Continuous Frequency Distributions & Summary Statistics

OUTLINE - DISTRIBUTION

- Probability distributions
- Continuous probability distributions by example

Probability Distribution

Discrete

- The probabilities associated with each specific value

Continuous

- The probabilities associated with a range of values

Continuous Probability Distributions

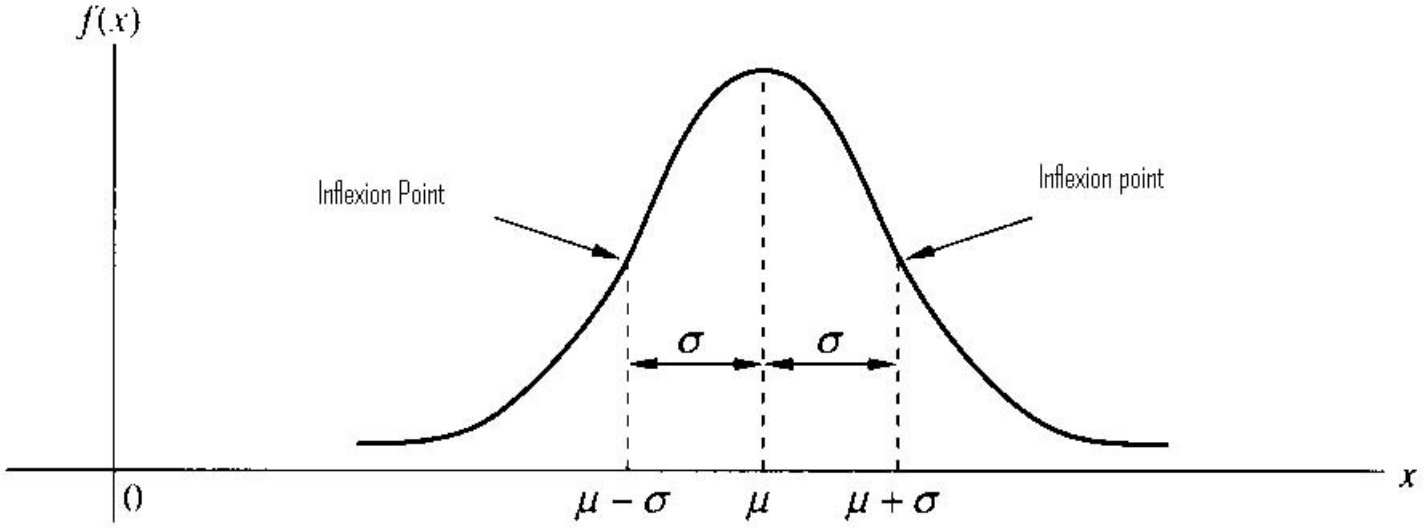
- We talk about probabilities for a range of values, not a particular value
- Probability for a range of values is determined by the area under the probability distribution curve

KNOWN CONTINUOUS DISTRIBUTIONS

- Normal Z (Gauss)
- STUDENT (t)
- PEARSON (χ^2)
- F (FISHER)

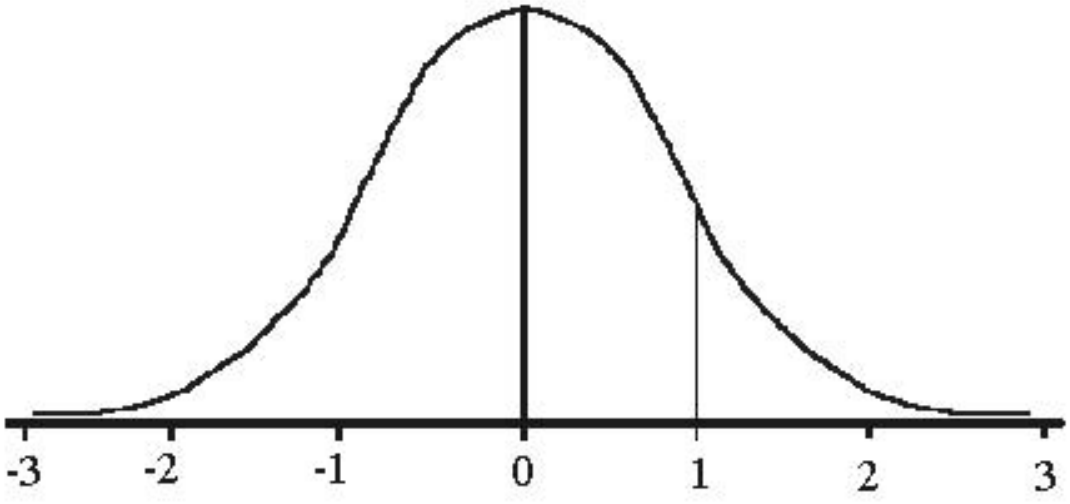
NORMAL DISTRIBUTION

- X random variable is normal of type N(μ,σ) if its distribution depend by two parameters: mean (μ) and standard deviation (σ)

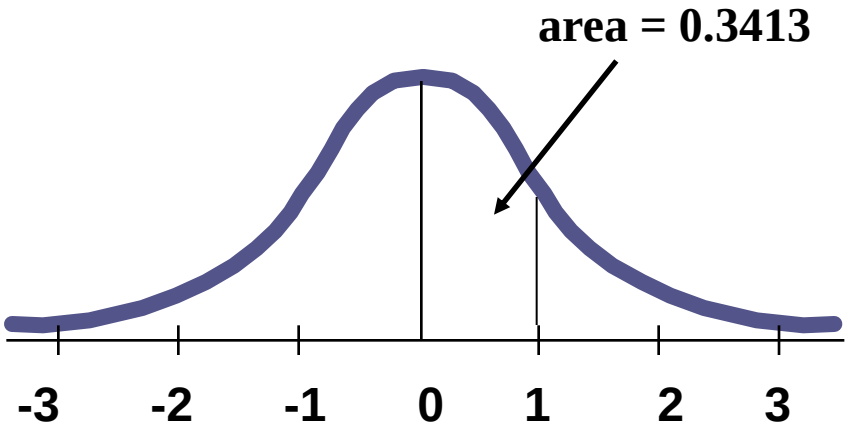


NORMAL DISTRIBUTION

- Normal distribution has mean μ and variance σ^2
- Standard normal distribution has the mean equal to 0 and the variance equal to 1



NORMAL DISTRIBUTION: COVERAGE



- μ 1*σ: contains ~ 68% of cases (34% from each part of distribution)
- μ 2*σ: contains ~ 95% of cases
- μ 3*σ: contains ~ 99.7% of cases

- Normal distribution is a limit case of binomial discrete distribution for sample with large sizes.

STUDENT DISTRIBUTION

- Student or t distribution
  - Probability distribution which appear in estimation of the mean of a normal distributed population when the sample size is small (<30)

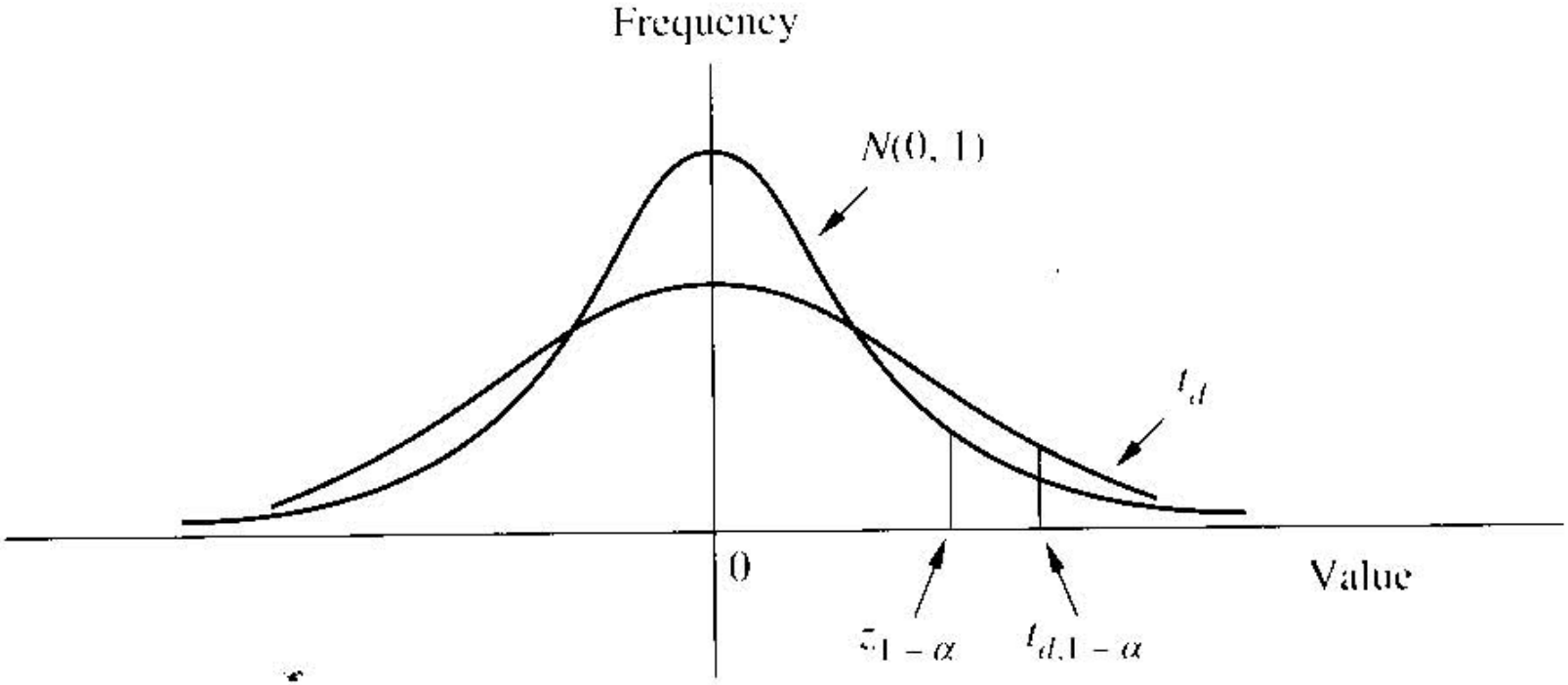
Student Distribution

- Properties
  - Is different for different sample sizes.
  - Is generally bell-shaped, but with smaller sample sizes shows increased variability (flatter).
    - The distribution is less peaked than a normal distribution and with thicker tails
    - As the sample size increases, the distribution approaches a normal distribution.
    - For $n > 30$, the differences are negligible.

Student Distribution

- Properties
  - The mean is zero (much like the standard normal distribution).
  - The distribution is symmetrical about the mean.
  - The variance is greater than one, but approaches one from above as the sample size increases (2=1 for the standard normal distribution).
  - It takes into account the fact that the population standard deviation is unknown.
  - The population is essentially normal (unimodal and basically symmetric)

STUDENT VS GAUSS DISTRIBUTIONS



CHI-SQUARED DISTRIBUTION

- Chi-square distribution (also chi-squared or χ^2-distribution)
- One of the most widely used theoretical probability distributions in inferential statistics
- It is used by
  - Chi-square tests for goodness of fit
    - of an observed distribution to a theoretical one
    - of the independence of two criteria of classification of qualitative data

F-DISTRIBUTION

- **Snedecor's *F* distribution** or the **Fisher-Snedecor distribution**
- A continuous probability distribution defined on $[0, +\infty)$
- arises as the null distribution of a test statistic:
  - likelihood-ratio tests
  - analysis of variance (F test)

SUMMARY STATISTICS

It's nice to have lots of data … but … sometimes it is too much for a good things

76	189	184	89	185	88	169
77	81	165	160	108	170	200
72	210	210	190	174	72	72
170	81	180	83	87	81	190
180	69	170	79	180	170	185
74	92	182	66	70	79	184
171	71	184	78	126	87	191
183	186	169	76	187	83	85
74	187	170	171	174	94	94
74	193	173	186	65	66	177
79	180	82	122	80	185	171
82	73	170	82	181	72	83
188	195	86	180	135	96	156
93	79	160	140	98	73	190
74	75	190	170	80	143	99
140	150	72	180	82	84	82
80	190	72	171	190	172	190
72	78	80	88	75	192	161
182	70	82	181	88	73	181
70	187	88	72	189	176	71
190	178	178	81	85	187	70
193	76	87	102	182	181	89
86	89	182	186	85	91	

OUTLINE

- Good Tables Practices
- Good Graphical Practices
- Numerical Summaries: 1 & 2 variables
- Ordinal Summaries: 1 & 2 variables

Summarizing Medical Data

- Large amounts of medical data are compressed into more easily assimilated summaries
  - Provide the user with a sense of the content
- There a number of ways data can be presented depending by the type of variables

GOOD TABLES PRACTICES

1. Simple: it is preferred to have 2 or 3 small tables instead of one big table
2. Must be information without reading the associated text:
   - Abbreviations and symbols must by explained at the bottom of the table
   - Definitions of rows and columns with units of measurements in headings (if it is applied)
   - Brief descriptive heading: what? when? where?
   - Must not duplicate material in the text or in illustration
   - Synthesis (total) rows and columns
3. If data are taken from another research the source of data must be referred.

GOOD GRAPHICAL PRACTICES

- Any graphical representation must to have:
  - Title
  - Definitions of axes
  - Units of measurements for each axe (if it is applied)
  - Legend (if it is applied)
- A good graphical representation must be as self-explanatory as possible!

GOOD GRAPHICAL PRACTICES

- The aim of a graphical representation is to transmit an information
- When drawing a graphical representation try to answer to the following question: Which is the aim of the graphical representation?
- Medical data must be represented graphically in a such a way in which to be useful for understanding the clinical phenomena
- Notice to:
  - The color composition (do not use color background)
  - The font size (it is suppose to be readable)

One Qualitative Variable: Frequency Tables

- Data are sort ascending
- The absolute frequency of each value is
- The distinct values and associated frequencies are included into a table :
  - Absolute frequency: the total amount of occurrences of one variable
  - Relative frequency = the absolute frequency divided by the total amount of occurrences

One Qualitative Variable: Frequency Tables

- Could contains the following types of frequencies:
  - Absolute frequency
  - Cumulative absolute frequency (ascending / descending)
  - Relative frequency
  - Cumulative relative frequency (ascending / descending)

- Microsoft Excel:
  - COUNTIF
  - Tabele Pivot
    - [Data - Pivot Table and Pivot Chart Report ...]

NUMERICAL SUMMARIES: ONE VARIABLE

- Quartiles
- Mean:
  - Population: μ (population's arithmetic mean)
  - Sample: m (sample's arithmetic mean)
    - ∑ means: add together all data elements whose symbol follows me
- Median (has no standard symbol):
  - Put the n observation in order of size
  - Median is the middle observation if n is odd
  - Median is the halfway between the two middle observations if n is even

Numerical Summaries: One Variable

- Mode (has no standard symbol)
  - Make a bar chart of the data
  - Mode is the center value of the highest bar
- Variance (the average of the squares of differences between the observations and theis mean):
  - Population: σ^2
  - Sample: s^2
- Standard deviation (the square roots of the respective variance):
  - Population: σ
  - Sample: s
- Standard error of the mean

NUMERICAL SUMMARIES: TWO VARIABLES

- Covariance (joint frequency distributions):
  - Required paired recordings (a reading on Y for each reading on X)
  - Interpretation:
    - If one variable tends to increase as the other increase (systolic and diastolic blood pressure) the covariance is positive and large.
    - If one variable tends to decrease as the other increase (PSA and prostate density) the covariance is negative and large.
    - If increases and decreases of one variable are unrelated to those of the other, the covariance tends to be small.
  - Useful in indication a shared behavior or independence between two variables (NO standard for interpreting it!).

NUMERICAL SUMMARIES: TWO VARIABLES

- Correlation coefficient:
  - Standardized covariance by dividing by the product of standard deviation of the two variables.
  - Interpretation:
    - If either variable is perfectly predictable from the other, the correlation coefficient is 1 when both increase together and -1 when one increases and other decreases.
    - If the two variables are independent (a value of one provide no information about the value of other) the correlation coefficient is 0.
    - A correlation coefficient of 0.10 is rather low, showing little predictable relationship
    - A correlation coefficient of 0.90 is rather high, showing that one increases rather predictably as the other increase.
  - Measure relationship along a straight line!

PICTORIAL SUMMARIES: ONE VARIABLE

- Bar chart:
  - The choice of interval is important (an unfortunate choice of intervals can change the apparent pattern of the distribution).
    - Enough intervals should be used so that the pattern will be minimally altering the beginning and ending positions.
  - The choice of number, width, and starting points of intervals arise from the user's judgment (they should be considered carefully before forming the chart).

PICTORIAL SUMMARIES: ONE VARIABLE

- Histogram:
  - Appears like the bar chart but differs in that the number of observations lying in an interval is represented by the area of a rectangular (or bar) rather than its height.
  - If all intervals are of equal width, the histogram is no different from the bar chart except cosmetically (no blank space between bars).
- Pie Chart:
  - Represents proportions rather than amounts.
  - Its main use is to visualize the relative prevalence of the phenomena.
  - Has the advantage of avoiding the illustration of sequence that sometimes is implied by the bars charts.

PICTORIAL SUMMARIES: ONE VARIABLE

- Line Chart:
  - The main use: to convey information similar to a bar chart but for intervals that form a sequence of time or order of events from left to right.
  - Relationship of a Line Chart to a Probability Distribution: as the sample size increases and the width of the intervals decreases, the line chart of a sample distribution approaches the picture of its probability distribution.

PICTORIAL SUMMARIES: ONE VARIABLE

- Mean-and-Standard Error Chart:
  - A diagram showing a set of means to be compared, augmented by an indication of the size of uncertainty associated with each mean.
  - If the data per group are distributed in a fairly symmetric and smooth bell-type curve, most of the relevant pattern may be discerned.
  - If the data per group are distributed irregular and/or asymmetrically this chart covers up important relationships and may lead to false conclusions.
  - Charts that are “data dependent” rather than “assumption dependent” as box-and-whisker charts often provide a batter understanding of the data.

PICTORIAL SUMMARIES: ONE VARIABLE

- Box-and-Whisker Chart:
  - Display typical (distribution center and spread) and atypical aspects (asymmetry, outlying values).
  - The whisker lengths that are similar and are about half the semibox length are evidence of symmetry and a near normal distribution.
  - Unequal whisker lengths indicate asymmetry in the outer part of the data distribution.
  - The presence of data far out in the tails, as well as the distance out, is shown by dots above and below the whisker ends.

PICTORIAL SUMMARIES: TWO VARIABLE

- Scatter Plot (depicting the relationship between variables):
  - Plot the pair of readings for each patient on perpendicular axes.
  - Indicate if the points are randomly scattered or clustered (we can see the location and shape of these clusters).
- Two-Dimensional frequency Distribution:
  - Several characteristics at once (3D image).
  - The frequency value of a point is readable but the viewer must extrapolate the height of a column (the extrapolation could be distorted by the perspective)

One Qualitative Variable: Frequency Tables

Absolute frequency

Relative frequency

Diagnosis	No. patients	Percent (%)
Asphyxia at birth	527	26.1
Obstetrical injuries	92	4.6
Septic status	7	0.3
Pneumonia	181	9.0
Diarrhea	8	0.4
Congenital malformations	598	29.6
Other causes	606	30.0
Total	2019	100

ONE QUALITATIVE VARIABLE: FREQUENCY TABLES

The sum of absolute frequencies of all values in the series that are less than or equal to x/n

The sum of absolute frequencies of all values in the series that are less than or equal to x

Diagnosis	f_a	f_r	f_a **cumulat ↑**	f_r **cumulat ↑**
Asphyxia at birth	527	26.10	527	26.10
Obstetrical injuries	92	4.56	619	30.66
Septic status	7	0.35	626	31.01
Pneumonia	181	8.96	807	39.97
Diarrhea	8	0.40	815	40.37
Congenital malformations	598	29.62	1413	69.99
Other causes	606	30.01	**2019**	**100**
Total	**2019**	**100**		

One Qualitative Variable: Frequency Tables

- Let have the following incubation time expressed in days for a infectious diseases: 5, 6, 7, 7, 8, 8, 5, 7, 8, 7. Which of the following values correspond to the ascending cumulative relative frequency of 0.7?

  A. 8

  B. 6

  C. 5

  D. 7

  E. No right answer

One Qualitative Variable: Frequency Tables

- Let have the following incubation time expressed in days for a infectious diseases: 5, 6, 7, 7, 8, 8, 5, 7, 8, 7. Which of the following values correspond to the ascending cumulative relative frequency of 0.7?

Value	f_a	f_r	f_a cc	f_r cc
5	2	0.20	2	0.20
6	1	0.10	3	0.30
7	4	0.40	7	0.70
8	3	0.30	10	1
Total	10	1		

Two Qualitative Variables: Contingency Table

	TBC+	TBC-	**Total**
sex=F	2	10	**12**
sex=M	24	54	**78**
Total	**26**	**64**	**90**

n Qualitative Variables: Frequency Tables

Table 1. Distribution of pulmonary pathologies associated with silicosis

Grade of silicosis	BrC	BPOC	Emphysema	CPC	TBC	**Total**
I	12	20	0	0	14	**46**
I/II	1	5	1	1	1	**9**
II	3	7	1	1	7	**19**
II/III	0	1	0	0	0	**1**
III	0	3	0	0	4	**7**
Total	**16**	**36**	**2**	**2**	**26**	**82**

BrC = chronic bronchitis; BPOC = broncho-pneumonitis chronic obstructive;
CPC = chronic pulmonary heart; TBC = pulmonary tuberculosis

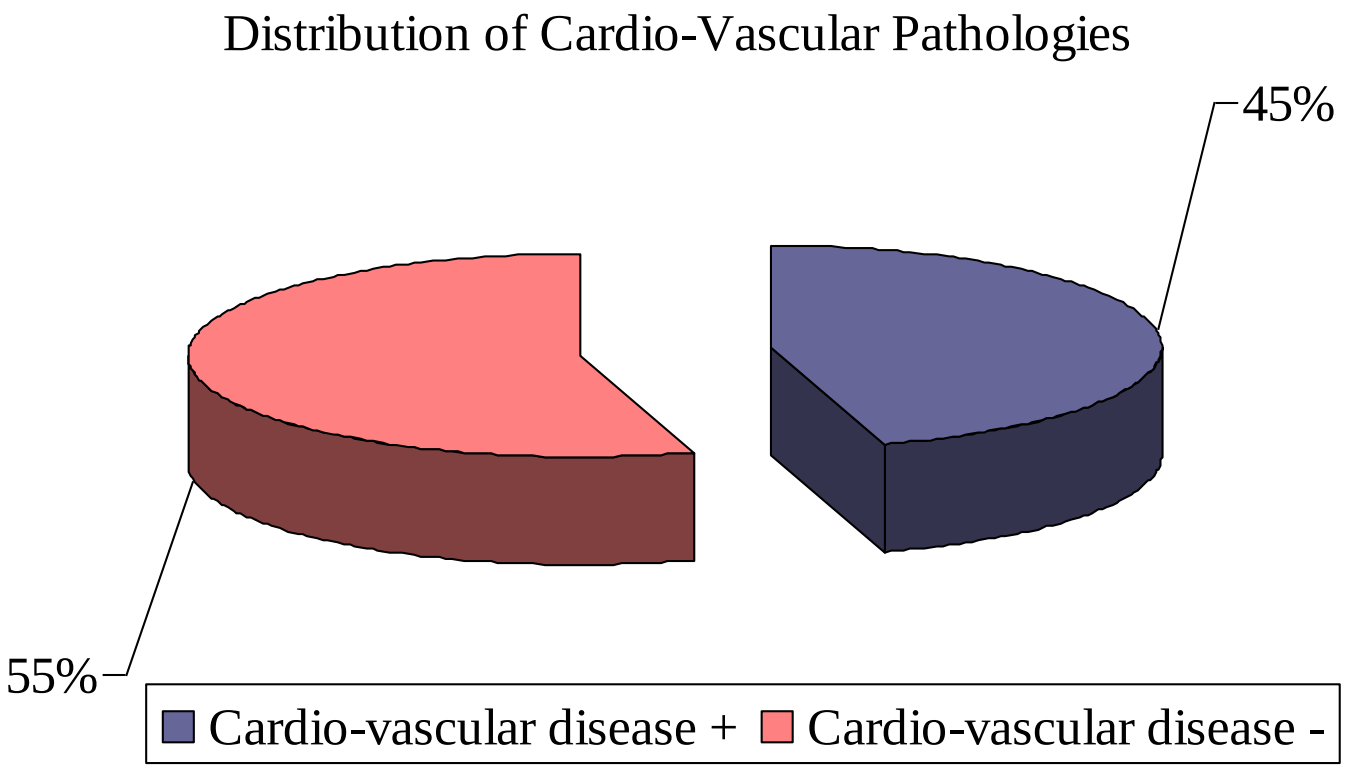
One Quantitative Variable: Frequency Classes Table

Weight (g)	$\mathbf{f_a}$	$\mathbf{f_r}$	$\mathbf{f_r}$ **cc** ↑
(2800 – 3200]	151	18.60	18.60
(3200 – 3400]	299	36.82	55.42
(3400 – 3600]	300	36.95	92.37
(3600 – 3800]	0	0.00	92.37
(3800 – 4000]	62	7.64	**100**
Total	**812**	**100**	

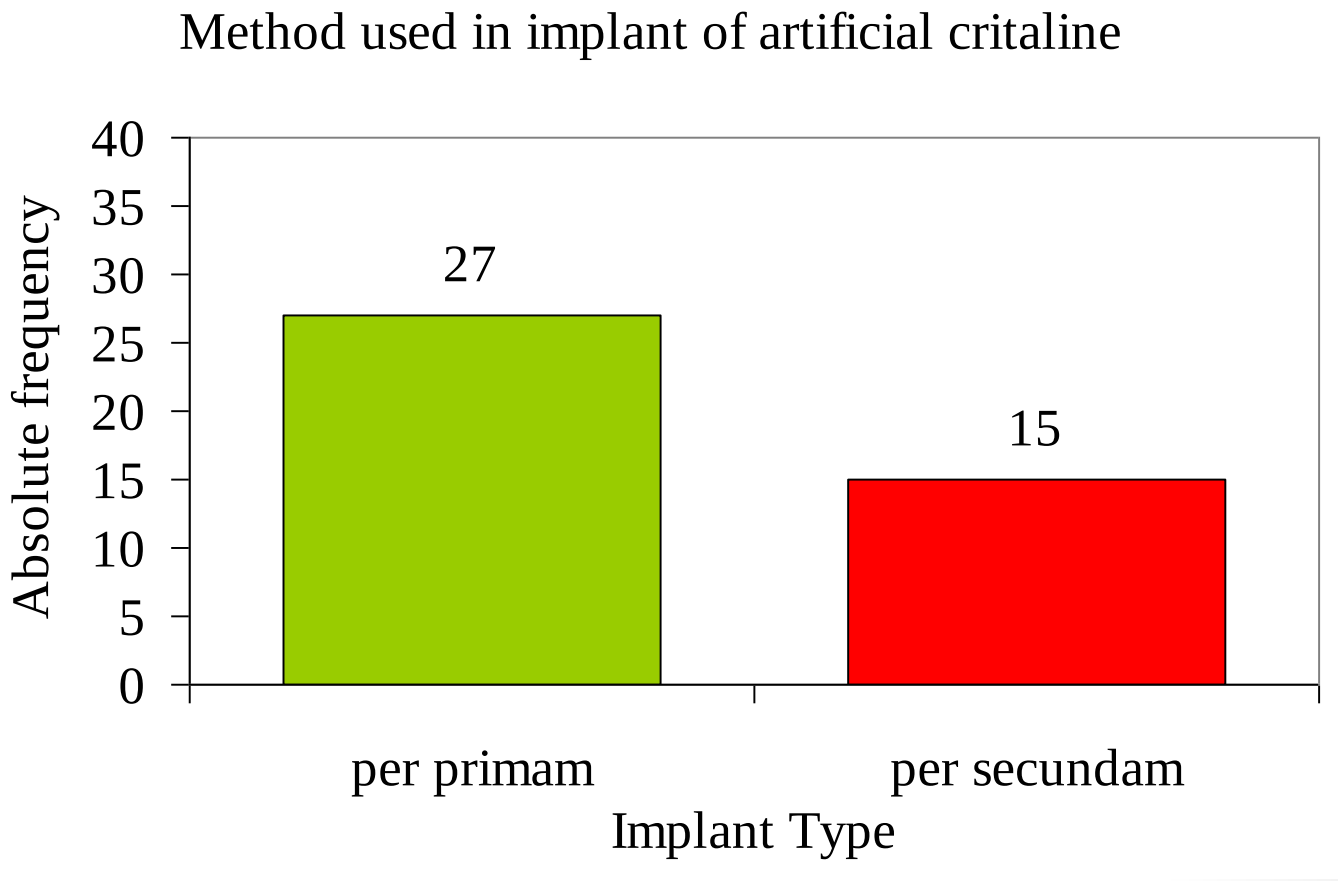
ONE VARIABLE: PIE

- Qualitative or Quantitative variables.
  - If it is quantitative could be drawn on frequency classes.
- It is used to represent absolute or relative frequencies:
  - Relative prevalence of a health phenomena
- Data are collected as absolute frequencies

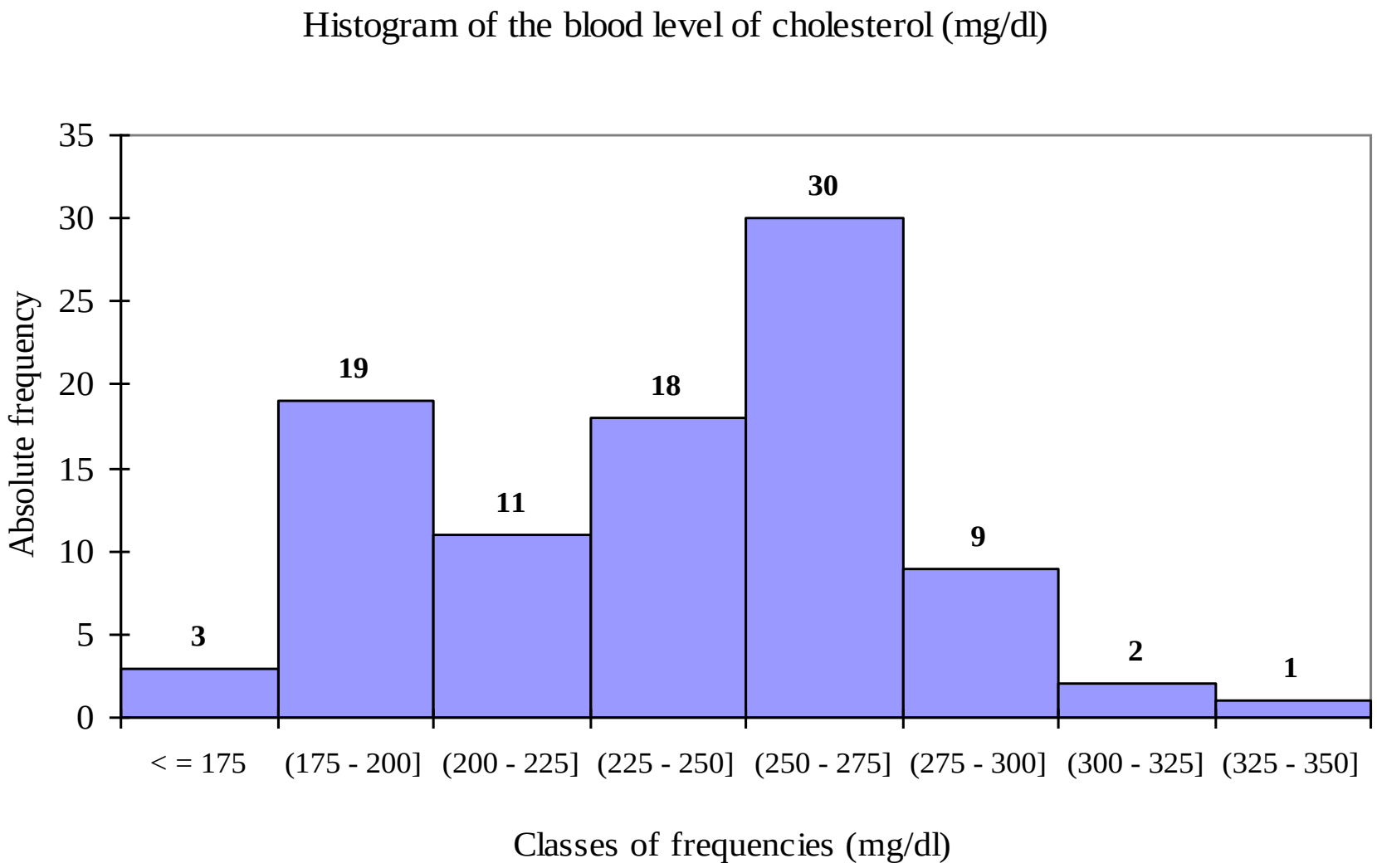
ONE VARIABLE: PIE



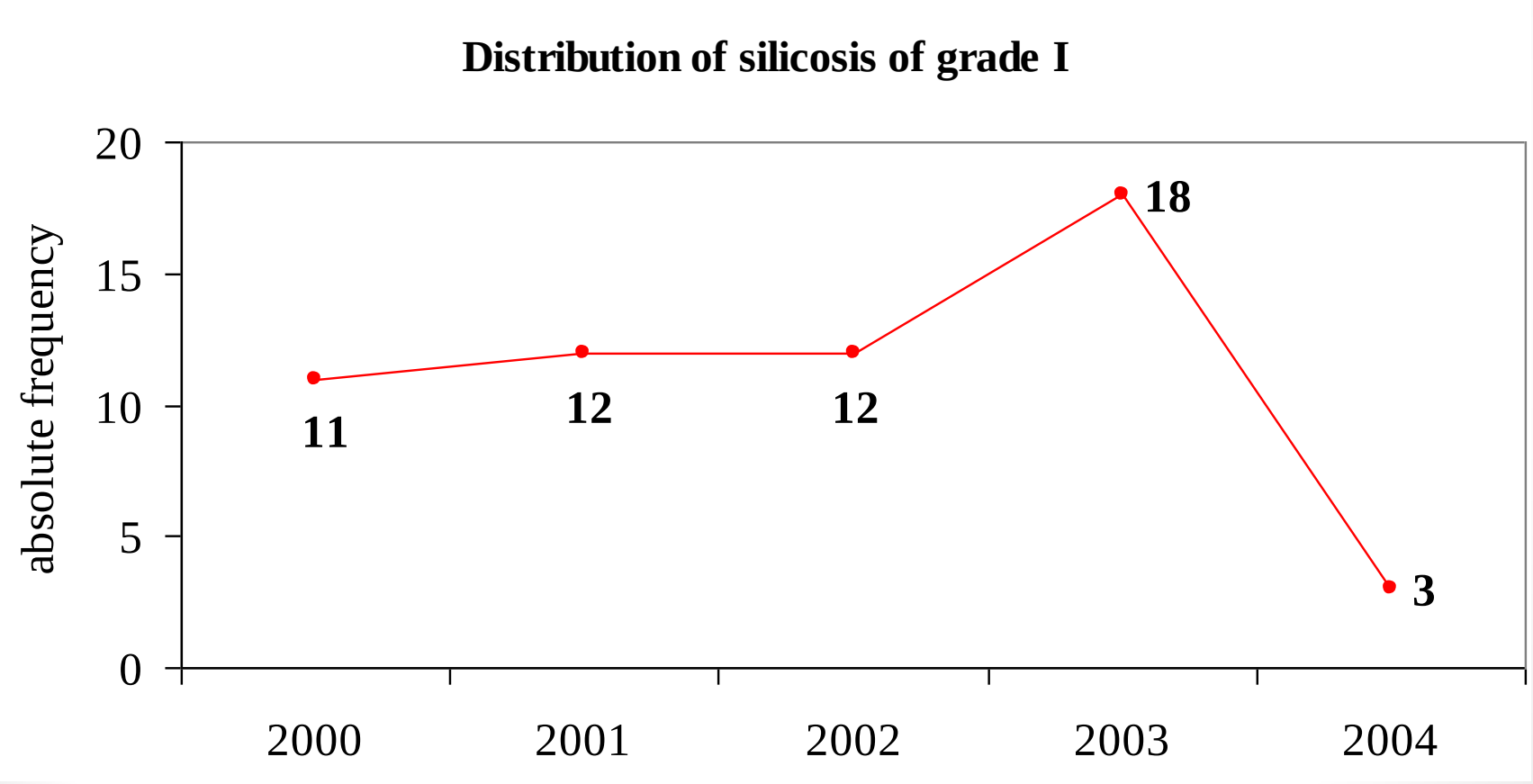
ONE VARIABLE: COLUMN



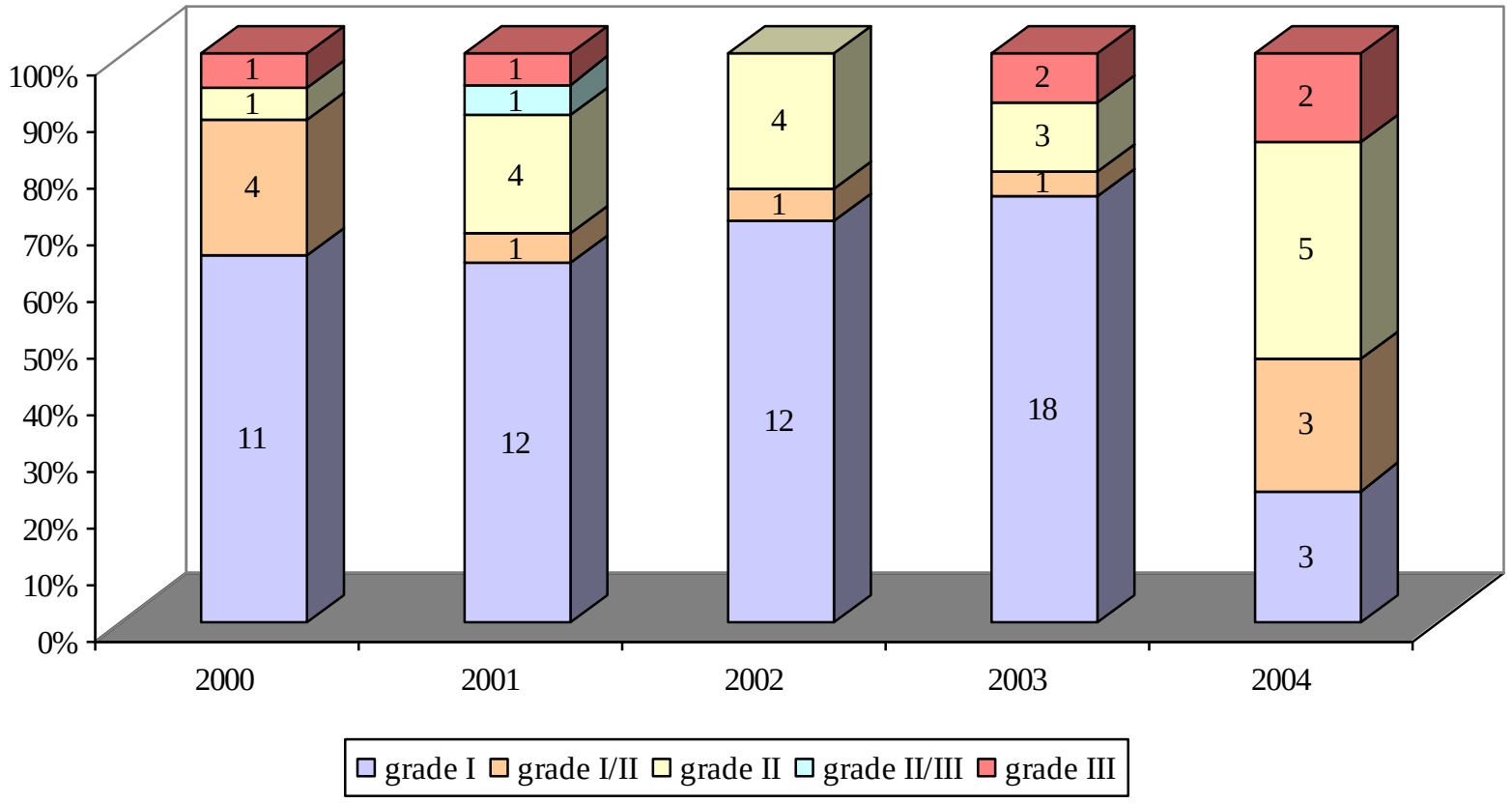
ONE VARIABLE: HISTOGRAM



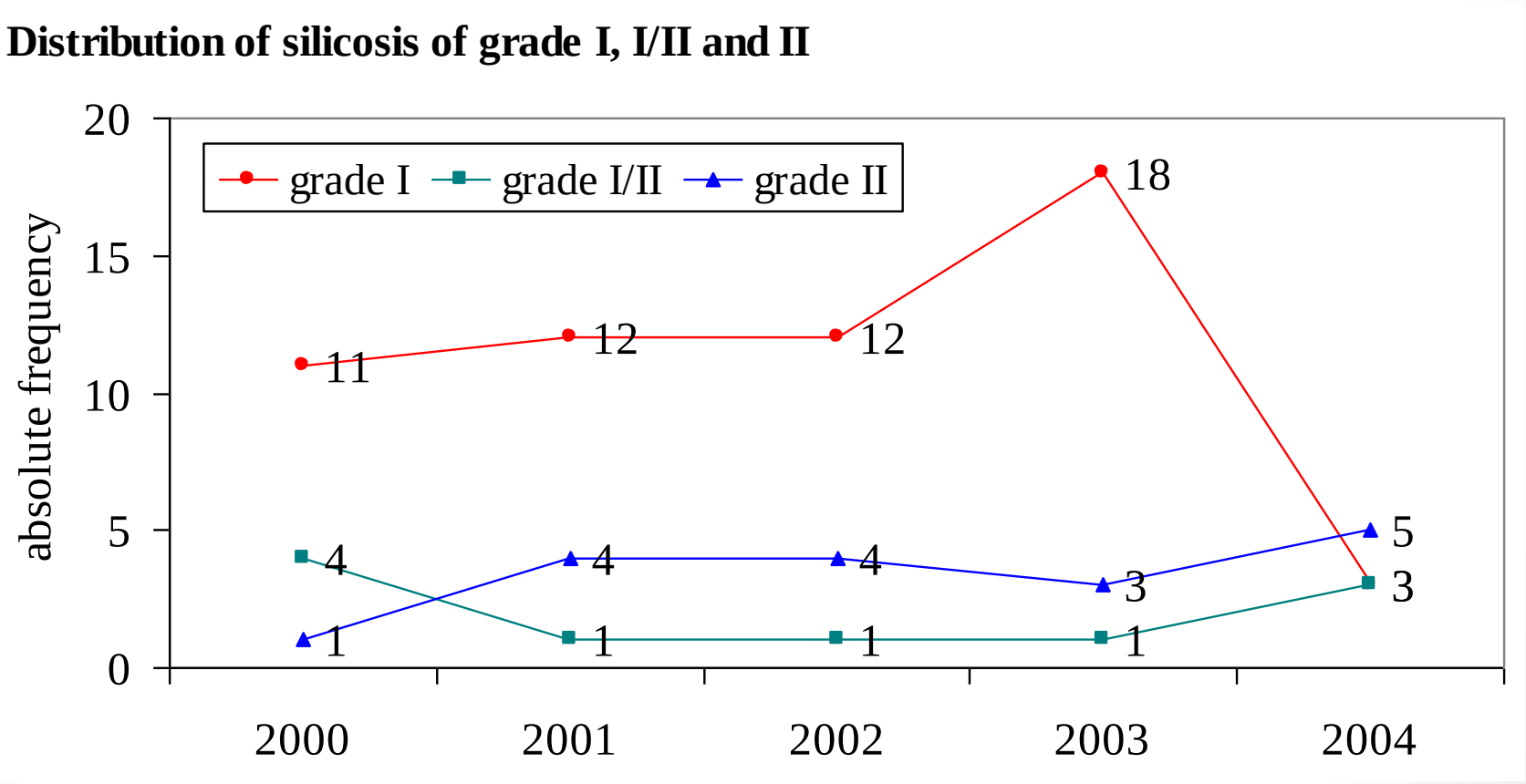
ONE VARIABLE: LINE



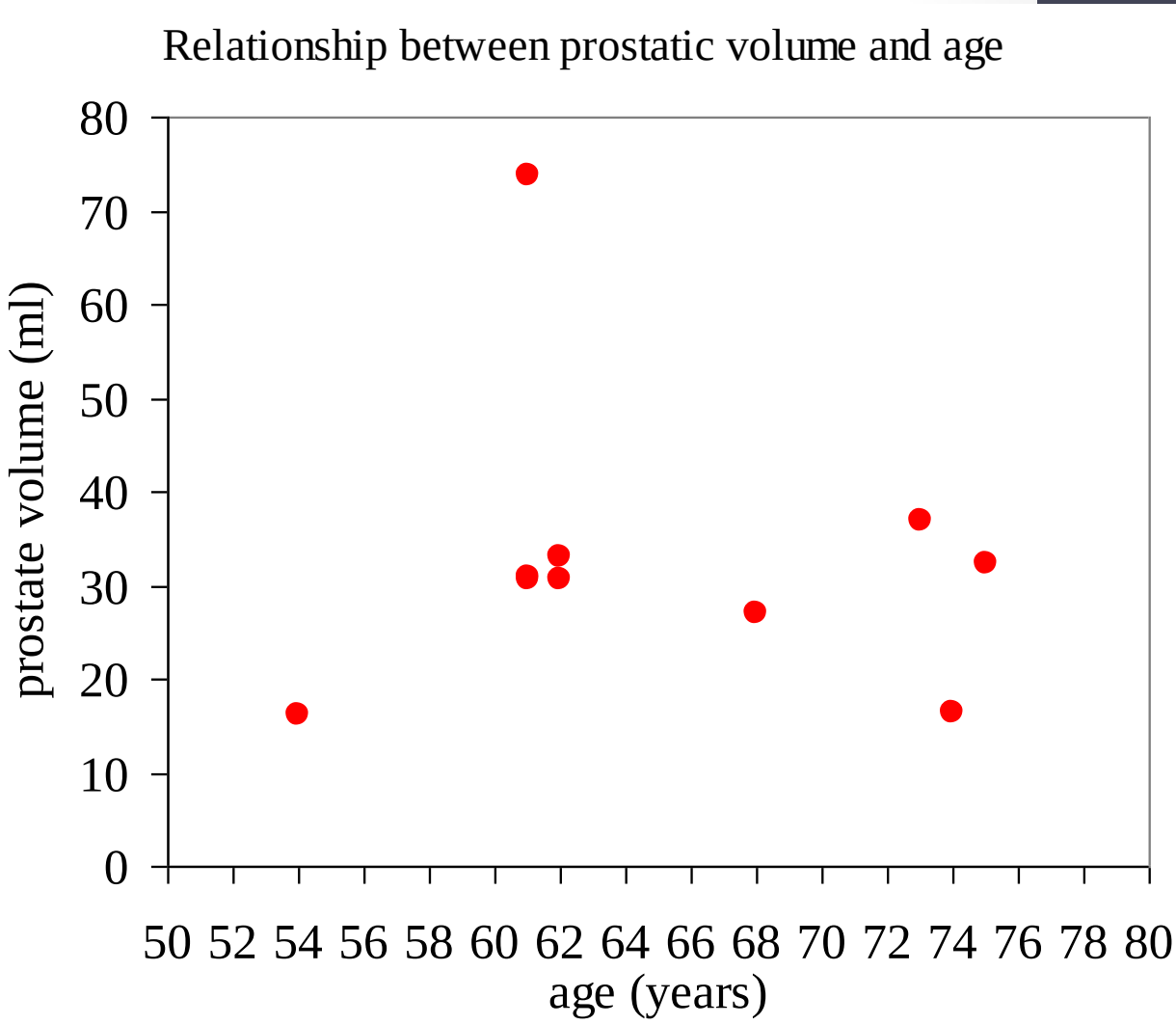
Two Qualitative Variable: COLUMNS



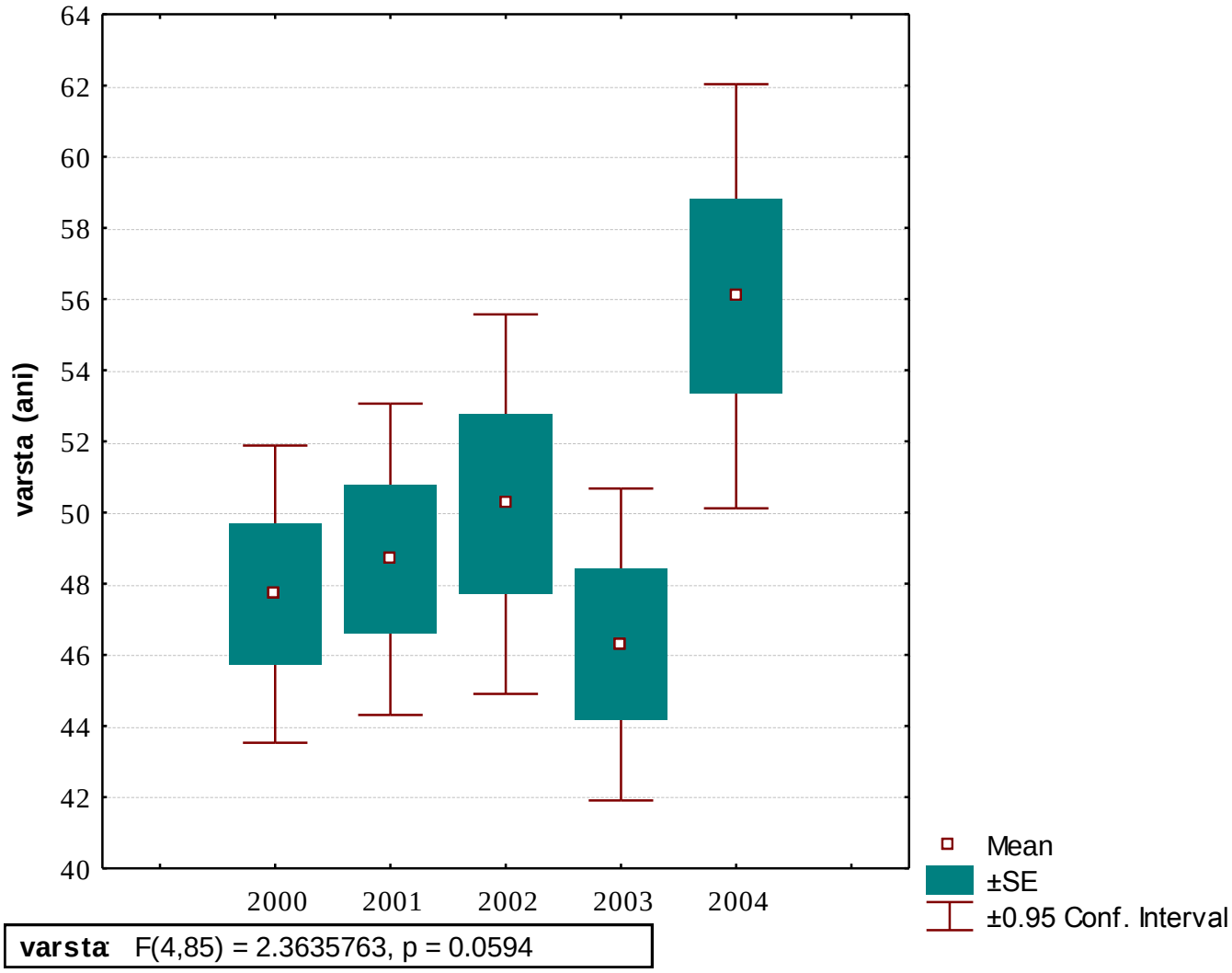
n QUALITATIVE VARIABLE: LINE



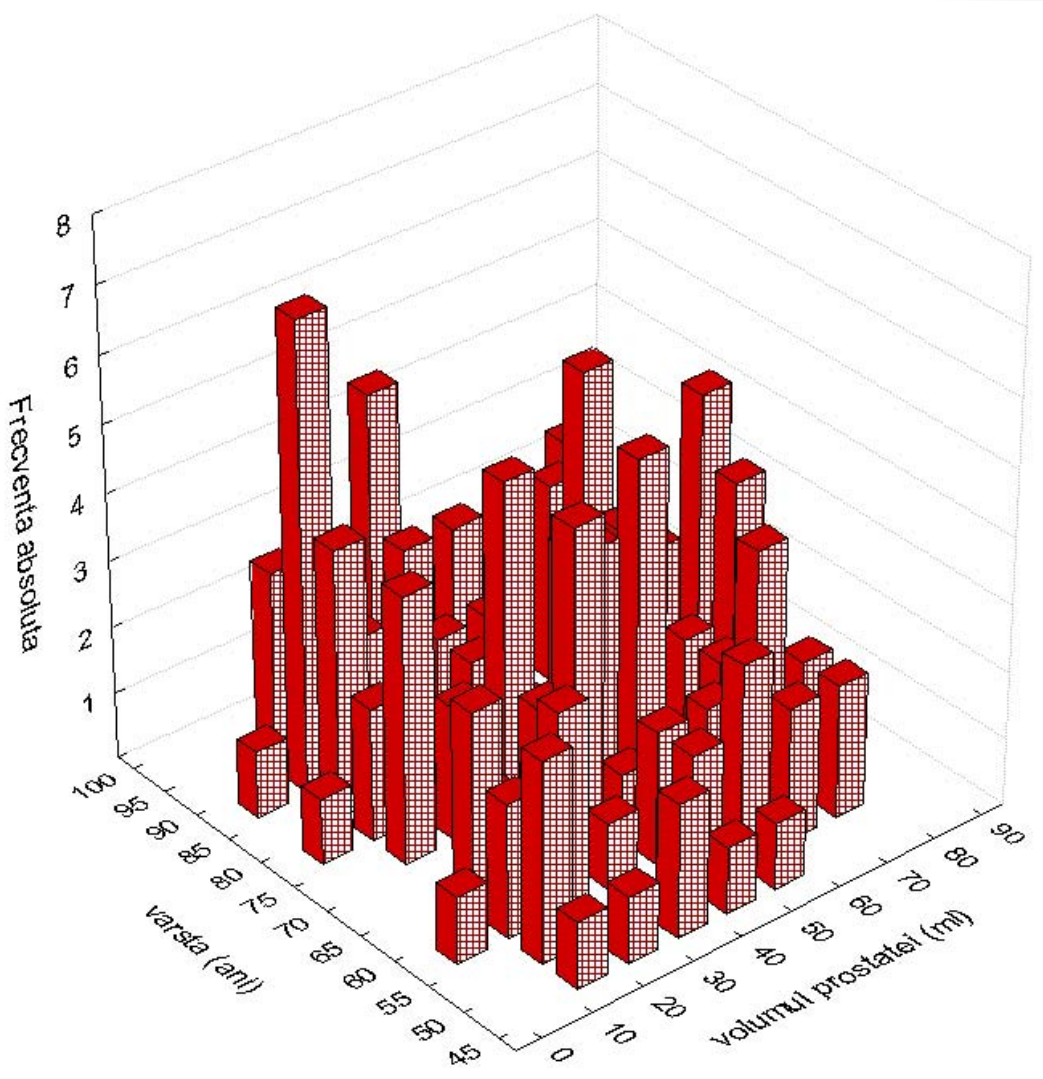
TWO QUANTITATIVE VARIABLES: SCATTER



TWO VARIABLES: BOX-AND-WHISKER



Two Variables: Tri-Dimensional Histograme



STATISTICAL SUMMARY BY EXAMPLE

Table 2

Test coverage and timeliness for sexually transmitted infections (STIs) and blood borne viruses (BBVs) by age group and location of prison of admission.

Test	Metropolitan adults % (n)	Regional adults % (n)	Juveniles* % (n)
STI coverage†	39.2 (163/416)	40.7 (120/295)	84.1 (195/232)
STI ≤ 7 days††	71.0 (110/155)	25.6 (30/117)	97.4 (185/190)
STI ≤ 28 days††	80.6 (125/155)	70.9 (83/117)	98.4 (187/190)
BBV coverage‡	47.2 (193/409)	45.5 (132/290)	15.8 (37/234)
BBV ≤ 7 days‡‡	9.1 (16/175)	15.6 (19/122)	69.7 (23/33)
BBV ≤ 28 days‡‡	43.4 (76/175)	63.9 (78/122)	97.0 (32/33)

*There are no juvenile correctional facilities in regional Western Australia

†excludes 3 refusals (1 metropolitan adult, 2 juveniles)

††proportion of those who had STI testing and information available on the time of testing

‡excludes 13 refusals (8 metropolitan adults, 5 regional adults)

‡‡proportion of those who had BBV testing and information available on the time of testing

Watkins *et al. BMC Public Health* 2009 **9**:385 doi:10.1186/1471-2458-9-385

STATISTICAL SUMMARY BY EXAMPLE

Table 1

Characteristics of study participants

Number	**72** ***Mean (SD)***
Age (years)	59.2 ± 8.3
Years since menopause (years)	12.0 ± 8.2
Number of pregnancies	5.2 ± 3.4
Body mass index (kg/m2)	27.7 ± 4.5
Physical activity score (min/week)	3448 ± 1053
Systolic blood pressure (mmHg)	137 ± 17
Serum level	
Triglyceride (g/l)	1.3 ± 0.7
Total Cholesterol (g/l)	2.1 ± 0.3
high-density lipoprotein (g/l)	0.5 ± 0.1
low-density lipoprotein (g/l)	1.2 ± 0.3
CA IMT (mm)	0.8 ± 0.4
FA IMT (mm)	0.8 ± 0.3
Lumbar spine BMD (g/cm^2)	0.917 ± 0.172
Trochanter BMD (g/cm^2)	0.669 ± 0.121
Femoral neck BMD (g/cm^2)	0.823 ± 0.109
Ward triangle BMD (g/cm^2)	0.645 ± 0.140
Femoral total BMD (g/cm^2)	0.860 ± 0.111
	Number (Percentage)
Current smoking	2 (2.8)
Osteoporosis	40 (55.6)
History of personal peripheral osteoporotic fractures	13 (18.1)

STATISTICAL SUMMARY BY EXAMPLE

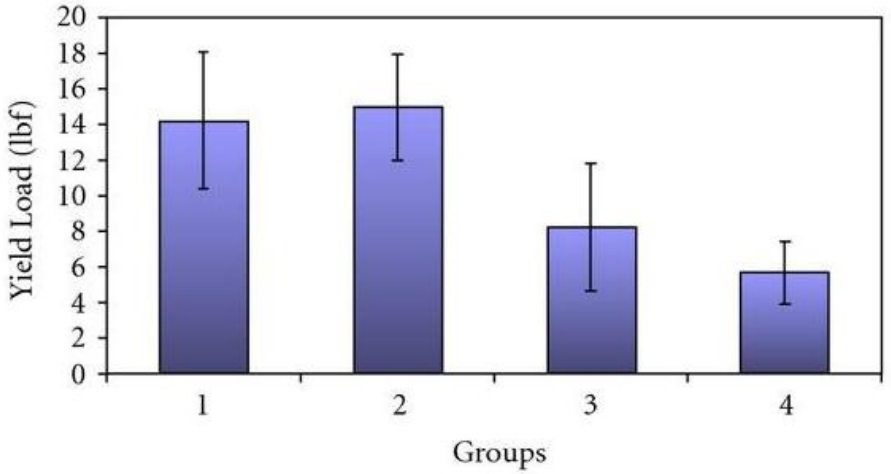
