Examples of Sampling Distributions

Distribution	***statistic whose variability it describes***
Binomial	proportion in SRS
Hypergeometric	proportion (finite N)
Poisson	small expected 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 **Binomial** prob. of ≥ 9 [9 or 10] +ve / 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 haven't 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

An estimate of 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($\bar{y}$)

Variability of the Proportion / Count in a Sample : The Binomial Distribution

What it is

- The n+1 probabilities p_0, p_1, ... p_y, ... p_n of observing

  0 "positives"
  1 "positive"
  2 "positives"
  . ..
  y "positives"
  .
  . ..
  n "positives"

  in n independent binary trials
  (such as in simple random sample of n individuals)

- Each observed element is binary (0 or 1)

- 2^n possible sequences ... but only n+1 possible observable <u>counts or proportions</u>
  i.e. 0 / n, 1 / n, ... , n / n
  (can think of y as sum of n Bernoulli random variables)

- Apart from sample size (n), the probabilities p_0 to p_n depend on only <u>1 parameter</u>

  the probability (individual element will be +)
  or
  the proportion of "+" individuals in the population being sampled from

- Generally refer to this (usually <u>unknowable</u>) parameter by Greek letter π (sometimes θ)

- Inferences concerning π through <u>observed</u> p

	<u>Parameter</u>	<u>Statistic</u>
Hanley et al.	π	p = y/n
M&M	p	$\hat{p}$ = y/n
Miettinen	P	p = y/n

The Binomial Distribution

Shorthand

if y = # positive out of n

then "y ~ Binomial(n , π)"

How it arises

Sample surveys
Clinical trials
Pilot studies
Genetics
Epidemiology ...

Use

- to make inferences about π

(after we have observed a proportion
p = y/n in a sample of n)

- to make inferences about more complex situations

e.g... in Epidemiology

Risk Difference RD $= \pi_1 - \pi_2$

Risk Ratio RR $= \dfrac{\pi_1}{\pi_2}$

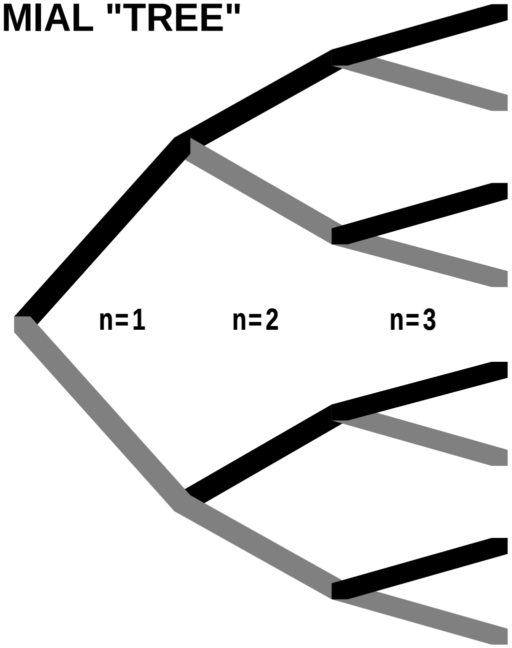
Odds Ratio OR $= \dfrac{\pi_1[1 - \pi_2]}{\pi_2[1 - \pi_1]}$

trend in several π's

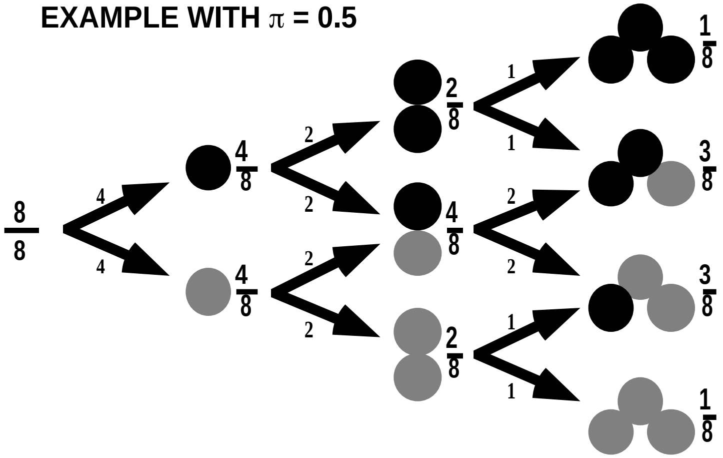
<u>NOTE</u> (see bottom of column opposite): M&M use the letter p for a population proportion and $\hat{p}$ or "p-hat" for the observed proportion in a sample. Others use the Greek letter π for the population value (parameter) and p for the sample proportion. Greek letters make the distinction clearer; note that when referring to a population mean, M&M do use the Greek letter μ (mu)!

