Examples of Sampling Distributions

Distribution	***statistic whose variability it describes***
Binomial	proportion in SRS
Hypergeometric	proportion (finite N)
Poisson	small proportion or rate
Gaussian	mean, proportion, differences, etc (n large)
Student's t	$\frac{\bar{y} - \mu}{SE\{\bar{y} - \mu\}}$
F	ratio of variances (used for ANOVA)
Chi-Square	proportion(s); rate(s) ($n\pi$ large)

Three ways of calculating sampling variability

1 directly from the relevant discrete distribution by adding probabilities of the variations in question

e.g only 0.01 + 0.001 = 0.011 probability of ≥ 9 positives / 10 if $\pi = 0.5$
2.5% probability of getting a Poisson count of 5 or more if $\mu = 1.624$
2.5% probability of getting a Poisson count of 5 or less if $\mu = 11.668$

2 from specially-worked out distributions for more complex statistics calculated from continuous or rank data --

e.g. student 's t, F ratio, χ^2, Wilcoxon,

3 [very common] from the Gaussian approximation to the relevant discrete or continuous distribution -- by using the standard deviation of the variation in question and assuming the variation is reasonably symmetric and bell-shaped [every sampling distribution has a standard deviation -- its just that it isn't very useful if the distribution is quite skewed or heavy-tailed]. *We give a special name (standard error) to the standard deviation of a sampling distribution in order to distinguish it from the measure of variability of individuals.* Interestingly, we havent given a special name to the square of the SD of a statistic -- we use Variance to denote both SE^2 and SD^2

Standard Error (SE) of a sample statistic

What it is

The SD of the different values of the sample statistic one would obtain in different random samples of a given size n.

Since we observe only one of the many possible different random samples of a given size, the SD of the sample statistic is not directly measurable. In this course, in computer simulations, and in mathematical statistics courses, we have the luxury of knowing the relevant information about each element in the population and thus the probabilities of all the possible sample statistics. e.g. we say if individual Y's are Gaussian with mean μ and standard deviation σ, then the different possible ybars will vary from μ in a certain known way. In real life, we don't know the value of μ and are interested in estimating it using the one sample we are allowed to observe. Thus the SE is usually an estimate or a projection of the variation in a conceptual distribution i.e. the sd of all the "might-have-been" statistics.

Use

If n large enough, the different possible values of the statistic would have a Gaussian distribution with a spread of 2-3 SE's on each side of the "true" parameter value [note the "would have"]

So, can calculate chance of various deviations from true value.
Can infer what parameter values could/ could not have given rise to the observed statistic

e.g.
if statistic is $\bar{y}$, we talk of SE of the mean (SEM)

SE($\bar{y}$) describes variation of $\bar{y}$ from μ

SD(y) describes variation of y from μ

Important, to avoid confusion in terms : See note [in material giving answer to Q5 of exercises on §5.2] on variations in usage of term $SE(\bar{y})$ vs $SD(\bar{y})$

M&M §5.2 Variability of the Mean of a Sample : Expectation / SE / Shape of its Sampling Distribution

- Quantitative variable (characteristic) of interest : Y
- N (effectively) infinite (or sampling with replacement)
- Mean of all Y values in population = μ
- Variance of all Y values in population = σ^2

- Sample of size n; observations $y_1, y_2, ..., y_n$
- Sample mean = $\frac{\sum y_i}{n} = \bar{y}$ (read "y-bar") or...

Statistic	E	SD($\bar{y}$) or Standard Error (Mean)
$\bar{y}$	μ_y	$\frac{\sigma_y}{\sqrt{n}}$

But what about the pattern (shape) of the variability?

The sampling distribution is frequency distribution (histogram, etc...) we would get if we could observe the mean (or other calculated statistic) of each of the (usually infinite number of) different random samples of a given size. It quantifies probabilistically how the statistic (used to estimate a population parameter) would vary around the "true parameter" from one possible sample to another. This distribution is strictly conceptual (except, for illustration purposes, in classroom exercises).

Relevance of knowing shape of sampling distribution:

We will only observe the mean in the one sample we chose; however we can, with certain assumptions, mathematically (beforehand) calculate how far the mean ($\bar{y}$) of a randomly selected sample is likely to be from the mean (μ) of the population. Thus we can say with a specified probability (95% for example) that the $\bar{y}$ that we are about to observe will be no more than Q (some constant, depending on whether we use 90%, 95%, 99%, ...) units from the population mean μ. If we turn this statement around, we can say that there is a 95% chance that the population mean μ (the quantity we would like to make inferences about) will not be more than Q units away from the sample mean ($\bar{y}$) we (are about to) observe. This information is what we use in a confidence interval for μ. We also use the samling distribution to assess the (probabilistic) distance of a sample mean from some "test" or "Null Hypothesis" value in statistical tests.

