Лекция 4

ОСНОВЫ МАТЕМАТИЧЕСКОЙ ТЕОРИИ ВЫБОРОЧНОГО МЕТОДА

ТОЧЕЧНОЕ И ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

Наблюдения могут быть:

Сплошные (изучаются все объекты совокупности)

Выборочные (изучается часть объектов)

Генеральная совокупность – совокупность всех гипотетически возможных мыслимых наблюдений, которые могли бы быть произведены при данном реальном комплексе условий

Выборочная совокупность, или ***выборка*** – часть объектов, которая отобрана из генеральной совокупности для непосредственного изучения

Объем генеральной совокупности – число объектов (наблюдений) генеральной совокупности. Будем обозначать его как N

Объем выборки – число объектов (наблюдений) выборки n

Сущность выборочного метода состоит в том, чтобы по свойствам выборки можно было бы судить о свойствах генеральной совокупности. Для этого выборка должна обладать всеми свойствами генеральной совокупности, значимыми с точки зрения задач исследования. Такая выборка называется ***репрезентативной***.

Преимущества выборочного метода

- Экономия затрат ресурсов (материальных, временных, трудовых)
- Иногда единственно возможный метод (например, в случае бесконечного объема генеральной совокупности, или когда исследование связано с уничтожением наблюдаемых объектов)
- Позволяет снизить ошибки регистрации (расхождения между истинным и зарегистрированным значениями признака)

Основной недостаток – ошибки репрезентативности (ошибки исследования)

Виды выборок

- Собственно-случайная (случайный выбор элементов без предварительного разделения на группы)
- Механическая (элементы отбираются через определенный интервал)
- Типическая (выбор элементов случайным образом из типических групп)
- Серийная (отбираются целые группы, на которые делится случайно, а внутри группы все элементы подвергаются сплошному наблюдению)

Способы образования выборки – *повторный отбор* (при этом возможно повторное исследование элемента) и *бесповторный отбор* (невозможно повторное исследование элемента)

Далее будем считать, что выборка является **собственно-случайной** и (повторной либо бесповторной).

Важнейшая задача выборочного метода в математической статистике – оценка параметров генеральной совокупности по данным выборки

Оценкой параметра называют всякую функцию результатов наблюдений над случайной величиной (***статистику***), с помощью которой судят о значении этого параметра

Например, для выборки из n элементов в результате наблюдений некоторой случайной величины X получили значения $X_1, X_2,...X_n$, по которой нужно оценить определенную характеристику θ (среднее значение, дисперсию, медиану и т.д.) всей генеральной совокупности. Оценкой будет любая функция полученных значений $\tilde{\theta}_n = \tilde{\theta}_n(X_1, X_2,...X_n)$, если по ней можно судить об истинном значении θ

Для разных выборок набор значений $X_1, X_2,...X_n$ будет получаться разным. Поэтому и значение оценки параметров $\tilde{\theta}_n$ тоже будет разным – это случайная величина. Чтобы оценка была «качественной», разброс значений оценки, полученных для разных выборок, должен быть несущественным.

Оценки параметров могут быть ***точечными*** и ***интервальными***

Точечная оценка есть определенное значение оцениваемого параметра, полученное по выборке.

Интервальная оценка есть интервал, который накрывает оцениваемый параметр с заданной вероятностью.

Рассмотрим сначала точечные оценки.

Точечные оценки

Некоторые важные свойства оценок:

Оценка $\tilde{\theta}_n$ параметра θ называется ***несмещенной***, если ее математическое ожидание равно оцениваемому параметру

$$M\left(\tilde{\theta}_n\right) = \theta$$

Требование несмещенности гарантирует отсутствие систематических ошибок при оценивании

Оценка $\tilde{\theta}_n$ параметра θ называется ***состоятельной***, если она удовлетворяет закону больших чисел, т.е. сходится по вероятности к оцениваемому параметру

$$\tilde{\theta}_n \xrightarrow[n\to\infty]{P} \theta$$

Если оценка несмещенная и ее дисперсия $\to 0$ при $n \to \infty$, то оценка является и состоятельной.