Some authors (e.g., Miettinen) use upper-case letters, [e.g. **P**, **OR**] for parameters and lower-case letters [e.g. **p** , **or**] for statistics (estimates of parameters)

BINOMIAL "TREE"



EXAMPLE WITH π = 0.5



Calculations greatly simplified by fact that $\pi_1 = \pi_2 = \pi_3$. Can calculate prob. of any one sequence of y +'s & (n–y) –'s. Since all such sequences have same prob $\pi^y(1-\pi)^{n-y}$, in lieu of adding, can multiply this prob. by number , i.e. ${}^nC_y$, of such sequences

The Binomial Distribution

Requirements for y to be Binomial(n , π)

- **Each element in "POPULATION" is binary (0 or 1), but interested only in estimating proportion (π) that are 1**

  (not interested in individuals per se)

- **fixed sample size n**

- **elements selected at random and independently of each other*;**

  all elements have same probability of being sampled.

- **(thus) prob (π) of a 1 is constant for each**

  sampling with replacement
  (if N large relative to n, SRS close to with replacement!)

  [generally we sample without replacement but makes little Δ when N is large rel. to n]

- **elements in population can be related to each other [e.g. spatial distribution of persons]**

  but if use simple random sampling, results in the sample elements are independent

? ? Binomial Variation ? ?

Interested in π the proportion of 16 year old girls in Québec protected against rubella

Choose 20 girls at random from each of
5 randomly selected schools ('n' = 100)

y number, out of total sample of 100, who are protected

Is** y **Binomial (n= 100 ,** π**) ??

Auto-analyzer ("SMAC")
18 chemistries on each person
y number of positive components

Is variation of** y **across persons Binomial (n=18 ,** π **= 0.03) ??
(from text Clinical Biostatistics by Ingelfinger et al.)

Interested in
π_u proportions in usual and
π_e exptl. exercise classes who 'stay the course'

Randomly Allocate
4 classes of
25 students to usual course

4 classes of
25 students to experimental course

Are numbers who stay the course in 'u' and 'e' samples Binomial with** n_u **= 100 and** n_e **= 100 ??

Sex Ratio 4 children in each family
y number of girls in family

Is variation of** y **across families Binomial (n=4 ,** π **= 0.49) ??

The Binomial Distribution

Calculating Binomial probabilities Bin(n , π)

- **Formula (or 1st principles)**

$$\boxed{\text{Prob(y out of n)} = \binom{n}{y}\, \pi^{y}\, (1-\pi)^{n-y} \quad \S}$$

e.g. if n=4 (so 5 probabilities) and π = 0.3

$$\text{Prob(0 / 4)} = \binom{4}{0}\, 0.3^{0}\,(1-0.3)^{4-0} = 0.2401$$

$$\text{Prob(1 / 4)} = \binom{4}{1}\, 0.3^{1}\,(1-0.3)^{4-1} = 0.4116$$

$$\text{Prob(2 / 4)} = \binom{4}{2}\, 0.3^{2}\,(1-0.3)^{4-2} = 0.2646$$

$$\text{Prob(3 / 4)} = \binom{4}{3}\, 0.3^{3}\,(1-0.3)^{4-3} = 0.0756$$

$$\text{Prob(4 / 4)} = \binom{4}{4}\, 0.3^{4}\,(1-0.3)^{4-4} = 0.0081$$

§ e.g. $\binom{8}{3}$, called '8 choose 3', $= \frac{8\times7\times6}{1\times2\times3}$; $\binom{8}{0} = 1$

Calculating Binomial probabilities Binomial(n , π) ... continued

- **Tables for various configurations of n and** π (M&M Table C)

  Table uses X as the r.v. , p as the expected proportion, and *k* as the possible realizations, while

  JH uses *Y*, π and *y* respectively

  e.g. n=4, p T-7... Table goes to p = 0.5 but note mirror images*

  $\pi = 0.3$

y	prob[y | π =0.3]
0	0.2401
1	0.4116
2	0.2646
3	0.0756
4	0.0081

  $\pi = 0.7$

y	prob[y | π =0.7]
0	0.0081
1	0.0756
2	0.2646
3	0.4116
4	0.2401

  *** for $\pi > 0.5$, Binomial_P(y | π) = Binomial_P(n - y, 1 – π)**

- **Other Tables**

  - **CRC Tables**
  - **Fisher and Yates Tables**
  - **Pearson and Hartley (Biometrika Tables..)**
  - **Documenta Geigy**

- **Spreadsheet --- e.g., Excel function**

  BINOMDIST(number_s, trials, probability_s, cumulative)

  BINOMDIST(*y* , *n* , π , cumulative)

  BINOMDIST(1 , 4 , 0.3 , FALSE) = 0.4116

  BINOMDIST(1 , 4 , 0.3 , TRUE) = 0.6517

  Cumulative Probability

  $$\text{Prob[} Y \le 1 \text{]} = \text{Prob[} Y = 0 \text{]} + \text{Prob[} Y = 1 \text{]}$$

  $$= 0.2401 + 0.4116 = 0.6517$$

  (the "s" stands for "success")

