Домашнее задание
вариант 6

от Татаринова Никиты Алексеевича, БПИ196

к 21.02.2022

Пункт 1

Выборка представляет из себя 76 наблюдений (без пропущенных значений) с данными о людях трудоспособного возраста из Волосовского района Ленинградской области (id6). Генеральной совокупностью, в таком случае, можно считать всех людей трудноспособного возраста из данного региона.

Для проверки репрезентативности выборки необходимо, в первую очередь, знать о методе построения выборки, то есть была выборка детерминированной или вероятностной (в случае детерминированной репрезентативность может быть сильно искажена). Кроме того, размер выборки оказывает существенное влияние на репрезентативность выборки – в данном случае имеется 76 наблюдений, что, как правило, мало для репрезентативной выборки, особенно при наличии других ошибок (например, ошибки измерения).

Пункт 2

Начнём с фомул рассчёта требуемых характеристик:

$$\begin{cases} \overline{X} = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i & \text{– среднее значение} \\ \hat{\sigma}^2 = \dfrac{1}{n-1}\sum\limits_{i=1}^{n} \left(x_i - \overline{X}\right)^2 & \text{– квадрат стандартного отклонения} \end{cases}$$

Для подсчёта требуемых характеристик используем Gretl.

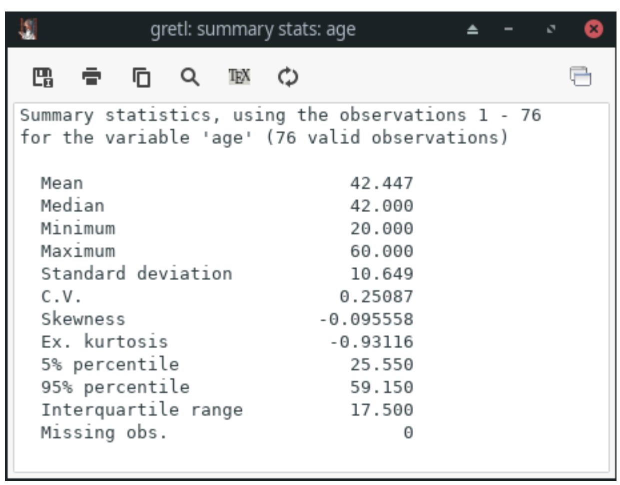


Рис. 1: Использование Gretl для подсчёта характеристик на примере переменной age.

	минимум	максимум	среднее значение	стандартное отклонение	размах = максимум - - минимум
public	0	1	~ 0.46	~ 0.50	1
lnwage	~ 7.60	~ 11.00	~ 10.05	~ 0.60	~ 3.40
educ	0	3	1.87	0.93	3
urban	0	1	~ 0.49	~ 0.50	1
male	0	1	~ 0.37	~ 0.49	1
age	20	60	~ 42.45	~ 10.65	40
children	0	4	~ 1.34	~ 0.87	4
internet	0	1	~ 0.78	0.42	1

Пункт 3

Для рассчёта p%-го перцентиля воспользуемся формулой:

$$i = \frac{p}{100} \cdot n$$

, где p – искомый перцентиль, n – число наблюдений. Причём, если i оказалось нецелым, то округлённое вверх i-ое значение в отсортированном по возрастанию списке является искомым p%-м перцентилем. Если i оказалось целым, то искомый перцентиль вычисляется как среднее из i-го и $(i+1)$-го значений в отсортированном по возрастанию списке.

$$\begin{cases} i_{25} = \frac{25}{100} \cdot 76 = 19 \\ i_{50} = \frac{50}{100} \cdot 76 = 38 \\ i_{75} = \frac{75}{100} \cdot 76 = 57 \end{cases}$$

К непрерывным переменным относится логарифм зарплаты и можно отнести возраст.

	Q_1	Q_2	Q_3	$IQR = Q_3 - Q_1$
lnwage	9.740969	10.12663	10.532002	0.791033
age	33.5	42	50.5	17

Пункт 4

Для подсчёта моды используем функцию MODE в Excel.

	среднее	медиана	мода
lnwage	10.05	10.12663	10.59663
age	42.45	42	41

На основании полученных значений можно предположить, что график распределения частот логарифма зарплаты имеет левостороннюю скошенность, в то время как аналогичный график возраста, наоборот, имеет правостороннюю скошенность.

Пункт 5

Прежде, чем строить box-plot с помощью Gretl, рассчитаем вручную границы "усов".

	Q_1	Q_3	IQR	$X_1 = Q_1 - 1.5 \cdot (Q_3 - Q_1)$	$X_3 = Q_3 + 1.5 \cdot (Q_3 - Q_1)$
lnwage	9.740969	10.532002	0.791033	8.5544195	11.7185515
age	33.5	50.5	17	8	76

Теперь построим box-plot для логарифма зарплаты.

