Significance Testing

Significance testing is the process of using statistics to determine how well data fit a particular pattern. Patterns can be very simple, such as "rates of obedience in the authority-present condition will be higher than rates of obedience in the authority-absent condition" or they can be more complex**.** The key to understanding significance testing is to understand that the strength of each pattern can be measured by a test statistic. What is a test statistic? First, let's look at what a statistic is. A **statistic** is a number representing some aspect of a group of numbers. For example, a **mean** is a statistic measuring "central tendency" in a group of numbers. The mean of the numbers 2, 7, and 9 is (2+7+9)/3 = 6. Another common statistic is **standard deviation**, which describes how spread out a group of numbers are. The numbers 2, 7, and 9 have a standard deviation of 3.6, while the numbers 4, 7, and 8 have a standard deviation of 2.1. These values reflect the fact that 2, 7, and 9 are more spread out than 4, 7, and 8. Both the mean and the standard deviation are statistics because they are numbers describing properties of groups of numbers.

Test Statistics

A **test statistic** is a number that represents the *strength of a pattern* in a group of numbers. It is a statistic because it reflects some property of a group of numbers, but unlike most statistics, it is sensitive to a particular pattern. There are many test statistics, and each is sensitive to a different pattern, but in general, test statistics get farther from zero when their pattern is present.

Test statistics are very useful for two reasons: 1) simplification: they enable researchers to express the strength of a pattern with a single number, and 2) falsifiability. If a hypothesis is **falsifiable**, then it is possible to test the hypothesis in a way that either supports or discredits the hypothesis. It is possible to use a test statistic to make a decision about whether or not a pattern is present. To understand how this is done, we must first discuss the concept of statistical significance.

Statistical Significance

Let's say that you are a detective and you get reports of a crooked gambling establishment. People are betting on a coin toss in which they gain one dollar for every heads and lose one dollar for every tails. The allegation is that the coin toss is somehow rigged to come up tails more than heads. This allegation becomes your **experimental hypothesis**: the pattern you expect to find. If your hypothesis is supported, you will charge the establishment with a crime.

The first decision you need to make is, how confident do I need to be before I conclude that the coin toss is rigged? If you observe a thousand coin tosses, find 538 tails, and compute that 538-or-more-tails would occur less than 1 in 100 times with a fair coin, would you feel confident charging the organization with a crime? What if you find 549 tails, which would occur less than 1 in 1000 times with a fair coin? The probability that you establish (e.g., 1 in 100, 1 in 1000) to make your decision of whether your hypothesis is supported or not supported is called the **alpha level**. In most psychological research, where the consequences of being wrong are not severe, the alpha level is set at either 0.05 or 0.01, which corresponds to either a 1 in 20 or 1 in 100 chance of being wrong: saying the coin is rigged when really it is fair and the string of tails was just bad luck. If you are conducting research where the consequences of being wrong are severe, such as saying a drug is safe when it might be deadly, you would set your alpha level lower, to perhaps 0.0001. So, your first step is to set the alpha level of your test.

Step two is to find a test statistic. You need a number that will reflect the strength of the pattern you suspect. In this case, the pattern you are observing is "number of tails" and you can

use that number as your test statistic. High values will indicate a stronger pattern. Let's say you observe 1000 coin tosses and 549 of those are tails. Your test statistic in this case is 549.

Step three is where the magic happens: computing the probability of your test statistic. More specifically, you want to know the probability of your test statistic, 549 tails out of 1000 coin tosses, occurring with a fair coin. If that probability is sufficiently low (below your alpha level), you will reject the idea that the toss is fair and conclude that it is rigged. But wait, do you want to know the probability of getting *exactly* 549 tails? It's probably pretty unusual to flip a coin 1000 times and get any one number, even 500. No, you want a sense of how unusual 549 is; how much it differs from the expected number of 500 tails for a fair coin. To answer this question, you could do 1000 coin tosses with a coin you know is fair and record the number of tails, then repeat that procedure ten million times, each time writing down the number of tails. Sometimes, you would get more than 549 tails. Most of the time, you would get around 500 tails. If you arranged the number of tails from smallest to largest and counted how often each result occurred, you would have a good sense of how unusual your obtained result of 549 is. If you drew a line at 548 tails and counted the number of results above 548, then divided that number by ten million, you would know the percentage of times that your result of 549 tails, or one more extreme than yours, occurs with a fair coin just by chance. That percentage would be the answer to your question: the probability of obtaining your test statistic, or one more extreme, just by chance. This number is called **statistical significance**. It is often written as the italicized lower-case letter *p*, or *p*-value, for probability value. If it is below your alpha level, you would conclude that the coin toss is rigged.

