Developing Intuition for the Standard Deviation via Computer Graphics

Bret Larget

Duquesne University

July 22, 1996

KEY WORDS: Graphical Spread; Graphics; Histogram; Probability Density Function; Sampling Distribution.

Abstract

While students generally have a good intuition for means prior to entering an introductory course in statistics or probability, the concept of a standard deviation is likely to be quite new. Formulas for the standard deviation in various situations are given which must either be accepted on faith or derived algebraically, depending on the mathematical strength of the student.

In this paper, we present an imprecise measure of variation, the graphical spread, which students can measure directly from plots of probability density functions, probability mass functions, or histograms of data. Without any algebraic work, students can develop an intuition for how reasonable measures of spread, such as the standard deviation, depend on the parameters of a specific probability distribution or on the sample size.

1 Introduction

Before introducing the binomial distribution, even to classes where the mathematical preparation of students is weak, most students can answer questions such as, "If I toss this coin 100 times, about how many heads do you think I will get?", or "If I draw 12 cards from a deck of cards, replacing each card and reshuffling between draws, how many Hearts do you expect to see?". By asking a respondent from the class how he or she arrived at the answer, it is possible to elicit from the class an intuitive formula for the expected value of the binomial distribution, $\mu = np$, which turns out, of course, to be correct.

Popular textbooks which introduce probability (see Pitman (1993), Rice (1995), or Ross (1994) as examples) motivate the definition for the standard deviation of a random variable X, $SD(X) = (E[(X - \mu)^2)])^{1/2}$, as a reasonable measure of spread, but the relationship between the parameters of the distribution and this formula is left to nonintuitive algebraic manipulation in most cases.

Educational researchers have found that visualization can assist the development of intuition. In discussing technology and educational reform, Barr (1990, page 79) writes, "Visualization software ... can support intuitive learning." Biehler (1994) advocates using visualization to enhance the understanding of statistical concepts. The Teaching Statistics Visually project of Mathieson et al. (1995) is an effort to teach statistical concepts through visualization. These references constitute a small sample of the published argument for increasing the use of visualization software in teaching statistics and probability.

In this paper we describe a method by which students may informally discover, with no algebraic manipulation, how a measure of spread for a probability distribution depends on the parameters of the distribution. After making this discovery, a rigorous computation of the standard deviation of a random variable should yield an answer consistent with newly developed intuition instead of a formula that can be judged solely on algebraic simplicity, or agreement with an answer in the textbook.

In Section 2, we define the *graphical spread* for probability distributions. Section 3 shows how the graphical spread may be used to determine the dependency of the spread of a probability distribution on its parameters. With slight modification, this is also an appropriate way to motivate the formula $SD(\bar{X}) = \sigma/\sqrt{n}$ as shown in Section 4.

2 The Graphical Spread

We define the *graphical spread* to be the difference between the highest and lowest values for which the plot of the probability density function, probability mass function, or histogram on a computer graphics terminal is noticeably above zero, when the vertical scale is such that the mode of the graph is close to maximal height.

Figure 1 contains a plot of the binomial(100,0.5) probability mass function. All probabilities from 29 to 71 are visible above the horizontal axis, resulting in a graphical spread of $71 - 29 = 42$.

[Figure 1 about here]

For a continuous example, Figure 2 contains a plot of the exponential(1) distribution. Beyond 9.2, the graph of the density function is not even a pixel above the horizontal axis. For this plot, the graphical spread is $9.2 - 0 = 9.2$.

[Figure 2 about here]

The graphical spread is dependent on the choice of software, the size and quality of the terminal, the size of the graphics window, and the judgement of the viewer. Undoubtedly, if you use your current favorite statistics package with the computer on your desk to measure the graphical spread of the two distributions just mentioned, you will determine different values. The purpose of the graphical spread is to determine how spread changes as parameters change in an informal manner with minimal mathematical overhead. As long as the technical factors affecting the graphical spread are held consistent, this ambiguity will not affect our purpose.

