南京大學

基于 SPSS 的病案数据分析

课 程： 管理统计学

教 师： 朱惠

课 题： 病案分析

院 系： 信息管理学院

专 业： 信息管理与信息系统

姓 名： 张晓宇

目录

1. 研究意义

在社会经济发展水平不断提升的今天，人们的生活质量和健康水平有了很大的改善和提高。但在医疗领域，医疗资源分配不均衡、医疗水平良莠不齐、医疗费用持续上涨等问题日益突出。本文基于病案分析的数据集，利用SPSS分析方法和工具，结合管理统计学的课程知识，分析住院费用的影响因素及各属性之间的相关关系。

2. 调查样本综合分析

调查样本综合分析

		計數	百分比 %
性别	男	1661	61.5%
	女	1039	38.5%
年龄	20岁以下	374	13.9%
	20-40岁	379	14.0%
	40-60岁	664	24.6%
	60-70岁	595	22.0%
	70 岁以上	688	25.5%
婚姻状况	未婚	438	16.3%
	已婚	2221	82.7%
	独身	26	1.0%
医院级别	省级	900	33.3%
	地级	900	33.3%
	县级	900	33.3%
医院代码	070	207	7.7%
	087	300	11.1%
	100	93	3.4%
	104	69	2.6%
	152	134	5.0%
	205	207	7.7%
	225	85	3.1%
	234	30	1.1%
	290	294	10.9%
	400	300	11.1%
	406	81	3.0%
	438	204	7.6%

	450	300	11.1%
	511	96	3.6%
	518	300	11.1%
	总计	2700	100.0%
地区	东部	900	33.3%
	中部	900	33.3%
	西部	900	33.3%
病情种类	1629	43	1.6%
	4018	275	10.2%
	4140	596	22.1%
	4349	318	11.8%
	4556	127	4.7%
	4659	391	14.5%
	4919	361	13.4%
	5409	322	11.9%
	7221	175	6.5%
	8738	92	3.4%
治疗类别	中医	438	16.2%
	西医	866	32.1%
	中西医	1396	51.7%
费用来源	公费	616	22.8%
	劳保	42	1.6%
	保险	229	8.5%
	自费	1752	64.9%
	统筹	5	.2%
	其他	56	2.1%
住院费用	1000以下	798	29.6%
	1000-3000	956	35.4%
	3000-5000	363	13.4%
	5000-10000	328	12.1%
	10000及以上	255	9.4%

由调查样本综合分析的图表可以看出，本次调查样本分布较广，共2700例病案。病人案例中男性较多。年龄范围中，60岁以上病人比例明显最多，随着年龄的增长，患病几率也逐渐上升。样本中已婚人数超过七成。涉及15家医院，同时按照省级、地级、县级医院，以及东部、中部、西部地区均为三分之一，样本数目持平且随机性较强。共选取10种疾病类型进行分析，其中4140病种病人数量较多。在治疗类别上，中医、西医、中西医均有涉及，且中西医结合占半数。在住

