Биологическая статистика

Математическая статистика-это раздел математики, посвященный математическим методам систематизации, обработки и использования данных для научных и практических выводов.
Генеральная совокупность-это совокупность объектов, которые отличаются друг от друга, но схожие определенным признаком.
Выборка-это часть генеральной совокупности.
О свойствах генеральной совокупности можно судить по свойствам выборки, поэтому она должна быть репрезентативной.
Вариационный ряд-это данные расположенные в порядке возрастания.
Для наглядности данные представляют в виде полигона или гистограммы распределения.
Гистограмма-это ступенчатая фигура, состоящая из прямоугольников, основании которых равны ширине класса, а высоты-функции плотности вероятности.

Построение гистограммы.
Предположим, что в результате эксперимента получен ряд значений случайной величины –X_i
X_1 X_2 X_3 …….. X_n
1. Строят вариационный ряд-все данные располагают в порядке возрастания.
2. Находят размах варьирования- $R=X_{max}-X_{min.}$
3. При большом ряде прибегают к группировке. Число групп или классов находят по формуле: К=2Ln**n.**
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 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).

Задача: Измерена концентрация сывороточного альбумина (г/.л) в крови 50 женщин, включённых в одно обследование. По полученным данным построить гистограмму.

42 41 42 44 44 36 38 41 42 44 42 39 49 40 45 32
34 43 37 39 41 39 48 42 43 33 43 35 32 39 35 43
44 47 40 39 42 41 46 37 49 41 39 43 42 47 48 51
52 34

Решение:

1. Строят вариационный ряд-все данные располагают в порядке возрастания:

32	32	33	34	35	35	35	35	36	37
37	38	39	39	39	39	39	39	40	40
41	41	41	41	41	41	42	42	42	42
42	42	43	43	43	43	43	44	44	44
46	46	47	47	48	48	49	49	51	52

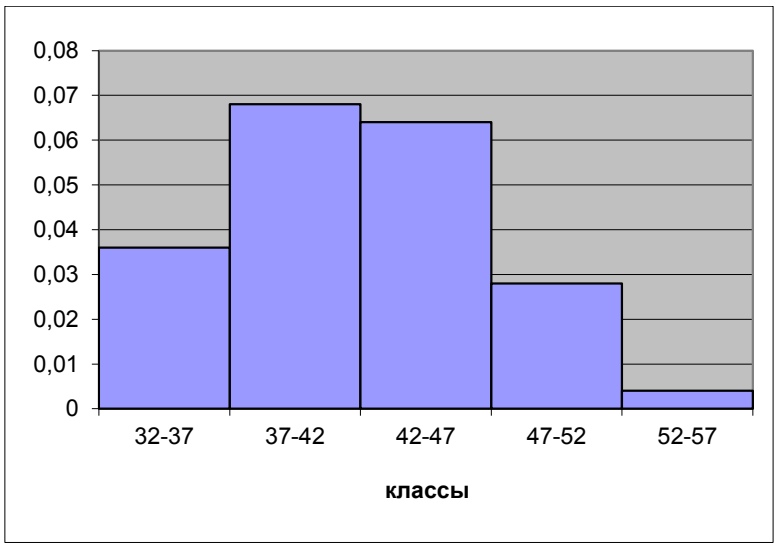
2. Находят размах выборки: $\mathbf{R=X_{max} - X_{min.}}$
   R=52-32=20
3. Выбирают количество классов: **к=4**;
4. Находят ширину одного класса по формуле: **d=R/k;** **d=20/4=5;**
5. Разбивают вариационный ряд на классы и находят частоту попадания в каждый класс:
   a) **32-37** $\mathbf{h_1=9}$
   b) **37-42** $\mathbf{h_2=17}$
   c) **42-47** $\mathbf{h_3=16}$
   d) **47-52** $\mathbf{h_4=7}$
   e) **52-57** $\mathbf{h_5=1}$

6.Расчитывают функцию плотности вероятности по каждому классу по формуле: $f(x)=\frac{h_i}{nd}$

1. $f_1=9/250 =0.036$
2. $f_2=17/250=0.068$
3. $f_3=16/250=0.064$
4. $f_4=7/250 =0.032$
5. $f_5=1/250 =0.004$

7 Строят гистограмму, откладывая по оси X значения случайной величины, а по Y-(F)-значения функции плотности вероятности:

№ класса	**1**	**2**	**3**	**4**	**5**
классы	32-37	37-42	42-47	47-52	52-57
F	0,036	0,068	0,064	0,028	0,004



Расчёт моды и медианы.

