

Лекция 8

Выборочный метод математической статистики. Проверка статистических гипотез

Выборочный метод

- В реальных условиях обычно бывает трудно или экономически нецелесообразно, а иногда и невозможно исследовать всю совокупность, характеризующую изучаемый признак (*генеральную совокупность*).
- На практике широко применяется выборочное наблюдение, когда обрабатывается часть генеральной совокупности (*выборочная совокупность*).

Выборочный метод

- Свойства (закон распределения и его параметры) генеральной совокупности неизвестны, поэтому возникает задача их оценки по выборке.
- Для получения хороших оценок характеристик генеральной совокупности необходимо, чтобы выборка была *репрезентативной* (представительной).
- Репрезентативность, в силу закона больших чисел, достигается случайностью отбора.

Выборочный метод

Различают 5 основных типов выборок.

1) *Собственно случайная*:

- *повторная* (элементы после выбора возвращаются обратно);
- *бесповторная* (выбранные элементы не возвращаются).

2) *Типическая* - генеральная совокупность предварительно разбивается на группы типических элементов, и выборка осуществляется из каждой.

Следует различать:

а) равномерные выборки (при равенстве объемов исходных групп в генеральной совокупности выбирается одинаковое количество элементов из каждой);

б) пропорциональные (численность выборок формируют пропорционально численностям или средним квадратическим отклонениям групп генеральной совокупности);

в) комбинированные (численность выборок пропорциональна и средним квадратическим отклонениям, и численностям групп генеральной совокупности).

Выборочный метод

Различают 5 основных типов выборок.

3) *Механическая* - отбор элементов проводится через определенный интервал.

4) *Серийная* - отбор проводится не по одному элементу, а сериями для проведения сплошного обследования.

5) *Комбинированная* - используются различные комбинации вышеуказанных методов, например, типическая выборка сочетается с механической и собственно случайной.

Выборочный метод

- После осуществления выборки возникает задача оценки числовых характеристик генеральной совокупности по элементам выборочной совокупности.
- Различают *точечные* и *интервальные оценки*.
- ***Точечная оценка*** характеристики генеральной совокупности - это число, определяемое по выборке.
- Пусть $\hat{\theta} = \widehat{\theta}_n$ - выборочная характеристика, вычисленная по результатам n наблюдений величины X , используемая в качестве оценки θ - характеристики генеральной совокупности (в качестве θ может быть $M(X)$, $D(X)$ и т. д.).

Выборочный метод

- Качество оценки $\hat{\theta}$ устанавливается по трем свойствам: **состоятельность, несмещенность, эффективность.**

1) Состоятельность. Оценка $\hat{\theta}_n$ является состоятельной оценкой генеральной характеристики θ , если для любого $\varepsilon > 0$ выполняется следующее равенство

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$$

Это означает, что при увеличении объема выборки n выборочная характеристика $\hat{\theta}_n \rightarrow \theta$.

2) Несмещенность. Оценка $\hat{\theta}$ генеральной характеристики θ называется несмещенной, если для любого фиксированного числа наблюдений и выполняется равенство $M(\hat{\theta}_n) = \theta$.

3) Эффективность. Несмещенная оценка $\hat{\theta} = \hat{\theta}_n$ генеральной характеристики θ называется несмещенной эффективной, если среди всех подобных оценок той же характеристики она имеет наименьшую дисперсию: $D(\hat{\theta}_n) \rightarrow \min$.

Выборочный метод

- Можно показать, что статистики \bar{X} , \hat{p} являются состоятельными, несмещенными и эффективными характеристиками математического ожидания $M(X)$ и вероятности p соответственно.
- Выборочная дисперсия \hat{D} (далее $\hat{D} = \sigma^2$) не обладает свойством *несмещенности*. На практике используют исправленную выборочную дисперсию S^2 , которая является несмещенной оценкой дисперсии генеральной совокупности:

$$S^2 = \frac{n}{n-1} \sigma^2(x) = \frac{n}{n-1} \cdot \frac{\sum (x_i - \bar{X})^2 \cdot n_i}{n} = \frac{\sum (x_i - \bar{X})^2 \cdot n_i}{n-1}, \quad (15)$$

где S - стандартное отклонение.

