

GGR 270Y First Term Lecture Notes

Fall 1996

By Harold Reynolds

Table of Contents

<u>I. Introductory Remarks.....</u>	<u>4</u>	<u>E. Types of Errors.....</u>	<u>9</u>
<u>A. Why study statistics?.....</u>	<u>4</u>	<u>F. Sources of Data.....</u>	<u>9</u>
<u>Daily exposure to statistics:.....</u>	<u>4</u>	<u>G. Populations & Samples.....</u>	<u>9</u>
<u>Summarizing information:.....</u>	<u>4</u>	<u>H. Desirable Characteristics of Sample Statistics.....</u>	<u>10</u>
<u>Polls:.....</u>	<u>4</u>	<u>III. SUMMARIZATION TECHNIQUES.....</u>	<u>10</u>
<u>Tools:.....</u>	<u>4</u>	<u>A. Ternary Diagrams.....</u>	<u>10</u>
<u>Describe and quantify error:.....</u>	<u>4</u>	<u>B. Frequency Distributions.....</u>	<u>10</u>
<u>Deduce and infer properties:.....</u>	<u>4</u>	<u>C. Population, Sampling, and Sample Sampling</u>	
<u>Analytic reasoning of experiment:.....</u>	<u>4</u>	<u>Distributions.....</u>	<u>11</u>
<u>B. What's to be afraid of?.....</u>	<u>4</u>	<u>Sampling distributions related to desirable statistic</u>	
<u>Statistics is an arcane art?.....</u>	<u>4</u>	<u>qualities.....</u>	<u>11</u>
<u>Source of fear:.....</u>	<u>4</u>	<u>Example of the creation of a sampling distribution.....</u>	<u>11</u>
<u>The math isn't hard!.....</u>	<u>4</u>	<u>The Sampling Distribution.....</u>	<u>12</u>
<u>I don't care! I can't do math anyways!.....</u>	<u>4</u>	<u>D. The Histogram.....</u>	<u>12</u>
<u>Hardest part is volume:.....</u>	<u>4</u>	<u>i) Guidelines For Grouping.....</u>	<u>12</u>
<u>Organization is the key to success.....</u>	<u>5</u>	<u>ii) Basic Procedure for Making a Histogram.....</u>	<u>12</u>
<u>Need help?.....</u>	<u>5</u>	<u>E. Simple Descriptive Summary Measures.....</u>	<u>13</u>
<u>A word of warning about group work on</u>		<u>IV. SIMPLE UNIVARIATE DESCRIPTIVE</u>	
<u>assignments:.....</u>	<u>5</u>	<u>STATISTICS.....</u>	<u>14</u>
<u>Finally, there is nothing to fear.....</u>	<u>5</u>	<u>A. Measures of Central Tendency.....</u>	<u>14</u>
<u>C. Some Mathematical Symbols.....</u>	<u>5</u>	<u>i) The Mean (Interval and Ratio Scales).....</u>	<u>14</u>
<u>i) The Summation operator.....</u>	<u>5</u>	<u>ii) The Median (Ordinal scale).....</u>	<u>15</u>
<u>Why is it used?.....</u>	<u>5</u>	<u>iii) Mode (Nominal Scale).....</u>	<u>15</u>
<u>What's this subscript thing, anyways?.....</u>	<u>5</u>	<u>iv) Which is better?.....</u>	<u>15</u>
<u>Clarify that, please!.....</u>	<u>5</u>	<u>B. Measures of dispersion.....</u>	<u>16</u>
<u>A DOUBLE sum? You mean it gets worse??.....</u>	<u>6</u>	<u>i) Range.....</u>	<u>16</u>
<u>Tree Size.....</u>	<u>6</u>	<u>ii) Interquartile Range.....</u>	<u>16</u>
<u>ii) The Product Operator.....</u>	<u>6</u>	<u>iii) Mean Deviation.....</u>	<u>16</u>
<u>iii) Factorial Operator.....</u>	<u>6</u>	<u>iv) Variance and Standard Deviation.....</u>	<u>16</u>
<u>D. Statistical Populations: Definitions.....</u>	<u>6</u>	<u>vi) The Coefficient Of Variation.....</u>	<u>17</u>
<u>II. DATA AND MEASUREMENT.....</u>	<u>7</u>	<u>vii) Skewness.....</u>	<u>17</u>
<u>A. Introduction.....</u>	<u>7</u>	<u>viii) Kurtosis.....</u>	<u>17</u>
<u>Deductive reasoning.....</u>	<u>7</u>	<u>C. The Lorenz Curve and Gini Coefficient.....</u>	<u>18</u>
<u>Inductive reasoning.....</u>	<u>7</u>	<u>i) The Lorenz Curve.....</u>	<u>18</u>
<u>B. Data.....</u>	<u>7</u>	<u>ii) Gini Coefficient.....</u>	<u>18</u>
<u>C. Scales of Measurement.....</u>	<u>8</u>	<u>V. CORRELATION.....</u>	<u>18</u>
<u>i) Nominal Scale ("having to do with names").....</u>	<u>8</u>	<u>A. Phi Coefficient (Nominal Scale Data).....</u>	<u>19</u>
<u>ii) Ordinal ("put into order").....</u>	<u>8</u>	<u>B. Chi-Squared Statistic (Nominal Scale Data).....</u>	<u>19</u>
<u>iii) Interval Scale (Meaningful unit of distance).....</u>	<u>8</u>	<u>C. Spearman Rank Correlation.....</u>	<u>21</u>
<u>iv) Ratio Scale (Meaningful zero).....</u>	<u>8</u>	<u>D. Scattergrams.....</u>	<u>21</u>
<u>D. Issues in Data Collection.....</u>	<u>8</u>		

GGR 270Y First Term Lecture Notes

E. Covariance.....	22	i) Example 1: The False Positive Paradox.....	35
F. Pearson's Correlation Coefficient.....	22	ii) Example 2: Urn and Ball problems.....	35
Comparing Pearson's with Spearman's for Ratio-scale data.....	22		
G. Interpreting Results.....	23		
VI. SIMPLE LINEAR REGRESSION.....	23		
VII. SPATIAL STATISTICS.....	24	IX. Some Theoretical Frequency Distributions.....	36
A. Geostatistics.....	26	Some Terms.....	36
i) Spatial Mean.....	26	A. The Uniform Distribution.....	37
ii) Standard Distance.....	26	B. The Binomial Distribution.....	37
iii) Areal Frequency Distribution.....	27	C. The Poisson Distribution.....	38
B. Point Pattern Analysis.....	27	D. The Normal Distribution.....	39
i) Nearest Neighbour Statistic.....	27	i) The Standard Normal Distribution.....	39
ii) Quadrat Analysis.....	27	ii) Using the table.....	39
C. Spatial Autocorrelation.....	28	iii) More advanced examples.....	39
i) Binary Connectivity Table or Matrix.....	28	E. Doing problems.....	40
ii) Join Count Statistic.....	28	F. How to Bell Grades.....	40
iii) Moran Coefficient.....	28		
iv) Geary Ratio.....	29	X. SAMPLING DESIGNS.....	40
v) Example of Moran Coefficient & Geary Ratio.....	29	A. Geographic Populations.....	41
How to Compute the Moran Coefficient and Geary Ratio.....	30	B. Sample Size and Sampling Error.....	41
VIII. PROBABILITY.....	30	C. Random Sampling.....	42
A. Counting Tools.....	31	D. Representative Samples and Sampling Frameworks.....	42
i) The Multiplicative Rule.....	31	i) Systematic Selection.....	42
ii) The Permutations Rule.....	31	ii) Stratification.....	42
iii) The Partitioning Rule.....	31	iii) Cluster Sampling.....	43
iv) The Combinations Rule.....	31	E. Geographic Sampling.....	43
v) Sampling with Replacement.....	32	F. Some other problems.....	44
B. Introduction to Set Theory.....	32		
i) Terminology.....	32	XI. Concepts for Inferential Statistics.....	44
ii) Example: Throwing 2 Dice.....	32	A. Estimation: Large-Sample Estimate of a Population Mean.....	44
iii) Application to Probability.....	33	B. Determining a Minimum Sample Size.....	46
iv) Mathematical Set Rules.....	33	C. Small-Sample Estimation of the Mean.....	46
v) Probability Rules using Sets.....	33	D. Introduction to Hypothesis Testing.....	47
C. Conditional Probability.....	33	i) Test Statistics.....	47
D. Example using Contingency Table.....	34	ii) Types of Tests.....	47
E. Bayes' Theorem.....	34	iii) Errors in Hypothesis Testing.....	47
		Type I Error.....	47
		Type II Error.....	48
		iv) The Power of a Test.....	48

I. Introductory Remarks

A. Why study statistics?

Contrary to popular belief, it is not to drive students crazy! The proximity of the Sidney Smith Building to the Clarke Institute (for Psychiatry) is only a coincidence. Really.

Daily exposure to statistics:

Whether you realize it or not, you are bombarded with statistics every day. Polls, average temperature for this time of year, crime rates, average income, traffic counts on city streets, all are statistics. Sports (when the players aren't on strike) are in my humble opinion one of the largest generators of useless statistics there is. (Governments are the biggest.)

Summarizing information:

Large amounts of information can be condensed into a few simple figures and/or statements using statistics. You can't tell someone the high temperature of today's date for the past 30 years, but you can say "the average temperature is..."

Polls:

Polls are probably the most familiar examples of the use of statistics. Since it is impractical to ask every Tom, Dick and Harriet his/her opinion on some issue, a certain number of Toms, Dicks and Harriets are chosen at random, and the polling company would have us believe their opinions reflect those of everyone.

Tools:

Statistical techniques are tools used to organize information and interpret observations. That's all they are!

Describe and quantify error:

Statistical analysis allows one to describe and quantify sources of uncertainty and/or error in experimental data. With an idea of the uncertainty in mind, one can assess the usefulness of the data. Obviously, a politician would like to know just how accurate a poll is, especially if his/her popularity is hovering around the area where s/he may not get re-elected.

Deduce and infer properties:

Statistics allow one to deduce or infer the properties of a population, based on the information we can derive from a sample of the population. As mentioned above, we can't always ask **everyone** their opinions.

Analytic reasoning of experiment:

Statistics forces you to do some analytic reasoning of the experiment you're considering while it's being planned. If there is some sort of problem with the experiment itself and how the data are collected, *no amount of analysis will give you worthwhile results!*

B. What's to be afraid of?

Statistics is an arcane art?

The majority of people view statistics as an arcane art practised by people with thick glasses and fancy calculators, who sit behind computers all day and have no social life. This is not always true. (Just kidding, it's not true at all!)

Source of fear:

Of all the courses students have to take, statistics is the one dreaded the most, because people perceive it as "really difficult". "Difficult" is such an unpleasant word--I prefer the term "*conceptually challenging*".

The math isn't hard!

The actual math involved is not terribly difficult. The most challenging mathematical concept is the exponent.

I don't care! I can't do math anyways!

Rubbish! In order to catch a ball, for example, your brain must solve an initial value differential equation problem based only on initial speed and direction in order to place your hands in the correct location. Furthermore, it must do so based on *estimates* of the distance of the thrower, the speed and trajectory of the ball!

Hardest part is volume:

Probably the hardest part about any stats course is the volume of material that has to be covered just to give the student the *basic* ideas. Just remember that what you will be seeing is only the tip of the iceberg. Instead of moaning about how much work there is (a valid complaint, of course), just remember that you're really getting off lightly!

Organization is the key to success.

As with any high-volume course, it is vitally important to keep your notes organized in some way. Bearing in mind that the exams are likely (but not necessarily--check with the instructor!) to be open book, or at least with a cheat sheet, maintaining an index where the important definitions and formulas are located can be a tremendous help! Re-reading your notes from last class and keeping up in the text are also strongly recommended.

Need help?

Forming a study group with some other classmates is extremely helpful, especially for doing the pesky assignments. Also, don't be afraid to ask questions in your tutorial! Asking a question means that you have the guts to ask what other people are also wanting to know about, not that you're a dummy! Stats concepts aren't always easy to grasp, and if you don't ask, you'll draw a blank when it comes to exam time! The TA's, overworked as they are, are here to help and will try their best.

A word of warning about group work on assignments:

ASSIGNMENTS MUST ULTIMATELY BE WRITTEN BY YOU ALONE! This means that although we expect (and even encourage) a certain amount of collaboration when doing the questions, YOU MUST WRITE THE ANSWERS TO EACH QUESTION ON YOUR ASSIGNMENT IN YOUR OWN WORDS! Copying someone's answers word for word is an academic offence and will result in a zero for all concerned. Besides, just copying doesn't mean you understand it! Exam questions are designed to test your understanding of the topic, and that can't be put in cheat sheets or open books! (Sorry if this sounds heavy-handed, but we want to spare you and us grief later!)

Finally, there is nothing to fear.

Numbers don't bite! (Only grumpy TA's do, and they rent Dobermans to do that for them!) (OK, just kidding...)

C. Some Mathematical Symbols

The purpose of this section is to introduce you to some of the mathematical symbols used in statistics. They are nothing more than shorthand notation, and are used to save lots of tedious writing. They may look strange to you now, but soon they'll be second nature. Honest.

i) The Summation operator

The symbol that most people associate with stats more than any other is Σ , the **summation operator**. This is the Greek capital letter *sigma*.

Why is it used?

Mathematicians are just as lazy as the rest of us. Writing $x_1 + x_2 + x_3 + \dots + x_{n-1} + x_n$ is tedious, and occurs frequently in statistics. To reduce the tedium (to a more manageable level) they defined $\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \dots + x_{n-1} + x_n$.

English translation: "The sum of x sub i , for all values of i from 1 to n ". The $i=1$ below the Σ and the n above it denote the **range** of the *index* I . I is the most commonly used index in summations by mathematical convention.

What's this subscript thing, anyways?

In regular algebra, one normally only deals with a few variables at a time, so it is usually convenient to call them a , b , x , y , or whatever. In statistics, however, most if not all the formulas used involve adding together an *indefinite* number of numbers; the formulas are indefinite because they can be used for almost **any** number of numbers. In this case, it is much simpler to use a (capital) letter, like X or Y , to represent a variable, and to use its lower case equivalent with subscripted numbers or letters representing numbers attached to it to represent specific *instances* of the variable.

Clarify that, please!

Suppose a geographer studying traffic patterns wanted to count how many cars passed through the intersection of Bloor and St. George streets. Obviously, this will depend on the traffic lights, so she decides to count cars passing through in a minute for several minutes and take the average. She represents the *variable* "number of cars" by the letter X . Suppose in the first minute 50 cars were counted. She records 50 beside x_1 . In the next minute, 43 go through, so she records 43 beside x_2 . For the following three minutes, she counts 51, 40, and 47 for the third, fourth, and fifth *instances* of X . To get the *average*, she sums the values and divides by 5.

If you think about it as adding a whole column of numbers, it will help when you look at the double sum.

A DOUBLE sum? You mean it gets worse??

A double sum is no different in concept than a regular sum! If you look upon a single sum as the sum of a *column* (a one-dimensional entity) of numbers, it is easy to look at a double sum as the sum of a *table* (a two-dimensional entity) of numbers. Suppose a forester is out in the bush counting trees, and each tree is also put in a category of small, medium or large. He could construct a table like this:

Tree Size				
Species	Small	Medium	Large	Species Total
Spruce	$x_{11} = 4$	$x_{12} = 8$	$x_{13} = 6$	$\sum_{j=1}^3 x_{1j} = 18$
Larch	$x_{21} = 10$	$x_{22} = 12$	$x_{23} = 2$	$\sum_{j=1}^3 x_{2j} = 24$
Poplar	$x_{31} = 18$	$x_{32} = 16$	$x_{33} = 10$	$\sum_{j=1}^3 x_{3j} = 44$
Aspen	$x_{41} = 5$	$x_{42} = 4$	$x_{43} = 12$	$\sum_{j=1}^3 x_{4j} = 21$
Size Total	$\sum_{i=1}^4 x_{i1} = 37$	$\sum_{i=1}^4 x_{i2} = 40$	$\sum_{i=1}^4 x_{i3} = 30$	$\sum_{i=1}^4 \sum_{j=1}^3 x_{ij} = 107$

As the table shows, each number fits into **two** categories, species and size. In order to find out how many spruce trees he counted, he must sum *across the spruce row*, that is summing using the **j subscript**. To find out how many large trees he counted, he must sum the *“large” column*, that is, summing using the **i subscript**.

To find out how many trees were counted, all the numbers in the table must be added up, by summing over BOTH i and j. You get the same results by summing the numbers of the different tree species (summing the *column sums*) or by summing the numbers of the different sizes (summing the *row sums*). That's all there is to it!

All you have to remember is that the i subscript represents the **row**, and the j subscript represents the **column** that the number occupies in the table.

ii) The Product Operator

The product operator is another shorthand notation, this time representing *multiplication* of variables instead of addition.

It is defined as $\prod_{i=1}^n x_i = x_1 \cdot x_2 \cdot x_3 \cdot \dots \cdot x_{n-1} \cdot x_n$. It is not used very frequently in statistics.

The rules for the summation and product operators are presented in the “Green Book (Griffith and Amrhein), pp. 11-15, and the “Black Book” (Burt and Barber) pp. 69-70.

iii) Factorial Operator

The factorial operator is ! (the exclamation mark). This isn't a way to add excitement to statistics, it's just another shorthand notation. $n! = n(n-1)(n-2)\dots(3)(2)(1)$. This will come up later when probability, combinations and permutations are discussed in more detail.

D. Statistical Populations: Definitions

This section consists of a few definitions relating to statistical populations. **Memorize these terms**, as they are fundamental to statistics!

Population: A population is a complete set of things, such as students in this classroom, soils, stores, road accidents, grades, ethnic groups, or voters. Statistics is the science and/or art of estimating information about an entire population based on only a small part of it, the sample (see below).

Variables: The properties of a *population* which can be measured are called *variables*. Examples: height, weight, shoe size, soil composition, damage done in an accident, and ethnic group distribution. The word "variable" is derived from the fact that the property can and will *vary among each element of the population*. Variable values are usually unknown and must be determined empirically by estimation.

Sample: A *sample* is a **subset** (i.e. a selection) of the members of a population. The students in the front row are a sample of the students in the classroom, who in turn are a sample of the students at the University of Toronto.

Representative: A sample is *representative* if it accurately reflects the population. For instance, if the class is roughly half male, the front row sample should be roughly half male if the sample is to be representative. The chief problem in collecting data is ensuring that it is representative.

Constant: A *constant* is a quantity that does not change. (duh!) It may be universal or set by theory, such as pi, e, G. It also may be set for a specific problem, but vary for different problems, such as speed limit, interest rate, g, or distance between places.

Parameter: A *parameter* is a constant measure which describes a characteristic of a **population**. This could be the average height or weight of students in the classroom, the number of blue cars on Bloor Street, or the average IQ of people who watch Married With Children.

Statistic: A *statistic* is the corresponding measure for a **sample**. It is an *estimate* of a population parameter. Different symbols are used to represent parameters and statistics.

	Mean	Variance	Standard Deviation	Correlation
Parameter (population)	μ	σ^2	σ	ρ
Statistic (sample)	\bar{x}	s^2	s	r

II. DATA AND MEASUREMENT

The purpose of this section is to list a lot of definitions that are commonly used in statistics, and certainly will come up frequently in this course. Boring as it may be, you still should learn what all this jargon means!

A. Introduction

Statistical methods are used to derive conclusions based on empirical data. Empirical data are measurements derived either from observations or experiments. Whenever you look at the thermometer outside, or step on the scale to see how much weight you've gained over Christmas, you have just recorded empirical data.

Deductive reasoning

Mathematics employs *deductive* reasoning, in which the conclusions follow logically from the preceding arguments and premises. For example: 1) Canada is larger than France. 2) France is larger than Bulgaria. 3) Therefore Canada is larger than Bulgaria.

