Терников Андрей

Статистический анализ занятости населения

Источник данных: http://www.economicswebinstitute.org/data/wagesmicrodata.xls

Описание: выборка состоит из 534 наблюдений по 11 переменным [зарплата, сфера занятости, рабочий сектор, членство в профсоюзе, количество лет обучения, рабочий стаж, возраст, пол, брачный статус, раса, место жительства] и представляет собой данные, собранные по американской переписи населения 1985 года.

Гипотеза (1): влияет ли членство в профсоюзе на размер зарплаты?
H0: з/п у членов профсоюза и у людей, не состоящих в профсоюзе, равны.
H1: з/п у членов профсоюза выше чем у людей, не состоящих в профсоюзе.

Проверка на нормальность выборки, расщепленной по членству в профсоюзе, дает отрицательный результат. Гипотеза H0 о том, что выборка распределена нормально отвергается на уровне значимости меньше 0,001 по критерию Шапиро-Уилка и по критерию Колмогорова-Смирнова для тех, кто не состоит в профсоюзе и на уровне значимости 0,018 (<0,05) для тех, кто состоит в профсоюзе.

Tests of Normality[b]

Union_ member	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
No	,142	438	,000	,839	438	,000
Yes	,101	96	,018	,930	96	,000

Для проверки гипотез воспользуемся критерием Манна-Уитни (для 2-х независимых выборок по членству в профсоюзе), который не требует проверки на нормальность. В нашем случае объем выборки большой, тогда критерий Манна-Уитни сводит распределение выборки к нормальному распределению, а некоторой неточностью по сравнению с t-критерием Стьюдента (используется для выборки, распределенной нормально) можно пренебречь. Выясняется, что гипотеза о том, что з/п у членов профсоюза и у людей, не состоящих в профсоюзе, равны, отвергается на уровне значимости меньшем 0,001 (сравнивая с односторонней значимостью для больших выборок). Следовательно, принимаем гипотезу H1.

Test Statistics[a]

	Wage
Mann-Whitney U	1,378E4
Wilcoxon W	1,099E5
Z	-5,292
Asymp. Sig. (2-tailed)	,000
Asymp. Sig. (1-tailed)	,000

a. Grouping Variable: Union_member

Вывод (1): з/п у членов профсоюза в среднем выше чем у людей, не состоящих в профсоюзе.

Гипотеза (2): существует ли связь между тем, в какой сфере занят работник, и рабочим сектором?
H0: связь между тем, в какой сфере занят работник, и рабочим сектором отсутствует.
H1: связь между тем, в какой сфере занят работник, и рабочим сектором существует.

Для проверки гипотез воспользуемся таблицей сопряженности, составленной для данных с номинальными шкалами.

Sector * Occupation Crosstabulation

			Occupation						
			Management	Sales	Clerical	Service	Prof	Other	Total
Sector	Other	Count	49	34	88	81	91	68	411
		Expected Count	42,3	29,2	74,7	63,9	80,8	120,1	411,0
	Manufacture	Count	6	4	7	2	12	68	99
		Expected Count	10,2	7,0	18,0	15,4	19,5	28,9	99,0
	Consruction	Count	0	0	2	0	2	20	24
		Expected Count	2,5	1,7	4,4	3,7	4,7	7,0	24,0
Total		Count	55	38	97	83	105	156	534
		Expected Count	55,0	38,0	97,0	83,0	105,0	156,0	534,0

Воспользуемся критериями Фи и Крамера для оценки меры тесноты связи. Мы отвергаем гипотезу H0 о том, что связь между тем, в какой сфере занят работник, и рабочим сектором отсутствует на уровне значимости меньше 0,001 (по обоим критериям). Следовательно, гипотеза H1 принимается.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,520	,000
	Cramer's V	,368	,000
N of Valid Cases		534	

Вывод (2): связь между тем, в какой сфере занят работник, и рабочим сектором существует.

Гипотеза (3): существует ли связь между зарплатой, количеством лет обучения и рабочим стажем?
H0: связь между зарплатой и количеством лет обучения (рабочим стажем) отсутствует.
H1: связь между зарплатой и количеством лет обучения (рабочим стажем) существует.

