名词解释

1.统计学
收集、处理、分析、解释数据并从数据中得出结论的科学

2.描述统计与推断统计
描述统计：数据收集、处理、汇总、图表描述、概括与分析
推断统计：利用样品推断总体特征

3.截面数据与时间序列
截面数据：同一或近似同一时间点上收集的数据
时间序列：不同时间上同一事物随时间变化的数据

4.品质数据与定量数据
品质数据：包括分类数据和顺序数据，主要是用文字描述的一类，亦称定性数据
定量数据：按照数字尺度表现为数值的一类数据

5.参数与统计量
参数：用来描述总体特征的概括性数字性度量
统计量：用来描述样本特征的概括性数字度量，也可以表述成关于样本的不含参数的可测实值函数。

6.总体的特征
同质性：样本总体都必须具有某一共同的品质标志属性或数量标志数值
大量性：构成总体的总体单位数目要足够多

7.概率抽样与非概率抽样
概率抽样：遵循随机原则进行的抽样，总体中每个单位都有一定的机会被选入样本。
非概率抽样：不依取随机原则，而是根据研究目的对数据的要求，采用某种方式从总体中抽出部分单位对其调查。

8.概率抽样与非概率抽样的区别
概率抽样按一定概率以随机原因抽取，每个被抽中概率是已知的，用样本对总体目标量可以估计。非概率抽样适合于探索性的研究，不能使用样本的结果对总体相应的参数进行推断，也无法得到总体参数的置信区别。

9.分层抽样与系统抽样
分层抽样：是将抽样单位按某种特征或某种规则划分为不同的层，然后从不同的层中独立随机抽样样本。
系统抽样：将总体按一定顺序排列，随机抽取一个初始单位，然后按事先规定好的规则确定其他样本单位。
分层抽样的样本结构与总体的特征相近，提高了估计的精度，系统抽样对估计量方差的估计比较困难。

10.多阶段抽样与整群抽样

多阶段抽样和整群抽样都是概率抽样的一种，整群抽样将所有单位合并成群，抽取群，然后对群中的所有样本全部实施调查。多阶段抽样时不断抽取群，直至最终抽样单位。

11.判断抽样及其分类

判断抽样是指研究人员根据经验、判断和对研究对象的了解，有目的地选择一些单位作为样本，实施时根据不同的目的有重点抽样、典型抽样、代表抽样等方式。判断抽样包括重点抽样、典型抽样和代表抽样。

12.方便抽样与判断抽样

方便抽样最大的特点是容易实施，调查的成本低，但是这种方法无法对总体的有关的参数进行推断。判断抽样是主观的，样本选择的好坏取决于调研者的判断、经验、专业程度和创造性。成本低易操作，但由于是人为确定的，没有依据随机的原则，无法对总体的有关参数进行估计。

13.系统抽样与配额抽样

将总体中的所有单位按一定顺序排列，在规定的范围内随机地抽取一个单位作为初始单位，然后按事先规定好的规则确定其他样本单位，这种方法叫系统抽样。配额抽样类似于概率抽样中的分层抽样，它首先将总体中的所有单位按照一定的标志分为若干类，然后在每个类中采用方便抽样或判断抽样的方式选取样本单位。这种抽样方式操作比较简单，而且可以保证样本的结构和总体的结构类似。但在抽取具体样本单位时，并不是随机原则。

14.抽样误差与非抽样误差定义及区别

抽样误差：是由于抽样的随机性引起的样本结果与总体真值之间的误差，主要存在于概率抽样中。抽样误差并不是针对某个具体样本的检测结果与总体真实结果的差异而言的，抽样误差描述的是所有样本可能的结果与总体真值之间的平均性差异。非抽样误差是相对抽样误差而言的，是指除抽样误差之外的，由于其他原因引起的样本观察结果与总体真值之间的差异。非抽样误差在任何抽样方式中都有可能产生。

抽样误差并不是针对某个具体样本的检测结果与总体真实结果的差异而言的，抽样误差描述的是所有样本可能的结果与总体真值之间的平均性差异非抽样误差在任何抽样方式中都有可能产生。非抽样误差分为：抽样框误差、回答误差和无回答误差。

15.抽样框误差

抽样框误差：抽样框信息不完整而导致的误差。一个好的抽样框应该是抽样框子中的单位和研究总体中的单位有一一对应的关系。

16.众数、中位数、平均数

众数是一组数据中出现次数最多的变量值，用 M_o 表示，众数主要用于测量分类数据的集中趋势；中位数：是一组数据排序后处于中间位置上的变量值，用

M_e表示，中位数主要用于测度顺序数据的集中趋势；平均数是集中趋势的最主要测度值，不适合分类数据和顺序数据。众数是一组数据分布的峰值，不受极端值的影响，其缺点就是不唯一性，而且只有当数据量较大时才有意义。中位数是一组数据中间位置上的代表值，不受极端值的影响，当数据分布偏斜程度较大时，使用中位数是好的选择。平均数是应用最广泛的集中趋势测度值，但对于偏态的数据，平均数代表性较差。

