Probabilistic models, useful to approximate the distribution of biological variables: **binomial** and **gaussian** distributions

Prof. Giuseppe Verlato

Unit of Epidemiology & Medical Statistics, Department of Diagnostics & Public Health, University of Verona

mean and variance

sample with size n	random variable: n?
Mean $\bar{x} = \Sigma x/n$	$\mu = E(X) = \Sigma x^*p$ discrete var. $\int x^*f(x)dx$ continuous var
variance $s^2 = \Sigma(x-\bar{x})^2/(n-1)$	$s^2 = E(X-\mu)^2$
$s^2 = \frac{\Sigma x^2 - (\Sigma x)^2/n}{(n-1)}$	$s^2 = E(X^2) - \mu^2$

THEORETICAL PROBABILITY DISTRIBUTIONS

DISTRIBUTION		expected (mean) E(X)	variance $E[X-E(X)]^2$
Binomial	$p(x)=\frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$	$n\pi$	$n\pi(1-\pi)$
Poisson	$p(x)=\frac{\mu^x e^{-\mu}}{x!}$	$\mu=n\pi$	$\mu=n\pi$
Normal	$f(x)=\frac{1}{\sigma\sqrt{(2\pi)}} e^{-(x-\mu)^2/2\sigma^2}$	μ	σ^2

Exercise on binomial distribution

Random variable = number of spontaneous abortions in 4 pregnancies.

n abortions	0	1	2	3	4
n women	24	28	7	5	6

Are spontaneous abortions (miscarriages) “randomly” distributed among women? or do they tend to concentrate in some women?

↓

If abortions are randomly distributed among women, the random variable “number of spontaneous abortions in 4 pregnancies” should follow the binomial distribution.

Exercise on binomial distribution

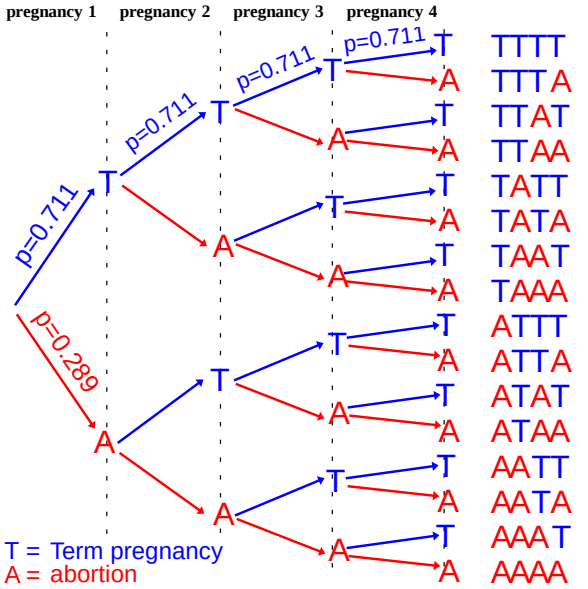
Underlying assumptions of the binomial distribution	**In the example:**
The variable summarizes the number of successes (or failures) in n trials	Number of spontaneous abortions in 4 pregnancies
Each trial results in one of two possible outcomes: success (1) or failure (0)	Each pregnancy can results in a spontaneous abortion (1) or in a living newborn (0)
The probability of success is the same in the whole sample	All women have the same probability of abortion
The replications are independent, i.e. the success in one trial does not affect the probability of success in the subsequent trial	The probability of abortion in one pregnancy is not affected by the outcomes of previous pregnancies

Exercise on binomial distribution

		TOTAL
women =	24 + 28 + 7 + 5 + 6	= 70
pregnancies =	70 * 4	= 280
N abortions =	24*0 + 28*1 + 7*2 + 5*3 + 6*4	= 81

p(abortion) = 81/280 = 0.289

Variable = n abortions in 4 pregnancies



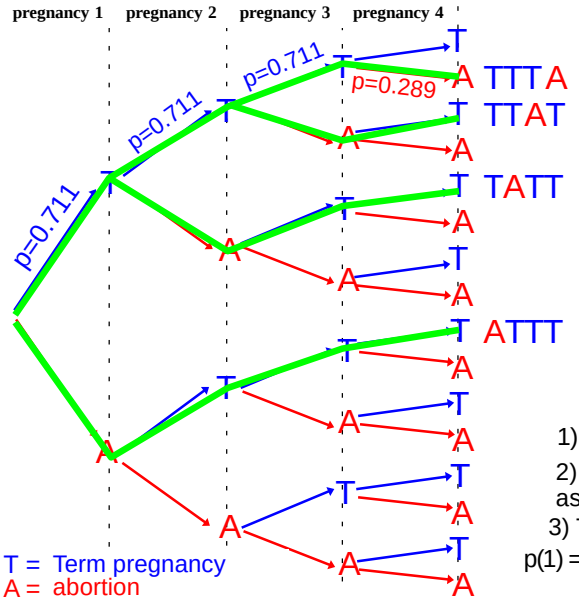
Which is the probability of having ZERO abortions, p(0), p(TTTT)?

