Probability

Meaning
Long Run Proportion
Estimate of (Un)certainty
Amount prepared to bet

Use
Describe likely behaviour of data
Communicate (un)certainty
Measure how far data are from
some hypothesized model

How Arrived At

Subjectively

Intuition, Informal calculation, consensus

Empirically

Experience (actuarial, ...)

Pure Thought

Elementary Statistical Principles

If necessary, breaking Complex outcomes into simpler ones

Advanced Statistical Theory

calculus e.g. Gauss' Law of Errors

References
• WMS5, Chapter 2 • Moore & McCabe Chapter 4 •Colton, Ch 3
• Freedman et al. Chapters 13,14,15 •Armitage and Berry, Ch 2
• Kong A, Barnett O, Mosteller F, and Youtz C. "How Medical Professionals Evaluate Expressions of Probability" NEJM 315: 740-744, 1986 ... ***on reserve***

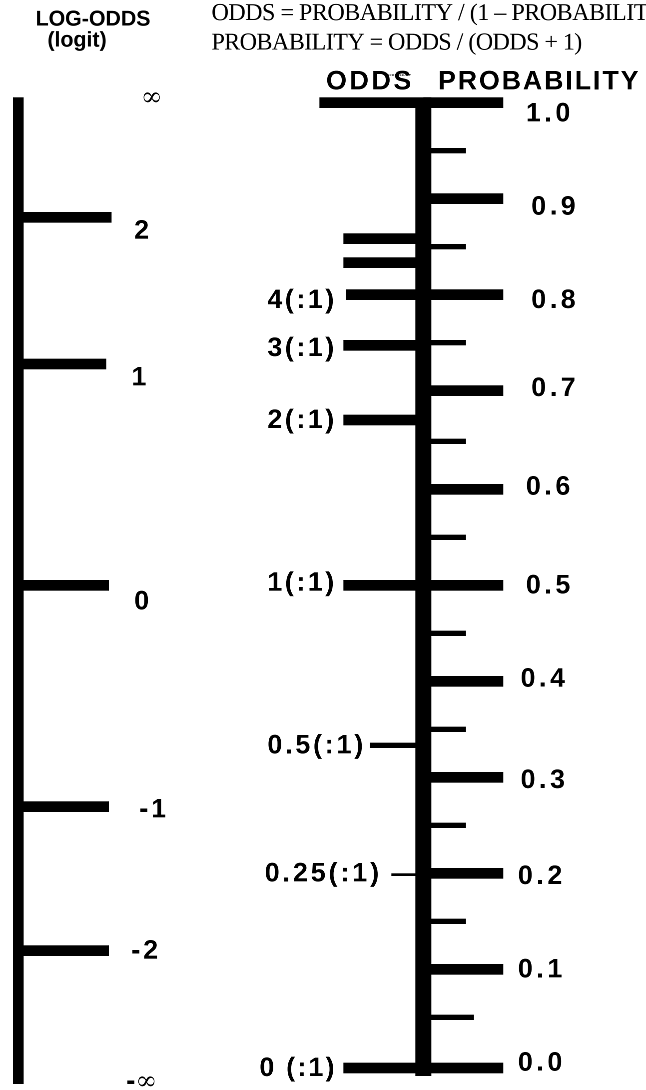
• Death and Taxes • Rain tomorrow • Cancer in your lifetime • Win lottery in single try • Win lottery twice • Get back 11/20 pilot questionnaires • Treat 14 patients get 0 successes • Duplicate Birthdays • Canada will use $US before the year 2010

- **OJ murdered his wife**
- **DNA matched**
- **OJ murdered wife | DNA matched**

" | " is shorthand for "given that.."

Probability Scales

ODDS = PROBABILITY / (1 – PROBABILITY)
PROBABILITY = ODDS / (ODDS + 1)



- **50 year old has colon ca**
- **50 year old with +ve haemoccult test has colon ca**
- **child is Group A Strep B positive**
- **8 yr old with fever & v. inflamed nodes is Gp A Strep B positive**
- **There is life on Mars**

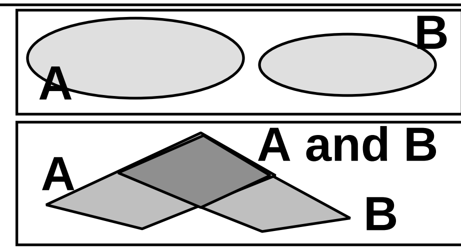
How to calculate probabilities

Wall Street Journal

"I figure there's a 40% chance of showers, and a 10% chance we know what we're talking about"

