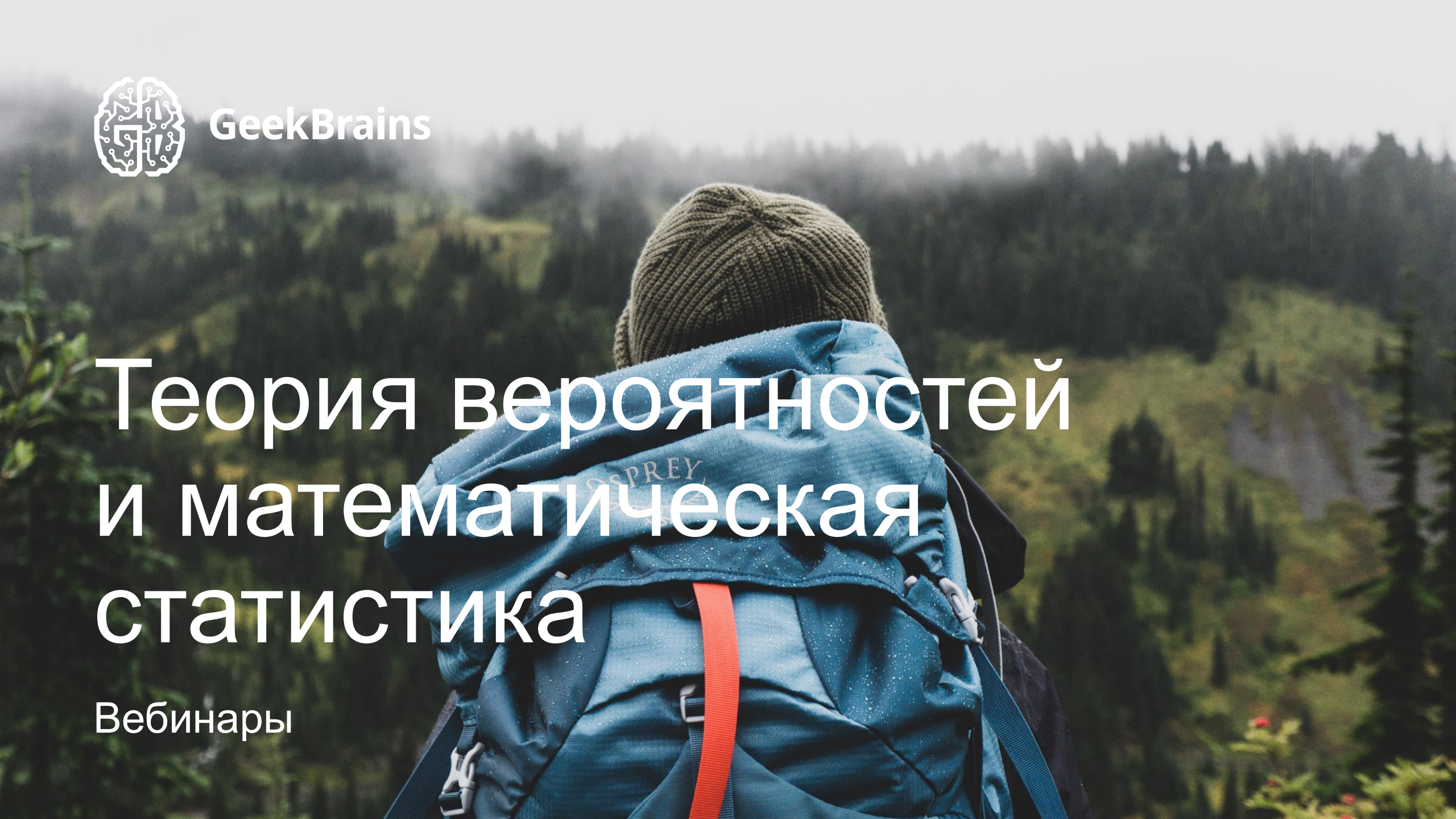
