The P-Value Concept in Hypothesis Testing and Its Application on Mortality Rate Data

[1]Pwasong, A.D. and [2]Kembe M.M.

[1]Department Of Mathematics, University Of Jos. Davougus@Gmail.Com
[2]Benue State University, Markurdi, Benue State. Email:kdzever@yahoo.com, Tel: 08036177129

Abstract

This study is aimed at comparing the probability value (p–value) of various hypotheses tested with the specified level of significance α at 5% level. The study used data obtained from Hajiya Gambo Sawaba Government General Hospital, Kofan Gaya, Zaria and other related examples to achieve this objective. The study involved the development and validation of the reasoning about p–values and statistical significance scale. The study finally recommends the use of p–value to take care of the probability of committing a type I error.

Keywords: P – value, hypothesis, significance and infarct

1.0 Introduction

More often than not, experiments are carried out primarily to discover new facts or to test the result of previous findings. One of the most important tools in the analysis of experiment is hypothesis testing. [2] Defined **test of hypotheses** or **test procedure** as a method for using sample data to decide between two competing claims (hypotheses) about a population characteristic. One hypothesis might be $\mu = 1000$ and the other $\mu \neq 1000$, or one might be $\pi = 0.01$ and the other $\pi < 0.01$. They further assert that, if it were possible to carry out a census of the entire population, we would know which of the two hypotheses is correct, but usually we must decide between them using information from samples.

According to [1] hypothesis testing is largely the product of Ronald Fisher, Jerzy Neyman, Karl Pearson and Egon Pearson. Fisher was an agricultural statistician who emphasized rigorous experimental design and methods to extract a result from few samples assuming Gaussian distributions. Neyman, who teamed with the younger brother Pearson, emphasized mathematical rigor and methods to obtain more results from many samples and a wider range of distributions.

The concept of p–values has been adopted widely in practice to avoid the imposition of the predefined level of significance that is always fixed at a specified level known as α – value. [8], defined the p–value as the probability of obtaining a value for the test statistic that is as extreme, or more extreme (taking account of the alternative hypothesis). They further explained that if the actual value of the statistics is too far from its expected value, the test is deemed to be significant and the decision to reject H_0 in favour of H_1. If the actual value of the statistics is close to its expected value the test is deemed not to be significant and the decision is not to reject $H_{0.}$ The set of values of the statistic that lead to rejection of H_0 is called the critical region or rejection region, and the set of values that do not lead to rejection of H_0 is called the acceptance region.

The p-value being a probability can take any value between 0 and 1. Values closed to 0 indicate that the observed difference is unlikely to be due to chance, whereas a p-value close to 1 suggests there is no difference between groups other than that due to random variations.

More technically, a p–value of an experiment is a random variable defined over the sample of the experiment, such that, its distribution under the null hypothesis is uniform on the interval [0, 1]. Many p– values can be defined for the same experiment.

Frequent uses of a fixed level of significance have become a thing of concern in hypothesis testing. Looking into this, the use of p-value as a level of significance has become necessary. Many statisticians or experimenters are not aware of the use of p– values and tend to stick to a pre–selected level of significance which would not bring the true picture of whether a given level of significance is barely into a rejection region or far into the region. Hence, this paper is aimed at determining how to calculate p–values using appropriate test statistics and why p-values are preferable to other fixed level of significance, as well as the relationship between p–values and other level of significance.

Hypothesis testing is so important because it provides an objective frame-work for making decisions using probabilistic methods, rather than relying on subjective impression. People can form different opinions by looking at data, but a hypothesis test provides a uniform decision making criterion that is consistent for all people. Hypothesis testing is also important and crucial in decision making. As part of statistical inference, it is widely applied in various aspects of discipline such as: economics, business and science. P– values as an extension of hypothesis testing is very important as it provides the true value of α for which the data is significant. Once the p–value is known the decision maker can determine how significant the data are without the data analyst formally imposing a pre-selected significance.

