Тема 1. Повторение теории вероятности.

В нашем курсе мы будем заниматься построением моделей, описывающих функциональные связи между экономическими переменными. Чтобы делать выводы о характере этих взаимосвязей необходимо использовать методы теории вероятности и математической статистики. Поэтому мы начнем с повторения нужных нам разделов этих дисциплин.

Случайные события и их вероятности.

Случайный эксперимент - процесс регистрации наблюдений на единице обследуемой совокупности, который обладает следующими тремя свойствами:

1) существует возможность (теоретическая или практическая) многократно повторить тот же самый эксперимент в тех же условиях

2) из-за большого числа случайных факторов, характеризующих условия проведения каждого эксперимента, невозможно точно предсказать его результат

3) чем большее число однотипных экспериментов мы произведем, тем ближе будут подсчитанные по результатам экспериментов относительные частоты появления интересующего нас события к некоторой постоянной величине, называемой вероятностью этого события.

Классический пример случайного эксперимента - бросание монеты.

Множество всех *взаимоисключающих* исходов случайного эксперимента называться *пространством элементарных событий*. Обозначим его $\Omega = \{\omega_1, \omega_2, ...\omega_N\}$.

Каждому элементарному событию ω_i соответствует вероятность p_i, которая характеризует частоту его появления. Она обладает следующими свойствами: $0 \le p_i \le 1 \forall i$ $\sum_{i:\omega_i \in \Omega} p_i = 1$.

Случайным событием A называют любое подмножество пространства элементарных событий Ω, $A = \{\omega_1^A, \omega_2^A, ...\omega_k^A ..\}$. То есть осуществление любого из элементарных событий ω_j^A ,входящего в A, влечет за собой осуществление события A. Вероятность события A определяется как $P\{A\} = \sum_{\omega_i \in A} P\{w_i\} = \sum_{i:\omega_i \in A} p_i$.

Пример 1.1. Бросается шестигранный кубик. $\Omega = \{1,2,3,4,5,6\}$, $P(\omega_i) = 1/6 \ \forall i$, Пусть событие A состоит в том, что выпадает число, которое меньше или равно 4, значит $A = \{1,2,3,4\}$, $P(A) = 4 \cdot \frac{1}{6} = \frac{2}{3}$.

Рассмотрим некоторые правила действий над случайными событиями и их вероятностями.

Пересечение (произведение) событий A и B-это событие, которое заключается в наступлении обоих этих событий. Обозначается $(A \cap B)$ или просто (AB).

События A и B называться *несовместными*, если $P(A \cap B) = 0$. Очевидно элементарные события ω_i, ω_j по определению являться несовместными для $\forall i \neq j$.

События A и B называться *независимыми*, если:

(1.1) $$P(A \cap B) = P(A) \cdot P(B).$$

Объединение (сумма) событий A и B (обозначается $(A \cup B)$ или $(A + B)$)-это такое событие, которое заключается в наступлении хотя бы одного из двух событий A и B. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Вероятность наступления события A при условии, что произошло событие B, определяется по формуле:

(1.2) $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

В формуле (1.2) предполагается, что вероятность условия $P(B) \neq 0$. В противном случае нет смысла говорить об условной вероятности. Если события A и B независимы, то подставляя (1.1) в (1.2) имеем $P(A|B) = P(A)$.

Полная система событий - это такой набор несовместных событий $(A_1, A_2..., A_k)$, которые в сумме исчерпывают все пространство элементарных событий. То есть

$$\begin{cases} \bigcup\limits_{i=1}^{k} A_i = \Omega \\ A_i \bigcap A_j = \varnothing \quad \forall i \neq j \end{cases}$$

Если $(A_1, A_2..., A_k)$ образуют полную систему событий, тогда справедлива формула полной вероятности

(1.3) $$P(B) = \sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)$$

Используя формулы (1.2) и (1.3) можно выразить $P(A_i|B)$ через $P(B|A_i)$. Из (1.2) имеем:
$P(B \cap A_i) = P(B|A_i) \cdot P(A_i) = P(A_i|B) \cdot P(B) \Rightarrow \quad P(A_i|B) = \dfrac{P(B|A_i) \cdot P(A_i)}{P(B)}$, подставляя $P(B)$ из формулы (1.3), получим

(1.4) $$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum\limits_{j=1}^{k} P(B|A_j) \cdot P(A_j)}$$ - это формула Байеса.

Пример 1.2.

Пусть известны следующие данные о студентах ВШЭ N-го курса.

Таблица 1.1

Место прописки	Факультет	
	Мировая экономика	Гос. управление
Москва	60	70
не Москва	40	30

Для случайно выбранного студента одного из этих факультетов рассчитаем совместные вероятности Р(Город Факультет). Для этого нужно поделить частоты на общее количество студентов:

таблица 1.2

	МЭ	ГУ	Сумма
Москва	0,3	0,35	0,65
Не Москва	0,2	0,15	0,35
Сумма	0,5	0,5	1

В строках и столбцах, которые называются сумма, находится безусловная вероятность, соответственно, факультетов и места жительства. Проверим, являются ли эти признаки независимыми. Для независимых событий должна выполняться формула (1.1), то есть например: $P(Moscow \cap International\,ec.) = P(Moscow) \cdot P(International\,ec.)$. Подставляем цифры 0,3=0,65 0,5-это не верно, значит признаки зависимы.

По формуле (1.2) рассчитаем вероятность факультета при условии, если мы знаем место прописки студента.

таблица 1.3

Условие	Факультет		
	МЭ	ГУ	Сумма
Москва	0,46	0,54	1
Не Москва	0,57	0,43	1

Как мы видим условные вероятности в таблице не сильно отличаются от безусловной 0,5. Аналогично по формуле (1.2) можно сосчитать вероятности места жительства при условии, если мы знаем факультет.

таблица 1.4

Место прописки.	Условие	
	МЭ	ГУ
Москва	0,6	0,7
Не Москва	0,4	0,3
Сумма	1	1

Рассмотрим, как в приложении к нашему случаю будет выглядеть формула Байеса, например: $P(Mos|Int.ec.) = \dfrac{P(Int.ec.|Mos) \cdot P(Mos)}{P(Int.ec.|Mos) \cdot P(Mos) + P(Int.ec.|Not\,Mos) \cdot P(Not\,Mos)}$,

подставим цифры и проверим $0.6 = \dfrac{0.46 \cdot 0.65}{0.46 \cdot 0.65 + 0.57 \cdot 0.35}$ -верно!

Случайные величины. Функция плотности. Функция распределения.

Случайная величина- это функция, определенная на множестве элементарных событий Ω и принимающая значения на множестве действительных чисел R. $X : \Omega \to R$.

Число возможных значений случайной величины X не превышает числа элементарных событий, входящих в множество Ω, но может быть и меньше. Поэтому, если число элементарных исходов конечно, мы получим дискретную случайную величину. Для описания дискретной с. в. достаточно знать вероятности, с которыми она принимает каждое из своих значений. Сумма этих вероятностей должна быть равна единице.

Вспомним **Пример 1.1** (бросание игрального кубика). Множество элементарных событий было задано как $\Omega = \{1,2,3,4,5,6\}$, определим случайную величину $X : x_i = \omega_i$. Соответственно, с. в. X можно описать следующим образом:

таблица 1.5

Значение x	1	2	3	4	5	6
Вероятность	1/6	1/6	1/6	1/6	1/6	1/6

Если у нас есть несколько дискретных случайных величин, они могут описываться совместным распределением вероятностей. Вспомним **Пример 1.2. О**пределим случайную величину $X : x = 1$, если факультет - мировая экономика; $x = 2$, если факультет - гос. управление. И аналогично случайную величину $Y : y = 1$, если место прописки

Москва; $y=2$, если не Москва. Совместное распределение этих случайных величин будет иметь вид: таблица 1.6 Совместное распределение с.в. X,Y.

	$x=1$	$x=2$
$y=1$	0,3	0,35
$y=2$	0,2	0,15

Таблица 1.6 полностью совпадает с таблицей 1.2, просто мы ее по-новому назвали. Из нее мы можем получить одномерные (частные) распределения с. в. X или Y. Для этого нужно сложить вероятности соответственно по строкам или по столбцам:

Таблица 1.7 частное распределение X.

