```
Using matplotlib backend: TkAgg
Populating the interactive namespace from numpy and matplotlib
```

Теорема Байеса (https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BE%D1

Теорема Байеса (или **формула Байеса**) — одна из основных теорем элементарной ***теории вероятностей***, которая позволяет определить ***вероятность*** какого-либо события при условии, что произошло другое статистически ***взаимозависимое*** с ним событие. Другими словами, по формуле Байеса можно более точно пересчитать вероятность, взяв в расчет как ранее известную информацию, так и данные новых наблюдений. Формула Байеса может быть выведена из основных аксиом теории вероятностей, в частности из условной вероятности. Особенность теоремы Байеса заключается в том, что для её практического применения требуется большое количество расчетов, вычислений, поэтому байесовские оценки стали активно использовать только после революции в компьютерных и сетевых технологиях.

Формулировка

Формула Байеса:

$$\boxed{P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}}$$

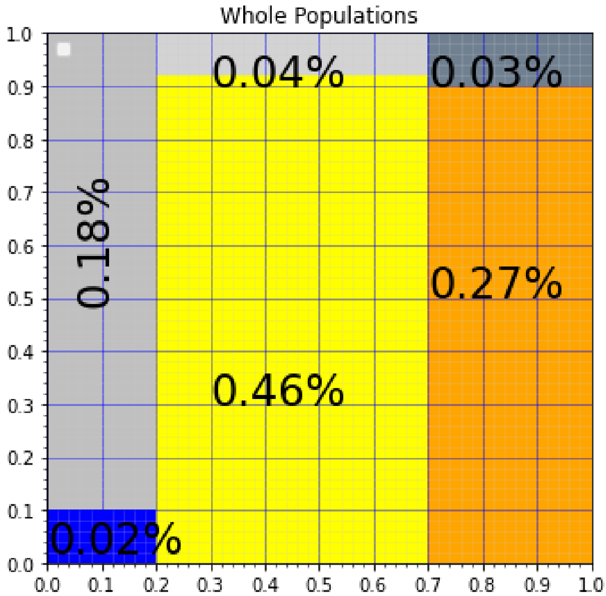
где

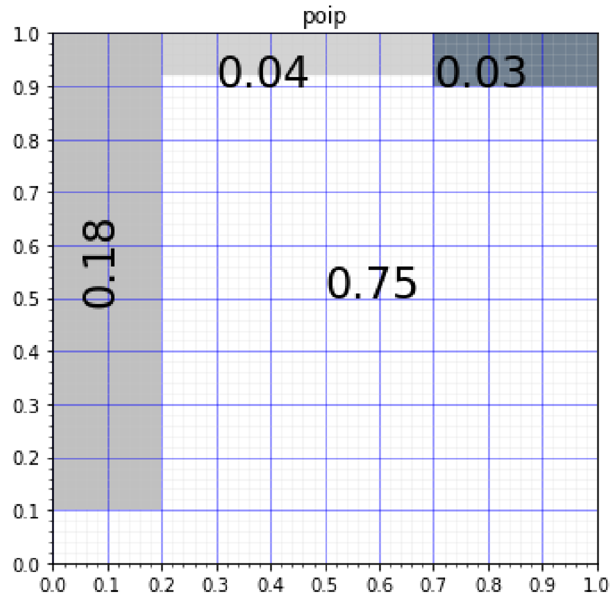
- $P(A)$ — априорная вероятность гипотезы A (смысл такой терминологии см. ниже);
- $P(A \mid B)$ — вероятность гипотезы A при наступлении события B (апостериорная вероятность);
- $P(B \mid A)$ — вероятность наступления события B при истинности гипотезы A;
- $P(B)$ — полная вероятность наступления события B.

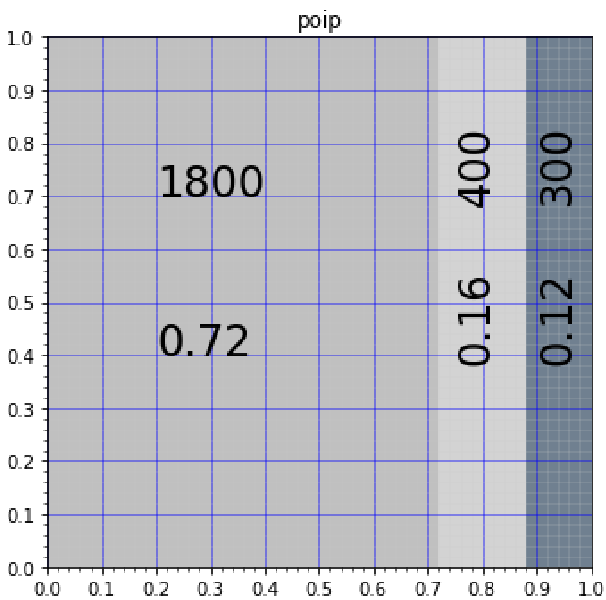
Байесовская вероятность (https://ru.wikipedia.org/wiki/%D0%91%D0%B0%D0%B9%D0%B5%D1%81%D0%BE%D0%B2%D1%81%D0%BA%

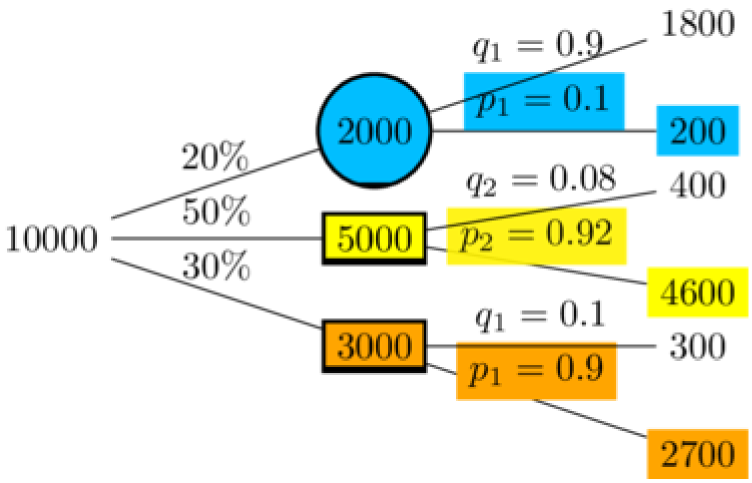
Байесовская вероятность — это интерпретация понятия ***вероятности***, используемая в байесовской теории. Вероятность определяется как степень уверенности в истинности ***суждения***. Для определения степени уверенности в истинности суждения при получении новой информации в байесовской теории используется **теорема Байеса**.

```
N= 10000
 200 +1800 == 2000  p1=0.10 q1=0.90
4600 + 400 == 5000  p2=0.92 q2=0.08
2700 + 300 == 3000  p3=0.90 q3=0.10
------------------
7500 +2500 ==10000
PA1|b ==0.72
PA2|b ==0.16
PA3|b ==0.12
PA1|b +PA2|b +PA3|b == 1.0

No handles with labels found to put in legend.
```









Пример 1

Производитель	Количество n_i	вероятность Брака модели $=q_i$	Количество Брака	Рабочие телефоны
Российский	2000	0.9	1800	200
iPhone	5000	0.08	400	4600
Samsung	3000	0.1	300	2700
Всего	10000		2500	7500

Все они продаются в единственном магазине в Москве. Радиолюбитель из глубинки проездом купил один. Телефон оказался поломанный. Спрашивается, с какой вероятностью этот телефон **Российский**, **iPhone**, **Samsung** ?

- Событие B — поломанный телефон,
- событие A_i — деталь произведена заводом i.

Тогда $P(A_i) = \dfrac{n_i}{N}$, где $N = n_1 + n_2 + n_3$, а $P(B|A_i) = q_i$.

По формуле полной вероятности

$$\boxed{P(B) = \sum_{i=1}^{3} P(B \mid A_i)P(A_i) = \frac{q_1 \cdot n_1}{N} + \frac{q_2 \cdot n_2}{N} + \frac{q_3 \cdot n_3}{N} = \frac{0.9 \cdot 2000 + 0.08 \cdot 5000 + 0.1 \cdot 3000}{10000} = = \frac{2500}{10000} = \boxed{0.25 = P(B)}}$$

По формуле Байеса получим:

$$
\boxed{
\begin{aligned}
P(A_1 \mid B) &= \frac{P(B \mid A_1)P(A_1)}{P(B)} &&= \frac{P(B \mid A_1)P(A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)} = \\
&= \frac{\frac{q_1 n_1}{N}}{\frac{q_1 n_1}{N} + \frac{q_2 n_2}{N} + \frac{q_3 n_3}{N}} &&= \frac{\frac{0.9 \cdot 2000}{10000}}{\frac{0.9 \cdot 2000}{10000} + \frac{0.08 \cdot 5000}{10000} + \frac{0.10 \cdot 30000}{10000}} = \\
&= \frac{\frac{1800}{10000}}{\frac{1800 + 400 + 300}{10000}} &&= \frac{1800}{2500} = \boxed{0.72 = P(A_1 \mid B)}
\end{aligned}
}
$$

$$
\boxed{
\begin{aligned}
P(A_2 \mid B) &= \frac{P(B \mid A_2)P(A_2)}{P(B)} &&= \frac{P(B \mid A_2)P(A_2)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)} = \\
&= \frac{\frac{q_2 n_2}{N}}{\frac{q_1 n_1}{N} + \frac{q_2 n_2}{N} + \frac{q_3 n_3}{N}} &&= \frac{\frac{0.08 \cdot 5000}{10000}}{\frac{0.9 \cdot 2000}{10000} + \frac{0.08 \cdot 5000}{10000} + \frac{0.10 \cdot 30000}{10000}} = \\
&= \frac{\frac{400}{10000}}{\frac{1800 + 400 + 300}{10000}} &&= \frac{400}{2500} = \boxed{0.16 = P(A_2 \mid B)}
\end{aligned}
}
$$

$$
\boxed{
\begin{aligned}
P(A_3 \mid B) &= \frac{P(B \mid A_3)P(A_3)}{P(B)} &&= \frac{P(B \mid A_3)P(A_3)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)} = \\
&= \frac{\frac{q_3 n_3}{N}}{\frac{q_1 n_1}{N} + \frac{q_2 n_2}{N} + \frac{q_3 n_3}{N}} &&= \frac{\frac{0.10 \cdot 3000}{10000}}{\frac{0.9 \cdot 2000}{10000} + \frac{0.08 \cdot 5000}{10000} + \frac{0.10 \cdot 30000}{10000}} = \\
&= \frac{\frac{300}{10000}}{\frac{1800 + 400 + 300}{10000}} &&= \frac{300}{2500} = \boxed{0.12 = P(A_1 \mid B)}
\end{aligned}
}
$$

Prob & Stats - Bayes Theorem (12 of 24) What if We Run the Test Again?

