Sánchez-López, L. (2010). El español para fines específicos: La proliferación de programas creados para satisfacer las necesidades del siglo XXI. *Hispania 93*(85-89): The John Hopkins University Press.

Slick, S. & Dunn, M. (2005). A Future Trend in Workforce Development: Occupational Spanish. *The Catalyst* 34.3 (Winter 2005), 20-23.

Thomason, S. G. (2001). *Language Contact: An Introduction*. Washington, D.C.: Georgetown UP.

Widdowson, H. G. (1983). *Learning purpose and language use.* Oxford; Oxford University Press.

Biographical Sketch

Dr. Maryjane Dunn is an Assistant Professor of Spanish in the Department of English, Foreign Languages, and Philosophy at Henderson State University. Although primarily a scholar of the Pilgrimage to Santiago, Spain, and 15th century allegorical literature, mid-career her path turned to the linguistic field of teaching language for special purposes while teaching Spanish to law enforcement officers, teachers, health professionals, and public safety workers across the United States.

2-Sample t-Distribution Approximation

Michael Lloyd, Ph.D.
Professor of Mathematics

Abstract

The *t*-distribution used for the 2-sample procedures introduced in elementary statistics is actually an approximation introduced by Welch and Satterthwaite in the late 1940s. We will explore how the error of this approximation depends on the sample sizes and the variances of the independent populations.

Motivation

We will examine the following example, which was extracted from a PowerPoint slide that accompanies a popular elementary statistics book:

> Does smoking damage the lungs of children exposed to parental smoking? Forced Vital Capacity (FVC) is the volume (in millimeters) of air that an individual can exhale in 6 seconds. FVC was obtained for two samples of children, one group exposed to parental smoking, and another group of children not exposed to parental smoking.

Parental smoking	$\bar{x}$ FVC	s FVC	n
Yes	75.5	9.3	30
No	88.2	15.1	30

We want to know if parental smoking decreases children's lung capacity as measured by the FVC test.

The following must be checked, or assumed, for the test statistic to have approximately a *t*-distribution:

1. The children from the smoking group, and the non-smoking group, are both simple random samples.
2. The FVC is Normally distributed for the smoking group, and for the non-smoking group.
3. The sample sizes for the smoking group, and the non-smoking group, are each 30, which is greater than 5.
4. The FVC responses for children in the smoking group are independent of those in the non-smoking group.

The *t* statistic is computed as follows: $t = \frac{\bar{x}_Y - \bar{x}_N}{\sqrt{\frac{s_Y^2}{n_Y} + \frac{s_N^2}{n_N}}} = \frac{75.5 - 88.2}{\sqrt{\frac{9.3^2}{30} + \frac{14.1^2}{30}}} = -3.9$

Degrees of Freedom for the *t* test Statistic

There are three ways to compute the degrees of freedom for a 2-sample *t* procedure.

1. Assuming that the population standard deviations for the smoking and nonsmoking groups are the same gives the largest degrees of freedom, $df = n_Y + n_N - 2 = 58$. For this example, it does not appear that the population standard deviations are the same since the sample standard deviations substantially differ. The hypotheses test for testing that the standard deviations are the same is not robust, so this method, called pooling, should not be done in practice. This will give the smallest margin of error, and the smallest p-value of the three methods for estimating the degrees of freedom.
2. The conservative estimate gives the smallest degrees of freedom, namely $df = \min(n_Y - 1, n_N - 1) = 29$. This simple method will give the largest margin of error, or the largest p-value of the three methods for estimating the p-value.
3. The Welch-Satterwaithe approximation, which is commonly used in introductory statistics courses, is between these extremes and uses the following degrees of freedom:

$$df = \frac{\left(\frac{s_Y^2}{n_Y} + \frac{s_N^2}{n_N}\right)^2}{\frac{1}{n_Y - 1}\left(\frac{s_Y^2}{n_Y}\right)^2 + \frac{1}{n_N - 1}\left(\frac{s_N^2}{n_N}\right)^2}$$

Applying this formula to the above example, gives $df = \frac{\left(\frac{9.3^2}{30} + \frac{14.1^2}{30}\right)^2}{\frac{1}{29}\left(\frac{9.3^2}{30}\right)^2 + \frac{1}{29}\left(\frac{14.1^2}{30}\right)^2} = 48.2$. The p-value is $P[t \geq -3.9] \approx 0.0001$. Students may not be aware that they are using this formula if they use software to compute the test statistics. Because the p-value is very small, we have strong evidence that the average lung capacity is impaired in children of adults who smoke.