Inductive reasoning

Statistics is *inductive*; that is, the conclusions exceed the contents of the preceding premises. For example, a poll of adults in a suburb reveals that 20% of those polled commute to work. From this, and assuming that the poll is "reliable" in a statistical sense (we learn about quantifying this later!) we could *infer* that 20% of **all** the adults in the suburb commute.

Given more precise limits on the inference (as explained later), one can use *inferential statistics*, which is the focus of much of this course.

B. Data

Data (plural of datum) are generated by the recording of measurements. Measurement involves either categorizing events (*qualitative*) or using numbers to give some sort of size to the event (*quantitative*). The different attributes of the thing of interest which one selects to measure are called **variables**, because individual values are expected to be different (for example, people's shoe size). Individual measurements of a variable are called **observations** (e.g. John Bigfoot wears a size 12).

C. Scales of Measurement

i) Nominal Scale ("having to do with names")

Observations are put into categories based on some criterion, such as rock type, social group / occupation, housing type.

Numbers are labels only!

Any number used in a nominal scale is a category label only. No mathematical operation can be performed on it because its assignment to the category is arbitrary. For example, if you were taking part in a computer survey which asked what colour your hair is, you might see a list like: 1. Brown 2. Black 3. Blonde 4. Red The numbers are for convenience only!

A **binary** or **dichotomous** variable can have 2 values--yes/no, male/female, dumb/dumber.

A **multichotomous** variable can have more than 2 values, such as ethnic background, street type, occupation.

No attempt is made to apply a scale to the categories.

ii) Ordinal ("put into order")

Observations are put into categories which can be **ranked** in some order. E.g. wealthy, middle-class, poor neighbourhoods; wallet-buster, expensive, moderate, cheap restaurants, hardness of minerals.

No precise value can be assigned to a difference between ranks. (When does "wealthy" become "middle-class"?)

Numbers as category names become more meaningful. For instance, diamond has a hardness on the Moh's scale of 10, while quartz has a hardness of 7, which means that diamond is harder than quartz and can scratch it.

Weak Ordering: Assign items to a category & rank categories.

Strong Ordering: Rank every observation in each category.

Assignments must be **mutually exclusive** (no overlapping categories) and **collectively exhaustive** (no gaps in the coverage by the categories). That is, one must be able to assign all the items to the various categories such that *each item fits into only one category.*

iii) Interval Scale (Meaningful unit of distance)

Numerical category names (which can be real as well as integers) now have a meaningful unit of distance separating them. We can now say that not only is $A > B$, but by how much, *but only in terms of addition and subtraction.*

Commonly cited examples are the Celsius and Fahrenheit temperature scales, IQ score, and elevation with respect to a local reference point.

iv) Ratio Scale (Meaningful zero)

In this scale there is a meaningful (absolute) zero, which allows division and multiplication to be used. The zeros in the Celsius and Fahrenheit scales are arbitrary, so you can't "really" say 20 C is twice as hot as 10 C.

The Kelvin temperature scale, weight, length, area, amount of milk produced, and value of crops are all ratio scales.

After all this, statistical analyses tend not to distinguish between interval and ratio scales. However, profs and TA's who ask trick questions on tests and assignments do, so think carefully before answering scale-related questions!

Page 41 of Griffith and Amrhein has examples of the different scales.

D. Issues in Data Collection

These three terms often get confused. Be sure to learn the difference!

Precision: How *exact* is the measurement instrument, i.e. how close to reality does it measure, or how many significant figures can it produce? The closer/more figures the better! For example, if someone tries to pass off a yardstick as a metrestick, some poor yokel's metric distances are going to be shorter than they should be!

Accuracy: Is the measurement free from a systematic bias or error? For instance, an upset dieter might turn back the bathroom scale by 5 pounds, thus giving subsequent measures that are consistently off. Note that this is not the same as precision! The number of significant figures has no relation to systematic bias.

Validity: Is an instrument measuring what it claims to be measuring? This becomes an issue when dealing with a survey or polls (yes, a survey can be considered an instrument!). Surrogate data and indices, i.e. data which have been collected are substituted in some way for data which have not, are often of questionable validity. An example of surrogate data is the use of tree rings to measure past weather, or the thickness of layers of clay on lake bottoms to measure local erosion effects.

E. Types of Errors

Again, these terms are often confused with one another, so be careful!

Calculation error: A goof is made in an arithmetic operation ($2+2=5$), or somehow numbers get miscopied in assignments (6435 written as 6345), etc. Probably one of the most maddening errors!

Measurement error: Incorrect numbers are present in the data set. Can be from undercounting (as in traffic counts), incorrect coding, use of surrogate variables, etc.

Specification error: Incorrect assumptions or formulae are used. For example, a family defined as mother, father and kids, ignoring single-parent or gay families; using $y=ax$ instead of $y=ax+b$.

Sampling error: The difference between a *sample* and the *population* from which it is drawn. This is due entirely to the variation between each sample. This is the fundamental problem in statistics. Example: a population has an average income of $\mu = \$15,000$, but a sample of 1000 has an average income of $\bar{x} = \$14,750$.

Random noise in the landscape: This is a type of sampling error that results from a number of factors which act together to produce apparently random disturbances into the data set. Meteorological data is especially prone to this sort of error. Consider a plot of UV intensity throughout a day. Under ideal circumstances, it should be a smooth curve, peaking in intensity at about 1:30 pm or so, but you very seldom get that. Some of the dips in intensity could be caused by passing clouds, or perhaps some low-level UV-absorbing pollutants passed by, or maybe some yutz sat beside it and ate his lunch!

F. Sources of Data

These definitions are pretty straightforward.

Primary data: Collected directly by the researcher through experiments, surveys, field work, etc.

Benefits: The researcher controls which data are measured, and will have a good idea of the sources of error associated with it.

Drawback: Can be expensive.

Secondary data: Data acquired from some other source, e.g. census data.

Benefits: Can be much cheaper and reduces chance of redundancy (i.e. that someone else has already collected it).

Drawbacks: One cannot be entirely sure of data quality; the specific variable of interest to researcher may not have been collected!

Tertiary data: Third-party publications of secondary data, such as an analysis of census data published in a newspaper.

Benefit: Analysis already done for you.

Drawback: You don't know how accurate the analysis is!

Direct: From direct observation.

Indirect: From a combination of direct sources. It's less accurate than direct sources. Possible sources of error include different levels of precision in measures and errors in calculation.

G. Populations & Samples

A **population**, as stated before, is the set of all possible observations. It must be defined clearly for inferential and sampling purposes. Unknown properties must be assumed.

A **sample** is a specifically derived subset of the population. In most cases it is too difficult/expensive/time consuming to examine an entire population so a portion of it is selected instead.

A sample should be **representative** in that it should be drawn from a good cross-section of the population. The catch-22 is that without enumerating the entire population, we can't be sure of the sample's representativeness!

Drawing **random** samples is probably the best way to ensure representative ones. Sampling error associated with random samples is well understood theoretically, so much so that probabilities can be attached to any resulting statistics.

Most random samples also produce statistics with **minimal deviation** from the population attribute in question. This means that the likelihood of the statistic being representative is much better than for other methods.

H. Desirable Characteristics of Sample Statistics

Unbiased: If the arithmetic mean of the statistic calculated for all possible samples of a given size n exactly equals its population parameter.

Sufficient: Summarizes all relevant information about the parent population contained in the sample, while ignoring any sample-specific information.

Efficient: The more the statistic values for various samples cluster around the true parameter value, the lower the sampling error and the greater the efficiency. Consider an archer shooting at a target. The archer wants to be accurate, but also wants the arrows to cluster as closely to the centre of the target as possible.

Consistent: The larger the sample size, the closer the statistic should be to its parameter value.

III. SUMMARIZATION TECHNIQUES

The quickest and most useful way to get some basic information about a data set is to display it graphically in some way. Probably the most familiar ways are bar charts (nominal and ordinal scales) and histograms (interval and ratio scales).

A. Ternary Diagrams

These diagrams are used to display data which has been divided into 3 (hence the name *ternary*) mutually exclusive, collectively exhaustive categories.

A ternary diagram is an equilateral triangle, on each side of which is a proportional or percent scale corresponding to one of the categories. Every point has 3 coordinates which **sum** to 1 (proportion) or 100 (percent). They are used primarily for soil analysis. See Griffith and Amrhein (p. 54) for more details if you really care. You'll probably never see one again unless you want to study physical geography, so don't worry about this too much!

B. Frequency Distributions

A **frequency distribution** is the number of times a random variable takes each of its possible values. The classic example is for the sum of two dice. The **random variable** in this case is the result you get when you throw the dice, and the possible values are the integers from 2 to 12. (The numbers 2-12 are distributed 1 2 3 4 5 6 5 4 3 2 1). **Every random variable has a frequency distribution.** The time you have to wait to get on an elevator or the subway, the shoe sizes of students in the class, raindrop sizes, and even the number of times Curly gets smacked by Moe in a Three Stooges film all have frequency distributions.

A **relative frequency distribution** is formed from a frequency distribution by taking the frequency of each observed value of the variable and dividing by the total number of observations. It is especially useful when there are a large number of observations which may make the vertical scale awkward.

A **cumulative frequency distribution** is formed by starting with the lowest value of the observed variable and its associated frequency. For each successive value of the variable, add its frequency to the total of the frequencies of all the previous values.

A **cumulative relative frequency distribution** is formed by the same process as above, except for the relative frequencies.

Die Roll →	2	3	4	5	6	7	8	9	10	11	12
Frequency	1	2	3	4	5	6	5	4	3	2	1
Cumulative	1	3	6	10	15	21	26	30	33	35	36
Relative	.028	.056	.083	.111	.139	.167	.139	.111	.083	.056	.028
Cumul. Rel.	.028	.083	.167	.278	.417	.583	.722	.833	.917	.972	1.00

An example of all four types of distribution is in the table above, illustrating the results of throwing two dice and adding the numbers.

Creating this distribution is easy, since all you have to do is create a category for each die roll. Most of the time, however, one has to create categories based on **ranges** of values. For more detail, see below.

C. Population, Sampling, and Sample Sampling Distributions

Suppose that I extract 15 rocks from my head and I weigh and label them and put them in a bag. Because we expect each rock to be different, each will have a different weight and we can construct a frequency distribution of the weights. We can define these rocks as a population in the statistical sense, so the frequency distribution of the weights of all 15 rocks can be called a **population (frequency) distribution**.

A **sampling distribution** is simply the frequency distribution of the values of a statistic which is computed for **all possible** samples of size *m* that are drawn from a population of size *n*. Note the three things which must be specified when referring to a sampling distribution: sample size, population size, and the statistic itself.

Suppose I now pick out **samples** at random of (*m*)6 rocks at a time, not caring in what order I select them, from the population of (*n*)15. The idea is that I want to estimate the mean weight of *all*15 of the rocks by finding the mean of a sample of them. It will be shown later that there are a total of 5005 *different* samples I can have, and for each I can compute a mean weight. We expect each of these to be different, hence I can construct a **sampling (frequency) distribution of these statistics!**

If a frequency distribution is constructed for **some** (but not all) of the possible samples of size *m* drawn from the population of size *n*, this distribution is called a **sample sampling distribution**. The name is derived from the idea that a set of fewer items than the total number available is a *sample*, any distribution made is a sample of the sampling distribution. This concept is not likely to turn up before the second term, when you will learn about a procedure (called the Kolmogorov-Smirnov test) which allows you to check to see if a set of numbers (sample) was drawn from a given (sampling) distribution.

Sampling distributions related to desirable statistic qualities

Three of the four desirable statistical properties (unbiased, efficient, and consistent) can be easily illustrated using sampling distributions. A statistic is **unbiased** if its sampling distribution is centred around the value of the parameter we're trying to estimate. If you have two different sample statistics trying to estimate the same parameter, the more **efficient** of the two is the one whose sampling distribution is less broad and more peaked around the parameter value. A statistic is **consistent** if the sampling distribution becomes more peaked (with the peak at the parameter value) as the sample size increases.

Consistency is a most important property! All of the students in the GGR 270 class can be considered a population. Suppose I want to estimate the mean weight of the students of the class by selecting a sample and finding its mean weight. Intuitively, we expect that as the sample size increases, the sample mean **for any sample** should tend to be closer to the true mean. For this to be true, the sampling distribution for a large sample size should be more peaked around the parameter value. The core concept of **inferential statistics** is looking at the location of a sample statistic within the sampling distribution and seeing how likely it is that the value would have occurred as a result of the natural variation of sample values. This will be covered in more detail later.

Example of the creation of a sampling distribution

Six rocks were extracted from someone's head and each was weighed, labeled, and put in a bag. This forms the **population** from which I can draw samples. Suppose I want to construct a sampling distribution of the *mean weight of 3 rocks from the population of 6*. To do this, I must enumerate **all** samples of size 3 which can be drawn from a population of size 6 (there are 20 in total) and compute the mean of each. The frequency distribution I can create from these 20 numbers is the **sampling distribution** I want. Below is the table I would use to create this distribution, and below that is the actual sampling distribution.

Example: Creation of a Sampling Distribution

Rock ID	1	2	3	4	5	6	
Weight (g)	11.24	13.48	16.9	24.28	20.89	10.43	Sample Mean
Sample 1	1	1	1	0	0	0	13.87
Sample 2	1	1	0	1	0	0	17.73

GGR 270Y First Term Lecture Notes

Sample 3	1	0	1	1	0	0	18.87
Sample 4	0	1	1	1	0	0	19.62
Sample 5	1	1	0	0	1	0	15.20
Sample 6	1	0	1	0	1	0	16.34
Sample 7	0	1	1	0	1	0	17.09
Sample 8	1	0	0	1	1	0	20.20
Sample 9	0	1	0	1	1	0	20.95
Sample 10	0	0	1	1	1	0	22.09
Sample 11	1	1	0	0	0	1	11.72
Sample 12	1	0	1	0	0	1	12.86
Sample 13	0	1	1	0	0	1	13.60
Sample 14	1	0	0	1	0	1	16.72
Sample 15	0	1	0	1	0	1	17.46
Sample 16	0	0	1	1	0	1	18.60
Sample 17	1	0	0	0	1	1	14.19
Sample 18	0	1	0	0	1	1	14.93
Sample 19	0	0	1	0	1	1	16.07
Sample 20	0	0	0	1	1	1	19.93

For this example, each row represents one of the possible 20 samples of 3 rocks from a population of 6. A 1 in the column means that the particular rock is part of the sample, and a 0 means that it is not. For example, Sample 13 consists of rocks 2, 3, and 6. The *Sample Mean* column is the mean of the weights of the rocks in the sample. Notice how all of the numbers are different.

The Sampling Distribution

Bin	>11, <=13	>13, <=15	>15, <=17	>17, <=19	>19, <=21	>21, <=23
Frequency	2	4	4	5	4	1

D. The Histogram

A histogram is a graphical method of displaying a frequency distribution. The **range** of values is on the horizontal (x) axis, while the **frequency** of occurrence is on the vertical (y) axis. Histograms are usually bar-type charts, but can be line graphs as well.

An **ogive** is a *cumulative relative frequency histogram* plotted as a line graph. (Know this! It's a favourite question on exams!)

The observations always have to be **grouped** in some way. The grouping chosen will depend on the data's range.

i) Guidelines For Grouping

There should be 6-12 classes, or about $1+3.3\ln N$ for "large" N. (This is NOT a strict rule! It is just a guideline!) A good initial guess is $N/3$ or $N/4$.

Each class **MUST** be the same width, i.e. same % of range. Uneven widths will lead to misleading displays!

Classes **MUST** be *mutually exclusive* (no overlapping classes) and *collectively exhaustive* (no gaps in the coverage) over the range.

Try to use a class width that fits a natural pattern in the data.

Watch for *outliers*, which can complicate things. An outlier is a number whose value is not close to any of the others. For example, if you have a bunch of rainfall values clustered around 5 mm, and one day when 20 mm fell, the 20 mm value is an outlier.

ii) Basic Procedure for Making a Histogram

This example will demonstrate how the sampling distribution of the previous example was created. Numbers in brackets indicate the results for the set of means.

Compute the range of values, which is (High-Low) or (High-Low+1 for integers). **(22.09-11.72=10.37)**

Choose the *initial* number of classes. In this case, since there are 20 numbers, it makes sense to have 5 classes, since it divides nicely into the range. **(5)**

Compute the range width, which is simply the range divided by the number of classes. $(10.37/5=2.074)$. Round this number down to 2. **(2)**

Choose a starting point below the minimum value of the set of numbers. It is helpful if this starting point and the range width are “simple”, i.e. multiples of 2, 5, 10, etc. In this case, 10 is a nice round number which is less than the minimum. **(10)**

Here is where the juggling begins. First of all, the total range spanned by the number of classes and class width is $5*2=10$. Unfortunately, the total range of the numbers is 10.37! This is what happens when you round down. For the sake of simplicity, add another class interval to make the spanned range 12. But $10+12=22$, and the highest number is 22.09. $11+12=23$, and 11 and 23 enclose the range of the number set, so we settle on starting at 11 with 6 classes of width 2.

The lowest range of the histogram should always be less than the lowest value of the numbers. As a result, the first class' boundaries should be values strictly >11 . The upper bound should be numbers ≤ 13 , since this pattern will be repeated for all classes to ensure they are mutually exclusive and collectively exhaustive.

The *actual bounds* of the classes will depend on the accuracy of the data. For example, the class of 10-19 would have limits of 9.5 to 19.49 if the numbers to be grouped were accurate to one decimal, like 10.1 or 19.2.

The **midpoint** of each class is simply $(\text{start}+\text{end})/2$, i.e. (12,14, 16, etc.). Midpoints are required to plot line histograms, such as ogives, as they will be used as the x-coordinates. Bar histograms simply have the bar's width be the width of the interval.

Assign the observations to the categories and create the diagram of choice.

E. Simple Descriptive Summary Measures

Most of these you have probably seen before, but they are presented here for reference anyways.

i) Ratio: The ratio between two variables is simply $(\# \text{ in A})/(\# \text{ in B})$. A ratio gives information about whether the variable A is less than, equal to, or greater than the variable B, but gives no info about what the actual values are.

ii) Proportion: A proportion relates one part or category of the data to the whole. The proportion is defined as

$x_i / \sum_{j=1}^n x_j$, where x_i is the count in group i, while $\sum_{j=1}^n x_j$ is the total count in all groups (the sum of the counts in all of

the groups). All proportions are between 0 and 1.0, and the sum of all proportions is 1.0.

For example, in the histogram example the proportion of numbers in the 7 category is $6/36$ or 0.167. Note that the proportion is the same as the *relative frequency*!

Proportions are useful when comparing 2 sets of data with different sizes and category counts. For example, comparing traffic counts at certain times of day on Bloor St. (a major artery) and St. George St. (a “collector” street) to see if peak traffic periods for both streets coincide or not.

In order for a proportion to be truly informative, you need not just the proportion, but also how many things are in the whole. “Half of all dentists surveyed recommend toothpaste X.” The question is, how many were surveyed? Two?

iii) Percentage: This is just the proportion*100, and is very commonly used everywhere. As with proportions, one must know the whole for it to be useful. For instance, unemployment in Toronto area (pop 3 million) of 7% vs London Ont. area (300,000) of 7% involve different numbers of people. **A warning:** percentages and proportions can be deceptive for small numbers. (1 out of 4 is 25%!). Percentages and proportions are probably best with about 50 or more observations.

iv) Rate of Change: A rate of change is defined as $R = \frac{x(t_2) - x(t_1)}{x(t_1)}$. This is read as (the change in the observed

variable) divided by (initial value of the variable). Given the rate of change and the new count, one can compute the old count by some simple algebraic manipulation.

v) Location Quotient: This rather esoteric summary measure is an index of *relative concentration in space*, that is, at a location in a region. It is a comparison of a **region's share of a particular variable** compared to its total over the map.

Example: Given a region of 1000 km², divided into 3 regions, A=200, B=300, and C=500 km². Suppose that in this region an outbreak of the debilitating disease infectious statspanikitis is occurring. Region A reports 150 cases, B has 100, and C has 350, for a total of 600 students afflicted.

