

## Практическое занятие 13: Гипотеза о законе распределения генеральной совокупности. Критерий согласия Пирсона

**Цель занятия:** приобретение навыков проверки гипотез о законе распределения генеральной совокупности.

### Задание 1

*Варианты задания 1*

### Задание 2

*Варианты задания 2*

Рассмотрим **генеральную совокупность**, распределение которой неизвестно. Однако есть основание полагать, что она распределена по некоторому закону **Z** (чаще всего, **нормально**). Это предположение может появиться как до, так и в результате статистического исследования, когда извлечена и изучена выборка объема **n**.

Требуется на уровне значимости **α** проверить нулевую гипотезу **H<sub>0</sub>** – о том, что **генеральная совокупность распределена по закону Z** против конкурирующей гипотезы **H<sub>1</sub>** о том, что **она по нему НЕ распределена**.

### Как проверить эту гипотезу?

Выборочные данные группируются в **дискретный или интервальный вариационный ряд** с **вариантами x<sub>i</sub>** и соответствующими **частотами n<sub>i</sub>**.

Поскольку эти данные взяты из практического опыта, то выборочный вариационный ряд называют **эмпирическим рядом**, а частоты **n<sub>i</sub> – эмпирическими частотами**.

Далее строятся графики, рассчитываются выборочные характеристики (**выборочная средняя**  $\bar{x}_e$ , **выборочная дисперсия**  $\sigma_e^2$  и другие).

$x_i$	$n_i$
$x_1$	$n_1$
$x_2$	$n_2$
$x_3$	$n_3$
...	...
$x_m$	$n_m$

На основе некоторых выборочных характеристик по специальным формулам, которые зависят от проверяемого закона **Z**, строится **теоретическое распределение**, где для тех же варианты  $x_1, x_2, \dots, x_m$  рассчитываются **теоретические частоты**  $n'_1, n'_2, n'_3, \dots, n'_m$ .

**Вопрос:** значимо или незначимо различие между эмпирическими  $n_1, n_2, \dots, n_m$  и теоретическими  $n'_1, n'_2, \dots, n'_m$  частотами?

Для ответа на это вопрос рассматривают различные **статистические критерии**, которые называют **критериями согласия**, и наиболее популярный из них разработал Карл Пирсон:

$$\chi^2 = \sum \frac{(n_i - n'_i)^2}{n'_i}$$

При достаточно большом  $n$  (*объёме выборки*) распределение этой случайной величины близко к **распределению хи-квадрат** с количеством степеней свободы  $k=m-r-1$ , где  $r$  – количество оцениваемых параметров закона **Z**.

Величина  $\chi^2$  – случайная по той причине, что в разных выборках будут получаться разные, заранее непредсказуемые эмпирические частоты.

Далее строится **правосторонняя** критическая область:



Критическое значение  $\chi^2_{kp} = \chi^2_{kp}(\alpha, k)$  можно отыскать с помощью таблицы «Критические точки распределения  $\chi^2$ » (см. Приложение) или MS Excel:

A	B	C	D	E	F	G	H	I
3	<b>Нахождение критической точки распределения <math>\chi^2</math></b>							
4								
5	<i>Введите количество степеней свободы <math>k=</math></i>						<b>9</b>	<i>и значение <math>\alpha=</math></i>
6								<b>0,025</b>
7	Таким образом,							
8	<b><math>\chi^2_{kp} = \chi^2_{kp}(\alpha, k) =</math></b>					<b>19,023</b>		
9								
10								
11								

$=\text{ХИ2ОБР}(I5;G5)$

Наблюдаемое значение критерия рассчитывается по эмпирическим и найденным теоретическим частотам:

$$\chi^2_{набл} = \sum_{i=1}^m \frac{(n_i - n'_i)^2}{n'_i}$$

Если  $\chi^2_{набл} < \chi^2_{kp}$ , то на уровне значимости  $\alpha$  нет оснований отвергать гипотезу  $H_0$  о том, что генеральная совокупность распределена по закону **Z**. То есть, различие между эмпирическими и теоретическими частотами **незначимо** и обусловлено случайными факторами (случайностью самой выборки, способом группировки данных и т.д.)

Если  $\chi^2_{набл} > \chi^2_{kp}$ , то нулевую гипотезу отвергаем, иными словами эмпирические и теоретические частоты отличаются **значимо**, и это различие **вряд ли** случайно.