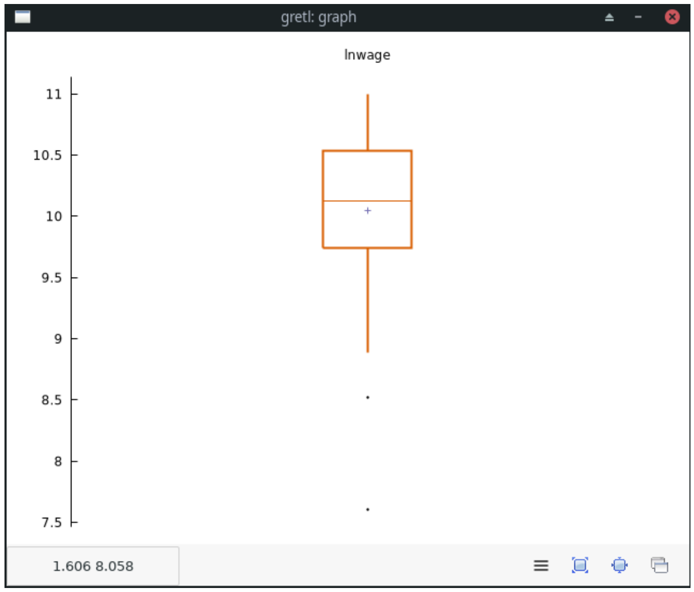


Рис. 2: Box-plot для переменной lnwage.

Обратим внимание, что вычисленные значения границ "усов" не совпадают с отображенными на графике. Связано это с тем, что итоговая граница верхнего (нижнего) "уса" ограничивается максимальным (минимальным) значением, попадаемым в изначально вычисленные границы. Если отсортировать данные по возрастанию логарифма заработной платы, то первыми 3мя значениями являются 7.600903, 8.517193, 8.888757, а изначально вычисленная граница нижнего "уса" – 8.5544195. Получается, что первые 2 точки являются выбросами, но при этом на полуинтервале $\left[8.5544195; 8.888757\right)$ нет точек, в связи с чем граница нижнего "уса" смещается до 8.888757. Что касается верхнего "уса", то изначальная граница имеет значение 11.7185515, в то время как максимальное значение в выборке – 11.0021. Поэтому, граница верхнего "уса" смещается до 11.0021.

Построим box-plot для возраста.

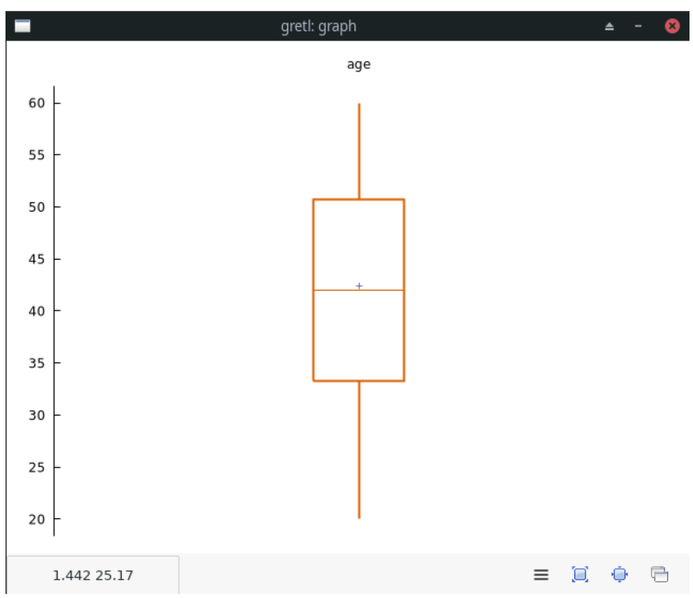


Рис. 3: Box-plot для переменной age.

Аналогично верхней границе box-plot для логарифма зарплат, обе границы box-plot для возраста ограничиваются минимальным и максимальным значениями выборки (выборосов нет, так как "усы" с запасом покрывают выборку).

Пункт 6

Формулы рассчёта коэффициентов ассиметрии (skewness) и островершинности (excess kurtosis) имеют вид:

$$\begin{cases} Sk = \dfrac{\frac{1}{n}\sum\limits_{i=1}^{n}\left(x_i - \overline{X}\right)^3}{s^3} \\ Ku = \dfrac{\frac{1}{n}\sum\limits_{i=1}^{n}\left(x_i - \overline{X}\right)^4}{s^4} - 3 \end{cases}$$

Коэффициент ассиметрии принимает значения:

- близкие к нулю, если гистограмма имеет симметричный вид;
- меньшие нуля, если гистограмма имеет левостороннюю скошенность;
- большие нуля, если диаграмма имеет правостороннюю скошенность.

Коэффициент островершинности принимает значения:

- близкие к нулю, если отсровершинность гистограммы схожа с островершинностью нормального распределения;
- большие нуля, если гистограмма имеет высокую островершнинность, то есть имеет длинные хвосты, что свидетельствует о возможном наличии выбросов;
- меньшие нуля, если гистограмма имеет низкую островершнинность, то есть имеет короткие хвосты, что свидетельствует о маловероятном наличии выбросов.

Построим гистограмму для логарифма зарплаты. Коэффициенты можно найти среди других описательных статистик (пункт 2).

