Тема 2

Повторение
теории вероятностей
и математической
статистики

1. Обзор основных понятий теории вероятностей

Ключевые понятия теории вероятностей

Понятия	***Пример***
случайный эксперимент	*сплошная перепись населения района*
случайная величина	*возраст респондента*
генеральная совокупность	*все полученные данные о возрастах*
случайное событие	*A={возраст респондента < 30 лет}*
вероятность события	P{A}=0.25 (если четверть опрошенных удовлетворяет этому критерию)

Понятие случайной величины

- Определена как действительная функция случайного события $X=f(\omega)$.
- Под этим термином, если отвлечься от строгих математических формулировок, понимается величина, которая в результате стечения обстоятельств (случайных событий) принимает то или иное заранее неизвестное значение. Каждое значение появляется с некой вероятностью:
- $P\{X=30$ лет$\} = m/n$
  - m-число 30-летних респондентов
  - n-общее число респондентов
  - Это – частотное определение вероятности

Вероятность

- Действительное число

$$0 \le P(A) \le 1$$

- Если $P(A) = 0$ - событие А называется невозможным

$$P(возраст = -10) = 0$$

- Если $P(A) = 1$ - событие А называется достоверным

$$P(0 \le возраст \le \infty) = 1$$

Вероятность практически невозможного события

Очень полезное понятие:

- пороговый уровень вероятности практически невозможного события
- устанавливается в зависимости от контекста
- как правило, используются значения вероятности

  0.1, 0.05, 0.01 или

  10% 5% 1%

Функция распределения вероятностей

- Наиболее полно случайная величина характеризуется своей ***функцией распределения вероятности***

  $$F(x)=P(X<x)$$

Вид этой функции определяется типом случайной величины, которая может быть:

- дискретной
- непрерывной
- смешанной

Наиболее популярные в эконометрике законы распределения вероятностей

- Непрерывные
  - нормальное
  - логнормальное
  - t-распределение, F-распределение,
  - хи-квадрат распределение
- Дискретные
  - бинарное
  - Пуассоновское

Дискретная СВ

- ***Дискретные распределения*** описывают СВ, принимающие дискретный набор значений – обычно несколько целых чисел. В этом случае мы можем перечислить все возможные значения и соответствующие им вероятности.

- Пример. *Продавец делает 6 телефонных звонков, каждый из которых заканчивается либо успехом, либо неудачей. Число успехов – СВ, которая принимает в данном случае значения 0, 1, 2, 3, 4, 5, 6. Поэтому она имеет дискретное распределение, ставящее в соответствие каждому из этих семи чисел их вероятности. Если продавец не реагирует на успех или неудачу, это распределение вероятностей может быть биномиальным.*

Биномиальное распределение

$$P_n(k) = C_n^k p^k (1-p)^{n-k}$$

- где p – вероятность успеха в отдельном испытании, k – число успехов, n – число испытаний, C_n^k - число сочетаний из n элементов по k, и его предельный случай для процессов, где вероятность успеха мала – ***распределение Пуассона***

Распределение Пуассона

$$P_n(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

- где $\lambda = np$.
- Оба распределения имеют большую область приложения на практике. используются, например, в теории массового обслуживания (вероятность, что из n покупателей k купят товар)

Непрерывная СВ

- Если СВ может принимать любое значение из некоторого промежутка (a, b), то она называется ***непрерывной*** и кроме функции распределения обладает ***функцией плотности*** f(x)=F'(x).
- Зная функцию плотности, мы можем вычислять вероятности всевозможных событий вида: $P(a < X < b) = \int_a^b f(x)dx$

Нормальное распределение

- Пример. *Вес коробки с кофе, сходящей с конвейерной линии. Он, скорее всего, будет подчиняться нормальному закону распределения вероятностей.*
- ***Нормальное распределение*** является наиболее важным с теоретической точки зрения и одновременно наиболее распространенным на практике.

