How to Choose the Level of Significance: A Pedagogical Note

Kim, Jae

31 August 2015

How to Choose the Level of Significance:
A Pedagogical Note

Jae H. Kim[*]
Department of Economics and Finance
La Trobe University, Bundoora, VIC 3086
Australia

Abstract

The level of significance should be chosen with careful consideration of the key factors such as the sample size, power of the test, and expected losses from Type I and II errors. While the conventional levels may still serve as practical benchmarks, they should not be adopted mindlessly and mechanically for every application.

Keywords: Expected Loss, Statistical Significance, Sample Size, Power of the test

1. Introduction

Hypothesis testing is an integral part of statistics from an introductory level to professional research in many fields of science. The level of significance is a key input into hypothesis testing. It controls the critical value and power of the test, thus having a consequential impact on the inferential outcome. It is the probability of rejecting the true null hypothesis, representing the degree of risk that the researcher is willing to take for Type I error. It is a convention to set the level at 0.05, while 0.01 and 0.10 levels are also widely used. Thoughtful students of statistics sometimes ask: "How do we choose the level of significance?" or "Can we always choose 0.05 under all circumstances?" Unfortunately, statistics textbooks do not provide in-depth answers to this fundamental question.

---

[*] Tel: +613 94796616; Email address: J.Kim@latrobe.edu.au

Students should be reminded that setting the level at 0.05 (0.01 or 0.10) is only a convention, based on R. A. Fisher's argument that one in twenty chance represents an unusual sampling occurrence (Moore and McCabe, 1993, p.473). However, there is no scientific basis for this choice (Lehmann and Romano, 2005, p.57). In fact, a few important factors must be carefully considered when setting the level of significance. For example, the level of significance should be set as a decreasing function of sample size (Leamer, 1978), and with a full consideration of the implications of Type I and Type II errors (see, for example, Skipper et al., 1967[1]). Although a good deal of academic research has been done on this issue for many years, these studies are not readily accessible to the students and teachers of basic statistics. In this paper, I present several examples that I use in my business statistic class at an introductory university level. To improve the readability, the references for academic research are given in a separate section.

2. Sample size (Power and Probability of Type II error)

Let α represent the level of significance which is the probability of rejecting the true null hypothesis (Type I error); and β the probability of accepting the false null hypothesis (Type II error), while $1 - \beta$ is the power of the test. For simplicity, we assume that the expected losses from Type I and II errors are identical, or the researcher is indifferent to the consequences of these errors. This assumption will be relaxed in the next section. Under this assumption, it is reasonable to set the level of significance as a decreasing function of sample size, as the following example shows.

Suppose $(X_1, \ldots, X_n)$ is a random sample from a normal distribution with the population mean μ and known standard deviation of 2. We test for H_0: $\mu = 0$ against H_1: $\mu > 0$. The test statistic is

---

[1] Reprinted in Morrison and Henkel (1970, p.160).

$Z = \dfrac{\overline{X}}{2/\sqrt{n}} = 0.5\sqrt{n}\,\overline{X}$, where $\overline{X}$ is the sample mean. At the 5% level of significance, H$_0$ is

rejected if Z is greater than the critical value of 1.645 or $\overline{X}$ is greater than $2(1.645)/\sqrt{n}$. Note

that the Z statistic is an increasing function of sample size or the critical value for $\overline{X}$ is a

decreasing function of sample size. This means that when the level of significance is fixed, the

null hypothesis is more likely to be rejected as the sample size increases. Let $\mu = 0.5$ be a value

of substantive importance under H$_1$. Table 1 presents $\beta = P(Z < 1.645| \mu = 0.5, \sigma = 2)$, along with

the power and critical values for a range of sample sizes. The upper panel presents the case

where α is fixed at 0.05 for all sample sizes, while the lower panel presents the case where α

is set as a decreasing function of sample size and in balance with the value of β. The upper

panel shows that, when the sample size is small, the value of β is unreasonably high compared

to $\alpha = 0.05$, resulting in a low power of the test. When the sample size is large, the power of

the test is high, but it appears that α is unreasonably high compared to β. For example, when

the sample size is 300, $\alpha = 0.05$ is 12.5 times higher than the value of β. In this case, a negligible

deviation from the null hypothesis may appear to be statistically significant (see Figure 1 and

the related discussion).