Тестирование **1000** человек на наличие болезни происходит в следущих условиях:

- **1%** являются больными
- **10** являются **больными**
- **990 здоровы**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **95%** == **SPECIFIC == TRUE Negative**, **5% == FALSE Positive**

```
#1(9.8) + #2( 49.5) == #5( 59.3)
#3(0.2) + #4(940.5) == #6(940.7)
P(D|+)=0.1652613827993255  P(H|+)=0.8347386172006745
P(D|+)=0.7950989320307973  P(H|+)=0.2049010679692027
P(D|+)=0.9870224116396645  P(H|+)=0.0129775883603355
-----------------------------------
P(D|+)=0.1652613827993255  P(H|+)=0.8347386172006745
P(D|+)=0.7950989320307973  P(H|+)=0.2049010679692027
P(D|+)=0.9870224116396645  P(H|+)=0.0129775883603355
```

TRUTH

TEST		**Disease**	**Healthy**	**Total**
	Test Posivive +	$\boxed{1}$ TRUE $\oplus$ 98% = 9.8	$\boxed{2}$ FALSE $\oplus$ 5% = 49.5	$\boxed{5}$ 59.3
	Test Negative −	$\boxed{3}$ FALSE $\ominus$ 2% = 0.2	$\boxed{4}$ TRUE $\ominus$ 95% = 940.5	$\boxed{6}$ 940.7
		10	990	1000

Первый способ из 13 параграфа ниже:

Вероятность того что больной является по Настоящему больным при

Первом ***положительном*** тестировании:

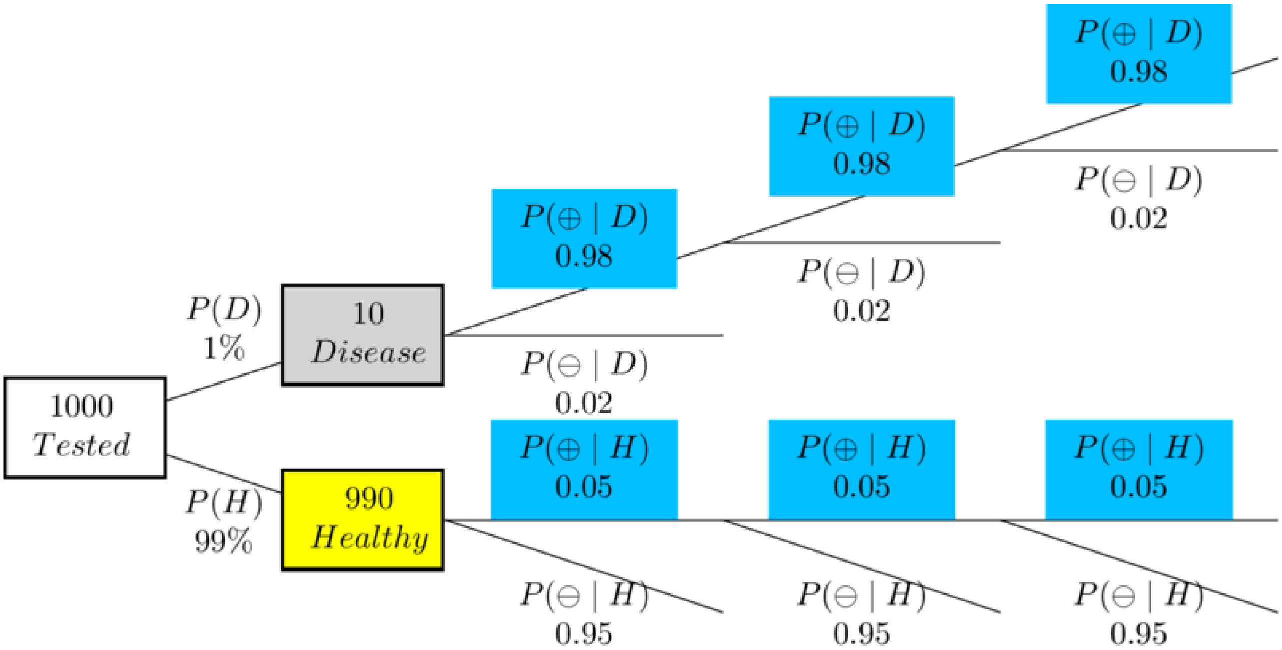
$$\boxed{P(D \mid +) = \frac{(98\%)\,(1\%)}{(98\%)(1\%) + (5\%)(99\%)} = 16.53\%} \quad \boxed{P(H \mid +) = 83.47\%}$$

Втором ***положительном*** тестированиии:

$$\boxed{P(D \mid +) = \frac{(98\%)\,(16.53\%)}{(98\%)(16.53\%) + (5\%)(83.47\%)} = 79.50989\%} \quad \boxed{P(H \mid +) = 20.49\%}$$

Третьем ***положительном*** тестированиии:

$$\boxed{P(D \mid +) = \frac{(98\%)\,(79.51\%)}{(98\%)(79.51\%) + (5\%)(20.49\%)} = 98.70224\%} \quad \boxed{P(H \mid +) = 1.2977\%}$$



Второй способ из русской Wikipwdia:

$$\begin{aligned} P(D \mid +) &= \boxed{\frac{A \cdot P}{A \cdot P + B \cdot (1-P)}} \\ P(D \mid +,+) &= \boxed{\frac{A^2 \cdot P}{A^2 \cdot P + B^2 \cdot (1-P)}} \\ P(D \mid +,+,+) &= \boxed{\frac{A^3 \cdot P}{A^3 \cdot P + B^3 \cdot (1-P)}} \end{aligned}$$

Вероятность того что больной является по Настоящему больным при

Первом ***положительном*** тестировании:

$$\boxed{P(D \mid +) = \frac{0.01 \cdot 0.98^1}{0.01 \cdot 0.98^1 + 0.99 \cdot 0.05^1} = 16.53\%} \quad \boxed{P(H \mid +) = 83.47\%}$$

Втором ***положительном*** тестировании:

$$\boxed{\begin{aligned} P(D \mid +,+) &= \frac{0.01 \cdot 0.98^2}{0.01 \cdot 0.98^2 + 0.99 \cdot 0.05^2} && = \\ &= \frac{0.01 \cdot \left[0.98 \cdot 0.98\right]}{0.01 \cdot \left[0.98 \cdot 0.98\right] + 0.99 \cdot \left[0.05 \cdot 0.05\right]} && = 79.50989\% \end{aligned}} \quad \boxed{P(H \mid +) = 20.49\%}$$

Третьем ***положительном*** тестированиии:

$$\boxed{\begin{aligned} P(D \mid +,+,+) &= \frac{0.01 \cdot 0.98^3}{0.01 \cdot 0.98^3 + 0.99 \cdot 0.05^3} && = \\ &= \frac{0.01 \cdot \left[0.98 \cdot 0.98 \cdot 0.98\right]}{0.01 \cdot \left[0.98 \cdot 0.98 \cdot 0.98\right] + 0.99 \cdot \left[0.05 \cdot 0.05 \cdot 0.05\right]} && = 98.70224\% \end{aligned}}$$

$$\boxed{P(H \mid +) = 1.2977\%}$$

$$\boxed{\begin{aligned} P(D \mid +) &= \boxed{\frac{A \cdot P}{A \cdot P + B \cdot (1 - P)}} \\ P(D \mid +,+) &= \boxed{\frac{A^2 \cdot P}{A^2 \cdot P + B^2 \cdot (1 - P)}} \\ P(D \mid +,+,+) &= \boxed{\frac{A^3 \cdot P}{A^3 \cdot P + B^3 \cdot (1 - P)}} \end{aligned}}$$

$$P(D \mid +) = \frac{A \cdot P(D)}{A \cdot P(D) + B \cdot (1 - P(D))} = \boxed{\frac{A \cdot P}{A \cdot P + B \cdot (1 - P)}}$$

$$
\begin{aligned}
P(D \mid +,+) &= \frac{A \cdot \left[\frac{A \cdot P}{A \cdot P + B \cdot (1-P)}\right]}{A \cdot \left[\frac{A \cdot P}{A \cdot P + B \cdot (1-P)}\right] + B \cdot \left(1 - \left[\frac{A \cdot P}{A \cdot P + B \cdot (1-P)}\right]\right)} &= \\
&= \frac{A \cdot \left[\frac{A \cdot P}{A \cdot P + B \cdot (1-P)}\right]}{A \cdot \left[\frac{A \cdot P}{A \cdot P + B \cdot (1-P)}\right] + B \cdot \left(1 - \left[\frac{A \cdot P}{A \cdot P + B \cdot (1-P)}\right]\right)} &= \\
&= \frac{\frac{A^2 \cdot P}{A \cdot P + B \cdot (1-P)}}{\frac{A^2 \cdot P}{A \cdot P + B \cdot (1-P)} + B \cdot \left(\frac{A \cdot P + B \cdot (1-P) - A \cdot P}{A \cdot P + B \cdot (1-P)}\right)} &= \\
&= \frac{\frac{A^2 \cdot P}{A \cdot P + B \cdot (1-P)}}{\frac{A^2 \cdot P}{A \cdot P + B \cdot (1-P)} + B \cdot \left(\frac{B \cdot (1-P)}{A \cdot P + B \cdot (1-P)}\right)} &= \\
&= \frac{\frac{A^2 \cdot P}{A \cdot P + B \cdot (1-P)}}{\frac{A^2 \cdot P}{A \cdot P + B \cdot (1-P)} + \frac{B^2 \cdot (1-P)}{A \cdot P + B \cdot (1-P)}} &= \\
&= \boxed{\frac{A^2 \cdot P}{A^2 \cdot P + B^2 \cdot (1-P)}}
\end{aligned}
$$

Type *Markdown* and LaTeX: α^2

[Prob & Stats - Bayes Theorem (1 of 24) What is Bayes Theorem? (https://youtu.be/gTaxZplxFEw)](https://youtu.be/gTaxZplxFEw)

$$
\boxed{P(A|B) = \frac{P(B \mid A)\, P(A)}{P(B)}}
$$

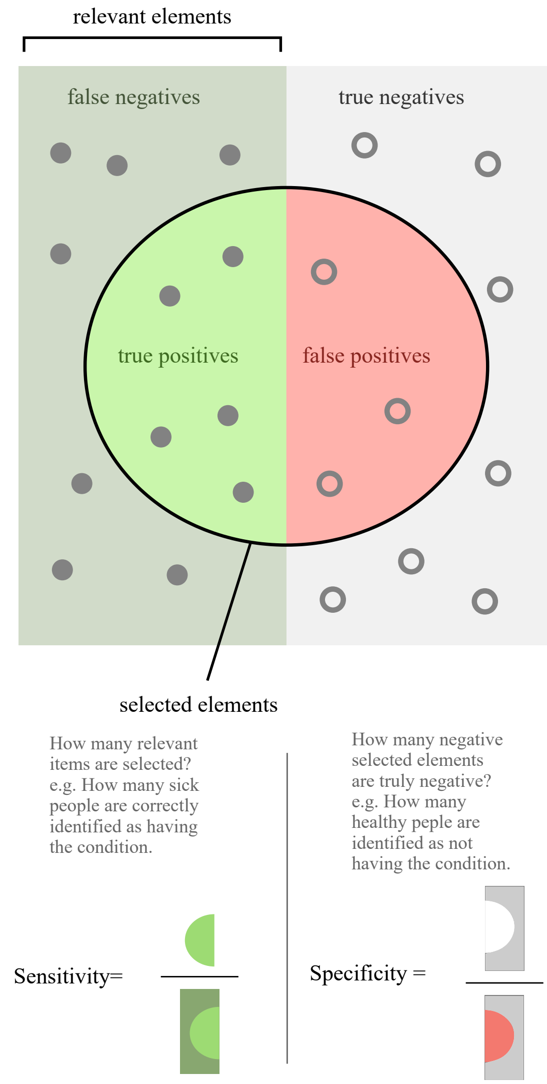
$$
\begin{aligned}
&\text{Where} \\
&P(A \mid B) = && \text{Вероятность события } A \\
& && \text{ когда прозошло событие } B = True \\
&P(B \mid A) = && \text{Вероятность события } B \\
& && \text{ когда прозошло событие } A = True \\
&P(A) = && \text{Вероятность события } A \text{ независимо от } B \\
&P(B) = && \text{Вероятность события } B \text{ независимо от } A
\end{aligned}
$$

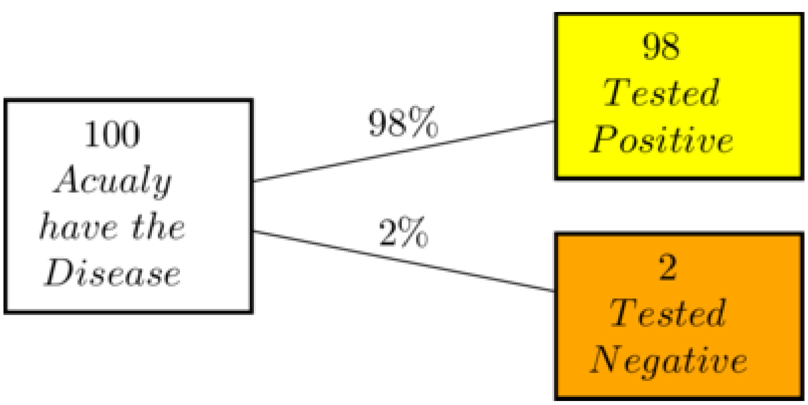
$$
\boxed{P(U \mid +) = \frac{P(+ \mid U)\, P(U)}{P(+)}} \qquad \boxed{P(D \mid +) = \frac{P(+ \mid D)\, P(D)}{P(+)}}
$$

$$
\begin{aligned}
P(+) &= P(TRUE+) + P(FALSE+) \\
+ &= \text{Test POSITIVE}
\end{aligned}
$$

SENSITIVITY (чувствительность, восприимчивость)
Чувствительность (истинно положительный)
Тест показывает вероятность того, что больной субъект будет классифицирован именно как больной.

