Simple Descriptive Statistics

Geography 450, Urban Research Elvin Wyly



Meaningless statistics... Many years ago, a New Yorker cartoon portrayed a man at home watching the evening news; the anchor reported, "Meaningless statistics were up one-point-five percent this month over last month." Numerical assertions pervade popular and policy discussion on almost any issue that matters. Too often, meaningless statistics are presented as unproblematic facts -- as objective statements of an objective, uncontested realities. Yet "Averages and relationships and trends and graphs are not always what they seem. There may be more in them than meets the eye, and there may be a good deal less." (Darrell Huff (1954). How to Lie With Statistics. New York: W.W. Norton, quote from p. 8.) Even if you think you'll never wind up doing anything even remotely related to statistical or quantitative work, you'll need to become a sophisticated, inquisitive consumer simply to survive in an information-saturated world. Edward Tufte puts it best: "Making a presentation is a moral act as well as an intellectual activity. The use of corrupt manipulations and blatant rhetorical ploys in a report or presentation -- outright lying, flagwaving, personal attacks, etc., setting up phony alternatives, misdirection, jargon-mongering, evading key issues, feigning disinterested objectivity, willful misunderstanding of other points of view -- suggests that the presenter lacks both credibility and evidence. To maintain standards of quality, relevance, and integrity for evidence, consumers of presentations should insist that presenters be held intellectually and ethically responsible for what they show and tell Thus consuming a presentation is also an intellectual and a moral activity." Edward Tufte (2006). Beautiful Evidence. Cheshire, Connecticut: Graphics Press, LLC, quote from p. 141. The cartoon was by Dana Fradon, published in *The New Yorker*, January 31, 1977. Image above: Abasaa (2010). "Statistics Bureau of Japan." Released into the public domain, via Wikimedia Commons.

"Don't confuse things that are merely countable with those that really count."¹ On the other hand, don't ignore the specificity and accountability that reliable, relevant numerical representations can offer. The purpose of simple statistical description is to provide rigorous, meaningful information about a set of observations regarding some aspect of the world. Descriptive statistics tell us about the amount of variability in this set of observations, and they provide several ways to express the "middle" or general central tendency of the observations; but descriptive statistics only provide information for the set of observations -- these techniques cannot be used to make inferences about anything other than the particular pieces of data we're working with. Consider an example: below are thirty-five numbers, each representing one response from a person filling out the Census of Canada questionnaire for their household in the spring of 2001.² People are asked many questions. The figures below represent responses to the question asking people with responsibility for the household finances, for those who rent their homes, how much they paid in total for rent, per month.

Descriptive statistics can be used to summarize the information contained in the responses offered by these thirty-five people. This can be very useful information. Of course, we might be far more interested in summarizing the rent payments of the entire Canadian population -- or, more precisely, the 3.95 million Canadian households who are renters. But to do that we have to use another approach that provides rules on how to

¹ Jonathan G. Koomey (2008). *Turning Numbers into Knowledge*. Oakland, CA: Analytics Press, quote from p. 62; Koomey attributes the statement to John Holdren, but a very similar sentiment was once also expressed by Albert Einstein.

² Data Source: Statistics Canada (2004). 2001 Census Canada, Public Use Microdata File. Ottawa: Statistics Canada.

make safe inferences from small samples that are designed to represent a larger population; this other approach is called **inferential statistics**. We'll deal with that later. Inferential statistics can be tricky and sometimes a bit difficult to interpret, in comparison to the easy-to-grasp simple descriptives. A great deal of confusion -- some by accident, some by design -- results when (flawed) inferential statistics masquerade as simple descriptive statistics.

Measures of Variation and Dispersion

But suppose we're interested in the rent payments reported by these thirty-five Canadian households in the spring of 2001. Five descriptive statistics are most commonly used to summarize the degree of variability, and to measure 'central tendency,' in a set of numbers.

1. The arithmetic **mean** or **average** is simply the sum of all the numbers, divided by the total number of observations. This is by far the most common and familiar measure of central tendency. For our dataset, the sum is 19,101, and of course we have thirty-five observations, so the mean is 19,101 divided by 35, or 545.74. As a shorthand instead of full sentences describing calculations like these, it is common to express the mean this way for any variable X:

$$\overline{X} = \frac{\sum X}{N}$$

This is simple shorthand for the mean (the X with the bar over it) is equal to the summation (the Greek letter Sigma) of all values of X, divided by N, the number of observations.