Значения X	$x=1$	$x=2$
Вероятность	0,5	0,5

Таблица 1.8 частное распределение Y

Значения Y	$y=1$	$y=2$
Вероятность	0,65	0,35

Две *случайные величины* называться *независимыми*, если независимы случайные события, которые определяют реализации этих случайных величин. Формально для дискретных случайных величин это можно записать следующим образом. Пусть $x_1, x_2.......x_n$ -все значения, которые принимает случайная величина X, а $y_1, y_2......y_m$ все значения, которые принимает случайная величина Y. Тогда с.в. X и Y являются независимыми, если выполняется равенство:

(1.5) $$P\left(X = x_i \cap Y = y_j\right) = P\left(X = x_i\right) \cdot P\left(Y = y_j\right) \ \forall i = 1...n; \ \forall j = 1....m.$$

Формула (1.5) очень похожа на (1.1), просто вместо вероятностей случайных событий в ней фигурируют вероятности, соответствующие реализации определенных значений случайной величины, что на самом деле одно и тоже.

Рассмотренные нами с.в. X,Y, очевидно, не являться независимыми, так как мы в примере 1.2 установили, что факультет и место жительство не являться независимыми признаками. Для зависимых случайных величин можно построить условные распределения. Для этого нужно, воспользовавшись формулой (1.2), вычислить вероятности, с которыми одна из случайных величин будет принимать все свои значения, если известно значение другой случайной величины. Эти вероятности уже сосчитаны в таблицах 1.3 и 1.4. Я их здесь еще раз воспроизведу, так как эти условные распределения нам потом потребуются.

Таблица 1.9 Условное распределение с.в. X.

Значение X	$x=1$	$x=2$
Вероятности при $y=1$	0,46	0,54
Вероятности при $y=2$	0,57	0,43

Таблица 1.10 Условное распределение с.в. Y.

Значение Y	$y=1$	$y=2$
Вероятности при $x=1$	0,6	0,4
Вероятности при $x=2$	0,7	0,3

Функцией распределения $F(x)$ *случайной величины* X -называют функцию, ставящую в соответствие любому значению x величину вероятности события $\{X \le x\}$, то есть:

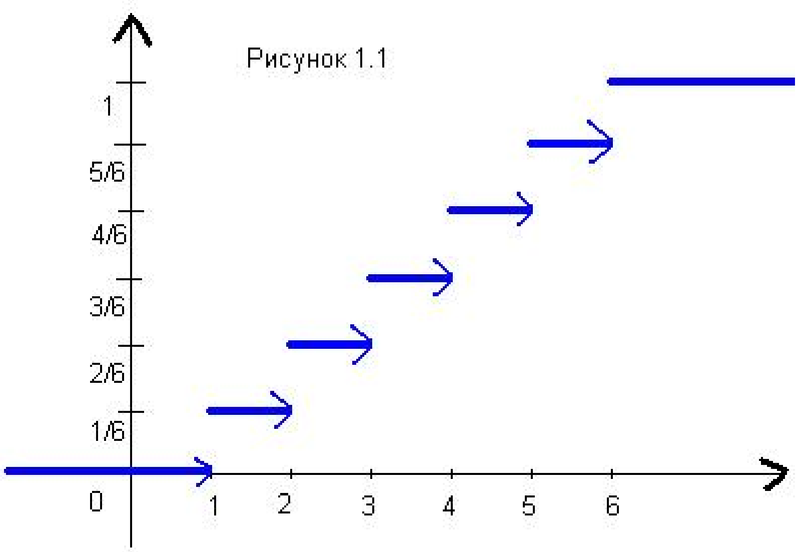
(1.6) $$F(x) = P\{X \le x\}$$

Функция распределения содержит всю информацию о случайной величине и однозначно ее определяет. То есть если у двух случайных величин совпадают функции распределения, значит у них будут совпадать и все остальные характеристики, такие как математическое ожидание, дисперсия и т. д. (что это такое будет сказано ниже).

Обозначим $\{x_1, x_2 ... x_i ... x_n\}$- значения, которые принимает дискретная случайная величина, а $\{p_1, p_2 .... p_i ..... p_n\}$ их вероятности, тогда можно записать функцию распределения дискретной случайной величины следующим образом:

(1.7) $$F(x) = P\{X \le x\} = \sum_{i:x_i \le x} P(X = x_i) = \sum_{i:x_i \le x} p_i$$

Из формулы (1.7) мы видим, что для дискретных случайных величин функция распределения всегда будет неубывающей ступенчатой функцией, состоящей из отрезков и лучей, параллельных оси OX. На рисунке 1.1 изображена функция распределения для случайной величины из Примера 1.1 (то есть для числа, которое выпадет при бросании игрального кубика).



(стрелки означают, что конечные точки не принадлежат отрезку)

Непосредственно из определения *функции распределения* следуют ее *свойства*. Они справедливы как для дискретных, так и для непрерывных случайных величин:

1) $F(x)$- неубывающая функция аргумента x.

2) $P(a < X \le b) = F(b) - F(a)$

3) $\lim_{x \to +\infty} F(x) = 1$ или $F(x) = 1$, для всех $x \ge x_{\max}$

4) $\lim_{x \to -\infty} F(x) = 0$ или $F(x) = 0$, для всех $x \le x_{\min}$ (где $x_{\min}, x_{\max}$- минимальное и максимальное возможные значения исследуемой случайной величины. Надо иметь в виду, что они не всегда существуют.)

Непрерывными называются случайные величины, у которых функция распределения непрерывна. Множество элементарных событий Ω в этом случае должно быть бесконечным. Чтобы строго объяснить, как при бесконечном Ω будет выглядеть функция определенная на множестве элементарных событий Ω и принимающая значения на множестве действительных чисел R необходимо знание специальных разделов теории множеств, которые в вышке не изучают (и правильно делают). В принципе логика там та же, что и для дискретных случайных величин.

Внимание! Вероятность того, что непрерывная случайная величина принимает конкретное числовое значение равна нулю. Это прямо следует из второго свойства функции распределения.

Если функция $F(x)$ непрерывна и имеет производную на всем промежутке изменения случайной величины, то $f(x) = F'(x)$- называется функцией плотности случайной величины.

Свойства функции плотности вытекают непосредственно из свойств функции распределения:

1) $f(x) \ge 0$
2) $\int\limits_{-\infty}^{+\infty} f(x)dx = 1$
3) $\int\limits_{a}^{b} f(x)dx = P\{a \le X \le b\}$
4) $F(x) = \int\limits_{-\infty}^{x} f(u)du$

Пример 1.3 Равномерное распределение (самое простое распределение непрерывной случайной величины). С.в. X называется распределенной равномерно на отрезке $[a,\ b]$, если ее функция плотности на этом отрезке постоянна, а за его пределами равна нулю, то есть

(1.8) $$f(x) = \begin{cases} \dfrac{1}{b-a} & x \in [a,\ b] \\ 0 & x \notin [a,\ b] \end{cases} \text{ соответственно } F(x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{a-b} & x \in [a,\ b) \\ 1 & x \ge b \end{cases}$$

Читателю рекомендуется построить график и проверить аналитически, что все свойства функций распределения и плотности здесь выполняются.

Функция плотности непрерывной случайной величины это в некоторым смысле тоже самое, что вероятность для дискретной случайной величины, поэтому у них есть некоторые общие свойства. Например, для нескольких случайных величин можно определить функцию совместной плотности: $f(x\,y)$. Смысл ее будет следующий: $P(a \le x \le b \cap c \le x \le d) = \int\limits_{a}^{b}\int\limits_{c}^{d} f(x\,y)dxdy$. По аналогии с формулой (1.5) две непрерывные случайные величины считаются независимыми, если выполняется формула:

(1.9) $$f(x,y) = f(x) \cdot f(y)$$

Одномерную (частную) функцию плотности можно из совместной получить следующим образом:

(1.10) $$f(x) = \int\limits_{-\infty}^{+\infty} f(x\,y)dy\,.$$

По аналогии с формулой для условной вероятности (1.2) можно определить условную плотность:

(1.11) $$f(x|y) = \frac{f(x\,y)}{f(y)}$$

Основные числовые характеристики случайных величин (математическое ожидание, дисперсия, ковариация, корреляция).