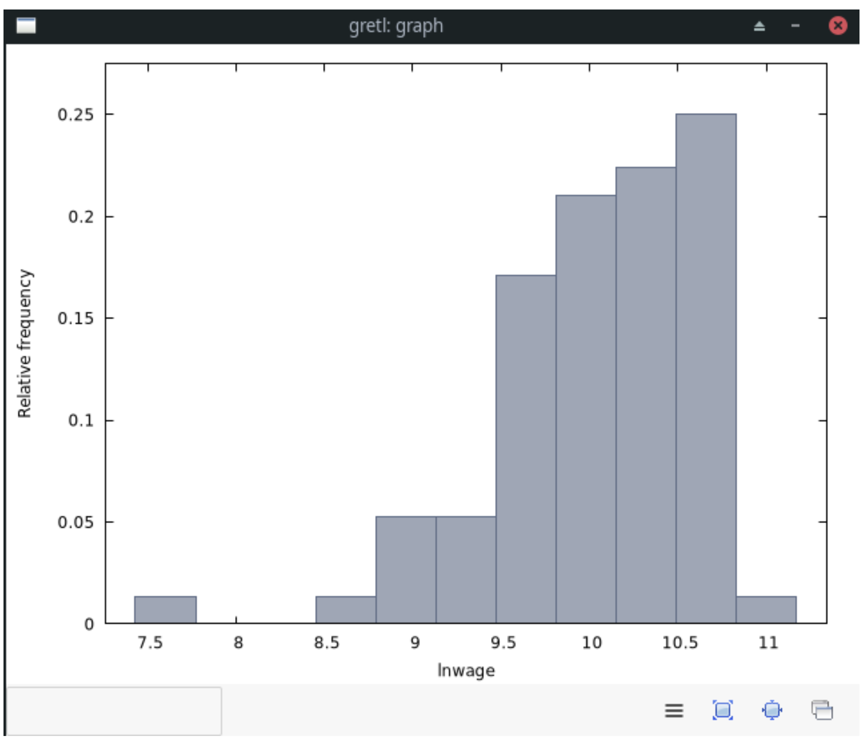


Рис. 4: График распределения частот для непрерывной переменной lnwage.

Как мы видим, гистограмма имеет явную левостороннюю скошенность и высокую островершинность – отрезок $(9,5;\ 10,7)$ содержит в себе большую часть наблюдений при наличии длинного хвоста $(7,5;\ 9,5)$. Об этом же свидетельствуют вышеописанные коэффициенты, имеющие значения $Sk = -1.2347$ и $Ku = 2.3808$.

Перейдём к построению гистограммы для возраста.

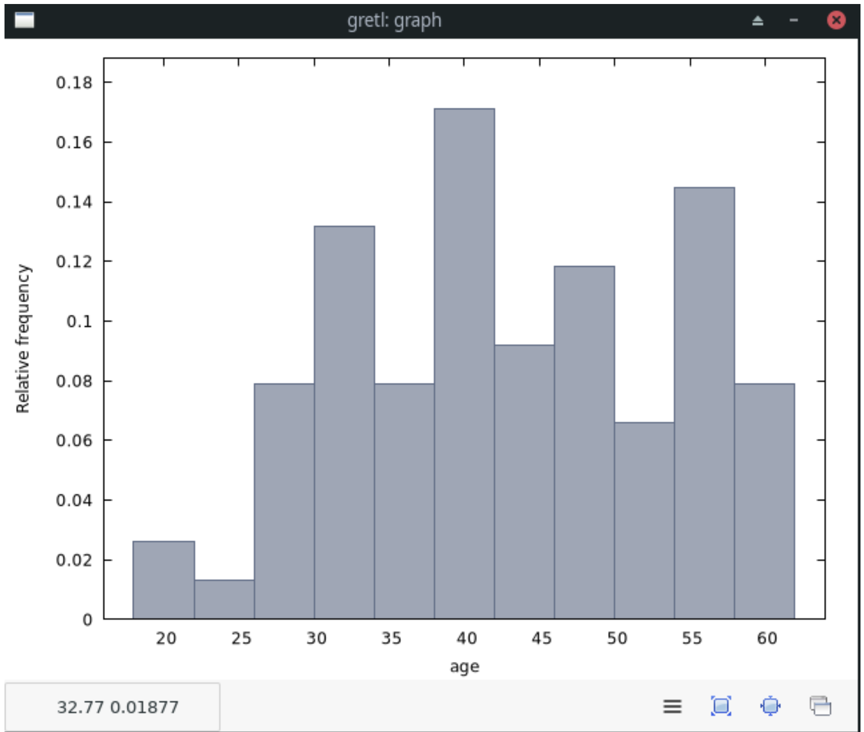


Рис. 5: График распределения частот для непрерывной переменной age.

Гистограмма выглядит достаточно симметрично, при этом интервал частовстречающихся возрастов достаточно широк, то есть островершинность низкая. Это подтверждается коэффициентами $Sk = -0.095558$ и $Ku = -0.93116$.

Пункт 7

Построим график распределения уровня образования.

