Статистический анализ

Тема7

Биометрия – это инструмент эмпирического познания живой природы.

- 1. Задача количественного представления биологических фактов (измерение)
- 2. Задача обобщенного описания множества фактов (статистическое оценивание)
- 3. Задача поиска закономерностей (проверка статистических гипотез).

Этапы биометрического исследования

- 1. Определить объект исследования.
- 2. Определить проблему (и актуальность) исследования. «Что плохо?»
- 3. Определить цель исследования. «Чего хочется?»
- 4. Определить задачи исследования.«Что сделать?»
- 5. Сбор и накопление данных, изучение биологического явления.
- 6. Решение биометрической задачи.

Таблица 1

Задача	Статистический показатель	Метод
<u>Оценить принадлежность…</u>		
варианты к выборке	средняя арифметическая и значение отдельной варианты (M, x)	оценка «выскакивающих» значений (критерий Стьюдента t)
<u>Оценить достоверность отличия…</u>		
двух выборок по величине признака	средняя арифметическая (M)	сравнение средних арифметических (критерий Стьюдента t)
двух выборок по изменчивости признака	дисперсия (S^2), стандартное отклонение (S), коэффициент вариации (CV)	сравнение дисперсий (критерий Фишера F)
двух выборок в целом	ранги (R)	сравнение степени упорядоченности вариант (критерий U Уилкоксона, критерий Q Розенбаума)
эмпирического и теоретического распределений	частоты встречаемости вариант (классов вариант) (a, A)	сравнение частотных распределений (критерий Пирсона χ^2)
<u>Оценить достоверность влияния…</u>		
фактора на величину признака	факториальная и случайная дисперсия (S^2), сила влияния (η^2)	дисперсионный анализ (критерий Фишера F)
одного признака на другой признак	коэффициент регрессии (a)	регрессионный анализ (критерий Фишера F и критерий Стьюдента t)
двух признаков друг на друга (взаимодействие)	коэффициент корреляции (r)	корреляционный анализ (критерий Стьюдента t)

- Нулевая гипотеза (H0) – это гипотетическое предположение об отношениях объектов, выраженное в терминах статистики и предназначенное для дальнейшей статистической проверки
- Статистический вывод. Статистический вывод, главный результат статистического анализа, – это заключение о справедливости или опровержении нулевой гипотезы.

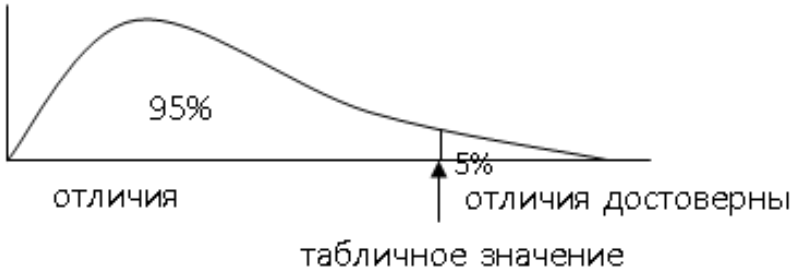


Рис. 1. *Схема использования критериев. Отмечены критические зоны для уровней значимости α = 0.05 и α = 0.01 (доверительные вероятности P = 0.95 и P = 0.99). Границами зон служат значения критериев из таблиц Приложения при данном уровне значимости. Если вычисленные величины критерия попадают в критическую зону (правее табличных), значит, отличие сравниваемых параметров достоверно*

ВЫБОРКА

- Выборка – множество значений случайной величины, совокупность вариант, набор чисел
- Генеральная совокупность – это множество всех вариант определенного типа (выборка бесконечного размера).

- Признак (свойство, показатель, величина, характеристика, переменная) – любая информация о наблюдаемом объекте, выраженная качественно или определенная количественно.

Существует целый ряд методов регистрации признаков биологических объектов.