SPECIFICITY
Специфичность





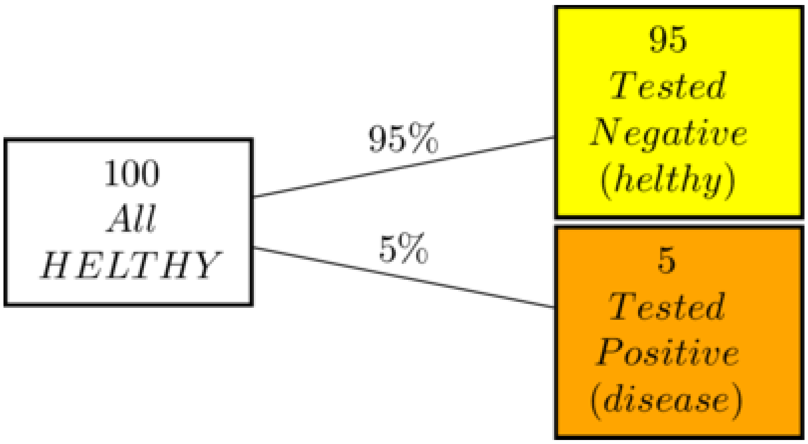
[Prob & Stats - Bayes Theorem (2 of 24) What is the Sensitivity of a Test? (https://youtu.be/pKE3v7tBp3w)](https://youtu.be/pKE3v7tBp3w)

Sensitivity (also called the **true positive rate**, the **recall**, or **probability of detection** in some fields) measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition). It is often mistakenly confused with the detection limit, while the detection limit is calculated from the analytical sensitivity, not from the epidemiological sensitivity.

Sensitivity/Чувствительность (истинно положительная пропорция) отражает долю положительных результатов, которые правильно идентифицированы как таковые. Иными словами, чувствительность диагностического теста показывает вероятность того, что больной субъект будет классифицирован именно как больной:

Тестируются **все больные**/положительные объекты 100:

- 98 больных будут определено правильно $+$
- 2 определяется неправильно $-$ 2 имеющих болезнь будут отмечены $FALSE\ NEGATIVE$



[Prob & Stats - Bayes Theorem (3 of 24) What is the Specificity of a Test? (https://youtu.be/8i62dc74mc0)](https://youtu.be/8i62dc74mc0)

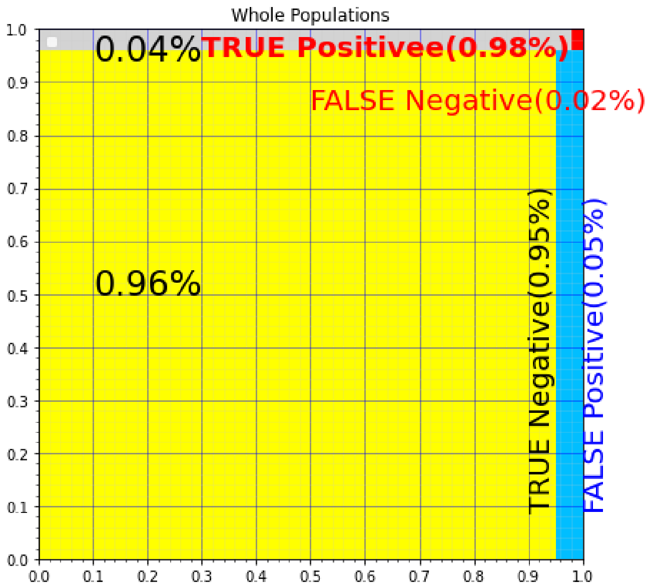
Specificity (also called the **true negative rate**) measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

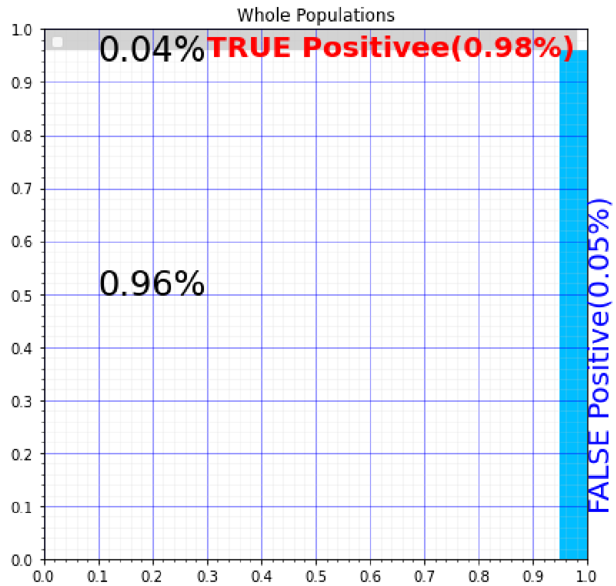
Specificity/Специфичность (истинно отрицательная пропорция) отражает долю отрицательных результатов, которые правильно идентифицированы как таковые, то есть вероятность того, что здоровые субъекты будут классифицированы именно как здоровые.:

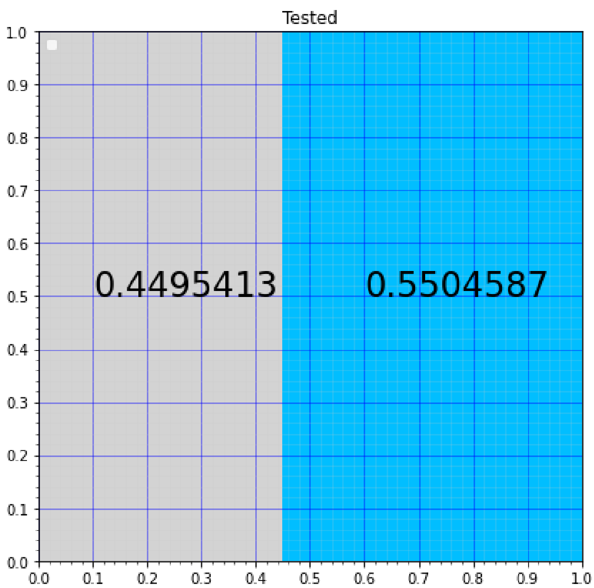
Тестируются **полностью здоровые**/положительные объекты 100:

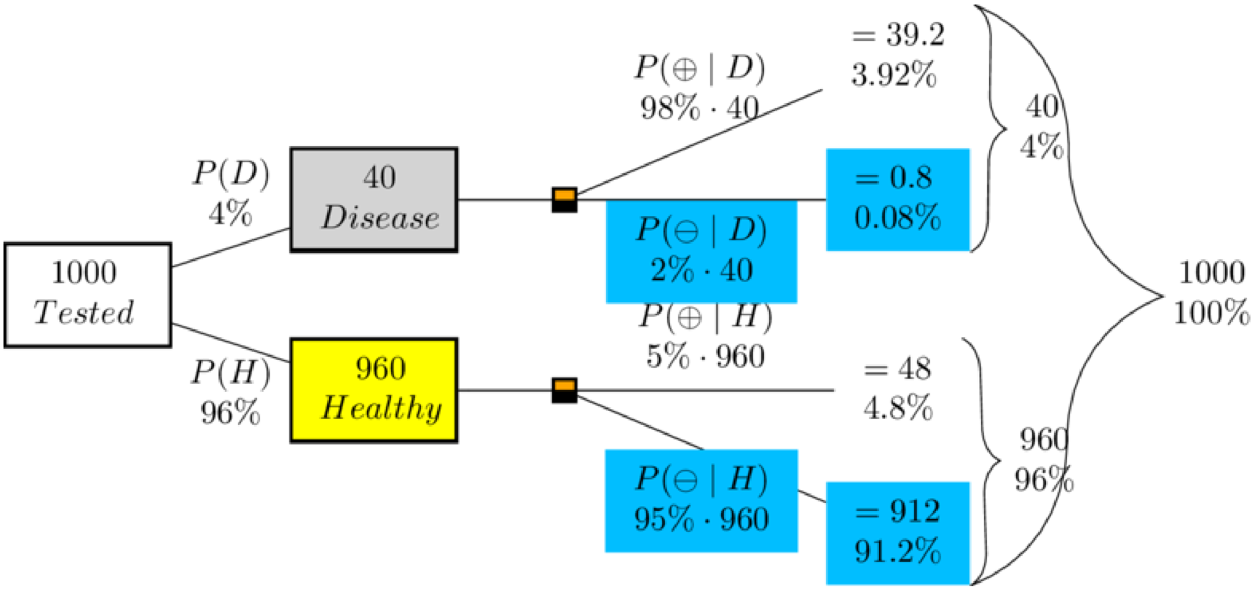
- 95 здоровых будут определено правильно $-$
- 5 определяется неправильно $+$ 5 здоровых будут отмечены как больные $FALSE\ POSITIVE$

```
No handles with labels found to put in legend.
No handles with labels found to put in legend.
No handles with labels found to put in legend.

0.0392 0.048 11.46788990825688 0.4495412844036697 0.5504587155963303

<matplotlib.legend.Legend at 0x207112e01f0>
```









[Prob & Stats - Bayes Theorem (4 of 24) A More Comprehensive Equation (https://youtu.be/QAYmrCuL2rQ)](https://youtu.be/QAYmrCuL2rQ)

После тестирования **1000** больных получили такой результат при качестве тестов:

- **4%** являются больными
- **40** тестированы как **больные**
- **960** оставлены **здоровыми**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **95%** == **SPECIFIC == TRUE Negative**, **5% == FALSE Positive**

$$\boxed{P(D \mid +) = \frac{P(+ \mid D)\ P(D)}{P(+)} \quad = \frac{\overbrace{P(+ \mid D)\ P(D)}^{\text{tested + TRUE POSITIVE}}}{\underbrace{P(+ \mid D)P(D) + P(+ \mid H)P(H)}_{\text{All tested +/POSITIVE}}}}$$

```
0.4495412844036697
```

Вероятность того что больной является по Настоящему больным при **положительном** тестировании:

$$P(D \mid +) = \frac{(98\%)\ (4\%)}{(98\%)(4\%) + (5\%)(96\%)} \quad = 44.95\%$$

[Prob & Stats - Bayes Theorem (5 of 24) A More Comprehensive Equation: Another Method (https://youtu.be/3cpEQEYn2PE)](https://youtu.be/3cpEQEYn2PE)

После тестирования **1000** больных получили такой результат при качестве тестов:

- **4%** являются больными
- **40** тестированы как **больные**

- **960** оставлены **здоровыми**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **95%** == **SPECIFIC == TRUE Negative**, **5% == FALSE Positive**

TRUTH

		Disease	**Healthy**	**Total**
TEST	Test Posivive $+$	$\boxed{1}$ TRUE $\oplus$ 98%	$\boxed{2}$ FALSE $\oplus$ 5%	
	Test Negative $-$	$\boxed{3}$ FALSE $\ominus$ 2%	$\boxed{4}$ TRUE $\ominus$ 95%	
		40	960	1000

$$P(D \mid +) = \frac{\boxed{1}}{\boxed{1} + \boxed{2}} = \frac{P(+ \mid D)\, P(D)}{P(+)} = \frac{\overbrace{P(+ \mid D)\, P(D)}^{\text{tested + TRUE POSITIVE}}}{\underbrace{P(+ \mid D)P(D) + P(+ \mid H)P(H)}_{\text{All tested +/POSITIVE}}}$$

```
#1(39.2) + #2( 48.0) == #5( 87.2)
#3( 0.8) + #4(912.0) == #6(912.8)
     40  +      960  ==   1000
```

[Prob & Stats - Bayes Theorem (6 of 24) A More Comprehensive Equation: Another Method (https://youtu.be/RYpejUyHvaY)](https://youtu.be/RYpejUyHvaY)

