Evaluating the Agreement of Methods using Gaillard-Makki Method

by

Golbou Makki

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
August 5, 2017

Keyword: Bland-Altman, Agreement, Regression, p-value

Approved by

Phillipe Gaillard, Associate Professor, Department of Mathematic & Statistics
Guanqun Cao, Assistant Professor, Department of Mathematic & Statistics
Bertram Zinner, Associate Professor, Department of Mathematic & Statistics

Abstract

In many laboratory studies such as medicine, fisheries, and chemistry comparison of two different methods of measurement is of great importance. These measurements are critical when the methods are changed, a new or alternative method proposed, or two exciting methods are not in complete agreement. The most common used methods such paired t-test, correlation and regression can be misleading, thus, Altman and Bland introduced a new method to address the shortages of those methods. However, Altman-Bland method is not introducing a test of significance. Moreover, threshold and intervals in this method are not set in prior. These two drawbacks are addressed in this study. P-value quantified for the Altman-Bland method and thresholds are set before the run of test.

Acknowledgments

I offer my most sincere gratitude to my major advisor Dr. Phillipe Gaillard. It has been a great honor and opportunity to build my academic career under his supervision at Auburn University.

Last but not least, I would like to thank my family. This was not possible without their love and support. I thank my sister Golbar. I dedicated this work to my parents Zohreh Hearizadeh and Mahmood Makki who always believed in me and supported me, and my husband Hamed Majidzadeh who was always by my side

Table of Contents

List of Tables

List of Figures

Chapter1: Introduction

In many laboratory studies such as medicine, fisheries, and chemistry comparison of two different methods of measurement is of great importance (Altman and Bland, 1983; Giavarina, 2015; Stevens et al., 2015). (Giavarina, 2015). The new methods (contender) can introduce benefits compared to the existing or accepted method (gold standard). For an instant, the new method may reduce the cost and time which would result in improved efficiency. The new method may improve the accuracy to avoid under and over estimations (Stevens et al., 2015). However, the magnitude of these changes in measurements between the new and old method needs to be quantified. To that end, various statistical approaches have been employed, including but not limited to paired t-test, correlation, regression, and Altman-Bland, (Balakrishnan et al., 2005; Barnhart et al., 2007; Marinovich et al., 2013). Among these methods, Altman-Bland, have been used the most and been cited more than 30,000 times (Stevens et al., 2015). The Altman-Bland method first introduced in 1983 in a manuscript entitled "Measurement in Medicine: the Analysis of Method Comparison Studies" (Altman and Bland, 1983). In this study, we review methods that have been used before Altman-Bland and will demonstrate the benefits of Altman-Bland method over them. A simple approach will be proposed at the end of the study to quantify p-value for the Altman -Bland method.

1.1. Paired t-test

A paired t-test (H_0: $\mu_d = 0$, H_0: $\mu_d \neq 0$) is a crude however widely used method for test of agreement (Balakrishnan et al., 2005). This method only evaluates if the averages of two methods have agreement while the random differences between these two methods can be large. It has been shown that t-test may reject the agreement of averages if the scatter around the 45^o line is near zero (indicating good agreement), and may fail to reject it if the scatter is very high (indicating poor agreement). There are two problems involved with this approach: First, this method is trying to provide supporting evidence to a claim by not rejecting H_0 means that low power works in our favor: Even if the two measures were to exhibit very little agreement, we could ensure retaining H_0 by having a sufficiently small sample size. This would not constitute credible evidence of agreement. The second problem is that this approach is focused on the mean difference. It is quite possible for the mean difference to be very close to zero, and yet for large differences to exist for some or even many subjects. Thus, we could retain H_0 even in the presence of substantial disagreement between the measures. In other words, the rationale is that to the extent that the two measures are in agreement, the average difference of the measurements for each subject will approach zero. Thus, failing to reject H0 would lead us to conclude in favor of agreement.

