Raphael "The School of Athens" (1509-1510?), Apostolic Palace, Vatican City

Methods for Obtaining Knowledge: Scientific Method

Raphael "Disputation of the Holy Sacrament" or perhaps "The Triumph of the Church" (1509-1510?), Apostolic Palace, Vatican City

Introduction to Experimental Psychology

Scientific Method

Epistemology (Study of knowledge)

Philosophers have identified at least three methods by which we can gain knowledge ("truth"). These are:

- Divine (non-physical) insight. We gain knowledge through "communication" with a higher, nonphysical being. Presumably this higher being knows much more than mortal humans. We then must have belief or "faith" in this divine truth.
- Pure logic and thought (Aristotle). If we are logical enough in our reasoning, we should be able to deduce all knowledge.
- Scientific manipulation. This is the only method of gaining knowledge that we shall study in this course. A biologist might want to know if water affects plant growth. The biologist then manipulates a variable, the amount of water that is given to a specific plant. The biologist *varies* (thus the word "variable") the amount of water given to the plant. He/she gives the plant more or less water. He/she then observes the *effect* of this variable (the amount of water that is given) on plant growth. The result might well be that plants that are given more water grow higher. We now have *knowledge*. We *know* that water will *cause* plants to grow.

Materialism versus Idealism

For many centuries, philosophers and scientists debated about the nature of human experience (consciousness). A pure materialist assumes that all that exists must exist in some physical form. This material existence is subject to the laws of the physical universe. There is thus no room in this model of the universe for non-material (non-physical) existence. A pure materialist is thus an atheist. He or she does not believe in any higher non-physical, existence, in god(s) or in the non-physical soul/mind. By contrast, a pure idealist does not believe in the physical reality of our existence. This is best exemplified by the Greek philosopher, Plato. How do we know that physical reality exists at all? How do I know that I am not simply experiencing dream-like existence? It was not until several centuries after Plato that a compromise solution was made by the French philosopher and mathematician, Descartes, who found room for a nonphysical existence (the mind or soul) and physical existence (the body). The laws of science operate *only* on a physical, material existence. Nevertheless, philosophers point out that we do have concepts such as love, hate, free will and in psychology we have concepts such as the mind, consciousness-unconsciousness. Are these entirely physical in nature?

Scientific Process

- *Observation of the universe*. There are many things we do not understand, for which we do not have an explanation. So…
- *Define a problem*. What is the problem that needs explanation? What questions are you asking? (What is the controversy?) Let's take an example from psychology. We know that a drug, alcohol, affects performance on many tasks. Why does it affect performance on these tasks? You might *speculate* (or have an "opinion" more about this later) that alcohol mainly affects tasks that are very complex. Simple tasks are thus spared. Even if this were true, it does not explain why alcohol affects complex tasks more than simple ones. Furthermore, you need to define what you

mean by a "complex" and an "easy" task. What is complex about a complex task? As you can see, we now have many problems that need to be solved.

- *Search the literature.* Go to the library. Read the scientific periodicals (called "journals"). Search the internet. Has anyone already asked the same questions? Many scientists have already done many, many studies on the effects of alcohol. While you might speculate that alcohol only affects complex tasks, you would be wrong. The fact is that we *know* (that word again) that alcohol may not affect all complex tasks. Speaking is a very complex cognitive function. Yet we manage to speak and to understand speech quite well under the influence of alcohol. Driving a car is also a relatively complex task. It takes several months (if not years) to master this task. Yet under the influence of alcohol, we seem to manage quite well. We can (unfortunately) get into the car, turn on the ignition, put the car in gear, accelerate, turn, navigate. Identifying red lights should be an easy task. Very young children can, after all, discriminate red from green. But, under the influence of alcohol, we sometimes do not do this task well. We sometimes miss the red light.
- *Development of theory*… a summary (synthesis of what we already know). Based on a complete review of existing studies (the "literature"), you might arrive at the conclusion that alcohol appears to mainly affect tasks that require short-term memory. Tasks that require long-term (or permanent) memory are not affected. Now you have developed a *theory* based on a thorough review of the literature. This is no longer speculation.
- Based on theory, scientists form an *hypothesis* (or prediction). In the example above, you would hypothesize that alcohol affects any task that requires short-term memory. It should not however affect tasks that require long-term memory.
- *Testing of hypothesis* (unbiased, objective). This involves the design of a study to answer the questions & resolve the controversy. Obviously you would need to include in your "design" certain tasks that require the use of short-term memory and others that do not.
- *Data collection.* Run the study.
- *Analyses of the data.* Does alcohol cause a *variation* in performance? Do the results *differ* (or vary) depending on which task was run?
- *Interpretation of the results.* Was the hypothesis supported by the results? Are there alternate explanations (theories) that might also explain the results?

Theories/Hypotheses

- Theory - predicts behaviour or events. Theories are formed after a review and synthesis of the relevant literature. It is based on existing *fact.*
- Theory vs speculation (opinions). A theory is only formed after an objective review and synthesis of already published research. It is not formed on the basis of guesswork or subjective opinion. We can all speculate about why things are the way they are. The problem with this is that your opinion is as good as mine. Which opinion is correct? Again, theory is based on fact. You may have an opinion that humans can get by on 2 hours of sleep. Your theory is wrong. It is a fact that humans require more sleep (a subject we will study in this course) and if they do not get, will suffer grave consequences.
- All hypotheses must be testable. (See next section). We must be able to design a study to test the hypotheses. Hypotheses make predictions. The results of the study will allow us to say whether your theory is true or false.
- A GOOD theory is one that potentially can be proven to be WRONG. There is no point in simply stating that alcohol affects performance. Even if I find that it does not affect certain tasks, you can always claim that I was not using the right tasks (the ones that would be affected by alcohol).

