Qualitative research

lecture 12

done by: Noor Abu Gnim

Frequency distributions

Usually in every research paper we have description for the sample which depend some time on the frequency distribution for presenting our result and it can be described in term of three characteristics: the shape of the distribution of the values, central tendency, and variability. It can be presented in tabular form or graphically.

Now we will explain each with examples:

1. **The shape of distributions:**

   If we take concepts of the income and concepts of age (concept mean variable) those variable must be measured on interval or ratio level (*continuous level*) because if variable are taken on nominal or ordinal level we can't apply normal distribution principle on it.

   It can be in many shapes:

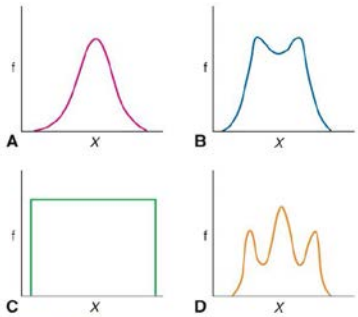

FIGURE 16.3 Examples of symmetric distributions.

❖ Symmetry

- Symmetry: the distribution of population is symmetrical if, when folded over, the two halves are superimposed on one another, there is a balance in both sides. Figure 16.3
- Skewed: asymmetric distributions, which mean that the peak is not in the center and one tail is longer than the other. Positive skew- if long tail point to the right (e.g. most people in Jordan have low to moderate incomes, with relatively few

people with high income in the tail), and negative skew- if long tail point to the left (e.g. age at death is an example of negatively skewed attribute: most people are at the upper end of distribution, with relatively few dying at an early age). Figure 16.4

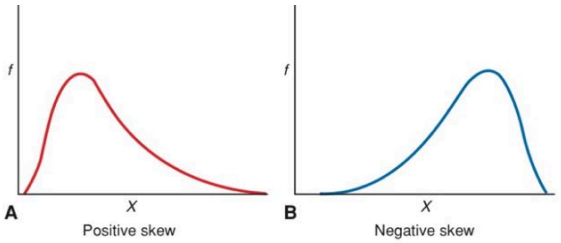

FIGURE 16.4 Examples of skewed distributions.

Note: the target always as researchers is to have a normal distribution for all our values, assumption of statics and if we have ve+ or ve- skew its mean we don't have normal distribution and a symmetrical and many assumption can't be applied on this. (Next lecture there will be farther explanation). To reach a normal distribution we should have a normal design by proper sample that presentable, larger sample better than small sample, closer to the population and reflect it. When you can discover all this? After you collect your data. (No one can reach to perfect distribution but there is an acceptable limit of error and we can do deletion or approximation for the extreme data).

❖ Modality

- A unimodal distribution has only one peak, whereas a multimodal distribution has two or more peaks. A distribution with two peaks is bimodal. Figure 16.3A is unimodal, and multimodal distribution are illustrated in Figure 16.3B and D.

Symmetry and modality are independent aspect of distribution. Skewness is unrelated to how many peaks the distribution has. So normal distribution as we mention is the symmetrical or close to it, which is unimodal, not skewed, not too peak and too flat. (too peak its mean the sample more homogenous vs. heterogeneous which too flat).

2. **Central tendency:**

Frequency distributions are a good way to organize data and clarify pattern. Often, however, a pattern is of a less interest than an overall summary. Researchers usually ask such a questions as, "what is the average body temperature of infants during bathing?" or "what is the average of weight loss of patients with cancer?" such questions seek a single number that best presents a distribution of values, and because number is more typically to come from the center of distribution than from extreme, so it's called central tendency which include: the mode, the median, and the mean.

- The mode: is the most frequently occurring score value in a distribution.
- The median: is the point in a distribution above which and below which 50% of cases fall. (The number should be sorted in a descending or ascending manner not randomly and take the middle number). Example: 2233456789

The value that exactly in the half is 4.5.

- ❖ The mean: equals the sum of all scores divided by total number of scores. For the previous example the mean is 4.9, which close to the median.
    - ✓ When to use median and when to use mode? In not normally distribute data (skewed data) we use median but if normally distributed we use mean because median not affected by extreme values or out layers, while mean affected. Because of that in the skewed data we cant go throw t-test, ANOVA, correlation and other tests so we have to go to something called non parametric.

3. **Variability:**

Two distributions with identical means could differ in variability- how spread out or dispread the data are. Figure 16.6,School A is more heterogeneous, which mean more variable than school B, and school B is more homogeneous than school A. Researchers compute an index of variability to express the extent to which scores in a distribution differ from one another. Two common indexes are the range and standard deviation.

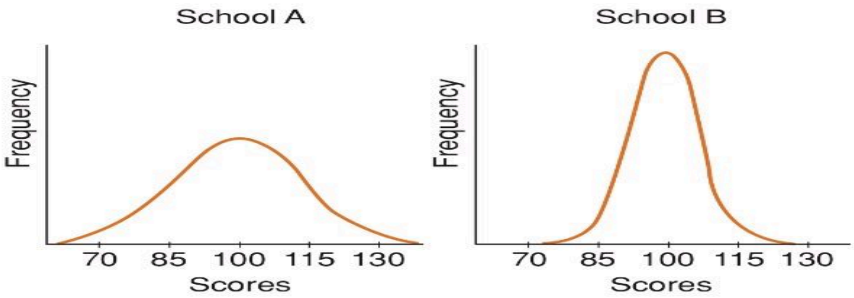

FIGURE 16.6 Two distributions of different variability.

