CHAPTER 7:
SAMPLING DISTRIBUTIONS

IN THE ASSIGNMENTS, WE USUALLY SAMPLE FROM A DISTRIBUTION (HAVING A FEW VALUES WITH SPECIFIC PROBABILITIES), IN 'REAL LIFE', SAMPLING IS USUALLY DONE FROM A POPULATION.

WE SHOULD REALIZE THAT, IN THIS CONTEXT, 'POPULATION' IS JUST A SPECIAL CASE OF 'DISTRIBUTION' (HAVING A HUGE NUMBER OF VALUES, ALL HAVING THE SAME PROBABILITY OF BEING SELECTED).

POPULATION (I.E. THE CORRESPONDING HISTOGRAM) CAN BE OF ANY SHAPE (SOMETIMES RESEMBLING THE NORMAL CURVE, BUT OFTEN FAR FROM IT).

FROM NOW ON, WE CONCENTRATE ON THE ISSUES RELATED TO SAMPLING FROM A POPULATION .

ANY QUANTITY WE COMPUTE BASED ON THE SAMPLE VALUES (E.G. THE SAMPLE MEAN, SAMPLE STANDARD DEVIATION) IS CALLED A **STATISTIC** (CLEARLY, A RANDOM VARIABLE WITH ITS OWN DISTRIBUTION - THESE ARE CALLED **SAMPLING DISTRIBUTIONS**).

THE MOST IMPORTANT RESULT CONCERNS THE SAMPLE MEAN $\bar{x}$. IT GOES UNDER THE NAME OF **CENTRAL LIMIT THEOREM**:

WHEN THE SAMPLE SIZE IS LARGE ($n > 30$), THE DISTRIBUTION OF THE SAMPLE MEAN IS, TO A GOOD APPROXIMATION, **NORMAL**, REGARDLESS OF THE SHAPE OF THE SAMPLED POPULATION.

ITS MEAN IS EQUAL TO :, ITS STANDARD DEVIATION IS $\frac{\sigma}{\sqrt{n}}$ (ALSO CALLED, IN THIS CONTEXT, THE **STANDARD ERROR** OF $\bar{x}$), WHERE : AND F ARE THE POPULATION'S MEAN AND STANDARD DEVIATION.

I WOULD LIKE TO EMPHASIZE AGAIN THAT THIS IS TRUE FOR A POPULATION OF ANY SHAPE (EVEN WHEN FAR FROM NORMAL). WHEN THE POPULATION ITSELF IS NORMAL, THE ABOVE STATEMENT IS EXACT, FOR ANY VALUE OF *n* (EVEN WHEN SMALL).

EXAMPLE: SAMPLING FROM A DISTRIBUTION WITH ： = 12.7 AND F = 3.2 (LET THE SAMPLE SIZE BE EQUAL TO 75), FIND THE PROBABILITY THAT THE SAMPLE MEAN WILL BE IN THE 12.6 TO 12.8 RANGE.

$$P(12.6<\bar{x}<12.8)=P\left(\frac{12.6-12.7}{\frac{3.2}{\sqrt{75}}}<\frac{\bar{x}-12.7}{\frac{3.2}{\sqrt{75}}}<\frac{12.8-12.7}{\frac{3.2}{\sqrt{75}}}\right)$$

$$=P(-0.27<Z<0.27)=0.6064-0.3936=21.28\%$$

THIS ANSWER IS ONLY APPROXIMATE (WITH A GOOD ACCURACY, THOUGH), UNLESS THE DISTRIBUTION FROM WHICH WE ARE SAMPLING IS ITSELF NORMAL (THE COMPUTATION WILL BE THE SAME).

SOMETIMES, THE QUESTION MAY CONCERN THE SAMPLE TOTAL RATHER THAN THE SAMPLE MEAN.

WE CAN DEAL WITH IT EASILY, SINCE THE TOTAL IS THE MEAN MULTIPLIED BY *n*.

EXAMPLE: LET A SAMPLE OF SIZE 12 BE TAKEN FROM A NORMAL POPULATION WITH A MEAN OF 35 LB. AND A STANDARD DEVIATION OF 4 LB. WHAT IS THE PROBABILITY THAT THE TOTAL OF THESE 12 VALUES WILL EXCEED 450 LB.

