Лекция 2. Интервальная статистика

1. Введение в интервальную статистику

Интервальная статистика – раздел математики, возникший на границе между интервальной математикой и математической статистикой.

Объектом исследования интервальной статистики являются интервальные наблюдения, т.е. наблюдения, заданные интервалом значений. Основной задачей интервальной статистики (также как и математической статистики) является восстановление статистических зависимостей (закономерностей).

1.1.Интервальная арифметика

Основная идея интервального анализа состоит в том, что вещественное чистло представляется не одним, а двумя числами – оценкой снизу и оценкой сверху, образующими интервальное число.

Арифметические операции над интервальными числами выполняются следующим образом:

$$[a_1, a_2] = [b_1, b_2] \circ [c_1, c_2], \text{ если } b \in [b_1, b_2], \text{ и } c \in [c_1, c_2], \text{ то } b \circ c \in [a_1, a_2],$$

где "$\circ$" – обычная арифметическая операция над вещественными числами (<+>,<->,<*>,</>).

Множество всех интервалов на R обозначается через IR.

Определение Если $r(x)$ – непрерывная унарная операция на R, то $r(X) = \left[\min_{x \in X} r(x), \max_{x \in X} r(x)\right]$, $X \in IR$, определяет соответствующую ей операцию на множестве IR.

Особенностью такого определения интервальных чисел является то, что произвольный невырожденный интервал из IR не имеет обратного ни по сложению, ни по умножению. Вместо *дистрибутивности* вещественных чисел для интервальных чисел выполняется свойство *субдистрибутивности*:

$$A(B + C) \subseteq AB + AC,$$

которое лежит в основе "*интервального расширения*".

Упражнения

- Пусть $A = [1,3]$, $B = [2,4]$. Найти $A + B$, $A - B$, AB, A / B.
- Вывести формулы для вычисления арифметических операций над интервальными числами через арифметические операции над вещественными числами и операции максимума и минимума.
- Доказать свойство субдистрибутивности интервальных чисел. Привести пример, когда нарушается дистрибутивность интервальных чисел.

1.2. Интервальная выборка

1.2.1. Абсолютная погрешность

Пусть в результате эксперимента наблюдается случайная величина $\xi + \eta$. Первая случайная величина задаёт статистическую неопределенность, а вторая – η задает измерительную погрешность, действующую аддитивно на результат измерения. Про погрешность измерения известно, что $|\eta| < \Delta$, где $\Delta > 0$ – максимальная абсолютная погрешность измерения.

Нас интересует распределение случайной величины ξ, при неизвестном распределении ошибки η.

Теорема 1.1. При сделанных выше предположениях

$$F_{\xi}(x-\Delta) \le F_{\xi+\eta}(x) \le F_{\xi}(x+\Delta)$$

Доказательство. Воспользуемся свойством монотонности функции распределения: если $x_1 \le x_2$, то $F(x_1) \le F(x_2)$. Имеем:

$$F_{\xi+\eta}(x) = P\{\xi+\eta < x\} = P\{\xi < x-\eta\} = F_{\xi}(x-\eta)$$

Так как $x-\eta \ge x-\Delta$, то $F_{\xi}(x-\eta) \ge F_{\xi}(x-\Delta)$. Аналогично, так как $x-\eta \le x+\Delta$, то $F_{\xi}(x-\eta) \le F_{\xi}(x+\Delta)$. Теорема доказана.

Таким образом, когда наблюдается случайная величина с аддитивной измерительной погрешностью, то в результате может получиться любое распределение в полосе от $F_\xi(x-\Delta)$, до $F_\xi(x+\Delta)$, как показано на рис. 1.1.

