Frequency Distribution Displays

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

1. Discrete vs. Continuous Variables
   - Introduction
   - Discrete Variables
   - Continuous Variables
   - Nominal Limits
   - Real Limits
2. Levels of Measurement
   - Introduction
   - Nominal Measurement
   - Ordinal Measurement
   - Interval Measurement
   - Ratio Measurement
   - Permissible Transforms
3. A Frequency Distribution Table
4. The Grouped Frequency Distribution Table
5. The Frequency Histogram
6. The Pie Chart
7. Quantiles

Introduction

- The variables we deal with in science can be divided into several categories according to some fundamental distinctions.
- One key distinction is between variables that are fundamentally *discrete* versus those that are fundamentally *continuous*.

Discrete Variables

- Some variables are inherently discrete. A discrete variable can take on only a countably finite number of values.
- Values for discrete variables are usually expressed in integers.

Example (Discrete Variables)

Some examples of discrete variables are:

- Number of children
- Number of students in a classroom

Discrete Variables

- When I record an observed value of a discrete variable, there is no "inherent round-off" in the data.
- For example, if I record that Cindy has two brothers, then I assume that she has exactly two brothers — not 2.2, or 1.8.

Continuous Variables

- Many variables in science are inherently continuous.
- A continuous variable can (for all practical purposes) take on an uncountably infinite number of values within a particular range.

Example (Continuous Variables)

Some examples of continuous variables are:

- A person's height in inches
- The altitude of Nashville in number of feet above sea level.

Continuous Variables

Implied Round-Off

- Since continuous variables are infinitely “fine-grained,” we must generally truncate the level of precision to which we record them.
- When we truncate, we generally round off to a particular number of decimal places.
- Generally, numbers recorded to k decimal places are “rounded up” if the $k + 1$ decimal place is 5 or larger, and are “rounded down” if the $k + 1$ decimal place is 4 or smaller.

Example (Rounding)

Suppose we were recording values to the nearest tenth of a point. Then 2.153... would be rounded up to 2.2, and 3.049... would be rounded down to 3.0.

Nominal Limits

- When we record data, there may (or may not) be an implicit roundoff process involved.
- For example, suppose I tell you that there are 4 women in my focus group who have either 1 or 2 children, and there is at least one woman with 1 child and at least one woman with 2.
- What is the minimum number of children that any woman might have in that group?
- Since number of children is a discrete quantity, the answer is 1.
- So, we can say that the nominal limits for that group are 1 to 2.

Nominal Limits

- Suppose I have a group of five men, and the heaviest weighs 190 pounds, the lightest weighs 142 on a scale that rounds to the nearest whole pound.
- We say that the nominal limits for this group are 142 and 190.
- However, the *real* limits for this group are different, as we see in the next section.

Real Limits

- In our immediately preceding example, we discussed a case in which there were 5 men, with the heaviest recorded as 190 pounds, the lightest as 142.
- Assuming perfect measurement with a roundoff to the nearest whole pound, what is the largest weight in pounds that the man listed at 190 might weigh?
- What is the lightest that the man listed at 142 might weigh?

Real Limits

- Strictly speaking, the man listed at 190 might weigh anywhere in the interval from 189.5 up to but not including 190.5. To keep matters simple, let's just say 190.5.
- In a similar vein, the man listed at 142 might weigh as little as 141.5 (and still be rounded up to 142).
- So we might say that the real limits for the group of men are $141.5 \geq X < 190.5$.
- Usually, the real limits extend beyond the nominal limits by *half the level of rounding*.

Real Limits

Example

Real Limits[Weight of Trucks] Suppose you have 6 trucks, and they are weighed on a scale that rounds to the nearest 100 pounds. The trucks' weights are recorded as 12200, 6400, 9800, 12900, 19100, and 11500. What are the real limits for this group?

Real Limits

Example

Real Limits[Weight of Trucks] Suppose you have 6 trucks, and they are weighed on a scale that rounds to the nearest 100 pounds. The trucks' weights are recorded as 12200, 6400, 9800, 12900, 19100, and 11500. What are the real limits for this group?

