第八章 数据的统计描述

教学内容	第一节 概率论基础
教学目标	**知识目标：** 了解随机试验与随机事件的概念；理解随机变量及分布函数的概念；掌握常见的数字特征的意义及其求法。 **能力目标：**通过实例演示、问题驱动等方式激发学生学习的积极性，通过观察对比、学生交流、师生交流、小组探究等多种形式， 培养学生的逻辑思维能力、分析判断能力、和解决问题能力。 **素质目标：** 培养学生敢于质疑、善于分析、勇于创新的精神；培养学生的自主学习意识和团队协作精神；培养学生脚踏实地、不畏艰辛、锲而不舍的精神。
授课学时	2 学时
教学重点	随机变量理论
教学难点	概率统计模型的建立。
教学方法	采用问题驱动法、案例演示法、启发式讲授法及自主学习法相结合，以教师的讲解为主，学生的课堂报告、分组讨论为辅，充分调动学生学习的主动性和思考问题的积极性。
教学手段	以课堂讲授为主，主要是多媒体课件和板书相结合的形式。同时借助在线资源，如慕课、雨课堂等平台。
教学过程	**（一）由身边的随机现象出发，引出随机试验和随机事件的概念（约 15 分钟）** 【教师活动】介绍“掷骰子观察点数”和“买彩票是否中奖”两个生活中的小现象，引导学生思考随机性现象和确定性现象的区别，引出概念。 知识点 1：随机性现象是指发生的结果是不可预知的现象， 也称为偶然现象。为了研究随机现象的规律需要进行试验，概率

论中通常讨论具有以下特点的试验：

（1）可重复：试验可以在相同的条件下多次重复进行。

（2）多结果：试验有多个可能的结果，这些结果在试验前是已知的。

（3）随机性：每次试验将会出现哪个结果在试验前是无法预知的。

具备上述三个特点的试验称为随机试验。

知识点 2： 随机试验的结果称为随机事件，简称为事件。随机试验的所有可能结果组成的集合称为样本空间，记作Ω 。样本空间中的每个元素（即随机试验的每个可能结果）称为样本点， 用 ω 表示。由单个样本点组成的集合称为基本事件，任意一个随机事件可以表示为由满足某些条件的样本点所组成的样本空间的一个子集合。

【学生活动】讨论身边存在的随机性现象，并举例说明。

【设计意图】从学生熟知的生活现象出发，更能产生代入感， 提高学生的学习主动性。

（二）采用讲授法，介绍随机变量及其分布函数的概念与性质。（约 15 分钟）

知识点 3: 设 Ω 是随机试验 E 的样本空间，称定义在Ω 上的实函数 $X=X(\omega)$ 为随机变量。对任意的实数 x ，称函数$F(x)=P(X\le x)$ 为随机变量 X 的分布函数。

分布函数 $F(x)$ 具有如下的性质：

（1）单调性：对任意的实数$a<b$ ，有 $F(a)\le F(b)$ ；

（2）规范性： $F(+\infty)=\lim\limits_{x\to+\infty}F(x)=1, F(-\infty)=\lim\limits_{x\to-\infty}F(x)=0$ ；

（3）右连续性：对任意的实数 x ， $F(x+0)=\lim\limits_{t\to x+0}F(t)=F(x)$ 。

如果随机变量 X 的可能取值为有限个或可列个，则称 X 为离散型随机变量。设 X 的所有可能取值为 $x_1,x_2,\cdots,x_n,\cdots$ ，则称

$$P(X=x_i)=p_i, i=1,2,\cdots$$

为随机变量 X 的分布列。其分布函数相应的可以表示为

$$F(x)=\sum_{x_i\le x}P(X=x_i).$$

如果随机变量 X 的分布函数为 $F(x)$，存在非负可积函数 $f(x)$， 使得对于任意实数 x， 都有

$$F(x)=\int_{-\infty}^{x}f(t)dt,$$

则称 X 为连续型随机变量，称 X 为 $f(x)$ 的密度函数。

知识点 4: 设Ω 是随机试验 E 的样本空间， $X=X(\omega)$ 和$Y=Y(\omega)$ 是定义在Ω 上的两个随机变量，则称(X,Y) 为二维随机变量。

