THE CARTOON GUIDE TO

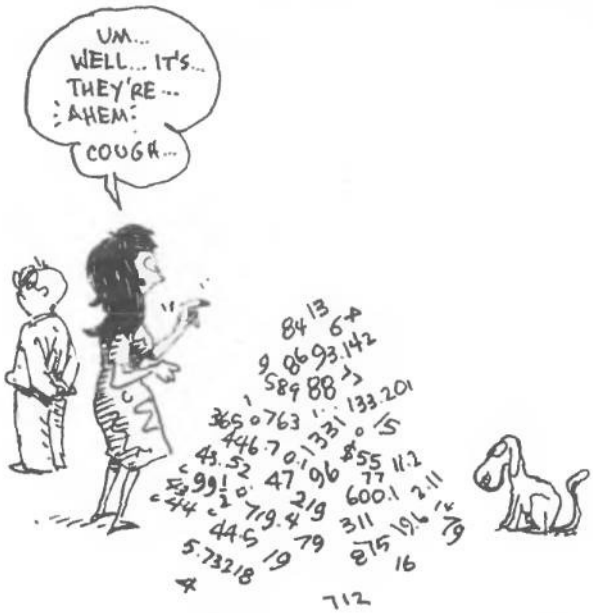
STATISTICS

LARRY GONICK
Author of *The Cartoon History of the Universe*
& WOOLLCOTT SMITH

Also by Larry Gonick

The Cartoon History of the Universe
The Cartoon Guide to Physics (with Art Huffman)
The Cartoon Guide to the Computer
The Cartoon Guide to Genetics (with Mark Wheelis)
The Cartoon History of the United States
The Cartoon Guide to (Non) Communication

THE CARTOON GUIDE TO STATISTICS



LARRY GONICK
& WOOLLCOTT SMITH

HarperPerennial
A Division of HarperCollinsPublishers

FIRST HARPERPERENNIAL EDITION

Illustrations by Larry Gonick

Library of Congress Cataloging-in-Publication Data

Gonick, Larry.
The cartoon guide to statistics / Larry Gonick & Woollcott Smith.
—1st HarperPerennial ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-06-273102-5 (pbk.)
1. Statistics—Caricatures and cartoons. I. Smith, Woollcott, 1941– . II. Title.
QA276.12.G67 1993
519.5—dc20 92-54683

97 LG/RRD 20 19 18 17 16

♦CONTENTS♦

CHAPTER 1 ---- 1
WHAT IS STATISTICS?

CHAPTER 2 ---- 7
DATA DESCRIPTION

CHAPTER 3 ---- 27
PROBABILITY

CHAPTER 4 ---- 53
RANDOM VARIABLES

CHAPTER 5 ---- 73
A TALE OF TWO DISTRIBUTIONS

CHAPTER 6 ---- 89
SAMPLING

CHAPTER 7 ---- 111
CONFIDENCE INTERVALS

CHAPTER 8 ---- 137
HYPOTHESIS TESTING

CHAPTER 9 ---- 157
COMPARING TWO POPULATIONS

CHAPTER 10 ---- 181
EXPERIMENTAL DESIGN

CHAPTER 11 ---- 187
REGRESSION

CHAPTER 12 ---- 211
CONCLUSION

BIBLIOGRAPHY ---- 221

INDEX ---- 224

Acknowledgments

WE WOULD LIKE TO THANK CAROL COHEN AT HARPERCOLLINS FOR SUGGESTING THIS PROJECT, OUR EDITOR ERICA SPABERG FOR PATIENTLY ENDURING THE LAST-MINUTE DASH TO THE DEADLINE, AND VICKY BIJUR, OUR LITERARY AGENT, FOR INITIATING THE GONICK/SMITH COLLABORATION BY INTRODUCING THE COAUTHORS.

WILLIAM FAIRLEY'S AND LEAH SMITH'S COMMENTS IMPROVED EARLIER DRAFTS OF THIS BOOK.

DONNA OKINO PROVIDED INVALUABLE ASSISTANCE AND ADVICE IN PRODUCING THE CARTOON PAGES. SHE SAYS THAT CREATING A CARTOON GUIDE IS HARDER THAN RUNNING A MARATHON, AND SHE SHOULD KNOW; SHE'S DONE BOTH.

THE ALTSYS CORPORATION CREATED FONTOGRAPHER, THE WONDERFUL SOFTWARE THAT ALLOWED US TO SIMULATE HAND-LETTERED TEXT AND FORMULAS ON THE MACINTOSH.

AND, SINCE EDUCATION IS ALWAYS A TWO-WAY STREET, A TIP OF THE HAT TO SMITH'S LONG-SUFFERING TEMPLE UNIVERSITY STUDENTS AND ESPECIALLY THE FALL '92 STUDY GROUP ORGANIZED BY ADRIANA TORRES. THE FUTURE IS THEIRS.

♦Chapter 1♦
WHAT IS STATISTICS?

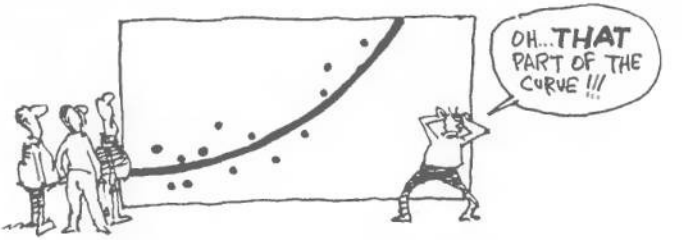
WE MUDDLE THROUGH LIFE MAKING CHOICES BASED ON *INCOMPLETE INFORMATION*...

MOST OF US LIVE COMFORTABLY WITH SOME LEVEL OF UNCERTAINTY.

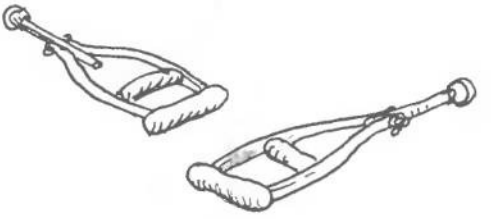
WHAT MAKES STATISTICS UNIQUE IS ITS ABILITY TO *QUANTIFY* UNCERTAINTY, TO MAKE IT PRECISE. THIS ALLOWS STATISTICIANS TO MAKE *CATEGORICAL STATEMENTS*, WITH COMPLETE ASSURANCE—ABOUT THEIR LEVEL OF UNCERTAINTY!

THIS IS NOT JUST A MATTER OR ORDERING SOUP! STATISTICS ALSO INVOLVES MATTERS OF *LIFE AND DEATH...*

FOR EXAMPLE, IN 1986, THE SPACE SHUTTLE *CHALLENGER* EXPLODED, KILLING SEVEN ASTRONAUTS. THE DECISION TO LAUNCH IN 29-DEGREE WEATHER HAD BEEN MADE WITHOUT DOING A SIMPLE ANALYSIS OF PERFORMANCE DATA AT LOW TEMPERATURE.



A MORE POSITIVE EXAMPLE IS THE *SALK POLIO VACCINE*. IN 1954, VACCINE TRIALS WERE PERFORMED ON SOME 400,000 CHILDREN, WITH STRICT CONTROLS TO ELIMINATE BIASED RESULTS. GOOD STATISTICAL ANALYSIS OF THE RESULTS FIRMLY ESTABLISHED THE VACCINE'S EFFECTIVENESS, AND TODAY POLIO IS ALMOST UNKNOWN.



TO ACCOMPLISH THEIR FEATS OF MATHEMATICAL LEGERDEMAIN, STATISTICIANS RELY ON THREE RELATED DISCIPLINES:

Data analysis

THE GATHERING, DISPLAY, AND SUMMARY OF DATA;

Probability

THE LAWS OF CHANCE, IN AND OUT OF THE CASINO;

Statistical inference

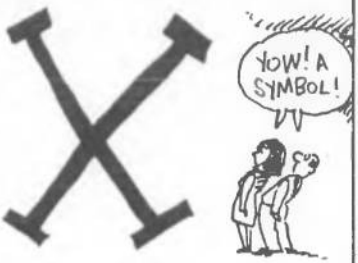
THE SCIENCE OF DRAWING STATISTICAL CONCLUSIONS FROM SPECIFIC DATA, USING A KNOWLEDGE OF PROBABILITY.

IN THIS BOOK, WE'LL LOOK AT ALL THREE, AS APPLIED TO A WIDE VARIETY OF SITUATIONS WHERE STATISTICS PLAYS A CRUCIAL ROLE IN THE MODERN WORLD.

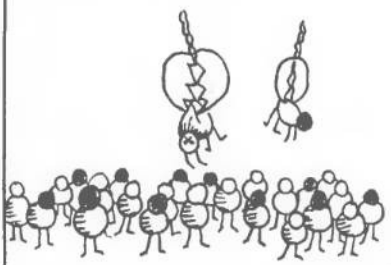
IN CHAPTER 2, WE'LL LOOK AT A SIMPLE DATA SET, THE REPORTED WEIGHTS OF A BUNCH OF COLLEGE STUDENTS.

IN CHAPTER 3, WE STUDY THE LAWS OF PROBABILITY IN THEIR BIRTHPLACE, THE GAMBLING DEN.

CHAPTERS 4 AND 5 SHOW HOW TO DESCRIBE THE WORLD WITH *PROBABILITY MODELS*, USING THE CONCEPT OF THE *RANDOM VARIABLE*.



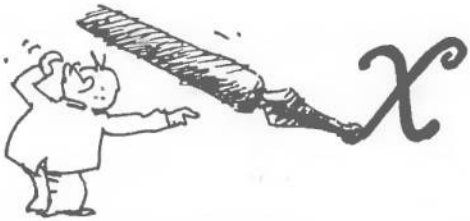
CHAPTER 6 INTRODUCES ONE OF THE STATISTICIAN'S ESSENTIAL PROCEDURES, TAKING *SAMPLES* OF A LARGE POPULATION.



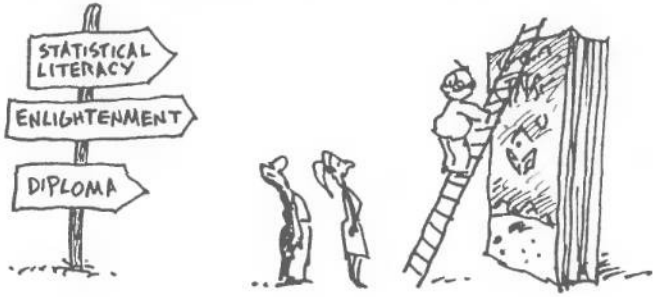
IN CHAPTER 7 AND BEYOND, WE DESCRIBE HOW TO MAKE STATISTICAL INFERENCES IN SUCH COMMON REAL-WORLD ARENAS AS *ELECTION POLLING, MANUFACTURING QUALITY CONTROL, MEDICAL TESTING, ENVIRONMENTAL MONITORING, RACIAL BIAS*, AND THE *LAW*.

FINALLY, IN DISCUSSING STATISTICS, IT'S HARD TO AVOID MENTIONING ONE OTHER THING: THE WIDESPREAD ***MISTRUST*** OF STATISTICS IN THE WORLD TODAY. EVERYONE KNOWS ABOUT "LYING WITH STATISTICS," WHILE GOOD STATISTICAL ANALYSIS IS NEARLY IMPOSSIBLE TO FIND IN DAILY LIFE. WHAT'S ONE TO DO?

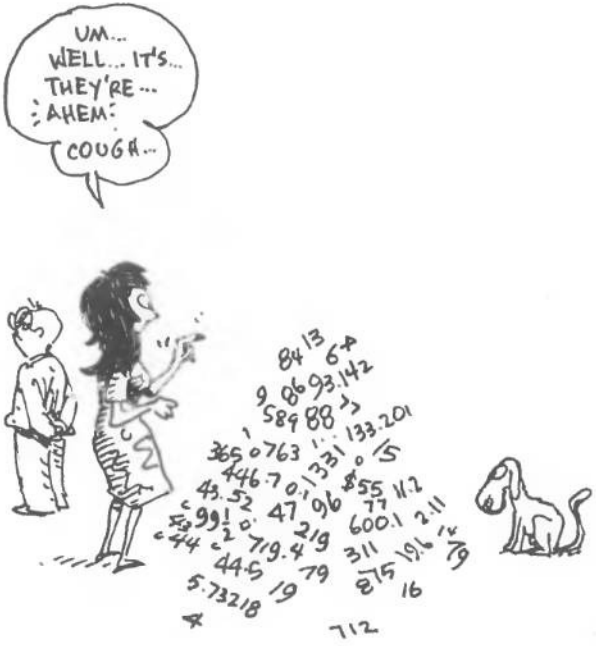
OUR HUMBLE OPINION IS THAT ***LEARNING A LITTLE MORE ABOUT THE SUBJECT*** MIGHT NOT BE SUCH A BAD IDEA.. AND THAT'S WHY WE WROTE THIS BOOK!



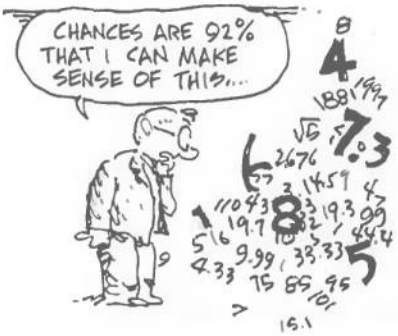
IN WHAT FOLLOWS, WE TRY TO PRESENT THE ELEMENTS OF STATISTICS AS GRAPHICALLY AND INTUITIVELY AS POSSIBLE. ALL YOU NEED TO GET THROUGH IT IS A LITTLE ***PATIENCE***, SOME ***THOUGHT***, AND A CERTAIN TOLERANCE FOR ***ALGEBRA***—OR, IF NOT THAT, THEN MAYBE A ***COURSE REQUIREMENT!!***



♦CHAPTER 2♦
DATA DESCRIPTION



DATA ARE THE STATISTICIAN'S RAW MATERIAL, THE NUMBERS WE USE TO INTERPRET REALITY. ALL STATISTICAL PROBLEMS INVOLVE EITHER THE COLLECTION, DESCRIPTION, AND ANALYSIS OF DATA, OR ***THINKING*** ABOUT THE COLLECTION, DESCRIPTION, AND ANALYSIS OF DATA.



THIS CHAPTER CONCENTRATES ON DATA ***DESCRIPTION.*** HOW CAN WE REPRESENT DATA IN USEFUL WAYS? HOW CAN WE SEE UNDERLYING PATTERNS IN A HEAP OF NAKED NUMBERS? HOW CAN WE SUMMARIZE THE DATA'S BASIC SHAPE?

WELL, TO DESCRIBE DATA, THE FIRST THING YOU NEED IS SOME ACTUAL DATA TO DESCRIBE... SO LET'S COLLECT SOME DATA!

HERE IS SOME REAL DATA: AS PART OF A CLASSROOM EXPERIMENT, 92 PENN STATE STUDENTS REPORTED THEIR WEIGHT, WITH THESE RESULTS:

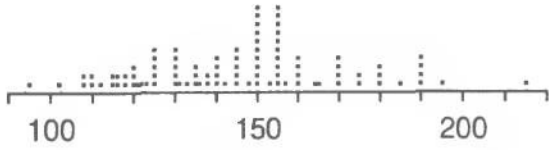
MALES

140 145 160 190 155 165 150 190 195 138 160 155 153 145 170 175 175 170 180 135 170 157 130 185 190 155 170 155 215 150 145 155 155 150 155 150 180 160 135 160 130 155 150 148 155 150 140 180 190 145 150 164 140 142 136 123 155

FEMALES

140 120 130 138 121 125 116 145 150 112 125 130 120 130 131 120 118 125 135 125 118 122 115 102 115 150 110 116 108 95 125 133 110 150 108

GETTING RIGHT DOWN TO BUSINESS, WE DRAW A *DOT PLOT*: ONE DOT PER STUDENT GOES OVER EACH STUDENT'S REPORTED WEIGHT.



Weight in Pounds

YOU MAY SEE A *PROBLEM* HERE: THE CLUMPS AT *150* AND *155* POUNDS. THE STUDENTS TENDED TO REPORT THEIR WEIGHT IN *FIVE-POUND INCREMENTS*. IN REAL-LIFE SITUATIONS LIKE THIS ONE, SUCH ROUNDING OFF CAN OBSCURE GENERAL PATTERNS IN DATA... BUT FOR NOW, WE'LL JUST WORK AROUND IT.

WE CAN SUMMARIZE THE DATA WITH A ***FREQUENCY TABLE.*** DIVIDE THE NUMBER LINE INTO INTERVALS AND COUNT THE NUMBER OF STUDENT WEIGHTS WITHIN EACH INTERVAL. THE ***FREQUENCY*** IS THE COUNT IN ANY GIVEN INTERVAL. THE ***RELATIVE FREQUENCY*** IS THE PROPORTION OF WEIGHTS IN EACH INTERVAL, I.E., IT'S THE FREQUENCY DIVIDED BY THE TOTAL NUMBER OF STUDENTS.

CLASS INTERVAL	MIDPOINT	FREQUENCY	RELATIVE FREQUENCY
87.5-102.4	95	2	.022
102.5-117.5	110	9	.098
117.5-132.4	125	19	.206
132.5-147.4	140	17	.185
147.5-162.4	155	27	.293
162.5-177.4	170	8	.087
177.5-192.4	185	8	.087
192.5-207.5	200	1	.011
207.5-222.4	215	1	.011
TOTAL		92	1.000

NOTE: WE KEPT THE INTERVAL BOUNDARIES AWAY FROM THOSE TROUBLESOME 5-POUND MULTIPLES. THIS GETS AROUND THE STUDENTS' REPORTING BIAS.

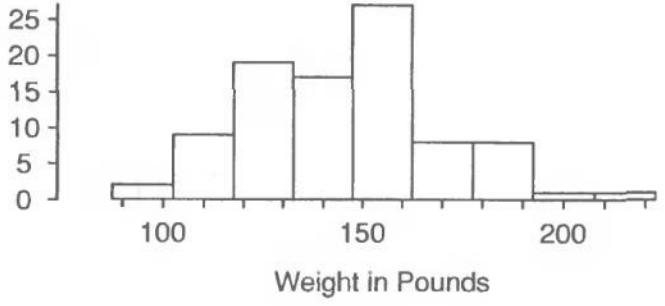
GUIDELINES FOR FORMING THE CLASS INTERVALS:

1) USE INTERVALS OF EQUAL LENGTH WITH MIDPOINTS AT CONVENIENT ROUND NUMBERS.

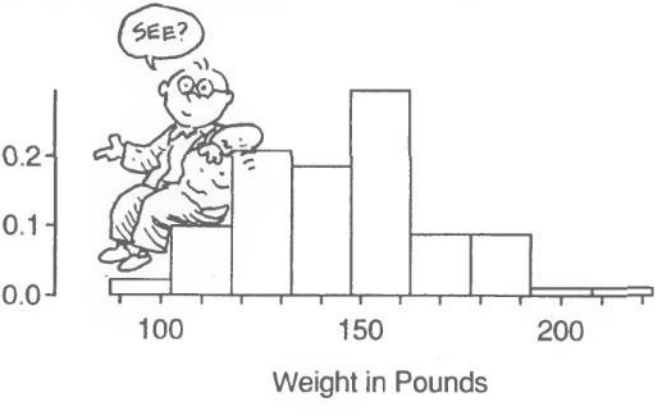
2) FOR A SMALL DATA SET, USE A SMALL NUMBER OF INTERVALS.

3) FOR A LARGE DATA SET, USE MORE INTERVALS!

IN THE FREQUENCY TABLE, WE ARE SHOWING HOW MANY DATA POINTS ARE "AROUND" EACH VALUE. WE CAN GRAPH THIS INFORMATION, TOO. THE RESULTING BAR GRAPH IS CALLED A *HISTOGRAM*. EACH BAR COVERS AN INTERVAL AND IS CENTERED AT THE MIDPOINT. THE BAR'S HEIGHT IS THE NUMBER OF DATA POINTS IN THE INTERVAL.



WE CAN ALSO DRAW A *RELATIVE FREQUENCY HISTOGRAM*, PLOTTING THE RELATIVE FREQUENCY AGAINST THE WEIGHT. IT LOOKS EXACTLY THE SAME, EXCEPT FOR THE VERTICAL SCALE.



THE STATISTICIAN *JOHN TUKEY* INVENTED A QUICK WAY TO SUMMARIZE DATA AND STILL KEEP THE INDIVIDUAL DATA POINTS. IT'S CALLED THE ***STEM-AND-LEAF*** DIAGRAM.

FOR THE WEIGHT DATA, THE STEM IS A COLUMN OF NUMBERS, CONSISTING OF THE WEIGHT DATA COUNTED BY TENS (I.E., WE LEAVE OFF THE LAST DIGIT).

```
9
10
11
12
13
14
15
16
17
18
19
20
21
```

NOW ADD THE FINAL DIGIT OF EACH WEIGHT IN THE APPROPRIATE ROW:

```
STEM : LEAVES
 9 :
10 :
11 : 628
12 : 0155005
13 : 080015
14 : 05
15 : 0
16 :
17 :
18 :
19 :
20 :
21 :
```

FILLED IN, IT LOOKS LIKE THIS:

```
9 : 5
10 : 288
11 : 628055060
12 : 01553005525
13 : 85008506001153
14 : 05505580502
15 : 505370550550505050050O
16 : 050004
17 : 055000
18 : 0500
19 : 00500
20:
21 : 5
```

AND FINALLY, PUT THE "LEAVES" IN ORDER.

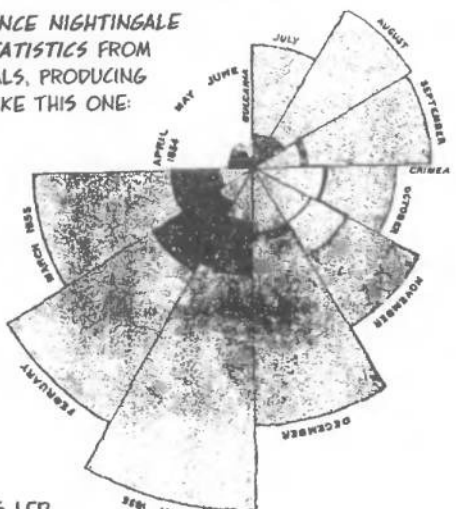
```
 9 : 5
10 : 288
11 : 002556688
12 : 00012355555
13 : 00000135556688
14 : 00002555558
15 : 000000000035555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20:
21 : 5
```

ALL THOSE ZEROES AND FIVES CLEARLY SHOW THE STUDENTS' REPORTING BIAS!

GOOD GRAPHIC DISPLAY IS PART ART AND PART SCIENCE

AND SOMETIMES, PART *POLITICS!*

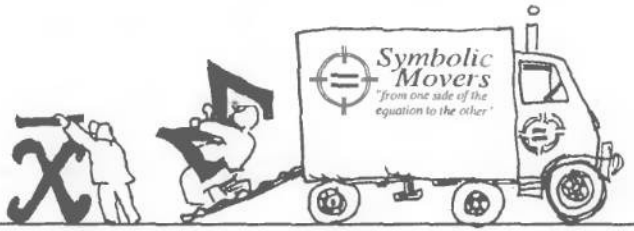
CRUSADING NURSE *FLORENCE NIGHTINGALE* COMPILED *MORTALITY STATISTICS* FROM BRITISH MILITARY HOSPITALS, PRODUCING SHOCKING HISTOGRAMS LIKE THIS ONE: THE RADIAL AXIS INDICATES DEATHS—IN HOSPITALS AS Wwell AS ON THE BATTLEFIELD—OF BRITISH SOLDIERS IN THE CRIMEAN WAR.



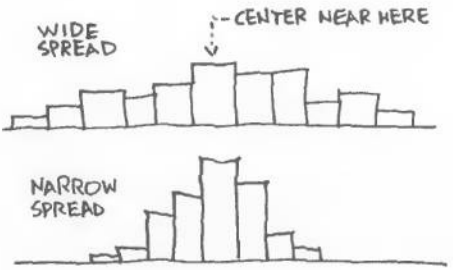
HER STATISTICAL EFFORTS LED DIRECTLY TO IMPROVED HOSPITAL CONDITIONS AND A REDUCTION IN THE DEATH RATE.

SUMMARY STATISTICS

NOW WE MOVE FROM PICTURES TO FORMULAS. OUR OBJECT IS TO GET SOME SIMPLE MEASUREMENTS OF THE CRUDEST CHARACTERISTICS OF A SET OF DATA....



ANY SET OF MEASUREMENTS HAS TWO IMPORTANT PROPERTIES: THE *CENTRAL OR TYPICAL VALUE*, AND THE *SPREAD* ABOUT THAT VALUE. YOU CAN SEE THE IDEA IN THESE HYPOTHETICAL HISTOGRAMS.



WE CAN GO A LONG WAY WITH A LITTLE NOTATION. SUPPOSE WE'RE MAKING A SERIES OF OBSERVATIONS... n OF THEM, TO BE EXACT... THEN WE WRITE

$$x_1, x_2, x_3, \ldots x_n$$

AS THE VALUES WE OBSERVE. THUS, n IS THE TOTAL NUMBER OF DATA POINTS, AND x_4 (SAY) IS THE VALUE OF THE FOURTH DATA POINT.

AN *ARRAY* IS A TABLE OF DATA:

OBSERVATION	1	2	3	4	....	n
DATA VALUE	x_1	x_2	x_3	x_4	....	x_n

A SMALL SET OF $n = 5$ DATA POINTS MAKES THE BOOKKEEPING EASY. SUPPOSE, FOR EXAMPLE, WE ASK FIVE PEOPLE HOW MANY HOURS OF *TELEVISION* THEY WATCH IN A WEEK... AND GET THE FOLLOWING ARRAY:

OBSERVATION	1	2	3	4	5
DATA VALUE	5	7	3	38	7

THEN $x_1 = 5$, $x_2 = 7$, $x_3 = 3$, $x_4 = 38$, AND $x_5 = 7$.

WHAT'S THE "CENTER" OF THESE DATA? THERE ARE ACTUALLY SEVERAL DIFFERENT WAYS TO MEASURE IT. WE'LL LOOK AT JUST TWO OF THEM.

THE **MEAN** (OR "AVERAGE")

THE ***MEAN*** OR AVERAGE VALUE IS REPRESENTED BY $\bar{x}$... IT'S OBTAINED BY ADDING ALL THE DATA AND DIVIDING BY THE NUMBER OF OBSERVATIONS:

$$\bar{x} = \frac{\text{SUM OF DATA}}{n}$$

$$= \frac{x_1 + x_2 + \ldots + x_n}{n}$$

FOR OUR EXAMPLE,

$$\bar{x} = \frac{5 + 7 + 3 + 38 + 7}{5} = \frac{60}{5}$$

$$= 12 \text{ HOURS}$$

WE HAVE A SHORTHAND FOR THAT $x_1 + x_2 + \dots + x_n$ USING THE GREEK CAPITAL LETTER *SIGMA*, FOR SUMMATION:

Σ

FOR THE SUM $x_1 + x_2 + \dots + x_n$ WE WRITE

$$\sum_{i=1}^{n} x_i$$

AND READ IT AS "THE SUM OF x_i AS i GOES FROM 1 TO n."

SAY IT TEN TIMES AND YOU'LL NEVER FORGET IT...

ALL RIGHT! NOW WE LOOKIN' LIKE A *STATISTICS* BOOK!

SO... TO REPEAT, THE AVERAGE, OR *MEAN*, OF A SET OF DATA x_i IS

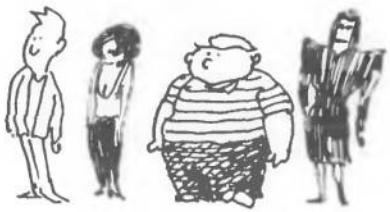
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{OR} \quad \sum_{i=1}^{n} \frac{x_i}{n}$$

IN THE CASE OF OUR 92 PENN STATE STUDENTS, THE MEAN WEIGHT IS

$$\sum_{i=1}^{92} \frac{x_i}{92} = \frac{13,354}{92}$$

$$=$$

145.15 POUNDS



THE MEDIAN

IS ANOTHER KIND OF CENTER: THE "MIDPOINT" OF THE DATA, LIKE THE "MEDIAN STRIP" IN A ROAD.

TO FIND THE MEDIAN VALUE OF A DATA SET, WE ARRANGE THE DATA IN ORDER FROM SMALLEST TO LARGEST. THE MEDIAN IS THE VALUE IN THE MIDDLE.

3 5 7 7 38

THE MEDIAN

IF THE NUMBER OF POINTS IS *EVEN*—IN WHICH CASE THERE IS NO MIDDLE, WE AVERAGE THE TWO VALUES AROUND THE MIDDLE... SO IF THE DATA ARE

3 5 7 7

MIDDLE SPACE

WE AVERAGE 5 AND 7 TO GET $\frac{5+7}{2} = 6$

THIS GIVES US A GENERAL RULE: ORDER THE DATA FROM SMALLEST TO LARGEST.

IF THE NUMBER OF DATA POINTS IS *ODD*, THE MEDIAN IS THE MIDDLE DATA POINT.

IF THE NUMBER OF POINTS IS *EVEN*, THE MEDIAN IS THE AVERAGE OF THE TWO DATA POINTS NEAREST THE MIDDLE.