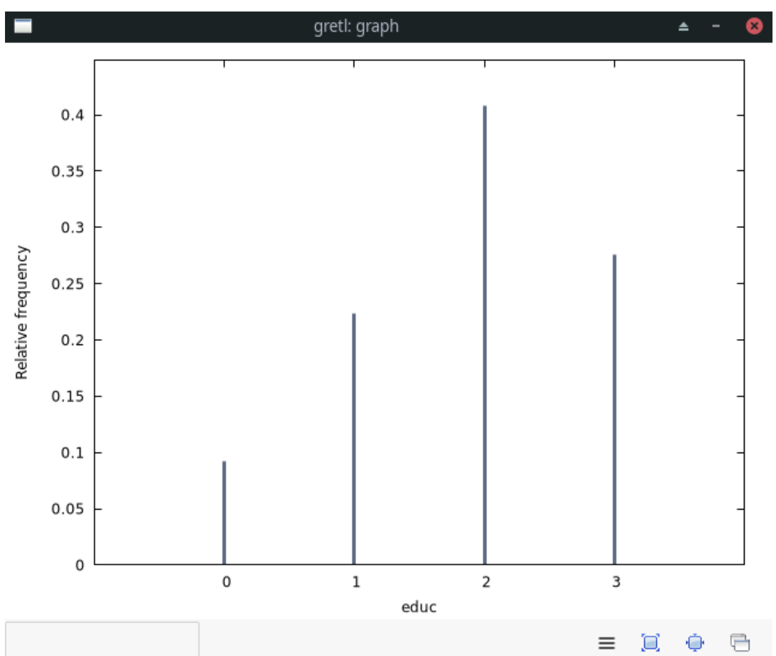


Рис. 6: График распределения частот для непрерывной переменной educ.

Получается, большинство респондентов ($\sim 40\%$) имеет среднее профессиональное образование, часть имеет высшее ($\sim 27.5\%$) или среднее общее ($\sim 22.5\%$) образование и меньшинство не имеет образования ($\sim 10\%$).

Пункт 8

Доверительные интервалы для математического ожидания (при неизвестной дисперсии) и для стандартного отклонения имеют вид:

$$\overline{X} - t_{n-1,\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$
$$\sqrt{\frac{(n-1)\cdot\hat{\sigma}^2}{\chi^2_{n-1,\frac{\alpha}{2}}}} < \sigma < \sqrt{\frac{(n-1)\cdot\hat{\sigma}^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}}$$

, где $(1-\alpha)$ – уровень доверия (α – вероятность ошибки), $t_{n-1,\frac{\alpha}{2}}:\ P\big\{U > t_{n-1,\frac{\alpha}{2}}\big\} = \frac{\alpha}{2}$; $U \sim t_{n-1}$, $\chi^2_{n-1,\frac{\alpha}{2}}:\ P\big\{Y > \chi^2_{n-1,\frac{\alpha}{2}}\big\} = \frac{\alpha}{2}$; $Y \sim \chi^2_{n-1}$. Тогда, при уровне доверия:

- $(1-\alpha) = 0.95$ доверительные интервалы имеют вид:

$$\begin{cases} 9.92 \approx 10.05 - 1.960 \cdot \frac{0.6}{\sqrt{76}} < \mu < 10.05 + 1.960 \cdot \frac{0.6}{\sqrt{76}} \approx 10.18 \\ 0.51 \approx \sqrt{\frac{(76-1)\cdot 0.6^2}{101,999}} < \sigma < \sqrt{\frac{(76-1)\cdot 0.6^2}{53.782}} \approx 0.71 \end{cases}$$

- $(1-\alpha) = 0.99$ доверительные интервалы имеют вид:

$$\begin{cases} 9.87 \approx 10.05 - 2.576 \cdot \frac{0.6}{\sqrt{76}} < \mu < 10.05 + 2.576 \cdot \frac{0.6}{\sqrt{76}} \approx 10.23 \\ 0.49 \approx \sqrt{\frac{(76-1)\cdot 0.6^2}{111,495}} < \sigma < \sqrt{\frac{(76-1)\cdot 0.6^2}{47.997}} \approx 0.75 \end{cases}$$

Пункт 9

Заметим, что переменная male принимает значение 1 для мужчины и 0 для женщины. В таком случае, среднее значение показывает выборочную долю мужчин, то есть для построения доверительного интервала для доли женщин нужно взять значение $\hat{p} = \big(1 - \overline{X}\big)$.

Доверительный интервал для доли имеет вид:

$$\hat{p} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(\hat{p}-1)}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(\hat{p}-1)}{n}}$$

, где $(1-\alpha)$ – уровень доверия (α – вероятность ошибки), $Z_{\frac{\alpha}{2}}:\ P\big\{Z > Z_{\frac{\alpha}{2}}\big\} = \frac{\alpha}{2}$; $Z \sim \mathcal{N}(0;1)$. Тогда, при уровне доверия:

- $(1-\alpha) = 0.90$ доверительный интервал имеет вид:

$$0.56 \approx 0.63 - 1.282 \cdot \sqrt{\frac{0.63\cdot(1-0.63)}{76}} < p < 0.63 + 1.282 \cdot \sqrt{\frac{0.63\cdot(1-0.63)}{76}} \approx 0.70$$

- $(1-\alpha) = 0.95$ доверительный интервал имеет вид:

$$0.54 \approx 0.63 - 1.645 \cdot \sqrt{\frac{0.63\cdot(1-0.63)}{76}} < p < 0.63 + 1.645 \cdot \sqrt{\frac{0.63\cdot(1-0.63)}{76}} \approx 0.72$$

Пункт 10

Для проверки гипотез в качестве эталонного уровня значимости возьмём $\alpha = 0.05$.

