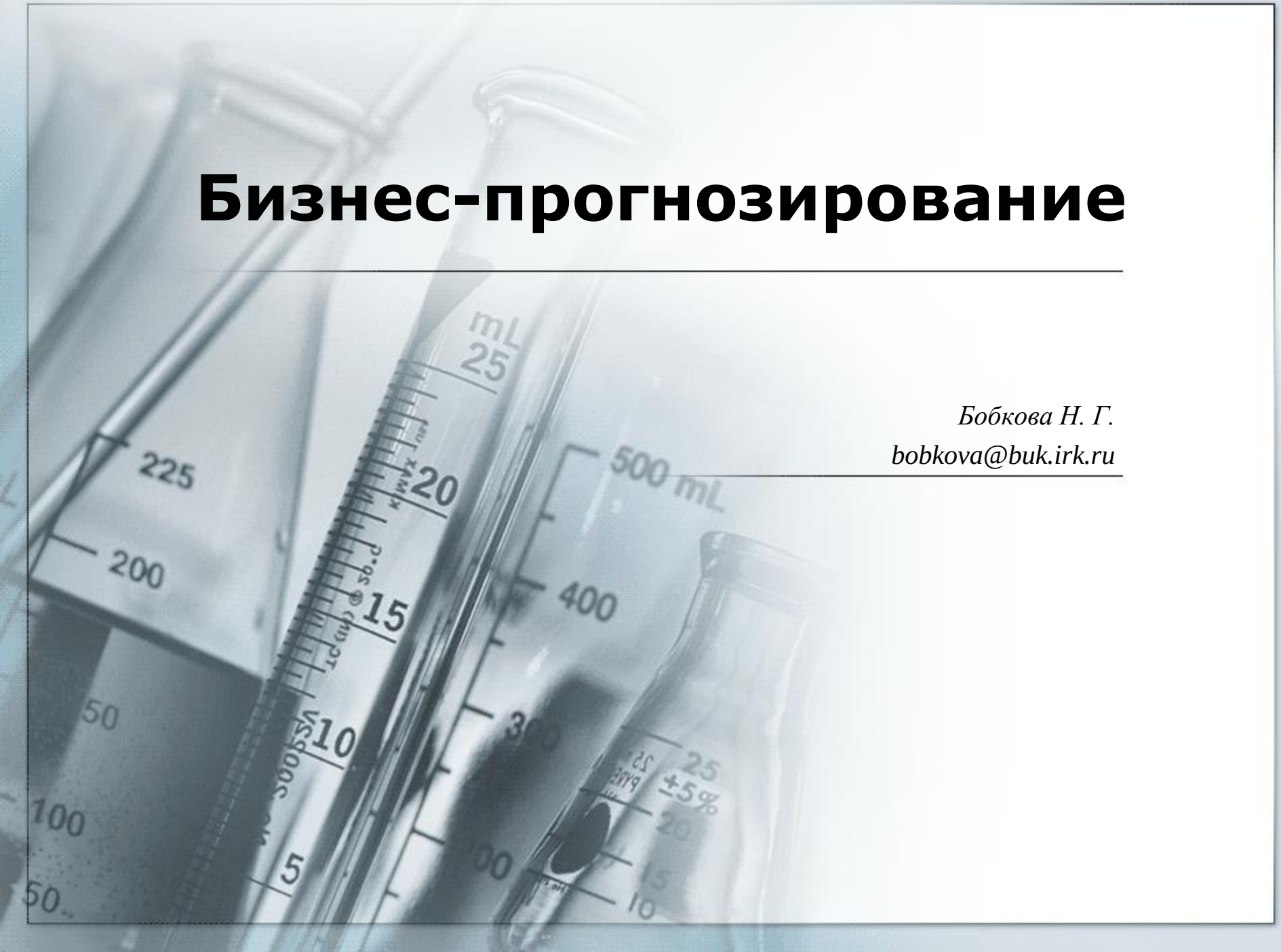
Бизнес-прогнозирование



Бобкова Н. Г.
bobkova@buk.irk.ru

КЛАССИФИКАЦИЯ ИСХОДНЫХ ДАННЫХ

Статистика

- Область математики, которая преобразует исходные данные в удобную и полезную информацию для принятия решений

- *Описательная статистика* – раздел статистики, изучающий методы сбора исходных данных, их описание и наглядное представление, а также вычисление количественных характеристик.

- *Аналитическая статистика* – раздел, посвященный обоснованиям правильности сделанных заключений о поведении генеральной совокупности на основе выборочных данных

ОПИСАТЕЛЬНАЯ СТАТИСТИКА

- Сбор данных
  - пример: опрос, перепись и т.д.
- Представление исходных данных
  - пример: таблицы и графики
- Характеристики исходных данных
  - пример: среднее, медиана, стандартное отклонение и т. д.

АНАЛИТИЧЕСКАЯ СТАТИСТИКА

- Оценка параметров генеральной совокупности
  - пример: оценивание средней зарплаты выпускников ВУЗа по имеющимся выборочным данным

- Проверка гипотез
  - пример: проверить справедливость утверждения о том, что средний объем жидкости, содержащийся в бутылке, равен 0,5 л

ОСНОВНЫЕ ПОНЯТИЯ СТАТИСТИКИ

- ПЕРЕМЕННАЯ – характеристика предмета или индивидуума
- ИСХОДНЫЕ ДАННЫЕ – набор различных значений, характеризующий данную переменную
- ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ – набор из всех предметов или индивидуумов, о котором Вы хотите сделать заключение

ОСНОВНЫЕ ПОНЯТИЯ СТАТИСТИКИ

- ВЫБОРКА – часть генеральной совокупности, отобранная для анализа
- ПАРАМЕТР – количественная мера, которая описывает характеристику генеральной совокупности
- СТАТИСТИКА – количественная мера, описывающая характеристику выборки

ЗАЧЕМ НАДО СОБИРАТЬ ИСХОДНЫЕ ДАННЫЕ?

- Маркетологам необходимо определить эффективность проводимой по ТВ рекламной компании
- Производители лекарств хотят определить, насколько новое лекарственное средство эффективно
- Менеджеры производственных фирм желают определить соответствие нового вида продукции существующим стандартам

ИСТОЧНИКИ ИСХОДНЫХ ДАННЫХ

- Исходные данные подразделяются на первичные и вторичные

- *Первичные исходные данные*: собираются тем, кто их будет анализировать (опрос, интервью, наблюдения, эксперимент)

- *Вторичные исходные данные*: собираются из каких-то источников (печатные издания, отчеты фирм и т.д.)

ТИПЫ ИСХОДНЫХ ДАННЫХ

- *Качественные исходные данные* –данные, которые характеризуют принадлежность элементов к каким-то классам (цвет, пол, название фирм) или определяют местонахождение элемента

- *Количественные исходные данные* – данные, которые характеризуют количественные значения исследуемых переменных (рост, вес, доход, спрос и т.д.)

УРОВНИ ИЗМЕРЕНИЯ ДАННЫХ

- *Шкала наименований* классифицирует качественные данные в отдельные группы, в которых не установлен порядок элементов

переменные	***группы***
наличие персональных компьютеров	да, нет
цвет	синий, красный
фирмы	Форд, Крайслер

УРОВНИ ИЗМЕРЕНИЯ ДАННЫХ

- *Порядковая шкала* используется для упорядочения (ранжирования) количественных исходных данных

переменные	***порядки***
студенты	1 курс, 2 курс,….
оценки студентов	5,4,3,2.
место в соревновании	1-е,2-е,…,7-е,..

- В шкале порядка отсутствуют понятия масштаба и начала отсчета

УРОВНИ ИЗМЕРЕНИЯ ДАННЫХ

- *Шкала интервалов* – упорядоченная шкала, в которой определена мера различия между количественными значениями признака. В общем случае шкала интервалов имеет произвольные точки отсчета и масштаб.

- *Шкала отношений* – упорядоченная шкала, в которой определена точка отсчета (нулевая точка)

УРОВНИ ИЗМЕРЕНИЯ ДАННЫХ

- Шкалы интервалов и отношений

Переменные	***Тип шкалы***
температура	интервальная
результаты тестов	интервальная
курс доллара	интервальная
вес	отношений
возраст	отношений
зар.плата	отношений
количество продаж	отношений

ТАБЛИЧНОЕ И ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ИСХОДНЫХ ДАННЫХ

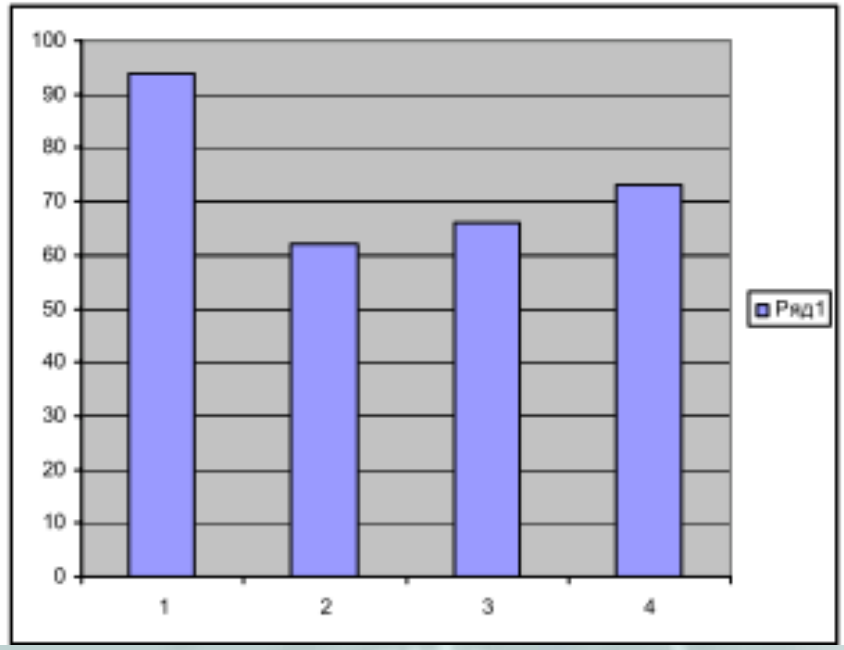
КАЧЕСТВЕННЫЕ ДАННЫЕ ИТОГОВЫЕ ТАБЛИЦЫ

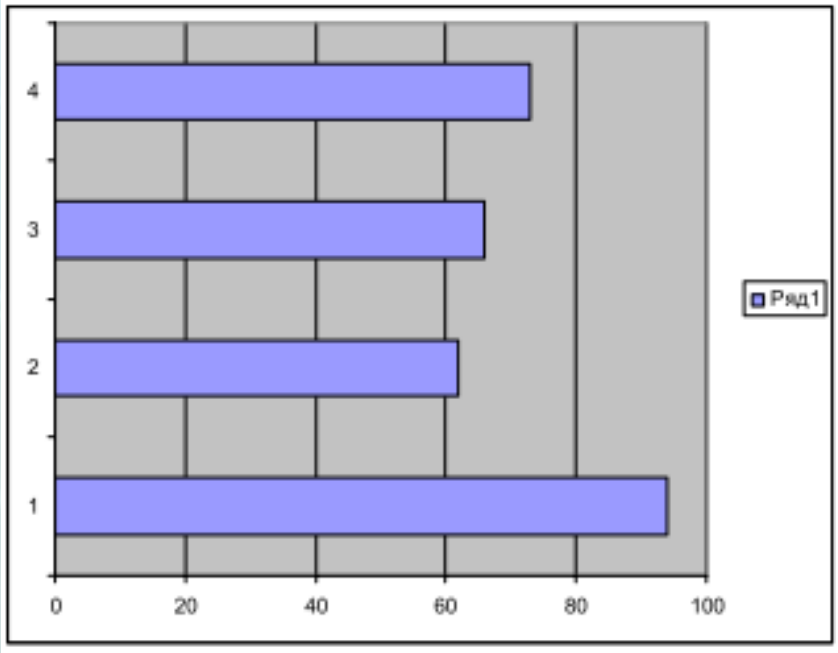
- *Итоговые таблицы* группируют частоту, количество или проценты объекта в виде набора категорий таким образом, что можно видеть разницу между категориями

