Статистика неколичественных данных

Лекция №1

для студентов 3 курса,
обучающихся по специальности 30.05.03 – Медицинская кибернетика

Романова Н.Ю., доцент кафедры медицинской и биологической физики

План лекции:

1. Актуальность темы. Анализ качественных признаков.
2. Вычисление параметров распределения качественных признаков.
3. Описание относительной частоты бинарного признака с использованием доверительного интервала.
4. Сравнение групп по качественному признаку.
5. Анализ таблиц сопряженности.
6. Заключение

Актуальность темы

Ряд медицинских исследований оперирует статистическими данными качественного характера:

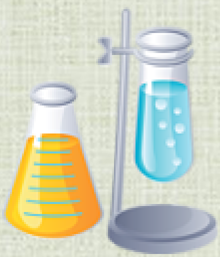
Да/нет, пол, буквенные аббревиатуры (коды заболеваний),

градации самочувствия и т.д.

Вспомнить типы измерительных шкал!

- **Качественные признаки (переменные, данные) могут быть номинальными (номинативные), в частности, бинарными, и порядковыми.**

Номинальными **называются признаки, значения которых представляют собой условные коды неизмеряемых категорий (например, коды пола или коды диагноза). Значения таких величин не могут быть упорядочены по какому-либо принципу по возрастанию или убыванию выраженности признака.**

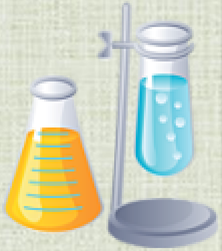


Порядковыми (ранговыми) называются признаки, значения которых отражают степень выраженности какой-либо характеристики объекта исследования (например, стадии заболевания, степени выраженности симптома).

При этом в отличие от количественных признаков "расстояния" между значениями порядковых признаков не могут быть оценены с помощью какой-либо известной шкалы, однако все же значения порядкового признака могут быть упорядочены (ранжированы).

Порядковые признаки являются качественными (иногда их называют "полуколичественными") оценками какой-либо характеристики.

При достаточно большом числе возможных значений качественного порядкового признака (обычно приближающемся к 20-30) на практике для анализа таких признаков могут применяться те же параметрические методы, что и для количественных признаков.


Все объекты исследования выборки могут быть разделены на подгруппы в соответствии со значениями любого из имеющихся качественных признаков, например, по полу. Число объектов исследования с определенным значением качественного признака называется ***абсолютной частотой* - n_i**.

Относительная частота - w_i— это отношение числа объектов с каким-либо значением признака к общему числу объектов ***n***.

Анализ качественных данных начинается с простейшего подсчета абсолютных и относительных частот для каждого значения (градации) такого признака.

Основные способы описания качественных данных:

вычисление параметров распределения качественных данных:

- *моды* (для номинальных данных)
- *медианы, моды и квартилей* (для порядковых данных);
- вычисление *абсолютных и относительных частот* (*пропорций, долей, процентов)* и *доверительных интервалов* для них.

Исследование распределения бинарного признака

- ***Доля, вероятность, пропорция*** **- это относительная частота события, выраженная в десятичных долях единицы или в процентах. Изменяется в интервале от 0 до 1, или от 0 до 100%.**

- ***Шанс*** **— отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет. Другое определение: шанс - это отношение частоты возникновения события к частоте его отсутствия. Значение шанса изменяется в интервале от нуля до бесконечности.**

Пример: **если вероятность возникновения осложнений после операции равна В (например, 0,3), то вероятность того, что оно не произойдет, равна 1 - В (в нашем примере 0,7). Шанс события равен отношению этих вероятностей (в нашем примере 0,3:0,7=0,43).**

Числовые характеристики распределения выборки

Распределение бинарного признака - это, по сути, биномиальное распределение с параметром в виде доли проявления данного признака в совокупности - *p*.