对任意的实数x,y，称二元函数$F(x,y)=P(X\le x,Y\le y)$ 为随机变量 (X,Y) 的分布函数。

（三）由“投资的收益与风险”案例出发，引出期望与方差的定义，介绍相关的数字特征。（约 25 分钟）

知识点 5：数学期望

设离散型随机变量X 的分布列为$P(X=x_i)=p_i,i=1,2,\cdots$，若级数 $\sum_{i=1}^{\infty}x_i p_i$ 绝对收敛，则称$\sum_{i=1}^{\infty}x_i p_i$ 为随机变量 X 的数学期望，简称为期望或均值，记为 $EX=\sum_{i=1}^{\infty}x_i p_i$ 。

设连续型随机变量 X 的密度函数为 $f(x)$，若积分$\int_{-\infty}^{+\infty}xf(x)dx$ 绝对收敛，则称$\int_{-\infty}^{+\infty}xf(x)dx$ 为随机变量 X 的数学期望，简称为期望或均值，记为 EX，即$EX=\int_{-\infty}^{+\infty}xf(x)dx$ 。

知识点 6：设X 为随机变量，若期望 $E(X-EX)^2$ 存在，则称其为随机变量X 的方差，记为 DX 。即

$$DX=E(X-EX)^2.$$

显然，有：

$$DX=E(X^2)-(EX)^2。$$

由定义可以看出，方差是刻画随机变量 X 取值的分散程度的数字特征。若随机变量 X 的方差越小，则 X 的取值越集中；若随机变量 X 的方差越大，则 X 的取值越分散。

知识点 7：协方差与相关系数设 (X,Y)

为二维随机变量，则称

$$Cov(X,Y)=E[(X-EX)(Y-EY)]$$

为随机变量 X 与 Y 之间的协方差。

称

$$\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{DX}\sqrt{DY}}$$

为随机变量 X 与 Y 之间的相关系数。

相关系数 ρ_{XY} 是一个反映 X 与 Y 之间的线性相关程度强弱的一个数字特征，其取值满足 $|\rho_{XY}|\le 1$。当 $|\rho_{XY}|=1$ 时，随机变量 X 与 Y 之间存在完全的线性关系。而当 $|\rho_{XY}|<1$ 时，X 与 Y 之间线性相关程度随着 $|\rho_{XY}|$ 的减小而越弱。当 $\rho_{XY}=0$ 时，X 与 Y 之间的线性相关性完全消失，此时，称 X 与 Y 不相关。

（四）采用案例演示法、问题驱动法，结合企业的产品检测间隔的实际问题，建立概率统计模型（30 分钟）

【教师活动】给出具体的概率统计模型案例，让学生自主练习，并随机选择学生到台上阐述自己建立的模型，最后老师总结步骤。

案例 1：在某企业的生产流水线中所使用的某种零件的寿命 X 服从参数为 λ 的指数分布 $E(\lambda)$。零件是否损坏往往需要通过检测才能确定。试确定该零件的最优检测间隔。

问题分析：因为零件是否损坏需要通过检测才能确定，如果检测时间间隔过大，导致流水线经常处于故障状态，会造成故障损失；如果检测时间间隔过小，则会产生过多不必要的检测费用， 因此应存在合理的检测间隔。

模型假设：

① 只考虑一类零件的检测问题，且零件的寿命$X \sim E(\lambda)$；

② 零件损坏需要通过检测才能确定，检测所需时间忽略不计；

③ 若检测发现零件损坏，则立即更换，更换时间忽略不计；若零件未损坏，则不进行更换；

④ 若零件损坏而未发现，会造成损失。

符号说明：

T：检测时间间隔；

c_i：零件的一次检测所需的费用；

c_f：零件的一次更换所需的费用；

c_d：零件损坏后的单位时间损失费。

模型建立： 设该零件相邻两次更换的时间长度为一个周期Y。显然，周期长度 Y 为零件寿命 X 的函数，其关系式为

$$Y=\begin{cases} T, 0<X\le T \\ 2T, T<X\le 2T \\ \cdots\cdots \\ (n+1)T, nT<X\le (n+1)T \\ \cdots\cdots \end{cases}.$$

一个周期的期望长度为

$$EY=\sum_{n=0}^{\infty}(n+1)T\int_{nT}^{(n+1)T}\lambda e^{-\lambda x}dx=\sum_{n=0}^{\infty}(n+1)T[e^{-n\lambda T}-e^{-(n+1)\lambda T}]$$
$$=\sum_{n=0}^{\infty}Te^{-n\lambda T}=\frac{T}{1-e^{-\lambda T}}$$

