

We have a **population of units** of population size **N**, i.e. the actual objects or subjects for which we want to collect observations. The **statistical population** is the population of measurements (**variables**) of interest corresponding to the units. Originally statistics involved describing or summarizing **numerical data** graphically or in tables ( the realm of **descriptive statistics**). Since we generally don't have the resources nor do we want to spend the effort to make a measurement for each unit of a very large population, instead more modern **inferential statistics** make **inferences** which are generalizations about the population as a whole based on a small **sample** of size **n** of the population (i.e. a small subset ).

*Note 1:* More generally we can have an **infinite conceptual population** such as the possible outcomes of choosing a random real number from the interval  $[0,1]$ . There are infinitely many real numbers in this interval.

*Note 2:* For **Categorical** data variables having values such as binary valued yes/no or success/failure ones (also called **Bernoulli random variables**) or the (ternary valued) colors red, green, or blue it is often useful to convert them into **ordinal** or **numeric** values by ordering or assigning numbers to the categories (e.g. 1,2,3 instead of red, green, blue).

There are risks associated with making a false generalization (that a hypothesis is false when it is true or vica versa) and statistics examines likelihoods and consequences of such errors. Statistics examines the sources of **variability** in the data and tries to determine how much of this variability (as manifested in the trends and relationships in the data) is due to chance (this is known as **experimental error**) and how much is due to natural laws or actual patterns and it looks at optimal ways to design experiments so as to minimize experimental error and determine real effects. In an experiment controllable input variables are known as **factors** and the output variables that result are known as **responses**. Good **experimental design** seeks to cope with **complexity**, allowing the examination of multiple factors simultaneously and the possible interactions between them. It also seeks to separate **correlation from causation** which may or may not be present. (Correlation is a measure of a linear relationship in the data). For example data from the town of Oldenburg in Germany during the years 1930 to 1936 supports the fairy tale that storks bring babies since both the stork population  $X$  and the human population  $Y$  observed at the end of each of 7 years shows a distinct linear trend in time so that the plot of stork population versus human population is nearly a straight line. But clearly the hidden causal variable factor here behind both of these correlated measurements is time.

If the (numeric valued) **variable** of interest was height, and the population **parameter** we were trying to determine was the population average height  $\mu$  of Madison Wisconsin citizens we might average the heights of a sample of 10 Madison residents (sample size  $n=10$ ) to get the **sample mean**  $\bar{x}$  as a measure of where the data is centered or located thus obtaining

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k = \frac{1}{10} (x_1 + x_2 + \dots + x_{10})$$

as our **estimate** of the true population mean  $\mu$  (also called the

**expected value** of a randomly chosen height  $X_1$  denoted by  $E[X_1]$  where the capital letter  $E$  stands for expectation). The population mean formula is similar to the sample mean formula except that we replace the sample size  $n$  by (in this example Madison's) population size  $N$  so

$$\mu = E[X_1] = \frac{1}{N} \sum_{i=1}^N x_i .$$

Notice for a fixed population this expected value or population mean  $\mu$  is no longer random but is instead a **fixed parameter** of the population whereas for a randomly chosen small sample, *prior* to choosing the sample,

the sample mean is a **random variable**  $\bar{X} = \frac{1}{n} [X_1 + X_2 + \dots + X_n] = \frac{S_n}{n}$  , WEEK 1 page 2

the average of  $n$  unknown heights **denoted by capital letters**  $X_i$  . Once selected, a particular sample is no longer random so we use **lower case letters**  $x_i$  for the  $i^{\text{th}}$  (non-random) height in a particular sample after selection. Notice that the sample mean  $\bar{x}$  is a kind of **summary** statistic of the typical (or average location or center value) of our sample data. Technically a **statistic** is a quantity calculated from (that is a function of) the sample observations, or the corresponding random variable prior to selecting the random sample.

In our example of height measurements from Madison's population of roughly  $N \approx 250,000$  , if we arrange height measurements in a sample in increasing order we get the **order statistics** denoted  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  .

In the above sum defining the population mean, rearranging the observed heights in increasing order will not change the sum. But now suppose all our height measurements have been rounded off to the nearest inch. The shortest baby or the tallest giant will likely lie somewhere between 1 foot (= 12 inches) and 8 feet (= 96 inches). Thus in this case there will only be  $K=84$  different height values  $z_k$  that the  $N = 250,000$  measurements can assume, which we can take arranged in increasing order, so we have

$$\mu = E[X_1] = \frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{N} \sum_{i=1}^N x_{(i)} = \sum_{k=1}^{K=84} z_k \frac{c(k)}{N} = \sum_{k=1}^{K=84} z_k p(z_k) .$$

Here  $c(k)$  is the count or **frequency** or number of  $x_i$ 's that assume the value  $z_k$  inches and

$p(z_k) = \frac{c(k)}{N}$  is the **proportion** or fraction or **relative frequency** of  $z_k$  values where *relative*

refers to the frequency count relative to the total population  $N$ . Since the counts  $c(k)$  add up to  $N$ , the proportions  $p(z_k)$  must add up to 1 when summed over all 84 values of  $k$ . For a fixed finite population as above, these relative frequencies are in fact the probabilities  $p(z_k) = P(X_1 = z_k)$  in our example that a randomly selected person's height (rounded to nearest inch) will equal the value  $z_k$  .