3 Examples for Distributions of Random Variables

The merit of the graphical spread is that it is trivial to calculate if one has the capability to easily plot distributions. In my teaching, I have written functions in S-PLUS, which allow students to graph various probability distributions for any parameter values they choose by point-and-click with a mouse. Other software packages may be substituted. It is certainly preferable to make viewing distributions as simple as possible so that students may concentrate on the probabilistic concepts rather than working the computer.

A general procedure for determining the relationship between spread and a parameter is to begin by assuming that the relationship has the form

$$\text{Graphical Spread} \propto \theta^a$$

where θ is a parameter and a is the exponent we wish to find empirically. (In some instances, we need to assume a more complicated relationship.) If a distribution has more than one parameter, fix the values of all but one and treat them one at a time.

By taking logarithms on both sides, we arrive at the expression

$$\log(\text{Graphical Spread}) = \text{constant} + a \log \theta$$

Thus, reading the slope of a plot of $\log(\text{Graphical Spread})$ vs. $\log \theta$ reveals a.

This method can be used to informally explore the relationship between the spread of a distribution and its parameters for any probability distribution likely to be introduced in an introductory course. We demonstrate its use on two common examples, the binomial and exponential distributions.

3.1 The Binomial(n, p) Distribution

To see how spread depends on the parameter n, first fix p to be an arbitrary value such as $p = 0.5$. Next, measure the graphical spread for several values of n. Table 1 contains a summary of these measurements.

Figure 3 shows the plot of the logarithm of graphical spread versus $\log n$. The points fall nearly exactly on a line with slope very close to 0.5. Students may conclude that the spread of the binomial distribution for fixed p increases as the square root of n.

Table 1: The Graphical Spread of the Binomial$(n, 0.5)$ Distribution for Various Values of n.

n	50	100	500	1000	5000	10,000
Graphical Spread	28	42	88	124	276	392

[Figure 3 about here]

As a practical matter, the relationship between the logarithm of the graphical spread and $\log n$ deviates sharply from linearity for small values of n. Prior to asking your students to undertake this activity, you will need to explore within your own teaching environment to determine an appropriate range of values from which to choose n.

If students have been exposed to regression, they can find the slope of a "good line" through the points by ordinary least squares. A slick alternative is easy-to-use software that allows students to visually overlay lines on a graphical display of the points. Another acceptable alternative is having the students use software to plot the points, and determine a good slope by eye with a straightedge on a paper printout. If students expect to find a round number, they should find the correct answer.

After determining how the graphical spread depends on n, we are ready to examine its relationship with p. In this case, an argument is needed to explain why the algebraic form of the relationship is more complicated. It is apparent that if p equals either 0 or 1 exactly, there will be no variability in the distribution and the graphical spread will be zero. This reasoning leads to the trial relationship

$$\text{Graphical Spread} \propto (p(1-p))^a$$

so that

$$\log(\text{Graphical Spread}) = \text{constant} + a\log(p(1-p))$$

Table 2 contains some values of the graphical spread where p varies and $n = 500$, and Figure 4 shows a plot of the logarithm of the graphical spread versus $\log(p(1-p))$. (This relationship will not be linear if n is too small or if p is too close to 0 or 1, so some guidance may be needed.) Again, the points fall almost exactly on a line whose slope is very close to 0.5, so we may conclude that the graphical spread varies as the square root of $p(1-p)$.

Combining the two pieces of information, we reach the conclusion that

$$\text{Graphical Spread} \propto \sqrt{np(1-p)}$$

After completing this exercise, a rigorous derivation of the standard deviation resulting in the same formula should agree with the student's newly developed intuition on the spread of the binomial distribution.

3.2 The Exponential Distribution

The exponential distribution has density function $f(x) = \lambda\exp(-\lambda x)$ for $x \geq 0$. Table 3 shows the value of the graphical spread for some values of λ and the graph of the logarithm of the graphical spread versus $\log\lambda$ is in Figure 5.