курс	количество	относительная частота	процент	относительная накопленная частота
1-ый	94	0,32	32%	0,32
2-ой	62	0,21	21%	0,53
3-ий	66	0,22	22%	0,75
4-ый	73	0,25	25%	1,00
итого	*295*	*1,00*	*100%*	*1,00*

ПОЛОСОВЫЕ ДИАГРАММЫ

- В *полосовых диаграммах* полоса (горизонтальная или вертикальная) соответствует каждой категории, длина – представляет количество, частоту или процент соответствующей категории

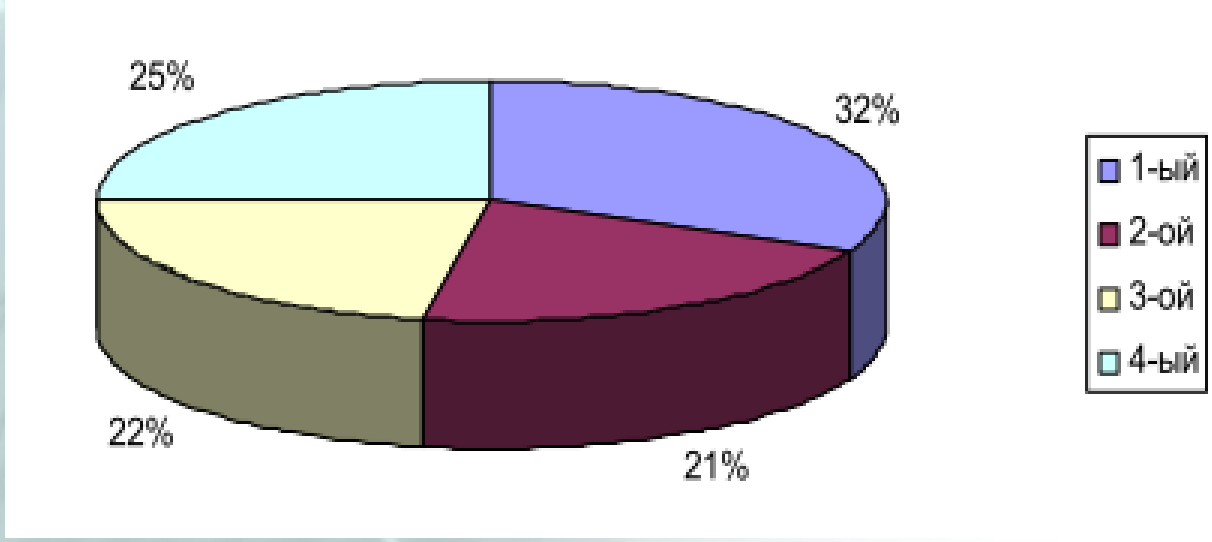




СЕКТОРНЫЕ ДИАГРАММЫ

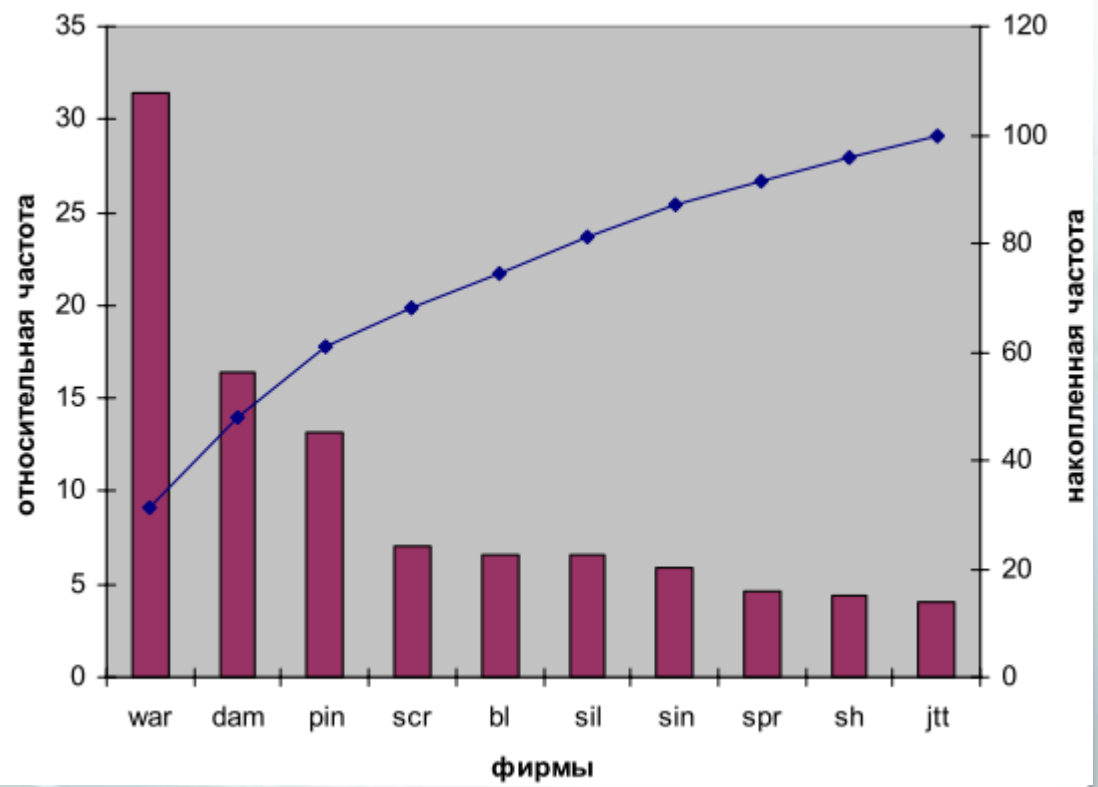
- *Секторные диаграммы* представляют собой круг, разбитый на секторы. Площадь каждого сектора пропорциональна удельному весу каждой отдельной категории

Распределение студентов по курсам



ДИАГРАММЫ ПАРЕТО

- *Диаграмма Парето* - диаграмма, в которой категории располагаются в порядке убывания частоты
  - используется для качественных данных
  - накопленный полигон показан на том же рисунке
  - предназначена для выделения наиболее важных классов



КОЛИЧЕСТВЕННЫЕ ИСХОДНЫЕ ДАННЫЕ

- *Частотное распределение* представляет собой таблицу, в которой данные сгруппированы в виде классов

- Необходимо определить количество классов исходя из подходящей ширины класса и установления границ классов

- Для определения ширины классов необходимо разделить размах данных (наибольшее значение – наименьшее значение) на число классов

ПРИМЕР ЧАСТОТНОГО РАСПРЕДЕЛЕНИЯ

- Данные представляют значения месячной оплаты за электроэнергию

96	171	202	178	147	102	153	197	127	82
157	185	90	116	172	111	148	213	130	165
141	149	206	175	123	128	144	168	109	167
95	163	150	154	130	143	187	166	139	149
108	119	183	151	114	135	191	137	129	158

ПОСТРОЕНИЕ ТАБЛИЦ РАСПРЕДЕЛЕНИЯ ЧАСТОТ

- Находим размах: 213 – 82 = 131
- Выбираем число классов: 5 (обычно между 5 и 15)
- Вычисляем ширину класса 131/5=26,2, округляем в большую сторону
- Определяем границы классов 82, 109, 136, 163, 190, 217
- Строим таблицу

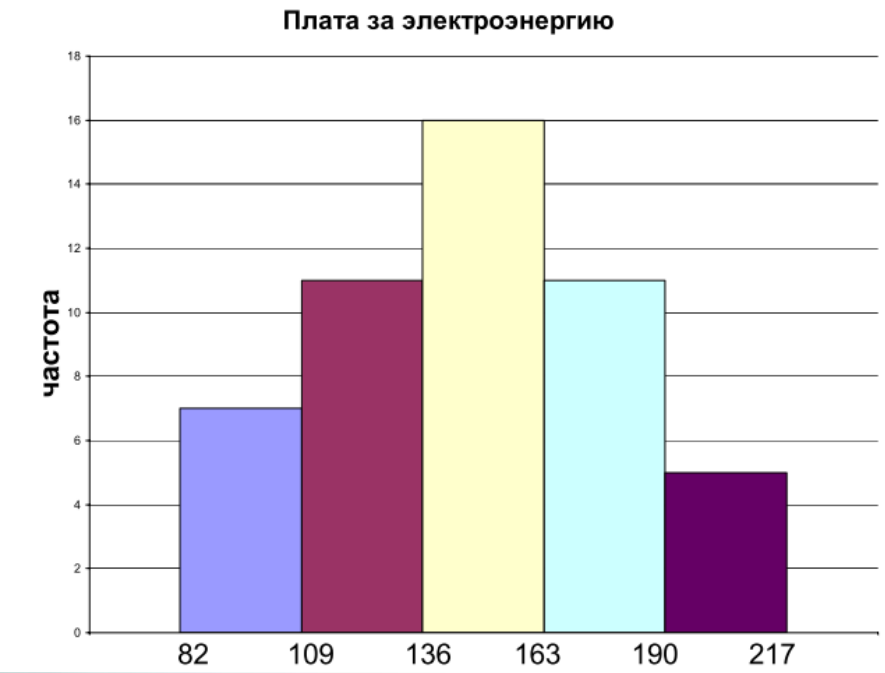
ТАБЛИЧНОЕ ПРЕДСТАВЛЕНИЕ ИСХОДНЫХ ДАННЫХ

величина оплаты	частота	относительная частота	процент	относительная накопленная частота
82-109	7	0,14	14%	0,14
110-136	11	0,22	22%	0,36
137-163	16	0,32	32%	0,68
164-190	11	0,22	22%	0,90
191-217	5	0,1	10%	1,00
Итого	50	1	100%	1,00

ИЗОБРАЖЕНИЕ КОЛИЧЕСТВЕННЫХ ДАННЫХ

- *Гистограмма*
  - Границы классов (или срединные точки) отмечаются на горизонтальной оси
  - На вертикальной оси

отображаются частота (или относительная частота, процент)