$P(x_1+x_2+x_3+....+x_{12} > 450) = P(\bar{x} > 450/12) = P(\bar{x} > 37.5) =$

$P(\frac{\bar{x}-35}{4/\sqrt{12}} > \frac{37.5-35}{4/\sqrt{12}}) = P(Z > 2.165) = 1.0000 - 0.9848 = 1.52\%$

THE NEXT EXAMPLE WILL BE A BIT MORE INVOLVED: LET US ASSUME THAT WE SAMPLE, INDEPENDENTLY, 50 TIMES FROM THE FOLLOWING DISTRIBUTION:

$X =$	-2	-1	0	1
Pr:	0.18	0.26	0.32	0.24

(IN MINITAB, WE WOULD STORE THE VALUES IN C1, THE PROBABILITIES IN C2, AND TYPE: >RANDOM 50 C3; >DISCRETE C1 C2.) WHAT IS THE PROBABILITY THAT THE SUM OF THE NUMBERS (>SUM C3) IS NEGATIVE?

WE MUST FIRST COMPUTE THE MEAN AND 'SIGMA' OF THE DISTRIBUTION, THUS: $\mu = -2 \times 0.18 - 1 \times 0.26 + 0 \times 0.32 + 1 \times 0.24 = -0.38$ AND

$\sigma = \sqrt{4 \times 0.18 + 1 \times 0.26 + 0 \times 0.32 + 1 \times 0.24 - (-0.38)^2} = 1.03711$

THEN, $\Pr(x_1 + x_2 + x_3 + \ldots + x_{50} < -0.5) = \Pr(\bar{x} < -0.01) =$

$$\Pr\left(\frac{\bar{x}-(-0.38)}{1.03711/\sqrt{50}} < \frac{-0.01-(-0.38)}{1.03711/\sqrt{50}}\right) = \Pr(Z < 2.52) = 99.41\ \%$$

(ALMOST CERTAIN - TRY IT). NOTE THE USE OF CONTINUITY CORRECTION.

TWO MORE EXAMPLES:

CONSIDER A (RANDOM INDEPENDENT) SAMPLE OF SIZE 58 FROM A POPULATION (OF ANY SHAPE) WITH : = -13.4 AND F = 7.3 WHAT IS THE PROBABILITY THAT THE SAMPLE MEAN WILL BE BIGGER (HIGHER) THAN -15.0 ?

$$\Pr(\bar{x} > -15) = \Pr\left[\frac{\bar{x}-(-13.4)}{7.3/\sqrt{58}} > \frac{-15-(-13.4)}{7.3/\sqrt{58}}\right] \quad . \quad \Pr(Z > -1.67) =$$

1.0000 - 0.0475 = 95.25%

ROLLING A DIE 100 TIMES, WHAT IS THE PROBABILITY THAT THE AVERAGE NUMBER OF DOTS OBTAINED WILL BE BIGGER THAN 4 ?

WE KNOW THAT THE DISTRIBUTION OF THE # OF DOTS (ROLLING ONCE) HAS THE MEAN OF 3.5, WE ALSO NEED F =

$$\sqrt{\frac{1+4+9+16+25+36}{6} - \left(\frac{7}{2}\right)^2} = \sqrt{\frac{35}{12}}$$

$$\Pr(\bar{x} > 4) = \Pr\left(\frac{\bar{x}-3.5}{\sqrt{\frac{35}{12}}} \times 10 > \frac{4-3.5}{\sqrt{\frac{35}{12}}} \times 10\right) \quad . \quad \Pr(Z > 2.93) = 0.17\%$$

CONSIDER A BINOMIAL EXPERIMENT (SUCCESS / FAILURE) OF n TRIALS. LET r BE THE NUMBER OF SUCCESSES ONE OBTAINS (A RANDOM VARIABLE HAVING THE BINOMIAL DISTRIBUTION, AS WE ALREADY KNOW). DIVIDING IT BY n RESULTS IN A SO CALLED **SAMPLE PROPORTION** $\hat{p} \equiv \frac{r}{n}$

ITS DISTRIBUTION IS STILL BINOMIAL (ONLY THE VALUES ARE NOW REDUCED BY THE $1/n$ FACTOR) WITH THE MEAN OF $np / n = p$ AND THE STANDARD DEVIATION OF $\sqrt{npq}\,/n = \sqrt{\frac{pq}{n}}$.

