Random variables
Probability distributions

OBJECTIVES

- **Discrete random variables**
- **Continuous random variable**

From the point of view of the study of probabilities, the events are the result of experiments.

Complex natural phenomena have multiple possibilities to obtain one or more results and the quantitative mathematical modeling of the possible results and their probabilistic study is done with the help of random variables.

The random nature of these variables is generated by the differences that appear between the members of the different samples, at each repetition of an experiment.

The name random variable does not imply that the values of the variable (characteristics) are random, but that the experiment is so complex that we cannot anticipate its outcome

Example: we cannot accurately calculate weight at maturity for a newborn.

Random variable

A random variable

- Is a quantification of a probability model that allows to model random data
- Is a quantity that may take any value of a given range that cannot be predicted exactly but can be described regarding their probability

Discrete: generally assessed by counting

Continuous: usually assessed by measurements

Discrete vs. continuous …

Random variables: arithmetic mean, standard deviation, frequency

Discrete:

- Can take a finite number of values
- Examples: The number of peoples with RH- from a sample | The number of children with flue from a collectivity | The number of anorexic students from university | Heartbeat

Continuous:

- Can take an infinite number of values into a defined range
- Vary continuously in a defined range
- Example: Body temperature | Blood sugar concentration | Blood pressure

A random variable X could take certain values with different probabilities

- Pr(X = a) – probability that X to take value a
- Pr(a ≤ X ≤ b) – probability that X to take value in the range [a, b]

Symbols:

$$x : \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ \Pr(x_1) & \Pr(x_2) & \dots & \Pr(x_n) \end{pmatrix}$$

The probabilities that appear in the distribution of a finite random variable verify the formula:

$$\sum_{i=1}^{n} \Pr(x_i) = 1$$

ABO and Rh blood type in Belgium

$$x:\begin{pmatrix} 0+ & A+ & B+ & AB+ & 0- & A- & B- & AB- \\ 0.380 & 0.340 & 0.085 & 0.041 & 0.070 & 0.060 & 0.015 & 0.009 \end{pmatrix}$$

$$\sum_{i=1}^{n} \Pr(x_i) = 0.380 + 0.34 + 0.085 + 0.041 + 0.07 + 0.06 + 0.015 + 0.009 = 1$$

Numerical characteristic of a finite random variable

- The **mean** of a discrete probability distribution (called also **expected value**) is given by the formula:

$$M(x) = \sum_{i=1}^{n} x_i \cdot \Pr(x_i)$$

- Represents the weighted average of possible values, each value being weighted by its probability of occurrence.

Numerical characteristic of a finite random variable

- **Variance**: is a weighted average of the squared deviations in X

$$V(x) = \sum_{i=1}^{n} (x_i - M(x))^2 \cdot \Pr(x_i)$$

- **Standard deviation**:

$$\sigma(x) = \sqrt{V(x)} = \sqrt{\sum_{i=1}^{n} (x_i - M(x))^2 \cdot \Pr(x_i)}$$

A new drug identified by a pharmaceutical company and trial on four subjects was conducted. The probability of not affect is 0.02, to work in one out of 4 cases is 0.05, to work in two subjects is 0.26, in 3 subjects is 0.45 and the probability of working in all subjects is 0.22.

$$x:\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0.02 & 0.05 & 0.26 & 0.45 & 0.22 \end{pmatrix}$$

- For a sample of 1.000 subjects which is the expected number of subjects to which the treatment works?
    - Mean
- Which is the worst scenario?
    - mean-variance
- Which is the best scenario?
    - mean+variance

For a sample of 1.000 subjects which is the expected number of subjects to which the treatment works?

$$x : \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0.02 & 0.05 & 0.26 & 0.45 & 0.22 \end{pmatrix}$$

Expected number of success for a sample of 4 subjects:

M(x) = 0*0.02+1*0.05+2*0.26+3*0.45+4*0.22 = 2.80

Expected number of success for a sample of 1.000 subjects:

2.80*1000/4 = 700

V(x) = $(0-2.8)^2$*0.02 + $(1-2.8)^2$*0.05 + $(2-2.8)^2$*0.26 + $(3-2.8)^2$*0.45 + $(4-2.8)^2$*0.22 = 0.82 ➔ σ(x) = 0.91

- The best scenario: (2.80+0.91)*1000/4 = 927.5
- The worst scenario: (2.80-0.91)*1000/4 = 472.5

Discrete probability distribution

Bernoulli: head versus tail (two possible outcomes)

Binomial: number of ‘head’ obtained by throwing a coin of *n* times

Poisson: number of patients consulted in an emergency office in one day

X	1	0
Pr(X=x)	p	1-p

Bernoulli random variable

The experiment has just two possible outcomes (success and failure – dichotomial variable): gender (boy or girl), results of a test (positive or negative)

