lab2实验报告

1. 实验目的

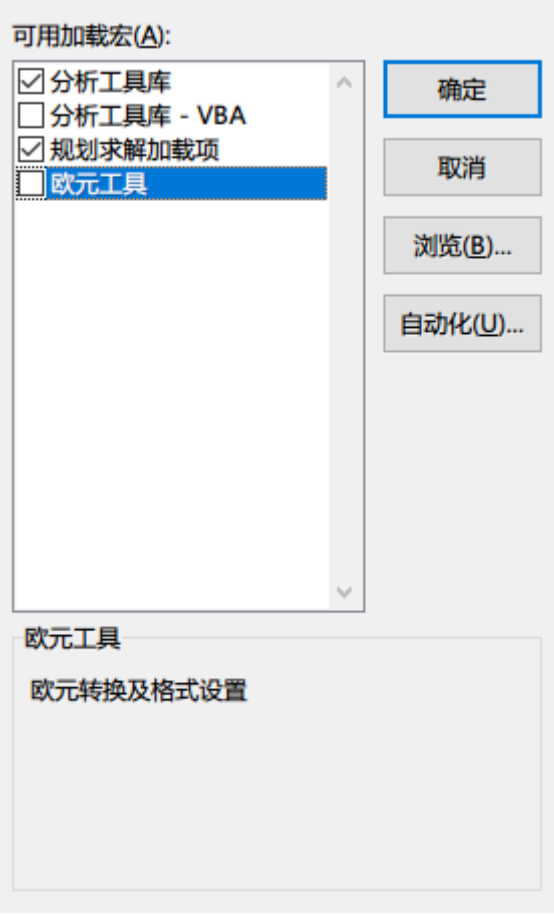
- 学习数据探索分析方法

2. 实验环境

- win10 64位
- excel 2016 64位

3. 实验步骤

- 在excel中的`选项-加载项`中选中分析工具库,点击下方的`转到`,勾选`分析工具库`和`规划求解加载项`.



打开数据分析功能

- 打开`马尾松腮扁叶蜂在林间表土层的水平分布调查数据.xlsx`,按照实验报告的要求,分别使用`AVERAGE`,`VAR`,`STDEVA`,`MEDIAN`,`MAX`,`MIN`,`MODE`,`SKEW`,`KURT`等函数分别求出数据集的平均值,方差,中位数,最大值,最小值,众数,偏度,峰度等,并使用`STDEVA/AVERAGE`,`MAX-MIN`以及`STDEVA/SQRT(COUNT)`求出变异系数,极差以及平均数的标准误.

- 接着再使用`数据-数据分析`中的`描述统计`来对整个数据集进行描述性统计,计算得出的结果与使用函数进行计算时是一样的.

	A	B	C	D	E	F
41		8	1	10		
42		7	4	2		
43		3	4	5		
44		3	4	2		
45		2	9	5		
46		2	5	4		
47		8	6	5		
48		4	1	1		
49		7	6	4		
50	平均数	3.422222	2.866667	3.422222		
51	方差	6.249495	6.027273	4.158586		
52	标准差	2.499899	2.45505	2.039261		
53	中位数	3	3	3		
54	最大值	9	11	10		
55	最小值	0	0	0		
56	极差	9	11	10		
57	众数	2	0	3		
58	变异系数	0.73049	0.856413	0.595888		
59	偏度	0.46175	1.019359	0.807282		
60	峰度	-0.55659	1.68012	1.573633		
61	标准误差	0.372663	0.365977	0.303995		
62	样地1		样地2		样地3	
63						
64	平均	3.422222	平均	2.866667	平均	3.422222
65	标准误差	0.372663	标准误差	0.365977	标准误差	0.303995
66	中位数	3	中位数	3	中位数	3
67	众数	2	众数	0	众数	3
68	标准差	2.499899	标准差	2.45505	标准差	2.039261
69	方差	6.249495	方差	6.027273	方差	4.158586
70	峰度	-0.55659	峰度	1.68012	峰度	1.573633
71	偏度	0.46175	偏度	1.019359	偏度	0.807282
72	区域	9	区域	11	区域	10
73	最小值	0	最小值	0	最小值	0
74	最大值	9	最大值	11	最大值	10
75	求和	154	求和	129	求和	154
76	观测数	45	观测数	45	观测数	45
77						
78						
79						
80						
81						
82						

DATA 1

描述性统计

- 接下来打开`新农药防治柑桔红蜘蛛.xlsx`进行单样本t检验.单样本T检验,主要用于检验单个变量的均值与指定的检验值之间是否存在显著性差异,再者,样本均值与总体均值之间的差异显著性检验,也属于单样本T检验.在进行单样本t检验时需要先给定总体均值.