Probability Calculations

Basic Rules



Probabilities add to 1

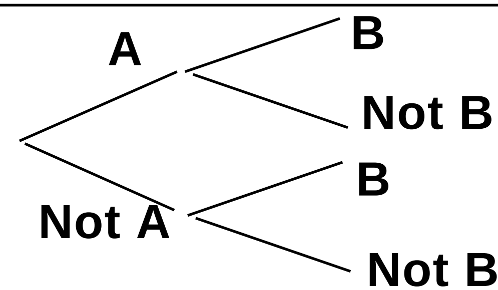
Prob(event) =
1 - Prob(complement)

ADDITION FOR "EITHER A OR B"

"PARALLEL"

If mutually exclusive
P(A or B) = P(A) + P(B)

If overlapping
P(A or B) = P(A) + P(B) - P(A and B)



MULTIPLICATION FOR "A AND B" OR "A THEN B"

"SERIAL"

If independent
P(A and B) = P(A) • P(B)

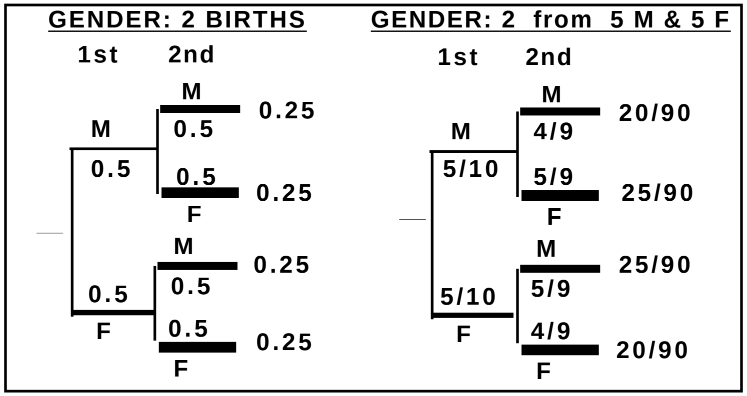
If dependent
P(A and B) = P(A) • P(B | A)

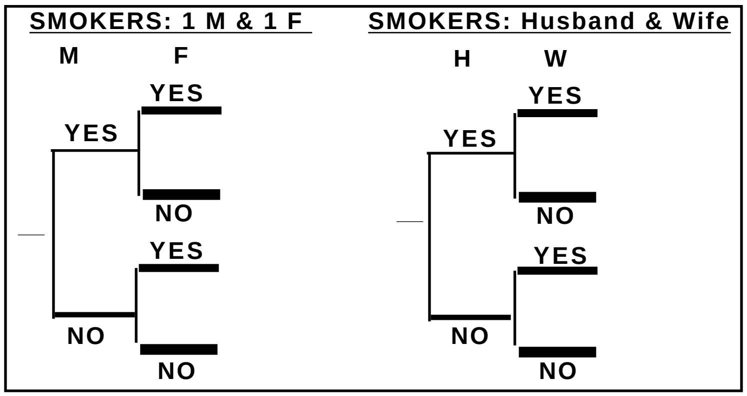
Conditional Probability P(B | A) = Probability of B "**given** A" or "conditional on A"

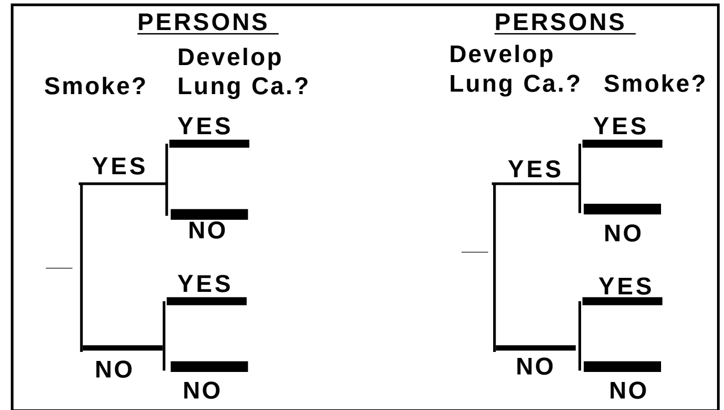
More Complex:

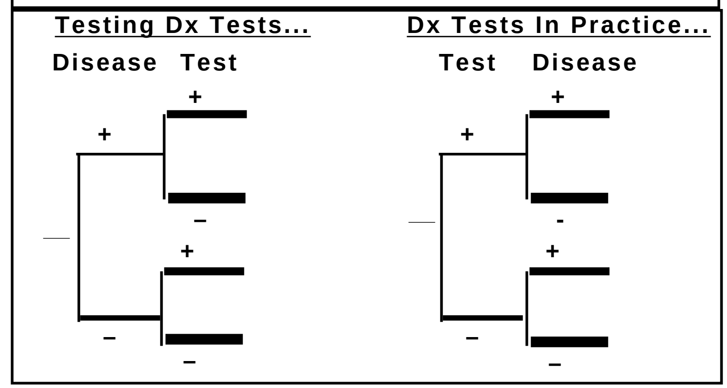
- Break up into elements
- Look for already worked-out calculations
- Beware of intuition, especially with "after the fact" calculations for non-standard situations

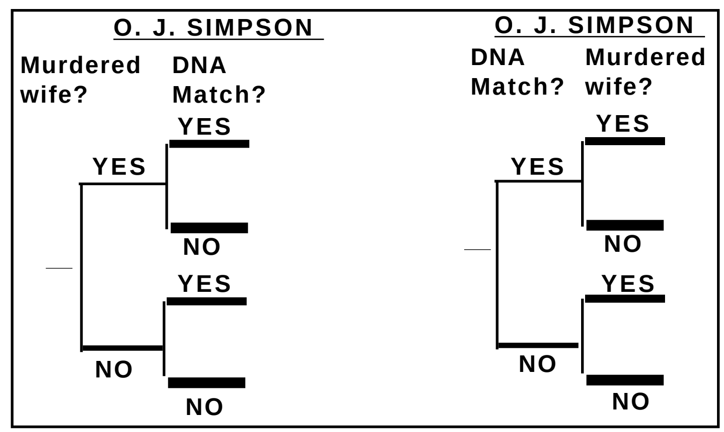
Examples of Conditional Probabilities...











Reverse Probabilities: Probability[data | Hypothesis] ≠ Probability[Hypothesis | data]

U.S. National Academy of Sciences under fire over plans for new study of DNA statistics:

Confusion leads to retrial in UK.

[NATURE p 101-102 Jan 13, 1994]

... He also argued that one of the prosecution's expert witnesses, as well as the judge, had **confused two different sorts of probability**.

One is the probability that DNA from an individual selected at random from the population would match that of the semen taken from the rape victim, a calculation generally based solely on the frequency of different alleles in the population.

The other is the separate probability that a match between a suspect's DNA and that taken from the scene of a crime could have arisen simply by chance [1] ***-- in other words that the suspect is innocent despite the apparent match. This probability depends on the other factors that led to the suspect being identified as such in the first place.***

[1] **Underlining is mine (JH). The wording of the singly-underlined phrase is imprecise; the doubly-underlined wording is much better .. if you read 'despite' as "given that" or "conditional on the fact of" JH**

During the trial, a forensic scientist gave the first probability in reply to a question about the second. Mansfield convinced the appeals court that the error was repeated by the judge in his summing up, and that this slip -- widely recognized as a danger in any trial requiring the explanation of statistical arguments to a lay jury -- justified a retrial.

In their judgement, the three appeal judges, headed by the Lord Chief Justice, Lord Farquharson, explicitly stated that their decision "should not be taken to indicate that DNA profiling is an unsafe source of evidence".

Nevertheless, with DNA techniques being increasingly used in court cases, some forensic scientists are worried that flaws in the presentation of their statistical significance could, as in the Deen case, undermine what might otherwise be a convincing demonstration of a suspect's guilt.

Some now argue, for example, that quantified statistical probabilities should be replaced, wherever possible, by a more descriptive presentation of the conclusions of their analysis. "The whole issue of statistics and DNA profiling has got rather out of hand," says one.

Others, however, say that the Deen case has been important in revealing the dangers inherent in the '**prosecutor's fallacy**'. They argue that this suggests the need for more sophisticated calculation and careful presentation of statistical probabilities.

"The way that the prosecution's case has been presented in trials involving DNA-based identification has often been very unsatisfactory," says David Balding, lecturer in probability and statistics at Queen Mary and Westfield College in London. "Warnings about the prosecutor's fallacy should be made much more explicit. After this decision, people are going to have to be more careful."

Reverse Probabilities: Probability[data | Hypothesis] ≠ Probability[Hypothesis | data]

"The prosecutor's fallacy"

Who's the DNA fingerprinting pointing at?

New Scientist, 29 Jan. 1994, 51-52. David Pringle

Pringle describes the successful appeal of a rape case where the primary evidence was DNA fingerprinting. In this case the statistician Peter Donnelly opened a new area of debate. He remarked that

forensic evidence answers the question

> ***"What is the probability that the defendant's DNA profile matches that of the crime sample, assuming that the defendant is innocent?"***

while the jury must try to answer the question

> ***"What is the probability that the defendant is innocent, assuming that the DNA profiles of the defendant and the crime sample match?"***

(JH) Donnelly's words make the contrast of the two types of probability much "crisper". The fuzziness of the wording on the previous page is sadly typical of the way statistical concepts often become muddied as they are passed on.