In the limit as the sample size  $n$  goes to infinity under suitable assumptions the **law of large numbers** says that the sample mean approaches the population mean :  $\bar{x} \rightarrow \mu$  as  $n \rightarrow \infty$  . From this it follows that the above relative frequencies converge as  $n \rightarrow \infty$  to **probabilities** which we can regard as limiting relative frequencies (proportions between 0 and 1) whose total sum is 1. For a random variable  $X$  that can take continuous values the probability  $P(X \in (a, b))$  that  $X$  lies in the interval  $(a, b)$  on the real number line with  $a \leq b$  is just the area lying under the **probability density function**  $f(x)$  for  $x$  between  $a$  and  $b$  . The **uniform U([0,1]) distribution** on the interval  $[0,1]$  for example is the one with density  $f(x) = 1$  for  $0 \leq x \leq 1$  and density 0 outside this interval. Then for  $0 \leq a \leq b \leq 1$  the probability of picking a random number between  $a$  and  $b$  according to this distribution is just the area under  $f(x) = 1$  on the interval  $(a, b)$  i.e. the area of a rectangle of height 1 and width  $b - a$  which equals  $b - a$  .

*Remarks : some further comments on probabilities :* Consider an infinite conceptual population of heights in which we shrink the distance  $\Delta z$  between adjacent  $z$  values down to billionths of an inch and finally to zero. Writing  $p(z_k) = f(z_k) \Delta z$  expresses the probability  $p(z_k)$  of obtaining a value between  $z_k$  and  $z_k + \Delta z$  as the area of a rectangle of vertical height  $f(z_k)$  (known as the *probability density function*) and small width  $\Delta z$  . In this continuous limit, the probability that our random variable (in this example height) takes a value between  $a$  and  $b$  is just the integral

$\int_a^b f(z) dz$  of the density function from  $a$  to  $b$  (which is the area under the density function). The

total area (total probability) under the density function must be  $\int_{-\infty}^{\infty} f(z) dz = 1$  . WEEK 1 page 3  
 The above sum for the population mean then becomes  $\sum z_k f(z_k) \Delta z$  which turns into the integral  

$$\mu = E[X] = \int z f(z) dz$$
 in the limit as  $\Delta z \rightarrow 0$  .

*Note* : the counts (viewed as random variables before the data has been collected) can be rewritten  
 $C(k) = \sum_{i=1}^N 1_{[z_k, z_k + \Delta z)}(X_i)$  where the indicator function  $1_{[z_k, z_k + \Delta z)}(X_i)$  equals 1 if the  $i^{th}$  random observation  $X_i$  lies in the interval  $[z_k, z_k + \Delta z)$  and equals 0 otherwise. Then  $p(z_k)$  is just the sample mean of the indicator random variables  $1_{[z_k, z_k + \Delta z)}(X_i)$  whose population mean is the limiting probability  $P(X_i \in [z_k, z_k + \Delta z))$  that the random variable  $X_i$  lies in the interval  $[z_k, z_k + \Delta z)$  . Thus the law of large numbers justifies the relative frequency interpretation of probability.

*Note on linearity of expectations* : when a (**discrete**) random variable  $Z$  takes values  $z_k$  with non-zero (discrete) probability  $p(z_k) = P(Z = z_k)$  we have the definition of expectation

$E[Z] = \sum_k z_k p(z_k)$  even when there are a (countably) infinite number of such values (whose probabilities must still sum to 1). For a **continuous** random variable with probability density  $f(z)$  (having zero probability of assuming any particular value) we have  $\mu = E[Z] = \int z f(z) dz$  . In either case by the distributive properties of finite sums (and their limiting infinite sums and integrals) for any constants  $a$  and  $b$  and random variables  $X$  and  $Y$  , the expectation satisfies the **linearity** property :

$$E[aX + bY] = a E[X] + b E[Y] .$$

**Need for randomization (random sampling)**: If we happen to live next door to the UW basketball team, it might be convenient for us to collect 10 heights by measuring the heights of team members next door. But such a self-selected **convenience sample** is not likely to be representative of the heights of the Madison population as a whole. Rather it is likely to be **biased** (not close to the actual parameter  $\mu$  of interest) . It is best to avoid such **self-selected samples** and to choose instead a **random sample**. In this case the sample mean  $\bar{X}$  will be a **random variable** which will be an **unbiased** estimator of the population mean (i.e. the **expected value**  $E\bar{X}$  of the sample mean  $\bar{X}$  is equal to the correct population mean :  $E\bar{X} = \mu$  ) We could for instance use a computer to randomly select 10 names out of the Madison phone directory and then attempt to measure the heights of these individuals (or better yet of a randomly selected member of their family so as to include children not listed in the phone book). Or we could use a **random number table** to choose random integer numbers corresponding to a random selection of these names (after ordering the names sequentially from 1 to the number of names in the phone book).