If both outcomes were EQUIPROBABLE,
p(0) = 0.5*0.5*0.5*0.5
= 0.0625 = 1/16

Considering that:
1) p(T)=0.711,
2) Trials (pregnancies) are assumed to be independent

p(TTTT) =
=0.711*0.711*0.711*0.711=
= 0.255

Variable = n abortions in 4 pregnancies



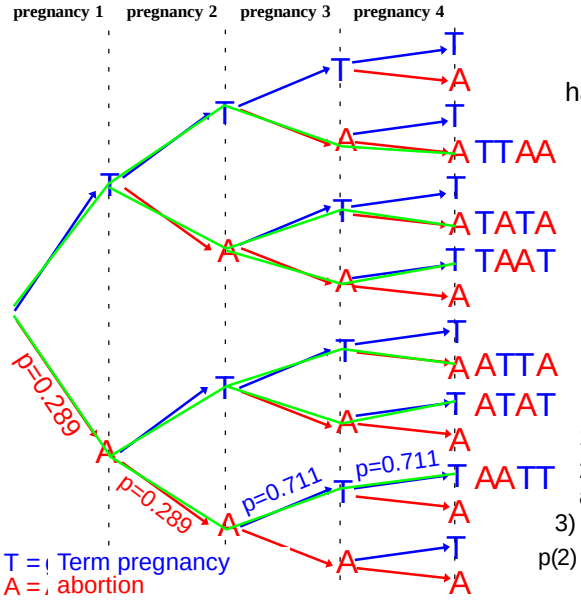
Which is the probability of having ONE abortion, p(1) ?

If both outcomes were EQUIPROBABLE,
p(1) = 4/16 = 1/4

Considering that:
1) p(A)=0.289, p(T)=0.711
2) Trials (pregnancies) are assumed to be independent
3) There are 4 possible paths

p(1) = 4*0.289*0.711*0.711*0.711 =
= 0.415

Variable = n abortions in 4 pregnancies



Which is the probability of having TWO abortions, p(2) ?

If both outcomes were EQUIPROBABLE,

p(2) = 6/16 = 3/8

Considering that:

1) p(A)=0.289, p(T)=0.711

2) Trials (pregnancies) are assumed to be independent

3) There are 6 possible paths

p(2) = 6*0.289*0.289*0.711*0.711 =

= 0.253

Variable = n abortions in 4 pregnancies

p(0) = (0.711*0.711*0.711*0.711) = 0.255

p(1) = 4 * 0.289 * (0.711*0.711*0.711) = 0.415

p(2) = 6 * (0.289*0.289) * (0.711*0.711) = 0.253

Is there a unifying formula ?

$$p(x) = \frac{n!}{x!(n-x)!}\, p^{x} (1-p)^{n-x}$$

Number of possible paths

p(abortion)

p(term pregnancy)

N.B. 4! = 4 factorial = 4* 3 * 2 * 1 = 24

0! = 1

In a binomial distribution $p(x)=\frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$

In the present example: n=4 p=0.289

$p(0)=\frac{4!}{0!\,4!}0.289^0(1-0.289)^4=0.255$

$p(1)=\frac{4!}{1!\,3!}0.289^1(1-0.289)^3=0.415$

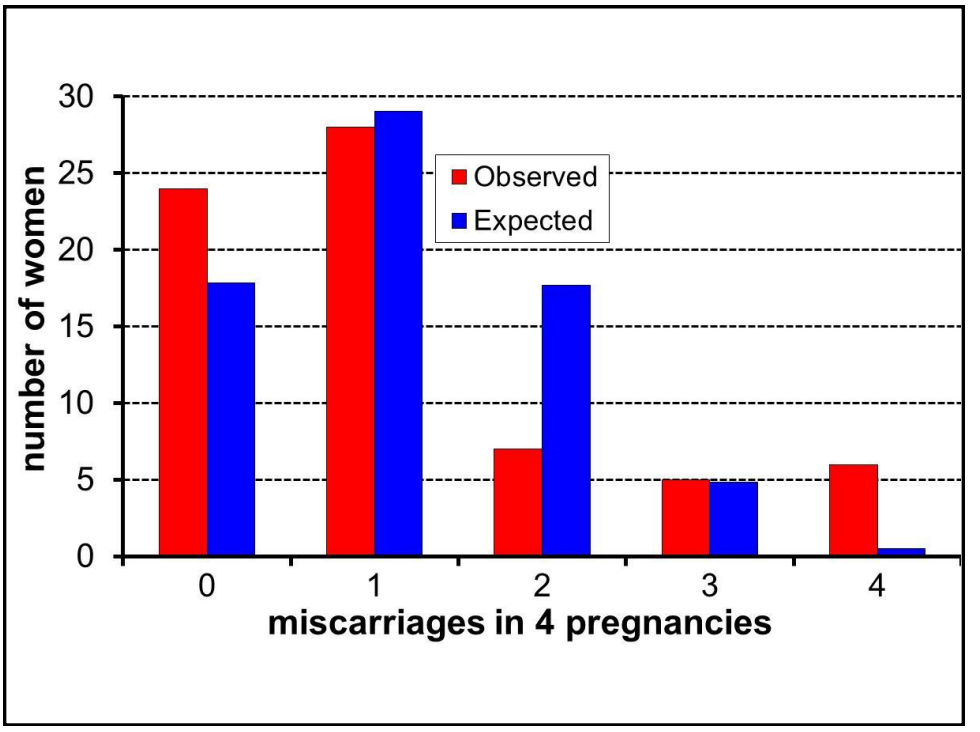
$p(2)=\frac{4!}{2!\,2!}0.289^2(1-0.289)^2=0.253$

$p(3)=\frac{4!}{3!\,1!}0.289^3(1-0.289)^1=0.069$

$p(4)=\frac{4!}{4!\,0!}0.289^4(1-0.289)^0=0.007$

Number of abortions in 4 pregnancies	OBSERVED	EXPECTED if the abortions were randomly distributed
0	24	0.255*70 = 17.85
1	28	0.415 * 70 = 29.05
2	7	0.253 * 70 = 17.71
3	5	0.069 * 70 = 4.83
4	6	0.007 * 70 = 0.49

OBERVED values differ from EXPECTED values: abortions are not randomly distributed among women, but rather they tend to concentrate in some women, for instance because of uterine hypoplasia or enlargement of the uterine cervix.



