# Непараметрическая статистика

**Лекция №8**
**для студентов 2 курса,**
**обучающихся по специальности 060609 – Медицинская кибернетика**
**доц. Шапиро Л.А.**
**Красноярск, 2015 г.**

# План лекции:

- **Актуальность темы**
- **Описательная статистика для признаков, не подчиняющихся нормальному закону распределения.**
- **Непараметрические критерии достоверности различия двух зависимых совокупностей**
- **Непараметрические критерии определения достоверности различия двух независимых совокупностей**
- **Заключение**

# Закон распределения-нормальный?

Параметрическая статистика

Да

M±σ, M±m,
M (95% ДИ)

Сравнение 2-х выборок по критерию Стьюдента

Корреляция по Пирсону

Нет

Me [25%-75%],
Mo, Min-Max

Сравнение 2-х выборок по критериям Манна-Уитни, Вилкоксона

Корреляция по Спирмену

Непараметрическая статистика

# Актуальность темы

**Параметрические** методы статистики – совокупность методов проверки статистических гипотез, основывающиеся на знании свойств генеральных совокупностей, из которых получены данные.

Однако часто свойства генеральных совокупностей неизвестны. Тогда следует применять **непараметрические** методы статистики.

Непараметрические методы требуют немногих предположений относительно генеральных совокупностей, из которых извлечены данные.

Непараметрические методы проще в применении, но менее чувствительны.

Непараметрические методы применимы в ситуациях, когда методы нормальной теории не работают.

# Описательная статистика для признаков, не подчиняющихся нормальному закону распределения.

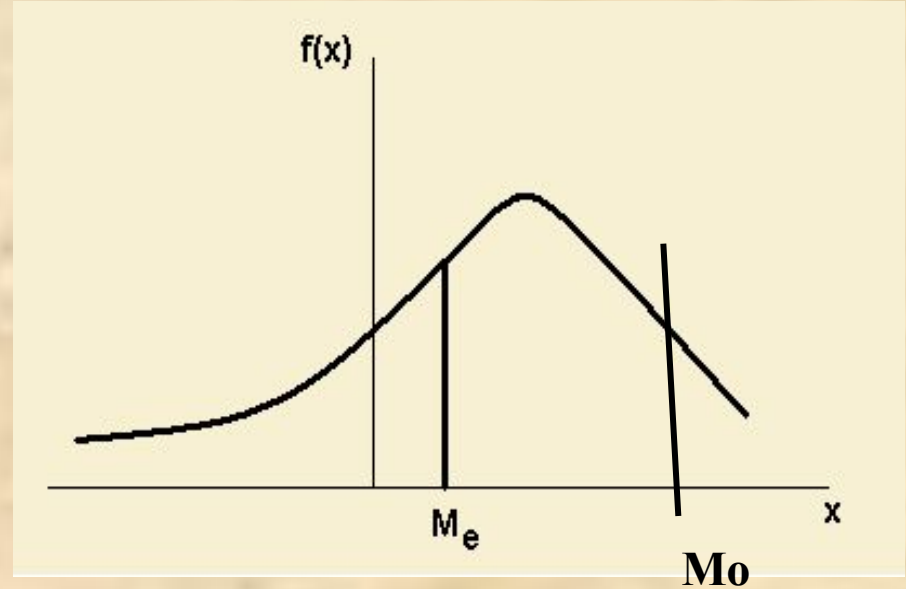
## Медиана и мода случайной величины

$M_e$ – такое значение случайной величины **x**, для которого выполняется следующее условие:

$$P(x < M_e) = P(x > M_e)$$

Геометрическая медиана - это абсцисса точки, в которой площадь ограниченная кривой плотности распределения, делится пополам.

$$\int_{-\infty}^{M_e} f(x)dx = \int_{M_e}^{+\infty} f(x)dx$$




**Мода** – значение СВ, при котором f(x)=max

Для характеристики структуры совокупности используются **квантили**.

**Квантили** характеризуют варианты значений признака, занимающие определенное место в ранжированной совокупности.

К квантилям относят такие характеристики как **медиана, квартили, квинтили, децили и перцентили.**

**Медианой** (англ. median) называется значение исследуемого признака, справа и слева от которого находится одинаковое число упорядоченных элементов выборки.

Также, как и среднее арифметическое, медиана дает общее представление о том, где находится центр выборки.

Рассмотрим способы определения медианы при различных значениях N. Для нахождения медианы измерения записывают в ряд по возрастанию значений. Если число измерений N нечетное, то медиана численно равна значению этого ряда, стоящему точно в середине, или на (N+1)/2 месте.

