The Central Limit Theorem

Consider a population that takes on the $(N = 5)$ values $X : \{1, 3, 4, 5, 7\}$, each with equal probability. If we simply observed individual values from this population, that would correspond to taking a sample of size 1, and determining the mean of the sample:

1	$\overline{X} = 1.0$	4	$\overline{X} = 4.0$	7	$\overline{X} = 7.0$
3	$\overline{X} = 3.0$	5	$\overline{X} = 5.0$		

The population, or equivalently the sample mean with a sample of size 1, has expected value equal to

$$\mu = (1)(.2) + (3)(.2) + (4)(.2) + (5)(.2) + (7)(.2) = 4.0$$

and variance

$$\sigma^2 = (1-4)^2(.2) + (3-4)^2(.2) + (4-4)^2(.2) + (5-4)^2(.2) + (7-4)^2(.2) = 4.0.$$

Now say we take a random sample of size 2. The possibilities are:

$1, 3$	$\overline{X} = 2.0$	$3, 4$	$\overline{X} = 3.5$	$4, 5$	$\overline{X} = 4.5$
$1, 4$	$\overline{X} = 2.5$	$3, 5$	$\overline{X} = 4.0$	$4, 7$	$\overline{X} = 5.5$
$1, 5$	$\overline{X} = 3.0$	$3, 7$	$\overline{X} = 5.0$	$5, 7$	$\overline{X} = 6.0$
$1, 7$	$\overline{X} = 4.0$				

This gives the probability function of $\overline{X}$:

$\overline{X}$	p	$\overline{X}$	p	$\overline{X}$	p
2.0	.1	3.5	.1	5.0	.1
2.5	.1	4.0	.2	5.5	.1
3.0	.1	4.5	.1	6.0	.1

Thus, $E\left(\overline{X}\right) = 4.0$ and $V\left(\overline{X}\right) = 1.5$. So, $\overline{X}$ is *unbiased*, but its variability is lower than that of X itself (the mean of a sample of size 1).

Now consider a sample of size 3:

$1, 3, 4$	$\overline{X} = 2.67$	$1, 4, 7$	$\overline{X} = 4.00$	$3, 5, 7$	$\overline{X} = 5.00$
$1, 3, 5$	$\overline{X} = 3.00$	$1, 5, 7$	$\overline{X} = 4.33$	$4, 5, 7$	$\overline{X} = 5.33$
$1, 3, 7$	$\overline{X} = 3.67$	$3, 4, 5$	$\overline{X} = 4.00$		
$1, 4, 5$	$\overline{X} = 3.33$	$3, 4, 7$	$\overline{X} = 4.67$		

The probability function is:

$\overline{X}$	p	$\overline{X}$	p	$\overline{X}$	p
2.67	.1	3.67	.1	4.67	.1
3.00	.1	4.00	.2	5.00	.1
3.33	.1	4.33	.1	5.33	.1

Thus, $E\left(\overline{X}\right) = 4.0$ and $V\left(\overline{X}\right) = .667$. So, $\overline{X}$ stays around μ, with progressively smaller variation as n (the sample size) increases. This result is the essence of statistical inference: that samples can provide information about populations, and that the accuracy of this information increases with an increase in the sample size. In fact,

$$SD\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}},$$

or

$$SD\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}$$

(for an infinite population). This standard deviation, since it is referring to a statistic ($\overline{X}$), is sometimes called a *standard error*.

A semingly counterintuitive implication of this formula is that a random sample of 500 people taken from the entire U.S. population of 250,000,000 (a 1 in 500,000 sampling rate) is far more predictive than a random sample of 50 people out of a population of 2500 (a 1 in 50 sampling rate). The standard errors of $\overline{X}$ in these two cases are, respectively,

$$SD\left(\overline{X}\right) = \frac{\sigma}{\sqrt{500}}\sqrt{\frac{249999500}{249999999}} = .044721\sigma$$

and

$$SD\left(\overline{X}\right) = \frac{\sigma}{\sqrt{50}}\sqrt{\frac{2450}{2499}} = .140028\sigma,$$

showing that the larger sample yields a sample mean that is far more accurate.