院费用上，大多在3000元以下。

3. 住院费用构成分析

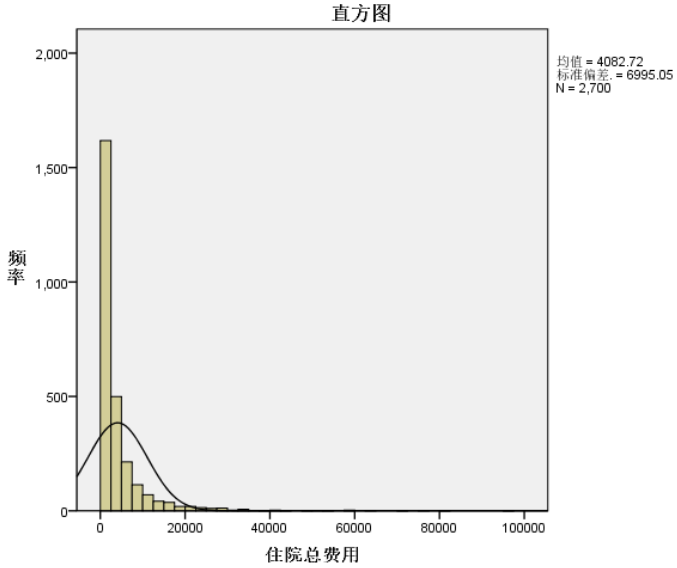
3.1 住院费用描述统计量

通过分析—描述统计—描述，对住院费用到治疗费等连续变量进行描述分析，得到下表。并对住院费用的分布情况进行描绘，得到住院总费用直方图。之后对住院总费用生成箱图。

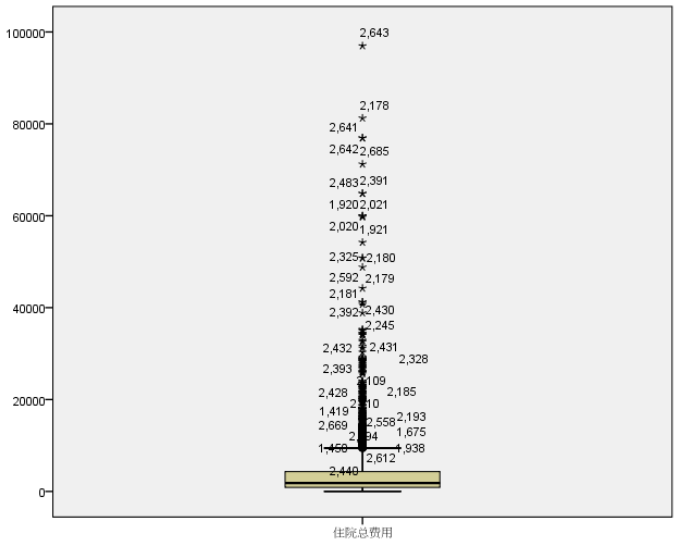
描述性統計資料

	N	最小值	最大值	平均數	標準偏差
住院总费用	2700	0	96965	4082.72	6995.050
床位费	2700	0	4400	234.33	346.383
中成药费	2700	0	18614	335.04	1002.231
西药费	2700	0	66891	1771.42	3690.751
检查费	2700	0	2981	120.08	223.033
血费	2700	0	1320	5.79	69.834
氧费	2700	0	2608	28.40	142.334
诊疗费	2700	0	28977	992.36	2565.113
手术费	2700	0	2758	62.57	231.823
接生费	2700	0	114	.04	2.194
其他费	2700	0	28978	160.13	723.545
放射费	2700	0	1810	35.84	93.823
化验费	2700	0	1958	119.88	147.038
中草药费	2700	0	7103	91.25	311.918
护理费	2700	0	5895	106.07	272.233
治疗费	2700	0	6223	19.73	135.242
有效的 N (listwise)	2700				

住院总费分布直方图



住院总费用箱图

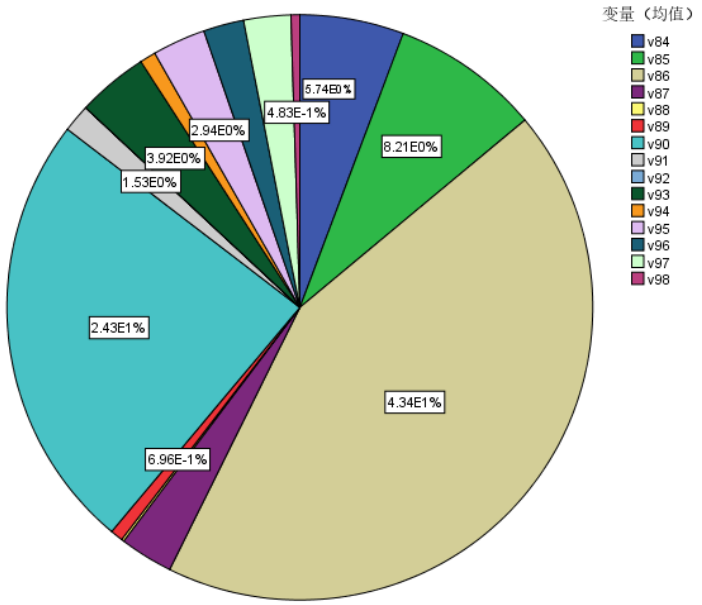


观察发现，住院总费用箱图中有大量的极端值，即各费用差异较大，且均值小于极大值，呈右偏分布，此外，我们还可以发现部分案例住院总费用为0，少量案例住院总费用达到近10万，这些都表明住院总费用的分布状况不满足一定的规律性。

3.2 住院费用占比

根据住院费用构成的描述统计量一表，构造住院费用均值饼图，以得到较为直观的费用占比情况。

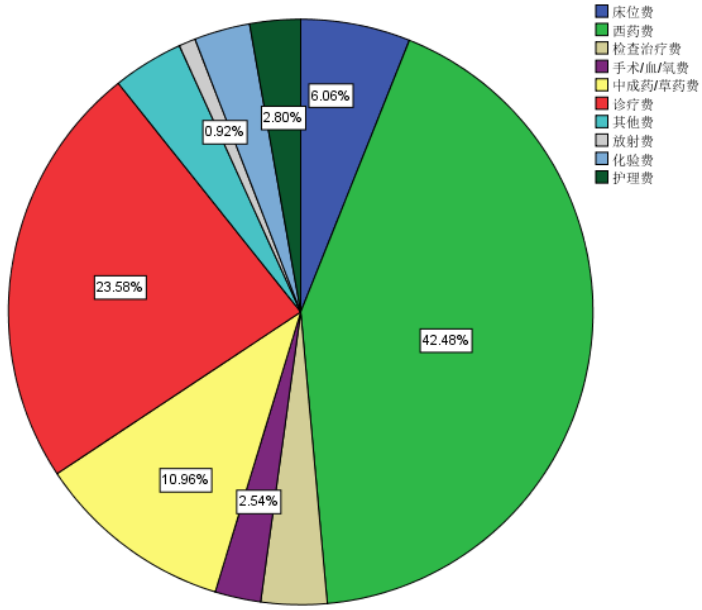
构成费用均值饼图



从饼图中可看出西药费(V86)占比最多，在诊疗费(V90)方面，中西医结合的方式比西医和中医的占比普遍偏高。可得出结论，西药费具有较高的比重，中成药费西医占比最少。西医的手术费有5%，中医、中西医并没有多少手术费。