WE ALSO KNOW THAT, WHEN BOTH $np > 5$ AND $nq > 5$, THIS DISTRIBUTION IS, TO A GOOD APPROXIMATION, NORMAL.

EXAMPLE: THE PERCENTAGE OF PEOPLE EXPERIENCING AN ADVERSE REACTION TO VACCINATION IS 22%. IF 47 PEOPLE ARE TO BE VACCINATED, WHAT IS THE PROBABILITY THAT MORE THAN 30% OF THEM WILL EXPERIENCE AN ADVERSE REACTION?

TO ANSWER THIS QUESTION, WE WILL USE NORMAL APPROXIMATION. TO SIMPLIFY THINGS, WE WILL IGNORE CONTINUITY CORRECTION.

$: = 0.22$ $\qquad$ $F = \sqrt{\frac{0.22 \times 0.78}{47}} = 0.060424$

$$\Pr(\hat{p} > 0.30) = \Pr\left(\frac{\hat{p} - 0.22}{0.060424} > \frac{0.30 - 0.22}{0.060424} \right) \approx \Pr(Z > 1.32) =$$

1.0000 - .9066 = 9.34%

SO FAR IN THIS COURSE, WE HAVE DISCUSSED ONLY 'DESCRIPTIVE STATISTICS' (HOW TO ORGANIZE AND PRESENT DATA), FOLLOWED BY A FEW CHAPTERS ON 'PROBABILITY', WHERE WE ASSUMED THAT ALL THE **PARAMETERS** OF A DISTRIBUTION (SUCH AS $:$ AND F OF THE NORMAL DISTRIBUTION) ARE KNOWN, AND WE WERE ASKED TO COMPUTE THE ODDS (PROBABILITIES) OF WHAT THE EXPERIMENT (YET TO BE CARRIED OUT) WILL YIELD.

‘STATISTICS’ DOES THE REVERSE: THE EXPERIMENT HAS BEEN DONE AND ITS OUTCOME(S) RECORDED. BASED ON THIS, WE WANT TO **ESTIMATE** THE VALUE OF THE UNKNOWN PARAMETER(S) AND, IF REQUIRED, MAKE A RELATED **DECISION** (E.G. SWITCH TO A SUPPLIER WHO’S PRODUCT IS BETTER THAN THE COMPETITION’S).

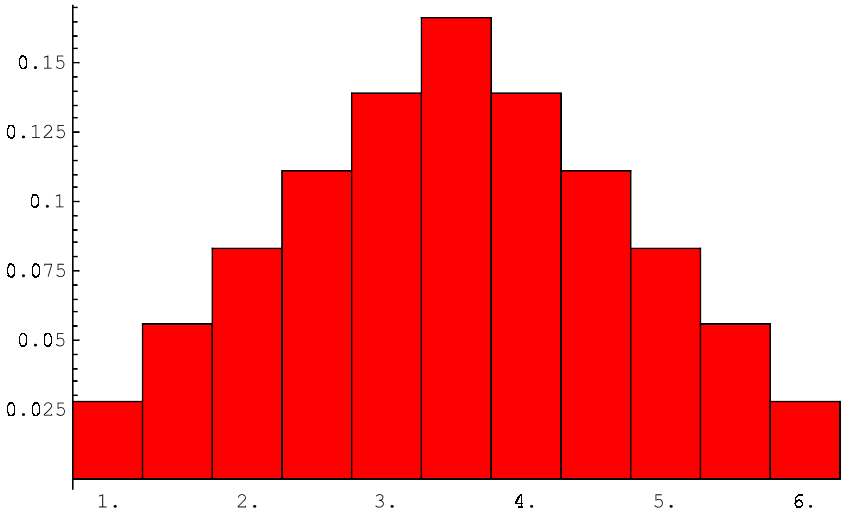
THIS IS THE APPROACH WE TAKE IN ALL SUBSEQUENT CHAPTERS - IT REQUIRES A QUALITATIVE CHANGE (ALMOST A REVERSAL) OF THE WAY WE PROCEED WHEN FACED WITH RESULTS OF A RANDOM EXPERIMENT - IT IS NOW LONGER A QUESTION OF FINDING THE ODDS OF SOMETHING WHICH HAS ALREADY HAPPENED; NOW WE WANT TO UTILIZE THE OUTCOME TO LEARN MORE ABOUT THE POPULATION FROM WHICH THE SAMPLE HAS BEEN DRAWN.