A review of the research literature from the field of statistics, statistical and mathematics education, psychology and educational psychology reveals difficulties or misconceptions students may have understanding probability and statistics. Researchers have examined how people's prior intuitions, heuristics, and biases may impact their reasoning about problems in probability, data analysis and descriptive statistics; for example, as in [3,5,6]. It was common in the past for researchers to classify results as statistically significant or non significant, based on whether the p–value was smaller than some pre–specified cut point, commonly 0.05. This practice is now becoming increasingly obsolete, and the use of exact p–values is much preferred. This is partly for practical reasons, because the increasing use of statistical software renders calculation of exact p–value simple as compared with the past when tabulated values were used.

The goal of this study is to develop an instrument for statistics education research that shows evidence of making inferences about student's inferential understanding. The study involved the development and validation of the reasoning about p–values and statistical significance scale.

2.0 Methodology

In another development, [7] asserts that the purpose of the p–value test is to facilitate statistics education research on students' conceptual understanding and misunderstanding of statistical inference and the effect of instructional approaches on the understanding. This section describes the method used to describe statistical significance and reasoning about p–values.

2.1 Data source

The data used for analysis in this paper was obtained from the secondary source, where the data is not originally collected by the investigator, but rather obtained from published or unpublished sources. The data was obtained from a recognized government institution of the Federal Government of Nigeria. The institution is a hospital called “ Hajiya Gambo Sawaba Government General Hospital”, Kofan Gaya, Zaria. The data were from the medical records unit of the hospital.

2.2 The p–value

Statistical analysis is most useful when one is looking for difference that is small compared to experimental impression and biological variability. A p–value is a measure of how much evidence one has against the null hypothesis. The null hypothesis, traditionally represented by the symbol H_0 is true. The type of hypothesis tests (right tailed test, left tailed test or two tailed test) will determine what “more extreme” means. The p–value measures consistency by calculating the probability of observing the result from your sample of data or a sample with result more extreme, assuming the null hypothesis is true. The smaller the p–value, the greater the inconsistency of the null hypothesis. The general rule is that a small p–value is evidence against the null hypothesis while a large p–value means little or no evidence against the null hypothesis.

2.3 Statistical significance

The term significant is seductive, and it is easy to misinterpret it. A result is said to be statistically significant when the p-value is less than a pre-set threshold value. It is easy to read into that word significant because the statistical use of the word has a meaning entirely distinct from its usual meaning. Just because a difference is statistically significant does not mean it is important or interesting. And a result that is not statistically significant (in the first experiment) may turn out to be very important.

If a result is statistically significant, there are two possible explanations:

(i) The populations are identical; so there is really no difference. You happened to randomly obtained larger values in one group and smaller values in the other, and the difference was large enough to generate a p- value less than the threshold you set. Finding a statistically significant result when the populations are identical is called making a type I error.

(ii) The populations really are different, so your conclusion is correct.

In writing up the result of a study, a distinction between scientific and statistical significance should be made, since the two terms do not necessarily coincide. The result of a study can be statistically significant but still not be scientifically important. This situation would occur if a small difference was found to be statistically significant because of a large sample size. Conversely, some statistically non-significant result can be scientifically important, encouraging researchers to perform large studies to confirm the direction of the findings and possibly reject H_0 with a larger sample size.

2.3.1 One vs two-tailed p-value method

When comparing two groups, you must distinguish between one and two-tailed p-values. Start with the null hypothesis that the two populations really are the same and that the observed differences between sample means is due to chance. The two-tailed p- value answers this question. Assuming the null hypothesis, what is the chance that randomly selected samples would have means as far apart as observed in this experiment with either group having the larger mean?

To interpret a one-tail p-value, you must predict which group would have the larger mean before collecting any data. The one-tail p-value answers these questions. Assuming the null hypothesis, what is the chance that randomly selected samples would have means as far apart as observed in this experiment with this specific group having the larger mean? A one-tail p-value is appropriate only when previous data, physical limitation or common sense tell you that a difference, if any, can only go in one direction. The issue is not whether you expect a difference to exist that is what you are trying to find out with the experiment. The issue is whether you should interpret increases and decreases the same.

One should only choose a one-tail p-value when one believes the following:

(i) Before collecting any data, you can predict which group will have the larger mean (if the means are in fact different)

(ii) If the other group ends up with the larger mean, then you should be willing to attribute that difference to chance, no matter how large the difference.

It is usually best to use a two tailed p-values for these reasons:

(i) The relationship between p-values and confidence interval is clearer with two-tailed p-value.