A **random sample from a finite population** (say picking a sample of size  $n = 5$  poker hand from a population of  $N = 52$  well shuffled cards in a deck) will mean one for which each possible sample of size  $n$  chosen from  $N$  is **equally likely** (i.e. has equal probability for fixed  $n$ ). A **random sample from an infinite population** is one which is **independent, identically distributed (IID)**. When flipping a fair coin (**fair** means probability  $P(heads) = P(tails) = 1/2$  ) for instance, to say coin flips are *independent* means roughly that the outcome of the previous flip should not influence the next flip. We will give a more precise definition of independence later. **Time series data** are generally **dependent** (i.e. not independent) and said to be **auto-correlated** since for example 5 consecutive days in Madison in July are likely to have similar temperatures compared to temperatures on 5 randomly chosen dates from all year round. In the poker hand example the individual cards viewed as samples of size 1 when **sampling without replacement** are not random (not **independent** of one another nor identically dist'd) unless we put each card back in the deck (**sampling with replacement**)

since the first card has probability 1/52 of being chosen but the second card has probability 1/51 as there are only 51 cards left in the deck unless we put cards back. WEEK 1 page 4

**Large sample statistics** : The law of large numbers tells us that the sample mean  $\bar{X}$  approaches the population mean  $\mu$  for large sample size but it does not tell us how fast it converges to it. The **Central Limit Theorem (CLT)** tells us more. For a sum  $S_n = X_1 + X_2 + \dots + X_n$  of a large number of independent, identically distributed (**IID**) observations, in practice for sample size  $n \geq 30$  under assumptions which usually hold (assuming the variance  $\sigma_{X_1}^2 = \text{var}(X_1) = E((X_1 - \mu)^2) < \infty$  exists), no matter what the probability distribution of the individual observations, the sum  $S_n$  and hence the sample mean  $\bar{X} = \frac{S_n}{n}$  will be approximately normally distributed. By subtracting off its mean and dividing by its **standard deviation** (= the square root of the variance) any normally distributed random variable becomes **standard normal**  $N(0,1)$ . I.e. it is normal with mean 0 and variance 1 with

(bell-shaped) **probability density function of a standard normal distribution**  $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

**Note on expectation and variance of sums:** Linearity of expectations  $E[aX + bY] = aE[X] + bE[Y]$  holds **whether or not independence** of the random variables X and Y holds and for IID independent, identically distributed sums gives  $E[S_n] = E[X_1 + X_2 + \dots + X_n] = E[X_1] + E[X_2] + \dots + E[X_n]$

$$= \mu + \mu + \dots + \mu = n\mu \text{ and hence } E[\bar{X}] = E\left[\frac{S_n}{n}\right] = \frac{1}{n}E[S_n] = \frac{n\mu}{n} = \mu \text{ or } E[\bar{X}] = \mu \text{ i.e. we say the}$$

sample mean is an **unbiased estimator** of the population mean  $\mu$ .

**Provided independence** of X and Y holds, the corresponding property for variance says

$$V[aX + bY] = a^2V[X] + b^2V[Y] \text{ which (for } a=b=1) \text{ gives } V[X + Y] = V[X] + V[Y] \text{ and (for } a=1, b=-1) \text{ gives } V[X - Y] = V[X] + V[Y].$$

$$\text{For independent sums this says } V[S_n] = V[X_1 + X_2 + \dots + X_n] = V[X_1] + V[X_2] + \dots + V[X_n] = \sigma^2 + \sigma^2 + \dots + \sigma^2 = n\sigma^2 \text{ and hence}$$

$$\sigma_{\bar{X}}^2 = V[\bar{X}] = V\left[\frac{S_n}{n}\right] = \frac{1}{n^2}V[S_n] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \text{ or } \sigma_{\bar{X}} = SD[\bar{X}] = \frac{\sigma}{\sqrt{n}} \text{ for the standard deviation of } \bar{X}.$$