- Replication - Others should be able to exactly replicate what you have done and what you have found. This is why science is said to be "objective" and "universal". If we repeat the experiment anywhere in the universe, the results should be the same. We thus obtain "universal" knowledge.

Hypothesis testing

There are many, many problems that remain unsolved. Different theories can, however, be developed, each of which might potentially explain a fact. Fact: men perform slightly better in math than women (to balance matters, women perform slightly better than men on verbal tasks). One theory to explain why men do better on math is that it is a result of evolution, and more specifically, hormonal differences. Another theory, however, maintains that the differences can be explained on the basis of social learning. Women learn that math skills are not admired. Men learn that they are admired. Both theories cannot be correct. We thus have a *controversy*.

- Scientists state an hypothesis. This is an expectation based on previous scientific research (fact). "If theory X is correct, we would expect that…" On the other hand, if theory Y is correct, we would expect that…"
- The hypothesis must be stated in a specific enough manner that we can prove it to be wrong
- Define variables of interest. *Operational definition*. If you claim alcohol affects performance, how do you define "performance"? If your hypothesis is that alcohol causes you to "feel good", how will you define "feeling good"?
- Means of measuring (quantifying) variables of interest. This is often an extremely difficult step in Psychology. Recall that science operates on the basis of material existence. There must be more or less of it. This implies that you must be able to *quantify* your variable of interest. You need to attach a number to it (to "measure" it). Logically, a variable must "vary". Even if alcohol does make one feel good, how would you measure "feelings" and in particular, "feeling good"?
- According to many philosophers of science, we should assume that all hypotheses are false unless proven otherwise. This is what is meant by the *null* hypothesis. The onus is thus on the researcher to prove their theory true. In this sense, science is very cynical and pessimistic. We assume god(s) does (do) not exist until proven otherwise. We assume there are no aliens on other planets. We assume there are no tigers roaming the campus of the University of Ottawa. If we come upon a tiger, then we have positive evidence. Our null hypothesis is wrong. If we do not find any tigers, it can always be claimed that we have not looked in the right places. In the example of alcohol, we assume that it does not affect performance. If it does, the null hypothesis is wrong. (Incidentally, this is exactly the same logic that we use in British-based "common law" legal systems. We assume that a suspect is not guilty… The *null* hypothesis is that the suspect is *not* guilty. Note we do not assume that they are innocent… it is philosophically difficult (if not impossible) to prove innocence. The onus is on the government (the courts) to prove the suspect to be guilty. There is enough positive evidence to judge the null hypothesis… the assumption that the suspect is not guilty… to be wrong. If there is not enough evidence to find the suspect guilty, they are not judged to be "innocent"; rather they are judged to be "not guilty".

Problems in Psychology

Often many of the questions we ask are exceedingly difficult to test. Frequently, we cannot easily define psychological concepts. We may not agree on a way of measuring the concept. Let's use the example of aggression. We know that certain children are more aggressive than others. As good scientists we need to ask the question, why? Why are some children more aggressive while other children are less aggressive? What is the cause of aggression? Perhaps it is because the very aggressive children have inherited a gene for aggressiveness. But perhaps certain children have also learned to become very aggressive. We go to the library/internet and do a search. We find that there are wide differences across the world. In particular, countries that do not censor very violent computer games, television and movies (the "action thrillers") appear to be much more violent than those that do. A "learning-modelling" theory claims that aggression is learned and indeed it is rewarded (reinforced). Behaviour that is reinforced is repeated. Those children that watch the most violence on television (or play the most violent computer games) tend to become the most aggressive. Another theory, a "genetic" or "biological" theory might however claim that the children that are the most violent are so because they have inherited a violent or aggressive gene.

- Theory: Watching violence on television *causes* children to become more aggressive. Biology has nothing to do with it.
- Alternate theory: It is biology (too much activity in certain "aggression" areas of the brain) that cause certain children (and boys in particular) to be aggressive. Learning has nothing to do with it.
- Both theories cannot be correct.
- *Operational definition*: Define *variables* of interest: Children? Violence? Aggression?
- Manipulate extent of violence to which children are exposed in the experiment? Presumably in the learning-modelling theory is true, then children who are exposed to more violence will be more aggressive than children who are not exposed to violence. How will actually do this manipulation?
- The manipulation of the amount of violence to which the children are exposed (they will be exposed to more or less violence) will provide a means to test the learning-modelling theory. But what about the genetic-biology theory. That will be more difficult to test. This is because we somehow will have to identify a gene that is associated with violence. Then, we need to manipulate genes in our children. Perhaps we identify those who have this gene. Then, in some of the children, we remove the gene (through genetic engineering). Do they now become less aggressive? In a group of children who do not have the gene, we add it. Do they now become more aggressive? Altering genetic structure is very difficult. And even if scientists could do it, is it morally ethical to do so? Please see the section in the textbook about the ethics of research for an in depth discussion of this issue. Thus, it is highly unlikely that we could actually test our alternate theory.
- Rather than altering genes, we could manipulate (alter) the brain. We remove the aggression areas of the brain. Now are the aggressive children less aggressive? Obviously, we could not ethically do this type of study. But perhaps we could identify those who are the most and least aggressive and then scan their brains using an MRI. Are the brains of the most aggressive different from those who are the least aggressive? Note that this is not a true experiment (because we have not manipulated anything; see next section). Further, even if the experimenter does find a difference in the brains, this does not tell us why the brains are different. We would not know the *cause* of the aggressive behaviour. Perhaps the brains are different because children who are aggressive have different upbringings. Perhaps their diets are different. Perhaps exposure to violence causes a change in the brain structures.
- Measure/quantification: How does one quantify violence? Aggression?