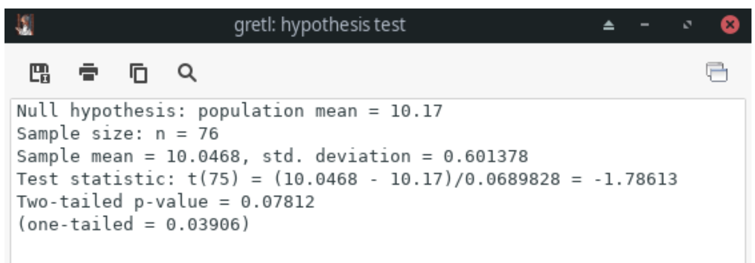
Основная гипотеза $H_0: \mu = 10.17$. Двусторонняя альтернативная гипотеза $H_A: \mu \neq 10.17$. Правосторонняя альтернативная гипотеза: $H_A: \mu > 10.17$.

Поскольку дисперсия неизвестна, в качестве критической статистики используется t-распределение:

$$t = \frac{\overline{X} - \mu_0}{{}^{\hat{\sigma}}/_{\sqrt{n}}} \overset{H_0}{\sim} t_{n-1}$$

Для проверки гипотезы вычислим расчётное (оно же критическое для уровня значимости, равного p-value) значение, критические значения на уровне значимости $\alpha = 0.05$ и p-значения для двусторонней и правосторонней альтернативных гипотез.

$$\begin{cases} t = \dfrac{10.05 - 10.17}{{}^{0.6}/_{76}} \approx -1.78613 \\ t_{n-1,\frac{\alpha}{2}} = 1.9921 \\ t_{n-1,\alpha} = 1.6654 \\ p{-}value = P\Big\{\big|t_{n-1}\big| > |t|\Big\} \approx 0.07812 & \text{для двусторонней альтернативной гипотезы} \\ p{-}value = P\Big\{t_{n-1} > t\Big\} \approx 0.96094 & \text{для правосторонней альтернативной гипотезы} \end{cases}$$




```
Null hypothesis: population mean = 10.17
Sample size: n = 76
Sample mean = 10.0468, std. deviation = 0.601378
Test statistic: t(75) = (10.0468 - 10.17)/0.0689828 = -1.78613
Two-tailed p-value = 0.07812
(one-tailed = 0.03906)
```

Рис. 7: Проверка гипотезы о равенстве матожидания lnwage значению 10.17.

Двусторонняя альтернативная гипотеза принимается, если $\big|t\big| > t_{n-1,\frac{\alpha}{2}}$. Как мы видим, при уровне значимости $\alpha = 0.05$ нулевая гипотеза о равенстве матожидания 10.17 не может быть отвергнута, то есть матожидание считается равным 10.17. Минимальный уровень значимости, при котором нулевая гипотеза отвергалась бы в пользу двусторонней альтернативной, равен $p - value = 0.07812$.

Правосторонняя альтернативная гипотеза принимается, если $t > t_{n-1,\alpha}$. Как мы видим, при уровне значимости $\alpha = 0.05$ нулевая гипотеза о равенстве матожидания 10.17 не может быть отвергнута, то есть матожидание считается равным 10.17. Минимальный уровень значимости, при котором нулевая гипотеза отвергалась бы в пользу правосторонней альтернативной, равен $p{-}value = 0.96094$, то есть на любом разумном уровне значимости нулевая гипотеза не отвергается. Обратим внимание, что Gretl выдал значение $0.03906 = 1 - 0.96094$. Связано это с тем, что он выдаёт $p - value$ для односторонней гипотезы вне зависимости от стороны. Так как у нас значение большое по модулю и отрицательное, левосторонняя гипотеза принималась бы как раз при низком уровне значимости, в то время как правосторонняя, наоборот. Примечательно, что при проверки левосторонней гипотезы нулевая гипотеза отвергалась бы в пользу левосторонней альтернативной, то есть можно было бы считать, что матожидание логарифма зарплаты меньше 10.17.

Пункт 11

Для проверки гипотезы создадим в Gretl две переменные: $lnwage\_male = lnwage \times male$, в которую сохраним логарифмы зарплат мужчин ($n_1 = 28$ наблюдений), и $lnwage\_female = lnwage \times (1 - male)$, в которую сохраним логарифмы зарплат женщин ($n_2 = 48$ наблюдений). Прежде, чем проверять гипотезу о матожидании, необходимо проверить гипотезу о равенстве дисперсий, так как у нас нет таких данных, а для проверки гипотезы о среднем они нужны.