This says **the larger the sample size n the smaller the variance of the sample mean  $\bar{X}$**  about  $\mu$  (i.e. the better the estimate  $\bar{X}$  of  $\mu$  is). This also says for *any* random variable X having mean

$$\mu \text{ and variance } \sigma^2, \text{ the } \mathbf{standardized variable } Z = \frac{X - \mu}{\sigma} \text{ has mean 0 and variance 1. If X was}$$

normal  $N(\mu, \sigma^2)$  the standardized Z will again be normal  $N(0,1)$ . Applying this and the central limit theorem (**CLT**) to the approximately normal  $S_n$  with mean  $n\mu$  and standard deviation  $\sqrt{n}\sigma$

the **CLT** result can be expressed as  $Z = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  is approximately standard normal

$N(0,1)$  for large  $n \geq 30$ . In reality we usually don't know the population standard deviation  $\sigma$  and have to replace  $\sigma$  with the *sample standard deviation*  $s$  in the CLT. This requires somewhat larger sample size ( $n \geq 40$  is usually OK)

**Small sample statistics** (sample size  $n$  small) must **either** make strong assumptions about the distribution of the individual observations (often that each is approximately normal since one can show even for few terms, *any* sum of IID normals is normal) **or** involve tests which are **robust** (not too much affected) to departures from normality **or** use statistical methods (so called **non-parametric methods** based on **order statistics**) which do not depend on the distribution of individual observations, but these methods, while more general purpose, may be less powerful (less accurate than other methods). If the observations  $X_i$  are normal, the Z in the CLT above is normal even for small  $n$ ; replacing

$\sigma$  by  $s$  in the CLT expression for  $Z$  gives for small  $n$  what is called a **t**-statistic with  $n-1$  **degrees of freedom**. The **t**-distribution density is also bell-shaped but wider than a normal  $Z$  distribution since  $s$  being a random variable adds variability not present in the fixed parameter  $\sigma$ .

**Statistics as an iterative (cyclical) process of deductive and inductive learning** : In Statistics we start from assumptions such as normality which represent a **probability model** and based on the model we make deductions about the world. We compare the deduced consequences of the **model** to the (real world) **data**. and then (in the inductive step) *revise the model* accordingly. We repeat (iterate) the process until satisfied with the results. Said differently we start with a **prior** model (i.e *before* data is added) and incorporate new data, which may entail updating parameters in the **posterior** model (i.e. *after* data is added). The *old posterior* then becomes the *new prior* in the **next cycle** of the process.

**Chapter 1 Examples** ( material relevant to Chapter 1 of text ) :

**Example 1** (like problem 1.1+1.3) : To measure the average miles per gallon that a typical car in New York city gets, the variable of interest, miles per gallon was measured for 254 randomly selected automobiles in New York City over the course of a month. Here the population of units are the set of all cars of New York City. The statistical population is the set of miles per gallon fuel efficiency measurements (or hypothetical measurements) associated with these cars. The (random) sample consists of the 254 mpg readings (sample size  $n=254$ ) corresponding to the 254 auto units randomly chosen.

**Example 2** (like problem 1.2) : A WORT-FM Madison radio host wants to know who the favorite Presidential candidates are amongst eligible voters in Madison. He asks listeners to download a questionnaire from the station's website and to fill it out online. What are the potential flaws of this (self-selected) method of surveying popular sentiment?

The desired population of interest consists of Madison voters. The actual population surveyed are both 1) listeners of WORT radio (whose political views may differ sharply from Madison residents as a whole) and 2) only those listeners who were sufficiently enthused about their candidate or about answering a questionnaire that they bothered to expend the effort needed to fill it out online. This last consideration may also be a source of bias. A better method might be to randomly select 100 Madison residents from the phone directory, call them and politely ask them to answer questions about their candidate. Such a survey is not limited to WORT listeners and hopefully if the person conducting the poll is polite most persons called will answer the questions without hanging up. If they do hang up we can hope that this behavior is not related to their political views.

**Example 3** (like problem 1.6) Using a **random number table** : Of 75 restaurants in a downtown district, use table 1.3 on p9 of the text to randomly select a sample of  $n=6$  restaurants for the health inspector to monitor for compliance with city food safety requirements.

I'll start with row 7 and columns 15 and 16 and read down the columns : The pairs of digits thus found in the random number table are 88 (we discard this since 88 is greater than 75) , and then 45, 61, 52, 75, 23, 68. Had we obtained a repeat of a given number we would also discard such repeats until 6 different numbers are obtained between 1 and 75. (I.e. We pick 45 since in row 8 we have a 4 in column 15 and a 5 in column 16, 61 since in row 8 we have a 6 in column 15 and a 1 in column 16 etc.) Thus the health inspector should visit the 6 restaurants numbered as above (with the number 45 being the first on his list).