Example of the distribution of a sample mean:

When summing or averaging n 0/1 values, there are only n+1 unique possibilities for the result. However, If we were studying a variable e.g. cholesterol or income that was measured on a continuous scale, the numbers of possible sample means would be very large and not easy to tabulate, so instead we take a simpler variable, that is measured on a discrete integer scale. However, the principle is the same as for a truly continuous variable.

Imagine we are interested in the average number of cars per household μ in a city area with a large number (N) of households. (With an estimate of the average number per household and the total number of households we can then estimate the total number of cars Nμ). It is not easy to get data on every single one of the N, so we draw a random sample, with replacement, of size n. [The sampling with replacement is simply for the sake of simplicity in this example -- we would use sampling without replacement in practice].

How much sampling variation can there be in the estimate we can get from the sample? What will the degree of "error" or "noise" depend on? Can we anticipate the magnitude of possible error and the pattern of the errors in estimation caused by use of a finite sample?

Suppose that 50% have no car, 30% have 1 and 20% have 2: i.e. of all the Y's, there are 0.5N 0's, 0.30N 1's and 0.20N 2's. [you would be correct to object "but we don't know this - this is the point of sampling"; however, as stated above, this is purely a conceptual or "what if" exercise; the relevance will become clear later]

- the average of the Y's is Σ Y x Prob(Y), i.e.

  $\mu = 0 \times 0.5 + 1 \times 0.3 + 2 \times 0.2$

  $= 0.7$; [the total of the Y's is 0.7N cars]

- half of the Y's are 0 and 20% are 2's, so the variance of Y, or

  $\sigma^2 = \Sigma(Y-\mu)^2 \times \text{Prob}(Y)$ is

  $\sigma^2 = (0 - 0.7)^2 \times 0.5 + (1 - 0.7)^2 \times 0.3 + (2 - 0.7)^2 \times 0.2$

  $= 0.49 \times 0.5 + 0.09 \times 0.3 + 1.69 \times 0.2$

  $= 0.61$ [sd, $\sigma = \sqrt{0.61} = 0.78$ is large relative to μ].

The variance of individuals is the variance of mean of sample with n=1

If we take a sample of size n = 2, and use the resulting $\bar{y} = \frac{y_1+y_2}{n}$ as our estimate of μ, [we write this as $\hat{\mu} = \bar{y}$], what can happen?

Distribution of sample mean when n=2

prob.	estimate of μ	direction	% in error
25%	$\bar{y} = \frac{0}{2}$ = **0.0**	low by 0.7	100
30%	$\bar{y} = \frac{1}{2}$ = **0.5**	low by 0.2	29
29%	$\bar{y} = \frac{2}{2}$ = **1.0**	high by 0.3	43
12%	$\bar{y} = \frac{3}{2}$ = **1.5**	high by 0.8	114
04%	$\bar{y} = \frac{4}{2}$ = **2.0**	high by 1.3	186

Consequently the possible estimates of the total number of cars i.e. 0, 0.5N, N, 1.5N and 2N will be "off" by the same percentages from 0.7N - and with the same high probability! It's not much good saying that "on average, over all possible samples" the sample will produce the correct estimate.

[Check:

average of all of the possible estimates = Σ estimate x Prob(estimate)

= 0 x 0.25 + 0.5 x 0.30 + 1.0 x 0.29 + 1.5 x 0.12 + 2.0 x 0.04

= 0.7 !!]

The possible errors in these estimates of 0.7 produced by a sample of 2 are very large. A sample of size n = 3 or n = 4 would give slightly less variable estimates. E.g. with n = 4, the distribution of the $3^4 = 81$ possible sample configurations, and their corresponding estimates of μ, can be enumerated manually as:

Distribution of sample mean when n=3

$\hat{\mu} = \bar{y}$	% Probability	% "Error"
0.00	06.25	- 100
0.25	15.00	- 64
0.50	23.50	- 29
0.75	23.40	+ 7
1.00	17.61	+ 43
1.25	09.36	+ 79
1.50	03.76	+ 114
1.75	00.96	+ 150
2.00	00.16	+ 186

There is still a good chance that the estimate will be a long way from the correct value of μ = 0.7.

If we are happy with an estimate that is not more than 50% in error, then the above table says that with a sample of n=4, there is a 23.50 + 23.40 + 17.61 or ≈ 65% chance that our sample will result in an "acceptable" estimate (i.e. within ±50% of μ). In other words, we can be 65% confident that our sample will yield an estimate within 50% of the population parameter μ.

