МАТЕМАТИЧЕСКАЯ СТАТИСТИКА(МС)

Математическая статистика изучает и разрабатывает методы сбора, обработки и интерпретации статистической информации для получения научных и практических выводов.

Теоретической основой математической статистики являются законы распределения и предельные теоремы закона больших чисел.

ВЫБОРОЧНЫЙ МЕТОД

Определение 1. Вся совокупность единиц, подвергаемых обследованию на какой-либо признак X, называется **генеральной совокупностью (Г.С.).**

Количество единиц генеральной совокупности называется **объемом Г.С. и** обозначается **N.**

Различают **сплошное и несплошное** обследования единиц Г.С. на некоторый признак X.

К **сплошному** наблюдению (обследованию) относятся, например, перепись населения, медосмотр студентов 1-го курса, ЕГЭ по русскому языку всех без исключения выпускников школ и т.д.

Виды **несплошного** наблюдения:

1) анкетное;
2) обследование основного массива;
3) монографическое (на какой-либо отдельный признак);
4) выборочное.

Определение 2. Единицы, отобранные из Г.С. для обследования на некоторый признак, образуют **выборочную совокупность** **(в.с.)** или **выборку**.

Количество единиц **в.с.** называется объемом **в.с.** и обозначается **n**.

Выборочное обследование применяется в тех случаях, когда:

1) Г.С. очень велика;

2) время, выделенное для обследования, ограничено;

3) обследование связано с уничтожением обследуемых объектов;

4) необходимо проверить точность сплошного наблюдения.

Преимущество выборочного метода в том, что:

1) он позволяет экономить силы, средства и время, т.е. является более дешевым и быстрым;

2) позволяет оперативно вмешиваться в ход процесса и вносить коррективы на промежуточных этапах;

3) является более точным и объективным.

Различают 2 способа отбора единиц в **в.с.:**

1) **повторный;**

2) **бесповторный.**

При повторном отборе отобранная единица регистрируется и после обследования на признак X возвращается в Г.С. и может участвовать в обследовании на другие признаки.

При бесповторном отборе отобранная единица регистрируется и после обследования на признак X не возвращается в Г.С.

Виды отбора:

не требующие разбиения Г.С. на группы:

1. Собственно-случайный отбор;
2. Механический отбор;

требующие разбиения Г.С. на группы:

3. Типический отбор;
4. Серийный отбор.

От того, правильно ли организован отбор единиц в в.с., зависят точность и качество результатов и выводов выборочного обследования.

При собственно-случайном отборе каждая единица Г.С. имеет равные шансы попасть в выборку: единицы Г.С. регистрируются, снабжаются номером и участвуют в жеребьевке.

При механическом отборе единицы Г.С. упорядочиваются, и в соответствии с процентом отбора извлекается определенное количество единиц, например, при 25%-й выборке отбирается каждая четвертая единица.

Типический отбор применяется в тех случаях, когда Г.С. неоднородна по составу. Тогда по некоторому признаку Г.С. разбивается на однородные типические группы и из каждой группы собственно-случайным или механическим способами извлекаются единицы в в.с.

Серийный отбор применяется в тех случаях, когда Г.С. однородна по составу. Тогда вся Г.С. разбивается на группы или **серии** и среди этих серий собственно-случайным или механическим способами извлекаются единицы в в.с.

Выборка должна быть **репрезентативной** (представительной),т.е. должна правильно отражать исследуемый признак Г.С.

Выборочный метод решает следующие задачи:

1) Организация выборочной совокупности;
2) Вычисление числовых характеристик (параметров) в.с.;
3) Оценка параметров Г.С. и выводы о них;
4) Определение необходимой численности выборки.

Статистическое распределение выборки

Пусть из Г.С. объемом **N** отобрана выборка объема **n**, которая обследуется на некоторый признак X (например, з/плата рабочих, % жирности молока, диаметр деталей и т. д.).

Определение[BT1]**.** Значения признака X: x_1, x_2,…, …, x_n называются вариантами, а упорядоченная последовательность вариант называется вариационным рядом.

Вариационный ряд можно задавать как в виде последовательности значений x_i, так и поинтервально.

Определение[BT1]**.** Число повторений варианты $\mathbf{x_i}$ называется ее частотой и обозначается $\mathbf{m_i}$, причем

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

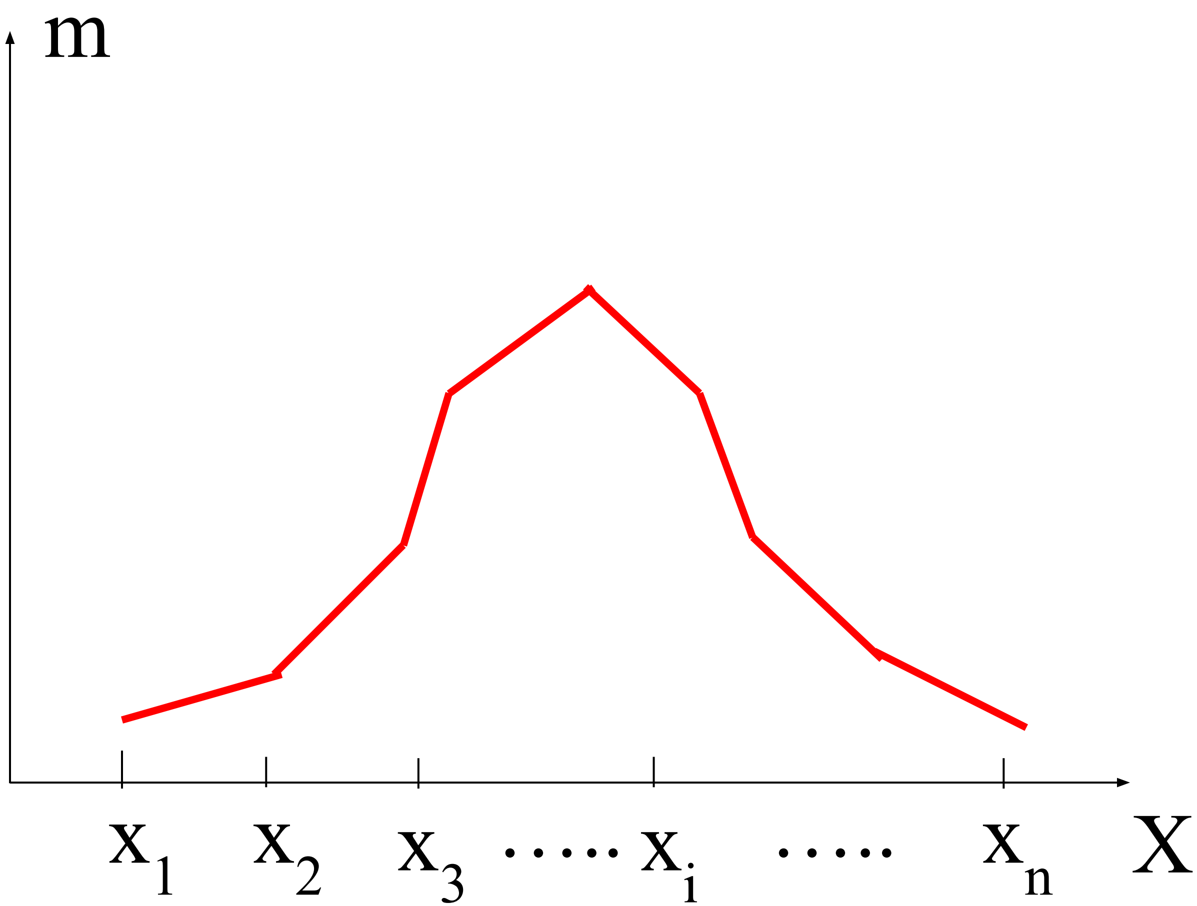
Последовательность частот $\mathbf{m_i}$ называется частотным рядом.

Определение[ВТ1]**.** Соответствие между вариационным и частотным рядами назыв. статистическим распределением выборки.

Способы задания статистического распределения выборки: **табличный и графический**.

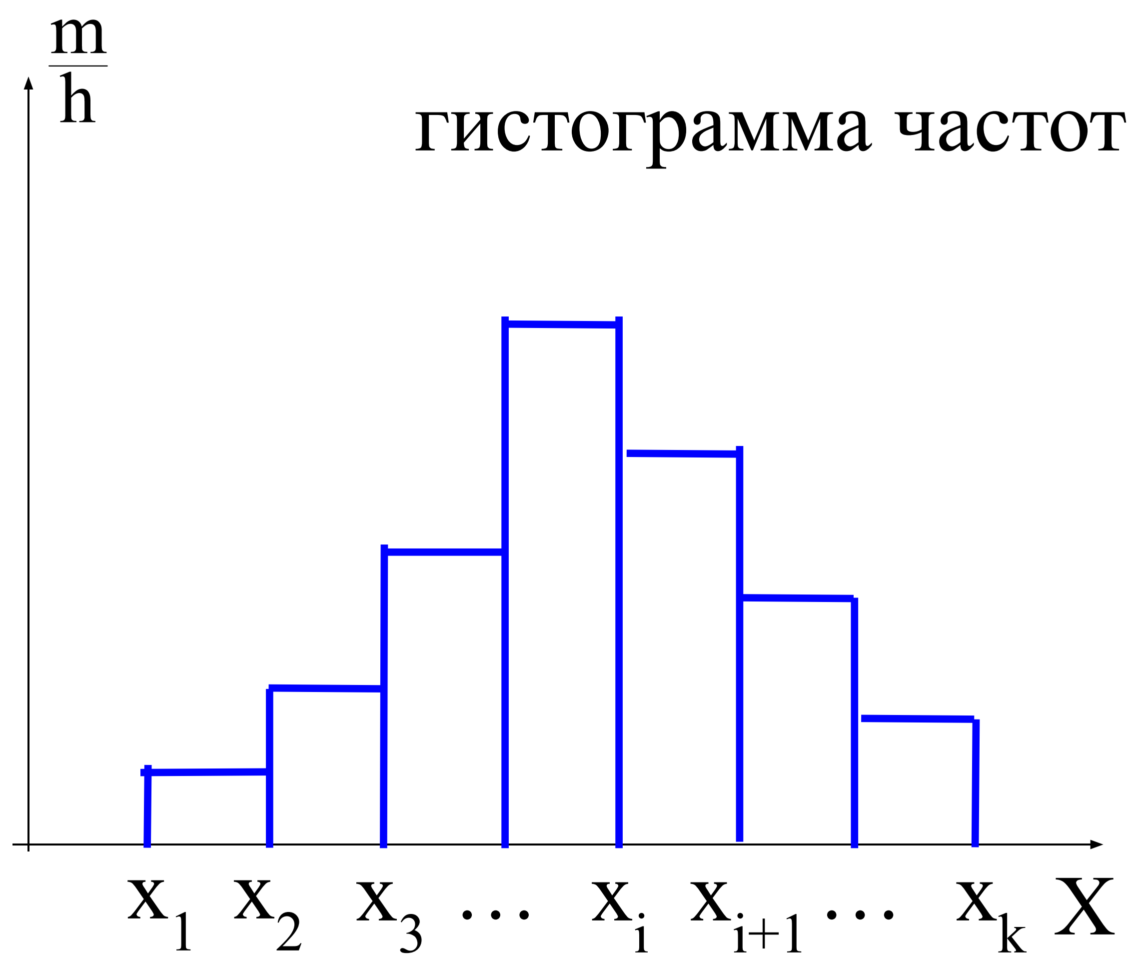
а) вариационный ряд задается в виде последовательности вариант:

X	m
x_1	m_1
x_2	m_2
……	…….
x_i	m_i
……	…….
x_n	m_n



б) вариационный ряд задается в виде последовательности интервалов:

X	m
$x_1 - x_2$	m_1
$x_2 - x_3$	m_2
....	......
$x_i - x_{i+1}$	m_i
....	......
$x_k - x_{k+1}$	m_k



$$h = \Delta x_i = x_{i+1} - x_i \quad (i = \overline{1,N}), \quad \frac{m_i}{h}$$ - высоты прям-ков

Числовые характеристики в.с.

Определение. Числовые характеристики (или статистики) – это параметры, которые в сжатой форме отражают особенности Г.С.

К ним относятся:

1) **выборочная средняя** – среднее взвешенное значение признака в.с.:

$$\boxed{\tilde{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{k} x_i m_i}, \quad \text{где} \quad n = \sum_{i=1}^{k} m_i$$

2) **выборочная дисперсия** – среднее взвешенное квадратов отклонений значений признака от среднего значения:

$$\boxed{\boldsymbol{\sigma}^2(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{k}(x_i - x)^2 m_i}$$

выборочная дисперсия – мера колеблемости значений признака около среднего значения.

3) **выборочная доля** – доля единиц в в.с., обла-дающих тем или иным признаком:

$$\boxed{w = \frac{m}{n}}$$

где **m** – число единиц в в.с., обладающих исследуемым признаком, **n** – объем выборки. Тогда (n – m) – число единиц в.с., не обладающих этим признаком, отсюда:

$$1 - w = 1 - \frac{m}{n} = \frac{n-m}{n}$$

Числовые характеристики Г.С.

1) **Генеральная средняя** – среднее взвешенное значение признака в Г.С.:

$$\boxed{\bar{x} = \frac{1}{N}\sum_{i=1}^{L} x_i M_i}, \qquad \text{где} \qquad N = \sum_{i=1}^{L} M_i$$

2) **Генеральная дисперсия** – дисперсия признака в Г.С.:

$$\boxed{D(X)=\sigma_0^2(X)=\frac{1}{N}\sum_{i=1}^{L}(x_i-x)^2 M_i}$$

3) **Генеральная доля** – доля единиц, обладающих тем или иным признаком в Г.С.:

$$\boxed{p=\frac{M}{N}},$$

где M – число единиц, обладающих этим признаком в Г.С.

Тогда $q = 1 – p = 1 - \dfrac{M}{N}$

- доля единиц, не обладающих этим признаком в Г.С.

Характеристики в.с. отличаются от соответствующих характеристик Г.С.

Определение. Отклонение характеристик в.с. от соответствующих характеристик Г.С. называется **ошибкой репрезентативности** или **ошибкой выборки**.

Определение. **Средняя ошибка** репрезентативности **μ** показывает, на сколько в среднем параметры в.с. отклонятся от соответствующих параметров Г.С.

Определение. Предельная ошибка репрезентативности **Δ** показывает наибольшее отклонение характеристики в.с. от соответствующей характеристики Г.С.

Предельная и средняя ошибки выборки связаны между собой соотношением:

$$\boxed{\Delta = t\mu,}$$

где **t** называется **коэффициентом надежности** или **коэффициентом достоверности**.

Оценки параметров распределения

Какую-либо характеристику Г.С. ($\bar{x}$, σ_0^2 или w) обозначим через $\bar{\theta}$, а соответствующую характеристику в.с. – через $\tilde{\theta}$.

Определение 1. Оценка генеральной характеристики, заданная одним числом, называется **точечной.**

Определение 2. Оценка генеральной характеристики, заданная двумя числами, называется **интервальной.**

Определение 3. Интервал

$$\boxed{\tilde{\theta} - \Delta_\theta \leq \bar{\theta} \leq \tilde{\theta} + \Delta_\theta}$$

или $\boxed{|\overline{\theta} - \tilde{\theta}| \le \Delta_\theta}$ называется **доверитель-ным интервалом** для $\overline{\theta}$, а Δ_θ называется **точностью оценки.**

Определение 4. Вероятность того, что $\overline{\theta}$ попадет в доверительный интервал, называется **надежностью** γ оценки $\overline{\theta}$, т.е.:

$$\boxed{\mathbf{P}(|\overline{\theta} - \tilde{\theta}| \le \Delta_\theta) = \gamma}$$

Требования к числовым характеристикам

Пусть из Г.С. объема **N** извлекаются всевозможные повторные выборки объемов **n** и вычисляются $\tilde{\theta}_i$ – значения $\tilde{\theta}$, затем находят $M(\tilde{\theta})$

Определение. Оценка $\tilde{\theta}$ называется **несмещенной**, если при любом объеме выборки:

$$\boxed{M(\tilde{\theta}) = \overline{\theta}.}$$

Если же $M(\tilde{\theta}) \neq \overline{\theta}$, то оценка называется смещенной.

Определение. **Состоятельной** называется статистическая оценка $\tilde{\theta}$, которая при $\mathbf{n} \longrightarrow \infty$ сходится по вероятности к оцениваемому параметру $\overline{\theta}$:

$$\boxed{\lim_{n \to \infty} P\left(\left| \overline{\theta} - \tilde{\theta} \right| \leq \Delta \right) = 1}$$

Если имеются несколько несмещенных оценок $\tilde{\theta}$, вычисленных при одинаковых $\mathbf{n}$, то лучшая из них – та, которая имеет минимальный разброс.

Определение. **Эффективной** называется несмещенная состоятельная оценка $\tilde{\theta}$, которая при заданном **n** имеет минимальную дисперсию $D(\tilde{\theta})$.

Теорема 1. Математическое ожидание среднего арифметического одинаково распределенных СВ равна математическому ожиданию каждой из этих СВ, а дисперсия среднего арифметического **n** СВ в **n** раз меньше дисперсии каждой из них:

$$M\left(\frac{X_1+X_2...+X_n}{n}\right)=M(X_i)$$

$$D\left(\frac{X_1+X_2...+X_n}{n}\right)=\frac{1}{n}D(X_i)$$

$$\sigma\left(\frac{X_1+X_2...+X_n}{n}\right)=\frac{1}{\sqrt{n}}\sigma(X_i)$$

Теорема 2. Выборочная средняя $\tilde{\mathbf{x}}$ является **несмещенной и состоятельной оценкой** генеральной средней $\bar{\mathbf{x}}$.

Теорема 3. Выборочная доля $\mathbf{w}$ является несмещенной и состоятельной оценкой генеральной доли $\mathbf{p}$.

Теорема 4. Исправленная выборочная дисперсия:

$$\boxed{S^2(X) = \frac{n}{n-1}\,\sigma^2(X)}$$

является несмещенной и состоятельной оценкой генеральной дисперсии $D(X) = \sigma_0{}^2(X)$.

Теорема Чебышева-Ляпунова

(для средней)

С вероятностью, равной $\mathbf{\Phi(t) = \gamma}$, где $\mathbf{t}$ – коэффициент надежности, можно утверждать, что отклонение генеральной средней от выборочной средней по абсолютной величине не превзойдет предельной ошибки выборки Δ_x:

$$\boxed{P(\left| \overline{x} - \tilde{x} \right| \le \Delta_x) = \Phi(t) = \gamma,}$$

где **предельная ошибка выборки для средней при собственно-случайном отборе:**

$$\boxed{\Delta_{x\text{повт}} = t\sqrt{\frac{\sigma_x^2}{n}}}$$ - для повторного отбора;

$$\boxed{\Delta_{x\text{бесповт}} = t\sqrt{\frac{\sigma_x^2}{n}\left(1 - \frac{n}{N}\right)}}$$ - для бесповторного отбора

Значение $\boldsymbol{\gamma = \Phi(t)}$ отыскивается в таблице Приложения 2.

Теорема Чебышева-Ляпунова

(для доли)

С вероятностью, равной $\boldsymbol{\Phi(t) = \gamma}$ можно утверждать, что абсолютная величина отклонения генеральной доли от выборочной доли не превзойдет предельной ошибки выборки $\boldsymbol{\Delta_w}$:

$$\boxed{\mathbf{P}(|\,\mathbf{p} - \mathbf{w}\,| \leq \boldsymbol{\Delta}_{\mathbf{w}}) = \boldsymbol{\Phi}(\mathbf{t}) = \boldsymbol{\gamma},}$$

где предельная ошибка выборки для доли при собственно-случайном отборе:

$$\boxed{\Delta_{\text{wповт}} = t\sqrt{\frac{w(1-w)}{n}}}$$ - для повторного отбора;

$$\boxed{\Delta_{\text{wбесп}} = t\sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}}$$ -для бесповторного отбора

Из теорем Чебышева-Ляпунова следует, что с вероятностью γ = **Ф(t)** **доверительный интервал для генеральной средней:**

$$\boxed{\tilde{x} - \Delta_x \leq \bar{x} \leq \tilde{x} + \Delta_x}\ ,$$

а **доверительный интервал для генеральной доли:**

$$\boxed{w - \Delta_w \leq p \leq w + \Delta_w}$$

Из формулы:

$$\Delta_x = t\sqrt{\frac{\sigma_x^2}{n}}$$

следует, что, чем больше объем в.с. **n**, тем меньше предельная ошибка $\Delta_{x,}$ а значит, и средняя ошибка

$$\boldsymbol{\mu}_{\mathbf{x}} = \sqrt{\frac{\sigma_x^2}{n}}$$

С другой стороны, чем выше надежность **Ф(t)**, тем больше коэффициент надежности **t** и тем больше ошибка $\boldsymbol{\Delta}_{\mathbf{x}}$. При бесповторном отборе:

$$\boldsymbol{\Delta}_{\mathbf{x\text{бесповт}}} < \boldsymbol{\Delta}_{\mathbf{x\text{повт}},}$$

а значит, доверительный интервал точнее, чем при повторном отборе. Такие же выводы справедливы для ошибок $\boldsymbol{\Delta}_{\mathbf{w}}$ по доле.

Приведенные выше формулы предельных ошибок для средней и для доли имеют место при собственно- случайном отборе. Все формулы, приведенные для собственно- случайного отбора, справедливы и для механического способа отбора.

Сводка формул для собственно- случайного и механического способов отбора

N – объем Г.С., **n** – объем в.с.

Характеристики в.с.

$\tilde{x} = \frac{1}{n}\sum_{i=1}^{k} x_i m_i$ - выборочная средняя, где k – число вариант вариационного ряда

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

$$\sigma_x{}^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \widetilde{x})^2 m_i$$ - выборочная дисперсия

$$w = \frac{m}{n}$$ - выборочная доля

Ошибки репрезентативности(выборки)

Для средней: $$\mathbf{\Delta}_{\mathbf{xповт}} = \mathbf{t}\sqrt{\frac{\sigma_x^2}{n}}, \quad \mathbf{\Delta}_{\mathbf{xбесп}} = \mathbf{t}\sqrt{\frac{\sigma_x^2}{n}\left(1 - \frac{n}{N}\right)}$$

$$\boldsymbol{\mu}_{\mathbf{xповт}} = \sqrt{\frac{\sigma_x^2}{n}}, \quad \boldsymbol{\mu}_{\mathbf{xбесп}} = \sqrt{\frac{\sigma_x^2}{n}\left(1 - \frac{n}{N}\right)}$$

Для доли: $$\mathbf{\Delta}_{\mathbf{wповт}} = \mathbf{t}\sqrt{\frac{\sigma_w^2}{n}}, \quad \mathbf{\Delta}_{\mathbf{wбесп}} = \mathbf{t}\sqrt{\frac{\sigma_w^2}{n}\left(1 - \frac{n}{N}\right)}$$

$$\mu_{\text{wповт}} = \sqrt{\frac{\sigma_w^2}{n}}\,, \qquad \mu_{\text{wбесп}} = \sqrt{\frac{\sigma_w^2}{n}\left(1-\frac{n}{N}\right)}$$