- **Statistical software - e.g. SAS** PROBBNML(p, n, y) **function**

- **Calculator ...**

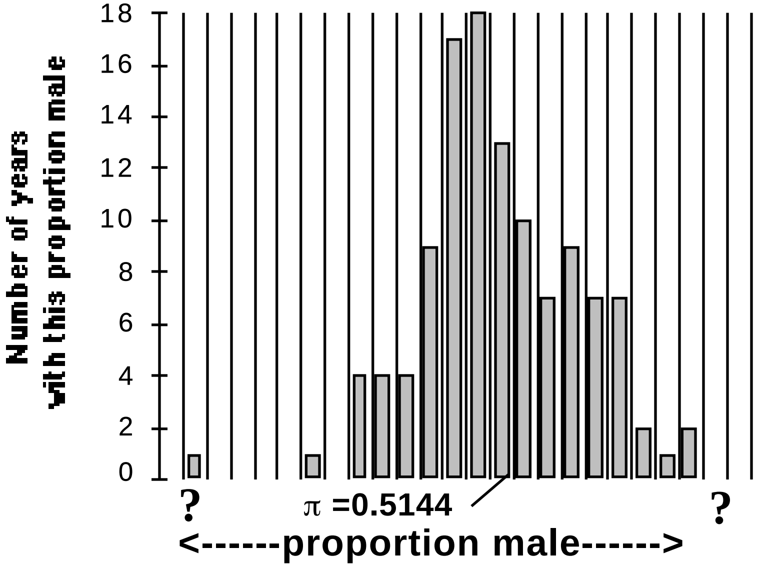
- **Approximations to Binomial**

  - **Normal (Gaussian) Distribution (n large or midrange π)**

  - **Poisson Distribution (n large and low π)**

PROPORTION of MALE BIRTHS Northern Ireland 1864-1979 {116 years}

(~ 24 000 - 39 000 live births per year)



Examination of the sex ratio was triggered by one very unusual year with very low percentage of male births; epidemiologists consider the male fetus more susceptible to environmental damage and searched for possible causes, such as radiation leaks from UK nuclear plants, etc.

Which raises the question: If we did not have historical data on the sex ratio, could we figure out what fluctuations there might be --- just by chance -- from year to year. The n's of births are fairly large so do you expect the % male to go below 45%, 48%, 50% some years?

Take an n of 32000. var[proportion male] = $\sqrt{\pi[1-\pi]}$ / $\sqrt{32000}$;

$\sqrt{\pi[1-\pi]}$ close to 0.5 since π close to 0.5; So, SD[proportion] = 0.5/√n = 0.0028

2SD[proportion] = 0.0056 ; 0.5144 ± 0.0056 = proportion 0.5088 to 0.5200 in 95% of years if no trend over time

[sex ratio has moved slightly downwards in most countries over the centuries]

See data on sex ratio in Canada and provinces 1931-90 (Resources Ch 5)

The Binomial Distribution

Prelude to using Normal (Gaussian) Distribution as approximation to Binomial Distribution

- **Need mean (E) and SD (or √VAR) of a proportion**

- **Have to specify SCALE i.e. whether summary is a**

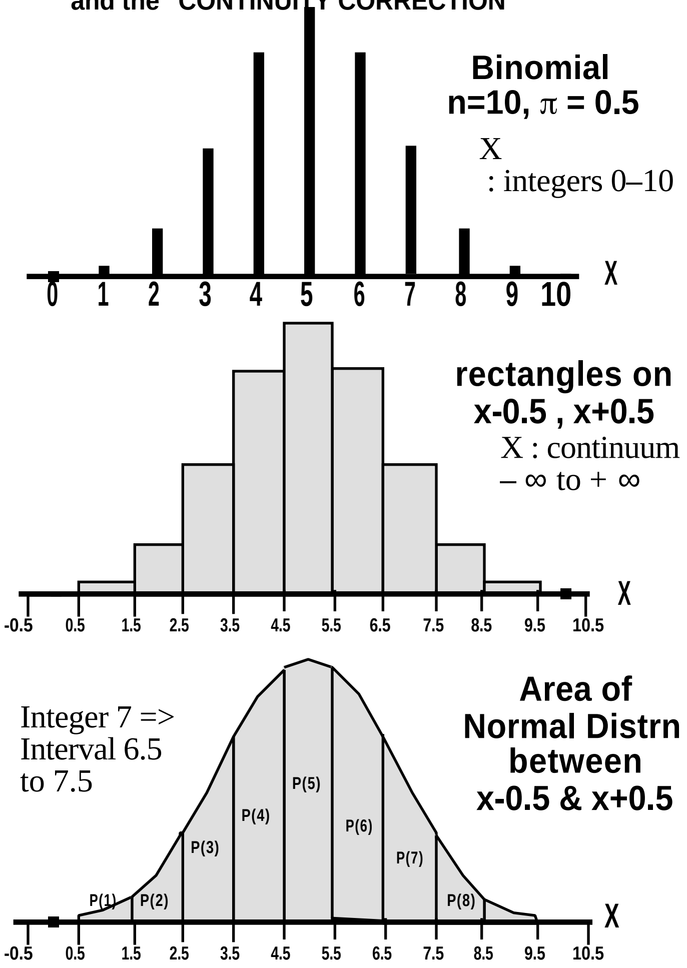
y	count	*e.g.* 2 in 10
p	proportion = y/n	*e.g.* 0.2
%	percentage = 100p%	*e.g.* 20%