Intuitively, one “expects” that the number of cases of the disease will be proportional to the area of each sub-region. The LQ tests to see if this is indeed the case.

	Area	Disease	Location Quotient
Sub-Region A	200/1000 = 0.20	150/600 = 0.250	0.250/0.20 = 1.250
B	300/1000 = 0.30	100/600 = 0.167	0.167/0.30 = 0.553
C	500/1000 = 0.50	350/600 = 0.584	0.584/0.50 = 1.168

In essence, the Location Quotient is the proportion of the variable divided by the proportion of the area it's in. If the number of cases is in fact proportional to the area, the LQ should be 1.0.

In this example, sub-regions A and C have more cases than one would “expect”, whereas B has significantly fewer cases.

IV. SIMPLE UNIVARIATE DESCRIPTIVE STATISTICS

What we would like are ways to summarize a distribution quickly and accurately. Graphs and plots can sometimes be a pain to deal with (i.e. to generate) and still do not help convey the required information in a way easily put into words.

The most common way to summarize a distribution is to state a measure of its **central tendency** and its **dispersion**, measures which will depend on the variable's *scale*.

A. Measures of Central Tendency

i) The Mean (Interval and Ratio Scales)

Mean is not just a term used to describe Stats profs! Also known as the **average**, the mean is the number which is in the middle of all the numbers in the distribution. This statement makes more sense if you plot all of the numbers on a number line; the mean will be right in the middle. This is probably the one statistic that everyone is familiar with.

We define:

Sample Mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and the **Population Mean:** $\mu = \frac{1}{N} \sum_{i=1}^N x_i$, where N is the size of the population and n is the size of the sample.

The mean has 2 useful properties: $\sum_{i=1}^n (x_i - \bar{x}) = 0$ (the sum of deviations about the mean is zero), and $\sum_{i=1}^n (x_i - y)^2$ is a minimum when $y = \bar{x}$.

Note that in the above formulas, every observation is treated as *equally significant*. When this is not the case, the **weighted mean** is used. It is defined by

$$\bar{x}_{\text{wgt}} = \left(\sum_{i=1}^n w_i x_i \right) / \left(\sum_{i=1}^n w_i \right)$$

w_i is the importance (weight) of observation x_i . Many types of data analysis attach different weights to the various values. The most common weight is simply the number of occurrences. For instance, if a survey of 100 students's house plants (not including their room-mates) revealed that 20 had 1, 30 had 2, 20 had 3, 16 had 4, and 14 had 5, the **weighted mean** number of plants is $(20*1 + 30*2 + 20*3 + 16*4 + 14*5)/(20 + 30 + 20 + 16 + 14) = 2.74$. The *unweighted* average of $1 + 2 + 3 + 4 + 5$ is 3, hence the weighting by frequency reduces the mean value (in this case).

If all of the weights are equal, the weighted mean formula reduces to the usual mean formula (try it yourself to verify). Just remember that the sum of a constant value is just n times that value.

The **Grouped Mean** is used when you are given data that have already been organized into groups. In other words, you have a set of (class) midpoint values and the count of occurrences (a frequency). Computation of the mean is directly analogous to the weighted mean, with the *frequency used as the weight*. The formal definition:

$$\bar{x}_{grp} = \left(\sum_{i=1}^n f_i m_i \right) / \left(\sum_{i=1}^n f_i \right), \text{ where } \sum_{i=1}^n f_i = n.$$

The m_i are the midpoint values for each group or class. For example, consider the following frequency distribution:

Group	1-2	2-3	3-4	4-5	5-6
Midpoint	1.5	2.5	3.5	4.5	5.5
Frequency	3	4	6	5	2

The grouped mean is thus $(1.5*3 + 2.5*4 + 3.5*6 + 4.5*5 + 5.5*2)/20 = 3.45$

ii) The Median (Ordinal scale)

The median observation is the one at which half of the observations are above it and half are below it. One must sort the observations and find the one which is the middle, an often tedious process. It can be applied to ordinal scale variables, as well as interval/ratio scale ones.

An interesting property is that $\sum_{i=1}^n |x_i - y|$ (the sum of the absolute values of the differences) is a minimum when y is the *median*.

The **position** of the median is observation $(n+1)/2$ (if n is odd). If n is even, the median will be at $n/2 + 0.5$ (once the obs are ordered).

The median can be computed for *grouped* data too, but it's a bit tricky. When you construct a cumulative frequency chart, note where it goes *above half the total observations*. The median will be in that group somewhere. "Exactly" where can be found in Blalock, pp. 64-65, by linear interpolation, but it isn't crucial to the course, so don't worry about it!

Note that the median is not a **sufficient** statistic for ratio data (i.e. it doesn't use all the information available in the data). Consider the set of numbers -3, -1, 4, 8, 13, whose median is 4. Suppose we replace the -3 with -13 and the 13 with 20. The median is still 4! The means are different (4.2 and 3.6 respectively) because the observation *values* are used to compute it. To compute a median, only the **order** of the numbers is used.

iii) Mode (Nominal Scale)

The **mode** is the value which occurs most frequently in a distribution. It can be found for all types of variables. If two values have a tie, the data is *bimodal*. There is a formula to compute the mode from grouped data, which is in the text and rather ugly. One could probably cheat and just say that the mode is in the group with the highest count.

iv) Which is better?

The mean is the most justifiable measure because of its useful mathematical properties. It uses all of the data. However, it is vulnerable to extreme values (outliers) in the set. e.g. the set {5,6,3,7,4} has a mean of 5. If the 7 were changed to 17, the mean would become 7, which is greater than all of the remaining numbers!

The median is usually better when the distribution is **skewed**. A **positively** skewed distribution has the tail pointing to the *right*, i.e. towards *increasing* x values. A **negatively** skewed distribution is the opposite. See Griffith and Amrhein, p. 82, or Burt and Barber, p. ??? for examples. It is always best to use the highest scale possible to keep the maximum amount of information.

Of course, there will always be some distributions in which there are no really good measures of central tendency. IT NEVER HURTS to have a picture (in your mind or on paper) of what the distribution looks like!

v) Quartiles, percentiles: A quartile is found by dividing the data range into 4.

1st quartile: location where 25% of data is below. *3rd quartile*: location where 25% of data is above, or where 75% is below. *2nd quartile* = median.

A **percentile** is found by dividing the data range into 100.

B. Measures of dispersion

Measures of dispersion measure the *spread* of the data values in the distribution. It should come as no surprise that there are several ways to measure dispersion.

i) Range

The **range** is simply the highest value-lowest value. By definition, it is affected by the extreme values in the data, and is thus useful only in a very general sense.

ii) Interquartile Range

Interquartile range is defined as Q3-Q1, the difference between the first and third quartiles. Sometimes you see the semi-interquartile: (Q3-Q1)/2. It essentially is a measure of the width of the peak of the distribution, and corrects for the extreme values, but it is an ordinal scale only.

iii) Mean Deviation

This is really useful only in descriptive stats. It is the average of the absolute deviations from the mean. The absolute value makes it very awkward to work with mathematically because it is not a continuous function.

$$\text{Definition: } MD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

iv) Variance and Standard Deviation

It was mentioned above that the sum of the squares of the deviations of the data from the mean produces a minimum value. We use an *average* of these squared deviations to compute the **variance** and **standard deviation**. As usual, there are two slightly different definitions of the two statistics, depending on whether you're looking at the population as a whole, or a sample from it.

Population Variance: $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$

Sample Variance: $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

Population Standard Deviation: σ

Sample Standard Deviation: s

The sample variance is divided by n-1 instead of n in order to make it **unbiased** (this can be proven mathematically, but I won't go into it here).

The computational formula for variance is $s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right)$

The advantages of using this formula are that you only have to make one pass through the data set, rather than making one pass to compute the mean, then another to compute the variance, and also that roundoff error is reduced.

The formula for σ is similar, but with 1/n replacing 1/(n-1) at the beginning.

For GROUPED DATA, such as that from a frequency distribution, the formula is a bit different:

$$s_{grp}^2 = \frac{1}{n-1} \sum_{i=1}^{n_g} n_i (x_i - \bar{x}_{grp})^2$$

Here, the x_i are the **midpoints** in the groups, n_i is the number of observations in class i , and n_g is the total number of classes. Note that we use the *grouped data mean* here, and that the differences are weighted by the **frequencies** of the groups. A similar formula for the population exists, just use N instead of $n-1$.

Example: Look at the sampling distribution of 3 rocks from a sample of 6 done a few classes ago. The midpoints of that distribution were 12, 14, 16, 18, 20, and 22, and the respective frequencies were 2, 4, 4, 5, 4, 1. The **grouped mean** is $(12*2 + 14*4 + 16*4 + 18*5 + 20*4 + 22*1)/20 = 16.8$ (compare to the "true" mean of 16.902). The **grouped (sample) standard deviation** is $\sqrt{[2*(12-16.8)^2 + 4*(14-16.8)^2 + \dots + 1*(22-16.8)^2]/(20-1)} = 2.858$ (as

opposed to the “true” sample std. dev. of 2.80). Note how information is lost when the values are grouped, and hence the statistics are slightly different!

Remember that the sum of the frequencies equals n, the total number of observations.

A computational formula similar to the above can be derived (you do it!).

Just as the mean can be distorted by large outliers, the standard deviation is even more prone to distortion, since it depends on the *square* of the differences.

v) **Z-Scores** (Will show up later in normal distribution!)

$$Z_i = \frac{x_i - \bar{x}}{s}$$

The Z-score of observation x_i is the answer to the question “how many standard deviations away from the mean the observation is.” Note that using this assumes the x_i are drawn from a normal distribution (to be discussed later!). The greater the number of standard deviations away from the mean the observation is, the less likely that it will have occurred by random chance. For example, any $Z > 1.65$ or < -1.65 is only likely to occur 10% of the time if the Z's are random, and any $Z > 1.96$ or < -1.96 is likely to occur only 5% of the time if the Z's are random. Testing to see if the value of a sample statistic is *significant* (i.e. is unlikely to happen by random chance) is the core of *inferential statistics*.

$Z = 1.5$ is about the threshold where you start to consider the observation a significant value.

vi) **The Coefficient Of Variation**

The CV = $\frac{s}{\bar{x}}$ for samples, $\frac{\sigma}{\mu}$ for populations. It is **dimensionless**, since the mean and the standard deviation have the same units. It can be expressed as a decimal or a percentage. It is the standard deviation *expressed as a fraction of the mean*.

What's it good for? It's useful because it allows comparisons of the relative dispersions of two samples which may have widely different means, use different units, and have different sample sizes.

How to compare dispersion of # of cars on the Don Valley Parking Lot compared to # of cars on Yonge Street? Say the DVP has 100,000 cars/day, $s=4000$; Yonge has 60,000, $s=600$. $CV(DVP)=0.04$, $CV(Yonge)=0.01$. Thus Yonge Street has a **smaller** relative variation than the DVP, which would show up in the distributions as Yonge Street's peak being narrower than that of the DVP.

vii) **Skewness**

In English, skewness is the (“3rd moment of distribution”)/(the cube of s).

$$\text{Skewness} = \frac{1}{ns^3} \sum_{i=1}^n (x_i - \bar{x})^3. \text{ Another way to compute it is } \text{Skewness} = \frac{3(\bar{x} - \text{median})}{s}$$

(which is usually faster!). If zero, the distribution is symmetric; if > 0 , the mean $>$ median and the distribution is positively skewed (tail on right); if < 0 , the distribution is negatively skewed.

viii) **Kurtosis**

This is not a disease statisticians get (though it sounds like one you'd like to inflict on one)! It is (the 4th moment of the distribution)/(s^4).

$$\text{Kurtosis} = \frac{1}{ns^4} \sum_{i=1}^n (x_i - \bar{x})^4$$

Kurtosis is a measure of how peaked or flat a distribution is. The more dispersed the values (i.e. the broader the curve is), the greater the average $x - \bar{x}$ will be, and the greater the kurtosis.

C. The Lorenz Curve and Gini Coefficient

i) The Lorenz Curve

This is used to compare the distributions of two different variables. It is easy to make: just plot the cumulative frequency distributions of the two variables against each other on a graph (one along the X-axis, the other along the Y-axis).

The classic (i.e. textbook) example is income distribution among a population. In an ideal situation, income should be distributed equally among each population class and you'd get a straight line. Because it's not equally distributed, a curve will be plotted above or below the straight line, depending on the ordering of the incomes.

Another example:

A survey was done to compare the age and average number of house plants in the homes of 45 people. The results are listed below:

Ages	# People	CF	CRF	# Plants	CF	CRF
20-21	5	5	0.11	4	4	0.11
22-23	8	13	0.29	7	11	0.31
24-25	15	28	0.62	6	17	0.48
26-27	11	39	0.87	8	25	0.71
28-29	6	45	1.00	10	35	1.00

(Unfortunately, I don't have time to create the plot! Sorry!)

There are 5 points to plot. Make the People CRF the x coordinate and the Plants CRF the y coordinate. Also be sure to plot the diagonal line from (0,0) to (1,1) which represents the perfect correspondence, and also **Label each point plotted so that it can be identified!**

If the curve is above the diagonal, it means that the y-axis variable is proportionally greater than the x-axis variable, and if it is below the diagonal, vice versa. Of course, the curve may be partly above the line and partly below, which indicates that the inequities shift back and forth.

Burt and Barber suggest that geographically referenced data be sorted by location quotient before being plotted, since any numbering system applied to the spatial units is arbitrary. Data which are presented as frequency distributions but are not geographically referenced should be plotted in the order of the distribution categories.

The x axis is usually human related (e.g. population), when such a category is to be examined.

ii) Gini Coefficient

The **Gini Coefficient** is a summary measure of the deviation in the Lorenz curve.

$$G = 0.5 \times \sum_{i=1}^n |x_i - y_i|$$

n is the number of groups (5 in this case), x_i and y_i are **relative** (NOT cumulative) frequencies for the x and y axis scales respectively.

G represents the area between the curve and the diagonal line.

For our example, $G = 0.5 \times (|0.11-0.11| + |0.18-0.20| + |0.33-0.17| + |0.24-0.23| + |0.13-0.29|) = 0.5 \times (0 + 0.02 + 0.16 + 0.01 + 0.16) = 0.175$

V. CORRELATION

When data are collected, usually more than one variable is recorded for a given event. For example, traffic counts may record number of cars, trucks, time of day, speed; a survey of people may look at their level of education and income; weather data may include temperatures, precipitation, hours of sunshine, and wind.

The researcher is usually interested in any *relationships* between the variables. Does education play a role in income level? Is the temperature related to how much sunshine a station gets? Is there more traffic at a certain time than at another?

The **strength** of the relationship is also of interest. Statistics give us measures of both the strength and significance of any relationship. The following items are various **test statistics** which we can compute to look for evidence of correlation between two variables.

Be aware of the old cliché “*Correlation does not prove causation*”!

There is nothing difficult about this--determining correlations is mostly tedious calculations.

A. Phi Coefficient (Nominal Scale Data)

Purpose: To determine the strength of association between 2 **nominal binary variables** (variables which can only have 2 values)

Definition:
$$\phi = \frac{AD - BC}{\sqrt{(A + B)(C + D)(A + C)(B + D)}}$$

A, B, C, and D are defined as follows:

		Variable 1		Total
		Category 1	Category 2	
Variable 2	Category 1	A	B	A+B
	Category 2	C	D	C+D
Total		A+C	B+D	N

ϕ ranges from -1 to 1 with **zero** meaning no relationship, **1** is a perfect direct relationship, **-1** a perfect negative relationship (i.e. Category 2 only occurs when Category 1 does not).

For example: The Smurfs are always looking for more smurfberry bushes, but they are hard to find. However, suppose that the barfberry, which resembles the smurfberry, but has the undesirable side effect suggested by its name, is easier to find. Papa Smurf thinks that there may be a relationship between the barfberry plant's and the smurfberry plant's proximity to one another, so he sends out four search parties to different parts of the forest to check. In the table, I'll use the short forms **S** for smurfberry, **B** for barfberry, and + for present and - for absent.

	Party 1			Party 2			Party 3			Party 4					
	B+	B-		B+	B-		B+	B-		B+	B-				
S+	30	0	30	S+	0	15	15	S+	20	2	22	S+	11	9	20
S-	0	6	6	S-	21	0	21	S-	6	8	14	S-	8	8	16
	30	6	36		21	15	36		26	10	36		19	17	36
ϕ	$\phi = (30*6 - 0*0) / \sqrt{(30*6*30*6)} = 1$			$\phi = (0*0 - 21*15) / \sqrt{(15*21*15*21)} = -1$			$\phi = (20*8 - 2*6) / \sqrt{(22*14*26*10)} = .52$			$\phi = (11*8 - 9*8) / \sqrt{(20*16*19*17)} = .05$					

Party 1 reports that whenever barfberry is present, so is smurfberry.

Party 2 reports that whenever barfberry is present, smurfberry is not.

Party 3 reports that there is a pretty strong correlation between the presence of barfberry and smurfberry which is apparent in the numbers.

Party 4 reports that there is little correlation between the two plants.

All of the terms are of roughly equal magnitude.

So Papa Smurf has to go back to the drawing board to see if another plant is a universal indicator of smurfberry presence. Meanwhile Greedy Smurf has to go to the infirmary because he gobbled a handful of the wrong berries.

NOTE: It is important to clearly specify variable definitions! By changing the variables around you can change the sign (but not the value) of the statistic.

B. Chi-Squared Statistic (Nominal Scale Data)

This statistic is an extension of the concept of the ϕ coefficient to **polychotomous** (i.e. more than two possible values) nominal scale variables.

The statistic is based upon the analysis of a **contingency** (or **cross-tabulation**) table. These are created when we have two nominal scale variables which have been *cross-classified*.

For example, suppose we wish to see if there is any relationship between the type of alcoholic beverage a person prefers and the political party he or she supports. The survey results are below:

		Political Party				
		PC	Liberal	NDP	Reform	Totals
Preferred Beverage	None	4 (6.3)	7 (6.4)	7(6.2)	10 (9.1)	28
	Beer	8 (16.4)	16 (16.8)	22 (16.1)	27 (23.7)	73
	Wine	21 (10.6)	10 (10.8)	6 (10.3)	10 (15.3)	47
	Spirits	12 (11.7)	13 (12.0)	9 (11.4)	18 (16.9)	52
	Totals	45	46	44	65	200

Interpreting the Totals row and column: The totals **row** shows the numbers of people (the *distribution*) who support each political party. The totals **column** shows the distribution of people who drink each type of booze. We can easily find the *proportion* of party supporters or drinkers by taking the appropriate total and **dividing by the sample total** which is in the lower right corner.

As always in correlation, we are looking for evidence of some sort of relationship between the two variables. The way to do this is by comparing the frequencies in the table to *expected frequencies*.

We start with the assumption that there is NO relationship between the variables. In other words, regardless of the actual proportion of Liberals there are in each beverage category, in the long run we can expect the same proportion of Liberals (which will be the proportion of Liberals in the ENTIRE sample) to be in ALL beverage categories.

To find the **expected number of Liberal spirits drinkers**, we take the proportion of Liberals in the sample, (46/200), and multiply by total number of spirits drinkers, 52, to get 12.0. The expected number of Liberal wine drinkers is the proportion of Liberals in the sample, (46/200), times the total number of wine drinkers, 47, to get 10.8. For Reform beer drinkers, take the proportion of Reformers in the sample, (65/200), and multiply by the total number of beer drinkers, 73, to get 23.7.

It is a good idea to **keep at least one decimal place** when computing the expected numbers, to help reduce roundoff error.

Note that this works BOTH WAYS. To find the number of teetotallers who are PCs, multiply the proportion of teetotallers in the sample (28/200) by the number of PCs in the sample, 45, to get 6.3.

See the pattern? To find the expected number for cell c_{ij} , multiply the row total r_i by the column total c_j , and divide by the total N in the sample.

We define the expected value for cell ij by $e_{ij} = \frac{r_i c_j}{N}$. All of the expected numbers will add up to the row and column totals.

