Контрольные вопросы к практической работе 7

(Ответы должны быть в письменном виде)

1. Что такое вариационный ряд?
2. Что такое ранжированный вариационный ряд?
3. Что такое средняя арифметическая величина?
4. Как в программе Excel вычисляется средняя величина?
5. Что такое мода вариационного ряда?
6. Что такое медиана вариационного ряда?
7. Где используется вычисление средних величин в медицине?

ТЕОРИЯ
Вариационные ряды. Средние величины.

В прикладной статистике объектом исследования являются полученные в результате наблюдений статистические данные. Статистические данные – это совокупность объектов (наблюдений, случаев) и признаков (переменных), их характеризующих.

Переменная (признак) – некоторое общее для всех изучаемых объектов, например людей, свойство, конкретные проявления которого могут меняться от объекта к объекту. Это может быть или физиологический признак (рост, вес, пол, острота зрения и т.д.), или психическое свойство (темперамент, характер, степень интеллектуальности и др.), которые характеризуют состояние рассматриваемого объекта. Переменные могут быть измерены в различных шкалах.

Различные проявления признака для разных объектов называют значениями. С точки зрения теории вероятностей и математической статистики изучаемая переменная также может трактоваться как случайная величина.

На сегодняшний день общепринятого согласия в вопросе «правильной» классификации типов переменных не достигнуто, поэтому для простоты мы примем за основу три типа переменных: *непрерывные*, *дискретные* и *категориальные* (*номинальные*):

Непрерывные переменные (continuous variables) могут принимать любые численные значения, которые естественным образом упорядочены на числовой оси (рост, вес, АД, острота зрения, РОЭ).

Дискретные переменные (discrete variables) могут принимать счётное множество упорядоченных значений, которые могут просто обозначать целочисленные данные или ранжировать данные по степени проявления на упорядоченной ранговой шкале (клиническая стадия опухоли, тяжесть состояния пациента).

Категориальные переменные (categorial variables) являются неупорядоченными и используются для качественной классификации (пол, цвет глаз, место жительства, тип темперамента); в частности, они могут быть бинарными (дихотомическими) и иметь категорические значения: 1/0, да/нет, имеется/отсутствует.

Тип переменных определяет набор статистических методов анализа данных.

Полученный в результате статистических наблюдений первичный материал необходимо сгруппировать для дальнейшего использования. При этом группировке подлежат только качественно однородные элементы.

Вариационный ряд – это однородная в качественном отношении статистическая совокупность, отдельные единицы которой характеризуют количественные различия изучаемого признака или явления.

Количественные (числовые) данные предполагают, что переменная принимает некоторое числовое значение. Из них выделяют дискретные данные и непрерывные.

Численное значение каждого отдельного признака или явления, входящего в вариационный ряд, называется вариантой и обозначается символом ***V*** или ***x,*** или ***y.***

Вариационный ряд, в котором каждая варианта встречается только один раз, называют ***простым***.

При увеличении числа наблюдений, как правило, встречаются повторяющиеся значения вариант. В этом случае создается сгруппированный

вариационный ряд, где указывается число повторений (частота, обозначается

буквой «***p***»).

Ранжированный вариационный ряд состоит из вариант, расположенных в порядке возрастания или убывания.

Интервальный вариационный ряд составляют при большом числе единиц наблюдения (более нескольких десятков) с целью упрощения последующих вычислений, выполняемых, как правило, без использования компьютера.

Также вариационные ряды делятся на ***дискретные*** и ***непрерывные***.

Непрерывный вариационный ряд включает значения вариант, которые могут выражаться любыми значениями.

Если в вариационном ряде значения признака (варианты) заданы в виде отдельных конкретных чисел, то такой ряд называют ***дискретным***.

Особое место в статистическом анализе принадлежит определению среднего уровня изучаемого признака или явления. Средний уровень признака измеряют средними величинами.