- **same core calculation for all 3 [only scale changes]**

summary	E	V=VAR	SD=√VAR
• count (y)	$n\pi$	$n \bullet \pi\,(1-\pi)$	$\sqrt{n\pi\,(1-\pi)} = \sqrt{n} \bullet \sqrt{\pi\,[1-\pi]}$ = √*n* •*SD(indiv. 0's and 1's)*
• prop'n (p) [most common statistic]	π	$\frac{\pi\ [1-\pi]}{n}$	$\sqrt{\frac{\pi\ [1-\pi]}{n}} = \frac{\sqrt{\pi\ [1-\pi]}}{\sqrt{n}}$ $= \frac{SD(indiv.\ 0's\ and\ 1's)}{\sqrt{n}}$
• percent (100p)	100π	100^2 Var(p)	100 SD(p)

Note that all the VAR's have the same "kernel" π (1-π) , which is the variance of a random variable that takes the value 0 with probability 1-π and the value 1 with probability π. Statisticians call this 0/1 or binary variable a Bernoulli Random Variable. Think of π (1-π) as the "unit" variance.

NORMAL (GAUSSIAN) APPROXIMATION TO BINOMIAL and the "CONTINUITY CORRECTION"



THE FIRST RECORDED P–VALUE???

by a physician no less!!

"AN ARGUMENT FOR DIVINE PROVIDENCE, TAKEN FROM THE CONSTANT REGULARITY OBSERV'D IN THE BIRTHS OF BOTH SEXES."

John Arbuthnot, 1667-1735
physician to Queen Anne

Arbuthnot claimed to demonstrate that divine providence, not chance, governed the sex ratio at birth.

To prove this point he represented a birth governed by chance as being like the throw of a two-sided die, and he presented data on the christenings in London for the 82-year period 1629-1710.

Under Arbuthnot's hypothesis of chance, for any one year male births will exceed female births with a probability slightly less than one-half. (It would be less than one-half by just half the very small probability that the two numbers are exactly equal.)

But even when taking it as one-half Arbuthnot found that a unit bet that male births would exceed female births for eighty-two years running to be worth only $(1/2)^{82}$ units in expectation, or

$$\frac{1}{4\ 8360\ 0000\ 0000\ 0000\ 0000\ 0000}$$

a **vanishingly small number**.

"From whence it follows, that it is Art, not Chance, that governs."

STIGLER : HISTORY OF STATISTICS

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")

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

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

The **sampling distribution** is the 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 [and speak loosely -- see later], 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 sampling 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 N**μ**). 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 estimates we might obtain 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 mean [or expected value] of the entire set of Y's is

$$\mu = \mathbf{0} \times 0.5 + \mathbf{1} \times 0.3 + \mathbf{2} \times 0.2 = \mathbf{0.7}$$

The variance of Y is

$$\sigma^2 = (0 - 0.7)^2 \times 0.5 + (1 - 0.7)^2 \times 0.3 + (2 - 0.7)^2 \times 0.2$$
$$= \mathbf{0.49} \times 0.5 + \mathbf{0.09} \times 0.3 + \mathbf{1.69} \times 0.2 = \mathbf{0.61}$$

[sd, $\sigma = \sqrt{0.61} = \mathbf{0.78}$ is slightly larger than μ].

Example of the distribution of a sample mean continued...

Suppose we take a sample of size **n = 2,** and use $\bar{y} = (y_1+y_2)/2$ as our $\hat{\mu}$, **what estimates might we obtain?** [we write estimate as $\hat{\mu} = \bar{y}$].

Distribution of all possible sample means when **n=2**

probability [frequency]	$\hat{\mu}$ [i.e., $\bar{y}$]	error [$\bar{y} - \hat{\mu}$]	% error [% of μ]
25%	$\frac{0}{2}$ = **0.0**	– 0.7	- 100
30%	$\frac{1}{2}$ = **0.5**	– 0.2	– 29
29%	$\frac{2}{2}$ = **1.0**	+ 0.3	+ 43
12%	$\frac{3}{2}$ = **1.5**	+ 0.8	+ 114
4%	$\frac{4}{2}$ = **2.0**	+ 1.3	+ 186

Most of the possible estimates of μ from samples of size 2 will be "off the target " by quite serious amounts. It's not much good saying that *"on average, over all possible samples"* the sample will produce the correct estimate.