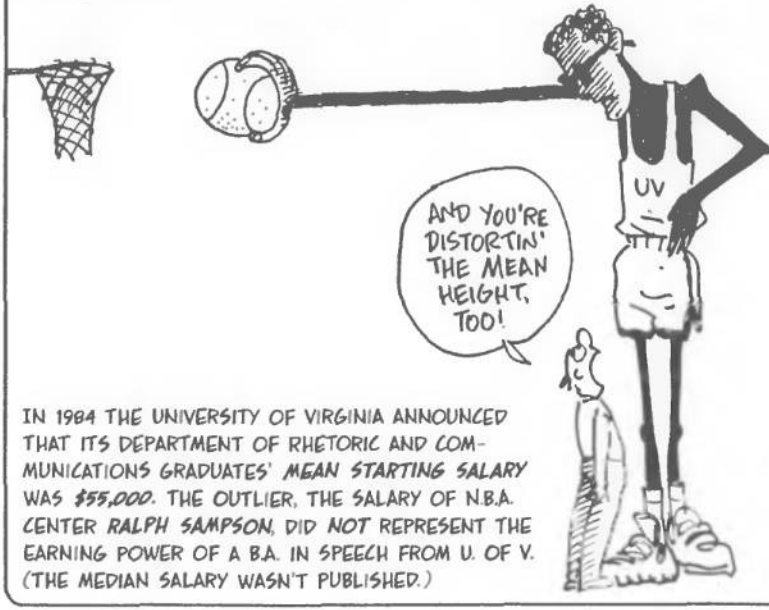
FOR THE $n = 92$ STUDENT WEIGHTS, WE CAN FIND THE MEDIAN FROM THE ORDERED STEM-AND-LEAF DIAGRAM: JUST COUNT TO THE 46^{TH} OBSERVATION. THE MEDIAN IS

$$\frac{x_{46} + x_{47}}{2} = \frac{145 + 145}{2}$$

$$= 145 \text{ POUNDS}$$

```
 9 : 5
10 : 288
11 : 002556688
12 : 00012355555
13 : 00000135556888
14 : 000025555588
15 : 00000000003555555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20 :
21 : 5
```

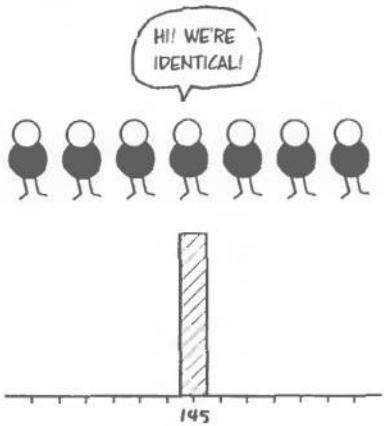
WHY MORE THAN ONE MEASURE OF THE CENTER? EACH HAS ADVANTAGES. FOR EXAMPLE, THE *MEDIAN* IS NOT SENSITIVE TO *OUTLIERS*, OR EXTREME VALUES NOT TYPICAL OF THE REST OF THE DATA. SUPPOSE IN OUR SMALL TV-WATCHING GROUP, ONE PERSON WATCHES *200* HOURS PER WEEK. THEN OUR DATA ARE 3, 5, 7, 7, 200. THE MEDIAN, 7, IS UNCHANGED, BUT THE MEAN IS NOW $\bar{x} = 45.8$!



IN 1984 THE UNIVERSITY OF VIRGINIA ANNOUNCED THAT ITS DEPARTMENT OF RHETORIC AND COMMUNICATIONS GRADUATES' *MEAN STARTING SALARY* WAS *$55,000*. THE OUTLIER, THE SALARY OF N.B.A. CENTER *RALPH SAMPSON*, DID *NOT* REPRESENT THE EARNING POWER OF A B.A. IN SPEECH FROM U. OF V. (THE MEDIAN SALARY WASN'T PUBLISHED.)

MEASURES OF SPREAD

BESIDES KNOWING THE CENTRAL POINT OF A DATA SET, WE'D ALSO LIKE TO DESCRIBE THE DATA'S *SPREAD*, OR HOW FAR FROM THE CENTER THE DATA TEND TO RANGE. FOR INSTANCE, IF THE STUDENTS ALL WEIGHED *EXACTLY* 145 POUNDS, THERE WOULD BE NO SPREAD AT ALL. NUMERICALLY, THE SPREAD WOULD BE *ZERO*, AND THE HISTOGRAM WOULD BE SKINNY.



BUT IF MANY OF THE STUDENTS WERE VERY LIGHT AND/OR VERY HEAVY, OBVIOUSLY WE'D SEE SOME SPREAD—SAY, IF THE *FOOTBALL TEAM* WAS PART OF THE SAMPLE...

THE HISTOGRAM WOULD BE WIDER, SOMETHING LIKE THIS:

AGAIN, THERE'S MORE THAN ONE WAY TO MEASURE A SPREAD. ONE WAY IS

INTERQUARTILE RANGE

THE IDEA IS TO DIVIDE THE DATA INTO FOUR EQUAL GROUPS AND SEE HOW FAR APART THE EXTREME GROUPS ARE.

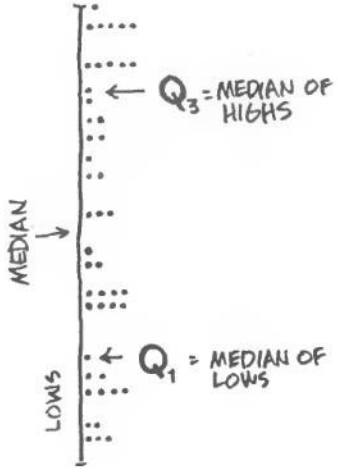
HERE'S THE RECIPE:

1) PUT THE DATA IN NUMERICAL ORDER.

2) DIVIDE THE DATA INTO TWO EQUAL HIGH AND LOW GROUPS AT THE MEDIAN. (IF THE MEDIAN IS A DATA POINT, INCLUDE IT IN BOTH THE HIGH AND LOW GROUPS.)

3) FIND THE MEDIAN OF THE LOW GROUP. THIS IS CALLED THE FIRST QUARTILE, OR Q_1.

4) THE MEDIAN OF THE HIGH GROUP IS THE THIRD QUARTILE, OR Q_3.



NOW THE INTERQUARTILE RANGE (IQR) IS THE DISTANCE (OR DIFFERENCE) BETWEEN THEM:

$$IQR = Q_3 - Q_1$$

HERE'S THE WEIGHT DATA WITH THE MIDPOINTS OF THE HIGH AND LOW GROUPS EMPHASIZED:

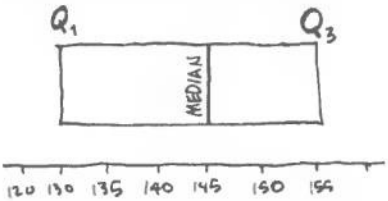
```
 9 : 5
10 : 288
11 : 002556688
12 : 000123555555
13 : 00000135556688
14 : 00002555558
15 : 000000000035555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20:
21 : 5
```

AND WE SEE THAT

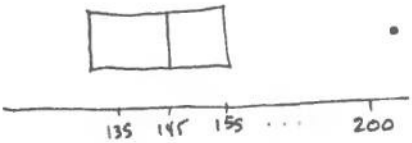
IQR = 156 − 125
= 31 POUNDS

AGAIN, THIS IS THE DIFFERENCE BETWEEN THE MEDIAN HEAVY STUDENT AND MEDIAN LIGHT ONE.

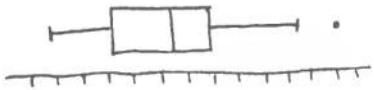
JOHN TUKEY INVENTED ANOTHER KIND OF DISPLAY TO SHOW OFF THE IQR, CALLED A *BOX AND WHISKERS* PLOT. THE BOX'S ENDS ARE THE QUARTILES Q_1 AND Q_3. WE DRAW THE MEDIAN INSIDE THE BOX.

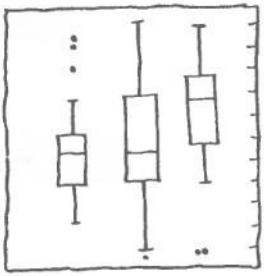


IF A POINT IS MORE THAN 1.5 IQR FROM AN END OF THE BOX, IT'S AN *OUTLIER*. DRAW THE OUTLIERS INDIVIDUALLY.



FINALLY, EXTEND "WHISKERS" OUT TO THE FARTHEST POINTS THAT ARE NOT OUTLIERS (I.E., WITHIN 1.5 IQR OF THE QUARTILES).



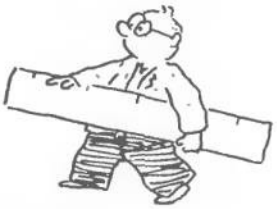


BOX-AND-WHISKERS PLOTS ARE ESPECIALLY GOOD FOR SHOWING OFF DIFFERENCES BETWEEN GROUPS.

THE STANDARD MEASURE OF SPREAD IS THE

STANDARD DEVIATION

UNLIKE THE IQR, WHICH IS BASED ON MEDIANS, THE STANDARD DEVIATION MEASURES THE SPREAD FROM THE *MEAN*. YOU CAN THINK OF IT, ROUGHLY SPEAKING, AS THE AVERAGE DISTANCE OF THE DATA FROM THE MEAN $\bar{x}$...

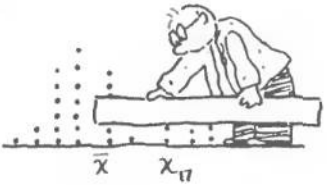


EXCEPT THAT WE USE THE *SQUARES* OF THE DISTANCES INSTEAD. THAT IS, IF THE SQUARED DISTANCE OF POINT x_i TO $\bar{x}$ IS $(x_i - \bar{x})^2$, THEN

$$\text{AVERAGE SQUARED DISTANCE} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

FOR TECHNICAL REASONS, WE USE $n-1$ IN THE DENOMINATOR RATHER THAN n, AND DEFINE THE *SAMPLE VARIANCE* s^2 AS

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

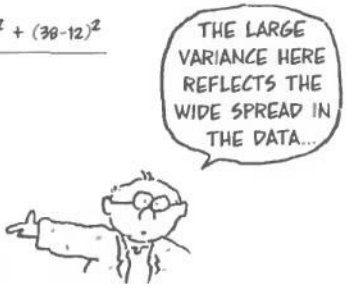


FOR THE DATA SET $\{3 \quad 5 \quad 7 \quad 7 \quad 38\}$, WITH $\bar{x} = 12$ AND $n = 5$ WE CALCULATE THE VARIANCE:

$$s^2 = \frac{(3-12)^2 + (5-12)^2 + (7-12)^2 + (7-12)^2 + (38-12)^2}{(5-1)}$$

$$= \frac{81 + 49 + 25 + 25 + 676}{4}$$

$$= 214$$



BUT A SPREAD MEASURE SHOULD HAVE THE SAME *UNITS* AS THE ORIGINAL DATA. IN THE EXAMPLE OF WEIGHTS, THE VARIANCE S^2 IS MEASURED IN POUNDS *SQUARED*... OOOPS!

THE OBVIOUS THING TO DO IS TO TAKE THE SQUARE ROOT, AND SO WE DO... TO DEFINE:

STANDARD DEVIATION

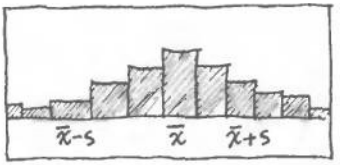
$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

WHICH, FOR OUR SIMPLE DATA SET, IS

$$s = \sqrt{214} = 14.63$$

EVEN FOR SMALL DATA SETS, THE ARITHMETIC CAN BE TEDIOUS! SO NOWADAYS, WE JUST HIT THE s BUTTON ON THE HAND CALCULATOR, OR CONSULT THE DATA REPORT GENERATED BY A COMPUTER SOFTWARE PACKAGE.

Properties of $\bar{x}$ and s

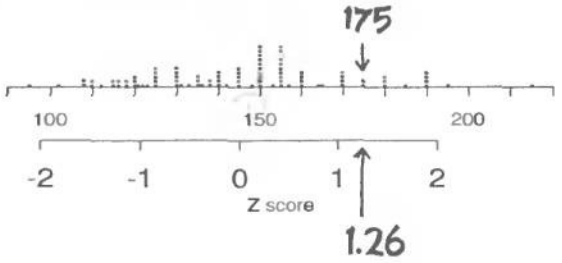


THE MEAN AND STANDARD DEVIATION ARE VERY GOOD FOR SUMMARIZING THE PROPERTIES OF FAIRLY *SYMMETRICAL HISTOGRAMS WITHOUT OUTLIERS*—I.E., HISTOGRAMS SHAPED LIKE MOUNDS.

IT'S OFTEN USEFUL TO KNOW HOW MANY STANDARD DEVIATIONS A DATA POINT IS FROM THE MEAN. WE DEFINE *z-SCORES*, OR STANDARDIZED SCORES, AS DISTANCE FROM $\bar{x}$ *PER STANDARD DEVIATION*.

$$z_i = \frac{x_i - \bar{x}}{s} \quad \text{FOR EACH } i.$$

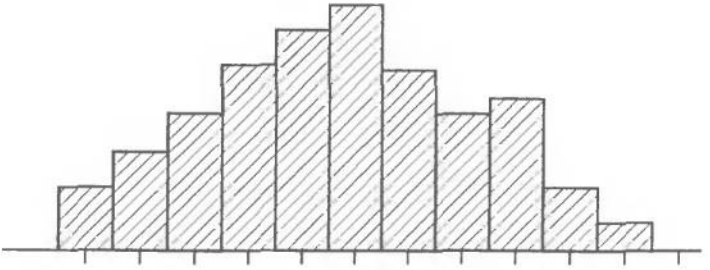
A z-SCORE OF +2 MEANS THAT AN OBSERVATION IS *TWO STANDARD DEVIATIONS ABOVE THE MEAN*. FOR THE WEIGHT DATA ($\bar{x}$=145.2 AND s=23.7), WE CAN PLOT THE DATA ON THE ORIGINAL x-AXIS IN POUNDS AND THE z-SCORE AXIS SIMULTANEOUSLY.



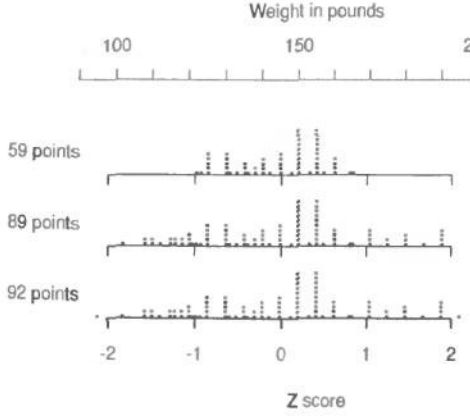
A STUDENT WEIGHING 175 POUNDS HAS A z-SCORE OF $\frac{175-145.2}{23.7} = 1.26$

an EMPIRICAL RULE:

FOR NEARLY SYMMETRIC MOUND-SHAPED DATA SETS, APPROXIMATELY **68%** OF THE DATA IS WITHIN *ONE* STANDARD DEVIATION OF THE MEAN AND **95%** OF THE DATA IS WITHIN *TWO* STANDARD DEVIATIONS OF THE MEAN.



FOR THE WEIGHTS, OUR EMPIRICAL RULE HOLDS UP PRETTY WELL: **64%** **(=59/92)** OF THE WEIGHTS ARE WITHIN ONE STANDARD DEVIATION OF THE MEAN, AND **97%** **(= 89/92)** OF THE WEIGHTS ARE WITHIN TWO STANDARD DEVIATIONS OF THE MEAN.



AND NOW FOR A REST FROM NUMBER CRUNCHING!

WE'VE COME A LONG WAY IN THIS CHAPTER! STARTING WITH A UNORGANIZED PILE OF NUMBERS, WE HAVE:

1) FOUND SEVERAL DIFFERENT WAYS TO DISPLAY THEM

2) LOOKED AT TWO DIFFERENT CONCEPTS OF THE CENTER OF DATA, THE MEDIAN AND THE MEAN

3) MEASURED THE SPREAD OF THE DATA AROUND THE CENTER IN TWO DIFFERENT WAYS

4) ENCOUNTERED MOUND-SHAPED HISTOGRAMS AND Z, A VARIABLE THAT INDICATES HOW MANY STANDARD DEVIATIONS YOU ARE FROM THE MEAN.

NOW, IN ORDER TO PROBE THE BEHAVIOR OF DATA MORE DEEPLY, WE'RE GOING TO MAKE A LITTLE DETOUR INTO THE REALM OF *RANDOMNESS*... A LAND WHERE THINGS *ALWAYS* WORK OUT IN THE LONG RUN, AND WHERE THE ONLY LAW IS THE LAW OF THE *GAMBLING CASINO*...

♦Chapter 3♦
PROBABILITY

NOTHING IN LIFE IS CERTAIN. IN EVERYTHING WE DO, WE GAUGE THE CHANCES OF SUCCESSFUL OUTCOMES, FROM BUSINESS TO MEDICINE TO THE WEATHER. BUT FOR MOST OF HUMAN HISTORY, *PROBABILITY*, THE FORMAL STUDY OF THE LAWS OF CHANCE, WAS USED FOR ONLY ONE THING: *GAMBLING*.

NOBODY KNOWS WHEN GAMBLING BEGAN. IT GOES BACK AT LEAST AS FAR AS *ANCIENT EGYPT*, WHERE SPORTING MEN AND WOMEN USED FOUR-SIDED *"ASTRAGALI"* MADE FROM ANIMAL HEELBONES.

THE ROMAN EMPEROR *CLAUDIUS* (10 BCE-54 CE) WROTE THE FIRST KNOWN TREATISE ON GAMBLING. UNFORTUNATELY, THIS BOOK, "HOW TO WIN AT DICE," WAS LOST.

MODERN DICE GREW POPULAR IN THE MIDDLE AGES, IN TIME FOR A RENAISSANCE RAKE, THE *CHEVALIER DE MERE*, TO POSE A MATHEMATICAL PUZZLER:

THE CHEVALIER REASONED THAT THE AVERAGE NUMBER OF SUCCESSFUL ROLLS WAS THE SAME FOR BOTH GAMBLES:

CHANCE OF ONE SIX = $\frac{1}{6}$

AVERAGE NUMBER IN FOUR ROLLS = $4 \cdot (\frac{1}{6}) = \frac{2}{3}$

CHANCE OF DOUBLE SIX IN ONE ROLL = $\frac{1}{36}$

AVERAGE NUMBER IN 24 ROLLS = $24 \cdot (\frac{1}{36}) = \frac{2}{3}$

WHY, THEN, DID HE *LOSE MORE OFTEN* WITH THE SECOND GAMBLE???

DE MERE PUT THE QUESTION TO HIS FRIEND, THE GENIUS *BLAISE PASCAL* (1623-1666).

ALTHOUGH PASCAL HAD EARLIER GIVEN UP MATHEMATICS AS A FORM OF SEXUAL INDULGENCE (!), HE AGREED TO TACKLE DE MERE'S PROBLEM.

PASCAL WROTE HIS FELLOW GENIUS *PIERRE DE FERMAT*, AND WITHIN A FEW LETTERS, THE TWO HAD WORKED OUT THE THEORY OF PROBABILITY IN ITS MODERN FORM—EXCEPT, OF COURSE, FOR THE CARTOONS.

BASIC DEFINITIONS

AS OUR GAMBLER PLAYS A GAME, WE PLAY SCIENTIST, OBSERVING THE OUTCOME:

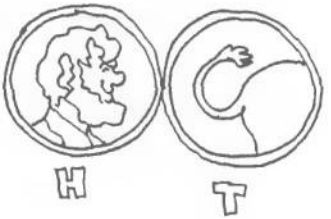
A **random experiment** IS THE PROCESS OF OBSERVING THE OUTCOME OF A CHANCE EVENT.

THE **elementary outcomes** ARE ALL POSSIBLE RESULTS OF THE RANDOM EXPERIMENT.

THE **sample space** IS THE SET OR COLLECTION OF ALL THE ELEMENTARY OUTCOMES.

IF THE EVENT WAS A COIN TOSS, FOR EXAMPLE, THE *RANDOM EXPERIMENT* CONSISTS OF RECORDING ITS OUTCOME...

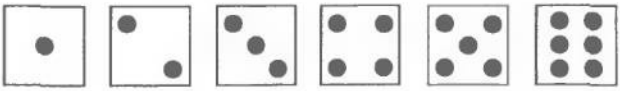
THE *ELEMENTARY OUTCOMES* ARE HEADS AND TAILS...



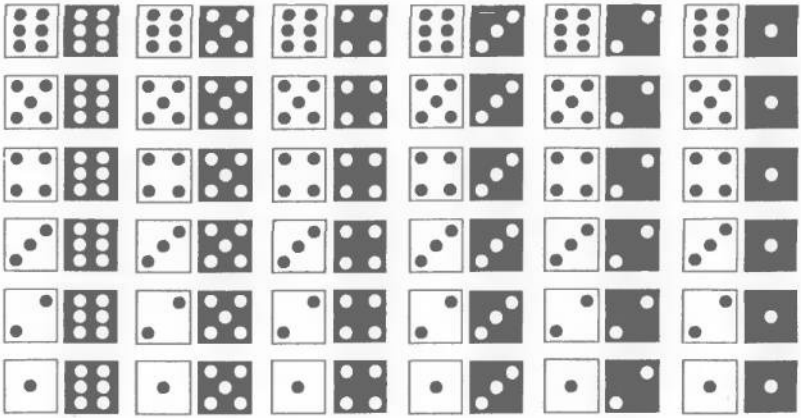
AND THE *SAMPLE SPACE* IS THE SET WRITTEN

$$\{H,T\}$$

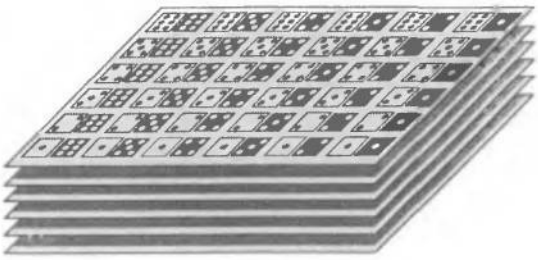
THE SAMPLE SPACE OF THE THROW OF A *SINGLE DIE* IS A LITTLE BIGGER.



AND FOR A *PAIR* OF DICE, THE SAMPLE SPACE LOOKS LIKE THIS (WE MAKE ONE DIE WHITE AND ONE BLACK TO TELL THEM APART):

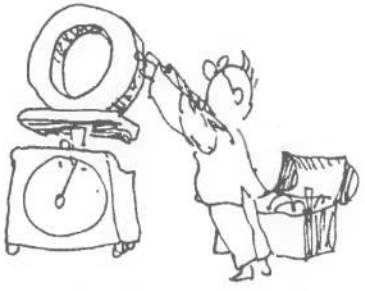


THIS SAMPLE SPACE HAS 36 (6X6) ELEMENTARY OUTCOMES. FOR THREE DICE, THE SPACE WOULD HAVE 216 ENTRIES, AS IN THIS 6X6X6 STACK. AND *FOUR DICE?*



AT SOME POINT, WE HAVE TO STOP LISTING, AND START THINKING...

NOW LET'S IMAGINE A RANDOM EXPERIMENT WITH n ELEMENTARY OUTCOMES $O_1, O_2, \ldots O_n$. WE WANT TO ASSIGN A ***NUMERICAL WEIGHT***, OR ***PROBABILITY***, TO EACH OUTCOME, WHICH MEASURES THE LIKELIHOOD OF ITS OCCURRING. WE WRITE THE PROBABILITY OF O_i AS $P(O_i)$.



FOR EXAMPLE, IN A FAIR COIN TOSS, HEADS AND TAILS ARE EQUALLY LIKELY, AND WE ASSIGN THEM BOTH THE PROBABILITY .5.

$$P(H) = P(T) = .5$$

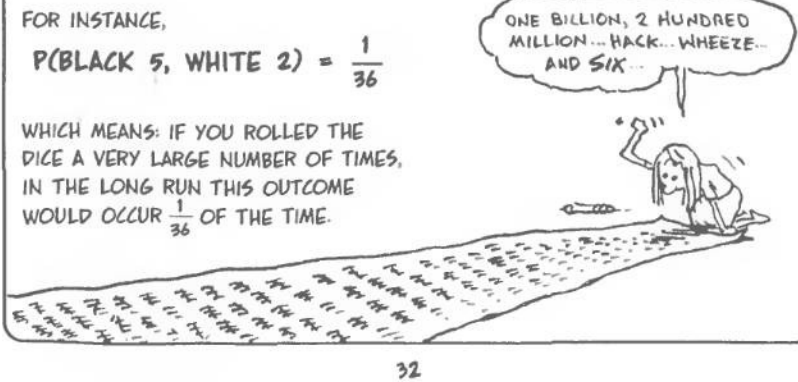
EACH OUTCOME COMES UP HALF THE TIME. ASK ANY ***FOOTBALL PLAYER!***

IN THE ROLL OF ***TWO DICE***, THERE ARE 36 ELEMENTARY OUTCOMES, ALL EQUALLY LIKELY, SO THE PROBABILITY OF EACH IS $\frac{1}{36}$.

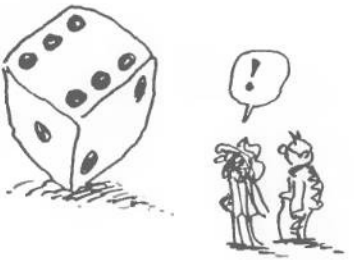
FOR INSTANCE,

$$P(\text{BLACK 5, WHITE 2}) = \frac{1}{36}$$

WHICH MEANS: IF YOU ROLLED THE DICE A VERY LARGE NUMBER OF TIMES, IN THE LONG RUN THIS OUTCOME WOULD OCCUR $\frac{1}{36}$ OF THE TIME.

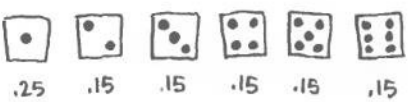


WHAT IF OUR GAMBLER *CHEATS* AND THROWS A *LOADED DIE?* FOR THE SAKE OF ARGUMENT, SUPPOSE THAT NOW A ONE COMES UP *25%* OF THE TIME (IN THE LONG RUN).



THE SAMPLE SPACE IS THE SAME AS FOR A FAIR DIE

$\{1, 2, 3, 4, 5, 6\}$



BUT THE PROBABILITIES ARE DIFFERENT. NOW $P(1) = .25$ AND THE REMAINING PROBABILTIES ADD UP TO .75. IF 2, 3, 4, 5, AND 6 WERE ALL EQUALLY LIKELY, THEN EACH ONE WOULD HAVE PROBABILITY $.15 = \frac{1}{5}(.75)$

IN GENERAL, ELEMENTARY OUTCOMES NEED NOT HAVE EQUAL PROBABILITY.

NOW WHAT CAN WE SAY ABOUT THE PROBABILITIES $P(O_i)$ IN AN ARBITRARY RANDOM EXPERIMENT? FIRST OF ALL,

$$P(O_i) \geq 0$$

PROBABILITIES ARE *NEVER NEGATIVE*. A PROBABILITY OF ZERO MEANS AN EVENT CAN'T HAPPEN. LESS THAN ZERO WOULD BE MEANINGLESS.

SECOND, IF AN EVENT IS *CERTAIN* TO HAPPEN, WE ASSIGN IT PROBABILITY 1. (IN THE LONG RUN, THAT'S THE PROPORTION OF TIMES IT WILL OCCUR!)

IN PARTICULAR, THE *TOTAL PROBABILITY OF THE SAMPLE SPACE* MUST BE 1. IF WE DO THE EXPERIMENT, *SOMETHING* IS BOUND TO HAPPEN!

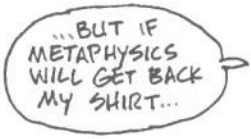
PUT THESE TWO TOGETHER, AND YOU HAVE THE *CHARACTERISTIC PROPERTIES OF PROBABILITY*:

$$P(O_i) \geq 0$$

PROBABILITY IS *NON-NEGATIVE*

$$P(O_1) + P(O_2) + \ldots + P(O_n) = 1$$

TOTAL PROBABILITY OF ALL ELEMENTARY OUTCOMES IS *ONE*.



LIKE A CLEVER POLITICIAN, WE HAVE AVOIDED CERTAIN *UNPLEASANT QUESTIONS*, SUCH AS A) WHAT DOES PROBABILITY *MEAN?* AND B) *HOW* DO WE ASSIGN PROBABILITIES TO OUTCOMES?

HERE ARE SOME APPROACHES THAT HAVE BEEN TAKEN:

Classical PROBABILITY: BASED ON GAMBLING IDEAS, THE FUNDAMENTAL ASSUMPTION IS THAT THE GAME IS FAIR AND ALL ELEMENTARY OUTCOMES HAVE THE SAME PROBABILITY.

Relative Frequency:

WHEN AN EXPERIMENT *CAN* BE REPEATED, THEN AN EVENT'S PROBABILITY IS THE PROPORTION OF TIMES THE EVENT OCCURS IN THE LONG RUN.

Personal PROBABILITY: MOST OF LIFE'S EVENTS ARE *NOT REPEATABLE*. PERSONAL PROBABILITY IS AN *INDIVIDUAL'S PERSONAL ASSESSMENT* OF AN OUTCOME'S LIKELIHOOD. IF A GAMBLER BELIEVES THAT A HORSE HAS MORE THAN A 50% CHANCE OF WINNING, HE'LL TAKE AN EVEN BET ON THAT HORSE.

AN *OBJECTIVIST* USES EITHER THE CLASSICAL OR FREQUENCY DEFINITION OF PROBABILITY. A *SUBJECTIVIST* OR *BAYESIAN* APPLIES FORMAL LAWS OF CHANCE TO HIS OWN, OR YOUR, PERSONAL PROBABILITIES.