Воспользуемся ранговым коэффициентом корреляции Кендалла, поскольку он не требует проверки на нормальность, не чувствителен к выбросам и применим для больших выборок. В обоих случаях (для образования и стажа) мы отвергаем нулевую гипотезу. Кроме того, статистическая достоверность выявленной связи подтверждается на уровне значимости 0,01 (отвергаем гипотезу об отсутствии связи).

Correlations

			Wage	Education
Kendall's tau_b	Wage	Correlation Coefficient	1,000	,282**
		Sig. (2-tailed)	.	,000
		N	534	534
	Education	Correlation Coefficient	,282**	1,000
		Sig. (2-tailed)	,000	.
		N	534	534

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

			Wage	Experience
Kendall's tau_b	Wage	Correlation Coefficient	1,000	,117**
		Sig. (2-tailed)	.	,000
		N	534	534
	Experience	Correlation Coefficient	,117**	1,000
		Sig. (2-tailed)	,000	.
		N	534	534

**. Correlation is significant at the 0.01 level (2-tailed).

Вывод (3): связь между зарплатой и количеством лет обучения (рабочим стажем) существует (на уровне значимости 0,01).

Гипотеза (4): существует ли связь между зарплатой, расой и брачным статусом?
H0: раса и брачный статус не имеют эффекта взаимодействия на з/п.
H1: раса и брачный статус имеют эффект взаимодействия на з/п.

H0: з/п не зависит от расовой принадлежности.
H1: з/п зависит от расовой принадлежности.

H0: з/п не зависит от брачного статуса.
H1: з/п зависит от брачного статуса.

Сперва проверим выборку на нормальность. Гипотеза о том, что выборка распределена нормально, отвергается.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Race	Statistic	df	Sig.	Statistic	df	Sig.
Wage	Other	,124	67	,012	,909	67	,000
	Hispanic	,234	27	,001	,635	27	,000
	White	,117	440	,000	,872	440	,000

a. Lilliefors Significance Correction

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Marital_status	Statistic	df	Sig.	Statistic	df	Sig.
Wage	Unmarried	,171	184	,000	,776	184	,000
	Married	,103	350	,000	,907	350	,000

a. Lilliefors Significance Correction

Однако необходимо проверить выборку на то, что стандартная ошибка выборки равна между зависимыми переменными, и то, что дисперсии для разных групп гомогенны. Это позволит нам говорить о том, что ошибки, отличающие распределения от нормальных, идентичны, и мы можем ими пренебречь в дальнейшем анализе. По критерию Ливеня нулевая гипотеза о том, что дисперсии для каждой из групп статистически достоверно не различаются, принимается (0,675>0,05). Кроме того, по критерию Шеффе (на гомогенность) нулевая гипотеза о том, что дисперсии по группам гомогенны, принимается (0,122>0,05).

Wage

Scheffe

Race	N	Subset 1
Hispanic	27	7,28
Other	67	8,06
White	440	9,28
Sig.		,122

Means for groups in homogeneous subsets are displayed.

Based on observed means.

The error term is Mean Square(Error) = 26,085.

Levene's Test of Equality of Error Variances[a]

Dependent Variable:Wage

F	df1	df2	Sig.
,633	5	528	,675

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Marital_status + Race + Marital_status * Race

Выводы (4):

Таким образом, есть возможность применения двухфакторного дисперсионного анализа.

