ОСНОВЫ НАУЧНЫХ ИССЛЕДОВАНИЙ

Задание для ГП-17

24.11.2020

Задание на 24 ноября 2020

1) Ознакомиться с практическим занятием №7 **«Статистические методы в задачах обработки экспериментальных данных»**.

Сделать письменный конспект практического занятия.

Практическое занятие №7
Статистические методы в задачах обработки экспериментальных данных

Методами статистической обработки результатов эксперимента называются математические приемы, формулы, способы количественных расчетов, с помощью которых показатели, получаемые в ходе эксперимента, можно обобщать, приводить в систему, выявляя скрытые в них закономерности. Речь идет о таких закономерностях статистического характера, которые существуют между изучаемыми в эксперименте переменными величинами.

Некоторые из методов математико-статистического анализа позволяют вычислять так называемые элементарные математические статистики, характеризующие выборочное распределение данных, например выборочное среднее, выборочная дисперсия, мода, медиана и ряд других. Иные методы математической статистики, например дисперсионный анализ, регрессионный анализ, позволяют судить о динамике изменения отдельных статистик выборки. С помощью третьей группы методов, скажем, корреляционного анализа, факторного анализа, методов сравнения выборочных данных, можно достоверно судить о статистических связях, существующих между переменными величинами, которые исследуют в данном эксперименте.

1. Методы первичной статистической обработки результатов эксперимента

Все методы математико-статистического анализа условно делятся на первичные и вторичные. Первичными называют методы, с помощью которых можно получить показатели, непосредственно отражающие результаты производимых в эксперименте измерений. Вторичными называются методы статистической обработки, с помощью которых на базе первичных данных выявляют скрытые в них статистические закономерности.

К первичным методам статистической обработки относят, например, определение выборочной средней величины, выборочной дисперсии, выборочной моды и выборочной медианы. В число вторичных методов обычно

включают корреляционный анализ, регрессионный анализ, методы сравнения первичных статистик у двух или нескольких выборок.

Рассмотрим методы вычисления элементарных математических статистик.

1.1 Мода

Числовой характеристикой выборки, как правило, не требующей вычислений, является так называемая мода. Модой называют количественное значение исследуемого признака, наиболее часто встречающееся в выборке. Для симметричных распределений признаков, в том числе для нормального распределения, значение моды совпадает со значениями среднего и медианы. Для других типов распределений, несимметричных, это не характерно. К примеру, в последовательности значений признаков 1, 2, 5, 2, 4, 2, 6, 7, 2 модой является значение 2, так как оно встречается чаще других значений - четыре раза.

Моду находят согласно следующим правилам:

1) В том случае, когда все значения в выборке встречаются одинаково часто, принято считать, что этот выборочный ряд не имеет моды. Например: 5, 5, 6, 6, 7, 7 - в этой выборке моды нет.

2) Когда два соседних (смежных) значения имеют одинаковую частоту и их частота больше частот любых других значений, мода вычисляется как среднее арифметическое этих двух значений. Например, в выборке 1, 2, 2, 2, 5, 5, 5, 6 частоты рядом расположенных значений 2 и 5 совпадают и равняются 3. Эта частота больше, чем частота других значений 1 и 6 (у которых она равна 1). Следовательно, модой этого ряда будет величина =3,5

3) Если два несмежных (не соседних) значения в выборке имеют равные частоты, которые больше частот любого другого значения, то выделяют две моды. Например, в ряду 10, 11, 11, 11, 12, 13, 14, 14, 14, 17 модами являются значения 11 и 14. В таком случае говорят, что выборка является бимодальной.

Могут существовать и так называемые мультимодальные распределения, имеющие более двух вершин (мод).

4) Если мода оценивается по множеству сгруппированных данных, то для нахождения моды необходимо определить группу с наибольшей частотой признака. Эта группа называется модальной группой.

1.2 Медиана

Медианой называется значение изучаемого признака, которое делит выборку, упорядоченную по величине данного признака, пополам. Справа и слева от медианы в упорядоченном ряду остается по одинаковому количеству признаков. Например, для выборки 2, 3, 4, 4, 5, 6, 8, 7, 9 медианой будет значение 5, так как слева и справа от него остается по четыре показателя. Если ряд включает в себя четное число признаков, то медианой будет среднее, взятое как полусумма величин двух центральных значений ряда. Для следующего ряда 0, 1, 1, 2, 3, 4, 5, 5, 6, 7 медиана будет равна 3,5.