- 根据单样本t检验的公式,我先使用`AVERAGE`函数计算出样本的平均值,使用`STDEVA`函数计算出标准差,使用`STEDVA/SQRT(COUNT)`公式计算出平均数的标准误,最后使用`ABS(平均值-给定值)/标准误`的公式计算出统计量t.接着使用`TINV(0.05, COUNT - 1)`以及`TINV(0.01, COUNT - 1)`分别计算出置信度在95%以及99%时的统计量t值,将其与统计量t进行比较.

	A	B
1	试验区	防治效果
2	1	95%
3	2	92%
4	3	93%
5	4	99%
6	5	89%
7	6	78%
8	7	92%
9	8	78%
10	9	93%
11		
12		
13		
14	平均数	90%
15	标准差	0.0725335
16	标准误差	2%
17	统计量t	5%
18	t(0.05, n-1)	2.3060041
19	t(0.01, n-1)	3.3553873

单样本t检验

- 接着打开`暴雨前、后棉田一代棉铃虫百株卵量调查.xlsx`进行配对样本的t检验.配对样本T检验用来检验来自两配对总体的均值是否在统计上有显著差异.格局实验报告的流程,先使用`MAX-MIN`函数计算暴雨前后虫卵数量的差值.接着使用`STDEVA`,`STDEVA/ SQRT(COUNT)`,`ABS(AVERAGE-0)/标准误`分别计算差值的标准差,变异系数以及统计量t.使用上个实验的方法计算出置信度在95%以及99%时的统计量t值,进行t值的比较并分析两类样本的均值是否存在显著的差异.
- 打开`数据-数据分析`,选择`t检验-平均值的成对双样本分析`,导入两类样本进行分析.
- 接着进行95%置信区间的估计.统计的目的是要对总体分参数进行估计.因此有必要在一定的概保证下,估计出一个范围或者区间能够覆盖总体参数.我们成这个区间为置信区间.根据实验报告给出的公式,95%置信区间的上限为`样本平均值+置信度在95%的t值 * 平均值的标准误`,而相对的,95%的置信区间的下限为`样本平均值-置信度在95%的t值 * 平均值的标准误`.最终结果如下图.

	A	B	C	D
4	样本编号	暴雨前	暴雨后	差值
5	1	110	90	20
6	2	115	116	-1
7	3	133	101	32
8	4	133	131	2
9	5	126	110	16
10	6	108	88	20
11	7	110	92	18
12	8	110	104	6
13	9	140	126	14
14	10	104	86	18
15	11	160	114	46
16	12	120	88	32
17	13	120	112	8
18			平均数	17.76923
19			标准差	13.11585
20			标准误	3.637684
21			统计量t	4.884765
22			t(0.05, n-1)	2.178813
23			t(0.01, n-1)	3.05454
24			95%置信下限	9.843399
25			95%置信上限	25.69506
26				
27				
28		t-检验: 成对双样本均值分析		
29				
30			暴雨前	暴雨后
31		平均	122.2307692	104.4615
32		方差	252.8589744	228.2692
33		观测值	13	13
34		泊松相关系数	0.643294352	
35		假设平均差	0	
36		df	12	
37		t Stat	4.884765406	
38		P(T<=t) 单尾	0.000187742	
39		t 单尾临界	1.782287556	
40		P(T<=t) 双尾	0.000375484	
41		t 双尾临界	2.17881283	
42				
43				
44				
45				

配对样本t检验

- 接着打开 `不同水稻品种对稻纵卷叶螟幼虫数.xlsx` ,并对实验数据进行单因素方差分析.单因素方差分析是用于两个以及两个以上样本的平均值差别的检验.在研究中,由于研究样本同时受研究因素以及不可控制因素的影响,其数值会出现上下波动.单因素方差分析旨在排除不可控制因素的导致的误差,而检验研究因素是否会对样本带来明显的影响.
- 打开 `数据-数据分析` 中的 `单因素方差分析` ,将实验样本导入到分析中.分析结果如下图.

