СТАТИСТИКА

Аналитическая статистика.

Лекция 2. Выборочное наблюдение.

Автор: Равичев Л.В.
РХТУ им. Д.И.Менделеева
Кафедра управления технологическими инновациями
Москва - 2013

Выборочное наблюдение

Под ***выборочным наблюдением*** понимается такое ***несплошное*** наблюдение, при котором статистическому обследованию (наблюдению) повергаются единицы изучаемой совокупности, отобранные ***случайным*** образом.

Совокупность отобранных для обследования единиц в статистике принято называть ***выборочной,*** а совокупность единиц, из которых производится отбор, - ***генеральной.***

Выборочное наблюдение

№	Характеристика	Генеральная совокупность	Выборочная совокупность
1	Объем совокупности (численность единиц)	N	n
2	Численность единиц, обладающих обследуемым признаком.	M	m
3	Доля единиц, обладающих обследуемым признаком.	$P = \frac{M}{N}$	$W = \frac{m}{n}$
4	Средний размер признака.	$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$	$\widetilde{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
5	Дисперсия количесвенного признака.	$\sigma_{\overline{x}}^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N}$	$S_{\widetilde{x}}^2 = \frac{\sum_{i=1}^{n} (x_i - \widetilde{x})^2}{n}$
6	Дисперсия альтернативного признака.	$\sigma_{an}^2 = p \cdot q$	$S_{an}^2 = W(1-W)$

Ошибка выборочного наблюдения

Ошибка выборочного наблюдения – представляет собой разность между величиной параметра в генеральной совокупности и его величиной, вычисленной по результатам выборочного наблюдения.

Предельная ошибка выборки:

$$\Delta_{\tilde{x}} = | \bar{x} - \tilde{x} |$$

где:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\tilde{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Теорема П.Л.Чебышева

При достаточно большом числе независимых наблюдений с вероятностью, близкой к единице, можно утверждать, что отклонение выборочной средней от генеральной будет сколь угодно малым. При этом величина предельной ошибки выборки не должна превышать tμ.

$$\Delta_{\tilde{x}} \leq t\mu_{\tilde{x}}$$

где μ_x - средняя ошибка выборки:

$$\mu_{\tilde{x}} = \frac{\sigma}{\sqrt{n}}$$

Теорема А.М.Ляпунова

Распределение выборочных средних (а следовательно, и их отклонений от генеральной средней) при достаточно большом числе независимых наблюдений приближенно нормально при условии, что генеральная совокупность обладает конечной средней и ограниченной дисперсией.

$$P\{|\bar{x}-\tilde{x}|\le\Delta_{\tilde{x}}\}=\frac{1}{\sqrt{2\pi}}\cdot\int_{-t}^{t}e^{-\frac{t^2}{2}}dt=F(t)$$

где:

$$\Delta_{\tilde{x}}=\pm t\mu$$

Предельная ошибка выборки дает возможность выяснить, в каких пределах находится величина генеральной средней.

Теорема А.М.Ляпунова

Значение интеграла F(t) для различных значений **коэффициента доверия** t в специальных математических таблицах:

Целые и десятые доли t	Сотые доли t						
	0	1	2	3	…	8	9
0,0	0,0000	0,0080	0,0160	0,3988	…	0,0638	0,0718
0,1	0,0797	0,0876	0,3961	0,3956	…	0,3925	0,3918
…	…	…	…	…	…	…	…
2,1	0,9643	0,9651	0,9660	0,9698	…	0,9707	0,9715
…	…	…	…	…	…	…	…
5,0	0,9999999	-	-	-	-	-	-

Полученное значение F(t) = 0,9698 показывает, что в 96,98% случаев разность между выборочной и генеральной средней не превысит 2,13*μ.

Зная выборочную среднюю величину признака и предельную ошибку выборки можно определить границы интервала, в котором заключена генеральная средняя:

$$\tilde{x} - \Delta_{\tilde{x}} \le \bar{x} \le \tilde{x} + \Delta_{\tilde{x}}$$

Расчет предельной ошибки выборки

Расчет значений предельной ошибки выборки может быть произведен с помощью стандартной функции Excel ***ДОВЕРИТ.***

ДОВЕРИТ(p;σ;n)

Пример. В результате выборочного обследования жилищных условий жителей города на основе собственно-случайной повторной выборки, получен ряд распределения:

Общая площадь, приходящаяся на 1 человека, $м^2$	До 5	5-10	10-15	15-20	20-25	25-30	30 и более
Число жителей	8	95	204	270	210	130	83

Требуется с уровнем надежности 95% определить границы интервала, в который попадает средний размер общей площади.