## Задание 1

По результатам выборочного исследования найдено распределение средних удоев молока в фермерском хозяйстве (литров) от одной коровы за день:

Литры	7,5- 10,5	10,5- 13,5	13,5- 16,5	16,5- 19,5	19,5- 22,5	22,5- 25,5	25,5- 28,5	28,5- 31,5	31,5 - 34,5
Коров	2	6	10	17	33	11	9	7	5

На уровне значимости 0,05 проверить гипотезу о том, что генеральная совокупность (*средний убой коров всей фермы*) распределена нормально. Построить гистограмму частот и теоретическую кривую.

### Методика выполнения

на уровне значимости  $\alpha$  проверим гипотезу  $H_0$  о **нормальном распределении** генеральной совокупности против конкурирующей гипотезы  $H_1$  о том, что она так НЕ распределена. Используем критерий согласия

$$\text{Пирсона } \chi^2 = \sum \frac{(n_i - n'_i)^2}{n'_i}$$

Эмпирические частоты известны из предложенного интервального ряда, и осталось найти теоретические.

Для этого нужно вычислить **выборочную среднюю**  $\bar{x}_s$  и **выборочное стандартное отклонение**  $\sigma_s$ .

Выберем в качестве *вариант*  $x_i$  середины частичных интервалов (*длина каждого интервала*  $h=3$ ) и заполним расчётную таблицу. По причине большого объёма выборки исправлением дисперсии можно пренебречь.

Интервалы	$x_i$	$n_i$	$x_i n_i$	$x_i^2 n_i$
7,5-10,5	9	2	18	162
10,5-13,5	12	6	72	864
13,5-16,5	15	10	150	2250
16,5-19,5	18	17	306	5508
19,5-22,5	21	33	693	14553
22,5-25,5	24	11	264	6336
25,5-28,5	27	9	243	6561
28,5-31,5	30	7	210	6300
31,5-34,5	33	5	165	5445
$\Sigma =$		100	2121	47979

### Выборочная средняя

$$\bar{x}_e = \frac{\sum_{i=1}^k x_i n_i}{n} = 21,21 \text{ литра}$$

### Выборочная дисперсия

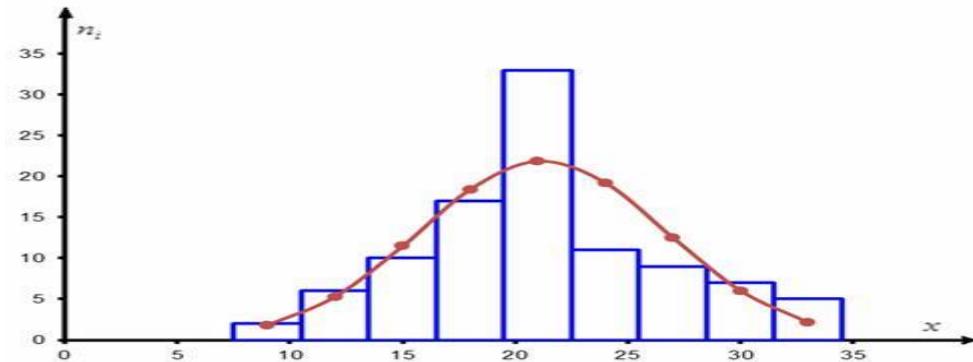
$$D_e = \frac{\sum_{i=1}^k x_i^2 n_i}{n} - (\bar{x}_e)^2 = 29,9259$$

Выборочное СКО = 5,47046 литра

Заполняем ещё одну расчётную таблицу:

A	B	C	D	E	F	
28	<b>Теоретические частоты</b>					
29	$n'_i = \frac{h \cdot n}{\sigma_e} \cdot f(z_i)$			$n=$	100	
30	$\sigma_e$			$h=$	3	
31						
32						
33	где	$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$	$z_i = \frac{x_i - \bar{x}_e}{\sigma_e}$			
34						
35						
36	=(\$B\$39-\$B\$49)/\$B\$51			=HORM.CT.PACPI(D39;0)		
37						
38	$x_i$	$n_i$	$z_i$	$f(z_i)$	$n'_i$	
39	9	2	-2,2321755	0,033033196	1,811693	
40	12	6	-1,6837294	0,096673986	5,302047	
41	15	10	-1,1352834	0,209428541	11,48603	
42	18	17	-0,5868373	0,33583761	18,41888	
43	21	33	-0,0383912	0,398648391	21,86371	
44	24	11	0,51005484	0,35028208	19,21108	
45	27	9	1,05850091	0,22783112	12,49531	
46	30	7	1,60694698	0,10969211	6,016021	
47	33	5	2,15539305	0,039093536	2,14407	
	$\bar{x}_e =$	21,21				
	$\sigma_e =$	5,47				

Построим гистограмму эмпирических частот и теоретическую кривую, которая проходит через точки  $(x_i, n_i)$



Найдём критическое значение  $\chi^2_{kp} = \chi^2_{kp}(a, k)$  критерия согласия Пирсона.