В качестве критической статистики используется F-распределение:

$$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} \sim F_{n_2-1,n_1-1}$$

Для проверки гипотезы вычислим расчётное (оно же критическое для уровня значимости, равного p-value) значение, критическое значение на уровне значимости $\alpha = 0.05$ и p-значение для двусторонней альтернативной гипотезы.

$$\begin{cases} F = \dfrac{0.43412}{0.22884} \approx 1.89705 \\ F_{n_2-1,n_1-1,\frac{\alpha}{2}} = 2.0396 \\ F_{n_2-1,n_1-1,1-\frac{\alpha}{2}} = 0.5222 \\ p-value = 2\times min\left(P\left\{F > F_{n_2-1,n_1-1,\frac{\alpha}{2}}\right\}; P\left\{F < F_{n_2-1,n_1-1,1-\frac{\alpha}{2}}\right\}\right) \approx 0.07738 \end{cases}$$

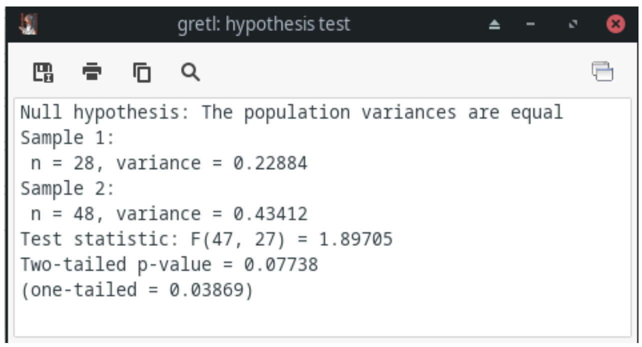


Рис. 8: Проверка гипотезы о равенстве дисперсий lnwage_male и lnwage_female

Двусторонняя альтернативная гипотеза принимается, если $F > F_{n_2-1,n_1-1,\frac{\alpha}{2}}$ или $F < F_{n_2-1,n_1-1,1-\frac{\alpha}{2}}$. Как мы видим, при уровне значимости $\alpha = 0.05$ нулевая гипотеза о равенстве дисперсий не может быть отвергнута, то есть дисперсии считаются равными. Минимальный уровень значимости, при котором нулевая гипотеза отвергалась бы в пользу двусторонней альтернативной, равен $p - value = 0.07738$.

Теперь проверим гипотезу превышении матожидания логарифма зарплаты мужчин над матожиданием логарифма зарплаты у женщин. Основная гипотеза $H_0\colon\ \mu_1 - \mu_2 = 0$. Правосторонняя альтернативная гипотеза $H_A\colon\ \mu_1 - \mu_2 > 0$.

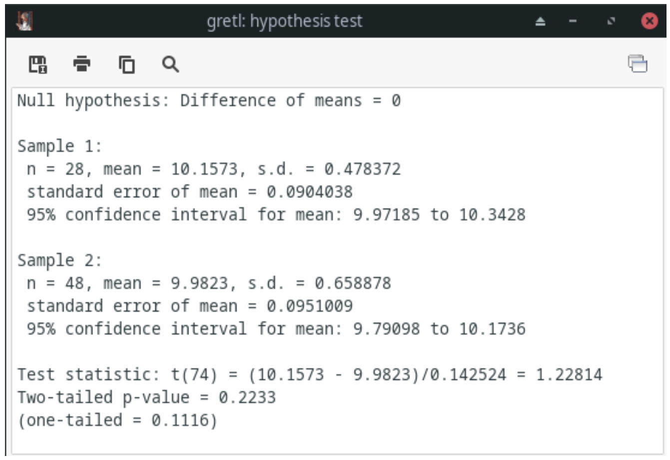
Поскольку дисперсии неизвестны, но равны, в качестве критической статистики используется t-распределение:

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\hat{\sigma}^2 \cdot \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \overset{H_0}{\sim} t_{n_1+n_2-2}$$

, где $\hat{\sigma}^2 = \dfrac{(n_1-1)\cdot\hat{\sigma}_1^2 + (n_2-1)\cdot\hat{\sigma}_2^2}{(n_1-1)+(n_2-1)}$.

Для проверки гипотезы вычислим расчётное (оно же критическое для уровня значимости, равного p-value) значение, критическое значение на уровне значимости $\alpha = 0.05$ и p-значение для правосторонней альтернативной гипотезы.

$$\begin{cases} t = \dfrac{(10.1573 - 9.9823) - 0}{0.142524} \approx 1.22814 \\ t_{n_1+n_2-2,\alpha} = 1.6657 \\ p-value = P\left\{t_{n-1} > t\right\} \approx 0.1116 \end{cases}$$



```
Null hypothesis: Difference of means = 0

Sample 1:
 n = 28, mean = 10.1573, s.d. = 0.478372
 standard error of mean = 0.0904038
 95% confidence interval for mean: 9.97185 to 10.3428

Sample 2:
 n = 48, mean = 9.9823, s.d. = 0.658878
 standard error of mean = 0.0951009
 95% confidence interval for mean: 9.79098 to 10.1736

Test statistic: t(74) = (10.1573 - 9.9823)/0.142524 = 1.22814
Two-tailed p-value = 0.2233
(one-tailed = 0.1116)
```

Рис. 9: Проверка гипотезы о равенстве матожиданий lnwage_male и lnwage_female

Правосторонняя альтернативная гипотеза принимается, если $t > t_{n_1+n_2-2,\alpha}$. Как мы видим, при уровне значимости $\alpha = 0.05$ нулевая гипотеза о равенстве матожиданий не может быть отвергнута, то есть матожидания считаются равным. Минимальный уровень значимости, при котором нулевая гипотеза отвергалась бы в пользу правосторонней альтернативной, равен $p{-}value = 0.1116$.