Выборочный метод

- Кроме того, в расчетах используют стандартную ошибку выборки:

$$S_x = \frac{S}{\sqrt{n}}. \quad (16)$$

- Точечные оценки получают обычно с помощью *метода моментов* и *метода максимального правдоподобия*.
- *Интервальной* называют оценку, которая определяется двумя числами - границами интервала. Она позволяет ответить на вопрос: внутри какого интервала и с какой вероятностью находится неизвестное значение оцениваемого параметра θ генеральной совокупности?

Выборочный метод

- Пусть $\hat{\theta}$ - точечная оценка параметра θ . Чем меньше разность $\hat{\theta}$ и θ , тем точнее и лучше оценка.
- Обычно говорят о *доверительной вероятности* (надежности оценки) $p = 1 - \alpha$, с которой θ будет находиться в интервале

$$\hat{\theta} - \Delta < \theta < \hat{\theta} + \Delta,$$

где: $\Delta (\Delta > 0)$ - предельная ошибка выборки, которая может быть либо задана наперед, либо вычислена; α - риск или уровень значимости (вероятность того, что неравенство будет неверным).

Выборочный метод

- Оценка указанного доверительного интервала может быть получена (с наименьшей вероятностью) с помощью неравенства Чебышева (при $\varepsilon = \Delta$).
- В качестве $1-\alpha$ принимают значения 0,90; 0,95; 0,99; 0,999.
- Доверительная вероятность показывает, что в $(1-\alpha)100\%$ случаев оценка θ будет покрываться указанным интервалом.

Выборочный метод

- Точечная оценка математического ожидания $M(X)=a$ определяется как средняя арифметическая:

$$\bar{X} = \frac{1}{n} \sum x_i n_i \quad (17)$$

- Точечная оценка вероятности p_i определяется как относительная частота:

$$\hat{p}_i = \frac{n_i}{n}. \quad (18)$$

Выборочный метод

- Для построения доверительного интервала параметра a - математического ожидания нормального распределения составляют выборочную характеристику (*статистику*), функционально зависимую от наблюдений и связанную с a , например, для повторного отбора:

$$u = \frac{\bar{X} - a}{\sigma(\bar{X})} = \frac{\bar{X} - a}{\frac{\sigma}{\sqrt{n}}} \quad (19)$$

Выборочный метод

- Статистика u распределена по нормальному закону распределения с математическим ожиданием $a = 0$ и средним квадратическим отклонением $\sigma=1$.

Отсюда

$$P(|u| < u_{\alpha/2}) = 1 - \alpha,$$

$$\text{или } 2\Phi(u_{\alpha/2}) = 1 - \alpha,$$

где Φ - функция Лапласа, $u_{\alpha/2}$ - *квантиль* нормального закона распределения, соответствующая уровню значимости α .

Выборочный метод

- *Доверительный интервал* для параметра a :

$$\bar{X} - u_{\alpha/2} \frac{\sigma}{\sqrt{n}} < a < \bar{X} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad (20)$$

где $\Delta_{\bar{X}} = u_{\alpha/2} \sigma(\bar{X})$ - предельная ошибка выборочной средней.

Выборочный метод

Формулы предельной ошибки и необходимого объема выборки для различных случаев отбора