Table 2: The Graphical Spread of the Binomial$(500, p)$ Distribution for various values of p.

p	0.1	0.2	0.3	0.4	0.5
Graphical Spread	52	71	81	87	88

Table 3: The Graphical Spread of the Exponential Distribution for various values of λ

λ	0.1	0.5	1	5	10
Graphical Spread	92.5	18.5	9.25	1.85	0.925

The points in Figure 5 fall exactly on a line with a slope of -1. Hence, we conclude that the graphical spread is inversely proportional to λ.

$$\text{Graphical Spread} \propto \frac{1}{\lambda}$$

At this point, the intuition of the student has been prepared to accept the solution to a rigorous derivation of the standard deviation of the exponential distribution.

4 The Sampling Distribution of the Sample Mean

When a student has become comfortable with finding variances of linear combinations of independent random variables, derivation of the formula

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

should be easily accomplished. An intuitive confirmation of this fact may be instilled by measuring the graphical spread of histograms from simulations from the sampling distribution of $\bar{X}$ for several sample sizes. This could effectively be done using essentially any choice for a population. (Asking students to do this for the Cauchy distribution might lead to an interesting class discussion!)

Here, we will use the Poisson(1) distribution. While this choice would allow us to easily plot the theoretical sampling distribution, we opt instead to plot a histogram of 1000 simulated sample means for each distribution sample size, an approach that would work for arbitrary finite populations or with distributions where i.i.d. sums are not easily described. Results are in Table 4 and Figure 6.

[Figure 6 about here]

Table 4: The Graphical Spread Measured from Histograms of Empirical Sampling Distributions for Various Sample Sizes

n	1	5	10	25	50	100	250	500
Graphical Spread	5	2.8	2.0	1.3	1.2	0.7	0.4	0.35

Since the software chooses boundaries of the histogram to be round numbers and we are estimating the sampling distribution empirically with 1000 randomly chosen points, the data plotted in Figure 6 is subject to much more variation from linearity than the previous examples. However, a line with a slope of -0.5 does pass through the middle of these points, and the formula for the standard deviation of $\bar{X}$ is empirically verified.

5 Conclusion

Students generally have a well-developed intuition for means prior to taking a course in probability or statistics. In contrast, intuition for standard deviations must be developed. The graphical spread allows students to develop an intuition for how the spread of a distribution changes when parameter values change, with very little mathematical background. The only requirement is the capability of easily plotting probability distributions. In the case of sampling distributions, students need to be able to easily plot distributions of sample means. There are many popular statistical software packages that may be used for this purpose. Developing an intuition for an informal measure of spread should add to the understanding gained in subsequent rigorous computations of standard deviation.

6 Acknowledgements

The author wishes to thank Joan Garfield and Nicholas Maxwell for their generous assistance in finding references for this article.

References

[1] Barr, D. (1990) "A solution in search of a problem: The role of technology in educational reform." *Journal for the Education of the Gifted*, 14(1), 79–95.

[2] Biehler, R. (1994) "Requirements for and Ideal Software Tool in Order to Support Learning and Doing Statistics." Paper presented at the Fourth International Conference on Teaching Statistics, Marrakech, Morocco.

[3] Mathieson, K., Doane, D.P., and Tracy, R.L. (1995) "A program for visualizing comparisons between two normal distributions," *Journal of Statistics Education*, 3(1).

[4] Pitman, Jim (1993). *Probability.* New York: Springer-Verlag.

[5] Rice, John A. (1995). *Mathematical Statistics and Data Analysis (2nd edition).* Belmont, CA: Duxbury Press.

[6] Ross, Sheldon (1994). *A First Course in Probability (4th edition).* New York: Macmillan.