Check: Average[estimates] = 0 × 0.25 + 0.5 × 0.30 + 1.0 × 0.29 + 1.5 × 0.12 + 2.0 × 0.04 = 0.7 = μ. Variance[estimates] = $(–0.7)^2$ × 0.25 + ... = 0.305 = σ^2 / 2 .

A sample of size **n = 4** would give less variable estimates. The distribution of the 3^4 = 81 possible sample configurations, and their corresponding estimates of μ, can be enumerated manually as:

Distribution of all possible sample means when **n=4**

probability [frequency]	$\hat{\mu}$ [i.e., $\bar{y}$]	error [$\bar{y} - \hat{\mu}$]	% error [% of μ]
6.25%	$\frac{0}{4}$ = **0.00**	– 0.70	- 100
15.00%	$\frac{1}{4}$ = **0.25**	– 0.45	– 64
23.50%	$\frac{2}{4}$ = **0.50**	– 0.20	– 29
23.40%	$\frac{3}{4}$ = **0.75**	+ 0.05	+ 7
17.61%	$\frac{4}{4}$ = **1.00**	+ 0.30	+ 43
9.36%	$\frac{5}{4}$ = **1.25**	+ 0.55	+ 79
3.76%	$\frac{6}{4}$ = **1.50**	+ 0.80	+ 114
0.96%	$\frac{7}{4}$ = **1.75**	+ 1.05	+ 150
0.16%	$\frac{8}{4}$ = **2.00**	+ 1.30	+ 186

Of course, there is still a good chance that the estimate will be a long way from the correct value of μ = 0.7. But the variance or scatter of the possible estimates is less than it would have been had one used n = 2.

Check:
Average[estimates] = 0 × 0.0625 + 0.25 × 0.15 + ... + 2 × 0.0016 = 0.7 = μ.
Variance[estimates] = $(–0.7)^2$ × 0.0625 + $(–0.45)^2$ × 0.15 ... = 0.1525 = σ^2 / 4 .

Example of the distribution of a sample mean continued...

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 403-404). All the examples show the following:

(1) As expected, the variation of possible sample means about the (in practice, unknown) target μ is less in larger samples. We can use variance or SD to measure this scatter. The SD (scatter) in the possible $\bar{y}$'s from samples of size n is $\sigma / \sqrt{n}$, where σ is the SD of the individual Y's.

This is true no matter what the shape of the distribution of the individual Y's.

(2) If the individual Y's DO are from a Gaussian distribution, then the distribution of possible $\bar{y}$'s will be Gaussian.

BUT ...

even if the individual Y's ARE NOT from a Gaussian distribution ...

the larger the n [and the more symmetric and unimodal the distribution of the individual Y's], the more the distribution of possible $\bar{y}$'s resembles a Gaussian distribution.

The fact that the sampling distribution of $\bar{y}$ [or of sample proportions, or sample slopes or correlations, or other statistics created by aggregation of individual observations ..] is, for a large enough n [and under other conditions*], close to Gaussian in shape no matter what the shape of the distribution of individual Y values, is referred to as the Central Limit Theorem.

* relating to the degree of symmetry and dispersion of the distribution of the individual Y's

We use the notation Y ~ Distribution(μ_y , σ_y)

as shorthand to say that "Y has a certain type of

distribution with mean μ_y and standard deviation σ_y".

In this notation, the Central Limit Theorem says that

if Y ~ ???????(μ_Y, σ_Y) , then

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

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

Individual (Bernoulli) 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

$$\sigma^2 = (0 - \pi)^2 \times (1 - \pi) + (1 - \pi)^2 \times \pi = \pi(1 - \pi).$$

If a sample of size n;

observations $y_1, y_2, ..., y_n$ (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 ~ Bernoulli(μ = π,   σ = $\sqrt{\pi[1 - \pi]}$) , then

$p = \bar{y}$ ~ GAUSSIAN(π , $\frac{\sqrt{\pi[1 - \pi]}}{\sqrt{n}}$) if n 'large' and π not extreme*.

*** i.e. E[# 'positive' = numerator = $\sum y_i$] sufficiently far from the minimum 0, and the maximum, n.**

Returning to the cars per apartment example above:

If n = 100, then the SD of possible $\bar{y}$'s from samples of size n=100 is σ / √100 = 0.78 / 10 = 0.078. Thus, we can approximate the distribution of possible $\bar{y}$'s by a Gaussian distribution with mean μ = 0.7 and standard deviation of 0.078, to get ...