- The range: is the highest score minus the lowest score in a distribution, e.g. the range is 5 (13-18), but in the exam if the question what is the range of for example 10 19 60 30 90, don't chose 80, which is the difference (its correct answer) but the more correct and accurate is to choose the exact range (the min and max score), 10-90, because other study may have the same difference 80 but different scores.
- Standard deviation (SD): indicates the *average amount* of deviation of values or scores in a distribution from the ***mean*** and is calculated using every score. (So always SD used with mean and we use them when data normally distributed).
    - ✓ Variance: is the value of SD squared (SD^2), and its mean how much one variable or group of variables can explained by other variables, which is the outcome (hopefully will explained in the next lectures).
    - ✓ Assume we have a normal distribution (standardized distribution) for the data; we always in a research deal with a different variables (age, gender, education…) to predict for example health outcome, we can do a combination of that variables (multivariable) but combination of different variables together is a mistake, so the program SPSS standardize the values and there will be no variation between the variables, now we have for each variable a standard deviation and one standard deviation cover one plus and one minus around mean. 1SD, cover 68% of the variables, 2SD covers 95%, 3SD cover 99%. Example on that, if we have a class grades and mean is 50, those who have grades from 40 to 60 (SD is 10, 1SD plus and minus mean) the number of student represent 68% of total sample, if

number of student are 100, 68 of them their grades between 40-60.if we take 2SD, the grades between 30 to70, the number of students in this range will be 95% of all student. Figure 16.7

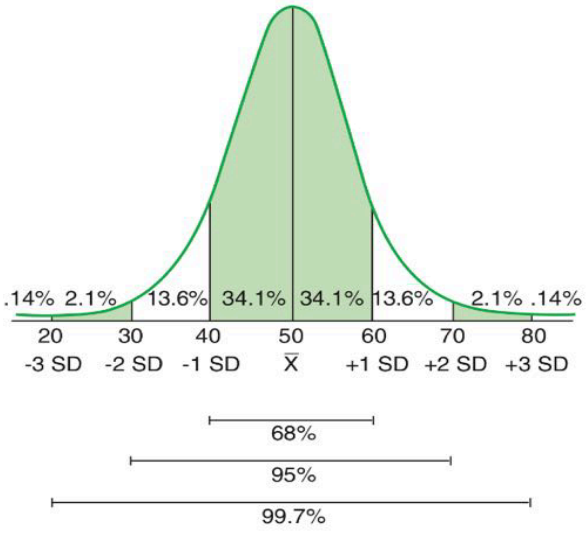

FIGURE 16.7 Standard deviations in a normal distribution.

Now we finish the frequency distributions and go to another subject.

Bivariate descriptive statistics

The mean, mode, and SD are univariate (one variable) descriptive statistics that describe one variable at a time. Most research is about relationships between variables, and bivariate (two variable) descriptive statistics describe such relationships. The commonly used method for describing two variable relationships is correlation index.

- **Correlation:** indicates direction and magnitude of relationship between two variables.

- Used with ordinal, interval, or ratio measures.
- Correlation coefficient (usually *Pearson*) summarizes information.
- Level of measurements is very important in correlation, if the two variables measured on continuous level *interval or ratio*not ordinal or others, we can apply *Pearson's correlation* equation. If one or both variables are on *ordinal* we use *Spearman's* correlation.
- With multiple variables a *correlation matrix* can be displayed.

Hypothesis testing

Researchers seek evidence through statistical analysis that their research hypothesis has a high probability of being correct. When they do statics; always they testing the hypothesis, statistical wise we are testing the null hypothesis (statistical hypothesis), which is in opposite formats or shape to the research hypothesis (confirmed with the literature). Our target is to reject the null hypothesis and to reject it the chance of error acceptable but we shouldn't exceed it, and this what we call it P- value statically, remember that we said in the previous lectures about alpha value; its defined in proposal, after collecting the real data we will get P-valueand then we compare it with alpha (5% mostly).

- A non-significant result means that any observed differences or relationship could have resulted from chance fluctuations.
- Statistical decisions are either correct or incorrect.

Errors in statistical decisions

- **Type 1 error**: rejection of a null hypothesis when it should not be rejected (false positive), its mean you get a significant result, e.g. when we get a p- value 4% and the alpha is 5%, it happens when for example we use T-test instead ANOVA test, which lead to something called inflation of alpha.
- **Type 2 error**: failure to reject null hypothesis when it should be rejected (false negative), when it happens? It occurs when we have smaller sample than the appropriate sample size, there will be no power for detecting the differences.

Forgive me for any mistake,
Your colleague: Noor A.G.

Sources:
Book: chapter 5 and 16.
Record.
Thanks for Salsabeela Bani Hamad for giving me her notes.