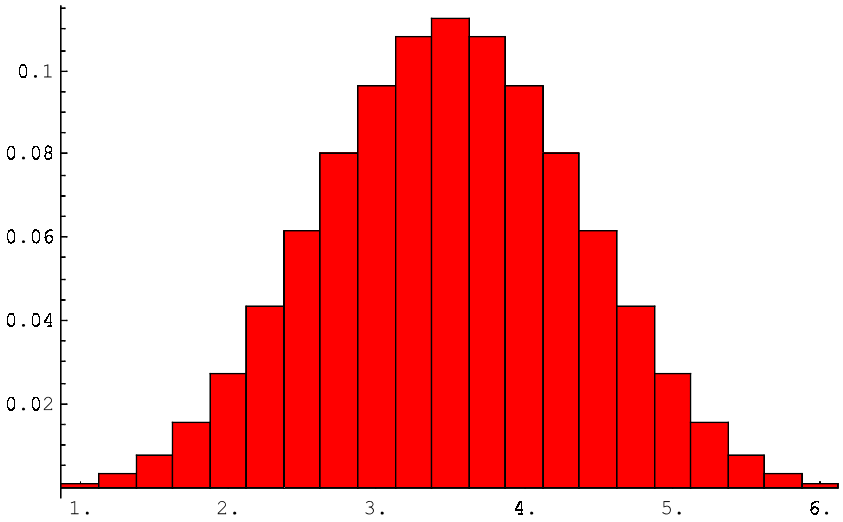
APPENDIX

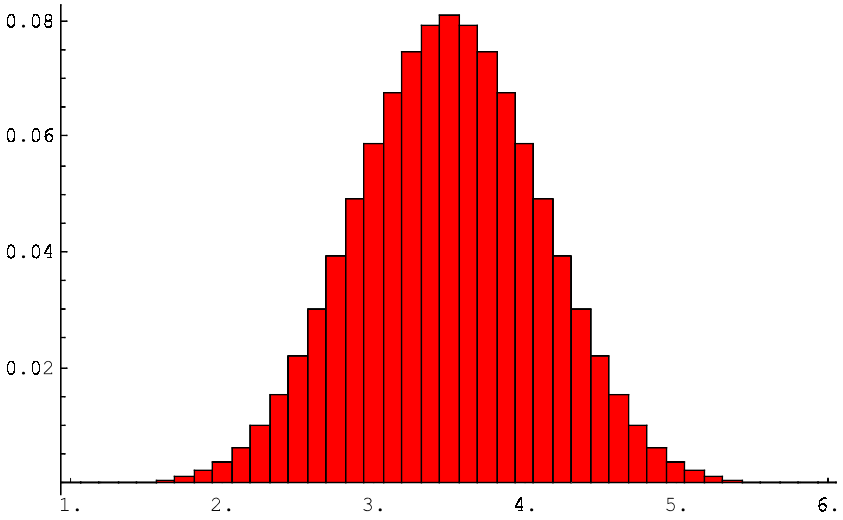
TO BETTER UNDERSTAND THE CENTRAL LIMIT THEOREM, LET US DISPLAY THE DISTRIBUTION'S HISTOGRAM FOR THE SAMPLE MEAN OF THE # OF DOTS OBTAINED IN 2, 4, 8, 16, 32 AND 64 ROLLS OF A REGULAR DIE (FOR 2 ROLLS, WE DERIVED THE CORRESPONDING PROBABILITIES TWO LECTURES AGO - A SIMILAR PROCEDURE WILL EXTEND THIS TO 4, 8, ...).

