Лабораторная работа 1. Статистическое описание данных.

Задание 1. Имеются следующие данные о 20 разговорах по телефону в минутах.

11	29	6	33	14	21	18	17	22	38
31	22	27	19	22	26	23	39	34	27

Построить вариационный ряд и функцию интегрального процента.

Указание. Пусть эксперимент состоит из серии независимых испытаний, которые проводятся в одних и тех же условиях. Совокупность наблюдаемых значений $x_1, \ldots, x_n$ называется *выборкой объема n.*

Вариационный ряд представляют собой упорядоченный ряд, состоящий из элементов выборки, т.е.

$$x_{(1)} \le \ldots \le x_{(n)}$$

Скопируем данные на лист MS Excel в один столбец. С помощью инструмента «*Сортировка от А до Я*» на вкладке «*Данные*», упорядочим полученный ряд чисел в порядке от меньшего к большему, т.е. построим вариационный ряд.

Интегральный процент это функция, для которой выполняется следующее соотношение:

$$Интегральный\%(x_{(k)}) = \frac{k}{n},$$

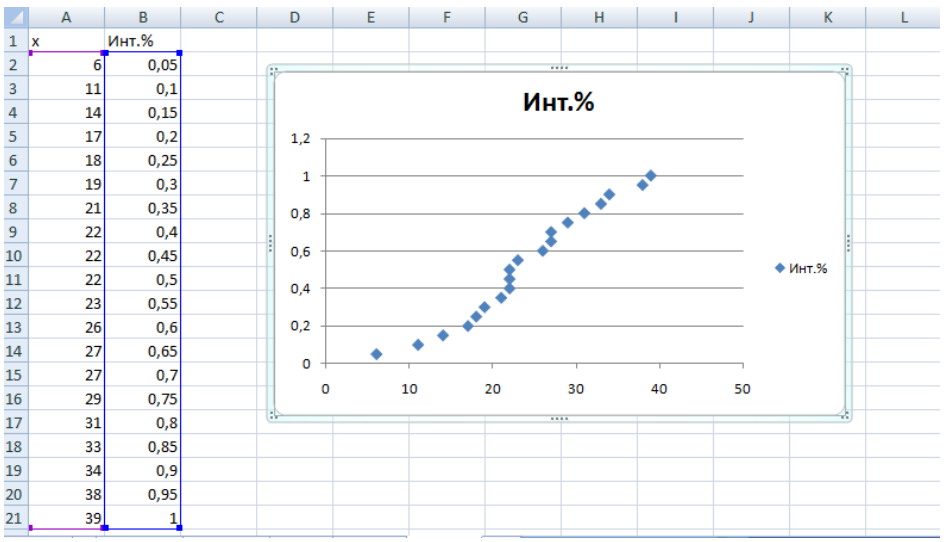
где $x_{(k)}$ – элемент *вариационного ряда* порядка *k.* Функция интегрального процента отражает характер изменения накопленных частот, т.е. долю элементов выборки не превышающих значение в точке *x.* Для данных непрерывного типа функция интегрального процента будет изменяться с шагом $1/n$. В рассматриваемом примере объем построенной выборки n = 20 и значения функции интегрального процента образуют арифметическую прогрессию вида: 0,05; 0,10; 0,15 …1,00. Запишем эти значения в соседнем столбце (см.рисунок ниже).

В MS Excel заполнение ячеек членами арифметической прогрессии можно осуществить следующим способом:

- необходимо ввести первые два числа
- выделить заполненные ячейки
- установить указатель мыши на маркере заполнения выделенного диапазона и протащить его вниз до тех пор, пока не получится нужный ряд чисел.

Для дальнейшего удобства, вставим пустую строку перед данными и присвоим столбцам соответствующие названия (см. на рисунке).

Построим график функции интегрального процента для полученной выборки. Для функции интегрального процента рекомендуется выбрать тип диаграммы «*Точечная*» (вкладка «Вставка», группа «Диаграммы»)



Задание 2. Сгруппируйте данные по интервалам и постройте гистограмму выборки. Постройте интегральный процент по интервальному распределению частот и сравните его с интегральным процентом, построенным по исходным данным в задании 1.

Указание. Для *построения гистограммы* необходимо найти размах выборки $R = x_{(n)} - x_{(1)}$ и разбить промежуток $[x_{(1)}, x_{(n)}]$ на N интервалов равной длины $l \approx R/N$, где $N \approx 1{,}72\sqrt[3]{n}$ или $N \approx \log_2 n$.

