Critical Reasoning 15 - Statistical Hypothesis Testing

with guest editor Katherine Eyal

Statistical Hypothesis Testing is a Set of Formal Decision Making Procedures Using the Method of Statistical Inferences to Test Statistical Hypotheses.

A **statistical hypothesis** is a statement about a population parameter that may or may not be true. A **statistical hypothesis test** then is a formal set of procedures by which to decide whether to accept or reject such an hypothesis. The statistical hypotheses that we are interested in are those that are testable based on direct measurement or the observation of a process that is modelled via a set of random variables. (Wikipedia: Statistical hypothesis testing)

While certain classical theories such as the theory of spontaneous generation may be discounted in the way Redi did by close experimental observation, this is not always possible when the constructs involved (such as IQ or conscientiousness) are not directly observable. In such cases we have to find a way of inferring or operationalising hidden or latent variables, usually by the use of a mathematical model. When making a statement about a population parameter it is seldom possible or even practical to test every member of the population to decide whether the result supports a particular hypothesis. Instead we proceed with the process of sampling. (See “Populations and Samples” Critical Reasoning 13)

Once we have drawn a sample from a population, as far as statistical hypothesis testing is concerned, we are usually interested in only one of two hypotheses at a time:

- The **null hypothesis** (H_0) or default hypothesis, usually that the observed or inferred sample is that of a purely random outcome

- The **alternative hypothesis** (H_1) that the observed or inferred sample results are of some non-random origin

Beginning with a research (or operational) hypothesis, the first step in proceeding with statistical hypothesis testing requires that we translate our hypothesis into a symbolic form that makes the null and alternative hypotheses explicit. Almost always, this involves a mathematical equality and/or inequality. Suppose, for example that we want to test whether a particular coin is biased. Our null hypothesis might be that if we flip the coin repeatedly, the proportion of heads (or tails) would be roughly half. The alternative hypothesis then would be that this proportion would be quite different. Together these hypotheses can be symbolised as,

H_0: P(H) = 0.5
H_1: P(H) ≠ 0.5

Note that, in this form, the alternative hypothesis is not specifying in which way a coin might be biased, only that is biased one way or the other. To test this we would require a **non-directional** or **two-tailed test** to determine whether there is a relationship between variables in either direction, *i.e.* either greater than or less than a certain value or range of values. If, n the other hand, we hypothesise that a particular coin is biased to land predominantly heads up, a **directional** or **one-tailed test** would be required to determine whether there is a relationship between variables in one direction only. Hence we would symbolise our alternative hypothesis as,

H_1: P(H) > 0.5

Rather than dwell on coin flipping, which is not very interesting or representative of real world research, we shall instead follow the example of Professors Kruger and Janeke (2012) of UNISA in which they propose the following simple, but more instructive research hypothesis:

UNISA students tend to have higher intelligence scores than the general population.

Such an hypothesis certainly appears reasonable because all universities have onerous admission requirements ensuring that only the most capable learners are admitted as students. On the other hand there are many highly intelligent people who are not or never were university students, perhaps because they never had the opportunity. Maybe they even outnumber those who did go to university, but that is not part of the hypothesis.

As it stands, this hypothesis is not sufficiently developed to undergo statistical hypothesis testing. Kruger and Janeke suggest that, “Part of the research process involves refining the research hypothesis until it suggests or implies the following:

- how the constructs involved will be measured
- what the research population is
- the nature of the relationship being investigated. (p. 71)

Before we decide how the constructs will be measured we must identify them. In this case we are looking at “group membership” which is a nominal variable taking on either of two exclusive the values of “UNISA student” or “member of the general population”. The other construct we need to measure is the dependant variable “intelligence”. Fortunately there are several reliable ways of operationalising intelligence as the outcome of an IQ test score, measured on an interval scale. As it happens IQ tests have been standardised so that they yield a population mean (μ) of 100 and a standard deviation (σ) of 15, so there is no need to calculate these parameters for the general population.

Four Types of Variables

Besides the distinction between random variables and observations referred to in Critical Reasoning 13, there are four types of measurement scales to take note of when categorising variables in statistics. **Nominal values** are used simply to name or label variables, such as “male” or “female”, “vaccinated” or “unvaccinated”. **Ordinal** scales are used to order or rank variables such as those used in customer satisfaction surveys, say from 1 for very unsatisfied to 5 for very satisfied. **Interval** scales on the other hand allow us to both order and quantify the difference between values, that comprise of units of equal size, such as the Celsius scale. **Ratio scales** moreover, allow us to do both of the above *and* allow us to calculate ratios because they have a defined zero value, such as for height or mass.

The table at the top of the next page summarises types of operation possible with each class of variable.

Clearly then, the research population is the population of UNISA students at any one time. So if we can calculate the mean IQ of UNISA students, we can compare it to that of the general population to test our hypothesis. The population mean is “the appropriate summary value to test when comparing the central value of two groups (or one group and a constant) when measures are on an interval scale, and we want to infer something about where the measurements for a particular group are concentrated.” (p. 74) According to our hypothesis, the mean IQ of UNISA students ought to be larger than that of the general population, which can be symbolised as,

H_1: μ > 100

The null hypothesis (H_0), on the other hand, can be symbolised as,

H_0: μ = 100

If it turns out that there is no difference between the mean IQ of the population of UNISA students and that of the general population then this is the value we should expect. Note that the null hypothesis is quite specific (μ = 100) whereas the alternative hypothesis only specifies a range of possible values (anything larger than μ). Also of note is that the null and alternative hypotheses must be mutually exclusive: they cannot both be true at the same time. Had the research hypothesis instead been:

> *UNISA students tend to have intelligence scores that differ from that of the general population*

we would not expect the population of UNISA students' mean IQ score to be specifically *either* higher *or* lower from that of the general population, only different in some way. By that difference we mean "just not equal to 100." Then our pair of hypothesis would be symbolised as,

H_0: μ = 100
H_1: $\mu \neq$ 100

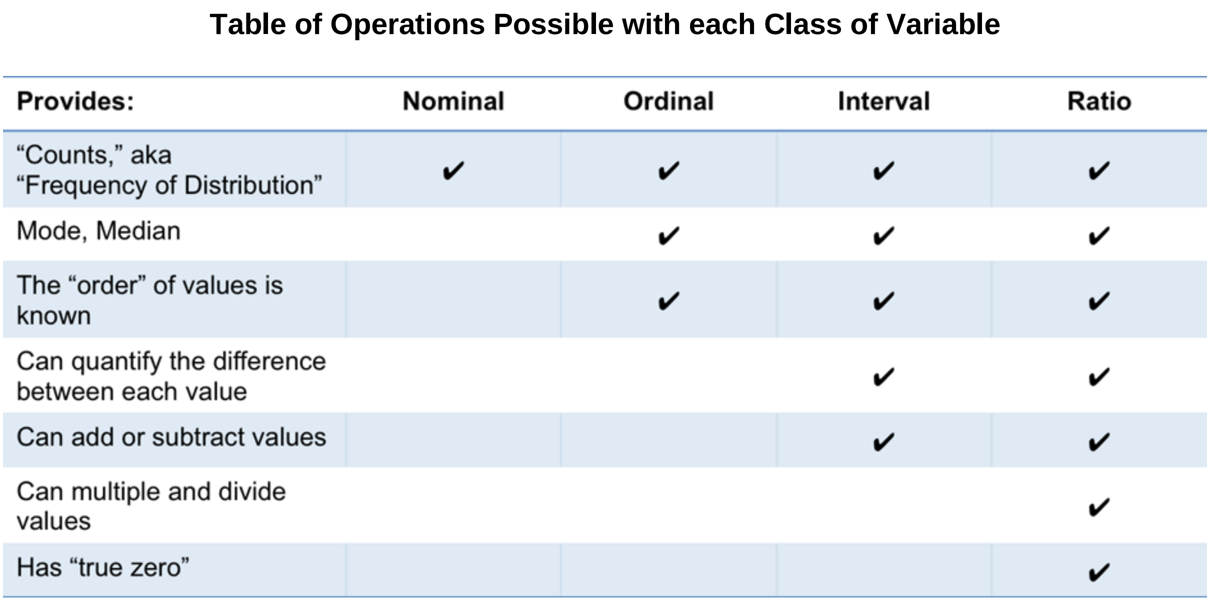


Table of Operations Possible with each Class of Variable

Provides:	Nominal	Ordinal	Interval	Ratio
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode, Median		✔	✔	✔
The "order" of values is known		✔	✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Found at: http://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/

Although it may be clear that the first, directional alternative hypothesis (H_1) above requires a directional or one-tailed test and that the second, non-directional alternative hypothesis (H_1) requires a non-directional two tailed test, it should be understood that what we are *actually* testing in both cases is the null or default hypothesis (H_0). On the assumption that the null hypothesis is true, we may reject it in favour of the alternative hypothesis (if the result of our test is "significantly" improbable), if not we fail to reject the null hypothesis. Although we decide which test to do based on H_1, it is actually H_0 that we are testing.

Let us return to our original pair of statistical hypotheses:

H_0: μ = 100
H_1: μ > 100

Recall that the general population parameter μ is already defined as 100. However in 2012 when Kruger and Janeke's study guide was published there were a whopping 336 286 students registered with UNISA, which is an impractically large population to measure in its entirely. If however we can find a way of testing a representative sample of the students we can calculate a sample mean ($\bar{x}$) as a proxy for the student population mean (μ). So although our hypotheses are stated in terms of parameters, we intend to test them in terms of sample statistics. (p. 76)

Kruger and Janeke's values are chosen for ease of calculation, therefore we will continue to follow their example.

> Suppose we select a random sample of 64 Unisa students. We then apply an intelligence test to each student and obtain his or her intelligence score. Let us assume that when we calculate the mean of the IQ scores of this random sample we find that the result is $\bar{x}$ = 104. (*l.c.*)

This looks like a promising result because our alternative hypothesis states that the student population mean IQ score (μ) is greater than 100 and the 104 we obtained is indeed greater than 100. However we cannot simply seize upon that the alternative hypothesis as true, because there is always the problematic question of sampling error whenever we try to infer a statement about a population from a sample statistic. On the other hand, we ought not to simply discard this result until we have performed further testing.

What if our result were $\bar{x}$ = 96 instead? The outcome could be the result of the sample chosen if the confidence intervals were wide enough to include 96. In that case, all we can say is that we can't reject H_0 but neither can we reject H_1 out of hand either because it is H_0 that we are actually testing. If, on the other hand, we were doing a *non-directional test* and we obtained a result *either side of* 100 this might imply that such a result favours the alternative hypothesis and that we ought to proceed with further testing.

Of course if we had obtained a value of $\bar{x}$ = 99 or $\bar{x}$ = 101, even if it were *statistically* significant, we would almost certainly not proceed with further testing because such a result would be *psychologically* insignificant *i.e.* there is nothing *psychologically* significantly different, as far as intelligence is concerned, between person *a* who scored 99 on an IQ test from person *b* who scored 100. For all we know the difference may very likely have been down to measurement error.