Freedman, Pisani, Purves, and Adhikari, in their 1991 second edition of the book *Statistics*, give a nice anology to explain this result. Suppose you took a drop of liquid from a bottle for a chemical analysis. If the liquid in the bottle is well–mixed, the chemical composition of the drop (i.e., the sample) would reflect quite faithfully the composition of the whole bottle (i.e., the population), and it really wouldn't matter if the bottle was a test tube or a gallon jug. On the other hand, we would expect to learn more from a test tube–sized sample than from a single drop.

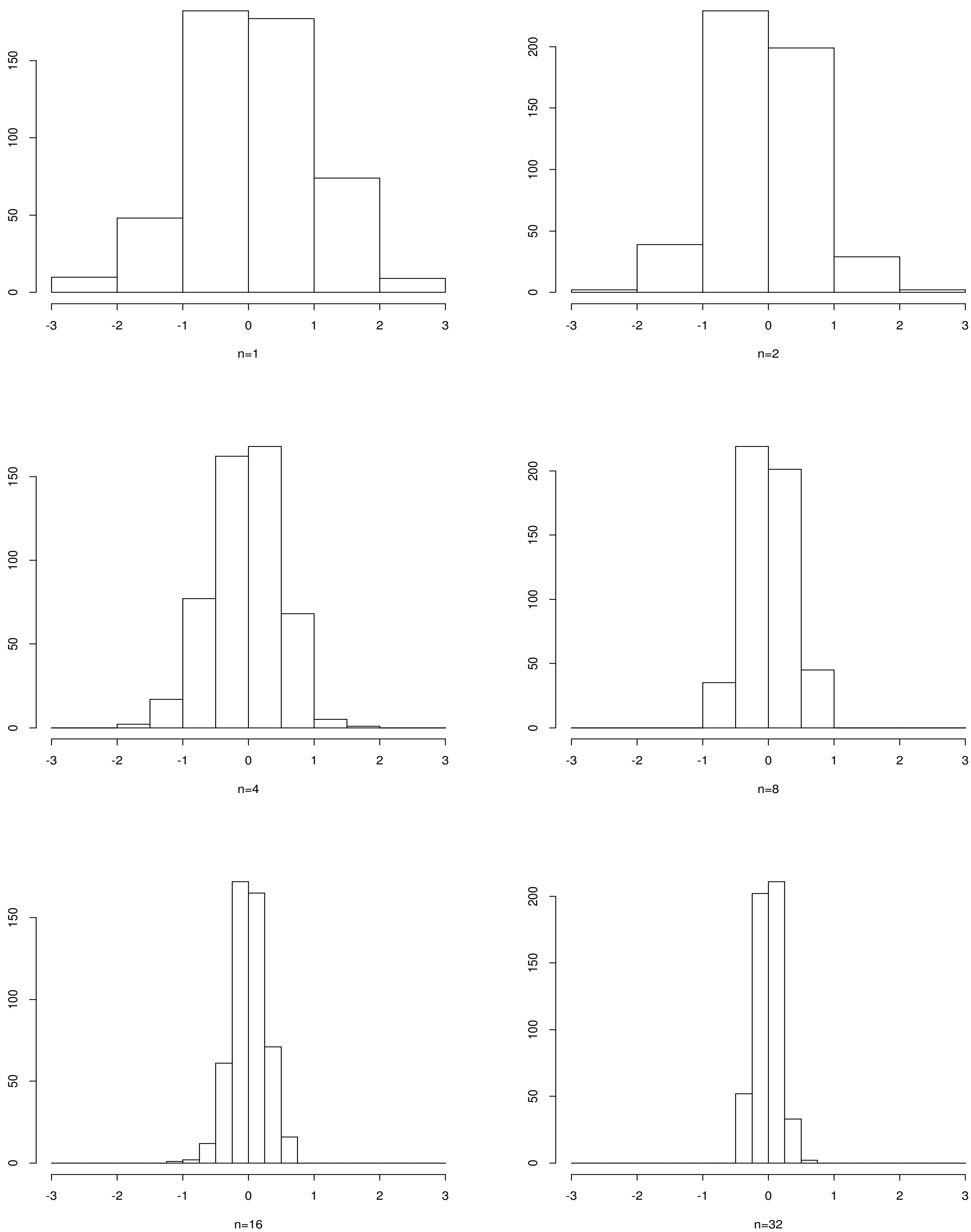
The **Central Limit Theorem** (CLT) adds one key result to the ones above. It says that for large enough sample size, the distribution of $\overline{X}$ (and, in fact, virtually any statistic) becomes closer and closer to Gaussian (normal), *no matter what the underlying distribution of X is.* This remarkable result implies that under virtually all circumstances it is possible to make probabilistic inferences about the values of population parameters based on sample statistics. It was first shown by Abraham de Moivre in the early 1700's for the Binomial process, and was extended to other situations by Carl Friedrich Gauss in the late 1700's.