The example given in the previous paragraph presents the logic behind all significance testing. To test your hypothesis that the coin toss is *rigged*, you begin by assuming a *fair* coin and seeing how rarely your results would be obtained under that assumption. The hypothesis you want to test is called the experimental hypothesis, while the assumption of no effect is called the **null hypothesis**. When you generate a large number of test statistics under the null hypothesis (as you did by repeating your coin toss experiment with a fair coin ten million times), the resulting distribution of test statistics is called the **null distribution**. By computing the percentage of the null distribution that is equal to or more extreme than your test statistic, you find the *p*-value. If the *p*-value is less than the alpha level, you would "reject the null hypothesis" and accept your experimental hypothesis.

One part of the process above seems terribly unwieldy: running millions of replications of a test under the null hypothesis to create the null distribution. In practice, researchers rely on mathematical approximations of this process. There is a formula for determining the exact probability of a 2-outcome event (heads or tails) with known probability (a fair coin comes up heads 50% of the time) and a given number of trials (1000 coin flips). It is called the binomial function. Each test statistic has a similar function that statistics programs use to estimate the probability of obtaining a given test statistic.

What does statistical significance tell you? Does it tell you the probability that a result will be repeated? No. Does it tell you the probability that the null hypothesis is true? No[1]. Does it tell you that a result is important? No. Statistical significance only tells you that a result is *unlikely* given the assumption of no effect (the null hypothesis). Regrettably, many people use the *p*-value as an indication of importance. To correct this problem, researchers have recommended additional ways to evaluate test statistics, such as effect size.

[1] Statistical significance gives you the probability of a test statistic (T), given the null hypothesis (H_0). Written as a conditional probability, it is $p(T|H_0)$. In contrast, the probability that the null hypothesis is true is $p(H_0)$. These can be very different.

Effect Size

Whereas statistical significance (the *p*-value) tells you how unlikely a test statistic is, effect size reflects the *strength* of the pattern that the test statistic is designed to detect. As mentioned in the reading on the scientific literature in psychology, meta-analyses use effect size to combine the results of many studies because all test statistics (*t, r, F,* etc.) can be converted to effect size. Several statistics have been developed to express effect size, but three of the most popular are *r,* η^2 (eta squared), and *d*.

r. *r* is on the same scale as the Pearson correlation coefficient *r,* ranging from -1 to +1 with 0 indicating no effect. It is most useful if the variables you are comparing are both *continuous*, that is, capable of being expressed by numbers along a continuum. For example, exposure to media violence exists along a continuum from none to a great deal, and degree of aggression also exists along a continuum from mild to severe. Thus, you could express the effect of exposure to media violence on aggression using an *r* statistic. Research on the effects of media violence on aggression finds effect sizes in the *r* = 0.2 to 0.3 range (Anderson et al., 2003).

η². Eta squared is analogous to r^2 but is used for *t*-tests and ANOVA. Like r^2, it reflects the percentage of variability in the dependent variable that can be explained by the independent variable.

d. *d* is generally used to describe the strength of the *difference* between two groups. It represents the number of standard deviations that separate the means of two groups. A *d* of 1.0 indicates that two means differ by one standard deviation, and a *d* of 0.5 indicates that two means differ by half a standard deviation. For most people, those units don't mean much, so Cohen (1988) suggests thinking about *d* in terms of percentiles. Consider two groups, experimental and control. With a *d* of 0, the mean of the experimental group exactly overlaps with the mean of the control group. At *d* = 0.8, the mean of the experimental group would fall at the 79th percentile of the control group, meaning that 79% of the scores of the control group are below the mean of the experimental group. At *d* = 1.7, the mean of the experimental group is at the 95.5th percentile.

Cohen (1988) offers the cutoff values in the table below as general guidelines for evaluating the strength of effect sizes.

Table 1

Cohen's (1988) Recommended Interpretations of Effect Size

	r	*d*
Small	.10	.2
Medium	.24	.5
Large	.37	.8

Four Test Statistics

Psychology is mostly concerned with four test statistics, which are known by their symbols or letters: χ^2 (Chi Square, pronounced "kai square"), *r*, *t*, and *F*.

Chi Square (χ^2)

The **Chi Square test statistic** measures the degree to which one categorical variable is distributed disproportionately across one or more other categorical variables. A **categorical variable** is a nominal-scale variable, a variable that can take on values in distinct categories (such as "Egyptian" or "Lebanese") that cannot be expressed as points along a continuum (as would be the case with a variable such as income or temperature). When we talk about it being *distributed* across another variable, we mean that we are looking at how often particular combinations of two categorical variables occur. For example, let's say we're studying whether support for gun control varies by political party. We call 100 registered Democrats and 100 registered Republicans and ask each person whether they think there should be more restrictions on gun use ("for gun control") or fewer restrictions ("against gun control"). We obtain the data in Table 2.