Пункт 12

Для проверки гипотезы создадим в Gretl две переменные: $lnwage\_internet = lnwage \times internet$, в которую сохраним логарифмы зарплат людей, пользовавшихся интернетом ($n_1 = 59$ наблюдений), и $lnwage\_no\_internet = lnwage \times (1 - internet)$, в которую сохраним логарифмы зарплат людей, не пользовавшихся интернетом ($n_2 = 17$ наблюдений).

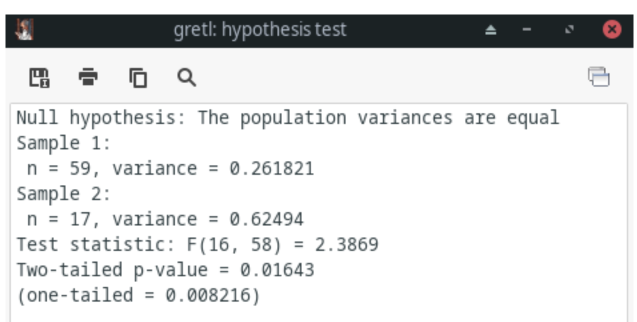
Основная гипотеза $H_0: \sigma_2^2 - \sigma_1^2 = 0$. Левосторонняя альтернативная гипотеза $H_A: \sigma_2^2 - \sigma_1^2 < 0$.

В качестве критической статистики используется F-распределение:

$$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} \sim F_{n_2-1,n_1-1}$$

Для проверки гипотезы вычислим расчётное (оно же критическое для уровня значимости, равного p-value) значение, критическое значение на уровне значимости $\alpha = 0.05$ и p-значение для левосторонней альтернативной гипотезы.

$$\begin{cases} F = \dfrac{0.62494}{0.261821} \approx 2.3869 \\ F_{n_2-1,n_1-1,1-\alpha} = 0.4742 \\ p{-}value = P\Big\{F > F_{n_2-1,n_1-1,\alpha}\Big\} \approx 0.991784 \end{cases}$$



```
Null hypothesis: The population variances are equal
Sample 1:
 n = 59, variance = 0.261821
Sample 2:
 n = 17, variance = 0.62494
Test statistic: F(16, 58) = 2.3869
Two-tailed p-value = 0.01643
(one-tailed = 0.008216)
```

Рис. 10: Проверка гипотезы о равенстве дисперсий lnwage_internet и lnwage_no_internet

Левосторонняя альтернативная гипотеза принимается, если $F < F_{n_2-1,n_1-1,1-\alpha}$. Как мы видим, при уровне значимости $\alpha = 0.05$ нулевая гипотеза о равенстве дисперсий не может быть отвергнута, то есть дисперсия логарифма зарплат людей, пользовавшихся интернетом, считается равной дисперсии логарифма зарплат людей, не пользовавшихся интернетом. Минимальный уровень значимости, при котором нулевая гипотеза отвергается в пользу левосторонней альтернативной, равен $p-value = 0.991784$. Обратим внимание, что дисперсия логарифмов зарплат людей, не пользовавшихся интернетом, в 2.3869 раз больше, чем дисперсия логарифмов зарплат людей, пользовавшихся интернетом, в связи с чем двусторонняя и правосторонняя альтернативные гипотезы принимаются на достаточно низких уровнях значимости: 0.01643 и $0.008216 = 1-0.991784$ соответственно.

Пункт 13

Для проверки гипотезы создадим в Gretl переменную $children\_bin = missing\Big(1-\dfrac{1-children}{1-children}\Big)$. Вычисления $\Big(1-\dfrac{1-children}{1-children}\Big)$ переводят единицу в missing value, а остальные числа в 0; функция missing переводит все missing value в 1, а все не missing value в ноль; получается, сначала мы переводим все единицы в missing value и остальные значения в нули, а потом заполняем пустые значения единицами, не меняя нули, то есть получаем бинарную переменную, в которой 1 – это семьи с одним ребёнком, 0 – семьи с любым другим количеством детей.

Основная гипотеза $H_0:\ p=0.5$. Односторонняя альтернативная гипотеза: $H_A:\ p>0.5$ или $H_A:\ p<0.5$.

В качестве критической статистики используется стандартное нормальное распределение:

$$Z = \frac{\hat{p}-p}{\sqrt{\dfrac{p\cdot(1-p)}{n}}} \overset{H_0}{\sim} \mathcal{N}(0;1)$$

Для проверки гипотезы вычислим расчётное (оно же критическое для уровня значимости, равного p-value) значение, критическое значение на уровне значимости $\alpha = 0.05$ и p-значение для односторонней альтернативной гипотезы.

$$\begin{cases} Z = \dfrac{0.5-0.5}{0.05} \approx 0 \\ Z_\alpha = 1.645 \\ p-value = P\Big\{F > F_{n_2-1,n_1-1,\alpha}\Big\} = 0.5 \end{cases}$$