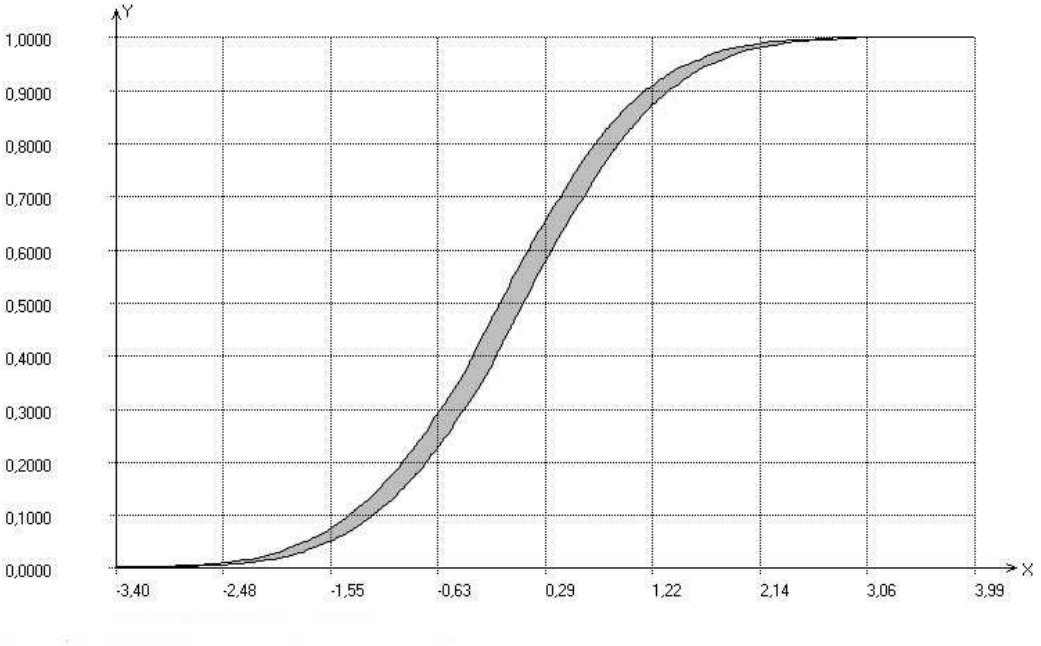


Рис. 1.1. Полоса распределений случайной величины с аддитивной погрешностью

1.2.1. Относительная погрешность

Пусть в результате эксперимента наблюдается случайная величина $\xi(1+\eta)$. Первая случайная величина задаёт статистическую неопределенность, а вторая – η задает измерительную погрешность, действующую мультипликативно на результат измерения. Про погрешность измерения известно, что $|\eta| \le \Delta < 1$, где $\Delta > 0$ – максимальная относительная погрешность измерения. Нас интересует распределение случайной величины ξ, при неизвестном распределении ошибки η.

Теорема 1.2. При сделанных выше предположениях

$$F_\xi\left(\frac{x}{1+\Delta}\right) \le F_{\xi(1+\eta)}(x) \le F_\xi\left(\frac{x}{1-\Delta}\right),\ x>0$$
$$F_\xi\left(\frac{x}{1-\Delta}\right) \le F_{\xi(1+\eta)}(x) \le F_\xi\left(\frac{x}{1+\Delta}\right),\ x<0$$

Доказательство. Воспользуемся свойством монотонности функции распределения: если $x_1 \le x_2$, то $F(x_1) \le F(x_2)$. Имеем:

$$F_{\xi(1+\eta)}(x) = P\{\xi(1+\eta) < x\} = P\left\{\xi < \frac{x}{1+\eta}\right\} = F_\xi\left(\frac{x}{1+\eta}\right)$$

Пусть $x > 0$. Тогда $\frac{x}{1+\Delta} \le \frac{x}{1+\eta} \le \frac{x}{1-\Delta}$ и, следовательно,

$$F_\xi\left(\frac{x}{1+\Delta}\right) \le F_\xi\left(\frac{x}{1+\eta}\right) \le F_\xi\left(\frac{x}{1-\Delta}\right).$$

Пусть $x < 0$. Тогда $\frac{x}{1-\Delta} \le \frac{x}{1+\eta} \le \frac{x}{1+\Delta}$ и, следовательно,

$$F_\xi\left(\frac{x}{1-\Delta}\right) \le F_\xi\left(\frac{x}{1+\eta}\right) \le F_\xi\left(\frac{x}{1+\Delta}\right).$$

Если же $x = 0$, то неравенства обратятся в равенство.
Теорема доказана.

На рисунке 1.2 показана полоса распределений в случае мультипликативной погрешности измерений. На практике, конечно, скорее всего будут наблюдаться оба вида погрешностей, и, если объединить результаты теорем 1.1 и 1.2, то при максимальной аддитивной погрешности Δ_1 и максимальной мультипликативной погрешности Δ_2 распределение случайной величины $\xi(1+\eta_1)+\eta_2$ будет находиться в полосе изображенной на рисунке 1.3.

