统计学知识点汇总

第一章：

统计学是收集、处理、分析、解析数据并从数据中得出结论的科学。

分类：描述统计、推断统计。

描述统计是研究数据收集、处理和描述的统计学方法。　推断统计是研究如何利用样本数据来推断总体特征的统计学方法（内容包括参数估计和假设检验）。

变量：每次观察都会得到不同结果的某种特征。　　分类变量：又称无序分类变量，观测结果表现为某种类别的变量。　　顺序变量：又称有序分类变量，观测结果表现为某种有序类别的变量。

数值变量：又称定量变量，观测结果表现为数字的变量。

数据：1、分类数据 2、顺序数据 3、数值型数据

总体：包含所研究的全部个体（数据）的集合。

样本：从总体中抽取的一部分元素的集合。

样本量：构成样本元素的数目。

抽样方法：1、简单随机抽样 2、分层抽样 3、系统抽样 4、整群抽样

简单随机抽样：从含有 N 个元素的总体中，抽取 n个元素组成一个样本，使得总体中的每一个元素都有相同的机会（概率）被抽中。

分层抽样：也称分类抽样，在抽样之前先将总体的元素划分为若干层（类），然后从各个层中抽取一定数量的元素组成一个样本。

软件应用：用 Excel抽取简单随机样本。

第二章：

一、定性数据的图示：1、条形图 2、帕累托图 3、饼图 4、环形图

条形图：是用宽度相同的条形来表示数据多少的图形，用于观察不同类别的多少或分布状况。

帕累托图：是按各类别出现的频数多少排序后绘制的条形图。通过对条形的排序，容易看出哪类频数出现的多，哪类出现的少。

饼图：主要用于表示一个样本（或总体）中各类别的频数占全部频数的比例。

用图表展示定量数据：

生成定量数据的频数分布表时，需要先将原始数据按照某种标准分成不同的组别，然后统计出各组别的数据频数即可。

一组数据所分的组数 K应不少于 5组且不多于 1 5组。

组距 =（最大值 最小值） 组数 组数 =全距 组距

每组组距均相等称为等距数列，反之则为异距数列 在比较等距数列与异距数列的次数分布时常用：

次数密度 =本组次数 本组组距

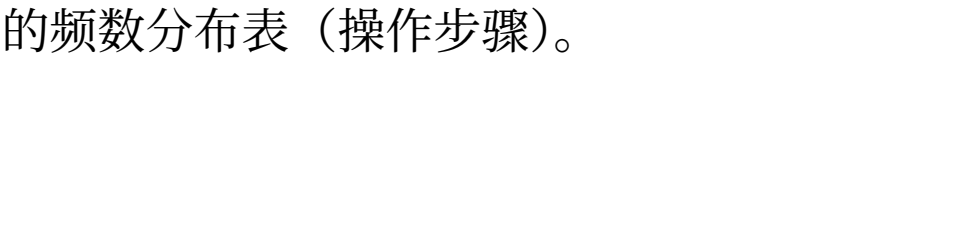
2组中值 c l a s s m i d p o i n t

组中值 =（本组上限 +本组下限） / 或组中值 =（本组假定上限 +本组假定下限） / 2

二、定量数据的图示：1、分组数据看分布：直方图 2、未分组数据看分布：茎叶图和箱线图、垂线图和误差图

最小值 2 5四分位数 中位数 7 5四分位数 最大值

箱线图的示意图：



3、两个变量间的关系：散点图是用二维坐标展示两个变量之间关系的一种图形。

4、比较多个样本的相似性：雷达图和轮廓图

雷达图是从一个点出发，用每一条射线代表一个变量，多个变量的数据点连接起来成线，即围成一个区域，多个样本围成多个区域，就是雷达图，利用它也可以研究多个样本之间的相似程度。