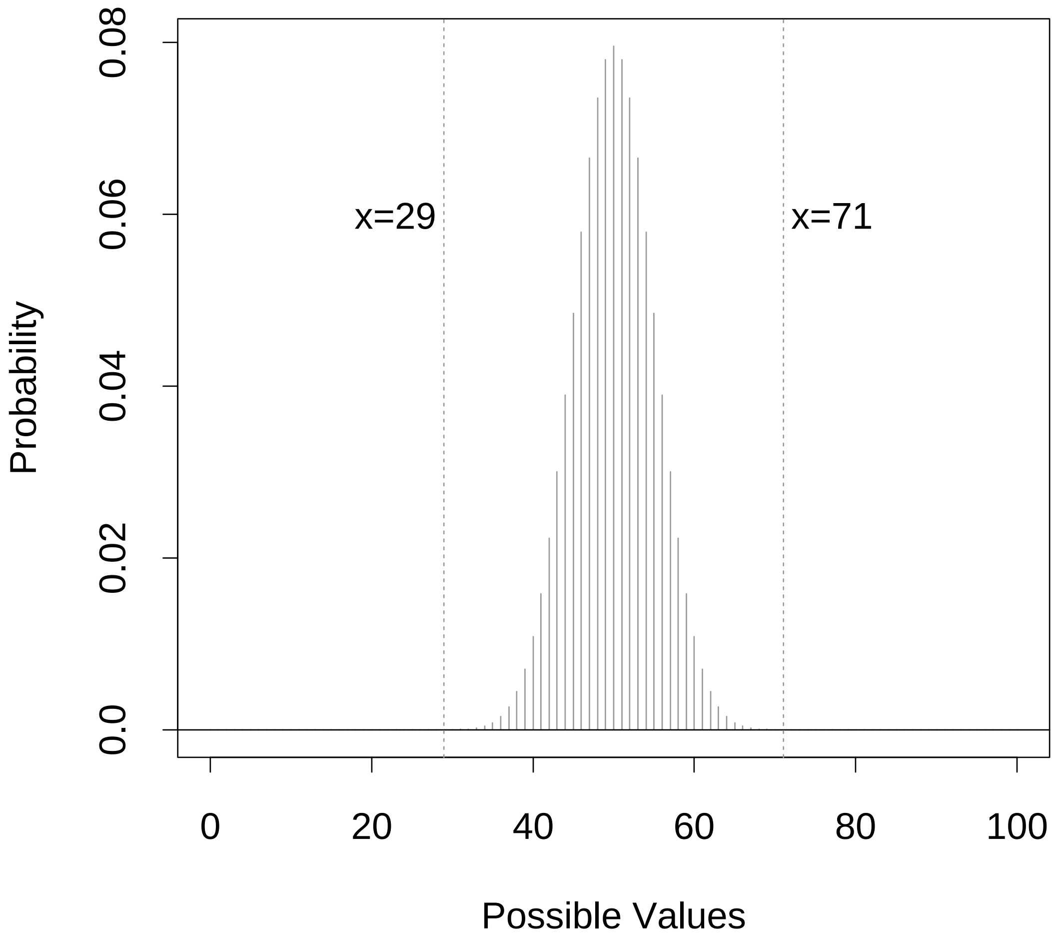

Figure 1: The Graphical Spread of the Binomial(100,0.5) Distribution

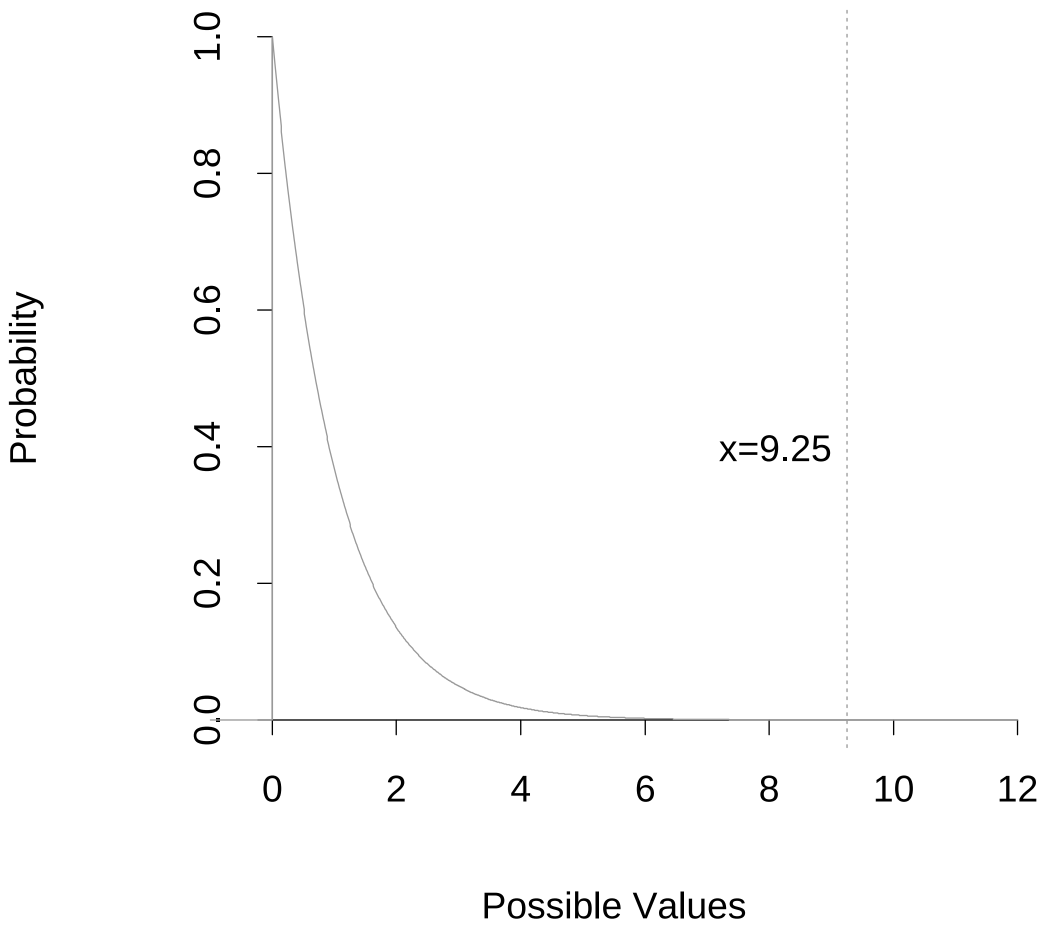

Figure 2: The Graphical Spread of