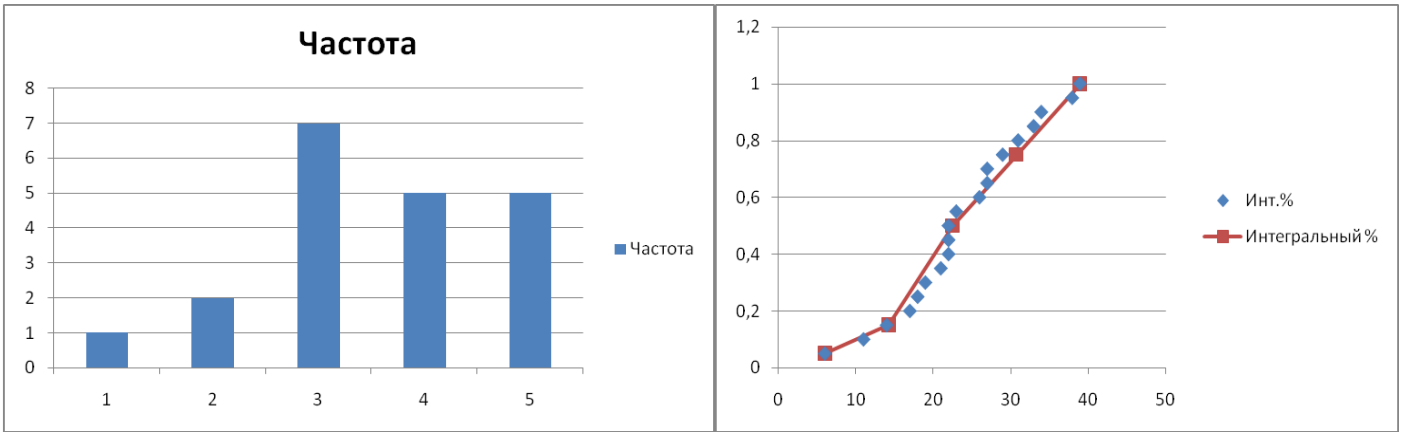
В нашем случае, размах выборки R= 39-6 = 33, возьмем N=4, тогда l=33/4=8,25.

Для нахождения частот интервального статистического ряда можно использовать инструмент «*Гистограмма*» из пакета «*Анализ данных*», предварительно построив *интервал карманов*. *Интервал карманов* – это диапазон ячеек, в котором указаны границы наших промежутков.

	A	B	C	D	E	F	G	H	I
1	x	Инт.%		R=	33		*Карман*	*Частота*	*Интегральный %*
2	6	0,05		N=	4		6	1	5,00%
3	11	0,1		l=	8,25		14,25	2	15,00%
4	14	0,15		Границы			22,5	7	50,00%
5	17	0,2		t0=	6		30,75	5	75,00%
6	18	0,25		t1=	14,25		39	5	100,00%
7	19	0,3		t2=	22,5		Еще	0	100,00%
8	21	0,35		t3=	30,75				
9	22	0,4		t4=	39				
10	22	0,45							
11	22	0,5							
12	23	0,55							

Результат применения инструмента «*Гистограмма*» будет иметь вид таблицы, расположенной в ячейках G1:I7 на рисунке. В столбце *Частота* вычисляется число попаданий в интервал, верхняя граница которого определяется значением указанным в столбце *Карман,* а нижняя граница – предыдущим по порядку значением (если такое существует). По представленной таблице можно выяснить, что в полученной выборке 1 значение не превосходит 6, 2 значения попали в диапазон от 6 до 14,25 и т.д. Столбец *Интегральный %* будет выведен только в случае, если был установлен флажок напротив пункта «Интегральный процент» в диалоговом окне «Гистограмма». В этом столбце вычисляются относительные накопленные частоты, т.е. число данных не превосходящих значение, указанное в столбце *Карман.* Так 5% данных принимают значения не превосходящие 6, 15% - не превосходят значение 14,25 и т.д.

Постройте гистограмму частот выборки по полученным данным, указав тип диаграммы «*Гистограмма*» (см.рисунок ниже). Для сравнения функции интегрального процента, построенного по сгруппированным данным, с функцией интегрального процента, построенной по исходным данным добавьте новые данные на ранее построенный график. Тип диаграммы для функции интегрального процента по сгруппированным данным укажите «*Точечная с прямыми отрезками и маркерами*» (см.рисунок)



Задание 3. Постройте гистограмму выборки и функцию интегрального процента, разбив выборку на $N=10$ интервалов. Сравните с результатами задания 2.