The χ^2 statistic itself is defined as follows:
$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(e_{ij} - o_{ij})^2}{e_{ij}} = \sum_{i=1}^n \sum_{j=1}^m \frac{o_{ij}^2}{e_{ij}} - N$$

where e_{ij} are the expected frequencies and o_{ij} are the actual observation frequencies. The double sum simply means *sum over the whole table*.

For this example we have
$$\chi^2 = \frac{(6.3 - 4)^2}{6.3} + \frac{(6.4 - 7)^2}{6.4} + \dots + \frac{(11.4 - 9)^2}{11.4} + \frac{(16.9 - 18)^2}{16.9} = 22.7$$

So what do we do with it? Recall that the assumption we used to compute the expected values was that the two variables were independent. Large values of χ^2 imply that the assumption is false, since some or all of the values are greatly

different from the expected values. To find the critical value of χ^2 , we must look at a table for the distribution. (Details will be provided next term.)

A contingency table of r rows and c columns has (r-1)(c-1) degrees of freedom. Here, it's 9. For a 5% confidence interval, we reject the hypothesis of independence if $\chi^2 > 16.9190$. Since our value is 22.77, we can say that there is definitely a preference for certain types of beverages among party members. (Don't worry, you'll be seeing this sort of thing all too often next term!)

C. Spearman Rank Correlation

Purpose: To discover possible correlation between 2 **ordinal** variables.

It measures **consistency**, because if 2 variables are consistently related, the ranks of their observations will be linearly related. E.g. larger object= higher price.

Definition:
$$r_s = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)}$$

D_i is the *difference in rank* between the first and second variables for observation i. The range of the statistic is from -1 to +1.

How to do it:

Rank the observations of each variable. Break any ties by assigning the tied observations a rank equal to the arithmetic mean of ranks they would have had if they were slightly different. E.g. suppose 3 people got 76 on an exam and 76 ranks 4th in the scores. Instead of assigning rank 4 to all of them, assign $(4+5+6)/3 = 5$. The next highest score (say 72) would have rank 7. Ties tend to reduce the coefficient's accuracy. For each event, find the square of the difference between the variables. Sum these values and stuff the result into the formula.

Example: City data

Size	Crime	Clima	Pollu	Siz-Cri d ²	Siz-Cli d ²	Siz-Poll d ²
1	1	6	3	0	25	4
2	2	5	1	0	9	1
3	3	4	6	0	1	9
4	4	3	5	0	1	1
5	5	2	2	0	9	9
6	6	1	4	0	25	4
				$r_s=1$	$r_s=-1$	$r_s=1-(6*28)/210= 0.2$

Note that you can use this statistic on interval and ratio scale data as well, but the results will not be as good as those from the **Pearson** correlation coefficient (see below). This is because the Spearman statistic is not **sufficient** for interval or ratio data, since only the order and not the actual values are important!

We now move into a discussion of *parametric techniques*, which allow one to examine both the direction and strength of interval/ratio data correlations.

D. Scattergrams

Purpose: Plotting the two variables as points on a graph will give an idea of how they might be related. This type of plot is called a scattergram.

Types of relations	Examples
Positive [line of + slope]	Slope & Erosion; Attendance & Grade
Negative [line of - slope]	PC popularity and time, income & crime
No relation [blob]	Temp & River drainage basin Area
Non-Linear [parabola]	

E. Covariance

Purpose: To describe the joint variation or dispersion of X and Y.

Definition: (sample) $COV_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \left[\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \right]$

For the population covariance, use 1/n in front.

Instead of the sum of squares of x or y, we have the product of their deviations from their respective means. This gives the way they vary **together**. If X and Y are *positively* related, then large value of X will normally be related to large values of Y. If $X > \bar{x}$, then $Y > \bar{y}$ usually. Therefore the product of the terms will usually be positive. In a *negative* relationship, a large value of X will normally be associated with a large negative value of Y. Certainly, if $X > \bar{x}$ then $Y < \bar{y}$ and the product of the terms will be negative. The sign of covariance can thus be either positive or negative. If it happens to be zero, there is no *linear* relation between the two variables.

Its value can be anything from $-\infty$ to $+\infty$, which helps make comparisons between covariances difficult. It's even worse when different sample sizes are involved.

F. Pearson's Correlation Coefficient

Purpose: Provides a normalized coefficient (i.e. between -1 and 1) that can provide much easier comparisons between different data sets. Note that a lot of statistics are that way, and that the sometimes bizarre-looking denominators are there solely to normalize the statistic.

Definition:

$$r = \frac{COV_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2}}$$

Explanation: This formula is more easily remembered as the sample covariance divided by the product of the sample (unbiased) standard deviations of X and Y. The second formula is the **computational version**, which will reduce roundoff error. The denominator is always positive. The numerator (covariance) controls the sign. The range of the correlation coefficient is from -1 to +1. Zero again means no *linear* relation between X and Y.

The formula looks hideous, but it's not overly complicated and is merely tedious to compute. This is where computers become so useful!

Comparing Pearson's with Spearman's for Ratio-scale data

Suppose we wish to see if there is a relationship between the average amount of snow over a river basin and the total precipitation it receives. The following figures were lifted from a dataset in the Griffith and Amrhein book and are in inches of snow and inches of water respectively. The following is a worktable that is typical of one used to hand-compute Pearson's and Spearman's correlation coefficients. Computing both will give an idea of why it is better to use the Pearson coefficient for ratio or interval data. As this is a *sample* of the rivers dataset for Ontario, the sample variance and covariance formulae must be used.

River Name	Snow	Precip	S ²	P ²	S*P	Rank(S)	Rank(P)	D ²
Ausable	76	22	5776	484	1672	5	1	16
Big Otter	60	35.5	3600	1260.25	2130	2	6	16
Black R	92	36	8464	1296	3312	7	7	0
East Oakville	48	30	2304	900	1440	1	3	4
Humber	60	29.5	3600	870.25	1770	3	2	1
Middle Maitland	108	35	11664	1225	3780	8	5	9
Nith	80	37	6400	1369	2960	6	9	9
Nottawasaga	72	31	5184	961	2232	4	4	0
Saugeen	120	36	14400	1296	4320	9	8	1

Sums	716	292	61392	9661.5	23616	56
------	-----	-----	-------	--------	-------	----

Variance of Snow: $(61392 - (1/9) * (716)^2) / (9 - 1) = 553.78$ Standard Deviation: 23.53

Variance of Precip: $(9661.5 - (1/9) * 292^2) / (9 - 1) = 23.40$ Standard Deviation: 4.84

Covariance of both: $(23616 - (1/9) * 716 * 292) / (9 - 1) = 48.22$

Pearson Correlation: $48.22 / (23.53 * 4.84) = 0.424$

Spearman Correlation: $1 - (6 * 56) / (9 * (81 - 1)) = 0.533$

Note how the Spearman coefficient implies a better correlation than the Pearson coefficient. Remember that the Spearman is not a **sufficient** statistic for ratio/interval data because it doesn't use the data values, only their *order*. This introduces inaccuracy. Also note that you should keep more significant figures than what I wrote down for your intermediate steps, to reduce roundoff error.

G. Interpreting Results

It is important to stress again that just because two variables have a high correlation, it is not necessarily true that “X causes Y”. Other, perhaps less important variables, may also be contributing.

Example: square feet of office space and number of trips to the building (lobby traffic). Secondary variables of importance are, among others, the type of business(es) in the building, how many of them there are, and their location in the office building. X and Y could mutually interact, as opposed to cause and effect. For example, low income, lack of education, unemployment, poor housing, and criminal activity are all interactive (exception: politicians...) The relationship may be spurious, caused by an unobserved third factor which controls both X and Y. Example: Cancer and number of groundhogs nearby.

A zero correlation does not necessarily mean there is no relationship between X and Y. It means there is no *linear* relationship! Unless we look at the scattergram, for all we know the points could lie on a parabola! This is why it is always a good idea to have a scattergram available to help with the interpretation.

VI. SIMPLE LINEAR REGRESSION

Purpose: to establish a precise mathematical relationship between the two variables X and Y, instead of just finding how well related they are. Once we have the relationship, we can use it for prediction and forecasting.

The simplest possible relation between two variables is a *linear* one, in the form $Y = aX + b$. Y is the **dependent** variable and X the **independent** variable. The form of the equation shows that Y depends on X.

a is the slope of the line. If $a > 0$, Y increases with increasing X; if $a < 0$, Y decreases with increasing X. It represents how fast Y changes with a change in X. **b** is the Y-intercept, the value of Y that occurs when $X = 0$.

Of course, when we plot our data, we don't expect that they will all fall nicely on a straight line. We've all been faced with a bunch of points on a graph, and the burning question “What is the best line that I can draw to match the data?”

This line should obviously be drawn in such a way as to **minimize the deviations between the line and the true data points**. This line, the *regression line*, can then be used to estimate the data points. The equation for the regression line is simply, where \hat{y}_i is the estimated value of the y_i which corresponds to the x_i .

a and b of the regression line are the *regression coefficients*. To find their formulae, we want to minimize the sum of the squares of the vertical distances between the points and the regression line; this is called the *ordinary least squares* approach. It is a simple calculus procedure, and the results are as follows:

$$a = \frac{COV_{xy}}{s_x^2} = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} ; \quad b = \bar{y} - a\bar{x}$$

Note that a regression line is guaranteed to pass through the point (\bar{x}, \bar{y}) . Note also that the formula for the slope can be easily remembered as the covariance of X and Y divided by the variance of X.

Consider the example in which the Pearson and Spearman coefficients were compared. We can easily compute $a = 48.22/553.78 = 0.0871$ and $b = (292 - 0.0871 * 716)/9 = 25.52$.

The **goodness of fit parameter** is called the *coefficient of determination*. It is a measure of how much of the total variation of Y about its mean is “explained” by the regression. See diagram in Black Book, page 442 or Green Book, page 325, to visualize what will be discussed next. Recall the formula used for computing the variance, which can be translated as the average of the sum of squares of deviations of the values from the mean. This *total sum of squares*,

$SST = \sum_{i=1}^n (y_i - \bar{y})^2$, is also a representation of the *total variability* within the variable Y. With a bit of messy

algebra, we can break up this term into two others, the variability “explained” by the regression line, a.k.a. *sum of*

squares of regression, $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ and the residual variability or *sum of squares of errors*,

$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, such that $SST = SSE + SSR$.

We define $r^2 = \frac{SSR}{SST}$, but normally use $r^2 = \frac{SST - SSE}{SST}$ since SSE and SST are more intuitive. r^2 has a range

between 0 and 1. The closer it is to 1, the better the fit, since SSR increases with the better fit. As it turns out, r^2 is just the Pearson correlation coefficient squared. It has a range between 0 and 1 as you might expect. The closer r^2 is to 1, the better the fit.

Another relationship worth noting is $r = \frac{as_x}{s_y}$. It follows from the original definitions of r^2 and the slope parameter.

The **residue** (or residual) is the difference between the observed value and the value predicted by the regression line.

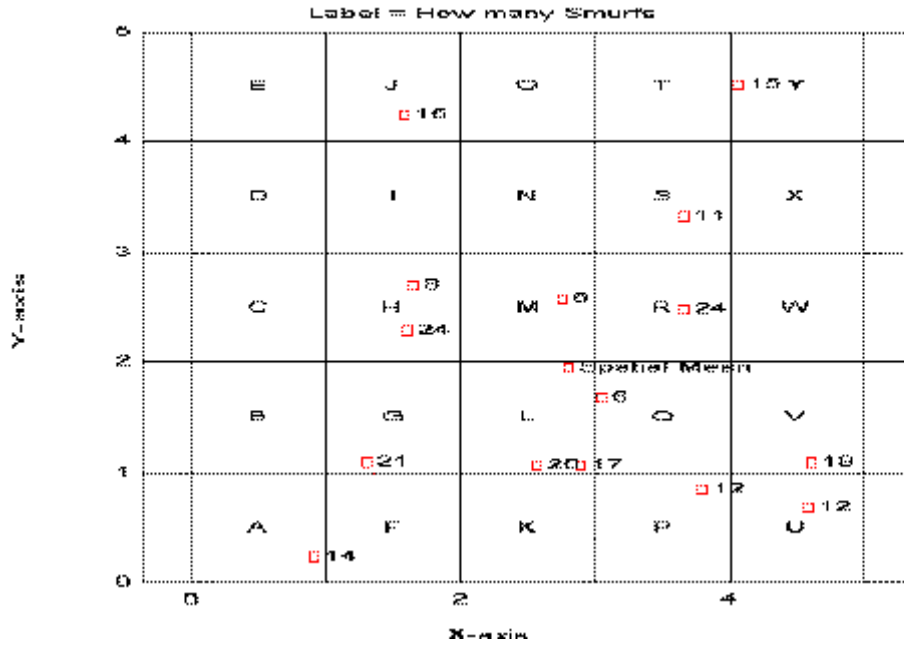
VII. SPATIAL STATISTICS

Geographical data are naturally spaced over an area of land. For example, a city can be divided into wards, census tracts, and political ridings. Often geographers are interested in computing various summary statistics which will take the geographical distribution of the data into account. If four tracts are arranged in a square pattern and the numbers (7,8,4,5) are assigned one per tract, there are 12 possible arrangements for the numbers (see Griffith and Amrhein, page 114). No matter how you arrange them, the mean is 6. This shows that the regular mean is **insufficient** to summarize spatial data.

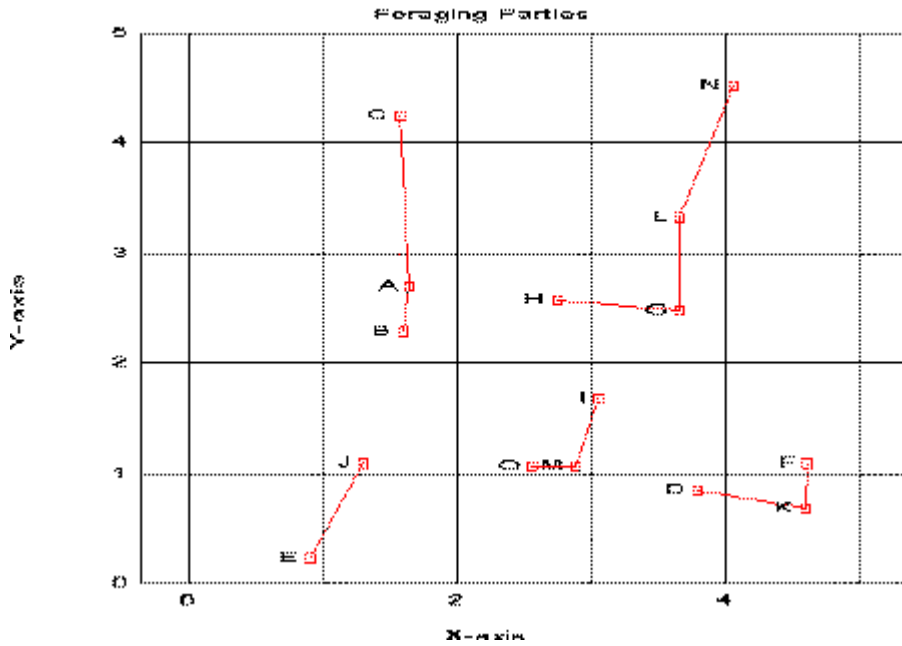
Spatial statistics fall into three categories: *geostatistics*, *point pattern analysis*, and *spatial autocorrelation*. A lot more has been done within all of these categories, but fortunately we only have to cover the basics in each.

The following example is the distribution of Smurf foraging parties in the forest some time one afternoon. The X and Y coordinates mark the locations and the numbers represent the number of Smurfs in each party.

Distribution of Smurfs in the Forest



Nearest-Neighbours for Smurfs



Point	X	Y	Smurfs	NN	Dist		
A	1.65	2.70	8	B	0.413		Spatial Mean is (2.71, 2.01)
B	1.60	2.29	24	A	0.413		Standard Distance is 2.38
C	1.58	4.25	16	A	1.551		Nearest N. Statistic: 1.14
D	3.79	0.85	12	K	0.815		
E	0.91	0.24	14	J	0.935		
F	4.61	1.09	19	K	0.400		
G	3.66	2.48	24	L	0.84		
H	2.75	2.58	6	G	0.915		
I	3.05	1.67	6	M	0.620		

	J	1.30	1.09	21	E	0.935	
	K	4.59	0.69	2	F	0.400	
	L	3.66	3.32	11	G	0.84	
	M	2.89	1.07	17	O	0.330	
	N	4.06	4.52	15	L	1.264	
	O	2.56	1.06	20	M	0.330	

A. Geostatistics

One has to be careful in the definitions of the areas used in all spatial statistics, because they are often (especially the point pattern analysis measures) **very sensitive to the areal pattern**.

i) Spatial Mean

Purpose: This statistic **locates** the “centre of mass” of the data. If you consider the area of study as being a thin plate (of zero mass), and the value of each data point as being a point mass on the plate, the spatial mean is the **location** you’d have to put your finger under to have the plate balance on it. It is two components, x and y, and **not** just one number like the mean!

Each coordinate of the spatial mean is computed separately, using the grouped data mean formulae:

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} \quad ; \quad \bar{y} = \frac{\sum_{i=1}^n f_i y_i}{\sum_{i=1}^n f_i} . \text{ Here, the } f_i \text{ are the } \textit{data values} \text{ at the coordinates } (x_i, y_i).$$

If the spatial mean of the data is significantly different from the geographic centre of the region, this indicates a **non-uniform distribution**. (Think about it!)

If the individual x and y coordinates of the data points are not known, the spatial mean can be estimated, at the cost of introducing measurement error, by the following procedure:

1. Impose a rectangular grid on the area. The grid spacing must be uniform on the x and y axis, but need not be the same for both.
2. Assign values (odd integers, beginning with 1) to the midpoints of the spaces on both axes. We do this so that zero doesn’t come in to complicate things. These values are the x_i ‘s and y_i’s used in the above summation formula.
3. Sum the frequencies (i.e. data point values) in each column along the x axis and row along the y axis. These values are the f_i ‘s used in the above summation formula.

ii) Standard Distance

This is the two-dimensional equivalent of the standard deviation, and is a common measure of *dispersion* in geostatistics. Note that a large value of standard distance means that the points are relatively scattered, while a small value means they are relatively clustered.

$$SD = \sqrt{s_x^2 + s_y^2}, \text{ where } s_x^2 = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum_{i=1}^n f_i} \text{ and } s_y^2 = \frac{\sum_{i=1}^n f_i (y_i - \bar{y})^2}{\sum_{i=1}^n f_i}$$

Note that as with the spatial mean, f_i is the value of the data point. \bar{x} and \bar{y} **spatial mean coordinates**.

This definition is in terms of sample variances, but it usually doesn’t matter if we use the population variance since the frequencies often add up to large values.

NOTE: The given formulas for spatial mean and standard distance are weighted by the data values. For the unweighted versions, just replace the denominators with n and all the f_i values with 1.

iii) Areal Frequency Distribution

This is nothing more than a 3-D map of the geographical distribution being studied. See Griffith and Amrhein p. 126 for an example.

B. Point Pattern Analysis

In point pattern analysis, as the name suggests, attention is restricted to the locational information of the data points, ignoring their values. Descriptive statistics have been created to describe the geographic distribution of data according to their spacing (the **nearest neighbour** statistic) and their density (**quadrat analysis**)

i) Nearest Neighbour Statistic

This statistic gives an idea of, on average, how far a point is from its nearest neighbour in the area of study. First we find the distance from each point to its nearest neighbour. Given the x and y coordinates of the points, we can compute the Euclidean distance, defined as $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. Obviously you don't have to compare a point to ALL of the others on the map! Only the ones that look like potential nearest neighbours. Note that just because A is the nearest neighbour to B does not mean that B will always be the nearest neighbour to A. A point may be the nearest neighbour to several points, or to no points at all.

Having found this average distance, we then compare it to the average point density you'd expect if there was a totally random distribution over the whole area (found by theory to be $\frac{1}{2\sqrt{n/area}}$). If we divide the average distance by the expected distance, we get $R = \frac{\bar{d}}{\frac{1}{2\sqrt{n/area}}}$, where R is the symbol used for the **nearest neighbour statistic**.

