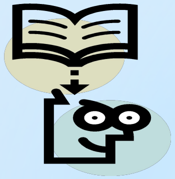
Выборочный метод в статистике (II)



Вопросы:

1. Средняя ошибка выборки для доли.
2. Виды ошибок выборки, распространение результатов выборочного наблюдения на генеральную совокупность, понятие репрезентативности выборки.
3. Определение необходимой численности выборки.
4. Малая выборка.

В выборочной совокупности дисперсия для доли рассчитывается по формуле:

$$\sigma^2 = w \cdot (1 - w)$$

Межсерийная дисперсия для доли рассчитывается по формуле:

$$\delta_w^2 = \frac{\sum (w_i - \widetilde{w})^2}{s}$$ **где** $$\widetilde{w} = \frac{\sum w_i}{s}$$

w_i – **доля единиц, обладающих признаком в i–$й$ серии выборочной совокупности**

<table>
<tr><th colspan="2" rowspan="2">Способ формирования выборочной совокупности</th><th colspan="2">Средняя ошибка выборки</th></tr>
<tr><th>Для среднего значения</th><th>Для доли</th></tr>
<tr><td rowspan="2">Собственно случайный</td><td>повторный</td><td></td><td></td></tr>
<tr><td>бесповторный</td><td rowspan="2"></td><td rowspan="2"></td></tr>
<tr><td colspan="2">Механический</td></tr>
<tr><td colspan="2">Типический (пропорциональный)</td><td></td><td></td></tr>
<tr><td colspan="2">Серийный</td><td></td><td></td></tr>
</table>

На основе выборочного обследования точно оценить изучаемый параметр генеральной совокупности нельзя. Можно только определить пределы, в которых он будет находиться:

$$\tilde{x} - \Delta \leq \bar{x} \leq \tilde{x} + \Delta$$

$$w - \Delta \leq p \leq w + \Delta$$

Δ – ***величина отклонения выборочной характеристики от генеральной называется предельной ошибкой выборки***