Table 2.
Attitudes toward Gun Control by Political Party

		Gun Control	
		For gun control	Against gun control
Political Party	Democratic	70	30
	Republican	20	80

When we talk about one variable being distributed disproportionately across another variable, we mean that the distribution of gun control opinion is not in the same proportion for Democrats (70:30) as it is for Republicans (20:80). To make sense of nominal-scale data, it is often useful to convert them to percentages. For the data in Table 1, we see that gun control is supported by 70% of our Democratic sample but only 20% of our Republican sample. Our next question is whether the difference between 70% and 20% is a significant difference; that is, whether the difference between 70% and 20% is so large that it is unlikely to occur by chance. Given the data above, we would obtain a χ^2 value of 48 with 1 degree of freedom, which would be significant at $p < .001$. Thus, we could conclude that Democrats were significantly more likely to be for gun control than Republicans. In general, Chi Square is the test statistic you would use if you are comparing two percentages and testing whether they are different.

*Correlation (*r*)*

Correlation is a statistical procedure used to measure the degree of linear relation between two variables. A linear relation is one that can be well described by a straight line. Figure 1 shows a linear relation between temperature and aggression. In these hypothetical data, participants are placed into rooms of different temperatures and their aggression is measured. Each data point refers to the same person, and each data point represents two pieces of information about that person: the temperature of their room and their level of aggression. Correlation requires that both of your variables are on an underlying continuum: interval- or ratio-scale. Correlation is often reported as a lower-case italicized "r": *r*. *r* ranges from -1 to +1, with scores farther from zero indicating a stronger relation. The sign of *r* (whether it is positive or negative) indicates the direction of the relation between the two variables.

Positive correlations indicate that high scores on one variable (temperature) tend to be found with high scores on another variable (aggression), and low scores on one variable tend to be found with low scores on the other variable. In the data presented in Figure 1, the correlation is +0.94, a very strong correlation because it is close to +1.

Figure 1. Correlation of +0.94

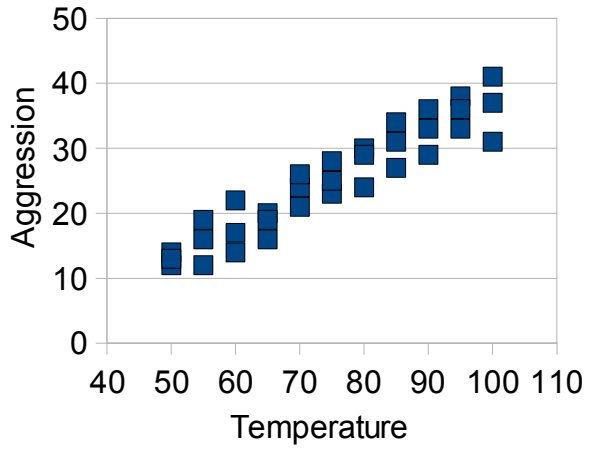


Figure 2. Correlation of +0.85

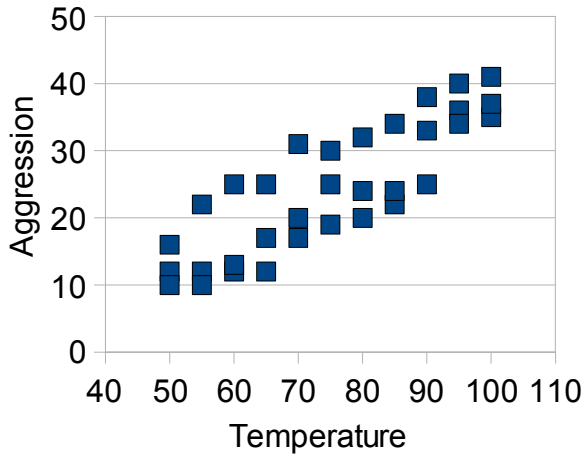
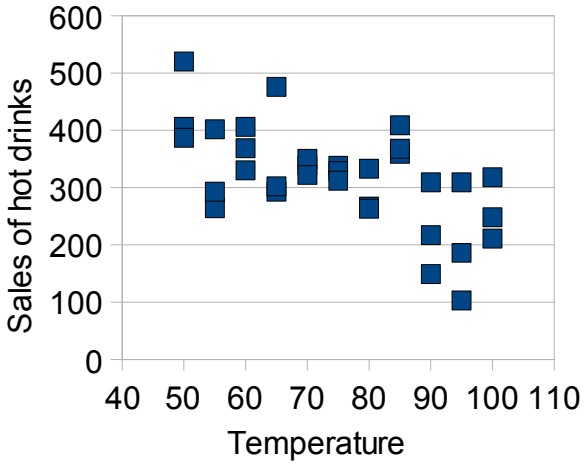


Figure 2 displays a correlation of r = +0.85. Compared to Figure 1, the data in Figure 2 are not as concentrated around a straight line; they are more spread out. As data points become less and less well-described by a straight line, the correlation decreases.