Interpretation:

The more closely the points are *clustered* together, the closer to 0 R will be, since the average nearest neighbour distance decreases. The closer R gets to 1, the more *randomly* spaced the points are. This follows from the definition of R. The value of R approaches 2.149 for perfectly uniformly spaced points. (Don't ask me how this was found!) Hence, the closer R is to 2.149, the more *uniformly spaced* the data are.

Drawbacks:

Dependence on area. If you take a given distribution of points and plunk it into a larger area, R will decrease until, for a sufficiently large area, R goes to zero. Conversely, if you manage to decrease the area by cutting lobes or chunks out from between the points, you can increase R. Therefore, the definition of the area is very important!

It works fine for homogeneous point patterns, in which the points are all relatively dispersed. Interpretation becomes more difficult for heterogeneous patterns (Griffith and Amrhein, p. 131; Burt and Barber, p. 213), in which there are distributions of *clustered* points.

The nearest-neighbour statistic for the Smurf foraging parties is 1.14, indicating that the distribution is pretty close to random (as it should be, since the numbers were all randomly generated!)

ii) Quadrat Analysis

This allows one to generate a statistic which describes the **density of the point distribution**, or the variability in the number of points per cell.

Superimpose a square grid over the area of study. The best (area) size for a square is apparently twice the expected frequency of points in a random distribution (i.e. $2 \cdot \text{area}/n$), so the length of a side is the square root of that. Construct a frequency distribution of the number of points per cell. That is, record how many cells there are with no points in them, 1 point, 2 points, and so on.

Using the formulas for grouped mean and grouped standard deviation, compute the variance to mean ratio (VTMR).

Note the difference from Griffith and Amrhein which incorrectly uses the coefficient of variation! (Because for the Poisson distribution, mean = variance = λ)

Interpretation: If the points are well dispersed, the cell frequencies will be similar, making the variance small and hence the VTMR close to zero. If the points are in clusters, with most cells containing no points and only a few with points, the variance will be large relative to the mean frequency, so the VTMR will be large. If the points are perfectly randomly dispersed, the VTMR will be 1. This is because the Poisson distribution describes the frequency of values for a randomly generated spatial pattern, and its variance is equal to its mean. (We will discuss this later!)

Problems: Both the size of the quadrats and the exact positioning of the quadrat grid will affect the value of the VTMR. Studies have shown that if the grid size is decreased (and hence the number of quadrats is increased), the variance decreases faster than the mean, meaning that the VTMR tends to decrease with decreasing quadrat size. It is **not a consistent statistic**.

The areas being analyzed are seldom perfectly regular, and it is almost inevitable that quadrat squares will overlap the area's boundary. How to deal with boundary squares can be a headache unto itself, and sometimes can affect the result.

C. Spatial Autocorrelation

This may be defined as the relationship among values of a single variable that comes from the geographic arrangement of the areas in which these values occur. Classical statistics assumes no such relationship (i.e. assumes all values are independent). These statistics are a real pain to compute, as you will soon find out!

i) Binary Connectivity Table or Matrix

This is a table which summarizes the connections between each region, i.e. which region borders which.

Label each region on the map. Create a table, whose row and column labels are A, B, C, etc. For each cell in the table, fill in a 1 if the regions in the row and column labels have a boundary, and 0 if they don't. Fill in 0 on the diagonal, since A doesn't have a boundary with A, etc. The more 1's in the table, the more highly connected the regions. Note that the table will always be square. As a matrix, it is denoted by C, with elements c_{ij} . The cell in row i is connected to the cell in column j if $c_{ij}=1$, and isn't if $c_{ij}=0$.

ii) Join Count Statistic

This is used when data is on a **binary** (nominal) scale (or can be degraded to one).

We can regard the connectivity table as an enumeration of all possible pairs of areas. However, since we have duplication, with A being next to B, and B being next to A, when we add up all the 1's in the table we then divide the

number by 2 to find the total number of connections. In other words, $J = 0.5 * \sum_{i=1}^n \sum_{j=1}^n c_{ij}$.

With our binary variable, we can say that an area either has or doesn't have the variable. At the risk of being politically incorrect, the usual jargon is to label areas "with" something as "white" and those without as "black". We record the number of "white" areas as n_1 and "black" areas with n_2 .

We can now look at the map and see how many boundaries there are between two "white" areas (WW), how many there are between two "black" areas (BB), and how many there are between a "white" and a "black" area (BW). The sum $WW+BB+WB=J$ since this is just a way of dividing up the boundaries. The expected values of the three variables are as follows:

$$E(WW) = \frac{Jn_1(n_1 - 1)}{n(n - 1)}, \quad E(BB) = \frac{Jn_2(n_2 - 1)}{n(n - 1)}, \quad E(BW) = \frac{2Jn_1n_2}{n(n - 1)}$$

These are the values we would expect WW, BB, and BW to have if the distribution of the white and black areas was *truly random*. If the actual values are close to the expected values, the distribution is random. Note that the sum of the expected values also equals J.

If the sum of $BB+WW$ is high compared to BW , the distribution is **clustered**, otherwise it is **dispersed**.

iii) Moran Coefficient

This statistic is used when the data are **interval or ratio scale**. It is written in terms of the type of cross-products found in the classical correlation coefficient formula, but instead of two variables X and Y, we use the *pairs of adjoining areas*.

The Moran Coefficient MC is defined as follows:
$$MC = \left(\frac{n}{\sum_{i=1}^n \sum_{j=1}^n c_{ij}} \right) \left(\frac{\sum_{i=1}^n \sum_{j=1}^n c_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

The **first term** is merely the ratio of the number of data point areas to the the total number of connections between the areas. The c_{ij} are the values from the binary connectivity matrix. The sum of all c_{ij} means, as before, to just add up all the 1's in the binary connectivity matrix. n is the number of observations (i.e. areas). Since the connectivity table is square by definition, it has n rows and columns, so the sums over i and j are from 1 to n .

The denominator of the second term is the variance of the data. The numerator of the second term is the cross-product of the deviations of the j 'th and the i 'th observations. Note that the double sum implies that you need to sum over all possible combinations of x_i and x_j , but the c_{ij} term, which is zero for non-connected pairs of regions, means that **the sum is really over all possible pairs of connected regions**.

Interpretation: Values tend to be between -1 and +1, though are not restricted to this range. Values near +1 indicate similar values tend to cluster (positive spatial autocorrelation); values near -1 indicate dissimilar values tend to cluster (negative spatial autocorrelation); values near $-1/(n-1)$ (which goes to 0 as n gets large) indicate values tend to be randomly scattered.

iv) Geary Ratio

This is written in terms of **paired differences**:
$$GR = \left(\frac{n-1}{2 \sum_{i=1}^n \sum_{j=1}^n c_{ij}} \right) \left(\frac{\sum_{i=1}^n \sum_{j=1}^n c_{ij} (x_i - x_j)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

The values of GR tend to be between 0 and 2, but are not restricted to this range. Values approaching 0 indicate similar terms tend to cluster ($x_i - x_j$ tends to be small); values near 2 indicate dissimilar values tend to cluster; values near the expected value of 1 indicate random patterns.

v) Example of Moran Coefficient & Geary Ratio

Schematic of area is below.

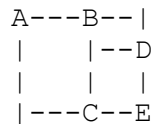


Table 1: Values

Area	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
A	20	6	36
B	16	2	4
C	15	1	1
D	10	-4	16
E	9	-5	25

Table 2: Binary Connectivity Table

	A	B	C	D	E
A	0	1	1	0	0
B	1	0	1	1	0
C	1	1	0	1	1
D	0	1	1	0	1
E	0	0	1	1	0

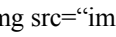
Table 3: Values for $\sum c_{ij} (x_i - \bar{x})(x_j - \bar{x})$ for computing the Moran Coefficient

		$(x_j - \bar{x})$				
$14 = \bar{x}$	$(x_i - \bar{x})$	6	2	1	-4	-5
$5 = n$		6	0	12	6	0
$14 = \sum c_{ij}$		2	12	0	2	-8
$82 = \sum (x_i - \bar{x})^2$		1	6	2	0	-4
$46 = \sum c_{ij}(x_i - \bar{x})(x_j - \bar{x})$		-4	0	-8	-4	0
$0.200 = \text{Moran Coeff}$		-5	0	0	-5	20

Table 4: Values for $\sum c_{ij}(x_i - x_j)^2$ for computing the Geary Ratio

		x_j				
$14 = \bar{x}$	x_i	20	16	15	10	9
$5 = n$		20	0	16	25	0
$14 = \sum c_{ij}$		16	16	0	1	36
$82 = \sum (x_i - \bar{x})^2$		15	25	1	0	25
$280 = \sum c_{ij}(x_i - x_j)^2$		10	0	36	25	0
$0.488 = \text{Geary Ratio}$		9	0	0	36	1

How to Compute the Moran Coefficient and Geary Ratio

1. Construct Table 1, containing the area data and values derived from the differences between the data and the mean. Compute the sum of column 3.
2. Construct Table 2, the binary connectivity table.
3. Construct Table 3. To do this, take the  values from column 2 of Table 1 and put them in the row and column labels of the table. Next, put zeros in the table corresponding to the zeros of Table 2. For the remaining entries, multiply the numbers in the appropriate row (i) and column (j).
4. Construct Table 4. To do this, take the x values from column 1 of Table 1 and put them into the row and column labels of the table. Next, put zeros in the table corresponding to the zeros of Table 2. For the remaining entries, find the square of the differences between the numbers in the appropriate row (x_i) and column (x_j).
5. Compute the Moran Coefficient, using the sum of all the values in Table 3 for the $\sum c_{ij}(x_i - \bar{x})(x_j - \bar{x})$ term.
6. Compute the Geary Ratio, using the sum of all the values in Table 4 for the $\sum c_{ij}(x_i - x_j)^2$ term.

VIII. PROBABILITY

Probability refers to the likelihood or chance that an event will occur.

An **event** is almost any observable phenomenon that can have at least 2 outcomes, such as the toss of a coin, a poker hand, test grade, flood.

P(theoretical outcome) is defined as the number of times the desired outcome can occur divided by the total number of events. This is used when we can use some theoretical way to compute all possible outcomes of an event.

P(empirical outcome) is defined as the number of times the outcome has occurred divided by the number of times the event has occurred. This is used for events which can only be recorded by observation, such as floods, droughts, fires, etc.

P can be stated as a ratio (12/100), a proportion (0.12) or a percent (12%).

The “*law of large numbers*” assumes that as the number of observations increases, the ratio P approaches some absolute or “a priori” probability. Thus a good estimate of an empirical probability requires a large number of observations. This naturally poses difficulties with natural hazards.

The key to probability statements is *uncertainty*. We cannot state for certain the outcome of any single or specific event, but we can state the likelihood an outcome will be observed.

A. Counting Tools

There are 4 different short-cuts we can use to count the number of outcomes of an event.

A meaning for the *factorial*: How many ways can we *arrange* a group of 4 students? In the first position, we can put any one of the 4. In the second position, we can put any one of the 3 remaining. In the third position, we can put any of the 2 remaining. In the last position we can only put the 1 remaining student. The total number of possible arrangements is $4 \times 3 \times 2 \times 1$, which we define as $4!$ and equals 24. You must multiply the combinations!

i) The Multiplicative Rule

Example: If I have 3 bags of rocks, with bag 1 having 6, bag 2 having 4, and bag 3 having 8, and I want to select 1 rock from each bag, there are $6 \times 4 \times 8 = 192$ ways I can do this. This is so because the selection from each bag is made **independently** of the others. If I select rock 1 from bag 1 and rock 1 from bag 2, I can then select any of the 8 rocks from the third bag. Then if I select R1 from B1 and R2 from B2, I can still select any of the 8 rocks from B3. If you work it out, for each of the 4 rocks in B2, I can select the 8 rocks in B3, giving $4 \times 8 = 32$ combinations. These 32 combinations can be repeated for each of the 6 rocks in B1, giving $6 \times 32 = 192$.

In general, if you have k sets of elements, with n_1 in the first set, n_2 in the second, and n_k in the k'th, and you want to select one from each set, there are a total of $n_1 \times n_2 \times \dots \times n_k$ possible samples to draw.

ii) The Permutations Rule

Example: How many ways can Papa Smurf select a group of 5 Smurfs to go foraging for “magic” mushrooms? If there are 200 other Smurfs in the village (the TV show was lying about there only being 100), the first position can be filled by any one of the 200. The second position can be filled by any one of the 199 remaining. And so on. The fifth position can be filled by any of the remaining 196 Smurfs. The total number of permutations for selecting 5 Smurfs is thus $200 \times 199 \times 198 \times 197 \times 196 = 3.04 \times 10^{11}$.

In general, if you have one group of size N, and you wish to select r elements **without replacement** (i.e. all are selected at once) where the **order of selection is important** from that group and arrange them into r positions, there are a total of $(N)(N-1)(N-2)\dots(N-r+1)$ ways to do this. This can be written in a more compact form using factorial notation, as

follows: $P(N, r) = \frac{N!}{(N-r)!}$ Remember that $N! = N(N-1)(N-2)\dots(3)(2)(1)$, so when you cancel out the common

terms from the numerator and denominator, you get the original formula.

iii) The Partitioning Rule

Example: Suppose Papa Smurf has 2 chores he has to get done, one of which requires 2 Smurfs and the other of which requires 3 Smurfs. As it happens, 5 Smurfs are loafing in the village square shooting the breeze and don't see him coming until it is too late to hide. How many ways can Papa assign the two chores to the 5 Smurfs?

If we initially treat the 5 available positions as separate, (i.e. abcde) there are $5 \times 4 \times 3 \times 2 \times 1 = 120$ ways to assign the Smurfs to the jobs. However, 3 of the jobs are identical, and the other 2 are also identical, so we are effectively reducing the number of unique arrangements to aaabb. The first 3 elements can be arranged in $3!$ ways, and the last 2 in $2!$ ways, so in effect we are losing $3! \times 2!$ permutations. To find how many remain, we divide $5!$ by $(3! \times 2!)$ to get 10.

In general, if we have a set of N elements that is *partitioned* into k subsets, each with n_k elements, there are

$\frac{N!}{n_1! n_2! \dots n_k!}$ distinct ways to do this.

iv) The Combinations Rule

This is a special case of the partitioning rule, in which the set of N elements is divided into 2 groups. We normally look at it from the perspective of taking a sample of size r **without replacement** from the population of size N, **in which the order of the sample elements is not important**. This means that there are N-r elements which are not sampled. Thus

by the partitioning rule, $C(N, r) = \binom{N}{r} = \frac{N!}{r!(N-r)!}$. This is read as “N choose r”, a short form for “how many ways

can we choose r elements from a population of N” (without replacement and order not important). For example, how many different combinations of 6 numbers can be chosen from 49 (as in the Lotto 6/49)? $C(49,6) = 49!/(43!6!) = 13,983,816$. Another example which combines the combination and multiplication rules follows from this one: how many ways can you select 4 of the 6 correct numbers from the Lotto 6/49? There are $C(6,4)$ ways to choose 4 of the 6 correct numbers, and $C(43,2)$ ways to choose 2 wrong numbers from the remaining 43. They must be multiplied for the total number of possibilities: $6!/(2!4!) * 43!/(2!41!) = 15*903 = 13545$. Go through all combinations, from 6 right to 0 right, and you’ll find out that there’s only a 2% chance of winning **any** money (i.e. getting 3 or more numbers right)! If any person tried something like this, it would be called a numbers racket and the person would be arrested. But it’s the government, so it’s called a lottery!

v) Sampling with Replacement

The above 4 rules assume that the sampling of the population is done without replacement. That is, when a sample is taken from the population, the items are taken one at a time, but are not replaced in the population after selection. Thus an element can only be selected once. With replacement, an element can be selected more than once. Suppose we have a population of size 4, and we want to enumerate all possible samples of size 2 that we can get by drawing an element, recording it, and then putting it back. In this case, order is important, since we are looking at an arrangement. For the first selection, we can draw any one of the 4 elements. But since we put the element back, for the second choice we can also draw any of the 4 elements. This gives us a possible $4^2=16$ samples. For samples of size 3, there would be $4^3=64$ possible samples.

In general, the number of possible arrangements (order important) of size r that can be drawn from a population of size N with replacement is N^r .

When we talk about *combinations*, i.e. order is unimportant, things get much trickier, because we have to look at the composition of each possible sample. For instance, a sample of size 5 taken with replacement could be of the form aaaaa, aaaab, aaabb, aaabc, aabbc, aabcd, or abcde. Since the order is not important, we sort any arrangement into one of these patterns! There is no easy rule for counting these things up. Because the probability of each sample composition (i.e. aabbc, aaabb) may very well be different, and because of the complexity, this is not frequently used! So don’t worry, nobody’s going to ask you to compute one!

B. Introduction to Set Theory

The use of sets and set theory is another way of looking at probability problems, especially when we’re looking at probabilities related to combinations of two or more events. For the most part, set theory appears more difficult than it really is because it is shrouded in a smokescreen of strange symbols and jargon. All you have to do is remember that a **set** is a *collection of objects*. In this case, the objects are samples of size n drawn from a population.

i) Terminology

Set theory problems can be graphically represented by the use of a **Venn diagram**. It consists of a rectangle, representing the sample space (the set of all possible samples of size n), and one or more inscribed circles, representing various groups of selected samples or elements (sets).

Consider two overlapping circles, one labelled A, the other B. The **union** of two sets consists of the area (representing elements or samples) enclosed by *both*. It is denoted $A \cup B$. If a sample is contained in $A \cup B$ it is in either A or B. The word *or* is used to denote a *union*. The area in which the two circles overlap is called the **intersection** of the two sets. It is denoted by $A \cap B$. If a sample is contained in $A \cap B$, it is in both A and B. The word *and* is used to denote an *intersection*. If the two circles do not intersect at all, they are **mutually exclusive** or **disjoint**. The “intersection” is the **empty set**, denoted by \emptyset . The part of the sample space **outside** a given circle, say A, is the **complement** of A, denoted by A^c . If A is a **subset** of B, all of the elements of set A are contained *within* set B. It is denoted by $A \subset B$.

ii) Example: Throwing 2 Dice

Throwing two 4-sided dice is equivalent of taking a sample of size 2 from a population of size 4, consisting of the integers 1 to 4, with replacement (since you can throw doubles), and the order important (if we have 2 distinct dice, i.e. red and white).

Define 4 sets:

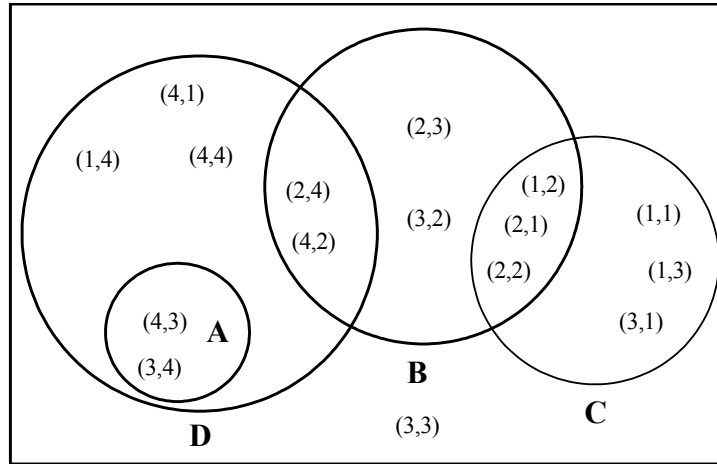
A (all samples whose sum is 7): $\{(3,4) (4,3)\}$;

B (all samples with a 2): $\{(1,2) (2,1) (2,2) (2,3) (2,4) (3,2) (4,2)\}$;

C (all samples whose sum is 4 or less): $\{(1,1) (1,2) (1,3) (2,1) (2,2) (3,1)\}$;

D (all samples with a 4 in them): $\{(1,4) (2,4) (3,4) (4,1) (4,2) (4,3) (4,4)\}$

Note that (3,3) is left out of these 4 sets. The Venn Diagram is displayed below. You can see from the diagram that $A \subset D$, $B \cap D = \{(2,4)(4,2)\}$, $B \cap C = \{(1,2), (2,1), (2,2)\}$, $C \cap D = \emptyset$.