Обозначим наличие признака за 1, отсутствие - за 0, m - количество объектов, обладающих признаком X, n - число наблюдений;

Тогда математическое ожидание такой величины будет рассчитываться следующим образом:

$$M(x)=\mu=\frac{1\cdot m+0\cdot(n-m)}{n}=\frac{m}{n}=p$$

А среднеквадратическое отклонение и стандартная ошибка выборки -

$$\sigma=\sqrt{\frac{(1-p)^2 m+(0-p)^2(n-m)}{n}}=$$

$$\boxed{=\sqrt{p\cdot(1-p)}},\ \text{учитывая, что}\ p=m/n$$

$$s_p=\frac{\sigma}{\sqrt{n}}=\sqrt{\frac{p(1-p)}{n}}$$

Что вполне согласуется с характеристиками биномиального распределения.

Поправка Ван-дер-Вардена (для p=0 и p=1)

$$p_в = \frac{(p_1+1)\cdot 100}{n+2}$$

$$s_{pв} = \sqrt{\frac{p(100-p)}{n+3}}$$

- для вероятности, выраженной в процентах

$$p_в = \frac{m+1}{n+2}$$

$$s_{pв} = \sqrt{\frac{p(1-p)}{n+3}}$$

- для вероятности, выраженной в долях

Пример:

n = 30

p = 0

$$p_в = \frac{(0+1)\cdot 100}{30+2} = 3{,}1\%$$

$$s_{pв} = \sqrt{\frac{3{,}1\cdot 96{,}9}{30+3}} = \sqrt{9{,}1} = 3{,}2\%$$

Результат представляется в виде:

p=0+3,2%

p=100-3,2%

Описание относительной частоты бинарного признака с использованием доверительного интервала (ДИ)

- **В современной научной литературе относительные частоты бинарных признаков — "событий" (т.е. признаков, имеющих только два возможных значения -"да" и "нет") — принято приводить вместе с указанием их ДИ.**
- **ДИ - это интервал, в котором с некоторой вероятностью (например, 95%) находится истинное популяционное значение. Границы ДИ вычисляют на основании анализа данных выборки.**

Доверительный интервал для доли

$$P \pm t \cdot s_p$$

или с поправкой за непрерывность:

$$I = P \pm t\left(\sqrt{\frac{P(1-P)}{n}} + \frac{1}{2n}\right)$$

Ограничения:
np>=5
nq>=5
0.3<p<0.7

Этот метод носит название «**метод Вальда**»

Коррекция по Агрести – Коуллу

Коррекция по Агрести – Коуллу представляет собой замену в формуле Вальда частоты встречаемости признака в выборке (p) на p`, при расчете которой к числителю добавляется 2, а к знаменателю добавляется 4, то есть p` = (X + 2) / (N + 4), где X – количество участников исследования, у которых имеется изучаемый признак, а N – объем выборки.

При больших выборках используются значения *z (1,96; 2,58)* для малых выборок рекомендуется подставлять значение *t Стьюдента для (n – 1)* степеней свободы

При частотах, не превышающих 25 % или превышающих 75 %, некоторые авторы рекомендуют рассчитывать доверительный интервал с помощью *arcsin-преобразования* (*угловое преобразование Фишера*):

$$\varphi = 2\arcsin\sqrt{p}$$

Стандартная ошибка вспомогательной переменной рассчитывается по формуле:

$$s_{\varphi} = \frac{1}{\sqrt{N}}$$

Новая переменная φ имеет нормальное распределение

Нижняя и верхняя границы для частоты встречаемости признака в генеральной совокупности будут выглядеть следующим образом:

$$p_н = sin^2 \frac{\varphi - z \cdot s_\varphi}{2}$$

$$p_в = sin^2 \frac{\varphi + z \cdot s_\varphi}{2}$$

Для малых выборок используется t-распределение с *n – 1* степенями свободы

Данный метод не дает отрицательных значений и позволяет более точно оценить доверительные интервалы для частот, чем метод Вальда.