Количество степеней свободы определяется по формуле  $k=m-r-1$ , где  $m$  — количество интервалов ( $m=9$ ), а  $r$  — количество оцениваемых параметров рассматриваемого закона распределения. У нормального закона оценивается  $r=2$  параметра.

**Пояснение:**  $\bar{x}_e$  — это оценка неизвестного генерального матоожидания, а  $\sigma_e$  — это оценка неизвестного генерального стандартного отклонения, итого два оцениваемых параметра.

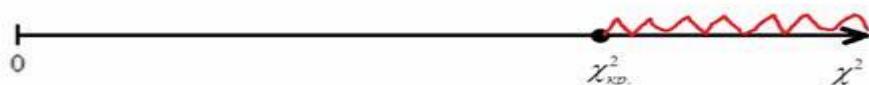
Таким образом,  $k=9-2-1=6$  и для уровня значимости  $a=0,05$ :

A	B	C	D	E	F	G	H	I
3	<b>Нахождение критической точки распределения <math>\chi^2</math></b>							
4								
5	<i>Ведите количество степеней свободы <math>k=</math></i>				6	<i>и значение <math>a=</math></i>	<b>0,05</b>	
6								
7	Таким образом,							
8	$\chi^2_{kp} = \chi^2_{kp}(a, k) =$				12,592			
9								
10								
11								

$=\text{ХИ2ОБР}(I5;G5)$

$$\chi^2_{kp} = \chi^2_{kp}(a, k) = \chi^2_{kp}(0,05, 6) \approx 12,592.$$

При  $\chi^2_{набл} > \chi^2_{kp}$  нулевая гипотеза отвергается, а при  $\chi^2_{набл} < \chi^2_{kp}$  таких оснований нет:



Вычислим наблюдаемое значение критерия  $\chi^2_{\text{набл}} = \sum_{i=1}^m \frac{(n_i - n'_i)^2}{n'_i}$ , и для этого удобно добавить в имеющуюся расчётную таблицу ещё один столбец:

A	B	C	D	E	F	G	H
28	<b>Теоретические частоты</b>						
29	$n'_i = \frac{h \cdot n}{\sigma_e} \cdot f(z_i)$			$n =$	100		
30				$h =$	3		
31							
32							
33	где	$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$	$z_i = \frac{x_i - \bar{x}_e}{\sigma_e}$				
34							
35							
36	$= (B39 - B$49) / B$51$			$= \text{НОРМ.СТ.РАСП}(D39; 0)$			
37							
38	$x_i$	$n_i$	$z_i$	$f(z_i)$	$n'_i$	$(n_i - n'_i)^2$	
39	9	2	-2,2321755	0,033033196	1,811693	$n'_i$	
40	12	6	-1,6837294	0,096673986	5,302047	0,019572662	
41	15	10	-1,1352834	0,209428541	11,48603	0,091877492	
42	18	17	-0,5868373	0,33583761	18,41888	0,192257383	
43	21	33	-0,0383912	0,398648391	21,86371	0,10930226	
44	24	11	0,51005484	0,35028208	19,21108	5,672268543	
45	27	9	1,05850091	0,22783112	12,49531	3,509530629	
46	30	7	1,60694698	0,10969211	6,016021	0,977741358	
47	33	5	2,15539305	0,039093536	2,14407	0,160939497	
						3,804138719	
	$\bar{x}_e =$	21,21			$= F\$30 * F\$29 / B\$51 * E39$		
	$\sigma_e =$	5,47				$\chi^2_{\text{набл}} = \sum_{i=1}^m \frac{(n_i - n'_i)^2}{n'_i} =$	14,53763

Ответ: поскольку  $\chi^2_{\text{набл}} \approx 14,54 > \chi^2_{\text{кр}} \approx 12,592$ , на уровне значимости 0,05 гипотезу  $H_0$  о нормальном распределении генеральной совокупности отвергаем.