Нормальное распределение

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

- где μ и σ - параметры распределения такие, что μ=M(x) – *математическое ожидание* СВ X, которое является неким аналогом среднего значения СВ, а σ2=D(x) – *дисперсия* СВ X, которая характеризует средний разброс СВ вокруг математического ожидания.

2. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СЛУЧАЙНЫХ ВЕЛИЧИН

Математическое ожидание и ***дисперсия***

- ***Математическое ожидание*** и ***дисперсия*** являются важнейшими числовыми характеристиками СВ.
- Их определения для дискретных СВ:

$$E(X)=\sum_{i=1}^{n}x_i p(x_i), \qquad V(X)=E(X-E(X))^2=\sum_{i=1}^{n}(x_i-E(x))^2 p(x_i)$$

- Для непрерывных СВ они определяются следующим образом:

$$E(X)=\int_{-\infty}^{\infty}xf(x)dx, \qquad V(X)=E(X-E(X))^2=\int_{-\infty}^{\infty}(x-E(X))^2 f(x)dx$$

Асимметрия и эксцесс

- Несимметричность распределения принято характеризовать ***асимметрией***
- Асимметрия = $\dfrac{E(X - E(X))^3}{(V(X))^{3/2}}$
- (для нормального распределения этот показатель равен нулю).
- Степень выраженности «хвостов», т.е. частоту появления удаленных от среднего значений, характеризует ***эксцесс***
- Эксцесс = $\dfrac{E(X - E(X))^4}{(V(X))^2}$ - 3
- для нормального распределения этот показатель равен 0

3. Статистическое оценивание и проверка гипотез

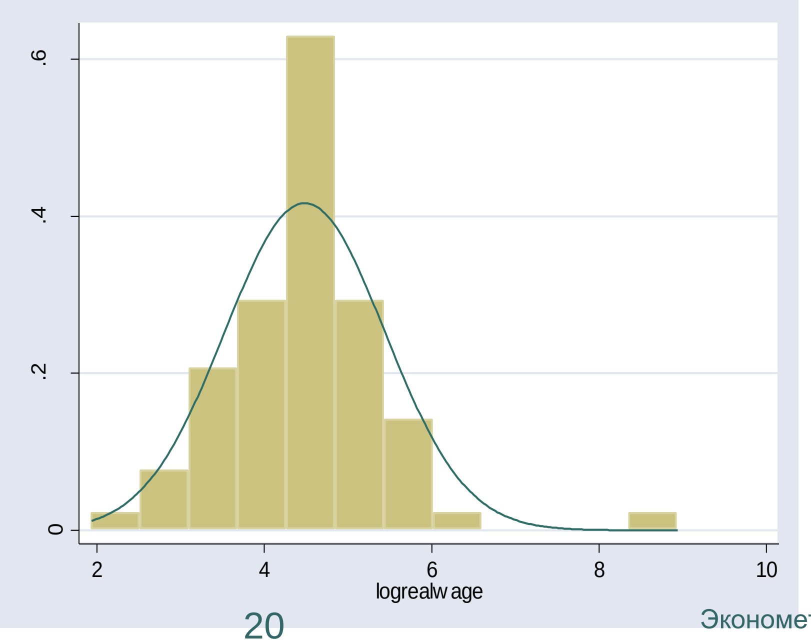
Гистограмма – эмпирический аналог закона распределения