- Качество (нечисловой дискретный признак) –выражаются словами или символами, они не имеют количественного содержания и выражают принадлежность данного объекта к определенной обширной группе объектов (зеленый, январь, ♀, ♪).
- Балл (оценка) – дискретный полуколичественный признак, численная характеристика объекта, присвоенная в соответствии с внешней заранее принятой шкалой баллов.
- Количество (число) – дискретный (счетный) количественный признак (число натурального ряда), характеризующий множество однородных объектов, черт, деталей строения, состав
- Проба – ограниченная совокупность разнокачественных объектов, которая характеризуется числом объектов одного определенного качества, это значение играет роль одной варианты выборки.
- Промер (ряд дробных или рациональных чисел) – непрерывный (мерный) количественный признак, характеризующий свойства объектов с помощью различных относительных количественных шкал – температурной, весовой, размерной, объемной и т. п.

Основная особенность выборки как множества значений случайной вели-чины – это отличие отдельных вариант друг от друга, явление изменчивости, варьирования, появления отличий между отдельными вариантами.

Один из источников, эндогенный, – это индивидуальные отличия по статусу и по состоянию.

Другой источник отличий между вариантами – факторы внешней среды.

$$x_i = x_{дом.} \pm x_{случ.},$$

где x_i – измеренное значение варианты,

i – индекс варианты (i = 1, 2, …, n),

n – объем (общее количество вариант) выборки,

$x_{дом.}$ – суммарный вклад j доминирующих факторов,

$x_{случ.}$ – суммарный вклад k случайных факторов.

ВЫЧИСЛЕНИЕ ПАРАМЕТРОВ ВЫБОРОК

- Средняя арифметическая

$$M = \frac{\sum x_i}{n}$$

- *Медиана* (Me)
- *Мода* (Mo)

- Стандартное отклонение

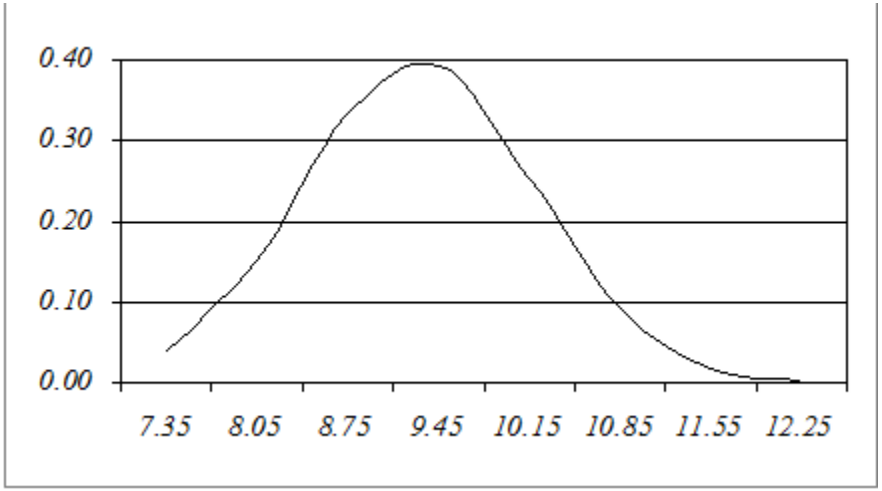
$$S = \sqrt{\frac{\sum (x - M)^2}{(n-1)}}$$

$$S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}}$$

- коэффициент вариации (CV)

$$CV = \frac{S}{M} \cdot 100\%$$

ОСНОВНЫЕ ТИПЫ РАСПРЕДЕЛЕНИЙ ПРИЗНАКОВ



Нормальное
Биномиальное
Распределение Пуассона
Альтернативное распределение
Полиномиальное распределение

Рис. 4. Нормальное распределение с параметрами n = 63, M = 9.3, S = 0.79. По оси абсцисс – вес тела землероек-бурозубок, по оси ординат – табличные значения для нормального распределения. Рассчитать ординаты нормальной кривой для конкретного значения xi можно по формуле:

$$p_i = (1 / \sqrt{2\pi}) \cdot e^{-(x_i - M)^2 / 2 \cdot S^2}$$

ОЦЕНКА РАЗЛИЧИЙ ДВУХ ВЫБОРОК

- Сравнение средних арифметических (параметрические критерии) t, F

1. Обе выборки взяты из одной генеральной совокупности, но средние отличаются в силу ошибки репрезентативности.

