\documentclass[aspectratio=169, usenames, dvipsnames, handout]{beamer}

%\documentclass[handout, aspectratio=169]{beamer}

% Russian Language

%--------------------------------------

\usepackage[T2A]{fontenc}

\usepackage[utf8]{inputenc}

\usepackage[russian]{babel}

%--------------------------------------

% Hyphenation rules

%--------------------------------------

\usepackage{hyphenat}

%--------------------------------------

% Plots

%--------------------------------------

\usepackage{pgfplots}

\pgfplotsset{width=8.5cm, compat=1.9}

%--------------------------------------

% colors: https://en.wikibooks.org/wiki/LaTeX/Colors

% Arrows

%--------------------------------------

\usepackage{tikz}

\usetikzlibrary{shapes.arrows}

%--------------------------------------

\tikzset{

myarrow/.style={

draw,

fill=MidnightBlue,

single arrow,

minimum height=3.5ex,

single arrow head extend=1ex

}

}

\newcommand{\arrowup}{%

\tikz [baseline=-0.5ex]{\node [myarrow,rotate=90] {};}

}

\newcommand{\arrowdown}{%

\tikz [baseline=-1ex]{\node [myarrow,rotate=-90] {};}

}

%--------------------------------------

% Add theme

\usetheme{Madrid}

\colorlet{beamer@blendedblue}{MidnightBlue!40!black}

% Theme settings

\setbeamertemplate{navigation symbols}{}

\makeatletter

\setbeamertemplate{footline}{

\leavevmode%

\hbox{%

\begin{beamercolorbox}[wd=.4\paperwidth,ht=2.25ex,dp=1ex,center]{author in head/foot}%

\usebeamerfont{author in head/foot}\insertshortauthor\expandafter

\end{beamercolorbox}%

\begin{beamercolorbox}[wd=.3\paperwidth,ht=2.25ex,dp=1ex,center]{title in head/foot}%

\usebeamerfont{title in head/foot}\insertshorttitle

\end{beamercolorbox}%

\begin{beamercolorbox}[wd=.3\paperwidth,ht=2.25ex,dp=1ex,right]{date in head/foot}%

\usebeamerfont{date in head/foot}\insertshortdate{}\hspace\*{2em}

\insertframenumber{} / \inserttotalframenumber\hspace\*{2ex}

\end{beamercolorbox}}%

\vskip0pt%

}

\makeatother

\makeatletter

\g@addto@macro\normalsize{%

\setlength\belowdisplayskip{-0pt}

}

% Integer part of the number

\usepackage{mathtools}

\DeclarePairedDelimiter\ceil{\lceil}{\rceil}

\DeclarePairedDelimiter\floor{\lfloor}{\rfloor}

% Information to be included in the title page:

\title{Теория Вероятностей и Статистика}

\subtitle{Оценивание характеристик распределения}

\institute{старший преподаватель, кандидат экономических наук}

\author{Потанин Богдан Станиславович}

\date{2021}

% Diagonals in tables

\usepackage{diagbox}

\usepackage{slashbox}

\begin{document}

\frame{\titlepage}

%

\begin{frame}

\frametitle{Оценивание характеристик распределения}

\framesubtitle{Мотивация}

\small

\setlength{\belowdisplayskip}{0pt}

\begin{itemize}

\item Зачастую сделать заранее достаточно точное предположение о семействе распределений, из которого была получена выборка, весьма затруднительно.

\item Тем не менее, некоторые отдельные характеристики (вероятности, медиана, математическое ожидание и т.д.) можно оценить, не накладывая существенных допущений о распределении и даже не учитывая его параметры.

\end{itemize}

\end{frame}

%

%

\begin{frame}

\frametitle{Оценивание вероятностей}

\framesubtitle{Безусловные вероятности}

\footnotesize

\setlength{\belowdisplayskip}{0pt}

\begin{itemize}

\item Имеется выборка $X\_{1},...,X\_{n}$ из распределения $\Theta(\theta)$.

\item Необходимо оценить вероятность $P(X\_{1}\in A)$, где $A\subset R$.