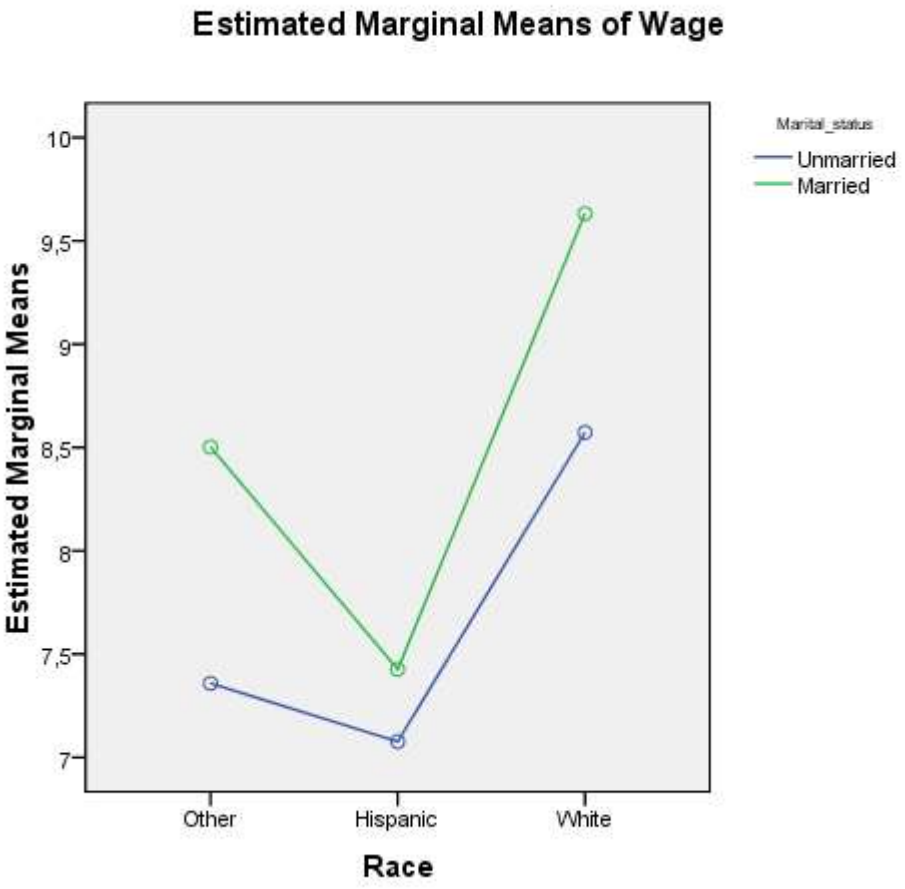
- Отвергаем гипотезу о том, что ни один из факторов модели не влияет на з/п (0,041<0,05).
- Принимаем гипотезу H0 о том, что з/п не зависит от брачного статуса (0,294>0,05).
- Мы можем принять на уровне значимости не меньшем, чем 0,063 гипотезу H1 о том, что з/п зависит от расовой принадлежности (0,062<0,063).
- Принимаем гипотезу H0 о том, что раса и брачный статус не имеют эффекта взаимодействия на з/п (0,939>0,05) (что подтверждается на графике средних).

Tests of Between-Subjects Effects

Dependent Variable:Wage

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	303,984[a]	5	60,797	2,331	,041
Intercept	10415,090	1	10415,090	399,280	,000
Marital_status	28,750	1	28,750	1,102	,294
Race	145,411	2	72,706	2,787	,062
Marital_status * Race	3,307	2	1,653	,063	,939
Error	13772,715	528	26,085		
Total	57562,308	534			
Corrected Total	14076,699	533			

a. R Squared = ,022 (Adjusted R Squared = ,012)



Проводя множественные сравнения, обнаруживается, что средние в группах при попарном сравнении значимо не различаются (принимается нулевая гипотеза о равенстве средних). Однако из полученных данных можно сделать вывод о том, что связь между «испанской» (латиноамериканской) и «белой» (европейской) расой статистически более значима (при уровне значимости 0,15 и ниже), чем при других попарных сравнениях.

Multiple Comparisons

Wage

Scheffe

(I) Race	(J) Race	Mean Difference (I-J)	Std. Error	Sig.	85% Confidence Interval	
					Lower Bound	Upper Bound
Other	Hispanic	,78	1,164	,801	-1,50	3,05
	White	-1,22	,670	,192	-2,53	,09
Hispanic	Other	-,78	1,164	,801	-3,05	1,50
	White	-1,99*	1,013	,145	-3,97	-,02
White	Other	1,22	,670	,192	-,09	2,53
	Hispanic	1,99*	1,013	,145	,02	3,97

Based on observed means.

The error term is Mean Square(Error) = 26,085.

*. The mean difference is significant at the ,15 level.