Chi-square test

$$\chi^2 = \Sigma\,(\text{observed} - \text{expected})^2 / \text{expected}$$

H_0: observed data follow a binomial distribution
H_1: observed data DO NOT follow a binomial distribution

Level of significance = 5%
Degrees of freedom =n° cells - n° parameters = 5-2 = 3

Critical threshold = $\chi^2_{3,\,0.05} = 7.81$

$$\chi^2 = \frac{(24-17.9)^2}{17.9} + \frac{(28-29.1)^2}{29.1} + \frac{(7-17.7)^2}{17.7} + \frac{(5-4.8)^2}{4.8} + \frac{(6-0.5)^2}{0.5}$$

$$= 2.12 + 0.04 + 6.48 + 0.01 + 61.96 = 70.60$$

Observed χ^2 > critical threshold → H_0 is rejected
70.60 7.81

from the BINOMIAL distribution to the POISSON distribution

BINOMIAL distribution

$$p(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

$$p(x) = \frac{n*(n-1)*\ldots*(n-x+1)}{x!} \left(\frac{\mu}{n}\right)^x \left(1-\frac{\mu}{n}\right)^{n-x}$$

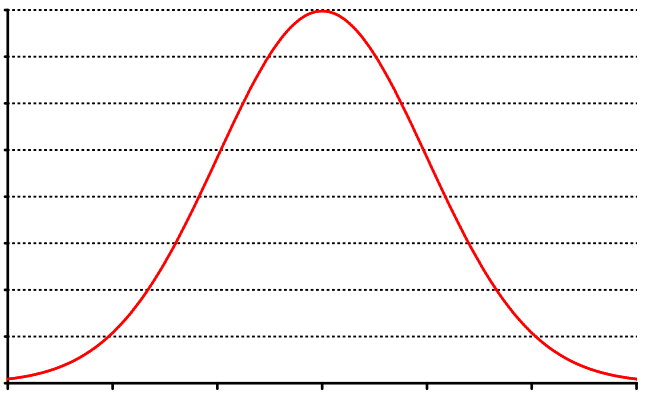
If $n \rightarrow \infty$ and $x \rightarrow 0$

$$p(x) = \frac{n^x}{x!} \frac{\mu^x}{n^x} \left(1-\frac{\mu}{n}\right)^n$$

POISSON distribution

$$p(x) = \frac{\mu^x}{x!} e^{-\mu}$$

The normal distribution, also known as Gaussian distribution or bell curve, is the most important and most widely used distribution in Medical Statistics.



Indeed:

1) If one sums 30 or more random variables, the distribution of the sum will be approximately normal, regardless of the distributions of the original variables (**central limit theorem**).

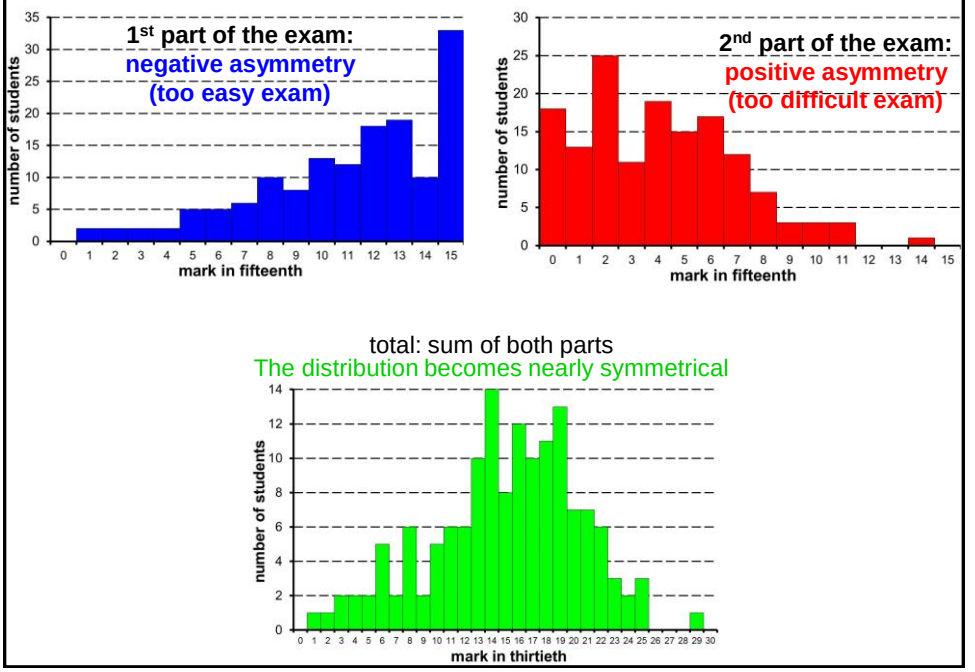
2) Most biological variables (weight, height, …) originate from the sum of several genetic and environmental variables.

