СТАТИСТИЧЕСКИЕ СПОСОБЫ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Методы статистической обработки результатов эксперимента:

математические приемы, формулы, способы количественных расчетов, с помощью которых показатели, получаемые в ходе эксперимента, можно обобщать, приводить в систему, выявляя скрытые в них закономерности.

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.

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

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 распределении, 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.

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 приближающимся 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.

Выборочное среднее

Выборочное среднее (среднее арифметическое) значение как статистический показатель представляет собой среднюю оценку изучаемого в эксперименте психологического качества. Эта оценка характеризует степень его развития в целом у той группы испытуемых, которая была подвергнута психодиагностическому обследованию. Сравнивая непосредственно средние значения двух или нескольких выборок, мы можем судить об относительной степени развития у людей, составляющих эти выборки, оцениваемого качества.

Выборочное среднее определяется при помощи следующей формулы:

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

где x - выборочная средняя величина или среднее арифметическое значение по выборке; n - количество испытуемых в выборке или частных психодиагностических показателей, на основе которых вычисляется средняя величина; x_k - частные значения показателей у отдельных испытуемых. 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

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.
- $R = x_{max} - x_{min}$
- 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 50X = 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

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.

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

где 5 - выборочная дисперсия, или просто дисперсия;

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.

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 (7).
- 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.

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.

Краткий обзор современных программных средств для проведения анализа данных.

MATLAB – это высокопроизводительный язык для технических расчетов. Он включает в себя вычисления, визуализацию и программирование в удобной среде, где задачи и решения выражаются в форме, близкой к математической. Типичное использование MATLAB – это:

- математические вычисления
- создание алгоритмов
- моделирование
- анализ данных, исследования и визуализация
- научная и инженерная графика
- разработка приложений, включая создание графического интерфейса

Краткий обзор современных программных средств для проведения анализа данных.

Mathcad – программное средство, среда для выполнения на компьютере разнообразных математических и технических расчетов, снабженная простым в освоении и в работе графическим интерфейсом, которая предоставляет пользователю инструменты для работы с формулами, числами, графиками и текстами.

В среде Mathcad доступны более сотни операторов и логических функций, предназначенных для численного и символьного решения математических задач различной сложности и применения этих функций для анализа данных.

Краткий обзор современных программных средств для проведения анализа данных.

STATISTICA – это универсальная интегрированная система, предназначенная для статистического анализа и визуализации данных, управления базами данных и разработки пользовательских приложений, содержащая широкий набор процедур анализа для применения в научных исследованиях, технике, бизнесе, а также специальные методы добычи данных.

С помощью реализованных в системе STATISTICA мощных языков программирования, снабженных специальными средствами поддержки, легко создаются законченные пользовательские решения и встраиваются в различные другие приложения или вычислительные среды.

Краткий обзор современных программных средств для проведения анализа данных.

Deductor

Аналитическая платформа Deductor реализует практически все современные подходы к анализу структурированной табличной информации: хранилища данных (Data Warehouse), многомерный анализ (OLAP), добыча данных (Data Mining), обнаружение знаний в базах данных (Knowledge Discovery in Databases). Лучшим способом изучить и понять целесообразность использования современных технологий анализа - это испытать все на практике.

Краткий обзор современных программных средств для проведения анализа данных.

STATGRAPHICS – это универсальный пакет для анализа и визуализации данных. Отличительной особенностью пакета является наличие такого инструмента как **StatAdvisor**, который помогает пользователям интерпретировать полученные результаты, обеспечивает возможность объединения в одном окне нескольких текстовых и графических подокон.

StatAdvisor дает пользователям понятные разъяснения полученных результатов, определяет, являются ли эти результаты существенными, и обращает особое внимание на любые возможные ошибки в анализе. Пользователи получают немедленную интерпретацию результатов в процедурах, доступных в как основной системе, так и в четырех специальных модулях, поставляемых по выбору: Quality Control (контроль качества), Experimental Design (планирование эксперимента), Time-Series Analysis (анализ временных рядов) и Advanced Multivariate Method (анализ вариаций).