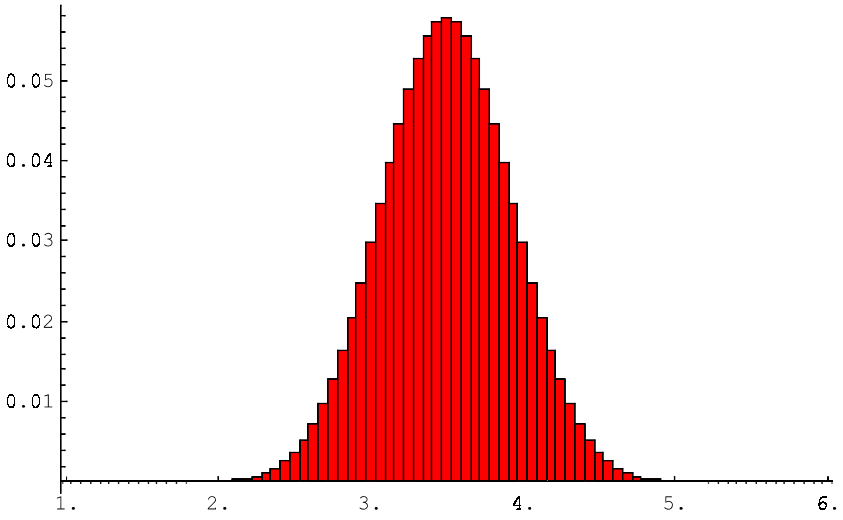
NOTE THAT THE RESULTS HAVE:

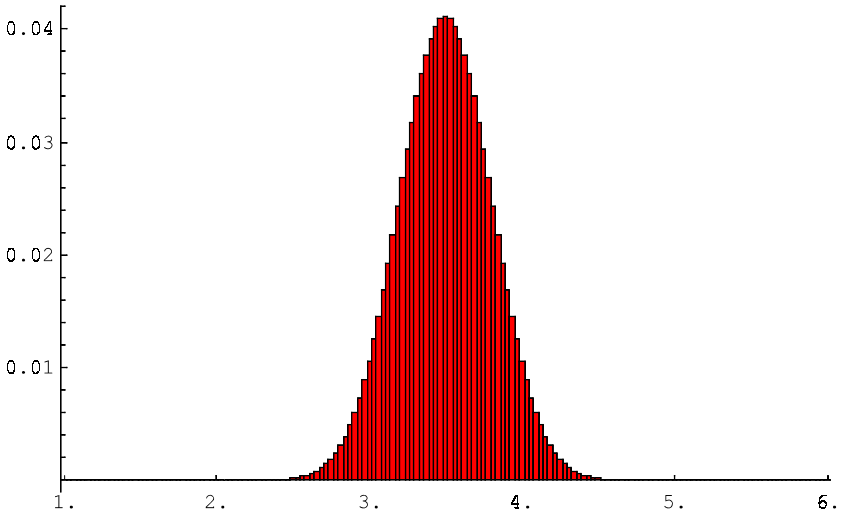
- THE SAME MEAN OF 3.5,
- BUT THEIR STANDARD DEVIATION (WIDTH) DECREASES INVERSELY PROPORTIONAL TO SQUARE ROOT OF *n* (NUMBER OF ROLLS,
- A SHAPE APPROACHING THE NORMAL (BELL SHAPED) CURVE.

