Индивидуальное домашнее задание №4

Студент: Балябин А.А.
Группа: 5362
Преподаватель: Медведев А.Н.

Задание 1

Возьмем в качестве параметров для генерации выборки из нормального распределения $N(a,\sigma^2)$ $a_0 = 7.3$, $\sigma = 1$ и сгенерируем выборки длины 10, 100 и 1000.

1.1

В качестве основной гипотезы возьмем $H_0 = \{a = a_0\}$, тогда в качестве альтернативы гипотезу $H_1 = \{a \neq a_0\}$.

Определим критерий: $\delta(\vec{X}) = \begin{matrix} H_0, если |G(\vec{X})| \leqslant C \\ H_1, если |G(\vec{X})| > C \end{matrix}$

Здесь $G(\vec{X})$ — статистика отклонения, C определяется уровнем значимости α, задаваемым при построении критерия.

$C = F_{и.р.}^{-1}(1 - \alpha/2)$, то есть, квантиль уровня $1 - \alpha/2$. $F_{и.р.}$ – известная функция распределения (в задании 1 — это Φ, то есть функция стандартного нормального распределения, в случае известной дисперсии).

Можно сразу же вычислить все необходимые значения C для каждого из уровней α для случая с известной дисперсией.

α	0.25	0.1	0.05	0.01	0.001
C	1.1503	1.6449	1.9599	2.5758	3.2905

Функция G такова, что в случае, если гипотеза H_0 верна, G имеет известное распределение. Если же H_0 неверна, то есть выборка $\vec{X}$ имеет какое-то другое распределение, то $|G(\vec{X})| \xrightarrow{p} \infty$ при $n \to \infty$ для любого такого распределения.

Критерий работает следующим образом. Вычисляется функция $G(\vec{X})$ от выборки $\vec{X}$, её значение сравнивается с граничным значением C. Если величина функции отклонения меньше этого порогового значения, то принимается основная гипотеза H_0, иначе — выбор делается в пользу альтернативной гипотезы H_1.

В случае, если дисперсия <u>известна</u>, функция отклонения имеет следующий вид:

$$G(\vec{X}) = \sqrt{n}\frac{\bar{X} - a}{\sigma} \quad .$$

Если же дисперсия <u>неизвестна</u>, то $G(\vec{X}) = \sqrt{n}\frac{\bar{X} - a}{\sqrt{s^2}}$, где s^2 – дисперсия, найденная по выборке. И тогда статистика отклонения распределена по закону Стьюдента с n-1 степенями свободы.

Можно составить таблицу, отражающую изменения в зависимости от длины выборки и степени известности дисперсии. Таблица составлена для случайно сгенерированных выборок.

	N = 10		N = 100		N = 1000	
Дисперсия	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**
|G(X)|	0.255	0.422	0.363	0.354	0.254	0.259

Результатом применения критерия при a = a_0 = 7.3 (при достаточно точном подсчете среднего выборочного значения и дисперсии в случае, если она неизвестна) будет принятие гипотезы H_0 для любых уровней значимости, ибо величина статистики отклонения будет по распределению стремиться к стандартному нормальному распределению и абсолютное отклонение будет мало, меньше С (здесь под С понимается квантиль стандартного нормального распределения, если дисперсия известна, и распределения Стьюдента – иначе).
Однако, на малых выборках значения среднего могут существенно отличаться друг от друга, настолько, что при определенных значениях, результатом применения критерия может быть отвержение гипотезы H_0 в пользу H_1. Но это происходит в основном для относительно больших уровней значимости (α = 0.25, α = 0.1). Чем больше выборка, тем точнее и ближе к истинному определено среднее значение.
Для полученных значений G можно заключить, что гипотеза H_0 будет принята во всех случаях, однако, на других выборках (в основном на малых) она может быть и отвергнута. Отвержение гипотезы H_0 при данных условиях на больших выборках крайне маловероятно. Как видно в таблице, значения статистики отклонения для N=10 сильно зависят от того, известна дисперсия или нет. Изменение видно уже в первом знаке после запятой. В случае N=100 изменение уже не такое сильное и видно только во втором знаке после запятой, а на выборке N=1000 значения статистики отклонения различаются лишь в третьем знаке после запятой.