Выборка		Собственно-случайная		Типическая		Серийная	
		повторная	бесповторная	повторная	бесповторная	повторная	бесповторная
Предельная ошибка, Δ	Средней, \bar{X}	$t \sqrt{\frac{\sigma^2}{n}}$	$t \sqrt{\frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right)}$	$t \sqrt{\frac{\sigma^2}{n}}$	$t \sqrt{\frac{\sigma^2}{n} \left(1 - \frac{n}{N}\right)}$	$t \sqrt{\frac{\delta_{M.C}^2}{n_c}}$	$t \sqrt{\frac{\delta_{M.C}^2}{n_c} \left(1 - \frac{n_c}{N_c}\right)}$
	Доли, \hat{P}	$t \sqrt{\frac{pq}{n}}$	$t \sqrt{\frac{pq}{n} \left(1 - \frac{n}{N}\right)}$	$t \sqrt{\frac{pq}{n}}$	$t \sqrt{\frac{pq}{n} \left(1 - \frac{n}{N}\right)}$	$t \sqrt{\frac{pq_{M.C}}{n_c}}$	$t \sqrt{\frac{pq_{M.C}}{n_c} \left(1 - \frac{n_c}{N_c}\right)}$
Необходимая численность, n	Средней, \bar{X}	$\frac{t^2 \sigma^2}{\Delta^2}$	$\frac{t^2 \sigma^2 N}{t^2 \sigma^2 + \Delta^2 N}$	$\frac{t^2 \sigma^2}{\Delta^2}$	$\frac{t^2 \sigma^2 N}{t^2 \sigma^2 + \Delta^2 N}$	$\frac{t^2 \delta_{M.C}^2}{\Delta^2}$	$\frac{t^2 \delta_{M.C}^2 N_c}{t^2 \delta_{M.C}^2 + \Delta^2 N_c}$
	Доли, \hat{P}	$\frac{t^2 pq}{\Delta^2}$	$\frac{t^2 Npq}{t^2 pq + \Delta^2 N}$	$\frac{t^2 pq}{\Delta^2}$	$\frac{t^2 Npq}{t^2 pq + \Delta^2 N}$	$\frac{t^2 pq_{M.C}}{\Delta^2}$	$\frac{t^2 N_c pq_{M.C}}{t^2 pq_{M.C} + \Delta^2 N_c}$

Выборочный метод

Формулы предельной ошибки и необходимого объема выборки для различных случаев отбора (пояснения к таблице)

- t - квантиль распределения, соответствующая уровню значимости α ,
 - а) при $n \geq 30$ $t = u_{\alpha/2}$ - квантиль нормального закона распределения (прил.1),
 - б) при $n < 30$ t - квантиль распределения Стьюдента с $\nu = n - 1$ степенями свободы для двусторонней области;
- σ^2 - выборочная дисперсия,
 - а) при $n \geq 30$ $\sigma^2 = \frac{\sum (x_i - \bar{x})^2 n_i}{n}$,
 - б) при $n < 30$ вместо σ^2 берут $S^2 = \frac{\sum (x_i - \bar{x})^2 n_i}{n - 1}$;
- pq - дисперсия относительной частоты в схеме повторных независимых испытаний;

Выборочный метод

Формулы предельной ошибки и необходимого объема выборки для различных случаев отбора (пояснения к таблице)

- N - объем генеральной совокупности;
- n - объем выборки;
- $\overline{\sigma^2}$ - средняя арифметическая групповых дисперсий (внутригрупповая дисперсия);
- \overline{pq} - средняя арифметическая дисперсий групповых долей;
- $\delta_{м.с}^2$ - межсерийная дисперсия;
- $R_{м.с}$ - межсерийная дисперсия доли;
- N_c - число серий в генеральной совокупности;
- n_c - число отобранных серий (объем выборки);

Проверка статистических гипотез. Основные определения

- *Статистической гипотезой* называется всякое высказывание о генеральной совокупности, проверяемое по выборке.
- Статистические гипотезы делятся на:
 - *параметрические* - это гипотезы, сформулированные относительно параметров (среднего значения, дисперсии и т. д.) распределения известного вида;
 - *непараметрические* - это гипотезы, сформулированные относительно вида распределения (например, определение по выборке степени нормальности генеральной совокупности).

Проверка статистических гипотез. Основные определения

- Процесс использования выборки для проверки гипотезы называется *статистическим доказательством*.
 - Основную выдвигаемую гипотезу называют *нулевой H_0* .
 - Наряду с нулевой гипотезой рассматривают *альтернативную* ей H_1 .
-

Например: $H_0: M(X)=1$, математическое ожидание генеральной совокупности равно 1;

$H_1: M(X)>1$, или $M(X)<1$, или $M(X) \neq 1$
(математическое ожидание больше 1, или меньше 1, или не равно 1).