Расчет предельной ошибки выборки

	B	C	D	E	F
2	Общая площадь, приходящаяся на 1 человека, м^2	Середина интервала, x	Число жителей, f	$(x - \bar{x})^2$	
3	До 5	2,5	8	272,42	=(C3-D11)^2
4	5-10	7,5	95	132,37	
5	10-15	12,5	204	42,32	
6	15-20	17,5	270	2,27	
7	20-25	22,5	210	12,22	
8	25-30	27,5	130	72,17	
9	30 и более	32,5	83	182,12	
10	Число жителей в выборочной совокупности, n		1000	=СУММ(D3:D9)	
11	Выборочная средняя		19,01	=СУММПРОИЗВ(C3:C9;D3:D9)/D10	
12	Дисперсия		51,11	=СУММПРОИЗВ(E3:E9;D3:D9)/(D10-1)	
13	Стандартное отклонение		7,15	=КОРЕНЬ(D12)	
14	Средняя ошибка выборки		0,23	=D13/КОРЕНЬ(D10)	
15	Коэффициент доверия, t		1,96	=НОРМСТОБР((0,95+1)/2)	
16	Предельная ошибка выборки		0,44	=D15*D14	
17	Предельная ошибка выборки (через ДОВЕРИТ)		0,44	=ДОВЕРИТ(0,05;D13;D10)	

Теорема Бернулли

При достаточно большом объеме выборки вероятность расхождения между долей признака в выборочной совокупности (w) и долей признака в генеральной совокупности (p) будет стремиться к единице.

$$P\{|w-p|\le t\mu\}\to 1$$

т.е. с вероятностью, сколь угодно близкой к 1, можно утверждать, что при достаточно большом объеме выборки частость признака (выборочная доля) сколь угодно мало будет отличаться от доли признака в генеральной совокупности.

Средняя ошибка выборки для альтернативного признака:

$$\mu_w=\sqrt{\frac{pq}{n}}\approx\sqrt{\frac{w(1-w)}{n}}$$

Теорема Бернулли

Предельная ошибка выборки альтернативного признака:

$$\Delta_w = t\mu_w$$

Доверительный интервал альтернативного признака:

$$w - \Delta_w \le p \le w + \Delta_w$$

Уточнение формулы средней ошибки выборки

Если отбор единиц из генеральной совокупности произведен **бесповторным** способом, т.е. способом при котором попавшая в выборку единица не возвращается в совокупность, то в формулы средней ошибки выборки вносится поправка:

$$\sqrt{1-\frac{n}{N}}$$

то есть:

$$\mu_{\tilde{x}} = \sqrt{\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)}$$

$$\mu_w = \sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}$$

Уточнение формулы средней ошибки выборки

Для приведенного выше примера, если предположить, что данные являются результатом бесповторного выбора из генеральной совокупности из 20000 единиц:

$$\mu_{\tilde{x}} = \sqrt{\frac{51{,}11}{1000}\left(1 - \frac{1000}{20000}\right)} = 0{,}22$$

При большом проценте выборке влияние поправки на бесповторность значительно возрастает.

$$\mu_{\tilde{x}} = \sqrt{\frac{51{,}11}{1000}\left(1 - \frac{1000}{10000}\right)} = 0{,}21$$

$$\mu_{\tilde{x}} = \sqrt{\frac{51{,}11}{1000}\left(1 - \frac{1000}{2000}\right)} = 0{,}16$$

Предельная ошибка альтернативного признака

Для приведенного выше примера, определим предельную ошибку выборки для лиц, обеспеченность жильем которых составляет менее 10 м^2.