	A	B	C	D	E	F	G
1	不同水稻品种对稻纵卷叶螟幼虫数						
2		水稻品种					
3	从复	1	2	3	4	5	
4	1	41	33	38	37	31	
5	2	39	37	35	39	34	
6	3	40	35	35	38	34	
7							
8							
9							
10	方差分析：单因素方差分析						
11							
12	SUMMARY						
13	组	观测数	求和	平均	方差		
14	1	3	120	40	1		
15	2	3	105	35	4		
16	3	3	108	36	3		
17	4	3	114	38	1		
18	5	3	99	33	3		
19							
20							
21	方差分析						
22	差异源	SS	df	MS	F	P-value	F crit
23	组间	87.6	4	21.9	9.125	0.002265	3.47805
24	组内	24	10	2.4			
25							
26	总计	111.6	14				

单因素方差分析

- 由图中可以看出,组间离均差平方和与均方差要远远大于组内的离均差平方和以及均方差.P值为0.0022也远远小于极显著的标准0.01.因此可以得出结论不同水稻之间的抗虫性存在着显著的差异.
- 打开 `一代粘虫幼虫发生量与环境因素的关系.xlsx` ,使用 `数据-数据分析` 中的 `方差分析-可重复双因素分析` 对数据集进行分析.结果如下图.

	A	B	C	D	E	F	G
16							
17							
18	方差分析：可重复双因素分析						
19							
20	SUMMARY	27	29	31	总计		
21	91.2						
22	观测数	4	4	4	12		
23	求和	335.9	292.5	255.4	883.8		
24	平均	83.975	73.125	63.85	73.65		
25	方差	7.949167	29.1425	0.83	84.13182		
26							
27	93.2						
28	观测数	4	4	4	12		
29	求和	332.8	296.3	289	918.1		
30	平均	83.2	74.075	72.25	76.50833		
31	方差	5.2	32.64917	96.03667	61.54447		
32							
33	100.2						
34	观测数	4	4	4	12		
35	求和	369	324.4	295.7	989.1		
36	平均	92.25	81.1	73.925	82.425		
37	方差	2.696667	13.20667	2.929167	67.14932		
38							
39	总计						
40	观测数	12	12	12			
41	求和	1037.7	913.2	840.1			
42	平均	86.475	76.1	70.00833			
43	方差	22.62205	34.25455	48.41356			
44							
45							
46	方差分析						
47	差异源	SS	df	MS	F	P-value	F crit
48	样本	480.7106	2	240.3553	11.34703	0.000265	3.354131
49	列	1663.601	2	831.8003	39.26879	1.02E-08	3.354131
50	交互	105.5611	4	26.39028	1.245869	0.31533	2.727765
51	内部	571.92	27	21.18222			
52							
53	总计	2821.792	35				

双因素方差分析

- 接着进行回归分析,分析结果如下图.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3		回归统计							
4	Multiple	0.893857392							
5	R Square	0.798981038							
6	Adjusted	0.725883234							
7	标准误差	0.619486459							
8	观测值	16							
9									
10	方差分析								
11		df	SS	MS	F	Significance F			
12	回归分析	4	16.7786	4.19465	10.9303	0.000793865			
13	残差	11	4.221398	0.383763					
14	总计	15	21						
15									
16		Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
17	Intercept	-0.182114072	0.441578	-0.41242	0.687958	-1.154021275	0.789793	-1.15402	0.789793
18	蛾量	0.142195305	0.158014	0.899892	0.387454	-0.205590557	0.489981	-0.20559	0.489981
19	卵量	0.244517176	0.213494	1.145311	0.276387	-0.22538011	0.714414	-0.22538	0.714414
20	降水量	0.210092567	0.224339	0.936495	0.369124	-0.283674777	0.70386	-0.28367	0.70386
21	雨日	0.604710831	0.244545	2.4728	0.030967	0.066471021	1.142951	0.066471	1.142951
22									
23									
24									
25	RESIDUAL OUTPUT								
26									
27	观测值	预测 幼虫密度	残差						
28	1	1.445987724	-0.44599						
29	2	1.50843616	-0.50844						
30	3	1.303792419	-0.30379						
31	4	4.623949447	-0.62395						
32	5	1.019401808	-0.0194						
33	6	1.161597113	-0.1616						
34	7	2.607630169	0.39237						
35	8	1.263918984	-0.26392						
36	9	3.272332478	0.727668						
37	10	3.062239912	-0.06224						
38	11	2.220917687	-0.22092						
39	12	2.743424691	-0.74342						
40	13	3.103618346	-0.10362						
41	14	2.371018776	0.628981						
42	15	3.272332478	0.727668						
43	16	1.019401808	0.980598						
44									
45									
46									
47									