Математическое ожидание случайной величины X (обозначается $E(X)$) определяется следующим образом:

(1.12) $$E(X) = \int\limits_{-\infty}^{+\infty} x \cdot f(x)dx$$ - для непрерывных случайных величин, (1.13) $$E(X)=\sum_{i=1}^{N} x_i \cdot p_i$$ - для дискретных случайных величин,

где $\{x_1, x_2 ... x_i ... x_N\}$ - значения, которые принимает случайная величина, а $\{p_1, p_2 .... p_i ..... p_N\}$ их вероятности. Иногда математическое ожидание называют средним значением случайной величины, это вполне отражает суть этой характеристики. В формуле (1.12) предполагается абсолютная сходимость интеграла.

Пусть $g(X)$- некоторая непрерывная функция от случайной величины X, тоже, естественно, случайная величина, тогда:

(1.14) $$E(g(X))=\int_{-\infty}^{+\infty} g(x)\cdot f(x)dx$$ - для непрерывных случайных величин,

(1.15) $$E(g(X))=\sum_{i=1}^{N} g(x_i)\cdot p_i$$ - для дискретных случайных величин.

В формуле (1.14) предполагается абсолютная сходимость интеграла.

Пусть X, Y -случайные величины, α -произвольное постоянное число, тогда непосредственно из определения математического ожидания (из выражений (1.12) и (1.13)) вытекают следующие его свойства:

1) $E(\alpha)=\alpha$

2) $E(\alpha \cdot X)=\alpha \cdot E(X)$

3) $E(X+Y)=E(X)+E(Y)$

4) Если случайные величины X, Y независимы, то $E(X \cdot Y)=E(X)\cdot E(Y)$.

Математическое ожиданием случайной величины X при условии, что некоторая другая случайная величина Y принимает известное нам значение y^*, (обозначается $E(X|Y=y^*)$) определяется следующим образом:

(1.16) $$E(X|Y=y^*)=\left.\int_{-\infty}^{+\infty} x\cdot f(x|y)dx\right|_{y=y^*}$$ - для непрерывных случайных величин,

(1.17) $$E(X|Y=y^*)=\sum_{i=1}^{N} x_i \cdot P(x_i|Y=y^*)$$ - для дискретных случайных величин.

Условная плотность $f(x|y)$ в (1.15) определяется по формуле (1.11). Условная вероятность $P(x_i|Y=y^*)$ в (1.16) - по формуле (1.2). Для непрерывной случайной величины необязательно знать конкретное значение условия y^*, условное математическое ожидание может быть определено как функция от y. Для условного математического ожидания выполняются все свойства обычного (безусловного) математического ожидания. Если случайные величины X, Y независимы, то $E(X|Y=y^*)=E(X)$.

Дисперсией случайной величины X (обозначается $D(X)$ или $\operatorname{var}(X)$) называется математическое ожидание от случайной величины: $(X-E(X))^2$. Согласно формулам (1.14) и (1.15) дисперсия вычисляется следующим образом:

(1.18) $$D(X)=\int_{-\infty}^{+\infty}(x-E(X))^2\cdot f(x)dx$$ -для непрерывных случайных величин,

(1.19) $$D(X)=\sum_{i=1}^{N}(x_i-E(X))^2\cdot p_i$$ - для дискретных случайных величин.

Стандартным отклонением случайной величины X (обозначается σ_X) называется корень из дисперсии: $\sigma_X = \sqrt{D(X)}$.

Пусть X, Y -случайные величины, α, β -произвольные постоянные числа, тогда из свойств математического ожидания можно вывести следующие свойства дисперсии:

1) $D(X) = E(X^2) - (E(X))^2$

2) $D(\alpha) = 0$

3) $D(\beta \cdot X) = \beta^2 D(X)$

4) $D(\alpha + \beta \cdot X) = \beta^2 \cdot D(X)$

5) Если случайные величины X, Y независимы, то $D(X+Y) = D(X) + D(Y)$

Дисперсией случайной величины X при условии, что некоторая другая случайная величина Y принимает известное нам значение y^*, (обозначается $D(X|Y=y^*)$) называется условное математическое ожидание $E\left(\left(X - E(X)\right)^2 \middle| Y = y^*\right)$. Вычисляется по формуле (1.16) или (1.17). Для условной дисперсии выполняться те же свойства, что для безусловной.

Ковариацией двух случайных величин X, Y (обозначается $\operatorname{cov}(X,Y)$) называется математическое ожидание случайной величины $\left((X - E(X)) \cdot (Y - E(Y))\right)$. Согласно (1.14) и (1.15) ковариация вычисляется следующим образом:

(1.20) $$\operatorname{cov}(X,Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left((X - E(X)) \cdot (Y - E(Y))\right) \cdot f(x\,y) dxdy$$

-для непрерывных случайных величин,

(1.21) $$\operatorname{cov}(X,Y) = \sum_{i=1}^{N} \sum_{j=1}^{M} \left((x_i - E(X)) \cdot (y_j - E(Y))\right) \cdot P(X = x_i \cap Y = y_j)$$

-для дискретных случайных величин.

Пусть X, Y -случайные величины, α, β -произвольные постоянные числа, тогда из свойств математического ожидания можно вывести следующие свойства ковариации:

1) $\operatorname{cov}(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y)$

2) $\operatorname{cov}(X,X) = D(X)$

3) $D(X+Y) = D(X) + D(Y) + 2 \cdot \operatorname{cov}(X,Y)$

4) $\operatorname{cov}(\alpha \cdot X, \beta \cdot Y) = \alpha \cdot \beta \cdot \operatorname{cov}(X,Y)$

5) $\operatorname{cov}(X, \alpha) = 0$

Ковариация показывает степень связи двух случайных величин. Она зависит от единиц измерения, поэтому удобнее использовать нормированный показатель- коэффициент корреляции (обозначается $cor(X,Y)$ или $r_{X,Y}$) определяется

(1.22) $$cor(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

$cor(X,Y)$- показывает степень тесноты линейной статистической связи случайных величин X, Y. Свойства коэффициента корреляции выводятся из свойств ковариации и дисперсии. Пусть, как обычно, X, Y -случайные величины, α, β -произвольные постоянные числа, тогда

1) $-1 \le cor(X,Y) \le 1$

2) Если $X = \alpha + \beta \cdot Y$, то $cor(X,Y) = \begin{cases} 1 & \beta > 0 \\ -1 & \beta < 0 \end{cases}$

3) $cor(\alpha \cdot X, \beta \cdot Y) = \dfrac{\alpha \cdot \beta}{|\alpha \cdot \beta|} \cdot cor(X,Y)$

4) Если случайные величины X,Y независимы, то $cor(X,Y) = 0$. Обратное утверждение в общем случае неверно. Если распределение с.в. X,Y нормальное, то из $cor(X,Y) = 0$ следует независимость X,Y.

Пример 1.4 Для дискретных случайных величин X,Y, распределение которых приведено в таблицах 1.6-1.10, вычислим все изученные нами числовые характеристики:

$E(X) = 0.5 \cdot 1 + 0.5 \cdot 2 = 1.5 \qquad E(Y) = 0.65 \cdot 1 + 0.35 \cdot 2 = 1.35$

$E(X \cdot Y) = 1 \cdot 1 \cdot 0.3 + 1 \cdot 2 \cdot 0.35 + 2 \cdot 1 \cdot 0.2 + 2 \cdot 2 \cdot 0.15 = 2$

Так как X,Y не являются независимыми $E(X \cdot Y) \neq E(X) \cdot E(Y)$.