2. The **median** is an alternative measure of central tendency. The mean is sensitive to outliers or extreme cases. If the first figure in the list above were 6,500 rather than 650, then the average would jump from 545.74 to 712.88. A single extreme or bizarre observation can exert considerable influence on the mean. The median, by contrast, is not affected by outliers. The median is the mid-point of the distribution: rank the numbers from smallest to largest, and count halfway down the list. So arrange the numbers like this:

and the halfway point in the list (shown here in bold) is the median. (In cases where we have an even number of observations, then the two figures tied for the middle place are averaged). Median figures are much more stable, and less influenced by idiosyncratic events, than averages.

3. The **range** is the difference between the largest and the smallest values. In our case, the highest reported rent is \$1,300 per month, and the lowest is \$200; the range is thus 1,100.

4. The **variance** is just what it sounds like -- a measure of the degree of total variability in a set of numbers. We begin by calculating the deviation between each observation and the mean; we'll call the deviation little x to distinguish it from the values of X,

 $x = X - \overline{X}$

If we do this for our dataset, we get this:

<u>X</u>	<u>x</u>
200	-345.74
260	-285.74
263	-282.74
301	-244.74
334	-211.74
350	-195.74
400	-145.74
400	-145.74
430	-115.74
431	-114.74
450	-95.74
475	-70.74
475	-70.74
501	-44.74
510	-35.74
517	-28.74
520	-25.74
530	-15.74
550	4.26
550	4.26
569	23.26
592	46.26
597	51.26
600	54.26
603	57.26
613	67.26
650	104.26
650	104.26
652	106.26
655	109.26

663	117.26
780	234.26
805	259.26
925	379.26
1300	754.26

Now, if we wanted a total measure of variability, the logical next step would be to add up all the deviations from the mean. Unfortunately, if we do this, the values will all sum to zero. This happens every time, no matter what dataset we choose: calculating deviations from the mean and then summing the deviations adds up to zero, because the magnitudes above the mean cancel out those below the mean.

If we square the deviations, however -- multiplying each deviation value by itself -- we can get around this problem. If we then add up the squared deviations, the result will always be a non-negative number that expresses the total amount of variability (although it does not adjust for the number of observations). This is called the "sum of squares," and you'll see it mentioned, usually as an intermediate step, in many statistical approaches.

The variance is simply the sum of squares adjusted for the number of observations. Variance is usually referred to as s^2 , and it is calculated by summing the squared deviations from the mean and dividing by the number of observations.

$$s^2 = \frac{\sum \left(X - \overline{X}\right)^2}{N}$$

The variance is the average squared deviation of the values from the mean. It is a widely used measure of the variability or "spread" of a set of numbers. If it seems a bit awkward or confusing, then we can take one more step to obtain something that is much more familiar.

5. The **standard deviation** is the square root of the variance. The standard deviation is thus *the average difference of all the observations from the mean; it captures how close or far all the observations are from the average*. The standard deviation is commonly denoted by lowercase sigma, σ

$$\sigma = \sqrt{\frac{\sum (X - \overline{X})^2}{N}}$$

For the dataset above, σ is 201.92. The standard deviation is a very useful statistic. It provides a way to determine how observations cluster around the average. It also provides a consistent way to adjust variables that are measured on different scales. Suppose we wanted to compare a particular household's rent (measured in dollars, with a range as noted above of 1,300) to the age of the householder (measured in years, with a range a lot less than 1,300!). This can be done if we express values in terms of **z**-scores, also known as standard scores.

$$z = \frac{X - \overline{X}}{\sigma}$$

If a household's rent is 1.0 standard deviations above the mean -- that is, 201.92 above the mean of 545.74, or 747.66 -- then this household has a z-score of 1.0 for rent. Household age, measured on a very different scale with a wildly different range, can similarly be expressed in terms of standard deviations: if the average age were, say, 42.5 years, with a standard deviation of 11.5, then someone reporting their age as 54 would have a z-score of 1.0 for age. Z-scores are widely used to compare variables that are measured on different scales. These standard scores help to put everything on the same, comparable measurement scale.

Contingency Tables

Measures of spread and dispersion provide useful information, but they don't get us very far in interpretation. For that, we're usually interested in exploring the associations between different characteristics. Suppose we're interested in a small town in the B.C. interior that's fortunate to have enjoyed population growth and new construction in recent years. We're interested in the association between the timing of residential construction and the boundary between the town and the surrounding rural areas, so we obtain the following information:

observed frequencies				
	Rural	Urban	Total	
Built before 1970	30	20	50	
Built 1970-2000	16	28	44	
Built after 2000	8	52	60	
Total	54	100	154	

Table 1.	Contingency Table for Rural/Urban Status and	1
	Period of Residential Construction.	