- Probability of success = p
- Probability of failure = q = 1-p

Mean of X: $M(x) = 1 \cdot p + 0 \cdot (1-p)$

Variance of X: $V(x) = p \cdot (1-p)$

Mean	M(x) = n·p
Variance	V(x) = n·p·q
Standard deviation	σ(x) = √(n·p·q)

Binomial random variable

A binomial experiment:

1. It consists of a fixed number *n* of identical experiments.
2. There are only two possible outcomes in each trial, a success that occurs with a probability *p* and failure which occurs with a probability q (where q=1-p)
3. The experiments are independent with the same probability of success (denoted p)
4. The number of successes *X* obtained by performing the test *n* times is a random variable of *n* and *p* parameters and is noted as *Bi(n,p)*

The probability that X to be equal with a value k is

$$\Pr(X=k)=C_n^k p^k q^{n-k} \qquad C_n^k=\frac{n!}{k!\cdot(n-k)!}$$

Binomial random variable

Suppose that 90% of adults with joint pain report symptomatic relief with a specific medication. If the medication is given to 10 new independent patients with joint pain, what is the probability that it is effective in exactly 7 subjects?

p=0.9 ➔ q = 1-0.9 = 0.1

$$\Pr(X=7)=C_{10}^{7}p^{7}q^{10-7}=120*0.9^{7}*0.1^{3}=120*0.478*0.001=0.057$$

$$C_{10}^{7}=\frac{10*9*8*7*6*5*4*3*2*1}{(7*6*5*4*3*2*1)*(3*2*1)}=10*3*4=120$$

→ there is a 5.7% chance that exactly 7 of 10 patients will report pain relief when the probability that any one reports relief is 90%

Poisson random variable

Mean	M(x) = θ = n*p
Variance	V(x) = θ
Standard deviation	σ(x) = √θ

Is characterized by theoretical parameter θ (expected average number of achievement for a given event in a given range)

Symbol: Po(θ)

Poisson Distribution:

$$X : \begin{pmatrix} k \\ e^{-\theta} \cdot \frac{\theta^k}{k!} \end{pmatrix} \qquad \Pr(x = k) = \frac{e^{-\theta} \cdot \theta^k}{k!}$$

The mortality rate for specific viral pathology is 7 per 1000 cases (p=7/1000=0.007). What is the probability that in a group of 400 (*n*) people this pathology to cause 5 deaths?

θ=n·p=400*0.007=2.8

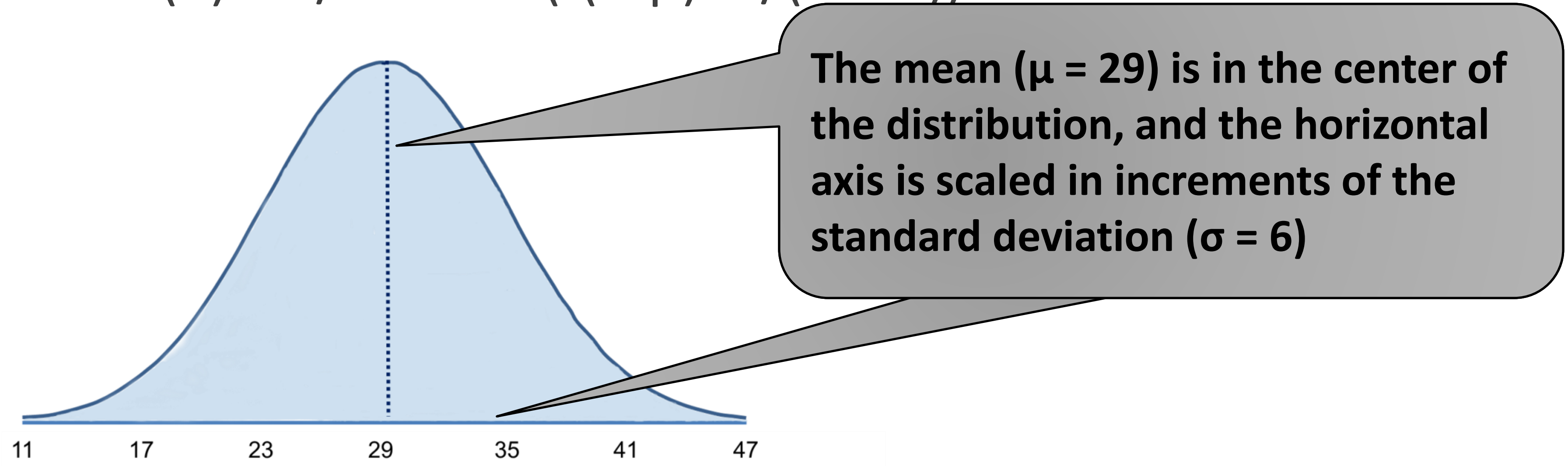
e=2.718281828=2.72 (Euler’s number, a mathematical constant, $\sum_{n=0}^{\infty} \frac{1}{n!}$)