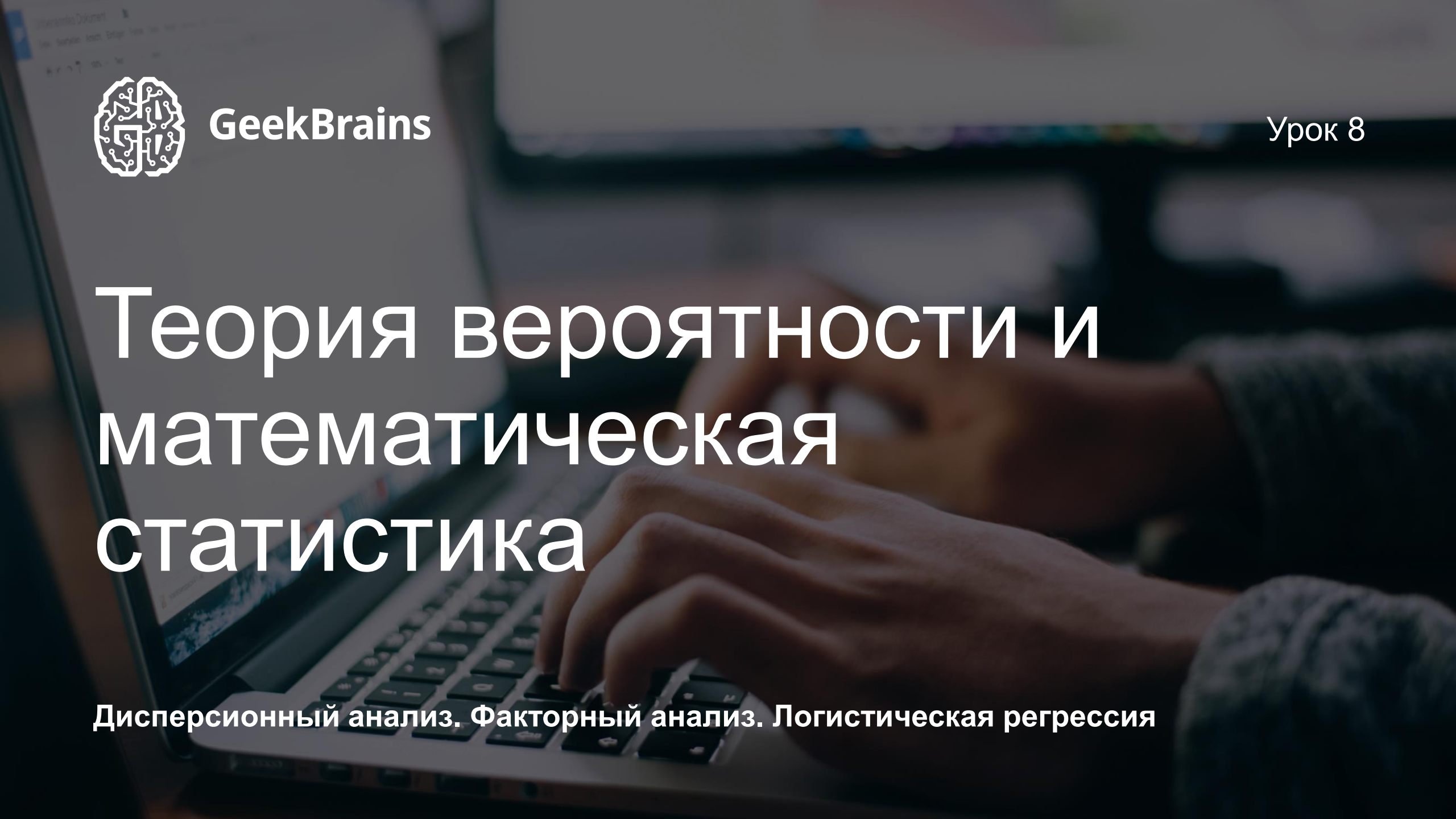