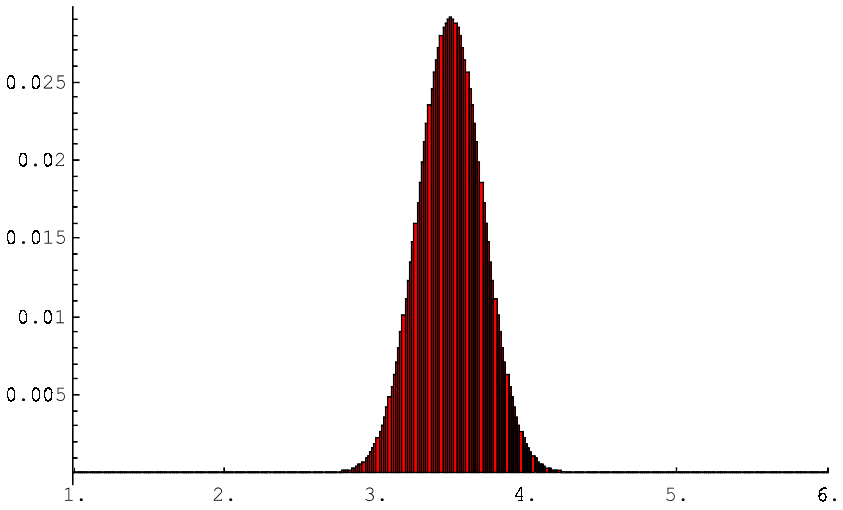

0.15
0.125
0.1
0.075
0.05
0.025
1.
2.
3.
4.
5.
6.


0.1
0.08
0.06
0.04
0.02
1.
2.
3.
4.
5.
6.

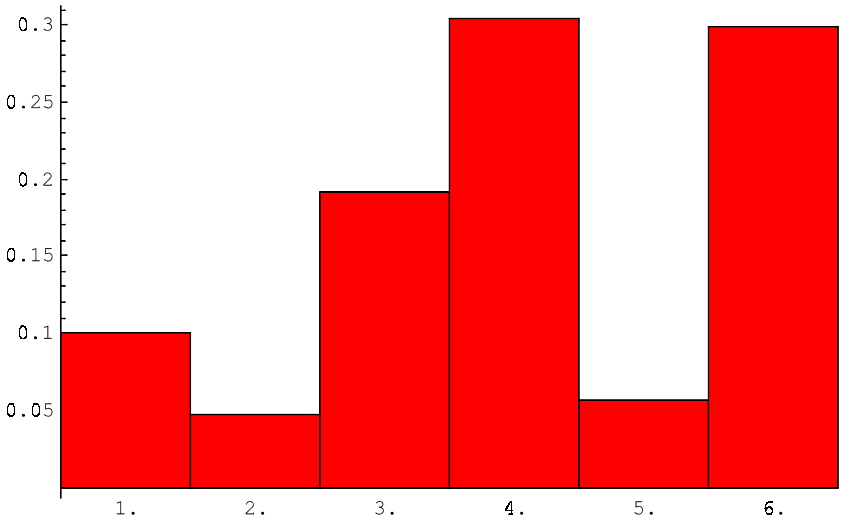

0.08
0.06
0.04
0.02
1.
2.
3.
4.
5.
6.


0.05
0.04
0.03
0.02
0.01
1.
2.
3.
4.
5.
6.


0.04
0.03
0.02
0.01
1.
2.
3.
4.
5.
6.

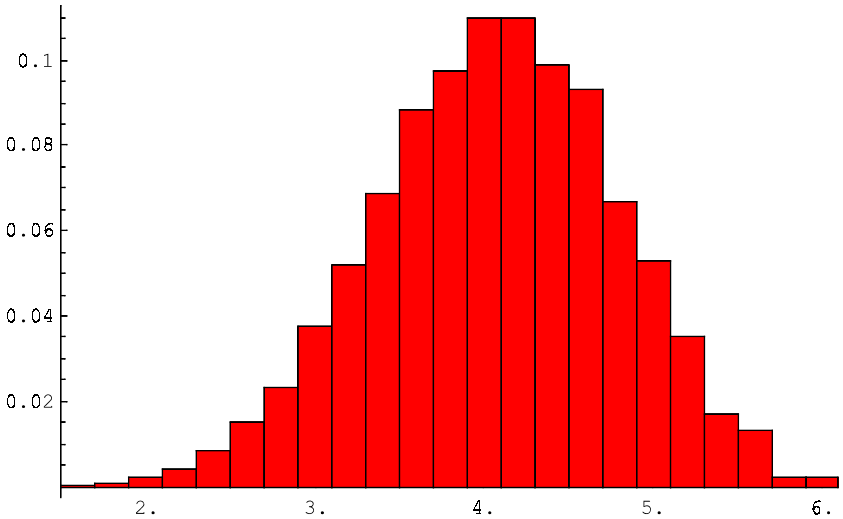

0.025
0.02
0.015
0.01
0.005
1.
2.
3.
4.
5.
6.