Например, медиана пяти измерений: 10, 17, 21, 24, 25 – равна 21 – значению, стоящему на третьем месте (N+1)/2=(5+1)/2=3.

Если число измерений четное, то медиана численно равна среднему арифметическому значений ряда, стоящих в середине, или на N/2 и (N/2)+1 местах.

Например, медиана восьми измерений: 5, 5, 6, 7, 8, 8, 9, 9 – равна 7,5 (7+8)/2=7,5 – среднему арифметическому значений ряда, стоящих на четвертом и пятом местах (N/2=8/2=4 и N/2+1=4+1=5).

# Мода (Mo)

Мода (англ. mode) представляет собой наиболее часто встречающееся значение переменной (иными словами, наиболее «модное» значение переменной). Сложность состоит в том, что редкая выборка имеет единственную моду. Если в выборке несколько мод, то говорят, что она мультимодальна или многомодальна (имеет два или более «пика»). Таким образом можно сказать, что мода характеризует не только положение выборки, но отчасти и форму ее распределения.

Например: 2, 6, 6, 8, 9, 9, 9, 10 –
мода = 9.

**Квартили** представляют собой значения, которые делят две половины выборки (разбитые медианой) еще раз пополам (от слова кварта — четверть).

**Нижнюю квартиль** часто обозначают символом 25% ($Q_1$), это означает, что 25% значений переменной меньше нижней квартили.

**Верхнюю квартиль** часто обозначают символом 75% ($Q_3$), это означает, что 75% значений переменной меньше верхней квартили.

**Интерквартильный размах:**
Me $[Q_1; Q_3]$

# Межквартильный размах

- **Пример:**

**1 2 4 7 8 9 10 12**

$Me=(7+8)/2=7,5$

$Q_1=(2+4)/2=3 \quad Q_3=(9+10)/2=9,5$

**Квинтили-**это значения признака в упорядоченной по возрастанию совокупности, которые делят совокупность на пять равных частей. Ниже $К_1$-20% значений.

**Децили**-это значения признака в упорядоченной по возрастанию совокупности, которые делят совокупность на 10 равных частей. Ниже $D_1$-10% значений.

**Перцентили**-это значения признака в упорядоченной по возрастанию совокупности, которые делят совокупность на 100 равных частей.

**Вариационный размах** (размах распределения) характеризует разницу между максимальным и минимальным значением признака в изучаемой совокупности:

$R=X_{min}-X_{max}$

# Выборочные характеристики: среднее, медиана и ранг

| выборка | № | значение 1 | значение 2 |
|---|---|---|---|
| 6 | 1 | 2 | 2 |
| 8 | 2 | 2 | 2 |
| 7 | 3 | 4 | 4 |
| 6 | 4 | 6 | 6 |
| 15 | 5 | 6 | 6 |
| 7 | 6 | 6 | 6 |
| 4 | 7 | 7 | 7 |
| 2 | 8 | 7 | 7 |
| 7 | 9 | 7 | 7 |
| 6 | 10 | 8 | 8 |
| 2 | 11 | 15 | 9 |
| | среднее | 6,364 | 5,818 |
| | медиана | 6 | 6 |

**Ранг**-место варианты в упорядоченном ряду.

| № | значение | ранг |
|---|---|---|
| 1 | 2 | (1+2)/2=1,5 |
| 2 | 2 | 1,5 |
| 3 | 4 | 3 |
| 4 | 6 | (4+5+6)/3=5 |
| 5 | 6 | 5 |
| 6 | 6 | 5 |
| 7 | 7 | (7+8+9)/3=8 |
| 8 | 7 | 8 |
| 9 | 7 | 8 |
| 10 | 8 | 10 |
| 11 | 15 | 11 |

# Основные задачи непараметрической статистики

Любое распределение можно охарактеризовать параметром положения, характеризующим центр группирования случайных величин, и параметром масштаба, характеризующим степень рассеяния случайных величин.

Когда закон распределения неизвестен, гипотезы о параметрах положения и масштаба производятся с помощью непараметрических критериев. Таким образом, в непараметрической статистике существуют **две основные задачи** – задача оценки сдвига положения, и задача оценки изменения масштаба.