回归分析

4. 实验总结

在本次实验中,我使用excel分别进行了描述性统计,单样本t检验,配对样本t检验,单因素方差分析,多因素方差分析以及回归分析.在进行这些实验的过程中,由于我本身对于统计学知识的不足,在理解分析结果的时候走了很大的弯路.所以继续巩固统计学的相关基础知识应该是以后的重点.

5. 操作习题

解释95%的置信区间

一个概率样本的置信区间（Confidence interval）是对这个样本的某个总体参数的区间估计。置信区间展现的是这个参数的真实值有一定概率落在测量结果的周围的程度。置信区间给出的是被测量参数的测量值的可信程度，即前面所要求的“一定概率”。这个概率被称为置信水平。置信区间是总体参数所在的可能范围,95%置信区间就是总体参数在这个范围的可能性大概是95%,或者说总体参数在这个范围,但其可信程度只有95%.以上述实验中的 回归分析 实验为例子,各个变量的权重的真实值由于样本容量的限制无法得知,但是根据变量权重的95%置信区间可以知道变量权重真实值的大概范围.

如何确定检验的显著性a?什么是假设检验的p值?如何根据p值来做出假设检验的结论?

显著性检验是用于检测科学实验中实验组与对照组之间是否有差异以及差异是否显著的办法.显著性检验要求在进行检验前先对数据进行一个假设,然后用检验来判断这个假设是否正确.一般而言，把要检验的假设称之为原假设，记为H0；把与H0相对应（相反）的假设称之为备择假设，记为H1.如果原假设实际上为真,而检验的结论却指出我们应该放弃原假设.此时我们把这类错误出现的概率记为a.概率a称为显著性水平.由于统计学中一般把在现实世界中发生概率小于5%的时事件认为是不可能事件,因此概率a的值常为0.05,如果要使得检验结果达到极为显著的水平,则可以将a值设置为0.01.

与a值相似,p值是一种在原假设为真的前提下出现观察样本以及更极端情况的概率.p值用于表示对原假设的支持程度,是用于确定是否应该拒绝原假设的一种方法.p值常常用于与给定的a值相比较.如果p值>a值,则在显著性水平a下接受原假设,如果p值<a值,则在显著性水平a下拒绝原假设.

区间估计与假设检验有何联系与区别?如何根据置信区间进行假设检验?

区间估计是参数估计的一种,参数估计指的是利用样本中的数据估计总体分布的某个或者某几个参数.点估计是另一种参数估计,它采用估计量的某个取值直接作为总体参数的估计值,点估计的缺点在于无法给出估计的可靠性,也不能指出估计值与参数真实值的接近程度.而区间估计在点估计的基础上给出总体参数估计的一个估计区间,在区间估计中,由样本估计量构造的总体参数在一定置信水平下的估计区间称为置信区间.

而假设检验正如上所示,是给定置信水平，根据假设值确定估计值可能出现的区间范围，该区间通常以假设值为中心.这里的假设值是指真实值的假设值.若计算得到的估计值在此区间范围内，则接受原假设，否则拒绝原假设.

他们之间的联系在于:都是根据样本信息推断总体参数;都以抽样分布作为理论依据,都具有一定的可信程度;而这可以相互转换.

他们之间的区别在于:参数估计是以样本资料估计总体参数的真值,假设检验是以样本资料检验对总体参数的先前假设是否成立;区间估计求得的是以样本估计值为中心的双侧置信区间,而假设检验既有双侧检验,也有单侧检验;区间估计立足于大概率,假设检验立足于小概率.