一个周期内的总费用为

$$Z=c_f+(n+1)c_i+[(n+1)T-X]c_d.$$

一个周期内的期望费用为

$$EZ=\sum_{n=0}^{\infty}\int_{nT}^{(n+1)T}\{c_f+(n+1)c_i+[(n+1)T-x]c_d\}\lambda e^{-\lambda x}dx$$
$$=c_f+\frac{c_i}{1-e^{-\lambda T}}+c_d\left(\frac{T}{1-e^{-\lambda T}}-\frac{1}{\lambda}\right)$$

单位时间内的期望费用为

<table>
<tr>
<td></td>
<td>

$$C(T)=\frac{EZ}{EY}=c_d+\frac{c_i+(c_f-c_d/\lambda)(1-e^{-\lambda T})}{T}$$

令$C'(T)=0$，得

$$(c_d-\lambda c_f)(1+\lambda T)e^{-\lambda T}=c_d-\lambda c_f-\lambda c_i$$

求解上述方程，便可得到最优的检测时间间隔T。

【学生活动】对上述案例进行讨论，自主练习，并到台上讲解自己的模型，阐述自己对问题的理解

【设计意图】学生到台上讲解，既能加深对模型的印象，也能激发学生的内在的学习动力，提高学习效果。

（五）课程思政（3 分钟）

结合“购买彩票是否中奖”的案例，引导学生树立正确的价值观，摒弃“一夜暴富”的幻想，树立认真工作，脚踏实地，爱岗敬业的价值观。

（六）课堂总结（2 分钟）

总结本次课的教学内容：

（1）：随机变量的理论体系；

（2）：概率统计案例。

</td>
</tr>
<tr>
<td>课后作业</td>
<td>本章的课后题 1。</td>
</tr>
</table>

教学内容	第二节 统计学的基本概念 第三节 参数估计
教学目标	**知识目标：** 理解统计学中的总体与样本的概念；掌握常见的统计量及正态总体的抽样分布；会用 Matlab 软件计算统计量；掌握点估计与区间估计的方法。 **能力目标：**通过实例演示、问题驱动等方式激发学生学习的积极性，通过观察对比、学生交流、师生交流、小组探究等多种形式， 培养学生的逻辑思维能力、分析判断能力、和解决问题能力。 **素质目标：** 培养学生敢于质疑、善于分析、勇于创新的精神；培养学生的自主学习意识和团队协作精神；培养学生脚踏实地、不畏艰辛、锲而不舍的精神。
授课学时	2 学时
教学重点	未知参数的点估计和区间估计。
教学难点	未知参数的区间估计的方法
教学方法	采用问题驱动法、案例演示法、启发式讲授法及自主学习法相结合，以教师的讲解为主，学生的课堂报告、分组讨论为辅，充分调动学生学习的主动性和思考问题的积极性。
教学手段	以课堂讲授为主，主要是多媒体课件和板书相结合的形式。同时借助在线资源，如慕课、雨课堂等平台。
教学过程	**（一）采用讲授法，介绍统计学的相关概念（约 30 分钟）** 知识点 1：把研究对象的全体称为总体，而把组成总体的每个单元称为个体。 从总体 X 中随机抽取 n 个个体，得到的 n 维随机变量 $(X_1, X_2, \cdots, X_n)$ 称为总体 X 的一个样本。样本$(X_1, X_2, \cdots, X_n)$ 中个体的数量 n 称为样本容量。一次抽取的结果是 n 个具体的数据 $(x_1, x_2, \cdots, x_n)$，称其为样本$(X_1, X_2, \cdots, X_n)$ 的一个观测值。

为使抽取的样本能很好地反应总体的特性，抽样一般应满足以下两个条件：

(1) 随机性 对每一次抽样，每个个体都有同样的机会被抽取；

(2) 独立性 每次抽取的结果既不影响其他各次抽取的结果，也不受其他各次抽取结果的影响。

满足以上两个条件的抽样方法称为简单随机抽样，由此得到的样本称为简单随机样本，简称为样本。显然，样本中的各个分量$X_1, X_2, \cdots, X_n$ 相互独立，且都与总体 X 同分布。

知识点2：设 $(X_1, X_2, \cdots, X_n)$ 是总体X 的样本，$g(x_1, x_2, \cdots, x_n)$ 是 n 元连续函数，且其中不含有任何未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 为统计量。

设 $(X_1, X_2, \cdots, X_n)$ 是取自正态总体$X \sim N(\mu, \sigma^2)$ 的容量为n 的样本，则

（1）$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ 或 $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$；