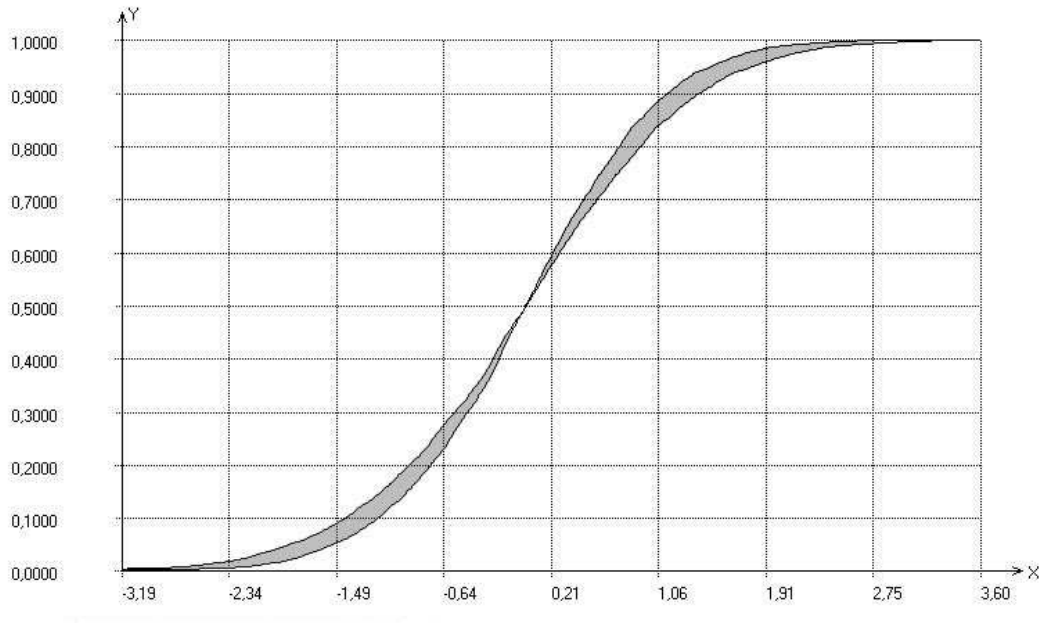


Рис. 1.2. Полоса распределений случайной величины с мультипликативной погрешностью

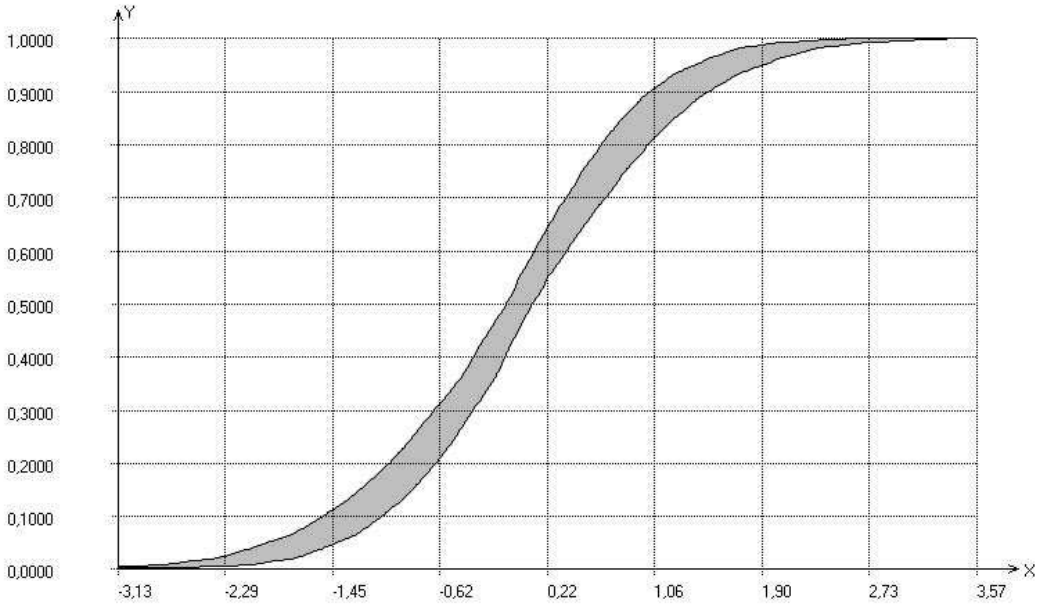


Рис. 1.3. Полоса распределений случайной величины с аддитивной и мультипликативной погрешностями

1.2.3. Интервальные наблюдения

Пусть на практике было произведено измерение x какой-либо случайной величины ξ. Естественно, что в результате измерений допущена погрешность, и, на самом деле, x – это реализация случайной величины $\xi(1+\eta_1)+\eta_2$. Причем распределения η_1 и η_2 не только неизвестны, но и могут меняться от эксперимента к эксперименту (например, при смене прибора, которым выполняются измерения). Пусть мы знаем максимально возможные погрешности Δ_1 и Δ_2 для η_1 и η_2 соответственно. Тогда мы можем утверждать, что при положительном x

$$x \in \left\lfloor x_{ист}(1-\Delta_1)-\Delta_2, x_{ист}(1+\Delta_1)+\Delta_2 \right\rfloor,$$

где $x_{ист}$ – истинное значение реализации случайной величины ξ. Однако в то же время можно утверждать, что

$$x_{ист} \in \left\lfloor x\ (1-\Delta_1)-\Delta_2, x\ (1+\Delta_1)+\Delta_2 \right\rfloor$$

Естественно, что эти интервалы пересекаются, но если первый интервал для нас неизвестен (т.к. неизвестно $x_{ист}$), то второй интервал известен и позволяет оценить $x_{ист}$ сверху и снизу (см. рис. 1.4).

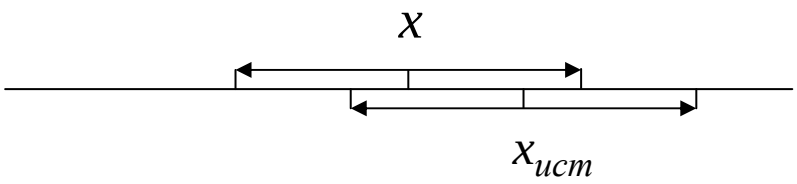


Рис. 1.4. Истинное значение случайной величины и наблюдение

Определение. *Интервальным наблюдением* называется интервал, содержащий значение реализации случайной величины.

