Variability

The average score in a distribution is important in many research contexts. So too is another set of statistics that quantify **how variable** (or "how dispersed") the scores tend to be. Do the scores vary a lot, or do they tend to be very similar or near each other in value? Sometimes **variability** in scores is the central issue in a research question. Variability is a quantitative concept, so none of this applies to distributions of qualitative data.

There are many intuitively appealing but little used measures of variability. The **range**, for example, is the difference between the largest and smallest score in the data set. The **interquartile range** or **IQR** is the difference between what we will later call the 25th and 75th percentile scores. By far the most widely used measures of variability are those to do with averaging how spread out the scores are from the mean. These are the **Sums of Squares** (SS), the **standard deviation** (s, or sd), and the **variance** (s^2 or "var").

Sums of squares

Consider for a minute transforming the original data in Table 4.1 to **deviations**. That is, each score is converted to the difference between that score and the mean. So all 1s become 1 minus 1.864 or -0.864. All 0s become 0 minus 1.864 or -1.864. All 2s become 2 minus 1.864 or 0.136. It might be obvious to you that if the scores tended to differ a lot from the mean, then these differences would tend to be large (ignoring the sign), whereas these differences would tend to be small (ignoring sign) if the scores tended vary little from the mean.

The measure typically used to quantify variability in a distribution is based on the concept of **average squared deviation** from the mean.

Let's take each difference from the mean and square it. Then, let's add up these squared deviations. When you do this you have the sum of the squared deviations (which is then reduced to "Sums of Squares", or SS). Its formula is

$$SS = \Sigma\left(X - \bar{X}\right)^2 = \Sigma X^2 - \frac{\left(\Sigma X\right)^2}{N}$$

The left-hand side of the equation is the definitional formula and the right hand side is the computational formula. SPSS output does not give the Sums of Squares for a variable when you choose Frequencies. However, many later statistical procedures do give this as part of the output. It's value lies in summarising the total amount of **variability** in the variable being examined. For the phone call data SS = 848.74 (calculated by the method below). The size of this number depends on the size of the numbers in the data and how much data there is (i.e., the sample size). There are no units for SS.

Sometimes there is confusion about the terms **variability** and **variance**. Variability refers to the Sums of Squares for a variable, while variance refers to the Sums of Squares divided by N-1. Sums of Squares are widely used because they are additive. Once we divide by N-1, the additive property disappears. When we later talk about the "proportion of variance explained" we really mean the "proportion of **variability** explained". If a variable X explains 56% of the variability in variable Y it refers to the proportion of Y's Sums of Squares that is attributable to variable X's Sums of Squares.

Variance

Variance (of a sample) is defined as

$$\begin{aligned}\text{var} &= \frac{\Sigma\left(X-\bar{X}\right)^2}{N-1}\\ &= \frac{SS}{N-1}\\ &= \frac{SS}{df}\end{aligned}$$

Once we divide the Sums of Squares by N-1 we get the sample **variance** which can be thought of as an **averaged sums of squares**. While important in statistical theory and in many statistical computations, it has the problem of being in squared units and is therefore difficult to manipulate and visualise.

To get the SS for a variable from the Frequencies information, you need to rearrange the above equation to get

$$\begin{aligned}\text{var} &= \frac{SS}{N-1}\\ SS &= \text{var}\times\left(N-1\right)\end{aligned}$$

Standard deviation

To overcome the problem of dealing with squared units, statisticians take the square root of the variance to get the standard deviation.

The **standard deviation** (for a sample) is defined symbolically as

$$s = \sqrt{\text{var}} = \sqrt{\frac{\Sigma\left(X-\bar{X}\right)^2}{N-1}}$$

So if the scores in the data were 5, 7, 6, 1, and 8, their squared differences from the mean would be 0.16 (from $[5\text{-}5.4]^2$), 2.56 (from $[7\text{-}5.4]^2$), 0.36 (from $[6\text{-}5.4]^2$), 19.36 (from $[1\text{-}5.4]^2$), and 6.76 (from $[8\text{-}5.4]^2$). The mean of these squared deviations is 5.84 and its square root is 2.41 (if dividing by N), which is the standard deviation of these scores. **The standard deviation is defined as the average amount by which scores in a distribution differ from the mean, ignoring the sign of the difference**. Sometimes, the standard deviation is defined as the average **distance** between any score in a distribution and the mean of the distribution.