Проверка статистических гипотез. Ошибки первого и второго родов

- Выбор между гипотезами H_0 и H_1 может сопровождаться ошибками двух родов.
- Ошибка *первого рода* α означает вероятность принятия H_1 , если верна гипотеза H_0 : $\alpha = p(H_1/H_0)$.
- Ошибка *второго рода* β означает вероятность принятия H_0 , если верна гипотеза H_1 : $\beta = p(H_0/H_1)$.
- Существует правильное решение двух видов:

$$p(H_0/H_0) = 1 - \alpha \quad \text{и} \quad p(H_1/H_1) = 1 - \beta$$

Принятая гипотеза	H_0	H_1
H_0 - верна	$P(H_0/H_0) = 1 - \alpha$	$P(H_1/H_0) = \alpha$
H_0 - неверна	$P(H_0/H_1) = \beta$	$P(H_1/H_1) = 1 - \beta$

Проверка статистических гипотез.

Статистический критерий

- Правило, по которому принимается решение о том, верна или не верна гипотеза H_0 , называется *критерием*, где:

$\alpha = P(H_1/H_0)$ - *уровень значимости критерия;*

$M = 1 - \beta = P(H_1/H_1)$ - *мощность критерия.*

- ***Статистическим критерием K*** называют случайную величину, с помощью которой принимают решение о принятии или отклонении H_0 .

Проверка статистических гипотез.

Статистический критерий

- Для проверки *параметрических* гипотез используют *критерии значимости*, основанные на статистиках: u , χ^2 , t , F .
- *Непараметрические* гипотезы проверяют с помощью *критериев согласия*, использующих статистики распределений: χ^2 , Колмогорова-Смирнова и т. д.

Проверка статистических гипотез. Статистический критерий

- Например, $H_0: M(X)=10$. В зависимости от альтернативной гипотезы рассматривают три случая.

1) Если $H_1: M(X) \neq 10$.

В этом случае рассматривают двустороннюю критическую область и используют дифференциальную функцию $f(K/H_0)$, для определения соответствующих квантилей (границ области принятия гипотезы - левой ($K_{1-\alpha/2}$) и правой ($K_{\alpha/2}$)).
и осью абсцисс, равна $(1 - \alpha)$



Проверка статистических гипотез. Статистический критерий

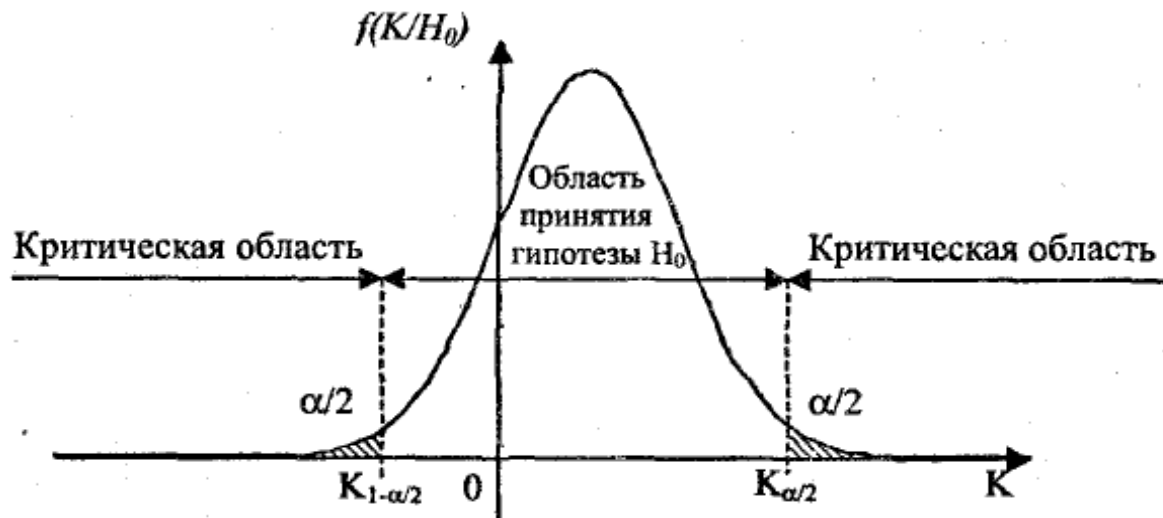
- Например, $H_0: M(X)=10$. В зависимости от альтернативной гипотезы рассматривают три случая.