Метод Клоппера-Пирсона

(с учетом биномиального распределения)

Нижняя граница

$$\underline{p} = \frac{n_a F_{1-P_2}\left(2n_a; 2\left(n-n_a+1\right)\right)}{n-n_a+1+n_a F_{1-P_2}\left(2n_a; 2\left(n-n_a+1\right)\right)} = \frac{n_a}{n_a+\left(n-n_a+1\right) F_{P_2}\left(2\left(n-n_a+1\right); 2n_a\right)},$$

Верхняя граница

$$\overline{p} = \frac{\left(n_a+1\right) F_{1-P_1}\left(2(n_a+1); 2\left(n-n_a\right)\right)}{n-n_a+\left(n_a+1\right) F_{1-P_1}\left(2(n_a+1); 2\left(n-n_a\right)\right)}.$$

В этих формулах $F_a\left(f_1; f_2\right)$ - квантиль F - распределения с f_1 и f_2 степенями свободы уровня a .

$p_1 = \alpha/2$

$p_2 = 1 - \alpha/2$

$n_a \equiv m$

FРАСПОБР($\alpha/2$, m, n)!

По мнению многих статистиков, наиболее оптимальную оценку доверительных интервалов для относительных частот осуществляет метод Уилсона, предложенный в 1927 г:

$$p_{1.2} = \frac{n}{t^2 + n}\left[w + \frac{t^2}{2n} \pm t \cdot \sqrt{\frac{w(1-w)}{n} + \left(\frac{t}{2n}\right)^2} \right]$$

где *w*–относительная частота события, t – нормированное отклонение для заданного уровня надежности (γ=0,95, 0,99…) .

Данный метод не только позволяет оценить доверительные интервалы как для очень малых и очень больших частот, но и применим для малого числа наблюдений.

Доверительные интервалы, рассчитанные шестью разными способами для двух примеров, описанных в тексте

Способ расчета доверительного интервала	95% ДИ для **X=1, N=20,** P=0,0500, или 5%	95% ДИ для **X=450, N=1000,** P=0,4500, или 45%
Вальда	–0,0455–0,2541	0,4192–0,4810
Вальда с коррекцией по Агрести – Коуллу	<,0001–0,2541	0,4194–0,4810
Уилсона	0,0089–0,2361	0,4194–0,4810
Уилсона с коррекцией на непрерывность	0,0026–0,2694	0,4189–0,4815
«Точный метод» Клоппера – Пирсона	0,0013–0,2487	0,4189–0,4814
Угловое преобразование	<0,0001–**0,1967**	0,4193–0,4809

Сравнение долей

Для сравнения выборочных долей двух несвязных совокупностей применяется критерий нормированного отклонения

$$z = \frac{\textit{Разность выборочных долей}}{\textit{Стандартная ошибка разности выборочных долей}}$$

где стандартная ошибка -

$$s_{p_1 - p_2} = \sqrt{s_{p_1}^2 + s_{p_2}^2}$$

$$z = \frac{p_1 - p_2}{s_{p_1 - p_2}} = \frac{p_1 - p_2}{\sqrt{s_{p_1}^2 + s_{p_2}^2}}$$

$$s_{p_1} = \sqrt{\frac{p_1(1 - p_1)}{n_1}} \qquad s_{p_2} = \sqrt{\frac{p_2(1 - p_2)}{n_2}}$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$p = \frac{m_1 + m_2}{n_1 + n_2}$$

Пример:

группа 1 - пациенты курят:

n_1=25 инфаркт миокарда (ИМ) наблюдается у m_1=18

группа 2 - пациенты не курят

n_2=19 ИМ m_2=6

Нулевая гипотеза: частота ИМ не зависит от курения

Решение:

$$p_1 = \frac{18}{25} = 0{,}72 \qquad p_2 = \frac{6}{19} = 0{,}316$$

$$p = \frac{18+6}{25+19} = \frac{24}{44} = 0{,}545$$

$$z = \frac{0{,}72 - 0{,}316}{\sqrt{0.545(1-0.545)\left(\frac{1}{25}+\frac{1}{19}\right)}} = \frac{0{,}404}{0{,}15} = 2{,}69$$