(ii) Some tests compare three or more groups, which makes the concepts of tail inappropriate.

In other situations, you will want to make a decision based on a single comparison. In these situations, follow the steps of statistical hypothesis testing:

(i) Set a threshold p-value before you do the experiment. Ideally, you should set this value based on the relative consequences of missing a true difference or falsely finding a difference. In fact, the threshold value called (alpha) is traditionally almost always set to 0.05.

(ii) Define the null hypothesis. If you are comparing two means, the null hypothesis is that the two populations have the same means.

(iii) Do the .appropriate statistical test to compute the p-value.

(iv) Compare the p-value to the present threshold value. If the p-value is less than the threshold, state that you "reject the null hypothesis" and that the difference is "statistically significant". If the p-value is greater than the threshold value, state that "do not reject the null hypothesis" the difference is not "statistically significant."

2.3.2 P-value under single value (One-tailed test)

A test of any statistical hypothesis, where the alternative hypothesis is expressed by means of a less than symbol (<) or greater than symbol (>) is called a one tailed test, since the entire critical region lies in one tail of the distribution of the test statistic. The symbols < or > point to the direction of the critical region. The steps for testing a hypothesis about a mean of a population with known variance against one sided alternative hypothesis may be summarized as follows:

(i) Ho: $\mu = \mu_0$

(ii) H_1: alternative is either $\mu < \mu_0$ or $\mu > \mu_0$

(iii) Choose a level of significance equal to α

(iv) Critical region $Z < -Z_\alpha$ for the alternative $\mu < \mu_0$ or $Z > Z_\alpha$ for the alternative $\mu < \mu_0$

where Z has a standard normal distribution. Compute $\bar{x}$ from a random sample of size n, and then find

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad (1)$$

(v) Conclusion: Reject H_0 if Z falls in the critical region otherwise accepts H_0.

We can solve for p as a function of Z by: $p = \phi(z_p) = \phi(z) = \phi\left[\frac{\bar{x} - \mu_0}{\sigma/\sqrt{}}\right]$ (2)

Example: A topic of recent clinical interest is the probability of using drugs to reduce infarct size in patients who have had a myocardial infarction within the past 24 hours. Suppose we know that in untreated patients the mean infarct size is 25 with a standard deviation of 10.

Furthermore, in 8 patients treated with the drug, the mean infarct size is 16. Is the drug effective in reducing infarct size? (Use α= 0.05).

Solution: The hypotheses are:
Ho: $\mu = 25$ versus H_1: $\mu < 25$, $\sigma = 10$ and

Hence, $p = 0.005 < 0.05$. Thus H_0 is rejected and we conclude that the drug reduces infarct size.

“

$$p=\phi\left(z_p\right)=\phi\left(z\right)=\left[\frac{16-25}{10/\sqrt{8}}\right]=\phi\left(-2.55\right)=1-\phi\left(2.55\right)=1-0.9945\approx 0.005$$

The importance of p-value is that it tells us exactly how significant the results are without performing repeated significance tests at different levels. In the above example the p-value is equal to 0.005 and thus the results are highly significant, which is known under the null hypothesis $\bar{x} \sim N\left(\mu_0, \sigma^2/n\right)$. Hence the probability of obtaining a sample that is no larger than $\bar{x}$ under the null hypothesis is:

$$\phi\left[\frac{x-\mu_0}{\sigma/\sqrt{n}}\right] = \phi(z) = \text{p-value} \qquad (3)$$

These are extracts from [3].

2.3.3 P-value under single value (two-tailed test)

A test of any statistical hypothesis where the alternative is written with a non-equal sign (≠) is called a two-tailed test, since the critical region is split into two equal parts, one in each tail of the distribution of the test statistic.