Пример: Отдельное выборочное значение является несмещенной оценкой генеральной средней $\overline{x}_г = \frac{\sum_{i=1}^{N} x_i}{N}$, но не является состоятельной оценкой. Выборочное среднее $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ несмещенной и состоятельной оценкой генеральной средней

Несмещенная оценка $\tilde{\theta}_n$ параметра θ называется эффективной, если она имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра θ, вычисленных по выборкам одного и того же объема n. Это решающее свойство, определяющее качество оценки. Степень эффективности любой оценки определяется отношением дисперсий данной и эффективной оценок. На практике часто в целях упрощения расчетов используются оценки, не обладающие высокой эффективностью. Например, генеральную среднюю можно оценивать медианой (при нормальном распределении асимптотически эффективность такой оценки стремится к $2/\pi \approx 0,64$), в то время как эффективной оценкой является выборочная средняя.

Точечная оценка генеральной средней

$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ - выборочная средняя

$\overline{x}_г = \frac{\sum_{i=1}^{N} x_i}{N}$ - генеральная средняя

Теорема. Выборочная средняя $\overline{x}$ повторной выборки есть несмещенная и состоятельная оценка генеральной средней $\overline{x}_г$, причем

$$\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}$$

где σ^2 - генеральная дисперсия

Теорема. Выборочная средняя $\overline{x}$ бесповторной выборки есть несмещенная и состоятельная оценка генеральной средней $\overline{x}_г$, причем

$$\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \approx \frac{\sigma^2}{n}\left(1-\frac{n}{N}\right)$$

Если численность выборки не превышает 5% генеральной совокупности, то можно пользоваться формулой повторной выборки

Точечная оценка генеральной дисперсии

$\sum_{i=1}^{n}\frac{(x_i-\overline{x})^2}{n}$ - выборочная дисперсия

$\sum_{i=1}^{N}\frac{(x_i-\overline{x}_г)^2}{N}$ - генеральная дисперсия

Теорема. Выборочная дисперсия $\sum_{i=1}^{n}\frac{(x_i-\overline{x})^2}{n}$ повторной и бесповторной выборок есть смещенная и состоятельная оценка генеральной дисперсии σ^2. Вычисленная по этой формуле дисперсия для разных выборок в среднем занижает генеральную дисперсию. **Исправленная выборочная дисперсия** вычисляется по формуле

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

Интервальные оценки

Интервальной оценкой параметра θ называется числовой интервал, который с заданной вероятностью γ накрывает неизвестное значение параметра θ. Соответствующий интервал называется ***доверительным***, а вероятность γ - ***доверительной вероятностью***, или ***надежностью оценки***. Обычно рассматривают симметричный интервал $(\theta-\Delta,\theta+\Delta)$

Величина доверительного интервала 2Δ уменьшается с ростом объема выборки и увеличивается с ростом доверительной вероятности. Величина Δ является ошибкой репрезентативности (представительства) выборки

Теорема. Если случайная величина X имеет нормальный закон распределения с параметрами $a_г=\overline{x}_г$ и σ, т.е. $N(a_г,\sigma^2)$, то выборочная средняя $\overline{x}$ имеет нормальный закон распределения $N(a_г,\frac{\sigma^2}{n})$.

Следовательно, параметр $\frac{\overline{x}-a_г}{\left(\frac{\sigma}{\sqrt{n}}\right)}$ имеет стандартное нормальное распределение. Теорема справедлива для любого объема выборки.

