第二章　数值变量资料的统计分析

Descriptive Statistics

统计描述

统计分析

统计推断

【例7-1】某地用随机抽样的方法对130名健康成年男性红细胞数进行了检测，资料如表1，请描述男性红细胞的情况。

表1 某地130名正常成年男子红细胞数（ $10^{12}/L$ ）

3.79*	4.57	5.19	4.86	4.28	4.67	5.37	4.98	4.45	5.88*
4.53	5.16	4.84	4.15	4.66	5.31	4.97	4.43	4.77	4.78
5.1	4.83	4.11	4.63	5.28	4.94	4.4	4.74	5.67	5.05
4.81	3.98	4.61	5.23	4.9	4.35	4.7	5.46	5.03	4.49
3.89	4.57	5.21	4.87	4.29	4.67	5.38	4.98	4.46	4.78
4.54	5.16	4.85	4.17	4.66	5.32	4.97	4.43	4.77	5.07
5.13	4.83	4.13	4.64	5.29	4.95	4.42	4.74	5.69	4.53
4.81	4.01	4.62	5.26	4.91	4.36	4.73	5.49	5.04	4.78
3.94	4.57	5.23	4.9	4.31	4.68	5.39	4.99	4.48	5.08
4.54	5.17	4.86	4.27	4.66	5.36	4.98	4.43	4.77	4.53
5.15	4.84	4.13	4.64	5.29	4.96	4.42	4.75	5.69	4.8
4.82	4.1	4.62	5.26	4.93	4.39	4.74	5.61	5.04	5.1
3.98	4.58	5.23	4.9	4.33	4.68	5.4	5	4.49	4.8

问题？？

- 表1的130个数据，无论多认真审视，也说不清这些人红细胞的情况怎样、特征如何。
- 你应该如何着手整理，整理的目的是什么？
- 结合学过的知识，你认为用什么方式描述这份资料能让人对资料内容一目了然？
- 统计描述就是解决此问题的方法，即用统计表、统计图和统计指标来描述样本数据的特征

第一节　计量资料的统计描述

- 频数表与频数分布
- 平均指标（算术均数、几何均数、中位数、众数）
- 变异指标（极差、百分位数与四分位间距、方差、标准差、变异系数）

一、频数表与频数分布
(frequency table and frequency distribution)

频数（frequency）：

变量值出现的次数，即例数

频数表（frequency distribution table）：

反映变量值与频数之间关系的统计表

表1 某地130名正常成年男子红细胞数（ $10^{12}/L$ ）

3.79*	4.57	5.19	4.86	4.28	4.67	5.37	4.98	4.45	5.88*
4.53	5.16	4.84	4.15	4.66	5.31	4.97	4.43	4.77	4.78
5.1	4.83	4.11	4.63	5.28	4.94	4.4	4.74	5.67	5.05
4.81	3.98	4.61	5.23	4.9	4.35	4.7	5.46	5.03	4.49
3.89	4.57	5.21	4.87	4.29	4.67	5.38	4.98	4.46	4.78
4.54	5.16	4.85	4.17	4.66	5.32	4.97	4.43	4.77	5.07
5.13	4.83	4.13	4.64	5.29	4.95	4.42	4.74	5.69	4.53
4.81	4.01	4.62	5.26	4.91	4.36	4.73	5.49	5.04	4.78
3.94	4.57	5.23	4.9	4.31	4.68	5.39	4.99	4.48	5.08
4.54	5.17	4.86	4.27	4.66	5.36	4.98	4.43	4.77	4.53
5.15	4.84	4.13	4.64	5.29	4.96	4.42	4.75	5.69	4.8
4.82	4.1	4.62	5.26	4.93	4.39	4.74	5.61	5.04	5.1
3.98	4.58	5.23	4.9	4.33	4.68	5.4	5	4.49	4.8