После тестирования больных получили такой результат при качестве тестов:

- **4%** являются больными
- **40** тестированы как **больные**
- **960** оставлены **здоровыми**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **95%** == **SPECIFIC == TRUE Negative**, **5% == FALSE Positive**

TRUTH

		Disease	**Healthy**	**Total**
TEST	Test Posivive $+$	$\boxed{1}$ TRUE $\oplus$ $98\% = 39.2$	$\boxed{2}$ FALSE $\oplus$ $5\% = 48$	87.2
	Test Negative $-$	$\boxed{3}$ FALSE $\ominus$ $2\% = 0.8$	$\boxed{4}$ TRUE $\ominus$ $95\% = 912$	912.8
		40	960	1000

$$P(D \mid +) = \frac{\boxed{1}}{\boxed{1}+\boxed{2}} = \frac{P(+\mid D)\; P(D)}{P(+)} = \frac{\overbrace{P(+\mid D)\; P(D)}^{\text{tested + TRUE POSITIVE}}}{\underbrace{P(+\mid D)P(D) + P(+\mid H)P(H)}_{\text{All tested +/POSITIVE}}}$$

Вероятность того что больной является по Настоящему больным при **положительном** тестировании:

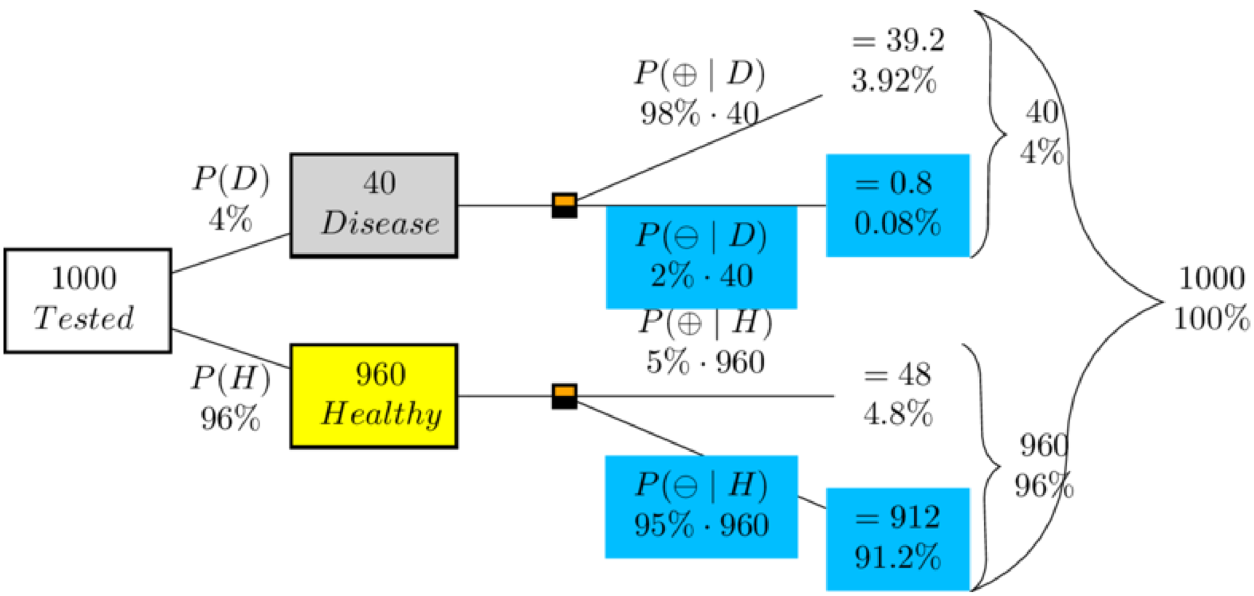
$$P(D \mid +) = \frac{39.2}{39.2 + 48} = \frac{(98\%)\,(4\%)}{(98\%)(4\%) + (5\%)(96\%)} = 44.95\%$$

[Prob & Stats - Bayes Theorem (7 of 24) The Tree Diagram (https://youtu.be/94J2yHPFvQc)](https://youtu.be/94J2yHPFvQc)

После тестирования больных получили такой результат при качестве тестов:

- **4%** являются больными
- **40** тестированы как **больные**
- **960** оставлены **здоровыми**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **95%** == **SPECIFIC == TRUE Negative**, **5% == FALSE Positive**



[Prob & Stats - Bayes Theorem (8 of 24) What Happens if the Disease is Rare? (https://youtu.be/yzTP-gKXoil)](https://youtu.be/yzTP-gKXoil)

После тестирования **1000** больных получили такой результат при качестве тестов:

- 1% ~~**4%**~~ являются больными
- 10 ~~**40**~~ тестированы как **больные**
- 990 ~~**960**~~ оставлены **здоровыми**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **95%** == **SPECIFIC == TRUE Negative**, **5% == FALSE Positive**

```
0.4495412844036697
0.16526138279932545
```

Вероятность того что больной является по Настоящему больным при **положительном** тестировании:

$$P(D \mid +) = \frac{(98\%)\ (4\%)}{(98\%)(4\%) + (5\%)(96\%)} = 44.95\%$$
$$P(D \mid +) = \frac{(98\%)\ (1\%)}{(98\%)(1\%) + (5\%)(99\%)} = 16.53\%$$

[Prob & Stats - Bayes Theorem (9 of 24) What Happens if the Disease is Very Rare? (https://youtu.be/O8TzqxFDsyw)](https://youtu.be/O8TzqxFDsyw)

После тестирования **1000** людей получили такой результат при качестве тестов:

- 0.1% ~~**1%**~~ ~~**4%**~~ являются больными
- 1 ~~**10**~~ ~~**40**~~ тестированы как **больные**
- 999 ~~**990**~~ ~~**960**~~ оставлены **здоровыми**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **95%** == **SPECIFIC == TRUE Negative**, **5% == FALSE Positive**

```
0.4495412844036697
0.16526138279932545
0.01924209699587669
```

Вероятность того что больной является по Настоящему больным при **положительном** тестировании:

$$P(D \mid +) = \frac{(98\%)\ (4\%)}{(98\%)(4\%) + (5\%)(96\%)} = 44.95\%$$
$$P(D \mid +) = \frac{(98\%)\ (1\%)}{(98\%)(1\%) + (5\%)(99\%)} = 16.53\%$$
$$P(D \mid +) = \frac{(98\%)\ (0.1\%)}{(98\%)(0.1\%) + (5\%)(99.9\%)} = 1.92\%$$

[Prob & Stats - Bayes Theorem (10 of 24) What Happens if the Disease is Very, Very Rare? (https://youtu.be/K3-IzBXB_0w)](https://youtu.be/K3-IzBXB_0w)

После тестирования **1000** больных получили такой результат при качестве тестов:

- 0.1% ~~**1%**~~ ~~**4%**~~ являются больными
- 1 ~~**10**~~ ~~**40**~~ тестированы как **больные**
- 999 ~~**990**~~ ~~**960**~~ оставлены **здоровыми**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test 99% ~~**95%**~~ == **SPECIFIC == TRUE Negative**, 1% ~~**5%**~~ **== FALSE Positive**

```
0.4495412844036697
0.16526138279932545
0.01924209699587669
0.08933454876937101
```

Вероятность того что больной является по Настоящему больным при **положительном** тестировании:

$$P(D \mid +) = \frac{(98\%)\ (4\%)}{(98\%)(4\%) + (5\%)(96\%)} = 44.95\%$$

$$P(D \mid +) = \frac{(98\%)\ (1\%)}{(98\%)(1\%) + (5\%)(99\%)} = 16.53\%$$

$$P(D \mid +) = \frac{(98\%)\ (0.1\%)}{(98\%)(0.1\%) + (5\%)(99.9\%)} = 1.92\%$$

$$P(D \mid +) = \frac{(98\%)\ (0.1\%)}{(98\%)(0.1\%) + (1\%)(99.9\%)} = 8.93\%$$

[Prob & Stats - Bayes Theorem (11 of 24) What Happens if Disease is Very, Very Rare & Better Testing? (https://youtu.be/nbJi3vNa6fA)](https://youtu.be/nbJi3vNa6fA)

TRUTH

TEST		**Disease**	**Healthy**	**Total**
	Test Posivive +	$\boxed{1}$ TRUE $\oplus$ 98% = 0.98	$\boxed{2}$ FALSE $\oplus$ 1% = 9.99	$\boxed{5}$ 10.97
	Test Negative −	$\boxed{3}$ FALSE $\ominus$ 2% = 0.02	$\boxed{4}$ TRUE $\ominus$ 99% = 989.01	$\boxed{6}$ 989.03
		1	999	1000

$$\boxed{P(D \mid +) = \frac{\boxed{1}}{\boxed{1} + \boxed{2}} = \frac{\boxed{1}}{\boxed{5}} = \frac{0.98}{10.97} = 0.08933 \quad (8.93\%)}$$

[Prob & Stats - Bayes Theorem (12 of 24) What if We Run the Test Again? (https://youtu.be/ncaCq6FIsTg)](https://youtu.be/ncaCq6FIsTg)

Тестирование **1000** человек на наличие болезни происходит в следущих условиях:

- **1%** являются больными
- **10** являются **больными**
- **990 здоровы**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **95%** == **SPECIFIC == TRUE Negative**, **5% == FALSE Positive**

		TRUTH		
		Disease	**Healthy**	**Total**
TEST	Test Posivive +	$\boxed{1}$ TRUE $\oplus$ 98% = 9.8	$\boxed{2}$ FALSE $\oplus$ 5% = 49.5	$\boxed{5}$ 59.3
	Test Negative −	$\boxed{3}$ FALSE $\ominus$ 2% = 0.2	$\boxed{4}$ TRUE $\ominus$ 95% = 940.5	$\boxed{6}$ 940.7
		10	990	1000

```
#1(9.8) + #2( 49.5) == #5( 59.3)
#3(0.2) + #4(940.5) == #6(940.7)
P(D|+)=0.1652613827993255  P(H|+)=0.8347386172006745
P(D|+)=0.7950989320307973  P(H|+)=0.2049010679692027
```

Вероятность того что больной является по Настоящему больным при

Первом ***положительном*** тестировании:

$$\boxed{P(D \mid +) = \frac{(98\%)\,(1\%)}{(98\%)(1\%) + (5\%)(99\%)} = 16.53\%} \quad \boxed{P(H \mid +) = 83.47\%}$$

Втором ***положительном*** тестировании:

$$\boxed{P(D \mid +) = \frac{(98\%)\,(16.53\%)}{(98\%)(16.53\%) + (5\%)(83.47\%)} = 79.50989\%} \quad \boxed{P(H \mid +) = 20.49\%}$$

[Prob & Stats - Bayes Theorem (13 of 24) What if We Run the Test Again and Again? (https://youtu.be/n_TvlbA--y8)](https://youtu.be/n_TvlbA--y8)

Тестирование **1000** человек на наличие болезни происходит в следущих условиях:

- **1%** являются больными
- **10** являются **больными**
- **990 здоровы**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **95%** == **SPECIFIC == TRUE Negative**, **5% == FALSE Positive**

```
#1(9.8) + #2( 49.5) == #5( 59.3)
#3(0.2) + #4(940.5) == #6(940.7)
P(D|+)=0.1652613827993255  P(H|+)=0.8347386172006745
P(D|+)=0.7950989320307973  P(H|+)=0.2049010679692027
P(D|+)=0.9870224116396645  P(H|+)=0.0129775883603355
```

		TRUTH		
		Disease	**Healthy**	**Total**
TEST	Test Posivive +	$\boxed{1}$ TRUE $\oplus$ 98% = 9.8	$\boxed{2}$ FALSE $\oplus$ 5% = 49.5	$\boxed{5}$ 59.3
	Test Negative −	$\boxed{3}$ FALSE $\ominus$ 2% = 0.2	$\boxed{4}$ TRUE $\ominus$ 95% = 940.5	$\boxed{6}$ 940.7
		10	990	1000

Вероятность того что больной является по Настоящему больным при

Первом ***положительном*** тестировании:

$$\boxed{P(D \mid +) = \frac{(98\%)\ (1\%)}{(98\%)(1\%) + (5\%)(99\%)} = 16.53\%} \quad \boxed{P(H \mid +) = 83.47\%}$$

Втором ***положительном*** тестированиии:

$$\boxed{P(D \mid +) = \frac{(98\%)\ (16.53\%)}{(98\%)(16.53\%) + (5\%)(83.47\%)} = 79.50989\%}$$

$$\boxed{P(H \mid +) = 20.49\%}$$

Третьем ***положительном*** тестированиии:

$$\boxed{P(D \mid +) = \frac{(98\%)\ (79.51\%)}{(98\%)(79.51\%) + (5\%)(20.49\%)} = 98.70224\%}$$

$$\boxed{P(H \mid +) = 1.2977\%}$$

Основной целью диагностического теста является постановка диагноза, поэтому мы должны знать вероятность того, что тест позволяет ставить правильный диагноз. Чувствительность и специфичность не дают нам подобной информации. Вместо этого необходимо анализировать результаты теста, используя прогностические значения.