Is there an association between the rural-urban boundary and the age of the housing stock? At first glance, there does seem to be a link. Thirty of 54 homes in the rural area were built before 1970, while only twenty out of 100 homes in the town were built that long ago. But how strong is this association? Could it have occurred purely by chance?

More than a century ago, a man named Pearson developed a reliable method of obtaining answers to questions like this.³ Carl Pearson was born in London of Yorkshire descent in 1857, and after graduating from Cambridge he enrolled at Heidelberg University; Heidelberg enrolled him as "Karl," when he arrived in 1879, and Pearson himself used both the "C" and "K" interchangeably until 1884, when he began using "KP" consistently. Pearson had wide-ranging interests in philosophy and mathematics, and he was soon swept up in the ferment of scientific innovation that was transforming observation, measurement, probability, and other aspects of what today would be regarded as modern statistics. In the 1890s, Pearson and many others were struggling to find ways of assessing the "goodness of fit" between distributions of things that were

³ For a fascinating account, see M. Eileen Magnello (2007). "Karl Pearson and the Origins of Modern Statistics: An Elastician Becomes a Statistician." *The Rutherford Journal* 1, 1-13.

observed, compared to what would be expected on the basis of various kinds of assumptions -- particularly, the reasonable expectation that the differences were simply the product of random chance. Consider the example of our contingency table. If there were no relationship between rural/urban status and period of construction, what would the table of frequencies look like? If there were no relationship, then we would expect, for instance, the proportion of rural homes built before 1970 to be the same as the proportion for urban homes. Since 50 out of all 154 homes in our little study area were built before 1970 (about 32.5 percent), this would mean we would expect that 32.5 percent of the 54 rural homes were built before 1970, and 32.5 percent of the 100 urban homes were also constructed before 1970. We can work these expected frequencies out (we'll call them f_e) if we multiply each row total by the column total, and divide by the grand total for the entire table. So for the first cell -- rural homes built before 1970 -- we would multiply the row total (50) by the column total (54), and then divide by the grand total of 154:

	Rural	Urban	Total
Built before 1970	fe		50
Built 1970-2000			44
Built after 2000			60
Total	54	100	154

expected frequency (fe)=(50 x 54)/154=17.5

If we do this for all the cells, we obtain:

observed frequencies			expe	ected freque	ncies		
	Rural	Urban	Total		Rural	Urban	Total
Built before 1970	30	20	50	Built before 1970	17.5	32.5	50
Built 1970-2000	16	28	44	Built 1970-2000	15.4	28.6	44
Built after 2000	8	52	60	Built after 2000	21.0	39.0	60
Total	54	100	154	Total	54	100	154

 Table 2. Contingency Table with Observed and Expected Frequencies.

The table on the right represents the breakdown we'd expect if there were no relationship between the variables in the rows and the columns. The table on the left is the breakdown we actually observed. The crucial question is: could the differences between the observed and expected outcomes result purely from chance? How much difference between the observed and expected frequencies is required before we're prepared to say that there is a significant relationship between the two variables?

Pearson's work demonstrated that it was possible to a) work out, in advance, the deviations that would occur purely by chance, b) calculate the differences between observed and expected frequencies, and c) compare the value of this difference to the deviations produced solely by chance, random factors. Pearson called this the chi-square test for goodness of fit, after the Greek letter for c.

$$\chi^2 = \sum \frac{\left(f_o - f_e\right)^2}{f_e}$$

Table 3. Calculating the Difference between Observed and Expected Frequencies.

