Кафедра Автоматизации управления медицинской службой (с военно-медицинской статистикой)

Основы выборочного метода статистического исследования

Лектор – кандидат мед. наук доцент
Кобзев Александр Сергеевич

ПЛАН ЛЕКЦИИ

1. Основные понятия выборочного метода статистического исследования. Генеральная совокупность и выборка.

2. Ошибки репрезентативности, оценка точности и надежности выборочных числовых характеристик.

3. Определение требуемого числа наблюдений в выборке.

ЛИТЕРАТУРА

- 1. Военно-медицинская статистика: учебник; под ред. В.И. Кувакина, В.В. Иванова. - СПб. : ВМедА, 2005. *Стр. 70-83 и 102-120*.
- 2. Математико-статистические методы в клинической практике. Учебное пособие. - С.-Петербург, 1993.
- Лядов В.Р. Основы теории вероятностей и математической статистики. Для студентов медицинских ВУЗов // Информационно-аналитическая библиотека. - Вып.2.-СПБ., 1998.-108 с.

Введение

- ***Военно-медицинская статистика***, изучает количественную сторону массовых явлений и процессов в различных областях военной медицины.

- Высокий ***научный уровень*** получения, обобщения, обработки и анализа количественной информации достигается путем проведения специально организованных статистических исследований.

- К ***статистическим исследованиям*** относятся такие исследования, которые предполагают сбор, накопление, обработку и анализ преимущественно количественных данных с использованием особых методов, разработанных в специальных разделах прикладной математики - теории вероятностей и математической статистике.

- Это "***математико-статистические методы исследования***".

1. Основные понятия выборочного метода статистического исследования.

Генеральная совокупность и выборка

- При проведении статистического исследования осуществляется ***статистическое наблюдение*** *(обследование).*
- Статистическое наблюдение может охватывать либо все без исключения единицы, из которых состоит изучаемое явление (***сплошное*** наблюдение), либо - только их часть (***несплошное*** наблюдение).
- Статистическая совокупность, включающая все единицы изучаемого явления или процесса, носит название ***генеральной совокупности***.
- Статистическая совокупность, включающая в себя лишь часть генеральной совокупности, носит название ***выборочной совокупности***, или ***выборки***.

Основные характеристики сплошного и выборочного исследования

ХАРАКТЕРИСТИКА	*СПЛОШНОЕ ИССЛЕДОВАНИЕ*	*ВЫБОРОЧНОЕ ИССЛЕДОВАНИЕ*
Изучаемая совокупность	Генеральная – включает все единицы исследования (численность – **N**)	Выборочная – включает часть единиц генеральной совокупности (численность – **n**)
Обобщающие числовые характеристики (статистические показатели)	Истинные: - **генеральная частота** (вероятность **P**); - **генеральная средняя** (математическое ожидание M_x);	Выборочные: - выборочная частота (частость $\bar{p}$) - выборочная средняя ($\bar{x}$)
Метод отбора	Сплошной	Случайный
Ошибки репрезентативности	Отсутствуют	Имеются: Для $\bar{p}$ $m_{\bar{p}} = \sqrt{\frac{\bar{p}(100 - \bar{p})}{n}}$ Для $\bar{x}$ $m_{\bar{x}} = \frac{S_x}{\sqrt{n}}$
Оценки	Точные	Вероятностные: α | 0,95 | 0,99 | 0,999 P_0 | 0,05 | 0,01 | 0,001 $P_0 = 1 - \alpha$
Оценка точности и надежности числовых характеристик	Не проводится	Проводится с помощью доверительного интервала : $I_P = \bar{p} \pm t_\alpha \cdot m_{\bar{p}}$ $I_{M_x} = \bar{x} \pm t_\alpha \cdot m_{\bar{x}}$ $n' = n - 1$ (3m < значения показателя)

ХАРАКТЕРИСТИКА	*СПЛОШНОЕ ИССЛЕДОВАНИЕ*	*ВЫБОРОЧНОЕ ИССЛЕДОВАНИЕ*
Изучаемая совокупность	**Генеральная** – включает все единицы исследования (численность – **N**)	Выборочная – включает часть единиц генеральной совокупности (численность – **n**)
Числовые характеристики	Истинные: - **генеральная частота** (вероятность **P**); - **генеральная средняя** (математическое ожидание $\mathbf{M_x}$);	Выборочные: - выборочная частота (частость $\overline{p}$) - выборочная средняя ($\overline{x}$)
Метод отбора	Сплошной	Случайный

Выборочный метод наблюдения

- Обобщающие Числовые характеристики (средние величины или относительные величины), получаемые на основе несплошного наблюдения, как правило, отличаются от аналогичных показателей, вычисленных на основе данных сплошного наблюдения, и не могут быть безоговорочно использованы для объективной количественной оценки изучаемого явления в целом.

