Comparison of Intervals

Dr. Brad Warner
Maj. Jim Rutledge

Department of Mathematical Sciences

United States Air Force Academy
Colorado Springs, Colorado 80840

JANUARY 1998

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

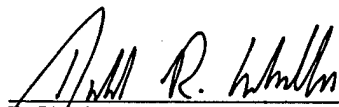DEAN OF THE FACULTY
UNITED STATES AIR FORCE ACADEMY
COLORADO 80840

This research report entitled "Comparison of Intervals" is presented as a competent treatment of the subject, worthy of publication. The United States Air Force Academy vouches for the quality of the research, without necessarily endorsing the opinions and conclusions of the author. Therefore, the views expressed in this article are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the US Government.

This report has been cleared for open publication and public release by the appropriate Office of Information in accordance with AFM 190-1, AFR 12-30, and AFR 80-3. This report may have unlimited distribution.

DONALD R. ERBSCHLOE, Lt Col, USAF
Director of Research

3 Feb 98
Date

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE 13 Jan 98	3. REPORT TYPE AND DATES COVERED Technical Report August 1997–December 1997

4. TITLE AND SUBTITLE
Comparison of Intervals

5. FUNDING NUMBERS

6. AUTHOR(S)
Dr. Brad Warner
Maj. Jim Rutledge

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Directorate of Education
2354 Fairchild Dr Suite 4K25
USAF Academy Co m80840

8. PERFORMING ORGANIZATION REPORT NUMBER
USAFA TR 98-1

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

This report deals with probability and statistics in regards to Chips Ahoy cookies and the and the "challenge" made by Nabisco.

14. SUBJECT TERMS
Intevals, Comparisons, probabilities

15. NUMBER OF PAGES
9

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE SAR	19. SECURITY CLASSIFICATION OF ABSTRACT SAR	20. LIMITATION OF ABSTRACT SAR

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

1 Introduction

Either being short on statisticians or employing a clever marketing scheme, Nabisco has recently issued the "Chips Ahoy! 1000 Chips Challenge"[1]. The cookie company is asking for the most "creative" way to confirm that there are 1000 chips in every 18 ounce bag of their chocolate chip cookies. This sounds like a simple enough challenge but several questions come to mind:

1. Where do we get the bags of cookies?

2. How many bags do we need?

3. How do we count the chips in a cookie?

4. What method of data analysis is most appropriate?

5. Who possibly has enough free time to partake in such a trivial pursuit?

Faculty members at the United States Air Force Academy are encouraged to involve cadets in interesting and sometimes relevant projects. Likewise, cadets are always eager to accept a challenge, especially if chocolate chip cookies are involved. Thus, we have the perfect mix for tackling the Chips Ahoy! challenge.

This paper will discuss how our Introductory Probability and Statistics course addressed the above questions in meeting the challenge. However, most of the attention will focus on the fourth question, the analysis, since as statisticians, this is our bread and butter, or milk and cookies. The first attempt to analyze the data involved a confidence interval for the mean. We quickly realized that this standard method presented in most introductory courses was inappropriate. We began to investigate other intervals which are appropriate for the problem but are **not** typically covered in an introductory course. We will share some insights on confidence, prediction, and tolerance intervals and use the Chips Ahoy! challenge to tie these concepts together.

2 Methods

An 18 ounce bag of Chips Ahoy! cookies contains approximately 16 servings at 3 cookies per serving. Nabisco claims that each bag of cookies contains over 1000 chips. A chip is defined as, "any distinct piece of chocolate that is baked into or on top of the cookie dough regardless of whether or not it is 100% whole."[1]

In order to obtain a representative sample of cookies, cadets had friends and relatives from across the country mail in bags of cookies to the Academy. In all, 275 bags were received from 46 different states. Once all the cookies were received from across the United States, it was time to count the chips. Chips were separated from the cookie by dissolving the cookie in water. Although the preferred method of cookie extraction is to dissolve the cookie in milk, water was used to minimize costs.