the Exponential(1) Distribution

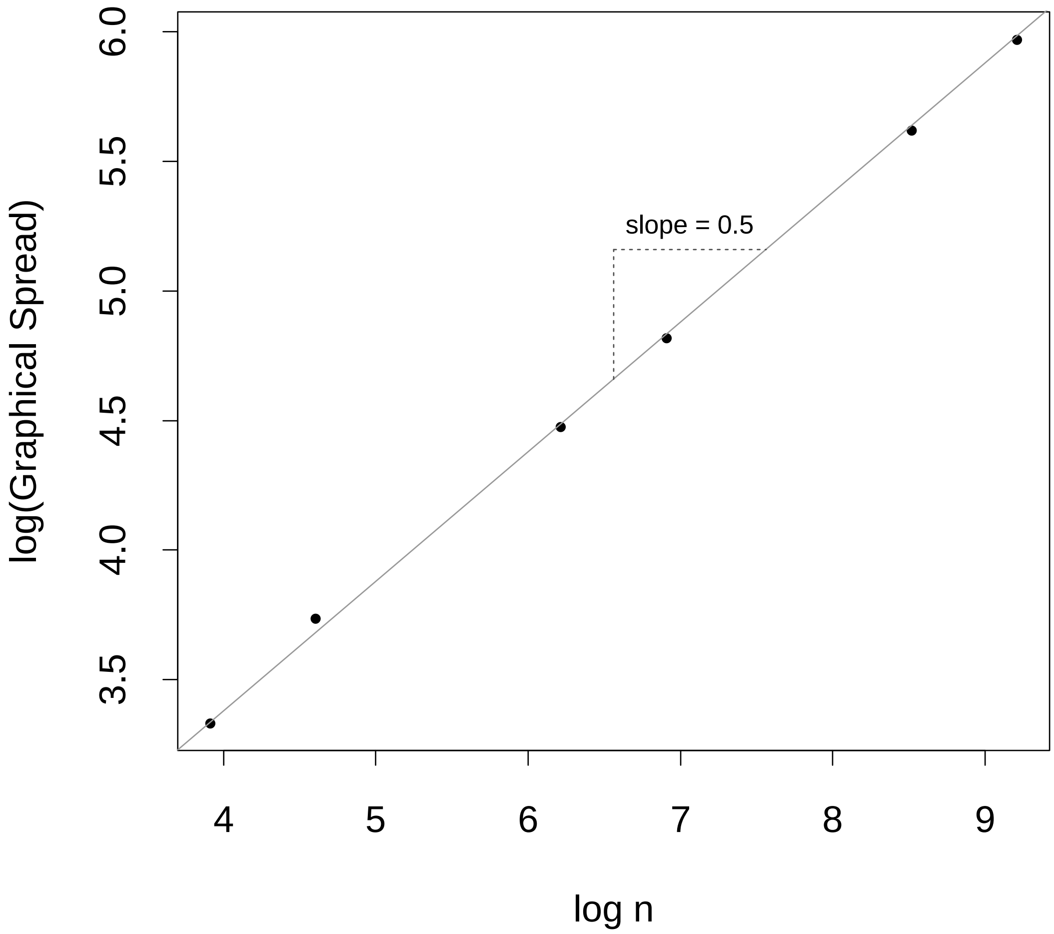

Figure 3: The Linear Relationship Between log(Graphical Spread) and $\log n$ in the Binomial$(n, 0.5)$ Distribution

