iClicker for Conditional Probability

Lec˙33˙iClicker˙conditionalProbability

Quick definition of probability p

- Example, Bio-360 students:

$$p = \frac{\text{3 Juniors}}{\text{50 } \textbf{Total} \text{ students in Bio-360}} = \frac{6}{100} = 0.06 = 6\%$$

- **Probability p is** **fraction of total**
  - Can use raw fraction (0 to 1)
  - Can use percent (0% to 100%)

- In words: Probability is the frequency of an event
  - "If you pick a random student from Bio-360, the probability of finding a Junior is 0.06 or 6%"
  - "If you flip a coin, the probability of getting heads is 0.5 or 50%."

Disease frequency in population

Disease	**Healthy**	**Total**
10	990	$10 + 990 = 1000$

iClicker:
What is the probability of having the disease?

A $p = 10 = 10\%$

B $p = \dfrac{10}{10 + 990} = 0.01 = 1\%$

C $p = \dfrac{10}{990} = 0.010101010101 = 1.0101010101\%$

D $p = 10 + 990 = 1000$

What's the probability of being healthy?

Same data:

Disease	**Healthy**	**Total**
10	990	$10 + 990 = 1000$

iClicker: probability of being **healthy**?

A $p = \dfrac{990}{10 + 990} = 0.99 = 99\%$

B $p = 10 + 990 = 1000$

C $p = \dfrac{10}{990} = 0.010101010101 = 1.0101010101\%$

D $p = 990$

Major medical breakthrough: Test for the disease

Numbers represent test results in 1000 patients.

	Disease	**Healthy**	**Total**
Test positive	8	99	107
Test negative	2	891	893
Total	10	990	1000

Ignoring testing, what's the probability of having the disease?

(A) $p = 8 + 2 = 10$

(B) $p = \dfrac{10}{10 + 990} = 1\%$

(C) $p = \dfrac{10}{990} = 1.01\%$

(D) $p = \dfrac{8}{1000} = 0.8\%$

Test isn't perfect: **False Positive** can happen

	Disease	**Healthy**	**Total**
Test positive	8	99	107
Test negative	2	891	893
Total	10	990	1000

How many **False Positive** in this table?

A 8

B 2

C 99

D 891

You test positive, what's the probability of having the disease?

	Disease	**Healthy**	**Total**
Test positive	8	99	107
Test negative	2	891	893
Total	10	990	1000

Doctor says

test is **positive**. OMG!!!

Big Hint: **Ignore the negative tests.**

Ask: **Of the positive tests, what proportion have the disease?**

(A) $\frac{8}{8+2}$

(B) $\frac{8}{8+99}$

(C) $\frac{8}{2+99}$

(D) $\frac{8}{99+891}$

So why are there so many **False Positive**?

	Disease	**Healthy**	**Total**
Test positive	8	99	107
Test negative	2	891	893
Total	10	990	1000

- If you have the disease, test is **True Positive** in $\frac{8}{8+2} = \frac{8}{10} = 80\%$. Seems pretty good?
- If you are healthy, test is **False Positive** in $\frac{99}{8+99} = \frac{99}{107} = 93\%!!!$
- How do you reconcile these?

We've seen similar bioinformatics situations with **False Positive**

Searching PubMed:

- Similar to disease test.
- **True Positive** rate is pretty good:
  - A good query finds many relevant documents
- **False Positive** count can be high
  - PubMed is huge ($\approx 27 \times 10^6$)
  - Suppose 1% of the documents incorrectly match query: $1\% \times (27 \times 10^6) = 270{,}000$ false positives!!

Searching with BLAST:

- We can get non-homologous sequences returned
- E-value tells us about sequence similarity and false positives

How's all this relate to E-values from BLAST?

	Homologous	**Not homologous**
BLAST hit	True Positives	**E-value = estimate of False Positives with same or better alignment scores**
BLAST negative	False Negatives	True Negatives

Discussion question:

How could we experimentally measure BLAST's True Positive and True Negative rates?

Another example: taking a school test

	Studied	**Didn't study**	**Total**
Passed	17	3	20
Failed	2	23	25
Total	19	26	45

What's the probability of passing?

A $p = \frac{17}{45}$

B $p = \frac{17+3}{45}$

C $p = \frac{17+3}{17+45}$

D $p = \frac{17+2}{45}$

What's the probability of passing if you study?

	Studied	**Didn't study**	**Total**
Passed	17	3	20
Failed	2	23	25
Total	19	26	45

If you study, what's the probability of passing?
Hint: Only consider students who study.

(A) $p = \frac{17}{45}$

(B) $p = \frac{17}{3 + 23}$

(C) $p = \frac{17}{17 + 3}$

(D) $p = \frac{17}{17 + 2}$

If you passed, what's the probability that you studied?

	Studied	**Didn't study**	**Total**
Passed	17	3	20
Failed	2	23	25
Total	19	26	45

If you passed, what's the probability that you studied?

(A) $p = \frac{17}{45}$

(B) $p = \frac{17}{3 + 23}$

(C) $p = \frac{17}{17 + 3}$

(D) $p = \frac{17}{17 + 2}$

For folks who've studied conditional probability: Today's exercise is just another view of Bayes theorem

	Studied	**Didn't study**	**Total**
Passed	17	3	20
Failed	2	23	25
Total	19	26	45

$$\Pr(\textit{Studied} \mid \textit{Passed}) = \frac{\Pr(\textit{Passed} \mid \textit{Studied})\ \Pr(\textit{Studied})}{\Pr(\textit{Passed})}$$

$$= \frac{\left(\frac{17}{19}\right)\left(\frac{19}{45}\right)}{\left(\frac{20}{45}\right)} = \frac{\left(\frac{17}{\cancel{19}}\right)\left(\frac{\cancel{19}}{\cancel{45}}\right)}{\left(\frac{20}{\cancel{45}}\right)} = \frac{17}{20}$$

$$= 85\%$$