# Задача оценки сдвига: измерения фактора IV по шкале депрессии до и после принятия транквилизатора

| пациент | X(i) | Y(i) |
|---|---|---|
| 1 | 1,83 | 0,878 |
| 2 | 0,5 | 0,647 |
| 3 | 1,62 | 0,598 |
| 4 | 2,48 | 2,05 |
| 5 | 1,68 | 1,06 |
| 6 | 1,88 | 1,29 |
| 7 | 1,55 | 1,06 |
| 8 | 3,06 | 3,14 |
| 9 | 1,3 | 1,29 |

Шкала депрессии Гамильтона характеризует уровень суицидальности пациента. Чем меньше коэффициент,тем лучше состояние больного.

# Критерий знаков

Статистическая модель: разность Z(i) является случайно выбранным наблюдением. Совокупности Z(i) имеют одну и ту же медиану. Нулевая гипотеза: общая медиана равна нулю.

Вычисление критериальной статистики:

1. Запишем знак разности для каждой пары значений признака.
2. Подсчитаем числа N(+) и N(-) разностей одного знака и
3. Выберем число $G_{эмп}=\min(N(+), N(-))$.
4. Найдем $G_{крит}$ для $n=N_{max}$ и $\alpha=0,05$

Если $G_{эмп} \leq G_{крит}$ нулевая гипотеза отвергается. Различия статистически значимы.

Если $G_{эмп} > G_{крит}$ нулевая гипотеза не отвергается. Различия статистически не значимы.

# Биномиальное распределение как основа статистики критерия знаков

Если курс лечения не приводит к изменениям, то характеристики пациента до и после лечения будут примерно одинаковыми, разница между этими величинами будет случайной, и число положительных значений разности будет близко к числу отрицательных значений

$$p_n(m) = C_n^m p^m (1-p)^{n-m}$$

$$p_n(n/2) = \frac{n!}{(n-n/2)!(n/2)!}\left(\frac{1}{2}\right)^{n/2}\left(\frac{1}{2}\right)^{n/2}$$

биномиальный критерий

# Критерий знаков

| пациент | X | Y | Z | sign |
|---|---|---|---|---|
| 1 | 1,83 | 0,878 | -0,952 | -1 |
| 2 | 0,5 | 0,647 | 0,147 | 1 |
| 3 | 1,62 | 0,598 | -1,022 | -1 |
| 4 | 2,48 | 2,05 | -0,43 | -1 |
| 5 | 1,68 | 1,06 | -0,62 | -1 |
| 6 | 1,88 | 1,29 | -0,59 | -1 |
| 7 | 1,55 | 1,06 | -0,49 | -1 |
| 8 | 3,06 | 3,14 | 0,08 | 1 |
| 9 | 1,3 | 1,29 | -0,01 | -1 |

$N(+)=2$; $N(-)=7$; $G_{эмп} = \min(2, 7)=2$; $G_{крит}(0,05,7)=0$

$G_{эмп} > G_{крит}$ $(2>0)$ Нулевая гипотеза не отвергается.
Различия статистически не значимы.

# КРИТИЧЕСКИЕ ЗНАЧЕНИЯ КРИТЕРИЯ *G* ЗНАКОВ

## (для проверки ненаправленных альтернатив)

| *n* | *p* | | *n* | *p* | | *n* | *p* | | *n* | *p* | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0,05 | 0,01 | | 0,05 | 0,01 | | 0,05 | 0,01 | | 0,05 | 0,01 |
| 5 | 0 | — | 27 | 8 | 7 | 49 | 18 | 15 | 92 | 37 | 34 |
| 6 | 0 | — | 28 | 8 | 7 | 50 | 18 | 16 | 94 | 38 | 35 |
| 7 | 0 | 0 | 29 | 9 | 7 | 52 | 19 | 17 | 96 | 39 | 36 |
| 8 | 1 | 0 | 30 | 10 | 8 | 54 | 20 | 18 | 98 | 40 | 37 |
| 9 | 1 | 0 | 31 | 10 | 8 | 56 | 21 | 18 | 100 | 41 | 37 |
| 10 | 1 | 0 | 32 | 10 | 8 | 58 | 22 | 19 | 110 | 45 | 42 |
| 11 | 2 | 1 | 33 | 11 | 9 | 60 | 23 | 20 | 120 | 50 | 46 |
| 12 | 2 | 1 | 34 | 11 | 9 | 62 | 24 | 21 | 130 | 55 | 51 |
| 13 | 3 | 1 | 35 | 12 | 10 | 64 | 24 | 22 | 140 | 59 | 55 |
| 14 | 3 | 2 | 36 | 12 | 10 | 66 | 25 | 23 | 150 | 64 | 60 |
| 15 | 3 | 2 | 37 | 13 | 10 | 68 | 26 | 23 | 160 | 69 | 64 |
| 16 | 4 | 2 | 38 | 13 | 11 | 70 | 27 | 24 | 170 | 73 | 69 |
| 17 | 4 | 3 | 39 | 13 | 11 | 72 | 28 | 25 | 180 | 78 | 73 |
| 18 | 5 | 3 | 40 | 14 | 12 | 74 | 29 | 26 | 190 | 83 | 78 |
| 19 | 5 | 4 | 41 | 14 | 12 | 76 | 30 | 27 | 200 | 87 | 83 |
| 20 | 5 | 4 | 42 | 15 | 13 | 78 | 31 | 28 | 220 | 97 | 92 |
| 21 | 6 | 4 | 43 | 15 | 13 | 80 | 32 | 29 | 240 | 106 | 101 |
| 22 | 6 | 5 | 44 | 16 | 13 | 82 | 33 | 30 | 260 | 116 | 110 |
| 23 | 7 | 5 | 45 | 16 | 14 | 84 | 33 | 30 | 280 | 125 | 120 |
| 24 | 7 | 5 | 46 | 16 | 14 | 86 | 34 | 31 | 300 | 135 | 129 |
| 25 | 7 | 6 | 47 | 17 | 15 | 88 | 35 | 32 | | | |
| 26 | 8 | 6 | 48 | 17 | 15 | 90 | 36 | 33 | | | |

