Лекция.

Элементы математической статистики.

План.

1. Статистика как наука. Этапы статистической работы.
2. I-й этап статистической работы. Генеральная совокупность и выборка.
3. I I-ой этап статистической работы. Описательная статистика:
- Гистограмма. Построение гистограммы.
- Характеристики статистического распределения: положения; рассеяния; формы.
4. I I I-й этап статистической работы. Оценка параметров генеральной совокупности:
- Точечная;
- Интервальная
- Интервальная оценка при малой выборке.
5. Планирование эксперимента. Определение необходимого объема выборочной совокупности.

1. Статистика как наука. Этапы статистической работы.

Математическая статистика – раздел математики, посвященный математическим методам систематизации, обработки и использования статистических данных для научных и практических выводов.

При этом **статистическими данными** называются сведения о числе объектов в какой-либо более или менее обширной совокупности, обладающих теми или иными признаками.

В процессе работы каждый врач, инженер, менеджер, научный работник непременно сталкивается со статистическими исследованиями (изучение спроса, анализ результатов эксперимента и т.д.). В любой статистической работе можно различить три этапа:

1. Сбор данных.
2. Обработка данных (описательная статистика). Все данные, полученные в результате эксперимента, необходимо записать в виде таблиц, графиков, схем.
3. Выводы и прогнозы. Один из самых важных разделов. По выборке наблюдений делают определенные выводы обо всем изучаемом процессе.

2. I-й этап статистической работы. Генеральная совокупность и выборка.

Рассмотрим 1-й этап статистической работы.

Обычно исследования проводятся не на единичных, а на групповых объектах, объединенных по какому-либо признаку.

Генеральная совокупность – это совокупность всех объектов, однородных по какому-либо признаку.

Число элементов генеральной совокупности называется ее **объемом** (N).

Часть генеральной совокупности, случайным образом, отобранная для наблюдений, называется **выборкой. n – объем выборки.**

Пример. Все студенты ОрГМУ – генеральная совокупность, 1 курс 2-ой поток леч.фак-та – выборка.

Чтобы высказать определенное суждение о свойствах генеральной совокупности, исследования проводят на выборке.

Чтобы проверить эффективность некоторого нового лекарственного препарата для лечения какого-либо заболевания, нет необходимости проводить исследования в отношении всех больных, страдающих этим заболеванием.

Выборка должна наиболее полно характеризовать свойства и особенности генеральной совокупности, т.е. она должна быть **репрезентативной** (достаточно представительной).

Способы формирования выборки.

1. **По способу лотереи** (случайный отбор). Т.е каждый элемент имеет одинаковую вероятность попасть в выборку
2. **Стратифицированная выборка**, при которой вся совокупность разбивается на группы, а затем в каждой группе делается случайный отбор.

3. Второй этап статистической работы. Описательная статистика.

Обработка данных начинается с упорядочения или систематизации собранных данных. Одна из форм систематизации данных – построение статистического ряда. Видное место занимают вариационные ряды.

Вариационным называют ряд, все значения которого располагают в порядке возрастания или убывания. Для того, чтобы наглядно представить закономерность варьирования количественных признаков, вариационные

ряды принято изображать в вид графиков. В медицине наиболее часто для наглядного изображения вариационного ряда используют гистограмму или полигон частот.

Рассмотрим построение гистограммы.

Предположим, что в результате эксперимента получен ряд значений случайной величины: X_1 X_2 X_3 …..X_n

1. Строят вариационный ряд. Все данные располагают в порядке возрастания.

2. Находят размах выборки (варьирования)

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

3. Определяют число классов. При большом ряде прибегают к группировке. Число групп или классов находят по формуле: К=2Lnn,где n – объем выборки. Если n≤40, то **к** принимают равным 3 или 4.