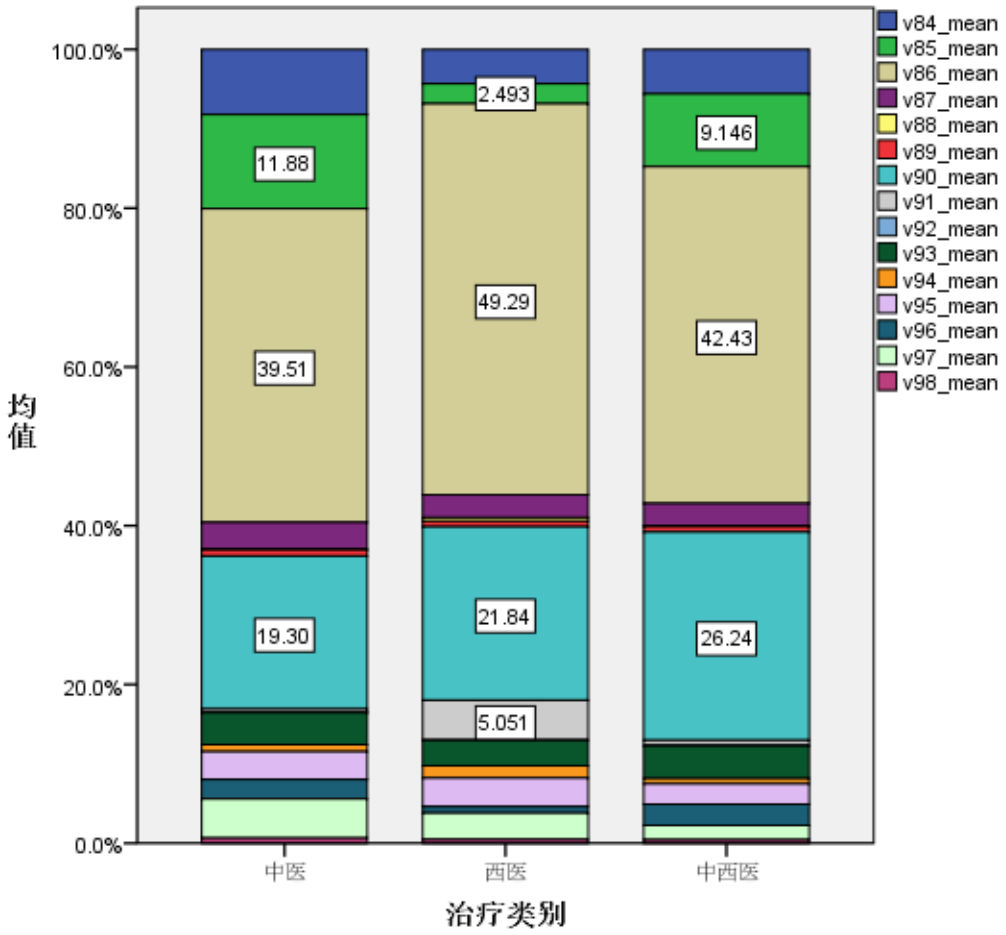
另外，由于一些费用占比较小或相关性较大，对其进行了合并。如手术费、学费、氧费合并为手术/血/氧费(V88_89_91)、检查费和治疗费合并为检查治疗费(V87_98)、中成药费和中草药费都与中药有关，且占比也都较小，所以统称为中成药/草药费(V85_96)。