\item Несмещенная и состоятельная оценка данной вероятности может быть получена как доля наблюдений в выборке, принадлежащих множеству $A$:

$$\hat{P}(X\_{1}\in A)=\frac{1}{n}\sum\limits\_{i=1}^{n}I(X\_{i}\in A)\sim B(n, P(X\_{1}\in A))$$

$$\text{ где: }I(X\_{i}\in A)=\begin{cases}1\text{, если }X\_{i}\in A\\0\text{, в противном случае}\end{cases}\sim Ber(P(X\_{1}\in A))$$

\item Подставляя вместо наблюдений $X\_{i}$ их реализации $x\_{i}$ можно получить реализацию оценки соответствующей вероятности.\\

\textbf{Доказательство:}

$$E\left(\hat{P}(X\_{1}\in A)\right)=\frac{1}{n}\sum\limits\_{i=1}^{n}E\left(I(X\_{i}\in A)\right)=\frac{1}{n}\sum\limits\_{i=1}^{n}P(X\_{i}\in A)=\frac{1}{n}\times nP(X\_{i}\in A)=P(X\_{i}\in B)$$

$$\lim\limits\_{n\to\infty}Var\left(\hat{P}(X\_{1}\in A)\right)=\lim\limits\_{n\to\infty}\frac{1}{n^2}\sum\limits\_{i=1}^{n}Var\left(I(X\_{i}\in A)\right)=\lim\limits\_{n\to\infty}\frac{n\times Var\left(I(X\_{1}\in A)\right)}{n^2}=0$$

\end{itemize}

\end{frame}

%

%

\begin{frame}

\frametitle{Оценивание вероятностей}

\framesubtitle{Пример оценивания обычных вероятностей}

\footnotesize

\setlength{\belowdisplayskip}{0pt}

\begin{itemize}

\item Имеется выборка с реализацией $x=(5,3,1, -2, 3)$. Найдите реализации оценок вероятностей $P(X=3)$, $P(X\_{1}\geq2.5)$, $P(X\_{1}=2)$, $P(-10\leq X\_{1}<5)$, $P(X\_{1}\in\{2,-2,1\}$ и $P(X\_{1}\leq100)$.\\

\textbf{Решение}:

$$\hat{P}(X=3)(x)=\frac{1}{5}\left(I(X\_{1}=3))+I(X\_{2}=3))+I(X\_{3}=3))+I(X\_{4}=3))+I(X\_{5}=3))\right)=$$

$$=\frac{1}{5}\left(0+1+0+0+1\right)=\frac{2}{5}=0.4$$

$$\hat{P}(X\geq2.5)(x)=\frac{1}{5}\left(1+1+0+0+1\right)=\frac{3}{5}=0.6$$

$$\hat{P}(X=2)(x)=\frac{1}{5}\left(0+0+0+0+0\right)=0$$

$$\hat{P}(-10\leq X<5)(x)=\frac{1}{5}\left(0+1+1+1+1\right)=\frac{4}{5}=0.8$$

$$\hat{P}(X\in\{2,-2,1\})(x)=\frac{1}{5}\left(0+0+1+1+0\right)=\frac{2}{5}=0.4$$

$$\hat{P}(X\leq 100)(x)=\frac{1}{5}\left(1+1+1+1+1\right)=\frac{5}{5}=1$$

\end{itemize}

\end{frame}

%

%

\begin{frame}

\frametitle{Оценивание вероятностей}

\framesubtitle{Условные вероятности}

\scriptsize

\setlength{\belowdisplayskip}{0pt}

\begin{itemize}

\item Имеется выборка $X\_{1},...,X\_{n}$ из распределения $\Theta(\theta)$.

\item Необходимо оценить условную вероятность $P(X\_{1}\in A|X\_{1}\in B)$, где $A,B\subset R$ и $P(X\_{1}\in B)>0$.

\item Несмещенная и состоятельная оценка данной вероятности может быть получена как доля наблюдений в выборке, принадлежащих множеству $A\cap B$, среди наблюдений, принадлежащих $B$:

$$\hat{P}(X\_{1}\in A|X\_{1}\in B)=\frac{1}{m}\sum\limits\_{i=1}^{n}I(X\_{i}\in A\cap B)=\frac{\hat{P}(X\_{1}\in A\cap B)}{\hat{P}(X\_{1}\in B)}\text{ , где }m=\sum\limits\_{i=1}^{n}I(X\_{i}\in B)$$

\item Подставляя вместо наблюдений $X\_{i}$ их реализации $x\_{i}$ можно получить реализацию оценки соответствующей условной вероятности.\\

\textbf{Доказательство:} докажем состоятельность, используя теорему Слуцкого:

$$\begin{cases}\hat{P}(X\_{1}\in A\cap B)\xrightarrow[]{p}P(X\_{1}\in A\cap B)\\ \hat{P}(X\_{1}\in B)\xrightarrow[]{p}P(X\_{1}\in B)\end{cases}\implies \hat{P}(X\_{1}\in A|X\_{1}\in B)\xrightarrow[]{p}\frac{P(X\_{1}\in A\cap B)}{P(X\_{1}\in B)}=P(X\_{1}\in A|X\_{1}\in B)$$