Поскольку параметры генеральной дисперсии ($a_г$ и σ) на практике неизвестны, их заменяют «наилучшими» оценками по данной конкретной выборке – выборочной средней (будем ее обозначать так же a) и выборочной исправленной дисперсией $\sigma_в$ $\sigma\to\sigma_в=\frac{\sigma\sqrt{n}}{\sqrt{n-1}}$. Тогда в результате замены параметр (статистику) $t=\frac{\overline{x}-a}{\left(\frac{\sigma_в}{\sqrt{n}}\right)}$. Необходимо выяснить вид

распределения этой статистики. Умножим числитель и знаменатель t на $\frac{\sigma}{\sqrt{n}}$ и выразим исправленную выборочную дисперсию, которая считается по формуле $\sigma_в^2 = \sum_{i=1}^{n} \frac{(x_i - a)^2}{n-1}$ через оценку «неисправленной» дисперсии $s^2 = \sum_{i=1}^{n} \frac{(x_i - a)^2}{n}$

$$t = \frac{\bar{x} - a}{\left(\frac{\sigma_в}{\sqrt{n}}\right)} = \frac{(\bar{x} - a) \Big/ \frac{\sigma}{\sqrt{n}}}{\sqrt{\frac{ns^2}{(n-1)\sigma^2}}}$$