Иными словами различие между эмпирическими и теоретическими частотами статистически **значимо** и вряд ли объяснимо случайными факторами.

В чём может быть **причина**? Ведь по **теореме Ляпунова** большинство коров не оказывают практически никакого влияния на убой других коров, и поэтому *распределение ген. совокупности должно быть близко к нормальному*.

Причины могут быть разными. Например, неоднородный состав совокупности (коровы разной породы), или на ферме есть VIP-хлев, где коровы получают улучшенное питание. А может быть, некоторые коровы больны и как раз оказывают существенное влияние на остальных, в связи с чем нарушается условие теоремы Ляпунова.

Интересно отметить, что при уменьшении уровня значимости до 0,01 критическое значение  $\chi^2_{кр} = \chi^2_{кр}(0,01; 6) \approx 16,81$  и  $\chi^2_{набл} \approx 12,592 < \chi^2_{кр}$  гипотеза о нормальном распределении уже принимается.

A	B	C	D	E	F	G	H	I
3	<b>Нахождение критической точки распределения <math>\chi^2</math></b>							
4								
5	<i>Введите количество степеней свободы <math>k =</math></i>				6	<i>и значение <math>\alpha =</math></i>	<b>0,01</b>	
6								
7	Таким образом,							
8	$\chi^2_{кр} = \chi^2_{кр}(\alpha, k) =$				16,812			
9								
10						=ХИ2ОБР(I5;G5)		
11								

Однако не нужно забывать, что здесь выросла  $\beta$  - вероятность того, что принята неправильная гипотеза.

И, конечно, в случае сомнений имеет смысл увеличить объём выборки, чтобы провести повторное исследование.

Постановка задачи в другой форме - *на основании исследования выборки выдвинуть гипотезу о законе распределения генеральной совокупности.*

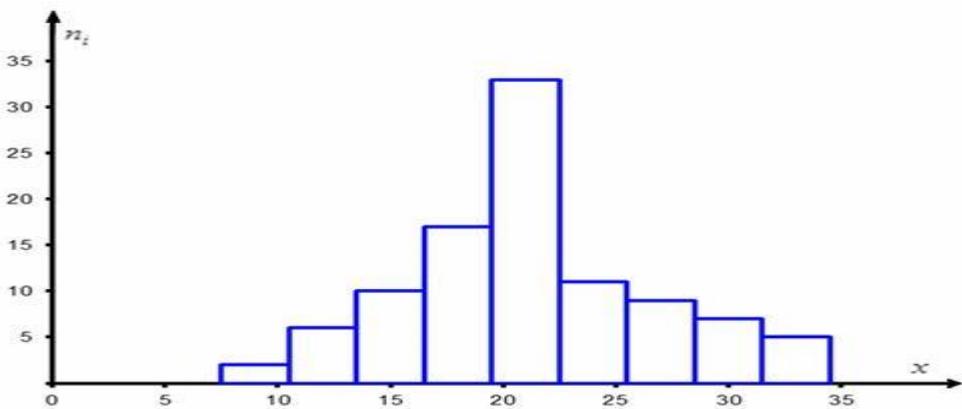
То есть, здесь не говорится о том, что предполагаемый закон нормальный (или какой-то другой) – этот вопрос предлагается проанализировать самостоятельно.

Каким образом это можно сделать?

**Во-первых, гипотезу можно выдвинуть априорно**, даже не исследуя выборку, и зависеть она будет от содержания задачи. Так, для коров используем упомянутую выше теорему Ляпунова: если каждый объект совокупности оказывается несущественное влияние на всю совокупность, то её распределение близко к **нормальному**. Если речь идёт о погрешностях округления, то распределены они обычно **равномерно**. Если распадаются радиоактивные изотопы, то, скорее всего, по **экспоненциальному закону**. И так далее.

Но по условию, требуют опираться на выборочные данные, и здесь есть сразу несколько признаков, чтобы «вычислить» этот закон.

Самый простой и наглядный способ – графический. Интервальный вариационный ряд чаще всего изображают **гистограммой**:



Построенная гистограмма по форме напоминает колоколообразный **график плотности нормального распределения**, и это является веской причиной предположить, что генеральная совокупность распределена нормально. Да, здесь есть слишком высокий средний столбик, но, возможно, это просто случайность выборки.

Если столбики примерно одинаковы по высоте, то предполагаем, что генеральная совокупность распределена **равномерно**.