Types of Studies

In the natural sciences, *true* experimentation is often carried out. This is often not possible in the humanities, in medicine and in social sciences. We shall examine the following types of studies: case studies, true experiments, quasi-experiments, naturalistic observation, survey studies and correlational studies. Please note that the list is somewhat arbitrary and not mutually exclusive. A survey might also, for example, be a correlational study. It might also be a true experiment.

Case Studies

- Study one or more individuals in detail to obtain data that would be true for all of us. Often these individuals are exceptional cases. For example, we might study a brilliant mathematician, an individual with outstanding memory, or perhaps an extreme autistic.
- Can suggest hypotheses for further testing
- BUT, can be misleading
- We can all think of a single case that would prove the theory right or wrong.

Sampling

When we study groups of individuals, we obviously cannot study all members of the group. If we wish to compare males and females, we cannot study every single member of the Canadian population. We therefore need to select certain individuals who are representative of the population.

- Population - the whole group you wish to study
- Sample - a smaller *random* (every member of the population has an equal likelihood of participating in the study) selection of individuals in the population
  - The sample must be *representative*. It is better to have a small but representative sample than a large, unrepresentative sample.

True Experimental Studies

- In a true experiment, the factor (or *variable*) of interest is *manipulated*. This variable is called the *independent* variable. Let's return to the study of the influence of alcohol. We might manipulate the dosage of alcohol. The scientist will give more or less alcohol. One group gets a large dosage (say 200 ml of vodka… that's about 7 or so oz!). A second group gets a smaller dosage (say 50 ml). A third group gets no alcohol (0 ml of vodka). We mix the alcohol with a mint drink. None of the participants in the experiment can tell which drink they are getting. The experimenter also does not know the dosage an individual subject was given (note: volunteers for experiments are at times called "participants" and at other times, "subjects". I prefer "subjects" and will use this term during this course). Thus, neither the subject nor the experimenter knows what dosage of the drug has been given. This is called a *"double blind"* study.
- Our hypothesis is that alcohol will affect performance (we obviously need to define "performance". What task will the participants perform? Larger dosages will affect performance more than small dosages.
- We assume our expectation to be wrong! Again, this is the null hypothesis. We thus assume that alcohol does not affect performance until proven otherwise.
- If our theory is true, the manipulation of the independent variable *causes* another variable to change. This is a *cause-and-effect* relationship. In the example, alcohol might cause performance to worsen. The variable that is affected by manipulation of the independent variable is called the *dependent* variable. In this example, our dependent variable is performance on a task. We need to define the task. Let us assume that the participant is asked to push one button if a red light is presented and another button if a green light is presented. We will measure how long it takes to respond to each of the lights. (The delay in responding is called the "reaction time" (RT)).

- The dependent variable is that which we measure. Thus, the dependent measure in this example is RT. The dependent variable, RT, might vary because of the independent variable, the dosage of alcohol. If the theory is correct, RT should be slower when a high dosage is given and faster when a low dosage is given .
- While certain variables are manipulated, others are held *constant*. Everything in the alcohol experiment is held constant, except for the dosage of alcohol. All participants, regardless of the amount they drink, see exactly the same lights. They all do exactly the same thing in the lab.
- Note that there are two things that vary. The dosage of alcohol that our three groups receive is varied (or manipulated). Performance (measured by RT) may also vary. We thus have an independent variable (it varies…we give more or less alcohol) and a dependent variable (RT *may* vary; it might be faster or slower. But it will only vary if our theory is correct. It will vary if the independent variable causes it to vary; it will not vary if the independent variable does not cause it to vary). If alcohol does indeed affect performance (RT becomes slower with higher dosages), we *reject* the null hypothesis as being *false*. Our conclusion is thus that it is not true that alcohol does not affect performance. In other words, alcohol does affect performance. Alcohol *causes* RT to vary. If RT is not influenced by alcohol, it will not vary. RT will be the same whether participants drink alcohol or not. In this case, we fail to reject the null hypothesis. We do not have evidence that alcohol affects performance.
- Note that if RT does vary, we will know why. We have obtained *knowledge* about the *effects* of alcohol. It *causes* RT to slow. Assuming this is the first time the experiment has ever been run, we now know something we did not know before. This is called scientific *progress*. When scientists do not know why certain variables vary, they are said to be *ignorant*. Ignorance is the absence of knowledge. Variance that cannot be explained is called *error*. A "mystery" is something we cannot explain (for which we do not have knowledge of why it varies). It is assumed that scientists can potentially explain everything in the universe. If, at the moment, they cannot, they must be making an error. Please note that contrary to popular belief, scientists can, in fact, only explain a very few things. We remain, therefore, very ignorant about most of the mysteries of the universe.
- There is also another very important source of variance in our experiment. We have a certain number of subjects in the experiment. Their reaction times will not be identical. Some will be faster. Some will be slower. In other words, there are individual differences (or variance) within each group. What is the cause of this variance? The only variance we can explain is that which we manipulated. We do not know why individual subjects' RTs vary. Variance as a result of individual differences is thus called error variance. We can speculate about why individual RTs vary. Perhaps some individuals were more interested in the study. Perhaps certain individuals are simply faster while others are slower. Perhaps some react to alcohol more than others, but we are only guessing. We do not really know any of this. To gain this knowledge, we need to manipulate these variables. So, if you think "interest in the study" explains RT, you now need to design another experiment. You need to define "interest" (and how would you do this?). You need to manipulate "interest". Then, you examine if "interest" in the task does indeed cause RT to vary.
- Our dependent variable (RT in this example) therefore varies for two reasons – because of the effects of the independent variable (that which we manipulated) and individual differences. We can explain the effects of the independent variable. Alcohol might cause RT to vary. We cannot explain variation among individual participants. In any experiment, there are thus two sources of variance: one that we can explain, and one that we cannot explain. As we shall later see, this becomes crucial when we attempt to determine if differences due to the independent variable are *statistically significant*.