Apparently, Donnelly suggested to the Lord Chief Justice and his fellow judges that they imagine themselves playing a game of poker with the Archbishop of Canterbury. If the Archbishop were to deal himself a royal flush on the first hand, one might suspect him of cheating. Assuming that he is an honest card player (and shuffled eleven times) the chance of this happening is about 1 in 70,000.

But if the judges were asked whether the Archbishop were honest, given that he had just dealt a royal flush, they would be likely to place the chance a bit higher than 1 in 70,000 *.

The error in mixing up these two probabilities is called the "the prosecutor's fallacy", and it is suggested that newspapers regularly make this error.

Apparently, Donnelly's testimony convinced the three judges that the case before them involved an example of this and they ordered a retrial

from Vol 3.02 of Chance News

*** (JH) This is a very nice example of the advantages of Bayesian over Frequentist inference .. it lets one take one's prior knowledge (the fact that he is the Archbishop) into account.**

Random Variables ; Probability Distributions ; Expectation and Variance of a Random Variable

Random Variables & Probability Distributions

What they are:

E.g. Random Variable	Possible Outcomes (abbreviated)	Corresponding Probabilities
the blood group of n = 1 randomly selected person	A	P(A)
	B	P(B)
	AB	P(AB)
	O	P(O)
		1.00
How many of n = 20 randomly selected persons will return questionnaire in pilot study	0	P(0)
	1	P(1)
	2	P(2)
	...	...
	20	P(20)
		1.00
Mean cholesterol level in n=30 randomly selected persons	<100	P(<100)
	100-101	P(100-101)
	...	...
	249-250	P(249-250)
	>250	P(>250)
		1.00
the value of the test-statistic if 2 populations sampled from had the same mean	< -2.0	.028
	-2 to -1	.136
	-1 to 0	.341
	0 to 1	.341
	1 to 2	.136
	> 2.0	.028
		1.000

- we use probabilities or fractions as relative frequencies (like a histogram with an infinite number of entries)
- typically, the random quantity is obtained from an aggregate of elements e.g. sum, mean , proportion, regression slope

Other References •Colton, Ch 3

Expectation (Mean) & Variance of Random Variable

- **If Y takes on the DISCRETE values**

y_0	with probability	p_0	
y_1	with probability	p_1	
...	...		...
y_k	with probability	p_k	

then the expected value of Y (written "E(Y)") is

$$y_0 \cdot p_0 + y_1 \cdot p_1 + y_2 \cdot p_2 + \ldots + y_k \cdot p_k \quad \text{or} \quad \sum_{i=1}^{i=K} y_i \cdot p_i$$

Compare the formula for E(Y) with that for xbar:-

- E(Y) is a mean that uses expected (i.e. unobservable or theoretical or long run) relative frequencies (p's)
- ybar uses observed relative frequencies (f / n)'s.

- **If Y takes on the CONTINUOUS values** $y - \frac{\Delta y}{2}$ to $y + \frac{\Delta y}{2}$ with probability $p = f(y) \cdot \Delta y$,

then $E(Y) = \int_{y_{min}}^{y_{max}} y \cdot f(y) \cdot \Delta y$

Variance of a Random Variable

$$Var(Y) = \sigma^2 = E[\,(Y - \mu)^2\,] = \sum_{i=1}^{i=K} [y_i - \mu]^2 \cdot p_i$$

i.e. the Expected Squared Deviation from μ

Just as there was a computational shortcut for calculating σ^2, we can write

$$Var(Y) = \sigma^2 = E[\,Y^2\,] - \mu^2$$

"ave(square) - squared ave"

Other References •Colton, Ch 3

Random Variables ; Probability Distributions ; Expectation and Variance of a Random Variable

Relevance of Expectation of a Random Variable

1 **ACTS AS A MEAN FOR A VARIABLE THAT HAS A (CONCEPTUAL) REPETITION OR AN INFINITE N**

2 **THE EXPECTED VALUE OF A RANDOM VARIABLE X WILL USUALLY BE IN TERMS OF POPULATION PARAMETERS**

A STATISTIC WITH EXPECTED VALUE Θ IS AN "UNBIASED ESTIMATOR" OF Θ.

e.g.1 Y = Proportion of YES' in sample

$E(Y) = \pi$ = PROPORTION of YES' in POP^{ln}

THEN $\hat{\pi} = Y$

(Y is an unbiased estimator of π)

e.g.2 Likewise, if we use divisor of n - 1,

$E(s^2) = \sigma^2$, so...