Знание медианы полезно для того, чтобы установить, является ли распределение частных значений изученного признака симметричным и при-

ближающимся к так называемому нормальному распределению. Средняя и медиана для нормального распределения обычно совпадают или очень мало отличаются друг от друга. Если выборочное распределение признаков нормально, то к нему можно применять методы вторичных статистических расчетов, основанные на нормальном распределении данных. В противном случае этого делать нельзя, так как в расчеты могут вкрасться серьезные ошибки.

1.3 Выборочное среднее

Выборочное среднее (среднее арифметическое) значение как статистический показатель представляет собой среднюю оценку изучаемого в эксперименте качества. Выборочное среднее определяется при помощи следующей формулы:

$$\overline{X} = \frac{(X_1 + X_2 + ... + X_n)}{n} = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} X_i \right)$$

где x - выборочная средняя величина или среднее арифметическое значение по выборке; n - количество показателей в выборке, на основе которых вычисляется средняя величина; xk - частные значения показателей. Всего таких показателей n, поэтому индекс k данной переменной принимает значения от 1 до n; ∑ - принятый в математике знак суммирования величин тех переменных, которые находятся справа от этого знака. Выражение соответственно означает сумму всех x с индексом k, от 1до n.

1.4 Разброс выборки

Разброс (иногда эту величину называют размахом) выборки обозначается буквой R. Это самый простой показатель, который можно получить для выборки - разность между максимальной и минимальной величинами данного конкретного вариационного ряда, т.е.

R= xmax - xmin

Понятно, что чем сильнее варьирует измеряемый признак, тем больше величина R, и наоборот. Однако может случиться так, что у двух выборочных рядов и средние, и размах совпадают, однако характер варьирования этих рядов будет различный. Например, даны две выборки:

X = 10 15 20 25 30 35 40 45 50 X = 30 R = 40
Y = 10 28 28 30 30 30 32 32 50 Y=30 R = 40

При равенстве средних и разбросов для этих двух выборочных рядов характер их варьирования различен. Для того чтобы более четко представлять характер варьирования выборок, следует обратиться к их распределениям.

1.5 Дисперсия

Дисперсия - это среднее арифметическое квадратов отклонений значений переменной от её среднего значения.

Дисперсия как статистическая величина характеризует, насколько частные значения отклоняются от средней величины в данной выборке. Чем больше дисперсия, тем больше отклонения или разброс данных.

$$D = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2 \qquad (1)$$

В статистическом понимании Дисперсия

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

есть среднее арифметическое из квадратов отклонений величин x_i от их среднего арифметического

$$\overline{x} = (x_1 + x_2 + \ldots + x_n) / n.$$

где 1 - выборочная дисперсия, или просто дисперсия;

(X_i-X) - выражение, означающее, что для всех x, от первого до последнего в данной выборке необходимо вычислить разности между частными и средними значениями, возвести эти разности в квадрат и просуммировать;

n - количество значений, по которым вычисляется дисперсия. Однако сама дисперсия, как характеристика отклонения от среднего, часто неудобна для интерпретации. Для того, чтобы приблизить размерность дисперсии к размерности измеряемого признака применяют операцию извлечения квадратного корня из дисперсии. Полученную величину называют **среднеквадратическим отклонением**.

Из суммы квадратов, делённых на число членов ряда извлекается квадратный корень.

$$S_x = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Среднеквадрати́ческое отклоне́ние (синонимы: среднее квадрати́ческое отклоне́ние, среднеквадрати́чное отклоне́ние, квадрати́чное отклоне́ние; близкие термины: станда́ртное отклоне́ние, станда́ртный разбро́с) — в теории вероятностей и статистике наиболее распространённый показатель рассеивания значений случайной величины относительно её математического ожидания (среднего значения случ. величины).

Среднеквадратическое отклонение измеряется в единицах измерения самой случайной величины и используется при расчёте стандартной ошибки среднего арифметического, при построении доверительных интервалов, при статистической проверке гипотез, при измерении линейной взаимосвязи между случайными величинами. Определяется как квадратный корень из дисперсии случайной величины.

Большое значение среднеквадратического отклонения показывает большой разброс значений в представленном множестве со средней величиной множества; маленькое значение, соответственно, показывает, что значения в множестве сгруппированы вокруг среднего значения.

Например, у нас есть три числовых множества: {0, 0, 14, 14}, {0, 6, 8, 14} и {6, 6, 8, 8}. У всех трёх множеств средние значения равны 7, а среднеквадратические отклонения, соответственно, равны 7, 5 и 1. У последнего множества среднеквадратическое отклонение маленькое, так как значения в множестве сгруппированы вокруг среднего значения; у первого множества самое большое значение среднеквадратического отклонения — значения внутри множества сильно расходятся со средним значением.