\textbf{Пример:} имеется реализация выборки $x=(5,3,1, -2, 3)$, найдем реализацию оценки $P(X\_{1}\leq 3|X\_{1}>0)$.

$$m=I(X\_{1}>0)+...+I(X\_{5}>0)=1+1+1+0+1=4$$

$$\hat{P}(X\_{1}\leq 3|X\_{1}>0)=\frac{1}{4}\left(I(0<X\_{1}\leq 3)+...+I(0<X\_{5}\leq 3)\right)=\frac{1}{4}\left(0+1+1+0+1\right)=\frac{3}{4}$$

\end{itemize}

\end{frame}

%

%

\begin{frame}

\frametitle{Оценивание вероятностей}

\framesubtitle{Выборочная функция распределения}

\small

\setlength{\belowdisplayskip}{0pt}

\begin{itemize}

\item \textbf{Выборочная функция распределения} определяется как:

$$\hat{F}\_{n}(t)=\hat{P}(X\_{1}\leq t)=\frac{1}{n}\sum\limits\_{i=1}^{n}I(X\_{i}\leq t)\sim B(n, P(X\_{1}\leq t))$$

\item Выборочная функция распределения в точке $t$ является несмещенной и состоятельной оценкой функции распределения в этой же точке, то есть $\hat{F}\_{n}(t)\xrightarrow[]{p}F\_{X\_{1}}(t),\forall t\in R$.\\

\textbf{Пример:}

Найдем реализацию выборочной функции распределения для выборки с реализацией $x=(5,3,1, -2, 3)$.

$$F\_{n}(t)|(X=x)=\begin{cases}(0+0+0+0+0)/5=0\text{, если }t<-2\\ (0+0+0+1+0)/5=0.2\text{, если }-2\leq t<1\\ (0+0+1+1+0)/5=0.4\text{, если }1\leq t<3\\ (0+1+1+1+1)/5=0.8\text{, если }3\leq t<5\\ (1+1+1+1+1)/5=1\text{, если }t\geq 5\\\end{cases}$$

\end{itemize}

\end{frame}

%

%

\begin{frame}

\frametitle{Оценивание вероятностей}

\framesubtitle{Верхняя граница погрешности}

\small

\setlength{\belowdisplayskip}{0pt}

\begin{itemize}

\item Погрешность при использовании выборочной функции распределения вместо истинной можно записать как $|\hat{F}\_{n}(t)-F\_{X\_{1}}(t)|$.

\item При помощи неравенства Чебышева можно найти верхнюю границу для вероятности того, что соответствующая погрешность превысит определенное значение:

$$P(|\hat{F}\_{n}(t)-F\_{X\_{1}}(t)|>\varepsilon)\leq\frac{F\_{n}(t)(1-F\_{n}(t))}{n\varepsilon^2}\leq\frac{1}{4n\varepsilon^2}$$

\item Чем больше объем выборки $n$, тем меньше максимально возможная вероятность получить погрешность, превышающую $\varepsilon$.

\item Например, если объем выборки составляет $n=1000000$ миллион наблюдений, то вероятность того, что погрешность превысит $\varepsilon=0.01$ окажется меньше, чем:

$$\frac{1}{4\times1000000\times0.01^2}=0.0025$$

\end{itemize}

\end{frame}

%

%

\begin{frame}

\frametitle{Выборочные характеристики}

\framesubtitle{Выборочные моменты}

\setlength{\belowdisplayskip}{0pt}

\small

\begin{itemize}

\item Выборочное среднее $\overline{X}\_{n}$ является несмещенной и состоятельной оценкой математического ожидания $E(X\_{1})$.\\

\textbf{Доказательство:}\\

$$E(\overline{X}\_{n})=E(X\_{1})$$

$$\lim\limits\_{n\to\infty}Var(\overline{X}\_{n})=\lim\limits\_{n\to\infty}Var(X\_{1})/n=0$$