- Наименьшее расхождение между обобщающими показателями сплошного и несплошного наблюдений может быть получено при использовании ***выборочного метода наблюдения***.

- Наиболее важным условием правильного применения выборочного метода является соблюдение ***принципа случайного отбора*** при формировании выборки, то есть такого отбора, когда каждая единица генеральной совокупности имеет одинаковые со всеми другими возможности попасть в выборочную совокупность.

Закон больших чисел

$$M_x = \lim_{n \to \infty} \bar{x}$$

$$P_x = \lim_{n \to \infty} \bar{p}$$

2. Ошибки репрезентативности, оценка точности и надежности выборочных числовых характеристик

- *Точность* и *надежность* числовых характеристик определяется с учетом ошибок репрезентативности выборочных характеристик и доверительных интервалов.
- *Ошибки репрезентативности* - это ошибки представительности, возникающие потому, что при выборочном исследовании изучается только часть генеральной совокупности, которая недостаточно точно воспроизводит, то есть представляет, генеральную совокупность.

ХАРАКТЕРИСТИКА	*СПЛОШНОЕ ИССЛЕДОВАНИЕ*	*ВЫБОРОЧНОЕ ИССЛЕДОВАНИЕ*
Изучаемая совокупность	Генеральная – включает все единицы исследования (численность – **N**)	Выборочная – включает часть единиц генеральной совокупности (численность – **n**)
Числовые характеристики	Истинные: - **генеральная частота** (вероятность **P**); - **генеральная средняя** (математическое ожидание $\mathbf{M_x}$);	Выборочные: - выборочная частота (частость $\overline{p}$) - выборочная средняя ($\overline{x}$)
Метод отбора	Сплошной	Случайный
Ошибки репрезентативности	Отсутствуют	Имеются: Для $\overline{p}$ $m_{\overline{p}} = \sqrt{\frac{\overline{p}(100-\overline{p})}{n}}$ Для $\overline{x}$ $m_{\overline{x}} = \frac{S_x}{\sqrt{n}}$

Средняя ошибка выборочной средней величины

- Выяснить близость выборочной средней к средней генеральной совокупности позволяет вычисление ошибки репрезентативности выборочной средней величины.
- Ошибку выборочной средней арифметической $m_{\bar{x}}$ вычисляют по формуле:

$$m_{\bar{x}} = \frac{S_x}{\sqrt{n}}$$

- где S_x - стандартное (среднее квадратическое) отклонение признака X, а n - число наблюдений в выборке.
- Ошибка выборочной средней - величина именованная, ее выражают в тех же единицах, что и среднюю арифметическую величину.

Доверительный интервал

- Пределы, в которых может находиться генеральная средняя (или вероятность случайного события), принято называть доверительными границами, а интервал, включающий ее возможное значение, - ***доверительным интервалом***.

Доверительный интервал для истинного значения средней арифметической

- Доверительный интервал для истинного значения средней арифметической (I_{M_x}) определяется исходя из величины выборочной средней с учетом ее предельной ошибки (т.е. ошибки, умноженной на некоторый коэффициент) по формуле:
- $I_{M_x} = \bar{x} \pm \varepsilon$, где $\varepsilon = t_\alpha \cdot m_{\bar{x}}$, т.е. $I_{M_x} = \bar{x} \pm t_\alpha \cdot m_{\bar{x}}$
- t_α носит название доверительного коэффициента и определяется по специальной таблице с заданной доверительной вероятностью.