From the 275 bags of cookies a random sample of 42 bags was selected for counting. The remaining 233 bags were used to maintain the energy level of the cadets doing the counting. Data for the number of chips found in 42 bags of cookies can be found in Table 1.

Table 1: Raw Data for Number of Chips per Bag						
1363	1269	1293	1356	1137	1228	1087
1307	1325	1214	1294	1213	1154	1514
1279	1239	1219	1545	1135	1121	1185
1258	1098	1377	1440	1219	1546	1270
1215	1402	1191	1132	1295	1419	1247
1166	1345	1200	1143	1199	1103	1244

$$\overline{x} = 1261.57 \text{ and } s = 117.58$$

3 Results

Now that the "cookie is crumbled", what do we do? In this section we will discuss 6 different approaches to answering this question. In general the approaches involve either a confidence interval, prediction interval, or tolerance interval.

3.1 Confidence interval for the mean approach

Being in an introductory statistics course, the cadets immediately felt that a confidence interval would be the most appropriate analysis. A lower confidence bound for the average number of chips in a bag can be found by

$$\overline{y} - t_{1-\alpha,n-1}\frac{s_y}{\sqrt{n}}.$$

where $t_{1-\alpha,n-1}$ is the $100(1-\alpha)th$ percentile of a t distribution with $n-1$ degrees of freedom, $\overline{y}$ is the sample mean, s_y is the sample standard deviation, and n is the sample size. For our data set the 95% lower confidence bound is 1231 chips. Therefore, we are 95% confident that the average number of chips per bag is at least 1231. By confident we mean that 95% of similarly constructed intervals from many independent random samples would actually capture the true average number of chips.

There are several draw backs with this method. First, it is an estimate of the average number of chips per bag. Nabisco's claim centers around the number of chips in an individual bag **not** on the average number of chips. That is, Nabisco is interested in the entire distribution of chips not just where the distribution is centered. Second, the interval does not address the issue of 1000 chips. Therefore, even though the confidence interval looks promising, it really does not meet the challenge.

3.2 Confidence interval for the proportion conforming

After scratching our heads and eating a few cookies, we decided to look at a confidence interval for the proportion of bags with more than 1000 chips. From the data in Table 1 we can see that 100% of the bags have over 1000 chips. If we define the random variable P as the proportion of bags with more than 1000 chips, we can find a lower confidence bound for π, the true population proportion. Most introductory statistics courses develop the confidence bound for the proportion based on the normal approximation. However, this approximation is poor for small or large values of π. An exact lower confidence bound can be found using

$$1 - \beta(1 - \alpha, n - x + 1, x)$$

where $\beta(1 - \alpha, a, b)$ is the $100(1 - \alpha)th$ percentile of a Beta distribution with parameters a and b [2]. Most popular spread sheets such as Microsoft Excel have this function built in. For the our data $n = 42$ and $x = 42$, yielding a 95% lower confidence bound of 0.93. Thus we are 95% confident that the true proportion of bags with more than 1000 chips is at least 93%. Again, by confident we mean that 95% of similarly constructed intervals will capture the true population proportion of bags with more than 1000 chips.

Their are two potential problems with this approach. First, the lower bound of 93% seems to be very conservative. Perhaps by making some distributional assumptions a tighter bound can be obtained. Second, the defined limit (1000 chips) is less than our smallest observed value, thus the lower confidence bound is not sensitive to this limit. In other words, suppose that the Nabisco challenge had been for more than 500 chips per bag. The analysis on our data set would allow us to conclude that we are 95% confident that the proportion of bags with more than 500 chips is at least 93%. Intuitively one would think that we should have a larger proportion conforming as the limit is reduced.