合并后各类费用占比



此外，对中医、西医和中西医这三种类型的住院费用占比情况进行对比，在分类汇总之后输出堆积条图如下。

不同治疗类别住院费用构成（均值）堆积条图

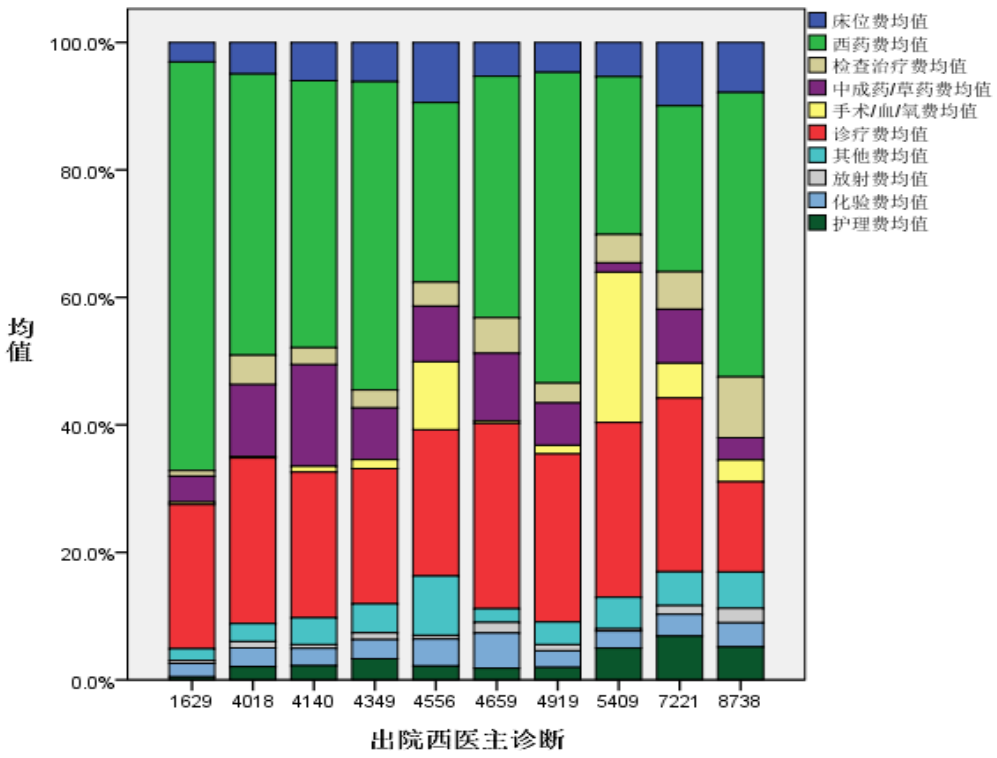


可以发现，西药费(V86)、诊疗费(V90)、中成药费(V85)、床位费(V84)、其他费分别在住院费用构成中居于前五位。

3.3 不同病情类型病人住院费用构成

对不同病情类型的病人住院费用种类进行分类汇总，在分类汇总之后利用堆积条图并缩放至100%，使堆积条图的纵坐标表示为百分比。

十病种人均住院费用构成



我们可以看出，不同病情类型的病人住院费用种类也各不相同，西药费、诊疗费占较大比例，基本占据各病情类型费用支出的前两位。且不同病情类型中，西药费、诊疗费变化幅度也有较大的不同。其中，1629、4349、4919、8738 病情类型中西药费开销较高，5409 病情类型对手术/血/氧费开销明显较高。

4. 住院费用影响因素分析

通过 K-S 检验判断住院总费用是否服从正态分布。H0 为住院总费用服从正态分布，得到下表，观察可知渐近显著性(双侧)小于 0.05，认为是小概率事件，所以拒绝 H0，即住院总费用不服从正态分布。

单样本 Kolmogorov-Smirnov 检验

		住院总费用
N		2660
正态参数[a,b]	均值	3817.65
	标准差	5420.567
最极端差别	绝对值	.243
	正	.218
	负	-.243
Kolmogorov-Smirnov Z		12.551
渐近显著性(双侧)		.000

a. 检验分布为正态分布。

单样本 Kolmogorov-Smirnov 检验

		住院总费用
N		2660
正态参数[a,b]	均值	3817.65
	标准差	5420.567
最极端差别	绝对值	.243
	正	.218
	负	-.243
Kolmogorov-Smirnov Z		12.551
渐近显著性(双侧)		.000

a. 检验分布为正态分布。

b. 根据数据计算得到。

因此，住院总费用不服从正态分布，不满足方差分析要求，故在单因素分析部分选用秩和检验。在此篇报告中，我将从性别、年龄、医院类型、疾病种类和治疗效果四个角度分析这些因素对于住院总费用的影响。

4.1 不同性别对住院费用的影响分析

首先输出不同性别住院费用分析表如下：

同性别住院总费用分析

			均值	表 N %
性别	男	住院总费用	4747	61.5%
	女	住院总费用	3021	38.5%

对不同性别与住院总费用的关联进行 t 检验。设 H0：假设两样本来自同一总体，得到如下结果。

独立样本检验

	方差方程的 Levene 检验		均值方程的 t 检验						
	F	Sig.	t	df	Sig.(双侧)	均值差值	标准误差值	差分的 95% 置信区间	
								下限	上限
假设方差相等	67.877	.000	6.283	2698	.000	1726.207	274.730	1187.505	2264.910
假设方差不相等			6.978	2696.996	.000	1726.207	247.366	1241.162	2211.253

观察可知，渐近显著性(双侧)小于 005，认为是小概率事件，拒绝两者无关的假设，认为不同性别对住院总费用产生了影响。

也可以通过 Mann-Whitney U 检验对不同性别住院费用进行非参数检验，得到统计量 Z=-6.234，P=0.000<0.05，同样可以说明男女病例的住院费用差别具有统计学意义。

Mann-Whitney U 检验

秩

	性别	N	秩均值	秩和
住院总费用	男	1631	1404.24	2290311.50
	女	1029	1213.62	1248818.50
	总数	2660		

检验统计量[a]

	住院总费用
Mann-Whitney U	718883.500
Wilcoxon W	1248818.500
Z	-6.234
渐近显著性(双侧)	.000

a. 分组变量: 性别

4.2 年龄对住院费用的影响分析

年龄和住院费用均为离散型变量，可对其等距分组可视离散化进行分析，并输出不同年龄段住院费用分析表如下：

不同年龄住院总费用分析

			均值	表 N %
年龄分组	20岁以下	住院总费用	692	13.9%
	21-40岁	住院总费用	1853	14.0%
	41-60岁	住院总费用	2955	24.6%
	61-80岁	住院总费用	6174	43.3%
	81岁以上	住院总费用	7755	4.2%

