四川大學

课 程 报 告

课　　程　分类数据分析

学　　院　商学院

学生姓名　万科

专　　业　工业工程

学　　号　2019225025117

年　　级　2019 级

指导教师　李宗敏

教务处制表

二〇二〇年 十一月 十六日

1 问题描述

某公司一直实行的是岗位等级制薪酬，薪酬水平与岗位直接相关，岗变薪变。此外，薪酬还与出差、加班等情况有关，员工如果想提高工资，只有两种方式：一是有空缺职位时的补缺，即升职；二是不断加班或出差。近几年随着一大批各岗位的老员工退休，公司要补充新人。新人的来源主要有两个是从毕业生中直接招聘，从头培养；二是招聘有经验的员工担任重要的管理和技术岗位。公司逐渐感觉到能力、经验比较强的员工不愿意被提拔到更高的岗位等级上，工作积极性不高，且有经验的女员工流失比例很大，社会上有关于该公司薪酬歧视女性员工的说法，导致招聘新的女职工也较难，公司男女比例越来越失调。管理层认识到，必须要进行薪酬体系的变革，要尽快解决性别公平性的问题、岗位等级薪酬差异问题。在制定薪酬改革方案之前，公司准备做一次全面的情况调查，并调查一下职工对于薪酬体系改革的态度，你如果是人事主管，应如何入手？具体要解决的问题如下所示：

(1). 是否女性的工资显著低于男性的工资?

(2). 性别与工资水平的相关程度如何？性别与工资水平的关联强度有多大?

(3). 岗位水平与工资水平关联的强度。

(4). 支持改革的比例是否超过了 50%？

(5). 男女支持改革的比例是否有差异？

2 问题解析

(1). 是否女性的工资显著低于男性的工资?

原假设 H0 是女性的工资水平与男性的工资水平间不存在显著性差异（95%的置信水平）。令女性的工资水平为 m1，男性的工资水平为 m2，得 H0：m1<m2，H1：m1>=m2。

(2). 性别与工资水平的相关程度如何？性别与工资水平的关联强度有多大?

本小题需探索性别和工资水平间关联性，将运用相关性分析的知识。

(3). 岗位水平与工资水平关联的强度。

本小题需探索岗位水平和工资水平间关联性，将运用相关性分析的知识。

(4). 支持改革的比例是否超过了 50%？

本小题将涉及到假设检验的知识，原假设 H0 是支持改革的比例超过了 50%（95%的置信水平）。令支持改革的比例为 m，得 H0：m>=0.5，H1：m<0.5。

(5). 男女支持改革的比例是否有差异？

原假设 H0 是男性和女性支持改革的比例不存在显著性差异（95%的置信水平）。令女性支持改革的比例为 m1，男性支持改革的比例为 m2，得 H0：m1=m2，H1：m1≠m2。此题可转换成独立性的开方检验，观察性

别和对改革的态度是否相关。

3 数据搜集

本报告数据来自已经提供的问卷调查结果，已经脱敏化。

4 数据分析及结果展示

4.1 探索性分析

对数据进行了描述性统计分析，共 220 条数据，无缺失值，其中岗位级别、性别和态度为分类变量。

ID	level	gender	salary	attitude
2	5	1	5800	2
3	5	1	5801	2
11	5	1	5802	2
12	5	1	5803	1
16	5	1	5804	2

```
df.describe()
```

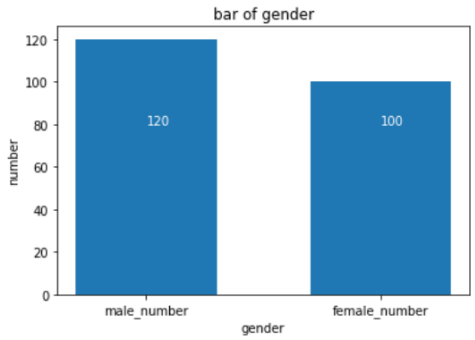
	level	gender	salary	attitude
count	220.000000	220.000000	220.000000	220.000000
mean	3.150000	1.454545	5026.531818	1.527273
std	1.430346	0.499065	857.032133	0.500394
min	1.000000	1.000000	3500.000000	1.000000
25%	2.000000	1.000000	4501.000000	1.000000
50%	3.000000	1.000000	5004.000000	2.000000
75%	4.000000	2.000000	5801.250000	2.000000
max	5.000000	2.000000	6709.000000	2.000000