Теория вероятностей и математическая статистика

Вебинары



GeekBrains

Урок 8

Теория вероятности и математическая статистика

Дисперсионный анализ. Факторный анализ. Логистическая регрессия

На этом уроке мы изучим:

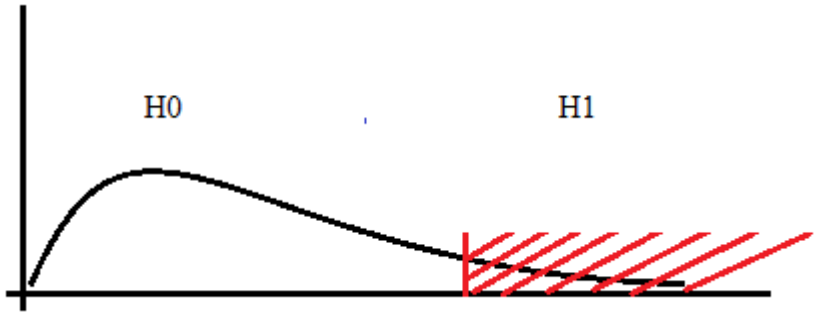
1. Однофакторный дисперсионный анализ
2. Двухфакторный дисперсионный анализ
3. Логистическая регрессия

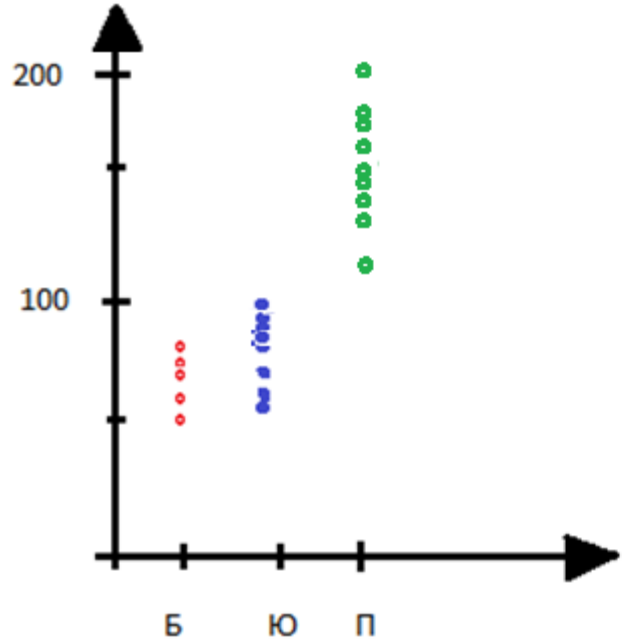
Дисперсный анализ используется для исследования влияния одного или нескольких качественных показателей на количественный показатель.

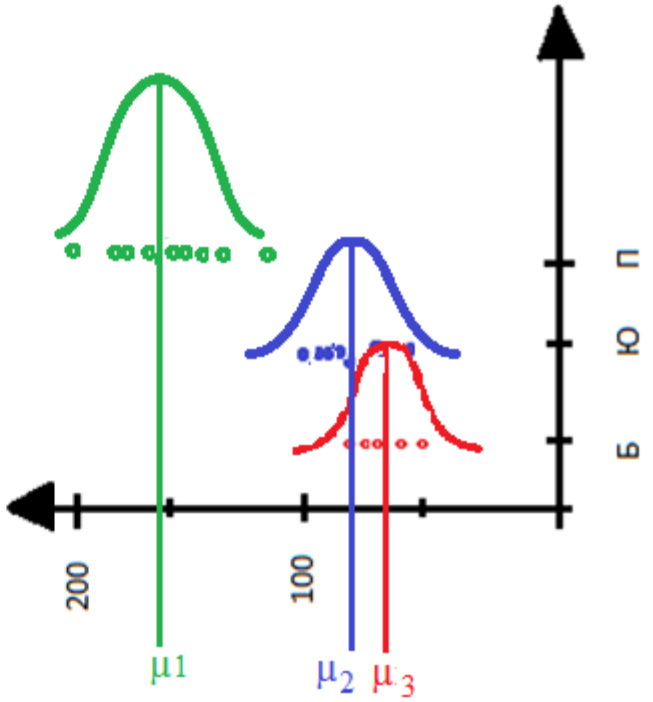
В **однофакторном дисперсионном анализе** на одну количественную переменную ***Y*** влияет один фактор (один качественный показатель), наблюдаемый на ***k*** уровнях, то есть имеем ***k*** выборок для переменной ***Y.***

Например, с помощью однофакторного дисперсионного анализа можно определить, является ли статистически значимым различие среднего размера заработной платы (количественный признак - переменная ***Y***) в трех разных группах людей, отличающихся по признаку профессии, которая в данном случае будет являться качественным фактором, наблюдаемым на ***k*** уровнях (этими уровнями могут быть, к примеру, профессии бухгалтера, юриста и программиста).