Задание 4. Найти выборочное среднее, медиану, квартили распределения, показатели вариации.

Указание. *Выборочное среднее* $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ является показателем общей тенденции и может быть найдено с помощью следующей статистической функции **СРЗНАЧ(число1**; число2; ...)

Медиану выборки можно найти с помощью статистической функции **МЕДИАНА(число1;число2;...)** или **КВАРТИЛЬ(массив;часть).** Параметр **часть** определяет значение, которое возвращает функция КВАРТИЛЬ (см.таблицу). Квартили выборки находятся только по данным вариационного ряда и разбивают выборку на четыре интервала, на каждый из которых приходится по 25% наблюдаемых значений.

Если часть равна	КВАРТИЛЬ возвращает
0	Минимальное значение
1	Первую квартиль (25-ю персентиль)
2	Значение медианы (50-ю персентиль)
3	Третью квартиль (75-ю персентиль)
4	Максимальное значение

Убедитесь, что указанные характеристики будут принимать следующие значения:

- выборочное среднее $\bar{x} = 23{,}95$;
- медиана Me=22,5;
- нижняя квартили $z_{1/4}$= 18,75;
- верхняя квартиль $z_{3/4}$ =29,5

Вариативность выборки можно оценить с помощью

$$\textit{выборочной дисперсии}\ s^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 \ \text{или}$$

$$\textit{несмещенной}\ \text{выборочной дисперсии}\ s_0^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = s^2\frac{n}{n-1},$$

которые можно найти с помощью следующих статистических функций

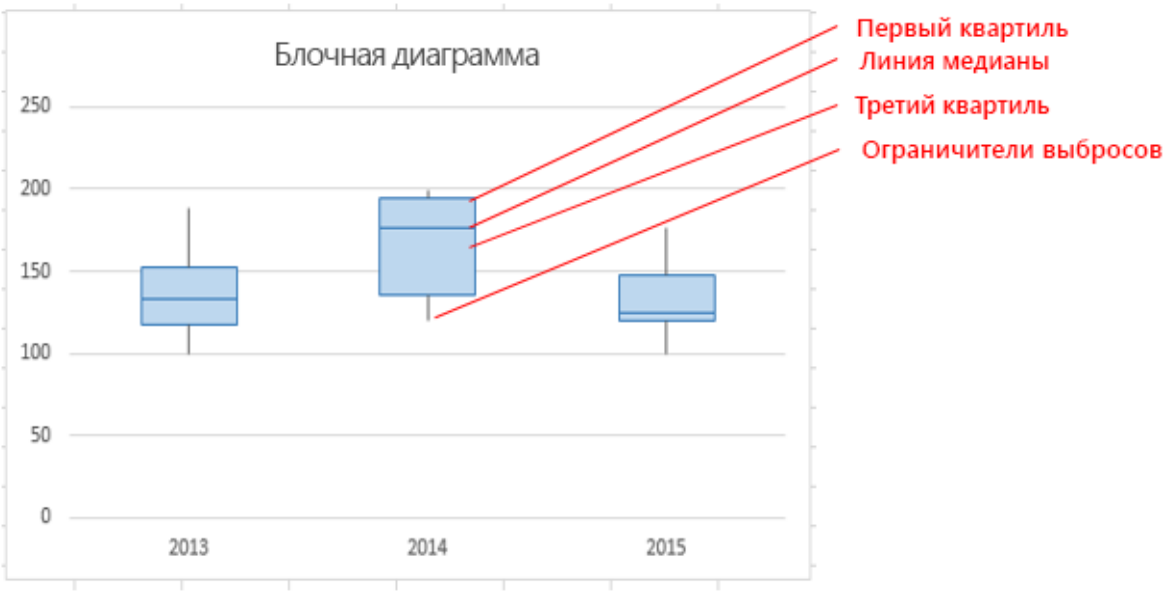
ДИСП.Г(число1;число2; ...) и ДИСП(число1;число2; ...)

Убедитесь, что выборочная дисперсия $s^2 = 71{,}1475$ и несмещенная выборочная дисперсия $s_0^2 = 74{,}8911$.

Задание 5. Постройте гистограмму выборки и функцию интегрального процента, разбив выборку на интервалы со следующими границами: $\bar{x}-3s$, $\bar{x}-2s$, $\bar{x}-s$, $\bar{x}$, $\bar{x}+s$, $\bar{x}+2s$, $\bar{x}+3s$, где $s=\sqrt{s^2}$ - среднеквадратическое отклонение (стандартное отклонение). Сравните с предыдущими результатами.