$E(X | Y=1) = 1 \cdot 0.46 + 2 \cdot 0.54 = 1.54 \quad E(X | Y=2) = 1 \cdot 0.57 + 2 \cdot 0.43 = 1.43$

$E(Y | X=1) = 1 \cdot 0.6 + 2 \cdot 0.4 = 1.4 \qquad E(Y | X=2) = 1 \cdot 0.7 + 2 \cdot 0.3 = 1.3$

$D(X) = (1-1.5)^2 \cdot 0.5 + (2-1.5)^2 \cdot 0.5 = 0.25 \qquad \sigma_X = 0.5$

$D(Y) = (1-1.35)^2 \cdot 0.65 + (2-1.35)^2 \cdot 0.35 = 0.228 \qquad \sigma_Y = 0.477$

$D(X | Y=1) = (1-1.54)^2 \cdot 0.46 + (2-1.54)^2 \cdot 0.54 = 0.248$

$D(X | Y=2) = (1-1.43)^2 \cdot 0.57 + (2-1.43)^2 \cdot 0.43 = 0.245$

$D(Y | X=1) = (1-1.4)^2 \cdot 0.6 + (2-1.4)^2 \cdot 0.4 = 0.240$

$D(Y | X=2) = (1-1.3)^2 \cdot 0.7 + (2-1.3)^2 \cdot 0.3 = 0.210$

Для вычисления $\operatorname{cov}(X,Y)$ воспользуемся свойством 1, так проще:

$\operatorname{cov}(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y) = 2 - 1.5 \cdot 1.35 = -0.025$

$D(X+Y) = D(X) + D(Y) + 2 \cdot \operatorname{cov}(X,Y) = 0.25 + 0.228 - 2 \cdot 0.025 = 0.428$

$cor(X,Y) = \dfrac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = -0.105$

Пример 1.5. Для случайной величины, равномерно распределенной на отрезке $[a, b]$, вычислим математическое ожидание и дисперсию. Функции распределения и плотности этой величины приведены выше (формула 1.8).

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x) dx = \int_a^b x \cdot \frac{1}{b-a} dx = \left. \frac{x^2}{2(b-a)} \right|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}$$

$$D(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 \cdot f(x) dx = \int_a^b \left(x - \frac{a+b}{2} \right)^2 \cdot \frac{1}{b-a} dx = \int_a^b \left(\frac{x^2}{b-a} - \frac{x \cdot (a+b)}{b-a} + \frac{(a+b)^2}{4 \cdot (b-a)} \right) dx =$$

$$= \left. \left(\frac{x^3}{3 \cdot (b-a)} - \frac{x^2 \cdot (a+b)}{2 \cdot (b-a)} + \frac{x \cdot (a+b)^2}{4 \cdot (b-a)} \right) \right|_b^a = \frac{b^3 - a^3}{3 \cdot (b-a)} - \frac{(a+b)^2}{2} + \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}$$

Квантили и процентные точки. Квантилем уровня q непрерывной случайной величины X с функцией распределения $F(x)$ (обозначается u_q) называется такое значение случайной величины X u_q, для которого выполняется формула:

(1.23) $$F(u_q) = P\{X \le u_q\} = q$$

Под $100 \cdot q\%$-ной точкой случайной величины X понимается значение случайной величины X - ω_q, которое удовлетворяет формуле:

(1.24) $$1 - F(\omega_q) = P\{X \ge \omega_q\} = q$$

Из (1.23) и (1.24) следует, что $u_q = \omega_{(1-q)}$.

Рассмотрим *числовые характеристики многомерных случайных величин*. Пусть θ-вектор размера $(n \times 1)$, состоящий из случайных величин. (Напомню, что при определении размерности матриц сначала указывается число строк потом - столбцов). Математическое ожидание $E(\theta)$ определяется как

(1.25) $$E(\theta) = \begin{pmatrix} E(\theta_1) \\ E(\theta_2) \\ \ldots\ldots \\ E(\theta_n) \end{pmatrix}$$

Ковариационной матрицей θ (обозначается $\mathrm{cov}(\theta)$ или Σ_θ) называется матрица размера $(n \times n)$, которая определяется как

(1.26) $$\Sigma_\theta = E\Big((\theta - E(\theta))(\theta - E(\theta))^T\Big) = \left\langle \begin{matrix} \mathrm{var}(\theta_1) & & \mathrm{cov}(\theta_n\theta_1) \\ & \mathrm{cov}(\theta_i\theta_j)\ \ \mathrm{var}(\theta_i) & \\ \mathrm{cov}(\theta_1\theta_n) & & \mathrm{var}(\theta_n) \end{matrix} \right\rangle$$

Ковариационная матрица обладает следующими свойствами.

1) Симметрична, так как $\mathrm{cov}(\theta_1\theta_n) = \mathrm{cov}(\theta_n\theta_1)$.
2) Неотрицательно определена (то есть для любого вектора $\overline{x}$ выполняется $\overline{x}^T \mathrm{cov}(\theta)\overline{x} \ge 0$)
3) Если компоненты случайного вектора независимы, то его ковариационная матрица будет диагональной. Обратное в общем случае не верно.

Определим случайный вектор η размера $(m \times 1)$ как $\eta = H \cdot \theta + h$, где H-это матрица размера $(m \times n)$, состоящая из неслучайных чисел, h-вектор констант размера $(m \times 1)$. Тогда из свойств математического ожидания и свойств матриц следуют соотношения:

(1.27) $$E(\eta) = H \cdot E(\theta) + h$$

(1.28) $$\Sigma_\eta = H \cdot \Sigma_\theta \cdot H^T$$

Распределения, используемые для проверки гипотез при построении эконометрических моделей.

Нормальное распределение.

Значение исследуемой случайной величины формируется под воздействием очень большого числа независимых случайных факторов, причем сила воздействия каждого отдельного фактора мала и не может превалировать среди остальных, а характер воздействия *аддитивный*. Если это так, получается нормально распределенная случайная величина.

Функция плотности случайной величины X, распределенной нормально с математическим ожиданием μ и дисперсией σ^2 (кратко распределение $N\left(\mu\ \sigma^2\right)$), имеет вид

(1.29) $$\varphi\left(x,\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\cdot\sigma^2}}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

функция распределения соответственно имеет вид

(1.30) $$\Phi\left(x,\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\cdot\sigma^2}}\cdot\int_{-\infty}^{x}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$$

Нормальный закон с параметрами $\mu=0$, $\sigma^2=1$ принято называть стандартным, обозначается $N\left(0;1\right)$.

В Excel для работы с нормальным распределением используются функции:

НОРМРАСП(X, среднее, стандартное_откл, интегральная)

$$\varphi\left(x,\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\cdot\sigma^2}}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}=\text{НОРМРАСП}\left(x,\mu,\sigma,0\right)$$

$$\Phi\left(x,\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\cdot\sigma^2}}\cdot\int_{-\infty}^{x}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt=\text{НОРМРАСП}\left(x,\mu,\sigma,1\right)$$

НОРМОБР(вероятность, среднее, стандартное_откл)

$u_q=\text{НОРМОБР}\left(q,\mu,\sigma\right)$, где u_q квантиль нормального распределения, определяется по формуле (1.23).

Пусть $X_1,X_2...X_i...X_N$ - независимые, нормально распределенные случайные величины с параметрами $N\left(\mu_i\ \ \sigma_i\right)$; $\alpha_0,\alpha_1.......\alpha_n$-действительные числа. Тогда случайная величина $\alpha_0+\alpha_1\cdot X_1+....+\alpha\cdot X_i.....+\alpha_n\cdot X_n$ также является нормальной с математическим ожиданием: $\alpha_0+\alpha_1\cdot\mu_1+....+\alpha\cdot\mu_i.....+\alpha_n\cdot\mu_n$ и дисперсией: $\alpha_1\cdot\sigma_1^2+....+\alpha_i\cdot\sigma_i^2.....+\alpha_n\cdot\sigma_n^2$.

Многомерное нормальное распределение описывает совместное распределение случайного вектора X размера $\left(n\times1\right)$:

(1.31) $$\phi\left(X,M_X,\Sigma_X\right)=\frac{1}{\left(2\pi\right)^{\frac{n}{2}}\sqrt{\det\left(\Sigma_X\right)}}\cdot\exp\left(-\frac{1}{2}\left(X-M_X\right)^T\cdot\Sigma_X^{-1}\cdot\left(X-M_X\right)\right)$$

где M_X,Σ_X - математическое ожидание и ковариационная матрица случайного вектора X определятся по формулам (1.25), (1.26), соответственно $\det\left(\Sigma_X\right)$- определитель ковариационной матрицы. Из формулы (1.31) выводятся свойства нормального распределения.

1) Если $\operatorname{cov}\left(x_ix_j\right)=0$ для $\forall i,j=1....n$, то компоненты вектора X независимы. То есть для нормального распределения из диагональности ковариационной матрицы следует

независимость компонентов случайного вектора. Доказывается это по определению (формула (1.9)).

2) Частное распределения нескольких компонент вектора X (определяются по формуле (1.10)) тоже будет нормальным.

3) Условное распределение любой из компонент вектора X при фиксированных остальных компонентах, тоже будет нормальным. (определяется по формуле (1.11))

Хиквадрат распределение.