17 平均数的统计思想

平均数是一组数据的重心所在，是数据误差相互抵消后的必然结果，反映出事物必然性的数量特征。

18.变异系数与方差

方差和标准差是反映数据分散程度的绝对值，其数值的大小一方面受原变量值本身水平高低的影响，也就是与变量的平均数大小有关。变异系数是测度数据离散程度的相对统计量，主要是用于比较不同样本数据的离散程度。

19.切比雪夫不等式

切比雪夫不等式原始形式：$P(X > r) \le \frac{E(X)}{r}$，r 为任意常数，X 为随机变量；而我们平时看到的切比雪夫不等式 $P(|X - \mu| \ge \varepsilon) \le \frac{Var(X)}{\varepsilon^2}$，$\mu$ 为随机变量 X 的期望，切比雪夫不等式推导出 3-sigma 原理。另外要会证明该不等式。

20.偏度与峰度

设 X 的三阶矩存在，则

$$\beta_1 = \frac{v_3}{v_2^{3/2}} \Rightarrow \text{X 分布的偏度}$$

β_1>0：称分布正偏或右偏；β_1=0：称分布对称；β_1<0：称分布左偏或负偏。

设 X 的四阶矩存在，则：

$$\beta_2 = \frac{v_4}{v_2^2} - 3 \Rightarrow \text{X 分布的峰度}$$

β_2<0，则标准化后的分布形状比标准正态分布更平坦，称为低峰度；

β_2=0，则标准化后的分布形状与标准正态分布相当；

β_2>0，则标准化后的分布形状比标准正态分布更尖峭，称为低峰度；

21.概率的统计定义

在相同条件下随机试验 n 次，某事件 A 出现 m 次$(m \le n)$，则 m/n 称为事件

A 发生的概率。随着 n 的增大，该频率围绕某一常数 p 上下波动，且波动的幅度逐渐减小，趋于稳定，这个频率的稳定值即为该事件的概率。

22.互斥与独立区别

互斥事件一定是不独立的，不独立的事件不一定是互斥的。不互斥的事件可能是独立的也可能是不独立的，然而独立事件不可能是互斥的。

23.二项分布的特点

特点：试验包括 n 个相同的试验；每次试验只有可能两个结果；出现一种结果 p 的概率相同；试验是相互独立的。

24.泊松定理

二项分布的 n 值比较大，成功的概率 p 比较小，np 适中的时候，二项分布用泊松分布近似具有很好的效果。

$$C_n^x p^x (1-p)^{n-x} \approx \frac{\lambda^x e^{-\lambda}}{x!}$$

25.次序统计量与充分统计量

次序统计量：样本观测值 $x_1, x_2, \cdots, x_n$ 经过排序后，$x_{(i)}$ 称为第 i 个次序统计量。在连续场合，随机变量 X 的密度函数为 $f(x)$，分布函数为 $F(x)$，则第 i 个次序统计量的密度函数为：

$$f_{x_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} f(x) \cdot (F(x))^{i-1} \cdot (1-F(x))^{n-i}$$

在这里注意最大值分布和最小值分布就好。

充分统计量：统计量加工过程中一点信息都不损失的统计量，判断充分统计量的方法称为因子分解定理。

26.枢轴量

设法构造样本和未知参数 θ 的函数 $G = G(x_1, x_2, \cdots, x_n, \theta)$ 使得 G 的分布不依赖与未知参数，一般称具有这种性质的函数 G 为枢轴量。必须分清楚枢轴量与统计量之间的区别，枢轴量不是统计量，因为枢轴量含有未知参数，只是枢轴量的分布不依赖于参数，能形成一个具体的分布，而统计量是不含有参数的。

27.柯西分布

柯西分布的密度函数 $f(x)$ 是形式如下：

$$f(x|\theta) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}, -\infty < x < \infty, -\infty < \theta < \infty,$$

注意柯西分布的一阶矩不存在。

28.中心极限定理以及其来源

1.林德伯格-勒维中心极限定理（独立同分布条件下）

设 $\{X_n\}$ 为独立同分布的随机变量序列，并且 $EX_i = \mu, DX_i = \sigma^2 > 0$,即方差有限,记$\forall x \in R$，恒有：当$n \to \infty$时,$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