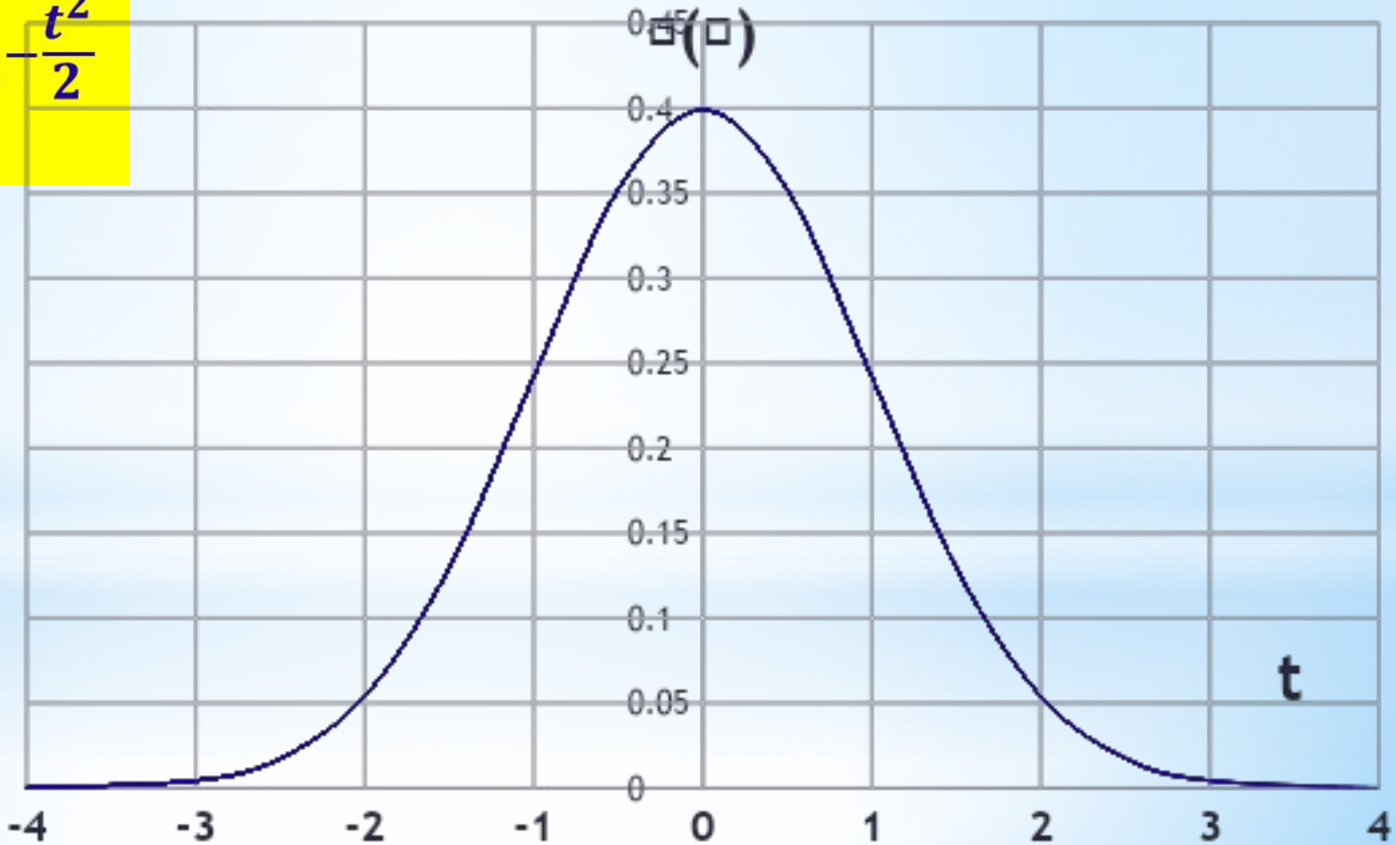
$$\Delta = t \times \mu$$

t – коэффициент доверия, зависящий от вероятности, с которой определяется предельная ошибка выборки

Независимо от характера распределения признака в генеральной совокупности при увеличении объема выборки распределение вероятностей появления $\varphi(t)$ того или иного значения выборочной средней приближается к нормальному распределению.