Средняя величина характеризует общий количественный уровень изучаемого признака и является групповым свойством статистической совокупности. Она нивелирует, ослабляет случайные отклонения индивидуальных наблюдений в ту или иную сторону и выдвигает на первый план основное, типичное свойство изучаемого признака.

В медицине средние величины широко используются:

- ***Для оценки состояния здоровья населения: характеристики физического развития*** (рост, вес, окружность грудной клетки и пр.), ***выявления распространенности и длительности различных заболеваний, анализа демографических показателей*** (естественного движения населения, средней продолжительности предстоящей жизни, воспроизводства населения, средней численности населения и др.).
- ***Для изучения деятельности ЛПУ, медицинских кадров и оценки качества их работы, планирования и определения потребности населения в различных видах медицинской помощи*** (среднее число обращений или посещений на одного жителя в год, средняя длительность пребывания больного в стационаре, средняя продолжительность обследования больного, средняя обеспеченность врачами, койками и пр.).
- ***Для характеристики санитарно-эпидемиологического состояния*** (средняя запыленность воздуха в цехе, средняя площадь на одного человека, средние нормы потребления белков, жиров и углеводов и т. д.).
- ***Для определения медико-физиологических показателей в норме и патологии, при обработке лабораторных данных, для установления достоверности результатов выборочного исследования в социально-гигиенических, клинических, экспериментальных исследованиях.***

Вычисление средних величин выполняется на основе вариационных рядов. Общими характеристиками значений признака, отражаемого в вариационном ряду, являются средние величины. К ним относятся:

Мода (Mo) - значение наиболее часто встречающейся варианты.

Медиана (Me) - значение варианты, делящей ранжированный вариационный ряд пополам (с каждой стороны медианы находится половина вариант).

Средняя арифметическая величина (M или $\bar{x}$) – это общая количественная характеристика определенного признака изучаемых явлений, составляющих качественно однородную статистическую совокупность. Различают *среднюю арифметическую простую* и *взвешенную*.

Простая средняя арифметическая вычисляется по формуле:

$$M = \frac{\sum V}{n},$$

где **Σ*V***- сумма вариант;
n - число наблюдений.

В сгруппированном вариационном ряду определяют ***взвешенную среднюю арифметическую***:

$$M = \frac{\sum Vp}{n},$$

Σ*Vp* - сумма произведений вариант на их частоты;
n - число наблюдений.

В редких случаях, когда имеется симметричный вариационный ряд, значения моды и медианы совпадают со значением средней арифметической.

Средние арифметические величины могут не в полной мере отражать свойства вариационного ряда, в особенности, когда необходимо сопоставление с другими средними. Близкие по значению средние могут быть получены из рядов с различной степенью рассеяния. Чем ближе друг к другу отдельные варианты по своей количественной характеристике, тем меньше ***рассеяние (вариабельность) ряда***, тем типичнее его средняя.

Основными параметрами, позволяющими оценить вариабельность признака:

- ✓ Размах;
- ✓ Амплитуда;
- ✓ Среднее квадратическое отклонение;
- ✓ Коэффициент вариации.

Приблизительно о вариабельности признака можно судить по размаху и амплитуде вариационного ряда. ***Размах*** указывает на максимальную ***(V_{max})*** и минимальную ***(V_{min})*** варианты в ряду. ***Амплитуда (A_m)*** является разностью этих вариант:

$$A_m = V_{\max} - V_{\min}$$

Еще одной важнейшей характеристикой вариабельности ряда является дисперсия (***D***). Она учитывает величину отклонения (***d***) каждой варианты вариационного ряда от его средней арифметической, т.е. (***d=V-M***). Поскольку отклонения вариант от средней могут быть положительными и отрицательными, то их сумма может оказаться равной нулю ($\sum d=0$). Чтобы избежать этого, величины отклонения (d) возводятся во вторую степень и усредняются.