Определим случайную величину $\chi^2(m)$ как сумму квадратов m независимых случайных величин, распределенных по стандартному нормальному закону $N(0,1)$. Функция плотности этой величины будет иметь вид

(1.32) $$f_{\chi^2(m)}(x)=\begin{cases}\dfrac{1}{2^{\frac{m}{2}}\Gamma\left(\dfrac{m}{2}\right)}\cdot x^{\frac{m}{2}-1}\cdot e^{-\frac{x}{2}} & x>0\\ 0 & x\le 0\end{cases}$$

m -принято называть числом степеней свободы. $\Gamma(\bullet)$ - это гамма функция

(1.33) $$\Gamma(y)=\int\limits_0^{+\infty} e^{-t}\cdot t^{y-1}dt\,.$$

Функция распределения, как и для нормального распределения, аналитически не выписывается:

(1.34) $$F_{\chi^2(m)}(x)=\begin{cases}\int\limits_0^x f_{\chi^2(m)}(t)\,dt & x>0\\ 0 & x\le 0\end{cases}$$

В Excel для работы с хи-квадрат распределением используются функции:

ХИ2РАСП(Х, степени_свободы)

$F_{\chi^2(m)}(x)=1-\text{ХИ2РАСП}(x,m)$

ХИ2ОБР(вероятность, степени_свободы)

$u_{1-q}\left(\chi^2(m)\right)=\omega_q\left(\chi^2(m)\right)=\text{ХИ2ОБР}(q,\,m)$, где $u_{1-q}\left(\chi^2(m)\right),\ \omega_q\left(\chi^2(m)\right)$ - соответственно, квантиль и процентная точка распределения хи-квадрат с m степенями свободы. Для определения смотри формулы (1.23) и (1.24).

Математическое ожидание и дисперсия $\chi^2(m)$

(1.35) $$E\left(\chi^2(m)\right)=m$$

(1.36) $$\operatorname{var}\left(\chi^2(m)\right)=2\cdot m$$

Распределение Стьюдента с m степенями свободы ($t(m)$-распределение).

Пусть $X_0,X_1,X_2...X_i...X_m$ – независимые случайные величины, распределенные по стандартному нормальному закону $N(0,1)$. Определим случайную величину $t(m)$:

(1.37) $$t(m)=\frac{X_0}{\sqrt{\dfrac{1}{m}\sum\limits_{i=1}^{m}X_i^2}}=\frac{N(0,1)}{\sqrt{\chi^2(m)\big/m}}$$

Функция распределения $t(m)$ имеет вид:

(1.38) $$f_{t(m)}(x)=\frac{1}{\sqrt{\pi\cdot m}}\cdot\frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)}\cdot\left(1+\frac{x^2}{m}\right)^{-\frac{m+1}{2}}$$

где $\Gamma(\bullet)$ - это гамма функция (смотри формулу (1.33)). Функция распределения $t(m)$ аналитически не выписывается, определяется численно как интеграл от (1.38).

(1.39) $$E\left(t(m)\right)=0$$

(1.40) $$D\left(t(m)\right)=\frac{m}{m-2}$$ (существует только при $m>2$)

Хвосты $t(m)$ распределения толще, чем у нормального. При $m\to+\infty$ функция плотности $f_{t(m)}(x)$ стремится к плотности стандартного нормального распределения $\varphi(x)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{x^2}{2}}$.

В Excel для работы с $t(m)$ распределением используются функции:

СТЬЮДРАСП(x, степени_свободы, хвосты)

СТЬЮДРАСП$(x,m,1)=\int_x^{+\infty} f_{t(m)}(u)\,du$

СТЬЮДРАСП$(x,m,2)=\int_{-\infty}^{-x} f_{t(m)}(u)\,du+\int_x^{+\infty} f_{t(m)}(u)\,du=2\cdot\int_x^{+\infty} f_{t(m)}(u)\,du$

СТЬЮДРАСПОБР(вероятность, степени_свободы)

СТЬЮДРАСПОБР$(\alpha,m)=x$ такое, что $\int_{-x}^{x} f_{t(m)}(u)\,du=1-\alpha$.

F-распределение.

Рассмотрим две независимые случайные величины $\chi^2(m_1)$ и $\chi^2(m_2)$, закон распределения которых определяется формулой (1.34). Определим случайную величину

(1.41) $$F(m_1,m_2)=\frac{\chi^2(m_1)/m_1}{\chi^2(m_2)/m_2}$$

Функция плотности этой случайной величины будет иметь вид

(1.42) $$f_{F(m_1 m_2)}(x)=\frac{\Gamma\left(\frac{m_1+m_2}{2}\right)\cdot m_1^{\frac{m_2}{2}}\cdot m_2^{\frac{m_1}{2}}}{\Gamma\left(\frac{m_1}{2}\right)\cdot\Gamma\left(\frac{m_2}{2}\right)}\cdot\frac{x^{\frac{m_1}{2}-1}}{\left(m_1\cdot x+m_2\right)^{\frac{m_1+m_2}{2}}}$$

Функция распределения $F(m_1,m_2)$ аналитически не выписывается, определяется численно как интеграл от (1.42).

Математическое ожидание и дисперсия $F(m_1,m_2)$:

(1.43) $$E\left(F(m_1,m_2)\right)=\frac{m_2}{m_2-2}$$ (существует только при $m_2>2$)

(1.44) $$D\left(F(m_1,m_2)\right)=\frac{2\cdot m_2^2\cdot(m_1+m_2-2)}{m_1\cdot(m_2-2)^2\cdot(m_2-4)}$$ (существует только при $m_2>4$)

Так как $F\left(m_1,m_2\right)=\frac{1}{F\left(m_2,m_1\right)}$, справедлива формула:

(1.46) $$\omega_q\left(F\left(m_1,m_2\right)\right)=\frac{1}{\omega_{1-q}\left(F\left(m_2,m_1\right)\right)}$$

В Excel для работы с $F\left(m_1,m_2\right)$ распределением используются функции:

FРАСП(x, степени_свободы1; степени_свободы2)

$$\int_x^{+\infty} f_{F(m_1 m_2)}\left(u\right)du=\text{FРАСП}\left(x,m_1,m_2\right)$$

FРАСПОБР(вероятность, степени_свободы1; степени_свободы2)

$u_{1-q}\left(F\left(m_1,m_2\right)\right)=\omega_q\left(F\left(m_1,m_2\right)\right)=\text{FРАСПОБР}\left(q,m_1,m_2\right)$,где $u_{1-q}\left(F\left(m_1,m_2\right)\right)$, $\omega_q\left(F\left(m_1,m_2\right)\right)$-соответственно, квантиль и процентная точка $F\left(m_1,m_2\right)$ распределения. Для определения смотри формулы (1.23) и (1.24).

Несколько слов о сходимостях.

Сходимость по вероятности $p\lim$ (convergence in probability).

Последовательность случайных величин $\left\{X_T\right\}_{T=1}^{\infty}$ сходится по вероятности к числу c, если для $\forall\varepsilon>0$ и $\forall\delta>0$ существует N такое, что для $T\geq N$ выполняется неравенство

(1.47) $$P\left\{\left|X_T-c\right|>\delta\right\}<\varepsilon$$

Если неравенство (1.47) выполняется, то c называют пределом по вероятности случайной последовательности $\left\{X_T\right\}_{T=1}^{\infty}$, обозначается $p\lim X_T=c$ или $X_T\xrightarrow{p}c$.

Свойства придела по вероятности:

1) Пусть X_T последовательность случайных чисел, а $g\left(\cdot\right)$- непрерывная, не зависящая от T функция, тогда $p\lim g\left(X_T\right)=g\left(p\lim X_T\right)$.

2) Пусть X_T и Y_T- последовательности случайных чисел, причем $p\lim X_T$ и $p\lim X_T$ существуют, тогда $p\lim\left(X_T\cdot Y_T\right)=p\lim X_T\cdot p\lim Y_T$ и $p\lim\left(\frac{X_T}{Y_T}\right)=\frac{p\lim X_T}{p\lim Y_T}$.

Сходимость в среднем квадратическом (convergence in mean square).