Определение. *Интервальной выборкой* объема n называется множество из n интервальных наблюдений:

$$\mathbf{X}_n = \left\{ [a_i, b_i] \in IR \,\middle|\, a_i \le x_i \le b_i, a_i \in R, b_i \in R, i = 1, \ldots, n \right\}$$

Замечание 1. К интервальной выборке могут привести процедуры группирования и цензурирования. Отличие заключается в том, что интервалы группирования задаются априори, а в модели с погрешностями измерений границы интервалов порождаются самими наблюдениями и, таким образом, также являются случайными.

Замечание 2. Интервалы $[a_i, b_i]$ могут быть бесконечными. Эта ситуация может возникнуть, например:

а) в случае, когда стрелка измерительного прибора зашкаливает и, поэтому установить точное значение границы невозможно;

б) при испытаниях на надежность фиксируется момент выхода прибора из строя. На момент окончания испытаний часть приборов все еще работает, поэтому время их поломки неизвестно.

1.3 Геометрическая интерпретация интервальной выборки

В пространстве R^n выборки, традиционно рассматриваемые в математической статистике, $X_n = \{x_i,\ i = 1,\ldots,n\}$ представляют собой точку. Будем называть такие выборки *точечными*. Интервальная выборка в пространстве R^n задает $n-$мерный параллелепипед $[a_1,b_1]\times[a_2,b_2]\times\ldots\times[a_n,b_n]$. Будем говорить, что точечная выборка принадлежит интервальной, $X_n \in \mathbf{X}_n$, если $a_i \le x_i \le b_i, i = 1,\ldots,n$.

1.4. Эмпирическая функция распределения и гистограмма

Основную информацию о распределении случайной величины ξ исследователь получает по эмпирической функции распределения и гистограмме, на которые опираются статистические методы анализа.

1.4.1. Интервальная гистограмма

Разобьём область определения случайной величины на k интервалов точками $X_0 < X_1 <\ldots< X_k$ (X_0 – левая граница области определения, X_k – правая граница области определения) и подсчитаем число наблюдений, попавших в каждый интервал $(X_j, X_{j+1}]$, $j = 1,\ldots,k-1$. Если интервальное наблюдение $[a_i, b_i]$ покрывает точку разбиения X_j (т.е. $X_j \in [a_i, b_i]$), то точеченое знаечние можно отнести как к интервалу $(X_{j-1}, X_j]$, так и к интервалу $(X_j, X_{j+1}]$. Таким образом

можно получить 2^p гистограмм, где p – число наблюдений, попавших на границы разбиения. Совокупность всех гистограмм дает нам интервальную гистограмму (см. рис. 1.5).

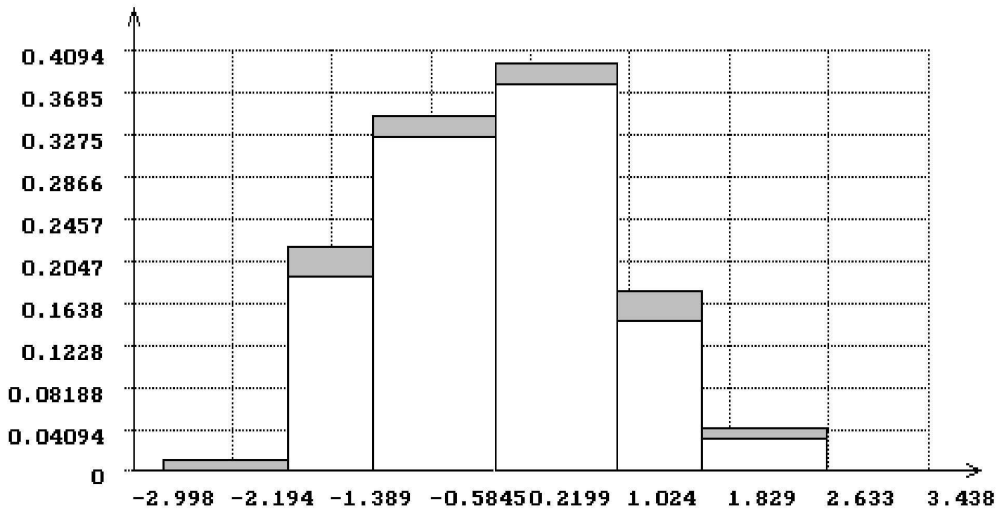


Рис. 1.5. Интервальная гистограмма

1.4.2. Интервальная эмпирическая функция распределения

Упорядочим граничные точки интервалов:

$$a_{(1)} \le a_{(2)} \le \ldots \le a_n \quad \text{и} \quad b_{(1)} \le b_{(2)} \le \ldots \le b_n$$

Предположим, что все точечные наблюдения x_i совпали с левыми границами интервалов. Тогда эмпирическая фуркция распределения будет иметь вид

$$\overline{F_n}(x) = \begin{cases} 0, \ x < a_{(1)} \\ i/n, \ a_{(i)} \le x < a_{(i+1)}, \ i = 1,2,\ldots,n-1 \\ 1, \ x \ge a_{(n)} \end{cases}$$