Quantifying the Probability of a Sample Result (Under the Null Hypothesis)

Kruger and Janeke's sample result of $\bar{x}$ = 104, just 4 points away from the general population mean of 100 appears, at first blush, to be close enough to have been the possible outcome of random errors of measurement; however we need to quantify this probability. In other words we need to calculate the probability of obtaining a sample result of $\bar{x}$ = 104 *under the assumption of the null hypothesis (H_0)* that the mean UNISA student population score for I.Q. is 100. This probability is given by a p*-value*, the very same p-values with which we became acquainted in Critical Reasoning 13.

It is paramount in inferential statistics that we understand what a p-value indicates (and what it does not): The **p-value** is a function of the observed sample results (a statistic) used for testing a statistical hypothesis. (Wikipedia: p-value) In the case of Kruger and Janeke's example, the p-value indicates the probability that the sample mean of 104 (or more) obtained was in fact derived by chance from a population with a mean of 100. Thus the p-value is a probability reflecting the likelihood of a sample result *under the assumption of the null hypothesis (H_0).* However, if the p-value is significantly improbable we might decide to reject H_0 and fall back on H_1 instead.

According to Kruger and Janeke, "The reason why we have to derive this p-value relative to an assumption that H_0 is true is that the probability that H_1 is true cannot really be calculated directly. It would be very convenient if we could calculate two p-values, one as if H_0 were true and another as if H_1 were true. Then one could simply choose the hypothesis that leads to the bigger p-value on the basis that this is the statement that is most probably true. Unfortunately, H_1 does not state an exact value of the population mean." (2012, p. 78) H_1 only says that it is larger, which means that the distribution of this range of possible means is infinite and therefore unmeasurable. Instead we proceed by our roundabout method of testing "under the null hypothesis" *i.e.* assuming that H_0 is true.

According to Kruger and Janeke, in determining the p-value we need to know about the general probability distribution of the means in order to make probability judgements about them. If you were fortunate enough to have worked through Critical Reasoning 13 you will remember that the Central Limit Theorem (CLT) provides us with just such information. In particular, recall that:

- Regardless of the shape, mean or standard deviation of the parent population, the distribution of the sampling means approaches a normal distribution as n increases. (In fact, it approaches very close to normal with an n of as low as 30.)

- The distribution of the sample means is described by the mean ($\mu_{\overline{x}} = \mu$) and its standard deviation is given by $\sigma/\sqrt{n}$.

The quantity given by $\sigma/\sqrt{n}$ is literally the **standard deviation of the sampling means** a.k.a. the **standard error** because it is an estimate of the size of the error we shall make if we use the mean of the sampling means ($\mu_{\overline{x}}$) as an estimate of the population mean (μ) . Because the standard error is so frequently used in statistics and in scientific reports it has its own special symbol: $\sigma_{\overline{x}}$

If we now apply the CLT to the information that we already have in hand for this example, namely that for IQ scores μ = 100, σ = 15 and that for our sample n = 64, then we get:

$$\mu_{\overline{x}} = \mu = 100$$

and

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{64}} = \frac{15}{8} = 1.875$$

Our next task is to standardise our score of $\overline{x}$ = 104 so that we can calculate the probability of obtaining the score that we did under the null hypothesis. In Critical reasoning 13 we learned how to convert any normally distributed variable x into a z-value of the standard normal distribution as follows:

$$z = \frac{x - \mu}{\sigma}$$

But we are not interested here in any particular raw datum, x; what we *are* interested in is converting the sample mean $\overline{x}$ into a z-score ($z_{\overline{x}}$) for which there is an associated probability that we can look up. We do this by replacing the symbols above with those that refer to the distribution of the sample mean $\overline{x}$, thus:

$$z_{\overline{x}} = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}}$$

We are now in a position to substitute the information that we have for our sample into this formula, thus:

$$z_{\overline{x}} = \frac{104 - 100}{1.875} = 2.133$$

Before we look up the associated probability, let us remind ourselves of the purpose of above calculation, namely that:

$$p(\overline{x} > 104) = p(z_{\overline{x}} > 2.133)$$

so that if we have the probability at right we also have the probability on the left. If you have not already downloaded or bookmarked a copy of the z-tables of standard normal probabilities, they are available here courtesy of the University of Florida. In this case we are looking for the lesser proportion ($\overline{x}$ > 104) to the right shaded, in the figure below. Therefore we need consult the first of the two z-tables. It does not matter that our table gives the probabilities for lesser proportion to the left because the graph is symmetrical about the centre z = 0. Therefore looking up the probability associated with z = -2.133 to the left will give us the same p-value of 0.0166 as would looking up z = +2.133 to the right. If you prefer to use a statistical package rather than a table you will obtain the same p-value, so long as you select the one-tailed output option.

Before we decide what our result means please look at the figure below (redrawn from Kruger and Janeke's figure 3.1 in their 2012 UNISA study guide.) What this figure does represents is the distribution of the mean, not the distribution of IQ scores within the general population. The position of the sample mean ($\overline{x}$ = 104) is shown here beneath (but not on) the x-axis because this is what we want to compare to the above distribution, not that of IQs in general. (p. 80)