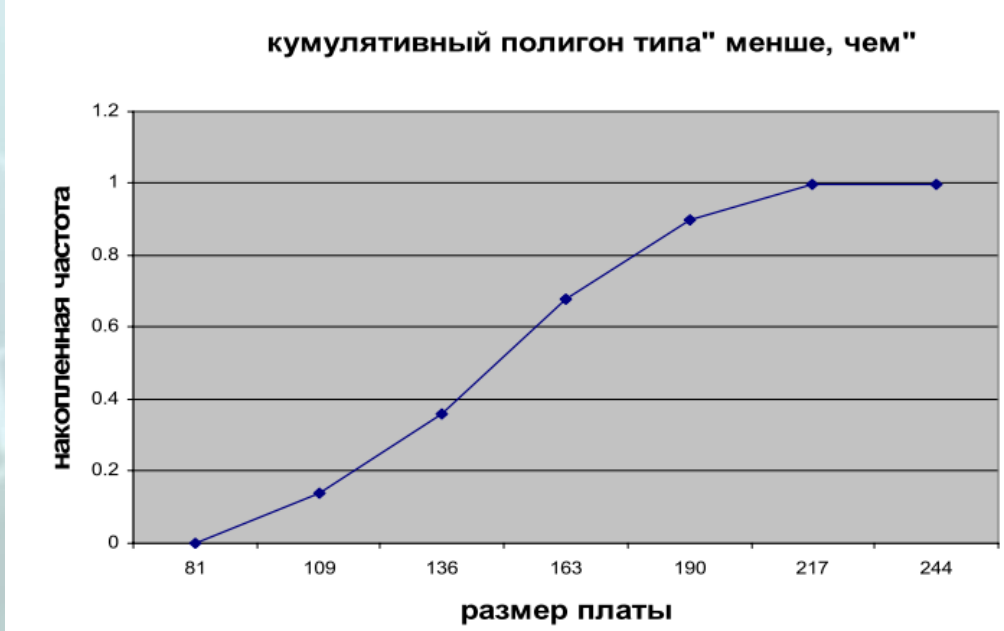


ИЗОБРАЖЕНИЕ КОЛИЧЕСТВЕННЫХ ДАННЫХ

- *Полигон*

ИЗОБРАЖЕНИЕ КОЛИЧЕСТВЕННЫХ ДАННЫХ

- *Кумулятивный полигон* (частот, относительных частот, процентов) помогает визуально определить:
  - сколько единиц наблюдений имеет значение признака, превышающее заданное значение
  - какой процент составляют значения исходных данных, меньшие определенного числа



ПЕРЕКРЕСТНЫЕ ТАБЛИЦЫ (ТАБЛИЦЫ СОПРЯЖЕННОСТИ)

- *Таблицы сопряженности* используются, когда имеется две качественные переменные.
- Классы одной переменной расположены по строкам, а классы другой переменной расположены по столбцам.
- На пересечении строки и столбца находится значение, принадлежащее как классу строки, так и классу столбца.

ТАБЛИЦА СОПРЯЖЕННОСТИ (пример)

- Опрошено 500 посетителей торгового центра. Среди вопросов был: “Нравится ли Вам ассортимент одежды в магазине?” Данные опроса представлены в таблице.

нравится или нет	**мужчины**	**женщины**	**итого**
да	136	224	360
нет	104	36	140
итого	240	260	500

ТАБЛИЦА СОПРЯЖЕННОСТИ (пример)

- Суммарный процент

нравится или нет	мужчины	женщины	итого
да	27,2	44,8	72
нет	20,8	7,2	28
итого	48	52	100

- Процент по строкам

нравится или нет	мужчины	женщины	итого
да	37,78	62,22	100
нет	74,29	25,71	100
итого	48	52	100

ТАБЛИЦА СОПРЯЖЕННОСТИ (пример)

- Диаграмма боковых полос для наглядного представления данных, находящихся в таблицах сопряженности

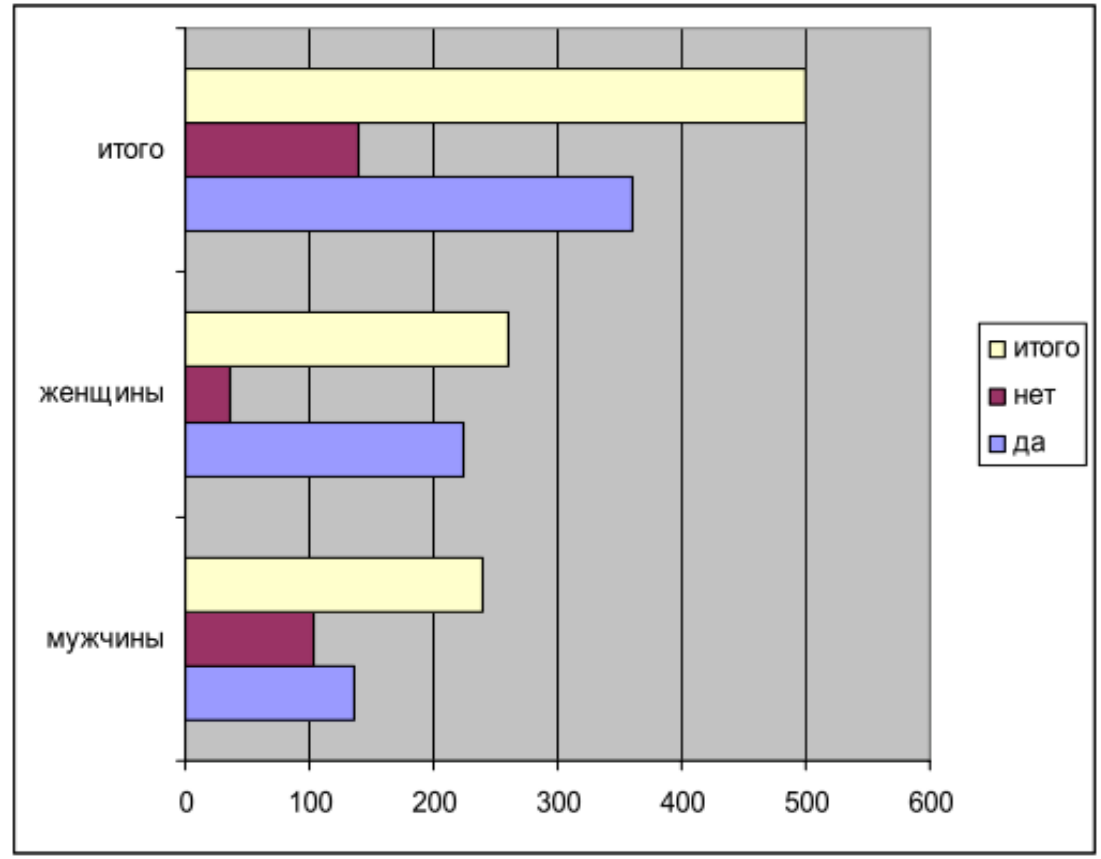
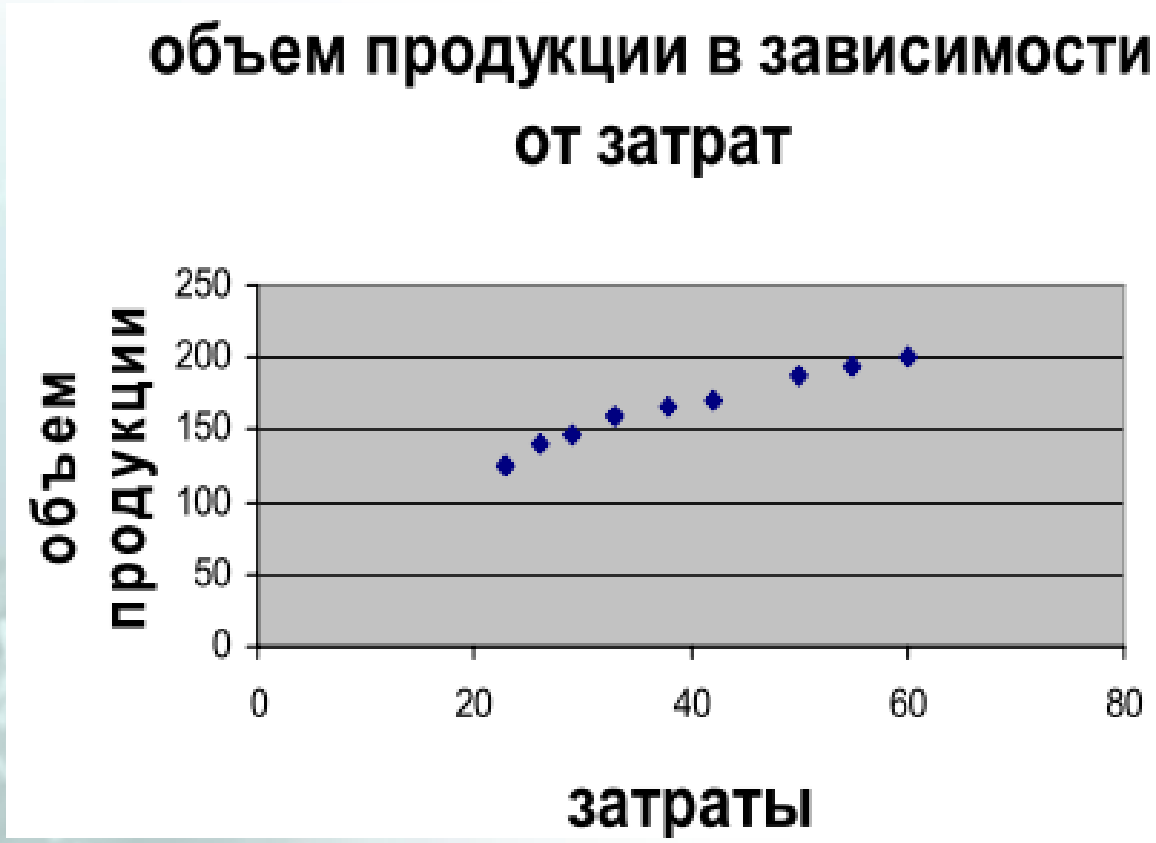


ГРАФИК РАССЕИВАНИЯ

- *График рассеивания* применяется в случае, когда имеются парные наблюдения, принадлежащие двум количественным переменным
- Значения одной переменной располагаются по горизонтальной оси, значения другой переменной – по вертикальной оси

ГРАФИК РАССЕИВАНИЯ

Объем за день	Издержки за день
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200



ВРЕМЕННЫЕ РЯДЫ

- *Временной (динамический) ряд* – совокупность значений какой-либо переменной за несколько последовательных периодов времени
- График временного ряда обычно изображается ломанной, причем временные периоды находятся на горизонтальной оси, а значения переменной на вертикальной оси

ВРЕМЕННЫЕ РЯДЫ (пример)

Квартальные данные об объеме продаж (млн. рублей)

Номер	Год	Квартал	Объем продаж
1	2007	1	20
2	2007	2	12
3	2007	3	27
4	2007	4	30
5	2008	1	23
6	2008	2	20
7	2008	3	34

ВРЕМЕННЫЕ РЯДЫ (пример)

ПРИНЦИПЫ ПОСТРОЕНИЯ ГРАФИКОВ

- График не должен искажать исходные данные
- График не должен содержать ненужных подписей, рисунков, украшений и т.д.
- Вертикальная ось должна начинаться с нуля
- Все оси должны иметь подписи
- График должен иметь название
- Необходимо использовать самый простой вид графика для исследуемых исходных данных

ПРИМЕРЫ

- Плохая презентация

1960 - $1.00

1970 - $1.60

1980 - $3.00

1990 - $3.80

- Хорошая презентация

ПРИМЕР (неправильно выбранная ось)

- Плохая презентация

- Хорошая презентация

ПРИМЕР
(сжатие / растяжение по вертикальной оси)