Pr(x=5) = ($2.72^{-2.8}$*2.85)/(5*4*3*2*1) =10.45/120 = 0.09

Normal distribution

When we have a normally distributed variable, and we know the population mean (μ) and population standard deviation (σ), we can compute the probability of particular values using the following formula:

$Pr(X) = 1/\sigma\sqrt{2}\pi \cdot e\hat{}(-(X-\mu)\hat{}2/(2\sigma\hat{}2))$



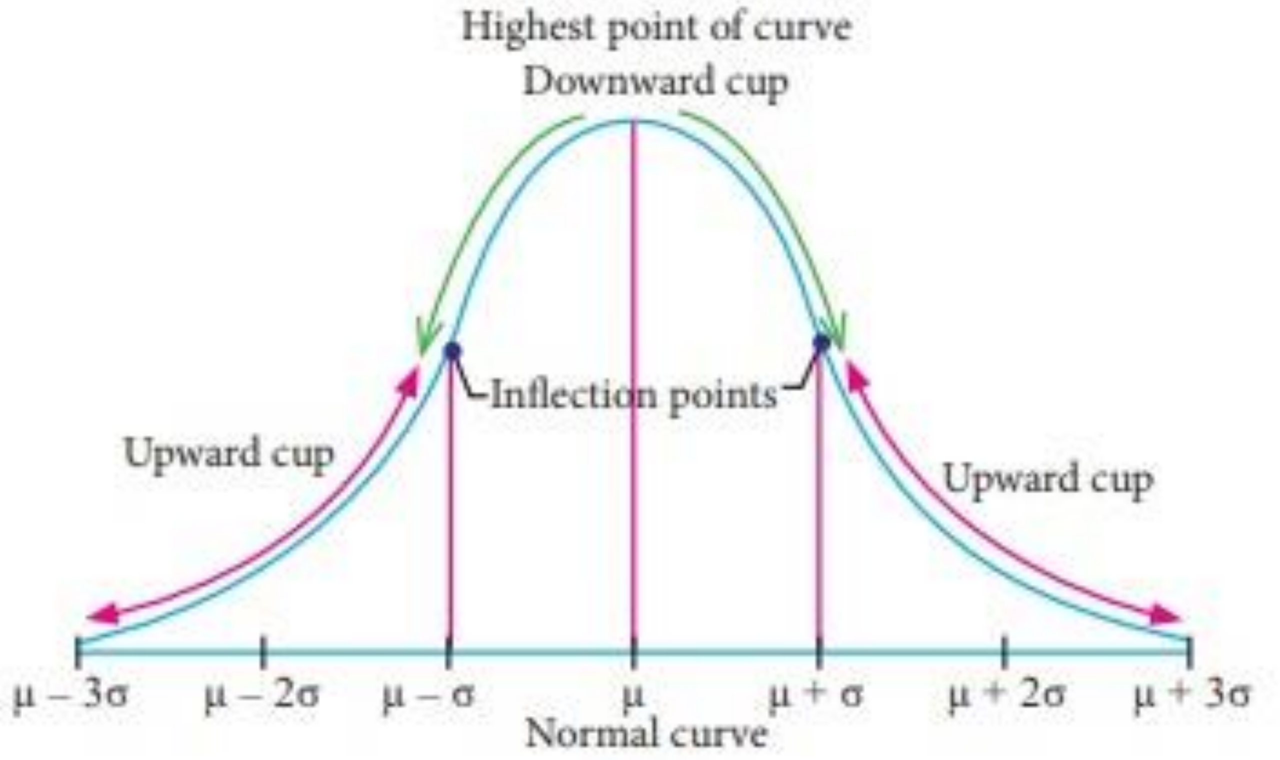
The mean (μ = 29) is in the center of the distribution, and the horizontal axis is scaled in increments of the standard deviation (σ = 6)