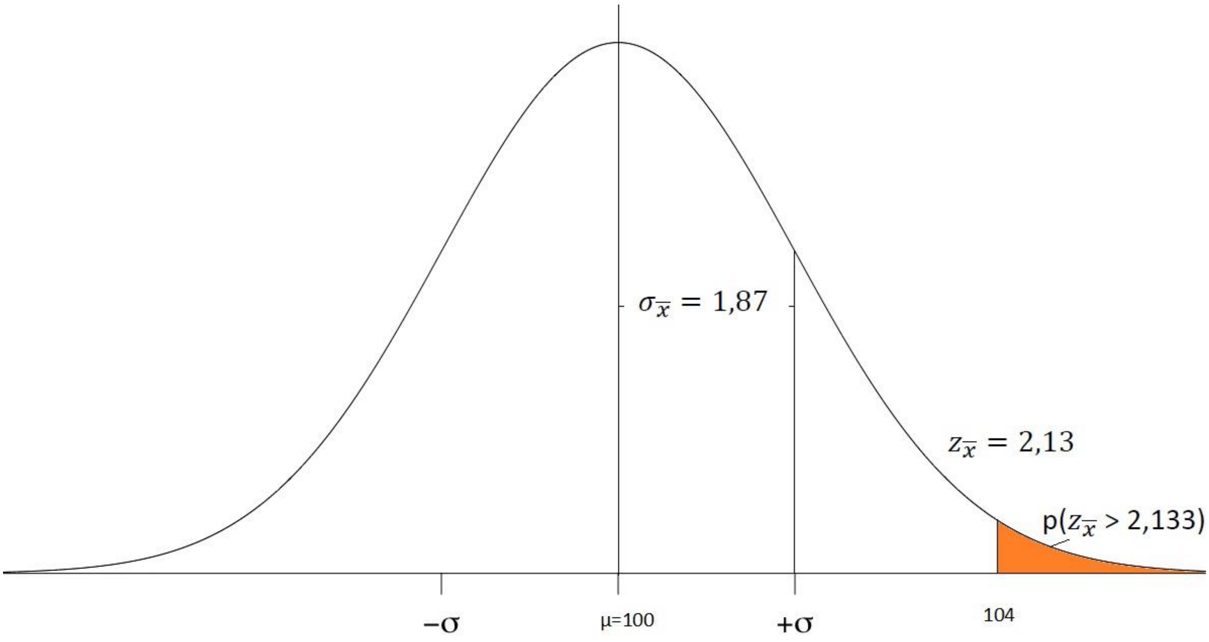


Figure: Distribution of the Mean of IQ Scores (not to scale)

The probability of $p(z_{\overline{x}}$ > 2,133) is given by the shaded area to the right. Can you identify the other quantities labelled on the sketch? If you have understood the concept of the standard error ($\sigma_{\overline{x}}$) you will see that its value corresponds to the width of one standard deviation (σ). Only just a little more tricky to appreciate is that $z_{\overline{x}}$ value of 2.13 is a standardised score that represents how many standard deviations our sample mean of 104 is away from the population mean of 100 shown beneath (but not on) the x-axis.

What are we to make of the p-value of 0,0166 (about 1 in 60)? Is the probability of obtaining the sample mean that we did, under the null hypothesis so low that we are prepared to reject it in favour of the alternative hypothesis? Ideally we should have chosen a cut-off value before we undertook out research, because otherwise the temptation is to cherry pick the results that we like by accepting them after the fact.

The name that we give to this cut-off point is the **significance level** ($\boldsymbol{\alpha}$), which is the probability at which a researcher(s) will be willing to reject the null hypothesis, given that the null hypothesis is true. α is thus a chosen, rather than calculated value. Typically in the psychological sciences we pre-select an α of 0.05 or 0.01 which "…specifies the maximum risk that we are willing to take of making an error if we reject the null hypothesis." (p. 83) The value of α that we choose depends very much upon the sort of research we are undertaking. If we were simply conducting market research into brand preferences for varieties of biscuits, we would be unlikely to agonise over a low α. If, on the other hand we were investigating the safety of a vaccine intended to be administered to babies, an α of 0,01 would be far too large, especially considering that most infant vaccine campaigns involve millions of doses in multiple countries and that even one baby coming to harm that could have been prevented is an unacceptable risk.

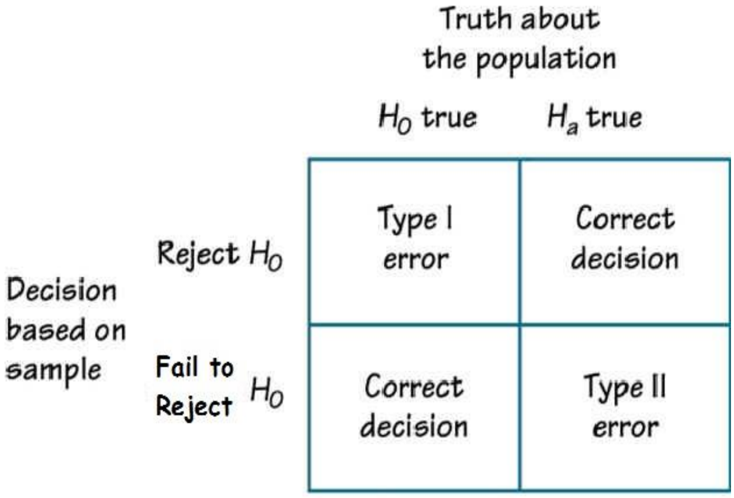
As it is, our p-value of 0.0166 is well below the α of 0.10 corresponding to the 90% confidence level that is commonly accepted in the Social Sciences; therefore we decide to reject the null hypothesis and fall back on the alternative hypothesis instead. Similarly, had we been investigating the non-directional alternative hypothesis, our p-value for the two-tailed test would have been precisely twice that of the one-tailed test *i.e.* 2 x 0.0166 = 0.0332. This value is still below the α of 0.05 commonly used in the psychological sciences in general, corresponding to a 95% confidence level; therefore we would still be justified in rejecting the null hypothesis, and falling back on the two tailed alternative hypothesis instead.

