The Scientific Method

Cognition Lab – Rutgers University

Learning Goals

- Learn about what science is
- Learn some basic methods used to measure how we perceive or represent objects
  - *These methods are relevant to the study of cognition*
- Learn the difference between accuracy and precision of judgments
- Learn about analyzing data, including mean, SD, and a simple statistical test
- Learn how to write some elements of a laboratory report

What is science?

Chances are when you've taken science classes previously, you learned to memorize facts and formulas.

For example, if you took an astronomy class you would learn that the speed of light is approximately 300 million meters/second, and that you can determine the distance of a celestial object if you know its apparent brightness (how bright it looks) and its absolute luminosity (how bright it really is).

Facts and formulas are very useful for assignments and exams, but they don't reflect what it really means to **do** science.

So, what is science then?

Science is about discovery! When scientists come up with questions to test, the outcome is unknown. Generally a scientist will come up with some predictions about the outcome and how to interpret the predicted outcomes, but the results often don't come out as expected. Even when scientists are wrong, they can be wrong in ways that they don't expect! This is sometimes more interesting than being right.

Science is always changing! Theories are evolving all the time, and every now and then they are overturned completely. For example, Newtonian physics dominated the field for hundreds of years until Einstein's theories came along. And even Einstein's theories have been challenged by quantum mechanics. The changes to physics over time is an example of what happens in **all** sciences.

So, what is science then?

Science is not about *proving* hypotheses! Scientists do experiments to find evidence in support of (or that contradicts) a hypothesis, and hypotheses come from theories. Theories describe what we think the world is like, given the evidence we have at this point and what we think that evidence means. Usually theories don't fit the evidence perfectly, which is why theories are always changing. The goal of science is to get as close as we can to discovering the true state of the world, while acknowledging that it may not be possible to know definitively what the truth is. Conversely, you could say that science is about becoming less and less *incorrect* about the true state of the world.

What counts as science?

In order for something to be scientific, it needs to meet three necessary criteria

1. It must use **empirical methods**
2. The hypothesis tested must be **falsifiable**
3. The results must be **verifiable**

Empirical Methods

Scientific experiments involve systematic **empiricism**. “Empiricism” is a method of collecting data through observation. This means that scientists observe events under carefully controlled circumstances in order to avoid bias. There are two primary sources of bias: bias from the experimenter’s beliefs and bias from outside influences that are not well controlled. Scientists use careful experimental design to insure the evidence generated by their experiments is solid.

Testable Hypotheses

Scientific experiments must test falsifiable hypotheses. A falsifiable hypothesis is a hypothesis that can be disproven.

Falsifiable example: All cats have stripes.

This can be easily falsified by observing a cat without stripes.

Unfalsifiable example: Cats have souls.

No observation can prove this false because of the way the hypothesis is posed.

Verifiable

The findings of scientific experiments must be verifiable. This means:

1. Scientists must publish their findings through peer-review to get credit for them. Peer-review means that other scientists in the same field must review and approve of the methods. This process prevents flawed experiments from being accepted as scientific.

2. Scientific experiments must be replicable. This means that another scientist could recreate your experiment and get the same results you did. If they follow your methods exactly and get different results, something is wrong.

3. Ideally, scientists should share access to the dataset that their results are based on so that other scientists can verify the results.

RUTGERS

Independent and Dependent Variables

Independent Variable(s)

The variable we are manipulating

Dependent Variable(s)

The response measures we are collecting

Both Independent and Dependent variables are concrete entities that can be measured.

Example: An experiment is done to address the question: Does caffeine improve running speed?

Independent variable: Amount of caffeine ingested (measured in milligrams).

Dependent variable: Time to run a 5 Kilometer race (measured in minutes).

Psychology Experiment Components

A **trial** refers to collection of a single response (or other type of data point) in an experiment. In the caffeine example, one 5 K race would constitute a single trial. In the experiment we'll be doing in class, a single trial will generally consist of a stimulus (for example, a word on the screen) or question and a response. You should only ever use the term ***trial*** to refer to an individual trial, and not to a group of trials.

A **block** or **session** refers to a group of trials that usually share the same experimental parameters, meaning they are in the same condition.

Experimental Design

Consider the caffeine experiment example described on the previous slide. If you wanted to test the effect of caffeine on running speed, how should you design your experiment?