↓

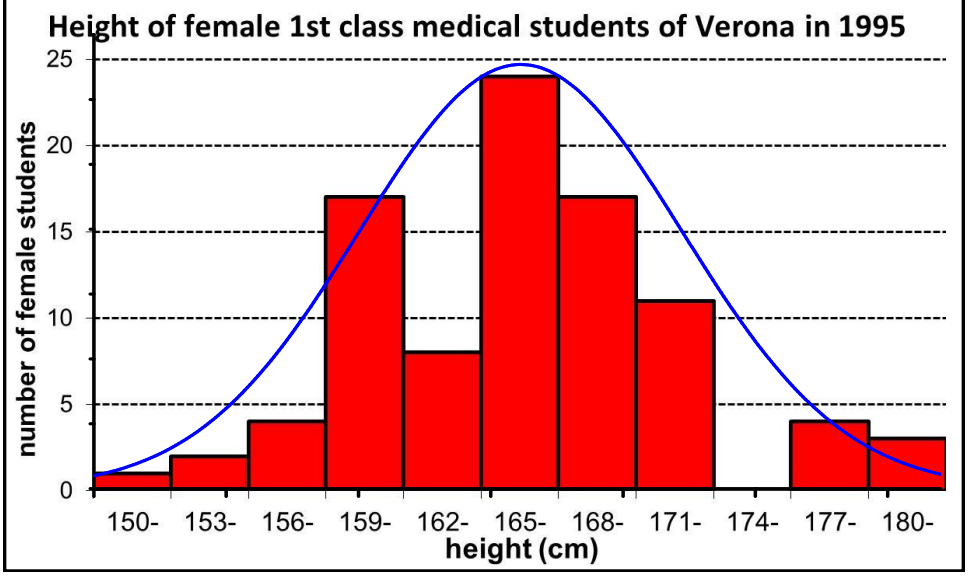
Most biological variables follow the normal distribution.

3) Most theoretical probability distributions (binomial, Poisson, Student's t) converge to the normal distribution, as sample size increases.

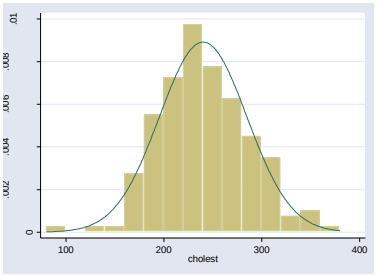
Exam for medical school students in November 1994



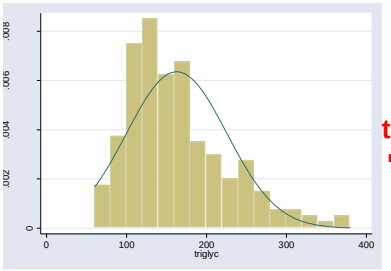
Empirical distribution (red columns) can be approximated by a normal distribution (blue line) with mean=166.1 cm and standard deviation=6.1 cm.



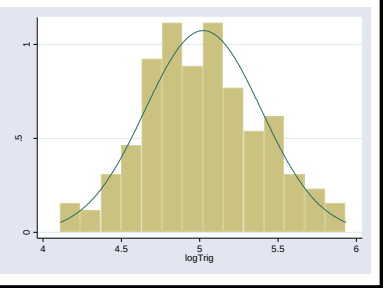
Distribution of total serum [cholesterol] (mg/dL) in 200 type 2 diabetic patients