- Gökhan Açıkgöz, Murat İnanç Cengiz, İlker Keskiner, Şereften Açıkgöz, Murat Can, and Aydan Açıkgöz. Correlation of Hepatitis C Antibody Levels in Gingival Crevicular Fluid and Saliva of Hepatitis C Seropositive Hemodialysis Patients. International Journal of Dentistry 2009; Article ID 247121.

Table 1: Crosstabulation of HCV antibodies Immunoreactivity in Gingival Crevicular fluid and Saliva, Kappa = 0.426; $p < .001$.

		Gingival Crevicular fluid							
		Positive		Gray Zone		Negative		Total	
		n	%	n	%	n	%	n	%
Saliva	Positive	2	5.1			3	7.7	5	12.8
	Gray Zone			3	7.7			3	7.7
	Negative	4	10.3	1	2.6	26	66.7	31	79.5
	Total	6	15.4	4	10.3	29	74.4	39	100

STATISTICAL SUMMARY BY EXAMPLE

- Park SE, Chao M, Raj PA. Mechanical Properties of Surface-Charged Poly(Methyl Methacrylate) as Denture Resins. 2009; Article ID 841431:6 pages



Group	Group 1 (control)	Group 2 (5% *m*PMMA)	Group 3 (10% *m*PMMA)	Group 4 (20% *m*PMMA)
Mean	14.23	14.96	8.23	5.66
S.D.	3.84	2.98	3.59	1.75

Figure 1: The bar graph represents the mean and standard deviation values for transverse strength or force at fracture for each of the experimental groups.

GOOD TABLES PRACTICES: SUMMARY!

- Tables:
  - Capture: information concisely and display it efficiently
  - Provide information at any desired level of detail and precision
  - Number tables consecutively in the order of their first citation in the text and supply a brief title for each
  - Give each column a short or an abbreviated heading. Authors should place explanatory matter in footnotes, not in the heading
  - Explain all nonstandard abbreviations in footnotes
  - Identify statistical measures of variations
  - If you use data from another published or unpublished source, obtain permission and acknowledge that source fully

GOOD GRAPHIC PRACTICES: SUMMARY!

- Figures should be made as self-explanatory as possible.
- Titles and detailed explanations belong in the legends-not on the illustrations themselves.
- Figures should be numbered consecutively according to the order in which they have been cited in the text.
- If a figure has been published previously, acknowledge the original source and obtain written permission from the copyright holder to reproduce the figure.
- Explain clearly in the legend each symbols, arrows, numbers, or letters used in a figure.
- Avoid 3D graphical representations!

SUMMARIZING DATA - GRAPHS

- **SCATTER PLOT:**
  - two continuous numerical values
- **BAR GRAPH:**
  - qualitative variables
- **LINE GRAPH:** one quantitative variable
- **HISTOGRAM:** one continuous variable
- **PIE CHART:** one/two qualitative variables