$z_{кр}=1{,}96$ p=0,95 (α=0,05) ($t_{кр}=2{,}02$)

$z_{кр}=2{,}58$ p=0,99 (α=0,01) ($t_{кр}=2{,}70$)

Различия достоверны для α=0,05,

(для α=0,01 при использовании распределения Стьюдента различия недостоверны)

http://linda6035.ucoz.ru/

Поправка Йетса

В случае, когда значение доли меньше 0,1 (но более 0,05) необходимо вводить поправку:

$$z=\frac{\left|p_1-p_2\right|-\frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}{\sqrt{p\left(1-p\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Для нашего примера:

$$z=\frac{(0{,}72-0{,}316)-0{,}046}{\sqrt{0{,}545(1-0{,}545)\left(\frac{1}{25}+\frac{1}{19}\right)}}=\frac{0{,}358}{0{,}15}=2{,}39$$

Различия достоверны 2,39>1,96 (α<0,05)

φ-преобразование Фишера

Сравниваемые доли выражают в процентах с введением *поправки Йетса на непрерывность*.

$$\varphi_{1,2} = 2\arcsin\sqrt{P_{1,2}}$$

$$t_\varphi = (\varphi_1 - \varphi_2)\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \geqslant t_{st}$$

- условие значимости различий

Поправка на группировку: если $p_1 > p_2$

$$p_1 \to p_1' = p_1 - \frac{1}{2n_1} \quad \text{или} \quad \frac{m_1}{n_1} \to \frac{m_1 - 0{,}5}{n_1}$$

$$p_2 \to p_2' = p_2 + \frac{1}{2n_2} \quad \text{или} \quad \frac{m_2}{n_2} \to \frac{m_2 + 0{,}5}{n_2}$$

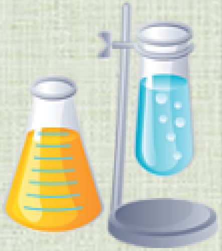


Таблица φ-критерия Фишера

0,а	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
,00	0,000	0,200	0,284	0,348	0,403	0,451	0,495	0,536	0,574	0,609
,10	0,644	0,676	0,707	0,738	0,767	0,795	0,823	0,850	0,876	0,902
,20	0,927	0,952	0,976	1,000	1,024	1,047	1,070	1,093	1,115	1,137
,30	1,159	1,182	1,203	1,224	1,245	1,266	1,287	1,308	1,328	1,349
,40	1,369	1,390	1,410	1,430	1,451	1,471	1,491	1,511	1,531	1,551
,50	1,571	1,591	1,611	1,631	1,651	1,671	1,691	1,711	1,731	1,752
,60	1,772	1,793	1,813	1,834	1,855	1,875	1,897	1,918	1,939	1,961
,70	1,982	2,004	2,026	2,049	2,071	2,094	2,118	2,141	2,165	2,190
,80	2,214	2,240	2,265	2,292	2,319	2,246	2,375	2,404	2,434	2,465
,90	2,498	2,532	2,568	2,606	2,647	2,691	2,739	2,793	2,858	2,941

Определить значимость различий в предыдущей задаче, используя угловое преобразование Фишера.

$$p_1 = \frac{18-0,5}{25} = 0,7 = 70\%$$

$$p_2 = \frac{6+0,5}{19} = 0,342 = 34,2\%$$

По таблице для φ Фишера $\varphi_1 = 1,982$, $\varphi_2 = 1,249$

$$t_\varphi = (1.982-1.249)\sqrt{\frac{25\cdot 19}{25+19}} = 0.733\sqrt{\frac{475}{44}} = 2,81 > 2,58$$

- То есть различия значимы на уровне < 0,01

Те же самые данные, у которых мы сравнивали доли, можно обработать с помощью такого статистического «инструмента», как *таблица сопряженности*.