1.2. Correlation

Correlation is to determine the relationship between a pair of variables. The result of the correlation is called correlation coefficient which is calculated as the ratio of covariance between the variables to the product of their standard deviations (Giavarina, 2015). This ratio ranging from +1 to -1 and closer to the two sides the stronger the linear relation is. Pearson product moment coefficient of correlation is the most common correlation method used which measures the degree of linear association between two variables. A common test of hypothesis involving Person's r consists of testing the H_0 that $\rho=0$. For the purpose of assessing agreement, this hypothesis test is rarely conducted, and for good reasons: simply rejecting H0: $\rho=0$ would constitute remarkably weak evidence of agreement.

$$r_{xy} = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right) \cdot \left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

Admittedly superior to the paired t-test for comparison of methods, the correlation approach is problematic in that the magnitude of Pearson's r depends on how close the data points are to the linear regression line, whereas agreement depends on closeness to the identity line. As the regression line is not constrained to have a zero intercept and a unit slope, these two lines can have quite different locations. This makes Pearson's r a possibly misleading measure of agreement, as it can overestimate (though not underestimate) the extent of the agreement between the two measures.

Correlation shows the relation and its strength, while it incapable of showing the agreement. A strong relation would be observed with any straight line between the variables while the agreement requires the points along the equality line (Bland and Altman, 2010). Moreover, the correlation would not be affected by a change in measurement scale while the agreement will be affected, and data with poor agreement have been shown to have strong correlations (Bland and Altman, 2010). Another shortage involved with the test of significance is that they only show that the two methods are related which seems obvious considering that both methods are to measure the same variable. However, this cannot be necessarily translated to agreement of two methods.

1.3. Regression

Regression also is to determine the relationship between variables, however, in this case, one of the variables is the dependent variable or outcome, while the other one is the predictor while in correlation both variables are considered the same (Crawford, 2006). In other words, linear regression finds the best line that predicts one variable from the other one and quantifies it using r^2 (Giavarina, 2015). Linear regression includes the assumption that should be met before analysis (Crawford, 2006). These assumptions are:" (i) normally distributed residuals with a mean of zero; (ii) constant variance of the residuals; and (iii) independence of residuals from different observations."

linear regression, with the gold standard measure (x2) as the dependent variable and the contender measure (x1) as the independent variable. For the purpose of assessment of agreement between two continuous measures, a two-part H0 is used, and the research hypothesis is that neither will be rejected:

- $H01: \beta_0 = 0$
- $HA1: \beta_0 \neq 0$
- $H02: \beta_1 = 1$
- $HA2: \beta_1 \neq 1$

A problem with this approach is that low power helps the research hypothesis of agreement (as the research hypothesis is aligned not with HA but with H0).

1.4.Altman and Bland Method

Since its first publication in 1983, a fourth approach, the Altman-Bland (Douglas Altman & Martin Bland) approach has been a very popular choice for assessing agreement. Unlike the other three approaches outlined above, it focuses on the entire distribution of differences. This method which is also called ‘‘limits of agreement’’ approach uses a scatter plot called “difference plot” to differentiate the measurements on a subject by two different methods (Myles and Cui, 2007). This plot can determine if the difference is related to averages and if there was

no relationship the distribution of the differences can be computed as follow (Stevens et al., 2015):

$$\bar{d} \pm 1.96\ S_d$$

sample of i= 1, 2, …, n subjects,

$\bar{d}$ = sample average

S_d = standard deviation of the observed differences

The limits are representing intervals that are expected to include 95% of the differences, assuming that the differences have a rough normal distribution. Horizontal reference lines corresponding to the upper and lower limits of agreement and the average difference $\bar{d}$, are added to the plot. The agreement between two methods are evaluated based on a maximum allowable difference between two measurements that have been determined by the scientist.

1.4.1. Drawbacks of Bland and Altman method

The difference plot is the main result of the Altman-Bland approach. It provides the 95% agreement limits, which are the estimated limits within which 95% of the differences in the population would fall, assuming a normal distribution for these differences. These limits are not set by the investigator; they are a function of the sample data. The question to be answered by the investigator is whether these limits are narrow enough and whether the bias (average

difference) is small enough to conclude in favor of the satisfactory agreement. Thus, the Altman-Bland method has two limitations:

1) Lack of a test of significance is the first drawback of this method. The software implementations that often are provided test the H_0: Bias = 0. However, these test that were not provided by Altman and Bland are exactly like paired t-test with same disadvantageous discussed above.
2) As with the correlation approach, since a threshold of what constitutes sufficient agreement does not need to be specified a priori, it is not clear how to translate the results (correlation coefficient or computed agreement limits) into a decision as to whether sufficient agreement has been reached

Chapter 2: Numerical examples from methods that have been used for agreement

As discussed in chapter one techniques such as paired t-test, correlation and regression are not appropriate techniques for method comparison and agreement. However, Bland-Altman provides a simple and appropriate approach for measure of agreement. In this chapter, numeric examples are used to show why methods such as regression, correlation, and paired t-test are inappropriate for measure of agreement and the results compared to outcomes of Altman-Bland method.