从表中可以得出，不同年龄段的病人住院总费用差异较大。其中，40岁以下病人占比较小，且住院费用均值最低。61-80岁病人占近半的比例，费用均值也较高，为6174。而81岁以上的病人虽然仅占比3.9%，但这部分年龄段病人的住院费用均值却高达7755，可见高龄老人住院费用开销往往更大。

选用秩相关系数，Spearman 等级相关对年龄和住院费用进行相关分析，分析结果如下：

Spearman 的 rho	年龄	相关系数	1.000	.480**
		Sig.（双侧）	.	.000
		N	2660	2660
	住院总费用	相关系数	.480**	1.000
		Sig.（双侧）	.000	.
		N	2660	2660

**. 在置信度（双测）为 0.01 时，相关性是显著的。

可以看出，显著性水平为0.000，具有统计学意义。且它们的相关系数为0.48，表明病人的年龄与住院费用呈正相关关系，即患者年龄越大，住院总费用越大。

4.3 不同医院类型对住院费用的影响分析

医院的类型主要分为级别和地域，可以对医院级别与地域分别进行探究，首先输出不同医院类型与住院总费用的交叉列表如下：

不同医院类型与住院总费用的交叉列表

				住院总费用	
				计数	均值
地区	东部	级别	省级	300	5532
			地级	300	2242
			县级	300	1575
	中部	级别	省级	300	4990
			地级	300	4303
			县级	300	1380
	西部	级别	省级	300	13747
			地级	300	2129
			县级	300	845

4.3.1 不同医院地区对住院费用的影响分析

医院地区对住院费用的影响

		住院总费用	
		均值	层总和 %
地区	东部	3117	27.6%
	中部	3668	31.5%
	西部	4676	40.8%

医院地区对住院费用的影响占比从高到低依次是西部、中部、东部。东部经

济更加发达，人口众多，医疗水平也更高，医疗设施较为完善，住院总费用也最低的。而西部地区的医院住院总费用均值最高，一定程度上说明了西部地区医疗设施不完善，反映了我国地区之间医疗设施和资源分配不平衡，也在提醒我国应加大对中西部医疗资源的投入，以减少医疗费用的不平衡，进而减轻人们的医疗经济负担。

4.3.2 不同医院级别对住院费用的影响分析

医院级别对住院费用的影响

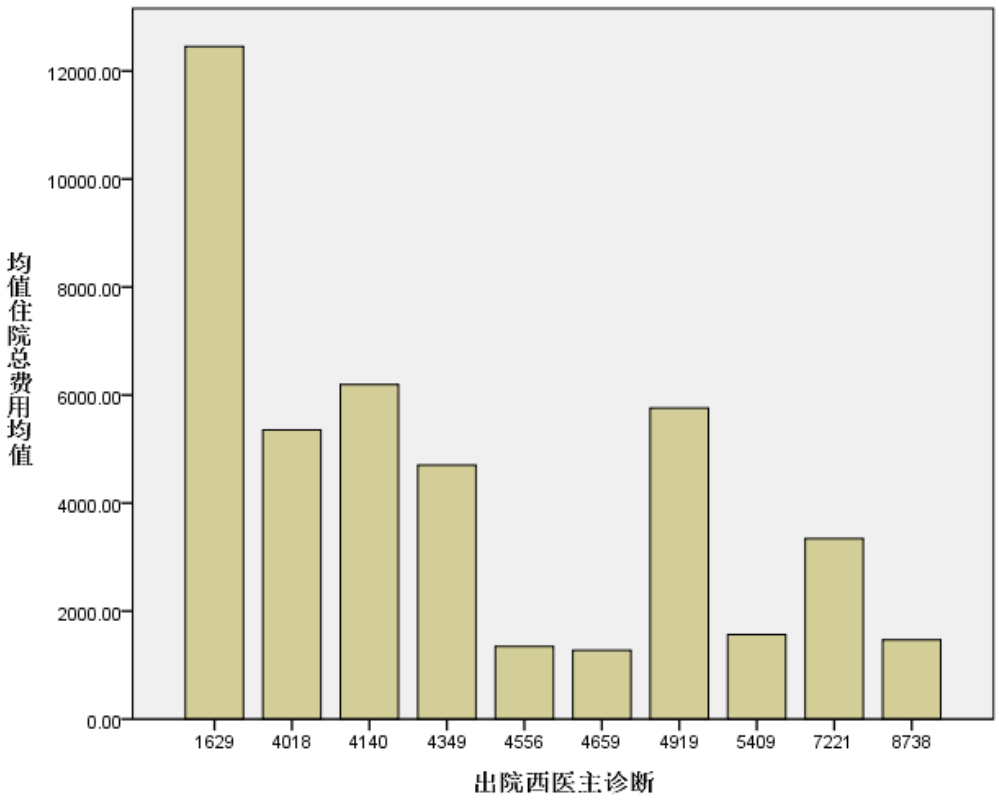
		住院总费用	
		均值	层总和 %
级别	省级	7229	63.1%
	地级	2981	25.6%
	县级	1267	11.2%

医院级别是住院费用重要的影响因素，省级医院对住院费用的影响占比高达63%，远远高于地级医院与县级医院。由此可见，省级医院收费较高，地级县级医院收费水平相对较低，这一情况应同样与经济发展水平、医疗设施、医疗资源均衡程度有关。