Тестируем нулевую гипотезу
H0 : $\mu_1 = \mu_2 = \mu_3$





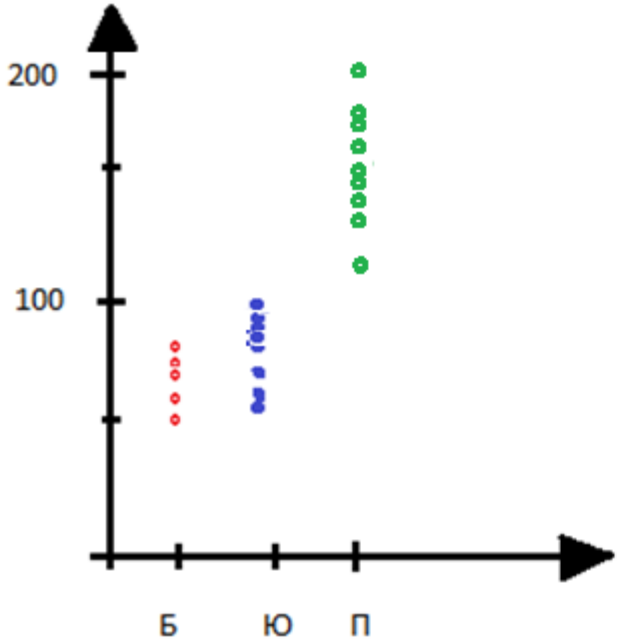


F_H ___ F_t

$$F_H = \frac{\sigma_F^2}{\sigma_{ост}^2}$$

$$\sigma_F^2 = \frac{S_F^2}{k-1}$$

$$\sigma_{ост}^2 = \frac{S_{ост}^2}{n-k}$$



k =3
n1 = 5 (бухгалтера)
n2 = 8 (юристы)
n3 = 7 (программисты)

n= n1+n2+n3

1) $\overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$

$\overline{y}_1$ =(70+50+65+60+75)/5 =64
$\overline{y}_2$ =(80+75+90+70+75+65+85+100)/8 =80
$\overline{y}_3$ =(130+100+140+150+160+170+200)/7=150

$\overline{Y}$ — среднее значение переменной **Y** по всем значениям:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{k}\overline{y}_i n_i$$

$\overline{Y}$ =(70+50+65+60+75+80+75+90+70+75+65+85+100+130 +100+140+150+160+170+200)/20=100,5

S^2 — сумма квадратов отклонений наблюдений от общего среднего:

$$S^2 : \sum_{i=1}^{k}\sum_{i=1}^{n_i}(y_{ij} - \overline{Y})^2$$

$S^2 =$ 34445

S_F^2 — сумма квадратов отклонений средних групповых значений от общего среднего значения $\overline{Y}$:

$$S_F^2 = \sum_{i=1}^{k}(\overline{y}_i - \overline{Y})^2 n_i$$

S_F^2 = (64-100,5)^2 * 5 +(80-100,5)^2*8+(150-100,5)^2*7 = 27175

$S^2_{\text{ост}}$ — остаточная сумма квадратов отклонений:

$$S^2_{\text{ост}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$$

$S^2_{\text{ост}}$ =(70-64)^2+(50-64)^2+(65-64)^2+.....+(200-150)^2= 7270

$$S^2 = S^2_F + S^2_{\text{ост}}$$

S^2 = 27175+ 7270=34445

Вычисляем факторную дисперсию: $\sigma^2_F = \frac{S^2_F}{k-1} = \frac{1}{k-1}\sum_{i=1}^{k}(\overline{y}_i - \overline{Y})^2 n_i$

σ^2_F =27175/(3-1) = 13587,5

Вычислим остаточную дисперсию: $\sigma^2_{\text{ост}} = \frac{S^2_{\text{ост}}}{n-k} = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$

$\sigma^2_{\text{ост}}$ = 7270/(20-3) = 427,65

$$F_H = \frac{\sigma^2_F}{\sigma^2_{\text{ост}}}$$ = 13587,5/427,65=31,772

Найдем значение $F_{крит}$ в таблице критических точек распределения Фишера-Снедекора для заданного уровня значимости $\alpha = 0.05$ и двух степеней свободы:

$df_{межд} = k - 1 = 3 - 1 = 2$ и $df_{внутр} = n - k = 20 - 3 = 17$.