For a given n, we can trade a larger % error for a larger degree of confidence and vice versa e.g. if n=4, we can be 89% confident that our sample will result in an estimate within 80% of μ or be 25% confident that our sample will result in an estimate within 10% of μ.

If we use a bigger n, we can increase the degree of confidence, or narrow the margin of error (or a mix of the two), since with a larger sample size, the distribution of possible estimates is tighter around μ. With n=100, we can associate a 20% error with a statement of 90% confidence or a 10% error with a statement of 65% confidence.

But one could argue that there are two problems with these calculations: first, they assumed that we knew both μ and the distribution of the individual Y's before we start ; second, they used manual enumeration of the possible configurations for a small n and Y's with a small number (3) of integer values.

What about real situations with samples of 10's or 100's from unknown distributions of Y's on a continuous scale?

The answer can be seen by examining the sampling distributions as a function of n in the 'cars per household' example, and in other examples dealing with Y's with a more continuous distribution (see Colton p103-108, A&B p80-83 and M&M 398-400). All the examples show the following:

(1) As expected, the variation of possible sample means about μ is less in larger samples. If we use the variance as a measure of scatter, then the variance (scatter) in the possible $\bar{y}$'s from samples of size n is σ^2/n, where σ^2 is the variance of the individual Y's. This is true regradless of the shape of the distribution of the individual Y's.

(2) If n is large enough, the distribution of possible $\bar{y}$'s resembles more and more a Gaussian distribution. This happens whether the individual Y's have a Gaussian or a Non-Gaussian distribution (see the various examples).

The importance of this is that the variability (variance or SD) of the possible sample means can be predicted just from the spread (σ^2) of the individual Y's in the population and from the sample size (n); we do not need to know exactly what the shape of the Y distribution is. The law that says that *the sampling distribution of a sample mean [or a sample proportion, or a sample slope or correlation..] is, for a large enough n [and under certain other conditions], Gaussian in shape no matter what the shape of the individual values*, is called the Central Limit Theorem.

If we use the notation " X ~ Distribution(μ_X, σ_X) " as shorthand to say that "X has a certain type of distribution with mean μ_X and standard deviation σ_X", then the **Central Limit Theorem** says that

When Y ~ ???????(μ_Y, σ_Y) , then

$\bar{y}$ **~ Gaussian(** μ_Y , $\frac{\sigma_Y}{\sqrt{n}}$ **) if n is large enough.**

The Gaussian approximation to certain Binomial distributions is an example of the Central Limit Theorem in action.

The individual Y's have a 2-point distribution: a proportion (1-π) have the value Y=0 and the remaining proportion π have Y=1.

The mean (μ) of all (0,1) Y values in population is π.

The variance, σ^2, of all Y values in population
$\sigma^2 = (0-\pi)^2 \times (1-\pi) + (1 - \pi)^2 \times \pi = \pi(1 - \pi)$.

Sample of size n; observations $y_1, y_2, \ldots, y_n$: a sequence of n 0's and 1's.

Sample mean $\bar{y} = \frac{\sum y_i}{n} = \frac{\text{number of 1's}}{n} = p$,

So ...

When Y ~ BINARY($\mu = \pi$, $\sigma = \sqrt{\pi[1-\pi]}$ **) , then**

p = $\bar{y}$ **~ GAUSSIAN(** π , $\frac{\sqrt{\pi[1-\pi]}}{\sqrt{n}}$ **) if n large enough**

Returning to the example above: if n = 100, then the SD of possible $\bar{y}$'s from samples of size n=100 is $\sigma / \sqrt{100} = 0.78 / 10 = 0.078$. Thus, we can approximate the distribution of possible $\bar{y}$'s by a Gaussian distribution with mean $\mu = 0.7$ and standard deviation of 0.078, to get ...

	Interval	Prob.	% Error
$\mu \pm 1.00SD(\bar{y}) = 0.7 \pm 0.078 =$	0.62 to 0.77	68%	-11% to +11%
$\mu \pm 1.50SD(\bar{y}) = 0.7 \pm 0.117 =$	0.58 to 0.81	87%	-17% to +17%
$\mu \pm 1.96SD(\bar{y}) = 0.7 \pm 0.143 =$	0.55 to 0.84	95%	-20% to +20%
$\mu \pm 3.00SD(\bar{y}) = 0.7 \pm 0.234 =$	0.46 to 0.93	99.7%	-33% to +33%

[The Gaussian-based intervals are only slightly different from the results of a computer simulation in which we drew samples of size 100 from the above Y distribution]

If this variability in the possible estimates is still not acceptable and we use a sample size of 200, the standard deviation of the possible $\bar{y}$'s is not halved (divided by 2) but rather divided by $\sqrt{2}$=1.4. We would need to go to n = 400 to cut the s.d. down to half of what it is with n = 100.