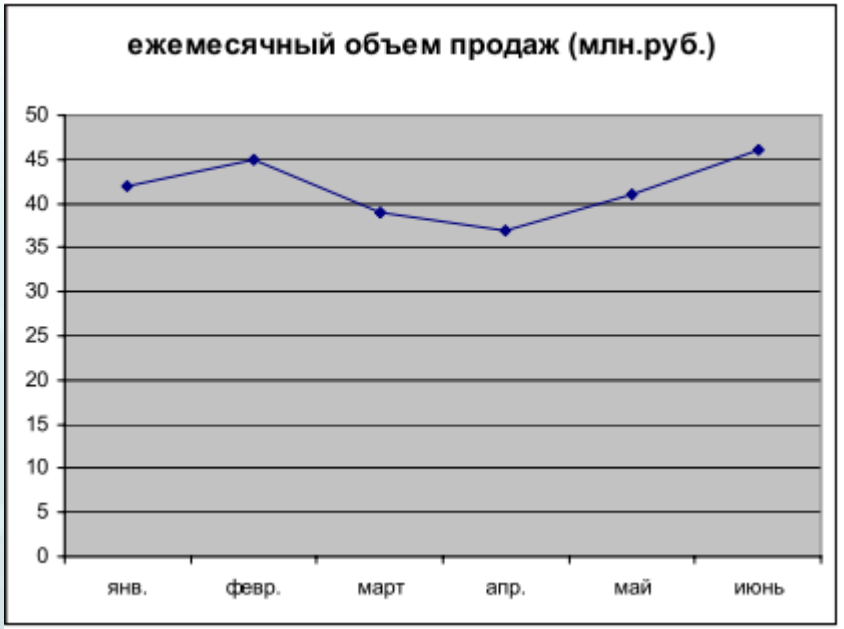
- Плохая презентация

- Хорошая презентация

ПРИМЕР
(нет нулевой точки по вертикальной оси)

- Плохая презентация

- Хорошая презентация



КОЛИЧЕСТВЕННЫЕ ОЦЕНКИ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

- *Центральная тенденция* характеризует свойство данных скапливаться вокруг какого-то центрального значения

- *Вариация* есть количество разброса или рассеивания значений исходных данных

СРЕДНЯЯ АРИФМЕТИЧЕСКАЯ

- *Средняя арифметическая (средняя)* – наиболее общая оценка центральной тенденции
- На значение средней влияют экстремальные значения

$$\overline{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + ... + X_n}{n}$$

n – количество наблюдений

X_i - значения

1; 2; 3; 4; 5

Средняя = 3

$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

1; 2; 3; 4; 10

Средняя = 4

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

СВОЙСТВА СРЕДНИХ

- Средняя может быть вычислена в шкалах интервалов и отношений (доход, вес, цена)
- При вычислении средней используются все значения
- Средняя арифметическая единственна
- Сумма отклонений каждого значения от средней равна нулю
- Средняя – наиболее удобная характеристика для сравнения нескольких совокупностей, одинаковых по качественному составу

МЕДИАНА

- Медиана характеризует величину, обладающую свойством: слева от медианы находится ровно половина всех данных, которые меньше ее, справа – половина всех данных, которые больше ее
- Для определения медианы необходимо все данные расположить в порядке возрастания
- Медиана расположена на (n + 1)/2 месте упорядоченного ряда (*примечание: (n + 1)/2 – это не значение медианы, а позиция медианы в ранжированных исходных данных*)
- Если количество значений равно нечетному числу, то медиана равна срединному числу
- Если количество значений четно, то медиана равна полусумме двух срединных чисел

СВОЙСТВА МЕДИАНЫ

- Если имеется одно или два экстремальных значения, то это не влияет на значение медианы
- Значение медианы единственное
- Медиана может быть определена, даже если используются не все данные
- Медиана может быть определена для данных, измеряемых как в шкалах отношений и интервалов, так и в порядковой шкале

МОДА

- Мода – наиболее часто встречающее значение переменной
- Значение моды не зависит от экстремальных значений
- Используется как для количественных, так и для качественных исходных данных
- Может моды не быть
- Может быть несколько мод

ХАРАКТЕРИСТИКИ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

Стоимость квартир	По возрастанию
500000	160000
160000	175000
175000	188000
210000	188000
188000	198000
232000	210000
198000	232000
188000	500000

- Средняя: (1851000/8) = 231375
- Медиана:
  (188000+198000)/2 = 193000
- Мода: 188000

КАКУЮ ХАРАКТЕРИСТИКУ ВЫБРАТЬ?

- Если нет экстремальных значений в исходных данных, то обычно используется средняя

- Так как медиана не чувствительна к экстремальным значениям, то при их присутствии медиана лучше представляет исходные данные

КВАРТИЛИ

- *Квартили* – это такие значения, которые делят ранжированные в порядке возрастания исходные данные на четыре равные по численности группы
- Первая квартиль Q1 – это такое значение, для которого 25% всех данных меньше Q1 и 75% больше Q1
- Q2 совпадает с медианой
- Только 25% значений больше третьей квартили Q3

ХАРАКТЕРИСТИКИ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

Стоимость квартир	По возрастанию
500000	160000
160000	175000
175000	188000
210000	188000
188000	198000
232000	210000
198000	232000
188000	500000

- Q1=184750
- Q2 =193000
- Q3 =215500

ОСНОВНЫЕ ХАРАКТЕРИСТИКИ ВАРИАЦИИ

- *Вариация* – это количественная оценка разброса значений исходных данных
  - размах
  - межквартильный размах
  - дисперсия
  - стандартное отклонение
  - коэффициент вариации

РАЗМАХ

- Самая простая характеристика вариации
- Размах равен разности между наибольшим значением исходных данных и наименьшим

$$R = X_{\max} - X_{\min}$$

- Пример: 1; 3; 9; 5; 13; 10; 12; 19; 6; 14
- R = 19 – 1 = 18

НЕДОСТАТКИ РАЗМАХА

- Не учитывает внутреннюю структуру исходных данных

7; 8; 9; 10; 11; 12

R = 12 – 7 = 5

7; 10; 11; 12; 12; 12

R = 12 – 7 = 5

- Чувствителен к экстремальным значениям

1; 1; 1; 1; 2; 2; 2; 3; 4; 5

R = 5 – 1 = 4

1; 1; 1; 1; 2; 2; 2; 2; 2; 120

R = 120 – 1 = 119

МЕЖКВАРТИЛЬНЫЙ РАЗМАХ

- *Межквартильный IQR размах* исключает проблемы, связанные с наличием экстремальных значений данных

- Межквартильный размах равен разности между третьей квартилью и первой квартилью

$$IQR = Q_3 - Q_1$$

ДИСПЕРСИЯ

- *Дисперсия* есть средняя (приблизительно) квадратов отклонений значений исходных данных от средней данных
- Дисперсия выборки

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

  - $\overline{X}$ – средняя выборки
  - n – размер выборки
  - X_i – i-ое значение переменной X

СТАНДАРТНОЕ ОТКЛОНЕНИЕ

- Наиболее часто используемая характеристика вариации
- Характеризует разброс данных относительно средней
- Имеет ту же размерность, что и исходные данные

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$$

СТАНДАРТНОЕ ОТКЛОНЕНИЕ

- Исходные данные: 11; 12; 13; 16; 16; 17; 18; 21
- Средняя = 15,5 S = 3,338

- Исходные данные: 14; 15; 15; 15; 16; 16; 16; 17
- Средняя = 15,5 S = 0,926

- Исходные данные: 11; 11; 11; 12; 19; 20; 20; 20
- Средняя =15,5 S = 4,570

ОБЩИЕ СВОЙСТВА ХАРАКТЕРИСТИК РАЗБРОСА

- Чем больше исходные данные рассредоточены, тем больше значения размаха, межквартильного размаха, дисперсии и стандартного отклонения
- Если все исходные данные равны между собой, то все характеристики разброса равны нулю
- Ни одна из вышеперечисленных характеристик не может быть отрицательной

КОЭФФИЦИЕНТ ВАРИАЦИИ

- *Коэффициент вариации* равен отношению стандартного отклонения к средней, умноженное на100
- Он всегда выражается в процентах (не зависит от единиц измерения)
- Используется при сравнительном анализе нескольких совокупностей, когда данные измеряются в качественно различных единицах или средние существенно отличаются между собой

$$CV = (\frac{S}{\overline{X}}) * 100\%$$

КОЭФФИЦИЕНТ ВАРИАЦИИ

- Рынок А:
  - Средняя цена за прошлый год = $50
  - Стандартное отклонение = $5
  - Коэффициент вариации = 10%
- Рынок В:
  - Средняя цена за прошлый год = $100
  - Стандартное отклонение = $5
  - Коэффициент вариации = 5%
- Оба рынка имеют одинаковое стандартное отклонение, но рынок В имеет меньший разброс относительно средней цены, чем рынок А

Z - ОЦЕНКИ

- *Z-оценки* определяют число стандартных отклонений, содержащихся между конкретным значением исходных данных и средней данных
- Конкретное значение считается экстремальным, если его Z-оценка меньше -3,0 или больше чем +3,0
- Чем больше абсолютное значение Z-оценки, тем дальше данное значение от средней

$$Z = \frac{X - \overline{X}}{S}$$

Z - ОЦЕНКИ

- Пусть средняя оценка по тесту равна 490 со стандартным отклонением 100. Результат господина Иванова – 620. Найти Z-оценку и дать ей интерпретацию.

$$Z = \frac{X - \bar{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1{,}3$$

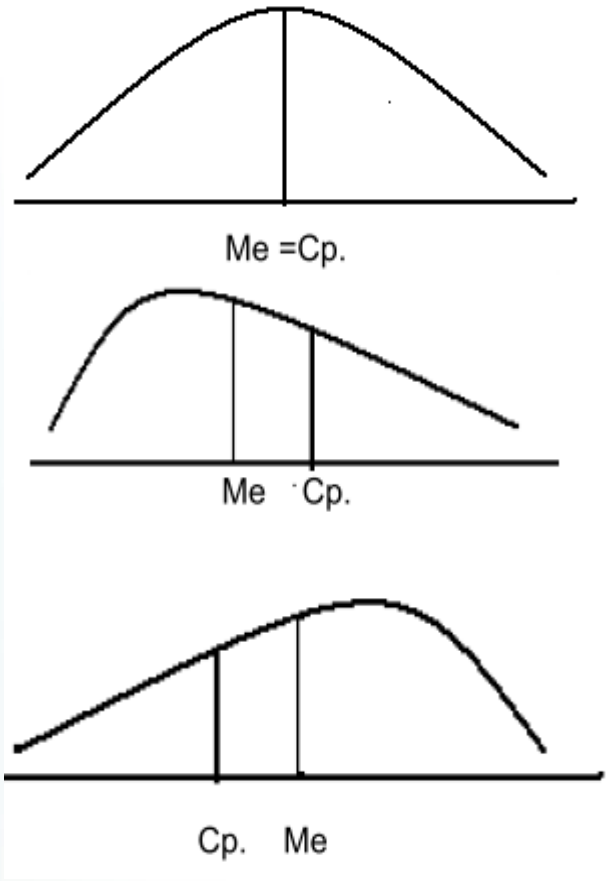
- Результат 620 на 1,3 стандартных отклонения больше средней оценки и поэтому не может быть отнесен к категории экстремальных

ПРОЦЕНТНЫЕ ТОЧКИ

- *P-процентная точка* – это такое значение, которое делит упорядоченные исходные данные таким образом, что P% всех значений меньше данной точки, а (1-P)% значений исходных данных больше
- 25%-процентная точка совпадает с первой квартильей, 75%-процентная точка равна третьей квартили
- 80% всех значений исходных данных находятся между 10%-процентной точкой и 90%-процентной точкой