2.1. Material and Methods

In this study, data generated using SAS version (9.4), to show how early methods including the use of regression, correlation, and t-test cannot be used to show the agreement of methods.

2.2. Results

2.2.1 Example showing inappropriacy of t-test and equivalence testing

Two datasets (n=30), with equal means, randomly generated (appendix 1). The characteristics of the generated data is depicted in Table 1. The t-test rejects the null hypothesis suggesting that two methods are in agreement (Table 2), while the two data sets are randomly generated and only have an equal average. Altman-bland method, however, reveals the difference between two methods (Fig. 2.1). The mean difference was zero since the mean values were equal. However, the limits of agreement (-82.4, 82.4) are not small enough that we can consider that two methods agree. “These limits are not set by the investigator; they are a function of the sample data. The

question to be answered by the investigator is whether these limits are narrow enough and whether the bias (average difference) is small enough to conclude in favor of sufficient agreement. Altman and Bland did not recommend a test of significance for these agreement limits.

2.2.2. Example showing inappropriacy of regression and correlation

Two random data set (n=100) with a significant correlation were generated (appendix 2). Pearson's correlation coefficient suggested a strong agreement between the two methods (r = 0.94). Similarly, linear regression methods showed an agreement between two methods (r^2 = 0.88, $p < 0.001$, Fig. 2.2).

However, the Bland-Altman method revealed that the two methods are not in agreement. Figure 2.3B compares two methods (generated data). Since a linear relation can be observed between average and the difference in two methods (Fig. 2.3B), a logarithmic transformation have been suggested in such cases (Bland and Altman, 2010). The mean difference was -0.016 on the log scale with limits of agreements of -0.1, 0.1. The antilog of this limit would be 0.70, and 1.26 respectively. It is of importance to consider that these values are dimensionless ratios and are suggesting that for 95% of cases the second method will be between 0.70 and 1.26 times of the first method. Thus, the second method may differ from the first method by 30% below and 26% above.

Tables and Graphs

Table 2.1. Summary of the data generated for evaluation of the t-test.

Variable	N	Mean	Std Dev	Minimum	Maximum
X1	30	58.000	25.380	20.000	100.000
X2	30	58.000	28.211	10.000	100.000

Table 2.2. Summary of the t-test. t-test rejects the null hypothesis suggesting that two methods are in agreement

Simulated Data
Conventional t-test (H0: mean difference = 0)