BASIC OPERATIONS

SO FAR, WE HAVE DISCUSSED ONLY THE PROBABILITY OF ELEMENTARY OUTCOMES. IN THEORY, THAT WOULD BE ENOUGH TO DESCRIBE ANY RANDOM EXPERIMENT, BUT IN PRACTICE IT'S PRETTY UNWIELDY. FOR EXAMPLE, EVEN SUCH AN ORDINARY OCCURRENCE AS ROLLING A SEVEN IS NOT AN ELEMENTARY OUTCOME... SO WE INTRODUCE A *NEW IDEA:*

AN ***EVENT*** IS A SET OF ELEMENTARY OUTCOMES. THE PROBABILITY OF AN EVENT IS THE SUM OF THE PROBABILITIES OF THE ELEMENTARY OUTCOMES IN THE SET. FOR INSTANCE, SOME EVENTS IN THE LIFE OF A TWO-DICED ROLLER ARE:

EVENT DESCRIPTION	EVENT'S ELEMENTARY OUTCOMES	PROBABILITY
A: DICE ADD TO 3	{(1,2), (2,1)}	$P(A) = \frac{2}{36}$
B: DICE ADD TO 6	{(1,5), (2,4), (3,3), (4,2), (5,1)}	$P(B) = \frac{5}{36}$
C: WHITE DIE SHOWS 1	{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)}	$P(C) = \frac{6}{36}$
D: BLACK DIE SHOWS 1	{(1,1), (2,1), (3,1), (4,1), (5,1), (6,1)}	$P(D) = \frac{6}{36}$

THE BEAUTY OF USING EVENTS, RATHER THAN ELEMENTARY OUTCOMES, IS THAT WE CAN *COMBINE* EVENTS TO MAKE OTHER EVENTS, USING *LOGICAL OPERATIONS*. THE RELEVANT WORDS ARE ***AND***, ***OR***, AND ***NOT***.

THAT IS, GIVEN EVENTS E AND F, WE CAN MAKE NEW EVENTS:

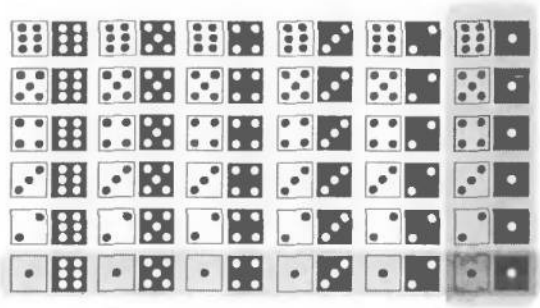
E **and** F : THE EVENT E AND THE EVENT F BOTH OCCUR.

E **or** F : THE EVENT E OR THE EVENT F OCCURS (OR BOTH DO).

not E : THE EVENT E DOES NOT OCCUR.

COMBINING OUR PRIMITIVE DEFINITIONS OF PROBABILITY WITH THESE LOGICAL OPERATIONS WILL GIVE US SOME POWERFUL FORMULAS FOR MANIPULATING PROBABILITIES.

LET'S RETURN TO THE DICE-THROWING EXAMPLE. IF C IS THE EVENT, WHITE DIE = 1, AND D IS THE EVENT, BLACK DIE = 1, THEN:



C OR D IS THE ENTIRE SHADED AREA (WHERE ONE DIE OR THE OTHER IS 1).

C AND D IS WHERE THE SHADED AREAS OVERLAP (BOTH DICE ARE 1).

THIS ILLUSTRATES THE *ADDITION RULE*: FOR ANY EVENTS E, F,

P(E OR F) = P(E) + P(F) – P(E AND F)

ADDING P(E) + P(F) *DOUBLE COUNTS* THE ELEMENTARY OUTCOMES SHARED BY E AND F, SO WE HAVE TO SUBTRACT THE EXTRA AMOUNT, WHICH IS P(E AND F).

IN THE ABOVE EXAMPLE,

$$P(C \text{ OR } D) = \frac{11}{36}$$

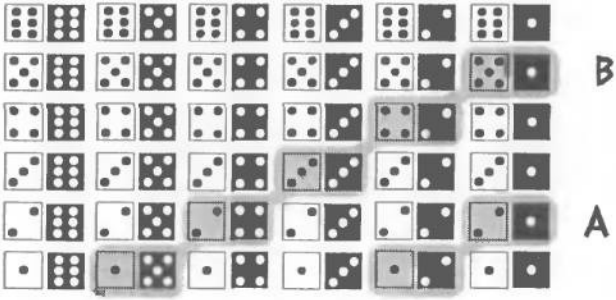
AS YOU CAN SEE BY COUNTING ELEMENTARY OUTCOMES. LIKEWISE,

$$P(C \text{ AND } D) = \frac{1}{36}$$

AND WE CONFIRM THE FORMULA:

$$P(C) + P(D) - P(C \text{ AND } D)$$
$$= \frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36}$$
$$= P(C \text{ OR } D)$$

SOMETIMES, THE OVERLAP **E AND F** IS EMPTY, AND THE TWO EVENTS HAVE NO ELEMENTARY OUTCOMES IN COMMON. IN THAT CASE, WE SAY E AND F ARE ***MUTUALLY EXCLUSIVE***, MAKING P(E AND F) = 0. HERE WE SEE THE MUTUALLY EXCLUSIVE EVENTS **A**, THE DICE ADD TO 3, AND **B**, THE DICE ADD TO 6.



FOR MUTUALLY EXCLUSIVE EVENTS, WE GET A *SPECIAL ADDITION RULE*: IF E AND F ARE MUTUALLY EXCLUSIVE, THEN

P(E OR F) = P(E) + P(F)

AND WE CHECK THAT P(A OR B) = $\frac{7}{36} = \frac{2}{36} + \frac{5}{36}$ = P(A) + P(B)

AND FINALLY, A *SUBTRACTION RULE*: FOR ANY EVENT E,

P(E) = 1 - P(NOT E)

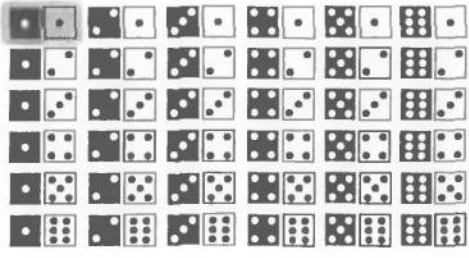
THIS IS USEFUL WHEN P(NOT E) IS EASIER TO COMPUTE THAN P(E). FOR INSTANCE, LET **E** BE THE EVENT, A DOUBLE-1 IS *NOT THROWN*. THE EVENT NOT-E, A DOUBLE-1 *IS* THROWN, HAS PROBABILITY P(NOT E) = $\frac{1}{36}$.

SO

P(E) = 1 − P(NOT E)

= $1 - \frac{1}{36}$

= $\frac{35}{36}$



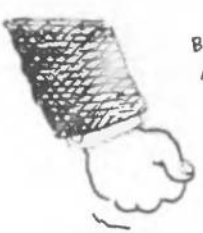
THE FORMULAS WE JUST DERIVED ARE, IN FACT, ADEQUATE FOR ANSWERING DE MERE'S QUESTION— BUT NOT EASILY! (YOU MIGHT TRY USING THEM ON A SIMPLER QUESTION: WHAT'S THE PROBABILITY OF ROLLING AT LEAST ONE SIX IN TWO ROLLS OF A SINGLE DIE?) WE NEED *MORE MACHINERY!*

SO WE INTRODUCE

conditional probability

(AN ESSENTIAL CONCEPT IN STATISTICS!)

SUPPOSE WE ALTER OUR EXPERIMENT SLIGHTLY, AND THROW THE WHITE DIE *BEFORE* THE BLACK DIE. WHAT'S THE PROBABILITY THAT THE FACES SUM TO 3?



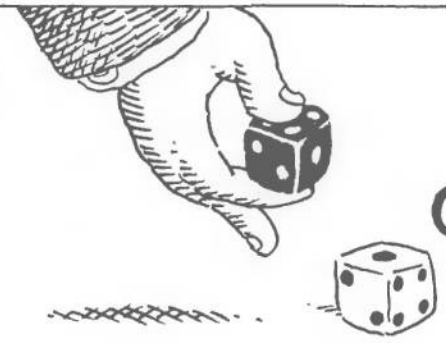
BEFORE THE DICE ARE THROWN, THE PROBABILITY IS $P(A) = \frac{2}{36}$

NOW SUPPOSE THE WHITE DIE COMES UP 1 (EVENT C). WHAT'S THE PROBABILITY OF A NOW?

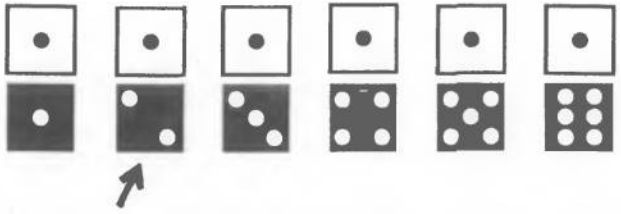
WE CALL IT THE *CONDITIONAL PROBABILITY* THAT EVENT A WILL OCCUR, GIVEN THE *CONDITION* THAT EVENT C HAS ALREADY OCCURRED. WE WRITE

P(A|C)

AND SAY "THE PROBABILITY OF A, *GIVEN* C."



BEFORE ANY DICE WERE THROWN, THE SAMPLE SPACE HAD 36 OUTCOMES, BUT NOW THAT THE EVENT C HAS OCCURRED, THE OUTCOME MUST BELONG TO THE *REDUCED SAMPLE SPACE C.*



IN THE REDUCED SAMPLE SPACE OF SIX ELEMENTARY OUTCOMES, ONLY ONE OUTCOME (1,2) SUMS TO 3. SO THE CONDITIONAL PROBABILITY IS 1/6.

IN GENERAL, TO FIND THE CONDITIONAL PROBABILITY **P(E|F)**, WE LOOK AT THE EVENT E AND F AS PART OF THE REDUCED SAMPLE SPACE F.

WE TRANSLATE THIS INTO A FORMAL DEFINITION: THE *CONDITIONAL PROBABILITY OF E, GIVEN F,* IS

$$P(E|F) = \frac{P(E \text{ and } F)}{P(F)}$$

FROM WHICH YOU CAN DIRECTLY VERIFY SOME INTUITIVE FACTS:

$P(E|E) = 1$ (ONCE E OCCURS, IT'S CERTAIN.)

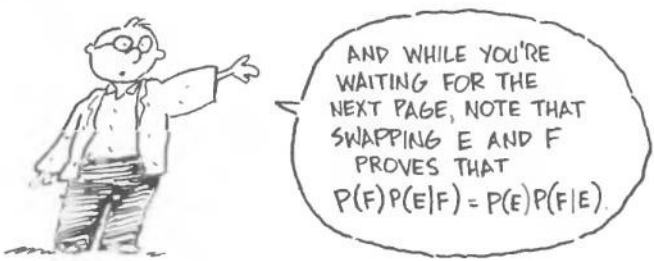
WHEN E AND F ARE MUTUALLY EXCLUSIVE,

$P(E|F) = 0$ (ONCE F HAS OCCURRED, E IS IMPOSSIBLE.)

REARRANGING THE DEFINITION GIVES US THE *MULTIPLICATION RULE:*

$$P(E \text{ AND } F) = P(E|F)P(F)$$

WHICH WE WOULD LIKE TO REDUCE TO A "SPECIAL" MULTIPLICATION RULE, UNDER THE FAVORABLE CIRCUMSTANCES THAT $P(E|F) = P(E)$. THAT WOULD BE EXCELLENT!



INDEPENDENCE and the special multiplication rule.

TWO EVENTS E AND F ARE *INDEPENDENT* OF EACH OTHER IF THE OCCURRENCE OF ONE HAS *NO INFLUENCE* ON THE PROBABILITY OF THE OTHER. FOR INSTANCE, THE ROLL OF ONE DIE HAS NO EFFECT ON THE ROLL OF ANOTHER (UNLESS THEY'RE GLUED TOGETHER, MAGNETIC, ETC.!).

IN TERMS OF CONDITIONAL PROBABILITY, THIS AMOUNTS TO SAYING $P(E) = P(E|F)$ OR, EQUIVALENTLY, $P(F) = P(F|E)$. WHEN E AND F ARE INDEPENDENT, WE GET A *SPECIAL MULTIPLICATION RULE*:

$$P(E \text{ AND } F) = P(E)P(F)$$

LET'S VERIFY THE INDEPENDENCE OF DICE, USING THE FORMULAS. *C* IS THE EVENT *WHITE DIE COMES UP 1*; *D* IS THE EVENT *BLACK DIE COMES UP 1*, AND WE HAVE:

$$P(C|D) = \frac{P(C \text{ AND } D)}{P(D)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} = P(C)$$

BUT THE WHITE DIE SHOWING 1 OBVIOUSLY DOES AFFECT THE CHANCES THAT THE SUM OF THE TWO DICE IS 3!

$$P(A|C) = \frac{P(A \text{ AND } C)}{P(C)} = \frac{P(1,2)}{P(C)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} \neq P(A) = \frac{1}{18}$$

SO THESE TWO EVENTS ARE *NOT INDEPENDENT*.

BEFORE GOING ON, LET'S SUMMARIZE ALL THE RULES WE'VE ACCUMULATED:

ADDITION RULE:

$$\mathbf{P(E \text{ OR } F) = P(E) + P(F) - P(E \text{ AND } F)}$$

SPECIAL ADDITION RULE: WHEN E AND F ARE MUTUALLY EXCLUSIVE,

$$\mathbf{P(E \text{ OR } F) = P(E) + P(F)}$$

SUBTRACTION RULE:

$$\mathbf{P(E) = 1 - P(\text{NOT } E)}$$

MULTIPLICATION RULE:

$$\mathbf{P(E \text{ AND } F) = P(E \mid F)P(F)}$$

SPECIAL MULTIPLICATION RULE: WHEN E AND F ARE INDEPENDENT,

$$\mathbf{P(E \text{ AND } F) = P(E)P(F)}$$

AND NOW, DE MERE'S PROBLEM AT LAST... LET E BE THE EVENT OF GETTING AT LEAST ONE SIX IN FOUR ROLLS OF A SINGLE DIE. WHAT'S P(E)? THIS IS ONE OF THOSE EVENTS WHOSE NEGATIVE IS EASIER TO DESCRIBE: *NOT E* IS THE EVENT OF *GETTING NO SIXES IN FOUR THROWS.*

IF A_i IS THE EVENT, GETTING NO SIX ON THE i^{TH} THROW, WE KNOW THAT $P(A_i) = \frac{5}{6}$. WE ALSO KNOW THAT ROLLS ARE INDEPENDENT, SO

$$P(\text{NOT } E) = P(A_1 \text{ AND } A_2 \text{ AND } A_3 \text{ AND } A_4)$$

MULTIPLICATION RULE →

$$= \left(\frac{5}{6}\right)^4 = .482,$$

SO

$$P(E) = 1 - P(\text{NOT } E) = .518$$

NOW THE SECOND HALF: LET F BE THE EVENT, GETTING AT LEAST ONE DOUBLE SIX IN 24 THROWS. AGAIN, **NOT F** IS EASIER TO DESCRIBE. IT'S THE EVENT OF GETTING NO DOUBLE SIXES.

IF B_i IS THE EVENT, NO DOUBLE SIX IS THROWN ON THE i^{TH} ROLL, THEN NOT F = B_1 AND B_2 AND... B_{24}. THE PROBABILITY OF EACH B IS

$$P(B_i) = \frac{35}{36}, \text{ SO}$$

$$P(\text{NOT F}) = \left(\frac{35}{36}\right)^{24} = .509$$

(BY THE MULTIPLICATION RULE) AND WE CONCLUDE THAT

$$P(F) = 1 - P(\text{NOT F}) = 1 - .509$$

$$= .491$$

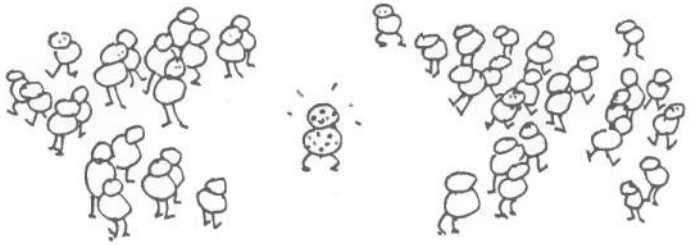
DE MERE TOLD PASCAL HE HAD ACTUALLY OBSERVED THAT EVENT F OCCURRED LESS OFTEN THAN EVENT E, BUT HE WAS AT A LOSS TO EXPLAIN WHY... FROM WHICH WE CONCLUDE THAT DE MERE GAMBLED OFTEN AND *KEPT CAREFUL RECORDS!!*

NOW LET'S LEAVE THE CASINO AND REJOIN THE REAL WORLD...

BAYES THEOREM and the case of the false positives...

FOR A MORE SERIOUS APPLICATION OF CONDITIONAL PROBABILITY, LET'S ENTER AN ARENA OF LIFE AND DEATH...

SUPPOSE A RARE DISEASE INFECTS ONE OUT OF EVERY 1000 PEOPLE IN A POPULATION...



AND SUPPOSE THAT THERE IS A GOOD, BUT NOT PERFECT, TEST FOR THIS DISEASE: IF A PERSON HAS THE DISEASE, THE TEST COMES BACK POSITIVE *99%* OF THE TIME. ON THE OTHER HAND, THE TEST ALSO PRODUCES SOME *FALSE POSITIVES*. ABOUT *2%* OF UNINFECTED PATIENTS ALSO TEST POSITIVE. AND YOU JUST TESTED POSITIVE. WHAT ARE YOUR CHANCES OF HAVING THE DISEASE?

WE HAVE TWO EVENTS TO WORK WITH:

A : PATIENT HAS THE DISEASE
B : PATIENT TESTS POSITIVE.

THE INFORMATION ABOUT THE TEST'S EFFECTIVENESS CAN BE WRITTEN

$P(A) = .001$ (ONE PATIENT IN 1000 HAS THE DISEASE)

$P(B|A) = .99$ (PROBABILITY OF A POSITIVE TEST, GIVEN INFECTION, IS .99)

$P(B|\text{NOT } A) = .02$ (PROBABILITY OF A FALSE POSITIVE, GIVEN NO INFECTION, IS .02)

AND WE ASK

$P(A|B) = \text{WHAT?}$ (PROBABILITY OF HAVING THE DISEASE, GIVEN A POSITIVE TEST)

SINCE THE TREATMENT FOR THIS DISEASE HAS SERIOUS SIDE EFFECTS, THE DOCTOR, HER LAWYER, AND HER LAWYER'S LAWYER CALL ON JOE BAYES, CP (CONSULTING PROBABILIST), FOR AN ANSWER. JOE DERIVES A THEOREM FIRST PROVED BY HIS ANCESTOR, THE REV. THOMAS *BAYES* (1744-1809).

JOE BEGINS WITH A 2X2 TABLE, WHICH DIVIDES THE SAMPLE SPACE INTO FOUR MUTUALLY EXCLUSIVE EVENTS. IT DISPLAYS EVERY POSSIBLE COMBINATION OF DISEASE STATE AND TEST RESULT.

	A	NOT A
B	A AND B	NOT A AND B
NOT B	A AND NOT B	NOT A AND NOT B

LET'S FIND THE PROBABILITIES OF EACH EVENT IN THE TABLE:

	A	NOT A	SUM
B	P(A AND B)	P(NOT A AND B)	P(B)
NOT B	P(A AND NOT B)	P(NOT A AND NOT B)	P(NOT B)
	P(A)	P(NOT A)	1

THE PROBABILITIES IN THE MARGINS ARE FOUND BY SUMMING ACROSS ROWS AND DOWN COLUMNS.

NOW COMPUTE:

$P(A \text{ AND } B) = P(B|A)P(A) = (.99)(.001) = .00099$

$P(\text{NOT A AND } B) = P(B|\text{NOT A})P(\text{NOT A}) = (.02)(.999) = .01998$

ALLOWING US TO FILL IN SOME ENTRIES:

	A	NOT A	SUM
B	.00099	.01998	.02097
NOT B	P(A AND NOT B)	P(NOT A AND NOT B)	P(NOT B)
	.001	.999	1

WE FIND THE REMAINING PROBABILITIES BY SUBTRACTING IN THE COLUMNS, THEN ADDING ACROSS THE ROWS.

THE FINAL TABLE IS:

	A	NOT A	
B	.00099	.01998	.02097 P(B)
NOT B	.00001	.97902	.97903 P(NOT B)
	.001	.999	1
	P(A)	P(NOT A)	

FROM WHICH WE DIRECTLY DERIVE

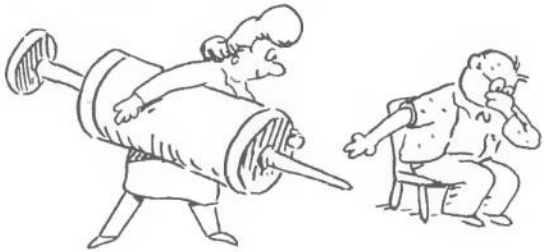
$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{.00099}{.02097} = .0472$$

DESPITE THE HIGH ACCURACY OF THE TEST, *LESS THAN 5%* OF THOSE WHO TEST POSITIVE ACTUALLY HAVE THE DISEASE! THIS IS CALLED THE *FALSE POSITIVE PARADOX.*

THIS TABLE SHOWS WHAT HAPPENS IN A GROUP OF A THOUSAND PATIENTS. ON AVERAGE, ONLY 21 PEOPLE WILL TEST POSITIVE—AND ONLY *ONE* OF THEM HAS THE DISEASE! 20 FALSE POSITIVES COME FROM THE *MUCH LARGER UNINFECTED GROUP.*

	DISEASE	NO DISEASE	
TESTS POSITIVE	1	20	21
TESTS NEGATIVE	0	979	979
	1	999	1000

WHAT'S THE PHYSICIAN TO DO? JOE BAYES ADVISES HER NOT TO START TREATMENT ON THE BASIS OF THIS TEST ALONE. THE TEST DOES PROVIDE INFORMATION, HOWEVER: WITH A POSITIVE TEST THE PATIENT'S CHANCE OF HAVING THE DISEASE INCREASED FROM 1 IN 1000 TO 1 IN 23. THE DOCTOR FOLLOWS UP WITH MORE TESTS.



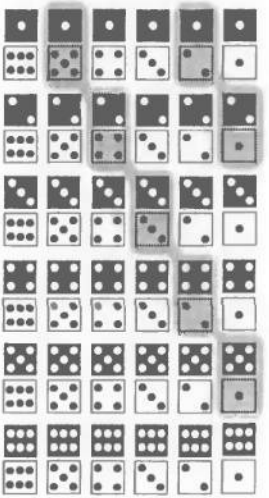
JOE BAYES COLLECTS HIS CONSULTING CHECK BEFORE ADMITTING THAT ALL THOSE STEPS HE WENT THROUGH CAN BE COMPRESSED INTO THE SINGLE FORMULA CALLED BAYES THEOREM:

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(\text{NOT A})P(B|\text{NOT A})}$$

IT COMPUTES P(A|B) FROM P(A) AND THE TWO CONDITIONAL PROBABILITIES P(B|A) AND P(B|NOT A). YOU CAN DERIVE IT BY NOTING THAT THE BIG FRACTION CAN BE EXPRESSED AS

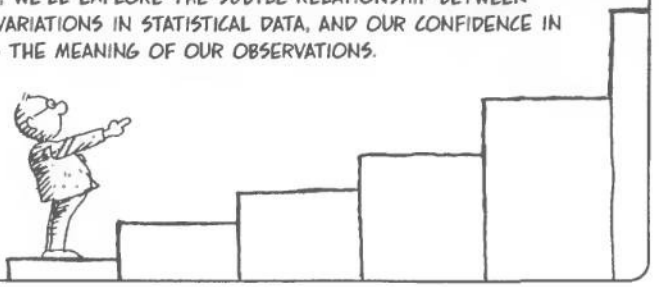
$$\frac{P(A \text{ and } B)}{P(A \text{ and } B)+P(\text{NOT A and } B)} = \frac{P(A \text{ and } B)}{P(B)} = P(A|B)$$

IN THIS CHAPTER, WE COVERED THE BASICS OF PROBABILITY: ITS DEFINITION, SAMPLE SPACES AND ELEMENTARY OUTCOMES, CONDITIONAL PROBABILITY, AND SOME BASIC FORMULAS FOR COMPUTING PROBABILITIES. WE ILLUSTRATED THESE IDEAS USING A 2-DICE SAMPLE SPACE. FOR THE MODERN GAMBLER, PROBABILITY IS THE POWER TOOL OF CHOICE.



AND FINALLY, IN THE MEDICAL EXAMPLE, WE SHOWED HOW THESE ABSTRACT IDEAS COULD HELP TO MAKE GOOD DECISIONS IN THE FACE OF *IMPERFECT INFORMATION* AND *REAL RISKS*—THE *ULTIMATE GOAL* OF STATISTICS.

BUT THIS IS JUST THE BEGINNING. FOR US, PROBABILITY IS ONLY A *TOOL*—AN ESSENTIAL TOOL, TO BE SURE—IN THE STUDY OF STATISTICS. IN THE CHAPTERS THAT FOLLOW, WE'LL EXPLORE THE SUBTLE RELATIONSHIP BETWEEN PROBABILITY, VARIATIONS IN STATISTICAL DATA, AND OUR CONFIDENCE IN INTERPRETING THE MEANING OF OUR OBSERVATIONS.



♦Chapter 4♦
RANDOM VARIABLES

IN CHAPTER 2, WE SAW THAT OBSERVATIONS OF NUMERICAL DATA, LIKE STUDENTS' WEIGHTS, CAN BE GRAPHED AND SUMMARIZED IN TERMS OF MIDPOINTS, SPREADS, OUTLIERS, ETC. IN CHAPTER 3, WE SAW HOW PROBABILITIES CAN BE ASSIGNED TO THE OUTCOMES OF A RANDOM EXPERIMENT.

IF WE IMAGINE A RANDOM EXPERIMENT REPEATED MANY TIMES, WE EXPECT THAT THE ACTUAL OUTCOMES OVER TIME WILL BE GOVERNED BY THEIR PROBABILITIES. THE PROBABILITIES FORM A *MODEL* FOR REAL-LIFE EXPERIMENTS... SO WHY NOT DO FOR THE MODEL WHAT WE'VE ALREADY DONE FOR THE DATA IT DESCRIBES?

THE KEY IDEA IS THE *RANDOM VARIABLE*, WHICH WE WRITE AS A LARGE

X

A RANDOM VARIABLE IS DEFINED AS THE *NUMERICAL OUTCOME OF A RANDOM EXPERIMENT.*

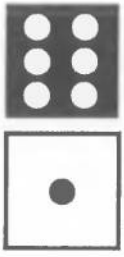
FOR EXAMPLE, IMAGINE DRAWING ONE STUDENT AT RANDOM FROM THE STUDENT BODY. THAT'S THE RANDOM EXPERIMENT. THE STUDENT'S *HEIGHT, WEIGHT, FAMILY INCOME, S.A.T. SCORE, AND GRADE POINT AVERAGE* ARE ALL NUMERICAL *VARIABLES* DESCRIBING PROPERTIES OF THE RANDOMLY SELECTED STUDENT. THEY'RE ALL *RANDOM VARIABLES.*

ANOTHER EXAMPLE: TOSS TWO COINS (THE RANDOM EXPERIMENT) AND RECORD THE *NUMBER* OF HEADS: 0, 1, OR 2.

OUTCOME	TT	HT TH	HH
x	0	1	2

NOTE THE NOTATION! THE VARIABLE IS WRITTEN WITH A CAPITAL X. THE LOWERCASE x REPRESENTS A SINGLE VALUE OF X, FOR EXAMPLE $x = 2$, IF HEADS COMES UP TWICE.

ANOTHER EXAMPLE IS BASED ON THE FAMILIAR TOSS OF TWO DICE. LET Y REPRESENT THE SUM OF THE DOTS ON THE TWO DICE. FOR THIS RANDOM VARIABLE, Y CAN BE ANY NUMBER BETWEEN 2 AND 12.



$Y = 7$

NOW WE WANT TO LOOK AT THE ***PROBABILITIES*** OF THE OUTCOMES. FOR THE PROBABILITY THAT THE RANDOM VARIABLE X HAS THE VALUE x, WE WRITE $Pr(X = x)$, OR JUST $p(x)$. FOR THE COIN-FLIPPING RANDOM VARIABLE X, WE CAN MAKE THE TABLE:

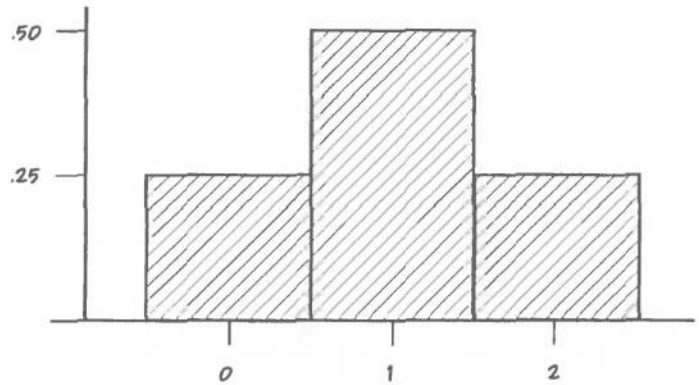
x	0	1	2
$Pr(X=x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

THIS TABLE IS CALLED THE ***PROBABILITY DISTRIBUTION*** OF THE RANDOM VARIABLE X.