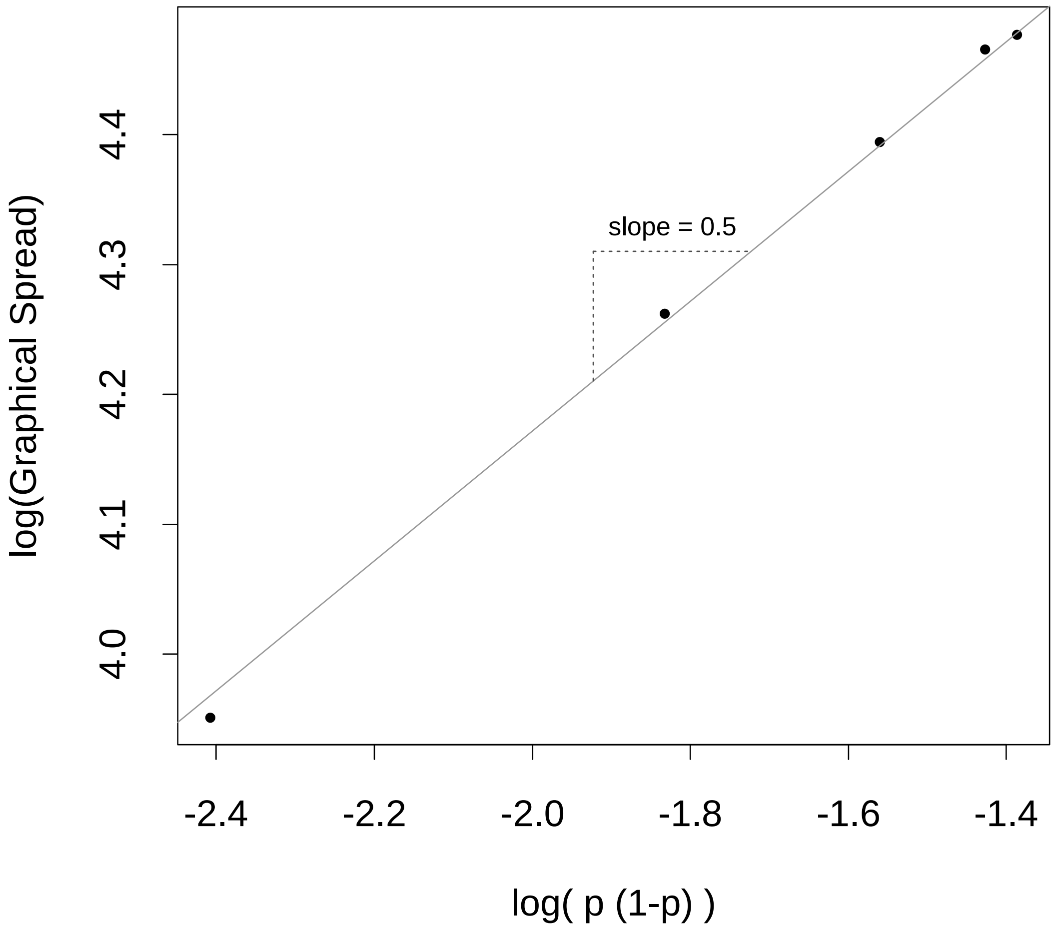

Figure 4: The Linear Relationship Between log(Graphical Spread) and $\log(p(1-p))$ in the $\text{Binomial}(500, p)$ Distribution

