Probabilistic Modeling of Perception and Cognition

Lecture Notes and Exercise Solutions

COLLECTED AND COMPILED BY

Sebastian Höffner shoeffner@uos.de
Lisa Goerke lgoerke@uos.de
Andrea Suckro asuckro@uos.de
Valentin Churavy churavy@uos.de
Kai Standvoss kstandvoss@uos.de

Summer semester 2014

Lecturer:
Juniorprof. Dr. Frank Jäkel

University Osnabrück
Institute of Cognitive Science

Dear reader,

These notes were originally taken during the summer semester 2014. We hope they are helpful while following along the course in the future, although they are partly incomplete, riddled, or really just a quick scribble of what was on the board.
Sadly we didn't write down all exercise solutions, however you should try to solve them yourself anyway.
If you find any mistakes, contact us or correct them if possible.
Feel free to take our notes and adapt them to your semester as well – or just expand this document.

We hope you enjoy reading the notes as much as we enjoyed writing and discussing them.

Sebastian, Lisa, Andrea, Kai, and Valentin

Contents

1 Probability Refresher I *2014-04-22*

What is Logic? What is Probability Theory?

Logic is reasoning under certainty, **Probability Theory** is reasoning under uncertainty. In Logic we can distinguish between descriptive and prescriptive approaches - in Probability Theory we distinguish between the frequentist and the Bayesian view.
The two views in Probability Theory are different in how probabilities are to be interpreted: The frequentist view interprets probabilities as **limits of relative frequencies**, while the Bayesian view interprets probabilities as **beliefs**. This course will try to make the distinction between both views clear.

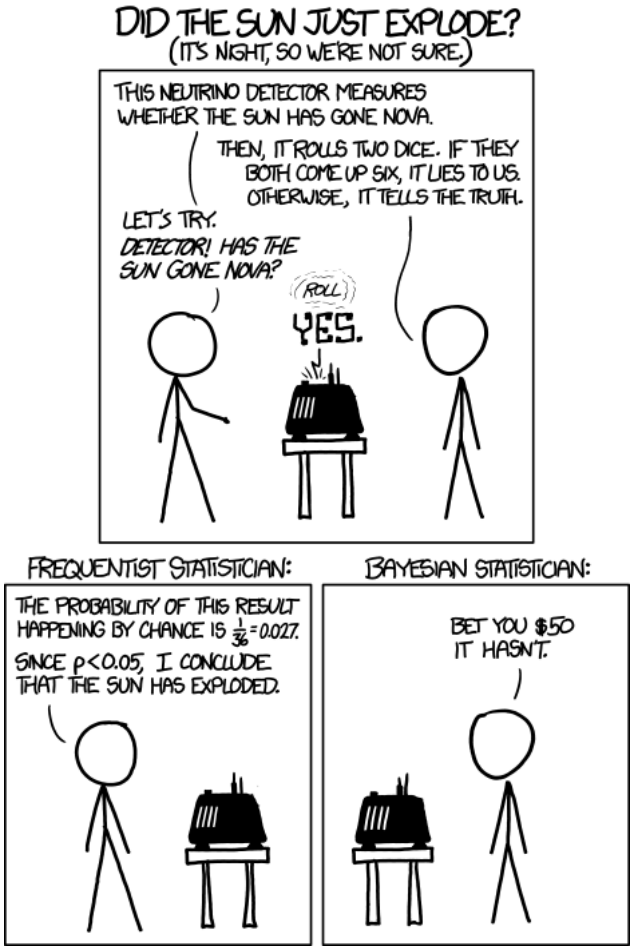


Figure 1: *Source: http://xkcd.com/1132/*

1.1 Games of Chance: Coin Toss & Thumbtack Toss

Coin Toss

Alice offers Bob a bet:

> *Let's toss a coin. I will give you $2 whenever it shows heads. But each time it shows tails, you will give me $3.*

Should Bob accept? Let's take Bob's point of view and see.

$$x \in \{0 = tails, 1 = heads\}$$
$$\text{If } x = 1 : \text{Alice gives Bob \$2.}$$
$$\text{If } x = 0 : \text{Bob gives Alice \$3.}$$

If they play once, the money Bob gains equals: $2x - 3(1 - x)$, with $x = 0$ on tails or $x = 1$ for heads, respectively. But since they will play n times Bob has to calculate the sum for n games:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(2x_i - 3\left(1 - x_i\right)\right) &= 2\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) - 3\left(\frac{1}{n}\sum_{i=1}^{n}\left(1 - x_i\right)\right) \\ &= 2\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) - 3\left(1 - \frac{1}{n}\sum_{i=1}^{n}x_i\right)\end{aligned}$$

Where $\frac{1}{n}\sum_{i=1}^{n}x_i$ is the **relative proportion of heads**.
The expected value (*read as: "Bob's expected gain"*) is, as can be seen above, $E = 2p - 3(1 - p)$, where p is the probability of heads. p can now easily be expressed as:

$$\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) = p$$

But what shall Bob do now, where he has a formula to derive p? Basically he has two choices: Trying out and tossing a coin n times or using his *a priori belief* and assigning a p. How he decides is the difference between the frequentist and the Bayesian view. Eventually Bob sets $p = 0.5$ and inserts it into the formula for the expected outcome. So Bob's expected gain is $E = 2 \cdot 0.5 - 3(1 - 0.5) = -0.5[\$]$. Hence Bob shouldn't play.

Thumbtack Toss

Alice has another bet for Bob:

> *Let's toss a thumbtack. I have heads, you have tails. You are allowed to choose your stakes, but I have to agree on them to play.*

What stakes should Bob choose? We will have a look at his situation again.

$$x \in \{0 = tails, 1 = heads\}$$

Figure 2: Thumbtacks - left: tails, right: heads. *Source: http://blog.sls-construction.com/*

The probability p is again:

$$\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) = p$$

What is different is Bob's expected gain. He now has to consider the stakes as well.

$$E = s_1 p - s_2(1-p)$$

s_1 is Alice's stake and s_2 is Bob's stake. To have a *"fair"* bet the expected gain should be zero. Bob uses this knowledge to derive his stake.

$$\begin{aligned} & E = s_1 p - s_2(1-p) \\ 0 = E \; & \Leftrightarrow s_1 p = s_2(1-p) \\ & \Leftrightarrow \frac{p}{1-p} = \frac{s_2}{s_1} \end{aligned}$$

This last formula $\frac{p}{1-p} = \frac{s_2}{s_1}$ are the **odds**. If Bob fixes one stake and inserts p, he can calculate the other stake needed for a fair bet. But again he has the problem of how to get to p.

Conclusion

To derive p you always have two possibilities: The frequentist view and the Bayesian view. The difference is how we measure p:

- **Frequentist:** measure p as a property of the coin/thumbtack by throwing it n times
- **Bayesian:** measure p as a property of the "agent" (i.e. the decision-maker) by asking which bets are "fair" for him/her

PMPC Tutorial Sheet 1

1. How are you going to get an A in this class? What is your strategy to achieve this goal? Write it down!

2. (a) I pay you \$10 if you roll a 6 with a six-sided die. What should your stake be for the bet to be fair? (b) You suggest to pay me \$25 if I roll a 6 with a six-sided die—but only if you like the stakes that I offer you in case I don't roll a six. For which stakes should you accept the bet? (c) I offer you a bet with my stake being s_1 and your stake being s_2. For what probability would the bet be fair?

3. Neural correlates of decision variables in parietal cortex [5, 1]. Single cell responses are recorded from LIP (lateral intra-parietal area) while a monkey is performing the following saccadic decision task. The monkey is fixating. With a change of colour of the fixation spot the monkey is instructed to perform a saccade to one of two possible targets. Say red indicates that the monkey should saccade to the red target—and blue to the blue target. Each of the two instructions can come up with different probabilities. Furthermore, the correct execution of the instruction (which is not difficult for the monkey after training) results in different juice-rewards—depending on whether the target was red or blue. It is suspected that the reward that the monkey expects to receive after having seen the target correlates with the firing rate of neurons in LIP.
   You want to set up an experiment in which the average amount of juice a monkey receives is 1 ml per trial. You want to fix the probability with which each instruction (red or blue) comes up. Say, the probability for red should be p. How much juice should be given on the red trials? And the blue trials?

4. Linda Problem [6]. Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
   Sort the following statements with regard to their probability, starting with the most probable statement:
   (a) Linda is a teacher in elementary school
   (b) Linda works at a bookstore and takes Yoga classes
   (c) Linda is active in the feminist movement
   (d) Linda is a psychiatric social worker
   (e) Linda is a bank teller
   (f) Linda is an insurance salesperson
   (g) Linda is a bank teller and is active in the feminist movement

5. In an experiment on visual working memory subjects are shown a display with N squares of different and easily distinguishable colors at random locations. The screen goes blank for some time during which the subjects are asked to remember the display. After a short while the screen is again showing a display with squares that is almost identical to the first display. The only

difference is that a randomly picked square has changed color. The task of the subjects is to report whether they have seen a change or not and to report which square has changed (if they didn't see a change they are forced to guess). For each trial the experimenter records whether the subject reported seeing a change or not and whether the square the subject chose was the correct one (the experimenter is lazy and does not keep records of the square that changed and the square that was chosen by the subject). Assume that the subject can remember the colors of exactly M squares. What do you expect the joint distribution of the responses to be? How does the probability of a correct response change as a function of the number of squares? (Exercise was inspired by [4])

6. Recommended reading to accompany the course: John Kruschke has written a very accessible book on Bayesian statistics for psychology and cognitive science students [2]. Michael Lee and Eric-Jan Wagenmakers' book is also very readable and features many direct applications in cognitive modeling [3]. Although I don't follow these two books each of them is a good complement to this course. Another good—and still developing—source for additional reading for the first half of this class is *Probabilistic Progamming and Bayesian Methods for Hackers.*[1] The second half of the course is somewhat similar to part IV in [3] and will focus more concretely on cognitive models. Another good online source for the second half of the course is *Probabilistic Models of Cognition.*[2]

References

[1] P. W. Glimcher. *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*. Bradford Book, 2004.

[2] J. K. Kruschke. *Doing Bayesian Data Analysis*. Academic Press, Burlington, MA, 2011.

[3] M. D. Lee and E.-J. Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, 2013.

[4] S. J. Luck and E. K. Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390:279–281, 1997.

[5] M. L. Platt and P. W. Glimcher. Neural correlates of decision variables in parietal cortex. *Nature*, 400:233–238, 1999.

[6] A. Tversky and D. Kahneman. Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315, 1983.

[1] https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers

[2] https://www.probmods.org/

3 Solution 1: Probability Refresher I *2014-04-28*

Exercise 2

Exercise 2.a

We use the formula for calculating odds:

$$\frac{p}{1-p} = \frac{s_2}{s_1}$$

The probability for rolling a 6 is $\frac{1}{6}$, s_1 can be set to 10.

$$\frac{\frac{1}{6}}{1-\frac{1}{6}} = \frac{s_2}{10} \Leftrightarrow s_2 = 2$$

So our stake should be \$2.

Exercise 2.b

We use the same formula again, this time with $s_1 = 25$.

$$\frac{\frac{1}{6}}{1-\frac{1}{6}} = \frac{s_2}{25} \Leftrightarrow s_2 = 5$$

So we accept all stakes $s_2 \geq \$5$.

Exercise 2.c

We use the formula for odds and solve it for p.

$$\begin{aligned}
& \frac{p}{1-p} = \frac{s_2}{s_1} \\
\Leftrightarrow \quad & p = \frac{s_2}{s_1}(1-p) \\
\Leftrightarrow \quad & p = \frac{s_2}{s_1} - \frac{s_2}{s_1}p \\
\Leftrightarrow \quad & ps_1 = s_2 - ps_2 \\
\Leftrightarrow \quad & ps_1 + ps_2 = s_2 \\
\Leftrightarrow \quad & p(s_1 + s_2) = s_2 \\
\Leftrightarrow \quad & p = \frac{s_2}{s_1 + s_2}
\end{aligned}$$

So the probability should be $p = \frac{s_2}{s_1+s_2}$.

Exercise 3

The average amount of juice over all trials shall be 1 ml. The formula for this is the known:

$$EV = s_1 p_1 + s_2 p_2$$

Where p_1 is the probability for a red trial and p_2 for a blue trial. Since we always have either a red or a blue trial, we get the following:

$$\begin{aligned}
EV = 1\ ml &= s_1 p_1 + s_2(1-p_1) \\
\Leftrightarrow s_1 &= \frac{1\ ml - s_2(1-p_1)}{p_1} \\
\Leftrightarrow s_2 &= \frac{1\ ml - s_1 p_1}{1-p_1}
\end{aligned}$$

In order to obtain concrete values we need to fix p_1 and one of the stakes (either s_1 or s_2) with a value $\leq 1\ ml$.

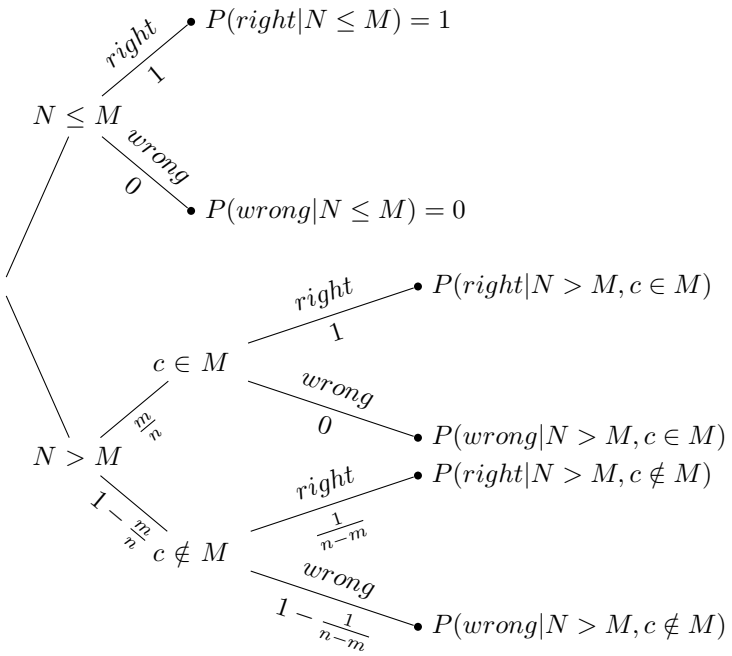


Figure 3: Probability tree

Exercise 4

In general the statements with a conjunction are less likely. So a possible order could be:
$(c) \Rightarrow (d) \Rightarrow (e) \Rightarrow (f) \Rightarrow (a) \Rightarrow (b) \Rightarrow (g)$
However, many people would assign higher probabilities to those statements with conjunctions (for example "Linda is a bank teller and is active in the feminist movement." is often seen as more probable than "Linda is a bank teller.") because of their contextual knowledge. This phenomenon is called "conjunction fallacy".

Exercise 5

Note: We call the *set of memorized squares* M, the *set of all squares* N and their respective *numbers of elements* $|M| = m$ and $|N| = n$. We also introduce c, the changing square, for which by definition holds $c \in N$.
In this experiment we have two cases. The first case is that the number m of squares the subject can memorize exceeds or is equal to the number n of total squares. This case is trivial: The subject will (under ideal circumstances) always report the right square. Hence the probability $P(right|N \leq M) = 1$.
The second case is the more interesting case. How is the probability for being right if $N > M$ ($P(right|N > M)$)?
We have to distinguish between two cases again: The first case is that the changing square is among those m squares the subject memorized, $c \in M$. The second case is when the change lies outside, $c \notin M$.
In the first case the subject will again always report the right c, but in the second case it has to guess *one out of those squares it did not memorize.*
Since we can not tell how big M and N are, we can not set probabilities for our first decision. But we can tell how big the probabilities are that $c \in M$ or $c \notin M$. For $c \in M$ this is simply $p = \frac{m}{n}$, we can understand this formula as *"there are m possibilities that c is among the m squares out of N"*. For $c \notin M$ we can just use the complementary event: $p = 1 - \frac{m}{n}$.
We don't have to follow $c \in M$ further, as we already found out the subject will find the change. In case $c \notin M$ the probabilities are a bit different: The subject now has to pick one of the elements out of $N \setminus M$, since it knows that $c \notin M$. Either it hits the correct one ($p = \frac{1}{n-m}$) or not ($p = 1 - \frac{1}{n-m}$).
All probabilities can be seen in the tree in figure 3 (page 11).
To get the total probabilities for *right* and *wrong* choices, we can sum up the paths.

Note that we ignore the upper half of the tree in summing the paths up, since we can not derive probabilities for the first branching. We just add those cases individually to our function.
We get the following results:

$$\begin{aligned}
P(right|N \leq M) &= 1 \\
P(wrong|N \leq M) &= 0 \\
P(right|N > M, c \in M) &= \frac{m}{n} \\
P(wrong|N > M, c \in M) &= 0 \\
P(right|N > M, c \notin M) &= \left(1 - \frac{m}{n}\right) \frac{1}{n-m} \\
P(wrong|N > M, c \notin M) &= \left(1 - \frac{m}{n}\right) \left(1 - \frac{1}{n-m}\right)
\end{aligned}$$

We can now pick $P(right|N > M, c \in M)$ and $P(right|N > M, c \notin M)$ to calculate the marginal probability $P(right|N > M) = P(right|N > M, c \in M) + P(right|N > M, c \notin M)$.

$$P(right|N > M) = \frac{m}{n} + (1 - \frac{m}{n})\frac{1}{n-m} = \frac{m+1}{n}$$

Finally we can define $P_{right}(N, M)$ with $P(right|N \leq M)$ and $P(right|N > M)$:

$$P_{right}(N, M) = \begin{cases} 1 & \text{if } N \leq M \\ \frac{|M|+1}{|N|} & \text{if } N > M \end{cases}$$

4 Probability Refresher II *2014-04-25*

4.1 Rules of Probability

Assume a hat filled with cards. Each card has a red and a blue side, the red sides are labeled from 1 to 6 and blue sides from 1 to 4, resulting in 24 different cards. We can describe this **sample space** (or *set of all possible outcomes*) Ω with

$$\Omega = \{[1,1], [1,2], ..., [6,4]\}$$

where $[1,2]$ represents the card with 1 on the red and 2 on the blue side.
To describe the following examples we first have to define some terms and relations.

- $x \in \Omega$ is the **random variable** x which can be drawn from Ω.
  - Example: $[1,2]$, i.e. the card with a red 1 and a blue 2.
- $|S|$ where S is a set is the **number of elements** in S.
  - Example: $|\{1,2\}| = 2$
- $E \subset \Omega$ is an **Event** E (*something you can bet on*).
  - Example: Drawing a card with the number on the red site smaller than number on its blue site.

How many events do we have? The number of events is simply $2^{|\Omega|} = 2^{24} = 2^{10} \cdot 2^{10} \cdot 2^4 \approx 16$ Mio.
If we now put another $[1,1]$ card into the hat, the *number of events stays the same, but the probabilities change*. This can be seen in the following table 1.

		red					
		1	2	3	4	5	6
blue	1	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$
	2	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$
	3	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$
	4	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$

Table 1: Probabilities of drawing a specific card from the hat

To get the probability of an event we just have to sum up the values in table 1. For example for the event "The red number is smaller than the blue number", let's call it $R < B$, the probability $P(R < B)$ can be calculated as follows:

$$\begin{aligned} P(R<B) &= P(R=1,B=2) + P(R=1,B=3) + P(R=2,B=3) \\ &\quad + P(R=1,B=4) + P(R=2,B=4) + P(R=3,B=4) \\ &= 6 \cdot \frac{1}{25} = \frac{6}{25} \end{aligned}$$

4.2 Axioms of Probability

1. $P(\{\}) = 0, P(\Omega) = 1$
The probability for the empty set is 0. The probability for any event to happen is 1.

2. $\forall E : 0 \leq P(E) \leq 1$
Probabilities are in the range from 0 to 1.

3. `if` $E = E_1 \cup E_2$ `and` $E_1 \cap E_2 = \{\}$ `then` $P(E) = P(E_1 \cup E_2) = P(E_1) + P(E_2)$
If two events don't intersect their combined probability is the sum of their individual probabilities.

3'. *(follows from 3)*

`if` $E = \bigcup_{i=1}^{n} E_i$ `and` $\forall_{i,j;i\neq j} E_i \cap E_j = \{\}$

`then` $P(E) = P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$

If n events don't intersect their combined probability is the sum of their individual probabilities.

Note that we can calculate P for all events by adding up singleton events.

Other rules that follow

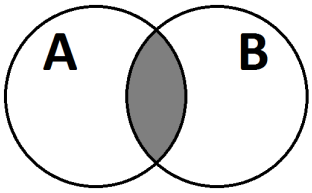


Figure 4: The intersection between two sets makes math a bit tricky

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A \setminus B) + P(A \cap B) + P(B \setminus A)$
The probability that one of two events happens is their individual probabilities minus the probability that both events happen simultaneously (otherwise we would account for that case twice, see also figure 4).

2. $P(A \cup \neg A) = P(\Omega) = 1 = P(A) + P(\neg A)$
The probability that an event happens or not is 1.

3. $P(A) = 1 - P(\neg A)$
The probability of an event to not happen is 1 minus the probability of the event (and vice versa).

4.3 Random Variables & Joint Distribution

An example for a joint distribution: You roll two dice, one is six-sided and red, the other one is four-sided and blue.

$$\text{R: } \Omega_R = \{1, ..., 6\} \qquad P(R = \omega) = \frac{1}{6} \ \forall\ \omega \in \Omega_R$$

$$\text{B: } \Omega_B = \{1, ..., 4\} \qquad P(B = \omega) = \frac{1}{4} \ \forall\ \omega \in \Omega_B$$

The joint sample space Ω is $\Omega = \Omega_R \times \Omega_B$.
Since R and B are independent the joint probability is $P(R, B) = \frac{1}{24}$ for each value of R, B.
More formally speaking it holds that $P(R = i, B = j) = \frac{1}{24}$ for all values of i, j since $P(R, B) = P(R) \cdot P(B)$ for all independent R, B.

4.4 Marginal and Conditional Probability

For the following examples please refer to table 1 (page 13).

Marginal Probability

The **Marginal Probability** for $P(R = j)$ is:

$$P(R = j) = \begin{cases} \frac{5}{25} & \text{if } j = 1 \\ \frac{4}{25} & \text{else} \end{cases}$$

This can be calculated by summing up one dimension of the table.

$$P(R = j) = \sum_{i \in \Omega_B} P(R = j, B = i)$$

This can be written a bit more casual (here for B now):

$$P(B) = \sum_{R} P(R, B) = \begin{cases} \frac{7}{25} & \text{if } B = 1 \\ \frac{6}{25} & \text{else} \end{cases}$$

Conditional Probability

In case a card was picked and we already know what number the red side shows, $P(R, B) \neq P(R) \cdot P(B)$ is *not independent*. $P(R, B)$ is now dependent on the already known red number.
The probabilities that follow are:

$$P(B|R = 1) = \begin{cases} \frac{2}{5} & \text{if } B = 1 \\ \frac{1}{5} & \text{else} \end{cases}$$

$$P(B|R = 2, ..., 6) = \frac{1}{4}$$

The conditional probability $P(A|B)$ (read: P of A given B) can be expressed as follows:

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

where

$P(A, B)$ is the joint pobability

$\sum_{A} P(A, B)$ is the marginal probability

$P(A, B)$ is a function of A (because B is fixed)

$P(B)$ is the renorm

With the product rule $P(A|B)P(B) = P(A, B) = P(B|A)P(A)$ we can derive **Bayes' rule**:

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

where we call

$P(B|A)$ posterior

$P(A|B)$ likelihood

$P(B)$ prior

$P(A)$ evidence

PMPC Tutorial Sheet 2

1. What is the sample space in roulette? What is the event space in roulette? How many possible events are there? What is the difference between betting \$18 on the even numbers and betting \$1 on each even number? How does the bank make money?

2. Prove the so-called union bound:

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P\left(E_i\right)$$

3. The following expression is a useful variant of Bayes' rule:

$$\frac{P(A \mid B)}{P(\text{not } A \mid B)} = \frac{P(B \mid A)}{P(B \mid \text{not } A)} \cdot \frac{P(A)}{P(\text{not } A)}.$$

   The left-hand side are the posterior odds that A is true. The first term on the right-hand side is called the likelihood ratio; the second term are the prior odds. Convince yourself that it is correct.

4. In a famous series of studies Daniel Kahneman and Amos Tversky examined judgements under uncertainty. In one of their experiments participants were given the following judgement task [4]:
   A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:
   (a) 85% of the cabs in the city are Green and 15% are Blue
   (b) a witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time. What is the probability that the cab involved in the accident was Blue rather than Green?
   A large number of subjects were tested and the mode and the median of the answers is 80%. What is the correct answer? What is going wrong?

5. Is medical screening sensible [1, 2]? To diagnose colorectal cancer the hemocult test is conducted to detect occult in the stool. For symptom-free people over 50 years old who participate in screening using the hemocult

test the following information is available:
Thirty out of 10,000 people have colorectal cancer. Of these 30 people with the cancer, 15 will have a positive test result (The hemocult test is not very sensitive, its hit-rate is only 50%). Of the remaining 9,970 people without the cancer, 300 will still have a positive hemocult test (the false-alarm-rate is about 3%). Imagine a sample of people (over 50, no symptoms) who have positive hemocult tests. How many of these people actually have colorectal cancer?
What are the advantages of screening, i.e. testing a large number of people who have no symptoms? What are the disadvantages?

6. For the last exercise write out the full joint probability table. Write out both conditional distributions. Calculate both marginal distributions. What would the joint distribution look like if the result of the hemocult test was independent of the patient having colorectal cancer but the marginal distributions were the same? What would the conditional distributions look like in this case?

7. The Monty Hall Problem (the presentation that Gerd Gigerenzer [1, p. 217] gives is worthwhile reading). For about three decades, Monty Hall hosted a popular American game show called *Let's Make a Deal*. The final contestant is given the choice of three doors. Behind one door is a car, behind the others goats. The contestant picks a door, say number 1, and the host, who knows what's behind the doors, opens another door, say number 3, which has a goat. He asks the contestant whether he or she wants to switch to door number 2. Should the contestant switch? Why?

8. Read the paper on learning Bayes' Rule by Sedlmeier and Gigerenzer [3].

References

[1] G. Gigerenzer. *Reckoning with Risk*. Penguin Books Ltd, 2003.

[2] G. Gigerenzer, W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz, and S. Woloshin. Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8(2):53–96, 2007.

[3] P. Sedlmeier and G. Gigerenzer. Teaching bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130(3):380–400, 2001.

[4] A. Tversky and D. Kahneman. Evidential impact of base rates. In D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgment under Uncertainty: Heuristics and Biases*, pages 153–160. Cambridge University Press, 1982.

6 Solution 2: Probability Refresher II *2014-05-05*

Exercise 1

There are different types of roulette, we will only deal with the so-called "European" one, which features only one 0, no 00.
The sample space Ω is $\Omega = \{0, 1, 2, 3, ..., 36\}$. The event space is $\mathcal{P}(\Omega) \setminus \Omega$, the number of events $2^{|\Omega|} = 2^{37} = 2^{10} \cdot 2^{10} \cdot 2^{10} \cdot 2^7 \approx 128,000,000,000$ (128 Bio).
The expected values are:

$$E(even) = 18 \cdot \frac{18}{37} - 18 \cdot \frac{19}{37} = -0.49$$
$$E(even\ numbers) = \left(35 \cdot \frac{1}{37} - 1 \cdot \frac{36}{37}\right) \cdot 18 = -0.49$$

So basically both variants are the same.

However, in European roulette there usually exists the so-called "en prison"-rule. This rule freezes ("imprisons") the stakes made on the small bets ("odd", "even", "red", "black", "high", "low") when 0 comes up until the specific bet was fulfilled twice in a row. When it came up twice in a row, you are allowed to get your stakes back. Alternatively, if you want to play with your bets, you can keep half of them instead of letting them being frozen, losing the other half. This modifies the expected value (assuming *return* and *freeze* are used with a chance of 0.5):

Note: *This paragraph is from our homework research. In the lecture we just dealt with the scenario above.*

$$E(even, 0, freeze) = 18 \cdot \frac{18}{37} - 18 \cdot \frac{18}{37} - 0 \cdot \frac{1}{37} = 0$$
$$E(even, 0, return) = 18 \cdot \frac{18}{37} - 18 \cdot \frac{18}{37} - 9 \cdot \frac{1}{37} = -0.24$$
$$E(even, 0) = \frac{1}{2}\left(E(even, 0, return) + E(even, 0, freeze)\right) = -0.12$$
$$E(even) = \frac{\frac{1}{37}E(even, 0) + \frac{36}{37}E(even, \neg 0)}{2} = -0.48$$

With the "en prison" rule betting on "even" is therefore better than betting on all even numbers.

The bank makes money because of the 0: For calculating the odds they assume not 37 but 36 numbers.

Exercise 2

The union-bound can be proved by induction.