Аналогично, если все точечные наблюдения x_i совпали с правыми границами интервалов, то эмпирическая фуркция распределения будет иметь вид

$$\underline{F_n}(x) = \begin{cases} 0, \ x < b_{(1)} \\ i/n, \ b_{(i)} \le x < b_{(i+1)}, \ i = 1,2,\ldots,n-1 \\ 1, \ x \ge b_{(n)} \end{cases}$$

Пример. Участникам статистического эксперимента предлагали оценить свой рост и вес сверху и снизу. Были получены следующие интервальные наблюдения (см. таблицу 1.1). Из таблицы хорошо видно, что интервалы неопределенности имеют разную длину, причем чем выше значение наблюдения, тем больше величина погрешности. Соответствующие интервальные эмпирические функции распределения приведены на рис. 1.6 и 1.7.

Таблица 1.1

Оценки роста и веса группы студентов 5-го курса

№	Рост		Вес	
	оценка снизу	оценка сверху	оценка снизу	оценка сверху
1.	179	181	72	78
2.	180	185	60	70
3.	182	187	80	90
4.	179	182	71	75
5.	155	157	46	48
6.	160	165	51	52,5
7.	174	176	80	90
8.	178	184	85	95
9.	177	181	55	65
10.	174	176	55	60

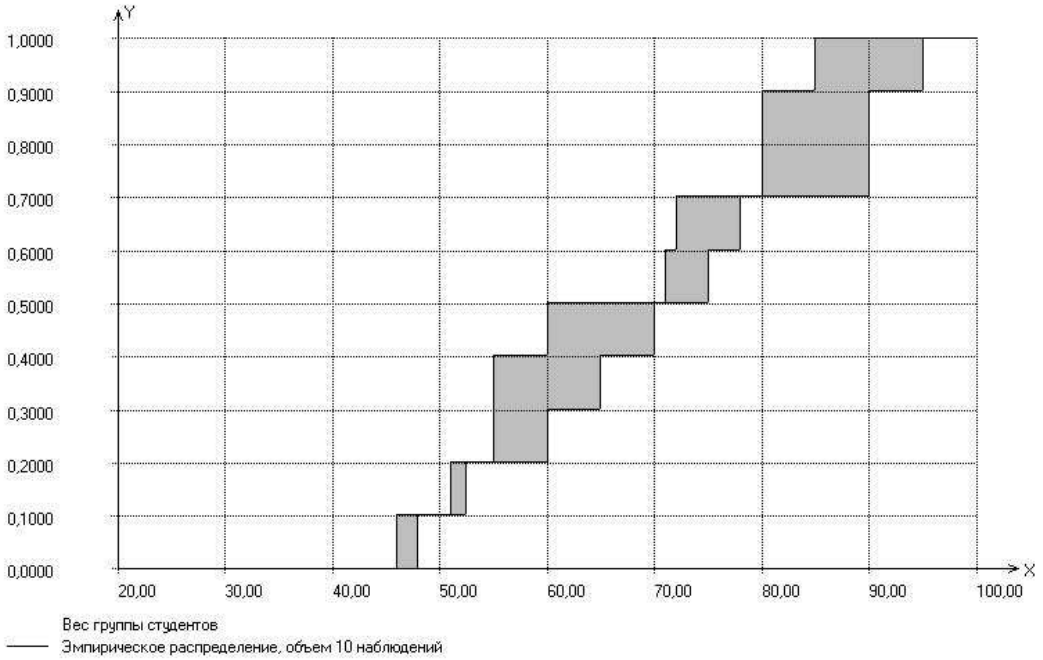


Рис. 1.6. Распределение группы студентов по весу

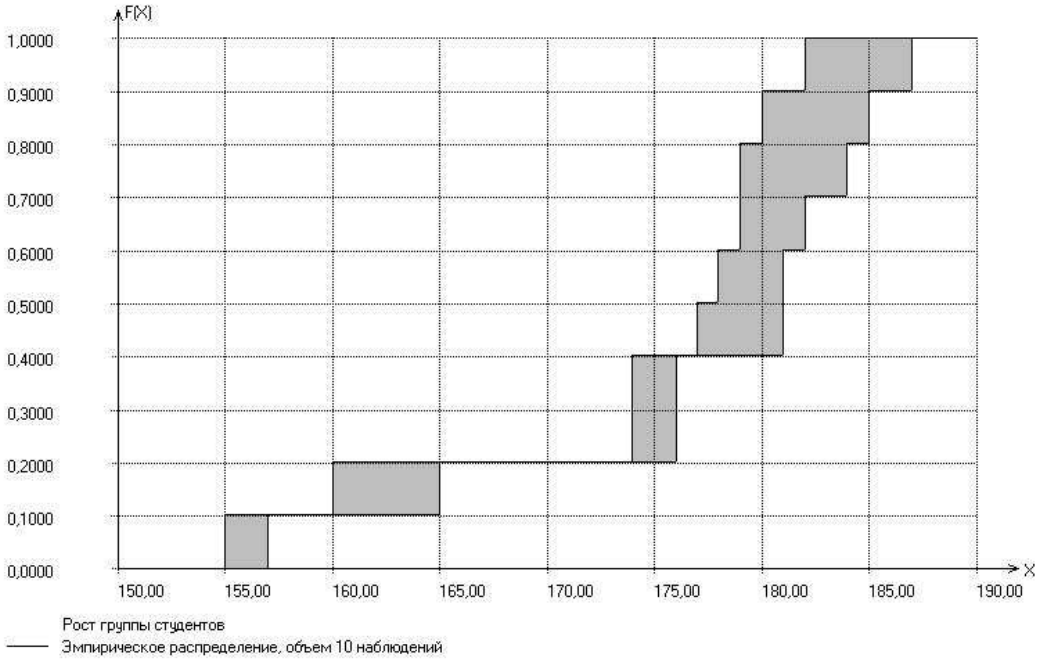


Рис. 1.7. Распределение группы студентов по росту

Упражнения

- Построить интервальную эмпирическую функцию распределения для группированной, цензурированной, частично группированной выборки.
- Каким образом можно моделировать интервальные выборки?

2. Проверка гипотез о согласии

2.1. Проверка гипотез о согласии по точечной выборке

Пусть проверяется гипотеза H_0: $F_\xi(x) = F(x)$ против гипотезы H_1: $F_\xi(x) \neq F(x)$ по заданной точечной выборке $X_n = \{x_i, i = 1, \ldots, n\}$ при уровне значимости α. Процедура проверки гипотезы о согласии заключается в следующем:

1. Формируется статистика $S_n(X_n, F)$.
2. Находится распределение статистики $g(S|H_0)$.
3. Вычисляется *вероятность согласия* (наибольший уровень значимости критерия, при котором гипотеза не отвергается)

$$P\{S > S_n(X_n, F)\} = \int_{S_n(X_n,F)}^{+\infty} g(S|H_0)dS = p(X_n, F)$$

4. Принимается решение о согласии выборки X_n с распределением $F(x)$: если $p(X_n, F) > \alpha$, то гипотеза H_0 не отвергается, если $p(X_n, F) < \alpha$, то гипотеза H_0 отвергается.