Последовательность случайных величин $\left\{X_T\right\}_{T=1}^{\infty}$ сходиться в среднем квадратическом к числу c, (обозначается $X_T\xrightarrow{m.s.}c$) если для $\forall\varepsilon>0$ существует N такое, что для $T\geq N$ выполняется неравенство

(1.49) $$E\left(X_T-c\right)^2<\varepsilon$$

Или, проще говоря, $\lim\limits_{T\to+\infty}\left(E\left(X_T-c\right)^2\right)=0$

Обобщенное неравенство Чебышева пусть X-случайная величина и $E\left|X\right|^r$- существует и конечно для некоторых $r>0$, тогда для$\forall\delta>0$ и $\forall c$ выполняется

(1.50) $$P\left\{\left|X-c\right|>\delta\right\}\leq\frac{E\left|X-c\right|^r}{\delta^r}$$

В более привычном нам виде неравенство Чебышева выписывается для $r=2,\ c=E\left(X\right)$,

(1.51) $$P\left\{\left|X-E\left(X\right)\right|>\delta\right\}\leq\frac{\operatorname{var}\left(X\right)}{\delta^2}$$

Используя (1.50), можно доказать, что из $X_T\xrightarrow{m.s.}c$ следует $X_T\xrightarrow{p}c$.

Сходимость в распределении (convergence in distribution).

Последовательность случайных величин $\{X_T\}_{T=1}^{\infty}$ сходится в распределении к случайной величине X (обозначается $X_T \xrightarrow{L} X$), если $\lim\limits_{T\to+\infty} F_{X_T}(x) = F_X(x)$, где $F_X(x)$ - функция распределения X, $F_{X_T}(x)$ - функция распределения X_T.

Некоторые свойства:

1) Пусть $\{X_T\}_{T=1}^{\infty}$ и $\{Y_T\}_{T=1}^{\infty}$ последовательности случайных величин, причем $Y_T \xrightarrow{L} Y$ и $(X_T - Y_T) \xrightarrow{P} 0$, тогда $X_T \xrightarrow{L} Y$

2) Пусть $\{X_T\}_{T=1}^{\infty}$ и $\{Y_T\}_{T=1}^{\infty}$ последовательности случайных величин, причем $Y_T \xrightarrow{L} Y$ и $X_T \xrightarrow{P} c$, тогда $X_T + Y_T \xrightarrow{L} Y + c$ и $X_T \cdot Y_T \xrightarrow{L} c \cdot Y$

3) Пусть $\{X_T\}_{T=1}^{\infty}$ последовательность случайных величин, причем $\{X_T\}_{T=1}^{\infty}$ $X_T \xrightarrow{L} X$, а $g(\cdot)$- непрерывная, не зависящая от T функция, тогда последовательность $g(X_T) \xrightarrow{L} g(X)$.

4) Пусть $\{X_T\}_{T=1}^{\infty}$ последовательность случайных величин, причем $\sqrt{T}\cdot(X_T - c) \xrightarrow{L} X$, $g(\cdot)$- непрерывная, не зависящая от T функция, тогда последовательность $\sqrt{T}\cdot\left(g(X_T) - g(c)\right) \xrightarrow{L} g'(c)\cdot X$, где $g'(c) = \left.\dfrac{\partial g(x)}{\partial x}\right|_{x=c}$.

Вы должны знать о двух сходимостях в распределении:

1) Центральная предельная теорема. Пусть случайные величины Y_t-независимы и одинаково распределены для $\forall t$ с математическим ожиданием μ и дисперсией σ^2. $\overline{Y}_T = \dfrac{\sum\limits_{t=1}^{T} Y_t}{T}$, тогда $\sqrt{T}\cdot\left(\overline{Y}_T - \mu\right) \xrightarrow{L} N\left(0,\ \sigma^2\right)$.

2) $t(m) \xrightarrow{L} N(0, 1)$, где $t(m)$ определяется формулой (1.37)

Статистические оценки и их свойства.

Генеральная совокупность (population)- совокупность всех мыслимых наблюдений, которые могли бы быть произведены при данном комплексе условий. (Это чисто математическая абстракция).

Под *выборкой (sample) размера* N из данной генеральной совокупности могут пониматься две вещи

1) Результаты N наблюдений случайных величин, образующих генеральную совокупность, то есть конкретные числа $x_1, x_2, ...x_N$.

2) Последовательность из N случайных величин $X_1, X_2, ...X_i.....X_N$, где X_i-результат, который мог бы быть получен на i-том шаге N кратного эксперимента.

Если случайные величины $X_1, X_2, ...X_i.....X_N$ независимы, то *выборка называется случайной*. Наша задача будет состоять в том, чтобы на основе выборки оценивать параметры генеральной совокупности.

Статистика- любая функция от наблюдений, составляющих выборку, и объема выборки N, не зависящая от параметров генеральной совокупности.

Оценка $\widehat{\theta} = f(x_1 ... x_N)$- статистика, использующаяся для оценки неизвестного параметра генеральной совокупности θ (estimator- формула оценки, estimate- значение оценки после подстановки наблюдений).

Свойства оценок.

1) Несмещенность $E\left(\hat{\theta}\right)=\theta$.(unbiased- несмещенный)

Свойство несмещенности проверяется по определению.

2) Состоятельность $p\lim\hat{\theta}=\theta$, где $p\lim$ определен в (1.47) (consistent- состоятельный)
Есть два способа доказательства состоятельности:
Способ 1. Достаточное условие: если оценка является несмещенной и ее дисперсия стремиться к нулю с ростом выборки, то она является эффективной.
Способ 2. По определению предела по вероятности $p\lim g(n)\cdot\hat{\theta}=\lim g(n)\,p\lim\hat{\theta}$.

Поэтому если мы можем представить оценку, как произведение состоятельной оценки и функции $g(n)$ и при этом $g(n)\longrightarrow 1$, то оценка будет состоятельной.

3) Эффективность $\operatorname{var}\left(\hat{\theta}\right)\le\operatorname{var}\left(\hat{\theta}'\right)$, где $\hat{\theta}'$-любая несмещенная оценка параметра θ. То есть $\hat{\theta}$ обладает наименьшей дисперсией среди несмещенных оценок (best-эффективный).
Эффективность обычно доказывается по теореме Крамера-Рао, которую мы пока не прошли. Либо по определению, но тогда надо выбрать класс функций, в котором мы ищем эффективную оценку.

Пример 1. Выборка из одинаково распределенных независимых случайных величин с параметрами $\left(\mu,\sigma^2\right)$. Рассмотрим различные линейные оценки параметра μ. При анализе их свойств важно понимать, что математическое ожидание суммы равно сумме математических ожиданий, а дисперсия суммы независимых случайных величин равна сумме дисперсий

1) $\bar{X}=\dfrac{\sum_{i=1}^{n}X_i}{n}$

$E\left(\bar{X}\right)=\dfrac{\sum_{i=1}^{n}E\left(X_i\right)}{n}=\dfrac{n\cdot\mu}{n}=\mu$ - оценка несмещенная.

Используем способ 1 доказательства состоятельности, оценка несмещенная и $D\left(\bar{X}\right)=\sum_{i=1}^{n}D\left(\dfrac{X_i}{n}\right)=n\cdot\dfrac{\sigma^2}{n^2}=\dfrac{\sigma^2}{n}\xrightarrow[\infty]{}0$ - значит состоятельная, то есть $p\lim\bar{X}=\mu$.

Докажем, что оценка $\bar{X}$ является эффективной в классе линейных оценок параметра μ.

Пусть $\hat{\mu}=\sum_{i=1}^{n}\alpha_i\cdot X_i$ - произвольная линейная оценка параметра μ. Если мы подберем α_i- таким образом, чтобы оценка $\hat{\mu}$- была несмещенной и ее дисперсия была минимальной, то мы получим эффективную оценку в классе линейных оценок. $E\left(\hat{\mu}\right)=\mu\cdot\sum_{i=1}^{n}\alpha_i=\mu\Rightarrow\sum_{i=1}^{n}\alpha_i=1$, $D\left(\hat{\mu}\right)=\sigma^2\cdot\sum_{i=1}^{n}\alpha_i^2$ Задача имеет вид:

$$\begin{cases}\sum_{i=1}^{n}\alpha_i^2\Rightarrow\min\\ \sum_{i=1}^{n}\alpha_i=1\end{cases}$$

Решая ее методом множителей Лагранжа, получаем $\alpha_i = \frac{1}{n}\ \forall i$, то есть эффективная оценка и есть $\bar{X}$. Решение задачи условной оптимизации единственное, поэтому других линейных эффективных оценок параметра μ не существует.