4.4 不同病情类型对住院费用的影响分析

不同病情类型的病人住院费用种类分布特征不一样，不同的病情类型之间也会存在着住院费用的差异，生成不同病情类型住院费用均值的条形图，如图：

不同病情类型的住院费用

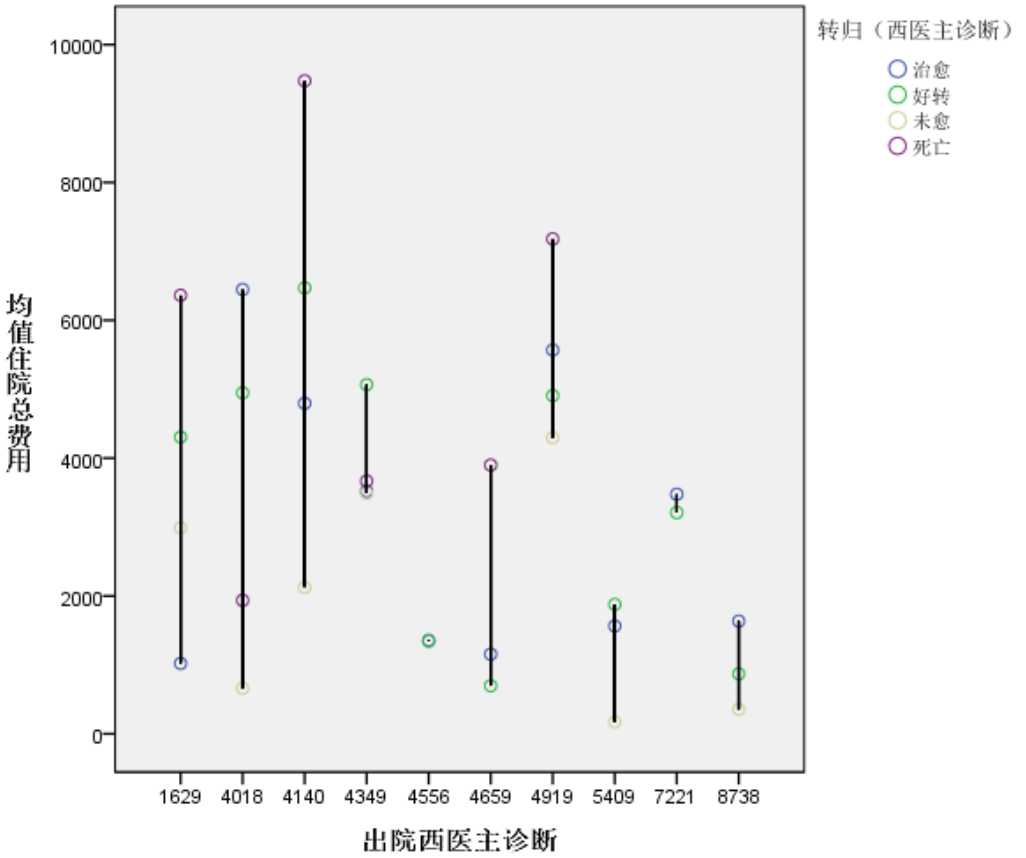


从图中可以看出，编号为1629的疾病需要的住院费用均值最高，高达12000，而编号为4569的疾病需要的住院费用均值最低，仅为1200左右，因此，可以明显看出不同病情类型影响着住院费用的高低，不同的病情类型之间存在着明显的收费差异。

4.5 不同治疗效果对住院费用的影响分析

不同病情类型会有不同的治疗方法和治疗效果，因此不妨控制病情类型这一变量，输出不同病情类型治疗效果的住院费用垂直线图，如下图所示：

不同病情类型治疗效果的住院费用



有垂直线图可知，在图中10种病情中，有3种是病情治愈所花费的费用最多，4种是病情好转所花费的费用最多。此外，还有4种是死亡花费的费用最多。因此可以得出结论，治疗效果对住院费用的影响不大。