$\hat{\sigma}^2 = s^2$ is an unbiased estimator of σ^2

{ $\hat{\sigma}^2$ stands for "estimate of " σ^2 }

If we use divisor of n

$E(s^2$ with divisor of n$) = \frac{n-1}{n}\ \sigma^2$ (too small on average)

e.g. Life Expectancy at birth (Québec 1990 mortality data)

"Y" = Length of life = age at death. Assume for sake of illustration that deaths in a decade are all at midpoint of interval (calculations done one year rather than one decade at a time would be more exact)

decade	mid-point age	**Males:** proportion (p) dying in this decade	age × p	**Females:** proportion (p) dying in this decade	age × p
0-10	5	0.010	0.050	0.008	0.040
10-20	15	0.006	0.089	0.002	0.030
20-30	25	0.012	0.295	0.004	0.099
30-40	35	0.016	0.544	0.007	0.242
40-50	45	0.030	1.335	0.017	0.749
50-60	55	0.074	4.079	0.040	2.223
60-70	65	0.180	11.697	0.096	6.233
70-80	75	0.301	22.610	0.214	16.049
80-90	85	0.279	23.680	0.358	30.442
90-100	95	0.093	8.822	0.254	24.136
All (Σ)		**1.000**	**73.2**	**1.000**	**80.2**

Expectation of Life at Birth (**average** longevity)

Males: 73.2 years **Females**: 80.2 years

Variance[longevity] = average[square] – squared average

Males: Ave[square] = $5^2 \bullet 0.010 + 15^2 \bullet 0.006 + \ldots 95^2 \bullet 0.093$
= 5619.38 , so

Var[longevity] = $5619.38 - 73.2^2 = 261.14$ or

SD[longevity] = $\sqrt{261.14}$ = 16.2

[Think of it as the SD when the 'n' is 1 000 or 1 000 000]

Note: Since distribution of longevity not Gaussian, SDdeviation not helpful in describing limits of individual variation (%-iles would be better)

Random Variables ; Probability Distributions ; Expectation and Variance of a Random Variable

If waiting for one of 3 unevenly spaced elevators
(all equally likely to arrive next),
where (?) do you stand? what criterion does it imply?

0	1				5	<--elevators
						average distance
?						
0	1				5	**2.00**
?						
0.5	0.5				4.5	**1.83**
	?					
1	**0**				**4**	**1.67**
	?					
1.5	0.5				3.5	**1.83**
		?				
2	1				3	**2.00**
		?				
2.5	1.5				2.5	**2.17**
			?			
3	2				2	**2.33**
			?			
3.5	2.5				1.5	**2.50**
				?		
4	3				1	**2.67**
				?		
4.5	3.5				0.5	**2.83**
					?	
5	4				0	**3.00**

? = mean position minimizes average squared deviation.
? = median minimizes the average absolute deviation*.

0	1				5	<--elevators
						average squared distance
?						
0	1				25.00	**8.67**
?						
0.25	0.25				20.25	**6.92**
	?					
1	0				16.00	**5.67**
	?					
2.25	0.25				12.25	**4.92**
		?				
4	**1**				**9.00**	**4.67**
		?				
6.25	2.25				6.25	**4.92**
			?			
9	4				4.00	**5.67**
			?			
12.25	6.25				2.25	**6.92**
				?		
16	9				1.00	**8.67**
				?		
20.25	12.25				0.25	**10.92**
					?	
25	16				0.00	**13.67**

* see elsewhere on 607 and 697 course pages

Random Variables ; Probability Distributions ; Expectation and Variance of a Random Variable

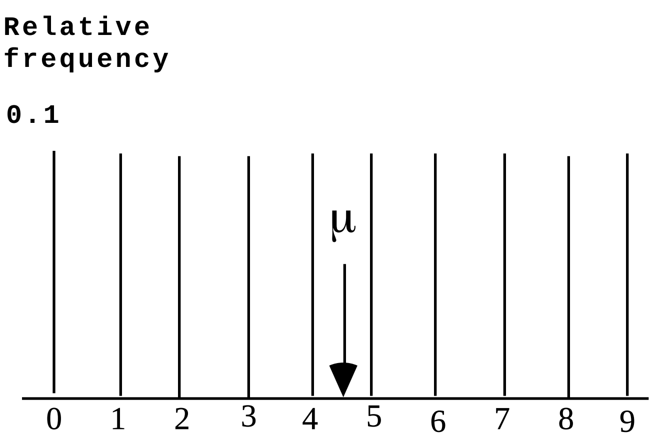
e.g. Expectation & Variance of Random Digits 0 - 9

y	Prob	y × prob	y^2	y^2 × prob
0	0.1	0.0	0	0.0
1	0.1	0.1	1	0.1
2	0.1	0.2	4	0.4
.	.	.	.	.
.	.	.	.	.
7	0.1	0.7	49	4.9
8	0.1	0.8	64	6.4
9	0.1	0.9	81	8.1
Σ	1.0	4.5		28.5