From the lower panel, we can see that, by achieving a balance between the probabilities of

committing Type I and II errors, the test enjoys a substantially higher power for nearly all cases.

For example, when the sample size is 30 with $\alpha = 0.05$, the power of the test is only 0.20.

However, if α is set at 0.35, the power of the test is 0.65. When n $= 300$, setting $\alpha = 0.015$

provides a balance with the value of β. In addition, the sum of the probabilities of Type I and

II errors $\alpha + \beta$ is always higher when α is fixed at 0.05. In general, a higher power of the test

can be achieved when α is set as a decreasing function of sample size and in balance with the

value of β (see also Figure 2 and the related discussion).

Figure 1 presents two scatter plots (labelled A and B) between random variables Y and X, both with sample size 1000. The two plots are almost identical, showing no linear association between the two. In fact, Y and X are independent in Plot A; but in Plot B, they are related with the correlation of 0.05. Regressing Y on X in Plot A, the slope coefficient is 0.04 with t-statistic 1.23 and p-value 0.22, indicating no statistical significance at any reasonable level. In Plot B, the regression slope coefficient is 0.09 with t-statistic 2.82 and p-value 0.004. In this case, although X and Y are related with a negligible correlation, the regression slope coefficient is statistically significant at 1% level of significance. That is, the t-statistic and p-value give a wrong impression or illusion that there is a strong association between the two variables, which can mislead the researcher into a belief that the degree of linear association is highly substantial (see further discussion in Section 4 with reference to Soyer and Hogarth; 2012). Considering the large sample size, a much lower level of significance (such as 0.005 or 0.001) should be adopted, which will deliver the decision of a marginal or no statistical significance (see further discussion in Section 4 with reference to Johnson; 2013).

3. Expected losses from Type I and II errors

Students should be reminded that Type I and II errors often incur losses which affect people's lives, such as ill health, false imprisonment, and economic recession (see, for example, Ziliak and McCloskey, 2008). The level of significance should be chosen taking full account of these losses. Setting α to a conventional level for every application may mean that the researcher does not explicitly consider the consequences or losses resulting from Type I and II errors in their decision-making.

Example: Testing for No Pregnancy

Consider a patient seeing a doctor to check if she is pregnant or not. The doctor maintains the belief that the patient is not pregnant until a medical test provides the evidence otherwise. The doctor is testing for the null hypothesis that the patient is not pregnant against the alternative that she is. Suppose two tests for pregnancy are available: Tests A and B. Test A has a 5% chance of showing evidence for pregnancy when the patient is not in fact pregnant (Type I error); but it has a 20% chance of indicating evidence for no pregnancy when in fact the patient is pregnant (Type II error). Test B has a 20% chance of Type I error and a 5% chance of Type II error. The consequence of Type I error is diagnosing a patient as pregnant when in fact she is not; while that of Type II error is that the patient is told that she is not pregnant when in fact she is. Test A has four times smaller chance of making the Type I error; but it has four times more chance of making the Type II error. If the doctor believes that Type II error has more serious consequences than Type I error since the former risks the lives of the patient and baby, Test B ($\alpha = 0.2$, $\beta = 0.05$) should be preferred as it is a safer option.

Example: Hypothesis Testing as a Legal Trial

Hypothesis testing is often likened with a trial where the defendant is assumed to be innocent (H_0) until the evidence showing otherwise is presented. The jury returns a guilty verdict when they are convinced by the evidence presented. If the evidence is not sufficiently compelling, then they deliver a "not guilty" verdict. In the court of law, there are different standards of evidence that should be presented, as Table 2 shows. For a civil trial, a low burden of proof (preponderance of evidence) is required since the consequences of wrong decisions are not severe. However, for a criminal trial where the final outcome may be the death penalty or imprisonment, a tall bar (beyond reasonable doubt) is required to reject the null hypothesis. This means that the legal system is using different levels of significance (or critical values) depending on the consequences of wrong decisions. That is, the level of significance for

"preponderance of evidence" may be as high as 0.40; and that for "clear and convincing evidence" can be as low as 0.01. To meet the level of "beyond reasonable doubt", the level of significance should be much lower (say 0.0001) which places a tall bar for a guilty verdict.