The above formula is the definition for a **sample** standard deviation. To calculate the standard deviation for a **population**, N is used in the denominator instead of N-1. Suffice it to say that in most contexts, regardless of the purpose of your data analysis, computer programs will print the result from the sample sd. **So we will use the second formula as our definitional formula for the standard deviation**, even though conceptually dividing by N makes more sense (i.e., dividing by how many scores there are to get the average). When N is fairly large, the difference between the different formulas is small and trivial.

Using the N-1 version of the formula, we still define the standard deviation as the average amount by which scores in a distribution differ from the mean, ignoring the sign of the difference, even though this isn’t a true average using this formula.

The standard deviation in our phone call data is 2.196, from the SPSS printout in Output 4.2. So the mean number of phone call is 1.864 with a standard deviation of 2.196. The units are now the same as the original data. But, is this a large standard deviation? It is hard to say. In a normal distribution the mean and standard deviation are independent of each other. That is one could be large or small and the other large or small without any influence on each other. However, in reality they are often linked so that larger, means tend to have larger standard deviations. This leads into the area of transformations that are a way of re-establishing this independence.

$$\text{Coefficient of variation} = \frac{\text{Standard deviation}}{\text{Mean}}$$

A useful measure of a distribution that is sometimes used is the ratio of the standard deviation to the mean

The standard deviation has one undesirable feature. Like the mean, one or two extreme scores easily influence the standard deviation. So really atypical scores in a distribution ("outliers") can wildly change the distribution’s standard deviation. Here, adding a score of 200 increases the sd from 2.196 to 15.0115, a seven-fold increase! Because both of these descriptive statistics are influenced by extreme cases, it is important to note when extreme values exist in your data and might be influencing your statistics. How to define "extreme," and what to do if you have extreme data points is a controversial and complex topic out of the scope of this class.

The Normal Distribution

One of the more important distributions in statistics is the "normal" distribution. The normal distribution is depicted in Figure 4.7 below. Notice a few things about its features.

First, it is a symmetrical distribution, meaning that the left half of the normal distribution is a mirror image of the right half.

Second, most of the scores in a normal distribution tend to occur near the center, while more extreme scores on either side of the center become increasingly rare. As the distance from the center increases, the frequency of scores decreases.

Third, the mean, median, and mode of the normal distribution are the same.

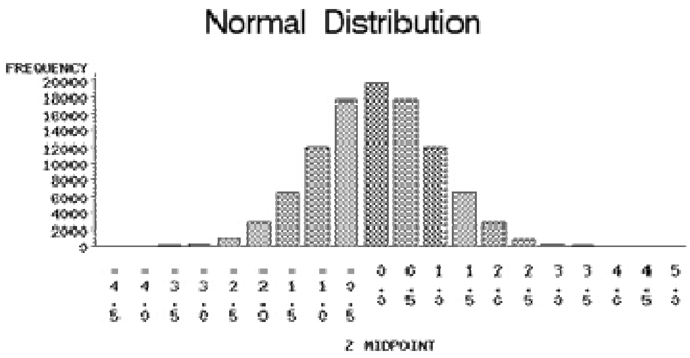

Figure. 4.7 Histogram of Computer-Generated Normal Data

Some people claim that the many variables are distributed normally, including such things are heights, weight, age. While there is some truth to the claim that many distributions are similar to the normal distribution, however, many of the things that appear normally distributed in fact are not. Still, many of the variables that we study, when measured, do **approximate a normal distribution**, so it is worth understanding its properties. As well, a prominent assumption or requirement for statistical tests is normality in the parent population from which the scores came. We can only check normality in the parent population by checking normality in the sample of scores we have. For most purposes, an approximately normal curve is fine. That is, one that does not deviate significantly from our symmetry.

Transformations

In many ways, the way in which we measure variables is very arbitrary. If we measured height, we could use a ruler and measure in feet and inches, or a ruler in centimetres and millimetres. Now these measurements can be converted from one to the other by a rule or formula. One hand = x inches = y centimetres. We do it all the time in the real world. The position of the hands of the clock (angle in degrees) measures the time (in hours and minutes,), level of mercury in a thermometer (centimetres) = temperature (in degrees), scores on a piece of paper (scores or points) = IQ (in some other arbitrary units). Therefore, the data you have are not sacred in terms of the actual numbers assigned. So, there are many options available to us in terms of converting scores from one metric to another or to another set of points on the same metric. The scaling of scores in the Higher School Certificate examination is a common example.