Normal distribution

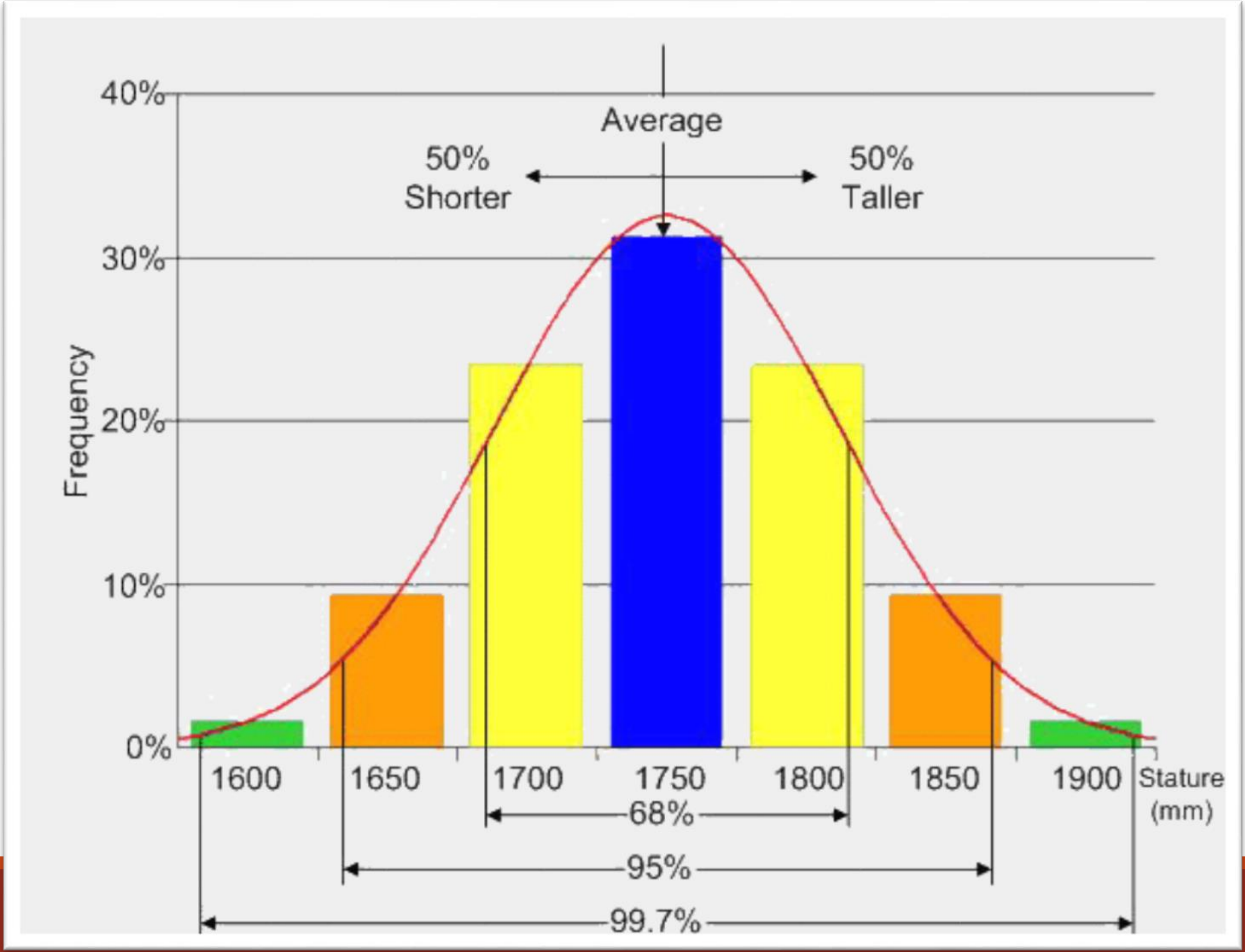
Normal distribution:

- Symmetric
- Not skewed
- Unimodal
- Described by two parameters:
  - Probability density function: $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- μ & σ are parameters
  - μ = mean
  - σ = standard deviation
  - π, e = constants

Normal distribution

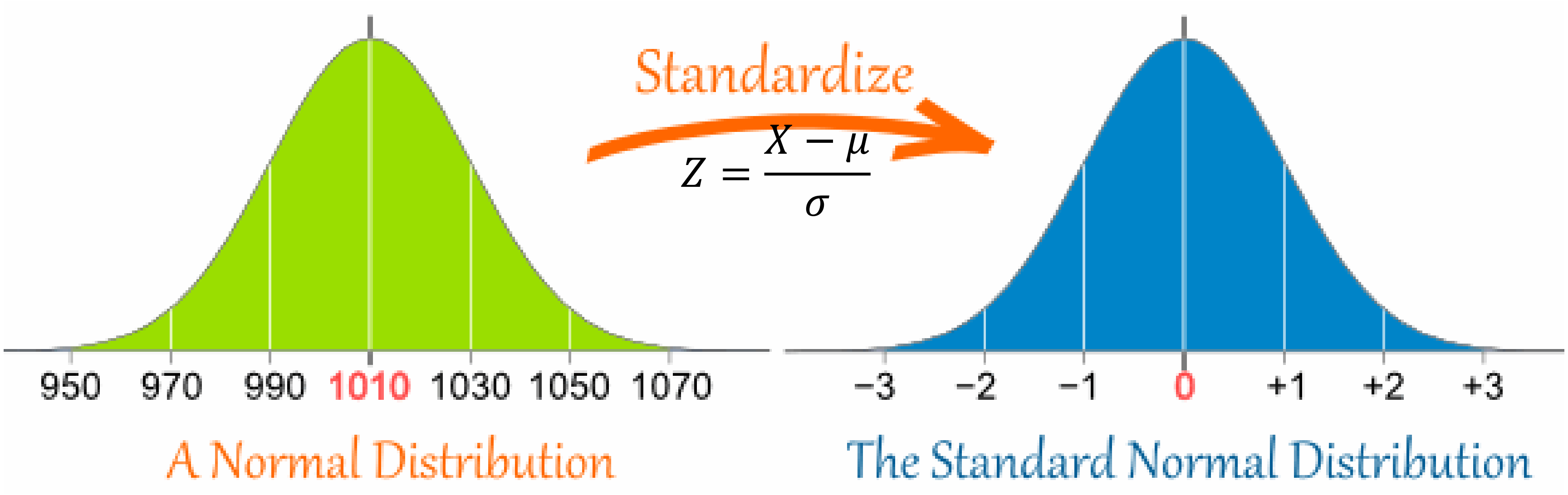


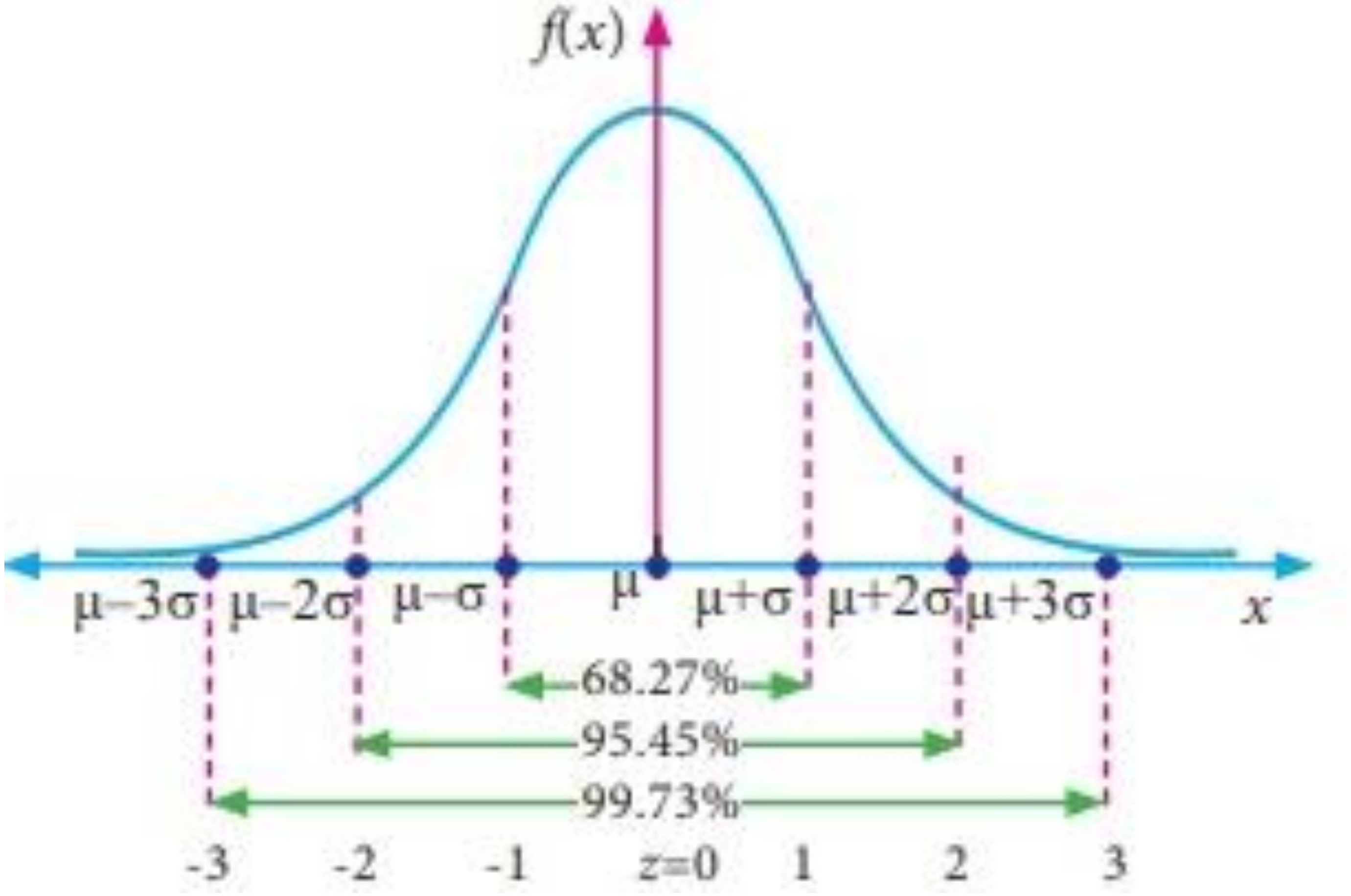
Normal distribution

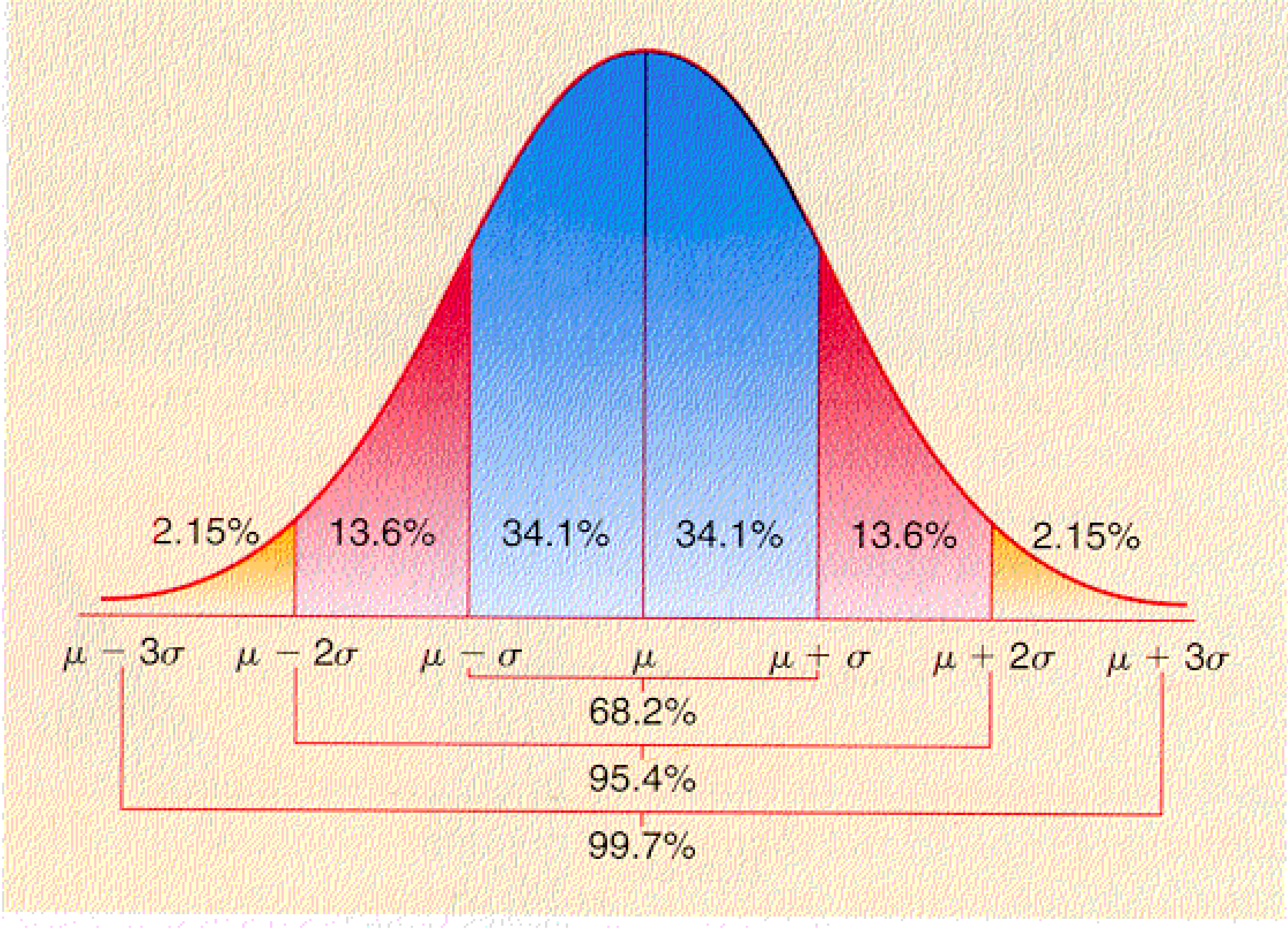


Standard normal distribution

The standard normal distribution is a normal distribution with a mean of zero and standard deviation of 1.