Задание 6. Построить коробковую диаграмму (box plot, ящик с усами).

Указание. Коробковая (блочная) диаграмма позволяет наглядно показать распределение данных полученных в ходе статистического исследования. На такой диаграмме статистические данные разделены на квартили, а между первым и третьим квартилем находится прямоугольник с дополнительной линией, обозначающей медиану (второй квартиль). На некоторых блочных диаграммах минимальные и максимальные значения, которые выходят за пределы первого и третьего квартилей, представлены в виде линий, которые часто называют *усами* (см.рис.[1], найдите на рисунке ошибку).

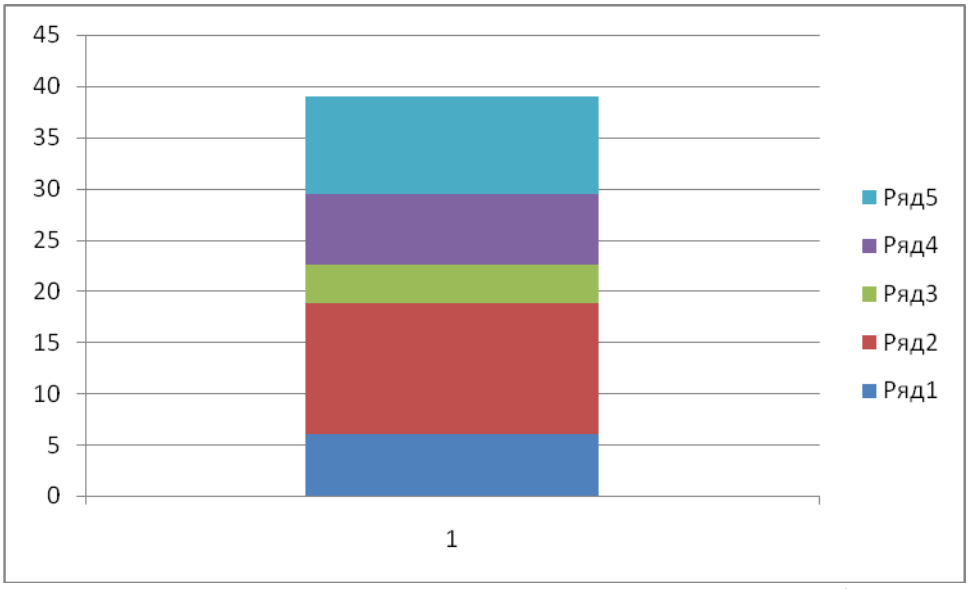


Продемонстрируем один из вариантов построения коробковой диаграммы в MS Excel[2]. Для построения такой диаграммы нам понадобятся характеристики выборки (см.таблицу ниже), которые можно найти с помощью функции КВАРТИЛЬ или взять из задания 4. Затем нужно вычислить разницу между полученными значениями

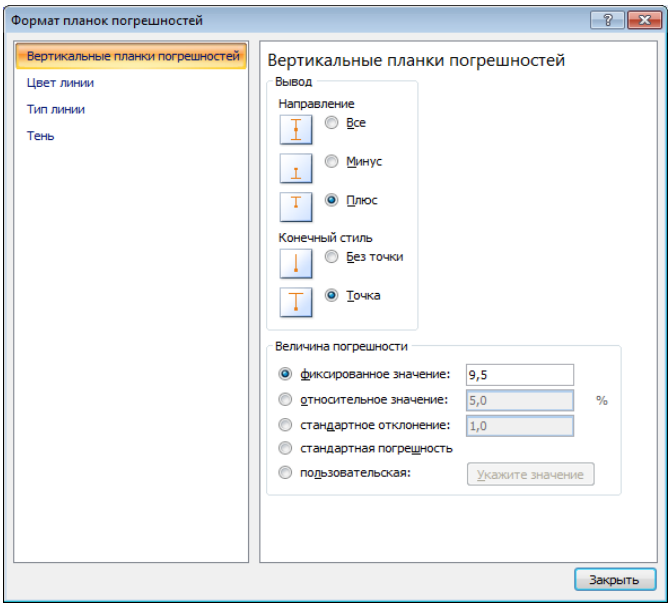
Если часть равна	КВАРТИЛЬ возвращает	Значение	Разница	Примечание
(1)	(2)	(3)	(4)	(5)
0	Минимальное значение	6	=6	
1	Первая квартиль	18,75	=18,75-6=12,75	длина нижнего уса
2	Медиана	22,5	=22,5-18,75=3,75	высота первой части коробки от нижнего квартиля до медианы
3	Третья квартиль	29,5	=29,5-22,5=7	высота второй части коробки от медианы до третьего квартиля
4	Максимальное значение	39	=39-29,5=9,5	высота верхнего уса

