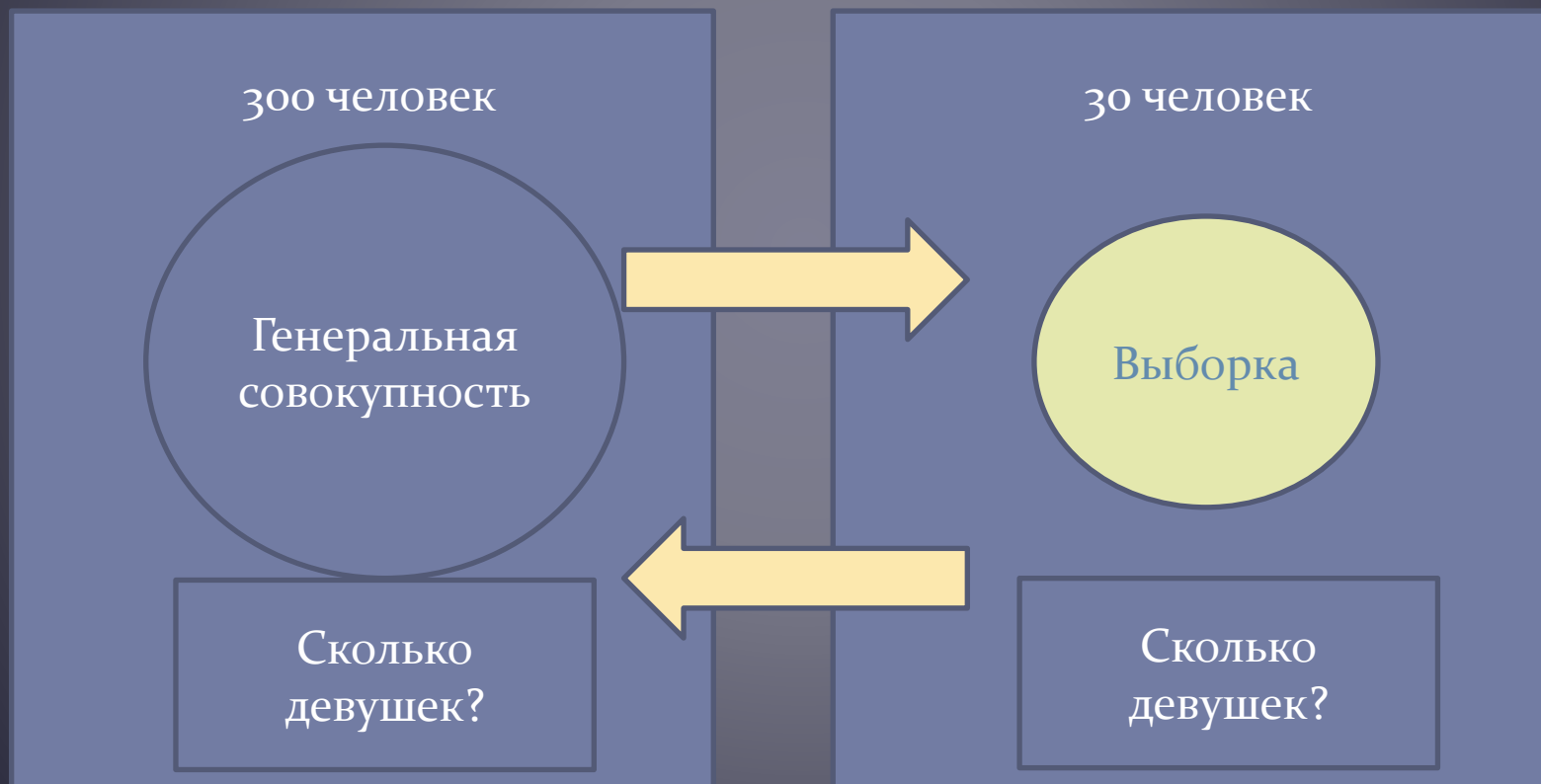


Математическая статистика

Раздел математики, в котором изучаются методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений

- Предметом математической статистики является изучение случайных величин по результатам наблюдений.
- Задачи:
 - 1. упорядочить данные
 - 2. оценить характеристики наблюдаемой величины
 - 3. проверить статистическую гипотезу
- Говорят, что «математическая статистика – это теория принятия решения в условиях неопределенности».