\item По аналогии несмещенные и состоятельные оценки начальных моментов $E(X\_{1}^{k})$, где $k\in N$, могут быть получены с использованием выборочных начальных моментов соответствующего порядка:

$$\hat{m}\_{k}=\overline{X^k}\_{n}=\frac{1}{n}\sum\limits\_{i=1}^{n}X\_{i}^{k}$$

\textbf{Пример:} рассчитаем реализацию третьего начального выборочного момента по выборке с реализациями $x=(5,2,5)$:

$$\hat{m}\_{k}(x)=\frac{1}{3}\left(5^3+2^3+5^3\right)=86$$

\end{itemize}

\end{frame}

%

%

\begin{frame}

\frametitle{Выборочные характеристики}

\framesubtitle{Выборочная дисперсия}

\footnotesize

\setlength{\belowdisplayskip}{0pt}

\begin{itemize}

\item \textbf{Выборочная дисперсия} рассчитывается как:

$$S\_{n}^2=\frac{1}{n}\sum\limits\_{i=1}^{n}(X\_{i}-\overline{X})^2$$

\item Она является смещенной, но состоятельной оценкой дисперсии $Var(X\_{1})$.\\

\textbf{Доказательство}: покажем, что оценка является смещенной:

$$E(S\_{n}^2)=\frac{1}{n}\sum\limits\_{i=1}^{n} E((X\_{i}-\overline{X}\_{n})^2)=\frac{1}{n}\times nE((X\_{1}-\overline{X}\_{n})^2)=E((X\_{1}-\overline{X}\_{n})^2)=$$

$$=Var(X\_{1}-\overline{X}\_{n})+E(X\_{1}-\overline{X}\_{n})^2=Var(X\_{1})+Var(\overline{X}\_{n})-2Cov(X\_{1},\overline{X}\_{n})+0^2=$$

$$=Var(X\_{1})+\frac{Var(X\_{1})}{n}-2Cov\left(X\_{1},\frac{1}{n}X\_{1}\right)=Var(X\_{1})\times\left(1+\frac{1}{n}-2\frac{1}{n}\right)=\frac{n-1}{n}Var(X\_{1})$$

\item Состоятельная и несмещенная оценка дисперсии может быть получена с использование \textbf{исправленной (скорректированной) выборочной дисперсии}:

$$\hat{\sigma}^2=\frac{n}{n-1}S\_{n}^2=\frac{1}{n-1}\sum\limits\_{i=1}^{n}(X\_{i}-\overline{X}\_{n})^2$$

\end{itemize}

\end{frame}

%

%

\begin{frame}

\frametitle{Выборочные характеристики}

\framesubtitle{Вариационный ряд}

\scriptsize

\setlength{\belowdisplayskip}{0pt}

\begin{itemize}

\item Обозначим через $X\_{(i)}$ наблюдение, которое является $i$-м по величине в выборке. Оно называется $\mathbf{i}$\textbf{-й порядковой статистикой}. Статистики $X\_{(1)}=\min(X\_{1},...,X\_{n})$ и $X\_{(n)}=\max(X\_{1},...,X\_{n})$ именуются \textbf{минимальной и максимальной порядковыми статистиками} соответственно.

\item Последовательность $X\_{(1)},...,X\_{(n)}$ \textbf{именуется вариационным рядом}.

\item Распределение максимальной экстремальной статистики определяется как:

$$F\_{X\_{(1)}}(x)=F\_{\max(X\_{1},...,X\_{n})}(x)=P(\max(X\_{1},...,X\_{n})\leq x)=P(X\_{1}\leq x,...,X\_{n}\leq x)=$$

$$=P(X\_{1}\leq x) \times... \times P(X\_{n}\leq x)=F\_{X\_{1}}(x) \times... \times F\_{X\_{n}}(x)=\left(F\_{X\_{1}}(x)\right)^{n}$$

\item Распределение минимальной экстремальной статистики определяется как:

$$F\_{X\_{(n)}}(x)=F\_{\min(X\_{1},...,X\_{n})}(x)=P(\min(X\_{1},...,X\_{n})\leq x)=1-P(\min(X\_{1},...,X\_{n})> x)=$$

$$=1-P(X\_{1}> x)\times ... \times P(X\_{n}> x)=1-\left(1-F\_{X\_{1}}(x)\right)\times ... \times \left(1-F\_{X\_{n}}(x)\right)=1-\left(1-F\_{X\_{1}}(x)\right)^n$$

\item Распределение $i$-й порядковой статистики определяется как:

$$F\_{X\_{(i)}}(x)=\sum\_{k=i}^{n}C\_{n}^{k}\left(1-F\_{X\_{1}}(x)\right)^{n-k}\left(F\_{X\_{1}}(x)\right)^{k}$$

\textbf{Пример:} найдем вероятность того, что в выборке из $n=5$ экспоненциальных случайных величин с параметром $\lambda=0.2$ наибольшее значение не превысит $10$:

$$F\_{X\_{(5)}}(10)=(1-e^{-0.2\times10})^5\approx0.483$$

\end{itemize}

\end{frame}

%

%

\begin{frame}

\frametitle{Оценивание функции плотности}

\framesubtitle{Гистограмма}

\scriptsize

\setlength{\belowdisplayskip}{0pt}

\begin{itemize}

\item Гистограмма является оценкой функции плотности.