2.棣莫弗-拉普拉斯定理（二项分布条件下）

设$\{Y_n\}$服从二项分布 B(n,p)，记$\forall x \in R$，则有：

$$\lim_{n\to\infty} P\{\frac{Y_n - np}{\sqrt{np(1-p)}} \le x\} = \Phi(x)$$

其$\Phi(x)$为正态分布。当用正态分布来作为二项分布的近似计算时，需要作出修正，$P(L \le \mu_n \le U) = P(L - 0.5 \le \mu_n \le U + 0.5)$

林氏中心极限定理证明可以选择矩母函数的泰勒公式二阶展开近似方法。

29.点估计与区间估计的区别

点估计就是用样本统计量的某个取值直接作为总体参数的估计值，任何值都可以做未知参数的点估计值，一个点估计值的可靠性由抽样标准误差来衡量的。

区间估计是在点估计的基础上，给出总体参数的一个区间范围，该区间通常由样本统计量加减估计误差得到的；区间估计三要素：点估计值，抽样平均度，估计的可靠度。

30.相合估计

设$\theta \in \Theta$为未知参数，$\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \cdots, x_n)$是$\theta$的一个估计量，n 为样本容量，若对$\forall \varepsilon > 0$，有：

$$\lim_{n\to\infty} P(\left|\hat{\theta}_n - \theta\right| \ge \varepsilon) = 0$$

则称$\hat{\theta}_n$为参数θ的相合估计。相合性是对估计的一个最基本的要求。

31.匹配样本

一个样本中的数据对应于另一个样本的数据，通常两样本的数目相同。

32.渐近无偏估计

$\hat{\theta}_n$为随机变量θ的估计量，且满足$E\hat{\theta}_n \ne \theta, \lim_{n\to\infty} \hat{\theta}_n = \theta$，此时称$\hat{\theta}_n$为$\theta$的渐近无偏估计量。

33.最小方差无偏估计

对参数估计问题，设$\hat{\theta}$为θ的一个无偏估计，如果对另外**任意**一个θ的无偏估计$\tilde{\theta}$，在参数空间Θ上都有：

$$D(\hat{\theta}) \le D(\tilde{\theta})$$

则称$\hat{\theta}$为θ的一致最小方差无偏估计，记为 UMVUE。一般而言，若 UMVUE 存在，则它一定是统计量的函数。

34.经验分布函数

经验分布函数：设$(x_1, x_2, \cdots, x_n)$为总体样本$(X_1, X_2, \cdots, X_n)$的一个观测值，若将样本观测值x_i由小到大进行排列，为$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$，对任意实数，称函数：

$$F_n(x) = \begin{cases} 0, x < x_{(1)} \\ \dfrac{k}{n}, x_{(k)} \le x < x_{(k+1)}, k = 1,2,\cdots,n-1 \\ 1, x \ge x(n) \end{cases}$$

为样本$(X_1, X_2, \cdots, X_n)$的经验分布函数。

35.费希尔信息量

费希尔信息量：$I(\theta) = E[\dfrac{\partial}{\partial\theta}\ln p(x;\theta)]^2$，当总体分布的$I(\theta)$越大，包含总体分布参数$\theta$的信息越多，这可以比较两个统计量哪一个具有总体参数的信息更加充分。

36.第一类错误与第二类错误的关系与区别

犯第一类错误的概率α与犯第二类错误的概率β可以用一个函数表示，即所谓的势函数。设检验问题：

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

的拒绝域为 W，则样本观测值落入 W 的概率称为该检验的势函数。记为：

$$g(\theta) = P_\theta(X \in W), \theta \in \Theta = \Theta_0 \cup \Theta_1 \Rightarrow g(\theta) = \begin{cases} \alpha(\theta), \theta \in \Theta_0 \\ 1-\beta(\theta), \theta \in \Theta_1 \end{cases}$$

在样本量给定的条件下，α与β增减互反相互引导，我们找不到一个能让α

与β同时减小的检验。同时使α与β同时减小的方法就是增大样本量

37.检验功效

$1-\beta$是反映统计检验判别能力大小的重要标志，我们称之为检验功效或检验力。在犯第一类错误的概率得到控制的条件下，犯取伪错误的概率也要尽可能地小，或者说，不取伪的概率$1-\beta$尽可能增大。$1-\beta$越大，意味着当原假设不真实时，检验判别出原假设不真实的概率越大，检验的判别能力就越好；$1-\beta$越小，意味着当检验的判别能力越差。

38.小概率原理

在一次试验中，小概率事件发生的概率几乎为 0，但是如果真的发生了，就有理由拒绝原假设，这是假设检验的思想。

39.p 值定义及作用

P 值是指当原假设为真时，所得到的样本观察结果或者更极端结果出现的概率。另一种定义就是，在一个假设检验中，利用样本观测值能够做出拒绝原假设的最小显著性水平。当$p<\alpha$时，拒绝原假设，当$p\geq\alpha$时，接受原假设。P 值越小，拒绝原假设的理由越充分。P 值的大小取决于：（1）样本数据域原假设的差异；（2）样本量的大小；（3）被假设参数的总体分布。