$\text{Var}[Y] = E[Y^2] - \{E[Y]\}^2 \approx 28.5 - 4.5^2 = 8.25$

[Variance = ave. square minus squared ave.]

$\text{SD}[Y] = \sqrt{\text{Var}[Y]} = 2.9$



Expectation, Variance & SD of a Binary [0 / 1] "Bernoulli" RV

Y = 0 with probability p(0) = 1 - π

Y = 1 with probability p(1) = π

In other words...

> A proportion π of the individual elements in the population are "**positive**" (Y = 1); the remaining fraction or proportion 1- π are "**negative**" (Y=0)

$$E(Y) = 0 \times p(0) + 1 \times p(1)$$
$$= 0 \times (1 - \pi) + 1 \times \pi$$
$$= \pi$$

$$VAR(Y) = E(Y^2) - \{E(Y)\}^2$$
$$= 0^2 \times p(0) + 1^2 \times p(1) - \pi^2$$
$$= 0 + 1 \times \pi - \pi^2$$
$$= \pi - \pi^2$$

ie $VAR(Y_{Bernoulli}) = \pi(1 - \pi)$ = prop. neg. × prop pos.

$SD(Y_{Bernoulli}) = \sqrt{VAR(?)} = \sqrt{\pi(1 - \pi)}$

*This "Bernoulli" Random Variable is a **key** one in **Epidemiology** -- it is the 'kernel' or 'atom' in the molecules called Binomial Random Variables. The unit variance π[1–π] and its square root show up whenever we deal with 0/1 data.*

Expectation and Variance of a Linear Combination of Random Variables (R.V's)

Expectation, Variance, and SD of a SUM of 2 (or more) UNCORRELATED Random Variables

R.V.	Mean ("Expectation")	Variance ("Var")
Y_1	μ_1	σ_1^2
Y_2	μ_2	σ_2^2
$Y_1 + ?_2$	$\mu_1 + \mu_2$	$\sigma_1^2 + \sigma_2^2$

Remember... SD's DON'T ADD; VARIANCES DO!!

In general... (using E as shorthand for Expected Value)

$E[\Sigma Y_i] = \Sigma E[Y_i]$ .. **whether correlated or not**

$Var[\Sigma Y_i] = \Sigma Var[Y_i]$ .. **if uncorrelated**

$Var[\Sigma Y_i] = \Sigma Var[Y_i] + \Sigma Covar[Y_i, Y_j]$ .. **otherwise**

Even more generally... if use weights w_i

$E[\Sigma w_i Y_i] = \Sigma w_i E[Y_i]$

$Var[\Sigma w_i Y_i] = \Sigma w_i^2 Var[Y_i] + \Sigma w_i w_j Covar[Y_i, Y_j]$

Thus if Y_1 and Y_2 are uncorrelated

$$Var[Y_1 \pm Y_2] = Var[Y_1] + Var[Y_2]$$

NOTE: Var[**Difference**]= **SUM** of Variances

Example of Variance and SD of a SUM

Planeloads of n=100 persons, randomly chosen from a population with

μ = 70 Kg σ = 8 Kg so σ^2 = 64 Kg^2

Y_i: weight of i-th passenger in sample

E[Combined weight of 100 passengers]

$= \Sigma E[Y_i] = \Sigma 70 = 100 ? 70 = 7000$Kg

Var[Combined weight of 100 passengers]

$= \Sigma Var[Y_i] = \Sigma 64 = 6400$ Kg^2

SD[Combined weight of 100 passengers]

$= \sqrt{6400} = 80$ Kg

$= \sqrt{n} \cdot$ SD[weight of individuals]

Example of Variance and SD of a Difference

Difference, $H_m - H_f$, in heights, H_m and H_f, of a randomly selected male and a randomly selected female from populations with

μ_m = 175 cm σ_m = 6.1 cm μ_f = 162 cm σ_f = 5.8 cm

[parameter values taken from 1972 Busselton (Australia) Study -- see course 678 web page]

Sampling variability of two common statistics - sample mean (ybar) and sample proportion $\hat{p}$