$F_{крит}$ =3,59

31,772 > 3,59
Следовательно, различие между группами статистически значимое

Условия для дисперсионного анализа:
1. Значения групп следуют нормальному распределению
2. Однородность дисперсий
3. Независимость

Уровень значимости $\alpha = 0,05$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12
1	161	200	216	225	230	234	237	239	241	242	244
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,90
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,50	2,38

Вывод:

профессия влияет на уровень заработной платы

ANOVA

In [4]:
```
import numpy as np
```

In [5]:
```
from scipy import stats
```

In [6]:
```
stats.f_oneway?
```

In [7]:
```
y1=np.array([70,50,65,60,75])
y1
```

Out[7]:
```
array([70, 50, 65, 60, 75])
```

In [8]:
```
y2= ([80,75,90,70,75,65,85,100])
y2
```

Out[8]:
```
[80, 75, 90, 70, 75, 65, 85, 100]
```

In [9]:
```
y3= ([130,100,140,150,160,170,200])
y3
```

Out[9]:
```
[130, 100, 140, 150, 160, 170, 200]
```

In [10]:
```
stats.f_oneway(y1,y2,y3)
```

Out[10]:
```
F_onewayResult(statistic=31.77269601100413, pvalue=1.8091304567650962e-06)
```

В **двухфакторном дисперсионном анализе** на одну количественную переменную ***Y*** влияют два фактора (два качественных показателя), наблюдаемый соответственно на ***k*** и ***m*** уровнях.

$$y_{ijk} = M + A_i + B_j + AB + E_{ijk}$$

$$y_{ijk} - M = A_i + B_j + AB + E_{ijk}$$

$$(y_{ijk} - M)^2 = A_i^2 + B_j^2 + AB^2 + E_{ijk}^2$$

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

SS_A	$F_{pA} \_?\_ F_{TA}$
SS_B	$F_{pB} \_\_ F_{TB}$
SS_{AB}	$F_{pAB} \_\_ F_{TAB}$
SS_E	

Фактор B J

Yijk

k=2-количество репликаций

Фактор A i

	1 уровень	2 уровень	
1 уровень	57 ; 59 58	56;58 57	57.5
2 уровень	32; 34 33	71;71 71	52
	45.5	64	54.75

```
2-WAY anova

# y ijn ,где i- уовень ФАКТОРА А,  j- уровень фактора Б, n- число репликаций

y111=57

y112=59

y11=(y111+y112)/2=58
y11=58

y121=56

y122==58

y12=(y121+y122)/2
y12=57

y211=32

y212=34

y21=(y211+y212)/2
y21=33

y221=71

y222=71

y22=(y221+yy222)/2
y22=71
```

```
YcpA1=(y11+y12)/2=(58+57)/2=57.5

YcpA2=(y21+y22)/2=(33+71)/2=52

YcpB1=(y11+y21)/2=(58+33)/2=45.5

YcpB2=(y12+y22)/2=(57+71)/2=64

Ycp=mean(YcpA1+YcpA2+YcpB1+YcpB2)/4

Ycp=54.75
```

```
SSt=sum((yijk)**2)-a*b*n*(Ycp**2)=57**2+59**2+.....+71**2+71**2 - 2*2*2*(54.75)**2=1512

SSA=a*n* sum((YcpA)**2)-a*b*n(Ycp)**2 = 2*2*((57.5)**2+(52.5)**2)-8*(54.75)**2=61

SSB=b*n*sum((YcpB)**2)-a*b*n*(Ycp)**2= 2*2*((45.5)**2+(64)**2)-8*(54.75)**2=685

SSAB=n*(sum((yij_cp)**2)-a*b*n*Ycp-SSA-SSB=2*((58)**2+(57)**2+(33)**2+(71)**2)-8*((54.74)**2)-61-685=759.5

SSE=SSt-SSA-SSB-SSAB=1512-61-685-759.5=6.5

dfA=2-1=1 ## (2 уровня a-1)

dfB= 2-1=1 ##2 уровня b-1
dfAB = (a-1)*(b-1)=(2-1)*(2-1)=1

dfE= a*b*(n-1)=2*2(2-1)=4

MSA=SSA/dfA = 61/1=61

MSB=SSB/dfB=685/1=685

MSAB=SSAB/dfAB=759.5/1=759.5

MSE=SSE/dfE=6.5/4=1.625

FA=MSA/MSE=61/1.625=37.54

FB=MSB/MSE=685/1.625=421.54

FAB=MSAB/MSE=759.5/1.625=467

F=7.71

# Сравниваем все рассчитанные  F с табличным ,наибольший эффект взаимодействие факторов AB
```