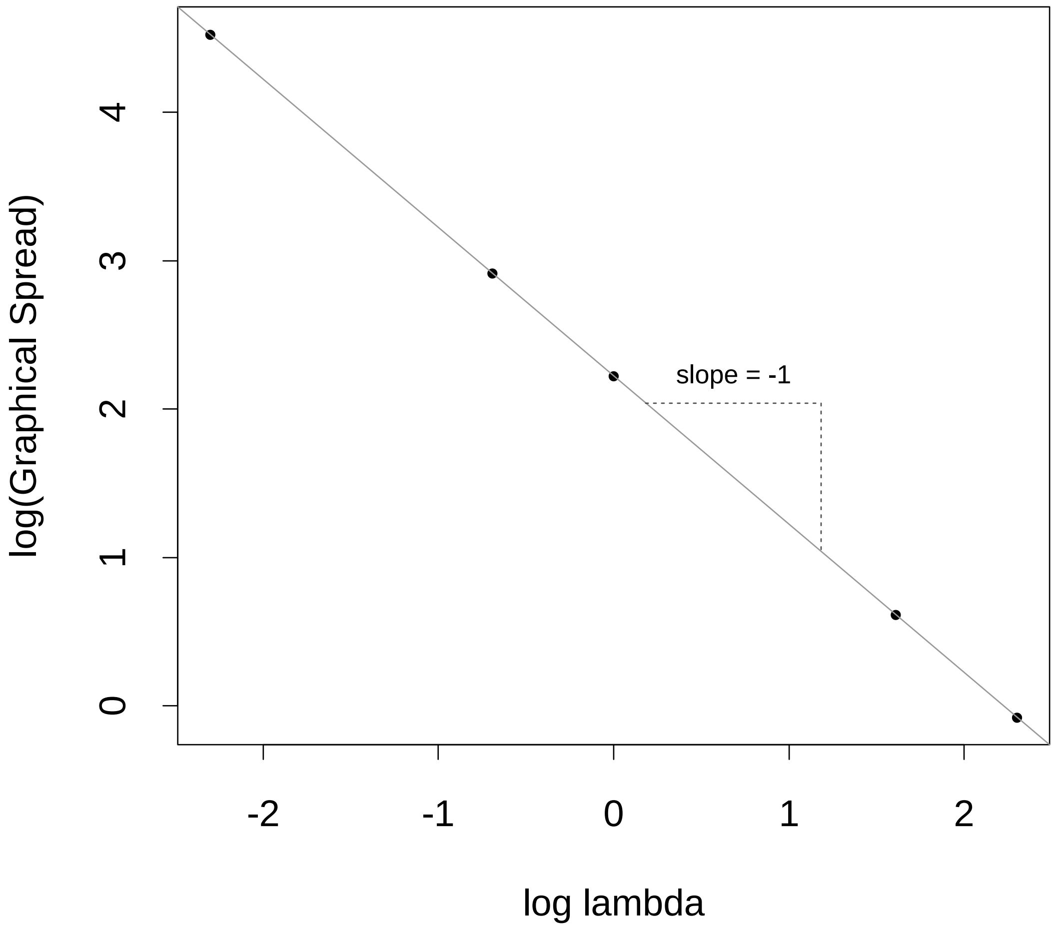

Figure 5: The Linear Relationship Between log(Graphical Spread) and $\log \lambda$ in the Exponential(λ) Distribution