# Критерий Уилкоксона для парных выборочных наблюдений (зависимые выборки)

Для того, чтобы проверить нулевую гипотезу, нужно:

1. Вычислить разности значений признака для каждого объекта (d).
2. Вычислить абсолютные разности |d| и расположить их в возрастающем порядке.
3. Вычислить ранги.
4. Выписать ранги положительных и отрицательных значений разностей.
5. Подсчитать суммы рангов отдельно для положительных и отрицательных значений разностей ($T^+$ и $T^-$).

6. За эмпирическое значение критерия $T_{эмп}$ принять наименьшее значение (T+ или T-).
7. Определить табличное значение $T_{крит}$ для $\alpha=0,05$ и n.

Если $T_{эмп} \leq T_{крит}$, нулевая гипотеза отвергается. Различие сравниваемых рядов статистически значимо.

Если $T_{эмп} > T_{крит}$, нулевая гипотеза не отвергается. Различие сравниваемых рядов статистически не значимо.

# Пример:

| Пациент | До лечения | После лечения |
|---|---|---|
| 1 | 1,83 | 0,878 |
| 2 | 0,5 | 0,647 |
| 3 | 1,62 | 0,598 |
| 4 | 2,48 | 2,05 |
| 5 | 1,68 | 1,06 |
| 6 | 1,88 | 1,29 |
| 7 | 1,55 | 1,06 |
| 8 | 3,06 | 3,14 |
| 9 | 1,3 | 1,29 |

| Пациент | До лечения | После лечения | d | \|d\| | | Ранг | Ранг d+ | Ранг d- |
|---|---|---|---|---|---|---|---|---|
| 1 | 1,83 | 0,878 | -0,952 | 0,952 | 0,01 | 1 | | 1 |
| 2 | 0,5 | 0,647 | 0,147 | 0,147 | 0,08 | 2 | 2 | |
| 3 | 1,62 | 0,598 | -1,022 | 1,022 | 0,147 | 3 | 3 | |
| 4 | 2,48 | 2,05 | -0,43 | 0,43 | 0,43 | 4 | | 4 |
| 5 | 1,68 | 1,06 | -0,62 | 0,62 | 0,49 | 5 | | 5 |
| 6 | 1,88 | 1,29 | -0,59 | 0,59 | 0,59 | 6 | | 6 |
| 7 | 1,55 | 1,06 | -0,49 | 0,49 | 0,62 | 7 | | 7 |
| 8 | 3,06 | 3,14 | 0,08 | 0,08 | 0,952 | 8 | | 8 |
| 9 | 1,3 | 1,29 | -0,01 | 0,01 | 1,022 | 9 | | 9 |
| | | | | | | | 5 | 40 |

Общая сумма рангов = 45; $T+ = 5$; $T^- = 40$.

$T = \min(T+, T-) = 5$ $T_{крит}(9,0,05)=5$

Нулевая гипотеза опровергается при $\alpha=0,05$.

Значения параметра у пациентов до и после лечения различаются статистически значимо.

# Сравнение двух независимых выборок. Критерий Манна-Уитни

Эмпирическое значение критерия Манна-Уитни U показывает насколько совпадают (пересекаются) два ряда значений измеренного признака. Нулевой гипотезе соответствует ситуация, когда значения одной выборки будут равномерно распределены среди другой.