The null hypothesis, H_0 will always be stated using the equality sign so as to specify a single value. In this way the probability of committing a type 1 error can be controlled. The steps for testing a hypothesis about a mean of a population with known variance ∂^2 against two-sided alternative hypothesis may be summarized as follows:

(i) Ho: $\mu = \mu_0$

(iii) H_1: $\mu \neq \mu_0$

n = 8

The p-value is computed using

$$=\phi\left(Z_p\right)=\phi_{(z)}=\left[\frac{16-25}{10/\sqrt{8}}\right]=\phi\left(-2.55\right)=1-$$

ϕ(2.55) = 1 -] = 0.9945 ≈ 0.005

(iv) Choose a level of significance equal to α

(v) Critical region $Z < -Z_\alpha/2$ and $Z >$

(vi) $Z_\alpha/2$ for the alternative $\mu \neq \mu_0$ where Z has a standard normal distribution. Compute x from a random sample of size n and then *find* $Z = \dfrac{x-\mu_0}{\sigma/\sqrt{n}}$

(vii) **Conclusion**: reject H_0 if Z falls in the critical region otherwise accept H_0.

These are extracts from [7].

2.3.4 Determining the p-value for one sample Z test

To determine the p-value for the one-sample Z test for the mean of a normal distribution with known variance (two – alternatives) is given :

$$\begin{cases} 2\phi(z), & if \quad z \le 0 \\ 2\left(1-\phi(z)\right), & if \quad z > 0 \end{cases} \qquad (4)$$

Thus, in words, if Z ≤ 0, then p = 2 times the area under an N (0, 1) distribution to the left of Z; if Z > 0, then p = 2 times the area under an N (0,1) distribution to the right of Z. Using the above example we have:

H_0: $\mu = 25$ versus H_1: $\mu \neq 25$, $\sigma = 10$

The p-value is computed using

p = 2 x ϕ(-2.55)
= 2 x [l - ϕ(2.55)]
= 2 x (1 - 0.9945)
= 2 x 0.005 = 0.01
p = 0.01 < 0.05

Therefore we reject H_0 and conclude that the drug reduces infarct size.

2.3.4 P-value between two values

(One-tailed test)

The steps for calculating p-value between two values with one-sided alternative is the same as the one considered in 2.3.3 above, the only difference is that the levels of significance alpha (α) are in two different forms. If we decide to use two different levels of α at 0.05 and 0.01 respectively, we may want to determine whether H_0 is rejected at both levels of significant or accepted in one of the levels and rejected in the other. From our previous example we know that the p-value, using one sided alternative is 0.005. If we compare this value with the two levels of α i.e. 0.05 and 0.01, we will notice that the p-value which is 0.005 is so small compared to these two values. So we reject H_0 and conclude that the result is highly significant.

2.3.5 P-value between two values (Two-tailed test)

To calculate p–value between two values with two-sided alternative requires the following steps:

(i) Set H_0: $\mu = \mu_0$

(ii) Set $H_1 : \mu \neq \mu_0$

(iii) Choose a level of significance α. In this case α is chosen at two levels (0.01 and 0.05)

(iv) Critical region $Z < -Z_{\alpha/2}$ and $Z > Z_{\alpha/2}$ for the alternative where Z has a standard zormal distribution. Compute X from a random sample of size n, and then find:

(v) $$Z = \frac{x - \mu_0}{\delta / \sqrt{n}} \qquad (5)$$

We can calculate p-value in terms of Z i.e. p = ϕ(Z). Since it is two-sided alternative we have:

$$\begin{cases} 2\phi(z), & if \quad z \leq 0 \\ 2(1-\phi(z)), & if \quad z \geq 0 \end{cases} \qquad (6)$$

Using our previous example we have

$$p = \phi(Z_p) = \phi_{(z)} = \left[\frac{16-25}{10 \big/ \sqrt{8}} \right] = \phi(-2.55)$$

Therefore, for two-sided alternative we have:

2 (1 – ϕ(2.55)) = 2 (1 - 0.9945) = 2 (0.005) = 0.01. Therefore p = 0.01.

Conclusion: Since 0.01 < P < 0.05, we reject H_0 and conclude that the result is statistically significant (i.e. the Drug may reduce the infarct size).