Можно рассчитать реально достигнутый уровень значимости.
Возьмем выборку $\bar{X}$, точно принадлежащую распределению *N(a,σ²)*.
Тогда $G(\vec{X})$ – функция отклонения для выборки $\bar{X}$, точно принадлежащей распределению *N(a, σ²)*, принадлежность которому выборки $\vec{x}$ проверяется.

$$\alpha^* = 1 - F_{u.p.}(|G(\vec{x})|) + F_{u.p.}(-|G(\vec{x})|) \ = \ 1 - P(\eta < |G(\vec{x})|) + P(\eta < -|G(\vec{x})|) =$$
$$= P(\eta > |G(\vec{x})|) + P(\eta < -|G(\vec{x})|) = P(|\eta| > |G(\vec{x})|) = P_{H_0}(|G(\vec{X})| > |G(\vec{x})|)$$

Чем больше α*, тем лучше это свидетельствует в пользу гипотезы H_0. То есть, тем больше вероятность, взяв выборку из распределения *N(a, σ²)*, получить по ней большее отклонение, чем у проверяемой выборки. А реально достигнутый уровень значимости, по сути, и есть вероятность получения такого результата.

	N = 10		N = 100		N = 1000	
Дисперсия	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**
α*	0.86	0.81	0.87	0.89	0.80	0.82

Значения реально достигнутого уровня значимости могут сильно колебаться в зависимости от конкретной сгенерированной выборки (например, для выборки длины 10 получались даже значения α* = 0.49). Обоснование этому дано в анализе результатов задания 3.

Конкретный результат α* можно интерпретировать так, что в среднем α* % контрольных выборок из известного распределения, удовлетворяющих основной гипотезе, будут обладать бо́льшим отклонением $|G(\vec{X})|$, чем тестируемая выборка, принадлежность которой этому распределению устанавливается, что в свою очередь означает, что проверяемая выборка ведет себя не хуже, чем α* % контрольных выборок.

1.2

Переопределим гипотезы так.

H_0: a = a_0 + ε

H_1: a ≠ a_0 + ε, где ε = 10, 0.1, 0.01

Построим таблицу значений статистики отклонений. Зеленым выделим ячейки, в которых гипотеза H_0 принимается, красным — в которых отвергается.

ε = 10	**N = 10**		**N = 100**		**N = 1000**	
Дисперсия	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**
|G(X)|	31.59	45.61	99.85	113.74	316.38	321.45
α = 0.25						
α = 0.1						
α = 0.05						
α = 0.01						
α = 0.001						

ε = 0.1	**N = 10**		**N = 100**		**N = 1000**	
Дисперсия	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**
|G(X)|	0.095	0.161	0.94	1.07	2.44	2.64
α = 0.25						
α = 0.1						
α = 0.05						
α = 0.01						
α = 0.001						

ε = 0.01	N = 10		N = 100		N = 1000	
Дисперсия	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**
|G(X)|	0.11	0.13	0.17	0.15	0.094	0.099
α = 0.25						
α = 0.1						
α = 0.05						
α = 0.01						
α = 0.001						

Стоит еще взглянуть на результаты при ε = 0.2. Здесь лучше видна зависимость значения критерия от длины выборки и уровня значимости.

ε = 0.2	N = 10		N = 100		N = 1000	
Дисперсия	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**
|G(X)|	0.51	0.88	1.92	2.14	6.16	6.05
α = 0.25						
α = 0.1						
α = 0.05						
α = 0.01						
α = 0.001						

Анализ полученных результатов:

Для ε = 10 видно, что значение статистики отклонения быстро расходится при увеличении длины выборки, что по определению соответствует тому, что гипотеза H_0 неверна. Поэтому вне зависимости от уровня значимости критерий отвергает основную гипотезу в пользу альтернативной.