Induction assumption

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n}(P(E_i)) \tag{1}$$

Base case
For $n = 1$ we have

$$P\left(\bigcup_{i=1}^{1} E_i\right) \leq \sum_{i=1}^{1}(P(E_i)) \tag{2}$$
$$\Leftrightarrow \quad P(E_1) \leq P(E_1) \tag{3}$$

Inductive Step
We know that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, so it follows that:

$$P\left(\bigcup_{i=1}^{n+1} E_i\right) = P\left(\bigcup_{i=1}^{n} E_i\right) + P(E_{n+1}) - P\left(\bigcup_{i=1}^{n} E_i \cap E_{n+1}\right) \tag{4}$$

$$\overset{\text{IA}}{\leq} \sum_{i=1}^{n}(P(E_i)) + P(E_{n+1}) = \sum_{i=1}^{n+1}(P(E_i)) \tag{5}$$

Exercise 3

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

Putting it together:

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

This is like:

$$\textit{Posterior Odds} \;=\; \textit{Likelihood Ratio} \;\cdot\; \textit{Prior Odds}$$

We can say this is updating your beliefs with new information.

Exercise 4

We use Bayes' rule to calculate P(Blue Car|Testified by Witness)

$$P(B|T) = \frac{P(T|B)P(B)}{P(T)}$$

We know that $P(T|B) = 0.8$ and $P(B) = 0.15$, so we only have to find out $P(T)$. Therefore we marginalize over B:

$$\sum_{B} P(T|B)P(B) = 0.8 \cdot 0.15 + 0.2 \cdot 0.85 = 0.29$$

Now we can take our values and calculate the correct answer:

$$P(B|T) = \frac{P(T|B)P(B)}{P(T)} = \frac{0.8 \cdot 0.15}{0.29} = 0.414$$

When testing a large group of people, the common mistake becomes obvious: they take the probability that the witness correctly identified the color, which is 80 %. This is because they neglect the base rate (*base rate neglect*).

Exercise 5

We take Bayes' rule to calculate P(Cancer|Positive Test)

$$P(C|P) = \frac{P(P|C)P(C)}{P(P)}$$

We know that $P(P|C) = 0.5$ and $P(C) = 0.003$, so we only have to find out $P(P)$. Therefore we marginalize over C:

$$\sum_C P(P|C)P(C) = 0.5 \cdot 0.003 + 0.03 \cdot 0.997 = 0.0314$$

Now we can take our values and calculate the correct answer:

$$P(C|P) = \frac{P(P|C)P(C)}{P(P)} = \frac{0.5 \cdot 0.003}{0.0314} = 0.048$$

As we can see, only 4.8 % of the people being tested positively actually have cancer. Nevertheless screening is a good thing because out of 1000 detections still 50 have cancer and those can be treated which probably makes up for the disadvantage of having 950 false alarms.

Exercise 6

$P(P, C)$

	Cancer	¬ Cancer	
Positive Test	$\frac{15}{10{,}000}$	$\frac{300}{10{,}000}$	$\frac{315}{10{,}000}$
Negative Test	$\frac{15}{10{,}000}$	$\frac{9{,}670}{10{,}000}$	$\frac{9{,}685}{10{,}000}$
	$\frac{30}{10{,}000}$	$\frac{9{,}970}{10{,}000}$	$\frac{10{,}000}{10{,}000}$

$P(P|C)$

	Cancer	¬ Cancer
Positive Test	$\frac{15}{30}$	$\frac{300}{9{,}970}$
Negative Test	$\frac{15}{30}$	$\frac{9{,}670}{9{,}970}$
	1	1

$P(C|P)$

	Cancer	¬ Cancer	
Positive Test	$\frac{15}{300+15}$	$1 - \frac{15}{300+15}$	1
Negative Test	$\frac{15}{15+9{,}670}$	$\frac{9{,}670}{15+9{,}970}$	1

Assume independence $P(C, P)$
$P(C|P) = P(C)$

	Cancer	¬ Cancer	
Positive Test	$\frac{15+300}{10{,}000} \cdot \frac{15+15}{10{,}000}$	$\frac{15+300}{10{,}000} \cdot \frac{300+9{,}670}{10{,}000}$	$\frac{15+300}{10{,}000}$
Negative Test	$\frac{15+9{,}670}{10{,}000} \cdot \frac{15+15}{10{,}000}$	$\frac{15+9{,}670}{10{,}000} \cdot \frac{300+9{,}670}{10{,}000}$	$\frac{15+9{,}670}{10{,}000}$

Exercise 7

As the probability for each door to contain a car is equal we know that our chance to win is given by $P(\text{Win}) = \frac{1}{3}$. Further the probability that the Quiz-master chooses one of the remaining doors is $P(\text{Open}) = \frac{1}{2}$ The Quiz-master would not open the door which contains the car, so $P(\text{notOpen}|\text{Win}) = 1$. Now if calculate the probability of the car being behind the unopened door using Bayes' we get:

$$P(Win|notOpen) = \frac{P(notOpen|Win) \cdot P(Win)}{P(notOpen)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Therefore the probability that the car is behind the door that the Quiz-master did not open is $\frac{2}{3}$ while the probability of our first chosen door still is $\frac{1}{3}$. So we would double our chances to win by changing our choice.

7 Measuring Beliefs I *2014-05-02*

7.1 Probability as Belief

We can measure probabilities for recurring events, how can we measure probabilities for unique events? Unique events are for example:

- How sure are people which population is bigger, the EU or the US population?
- How can bookmakers set the odds for soccer games?
- How high is the probability for a nuclear reactor to blow up?

The frequentist view is not really helpful here: Since those events don't appear numerous times, you can not measure any limits of relative frequencies. But the Bayesian view helps us to use the same math to determine these probabilities.

7.2 What do you accept as a fair bet?

Figure 5: A horse race ticket. *Source: Reuben Goossens, ssmaritime.com*

Let's assume for the next examples that people are honest (Otherwise they would lie to win).
Assume you have a ticket you can exchange for \$1 if A happens, otherwise it's worth nothing.

$$Ticket = \begin{cases} \$1 \text{ if } A \\ \$0 \text{ else} \end{cases}$$

What would be a fair price for that ticket?

$$\begin{aligned} (\$1 - c)\, P(A) - cP(\neg A) &= 0 \\ \Leftrightarrow P(A) &= c \end{aligned}$$

Coherence (fair pricing)

1. $P(certain) = 1$, $P(impossible) = 0$
2. $\forall A\ 0 \leq P(A) \leq 1$
3. $P(A \cap B) = \{\} \rightarrow P(A \cup B) = P(A) + P(B)$

These rules follow from some logical thoughts.
Imagine $P(A) + P(\neg A) > 1$. Then the bookmaker would make money and the bet wasn't fair. If you are not the bookmaker, you want to have something like $P(A) + P(\neg A) < 1$.
Another case is $A \cap B = \{\}$, i.e. A and B are mutually exclusive. Then you want to buy $P(A) + P(B)$ but sell $P(A \cup B)$ if $P(A) + P(B) < P(A \cup B)$.

7.3 Conditional Bets

If we don't have repeatable events, how can we justify conditional probabilities?
Assume a ticket again, this time of the following form:

$$Ticket = \begin{cases} \$1 \text{ if } A \cap B \\ \$c \text{ if } \neg B \textit{ (refund)} \\ \$0 \text{ else} \end{cases}$$

A is dependent on B now.

$$\begin{aligned} P(A|B) &= P(A \cap B) + P(\neg B)P(A|B) \\ 1 &= \frac{P(A \cap B)}{P(A|B)} + 1 - P(B) \\ P(B) &= \frac{P(A \cap B)}{P(A|B)} \\ P(A|B)P(B) &= P(A \cap B) \\ P(A|B) &= \frac{P(A \cap B)}{P(B)} \end{aligned}$$

Philosophical differences matter

Alice has two coins, coin 1 with a probability of 0.5 for heads and tails, coin 2 with probability 0.4 for heads (and 0.6 for tails).
She chooses a coin and tells Bob she would flip it n times now. Then Bob has to guess which coin she flipped.
Bob has two hypotheses, one for each coin. To check his hypotheses, he can now use the data (i.e. the n coin flips) and calculate the probabilities for his hypotheses - then he can compare those and choose the one with higher probability.

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

(with H = Hypothesis, D = Data)

Calibration and Coherence

Note that there is a difference between coherence and calibration. You are well calibrated if you answer according to your real beliefs and knowledge. For example if you play Roulette you do bet although you know that you can lose because of the 0, so you are not well calibrated (a well calibrated person in that case would not play). Being coherent means that you follow the rules of probability, for example that you don't trip into the conjunction fallacy trap (see the chapter about "Conditional Bets" above).

Note: *In short: Being ill-calibrated means you lose money on average, while being incoherent means you lose it.*

PMPC Tutorial Sheet 3

1. Go back to last week's tutorial sheet. Why did you make the mistakes you did? What did you not understand before?

2. DNA evidence in court [1]. In a rape-murder case the police found traces of DNA on the victim that did not belong to the victim. The police had a couple of suspects but their DNA did not match the trace on the victim. Further examination revealed that the DNA did match with DNA found in an unsolved murder case that happened in the same region a year before. The police decides to ask all men in the region that are aged between 20 and 50 to participate in voluntary DNA testing. There are about 100,000 of them—but of course not all of them participate. If two DNA samples come from the same person a test will reveal this with practically perfect certainty. Sometimes a match will be found even if the samples do not come from the same person. With current DNA technology this happens in about 0.001 percent of the cases. A match is found in one of the men who were screened. What is the probability that the DNA on the victim is really from this man? If you were his lawyer how would you argue?

3. If you are philosophically inclined and want to know more about the subjective interpreation of probabilities and how probabilities, betting, and logic relate to each other, read chapter 1 of [2]. I closely followed the presentation given there in class.

4. In 2004 the European soccer championships took place in Portugal. In the final Greece won over Portugal with a score of one to zero. Before Greece reached the final hardly anyone had thought it possible that Greece could become European champion. In fact the odds for this to happen were judged to be about 1:100 before the tournament started. Even though Greece had beaten Portugal in the opening game the bookmakers set the odds for the final as: 7:2 for Portugal to win and 2:5 for Greece to win. Two men (betters are mostly men) placed bets with the bookmaker. Adam bet 10 Euros on Portugal and Boris bet 10 Euros on Greece. Use the odds to calculate how much money Boris has won and how much money Adam would have won.

5. If the odds for Portugal are 7:2 then the odds for Greece should be 2:7. Why is this not so? Charly placed two bets: 4.90 Euros on Portugal and 1.80 Euros on Greece. How much did he win and how much would he have won in case Portugal had become champion?

6. Before the match your belief in Greece winning was p. For which values of p would you have been tempted to place a bet? Why do people bet at all?

7. The conjunction fallacy revisited. Someone believes that $P(A \cap B) > P(A)$. Accordingly, he is willing to buy or sell a ticket worth \$1 if $A \cap B$ and \$0 otherwise at a price of $\$P(A \cap B)$. And similarly for $P(A)$. How can you take advantage of this person?

8. Geometric distribution. Repeatedly toss a thumbtack until it lands on the pin. Let N be the random variable that gives the number of tosses needed until the thumbtack landed on the pin. What is the sample space for N? What is the probability distribution $p(N)$? Convince yourself that the distribution is normalized to 1. What is the probability for N being odd?

9. St. Petersburg paradox. Someone offers you the following gamble: A fair coin is tossed until head comes up. Call the number of tosses needed N. You get 2^N Euros. But before you are allowed to play the game you have to pay 1000 Euros. For which N would you win money? What is the probability that you lose money? Would you play? What is your expected gain for this gamble? What does this tell you about using fair bets as a way to elicit beliefs?

10. Imagine playing the following game with a friend. She has two dice, a red one and a blue one. Behind your back she rolls the dice. After each roll she tells you a number between 1 and 6. The rules of the game are that your friend either always reports the number that the red die shows or she always reports the minimum of the two dice. Your task is to figure out what she is doing. The results for the first 10 rolls have been 4, 2, 5, 6, 6, 3, 2, 6, 6, 1. What are your beliefs given these data?

References

[1] G. Gigerenzer. *Reckoning with Risk*. Penguin Books Ltd, 2003.

[2] R. Jeffrey. *Subjective Probability. The Real Thing*. Cambridge University Press, 2004.

9 Solution 3: Measuring Beliefs I *2014-05-19*

Exercise 2

Before we can calculate any probabilities, we make some assumptions about the circumstances of the case.
First we assume the number of subjects taking part in the DNA test is still 100,000.
Further we assume that the murderer really is one of the men tested (and not an outsider or deceased or...) which renders the probability of being guilty ($P(G)$) as

$$P(G) = \frac{1}{100,000}$$

and therefore

$$P(\neg G) = \frac{99,999}{100,000}$$

A positive match with the correct DNA (i.e. matching the guilty person) has the probability of $P(P|G) = 1$.
The actual probability we are interested in is the one of really being guilty given a positive test result:

$$P(G|P) = \frac{P(P|G) \cdot P(G)}{P(P)}$$
$$P(G|P) = \frac{1 \cdot \frac{1}{100,000}}{P(P)}$$

To get the missing factor, $P(P)$ we can use the marginal probability:

$$\begin{aligned} P(P) &= \sum_{i=1}^{N} (P(P|G_i) \cdot P(G_i)) \\ &= P(P|G) \cdot P(G) + P(P|\neg G) \cdot P(\neg G) \\ &= 1 \cdot \frac{1}{100,000} + 0.00001 \cdot \frac{99,999}{100,000} \\ &= \frac{1.99999}{100,000} \end{aligned}$$

Now we can calculate $P(G|P)$:

$$P(G|P) = \frac{1 \cdot \frac{1}{100,000}}{\frac{1.99999}{100,000}} = \frac{1}{1.99999} > \frac{1}{2}$$

That means the actual probability that the accused person really is guilty is slightly above 50 %. At first glance in court it may seem very likely that the defendant is the murderer. But we as clever attorneys can argue that with 100,000 participants taking part in the DNA test, the chance for a false alarm is still pretty high, even though the test is so accurate. This is again a case of neglecting the base rate.

Of course the probability of 50 % is just an upper bound and can be much less for fewer participants but nevertheless shows the huge mistake one makes by neglecting the effect of a high base rate.

Exercise 4

Boris bet 10€ on Greece with odds of 2 : 5. The bookmaker bet $10 \cdot \frac{5}{2}€ = 25€$ on Portugal. Boris won 35€, which is a gain of 25€.
Adam bet 10€ on Portugal with odds of 7 : 2. The bookmaker bet $10 \cdot \frac{2}{7}€ = 2.86€$ on Greece. Adam would have won 12.86€, which would have been a gain of 2.86€.

Exercise 5

Charly's bets were $4.90€ + 1.80€ = 6.70€$ in total.

$$E(Portugal) = 4.90 \cdot \frac{2}{7} - 1.80 = -0.4[€]$$
$$E(Greece) = 1.80 \cdot \frac{7}{2} - 4.90 = 1.4[€]$$

If the odds were a fair bet, i.e. $7 : 2$ for Portugal and $2 : 7$ for Greece, Charly had either lost $0.40€$ in case Portugal won, or he'd gained $1.40€$ in case Greece won.

The bookmaker however wants to make money, that's why he fixes the odds in a way that he will earn some.

> With $7 : 2$ for Portugal Charly's gain in case of Portugal's victory will be $4.90 \cdot \frac{2}{7} - 1.80 = -0.40[€]$, he in fact loses money.
>
> With $2 : 5$ for Greece Charly's gain in case of Greece's victory will be $1.80 \cdot \frac{5}{2} - 4.90 = -0.40[€]$, again he loses money in total.

So the bookmaker would always win some money (those $0.40€$ Charly lost in each case).

Exercise 6

$$\begin{aligned} &\frac{p}{1-p} \geq \frac{2}{5} \\ &\Leftrightarrow 5p \geq 2 - 2p \\ &\Leftrightarrow 7p \geq 2 \\ &\Leftrightarrow p \geq \frac{2}{7} \end{aligned}$$

If the probability for Greece to win was higher than $\frac{2}{7}$, we might have been tempted to place a bet.
People bet because they hope to win or assume they have a dutch book (i.e. a guaranteed win). And they will bet if they really have a dutch book.
To check if a dutch book is possible one can do a simple calculation:

$$\frac{7}{7+2} + \frac{2}{2+5} \overset{?}{<} 1$$

That is checking if the probabilities from the bookmaker's view are less or greater than 1. If they are greater, you will lose - if their sum is smaller, you can find a dutch book.

Exercise 7

The person believes that $P(A \cap B) > P(A)$.
We can buy cheap $P(A)$ tickets and sell expensive $P(A \cap B)$ tickets, so even before we know the outcome we make money.
In practice now the other person has $P(A \cap B)$, we have $P(A)$.
Now we see what could happen:

$$\begin{aligned} A \cap B &: \text{We are even.} \\ A \cap \neg B &: \text{I get \$ 1.} \\ \neg A \cap B &: \text{All tickets lose.} \\ \neg A \cap \neg B &: \text{All tickets lose.} \end{aligned}$$

So either both parties have the same outcome or we win.

Exercise 8

Note: pin $:= 1 - p$*, head* $:= p$
The sample space is $\Omega = \mathbb{N}$ because we can get pin on the first, second, third, or n-th throw.
We calculate the probability distribution.

$$p(N = n) = p^{n-1}(1 - p)$$

We can use *countable additivity* and add the numbers up to infinity.

$$\begin{aligned} 1 &\overset{?}{=} \sum_{n=1}^{\infty} p(N = n) = \sum_{n=1}^{\infty} p^{n-1}(1 - p) \\ &= (1 - p) \sum_{n=0}^{\infty} p^n = (1 - p) \frac{1}{(1 - p)} \overset{!}{=} 1 \end{aligned}$$

For odd N the probability is accordingly:

$$\begin{aligned} P(\text{N is odd}) &= \sum_{n=0}^{\infty} p(N = 2n + 1) \\ &= \sum_{n=0}^{\infty} (1 - p) p^{2n} \\ &= (1 - p) \sum_{n=0}^{\infty} \left(p^2\right)^n \\ &= \frac{1 - p}{1 - p^2} = \frac{1}{1 + p} \end{aligned}$$

Exercise 9

N	1	2	3	...	n
probability	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2^3}$	...	$\frac{1}{2^n}$
Win	2^1	2^2	2^3	...	2^n

The probability that we lose money is for all N where we get less than 1000 $.

$$2^N > 1000 \rightarrow N \geq 10$$

$$p(\text{lose money}) = p(N < 10) = \sum_{n=1}^{9} \frac{1}{2^n} \approx 0.998$$

Now the expected gain is:

$$\begin{aligned} E &= -1000 + \frac{1}{2} \cdot 2^1 + \frac{1}{4} \cdot 2^2 + \frac{1}{8} \cdot 2^3 + ... \\ &= -1000 + \sum_{n=1}^{\infty} \frac{1}{2^n} \cdot 2^n \\ &= -1000 + \sum_{n=1}^{\infty} 1 \\ &= +\infty \end{aligned}$$

Although this looks rather promising, of course we don't play the game.

Exercise 10

We just check with which model (i.e. with which die) the given data set is more probable - that is the best guess we can make.

$$
\begin{aligned}
P(min|D) &= \frac{P(D|min)P(min)}{P(D)} \\
P(red|D) &= \frac{P(D|red)P(red)}{P(D)} \\
P(red) &= \frac{1}{2} \\
P(min) &= \frac{1}{2} \\
P(D|min) &= \frac{5}{36} + 2 \cdot \left(\frac{9}{36}\right) + \frac{3}{36} + 4 \cdot \left(\frac{1}{36}\right) + \frac{7}{36} + \frac{11}{36} = \frac{48}{36} = \frac{4}{3} \\
P(D|red) &= 10 \cdot \left(\frac{1}{6}\right) = \frac{10}{6} = \frac{5}{3}
\end{aligned}
$$

$P(D|min) < P(D|red)$, so she most probably decided on the red die.

10 Measuring Beliefs II *2014-05-09*

10.1 Probabilities of Continuous Random Variables

How tall is Frank Jäkel? 1.80m, 1.70m, 1.68m, 1.69m, or even 1.7034241m?
Not only are there problems with real numbers like 1.7034241, but also with the question the Bayesian view inevitably asks: "What do you think is the probability for that size?"
One sees: continuous random variables are difficult. There is an infinite uncountable range of numbers and one shall assign probabilities for them. This leads straight to the question: "What's the probability of a real number?"
In the following section this problem gets tackled in three ways.

10.1.1 Solution 1: Histograms (Probability Density Function, PDF)

The first and naive way is to discretize the sample space $\mathbb{R}$ into bins and assign probabilities to those bins (figure 6).

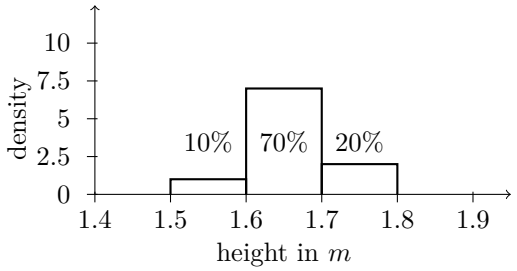


Figure 6: Histogram

Note that **the area describes the probability**. For the second box, one would assign a y-value of 7, such that the width (0.1) times the height equals the probability ($0.1 \cdot 7 = 0.7$).
For a finer granularity one can now change the resolution of the bins and split the probabilities among them (figure 7.

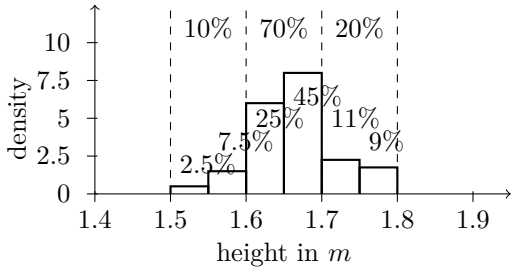


Figure 7: Histogram with higher resolution

Extreme cases Usually this will yield a nice distribution of how beliefs are. However, there are two special extreme cases.

> The first case is that all values are equally probable: Since we have infinitely many values on the real number line, the probability for each individual value is $\frac{1}{n} \Rightarrow \lim_{n\to\infty} \frac{1}{n} = 0$.
>
> The second case is that the full probability gets assigned to one single value. Since a single value has the width 0, again the probability will become 0 for all values.

The probability density function However, if the resulting histogram is somewhere between those extreme cases, then the limit of the distribution yields the probability density function (figure 8).

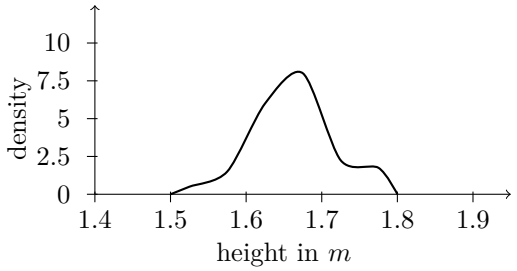


Figure 8: Probability density function

As mentioned before it's still not possible to calculate the probability of a specific number. What is possible though, is calculating the probability of an interval. This is useful, since people always bet on intervals. For instance, betting on "2" means to bet on the interval $[2, 3[$.

10.1.2 Solution 2: Cumulative Density Function, CDF

To calculate the probability of an interval it's possible to simply sum up all probabilities in the specific interval.
This can be done by integration of the PDF (see figure 9).

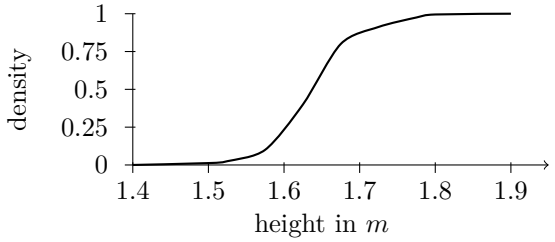


Figure 9: Cumulative density function

10.1.3 Solution 3: Parametric Distribution

The only remaining problem is deriving the correct probability density function. We can avoid this by using a very common statistics method and model the PDF as a Gaussian distribution.
This way we reduce the problem finding the correct function to finding the correct parameters.

The Gaussian distribution (see figure 10) is defined as

$$p(X = x) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\left(\frac{\mu - x}{\sigma}\right)^2} = \phi(x; \mu, \sigma) = \phi\left(\frac{\mu - x}{\sigma}; 0, 1\right)$$

$$\frac{\mu - x}{\sigma} = z$$

The corresponding integral is Φ (see figure 10).

$$P(X \leq t) = \int_{-\infty}^{t} p(X = x)dx = \Phi(t; \mu, \sigma)$$

The area of the standard deviation ($\mu - \sigma \leq X \leq \mu + \sigma$) has a probability of approximately 68 % (see figure 11).

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = \Phi(\mu + \sigma; \mu, \sigma) - \Phi(\mu - \sigma; \mu, \sigma) \approx 68\%$$

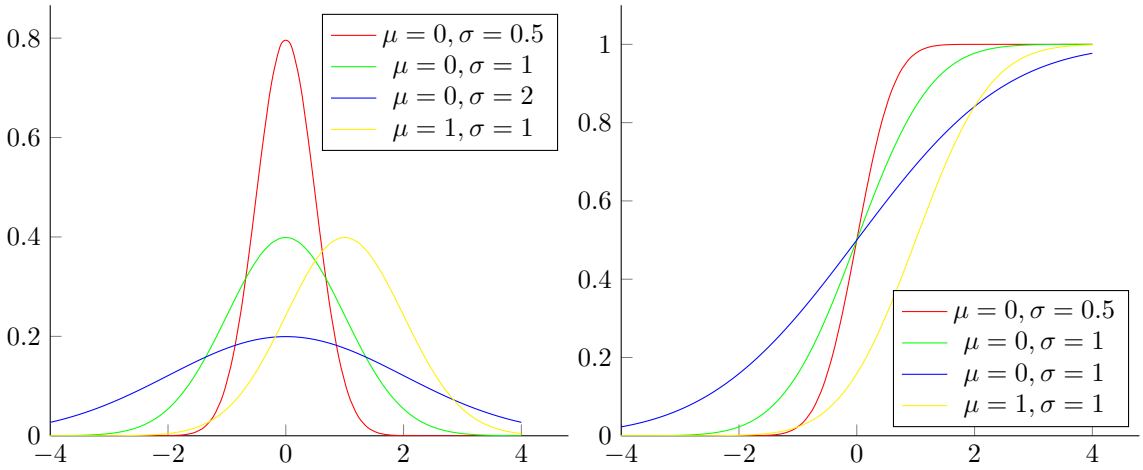


Figure 10: Left: Gaussian distributions. Right: Their corresponding integrals. Parameters: $\mu = 0$ and $\sigma = 0.5, 1, 2$, and $\mu = 1, \sigma = 1$.

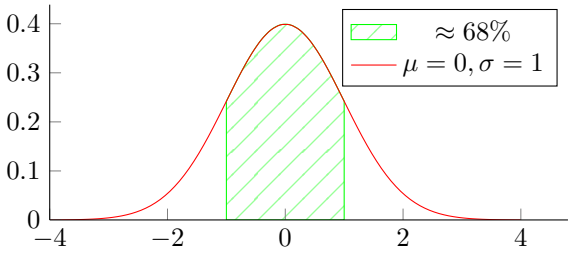


Figure 11: The area of the standard deviation yields approximately 68 %

We can also find other useful probabilities which are commonly used to do statistics:

- $\mu \pm \sigma \approx 68\%$
- $\mu \pm 2\sigma \approx 95\%$
- $\mu \pm 3\sigma \approx 99\%$

10.2 Proper Scoring Rules

Multiple choice tests would be better if you would state "how your belief is, that this is right", rather than just answering the question (For more about this see page 39).

Take a look at this example:
The EU population is bigger than the US population.
Give the belief for this to be true.
(This means 0 = "I believe this is false", 1 = "I believe this is true", 0.5 = "I don't know")

The aim of a proper scoring rule is to yield a high gain (a minimum loss) if the answer is true and the belief in it is high, but yield no gain if the answer is false but the belief in it high.
In the loss-function L given below q is the belief assigned to the answer given and X is 1 if the statement was true or 0 if it was false.

$$\begin{aligned} L(X, q) &= (X - q)^2 \\ &= X^2 - 2qX + q^2 \text{ (note: } X^2 = X, \text{ since only 0 or 1)} \\ &= X(1 - 2q) + q^2 \end{aligned}$$

Penalty for lying

If we now assume the subject is not stating her actual belief p, but another value q ($p \neq q$), the formula changes in the following way:

$$\begin{aligned} E\left(L\left(p,q\right)\right) &= p - 2pq + q^2 = \left(p - p^2\right) + \left(p^2 - 2pq + q^2\right) \\ &= \underbrace{p\left(1-p\right)}_{\text{basic loss}} + \underbrace{\left(p-q\right)^2}_{\text{penalty for lying}} \end{aligned}$$

Note that the basic loss is maximal if the subject has no clue ($p = 0.5$).