Генеральная совокупность и выборка



выборка

- Повторная
- Бесповторная

Способ
отбора

- Простой
- Типический
- Механический
- Серийный

Пусть из генеральной совокупности извлечена выборка, причем x_1 наблюдалось n_1 раз, $x_2 - n_2$ раз, $x_k - n_k$ раз и $\sum n_i = n$ – объем выборки. Наблюдаемые значения x_i называют **вариантами**, а последовательность вариантов, записанных в возрастающем порядке, – **вариационным рядом**. Числа наблюдений n_i называют **абсолютными частотами**, а их отношения к объему выборки $n_i / n = w_i$ – **относительными частотами** или **частностями**.

Соответствие, установленное между наблюдаемыми вариантами и их частотами (абсолютными или относительными), называют статистическим распределением.

При этом должны выполняться два условия нормировки:

1) $n_1 + n_2 + \dots + n_k = n$ (объем выборки);

2) $w_1 + w_2 + \dots + w_k = 1$.

Удобной формой записи статистического распределения является таблица. В верхней строке таблицы записывают последовательность вариантов, в нижней – соответствующие им частоты (абсолютные или относительные).

Пример 1. Имеются данные о количестве дежурств сотрудниками кафедры за месяц. Произведена выборка объемом $n = 15$:

3 0 5 7 4 3 1 9 5 3 4 4 2 8 5.

Составить статистический вариационный ряд распределения частот (абсолютных и относительных).

Решение

1. Расположить значения выборки в возрастающем порядке:

0 1 2 3 3 3 4 4 4 5 5 5 7 8 9.

Имеем девять различных значений.

2. Найти абсолютные частоты появления каждого значения выборки:

$n_1 = 1, n_2 = 1, n_3 = 1, n_4 = 3, n_5 = 3, n_6 = 3, n_7 = 1, n_8 = 1, n_9 = 1.$

Проверить первое условие нормировки:

$$n = \sum_{i=1}^9 n_i = 1 + 1 + 1 + 3 + 3 + 3 + 1 + 1 + 1 = 15$$

3. Вычислить относительные частоты появления каждого значения выборки по формуле $w_i = n_i / n$:

$w_1^* = 1/15, w_2 = 1/15, w_3 = 1/15, w_4 = 3/15, w_5 = 3/15, w_6 = 3/15, w_7 = 1/15, w_8 = 1/15, w_9 = 1/15.$

Проверить второе условие нормировки:

$$W = \sum_{i=1}^9 w_i^* = \frac{1}{15} + \frac{1}{15} + \frac{1}{15} + \frac{3}{15} + \frac{3}{15} + \frac{3}{15} + \frac{1}{15} + \frac{1}{15} + \frac{1}{15} = 1$$

4. Внести полученные данные в таблицу:

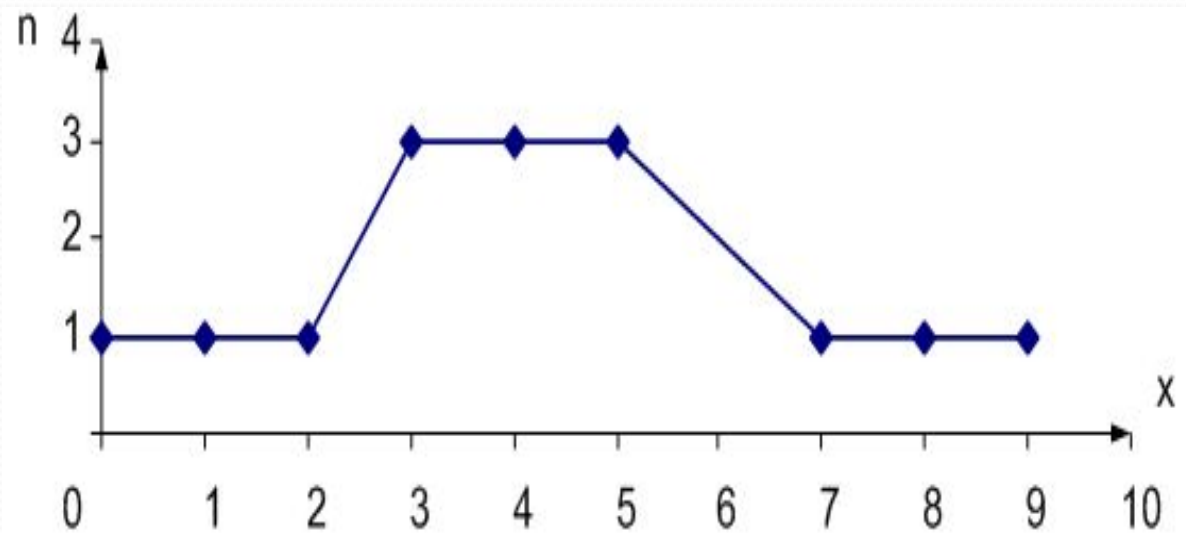
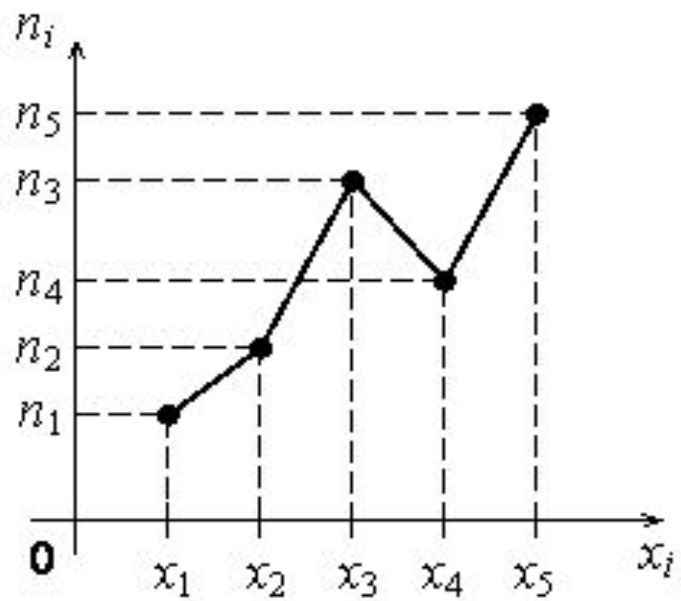
X_i	0	1	2	3	4	5	7	8	9
n_i	1	1	1	3	3	3	1	1	1
w_i	1/15	1/15	1/15	3/15	3/15	3/15	1/15	1/15	1/15

ПОЛИГОН

Для геометрического изображения такого статистического распределения служит *полигон частот* или *полигон относительных частот*.

Полигоном частот называют ломаную линию, отрезки, которой соединяют точки $(x_1; n_1), (x_2; n_2), \dots, (x_k; n_k)$. Для построения полигона частот на оси абсцисс откладывают варианты x_i , а на оси ординат — соответствующие им частоты n_i .

Полигоном относительных частот называют ломаную линию, отрезки которой соединяют точки $(x_1; w_1), (x_2; w_2), \dots, (x_k; w_k)$. Для построения полигона частот на оси абсцисс откладывают варианты x_i , а на оси ординат — соответствующие им относительные частоты p_i .



Гистограмма

Гистограммой частот называется ступенчатая фигура, основанием i -го прямоугольника которой являются частичные интервалы длиной Δ_i , и высотой n_i .

Для построения гистограммы частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс на расстоянии n_i .

В практике для удобства вычислений обычно используют ряды с равными интервалами (Δ), которые называют шагом интервала.

Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиной Δ_i , а высоты равны отношению w_i .