FOR THE RANDOM VARIABLE Y (THE SUM OF TWO DICE), THE PROBABILITY DISTRIBUTION LOOKS LIKE THIS:

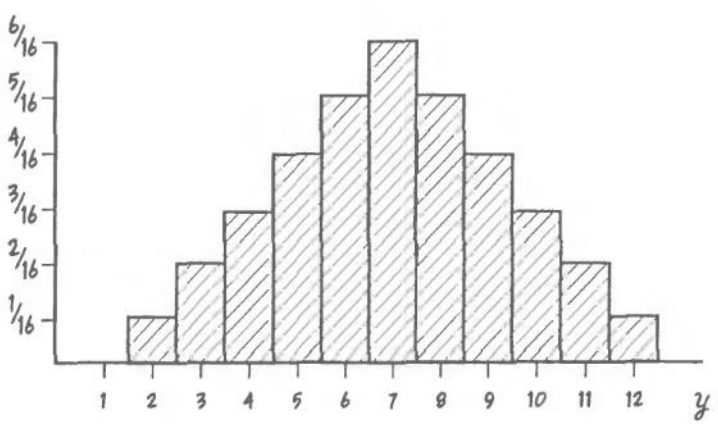
y	2	3	4	5	6	7	8	9	10	11	12
$Pr(Y=y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

NOW LET'S DRAW GRAPHS, OR ***HISTOGRAMS***, SHOWING THESE PROBABILITY DISTRIBUTIONS. FOR EACH VALUE OF X, WE DRAW A BAR EQUAL IN HEIGHT TO $p(x)$.

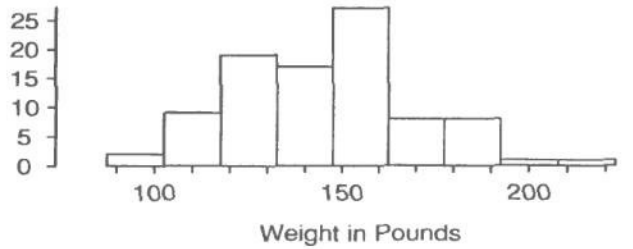


IT'S EASY TO SEE THAT THE TOTAL AREA OF THESE BOXES IS 1: EACH BOX HAS BASE 1 AND HEIGHT $p(x)$, SO THE TOTAL AREA IS THE SUM OF THE PROBABILITIES OF ALL OUTCOMES, I.E. 1.

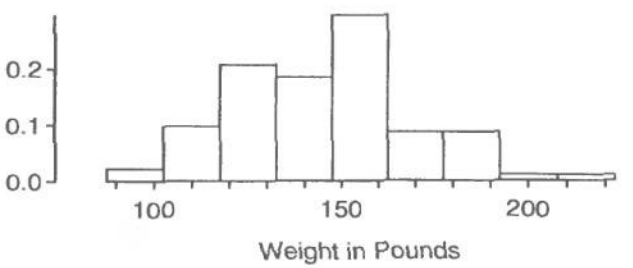
HERE'S THE PROBABILITY HISTOGRAM OF THE RANDOM VARIABLE Y, SHOWING THE PROBABILITY DISTRIBUTION OF THE SUM OF TWO DICE:



WHY DO WE CALL THESE GRAPHS HISTOGRAMS? YOU'LL RECALL THAT IN CHAPTER 2, A HISTOGRAM WAS A GRAPH THAT DISPLAYED HOW MANY DATA POINTS LAY IN EACH OF A SERIES OF INTERVALS:



FROM THIS ***FREQUENCY*** HISTOGRAM, WE DERIVED THE ***RELATIVE*** FREQUENCY HISTOGRAM, SHOWING THE ***PROPORTION*** OF DATA IN EACH INTERVAL:

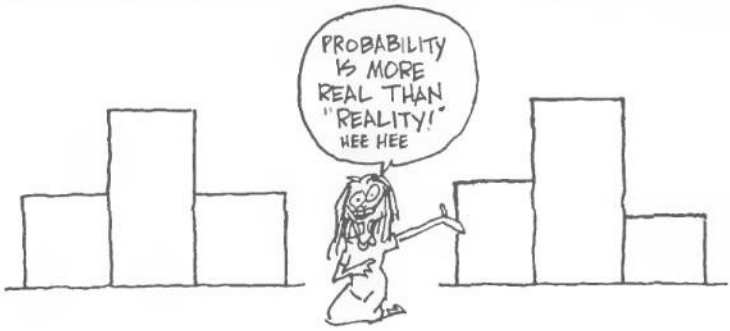


BUT YOU'LL RECALL THAT, BY ONE DEFINITION, PROBABILITY IS THE ***RELATIVE FREQUENCY OF AN EVENT "IN THE LONG RUN."*** IF WE REPEAT THE RANDOM EXPERIMENT MANY TIMES, THE ***RELATIVE FREQUENCY*** HISTOGRAM OF THE OUTCOMES SHOULD COME TO LOOK VERY MUCH LIKE THE RANDOM VARIABLE'S ***PROBABILITY*** HISTOGRAM!

WE KNOW X'S PROBABILITY DISTRIBUTION, AND WE ALSO KNOW THAT THE ACTUAL COIN FLIPS WILL MATCH THE PROBABILITIES APPROXIMATELY. AFTER 1000 TOSSES, THE MAD TOSSER TALLIES HER DATA:

PROBABILITY MODEL		OBSERVED DATA	
$p(x)$	x	n_x = NUMBER OF OCCURRENCES	$\frac{n_x}{n}$ = RELATIVE FREQUENCY
.25	0	260	.260
.5	1	517	.517
.25	2	223	.223

AND WE SEE THAT THE PROBABILITY HISTOGRAM OF X LOOKS LIKE THE "PURE FORM" OR MODEL OF THE RELATIVE FREQUENCY HISTOGRAM OF THE DATA.

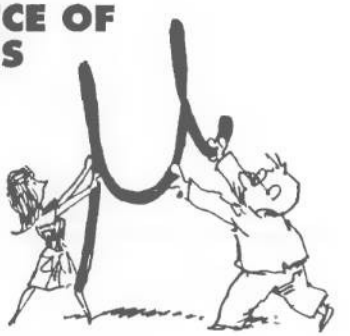


TO EXTEND THE ANALOGY BETWEEN RELATIVE FREQUENCY AND DATA, WE SHOULD NOW BE WILLING TO TALK ABOUT THE MEAN AND VARIANCE (OR STANDARD DEVIATION) OF A PROBABILITY DISTRIBUTION...

AND JUST TO REMIND OURSELVES THAT WE'RE IN THE REALM OF THE ABSTRACT, WE BREAK OUT SOME *GREEK LETTERS*...

MEAN AND VARIANCE OF RANDOM VARIABLES

WE USE SPECIAL TERMINOLOGY AND SYMBOLS TO DISTINGUISH BETWEEN THE PROPERTIES OF DATA SETS AND PROBABILITY DISTRIBUTIONS:



PROPERTIES OF DATA ARE CALLED ***SAMPLE*** PROPERTIES, WHILE PROPERTIES OF THE PROBABILITY DISTRIBUTION ARE CALLED ***MODEL*** OR ***POPULATION*** PROPERTIES. WE USE THE GREEK LETTER μ (MU) FOR THE POPULATION MEAN, AND σ (LOWERCASE SIGMA) FOR THE POPULATION STANDARD DEVIATION. (FOR DATA, WE USE THE ROMAN SYMBOLS $\bar{x}$ AND s.)

THE SAMPLE MEAN WAS DEFINED BY THE EQUATION

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

NOW SOME OF THESE DATA POINTS x_i MAY WELL HAVE EQUAL VALUES. THINK OF THE MAD COIN TOSSER: THE ONLY AVAILABLE VALUES WERE 0, 1, AND 2, AND SHE MADE 1000 TOSSES. THE VALUE 0 WAS TAKEN ON 260 TIMES, 1 HEAD CAME UP 517 TIMES, AND 2 HEADS, 223 TIMES.

AS WE LET x RANGE OVER ALL VALUES OF X, CALL n_x THE NUMBER OF DATA POINTS WITH THE VALUE x. THEN WE CAN REWRITE THAT FORMULA AS

$$\bar{x} = \frac{1}{n}\sum_{all\ x} n_x x$$

OR

$$\bar{x} = \sum_{all\ x} x\frac{n_x}{n}$$

AH! BUT NOW $\frac{n_x}{n}$ IS THE RELATIVE FREQUENCY... THE "APPROXIMATE PROBABILITY..." THE NUMBER THAT APPROACHES $p(x)$...SO, BY ANALOGY, WE FORM THE EXPRESSION

$$\sum_{all\ x} x p(x)$$

AND DEFINE THAT AS THE *MEAN OF THE PROBABILITY DISTRIBUTION.*

DEFINITION: THE **mean** OF THE RANDOM VARIABLE X IS DEFINED AS

$$\mu = \sum_{\text{all } x} x p(x)$$

THIS IS ALSO CALLED THE ***EXPECTED VALUE*** OF X, OR E[X]. THINK OF IT AS THE SUM OF THE POSSIBLE VALUES, EACH WEIGHTED BY ITS PROBABILITY.

THE MAD COIN TOSSER'S EXPERIMENT ALLOWS US TO COMPARE HER SAMPLE MEAN $\bar{x}$ WITH OUR MODEL MEAN μ:

SAMPLE

x	$\frac{n_x}{n}$	$x\frac{n_x}{n}$
0	.26	0
1	.517	.517
2	.223	.446
		$.963 = \bar{x}$

MODEL

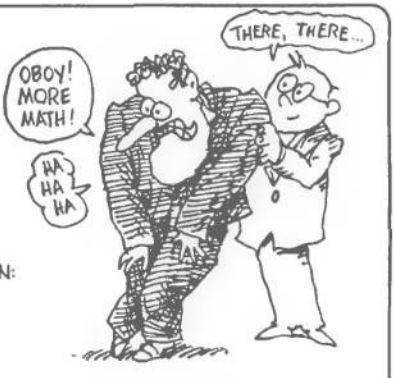
x	$p(x)$	$xp(x)$
0	.25	0
1	.5	.5
2	.25	.5
		$1 = \mu$

NOW LET'S DO THE SAME THING TO THE ***VARIANCE***. MAYBE YOU REMEMBER THE FORMULA

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

IT (ALMOST) MEASURES THE AVERAGE SQUARED DISTANCE OF DATA FROM THE MEAN. AS ABOVE THIS CAN BE REWRITTEN:

$$s^2 = \sum_{\text{all } x}(x-\bar{x})^2 \frac{n_x}{n-1}$$



EXCEPT FOR THAT ANNOYING DENOMINATOR $n-1$ INSTEAD OF n, THIS ALSO LOOKS LIKE A WEIGHTED SUM OF SQUARED DISTANCES... SO WE MAKE ANOTHER *DEFINITION:*

THE **variance** OF A RANDOM VARIABLE X IS THE EXPECTED SQUARED DISTANCE FROM THE POPULATION MEAN:

$$\sigma^2 = \sum_{all\ x} (x-\mu)^2 p(x)$$

THE **standard deviation** σ IS THE SQUARE ROOT OF THE VARIANCE.

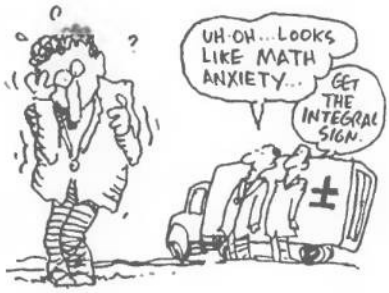
WE USE THE TABLE FROM THE LAST PAGE TO FIND THE VARIANCE OF THE TWO-COIN TOSS (FOR WHICH $\mu = 1$).

x	$p(x)$	$(x-\mu)^2 p(x)$
0	.25	$(0-1)^2 .25 = .25$
1	.5	$(1-1)^2 .50 = 0$
2	.25	$(2-1)^2 .25 = .25$
TOTAL		$.50 = \sigma^2$

TO SUM UP: μ AND σ, THE POPULATION *MEAN* AND *STANDARD DEVIATION*, ARE PROPERTIES WE CAN COMPUTE FROM *PROBABILITY DISTRIBUTIONS*. THEY ARE COMPLETELY ANALOGOUS TO THE SAMPLE MEAN $\bar{x}$ AND STANDARD DEVIATION s COMPUTED FROM SAMPLE DATA.

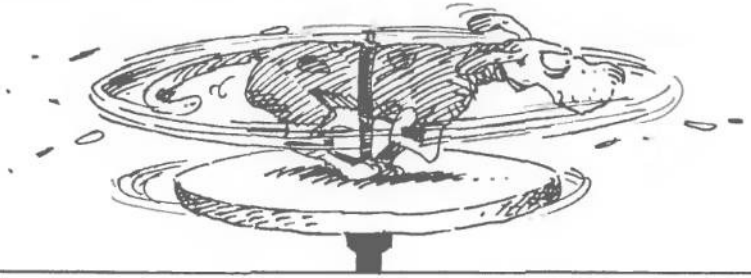
OUR EXAMPLES SO FAR HAVE BEEN *DISCRETE* RANDOM VARIABLES. THEIR OUTCOMES ARE A SET OF ISOLATED ("DISCRETE") VALUES, LIKE THOSE WE SAW IN CHAPTER 3, BUT THERE ARE ALSO

Continuous Random Variables

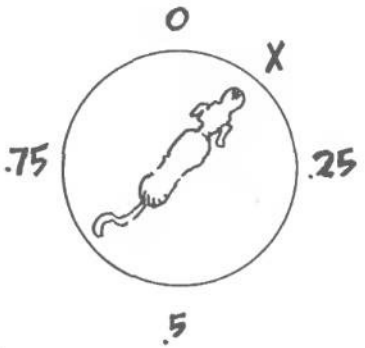


LET'S IMAGINE A RANDOM EXPERIMENT IN WHICH ***ALL OUTCOMES HAVE PROBABILITY ZERO.*** THAT'S RIGHT, $p(x) = 0$ FOR EVERY x.

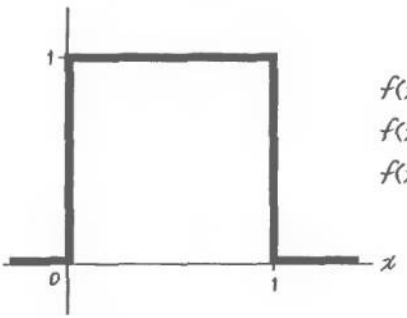
A SIMPLE EXAMPLE IS A BALANCED, SPINNING POINTER. IT CAN STOP ANYWHERE IN THE CIRCLE. IF X REPRESENTS THE PROPORTION OF THE TOTAL CIRCUMFERENCE IT LANDS ON, THE RANDOM VARIABLE X CAN TAKE ON ANY VALUE BETWEEN 0 AND 1—AN ***INFINITE*** RANGE OF VALUES.



SOME PROBABILITIES ARE EASY TO FIND, LIKE THE PROBABILITY THAT X FALLS WITHIN A RANGE: FOR EXAMPLE, $Pr(.25 \leq X \leq .75) = .5$, BECAUSE IT'S HALF THE CIRCLE. BUT WHAT ABOUT $Pr(X = .5)$? SINCE X CAN TAKE ON AN INFINITE NUMBER OF VALUES, AND ALL OF THESE VALUES ARE EQUALLY LIKELY, THE PROBABILITY THAT X IS EXACTLY .5 (OR EXACTLY ANYTHING) IS PRECISELY 0.



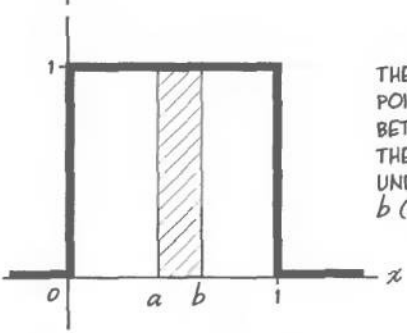
HOW CAN WE DRAW A PICTURE OF THIS? BY ANALOGY WITH THE CASE OF DISCRETE PROBABILITIES, WE TRY TO SEE CONTINUOUS PROBABILITIES AS ***AREAS UNDER SOMETHING***. FOR THE SPINNING POINTER, THE "SOMETHING" LOOKS LIKE THIS:



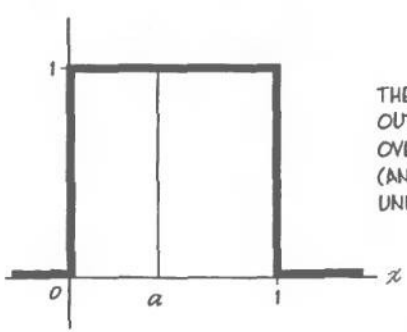
$f(x) = 0$ WHEN $x < 0$

$f(x) = 1$ WHEN $0 \leq x \leq 1$

$f(x) = 0$ WHEN $x > 1$



THE PROBABILITY THAT THE POINTER POINTS ANYWHERE BETWEEN a AND b IS PRECISELY THE AREA OF THE SHADED REGION UNDER THE CURVE BETWEEN a AND b (IN THIS CASE, $b-a$).

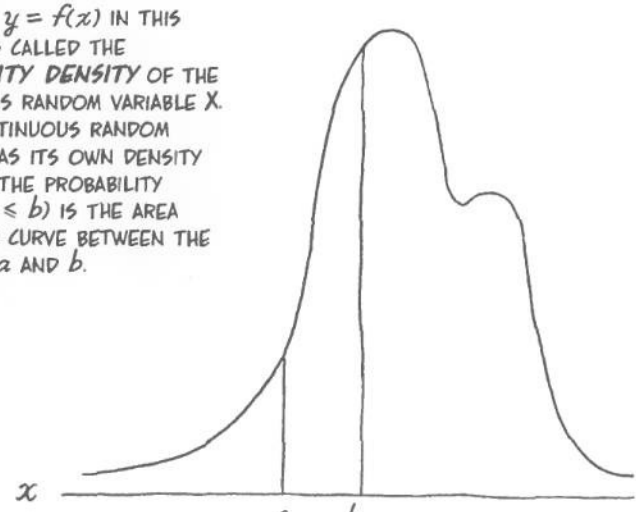


THE PROBABILITY OF AN EXACT OUTCOME, HOWEVER, IS THE "AREA" OVER A POINT, WHICH IS ***ZERO***. (AND NOTE THAT THE TOTAL AREA UNDER THE CURVE IS EXACTLY 1.)

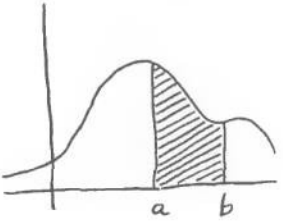
THE SAME PICTURE DESCRIBES THE *RANDOM NUMBER GENERATOR* FOUND ON MOST COMPUTERS AND SOME CALCULATORS. PRESS THE BUTTON; OUT POPS A NUMBER BETWEEN 0 AND 1; AND ALL THE NUMBERS ARE EQUALLY LIKELY, JUST AS WITH THE SPINNING POINTER.

BUT SADLY, THEY AREN'T TRULY RANDOM. THEY'RE PRODUCED BY SOME ALGORITHM, SO, TO BE ACCURATE, WE CALL THEM *PSEUDO-RANDOM* NUMBERS.

THE CURVE $y = f(x)$ IN THIS EXAMPLE IS CALLED THE ***PROBABILITY DENSITY*** OF THE CONTINUOUS RANDOM VARIABLE X. EVERY CONTINUOUS RANDOM VARIABLE HAS ITS OWN DENSITY FUNCTION. THE PROBABILITY $Pr(a \leq X \leq b)$ IS THE AREA UNDER THE CURVE BETWEEN THE x-VALUES a AND b.



IN GENERAL, THE PROBABILITY DENSITY WON'T BE SO SIMPLE, AND COMPUTING THE AREAS CAN BE FAR FROM TRIVIAL.



$$\int_a^b f(x)\,dx$$

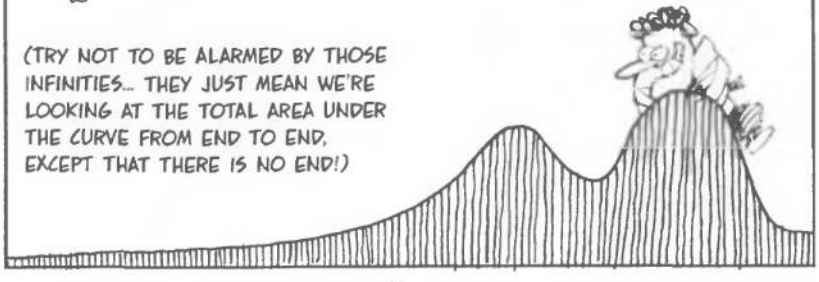
WE HAVE TO USE CALCULUS NOTATION TO DESCRIBE THE AREA UNDER THE CURVE $f(x)$. THIS SYMBOL IS READ "THE INTEGRAL OF f FROM a TO b."

LIKE DISCRETE PROBABILITIES, CONTINUOUS DENSITIES HAVE TWO FAMILIAR PROPERTIES:

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

(TRY NOT TO BE ALARMED BY THOSE INFINITIES... THEY JUST MEAN WE'RE LOOKING AT THE TOTAL AREA UNDER THE CURVE FROM END TO END, EXCEPT THAT THERE IS NO END!)



ALTHOUGH THE NOTATION MAY BE UNFAMILIAR, ALL IT MEANS IS AN *AREA*.. THE INTEGRAL SIGN ITSELF IS A STRETCHED "S," FOR SUM, WHICH THE INTEGRAL, IN SOME SENSE, IS.

AS A SUMLIKE SOMETHING, THE INTEGRAL SERVES TO DEFINE THE

MEAN AND VARIANCE of a continuous random variable.

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx$$

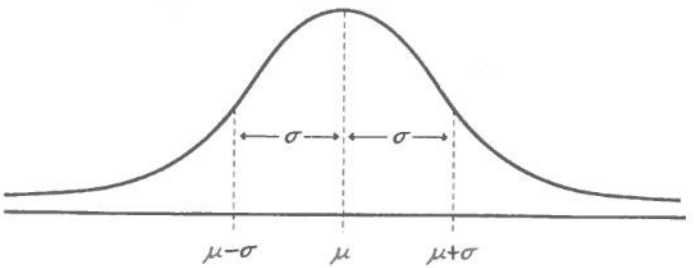
$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

BY ANALOGY WITH THE DISCRETE FORMULAS:

$$\mu = \sum_{\text{all } x} x p(x)$$

$$\sigma^2 = \sum_{\text{all } x} (x-\mu)^2 p(x)$$

ALTHOUGH IT MAY NOT BE OBVIOUS FROM THE FORMULAS, THESE DEFINITIONS OF MEAN AND VARIANCE ARE ENTIRELY CONSISTENT WITH THEIR ROLE AS CENTER AND AVERAGE SPREAD OF THE PROBABILITIES GIVEN BY THE DENSITY $f(x)$. THE PICTURE TO KEEP IN MIND IS THIS:



ADDING
random variables

ONCE YOU KNOW THE MEAN AND VARIANCE OF A RANDOM VARIABLE, WHAT CAN YOU DO WITH THEM? WELL, FOR ONE THING, YOU CAN FIND THE MEAN AND VARIANCE OF SOME *OTHER* RANDOM VARIABLES...

FOR EXAMPLE, LOOK AT A FAIR COIN TOSS. LET X = 1 IF THE COIN COMES UP HEADS AND 0 IF IT COMES UP TAILS.

x	0	1
$p(x)$	.5	.5

BY NOW, YOU SHOULD BE ABLE TO FIND THE MEAN

$$E[X] = 0 \cdot p(0) = 1 \cdot p(1)$$
$$= 0 + .5$$
$$= .5$$

AND THE VARIANCE

$$\sigma^2 = (0-.5)^2 p(0) + (1-.5)^2 p(1)$$
$$= .25$$

NOW LET'S PLAY A SIMPLE GAMBLING GAME: YOU ANTE UP $6.00 TO PLAY; I FLIP A COIN; YOU WIN $10 IF THE COIN COMES UP HEADS, ZERO IF TAILS. THEN YOUR WINNINGS W ARE

$$W = 10X - 6$$

A NEW RANDOM VARIABLE! WHAT ARE ITS MEAN AND VARIANCE?

A LITTLE THOUGHT SHOULD CONVINCE YOU THAT $E[W]$ IS GIVEN BY

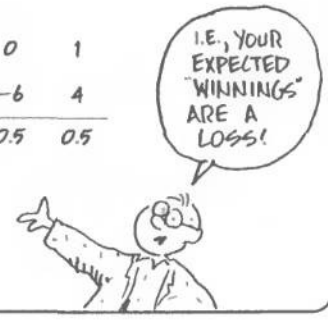
$$E[W] = E[10X - 6]$$
$$= 10E[X] - 6$$

WHICH WORKS OUT TO

$$10(0.5) - 6 = -1$$

YOU CAN CHECK IT USING THIS TABLE:

x	0	1
w	-6	4
$p(w)$	0.5	0.5

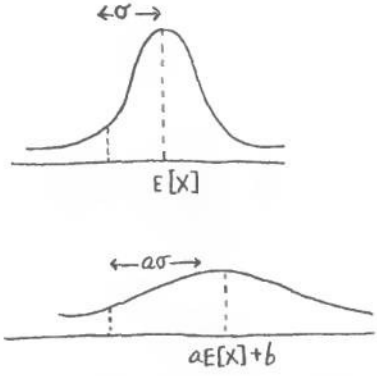


IN GENERAL, IT IS NOT HARD TO SHOW THAT

$$E[aX+b] = aE[X]+b$$

WHEN a AND b ARE ANY NUMBERS AND X IS ANY RANDOM VARIABLE. FOR THE VARIANCE, THERE'S ALSO A GENERAL RESULT:

$$\sigma^2(aX+b) = a^2\sigma^2(X)$$



IN THE GAMBLING GAME ABOVE, THE POSSIBLE OUTCOMES ARE -6 AND 4, SO IT'S CLEAR THAT THE VARIANCE OF W MUST BE GREATER THAN THE VARIANCE OF X. IN FACT,

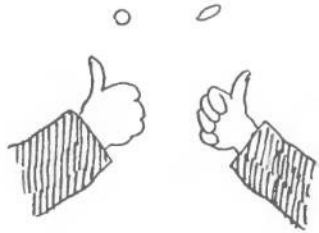
$$\sigma^2(W) = \sigma^2(10X+6)$$
$$= 100\sigma^2(X)$$
$$= 25$$

AND

$$\sigma(W) = 5$$

YOU CAN ALSO *ADD* TWO RANDOM VARIABLES TOGETHER. FOR INSTANCE, SUPPOSE WE TOSS A COIN *TWICE*. THE NUMBER OF HEADS ON BOTH TOSSES IS X_1+X_2, WHERE X_1 AND X_2 ARE THE RANDOM VARIABLES GIVING THE RESULTS OF THE FIRST AND SECOND TOSSES.

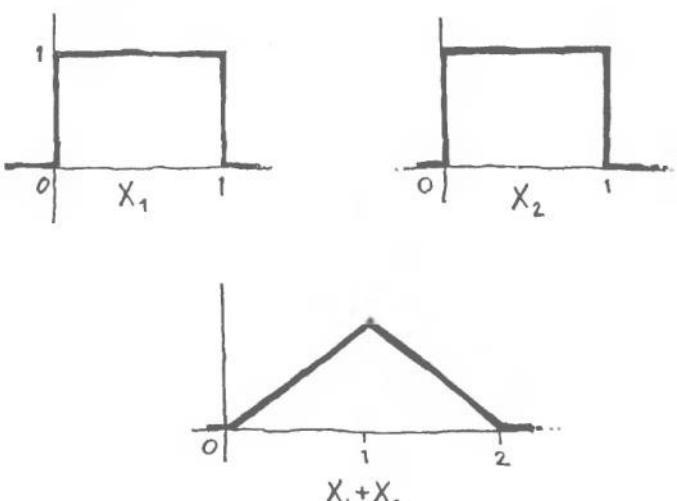
x_1+x_2	0	1	2
$p(x_1+x_2)$	.25	.5	.25



AGAIN, IT'S EASY TO SEE THAT

$$E[X_1+X_2] = E[X_1]+E[X_2]$$

(DON'T ASK ABOUT THE PROBABILITY DISTRIBUTION OF X_1+X_2, BECAUSE IT DEPENDS IN A COMPLICATED WAY ON THE TWO ORIGINAL DISTRIBUTIONS. FOR EXAMPLE, IF X_1 AND X_2 ARE BOTH THE SPINNING POINTER DISTRIBUTION, THE HISTOGRAMS ACT LIKE THIS:)



THE VARIANCE OF THE SUM OF RANDOM VARIABLES HAS A SIMPLE FORM IN THE SPECIAL CASE WHEN THE VARIABLES X AND Y ARE ***INDEPENDENT***. THE TECHNICAL DEFINITION OF INDEPENDENCE IS BASED ON THE PROBABILITY PROPERTY P(A AND B) = P(A)P(B)... BUT FOR US, INDEPENDENCE JUST MEANS THAT X AND Y ARE GENERATED BY ***INDEPENDENT MECHANISMS***, SUCH AS FLIPS OF A COIN, ROLLS OF A DIE, ETC.