Difference: X1 - X2

N	Mean	Std Dev	Std Err	Minimum	Maximum
30	0	42.0181	7.6714	-60.0000	90.0000

DF	t Value	Pr > |t|
29	0.00	1.0000

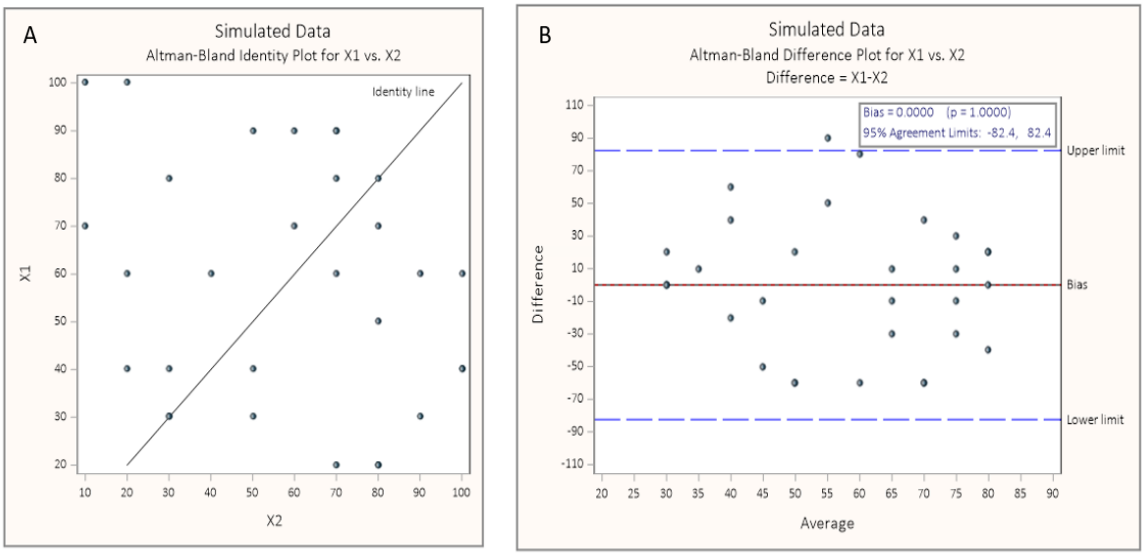

Figure 2.1. Altman-Bland graphs for the generated data with equal means.