f(x)
x
μ–3σ
μ–2σ
μ–σ
μ
μ+σ
μ+2σ
μ+3σ
68.27%
95.45%
99.73%
-3
-2
-1
z=0
1
2
3



μ=10 age

σ=2 age

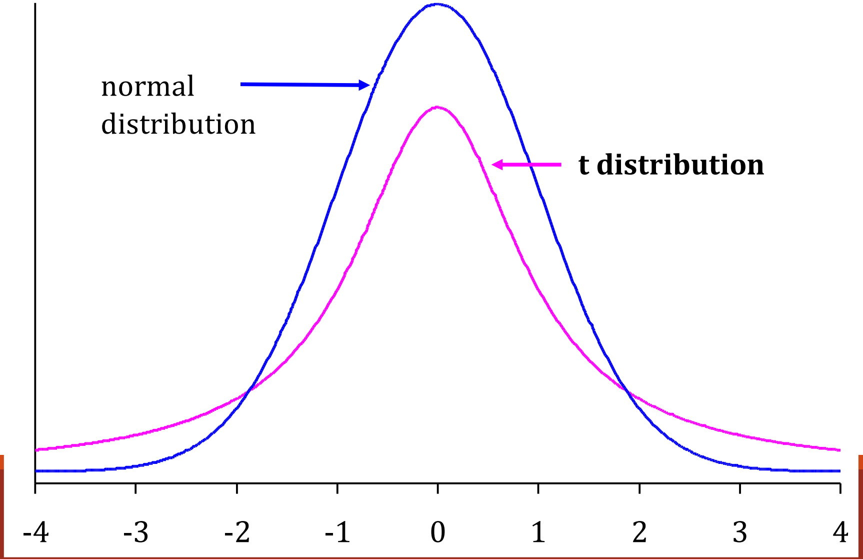
how many subjects are older than 12 years (aprox)?

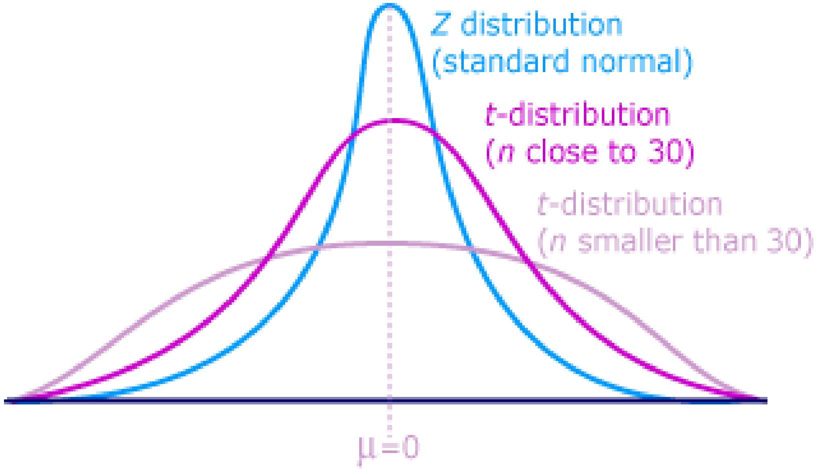
(100%-68%)/2=34%/2=17%

Student t-distribution

Student's t-distribution (or simply the t-distribution):

- A member of a family of continuous probability distributions that arise when estimating the mean of a normally distributed population in situations where the sample size is small, and population standard deviation is unknown.
- Used by Student t-test, to construct confidence intervals, in linear regression analysis





Normal distribution

A family doctor with ~ 3,000 subjects on the list measure over one year the heart rates (expected to be normal distributed). Three statistics were reported: mean = 75, minimum = 45, and maximum = 105. Which of the following is most likely to be the standard deviation of the distribution?

A. 2 → 75 ± 3×2 = (69; 81)

B. 5 → 75 ± 3×5 = (60; 90)

C. 10 → 75 ± 3×10 = (45; 105)

D. 12 → 75 ± 3×12 = (39; 111)

E. 15 → 75 ± 3×15 = (30; 120)

Normal distribution

n=3000; mean = 75; minimum = 45; maximum = 105

Which of the following is most likely to be the standard deviation of the distribution?

100% of data in the range of [45; 105]

99.7% of data in the range of ?