5、掌握各种图标的绘制，直方图与条形图的区别、茎叶图与直方图的区别。

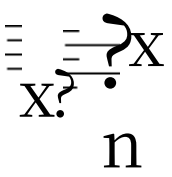
三、合理使用图表

E x c应用：生成定性 定量数据的频数分布表（操作步骤）。

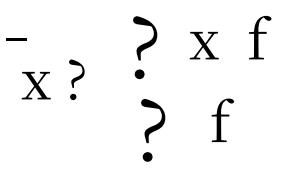
第三章：用统计量描述数据

一、水平的度量：平均数：计算形式： =总体标志总量 总体单位总量

（一）简单均数



（二）加权均数



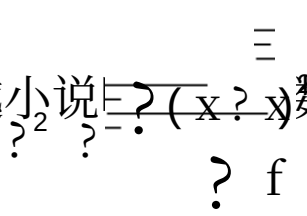
中位数：是一组数据排序后处于中间位置的数值，用 M e表示。

众数：是一组数据中频数最大的变量值，直观地反映了数据的集中趋势。是度量定类数据集中趋势的测度。一般用 M o表示。

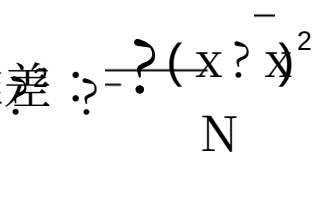
四分位数：是一组数据排序后处于 2 5%和 7 5%位置上的值。它是通过 3个点将全部数据等分为四部分，其中每部分包含 2 5%的数据。显然，中间的四分位数就是中位数，因此通常所说的四分位数是指处在 2 5%位置上和处在 7 5%位置上的数值。

二、差异的度量：1、极差是一组数据的最大值与最小值之差，也称全距，用 R 表示。由于极差只是利用了一组数据两端的信息，因而容易受极值端的影响，不能全面反映差异状况。

2、四分位差是一组数据 7 5%位置上的四分位数与 2 5%位置上的四分位数之差，也称为内距或四分间距，用 Q d表示，反映了中间 5 0%数据的离散程度，其数值越小说明数值越集中，数值越大说明中间的数值越分散，四分位差不受极值的影响。



3、样本方差和标准差：



方差是度量数值变量离散程度的基本测度。 n个同性质独立变量和的方差等于各个变量方差之和。

n个同性质独立变量平均数的方差等于各变量方差平均数的 1 / n

4、标准分数：测度每个数值在该组数据中的相对位置，并可以用它来判断一组数据中是否有离群点，它是某个数据与其平均数的离差除以标准差后的值。

三、比较几组数据的离散程度：离散系数是一组数据的标准差与其相应的平均数之比，它消除了数据水平高低和计量单位对标准差大小的影响。主要用于比较不同样本数据的离散程度，离散系数越大说明数

据的离散程度也越大，离散系数越小说明数据的离散程度也越小。

计算公式是：　　V s = S / x

四、分布形状的度量

偏态系数　　$SK=\frac{\sum(x-\bar{x})^3}{(n-1)(n-2)s^3}$ or $SK=\frac{\sum(x-\bar{x})^3 f}{(n-1)(n-2)s^3}$

偏态系数为 0时，数据是对称分布；偏态系数为负数时，数据是左偏分布，也称为负偏态；偏态系数为正数时，数据是右偏分布，也称为正偏态。偏态系数越大表明偏离程度越大。

峰态系数　　$K=\frac{\sum(x-\bar{x})^4}{(n-1)(n-2)(n-3)s^4}$ or $K=\frac{\sum(x-\bar{x})^4 f}{(n-1)(n-2)(n-3)s^4}$

峰度系数为 3时，数据是对称分布；峰度系数大于 3时，数据是尖峰分布；峰度系数小于 3时，数据是平峰分布。

软件应用：用 E x c l计算描述统计量。

第一步：选择【工具】 -【数据分析】。在分析工具中选择【描述统计】。单击【确定】。

第二部：将原始数据所在的区域输入【输入区域】；在【输出选项】中选择结果的输出位置；选择【汇总统计】。单击【确定】

第四章：概率分布

事件发生可能性大小的度量就是概率。

随机变量的概率分布 1、有些随机变量只能取有限个值, 称为离散型随机变量。 2、有些则可以取一个或

多个区间中的任何值，称为连续性随机变量。

描述随机变量集中程度的统计量称为期望值。

一、离散型随机变量的概率分布（二项分布、超几何分布、泊松分布）

1、二项分布（binomial：binary 二项的；nomial 命名的）：n次独立事件，每次成功概率为 p（不成功概率为 q）。n次试验，成功 x次，每次成功的概率 p，则成功 x次的概率 P为 $P = C_n^x p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}$