Все критерии проверки гипотез о согласии отличаются только видом статистики $S_n(X_n, F)$. Предложив новую статистику можно получить новый критерий согласия. Однако тут есть две сложности.

Во-первых, необходимо найти распределение статистики при верной нулевой гипотезе $g(S|H_0)$. Если это распределение не зависит от вида распределения $F(x)$ (параметра), то критерий называют "свободным от распределения" (непараметрическим). Если же для каждого распределения $F(x)$ (параметра) получается свое распределение, то такой критерий является "зависящим от распределения" (параметрическим).

Во-вторых, критерий должен обладать высокой мощностью, то есть с большой вероятностью отвергать неверную гипотезу.

2.2. Проверка простых гипотез по интервальной выборке

Пусть дана интервальная выборка $\mathbf{X}_n$. Тогда мы можем определить границы для статистики критерия:

$$\underline{S_n}(\mathbf{X}_n, F) = \inf_{X_n \in \mathbf{X}_n} S(X_n, F) \le S_n(\mathbf{X}_n, F) \le \sup_{X_n \in \mathbf{X}_n} S(X_n, F) = \overline{S_n}(\mathbf{X}_n, F)$$

Тогда вероятность согласия будет лежать в интервале $[p_{\min}, p_{\max}]$, где $p_{\min} = \int_{\overline{S_n}(X_n,F)}^{+\infty} g(S|H_0)dS$, $p_{\max} = \int_{\underline{S_n}(X_n,F)}^{+\infty} g(S|H_0)dS$ (см. рис. 2.1)

Рис. 2.1.

Тогда о согласии можно сделать следующие выводы:

- $p_{\max} < \alpha$ – гипотеза H_0 отвергается;
- $p_{\min} > \alpha$ – гипотеза H_0 не отвергается;
- $p_{\min} \le \alpha \le p_{\max}$ – гипотеза H_0 может быть либо отвергнута либо не отвергнута (зона нечувствительности критерия).

В последнем случае возможны разные варианты принятия решения: можно задать степень доверия исследователя к наблюдаемым данным, выполнить процедуру рандомизации... Однако, из полученной в последующих пунктах теоремы об асимптотических свойствах границ статистики Колмогорова по интервальной выборке следует, что для истинной модели и для любой модели, близкой к истинной в пределах погрешности измерений, интервал $[p_{\min}, p_{\max}]$ стремится к интервалу $[0,1]$ с ростом объема выборки. Таким образом, если считать, что $p = p_{\min}$, то рано или поздно истинная гипотеза будет отвергнута, а если считать, что $p = p_{\max}$, то рано или поздно может быть принята любая близкая гипотеза.