A **negative correlation** indicates that high scores on one variable tend to be found with low scores on the other variable. Figure 3 shows an example of a negative correlation: the relation between temperature and the sales of hot drinks. As the temperature increases, people buy fewer hot drinks: high values on one variable are found with lower values on the other.

Figure 3. Correlation of -0.61



One way to convert the correlation coefficient r into a more usable form is to square it, creating r^2. r^2 is the percentage of one variable that can be explained, accounted for, or predicted from a linear relation with the other variable. If the correlation between temperature and aggression is r = +.85, then r^2 = 0.72, meaning that temperature has accounted for 72% of aggression. This means that 28% is unaccounted for due to measurement error and alternative causes of aggression. Accounting for 72% of any human behavior would be a monumental achievement given the myriad causes of behavior. Many psychological correlations are around r = +0.3, which means that they account for only about 9% of behavior. Although it may not seem like much, being able to predict 9% of behavior can be an enormous advantage when you are dealing with large numbers of people, such as customers on eBay or visitors to Disneyworld.

Independent t-*test*

An **independent *t*-test** is used to compare the means of two separate (independent) groups of people. It is used when your independent variable is nominal-scale and has two levels (such as “experimental” and “control”) and when your dependent variable is on a continuum and is interval- or ratio-scale (such as “anxiety”). When the two groups have means

that are exactly equal, *t* is zero. As the means of the two groups diverge, *t* increases. The formula for *t* begins by subtracting the mean of one group from the mean of the other group. Thus, if the first group has a higher mean, *t* will be positive, but if the second group has a higher mean, *t* will be negative.

t gets larger as the magnitude of the difference between means increases, and gets smaller as the variability in scores within each group increases. This means that when scores are very spread out within each group, *t* will be smaller and you will be *less* likely to obtain a statistically significant result. If you were comparing the running times of two groups of Olympic athletes, one wearing a special shoe and one wearing a regular shoe, there would be very little variability within each group; each runner would probably be within a few hundredths of a second of each other. Under those circumstances, you would be more likely to find a statistically significant difference between the groups because the variability within groups would be small. Constrast that situation to a comparison of two groups of first-graders, one wearing the special shoe and one not. The running times for the first graders may differ by several minutes, spreading out the scores so much that it would not be possible to see the effects of the running shoe.

ANOVA

ANOVA stands for ANalysis Of VAriance. It is not a test statistic per se, but rather a statistical procedure. Variance consists of the differences between scores. Greater differences among scores is greater variance. Analysis refers to a “cutting into pieces” of the variance. In this case, our goal is to divide variance into two major pieces: the variance within each group and the variance between the groups. Let's say you survey ten members each of three fraternities and ask them how satisfied they are with college. The differences among the responses within each fraternity comprise the within-group variance. The differences between the means of the three groups comprise the between-groups variance. The test statistic for ANOVA is an ***F*** and it is the ratio of between-group variance to within-group variance. As between-groups variance increases and within-group variance decreases, *F* gets larger and is more likely to indicate a significant difference among the means of the groups.

ANOVA requires that your independent variable is nominal-scale, but unlike the independent *t*-test, ANOVA can handle more than two groups. Like the *t*-test, the dependent variable for ANOVA must be continuous and on an interval or ratio scale.

One-way ANOVA refers to ANOVA involving a single independent variable, such as which fraternity a participant is in. **Factorial ANOVA** involves more than one independent variable, such as a study investigating both the effects of being Greek or independent and the effects of gender. In that case, the experimental design would require a **two-way ANOVA** because there are two independent variables: Greek status and gender.

Summary of test procedures

The appropriate test is determined by the kind of data you have and the pattern you want to look for, as outlined in Table 3.

Table 3

Type of Data Required for Each Test

IV	DV	Test	Example
Nominal	Nominal	χ^2	Do a higher percentage of Republicans than Democrats favor the death penalty?
Interval	Interval	*r*	As temperature increases, are violent crimes more likely?
2 categories	Interval	*t*	Are males more aggressive than females?
2+ categories	Interval	ANOVA	Do four fraternities differ in their average GPA?