Explanation of the degrees of freedom formula

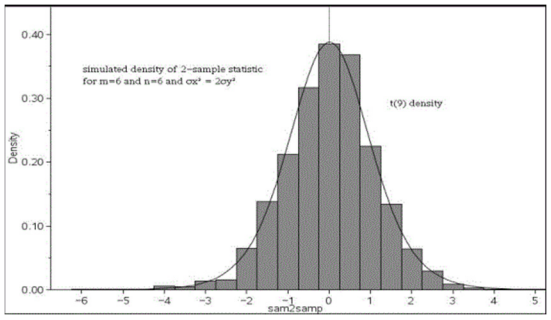
We will assume throughout the rest of the paper that X and Y are independent, Normally distributed random variables. Also assume that the sample sizes from X and Y are m and n,

respectively. The random variable $T = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{\mathrm{n}}}}$ is actually only approximated by the *t*-distribution with degrees of freedom of given by $df = \frac{\left(\frac{s_X^2}{\mathrm{m}}+\frac{s_Y^2}{\mathrm{n}}\right)^2}{\frac{1}{\mathrm{m}-1}\left(\frac{s_X^2}{\mathrm{m}}\right)^2+\frac{1}{\mathrm{n}-1}\left(\frac{s_Y^2}{\mathrm{n}}\right)^2}$.

For example, assume $\sigma_X^2 = 2\sigma_Y^2$ and $m = n = 6$. Then, T should have approximately the *t*-distribution with $\frac{\left(\frac{2\sigma_Y^2}{6}+\frac{\sigma_Y^2}{6}\right)^2}{\frac{1}{5}\left(\frac{2\sigma_Y^2}{6}\right)^2+\frac{1}{5}\left(\frac{\sigma_Y^2}{6}\right)^2} = 9$ degrees of freedom. Although the exact distribution of T is unknown, it can be simulated because of the following:

1. The numerator and denominator of T are independent.
2. The numerator of T has a Normal distribution with mean 0 and variance $\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{\mathrm{n}} = \frac{2\sigma_Y^2}{6}+\frac{\sigma_Y^2}{6} = \frac{1}{2}\sigma_Y^2$.
3. The variances in its denominator are multiples of the chi-squared distribution. Specifically, $S_X^2 = \frac{\sigma_X^2}{5}A = \frac{2\sigma_Y^2}{5}A$, where A has the chi-squared distribution with 5 degrees of freedom. Also, $S_Y^2 = \frac{\sigma_Y^2}{5}B$, where B has the chi-squared distribution with 5 degrees of freedom, and is independent of A.

The accompanying figure shows an empirical histogram for T using 2999 simulations. The theoretical *t*-distribution with 9 degrees of freedom appears to approximate the empirical distribution of T well.



Derivation of the degrees of freedom formula

We will reproduce a derivation of the Welch-Satterwaithe approximation that appears in many advanced statistics texts. Let $D = \bar{X} - \bar{Y}$. By independence, $\sigma_D^2 = \sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 = \frac{\sigma_X^2}{\mathrm{m}}+\frac{\sigma_Y^2}{\mathrm{n}}$. The random variable T can be rewritten as $\frac{Z}{\sqrt{U/r}}$, where $Z = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sigma_D}$, $U = \frac{r\left(\frac{S_X^2}{m}+\frac{S_Y^2}{\mathrm{n}}\right)}{\sigma_D^2}$, and r is any positive number. The random variable Z will have the standard Normal distribution, and Z and U are independent. The constant r will be chosen so that U will have approximately the chi-squared distribution with r degrees of freedom using the method of moments. Let V actually have a chi-squared distribution with *r* degrees of freedom. Then, T will be