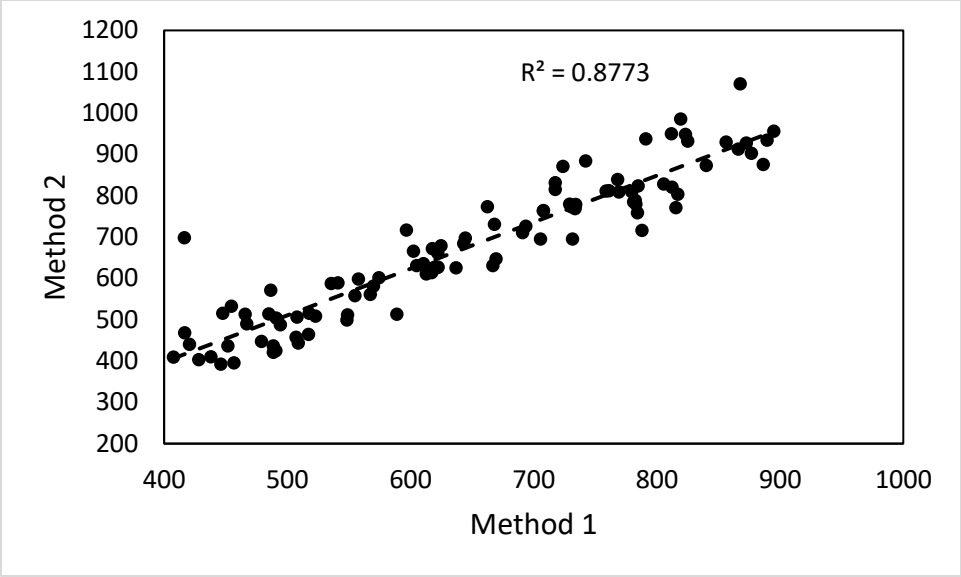

Figure 2.2. The linear regression between to generated data

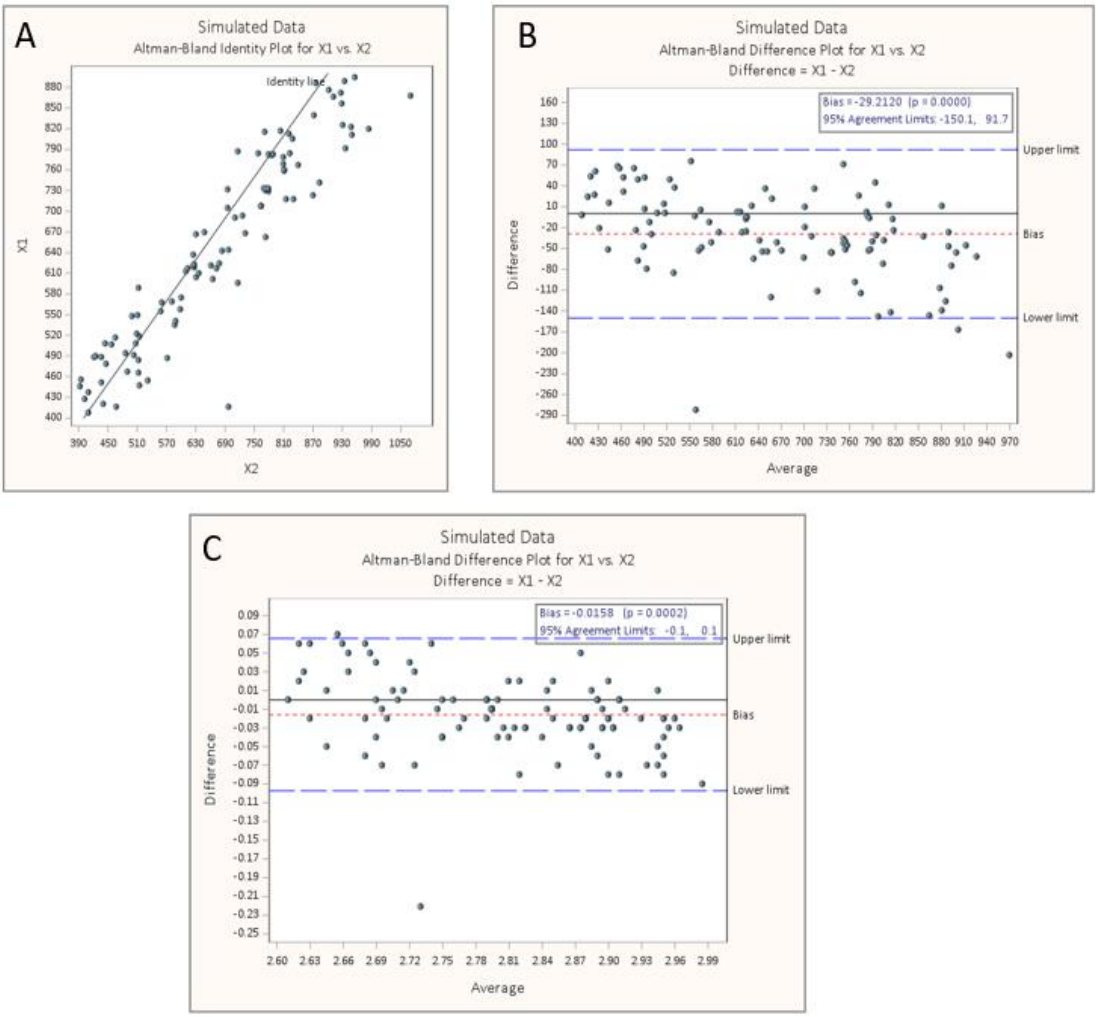

Figure 2.3. Altman-Bland graphs for the generated data with significant relation

Chapter 3: Upper-tail one-sample binomial proportion test

Computing a two sided interval estimate for a population proportion is of importance in applied statistics (Pires and Amado, 2008). The One-Sample Proportion Test is used to assess whether a population proportion (P1) is significantly different from a hypothesized value (P0). The hypotheses may be stated regarding the proportions, their difference, their ratio, or their odds ratio, but all four hypotheses result in the same test statistics.

In this study, a one-sample binomial proportion test was used to address the question of agreement. Let's say we want 0.75 of the differences to be located within the interval we have specified. The number of successes divided by the number of trials defines the binomial proportion. Thus, greater proportion inside the interval is preferred. The data sets with values inside the interval were considered "1", otherwise they were considered "0". This approach helps us to quantify the p-value with the binomial distribution. H0 can be tested that this proportion is smaller than .75 and set out to reject it.

$$H0: \pi \leq .75$$

$$HA: \pi > .75$$

The null hypothesis is that of the insufficient agreement, and its rejection constitutes the evidence in favor of agreement. This proportion test approach has the following desirable features:

1) It requires the investigator to define a priori what is meant by agreement, regarding limits and proportion, before being possibly influenced by the data.
2) It provides a p-value so that the agreement questions can be answered.
3) This p-value pertains to a test of hypothesis in which the research hypothesis is aligned with the alternative hypothesis so that the claim of the agreement is made more difficult when statistical power is low.
4) This test is focused on the entire distribution of differences, not on the mean of these differences.

To further clarify this method, a numeric example is provided below. The characteristics of the generated data are depicted in table 3.1. In this case, we want 0.75 of the differences to be located within the interval the we have specified (-0.5, 0.5, Fig.3.1). The null hypothesis explained above, can be tested for this data set (Fig.3.2). This sampling distribution has been generated under H0 ($\pi = .75$). The p-value for our binomial significance test of proportion is the area under the curve that is as or more extreme than the observed value. As depicted in figure 3.1, 9 data points fall outside the difference limits, which means that the observed proportion of interest is (30-9)/30 = 0.7. This is an upper-tail test, so more extreme means larger (we could reject H0 with a sample proportion larger than .75 but not smaller). For an upper-tail one-sample binomial test of proportion, the p-value is given by $\sum,x=k,n(\text{combined}(n,x)\cdot p^{x}\cdot(1-p)^{(n-x)})$. With an observed proportion of 0.7000 and an upper-tail p-value of 0.803407 (not shown on SAS

output, table 3.2). We do not reject the H0 of $\pi \leq .75$ and conclude that it is not in favor of the agreement, with difference limits of ± 0.5 (table. 3.2).