（2）$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$；

（3）$\frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n-1)$。

设 $(X_1, X_2, \ldots, X_{n_1})$ 是取自正态总体$X_1 \sim N(\mu_1, \sigma_1^2)$ 的容量为n_1 的样本，$(Y_1, Y_2, \ldots, Y_{n_2})$ 是取自正态总体$Y \sim N(\mu_2, \sigma_2^2)$ 的容量为n_2 的样本，且 $(X_1, X_2, \ldots, X_{n_1})$ 与 $(Y_1, Y_2, \ldots, Y_{n_2})$ 相互独立。样本均值分别记为 $\bar{X}, \bar{Y}$.样本方差分别记为 S_1^2, S_2^2.则

（1）$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim$

$N(0,1)$ ；	（2） $\frac{1}{S_2^2/\sigma_2^2} \quad {}^{1} F(n_1-1, n_2 \quad -1)$ ；

（3）当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时，$\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$，

其中，$S_w=\sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$ 称为混合样本标准差。

（二）采用案例分析法、讲授法，介绍利用 Matlab 软件计算统计量的方法。（约 15 分钟）

案例 1：学校随机抽取 100 名学生，测量他们的身高和体重， 所得数据如下表所示。分别求身高的均值、中位数、极差、方差、标准差；计算身高与体重的协方差与相关系数。

100 名学生的身高和体重数据

身高	体重	身高	体重	身高	体重	身高	体重	身高	体重
172	75	169	55	169	64	171	65	167	47
171	62	168	67	165	52	169	62	168	65
166	62	168	65	164	59	170	58	165	64
160	55	175	67	173	74	172	64	168	57
155	57	176	64	172	69	169	58	176	57
173	58	168	50	169	52	167	72	170	57
166	55	161	49	173	57	175	76	158	51
170	63	169	63	173	61	164	59	165	62
167	53	171	61	166	70	166	63	172	53
173	60	178	64	163	57	169	54	169	66
178	60	177	66	170	56	167	54	169	58
173	73	170	58	160	65	179	62	172	50
163	47	173	67	165	58	176	63	162	52
165	66	172	59	177	66	182	69	175	75
170	60	170	62	169	63	186	77	174	66
163	50	172	59	176	60	166	76	167	63
172	57	177	58	177	67	169	72	166	50
182	63	176	68	172	56	173	59	174	64
171	59	175	68	165	56	169	65	168	62
177	64	184	70	166	49	171	71	170	59

（三）采用讲授法、启发式教学法，介绍参数估计的理论。（约 40 分钟）

知识点 3：设 $\theta=(\theta_1,\theta_2,\cdots,\theta_m)$ 是总体 X 分布中的未知参数（或数字特征）用样本 $(X_1,\cdots,X_n)$

$X_2, , X_n$) 的一个统计量$\theta = \theta(X_1, X_2,$

来估计 θ ，则称统计量 $\theta=\theta(X_1,X_2,\cdots,X_n)$ 为参数 θ 的估计量。将样本的观测值 $(x_1,x_2,\cdots,x_n)$ 代入估计量 $\theta=\theta(X_1,X_2,\cdots,X_n)$ 得

$\theta=\theta(x_1,x_2,\cdots,x_n)$ ，称其为参数 θ 得一个估计值。

常用的两种估计方法为矩估计法和极大似然估计法。

（1）矩估计法

矩估计的原理是用样本的矩替换总体的相应阶的矩，进而用样本的矩的连续函数来替换总体的矩的同一函数。

设 $\theta=(\theta_1,\theta_2,\cdots,\theta_m)$ 为总体 X 的分布中的未知参数。设总体 X 存在直到 m 阶的原点矩，易见总体的各阶原点矩都是未知参数

$\theta=(\theta_1,\theta_2,\cdots,\theta_m)$ 的函数，即

$$E(X^k)=\alpha_k(\theta_1,\theta_2,\cdots,\theta_m),\ 1\le k\le m .$$

设 $(X_1,X_2,\cdots,X_n)$ 为总体 X 的样本，则以样本的 k 阶原点矩分别代替总体的 k 阶原点矩，$1\le k\le m$ ，得到一个关于未知参数