**Example 4** (like problem 1.7) Extending the 16 slot depth data in Table 1.1 on p4 and the x-bar chart in Figure 1.1 on p5, suppose two new samples of 3 ceramic parts each were measured after the machine was repaired yielding sample 17 : 215, 217, 216 having  $\bar{x}_{17}=216$  and  $\bar{x}_{18}=217$ . To get the new value of the mean of the sample means  $\bar{x}_{new}^{(18)}$  based on the two new sample mean values (for a total of 18 sample mean values) we don't want to have to recompute the sums used previously. Rather use

$$\bar{x}_{new}^{(18)} = \frac{1}{18} (16 * \bar{x}_{old}^{(16)} + \bar{x}_{17} + \bar{x}_{18})$$

which recursively updates the new value of the mean of the sample means based on old computed value and the 2 new observations only. This simple idea has a higher dimensional generalization useful for prediction of time series called a **Kalman Filter** (which itself has a generalization called the **extended Kalman filter**).

## Lecture 2 : Descriptive Statistics

**Measures of center or location** : Given a collection of **n** data values assigned to variables

$$x_1, x_2, \dots, x_n \text{ the sample mean } \bar{x} = \frac{1}{n} S_n = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$

is commonly used as a single number description summarizing the center (or location or central tendency) of the data and as an estimate of the actual population parameter namely the population mean  $\mu$ . (Here

$S_n = x_1 + x_2 + \dots + x_n$  is the  $n^{\text{th}}$  partial sum of the variables.) But other choices are possible. Another commonly used measure of location or center is the **sample median**  $\tilde{x}$ . If we rearrange the sample data in increasing order we can assign these increasing values to the **order statistics** variables denoted

$$\min = x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} = \max$$

Then when the number of data values  $n$  is odd the sample median is defined as the value in the middle of the ordered data, while when  $n$  is even, the sample median is the average of the two middle values. Equivalently  $\tilde{x} = Q_2$  the sample median is the same as the second quartile i.e. the value 50% of the way through the ordered data. To be precise we use Richard Johnson's

**Definition** of the **sample median** for  $p = 1/2$  (and of **quartiles & percentiles** for other  $p$ ) :

If  $np = k$  is an integer (i.e. when  $n$  even) then we take the average of the two values in the middle :

$$\text{median } \tilde{x} = Q_2 = \frac{x_k + x_{k+1}}{2}$$

while if  $np$  is not an integer we round up to  $k$  and define the second quartile as :

$$\text{median } \tilde{x} = Q_2 = x_{(k)} \text{ the } k\text{-th ordered value}$$

which is the single data value in the middle of the ordered data. The other quartiles (and more generally percentiles) are defined by exactly the same procedure except that for the first quartile  $Q_1$  (the 25% mark) we take  $p = 1/4 = 25\%$  while for the third quartile  $Q_3$  (the 75% mark) we take  $p = 3/4$  etc.

Note: Other authors may use a slightly different definition of the quartiles than Johnson's. When the sample size is large it won't make much difference.

One other measure of location sometimes encountered is the **mode** which is the most frequently occurring data value. In the case of a frequency distribution which is bell shaped this value is the value

on the x-axis associated to the unique peak of frequency (the y-coordinate) but it is possible that the mode is not unique such as in a **bimodal** distribution in which case one has two data values where the peaks of identical frequencies (i.e. heights) occur.

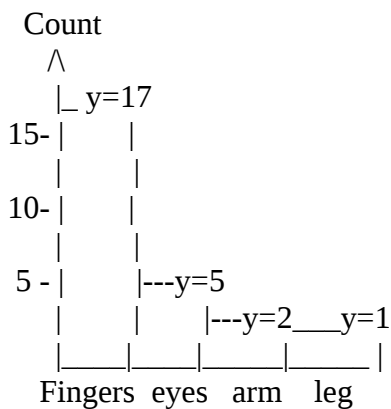
**Visualizing data :** (sometimes part of what is called **exploratory data analysis** )

**Dotplots and Pareto diagrams :**

A Pareto chart is a bar chart showing the largest counts of categories that the data falls under in decreasing order from right to left with a possibly larger category on the right of everything else left over.

**Problem 2.1** Accidents at a potato chip plant are characterized by the area of the human body injured. For the accident body location counts broken down into fingers :17, eyes: 5, arm: 2 , leg: 1

a) Draw a Pareto chart : I have yet to get R graphics images into pdf files so apologies for poor bar charts



What percentage of accidents occur for

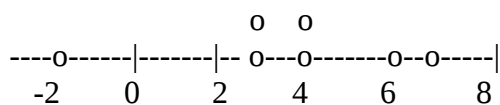
b) Fingers? Note that the total accident count is

total count = 17+5+2+1 = 25. Thus the fraction for fingers is 17/25 = 68%

c) Fingers and eyes : (17+5) / 25 = 22/25 = 88%

The dotplot of the 7 deviations : 3, 6, -2, 4, 7, 4, 3

(observed speed minus target speed ) of cutting speed of a lathe given in Figure 2.2 on p14 looks something like



If the lathe were behaving exactly at the target speed set by the controller, the deviations would all be zero. Ideally the deviations would be centered about zero with half negative and half positive but here almost all are positive with the sample mean of the deviations being

$$\bar{x} = \frac{3+6-2+4+7+4+3}{7} = \frac{25}{7} \text{ which slightly more than 3 and } \frac{1}{2}.$$

Thus we conclude that the lathe is running fast.