Для ε = 0.1 видно, что на выборке длины 10 основная гипотеза принимается при всех заданных уровнях значимости. Чем больше значение α, тем уже доверительный интервал, в который могут попадать оцениваемые значения, и тем меньше значение C, с которым сравнивается значение статистики отклонения |G(X)|. В таблице с ε = 0.2 лучше виден этот постепенный переход.

При ε = 0.01 на данных выборках величина этого отклонения не влияет на результат применения критерия, так как она относительно мала по сравнению с истинным средним значением 7.3.

Очевидно, что чем больше ε, тем больше значение статистики отклонения и тем меньше вероятность принятия основной гипотезы H_0: $a = a_0 + ε$. Однако, даже при малых ε H_0 может быть отвергнута в пользу альтернативы, если уровень α

достаточно велик или выборка достаточно большого объема, чтобы такие отклонения могли повлиять на значение критерия.
Неизвестная дисперсия, а именно, необходимость ее вычисления, чаще всего ведет к увеличению значения статистики отклонения и иногда это меняет значение критерия. Так, например, это произошло при $\varepsilon = 0.2$ на выборке длины 100. Значение $|G(X)|$ увеличилось с 1.92 до 2.14 (хотя и единичный случай несет в себе мало полезной информации). Поэтому в первом случае H_0 принимается, а во втором уже отвергается. Однако, если длина выборки велика, то применяемая оценка дисперсии мало отличается от истинного значения в силу ее состоятельности.

Вычислим реально достигнутые уровни значимости при каждом из заданных ε = 10, 0.1 и 0.01.

ε = 10	**N = 10**		**N = 100**		**N = 1000**	
Дисперсия	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**
α*	$\to 0$	$\to 0$	$\to 0$	$\to 0$	$\to 0$	$\to 0$

ε = 0.1	**N = 10**		**N = 100**		**N = 1000**	
Дисперсия	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**
α*	0.71	0.76	0.36	0.33	0.0029	0.0032

ε = 0.01	**N = 10**		**N = 100**		**N = 1000**	
Дисперсия	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**	**Известна**	**Неизвестна**
α*	0.97	0.97	0.96	0.96	0.88	0.88

Видно, что при ε порядка a_0 происходит довольно быстрый рост значений статистики отклонения при увеличении объема выборки (что свидетельствует в пользу альтернативной гипотезы H_1: $a \neq a_0$), в следствие чего реально достигнутый уровень значимости стремится к 0. Это объясняется тем, что вероятность того, что контрольная выборка будет вести себя так же или хуже, чем тестируемая, стремится к нулю с ростом значений $G(x)$, так как она точно взята из известного распределения. При $\varepsilon = 0.1$ $|G(X)|$ тоже расходится, хотя и медленнее. Это видно по тому, как уменьшается реально достигнутый уровень значимости в зависимости от выборки. А $\varepsilon = 0.01$ дает еще меньший вклад, хотя все так же заставляет $|G(X)|$ расходиться.
В целом, изменение реально достигнутого уровня значимости при использовании различных ε зависит от того, какие значения параметров задавались при генерации. Если ε имеет тот же порядок, что и параметр, то расхождение будет существенно. Чем меньше порядок ε по сравнению с порядком параметра, тем медленнее будет происходить рост $|G(X)|$ при увеличении размера выборки.

Задание 2

Рассмотрим выборку из файла «31_norm_13.554397.csv».
В качестве основной гипотезы возьмем $H_0 = \{a = a_0\}$, тогда в качестве альтернативы гипотезу $H_1 = \{a \neq a_0\}$.
Определим критерий: $\delta(\vec{X}) = \begin{cases} H_0, если |G(\vec{X})| \leqslant C \\ H_1, если |G(\vec{X})| > C \end{cases}$
Рассчитаем среднее значение по выборке. Получим $\bar{X} = 13.622882$
Рассчитанная по выборке дисперсия (s^2 = 0.46) отличается от заданной в условии (σ^2 = 1).
Статистика отклонения определяется так же, как и в задании 1.
Вычислим значение функции отклонения при известной и неизвестной дисперсиях.