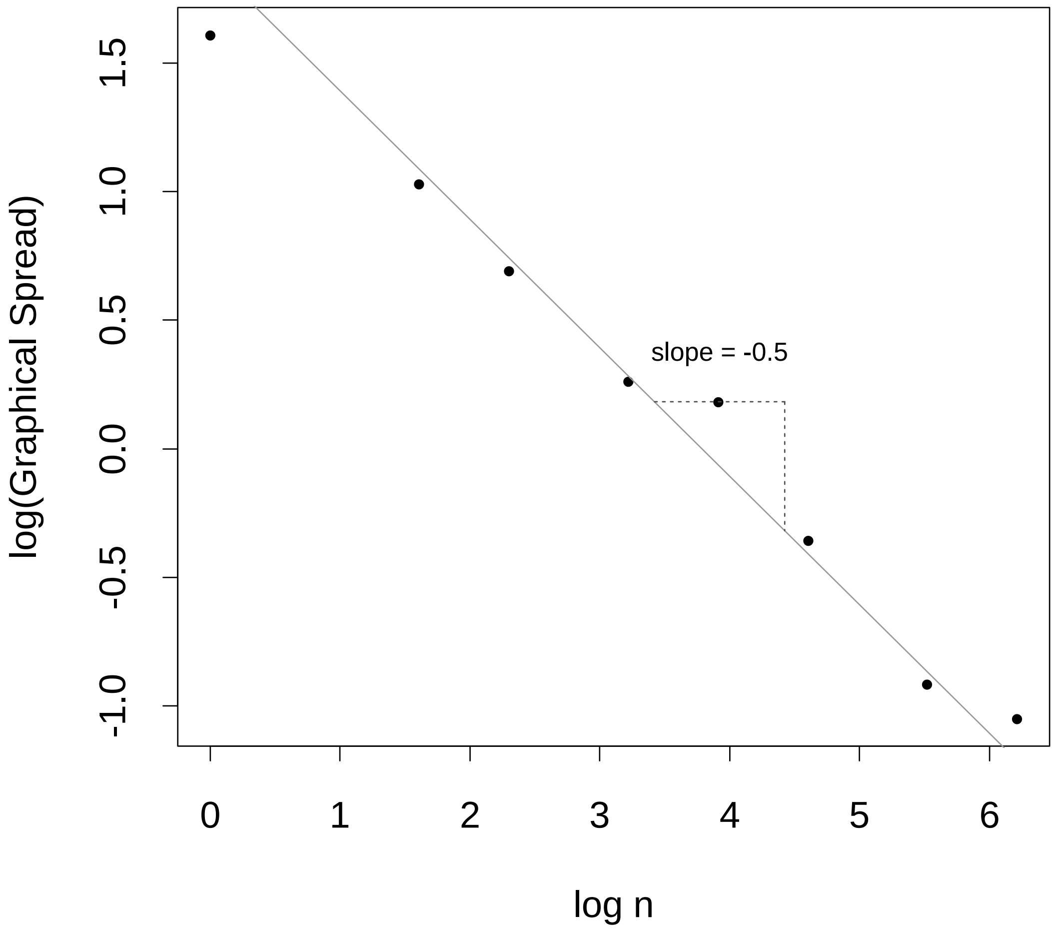

Figure 6: The Linear Relationship Between log(Graphical Spread) and $\log n$ in the Sampling Distribution of $\bar{X}$