ФОРМА КРИВОЙ РАСПРЕДЕЛЕНИЯ

- Если кривая распределения симметрична, то средняя совпадает с медианой
- Если график имеет хвост, вытянутый вправо, то говорят о положительной асимметрии и в этом случае медиана меньше средней
- В случае отрицательной асимметрии (хвост графика вытянут влево) медиана больше средней



Коэффициент асимметрии

- Коэффициент асимметрии определяет степень асимметрии
- Если коэффициент асимметрии положителен, то асимметрия положительна
- Если коэффициент асимметрии отрицателен, то асимметрия отрицательна
- Если равен нулю, то распределение симметрично

$$S_k = \frac{3(\overline{X} - Me)}{S}$$

КОЛИЧЕСТВЕННЫЕ ХАРАКТЕРИСТИКИ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

- Характеристики выборки называются статистиками
- Характеристики генеральной совокупности называются параметрами и обозначаются греческими буквами
- Наиболее важные параметры генеральной совокупности – это средняя, дисперсия и стандартное отклонение

СРЕДНЯЯ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

- Средняя генеральной совокупности – это сумма всех значений, деленная на размер генеральной совокупности N

$$\mu = \frac{\sum X_i}{N} = \frac{X_1 + X_2 + \ldots + X_N}{N}$$

μ - средняя генеральной совокупности

N – размер генеральной совокупности

X_i - i-ое значение переменной X

ДИСПЕРСИЯ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

- Дисперсия генеральной совокупности – это средняя квадратов отклонений значений исходных данных от средней данных

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

- Стандартное отклонение

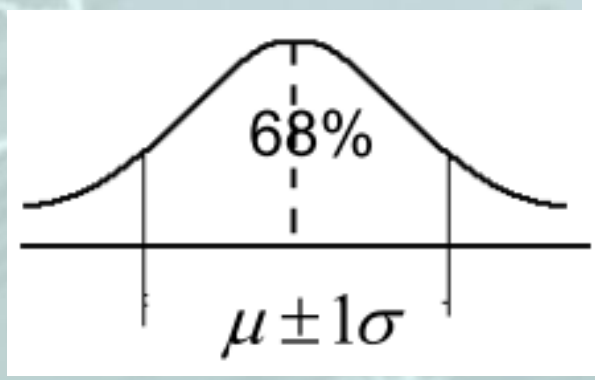
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

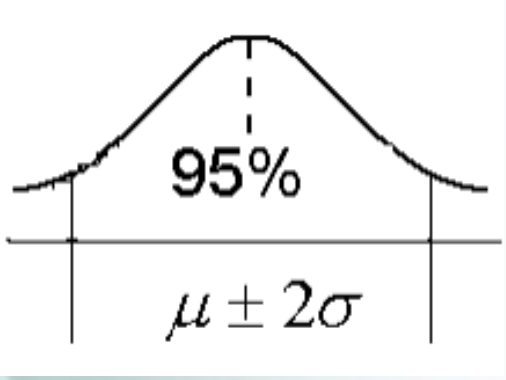
СТАТИСТИКИ ВЫБОРКИ И ПАРАМЕТРЫ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

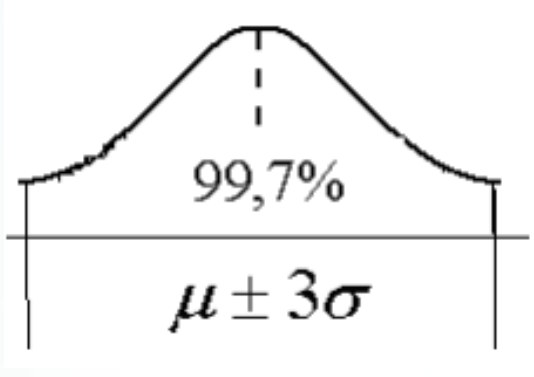
Название	Параметры генеральной совокупности	Статистики выборки
Средняя	μ	$\overline{X}$
Дисперсия	σ^2	S^2
Стандартное отклонение	σ	S

ЭМПИРИЧЕСКОЕ ПРАВИЛО

- Если исходные данные имеют распределение в виде колоколообразной кривой, то:
  - приблизительно 68% всех данных находятся на расстоянии одного стандартного отклонения от средней
  - приблизительно 95% всех данных – на расстоянии двух стандартных отклонений от средней
  - приблизительно 99,7% всех данных – на расстоянии трех стандартных отклонений от средней







ПРАКТИЧЕСКОЕ ИСПОЛЬЗОВАНИЕ ЭМПИРИЧЕСКОГО ПРАВИЛА

- Тестирование большого количества батареек показало, что их средний срок службы – 19 часов, имеет колоколообразное распределение со стандартным отклонением 1,2 часа.
- Примерно 68% имеют срок службы в границах от 17,8 до 20,2 часов (19 +/- 1,2)
- Примерно 95% имеют срок службы от 16,6 до 21,4 часов (19 +/- 2*1,2)
- Около 99,7% всех батареек имеют срок службы в пределах от 15,4 до 22,5 часов (19 +/- 3*1,2)

ПРАВИЛО ЧЕБЫШЕВА

- Независимо от того, как распределены исходные данные (симметрично или асимметрично), по крайней мере, $\left(1 - \frac{1}{K^2}\right)$ значений находится в пределах K стандартных отклонений от средней для любых K, больших единицы

КОВАРИАЦИЯ ВЫБОРКИ

- *Ковариация* есть мера измерения линейной зависимости между двумя количественными переменными

$$COV(X,Y) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

- Ковариация только определяет наличие зависимости, но не выявляет причину зависимости
  - *COV (X,Y)* > 0 => X и Y одновременно или возрастают, или убывают
  - *COV (X,Y)* < 0 => если X возрастает, то Y убывает и наоборот
  - *COV (X,Y)* = 0 => X и Y независимы

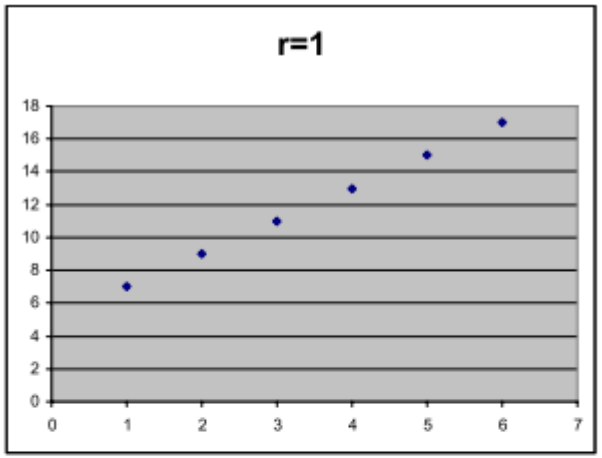
КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

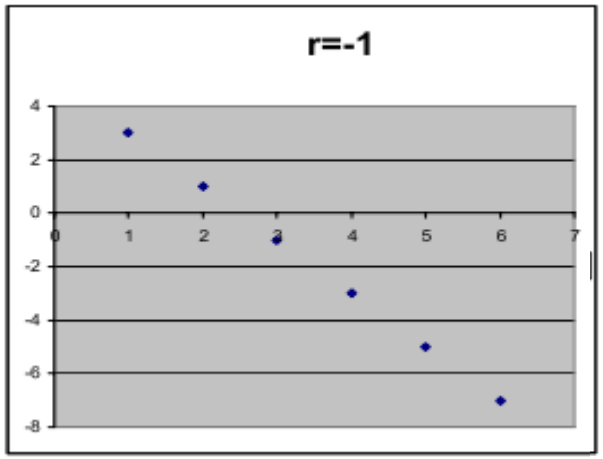
- *Коэффициент корреляции* служит мерой относительной линейной зависимости между двумя переменными

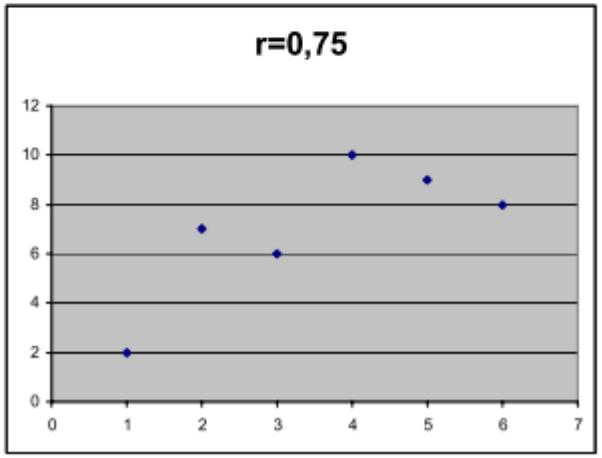
$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}\sqrt{\sum (Y_i - \overline{Y})^2}} = \frac{COV(X,Y)}{S_x S_y}$$

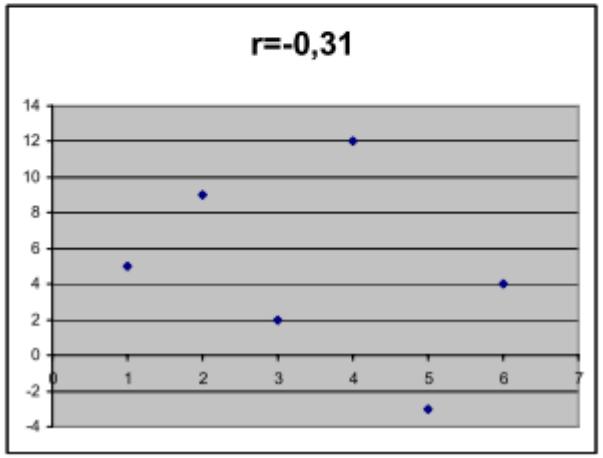
- Безразмерная величина
- Принимает значения в промежутке [-1;+1]
- Чем ближе r к (-1), тем сильнее отрицательная линейная зависимость
- Чем ближе значения коэффициента корреляции к (+1), тем сильнее положительная линейная взаимосвязь
- Любая линейная взаимосвязь слабая, если значение r близко к нулю

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ









КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ *(пример)*

Фирма сбывает свой товар в разных регионах. Случайно отобраны данные по 10 регионам. Пары наблюдений по регионам: затраты на рекламу X (дес. тыс. долл.) и объем продаж Y (тыс.ед.)