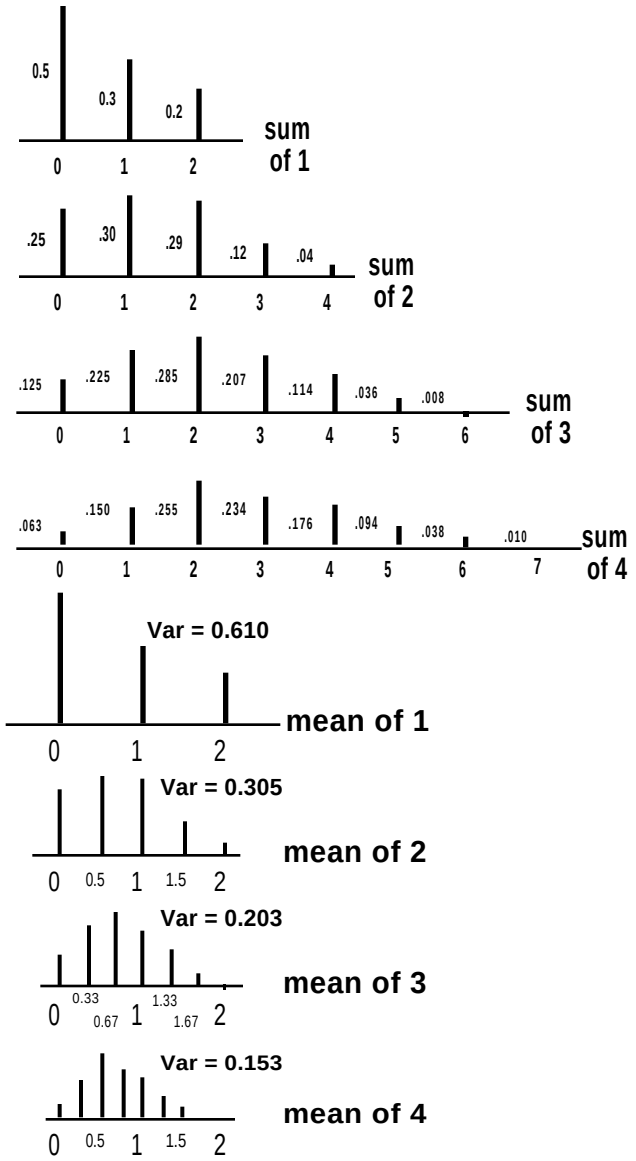
[Notice that in all of this (as long as we sample with replacement, so that the n members are drawn independently of each other), the size of the population (N) didn't enter into the calculations at all. The errors of our estimates (i.e. how different we are from μ on randomly selected samples) vary directly with σ and inversely with $\sqrt{n}$. However, if we were interested in estimating $N\mu$ rather than μ, the absolute error would be the same in the two scales.]

Message from diagram opposite:

Var (Sum) > Var of Individuals by factor of $\sqrt{n}$
Var (Mean) < Var of individuals by same factor

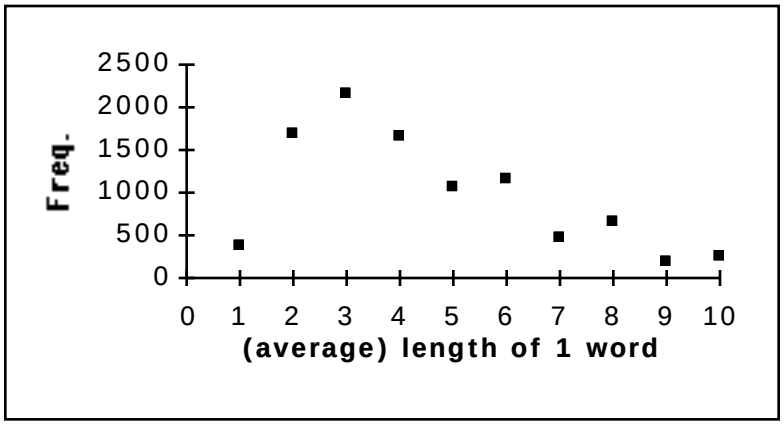
In addition, and also very important !!: Variation of sample means (or sums) is more Gaussian than Variation of individuals