The plots on the next three pages reinforce what is going on here. In each case, histograms are given for the distribution of the sample mean, based on various sample sizes, and taken from different populations (the histograms are based on 500 random draws of each sample size from each population; this is an example of *Monte Carlo simulation*). The first set of plots refers to a Gaussian population. Note that even for a sample of size 1 the sample mean has a normal distribution (of course — this is just the population itself). But, note also that the distribution of the sample mean gets progressively more concentrated around the population mean, reflecting its increasing accuracy.

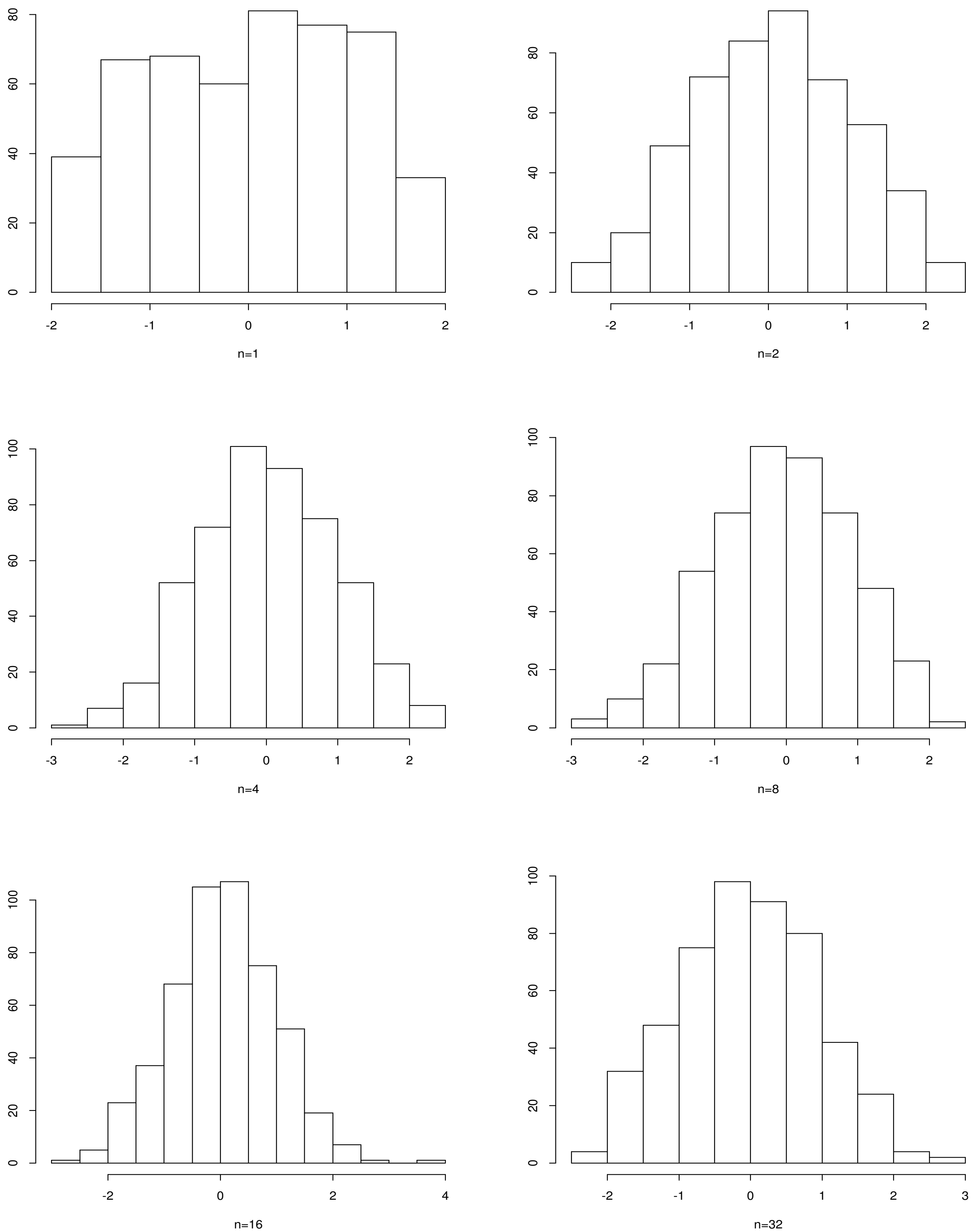
The second set of plots refer to a uniform population (the histograms refer to the standardized distribution, so the narrowing effect is not present). It is apparent that the Central Limit Theorem takes over for extremely small samples — perhaps even as small as $n = 2$.