Построение гистограммы

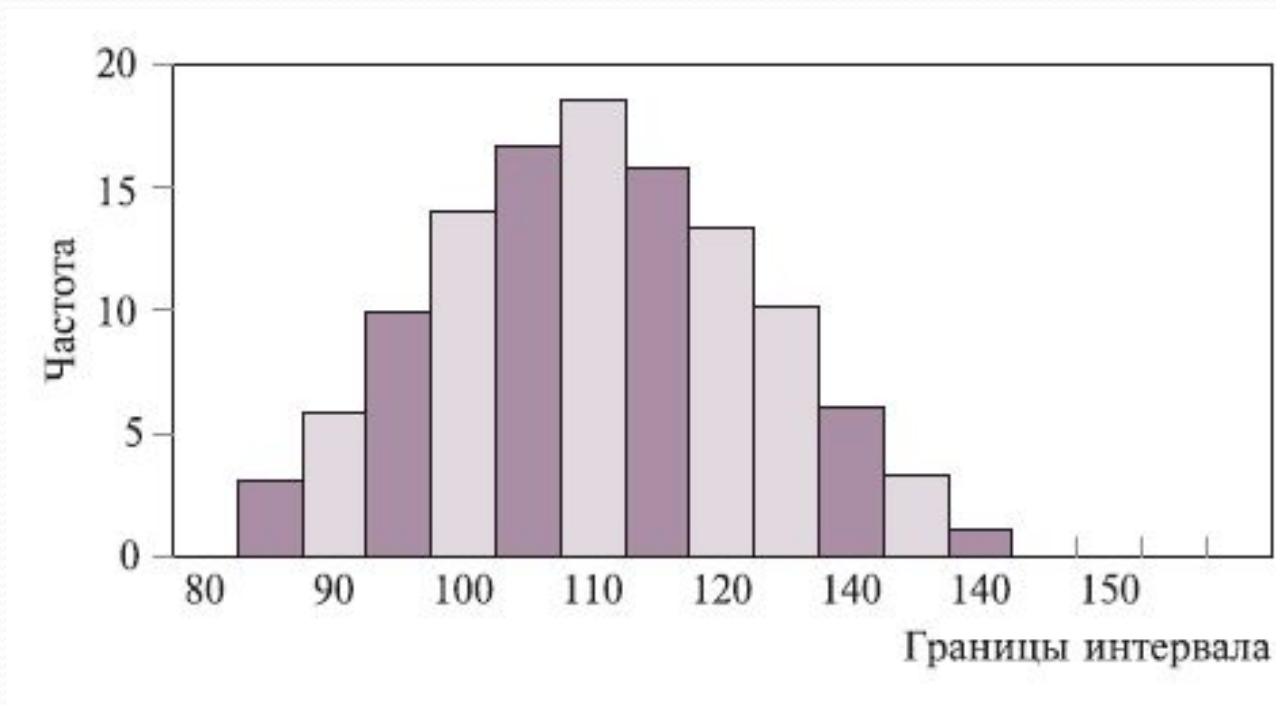
Порядок построения гистограммы

1. Собрать данные, выявить максимальное и минимальное значения и определить диапазон (размах) гистограммы.
2. Полученный диапазон разделить на интервалы, предварительно определив их число (обычно 5-20 в зависимости от числа показателей) и определить ширину интервала.

$$\Delta = (x_{\max} - x_{\min}) / k$$

3. Все данные распределить по интервалам в порядке возрастания: левая граница первого интервала должна быть равна наименьшему из имеющихся значений.
4. Подсчитать частоту каждого интервала.
5. Вычислить относительную частоту попадания данных в каждый из интервалов.
6. По полученным данным построить гистограмму - столбчатую диаграмму, высота столбиков которой соответствует частоте или относительной частоте попадания данных в каждый из интервалов:

Гистограмма нормального распределения

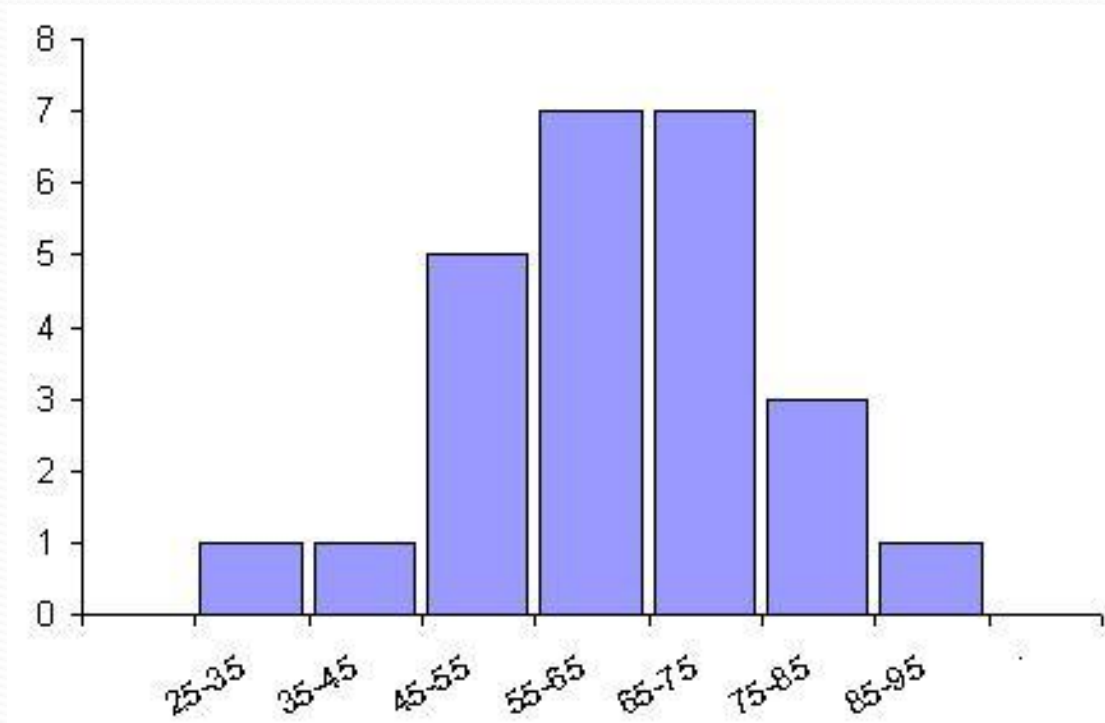


Пример.2. По результатам тестирования по анатомии студентов 2-го курса получены данные о доступности заданий теста (отношение числа студентов, правильно выполнивших задания, к числу тестируемых студентов), представленные ниже, в таблице. Тест содержал 25 заданий. Получены следующие данные: 25, 37, 46, 46, 50, 54, 55, 57, 58, 60, 60, 61, 64, 65, 66, 66, 67, 70, 71, 72, 75, 77, 85, 85, 95. Построить гистограмму, распределив данные в 7 интервалов.

Доступность задания x , %	25-35	35-45	45-55	55-65	65-75	75-85	85-95
Количество задач n	1	5	5	7	7	3	1

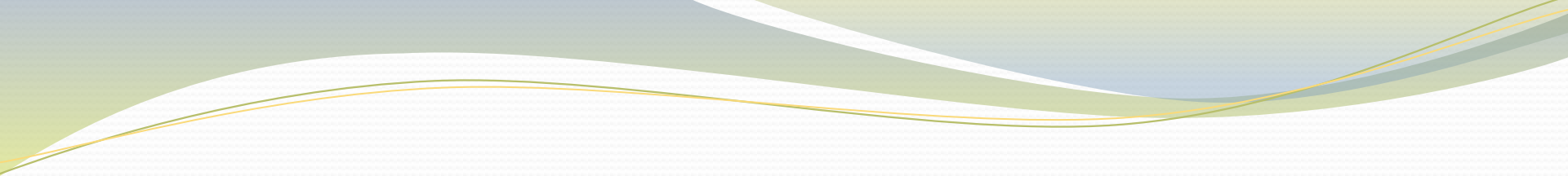
Решение.

Откладываем на оси абсцисс 7 отрезков длиной 10. На них, как на основаниях, строим прямоугольники, высоты которых соответственно равны 1, 1, 5, 7, 7, 3, 1. Полученная ступенчатая фигура и является искомой гистограммой.



Статистические оценки

- Оценка
 - Точечная
 - Интервальная

- 
- Точечная
 - смещенная
 - несмещенная
 - эффективная
 - состоятельная
 - Оценка
 - Хар-ка

Пусть изучается дискретная генеральная совокупность относительно количественного признака. **Генеральной средней** называется среднее арифметическое значений признака генеральной совокупности. Она вычисляется по формуле

$$\bar{x}_2 = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{или} \quad \bar{x}_2 = \frac{1}{n} \sum_{i=1}^n x_i m_i$$

где x_i — значения признака генеральной совокупности объема n ; m_i — соответствующие частоты, причем

$$\sum_{i=1}^n m_i = n$$

Если генеральная средняя неизвестна и требуется оценить ее по данным выборки, то в качестве оценки генеральной средней принимают выборочную среднюю, которая является несмещенной и состоятельной оценкой. Отсюда следует, что если по нескольким выборкам достаточно большого объема из одной и той же генеральной совокупности будут найдены выборочные средние, то они будут приблизительно равны между собой. В этом состоит свойство **устойчивости выборочных средних**.

$$\bar{x}_v = \bar{x}_z$$

Для того чтобы охарактеризовать рассеяние значений количественного признака X генеральной совокупности вокруг своего среднего значения, вводят сводную характеристику D_g — генеральную дисперсию. **Генеральной дисперсией** называется среднее арифметическое квадратов отклонений значений признака генеральной совокупности от их среднего значения, которое вычисляется по формуле

$$D_g = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$D_g = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 m_i$$

Для того чтобы охарактеризовать рассеяние наблюдаемых значений количественного признака выборки вокруг своего среднего значения \bar{x}_B , вводят сводную характеристику D_B — выборочную дисперсию. **Выборочной дисперсией** называется среднее арифметическое квадратов отклонений наблюдаемых значений признака от их среднего значения, которое вычисляется по формуле

$$D_B = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_B)^2$$

$$D_B = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_B)^2 m_i$$

Кроме дисперсии для характеристики рассеяния значений признака генеральной (выборочной) совокупности вокруг своего среднего значения используют сводную характеристику — среднее квадратическое отклонение.