4. Находят величину класса: $d = \frac{R}{K}$

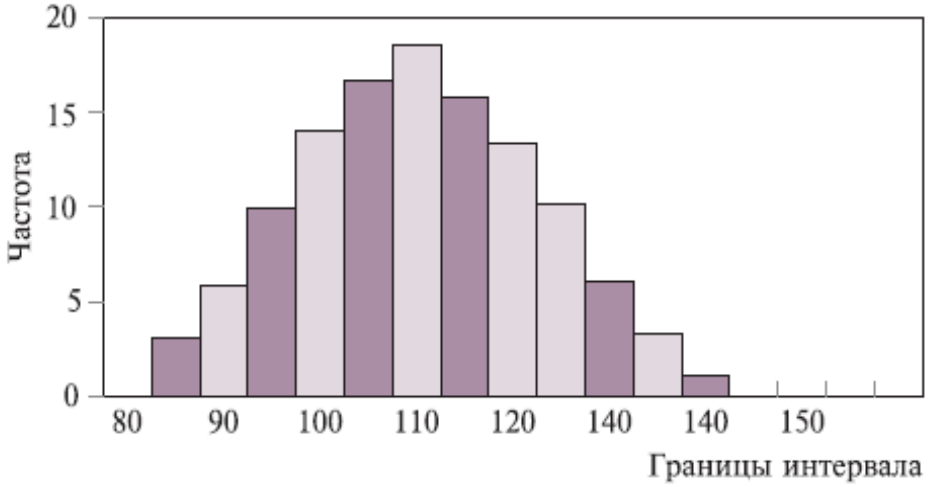
5. Разбивают выборку на классы:
   1. X_{min}– X_{min}+d
   2. X_{min}+d– X_{min}+2d
   3. X_{min}+2d– X_{min}+3d и т.д.

6. Находят число измерений, попавших в каждый класс (частота попадания–h_i).

7. Определяют функцию плотности вероятности (эмпирическую плотность вероятности случайной величины)

$$f(x) = \frac{h_i}{nd}$$

8. Строят гистограмму: по оси абсцисс откладывают границы классовых интервалов, по оси ординат – значения функции плотности вероятности – f(x).



Характеристики распределения.

Вариационные ряды и их графики дают наглядное представление о варьировании признаков, но они не достаточны для полного описания варьирующих объектов. Для этой цели служат особые числовые показатели, называемые статистическими характеристиками.

Характеристики положения – характеризующие центральную тенденцию или уровень ряда. К ним относятся: среднее арифметическое, медиана и мода.

Среднее арифметическое ($\overline{x}$) – это сумма всех членов совокупности, деленная на их общее число.

$$\overline{x}_B = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Медиана ($\tilde{x}$) – средняя, относительно которой ряд распределения делится на две равные части. В обе стороны от медианы располагается одинаковое число значений (вариант).

Пусть, например, имеется ранжированная выборка, содержащая нечетное число членов n = 9: 12 14 14 18 **20** 22 22 26 28. Тогда медиана будет равна **20.**

Если выборка содержит четное число членов, то медиана не может быть определена столь однозначно. Например, получен ряд из 10 членов: 6 8 10 12 **14 16** 18 20 22 24.

Медианой в этом случае может быть любое число между 14 и 16 (5-м и 6-м членами ряда). Для определенности принято считать в качестве медианы среднее арифметическое этих значений, т. е. $Me = \frac{14+16}{2} = 15.$

Мода ($\hat{x}$) – вершина распределения. Величина, встречающаяся в выборке наиболее часто.

Дана выборка: 12 14 14 18 18 18 18 18 20 22 22 26 28. Мода равна **18.**

При нормальном распределении случайной величины все три характеристики положения совпадают. $\overline{x} = \tilde{x} = \hat{x}$

Характеристики рассеяния

Средние значения не дают полной информации о варьирующем (изменяющемся) признаке. Нетрудно представить себе два эмпирических распределения, у которых средние одинаковы, но при этом у одного из них значения признака рассеяны в узком диапазоне вокруг среднего, а у другого – в широком. Поэтому наряду со средними значениями вычисляют и характеристики рассеяния выборки. Рассмотрим наиболее употребительные из них.

Дисперсия – средний квадрат отклонения значений признака от среднего арифметического. Характеризует степень рассеяния случайной величины вокруг ее математического ожидания (среднего арифметического).

$$D = \frac{\sum (x_1 - \bar{x})^2}{n}$$

Стандартное отклонение (среднеарифметическое отклонение). Характеризует степень рассеяния случайной величины вокруг ее математического ожидания (среднего арифметического). Эта величина оказывается более удобной для расчетов, т.к. выражена в линейных единицах.

$$\delta = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Коэффициент вариации – численно равен стандартному отклонению, выраженному в процентах, от величины средней арифметической.

$$Cv = \frac{\delta}{\bar{x}} * 100\%$$

Этот показатель позволяет сравнивать изменчивость признаков, выраженных разными единицами.