$\theta=(\theta_1,\theta_2,\cdots,\theta_m)$ 的方程组，即

$$\frac{1}{n}\sum_{i=1}^{n}X_i^k=E(X^k),\ 1\le k\le m$$

求解上述方程组，可得各参数 $\theta_1,\theta_2,\cdots,\theta_m$ 的矩估计量

$\theta_k=\theta_k(X_1,X_2,\cdots,X_n)$ ，$1\le k\le m$. 若已知样本的一组观测值为

$(x_1,x_2,\cdots,x_n)$ ，则可以得到各参数 $\theta_1,\theta_2,\cdots,\theta_m$ 的矩估计值为

$$\theta_k=\theta_k(x_1,x_2,\cdots,x_n),1\le k\le m .$$

（2）极大似然估计法

极大似然估计法的直观想法是：我们选取的未知参数的估计值，应当使我们得到的样本观测值出现的概率在未知参数的所有可能取值中达到最大。

	设 $\theta=(\theta_1,\theta_2,\cdots,\theta_m)$ 为总体 X 的概率密度函数 $f(x;\theta)$ 中的未知参数（当总体 X 服从离散型分布时，$f(x;\theta)$ 为 X 的概率分布列），其中 $\theta\in\Theta$，Θ 为参数空间。$(x_1,x_2,\cdots,x_n)$ 为样本 $(X_1,X_2,\cdots,X_n)$ 的一组观测值。称 … … … … … – … … … … … … … …

	$$L(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n} f(x_i;\theta)$$ 为样本的似然函数。 若存在某个 $\theta=(\theta_1,\theta_2,\ \cdots,\theta_m)$，$\theta\in\Theta$ 使得 $$L(x_1,x_2,\dots,x_n;\theta)=\max_{\theta\in\Theta}L(x_1,x_2,\dots,x_n;\theta),$$ 其中 $\theta_i=\theta_i(x_1,x_2,\cdots,x_n),1\le i\le m$，则称 θ 为参数 θ 的极大似然估计值。进而称 $$\theta=\theta(X_1,X_2,\ \cdots,X_n)$$ $$=(\theta_1(X_1,X_2,\ \cdots,X_n),\theta_2(X_1,X_2,\cdots,X_n),\cdots,\theta_m(X_1,X_2,\cdots,X_n))$$ 为参数θ 的极大似然估计量。 知识点4: 设 θ 为总体X 的分布中的未知参数，$(X_1,X_2,\ \cdots,X_n)$ 是取自总体 X 的样本。对给定的 $\alpha(0<\alpha<1)$，若统计量 $\underline{\theta}=\underline{\theta}(X_1,X_2,\cdots,X_n)$ 和 $\overline{\theta}=\overline{\theta}(X_1,X_2,\cdots,X_n)$ 满足 $$P(\underline{\theta}<\theta<\overline{\theta})=1-\alpha,$$ 则称 $(\underline{\theta},\overline{\theta})$ 为 θ 的置信水平为$1-\alpha$ 的置信区间，$\underline{\theta}$ 称为置信下限，$\overline{\theta}$ 称为置信上限，$1-\alpha$ 称为置信水平 设 $(X_1,X_2,\cdots,X_n)$ 是取自正态总体$X\sim N(\mu,\sigma^2)$ 的样本，其中总体均值 μ 为未知参数。置信水平为$1-\alpha$。 (1) 总体方差σ^2 已知时均值μ 的置信区间 由$\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$ 可知， $$P(-u_{\alpha/2}<\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}<u_{\alpha/2})=1-\alpha,$$ 其中$u_{\alpha/2}$ 为标准正态分布的上侧$\alpha/2$ 分位数。 整理可得 $$P(\overline{X}-u_{\alpha/2}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+u_{\alpha/2}\frac{\sigma}{\sqrt{n}})=1-\alpha.$$ 所以，当方差σ^2 已知时均值μ 的置信水平为$1-\alpha$ 的置信区间

为

$$(\bar{X}-u_{\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{X}+u_{\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

(2) 总体方差 σ^2 未知时均值 μ 的置信区间由 $\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$ 可知，

$$P(-t_{\alpha/2}(n-1)<\frac{\bar{X}-\mu}{S/\sqrt{n}}<t_{\alpha/2}(n-1))=1-\alpha,$$

其中 $t_{\alpha/2}(n-1)$ 为自由度为 $(n-1)$ 的 t 分布的上侧 $\alpha/2$ 分位数。

整理可得

$$P(\bar{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}<\mu<\bar{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}})=1-\alpha.$$

所以，当方差 σ^2 未知时均值 μ 的置信水平为 $1-\alpha$ 的置信区间为

$$(\bar{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\bar{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}).$$

设 $(X_1,X_2,\cdots,X_n)$ 是取自正态总体 $X\sim N(\mu,\sigma^2)$ 的样本，其中总体方差 σ^2 为未知参数。置信水平为 $1-\alpha$。