Example: Minimizing Expected Losses

Consider a business analyst testing for the null hypothesis that a project is not profitable against the alternative that it is. Suppose for the sake of simplicity that $P(H_0$ is true$) = P(H_1$ is true$) = 0.5$. Let L_1 and L_2 be the losses from Type I error and Type II error, then the expected loss from wrong decisions is $0.5\alpha L_1 + 0.5\beta L_2$. Table 3 presents these values using two different scenarios of (L_1, L_2). In the first scenario, the loss from Type II error is five times higher than that of Type I error, i.e., $(L_1, L_2) = (20, 100)$; and the opposite is the case for the second scenario. When the analyst chooses α of 0.05, the corresponding value of β is assumed to be 0.25; and if the analyst sets α at 0.25, and it is assumed to be 0.05.

Suppose the analyst wishes to minimize the expected loss. Then, when $(L_1, L_2) = (20, 100)$, $(\alpha, \beta) = (0.25, 0.05)$ should be chosen since it is associated with a lower expected loss. Since the loss from Type II error is substantially higher, a higher α level should be chosen so that a lower probability is assigned to Type II error. Similarly, under $(L_1, L_2) = (100, 20)$, $(\alpha, \beta) = (0.05, 0.25)$ should be chosen. This illustrative example demonstrates that when the losses from Type I and II errors are different, the level of significance should be set in consideration of their relative losses.

4. Summary of Selected Academic Research

Leamer (1978; Chapter 4) makes the most notable academic contribution to this issue by

presenting a detailed analysis as to how the level of significance should be chosen in consideration of sample size and expected losses. He introduces the line of enlightened judgement, which is obtained by plotting all possible combinations of (α, β) given the sample size. In the context of the example in Table 1, the line of enlightened judgement is all possible combinations of (α_i, β_i) where $\beta_i = P(Z < CR_i \mid \mu = 0.5, \sigma = 2)$ and CR_i is the critical value corresponding to α_i. Leamer (1978) shows how the optimal level of significance can be chosen by minimizing the expected losses from Type I and II errors, and demonstrates that the optimal significance level is a function of sample size and expected losses.

Figure 2 presents three lines of judgement corresponding to the (α, β) values in Table 1 when the sample size is 10, 50, and 100. Given the sample size, the line depicts a trade-off between α and β. As the sample size increases, the line shifts towards the origin as the power increases. The green line represents the case where the level of significance is fixed at 0.05. The (α, β) values in the upper panel of Table 1 correspond to the points where this line and the lines of enlightened judgement intersect. The 45-degree line connects the points where the value of $\alpha+\beta$ is minimized for each line of enlightened judgement (assuming $L_1=L_2$), which correspond to the (α, β) values in the lower panel of Table 1. Kim and Ji (2015) also discuss the line of enlightened judgement with an example in finance.

In earlier studies, a number of authors argue that the level of significance should be chosen as a function of sample size and expected losses. Labovitz (1968)[2] argues that sample size is one of the key factors for selecting the level of significance, along with the power or probability of Type II error (β) of the test. Kish (1959)[3] states that when the power is low, the level of

---

[2] Reprinted in Morrison and Henkel (1970, p.168).
[3] Reprinted in Morrison and Henkel (1970, p.139).

significance much higher than the conventional levels may be more appropriate. Winer (1962) also states that "when the power of the tests is likely to be low …, and when Type I and Type II errors are of approximately equal importance, the 0.3 and 0.2 levels of significance may be more appropriate than the .05 and .01 levels" (cited in Skipper et al., 1967)[4]. Ziliak and McCloskey (2008, p.8) state that "without a loss function, a test of statistical significance is meaningless", arguing that hypothesis testing without considering the potential losses is not ethically and economically defensible.

Keuzenkamp and Magnus (1995, p.20) conduct a survey of economics papers and report that "the choice of significance levels seems arbitrary and depends more on convention and, occasionally, on the desire of an investigator to reject or accept a hypothesis". They also note that Fisher's theory of significance testing is intended for small samples, stating that "Fisher does not discuss what the appropriate significance levels are for large samples". Kish (1959)[5] asserts that (at the conventional level of significance) "in small samples, significant, that is, meaningful, results may fail to appear statistically significant. But if the sample size is large enough, the most insignificant relationships will appear statistically significant". From a recent survey of papers published in finance journals, Kim and Ji (2015) report that the conventional levels of significance are almost exclusively used in finance research, despite the widespread use of large or massive sample size. Gigerenzer (2004, p.601) argues that "the combination of large sample size and low p-value is of little value in itself". Engsted (2009, p.401) points out that using the conventional level "mechanically and thoughtlessly in each and every application" is meaningless.