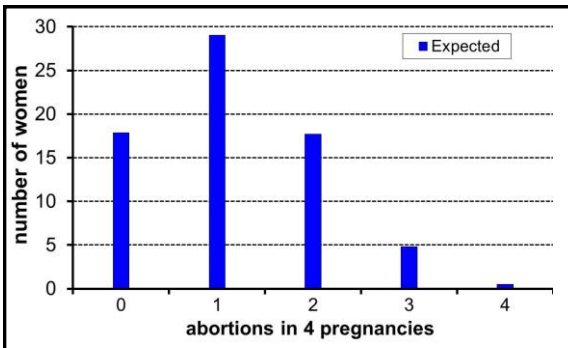


Distribution of serum [triglycerides] in mg/dL in 200 type 2 diabetic patients



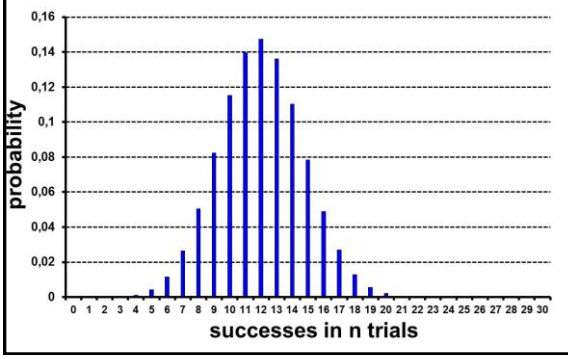
Logarithmic transformation





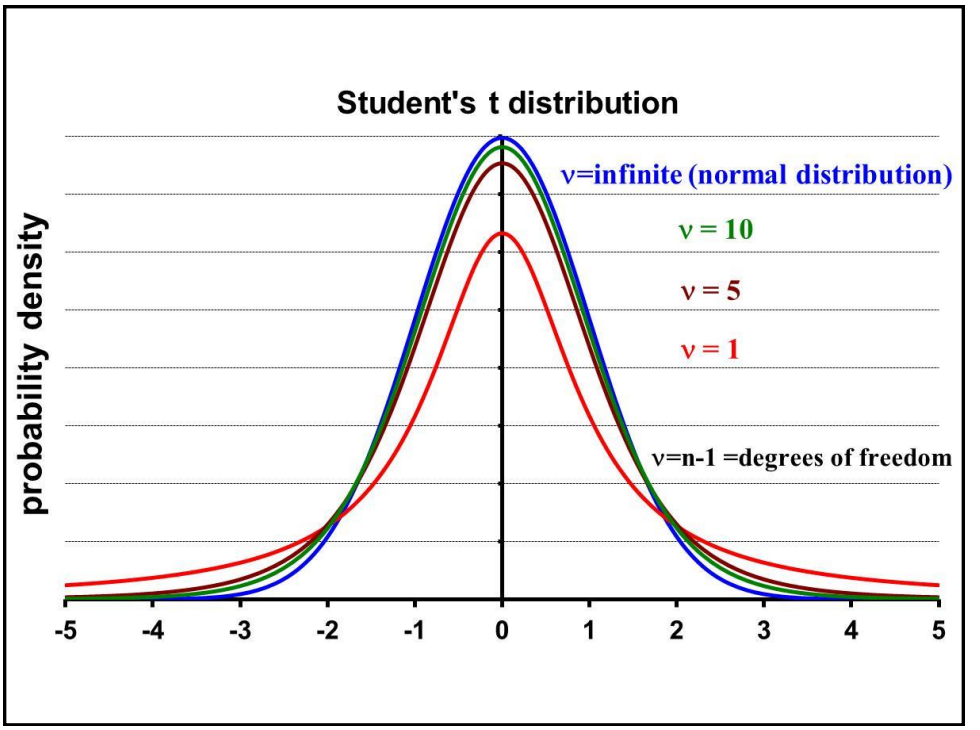
Binomial distribution with π=0.289 ed n=4.

Positively skewed distribution.

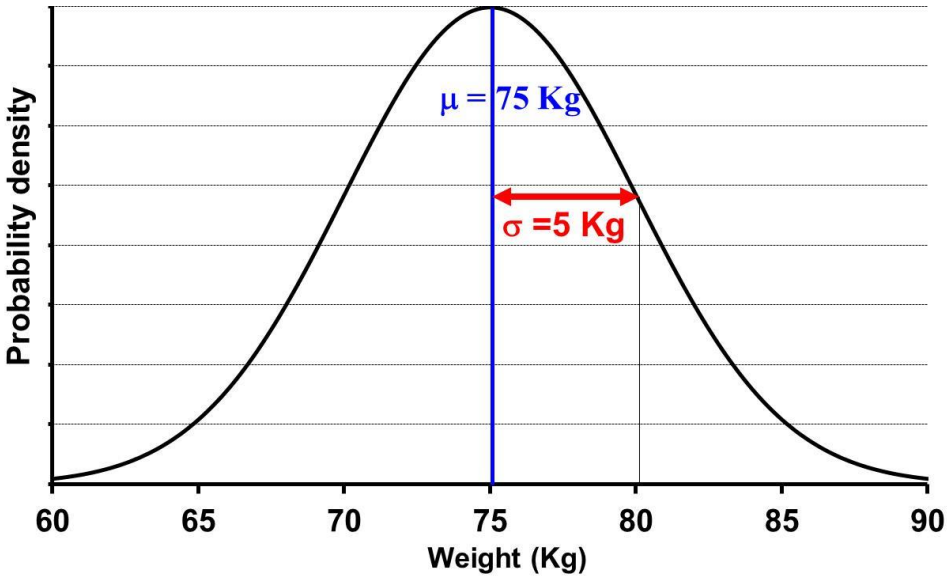


Binomial distribution with π=0.4 ed n=30.

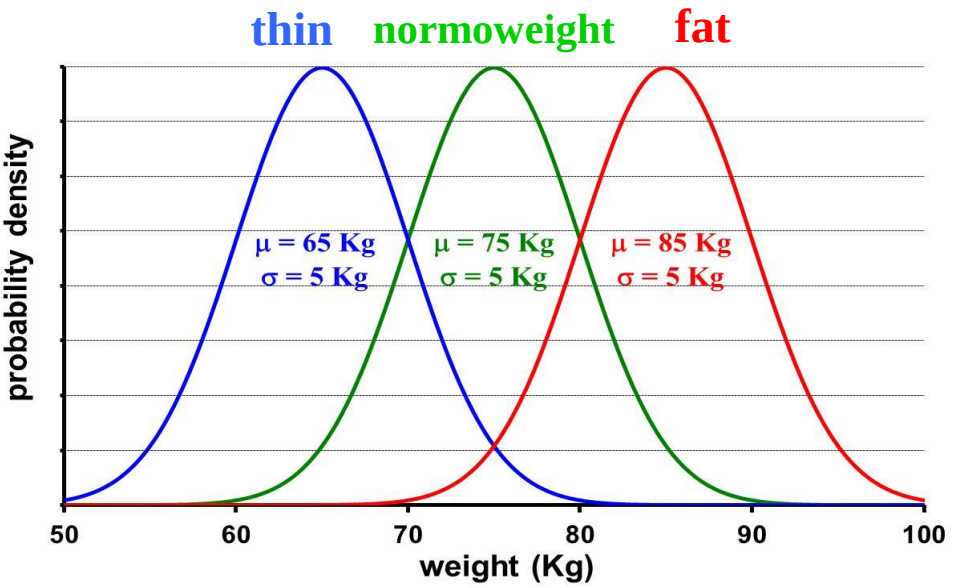
Symmetrical distribution, tending to the normal distribution.