Критические значения *t* по распределению Стьюдента

Число степеней свободы *n'*	Уровни значимости P_0		
	0,05 (5%)	0,01 (1%)	0,001 (0,1%)
1	12,71	63,66	636,58
2	4,30	9,92	31,60
3	3,18	5,84	12,92
4	2,78	4,60	8,61
5	2,57	4,03	6,87
6	2,45	3,71	5,96
7	2,36	3,50	5,41
8	2,31	3,36	5,04
9	2,26	3,25	4,78
10	2,23	3,17	4,59
12	2,18	3,05	4,32
14	2,14	2,98	4,14
16	2,12	2,92	4,01
18	2,10	2,88	3,92
20	2,09	2,85	3,85
22	2,07	2,82	3,79
24	2,06	2,80	3,75
30	2,04	2,75	3,65
40	2,02	2,70	3,55
120	1,98	2,62	3,37
∞	1,96	2,58	3,29
Доверительные вероятности $[1-P_0]$	0,95 (95%)	0,99 (99%)	0,999 (99,9%)

Доверительная вероятность

- *Доверительная вероятность* (α) характеризует надежность (достоверность, правильность) результатов выборочных медико-статистических исследований.
- В статистических исследованиях минимально приемлемым уровнем доверительной вероятности (надежности) выборочных показателей считается уровень 95 %.
- Величина, характеризующая вероятность ошибки статистического результата, носит название *уровня значимости*. Уровень значимости принято обозначать буквой *P*.
- Между доверительной вероятностью и уровнем значимости имеются следующие соотношения:

Доверительная вероятность α	0,95 (95%)	0,99 (99%)	0,999 (99,9%)
Уровень значимости $P = 1 - \alpha$	0,05 (5%)	0,01 (1%)	0,001 (0,1%)

- Для достаточно надежных статистических выводов вероятность ошибки (уровень значимости) не должна превышать 0,05 ($P < 0,05$, или $P < 5\%$).

Оценка точности

- Величина доверительного интервала позволяет произвести ***оценку точности*** обобщающей числовой характеристики, т.е. установить в каком интервале возможных значений находится истинное (для генеральной совокупности) значение данной числовой характеристики.

- На практике для построения доверительного интервала выборочной средней арифметической величины доверительный коэффициент t_α определяется по таблице с доверительной вероятностью не менее 0,95 (уровнем значимости не более 0,05) при числе степеней свободы $n' = n - 1$.

Пример 1

- Средняя длительность лечения некоторого заболевания ($\bar{X}$) у 10 больных составила 30 суток, а ошибка средней ($m_{\bar{x}}$) равна 1,7 суток.
- Доверительный интервал средней величины длительности лечения при $n' = 10 - 1 = 9$, $P_0 = 0{,}05$; $t_\alpha = 2{,}26$ будет составлять: 30 ± 2,26 · 1,7 суток
- Границы доверительного интервала будут иметь значения от 26,2 до 33,8 суток.

- Вывод: с вероятностью 95% истинное среднее значение (математическое ожидание) длительности лечения для данного заболевания будет находиться, при округлении до целых, в интервале от 26 до 34 суток.

Доверительный интервал для частости случайного события

- Как и для средних величин, с учетом найденного значения $m_{\bar{p}}$

$$m_{\bar{p}} = \sqrt{\frac{\bar{p}(100 - \bar{p})}{n}}$$

может быть определен доверительный интервал для истинного значения относительного показателя частоты ($\boldsymbol{I_P}$) с нужной (выбранной нами) доверительной вероятностью, т.е. может быть проведена оценка точности относительной величины:

$$I_P = \bar{p} \pm t_\alpha \cdot m_{\bar{p}}$$

Пример 2.

- С целью изучения эффективности противогриппозной вакцины были привиты 280 военнослужащих срочной службы. Из них гриппом заболели 60 человек. Вычислить ошибку репрезентативности и определить 99% доверительный интервал для показателя частости заболевания гриппом.
- По условию задачи частость заболевания гриппом военнослужащих в данной выборке составляет:

$$\overline{p} = \frac{m}{n} = \frac{60}{280} = 0{,}214$$

- или 21,4%. Подставим данные в формулу расчёта величины ошибки показателя частоты:

$$m_{\overline{p}} = \sqrt{\frac{\overline{p} \cdot (100 - \overline{p})}{n}} = \sqrt{\frac{21{,}4 \cdot (100 - 21{,}4)}{280}} = 2{,}45\%$$

- Доверительный интервал для вероятности (при $n' \to \infty$, P = 0,01, t_{99} = 2,58)
- составит: $I_P = \overline{p} \pm t_{99} \cdot m_{\overline{p}} = 21{,}4 \pm 2{,}58 \cdot 2{,}45(\%)$

т.е. $(15{,}08 \div 27{,}72)\%$.