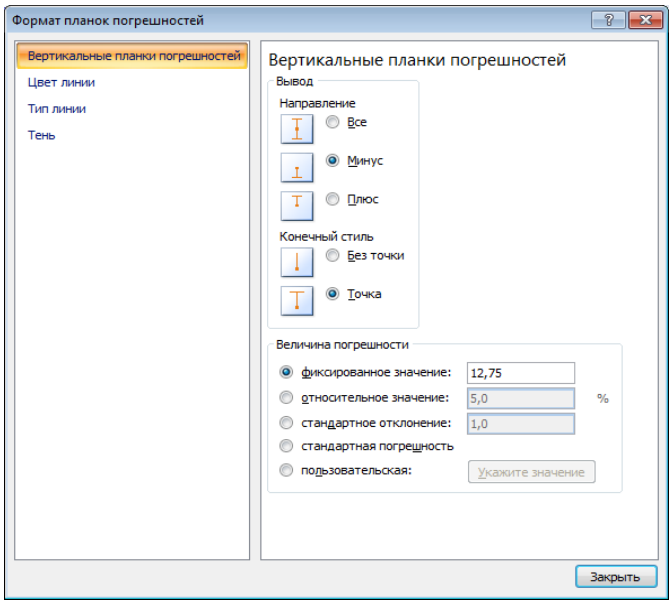
По полученным значениям в столбце (4) Разница построим *гистограмму с накоплением*. Так как Excel по умолчанию рисует столбцы с накоплением на основе наборов данных по горизонтали, а не по вертикали, необходимо поменять местами *оси диаграммы*, нажав на кнопку *Строка/столбец* на вкладке *Конструктор/Данные* или через контекстное меню *Выбрать данные*. Диаграмма должна выглядеть как представлено на рисунке ниже.

[1] URL: https://goo.gl/zvHpfW [дата обращения 07.03.2017]
[2] В MS Excel 2016 данная диаграмма включена в набор стандартных диаграмм.



Далее преобразуем *Макет* диаграммы, добавив *Планки погрешностей* (Пределы погрешностей), которые будут изображать *усы коробки*. Для добавления верхнего уса выберите на диаграмме *ряд 4*, который соответствует второй части коробки от медианы до верхнего квартиля, откройте *Планки погрешностей/Дополнительные параметры …* и установите в области *Формат* параметры, указанные на рисунке ниже (слева). Величину погрешности нужно установить равной длине уса. Для добавления нижнего уса нужно выбрать *ряд 2* и установить величину погрешности равной длине нижнего уса (см.рисунок справа)





Убедитесь, что диаграмма примет вид представленный на рисунке ниже (слева). Остается скрыть отображение рядов 1, 2 и 5, отменив *заливку фигуры* соответствующих частей диаграммы на вкладке *Формат*. Завершить оформления коробки, выбрав один цвет заливки для рядов 3 и 4 и настроив *контур фигуры*. Убедитесь, что диаграмма примет вид представленный на рисунке ниже (справа). Точкой внутри коробки отмечено значение выборочного среднего, которое добавлено как новый ряд 6 и тип диаграммы ряда изменен на *Точечная с маркерами*.

Задание 7. Смоделировать выборку объема $n=50$ с помощью функции

СЛУЧМЕЖДУ(нижн_граница;верхн_граница),

где в качестве нижней границы возьмем значение =6, верхняя граница = 39. Постройте гистограмму выборку и функцию интегрального процента, найдите числовые характеристики и постройте

коробковую диаграмму. Сравните полученные результаты с соответствующими характеристиками для выборки из задания 1.

Указание. Так как значение функции **СЛУЧМЕЖДУ(…)** изменяется при каждом пересчете, после генерации выборки скопируйте получившиеся значения (в режиме «только значения») на отдельный лист.

Контрольные вопросы.

1. Дайте определение вариационного ряда и функции интегрального процента.
2. Перечислите числовые характеристики выборки, с помощью которых можно выявить общую тенденции. Дайте их определения.
3. Перечислите числовые характеристики выборки, с помощью которых можно выявить общую вариативность показателя. Дайте их определения.
4. Опишите процедуру построения гистограммы выборки.