В общем смысле среднеквадратическое отклонение можно считать мерой неопределённости. К примеру, в физике среднеквадратическое отклонение используется для определения погрешности серии последовательных измерений какой-либо величины. Это значение очень важно для определения правдоподобности изучаемого явления в сравнении с предсказанным теорией значением: если среднее значение измерений сильно отличается от предсказанных теорией значений (большое значение среднеквадратического отклонения), то полученные значения или метод их получения следует перепроверить.

На практике среднеквадратическое отклонение позволяет оценить, насколько значения во множестве могут отличаться от среднего значения.

2. Методы вторичной статистической обработки результатов эксперимента

С помощью вторичных методов статистической обработки экспериментальных данных непосредственно проверяются, доказываются или опровергаются гипотезы, связанные с экспериментом. Эти методы, как правило, сложнее, чем методы первичной статистической обработки, и требуют от исследователя хорошей подготовки в области элементарной математики и статистики.

Обсуждаемую группу методов можно разделить на несколько подгрупп:

1. Регрессионное исчисление.

2. Методы сравнения между собой двух или нескольких элементарных статистик (средних, дисперсий и т.п.), относящихся к разным выборкам.

3. Методы установления статистических взаимосвязей между переменными, например их корреляции друг с другом.

4. Методы выявления внутренней статистической структуры эмпирических данных (например, факторный анализ). Рассмотрим каждую из выделенных подгрупп методов вторичной статистической обработки на примерах.

2.1 Регрессионное исчисление

Регрессионное исчисление - это метод математической статистики, позволяющий свести частные, разрозненные данные к некоторому линейному графику, приблизительно отражающему их внутреннюю взаимосвязь, и получить возможность по значению одной из переменных приблизительно оценивать вероятное значение другой переменной.

Графическое выражение регрессионного уравнения называют линией регрессии. Линия регрессии выражает наилучшие предсказания зависимой переменой (Y) по независимым переменным (X).

Регрессию выражают с помощью двух уравнений регрессии, которые в самом прямом случае выглядят, как уравнения прямой.

Y = a 0 + a 1 * X (1)
X = b 0 + b 1 * Y (2)

В уравнении (1) Y - зависимая переменная, X - независимая переменная, a 0 - свободный член, a 1 - коэффициент регрессии, или угловой коэффициент, определяющий наклон линии регрессии по отношению к осям координат.

В уравнении (2) X - зависимая переменная, Y - независимая переменная, b 0 - свободный член, b 1 - коэффициент регрессии, или угловой коэффициент, определяющий наклон линии регрессии по отношению к осям координат.

Количественное представление связи (зависимости) между X и Y (между Y и X) называется **регрессионным анализом**. Регрессио́нный анализ — статистический метод исследования влияния одной или нескольких независимых переменных $X_1, X_2, ..., X_p$ на зависимую переменную Y. Независимые переменные иначе называют регрессорами или предикторами, а зависимые переменные критериальными. Терминология *зависимых* и *независимых* переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

Главная задача регрессионного анализа заключается в нахождении коэффициентов a 0, b 0, a1и b 1 и определении уровня значимости полученных аналитических выражений, связывающих между собой переменные X и У.

При этом коэффициенты регрессии a 1 и b 1 показывают, насколько в среднем величина одной переменной изменяется при изменении на единицу меры другой. Коэффициент регрессии a 1 в уравнении можно подсчитать по формуле:

$$a1 = r_{xy} \cdot \frac{S_y}{S_x}$$

а коэффициент b 1 в уравнении по формуле

$$b1 = r_{yx} \cdot \frac{S_x}{S_y}$$

где ryx - коэффициент корреляции между переменными X и Y;

Sx - среднеквадратическое отклонение, подсчитанное для переменной X;

Sy - среднеквадратическое отклонение, подсчитанное для переменной У/

Для применения метода линейного регрессионного анализа необходимо соблюдать следующие условия:

1. Сравниваемые переменные X и Y должны быть измерены в шкале интервалов или отношений.

2. Предполагается, что переменные X и Y имеют нормальный закон распределения.

3. Число варьирующих признаков в сравниваемых переменных должно быть одинаковым.

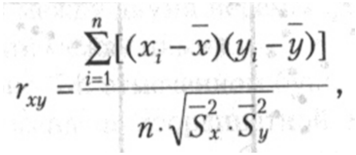
2.2 Корреляция

Следующий метод вторичной статистической обработки, посредством которого выясняется связь или прямая зависимость между двумя рядами экспериментальных данных, носит название метод **корреляций**. Он показывает, каким образом одно явление влияет на другое или связано с ним в своей динамике. Подобного рода зависимости существуют, к примеру, между величинами, находящимися в причинно-следственных связях друг с другом. Если выясняется, что два явления статистически достоверно коррелируют друг с другом и если при этом есть уверенность в том, что одно из них может выступать в качестве причины другого явления, то отсюда определенно следует вывод о наличии между ними причинно-следственной зависимости.