Коэффициент корреляции r = 0,845. Это значит, что между затратами на рекламу и объемом продаж существует достаточно тесная линейная положительная зависимость, т.е. с ростом затрат на рекламу объем продаж в среднем возрастает

регион	X	Y
1	22	16
2	25	17
3	45	25
4	37	24
5	28	22
6	50	21
7	56	32
8	34	18
9	60	30
10	40	20

НЕКОТОРЫЕ ЭТИЧЕСКИЕ АСПЕКТЫ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ

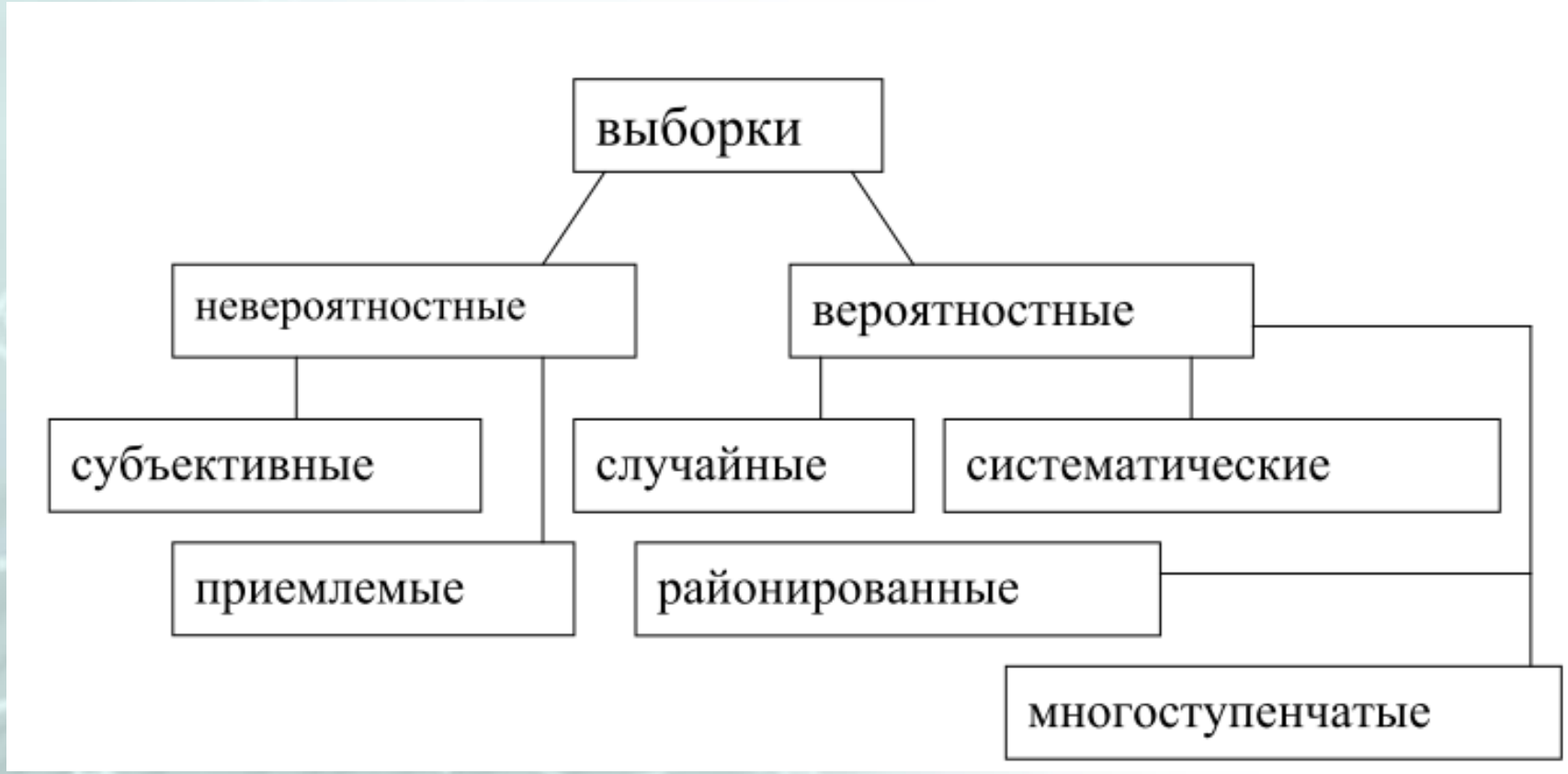
- Количественные описательные оценки:
  - Должны обосновывать как “хорошие” результаты, так и “плохие”
  - Должны быть представлены в честной, объективной и нейтральной форме
  - Не должны использовать суммарные оценки, искажающие реальную действительность
  - Анализ исходных данных должен быть объективным
  - Интерпретация исходных данных есть субъективный процесс, который не должен приводить только к желательным выводам

ВЫБОРКИ И ВЫБОРОЧНЫЕ РАСПРЕДЕЛЕНИЯ

ПОЧЕМУ ВЫБОРКА?

- Формирование выборочных данных требует меньше времени
- Формирование выборки менее затратно с финансовой точки зрения
- Анализ выборки менее громоздкий и более практичный, чем анализ всей генеральной совокупности

ТИПЫ ВЫБОРОК



ТИПЫ ВЫБОРОК

- Формирование “*приемлемых*” выборок предполагает, что отбор элементов выборки осуществляется наиболее простым, недорогим и удобным способом

- Отбор элементов при формировании “*субъективных*” выборок осуществляется на основе мнения эксперта, здравого смысла

ПРОСТАЯ СЛУЧАЙНАЯ ВЫБОРКА

- Любой элемент генеральной совокупности имеет одинаковый шанс быть выбранным
- Выбор может быть с возвратом (выбранный элемент может быть возвращен для возможного повторного отбора) или без возврата
- Для формирования простой случайной выборки рекомендуется использовать таблицы случайных чисел

ФОРМИРОВАНИЕ СИСТЕМАТИЧЕСКИХ ВЫБОРОК

- Определяется размер выборки: n
- Размер совокупности N делится на n и определяется количество групп k = N/n
- Случайным образом выбирается элемент из первой группы, затем выбираются каждые k-ые элементы соответственно
- Например, предположим, что размер выборки n=9 и N=72. Тогда k = 72/9 = 8 группам. Случайным образом выбирается, например, 3-ий элемент из первой группы. Затем выбираются каждый 8-ой элемент соответственно (т.е. 3, 11, 19, 27, 35, 43, 51, 59, 67)

РАЙОНИРОВАННЫЕ ВЫБОРКИ

- Исследователь делит генеральную совокупность на несколько “районов”
- Элементы выборки отбираются случайным образом не из всей совокупности как целого, а из каждого “района” отдельно
- Районированный отбор используется, например, при социологических опросах, когда районирование может производиться по территориальному, социальному и демографическому признакам

МНОГОСТУПЕНЧАТЫЙ ОТБОР

- Многоступенчатый отбор предполагает последовательность случайных отборов, причем извлечение единиц в выборку происходит на последней стадии отбора

- Например, необходимо обследовать областные города. Такой отбор может быть проведен в три этапа: единицы отбора первого этапа – края, единицы отбора второго этапа – области, третьего этапа (составляющие выборку) – областные города

ОЦЕНИВАНИЕ ЦЕННОСТИ ОБСЛЕДОВАНИЯ

- Какова цель обследования?
- Какие формировались выборки – вероятностные или невероятностные?
- Ошибки охвата – подходящая ли совокупность?
- Ошибки, связанные с нечестными ответами – разобраться почему
- Ошибки измерения – хорошие вопросы подразумевают честные ответы
- Ошибки выборки – всегда существуют

ТИПЫ ОШИБОК ОПРОСОВ

- Ошибки охвата возникают в следствии того, что отбор был пристрастным (некоторые группы были исключены и не имеют шанса быть выбранными)
- Ошибки из-за нечестных ответов (все люди разные)
- Ошибки выборки (разные выборки могут давать разную информацию)
- Ошибки измерения (неточные вопросы, зависимость опрашиваемых, некомпетентность опрашиваемых)

ВЫБОРОЧНЫЕ РАСПРЕДЕЛЕНИЯ

- *Выборочное распределение* – это распределение всех возможных значений статистик для данного размера выборки, сформированной из генеральной совокупности

- Например, предположим, что Вы выбираете 50 студентов из всех студентов университета и Вас интересует их средний балл. Если Вы сформируете несколько разных выборок размером 50, то получите различные значения средних. Распределение всех возможных средних и будет выборочным распределением

ТОЧЕЧНЫЕ ОЦЕНКИ

- Под *точечной оценкой* понимается число (точка), которое используется в качестве оценки параметра генеральной совокупности

- Чтобы оценить качество точечных оценок в статистическом анализе рассматриваются четыре критерия:
  - несмещенность
  - эффективность
  - состоятельность
  - достаточность

КРИТЕРИИ КАЧЕСТВА ТОЧЕЧНЫХ ОЦЕНОК

- *Несмещенность*: статистика называется несмещенной, если все выборочные значения располагаются симметрично относительно истинного значения оцениваемого параметра
- *Эффективность:* критерий эффективности характеризует минимальность стандартной ошибки статистики, используемой в качестве точечной оценки параметра генеральной совокупности
- *Состоятельность:* оценка истинного значения параметра является состоятельной, если по мере увеличения объема выборки ее значение приближается к истинному значению параметра
- *Достаточность*: оценка является достаточной, если при ее вычислении используется вся содержащаяся в выборке информация
- Выборочная средняя удовлетворяет всем критериям

ВЫБОРОЧНЫЕ РАСПРЕДЕЛЕНИЯ СТАНДАРТНАЯ ОШИБКА СРЕДНЕЙ

- Разность между значением статистики и соответствующим параметром называется ошибкой выборки
- Разные выборки одного и того же размера из одной совокупности как правило имеют разные значения средних
- Мерой разброса средней выборочного распределения служит стандартная ошибка средней

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА

- Если генеральная совокупность имеет нормальный закон распределения со средней и стандартным отклонением выборочное распределение средних также распределено по нормальному закону

- Если распределение генеральной совокупности не является нормальным, то распределение выборочных средних при достаточно больших размерах выборок почти нормально

$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

ВЫБОРОЧНЫЕ РАСПРЕДЕЛЕНИЯ

- Если генеральная совокупность имеет распределение отличное от нормального, то:
  - Для большинства распределений размер выборки n>30 приводит к тому, что выборочное распределение средних примерно нормально
  - Для распределений близких к симметричным n>15 приводит к тому, что выборочное распределение средних примерно нормальное
  - Чем больше n, тем меньше значение стандартной ошибки средней

ВЫБОРОЧНЫЕ РАСПРЕДЕЛЕНИЯ (пример)

- Пусть средняя генеральной совокупности = 8 со стандартным отклонением = 3. Выбрана случайная выборка размером n = 36. Какова вероятность, что средняя выборки будет между 7,75 и 8,25?
- О распределении генеральной совокупности нам ничего не известно, но n>30. В силу центральной предельной теоремы распределение выборочных средних примерно нормально, т.е.

$$\mu_{\bar{X}} = 8 \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0{,}5$$

$$P(7{,}75 < \mu_{\bar{X}} < 8{,}25) = P(-0{,}5 < Z < 0{,}5) = 0{,}3830$$

ВЫБОРОЧНЫЕ РАСПРЕДЕЛЕНИЯ ДОЛЯ (ПРОПОРЦИЯ)

- Под *долей* (или *пропорцией*) *p* понимается относительная или процентная характеристика, определяющая часть элементов совокупности, обладающих некоторым признаком (свойством)
- Доля выборки является точечной оценкой p
- Если X – число элементов конкретного признака в выборке размером n, то

$$\bar{p} = \frac{X}{n} \qquad 0 \le \bar{p} \le 1$$

- Стандартная ошибка доли

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

ОБЩИЕ ПРИНЦИПЫ ИНТЕРВАЛЬНЫХ ОЦЕНОК

- *Доверительный интервал* – это промежуток, внутри которого с известной вероятностью находится истинное значение параметра
- Базируется на данных одной выборки
- Зависит от *уровня значимости* (доверия) α. По заданному значению α определяется доверительная вероятность (1-α). Например, если α=0,05, то это значит, что степень доверия к полученному интервалу 95%
- Никогда не может быть 100% степени доверия