In statistical practice there are a number of transformations that are in common use. The ones we will mention are dichotomisation, standardisation, normalising.

Dichotomisation

Variable can be classified in many ways. One way is continuous or discrete. A continuous variable (e.g., length) takes on many values, it is not restricted to just a few values such as gender (takes two values) or days of the week (takes on seven values). A variable that takes on only two values is **a dichotomous variable**. Male/female, yes/no, agree/disagree, true/false, present/absent, less than/more than, lowest half/highest half, experimental group/control group, are all examples of dichotomous variables.

A continuous variable is said to contain more information about a construct because it measures it more accurately or more sensitively. Asking a person if they Agree or Disagree to a question does not give as much information about that person’s level of agreement as does a seven point scale

1	2	3	4	5	6	7
Strongly Disagree	Moderately Disagree	Somewhat Disagree	Ambivalent	Somewhat Agree	Moderately Agree	Strongly Agree

So, in general, you should use continuous variables wherever possible. However, there are times when only dichotomous measures are possible. There are also times when dichotomous variables are useful in their own right. This is mainly when you are only interested in broad comparisons. To compare two or three or four broad groupings can sometimes lead to a clearer understanding of relationships in data than considering continuous data.

It is possible to convert continuous measurements to smaller numbers of categories by **recoding** the variable. We could recode height into below average or above average (call them 0 and 1, or 1 and 2). We could convert age into three categories of young, middle, and old. We could recode the above scale to Disagree (all responses of 1, 2, and 3) and Agree (all responses of 5, 6, and 7) and ignore a response of 4. Whenever you do this you are losing information about the construct being measured. Note it is not possible to convert from a dichotomous variable to a higher number of categories. If you only have Agree/Disagree data, you cannot recode into a seven-point scale. This is because you do not have the information needed to regain that level of sensitivity.

Report Writing

The beauty of scientific studies is the ability to convert data to information that advances human learning either by supporting or rejecting earlier theories and hypotheses. This can only be achieved through the art of good report writing. Report writing is an art that must be mastered by every student in order to be able to translate research findings into recommendations. Often student fail woefully not due to poor experiment but due poor reporting. This becomes evidence during seminar presentation and project write-ups.

A good report must convey the followings:

- The rationale for the study
- The methods used
- The results/findings
- Conclusiom

Rationale

In any study there must be convincing reasons for the study. This is because research involves human resources, time and energy as such it is nonsensical to conduct a research without a good justification or rational. A good rational must state there underlining reasons for this research and the advances it would make if conducted. For example consider a study to investigate the prevalence of HIV in Abeokuta. The question to ask is why must we know the prevalence of HIV in Abeokuta? There many be lots of rationales for these. It could be that we do not have such information, or we would wan to advocate for more intervention in HIV, perhaps we are curious to know if there are more HIV prevalence in male or female or in children. These are good rationale for your study.

The rationale will lead to the objective of the study.

Methods

In any study there must be testable methods that are reliable, repeatable and verifiable. A good report must provide clear methodology of the study that can be easily followed by another research any way in the world. Often research who do not disclose their methods sufficiently are not only killing research but are also undermining them self. If you want to hide information, hide your results and not your methods. All methods must be repeatable.

Results/Findings

Converting data into meaningful information is the hallmark of good reporting. Decide on the best methods to convey this information as simple as possible. Your results need not to be confusing. If a table will better convey more information than a chart, then use a table.

Sometimes you are limited to type of presentation by the kind of variables in the study. Result must address stated objectives of the study. Statistical inferences must be properly stated and any significant relationship must be vigorously justified. Poor statistical report is very dangerous as it could give false conclusion. Student must not “pad” data to move toward a predetermine conclusion just to impress their supervisors

Conclusion

A good report comes with a sound conclusion. A conclusion essential restated key findings and recommendation either for further study or accept or reject a hypothesis. The conclusion must justify why the study was conducted.