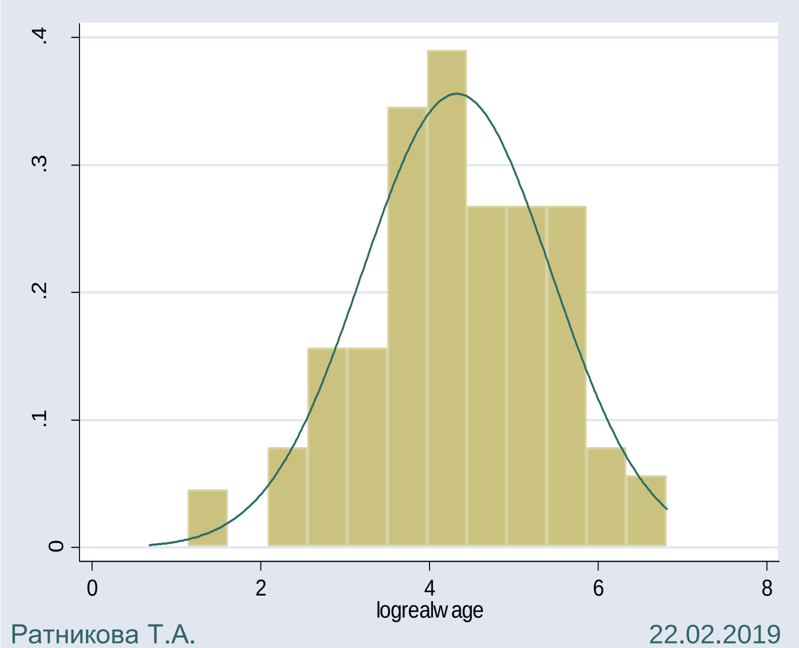
Принцип построения гистограммы:

- интервал наблюдений разбивается на m интервалов, называемых интервалами группировки.
- Графическое изображение зависимости частоты попадания элементов выборки в интервал группировки от соответствующего интервала группировки называется ***гистограммой выборки***.

<u>Пример</u>: *гистограммы логарифма заработной платы*

жителей Москвы и жителей республики Коми на 2000 год по данным РМЭЗ





Цель построения гистограммы

- Цель построения гистограммы – понять можно ли считать распределение СВ из выборки близким к нормальному, т.к. это бывает важно знать при проверке гипотез и построении доверительных интервалов.

Оценивание параметров генеральной совокупности на основании выборки

- Обо всей генеральной совокупности мы, как правило, ничего не знаем точно и можем строить лишь догадки - ***гипотезы***. Для проверки своих гипотез мы исследуем ***независимую выборку*** из генеральной совокупности и строим на основании выборки ***выборочные оценки*** неизвестных теоретических параметров.
- Различают ***точечные*** и ***интервальные оценки.***

Примеры точечных оценок

Выборочное среднее (является оценкой для E(X)):

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Выборочная оценка для V(x):

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

Закон больших чисел

- На основании приведенных выше оценок судят об истинных значениях параметров генеральной совокупности.
- ***Закон больших чисел*** утверждает, что для больших выборок приведенные выше оценки становятся очень близкими к оцениваемым параметрам.

Методы получения оценок

- Существуют специальные методы получения оценок, например, метод моментов и метод максимального правдоподобия.
- Суть метода моментов состоит в приравнивании теоретических и выборочных моментов распределения и последующем выражении неизвестных теоретических параметров через наблюдаемые величины.
- Суть метода максимального правдоподобия - в отыскании значений неизвестных теоретических параметров, при которых совместная функция распределения (или функция плотности) выборочных СВ достигает максимума.

Свойства оценок

- Точечные оценки считаются «хорошими», если они обладают определенными свойствами:
- ***несмещенностью*** (в этом случае математическое ожидание оценки совпадает с оцениваемым теоретическим параметром);
- ***состоятельностью*** (это означает, что для больших выборок вероятность значимых отклонений величины оценки от значения оцениваемого теоретического параметра равна нулю);
- ***эффективностью*** (чем меньше дисперсия оценки, тем она считается эффективнее).
- Исследование свойств оценок – это отдельная теоретическая задача.

Доверительные интервалы

- Интервальные оценки строятся на основании точечных оценок и ***доверительной вероятности***, которая позволяет судить, на сколько мы можем быть уверены, что построенный интервал будет содержать в себе неизвестный теоретический параметр.