3.3 Prediction interval

Although we are not dealing with fortune cookies, another approach to this problem is to try and predict the number of chips in future purchases of bags of cookies. One way to do this is with a prediction interval. The prediction interval is based on the difference between the next observation and the mean. Since these variables are independent, the variance of this difference is $\sigma^2\sqrt{1 + \frac{1}{n}}$. Thus the appropriate lower prediction bound is

$$\bar{y} - t_{1-\alpha,n-1}s_y\sqrt{1 + \frac{1}{n}}. \tag{1}$$

This interval is based on the assumption that individual observations are normally distributed [3]. Figure 1 is normal Q-Q plot of the number of chips per bag. This figure along with a Kolmogorov-Smirnov goodness of fit test [4], reveals there is no reason to reject the claim that the data are normally distributed. Therefore, we are 95% confident that the next bag of Chips Ahoy!
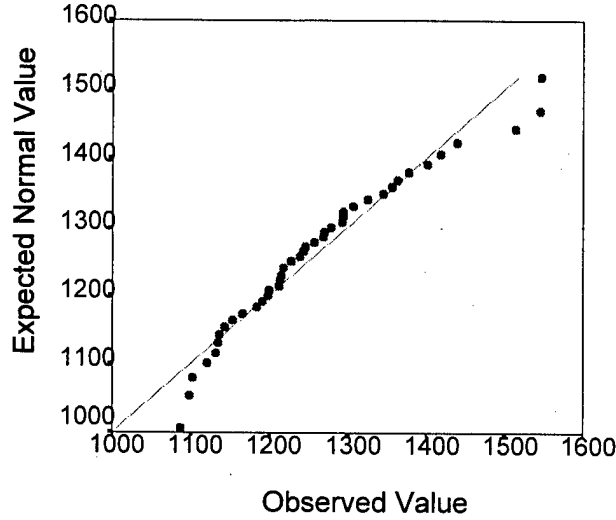
3

Figure 1: Normal Q-Q Plot of the Number of Chips in a Bag of Cookies.

cookies will contain more than 1061 chips. In this case confident means 95% of similarly constructed prediction intervals from independent samples will contain the future value from another independent sample.

A problem with this method is that it is based on the next bag of cookies and does not account for the population of bags of cookies. This method can be extended to the next m bags of cookies using a Bonferonni approach where Equation 1 becomes [5]

$$\bar{y} - t_{1-\alpha/m, n-1} s_y \sqrt{1 + \frac{1}{n}}.$$

Notice that the width of the interval increases with the number of predicted bags m. Another problem is that, again, this approach doesn't focus on 1000 chips.

3.4 Parametric tolerance interval

After consuming several bags of cookies, our *tolerance* for chocolate chip cookies was gone. But, it occurred to us that a tolerance interval may be appropriate for this analysis. A tolerance interval establishes bounds between which a fixed proportion of individual observations will occur. If we knew that the total number of chips per bag was normally distributed with mean, μ, and variance, σ^2, then 95% of the bags of cookies will contain more than $\mu - 1.645\sigma$ chips. Since we do not know μ and σ^2, we must estimate them with sample statistics. The use of the sample statistics in place of the population parameters adds

4

uncertainty in the proportion of the population covered by the estimate $\bar{y} - Z_{(1-\alpha)}s_y$. The tolerance interval accounts for this uncertainty by assigning a level of confidence to the proportion of the population falling in the interval. If the parent population is normally distributed, the tolerance interval has the form

$$\bar{y} - ks_y \tag{2}$$

where k adjusts for the use of sample statistics in place of population parameters. As the sample size (n) increases, k approaches the standard normal percentile (for a 5% level of significance, as n approaches ∞, k decreases to 1.645). Tables for k can be found in Eisenhart [6]. For our data, we can be 95% confident that **at least** 95% of the bags of Nabisco Chips Ahoy! cookies contain more than 1012 chips. It is important to note that the claim is **at most** 5% of all bags of cookies will contain less than 1012 chips. By confident we mean that 95% of similarly constructed intervals from independent random samples would cover at least 95% of the total number of chips per bag.

Notice that the tolerance statement centers on 1012 chips and not 1000. This method also assumes normality for the distribution of the number of chips in a bag of cookies. While this assumption is plausible for this data set (see figure 1), one may not always be willing to assume normality. In which case a distribution-free method may be preferable.