Type I and Type II Errors

Set out as it is above may give the false impression that statistical hypothesis testing is a purely mechanical procedure for deciding between competing hypotheses in a way that could be programmed into a spreadsheet, without the need for rational deliberation. Recall, however that "the p-value represents the probability that the null hypothesis is true: that the effect we see in our observation is due to chance effects like measurement error." The smaller the p-value the more likely we are to reject the null hypothesis. Put another way the p-value is a direct measure of the probability that the null hypothesis is being *mistakenly* rejected, whereas all along it has *actually been true*. Such a mistake, that of "rejecting the null hypothesis when in fact it is true - is referred to as a ***Type I error***." (p. 84 original emphasis)

Setting α at a particular level: If an α of 0,05 corresponds to a confidence level of 95% and setting α to 0.01 corresponds to a confidence level of 99%, why do we not just always set α to the smallest possible value in order to avert any possible type I errors and thus insure the maximum possible confidence in our research? The short answer is that we would risk mistakenly rejecting a great many significant results. Suppose our p-value comes in larger than our relatively low α and we decide not to reject the null hypothesis *when in fact it was false* all along, what we *should have done is rejected* it in favour of the alternative hypothesis. This mistake is referred to as ***Type II error***.

The probability of failing to reject the null hypothesis when it is false is given by β and understandably there is a relation between β and α: The smaller the level of α the higher β. In other words if we want to avoid a type I error we should set α low; however the smaller α the higher the likelihood that we will make a type II error, β. Unfortunately we cannot calculate the value of β directly because the alternative hypothesis does not specify a mean for the population, only a range such as μ > 100.



		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Fail to Reject H_0	Correct decision	Type II error

The Relationship Between Decicions about Hypotheses Based on a Samle vs. the Truth or Error Concerning the Population

Consider an intruder alarm system by way of analogy. The alarm system has two basic states: the *default state*, in which the system is on but not detecting anything and the *alarm state* in which a sensor is tripped causing the siren to sound. If we allow for anthropomorphism, the alarm has two basic, mutually exclusive “beliefs” or “hypotheses” about the zone it is monitoring, the default hypothesis (H_0) that nothing is being detected and the alternative hypothesis (H_1) that something is being detected. As any owner of an alarm system will know, there are two challenges involved in calibrating such a system: sensitivity and false alarms. On the one hand, the alarm system must be sensitive enough to detect even the stealthiest intruder; on the other hand it must not sound the siren every time the family cat jumps up on the couch or a twig taps on the window pane. Failing to detect an actual intruder, (a **false negative**) corresponds to a type II error, while falsely detecting a non-existent intruder, (a **false positive**) corresponds to a type I error.

In the case of an alarm system, we can mitigate the rate of false negatives by dialling up the sensitivity of the sensors, however we cannot always set them to be maximally sensitive because then we are going to be literally alarmed by too many false positives. Consider the “boy who cried wolf”. And while the problem or false positives may simply be an annoyance in some circumstances, a false positive pathology test for the presence of cancerous cells, for example may result in the dangerous and unnecessary surgical removal a healthy organ such a prostate or even an entire breast. The decision, therefore as to what we are least prepared to tolerate, type I *vs.* type II errors, must be the outcome of careful deliberation, not simply a calculation.

According to Kruger and Janeke "The ability of a statistical test to detect a significant relationship between variables when such a relationship does in fact exist, is referred to as its **power**." (p. 85) As the power of a statistical test increases, so the chances of making type II errors decreases. This is given by the formula:

$$power = 1 - \beta$$

Even if we cannot directly calculate β, we can appreciate the relation.

Factors Influencing Statistical Power (Sensitivity)

- The choice of level of significance (α) for a particular statistical test, is a statement of how improbable a positive result must be, assuming the null hypothesis to be true, for us to reject the null hypothesis. Increasing α from say 0,05 to 0,1 increases the chance of rejecting the null hypothesis (*i.e.* of obtaining a statistically significant result) when the null hypothesis is false, thereby reducing the chance of a type II error (false negative.) However this also increases the chance of rejecting the null hypothesis when it is actually true, hence increasing the chance of a type I error (false positive.) (Wikipedia: Statistical power)

- Sampling error can be caused by inappropriate sampling techniques, measurement error or due to external variables. By choosing appropriate sampling techniques and more reliable tests as well as controlling for or eliminating external variables in the research design effectively reduces the standard error of the sampling distribution of the mean. (p. 86)

- The choice of statistical test also influences the power of statistical testing procedures. "In general, **parametrical statistical tests** (based on assumptions about the distribution of populations or sampling estimates of them…) tend to be more powerful than equivalent **non-parametric tests**…" (which make no such assumptions) (p. 86)

- Since the sample size also determines the level of sampling error in a test result, increasing the sample size is by far the easiest way of boosting the statistical power a of a statistical test; however it is not always the most *efficient* measure. One does not necessarily need to choose the maximum achievable sample size for a given power due to the diminishing returns on investing in obtaining ever larger samples. (Wikipedia: Statistical power)

- On the other hand, when the sample size is large, even small effects can have statistical significance. According to the law of large numbers, "on average the result obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed." In a large sample therefore, the error variance (that portion of the variance that is due purely to randomness) ought to decrease. This can result in smaller p-values for sample effects that appear to be insignificant, which might persuade us to reject the hull hypothesis. (p. 86)

Kruger and Janeke (p. 86 - 87) provide a hypothetical example of how a psychologically insignificant result (a difference of just 2 points between the mean IQ of UNISA students *vs*. the general population) can quickly tend to a p-value limit of zero as the sample size increases, first from n = 10 then to n = 100 and then to n = 1000. Clearly, we don't want to seize upon every statistically significant result just because we are impressed by large sample sizes, when effect size (below) should be an important consideration.