	SS	df	MS	F
A	61	1	61	MSA/MSEr= 61/1.625=37.54
B	685	1	685	421.54
AB	759.5	1	759.5	467
Er	6.5	4	1.625	

Фактор B J

Yijk

k=2-количество репликаций

Фактор A i	1 уровень	2 уровень	
1 уровень	57 ; 59 58	56;58 57	57.5
2 уровень	32; 34 33	71;71 71	52
	45.5	64	54.75

In [36]:
```
1 import numpy as np
```

In [37]:
```
1 import statsmodels.api as sm
```

In [38]:
```
1 from statsmodels.formula.api import ols
```

In [35]:
```
df=pd.DataFrame({'fA':fA,'fB':fB,'treatments':treatments})
df
```

Out[35]:

	fA	fB	treatments
0	low	low	57
1	low	low	59
2	low	high	56
3	low	high	58
4	high	low	32
5	high	low	34
6	high	high	71
7	high	high	71

In [63]:
```
lm_model = ols('treatments ~ C(fA)*C(fB)',
               data=df).fit()
```

In [64]:
```
table = sm.stats.anova_lm(lm_model, typ=2)
table
```

Out[64]:

	sum_sq	df	F	PR(>F)
C(fA)	60.5	1.0	40.333333	0.003150
C(fB)	684.5	1.0	456.333333	0.000028
C(fA):C(fB)	760.5	1.0	507.000000	0.000023
Residual	6.0	4.0	NaN	NaN

	SS	df	MS	F
A	61	1	61	MSA/MSEr= 61/1.625=37.54
B	685	1	685	421.54
AB	759.5	1	759.5	467
Er	6.5	4	1.625	

Логистическая регрессия

Статистический метод, с помощью которого можно решать задачу бинарной классификации.

Логистическая регрессия

С помощью этого метода можно не только отнести объект

к одному из двух классов, но и оценить вероятности того,

что объект относится к данному классу для каждого из классов.

	x1	x2	x3	y
	zp	prod	poezdky	vozvrat
1	100	30	1	1
2	40	20	0	0
3	50	20	1	1
4	70	40	2	1
5	50	30	0	0
6	80	70	3	1
7	75	25	4	1

```
> Z =-0.18839+ 0.01115 *x1 -0.00279*x2+ 0.16286*x3 # модель
```

1 клиент

```
x1=68
x2=27
x3=2
Z
[1] 0.8202

> sigmoid= 1/(1+ e^(- Z ))
> sigmoid
[1] 0.6937473
```

2 клиент

```
x1=72
x2=40
x3=0
Z
[1] 0.50281

> sigmoid= 1/(1+ e^(- Z ))
> sigmoid
[1] 0.6227591
```

3 клиент

```
x1=120
x2=40
x3=1
Z
[1] 1.20087

> sigmoid= 1/(1+ e^(- Z ))
> sigmoid
[1] 0.7680273
```

Логистическая регрессия

Используется в банковском бизнесе для определения кредитоспособности заемщика. На основе показателя вероятности события “клиент отдаст долг”, полученного с помощью логистической регрессии, вычисляется скоринговый балл клиента
и принимается решение о выдаче кредита.

Итоги

1. Однофакторный дисперсионный анализ
2. Двухфакторный дисперсионный анализ
3. Логистическая регрессия