How to answer?

If one's belief is p, which q will yield the best gain (i.e. will minimize the expected loss)?
To answer this the expected loss function can be minimized, i.e. one can search the first derivative and set it to zero.

$$\begin{aligned} E\left(L\left(p,q\right)\right) &= p\left(1-2q\right) + q^2 \\ \frac{\partial E\left(L\left(p,q\right)\right)}{\partial q} &= -2p + 2q \\ \rightarrow 0 &= -2p + 2q \\ \Leftrightarrow p &= q \end{aligned}$$

As can be seen the loss function is minimal if $p = q$, i.e. if the person answering is telling the truth.

PMPC Tutorial Sheet 4

1. On average I arrive at the office at a certain time in the morning. However, there is some variability due to traffic. I believe that I'm never more than w minutes late and I never show up earlier than w minutes before my average time of arrival. Let's call X the difference between my actual time of arrival and my average time of arrival. Here's my belief about X:

$$p(X = x) \quad = \quad \begin{cases} \frac{x}{w^2} + \frac{1}{w} & \text{if } -w \leq x \leq 0 \\ -\frac{x}{w^2} + \frac{1}{w} & \text{if } 0 < x \leq w \\ 0 & \text{otherwise} \end{cases}$$

   Convince yourself that this probability density function is normalized to 1. What is the value of the density function at 0? Why can this be greater than 1? I have an appointment at my office at $w/2$ minutes after my average time of arrival. What's the probability that I'll be late for the appointment? What is the probability that I arrive exactly at $w/2$? For which time point am I 95% certain that I'll be there by then [4, 7-2, 7-3].

2. X has the probability density function (pdf) $p(X = t) = f(t)$ and the cumulative distribution function (cdf) $P(X \leq t) = F(t)$. Define a new random variable $Y = G^{-1}(X)$ for a strictly increasing function G with an inverse G^{-1} and derivative g. What is the cdf for Y? What is the pdf for Y?

3. Sampling from a distribution with pdf g, cdf G, and inverse cdf G^{-1}. X has density

$$p(X = t) = f(t) = \begin{cases} 1 & \text{for } 0 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

   i.e. X is distributed uniformly between 0 and 1. Show that $Y = G^{-1}(X)$ has the pdf $p(Y = t) = g(t)$. Use this fact to write a function that generates random samples from the triangular distribution from above.

4. I don't like multiple choice tests. But I would like them better if there wasn't a lot of guessing involved. Let's try out a method for academic testing that asks you for your belief in a statement, instead of just stating whether it's true or false. I want to try out with you whether such a test can be used for an exam. So we'll simulate an exam by you taking a general knowledge test. The idea is not so much that we really want test your general knowledge here, the idea is to see how well you can quantify your uncertainty about what you know and how well you are calibrated. So don't google the answers until you're done with the test. Nevertheless, the test is simulating an academic testing situation after all, and you do want to score as well as possible. You will be scored with the quadratic loss function that we discussed in class. This means you should be honest and try to be well calibrated. Taking the test should take between 30 and 45 minutes. Please take the test at https://ikw.uni-osnabrueck.de/limesurvey

5. A statement X in the test can be true ($X = 1$) or false ($X = 0$). Say, you answered that your probability is q for $X = 1$. You will be scored using the quadratic loss function

$$L_1(X, q) = (X - q)^2 .$$

This is, however, not the only loss function one could use. One could also use

$$L_2(X, q) = -X \cdot \log(q) - (1 - X) \cdot \log(1 - q) ,$$

i.e. the negative log likelihood. Your true belief in the statement is p. Convince yourself that L_2 is a proper scoring rule, i.e. your expected loss will be minimal if you are honest and $p = q$. Find the minimum of the expected loss with respect to q by taking first and second derivatives. Compare the expected loss for L_1 and L_2 for an honest and well-calibrated test-taker. How does the expected loss vary as a function of p? Make a plot for both loss functions. Which loss function is better? What happens when you are not well calibrated?

6. Bibliographical comments: A good book on how experts' beliefs can be elicited is [3]. This book also has useful background information on subjective probability. Often experts have different subjective beliefs and a decision maker has to come up with a decision based on conflicting expert opinions. If you just want to get a flavour of how experts' forecasts and opinions are scored and combined in practice, you can also look at [2, 1].

References

[1] W. Aspinall. A route to more tractable expert advice. *Nature*, 463(7279):294–295, Jan 2010.

[2] G. W. Brier. Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, 78(1):1–3, 1950.

[3] R. M. Cooke. *Experts in Uncertainty*. Oxford University Press, Oxford, 1991.

[4] F. Mosteller, R. E. Rourke, and G. B. Thomas. *Probability with statistical applications*. Addison-Wesley, 1970.

12 Solution 4: Measuring Beliefs II *2014-05-19*

Exercise 1

In order to calculate the area under the density function we have to take the integral. As the function is defined differently for separate segments we can first calculate the integral from $-\omega$ to 0 and then from 0 to ω (the rest is zero anyways).

$$\begin{aligned}
\int_{-w}^{0} p(x)dx &= \int_{-w}^{0} \frac{x}{\omega^2} + \frac{1}{w}dx \\
&= \left[\frac{1}{2\omega^2} \cdot x^2 + \frac{1}{\omega} \cdot x\right]_{-w}^{0} \\
&= 0 - \left(\frac{1}{2\omega^2} \cdot (-\omega)^2 + \frac{1}{x} \cdot (-\omega)\right) \\
&= -\left(\frac{1}{2} - 1\right) \\
&= \frac{1}{2} \\
\int_{0}^{\omega} p(x)dx &= \int_{0}^{\omega} -\frac{x}{\omega^2} + \frac{1}{w}dx \\
&= \left(-\frac{1}{2\omega^2} \cdot \omega^2 + \frac{1}{x} \cdot \omega\right) - 0 \\
&= -\left(\frac{1}{2} - 1\right) \\
&= \frac{1}{2} \\
\int_{-\omega}^{\omega} p(x)dx &= \int_{-\omega}^{0} p(x)dx + \int_{0}^{\omega} p(x)dx \\
&= \frac{1}{2} + \frac{1}{2} \\
&= 1
\end{aligned}$$

The value of the density function at 0 is given by:

$$p(X = 0) = \frac{1}{\omega^2} \cdot 0 + \frac{1}{\omega} = \frac{1}{\omega}$$

Therefore the value at zero can be greater than 1 for $\omega > 1$. This is possible because the density is not equal to the probability (which cannot be greater than 1) but has to be multiplied with the width of interval for which the probability should be calculated.
In order to get the probability for coming late at time point $\frac{\omega}{2}$ we have to calculate the probability for arriving after that time point, or in other words the integral from $\frac{\omega}{2}$ to ω (again the remaining area is

zero anyways):

$$\begin{aligned}
p\left(X > \frac{\omega}{2}\right) &= \int_{\frac{\omega}{2}}^{\omega} p(x)dx \\
&= \int_{\frac{\omega}{2}}^{\omega} -\frac{x}{\omega^2} + \frac{1}{w}dx \\
&= \left[-\frac{1}{2\omega^2} \cdot x^2 + \frac{1}{w} \cdot x\right]_{\frac{\omega}{2}}^{\omega} \\
&= -\frac{1}{2\omega^2} \cdot \omega^2 + \frac{1}{\omega} \cdot \omega - \left(-\frac{1}{2\omega^2} \cdot \left(\frac{\omega}{2}\right)^2 + \frac{1}{\omega} \cdot \frac{\omega}{2}\right) \\
&= \frac{1}{2} + \frac{1}{2\omega^2} \cdot \frac{\omega^2}{4} - \frac{1}{2} \\
&= \frac{1}{8} \\
&= 12.5\%
\end{aligned}$$

Therefore the probability to come late to the appointment is 12.5%.
The probability to arrive exactly at time point $\frac{\omega}{2}$ on the other hand is zero, because probabilities of continuous variables can only be calculated for intervals and not for points. Or mathematically:
$\int_{\frac{\omega}{2}}^{\frac{\omega}{2}} p(x)dx = 0$
As we know from the first part of the exercise, the probability to arrive before time point 0 is 50%. Thus we can just calculate the area after 0 which makes up for 45% to get the time point at which we can be sure with 95% that we will arrive before:

$$\begin{aligned}
\int_0^t p(x)dx &\stackrel{!}{=} 0.45 \\
\left[-\frac{1}{2\omega^2} \cdot x^2 + \frac{1}{\omega}\right]_0^t &= 0.45 \\
-\frac{1}{2\omega^2} \cdot t^2 + \frac{1}{\omega} - 0 &= 0.45 \\
\text{substitute: } z &= \frac{t}{\omega} \\
\frac{1}{2} \cdot z^2 + z &= 0.45 \\
z^2 - 2z &= -0.9 \\
z^2 - 2z + 1 &= 0.1 \\
(z-1)^2 &= 0.1 \\
z - 1 &= \sqrt{0.1} \\
z &= \sqrt{0.1} + 1 \\
z &= 1.3162 \vee z = 0.6838 \\
\frac{t}{\omega} &= 1.3162 \vee \frac{t}{\omega} = 0.6838 \\
t &= 1.3162 \cdot \omega \vee t = 0.6838 \cdot \omega
\end{aligned}$$

Since we know that we have to arrive before ω it can only be the second value. Therefore we can be 95% sure that we arrive before $0.6838 \cdot \omega$.

Exercise 2

$$\begin{aligned} Y &= G^{-}1(X) \\ \text{cdf: } P(Y \leq t) &= P\left(G^{-}1(X) \leq t\right) \\ &= P(X \leq G(t)) \\ &= F(G(t)) \\ \text{pdf: } \frac{\partial F(G(t))}{\partial t} &= f(G(t)) \cdot g(t) \end{aligned}$$

Exercise 3

$$\text{pdf: } \frac{\partial F(G(t))}{\partial t} = f(G(t)) \cdot g(t)$$

$f(t)$ is 1 for $0 \leq t \leq 1$, so $f(G(t)) \cdot g(t) = 1 \cdot g(t) = g(t)$.

Exercise 4

See page 39 for details on this exercise.

Exercise 5

We can replace X with p in L_2 *(this only works if the function is linear in X, see Tutorial Sheet 5)* and check if it's a good loss function.
First derivative:

$$\begin{aligned} L_2(X,q) &= -X \cdot \log(q) - (1-X) \cdot \log(1-q) \\ E(L_2(p,q)) &= -p \cdot \log(q) - (1-p) \cdot \log(1-q) \\ \frac{\partial E(L_2(p,q))}{\partial q} &= -p \cdot \frac{1}{q} - (1-p)\frac{-1}{1-q} \\ &= -\frac{p}{q} + \frac{(1-p)}{1-q} \end{aligned}$$

Check for p = q:

$$\begin{aligned} 0 &= -\frac{p}{q} + \frac{(1-p)}{1-q} \\ \Leftrightarrow \frac{p}{q} &= \frac{1-p}{1-q} \\ \Leftrightarrow p \cdot (1-q) &= (1-p) \cdot q \\ \Leftrightarrow p - pq &= q - pq \\ \Leftrightarrow p &= q \end{aligned}$$

Second derivative:

$$\begin{aligned} \frac{\partial E(L_2(p,q))}{\partial q} &= \frac{-p}{q} + \frac{1-p}{1-q} \\ \frac{\partial E(L_2(p,q))}{\partial^2 q} &= \frac{p}{q^2} + \frac{1-p}{(1-q)^2} \end{aligned}$$

$$\begin{aligned} &\text{use: } p = q \\ \Rightarrow \frac{q}{q^2} + \frac{1-q}{(1-q)^2} &= \frac{1}{q} + \frac{1}{1-q} \\ &= \frac{1}{q-q^2} > 0 \end{aligned}$$

Since the second derivative is greater than zero in our definition from 0...1 we found a minimum, so the loss function is good for honest people.
Still usually the first loss function is better, since this second function is unforgiving (a single mistake gives a loss of ∞).

13 Bayesian Inference Examples *2014-05-23*

This section is basically about Exercise 4 on the 4th Tutorial Sheet (see page 33). The idea is that we have a multiple choice test where the answers are not simply true or false but can be any value between 0 and 1 representing your belief in this statement to be true (where 0 corresponds to your belief in this statement being false, 1 that you belief it's true). What does a proper scoring do in this example? What is calibration?

13.1 Honesty

If we use a proper scoring rule the participant can minimize her error if she always states her honest belief in the statement. It is obvious that this does not help in getting any points for statements where you have no clue about its real truth value and you can still get lucky if you gamble and just guess a value for a statement.
But for a huge number of questions it is highly unlikely that you gain anything and we proved in the other exercises that it is optimal to state your true belief.

13.2 Calibration

If you are well-calibrated then 80% of the statements you marked with 80% should be true. Calibration is the bridge between frequentist and Bayesian view on this topic. Calibration can only be measured if you have a huge enough sample space, which is not often the case since you have rather twenty than two hundred questions.
Afterwards it is not possible (in our setup) to tell whether wrong answers are due to lying or bad calibration. There is another problem: It is very hard for a normal person to be well calibrated (even if they try). Psychological studies show that most people systematically overestimate their own belief. Meaning that they would write 1 where their true belief is rather 0.92.
People also overestimate small probabilities. For example, people think that it is rather likely to die in a plane crash than in a car accident, although this is rather unlikely compared to car accidents. One could take such studies into account and try to come up with correction terms for the common failures, but this is rather complicated. This is part of the reason why these proper scoring or multiple choice tests of this form are not widely used. The other reason is that not many people know about this stuff and the effect is not that great to even start with all the trouble.

14 Frequentist Inference Examples *2014-05-26*

14.1 Bayesian Inference for Thumbtacks

We return to the example from the first lecture (see page 6). Let us imagine we have thrown a thumbtack n times. The series of data we get from that may be: 01110101101, a binary string consisting of n entries (0 encodes tail, 1 head). Let $x_1....x_n =: X$ be a random variable. What is the probability of the data given our vague belief about q (for a coin we have the strong intuition of believing that $q = 0.5$, but for the thumbtack we are not sure)? The probability of the whole data (X) is the product of the probability of all single events (x_i):

$$P(X|q) = \prod_{i=1}^{n} \underbrace{q^{x_i}\,(1-q)^{1-x_i}}_{\text{Bernoulli}} = q^h\,(1-q)^t$$

h: #heads t: #tails
But since we do not know q we would also like to find the best q that explains the observed data. In other words: what is the probability of a specific q given the data? This can again be expressed with Bayes' Rule:

$$\overbrace{p(q|X)}^{\text{posterior}} = \frac{P(X|q)\overbrace{p(q)}^{\text{prior}}}{\int_0^1 P(X|q)p(q)dq}$$

The question arising here is what shall we take as the prior? In principle it is just our personal belief about the thumbtack, one experimenter might believe it is 0.7, others believe different values. This *subjectivity* troubles many Non-Bayesians. The good thing is, that it does not really matter which prior you choose, if you have enough data the result will still converge to the real $p(q|X)$!

Note: $\alpha, \beta \in \mathbb{N}_+$: $B(\alpha,\beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$

Since we have no concrete clue and it is not that important anyway, we may choose a distribution by pure convenience for $p(q)$. The distribution is called β *Distribution* :

$$p(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{\int_0^1 q^{\alpha-1}(1-q)^{\beta-1}dq} = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)}$$

We can now neglect the normalization term in the original $p(q|X)$. So we get rid of the integral in the denominator, since it is independent of q.

$$p(q|X) \propto \underbrace{q^h(1-q)^t}_{P(X|q)}\underbrace{q^{\alpha-1}(1-q)^{\beta-1}}_{p(q)} = \underbrace{q^{h+\alpha-1}(1-q)^{t+\beta-1}}_{\text{new } \beta\text{-distribution}}$$

If we now choose $\alpha_n = h + \alpha, \beta_n = t + \beta$, we can express the new β Distribution as:

$$p(q|X) = \frac{q^{\alpha_n-1}(1-q)^{\beta_n-1}}{B(\alpha_n,\beta_n)}$$

If the prior and posterior have the same distribution (like in this case) they are called *conjugate*.

14.2 Map estimate (maximum a posteriori)

In order to find the maximum posteriori term we need to calculate the first derivative of $p(q|X)$. That seems quite hard but we can reduce the problem by ignoring the normalization term $\frac{1}{B(\alpha_n,\beta_n)}$ (since it is independent from q) and taking the logarithm of the numerator, since this does not change the location of any maxima. By this we just have to maximize $\log(q^{\alpha_n-1}(1-q)_n^{\beta}-1)$.

$$\log\left(q^{\alpha_n-1}(1-q)^{\beta_n-1}\right) = (\alpha_n-1)\log\hat{q} + (\beta_n-1)\log(1-\hat{q})$$
$$\frac{\partial\left((\alpha_n-1)\log\hat{q} + (\beta_n-1)\log(1-\hat{q})\right)}{\partial\hat{q}} = \frac{\alpha_n-1}{\hat{q}} - \frac{\beta_n-1}{1-\hat{q}} = 0$$
$$\frac{\alpha_n-1}{\hat{q}} = \frac{\beta_n-1}{1-\hat{q}} \Leftrightarrow \frac{1-\hat{q}}{\hat{q}} = \frac{\beta_n-1}{\alpha_n-1}$$
$$\frac{1}{\hat{q}} = \frac{\beta_n-1}{\alpha_n-1} + 1 = \frac{\beta_n-1}{\alpha_n-1} + \frac{\alpha_n-1}{\alpha_n-1} = \frac{\beta_n+\alpha_n-2}{\alpha_n-1}$$
$$\hat{q} = \frac{\alpha_n-1}{\beta_n+\alpha_n-2} = \frac{\alpha+h-1}{\alpha+\beta+h+t-2}$$

For $\alpha = \beta = 1$ we get what we have already suspected before: $\hat{q} = \frac{h}{h+t}$. α and β are also called pseudo-counters. They represent data points you have not seen but believe to be realistic. This is a way to put your prior belief about the problem in the model - but it is also dangerous. If you have a strong belief in a hypothesis it will need more and more data to prove in the limit that you are wrong.

Note: *The approach of using Bayesian statistics with a prior that does not assume or put in any information is called 'Objective Bayes'. It is somehow the middle ground between the two opposing camps.*

14.3 NHST Null Hypothesis Significance Testing

In the previous section we examined how Bayesian people tackle the problem of finding a good model for a problem. For frequentists it is a bit more complicated. Remember that probabilities (like the probability for heads for a fair coin) are objective/fixed properties of the object. Writing $p(q)$ (as well as $p(q|X) \propto p(q)p(X|q)$) makes no sense for this reason. q is not a random variable but a property and we need to find out its concrete value.

Experiment: Can someone discriminate between coke and pepsi?

We would like to know what $p(q)$ is. But as frequentists we can't do that. So we start with the null hypothesis H_0: Subjects can't discriminate: $q = \frac{1}{2}, n = 25$. Where q denotes the discrimination factor for the subjects and n is the number of trials. We now measure H (# "Heads": correct discriminations).

Note: *This section is wrong and needs to be updated!*

$$P(H = h|q) = \underbrace{\binom{n}{h} q^h (1-q)^{n-h}}_{\text{Binomial distribution}}$$

Now we introduce a criterion: subjects can discriminate if they get >20 correct answers. In that case you reject the null hypothesis.

Note: *Obviously WIP!*

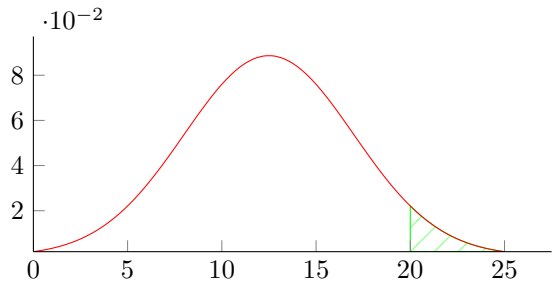
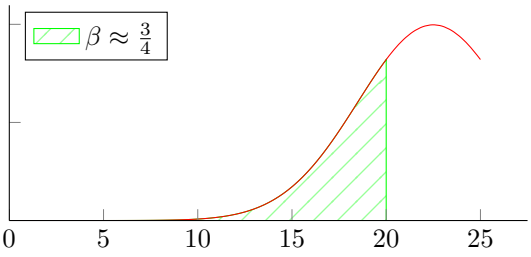


Figure 12: The corresponding Gaussian

$$P\left(H = h | q = \frac{1}{2}, n = 25\right)$$
$$P\left(H > 20 | q = \frac{1}{2}, n = 25\right) \text{ (p-Value)}$$

α is the signal level $\rightarrow$ type I error rate that's acceptable (usually 5%). This is the probability that you say there is an effect even if there is none.
say $q = \frac{4}{5}$



type II error $\beta \approx \frac{3}{4}$
tradeoff between α and β: "easier" for $q \rightarrow \frac{1}{2}$ for n big: the power is $1 - \beta$.

PMPC Tutorial Sheet 5

1. Plot the beta distribution.[1] Vary the parameters, α and β, systematically to get a feeling for how they change the distribution. What do you observe? What will happen if you use the beta distribution to learn about the probability q that a thumbtack will land on its head and you get more and more data?

2. There are two candidates for an election. In a poll of 1000 people 533 were in favor of candidate A and the rest preferred candidate B. Do you think candidate A is going to win? How sure are you? What is your prior distribution for the proportion of people voting for A? Can you express your prior belief as a beta distribution? Make a plot of your posterior probability density function and cumulative density function.

3. In 1994 CNN announced a poll of 500 people aged 18-34. An incredible 46% (230/500) say they believe in unidentified flying objects (UFOs). Assuming a beta distribution with $\alpha = 1$ and $\beta = 9$ prior density for the population proportion who believe in UFOs, find a 95% posterior probability interval for this proportion. Produce a plot with the prior and the posterior probability distribution. Mark the central 95% posterior probability interval.

4. Null hypothesis significance testing. Is it possible to distinguish Coca Cola from Pepsi? On each trial a subject is either given Coca Cola or Pepsi and has to decide which one it is. This is repeated for 25 trials and the subject has to give a response on each trial ("no idea" is not an option, only "Coca Cola" or "Pepsi" are possible responses). The experimenter decides to reject the null-hypothesis that the subject is merely guessing if she gets more then 20 correct responses. Imagine that the subject cannot distinguish the two drinks at all. Imagine also that the experiment was repeated a great many times (every time with 25 trials). How often do you expect it will happen that the subject got more than 20 correct responses despite a total inability to discriminate the drinks? Run a simulation to answer the question.

5. Usually adopted interpretations of $p < .01$ by 70 academic psychologists [3].

Statement	f	%
1. The null hypothesis is absolutely disproved.	1	1.4
2. The probability of the null hypothesis has been found.	32	45.7
3. The experimental hypothesis is absolutely proved.	2	2.9
4. The probability of the experimental hypothesis can be deduced.	30	42.9
5. The probability that the decision taken is wrong is known.	48	68.6
6. A replication has a .99 probability of being significant.	24	34.3
7. The probability of the data given the null hypothesis is known.	8	11.3

Which of these statements are correct and which are wrong? Explain why!

6. Is most research wrong [2]? Psychologists almost universally adopt an α level of 5% to reject the null hypothesis of no effect. Let's assume the power $1 - \beta$ of most studies in psychology is 75%.

[1] The beta distribution and related functions are `betapdf, betacdf, betainv` in matlab and octave. The parameters that we called α and β in class are called `A` and B in the functions. See `help betapdf`.

We don't know what the proportion q of non-null effects is among all the effects that are tested in psychology. An effect was found significant by null hypothesis significance testing (NHST). For what proportion q is the probability that the effect is real less than $\frac{1}{2}$?

7. Some people think that null hypothesis significance testing (NHST) should be banned. If you want to read up on the controversy surrounding NHST I highly recommend the (very polemic) paper by Cohen [1] and the paper by Ioannides [2]. For a more balanced account you should read the paper by Krantz [4].

8. The expectation of a function $f(X)$ of a random variable X over a discrete sample space Ω is defined as

$$E(f(X)) = \sum_{x \epsilon \Omega} p(x) f(x).$$

You can think about this as a weighted average: Which values of $f(X)$ will you see on average. Show that

$$E(a \cdot f(X) + b) = a \cdot E(f(X)) + b.$$

For two random variables X and Y with sample spaces Ω_X and Ω_Y and a function $f(X, Y)$ of both variables, the joint expectation is defined as

$$E(f(X,Y)) = \sum_{x \epsilon \Omega_X} \sum_{y \epsilon \Omega_Y} p(x,y) f(x,y).$$

Show that if X and Y are independent

$$E(X + Y) = E(X) + E(Y).$$

The mean of a random variable is defined as $E(X)$. The variance of a random variable is defined as

$$\text{var}(X) = E\left((X - E(X))^2 \right),$$

the expectation of the squared distance from the mean. Show that

$$\text{var}(X) = E\left(X^2\right) - E\left(X\right)^2.$$

Show that for two independent random variables X and Y

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

A thumbtack has probability p of landing on its head. You toss it n times and get the outcomes $X_1 ... X_n$. What is the mean and the variance of each X_i? What is the mean and the variance of $N = \sum_{i=1}^{n} X_i$?

References

[1] J. Cohen. The earth is round (p<.05). *American Psychologist*, 49(12):997–1003, 1994.

[2] J. P.A. Ioannides. Why most published research findings are false. *PLoS Medicine*, 2(8), 2005.

[3] R. B. Kline. *Beyond Significance Testing.* American Psychological Association, Washington, DC, 2004.

[4] D. H. Krantz. The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, 94(448):1372–1381, 1999.

16 Solution 5: Bayesian and Frequentist Inference I *2014-05-30*

Exercise 1

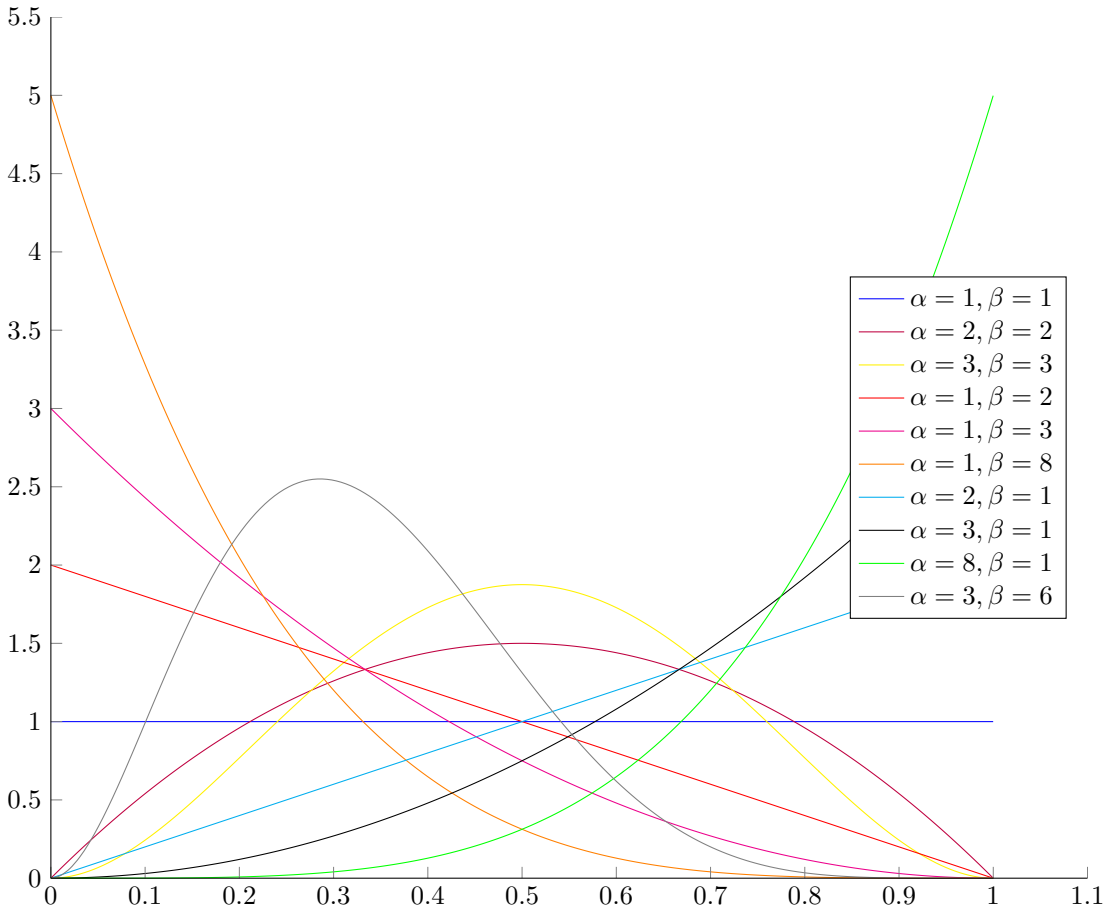


Figure 13: Different β-distributions

We can easily observe that the β-distribution is symmetric for $\alpha = \beta$ and always 1 in case $\alpha = \beta = 1$. If we increase one parameter the distribution's peak wanders to one side, towards 0 for higher βs and towards 1 for higher αs.
If we increase both parameters unequally, the distribution's peak moves towards the higher parameter (similar to the movement mentioned before, see $\alpha = 3, \beta = 6$) and gets for high values much narrower and steeper (not shown in the plot, see figure 16 for an example).

