ProcessGPS

Guaranteed Project Success

A Layperson's Guide to Hypothesis Testing

By Michael Reames and Gabriel Kemeny – ProcessGPS

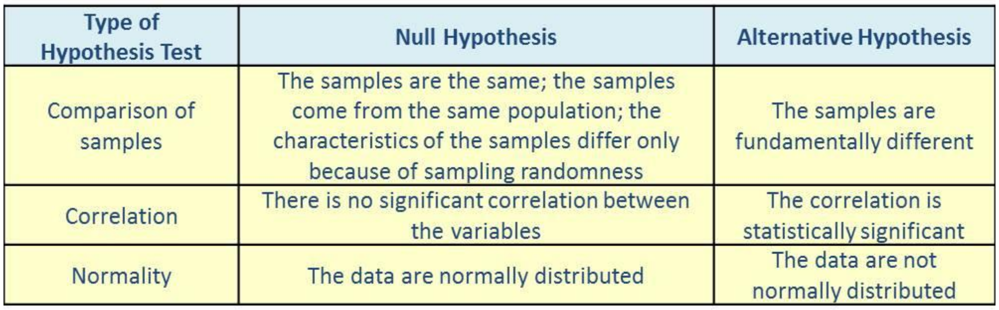
In a recent Black Belt Class, the partners of ProcessGPS had a lively discussion about the topic of hypothesis testing. Sadly, many individuals (yes, even Black Belt candidates) start perspiring when confronted with such daunting topics as p-values, sample size determination, t-tests, analysis of variance (ANOVA), and Chi-Square analysis. Some even have difficulty pronouncing them, much less performing them!

In this article we set out to make the concept of hypothesis testing logical and comprehensible for those who are convinced they'll never figure it out. We'll approach the topic qualitatively (i.e., not statistically) by means of two familiar analogies: the American justice system and airport passenger security screening. We look at the set-up of these systems, the assumptions being made in each, and the way that they qualitatively mirror the foundations of the statistically rigorous topic of hypothesis testing.

The Development of Hypothesis Testing

The idea of hypothesis testing is based on the scientific method, where a conscientious experimenter identifies a "hypothesis," a guess as to what is true about the nature of something. She may think that something is true, but as a responsible scientist, she knows that she must test that hypothesis in some rigorous way. She has a sample of data available, one that is appropriately sized and randomly selected so as to minimize the chance of bias. This sample is therefore very likely to be representative of the total population. Now she performs a test of the data to draw some conclusions. The decision she makes is either a correct decision or an incorrect decision, based on the true state of nature (which we assume that we cannot know with certainty).

The "guess" that the investigator makes has an important designation. In hypothesis testing, we speak of the null hypothesis and an alternative hypothesis. Although in theory a scientist could set up these hypotheses any way she chooses, the accepted convention is to make the null hypothesis the unexciting conclusion ("*Null is Dull*"); that things are as they usually appear in nature. In general, the investigator desires that the data prove that in fact things are not happening as they occur in nature, and that an important cause-and-effect relationship is being revealed. For instance, in Figure 1 we see some typical hypothesis statements:



Type of Hypothesis Test	Null Hypothesis	Alternative Hypothesis
Comparison of samples	The samples are the same; the samples come from the same population; the characteristics of the samples differ only because of sampling randomness	The samples are fundamentally different
Correlation	There is no significant correlation between the variables	The correlation is statistically significant
Normality	The data are normally distributed	The data are not normally distributed

Figure 1- Hypothesis Statement Examples

Now, if the scientist were not conscientious, but simply heedless, she could make a pure guess as to the applicability of her hypothesis to the true state of nature. An uneducated guess - in other words, a random choice – may good or bad, depending on the true state of nature. Random decisions are not considered an attractive way to proceed (see Figure 2).

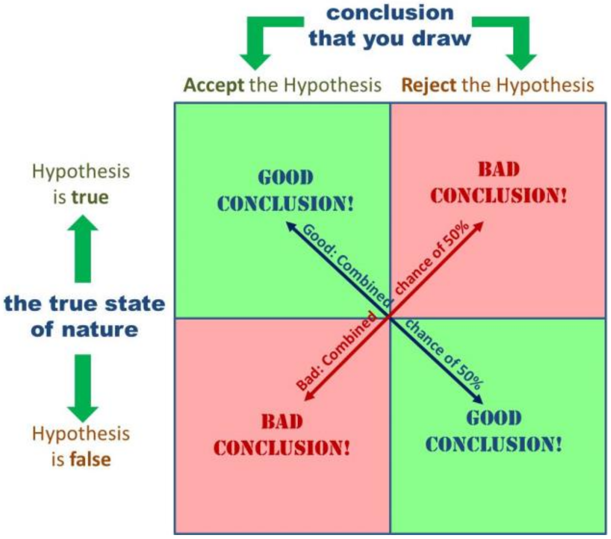


Figure 2 - Making decisions based on random chance (not a good method)