---

[4] Reprinted in Morrison and Henkel (1970, p.157).
[5] Reprinted in Morrison and Henkel (1970, p.139).

From a survey of academic economists, Soyer and Hogarth (2012) find that regression statistics can create an illusion of strong association. They find that the surveyed economists provide better predictions when they are presented with a simple visual representation of the data than when they are confronted only with regression statistics (as in Figure 1). By reconciling the classical and Bayesian methods of significance testing for a large number of the papers published in psychology journals, Johnson (2013) finds that p-values of 0.005 and 0.001 correspond to strong and very strong evidence against H_0, while the p-values in the neighbourhood of 0.05 and 0.01 reflect only modest evidence. Based on this, Johnson (2013) recommends adoption of the "revised standards for statistical evidence" by setting the level of significance at 0.005 or 0.001, instead of 0.05 and 0.01 (as in the example in Figure 1).

6. Concluding Remarks

Although the level of significance is an important input to hypothesis testing, modern statistical textbooks allocate surprisingly little space on the discussion as to how it should be chosen for sound statistical inference. This paper presents such a discussion with several examples for students, along with the selected references to the past and recent academic research. While the conventional levels may still serve as useful benchmarks, mindless and mechanical choice of these levels should be avoided. Students of basic statistics should understand that the level of significance should be chosen with relevant contexts in mind, in careful consideration of the key factors such as sample size and expected losses.

References

Engsted, T. 2009, Statistical vs. economic significance in economics and econometrics: Further comments on McCloskey and Ziliak, *Journal of Economic Methodology*, 16, 4, 393-408.

Gigerenzer, G. 2004, Mindless statistics: Comment on "Size Matters", *Journal of Socio-Economics*, 33, 587-606.

Johnson, V. E. 2013, Revised standards for statistical evidence, Proceedings of the National Academy of Sciences, www.pnas.org/cgi/doi/10.1073/pnas.1313476110

Keuzenkamp, H.A. and Magnus, J. 1995, On tests and significance in econometrics, *Journal of Econometrics*, 67, 1, 103–128.

Kim, J. H. and Ji, P. 2015, Significance Testing in Empirical Finance: A Critical Review and Assessment, Journal of Empirical Finance, in press. (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2410049)

Kish, L. 1959, Some statistical problems in research design, *American Sociological Review*, 24, 328-338.

Labovitz, S. 1968, Criteria for selecting a significance level: a note on the sacredness of 0.05, *The American Sociologist*, 3, 200-222.

Leamer, E. 1978, *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley, New York.

Lehmann E.L. and Romano, J.S. 2005, *Testing Statistical Hypothesis*, 3rd edition, Springer, New York.

Moore, D.S. and McCabe, G.P. 1993, *Introduction to the Practice of Statistics*, 2nd edition, W.H. Freeman and Company, New York.

Morrison, D. E. and Henkel, R. E. 1970, *The Significance Test Controversy: A Reader*, edited by D. E. Morrison and R. E. Henkel. Aldine Transactions, New Brunswick, NJ.

Skipper, J. K. JR., Guenther, A. L. and Nass, G. 1967, The sacredness of .05: a note on concerning the use of statistical levels of significance in social science, *The American Sociologist*, 2, 16-18.

Soyer, E. and Hogarth, R. M. 2012, The illusion of predictability: How regression statistics mislead experts, *International Journal of Forecasting*, 28, 695-711.

Winer, B. J. 1962, *Statistical Principles in Experimental Design*, New York, McGraw-Hill.

Ziliak, S. T. and McCloskey, D.N. 2008, *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, The University of Michigan Press.