When using the β-distribution to learn about the probability q for the thumbtack example (page 6) we would start with $\alpha = \beta = 1$, i.e. uninformed. After gathering some data we can use it to update our prior belief and adjust the parameters according to our posterior distribution. If we continue doing this iteratively the distribution will get closer to the real q of the thumbtack.

Exercise 2

We have only very limited information and a best guess is that A will win the elections, since there are 53.3% of the people supporting him. We are not really sure about that because of the lack of information. Our prior belief is uninformed, i.e. we have no idea about the outcome (Assuming we don't know the poll). The likelihood is our data sample, i.e. the poll's data. It is a binomial distribution: $P(Poll|A) =$

$\binom{1000}{533}p^{533}(1-p)^{1000-533}$. We use the β-distribution to model our prior belief: it is conjugate to the Binomial distribution.
We try to find to find $P(A|Poll)$, i.e. the probability that A wins the election given the poll results.

$$P(A|Poll) = \frac{P(Poll|A)P(A)}{P(Poll)}$$
$$P(A|Poll) = \frac{\binom{1000}{533}p^{533}(1-p)^{467} \cdot p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$$

using an uninformed prior, i.e. $\alpha = 1, \beta = 1$

$$P(A|Poll) = \frac{\binom{1000}{533}p^{533}(1-p)^{467} \cdot p^{1-1}(1-p)^{1-1}}{B(1,1)}$$
$$P(A|Poll) \propto \frac{p^{533+1-1}(1-p)^{467+1-1}}{B(534,468)}$$

We can plot this posterior belief $P(A|Poll)$ (figure 14) and see how likely it is that A wins. Since the highest density is above 0.5, it's more likely that A wins the election.

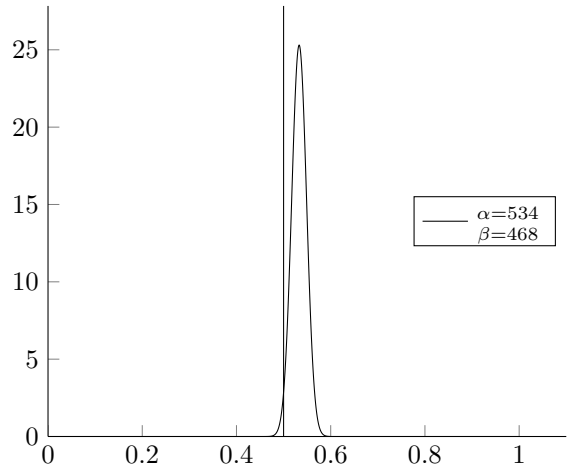


Figure 14: How likely is it, that A wins the election?

Exercise 3

We denote B as the number of people believing in UFOs and D as the data (i.e. the poll results) given. Using the β-Distribution we come up with the following model:

$$P(B|D) = \frac{P(D|B)P(B)}{P(D)}$$
$$\Leftrightarrow P(B|D) = \frac{\binom{500}{230}q^{230}\,(1-q)^{500-230}\,P(B)}{P(D)}$$
$$P(B|D) = \frac{\overbrace{\binom{500}{230}q^{230}\,(1-q)^{270}}^{\text{Binomial}}\overbrace{q^{\alpha-1}\,(1-q)^{\beta-1}}^{\beta\text{-Distribution}}}{B(\alpha,\beta)}$$
$$P(B|D) \propto q^{230+1-1}\,(1-q)^{270+9-1}$$
$$P(B|D) \Rightarrow \frac{q^{231-1}\,(1-q)^{279-1}}{B(231,279)}$$

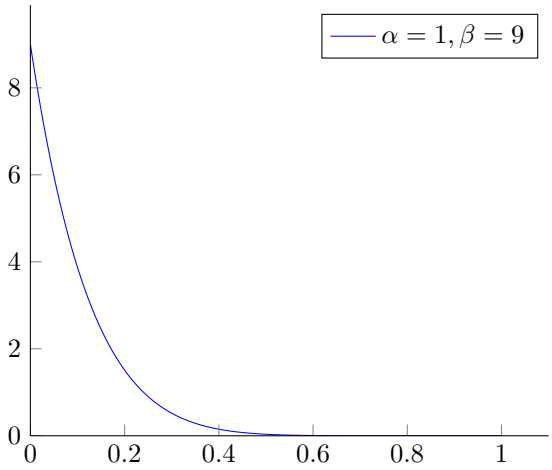


Figure 15: Prior β-Distribution

The prior distribution (figure 15) is far to the left, as we have a relatively huge β compared to α. With the new $\alpha = 231$ and $\beta = 279$ we can find the 95% interval by calculating the CDF of the posterior β-distribution and searching for the x-values for $y_{min} = 0.025$ and $y_{max} = 0.975$. Using these x-values for the corresponding PDF yields the 95% interval (figure 16). The left figure shows the CDF and the intersections for a visual explanation: By using the inverse of the CDF we can calculate the x-values much easier.

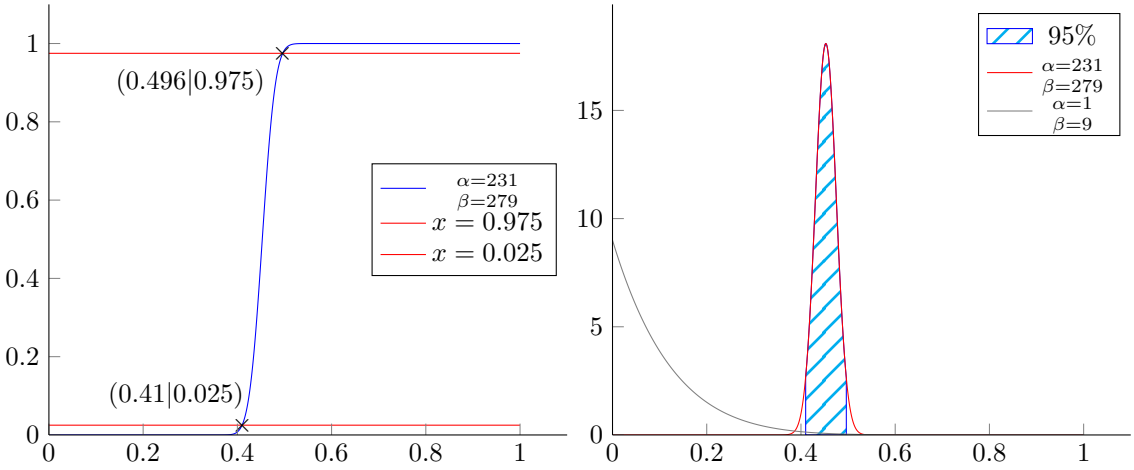


Figure 16: Left: Posterior CDF and 95% intersections. Right: Posterior PDF and 95% interval, prior PDF for comparison.

`MATLAB` code to try out exercise 3:

```
x = linspace(0, 1, 1000);              % samples

alphaPrior = 1; alphaPosterior = 231; % prior/posterior alpha/beta
betaPrior  = 9; betaPosterior  = 279; % from sheet and calculated

yPriorPDF     = betapdf(x, alphaPrior,     betaPrior);
yPosteriorPDF = betapdf(x, alphaPosterior, betaPosterior);
yPosteriorCDF = betacdf(x, alphaPosterior, betaPosterior);

yLow  = 0.025;                                          % 95 % threshold
yHigh = 0.975;
```

```
xLow  = betainv(yLow,  alphaPosterior, betaPosterior); % corresponding x
xHigh = betainv(yHigh, alphaPosterior, betaPosterior);

subplot(2, 2, 1);
    plot(x, yPriorPDF);               % plot prior
    title(['Prior PDF, \alpha = ' num2str(alphaPrior)...
           ', \beta = ' num2str(betaPrior)]);
subplot(2, 2, 3);
    hold on;
    plot(x, yPosteriorCDF);           % plot posterior cdf
    plot(xlim, [yHigh yHigh], 'r');   % plot 95% lines and intersections
    plot(xlim, [yLow  yLow],  'r');
    plot(xHigh, yHigh, 'og');
    plot(xLow,  yLow,  'og');
    title(['Posterior CDF with 95 % intersections, \alpha = '...
           num2str(alphaPosterior) ', \beta = ' num2str(betaPosterior)]);
    hold off;
subplot(2, 2, [2 4]);
    hold on;
    for i = linspace(xLow, xHigh, 100)  % plot area
        plot([i i], [0 betapdf(i, alphaPosterior, betaPosterior)], 'r');
    end

    plot(x, yPosteriorPDF);           % plot pdf
    title(['Posterior PDF with 95 % interval, \alpha = '...
           num2str(alphaPosterior) ', \beta = ' num2str(betaPosterior)]);
    hold off;
```

Exercise 4

To make a guess we can run a simulation with n subjects who have to do the task 25 times. We measure how often they answer correctly and in the end we check how many people g out of n had more than 20 correct answers.
One possibly result was:
`0.0473% (473) of 1000000 subjects scored more than 20 out of 25.`

`MATLAB` code to try out exercise 4:

```
% values for the experiment
numberTrials   = 25;
threshold      = 20;
numberSubjects = 1e6;

% generate trials for each subject, count the successes per subject
trials = sum(rand(numberSubjects, numberTrials) > 0.5, 2);

% plot a histogram
hist(trials);

% count how many are very good and what their proportion is
numHigh    = sum(trials > threshold);
proportion = numHigh / numberSubjects;

display([num2str(proportion*100) '% (' num2str(numHigh) ') of '...
         num2str(numberSubjects) ' subjects scored more than '...
         num2str(threshold) ' out of ' num2str(numberTrials) '.'])
```

Exercise 5

1. Since $p < 0.01$ it's not absolutely disproved, because we still have to take β into account.
2. You don't find the probability of the H_0 ($p(q = \frac{1}{2}|H = h, n = 25)$), but instead you find $p(H > 20|q = \frac{1}{2}, n = 25)$ - the p-value. But the p-value is not sufficient to know the probability of H_0.

3. As in 1: it's not certain.
4. Finding $p(q > \frac{1}{2}|H = h, n = 25)$ is only possible with Bayes' rule, so it's impossible from a frequentist point of view.
5. To know the probability that the decision taken was wrong you need to know the type II (β-) error. This is not possible if you don't know q, which you don't in an NHST.
6. It's not possible to know what happens if you repeat an experiment. The results might be similar or totally different, it is dependent on the power of the test, which we only know if we know the effect size, too.
7. This is the only true fact in this list that we can derive from an NHST.

Exercise 6

We want to find out the relative proportion of true hypotheses to hypotheses which were wrongly found out to be true, i.e. $\frac{\text{true hypotheses}}{\text{hypotheses found true}}$.
By assuming that the probability for a hypotheses to be true is q, we can come up with a simple probability tree (figure 17).

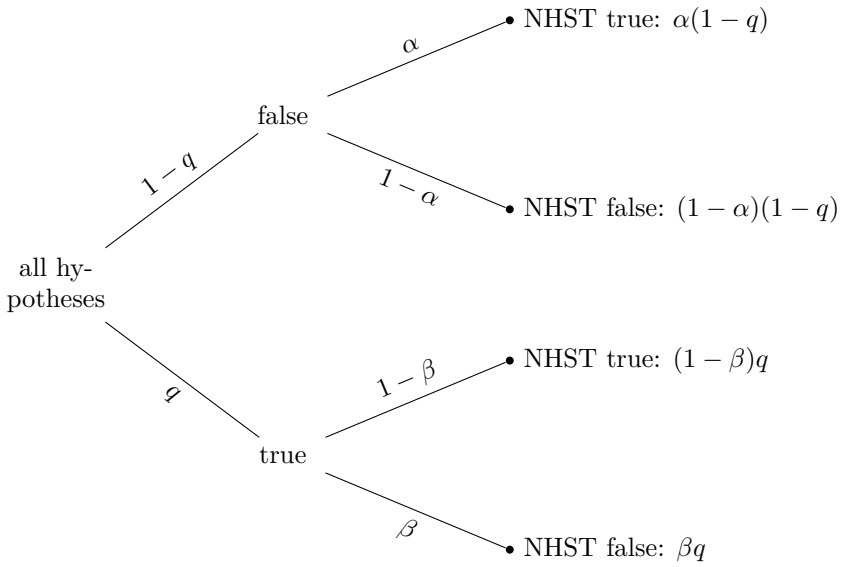


Figure 17: Probability tree

By using the formulas for hypotheses which were found to be true we can fill the formula mentioned above.

$$\frac{\overbrace{(1-\beta)q}^{\text{true hypotheses}}}{\underbrace{(1-\beta)q + \alpha(1-q)}_{\text{hypotheses found true}}}$$

We can solve this for q by plugging in the given values $\alpha = 5\%, 1 - \beta = 75\%$ and setting it to be smaller

than $\frac{1}{2}$.

$$\begin{aligned}
\frac{1}{2} &> \frac{(1-\beta)q}{(1-\beta)q + \alpha(1-q)} \\
\frac{1}{2} &> \frac{0.75q}{0.75q + 0.05(1-q)} \\
\frac{1}{2} &> \frac{0.75q}{0.75q + 0.05 - 0.05q} \\
\frac{1}{2} &> \frac{0.75q}{0.7q + 0.05} \\
0.7q + 0.05 &> 1.5q \\
0.05 &> 0.8q \\
0.0625 &> q
\end{aligned}$$

So if q is smaller than $\frac{1}{16}$, the probability that the effect is real is less than 50%.

Exercise 8

This exercise was posed again later, since it was not crucial for the midterm exam. For the solution please refer to page 22.

17 Midterm Exam Questions

There are 8 questions and each question is worth 4 points. You only have to answer 6 out of the 8 questions. Hence, the maximum score is 24 points. If you answer more than 6 questions the answers with the lowest scores will be discarded. You have from 8:00 to 10:00 to work on your responses. Please respond in full sentences.

Question 1

A. took an HIV test. The test turned out to be positive. About 0.01% of men like him are infected with HIV. If someone has the virus there is a 99.9% chance that the test result will be positive. If someone is not infected there is a 99.99% chance that the result will be negative. What is the probability that A. is infected with HIV?

Question 2

You have the choice of flying with a 2-engine or a 4-engine plane. Both are old models with the same kind of engines that sometimes fail with probability p. A plane can still fly if at least half of the engines are working. Which of the two is safer?

Question 3

To encourage Elmer's promising tennis career, his father offers him a prize if he wins (at least) two tennis sets in a row in a three-set series to be played with his father and the club champion alternately: father-champion-father or champion-father-champion, according to Elmer's choice. The champion is a better player than Elmer's father. Which series should Elmer choose?

Question 4

Bookie A has set the odds for Real Madrid winning the Champions League to $1:3$ (i.e. if you bet 1 € on Real Madrid you win 3 € if Real Madrid wins) and the odds for Real Madrid not winning to $4:1$. Bookie B has set the odds for Madrid to $1:5$ and the odds against Madrid to $6:1$. How can you take advantage of this situation?

Question 5

In a simple word learning experiment subjects are shown bug-like stimuli. They can have wings or not, they can have antennas or not, they can have dots or not and they can have a sting or not (see figure 18 for one bug that has none of these features and one bug that has all of these features). The experimenter picks one of the features uniformly at random, for example "has wings". The subject is told that they are learning a foreign language and that their task is to figure out what the word "dax" means. They are also told that the word "dax" refers to one of the four features (wings, antennas, dots, or sting) and that their task is to find out which one it is, for example "has wings". To this end, on every trial one of the 16 possible stimuli is picked by the experimenter uniformly at random and shown to the subject. The subject has to say whether the word "dax" applies to this stimulus or not. On each trial the experimenter gives feedback as to whether the word "dax" was used correctly or not, for example whether the stimulus has wings or not. Let's assume subjects have the following strategy: First, they pick one of the four features. They assume that this is the meaning of the word "dax" until they make a mistake. If they

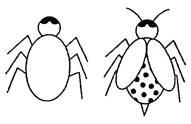


Figure 18: Bugs

did not pick the right feature, how many trials will it take until they make a mistake (more precisely: What is the distribution over the number of trials that it takes)? If they made a mistake they will try the next feature. They don't try another feature unless they make a mistake. How many mistakes will they make until they found the correct interpretation of the word "dax" (i.e. what is the distribution over the number of mistakes)?

Question 6

X is a continuous random variable with uniform probability density function

$$p(X = x) = \begin{cases} e^{-x} & \text{for } 0 \leq x \\ 0 & \text{otherwise} \end{cases}$$

What is the cumulative distribution function for X? Y is a random variable that is defined as $Y = \sqrt{X}$. What is the cumulative distribution function of Y? What is the probability density function of Y?

Question 7

Let X be the truth-value of a statement (either 0 or 1). You want to elicit the beliefs of several experts about this statement. A friend of yours suggests to use the following loss function

$$L(X, q) = -Xq^2 - (1 - X)(1 - q)^2$$

to score the responses (note that negative loss is gain). Is this loss function a proper scoring rule? Explain.

Question 8

The following are statements about p-values in published or online work. Explain why each is wrong and offer a correct version:

(a) "If $p < 0.05$, the researcher knows that there is less than 1 chance in 20 that the observed difference between the experimental and control groups occurred by chance."

(b) "A hypothesis is rejected if the probability that it is a true statement is very low - lower than some predetermined probability ... the level of significance."

(c) "Significance levels show you how likely a result is due to chance. The most common level, used to mean something is good enough to be believed, is .95. This means that the finding has a 95% chance of being true."

(d) "By convention, a p-value higher than 0.05 usually indicates that the results of the study, however good or bad, were probably due only to chance."

18 Midterm Solutions *2014-06-06*

Question 1

- T = positive test
- $\neg T$ = negative test
- HIV = infected with HIV
- $\neg HIV$ = not infected with HIV

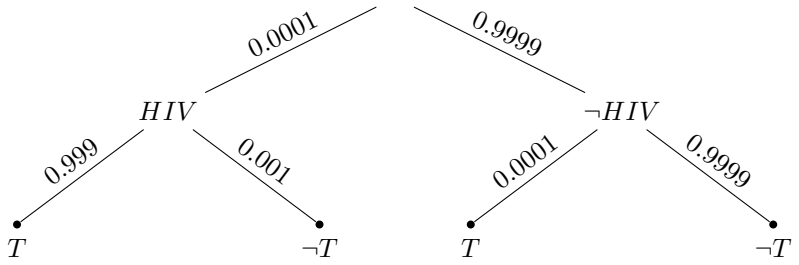


Figure 19: Probability tree

We search for

$$P(HIV|T) = \frac{P(T|HIV)P(HIV)}{P(T)} = \frac{0.999 \cdot 0.0001}{0.999 \cdot 0.0001 + 0.0001 \cdot 0.9999} = \frac{1110}{2221} < \frac{1}{2}$$

The probability that A is infected with HIV is just below 0.5.

Question 2

We can draw tables (See table 2) to visualize in which cases the airplane will crash.

#	E1	E2
1	O	O
2	X	O
3	O	X
4	X	X

(a) two engine plane

#	E1	E2	E3	E4
1	O	O	O	O
2	X	O	O	O
3	O	X	O	O
			⋮	
12	O	X	X	X
13	X	O	X	X
14	X	X	O	X
15	X	X	X	O
16	X	X	X	X

(b) four engine plane

Table 2: Tables for the two airplanes. Note that an X indicates an engine failure while the marked rows are those in which more than 50% of the engines fail, i.e. the planes crash.

Probability for a crash with

- two engines: p^2
- four engines: $p^4 + 4p^3\,(1-p)$

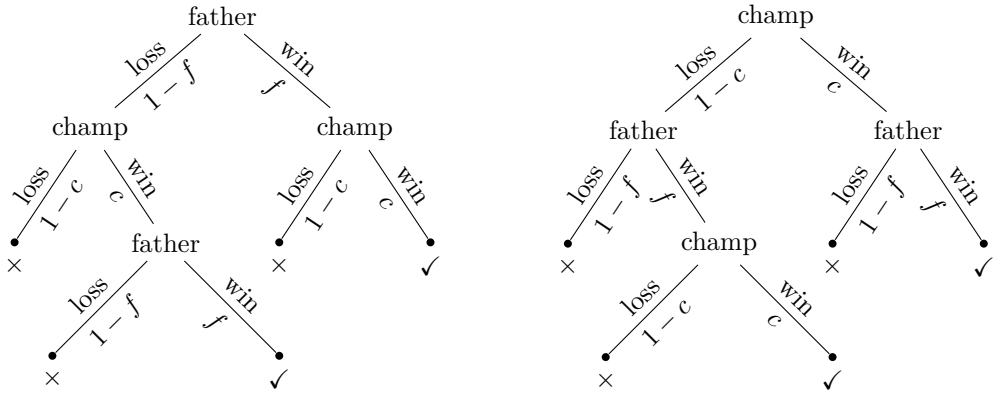
To check which case is better we plug these probabilities into an inequality and simplify it until we are sure it either holds or not. *Note that we divide by p^2, which is only possible under the assumption that $p > 0$!*

$$\begin{aligned} p^2 &\overset{?}{\geq} p^4 + 4p^3\,(1-p) \\ &= p^4 + 4p^3 - 4p^4 \\ &= -3p^4 + 4p^3 \\ 1 &\geq -3p^2 + 4p \\ 1 - 4p + 3p^2 &\overset{?}{\geq} 0 \\ (1-p)(1-3p) &\overset{?}{\geq} 0 \\ p &\geq \frac{1}{3} \end{aligned}$$

If $p \geq \frac{1}{3}$, we should fly with the four engine airplane (it's more likely that the two engine airplane will crash in that case), otherwise with the two engine airplane. The choice is dependent on p.

Question 3

The important clue for this question is that Elmer has two win *two consecutive* sets.
We can draw two simple probability trees(figure 20) to calculate Elmer's chances. We use f and c (probability to win against father and champ, respectively) where $f > c$ denote the chances to win each set.



(a) father-champ-father

(b) champ-father-champ

Figure 20: Two possible choices

We can see that in the first setup, father-champ-father, the probability to win two consecutive matches (those marked with ✓) is $fc+(1-f)cf = cf(2-f)$, and in the second setup accordingly: $cf+(1-c)fc = fc(2-c)$.
Since we know $f > c$ we can use this information to evaluate the odds:

$$\frac{P(success|f-c-f)}{P(success|c-f-c)} = \frac{cf(2-f)}{cf(2-c)} = \frac{2-f}{2-c} < 1$$

Since $\frac{2-f}{2-c} < 1$ the odds are in favor of the second setup: **champ-father-champ**.
This also makes sense intuitively because playing against the champ first gives a higher probability for a second chance in case the first match is lost. This of course only is valid since Elmer has to win two sets in a row.

Question 4

We can take advantage if we are able to set up a dutch book.

- Bookie A: Real wins 1:3
- Bookie A: Real loses 4:1
- Bookie B: Real wins 1:5
- Bookie B: Real loses 6:1

As we can see Bookie B pays more for a win than we have to pay for a loss at Bookie A. We should be able to exploit this.

$$\frac{5}{1+5} + \frac{1}{1+4} = \frac{5}{6} + \frac{1}{5} = \frac{25}{30} + \frac{6}{30} = \frac{31}{30} > 1 \Rightarrow \text{A dutch book is possible}$$

The expected value for win and loss when betting the scenario above are:

$$E(win) = 5x - y$$
$$E(loss) = \frac{1}{4}y - x$$

We can calculates for which interval we can make money:

$$E(win) = 5x - y > 0 \Leftrightarrow 5x > y \Leftrightarrow y < 5x$$
$$E(loss) = \frac{1}{4}y - x > 0 \Leftrightarrow -x > -\frac{1}{4}y \Leftrightarrow 4x < y$$

So we can take advantage by betting x on win at B and y on loss at A as long as x and y fulfill this inequality: $4x < y < 5x$.

Question 5

For each trial the picture either shows the stimulus or not. So the probability to get the stimulus on the n^{th} trial is $\frac{1}{2}^n$. That means the distribution over the number of trials n is simply: $P(n) = \frac{1}{2^n}$.
The probabilities to make x mistakes can be simply derived: there are four features. So the subject will pick the right feature directly with a probability of $\frac{1}{4}$ (this means 0 mistakes). If the subject was wrong (1 mistake), which happens in $\frac{3}{4}$ of the cases, it has a chance of $\frac{1}{3}$ to get it correct the second time, and so on.

x=0: $\frac{1}{4} = \frac{1}{4}$

x=1: $\frac{3}{4} \cdot \frac{1}{3} = \frac{1}{4}$

x=2: $\frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{4}$

x=3: $\frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} \cdot \frac{1}{1} = \frac{1}{4}$

As one can easily see: the probability for each number of mistakes is the same. So the distribution is uniform, i.e. $P(X) = \frac{1}{4}$.

Question 6

$$P(X \leq t) = \left[-e^{-x}\right]_0^t = 1 - e^{-t}$$
$$P(Y \leq t) = P(X \leq t^2) = 1 - e^{-(t^2)}$$
$$\text{pdf: } 2ye^{-(y^2)}$$

Question 7

To check if a proper scoring rule is given, we can check if it is minimal (since our function is a loss function) for $p = q$. So we take the 1^{st} derivative and set it to zero to see if $p = q$ holds. We have to check the second derivative to see whether we found a mini- or maximum.

$$\begin{aligned} L(X,q) &= -Xq^2 - (1-X)(1-q)^2 \\ \Leftrightarrow L(X,q) &= -Xq^2 - (1-X)(1-2q+q^2) \\ \Leftrightarrow L(X,q) &= -Xq^2 - 1 + 2q - q^2 + X - 2Xq + Xq^2 \\ \Leftrightarrow L(X,q) &= -q^2 + 2q - 2Xq + X - 1 \end{aligned}$$

$$\begin{aligned} \text{Set } X &= p \\ L(p,q) &= -q^2 + 2q - 2pq + p - 1 \\ \frac{\partial L(p,q)}{\partial q} &= -2q + 2 - 2p \end{aligned}$$

$$\begin{aligned} \text{Set } \frac{\partial L(p,q)}{\partial q} &= 0 \\ -2q + 2 - 2p &= 0 \\ \Leftrightarrow 2 - 2p &= 2q \\ \Leftrightarrow 1 - p &= q \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 L(p,q)}{\partial^2 q} &= -q \\ \frac{\partial^2 L(p,q)}{\partial^2 q} &< 0 \end{aligned}$$

The 1^{st} derivative has an extreme value at $1 - p = q$, however since this is a maximum (2^{nd} derivative $<$ 0) the scoring rule is not proper.

Question 8

Sorry, we don't have the solution for this. If you have it, feel free to contact us.

19 Signal Detection Theory I *2014-06-13*

19.1 Detection tasks

Detection tasks are simple yes/no response tasks. Yes and no corresponds most of the time to the existence or absence of a signal. The difficulty of such a task stems from the fact that there is always noise in the system, so we can't be sure whether or not the given answer is correct.

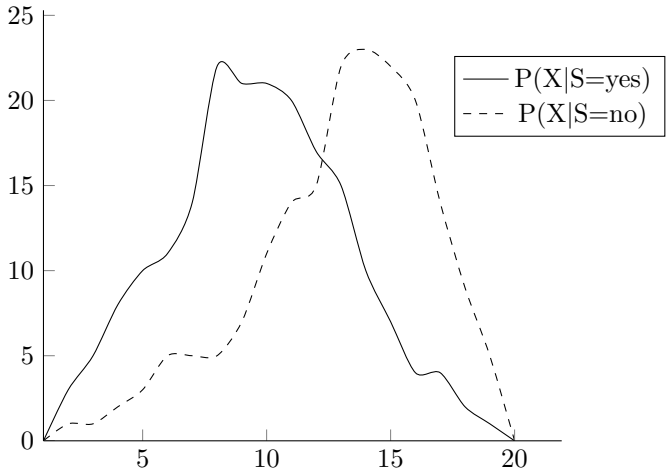
19.1.1 Examples

- Binary signal transmission over noise channel (cable, radio)
- Information retrieval
- airport security scans (is this a weapon or just a hair dryer)

The results of one single trial in such an experiment may take one of the values from the following table:

Response \ Signal	S = yes	S = no
R = yes	Hit	False Alarm
R = no	Miss	Correct Rejection

The probabilities to get a certain response given the existence of a signal may look like the following pdf's:



19.1.2 Response strategy

We are now looking for a set of values for which our response will be *yes*. This is called the response strategy. It somehow determines from what signal strength onward you would report the signal was there. Formally: `if` $x \in A$ `then YES else NO`. Our response strategy should not only depend on the probabilities, but also on the cost of being wrong. The loss function $L(S, R)$ depends now on the costs for the different cases.