2. Выборки взяты из разных генеральных совокупностей, отличие средних вызвано, в основном, действием разных доминирующих факторов

Формула для t критерия отличия средних:

$$t = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}}$$

$\sim t_{(\alpha,\, df)}$.

Сравнение выборок с помощью непараметрических критериев

- Ранг – это число натурального ряда, которым обозначается порядковый номер каждого члена упорядоченной совокупности вариант

1 упорядочивание и ранжирование вариант, 2 подсчет сумм рангов в соответствии с правилами данного критерия, 3 сравнение полученной величины с табличным значением критерия

Критерий U Уилкоксона – Манна – Уитни

Ранги вариант суммируют отдельно по каждой выборке:

$R_1 = \Sigma r_i$, $R_2 = \Sigma r_j$, $i = 1, 2, \ldots, n_1$, $i = 1, 2, \ldots, n_2$

и вычисляют величину критерия:

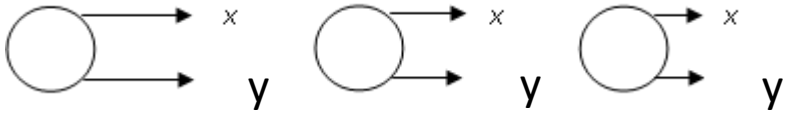
,

$$t = \frac{U - 0.5 \cdot n_1 \cdot n_2}{\sqrt{(n_1 \cdot n_2 \cdot (n+1)/12)}}$$

где $U = \max(U_1, U_2)$ – максимальное значение из двух величин:

- $U_1 = n_1 \cdot n_2 + 0.5 \cdot n_1(n_1 + 1) - R_1$
- $U_2 = n_1 \cdot n_2 + 0.5 \cdot n_2(n_2 + 1) - R_2$

ОЦЕНКА ЗАВИСИМОСТИ МЕЖДУ ПРИЗНАКАМИ



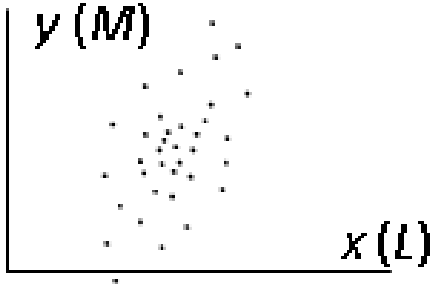


Рис. 10. Область рассеяния вариант

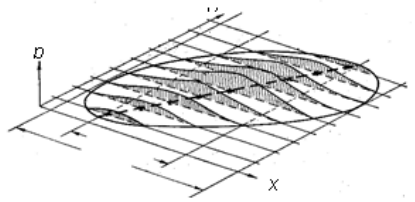


Рис. 11. Двумерное распределение

Взаимная связь (взаимная зависимость) двух признаков при их изменчивости, т. е. сопряженность их вариации, называется корреляцией.

$$r = \sqrt{\frac{ковариация}{изменчивость}} = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{\sum (y - M_y)(x - M_x)}{\sqrt{\sum (y - M_y) \cdot \sum (x - M_x)}}$$

Таблица 13

i	y	x	y^2	x^2	x·y
1	25	352	625	123904	8800
2	26	376	676	141376	9776
3	31	402	961	161604	12462
4	32	453	1024	205208	14496
5	34	484	1156	234256	16456
6	38	528	1444	278784	20064
7	38	555	1444	308025	21090
Σ	224	3150	7330	1453158	103144