- "ybar" (mean of n sample values)
- "p-hat" (sample proportion -- mean of n 0's and 1's)

calculated from the "Y" values in a simple random sample of size n from a 'universe' where ...

mean(Y) = E(Y) = μ ,

Variance(Y) = σ^2 (so SD(?) = σ)

We can express the variability using the standard deviation (square root of the variance) of the statistic, once we represent the statistic as a mean of n independent identically distributed random variables

$$Y_1,\ Y_2,\ \ldots\ ,\ Y_n,$$

each with mean μ , variance σ^2 , (i.e., SD σ)

(Some statisticians think of the n observed values in the sample as n 'realizations' of the single random variable Y)

$$\text{ybar} = \frac{Y_1 + Y_2 + \quad \ldots \quad + Y_n}{n}$$

$$= \frac{1}{n} \times (Y_1 + Y_2 + \quad \ldots \quad + Y_n)$$

So... $\text{Var(ybar)} = \frac{1}{n^2} \times \text{Var [Sum of n Y's]}$

$$= \frac{1}{n^2} \Sigma \text{ var} [\ Y_i\ ] = \frac{1}{n^2}\ n\ \sigma^2$$

$$= \frac{\sigma^2}{n}$$

$\therefore$ the variance of the means of (all possible simple random) samples of n values is n times smaller than the variance of all of the individual values in the 'universe' of Y's.

$\therefore$ the SD of the means of (all possible simple random) samples of n values is $\sqrt{n}$ times smaller then the SD of individual values.

Same rule for p ... a proportion (numerator/n) is the mean of n binary (0/1) RV's $Y_1, Y_2, \quad \ldots , Y_n$, with

$$\sigma^2 = \text{var} [\ Y_1\ ] = \pi\ (\ 1 - \pi\ )$$

"Cancellation of extremes" and reduction of uncertainty: how insurance companies stay solvent

Possible Earnings from <u>single</u> insurance policy and from <u>pool of n</u> insurance policies:

Earnings from a single policy (n=1)

Y=Earnings	Prob	Y × Prob	Y^2 × Prob
–$19,900	0.00183	–$36.417	724,698.3
–$19,800	0.00186	–$36.828	729,194.4
–$19,700	0.00189	–$37.233	733,490.1
–$19,600	0.00191	–$37.436	733,745.6
–$19,500	0.00193	–$37.635	733,882.5
$500	0.99058	$495.290	247,645.0
	1.00000	$309.741	3,902,655.9

Expected (i.e. average) Earnings per policy
= Σ Earnings × Probability = $309.74

Variance(Earnings) = ave(Earnings²) – (ave Earnings)²

= 3,902,655.9 – 309.74²

= 3,806,717 ($²)

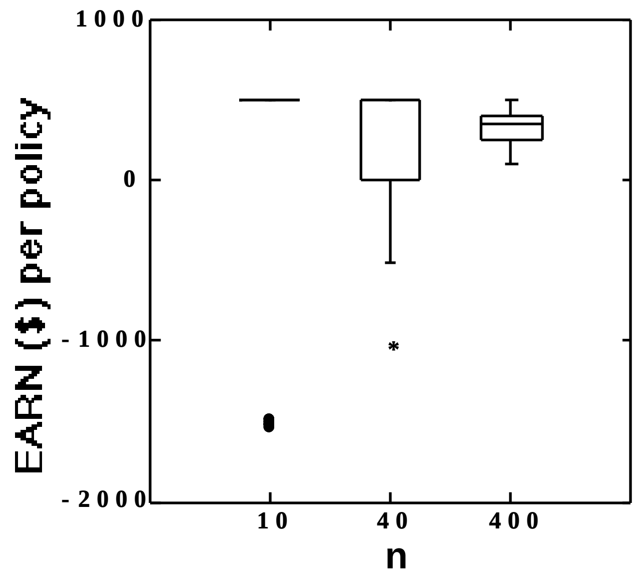
SD(Earnings) = $\sqrt{\text{Var(Earnings)}}$ = $1,951

Earnings per policy from a pool of n policies
Statistics for earnings from pooled policies based on several simulations per pool size

n:	1	10	40	400
MINIMUM	-29,900	-1,540	-1,022	96
MAXIMUM	500	500	500	500
MEAN	309	268	318	320
STD DEV	1,951	645	309	92
$\frac{\$1951}{\sqrt{n}}$	$1,951	$617	$308	$98

<u>Note</u>: This example is from Q5.22 page 358 of 1st Edition of Moore and McCabe. Q4.48 in 2nd edition and Q4.52 p 341 in 3rd edition have $100,000 policy and $250 premium per year, but principle is same.