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

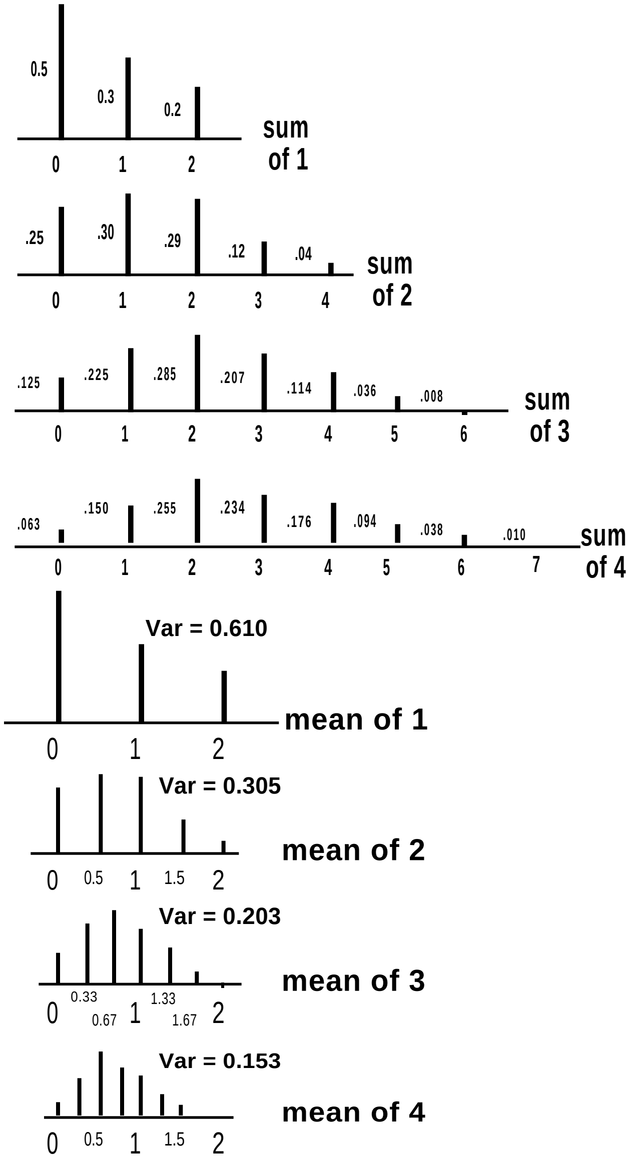
[Notice that in all of this (as long as we sample with replacement, so that the n members are drawn independently of each other), the size of the population (N) didn't enter into the calculations at all. The errors of our estimates (i.e. how different we are from μ on randomly selected samples) vary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

Message from diagram opposite:

Var (Sum) > Var of Individuals by factor of √ n
Var (Mean) < Var of individuals by same factor of √ n

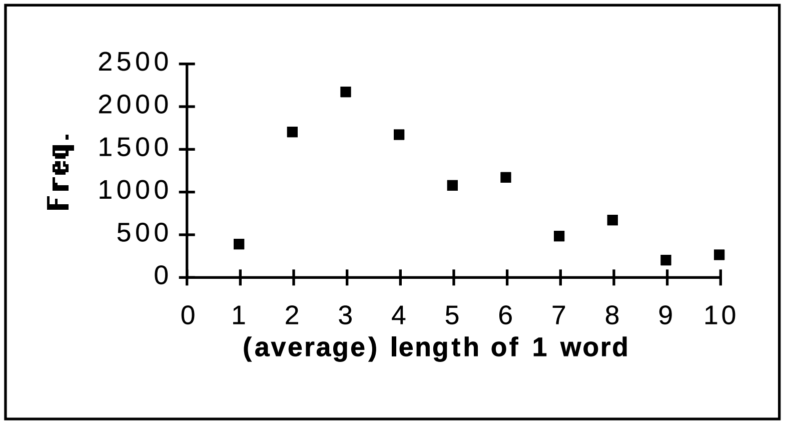
In addition, and also very important: Variation of sample means (or sums) is more Gaussian than variation of individuals