2.3. Критерий согласия Колмогорова

Статистика Колмогорова имеет вид:

$$D_n = \sup_x \left| F_n(x) - F(x) \right|$$

Распределение этой статистики было получено Колмогоровым в 1933 г. Найдем верхнюю и нижнюю границу для этой статистики.

$$\underline{F_n}(x) \le F_n(x) \le \overline{F_n}(x)$$

Отсюда

$$\underline{F_n}(x) - F(x) \le F_n(x) - F(x) \le \overline{F_n}(x) - F(x)$$
$$F(x) - \overline{F_n}(x) \le F(x) - F_n(x) \le F(x) - \underline{F_n}(x)$$

Эти неравенства выполняются для всех x, значит они будут выполняться и для супремумов

$$\sup_x \left(\underline{F_n}(x) - F(x) \right) \le \sup_x \left(F_n(x) - F(x) \right) \le \sup_x \left(\overline{F_n}(x) - F(x) \right)$$
$$\sup_x \left(F(x) - \overline{F_n}(x) \right) \le \sup_x \left(F(x) - F_n(x) \right) \le \sup_x \left(F(x) - \underline{F_n}(x) \right)$$

Объединим эти неравенства в одно, и учтем, что статистика не может быть отрицательной:

$$\underline{D_n} = \max\left\{\sup_x\left(\underline{F_n}(x) - F(x)\right), \sup_x\left(F(x) - \overline{F_n}(x)\right), 0\right\} \le$$

$$\le D_n = \max\left\{\sup_x\left(F_n(x) - F(x)\right), \sup_x\left(F(x) - F_n(x)\right)\right\} \le$$

$$\le \overline{D_n} = \max\left\{\sup_x\left(\overline{F_n}(x) - F(x)\right), \sup_x\left(F(x) - \underline{F_n}(x)\right)\right\}$$

2.4 Асимптотические свойства критерия согласия Колмогорова по интервальной выборке

Естественно, что чем меньше длина интервала $\left[p_{\min}, p_{\max}\right]$, тем более определенные выводы можно сделать.

На величину $\Delta p = p_{\max} - p_{\min}$ в случае верной основной гипотезы H_0 влияют:

* диаметр множества $\mathbf{X}_n$ $(d(\mathbf{X}_n) = 0 \Rightarrow \Delta p = 0)$
* закон распределения $F(x)$
* критерий согласия
* количество наблюдений

Теорема Пусть дана последовательность интервальных выборок $\{\mathbf{X}_n, n = 1,2,\ldots\}$ и $\exists \underline{F}(x) \neq \overline{F}(x)$: $\forall \varepsilon > 0$

$$P\left\{\sup_x \left|\underline{F_n}(x) - \underline{F}(x)\right| > \varepsilon\right\} = O(1/n)$$

$$P\left\{\sup_x \left|\overline{F_n}(x) - \overline{F}(x)\right| > \varepsilon\right\} = O(1/n)$$

Тогда, если $\forall x \quad \underline{F}(x) \le F(x) \le \overline{F}(x)$, то $\Delta p \to 1$, иначе $\Delta p \to 0$.

Доказательство. Статистика $S = \dfrac{(6nD_n + 1)^2}{18n}$ при достаточно большом n имеет распределение

$$P\{S > S^*\} = 1 - K\left(\sqrt{S^*/2}\right),$$

где $K(y) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 y^2}$ – функция распределения Колмогорова.

Для оценок границ $\underline{D_n}$ и $\overline{D_n}$ статистики D_n имеем:

$$p_{\min} = 1 - K\left(\frac{6n\overline{D_n} + 1}{6\sqrt{n}}\right), \; p_{\max} = 1 - K\left(\frac{6n\underline{D_n} + 1}{6\sqrt{n}}\right)$$

Тогда $\Delta p \to 1$, если $p_{\max} \to 1, p_{\min} \to 0$; и $\Delta p \to 0$, если $p_{\max} \to 0$.
В свою очередь, $p_{\min} \to 0$, если статистика $\overline{D_n}$ не будет стремиться к нулю; $p_{\max} \to 0$, если статистика $\underline{D_n}$ также не будет стремиться к нулю; и $p_{\max} \to 1$, если $\underline{D_n}$ стремиться к нулю со скоростью $O(1/n)$.

Рассмотрим теперь два случая: 1). $\forall x \;\; \underline{F}(x) \le F(x) \le \overline{F}(x)$ и 2). $\exists x_0 \; F(x_0) \notin \left[\underline{F}(x_0), \overline{F}(x)\right]$.

1). Пусть $\forall x \;\; \underline{F}(x) \le F(x) \le \overline{F}(x)$. Нижняя граница статистики вычисляется по формуле

$$\underline{D_n} = \max\left\{\sup_x\left(\underline{F_n}(x) - F(x)\right), \sup_x\left(F(x) - \overline{F_n}(x)\right), 0\right\}$$

Если неравенство строгое, $\underline{F}(x) < F(x) < \overline{F}(x)$, то первые две величины в фигурных скобках, становятся отрицательными с вероятностью 1 при достаточно большом n, и поэтому максимум будет

равен нулю. Если же $F(x)$ совпадает с $\underline{F}(x)$ или $\overline{F}(x)$, то, сделав соответстующую замену, мы получим, что $\forall \varepsilon > 0\ P\{\underline{D_n} > \varepsilon\} = O(1/n)$. Таким образом мы доказали, что верхняя граница вероятности согласия стремится к единице.