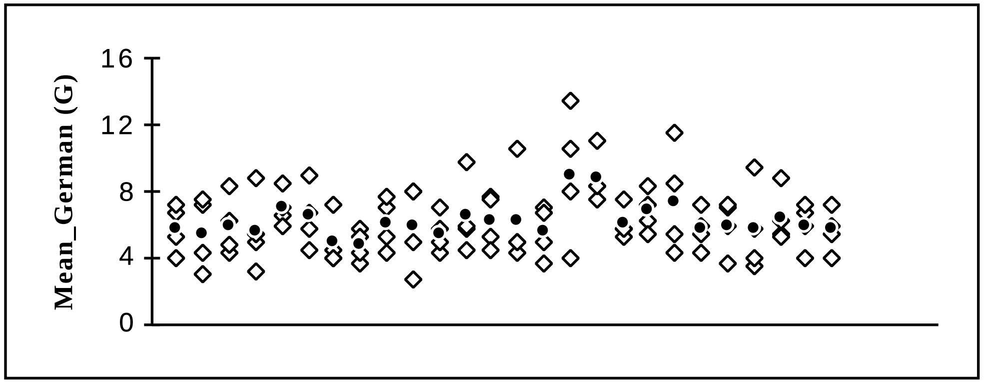
Earnings from pool of n policies



Notice that the SD[mean of n policies] in the simulations is quite close to that predicted theoretically, namely

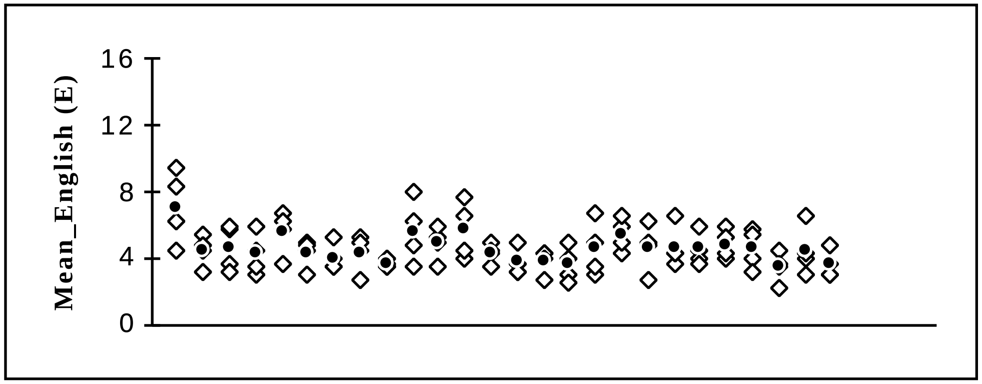
<---- $1,951/√n

Variation in the mean word length in samples of sizes n=4(◊) and n=16(•) , and in the differences of two means (G - E)
[each ◊ and • represents a sample from a student in course 607 in a previous year]



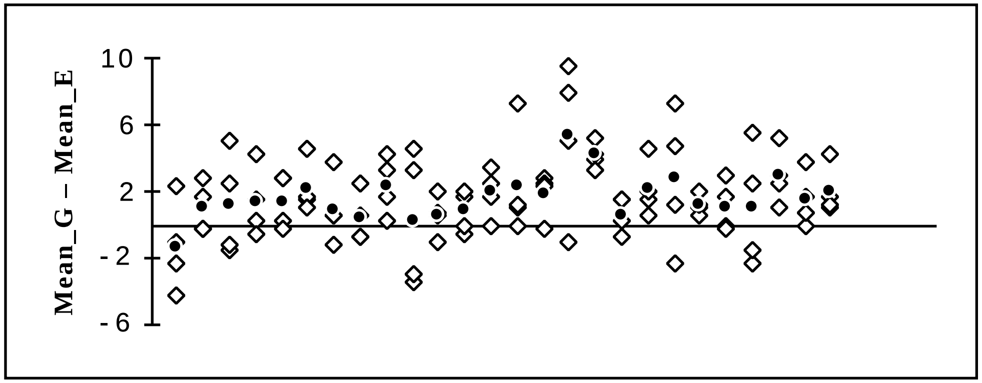
---- GERMAN ----

SD of means based on	SD of means based on
(◊) n=4	(•) n=16
1.97	0.97



---- ENGLISH ----

SD of means based on	SD of means based on
(◊) n=4	(•) n=16
1.33	0.81



GERMAN – ENGLISH

SD of Δ means based on	SD of Δ means based on
(◊) n=4	(•) n=16
2.40	1.30