Venn diagram for example

iii) Application to Probability

Since probability refers to *relative frequencies*, we can read off the probability of getting a certain sample. For instance, since 7 of the 16 samples have the number 4 (set D) in them, $P(D)=7/16=0.44$. $P(C)$ and $P(B)$ also are $7/16$. Since 3 samples are in C and B, $P(B \cap C) = 3/16=0.19$. $P(A')=14/16=0.88$. And so on.

iv) Mathematical Set Rules

The Venn diagram can also be used to illustrate mathematical rules of sets. $\subset \cap \cup$

1) Commutative Rule: $A \cap B = B \cap A$; $A \cup B = B \cup A$.

2) Associative Rule: $A \cup (B \cap C) = (A \cup B) \cap C$; $A \cap (B \cup C) = (A \cap B) \cup C$.

3) Distributive Rule: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$; $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

(Sorry, no pictures available)

v) Probability Rules using Sets

Set theory can take a lot of the pain out of computing compound probabilities. Inspection of simple Venn diagrams will reveal the following probability rules:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \text{ and } P(A') = 1 - P(A) \text{ and } P(A \cap B') = P(A) - P(A \cap B)$$

Note that in the first rule, the last term is subtracted because it occurs **twice**, in both A and B.

C. Conditional Probability

The probabilities we have discussed before are **unconditional**, because no special conditions have been assumed aside from the ones which defined the experiment.

If we have additional knowledge about the event, we can alter the probability of its occurrence based upon that knowledge. For instance, we know that the probability of observing an even number on the toss of a die (call it event A) is $3/6 = 1/2$. Suppose that we are told that on a particular throw, the number which came up was ≤ 3 (call this event B).

We now no longer have reason to believe that the probability of an even number given that the number is ≤ 3 , is $1/2$, since of the numbers 1,2,3 only 2 is even. Instead, it has become $1/3$. Our extra information (event B) has changed our value of the probability by *reducing the sample space*, i.e. possible outcomes, by imposing a condition on the event A.

We use the symbol $P(A|B)$ to refer to “The probability that A occurs given that B occurs”. We define it as:

$P(A|B) = \frac{P(A \cap B)}{P(B)}$. This formula adjusts the probability of $A \cap B$ from its original value in the complete sample space (set of outcomes) to a conditional value in the reduced sample space B (outcomes which fit into result defined by B).

Similarly, we can write $P(B|A) = \frac{P(B \cap A)}{P(A)}$. The first can be rewritten as $P(A \cap B) = P(A|B) * P(B)$. The second can be written as $P(B \cap A) = P(B|A) * P(A)$. But since $P(A \cap B) = P(B \cap A)$, we can write $P(A \cap B) = P(A|B) * P(B) = P(B|A) * P(A)$.

This is the **multiplication rule**. It states that if we know $P(B|A)$ and $P(A)$, or $P(A|B)$ and $P(B)$, we can find $P(A \text{ and } B)$. If $P(A|B)=P(A)$, then the probability of A is *unconditional*. Similarly, if $P(B|A)=P(B)$, then the probability of B is unconditional. This means that the occurrence of A does not influence B, and vice versa. In other words, A and B are *statistically independent*. If A and B are independent, $P(A \cap B) = P(A) * P(B)$. This follows from the re-written equations: $P(A \cap B) = P(A|B)P(B) = P(A)P(B)$, since $P(A|B)=P(A)$. This is an extremely important property!

D. Example using Contingency Table

Let’s go back to the table used in the χ^2 example, where we looked at political parties and alcohol consumption.

Political Party		PC	Liberal	NDP	Reform	Totals
Preferred Beverage	None	4 (6.3)	7 (6.4)	7 (6.2)	10 (9.1)	28
	Beer	8 (16.4)	16 (16.8)	22 (16.1)	27 (23.7)	73
	Wine	21 (10.6)	10 (10.8)	6 (10.3)	10 (15.3)	47
	Spirits	12 (11.7)	13 (12.0)	9 (11.4)	18 (16.9)	52
	Totals	45	46	44	65	200

Dividing each row total by 200 (n) gives the probability of selecting a person who prefers a certain type of beverage. Dividing each column total by 200 gives the probability of selecting a person who prefers a certain political party. Dividing each element of the table by the appropriate row or column total will produce two **conditional probabilities**. Dividing by row total will give $P(\text{party}|\text{beverage})$. Dividing by column total will give $P(\text{beverage}|\text{party})$.

Why is this? Go back to the conditional probability formula. A given cell entry / n gives a $P(\text{Party and bev})$. A given row total / n gives $P(\text{bev})$. Hence, $P(\text{party and bev}) / P(\text{bev}) = P(\text{party}|\text{bev})$. Dividing the cell entry by the row total will give the same results as dividing the probabilities, because the n’s will cancel out (and you can save some work).

Now we have another way to test for independence of the two variables. For example, let’s check NDP supporters and wine drinkers. Does $P(\text{NDP \& Wine}) = P(\text{NDP}|\text{Wine})P(\text{Wine}) = (6/47)(47/200) = 6/200$ in fact equal $P(\text{NDP})P(\text{Wine}) = (44/200)(47/200)$? Well, they don’t, so NDP and wine drinkers are not independent, and hence party affiliation and booze consumption are not independent. This agrees with our previous finding using the χ^2 statistic. If the numbers in the table had been distributed with their expected values, we would have found that $P(\text{Party \& bev}) = P(\text{Party}|\text{bev})P(\text{bev})$. By dividing the row and columns sums by n, we have generated 2 separate relative frequency distributions. They are called **marginal probabilities** because they appear in the margins of the tables.

E. Bayes’ Theorem

So far we have been looking at *prior* probabilities, which are assignments of numbers to events before the samples are selected. We now look at *posterior* probabilities, in which we’ve drawn the sample and look back to figure the probability it was drawn in a certain way. In essence, we will be trying to compute a conditional probability when we are given information involving at least one other conditional probability. This is how to recognize a Bayes’ Theorem question.

i) Example 1: The False Positive Paradox

(stolen from the Cartoon Guide to Statistics!). Suppose there is a good, but not perfect, test for a disease which affects 1 Smurf in 200. If a Smurf has the disease, the test is positive 99% of the time. However, it returns false positives (Smurf not sick) 2% of the time. If Hypochondriac Smurf tests positive, what is the chance he has the disease?

We have two events to examine: B Hypochondriac Smurf has the disease, A he tests positive. The information we have been given is: $P(B)=0.005$ (1 Smurf in 200 has it); $P(A|B)=0.99$ (positive test given sick Smurf); $P(A|B')=0.02$ (positive test given healthy [i.e. not sick] Smurf). We want to know $P(B|A)$, the chance that the Smurf has the disease given a positive test.

The best way to solve one of these problems is to construct a table of probabilities.

Test \ Health	Sick Smurf (B)	Well Smurf (B')	Sum
Pos Test (A)	$P(B \& A)$ [0.00495]	$P(B' \& A)$ [0.0199]	$P(A)$ [0.02485]
Neg Test (A')	$P(B \& A')$ {0.00005}	$P(B' \& A')$ {0.9751}	$P(A')$ {0.97515}
Sum	$P(B)$ [0.005]	$P(B')$ [0.995]	1

Numbers in [] are given; numbers in { } are computed.

Let's fill in the table entries that we know or can calculate. First, we were given $P(B)=0.005$, so we can compute $P(B')$ immediately as being $1-0.005$ or 0.995 . Next, we know $P(B \& A) = P(A \& B) = P(A|B)P(B) = 0.99*0.005 = 0.00495$. $P(B' \& A) = P(A \& B') = P(A|B')P(B') = 0.02*0.995 = 0.0199$.

Using the multiplicative rule, $P(B|A)=P(B \& A)/P(A) = 0.00495/0.02485 = 0.199$. In this particular case, it wasn't necessary to construct the entire table, but it never hurts to do so!

What this example shows is that if Hypochondriac Smurf tests positive, there is really only a 20% chance (1 in 5) that he is sick! At least we have increased his chances of diagnosis from 1 in 200 to 1 in 5. Note that the false positives come from the much larger uninfected group.

Simple Bayes' Theorem:
$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B')P(B')}$$

What does this mean? The formula was derived right from the table. $P(A|B)P(B)$ is just $P(B \& A)$, and the denominator is just $P(A)$, expressed as the sum of the probabilities in the row.

ii) Example 2: Urn and Ball problems

These are old standbys that always appear when you least want them to (such as on an exam). (Note that an urn is a ceramic container used in the old days to hold liquids like wine or oil.) A typical problem has a number of urns each containing different numbers of coloured balls. If you draw a coloured ball, what is the probability you got it from a particular urn. However, they can come in other forms, such as this: Suppose that 3 guys share a house somewhere, and that the first guy up makes the coffee so they can get started on another hard day's studying. Joey is first up 30% of the time, Danny 50%, and Eddie 20% of the time. Joey, Danny and Eddie make lousy coffee 10%, 15%, and 20% of the time. If the coffee is lousy, what is the probability of its having been made by Danny? Joey? Eddie?

Problems of this sort are best done by drawing a probability tree. First, let's label the event Joey makes coffee J, Eddie E, and Danny D, and let L be the event lousy coffee (L'=good brew). We are asked to find $P(D|L)$, $P(J|L)$, and $P(E|L)$. L is the prior event, i.e. the "given", and it forms the first set of branches of the tree. The others are the conditional probabilities (i.e. prob of lousy given Danny); they go in the second level.

Prior probs	cond probs	events	Joint probs
--P (D) = 0.5	*D ---P (L D) = 0.15	D&L	P (D&L) = 0.5*0.15=0.075
	--P (L' D) = 0.85	D&L'	
* --P (J) = 0.3	*J ---P (L J) = 0.10	J&L	P (J&L) = 0.3*0.10=0.03
	--P (L' J) = 0.90	J&L'	
--P (E) = 0.2	*E ---P (L E) = 0.20	E&L	P (E&L) = 0.2*0.2=0.04
	--P (L' E) = 0.80	E&L'	

The probability of lousy coffee, $P(L)=P(D\&L)+P(J\&L)+P(E\&L)=0.145$. To answer the question,

$$P(D|L)=P(D\&L)/P(L)=0.075/0.145=0.517.$$

$$P(J|L)=0.03/0.145=0.207.$$

$$P(E|L)=0.04/0.145=0.276.$$

The **general** Bayes' Theorem is:
$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A|B_1)P(B_1) + P(A|B_2) + \dots + P(A|B_n)}$$

You get $P(A)$ by summing all the joint probabilities which involve A .

IX. Some Theoretical Frequency Distributions

There exist many different ideal probability distributions that come up when different things are studied. Only 4 of them are covered in the text, since they are the ones most likely to occur in a geographical context.

Some Terms

A **random variable** is variable whose observed values are determined by chance. We can usually determine the probability that a random variable will assume a certain value. Random variables are denoted by a capital letter (e.g. X), while outcomes are denoted by a lower-case letter, often with a subscript attached to denote the observation number (e.g. x_i).

A **discrete random variable** has a countable number of outcomes. For instance, the result of die rolls, poker hands, coin tosses. The number may be large, but it is not infinite, so that the probability of an event will be (usually) greater than zero.

A **continuous random variable** can assume any value within one or more ranges. For instance, a textbook's weight can be anything between 0.5 lbs and about 15 lbs, to as many decimal places as you want.

Because there are an infinite number of possible values in the range(s), the probability of the variable equalling a particular value is **zero!** Probabilities for continuous variables are expressed as $P(a \leq X \leq b)$, the chance the variable's value falls within the range $[a, b]$.

A **probability distribution** for a random variable refers to the relative number of times we would expect to get each value of a random variable, based on a very large number of outcomes. All distributions must follow these 2 rules: **a)** the sum of all the probabilities must be 1, and **b)** the probability that the variable assumes a certain value [written as $P(X = x_i)$] must be ≥ 0 .

The probability distribution of a continuous random variable is a *continuous distribution*, while that of a discrete random variable is a *discrete distribution*. The equation which generates the characteristic curve of a continuous distribution is called the **probability density function**.

A **finite distribution** is one in which the range of values that can occur in it is limited, i.e. between some a and b . The values of an **infinite** distribution are unlimited, i.e. from $-\infty$ to $+\infty$ (or $-\infty$ to b or a to $+\infty$).

The **mean of a probability distribution** can be found by multiplying each value in the distribution by the probability of its occurrence, and adding all the results together. We use μ to denote the mean because it can be thought of as the mean value of the variable found from a very large number of repetitions of the experiment which produces the distribution. A mathematical term for the mean is the *expected value*, but typical to mathematics the word "expected" doesn't have its regular meaning. It is written $E(X)$.

Definition: The mean of a **discrete** random variable is $\mu = E(X) = \sum_{i=1}^n x_i p(x_i)$. Note that we usually use the short form $p(x_i)$ instead of $P(X = x_i)$. For a continuous distribution, we have to use an integral expression, similar to the summation expression.

Definition: The population variance of a discrete random variable is the average of the squared distance between x_i and

$$\mu, \text{ i.e. } \sigma^2 = E[(x_i - \mu)^2] = \left[\sum_{i=1}^n (x_i - \mu)^2 p(x_i) \right]$$

The **standard deviation** of a random variable is defined as usual, i.e. the square root of the variance.

You will not be expected to compute these yourself, but I thought you'd at least like to know where the formulas come from.

A. The Uniform Distribution

This is probably the simplest of the distributions. Each value of the variable occurs with equal probability for a discrete distribution. The probability density function is a constant over the range of values.

It is a *finite* distribution, and can be discrete or continuous.

When used: This shows up when you have sampling with replacement, since each element has an equal probability of being chosen. Also, if you have no idea how a variable is distributed, the usual initial assumption is uniformly. For instance, the location of a short in a 5-m wire could have an equal probability of being in any given 1-cm length of wire.

Continuous: $f(x)=1/(b-a), a \leq x \leq b; =0$ otherwise. $\mu = \frac{b+a}{2}$ and $\sigma^2 = \frac{(b-a)^2}{12}$

Discrete: If the range $[a,b]$ is divided into k subintervals, $p(x_i) = 1/k$. If the range $[a,b]$ is a set of integers, $k=(b-a+1)$.

$$\mu = \frac{b(b+1) - a(a-1)}{2k} \text{ and } \sigma^2 = \frac{(b-a)(b-a+2)}{12} = \frac{(k-1)(k+1)}{12}.$$

B. The Binomial Distribution

When used: This distribution occurs when the event can have only 2 outcomes (hit or miss, male or female, heads or tails, etc). It is a *finite, discrete* distribution.

Binomial variables have certain characteristics:

1. The sample consists of n independent outcomes.
2. Only 2 outcomes are possible: "Success" and "Failure".
3. $P(\text{success})$ is the same from outcome to outcome, defined as p .
4. The binomial variable itself is the number of successes in the sample.
5. Note that the order of successes and failures is not important.

Definition: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

The factor $\binom{n}{x}$ represents the number of combinations of x successes and $n-x$ failures. This is because the binomial variable is equivalent to drawing x successes and $n-x$ failures without replacement from a population, with order not important.

The other terms come from the multiplication rule for independent events.

Example: Aside from hating Smurfs, Gargamel also hates Christmas. By November 15, street-corner Santas were infesting the city. He decided to make his anti-Christmas statement (disguised as Papa Smurf of course) with his paint pellet gun. However, he'd had a little too much "medicine" the night before, limiting his probability of hitting a Santa to 0.3 (he can't figure out which of the two Santas he sees is the real one!). Suppose he gets the chance to shoot at 5 of them before the authorities close in and he has to use his teleport spell to get away. What is $P(\text{he nails } 2)$? 3?

The problem has defined $n=5$, and $p=0.3$. We want $P(X=2)$ and $P(X=3)$. Substitute into the formula: $P(X=2) = C(5,2) (0.3)^2(0.7)^3 = 10(0.09)(0.343) = 0.3087$. $P(X=3) = 10(0.3)^3(0.7)^2 = 0.1323$. As you can see, the probability drops sharply from $P(X=2)$ to $P(X=3)$. The shape of the distribution depends on p and n , with it being symmetric for $p=0.5$ and skewed otherwise.

If I wanted to find Gargamel's chance of bagging 3 or more Santas, I'd have to sum $P(X=3) + P(X=4) + P(X=5)$. This could get cumbersome if I had $n=100$ and wanted $P(X \geq 40)$, so tables are available, or certain approximations could be made using other distributions.

If p is small and n is large, and $np \leq 7$, the Poisson distribution gives a good approximation to the binomial. If $10 \leq n \leq 25$ and $p=0.5$, or $n \geq 25$ and $0.2 \leq p \leq 0.8$, the normal distribution gives a good approximation.

Mean: $\mu = np$ **Variance:** $\sigma^2 = np(1-p)$

Griffith and Amrhein, pp. 173-174 gives an example of another use for the distribution, as a way to test the representativeness of a sample.

C. The Poisson Distribution

When used: Primarily in describing the number of events that will occur in a *specific period of time, length, area or volume*. **Examples:** The number of traffic accidents per month at an intersection, number of surface defects on a new car, number of diseased trees in an acre of woodland, people in lineups, thunderstorm occurrences in an area, and in **quadrat analysis**, which was discussed earlier.

Characteristics

1. The number of times an event occurs during a given period of time, or given area, volume, distance or some other unit of measurement) is counted.
2. The probability that an event occurs in a given unit of time, distance, etc. is the same for all units.
3. The number of events that occur in a given unit of time, area, distance, etc. is independent of the number that occurs in other units.
4. The mean or expected number of events in each unit is denoted by the Greek letter λ . It is either known from past experience (for instance, a hospital would have records of the average number of people coming to emergency per hour), or for a binomial approximation $\lambda = np$ (i.e. the mean of the binomial distribution).

Definition: $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$

Example: Hailstorms in a Midwest county. In the American Midwest, hailstorms are a constant threat during the summer. Over a 35-year period, there were 10 years without storms, 12 with one storm, 9 with 2, 3 with 3, and 1 with 4. The observed probability of a year with 0 storms is thus $10/35$, 1 storm $12/35$, etc. Since 43 storms occurred, the temporal average is $43/35=1.23$ storms per year.

To find out if the pattern is indeed random, we need to compare the observed probabilities and/or frequencies with the Poisson distribution of expected hailstorms per year. We use the observed temporal average of 1.23 for λ , and substitute the values of 0, 1, ..., 5 into the Poisson formula.

Storms per year	Obs Freq of Years	Total Storms	Obs Prob of Years	Poisson Prob	Expctd Freq (years)
0	10	0	.285	.292	10.2
1	12	12	.343	.360	12.6
2	9	18	.257	.222	7.8
3	3	9	.086	.091	3.2
4	1	4	.029	.028	1.0
5+	0	0	.000	.007	0.2
TOTAL	43	1.00	1.000	35.0	35.0

Mean: $\mu = \lambda$ **Variance:** $\sigma^2 = \lambda$

The mean should be λ , of course, since that's how we defined the distribution. The interesting part is that the variance also equals λ . Beware of the error in the Griffith and Amrhein text, p. 177, which defines the coefficient of variation σ/μ as 1 for a Poisson distribution. It can't be so, according to the definitions!

D. The Normal Distribution

This is probably the most important of all the probability distributions. Among other things, people's weights, heights and shoe sizes are normally distributed, as are annual rainfall and temperatures of a region, IQ scores, test scores, and most natural phenomena in general. Many more variables can be approximated very well by normal distributions.

Measurement errors are also normally distributed, and this is in fact how the distribution was initially discovered. In the 1700's there was a brilliant German mathematician named Gauss. One night he was in an observatory with some students charting star locations. He got annoyed at one whose measurements were inconsistent, but to his surprise found that he couldn't get the same measurement twice either. When he sat down to look at the results, he found that the errors fell into a bell-shaped pattern. The curve is also called "Gaussian" after its discoverer.