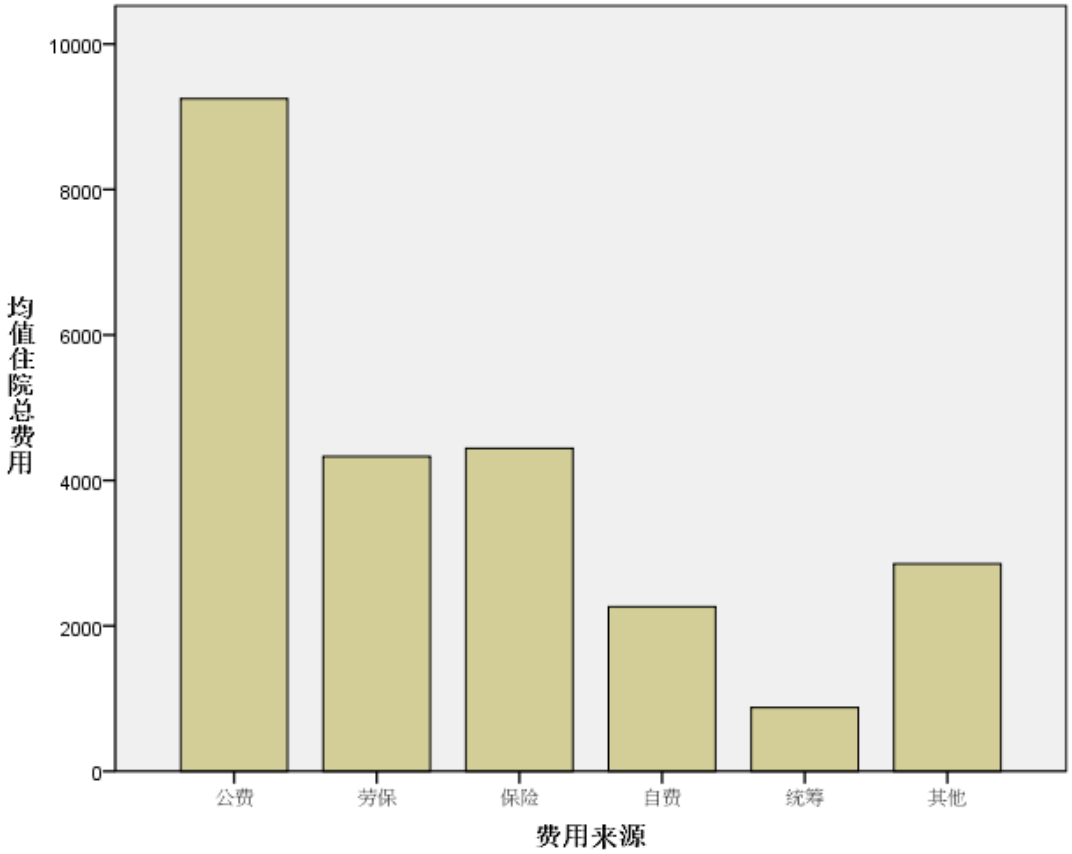
4.6 不同费用来源对住院费用的影响分析

对不同费用来源的住院总费用进行分类汇总，输出以下表格：

不同费用来源的住院总费用情况

		住院总费用	
		均值	列 N %
费用来源	公费	9249	22.8%
	劳保	4327	1.6%
	保险	4440	8.5%
	自费	2262	64.9%
	统筹	876	0.2%
	其他	2853	2.1%

不同费用来源的住院总费用情况



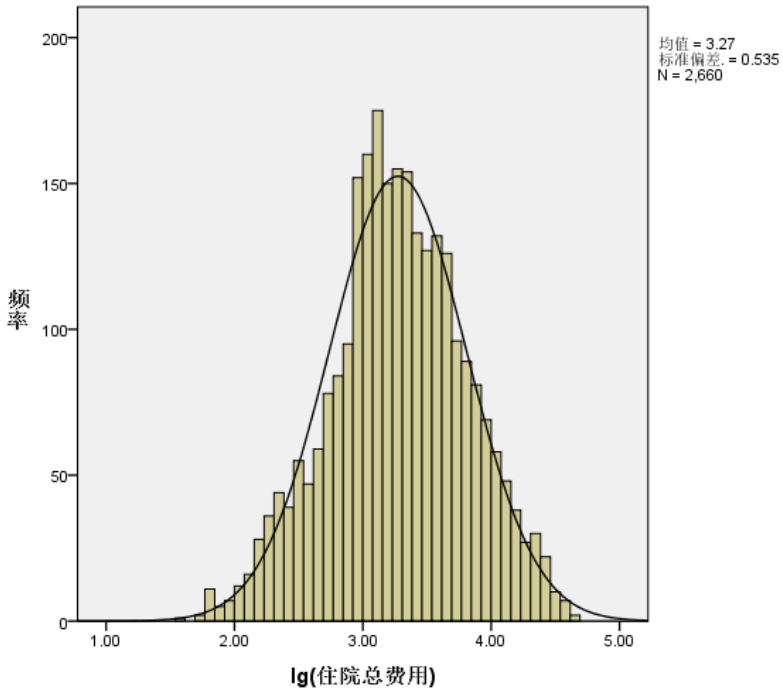
由此可知，自费医疗人数所占比例最高，为64.9%，而公费这一费用来源的住院费用均值最高，超过9000元。

5、对住院费用影响因素的回归分析

5.1 变量选择

影响住院费用的因素有很多，为了确定个属性对住院费用是否有影响及影响的具体形式，此篇报告将以住院费用为因变量，选取性别、年龄、医院地区、医院级别、病情类型、费用来源、入院情况、转归情况、治疗天数共 9 个变量作为自变量。由于住院费用呈偏态分布，因此需要对住院费用采用对数变换，取以 10 为底的对数 log10 进行标准化处理，将非线性关系转变为线性关系，使住院费用对数化之后呈正态分布，输出对数化的住院费用分布直方图如下图：

对数化的住院费用分布直方图



5.2 回归分析

选取的变量以及变量的度量方式如图：

选取的变量及其度量方式

变量代码	变量名称	度量方式
X1	性别	男=1，女=2
X2	年龄	岁数
X3	入院情况	危=1，急=2，一般=3
X4	住院天数	实际天数
X5	转归	治愈=1，好转=2，未愈=3，死亡=4
X6	治疗类别	中医=1，西医=2，中西医=3
X7	费用来源	公费=1，劳保=2，保险=3，自费=4，统筹=5，其他=9
X8	医院地区	东部=1，中部=2，西部=3
X9	医院级别	省级=1，地级=2，县级=3