Effect of n on Sampling behaviour of Sums & Means

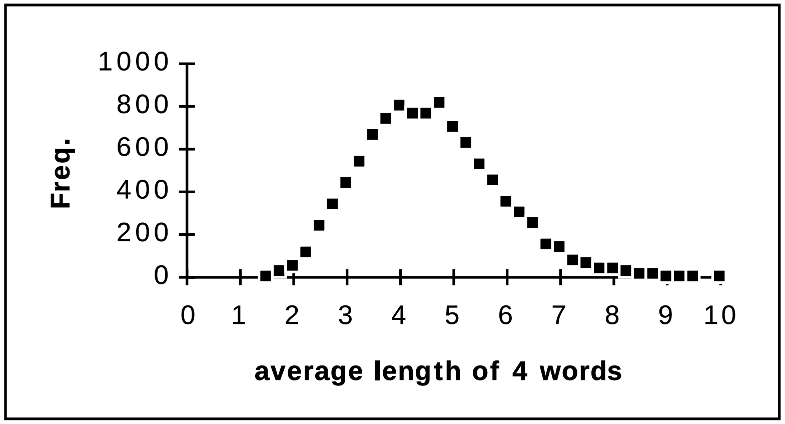


Another Example of Central Limit Theorem at work

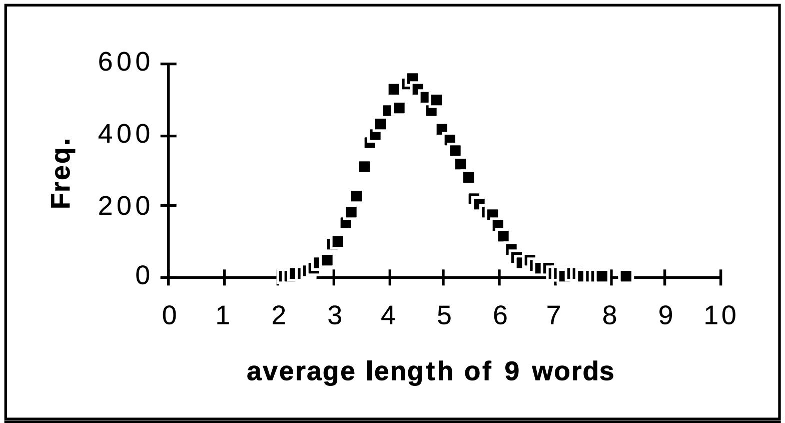
The variability in length of individual words...

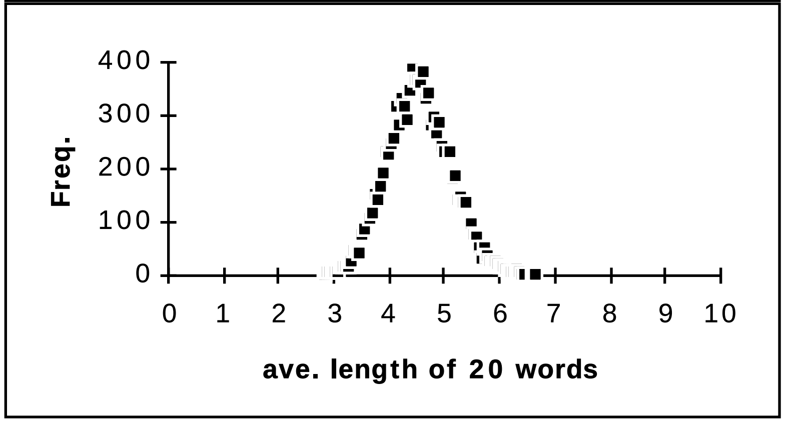


The variability in the average word length in samples of 4, 9, 20 words [Monte Carlo simulation]