由 $\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$ 可知，

$$P(\chi^2_{1-\alpha/2}(n-1)<\frac{(n-1)S^2}{\sigma^2}<\chi^2_{\alpha/2}(n-1))=1-\alpha,$$

其中 $\chi^2_{1-\alpha/2}(n-1)$ 与 $\chi^2_{\alpha/2}(n-1)$ 分别为自由度为 $(n-1)$ 的 χ^2 分布的上侧 $1-\alpha/2$ 分位数与上侧 $\alpha/2$ 分位数。

整理可得

$$P(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}<\sigma^2<\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)})=1-\alpha.$$

所以，方差 σ^2 的置信水平为 $1-\alpha$ 的置信区间为

$$(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}).$$

（四）课堂总结（5 分钟）

	总结本次课的教学内容： 1)(：点估计的常用方法； 2)(：区间估计的常用公式。
课后作业	课下自学 Matlab 软件求解点估计和区间估计问题。

教学内容	第四节 假设检验 第五节 数理统计模型—专家打分的可信度评价问题
教学目标	**知识目标：** 掌握常用的参数假设检验的方法；会建立并求解数理统计模型。 **能力目标：**通过实例演示、问题驱动等方式激发学生学习的积极性，通过观察对比、学生交流、师生交流、小组探究等多种形式， 培养学生的逻辑思维能力、分析判断能力、和解决问题能力。 **素质目标：** 培养学生敢于质疑、善于分析、勇于创新的精神；培养学生的自主学习意识和团队协作精神；培养学生脚踏实地、不畏艰辛、锲而不舍的精神。
授课学时	2 学时
教学重点	假设检验方法。
教学难点	假设检验方法。
教学方法	采用问题驱动法、案例演示法、启发式讲授法及自主学习法相结合，以教师的讲解为主，学生的课堂报告、分组讨论为辅，充分调动学生学习的主动性和思考问题的积极性。
教学手段	以课堂讲授为主，主要是多媒体课件和板书相结合的形式。同时借助在线资源，如慕课、雨课堂等平台。
教学过程	**（一）采用讲授法，介绍假设检验的思想与方法（约 25 分钟）** 知识点 1：单个正态总体均值的假设检验 设 $(X_1, X_2, \cdots, X_n)$ 是取自正态总体 $X \sim N(\mu, \sigma^2)$ 的样本。检验问题为 $$H_0: \mu = \mu_0; H_1: \mu \neq \mu_0.$$ (1) 总体方差 σ^2 已知时均值 μ 的假设检验 当原假设 H_0 为真时，检验统计量 $U = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$。对给定的显著性水平 α，有

	$$P(-u_{\alpha/2} < \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < u_{\alpha/2}) = 1-\alpha,$$ 即 $$P(\left\|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right\| \ge u_{\alpha/2}) = \alpha,$$ 其中$u_{\alpha/2}$为标准正态分布的上侧$\alpha/2$ 分位数。 故检验的否定域为 $$W = \{(x_1, x_2, \cdots, x_n) \mid \|u\| \ge u_{\alpha/2}\}.$$ 计算检验统计量U 的观测值u，若给定的样本落入否定域W， 即有$\|u\| \ge u_{\alpha/2}$，则否定原假设 H_0；若有$\|u\| < u_{\alpha/2}$，则接受原假设 H_0 。 (2) 总体方差σ^2未知时均值μ 的假设检验 当原假设H_0 为真时，检验统计量 $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t(n-1)$ 。对给定的显著性水平α，有 $$P(-t_{\alpha/2}(n-1) < \frac{\overline{X}-\mu_0}{S/\sqrt{n}} < t_{\alpha/2}(n-1)) = 1-\alpha,$$ 即 $$P(\left\|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right\| \ge t_{\alpha/2}(n-1)) = \alpha,$$ 其中$t_{\alpha/2}(n-1)$ 为自由度为$(n-1)$ 的t 分布的上侧$\alpha/2$ 分位数。 故检验的否定域为 $$W = \{(x_1, x_2, \cdots, x_n) \mid \|t\| \ge t_{\alpha/2}(n-1)\}.$$ 计算检验统计量T 的观测值t，若给定的样本落入否定域W， 即有 $\|t\| \ge t_{\alpha/2}(n-1)$，则否定原假设 H_0；若有 $\|t\| < t_{\alpha/2}(n-1)$，则接受原假设 H_0 。 知识点 2：单个正态总体方差的假设检验 设 $(X_1, X_2, \cdots, X_n)$ 是取自正态总体$X \sim N(\mu, \sigma^2)$ 的样本。检验问题为 $H_0: \sigma^2 = \sigma_0^2; H_1: \sigma^2 \ne \sigma_0^2$.