approximated by $\frac{Z}{\sqrt{V/r}}$, which has the *t*-distribution with r degrees of freedom. This method is analogous to approximating a function by its Taylor polynomial by making some of its derivatives agree at a fixed point. Here, we will make the maximal number of moments of U and V agree by judiciously choosing r. The first moments are the same, namely $E[U] = \frac{r}{\sigma_D^2}\left(\frac{\sigma_X^2}{\mathrm{m}} + \frac{\sigma_Y^2}{\mathrm{n}}\right) = r$ and $E[V] = r$. We will choose r so that the second moments are the same. That is, $E[U^2] = E[V^2]$, or equivalently, $\mathrm{Var}(U) = \mathrm{Var}(V)$. By independence, $\mathrm{Var}(U) = \frac{r^2}{\sigma_D^4}\left(\frac{\mathrm{Var}(S_X^2)}{\mathrm{m}^2} + \frac{\mathrm{Var}(S_Y^2)}{\mathrm{n}^2}\right)$. The random variable $\frac{(m-1)S_X^2}{\sigma_X^2}$ has the chi-squared distribution with $m - 1$ degrees of freedom. So, $\frac{(m-1)^2\mathrm{Var}(S_X^2)}{\sigma_X^4} = 2(m-1)$, which implies $\mathrm{Var}(S_X^2) = \frac{2\sigma_X^4}{m-1}$. Similarly, $\mathrm{Var}(S_Y^2) = \frac{2\sigma_Y^4}{n-1}$. Thus, $\mathrm{Var}(U) = \frac{r^2}{\sigma_D^4}\left(\frac{2\sigma_X^4}{\mathrm{m}^2(m-1)} + \frac{2\sigma_Y^4}{\mathrm{n}^2(n-1)}\right)$. The variance of V is simply $\mathrm{Var}(V) = 2r$. Solve the equation $\mathrm{Var}(U) = \mathrm{Var}(V)$ for r to get the Welch-Satterthwaite approximation for the degrees of freedom:

$$df = \frac{\left(\frac{\sigma_X^2}{\mathrm{m}} + \frac{\sigma_Y^2}{\mathrm{n}}\right)^2}{\frac{1}{m-1}\left(\frac{\sigma_X^2}{\mathrm{m}}\right)^2 + \frac{1}{\mathrm{n}-1}\left(\frac{\sigma_Y^2}{\mathrm{n}}\right)^2}$$

In practice, the population standard deviations are approximated using the sample standard deviations. This idea can be used to approximate any linear combination of independent chi-squared random variables with a single chi-squared random variable.

We will show that the third moments of U and V disagree.

$$E[U^3] = \frac{r^3}{\sigma_D^6} E\left[\frac{S_X^6}{m^3} + 3\frac{S_X^4 S_Y^2}{m^2 n} + 3\frac{S_X^2 S_Y^4}{mn^2} + \frac{S_Y^6}{n^3}\right]$$

$$= \frac{r^3}{\sigma_D^6} E\left[\frac{U_X^6\sigma_X^6}{m^3(m-1)^3} + \frac{3U_X^4 U_Y^2 \sigma_X^4 \sigma_Y^2}{m^2 n(m-1)^2(n-1)} + \frac{3U_X^2 U_Y^4 \sigma_X^2 \sigma_Y^4}{mn^2(m-1)(n-1)^2} + \frac{U_Y^6\sigma_Y^6}{n^3(n-1)^3}\right]$$

$$= \frac{r^3}{\sigma_D^6} E\left[\frac{(m+3)(m+1)\sigma_X^6}{m^3(m-1)^3} + \frac{3(m+1)\sigma_X^4\sigma_Y^2}{m^2 n(m-1)} + \frac{3(n+1)\sigma_X^2\sigma_Y^4}{mn^2(n-1)} + \frac{(n+3)(n+1)\sigma_Y^6}{n^3(n-1)^3}\right]$$

and

$$E[V^3] = \frac{2^3\Gamma(3 + r/2)}{\Gamma(r/2)} = 8(2 + r/2)(1 + r/2)r/2 = (r+4)(r+2)r$$

Assume that $\sigma_X = \sigma_Y$. The following was obtained using a computer algebra system:

$$E[U^3] - E[V^3]$$

$$= \frac{8m(m-1)(m+n)^2 n(n-1)[m^4 - 2m^3 - 2m^2n^2 + 2m^2n + m^2 + 2mn^2 - 2mn + n^4 - 2n^3 + n^2]}{(m^3 - m^2 + n^3 - n^2)^3}$$

This last expression is zero if $m = n$. If $m \neq n$, then it is almost never zero by the Fundamental Theorem of Algebra.