Дисперсия	**Известна**	**Неизвестна**
|G(X)|	2.1657	3.1941
α = 0.25		
α = 0.1		
α = 0.05		
α = 0.01		
α = 0.001		

Реально достигнутые уровни значимости:

Дисперсия	**Известна**	**Неизвестна**
α*	0.03	0.0014

<u>Анализ полученных результатов:</u>

Как можно заметить, реально достигнутый уровень значимости и при известной и при неизвестной дисперсиях достаточно низкий. Это можно объяснить тем, что значение реального среднего выборочного значения, отличающееся от предполагаемого на 0.1, в совокупности с достаточно большой длиной выборки N = 1000 дает такое значение статистики отклонения, что гипотеза принимается лишь при уровнях доверия α = 0.01 и α = 0.001. Однако, α* в первом случае все равно больше. Но этому также есть объяснение. В случае неизвестной дисперсии ее значение приходилось рассчитывать по выборке, в следствие чего, оно получалось равным s^2 = 0.46, что достаточно сильно отличается от заданной в условии, равной 1.

Задание 3

Проверим простую гипотезу согласия о параметрах распределения Бернулли.
Сгенерируем выборки размеров 10, 100 и 1000 из распределения Бернулли с параметром $p_0 = 0.73$.
В качестве основной гипотезы возьмем $H_0 = \{p = p_0\}$, тогда в качестве альтернативы гипотезу $H_1 = \{p \neq p_0\}$.
Для построения критерия можно воспользоваться центральной предельной теоремой. Если основная гипотеза верна, то $X_i \sim B_{p0}$. Тогда по теореме Муавра-Лапласа распределение случайной величины $\frac{n\bar{X} - np_0}{\sqrt{np_0(1-p_0)}}$, где $n\bar{X} = X_1 + \ldots + X_n$ - количество успехов в n испытаниях Бернулли, с ростом n стремится к стандартному нормальному. Возьмем эту функцию в качестве G(X).
Определим критерий: $\delta(\vec{X}) = \begin{matrix} H_0, если |G(\vec{X})| \leqslant C \\ H_1, если |G(\vec{X})| > C \end{matrix}$, где C — квантиль стандартного нормального распределения уровня 1-α/2.
Составим таблицу, отражающую результат применения критерия для выборок длины 10, 100 и 1000 и различных уровней значимости. Зеленым выделим поля, на которых критерий принял основную гипотезу H_0, красным — поля, на которых критерий отвергнул H_0 в пользу альтернативы H_1.

	N = 10	**N = 100**	**N = 1000**
nX	8	70	738
|G(X)|	0.4986	0.6757	0.5698
α = 0.25			
α = 0.1			
α = 0.05			
α = 0.01			
α = 0.001			

Рассчитаем реально достигнутый уровень значимости

	N = 10	**N = 100**	**N = 1000**
α*	0.62	0.65	0.62

Но в данном случае попались достаточно «хорошие выборки». Можно посмотреть, что произойдет с реально достигнутым уровнем значимости, если выборка будет немного «хуже».

	N = 10	N = 100	N = 1000
nX	6	77	744
|G(X)|	0.9259	0.9009	0.9298
α = 0.25			
α = 0.1			
α = 0.05			
α = 0.01			
α = 0.001			

	N = 10	N = 100	N = 1000
α*	0.35	0.36	0.34

Под словом «хуже» понимается ситуация, когда nX при генерации несколько отличается от ожидаемого значения, задаваемого параметром p. Однако, значением критерия все равно остается принятие основной гипотезы для всех данных уровней значимости при данных объемах выборок.