Но наиболее часто применяется более удобный параметр, вычисляемый на основе дисперсии - среднее квадратическое отклонение ***(σ)***.

Поскольку дисперсия выражается квадратом отклонений, ее величина не может использоваться в сопоставлении со средней арифметической. Для этих целей применяется ***среднее квадратическое отклонение*** (***σ***).

$$\sigma = \sqrt{\frac{\sum d^2}{n}}. \qquad (1)$$

Оно характеризует среднее отклонение всех вариант вариационного ряда от средней арифметической величины в тех же единицах, что и сама средняя величина, поэтому они могут использоваться совместно.

Формула (1) применяется при числе наблюдений (***n***) больше 30. В противном случае значение среднего квадратического отклонения будет иметь погрешность,

связанную с математическим смещением ***(n-1)***. В связи с этим, более точный результат может быть получен с помощью учета такого смещения в формуле расчета *стандартного отклонения*:

$s = \sqrt{\sum d^2}$.

Стандартное отклонение (s) это оценка среднеквадратического отклонения случайной величины *X* относительно её математического ожидания на основе несмещённой оценки её дисперсии.

При значениях ***n>30*** среднее квадратическое отклонение ***(σ)*** и стандартное отклонение ***(s)*** будут одинаковыми (***σ=s***). Поэтому в большинстве практических пособий эти критерии рассматриваются как разнозначные. В программе Excel вычисление стандартного отклонения может быть выполнено функцией =СТАНДОТКЛОН(диапазон). А с целью расчета среднего квадратического отклонения требуется создать соответствующую формулу.

Величина среднего квадратического отклонения позволяет судить о характере однородности вариационного ряда и исследуемой группы:

- Если величина ***σ*** небольшая, то это свидетельствует о достаточно высокой однородности изучаемого явления. В этом случае среднюю арифметическую следует признать вполне характерной для данного вариационного ряда.
- Слишком малая величина ***σ*** заставляет думать об искусственном подборе наблюдений.
- При очень большой величине ***σ*** средняя арифметическая в меньшей степени характеризует вариационный ряд, что говорит о значительной вариабельности изучаемого признака или явления или о неоднородности исследуемой группы.

Заметим, что сопоставление величины среднего квадратического отклонения возможно только для признаков одинаковой размерности. Действительно, при сравнении разнообразия веса новорожденных детей и взрослых, мы всегда получим более высокие значения ***σ*** у взрослых.

Значения коэффициента вариации менее 10% свидетельствует о малом рассеянии, от 10 до 20% – о среднем, более 20% – о сильном рассеянии вариант вокруг средней арифметической.

Для формулировки общего вывода об изучаемом явлении, результаты, полученные на основе выборочной совокупности, должны быть, перенесены на генеральную совокупность статистическими методами. При этом следует помнить, что найденные на основе данных выборочной совокупности величины, например, среднее арифметическое, как правило, при повторных исследованиях под влиянием случайных явлений может меняться.

Чтобы определить степень совпадения выборочного исследования и генеральной совокупности, необходимо оценить величину ошибки, которая неизбежно возникает при выборочном наблюдении – ***ошибки репрезентативности (m)*** (или ***средней ошибки средней арифметической***). Она является разностью между средними, полученными при выборочном статистическом наблюдении, и аналогичными величинами, которые были бы получены при изучении генеральной совокупности.

Ошибку репрезентативности нельзя смешивать с ошибками регистрации или ошибками внимания (описки, просчеты, опечатки и др.). Величина ошибки репрезентативности зависит от объема выборки и от вариабельности признака. Чем больше число наблюдений, тем ближе выборка к генеральной совокупности и тем меньше ошибка. Чем более изменчив признак, тем больше величина статистической

ошибки.

Прогнозирование величины средней арифметической в генеральной совокупности выполняется с указанием ***доверительных границ*** – минимального и максимального значений. Поскольку выборочная средняя является случайной величиной, такой прогноз выполняется с приемлемым для исследователя уровнем вероятности. В медицинских исследованиях он составляет не менее 95%.