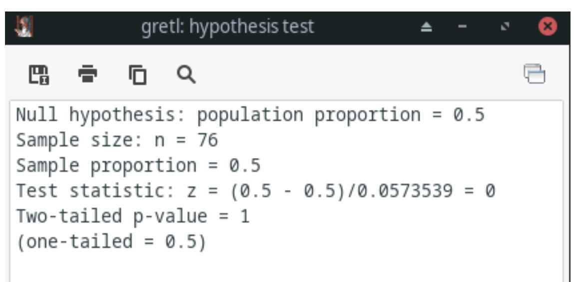
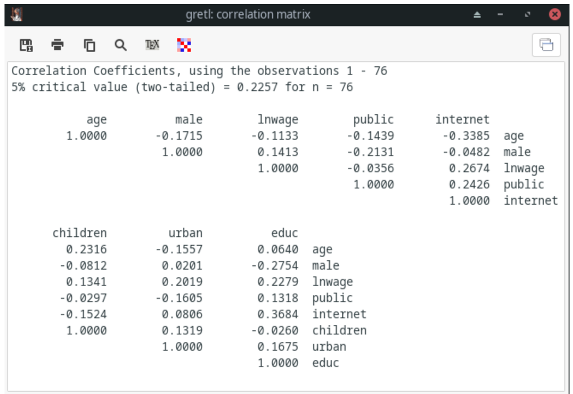


Рис. 11: Проверка гипотезы о равенстве доли children_bin 0.5

Односторонняя альтернативная гипотеза принимается, если $Z > Z_\alpha$ или $Z < -Z_\alpha$. Как мы видим, при уровне значимости $\alpha = 0.05$ нулевая гипотеза о равенстве доли 0.5 не может быть отвергнута, то есть доля семей с одним ребёнком считается равной 0.5. Минимальный уровень значимости, при котором нулевая гипотеза отвергается в пользу односторонней альтернативной, равен $p - value = 0.5$. Обратим внимание, что в данном случае сторона не имеет значение, так как выборочная доля совпадает с эталонной, что делает расчётное значение равным 0, то есть точкой симметрии стандартного нормального распределения.

Пункт 14

Построим корреляционную таблицу.



gretl: correlation matrix

Correlation Coefficients, using the observations 1 - 76
5% critical value (two-tailed) = 0.2257 for n = 76

age	male	lnwage	public	internet	
1.0000	-0.1715	-0.1133	-0.1439	-0.3385	age
	1.0000	0.1413	-0.2131	-0.0482	male
		1.0000	-0.0356	0.2674	lnwage
			1.0000	0.2426	public
				1.0000	internet

children	urban	educ	
0.2316	-0.1557	0.0640	age
-0.0812	0.0201	-0.2754	male
0.1341	0.2019	0.2279	lnwage
-0.0297	-0.1605	0.1318	public
-0.1524	0.0806	0.3684	internet
1.0000	0.1319	-0.0260	children
	1.0000	0.1675	urban
		1.0000	educ

Рис. 12: Корреляционная таблица

Как мы видим, коэффициенты корреляции имеют невысокие значения. Всего у двух пар переменных коэффициенты по модулю превышают 0.3. Первая из них пар – *interner* – *age* с коэффициентом корреляции -0.3358. Взаимосвязь может означать, что взрослые чаще могут не использовать интернет длительное время, чем молодые люди. Вторая пара – *educ* – *internet* с коэффициентом корреляции 0.3684. Данное значение коэффициента может говорить о том, что для более высоких уровней образования чаще требуется интернет (в школе предметы подробно разбираются учителями, в ПТУ мало предметов, где требуется интернет).

Пункт 15

Крайне интересно было бы, на мой взгляд, посмотреть зависимость между логарифмом зарплаты и образованием. Во-первых, обратим внимание, что среди взаимосвязей логарифма зарплаты с другими переменными связь с образованием имеет 2е по модулю величины значение коэффициента корреляции – 0.2279 (после связи с интернетом – 0.2674), причём коэффициент положителен, что говорит о наличии положительной линейной связи. Во-вторых, взглянем на boxplot логарифма зарплат, факторизованный по образованию.

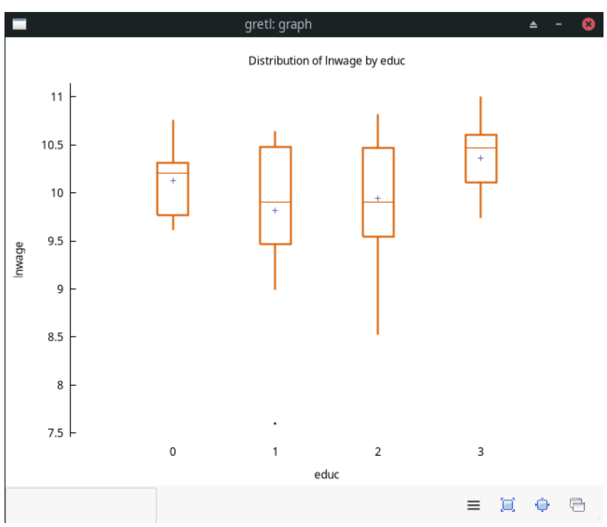


Рис. 13: Boxplot переменной lnwage, факторизованной по переменной educ

Как мы видим, границы размаха и средние смещаются в сторону увеличения с ростом уровня образования (разве что, средние у людей с нулевым образованием оказались выше, чем у людей со средним общим или средним профессиональным образованием).