Experimental Designs

Researchers employ different designs to test their theories. Studies are designed to assure that the results that are obtained cannot be explained by "confounds"

- One design uses *control and experimental* conditions (or perhaps groups). *Control and Experimental groups.*
    - The control condition provides a baseline to which the experimental condition can be compared. No experimental manipulation is carried out with the control group. In the alcohol example above, the control group receives no alcohol.
    - The experimental manipulation is carried out with the experimental group. In the example above, the experimental groups receive different dosages of alcohol.
    - *Random assignment* -- participants will be randomly assigned to either the control or experimental conditions.
    - Again, an example will be used. We want to know the effects of sleep deprivation on RT. We will manipulate the amount of sleep that participants will have and then examine the effects on performance (on the speed of responding or RT). The independent variable is thus the amount of sleep. The dependent variable is thus RT. Twenty individuals will participate in the study. Half (10) of them are assigned *at random* to the experimental group. The other half are assigned to the control group. The experimental group is tested following 24 hours of sleep deprivation. The control group is not sleep deprived. The amount of sleep is thus manipulated between the two groups.
- *Pre-post designs.* A problem with the use of 2 different groups (experimental and control) is that whatever differences we find in our independent variable might be due to chance. Let us assume that RT is much slower in the experimental than the control group. Can we then conclude that sleep deprivation causes a slowing of decision-making? Not necessarily. Perhaps the experimental group was simply a very slow group. Had this group not been sleep deprived, the participants would still have had slow RTs. The fact that we randomly assigned participants to one group or the other should control for this finding, but it is possible that just by chance the slow responders were randomly assigned to the experimental group (just as it is possible to sometimes flip a coin and get 5 heads in a row).
    - The solution to this dilemma is to use the same group of participants in both *conditions.* The group is then tested prior to sleep deprivation and again following it. They are thus tested pre- and post-deprivation. The *pre-post design* thus controls for possible random differences in the selection of different groups. But… there are also problems with the pre-post design. Differences between pre- and post- conditions might be due to a "confound" such as practice (with repetition of a task performance might be better), or perhaps participants become bored of having to do the same task a second time, in which case, performance deteriorates.
- A special design is used to examine the effects of "treatment". Let us assume that I have a theory that suggests that a new drug I have discovered will be effective in the treatment of cancer. I give the pill to a group of cancer patients. After 3 weeks, it is found that the pill is quite effective. More than 40% of the patients' symptoms are reduced.
- Does this necessarily mean my drug was effective? No! It is entirely possible that the reductions in the symptoms might have happened even if there were no drug treatment. Or, perhaps the fact that patients knew they were being treated for cancer caused the change in their symptoms. These types of changes are called the *"placebo" effect.*
- To control for this, one group of patients is given the actual drug (or any other treatment) and the other half is given what they think is a valid treatment. They might be given a sugar pill but are given exactly the same expectations as the other group. This control condition is called the

placebo condition. To ensure that all patients have the same expectations, neither the experimenters nor the patients know which treatment (the actual drug or the placebo) is being given. This is called a *double blind* design. Placebo effects can be extremely powerful. Placebos have been shown to be very effective in decreasing feelings of extreme pain, almost as powerful as a potent anaesthetic. Many medical disorders improve upon administration of a placebo. No medical treatment that is approved by the Ministry of Health can be used with the general public before a placebo study is carried out. In clinical psychology, the placebo effect is a particularly good explanation of the supposed benefits of psychotherapy and counseling. Many scientists claim that the positive effects of so-called natural medicines and homeopathic treatments are due entirely to the placebo effect. In short, these treatments cannot be claimed to "work" at all because controlled, placebo studies have not been carried out. Until the placebo studies are carried out, scientists assume the null hypothesis. The treatment is assumed not to work.

Problems with True Experiments

- Often the sample size is small. Can we generalize to the population as a whole from such small samples?
- Experiment must be carried out in a controlled setting (often in a laboratory). Is this typical of the real world? Let us suppose we are studying changes in personality as a result of alcohol. We hypothesize that alcohol will make participants less inhibited. They will socialize more and perhaps be more aggressive. Lab results indicate, however, that the participants are not more aggressive or sociable. Perhaps this can be explained by the fact that they are in a lab environment and do not act as they would in a more normal, social bar-type situation. Is it possible to generalize from the results of a lab-based study to the real world?