Recalling that we cannot know with certainty the true state of nature but are trying to make an educated guess, we perform a hypothesis test. The nature of a hypothesis test is that it allows us to minimize the chance of an error. In any hypothesis test there are two types of errors:

1. **Error Type 1**: Rejecting a hypothesis which is actually true in nature

2. **Error Type 2**: Accepting a hypothesis when it is actually false in nature

Both types of error are bad. If we could minimize the chance of making either type of error, that would be ideal. We can decrease the chance of either or both types of error by increasing our sample size. If the chance (probability) of Type 1 and Type 2 error were small, we would have the situation depicted in Figure 3.

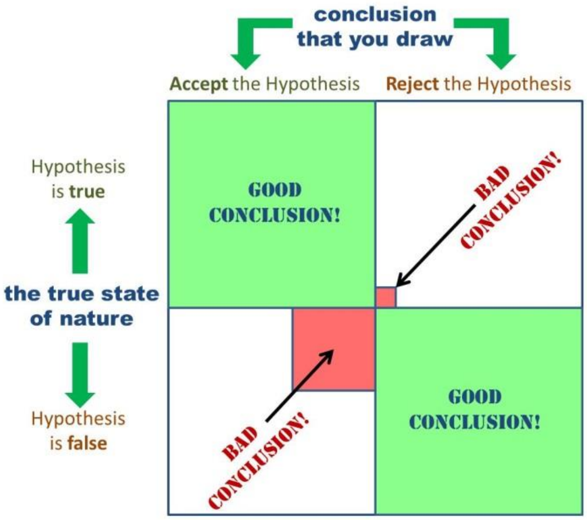


Figure 3 - A better way to make decisions: minimizing the chance of bad conclusions

This, then, is the essence of hypothesis testing: to maximize the chances that the conclusions we draw are in accordance with the true state of nature. While there are lots of statistics and theory to back this up, this is what the Black Belt does when he tries to make a decision.

The two red quadrants are both bad conclusions, but their characteristics are different. We call rejecting a true hypothesis to be a **Type 1 error**; and when we accept a false hypothesis we call this a **Type 2 error** (see Figure 4)

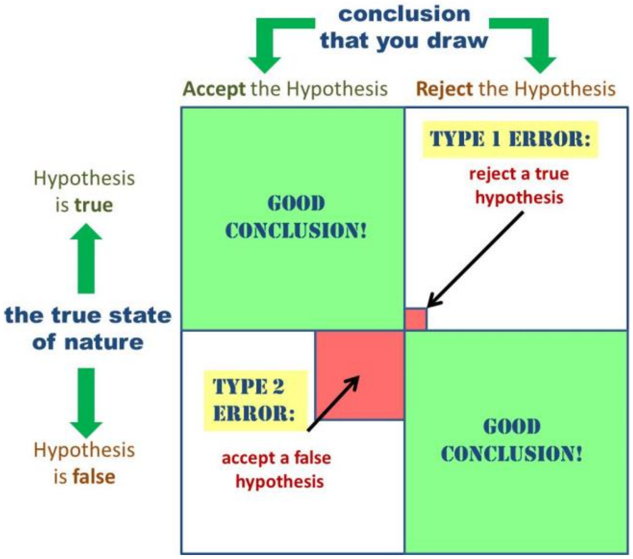


Figure 4 - Type 1 errors and Type 2 errors defined

Hypothesis Testing Applied to the American Justice System

Understanding the nature of many individuals to shy away from the background statistics, let's proceed to cement our understanding of hypothesis testing by applying it qualitatively to the American justice system. Its embodiment is that an accused person is "presumed innocent unless proven guilty." Furthermore, in a criminal trial, the standard for guilt is "beyond a reasonable doubt."

In terms of the null versus the alternative hypothesis, we have the following:

Type of Test	Null Hypothesis	Alternative Hypothesis
American Justice System	The defendant is innocent (Presumption of innocence until [unless] proven guilty)	The defendant is guilty (beyond a reasonable doubt)

These principles are neatly embodied within the statistics of hypothesis testing. For example, let's say that the accused really is innocent of the criminal charge. Many times only the accused knows whether this is actually true or false. The prosecuting attorney doesn't know; the judge doesn't know; the jury doesn't know. The presumption of innocence means that the jury starts out believing that the "true state of nature" is that the defendant is innocent (upper left quadrant of Figure 4). Therefore, the trial is a means of determining if there is sufficient evidence to prove that the defendant is guilty "beyond a reasonable doubt" (lower right quadrant). In any case, we'd ideally like to minimize the chance of any error: convicting an innocent person (Type 1 error) or freeing a guilty person (Type 2 error). Understanding the interconnected nature of errors, however, we place a very strong preference on truly minimizing the chance that we would ever convict an innocent person. The standard of "beyond a reasonable doubt", combined with the requirement of a unanimous decision by the jury, ensures that we have minimized this chance.