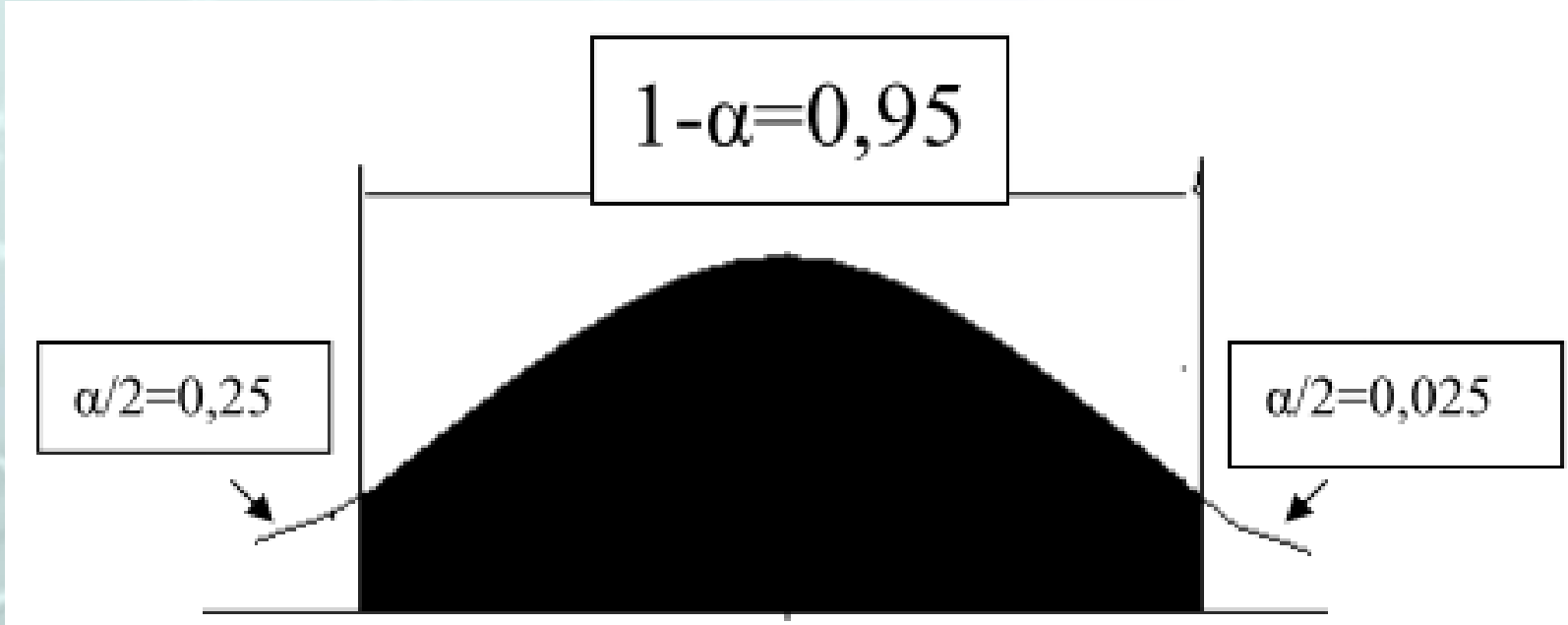
ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ μ (σ известно)

- Допущения:
  - стандартное отклонение генеральной совокупности σ известно
  - генеральная совокупность распределена по нормальному закону
- Доверительный интервал имеет вид

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ μ (σ известно)

- Представим 95% доверительный интервал



НАХОЖДЕНИЕ КРИТИЧЕСКОГО ЗНАЧЕНИЯ Z

- Обычно находятся 90%, 95%, и 99% доверительные интервалы

Уровень значимости	Доверительная вероятность	Значения Z
0,2	80%	1,28
0,1	90%	1,645
0,05	95%	1,96
0,02	98%	2,33
0,01	99%	2,58

ДОВЕРИТЕЛЬНЫЙ ИТЕРВАЛ ДЛЯ μ, σ ИЗВЕСТНО (пример)

- Компания производит определенный тип электроприборов, которые имеют разный срок службы. Ранее было установлено, что стандартное отклонение составляет 50 часов. Из вновь произведенной партии была извлечена выборка объемом n =10 и их средний срок службы составил 384 часа. Определить 90% и 95% доверительные интервалы.

ДОВЕРИТЕЛЬНЫЙ ИТЕРВАЛ (пример)

- 90% интервал

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 384 \pm 1{,}645 \frac{50}{\sqrt{10}} = 384 \pm 16$$

- С уверенностью 90% можно утверждать, что истинное значение среднего срока службы приборов находится в промежутке от 368 до 400 часов. Аналогично находим 95% доверительный интервал. Он будет от 353 до 415 часов.
- Заметим, что чем больше степень доверия, тем доверительный интервал шире.

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ μ (σ неизвестно)

- Если стандартное отклонение генеральной совокупности σ неизвестно, то используется стандартное отклонение выборки S
- Это приводит к дополнительной неопределенности, так как S может иметь разные значения при разных выборках
- Поэтому в данном случае необходимо использовать t распределение вместо нормального

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ μ (σ неизвестно)

- Допущения:
  - стандартное отклонение генеральной совокупности неизвестно
  - генеральная совокупность распределена по нормальному закону
- Доверительный интервал

$$\overline{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

$t_{\alpha/2,n-1}$ - критическое значение t распределения с n-1 степенями свободы при уровне значимости α

t распределение Стьюдента

- Значение t распределения зависит от числа степеней свободы. Число степеней свободы для t распределения равно количеству элементов выборки минус единица, т. е. d.f. = n-1 (количество наблюдений, которые свободны изменяться после того, как вычислена средняя выборки).

- При n>30 t распределение почти полностью совпадает с нормальным.

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ (пример)

- Средняя случайной выборки размером n = 25 равна 50, S = 8. Определить 95% доверительный интервал для μ.
- Число степеней свободы d. f. = n-1 = 24, α = 0,05
- Доверительный интервал

$$\overline{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}} = 50 \pm (2{,}0639) \frac{8}{\sqrt{25}} = (46{,}7;\ 53{,}3)$$

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ДОЛИ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

- Доверительные интервалы для доли генеральной совокупности (p) определяются следующим образом:

  1. По данным выборки находится выборочная доля
  2. Находится стандартная ошибка доли

$$S_p = \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

  3. Доверительный интервал вычисляется по формуле

$$\overline{p} \pm Z_{\alpha/2} S_p$$

где n – размер выборки;

α – уровень значимости

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ДОЛИ (пример)

- Случайным образом выбрано 100 студентов. 25 из них сдали экзамен по статистике на отлично. Каков 95% доверительный интервал для всех студентов, которые получили отличную оценку по статистике?

- Находим

$$\overline{p} = \frac{25}{100} = 0{,}25;\ S_p = \sqrt{\overline{p}(1-\overline{p})/n} = \sqrt{0{,}25(0{,}75)/100} = 0{,}0433$$

- Тогда 95% доверительный интервал

$$0{,}25 \pm 1{,}96(0{,}0433) \textit{ или } (0{,}1661 - 0{,}3349)$$

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ДОЛИ (пример)

- Мы имеем 95% уверенности, что истинная доля тех студентов, которые получили отличную оценку по статистике составляет от 16,51% до 33,49% от всех студентов

- Несмотря на то, что истинное значение доли отличных оценок может и не быть в пределах от 0,1651 до 0,3349, 95% всех возможных интервалов, сформированных из выборок размером 100, будут содержать истинное значение доли

ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРКИ, ОБЕСПЕЧИВАЮЩЕГО ДОПУСТИМУЮ ОШИБКУ

- Обозначим ошибку выборки, то есть разность между истинным значением параметра и его точечной оценкой, через E (не путать со стандартной ошибкой)
- Для средней ошибка выборки

$$E = \left|\mu - \overline{X}\right| = Z_{\alpha/2}\frac{S}{\sqrt{n}}$$

- Выразим отсюда значение n

$$n = \frac{Z^2{}_{\alpha/2}S^2}{E^2}$$

ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРКИ (пример)

- Торговая фирма хочет открыть новый супермаркет. Для проведения маркетинговых исследований необходимо иметь информацию о годовых доходах семей, живущих в прилегающем районе.
- Предварительные исследования показали, что доходы варьируются в пределах от 9000 до 29000 долл. Однако для более надежных прогнозов необходима точность в размере 200 долл. С доверительной вероятностью 95%.
- Для определения стандартного отклонения S было проведено обследование 50 семей. В результате была получена оценка стандартного отклонения S = 3000 долл.
- Требуется определить сколько семей надо обследовать, чтобы получить заданную точность

ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРКИ (пример)

- В нашем случае $\alpha = 0{,}05$, $E = 3000$, $S = 3000$.
- Тогда

$$n = \frac{Z_{\alpha/2}^2 S^2}{E^2} = \frac{1{,}96^2 * 3000^2}{200^2} = 864{,}36$$

- Очевидно, что выборка из 865 семей может обеспечить заданную точность

ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРКИ (для доли)

- Чтобы определить необходимый размер выборки, обеспечивающий заданную точность, как и в случае для средней надо знать:
  - уровень значимости α, по которому определяется критическое значение Z
  - величину допустимой ошибки E
  - истинное значение доли генеральной совокупности (если она неизвестна, то ее берут равной 0,5)
- Тогда

$$E = Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$n = \frac{Z^2_{\alpha/2}\,p(1-p)}{E^2}$$

ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРКИ ДЛЯ ДОЛИ (пример)

- Сколько избирателей необходимо опросить, чтобы ошибка прогноза составляла 2% с вероятностью 0,95?
- В данном случае мы не имеем информации о значении доли p, поэтому полагаем p = 0,5.
- Тогда

$$n = \frac{1{,}96^2}{4 * 0{,}02^2} = 2401$$

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ (генеральная совокупность конечного объема)

- В случае конечной генеральной совокупности объема N при условии n/N>0,05 следует учитывать поправку на конечность объема
- Это делается с помощью поправочного множителя следующим образом:
- для средней

$$\overline{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

- для доли

$$\overline{p} \pm Z_{\alpha/2} \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \sqrt{\frac{N-n}{N-1}}$$

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ (пример)

- В фирме работают 645 служащих. Составлена случайная выборка из 49 человек. Средний недельный заработок в выборке равен 110 долл. со стандартным отклонением 10,5 долл. Требуется найти 95-процентные доверительные пределы для среднего недельного заработка всех служащих фирмы

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ (пример)

- В данном случае n = 49, N = 645, следовательно n/N = 49/645 = 0,07>0,05.
- Поэтому для нахождения 95%-процентных доверительных пределов следует учитывать поправочный множитель. Тогда

$$110 \pm 1{,}96 \frac{10{,}5}{\sqrt{49}} \sqrt{\frac{645-49}{645-1}} = 110 \pm 2{,}82$$

- С вероятностью 0,95 можно утверждать, что средний недельный заработок служащих фирмы находится в пределах от 107,12 до 112,82 долл.