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

$$t = \frac{\tilde{x} - \bar{x}}{\mu}$$

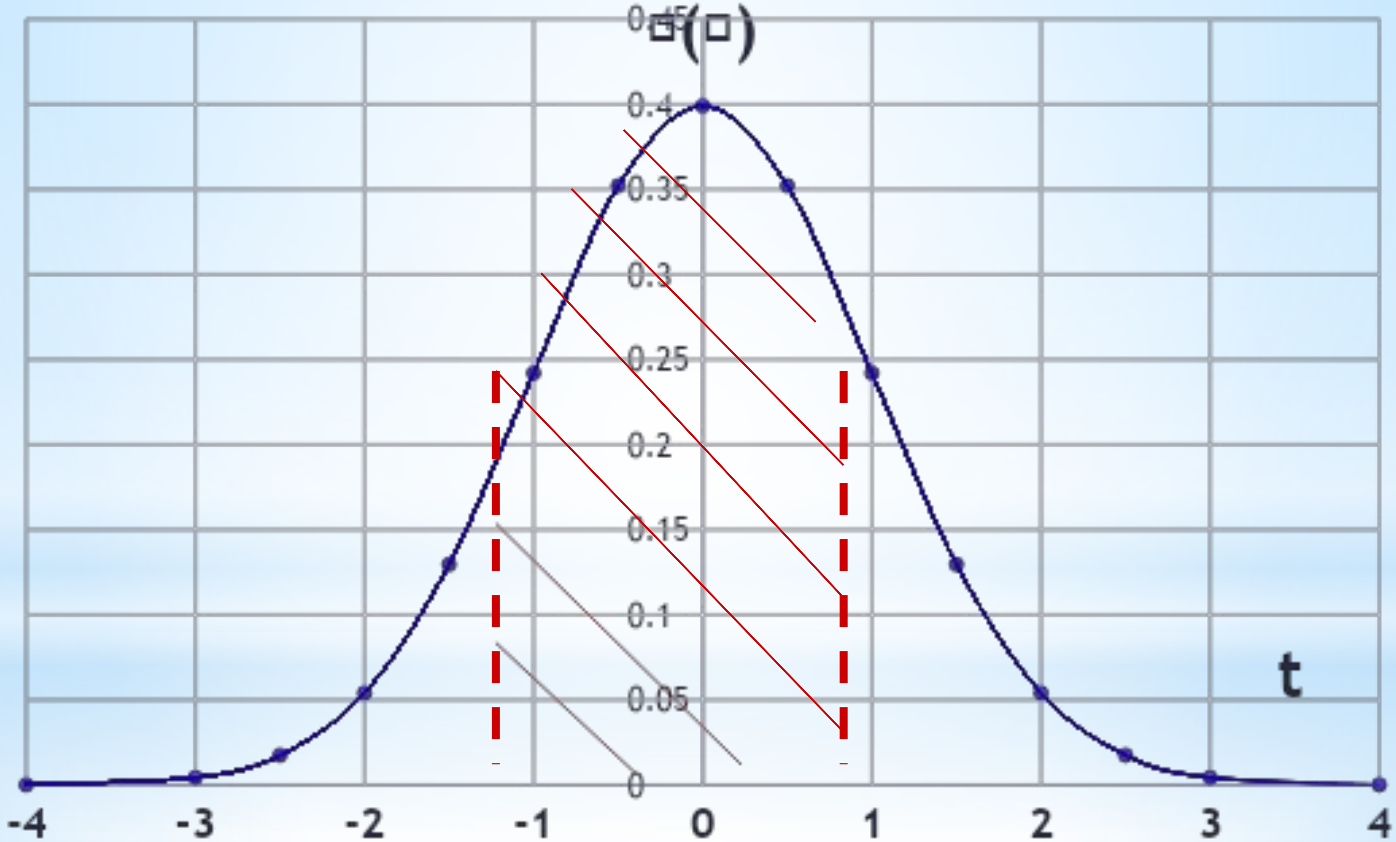


Вероятность *P(t)* появления заданной предельной ошибки $\Delta = t\mu$ также подчиняется указанному закону и может быть найдена с помощью интеграла вероятностей Лапласа:

$$P(t) = P(|\widetilde{x} - \overline{x}| \leq t\mu) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt$$

При t = 1

$$P(t) = P\left(\left|\tilde{x} - \bar{x}\right| \le \mu\right) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{+1} e^{-\frac{t^2}{2}} dt$$



Наиболее часто встречающиеся значения *t* и *P(t)*

Коэффициент доверия *t*	Доверительная вероятность *P(t)*	Интерпретация: *появление заданной предельной ошибки возможно с вероятностью…*
1	0,683	68,3%
2	0,954	95,4%
3	0,997	99,7%
1,96	0,950	95,0%
2,58	0,990	99,0%

Значения интеграла Лапласа для разных значений t рассчитаны и приводятся в специальных таблицах

t	Сотые доли t									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0,0	0,0000	0,0080	0,0160	0,0239	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,1	0,0797	0,0876	0,0955	0,1035	0,1113	0,1192	0,1271	0,1350	0,1429	0,1507
0,2	0,1585	0,1663	0,1741	0,1819	0,1897	0,1974	0,2051	0,2128	0,2205	0,2282
0,3	0,2358	0,2434	0,2510	0,2586	0,2661	0,2737	0,2812	0,2886	0,2961	0,3035
0,4	0,3108	0,3182	0,3255	0,3328	0,3401	0,3473	0,3545	0,3616	0,3688	0,3759
0,5	0,3829	0,3900	0,3969	0,4039	0,4108	0,4177	0,4245	0,4313	0,4381	0,4448
0,6	0,4515	0,4581	0,4647	0,4713	0,4778	0,4843	0,4908	0,4971	0,5035	0,5098
0,7	0,5161	0,5223	0,5285	0,5346	0,5407	0,5468	0,5528	0,5587	0,5646	0,5705
0,8	0,5763	0,5821	0,5878	0,5935	0,5991	0,6047	0,6102	0,6157	0,6211	0,6265
0,9	0,6319	0,6372	0,6424	0,6476	0,6528	0,6579	0,6629	0,6680	0,6729	0,6778
1,0	0,6827	0,6875	0,6923	0,6970	0,7017	0,7063	0,7109	0,7154	0,7199	0,7243
1,1	0,7287	0,7330	0,7373	0,7415	0,7457	0,7499	0,7540	0,7580	0,7620	0,7660
1,2	0,7699	0,7737	0,7775	0,7813	0,7850	0,7887	0,7923	0,7959	0,7995	0,8030
1,3	0,8064	0,8098	0,8132	0,8165	0,8198	0,8230	0,8262	0,8293	0,8324	0,8355
1,4	0,8385	0,8415	0,8444	0,8473	0,8501	0,8529	0,8557	0,8584	0,8611	0,8638
1,5	0,8664	0,8670	0,8715	0,8740	0,8764	0,8789	0,8812	0,8836	0,8859	0,8882
1,6	0,8904	0,8926	0,8948	0,8969	0,8990	0,9011	0,9031	0,9051	0,9070	0,9090
1,7	0,9109	0,9127	0,9146	0,9164	0,9181	0,9199	0,9216	0,9233	0,9249	0,9266
1,8	0,9281	0,9297	0,9312	0,9328	0,9342	0,9357	0,9371	0,9385	0,9399	0,9412
1,9	0,9426	0,9439	0,9451	0,9464	0,9476	0,9488	0,9500	0,9512	0,9523	0,9534
2,0	0,9545	0,9556	0,9566	0,9576	0,9587	0,9596	0,9606	0,9616	0,9625	0,9634
2,1	0,9643	0,9651	0,9660	0,9668	0,9677	0,9684	0,9692	0,9700	0,9707	0,9715
2,2	0,9722	0,9729	0,9736	0,9743	0,9749	0,9756	0,9762	0,9768	0,9774	0,9780
2,3	0,9786	0,9791	0,9797	0,9802	0,9807	0,9812	0,9817	0,9822	0,9827	0,9832
2,4	0,9836	0,9841	0,9845	0,9849	0,9853	0,9857	0,9861	0,9865	0,9869	0,9872
2,5	0,9876	0,9879	0,9883	0,9886	0,9889	0,9892	0,9895	0,9898	0,9901	0,9904