Response \ Signal	S = yes	S = no
R = yes	C_H	C_{FA}
R = no	C_M	C_{CR}

The probabilities for Hit and false Alarm are the integrals over the pdf's.

$$P(H) = \int_A p(X \mid S = yes)dx = P(R = y \mid S = y)$$
$$P(FA) = \int_A p(X \mid S = no)dx = P(R = y \mid S = n)$$

The probabilities for miss and correct response can be directly computed from these.

$$P(M) = 1 - P(H) = P(R = n \mid S = y)$$
$$P(CR) = 1 - P(FA) = P(R = n \mid S = n)$$

19.1.3 Minimize expected loss

As with the proper scoring rules we want to give responses in order to minimize our expected loss for certain costs.

$$E(L(S = y, R)) = P(H) \cdot C_H + \underbrace{P(M)}_{1-P(H)} \cdot C_M$$
$$E(L(S = n, R)) = P(FA) \cdot C_{FA} + \underbrace{P(CR)}_{1-P(FA)} \cdot C_{CR}$$

In the following we use this shorthand notation:

$$\pi_y = P(S = y)$$
$$\pi_n = P(S = n)$$

$$\begin{aligned} E(L(S, R)) &= \pi_y E(L(S = y, R)) + \pi_n E(L(S = n, R)) \\ &= \pi_y C_H P(H) + \pi_y C_M - \pi_y C_M P(H) + \pi_n C_{CR} - \pi_n C_{CR} P(FA) + \pi_n C_{FA} P(FA) \\ &= \pi_y P(H)(C_H - C_M) + \pi_n P(FA)(C_{FA} - C_{CR}) + \underbrace{(\pi_y C_M + \pi_n C_{CR})}_{\text{independent of } A} \end{aligned}$$

Now we minimize only the parts dependent on A.

$$\begin{aligned} &\pi_y P(H)(C_H - C_M) + \pi_n P(FA)(C_{FA} - C_{CR}) \\ &= \pi_y \left(\int_A p(X \mid S = yes)dx \right) (C_H - C_M) + \pi_n \left(\int_A p(X \mid S = no)dx \right) (C_{FA} - C_{CR}) \\ &= \int_A [\pi_y p(X \mid S = yes)(C_H - C_M) + \pi_n p(X \mid S = no)(C_{FA} - C_{CR})dx] \end{aligned}$$

Choose A such that we only integrate over the negative part.

$$\begin{aligned} \pi_y p(X \mid S = yes)(C_H - C_M) + \pi_n p(X \mid S = no)(C_{FA} - C_{CR}) < 0 \\ \pi_y p(X \mid S = yes)(C_H - C_M) \leq -\pi_n p(X \mid S = no)(C_{FA} - C_{CR}) \\ \pi_y p(X \mid S = yes)(C_H - C_M) \leq \pi_n p(X \mid S = no)(C_{CR} - C_{FA}) \end{aligned}$$

$$\underbrace{\frac{\pi_y p(X \mid S = yes)}{\pi_n p(X \mid S = no)}}_{\text{posterior odds}} \geq \underbrace{\frac{C_{CR} - C_{FA}}{C_H - C_M}}_{\text{costs threshold}}$$

Interpretation: We choose A so that the posterior odds are greater then the costs.
Other way of writing:

$$\underbrace{\frac{p(X \mid S = yes)}{p(X \mid S = no)}}_{\text{likelihood ratio}} \geq \underbrace{\frac{\pi_n(C_{CR} - C_{FA})}{\pi_y(C_H - C_M)}}_{\beta}$$

19.1.4 Use Gaussians for modelling

We can model the probability of Hits and False Alarms with Gaussians respectively:

$$P\left(X|s = yes\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\frac{(x-\mu_y)^2}{\sigma^2}}$$
$$P\left(X|s = no\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2}}$$

This is not easy to calculate, so we simplify by applying the log thereby obtaining the log-likelihood ratio and compare that to our previous calculated β:

$$-\frac{1}{2\sigma^2} \cdot \left[(x-\mu_y)^2 - (x-\mu_n)^2\right] \geq log\left(\beta\right)$$
$$x^2 - 2x\mu_y + \mu_y^2 - x^2 + 2x\mu_n - \mu_n^2 \leq -2\sigma^2 \cdot log\left(\beta\right)$$
$$2x\left(\mu_n - \mu_y\right) + \mu_y^2 - \mu_n^2 \leq -2\sigma^2 \cdot log\left(\beta\right)$$

By convention the mean of the noise distribution μ_n is smaller than μ_y, such that by solving for x and dividing by $(\mu_n - \mu_y)$ the inequality turns and we get:

$$x \geq \underbrace{\frac{-2\sigma^2 \cdot log\left(\beta\right) + \mu_n^2 - \mu_y^2}{2\left(\mu_n - \mu_y\right)}}_{\Theta}$$

That we can interpret such that we answer with yes in the case the x we perceive is bigger than the threshold (criterion) Θ. In the case of $\beta = 1$ and equal prior probabilities that means:

$$x \geq \frac{(\mu_n - \mu_y) \cdot (\mu_n + \mu_y)}{2\left(\mu_n - \mu_y\right)} = \frac{(\mu_n + \mu_y)}{2}$$

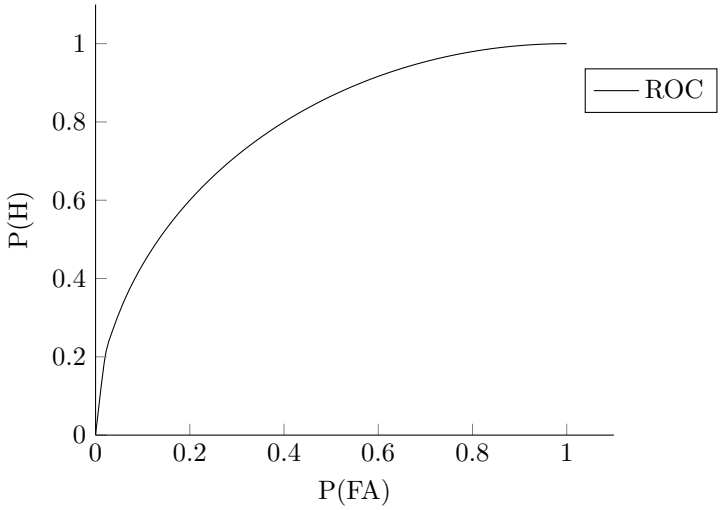
i.e. the best threshold is just in the middle of the two Gaussians, given the condition that both Gaussians have the same variance.

19.2 Signal to Noise Ratio

Now we want to find the limits of perception - how few light can we detect, or in other words: What is a persons signal to noise ratio for light detection. For the case where we again assume two Gaussian distributions with equal deviation, the signal to noise ratio is defined as:

$$SNR = \frac{\mu_y - \mu_n}{\sigma}$$

As we care for the actual perception (the distance of the Gaussians) of a subject and not his decision-criterion when to say yes, we have to come up with a measure independent of the subjects threshold. Receiver operator characteristics (ROC) allow for that by systematically varying the threshold, such that all possible criteria are covered (from always saying yes, to always saying no). Then we get a curve describing the perception of a subject independently of his criteria.



Varying the threshold may be achieved by the experimenter, by manipulating the costs and pay-offs for False Alarms or Hits. Note that the above depicted graph is a theoretical model. In a real experiment we would get single data points, scattered around such a curve. The further left we get on the $P(FA)$ axis, the further right the subject placed her criterion.

20 Signal Detection Theory II *2014-06-16*

20.1 Objective Sensitivity

In the previous lecture we saw how ROC curves helped us to measure the sensitivity of a subject in a decision task. We could compare the curves for several subjects and find out which one has the 'better' sensitivity.

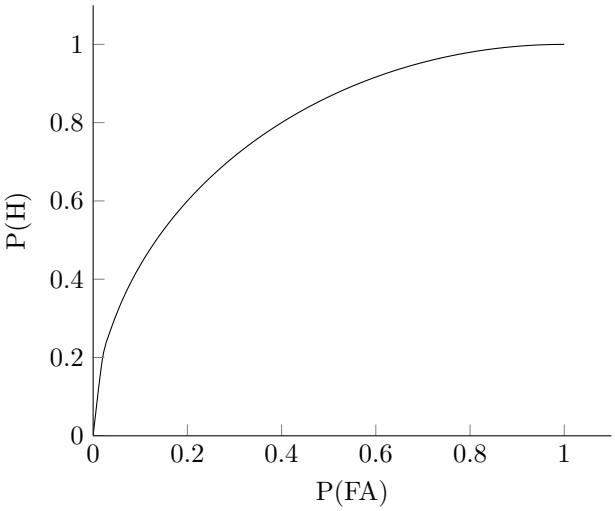


Figure 21: ROC-Curve

But it would be better to have a single value to compare the sensitivity. For the case where signal and noise are Gaussians with same variance this value is the SNR (Signal to Noise Ratio). We can calculate that!

First we subtract the mean of the No-responses to make it zero-centered.

$$\begin{aligned} p(FA) = \int_{\theta}^{\infty} \varphi\left(x, \mu_n, \sigma\right) dx &= 1 - \Phi\left(\theta, x_n, \sigma\right) \\ &= 1 - \Phi\left(\theta - \mu_n, 0, \sigma\right) \\ p(H) = \int_{0}^{\infty} \varphi\left(x, \mu_y, \sigma\right) dx &= 1 - \Phi\left(\theta, x_y, \sigma\right) \\ &= 1 - \Phi\left(\theta - \mu_n, \mu_y - \mu_n, \sigma\right) \end{aligned}$$

We may also adapt on the deviation by dividing through σ, thus arriving on a Gaussian with variance $\sigma^2 = 1$ and mean $\mu = 0$ in statistics this is called standardising or normalising.

$$\begin{aligned} p(FA) = 1 - \Phi\left(\frac{\theta - \mu_n}{\sigma}, 0, 1\right) &= 1 - \Phi\left(\theta'\right) \\ p(H) = 1 - \Phi\left(\frac{\theta - \mu_n}{\sigma}, \frac{\mu_y - \mu_n}{\sigma}, 1\right) &= 1 - \Phi\left(\theta' - d', 0, 1\right) \\ &= 1 - \Phi\left(\theta' - d'\right) \\ d' &= \frac{\mu_y - \mu_n}{\sigma} \end{aligned}$$

We can rearrange the formula for $P(FA)$ to:

$$\begin{aligned}\Phi(\theta') &= 1 - P(FA)\\ \theta' &= \Phi^{-1}(1 - P(FA))\\ &= -\Phi^{-1}(P(FA))\end{aligned}$$

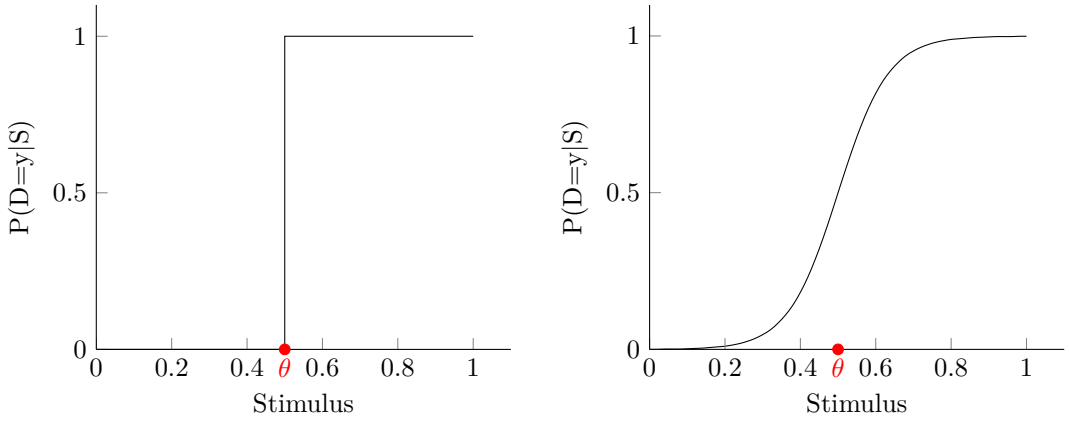
The last step is possible because the Gaussian is symmetric. Now we try to find a formula for d' as well.

$$\begin{aligned}\Phi(\theta' - d') &= 1 - P(H)\\ \theta' - d' &= \Phi^{-1}(1 - P(H))\\ d' &= \theta' + \Phi^{-1}(P(H))\\ &= \Phi^{-1}(P(H)) - \Phi^{-1}(P(FA))\end{aligned}$$

By this we disentangled the sensitivity and the response bias of the subject. θ is rather a bias than a threshold.

20.2 Is there a sensory threshold?

A long time ago, people thought that our sensors work in a 0-1 like manner. There is an internal threshold and either the signal is strong enough to surpass this threshold or not. Detection experiments were conducted to find that threshold of consciousness.



(a) Model of a Sensory Threshold

(b) The Real Data

Figure 22: Prediction of the Threshold-Theory

As seen in the plots, the real data that was measured does not really fit the model. Two reasons could explain this difference. First of all there could be noise in the threshold (depending on some hidden neuron mechanisms). Another explanation could be noise in the stimulus.

The function $P(D = y|S = s) = F_\theta(s), F_\theta(\theta) = \frac{1}{2}$ is called the psychometric function. This is all fine, as long as the subject is honest and not lying about her sensation (R is the response and D is whether the subject detected the stimulus):

$$P(R = y|D = y) = P(R = n|D = n) = 1$$

To measure the subjects honesty we introduce catch-trials! For 50% of the trials we have a stimulus and for the other half we do not. Our model for the subject may be:

$$\begin{aligned}P(R = y|D = y) &= 1\\ P(R = y|D = n) &= q\end{aligned}$$

So the subject only lies when she detected nothing with a certain probability. The probability for hit(H) and false alarm (FA) is therefore:

$$\begin{aligned} P(H) &= P(D = y|S = s) \cdot 1 + (1 - P(D = y|S = s)) \cdot q \\ &= q + (1 - q)P(D = y|S = s) \\ &= q + (1 - q)F_\theta(s) \\ &= P(FA) + (1 - P(FA))F_\theta(s) \\ P(FA) &= P(D = y|s = 0) \cdot 1 + (1 - P(D = y|s = 0)) \cdot q \\ &= q + (1 - q)P(D = y|s = 0) \\ &= q \end{aligned}$$

We assume that we have a high threshold so that $P(D = y|s = 0) \approx 0$. If we solve the first equation for $F_\theta(s)$, we get:

$$F_\theta(s) = \frac{P(H) - P(FA)}{1 - P(FA)}$$

Since we know from the psychometric that for θ this equation should be $\frac{1}{2}$ we can find θ!

20.3 Why High Threshold Theory is wrong!

1. ROC-Curves

If HT-Theory would be right, the ROC-Curves would be straight lines and not curves. But we get curves from the real data.

2. Relation between Y-N and 2-AFC

In a Y-N experiment I may state for each frame if there was a stimulus (Y) or not (N). In 2-AFC the subject sees 2 frames and has to decide in which frame the stimulus was. If we did not see the stimulus we have a 50% chance to get the answer right. The probability of a correct answer should be:

$$\begin{aligned} P(H) &= F_\theta(s) + (1 - F_\theta(s)) \cdot \frac{1}{2} \\ &= \frac{1}{2} + \frac{1}{2}F_\theta(s) \\ F_\theta(s) &= 2P(H) - 1 \end{aligned}$$

But these formulas do not match up with the real data.

3. 2^{nd}Choice in 4-AFC Task

In this case the difference becomes even more obvious. In the experiment you have 4 screens where the stimulus could appear on. If you got it wrong on the first try you may choose again. In HT-Theory one can expect that the chance to get it right in the second round should be $\frac{1}{3}$, because I saw nothing on those screens. But in reality the data shows that people are way better than $\frac{1}{3}$!

4. Rating Data

This experiment has been conducted with Y-N tasks with 50% catch trials.The subjects should always rate their answers on a scale from 1 (unsure) to 5 (super sure). Results show that people seem to be able to rate how sure they are about their perception. And this rating fits the data very well. HT-Theory can not account for this, since there are only all-or-none responses.

PMPC Tutorial Sheet 6

1. Go back to exercise 1 on the first tutorial sheet. Read again about what you planned to do to get an A in this class. Did you follow your plan so far? If not, why did you deviate from it? What would you do differently now after you have taken the mid-term exam?

2. We haven't discussed exercise 8 on the last tutorial sheet yet. So if you haven't worked on it, you can still do it.

3. Memory and ROC [7]. You get a study list of nonsense words. A few days later you get a test list of words, half of which are new and half of which are old. Your task is to say whether each word on the test list is old or new. Assume that memory works in an all-or-none fashion, i.e. there is a certain probability that an item on the study list will be stored and will be recognized later. Items that are not on the study list aren't stored and therefore cannot be recognized. Assume your strategy is that you always say old to words on the test list that you recognize. For the words that you don't recognize you say that they are old with probability q. What does the ROC curve look like when you vary q? Can you think of experiments that test this model?

4. Luce's low threshold model for simple detection tasks [2]. Assume there is a sensory threshold and a human observer will go into a detect state ($D = y$) if the threshold is crossed and into a no-detect state otherwise ($D = n$). If a stimulus of a certain magnitude is presented this will happen with $P(D = y \mid S = y) = p_y$. Contrary to what we assumed in class the threshold is *low* and the observer will sometimes end up in a detect state even if there was no stimulus, i.e. $P(D = y \mid S = n) = p_n$ with $p_n > 0$ (for a *high* threshold model $p_n = 0$). If the observer always reports her detection state (she will say yes if she detects something and no if she doesn't) the hit rate will be $P(H) = p_y$ and the false alarm rate will be $P(FA) = p_n$. However, the subject might want to adapt her hit rate and false alarm rate to different pay-off situations. How could she do this? If she wanted a lower false alarm rate (i.e. $P(FA)$ should be smaller than p_n) she could report no detection in no-detect states and say yes only with probability $t < 1$ in detect states. Similarly, if she wanted a higher false alarm rate (i.e. $P(FA)$ should be greater than p_n) she could always say yes in detect states but also say yes with probability $u > 0$ in no-detect states. What does the ROC curve for such an observer look like? Make a plot.

5. Consider a subject in a detection experiment. On each trial there is a signal or not. Simulate an observer under the assumption of the equal variance Gaussian model, i.e. there is a decision axis X with the following signal-plus-noise and noise-only distributions:[1]

$$\begin{aligned} P(X = x \mid S = y) &= \frac{e^{-\frac{1}{2}\left(x-d'\right)^2}}{\sqrt{2\pi}} \\ P(X = x \mid S = n) &= \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} \end{aligned}$$

[1] in Matlab and Octave the standard normal distribution is given by `p=normpdf(x,0,1)`. The cumulative distribution function and its inverse are `P=normcdf(x,0,1)` and `x=norminv(P,0,1)`. You can draw a random sample with `r=normrnd(0,1)`. You can change the mean and the standard deviation by changing the parameters to something different than 0 and 1.

First plot the ROC curves for an observer with $d' = 0, \frac{1}{2}, 1, 2, 3$. How does the ROC curve relate to d'?
Now, assume that the subject has a d' of 1 and you are running an experiment with three different conditions. The stimulus is always the same in all conditions but the pay-offs, and hence the subject's criterion, are changed in each condition. Each condition has 200 trials, half of which are signal-plus-noise trials and half of which are noise-only trials. Assume that in condition 1 the subject's decision criterion is $\frac{1}{4}$, i.e. the subject will say yes if $X > \frac{1}{4}$. In condition 2 the criterion is $\frac{1}{2}$ and in condition 3 it's $\frac{3}{2}$. Simulate this subject for the three conditions: For each trial draw an X from the right Gaussian distribution and check whether it is greater than the criterion to generate the subject's response. Plot the empirical hit rate and false alarm rate on top of the theoretical ROC curve.

6. Rating data in detection experiments. Assume the same setup as in the previous exercise. Here, instead of reporting whether there was a signal or not the subject has a 4-point categorical response scale (0 = pretty certain that there was no signal, 1 = I lean towards no, 2 = I lean towards yes, 3 = pretty certain that there was a signal). There is just one condition with 600 trials and no pay-offs. Simulate a subject that says 0 if $X < \frac{1}{4}$, 1 if $\frac{1}{4} \leq X < \frac{1}{2}$, 2 if $\frac{1}{2} \leq X < \frac{3}{2}$, and 4 if $X \geq \frac{3}{2}$. Make a plot of the signal-plus-noise and noise-only distributions and mark the decision criteria for the 4 possible responses. How does the rating data relate to the detection data in the previous exercise? Can you translate the rating data to hit rate and false alarm rate and plot it on top of the ROC curve?

7. In case you later want to read up on signal detection theory, the classic book is [1]. A useful book is also [3]. Papers that provide an introduction to basic ideas are [4, 6]. In the lecture on Monday I closely followed [5].

References

[1] D. M. Green and J. A. Swets. *Signal Detection and Psychophysics (Reprint Edition).* Peninsula Publishing, 1988.

[2] R. D. Luce. A threshold theory for simple detection experiments. *Psychological Review*, 70(1):61–79, 1963.

[3] N. A. Macmillan and C. D. Creelman. *Detection Theory: A User's Guide.* Cambridge University Press, 1991.

[4] J. Swets, W. P. Tanner, and T. G. Birdsall. Decision processes in perception. *Psychological Review*, 68:301–340, Sep 1961.

[5] J. A. Swets. Is there a sensory threshold? *Science*, 134:168–77, 1961.

[6] J. A. Swets, R. M Dawes, and J. Monahan. Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1(1):1–26, 2000.

[7] A. P. Yonelinas and C. M. Parks. Receiver operating characteristics (ROCs) in recognition memory: a review. *Psychological Bulletin*, 133(5):800–832, 2007.

22 Solution 6: Signal Detection Theory I *2014-06-23*

Excercise 2

We want to get a better feeling of what the Expected Value of a random variable is and how to work with it. As we have already seen intuitively the Expected Value is just a weighted average over the sample space, so formally we get:

$$E\left(f\left(X\right)\right) = \sum_{x \in \Omega} f\left(x\right) \cdot p\left(x\right)$$

For equal probabilities we get the average as we know it. One important use of the Expected Value is that we can use it as moments of a distribution and thereby fully describe the distribution. The moments of a distribution are:

$$\begin{aligned} E\left(X\right) \\ E\left(X^2\right) \\ E\left(X^3\right) \\ . \\ . \end{aligned}$$

Using only $E\left(X\right)$ and $E\left(X^2\right)$ we can for example completely characterize the normal distribution as they describe mean and variance. Noting the importance of the expected Value we are going to look at some rules. First we show the linearity:

$$\begin{aligned} E\left(a \cdot f\left(X\right) + b\right) &= \sum_{x \in \Omega} \left[a \cdot f\left(x\right) + b\right] \cdot p\left(x\right) \\ &= a \cdot \underbrace{\sum f\left(x\right) p\left(x\right)}_{E(f(X))} + \underbrace{\sum b \cdot p\left(x\right)}_{b} \\ &= a \cdot E\left(f\left(X\right)\right) + b \end{aligned}$$

From the definition of the Expected Value follows directly:

$$E\left(f\left(X, Y\right)\right) = \sum_{x \in \Omega} \sum_{y \in \Omega} p(x, y) \cdot f\left(x, y\right)$$

And thereby we can deduce for independent random variables:

$$\begin{aligned} E\left(X + Y\right) &= \sum_{x \in \Omega} \sum_{y \in \Omega} p(x, y) \cdot f\left(x + y\right) \\ &= \sum_{x \in \Omega} \sum_{y \in \Omega} p\left(x\right) p\left(y\right) \cdot \left(x + y\right)) \\ &= \sum_{x \in \Omega} \sum_{y \in \Omega} p\left(x\right) p\left(y\right) x + p\left(x\right) p\left(x\right) p\left(y\right) y \\ &= \underbrace{\sum_{x \in \Omega} \sum_{y \in \Omega} p\left(x\right) p\left(y\right) x}_{E(X)} + \underbrace{\sum_{x \in \Omega} \sum_{y \in \Omega} p\left(x\right) p\left(y\right) y}_{\sum_x p\left(x\right) \sum_y p\left(y\right) y} \\ &= E\left(X\right) + E\left(Y\right) \end{aligned}$$

[1] Because $\sum_x p\left(x\right)$ normalizes to one.

Further we note:

$$\begin{aligned} E(X \cdot Y) &= \sum_{x \in \Omega} \sum_{y \in \Omega} p(x)\, p(y) \cdot (x \cdot y) \\ &= \sum_{x} p(x)\, x \cdot \sum_{y} p(y)\, y \\ &= E(X) \cdot E(Y) \end{aligned}$$

That we can use to derive an expression for the variance starting with the standard definition:

$$\begin{aligned} var(X) &= E\left((X - E(X))^2\right) \\ &= E\left(X^2 - 2E(X)\, X + E(X)^2\right) \\ &= E\left(X^2\right) + 2E(X)\, E(X) + E(X)^2 \\ &= E\left(X^2\right) - E(X)^2 \end{aligned}$$

And from this we get:

$$\begin{aligned} var(X+Y) &= E\left((X+Y)^2\right) - E(X+Y)^2 \\ &= E\left(X^2 + 2XY + Y^2\right) - (E(X) + E(Y))^2 \\ &= E\left(X^2\right) + 2E(XY) + E\left(Y^2\right) - E(X)^2 + 2E(X)\, E(Y) + E\left(Y^2\right) \\ &= E\left(X^2\right) - E(X)^2 + E\left(Y^2\right) - E(Y)^2 \\ &= var(X) + var(Y) \end{aligned}$$

Applying the gained knowledge to a thumbtack experiment with n tosses and $X_1 ... X_n$ outcomes as random variables. Now we can calculate the Expected Value and the Variance for the X_i:

$$\begin{aligned} E(X_i) &= (1-p) \cdot 0 + p \cdot 1 \\ &= p \end{aligned}$$

$$\begin{aligned} var(X_i) &= \underbrace{E\left(X_i^2\right)}_{E(X_i)} - \underbrace{E(X_i)^2}_{p^2} \\ &= p - p^2 \\ &= p \cdot (1-p) \end{aligned}$$

For the sum of the X_i as new random variable $N = \sum_{i=1}^{n} X_i$ we get:

$$\begin{aligned} E(N) &= n \cdot p \\ var(N) &= n \cdot p \cdot (1-p) \end{aligned}$$

Excercise 3

As we want to get the ROC we have to look at the Hits and the False Alarms which we get assuming a high threshold model:

$$\begin{aligned}
P(H) &= P(Response = old | Stimulus = old) \\
&= \underbrace{P(R = old | Memory = old)}_{1} \underbrace{P(M = old | S = old)}_{P_r} \\
&+ \underbrace{P(R = old | M = new)}_{q} \underbrace{P(M = new | S = old)}_{1-P_r} \\
&= \underbrace{P(M = old | S = old)}_{P_r} + q \cdot (1 - P_r)
\end{aligned}$$

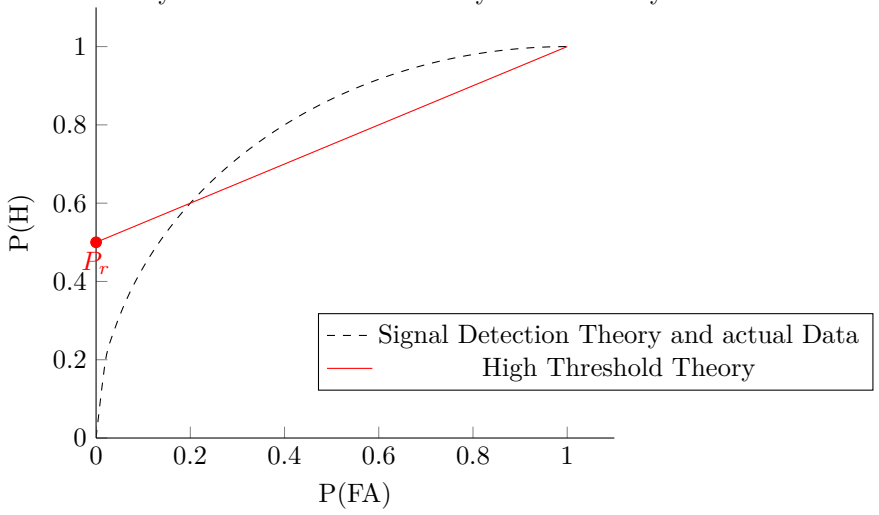
With q as the probability with which we say we remember a Stimulus even though we don't.

$$\begin{aligned}
P(FA) &= P(R = old | S = new) \\
&= \underbrace{P(R = old | M = old)}_{1} \cdot \underbrace{P(M = old | S = new)}_{0} \\
&+ \underbrace{P(R = old | M = new)}_{q} \cdot \underbrace{P(M = new | S = new)}_{1} \\
&= q
\end{aligned}$$

That means we can express $P(H)$ as a linear function of the False Alarms:

$$P(H) = P_r + P(FA) \cdot (1 - P_r) \tag{6}$$

And therefore we would get a linear ROC curve which, as we already have seen, does not correspond to the data. It would mean that we only have false alarms because we chose to answer opposite to our actual memory and not because we falsely remember any stimuli.