**Histogram example** (somewhat like **problem 2.9:**) As another example of a bar chart we look at the histogram example obtained from the 58 data values given in increasing order on p22 of the text and summarized in Figure 2.8. The picture in figure 2.8 of the text was computer generated but we were not told what the identical width used was for each of the seven class intervals pictured. We take a guess that comes close to reproducing the picture. Namely the intervals represented by the 7 bars of the histogram in the graph extend from slightly before the smallest data value 66.4 to slightly larger than the largest data value 75.3 so we will divide up the interval from 66.3 to 75.4 into 7 equal **class intervals** each of **width** 1.3 ( = (75.4-66.3)/7 ) So that the intervals do not overlap, we follow the **left-endpoint convention** of including the left but not the right endpoint within each class interval. Note that the difference of the endpoint values is the width 1.3. Since the data are already ordered it is easy to compute the frequency counts (i.e. the number of the 58 values listed which lie in that class interval) for each of the seven intervals so obtained :

Interval	Frequency Count	cumulative frequency count	relative frequency (percent or fraction of total)	cumulative relative frequency percent
[66.3, 67.6)	1	1	1/58 = 1.7%	1.7%
[67.6, 68.9)	7	8	7/58 = 12.0 %	13.7%
[68.9, 70.2)	16	24	16/58 = 27.6%	41.3%
[70.2, 71.5)	14	38	14/58 = 24.2%	65.5%
[71.5, 72.8)	13	51	13/58 = 22.4%	87.9%
[72.8, 74.1)	4	55	4/58 = 6.9%	94.8%
[74.1, 75.4)	3	58	3/58 = 5.2%	100 %
total count :	58	total percent	1 = 100%	

These values are not exactly as shown in the graph in Figure 2.8 but they are close.

Note that the continuous **bell shaped curve** fitted over the discontinuous histogram bar graph pictured in Figure 2.8 is a common situation : as the number of data values n gets large, a bar graph histogram typically approaches some continuous curve distribution more and more closely.

An example of **cumulative frequency** is plotted in Figure 2.9 on the next page of the book for the different set of data of sulfur oxide emissions given on p16 of the text.

The dots are placed at the beginning of each class interval with the height of the dot representing the cumulative frequency up to that point. The dots are connected by line segments to give the **ogive** or cumulative frequency graph.

(Note that the cumulative counts given in the Pareto diagram in figure 2.1 on p14 were drawn differently with the dots falling in the middle of each bar which would correspond to the so called **class mark** which is the value in the middle of each class interval if we were dealing with class intervals rather than categories as in the Pareto diagram. )

**Stem and leaf display** : A stem and leaf diagram is kind of like a histogram but provides more detailed information. If we were trying to summarize the (ordered) data set of 16 values

13, 13, 15, 16, 17, 17, 19, 23, 25, 26, 26, 28, 28, 28, 30, 32

we could break off the tens column which represent the class intervals [10-19) , [20-29) [30-39)

and describe the above data as

1| 3, 3, 5, 6, 7, 7, 9 ( for a count of 7 values between 10 and 19)  
 2| 3, 5, 6, 6, 8, 8, 8 ( a count of 7 values between 20 and 29 )  
 3| 0, 2 ( a count of 2 values between 30 and 39 )

Note although a stem and leaf display is like a histogram turned on its side , a histogram would only plot the counts in each class interval (in this case the counts 7, 7 and 2 ) and would loose track of the individual data values contained in the stem and leaf display.