Answer. The lightest truck could weigh as little as 6350 and still be rounded up to 6400, and the heaviest could weigh up to (but not including) 19150 and still be recorded as 19100. So the real limits are 6350 and 19150.

Levels of Measurement

- Measurement is the process of assigning numbers to objects in order to represent properties of the objects. The process is so familiar that we often overlook its fundamental characteristics.
- In this discussion, we introduce the notion of *level of measurement.*
- The level of measurement that numbers achieve is, roughly speaking, the extent to which certain properties of the numbers match up with corresponding properties of the objects they are measuring.

Levels of Measurement

Dangers to Avoid

Understanding levels of measurement can help us avoid certain dangers.

- Attaching unwarranted significance to aspects of the numbers that do not convey meaningful information.
- Failing to simply data when would easily do so.
- Manipulating our data in ways that destroy information.
- Performing meaningless statistical operations on the data.

Nominal Measurement

- Some attributes are not inherently ordered. They are the same, or different.
- Others might be ordered, but the numerical assignment does not capture the ordering.
- In such cases, we say that our numbers have achieved a *nominal level of measurement.*

Nominal Measurement

Example (Nominal Measurement)

Here are some examples of nominal measurement:

- *Football Players' Numbers.* There is no necessary connection between a football player's number and his value. The numbers simply identify different players with different numbers.
- *Student ID Numbers.*
- *A Numerical Code for a Student's Gender.* Suppose your data bank had a 1 if the student is male, a 2 if the student is female. These codes do not imply any inherent ordering of the sexes, but simply indicate which student is which.

Ordinal Measurement

- Many attributes have an ordering. For example, if you select two people, it is almost certainly the case that one is either taller than the other or shorter.
- If an assignment of numbers to objects captures only the Same-Difference property and the Ordering property, then we say that the numbers have achieved an *ordinal level of measurement*.

Ordinal Measurement

Example (Ordinal Measurement)

Some examples of ordinal measurement:

- *Class Ranks in a University Course.* There is almost certainly additional useful information in the actual course grades than is displayed in the Class Ranks. For example, suppose 3 students had percentage grades of 95,94, and 67. There is a very small difference between the first two students, and a huge difference between the second and third ranking students. Knowing the rankings of 1,2,3 does not convey that information.
- *Rankings in a Beauty Contest.*
- *Finishing Positions in a NASCAR Race.*

Interval Measurement

- A set of numbers assigned to measure a group of objects on some attribute achieves an *interval level of measurement* if, besides correctly characterizing the same-difference and ordering properties, the numbers also convey correct information about the relative spacing of the objects with respect to the attribute of interest.

Interval Measurement

Example (Interval Measurement)

- Suppose 5 women have given birth to 0,0,1,2 and 4 children.
- The five women are assigned *child production scores* of 1,1,2,3,and 4 respectively.
- Notice that the relative spacing of the women is 0 (between the first two women), 1 (between the second and third women), 1 (between the third and fourth women), and 2 (beween the fourth and fifth women).
- This *relative spacing* is lost with the numerical assignment. It correctly captures same-difference and order, but not the relative spacing.
- On the other hand, we can say that the numerical assignment 1,1,2,3,5 does achieve an interval level of measurement.
- *IMPORTANT*. An assignment of 2,2,4,6,10 would also achieve an interval level of measurement, because the *relative spacing* of the assigned scores matches the relative spacing of the number of children, *and* the order and same-difference properties are also correctly captured.

Ratio Measurement

- If a set of numbers correctly captures all the properties of an interval level of measurement, *and also has a correct zero point and correctly matches ratios of the attribute being measured*, then it has achieved a ratio scale of measurement.