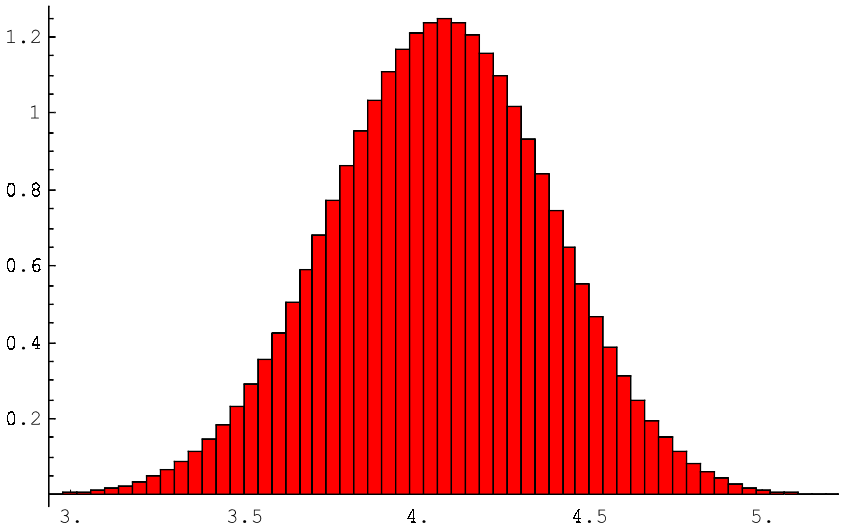
TO SEE THAT THIS IS NOT A COINCIDENCE, WE WILL TRY THE SAME THING WITH A FAIRLY IRREGULAR DISTRIBUTION, SAY:

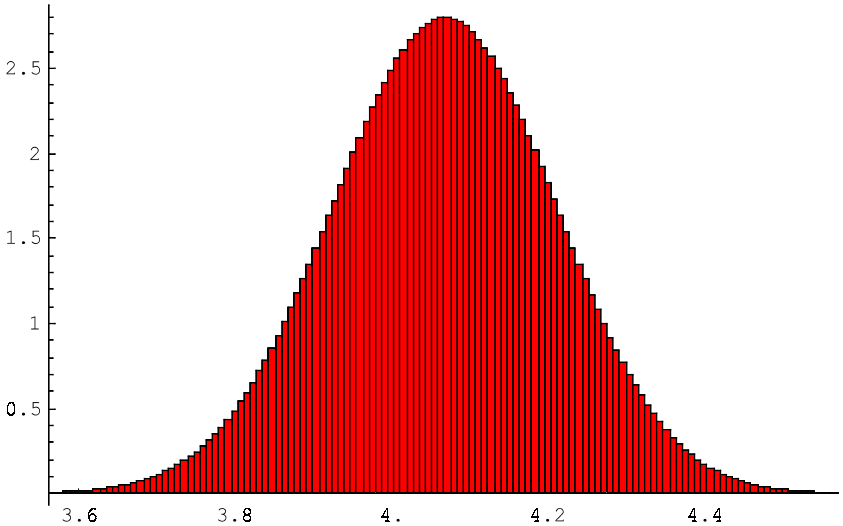


TO SPEED UP THE CONVERGENCE TO (THE PROCESS OF APPROACHING) THE NORMAL DISTRIBUTION, WE WILL DISPLAY THE RESULTS OF $n = 5$, n = 25 AND $n = 125$ ONLY.

THIS TIME, WE WILL ALSO ADJUST THE SCALE OF EACH HISTOGRAM (STRETCHING IT HORIZONTALLY), SO THAT IT DOES NOT BECOME TOO NARROW, AND THE DETAILS OF ITS SHAPE CAN BE MORE READILY OBSERVED.


0.1
0.08
0.06
0.04
0.02
2.
3.
4.
5.
6.


1.2
1
0.8
0.6
0.4
0.2
3.
3.5
4.
4.5
5.


2.5
2
1.5
1
0.5
3.6
3.8
4.
4.2
4.4

LET'S MAGNIFY THE LAST HISTOGRAM: