Data, Randomness and Probability

September 11, 2022

Instructor and TA's (by appointments only)

Denis HY LEUNG SOE 5047

denisleung@smu.edu.sg
office hours: Tuesday 2 - 5pm

TA's SOE/SCIS2 GSR 3-16

Bhavika Agrawal (G1)
bhavikaa.2021@economics.smu.edu.sg
Thursday 3:30 - 6:30pm

Lim Fang Qi (G2)
fangqi.lim.2021@economics.smu.edu.sg
Thursday 12 - 3pm

Bharat Gangwani (G3)
bharatg.2020@economics.smu.edu.sg
Wednesday 4 - 7pm

Essentials

- Course webpage: `http://economics.smu.edu.sg/faculty/profile/9699/Denis%20LEUNG` (NOT eLearn!)
- Understanding of basic Calculus and Algebra – Appendix in course notes
- Readings *before* each class
- Projects vs Homework
- If you missed a class, it is **your** responsibility to find out what you have missed from your classmates or course webpage
- **Do not copy down/memorise formulae blindly. Discard as many formulae as possible as you progress**

Projects

(1) There will be 2 group projects

(2) For project administration (submission of reports, *etc.*), each class is assigned a TA, shown previously. However, any TA may be approached for general consultation

(3) You will work with the same group of fellow students on both projects

(4) Those who wish to be in the same group should submit ONE email with their names to the TA before the end of Week 4

(5) Each group submits a schedule of their project meetings (online or in person) to the TA by Week 7

(6) A group member may miss 1 meeting for each project. If any group member misses more than 1 meeting, his/her project grade as well as class participation grade will be pro-rated by # meetings attended/total # meetings

(7) Anyone not contactable by email/phone/etc after 3 attempts from their group members will be considered to have agreed on dates/times of meetings; subsequent absence of such individual from meetings follow the same grading guidelines as (6)

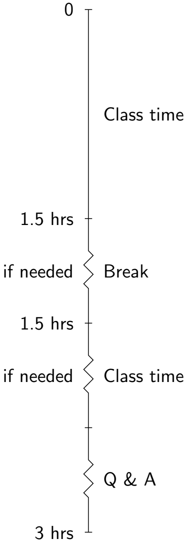
(8) Project reports must be type written in google doc with time stamp indicating each group member's contribution

(9) Each group receives one grade, notwithstanding (6)-(8) above. I reserve the right to ask any individual(s) to submit separate report(s)

Assessments

- Class Participation (10%)
- Projects (40%)
  - 2 projects with presentation 20% each
  - Each project's grade includes 8% individual assessment (quizzes)
- Exam (50%)
  - Closed book but one 2-sided A-4 "cheat sheet" is allowed

Timeline of classes



Data (Women's wage data)

University of Michigan Panel Study of Income Dynamics on 753 white married women in the US (1975-76):

Woman	Workforce status (1=Yes, 0=No)	Hrs worked	#kids $<$ 6 yrs	Age	Education (yrs)	Hourly wage rate	Husband's wage rate	Experience (yrs)
1	1	1610	1	32	12	3.3540	4.0288	14
2	1	1656	0	30	12	1.3889	8.4416	5
3	1	1980	1	35	12	4.5455	3.5807	15
4	1	456	0	34	12	1.0965	3.5417	6
5	1	1568	1	31	14	4.5918	10.0000	7
6	1	2032	0	54	12	4.7421	6.7106	33
7	1	1440	0	37	16	8.3333	3.4277	11
8	1	1020	0	54	12	7.8431	2.5485	35
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
750	0	0	2	31	12	0.0000	4.8638	14
751	0	0	0	43	12	0.0000	1.0898	4
752	0	0	0	60	12	0.0000	12.4400	15
753	0	0	0	39	9	0.0000	6.0897	12

Sample *vs.* Population

(a) Data are a **sample** (subset) from a **population** that we want to study

e.g., 753 women (sample) out of all white married women in 1975-1976 (population)

(b) We are interested in some characteristics of the population

e.g., average wage or percentage of women who earned more than minimum wage in the population

(c) We use a sample to answer questions about the population

(d) Data = Sample

Data structure and terminologies

- Sample size - n: number of units (observations) in the sample
- **Variables** - characteristics of the units

  e.g., Workforce status, hrs worked, age, wage rate, *etc.*
  - Often represented by symbols, X, Y, Z, *etc.*
  - **Quantitative**: numeric
    eg., Age, wage rate, hrs worked
  - **Qualitative (Categorical)**: Not quantitative (no natural ordering)
    eg., Gender, colour, race
  - **Discrete**: countable number of values
    eg., Gender, # kids, # days
  - **Continuous**: uncountably many in a range (a, b)
    eg., Wage rate, age (real, not rounded), temperature
- A sample is n observations, $X_1, ..., X_n$, of X

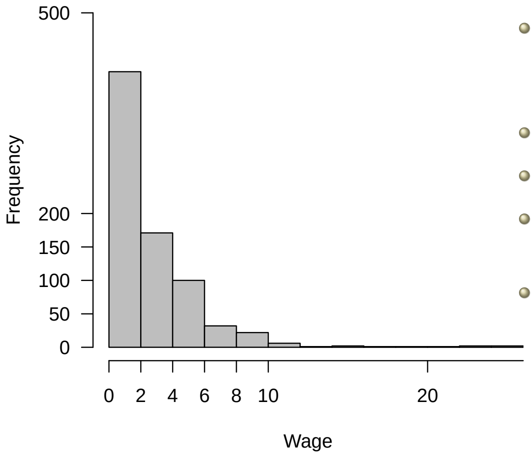
Summarising qualitative (categorical) or quantitative data with a few values

Variable	Levels	n	%
in workforce	0	325	43.2
	1	428	56.8
	all	753	100.0
# kids < 6	0	606	80.5
	1	118	15.7
	2	26	3.4
	3	3	0.4
	all	753	100.0
education	5	4	0.5
	6	6	0.8
	7	8	1.1
	8	30	4.0
	9	25	3.3
	10	44	5.8
	11	43	5.7
	12	381	50.6
	13	44	5.8
	14	51	6.8
	15	14	1.9
	16	57	7.6
	17	46	6.1
	all	753	100.0

- **frequency distribution**
  - tells us everything about a categorical variable
  - gives # observations within each category/level
- Most easily displayed using a **table**
- Proportions or percentage in each category or level, *eg.*,

$$\frac{325}{753} = \frac{325}{753} \times 100 \text{ percent} \approx 43.2 \text{ percent}$$

Summarising quantitative data (continuous or discrete with many values



- **Histogram** is a useful **graphical** summary for quantitative data
- Groups observations into **bins**
- **Bin width** defines grouping
- Height (area) of bin proportional to group size
- Most women earned $<$ \$2 an hour; few earned $>$ \$10

Summarising quantitative data – average

- **Numerical summaries**
- Sample **mean**, **median** – measure "typical" or "average" value of the data

  eg., $X_1, \dots, X_{10} = 1, 1, 4, 2, 5, 2, 2, 3, 3, 4$

(a) Sample mean

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1 + 1 + 4 + \dots + 3 + 4}{10} = 2.7$$

(b) Sample median - "middle" observation when data are ranked from lowest to highest. If $n =$ odd, sample median$=$ middle value. If $n =$ even, sample median $=$ the average of the two middle values

1,1,2,2, 2,3 ,3,4,4,5

$$\frac{2+3}{2} = 2.5$$

Summarising quantitative data – average (2)

Add an observation

$X_1, ..., X_{11} = 1, 1, 4, 2, 5, 2, 2, 3, 3, 4, \boxed{50}$

(a) Sample mean

$$\bar{X} = \frac{1+1+4+...+3+4+\boxed{50}}{11} = 7.7$$

(b) Sample median $= 3$

1,1,2,2,2, $\boxed{3}$,3,4,4,5,50

(c) Mean is sensitive to the change but median is not

- Mean uses 1,1,4,2,5,2,2,3,3,4,50
  Median uses 3
- Mean uses information from every observation
- Mean not representative of the average when there are extremes
- Mean better represents the average when there are no extremes

Summarising quantitative data – spread

- Spread describes how the value of a variable changes over n observations
- Sample **range**, **variance** (s^2), **interquartile range (IQR)**

eg. $X_1, \ldots, X_{10} = 1, 1, 4, 2, 5, 2, 2, 3, 3, 4$

(a) Sample range $=$ largest $-$ smallest $= 5 - 1 = 4$

(b) Sample variance $=$ average "distance" between observations and $\bar{X}$

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1^a} = \frac{(1 - 2.7)^2 + \ldots + (4 - 2.7)^2}{10 - 1} \approx 1.79$$

Taking square root gives **standard deviation** (s)

(c) IQR$=$ $\overbrace{\text{upper}}^{\text{top}}$ $\overbrace{\text{quartile}}^{25\%}$ (75-th percentile)$-$ $\overbrace{\text{lower}}^{\text{bottom}}$ $\overbrace{\text{quartile}}^{25\%}$ (25-th percentile)

$$1,1,\overbrace{\boxed{}}^{25\%}2,2,2,3,3,\overbrace{\boxed{}}^{75\%}4,4,5 \qquad \text{IQR} = \underbrace{\frac{3+4}{2}}_{\text{top } 25\%} - \underbrace{\frac{1+2}{2}}_{\text{bottom } 25\%} = 3.5 - 1.5 = 2$$

---

[a] Alternatively use n

Summarising quantitative data – spread (2)

Range uses	1 5
s^2 and s use	1,1,4,2,5,2,2,3,3,4
IQR uses	1,2 3,4

- s^2 (s) uses information from every observation
- Range uses only the *extreme* observations
- s^2 (s) and range not representative of the spread when there are extremes
- s^2 (s) is better than range to represent the spread when there are no extremes since s^2 (s) uses more information
- When there are extremes, IQR is the best because it ignores the extremes

Randomness

Pandemic data

Treatment outcome from $n = 100$ patients in a pandemic:
1 = "recovered" and 0 = "not recovered"

1 1 1 0 0 0 1 1 1 0 0 1 0 1 0 0 1 1 1 1
1 1 1 0 1 1 0 0 1 1 1 1 0 1 1 0 0 1 1 0
0 1 0 1 1 0 1 0 0 1 1 1 1 0 1 0 1 1 1 0
1 0 1 0 1 1 0 0 0 0 1 1 1 0 1 0 0 0 1 1
1 0 1 1 1 1 0 0 1 1 1 1 0 1 0 0 1 0 1 1

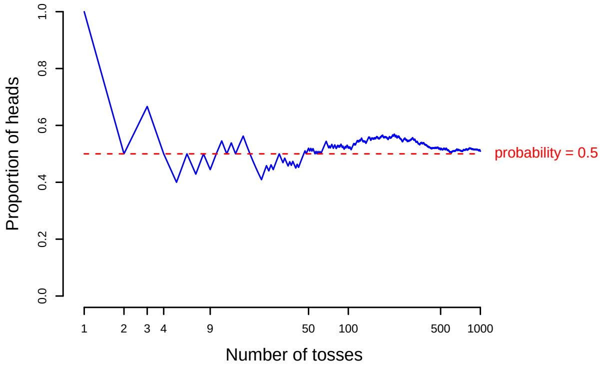
Outcome	0	1
n	40	60

- Why did some patients recovered and others not? pattern of 1 and 0's not easy to predict – **random**
- **Probability** helps explain randomness

Definition of probability - Tossing a fair coin

A fair coin has a $\frac{1}{2}$ "probability" of observing heads, what does it mean?

Toss	1	2	3	4	5	6	7	8	9	$\cdots$
Outcome	H	T	H	T	T	H	T	H	T	$\cdots$



The long run proportion (frequency) of heads is the probability of heads

Insight from coin tossing experiment

- Probability is the long run frequency of an outcome
- Probability cannot predict individual outcomes
- However, it can be used to predict long run trends
- Probability always lies between 0 and 1, with a value closer to 1 meaning a higher frequency of occurrence
- Probability is numeric in value so we can use it to:

  - compare the relative chance between different outcomes (events)

  - carry out calculations

Proportion and probability

- Toss of fair coin: H, T, H, T, T, H, T, H, T

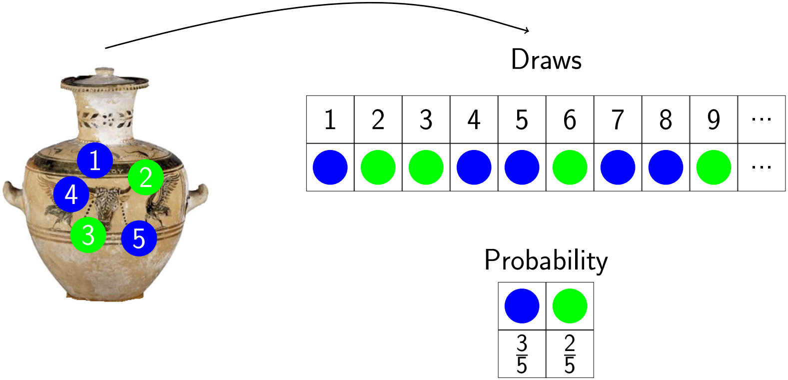
Outcome	H	T
Sample Proportion	$4/9$	$5/9$
Probability	$1/2$	$1/2$

- Treatment outcome

Outcome	0	1
Sample Proportion	$40/100$	$60/100$
Probability	$\mathrm{P}(0)$	$\mathrm{P}(1)$

- Probabilities are *population* proportions

Probability Axioms - Urn model (1)- drawing marbles from an urn (with replacement)



Urn model (2)

- Five possible **Outcomes**: 1 2 3 4 5
- Interested in **Event** A:
- A={ 1 4 5 }; hence an event is a collection of outcomes
- $\mathrm{P}(A) = \frac{3}{5} = 0.6(60\%) = \frac{\text{Number of marbles in } A}{\text{Total number of marbles}}$

Complementary events

- Marbles in urn: 1 2 3 4 5
- Interested in $\bar{A}$: (Not A)
- $\bar{A}$={ 2 3 }
- $\bar{A}$, sometimes written as A^C, is called the **complementary** event of A
- Chance of $=1-$ chance of

$$\Rightarrow \mathrm{P}(\bar{A}) = 1 - \mathrm{P}(A) = 1 - \frac{3}{5} = \frac{2}{5}$$

Joint probability and disjoint events

- A **joint probability** between two events A and B is one of the following:

$$\mathrm{P}(A \text{ and } B) \qquad \mathrm{P}(A \cap B) \qquad \mathrm{P}(AB)$$

- The simplest form of joint probability between A and B is when they are **disjoint** (also called **mutually exclusive**). Disjoint events cannot occur simultaneously: $\mathrm{P}(A \text{ and } B){=}0$

Example

$$A = \{{\color{blue}\bullet} \text{ in 1st draw}\}$$
$$B = \{{\color{green}\bullet} \text{ in 1st draw}\}$$

$$\mathrm{P}(A \text{ and } B) = 0$$

Partition rule

- For any A, a partition is any collection of disjoint subsets of A that together make up A, *eg.*,

  If A= {all white women}, a partition of A is A_1 ={all white women with wage < 3} and A_2 ={all white women with wage ≥ 3}

- For any A, $\mathrm{P}(A)$ can be written as the sum of the probabilities of its disjoint subsets, *eg.*,

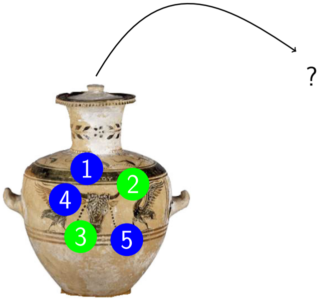
$$
\begin{aligned}
\mathrm{P}(\bullet) &= \mathrm{P}(\bullet \text{ and odd}) + \mathrm{P}(\bullet \text{ and even}) \\
&= \mathrm{P}(\{1 \quad 5\}) + \mathrm{P}(\{4\}) \\
&= \frac{2}{5} + \frac{1}{5} \\
&= \frac{3}{5}
\end{aligned}
$$

Conditional probability

Conditional probability is a useful quantification of how the assessment of chance changed due to new information: "If A happened, what is the chance of B?"

The conditional probability of "B given A" is written as $\mathrm{P}(B|A)$

Example Drawing marbles WITHOUT replacement



$$A = \{● \text{ is drawn}\}$$
$$B = \{● \text{ is drawn}\}$$
$$\mathrm{P}(B) = \frac{2}{5}$$
$$\mathrm{P}(B|A) = \frac{2}{4}$$

Joint probability and the multiplication rule

- The most general way of calculating joint probability is the **Multiplication Rule**

$$\mathrm{P}(AB) = \mathrm{P}(B|A)\mathrm{P}(A) = \mathrm{P}(A|B)\mathrm{P}(B)$$

Example

Marbles in urn: 1 2 3 4 5

$$\begin{aligned}\mathrm{P}(\textcolor{blue}{\bullet}\text{ and even}) &= \mathrm{P}(\text{even}|\textcolor{blue}{\bullet})\mathrm{P}(\textcolor{blue}{\bullet}) = \left(\frac{1}{3}\right)\left(\frac{3}{5}\right) = \frac{1}{5} \\ &= \mathrm{P}(\textcolor{blue}{\bullet}|\text{even})\mathrm{P}(\text{even}) = \left(\frac{1}{2}\right)\left(\frac{2}{5}\right) = \frac{1}{5}\end{aligned}$$

- Rearranging the multiplication rule:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(AB)}{\mathrm{P}(B)} \quad \text{and} \quad \mathrm{P}(B|A) = \frac{\mathrm{P}(AB)}{\mathrm{P}(A)}$$

Independence

- A and B are **independent** means they don't offer information about one another

- If A and B are independent, conditional probability becomes unconditional:

  (i) $\mathrm{P}(A|B) = \mathrm{P}(A)$ "B says nothing about A"

  (ii) $\mathrm{P}(B|A) = \mathrm{P}(B)$ "A says nothing about B"

- Independence is NOT the same as mutually exclusive (disjoint), which is $\mathrm{P}(A \text{ and } B) = 0$. In fact when A and B are disjoint, they are very dependent

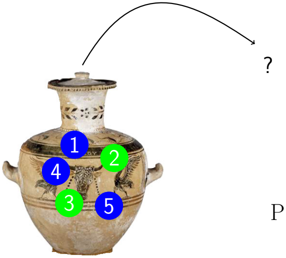
Joint probability of independence events

When A and B are independent, using the multiplication rule and, (i) or (ii) from previous slide:

$$\text{(iii)}\quad \mathrm{P}(AB) = \overbrace{\mathrm{P}(A|B)}^{=\mathrm{P}(A)}\mathrm{P}(B) = \mathrm{P}(A)\mathrm{P}(B)$$

We can use (i), (ii) or (iii) for any two independent events A and B

Example Drawing marbles WITH replacement

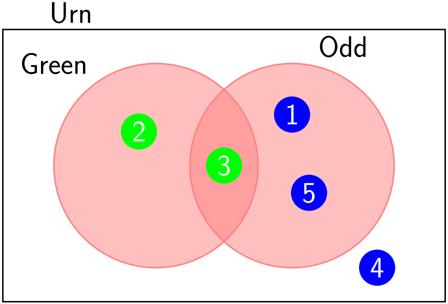


$$\begin{aligned}
A &= \{\bullet \text{ is drawn}\}\\
B &= \{\bullet \text{ is drawn}\}\\
\mathrm{P}(B) &= \frac{2}{5}\\
\mathrm{P}(B|A) &= \frac{2}{5}\\
\mathrm{P}(AB) = \mathrm{P}(B|A)\mathrm{P}(A) &= \mathrm{P}(B)\mathrm{P}(A)\\
&= \frac{2}{5} \times \frac{3}{5} = \frac{6}{25}
\end{aligned}$$

Union of events (1)

Union of events can sometimes be best visualized using a **Venn diagram** (John Venn, 1834-1923)

Example What is the probability of drawing a or an odd number ?



Union of events (2)

$$
\begin{aligned}
\mathrm{P}({\color{green}\bullet} \text{ or odd}) &= \mathrm{P}({\color{green}\bullet}) + \mathrm{P}(\text{Odd}) - \mathrm{P}({\color{green}\bullet} \text{ and odd}) \\
&= \frac{2}{5} + \frac{3}{5} - \frac{1}{5} \\
&= \frac{4}{5}
\end{aligned}
$$

In general, if A and B are:

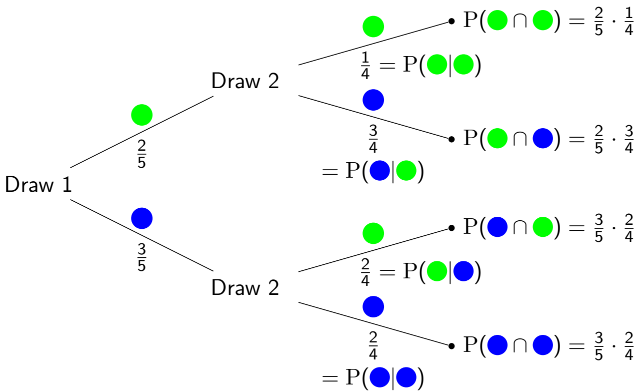
- disjoint, then $\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B)$
- not disjoint, then $\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(AB)$

Probability tree

Probability tree is useful for studying combinations of events. Branches of a tree are *conditional* probabilities.

Example

Drawing two marbles from urn without replacement:

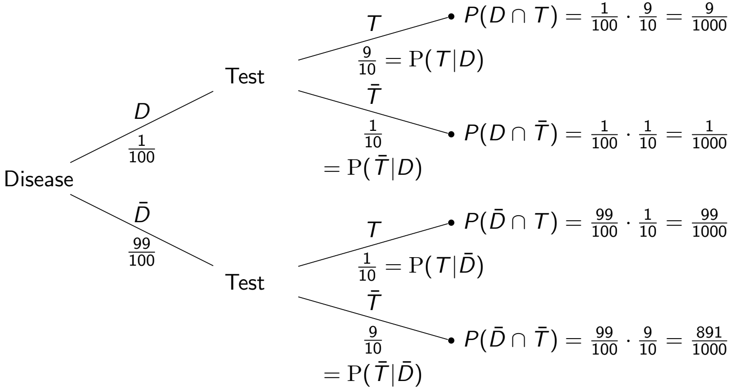


Bayes Theorem (Thomas Bayes, 1701-1761)

Trees are useful for visualizing $\mathrm{P}(B|A)$ when B follows from A in a natural (time) order. Many problems require $\mathrm{P}(A|B)$, **Bayes Theorem** provides an answer.

Example

Testing for an infectious disease.



What is $\mathrm{P}(D|T)$ or $\mathrm{P}(\bar{D}|\bar{T})$?

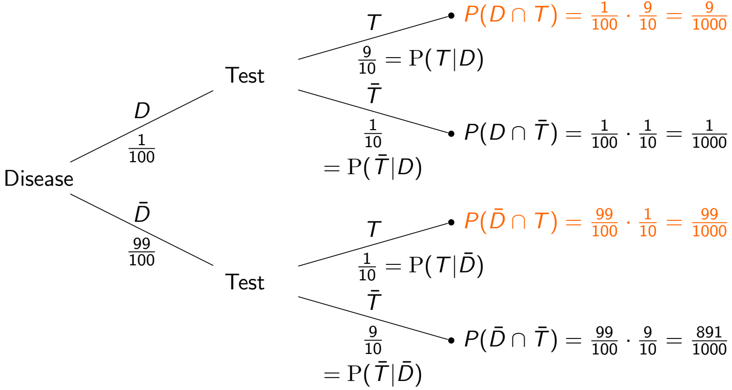
Bayes Theorem (2)

$$
\begin{aligned}
\mathrm{P}(D|T) = \frac{\mathrm{P}(D \cap T)}{\mathrm{P}(T)} &= \frac{\mathrm{P}(T|D)\mathrm{P}(D)}{\mathrm{P}(T)} \\
&= \frac{\mathrm{P}(T|D)\mathrm{P}(D)}{\underbrace{\mathrm{P}(T \cap D) + \mathrm{P}(T \cap \bar{D})}_{\text{Partition rule}}} \\
&= \frac{\mathrm{P}(T|D)\mathrm{P}(D)}{\underbrace{\mathrm{P}(T|D)\mathrm{P}(D) + \mathrm{P}(T|\bar{D})\mathrm{P}(\bar{D})}_{\text{Multiplication rule}}} \\
&= \frac{\frac{9}{10} \cdot \frac{1}{100}}{\frac{9}{10} \cdot \frac{1}{100} + \frac{1}{10} \cdot \frac{99}{100}} = \frac{9}{108}
\end{aligned}
$$

In general,

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

$\mathrm{P}(D|T)$ using tree



$$\mathrm{P}(D|T) = \frac{\mathrm{P}(DT)}{\mathrm{P}(T)} = \frac{\frac{9}{1000}}{\frac{9}{1000} + \frac{99}{1000}} = \frac{9}{108}$$