On the other hand, while it is certainly not a desirable outcome to allow a guilty person to go free, we accept this as the price that American society pays for the potential tragedy of putting an innocent person in prison (or worse, putting an innocent person to death).

Thus, our justice system skews the conclusion that we draw in favor of minimizing the chance of a Type 1 error: convicting an innocent person. This cannot by nature be a statistical test. For example, we can never say with 95% [or 99%] certainty that the system will not convict an innocent person. Instead, we say qualitatively "beyond a reasonable doubt" and require a unanimous jury decision. And thereby the system allows for a greater probability that a guilty person will fail to be convicted (not a good outcome, but tolerated).

In a criminal case, the prosecution must prove the guilt of the accused. The Defense need not prove innocence, but only needs to place a reasonable doubt in the mind of one juror. Thus, as an interesting and important aside, note that we don't declare the person "innocent." Indeed, only somebody who was at the scene of the crime could declare innocence with certainty. And by design, nobody on the jury could have been there. Thus, we declare the person "not guilty."

In hypothesis testing, similarly, we don't say that we've proven that the true state of nature is as it was hypothesized. Instead, we say that we have failed to disprove this hypothesis; another way of saying this is that we don't have sufficient evidence to prove otherwise. The practical aspect is that we must consider it is as likely to be true as not; but still this cannot be proven.

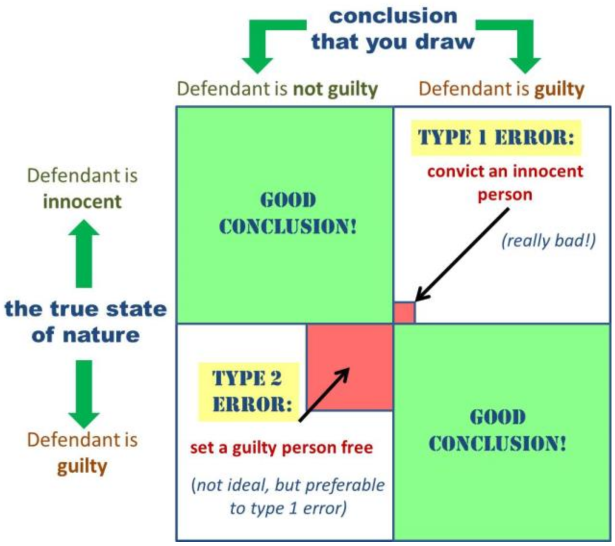


Figure 5 - The American System of Jurisprudence as a form of Hypothesis Test

Hypothesis Testing Application to the Airline Passenger Screening System

Now let us contrast the justice system with another process familiar to many of us: screening passengers prior to boarding commercial aircraft. In this process, the Transportation Security Administration (and, presumably, the flying public) takes the following as a principled stand: That the screening process will minimize the possibility of allowing an armed (or dangerous) passenger on board an aircraft. In terms of the hypothesis testing model, this can be defined as Error Type 2: accepting a hypothesis as true when in reality it is not true. The conclusion that we draw here is that a person is not dangerous while the true state of nature is that he/she is concealing a weapon.

In terms of the null and alternative hypotheses, we have:

Type of Test	Null Hypothesis	Alternative Hypothesis
Airline Passenger Screening	The passenger is unarmed and not dangerous (as are most)	The passenger is armed and potentially dangerous

This situation presents an interesting twist on the American justice system; with passenger screening, the TSA would rather detain a passenger who is not dangerous in favor of increasing the odds of detaining <u>every</u> single passenger who constitutes a threat to airline safety. Figure 6 shows this visually: minimize the chance of a Type 2 error, even at the expense of increasing the chance of a Type 1 error. Of course, if you are the passenger being screened out (and in truth are not posing any danger), then it’s a terrible inconvenience to you individually; but it would be much worse were The TSA to minimize the chance of individual inconvenience for the increased likelihood of allowing a dangerous person on board.