Таблица сопряженности 2х2

признак 1	признак 2		всего
	есть	нет	
Есть	a	b	a+b
Нет	c	d	c+d
Всего	a+c	b+d	n= =a+b+c+d

Задача анализа таблицы сопряженности – поиск математического описания закономерностей, содержащихся в данных

Нулевая гипотеза H_0:

между признаками 1 и 2 нет зависимости.

Альтернативная:

Зависимость есть.

Применение критерия «хи-квадрат» для проверки H_0 в таблицах сопряженности

$$\chi^2 = \sum \frac{(\textit{наблюдаемое число} - \textit{ожидаемое число})^2}{\textit{ожидаемое число}} -$$

сумма по всем клеткам

$$\chi^2 = \sum \frac{(n_н - n_о)^2}{n_о}$$

В этом случае критическая область значения χ^2 – область, лежащая за 95% (99%) интервалом распределения χ^2 с (r-1)(c-1) степенями свободы ($r \times c$ - размерность таблицы).

Если величина χ^2 больше критического значения, нуль-гипотеза о независимости признаков опровергается.

Ограничения:

- Общее количество данных не менее 30
- Значения частот в ячейках - не менее 5

Пример.

Курение	Инфаркт миокарда		
	Да	Нет	Сумма
Да	18	7	25
Нет	6	13	19
Сумма	24	20	44

Рассчитаем ожидаемые частоты:

Курение	Инфаркт миокарда		
	Да	Нет	Сумма
Да	18 0,545*25=13,64	7 0,455*25=11,36	25
Нет	6 0,545*19=10,36	13 0,455*19=8,64	19
Сумма	24	20	44

$$p = \frac{24}{44} = 0{,}545 \qquad q = \frac{20}{44} = 0{,}455$$

Или рассчитываем теоретические частоты, перемножая соответствующие маргинальные и деля это произведение на общую сумму частот.

$$\chi^2 = \sum \frac{(n_{\text{н}} - n_{\text{о}})^2}{n_{\text{о}}}$$

$$\chi^2 = \frac{(18 - 13{,}64)^2}{13{,}64} + \frac{(7 - 11{,}36)^2}{11{,}36} + \frac{(6 - 10{,}36)^2}{10{,}36} + \frac{(13 - 8{,}64)^2}{8{,}64} = 7{,}10$$

$$\chi^2 = \sum \frac{\left(|n_{\text{н}} - n_{\text{о}}| - \frac{1}{2}\right)^2}{n_{\text{о}}}$$

- с поправкой Йетса

$$\chi^2 = \frac{\left(|18 - 13{,}64| - \frac{1}{2}\right)^2}{13{,}64} + \frac{\left(|7 - 11{,}36| - \frac{1}{2}\right)^2}{11{,}36} + \frac{\left(|6 - 10{,}36| - \frac{1}{2}\right)^2}{10{,}36} + \frac{\left(|13 - 8{,}64| - \frac{1}{2}\right)^2}{8{,}64} = 5{,}57$$

$\chi^2 = 5{,}57 > \chi^2$ (крит) $= 3{,}84$

Таким образом, курение и частота инфарктов связаны.

Процедура использования критерия хи-квадрат в случае таблицы (r x c) точно такая же, как и при анализе таблицы (2 x 2).

Точный критерий Фишера

Критерий хи-квадрат применяется, когда общее число N > 30; a, b, c, d >5. Если эти условия не выполняются, применяется точный критерий Фишера

$$p=\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}.$$

Достоинством метода является соответствие полученного критерия точному значению уровня значимости *p*. Необходимо лишь сопоставить данное число с критическим уровнем значимости, обычно принимаемым в медицинских исследованиях за 0,05 или 0,01.

Пример: изучается зависимость частоты рождения детей с врожденными пороками развития (ВПР) от курения матери во время беременности.

	Исход есть (Наличие ВПР)	Исхода нет (Отсутствие ВПР)	**Всего**
Фактор риска есть (Курящие)	A = 10	B = 70	**(A + B) = 80**
Фактор риска отсутствует (Некурящие)	C = 2	D = 88	**(C + D) = 90**
Всего	**(A + C) = 12**	**(B + D) = 158**	**(A + B + C + D) = 170**

В результате вычислений находим, что Р = 0,0137

То есть, полученное в нашем примере значение 0,0137 и есть уровень значимости различий сравниваемых групп по частоте развития ВПР плода.