Верхняя граница статистики вычисляется по формуле

$$\overline{D_n} = \max\left\{\sup_x\left(\overline{F_n}(x) - F(x)\right), \sup_x\left(F(x) - \underline{F_n}(x)\right)\right\}$$

Возьмем любую точку x_0, в которой $\overline{F}(x_0) - \underline{F}(x_0) = c > 0$. Тогда $P\{\overline{D_n} > c/2\} \to 1$. Значит $p_{\min} \to 0$ и $\Delta p \to 1$.

2). Пусть x_0 – точка, в которой $F(x) > \overline{F}(x)$ (аналогично можно рассмотреть случай, когда $F(x) < \underline{F}(x)$). Обозначим $d = F(x_0) - \overline{F}(x_0)$.

Тогда

$$P\{\underline{D_n} \ge d/2\} = P\left\{\max\left\{\sup_x\left(\underline{F_n}(x) - F(x)\right), \sup_x\left(F(x) - \overline{F_n}(x)\right), 0\right\} \ge d/2\right\} \ge$$

$$\ge P\{F(x_0) - \overline{F_n}(x_0) \ge d/2\} = P\{F(x_0) - \overline{F}(x_0) + \overline{F}(x_0) - \overline{F_n}(x_0) \ge d/2\} =$$

$$= P\{d + \overline{F}(x_0) - \overline{F_n}(x_0) \ge d/2\} = P\{\overline{F}(x_0) - \overline{F_n}(x_0) \ge -d/2\} =$$

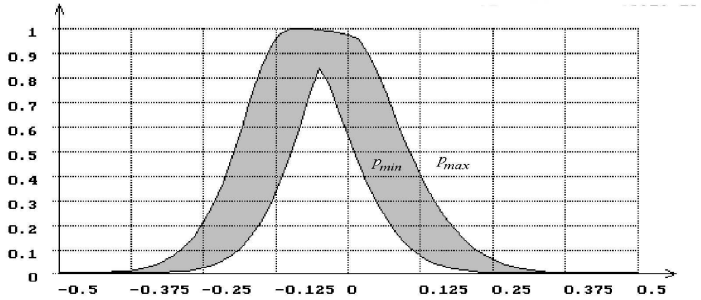
$$=1-P\left\{\overline{F}(x_0)-\overline{F_n}(x_0)<-d/2\right\}\geq 1-P\left\{\sup_x\left|\overline{F}(x)-\overline{F_n}(x)\right|>d/2\right\}=$$

$=1-O(1/n)$.

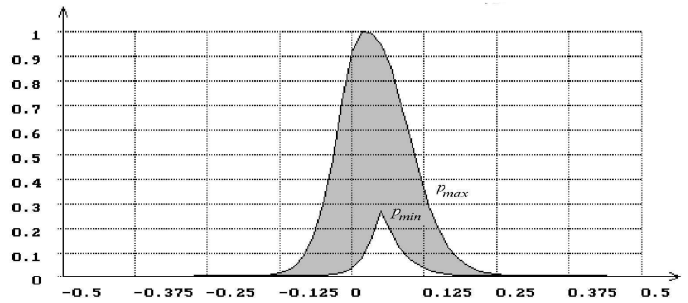
Значит $p_{\max}\to 0$ и $\Delta p\to 0$. Теорема доказана.

Таким образом, с ростом объема выборки зона нечувствительности критерия растёт.

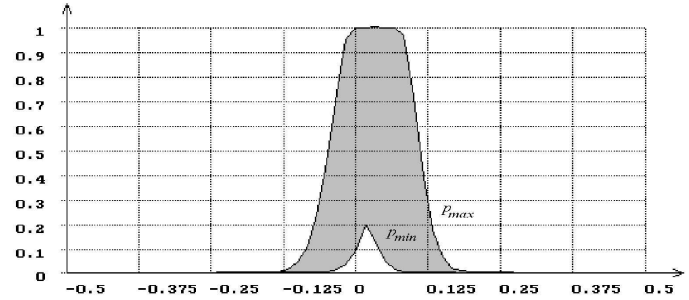
Пример. Рассмотрим случай, когда альтернативная гипотеза имеет то же распределение, что и основная, но с другим параметром. Было смоделировано три интервальных выборки по нормальному закону объемом 100, 500 и 100 наблюдений. На рис. 2.2 показано, какое было бы согласие, если бы мы проверяли согласие с нормальным распределением с параметром сдвига от -0.5 до 0.5.



а.



б.



в.

Рис. 2.2. Согласие интервальных выборок разного объема с нормальным распределением по критерию Колмогорова (а) 100 наблюдений, (б) 500 наблюдений, (в) 1000 наблюдений

Из графиков хорошо видно, что с ростом n расстояние между $p_{\max}$ и $p_{\min}$ увеличивается, график $p_{\max}$ становится более крутым на краях и плоским в середине. Все это подтверждает выводы теоремы.