Случайная величина в числителе имеет стандартное нормальное распределение, как было отмечено выше. А случайная величина $\frac{ns^2}{(n-1)\sigma^2}$ имеет хи-квадрат распределение с $k = n-1$ степенями свободы (так как это сумма квадратов нормально распределенных величин). Следовательно, статистика t имеет распределение Стьюдента.

Таким образом, **статистика** $t = \frac{\bar{x} - a}{\sigma_{\bar{x}}}$**,** где a - среднее (точечная оценка), $\sigma_a = \frac{\sigma_в}{\sqrt{n}}$ - среднеквадратичная ошибка среднего (точечная оценка), $\sigma_в$ - стандартное отклонение (точечная оценка) **имеет распределение Стьюдента с** $k = n-1$ **степенями свободы**, где n - объем выборки.

Значение t^* определяют границы интервала $\Delta = t^* \sigma_a$, который с заданной доверительной вероятностью γ накрывает истинное среднее значение. Величина t^* зависит от доверительной вероятности и объема выборки $t^* = t_{\gamma, n-1}$. Будем называть ее коэффициентом Стьюдента. Можно найти по таблицам, либо с помощью встроенных функций SPSS

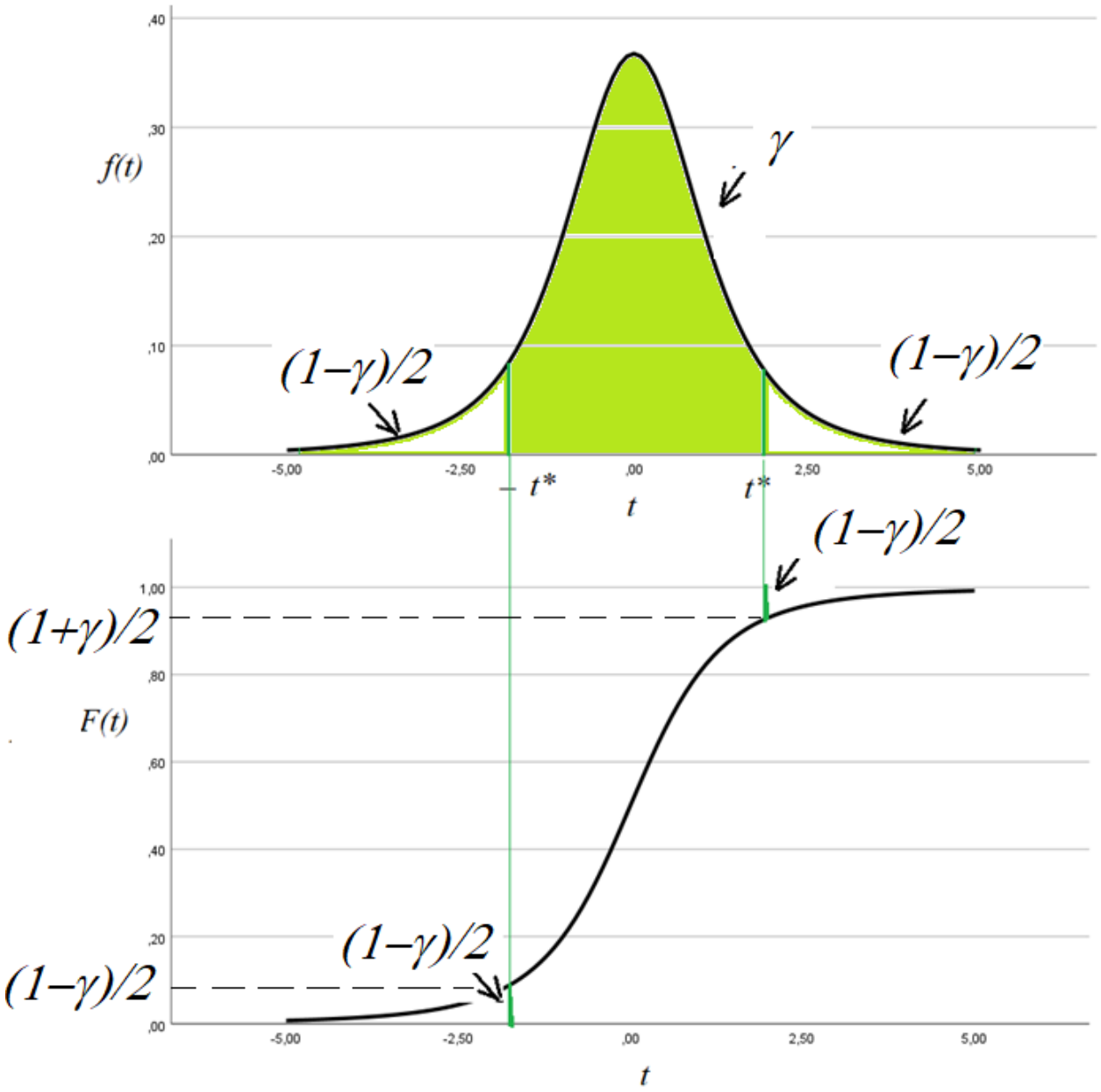
$$t^* = t_{\gamma, n-1} = \text{IDF.T}\left(\frac{1+\gamma}{2}, \text{n}-1\right)$$

