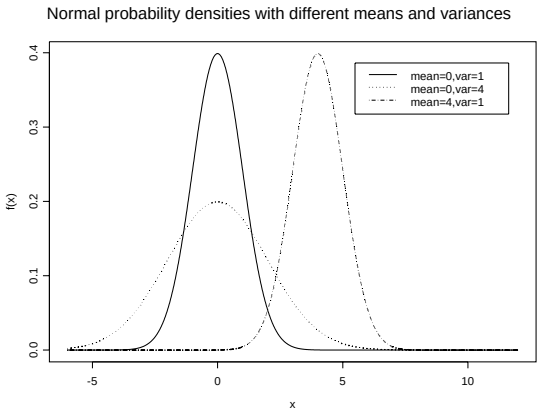
$$X \sim N(\mu, \sigma^2)$$

    to denote that the r.v. X folllows a normal distribution with mean μ and variance σ^2.

  - E.g., $X \sim N(3, 9)$ means that X has a normal distribution with mean 3 and variance 9 (or SD=3).

- The normal curve given above has mean 0 and variance 1. I.e., it is $N(0, 1)$, which is called the **standard normal distribution.**

- Normal distributions with different means have different locations.
- Normal distributions with different variances have different degrees of spread (dispersion).
    - Below are three normal probability distributions with different means and variances.



The normal p.d.f. is a function of x that maps out a bell-shaped curve. That is, it is a function $f(x)$ that gives a probability density (a value along the vertical axis) for each value of x (a value along the horizontal axis).

For a r.v. $X \sim N(\mu, \sigma)$, the specific mathematical form of the normal probability density of X is

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

where again, e denotes the constant 2.71828... and π denotes the constant 3.14159....

Facts about the normal distribution:

1. It is symmetric and unimodal.
   - As a consequence of this, the mean, median and mode are all equal and occur at the peak of the normal p.d.f.
2. The normal p.d.f. can be located (have mean) anywhere along the real line between $\pm\infty$ and extends indefinitely away from its mean in either direction without ever touching the horizontal axis.
   - That is, if $X \sim N(\mu, \sigma^2)$, then any value of X is possible, although values far from μ will not be very probable.
3. As with any p.d.f., the area under the normal curve between any two numbers x_1, x_2 where $x_1 < x_2$ gives

   $$P(x_1 < X < x_2)$$

   and the total area under the p.d.f. is 1.

   In particular, here are a few notable normal probabilities:

   - For $x_1 = \mu - 1\sigma$, $x_2 = \mu + 1\sigma$,

     $$P(\mu - 1\sigma < X < \mu + 1\sigma) = .6826$$

     That is, 68.26% of the time a normally distributed r.v. falls within 1 SD of its mean (i.e., has z score between -1 and 1).
   - For $x_1 = \mu - 2\sigma$, $x_2 = \mu + 2\sigma$,

     $$P(\mu - 2\sigma < X < \mu + 2\sigma) = .9544$$

     That is, 95.44% of the time a normally distributed r.v. falls within 2 SDs of its mean (i.e., has z score between -2 and 2).
   - For $x_1 = \mu - 3\sigma$, $x_2 = \mu + 3\sigma$,

     $$P(\mu - 3\sigma < X < \mu + 3\sigma) = .9972$$

     That is, 99.72% of the time a normally distributed r.v. falls within 3 SDs of its mean (i.e., has z score between -3 and 3).

- These results are where the "empical rule" comes from.

Example — Height

Suppose that US adult women have heights that are normally distributed where the population mean height is 65 inches and the population standard deviation for women's height is 2.5 inches.

Suppose that US adult men have heights that are normally distributed with population mean 70 inches and population SD of 3 inches.

Let X = the height of a randomly selected adult US woman, and Y = the height of a randomly selected adult US man. Then

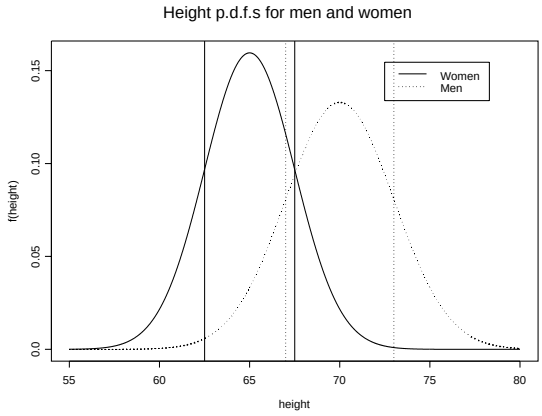
$$X \sim N(\mu_X, \sigma_X^2) = N(65, 2.5^2), \quad Y \sim N(\mu_Y, \sigma_Y^2) = N(70, 3^2).$$

- For women, one SD below the mean is $\mu_X - 1\sigma_X = 65 - 1(2.5) = 62.5$. One SD above the mean is $\mu_X + 1\sigma_X = 65 + 1(2.5) = 67.5$.
  - So, the probability that a randomly selected woman has height between 62.5 and 67.5 inches is

    $$P(62.5 < X < 67.5) = .6826$$

    (68.26% of women have heights between 62.5 and 67.5 inches).

The height p.d.f.s for men and women are given below.



- Clearly, the area under the curve between 62.5 and 67.5 inches for men is much less than 68.26%.
  - In fact the area under the male height curve between 62.5 and 67.5 inches turns out to be .1961 or 19.61%.