The advent of the internet has allowed researchers to conduct studies with extremely large sample sizes (in the tens of thousands) where "p-values go quickly to zero, and solely relying on p-values can lead the researcher to claim support for results of no practical significance." In their ironically named article, "Too Big to Fail: Large Samples and the p-Value Problem" Lin and colleagues describe why an enormous sample size and a vanishing p-value on their own can be problematic to meaningless. Chief among their recommendations to mitigate the problem is to understand the effect size. (Lin *et al*. 2013)

Effect Size

Effect size is a statistical quantification of the strength of the relationship between two variables of a population, or a sample-based estimate of the same. In the case that we are looking at difference between two groups, the difference between the two means $\overline{x_1} - \overline{x_2}$ would be a direct measure of effect size; however this does not take into account the variability in the population. By dividing this difference by the **pooled standard deviation** ($\boldsymbol{\sigma_p}$) of both groups we can arrive at a standardized measure of effect size, which together with the sample size completely determines the power of a statistical test. (Wikipedia: Statistical power)

In a two group situation, **Cohen's $\boldsymbol{d}$**, as this measure is known is given by:

$$d = \frac{\overline{x_1} - \overline{x_2}}{\sigma_p}$$

where $\overline{x_1}$ and $\overline{x_2}$ are the means of the two groups respectively and σ_p is the pooled standard deviation. A Cohen's d of 1 or greater indicates that the two groups' means differ from one another by more than 1 standard deviation - quite substantially. Kruger and Janeke, (p. 88) offer the following rule of thumb by which to interpret effect size for a given Cohen's d:

d ≈ 0,2 : effect size 'small'
d ≈ 0,5 : effect size 'medium'
d ≈ 0,8 : effect size 'large'

Cohen's d can also be used to estimate the sample size required for statistical testing: a low Cohen's d indicates that a larger sample size is required and *vice versa*. For completeness sake we reproduce the formula for the pooled standard deviation (σ_p) for two independent samples; though it does not form part of the syllabus for undergraduate courses in humanities.

$$\sigma_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

(Wikipedia: Effect size)

Glass' Δ meanwhile can be used when treating one group as a control and testing various treatment groups against the control. It can also be used as an estimator of effect size using only the standard control group (s_2 below), thus:

$$\Delta = \frac{\overline{x_1} - \overline{x_2}}{s_2}$$

Glass' Δ has the advantage that effect sizes do not differ under equal means and different variances; however, "under a correct assumption of equal population variances a pooled estimate for σ is more precise." (Wikipedia: Effect size)

General Strategy for Addressing Problems Involving Hypothesis Testing

Kruger and Janeke provide the following guidance when addressing research scenarios involving hypothesis testing, such as those in the tasks below. With only minor edits then, they advise:

A. Formulate the research or operational hypothesis

Try to formulate this hypothesis so that the following aspects are implied:

- the constructs between which a relation is being postulated
- the nature or rule of the relation
- the research population
- how constructs are being measured
- the research design

B. Translate the research hypothesis into statistical hypotheses and test these on the basis of data from sample(s)

Step 1: State the statistical hypotheses and set the value of α

Step 2: Select a random sample(s) and calculate appropriate statistics. Look at the data from a common-sense, non-statistical point of view. What does it seem to tell you?

Step 3: Select an appropriate test statistic. (We have learned of only one type so far.) Calculate the appropriate p-value (*e.g.* decide if this should be a directional or non-directional p-value) and compare it with the α value. If the p-value is smaller than α, reject H_0 and accept H_1. If not, do not reject H_0.

C. Draw a conclusion regarding the research/operational hypothesis (2012 p. 89)

In real world examples it would also be necessary to examine effect size and report your findings alongside.

Further Example

Suppose you are a high school Biology teacher at a boarding school for girls. All 25 girls in your class decide *en masse* to become vegetarian after watching a disturbing documentary about the appalling conditions of local slaughter houses. The Matron accordingly instructs the kitchen staff to prepare nutritious and appetising vegetarian meals with suitable meat substitutes. After one month you notice several girls regularly asleep in your class and you wonder whether they might not be getting sufficient iron from their new diet, which might account for their fatigue.

Since you do not know whether the girls may in fact be getting more than enough or less iron in their new diet or whether your lessons might have become more boring, you opt for a non-directional test. You decide to operationalise your hypothesis using Haemoglobin (Hb) levels as a proxy for iron levels in the body because determining them involves only a finger prick test, which yields a result in grams per Litre accurate to 1 decimal place.

You look up the population mean and standard deviation of Hb levels for girls age 12 - 18 years and discover that the mean Hb level is 14.0 g/dL with a standard deviation of 1.0 g/dL. You hypothesise that for your class:

H_0: Hb μ = 14.0 g/dL

H_1: Hb $\mu \neq$ 14.0 g/dL

and that a single sample z-test would be appropriate. You decide to test your hypothesis at the 95% confidence level, therefore you set your α at 0.05.

During the following practical each girl uses a sterile lancet to draw a single drop of blood from her middle finger and places it in the haemoglobin testing unit. Each girl then writes her result on the white board in front of the class. During the exercise you notice that two of the girls have recorded a Hb level of just below 12 g/dL. You suspect they may be borderline iron-deficient anaemic and you make a mental note to speak to the Matron about supplementation for them.

When everybody's result is up on the board, the girls work out the mean and standard deviation of the class using only their scientific calculators. They get a mean Hb level of 13,6 g/dL and a standard deviation of 0,6 g/dL.

At first blush this looks like a plausible result. The fact the girls are all eating the same meals and residing at the same altitude could account for their narrower standard deviation of Hb levels compared to the general population. Their lower mean level of Hb might also be accounted for by their vegetarian diet; however unfortunately you did not do a before and after study, so you might never know. Never the less, you decide to press on with the calculations.