Table 1. Sample Size, Probabilities of Type I and II errors, Power, and Critical Values

n	α	β	1- β	Critical Value
The value of α is fixed				
10	0.05	0.80	0.2	1.645
50	0.05	0.45	0.55	1.645
100	0.05	0.20	0.80	1.645
200	0.05	0.03	0.97	1.645
300	0.05	0.004	0.996	1.645
The value of α decreases with sample size				
10	0.35	0.35	0.65	0.40
50	0.19	0.19	0.81	0.89
100	0.11	0.11	0.89	1.25
200	0.04	0.04	0.96	1.76
300	0.015	0.015	0.985	2.17

n: sample size; α: the level of significance; β: Probability of Type II error, 1-β: power of the test

Table 2. Burden of Proof in Legal Trials

Burden of Proof	Description	Trials
Preponderance of Evidence	Greater than 50% chance	Civil, Family: Child support, unemployment benefit
Clear and Convincing Evidence	Highly and substantially probable	Civil, Criminal: Paternity, Juvenile delinquency, Probate, Decision to remove life support
Beyond Reasonable Doubt	No plausible reason to believe otherwise	Criminal: Imprisonment, Death Penalty

Table 3. Expected Losses from Hypothesis Testing

	$(L_1, L_2) = (20, 100)$	$(L_1, L_2) = (100, 20)$
$(\alpha, \beta) = (0.05, 0.25)$	13	5
$(\alpha, \beta) = (0.25, 0.05)$	5	13

The entries of the table are the value of expected loss $0.5\alpha L_1 + 0.5\beta L_2$

Figure 1. Statistical Significance and Sample Size
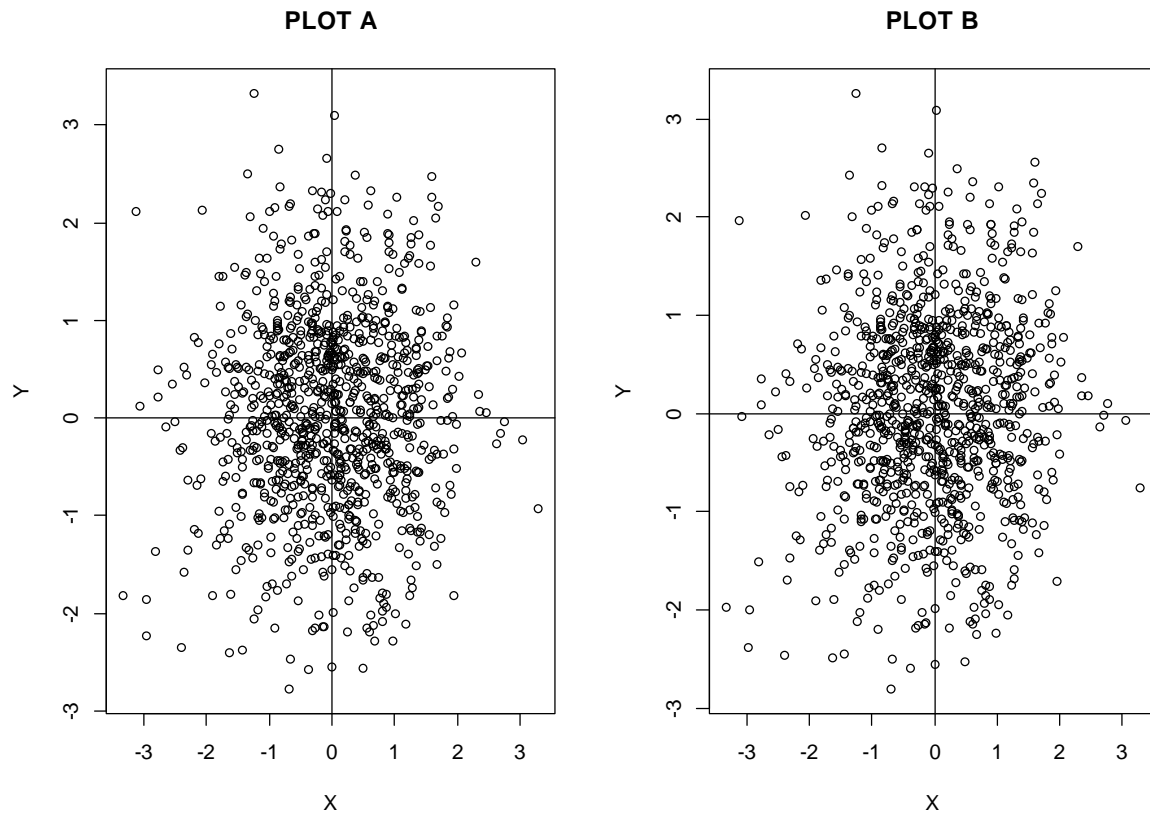
PLOT A

PLOT B

X ~ N(0,1) and Y ~ N(0,1) with sample size 1000.
Plot A: Y and X independent; and the regression slope coefficient is statistical insignificant.
Plot B: Y and X are related with negligible correlation of 0.05, but the regression slope coefficient is statistically significant at the 1% level.
The same random numbers are used for both plots.