P(t) = 0,9385

⇓

t = 1,87

t = 2,44

⇓

P(t) = 0,9853

Наряду с абсолютной величиной средней и предельной ошибки выборки в статистической практике используется ***относительная ошибка выборки***, рассчитываемая как отношение предельной ошибки к исследуемому параметру:

$$\Delta_{отн} = \frac{\Delta}{\tilde{x}} \times 100\%$$

или

$$\Delta_{отн} = \frac{\Delta}{w} \times 100\%$$

Теоретически в знаменателе должно стоять значение исследуемого параметра генеральной совокупности.

Однако, учитывая, что оно неизвестно, относительная ошибка рассчитывается через соответствующий параметр выборки.

Распространение результатов выборочного наблюдения на генеральную совокупность

$$\tilde{x} - \Delta \leq \bar{x} \leq \tilde{x} + \Delta$$

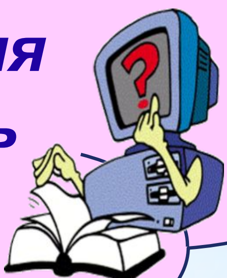
$$N(\tilde{x} - \Delta) \leq N\bar{x} \leq N(\tilde{x} + \Delta)$$

$$w - \Delta \leq p \leq w + \Delta$$

!

При распространении результатов выборочного наблюдения на генеральную совокупность ВСЕГДА необходимо указывать вероятность, с которой гарантируется этот результат

Всегда ли результаты выборочного наблюдения можно распространять на генеральную совокупность



Если результаты выборочного наблюдения можно распространять на генеральную совокупность, то такая выборка называется *репрезентативной*

Критерий репрезентативности выборочной совокупности

$$\left.\begin{array}{c}\Delta_{\text{отн}} = \dfrac{\Delta}{\tilde{x}} \times \mathbf{100\%} \\ \text{или} \\ \Delta_{\text{отн}} = \dfrac{\Delta}{w} \times \mathbf{100\%}\end{array}\right] \leq \mathbf{5\%}$$

Определение необходимой численности выборки

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

$$\mu^2 nN = \sigma^2 N - \sigma^2 n \Rightarrow \boxed{\mathbf{n} = \frac{\sigma^2 N}{\mu^2 N + \sigma^2}}$$

$$\Delta = \mu t$$

$$\Downarrow$$

$$\mu = \frac{\Delta}{t} \Rightarrow \mathbf{n} = \frac{\sigma^2 N}{\frac{\Delta^2}{t^2} N + \sigma^2} \Rightarrow \boxed{\mathbf{n} = \frac{\sigma^2 N t^2}{\Delta^2 N + \sigma^2 t^2}}$$