The range of the Satterthwaite approximation df

We will show that the Satterthwaite approximation for the degrees of freedom lies between the conservative and pooling estimates. A computer algebra system was used to show the following:

$$(m+n-2)-r=\frac{[\sigma_X^2 n(n-1)-\sigma_Y^2 m(m-1)]^2}{\sigma_X^4 n^2(n-1)+\sigma_Y^4 m^2(m-1)}$$

The right side is nonnegative, so $r \le m+n-2$, the pooling estimate for r. In fact, $r = m+n-2$ if and only if $\frac{\sigma_Y^2}{\sigma_X^2}=\frac{n(n-1)}{m(m-1)}$. Note that the larger variance corresponds to the larger sample size.

We will now show that the lower bound for r is the conservative estimate. Without loss of generality, assume that $n \le m$. A computer algebra system was used to show the following:

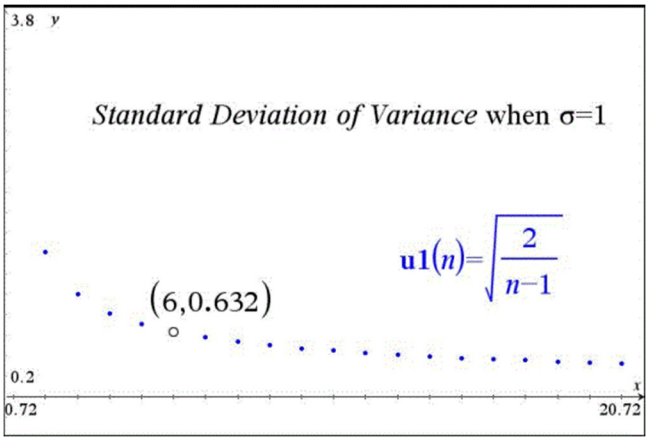
$$r-(n-1)=\frac{\sigma_X^2[\sigma_X^2(m-n)n+2\sigma_Y^2 m(m-1)]n(n-1)}{\sigma_X^4 n^2(n-1)+\sigma_Y^4 m^2(m-1)}$$

The right side is positive, so $r > n-1$. Since $n \le m$, $r > \min(m-1, n-1)$. Since the numerator on the right side of the expression for $r-(n-1)$ is a quadratic in m and the denominator is cubic in m, $r-(n-1)$ will converge to zero as $m \to \infty$ if all the other variables are fixed. Therefore the range of r is precisely $\min(m-1, n-1) < r \le m+n-2$.

Using the sample variances in practice

Recall that we actually approximate the population variances with the sample variances in the Satterthwaite approximation for the degrees of freedom.

Assume that S^2 is computed using n independent, identically distributed random variables with $\sigma^2 = 1$. Since $\mathrm{Var}\big(\chi^2(n-1)\big) = 2(n-1)$, it will follow that $\mathrm{Var}(S^2) = \frac{2}{n-1}$. The standard deviation of S^2 versus n is shown in the accompanying graph. The condition that $m \ge 6$ and $n \ge 6$ is plausible because 6 is the minimal n where the standard deviation of S^2 is less than half its maximal value at $n = 2$.

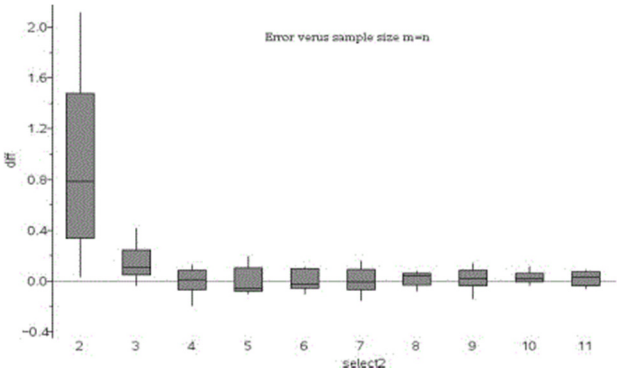


Satterthwaite error dependence on sample sizes

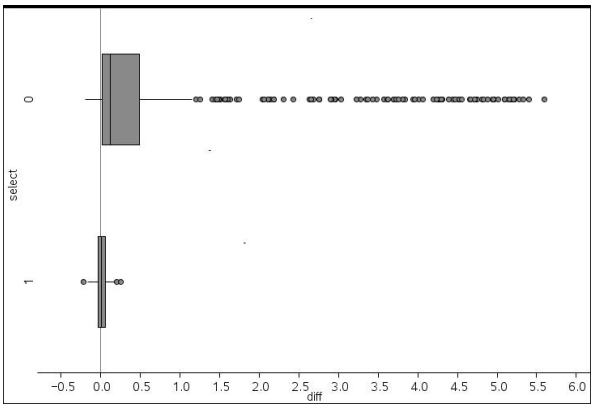
First, we will look at the special case where $\sigma_X = \sigma_Y$ and $m = n$. In this case, the Welch-Satterthwaite value for r is $\frac{4/n^2}{\frac{2}{n-1}\cdot\frac{1}{n^2}} = 2n-2$. This is the same value for the degrees of freedom as that obtained by pooling the samples. There is no error in this case since $U = \frac{2n-2}{\sigma_D^2}$.