- Следовательно, с надежностью (0,99) 99% (уровнем значимости 0,01(1%) можно утверждать, что частота заболеваемости гриппом среди всех военнослужащих срочной службы при применении данной вакцины будет находиться в интервале от 15,1% до 27,7%.

Определение требуемого числа наблюдений в выборке

- В случае количественного признака (X) предельная ошибка определяется по формуле:

$$\varepsilon = t_\alpha \cdot m_{\bar{x}} = \frac{t_\alpha \cdot S_x}{\sqrt{n}}$$

в которую входит величина n - число наблюдаемых случаев.

- Решая приведенное равенство относительно n, получим формулу для определения требуемого числа наблюдений:

$$n = \frac{t_\alpha^2 \cdot S_x^2}{\varepsilon^2}$$

- Величина ε определяется исследователем, исходя из необходимой точности результатов. Исследователь сам устанавливает, для какой доверительной вероятности необходимо определить величину предельной ошибки, что находит отражение при выборе значения коэффициента t по таблице. Среднее квадратическое (стандартное) отклонение определяется либо на небольшой (пробной) выборке, либо на основании ранее проведенных исследований.

Пример 3

- Изучалась эффективность новой методики лечения острых гнойных заболеваний пальцев и кисти по показателю средней длительности лечения больных. У 16 пациентов (*n = 16*) она составила $\overline{x}$ = 15,7 дня, при S_x = 8,9 дня и $m_{\overline{x}}$ = 5,4 дня (ошибка большая).

- Следует определить минимально необходимое число наблюдений, при котором с надежностью 95% предельная ошибка длительности лечения ε не превысила бы 3 дня.

- Для решения воспользуемся формулой (при $n' = n - 1 = 15$, $t_{95} = 2{,}13$):

- $$n = \frac{2{,}13^2 \cdot 8{,}9^2}{3^2} \approx 40 \text{ наблюдений}$$

- *Вывод:* с вероятностью 95% предельная ошибка средней длительности лечения данной категории больных не превысит 3-х дней, при минимальной численности выборки ≥ 40 наблюдений.

Определение требуемого числа наблюдений в выборке

Для показателей частоты:
предельная ошибка определяется по формуле:

$$\varepsilon = t_{\alpha} \cdot m_{\bar{p}} = t_{\alpha} \cdot \sqrt{\frac{\bar{p} \cdot (100 - \bar{p})}{n}}$$

Следовательно, минимальное число необходимых наблюдений для оценки вероятности случайного события при условии, что максимальная ошибка при этом составит не более заданной величины ε, может быть рассчитано по формуле:

$$n = \frac{t_{\alpha}^{2} \cdot \bar{p} \cdot (100 - \bar{p})}{\varepsilon^{2}}$$

ПРИМЕР 4

- Уровень заболеваемости ОРВИ рядового состава в одном из подразделений воинской части в начальный период вспышки *(n = 30)* составил 21,0 ± 7,4%. Учитывая большую величину ошибки репрезентативности, определить требуемое число наблюдений, чтобы предельная ошибка ε показателя уровня заболеваемости не превысила 6%, с доверительной вероятностью 95% (при n' = 30 -1 = 29; t_{95} = 2,04).
- **Решение:**
- Подставим исходные данные в формулу определения требуемого числа наблюдений:

$$n = \frac{2{,}04^2 \cdot 21 \cdot (100 - 21)}{6^2} \approx 192$$

- **Вывод 1:** с вероятностью 95% при числе наблюдений в выборке 192 предельная ошибка уровня заболеваемости ОРВИ не превысит 6%
- **Вывод 2:** уменьшение ошибки влечёт за собой увеличение числа наблюдений в выборке.

Лекция окончена