Quasi-Experiments

- Often in the humanities and health sciences, it is not possible to manipulate the independent variable.
- As an example, a scientist has done an extensive search of the literature and observed that there is good evidence that that men do perform slightly better in math than women. The scientist then develops a theory that men are better at math because of their male hormones. To test this theory, we need to manipulate the independent variable, the quantity of male hormones, in both males and females. A study is designed. We arrange to test a sample of undergraduate students on a math test. The results do indeed indicate that the men do better than the women. Why? (Again, in science we want to know the cause of variation. The scores on our math test vary and we do know that sex or gender will cause the scores to vary, but why?). Recall in a true experiment, we manipulate the independent variable. But in this experiment, nothing was manipulated. To prove that the *cause* of the difference is due to hormones, we need to *manipulate* the quantity of male and female hormones in our participants. In a true experiment, some women would be given massive dosages of male hormones (and thus become male-like) and some men would be given massive dosages of female hormones (and thus become female-like). If the differences in math were due to hormones, then the women should now do better in math than the men. How many humans would however volunteer to participate in this study? More importantly, is it ethical to carry out such a study?
- Several other examples could be given – differences between younger and older participants; differences between normal controls and patients.
- In a quasi-experiment, it is not possible to manipulate the independent variable (example, sex, age, race…)

- It is however assumed that the differences that are found are *caused* by the independent variable. This can be a very questionable and controversial assumption.

Natural Observation.

Not all studies in the humanities are carried out in a laboratory. As we noted above, it may be difficult to generalize from a lab to true life situations. Human participants may act differently in a lab situation than they would in their natural environment. For example, a participant may be able to compensate for the effects of sleep deprivation in a lab context. However, put them on a real highway and they fall asleep. In the alcohol studies described above, very few participants ever become violent or aggressive. Does this mean that alcohol has no effect on aggression or other emotions? Go to a real bar and there are only too many examples of violence.

- Studies carried out in the field or "natural" environments attempt to overcome the limitation of generalization that is imposed on lab studies.
- But, there are problems with studies in the natural environment. It is exceedingly difficult to implement the control of other variables that might affect the results (confounding variables) afforded by the lab context. Thus, while there might be considerably more violence in a real bar, what accounts for the violence? Is it only alcohol? How can these confounding variables be controlled?
- Again, it is very difficult to carry out true experiments in the natural environment. As such, the researcher cannot easily know why differences occur.

Surveys

In the social sciences and in particular in social psychology, we often gain knowledge of human beliefs and attitudes by directly asking about them with *surveys*.

- Participants are asked to report their behaviour, attitudes or beliefs
- The participant is asked a question with which they can either agree or disagree.
- Wording of questions can be crucial.

    Example of survey questions: *Do you agree or disagree that*:
    a. Canada should not allow pornography.
    b. Canada should censor pornography.

    a. Canada should not put marijuana users in prison.
    b. Canada should legalize marijuana usage.

  The phrasing about your opinions about pornography and drug usage varies ("should not allow" and "should censor") but in essence, the two statements are identical ("should not allow = "should censor"; "Should not put… in prison" = "legalize"). Yet we might find that the proportions of respondents that answer "Yes" will vary depending on whether they are given question a or b.

- Survey studies are often quasi-experiments. If a researcher determines that 65% of men and only 35% of women are opposed to gun control in Canada, this would be an example of a quasi-experiment. The sex (or gender) of the participant could not be manipulated. Therefore, we do not know why males and females answered the way they did.
- Surveys can also be used in a true experiment. Suppose we do the same survey on gun control and obtain the same results. Now, the participants are shown a film depicting just how many

people are murdered by guns. Now, the results change. Most men might now no longer be opposed to gun control. This study carried out a true manipulation. As a result, it is possible to say that we now *know* that attitudes toward gun control will vary depending on prior learning.

Measures of Central Tendency

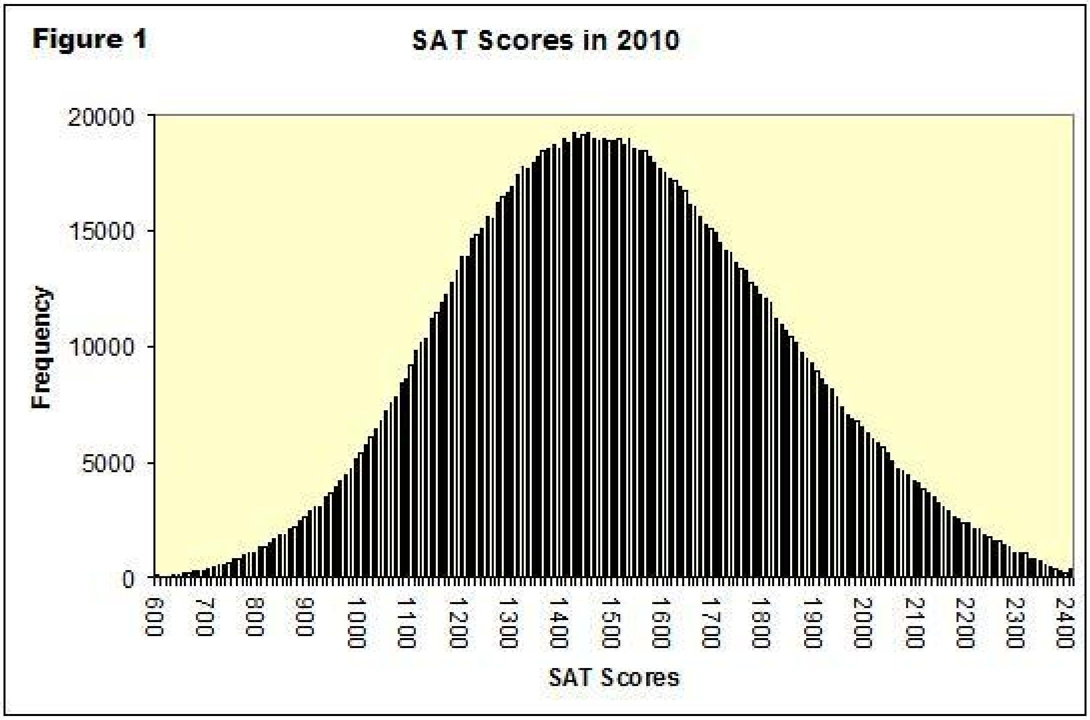
- There are usually (almost always) individual differences in the dependent measure. Some individuals score high; some score low.
- Statisticians typically employ three different measures of central tendency (the "typical" score).
- These are the mode, mean and median
- The *mode* is the score that occurs most often.
- The *mean* is the average of all scores.
- The *median* is the score at which half the individuals score above and half score below.