Доверительные интервалы

Для средней: $\tilde{\mathbf{x}} - \Delta_{\mathbf{x}} \le \bar{\mathbf{x}} \le \tilde{\mathbf{x}} + \Delta_{\mathbf{x}}$

Для доли: $\mathbf{w} - \Delta_{\mathbf{w}} \le \mathbf{p} \le \mathbf{w} + \Delta_{\mathbf{w}}$

Сводка формул для типического способа отбора

N – объем Г.С., n – объем в.с.;

l – количество типических групп;

k_j – число вариант в j–й типической группе$(j = \overline{1, l})$

$\mathbf{N_j}$ **- объем j**–й типической группы в Г.С.;

n_j - объем **j**–й типической группы в в.с.;

$$n_j = \sum_{i=1}^{kj} m_i , \qquad \mathbf{n} = \sum_{j=1}^{l} n_j$$

Внутригрупповые характеристики
j–й типической группы

$$\tilde{x}_j = \frac{1}{n_j}\sum_{i=1}^{k_j} x_i m_i$$ - внутригрупповая средняя j–й типической группы;

$$\sigma_j^2 = \frac{1}{n_j}\sum_{j=1}^{k_j}(x_i - \tilde{x}_j)^2 m_j$$ - внутригрупповая дисперсия;

$$w_j = \frac{m_j}{n_j}$$ - внутригрупповая доля

<u>Общие(межгрупповые) характеристики в.с.</u>

$$\tilde{x} = \frac{1}{n}\sum_{j=1}^{1} \tilde{x}_j n_j$$ - общая(межгрупповая) средняя;

$$\sigma_x{}^2 = \frac{1}{n}\sum_{j=1}^{1} \sigma_j{}^2 n_j$$ - общая дисперсия по средней;

$$w = \frac{1}{n}\sum_{j=1}^{1} w_j n_j$$ - общая доля;

$$\sigma_w{}^2 = \frac{1}{n}\sum_{j=1}^{1} w_j(1 - w_j) n_j$$ - общая дисперсия по доле;

Предельные ошибки выборки

Для средней: $\Delta_{\text{хповт}} = t\sqrt{\frac{\sigma_x^2}{n}}$, $\Delta_{\text{хбесповт}} = t\sqrt{\frac{\sigma_x^2}{n}(1-\frac{n}{N})}$

Для доли: $\Delta_{\text{wповт}} = t\sqrt{\frac{\sigma_w^2}{n}}$, $\Delta_{\text{wбесп}} = t\sqrt{\frac{\sigma_w^2}{n}(1-\frac{n}{N})}$

Доверительные интервалы записываются так же, как и при собственно-случайном отборе.

Сводка формул для серийного способа отбора

S – общее число серий в Г.С.
s – число серий, отобранных в в.с.

j – номер серии, $j = \overline{1, s}.$

$\tilde{x}_j$ – средняя в j –й серии;

w_j – доля в j –й серии;

Общие характеристики в.с.

$$\tilde{\mathbf{x}} = \frac{1}{s}\sum_{j=1}^{s}\tilde{x}_j; \qquad \sigma_x^{\,2} = \frac{1}{s}\sum_{j=1}^{s}(\tilde{x}_j - \tilde{x})^2$$

$$w = \frac{1}{s}\sum_{j=1}^{s}w_j \quad ; \qquad \sigma_w^{\,2} = \frac{1}{s}\sum_{j=1}^{s}(w_j - w)^2$$

Предельные ошибки

$$\Delta_{x\ \text{повт}} = \mathbf{t} \cdot \sqrt{\frac{\sigma_x^2}{s}}, \qquad \Delta_{x\ \text{повт}} = \mathbf{t} \cdot \sqrt{\frac{\sigma_x^2}{s}},$$

$$\Delta_{x\ \text{повт}} = \mathbf{t} \cdot \sqrt{\frac{\sigma_x^2}{s}}, \qquad \Delta_{x\ \text{повт}} = \mathbf{t} \cdot \sqrt{\frac{\sigma_x^2}{s}},$$