Нормированное отклонение – это отклонение той или иной варианты от средней арифметической, отнесенное к величине среднеквадратического отклонения (стандартного отклонения).

$$t_p = \frac{x_i - \overline{x}}{\delta}$$

Этот показатель позволяет «измерять» отклонение отдельных вариант от среднего уровня и сравнивать их для разных признаков.

Характеристики формы.

$\overline{x}$, D,σ не содержат информации о законе распределения. Не все признаки распределяются по нормальному закону. Некоторые обнаруживают явную ассиметрию. Возможны и другие случаи отклонений от нормального закона. Приближенную оценку закона распределения можно получить при помощи коэффициентов ассиметрии и эксцесса.

Коэффициент ассиметрии является мерой скошенности рядов и вычисляется по формуле:

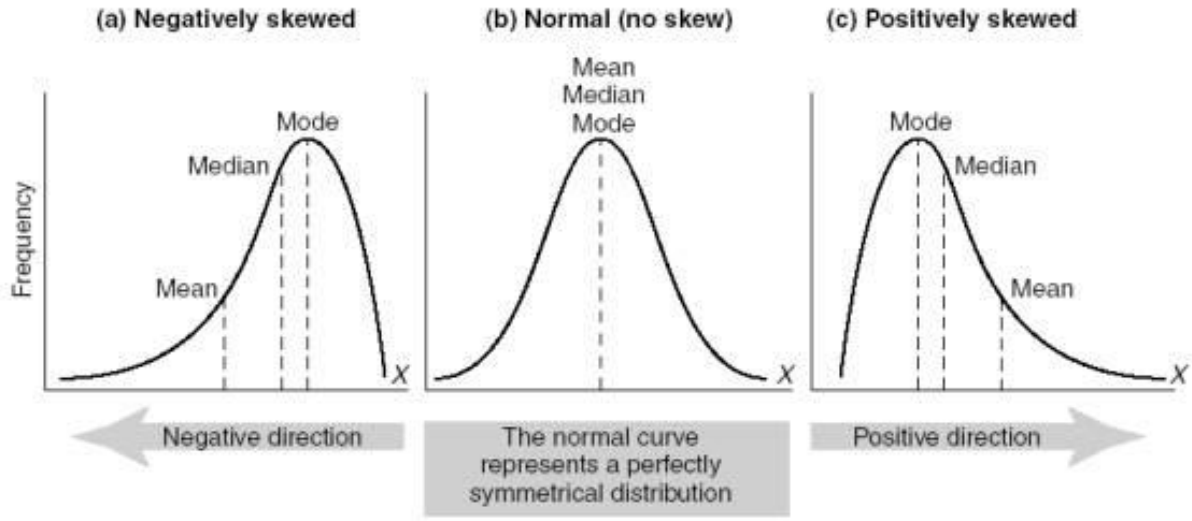


FIGURE 15.6 Examples of normal and skewed distributions

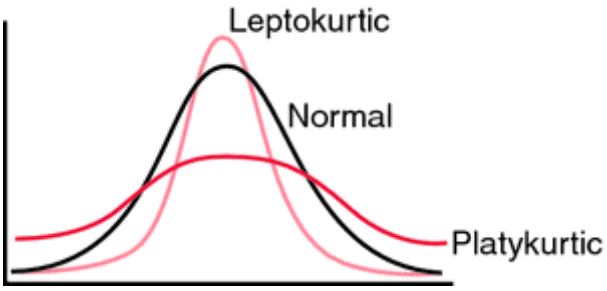
Коэффициент эксцесса

Коэффициент эксцесса в статистике — мера остроты пика в распределении случайной величины.

Эксцесс характеризует распределения, в которых значения величин либо сосредоточены близко к средней величине, либо наоборот распределены далеко от нее.

Положительный эксцесс (leptokurtic) — острая вершина, когда пик выше, чем пик нормального распределения.

Отрицательный эксцесс (platykurtic) — тупая вершина, когда пик ниже пика нормального распределения).



Коэффициент эксцесса вычисляется по формуле.

4. Третий этап статистической работы.

Оценка параметров генеральной совокупности.

Числовые показатели, характеризующие генеральную совокупность, называют генеральными параметрами, характеризующими выборку – выборочными характеристиками.