WHEN X AND Y ARE INDEPENDENT, ***THEIR VARIANCES ADD:***

$$\sigma^2(X+Y) = \sigma^2(X)+\sigma^2(Y)$$

IN THE CASE OF TWO COIN TOSSES,

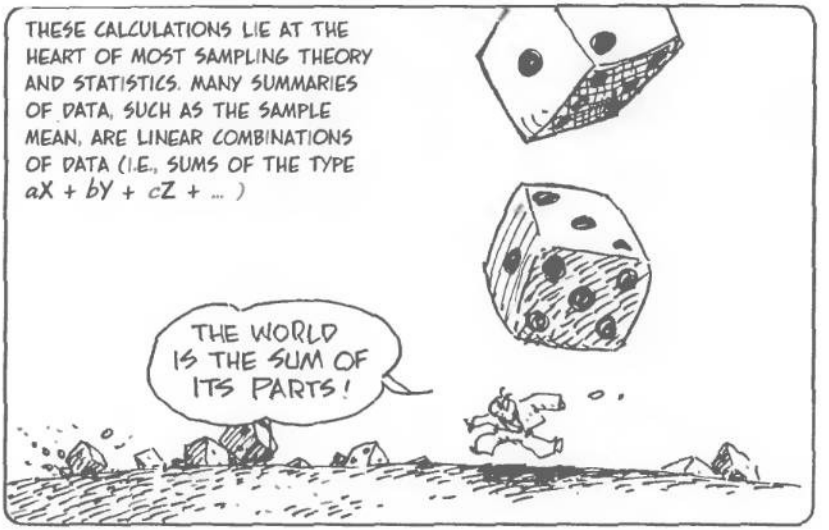
$$\begin{aligned}\sigma^2(X_1+X_2) &= \sigma^2(X_1)+\sigma^2(X_2)\\ &= .25+.25\\ &= .5\end{aligned}$$

ALL OF THIS CAN BE GENERALIZED TO THE SUM OF MANY RANDOM VARIABLES:

$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i]$$

AND, WHEN THE X_i ARE ALL INDEPENDENT,

$$\sigma^2(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \sigma^2(X_i)$$

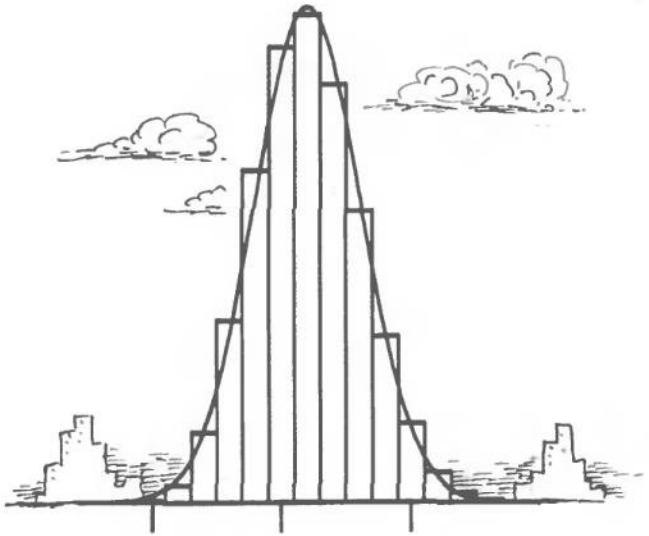


IN THE NEXT CHAPTER, WE WILL SEE TWO IMPORTANT EXAMPLES OF RANDOM VARIABLES: ONE, THE ***BINOMIAL***, IS THE SUM OF MANY REPEATED INDEPENDENT RANDOM VARIABLES. THE OTHER, THE ***NORMAL***, IS A CONTINUOUS RANDOM VARIABLE THAT HAS A SURPRISING RELATIONSHIP TO THE BINOMIAL, AND ANY OTHER SUM OF INDEPENDENT RANDOM VARIABLES AS WELL.

♦Chapter 5♦

A TALE OF TWO DISTRIBUTIONS

NOW WE LOOK AT TWO IMPORTANT EXAMPLES OF RANDOM VARIABLES, ONE DISCRETE AND ONE CONTINUOUS.



WE BEGIN WITH THE DISCRETE ONE, CALLED THE ***BINOMIAL*** RANDOM VARIABLE. SUPPOSE WE HAVE A RANDOM PROCESS WITH JUST ***TWO POSSIBLE OUTCOMES:*** A HEADS-OR-TAILS COIN TOSS, A WIN-OR-LOSE FOOTBALL GAME, A PASS-OR-FAIL AUTOMOTIVE SMOG INSPECTION. WE ARBITRARILY CALL ONE OF THESE OUTCOMES A ***SUCCESS*** AND THE OTHER A ***FAILURE.***

WHAT WE DO IS TO REPEAT THIS EXPERIMENT... WELL, REPEATEDLY. SUCH A REPEATABLE EXPERIMENT IS CALLED A

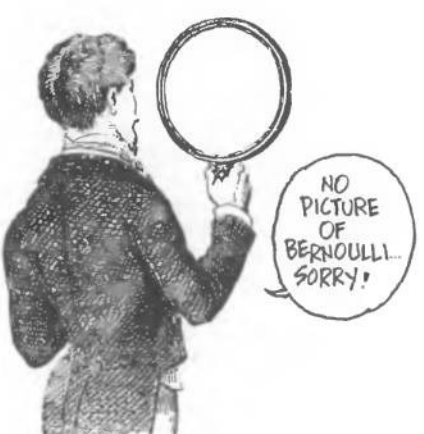
Bernoulli trial,

PROVIDED IT HAS THESE CRITICAL PROPERTIES:

1) THE RESULT OF EACH TRIAL MAY BE EITHER A SUCCESS OR A FAILURE

2) THE PROBABILITY p OF SUCCESS IS THE SAME IN EVERY TRIAL.

3) THE TRIALS ARE ***INDEPENDENT:*** THE OUTCOME OF ONE TRIAL HAS NO INFLUENCE ON LATER OUTCOMES.



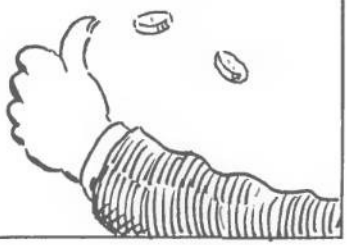
STARTING WITH A BERNOULLI TRIAL, WITH PROBABILITY OF SUCCESS p, LET'S BUILD A NEW RANDOM VARIABLE BY REPEATING THE BERNOULLI TRIAL.

The binomial random variable

X IS THE *NUMBER OF SUCCESSES* IN n REPEATED BERNOULLI TRIALS WITH PROBABILITY p OF SUCCESS.

AN EXAMPLE OF A BINOMIAL RANDOM VARIABLE IS THE NUMBER OF HEADS (SUCCESSES) IN TWO FLIPS OF A COIN. HERE $n=2$ AND $p=.5$

k= NUMBER OF SUCCESSES	0	1	2
$Pr(X=k)$	.25	.5	.25

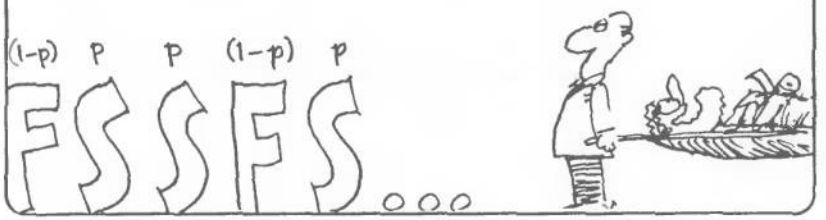


ANOTHER EXAMPLE IS DE MERE'S FIRST GAMBLE: TOSSING A SINGLE DIE FOUR TIMES IN A ROW. SUCCESS MEANS ROLLING A 6. THE DISTRIBUTION IS:

IN GENERAL, WHAT'S THE PROBABILITY DISTRIBUTION OF THE BINOMIAL FOR *ANY* PROBABILITY p AND NUMBER OF TRIALS n? A PROBABILITY CALCULATION GIVES THE ANSWER: THE PROBABILITY OF OBTAINING k SUCCESSES IN n TRIALS, $Pr(X=k)$, IS

$$Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

HERE $\binom{n}{k}$, READ "n CHOOSE k," IS THE ***BINOMIAL COEFFICIENT.*** IT COUNTS ALL POSSIBLE WAYS OF GETTING k SUCCESSES IN n TRIALS. EACH INDIVIDUAL SEQUENCE OF k SUCCESSES AND $n-k$ FAILURES HAS PROBABILITY $p^k(1-p)^{n-k}$, BY THE MULTIPLICATION RULE. THERE ARE $\binom{n}{k}$ OF THESE SEQUENCES.



THE FORMULA FOR $\binom{n}{k}$ IS

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

WHERE

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 1$$

AND $0!$ IS TAKEN TO BE 1. FOR INSTANCE, $\binom{4}{2}$, THE NUMBER OF POSSIBLE WAYS TO CHOOSE TWO LETTERS FROM A SET OF FOUR LETTERS, IS

$$\binom{4}{2} = \frac{4!}{2!2!} = \frac{24}{4} = 6$$

ANOTHER VIEW OF THE BINOMIAL COEFFICIENTS IS IN *PASCAL'S TRIANGLE*. EACH ENTRY IS THE SUM OF THE TWO NUMBERS JUST ABOVE IT.

$\binom{3}{1}$ $\binom{7}{2}$ $\binom{5}{3}$ $\binom{10}{7}$ $\binom{8}{0}$ $\binom{10}{10}$

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1
1 10 45 120 210 252 210 120 45 10 1
1 11 55 165 330 462 462 330 165 55 11 1
1 12 66 220 495 792 924 792 495 220 66 12 1

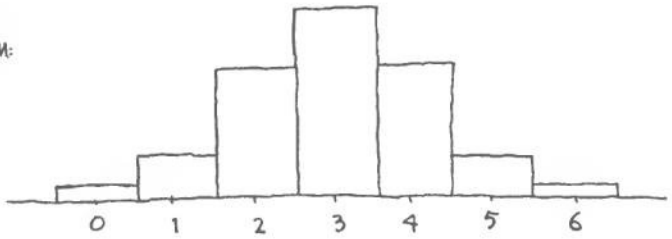
ETC.

TO FIND $\binom{n}{k}$, JUST COUNT DOWN TO ROW n AND OVER TO ENTRY k (REMEMBERING ALWAYS TO START COUNTING FROM ZERO).

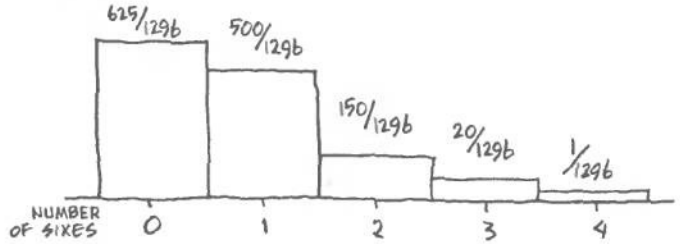
WHEN $p = .5$, THE BINOMIAL'S PROBABILITY DISTRIBUTION IS PERFECTLY SYMMETRICAL. FOR 6 COIN FLIPS, FOR INSTANCE, IT'S

k = #HEADS	0	1	2	3	4	5	6
$Pr(X=k)$	$\left(\frac{1}{2}\right)^6$	$\left(\frac{1}{2}\right)^6 \cdot 6$	$\left(\frac{1}{2}\right)^6 \cdot 15$	$\left(\frac{1}{2}\right)^6 \cdot 20$	$\left(\frac{1}{2}\right)^6 \cdot 15$	$\left(\frac{1}{2}\right)^6 \cdot 6$	$\left(\frac{1}{2}\right)^6$

WITH THIS HISTOGRAM:



FOR DE MERE'S ROLL OF FOUR DICE, THE DISTRIBUTION IS MORE LOPSIDED:



THE *MEAN* AND *VARIANCE* OF THE BINOMIAL DISTRIBUTION ARE

$$\mu = np$$

$$\sigma^2 = np(1-p)$$

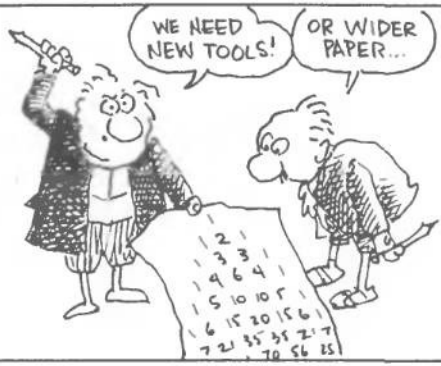
NOTE THAT THE MEAN MAKES INTUITIVE SENSE: IN n BERNOULLI TRIALS, THE EXPECTED NUMBER OF SUCCESSES SHOULD BE np. THE VARIANCE FOLLOWS FROM THE FACT THAT THE BINOMIAL IS THE SUM OF n INDEPENDENT BERNOULLI TRIALS OF VARIANCE $p(1-p)$.

THE PARAMETERS OF THE BINOMIAL DISTRIBUTION ARE n AND p. THE DISTRIBUTION, MEAN, AND VARIANCE DEPEND *ONLY* ON THESE TWO NUMBERS. TABLES OF THE BINOMIAL DISTRIBUTION APPEAR IN MOST TEXTBOOKS AND COMPUTER PROGRAMS. HERE IS A TABLE FOR $n=10$.

VALUES OF $Pr(X=k)$

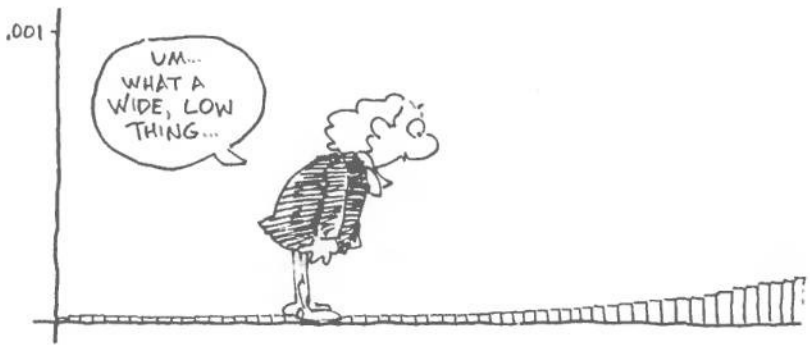
		k										
		0	1	2	3	4	5	6	7	8	9	10
p	.1	0.349	0.387	0.194	0.057	0.011	0.001	0.000	0.000	0.000	0.000	0.000
	.25	0.056	0.188	0.282	0.250	0.146	0.058	0.016	0.003	0.000	0.000	0.000
	.50	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044	0.010	0.001
	.75	0.000	0.000	0.000	0.003	0.016	0.058	0.146	0.250	0.282	0.188	0.056
	.9	0.000	0.000	0.000	0.000	0.000	0.001	0.011	0.057	0.194	0.387	0.349

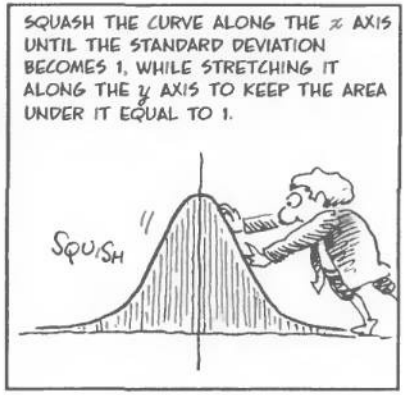
BUT CALCULATING THESE THINGS FOR LARGE VALUES OF n CAN BE A PAIN... OR AT LEAST, IT WAS BACK IN THE 18TH CENTURY, WHEN *JAMES BERNOULLI* AND *ABRAHAM DE MOIVRE* WERE TRYING TO DO IT WITHOUT A COMPUTER.



DEPLOYING A NEWLY INVENTED WEAPON, THE *CALCULUS*, DE MOIVRE SHOWED THAN WHEN $p = .5$, THE BINOMIAL DISTRIBUTION WAS CLOSELY APPROXIMATED BY A *CONTINUOUS DENSITY FUNCTION* WHICH COULD BE DESCRIBED VERY SIMPLY.

TO SEE HOW THIS WORKS, IMAGINE THE BINOMIAL DISTRIBUTION WITH $p = .5$ AND n VERY LARGE—A MILLION, SAY...



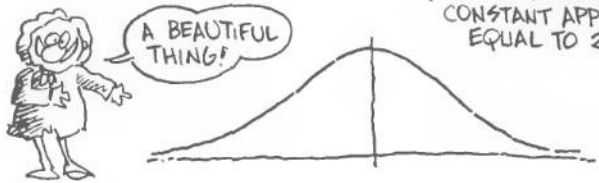


THE RESULT IS VERY CLOSE TO A *SMOOTH, SYMMETRICAL, BELL-SHAPED CURVE*, WHICH DEMOIVRE SHOWED WAS GIVEN BY THE SIMPLE FORMULA:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

THIS FUNCTION IS CALLED THE **standard normal distribution.**

(*e* IS A USEFUL MATHEMATICAL CONSTANT APPROXIMATELY EQUAL TO 2.718.)



(CONVINCE YOURSELF THAT THIS FUNCTION REALLY HAS A BELL-SHAPED GRAPH. FOR z FAR FROM ZERO, $f(z)$ IS VERY NEARLY ZERO—IT HAS A BIG DENOMINATOR; IT'S SYMMETRICAL, SINCE $f(z) = f(-z)$, AND IT HAS A MAXIMUM AT $z = 0$.)

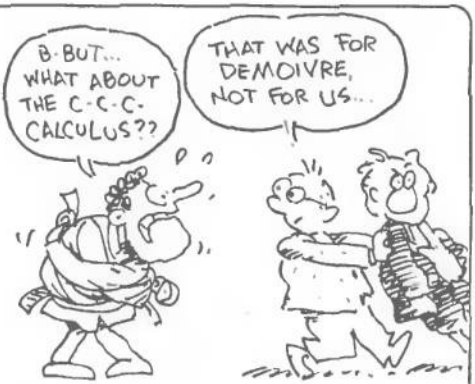
THE DISTRIBUTION IS CALLED THE *STANDARD* NORMAL BECAUSE ALL THAT SQUASHING AND STRETCHING WAS SPECIALLY ARRANGED TO GIVE IT THESE SIMPLE PROPERTIES, WHICH WE PRESENT WITHOUT PROOF:

$$\mu = 0$$

$$\sigma = 1$$

TO SUMMARIZE DE MOIVRE, IF YOU *"NORMALIZE" THE BINOMIAL* DISTRIBUTION WITH $p = 1/2$—I.E., CENTER IT ON ZERO AND MAKE ITS STANDARD DEVIATION = 1, THEN IT CLOSELY FITS THE *STANDARD NORMAL DISTRIBUTION*

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

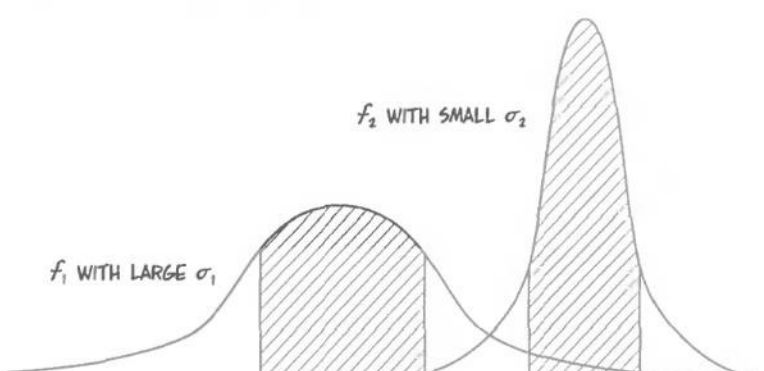


OTHER NORMALS, WITH DIFFERENT MEANS AND VARIANCES, ARE OBTAINED BY STRETCHING AND SLIDING THE STANDARD NORMAL. IN GENERAL, WE WRITE THE FORMULA

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

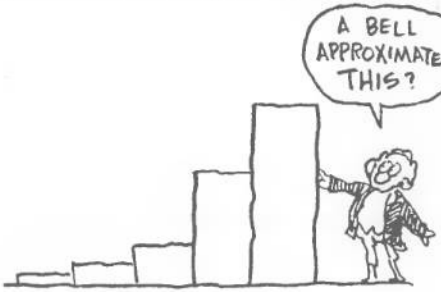
THIS GIVES A SYMMETRIC, BELL-SHAPED DISTRIBUTION CENTERED ON THE MEAN μ WITH THE STANDARD DEVIATION σ.

HERE ARE TWO DIFFERENT NORMALS WITH THE REGIONS WITHIN THEIR STANDARD DEVIATIONS SHADED.



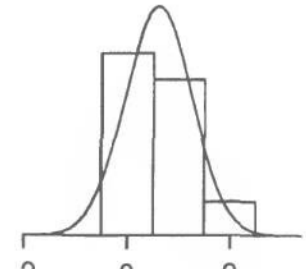
DE MOIVRE PROVED THAT THE STANDARD NORMAL FITS THE (NORMALIZED) BINOMIAL WITH $p = .5$, BUT, IN FACT, IT WORKS FOR *ANY VALUE OF* p.

GENERALLY: FOR ANY VALUE OF p, THE BINOMIAL DISTRIBUTION OF n TRIALS WITH PROBABILITY p IS APPROXIMATED BY THE NORMAL CURVE WITH $\mu = np$ AND $\sigma = np(1-p)$.

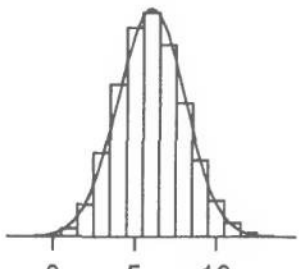


THIS IS ACTUALLY A LITTLE STRANGE. ALL NORMALS ARE SYMMETRICAL AND BELL SHAPED... BUT, AS WE SAW, BINOMIAL DISTRIBUTIONS ARE *NOT SYMMETRICAL* WHEN $p \neq .5$.

BUT IT TURNS OUT THAT AS n GETS LARGE, THE BINOMIAL'S ASYMMETRY IS OVERWHELMED, AS YOU SEE IN THIS EXAMPLE:



Binomial: n = 2 and p = 0.3



Binomial: n = 20 and p = 0.3

IN FACT, DEMOIVRE'S DISCOVERY ABOUT THE BINOMIAL IS A SPECIAL CASE OF AN EVEN *MORE* GENERAL RESULT, WHICH HELPS EXPLAIN WHY THE NORMAL IS SO IMPORTANT AND WIDESPREAD IN NATURE. IT IS THIS:

"Fuzzy Central Limit Theorem":

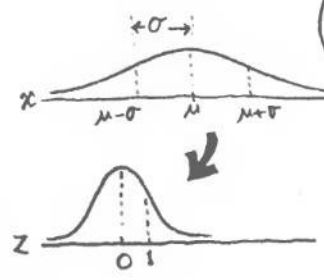
DATA THAT ARE INFLUENCED BY *MANY SMALL AND UNRELATED RANDOM EFFECTS* ARE APPROXIMATELY *NORMALLY DISTRIBUTED.*

THIS EXPLAINS WHY THE NORMAL IS *EVERYWHERE:* STOCK MARKET FLUCTUATIONS, STUDENT WEIGHTS, YEARLY TEMPERATURE AVERAGES, S.A.T. SCORES: ALL ARE THE RESULT OF MANY DIFFERENT EFFECTS. FOR EXAMPLE, A STUDENT'S WEIGHT IS THE RESULT OF GENETICS, NUTRITION, ILLNESS, AND LAST NIGHT'S BEER PARTY. WHEN YOU PUT THEM ALL TOGETHER, YOU GET THE NORMAL! (REMEMBER, THE BINOMIAL IS THE RESULT OF n INDEPENDENT BERNOULLI TRIALS.)

THE z TRANSFORMATION

$$z = \frac{x-\mu}{\sigma}$$

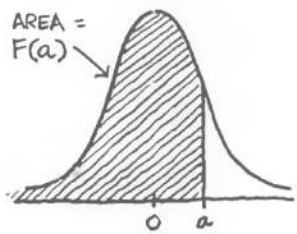
CHANGES A NORMAL RANDOM VARIABLE WITH MEAN μ AND STANDARD DEVIATION σ INTO A *STANDARD* NORMAL RANDOM VARIABLE WITH MEAN 0 AND STANDARD DEVIATION 1.



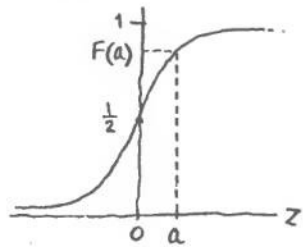
THEN ALL WE NEED TO FIND PROBABILITIES FOR *ANY* NORMAL DISTRIBUTION IS THE SINGLE TABLE FOR THE *STANDARD NORMAL* $F(z)$.

z	-2.5	-2.4	-2.3	-2.2	-2.1	-2.0	-1.9	-1.8	-1.7	-1.6
F(z)	0.006	0.008	0.011	0.014	0.018	0.023	0.029	0.036	0.045	0.055
z	-1.5	-1.4	-1.3	-1.2	-1.1	-1.0	-0.9	-0.8	-0.7	-0.6
F(z)	0.067	0.081	0.097	0.115	0.136	0.159	0.184	0.212	0.242	0.274
z	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4
F(z)	0.309	0.345	0.382	0.421	0.460	0.500	0.540	0.579	0.618	0.655
z	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4
F(z)	0.691	0.726	0.758	0.788	0.816	0.841	0.864	0.885	0.903	0.919
z	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4
F(z)	0.933	0.945	0.955	0.964	0.971	0.977	0.982	0.986	0.989	0.992
z	2.5									
F(z)	0.994									

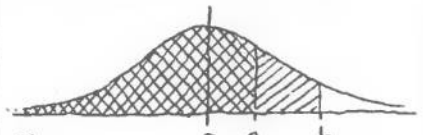
HERE $F(a) = Pr(z \leq a)$, THE AREA UNDER THE DENSITY CURVE TO THE LEFT OF $z = a$.



(WE CAN ALSO GRAPH THE CURVE $y = F(z)$, THE *CUMULATIVE PROBABILITY*. IT LOOKS LIKE THIS.)



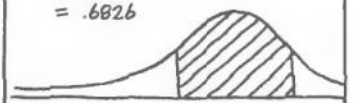
THE TABLE ALLOWS US TO FIND THE PROBABILITY OF z BEING IN ANY INTERVAL $a \leq z \leq b$. IT IS JUST THE DIFFERENCE BETWEEN THE AREAS $F(b)$ AND $F(a)$.



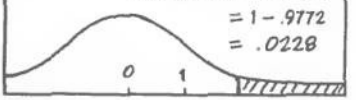
$$Pr(a \leq z \leq b) = F(b) - F(a)$$

SO, FOR EXAMPLE,

$$Pr(-1 < z < 1) = F(1) - F(-1)$$
$$= .8413 - .1587$$
$$= .6826$$

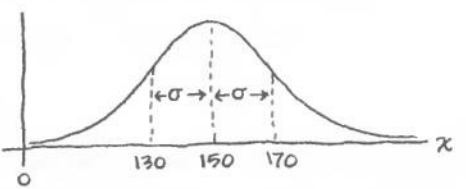


$$Pr(z \geq 2) = 1 - F(2)$$
$$= 1 - .9772$$
$$= .0228$$



USING THE SUBSTITUTION $z = \frac{x-\mu}{\sigma}$, WE CAN USE THE SAME TABLE TO FIND PROBABILITIES FOR OTHER NORMAL DISTRIBUTIONS.

FOR EXAMPLE, SUPPOSE STUDENT WEIGHTS ARE NORMALLY DISTRIBUTED WITH A MEAN $\mu = 150$ POUNDS AND STANDARD DEVIATION $\sigma = 20$:

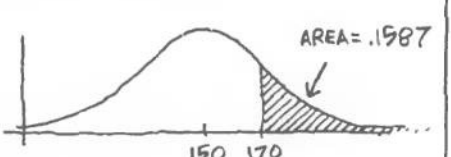


THEN WHAT'S THE PROBABILITY OF WEIGHING MORE THAN 170 POUNDS?

NOW IT'S "JUST" ALGEBRA.

$$Pr(X > 170) =$$
$$Pr\left(\frac{X-\mu}{\sigma} > \frac{170-150}{20}\right) =$$
$$Pr\left(z > \frac{20}{20}\right) =$$
$$Pr(z > 1)$$

THAT'S $1 - F(1)$, WHICH WE CAN READ FROM THE TABLE AS $1 - .8413 = .1587$

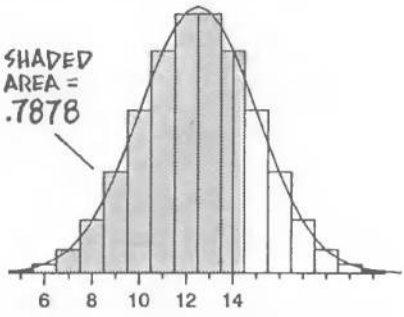


A LITTLE LESS THAN ONE STUDENT IN SIX TIPS THE SCALES ABOVE 170 POUNDS.

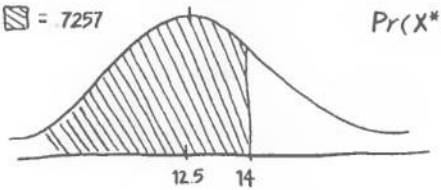
THE GENERAL RULE FOR COMPUTING NORMAL PROBABILITIES IS THEREFORE:

$$Pr(a \leq X \leq b) = F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right)$$

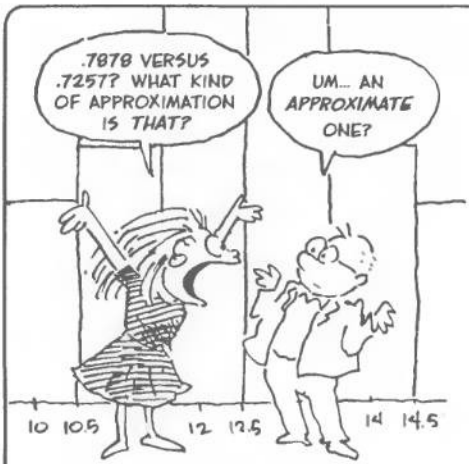
NOW BACK TO DE MOIVRE AND HIS BINOMIAL APPROXIMATION... LET'S LOOK AT A BINOMIAL DISTRIBUTION WITH $n = 25$ TRIALS AND $p = .5$ (25 COIN FLIPS, SAY). WE CAN COMPUTE (OR LOOK UP IN A TABLE) ANY PROBABILITY, FOR EXAMPLE, $Pr(X \leq 14)$. IT IS **.7878** EXACTLY.