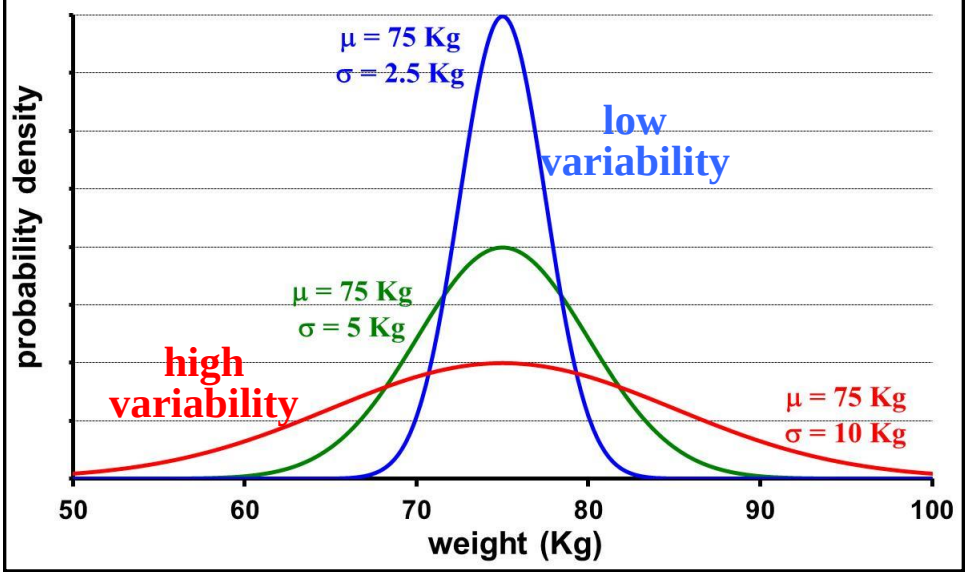
A normal distribution is fully characterized by just two parameters, the mean (μ) and the standard deviation (σ)



These three distributions differ only for the mean μ (measure of central tendency)



These three distributions differ only for the standard deviation σ (measure of variability)

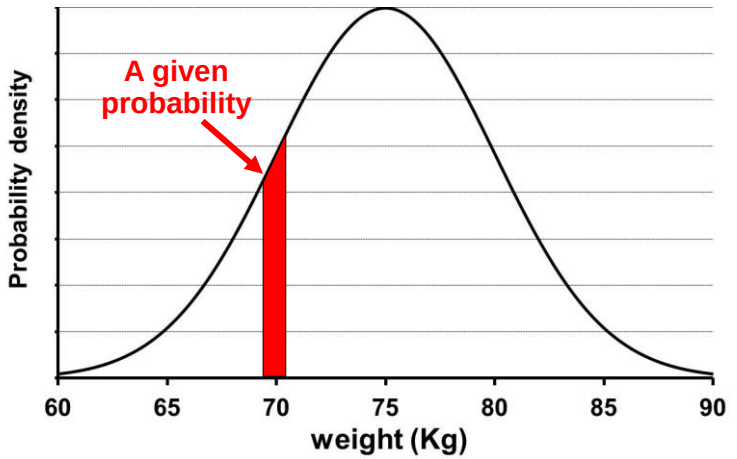


Probability and probability density

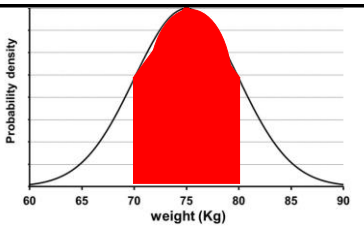
Weight is a quantitative continuous variable.

Which is the probability that the weight of a given individual could be exactly 73Kg 133g 917mg 715μg 822ng? Virtually zero.

It is not possible to attribute a probability to a single value, but rather to an interval.