Оценка дисперсии изучаемого признака при определении необходимой численности выборки

Возможные подходы:

* исходя из результатов предыдущих обследований (выборочных или сплошных), либо специально организованного пробного обследования
* исходя из гипотезы о законе распределения изучаемого признака в генеральной совокупности. Если распределение близко к нормальному, то размах вариации (*R*) в 6 раз больше среднего квадратического отклонения

$$\sigma^2 = \frac{R^2}{36} = \frac{\left(x_{\max} - x_{\min}\right)^2}{36}$$

* если в результате обследования необходимо установить долю единиц, обладающих определенным значением альтернативного признака. Дисперсия альтернативного признака равна $\sigma^2 = pq$, где p – доля единиц, обладающих изучаемым признаком; q – доля единиц, не обладающих им. Максимальное значение дисперсии альтернативного признака равно 0,25 при $p = q = 0,5$

Малая выборка

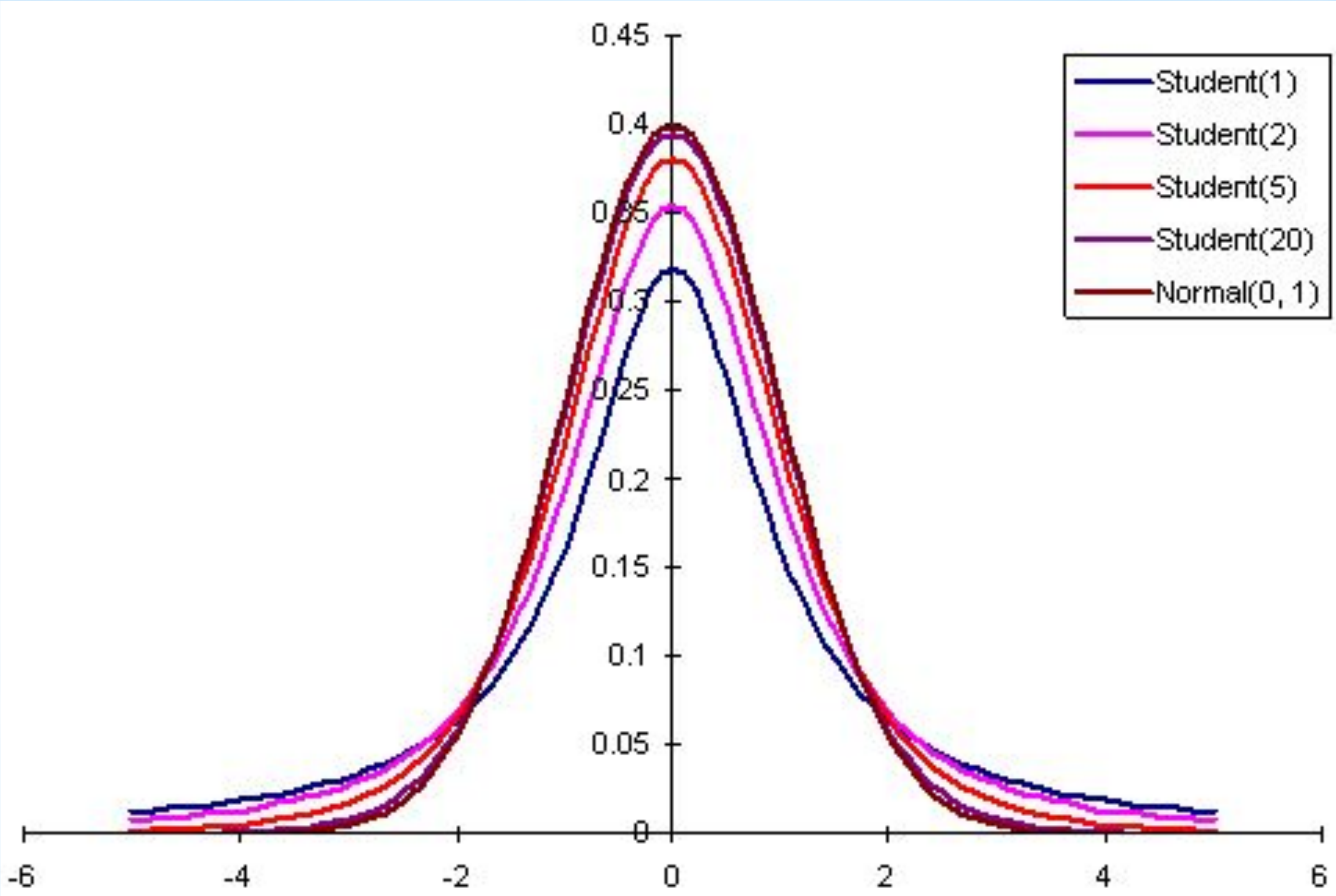
Выборка считается малой, если обследуется не более 30 единиц