1. Выборочная доля:

$$w = \frac{103}{1000} = 0{,}103$$

2. Дисперсия:

$$\sigma_w^{\,2} = w(1-w) = 0{,}103 \cdot 0{,}897 = 0{,}0924$$

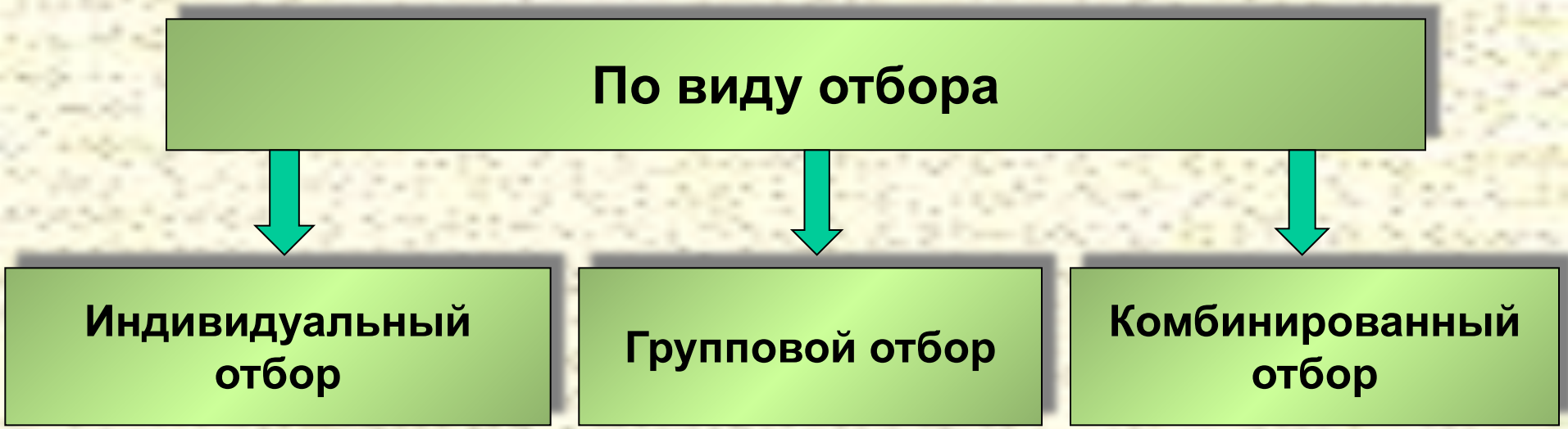
3. Средняя ошибка выборки:

$$\mu_w = \sqrt{\frac{0{,}0924}{1000}\left(1-\frac{1000}{20000}\right)} = 0{,}0094$$

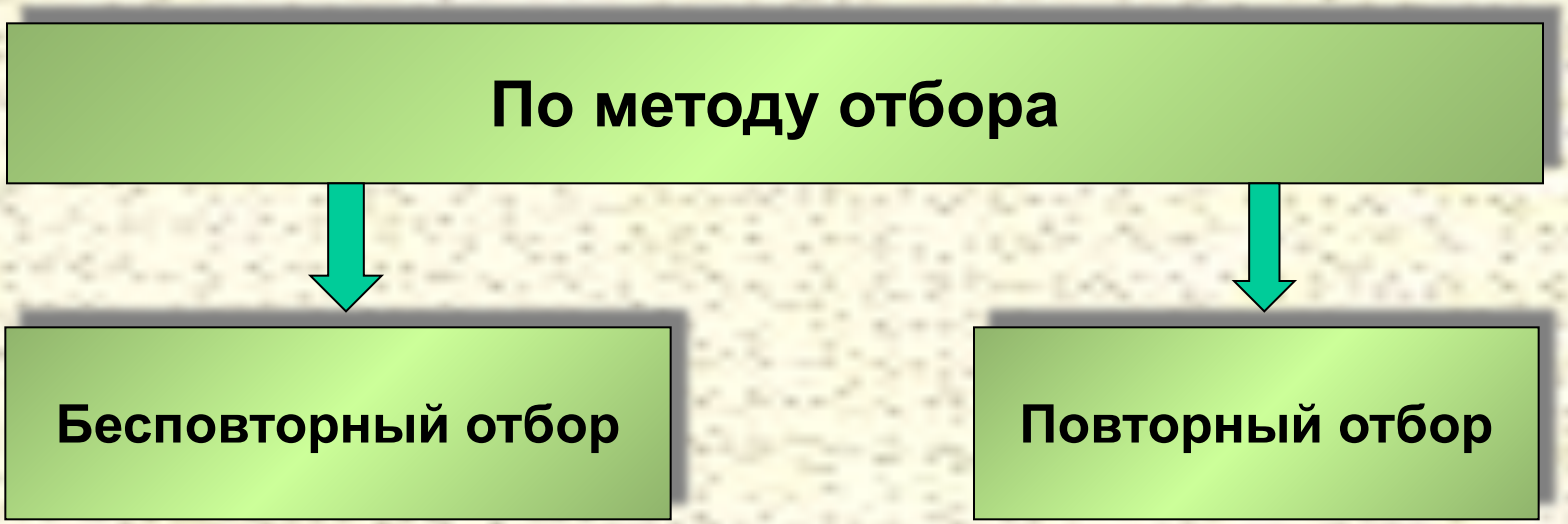
4. Предельная ошибка выборки:

$$\Delta_w = 1{,}96 \cdot 0{,}0094 = 0{,}0184$$

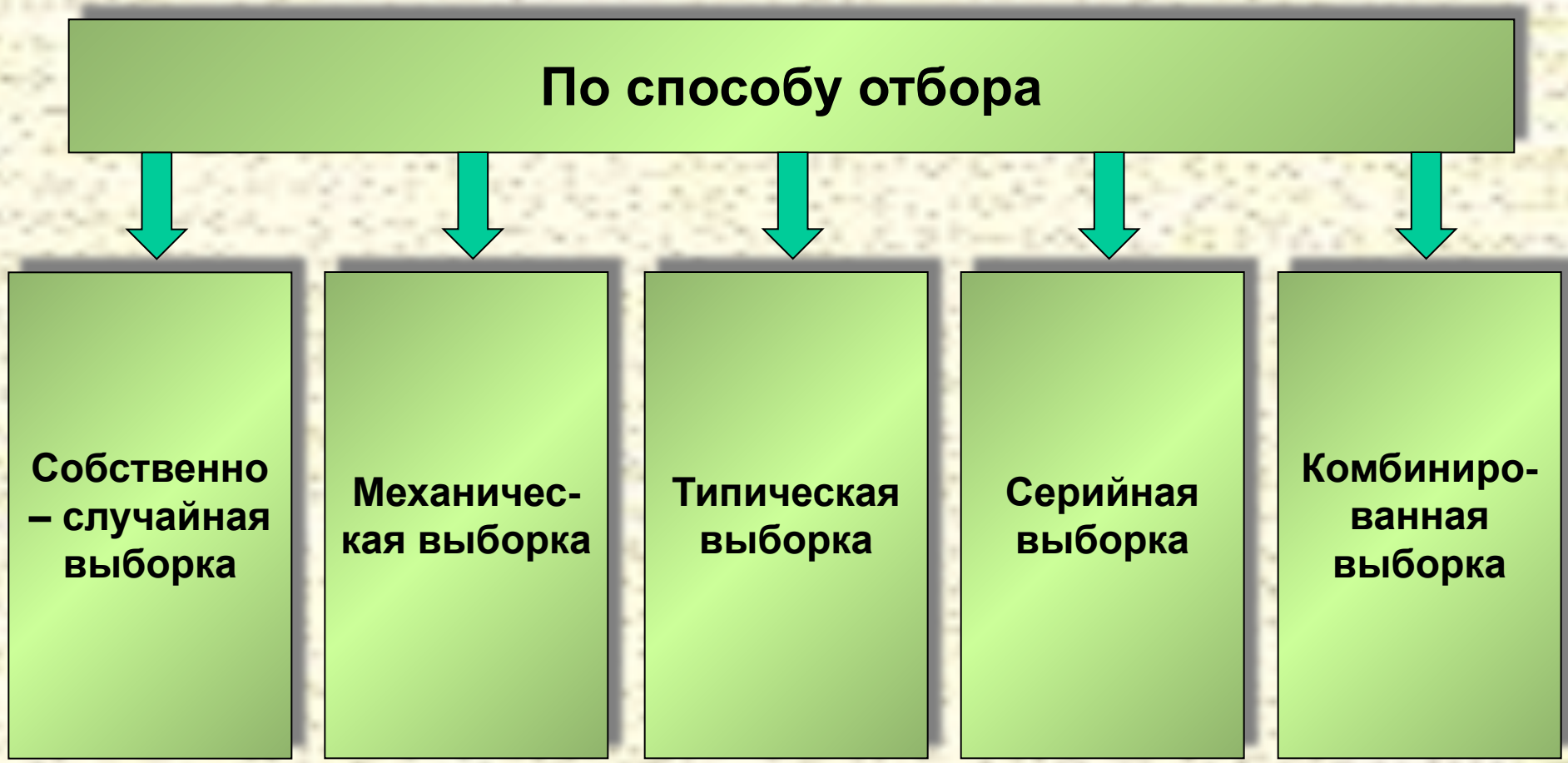
Способы формирования выборочной совокупности



Способы формирования выборочной совокупности



Способы формирования выборочной совокупности



Типическая выборка

Выборка, пропорционально объему типических групп.