2、超几何分布

（hyper geometric）：样本抽取后不放回时的离散型概率分布。N个总体有 T次成功次数，则抽取 n次中有 x次成功的概率。例：6名业务骨干中的 3人在职时间超过了 5年。随机抽取这 6人中的 4人，恰好有 2人在职时间超过了 5年的概率。

$$P = \frac{\binom{T}{x}\binom{N-T}{n-x}}{\binom{N}{n}} = \frac{\binom{3}{2}\binom{6-3}{4-2}}{\binom{6}{4}} = 0.6$$

3、泊松分布（Poisson distribution）

事件在一段时（空）间内连续发生时指定次数事件的概率。

例：某网店平均每小时接单 5个。现在随机抽取 1小时观察，恰好接 3个定单的概率是

$$P = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{5^3 e^{-5}}{3!} = 0.1404$$

λ为事件的均值

二、连续性随机变量的概率分布（正态分布、指数分布、均匀分布）

正态曲线的性质：1、正态曲线的图形是关于 x μ对称的钟形曲线，且峰值在x μ处。

多个区间中的任何值，称为连续性随机变量。

描述随机变量集中程度的统计量称为期望值。

一、离散型随机变量的概率分布（二项分布、超几何分布、泊松分布）

1、二项分布（binomial：binary 二项的；nomial 命名的）：二项分布的实验是指符合以下条件的实验：每次实验只有两个互斥的结果，即成功和不成功；每次实验成功概率为 p（不成功概率为 q）。n次试验，成功 x次，每次成功的概率 p，则成功 x次的概率 P为 $P = C_n^x p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x}$

2、超几何分布

（hyper geometric）：样本抽取后不放回时的离散型概率分布。N个总体有 T次成功次数，则抽取 n次中有 x次成功的概率。例：6名业务骨干中的 3人在职时间超过了 5年。随机抽取这 6人中的 4人，恰好有 2人在职时间超过了 5年的概率。

$$P = \frac{\binom{T}{x}\binom{N-T}{n-x}}{\binom{N}{n}} = \frac{\binom{3}{2}\binom{6-3}{4-2}}{\binom{6}{4}} = 0.6$$

3、泊松分布（Poisson distribution）

事件在一段时（空）间内连续发生时指定次数事件的概率。

例：某网店平均每小时接单 5个。现在随机抽取 1小时观察，恰好接 3个定单的概率是

$$P = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{5^3 e^{-5}}{3!} = 0.1404$$

λ为事件的均值

二、连续性随机变量的概率分布（正态分布、指数分布、均匀分布）

正态曲线的性质：1、正态曲线的图形是关于 x=μ对称的钟形曲线，且峰值在x=μ处。

减估计误差。

在区间估计中，由样本估计量构造出的总体参数在一定置信水平下的估计区间称为置信区间，其中区间的最小值称为置信下限，最大值称为置信上限。

一般的，如果将构造置信区间的步骤重复多次，置信区间中包含总体参数真值的次数所占的比例称为置信水平，也称为置信度或置信系数。

置信水平 = 1α-

3、评价估计量的标准

无偏性：是指估计量抽样分布的期望值等于被估计的总体参数。

有效性：是指估计量的方差尽可能小。

一致性：是指随着样本量的增大，点估计量的值越来越接近被估计总体的参数。

二、一个总体参数的区间估计

母体条件	区间估计
正态，大样本，方差已、未知	$x \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
正态，小样本，方差未知	$\bar{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$
分布未知，大样本时	$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$

类比总体比例 方差的区间估计

三、两个总体参数的区间估计

母体条件	区间估计
正态分布，方差已知	$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
正态分布，方差未知	$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
正态分布，方差未知且不相等，小样本	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}(f)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

类比两个总体比例之差 方差比的区间估计

四、样本量的确定

1、估计总体均值时样本量的确定。

2、估计总体比例时样本的确定。（熟练掌握其公式）