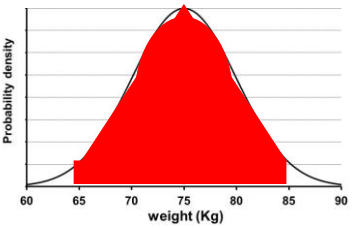


mean ± 1 standard dev. =
75 ± 5 = 70-80 Kg



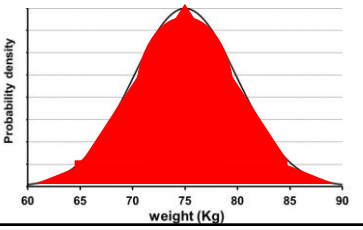
68.26%

mean ± 2 standard dev.=
75 ± 10 = 65-85 Kg



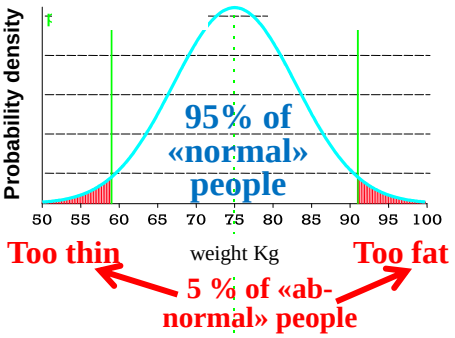
95.44%

mean ± 3 standard dev.=
75 ± 15 = 60-90 Kg

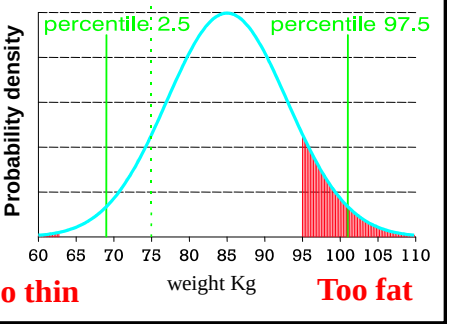


99.74%

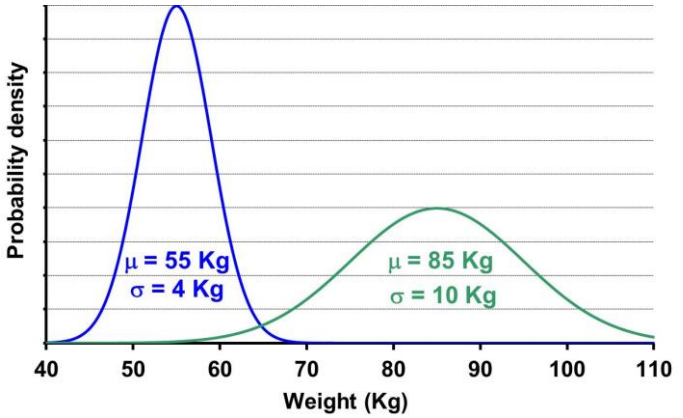
STATISTICAL NORMAL RANGE



CLINICAL NORMAL RANGE:
Population is obese as a whole
"*The American population is constant in number, but it is ballooning in mass*" (CDC, Atlanta, USA)



There is an infinite number of normal distributions, different one from another

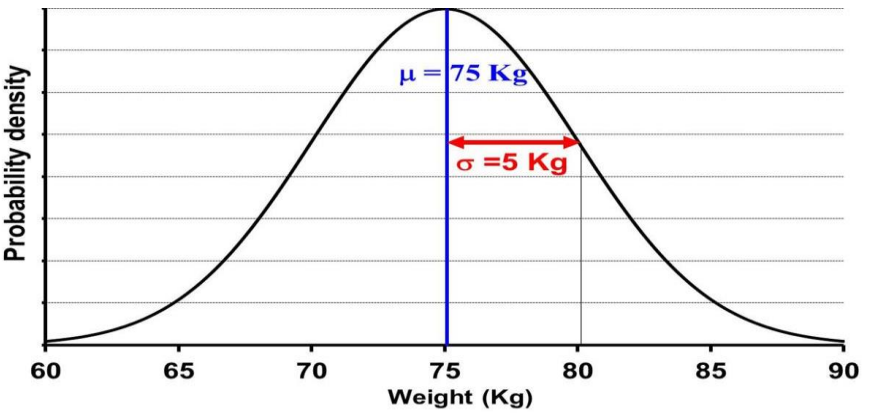


Is it possible to put together all different normal distributions in a single standard distribution ?

Yes, it is possible through the normal transformation

z (normal standard deviate) = $(x-\mu)/\sigma$

z (normal standard deviate) = $(x-\mu)/\sigma$



z	-3	-2	-1	0	+1	+2	+3

weight=60 Kg z = (60-75)/5 = -15/5 = -3

weight=65 Kg z = (65-75)/5 = -10/5 = -2

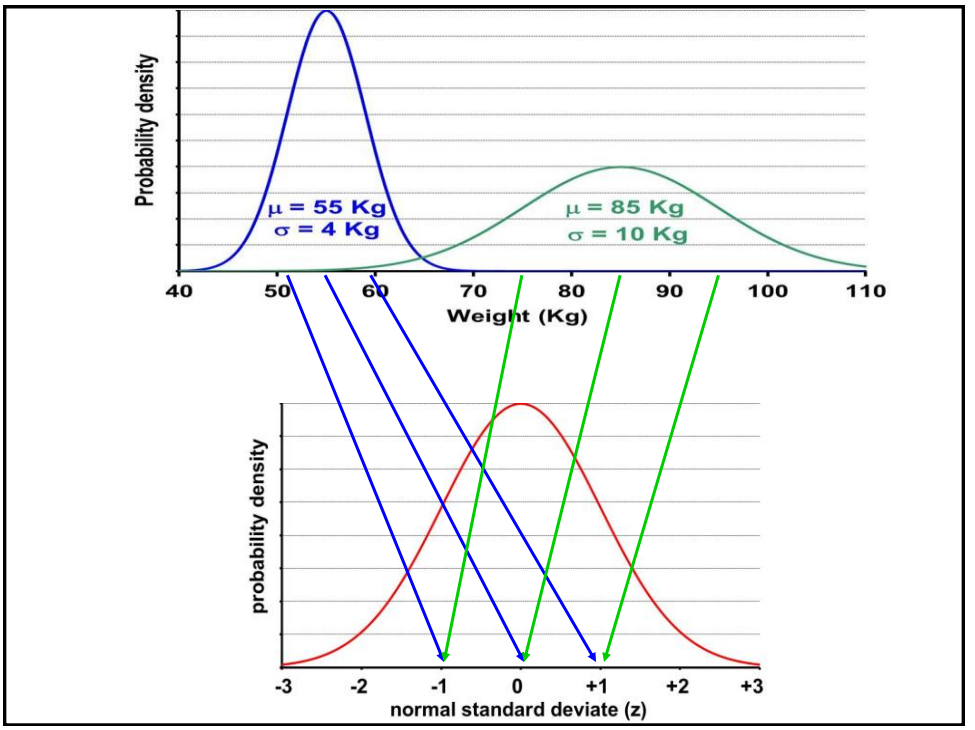
weight=70 Kg z = (70-75)/5 = -5/5 = -1

weight=75 Kg z = (75-75)/5 = 0/5 = 0

weight=80 Kg z = (80-75)/5 = +1

weight=85 Kg z = (85-75)/5 = +2

weight=90 Kg z = (90-75)/5 = +3



This z-table (standard normal distribution table) shows the probability that Z (standard normal VARIABLE) will be greater than (or equal to) a given z-score

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.06430	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233

Which is the probability of Z being greater than (or equal to) 1.87?