NOW CALCULATE A NORMAL RANDOM VARIABLE X^* WITH THE SAME MEAN $\mu = np = (25)(.5) = 12.5$ AND STANDARD DEVIATION $\sigma = np(1-p) = 2.5$.



$$Pr(X^* \leq 14) = Pr\left(Z \leq \frac{14-12.5}{2.5}\right)$$
$$= Pr(Z \leq .6)$$
$$= \mathbf{.7257}$$

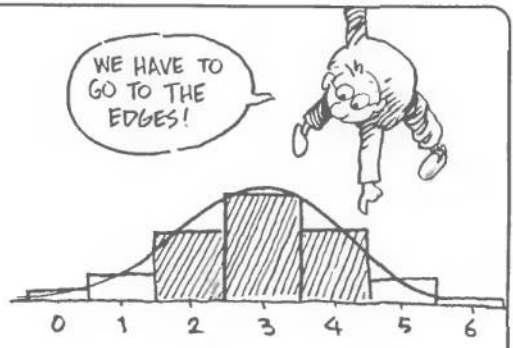


AH, BUT WE CAN DO BETTER! IF YOU LOOK CLOSELY AT THE FIRST HISTOGRAM, YOU SEE THE BARS ARE *CENTERED* ON THE NUMBERS. THIS MEANS $Pr(X^* \leq 14)$ IS ACTUALLY THE AREA UNDER THE BARS LESS THAN $x = \mathbf{14.5}$. WE NEED TO ACCOUNT FOR THAT EXTRA .5, AND IN FACT,

$$Pr(X^* \leq 14.5) = Pr(Z \leq .8)$$
$$= \mathbf{.7881}$$

A VERY GOOD APPROXIMATION TO .7878 INDEED!

THAT LITTLE EXTRA .5 WE ADDED IS CALLED THE **continuity correction.** WE HAVE TO INCLUDE IT TO GET A GOOD CONTINUOUS APPROXIMATION TO OUR DISCRETE BINOMIAL RANDOM VARIABLE X. IT'S SUMMARIZED BY THIS ONE HIDEOUS FORMULA:

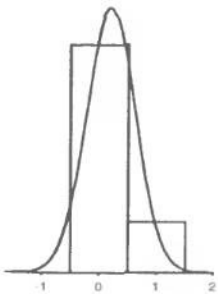


$$Pr(a \le X \le b) \simeq Pr\left(\frac{a-\frac{1}{2}-np}{\sqrt{np(1-p)}} \le Z \le \frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

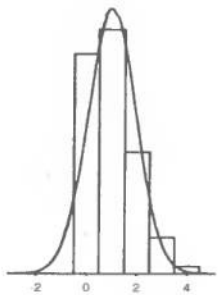
WHEN IS THIS APPROXIMATION "GOOD ENOUGH?" FOR STATISTICIANS, THE RULE OF THUMB IS: WHENEVER n IS BIG ENOUGH TO MAKE THE NUMBER OF EXPECTED SUCCESSES AND FAILURES BOTH GREATER THAN *FIVE*:

$$np \ge 5 \quad and \quad n(1-p) \ge 5$$

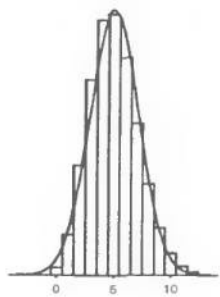
YOU CAN SEE FROM THESE HISTOGRAMS THAT THE FIT WHEN $p = 0.1$ IS MEDIOCRE OR WORSE UNTIL n REACHES 50, MAKING $np = 5$.



$n=2$, $p=0.1$



$n=10$, $p=0.1$

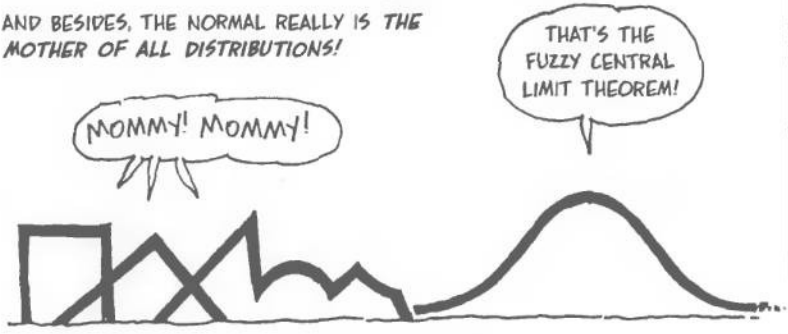


$n=50$, $p=0.1$

WHAT'S SO GREAT ABOUT THIS NORMAL APPROXIMATION? THE BINOMIAL DISTRIBUTION OCCURS COMMONLY IN NATURE, AND IT ISN'T HARD TO UNDERSTAND, BUT IT CAN BE TIRESOME TO CALCULATE.

THE NORMAL WHICH APPROXIMATES IT MAY BE LESS INTUITIVE, BUT IT'S VERY EASY TO USE. THE z-TRANSFORM CONVERTS *ANY* NORMAL TO THE *STANDARD NORMAL*, ALLOWING US TO READ PROBABILITIES STRAIGHT OUT OF A SINGLE NUMERICAL TABLE.

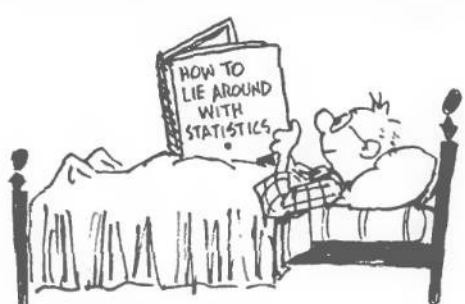
AND BESIDES, THE NORMAL REALLY IS *THE MOTHER OF ALL DISTRIBUTIONS!*



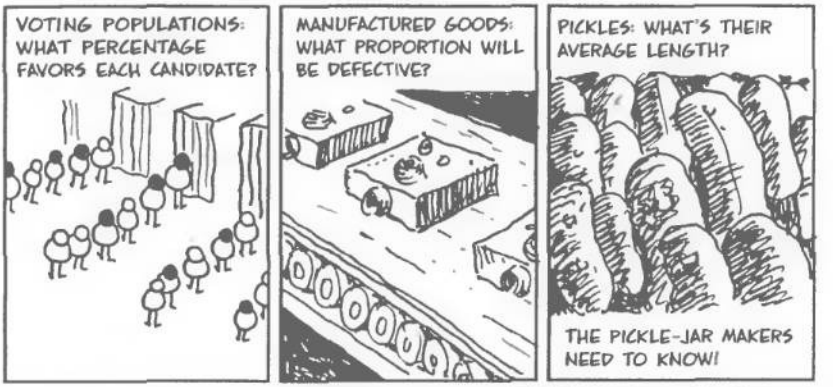
♦Chapter 6♦
SAMPLING

BY NOW, AFTER A STEADY DIET OF COINS, DICE, AND ABSTRACT IDEAS, YOU MAY BE WONDERING WHAT ALL THIS STATISTICAL EQUIPMENT WE'VE BEEN BUILDING HAS TO DO WITH THE *REAL WORLD.* WELL, NOW WE'RE FINALLY GOING TO FIND OUT...

IN THIS CHAPTER, WE BEGIN LOOKING AT THE *REAL* BUSINESS OF STATISTICS, WHICH IS, AFTER ALL, TO SAVE PEOPLE *TIME* AND *MONEY.* PEOPLE HATE TO WASTE TIME DOING *UNNECESSARY WORK,* AND ONE THING STATISTICS CAN DO IS TELL US EXACTLY HOW LAZY WE CAN AFFORD TO BE.

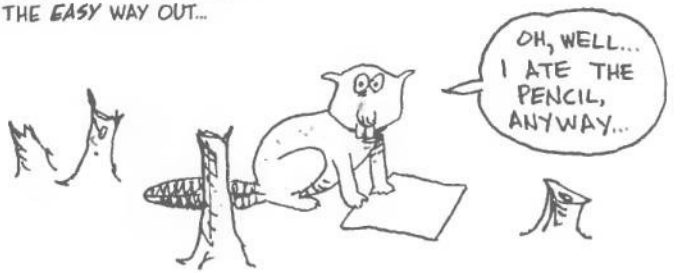


THE PROBLEM WITH THE WORLD IS THAT THE COLLECTIONS OF STUFF IN IT ARE SO LARGE, IT'S HARD TO GET THE INFORMATION WE WANT:



THE INDUSTRIOUS, HARD-WORKING, SIMPLE-MINDED BEAVERLIKE WAY TO ANSWER THESE QUESTIONS WOULD BE TO MEASURE EVERY SINGLE PICKLE IN THE WORLD (SAY) AND DO SOME ARITHMETIC.

BUT WE AREN'T BEAVERS—WE'RE ***STATISTICIANS!*** WE'RE LOOKING FOR THE ***EASY*** WAY OUT...



OUR METHOD IS TO TAKE A ***SAMPLE...*** A RELATIVELY SMALL SUBSET OF THE TOTAL POPULATION, THE WAY POLLSTERS DO AT ELECTION TIME.

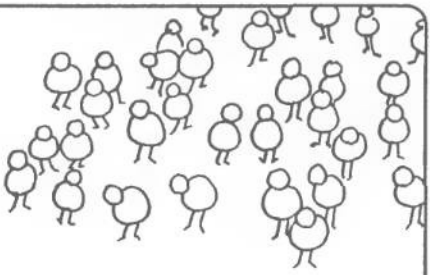
AN OBVIOUS QUESTION IS: HOW BIG A SAMPLE DO WE HAVE TO TAKE TO GET MEANINGFUL RESULTS?

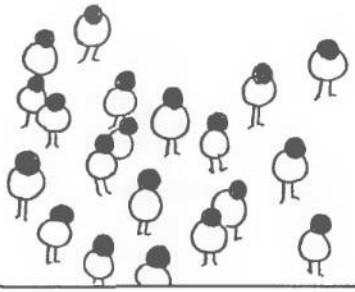
AND THE ANSWER, WHICH YOU SHOULD INSCRIBE IN YOUR BRAIN FOREVERMORE, WILL TURN OUT TO BE: IF n IS THE NUMBER OF ITEMS IN THE SAMPLE, THEN EVERYTHING IS GOVERNED BY

$$\frac{1}{\sqrt{n}}.$$

SAMPLING DESIGN

BEFORE DOING THE NUMBERS, WE SHOULD POINT OUT THAT THE *QUALITY* OF THE SAMPLE IS AS IMPORTANT AS ITS *SIZE*. HOW DO WE ASSURE OURSELVES THAT WE'RE CHOOSING A REPRESENTATIVE SAMPLE?





THE *SELECTION PROCESS ITSELF* IS CRITICAL. FOR EXAMPLE, A VOTER SURVEY THAT SYSTEMATICALLY EXCLUDED BLACK PEOPLE WOULD BE WORTHLESS, AND THERE ARE A HOST OF OTHER WAYS TO RUIN, OR BIAS, A SAMPLE.

NOT TO PROLONG THE MYSTERY, THE WAY TO GET STATISTICALLY DEPENDABLE RESULTS IS TO CHOOSE THE SAMPLE AT **random.**

THE **SIMPLE** RANDOM SAMPLE

SUPPOSE WE HAVE A LARGE POPULATION OF OBJECTS AND A PROCEDURE FOR SELECTING n OF THEM. IF THE PROCEDURE ENSURES THAT *ALL POSSIBLE SAMPLES* OF n OBJECTS ARE *EQUALLY LIKELY*, THEN WE CALL THE PROCEDURE A **simple random sample.**

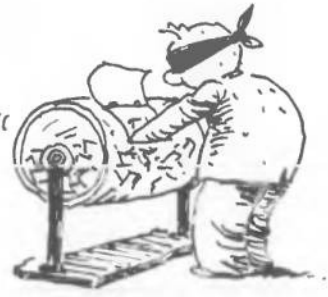
THE SIMPLE RANDOM SAMPLE HAS TWO PROPERTIES THAT MAKE IT THE STANDARD AGAINST WHICH WE MEASURE ALL OTHER METHODS:

1) UNBIASED: EACH UNIT HAS THE SAME CHANCE OF BEING CHOSEN.

2) INDEPENDENCE: SELECTION OF ONE UNIT HAS NO INFLUENCE ON THE SELECTION OF OTHER UNITS.

UNFORTUNATELY, IN THE REAL WORLD, COMPLETELY UNBIASED, INDEPENDENT SAMPLES ARE HARD TO FIND. FOR INSTANCE, SURVEYING VOTERS BY RANDOMLY DIALING TELEPHONE NUMBERS IS BIASED: IT IGNORES VOTERS WITHOUT A TELEPHONE AND OVERSAMPLES PEOPLE WITH MORE THAN ONE NUMBER.

IT'S THEORETICALLY POSSIBLE TO GET A RANDOM SAMPLE BY BUILDING A ***SAMPLING FRAME***: A LIST OF EVERY UNIT IN THE POPULATION. BY USING A RANDOM NUMBER GENERATOR, WE CAN PICK n OBJECTS AT RANDOM.



EQUIVALENTLY, WE CAN PUT ALL THE NAMES ON CARDS AND PULL n OF THEM OUT OF A DRUM.

BUT THIS IS NOT ALWAYS EASY. MAKING THE FRAME MAY BE PROHIBITIVELY COSTLY, CONTROVERSIAL, OR EVEN IMPOSSIBLE. FOR EXAMPLE, AN E.P.A. WATER QUALITY STUDY NEEDED A SAMPLING FRAME OF LAKES IN THE U.S., SO THEN SOMEBODY HAS TO DECIDE:

ARE THERE OTHER WAYS TO SAMPLE THAT ARE MORE ***EFFICIENT*** AND ***COST-EFFECTIVE*** THAN A SIMPLE RANDOM SAMPLE? YES—IF YOU ALREADY KNOW SOMETHING ABOUT THE POPULATION. FOR INSTANCE...

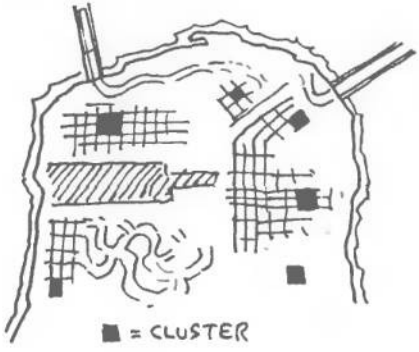
Stratified

SAMPLING: DIVIDE THE POPULATION UNITS INTO HOMOGENEOUS GROUPS *(STRATA)* AND DRAW A SIMPLE RANDOM SAMPLE FROM EACH GROUP.

FOR EXAMPLE, THE POPULATION OF ALL *PICKLES* CAN BE STRATIFIED BY *TYPE OF PICKLE*. WITHIN EACH TYPE OR STRATUM, THE SIZE SHOULD BE LESS VARIABLE.

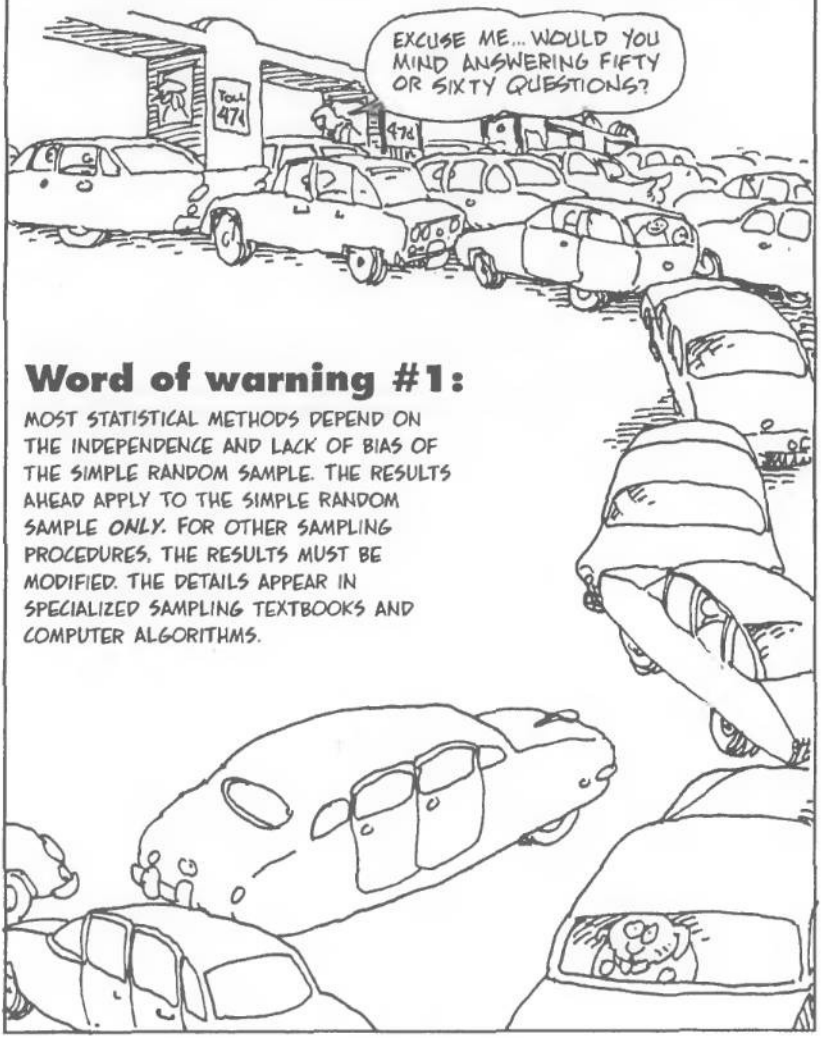
Cluster

SAMPLING GROUPS THE POPULATION INTO SMALL CLUSTERS, DRAWS A SIMPLE RANDOM SAMPLE OF CLUSTERS, AND OBSERVES EVERYTHING IN THE SAMPLED CLUSTERS. THIS CAN BE COST-EFFECTIVE IF TRAVEL COSTS BETWEEN RANDOMLY SAMPLED UNITS IS HIGH.



AN EXAMPLE IS A CITY HOUSING SURVEY WHICH DIVIDES A CITY INTO BLOCKS, RANDOMLY SAMPLES THE BLOCKS, AND LOOKS AT EVERY HOUSING UNIT IN EACH SAMPLED BLOCK.

Systematic SAMPLING STARTS WITH A RANDOMLY CHOSEN UNIT AND THEN SELECTS EVERY k^{TH} UNIT THEREAFTER. FOR INSTANCE, A *HIGHWAY TRAFFIC STUDY* MIGHT CHECK EVERY HUNDREDTH CAR AT A TOLL BOOTH. THIS PLAN IS EASY TO IMPLEMENT AND CAN BE MORE EFFICIENT IF TRAFFIC PATTERNS VARY SMOOTHLY OVER TIME.



Word of warning #1:

MOST STATISTICAL METHODS DEPEND ON THE INDEPENDENCE AND LACK OF BIAS OF THE SIMPLE RANDOM SAMPLE. THE RESULTS AHEAD APPLY TO THE SIMPLE RANDOM SAMPLE *ONLY*. FOR OTHER SAMPLING PROCEDURES, THE RESULTS MUST BE MODIFIED. THE DETAILS APPEAR IN SPECIALIZED SAMPLING TEXTBOOKS AND COMPUTER ALGORITHMS.

Word of warning #2:

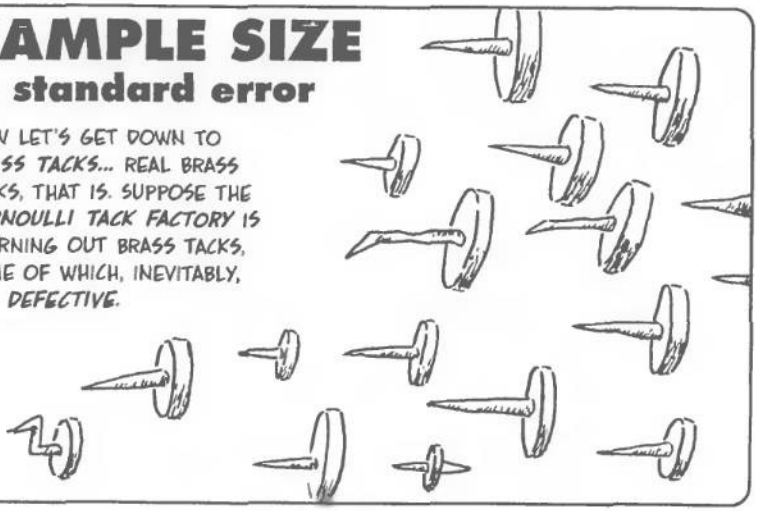
WITHOUT RANDOMIZED DESIGN, THERE CAN BE NO DEPENDABLE STATISTICAL ANALYSIS, NO MATTER HOW IT IS MODIFIED. THE BEAUTY OF RANDOM SAMPLING IS THAT IT "STATISTICALLY GUARANTEES" THE ACCURACY OF THE SURVEY.

A COMMONLY USED METHOD IS ESPECIALLY PRONE TO BIAS: IT'S CALLED AN **opportunity** SAMPLE. AVOIDING ALL THE BOTHER OF DESIGNING A PROCEDURE, THE OPPORTUNITY SAMPLER JUST GRABS THE FIRST n POPULATION UNITS TO COME ALONG.

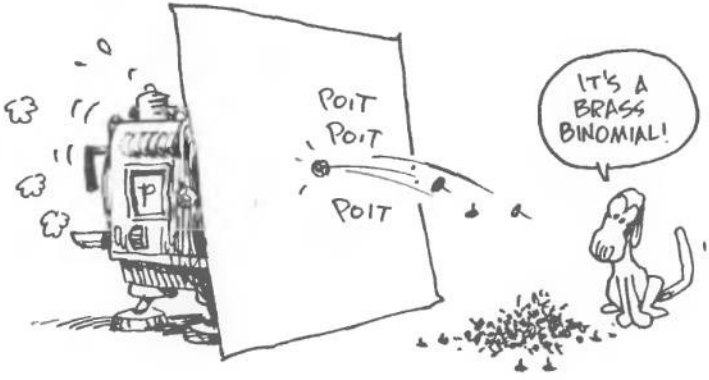
A CLASSIC EXAMPLE IS SHERE HITE'S BOOK, *WOMEN AND LOVE*. 100,000 QUESTIONNAIRES WENT TO WOMEN'S ORGANIZATIONS (AN *OPPORTUNITY SAMPLE*), ONLY *4.5%* WERE FILLED OUT AND RETURNED (*RESPONSE BIAS*). SO HER "RESULTS" WERE BASED ON A SAMPLE OF WOMEN WHO WERE HIGHLY MOTIVATED TO ANSWER THE SURVEY'S QUESTIONS, FOR WHATEVER REASON.

SAMPLE SIZE
& standard error

NOW LET'S GET DOWN TO *BRASS TACKS*... REAL BRASS TACKS, THAT IS. SUPPOSE THE *BERNOULLI TACK FACTORY* IS CHURNING OUT BRASS TACKS, SOME OF WHICH, INEVITABLY, ARE *DEFECTIVE*.



THE ASTUTE READER WILL RECOGNIZE THIS AS A *BERNOULLI SYSTEM*: EACH NEW TACK IS THE OUTCOME OF A BERNOULLI TRIAL WITH SOME PROBABILITY p OF SUCCESS (I.E., BEING DEFECT-FREE) AND PROBABILITY $1-p$ OF FAILURE (I.E., BEING DEFECTIVE).



WE THINK OF THIS SITUATION AS IF THERE WERE A *HIDDEN BUT REAL "BERNOULLI MACHINE"* WHOSE PROBABILITY p GOVERNS THE OUTCOMES WE OBSERVE IN THE SO-CALLED "REAL WORLD."

SINCE THE BERNOULLI MACHINE IS INVISIBLE, WE DON'T KNOW WHAT p IS, BUT WE'D LIKE TO FIND OUT. SO WE TAKE A ***RANDOM SAMPLE*** OF n TACKS, AND FIND THAT x OF THEM ARE O.K.

NOW THE PROPORTION OF SUCCESSES IN THE SAMPLE SHOULD BE SOMEWHERE AROUND p... SO WE CALL IT $\hat{p}$, PRONOUNCED "P-HAT."

$$\hat{p} = \frac{x}{n}$$

$\hat{p}$ IS THE NUMBER OF SUCCESSES x IN THE SAMPLE, DIVIDED BY THE SAMPLE SIZE n. FOR EXAMPLE, IF p WAS .85, AND WE SAMPLED $n=1000$ TACKS, MAYBE WE FOUND $x=832$ GOOD ONES, MAKING $\hat{p}=.832$.

WE ASK: HOW GOOD IS THIS ESTIMATE?

AND WE ANSWER WITH ANOTHER QUESTION: WHAT DOES THE FIRST QUESTION MEAN?

WE CAN'T KNOW THE PRECISE DIFFERENCE BETWEEN $\hat{p}$ AND p, BECAUSE WE DON'T KNOW THE VALUE OF p. THE *REAL* QUESTION IS THIS: IF WE TOOK *MANY SAMPLES OF 1000 TACKS* AND OBSERVED $\hat{p}$ FOR EACH SAMPLE, HOW WOULD THOSE VALUES OF $\hat{p}$ BE *DISTRIBUTED* AROUND p?

IN FACT, THESE $\hat{p}$ VALUES ARE LOOKING MORE AND MORE LIKE A *RANDOM VARIABLE*: THE SELECTION OF THE n-UNIT SAMPLE IS A RANDOM EXPERIMENT, AND THE OBSERVATION $\hat{p}$ IS A NUMERICAL OUTCOME!

TO BE PRECISE, IF X IS THE NUMBER OF SUCCESSES IN THE SAMPLE, THEN X IS NOTHING BUT OUR OLD FRIEND THE BINOMIAL RANDOM VARIABLE (n TRIALS, PROBABILITY p)... AND WE DEFINE THE *OBSERVED PROPORTION* TO BE THE RANDOM VARIABLE

$$\hat{P} = \frac{X}{n}$$

KNOWING ALL ABOUT X, WE QUICKLY CONCLUDE A FEW FACTS ABOUT $\hat{P}$:

1) THE MEAN OF $\hat{P}$ IS $E[\hat{P}] = p$

2) THE STANDARD DEVIATION OF $\hat{P}$ IS

$$\sigma(\hat{P}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

3) FOR LARGE n, $\hat{P}$ IS APPROXIMATELY NORMAL.

AND THERE YOU HAVE IT ALL! THE OBSERVED VALUES OF $\hat{P}$ WILL BE CENTERED ON p (NOT SURPRISINGLY), AND THEIR STANDARD DEVIATION, OR SPREAD, IS PROPORTIONAL TO THAT MAGIC NUMBER WE MENTIONED AT THE BEGINNING OF THE CHAPTER:

AND, SINCE $\hat{P}$ IS NEARLY NORMAL, WE CAN USE OUR RULE OF THUMB TO CONCLUDE THAT APPROXIMATELY 68% OF ALL ESTIMATES WILL FALL WITHIN ONE STANDARD DEVIATION OF THE TRUE VALUE p.

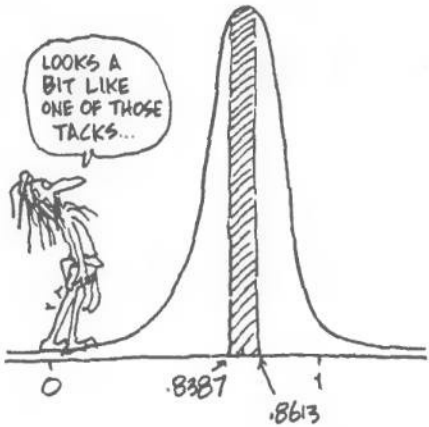
GOING BACK TO THE TACKS, WITH $n = 1000$ AND $p = .85$, WE GET A STANDARD DEVIATION OF

$$\sigma(\hat{P}) = \sqrt{\frac{(.85)(.15)}{1000}}$$

$$= .0113$$

SO WE EXPECT ABOUT 68% OF OUR ESTIMATES TO FALL IN THE NARROW INTERVAL

$$.8387 \leq \hat{p} \leq .8613$$



THE STANDARD DEVIATION OF $\hat{P}$ IS A MEASURE OF THE **sampling error.** AS WE'VE SEEN, FOR THE BINOMIAL $\hat{P}$, THIS SAMPLING ERROR IS INVERSELY PROPORTIONAL TO $\sqrt{n}$. INCREASING THE SAMPLE SIZE BY A FACTOR OF 4 REDUCES THE SPREAD $\sigma(\hat{P})$ BY A FACTOR OF 2.

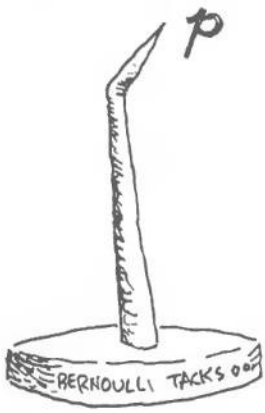
SAMPLE SIZES FOR TACKS, $p = 0.85$

n	1	4	16	25	100	10,000
$\sqrt{n}$	1	2	4	5	10	100
$\sigma(\hat{p})$	.357	.1785	.089	.071	.0357	.0036

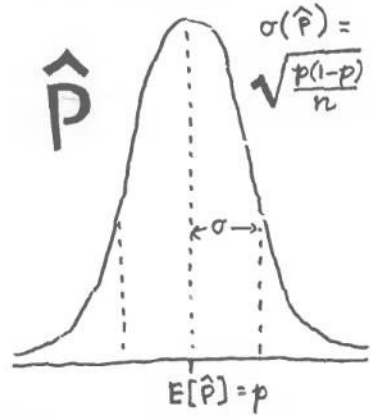
LINGUISTIC NOTE: AN ***ESTIMATE*** IS A SINGLE MEASURE OR OBSERVATION. AN ***ESTIMATOR*** IS A RULE FOR GETTING ESTIMATES. IN THIS CASE, THE ESTIMATOR IS THE RANDOM VARIABLE $\hat{P} = \frac{X}{n}$.

