Conditional Probability

Peter Büchel

HSLU W

SA: W09

Qualitative predictors

- So far: All variables *quantitative* in linear regression system
- But: Often some predictor variables are *qualitative*

Example

- Data set `Credit` was collected in the USA
- Contains for a larger number of individuals:
  - `Balance` (monthly credit card invoice): Response variable, quantitative
  - `Age` (age): Predictor, quantitative
  - `Cards` (number of credit cards): Predictor, quantitative
  - `Education` (number of years of education): Predictor, quantitative
  - `Income` (Income in thousands of dollars): Predictor, quantitative
  - `Limit` (credit card limit): Predictor, quantitative
  - `Rating` (creditworthiness): Predictor, quantitative

- Data set:

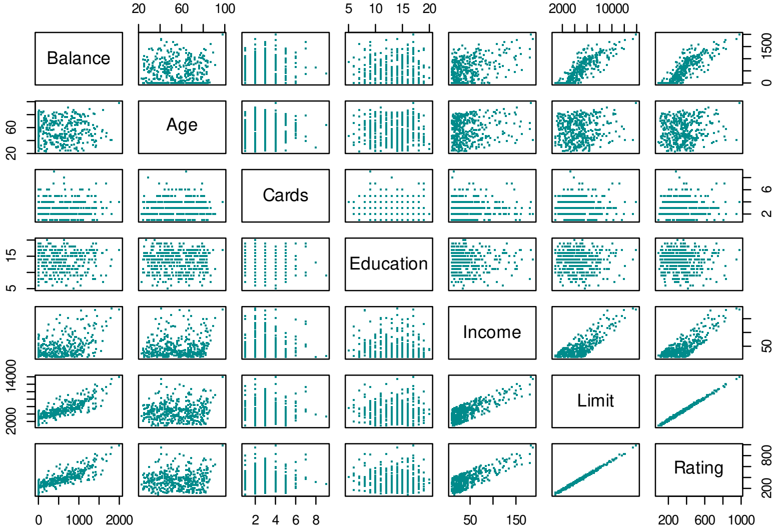
```
Credit <- read.csv("../Data/Credit.csv")[, -1]
head(Credit)
```

```
##    Income Limit Rating Cards Age Education Gender Student
## 1  14.891  3606    283     2  34        11   Male      No
## 2 106.025  6645    483     3  82        15 Female     Yes
## 3 104.593  7075    514     4  71        11   Male      No
## 4 148.924  9504    681     3  36        11 Female      No
## 5  55.882  4897    357     2  68        16   Male      No
## 6  80.180  8047    569     4  77        10   Male      No
##   Married Ethnicity Balance
## 1     Yes Caucasian     333
## 2     Yes     Asian     903
## 3      No     Asian     580
## 4      No     Asian     964
## 5     Yes Caucasian     331
## 6      No Caucasian    1151
```

```
colnames(Credit)
```

```
##  [1] "Income"    "Limit"     "Rating"    "Cards"
##  [5] "Age"       "Education" "Gender"    "Student"
##  [9] "Married"   "Ethnicity" "Balance"
```

- Figure:


Balance
Age
Cards
Education
Income
Limit
Rating
20
60
100
5
10
15
20
2000
10000
1500
0
60
20
6
2
15
5
50
14000
2000
800
200
0
1000
2000
2
4
6
8
50
150
200
600
1000

- Code:

```r
Credit <- read.csv("../Data/Credit.csv")
pairs(~Balance + Age + Cards + Education + Income + Limit + Rating,
    data = Credit, pch = ".", col = "dark cyan")
```

- Scatter plots of pairs of variables: Given by appropriate column and row labels

- Plot directly to the right of „balance": Scatterplot of the variables `Age` and `Balance`

- Scatter plots:
  - `Age` - `Balance`: No correlation
  - `Education` - `Balance`: No correlation
  - `Income` - `Balance`: Weak link
  - `Limit` - `Balance`: Strong correlation

- In addition to quantitative variables: Four qualitative predictors:
  - `Gender`
  - `Student`
  - `Ethnicity`
  - `Married`
- Qualitative predictors: Also called *factors*
- Factors assume *levels*:
  - `Gender`: Male, female
  - `student`: Yes, no
  - `Ethnicity`: Caucasian, African-American, Asian
  - `Married`: Yes, no

Qualitative predictor with only two levels

- Example `balance`: Difference between men and women
- Other variables are ignored for the moment
- Qualitative predictor with two *levels* (possible values)
- Addition of this variable in regression model very simple
- Introducing indicator variable (or *dummy variable*) which can only take two possible numerical values

Example

- For `Gender`:

$$x_i = \begin{cases} 1 & \text{if } i\text{-th person female} \\ 0 & \text{if } i\text{-th person male} \end{cases}$$

- Using this variable as an predictor variable in the regression model
- Model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{-th person female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{-th person male} \end{cases}$$

- β_0: Average credit card bills of men
- $\beta_0 + \beta_1$: Average credit card bills of women
- β_1: Average *difference* of bills men/women

- Table: Coefficient estimates for our model:

	coefficient	Std.error	t statistics	p value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[female]`	19.73	46.05	0.429	0.6690

```r
balance <- Credit[, "Balance"]
gender <- Credit[, "Gender"] == "Female"
round(summary(lm(balance ~ gender))$coef, digits = 5)
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 509.80311   33.12808 15.38885  0.00000
## genderTRUE   19.73312   46.05121  0.42850  0.66852
```

- Estimated average bills for men: \$ 509.80
- Estimated difference to women: \$ 19.73
- women: $\$ 509.80 + \$ 19.73 = \$ 529.53$
- *p*-value for indicator variable β_1 with 0.6690 very high
- No statistically significant difference in `Balance` between women and men

- Example before: Women coded as 1 and men as 0
- Completely arbitrary
- Coding: *No* Influence on degree of estimation of the model to data
- Different coding: Different interpretation of the coefficients
- Coding men with 1 and women with 0
- Estimate for the parameters β_0 and β_1 \$ 529.53, resp. \$ -19.73
- Corresponds in turn to invoices from:
  - Women: \$ 529.53
  - Men: \$ 529.73 – \$ 19.73 = \$ 509.80
- Same result as before

Example

- Instead of the $0/1$ coding:

$$x_i = \begin{cases} 1 & \text{if } i\text{-th person female} \\ -1 & \text{if } i\text{-th person male} \end{cases}$$

- Regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{-th person female} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{if } i\text{-th person male} \end{cases}$$

- β_0: Average bills without consideration of gender

- β_1: Value of women above average and below average for men

- β_0 estimated by \$ 519.665: Average bills of \$ 509.80 for men and of \$ 529.53 for women
- Estimate \$ 9.865 for β_1: half of the difference \$ 19.73 between men and women
- Important: Predictions for response variable do not depend on coding
- Only difference: Interpretation of the coefficients

Qualitative predictor with more than two levels

- Qualitative predictor can have more than two levels
- *One* indicator variable for all possible values is not enough
- In this situation: Add additional indicator variable

Example

- Variable `Ethnicity`: *Three* possible levels

- Choose *two* different indicator variables

- *Choice* of the 1st indicator variables:

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{-th person asian} \\ 0 & \text{if } i\text{-th person not asian} \end{cases}$$

- 2nd indicator variable:

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{-th person caucasian} \\ 0 & \text{if } i\text{-th person not caucasian} \end{cases}$$

- Include both variables in regression equation:

$$y_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\varepsilon_i = \begin{cases} \beta_0+\beta_1+\varepsilon_i & \text{if } i\text{-th person asian} \\ \beta_0+\beta_2+\varepsilon_i & \text{if} i\text{-th person caucasian} \\ \beta_0+\varepsilon_i & \text{if } i\text{-th person african-american} \end{cases}$$

- β_0: Average credit card bills of African Americans
- β_1: Difference in average bills of African Americans and Asians
- β_2: Difference in average bills of African Americans and Caucasians

Remarks

- There is always one indicator variable less than it has levels
- Level without indicator variable (here African American): *Baseline*
- The following equation makes *no* sense:

$$y_i = \beta_0 + \beta_1 + \beta_2 + \varepsilon_i$$

  - Should be asian *and* caucasian

- Output: Estimated `balance` \$ 531.00 as baseline (african american):

```r
balance <- Credit[, "Balance"]
ethnicity <- Credit[, "Ethnicity"]
summary(lm(balance ~ ethnicity))
```

```
##
## Call:
## lm(formula = balance ~ ethnicity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -531.00 -457.08  -63.25  339.25 1480.50
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           531.00      46.32  11.464   <2e-16 ***
## ethnicityAsian        -18.69      65.02  -0.287    0.774
## ethnicityCaucasian    -12.50      56.68  -0.221    0.826
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.9 on 397 degrees of freedom
## Multiple R-squared:  0.0002188, Adjusted R-squared:  -0.004818
## F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

- Estimate for asians: \$ -18.69
- Average bills smaller by this amount than those of african americans
- Caucasians have averaged around \$ 12.50 smaller bills than the african americans
- p-values large: Random deviations
- No significant difference in credit card bills between ethnic groups
- Level for baseline arbitrary
- Prediction of the target variable does not depend on the coding

- p-values depend on the encoding
- View F statistics
- F-test and test

$$H_0: \quad \beta_1 = \beta_2 = 0$$

- p-value of this statistic depends *not* on the coding
- p value 0.96: Relatively high
- Assumption confirmed: Null hypothesis *not* reject
- There is no connection between `Balance` and `Ethnicity`

Qualitative and quantitative predictors

- Indicator variables: Integrate qualitative *and* quantitative predictor into regression model
- Regression of `Balance` with quantitative predictor `Income` and qualitative predictor `student`
- `Student` with indicator variables
- Multiple linear regression

Example: Data set `Credit`

- Predict response variable `Balance` by predictor variables `Income` (quantitative) and `Student` (qualitative)

- Without interaction term:

$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \cdot \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{-th person student} \\ 0 & \text{if } i\text{-th person no student} \end{cases} \\
&= \beta_1 \cdot \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{-th person student} \\ \beta_0 & \text{if } i\text{-th person not a student} \end{cases}
\end{aligned}$$

- Output:

```
student <- Credit[, "Student"]
income <- Credit[, "Income"]
summary(lm(balance ~ income + student))
```

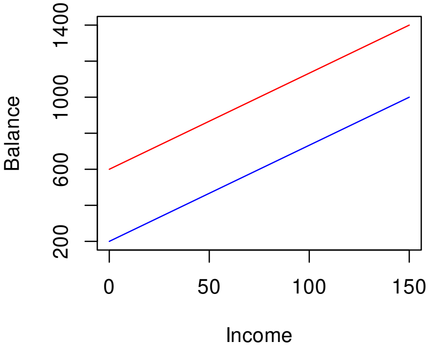
```
##
## Call:
## lm(formula = balance ~ income + student)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -762.37 -331.38  -45.04  323.60  818.28
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 211.1430    32.4572   6.505 2.34e-10 ***
## income        5.9843     0.5566  10.751  < 2e-16 ***
## studentYes  382.6705    65.3108   5.859 9.78e-09 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 391.8 on 397 degrees of freedom
## Multiple R-squared:  0.2775, Adjusted R-squared:  0.2738
## F-statistic: 76.22 on 2 and 397 DF,  p-value: < 2.2e-16
```

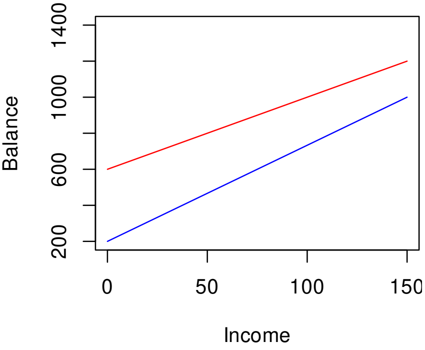
- $\widehat{\beta}_0$:
  Without income and as a non-student you pay \$211 monthly credit card bill

- $\widehat{\beta}_1$:
  For every \$1000 more income, you pay \$6 more for credit card bill (regardless of student status)

- $\widehat{\beta}_2$:
  Students pay \$383 more for credit card bills than non-students (regardless of income)

- Model describes two parallel straight lines: One for students and one for non-students
  - Slope β_1 is the same for both
  - y-axis sections are different ($\beta_0 + \beta_2$ and β_0)
- Figure left:





- Average increase of `Balance` for increase of `Income` by one unit does not depend on whether the respective individual is a student or not
- Possible limitation of the model: Change in `Income` may have a different effect on bills whether someone is student or not
- Easing this restriction: Introduction of an interaction variable
- `Income` is combined with the indicator variable for `Student`

- Model:

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \cdot \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \cdot \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$
$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \cdot \texttt{income}_i & \text{if not student} \end{cases}$$

- Two different regression lines for students and non-students (top right figure):
  - Different slopes $\beta_1 + \beta_3$ and β_1
  - Different *y*-axle sections $\beta_0 + \beta_2$ and β_0

- Possibility to consider changes in response variable (credit card bills) due to changes in income for students and non-students separately

- Right side of figure above: Estimated relationship between `Income` and `Balance` for students (blue) and non-students (red)
- Slope for non-students greater than for students
- Suggests: Increase in student income results in a greater increase in credit card bills than for non-students

- Output:

```
summary(lm(balance ~ income * student))
##
## Call:
## lm(formula = balance ~ income * student)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -773.39 -325.70  -41.13  321.65  814.04
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         200.6232    33.6984   5.953 5.79e-09 ***
## income                6.2182     0.5921  10.502  < 2e-16 ***
## studentYes          476.6758   104.3512   4.568 6.59e-06 ***
## income:studentYes   -1.9992     1.7313  -1.155    0.249
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 391.6 on 396 degrees of freedom
## Multiple R-squared:  0.2799, Adjusted R-squared:  0.2744
## F-statistic:  51.3 on 3 and 396 DF,  p-value: < 2.2e-16
```

- p-value of the interaction is not statistically significant
- Thus there is no interaction
- Slopes of the two straight lines are not significantly different

Example for Conditional Probability

- Group of 20 people:
  - Some are smokers, the others non-smokers
  - Some are women, the rest are men

- Denote:

$$F\text{: Female,} \qquad M\text{: Male,} \qquad S\text{: Smoker,} \qquad \overline{S}\text{: Non-smoker}$$

- Table:

$|S \cap M|$ $\qquad$ $|S|$

	M	F	
S	3	1	4
$\overline{S}$	9	7	16
	12	8	20

- There are:
  - 4 smokers and 16 non-smokers
  - 8 women and 12 men
- Value 3 top left: Number of people who are male *and* are smokers
- Notation:

$$\big|S \cap M\big| = 3$$

- To get probabilities: Divide all values in table by 20
- Table with probabilities:

$P(S \cap M)$ $P(S)$

	M	F	
S	0.15	0.05	0.20
$\overline{S}$	0.45	0.35	0.8
	0.6	0.4	1

- Value 0.15 top left: Probability that a randomly chosen person is a man and a smoker

- Calculation:

$$P(S \cap M) = \frac{|S \cap M|}{|\Omega|} = \frac{3}{20} = 0.15$$

- Value 0.2 last column: Probability that a randomly chosen person is a smoker

- Hence:

$$P(S) = \frac{|S|}{|\Omega|} = 0.2$$

- Consider only a part of the table: Smokers

	M	F	
S	0.15	0.05	0.2
$\overline{S}$	0.45	0.35	0.8
	0.6	0.4	1

- May ask for probability that a randomly chosen person *among the* smokers is a man

- From 1st table (absolute numbers), this probability is:

$$\frac{|S \cap M|}{|S|} = \frac{3}{4} = 0.75$$

- Using 2nd table (probabilities):

$$\frac{P(S \cap M)}{P(S)} = \frac{0.15}{0.20} = 0.75$$

- This means that 75 % of smokers are men

- This fact is called *conditional probability*

- Notation:

$$P(M \mid S)$$

- Variable S *after* vertical dash: New sample space

- Term "conditional": Not whole sample space is considered but only part of it

- New sample space: Smokers S

- It follows:

$$P(M \mid S) = \frac{P(S \cap M)}{P(S)} \qquad (*)$$

- Formula is used to define conditional probability

- Calculation of conditional probability:

$$P(S \mid M)$$

- Probability that a randomly chosen man is a smoker

- Table: Only men are considered:

	M	F	
S	0.15	0.05	0.2
$\overline{S}$	0.45	0.35	0.8
	0.6	0.4	1

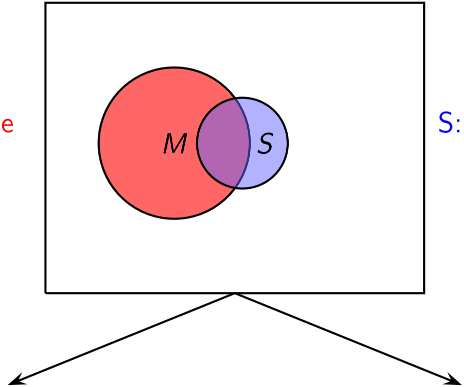
- Calculation of probability: Swap variable M and S in equation $(*)$
- Result:

$$P(S \mid M) = \frac{P(M \cap S)}{P(M)} = \frac{P(S \cap M)}{P(M)} = \frac{0.15}{0.6} = 0.25$$

Conditional Probability

Ω: Students in this lecture

M: Male
$P(M)$



S: Smoker
$P(S)$

$P(S \mid M)$

when a man has been chosen

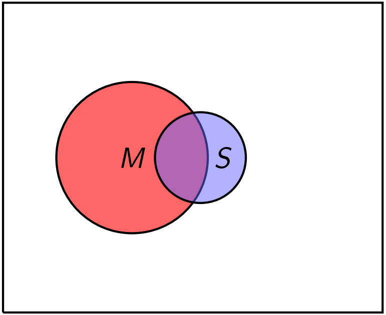
$P(M \mid S)$

when a smoker was chosen

Conditional Probability

Ω: Students in this lecture

M: Male
$P(M)$

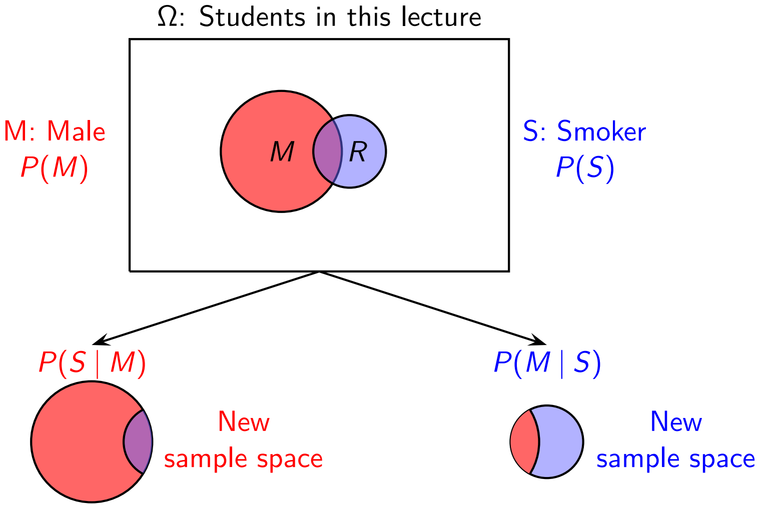


S: Smoker
$P(S)$

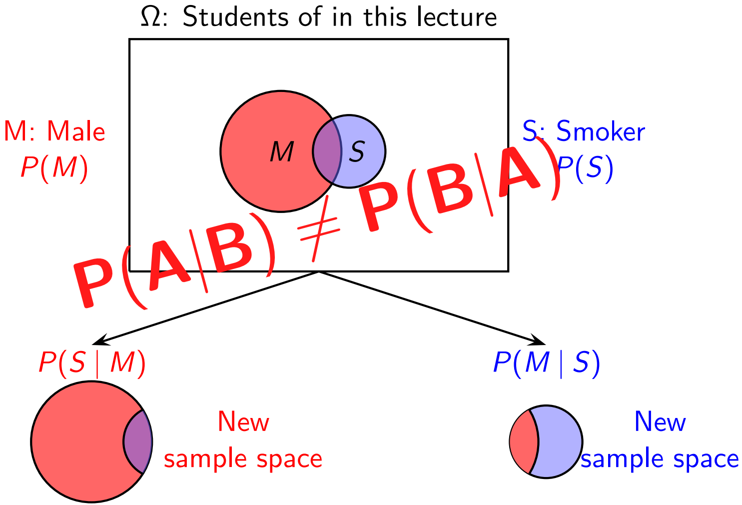
Which statement is correct?

1. $P(M \mid S) = P(S \mid M)$
2. $P(M \mid S) > P(S \mid M)$
3. $P(M \mid S) < P(S \mid M)$

Conditional Probability



Conditional Probability



Example

- There is:
$$P(\text{woman} \mid \text{pregnant}) = 1$$
- All pregnant people are women
- However:
$$P(\text{pregnant} \mid \text{women}) = 1$$
- All women are pregnant**?**
- In reality:
$$P(\text{pregnant} \mid \text{woman}) = 0.03$$
- 3 percent of women of childbearing age are pregnant

Conditional Probability: Definition

- *Conditional probability* is probability that event A occurs when one already knows that B has occurred
- Notation:

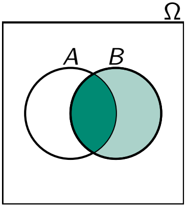
$$P(A \mid B)$$

- Vertical bar is read as “under the condition” or “given”
- Conditional probability $P(A \mid B)$ is defined by

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

Definition of Conditional Probability using Venn Diagrams

- Figure:



- $P(\Omega) = 1$
- $P(A \cap B)$ area of dark colored area
- $P(B)$ area of total colored area B
- Proportion of dark colored area to colored area is:

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

Example: Medical Test

- Medical Test: Designed to determine whether a person has a disease or not
- Naturally this test is not quite accurate:
  - Sometimes indicates disease even though person is healthy
  - It does not indicate disease, even though person is ill
- Question:
  - You go to physician and do this test for a deadly disease
  - Test is positive: You have disease according to test but do not necessarily have disease
  - What is the probability that you really have disease?

- Notation:
  - D: Person has disease $\overline{D}$: Person does not have disease
  - $+$: Test indicates disease $-$: Test does not indicate disease

- Probabilities in Table are known by experiments:

	D	$\overline{D}$
$+$	0.009	0.099
$-$	0.001	0.891

- For example: Probability that person has disease *and* test is positive

$$P(D \cap +) = 0.009$$

- This probability is quite small
- Reason: Only a small proportion of population has disease
- Various conditional probabilities:
  - $P(+ \mid D)$: Probability that a ill person is really tested positive
  - $P(- \mid \overline{D})$: Probability that a healthy person is correctly tested negative
  - $P(D \mid +)$: Probability that a person tested positive is really ill
  - etc.
- First calculate probability $P(+ \mid D)$:

$$P(+ \mid D) = \frac{P(+ \cap D)}{P(D)} = \frac{0.009}{0.009 + 0.001} = 0.9$$

- For $P(D)$ following fact was used:

$$P(D) = P(D \cap +) + P(D \cap -) = 0.009 + 0.001$$

- Sum of entries in Table in column D
- Patients are tested either positive or negative
- Conditional probability $P(- \mid \overline{D})$:

$$P(- \mid \overline{D}) = \frac{P(- \cap \overline{D})}{P(\overline{D})} = \frac{0.891}{0.891 + 0.099} = 0.9$$

- Seems that this test is quite accurate

- Ill people are tested positive with 90 % and healthy people are tested negative with 90 %
- *Reverse* question
- Suppose you do test and result is positive
- What is the probability that you really have disease?
- Most people will answer 0.9 (even physicians)
- Do you have to worry a lot and write a will?

- *Correct* answer is conditional probability $P(D \mid +)$:

$$P(D \mid +) = \frac{P(+ \cap D)}{P(+)} = \frac{0.009}{0.009 + 0.099} = 0.08$$

- What does this result mean?
- Conditional probability $P(D \mid +)$ is probability that you are really ill if test is positive
- This probability is only 8 %
- If test is positive, you only have disease with probability 8 %
- Positive test tells very little about whether you have disease or not

- *Why* this surprising result?
- Reason: Disease is *rare*
- Numerical example: 100 000 people
  - 1000 people have disease (1 %)
  - 90 % of these will test positive: 900 people
  - 99 000 have not contracted disease
  - 10 % of these will test positive: 9900 people
  - Total number of people whose test was positive:

$$900 + 9900 = 10\,800$$

  - However: Among those who tested positive there are far more healthy people who were *falsely* tested positive
  - Probability that a person who has tested positive is really ill:

$$\frac{900}{10\,800} = 0.0833$$

Bayes Theorem

- Bayes theorem: Describes relation between $P(A \mid B)$ and $P(B \mid A)$

Bayes Theorem

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

- *Example:* Bayes theorem returns same solution as before:

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+)} = \frac{0.9 \cdot (0.009 + 0.001)}{0.009 + 0.099} = \frac{0.009}{0.009 + 0.099} = 0.08$$

Proof of Bayes Theorem

- Apply definition of conditional probability *twice*
  - It applies:

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} \quad \Rightarrow \quad P(B \cap A) = P(B \mid A) \cdot P(A)$$

  - And:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \Rightarrow \quad P(A \cap B) = P(A \mid B) \cdot P(B)$$

- Because $A \cap B = B \cap A$, it follows:

$$P(A \cap B) = P(B \cap A)$$

- And thus:

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

- Divide both sides by $P(B)$ and get Bayes theorem:

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

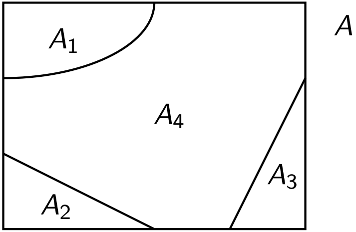
Law of Total Probability

- Another useful law: *Total probability*

- Set A divided into sets $A_1, \ldots, A_k$, which do not intersect and together (union) form whole set A

- Such a division is called *partition*

- Graphical example:



Example

- Rolling of die: Possible partition of $A = \{1, 2, 3, 4, 5, 6\}$:

$$A_1 = \{1\}, \qquad A_2 = \{2, 4\}, \qquad A_3 = \{3, 5, 6\}$$

- Because:

$$A_1 \cap A_2 = \{\}, \quad A_1 \cap A_3 = \{\}, \quad A_2 \cap A_3 = \{\}$$

- And:

$$A_1 \cup A_2 \cup A_3 = A$$

Law of Total Probability

- If $A_1, \ldots, A_k$ is a partition of A and B an event, then

  $$\begin{aligned} P(B) &= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \ldots + P(B \mid A_k)P(A_k) \\ &= \sum_{i=1}^{k} P(B \mid A_k)P(A_k) \end{aligned}$$

- For $k = 2$:

  $$P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2)$$

- For $k = 3$:

  $$P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)$$

- Proof: See lecture notes

Example: Spam-Mail

- Divide emails into three categories:

$$A_1\text{: spam,}\quad A_2\text{: low priority,}\quad A_3\text{: high priority}$$

- Known from earlier observations:

$$P(A_1) = 0.7, \qquad P(A_2) = 0.2, \qquad P(A_3) = 0.1$$

- It applies

$$P(A_1) + P(A_2) + P(A_3) = 1$$

  as it should for a partition

- Event B: Word "free" appears in email

- This word occurs very often in spam mails, but not only
- Known from earlier observations:

$$P(B \mid A_1) = 0.9, \quad P(B \mid A_2) = 0.01, \quad P(B \mid A_3) = 0.01$$

- In this case, the sum is not 1, doesn't have to be
- These are the probabilities in which word "free" occurs in three mail categories
- Suppose you receive an email containing word "free"
- What is the probability that it is spam?

- Solution with Bayes theorem and law of total probability:

$$
\begin{aligned}
P(A_1|B) &= \frac{P(B \mid A_1)P(A_1)}{P(B)} \\
&= \frac{P(B \mid A_1)P(A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)} \\
&= \frac{0.9 \cdot 0.7}{(0.9 \cdot 0.7) + (0.01 \cdot 0.2) + (0.01 \cdot 0.1)} \\
&= 0.995
\end{aligned}
$$

- Many spam filters are actually based on this principle
- Mails are searched for words like “free”, “credit”, etc., which are frequently found in spam mails, but are not likely to be found in others

Example

- Some children are born with Down syndrome
- There are tests that pregnant women can take to see if their baby could suffer from this disease
- Study (University of Liverpool) on how well test results are interpreted by those involved: Pregnant women, their partners, midwives and obstetricians
- Following scenario was shown to 85 people:

  Serum test examines pregnant women for babies with Down syndrome. Test is a very good but not perfect. About 1 % of babies have Down syndrome. If baby has Down syndrome, there is a 90 % probability that result will be positive. If baby is not affected, there is still a 1 % probability that result will be positive. A pregnant woman has been tested and result is positive. What is the probability that her baby actually has Down syndrome?

- Table, how well the 85 people performed:

	Correct	Too high	Too low	
Pregnant women	1	15	6	22
Partners	3	10	7	20
Midwives	0	10	12	22
Obstetricians	1	16	4	21
	5	51	29	85

- Only five of the 85 gave the correct answer
- Health professionals were no better than pregnant women and their partners
- Especially remarkable: Only one in 21 obstetricians gave correct answer

- Other group of 81 people: Alternative scenario was shown:

  Serum test examines pregnant women for babies with Down syndrome. Test is a very good but not perfect. About 100 of 10 000 *babies have Down syndrome. Of these 100 babies with Down syndrome, 90 will have a positive test result. Of remaining 9900 unaffected babies, 99 will still have a positive test result. How many pregnant women who test positive will actually have a baby with Down's syndrome?*

- Result:

	Correct	Too high	Too low	
Pregnant women	3	3	10	21
Companion	3	8	9	20
Midwives	0	7	13	20
Obstetricians	13	3	4	20
	19	26	36	81

- Clearly an improvement
- Reformulation of scenario: Absolute numbers used instead of percentages
- Makes it an easier problem
- Must only consider the two numbers 90 and 99:

$$\frac{90}{90+99} \approx 48\,\%$$

- But: Still, only about a quarter of participants gave correct answer
- After all: Obstetricians scored significantly better