选取对数化的住院费用为因变量，对上述9个变量进入自变量块进行回归分析。

5.3 结果解释

相关系数及标准误差表 模型摘要

模型	R	R 方	调整 R 方	标准估计的误差
6	.715[f]	.511	.510	.37488
7	.732[g]	.535	.534	.36538
8	.737[h]	.543	.541	.36258
9	.805[i]	.648	.647	.31798

i. 预测变量：(常量)，性别，年龄，入院情况，住院天数，转归（西医主诊断），治疗类别，费用来源，地区，级别。

模型摘要表中R表示线性回归的相关系数，体现了Y被X的确定程度。由模型摘要表可看出有4个模型，此处选择模型9，模型9的R为0.805，拟合度相对较高，可见这个方程拟合得较好。该模型有9个预测变量，其复相关系数R值为0.805，决定系数R方的值为0.648，可以看出自变量对因变量有一定的影响。

方差分析表 Anova[j]

模型		平方和	df	均方	F	Sig.
9	回归	492.939	9	54.771	541.684	.000[i]
	残差	267.240	2643	.101		
	总计	760.179	2652			

a. 因变量：lg(住院总费用)

b. 预测变量：(常量)，住院天数，入院情况，性别，出院西医主诊断，地区，转归（西医主诊断），级别，费用来源，年龄。

由方差分析表可以看出回归方程的相伴概率值（显著性）为 0.000，小于 0.01，说明回归方程高度显著，在置信度为 99%的情况下通过了检验。即说明预测变量整体上对因变量有高度显著的线性影响。

回归系数及检验统计量 t、相伴概率值

系数[a]

模型		非标准化系数		标准系数		
		B	标准 误差	试用版	t	Sig.
9	(常量)	3.615	.053		68.133	.000
	性别	-.037	.013	-.034	-2.898	.004
	年龄	.006	.000	.279	20.139	.000

入院情况	-.006	.010	-.007	-.577	.564
住院天数	.007	.000	.344	25.895	.000
转归（西医主诊断）	.005	.011	.006	.446	.656
治疗类别	-.005	.009	-.007	-.569	.569
费用来源	-.045	.005	-.119	-9.050	.000
地区	-.053	.008	-.081	-6.755	.000
级别	-.240	.009	-.367	-28.189	.000

a. 因变量: lg(住院总费用)

由回归系数及检验统计量表看出，自变量住院天数、入院情况、性别、出院、地区、转归、级别、费用来源、年龄，Sig 值表示显著水平，Sig 为 0.000，小于 0.05，说明自变量的回归系数在 99%的置信度上通过了检验，回归方程显著有意义。此时，疾病种类、入院情况、转归的显著系数分别为 0.569、0.564、0.656 均大于 0.05，所以应当剔除这些变量。同时，由于性别与对数化的住院费用进行回归分析时 R 值只有 0.115，大于 0.1，不能通过检验，说明相关性较小，故将性别变量也剔除。

剔除以上变量后重新进行回归拟合，得到下表：

第二次回归拟合的输出表

系数[a]

模型	非标准化系数		标准系数	t	Sig.
	B	标准 误差	试用版		
(常量)	3.546	.037		96.785	.000
年龄	.006	.000	.280	21.853	.000
住院天数	.007	.000	.347	26.283	.000
费用来源	-.047	.005	-.123	-9.576	.000
地区	-.053	.008	-.081	-6.823	.000
级别	-.239	.008	-.367	-28.328	.000

a. 因变量: lg(住院总费用)

5.4 得出结论

由以上分析，可以建立多元线性回归模型，标准化后为：
Y=0.28X2+0.347X4-0.123X7-0.081X8-0.367X9。

由方程可以看出年龄每增加一个标准差，住院费用增加-0.28个标准差；住院天数每增加一个标准差，住院费用就会增加0.347个标准差；医院地区每增加

-0.081个标准差，住院费用增加-0.367个标准差。由此可以看出，住院费用与年龄、住院天数呈正相关，与费用来源、医院地区、医院级别呈负相关。影响住院费用的主要因素从大到小为：医院级别、住院天数、年龄、费用来源、医院地区。

6、结语

通过对数据从各方面进行分析，可以得出以下结论：

1）从住院费用的基本构成情况这一角度，住院总费用的分布状况不满足一定的规律性；西药费、诊疗费、中成药费、床位费、其他费在住院费用构成中居于前五位；不同病情类型的病人住院费用种类也各不相同。

2）从影响住院费用的因素这一角度，性别、年龄、医院地区、医院级别等都可能在不同程度上影响住院费用。单因素分析可以比较直观地看出各因素对住院费用的影响。

3）从对住院费用影响因素的多元回归分析这一角度，住院费用与年龄、住院天数呈正相关，与费用来源、医院地区、医院级别呈负相关。影响住院费用的主要因素从大到小为：医院级别、住院天数、年龄、费用来源、医院地区。