MOST OF STATISTICS INVOLVES THE 4-STEP PROCESS WE'VE JUST WALKED THROUGH:

DEFINE POPULATION WITH UNKNOWN PARAMETER



FIND AN ESTIMATOR, ITS THEORETICAL SAMPLING DISTRIBUTION AND STANDARD DEVIATION.

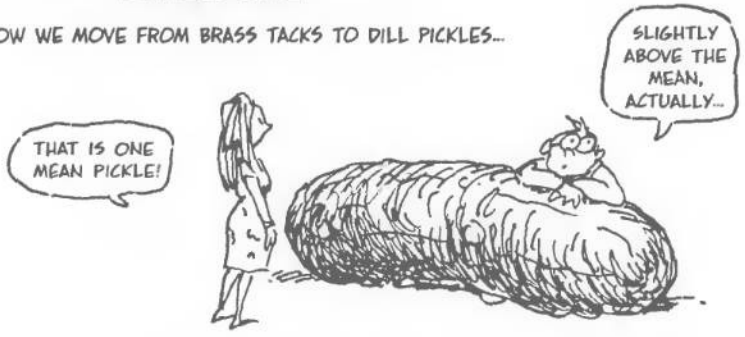


ACTUALLY DRAW A RANDOM SAMPLE AND FIND THE ESTIMATE.

REPORT THE RESULT AND ITS STATISTICAL OR SAMPLING ERROR.

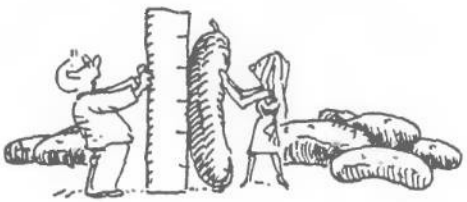
Sampling Distribution of the MEAN

NOW WE MOVE FROM BRASS TACKS TO DILL PICKLES...



JAR MANUFACTURERS WOULD LIKE TO KNOW THE ***AVERAGE LENGTH*** OF A PICKLE WITHOUT EXAMINING EVERY CUCUMBER IN CALIFORNIA. THEY RANDOMLY SELECT n PICKLES AND MEASURE THEIR LENGTHS $x_1, x_2, \ldots, x_n$.

BY NOW YOU MAY BE USED TO THE IDEA THAT EACH X_i IS A ***RANDOM VARIABLE:*** THE NUMERICAL OUTCOME OF A RANDOM EXPERIMENT.



IF μ IS THE (UNKNOWN) MEAN PICKLE LENGTH, AND σ IS THE STANDARD DEVIATION OF THE ***PICKLE LENGTH DISTRIBUTION,*** THEN

$$E[X_i] = \mu$$
$$\sigma(X_i) = \sigma$$

FOR EVERY i (BECAUSE x_i COULD HAVE BEEN THE LENGTH OF ANY PICKLE).

NOW WE LOOK AT THE SAMPLE MEAN: THE AVERAGE LENGTH OF THE SELECTED PICKLES. IT'S A NEW RANDOM VARIABLE GIVEN BY:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

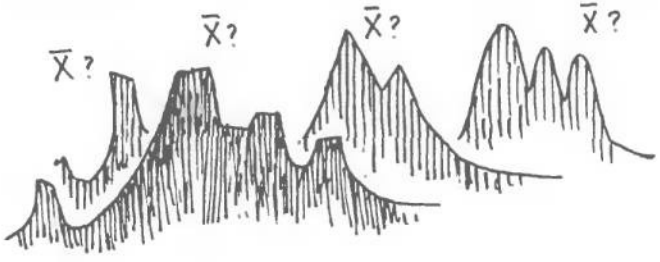
AS BEFORE, WE'D LIKE TO KNOW "HOW CLOSE" THIS IS TO μ, MEANING, IF THIS SAMPLING WERE DONE MANY TIMES, WHAT'S THE DISTRIBUTION OF $\bar{X}$? BECAUSE WE KNOW ABOUT X_1, X_2, ..., AND X_n, WE ALSO KNOW THAT

$$E[\bar{X}] = \mu$$

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

ONCE AGAIN, WE SEE THE MAGIC DENOMINATOR! THE SPREAD OF OBSERVED SAMPLE MEANS GOES AS $\frac{1}{\sqrt{n}}$.

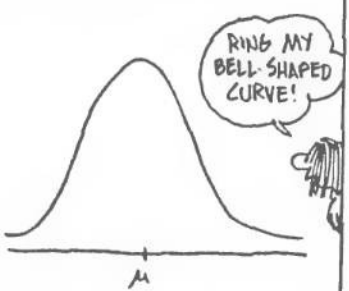
BUT WE DON'T KNOW THE SHAPE OF $\bar{X}$'S DISTRIBUTION. THE SAMPLE *PROBABILITY DISTRIBUTION* $\hat{p}$ WAS ALMOST NORMAL, BECAUSE IT WAS BASED ON A BINOMIAL RANDOM VARIABLE. BUT WHAT ABOUT $\bar{X}$, THE SAMPLE *MEAN* ESTIMATOR???



IT TURNS OUT THAT $\bar{X}$ IS ALSO APPROXIMATELY NORMAL! THIS FAMOUS RESULT IS CALLED THE

CENTRAL LIMIT THEOREM

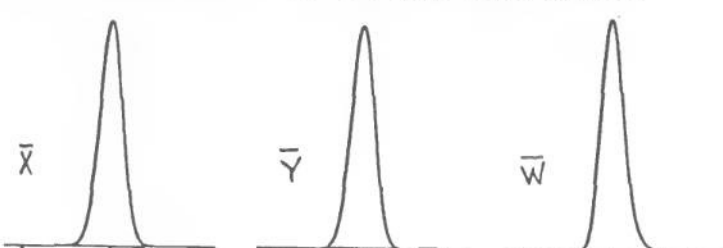
IT SAYS: IF ONE TAKES RANDOM SAMPLES OF SIZE n FROM A POPULATION OF MEAN μ AND STANDARD DEVIATION σ, THEN, AS n GETS LARGE, $\bar{X}$ APPROACHES THE *NORMAL DISTRIBUTION* WITH MEAN μ AND STANDARD DEVIATION $\sigma/\sqrt{n}$. THEN



$$Pr(a \leq \bar{X} \leq b) \approx Pr\left(\frac{a-\mu}{\sigma/\sqrt{n}} \leq Z \leq \frac{b-\mu}{\sigma/\sqrt{n}}\right)$$

WHAT IS REMARKABLE ABOUT THIS? IT SAYS THAT REGARDLESS OF THE SHAPE OF THE ORIGINAL DISTRIBUTION (IN THIS CASE, OF PICKLE LENGTHS), THE TAKING OF *AVERAGES* RESULTS IN A *NORMAL*. TO FIND THE DISTRIBUTION OF $\bar{X}$, WE NEED KNOW *ONLY* THE POPULATION MEAN AND STANDARD DEVIATION.

THE THREE PROBABILITY DENSITIES ABOVE ALL HAVE THE SAME MEAN AND STANDARD DEVIATION. DESPITE THEIR DIFFERENT SHAPES, WHEN n=10, THE SAMPLING DISTRIBUTIONS OF THE MEAN, $\bar{X}$, ARE NEARLY IDENTICAL.



The t-distribution

AMAZING AS THE CENTRAL LIMIT THEOREM IS, IT HAS AT LEAST TWO PROBLEMS.

ONE: IT DEPENDS ON A LARGE SAMPLE SIZE.

TWO: TO USE IT, WE NEED TO KNOW σ, THE STANDARD DEVIATION.

BUT SAMPLE SIZES ARE OFTEN SMALL, AND σ IS USUALLY UNKNOWN. CERTAINLY, IN THE CASE OF THE PICKLES, WE HAVE *NO IDEA* HOW WIDELY THEIR LENGTHS VARY AROUND THE AVERAGE.

WHAT WE CAN DO IN THIS CASE IS TO *ESTIMATE* σ BY TAKING THE *STANDARD DEVIATION OF THE SAMPLE*, WHICH, YOU'LL RECALL, IS GIVEN BY THE FORMULA

$$s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

THEN, IN PLACE OF THE RANDOM VARIABLE

$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

WE SUBSTITUTE s FOR σ, AND DEFINE A *NEW RANDOM VARIABLE* t BY

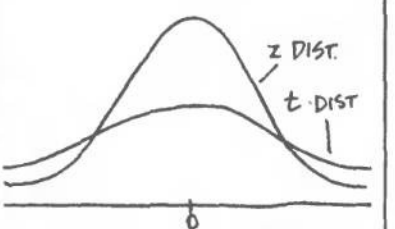
$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

YOU CAN THINK OF THE RANDOM VARIABLE t AS *THE BEST WE CAN DO UNDER THE CIRCUMSTANCES*. ITS DISTRIBUTION IS CALLED *STUDENT'S* t, BECAUSE ITS INVENTOR, *WILLIAM GOSSET*, PUBLISHED UNDER THE PSEUDONYM "STUDENT."

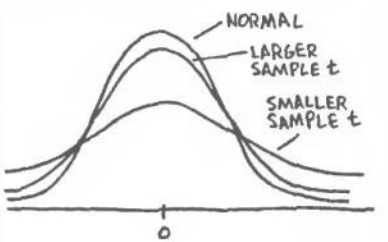
(GOSSET WAS EMPLOYED BY THE *GUINNESS BREWERY*, WHICH REQUIRED HIM TO USE A PSEUDONYM, FOR SOME REASON.)

MAKING THE ASSUMPTION THAT THE *ORIGINAL POPULATION DISTRIBUTION WAS NORMAL*, OR NEARLY NORMAL, "STUDENT" WAS ABLE TO CONCLUDE:

t IS MORE SPREAD OUT THAN z. IT'S "FLATTER" THAN NORMAL. THIS IS BECAUSE THE USE OF s INTRODUCES MORE UNCERTAINTY, MAKING t "SLOPPIER" THAN z.

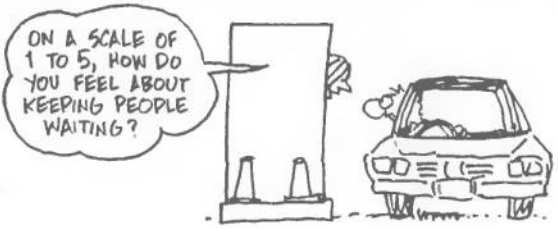


THE AMOUNT OF SPREAD DEPENDS ON THE *SAMPLE SIZE*. THE GREATER THE SAMPLE SIZE, THE MORE CONFIDENT WE CAN BE THAT s IS NEAR σ, AND THE CLOSER t GETS TO z, THE NORMAL.



GOSSET WAS ABLE TO COMPUTE TABLES OF t FOR VARIOUS SAMPLE SIZES, WHICH WE WILL SEE HOW TO USE IN THE FOLLOWING CHAPTER.

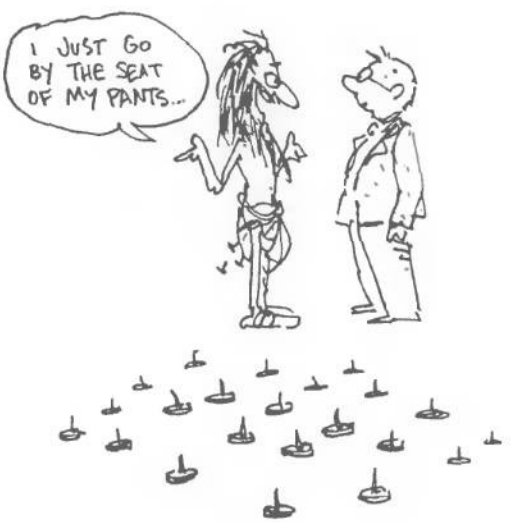
IN THIS CHAPTER, WE CONSIDERED A CENTRAL PROBLEM OF *REAL-WORLD STATISTICS*: HOW TO SELECT A *SAMPLE* FROM A LARGE POPULATION SO THAT STATISTICAL ANALYSIS CAN BE VALID. BESIDES THE "GOLD STANDARD" OF THE SIMPLE RANDOM SAMPLE, WE ALSO DESCRIBED SOME OTHER SAMPLING SCHEMES THAT ARE USED IN THE INTERESTS OF EFFICIENCY, COST, AND PRACTICALITY.



THEN, ASSUMING A SIMPLE RANDOM SAMPLE, WE CONSIDERED HOW VARIOUS SAMPLE STATISTICS WERE *DISTRIBUTED*. THAT IS, WE REGARDED THE ACT OF TAKING THE SAMPLE AS A *RANDOM EXPERIMENT*, SO THAT ITS STATISTICS BECAME *RANDOM VARIABLES*.

WE FOUND THAT SAMPLE *PROPORTIONS* $\hat{p}$ WERE APPROXIMATELY NORMALLY DISTRIBUTED, WHILE THE DISTRIBUTION OF THE SAMPLE *MEAN* $\bar{X}$ DEPENDED ON THE SAMPLE SIZE. FOR LARGE SAMPLES, THE DISTRIBUTION WAS APPROXIMATELY NORMAL, WHILE FOR SMALL SAMPLES, WE USE THE STUDENT'S t DISTRIBUTION.

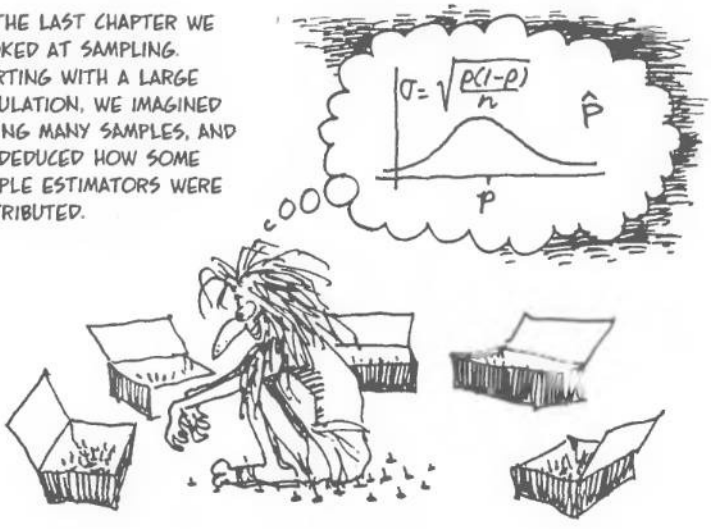
IN THE NEXT TWO CHAPTERS, WE LOOK AT HOW TO USE THESE DISTRIBUTIONS TO MAKE *STATISTICAL INFERENCES*: GIVEN A SINGLE OBSERVATION, LIKE A POLITICAL POLL, HOW DO WE USE OUR KNOWLEDGE OF $\hat{p}$ AND $\bar{X}$ TO EVALUATE IT?



♦Chapter 7♦

CONFIDENCE INTERVALS

IN THE LAST CHAPTER WE LOOKED AT SAMPLING. STARTING WITH A LARGE POPULATION, WE IMAGINED TAKING MANY SAMPLES, AND WE DEDUCED HOW SOME SAMPLE ESTIMATORS WERE DISTRIBUTED.



IN THIS CHAPTER, WE DO THE REVERSE. GIVEN *ONE* SAMPLE, WE ASK THE QUESTION, WHAT WAS THE RANDOM SYSTEM THAT GENERATED ITS STATISTICS?

IN *DEDUCTIVE REASONING*, WE REASON FROM A HYPOTHESIS TO A CONCLUSION: "IF LORD FASTBACK COMMITTED MURDER, THEN HE WOULD WIPE THE FINGERPRINTS OFF THE GUN."

INDUCTIVE REASONING, BY CONTRAST, ARGUES *BACKWARD* FROM A SET OF OBSERVATIONS TO A REASONABLE HYPOTHESIS:

IN MANY WAYS, SCIENCE, INCLUDING STATISTICS, IS LIKE DETECTIVE WORK. BEGINNING WITH A SET OF OBSERVATIONS, WE ASK WHAT CAN BE SAID ABOUT THE SYSTEMS THAT GENERATED THEM.

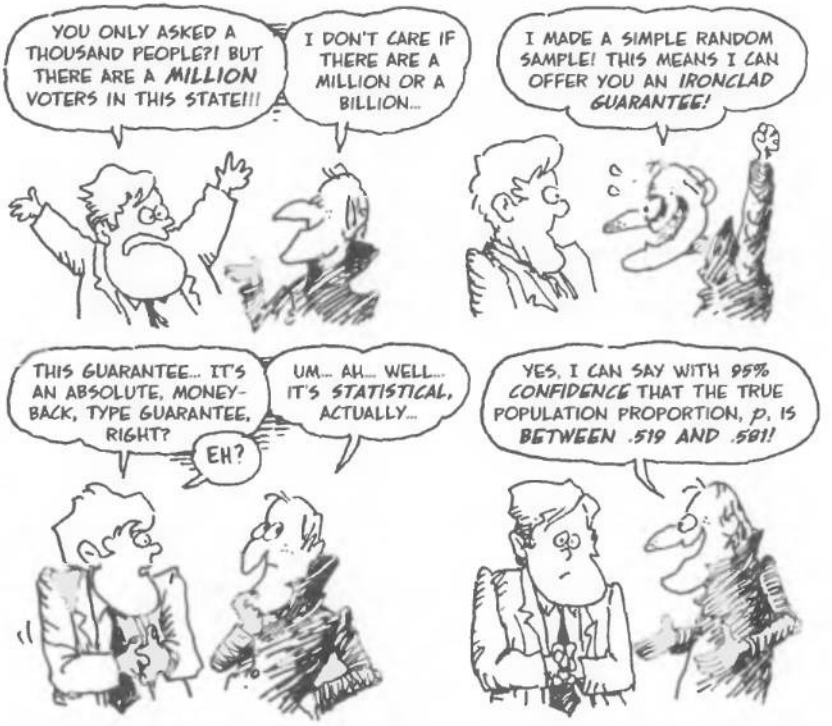
ESTIMATING
CONFIDENCE INTERVALS

IS ONE OF THE MOST EFFECTIVE FORMS OF STATISTICAL INFERENCE, AND ONE YOU SEE EVERY DAY BEFORE *ELECTION TIME*...

IN A RECENT ELECTION SOMEWHERE, INCUMBENT SENATOR *ASTUTE* (ACCENT ON THE LAST SYLLABLE, PLEASE!) COMMISSIONED A POLL BY *BETTER HOLMES RESEARCH*. POLLSTER HOLMES DRAWS A SIMPLE RANDOM SAMPLE OF 1000 VOTERS AND ASKS THEM WHAT THEY THINK OF ASTUTE.

AFTER CENSORING THE REMARKS OF A FEW GRUMPY OUTLIERS, HOLMES FINDS THAT *550* VOTERS FAVOR HIS CLIENT, SENATOR ASTUTE.

THIS IS THE SINGLE OBSERVATION.



AFTER ASTUTE CALMS DOWN, HOLMES EXPLAINS WHAT HE MEANS BY *95% CONFIDENCE*: HE KNOWS THAT HIS ESTIMATION PROCEDURE HAS A 95% *PROBABILITY* OF PRODUCING AN INTERVAL CONTAINING p, I.E., IN HIS MANY YEARS OF POLLING, p HAS FALLEN WITHIN THE CONFIDENCE INTERVAL AROUND THE OBSERVED VALUE, $\hat{p}$, 95% OF THE TIME.

SENATOR ASTUTE IS STILL CONFUSED! SO HOLMES GIVES HIM AN **archery lesson.**

SHOOT! ANYTHING TO TAKE MY MIND OFF THEM DANG STATISTICS!

CONSIDER AN ARCHER-POLLSTER SHOOTING AT A TARGET. SUPPOSE THAT SHE HITS THE 10 CM RADIUS BULL'S-EYE 95% OF THE TIME. THAT IS, ONLY ONE ARROW OUT OF 20 MISSES.

SITTING BEHIND THE TARGET IS A BRAVE DETECTIVE, WHO CAN'T SEE THE BULL'S-EYE. THE ARCHER SHOOTS A SINGLE ARROW.

KNOWING THE ARCHER'S SKILL LEVEL, THE DETECTIVE DRAWS A CIRCLE WITH 10 CM RADIUS AROUND THE ARROW. HE NOW HAS 95% *CONFIDENCE* THAT HIS CIRCLE INCLUDES THE CENTER OF THE BULL'S-EYE!

10 CM

HE REASONED THAT IF HE DREW 10 CM RADIUS CIRCLES AROUND *MANY ARROWS*, HIS CIRCLES WOULD INCLUDE THE CENTER 95% OF THE TIME.

TRUE CENTER

X=ARROW

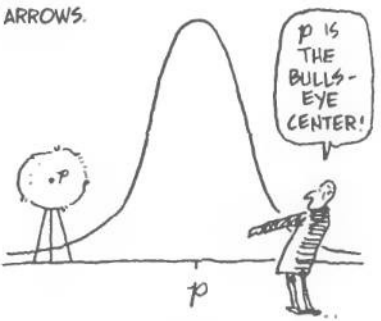
(PROBABILISTS USE THE TERM *STOCHASTIC* TO DESCRIBE RANDOM MODELS. IT'S DERIVED FROM THE GREEK *STOCHAZES-THAI*, MEANING TO AIM AT A TARGET, OR GUESS. FROM *STOCHOS*, A TARGET.)

HOLMES NOW TRANSLATES THE ARCHERY LESSON INTO THE LANGUAGE WE DEVELOPED LAST CHAPTER.

Step One: SHOOT A LOT OF ARROWS.

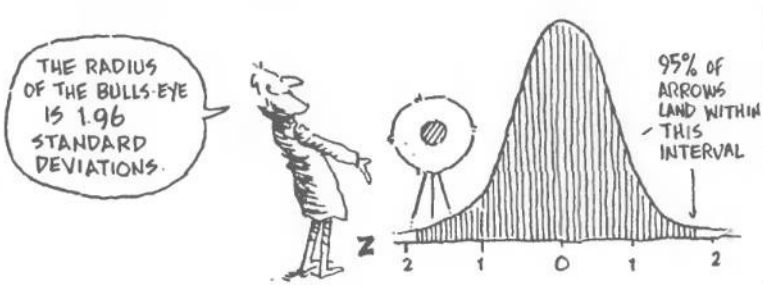
A PROBABILITY CALCULATION FINDS THE WIDTH OF THE "BULL'S-EYE." THE ESTIMATES $\hat{p}$ ARE OUR ARROWS. WE SAW THAT THE SAMPLING DISTRIBUTION OF $\hat{p}$ IS NEARLY *NORMAL* WITH MEAN p AND STANDARD DEVIATION

$$\sigma(\hat{p}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$



SINCE THE CURVE IS NORMAL, WE USE THE Z-TRANSFORM AND A STANDARD TABLE TO FIND THE WIDTH OF THE INTERVAL WITHIN WHICH 95% OF THE "ARROWS" HIT. (WE'LL SEE EXACTLY HOW TO DO THIS IN A FEW PAGES.) WE FIND THIS WIDTH TO BE **1.96** STANDARD DEVIATIONS:

$$.95 = Pr(-1.96 \leq Z \leq 1.96)$$



NOW WE DO SOME ALGEBRA. BY DEFINITION OF THE Z-TRANSFORM,

$$.95 \simeq \Pr\left(-1.96 \le \frac{\hat{p}-p}{\sigma(p)} \le 1.96\right)$$

WHICH BECOMES

$$.95 \simeq \Pr\left(p-1.96\sigma(p) \le \hat{p} \le p+1.96\sigma(p)\right)$$

WHICH IS JUST ANOTHER WAY OF SAYING THAT 95% OF THE $\hat{p}$ "ARROWS" LAND BETWEEN $p - 1.96\sigma(p)$ AND $p + 1.96\sigma(p)$.

NOW WE'RE IN A POSITION TO VIEW THE TARGET FROM BEHIND! ONE MORE TURN OF THE ALGEBRA CRANK MAKES IT

$$.95 \simeq \Pr\left(\hat{p}-1.96\sigma(p) \le p \le \hat{p}+1.96\sigma(p)\right)$$

HERE WE ARE DRAWING CIRCLES AROUND A LOT OF ARROWS (I.E., MAKING INTERVALS AROUND $\hat{p}$) AND SAYING THAT 95% OF THEM COVER p.

BUT THERE IS ONE TINY PROBLEM... ***WE DON'T ACTUALLY KNOW THE SIZE OF THE BULL'S-EYE,*** BECAUSE WE DON'T KNOW p, AND THE WIDTH IS A MULTIPLE OF $\sigma(p)$.

SO WE FUDGE A LITTLE AND USE THE ***STANDARD ERROR*** OF $\hat{P}$:

$$SE(\hat{P}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

IN ITS PLACE... IT'S CLOSE ENOUGH... IT'S THE BEST WE CAN DO... AND IT CAN EVEN BE THEORETICALLY JUSTIFIED!

NOW THE FORMULA IS

$$.95 \approx \Pr\left(\hat{p} - 1.96\,SE(\hat{p}) \le p \le \hat{p} + 1.96\,SE(\hat{p})\right)$$

AGAIN, THIS EQUATION DESCRIBES THE PROBABILITY THAT THE TRUE, FIXED POPULATION PROPORTION FALLS WITHIN THE *RANDOM INTERVAL*

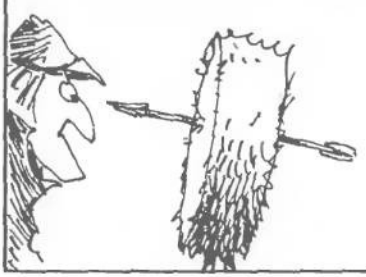
$$(\hat{p} - 1.96\,SE(\hat{p}),\ \hat{p} + 1.96\,SE(\hat{p})).$$

IF WE SAMPLED REPEATEDLY, THESE INTERVALS WOULD COVER p 95% OF THE TIME.

NOW OUR PROBABILITY CALCULATION IS DONE, AND IT'S TIME FOR...

Step Two:

THE DETECTIVE WORK. IN A REAL POLL, HOLMES TAKES JUST ONE SIMPLE RANDOM SAMPLE OF 1000 VOTES, FINDS $\hat{p} = .550$, AND WANTS TO INFER p.



HE MAKES USE OF STEP ONE TO COMPUTE

$$SE(\hat{p}) = \sqrt{\frac{(.55)(.45)}{1000}} = .0157$$

HE CONCLUDES THAT WE CAN HAVE 95% CONFIDENCE THAT p IS WITHIN THE RANGE

$$\hat{p} \pm 1.96\,SE(\hat{p})$$

$$= .550 \pm (1.96)(.0157)$$

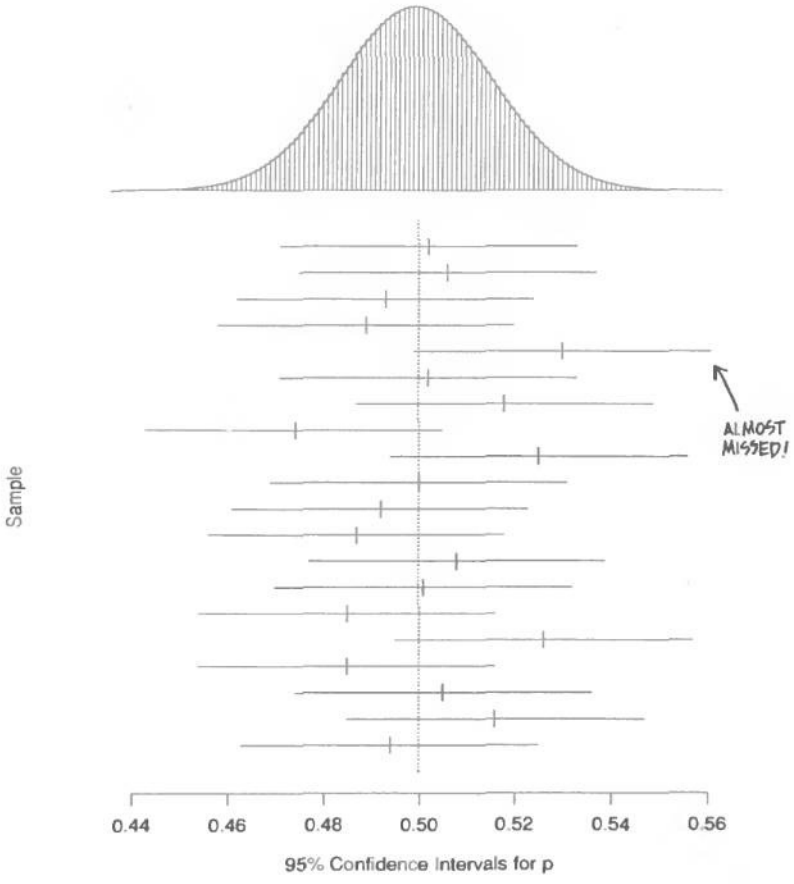
$$= .550 \pm .031$$

THIS IS WHAT POLLS MEAN WHEN THEY REFER TO THEIR "MARGIN OF ERROR." IN THIS CASE, HOLMES FOUND THAT

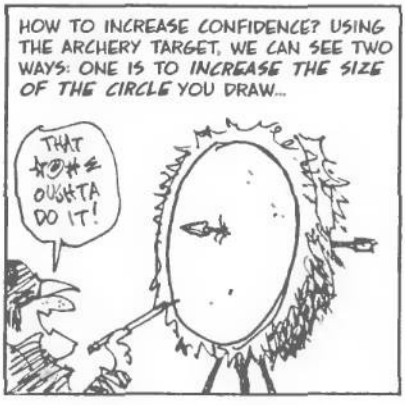
$$.519 \le p \le .581,$$

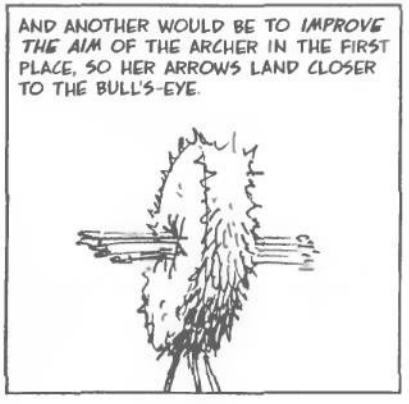
IN OTHER WORDS THAT $p = 55\%$ WITH A 3% MARGIN OF ERROR. (POLLS TYPICALLY USE A 95% CONFIDENCE LEVEL.)