Для **показательного распределения** тоже будет своя, характерная гистограмма.

#### Аналитические признаки нормального распределения:

- 1) У нормального распределения **математическое ожидание** совпадает с **модой** и **медианой**. В нашем случае соответствующие выборочные показатели весьма близки друг к другу (матожидание оценивается **выборочной средней**):  $\bar{x}_e = 21,21$ ,  $M_o \approx 20,76$ ,  $m_e \approx 20,86$ .
- 2) Выполнение **правила «трёх» сигм**. Практически все значения нормальной случайной величины находятся в интервале  $(a-3\sigma; a+3\sigma)$ . Найдём этот интервал для нашей выборки. Матожидание « $a$ » оценивается выборочной средней  $\bar{x}_e = 21,21$ , а стандартное отклонение «сигма» – **выборочным стандартным отклонением**  $\sigma_e \approx 5,47$ . Таким образом, наш эмпирический интервал:  $(21,21-3*5,47; 21,21+3*5,47)$   $(4,8; 37,62)$  – и в него действительно попадают все коровы!
- 3) **Коэффициенты асимметрии и эксцесса** нормального распределения равны нулю. В нашем случае эти характеристики не сказать что сильно, но довольно близки к нулю:  $A_s \approx 0,223$ ,  $E_k \approx -0,17$ .

На практике в исследование желательно включить все пункты за исключением, возможно, третьего (т.к. асимметрию и эксцесс рассчитывают далеко не всегда).

Следует отметить, что перечисленные выше предпосылки ещё не означают, что распределение нормально или то, что соответствующая гипотеза будет принята.

## Варианты задания 1

На уровне значимости 0,05 проверить гипотезу о том, что генеральная совокупность (средний удой коров всей фермы) распределена нормально															
№ вар.	Надои, л							№ вар.	Надои, л						
	7,5- 10,5	10,5- 13,5	13,5- 16,5	16,5- 19,5	19,5- 22,5	22,5- 25,5	25,5- 28,5		7,5- 10,5	10,5- 13,5	13,5- 16,5	16,5- 19,5	19,5- 22,5	22,5- 25,5	25,5- 28,5
	Число коров, $n_i$ , шт.														Число коров, $n_i$ , шт.
1	8	13	20	20	20	8	8	14	5	8	22	28	10	13	5
2	8	10	13	15	10	11	7	15	6	11	22	23	21	12	5
3	6	15	29	12	29	13	5	16	6	8	16	29	27	12	7
4	7	8	18	23	29	9	6	17	8	13	24	26	24	13	7
5	6	11	17	24	20	13	5	18	7	10	10	21	11	12	9
6	6	13	24	11	19	11	7	19	5	10	12	17	13	15	5
7	6	13	18	26	14	14	8	20	7	11	18	29	12	8	8
8	5	14	26	12	21	11	7	21	8	11	18	18	16	10	9
9	6	10	11	19	24	15	7	22	8	10	24	30	25	10	7
10	8	14	23	19	28	11	5	23	5	15	16	14	23	8	6
11	7	10	12	22	16	10	7	24	8	11	28	22	24	14	9
12	5	12	30	26	15	8	6	25	9	15	28	20	30	10	8
13	6	12	16	30	14	9	7	26	8	11	16	30	29	11	8

## Задание 2

В результате проверки  $n$  (500) контейнеров со стеклянными изделиями установлено, что число повреждённых изделий  $X$  имеет следующее эмпирическое распределение:

$x_i$	0	1	2	3	4	5	$n$
$n_i$	270	166	49	10	3	2	500

( $x_i$  – количество повреждённых изделий в контейнере,  $n_i$  – количество контейнеров)

С помощью критерия согласия Пирсона на уровне значимости 0,05 проверить гипотезу о том, что случайная величина  $X$  – число повреждённых изделий распределена по закону Пуассона.