3.5 Distribution-free tolerance interval

To avoid the assumption of normality in the parent population, a distribution-free tolerance interval based on order statistics can be formulated. Based on the first order statistic, the smallest value in our data set, the lower tolerance interval has the following surprisingly simple relationship (see appendix for details):

$$p^n = \alpha. \tag{3}$$

Here p is the proportion of the population greater than the smallest data point (first order statistic), α is the level of significance, and n is the sample size. The distribution-free method will be demonstrated on our data even though the normality assumption is plausible. This is done for demonstration purposes only. For our data the first order statistic is 1087 chips; using a .05 level of significance in Equation 3 yields a value of .93 for p. Thus we are 95% confident that at least 93% of all bags of Chips Ahoy! cookies contain more than 1087 chips. This interval is exactly the same as we would have obtained by the method in Section 3.2 if we had defined the limit as 1087 chips instead of 1000.

Equation 3 also provides a simple formula to calculate sample size. Solving for n, the sample size, the result is:

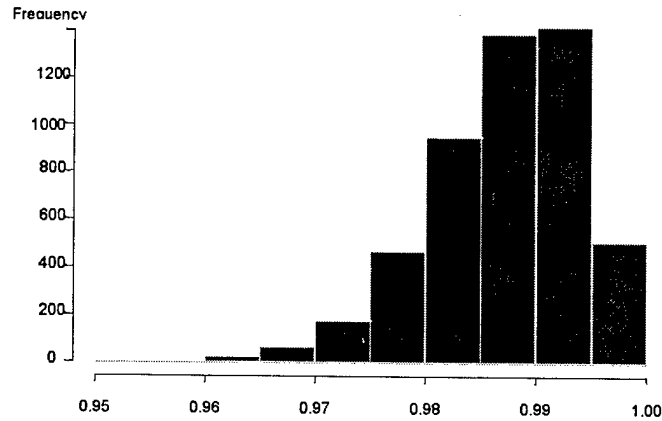$$n = \frac{\log_\pi(\alpha)}{\log_\pi(p)}. \tag{4}$$

5

Figure 2: Bootstrap Distribution of the Proportion with More than 1000 Chips.

3.6   Bootstrap based on the naive approach

A relatively simple approach that first semester statistics students want to take is to calculate $P(Y > 1000)$ where Y is the number of chips per bag of cookies. This probability can be interpreted as the proportion of all bags of cookies with more than 1000 chips. The probability calculation is based on the assumption that Y is normally distributed and that we know μ and σ^2. If we naively use the sample statistics as estimates of the parameters, this probability calculation is only a point estimate of the true probability. To account for the uncertainty associated with replacing the parameters with sample statistics, a bootstrap method [7] was used. In this method, we sample with replacement from our original data. Based on this sample, we calculate the sample mean and sample variance and use them as parameter estimates to calculate $P(Y > 1000)$ where we assume $Y \sim N(\bar{y}, s_y^2)$. This process is repeated to obtain the distribution of the proportion of bags of cookies with more than 1000 chips. A bootstrap lower confidence bound for this proportion was then calculated. Based on our data, Figure 2 is the bootstrap distribution of the proportion of bags with more than 1000 chips. The 5th percentile for this distribution is 97%. In other words, we are 95% confident that at least 97% of all bags of Chips Ahoy! cookies will contain more than 1000 chips. Notice that the interpretation is similar to Section 3.2; however, because of the normality assumption the confidence of this section is less conservative; it is narrower.

6

4 Discussion

One of the fascinating and fun aspects of statistics is the many approaches that can be used to solve the same problem. These varied approaches often have very subtle differences. In this paper, we were interested in finding out if all 18 ounce bags of Chips Ahoy! cookies contain more than 1000 chips. We first collected bags of cookies from across the United States and then extracted the chips in 42 randomly selected bags. For the analysis, we looked at confidence intervals, prediction intervals for a single bag, tolerance intervals (parametric and distribution-free), and a bootstrap confidence bound for the proportion of bags exceeding the specified 1000 chips (for an excellent discussion of the different types of intervals see Hahn and Meeker [5]). To reiterate, the results are as follows:

1. Using confidence intervals for the mean, we can be 95% confident that the average number of chips per bag is more than 1231 chips.