Р < 0,05, в связи с чем делаем вывод о наличии прямой взаимосвязи курения и вероятности развития ВПР плода. Частота возникновения врожденной патологии у детей курящих женщин *статистически значимо выше*, чем у некурящих

Таблицы сопряженности (r x c)

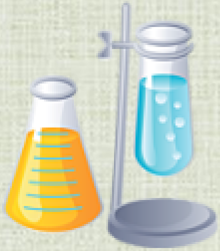
Конституция	Мужчины	Женщины	
Астеники	5/	8/	
Нормостеники	25/	30/	
Гиперстеники	12/	20/	

Процедура использования критерия хи-квадрат в случае таблицы (r x c) точно такая же, как и при анализе таблицы (2 x 2).

СЧИТАЕМ!

Сравнение *зависимых выборок* путем анализа таблиц сопряженности

Критерий Мак-Немара - является аналогом параметрического критерия Стьюдента и непараметрического критерия Т-Вилкоксона, применяется для анализа связанных измерений в случае изменения реакции с помощью дихотомической переменной.



По результатам такого исследования строится результирующая таблица 2x2 в виде:

ДО/ПОСЛЕ	0	1	Всего
1 0	A C	B D	A + B C + D
Всего...	A + C	B + D	N

В клетках A и D представлены изменения от ДО к ПОСЛЕ, причем в клетке A изменения благоприятных результатов на неблагоприятные, а в клетке D - наоборот. **Нулевая гипотеза** состоит в том, что в генеральной совокупности доля тех, кто изменяет благоприятную реакцию на неблагоприятную в результате воздействия, равна доле тех, кто изменяет реакцию в обратном порядке. Объем выборки N определяется как сумма частот в диагональных клетках A и D.

Для проверки гипотезы в случае с N > 50 рассчитывается статистика Хи-квадрат.

$$\chi^2 = \frac{\left(|A-D|-1\right)^2}{A+D}$$

Если рассчитанное значение статистики превосходит критическое значение (рассчитанное исходя из объема выборки N и уровня значимости), нулевая гипотеза отвергается.

__Пример__. В результате исследования была получена таблица:

ДО/ПОСЛЕ	0	1	Всего
1 0	26 38	30 32	56 70
Всего	64	62	126

$$\chi^2_{эмп} = \frac{\left(|26-32|-1\right)^2}{26+32} = 0,43$$

$\chi^2_{кр}$ (0,01)=6,63

$\chi^2_{кр}$(0,05)=3,84

Рассчитанное значение критерия меньше критического табличного: мы не можем отвергнуть нулевую гипотезу об отсутствии различий между показателями ДО и ПОСЛЕ на выбранном уровне значимости.

Итак, нами рассмотрены:

- Методы анализа качественных признаков.
- Способы сравнения групп по качественному признаку.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА:

Основная литература:

1. Попов А.М. Теория вероятней и математическая статистика /А.М. Попов, В.Н. Сотников. – М.: ЮРАЙТ, 2011. – 440 с.
2. Герасимов А. Н. Медицинская статистика: учебное пособие / А. Н. Герасимов. – М. : Мед. информ. агентство, 2007. – 480с.
3. Балдин К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин. – М. : Флинта, 2010. – 488с.

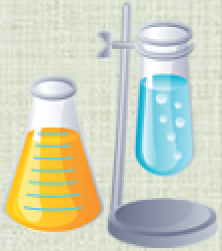
Экология человека 2008.05 Практикум

УДК 31:61

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ ДЛЯ ЧАСТОТ И ДОЛЕЙ

© 2008 г. **А. М. Гржибовский**

Национальный институт общественного здоровья, г. Осло, Норвегия


БЛАГОДАРЮ ЗА ВНИМАНИЕ!