Because you are only interested in the population of girls in your class rather than them as sample, you decide that you can dispense with calculating the standard error and proceed to calculate a z-score directly.

$$z = \frac{x-\mu}{\sigma} = \frac{13.6 - 14.0}{0.6} = -0.667$$

However we need to put this value in context. Because H_1 is a non-directional or two-tailed hypothesis you must divide your α between the left and right tails i.e. 0.05 ÷ 2 = 0.025. Looking at the right tail first, we use the second z-table, available here to find the z-score associated with our α of 0.025 *i.e.* (1 - 0.025 = 0.975) which is z = 1.96. However as this is a two-tailed test, you also need to consider the left tail z = - 1.96. These "two tails" can be visualised this as follows:

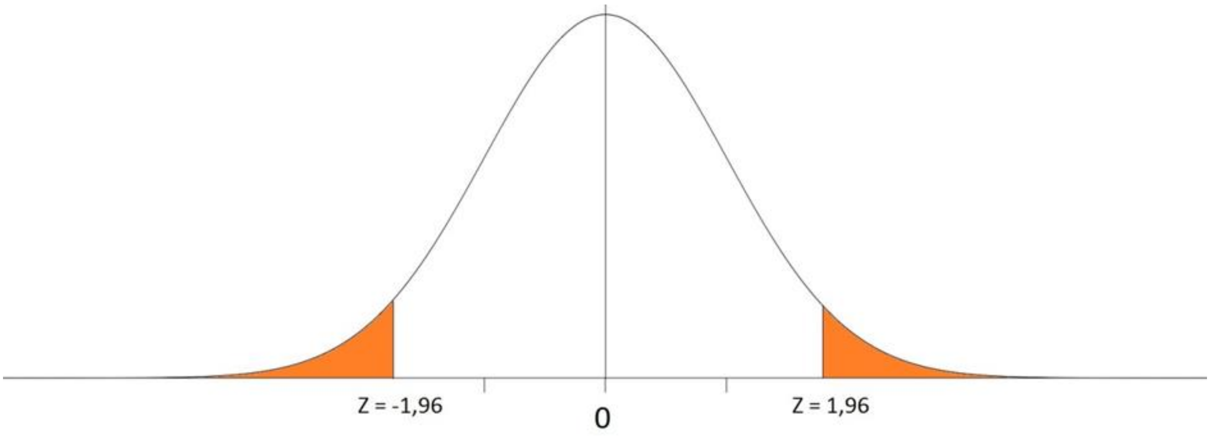


Figure depicting the two areas of interest (shaded) in the example above

Your z-value is much less than the 1,96 required at the 95% confidence level. In fact a z-value of this magnitude is associated with a two tailed p-value 0.5054 which is not significant even looking at just the left tail. Therefore you decide that you cannot simply reject the null hypothesis. Nor is there any need to calculate the size of an effect that you have failed to detect.

You share your results with the class tell them that as far as their iron levels are concerned, they appear to be a very typical bunch of teenagers and that henceforth there will be no excuse for sleeping in class. You suggest repeating the exercise after six months to ascertain whether there are any long term changes to their iron levels as an outcome of their vegetarian diet.

The Truth about Single Sample z-Tests

Having dwelled so long on the z-test for single samples, it may surprise you to know that it is almost never used in scientific research: for one, we seldom know the mean and standard deviation of any real world population, unless they have been rigged them that way, as in IQ. Secondly, in Science we are interested in conducting experiments or observing "natural" experiments. (See Critical Reasoning 12) This involves comparing a control group, which we endeavour to keep constant, with one or more experimental groups, in which we allow just one or as few as possible variables to fluctuate. This is clearly not possible in a single sample setup; although we shall meet several more complex yet familiar statistical tests in the study units ahead.

The reason for working through the single sample z-test above is that all the subsequent statistical tests that form part of the undergraduate humanities syllabi, with one notable exception, resemble the z-test procedurally and in the way their test statistics are calculated. Moreover all of their distributions can be mapped onto that of the z-distribution, so that we do not require separate tables for each type of statistical test. Although such tables do exist, it is easier to use a statistical package to return the value(s) you seek, so long as you are quite sure what sort of test is involved.

E.g. Is it for independent or dependent samples? Is it for equal or unequal variance? Do you want a one, or two tailed p-value? More of which in Critical Reasoning 17.

Task:

Go back to the section entitled "General Strategy for Addressing Problems Involving Hypothesis Testing". Kruger and Janeke offer some sound and very important recommendations in only a few words. Try to motivate each of their points. Why do you think they were at pains to include each of them? What to do think might be the result of ignoring any of their recommendations?

Feedback:

The headings "A.", "B." and "C." and their order is cardinal. Recall:

A. Formulate the research or operational hypothesis

B. Translate the research hypothesis into statistical hypotheses and test these on the basis of data from a sample(s)

C. Draw a conclusion regarding the research/operational hypothesis

Sadly, many students, in their enthusiasm to launch a research project, begin by collecting and analysing as much data as possible in the hope that this will suggest a suitable hypothesis. This however is both wrongheaded and mildly intellectually dishonest. On a practical level one can end up with a jumble of data, just days before a project is due and still be on the lookout for an hypothetical mould in to which to press one's "findings".

Surely nobody, you would think, would be so daft as to begin with point C (draw a conclusion *etc.*) and then go in search of evidence; however that is what we all do from time to time when exercising the confirmation bias. (See Critical Reasoning 06) We may decide what we already believe, based on a hunch, and then selectively take note of corroborating instances while ignoring disconfirming cases.

With the advent of the internet and access to ubiquitous "information" one can "corroborate" virtually any thesis (or its antithesis,) by cherry picking confirming instances and selectively ignoring contrary evidence, which is why it is all the more important to formulate and formally test your hypotheses as part of your *own* research.

The specific recommendations can be addressed point by point.

- We must to try to formulate an hypothesis so that the constructs between which a relation is being postulated are implied, because most of the constructs dealt with in the Humanities are latent (hidden) and must be made theoretically explicit.