Примеры интервальных оценок

- ***доверительный интервал для математического ожидания***, который будет содержать в себе неизвестное математическое ожидание с вероятностью $1-\alpha$

$$\bar{x} - t_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}} < E(X) < \bar{x} + t_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}$$

Примеры интервальных оценок

- ***доверительный интервал для дисперсии***

$$\frac{(n-1)\hat{\sigma}^2}{\chi^2_{\alpha/2}} < V(X) < \frac{(n-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}$$

Примеры интервальных оценок

- Здесь n – размер выборки,
- $t_{\alpha/2}$, $\chi^2_{\alpha/2}$, $\chi^2_{1-\alpha/2}$
  - квантили соответствующих распределений с n-1 степенью свободы

Проверка статистических гипотез

- Теоретические основы для проверки гипотез предоставляет ***центральная предельная теорема***:
- Пусть $x_1, \ldots, x_n$ – независимая выборка из генеральной совокупности с произвольным законом распределения, характеризующимся параметрами

$$E(X) = \mu, \quad V(X) = \sigma^2$$

- Тогда при достаточно больших значениях n
- выборочное среднее $\bar{x} \sim N(\mu, \sigma^2 / n)$,
- а

$$z = \frac{\bar{x} - \mu}{\sigma^2 / n} \sim N(0,1)$$

Проверка гипотез

- Помимо стандартного нормального закона для проверки гипотез используется еще ряд производных от нормального распределений:
- χ^2 - распределение Пирсона,
- t - распределение Стьюдента,
- F-распределение Фишера и др.

Несколько слов о происхождении тестовых статистик

- если $z_1, \ldots, z_k \sim N(0,1)$ и они независимы,
- то $$\sum_{i=1}^{k} z_i^2 = \chi^2(k)$$
- если $Z \sim N(0,1), \quad Y \sim \chi^2(k)$ и они независимы,

  то $$t = \frac{Z}{\sqrt{Y/k}} \sim t(k)$$
- если $Y_1 \sim \chi^2(k_1), Y_2 \sim \chi^2(k_2)$ и они независимы,

  то $$F = \frac{Y_1/k_1}{Y_2/k_2} \sim F(k_1, k_2)$$

Таблица 1. Наиболее употребительные параметрические критерии проверки статистических гипотез

- См. файл Проверка гипотез.doc

Пример

- В фирме работают два менеджера, управляющие проектами
- В прошлом году менеджер А управлял 11 проектами и получил по ним следующие значения доходности:
- -1,74 -4,21 4,59 3,56 1,71 0,69 -2,09
- -1,16 4,05 1,86 -0,88
- Менеджер Б управлял 15 проектами и получил такие результаты:
- -1,05 4,97 -2,04 3,59 -0,21 0,45 -0,24
- -1,04 4,04 4,39 0,45 -0,49 2,52 3,74 4,02

Пример

- Начальник отдела хочет уволить менеджера А по личным причинам. Он рассматривает следующую возможную формулировку причины увольнения: «Проекты, руководимые менеджером А, давали в среднем меньшую доходность, чем проекты менеджера Б».
- Оправдана ли эта формулировка и нижеследующие?
- «Проекты, руководимые менеджером А, давали в среднем доходность, меньшую, чем плановая (по статье 38 Устава плановая доходность составляет 1.5»
- «Проекты, руководимые менеджером А, более рискованные (в терминах дисперсии доходности), чем проекты менеджера Б»
- «Проекты, руководимые менеджером А, более рискованные, чем предполагается Уставом (по статье 39 Устава максимально приемлемый риск равен 4)».

Решение

Результаты двух выборочного t-теста с одинаковыми дисперсиями

Выборки	А	Б
Среднее	0,58	1,54
Выборочная дисперсия	8,05	5,80
Число наблюдений	11	15
Объединенная дисперсия		6,74
Гипотетическая разность средних		**0**
df		24
t-статистика		-0,93
P(T<=t) одностороннее		0,18
t критическое одностороннее		1,71
P(T<=t) двухстороннее		0,36
t критическое двухстороннее		2,06

Решение

- Поскольку абсолютное значение выборочной t-статистики равно 0,93, что меньше и t критического одностороннего (1,71), и t критического двухстороннего (2,06), то у нас нет статистических оснований отвергать основную гипотезу о равенстве доходностей, обеспечиваемых обоими менеджерами.
- Таким образом, предлагаемая формулировка причины увольнения не имеет под собой достаточных оснований.