Running speeds vary from person to person for reasons that you don't care about (height and stride length, athletic background, etc.) so you don't want to directly compare one person's time to another. Instead, you should measure a baseline for each participant and compare the difference in time.

The independent variable (caffeine) should be manipulated systematically. For example, you could test the effect *0 mg*, *10 mg*, *20 mg*, *30 mg*, etc., but it would be difficult to justify testing *0 mg*, *25 mg*, *38 mg*, and *60 mg*, because there is no obvious systematicity to those dosages.

Experimental Design

Things to watch out for:

You may have noticed on the last slide that there was a *0 mg* caffeine condition. Why would you need that if you are already measuring a baseline condition without caffeine? Participants get better at a task with practice, and you don't want to mistake a **practice effect** for the effect of your manipulation. Including a *0 mg* test condition serves as a **control**.

Sometimes the order of the conditions can affect behavior in ways that you would not expect. When this happens, it's called an **order effect**. These can be avoided by giving different participants different condition orders, or by selecting conditions at random.

Experimental Design

Things to watch out for:

Anything that varies with your manipulation but is not part of your manipulation is a **confound**. In the caffeine example, let's say you tested different dosage conditions on different days, and on some of those days the temperature was cold. On the colder days the run times would be faster because (counterintuitively!) cold weather increases running speed. Depending on what condition your participants completed on that day, the confound could either mask the effect of your manipulation or inflate it. Both are bad news! Experimenters need to anticipate and control for confounds as much as they possibly can.

Participant Selection

Because only members of this class are allowed to participate in our experiments, we won't worry about how we select our participants. But in research the way participants are selected can bias the outcome.

For example, in 1936 the *Literary Digest* famously projected Alfred Landon as the winner of the presidential election over Franklin D. Roosevelt. There is no President Landon in America's history, so obviously the projection was wrong. What happened?

The *Literary Digest* selected poll participants using phone books and other listings for luxury items during the Great Depression, which biased the sample by excluding the lower class.

Common Measures

In the experiments we will do this semester, we will frequently measure **decision accuracy**. This is exactly what it sounds like: the percent correct for an experimental condition, calculated by tallying the correct responses and dividing by the total. Decision accuracy can be used to assess the difficulty of a task, and can be used as the sole response measure or in conjunction with another measure.

$$\% \text{ correct} = \frac{n_{correct}}{n_{total}}$$

Common Measures

You've probably heard of **reaction time** (RT) before, but you may not know what it means to psychologists. Reaction time (usually measured in milliseconds) refers to the time it takes between the appearance of a stimulus and the participant's response to that stimulus. The time it takes the participant to respond reflects the processing time required to complete the task. This allows us to measure the time course of cognitive processes.

Reaction time can also reflect response difficulty: more difficult judgments take longer to make, even if the ultimate decision accuracy is very high. In other words, just because the participant performs well doesn't mean the task was necessarily easy.

$$RT = t_{response} - t_{stimulus\ onset}$$

Descriptive Statistics

Descriptive statistics tell you about the basic characteristics of your data.

The **mean** is the average of the response values in the sample, taken by summing the values and dividing by the total number of values. The mean tells you where values tend to be located in the sample.

The **standard deviation** reflects the variability in the data. It is calculated by getting the difference from each data point to the mean, adding up and then squaring those differences, and dividing all of that by the total number of values. It tells you how similar the values in your sample are to one another.

You will be learning to calculate these values using spreadsheet software.

Why does variability matter?

Even **when the experimental parameters are the same**, we expect there to be some response variability due to noise. There are three sources of this kind of variability:

1. **Physical**: due to variability in the equipment
2. **Neural**: due to variability in the brain
3. **Psychological**: due to variability in attention, memory, decision criteria, etc.

Realistically, some variability due to all 3 sources is unavoidable. It's often hard to know how much variability is due to each of these 3 sources, or to pinpoint the exact source of variability within each source (and trying to do so is an interesting line of research). But cheer up: Even if we can't know the exact sources of the variability, we can certainly measure it.

Standard Deviation: Example

Here's an example of variability when applied to test scores. The standard deviation reflects to what degree measurements vary from one another.