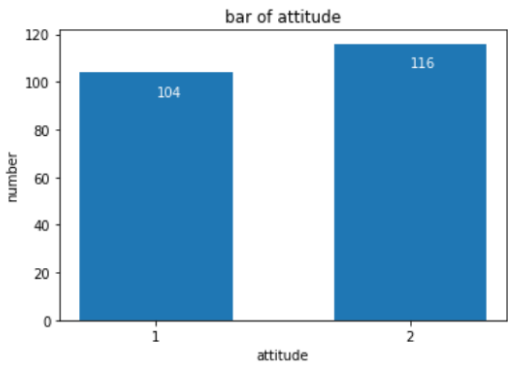
```
for i in ['level','gender','attitude']:
    print('the unique values of {}'.format(i)+' is {}'.format(df[i].unique()))
```

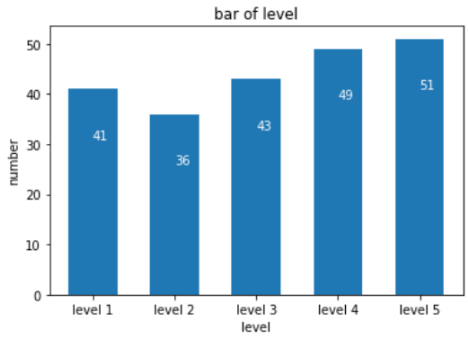
```
the unique values of level is [3 5 4 1 2]
the unique values of gender is [1 2]
the unique values of attitude is [2 1]
```

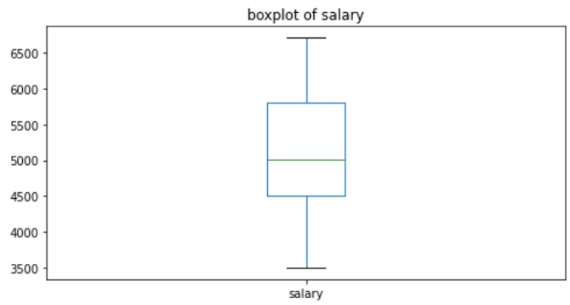
```
df.count()
```

```
level       220
gender      220
salary      220
attitude    220
dtype: int64
```









4.2 是否女性工资显著低于男性工资

对工资进行分段，大致分为三个区间，低收入群体[3500,4500]，中等收入群体[4500,5500]，高收入群体[5500,7000]。

结果解析：在 95%的置信水平下，

1) 首先将男性和女性分组，分别计算各自的平均工资；

2) 其次再进行假设检验，由于不知道总体的方差，这里采取 t 检验

低收入群体：[3500,4500]

```
male_df1=d1.loc[df['gender']==1]
female_df1=d1.loc[df['gender']==2]
from scipy import stats
stats.t.ppf(0.05, 219)
stat, p = stats.ttest_ind(male_df1['salary'],female_df1['salary'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('不能拒绝原假设，低收入群体中，女性工资显著低于男性。')
else:
    print('拒绝原假设，低收入群体中，女性工资不显著低于男性。')
```

```
stat=-3.967, p=0.000
拒绝原假设，低收入群体中，女性工资不显著低于男性。
```

中等收入群体：[4500,5500]

```
male_df3=d3.loc[df['gender']==1]
female_df3=d3.loc[df['gender']==2]
from scipy import stats
stats.t.ppf(0.05, 219)
stat, p = stats.ttest_ind(male_df3['salary'],female_df3['salary'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('不能拒绝原假设，中等收入群体中，女性工资显著低于男性。')
else:
    print('拒绝原假设，中等收入群体中，女性工资不显著低于男性。')
```

```
stat=-0.631, p=0.530
不能拒绝原假设，中等收入群体中，女性工资显著低于男性。
```

高等收入群体：[5500,7000]

```
from scipy import stats
stats.t.ppf(0.05, 219)
stat, p = stats.ttest_ind(male_df['salary'],female_df['salary'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('不能拒绝原假设，女性工资显著低于男性。')
else:
    print('拒绝原假设，女性工资不显著低于男性。')
```

```
stat=-1.693, p=0.092
不能拒绝原假设，女性工资显著低于男性。
```

不划分区间的总体收入

```
male_df2=d2.loc[df['gender']==1]
female_df2=d2.loc[df['gender']==2]
from scipy import stats
stats.t.ppf(0.05, 219)
stat, p = stats.ttest_ind(male_df2['salary'],female_df2['salary'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('不能拒绝原假设，高收入群体中，女性工资显著低于男性。')
else:
    print('拒绝原假设，高收入群体中，女性工资不显著低于男性。')
```

```
stat=-1.875, p=0.064
不能拒绝原假设，高收入群体中，女性工资显著低于男性。
```