	当原假设 H_0 为真时，检验统计量 $\chi^2=\frac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$。对给定的显著性水平 α，有 $$P\{(\chi^2\le\chi^2_{1-\alpha/2}(n-1))\cup(\chi^2\ge\chi^2_{\alpha/2}(n-1))\}=\alpha,$$ 其中 $\chi^2_{\alpha/2}(n-1),\chi^2_{1-\alpha/2}(n-1)$ 分别为自由度为 $(n-1)$ 的 χ^2 分布的上侧 $\alpha/2$ 分位数和上侧 $1-\alpha/2$ 分位数。 故检验的否定域为 $$W=\{\chi^2\le\chi^2_{1-\alpha/2}(n-1)\}\cup\{\chi^2\ge\chi^2_{\alpha/2}(n-1)\}.$$ 计算检验统计量 χ^2 的观测值，若给定的样本落入否定域 W，则否定原假设 H_0；否则接受原假设 H_0。 **（二）采用案例分析法、讲授法，介绍利用 Matlab 软件进行假设检验的方法。（约 20 分钟）** 案例 1：某厂加工的某种铝材的长度 X 服从正态分布 $N(\mu,\sigma^2)$，其均值规定为 $120cm$。现从该厂加工的铝材中随机抽取 5 件，测得其长度分别为（单位：cm）119,120,119.2,119.7,119.6. 试检验该厂加工的铝材长度是否符合规定？（显著性水平为 0.05） 解：由题设，铝材长度 $X\sim N(\mu,\sigma^2)$.检验假设： $$H_0:\mu=120;H_1:\mu\ne120.$$ 由于方差 σ^2 未知，可选择 $T=\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\sim t(n-1)$ 为检验统计量。 检验的否定域为 $W=\{(x_1,x_2,\cdots,x_n)\ \lvert t\rvert\ge t_{0.025}(4)\}$. 查 t 分布表可得 $t_{0.025}(4)=2.7764$ 由样本计算可得，$\bar{x}=119.5, s=0.4$.故检验统计量 T 的观测值为 $t=-2.7951$，落入否定域 W 内，故应否定原假设 H_0，即认为该厂加工的铝材长度不符合规定。 案例 2：设某厂生产的维尼纶产品的纤度 X 服从正态分布

	$N(\mu,\sigma^2)$，方差$\sigma^2=0.048^2$。采用新工艺后，从该厂生产的维尼纶产品中随机抽取5 件，测得纤度分别为$1.32,1.55,1.36,1.40,1.44$．问新工艺对维尼纶产品的纤度的方差有无显著性影响？（显著性水平为0.1） 解：由题设，产品的纤度$X \sim N(\mu,\sigma^2)$．检验假设： $$H_0:\sigma^2=0.048^2;\ H_1:\sigma^2\neq 0.048^2.$$ 选择$\chi_0^2=\frac{(n-1)S^2}{\sigma^2}$为检验统计量。检验的否定域为 $$W=\{\chi^2\le\chi^2_{1-\alpha/2}(n-1)\}\cup\{\chi^2\ge\chi^2_{\alpha/2}(n-1)\}.$$ 查χ^2分布表可得 $\chi^2_{0.05}(4)=9.4877,\chi^2_{0.95}(4)=0.7107$．由样本计算可得，$s^2=0.0078$．故检验统计量的观测值为 $\chi^2=13.5069$，落入否定域W 内，故应否定原假设 H_0，即认为新工艺对维尼纶产品的纤度的方差有显著性影响。 **（三）采用案例分析法、启发式教学法，介绍建模竞赛真题。（约 40 分钟）** 案例 3：确定葡萄酒质量时一般是通过聘请一批有资质的评酒员进行品评。每个评酒员在对葡萄酒进行品尝后对其分类指标打分，然后求和得到其总分，从而确定葡萄酒的质量。现有 27 个葡萄酒样品，由两组（每组 10 名）评酒员独立对各葡萄酒样品进行品评后打分。则两组评酒员的评价结果有无显著性差异？哪一组结果更可信？ **问题分析：** 原问题所给的数据中存在缺失数据，因此在进行数据分析之前应先对缺失数据或异常数据进行处理。在原题所给数据中，第一组红葡萄酒第 20 号样本中第 4 个评酒员缺失色调打分数据，可以利用其他评酒员的同一项打分数据的均值代替。 葡萄酒的质量取决于好的酿酒葡萄，而好的酿酒葡萄受多种因素影响，如品种、土壤、光照、湿度、降水、工艺等。对每个评酒员而言，其评价结果不应受个人偏好的影响，即其打分结果