Barbie's exam scores	Ken's exam scores
80	89
90	90
100	91

Both Barbie and Ken have mean scores of 90, but look how Barbie's scores are spread out more than Ken's. Barbie is more variable. Her scores have a higher standard deviation because of that variability. If we only saw the mean for Barbie's and Ken's exam scores, we might think they got the same scores on the exams. But by looking at the standard deviation we can clearly see Barbie and Ken are doing different things.

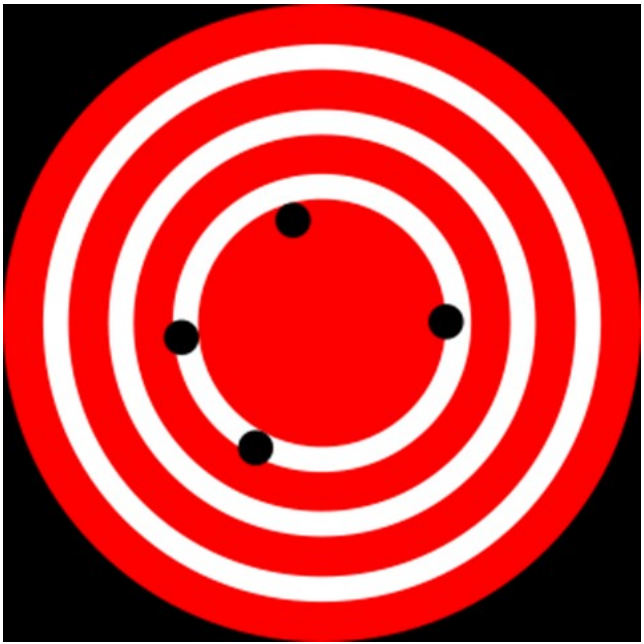
Accuracy vs. Precision

Let's briefly introduce some terminology because it might be helpful later.

Sometimes we know what the true mean is, and we can compare our sample mean to the true mean. For example, if you asked participants to estimate the size of a chair in an image you could compare their estimates to the actual size of the chair. If the **mean** is very close to the true value we can say that the judgments are very accurate.

If the **standard deviation** of the judgments is low, we say that the judgments are very precise.

Example: High accuracy, low precision



Example: Low accuracy, high precision



More examples



Variability

We just talked about the variability we need to take into account that occurs due to forces outside our control. As experimenters we want to know what variability in the data is due to our manipulation, not due to noise.

This brings us to hypothesis testing, which is something you learned about in quantitative methods. We'll only review hypothesis testing at a very basic level here, and will revisit it some when we do more sophisticated data analysis.

Hypothesis Testing Basics

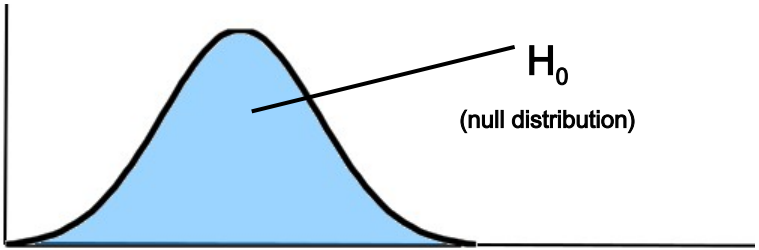
You've collected your data and now you want to know whether your manipulation worked. The mean value for your manipulation *looks* like it's higher than the mean for your control condition, but there's a chance that difference occurred due to random variation rather than your manipulation. Remember, we expect there will be some variation in the data due to noise, and if that variance is large then the means can be misleading.

In other words, the difference between the means *could* be a real effect, or it might not be. We can't tell unless we conduct a statistical test.

Hypothesis Testing Basics

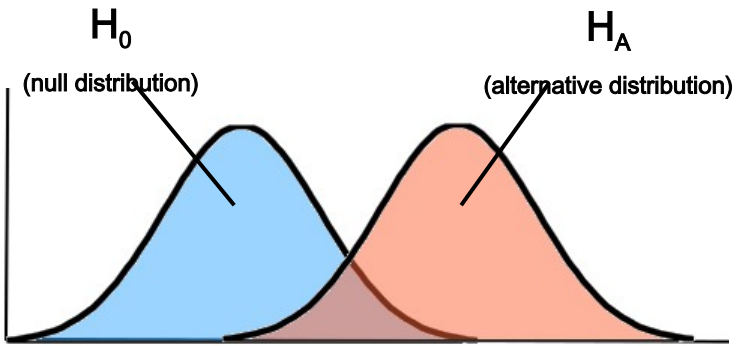
So, there's a chance our manipulation had no effect. We need to formulate that as a falsifiable hypothesis and then attempt to reject it. This is called the **null hypothesis (H_0)**: the hypothesis that the manipulation had no effect, and any apparent difference between the groups isn't real.