$\overline{X}$ (выборочная средняя) может служить оценкой для **μ** – средняя генеральной совокупности;

σ^2 (выборочная дисперсия) – для S_x^2 (генеральная дисперсия);

σ(выборочное стандартное отклонение) – для S_x генеральной совокупности.

	Параметры генеральной совокупности	Параметры выборки
Среднее значение	μ	$\overline{X}$
Дисперсия	S_x^2	σ^2
Стандартное отклонение	S_x	σ

Это **точечные оценки**, представляющие собой числа («точки»), вычисляемые по случайной выборке.

Точечной называют оценку, которая определяется одним числом.

Выборочные характеристики, как величины случайные, варьирующие вокруг своих генеральных параметров, в основном не совпадают с ними по абсолютной величине.

Поэтому, применяют второй вид оценки параметров генеральной совокупности – интервальную оценку.

Интервальная оценка параметров генеральной совокупности.

В некоторых случаях представляет интерес не получение точечной оценки неизвестного параметра генеральной совокупности, а определение некоторого интервала, в котором может находиться этот параметр с заданной вероятностью.

Доверительным интервалом называют интервал, в котором с той или иной вероятностью находится генеральный параметр.

Вероятность, с которой гарантируется попадание параметра генеральной совокупности внутрь доверительного интервала, называется **доверительной.**

Обычно в качестве доверительных используют вероятности: $\mathbf{P_1=0,95;}$ $\mathbf{P_2=0,99;\ P_3=0,999.}$

Это означает, что параметр генеральной совокупности попадет в указанный интервал в первом случае в 95 случаев из 100, во втором – в 99 случаях из 100, в третьем случае – в 999 случаев из 1000.

В некоторых случаях указывается не доверительная вероятность, а вероятность обратных случаев, когда параметр не попадает в указанный интервал. Вероятность таких маловероятных случаев, называется **уровнем значимости α** и равна**: α = 1-P.**

Уровень значимости – это вероятность, которой можно пренебречь.

Доверительным вероятностям соответствуют следующие величины нормированных отклонений ($\mathbf{t_p}$).

Доверительная вероятность **(P)**	Нормированное отклонение ($\mathbf{t_p}$)	Уровень значимости **α = 1-P.**
0,95	1,96	0,05
0,99	2,58	0,01
0,999	3,29	0,001

Для нормального закона распределения, зная величину выборочной средней и ее ошибку, можно определить границы, в которых с той или иной вероятностью находится параметр генеральной совокупности – среднее значение **μ.**

$$\bar{x} - t_p\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + t_p\frac{\sigma}{\sqrt{n}},$$ где

$\overline{x}$ – среднее значение выборки;

t_p – нормированное отклонение;

σ – стандартная ошибка;

n – объем выборки;

μ –среднее значение генеральной совокупности.

Интервальная оценка при малой выборке.

Если объем выборки ≤30, то такая выборка считается малой. Для нахождения доверительного интервала для среднего значения генеральной совокупности значение нормированного отклонения берут из таблицы «Значение коэффициента Стьюдента» в зависимости от объема выборки и выбранной доверительной вероятности.

5. Планирование эксперимента. Определение необходимого объема выборочной совокупности.

Планирование эксперимента и обработка их результатов – это две тесно связанные между собой задачи статистического анализа.

Термин «эксперимент» означает искусственно организуемый комплекс условий, в которых испытывают воздействие того или иного фактора или одновременно нескольких факторов на результативный признак. В фармакологии – это испытание эффективности новых лечебных препаратов, в медицине – проверка разных способов лечения больных и т.д.

Термин «эксперимент» можно применять в более широком смысле, понимая под ним любые испытания, проводимые исследователем в отношении изучаемого объекта.

При всем многообразии методов исследовательской работы задача планирования сводится к тому, чтобы при возможно минимальных объемах наблюдений получить достаточно полную информацию об изучаемых объектах. Практический опыт подсказывает, что неразумно стремиться к неоправданно большому числу испытаний, если убедительный результат можно получить при минимально допустимом объеме выборки – n.

Формула необходимого объема выборочной совокупности:

$$\mathbf{n} = \frac{t^2 6^2}{\Delta^2},$$

где Δ – точность эксперимента, разность между средними значениями генеральной совокупности и выборки.