It is a *continuous infinite* distribution. It is defined in terms of its mean and standard deviation by the following

equation:
$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right]$$
, where, as usual, μ is the mean and σ is the standard deviation.

i) The Standard Normal Distribution

The shape of the distribution for any normal random variable depends on both μ and σ . Since these can take any combination of values, it is obviously impractical to compute tables for them all. However, it is possible to mathematically transform any normal distribution to a standard one with a mean of 0 and a std dev of 1. We use the **z-score** for this: $z = \frac{x - \mu}{\sigma}$ or $z = \frac{x - \bar{x}}{s}$. Recall that earlier in the course, the z-score was defined as the number of

standard deviations away from the mean the value x is. This is still the case. Because there is no closed form solution for the integral of the equation, values have to be computed numerically and are presented in standard tables, including the one on p. 458 of Griffith and Amrhein.

ii) Using the table

Using the table is easy. The numbers represent the area under the curve from 0 to z , where z is positive. Suppose you want to see what $P(0 \leq z \leq 1.45)$ is. Look down the leftmost column of the table until you see 1.4. Then move across the row until you are in the 0.0500 column. The value 0.4265 is the result we seek. Notice how the numbers in the table increase from top to bottom and from left to right.

iii) More advanced examples

It is always beneficial to draw a sketch of what you are trying to do to avoid confusion!

$P(0 \leq z \leq a)$: Use the table value.

$P(z \leq a)$: This is $0.5 +$ table value. Since the curve is symmetric, and since the area under it is by definition 1, the area from $-\infty$ to 0, which is half the area under the curve, is 0.5.

$P(z \geq a)$: This is $0.5 -$ table value. Since we know the table value is the area from 0 to a , and the total area from 0 to infinity is 0.5, we find the unknown area by subtraction

$P(z \leq -a)$: Here we exploit the symmetry of the curve. $P(z \leq -a)$ is exactly the same as $P(z \geq a)$, i.e. $0.5 -$ table value for $z=a$.

$P(-a \leq z \leq 0)$: Again using symmetry, this equals the table value for $P(0 \leq z \leq a)$.

$P(z \geq -a)$: Again using symmetry, this equals $P(z \leq -a)$ or $0.5 +$ table value.

$P(a \leq z \leq b)$: This equals $P(z \leq b) - P(z \leq a)$, where the latter two probabilities are found using the above methods.

Example: Al Bundy has to keep enough shoes in stock to make a success of his shoe store (ha-ha). Let's say that the length of a woman's foot is a normal random variable with mean of 20 cm and a standard deviation of 5 cm. If, on average, he sells 300 pairs of shoes of a certain popular style a week, how many of these pairs will be size 25? 18?

Between 25 and 18? Note that a shoe “size” for this problem is designed to fit a foot of length “size” plus or minus 0.5 cm.

- a) First we want to find $P(\text{size } 25) = P(24.5 \leq \text{length} \leq 25.5)$. When we convert the two numbers to z-scores, we get $P(0.9 \leq z \leq 1.1) = .3643 - .3159 = .0484$. To **convert to an expected number**, multiply $p \cdot n$: $0.0484 \cdot 300 = 14.52$ or 15.
- b) Next we want to find $P(\text{size } 18) = P(17.5 \leq \text{length} \leq 18.5)$. Again converting to z-scores we get $P(-0.5 \leq z \leq -0.3)$. Using the symmetry of the curve, this is the same as $P(0.3 \leq z \leq 0.5) = .1915 - .1179 = .0736$, and the expected value is $p \cdot n = .0736 \cdot 300 = 22.08$ or 22.
- c) Here we want to find $P(\text{size } 18 \leq \text{size} \leq \text{size } 25)$. The absolute minimum of foot size in this range is thus 17.5 (the smallest foot size 18 will fit) and the maximum foot size is 25.5 (the largest foot a size 25 will fit). So the real problem is $P(17.5 \leq \text{length} \leq 25.5)$. As z-scores, it is $P(-0.5 \leq z \leq 1.1)$.

Because this involves both a positive and negative z-score, we must derive the result step by step. First, $P(-0.5 \leq z \leq 1.1) = P(z \leq 1.1) - P(z \leq -0.5)$. From above, we know $P(z \leq 1.1) = 0.5 + P(0 \leq z \leq 1.1)$, the table value. Also from above, we know $P(z \leq -0.5) = 0.5 - P(0 \leq z \leq 0.5)$, the table value. Now we substitute these results into the first to get $[0.5 + P(0 \leq z \leq 1.1)] - [0.5 - P(0 \leq z \leq 0.5)]$ to get the formula $P(-0.5 \leq z \leq 1.1) = P(0 \leq z \leq 1.1) + P(0 \leq z \leq .5) = .3643 + .1915 = .5558$. The number of pairs of shoes Al can expect to sell is thus $0.5558 \cdot 300 = 166.74$ or 167.

E. Doing problems

Like with general probabilities, the hardest part of doing any problem involving probability distributions is identifying which one to use to solve the problem. Sometimes the problem’s wording will include a phrase like “variable xxx follows a yyy distribution”, in which case you don’t have to worry about it!

You may find it helpful to remember that *binomial* variables involve counting the number of successes and failures, *Poisson* variables involve counting the occurrences of something in space or time, and **normal** variables describe the variation of some natural quantity. Once the distribution has been determined, the rest is just mechanical, substituting the appropriate values into the formula and/or looking the number up in a table to find the result.

It is worth noting that the area under the normal curve from -1 to +1 is .6413; -1.5 to +1.5 is .8664; and -2 to +2 is .9544. In other words, about 64% of observations of a normal random variable should occur within 1 standard deviation of the mean; about 87% should be within 1.5 std. devs, and about 95% should be within 2 std. devs of the mean. Hence, the further away from the mean an observation is, the less likely it is to have occurred, and the more significant it becomes.

F. How to Bell Grades

The underlying assumption is that marks are normally distributed. Some departments, whose names shall not be mentioned, are more fanatical than others in determining that class marks should have a predetermined mean and standard deviation. If these arbitrary demands aren’t met, or if an exam has an unexpectedly poor (or good!) showing, the marks can be adjusted from the original distribution to a new, more satisfactory distribution, a process commonly referred to as “belling” or “marking on the curve”.

Doing this is actually a simple procedure. First you compute the mean and standard deviation of the actual grades in the usual way. Next, compute the z-scores for each mark in the usual way. Finally, given the z-score and the desired mean μ_2 and standard deviation σ_2 , the new grade can be computed using $x'_i = \sigma_2 z_i + \mu_2$. To do it all in one step, use

$$x'_i = \frac{\sigma_2}{\sigma_1} (x_i - \mu_1) + \mu_2. \text{ Note how this formula can be written in the same form as to the z-score formula:}$$

$$x'_i = \frac{x_i - \mu_1}{\sigma_1} \sigma_2 + \mu_2, \text{ where } \sigma' = \frac{\sigma_2}{\sigma_1} \text{ and } \mu' = \mu_1 - \sigma_1 \mu_2. \text{ Instead of converting the observation to the usual}$$

standardization (mean of 0 and standard deviation of 1), we are converting it to an **alternative standardization** with mean μ' and standard deviation of σ' .

X. SAMPLING DESIGNS

As was mentioned near the beginning of the year, statistics serves two purposes, to describe the characteristics of a data set, and to make inferences about a population based upon a sample drawn from it. Having discussed descriptive

statistics, we are now ready to ready to enter the great and glorious world of inferential statistics, which we will be spending the rest of the year reviewing. The most important aspect of inferential statistics is the **selection of the sample**. If you don't have a reliable way to get a representative sample, all the statistics in the world won't help you derive any useful information from it. In the real world, the best way to design an experiment and collect data is often not obvious and considerable time must be spent looking at all the various factors which may try to foul things up.

Important factors in the sampling design include identification of the population you're sampling, ensuring representativeness and randomness, the sampling perspective (i.e. replacement, order important) and framework you use, and the likelihood of the resulting statistics. Geographical statistics have their own unique problems, most of which are spatially related and difficult to deal with.

A. Geographic Populations

A **statistical population** is a set of numerical values corresponding to measures taken from a parent population. If geographic references are attached to the values, we will have a statistical geographic population. An example is the average income for each county in Canada. Defining the population is in itself a challenging task. For instance, suppose you want to find the origins of the students here at U of T. How do you define "student"? How do you define "origin", as place of birth, or something based on ethnic group? Other examples: Study residential patterns of the illiterate. How do you find illiterate people? How do you define illiterate? Mobility patterns of the elderly: what types of trips do you study?

Another way to define a geographical population is the way in which the region is divided up into units, and using the observed map pattern as a sample. The observed spatial pattern can be regarded as one out of all possible spatial configurations that could be formed by the given data. Consider how Metro is divided up into political ridings, on the municipal, provincial and federal levels. If I'm not mistaken, each riding is designed in such a way as to hold approximately the same number of people. As you can probably imagine, there are a lot of ways you can do this, and the current one is just one of them. Spatial units such as ridings, census tracts, postal codes and so on are known as *Modifiable Area Units* because their shape and size are essentially arbitrary.

As long as the phenomenon being studied is not responsible in some way for the distribution of the boundaries, then the map pattern can be reasonably assumed to be a random partitioning of the surface and using it to define a population is fine. An example of where the phenomenon is responsible for the pattern is "gerrymandering", the (illegal) art of redefining electoral districts to include as many of the (usually incumbent) candidate's supporters as possible.

Another way of looking at it is the *randomization outlook*, in which the number of different ways the numerical values can be assigned to the areal units is examined. The parent population consists of $n!$ possible map patterns, all of which are considered equally likely; the current one is one of these patterns.

There is also a third way of viewing the existence of samples from a geographic population. Social, economic, physical and/or other forces will create a spacial distribution of people over a region. For example, the building of the railway played a major role in defining population patterns in the prairie provinces. (See, for example, Pierre Berton's *The Last Spike*.) However, other factors which operate independently of each other and the primary forces will act to modify the distribution. For instance, a farmer on his way to his new farm might step in a gopher hole and break his leg, and wind up selling the property to a neighbour, thus changing the local distribution of farms.

In this case, the geographic population is the spatial distribution without the noise component.

B. Sample Size and Sampling Error

As you should recall from earlier in the year, *statistical error* is introduced when computations (statistics) are based upon a subset of the population (a sample), rather than the entire population. The object of inferential statistics is to try to deduce information about the population in the presence of statistical error, while taking it into account. As you should remember, the definition of a **consistent statistic** is one whose sampling distribution concentrates at and around the value of the corresponding population parameter. As the sample size n approaches N , the sampling distribution should get more concentrated.

The main question of interest is of course, what is the minimum size of n which will get the best tradeoff between maximizing the information content of the statistic and minimizing sampling error, while keeping external forces like cost of doing the survey at a reasonable level? This is not easily answered, because the optimal value will depend on the underlying population frequency distribution. This is discussed at great length in Griffith and Amrhein, pp. 202-208.

The conclusion drawn is that the minimum value for what is considered a “large” sample size is between 30 and 100, with larger values required for more skewed population distributions.

The text example is a demonstration of the *central limit theorem*, which states that if a random sample of size n is drawn from any population, the sampling distribution of \bar{x} will be approximately normal for a “large enough” n . The bigger the n , the better the approximation. The more skewed the population distribution, the larger n must be.

C. Random Sampling

In order to make a meaningful inference about a population, it must be defined carefully, and the sample must then be selected randomly in some way. Unfortunately, however, randomness often conflicts with the desire for representativeness, because a truly random sample may not necessarily be representative. For example, most of the points randomly selected from an area could be clustered in one corner.

Selection of random numbers is not difficult. Many scientific calculators now have buttons which will produce a uniform pseudo-random number between 0 and .999 (uniform meaning that each number has an equal chance of being generated). Spreadsheets and most computer languages also have pseudo-random number generators. The term “pseudo-random” is used because an algorithm is used to generate the sequence of numbers. If the same number generator “seed” is fed to the algorithm, it will generate the same sequence of numbers. You can stretch the range of random numbers from [0.000,0.999] to [a,b] by performing the following operation: $X1=(b-a)*X+a$. That is, multiply by the range and add the starting value. If no calculator is handy, use the random number table. Start anywhere in the table (not necessarily on an edge!) and take any path you like, so long as it doesn’t backtrack or otherwise intersect with itself. You can select strings of more than one digit, too. To convert to a range [a,b], take a series of 3-digit numbers, pretend each is a 3-digit decimal, and use the above transformation.

Note that by using a “truly” random process to select the areas, we are sampling with replacement, since with random numbers it is theoretically possible to select the same point more than once.

D. Representative Samples and Sampling Frameworks

As mentioned above, just because a sample is randomly selected does not mean that it is going to be representative. The very nature of randomness means that conceivably all of the points could be in only one part of the area of interest, like the upper left corner.

There are ways to ensure greater representativeness, which will be discussed below. All of these ways involve adjusting the probabilities of selecting each of the points in order to reduce the chances of selecting an unrepresentative sample. However, all of these ways restrict the number of ways in which samples can be drawn, which in turn reduces the number and types of meaningful statistics which can be computed. If the unequal probabilities that are used are known (as from the methods used below), the results can be calculated to take them into account. If they are not known, you can’t make any sort of inference from the statistics because you don’t have any idea of the sampling error. You always have to know how the data was collected in order to make valid inferences!

i) Systematic Selection

In this method, we select elements methodically, without replacement. For a finite data set with N elements, we can always number the elements 1 to N . We then divide the data set into N/n sub-units, pick a number k between 1 and N/n , and pick the k ’th element from each of the sub units. We will thus be picking the elements $(k, k+N/n, k+2N/n, \dots, k+(n-1)N/n)$. The sample size n is restricted in that it must divide evenly into N . If it doesn’t, not all elements will have an equal probability of being selected. If there are 14 points to choose from, the last 2 points here will only be selected if $k=1$ or 2. Systematic selection is not so good if the data happens to have a periodic trend equal to N/n . (Consider a sinusoid where we sample at same point on each wave.) After each sample of size n is selected, the ordering of the data should be scrambled up before the next sample is selected, to preserve randomness.

A **geographical area** will be divided evenly in both x and y directions, with widths dx and dy , though the number of divisions need not be equal. A point (x,y) will be selected at random in the first cell, and the other points will be taken from the same locations in the other cells, i.e. at $(x+k*dx, y+h*dy)$. The sample size n will be restricted by the resulting grid imposed on the area, since to be systematic, all points in the grid must be sampled.

ii) Stratification

Another alternative to random sampling is to *stratify* the population. In other words, the data are divided up into a relatively small number of mutually exclusive, collectively exhaustive categories (strata), and then each category is sampled randomly. In a **geographic area**, these strata may be lowlands and highlands, or river valley, grasslands, and

forest, or urban, suburban and rural. With a group of people, one might stratify by occupation, age, ethnic background, religion, or belief in Smurfs. The primary advantage is that you are ensuring that each subset will be represented in the sample, which increases the sample's chances of being more representative. By reducing the variability of the sample, you are reducing the sampling error as well.

For example, suppose I have an area that is divided into 40 units and I want to sample 11 of them. The area is organized into 5 regions as follows: R1 has 6 units, R2 has 7, R3 has 12, R4 has 10, R5 has 5. Employing selection with replacement (i.e. "totally random"), there are 40^{11} or 4.1943×10^{17} possible samples of size 11.

In a stratified sample, I might pick 2 from R1, 2 from R2, 3 from R3, 3 from R4 and 1 from R5 to get a sample of size 11. Again using a "totally random" scheme, there are only $62 \times 72 \times 123 \times 103 \times 51 = 1.5421 \times 10^{10}$. In other words, according to this stratification scheme, only $3.6 \times 10^{-6}\%$ of the possible random samples are considered to be representative!

Stratification may be **proportional**, when the same number of samples is taken from each stratum, or **disproportional**. The latter is best used when there are significant differences between the sizes and or homogeneity of the subpopulations in the strata, and/or when there is a significant difference in cost collecting data between the subregions. For example, you can expect it to be cheaper to collect data on plant diversity from a field than from a swamp. You can also get away with collecting fewer points from a very homogeneous stratum, since a smaller sample is almost guaranteed to be as representative as a larger one. One must beware of course that a stratum may be very homogeneous with respect to one variable, but heterogeneous with respect to another. For instance, a group of people with the similar heights can have widely varying weights.

iii) Cluster Sampling

In stratified sampling, the population is divided into groups (strata) and we sample from every stratum. In **cluster sampling**, we divide the population into groups called *clusters* and sample among the clusters. The primary reason for this is to reduce cost in data collection, especially for surveying people. The easiest example is dividing a city into census tracts. In *cluster* sampling, we would select some of the census tracts to collect lots of data from (perhaps sampling each person or area). In *stratified* sampling, we would sample a few points from each tract. Stratified sampling works best if the regions are as homogeneous as possible so that relatively few points can represent each region well. Sampling error will arise primarily from variability *within the strata*. Cluster sampling works best if the regions are as heterogeneous as possible. Sampling error will occur because of variability *between clusters*, since each element in the cluster (ideally) would be sampled.

For the majority of the case, cluster sampling will produce greater sampling error (less efficient) than random or stratified sampling. Consider a census tract type survey of incomes. Cluster sampling could give rather skewed results if, say, all of your selected tracts happened to be in swishy areas of town or in areas with lots of subsidized housing projects.

E. Geographic Sampling

Spatially distributed samples seek to provide information about the geographic distribution of the variable being studied. Any sample that is selected should allow us to infer information about its parent population. As usual, the spatial aspect makes everything a lot more complicated. Griffith and Amrhein, p. 215, illustrates several of the techniques.

Random spatial samples are not likely to be as representative as we would be like. It can be shown that selecting the x and y coordinates at from a uniform random distribution will produce a Poisson distribution of points, which is not spatially uniform. It may also be difficult to locate a point randomly selected from a map when you are in the field (especially if it is in the middle of a swamp or dense bush).

Systematic sampling can be done over the area, sampling points on a grid. As with regular systematic sampling, if the data are periodic in some way then systematic sampling may give an unrepresentative sample. As with the regular systematic sampling, the number of points in the sample will be constrained by the grid size. It is apparently better to use squares than rectangles.

Random stratified sampling can be done by dividing the area into different regions and randomly sampling from each region. The regions could be based on natural features such as landforms, or be totally artificial, such as quadrats. The distribution of points in each quadrat would be Poisson, but at least you'd be guaranteed a sampling from the entire area.

Cluster sampling can be done either of two ways, both of which involve the use of a square as the cluster. In the first way, a square grid is superimposed on the area and squares (clusters) are selected at random. In this case, the size of the

squares and locations of the grid lines are issues. In the second way, the centres of the squares and their orientations are selected in some manner so that overlapping is possible. Location can be selected at random or in some stratified random way.

Transects are a useful way to collect data in an area. All you do is collect data from a bunch of points in a line (or several lines) which cross the area. This is the best way to sample things like vegetation or rock types in a previously unexplored area. Transect lines can be chosen with the barest minimum of information and still yield statistically good results. The ability to control the orientation, placement and point spacing may improve the representativeness of a sample from an area which has a high degree of spacial autocorrelation. An area can always be stratified and transect lines drawn through the regions, and/or the lines can be drawn in a regular pattern. Of course, the orientation, length, and number of traverse lines, as well as the frequency and spacing of points along the lines all have to be selected, and carefully. Random traverse lines will tend to have a Poisson distribution, and as usual may not give a representative sample. Lines that are drawn in corners are not as useful as those through the middle. For these reasons, placement and orientation of lines are usually nonrandom. Selecting the points along the transect is a whole different kettle of fish.

F. Some other problems

Aside from other drawbacks mentioned above, here is another. The presence of positive spatial autocorrelation (i.e. similar values tending to cluster) means that there is less information in the geographic data than one might expect, since there is a pattern. In this case, systematic point sampling is the best option, followed by stratified random sampling or systematic traverses.