In this example, out of 30 points, 21 are within the interval. The exact p-value was the red area in figure 3.2 (0.1298 + 0.1593 + 0.1662 + 0.1455 + 0.1047 + 0.0604 + 0.0269 + 0.0086 + 0.0018 + 0.0002 = 0.8030407). If it were a continuous distribution it would be the area. However, since it is a discrete distribution, the p-value would be the sum of them. It is important to consider that the purpose was to conduct an exact binomial test so not a χ^2 approximation and SAS does not give us the exact p value for the binomial test.

One of the advantages of this proposed method is that the intervals have to be set before the test run. While in Altman-Bland method nothing is determined ahead of time which may have impacts on the decision made by scientists and be bias.

Tables and Graphs

Table 3.1. Characteristics of the generated data

Variable	N	Mean	Std Dev	Minimum	Maximum
X1	30	10.298	1.026	7.840	12.230
X2	30	10.298	1.005	8.340	12.130
Difference	30	0.000	0.487	-0.760	1.200

Table 3.2. Proportion test for the generated data

Proportion test

Within	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	9	30.00	9	30.00
Yes	21	70.00	30	100.00

Binomial Proportion	
Within = Yes	
Proportion	0.7000
ASE	0.0837
95% Lower Conf Limit	0.5360
95% Upper Conf Limit	0.8640
Exact Conf Limits	
95% Lower Conf Limit	0.5060
95% Upper Conf Limit	0.8527

Test of H0: Proportion = 0.75	
ASE under H0	0.0791
Z	-0.6325
One-sided Pr <= Z	0.2635
Two-sided Pr > |Z|	0.5271