Finally, the third set of plots refers to an exponential distribution. This distribution is highly skewed and long–tailed, so it takes much longer for the Central Limit Theorem to take over. Still, by $n = 32$, the Gaussian shape of the distribution of the mean is apparent.

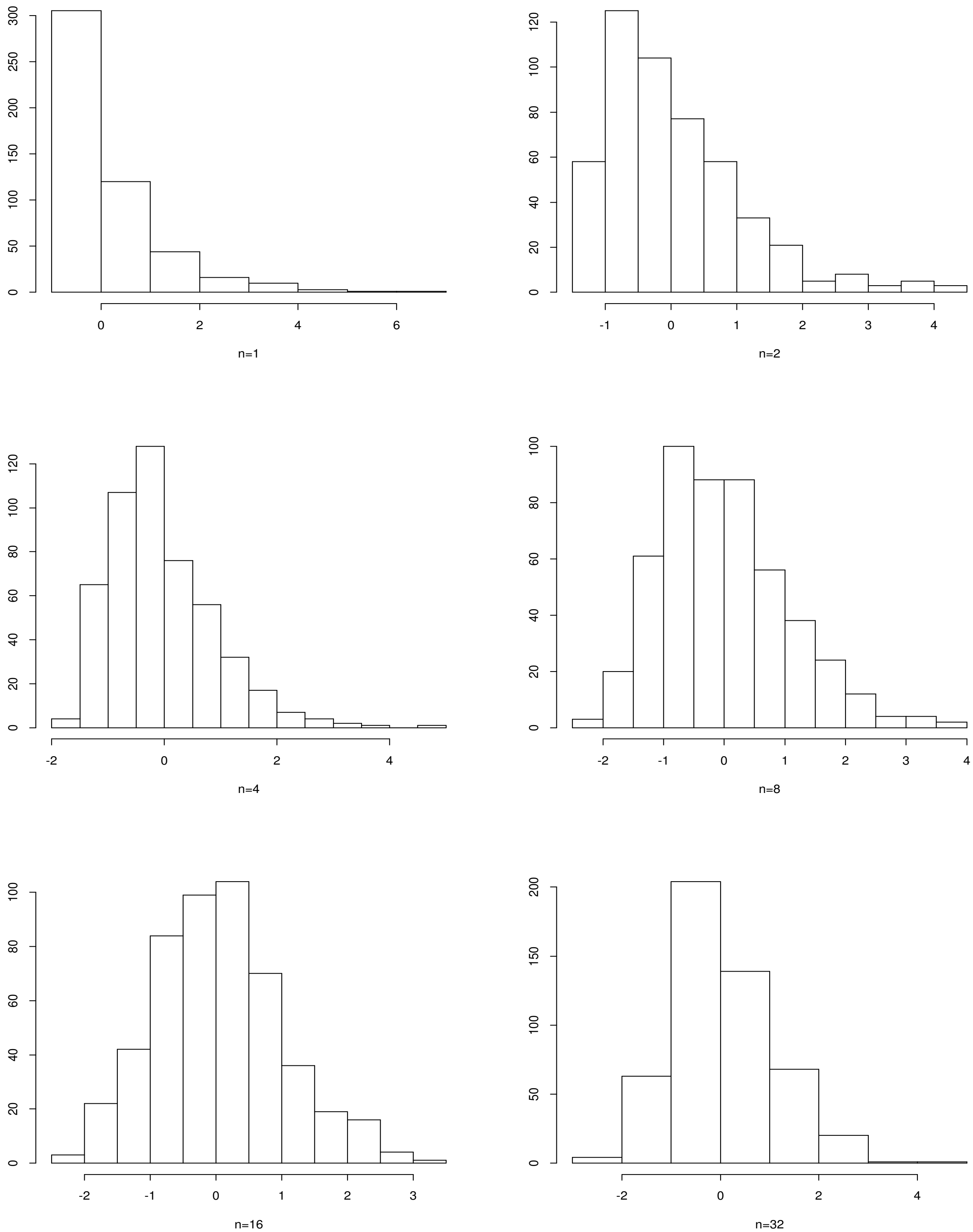
Sampling distributions of the mean from 500 samples of varying sizes from a normal population.