1. Число единиц, подлежащих отбору из каждой группы:

$$n_i = n\frac{N_i}{N}$$

2. Средняя ошибка выборки:

повторный отбор

$$\mu = \sqrt{\frac{\overline{\sigma}^2}{n}}$$

бесповторный отбор

$$\mu = \sqrt{\frac{\overline{\sigma}^2}{n}\left(1-\frac{n}{N}\right)}$$

Выборка, пропорционально дифференциации признака.

1. Число единиц, подлежащих отбору из каждой группы:

$$n_i = n\frac{\sigma_i N_i}{\sum \sigma_i N_i}$$

2. Средняя ошибка выборки:

повторный отбор

$$\mu = \frac{1}{N}\sqrt{\sum \frac{\sigma_i^2 N_i^2}{n_i}}$$

бесповторный отбор

$$\mu = \frac{1}{N}\sqrt{\sum \frac{\sigma_i^2 N_i^2}{n_i}\left(1-\frac{n_i}{N_i}\right)}$$

Типическая выборка

Пример. 10%-ный бесповторный типический отбор рабочих предприятия, пропорциональный размерам цехов, проведенный с целью оценки потерь из-за временной нетрудоспособности, привел к следующим результатам:

Цех	Всего рабочих	Обследовано человек	Число дней временной нетрудоспособности за год	
			средняя	дисперсия
1	1000	100	18	49
2	1400	140	12	25
3	800	80	15	16

Необходимо определить пределы среднего числа дней временной нетрудоспособности одного рабочего в целом по предприятию.

Типическая выборка

1. Расчет пропорционально объему типических групп.

Средняя из внутригрупповых дисперсий:

$$\overline{\sigma}^2 = \frac{\sum \sigma_i^2 n_i}{\sum n_i} = \frac{49 \cdot 100 + 25 \cdot 140 + 16 \cdot 80}{100 + 140 + 80} = 30{,}25$$

Средняя и предельная ошибки выборки (с вероятностью 0,954):

$$\mu_{\tilde{x}} = \sqrt{\frac{30{,}25}{320}\left(1 - \frac{320}{3200}\right)} = 0{,}29$$

$$\Delta_{\tilde{x}} = 2 \cdot 0{,}29 = 0{,}58$$

Выборочная средняя:

$$\tilde{x} = \frac{\sum x_i n_i}{\sum n_i} = \frac{18 \cdot 100 + 12 \cdot 140 + 15 \cdot 80}{100 + 140 + 80} = 14{,}6$$

$$14{,}6 - 0{,}58 \le \overline{x} \le 14{,}6 + 0{,}58$$

Типическая выборка

2. Расчет пропорционально дифференциации признака.

Необходимый объем выборки по каждому цеху:

$$\sum \sigma_i N_i = \sqrt{49} \cdot 1000 + \sqrt{25} \cdot 1400 + \sqrt{16} \cdot 800 = 17200$$

$$n_1 = 320 \cdot \frac{\sqrt{49} \cdot 1000}{17200} = 130$$

$$n_2 = 320 \cdot \frac{\sqrt{25} \cdot 1400}{17200} = 130$$

$$n_3 = 320 \cdot \frac{\sqrt{16} \cdot 800}{17200} = 60$$

Средняя и предельная ошибки выборки (с вероятностью 0,954):

$$\mu_{\tilde{x}} = 0,28$$

$$\Delta_{\tilde{x}} = 2 \cdot 0,28 = 0,56$$

$$14,6 - 0,56 \le \bar{x} \le 14,6 + 0,56$$

Серийная выборка

Средняя ошибки выборки:

повторный отбор

$$\mu = \sqrt{\frac{D_{мг}}{r}}$$

бесповторный отбор

$$\mu = \sqrt{\frac{D_{мг}}{r}\left(1 - \frac{r}{R}\right)}$$

Межгрупповая дисперсия:

$$D_{мг} = \frac{\sum(\tilde{x}_i - \tilde{x})^2}{r}$$

Определение необходимого объема выборки

Вид выборочного наблюдения	Повторный отбор	Бесповторный отбор
Собственно-случайная и механическая выборки		
а) при определении среднего размера признака	$n = \frac{t^2 \cdot \sigma_{\tilde{x}}^2}{\Delta_{\tilde{x}}^2}$	$n = \frac{t^2 \cdot \sigma_{\tilde{x}}^2 \cdot N}{\Delta_{\tilde{x}}^2 \cdot N + t^2 \cdot \sigma_{\tilde{x}}^2}$
б) при определении доли признака	$n = \frac{t^2 \cdot W(1-W)}{\Delta_W^2}$	$n = \frac{t^2 \cdot W(1-W) \cdot N}{\Delta_W^2 \cdot N + t^2 \cdot W(1-W)}$
Типическая выборка		
а) при определении среднего размера признака	$n = \frac{t^2 \cdot \overline{\sigma}_{\tilde{x}}^2}{\Delta_{\tilde{x}}^2}$	$n = \frac{t^2 \cdot \overline{\sigma}_{\tilde{x}}^2 \cdot N}{\Delta_{\tilde{x}}^2 \cdot N + t^2 \cdot \overline{\sigma}_{\tilde{x}}^2}$
б) при определении доли признака	$n = \frac{t^2 \cdot \overline{W(1-W)}}{\Delta_W^2}$	$n = \frac{t^2 \cdot \overline{W(1-W)} \cdot N}{\Delta_W^2 \cdot N + t^2 \cdot \overline{W(1-W)}}$

Определение необходимого объема выборки

Вид выборочного наблюдения	Повторный отбор	Бесповторный отбор
Серийная выборка		
а) при определении среднего размера признака	$r = \frac{t^2 \cdot D_{мг}}{\Delta^2_{\bar{x}}}$	$r = \frac{t^2 \cdot D_{мг} \cdot R}{\Delta^2_{\bar{x}} \cdot R + t^2 \cdot D_{мг}}$
б) при определении доли признака	$r = \frac{t^2 \cdot W_r(1-W_r)}{\Delta^2_W}$	$r = \frac{t^2 \cdot W_r(1-W_r) \cdot R}{\Delta^2_W \cdot R + t^2 \cdot W_r(1-W_r)}$

Определение необходимого объема выборки

Пример 1. В микрорайоне проживает 5000 семей. В порядке случайной бесповторной выборки предполагается определить средний размер семьи при условии, что ошибка выборочной средней не должна превышать 0,8 человека с вероятностью P=0,954 и при среднем квадратичном отклонении 3,0 человека.

$$n = \frac{t^2 \cdot \sigma_{\tilde{x}}^2 \cdot N}{\Delta_{\tilde{x}}^2 \cdot N + t^2 \cdot \sigma_{\tilde{x}}^2} = \frac{2^2 \cdot 3^2 \cdot 5000}{0{,}64 \cdot 5000 + 2^2 \cdot 3^2} = \frac{180000}{3236} \approx 56$$

Пример 2. Для определения средней длины детали следует провести выборочное обследование методом случайного повторного отбора. Какое количество деталей надо отобрать, чтобы ошибка выборки не превышала 3 мм с вероятностью 0,997 при среднем квадратическом отклонении 6 мм.

$$n = \frac{t^2 \cdot \sigma_{\tilde{x}}^2}{\Delta_{\tilde{x}}^2} = \frac{3^2 \cdot 6^2}{3^2} = 36$$

Определение необходимого объема выборки

Пример 3. В фермерских хозяйствах области 10 000 коров. Из них в районе А – 5000, в районе Б – 3000, в районе В - 2000. Чтобы определить средний надой предполагается провести типическую выборку коров с пропорциональным отбором внутри групп (механическим). Какое количество коров следует отобрать, чтобы с вероятностью 0,954 ошибка выборки не превышала 5 л, если на основе предыдущих обследований известно, что дисперсия типической выборки равна 1600?

$$n = \frac{t^2 \cdot \overline{\sigma}_{\tilde{x}}^2 \cdot N}{\Delta_{\tilde{x}}^2 \cdot N + t^2 \cdot \overline{\sigma}_{\tilde{x}}^2} = \frac{2^2 \cdot 1600 \cdot 10000}{5^2 \cdot 10000 + 2^2 \cdot 1600} \approx 250$$