1) Если $H_1: M(X) \neq 10$.

Площадь под криволинейной трапецией дифференциальной функции слева от $K_{1-\alpha/2}$ и справа от $K_{\alpha/2}$ равна $\alpha/2$.

Общая площадь ограниченная криволинейной трапецией дифференциальной функции, квантилями и осью абсцисс, равна $(1 - \alpha)$

$$P(K_{1-\alpha/2} < K < K_{\alpha/2}) = 1 - \alpha. \quad (1)$$

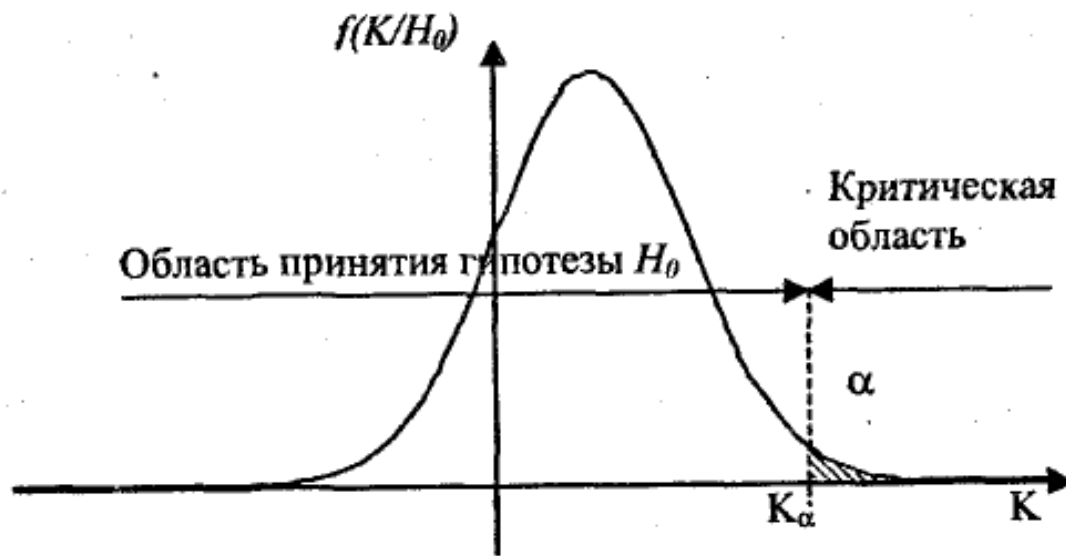


Двусторонняя критическая область

Проверка статистических гипотез. Статистический критерий

2) Если $H_1: M(X) > 10$, то рассматривается правосторонняя критическая область (площадь под криволинейной трапецией справа от K_α равна α):

$$P(K > K_\alpha) = \int_{K_\alpha}^{+\infty} f\left(\frac{K}{H_0}\right) dK = \alpha. \quad (2)$$



Правосторонняя критическая область

Проверка статистических гипотез. Статистический критерий

3) Если $H_0: M(X) < 10$, то рассматривается левосторонняя критическая область (площадь под криволинейной трапецией слева от $K_{1-\alpha}$ равна α):

$$P(K < K_{1-\alpha}) = \int_{-\infty}^{K_{1-\alpha}} f(K/H_0) dK = \alpha. \quad (3)$$



Левосторонняя критическая область

Проверка статистических гипотез. Алгоритм

1. Располагая выборочными данными (x_1, x_2, \dots, x_n) , формируют нулевую гипотезу H_0 и конкурирующую гипотезу H_1 .
2. Задают уровень значимости α (обычно принимают $\alpha = 0,1; 0,01; 0,05; 0,001$).
3. Рассматривается выборочная статистика наблюдений (критерий) K , обычно одна из перечисленных ниже:
 - u - нормальное распределение;
 - χ^2 - распределение Пирсона (хи - квадрат);
 - t - распределение Стьюдента;
 - F - распределение Фишера - Снедекора.