\item Разобьем выборку на $m$ интервалов равной длины $h$, где интервал обозначим как $b\_{k}$.

\item Гистограмма предполагает следующую (как правило смещенную) оценку функции плотности в точке $t\in b\_{k}$:

$$\hat{f}\_{X\_{i}}(t)=\frac{1}{nh}\sum\limits\_{i=1}^{n}I(X\_{i}\in b\_{k})$$

\item Эффективность оценки зависит от параметра $h$. Существуют различные подходы к подбору оптимального значения данного параметра, например, по правилу Райса (Rice rule) $h=2n^{1.5}$.\\

\begin{tikzpicture}

\begin{axis}[

width=0.88\textwidth,

height=0.35\textwidth,

ymin=0, ymax=0.5,

xmin = -3.5, xmax = 3.5,

minor y tick num = 3,

area style,

color=MidnightBlue,

]

\addplot+[ybar interval,mark=no] plot coordinates { (-3.5, 0.002) (-3.0, 0.02) (-2.5, 0.02) (-2.0, 0.056) (-1.5, 0.170) (-1.0, 0.294) (-0.5, 0.394) (0, 0.426) (0.5, 0.280) (1.0, 0.180) (1.5, 0.116) (2.0, 0.032) (2.5, 0.006) (3.0, 0.004)};

\addlegendentry{Гистограмма}

\addplot [

domain=-5:5,

samples=100,

color=Maroon,

line width=2pt

]

{1 / ((2 \* pi) ^ 0.5 \* 1) \* e ^ (-(x - 0) ^ 2 / (2 \* 1 ^ 2))};

\addlegendentry{Функция плотности}

\end{axis}

\end{tikzpicture}

\end{itemize}

\end{frame}

%

%

\begin{frame}

\frametitle{Оценивание функции плотности}

\framesubtitle{Ядерное оценивание}

\scriptsize

\setlength{\belowdisplayskip}{0pt}

\begin{itemize}

\item Функция плотности при помощи ядер оценивается следующим образом:

$$\hat{f}\_{X\_{1}}(t)=\frac{1}{nh}\sum\limits\_{i=1}^{n}K\left(\frac{t-x\_{i}}{h}\right)$$

\item Функция $K$ является неотрицательной и именуется \textbf{ядром}. В качестве нее часто используют функцию плотности стандартного нормального распределения $K=\phi(t)$ (гауссовское ядро).

\item Параметр $h$ именуется \textbf{шириной окна} и его подбирают из соображения максимизации эффективности.

\begin{tikzpicture}

\begin{axis}[

width=0.88\textwidth,

height=0.35\textwidth,

ymin=0, ymax=0.5,

xmin = -3.5, xmax = 3.5,

minor y tick num = 3,

area style,

color=MidnightBlue,

]

\addplot+[ybar interval,mark=no] plot coordinates { (-3.5, 0.002) (-3.0, 0.02) (-2.5, 0.02) (-2.0, 0.056) (-1.5, 0.170) (-1.0, 0.294) (-0.5, 0.394) (0, 0.426) (0.5, 0.280) (1.0, 0.180) (1.5, 0.116) (2.0, 0.032) (2.5, 0.006) (3.0, 0.004)};

\addlegendentry{Гистограмма}

\addplot [

domain=-5:5,

samples=100,

color=Maroon,

line width=2pt

]

{1 / ((2 \* pi) ^ 0.5 \* 1) \* e ^ (-(x - 0) ^ 2 / (2 \* 1 ^ 2))};

\addlegendentry{Функция плотности}

\addplot [

domain=-5:5,

samples=100,

color=lime,

line width=2pt

]

{1 / ((2 \* pi) ^ 0.5 \* 1) \* 0.5 \* (e ^ (-(x - 0.2) ^ 2 / (2 \* 1 ^ 2)) + e ^ (-(x + 0.5) ^ 2 / (2 \* 1 ^ 2)))};

\addlegendentry{Ядерная оценка плотности}

\end{axis}

\end{tikzpicture}

\end{itemize}

\end{frame}

%

\end{document}