Prob & Stats - Bayes Theorem (14 of 24) What is Positive Predictive Value (PPV)? (https://youtu.be/JX4Je4bO4Zw)

Положительное прогностическое значение - доля пациентов с положительными результатами теста, которые были правильно диагностированы.

Положительные и отрицательные прогностические значения (https://sites.google.com/site/konstbel/knigi/zametki-po-medicinskoj-statistike/diagnosticeskie-testy-2-prognosticeskie-znacenia)

Положительные и отрицательные прогностические значения (**PPV** и **NPV** соответственно) являются пропорциями положительных и отрицательных результатов статистических и диагностических тестов, которыми являются истинно положительные и истинно отрицательные результаты, соответственно. **PPV** и **NPV** описывают характеристики диагностического теста или другой статистической мерой. Высокий результат можно интерпретировать как указание на точность такой статистики. **PPV** может быть получен с помощью теоремы Байеса .

В поиске информации , **PPV** часто называют **точностью**.

$$PPV = \frac{\text{the number of TRUE POSITIVES}}{\text{the TOTAL number of POSITIVES}}$$

$$PPV = \frac{\text{number of TRUE POSITIVES}}{\text{number of TRUE POSITIVES} + \text{number of FALSE POSITIVES}}$$

$$PPV = \frac{\text{TRUE POSITIVES} \oplus}{\text{TOTAL POSITIVES} \oplus} = \frac{P(+ \mid D)\ P(D)}{P(+)} = \frac{P(+ \mid D)\ P(D)}{P(+ \mid D)P(D) + P(+ \mid H)P(H)} = P(D \mid +)$$

Положительное прогностическое значение $\boxed{PPV = P(D \mid +)}$ Вероятность того что объект является Правильно диагностирован при его положительном результате тестирования.

[Prob & Stats - Bayes Theorem (15 of 24) What is Negative Predictive Value (NPV)? (https://youtu.be/QFScbw9WKpM)](https://youtu.be/QFScbw9WKpM)

Отрицательное прогностическое значение - доля пациентов с отрицательными результатами теста, которые были правильно диагностированы.

[Положительные и отрицательные прогностические значения (https://sites.google.com/site/konstbel/knigi/zametki-po-medicinskoj-statistike/diagnosticeskie-testy-2-prognosticeskie-znacenia)](https://sites.google.com/site/konstbel/knigi/zametki-po-medicinskoj-statistike/diagnosticeskie-testy-2-prognosticeskie-znacenia)

$$NPV = \frac{\text{the number of TRUE NEGATIVES}}{\text{the TOTAL number of NEGATIVES}}$$

$$NPV = \frac{\text{number of TRUE NEGATIVES}}{\text{number of TRUE NEGATIVES} + \text{number of FALSE NEGATIVES}}$$

$$NPV = \frac{\text{TRUE NEGATIVES} \ominus}{\text{TOTAL NEGATIVES} \ominus} = \frac{P(- \mid H)\ P(H)}{P(-)} = \frac{P(- \mid H)\ P(H)}{\underbrace{P(- \mid H)P(H)}_{\text{TRUE NEGATIVES}} + \underbrace{P(- \mid D)P(D)}_{\text{FALSE NEGATIVE}}} = P(H \mid -)$$

Отрицательное прогностическое значение $\boxed{NPV = P(H \mid -)}$ Вероятность того что объект является Правильно диагностирован при его отрицательном результате тестирования.

[Prob & Stats - Bayes Theorem (16 of 24) PPV & NPV Numerical Examples (https://youtu.be/2WcSd7FfUFI)](https://youtu.be/2WcSd7FfUFI)

После тестирования больных получили такой результат при качестве тестов:

- **4%** являются больными
- **40** тестированы как **больные**
- **960** оставлены **здоровыми**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **95%** == **SPECIFIC == TRUE Negative**, **5% == FALSE Positive**

TRUTH

TEST		**Disease**	**Healthy**	**Total**
	Test Posivive $+$	$\boxed{1}$ TRUE $\oplus$ $98\% = 39.2$	$\boxed{2}$ FALSE $\oplus$ $5\% = 48$	$\boxed{5}$ TOTAL $\oplus$ 87.2
	Test Negative $-$	$\boxed{3}$ FALSE $\ominus$ $2\% = 0.8$	$\boxed{4}$ TRUE $\ominus$ $95\% = 912$	$\boxed{6}$ TOTAL $\ominus$ 912.8
		40	960	1000

```
#1(39.2) + #2(48.0) == #5(87.2)
#3(0.8) + #4(912.0) == #6(912.8)
PPV = 0.44954128440366975
NPV = 0.9991235758106924
0.4495412844036697
0.5504587155963303
Sensetivity = 0.9800000000000001
Specifity = 0.95
```

$$PPV = \frac{\text{TRUE } \oplus}{\text{TOTAL } \oplus} = \frac{39.2}{87.2} = 44.954\%$$

$$NPV = \frac{\text{TRUE } \ominus}{\text{TOTAL } \ominus} = \frac{912}{912.8} = 99.912\%$$

[Prob & Stats - Bayes Theorem (17 of 24) Prevalence, Sensitivity, Specificity, PPV, NPV (https://youtu.be/87coLIgU_us)](https://youtu.be/87coLIgU_us)

- **4%** являются больными
- **40** тестированы как **больные**
- **960** оставлены **здоровыми**

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **95%** == **SPECIFIC == TRUE Negative**, **5% == FALSE Positive**

TRUTH

TEST		**Disease**	**Healthy**	**Total**
	Test Posivive $+$	$\boxed{1}$ TRUE $\oplus$ $98\% = 39.2$	$\boxed{2}$ FALSE $\oplus$ $5\% = 48$	$\boxed{5}$ TOTAL $\oplus$ 87.2
	Test Negative $-$	$\boxed{3}$ FALSE $\ominus$ $2\% = 0.8$	$\boxed{4}$ TRUE $\ominus$ $95\% = 912$	$\boxed{6}$ TOTAL $\ominus$ 912.8
		40	960	1000

$$P(D \mid +) = PPV = \frac{P(+ \mid D)\, P(D)}{P(+)} = \frac{\overbrace{P(+ \mid D)\, P(D)}^{\text{tested + TRUE POSITIVE}}}{\underbrace{P(+ \mid D)P(D) + P(+ \mid H)P(H)}_{\text{All tested +/POSITIVE}}}$$

$$= \frac{(98\%)\,(4\%)}{(98\%)(4\%) + (5\%)(96\%)} = \frac{39.2}{39.2 + 48} = \frac{39.2}{87.2} = 44.95\%$$

$$\text{Распространенность заболевания} = \frac{TOTAL\ \text{больных}}{TOTAL\ \text{людей}} = \frac{40}{1000} = 4\%$$

$$SENSITIVITY = \frac{\text{TRUE} \oplus}{\text{TRUE} \oplus + \text{FALSE} \ominus} = \frac{\text{TRUE} \oplus}{\text{TOTAL D}} = \frac{39.2}{40} = 98\%$$

$$SPECIFITY = \frac{\text{TRUE} \ominus}{\text{TRUE} \ominus + \text{FALSE} \oplus} = \frac{\text{TRUE} \ominus}{\text{TOTAL H}} = \frac{912}{960} = 95\%$$

[Prob & Stats - Bayes Theorem (18 of 24) Simple Form of the Definitions (https://youtu.be/DoJzXmZpDP0)](https://youtu.be/DoJzXmZpDP0)

$$PPV = P\left(D \mid \oplus\right) = \frac{\text{TRUE} \oplus}{\text{TOTAL} \oplus} = \frac{39.2}{87.2} = 44.954\%$$

$$NPV = P\left(H \mid \ominus\right) = \frac{\text{TRUE} \ominus}{\text{TOTAL} \ominus} = \frac{912}{912.8} = 99.912\%$$

[Prob & Stats - Bayes Theorem (19 of 24) Determining Sensitivity and Specificity (https://youtu.be/FqTG-Eg5t9s)](https://youtu.be/FqTG-Eg5t9s)

TRUTH

	Disease	**Healthy**	
TEST	$\boxed{1}$ TRUE $\oplus$ 19	$\boxed{2}$ FALSE $\oplus$ 5	$= 24$
	$\boxed{3}$ FALSE $\ominus$ 1	$\boxed{4}$ TRUE $\ominus$ 75	$= 76$
	20	80	100

$$SENSITIVITY = \frac{\text{TRUE} \oplus}{\text{TOTAL D}} = \frac{\text{TRUE} \oplus}{\text{TRUE} \oplus + \text{FALSE} \ominus} = \frac{19}{20} = 95\%$$

$$SPECIFITY = \frac{\text{TRUE} \ominus}{\text{TOTAL H}} = \frac{\text{TRUE} \ominus}{\text{TRUE} \ominus + \text{FALSE} \oplus} = \frac{75}{80} = 93.75\%$$

$$PPV = P\left(D \mid \bigoplus\right) = \frac{\text{TRUE } \bigoplus}{\text{TOTAL } \bigoplus} = \frac{19}{24} = 79.17\%$$

$$NPV = P\left(H \mid \ominus\right) = \frac{\text{TRUE } \ominus}{\text{TOTAL } \ominus} = \frac{75}{76} = 98.68\%$$

```
#1(20) + #2( 0) == #5(20)
#3( 0) + #4(80) == #6(80)
PPV = 1.0
NPV = 1.0
Sensetivity = 20/20 == 1.0
Specifity = 80/80 == 1.0
--------------------------
#1(19) + #2( 0) == #5(19)
#3( 1) + #4(80) == #6(81)
PPV = 1.0
NPV = 0.9876543209876543
Sensetivity = 19/20 == 0.95
Specifity = 80/80 == 1.0
--------------------------
#1(18) + #2( 0) == #5(18)
#3( 2) + #4(80) == #6(82)
PPV = 1.0
NPV = 0.975609756097561
Sensetivity = 18/20 == 0.9
Specifity = 80/80 == 1.0
--------------------------
#1(17) + #2( 0) == #5(17)
#3( 3) + #4(80) == #6(83)
PPV = 1.0
NPV = 0.963855421686747
Sensetivity = 17/20 == 0.85
Specifity = 80/80 == 1.0
--------------------------
#1(16) + #2( 0) == #5(16)
#3( 4) + #4(80) == #6(84)
PPV = 1.0
NPV = 0.9523809523809523
Sensetivity = 16/20 == 0.8
Specifity = 80/80 == 1.0
--------------------------
#1(15) + #2( 0) == #5(15)
#3( 5) + #4(80) == #6(85)
PPV = 1.0
NPV = 0.9411764705882353
Sensetivity = 15/20 == 0.75
Specifity = 80/80 == 1.0
--------------------------
#1(14) + #2( 0) == #5(14)
#3( 6) + #4(80) == #6(86)
PPV = 1.0
NPV = 0.9302325581395349
Sensetivity = 14/20 == 0.7
Specifity = 80/80 == 1.0
--------------------------
#1(13) + #2( 0) == #5(13)
#3( 7) + #4(80) == #6(87)
PPV = 1.0
NPV = 0.9195402298850575
Sensetivity = 13/20 == 0.65
Specifity = 80/80 == 1.0
```