Ratio Measurement

Example (Ratio Measurement)

- In the preceding example, we considered the case five women with 0,0,1,2, and 4 children who were assigned *child production scores* of 1,1,2,3 and 5.
- These scores achieved an interval level of measurement because they correctly capture the same-difference property, the ordering, and the relative spacing of the attribute of interest.
- However, they did not achieve a ratio level of measurement, because, for example, the 4th woman had twice as many children as the third woman, but did not have a child production score that was twice as high.
- Also, the assigned numbers did not reflect a correct zero point. The first two women did not have any children, but did not receive a child production score of zero.
- Consider child production scores of 0,0,20,40, and 80 assigned to the five women. Would they achieve a ratio scale of measurement?

Permissible Transforms

- Measurement scales form a *hierarchy*.
- Ratio is better than Interval, which in turn is better than Ordinal, which in turn is better than Nominal.
- As you move up the hierarchy, more and more aspects of the attributes being measured are correctly captured by the numbers assigned.
- In this section, we ask the question, "What can you do to a list of numbers without reducing their level of measurement?"

Permissible Transforms

- In general, the higher the level of measurement, the less you can do.
- For nominal measurement, any transformation that preserves same-difference properties is permissible.
- For ordinal measurement, any listwise transformation of the form $Y = f(X)$ where f is a montonic, strictly increasing function, is permissible.
- For interval measurement, any listwise transformation of the form $Y = aX + b$, with $a > 0$ is permissible.
- For ratio measurement, only a listwise transformation of the form $Y = aX$, with $a > 0$ is permissible.

Permissible Transforms

- What will happen if you apply a transform to a list of numbers that is only permissible at a lower level?
- The answer is, you will drop the numbers to the level of measurement corresponding to the transform.

Permissible Transforms

Example (Appying a Non-Permissible Transform)

Suppose you have a list of numbers that have achieved an interval level of measurement, and you cube them all, i.e., apply a transform $Y = X^3$ to the entire list of numbers. In that case, since the function $f(X) = X^3$ is monotonic and strictly increasing, which is only permissible at the ordinal level or below, the numbers will be dropped to an ordinal level of measurement.

A Frequency Distribution

- Our first step in descriptive statistics is to characterize the data in a single group of observational units, with only one variable measured per unit.
- Where are the numbers on the number line?
- We can summarize this in tables, or (generally more effectively) in graphs.
- For example, suppose we have 15 students in a seminar, and their grades are

$$71, 71, 77, 80, 79, 75, 76, 72, 72, 74, 73, 71, 73, 78, 79$$

- We can summarize these grades in a *frequency distribution table* as shown in the table on the next slide.

A Frequency Distribution

X	f
80	1
79	2
78	1
77	1
76	1
75	1
74	1
73	2
72	2
71	3

A Frequency Distribution

- In the preceding table, X stands for the value, f for its frequency of occurrence.
- Using our summation notation from the previous lecture, we can deduce some well known facts about such tables.

A Frequency Distribution

- First of all, suppose we use the letter n to stand for the total number of observations in the data. (In this case, $n = 15$.)
- Furthermore, let k be the number of different values in the table. (In this case, $k = 10$, since all integer values from 71 to 80 are represented.)
- Then, in such tables, we can say that

$$\sum_{i=1}^{k} f_i = n \tag{1}$$

A Frequency Distribution

- In a similar vein, we can compute the sum of the observations summarized in the table as

$$\sum_{i=1}^{k} X_i f_i \tag{2}$$

and therefore their arithmetic average, or *mean* as

$$\bar{X}_{\bullet} = \frac{\sum_{i=1}^{k} X_i f_i}{\sum_{i=1}^{k} f_i} \tag{3}$$

Introduction

- Suppose we record weights in pounds of a group of 50 adult males, and obtain the following data.

Table : Weights of 50 Adult Males

165	105	147	170	169	195	170	162	178	187
168	166	195	128	186	138	205	142	90	115
147	143	159	142	134	166	201	108	123	164
139	143	163	178	111	165	191	194	173	165
165	127	131	157	110	146	156	151	171	117

- There are too many distinct values with too few replications to construct a meaningful frequency distribution table from the individual values.
- Instead, we break the number line into intervals, count how many numbers fall in each interval, and construct a *grouped frequency distribution table* from these data.