Figure 2. Examples of the Line of Enlightened Judgement
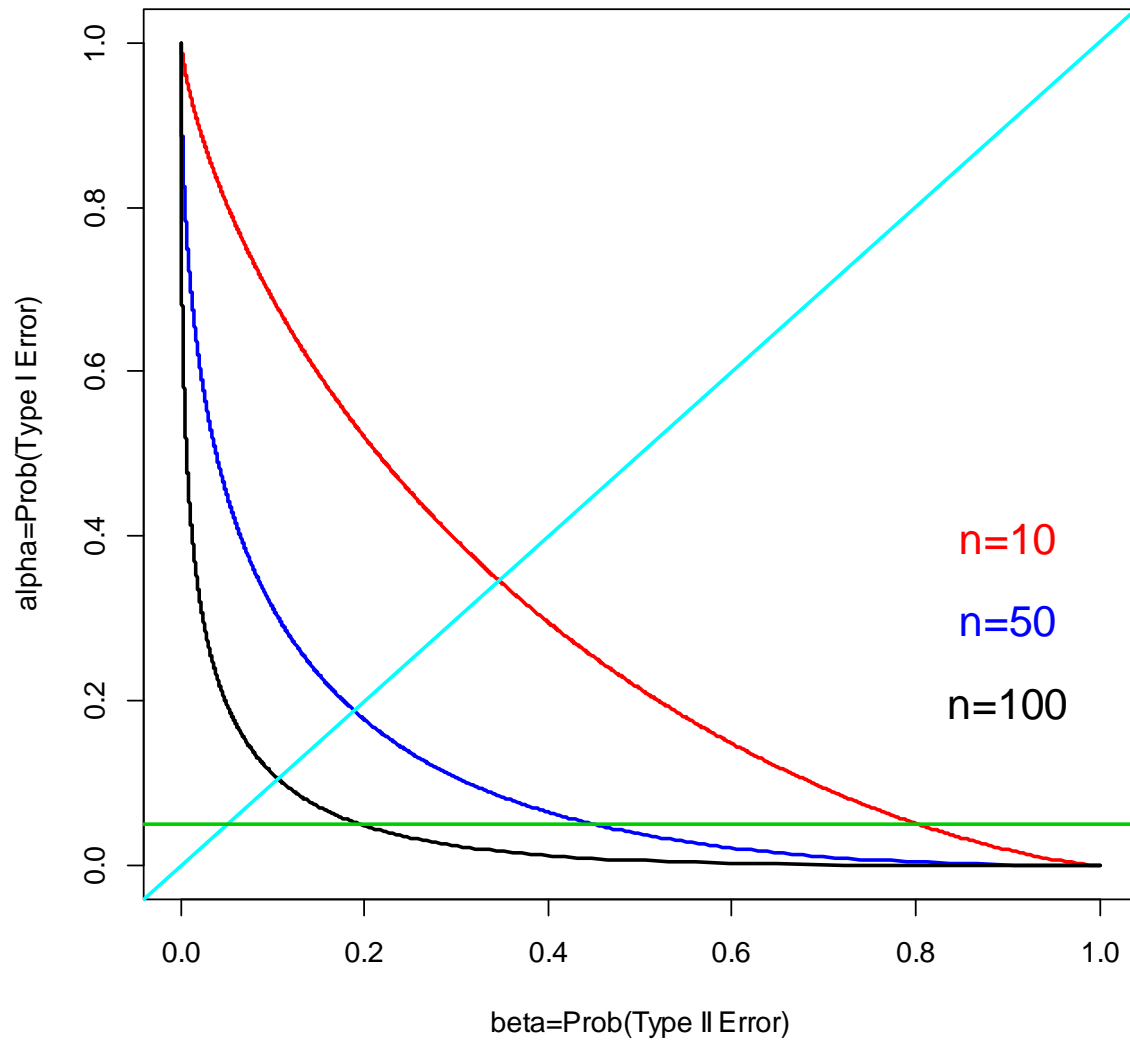
The horizontal line corresponds to α = 0.05.
The 45-degree line corresponds to the points where α+β is minimized (assuming $L_1=L_2$).