[Prob & Stats - Bayes Theorem (20 of 24) Effects of the Test Results: Example 1 (https://youtu.be/9E2mWYBbsuE)](https://youtu.be/9E2mWYBbsuE)

D	H
TRUE ⊕ 20	FALSE ⊕ 0
FALSE ⊖ 0	TRUE ⊖ 80

D	H
TRUE ⊕ 19	FALSE ⊕ 0
FALSE ⊖ 1	TRUE ⊖ 80

D	H
TRUE ⊕ 18	FALSE ⊕ 0
FALSE ⊖ 2	TRUE ⊖ 80

D	H
TRUE ⊕ 17	FALSE ⊕ 0
FALSE ⊖ 3	TRUE ⊖ 80

Sensitivity	$\frac{20}{20} = 100\%$	$\frac{19}{20} = 95\%$	$\frac{18}{20} = 90\%$	$\frac{17}{20} = 85\%$
Specifity	$\vert$	$\vert$	$\vert$	$\vert$
PPV	$\vert$	$\vert$	$\vert$	$\vert$
NPV	$\vert$	$\frac{80}{81} = 98.76\%$	$\frac{80}{82} = 97.56\%$	$\frac{80}{83} = 96.38\%$

D	H
TRUE ⊕ 16	FALSE ⊕ 0
FALSE ⊖ 4	TRUE ⊖ 80

D	H
TRUE ⊕ 15	FALSE ⊕ 0
FALSE ⊖ 5	TRUE ⊖ 80

D	H
TRUE ⊕ 14	FALSE ⊕ 0
FALSE ⊖ 6	TRUE ⊖ 80

D	H
TRUE ⊕ 13	FALSE ⊕ 0
FALSE ⊖ 7	TRUE ⊖ 80

Sensitivity	$\frac{16}{20} = 80\%$	$\frac{15}{20} = 75\%$	$\frac{14}{20} = 70\%$	$\frac{13}{20} = 65\%$
Specifity	$\vert$	$\vert$	$\vert$	$\vert$
PPV	$\vert$	$\vert$	$\vert$	$\vert$
NPV	$\frac{80}{84} = 95.24\%$	$\frac{80}{85} = 94.12\%$	$\frac{80}{86} = 93.02\%$	$\frac{80}{87} = 91.95\%$

```
#1(20) + #2( 0) == #5(20)
#3( 0) + #4(80) == #6(80)
PPV = 1.0
NPV = 1.0
Sensetivity = 20/20 == 1.0
Specifity = 80/80 == 1.0
--------------------------
#1(19) + #2( 5) == #5(24)
#3( 1) + #4(75) == #6(76)
PPV = 0.7916666666666666
NPV = 0.9868421052631579
Sensetivity = 19/20 == 0.95
Specifity = 75/80 == 0.9375
--------------------------
#1(18) + #2(10) == #5(28)
#3( 2) + #4(70) == #6(72)
PPV = 0.6428571428571429
NPV = 0.9722222222222222
Sensetivity = 18/20 == 0.9
Specifity = 70/80 == 0.875
--------------------------
#1(17) + #2(15) == #5(32)
#3( 3) + #4(65) == #6(68)
PPV = 0.53125
NPV = 0.9558823529411765
Sensetivity = 17/20 == 0.85
Specifity = 65/80 == 0.8125
--------------------------
#1(16) + #2(20) == #5(36)
#3( 4) + #4(60) == #6(64)
PPV = 0.4444444444444444
NPV = 0.9375
Sensetivity = 16/20 == 0.8
Specifity = 60/80 == 0.75
--------------------------
#1(15) + #2(25) == #5(40)
#3( 5) + #4(55) == #6(60)
PPV = 0.375
NPV = 0.9166666666666666
Sensetivity = 15/20 == 0.75
Specifity = 55/80 == 0.6875
--------------------------
#1(14) + #2(30) == #5(44)
#3( 6) + #4(50) == #6(56)
PPV = 0.3181818181818182
NPV = 0.8928571428571429
Sensetivity = 14/20 == 0.7
Specifity = 50/80 == 0.625
--------------------------
#1(13) + #2(35) == #5(48)
#3( 7) + #4(45) == #6(52)
PPV = 0.2708333333333333
NPV = 0.8653846153846154
Sensetivity = 13/20 == 0.65
Specifity = 45/80 == 0.5625
```

[Prob & Stats - Bayes Theorem (21 of 24) Effects of the Test Results: Example 2 (https://youtu.be/rMIL4d64RUg)](https://youtu.be/rMIL4d64RUg)

D	H
TRUE $\oplus$ 20	FALSE $\oplus$ 0
FALSE $\ominus$ 0	TRUE $\ominus$ 80

D	H
TRUE $\oplus$ 19	FALSE $\oplus$ 5
FALSE $\ominus$ 1	TRUE $\ominus$ 75

D	H
TRUE $\oplus$ 18	FALSE $\oplus$ 10
FALSE $\ominus$ 2	TRUE $\ominus$ 70

D	H
TRUE $\oplus$ 17	FALSE $\oplus$ 15
FALSE $\ominus$ 3	TRUE $\ominus$ 65

Sensitivity	$\frac{20}{20} = 100\%$	$\frac{19}{20} = 95\%$	$\frac{18}{20} = 90\%$	$\frac{17}{20} = 85\%$
Specifity	1	$\frac{75}{80} = 93.75\%$	$\frac{70}{80} = 87.5\%$	$\frac{65}{80} = 81.25\%$
PPV	1	$\frac{19}{24} = 79.16\%$	$\frac{18}{28} = 64.29\%$	$\frac{17}{32} = 53.13\%$
NPV	1	$\frac{80}{81} = 98.68\%$	$\frac{80}{82} = 97.22\%$	$\frac{80}{83} = 95.59\%$

D	H
TRUE $\oplus$ 16	FALSE $\oplus$ 20
FALSE $\ominus$ 4	TRUE $\ominus$ 60

D	H
TRUE $\oplus$ 15	FALSE $\oplus$ 25
FALSE $\ominus$ 5	TRUE $\ominus$ 55

D	H
TRUE $\oplus$ 14	FALSE $\oplus$ 30
FALSE $\ominus$ 6	TRUE $\ominus$ 50

D	H
TRUE $\oplus$ 13	FALSE $\oplus$ 35
FALSE $\ominus$ 7	TRUE $\ominus$ 45

Sensitivity	$\frac{16}{20} = 80\%$	$\frac{15}{20} = 75\%$	$\frac{14}{20} = 70\%$	$\frac{13}{20} = 65\%$
Specifity	$\frac{60}{80} = 75\%$	$\frac{55}{80} = 68.75\%$	$\frac{50}{80} = 62.5\%$	$\frac{45}{80} = 56.25\%$
PPV	$\frac{16}{36} = 44.44\%$	$\frac{15}{40} = 37.5\%$	$\frac{14}{44} = 31.82\%$	$\frac{13}{48} = 27.08\%$
NPV	$\frac{80}{84} = 93.75\%$	$\frac{80}{85} = 91.67\%$	$\frac{80}{86} = 89.28\%$	$\frac{80}{87} = 86.54\%$

```
TRUE+ [0.98]  FALSE+[0.04]
FALSE-[0.02]  TRUE- [0.96]
Disease[5000] + Helthy[95000]  == 100000
#1[ 4900] + #2[ 3800] == #5[ 8700]
#3[  100] + #4[91200] == #6[91300]
Sensetivity = 0.98
Specifity = 0.96
PPV = 0.5632183908045975
NPV = 0.9989047097480832
P(D)=0.5632183908045975
P(H)=0.43678160919540254
```

[Prob & Stats - Bayes Theorem (22 of 24) Example of Table Format: Step 1 (https://youtu.be/sNRsq__ol0Y)](https://youtu.be/sNRsq__ol0Y)

Всего в первый раз **100000** тестируемых при качестве тестов:

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **96%** == **SPECIFIC == TRUE Negative**, **4% == FALSE Positive**

- **5%** распространённость болезни **Prevalence** тогда:
- **5000 = D** являются **больными**
- **95000 = H** являются **здоровыми**

	Disease	Healthy	Total
$\oplus$ $+$	$\boxed{1}$ TRUE $\oplus$ $98\% = 4900$	$\boxed{2}$ FALSE $\oplus$ $4\% = 3800$	$\boxed{5}$ TOTAL $\oplus$ 8700
$\ominus$ $-$	$\boxed{3}$ FALSE $\ominus$ $2\% = 100$	$\boxed{4}$ TRUE $\ominus$ $96\% = 91200$	$\boxed{6}$ TOTAL $\ominus$ 91300
	5000	95000	100000

$$PPV = P\left(D \mid \oplus\right) = \frac{\text{TRUE } \oplus}{\text{TOTAL } \oplus} = \frac{4900}{8700} = 56.32\%$$

$$NPV = P\left(H \mid \ominus\right) = \frac{\text{TRUE } \ominus}{\text{TOTAL } \ominus} = \frac{91200}{91300} = 99.89\%$$

```
TRUE+ [0.98]  FALSE+[0.04]
FALSE-[0.02]  TRUE- [0.96]
Disease[5000] + Helthy[95000]  == 100000
#1[ 4900] + #2[ 3800] == #5[ 8700]
#3[  100] + #4[91200] == #6[91300]
Sensetivity =0.98
Specifity =0.96
PPV = 0.5632183908045975
NPV = 0.9989047097480832
P(D)=0.5632183908045975
P(H)=0.43678160919540254
--------------------------------------------
Disease[4900] + Helthy[3800]  == 8700
#1[ 4802] + #2[  152] == #5[ 4954]
#3[   98] + #4[ 3648] == #6[ 3746]
Sensetivity =0.98
Specifity =0.96
PPV = 0.9693177230520791
NPV = 0.9738387613454351
P(D)=0.9693177230520791
P(H)=0.030682276947920917
```

[Prob & Stats - Bayes Theorem (23 of 24) Example of Table Format: Step 2 (https://youtu.be/Ohaw8PG0NpQ)](https://youtu.be/Ohaw8PG0NpQ)

Тест 2 раза

Всего в первый раз **100000** тестируемых при качестве тестов:

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **96%** == **SPECIFIC == TRUE Negative**, **4% == FALSE Positive**

- **5%** распространённость болезни **Prevalence** тогда:
- **5000 = D** являются **больными**
- **95000 = H** являются **здоровыми**

	Disease	**Healthy**	**Total**
$\oplus$ +	$\boxed{1}$ TRUE $\oplus$ 98% = 4900	$\boxed{2}$ FALSE $\oplus$ 4% = 3800	8700
$\ominus$ −	$\boxed{3}$ FALSE $\ominus$ 2% = 100	$\boxed{4}$ TRUE $\ominus$ 96% = 91200	91300
	5000	95000	100000

Disease	**Healthy**	**Total**
$\boxed{1}$ TRUE $\oplus$ 98% = 4802	$\boxed{2}$ FALSE $\oplus$ 4% = 152	4954
$\boxed{3}$ FALSE $\ominus$ 2% = 98	$\boxed{4}$ TRUE $\ominus$ 96% = 3648	3746
4900	3800	8700

Первый тест:

$$PPV = P(D \mid \oplus) = \frac{\text{TRUE } \oplus}{\text{TOTAL } \oplus} = \frac{4900}{8700} = 56.32\%$$

$$NPV = P(H \mid \ominus) = \frac{\text{TRUE } \ominus}{\text{TOTAL } \ominus} = \frac{91200}{91300} = 99.89\%$$