The Grouped Frequency Distribution Table

- The purpose of this kind of a table is to select a set of intervals on the number line, then count the number of values that fall into each interval.
- Connected with a grouped frequency distribution table are a number of glossary terms that we will define and illustrate with the current example.

The Grouped Frequency Distribution Table

- The first task in constructing a grouped frequency distribution table is to examine the numbers, and divide the range of the values into a reasonable number of intervals.
- Generally, we want to have between 8 and 12 intervals, but there are no hard and fast rules (although there are computer algorithms to select the number of intervals automatically).
- You want enough intervals to display a reasonable picture of distribution shape, and not so many intervals that there are many intervals with low counts. Generally, you also want all intervals to be the same width.
- Connected with a grouped frequency distribution table are a number of glossary terms that we will define and illustrate with the current example.

The Grouped Frequency Distribution Table

- Examining the data in the table of weights, we see that the recorded values range from 90 to 205. If we desire nice even interval endpoints, we might select an interval width of 10.
- This would result in 12 intervals each of width 10.
- We can count the number of numbers in each of the 12 intervals by hand, or, as we shall see later in the course, we can let a statistical program like R do it for us.

Grouped Frequency Distribution Table

Nominal vs. Real Intervals

- In setting up our interval limits, it is important to remember the distinction between *nominal limits* and *real limits* for the intervals
- Suppose that our first interval includes weights that are recorded as being between 90 and 99 pounds. These weights are rounded to the nearest whole pound.
- If your scale is perfectly accurate, but rounds to the nearest whole pound, then a weight of 99 pounds might stand for any value up to but not including 99.5, because, for example, a weight of 99.34 would be rounded down to 99.
- In a similar vein, we realize that a person whose weight is 90 pounds could be as light as 89.5, in which case the weight would be rounded up to 90.
- Putting these facts together, an interval with *nominal limits* of 90 and 99 has real limits of 89.5 and 99.5, and has a real width of 10.

Grouped Frequency Distribution Table

R Code

- Here is the code for the data, which, as you might recall, are in a variable called `wts`. This code uses a function, `grouped.frequency.table`, which I have created for you:

```
> lower <- seq(90,200,10)
> upper <- seq(99,209,10)
> grouped.frequency.table(wts,lower,upper,round.off=1)
   lower upper lower.real upper.real  f cum.f rel.f cum.rel.f
1    200   209      199.5      209.5  2    50  0.04      1.00
2    190   199      189.5      199.5  4    48  0.08      0.96
3    180   189      179.5      189.5  2    44  0.04      0.88
4    170   179      169.5      179.5  6    42  0.12      0.84
5    160   169      159.5      169.5 11    36  0.22      0.72
6    150   159      149.5      159.5  4    25  0.08      0.50
7    140   149      139.5      149.5  7    21  0.14      0.42
8    130   139      129.5      139.5  4    14  0.08      0.28
9    120   129      119.5      129.5  3    10  0.06      0.20
10   110   119      109.5      119.5  4     7  0.08      0.14
11   100   109       99.5      109.5  2     3  0.04      0.06
12    90    99       89.5       99.5  1     1  0.02      0.02
```

Grouped Frequency Distribution Table

- In the table, we see each interval's nominal and real limits.
- There are a number of other quantities connected with the i^{th} interval.

1. The *frequency*, f_i, i.e., the number of values that occur in the i^{th} interval.
2. The *cumulative frequency*, $cum.f_i$, the number of values that occur *at or below* the i^{th} interval.
3. The *relative frequency*, $rel.f_i$, the proportion of values that occur in the i^{th} interval. The relative frequency is obtained by dividing the frequency by n, the total number of cases. *NOTE.* The Gravetter-Walnau textbook uses the letter p to stand for relative frequencies.
4. The *cumulative relative frequency*, $cum.rel.f_i$, the proportion of values that occur *at or below* the i^{th} interval. The cumulative relative frequency is obtained by dividing the cumulative frequency by n, the total number of cases.