## Методика выполнения

### Алгоритм решения.

- 1) Находим выборочную среднюю  $\bar{x}_\epsilon$ . Это значение будет оценкой параметра «лямбда» теоретического распределения  $p(i) \approx \frac{\lambda^i}{i!} e^{-\lambda}$ .
- 2) Находим значения  $p(i) \approx \frac{\bar{x}_\epsilon^i}{i!} e^{-\bar{x}_\epsilon}$  для  $i=0, 1, 2, 3, 4, 5$ . Вычисления можно проводить на обычном калькуляторе, но удобнее использовать MS Excel (функцию ПУАССОН).
- 3) Находим теоретические частоты  $n'_i = p(i)n$ .
- 4) Находим критическое значение  $\chi^2_{kp} = \chi^2_{kp}(a, k)$  критерия согласия Пирсона, где  $k=m-r-1$ . В данной задаче  $m=6$ . Оценивается один параметр («лямбда»), поэтому  $r=1$ .
- 5) Рассчитываем наблюдаемое значение критерия  $\chi^2_{набл} = \sum_{i=1}^m \frac{(n_i - n'_i)^2}{n'_i}$  и делаем вывод.

**Решение:** проверим гипотезу  $H_0$  о том, что генеральная совокупность распределена по закону Пуассона.

Вычисления сведём в таблицу:

A	B	C	D	E	F	G	
2	$x_i$	$n_i$	$x_i n_i$	$p_i$	$n'_i$	$\frac{(n_i - n'_i)^2}{n'_i}$	
3	0	270	0	0,5315	265,76	0,067522517	
4	1	166	166	0,3359	167,96	0,022935905	
5	2	49	98	0,1062	53,076	0,31305232	
6	3	10	30	0,0224	11,181	0,124822314	
7	4	3	12	0,0035	1,7667	0,861018945	
8	5	2	10	0,0004	0,2233	14,13596052	
9	$\Sigma =$	500	316				
10						$=\$D\$11^B3*EXP(-\$D\$11)/ФАКТР(B3)$	
11	$\bar{x}_\epsilon =$	<b>0,632</b>					
12				$\chi^2_{набл} = \sum_{i=1}^m \frac{(n_i - n'_i)^2}{n'_i} =$	<b>15,5253</b>		
13							
14							
15	$n'_i = p(i) \cdot n$						
16							
17							
18							
19	$p(i) \approx \frac{\bar{x}_\epsilon^i}{i!} e^{-\bar{x}_\epsilon}$						

Находим критическое значение для уровня значимости  $\alpha=0,05$  и количества степеней свободы  $k=m-r-1=6-1-1=4$ :

A	B	C	D	E	F	G	H	I
3	<b>Нахождение критической точки распределения <math>\chi^2</math></b>							
4								
5	Ведите количество степеней свободы $k=$				4	и значение $\alpha=$		
6								
7	Таким образом,							
8	$\chi^2_{kp} = \chi^2_{kp}(\alpha, k) =$				9,4877			
9								
10	$=\text{ХИ2ОБР}(I5;G5)$							
11								

Таким образом,  $\chi^2_{kp} = \chi^2_{kp}(0,05, 4) = 9,4877 < \chi^2_{\text{набл}} = 15,5253$ , поэтому на уровне значимости гипотезу  $H_0$  о том, что генеральная совокупность распределена по закону Пуассона отвергаем.

### Варианты задания 2

С помощью критерия согласия Пирсона на уровне значимости 0,05 проверить гипотезу: случайная величина  $X$  (число повреждённых изделий) распределена по закону Пуассона

№ вар.	К-во поврежд.изд. в контейнере $x_i$ , шт.							№ вар.	К-во поврежд.изд. в контейнере $x_i$ , шт.						
	0	1	2	3	4	5	6		0	1	2	3	4	5	6
	Число контейнеров $n_i$ , шт.								Число контейнеров $n_i$ , шт.						
1	170	66	35	23	10	7	8	14	83	74	39	23	17	10	6
2	193	96	47	31	12	9	3	15	125	73	59	28	12	7	6
3	213	69	46	32	11	8	6	16	108	70	52	26	20	8	2
4	95	69	47	23	15	10	3	17	81	83	41	24	20	10	2
5	143	91	57	22	16	7	3	18	216	80	46	27	16	7	6
6	179	79	36	32	16	9	5	19	247	80	51	29	13	10	7
7	142	80	36	22	19	10	7	20	150	88	43	22	18	9	7
8	121	88	50	21	13	9	8	21	181	84	41	28	10	10	3
9	217	95	43	33	17	10	7	22	221	89	35	24	20	7	5
10	135	95	51	21	20	8	4	23	146	87	51	33	16	7	2
11	117	78	51	28	16	10	3	24	137	98	52	24	13	8	5
12	105	97	57	22	19	8	6	25	247	99	34	31	20	8	5
13	93	67	55	26	19	7	3	26	254	74	53	23	15	8	8