Когда повышение уровня одной переменной сопровождается повышением уровня другой, то речь идёт о положительной корреляции. Если же рост одной переменной происходит при снижении уровня другой, то говорят об отрицательной корреляции. При отсутствии связи переменных мы имеем дело с нулевой корреляцией.

Имеется несколько разновидностей данного метода: линейный, ранговый, парный и множественный. Линейный корреляционный анализ позволяет устанавливать прямые связи между переменными величинами по их абсолютным значениям. Эти связи графически выражаются прямой линией, отсюда название "линейный". Ранговая корреляция определяет зависимость не между абсолютными значениями переменных, а между порядковыми местами, или рангами, занимаемыми ими в упорядоченном по величине ряду. Парный корреляционный анализ включает изучение корреляционных зависимостей только между парами переменных, а множественный, или многомерный, - между многими переменными одновременно. Распространенной в прикладной статистике формой многомерного корреляционного анализа является факторный анализ.

Коэффициент линейной корреляции определяется при помощи следующей формулы:



$$r_{xy} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{n \cdot \sqrt{\bar{S}_x^2 \cdot \bar{S}_y^2}},$$

где rxy - коэффициент линейной корреляции;

x, y - средние выборочные значения сравниваемых величин;

xi, yi - частные выборочные значения сравниваемых величин;

n - общее число величин в сравниваемых рядах показателей;

S2x, S2y - дисперсии, отклонения сравниваемых величин от средних значений.

Метод множественных корреляций в отличие от метода парных корреляций позволяет выявить общую структуру корреляционных зависимостей, существующих внутри многомерного экспериментального материала, включающего более двух переменных, и представить эти корреляционные зависимости в виде некоторой системы.

Для применения частного коэффициента корреляции необходимо соблюдать следующие условия:

1. Сравниваемые переменные должны быть измерены в шкале интервалов или отношений.

2. Предполагается, что все переменные имеют нормальный закон распределения.

3. Число варьирующих признаков в сравниваемых переменных должно быть одинаковым.

4. Для оценки уровня достоверности корреляционного отношения Пирсона следует пользоваться формулой (11.9) и таблицей критических значений для t-критерия Стьюдента при k = n - 2.

2.3 Факторный анализ

Факторный анализ - статистический метод, который используется при обработке больших массивов экспериментальных данных. Задачами факторного анализа являются: сокращение числа переменных (редукция данных) и определение структуры взаимосвязей между переменными, т.е. классификация переменных, поэтому факторный анализ используется как метод сокращения данных или как метод структурной классификации.

Важное отличие факторного анализа от всех описанных выше методов заключается в том, что его нельзя применять для обработки первичных, или, как говорят, "сырых", экспериментальных данных, т.е. полученных непосредственно при обследовании испытуемых. Материалом для факторного анализа служат корреляционные связи, а точнее - коэффициенты корреляции Пирсона, которые вычисляются между переменными, включенными в обследование. Иными словами, факторному анализу подвергают корреляционные матрицы, или, как их иначе называют, матрицы интеркорреляций. Наименования столбцов и строк в этих матрицах одинаковы, так как они представляют собой перечень переменных, включенных в анализ. По этой причине матрицы интеркорреляций всегда квадратные, т.е. число строк в них равно числу столбцов, и симметричные, т.е. на симметричных местах относительно главной диагонали стоят одни и те же коэффициенты корреляции.

Главное понятие факторного анализа - фактор. Это искусственный статистический показатель, возникающий в результате специальных преобразований таблицы коэффициентов корреляции между изучаемыми психологическими признаками, или матрицы интеркорреляций. Процедура извлечения факторов из матрицы интеркорреляций называется факторизацией матрицы. В результате факторизации из корреляционной матрицы может быть извлечено разное количество факторов вплоть до числа, равного количеству исходных переменных. Однако факторы, выделяемые в результате факторизации, как правило, неравноценны по своему значению.

С помощью выявленных факторов объясняют взаимозависимость явлений.

Чаще всего в итоге факторного анализа определяется не один, а несколько факторов, по-разному объясняющих матрицу интеркорреляций переменных. В таком случае факторы делят на генеральные, общие и единичные. Генеральными называются факторы, все факторные нагрузки которых значительно отличаются от нуля (нуль нагрузки свидетельствует о том, что данная переменная никак не связана с остальными и не оказывает на них никакого влияния в жизни). Общие - это факторы, у которых часть факторных нагрузок отлична от нуля. Единичные - это факторы, в которых существенно отличается от нуля только одна из нагрузок.