Генеральным средним квадратическим отклонением называют квадратный корень из генеральной дисперсии:

$$\sigma_z = \sqrt{D_z}$$

Выборочным средним квадратическим отклонением называют квадратный корень из выборочной дисперсии:

$$\sigma_v = \sqrt{D_v}$$

Пусть из генеральной совокупности в результате n независимых наблюдений над количественным признаком x извлечена выборка объема n . Требуется по данным выборки оценить неизвестную генеральную дисперсию D_{σ} . Если в качестве оценки генеральной дисперсии принять выборочную дисперсию, то эта оценка приведет к систематическим ошибкам, давая заниженное значение генеральной дисперсии. Объясняется это тем, что выборочная дисперсия является смещенной оценкой D_{σ} . Другими словами, математическое ожидание выборочной дисперсии не равно оцениваемой генеральной дисперсии, а равно $\frac{n-1}{n} D_{\sigma}$.

$$M(D_{\sigma}) = \frac{n-1}{n} D_{\sigma}$$

Легко исправить выборочную дисперсию так, чтобы ее математическое ожидание было равно генеральной дисперсии. Для этого нужно умножить D_{σ} на дробь $\frac{n}{n-1}$.

В результате получим исправленную дисперсию S^2 , которая будет несмещенной оценкой генеральной дисперсии:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_{\sigma})^2 m_i$$

Интервальные оценки

Задачу интервального оценивания можно сформулировать так: по данным выборки построить числовой интервал, относительно которого с заранее выбранной вероятностью можно сказать, что внутри него находится оцениваемый параметр. Интервальное оценивание особенно необходимо при малом количестве наблюдений, когда точечная оценка малонадежна.

Доверительным интервалом $\tilde{\theta}_n^{(1)} \tilde{\theta}_n^{(2)}$ для параметра θ называется такой интервал, относительно которого с заранее выбранной вероятностью $p=1-\alpha$, близкой к единице, можно утверждать, что он содержит неизвестное значение параметра θ , то есть $P\{\tilde{\theta}_n^{(1)} < \theta < \tilde{\theta}_n^{(2)}\} = 1 - \alpha$. Чем меньше для выбранной вероятности число $|\tilde{\theta}_n^{(1)} - \tilde{\theta}_n^{(2)}|$, тем точнее оценка неизвестного параметра θ . И, наоборот, если это число велико, то оценка, проведенная с помощью данного интервала, малопригодна для практики.

Так как концы доверительного интервала зависят от элементов выборки, то значения $\tilde{\theta}_n^{(1)}$ и $\tilde{\theta}_n^{(2)}$ могут изменяться от выборки к выборке. Вероятность принято называть **доверительной** (надежностью). Обычно надежность оценки задается наперед, причем в качестве α берут число, близкое к единице. Выбор доверительной вероятности не является математической задачей, а определяется конкретной решаемой проблемой. Наиболее часто задают надежность, равную 0,95; 0,99; 0,999.

Доверительный интервал для генеральной средней нормального распределения признака при неизвестном значении среднего квадратического отклонения задается выражением

$$(\bar{x}_2 - \delta < \mu < \bar{x}_2 + \delta)$$

$$\delta = t_{\alpha, n} \frac{S}{\sqrt{n}}$$

$t_{\alpha, n}$

Коэффициент Стьюдента

Определение необходимого объема выборки для получения оценок заданной точности

При планировании выборочного наблюдения с заранее заданным значением допустимой ошибки выборки необходимо правильно оценить требуемый **объем выборки**. Этот объем может быть определен на основе допустимой ошибки при выборочном наблюдении исходя из заданной вероятности p , гарантирующей допустимую величину уровня ошибки (с учетом способа организации наблюдения). Формулы для определения необходимой численности выборки n легко получить непосредственно из формул предельной ошибки выборки. Так, из выражения для предельной ошибки:

$$\delta = t \sqrt{\frac{S^2}{n}}$$

непосредственно вычисляется необходимый объем выборки n :

$$n = \frac{t^2 S^2}{\delta^2}$$

- Поясним смысл, который имеет заданная надежность. Надежность $\gamma=0,95$ указывает, что если произведено достаточно большое число выборок, то 95% из них определяет такие доверительные интервалы, в которых параметр действительно заключен, лишь в 5 % случаев он может выйти за границы доверительного интервала.