2) $\hat{\mu}_1 = \frac{X_1 + X_2}{4} + \frac{\sum_{i=3}^{n} X_i}{2\cdot(n-2)}$

$E(\hat{\mu}_1) = \frac{2\cdot\mu}{4} + \frac{\mu\cdot(n-2)}{2\cdot(n-2)} = \mu$ - оценка несмещенная

Проверим состоятельность способом 1 $D(\hat{\mu}_1) = \frac{2\cdot\sigma^2}{16} + \frac{(n-2)\sigma^2}{4\cdot(n-2)^2}$ - к нулю не стремиться значит доказать состоятельность не удается. По определению доказать состоятельность тоже не получается.

Эффективной оценка не будет, так как в классе линейных несмещенных оценок единственная эффективная оценка - $\bar{X}$.

3) $\hat{\mu}_2 = \frac{n\cdot\sum_{i=1}^{n} X_i}{(n-2)^2}$

$E(\hat{\mu}_2) = \frac{n^2\cdot\mu}{(n-2)^2} \neq \mu$ - оценка смещенная.

Попробуем доказать состоятельность по определению (способ 2): $p\lim\hat{\mu}_2 = p\lim\frac{n^2\cdot\bar{X}}{(n-2)^2} = \lim\frac{n^2}{(n-2)^2}\,p\lim\bar{X} = 1\cdot\mu = \mu$ - оценка состоятельна

Эффективной оценка не будет, так как в классе линейных несмещенных оценок единственная эффективная оценка - $\bar{X}$.

Пример 2. Выборка из нормально распределенных независимых случайных величин с параметрами $N(\mu, \sigma^2)$. Рассмотрим различные оценки параметра σ^2.

1) $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$

$E(s^2) = \sigma^2$ - оценка несмещенная для доказательства смотри Шведова.

Проверим состоятельность, используя достаточное условие (способ 1). Оценка несмещенная значит нужно проверить стремиться ли с увеличением выборки дисперсия к нулю. Вспомним, что, так как выборка сделана из нормального распределения $\frac{s^2}{\sigma^2}\cdot(n-1) \approx \chi^2(n-1)$, $s^2 \approx \frac{\sigma^2}{(n-1)}\cdot\chi^2(n-1)$,

$D(s^2) = \frac{\sigma^4}{(n-1)^2}\cdot D(\chi^2(n-1)) = \frac{\sigma^4}{(n-1)^2}\cdot 2\cdot(n-1) = \frac{2\cdot\sigma^4}{(n-1)} \to 0$ -стремиться к нулю при больших n, значит оценка состоятельна, то есть $p\lim s^2 = \sigma^2$.

Про эффективность оценок дисперсии мы пока ничего не знаем.

2) $d^2 = \dfrac{n \cdot \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{\left(n + 30 \right)^2}$

При анализе свойств оценок дисперсии удобно выражать их через оценку с известными нам свойствами s^2. Например в нашем случае $d^2 = \dfrac{n \cdot (n-1)}{(n+30)^2} \cdot s^2$.

$E\left(d^2\right) = \dfrac{n \cdot (n-1)}{(n+30)^2} \cdot \sigma^2 \neq \sigma^2$ - оценка смещенная.

Состоятельность для оценок дисперсии обычно доказываем по определению (способ 2)

$p\lim d^2 = p\lim \dfrac{n \cdot (n-1)}{(n+30)^2} \cdot s^2 = \lim \dfrac{n \cdot (n-1)}{(n+30)^2} \cdot p\lim s^2 = 1 \cdot \sigma^2 = \sigma^2$ - оценка состоятельна.

В тесте задачи будут аналогичные.

Для самопроверки решите:

Выборка из одинаково распределенных независимых случайных величин с параметрами $\left(\mu, \sigma^2\right)$. Рассмотрим различные линейные оценки параметра μ, какие из них будут несмещенными, состоятельными, эффективными?

$\widehat{\mu}_3 = \dfrac{X_1 + X_2 + X_3}{6} + \dfrac{\sum_{i=4}^{n} X_i}{2 \cdot (n-3)}$, $\widehat{\mu}_4 = \dfrac{X_1 + X_2 + X_3}{3}$, $\widehat{\mu}_5 = \dfrac{\sum_{i=2}^{n} X_i}{(n-1)}$ $\widehat{\mu}_6 = \dfrac{\sum_{i=1}^{n} X_i}{n} - \dfrac{1}{n+5}$

$\widehat{\mu}_7 = \dfrac{X_1}{5} + \dfrac{4}{5} \cdot \dfrac{\sum_{i=2}^{n} X_i}{(n-1)}$ $\widehat{\mu}_8 = \dfrac{n \cdot \sum_{i=1}^{n} X_i}{(n+4)^2}$ $\widehat{\mu}_9 = \dfrac{n \cdot \sum_{i=1}^{n} X_i}{(n-20)}$

Выборка из нормально распределенных независимых случайных величин с параметрами $N\left(\mu, \sigma^2\right)$. Рассмотрим различные оценки параметра σ^2, какие из них будут несмещенными, состоятельными?

$d^2 = \dfrac{\ln(n) \cdot \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{(n+30)^2}$ $\delta^2 = \dfrac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{(n-1)} - \dfrac{1}{n+4}$ $s^2 = \dfrac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{(n-1)}$

Проверка гипотез. Доверительные интервалы.

Нам нужно установить свойства генеральной совокупности. Так как у нас есть только выборка, мы можем это сделать только с определенной вероятностью, используя статистическую процедуру проверки гипотез. Большая часть гипотез, которые мы будем проверять в нашем курсе, будет касаться значений числовых параметров функциональных зависимостей, которые существуют между случайными величинами, образующими генеральную совокупность.

Процедура проверки гипотез.

1) Выдвигается нулевая гипотеза H_0 и альтернативная гипотеза H_A.

2) Задается *уровень значимости* α - вероятность отвергнуть H_0, если она на самом деле является справедливой.

3) Строиться критическая статистика $\gamma\left(x_1 x_2 \quad x_N\right)$, которая при выполнении гипотезы H_0 имеет известное нам распределение с функцией распределения $F_\gamma(t)$.

4) По имеющимся у нас наблюдениям вычисляется значение $\gamma\left(x_1 x_2 \quad x_N\right) = \widehat{\gamma}$.

5) В зависимости от вида альтернативной гипотезы H_A мы можем сделать следующие утверждения о значении статистики $\gamma(x_1 x_2 \quad x_N)$ при ее выполнении. (а) $\gamma(H_A) < \gamma(H_0)$ (б) $\gamma(H_A) > \gamma(H_0)$ (с) $\gamma(H_A) \neq \gamma(H_0)$.

6) Гипотеза H_0 не отвергается на уровне значимости α против альтернативной гипотезы H_A, если (рассмотрим 3 случая из предыдущего пункта) (а) $\hat{\gamma} \in [u_\alpha \quad +\infty)$ (б) $\hat{\gamma} \in (-\infty \quad u_{1-\alpha}]$ (с) $\hat{\gamma} \in \left[u_{\frac{\alpha}{2}}, \ u_{1-\frac{\alpha}{2}}\right]$, где u_q-это квантиль распределения $F_\gamma(t)$ определяется как $F_\gamma(u_q) = P\{\gamma \le u_q\} = q$.

7) Если утверждения, приведенные в пункте 6, не выполняются, говорят, что гипотеза H_0 отвергается на уровне значимости α против альтернативной гипотезы H_A.

Несколько понятий связанных с проверкой гипотез:

1) *p-value теста* (как перевести на русский язык науке не известно, иногда говорят - пи-значение)- это такой уровень значимости α, на котором при полученном нами по выборке значении $\hat{\gamma}$, нулевая гипотеза H_0 будет отвергнута против принятой нами H_A.

2) *Мощность теста*- вероятность отвергнуть H_0, если она не верна.

Если мы сравниваем два теста, в которых для проверки одинаковых H_0 и H_A используются разные $\gamma(x_1 x_2 \quad x_N)$ и $F_\gamma(t)$, может оказаться, что при одинаковом α мощность одного теста больше, чем у другого. В этом случае нужно использовать тест с большей мощностью.