Stem and leaf displays can also record more complicated data sets such as

1.4 | 51 68 74 ( the class interval here is [1.4 , 1.5) )

1.5 | 23 34 89 ( class interval [1.5, 1.6) )

for representing the data values 1.451, 1.468, 1.474, 1.523, 1.534, 1.589

or the stem and leaf display 23|1 , 3, 6 for the values 231, 233, 236 etc. ( class interval [230, 239 ) say )

**Scatter plot diagram :** For a final example of a graphical display we consider the scatter plot diagram for the **Space Shuttle Challenger disaster data** (which plots one variable against another useful for viewing correlations between data sets which we will study in chapter 11) . You can view the graph of this data on page 2 of the postscript file with link

<http://www.stat.duke.edu/Spring03/sta113/Notes/lec11.ps>

and see <http://www.math.yorku.ca/SCS/Gallery/missed.html>

and <http://web.grinnell.edu/individuals/kuipers/stat2labs/topics.html> (<--under topic: categorical data )

This disaster resulted in the deaths of the seven astronauts, cost billions of dollars and set the space program back year. The black curve is an extrapolation from the data represented by dots of the likelihood of O-ring failure which at 30 degrees Farenheit is around 80%. It is surrounded by two other curves which tell us that with say 95% probability the correct failure probability of the extrapolated curve lies between the upper and lower curves. The engineers plotted 7 data points : the y coordinate of the points represented the number of distressed O-rings for failed O-ring data. Thinking it was not informative they ignored 17 data points, the larger part of the relevant data that were the dots representing zero stressed O-rings, i.e. no O-ring failures (dots along the x-axis, where y=0 failures occurred) . These dots appear underneath the U shaped sequence of dots which represent flights that experienced 1 , 2 , or 3 stressed O-rings and you will note that the x-axis represents temperature. The engineers had no existing data for temperatures below 55 degrees Farenheit yet the Challenger took off when the temperature was 30 degrees Farenheit. The temperatures at which no O-ring failures occurred were all in the range from 65 degrees to around 82 degrees Farenheit. To be fair, engineers were aware of and complained about the O-ring problem but managers didn't listen.

Mark Twain, author of Huckleberry Finn and the Adventures of Tom Sawyer once said

“Lies, Damn Lies and Statistics”.

You might want to read the Statistics book by this name. In the case of the Challenger, the engineers weren't exactly lying, rather they were using a lawyer's definition of truth: they ignored more than half of the relevant data ! Applied statisticians like R. Snedecor have told us time and time again :

“ In God We Trust, Others Must Have Data !”

The Challenger disaster emphasizes this point.

**Boxplots** and their relatives are discussed below.

**Measures of spread about the center** ( variation/dispersion/width) : If you told a visitor from another galaxy that the average height of an adult human on earth is 71” they might wonder if the smallest is one milimeter tall and the largest is a kilometer in height since the mean value does not include the variability (variance) of the data

When the center is given by the median, which to be able to compute WEEK 1 page 10 requires the ordered data from smallest to largest, such variability measures include the **range**

$$\text{range} = \max - \min = x_{(n)} - x_{(1)}$$

which measures distance between extremes, and another, also in terms of the ordered data, is the **inter-quartile range** also known as the **fourth spread**  $f_s$  :

$$f_s = \text{inter-quartile range} = Q_3 - Q_1$$

which is the distance between the 3<sup>rd</sup> quartile and the 1<sup>st</sup> quartile (i.e. between the data value 75% of the way through the ordered data and the value 25% of the way through).

**Outliers** : All of the quartiles including the median and hence also the inter-quartile range measure of spread are not sensitive to **outliers** i.e. to extreme values of the data not representative of the typical behavior of the data, since for example the later disregards the largest and smallest 25% of the (ordered) data values. (Of course the range is sensitive since it uses the most extreme values.) This is not true of the sample mean, sample variance and sample standard deviation discussed below which are all sensitive to outliers since they use *all* the data including the outliers in their calculation. One can however speak of a **trimmed sample mean** or **trimmed sample variance**, for example a 5% trimmed mean  $\bar{x}_{tr(5)}$  disregards (trims) the largest and smallest 5% of the ordered data in its calculation. As in the discussion of modified boxplots in the text we can speak of a **mild outlier** as being more than 1.5 times but less than 3 times the inter-quartile range to the left of the 1<sup>st</sup> quartile or to the right of the 3<sup>rd</sup> quartile, and of an **extreme outlier** if more than 3 times this inter-quartile distance  $f_s$  .

**Sample Variance** and **sample standard deviation** : When the center is given by the mean, spread is usually measured in terms of the sum of squares of the **deviations from the (sample) mean**  $(x_i - \bar{x})$  of the data . Namely the **sample variance**  $s^2$  (with  $n-1$  **degrees of freedom**) is defined by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} S_{xx} \quad \text{where} \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{is the sum of squared deviations from}$$

the mean of the  $x_i$  data values. Dividing by  $n-1$  , not  $n$  above, is needed to insure  $E[s^2] = \sigma^2$  i.e. sample variance  $s^2$  is an unbiased estimator of population variance  $\sigma^2$  . You were asked to show

in problem 2.51 the alternate but equivalent computational formula  $s^2 = \frac{1}{n-1} \left( \sum x_k^2 - \frac{(\sum x_k)^2}{n} \right)$  .