Нужно отобрать 250 коров, из них

в районе А: $n_1 = 250 \cdot \frac{5000}{10000} = 125$

в районе Б: $n_2 = 250 \cdot \frac{3000}{10000} = 75$

в районе В: $n_3 = 250 \cdot \frac{2000}{10000} = 50$

Определение необходимого объема выборки

Пример 4. На склад поступило 100 ящиков деталей по 80 шт. в каждом. Для установления среднего веса деталей следует провести серийную выборку деталей методом механического отбора так, чтобы с вероятностью 0,954 ошибка выборки не превышала 2 г. На основе предыдущих обследований известно, что дисперсия серийной выборки равна 4. Определить необходимый объем выборки.

$$r = \frac{t^2 \cdot D_{мг} \cdot R}{\Delta_{\bar{x}}^2 \cdot R + t^2 \cdot D_{мг}} = \frac{2^2 \cdot 4 \cdot 100}{2^2 \cdot 100 + 2^2 \cdot 4} \approx 4$$

Методики, разработанные в рамках конкретных обследований и определенных способов формирования выборочной совокупности, требу-ют дальнейшего теоретического обоснования и практической провер-ки.

Малая выборка
Распределение Стьюдента

Под **малой выборкой** понимается такое выборочное наблюдение, численность единиц которого не превышает 30.

Критерий Стьюдента:

$$t = \frac{\tilde{x} - \bar{x}}{\mu_{мв}}$$

где:

$$\mu_{мв} = \frac{\sigma}{\sqrt{n-1}}$$

мера случайных колебаний выборочной средней в малой выборке.

$$\sigma = \sqrt{\frac{\sum (x_i - \tilde{x})}{n}}$$

$$\Delta_{МВ} = t \cdot \mu_{МВ}$$

Малая выборка
Распределение Стьюдента

Способы нахождения критерия Стьюдента.

1. С помощью таблиц распределения Стьюдента (t - распределение):

Число степеней свободы k=n-1	Уровень значимости					
	0,9	0,8	…	0,02	0,01	0,001
1	0,158	0,325	…	31,821	63,657	636,619
2	0,142	0,289	…	6,965	9,925	31,589
…	…	…	…	…	…	…
9	0,129	0,261	…	2,821	3,250	4,781
…	…	…	…	…	…	…
30	0,127	0,256	…	2,457	2,750	3,646
…	…	…	…	…	…	…
120	0,126	0,254	…	2,358	2,617	3,373
∞	0,126	0,253	…	2,326	2,576	3,291

Малая выборка
Распределение Стьюдента

2. С помощью стандартной функции Excel **СТЬЮДРАСПОБР**.

СТЬДРАСПОБР(p;k).

Для расчета t – распределения, т.е. значения уровня значимости при известных значениях t и k, необходимо воспользоваться стандартной функцией Excel **СТЬЮДРАСП**.

СТЬДРАСП(t;k;r).

где r может принимать два значения : 1 или 2. При r=1 функция **СТЬЮДРАСП** рассчитывает одностороннее t – распределение, при r=2, двустороннее t – распределение.

Малая выборка
Распределение Стьюдента

Пример. При контрольной проверке качества поставленного в торговлю маргарина получены следующие данные о содержании консерванта Е205 в 10 пробах, %: 4,3; 4,2; 3,8; 4,3; 3,7; 3,9; 4,5; 4,4; 4,0; 3,9. Определить вероятность того, что среднее содержание консерванта Е205 во всей партии не выйдет за пределы 0,1% его среднего содержания в представленных пробах.

	A	B	C	D	E	F	G	H	I	J
1	Содержание консерванта Е205 в пробах, %									
2	4,3	4,2	3,8	4,3	3,7	3,9	4,5	4,4	4,0	3,9
3										
4	Предельная ошибка выборки, %				0,1					
5	Число степеней свободы				9	=СЧЁТ(A2:J2)-1				
6	Выборочная средняя				4,1	=СРЗНАЧ(A2:J2)				
7	Нижняя граница интервала				4,0	=E5-E4				
8	Верхняя граница интервала				4,2	=E5+E4				
9	Стандартное отклонение				0,261	=СТАНДОТКЛОНП(A2:J2)				
10	Средняя ошибка выборки				0,087	=E9/КОРЕНЬ(10-1)				
11	Коэффициент доверия				1,150	=E4/E10				
12	Уровень значимости				0,280	=СТЬЮДРАСП(E11;E5;2)				
13	Доверительная вероятность				0,720	=1-E12				
14										
15	Коэффициент доверия (через СТЬЮДРАСПОБР)				1,150	=СТЬЮДРАСПОБР(E12;E5)				
16										