3.0 Data Presentation

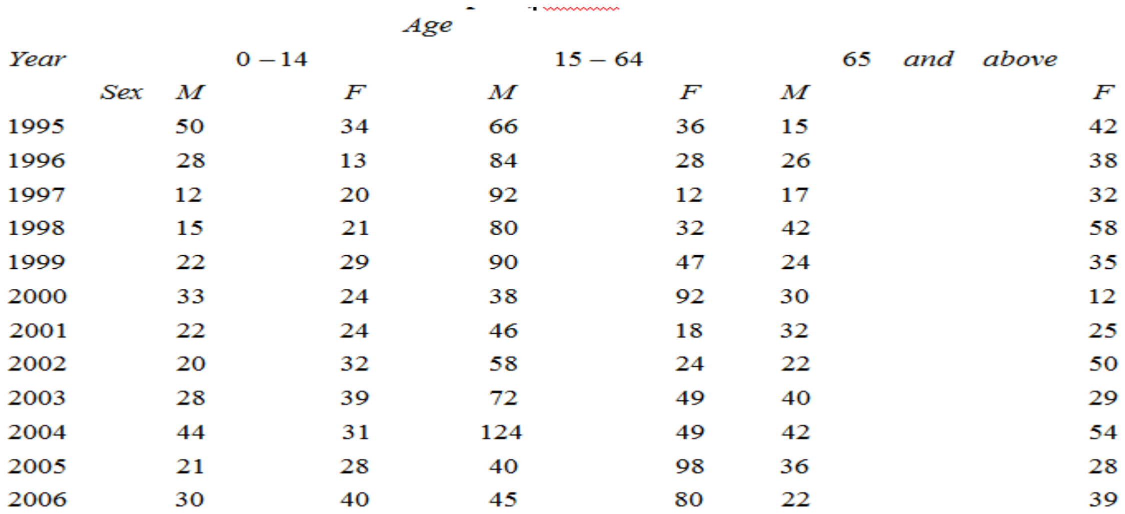
The data in tables 1 and 2 are part of the secondary data collected from the Hajiya Gambo Sawaba Government General Hospital Kofan-Gayan, Zaria, Kaduna State. The data were subjected to analysis using Predictive Analytic Software (PASW), with a view to calculating the p-value for all the hypotheses under consideration

.

Table 1: Mortality rate by sex distribution of Hajiya Gambo Sawaba Government General Hospital, Zaria in 1999.

Male	8	3	12	11	6	8	10	12	6	17	2	8	9	9	8	2
Female	10	14	6	3	9	6	6	3	4	20	2	4	6	7	4	8

Table 2: Mortality rate by age\ sex distribution of Hajiya Gambo Sawaba Government General Hospital, Zaria



		Age					
Year		0 – 14		15 – 64		65 and above	
	Sex	M	F	M	F	M	F
1995		50	34	66	36	15	42
1996		28	13	84	28	26	38
1997		12	20	92	12	17	32
1998		15	21	80	32	42	58
1999		22	29	90	47	24	35
2000		33	24	38	92	30	12
2001		22	24	46	18	32	25
2002		20	32	58	24	22	50
2003		28	39	72	49	40	29
2004		44	31	124	49	42	54
2005		21	28	40	98	36	28
2006		30	40	45	80	22	39

3.1 Analysis and Results.

The data was analyzed using the Predictive Analytic Software (PASW) and tables 3(a), 3(b), 4 and 5 revealed the results obtained. The level of significance is 0.05 and the various hypotheses to be tested are listed below:

(*i*) H_0: There is no significant difference between male and female with respect to mortality rate.

.

(ii) H_1: There is significant difference between male and female with respect to mortality rate.

(iii) H_0: There is no significant difference between sex and age group with respect to mortality rate.

(iv) H_1: There is significant difference between sex and age group with respect to mortality rate.

Table 3(a): t – test for sex with respect to mortality rate (Group Statistics)

	Sex	N	Mean	Std. Deviation	Std. Error Mean
Mortality rate	Male	16	8.1875	3.93647	0.9841
	Female	16	7.000	4.6188	1.1547

Table 3(b): Independent sample test

		Levene's test forEquality of variances		t – test for equality of means						
Mortality Rate		**F**	**Sig.**	**t**	**df**	**Sig.(2-tailed)**	**Mean difference**	**Std.error difference**	**95% Confidence**	
	Equal variances assumed	0.162	0.691	0.783	30	0.44	1.1875	1.51718	-1.9	4.29
	Equal variances not assumed			0.783	29.265	0.44	1.1875	1.51718	-1.9	4.29

3.2 Interpretation

In comparing mortality rate between male and female in Sawaba Government General Hospital, Zaria in 1999, the result above clearly portrayed there is no significant difference between male and female with respect to mortality rate. At 5% significance level, the p-value which is 0.44 is greater than 0.05. Hence we conclude by accepting the null hypothesis that there is no significant difference existing between male and female mortality rate. For the chi-square test, the results in tables 4 and 5 were obtained.