	<u> </u>		
	Rural	Urban	Total
Built before 1970	30	20	50
Built 1970-2000	16	28	44
Built after 2000	8	52	60
Total	54	100	154
(squared differe	nces between obser	ved and exp	pected)
(sauared differe	nces between obser	ved and ext	pected)
(squared differe	nces between obser divided by expecte	ved and exp ed	pected)
(squared differe Built before 1970 Built 1970-2000	nces between obser divided by expecte 8.929 0.023	<u>ved and exp</u> ed 4.808 0.013	pected)
(squared differe Built before 1970 Built 1970-2000 Built after 2000	nces between obser divided by expecte 0.023 8.048	ved and exp d 4.808 0.013 4.333	pected)

expected frequencies				
	Rural	Urban	Total	
Built before 1970	17.5	32.5	50	
Built 1970-2000	15.4	28.6	44	
Built after 2000	21.0	39.0	60	
Total	54	100	154	

After we calculate the squared differences and divide by the expected, we add up all the results to obtain one number for the entire table -- in this case, 26.153. This is the χ^2 statistic, and it represents a summary measure of how far a particular set of numbers departs from what we would expect based purely on chance. Each distinct table of numbers will have its own unique χ^2 value, and since χ^2 is calculated by summing values for each cell, it is affected by how many rows and columns we have in a particular table: all else constant, adding more rows and/or columns will result in a higher χ^2 value. This means that to interpret any particular χ^2 value, we have to adjust for the number of rows and columns using a concept called "degrees of freedom":

df = (number of rows minus 1) x (number of columns minus 1)

For our table, df=(3-1)x(2-1)=2. 'Degrees of freedom' refers to the unique information contained in a set of numbers: once we know the values in all but the last row of a particular column, then there is no 'freedom' in deciding what value is in the last -- it's pre-determined by the difference between the total and the sum of all the other cell values.

With a χ^2 value and the corresponding *df* for a particular contingency table, we can now answer our key question: are the differences between the observed and expected frequencies large enough to have not occurred by chance? Pearson worked out the values of χ^2 for many different *dfs* that resulted from purely random, chance fluctuations, and today, nearly every introductory statistics textbook on the planet includes an appendix table for the "Critical Values of the Chi-Square Test." Typically, these tables provide thresholds for various significance levels. With *df*=2, if there is no association between two variables in any table, then there is only a 10 percent probability that the calculated χ^2 will be more than 4.60; there's only a 5 percent chance that it will exceed 5.99; there's only a 1 percent chance it will be more than 9.21; and there is only a 0.01 chance -- one time out of a thousand -- that the χ^2 value will exceed 13.8. Since our calculated χ^2 value (26.1) far exceeds any of these values, we can have considerable confidence that the differences are not simply the result of pure, random chance fluctuations. There does seem to be an association between these two characteristics.

An Example in Stata

Look for the file named "2001hh.dta" in the geog450/commute directory on the G: drive. Copy the file to

 $c:\data\pumf\2001hh.dta$

Now open STATA and issue these commands in the command panel:

set memory 200m
use "c:\data\pumf\2001hh.dta"

You should see a screen that look something like this.



The dataset you have on the screen is the 2001 Census of Canada, Public-Use Microdata File, or PUMF. This is an anonymized, random sample of the responses provided by each household to the Census Questionnaires distributed in the Spring of 2001. If you issue the

describe

command you'll quickly notice that we have 151 variables, for a total of 312,513 observations. Each observation represents the responses for the people in a particular household, and these households were carefully chosen through random sampling procedures that we'll explore next week. For now, let's simply take a look at a few simple descriptive statistics on these households.

Issue the command

summarize hhsize, detail

And you'll notice that the mean number of persons per household is 2.55, with a variance of 1.93 and a standard deviation of 1.39. The median is not labeled specifically, but you can figure it out quickly if you look at the 'percentiles' on the left -- this is simply the ranked list of all observations from smallest to largest, with various points on the distributions indicated. The median, recall, is the halfway point -- the 50th percentile. So the median number of persons per household for the PUMF sample is 2.

	Number of	persons in the	household	
	Percentiles	Smallest		
1%	1	1		
5%	1	1		
10%	1	1	Obs	312513
25%	1	1	Sum of Wgt.	312513
50%	2		Mean	2.551507
		Largest	Std. Dev.	1.39274
75%	4	8		
90%	4	8	Variance	1.939724
95%	5	8	Skewness	.8702244
99%	6	8	Kurtosis	3.397734

Now let's explore a contingency table. Issue the command

tabulate builth tenurh

and you'll obtain this.