$$n \leq 30$$

Средняя ошибка малой выборки при собственно случайном повторном отборе

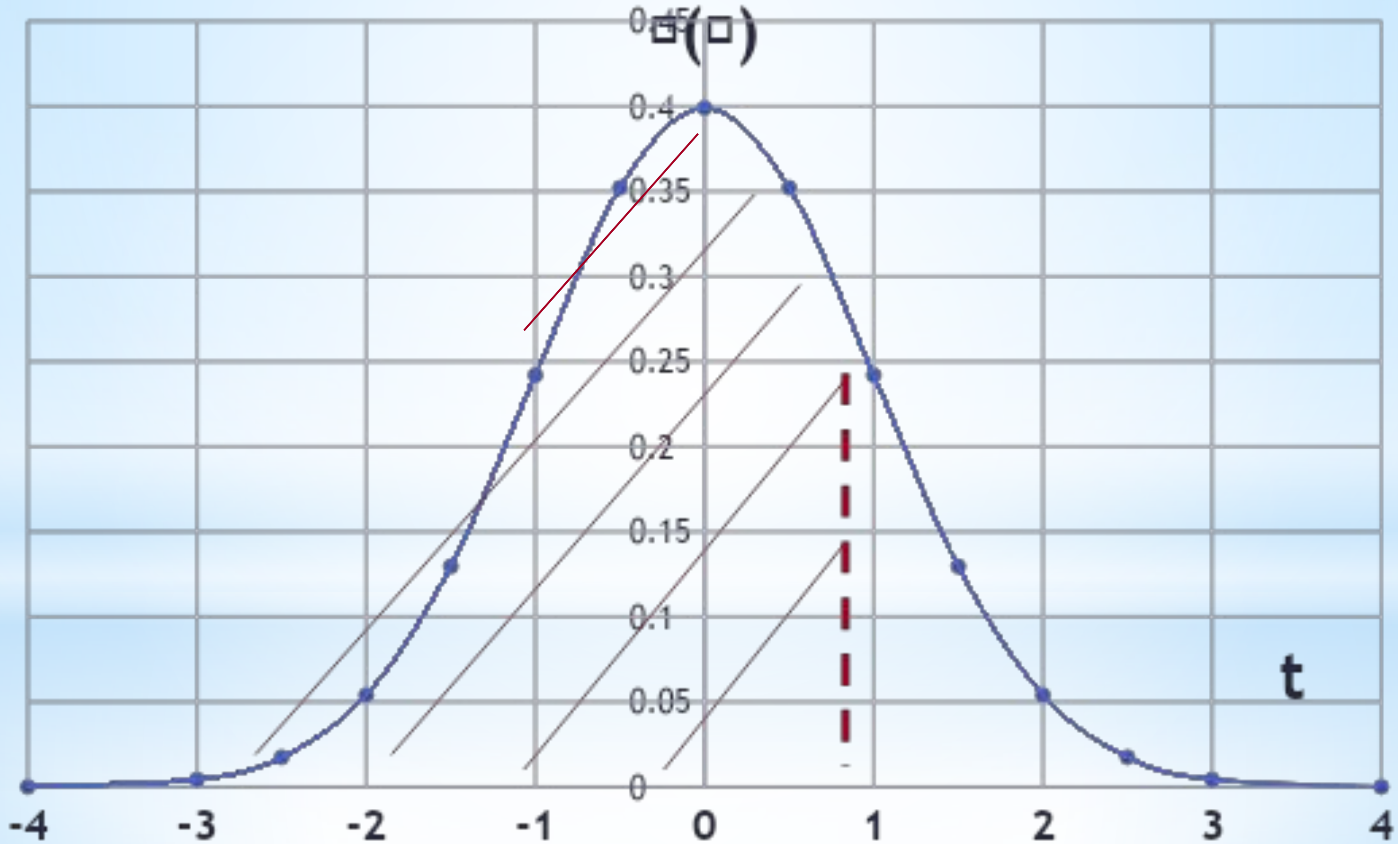
$$\mu_{\text{м.в.}} = \sqrt{\frac{\sigma^2}{n-1}}$$