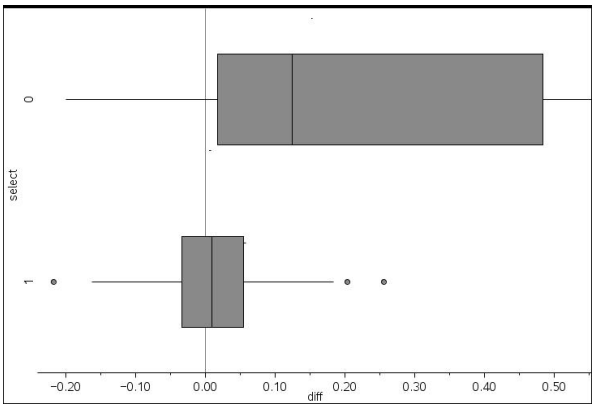
$\left[\frac{S_X^2}{m}+\frac{S_Y^2}{\mathrm{n}}\right]=\frac{\frac{2}{n}\cdot[(n-1)S_X^2+(n-1)S_Y^2]}{\frac{\sigma_X^2}{\mathrm{n}}+\frac{\sigma_Y^2}{\mathrm{n}}}$ actually has the chi-squared distribution with $2(n-1)$ degrees of freedom. That is, T will have exactly the *t*-distribution with $2n-2$ degrees of freedom.

The 2½ percentile for T was simulated and computed using the Satterthwaite approximation using $2 \le m \le 11$, $2 \le n \le 11$, and $\frac{\sigma_Y^2}{\sigma_X^2} \in \left\{\frac{1}{5},\frac{1}{4},\frac{1}{3},\frac{1}{2},1,2,3,4,5\right\}$. The error computed by subtracting these two approximations is shown in the accompanying figure when the sample sizes are the same. This percentile was chosen because it is needed when computing a 95 percent confidence interval.



The interquartile ranges are likely fairly constant for $n \ge 4$ because a simulation was used. The errors for $n = 3$ are significantly higher than those for $n = 4$ based on a Mann-Whitney test ($p = 0.03$). Also, the simulated percentile is usually larger than the Satterthwaite approximation, particularly for $n = 2$ or $n = 3$. This suggests that the Satterthwaite approximation will tend to give a smaller margin of error than the actual distribution of T would if m and n are small. The following boxplots are for the same error difference. Group 0 is when the sample sizes condition is not satisfied ($m \le 5$ or $n \le 5$), and Group 1 is when the sample sizes condition is satisfied ($6 \le m \le 11$ and $6 \le n \le 11$). The second graph shows more detail for the Group 1 boxplot.



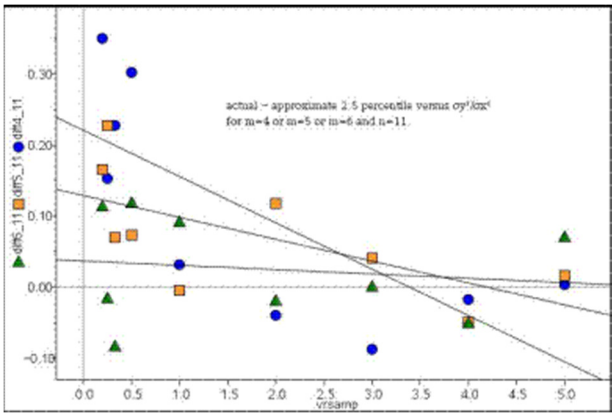


Group 0 had many large outliers and was skewed right, while Group 1 had only a few outliers and was symmetric. A nonparametric 95 percent confidence interval for the median of the Group 0 differences based on binomial distribution is $(0.106,0.155)$. Hence, we have evidence that the Satterthwaite approximate percentile tends to be too small when the conditions are not satisfied. The nonparametric confidence interval for the median of the Group 1 differences is $(0.001,0.019)$. Since this interval consists of all positive numbers, the errors also tend tobiased