THIS PAGE SHOWS THE RESULTS OF A COMPUTER SIMULATION OF TWENTY SAMPLES OF SIZE $n = 1000$. WE ASSUMED THAT THE TRUE VALUE OF $p = .5$. AT THE TOP YOU SEE THE SAMPLING DISTRIBUTION OF $\hat{p}$ (NORMAL, WITH MEAN p AND $\sigma = \sqrt{\frac{p(1-p)}{n}}$). BELOW ARE THE 95% CONFIDENCE INTERVALS FROM EACH SAMPLE. ON AVERAGE, ONE OUT OF TWENTY (OR 5%) OF THESE INTERVALS WILL NOT COVER THE POINT $p = .5$.



ALTHOUGH 95% CONFIDENCE IS GOOD ENOUGH FOR NEWSPAPER POLLS, IT ISN'T GOOD ENOUGH FOR *SENATOR ASTUTE*. HE WANTS 99%!



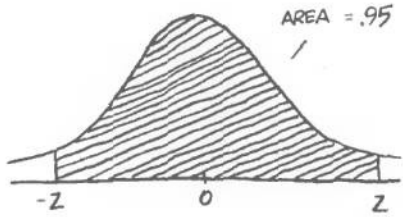


THE FIRST METHOD IS EQUIVALENT TO *WIDENING THE CONFIDENCE INTERVAL*. THE GREATER THE MARGIN OF ERROR, THE MORE CERTAIN YOU ARE THE TRUE VALUE OF p LIES IN THE INTERVAL.

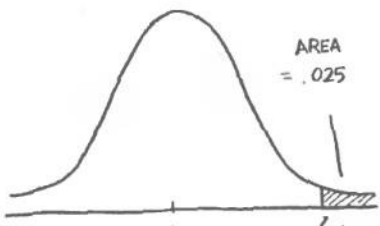
MAYBE IT'S TIME TO SEE EXACTLY HOW WE FIND THE ENDS OF THESE CONFIDENCE INTERVALS...

THE RELEVANT NUMBER HERE WE USUALLY CALL α. IT MEASURES THE DIFFERENCE BETWEEN THE DESIRED CONFIDENCE LEVEL AND CERTAINTY. FOR EXAMPLE, WHEN THE CONFIDENCE LEVEL IS 95%, OR 0.95, α IS .05. SO WE SPEAK OF THE $(1-\alpha)\cdot 100\%$ CONFIDENCE LEVEL.

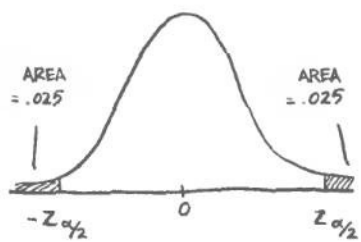
FINDING THE $(1-\alpha)\cdot 100\%$ CONFIDENCE INTERVAL MEANS: LOOK AT A STANDARD NORMAL CURVE, AND FIND THE POINTS $\pm z$ BETWEEN WHICH THE AREA IS $1-\alpha$.



THIS POINT, CALLED $z_{\frac{\alpha}{2}}$, IS THE z-VALUE BEYOND WHICH THE AREA IS $.025 = \frac{\alpha}{2}$.



THAT'S BECAUSE WE'RE CHOPPING OFF "TAILS" AT BOTH ENDS OF THE CURVE, WHICH HAVE A TOTAL AREA OF $\frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$.



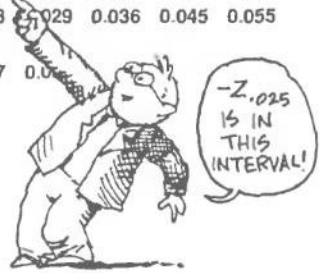
WE CAN FIND $z_{\alpha/2}$ STRAIGHT FROM THE STANDARD NORMAL TABLE (PAGE 84). IT'S THE POINT WITH THE PROPERTY

$$Pr(z \geq z_{\alpha/2}) = \frac{\alpha}{2}$$

IN PARTICULAR,

$$Pr(z \geq z_{.025}) = .025$$

z	-2.5	-2.4	-2.3	-2.2	-2.1
F(z)	0.006	0.008	0.011	0.014	0.018
z	-2.0	-1.9	-1.8	-1.7	-1.6
F(z)	0.023	0.029	0.036	0.045	0.055
z	-1.5				
F(z)	0.067				



HERE'S A LITTLE TABLE OF THE CRITICAL VALUES FOR VARIOUS LEVELS OF CONFIDENCE...

$1-\alpha$	.80	.90	.95	.99
α	.20	.10	.05	.01
$\alpha/2$	.10	.05	.025	.005
$z_{\frac{\alpha}{2}}$	1.28	1.64	1.96	2.58

TO MAKE A 99% CONFIDENCE INTERVAL, WE USE THAT TABLE TO WRITE

$$.99 = Pr(\hat{p} - 2.58SE(\hat{p}) \leq p \leq \hat{p} + 2.58SE(\hat{p}))$$

WHICH WE SLOPPILY ABBREVIATE AS

$$p = \hat{p} \pm 2.58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= .55 \pm 2.58\sqrt{\frac{(.55)(.45)}{1000}}$$

$$= .55 \pm .041$$

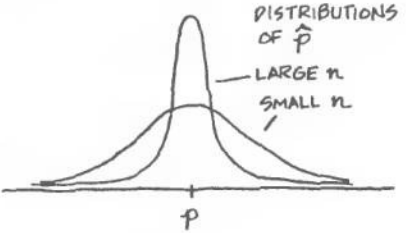
WITH 99% CONFIDENCE.

WIDENING THE INTERVAL IS ONE WAY TO INCREASE OUR CONFIDENCE IN THE RESULT. AS WE MENTIONED, ANOTHER WAY WOULD BE TO SHOOT OUR ARROWS ***MORE ACCURATELY.*** IF WE KNEW THAT THE ARCHER GOT 95% OF HER ARROWS WITHIN ***1 CM*** OF THE BULL'S-EYE, OUR ESTIMATES COULD BE A LOT SHARPER!

HOW DO WE DO THIS? BY INCREASING THE SAMPLE SIZE! THE WIDTH OF THE CONFIDENCE INTERVAL DEPENDS ON THE SAMPLE SIZE: THE INTERVAL HAS THE FORM $\hat{p}$ + E, WHERE E, THE ERROR, IS GIVEN BY

$$E = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

SO THE BIGGER WE MAKE n, THE SMALLER THE ERROR. (E.G., QUADRUPLING n HALVES THE INTERVAL WIDTH.)



ASTUTE ASKS HOLMES TO GIVE HIM A SMALL ERROR WITH HIGH CONFIDENCE—SAY 99% CONFIDENCE WITH E = ±.01. HOLMES SOLVES FOR n.

$$n = \frac{z_{\frac{\alpha}{2}}^{\,2}\, p^*(1-p^*)}{E^2}$$

(WHERE p^* IS A GUESS AT THE TRUE PROPORTION p—REMEMBER, WE HAVEN'T TAKEN THE SAMPLE YET!)

TAKING A CONSERVATIVE GUESS OF $p^* = .5$, HOLMES FINDS

$$n = \frac{(2.58)^2(.5)^2}{(.01)^2}$$

$$= \frac{(6.65)(.25)}{.0001}$$

$$= 16{,}641$$

1000 VOTERS GAVE A 3% ERROR WITH 95% CONFIDENCE. TO GET A 1% ERROR WITH 99% CONFIDENCE, HOLMES HAS TO SAMPLE 16,641 VOTERS!

SO THEY DO THE POLL, AND GO INTO THE ELECTION WITH *99% CONFIDENCE.*

BUT... ALL THIS PROBABILITY STUFF IS ONLY GOOD *BEFORE* AN ELECTION. *AFTER* THE ELECTION, THE SENATOR IS EITHER *100% IN* OR *100% OUT!* AND DESPITE EVERYTHING, SENATOR ASTUTE LOSES THE ELECTION...

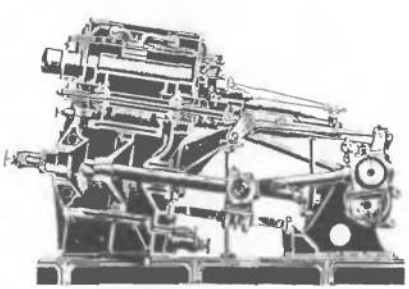
WHAT HAPPENED IS THAT POLITICIANS ARE NOT ELECTED BY POLLS!

SOME PROBLEMS WITH POLLS, AS OPPOSED TO ELECTIONS:

THERE IS NO WAY FOR A POLLSTER TO GET INSIDE A POTENTIAL VOTER'S HEAD AND KNOW IF SHE'S GOING TO VOTE, IF SHE'S LYING, OR IF SHE'S GOING TO CHANGE HER MIND BEFORE ELECTION DAY. LARGE SAMPLE SIZES CANNOT REDUCE THESE KINDS OF ERRORS.

SINCE THESE ERRORS CAN BE LARGE, IT SELDOM PAYS TO TAKE A VERY LARGE RANDOM SAMPLE.

IN THE LAST FIVE PRESIDENTIAL ELECTIONS, THE GALLUP POLL HAS INTERVIEWED FEWER THAN 4,000 VOTERS FOR EACH ELECTION. YET IN ALL FIVE ELECTIONS, THE GALLUP ORGANIZATION'S ERRORS IN PREDICTING THE PRESIDENTIAL ELECTION OUTCOME HAVE BEEN LESS THAN 2%.



THEIR SUCCESS IS DUE TO THEIR USE OF ESTIMATORS THAT ACCOUNT FOR NON-RESPONSE, AND THEY SCREEN OUT ELIGIBLE VOTERS WHO ARE NOT LIKELY TO VOTE.

TO SUMMARIZE, ESTIMATED PROPORTION = TRUE PROPORTION + BIAS + RANDOM SAMPLING ERROR. EVEN POLLSTERS HAVE LIMITED FUNDS. THEY WISELY CHOOSE TO SPEND THEIR MONEY *REDUCING BIAS*, RATHER THAN INCREASING THE SAMPLES BEYOND 4,000 VOTERS.

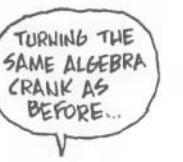
Confidence Intervals for μ

UP TO NOW, WE'VE BEEN LOOKING AT CONFIDENCE INTERVALS FOR A PROPORTION p OF A POPULATION. EXACTLY THE SAME REASONING WORKS FOR THE POPULATION MEAN μ.

IN THE LAST CHAPTER (P. 105), WE SAW THAT THE DISTRIBUTION OF SAMPLE MEANS $\bar{X}$ IS APPROXIMATELY ***NORMAL***, CENTERED ON THE ACTUAL POPULATION MEAN μ, WITH STANDARD DEVIATION $\sigma/\sqrt{n}$, WHERE σ IS THE POPULATION STANDARD DEVIATION. SO, FOR LARGE n,

$$.95 = Pr(-1.96 \leq Z \leq 1.96)$$

$$\approx Pr\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right)$$



AGAIN, NOT KNOWING σ, WE REPLACE σ WITH s, THE SAMPLE STANDARD DEVIATION:

$$.95 \approx Pr\left(-1.96 \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq 1.96\right)$$

THE TERM $s/\sqrt{n}$ IS CALLED THE ***SAMPLE STANDARD ERROR***, AND WRITTEN $SE(\bar{X})$. WE CONCLUDE THAT

$$.95 \approx Pr(\bar{X} - 1.96\,SE(\bar{X}) \leq \mu \leq \bar{X} + 1.96\,SE(\bar{X}))$$

WHERE

$$SE(\bar{X}) = \frac{s}{\sqrt{n}}$$

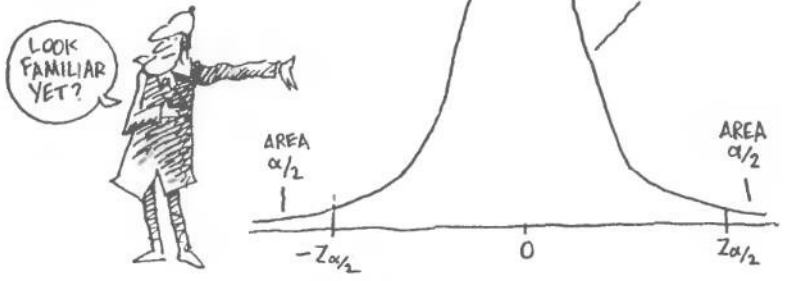
JUST AS BEFORE, WE HAVE FOUND THAT THE RANDOM INTERVAL

$$\bar{X} \pm 1.96\,SE(\bar{X})$$

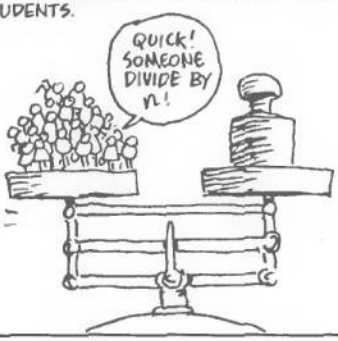
COVERS THE TRUE MEAN, μ, WITH PROBABILITY .95... SO NOW WE CAN CALL IN SHERLOCK HOLMES TO MAKE A STATISTICAL INFERENCE BASED ON A SINGLE SAMPLE OF SIZE n WITH MEAN $\bar{x}$.

HE (AND WE) ARE 95% CONFIDENT THAT THE MEAN μ IS WITHIN THE INTERVAL $\bar{x} \pm 1.96\,SE(\bar{X})$.

AS BEFORE, FOR AN ARBITRARY LEVEL OF CONFIDENCE $1-\alpha$, WE REPLACE 1.96 BY $z_{\frac{\alpha}{2}}$.



LET'S REVISIT THE STUDENT WEIGHT DATA FROM CHAPTER 2, ASSUMING THAT THE $n = 92$ STUDENTS WERE A SIMPLE RANDOM SAMPLE OF ALL PENN STATE STUDENTS.



THE SAMPLE MEAN $\bar{x}$ WAS 145.2 LBS. AND SAMPLE STANDARD DEVIATION s WAS 23.7. SO THE STANDARD ERROR IS

$$SE(\bar{x}) = \frac{23.7}{\sqrt{92}} = 2.47$$

AND WE NOW HAVE 95% CONFIDENCE THAT THE MEAN WEIGHT OF ALL PENN STATE STUDENTS FALLS IN THE INTERVAL

$$\bar{x} \pm 1.96 SE(\bar{x})$$
$$= 145.2 \pm (1.96)(2.47)$$
$$= \mathbf{145.2 \pm 4.8 \text{ POUNDS}}$$

TO SUMMARIZE: FOR A SIMPLE RANDOM SAMPLE (SRS) OF LARGE SIZE, THE $(1-\alpha)\cdot 100\%$ CONFIDENCE INTERVAL IS:

POPULATION MEAN, μ

$$\mu = \bar{x} \pm z_{\frac{\alpha}{2}} SE(\bar{x})$$

WHERE $SE(\bar{x}) = s/\sqrt{n}$

POPULATION PROPORTION, p

$$p = \hat{p} \pm z_{\frac{\alpha}{2}} SE(\hat{p})$$

WHERE $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

THE SIZE OF BOTH INTERVALS IS CONTROLLED BY THE LEVEL OF CONFIDENCE $(1-\alpha)\cdot 100\%$ AND THE SAMPLE SIZE, n.

Student's t (again!)

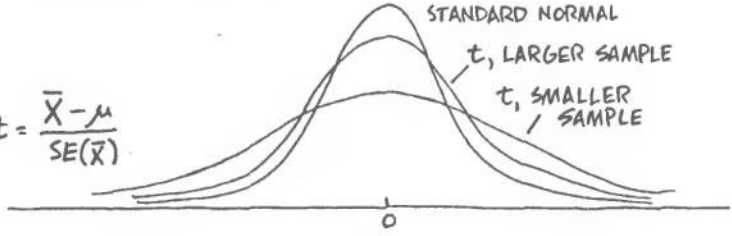
AS WE SAW IN CHAPTER 6, THE STATISTIC

$$\frac{\bar{X}-\mu}{SE(\bar{X})}$$

HAS AN APPROXIMATELY NORMAL DISTRIBUTION ONLY WHEN IT IS COMPUTED USING A *LARGE SAMPLE*. FOR SMALL SAMPLES (n=5, 10, 25...), THIS IS NO LONGER THE CASE, AND WE HAVE TO USE THE STUDENT'S t.

LET'S LOOK AT t A LITTLE MORE CLOSELY. WE MENTIONED THAT THE t DISTRIBUTION IS MORE SPREAD OUT THAN THE NORMAL, AND THAT THE AMOUNT OF SPREAD DEPENDS ON THE SAMPLE SIZE.

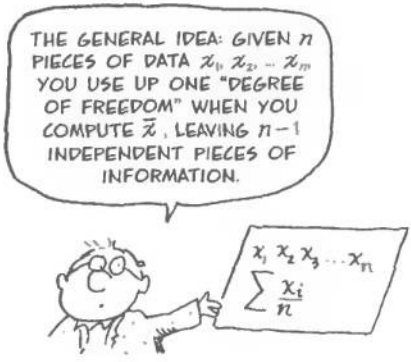
$$t = \frac{\bar{X}-\mu}{SE(\bar{X})}$$



WHAT ITS DISCOVERER GOSSET DID WAS TO *QUANTIFY* THIS RELATIONSHIP. IF n IS THE SAMPLE SIZE, HE SAID, THEN CALL $n-1$ THE NUMBER OF

degrees of freedom

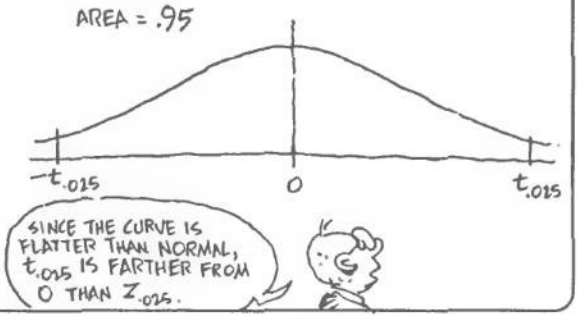
OF THE SAMPLE.



GOSSET COMPUTED TABLES OF THE t DISTRIBUTION FOR DIFFERENT SAMPLE SIZES—I.E., DEGREES OF FREEDOM. WE REPEAT, THE *MORE DEGREES OF FREEDOM*, THE CLOSER t BECOMES TO THE *STANDARD NORMAL*.

KNOWING THE SAMPLE SIZE n, WE CHOOSE THE t DISTRIBUTION WITH $n-1$ DEGREES OF FREEDOM.

AS WITH THE z DISTRIBUTION (I.E., THE STANDARD NORMAL), WE GET A 95% CONFIDENCE LEVEL BY FINDING THE CRITICAL VALUE $t_{.025}$ BEYOND WHICH THE AREA UNDER THE CURVE IS .025.



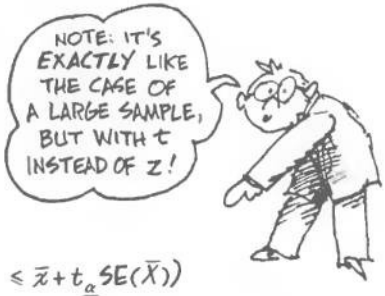
FOR A $(1-\alpha)\cdot 100\%$ CONFIDENCE INTERVAL, WE FIND THE CRITICAL VALUE $t_{\frac{\alpha}{2}}$ SUCH THAT $Pr(t \geq t_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$. HERE IS A SHORT TABLE OF CRITICAL VALUES FOR THE t DISTRIBUTION:

	$1-\alpha$	.80	.90	.95	.99
	α	.20	.10	.05	.01
	$\alpha/2$	.10	.05	.025	.005
DEGREES OF FREEDOM	1	3.09	6.31	12.71	63.66
	10	1.37	1.81	2.23	4.14
	30	1.31	1.70	2.04	2.75
	100	1.29	1.66	1.98	2.63
	∞	1.28	1.65	1.96	2.58

EACH COLUMN REPRESENTS A FIXED LEVEL OF CONFIDENCE, WITH INCREASING NUMBERS OF DEGREES OF FREEDOM. THE HIGHER THE DEGREES OF FREEDOM, THE CLOSER THE CRITICAL VALUE GETS TO $z_{\alpha/2}$, THE CRITICAL VALUE OF THE NORMAL DISTRIBUTION.

WE DERIVE THE WIDTH OF OUR CONFIDENCE INTERVAL DIRECTLY FROM THE DEFINITION OF t:

$$t = \frac{\bar{X} - \mu}{SE(\bar{X})}$$



THEN, FOR CONFIDENCE LEVEL $(1-\alpha)\cdot 100\%$,

$$(1-\alpha) = Pr\left(\bar{x} - t_{\frac{\alpha}{2}} SE(\bar{X}) \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2}} SE(\bar{X})\right)$$

FROM WHICH WE INFER: GIVEN A SINGLE SAMPLE OF SIZE n AND MEAN $\bar{x}$, WE CAN BE $(1-\alpha)\cdot 100\%$ CONFIDENT THAT THE POPULATION MEAN μ FALLS IN THE RANGE

$$\mu = \bar{x} \pm t_{\frac{\alpha}{2}} SE(\bar{x})$$

WHERE $SE(\bar{x}) = s/\sqrt{n}$ AND $t_{\frac{\alpha}{2}}$ IS THE CRITICAL VALUE OF THE t DISTRIBUTION WITH $n-1$ DEGREES OF FREEDOM.

NOTE: STRICTLY SPEAKING, THE DERIVATION OF THE t DISTRIBUTION DEPENDED ON THE ASSUMPTION THAT THE SAMPLE WAS FROM A NORMAL POPULATION. IN PRACTICE, CONFIDENCE INTERVALS BASED ON THE t WORK REASONABLY WELL, EVEN WHEN THE POPULATION DISTRIBUTION IS ONLY APPROXIMATELY MOUND-SHAPED.

example: SUPPOSE *CHAMELEON MOTORS* HAS TO CRASH TEST ITS CARS TO DETERMINE THE AVERAGE REPAIR COST OF A 10 M.P.H. HEAD-ON COLLISION. THIS IS EXPENSIVE! THEY DECIDE TO TRY IT ON JUST *FIVE* CHAMELEONS.

THEY FIND THE DAMAGE DATA TO BE \$150, \$400, \$720, \$500, AND \$930.

THE SAMPLE MEAN:

$$\bar{x} = \$540$$

THE STANDARD DEVIATION:

$$s = \$299$$

YOU CAN CHECK s WITH A HAND CALCULATOR. IT'S

$$\sqrt{\frac{1}{4}\left((150-540)^2+(400-540)^2+(720-540)^2+(500-540)^2+(930-540)^2\right)}$$

SO WHERE CAN WE PLACE THE MEAN WITH 95% CONFIDENCE? WE FIND OUR CRITICAL VALUE $t_{.025}$ WITH 4 DEGREES OF FREEDOM:

	$1-\alpha$	.80	.90	.95	.99
	α	.20	.10	.05	.01
	$\alpha/2$	.10	.05	.025	.005
DEGREES OF FREEDOM	1	3.09	6.31	12.71	63.66
	2	1.89	2.92	4.30	9.92
	3	1.64	2.35	3.18	5.84
	4	1.53	2.13	2.78	4.60
	5	1.48	2.01	2.57	4.03

AND PLUG IT IN:

$$\mu = \bar{x} \pm 2.78 \, s/\sqrt{n}$$
$$= 540 \pm 2.78 (299/\sqrt{5})$$
$$= 540 \pm 372$$

SO THE BEST WE CAN SAY WITH 95% CONFIDENCE IS THAT THE AVERAGE DAMAGE WILL LIE BETWEEN $168 AND $912.

THE COMPANY CAN EITHER BE SATISFIED WITH THAT, OR DO FURTHER TESTS...

TO COMPUTE THIS CONFIDENCE INTERVAL USING STUDENT'S t, WE HAVE MADE AN ***UNSTATED ASSUMPTION:*** WE ASSUMED THAT CRASH REPAIR COSTS ARE APPROXIMATELY ***NORMALLY DISTRIBUTED,*** I.E., IF WE CRASHED 1000 CHAMELEONS, THE HISTOGRAM OF REPAIR COSTS WOULD BE SYMMETRICAL AND MOUND-SHAPED. WE ***CAN NOT KNOW THIS*** FROM 5 DATA POINTS ALONE... BUT MAYBE YEARS OF EXPERIENCE WITH EARLIER MODELS PROVIDE NORMALLY DISTRIBUTED COST HISTOGRAMS FOR FRONT END REPAIRS: INFORMATION WHICH WOULD TEND TO SUPPORT OUR USE OF STUDENT'S t.

TO SUM UP (!), WE NOW HAVE THREE SIMPLE RECIPES FOR FINDING CONFIDENCE INTERVALS. FOR PROPORTIONS, OR MEANS WITH LARGE SAMPLE SIZES, WE LOOK UP $z_{\frac{\alpha}{2}}$ IN A NORMAL TABLE. FOR MEANS OF SMALL SAMPLE SIZES (SAY $n \leq 30$), WE FIND $t_{\frac{\alpha}{2}}$ IN THE t TABLE.

IN ALL CASES, THE WIDTH OF THE INTERVAL IS THAT CRITICAL VALUE TIMES THE STANDARD ERROR:

$$z_{\frac{\alpha}{2}} SE(\hat{p}) \qquad z_{\frac{\alpha}{2}} SE(\overline{X}) \qquad t_{\frac{\alpha}{2}} SE(\overline{X})$$

AND EACH OF THOSE STANDARD ERRORS IS PROPORTIONAL TO THAT MAGIC NUMBER:

♦Chapter 8♦
HYPOTHESIS TESTING

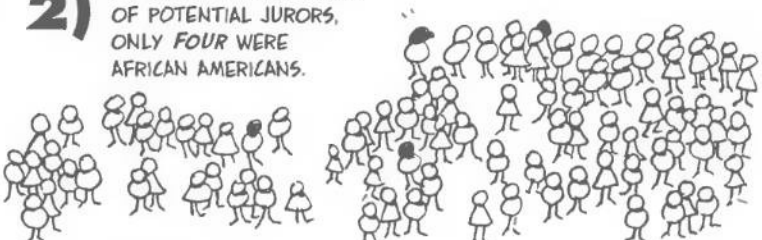
NOW WE ENTER A NEW AREA... GOVERNMENT, BUSINESS, AND THE HARD AND SOFT SCIENCES ALL USE AND OFTEN ABUSE THESE TESTS OF SIGNIFICANCE. IT'S ALL ABOUT ANSWERING THE QUESTION, *"COULD THESE OBSERVATIONS REALLY HAVE OCCURRED BY CHANCE?"*

WE BEGIN WITH AN EXAMPLE FROM THE LAW: A COMPOSITE OF SEVERAL CASES ARGUED IN THE SOUTH BETWEEN 1960 AND 1980, IN WHICH EXPERT WITNESSES PRESENTED THE CASE FOR *RACIAL BIAS IN JURY SELECTION.*

PANELS OF JURORS ARE THEORETICALLY DRAWN AT RANDOM FROM A LIST OF ELIGIBLE CITIZENS. HOWEVER, IN SOUTHERN STATES IN THE '50S AND '60S, FEW AFRICAN AMERICANS WERE FOUND ON JURY PANELS, SO SOME DEFENDANTS CHALLENGED THE VERDICTS. ON APPEAL, AN EXPERT STATISTICAL WITNESS GAVE THIS EVIDENCE:

1) 50% OF ELIGIBLE CITIZENS WERE AFRICAN AMERICAN.

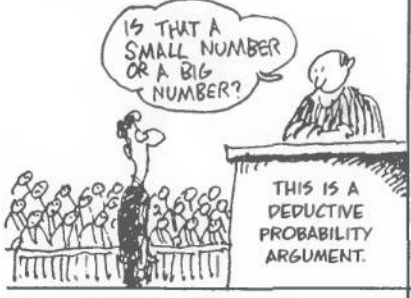
2) ON AN 80-PERSON PANEL OF POTENTIAL JURORS, ONLY *FOUR* WERE AFRICAN AMERICANS.



COULD THIS BE THE RESULT OF *PURE CHANCE?*

FOR THE SAKE OF ARGUMENT, SUPPOSE THAT THE SELECTION OF POTENTIAL JURORS WAS *RANDOM*. THEN THE NUMBER OF AFRICAN AMERICANS ON THE 80-PERSON PANEL WOULD BE THE *BINOMIAL* RANDOM VARIABLE X WITH $n = 80$ TRIALS AND $p = .5$.

THUS, THE CHANCES OF GETTING A JURY WITH ONLY 4 AFRICAN AMERICANS IS $Pr(X \leq 4)$, WHICH WORKS OUT TO ABOUT .00000000000000000014 (!).



SINCE THE PROBABILITY IS SO SMALL, THE PARTICULAR PANEL WITH ONLY FOUR BLACK MEMBERS IS *STRONG EVIDENCE* AGAINST THE *HYPOTHESIS* OF RANDOM SELECTION.

TO DRIVE THE POINT HOME, THE STATISTICIAN NOTES THAT THIS PROBABILITY IS LESS THAN THE CHANCES OF GETTING *THREE CONSECUTIVE ROYAL FLUSHES* IN POKER.

SO THE JUDGE *REJECTS THE HYPOTHESIS* OF RANDOM SELECTION.