23 Solution 6: Signal Detection Theory II *2014-06-27*

Excercise 4

Luce's idea is: If High Threshold Theory does not work, maybe by using a low threshold, ROC curves will be explainable. This idea translates as follows:

$$\begin{aligned} S \in \{y,n\}, D \in \{y,n\}, R \in \{y,n\} \\ P\left(D = y | S = y\right) = p_y \\ P\left(D = y | S = n\right) = p_n > 0 \end{aligned}$$

We have to consider to different cases of the subject's behaviour.
Case 1: The subject want to lower his false-alarm rate $p(FA)$:

$$\begin{aligned} & P\left(R = y | D = n\right) = 0 \\ & P\left(R = y | D = y\right) = t < 1 \\ \Rightarrow P(FA) &= P\left(R = y | D = y\right) && \cdot P\left(D = y | S = n\right) && + P\left(R = y | D = n\right) && \cdot P\left(D = n | S = n\right) \\ &= t && \cdot p_n && + 0 && \cdot \left(1 - p_n\right) \\ &= t \cdot p_n \\ \Rightarrow P(H) &= P\left(R = y | D = y\right) && \cdot P\left(D = y | S = y\right) && + P\left(R = y | D = n\right) && \cdot P\left(D = n | S = y\right) \\ &= t && \cdot p_y && + 0 && \cdot \left(1 - p_y\right) \\ &= t \cdot p_y \\ &= \frac{P(FA)}{p_n} \cdot p_y \end{aligned}$$

Now the first part of the plot looks like this:

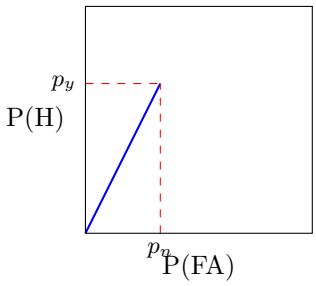


Figure 23: Low Threshold Model, first part of plot with slope of $\frac{p_y}{p_n}$

Case 2: The subject wants a higher hit rate $p(H)$:

$$
\begin{aligned}
& P\left(R=y|D=Y\right)=1 \\
& P\left(R=y|D=n\right)=u>0 \\
\Rightarrow P(FA) &= P\left(R=y|D=y\right) && \cdot P\left(D=y|S=n\right) && + P\left(R=y|D=n\right) && \cdot P\left(D=n|S=n\right) \\
&= 1 && \cdot p_n && + u && \cdot \left(1-p_n\right) \\
&= p_n + u \cdot \left(1-p_n\right) \\
\Leftrightarrow u &= \frac{P(FA)-p_n}{1-p_n} \\
\Rightarrow P(H) &= P\left(R=y|D=y\right) && \cdot P\left(D=y|S=y\right) && + P\left(R=y|D=n\right) && \cdot P\left(D=n|S=y\right) \\
&= 1 && \cdot p_y && + u && \cdot \left(1-p_y\right) \\
&= p_y + u \cdot \left(1-p_y\right) \\
&= p_y + \frac{P(FA)-p_n}{1-p_n} && \cdot \left(1-p_y\right) \\
&= \frac{1-p_y}{1-p_n} \cdot P(FA) && + \frac{p_y-p_n}{1-p_n}
\end{aligned}
$$

We can finish our plot:

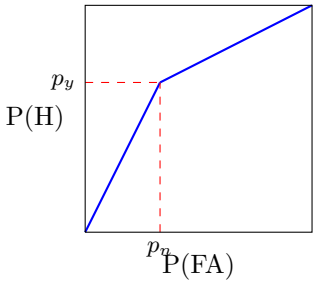


Figure 24: Low Threshold Model Plot

As can be seen, Low Threshold Theory better fits the actual data since its form is, compared with the linear function of High Threshold Theory, more like that of an ROC curve. Nevertheless it is not the right model for what is really going on.

Excercise 5

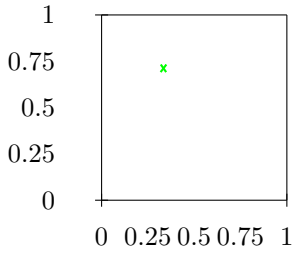


Figure 25: One datapoint of a Signal Detection Experiment

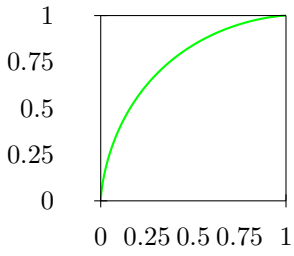


Figure 26: ROC curve with distance of gaussians d = 1

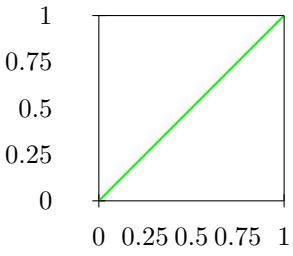


Figure 27: ROC curve with distance of gaussians d = 0

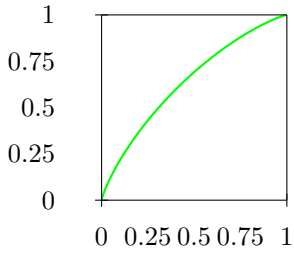


Figure 28: ROC curve with distance of gaussians d = 1/2

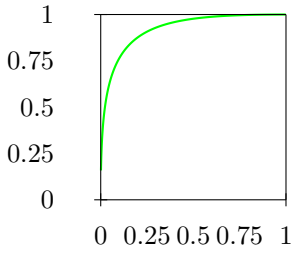


Figure 29: ROC curve with distance of gaussians d = 2

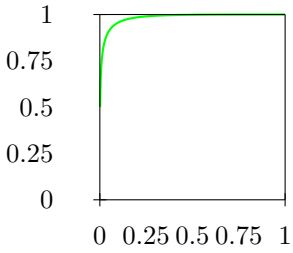


Figure 30: ROC curve with distance of gaussians d = 3

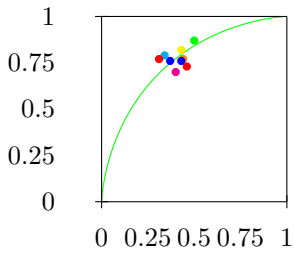


Figure 31: Subject with criterion c=1/4

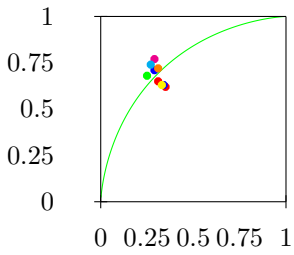


Figure 32: Subject with criterion c=1/2

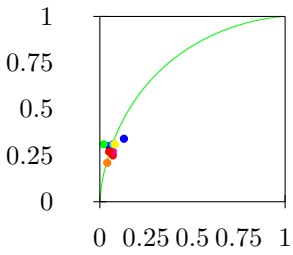


Figure 33: Subject with criterion c=3/2

Excercise 6

Signal detection experiments are pretty extensive – 200 trials mean nothing at all. Before the experimenter has 15,000 trials it is not worth even starting (or so a wise man said). Thus we can introduce a method that gives more data for each. To do this, we introduce a scale from 0 to 3, where 0 is absolutely not and 3 absolutely yes. It could look like this:

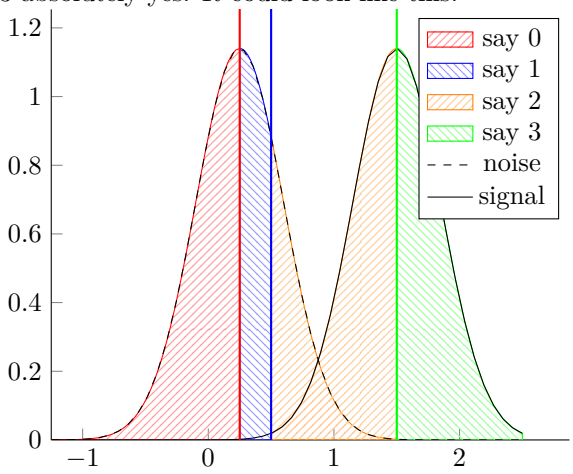


Figure 34: Gaussians with different criterions

Now varying the range on the scale we count as an 'no'-response, we get more points for our ROC curve.
Example: Take 0 as 'no', 1-3 as 'yes → plot this point
Take 0-1 as 'no', 2-3 as 'yes → plot this point
Take 0-2 as 'no', 3 as 'yes → plot this point

Note: *Instead using ROC curves is a great idea, but be aware that more than 7 response categories are hard to handle by the subjects.*

24 Signal Detection Theory III *2014-06-20*

24.1 From YN to 2AFC

In comparison to simple Yes-No-task there exists an alternative task design which is the 2-Alternative-Forced-Choice-task. In each trial the subject is presented with two intervals with a light stimulus in one of it, therefore there are two "stimulations" X_1 and X_2. The subject is then forced to state in which interval the stimulus appeared. By this we get a probability distribution for the stimulation in each interval. The probabilities for this experiment are given in the following.

$$
\begin{aligned}
(X_1|S=1) &\sim N\left(\Delta\mu, \sigma^2\right) \\
(X_2|S=1) &\sim N\left(0, \sigma^2\right) \\
(X_1|S=2) &\sim N\left(0, \sigma^2\right) \\
(X_2|S=2) &\sim N\left(\Delta\mu, \sigma^2\right)
\end{aligned}
$$

We are using again the same Gaussian's with different means. This is also referred to as *equal variance signal detection model* and may be plotted like this:

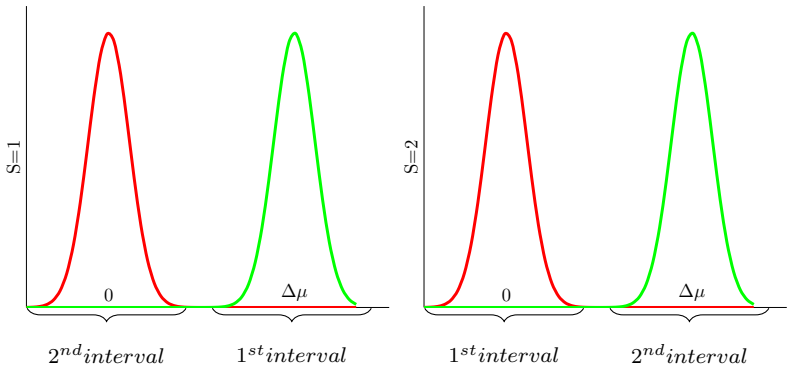


Figure 35: Mean-shifted Gaussian distributions

If the stimulus was presented in the 1^{st} interval X_2 (our sensation for the 2^{nd} interval) is so to say the noise distribution and the other way around if the stimulus is shown in the 2^{nd} interval. If we now choose the variables X_1 and X_2 as the axis we get the following plot:

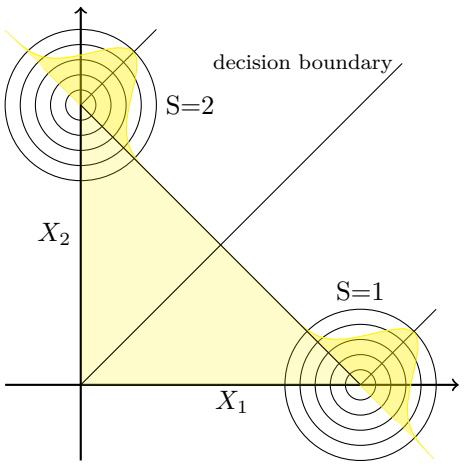


Figure 36: The distance between the two distributions is $\sqrt{2}\Delta\mu$.

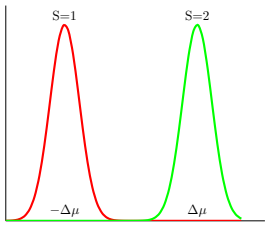
If we could discriminate perfectly our data points would lie on the x or y axis for each trial (depending on the interval). When we discussed why HT-Theory is wrong (20.3), we already stated that subjects perform better in 2AFC than in YN-Tasks. Now we see why: the distance of the two distributions is $\Delta\mu_{2AFC} = \sqrt{2}\Delta\mu$. This is $> \Delta\mu$! The strategy for the best performance in 2AFC is the following:

- Say 1 if $X_1 > X_2$
- Say 2 if $X_2 \geq X_1$

We see that $\Delta X = X_2 - X_1 \overset{!}{>} 0$. What is now the distribution of ΔX? Note that if you scale or add normal distributions you always get again a normal distribution with scaled standard deviations and means and added variances and means.

$$(\Delta X|S=1) \sim N\left(0,\sigma^2\right) - N\left(\Delta\mu, 2\sigma^2\right) = N\left(-\Delta\mu, 2\sigma^2\right)$$
$$(\Delta X|S=2) \sim N\left(\Delta\mu,\sigma^2\right) - N\left(0, 2\sigma^2\right) = N\left(\Delta\mu, 2\sigma^2\right)$$

We may now calculate the Signal to Noise Ratio of this two distributions.



$$\begin{aligned} SNR &= \frac{\Delta\mu - (-\Delta\mu)}{\sqrt{2\sigma^2}} \\ &= \frac{2\Delta\mu}{\sqrt{2}\sigma} \\ &= \sqrt{2}\Delta\mu \quad \text{(same result as in the geometric solution)} \end{aligned}$$

24.2 Cue Combination
Ernst & Banks (2002): Visio-haptic cue combination

The task in this experiment is to judge the size of a bar when you can see and feel it. Your two measurements of s may be defined as the following:

$$V \sim N\left(s, \sigma_V^2\right)$$
$$H \sim N\left(s, \sigma_H^2\right)$$

In this is example it is not wise to choose the same distribution for both measurements, since we would expect that our visual system is more accurate than the haptic one (imagine $s = 2cm$, figure 37). Different than in the 2AFC the two distributions have the same mean – leading to the plot in figure 38. Given the length of the bar (s) this is the probability for our haptic (h) and visual (v) impression:

$$p\left(V=v; H=h|s\right) = \frac{1}{\sqrt{2\pi}\sigma_V} e^{-\frac{1}{2}\left(\frac{v-s}{\sigma_V}\right)^2} \frac{1}{\sqrt{2\pi}\sigma_H} e^{-\frac{1}{2}\left(\frac{h-s}{\sigma_H}\right)^2}$$

Together with the log-likelihood we can calculate a ML-Estimate $\hat{s}$ for s:

$$\Rightarrow -\frac{1}{2}\left(\left(\frac{v-\hat{s}}{\sigma_V}\right)^2 + \left(\frac{h-\hat{s}}{\sigma_H}\right)^2\right) = -\frac{1}{2}\left(\frac{v-\hat{s}}{\sigma_V}\right)^2 - \frac{1}{2}\left(\frac{h-\hat{s}}{\sigma_H}\right)^2$$

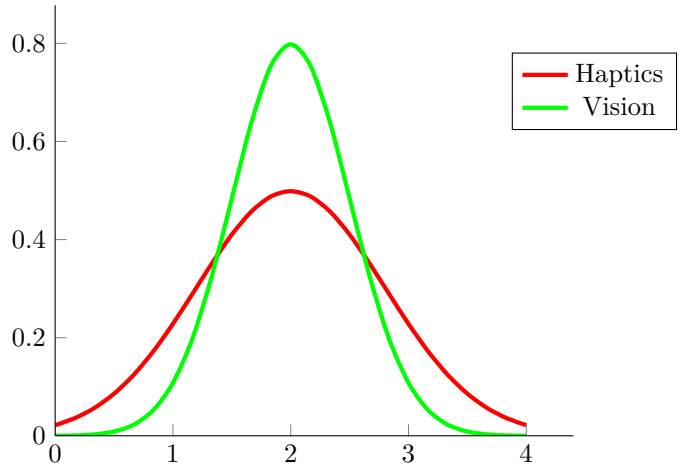


Figure 37: The visual system is more accurate than the haptic, thus the normal distribution of the visual system has a smaller variance.

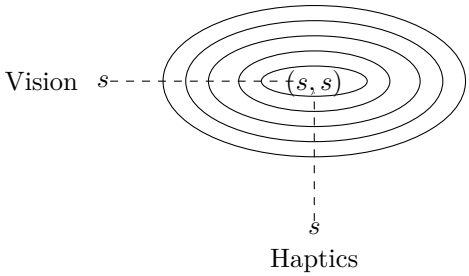


Figure 38: Vision and haptic systems' distributions have the same mean.

Use first derivative:

$$\left(\frac{v-\hat{s}}{\sigma_V}\right)\frac{2}{2\sigma_V}+\left(\frac{h-\hat{s}}{\sigma_H}\right)\frac{2}{2\sigma_H}\stackrel{!}{=}0$$

$$\Leftrightarrow \quad \frac{v-\hat{s}}{\sigma_V^2}+\frac{h-\hat{s}}{\sigma_H^2}=0$$

$$\Leftrightarrow \quad \frac{v}{\sigma_V^2}+\frac{h}{\sigma_H^2}-\hat{s}\left(\frac{1}{\sigma_V^2}+\frac{1}{\sigma_H^2}\right)=0$$

$$\Leftrightarrow \quad \frac{v}{\sigma_V^2}+\frac{h}{\sigma_H^2}=\hat{s}\left(\frac{1}{\sigma_V^2}+\frac{1}{\sigma_H^2}\right)$$

$$\Leftrightarrow \quad \hat{s}=\left(\frac{v}{\sigma_V^2}+\frac{h}{\sigma_H^2}\right)\frac{\sigma_V^2\sigma_H^2}{\sigma_V^2+\sigma_H^2}$$

$$\Leftrightarrow \quad \hat{s}=\frac{v\sigma_H^2}{\sigma_V^2+\sigma_H^2}+\frac{h\sigma_V^2}{\sigma_V^2+\sigma_H^2}$$

This estimate seems logical, since the variances are used as a normalization term in the denominator and the numerator weights our sensation according to their internal variance. In our example we assumed the visual system to have a small variance compared to the haptic system, so the v has greater impact on $\hat{s}$.

PMPC Tutorial Sheet 7

1. The Poisson process is frequently used to model temporal events, e.g. the occurence of earthquakes, phone calls or action potentials. The Poisson process has one free parameter: the rate r at which events occur. The probabity that N events happen during an interval of duration t is given by[1]

$$P(N = n|r, t) = \frac{(rt)^n}{n!}\exp(-rt).$$

   Derive the maximum likelihood estimator for the rate parameter r in the Poisson model. Assume the number of events and the time interval to be fixed by an experiment. For which value of the rate parameter is the likelihood function at a maximum? Bonus question: Show that the expected number of events and the variance are both rt.[2]

2. Stimulus decoding [1, Chapters 1 and 3]. In the lobular plate of the fly (*calliphora vicina*) brain there are only two spiking neurons. These neurons are very easy to record from which has made them a popular model system for theoretical studies [2]. These neurons are selective for horizontal visual motion and they are called H1 neurons. One of them is selective for leftward motion, the other one for rightward motion. You record from one of them and find that it has a firing rate of 20 Hz for leftward motion and 80 Hz for rightward motion. In a follow-up experiment either a leftward or a rightward motion (each with probability $\frac{1}{2}$) is presented to the fly in each trial while you are still recording from the same neuron. Assume that the firing of the neuron is a Poisson process for leftward and rightward motion that differs only in the firing rate. Make a plot of the two distributions for the number of spikes in a 100 ms time interval. Where do the two distributions intersect? Calculate the log likelihood ratio for leftward and rightward motion when you observe n spikes in the 100 ms after the stimulus onset. Make a plot of the log likelihood ratio as a function of n.[3]

3. Imagine a homunculus downstream from the the above neuron. The homunculus counts the neurons for 100 ms after the onset of each trial. The homunculus isn't able to see the input to the fly's eye. For simplicity we will also assume that he has no access to the other H1 neuron. His only access to the visual world is through this one neuron. He needs to make a decision as to whether the motion is leftward or rightward because he

[1] Often one finds the Poisson distribution parametrized with only one parameter $\lambda := rt$

[2] Hint: Since the distribution is normalized to one you know that $\sum_{n=0}^{\infty} \frac{(rt)^n}{n!} = \exp(rt)$.

[3] The Poisson distribution is called `poisspdf` in Matlab and Octave. It has one parameter $\lambda := rt$, the expected number of events.

needs to steer the fly away from a wall, for example. How well is he able to do this? Plot the Receiver Operating Characteristic, i.e. the hit-rate vs the false-alarm rate for varying decision criteria.

4. Check that the characteristic function of a Gaussian random variable X with mean μ and variance σ^2 is

$$\chi_X(\omega) = e^{i\mu\omega - \frac{1}{2}\sigma^2\omega^2}.$$

Reminder: The Gaussian distribution is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

and the characteristic function is defined as

$$\begin{aligned} \chi_X(\omega) &= E\left(e^{i\omega X}\right) \\ &= \int_{-\infty}^{\infty} f_X(x) e^{i\omega x} dx \end{aligned}$$

where $i^2 = -1$.

5. Show that the n'th derivative of the characteristic function evaluated at zero is related to the n'th moment of the distribution in the following way:

$$\chi_X^{(n)}(0) = i^n E\left(X^n\right).$$

Use this relationship to derive the first two moments of the Gaussian distribution.

References

[1] P. Dayan and L. F. Abbott. *Theoretical Neuroscience.* MIT Press, Cambridge, MA, 2001.

[2] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code.* MIT Press, Cambridge, MA, 1997.

26 Solution 7: Signal Detection Theory III *2014-06-30*

Exercise 1

Why do we always use a Gaussian to model all those things? Can't we just use other models as well?
We think about neurons as Poisson processes.
If we just look at the spikes, a neuron's firing behavior looks roughly like figure 39a. To model it, we discretize over time and say for each cell whether the neuron fires or not (figure 39b) and can then calculate the probability.
We can then find the distribution over time for the neuron by taking an interval of time and counting the number of spikes. This will result in the Poisson distribution: $P(N = n|r, t) = \frac{(rt)^n}{n!}e^{rt}$.
Assuming we have 10 spikes in a second and plug this into the Poisson distribution, we should intuitively end up with 10 Hz.

$$
\begin{aligned}
P(N = n|r, t) &= \frac{(rt)^n}{n!}e^{-rt} \\
\log P(N = n|r, t) &= n \log rt - \log n! - rt \\
\frac{\partial \log P(N = n|r, t)}{\partial r} &= \frac{n}{rt}t - t \\
\frac{\partial^2 \log P(N = n|r, t)}{\partial^2 r} &= -\frac{n}{r^2}
\end{aligned}
$$

Use the first derivative to maximize it:

$$
\begin{aligned}
\frac{\partial \log P(N = n|r, t)}{\partial r} &= 0 \\
\frac{n}{\hat{r}t}t - t &= 0 \\
\frac{n}{\hat{r}t}t &= t \\
\frac{n}{\hat{r}t} &= 1 \\
\frac{n}{t} &= \hat{r}
\end{aligned}
$$

Bonus Question

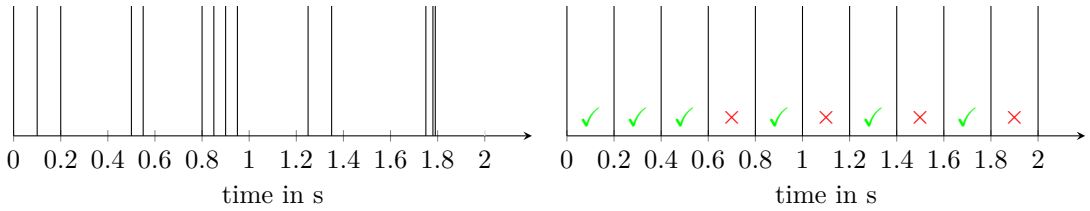
$$
\begin{aligned}
E(N) &= \sum_{n=0}^{\infty} n\frac{(rt)^n}{n!}e^{-rt} \\
&= e^{-rt}\sum_{n=0}^{\infty}\frac{(rt)^n}{(n-1)!}
\end{aligned}
$$

use:

$$
\sum_{n=0}^{\infty}\frac{(rt)^n}{n!}e^{-rt} = 1
$$

for:

$$
\begin{aligned}
E(N) &= e^{-rt}(rt)\sum_{n=1}^{\infty}\frac{(rt)^{(n-1)}}{(n-1)!} \\
&= e^{-rt}(rt)\sum_{n=0}^{\infty}\frac{(rt)^n}{n!} \\
&= e^{-rt}(rt)e^{rt} = rt
\end{aligned}
$$



(a) "Data" for firing neurons: each line stands for a firing.

(b) Model for firing neurons: Bins to show whether a neuron fired during that time or not.

Figure 39: Example data for firing neurons and the respective model

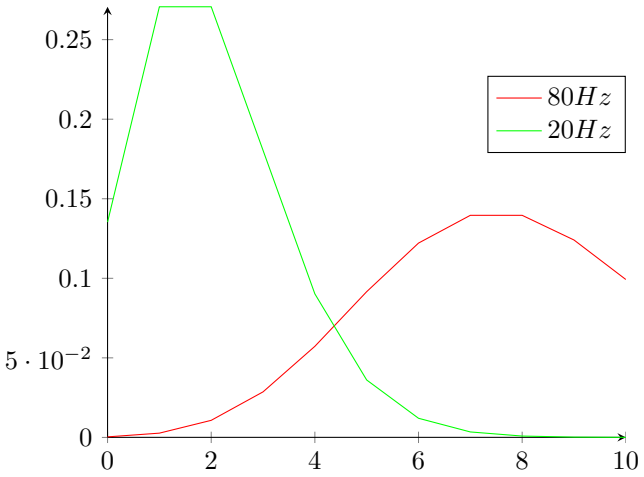


Figure 40: The poisson distributions for $20Hz$ and $80Hz$.

So it's exactly what we said above: If we have 10 spikes in a second, we come up with 10 Hz.

$$
\begin{aligned}
var(N) &= E(N^2) - E(N)^2 \\
&= \sum_{n=0}^{\infty} n^2 \frac{(rt)^n}{n!} e^{-rt} - (rt)^2 \\
&= \sum_{n=0}^{\infty} (n+1) \frac{(rt)^{n+1}}{n!} e^{-rt} \\
&= (rt) e^{-rt} \left[\underbrace{\sum_{n=0}^{\infty} n \frac{(rt)^n}{n!}}_{rt \cdot e^{rt}} + \underbrace{\sum_{n=0}^{\infty} \frac{(rt)^n}{n!}}_{e^{rt}} \right] - (rt)^2 \\
&= (rt) e^{-rt} \left(rt e^{rt} + e^{rt} \right) - (rt)^2 \\
&= (rt)(rt+1) - (rt)^2 = rt
\end{aligned}
$$

This variance corresponds with the Weber-Fechner-law (for details see Wikipedia): If the signal increases, also the noise (i.e. the variance) increases.

Exercise 2

$$
\frac{P(N=n|r=80Hz, t=100ms)}{P(N=n|r=20Hz, t=100ms)} = \frac{\frac{8^n}{n!} e^{-8}}{\frac{2^n}{n!} e^{-2}} = \left(\frac{8}{2}\right)^n e^{-8+2}
$$

For calculating the intersection we set $\left(\frac{8}{2}\right)^n e^{-8+2} = 1$.

$$\begin{aligned}\left(\frac{8}{2}\right)^n e^{-8+2} &= 1\\ 4^n &= e^6\\ n\log 4 &= 6\\ n &\approx 4.3\end{aligned}$$

Exercise 3

We don't have this solution. If you have it, feel free to contact us so we can add it.

Exercise 4

We are given $\chi_X(\omega) = E(e^{i\omega X}) = \int_{-\infty}^{\infty} f_X(x)e^{i\omega x}dx$, which is the Fourier transformation $\mathcal{F}(f_X)(\omega)$.

We assume two independent random variables X, Y with $Z = X+Y$ and the functions f_X, f_Y, and f_Z, respectively and derive the convolution $f_Z(z)$.

Note: $*$ *denotes a convolution.*

$$\begin{aligned}f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(\underbrace{z-x}_{!})dx\\ &= \int_{-\infty}^{\infty} f_X(x) f_X(\underbrace{z-y}_{!})dy\\ &= f_X(z) * f_Y(z)\end{aligned}$$

We can use this information to derive the characteristic function $\chi_Z(\omega)$.

$$\begin{aligned}\chi_Z(\omega) &= \chi_{X+Y}(\omega) = E(e^{i\omega(X+Y)})\\ &= E(e^{i\omega X}e^{i\omega Y}) = E(e^{i\omega X})E(e^{i\omega Y})\\ &= \chi_X(\omega) + \chi_Y(\omega).\end{aligned}$$

Exercise 5

This exercise was not discussed in class.