.

Table 4: Sex and age group cross tabulation

Crosstab

			Age			Total
			0 -14	15 - 64	65 - above	
Sex	M	Count	325	836	348	1509
		Expected Count	349.3	741.5	418.1	1509
	F	Count	335	565	442	1342
		Expected Count	310.7	859.5	371.9	1342
Total		Count	660	1401	790	2851
		Expected Count	660	1401	790	2851

Table 5: Chi-Square Test

	Value	df	Asymp. Sig(2-sided)
Pearson Chi-Square	54.160[a]	2	0.000
Likelihood Ratio	54.327	2	0.000
Linear-by-Linear Association	5.83	1	0.016
No of Valid Cases	2851		

[a] cells (0%) have expected count less than 5. The minimum expected count

From the table 5 above we can say that the difference between sex and age group with respect to mortality rate is highly significant. Using p-value approach the null hypothesis is rejected since p-value at 0.000 is less than 0.05 level of significance. We therefore, conclude that the mortality rate does not depend on age and sex.

4.0 Conclusion and Recommendation

From the above results, it is seen that the conclusion reached using p-value at α (alpha) level of testing is more accurate and more reliable tha those reached using other criteria.

The use of p-value did not only show the significant difference but also revealed how significant the observed difference is in statistical point of view. We can deduce from the p-value whether the observed difference is far into the critical regions or merely into the regions and with this information, we can decide whether the data should be adjusted to meet up with the desired accuracy or be thrown aside completely.

Finally p-value do not only offer experimenters or investigators varieties of choices, but also eliminates the fear of imposition of pre-set level of significance that always result in reaching partial or inadequate conclusion whenever an experiment or trial is conducted. The main goal in this research work is to compare the probability value (p-value) of various hypotheses tested with the specified level of significance α (alpha) at 5% level with a view to determining whether the null hypothesis should be rejected or not.

One of the most important advantages of p-value in hypothesis testing is its ability to eliminate the fear of doubt as regarding the validity of conclusions being drawn from the result of an experiment. Since performing experiments cost money, time, energy and resources. Utmost caution has to be exercised before drawing conclusions. Unfortunately, many experimenters or investigators are not aware of the use of p-values and for this reason, it is recommended that p-values should always be used in hypothesis testing rather than the pre-set level of significance that we are conversant with.

Moreover, since a p-value conveys much information about the weight of evidence against the null hypothesis and knowing fully from the fact that rejecting a true null hypothesis H_0 implies probability of committing type I error, we therefore, recommend the use of p-value to take care of this error.

References

[1] Cobb, G. (1992). Teaching statistics. In L.A. Steen (Ed:), *Heading the case for change : Suggestions for curricular action (pp. 3 – 43).*Wasshington, D.C: the Mathematical Association of America.

[2] Devore, J., & Peck, R. (2006). *Statistics: The exploration and analysis of data (5th ed:),* Belmont; CA: Brooks/ Cole – Thomson Learning.

[3] Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: *Implications for research. Journal for Research in Mathematics Education 19(1),* 44 – 63.

[4] Huberty, C.J. (1993). Historical origins of statistical testing practices: *The treatment of Fisher versus Neyman – Pearson view in textbooks. Journal of Experimental Education, 61(4),* 317 – 333.

[5] Kahneman, D., & Tversky, A. (1982). Subjective probability: A judgement of representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *judgement under uncertainty: Heuristics and biases (pp. 32 – 47)*: Cambridge: Cambridge University Press

[6] Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education, 3(1). Retrieved march 20, 2007,* from http://www.amstat.org/publications/jse/v3n1/konold.htm/

[7] Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological methods, 5(2), 241* - 301

[8] Saldanha, L.A., & Thompson, P.W. (2006). Investigating statistical unusualness in the context of a resampling activity: Students exploring connections between sampling distribution and statistical inference. In A. Rossman & B. Chance (Eds). *Working cooperatively in statistics education: Proceedings of the Seventh International*