2. Using confidence intervals for the proportion exceeding a limit, we can be 95% confident that the proportion of bags with more than 1000 chips is at least 93%.

3. Using a prediction interval, we can be 95% confident that the next single bag of Chips Ahoy! cookies will contain more than 1061 chips.

4. The parametric tolerance interval suggests that we can be 95% confident that at least 95% of the bags contain more than 1012 chips.

5. The nonparametric tolerance interval suggests that we can be 95% confident that at least 93% of the bags contain more than 1087 chips.

6. Using the bootstrap approach we can be 95% confident that at least 97% of the population of Chips Ahoy! cookie bags contain more than 1000 chips.

It is reassuring to see that all the methods yield consistent answers. However the question still remains, did we meet the Nabisco challenge? From a statistical point of view we need to acknowledge that we can never be 100% confident that all the bags contain more than 1000 chips. As an example, say Nabisco wants to be 99.99% confident that 99.9999% of the bags of Chips Ahoy! cookies exceed the smallest observed count. Using Equation 4, they would need to count the chips in 9,210,336 bags of cookies. If all of the 9.2 million bags had over 1000 chips, they still would not be 100% confident that all bags have over 1000 chips. However, from a practical point of view, the bootstrap approach states that we are 95% confident that **at least** 97% of the population of Chips Ahoy! cookie bags contain more than 1000 chips. So, yes we did meet, beat, defeat, and eat the Nabisco Chips Ahoy! challenge!

References

[1] Chips Ahoy! Introductory Page [Online] Available: http://www.chipsahoy.com.

[2] Rutledge, J., and Bodenschatz, C., Confidence Intervals for Proportions when the Normal Approximation to the Binomial Distribution is Inappropriate. Journal of Air Force Operational Test and Evaluation, AFOTEC RP 99-1 Vol 1 No 1: 41-43.

[3] Montgomery, D. and Runger, G., Applied Statistics and Probability for Engineers. John Wiley and Sons, New York, 1994.

[4] Massey, F. J., The Kolmogorov-Smirnov Test for Goodness of Fit, Journal of the American Statistical Association, vol. 46, 1951.

[5] Hahn, G. and Meeker, W., Statistical Intervals: A Guide for Practitioners. John Wiley and Sons, New York, 1991.

[6] Eisenhart, C., Hastay, M. W., and Wallis, W. A., Techniques of Statistical Analysis. McGraw-Hill, New York, 1947.

[7] Efron, B., and Tibshirani, R., An Introduction to the Bootstrap. Chapman Hall, New York 1993.

[8] David, H. A., Order Statistics, John Wiley and Sons, New York. 1981.

Appendix

The distribution of the rth order statistic [8] is:

$$g(y_r) = \frac{n!}{(r-1)!(n-r)!} \left[\int_{-\infty}^{y_r} f(x)dx \right]^{r-1} f(x) \left[\int_{y_r}^{\infty} f(x)dx \right]^{n-r} \qquad (5)$$

where n is the sample size and $f(x)$ is the probability density function of the parent random variable X. Define a new random variable p as the proportion of the population that is greater than the rth order statistic, y_r:

$$p = \int_{y_r}^{\infty} f(x)dx. \qquad (6)$$

Transforming Equation 5 using Equation 6, yields the probability density function of p as

$$h(p) = \frac{n!}{(r-1)!(n-r)!} \left[1-p \right]^{r-1} \left[p \right]^{n-r}. \qquad (7)$$

Equation 7 is a Beta distribution with parameters $n - r + 1$ and r. Notice that Equation 7 does not depend on the distribution of X! For the case of the first order statistic $(r = 1)$, Equation 7 becomes

$$h(p) = np^{n-1} \quad 0 \le p \le 1. \qquad (8)$$

To find the proportion p of the population that exceeds the rth order statistic with $(1 - \alpha)\%$ confidence, use Equation 8 in

$$\int_0^p h(p)dp = \alpha$$

which reduces to

$$p^n = \alpha.$$