Type I and Type II Errors

Type I and Type II errors refer to two possible ways that the conclusions from statistical significance tests can be mistaken. Perhaps the easiest way to explain this is to begin with the chart in Table 4, which shows the relationship between the "true state of nature" and your conclusions. Recall that the null hypothesis is the assumption of no effect. In Table 4, we see that there are two ways to be correct: to reject the null when there is an effect; and not to reject the null when there is no effect. There are also two ways to make mistakes, and these are the Type I and Type II errors. A Type I error occurs when we reject the null hypothesis and claim there is an effect even though the true state of nature is that there is no effect. Thus, a Type I error occurs when *p* is below .05 just by chance, not because there is a real effect. Here we see another definition of statistical significance (the *p*-value): the risk of making a Type I error. A Type II error occurs when we find that *p* is above .05 and conclude that there is no significant relationship when, in fact, there is a real relationship and we just didn't detect it.

Table 4

Explaining Type I and Type II Errors

		Your conclusions	
		Do not reject null	Reject null, conclude H_e
True state of nature	Null is true (there really is no effect)	Say no effect when there is no effect (correct)	Type I Error: False positive (*p*-value = prob. of making a Type-I Error)
	Null is false (there really is an effect)	Type II Error: False negative	Say there is an effect when there is an effect (correct)

"Accepting" the Null Hypothesis?

The null hypothesis is a prediction of no effect. If your analysis yields $p > .05$ and you have set .05 as your alpha level, you cannot reject the null hypothesis. Why don't we say "accept" the null hypothesis rather than the more cumbersome "do not reject"? The reason is that "to accept the null hypothesis" would be to state as empirical fact that there is no effect, when all you really know is that you do not have sufficient evidence of an effect. It may be that an effect exists but your measures were not sensitive enough to pick it up or you did not have enough participants. For example, a researcher studying gender differences in preferences for romantic films collects data from 10 men and 10 women and finds that the mean preference

rating for romantic films is higher for women than for men, but this difference is not significant at $p = .08$. The researcher should not conclude that there is no gender difference (accept the null hypothesis), but rather should conclude that no gender difference was found.

(Statistical) Power

In the context of statistics, **power** refers to the probability of achieving a specific *p*-value, given a specified effect size, sample size, and variability in the dependent variable. Thus, power is a probability, a number between 0 and 1. Power increases as effect size and sample size increase and as variability in the dependent variable decreases. For example, consider a study that compares the means of two groups of 10 people each. Assume that the mean of one group is a 4 and the mean of the other group is a 4.5 and that the variability of each group is a standard deviation of 1.0. With a *p*-value cutoff of .05, the power of that study is only .201. That is, there is only a 20.1% probability that the study will achieve a *p*-value less than .05. That information might make you pessimistic about actually conducting the study because even if everything goes smoothly, you will only find $p < .05$ 20.1% of the time. If the number of people in each group is increased to 100, power in the study just mentioned jumps to 0.942. This shows that power is influenced by sample size and tells you why researchers often try to obtain as many participants as possible. If we keep sample size in each group to 10 and increase our estimate of the difference between the means to 1.0 (by predicting that the means will be 4.0 for one group and 5.0 for the other, for example), power jumps to 0.609. This shows that larger effects are easier to detect than smaller effects. A common application of statistical power is in computing the necessary sample size to have power at a certain level. For example, if you would like to have an 80% probability of obtaining $p < .05$, how many people would you need in your sample? To answer that question, you need reasonably accurate estimates of effect size and variability, such as from an earlier study. With that information, you could learn that to have 80% power, you would need only 20 participants, in which case you should do the study, or you could learn that you would need 2,000 participants, in which case you might give up.

Summary: Definitions of Statistical Significance

In this reading, we have encountered at least five ways to define statistical significance. All of them are equivalent and are just different ways of saying the same thing. Below is a summary of these definitions. Statistical significance is the probability of...

1. ...obtaining a test statistic as large or larger than the one you obtained, just by chance (as the result of a completely random process)
2. ...any observed difference between groups (or correlation between variables) being due to a random organization of the data
3. ...making a Type I (false positive) error
4. ...concluding that an effect exists, when in fact no effect exists
5. ...rejecting the null hypothesis (and accepting the experimental hypothesis), when the null hypothesis should not be rejected

References

Anderson, C. A., Berkowitz, L., Donnerstein, E., Huesmann, L. R., Johnson, J. D., Linz, D. G., Malamuth, N. M., & Wartella, E. (2003). The Influence of Media Violence on Youth. *Psychological Science in the Public Interest, 4*(3), 81-110.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.