Problems with the Mean as a Measure of Central Tendency

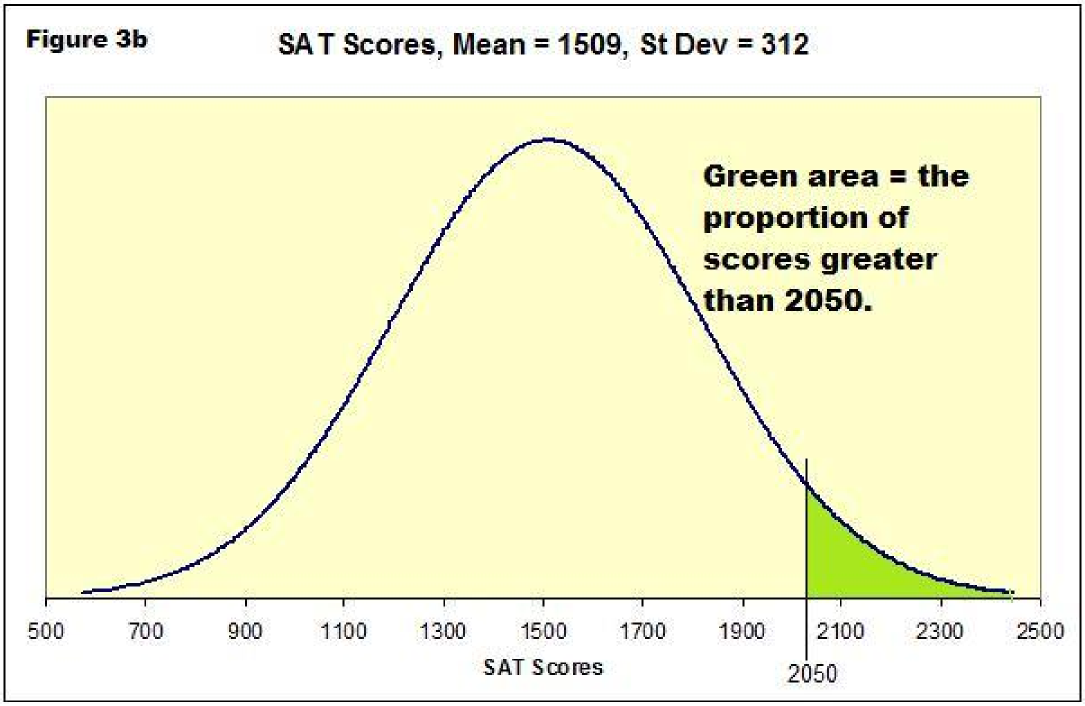
- At times, our measures are not normally distributed. Extremely high (or low) scores might distort the average (i.e., the mean). Most university students are in their late teens or early 20s. However, some are in their 70s. The older students would tend to "pull up" the mean.
- In this case, the median might be a better measure of central tendency. Why?

The Normal Curve

- A *distribution of scores* is a graphic representation of how many individuals have a particular score.
- If I ask the age of students in this class, perhaps 40 are 17, 120 are 18, 50 are 19, 30 are 20, 12 are 21, and perhaps 5 more than 21.
- Many psychological measures are distributed according to a bell-shaped *"normal" curve.*
- In the example above (age of students in this class), the age of students would not be distributed normally, but if we survey your grade point average (GPA) in high school, we should obtain a normal curve. Let us suppose the mean (or average) is 83%. Most students have a GPA that is about 83%. A small percentage might have a GPA that is above 95% and a small percentage below 71%.
- The image below represents the SAT scores in the U.S. in 2010. This example can be found on the web (http://introductorystats.wordpress.com/2011/09/24/an-example-of-a-normal-curve/). In the U.S., the SAT is a test taken by all high students who hope to go on to universities. In 2010, about 1.5 million (that is not a typo) students wrote the exam. The minimum mark on the exam is 0 (theoretically) and the maximum mark is 2400. The mean on the SAT for the 1.5 million students is about 1450. Figure 1 presents the "frequency distribution" of the students. There are 181 bars in this histogram, corresponding to the 181 distinct possible scores in the data (the scores come with an increment of 10, ranging from 600, 610, 620, and all the way to 2390, and 2400). Because of many bars crammed into a small graph, each bar appears as a thin vertical line. The histogram is symmetrical around a single peak, the actual mean of all the students (around 1500) and it tapers down smoothly on each side. Most of the data are clustered in the middle. The bars around the middle are very tall (i.e. most students score in the middle range). On the other hand, the bars at either the extreme left side or the right side are very short (very few students have extremely low or extremely high scores). As a result, the histogram has a "bell" shape. This is the shape of a *normal* curve.