XI. Concepts for Inferential Statistics

As has been mentioned previously, one of the objectives of statistics is to be able to make a statement about a population parameter once the corresponding sample statistic has been calculated. This is an example of **inductive** reasoning: given one sample, we ask what was the random system that generated its statistics?

Inferential statistics comes in two flavours, **estimation** and **hypothesis testing**. In *estimation*, we use our sample to compute a guess, also known by the more politically correct term of “statistic”, as well as an idea of how close we think it is to the parameter. *Hypothesis testing* involves checking to see if a claim about the true value of a parameter is valid or not.

A. Estimation: Large-Sample Estimate of a Population Mean

Another term for statistic is **point estimate**, since we are estimating the parameter value. A **point estimator** is the

mathematical way we compute the point estimate. For instance, $\frac{1}{n} \sum_{i=1}^n x_i$ is the point estimator used to compute the

estimate of the population mean \bar{x} . Because of sampling error, we know that it’s not likely that our sample statistic will be equal to the population parameter, but instead will fall into an interval of values. We will have to be satisfied knowing that the statistic is “close to” the parameter. That leads to the obvious question, what is “close”? We can phrase the latter question differently: How **confident** can we be that the value of the statistic falls within a certain “distance” of the parameter? Or, what is the probability that the parameter’s value is within a certain range of the statistic’s value? This range is the **confidence interval**, which is the *probability* that the value of the parameter falls within the range specified by the confidence interval surrounding the statistic.

Example of Confidence Interval: Suppose that Gargamel’s friend Amy Surplus gives a warranty that her shotguns shoot pellets that will be within 15 cm of the intended target 95% of the time at a range of 10 m. In other words, if Gargamel fires at a paper target’s bull’s-eye 10 m away, and if he were to draw circles of 15 cm radius around each hole, 95% of those circles will include the bull’s-eye *in the long run*. The exact number will vary, of course, from sample to sample (i.e. shot to shot). Another example is on p. 225 of Griffith and Amrhein.

Recall the *Central Limit Theorem*, which applies to the sampling distribution of the mean of a sample. Consider samples of size n drawn from a population, whose mean is μ and standard deviation is σ with replacement and order important. The population can have any frequency distribution. The sampling distribution of \bar{x} will have a mean $\mu_{\bar{x}} = \mu$ and a standard deviation $\sigma_{\bar{x}} = \sigma / \sqrt{n}$, and *approaches a normal distribution as n gets large*. This allows us to use the normal distribution curve for computing confidence intervals.

The unit of measurement of the confidence interval is the **standard error**. This is just the standard deviation of the sampling distribution of the statistic.

GGR 270Y First Term Lecture Notes

So how do we go about computing a confidence interval? The first step is to find how many standard errors away from the mean we need to go (on both sides of the mean) in order to find the area under the normal curve which is equal to the value of our confidence level. The z which fills this criterion is called $z_{\alpha/2}$. For example, for a 95% confidence interval, we want to find out which z will give us $P(-z_{\alpha/2} \leq z \leq z_{\alpha/2}) = 0.95$.

Where does the $z_{\alpha/2}$ come from? In order to confuse the layman, statisticians don't refer to the area under the normal curve they want, but instead refer to the **area that is left out by the confidence interval**. When we want a 95% confidence level, they say we want to ignore 5% of the area under the curve. This ignored area is the α . Furthermore, the ignored area is under the tails of the curve, so half of it is assigned to be under each tail, hence the $\alpha/2$. Basically, when we think "95% confidence level", the statistician thinks "100(1-0.05)% confidence level", or in general 100(1- α)%.

To find we use the table to find the z whose value is such that $P(0 \leq z \leq z_{\alpha/2}) = 0.5 - \alpha/2$. Remember, we're *subtracting* the area under the tail from the whole area from 0 to infinity. Look in the table until you see a number equal or close to $0.5 - \alpha/2$, then find the z corresponding to it. To save you from doing this all the time, here are the z values corresponding to the most commonly used confidence levels.

100(1- α)%	α	$\alpha/2$	$z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.58

Now that we know how many standard deviations away from the mean we have to go to get our confidence interval, we can compute it using the formula $\left[\left(\mu - z_{\alpha/2} \sigma / \sqrt{n} \right) \leq z \leq \left(\mu + z_{\alpha/2} \sigma / \sqrt{n} \right) \right]$ if we know the population parameters (for instance, from a distribution problem), or $\left[\left(\bar{x} - z_{\alpha/2} s / \sqrt{n} \right) \leq z \leq \left(\bar{x} + z_{\alpha/2} s / \sqrt{n} \right) \right]$ if we are estimating the parameters. The term $z_{\alpha/2} \sigma / \sqrt{n}$ is also called the **margin of error**.

Important note! There is an "official" name for what we are doing here: *Large-sample estimation of a population mean*. We are computing an estimate and confidence interval for a population mean, and the Central Limit Theorem allows us to use the normal distribution table to compute the CI.

Example

Bob wants to get elected as Sesame Street's Dictator-for-Life. Of the 1000 voters polled, 550 claimed they would vote for him. Construct the 90%, 95%, and 99% confidence intervals for p , the probability he gets elected.

The election result for Bob is a binomial random variable, with $n=1$ --either the voter votes for him or doesn't in the one election. The poll is trying to estimate the true probability the person will vote for Bob, p . Our estimate for p is $550/1000=0.550$. The standard deviation of the sampling distribution (standard error) is $\sigma / \sqrt{n} = \sqrt{(1)p(1-p)} / \sqrt{n}$, which equals $\sqrt{.55(.45)} / \sqrt{1000}$. **NOTE: The n in the binomial distribution's σ is different from the sample size n !**

To compute the confidence interval for each confidence level, multiply this number by the appropriate $z_{\alpha/2}$.

CI	$z_{\alpha/2}$	Mrg Err	Lower	Upper
90%	1.645	.026	.524	.576
95%	1.96	.031	.519	.581
99%	2.58	.041	.509	.591

As you can see, the greater the confidence level, the wider the confidence interval will be. In the long run, we can expect the probability of a vote for Bob to be in the first range 90% of the time, in the second range 95% of the time, and in the third range 99% of the time. (Not that the polls really matter, since Sesame St. elections are won by the number of ballot boxes you can misappropriate and stuff.)

Doing these large-sample estimates of the mean problems is easy. All you need is the sample mean and standard deviation, or some way to get them.

B. Determining a Minimum Sample Size

As can be seen by the formula, the standard error will decrease as n increases. If we decide to fix the margin of error at some “upper limit”, given the confidence level and the standard deviation we can compute the n that will give this upper limit. $n = \left(z_{\alpha/2} \sigma / ME \right)^2$ where ME is the required margin of error.

Example

A hospital studied the records of 100 patients to determine the average length of stay. It was found to be 4.65 days, with an s of 4.9 days. The margin of error for a 95% confidence interval is $1.96 * 4.9 / 100.5$ or .96 days. Suppose we want to estimate μ to within 0.25 days. How many records would have to be reviewed? (Note: with a large sample size, we can safely approximate σ with s .)

Margin of error is .25, $s = 4.9$, and $z_{\alpha/2} = 1.96$. Substituting these values into the formula gives $n = 1475.8$ or 1476. As you can see, the inverse square root dependence means that you have to get increasingly large sample sizes in order to get smaller margins of error.

C. Small-Sample Estimation of the Mean

More often than not, we won't have the luxury of a large sample, i.e. $n < 30$, or even $n < 100$ for some underlying distributions. We can no longer assume the sampling distribution of \bar{x} is approximately normal, since the Central Limit Theorem assumes normality only for large samples. Not only that, but s becomes an increasingly poorer approximation for σ as n decreases.

We can dodge around the first problem with the knowledge that if the population distribution is normal or nearly normal, the sampling distribution of \bar{x} will also be normal or nearly normal. We can dodge the second problem by using the **student t distribution** instead of the normal distribution.

Student's t distribution is the sampling distribution of the **t statistic** $t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$ when we are sampling from a normal

distribution. It is actually a family of distributions, with the shape of each determined by the *degrees of freedom*, which are determined by $v = n - 1$. Using the degrees of freedom allows us to take the greater variability of the samples due to their small size into account. See Griffith and Amrhein, p. 231 for more details. As the number of degrees of freedom increases, the t distribution approaches the normal distribution.

Constructing a confidence interval for a small-sample estimate of a mean is almost the same as for a large-sample estimate, except that you use $t_{\alpha/2, v}$ instead of $z_{\alpha/2}$ and you use s instead of σ , which is usually unknown. In other words, a small sample CI is constructed by $\left[\bar{x} - t_{\alpha/2, v} s / \sqrt{n}, \bar{x} + t_{\alpha/2, v} s / \sqrt{n} \right]$.

So how do you find a t value? In Griffith and Amrhein, p. 459, there is a table of t values, with degrees of freedom listed on the left margin and probabilities across the top margin. Some tables list the α value (i.e. the area under the tail from t to infinity) in the top margin, but this table lists the $1 - \alpha$ values (i.e. the area under the curve from minus infinity to t).

Example

A drug company is testing a new drug which is supposed to reduce blood pressure. From the six people who are used as subjects, it is found that the average drop in blood pressure is 2.28 points, with an s of .95 points. What is the 95% confidence interval for the mean change in pressure?

For this problem, $n=6$, so there are 5 degrees of freedom. $\alpha=.05$, so $\alpha/2=.025$. In the table we look up $t(.975, 5)$ (since the text uses $1 - \alpha$) by looking down the .975 column until the 5 df row is reached. This t is 2.571. The confidence interval is thus $2.28 \pm 2.571 * .95 / \text{sqrt}(6)$, or [1.28, 3.28]. That is, we can be 95% confident that the mean decrease in blood pressure is between 1.28 and 3.28 points. In other words, the testing procedure will produce an interval that contains the true mean 95% of the time.

Note that because we had a small sample, we had to use the t value 2.571 to form the confidence interval instead of the smaller z value of 1.96. The greater variability due to small sample size requires that we have a larger confidence interval. If we want to narrow the interval, we can always try to get a larger sample.

Another reason for small samples may be due to the need for destructive sampling techniques. For example, testing the average lifetime of appliances so that the manufacturers can set the warranties to expire just before they self-destruct.

D. Introduction to Hypothesis Testing

As was previously mentioned, evaluating (or testing) a claim (or *hypothesis*), about the true value of a parameter is called **hypothesis testing**. The hypothesis we are testing is called the **null hypothesis** and is denoted by H_0 . The use of the word “null” suggests the sample statistic we are testing is **not different from the parameter value**, there is no relationship, no improvement, etc.

The statistic is **evaluated under the assumption that the null hypothesis is true**, kind of like a trial: “innocent until proven guilty”.

A second hypothesis about the value, called the **alternate hypothesis**, denoted by H_a , must also be included with the null hypothesis. The two hypotheses must be mutually exclusive: the null hypothesis says there is no difference, while the alternate hypothesis says there is difference. Rejection of the null hypothesis implies acceptance of the alternate.

i) Test Statistics

To test the hypothesis, you need to construct a test statistic. It is frequently in a form similar to $z=(x-\mu)/\text{std err}$, but each test (and there are a lot of them!) has its own special test statistic and distribution with which to test it for you to memorize. (Stop whining. This is good for you. Really.)

The null hypothesis, with its definition, allows one to construct a sampling distribution that can be used to test it. This distribution is divided into two parts. Test statistics that fall within what I call the **acceptance region**, which corresponds to the confidence interval, give no reason to disprove H_0 . The probability of this occurrence is $1-\alpha$.

Convincing evidence in favour of H_a will exist when the sample statistic exceeds the hypothesized value by an amount that cannot be readily attributed to sample variability. These test statistics fall within the critical or rejection region, are not consistent with H_0 , and therefore consistent with H_a , and occur in the tail region (or regions for a two-tailed test).

ii) Types of Tests

The probability associated with the critical region is called the **level of significance**, α . This is the *probability of rejecting the null hypothesis*. The usual values are 10%, 5% and 1%. The **critical value** is the *value which marks the boundary of the rejection region*.

A **right-tailed test** will be used in a situation when we are testing if the parameter is *greater than* some value. For example, $H_a: \mu > 2000$. The level of significance will be $\alpha = \alpha_2$. A **left-tailed test** will be used in a situation when we are testing if the parameter is *less than* some value. For example, $H_a: \mu < 2000$. The level of significance will be $\alpha = \alpha_1$. A **two-tailed test** will be used when we are trying to show that the parameter is either larger or smaller than a given value. For example, $H_a: \mu \neq 2000$. In this case, the level of significance must be divided equally among the two tails of the distribution, since we must have a rejection region in *both directions*. Hence, it will be $\alpha/2$.

iii) Errors in Hypothesis Testing

Because our hypotheses divide all possible sample outcomes into 2 groups, or populations, any sample we may draw will fall in either the population described by the null hypothesis or that of the alternate hypothesis. However, sample variability is such that we may occasionally get a sample which belongs to the population of H_0 but which is quite *unlikely* to occur, and hence falls into the rejection region.

Both of the examples to follow are examples of Large-Sample Hypothesis test of a population mean.

Type I Error

Because the sample falls into the rejection region, we reject H_0 even though the sample really belongs to that population. This is what is called a **Type I error**. The probability of this occurring is just α , the size of the rejection region or level of significance.

Example of Type I error: We know that the the mean value you get when throwing 2 dice is 7. Suppose I threw 2 dice 40 times and got 10 nines, 10 tens, 10 elevens, and 10 twelves, with $\bar{x} = 10.5$ and $s = 1.132$. This is a relatively large sample drawn from a population (of sums of die rolls) which is essentially normal, so we can use the normal distribution to approximate the sampling distribution of the sample mean.

We can define $H_0: \mu = 7$. Our sample suggests an $H_a: \alpha > 7$. The test statistic is $z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{10.5 - 7}{1.32/\sqrt{40}} = 19.6$.

The critical value for a 5% level of significance in a right-tailed test is $z_\alpha = 1.645$ (i.e. the value of z which gives $P(0 \leq z \leq z_\alpha) = 0.95$). It is obvious that 19.6 is way beyond the critical value, and this would lead us to accept H_a and reject H_0 and create a Type I error.

Type II Error

Similarly, if we draw a sample which belongs to H_a , but appears to occur as H_0 , we would believe it belonged to H_0 even though it didn't. This type of error is called a **Type II error** and is usually denoted by β . It is not at all straightforward to compute, as the examples below will show.

Example of Type II error: Back to the dice. Suppose I hypothesize that the mean of two dice is 8. I throw 25 times and get 10 sevens, 10 eights, and 5 nines, with $\bar{x} = 7.8$, $s = 0.764$. I define my experiment as $H_0: \mu = 8$ and $H_a: \mu < 8$. Again, since the population is essentially normal, I can get away with using z even with this small a sample size. The z statistic is -1.31, which is greater (i.e. more positive) than the critical value of $z_\alpha = -1.645$. This means that although our mean is less than the hypothesized one, it is not significantly so, so we must accept H_0 , even though it is false, and thus create a Type II error.

The whole problem with errors of course is that we can't know for sure if we've made one or not. We can only compute the probability of committing them and hope that we don't!

iv) The Power of a Test

The **power** of a test is the probability that the test will correctly lead to a rejection of H_0 for a particular value of μ in H_a . It is equal to $1 - \beta$ for the particular alternative considered. The greater the power, the better the test.

Example 1: Variation of Power with μ

Suppose that building specifications for a city require that the strength of residential sewer pipe be more than 2400 pounds per foot of length. Suppose 50 sections of pipe are tested and found to have a mean strength of 2425 pounds/foot with an s of 200. Use $\alpha = .05$.

From the point of view of the city, the strength of a supplier's pipe is assumed to be less than its requirements until proven otherwise. Thus, $H_0: \mu < 2400$ and $H_a: \mu > 2400$ are the null and alternate hypotheses. To compute the power, we must find β the probability we accept H_0 (i.e. we think the pipe is too weak) when in fact it is false (i.e. it meets the specs). So how do we do this?

Find the z_α corresponding to the significance level. Since we are doing an upper-tailed test with $\alpha = .05$, $z = 1.645$.

Find the value of \bar{x} which corresponds to the critical z value in the null distribution (mean of 2400): i.e.

$\bar{x}_0 = \mu_0 + z_\alpha \sigma_{\bar{x}} = 2400 + 1.645(\sigma / \sqrt{n})$. This is the largest value of \bar{x} that supports the null hypothesis. With the large sample size we can approximate σ with s . Using $s = 200$ and $n = 50$, we get $\bar{x} = 2446.5$.

Now we find the z statistic for this \bar{x} in the **alternate distribution**, which has a mean of 2425 (the value we found) and the same standard error as the null distribution. This is $z = (2446.5 - 2425)/(200/\sqrt{50}) = 0.76$.

Finally, we find the area of the acceptance region under the alternate distribution. Remember that since the rejection region for H_0 is for values **greater than** the critical value for its distribution, $z_0 = 1.645$, the acceptance region for H_a is for values **less than** the critical value for its distribution, $z_a = .76$. This area is .7764.

With $\beta = .7764$, the power is .2236. This means that the test will lead to the *correct rejection of the null hypothesis* only 22% of the time if the specifications are exceeded by only 25 pounds/foot.

If the sample mean had been 2450 instead of 2425, β would have been .4522 and the power would have increased to .5478. If 2475, $\beta = .1562$ and power = .8438. Thus, if the mean strength of pipe exceeds the standard by 75 pounds per foot, the chance of correctly rejecting H_0 increases to 84%.

Conclusion: Power *increases* as the distance between the null and sample means increases.

Example 2: Variation of power with α

GGR 270Y First Term Lecture Notes

Suppose Papa Smurf is testing the effects of an improved version of a certain white powder he has created in his lab. His 100 Smurf “volunteers” reacted an average of 1.2 seconds after the old stuff was inhaled through the nose. Suppose that the reaction time of the 100 “volunteers” with the new powder is 1.1 seconds, with an s of 0.5 seconds. He wants to find out if this represents a significant departure from the the old powder. Find β and the power for significance levels $\alpha = .05$ and $.01$.

We can see that $H_0: \mu = 1.2$ (assume no change), and $H_a: \mu \neq 1.2$. This is a **two-tailed** test, since if μ isn't equal to 1.2, it is either greater than 1.2 or less than 1.2. For the rejection region corresponding to $\alpha = .05$, $z_0 = 1.96$.

As before, find the values of \bar{x} corresponding to the critical values. $\bar{x}_0 = \mu_0 \pm 1.96\left(s/\sqrt{n}\right) = 1.2 \pm 1.96\left(.5/\sqrt{100}\right)$ and the values are thus 1.102 and 1.298. Remember, these are the values of \bar{x} that *support* H_0 .

Now find the z values in the alternate distribution, where $\mu_a = 1.1$. Lower bound: $z(a,L) = (1.102-1.1)/(.5/10) = .04$. Upper bound: $z(a,U) = (1.298-1.1)/(.05) = 3.96$.

Rejection region for H_0 is $z < -z_0$ or $z > +z_0$. Hence the *acceptance* region for H_a is $z(a,L) \leq z \leq z(a,U)$. This is equal to $.5 - .0160 = .4840$ (the $.5$ is because the area from 0 to 3.96 is essentially $.5$).

Hence, $\beta = .4840$ and power is $.5160$. Thus, the test will lead to a Type II error 48% of the time for an α of $.05$.

For $\alpha = .01$, $z_0 = \pm 2.575$, bounds on \bar{x} are 1.0712 and 1.3288, new z values with $\mu_a = 1.1$ are $-.58$ and 4.58 , and $\beta = .7190$, power = $.2810$ (check these yourself!).

Conclusion: As α is decreased, so is the power of the test.

So even though you are decreasing your chances of incorrectly rejecting H_0 by reducing α , you are simultaneously decreasing your chances of correctly accepting H_0 for a given alternative! Thus, *your level of significance must be selected carefully, depending on the problem at hand.*

This is the end of the lecture notes

I sincerely hope that you found them useful. Please let me know your opinion of them, and if there are any corrections or clarifications to make.

Harold Reynolds