Period of	Ten	ure	
construction	Owned (wi	Rented (f	Total
1920 or before 1921-1945 1946-1960 1961-1970 1971-1980 1981-1985 1986-1990 1991-1995 1996-2001	14,528 14,586 31,163 27,252 42,737 17,650 21,477 17,975 18,315	6,406 9,193 18,292 22,630 23,510 9,433 7,579 6,041 3,746	20,934 23,779 49,455 49,882 66,247 27,083 29,056 24,016 22,061
Total	205,683	106,830	312,513

Is there any relationship between owning and renting, on the one hand, and eras of construction of Canada's housing stock? We can test for this association if we ask for a chi-square statistic. Issue the command

tabulate builth tenurh, chi2

Period of construction	Tenu Owned (wi	ire Rented (f	Total
1920 or before 1921-1945 1946-1960 1961-1970 1971-1980 1981-1985 1986-1990 1991-1995 1996-2001	14,528 14,586 31,163 27,252 42,737 17,650 21,477 17,975 18,315	6,406 9,193 18,292 22,630 9,433 7,579 6,041 3,746	20,934 23,779 49,455 49,882 66,247 27,083 29,056 24,016 22,061
Total	205,683	106,830	312,513
Pearso	on chi2(8) =	8.0e+03	Pr = 0.000

Notice that the χ^2 value is huge -- so large that STATA expresses it in scientific notation (8.0 times ten raised to the power of 3, or about 8,000). The corresponding probability of obtaining a figure this large solely from random chance fluctuations is extremely low -- rounded off to 0.000.

The χ^2 test is extremely versatile, and it is widely used, but it is also affected by sample size.

Think carefully about that last word, though -- *sample*. Our discussion today focused on simple descriptive statistics -- numerical summaries of a set of numbers, without any inference to the broader "population" that those numbers purport to represent. But of course these inferences are crucial, and so we will need to move from the realm of descriptive statistics to inferential techniques. Let's begin by considering the effect of sampling on our analysis of Canadian households. Note the grand total in the output above -- 312,513. This is the number of households who were sampled and interviewed to obtain responses to the many questions on the Census questionnaire. Statistics Canada also provides information allowing us to figure out how many households in the population are represented by each sampled household. Issue the command

Period of	Ten	ure	
construction	Owned (wi +	Rented (f	Total
1920 or before 1921-1945 1946-1960 1961-1970 1971-1980 1981-1985 1986-1990 1991-1995 1996-2001	537,536 539,682 1,153,031 1,008,324 1,581,269 653,050 794,649 665,075 677,655	237,022 340,141 676,804 837,310 869,870 349,021 280,423 223,517 138,602	774,558 879,823 1,829,835 1,845,634 2,451,139 1,002,071 1,075,072 888,592 816,257
Total	+	3,952,710	+

tabulate builth tenurh [fweight=weighth]

Now that's more like it. The PUMF sample -- 312 thousand sampled households -- corresponds to a total population of about 11.5 million households. The "fweight" option

asks STATA to tabulate the variables of interest, "weighting" each sampled record according to the value of "weighth" -- which is the number assigned by Statistics Canada to each household, representing the number of population households represented by each sample observation. For the largest social surveys that provide large samples chosen according to strict rules of random selection, the weight value will typically be the same for each observation (here it is 37 for all households). Smaller surveys, with fewer sampled observations, are much more likely to "under-sample" some kinds of phenomenon, and to "over-sample" others; those who design these smaller surveys, then, must assign differential weights to various observations to compensate for this sampling variability.

One final note of caution. Issue the command

tabulate builth tenurh [fweight=weighth], chi2

Period of | Tenure construction | Owned (wi Rented (f | Total 1920 or before 537,536 237,022 774,558 1921-1945 539,682 340,141 879,823 1946-1960 1,153,031 676,804 1,829,835 1,845,634 1961-1970 1,008,324 837,310 1971-1980 1,581,269 869,870 2,451,139 1981-1985 349,021 653,050 1,002,071 280,423 1986-1990 794,649 1,075,072 1991-1995 888,592 665,075 223,517 1996-2001 677,655 138,602 816,257 Total | 7,610,271 3,952,710 |11,562,981 Pearson chi2(8) = 2.9e+05Pr = 0.000

Now compare the χ^2 value above -- 2.9 x 10⁵ -- to what we obtained with exactly the same variables using the unweighted observations earlier. The proportions in various parts of the table are exactly the same; we've simply multiplied everything by the household weight (37), inflating the χ^2 statistic, and greatly increasing the likelihood that we will detect strong associations even when they may not really exist. This is why it's so important to think through the *meaning* of statistical tests and procedures before you go anywhere near the computer. As computer processing speeds accelerate, it's possible to do more and more stuff, faster and faster. This also means it's possible to make massive, stupid mistakes faster than ever before. And thus meaningless statistics may have increased by much more than one point five percent this month over last month!