的均值应尽可能接近总平均值，同时还应能够区分酒的品质的好坏，这就要求其打分结果应有差异性，即方差越大越好。

模型假设 ：

（1） 每个评酒员都独立打分；

（2） 葡萄酒的抽样是随机的；

（3） 葡萄酒的总体质量服从正态分布。

符号说明：

M：样本容量；

N：评酒员人数；

x_{ij}： 第一组的第 i 个评酒员对第 j 个样本的打分数据，

$i=1,2,\ \dots,N,\ j=1,2,\dots,M$；

y_{ij}： 第二组的第 i 个评酒员对第 j 个样本的打分数据，

$i=1,2,\ \dots,N,\ j=1,2,\dots,M$.

模型建立：由统计学的理论，可建立如下的描述性统计指标：

$$\bar{x}_i=\frac{1}{M}\sum_{j=1}^{M}x_{ij},\ \bar{y}_i=\frac{1}{M}\sum_{j=1}^{M}y_{ij},$$

$$s_{1i}^2=\frac{1}{M-1}\sum_{j=1}^{M}(x_{ij}-\bar{x}_i)^2,\ s_{2i}^2=\frac{1}{M-1}\sum_{j=1}^{M}(y_{ij}-\bar{y}_i)^2,$$

$$R_{1i}=\max_{1\le j\le M}x_{ij}-\min_{1\le j\le M}x_{ij},\ R_{2i}=\max_{1\le j\le M}y_{ij}-\min_{1\le j\le M}y_{ij},$$

$$\mu_1=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}x_{ij},\ \mu_2=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij},$$

$$s_1^2=\frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(x_{ij}-\mu_1)^2,\ s_2^2=\frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu_2)^2,$$

$$R_1=\max_{1\le i\le N,1\le j\le M}x_{ij}-\min_{1\le i\le N,1\le j\le M}x_{ij},\ R_2=\max_{1\le i\le N,1\le j\le M}y_{ij}-\min_{1\le i\le N,1\le j\le M}y_{ij}.$$

其中， $\bar{x}_i,\bar{y}_i$ 分别表示第一组、第二组第 i 个评酒员的打分均值；

s_{1i}^2,s_{2i}^2 分别表示第一组、第二组第 i 个评酒员的打分方差；

R_{1i},R_{2i} 分别表示第一组、第二组第 i 个评酒员的打分极差；，μ_1,μ_2 分别表示第一组、第二组所有评酒员的打分均值；s_1^2,s_2^2 分别表示第一组、第二组所有评酒员的打分方差； R_1,R_2 分别表示第一组、第二组所有评酒员的打分极差。

	（1） 两组评酒员的打分结果是否有显著性差异，等价于检验两组样本均值是否相等，即等价于检验假设 $$H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2.$$ 若把两组样本视为取自两个相互独立的正态总体，考虑到两组样本的容量相同，可以采用基于成对数据的t 检验方法。 （2） 鉴于方差的大小体现了打分结果的区分度，即每组样本打分结果的方差越小，则样本的区分度越差，即无法区分酒的品质的差异度，因此可信的组别应为方差最大的一组。模型描述为$\max\{s_1^2, s_2^2\}$。 **模型求解**：利用 Matlab 软件，可以得到上述描述统计量的计算结果，从计算结果来看，两组评酒员的打分均值略微有差异。为验证两组结果是否有本质上的差异，将每组评酒员对每个样本的打分值取平均，得到每组样本的质量得分。作出假设： $$H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2.$$ 在假定两个总体方差未知且相等的情形下，利用两个总体均值（成对样本）的 t 检验方法，可以得到： $p = 6.4271\times10^{-4}$，即拒绝原假设，认为两组结果有差异。 从方差统计结果来看，第一组评酒员的总体方差为104.9716，远大于第二组评酒员的方差，因此第一组评酒员更可信。 **（四）课堂总结（5 分钟）** 总结本次课的教学内容： （1）：假设检验的常用方法； （2）：利用 Matlab 软件进行假设检验的方法。
课后作业	利用 Matlab 软件求解本章课后题 2、3。