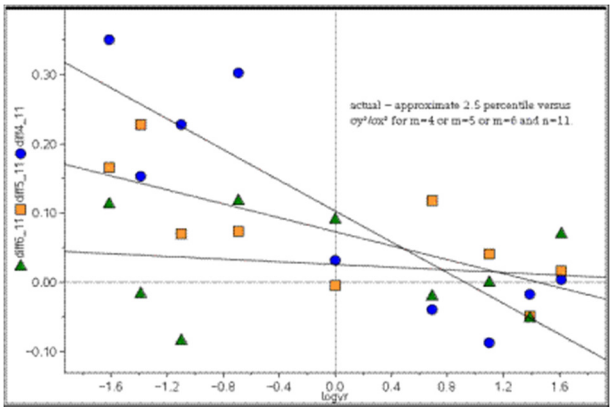
when the conditions are satisfied. However, ninety-five percent of the Group 1 differences are between -0.115 and 0.140, which is reasonable for an approximation.

Satterthwaite error dependence on variances

The accompanying plot shows the error in using the Satterthwaite approximation versus the variance ratio $\frac{\sigma_Y^2}{\sigma_X^2}$. The samples sizes are $n = 11$ while $m = 4$ (circle), $m = 5$ (square), and $m = 6$ (triangle). It is not surprising that the error tends to be less for larger m. Also, the error tends to decrease as the variance ratio increases if $m < 6$.



The linear relationship is stronger if the variance ratio is transformed using a logarithm. However, a nested F test did not show that the slopes were collectively significantly different in predicting the error ($F = 0.59$, $df = (2,21)$, $p = 0.58$).



Conclusions

The Satterthwaite approximate can be very inaccurate if the sample size condition is not satisfied, especially in the direction of underestimating the margin of error. If the sample size condition is satisfied, then the absolute error in the 2½ percentile is likely less than 0.14. The sample size condition is also needed so that the population standard deviations can be reasonably approximated by the sample standard deviations. Although it was not significant for our small number of simulations, the error appeared to be less if the larger sample size corresponded to the larger variance.

References

Allwood, M. (2008), "The Satterthwaite Formula for Degrees of Freedom in the Two-Sample *t*-Test."
http://apcentral.collegeboard.com/apc/public/repository/ap05_stats_allwood_fin4prod.pdf

Moore, D.S. (2013), "The Basic Practice of Statistics." WH Freeman

Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components." Biometrics Bulletin 2: 110–114

Welch, B. L. (1947), "The generalization of "student's" problem when several different population variances are involved." Biometrika 34: 28–35

http://en.wikipedia.org/wiki/Welch%E2%80%93Satterthwaite_equation, "Welch-Satterthwaite Equation"

Biographical Sketch

Michael Lloyd received his B.S in Chemical Engineering in 1984 and accepted a position at Henderson State University in 1993 shortly after earning his Ph.D. in Mathematics from Kansas State University. He has presented papers at meetings of the Academy of Economics and Finance, the American Mathematical Society, the Arkansas Conference on Teaching, the Mathematical Association of America, and the Southwest Arkansas Council of Teachers of Mathematics. He has also been an Advanced Placement statistics consultant since 2002.

Statistical Oddities in Baseball History

Fred Worth, Ph.D.
Professor of Mathematics

Abstract

When I was in first grade, I came home one day and explained to my mother how frustrated I was. They had not yet taught us how to do long division. That really bothered me because I wanted to be able to calculate batting averages for baseball players. So my mother taught me long division. Baseball is the ideal sport for people like me since statistics are far more a part of baseball than they are in any other sport. This paper is simply a list of some of the baseball statistical oddities I have found amusing over the years.

Players with at least 40 Home Runs but fewer than 100 Runs Batted In				
Player	**Year**	**Team**	**HR**	**RBI**
Duke Snider	1957	Dodgers	40	92
Mickey Mantle	1958	Yankees	42	97
Mickey Mantle	1960	Yankees	40	94
Harmon Killebrew	1963	Twins	45	96
Hank Aaron	1969	Braves	44	97
Rico Petrocelli	1969	Red Sox	40	97
Hank Aaron	1973	Braves	40	96
Davey Johnson	1973	Braves	43	99
Darrell Evans	1985	Tigers	40	94
Matt Williams	1994	Giants	43	96
Ken Griffey Jr.	1994	Mariners	40	90
Barry Bonds	2003	Giants	45	90