Анализ полученных результатов:

Распределение Бернулли — это дискретное распределение. Случайная величина, имеющая такое распределение может принимать только два значения: 0 или 1. Параметром служит вероятность успеха в одном испытании Бернулли — p.

$P(X=1) = p$

$P(X=0) = 1-p$

В этом задании проверялась гипотеза о том, что распределение, из которого была взята выборка, имело параметр p равный p_0, которое задавалось при генерации (p_0 = 7.3). При p = 7.3 (= p_0) статистика отклонения дает небольшие значения, меньшие значений C (квантили стандартного нормального распределения), с которыми происходит сравнение. Величина статистики отклонения зависит от конкретной выборки. Чем оцениваемый параметр ближе к истинному значению, тем она меньше и тем, соответственно, больше вероятность того, что контрольная выборка, взятая из распределения с истинным значением параметра, будет иметь большее отклонение, чем тестируемая. А это и есть смысл реально достигнутого уровня значимости. То есть, можно считать, что изменение реально достигнутого уровня значимости при переходе от «хороших» выборок к «менее хорошим» обосновано.

Зависимость от уровня значимости критерия и от длины выборки не наблюдается в том смысле, что независимо от уровня значимости и длины выборки значение статистики отклонения при p, совпадающим с p_0, будет минимальным из возможных, а это значит, что |G(X)| не будет расходиться на бесконечности (подтверждение отсутствия зависимости от n), а будет иметь асимптотически нормальное распределение, и поэтому даже самый строгий из имеющихся критериев (при $\alpha = 0.25$) будет принимать гипотезу H_0 (подтверждение отсутствия зависимости от α).

Задание 4

Переопределим гипотезы $H_0 = \{p = p_0+\varepsilon\}$; $H_1 = \{p \neq p_0+\varepsilon\}$.
Теперь можно применить этот же критерий для соответствующих ε = 0.2, 0.1, 0.01 (ни при каком из данных ε параметр p не становится > 1). Составим таблицы результатов работы критерия.

ε = 0.2	**N = 10**	**N = 100**	**N = 1000**
nX	6	79	720
|G(X)|	4.09	5.48	26.03
α = 0.25			
α = 0.1			
α = 0.05			
α = 0.01			
α = 0.001			

ε = 0.1	**N = 10**	**N = 100**	**N = 1000**
nX	6	68	722
|G(X)|	1.94	3.99	9.09
α = 0.25			
α = 0.1			
α = 0.05			
α = 0.01			
α = 0.001			

ε = 0.01	**N = 10**	**N = 100**	**N = 1000**
nX	6	79	725
|G(X)|	1.01	1.14	1.08
α = 0.25			
α = 0.1			
α = 0.05			
α = 0.01			
α = 0.001			

Реально достигнутые уровни значимости:

0.2	**N = 10**	**N = 100**	**N = 1000**
α*	$4.3*10^{-5} \to 0$	$4.12*10^{-9} \to 0$	$1.21*10^{-149} \to 0$

0.1	**N = 10**	**N = 100**	**N = 1000**
α*	0.053	0.0002	$1.01*10^{-20} \to 0$

0.01	**N = 10**	**N = 100**	**N = 1000**
α*	0.31	0.25	0.28

Анализ полученных результатов:

Из таблиц видно, что чем больше ε, тем быстрее растет значение статистики отклонения при увеличении длины выборки и, соответственно, уменьшается реально достигнутый уровень значимости. Это объясняется так же, как и в предыдущих заданиях.

Так же, скорость роста статистики отклонения сильно зависит от соотношения величины ε и величины p_0. Чем ближе ε к p_0, тем быстрее будет происходить рост статистики отклонения и тем меньше вероятность принятия гипотезы H_0 с ростом величины выборки, следовательно, тем меньше реально достигнутый уровень значимости. Это подтверждается приведенными в таблицах результатами симуляций. Малое значение ε по сравнению с заданным p_0 практически не оказывает влияния на результат работы критерия (по крайней мере, на данных выборках и при данных уровнях значимости).