Согласно теории вероятности, в явлениях, подчиняющихся нормальному закону распределения (рис. 1), между значениями средней арифметической, среднего квадратического отклонения и вариантами существует строгая зависимость - выполняется ***правило трех сигм***: 99,7% значений варьирующего признака находятся в пределах M ± 3σ.

Кроме того 68,3% варьирующего признака находятся в пределах M ± 1σ, 95,5% — в пределах M ± 2σ.

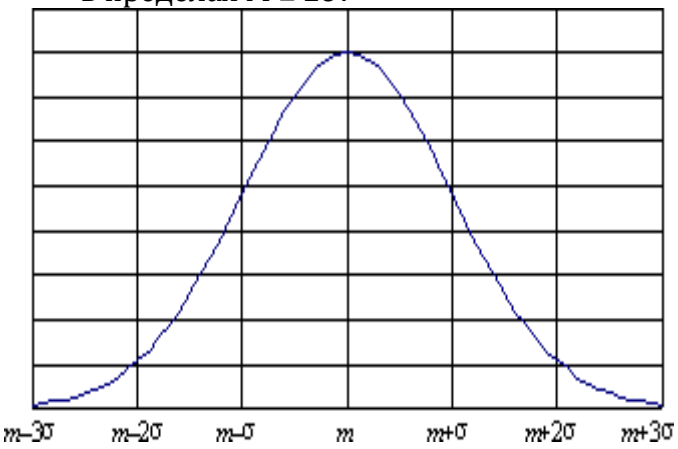

Рис. 1. Плотность вероятностей нормального распределения.

Большинство экспериментальных исследований, в том числе и в области медицины, связано с измерениями, результаты которых могут принимать практически любые значения в заданном интервале, поэтому, как правило, описываются моделью непрерывных случайных величин. В связи с этим в большинстве статистических методов рассматриваются непрерывные распределения. Одним из таких распределений, имеющим основополагающую роль в математической статистике, является ***нормальное*** или ***Гауссово распределение***.

Это объясняется целым рядом причин.

1. Прежде всего, многие экспериментальные наблюдения можно успешно описать с помощью нормального распределения (измерения веса, роста и других физиологических параметров организма человека).
2. Многие распределения, связанные со случайной выборкой, при увеличении объема последней переходят в нормальное.
3. Нормальное распределение хорошо подходит в качестве приближенного описания других непрерывных распределений (например, асимметричных).

В то же время следует отметить, что в медицинских данных встречается много экспериментальных распределений, описание которых моделью нормального распределения невозможно. Для этого в статистке разработаны методы, которые принято называть «Непараметрическими».

Выбор статистического метода, который подходит для обработки данных конкретного эксперимента, должен производиться в зависимости от принадлежности полученных данных к нормальному закону распределения. Проверка гипотезы на подчинение признака нормальному закону распределения выполняется с помощью гистограммы распределения частот (графика), а также ряда

статистических критериев. Среди них:

- Критерий асимметрии (β);
- Критерий проверки на эксцесс (γ);
- Критерий Шапиро – Уилкса (W) .

Анализ характера распределения данных (его еще называют проверкой на нормальность распределения) осуществляется по каждому параметру. Чтобы уверенно судить о соответствии распределения параметра нормальному закону, необходимо достаточно большое число единиц наблюдения (не менее 30 значений).

Для нормального распределения критерии асимметрии и эксцесса принимают значение 0. Если распределение смещено вправо $\beta > 0$ (положительная асимметрия), при $\beta < 0$ - график распределения смещен влево (отрицательная асимметрия). Критерий асимметрии проверяет форму кривойраспределения. В случае нормального закона γ=0. При $\gamma > 0$ кривая распределения острее, если $\gamma < 0$ пик более сглаженный, чем функция нормального распределения.