Повторое тестирование тех у кого тест дал $\oplus$:

$$PPV = P(D \mid \oplus) = \frac{\text{TRUE } \oplus}{\text{TOTAL } \oplus} = \frac{4802}{4954} = 96.93\%$$

$$NPV = P(H \mid \ominus) = \frac{\text{TRUE } \ominus}{\text{TOTAL } \ominus} = \frac{3648}{3746} = 97.38\%$$

```
TRUE+ [0.98]  FALSE+[0.04]
FALSE-[0.02]  TRUE- [0.96]
Disease[5000] + Helthy[95000]  == 100000
#1[ 4900] + #2[ 3800] == #5[ 8700]
#3[  100] + #4[91200] == #6[91300]
Sensetivity =0.98
Specifity =0.96
PPV = 0.5632183908045975
NPV = 0.9989047097480832
P(D)=0.5632183908045975
P(H)=0.43678160919540254
--------------------------------------------
Disease[4900] + Helthy[3800]  == 8700
#1[ 4802] + #2[  152] == #5[ 4954]
#3[   98] + #4[ 3648] == #6[ 3746]
Sensetivity =0.98
Specifity =0.96
PPV = 0.9693177230520791
NPV = 0.9738387613454351
P(D)=0.9693177230520791
P(H)=0.030682276947920917
--------------------------------------------
Disease[4802] + Helthy[152]  == 4954
#1[ 4706] + #2[    6] == #5[ 4712]
#3[   96] + #4[  146] == #6[  242]
Sensetivity =0.98
Specifity =0.96
PPV = 0.9987096883727643
NPV = 0.6030748884113077
P(D)=0.9987096883727642
P(H)=0.0012903116272358073
```

[Prob & Stats - Bayes Theorem (24 of 24) Example of Table Format: Step 3 (https://youtu.be/E-QNjj6REXo)](https://youtu.be/E-QNjj6REXo)

Тест 3 разатехктоопределяетсякакбольной

Всего в первый раз **100000** тестируемых при качестве тестов:

- Test **98%** == **SENSITIVE == TRUE Positive**, **2% == FALSE Negative пропущено**
- Test **96%** == **SPECIFIC == TRUE Negative**, **4% == FALSE Positive**

- **5%** распространённость болезни **Prevalence** тогда:
- **5000 = D** являются **больными**
- **95000 = H** являются **здоровыми**

	Disease	Healthy	Total
$\oplus$ +	TRUE $\oplus$ 4900	FALSE $\oplus$ 3800	8700
$\ominus$ −	FALSE $\ominus$ 100	TRUE $\ominus$ 91200	91300
	5000	95000	100000

D	H	Total
4802	152	4954
98	3648	3746
4900	3800	8700

D	H	Total
4706	6	4712
96	146	242
4802	152	4954

Первый тест:

$$PPV = P\left(D \mid \oplus\right) = \frac{\text{TRUE } \oplus}{\text{TOTAL } \oplus} = \frac{4900}{8700} = 56.32\%$$

$$NPV = P\left(H \mid \ominus\right) = \frac{\text{TRUE } \ominus}{\text{TOTAL } \ominus} = \frac{91200}{91300} = 99.89\%$$

Повторое тестирование тех у кого тест дал $\oplus$:

$$PPV = P\left(D \mid \oplus\right) = \frac{\text{TRUE } \oplus}{\text{TOTAL } \oplus} = \frac{4802}{4954} = 96.93\%$$

$$NPV = P\left(H \mid \ominus\right) = \frac{\text{TRUE } \ominus}{\text{TOTAL } \ominus} = \frac{3648}{3746} = 97.38\%$$

Третье тестирование тех у кого тест дал $\oplus$:

$$PPV = P\left(D \mid \oplus\right) = \frac{\text{TRUE } \oplus}{\text{TOTAL } \oplus} = \frac{4706}{4712} = 99.87\%$$

$$NPV = P\left(H \mid \ominus\right) = \frac{\text{TRUE } \ominus}{\text{TOTAL } \ominus} = \frac{146}{242} = 60.31\%$$

$$\begin{aligned}
&\text{Sensitivity= Recall} &&= TPR &&= \tfrac{TP}{TP+FN} \\
&\text{Specificity} &&= TNR &&= 1 - FPR = 1 - \tfrac{FP}{FP+TN} \\
&\text{Precision} &&= PPV &&= \tfrac{TP}{TP+FP}
\end{aligned}$$

Доказательство

Формула Байеса вытекает из определения [условной вероятности (https://ru.wikipedia.org/wiki/%D0%A3%D1%81%D0%BB%D0%BE%D0%B2%D0%BD%D0%B0%D1%8F_%D0%B2%D0%](https://ru.wikipedia.org/wiki/%D0%A3%D1%81%D0%BB%D0%BE%D0%B2%D0%BD%D0%B0%D1%8F_%D0%B2%D0%)

Вероятность совместного события AB двояко выражается через условные вероятности

$$P(AB) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

Следовательно

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B \mid A)\,P(A)}{P(B)}$$

Вычисление $P(B)$

В задачах и [статистических](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0) (https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0
приложениях $P(B)$ обычно вычисляется по [формуле полной вероятности](https://ru.wikipedia.org/wiki/%D0%A4%D0%BE%D1%80%D0%BC%D1%83%D0%BB%D0%B0_%D0%BF%D0%BE%D0%B) (https://ru.wikipedia.org/wiki/%D0%A4%D0%BE%D1%80%D0%BC%D1%83%D0%BB%D0%B0_%D0%BF%D0%BE%D
события, зависящего от нескольких [несовместных](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D1%81%D0%BE%D0%B2%D0%BC%D0%B5%D1%81%D1%82%D0) (https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D1%81%D0%BE%D0%B2%D0%BC%D0%B5%D1%81%D1%82%D0
гипотез, имеющих суммарную вероятность 1.:

$$P(B) = \sum_{i=1}^{N} P(A_i)P(B \mid A_i)$$

где вероятности под знаком суммы известны или допускают экспериментальную оценку. В этом случае формула Байеса записывается так:

$$P(A_j \mid B) = \frac{P(A_j)P(B \mid A_j)}{\sum_{i=1}^{N} P(A_i)P(B \mid A_i)}$$

Type *Markdown* and LaTeX: α^2

Теорема Байеса

В основе NBC (Naïve Bayes Classifier) лежит, как вы уже могли догадаться, теорема Байеса. где,

$$P(c|d) = \frac{P(d|c) \cdot P(c)}{P(d)}$$

- $P(c|d)$ — вероятность что документ принадлежит классу , именно её нам надо рассчитать;
- $P(d|c)$— вероятность встретить документ среди всех документов класса ;
- $P(c)$ — безусловная вероятность встретить документ класса в корпусе документов;
- $P(d)$ — безусловная вероятность документа в корпусе документов.

Смысл на обывательском уровне можно выразить следующим образом. Теорема Байеса позволяет переставить местами причину и следствие. Зная с какой вероятностью причина приводит к некоему событию, эта теорема позволяет расчитать вероятность того что именно эта причина привела к наблюдаемому событию. Цель классификации состоит в том чтобы понять к какому классу принадлежит документ, поэтому нам нужна не сама вероятность, а наиболее вероятный класс.

Байесовский классификатор использует оценку апостериорного максимума (Maximum a posteriori estimation) для определения наиболее вероятного класса. Грубо говоря, это класс с максимальной вероятностью.

$$c_{map} = \operatorname{arg\,max}_{c \in C} \frac{P(d|c) \cdot P(c)}{P(d)}$$

То есть нам надо рассчитать вероятность для всех классов и выбрать тот класс, который обладает максимальной вероятностью. Обратите внимание, знаменатель (вероятность документа) является константой и никак не может повлиять на ранжирование классов, поэтому в нашей задаче мы можем его игнорировать.

$$c_{map} = \texttt{arg max}_{c \in C} P(d|c) \cdot P(c)$$

Формула №1

Далее делается допущение которое и объясняет почему этот алгоритм называют наивным.

Предположение условной независимости

Байесовский же классификатор представляет документ как набор слов вероятности которых условно не зависят друг от друга. Этот подход иногда еще называется bag of words model. Исходя из этого предположения условная вероятность документа аппроксимируется произведением условных вероятностей всех слов входящих в документ.

$$P(d|c) \approx P(w_1|c) \cdot P(w_2|c) \cdots P(w_n|c) = \prod_{i=1}^{n} P(w_i|c)$$

Этот подход также называется Unigram Language Model. Языковые модели играют очень важную роль в задачах обработки натуральных языков, но выходят за пределы этой заметки. Подставив полученное выражение в формулу №1 мы получим:

$$c_{map} = \texttt{arg max}_{c \in C} \left[P(c) \cdot \prod_{i=1}^{n} P(w_i|c) \right]$$

Проблема арифметического переполнения

При достаточно большой длине документа придется перемножать большое количество очень маленьких чисел. Для того чтобы при этом избежать арифметического переполнения снизу зачастую пользуются свойством логарифма произведения $\texttt{log}(ab) = \texttt{log}\ a + \texttt{log}\ b$. Так как логарифм функция монотонная, ее применение к обоим частям выражения изменит только его численное значение, но не параметры при которых достигается максимум. При этом, логарифм от числа близкого к нулю будет числом отрицательным, но в абсолютном значении существенно большим чем исходное число, что делает логарифмические значения вероятностей более удобными для анализа. Поэтому, мы переписываем нашу формулу с использованием логарифма.

$$c_{map} = \texttt{arg max}_{c \in C} \left[\texttt{log}\ P(c) + \sum_{i=1}^{n} \texttt{log}\ P(w_i|c) \right]$$

Формула №2

Основание логарифма в данном случае не имеет значения. Вы можете использовать как натуральный, так и любой другой логарифм.

Оценка параметров Байесовской модели

Оценка вероятностей $P(c)$ и $P(w_i|c)$ осуществляется на обучающей выборке. Вероятность класса мы можем оценить как:

$$P(c) = \frac{D_c}{D}$$

где, D_c – количество документов принадлежащих классу c , а D – общее количество документов в обучающей выборке.

Оценка вероятности слова в классе может делаться несколькими путями. Здесь я приведу multinomial bayes model.

$$P(w_i|c) = \frac{W_{ic}}{\sum_{i' \in V} W_{i'c}}$$

Формула №3

- W_{ic} — количество раз сколько i-ое слово встречается в документах класса c ;
- V — словарь корпуса документов (список всех уникальных слов).

Другими словами, числитель описывает сколько раз слово встречается в документах класса (включая повторы), а знаменатель – это суммарное количество слов во всех документах этого класса.

Проблема неизвестных слов

С формулой №3 есть одна небольшая проблема. Если на этапе классификации вам встретится слово которого вы не видели на этапе обучения, то значения W_{ic} , а следственно и $P(w_i|c)$ будут равны нулю. *Это приведет к тому что документ с этим словом нельзя будет классифицировать, так как он будет иметь нулевую вероятность по всем классам*. Избавиться от этой проблемы путем анализа большего количества документов не получится. Вы никогда не сможете составить обучающую выборку содержащую все возможные слова включая неологизмы, опечатки, синонимы и т.д. Типичным решением проблемы неизвестных слов является аддитивное сглаживание (сглаживание Лапласа). Идея заключается в том что мы притворяемся как будто видели каждое слово на один раз больше, то есть прибавляем единицу к частоте каждого слова.

$$P(w_i|c) = \frac{W_{ic} + 1}{\sum_{i' \in V}(W_{i'c} + 1)} = \frac{W_{ic} + 1}{|V| + \sum_{i' \in V} W_{i'c}}$$

Логически данный подход смещает оценку вероятностей в сторону менее вероятных исходов. Таким образом, слова которые мы не видели на этапе обучения модели получают пусть маленькую, но все же не нулевую вероятность.

Собираем все вместе

Подставив выбранные нами оценки в формулу №2 мы получаем окончательную формулу по которой происходит байесовская классификация.

$$c_{map} = \operatorname{arg\,max}_{c \in C} \left[\log \frac{D_c}{D} + \sum_{i=1}^{n} \log \frac{W_{ic} + 1}{|V| + \sum_{i' \in V} W_{i'c}} \right]$$

Формула №4

Реализация классификатора

Для реализации Байесовского классификатора нам необходима обучающая выборка в которой проставлены соответствия между текстовыми документами и их классами. Затем нам необходимо собрать следующую статистику из выборки, которая будет использоваться на этапе классификации:

- относительные частоты классов в корпусе документов. То есть, как часто встречаются документы того или иного класса;
- суммарное количество слов в документах каждого класса;
- относительные частоты слов в пределах каждого класса;
- размер словаря выборки. Количество уникальных слов в выборке.