1. 频数表的编制步骤

（1）求极差（range）：即最大值与最小值之差，又称为全距。

本例极差： R=5.88-3.79=2.09（10^{12}/L）。

（2）决定组数、组段和组距：根据研究目的和样本含量n确定。组距=极差/组数，通常分10-15个组，为方便计，组距参考极差的十分之一，再略加调整。

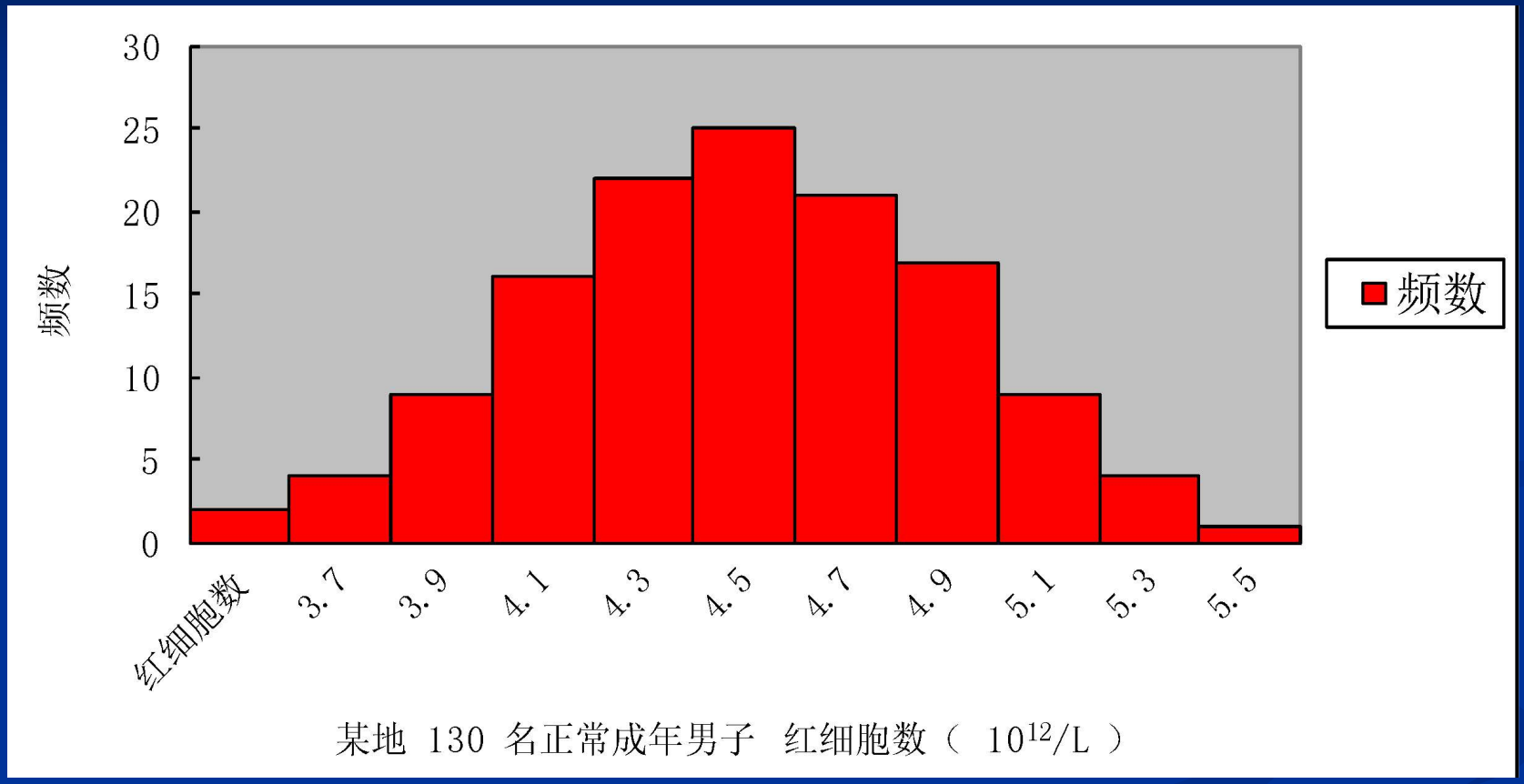
本例i= R /10=2.09/10=0.209≈0.2。

（3）列出组段：第一组段的下限略小于最小值，最后一个组段上限必须包含最大值，其它组段上限值忽略。

（4）划记计数：用划记法将所有数据归纳到各组段，得到各组段的频数。

表2 某地130名正常成年男子红细胞数频数分布

组段 （1）	划 记 （2）	频数，f （3）	组中值，X （4）	fX (5)= (3)×(4)
3.7～	丅	2	3.8	7.2
3.9～	下	4	4.0	16.0
4.1～	正下	9	4.2	37.8
4.3～	正正正–	16	4.4	70.4
4.5～	正正正正丅	22	4.6	101.2
4.7～	正正正正正	25	4.8	120
4.9～	正正正正–	21	5.0	105
5.1～	正正正丅	17	5.2	88.4
5.3～	正下	9	5.4	48.6
5.5～	下	4	5.6	22.4
5.7～5.9	–	1	5.8	5.8
合计		130		622.8


30
25
20
15
10
5
0
频数
红细胞数
3.7
3.9
4.1
4.3
4.5
4.7
4.9
5.1
5.3
5.5
频数
某地 130 名正常成年男子 红细胞数（ 10^{12}/L ）