0.0307 = 3.07%

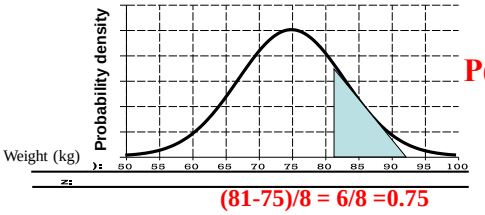
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.06430	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233

Which is the probability of Z being greater than (or equal to) 0.75 ?

0.2266 = 22.66%

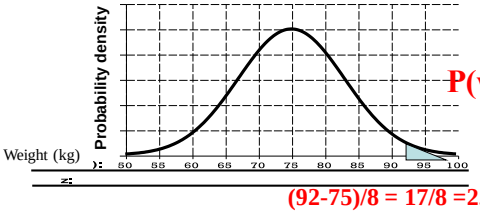
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.06430	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233

NORMAL DISTRIBUTION: mean =75 Kg, standard deviation = 8 Kg



P(weight≥ 81 kg) = 22.66%

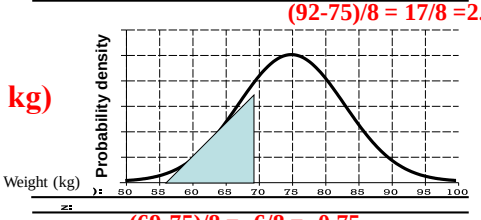
(81-75)/8 = 6/8 =0.75



P(weight ≥ 92 kg) = 1.68%

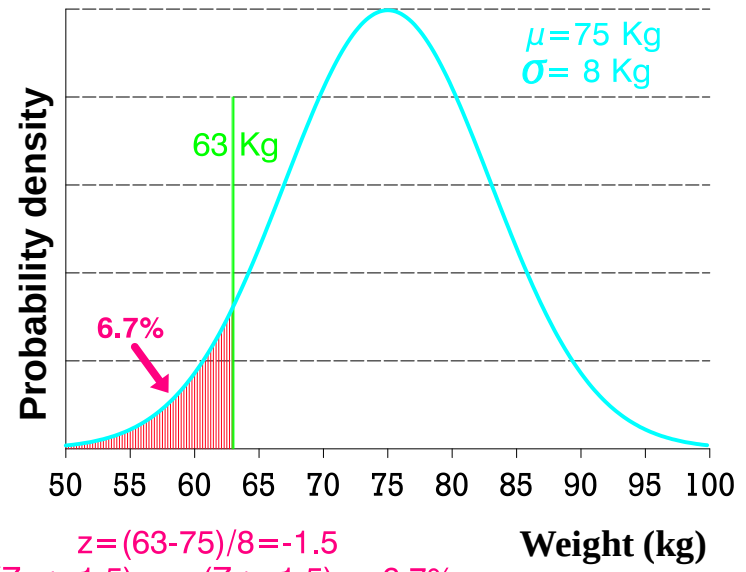
(92-75)/8 = 17/8 =2.125

P(weight≤ 69 kg) = 22.66%



(69-75)/8 = -6/8 = -0.75

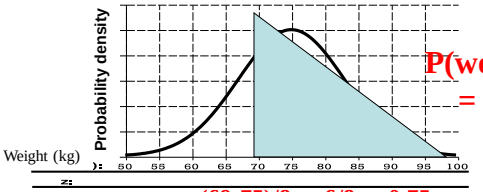
WEIGHT ≤ 63 Kg ?



z=(63-75)/8=-1.5

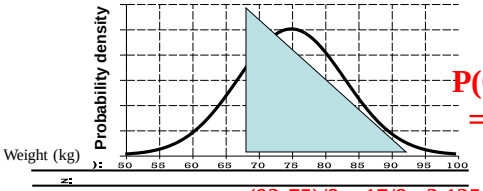
p (Z ≤ -1.5) = p(Z ≥ 1.5) = 6.7%

NORMAL DISTRIBUTION: mean =75 Kg, standard deviation = 8 Kg



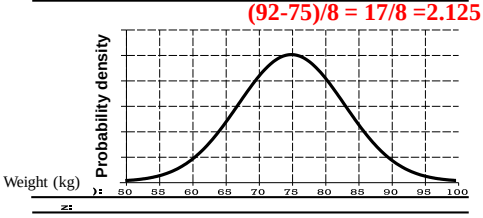
P(weight>=69 kg) = 100-22.66= 77.34%

(69-75)/8 = -6/8 = -0.75

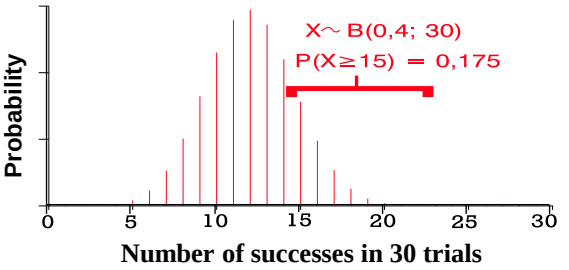


P(69<weight<92 kg) =100-22.66-1.68 = 75.6 %

(92-75)/8 = 17/8 =2.125



CONTINUITY CORRECTION



z= (15-12)/2.68 =1,12 P(Z>=1,12)=0,131

z= (15-12-0,5)/2.68 =0,93 P(Z>=0,93)=0,176