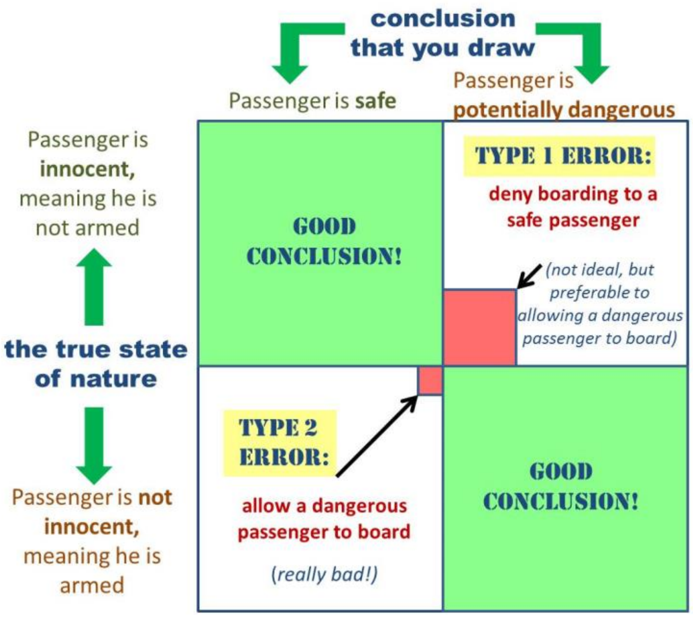

conclusion that you draw
Passenger is safe
Passenger is potentially dangerous
Passenger is innocent, meaning he is not armed
the true state of nature
Passenger is not innocent, meaning he is armed
GOOD CONCLUSION!
TYPE 1 ERROR:
deny boarding to a safe passenger
(not ideal, but preferable to allowing a dangerous passenger to board)
TYPE 2 ERROR:
allow a dangerous passenger to board
(really bad!)
GOOD CONCLUSION!

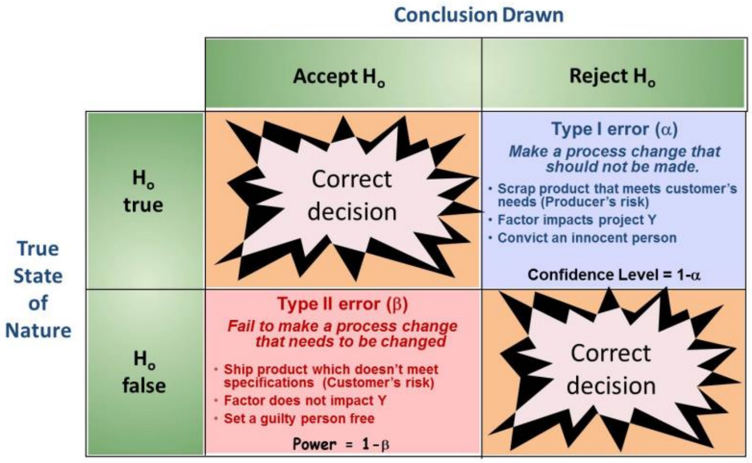
Figure 6 - The TSA Airline Passenger Screening Process as a form of Hypothesis Test

How Black Belts Use Hypothesis Testing in Lean Six Sigma Projects

In Lean Six Sigma problem solving, the Black Belt seeks root causes for a particular process issue, and wants to statistically validate that something occurring in the process (a particular factor) affects the output (a response) adversely, and thus fails to meet customer requirements consistently. Validation of this cause-and-effect relationship is achieved if the Black Belt has enough statistical evidence to reject the null hypothesis (that there is no correlation between the factor and the response). Obviously, in making the decision to reject the null hypothesis, the Black Belt must ensure that the probability of Error Type I is small: in other words, that there is a high level of confidence that the alternative hypothesis is in fact the truth.

On the other hand, if the probability of Error Type II is significant (statistically, we call this low power of the test), the Black Belt may miss the opportunity to identify a strong correlation between a factor (x) and a response (y) when it really exists (see Figure 7).

In order to decrease the probability of Errors Type I and II (i.e., to increase both the confidence level and power of the test), the Black Belt must increase the sample size of the data collected to validate root causes.



		Conclusion Drawn	
		Accept H_o	Reject H_o
True State of Nature	H_o true	Correct decision	Type I error (α) *Make a process change that should not be made.* • Scrap product that meets customer's needs (Producer's risk) • Factor impacts project Y • Convict an innocent person Confidence Level = $1-\alpha$
	H_o false	Type II error (β) *Fail to make a process change that needs to be changed* • Ship product which doesn't meet specifications (Customer's risk) • Factor does not impact Y • Set a guilty person free Power = $1-\beta$	Correct decision

Figure 7 – Types of Errors in Hypothesis Testing

Summary

This article only touches the surface of the very interesting field of hypothesis testing. The Black Belt, or Project Leader in a process improvement effort, learns the statistics behind the theory of hypothesis tests, and learns the rich variety of types of hypothesis testing that can assist in validating root causes. Nevertheless, the analogies drawn in this article are very useful in grounding the trainees in the theory. We hope that it has been useful for you also.

ProcessGPS
Guaranteed Project Success

Gabe Kemeny and Michael Reames
Partners and Co-Founders, ProcessGPS, LLC

gabe@ProcessGPS.com (801) GPS-0606
michael@ProcessGPS.com (484) 2020-GPS

Website: www.ProcessGPS.com

Your guaranteed success defines ProcessGPS' success