ОБЩИЕ ПРАВИЛА ПОСТРОЕНИЯ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ (для μ)

- Если объем случайной выборки n такой, что n/N<0,05 и известно стандартное отклонение σ генеральной совокупности, имеющей нормальное распределение, то доверительный интервал с уровнем значимости α имеет вид

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- При тех же условиях, но при n/N>0,05

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

ОБЩИЕ ПРАВИЛА ПОСТРОЕНИЯ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ (для μ)

- Генеральная совокупность распределена нормально, σ неизвестно и n/N<0,05, тогда находится стандартное отклонение выборки s и доверительный интервал для n>30

$$\overline{X} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}}$$

- Если n<30, то

$$\overline{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

ОБЩИЕ ПРАВИЛА ПОСТРОЕНИЯ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ (для μ)

- Если генеральная совокупность имеет нормальный закон распределения, σ неизвестно, n/N>0,05, то для n>30

$$\overline{X} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

- Для n<30

$$\overline{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

ОБЩИЕ ПРАВИЛА ПОСТРОЕНИЯ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ (для μ)

- Если распределение генеральной совокупности отличается от нормального, то в случае, когда σ неизвестно, вышеперечисленные правила могут использоваться только в случае $n>30$

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ РАЗНОСТИ СРЕДНИХ ДВУХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ

- Пусть имеются две разные генеральные совокупности и из каждой случайным образом извлекаются две выборки объемом n_1 и n_2. Если $n_1 > 30$ и $n_2 > 30$, то доверительный интервал для разности средних $\mu_1 - \mu_2$

$$\overline{X}_1 - \overline{X}_2 \pm Z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ РАЗНОСТИ СРЕДНИХ (пример)

- Фирма имеет филиалы в двух разных городах. Руководству необходимо выяснить, как отличаются средние почасовые ставки в этих филиалах. В первом филиале была сделана случайная выборка из 200 рабочих и было подсчитано, что средняя ставка X_1 =8,93 долл., S_1 =0,4. Аналогично для второго филиала n_2 =175, средняя X_2 =9,1 долл., S_2 =0,6.
- Определим 95% доверительный интервал

$$(8{,}83-9{,}1)\pm 1{,}96\sqrt{\frac{0{,}4^2}{200}+\frac{0{,}6^2}{175}}=-0{,}17\pm 0{,}104$$

- Как видно, с уверенностью 95% можно утверждать, что средняя ставка во втором филиале больше средней ставки в первом на 0,07-0,27 долл.

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ РАЗНОСТИ СРЕДНИХ ДВУХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ

- Если какое либо из n_i (i=1,2) меньше 30, то доверительный интервал для разности средних $\mu_1 - \mu_2$

$$\overline{X}_1 - \overline{X}_2 \pm t_{\alpha/2,df}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

- где

$$df = \frac{[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}]^2}{\frac{\frac{S_1^4}{n_1^2}}{n_1 - 1} + \frac{\frac{S_2^4}{n_2^2}}{n_2 - 1}}$$

- Значение df следует округлить до целой части

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ РАЗНОСТИ СРЕДНИХ ДВУХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ

- Если дисперсии двух генеральных совокупностей равны, то доверительный интервал

$$\overline{X}_1 - \overline{X}_2 \pm t_{\alpha/2,df}\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}$$

- где

$$df = n_1 + n_2 - 2$$

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 1}$$

ПРОВЕРКА ГИПОТЕЗ

ГИПОТЕЗЫ

- Под *гипотезой* понимается некое предположение о параметре генеральной совокупности

- Пример (для средней ген.совокупности):
  - средняя месячная оплата за телефон в данном городе равна 52 доллара

- Пример (для доли ген. совокупности):
  - количество дефектных деталей в закупаемой партии не должно превышать 4%

НУЛЕВАЯ ГИПОТЕЗА H_0

- Содержит предположение (количественное) о параметре ген. совокупности
- Всегда выражается в виде, содержащим знаки равенства или нестрогого неравенства, т.е. H_0 можеттбыть записана в одном из видах:

$$H_0 : \mu = \mu_0; H_0 : \mu \leq \mu_0; H_0 : \mu \geq \mu_0$$

АЛЬТЕРНАТИВНАЯ ГИПОТЕЗА H_a

- Противоположна нулевой гипотезе, т.е. содержит предположение о параметре, противоположное предположению нулевой гипотезы
- Всегда записывается в виде строгого неравенства, т.е. в одном из следующих видов

$H_a: \mu > \mu_0$; $H_a: \mu < \mu_0$; $H_a: \mu \neq \mu_0$

ПРОЦЕСС ПРОВЕРКИ ГИПОТЕЗ

- Делается предположение о параметре генеральной совокупности, которое необходимо проверить
- Формулируются H_0 и H_a
- По данным выборки подсчитываются статистики (средняя или доля, стандартное отклонение) и вычисляется критерий теста

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

ПРОЦЕСС ПРОВЕРКИ ГИПОТЕЗ

- Делается предположение о параметре генеральной совокупности, которое необходимо проверить
- Формулируются H_0 и H_a
- По данным выборки подсчитываются статистики (средняя или доля, стандартное отклонение) и вычисляется критерий теста

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

- Задается уровень значимости α, по которому можно судить о степени доверия к полученному выводу
- По заданному α определяется (из таблиц) значение Z табл. (или t табл.)
- Сравниваются Z и Z табл. . Если значение Z (по модулю) меньше значения Z табл. , то нулевая гипотеза не отвергается, т.е. считается правильной. В противном случае нулевая гипотеза отвергается, т.е. правильной считается альтернативная гипотеза

ОШИБКИ В ПРИНЯТИИ РЕШЕНИЯ

- *ошибка первого вида* – непринятие нулевой гипотезы, в то время как она верна. Вероятность ошибки первого вида равна α. Эта вероятность контролируется исследователем

- *ошибка второго вида* – принятие нулевой гипотезы, в то время как она неверна

ВИДЫ ПРОВЕРОК ГИПОТЕЗ

- *Односторонние проверки*:
  - если альтернативная гипотеза записывается в виде $H_a : \mu < \mu\ 0$ или $H_a : \mu > \mu\ 0$
  - $Z_{табл.} = Z_\alpha$, где Z_α характеризует такое значение Z для стандартного нормального распределения, которое отделяет хвостовую часть кривой с долей площади, равной α%
  - величина Z_α задается уровнем значимости α

ВИДЫ ПРОВЕРОК ГИПОТЕЗ

- *Двусторонние проверки*:
  - проводятся в случае, когда альтернативная гипотеза записывается в виде

$$H_a : \mu \neq \mu_0$$

  - $Z_{табл.} = Z_{\alpha/2}$
  - величина $Z_{\alpha/2}$ определяется из таблиц по заданному уровню значимости α

ПРОВЕРКА ГИПОТЕЗ (σ известно) (пример)

- Автоматическая линия разливает напиток по бутылкам объемом 0,5 л. Известно, что объем разливаемой жидкости распределен по нормальному закону с σ = 0,01л. Для проверки работы линии была извлечена партия бутылок в количестве 49. Средний объем оказался равным 0,49 л. Требуется проверить при 5-процентном уровне значимости гипотезу, что средний объем разливаемого напитка равен 0,5л.

ПРОВЕРКА ГИПОТЕЗ (σ известно) (пример)

1. Сформулируем гипотезы

   $H_0: \mu = 0{,}5$

   $H_a: \mu \neq 0{,}5$

2. Подсчитаем статистику теста

   $$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{0{,}49 - 0{,}5}{0{,}01 / \sqrt{49}} = -7$$

3. Так как α=0,05, то табличное значение $Z_{\alpha/2}$ =1,96.
4. Сравниваем значение статистики теста (по модулю) Z с $Z_{табл.}$ Так как 7 > 1,96, то отвергаем нулевую гипотезу.
5. ВЫВОД: объем разливаемой жидкости в среднем не равен 0,5 л, а это значит, что линия требует переналадки. Уверенность в правильности сделанного вывода составляет 95%.

ПРОВЕРКА ГИПОТЕЗ (σ неизвестно)

- На практике значение стандартного отклонения генеральной совокупности, как правило, неизвестно.
- В этом случае его можно заменить на соответствующую статистику – стандартное отклонение выборки S.
  - Если объем выборки большой (n>30), то процедура проверки та же самая, только вместо σ используется S.
  - Если объем выборки мал (n<30), то в случае, когда генеральная совокупность распределена по нормальному закону, используется t-распределение, т.е. вместо $Z_{табл.}$ находится по таблицам $t_{табл.}$

ПРОВЕРКА ГИПОТЕЗ (σ неизвестно) (пример)

- Известно, что средний пробег автошин составляет 40000 км. Был использован новый вариант резины с целью увеличения срока службы. Случайным образом было отобрано 25 шин, которые были протестированы. Результаты проверки показали, что средний пробег новых шин равен 42000 км. со стандартным отклонением S = 1500км.
- Необходимо проверить предположение, что новые шины имеют средний пробег, больший чем 40000 км.
- Проверку необходимо осуществить при 5-процентном уровне значимости (предполагается, что срок службы шин имеет нормальный закон распределения)

ПРОВЕРКА ГИПОТЕЗ (σ неизвестно) (пример)

1. Сформулируем гипотезы

$$H_0: \mu \leq 40000$$
$$H_0: \mu > 40000$$

2. Находим статистику теста

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{42000 - 40000}{1500/\sqrt{25}} = 6{,}67$$

3. По таблицам находим $t_{0,05/24}$ =1,711.
4. Сравниваем статистику теста с $t_{табл.}$. Так как 6,67>1,711, то отвергаем нулевую гипотезу, то есть с уверенностью в 95% утверждаем, что средний пробег новых шин больше чем 40000км.

ПРОВЕРКА ГИПОТЕЗ ДЛЯ ДОЛИ

- *Двусторонняя проверка*:
  - пусть выдвигается нулевая гипотеза о том, что доля генеральной совокупности принимает определенное значение.
- Тогда задача двусторонней проверки имеет следующий вид

$$H_0: p = p_0$$
$$H_a: p \neq p_0$$

- Если выполняются условия $np \geq 5$ и $n(1-p) \geq 5$, то процедура проверки аналогична рассмотренной выше.

ПРОВЕРКА ГИПОТЕЗ ДЛЯ ДОЛИ

- По данным выборки находится статистика теста по формуле

$$Z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$\bar{p}$ - значение доли выборки, n – размер выборки

ПРОВЕРКА ГИПОТЕЗ ДЛЯ ДОЛИ

- По данным выборки находится статистика теста по формуле

$$Z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$\bar{p}$ - значение доли выборки, n – размер выборки

- Значение статистики теста сравнивается со значением $Z_{\alpha/2}$, найденным по таблицам.
- Если $Z<Z_{\alpha/2}$, то нулевая гипотеза не отвергается, если $Z>Z_{\alpha/2}$, то нулевая гипотеза отвергается, т.е. считается верной альтернативная гипотеза.
- В случае односторонней проверки все действия те же самые, за исключением того, что вместо $Z_{\alpha/2}$ из таблиц находится Z_{α}.

ПРОВЕРКА ГИПОТЕЗ ДЛЯ ДОЛИ (пример)

- Доля бракованных деталей не должна превышать 5%. Для проверки наугад было выбрано 400 деталей и среди них обнаружено 25 бракованных. Необходимо проверить соответствует ли это требованиям, в противном случае вся партия должна быть возвращена поставщику.
- Проверку провести при 5-процентном уровне значимости.

ПРОВЕРКА ГИПОТЕЗ ДЛЯ ДОЛИ (пример)

- Прежде всего проверим правомочность проверки. Для этого найдем долю брака в выборке

$$\overline{p} = \frac{25}{400} = 0{,}0625$$

- Подсчитаем np=0,05*400=20 > 5, 400(1-0,05) > 5. Условия проверки выполняются.
- Сформулируем гипотезы

$H_0: p \leq 0{,}05$

$H_a: p > 0{,}05$

ПРОВЕРКА ГИПОТЕЗ ДЛЯ ДОЛИ (пример)

- Подсчитаем статистику теста

$$Z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0{,}0625 - 0{,}05}{\sqrt{\frac{0{,}05(1-0{,}05)}{400}}} = 1{,}157$$

- Сравним значение Z с Z_α =1,64. Так как $Z < Z_\alpha$, то нулевая гипотеза не отвергается, а значит процент бракованных деталей не превышает нормы.