A. 2 → 75 ± 3×2 = (69; 81)

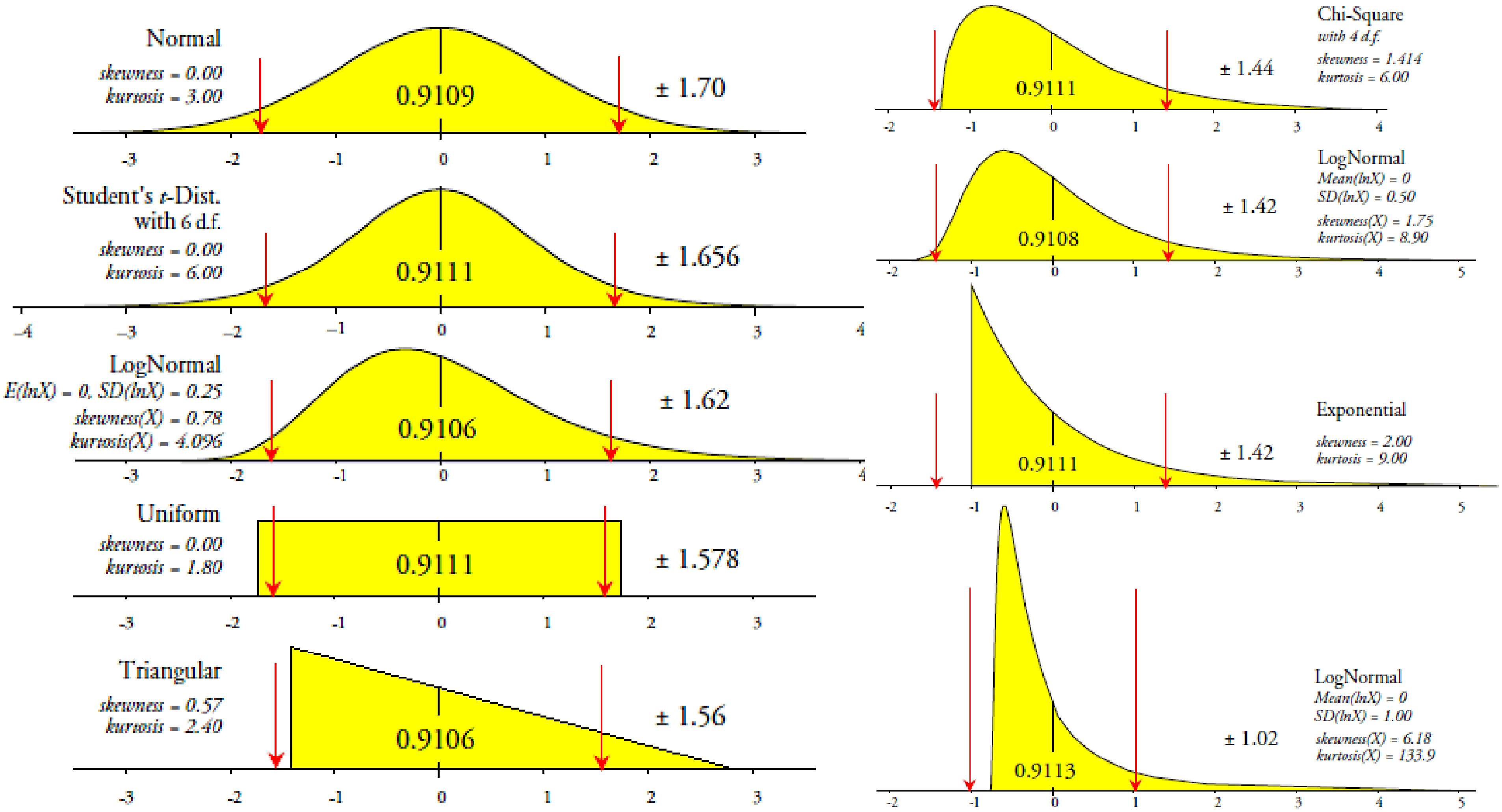
B. 5 → 75 ± 3×5 = (60; 90)

[✓ C.] 10 → 75 ± 3×10 = (45; 105)

D. 12 → 75 ± 3×12 = (39; 111)

E. 15 → 75 ± 3×15 = (30; 120)

Probability distributions by example



Recall

Random variables could be discrete or continuous.

For random variables we have:

- Discrete probability distributions
- Continuous probability distribution

It is possible to compute the expected values (means), variation and standard deviation for discrete random variables.

Recall

Normal distribution

- Is bell-shaped, unimodal, symmetric, and continuous; its mean, median, and mode are equal
- Its standard form has a mean of 0 and a standard deviation of 1
- Can be used to approximate other distributions to simplify the analysis of data (e.g. binomial distribution)