- Next we must try to formulate an hypothesis so that the *nature* of the relation is implied because this will determine the type of statistical test that is appropriate. *E.g.* Is the relation directional or non-directional? Is there more than one sample involved? If so, are they dependant or independent *etc.*? Failing to take this advice into account may result in one calculating an inappropriate, distorted or even meaningless test statistic.

- When formulating an hypothesis we must also try to make the criteria for membership of the research population explicit. An hypothesis about “people,” “pets” or “computers” for example, is likely to be highly ambiguous because there are so many *different* and sometimes *incompatible* ways of defining membership of such groups. What such categories lack is specificity that can be clearly stated.

- Any research hypothesis should explicitly state or clearly entail how the constructs are to be measured and on what scale(s): nominal, ordinal, interval or ratio. The instrument by which such a scale is to be measured must also be described if necessary. *E.g.* an anthropometer requires no explanation, while a structured interview or survey questionnaire requires a detailed explanation. Also one needs to have some idea about **construct validity**: Does a construct consistently represent what it is claimed to represent and if it is quantifiable, does it measure what is claimed to measure? (See Critical Reasoning 26 for further discussion.) Famous historical examples of constructs that represent absolutely nothing include Phlogiston, the Luminiferous Ether and Élan vital. Contemporary examples include “The Secret Law of Attraction” and “The Millionaire Mind-Set”.

- Our hypothesis must be formulated in such a way that our research design is made explicit. Carrying out appropriately designed research is what allows the research hypothesis to be empirically tested either directly or via statistical hypothesis testing.

Step 1:

Although the above points may take up a chapter length discussion in any dissertation, the actual scientific hypothesis should be a succinct, unambiguous, contingent statement, without modifiers (such as somewhat, very, quite, almost *etc.*) and whose truth conditions are known.

Translating the research hypothesis into a statistical hypothesis should be relatively straightforward. As we saw above, this almost always involves a mathematical inequality for directional tests or an equality for non-directional test. Selecting an appropriate value for α is trickier. Either by convention or through lack of imagination some scientists always set α at 0.05; however we cannot adopt such a “one size fits all” approach. The smaller we set the value of α the less likely we are to make a Type I error *i.e.* of incorrectly rejecting the null hypothesis when in fact it is true. However, then we are then more likely to make a Type II error *i.e.* of failing to reject the null hypothesis when in fact it is false.

One question we might ask ourselves in deciding on an appropriate α is “What sort of mistake can we least afford to make, a false positive (detecting an effect that isn’t there) or a false negative (failing to detect an effect that is present)?” Recall that in statistical hypothesis testing, a false positive represents a type I error while a false negative represents a type II error. Therefore if we are more concerned with avoiding false positives we should choose a lower α; however if we are more concerned with avoiding false negatives we should choose a slightly larger α; though not so large that avoiding type I errors becomes our new overriding concern.

Step 2:

"Sampling is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population." (Wikipedia: Sampling (statistics)) In order to achieve this, our sample should be as representative of the parent population as is fit for our purpose. Although statistical sampling is a vast topic, we wish to stress just a few points here: We have all been told that larger samples are better; however quantity alone does not always prevail over quality. A very large biased sample is much more likely to be distorted than a modest sample in which every member of the population has an equal chance of being randomly sampled. Although several sources of everyday bias were discussed in Critical Reasoning 06 under the heading of "Heuristics", sampling bias has the potential to be even more deceptive in formal research.

Once we have collected our samples, captured our data and calculated the appropriate statistics it is worthwhile looking at the data from a common-sense, non-statistical point of view, especially if the statistics in question were generated automatically by a statistical package. One question we might ask ourselves is whether such output is realistic or even possible given the range of values sampled. If our hypothesis is directional, the next logical question to ask is whether the result is in the right direction. If not, no further testing may be required; however we cannot simply reject H_1 out of hand since it is H_0 that we are actually testing. A further question is whether we understand or indeed need all of the descriptive statistics so generated. (Statistical packages typically generate more output than we require. *E.g.* Do we really need to report the kurtosis of our sample distribution when all we might be after, at this stage, is the mean and standard deviation?)

Step 3:

Choosing an appropriate test based on the assumptions made by each test is essential, however we have yet to learn of a variety of tests beyond the z-test covered in this study unit. Suppose however that we have chosen the correct test for our purposes and have calculated the correct test statistic and its associated p-value. (Remember, we may have to halve the p-value if we are conducting a directional test and the programme we are using always returns a two-tailed p-value.)

Next, comparing the p-value with our α allows us to decide whether or not to reject the null hypothesis. If the p-value is smaller than α, then we reject the null hypothesis and accept the alternative hypothesis by default. If not, we do not reject the null hypothesis. Note: The p-value is a direct measure of the probability of our result under the assumption of the null hypothesis (H_0) therefore we never test the alternative hypothesis directly.

Finally, we need to draw a conclusion regarding or research hypothesis. If we had to operationalise the hypothesis, we need to translate the statistical hypothesis back into the form of the research hypothesis, report it and stop there. However, because most academic journals are reluctant to publish negative or null results, what some researchers do, especially if their p-value is only slightly larger than their α, is to go back and tweak their α to a slightly larger value and report a positive result. Although such a practice is technically dishonest, the immense pressure most academics are under to "publish *or* perish" makes this a common practice.

References:

KRUGER, P. & JANEKE, H. C. (2012) *Psychological Research - Study Guide for PYC3704*. UNISA Department of Psychology

LIN, M., *et* al (2013) Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research, Articles in Advance, pp. 1–12*. Available at http://www.galitshmueli.com/system/files/Print%20Version.pdf

z-tables originally provided by the University of Florida at http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf

Mirror at http://philosophy.org.za/uploads_other/Ztables.pdf

The next Critical Reasoning study unit concerns deductive systems.