A way to reason about this mathematically is to say that all of the data in your sample, both in your control and your manipulation, came from the **null distribution**.



Hypothesis Testing Basics

There's also a chance that our manipulation worked, and the difference is real. We'll call this the **alternative hypothesis (H_A)**: the hypothesis that the manipulation did have an effect, and the difference we observe is real. If so, then the experimental data came from the **alternative distribution**.



Hypothesis Testing Basics

So now we have two hypotheses: the **null hypothesis** and the **alternative hypothesis**. How do we proceed?

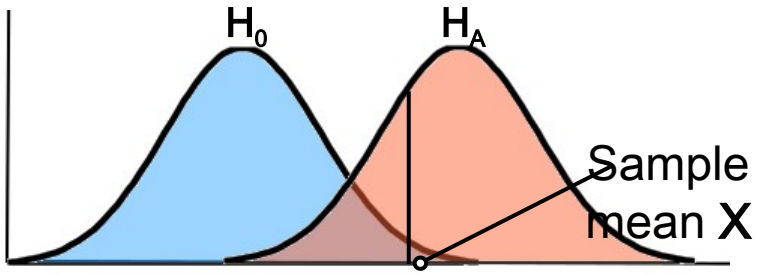
Remember, we can't prove a hypothesis. All we can do is reject a hypothesis by falsification (giving evidence that it's not true). Our goal in conducting a statistical test is to reject the null hypothesis. We do that by providing evidence that it would be extremely unlikely for the experimental data to have come from the null distribution.

We can do this by comparing the *p*-value of our statistical test to a predetermined threshold. If the *p*-value falls below the threshold, we can say the test is significant. Let's demonstrate that...

Hypothesis Testing Basics

Let's revisit our distributions. There is an area of overlap between the two distributions, and a data point X (representing a sample mean) is in that region.

X lies in the tail of the null distribution, but it's near the mean of the alternative distribution. The probability that X came from the null distribution is less than 5%. Because it's so unlikely that X came from the null distribution (< 0.05), we can reject the null hypothesis.



Hypothesis Testing Basics

When stating your results, there are some important phrasing conventions that you need to use.

If your statistical test is significant, you can say that you **reject the null hypothesis**. You can't say you've proven your alternative hypothesis because, again, we can't prove hypotheses.

If your statistical test is not significant, you can say that you **failed to reject the null hypothesis**.

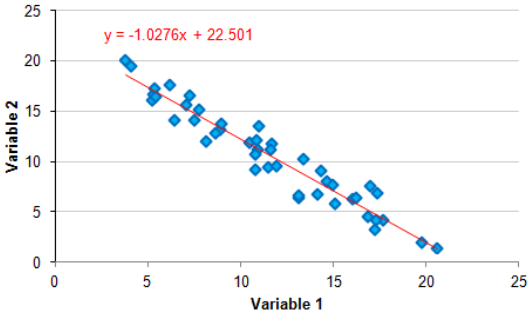
These conventions may sound silly, but they accurately reflect what scientists do.

Let's look at an example!

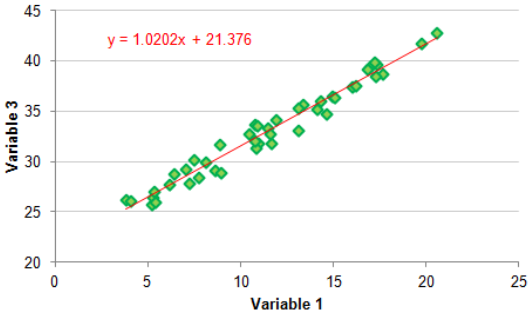
Correlation: A Statistical Test

A **correlation** is a statistical test for covariance between two variables. Covariance means that as one variable changes the other also changes, either in the same direction (a positive correlation) or in the opposite direction (a negative correlation).