1. Значения двух выборок объединяются в один упорядоченный ряд.
2. Значения объединенного ряда ранжируются.
3. Записываются ранги отдельно для первой и второй выборки.
4. Вычисляются суммы рангов для каждой выборки ($R_1$ и $R_2$).
5. Вычисляются $U_1$ и $U_2$ по формулам:

$$U_1 = n_1 n_2 + \frac{1}{2} n_1 (n_1 + 1) - R_1$$

$$U_2 = n_1 n_2 + \frac{1}{2} n_2 (n_2 + 1) - R_2$$

6. Находится минимальное значение критерия U=min ($U_1$, $U_2$)

а) **для малых n:**

Величина U сравнивается с табличным значением $U_{кр}$ (α=0,05, n) распределения Манна-Уитни.

Если U > $U_{кр}$ (α=0,05), нулевая гипотеза не опровергается. Уровни признака статистически значимо не различаются.

Если U <$U_{кр}$(α=0,05), нулевая гипотеза опровергается. Уровни признака различаются статистически значимо.

# Пример:

| № | пол | наблюдение | пол | наблюдение |
|---|---|---|---|---|
| 1 | м | 226,5 | ж | 221,5 |
| 2 | м | 224,1 | ж | 230,2 |
| 3 | м | 218,6 | ж | 223,4 |
| 4 | м | 220,1 | ж | 224,3 |
| 5 | м | 228,8 | ж | 230,8 |
| 6 | м | 229,6 | ж | 223,8 |
| 7 | м | 222,5 | | |

| № | пол | параметр |
|---|---|---|
| 1 | м | 226,5 |
| 2 | м | 224,1 |
| 3 | м | 218,6 |
| 4 | м | 220,1 |
| 5 | м | 228,8 |
| 6 | м | 229,6 |
| 7 | м | 222,5 |
| 8 | ж | 221,5 |
| 9 | ж | 230,2 |
| 10 | ж | 223,4 |
| 11 | ж | 224,3 |
| 12 | ж | 230,8 |
| 13 | ж | 223,8 |

| параметр | пол | ранг | ранг(м) | ранг(ж) |
|---|---|---|---|---|
| 218,6 | м | 1 | 1 | |
| 220,1 | м | 2 | 2 | |
| 221,5 | ж | 3 | | 3 |
| 222,5 | м | 4 | 4 | |
| 223,4 | ж | 5 | | 5 |
| 223,8 | ж | 6 | | 6 |
| 224,1 | м | 7 | 7 | |
| 224,3 | ж | 8 | | 8 |
| 226,5 | м | 9 | 9 | |
| 228,8 | м | 10 | 10 | |
| 229,6 | м | 11 | 11 | |
| 230,2 | ж | 12 | | 12 |
| 230,8 | ж | 13 | | 13 |

$R_1=44$ $R_2=47$

# Критерий Манна-Уитни

$$U_1 = 7 \cdot 6 + \frac{1}{2} \cdot 7 \cdot 8 - 44 = 26$$

$$U_2 = 7 \cdot 6 + \frac{1}{2} 6 \cdot 7 - 47 = 16$$

$$U = \min(U_1, U_2) = 16$$

$n_1 = 7$; $n_2 = 6$; $R_1 = 44$; $R_2 = 47$; $U_1 = 26$; $U_2 = 16$; $U_1 + U_2 = n_1 \cdot n_2$

$U = \min(26,16) = 16$; $U_{кр}(7,6) = 6$; $U > U_{кр}$ $(16 > 6)$.

Нулевая гипотеза не опровергается. Различия параметра у мужчин и женщин статистически не значимо ($\alpha > 0{,}05$).

б) **для больших n:**

применяется критерий z для нормального закона распределения

$$Z = \frac{\left(U - \frac{m \cdot n}{2}\right)}{\sqrt{\frac{m \cdot n(m + n + 1))}{12}}}$$

# Заключение

Нами рассмотрены:

1. Описательная статистика для признаков, не подчиняющихся нормальному закону распределения.
2. Непараметрические критерии достоверности различия двух совокупностей.

# РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА:

Основная литература:


- **Наследов А.Д. Математические методы психологического исследования – СПб.: Речь, 2008. – 392 с.**
- **Герасимов А. Н. Медицинская статистика: учебное пособие / А. Н. Герасимов. – М. : Мед. информ. агентство, 2007. – 480 с.**
- **Балдин К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин. – М. : Флинта, 2010. – 488с.**

БЛАГОДАРЮ ЗА ВНИМАНИЕ