- Note that the normal curve is balanced. Half of the individuals score at or above the mean and half score at or below the mean. Again, this is the definition of the *median*. Most score around the mean. Again, this is the definition of the *mode*. Only a small percentage will score well above or well below the mean.
- In a perfectly normal curve (as in the example above), the mean, the median and the mode are therefore identical.
- In Figure 3b below, we isolate the top 5% of the scores (shaded in green). This is a score of 2050 or higher.
- Thus, 95% of the students score below 2050.
- In this sense, the top 5% of the students (those scoring 2050 or above) are said to be statistically “abnormal”. It is critical to realize that what we mean by statistical abnormality is *arbitrary*. Is it the top (or on the other side of the curve, the bottom) 1%, 5%, 10%, 25% of individuals?
- As we shall see, the concept of statistical “significance” does require an understanding of the normal curve.
- In a true experiment, we manipulate the independent variable. Let’s assume that prior to the SAT, we give one group of students a memory drug. In principle, memory should improve. We give this drug to 50 students. We then compare their scores on the SAT with 50 students who did not take the drug. If we look at the normal curve in Figure 1, we would predict that the mean SAT score for the non-drug group would be about 1500. However, it is possible (but unlikely) that the nondrug group could have a mean of 2050 or above. By chance, we might have selected out 50 of the best students. What is the probability that we would have selected out the top students by random selection alone? .05. Thus, if we ran our study 100 times, we might have selected out a group that has a mean of 2050 or higher 5 times. Now, let’s look at our drug group. After ingesting the drug, their mean score is 2050. This is highly unlikely to occur by chance alone (but will happen 5 times out of 100) and is thus said to be “abnormal”. Statisticians will thus say that a different this large will not happen by chance 95 times out of 100.



Statistical Significance

- Statisticians have developed a measure called statistical significance.
- This means that differences of a certain magnitude could not occur by chance (see Discussion of Figure 3b above).
- A statistician might claim that differences this large could only occur by chance on 5% of occasions. This means that if we would run the experiment 100 times, we would find differences this large just by chance on 5% of occasions (i.e., 5 times).
- Let's use the example of alcohol and reaction time. The experimenter chooses 10 individuals at random to be in the control group and 10 individuals to be in the experimental group. The experimental group is asked to drink the equivalent of 10 beers. Subjects are asked to push a button whenever a red light is flashed on a monitor. The overall (or "grand") mean of all 20 subjects is 390 but this varies. Some individuals are above the mean; others below. The mean RT of the control group is 320 ms (milliseconds) while that of the experimental group is much slower, 460 ms. In short, the experimental manipulation appears to have caused the RT to increase. However, we might randomly select 10 subjects and place them in a group and select another 10 subjects and place them in a second group. We do not give any of the subjects alcohol. The two groups should thus, on average, have the same reaction times as our original control group, but just by chance (by the luck of random selection), they might not. What is the probability that the difference in means of the two groups will vary by 140 ms (exactly the same as in our experiment)? Statisticians will always provide a measure of chance probability. They might thus indicate that a difference of this magnitude could be found by chance (by random chance selection) only on 1% of occasions. In short, if we chose two groups at random (and did not give them any alcohol) and we did this100 times, we would obtain differences of this magnitude only 1 time. This is what we mean by a *chance* finding. There is no reason to expect that the mean RT of the two groups should vary, other than pure chance. Indeed, we have 99% likelihood that these differences could NOT be due to chance. *Statistical significance* thus provides

a measure of how often a difference could be found purely by chance. The researcher must always state what they consider to be statistically significant, .001, .01, .05, .25. Again, the probability is somewhat arbitrary. If I develop a cure for cancer, I might want to be more liberal about what is considered to be significant. Let's set the level of significance at .25. Thus, there are 25 chances in 100 that the cure does not actually work. On the other hand, I develop new brakes for a car that can stop your car 1 cm shorter than other brakes. However, there is one chance in a thousand (.001) that the brakes may fail completely. Would you use these brakes?