Note : we could **not** have use the sum of the deviations themselves to define a measure of spread since problem 2.50 of the homework shows these deviations (also called **residuals**) add to zero

$$\sum (x_i - \bar{x}) = 0$$

and zero is not a terribly informative measure of spread. (We say this constraint “eliminates one degree of freedom” since if we know the sample mean, the  $n$  data values are no longer independent.)

We could have used the absolute value of the deviations however or some power of that and this is sometimes done but is not as easy nor as customary to work with mathematically. Finally the **sample standard deviation** is another measure of spread defined as the square root of the sample variance

$$s = \sqrt{s^2}$$

which has the advantage over the variance that if we are say measuring deviations of height in meters from the mean height of some height data values, the variance will be in square meters (the wrong units) whereas the standard deviation will be in the correct units : meters for height.

**Boxplots** : We plotted the neutrino data given on p15 and p36 of the text on a line and discussed the boxplot and modified boxplot given on page 36 of the text. We verified the quartile computation

procedure of Johnson on this data which illustrates the case where  $np$  is not an integer : i.e. with the  $n=11$  ordered values of neutrino inter-arrival times

.021, .107, .179, .19, .196, .283, .58, .854, 1.18, 2.0, 7.3

for the 1<sup>st</sup> quartile we take  $p=1/4$  so  $np = 11/4$  gets rounded up to 3 and then

$$x_{(3)} = .179 = Q_1$$

gives the 1<sup>st</sup> quartile. Similarly for the median (2<sup>nd</sup> quartile) we take  $p=1/2$  so  $np= 11/2$  gets rounded up to 6 and so

$$Q_2 = x_{(6)} = .283$$

and finally for the 3<sup>rd</sup> quartile (marking 75% of the ordered data)  $p=3/4$  so  $np=33/4$  gets rounded up to 9 or

$$Q_3 = x_{(9)} = 1.18$$

gives the 3<sup>rd</sup> quartile. The rectangle (**box**) portion of the boxplot extends from the first to the third quartile with the vertical line dividing the rectangle at the median (2<sup>nd</sup> quartile) . A line segment extends from the minimum value .021 to the left of the rectangle (1<sup>st</sup> quartile) and another line segment extends to the right of the rectangle (3<sup>rd</sup> quartile) to the maximum value 7.3. In the **modified boxplot** this line segment only extends to the 2<sup>nd</sup> to last data value 2.0 since the value 7.3 is an outlier which is more than 1.5 times the interquartile range ( fourth spread )  $Q_3 - Q_1 = 1.001$  from the 3<sup>rd</sup> quartile (i.e from the right of the box )

**Example :** Using the data of **problem 2.31** of the text illustrates the quartile calculation procedure when  $np$  is an integer. Here the data set consists of the  $n=4$  deviations (observation – specification) of critical crank bore diameter in ten thousandths of an inch : -6, 1, -4, -3. Thus the ordered data are

$$x_{(1)} = -6, x_{(2)} = -4, x_{(3)} = -3, \text{ and } x_{(4)} = 1 .$$

Since for all quartiles including the median,  $p$  is a multiple of  $1/4$  and  $n=4$ , we have  $np$  is an integer in each case so the procedure says to average the integer and the next so that the quartiles are

$$Q_1 = \frac{x_{(1)} + x_{(2)}}{2} = -5, \quad Q_2 = \frac{x_{(2)} + x_{(3)}}{2} = -\frac{7}{2}, \quad Q_3 = \frac{x_{(3)} + x_{(4)}}{2} = -1 .$$

We will now do the **problem 2.31** of the text :

a) The sample mean is  $\bar{x} = \frac{-6+1-4-3}{4} = \frac{(-12)}{4} = -3$

b) We compute the sample standard deviation in two ways using the two formulas for the sample variance

First the sample variance is (with  $n=4$ ) :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{3} ((-6 - (-3))^2 + (1 - (-3))^2 + (-4 - (-3))^2 + (-3 - (-3))^2) = \frac{26}{3}$$

Then the sample standard deviation is the square root of this or  $s = \sqrt{s^2} \approx 2.94$  .

We can also use the alternate formula for the sample variance

$$s^2 = \frac{1}{n-1} \left( \sum x_k^2 - \frac{(\sum x_k)^2}{n} \right) \text{ where for both sums the index } k \text{ ranges from } k=1 \text{ to } k=n \text{ so with } n=4$$

our example gives  $s^2 = \frac{1}{3} ((6^2 + 1^2 + 4^2 + 3^2) - \frac{(-12)^2}{4}) = \frac{26}{3}$  as before.

- c) The sample mean given in part a above says that the average deviation is 3 ten thousandths of an inch smaller than the specified bore diameter so the answer is that the bore hole is too small by 3/10000 of an inch.

Finally we look at another variance computation example which illustrates that the variance does not