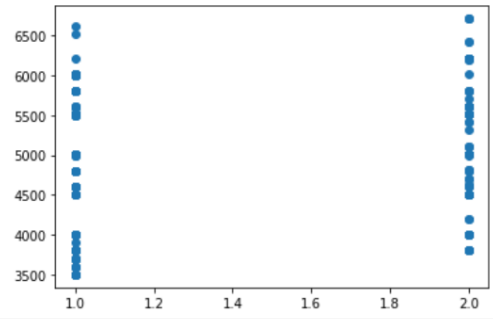
综上所述，仅低收入群体中女性收入不显著低于男性，而中等收入、高收入群体中女性收入显著低于男性，针对不划分区间的总体收入而言，女性收入是显著低于男性的。

4.3 性别与工资水平的相关程度如何？性别与工资水平的关联强度有多大?

因为性别是离散型变量，因此检验性别与工资之间的相关性采用点二列相关

```
plt.scatter(df['gender'], df['salary'])
```

```
<matplotlib.collections.PathCollection at 0x23f63564780>
```




```
from scipy import stats
stats.pointbiserialr(df['gender'],df['salary'])
```

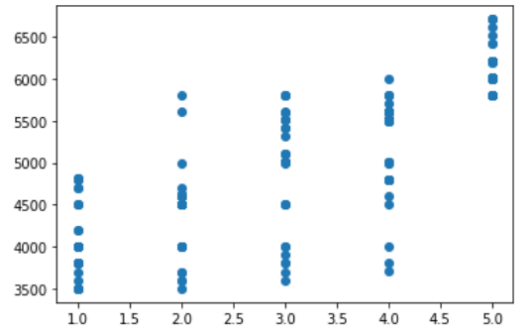
```
PointbiserialrResult(correlation=0.11390914016420847, pvalue=0.09190934179189782)
```

p 值大于 0.05，故接受原假设，工资与性别不太相关，相关系数为 0.1139

4.4 岗位水平与工资水平关联的强度

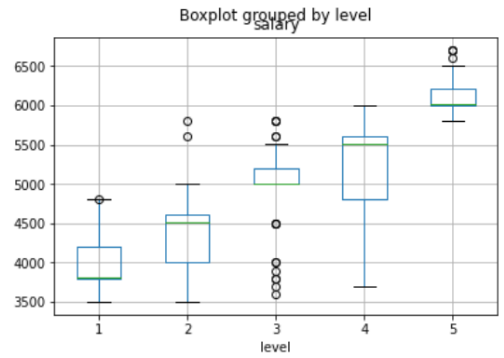
```
from matplotlib import pyplot as plt
plt.scatter(df['level'], df['salary'])
```

```
<matplotlib.collections.PathCollection at 0x23f634f2780>
```




```
df.boxplot('salary','level')
```

```
<AxesSubplot:title={'center':'salary'}, xlabel='level'>
```




```
h=df['level'].corr(df['salary'],method='spearman')
print('相关系数为：{}'.format(h))
print('p值为：{}'.format(p))
```

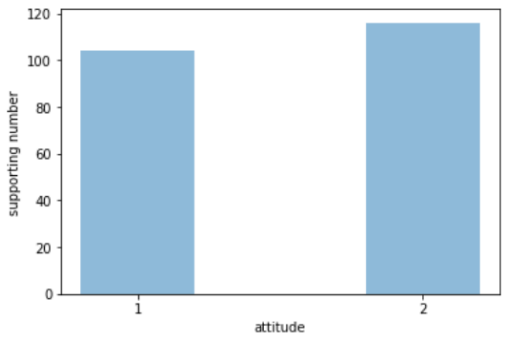
```
相关系数为：0.838215660512987
p值为：0.00023740610567732405
```

```
stats.pointbiserialr(df['level'],df['salary'])
```

```
PointbiserialrResult(correlation=0.8340125059900011, pvalue=3.230327151035903e-58)
```

因为岗位水平为离散型变量，因此建立工资水平与岗位水平之间的相关性采用斯皮尔曼检验与点二列检验。p 值小于 0.05 故拒绝原假设，岗位水平与工资水平相关，相关系数为 0.834

4.5 支持改革的比例是否超过了 50%？




```
df.groupby('attitude').size()
```

```
attitude
1    104
2    116
dtype: int64
```

```
#样本比例
percentage=104/220
percentage
```

```
0.4727272727272727
```

```
#支持比例
aim=0.5
```

