***Тема №8. Асимметрично распределенные данные. Основные характеристики.***

**Медиана и перцентили**

Для порядковых (ранговых) распределений, где критерием середины ряда является медиана, среднеквадратическое отклонение и дисперсия не могут служить характеристиками рассеяния вариант (т.к. их расчет ведется относительно среднего арифметического значения).

То же свойственно и для открытых вариационных рядов. Указанное обстоятельство связано с тем, что отклонения, по которым вычисляются дисперсия и σ, отсчитываются от среднего арифметического, которое не вычисляется в открытых вариационных рядах и в рядах распределений качественных признаков. Поэтому для сжатого описания распределений используется другой параметр разброса – **квантиль** (синоним - «nерцентиль»), пригодный для описания качественных и количественных признаков при любой форме их распределения. Этот параметр может использоваться и для перевода количественных признаков в качественные. В этом случае такие оценки присваиваются в зависимости от того, какому по порядку квантилю соответствует та или иная конкретная варианта.

В практике медико-биологических исследований наиболее часто используются следующие квантили:

$V\_{0,5}$ – медиана;

$V\_{0,25}$, $V\_{0,75}$ – квартили (четверти), где $V\_{0,25}$ – нижний квартиль, $V\_{0,75}$ *–* верхний квартиль.

Квантили делят область возможных изменений вариант в вариационном ряду на определенные интервалы. Медиана (квантиль $V\_{0,5}$) – это варианта, которая находится в середине вариационного ряда и делит этот ряд пополам, на две равные части (*0,5* и *0,5*). Квартиль делит ряд на четыре части: первая часть (нижний квартиль $V\_{0,25}$) – это варианта, отделяющая варианты, числовые значения которых не превышают 25% максимально возможного в данном ряду, квартиль $V\_{0,5}$ отделяет варианты с числовым значением до 50% от максимально возможного. Верхний квартиль ($V\_{0,75}$) отделяет варианты величиной до 75% от максимально возможных значений.

**В случае асимметричности распределения переменной относительно среднего арифметического для его характеристики используются медиана и квартили.** В этом случае используется следующая форма отображения средней величины – *Ме* ($V\_{0,25}$; $V\_{0,75}$). *Например*, исследуемый признак – «срок, в котором ребенок начал самостоятельно ходить» - в исследуемой группе имеет ассиметричное распределение. При этом, нижнему квартилю ($V\_{0,25}$) соответствует срок начала ходьбы – 9,5 месяцев, медиане – 11 месяцев, верхнему квартилю ($V\_{0,75}$) – 12 месяцев. Соответственно, характеристика средней тенденции указанного признака будет представлена, как 11 (9,5; 12) месяцев.

**Оценка статистической значимости результатов исследования**

Под статистической значимостью данных понимают степень их соответствия отображаемой действительности, т.е. статистически значимыми данными считаются те, которые не искажают и правильно отражают объективную реальность.

**Оценить статистическую значимость результатов исследования – означает определить, с какой вероятностью возможно перенести результаты, полученные на выборочной совокупности, на всю генеральную совокупность.** Оценка статистической значимости необходима для понимания того, насколько по части явления можно судить о явлении в целом и его закономерностях.

**Оценка статистической значимости результатов исследования складывается из:**

1. ошибок репрезентативности (ошибок средних и относительных величин) — *m*;

2. доверительных границ средних или относительных величин;

3. достоверности разности средних или относительных величин по критерию *t*.

**Стандартная ошибка средней арифметической** или **ошибка репрезентативности** характеризует колебания средней. При этом необходимо отметить, что чем больше объем выборки, тем меньше разброс средних величин. Стандартная ошибка среднего вычисляется по формуле:

$$m=\frac{σ}{\sqrt{n}}$$

В современной научной литературе средняя арифметическая записывается вместе с ошибкой репрезентативности:

$$M\pm m$$

или вместе со среднеквадратическим отклонением:

$$M\pm σ$$

В качестве примера рассмотрим данные по 1500 городских поликлиник страны (генеральная совокупность). Среднее число пациентов, обслуживающихся в поликлинике равно 18150 человек. Случайный отбор 10 % объектов (150 поликлиник) дает среднее число пациентов, равное 20051 человек. Ошибка выборки, очевидно связанная с тем, что не все 1500 поликлиник попали в выборку, равна разности между этими средними – генеральным средним (*M*ген) и выборочным средним (*М*выб). Если сформировать другую выборку того же объема из нашей генеральной совокупности, она даст другую величину ошибки. Все эти выборочные средние при достаточно больших выборках распределены нормально вокруг генеральной средней при достаточно большом числе повторений выборки одного и того же числа объектов из генеральной совокупности. Стандартная ошибка среднего *m* - это неизбежный разброс выборочных средних вокруг генеральной средней.