A **negative correlation**.
$r(43) = -.96$, $p < .05$



A **positive correlation**.
$r(43) = .98$, $p < .05$



Reporting a Statistical Test

When reporting a statistical result, we need at least two values:

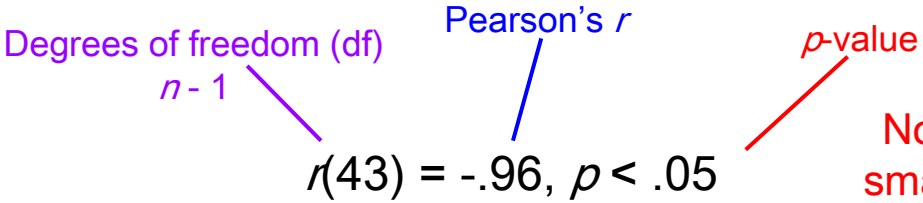
- The value of the statistical test (in our case, Pearson's *r*)
- The *p*-value (significance) of the statistic itself
  - This will be a number between 0 and 1
  - We can't claim the effect may be real unless *p* is very small (e.g., less than 0.05)

- We may also be required to report other values, such as
- The total number (*n*) of values compared
  - This is the number of values that we used to make the calculations
  - If you get a value for the test every trial, the number of values would be…
    - (number of trials)x(number of subjects)
    - …if we are using the results from all subjects

Correlation: A Statistical Test

What do we need to report for correlation:

- The value of the Pearson's *r*
  - This will be a number ranging from -1 to 1
- The total number (*n*) of values compared
  - This will be the number of values in each of our variables, and both should be the same
  - We will be reporting this as degrees of freedom (df), which is *n*-1
- The *p*-value (significance) of the statistic itself

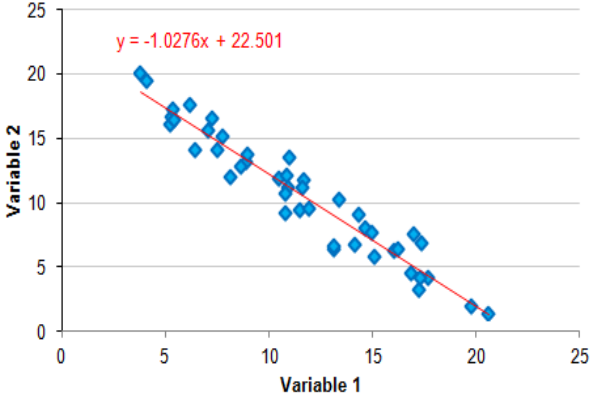


Note: since *p* is very small in this case, we can just say it's less than our threshold.

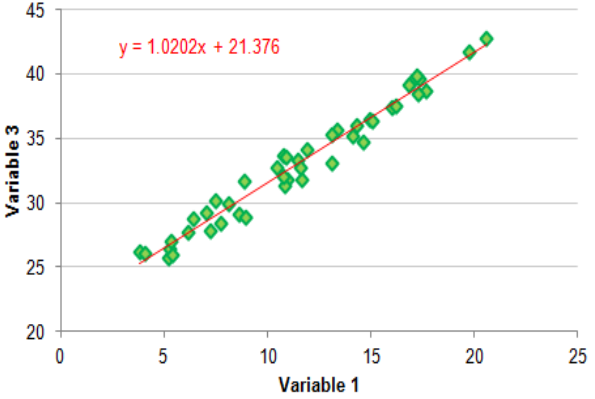
Correlation: A Statistical Test

Notice that everything is the same (or very nearly the same) in these two examples except the direction of Pearson's *r*.

$r(43) = -.96$, $p < .05$



$r(43) = .98$, $p < .05$



Correlation: A Statistical Test

But we aren't comparing means when we do a correlation. So, what's the rationale for this kind of analysis?

For a correlation, the **null hypothesis** is that there is no covariance between the variables, and anything that might look like covariance is noise.

Likewise, the **alternative hypothesis** for a correlation is that the variables do covary.

If the *p*-value associated with the correlation statistic is very small (less than 0.05), we can reject the null hypothesis because it is highly unlikely that the covariance is due to noise.