Sample Size = 30

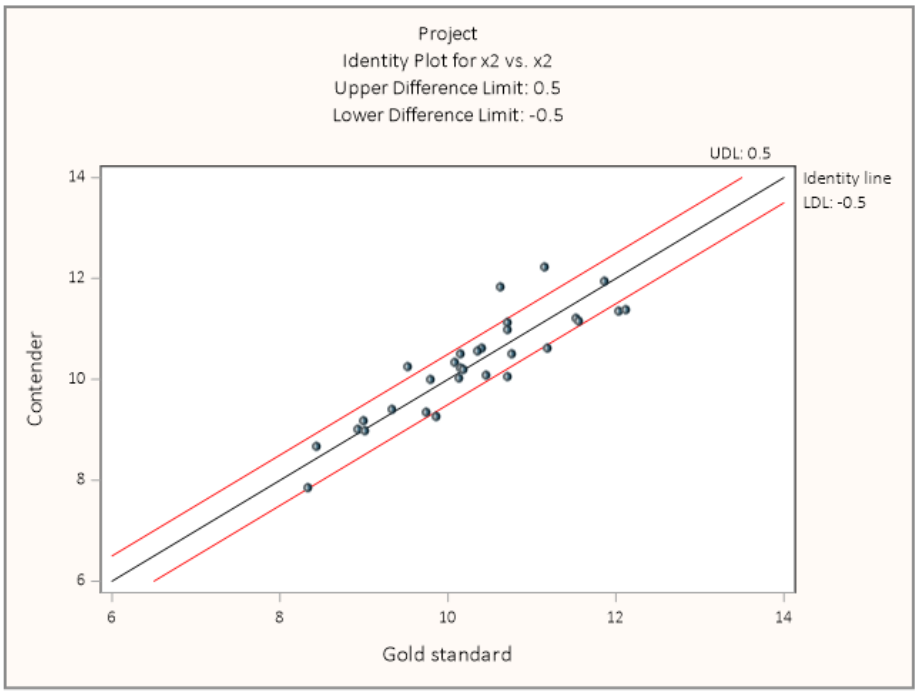

Figure 3.1. 0.75 of the differences to be located within the interval

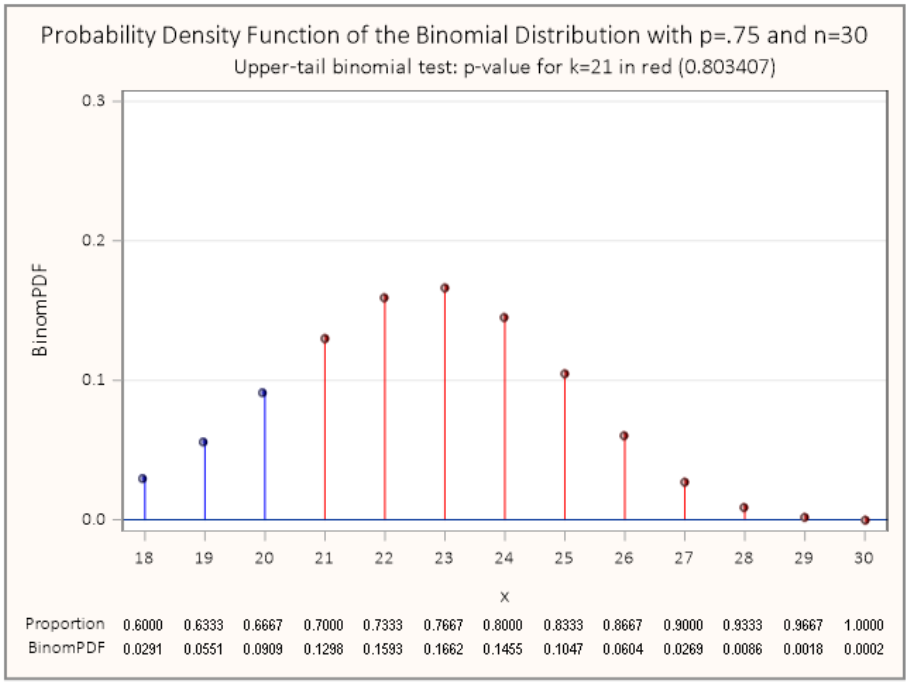

Figure 3.2. This sampling distribution has been generated under H0 ($\pi = .75$) so the p-value for our binomial significance test of proportion is the area under the curve that is as or more extreme than the observed value (21/30=0.7000). This is an upper-tail test, so more extreme means larger (we could reject H0 with a sample proportion larger than .75 but not smaller).

References

Altman, D.G., Bland, J.M., 1983. Measurement in Medicine: The Analysis of Method Comparison Studies. The Statistician 32, 307. doi:10.2307/2987937

Balakrishnan, N., Kannan, N., Nagaraja, H.N. (Eds.), 2005. Advances in ranking and selection, multiple comparisons and reliablility: methodology and applications, Statistics for industry and technology. Birkhäuser, Boston.

Barnhart, H.X., Haber, M.J., Lin, L.I., 2007. An overview on assessing agreement with continuous measurements. J. Biopharm. Stat. 17, 529–569.

Bland, J.M., Altman, D.G., 2010. Statistical methods for assessing agreement between two methods of clinical measurement. Int. J. Nurs. Stud. 47, 931–936.

Crawford, S.L., 2006. Correlation and Regression. Circulation 114, 2083–2088. doi:10.1161/CIRCULATIONAHA.105.586495

Giavarina, D., 2015. Understanding Bland Altman analysis. Biochem. Medica 25, 141–151. doi:10.11613/BM.2015.015

Marinovich, M.L., Macaskill, P., Irwig, L., Sardanelli, F., von Minckwitz, G., Mamounas, E., Brennan, M., Ciatto, S., Houssami, N., 2013. Meta-analysis of agreement between MRI and pathologic breast tumour size after neoadjuvant chemotherapy. Br. J. Cancer 109, 1528–1536. doi:10.1038/bjc.2013.473

Myles, P.S., Cui, J., 2007. I. Using the Bland Altman method to measure agreement with repeated measures. Br. J. Anaesth. 99, 309–311. doi:10.1093/bja/aem214

Pires, A.M., Amado, C., 2008. Interval estimators for a binomial proportion: Comparison of twenty methods. REVSTAT–Statistical J. 6, 165–197.

Stevens, N.T., Steiner, S.H., MacKay, R.J., 2015. Assessing agreement between two measurement systems: An alternative to the limits of agreement approach. Stat. Methods Med. Res. 0962280215601133. doi:10.1177/0962280215601133