В случае, когда результаты исследования представлены относительными величинами (например, процентными долями) – рассчитывается **стандартная ошибка доли:**

$$m=\sqrt{\frac{P(100-P)}{n}}$$

где P – показатель в %, n – количество наблюдений.

Результат отображается в виде *(P ± m)%.* *Например,* процент выздоровления среди больных составил (95,2±2,5)%.

**В том случае, если число элементов совокупности** $n<30$**, то при расчете стандартных ошибок среднего и доли в знаменателе дроби вместо** $n$ **необходимо ставить** $(n-1)$**.**

Для нормального распределения (распределение выборочных средних является нормальным) известно, какая часть совокупности попадает в любой интервал вокруг среднего значения. В частности:

* 68,3% всех выборочных средних попадают в интервал $M\pm m$
* 95,5% **-** в интервал $ M\pm 2∙m$
* 99,7% **-**в интервал $ M\pm 3∙m$

На практике проблема заключается в том, что характеристики генеральной совокупности нам неизвестны, а выборка делается именно с целью их оценки. Это означает, что если мы будем делать выборки одного и того же объема *n* из генеральной совокупности, то в 68,3% случаев на интервале будет находиться значение *M* (оно же в 95,5% случаев будет находиться на интервале и в 99,7% случаев – на интервале).

Поскольку реально делается только одна выборка, то формулируется это утверждение в терминах вероятности: с вероятностью 68,3% среднее значение признака в генеральной совокупности заключено в интервале, с вероятностью 95,5% **-** в интервале и т.д.

На практике вокруг выборочного значения строится такой интервал, который бы с заданной (достаточно высокой) вероятностью – **доверительной вероятностью –** «накрывал» бы истинное значение этого параметра в генеральной совокупности. Этот интервал называется **доверительным интервалом**.

**Доверительная вероятность *P* –** это степень уверенности в том, что доверительный интервал действительно будет содержать истинное (неизвестное) значение параметра в генеральной совокупности.

Например, если доверительная вероятность *Р* равна 90%, то это означает, что 90 выборок из 100 дадут правильную оценку параметра в генеральной совокупности. Соответственно, вероятность ошибки, т.е. неверной оценки генерального среднего по выборке, равна в процентах: $100\%-P$. Для данного примера это значит, что 10 выборок из 100 дадут неверную оценку.

Очевидно, что степень уверенности (доверительная вероятность) зависит от величины интервала: **чем шире интервал, тем выше уверенность, что в него попадет неизвестное значение для генеральной совокупности**. На практике для построения доверительного интервала берется, как минимум, удвоенная ошибка выборки, чтобы обеспечить уверенность не менее 95,5%.

Определение доверительных границ средних и относительных величин позволяет найти два их крайних значения – минимально возможное и максимально возможное, в пределах которых изучаемый показатель может встречаться во всей генеральной совокупности. Исходя из этого, **доверительные границы (или доверительный интервал)** - **это границы средних или относительных величин, выход за пределы которых вследствие случайных колебаний имеет незначительную вероятность.**

Доверительный интервал может быть переписан в виде: $M\pm t∙m$, где *t* – доверительный критерий.

Доверительные границы средней арифметической величины в генеральной совокупности определяют по формуле:

*Мген = Мвыб+ t mM*

для относительной величины:

*Рген = Рвыб+t mР*

где *Мген* и *Рген* - значения средней и относительной величины для генеральной совокупности; *Мвыб* и *Рвыб* - значения средней и относительной величины, полученные на выборочной совокупности; *mM* и *mP* - ошибки средней и относительной величин; *t* - доверительный критерий (критерий точности, который устанавливается при планировании исследования и может быть равен 2 или 3); *t m* - это доверительный интервал или Δ – предельная ошибка показателя, полученного при выборочном исследовании.

Следует отметить, что величина критерия *t* в определенной мере связана с вероятностью безошибочного прогноза (р), выраженной в %. Ее избирает сам исследователь, руководствуясь необходимостью получить результат с нужной степенью точности. Так, для вероятности безошибочного прогноза 95,5% величина критерия *t* составляет 2, для 99,7% - 3.

Приведенные оценки доверительного интервала приемлемы лишь для статистических совокупностей с количеством наблюдений более 30. При меньшем объеме совокупности (малых выборках) для определения критерия t пользуются специальными таблицами. В данных таблицах искомое значение находится на пересечении строки, соответствующей численности совокупности *(n-1)*, и столбца, соответствующего уровню вероятности безошибочного прогноза (95,5%; 99,7%), выбранному исследователем. В медицинских исследованиях при установлении доверительных границ любого показателя принята вероятность безошибочного прогноза 95,5% и более. Это означает, что величина показателя, полученная на выборочной совокупности должна встречаться в генеральной совокупности как минимум в 95,5% случаев.