В случае малой выборки в условиях нормально распределенной генеральной совокупности вероятность появления нормированного отклонения выборочной средней от генеральной подчиняется закону распределения Стьюдента


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-6
-4
-2
0
2
4
6
Student(1)
Student(2)
Student(5)
Student(20)
Normal(0, 1)

$$S(t) = S(\tilde{x} - \bar{x} \le t\mu) = \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\Gamma\left[\frac{\nu}{2}\right]\sqrt{\pi\nu}} \int_{-\infty}^{t} \left[1 + \frac{t^2}{\nu}\right]^{-\frac{\nu+1}{2}} dt$$

ν = n-1 – число степеней свободы



$S(t)$ в распределении Стьюдента

t \ ν	1	2	3	4	5	6–7	8–10	11–15	16–25	26–30	∞
0,0	500	500	500	500	500	500	500	500	500	500	500,000
0,1	532	535	537	537	538	538	539	539	539	539	539,827
0,2	563	570	573	574	576	575	578	578	578	578	579,259
0,3	593	606	608	610	612	613	615	616	616	616	617,911
0,4	621	636	642	645	647	649	651	652	653	654	655,421
0,5	648	667	674	678	681	683	685	687	689	689	691,462
0,6	672	695	705	710	713	715	718	721	722	724	725,746
0,7	694	723	733	739	742	746	749	752	754	756	758,036
0,8	715	746	759	766	770	774	778	781	783	785	788,144
0,9	733	768	783	790	795	800	804	808	811	813	815,939
1,0	750	789	804	813	818	823	828	832	835	838	841,344
1,1	765	807	824	834	839	844	850	854	858	860	864,333
1,2	779	824	842	852	858	864	870	874	878	881	884,930
1,3	791	838	858	868	875	881	887	892	896	899	903,199
1,4	803	852	872	883	890	896	902	907	912	915	919,243
1,5	813	864	885	896	903	909	916	921	925	928	933,192
1,6	822	875	896	908	915	921	928	933	937	940	945,200
1,7	831	884	906	918	925	932	938	943	948	951	955,434
1,8	839	893	915	927	934	941	947	952	956	959	964,069
1,9	846	901	923	935	942	948	955	960	964	967	971,283
2,0	852	908	930	942	949	955	962	967	970	973	977,249
2,1	858	915	937	948	955	961	967	972	976	978	982,135
2,2	864	921	942	954	960	966	972	977	980	982	986,096
2,3	870	926	948	958	965	971	977	981	984	986	989,275
2,4	874	931	952	963	969	975	980	984	987	989	991,802
2,5	879	935	956	976	973	978	983	987	989	991	993,790
2,6	883	939	960	970	976	981	986	989	991	993	995,338
2,7	887	943	963	973	979	983	988	991	993	995	996,533
2,8	891	946	966	976	981	985	990	993	995	996	997,444
2,9	894	949	969	978	983	987	991	994	996	997	998,134
3,0	898	952	971	980	985	989	993	995	997	997	998,650
3,1	901	955	973	982	987	990	994	996	997	998	999,032
3,2	904	957	975	984	988	991	995	997	998	998	999,312
3,3	906	960	977	985	989	992	995	997	998	999	999,516
3,4	909	962	979	986	990	993	996	998	998	999	999,663

$t = 1{,}8$

$n = 20$

$\nu = n-1 = 19$

$S(t) = 0{,}956$

S(t) – вероятность того, что фактическое значение нормированного отклонения выборочной средней от генеральной не превышает значения предельной ошибки

P(t) – вероятность того, что фактическое значение нормированного отклонения выборочной средней от генеральной *по абсолютной величине* не превышает значения предельной ошибки

$$\mathbf{1 = 2[1 - \mathit{S}(\mathit{t})] + P(\mathit{t})}$$

$$\mathbf{P(\mathit{t}) = 2\mathit{S}(\mathit{t}) - 1}$$

СПАСИБО ЗА ВНИМАНИЕ!