- Note that statistical significance and practical significance are not the same thing. A result can be statistically significant but have no practical significance.
- There are two sources of variance in any experiment. One we can explain (because of the experimental manipulation); one we cannot (individual differences).
- How do we determine if our experimental effect is statistically significant?
- The explained variance is divided by the unexplained variance (knowledge is divided by what we do not know).
- But how do we measure variance? We need to compute explained variance and unexplained variance. As the name implies, variance is computed by examining the extent to which scores deviate from the mean. You need to think about the following concepts. They are the essence of ALL science, whether you study Psychology, Physics, Biology, Sociology, Criminology, Medicine or Business. Note that if the experimental manipulation has a large effect, the explained variance will be very large. On the other hand, if the experimental manipulation has no effect at all, the explained variance would be zero (0). Unexplained variance examines the extent of individual differences. If all individuals scored exactly the same, there would be no individual differences. Thus unexplained variance would be zero. On the other hand, if individuals have very different scores, there would be large unexplained variance.
- Unexplained variance is computed by subtracting the mean from each individual's score, telling us how much an individual varies from the mean. (Note that we do not know why these individuals deviate from the mean… this is "unexplained"). As would be expected from the normal curve, some individuals will be above the mean, and some below the mean, giving negative and positive *deviance*. We square these scores (to eliminate the sign), sum up these squared deviances, and then divide by the number of individuals in the group (thus we have a mean or average deviance).This provides a measure called (not surprisingly) "variance". However, recall that we initially subtracted each individual's score from the overall mean, then we squared this difference. Thus, the variance is actually a measure of the squared difference. We still therefore need to take the square root of variance. The square root of variance provides another measure, the standard deviation (SD).
- In an experiment, there is an overall (or "grand") mean. In the example above, the grand mean was 390 ms (320 for one condition +460 ms in another condition, divided by 2). We then subtract this grand mean from the group means (320-390 and 460-390), square the differences, and sum these squares, then divide this sum by the number of groups to give a mean or average explained variance.
- If the ratio of explained to unexplained variance is large enough, the result is probably statistically significant. It is not likely that this difference could be due to chance.

Factors that affect statistical significance

- Recall that significance is a reflection that differences this size are not likely to be found by random chance.
- Explained variance (knowledge) is divided by unexplained variance (individual differences or ignorance).

- Thus, the likelihood of finding statistical significance increases according to the size of the experimental effect. Larger differences are more likely to be significant. The likelihood of finding significance decreases when the size of the experimental effect is small.
- The likelihood of finding significance increases when unexplained (individual) variance is small, and decreases when individual variation (individual differences) is large. In a pre-test, if all individuals score the same, and then in a post-test, they score only slightly higher, this is not due to chance. If the experimental manipulation has a highly consistent effect across all individuals, the effect is likely to be significant. In the alcohol study, we noted that the drug might slow reaction time. If only 10 participants are sampled, the experimental effect of the drug (alcohol) will probably be significant if all 10 participants show slow reaction times following ingestion of alcohol. It would be quite unlikely to flip a coin 10 times, and get heads on each flip.
- The size of the sample. If a very large sample is used, very small differences might be statistically significant. On college entrance exams in the U.S., men score minutely higher than women on math, but this is statistically significant. This reflects the representation of the sample. Large samples, because they are composed of so many members of the population, are much more likely to be representative of the population. In this case, many women actually score higher than men, but on average men score just slightly higher than women. Again, statistical significance does not imply practical significance.

Correlational Studies

Many studies that are carried out in the social and medical sciences are not experiments at all. Rather, they are best described as "correlational" studies. As an example, it is known that poorer people commit more crime than richer people. Does this mean that poverty causes people to commit more crime? Perhaps; perhaps not. Also, the more one smokes, the more likely it is that they will get cancer. Does this mean smoking causes cancer? It might, but there may be other factors that cause cancer. Why do people smoke to begin with? Maybe what causes people to smoke is what also causes the cancer. In these examples, a change in one variable is associated with a change in another. There is thus a co-relationship (or a *correlation*) between smoking and cancer and between poverty and crime.

- Correlation - A statistical measure of the extent of a relationship between two variables
- Correlation allows one to predict scores on one variable if the scores on another variable are known.
- Correlations vary from -1.0 to +1.0. A negative correlation indicates that as the scores on one variable increase, the scores on another decrease. The more alcohol one drinks, the less able one is to drive a car. A positive correlation indicates that as one score increases, the other score increases as well. The more one studies, the higher the marks; the more calories one eats, the heavier one becomes.
- A correlation permits a prediction. If there is a high correlation between eating calories and weight gain, knowing how much you eat, I can predict how much weight you will gain. Knowing how much alcohol you have consumed allows me to predict how well you can drive a car. Both positive and negative correlations allow the researcher to predict. The sign (+ or -) is thus incidental. It is the size of the correlation that is important. With a correlation of 1.0, there is perfect predictability. The closer the correlation is to 1.0, the closer the association between the two variables.
- On the other hand, if there is no relationship, the correlation is 0.0. Knowing the scores on one variable does not help predict the scores on another. Height is a very poor predictor of marks in Intro Psychology. The correlation between height and marks is thus close to 0.

- *Correlation does not prove causality*. While there might be a high correlation between smoking and the incidence of cancer, this does not mean smoking *causes* cancer. Perhaps smokers are under more stress than nonsmokers, and it is the stress that causes cancer, not smoking. There is a strong negative correlation between age and memory. The older we become, the more we forget. This does not mean however that ageing *causes* memory loss. Because the elderly have lived so long, they have stored much more than younger adults. They thus have a more difficult time searching through all these memories. Perhaps if a younger adult could be identified that also has stored a tremendous amount, they could have difficulty retrieving these memories. In Criminology in certain countries, it is hypothesized that capital punishment should reduce the incidence of murder. However, countries that have capital punishment also have very high murder rates. Often, countries that do not have capital countries do not have high rates of murder. Does the fact that the sentence for murder is capital punishment actually *cause* more murders to be committed?
- Again, in order to prove causality, the researcher must carry out a true experiment. In the correlational studies, no manipulation is carried out.