3) *Критическое множество*- множество значений $\hat{\gamma}$, при которых H_0 отвергается против принятой нами H_A. Критическое множество зависит от уровня значимости α. В пункте 6 приведены множества обратные к критическим.

Доверительный интервал- это случайная область, построенная по выборке, которая с заданной вероятностью накрывает фиксированное, но не известное нам значение параметра генеральной совокупности.

Построение доверительного интервала возможно, если удается подобрать статистику $\gamma(X_1 X_2 \quad X_N)$, которая обладает следующими свойствами:

1) Распределение $\gamma(X_1 X_2 \quad X_N)$ не зависит от параметра θ и описывается одним из стандартных затабулированных законов распределения.

2) Из того факта, что значения данной статистики заключены в определенных пределах с заданной вероятностью, можно сделать вывод, что неизвестный параметр θ должен лежать между некоторыми границами с той же самой вероятностью.

Примеры проверки гипотез и построения доверительных интервалов.

Проверка гипотезы о среднем значении нормально распределенной случайной величины с известной дисперсией.

Выборка размера N из нормально распределенных случайных величины с известной дисперсией σ^2. $H_0: \ \mu = \mu'$, $\gamma(x_1 x_2 \quad x_N) = \dfrac{\bar{X}_N - \mu'}{\sqrt{\sigma^2/N}}$, где $\bar{X}_T = \dfrac{\sum_{i=1}^{N} X_i}{N}$ при H_0 $\gamma(x_1 x_2 \quad x_N)$, распределена $N(0,1)$. При проверки $\hat{\gamma}$ нужно будет сверять с областями $\hat{\gamma} \in \left[u_\alpha\left(N(0,1)\right) \quad +\infty\right)$, если $H_A: \mu < \mu'$; $\hat{\gamma} \in \left(-\infty \quad u_{1-\alpha}\left(N(0,1)\right)\right]$, если $H_A: \mu > \mu'$

$\hat{\gamma} \in \left[u_{\frac{\alpha}{2}}\left(N(0,1)\right),\ \ u_{1-\frac{\alpha}{2}}\left(N(0,1)\right) \right]$, если $H_A: \mu \neq \mu'$. Если $\hat{\gamma}$ в эти области попадает, гипотеза H_0 не отвергается против соответствующей H_A.

Доверительный интервал для μ:

$$\bar{X}_T + u_{\alpha/2}\left(N(0,1)\right) \cdot \sqrt{\sigma^2 / N} \le \mu \le \bar{X}_T + u_{1-\frac{\alpha}{2}}\left(N(0,1)\right) \cdot \sqrt{\sigma^2 / N}$$

$$\bar{X}_T - u_{1-\frac{\alpha}{2}}\left(N(0,1)\right) \cdot \sqrt{\sigma^2 / N} \le \mu \le \bar{X}_T + u_{1-\frac{\alpha}{2}}\left(N(0,1)\right) \cdot \sqrt{\sigma^2 / N}$$

где $u_{\alpha/2}\left(N(0,1)\right)$-квантиль стандартного нормального распределения.

Проверка гипотезы о среднем значении нормально распределенной случайной величины с неизвестной дисперсией.

Выборка размера N из нормально распределенных случайных величин с неизвестной дисперсией $H_0:\ \mu = \mu'$ $\gamma\left(x_1\, x_2\ \ x_N\right) = \dfrac{\bar{X}_N - \mu'}{\sqrt{s^2 / N}}$ где $\bar{X}_T = \dfrac{\sum_{i=1}^{N} X_i}{N}$, $s^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}_N\right)^2$, распределена $t(N-1)$. При проверке $\hat{\gamma}$ нужно будет сверять с областями $\hat{\gamma} \in \left[u_\alpha\left(t(N-1)\right)\ \ +\infty\right)$, если $H_A: \mu < \mu'$; $\hat{\gamma} \in \left(-\infty\ \ u_{1-\alpha}\left(t(N-1)\right)\right]$, если $H_A: \mu > \mu'$ $\hat{\gamma} \in \left[u_{\frac{\alpha}{2}}\left(t(N-1)\right),\ \ u_{1-\frac{\alpha}{2}}\left(t(N-1)\right) \right]$, если $H_A: \mu \neq \mu'$. Если $\hat{\gamma}$ в эти области попадает, гипотеза H_0 не отвергается против соответствующей H_A.

Доверительный интервал для μ:

$$\bar{X}_T - u_{1-\frac{\alpha}{2}}\left(t(N-1)\right) \cdot \sqrt{s^2 / N} \le \mu \le \bar{X}_T + u_{1-\frac{\alpha}{2}}\left(t(n-1)\right) \cdot \sqrt{s^2 / N}$$

Проверка гипотезы о дисперсии нормально распределенной величины.

Выборка размера N из нормально распределенных случайных величины с неизвестной дисперсией $H_0:\ \sigma^2 = \delta^2$ $\gamma\left(x_1\, x_2\ \ x_N\right) = \dfrac{\sum_{i=1}^{N}\left(X_i - \bar{X}_N\right)^2}{\delta^2}$ -при H_0 распределена $\chi^2(N-1)$. При проверки $\hat{\gamma}$ нужно будет сверять с областями $\hat{\gamma} \in \left[0\ \ u_{1-\alpha}\left(\chi^2(N-1)\right)\right]$, если $H_A: \sigma^2 > \delta^2$; $\hat{\gamma} \in \left[u_\alpha\left(\chi^2(N-1)\right)\ \ +\infty\right)$, если $H_A: \sigma^2 < \delta^2$ $\hat{\gamma} \in \left[u_{\frac{\alpha}{2}}\left(\chi^2(N-1)\right),\ \ u_{1-\frac{\alpha}{2}}\left(\chi^2(N-1)\right) \right]$, если $H_A: \sigma^2 \neq \delta^2$. Если $\hat{\gamma}$ в эти области попадает, гипотеза H_0 не отвергается против соответствующей H_A. Доверительный интервал для дисперсии σ^2:

$$\frac{\sum_{i=1}^{N}\left(X_i - \bar{X}_N\right)^2}{u_{1-\frac{\alpha}{2}}\left(\chi^2(N-1)\right)} \le \sigma^2 \le \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}_N\right)^2}{u_{\alpha/2}\left(\chi^2(N-1)\right)}$$

Проверка гипотезы о равенстве дисперсий двух нормально распределенных величин.

Пусть есть независимые выборки из двух нормально распределенных генеральных совокупностей $X_1 X_2 ....X_n$ и $Y_1, Y_2 ....Y_m$ с параметрами $\left(\mu_X \ \sigma_X^2\right)$ и $\left(\mu_Y \ \sigma_Y^2\right)$. Обозначим $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \quad s_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{Y}_m\right)^2$. Пусть $s_X^2 > s_Y^2$, тогда проверяется $H_0 : \sigma_Y^2 = \sigma_X^2$ против $H_A : \sigma_Y^2 < \sigma_X^2$ статистика $\gamma\left(x_1..x_n \ y_1..y_m\right) = \frac{s_X^2}{s_Y^2}$ при H_0 имеет распределение $F\left(\left(n-1\right)\left(m-1\right)\right)$, $\widehat{\gamma}$ нужно будет сверять с областью $\widehat{\gamma} \in \left[0 \quad u_{1-\alpha}\left(F\left(\left(n-1\right)\left(m-1\right)\right)\right)\right]$. Если $\widehat{\gamma}$ в эту области попадает, гипотеза H_0 не отвергается против соответствующей H_A.

Использованная литература.

При написании этой главы я не пытался оправдать свою фамилию, поэтому пользовался вспомогательной литературой. Из книг мне пригодились 1 том «Теории вероятности и прикладной статистики» С. А. Айвазяна и «Теория вероятности и математическая статистика» А. С. Шведова. Материал про сходимости я взял из прекрасной книжки «Time Series Analysis» James Hamilton. Еще заглядывал в конспекты лекций Канторовича Г. Г. и семинаров Мамонтова А. А. по курсу эконометрика, который шел на факультет экономика в 2000-2001 году. В своих конспектах я, правда, ничего не понял, поэтому пришлось взять у Анны Бланк. Я также благодарен Елене Владимировне Коссовой за замечания по тексту главы. Никто из перечисленных выше лиц не является ответственным рецензентом и не несет ответственности за ошибки автора.

Платон.
platonhse@mail.ru