Effect of n on Sampling behaviour of Sums & Means

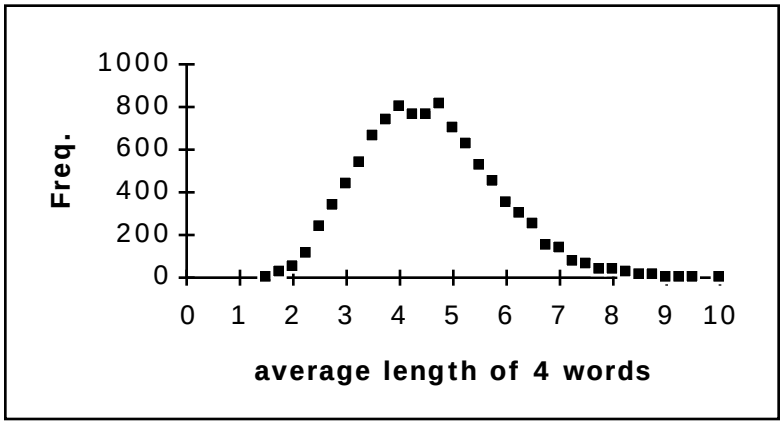


Another Example of Central Limit Theorem at work

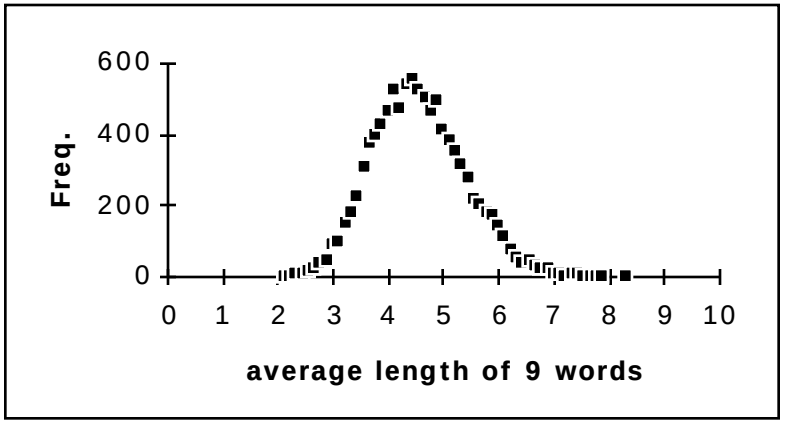
The variability in length of individual words...

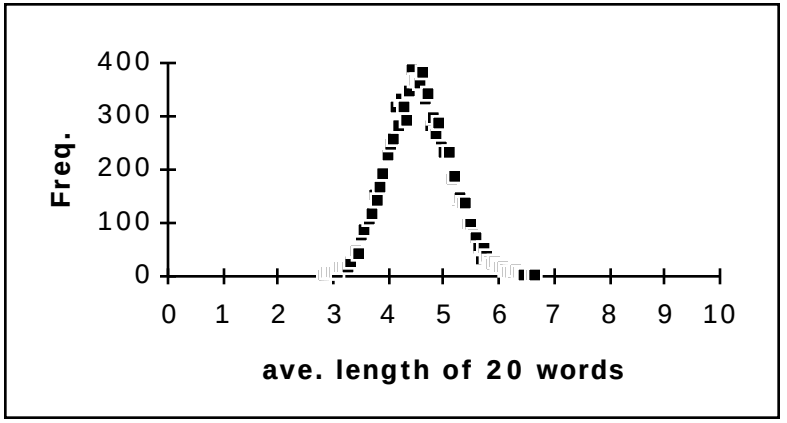


The variability in the average word length in samples of 4, 9, 20 words [Monte Carlo simulation]



The variation of means is closer to Gaussian than the variation of the individual observations, and the bigger the sample size, the closer to Gaussian. [i.e. with large enough n, you could not tell from the sampling distribution of the means what the shape of the distribution of the individual 'parent' observations. Averages of n=20 are essentially Gaussian (see observed vs fitted at right).





Variability in mean length of n=20 words

Mean [of means] 4.56
SD[of means] 0.56 Variance[of means] 0.3148

Quantiles	%ile	observed	fitted: mean+zSD	(z)
	99%	5.95	5.86	(2.32)
	95%	5.5	5.48	(1.96)
	90%	5.3	5.28	(1.28)
	75%	4.95	4.94	(0.67)
	50%	4.55	4.56	(0.00)
	25%	4.15	4.18	(–1.67)
	10%	3.85	3.84	(–1.28)
	5%	3.65	3.64	(–1.96)
	1%	3.35	3.26	(–2.32)