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)
```

Standard Error (SE) of the mean ... "SEM"

$$\mathrm{Var}(\bar{y}) = \mathrm{Var}\left[\frac{\sum y_i}{n}\right]$$

$$= \frac{1}{n^2}\mathrm{Var}\left[\sum y_i\right]$$

$$= \frac{1}{n^2}\left[\sum \mathrm{var}[y_i]\right]$$

$$= \frac{1}{n^2}\left[n\ \mathrm{var}[y_i]\right] \quad \textit{\{..if y's uncorrelated\}}$$

$$= \frac{1}{n}\ \mathrm{var}[y]$$

$$= \frac{\mathrm{var}[y]}{n}$$

$$\mathrm{SD}(\bar{y}) = \sqrt{\mathrm{Var}[(\bar{y})]} = \sqrt{\frac{\mathrm{var}[y]}{n}}$$

$$= \frac{\sqrt{\mathrm{Var}[y]}}{\sqrt{n}}$$

$$= \frac{\mathrm{SD}[y]}{\sqrt{n}}$$

$$\begin{aligned}\text{SEM} &= \text{Standard Error(sample mean)}\\ &= \text{SD(sample mean)}\\ &= \frac{\text{SD(individuals)}}{\sqrt{\text{sample size}}}\end{aligned}$$

Standard Error (SE) of commonly used estimates†

	Statistic	Standard Error (SE)
mean	$\bar{y}$	$\frac{\sigma_y}{\sqrt{n}}$
proportion **(binomial)**	p	$\sqrt{\frac{\pi\{1-\pi\}}{n}}$ $\left(= \frac{SD\{0s\&1s\}}{\sqrt{n}}\right)$
proportion **(finite N)**	p	$SE_{binomial}(p) \cdot \sqrt{1 - \frac{n}{N}}$ ¶ (= regular SE • $\sqrt{\text{1-sampling fraction}}$)
Sum / Difference		
	$p_1 \pm p_2$	$\sqrt{[SE\{p_1\}]^2 + [SE\{p_2\}]^2}$ §
	$\bar{y}_1 \pm \bar{y}_2$	$\sqrt{[SE\{\bar{y}_1\}]^2 + [SE\{\bar{y}_2\}]^2}$ §

§ Remember...

*** SD's and SE's (which are SD's too) DO NOT ADD THEIR SQUARES, i.e. VARIANCES, DO!**

• SE(SUM <u>or</u> DIFFERENCE)

= $\sqrt{SE^2 \ \underline{\text{PLUS}}\ SE^2}$ if estimates uncorrelated

¶ $\frac{n}{N}$ often close to zero, so downward correction negligible.

† Ref : A & B Ch 3 (they also deal with SE's of ratios and other functions and transformations of estimates)

Standard Error (SE) of combination or weighted average of estimates

$$\text{SE}\{\sum \text{estimates}\} = \sqrt{\sum\{[\text{SE of each estimate}]^2\}}$$

$$\text{SE}\{\text{constant } \mathbf{x} \text{ estimate}\} = \text{constant x SE}\{\text{estimate}\}$$

$$\text{SE}\{\text{constant } \mathbf{+} \text{ estimate}\} = \text{SE}\{\text{estimate}\}$$

$$\text{SE}\{\sum w_i \mathbf{x} \text{ estimate}_i\} = \sqrt{\sum\{w_i^2 \mathbf{x} [\text{SE estimate}_i]^2\}}$$

This last one is important for combining estimates from stratified samples, and for meta-analyses:

In an estimate for the overall population, derived from a **stratified sample**, the weights are chosen so that the overall estimate is unbiased i.e. the w's are the relative sizes of the segments (strata) of the overall population (see "combining estimates ... entire population" below). The **parameter values** will likely **differ between strata**. (this is why stratified sampling helps). The estimate for the entire population parameter is formed as a weighted average of the age-specific parameter estimates, with weights reflecting the proportions of population in the various strata.

If instead, one had several estimates of the **same** parameter value (a big assumption in the 'usual' approach to **meta-analyses**), but each estimate had a different uncertainty (precision), one should take a weighted average of them, but with the weights inversely proportional to the amount of uncertainty in each. from the formula above one can verify by algebra or trial and error that the smallest variance for the weighted average is obtained by using weights proportional to the inverse of the variance (squared standard error) of each estimate.
If there is **variation in the parameter value**, this SE is too small. The 'random effects' approach to meta-analyses weights each estimate in inverse relation to an amalgam of (i) each SE and (ii) the 'greater-than-random' variation between estimates [it allows for the possibility that the parameter estimates from each study would not be the same, even if each study used huge n's). The SE of this weighted average is larger than that using the simpler (called fixed effects) model; as a result, CI's are also wider.

COMBINING ESTIMATES FROM SUBPOPULATIONS TO FORM AN ESTIMATE FOR THE ENTIRE POPULATION

If several (say k) sub-populations or "strata" of sizes N_1, N_2, ... N_k, form one entire population of size $\sum N_k = N$. Interested in quantitative or qualitative characteristic of entire population. Denote this numerical or binary characteristic in each individual by Y, and an aggregate or summary (across all individuals in population) by θ, which could stand for an average (μ), a total ($T_{amount} = N\mu$), a proportion (π), a percentage (% = 100π) or a total count ($T_c = N\pi$). Examples:

If Y is a measured variable (i.e. "numerical")

μ:	the annual (per capita) consumption of cigarettes
T_{amount}:	the total undeclared yearly income ($T_{amount} = N\mu$ and conversely that $\mu = T_{amount} \div N$)

If Y is a binary variable (i.e. "yes/no")

π:	the proportion of persons who exercise regularly
100π %:	the percentage of children who have been fully vaccinated
Nπ:	the total number of persons who need R_x for hypertension ($T_c = N\pi$; $\pi = T_c \div N$)

The sub-populations might be age groups, the 2 sexes, occupations, provinces, etc. There is a corresponding θ for each of the K sub-populations, but one needs subscripts to distinguish one stratum from another. Rather than study every individual, one might instead measure Y in a *sample* from each stratum.

- **Estimate overall μ, π, or π%, combine estimates :**

Sub Popln	Size	Relative Size Wi = Ni ÷ N	Sample Size	Estimate of θi	SE of estimate
1	N_1	W_1	n_1	e_1	$SE(e_1)$
...	...	...	...	...	......
k	N_k	W_k	n_k	e_k	$SE(e_k)$
Total	$\sum N = N$	$\sum W=1$	$\sum n=n$	$\sum W_i e_i$	$\sum W_i^2[SE(e_i)]^2$

Note1- To estimate T_{amount} or T_c , use weights $W_i = N_i$;
Note2: If any sampling fraction $f_i = n_i \div N_i$ is sizable, the SE of the e_i should be scaled down i.e. it should be multiplied by $\sqrt{(1-f_i)}$
Note3: If variability in Y within a stratum is smaller than across strata, the smaller SE obtained from the SE's of the individual stratum specific estimates more accurately reflects the uncertainty in the overall estimate. Largest gain over SRS is when large inter-stratum and low intra-stratum variability