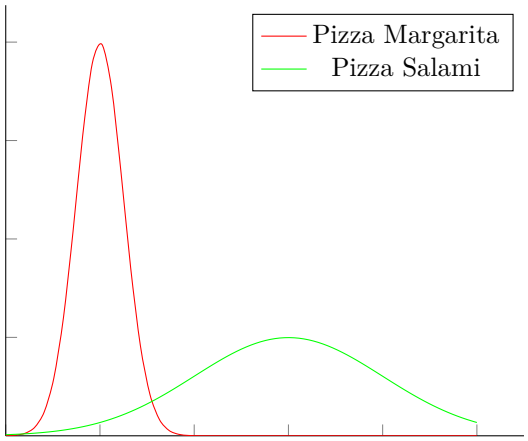


Figure 42: Red: Pizza margarita is quite popular among all subjects. Green: Pizza salami is not that popular among subjects, it has a higher variance.

27 Choice Models I *2014-07-04*

Utility

The *utility* is the variability in choices. It can either refer to the variability in choices of several subjects ("How many subjects prefer pizza tonno over pizza salami?") or the variability in choices of a single subject over time ("On how many days prefers the subject pizza tonno over pizza salami?"). A utility of e.g. 70% means that a subject chooses pizza tonno over pizza salami in 70 out of 100 times it's asked.
With choice models we try to find the utility of possible choices in order to make accurate predictions.
Note that there might be *polarizing* options, this means the variance changes. For example pizza margarita might be very popular, so many people like it thus the variance for a choice of pizza margarita gets smaller. However, pizza salami might be less popular and therefor its utility's variance is wider (see figure 42).
Since the utility is dependent on the choice to made, there can only be *relative* utilities.

Figure 41: Pizzaaaa!

27.1 Paired Comparison Experiment

A very common technique to check whether subjects prefer an option over another is a paired comparison experiment. Subjects are shown *all possible pairs* and say for each pair which option they prefer. This results in a matrix where we can find out which options are more popular than others. See table 3 for an example.

Two options

Note: *To make it easier to understand the math we will assume equal variance for given options unless noted otherwise.*

Assume we ask a subject to make a choice between two options. We consider two options i and j with the random variables x_i and x_j as their utilities (the subject's utilities for each option respectively) with $x_i \sim \mathcal{N}(\mu_i, 1)$ and $x_j \sim \mathcal{N}(\mu_j, 1)$. The subject "computes" $\Delta x_{ij} = x_i - x_j$ if $\Delta x_{ij} > 0$ (otherwise we would need Δx_{ji}). Δx_{ij} is also normal distributed, i.e. $\Delta_x ij \sim \mathcal{N}(\mu_i - \mu_j, 2)$. Δx_{ij} is the distribution of how likely it is, that the subject chooses i over j. A visualization of this can be found in figure 43.

these options are chosen \ over these options		
—	$\frac{15}{60}$	$\frac{40}{60}$
$\frac{45}{60}$	—	$\frac{30}{60}$
$\frac{20}{60}$	$\frac{30}{60}$	—

Table 3: Example paired comparison outcome. In this example we assume we asked 60 subjects, hence the denominator.
For computations we often set the diagonal (compare each option with itself) to $\frac{1}{2}$, which means there is no preference – this already makes sense intuitively.

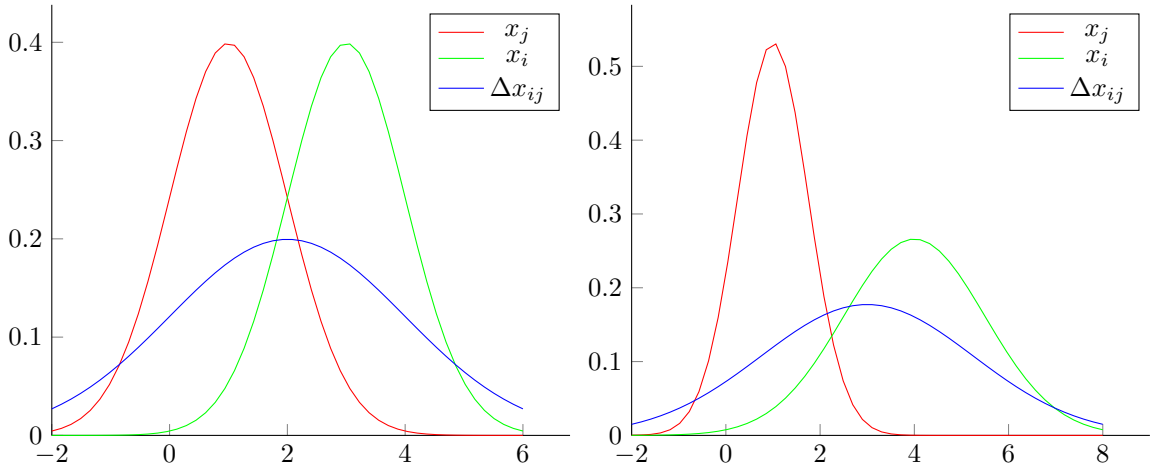


Figure 43: Relation of x_i, x_j and Δx_{ij}. Left: two options with equal variance. Right: two options with different variances. Note that the variances add up, hence Δx_{ij} gets flat and wide.

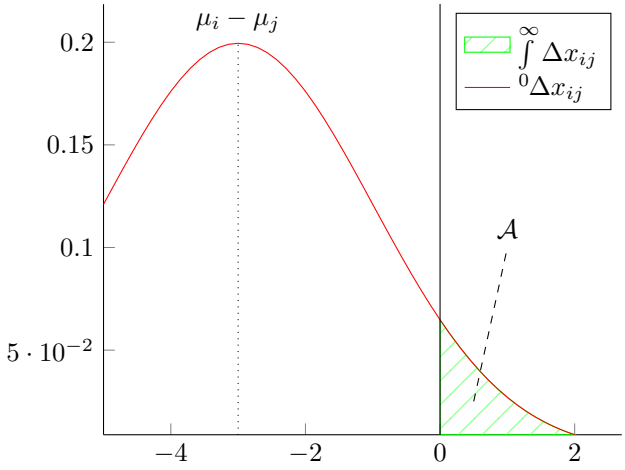


Figure 44: Δx_{ij}: If we know this area $\mathcal{A}$, we can guess $\Delta \mu_{ij}$.

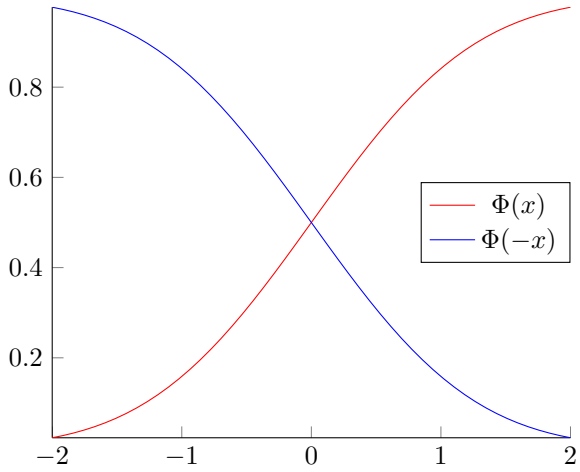


Figure 45: $\Phi(x) = 1 - \Phi(-x)$

Definitions

We define:

- $d_{ij} = \mu_i - \mu_j$
- p_{ij} is the probability that the subject chooses i over j.
- $P_{ij} = 1 - P_{ji} = 1 - \Phi(d_{ij}; 0, \sqrt{2}) = 1 - \Phi(-d_{ij}; 0, \sqrt{2}) \overset{\text{z-tf.}}{=} 1 - \Phi\left(-\frac{d_{ij}}{\sqrt{2}}; 0, 1\right)$
  Figure 45 shows visually that $\Phi(x) = 1 - \Phi(-x)$, so we can write $1 - \Phi\left(-\frac{d_{ij}}{\sqrt{2}}; 0, 1\right) = \Phi\left(\frac{d_{ij}}{\sqrt{2}}; 0, 1\right)$.
  To further simplify the notation we just write $\Phi\left(\frac{d_{ij}}{\sqrt{2}}\right)$.
- q_{ij}

Optimizing Paired Comparison Experiment

We search for an estimate of d_{ij}. We can use $q_{ij} = \Phi\left(\frac{\hat{d}_{ij}}{\sqrt{2}}\right)$ and derive $\Phi^{-1}(q_{ij}) = \frac{\hat{d}_{ij}}{\sqrt{2}} \Leftrightarrow \sqrt{2}\Phi^{-1}(q_{ij}) = \hat{d}_{ij}$.
$\hat{d}$ is a matrix (figure 46) with the estimated distances for $\hat{d}_{ij} = \hat{\mu}_i - \hat{\mu}_j$. **Note:** $\forall i \colon \hat{d}_{ii} = 0$
Hence $\hat{d}_{ij} = -\hat{d}_{ji}$. But what are good estimates for $\hat{\mu}_i$ and $\hat{\mu}_j$?

$$\hat{d} = \begin{pmatrix} 0 & & \cdots & \hat{d}_{ij} & \\ & 0 & & & \\ \vdots & & 0 & & \vdots \\ \hat{d}_{ji} & & & 0 & \\ & & \cdots & & 0 \end{pmatrix}$$

Figure 46: $\hat{d}$, note that $d_{ji} = -d_{ij}$ and the diagonal is 0.

We have $\frac{n(n-1)}{2}$ pairs, that means we have $(n-1)$ free parameters. This means we will not be able to determine the x-shift. This shouldn't bother us too much, since we are only interested in the difference between μ_i and μ_j anyway.
A method for minimizing (thus optimizing) the error in our estimates we can use the least squares estimate (LSE).

27.2 Least Squares Estimate

The idea of the least squares estimate is to *minimize the sum of all squared differences.*

$$Q = \frac{1}{2}\left(\sum_j \sum_i \left(\hat{\mu}_i - \hat{\mu}_j - \hat{d}_{ij}\right)\right)$$

So we want to minimize Q with respect to $\hat{\mu}_i$ and $\hat{\mu}_j$.
This can be done easily by taking the first derivative and setting it to 0.

$$\begin{aligned}
\frac{\partial Q}{\partial \hat{\mu}_k} &= \left(\sum_j \hat{\mu}_k - \hat{\mu}_j - \hat{d}_{kj}\right) - \left(\sum_i \hat{\mu}_i - \hat{\mu}_k - \hat{d}_{ik}\right) = 0 \\
&\Leftrightarrow \left(\sum_j \hat{\mu}_k - \hat{\mu}_j - \hat{d}_{kj}\right) + \left(\sum_i \hat{\mu}_k - \hat{\mu}_i \underbrace{+\hat{d}_{ik}}_{\text{Remember: } d_{ij}=-dji}\right) = 0 \\
&\Leftrightarrow \underbrace{\left(\sum_j \hat{\mu}_k - \hat{\mu}_j - \hat{d}_{kj}\right) + \left(\sum_i \hat{\mu}_k - \hat{\mu}_i - \hat{d}_{ki}\right)}_{\text{twice the same}} = 0 \\
&\Leftrightarrow 2\left(\sum_i \hat{\mu}_k - \hat{\mu}_i - \hat{d}_{ki}\right) = 0 \\
&\Leftrightarrow \sum_i \hat{\mu}_k - \hat{\mu}_i - \hat{d}_{ki} = 0 \\
&\Leftrightarrow n\hat{\mu}_k - \sum_i \hat{\mu}_i - \sum_i \hat{d}_{ki} = 0 \\
&\Leftrightarrow \hat{\mu}_k - \frac{1}{n}\sum_i \hat{\mu}_i = \frac{1}{n}\sum_i \hat{d}_{ki}
\end{aligned}$$

We end up with n equations (one for each k) in n unknowns. However the system's rank is $n-1$, so the system of equations is underdetermined. This means that to solve it we are free to choose something as we want, and obviously we set the average of the μ_is to 0 and get a nice formula to calculate the average over all distances, $\hat{\mu}_k$.

$$\frac{1}{n}\sum_i \hat{d}_{ki} \stackrel{!}{=} 0 \Rightarrow \hat{\mu}_k = \frac{1}{n}\sum_i \hat{d}_{ki}$$

Simple example

We have three normal distributions i, j, and k with the means $\mu_i = 1, \mu_j = 0$, and $\mu_k = -1$ (figure 47). For these distributions we can simply derive the matrix d and calculate the average distance between two plots.

$$d = \begin{pmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{pmatrix}$$

$$\mu_i = \frac{1}{n}\sum_l d_{il} = \frac{1}{3}\left(d_{i1} + d_{i2} + d_{i3}\right) = \frac{1}{3}\left(0 + 1 + 2\right) = 1$$

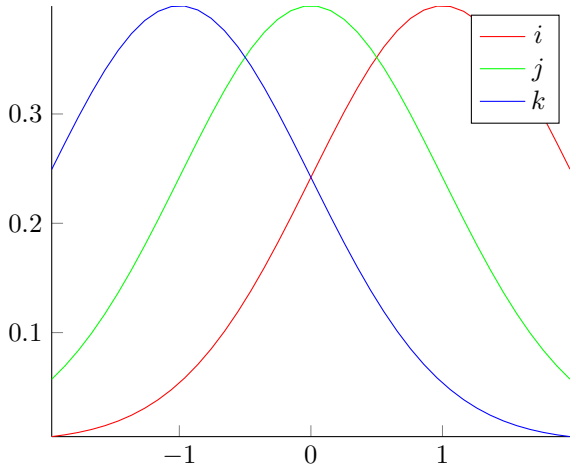


Figure 47: Three normal distributions i, j, and k.

28 Choice Models II *2014-07-07*

28.1 Thurstone Scaling

In 1920 Thurstone thought about measurements in Psychology. He conducted an experiment where he asked the subjects whether they think that one crime is more serious than another. Of course there exists no such thing like a crime-seriousness scale, but by comparing all pairs of answers, Thurstone could construct one.

This technique is also used in the Elo rating (famous amongst chess players) or the similar X-Box's Trueskill. These scales are used to match players of equal skill. The problem is, that you lack enough data to apply our method (you will never have a nearly complete matrix of all X-Box players competing against each other in one specific game). Good thing: we do not need the whole matrix! For subsets of the matrix we can predict new matches based on common past enemies. And (considering the X-Box setting) the matchmaker can also optimize their information by matching the right people together. But how do we know, that all this rating is formally correct?

28.2 A little bit of Measurement Theory

Consider the problems of an IQ-Test. You lack a concrete scale for the intelligence of a person as well as a 'suitable' opponent for a match up. The solution: to match the subject against the test items.

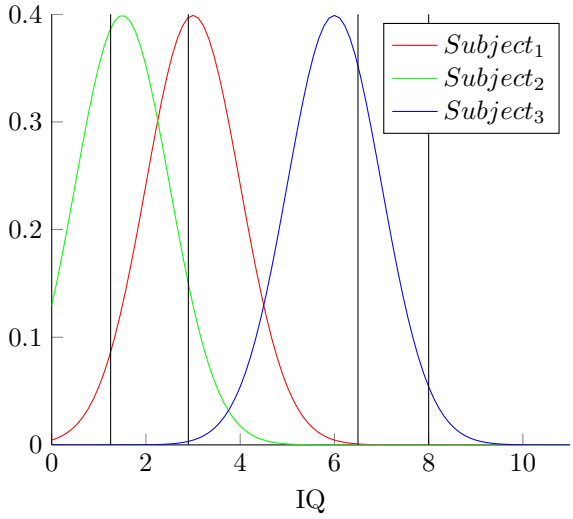


Figure 48: Performance of 3 subjects in an IQ-Test.

The test items mark thresholds similar to signal detection theory: you answer a question correctly and you are right of it, you fail you are left. Now it is possible to calculate simultaneously the position of the thresholds on the IQ-scale, as well as the IQ-scale itself (we do not discuss how to do this in detail).

What are the underlying assumptions of our "Measurement Model"?

Obviously we assume some kind of ordering between the different items. There is a fancy word for this:

28.2.1 Weak Stochastic Transitivity

If $\mu_i \geq \mu_j, \mu_j \geq \mu_k$ then $\mu_i \geq \mu_k$. In this case transitivity holds. We can rewrite this:

$$\underbrace{\mu_i - \mu_j \geq 0}_{d_{ij}}, \underbrace{\mu_j - \mu_k \geq 0}_{d_{jk}} \Rightarrow \underbrace{\mu_i - \mu_k \geq 0}_{d_{ik}}$$
$$\Leftrightarrow p_{ij} \geq \frac{1}{2}, p_{jk} \geq \frac{1}{2} \Rightarrow p_{ik} \geq \frac{1}{2}$$

So weak stochastic transitivity is about ordering of the different choices, but is less restrictive about the values. This constraint is exploited by:

28.2.2 Strong Stochastic Transitivity

If we know that choice i is preferred over choice j and j is chosen over k, the resulting choice probability of i over k can not be less than the maximum of the single probabilities:

$$d_{ij} \geq 0, d_{jk} \geq 0 \Rightarrow d_{ik} = d_{ij} + d_{jk} \geq max(d_{ij}, d_{jk})$$
$$p_{ij} \geq \frac{1}{2}, p_{jk} \geq \frac{1}{2} \Rightarrow p_{ik} \geq max(p_{ij}, p_{jk})$$

Strong stochastic transitivity may be violated in the case where the variances of the different choices differ:

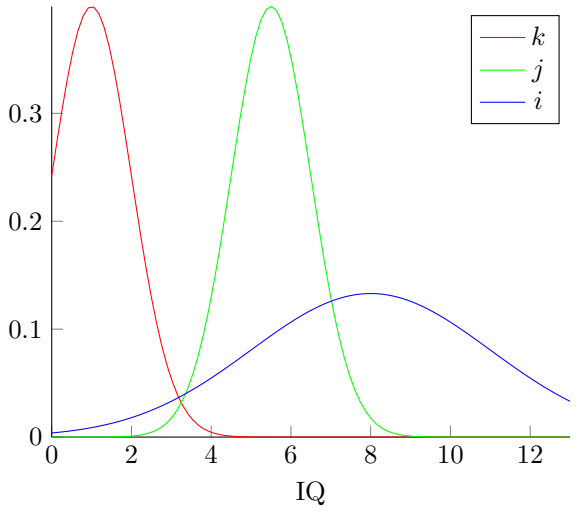


Figure 49: A problem for strong stochastic probability

In this case $p_{ij} = 0.6$, $p_{jk} = 0.95$ but $p_{ik} = 0.85$ which is less than the maximum of the other probabilities 0.95! But we can also think of different examples where the whole concept of transitivity is questionable.

Is transitivity reasonable?

Assume a situation of three chess players A, B, and C. A more often beats B than losing against him, B more often beats C but C also more often beats A than losing against her. This scenario is visualized in figure 50. If we try to find Gaussian distributions for each player's utility it gets clear quite quickly, that we will fail, since we don't know "where" to put the μ for the last competitor: left or right of the other two?
It seems our measurement model is not appropriate for this kind of situation. But how can we decide whether our model is appropriate or not?

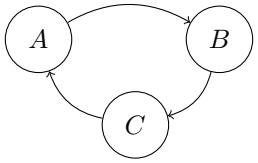


Figure 50: Three chessplayers dominate each other in a cyclic way.

28.2.3 Restle's Choice Model

Another model that does not assume one dimensional scaling for choices was proposed by Restle. In 1961 he showed with a gedankenexperiment why our previous model is maybe not that accurate and intuitive as it first sounded. Consider the following setting: we would like to go on holiday and have the following alternatives:

- Rome
- Paris
- Paris + an apple

If we are indifferent between Paris and Rome ($p_{21} = p_{12} = \frac{1}{2}$) – what is p_{32}? Actually the one extra apple should not change our basic decision between Paris and Rome, so $p_{32} \approx \frac{1}{2}$, but strong stochastic transitivity would predict $p_{32} = 1$! So if our previous model would be right every travel agent could simply persuade you to book any vacation by simply having an apple at hand.
Restle proposes a binary feature vector that describes each option. Ours look like $(Paris, Rome, Apple)$:

$$\begin{aligned} &Rome: && f_1 = (0,1,0) \\ &Paris: && f_2 = (1,0,0) \\ &Paris + an\ apple: && f_3 = (1,0,1) \end{aligned}$$

Each feature has a utility μ_1, μ_2, μ_3 and the probability to choose one over the other is dependent on the sum of all the features one choice has compared to the other. In the following formula m is the dimension of the feature vector.

$$\begin{aligned} p_{ij} &\propto \sum_{k=1}^{m} \mu_k (f_{ik} - f_{ik} \cdot f_{jk}) = u_{ij} \\ p_{ij} &= \frac{u_{ij}}{u_{ij} + u_{ji}} \end{aligned}$$

Let's calculate the probability with which we choose Rome over Paris:

$$\begin{aligned} p_{12} &= \frac{\sum_{k=1}^{3} \mu_k (f_{1k} - f_{1k} \cdot f_{2k})}{u_{12} + u_{21}} \\ &= \frac{\mu_2}{\mu_2 + \mu_1} = \frac{1}{2} \text{ , if } \mu_1 = \mu_2 \end{aligned}$$

Now we calculate the interesting choice Paris+Apple over Rome:

$$\begin{aligned} p_{31} &= \frac{\sum_{k=1}^{3} \mu_k (f_{3k} - f_{3k} \cdot f_{1k})}{u_{31} + u_{13}} \\ &= \frac{\mu_1 + \mu_3}{\mu_1 + \mu_3 + \mu_2} \approx \frac{1}{2} \text{ , since } \mu_3 << \mu_2 \end{aligned}$$

So Restle's model can predict this scenario much better.

PMPC Tutorial Sheet 8

1. Data are collected in a paired-comparison experiment. On each trial subjects have to say which of two options they prefer. Assume that the utility of each option i in a set of n options varies randomly on each trial and that the subject picks the options with the higher utility. The utility for each option has a Normal distribution with mean μ_i and standard deviation 1. Let $d_{ij} := \mu_i - \mu_j$ be the difference in utility between the mean utilities of two options. The probability of choosing i over j is then[1]

$$p_{ij} = 1 - \Phi(0; d_{ij}, \sqrt{2}) = \Phi(d_{ij}/\sqrt{2}; 0, 1).$$

   Let there be three options with mean utilities $\mu_1 = 1$, $\mu_2 = 1.5$, $\mu_3 = 2.5$. Calculate the choice probabilities for all possible pairs of options and put them into a matrix. How often will it happen that a subject prefers 2 over 1 and 3 over 2 but also 1 over 3?

2. Rumelhart and Greeno [2] conducted a paired-comparison experiment to test various choice models. They considered all pairs of 9 celebrities: Lyndon Johnson, Harold Wilson, Charles DeGaulle, Johnny Unitas, Carl Yastrzemski, A. J. Foyt, Brigitte Bardot, Elizabeth Taylor, Sophia Loren. The first three were politicians, the second three athletes, and the remaining three actresses. They presented each of 234 subjects with hypothetical choices between all possible pairings of the nine, asking for each pair with whom they would rather spend an hour of conversation with. The data are shown in table 1 (that you can also download from Stud.IP). Rows were chosen over columns. Hence, the first row shows how often Johnson was chosen over Wilson, over DeGaulle, and so on. A celebrity was, of course, never paired with itself and data on the diagonal

[1] The cumulative normal distribution Φ is called `normcdf` in Matlab and Octave and it's inverse is `norminv`.

	LJ	HW	CD	JH	CY	AF	BB	ET	SL
LJ	117	159	163	175	183	179	173	160	142
HW	75	117	138	164	172	160	156	122	122
CD	71	96	117	145	157	138	140	122	120
JH	59	70	89	117	176	115	124	86	61
CY	51	62	77	58	117	77	95	72	61
AF	55	74	96	119	157	117	134	92	71
BB	61	78	94	110	139	100	117	67	48
ET	74	112	112	148	162	142	167	117	87
SL	92	112	114	173	173	163	186	147	117

Table 1: Data from a paired comparison experiment [2]. The data show how often the rows were preferred over the columns for 234 subjects that did all pairwise comparisons each. The diagonal was simply filled up with $\frac{234}{2}$ for convenience.

of this matrix were actually not collected. I filled in the theoretical value of $\frac{1}{2} \cdot 234$ to simplify the analysis. Apply Thurstonian scaling to these data. Who is the most popular celebrity in this set, who is the least popular? Make a plot of the observed relative frequency of the choices as a function of the fitted probabilities.

3. The triple HW (2), SL (9), ET (8) violates strong stochastic transitivity. $p_{29} \geq \frac{1}{2}$ and $p_{98} \geq \frac{1}{2}$ but $p_{28} < p_{98}$. Explain what this means in your own words. Identify all triples in the data that violate strong stochastic transitivity. Why are there violations of strong stochastic transitivity in these data?

4. Bonus question: Write a function `restle.m` in Matlab or Octave that implements Restle's choice model [1]. The function gets the log of the utilities for each feature (the log is important so that the inputs can be positive and negative), a feature matrix, and a choice matrix (like the one in table 1) as inputs and returns the negative log likelihood and the theoretical choice probabilities as outputs.[2] What is an appropriate feature matrix for the data in table 1? Use the function `fminunc` to find the values for the feature utilities that minimize the negative log likelihood. For a function `f` of an n-dimensional vector `x` the following call finds the `x` that minimizes `f`: `x = fminunc(f, zeros(1,n))`[3]. In this example the function `fminunc` starts a numerical search for the best value with an initial value of all zeros. In order to define an appropriate `f` as an anonymous function that can be used to minimize the negative log likelihood the following line will be useful: `f = @(logutilities) restle(logutilities, featureMatrix, data)`. `f` is now a variable that refers to a function that takes `logutilites` as input but fixes the feature matrix and the data. You can use such an `f` as input to `fminunc`. Make a plot of the observed relative frequency of the choices as a function of the fitted probabilities.

References

[1] F. Restle. *Psychology of Judgment and Choice: A Theoretical Essay.* John Wiley & Sons, 1961.

[2] D. L. Rumelhart and J. G. Greeno. Similarity between stimuli: An experimental test of the luce and restle choice models. *Journal of Mathematical Psychology*, 8:370–381, 1971.

[2] Or download the function `restle.m` from Stud.IP; If you do implement it yourself, make sure that the diagonal of the choice matrix contains not `NaN` but $\frac{1}{2}$.

[3] `fminunc` is for Octave or Matlab if you have the optimization toolbox. Normally, in Matlab you will have to use `fminsearch` instead. Use the following code: `o = optimset; o.MaxFunEvals = 100000; o.MaxIter = 100000; o.TolX = 10^-8; o.TolFun = 10^-8; x = fminsearch(f,zeros(1,n),o);`

30 Solution 8: Choice Models I+II *2014-07-11*

Exercise 1

`MATLAB` code for exercise 1:

```
% create the three mu's
mu = [1 1.5 2.5]';

% calculate their distance matrix
M = repmat(mu, 1, 3);
D = M - M'
% result:
%D =      0   -0.5000   -1.5000
%    0.5000         0   -1.0000
%    1.5000    1.0000         0

% calculate the probability for each choice
P = normcdf(D / sqrt(2), 0, 1)
% result:
%P = 0.5000    0.3618    0.1444
%    0.6382    0.5000    0.2398
%    0.8556    0.7602    0.5000

probability = P(2,1) * P(3,2) * P(1,3)
% result:
%probability = 0.0701
```

This code simply initializes the three μ and creates the distance matrix between them. Then it simply calculates the probability for each event (i.e. how high is the probability that a subject chooses 1 over 2?). The probability for the case that a subject chooses 2 over 1, 3 over 2, and 1 over 3 is then just the multiplication of these three individual events: $p = P(2,1) \cdot P(3,2) \cdot P(1,3) \approx 0.07$.

So now we know the probability for one single subject. How is the probability for n subjects?
To answer that question we continue our script from before and add a simulation for the case that 2 is chosen over 1, 3 is chosen over 2, and 1 is chosen over three, to see how often the strong stochastic transitivity gets violated.

Reminder: Strong Stochastic Transitivity

$$\begin{aligned} \texttt{if } & p_{jk} \geq 0.5 \ \& \ p_{ij} \geq 0.5 \\ \texttt{then } & p_{ik} \geq \max\left(p_{jk}, p_{ij}\right) \end{aligned}$$

The following `MATLAB` code simulates the choices with the probabilities taken from the code above (Matrix P) for m subjects and counts the violations of strong stochastic transitivity. In the end it just divides the number of violations by the number of simulated trials to come up with a probability.

```
% probabilities for each event
P = normcdf(( repmat([1 1.5 2.5]', 1, 3)...
             -repmat([1 1.5 2.5]', 1, 3)') / sqrt(2), 0, 1);

% play around with m and n to see that many subjects is better than many
% trials with fewer subjects
m = 10;   % number of subjects
n = 1000; % number of trials with m subjects
violation = zeros(n, 1); % we also save when the violations occured

for t = 1:n % for each trial
    q21 = mean(rand(1, m) < P(2, 1)); % generate randomly how many persons
    q32 = mean(rand(1, m) < P(3, 2)); % prefer a choice over the other and
    q31 = mean(rand(1, m) < P(3, 1)); % take the mean over this boolean
                                      % vector to achieve a probability
```

```
    if q21 >= .5 && q32 >= .5          % check premise for s-s-t
        if q31 >= max(q21, q32)        % if the conclusion does hold: noop
        else                           % else
            violation(t) = 1;          % save the violation
        end
    end
end
probability = sum(violation)/n
% result: (for example)
%probability = 0.226
```

We can also invert this process (at least parts of it). Assuming we have the q values for a matrix (e.g. by asking random people whether they prefer pizza tonno over pizza salami, pizza tonno over pizza margarita, etc.) we can calculate the μs, i.e. the utilities for those choices. However, we are not able to find the x-shift, so if we have two μs, e.g. $\mu_1 = 1.5$ and $\mu_2 = 2$, the result of our calculations will turn out to be $\mu_1 = 0$ and $\mu_2 = 0.5$. An example `MATLAB` code can be found in the appendix on page I.