Совокупность этой информации мы будем называть моделью классификатора. Затем на этапе классификации необходимо для каждого класса рассчитать значение следующего выражения и выбрать класс с максимальным значением.

$$\log \frac{D_c}{D} + \sum_{i \in Q} \log \frac{W_{ic} + 1}{|V| + L_c}$$

Упрощенная запись формулы №4

в этой формуле:

- D_c — количество документов в обучающей выборке принадлежащих классу ;
- D — общее количество документов в обучающей выборке;
- $|V|$ — количество уникальных слов во всех документах обучающей выборки;
- L_c — суммарное количество слов в документах класса в обучающей выборке;
- W_{ic}— сколько раз i-ое слово встречалось в документах класса c в обучающей выборке;
- Q– множество слов классифицируемого документа (включая повторы).

Пример

Допустим, у нас есть три документа для которых известны их классы (HAM означает – не спам):

- [SPAM] предоставляю услуги бухгалтера;
- [SPAM] спешите купить виагру;
- [HAM] надо купить молоко.

Модель классификатора будет выглядеть следующим образом:

	spam	**ham**
частоты классов	2	1
суммарное количество слов	6	3

	spam	**ham**
предоставляю	1	0
услуги	1	0
бухгалтера	1	0
спешите	1	0
купить	1	1
виагру	1	0

	spam	ham
надо	0	1
молоко	0	1

Теперь классифицируем фразу "**надо купить сигареты**". Рассчитаем значение выражения для класса SPAM:
$log\frac{D_c}{D} + \sum log\frac{W_{ic}+1}{|V|+L_c} =$

$log\frac{\text{всего классов SPAM(2)}}{\text{всего примеров(3)}} +$

$log\frac{\text{надо(0+1)}}{\text{всего слов в словаре(8)+количество всех слов в SPAM(6)}} + log\frac{\text{купить(1+1)}}{\text{всего слов в словаре(8)+количество всех слов в SPAM(6)}}$
$+ log\frac{\text{сигареты(0+1)}}{\text{всего слов в словаре(8)+количество всех слов в SPAM(6)}} = log\frac{2}{3} + log\frac{1}{8+6} + log\frac{2}{8+6} + log\frac{1}{8+6} \approx -7.629$

```
-7.629489916393996
```

Теперь сделаем то же самое для класса HAM:

$log\frac{\text{всего классов HAM(1)}}{\text{всего примеров(3)}} +$

$log\frac{\text{надо(1+1)}}{\text{всего слов в словаре(8)+количество всех слов в HAM(3)}} + log\frac{\text{купить(1+1)}}{\text{всего слов в словаре(8)+количество всех слов в HAM(3)}}$
$+ log\frac{\text{сигареты(0+1)}}{\text{всего слов в словаре(8)+количество всех слов в HAM(3)}} = log\frac{1}{3} + log\frac{2}{8+3} + log\frac{2}{8+3} + log\frac{1}{8+3} \approx -6.906$

```
-6.906003745943331
```

Формирование вероятностного пространства

В простейшем случае вы выбираете класс который получил максимальную оценку. Но если вы например хотите помечать сообщение как спам только если соответствующая вероятность больше 80%, то сравнение логарифмических оценок вам ничего не даст. Оценки которые выдает алгоритм не удовлетворяют двум формальным свойствам которым должны удовлетворять все вероятностные оценки:

- они все должны быть в диапазоне от нуля до единицы;
- их сумма должна быть равна единице.

Для того чтобы решить эту задачу, необходимо из логарифмических оценок сформировать вероятностное пространство. А именно: избавиться от логарифмов и нормировать сумму по единице.

Здесь q_c — это логарфмическая оценка алгоритма для класса c , а возведение e(основание натурального логарфима) в степерь оценки используется для того чтобы избавиться от логарифма ($a^{log_a x} = x$). Таким образом, если вы в рассчетах использовали не натуральный логарифм, а десятичный, вам необходимо использовать не e , а 10.

Если сократить экспоненту оценки текущего класса в числителе и знаменателе, то в общем виде получим:

$$P(c|d) = \frac{1}{1 + \sum_{c' \in C ⍁} e^{q_{c'} - q_c}}$$

Обратите внимание, что сумма в знаменателе выполняется только по классам отличным от того для которого мы считаем вероятность. Но в каждом из слагаемых присутствует логарифмическая оценка оцениваемого класса.

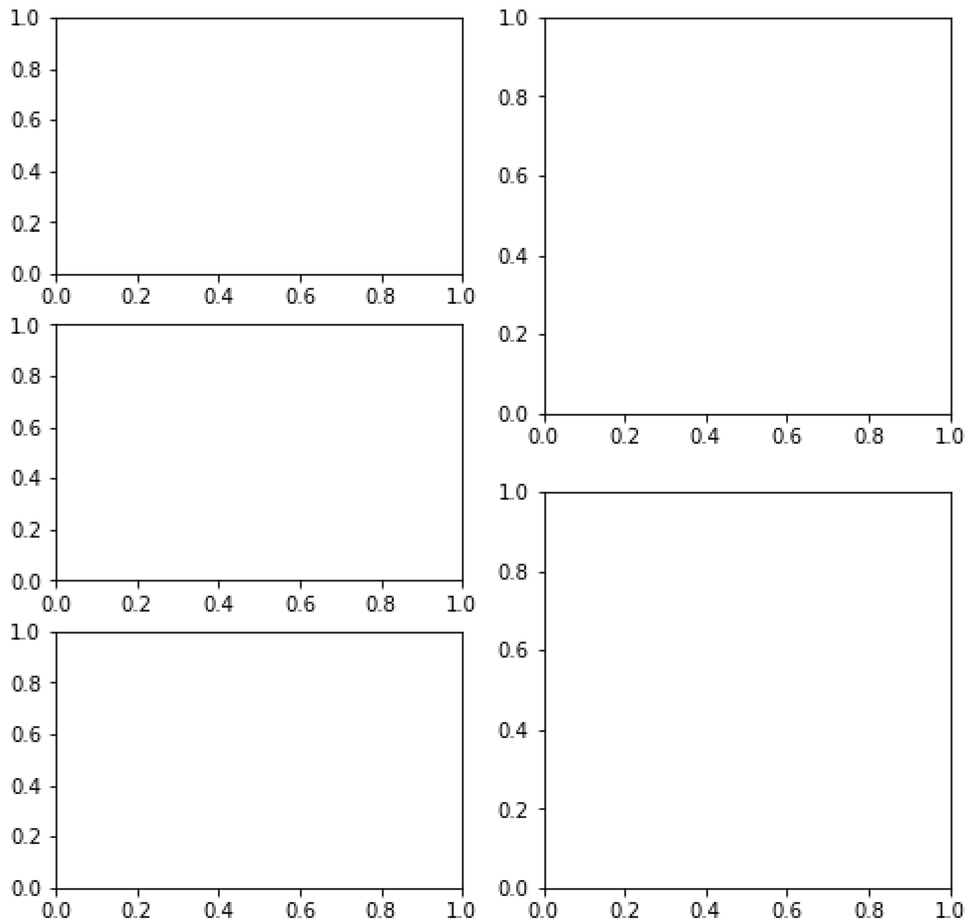
Для вышеприведенного примера вероятность что сообщение спам равно:

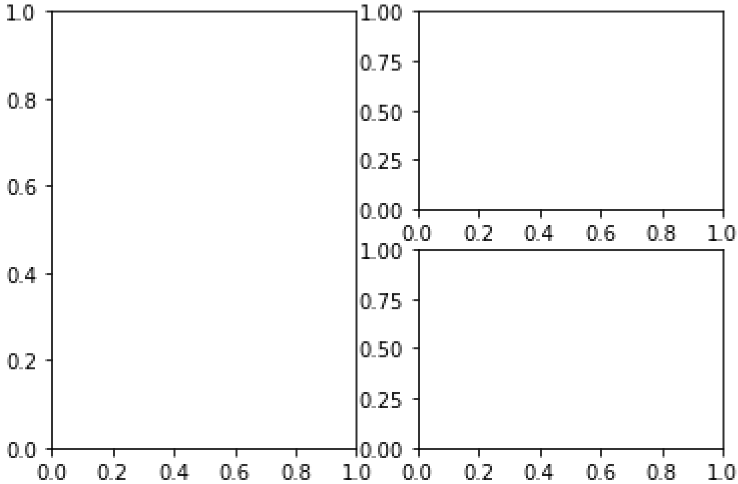
$$\frac{e^{-7.629}}{e^{-7.629} + e^{-6.9606}} = \frac{1}{1 + e^{-6.9606+7.629}} = 0.327$$

① ② ③ ④

```
0.32662576687116557
```

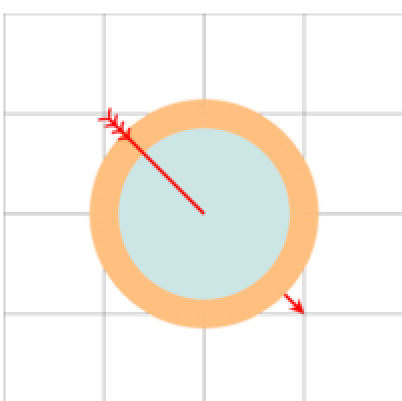
```
<matplotlib.axes._subplots.AxesSubplot at 0x207127adfa0>
```

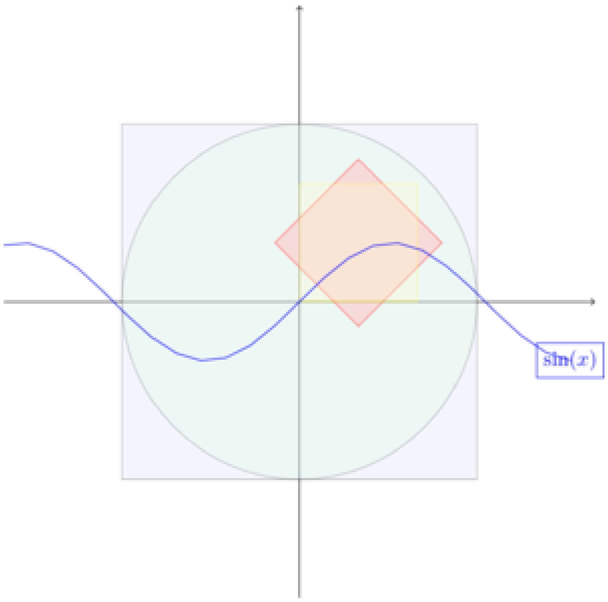


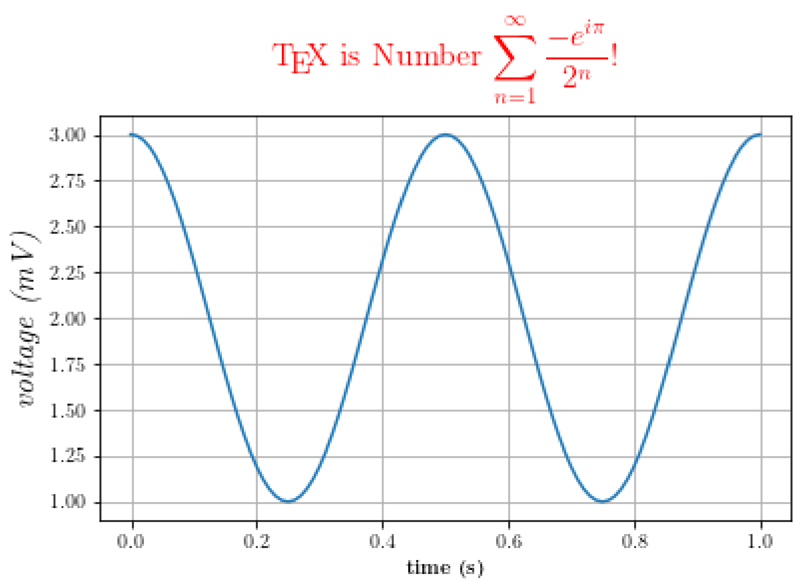



```
3.0
[nan  1.   2.]
3.0
```

```
<ipython-input-30-61ac654271b4>:19: RuntimeWarning: invalid value encountered in l
og2
  print(np.log2([-1, 2, 4]))
```







$$\text{Text here} \begin{cases} \text{Some text1} \\ \text{Some text2} \\ \text{Some text3} \end{cases}$$

Everything hide

`Markdown not work`

Click here to toggle on/off the raw code.