Grouped Frequency Distribution Table

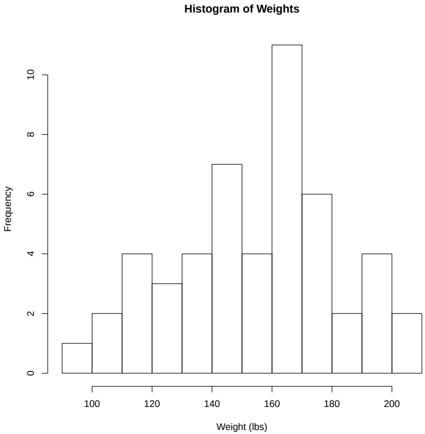
- Notice that the uppermost cumulative frequency (50 in this case) is always equal to n, the total number of observations.
- Notice also that the uppermost cumulative relative frequency is always equal to 1, and, of course, the sum of relative frequencies is equal to 1.

The Frequency Histogram

- The grouped frequency distribution table can provide some quick summary information about how values are distributed on the number line.
- The *frequency histogram* provides an accompanying visual representation.
- Here is a histogram of the weights data.

The Frequency Histogram

Matching Histogram and Frequency Table Intervals

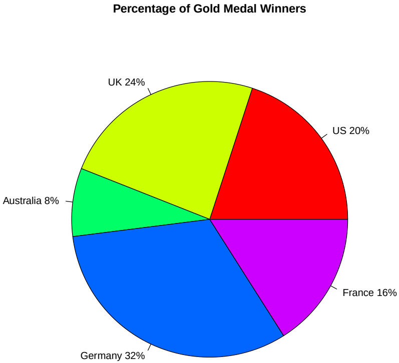


Categorical Data Plots

Bar Plots and Pie Charts

- When data are categorical, ordering the values along the X-axis might not make sense.
- One approach, discussed by Gravetter and Walnau in their Chapter 2, is a bar plot, which is very much like a histogram in appearance, except that the bars representing the frequencies of the various categories represented on the X-axis are separated by spaces.
- Another approach is the pie chart, shown on the next slide.

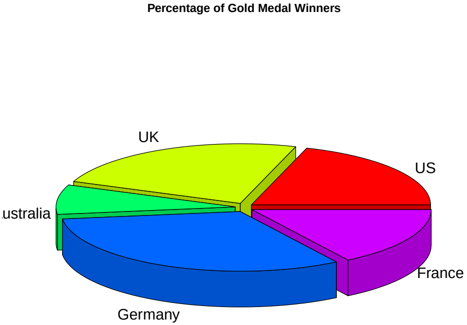
The Pie Chart



Exploded 3D Pie Charts

- Pie charts can be more dramatic when viewed in perspective.
- There is an “exploded” pie chart on the next slide.
- These are difficult to construct by hand (unless you are very artistic), but easy to create with statistical software like R.

Exploded 3D Pie Charts



Quantiles

- A *quantile* is that point in a distribution at or below which a certain proportion of cases fall.
- For example, the .25 quantile is a point at or below which 25% of the cases fall.

Quantiles

Percentiles

- The most famous kind of quantile is the *percentile*. A percentile is that point in a distribution at or below which a certain percentage of the cases fall. For example, P_{50}, the 50^{th} percentile, is that point at or below which half the cases fall.
- Since, in finite distributions, quantiles are not uniquely defined, it is common to talk about the percentile value of a given score, or the score that is at a certain percentile.
- Generally the percentile value of a given score is uniquely defined, while the score that is a t a given quantile may not be.
- Consider this concrete example. Four women have 0,1,6, and 8 children, respectively. What is the percentile value of the score 6 in this distribution?
- On the other hand, what score is at the 50^{th} percentile?