Note: *Not directly covered in the lecture.*

Exercises 2-4

The commented `MATLAB` code which was provided as a solution can be found in the appendix on page I. It employs the functions `thurstone` (page III) and `restle` (page III) which are solutions to the exercises two to four.
The `choiceplot` function (page IV) makes a plot which shows how good a model fits the data. The more data points are inside the red borders, the better the data is fitted by the model.
Please refer to the respective pages in the appendix for further information.

31 Everyday Predictions *2014-07-14*

Galton (1907) went to a fair and observed a simple guessing game. There was a bull displayed and you could get a price if you guessed the right weight of it. Galton had a look at all guesses and found, that the knowledge of the mass could display the real value for the bull's weight (1198 lbs) pretty good as can be seen in the following graph:

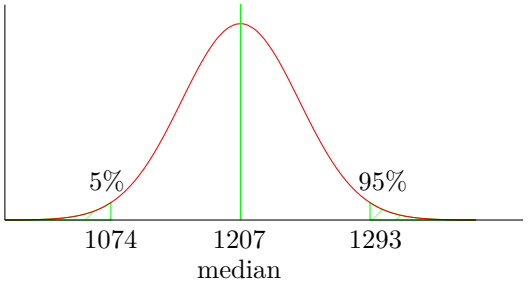


Figure 51: A popular fair game: guess the bull's weight. The *wisdom of the crowds* leads to quite a good solution.

In 2006 Griffiths & Tenenbaum did a "simple" Bayesian inference with "real-world" priors and only one data point. An example for this is the distribution of age of death of men in Germany.

Note: *Another famous example: the German Tank Problem.*

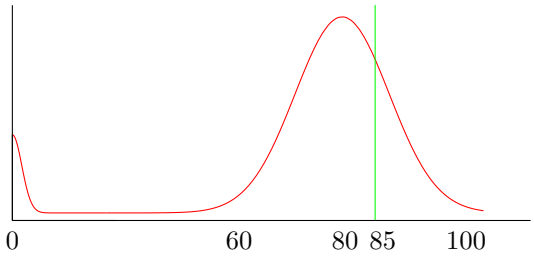


Figure 52: Example distribution of men's age of death in Germany.

Assume the following two scenarios.

- You meet someone who is 25 years old. When will he die? In this case you have to go with your prior, which then is your posterior.
- You meet someone who is 85 years old. When will he die? Every age below 85 is now not possible any more, you have to update your prior.

When X is the total value and Y the observed value, P(X) is our prior. It follows that:

$$P(Y = y|X = x) = \frac{1}{X} \cdot I(y \leq x) \qquad \rightarrow I \; is \; 1 \; if \; y \leq x; 0 \; otherwise$$

$$\begin{aligned} P(X = x|Y = y) &= \frac{P(Y = y|X = x) \cdot P(X = x)}{P(Y = y)} \\ &= \frac{\frac{1}{X} I(y \leq x) p(X = x)}{\int_{-\infty}^{\infty} \frac{1}{X} I(y \leq x) p(X = x) dx} \\ &= \frac{\frac{1}{X} I(y \leq x) p(X = x)}{\int_{y}^{\infty} \frac{p(X=x)}{X} dx} \end{aligned}$$

PMPC Tutorial Sheet 9

1. Everyday prediction [2]. If you heard a member of the House of Representatives had served for Y years, what would you predict his total term X in the House would be? Assume that X has the following prior distribution

$$p(X = x) = \beta^{-2} \cdot x \cdot e^{-x/\beta}$$

(this is a special case of the Erlang distribution which itself is a special case of the Gamma distribution). Make a plot of this distribution for different values of β. Assume that the distribution for Y conditional on X is

$$p(Y = y \mid X = x) = \begin{cases} \frac{1}{x} & \text{for } 0 \leq y \leq x \\ 0 & \text{otherwise} \end{cases}.$$

What is the posterior distribution for X given Y?

2. Read the paper by Griffiths and Tenenbaum (you can find it on Stud.IP). We've seen several cases where human subjects failed to respond according to the rules of probability theory, e.g. the conjunction fallacy or base rate neglect. Why do Griffiths and Tenenbaum find behavior that is in accordance with probability theory?

3. Is each subject in Griffiths' and Tenenbaum's study rational or is this a case of "wisdom of the crowd"? [1, 3]

References

[1] F. Galton. Vox populi. *Nature*, 75(1949):450–451, 1907.

[2] T. L. Griffiths and J. B. Tenenbaum. Optimal predictions in everyday cognition. *Psychological Science*, 17(9):767–773, 2006.

[3] M. C. Mozer, H. Pashler, and H. Homaei. Optimal preditions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, 32:1133–1147, 2008.

33 Solution 9: Everyday Predictions *2014-07-18*

Exercise 1

First we plot the prior distribution for some β to get a feeling for it. We can easily see that a smaller β causes a very high peak, while bigger β caus the peak to flatten and move to the right.

After we got a feeling for the prior distribution, let's calculate the posterior distribution to make a guess, how long the total term of the member of the House of Representatives will be.
Let X be the total time and Y the observed value, i.e. the time we heard.
Then we can use Bayes' rule to come up with a solution. We use the distribution from the sheet for $p(Y = y|X = x) = \frac{1}{x}\mathcal{I}(y \leq x)$ as the likelihood and also plug in the given prior distribution $\beta^{-2}xe^{-\frac{x}{\beta}}$. Note that $\mathcal{I}(y \leq x)$ is the indication function, i.e. it is 1 if $y \leq x$ and 0 otherwise.

$$\begin{aligned} p(X = x|Y = y) &= \frac{p(Y = y|X = x)p(X = x)}{p(Y = y)} \\ &= \frac{\frac{1}{x}\mathcal{I}(y \leq x)\beta^{-2}xe^{-\frac{x}{\beta}}}{p(Y = y)} \\ &= \frac{\mathcal{I}(y \leq x)\beta^{-2}e^{-\frac{x}{\beta}}}{p(Y = y)} \end{aligned}$$

We can derive $p(Y = y)$ (normalization) by taking the integral of $e^{-\frac{x}{\beta}}\mathcal{I}(y \leq x)$.

$$\begin{aligned} p(Y = y) &= \int_0^\infty e^{-\frac{x}{\beta}}\mathcal{I}(y \leq x)dx \\ &= \int_y^\infty e^{-\frac{x}{\beta}}dx = \left[-\beta e^{-\frac{x}{\beta}}\right]_y^\infty \\ &= -\beta \cdot 0 + \beta e^{-\frac{y}{\beta}} \end{aligned}$$

By employing $p(Y = y)$ in the formula, we finally come up with our posterior distribution:

$$\begin{aligned} p(X = x|Y = y) &= \frac{\mathcal{I}(y \leq x)\beta^{-2}e^{-\frac{x}{\beta}}}{p(Y = y)} \\ &= \frac{\mathcal{I}(y \leq x)e^{-\frac{x}{\beta}}}{\beta^3 e^{-\frac{y}{\beta}}} \\ &= \mathcal{I}(y \leq x)\frac{e^{-\frac{x}{\beta}}}{\beta^3 e^{-\frac{y}{\beta}}} \end{aligned}$$

The plot of the posterior distribution is as follows (for $y = 2$ and $\beta = 1$):

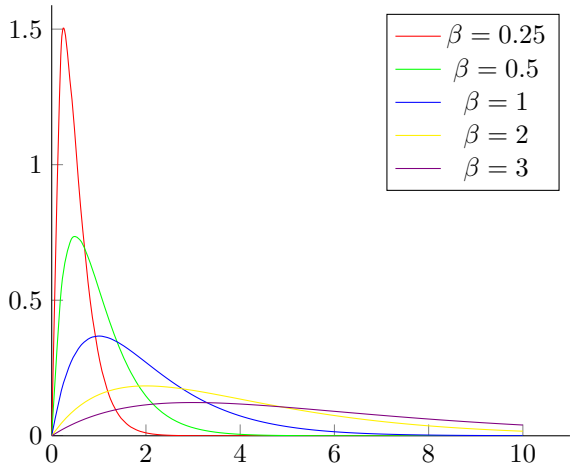


Figure 53: Special case of the Erlang distribution $\beta^{-2}xe^{-\frac{x}{\beta}}$ with different β.

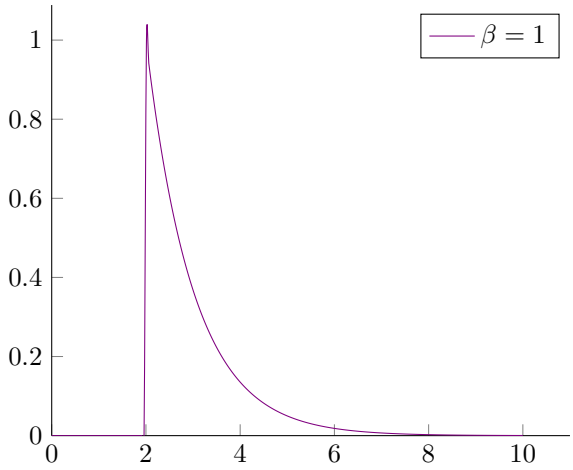


Figure 54: Final posterior distribution.

The idea for our prediction is now to take the mean of the "remaining" distribution, i.e. the mean of the part which is greater than y. The mean is where the integral over our distribution is 0.5.

$$
\begin{aligned}
\frac{1}{2} &\overset{!}{=} \int_y^{\hat{x}} \frac{1}{\beta^3} e^{-\frac{x-y}{\beta}} dx = \frac{1}{\beta^3} \int_0^{\hat{x}-y} e^{-\frac{x}{\beta}} dx \\
&= \frac{1}{\beta^3} \left[-\beta e^{-\frac{x}{\beta}} \right]_0^{\hat{x}-y} = \frac{1}{\beta^3} \left(-\beta e^{-\frac{\hat{x}-y}{\beta}} + \beta \right) \\
&= \frac{1}{\beta^2} (1 - e^{-\frac{\hat{x}-y}{\beta}}) \overset{!}{=} \frac{1}{2} \\
e^{-\frac{\hat{x}-y}{\beta}} &= 1 - \frac{\beta^2}{2} \\
-\frac{\hat{x}-y}{\beta} &= \log\left(1 - \frac{\beta^2}{2}\right) \\
-\hat{x} + y &= \beta \log\left(1 - \frac{\beta^2}{2}\right) \\
y &= \beta \log\left(1 - \frac{\beta^2}{2}\right) + \hat{x} \\
\hat{x} &= y - \beta \log\left(1 - \frac{\beta^2}{2}\right) \\
\hat{x} &= \underbrace{y}_{\text{current}} + \underbrace{\beta \log\left(\frac{1}{1 - \frac{\beta^2}{2}}\right)}_{\text{additional}}
\end{aligned}
$$

So our best guest for how long the member of the House of Representatives will stay in congress is $\beta \log \frac{2}{\beta^2}$.

Exercises 2 and 3

We don't have these solutions.

34 Final Exam Questions

There are 8 questions and each question is worth 4 points. You only have to answer 6 out of the 8 questions. Hence, the maximum score is 24 points. If you answer more than 6 questions the answers with the lowest scores will be discarded. You have from 8:00 to 10:00 to work on your responses. Please respond in full sentences.

Exercise 1

To diagnose colorectal cancer the hemoccult test is conducted to detect occult blood in the stool. For symptom-free people over 50 years old who participate in screening using the heoccult test the following information is available: 0.3% have colorectal cancer. Of these people with cancer, half of them will have a positie test result (The hemoccult test is not very sensitive, its hit-rate is only 50%). Of the people without cancer, 3% will still have a positive hemoccult test (the false-alarm-rate is about 3%). What is the posterior probability of having colorectal cancer for a symptom-free person over 50 with a positive test result?

Exercise 2

A statement X in a knowledge test can be true ($X = 1$) or false ($X = 0$). Say, you answered that your probability is q for $X = 1$. You will be scored using the following loss function

$$L(X, q) = \frac{1}{2}q^2 - Xq,$$

Your true belief in the statement is p. Show that L is a proper scoring rule.

Exercise 3

X is a random variable with probability density function

$$f(x) = \begin{cases} 1 \text{ if } 0 \leq x \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

Define a new random variable $Y = X^2$. What is the cumulative distribution function for Y? What is the probability density function for Y?

Exercise 4

The exponential distribution has the following probability density function

$$p(T_i = t_i|\lambda) = \lambda e^{-\lambda t_i}$$

for positive values of t_i, and strictly positive values of λ. Assume you have seen n independent samples $T_1, \ldots, T_n$ from an exponential distribution. What is the joint distribution $p(T_1 = t_1, \ldots, T_n = t_n|\lambda)$? What is the maximum likelihood estimate for λ after you've seen n samples.

Exercise 5

Your are a subject in a reaction time experiment. Your task is to press a button as fast as possible when a clearly visible light flashes up. The time point T of the next flash in the experiment depends only on the time point of the last flash. Without loss of generality we can assume that the last flash happened at time zero. The distribution for the next flash at time T is an exponential distribution with parameter λ

$$p(T = t|\lambda) = \lambda e^{-\lambda t}$$

It is now time $t^\star$ after the last flash and the next flash hasn't happened yet. What is the new distribution of T given that the next flash has not happened until time $t^\star$, $p(T = t|T > t^\star, \lambda)$? Why is it clever of the experimenter to use the exponential distribution?

Exercise 6

In a detection experiment the distribution for a trial without a signal ($signal = no$) is given by

$$p(X = x|signal = no) = \begin{cases} 1 \text{ if } 0 \leq x \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

The distribution for the trials where there is a signal ($signal = yes$) is given by

$$p(X = x|signal = yes) = \begin{cases} 1 \text{ if } 0.5 \leq x \leq 1.5 \\ 0 \text{ otherwise} \end{cases}$$

Assume a subject responds with *yes* if the observed value X is larger than a criterion c and *no* otherwise. How does the hit rate change as a function of c? How does the false alarm rate change as a function of c? Make a plot of the ROC curve.

Exercise 7

Explain why a high-threshold theory of detection is not tenable. Name and explain four findings that together speak against high-threshold theory.

Exercise 8

Let the probability p_{ij} that an option i is chosen over an option j in a paired comparison experiment be a function of the utilities μ_i and μ_j:

$$p_{ij} = F(\mu_i - \mu_j)$$

where F is a strictly increasing cumulative distribution function with $F(0) = \frac{1}{2}$. Choice models that can be expressed in this way are said to satisfy *simplescalability*. Show that simple scalability implies strong stochastic transitivity.

35 Final Exam Solutions *2014-07-25*

36 Appendix

36.1 MATLAB codes

Tutorial Sheet 8, Exercise 1: Find μ

This code is an example for the claim on page 91. It calculates the μs from given qs in a paired-comparison experiment.

```
% prepare q with "known" values
q = [0.5 0.75 0.345; 0.25 0.5 0.45; 0.655 0.55 0.5];
% calculate the inverse of the normal distribution and normalize it
P = norminv(q, 0, 1) * sqrt(2)
% extract the mu's
mu = P(:,1)'
```

Tutorial Sheet 8, Exercises 2-4: Choice Models

This code is the solution for tutorial sheet 8, exercises 2-5. It also demonstrates with a plot, how well the two compared models fit the data. The code was referenced on page 91.
To use it you need the files `thurstone.m` (page III), `restle.m` (page III), and `choiceplot.m` (page IV) as well.

```
load celebrities.txt
c = celebrities;

% empirical probabilities
q = c./234;

% estimates for differences
d = sqrt(2) * norminv(q,0,1);

% estimates for the means
mu = mean(d,2);

% fitted probabilities:
p = normcdf((repmat(mu,1,9)-repmat(mu',9,1))/sqrt(2),0,1);

% plot data versus predictions
figure(1)
choiceplot(p,q,234)
title('Thurstone with least squares')

% now do the same for a ml fit
c = celebrities;
negloglik = @(mu) thurstone(mu,c);
muml = zeros(9,1);

% OPTIMIZATION IN OCTAVE:
% muml = fminunc(negloglik,muml);

% OPTIMIZATION IN MATLAB:
o = optimset; % make sure we have sufficient precision in matlab
o.MaxFunEvals = 100000;
o.MaxIter = 100000;
o.TolX = 10^-8;
o.TolFun = 10^-8;
[muml,fval,exitflag,output] = fminsearch(negloglik,muml);

% plot data versus predictions
figure(2)
choiceplot(p,q,234)
title('Thurstone ML fit')

% compare the estimates of least square and ml fit
```

```
muml = muml-mean(muml); % subtract mean as for lsq fit
[negloglik,p] = thurstone(mu,c);
[negloglikml,p] = thurstone(muml,c);
fprintf('\n')
fprintf('Thurstonian Scaling\n')
fprintf('-------------------------------------------------------------------\n')
fprintf('      LJ1   HW2   CD3   JH4   CY5   AF6   BB7   ET8   SL9 NLL\n')
fprintf('lsq ')
fprintf('%+5.2f ', mu')
fprintf('%7.2f\n', negloglik)
fprintf('ml  ')
fprintf('%+5.2f ', muml')
fprintf('%7.2f\n', negloglikml)
fprintf('-------------------------------------------------------------------\n\n')

% now identify the cases that violate strong stochastic transitivity. If there
% are many violations this will be an indication that the assumption of s.s.t.
% does not hold for the data and therefore fitting a Thurstonian model was not
% the right thing to do.
fprintf('\n')
fprintf('Violations of strong stochastic transitivity\n')
fprintf('-------------------------------------------------------------------\n')
fprintf('a b c | q(a,b)         q(b,c)         q(a,c)\n')
fprintf('-------------------------------------------------------------------\n')
for a = 1:9
  for b = 1:9
    for c = 1:9
       if (q(a,b)>0.5) && (q(b,c)>0.5)
          if not(q(a,c)>q(a,b) & q(a,c)>q(b,c))
             % check whether the violation is significant:
             % assuming that q(a,c) was exact we can check how far the
             % other two are away from q(a,c) and whether one of them is
             % significantly bigger. This is just a quick hack but should
             % give you some idea as to which differences are relatively
             % big:
             s = sqrt(q(a,c)*(1-q(a,c))/234);
             d1 = (q(a,b)-q(a,c))/s;
             d2 = (q(b,c)-q(a,c))/s;
             sig = '';
             if (d1 > norminv(0.95,0,1)) || (d2 > norminv(0.95,0,1))
                 sig = '*';
                 if (d1 > norminv(0.99,0,1)) || (d2 > norminv(0.99,0,1))
                     sig = '**';
                 end
         end
             fprintf(['%d %d %d | '...
                      '  %.2f (%+1.2f)   %.2f (%+1.2f)   %.2f %s\n'],...
                     a,b,c,q(a,b),d1,q(b,c),d2,q(a,c),sig);
          end
       end
    end
  end
end
fprintf('-------------------------------------------------------------------\n\n')

% fit Restle's model to the same data but add features for 3 clusters
F = [eye(9); 1 1 1 0 0 0 0 0 0; 0 0 0 1 1 1 0 0 0; 0 0 0 0 0 0 1 1 1]';
logmu = ones(1,12)'; % as mu has to be positive, we optimize the log of mu
c = celebrities;
negloglik = @(logmu) restle(logmu,F,c);

% OPTIMIZATION IN OCTAVE
% logmu = fminunc(negloglik,logmu);

% OPTIMIZATION IN MATLAB
o = optimset;
o.MaxFunEvals = 100000;
o.MaxIter = 100000;
o.TolX = 10^-8;
```

```
o.TolFun = 10^-8;
[logmu,fval,exitflag,output] = fminsearch(negloglik,logmu,o);

[negloglik,p] = restle(logmu,F,c);
mu = exp(logmu);      % the actual mu's are only positive
mu = mu./sum(mu)*100; % and the scale is arbitrary

fprintf('\n')
fprintf('Restle''s Choice Model\n')
fprintf('-------------------------------------------------------------------\n')
fprintf(' LJ1  HW2  CD3  JH4  CY5  AF6  BB7  ET8  SL9 POLI SPOR ACTO NLL\n')
fprintf('%4.1f ', mu')
fprintf('%4.0f\n', negloglik)
fprintf('-------------------------------------------------------------------\n')

% now make a plot for Restle's model
figure(3)
choiceplot(p,q,234)
title('Restle')
```

Tutorial Sheet 8, Exercises 2-4: Thurstone Scaling

This function implements Thurstone's choice model, *Thurstone scaling*. Used on page 91.

```
function [negloglik p] = thurstone(mu,k)
% [negloglik p] = thurstone(mu,k)
%
% returns the negative log likelihood and the choice probabilities for
% feature utilities mu. k is the number of choices for each pair where rows
% are chosen over columns.
%

% do a few checks, mu is a column vector
n = length(mu); % number of objects
if all(size(mu)==[1,n])
   mu = mu';
end
if not(all(size(k)==[n,n]))
   error('k has to be n by n')
end

% calculate the difference in mean utility for each pair
d = (repmat(mu,1,n)-repmat(mu',n,1));

% and translate this to choice probabilities
p = normcdf(d,0,sqrt(2));

% finally, calculate the negative log likelihood
L = k .* log(p);
negloglik = -sum(L(:));
```

Tutorial Sheet 8, Exercises 2-4: Restle's Choice Model

This function implements Restle's choice model. Used on page 91.

```
function [negloglik p] = restle(logmu,F,k)
% [negloglik p] = restle(logmu,F,k)
%
% returns the negative log likelihood and the choice probabilities for
% log feature utilities logmu and feature matrix F. F has objects in the rows
% and features as columns. k is the number of choices for each pair where rows
% are chosen over columns.
%
```

```
n = size(F,1); % number of objects
m = size(F,2); % number of features
p = zeros(n,n);% choice probabilities

mu = exp(logmu);

% do a few checks
if all(size(mu)==[1,m])
   mu = mu';
end
if not(all(size(mu)==[m,1]))
   error('mu and F have to agree in dimensions')
end
if not(all(size(k)==[n,n]))
   error('k has to be n by n')
end

% now calculate the nominator in Restle's model for each pair
for i = 1:n
  for j = 1:n
    p(i,j) = (F(i,:)-F(i,:).*F(j,:)) * mu;
  end
end

% and normalize choice probabilities so that p+p'=1
p = p ./ (p+p');
% by definition the diagonal contains only zeros
p(eye(n)>0) = 0.5;

% finally, calculate the negative log likelihood
L = k .* log(p);
negloglik = -sum(L(:));
```

Tutorial Sheet 8, Exercises 2-4: Restle's Choice Model

This function implements Restle's choice model. Used on page 91.

```
function [negloglik p] = restle(logmu,F,k)
% [negloglik p] = restle(logmu,F,k)
%
% returns the negative log likelihood and the choice probabilities for
% log feature utilities logmu and feature matrix F. F has objects in the rows
% and features as columns. k is the number of choices for each pair where rows
% are chosen over columns.
%

n = size(F,1); % number of objects
m = size(F,2); % number of features
p = zeros(n,n);% choice probabilities

mu = exp(logmu);

% do a few checks
if all(size(mu)==[1,m])
   mu = mu';
end
if not(all(size(mu)==[m,1]))
   error('mu and F have to agree in dimensions')
end
if not(all(size(k)==[n,n]))
   error('k has to be n by n')
end

% now calculate the nominator in Restle's model for each pair
for i = 1:n
  for j = 1:n
    p(i,j) = (F(i,:)-F(i,:).*F(j,:)) * mu;
```

```
  end
end

% and normalize choice probabilities so that p+p'=1
p = p ./ (p+p');
% by definition the diagonal contains only zeros
p(eye(n)>0) = 0.5;

% finally, calculate the negative log likelihood
L = k .* log(p);
negloglik = -sum(L(:));
```

36.2 Recommended Readings

[Asp10] W. Aspinall. A route to more tractable expert advice. *Nature*, 463:294f, 2010.

[Bri50] G. Brier. Verification of Forecasts expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1–3, 1950. http://docs.lib.noaa.gov/rescue/mwr/078/mwr-078-01-0001.pdf.

[Car83] C. Carrier. Notetaking Research – Implications for the Classroom. *Journal of Instructional Development*, 6(3):19–26, 1983.

[Coh94] J. Cohen. The Earth Is Round (p < .05). *American Psychologist*, 49(12):997–1003, 1994. http://ist-socrates.berkeley.edu/~maccoun/PP279_Cohen1.pdf.

[Ear92] J. Earman. *Bayes or bust?* MIT Press, 1992. http://joelvelasco.net/teaching/120/Earman_1992BayesOrBust.pdf.

[EB02] M. Ernst and M. Banks. Humans integrate visual and haptic information in a statisticaly optimal fashion. *Nature*, 415:429–433, 2002. http://www.cns.nyu.edu/~david/courses/perceptionGrad/Readings/ErnstBanks-Nature2002.pdf.

[G+07] G. Gigerenzer et al. Helping Doctors and Patients Make Sense of Health Statistics. *Psychological Science in the Public Interest*, 8(2):53–96, 2007. http://library.mpib-berlin.mpg.de/ft/gg/GG_Helping_2008.pdf.

[Gal07] F. Galton. Vox populi. *Nature*, 75:450f, 1907. http://galton.org/essays/1900-1911/galton-1907-vox-populi.pdf.

[GT06] T. Griffiths and J. Tenenbaum. Optimal Predictions in Everyday Cognition. *Psychological Science*, 17(9):767–773, 2006. http://web.mit.edu/cocosci/Papers/Griffiths-Tenenbaum-PsychSci06.pdf.

[Ioa05] J. Ioannides. Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8):696–701, 2005. http://www.plosmedicine.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pmed.0020124&representation=PDF.

[Jay03] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. http://f3.tiera.ru/2/P_Physics/PT_Thermodynamics,%20statistical%20physics/Jaynes%20E.T.%20Probability%20theory%20-%20the%20logic%20of%20science%20(book%20draft,%201998)(592s).pdf.

[Jef04] R. Jeffrey. *Subjective Probability: The Real Thing*. Cambridge University Press, 2004. http://www.princeton.edu/~bayesway/Book*.pdf.

[Kra99] D. Krantz. The Null Hypothesis Testing Controversy in Psychology. *Journal of the American Statistical Association*, 94(488):1372–1381, 1999. http://www.unt.edu/rss/class/mike/5030/articles/krantznhst.pdf.

[M+08] M. Mozer et al. Optimal Predictions in Everyday Cognition: The Wisdom of Individuals or Crowds? *Cognitive Science*, 32:1133–1147, 2008. http://csjarchive.cogsci.rpi.edu/proceedings/2008/pdfs/p1051.pdf.

[RG71] D. Rumelhart and J. Greeno. Similarity Between Stimuli: An Experimental Test of the Luce and Restle Choice Models. *Journal of Mathematical Psychology*, 8:370–381, 1971. http://deepblue.lib.umich.edu/bitstream/handle/2027.42/33598/0000102.pdf.

[S+00] J. Swets et al. Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest,*, 1(1):1–26, 2000. http://peterhancock.ucf.edu/Downloads/humanfactors_2/Advanced%20Signal%20Detection%20Lecture/Swets%20Dawes%20and%20Monahan%202000.pdf.

[SG01] P. Sedlmeier and G. Gigerenzer. Teaching Bayesian Reasoning in Less Than Two Hours. *Journal of Experimental Psychology*, 130(3):380–400, 2001. http://library.mpib-berlin.mpg.de/ft/ps/PS_Teaching_2001.pdf.

[Swe61] J. Swets. Is There a Sensory Threshold? *Science, New Series*, 134(3473):168–177, 1961. http://www.phon.ucl.ac.uk/courses/spsci/AUDL4007/threshold.pdf.

[TK83] A. Tversky and D. Kahneman. Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, 90(4):293–315, 1983. http://psy2.ucsd.edu/~mckenzie/TverskyKahneman1983PsychRev.pdf.

[Tra03] K. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2003. http://eml.berkeley.edu/books/train1201.pdf.

[YP07] A. Yonelinas and C. Parks. Receiver Operating Characteristics (ROCs) in Recognition Memory: A Review. *Psychological Bulletin*, 133(5):800–832, 2007. https://faculty.unlv.edu/cparks/PDFs/Yonelinas%20&%20Parks,%202007%20%5bROC%20review%5d.pdf.

Index

The end of the day.