Для величин, по которым построена гистограмма, медиану можно определить следующим способом. Необходимо найти класс, в котором содержится медиана. Для этого необходимо складывать частоты встречаемости по классам до тех пор, пока сумма частот не превзойдет половину всех членов ряда. Данный класс называется медианным. Тогда медиану можно найти по формуле: $Me = x_{н} + \lambda(\frac{\frac{n}{2} - \sum f_i}{f_{Me}})$

где x_n- нижняя граница интервала, содержащего медиану,

$\sum f_i$- сумма накопленных частот, стоящая перед медианным классом,

λ- величина классового интервала,

f_{Me} - частота медианного класса,

n- общее число наблюдений.

Подставим числовые данные в формулу и рассчитаем медиану:

$Me = 37 + 5(\frac{\frac{25}{2} - 9}{17}) = 41.7$

Мода- это величина, наиболее часто встречающаяся в данной совокупности. Класс с наибольшей частотой называется модальным. Моду можно найти по формуле: $Mo = x_n + \lambda(\frac{f_2 - f_1}{2f_2 - f_1 + f_3})$

Где: x_n- нижняя граница модального класса,

f_2, f_1- частота класса, предшествующего модальному,

f_3- частота класса, следующего за модальным,

λ- ширина классового интервала.

Подставим числовые данные в формулу и рассчитаем моду:

$Ìî = 37 + 5(\frac{17 - 9}{2 \cdot 17 - 9 + 16}) = 38$

Расчёт коэффициентов ассиметрии и эксцесса.

Коэффициент ассиметрии определяется по формуле:

$$As = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^{\frac{3}{2}}}$$

Эксцесс определяется по формуле:

$$\acute{Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2} - 3$$

Интервальная оценка параметров генеральной совокупности.

По известным выборочным характеристикам можно построить интервал, в котором с той или иной вероятностью находится генеральный параметр. Вероятности, признанные достаточными для уверенного суждения о генеральных параметрах на основании известных выборочных показателей, называют доверительными. Обычно в качестве доверительных используют вероятности **P_1=0.95, P_2=0.99, P_3=0.999.**

Это означает, что при оценке генеральных параметров по известным выборочным показателям существует риск ошибиться в первом случае один раз на 20 испытаний, во втором- один раз на 100 испытаний и в третьем- один раз на 1000 испытаний.

Доверительным вероятностям соответствуют следующие величины нормированных отклонений:

вероятности **P_1=0.95** соответствует **t_1=1.96;**
вероятности **P_2=0.99** соответствует **t_2=2.58;**
вероятности **P_3=0.999** соответствует **t_3=3.29;**

$\bar{x} - t_p \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + t_p \frac{\sigma}{\sqrt{n}}$ -формула доверительного интервала

где $\bar{X}$ - среднее значение выборки;

t_p – нормированное отклонение;

S_x- стандартная ошибка на генеральной совокупности;

σ_x - стандартная ошибка на выборке;

n – объём выборки;

μ-среднее значение генеральной совокупности.

Задача:

Распределение кальция в сыворотке крови обезьян, как было установлено выше, характеризуется следующими выборочными показателями: $\bar{X}$ =11.94 мг, σ =1.27 мг, n =100. Построить 95% доверительный интервал для генеральной средней **μ** этого распределения.

Дано:

$\bar{X}$ =11.94 мг

σ=1.27 мг

n =100

P=0.95

_____

μ=?

Решение:

Применяют формулу доверительного интервала.

$\bar{x} - t_p \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + t_p \frac{\sigma}{\sqrt{n}}$

Подставляют численные данные:

$11.94-1.96\frac{1.27}{\sqrt{100}} \le \mu \le 11.94+1.96\frac{1.27}{\sqrt{100}}$

или $11.70 \le \mu \le 12.18$

Следовательно, с вероятностью P=0.95 можно утверждать, что генеральная средняя данного нормального распределения находится между 11.70 и 12.18 мг.

Ответ: $11.70 \le \mu \le 12.18$

Задача:

Исследователь хочет установить средний уровень гемоглобина в определенной группе населения. Сколько человек он должен обследовать, если в 99 случаев из 100 Δ=±5г/л, а **σ**=32г/л.?

Решение:

Дано:

σ_x =32

n =10

P=0.99

Δ=5_____

μ=?

Решение:

Применяем формулу необходимого объёма выборочной совокупности:

$n = \frac{t^2 \cdot \sigma_x^2}{\Delta^2}$

Где: $\Delta = X - \mu$-ошибка эксперимента

$\bar{X}$ - среднее значение выборки;

t_p – нормированное отклонение;

σ_x - стандартная ошибка на выборке;

n – необходимый объём выборки

μ-среднее значение генеральной совокупности.

$n=(\frac{2.58 \cdot 32}{5})^2 \approx 273$

Ответ: n=273