Либо

$$t^* = t_{\gamma, n-1} = -\text{IDF.T}\left(\frac{1-\gamma}{2}, \text{n}-1\right)$$

При больших объемах выборки n распределение Стьюдента приближается к стандартному нормальному. На практике при $n > 30$ можно приближенно считать

$$t^* \approx -\text{IDF.Normal}\left(\frac{1-\gamma}{2}, 0, 1\right) = \text{IDF.Normal}\left(\frac{1+\gamma}{2}, 0, 1\right)$$



Интервальной оценкой генеральной средней будет интервал

$$a - \sigma_{\bar{x}} t_{\gamma,n-1} \le a_{\varepsilon} \le a + \sigma_{\bar{x}} t_{\gamma,n-1}$$

В SPSS для вычисления статистик выборки можно использовать опции

Анализ → *Описательные статистики* → *Частоты*

Анализ → *Описательные статистики* → *Описательные статистики*

Анализ → *Описательные статистики* → *Разведочный анализ*

Первый вариант (→ *Частоты*) рассматривали ранее. Познакомимся с возможностями опций *Анализ* → *Описательные статистики* → *Описательные статистики*

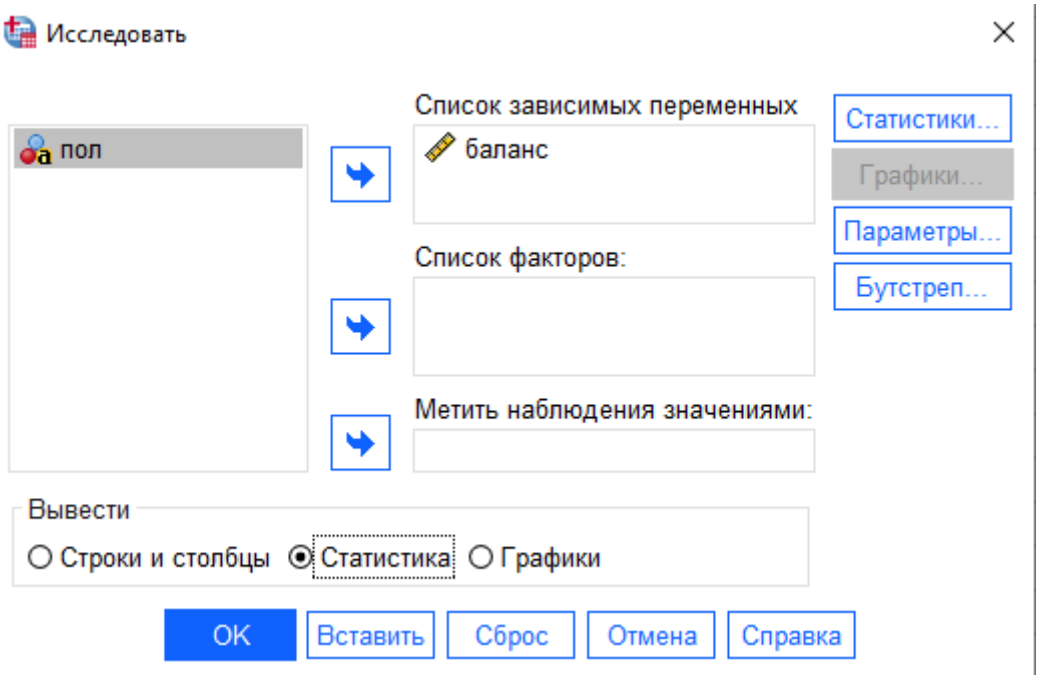
Пример. Файл Баланс.sav

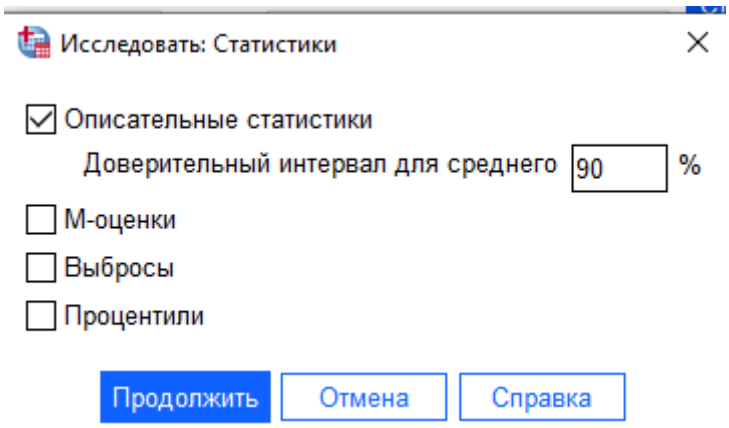
Описательные статистики

	N	Среднее		Стандартная отклонения	Дисперсия
	Статистика	Статистика	Стандартная ошибка	Статистика	Статистика
баланс	25	189,64	38,464	192,321	36987,240
N валидных (по списку)	25				

Для более подробного описания используются опции

Анализ → *Описательные статистики* → *Разведочный анализ*





Сводный отчет по наблюдениям

	Наблюдения					
	Допустимо		Пропущенные		Всего	
	N	Проценты	N	Проценты	N	Проценты
баланс	25	100,0%	0	0,0%	25	100,0%

Описательные статистики

			Статистика	Стандартная ошибка
баланс	Среднее		189,64	38,464
	90% Доверительный интервал для среднего	Нижняя граница	123,83	
		Верхняя граница	255,45	
	Среднее по выборке, усеченной на 5%		169,04	
	Медиана		104,00	
	Дисперсия		36987,240	
	Стандартная отклонения		192,321	
	Минимум		0	
	Максимум		805	
	Диапазон		805	
	Межквартильный диапазон		269	
	Асимметрия		1,638	,464
	Эксцесс		3,079	,902

$\Delta = 65,81$

$t_{\gamma,n-1} = 1,71 \left(= 65,81 / 38,464\right)$

$\gamma = 0,9$

$n = 25$

Проверим:

$t_{\gamma,n-1} = -\text{IDF.T}\left(0.05, 24\right) = 1,71$