Проверка статистических гипотез. Алгоритм

4. На основании выборки x_1, x_2, \dots, x_n - определяют значение критерия (статистики) K (приложения 2-4). В зависимости от вида альтернативной гипотезы выбирают по соответствующей таблице квантили критерия для двусторонней ($K_{1-\frac{\alpha}{2}}$ и $K_{\frac{\alpha}{2}}$) или односторонней области ($K_{1-\alpha}$ или K_{α}).

Если значения критерия попадают в критическую область, то H_0 отвергается; в противном случае принимается гипотеза H_0 и считается, что H_0 не противоречит выборочным данным (при этом существует возможность ошибки с вероятностью, равной α).

Проверка статистических гипотез. Алгоритм

- **Замечание.** Следует отметить, что возможность принятия гипотезы происходит из *принципа невозможности наступления маловероятных событий*. Те же события, вероятность которых близка к 1, принимаются за *достоверные*. Возникает *проблема выбора уровня риска* (уровня значимости α).
- В одних случаях возможно пренебрегать событиями $p < 0,05$, в других нельзя пренебрегать событиями, которые могут появиться с $p = 0,001$ (разрушение сооружений, транспортных средств и т. д.).

Проверка статистических гипотез. Алгоритм

- **Пример.** Два сорта озимой пшеницы испытывались на одинаковом количестве участков на протяжении семи лет (табл.) При уровне значимости $\alpha = 0,05$ проверить нулевую гипотезу о существенности различий в урожайности двух сортов озимой пшеницы.

Годы	Урожайность, ц/га	
	x_{2i}	x_{1i}
1995	47	53
1996	48	43
1997	46	45
1998	51	56
1999	52	58
2000	48	55
2001	52	59

Решение. Так как имеются две зависимости выборки, т.е. существует определенная корреляция между урожайностью сортов по годам, то необходимо оценить значимость не разности двух выборочных средних, а средней разности.

Выдвигаем нулевую гипотезу: средняя величина различий в урожайности пшеницы равна нулю, $H_0: \overline{X_1 - X_2} = 0$ при $H_1: \overline{X_1 - X_2} \neq 0$.

Проверка статистических гипотез. Алгоритм

- По данным таблицы найдем среднюю разность \bar{d} и ошибку средней разности $S_{\bar{d}}$:

$$\bar{d} = \frac{\sum d_i}{n} = \frac{35}{7} = 5; \quad S_{\bar{d}} = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n(n-1)}} = \sqrt{\frac{46}{7(7-1)}} = 1,047;$$

где $d_i = x_{1i} - x_{2i}$, n - число пар наблюдений. Для двусторонней критической области при $\alpha = 0,05$; $k = n - 1 = 7 - 1 = 6$, $t_{кр.} = 2,45$.

- Сопоставив расчётное значение t с критическим, можно сделать вывод, что два сорта существенно различаются по уровню урожайности

Вспомогательная таблица для расчета ошибки средней разности

Годы	Урожайность, ц/га		Разность $d_i = x_{1i} - x_{2i}$	$(d_i - \bar{d})$	$(d_i - \bar{d})^2$
	x_{2i}	x_{1i}			
1995	47	53	6	1	1
1996	48	43	5	0	0
1997	46	45	-1	-6	36
1998	51	56	5	0	0
1999	52	58	6	1	1
2000	48	55	7	2	4
2001	52	59	7	2	4
Сумма			35	0	46