2. 频数表的分布特征

①集中趋势(central tendency):变量值集中位置。本例在组段“4.7～4.9”。

——平均水平指标

②离散趋势(tendency of dispersion):变量值围绕集中位置的分布情况。本例4.3～5.1，共有101人，占77.7%；离“中心”位置越远，频数越小；且围绕“中心”左右对称。

——变异水平指标

二、平均指标

总称为平均数（average）反映了资料的集中趋势（*central tendency*）。常用的有：

1. 算术均数(arithmetic mean)，简称均数 (mean)

2. 几何均数(geometric mean)

3.中位数 (median)

1. 均数（mean）

(1)直接法 $\overline{X} = \frac{X_1 + X_2 + \square + X_n}{n} = \frac{\Sigma X}{n}$

(2)加权法（频数表法）基本思想：

以组中值代表组内的变量值（近似法），简化计算

$$\overline{X} = \frac{f_1 X_1 + fX_2 + fX_3 + \square + f_k X_k}{f_1 + f_2 + f_3 + \square + f_k} = \frac{\Sigma f X_i}{\Sigma f_i}$$

Σ为求和符号，读成sigma

适用条件：资料呈正态或近似正态。

表2 某地区130 名正常成年男子红细胞数（ 10^{12}/L ）的均数和标准差的加权计算

红细胞数 (1)	组中值X (2)	频数f (3)	fX_i (4)=(2)(3)	fX_i^2 (5)=(2)(4)
3.70~	3.80	2	7.60	28.88
3.90~	4.00	4	16.00	64.00
4.10~	4.20	9	37.80	158.76
4.30~	4.40	16	70.40	309.76
4.50~	4.60	22	101.20	465.22
4.70~	4.80	25	120.00	576.00
4.90~	5.00	21	105.00	525.00
5.10~	5.20	17	88.40	459.68
5.30~	5.40	9	48.60	262.44
	5.60	4	22.40	125.44
	5.80	1	5.80	33.64
	-	130	623.0	3009.12

均数＝623.0/130＝4.794

2. 几何均数（geometric mean）

$$\overline{X}_G = \sqrt[n]{X_1 X_2 \square X_n}$$

$$\lg \overline{X}_G = \frac{1}{n}(\lg X_1 + \lg X_2 + \square + \lg X_n) = \frac{\sum \lg X}{n}$$

$$\overline{X}_G = \lg^{-1} \frac{\sum \lg X}{n}$$

几何均数：变量对数值的算术均数的反对数。

几何均数的适用条件与实例

适用条件：呈倍数关系的等比资料或对数正态分布（正偏态）资料；如抗体滴度资料

血清的抗体效价滴度的倒数分别为：10、100、1000、10000、100000，求几何均数。

$$G = \lg^{-1}\left(\frac{\lg 10^1 + \lg 10^2 + \lg 10^3 + \lg 10^4 + \lg 10^5}{5}\right) = 1000$$

此例的算术均数为22222，显然不能代表滴度的平均水平。同一资料，几何均数<均数

频数表资料的几何均数

$$G = \lg^{-1}\left(\frac{\sum f_i \lg X_i}{\sum f_i}\right) = \lg^{-1}\left(\frac{f_1 \lg X_1 + f_2 \lg X_2 + \square + f_n \lg X_n}{\sum f_i}\right)$$

抗体滴度 (1)	人数,*f* (2)	滴度倒数,*X* (3)	lg*X* (4)	*f*·lg*X* (5)
1:2.5	14	2.5	0.3979	5.5706
1:10	18	10.0	1.0000	18.0000
1:40	22	40.0	1.6021	35.2462
1:160	12	160.0	2.2041	26.4492
1:640	6	640.0	2.8062	16.8372
合计	72			102.1032

3.中位数（median）
百分位数（percentile）

- 中位数：一组观察值按大小顺序排列，位置居中的那个数值称为中位数，记为*M*。
- 百分位数：一组数据从小到大排列，并分成100等份，第*x*等份之分割位置的数值称为第*x*百分位数，记为P_x
- 例如：$x=50$，记为P_{50}，读作“第五十百分位数”（即中位数）

适用情形：

适用于任意分布，常用于 ：

① 偏态分布（如发汞、尿铅）；

② 一端或两端无确定数值；

③ 分布情况不明。

常用百分位数：

P_{50}，P_{25}，P_{75}，，P_{5}，P_{95}，$P_{2.5}$，$P_{97.5}$

怎样求解中位数和百分位数？

（1）中位数计算公式与实例

先将观察值按从小到大顺序排列，再按以下公式计算：

$$Md=\begin{cases} x_{(n+1)/2} & n\text{为奇数} \\ \left(x_{n/2}+x_{1+n/2}\right)/2 & n\text{为偶数} \end{cases}$$

特点：仅仅利用了中间的1～2个数据