- np,n(1-p)>5,大样本情形

```
def percentage_test(d,D,n):
    z=(d-D)/np.sqrt(D*(1-D)/n)
    if np.abs(z)>1.65:
        print('拒绝原假设，支持改革的比例不超过50%')
    else:
        print('接受原假设，支持改革的比例超过50%')
```

```
percentage_test(percentage,aim,220)
```

```
接受原假设，支持改革的比例超过50%
```

因为检验是支持改革的比例，因此采用总体比例的假设检验

p 值大于 0.05，故接受原假设，支持改革的比例超过 50%

4.6 男女支持改革的比例是否有差异？

男女支持改革列联表的联合分布

```
import pandas as pd
sc_contingencytable=pd.crosstab(df['gender'],df['attitude'],margins=True)
sc_contingencytable
```

attitude	1	2	All
gender			
1	60	60	120
2	44	56	100
All	104	116	220

```
#可以计算各个单元格占总人数的百分比:联合分布
sc_contingencytable/sc_contingencytable.loc['All']['All']
```

attitude	1	2	All
gender			
1	0.272727	0.272727	0.545455
2	0.200000	0.254545	0.454545
All	0.472727	0.527273	1.000000

男女支持改革的条件分布

```
#条件分布
def percent_observed(data):
    return data/data['All']
cross=pd.crosstab(df['gender'],df['attitude'],
           margins=True).apply(percent_observed,axis=1)
```

```
gen_att=cross.iloc[:-1,:-1]
gen_att
```

attitude	1	2
gender		
1	0.50	0.50
2	0.44	0.56

相对风险：

```
male_risk=gen_att.iloc[:,0][1]/gen_att.iloc[:,0][2]
female_risk=gen_att.iloc[:,1][1]/gen_att.iloc[:,0][2]
print('支持改革的相对风险为'+str(male_risk))
print('不支持改革的相对风险为'+str(female_risk))
```

```
支持改革的相对风险为1.1363636363636365
不支持改革的相对风险为1.1363636363636365
```

优势比：男性支持改革与不支持的概率相同，女性不支持改革的较多，但目前女性中高等收入显著低于男性，理论上说支持改革的女性应该较多，但事实恰好相反，这点值得企业深思。

```
male_odds=gen_att.iloc[0,:][1]/gen_att.iloc[0,:][2]
female_odds=gen_att.iloc[1,:][1]/gen_att.iloc[1,:][2]
print('男性支持改革的优势比为'+str(male_odds))
print('女性支持改革的优势比为'+str(female_odds))
```

```
男性支持改革的优势比为1.0
女性支持改革的优势比为0.7857142857142857
```

```
sc_contingencytable.iloc[:-1,:-1]
```

attitude	1	2
gender		
1	60	60
2	44	56

```
from  scipy.stats import chi2_contingency
chi2_contingency(sc_contingencytable.iloc[:-1,:-1])
```

```
(0.5654702696728561, 0.45206454609626645, 1, array([[56.72727273, 63.27272727],
       [47.27272727, 52.72727273]]))
```

结果分析：在 95%的置信水平下，

1) 首先，按照性别进行态度的列联表化。

2) 其次，进行独立性的开方检验，得出的 p 值是 0.452，远大于 0.05，接受原假设，说明性别和态度是独立的，不存在显著差异。

综上所述，男女支持改革的比例不存在差异。

5 建议

通过上述的数据分析，在 95%的置信水平下，有几个有趣的发现，如下：

(1). 关于性别对薪酬的影响：性别对工资水平有低度的相关关系。该公司中等和高收入的女性工资显著低于男性工资，但是性别与工资水平又不存在显著的相关性，这说明公司里不存在明显的工资性别歧视，但需提高中高等收入女员工的收入水平。

(2). 工资水平和岗位级别有很强的关联性， 但同一级别之间工资水平的差异也很明显，这也解释了为什么很多员工不愿意被提拔到更高的岗位级别

(3). 公司内支持进行薪酬改革的比例占到 50%，且对改革的态度不因性别而发生差异，说明公司内支持薪酬改革已成为大势所趋，公司高管应当考虑重新构建薪酬体制。

综上所述，公司内不存在明显的性别薪酬差异，工资水平和职位水平有关联性，调查发现公司内职员均希望进行薪酬改革，这告诉公司需要不定期对员工薪酬态度进行调查，进行薪酬工资体制的再合理化。同时增加女员工福利，如带薪产假，妇女节礼物等方式，在宣讲会中宣传，吸引更多的女员工，解决性别失衡问题。