40.参数检验与假设检验的异同点

相同点：

（1）都根据样本推断总体信息；

（2）都以抽样分布为理论依据，推断结果均有风险；

（3）对同一问题的参数，使用同一样本、同一统计量、同一分布，因而二者可以互换；

不同点：

（1）参数估计是以样本资料估计总体参数的可能范围，假设检验是以样本资料检验对总体参数的先验假设是否成立。

（2）区间估计只是以估计值为中心的双侧的置信区间，假设检验既有双侧检验又有单侧检验。

（3）区间估计立足于大概率，以$1-\alpha$的可信度去估计总体参数的置信区间；假设的的检验立足于小概率，通常是给定的很小的显著性水平α去检验对总体参数的先验假设是否成立。

41.示性函数

$$I(x)=\begin{cases}1, x\in\Phi\\0, x\notin\Phi\end{cases}$$

示性函数在以后的学习中经常用到，体会用处就可以了。

42.方差分析

方差分析：通过检验各总体均值是否相等来判断分类型自变量对数值型因变量是否有显著影响。基本思想：实际是通过对数据误差来源的分析来判断不同总体的均值是否相等。

43.方差分析的基本假定与多重比较

方差分析的假定：

（1）每个总体（水平）均应服从正态分布；

（2）各个总体的方差σ^2相同；

（3）观测值是独立的；

44.多重判定系数

$R^2 = \frac{SSR + SSC}{SST}$（多重判定系数）

45.相关分析与回归分析、函数关系与相关关系

相关分析目的在于测度变量之间的关系强度，它所使用的测度工具就是相关系数；而回归分析侧重于考察变量之间的数量伴随关系，并通过一定的数学表达式将这种关系描述出来。

变量关系分函数关系和相关关系，函数关系是指一一对应的确定关系；相关关系是指变量之间客观存在的不确定的数量关系。相关分析是指对两个变量之间线性关系的描述与度量。

46.一元线性回归模型的基本假定

①y 与 x 具有线性关系；

②在重复抽样中，自变量 x 的取值是固定的，即假定 x 是非随机的。在①和②的假定下，给定的 x 值，y 取值对应一个分布，$E(y) = \beta_0 + \beta_1 x$；

③误差项ε是一个期望值为 0 的随机变量，$E(\varepsilon) = 0$；

④对于所有 x，ε的方差σ^2相同；

⑤误差项$\varepsilon \sim N(0, \sigma^2)$，且独立。

47.多元回归分析中，t 检验与 F 检验的作用

在一元线性回归中，自变量只有一个，相关系数检验，F 检验和 t 检验是等价的，但在多元回归分析中，F 检验只用来检验总体回归关系的显著性，而 t 检验则是检验各回归系数的显著性。

48 多重共线性及其表现形式、补救

定义：当回归模型中两个或两个或两个以上的自变量彼此相关时。在多元线性回

归中无多重共线性假定：Rank（X）=k。多重共线性带来的主要麻烦是对单个回归系数的解释和检验。

形式：

①变量之间高度相关时，可能会使回归的结果造成混乱，甚至会把分析引入歧途；

②可能对回归参数的估计值正负号产生影响；

③参数估计值不确定（因为 R(X)<k，此时 $|X'X| = 0$），参数估计值的方差无限大。

补救：

①修正多重共线性：这种方法有很多，比如：剔除变量；变换模型的形式，一般有差分形式，赋权重形式；利用非样本先验信息、截面数据域时间序列数据并用、变量变换，一般有计算相对指标、小类指标合并成大指标，名义数据转换成实际数据等等。

②逐步回归法：用被解释变量对每个考虑的解释变量作简单回归，将贡献最大的解释变量留住，以此为基础，再次逐个回归。除此之外，还有向前选择，向后剔除，最优子集。

③岭估计法：$\tilde{\beta}(k) = (X'X + kI)X'Y$，I 为单位矩阵，k 为常数，k 的选择原则是使得均方误差 $MSE[\tilde{\beta}(k)]$ 达到最小，最优的 k 依赖于 β 和 σ^2。

49.随机游走过程

一阶自回归模型：t 期观测值与 t-1 期观测值之间的线性关系 $Y_t = rY_{t-1} + \varepsilon_t$，即当 r=1 时，称 $Y_t = Y_{t-1} + \varepsilon_t$ 为随机游走过程，而且随机游走过程是非平稳的。

50.季节指数

季节指数：刻画了序列在一年度内各个月份或各季度的典型季节特征。在乘法模型中，季节指数是以其平均数等于 100%为条件而构成的，它反映了某一月份或季度的数值占全年平均数值的大小。若无季节变动，则各期的季节指数为 100%。