Standardized sampling distributions of the mean from 500 samples of varying sizes from a uniform population.



Standardized sampling distributions of the mean from 500 samples of varying sizes from an exponential population.



Consider the following application of these results: a labor contract states that workers will receive a monetary bonus if the average number of defect–free products is at least 100 units per worker per day (thus, individual worker quality is the random variable here). While the mean defect–free number per day is open to question (and is the subject of the contract issue), it is known that the standard deviation of this number is $\sigma = 10$ (one way this might be possible is if the *location* of quality is dependent on worker skill, while the *variability* of quality is based on the machinery being used; still, this is admittedly somewhat unrealistic, and we will discuss ways of proceeding when σ is unknown in a little while). Management proposes the following way to evaluate quality: a random sample of 50 workers will be taken, and their quality (number of defect–free units produced) will be examined for a recent day. The average of the defect–free values, $\overline{X}$, will be calculated, and if it is greater than 104, all workers will receive the bonus. Should labor agree to this proposal?

The Central Limit Theorem provides a way to answer this question. Say workers have, in fact, achieved the target average quality of 100. Given this, what are the chances that they will get the bonus? We are interested in

$$\begin{aligned} P\left(\overline{X} > 104\right) &= P\left(Z > \frac{104 - 100}{10/\sqrt{50}}\right) \\ &= P(Z > 2.83) \\ &= .0023 \end{aligned}$$

So, even if labor has achieved its goal, the chances that they'll get the bonus are exceedingly small. They should have the $\overline{X}$ target set to a smaller number — say 101 (what would the probability be then?). Say labor's counteroffer of awarding the bonus when $\overline{X}$ is greater than 101 is adopted; what are the chances of management awarding the bonus when average quality is, in fact, 99 defect–free units per worker per day? Can you think of any way of devising a method that would be more likely to satisfy **both** management and labor? (that is, will be likely to lead to a bonus when defect–free productivity is high enough, but unlikely to lead to a bonus when it is too low).

Note, by the way, that the Central Limit Theorem also can be used to derive probability statements about sums of independent observations, since the two probabilities $P(\overline{X} > c)$ and $P(\sum x_i > nc)$, for example, are identical. It also applies for weighted sums. Consider a portfolio of stocks, with p_i of the portfolio invested in stock i ($\sum_i p_i = 1$). The total return of the portfolio is $R = \sum_i p_i r_i$, where r_i is the return of the ith stock. The expected portfolio return is $E(R) = \sum_i p_i E(r_i)$, while the variance of the portfolio return is $V(R) = \sum_i p_i^2 V(r_i) + 2\sum_{i>j} p_i p_j Cov(r_i, r_j)$. The Central Limit Theorem says that if enough stocks are in the portfolio, the portfolio return will be (roughly) normally distributed, with mean $E(R)$ and variance $V(R)$.

"I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'Law of Frequency of Error.' The law would have been personified by the Greeks and deified, if they had known it. It reigns with serenity and complete self–effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect its sway. It is the supreme law of unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along."

— Francis Galton

"Psychohistory dealt not with man, but with man–masses. It was the science of mobs; mobs in their billions. It could forecast reactions to stimuli with something of the accuracy that a lesser science could bring to the forecast of a rebound of a billiard ball. The reaction of one man could be forecast by no known mathematics; the reaction of a billion is something else again... Implicit in all these definitions is the assumption that the human conglomerate being dealt with is sufficiently large for valid statistical treatment."

— Isaac Asimov, *The Foundation Trilogy*

"'Winwood Reade is good on the subject,' said Holmes. 'He remarks that, while the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician.'"

— A. Conan Doyle, *The Sign of Four*