Probabilistic Modeling of Perception and Cognition

Lecture Notes

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

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.

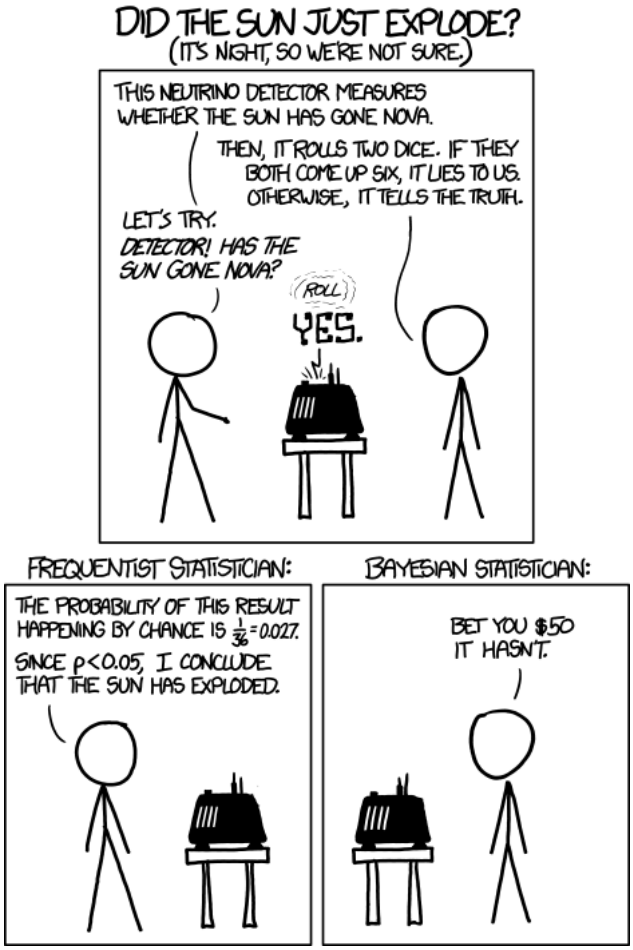


Figure 1: *Source: http://xkcd.com/1132/*

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

2 Probability Refresher II

2.1 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:

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

2.2 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

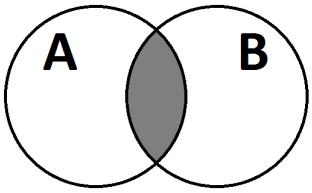


Figure 3: The intersection between two sets makes math a bit tricky

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 3).

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).

2.3 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.

2.4 Marginal and Conditional Probability

For the following examples please refer to table 1 (page 8).

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

3 Measuring Beliefs I

3.1 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.

3.2 What do you accept as a fair bet?

Figure 4: 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 $P(A) + P(B) < P(A \cup B)$.

3.3 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.*

4 Measuring Beliefs II

4.1 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.

4.1.1 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 5).

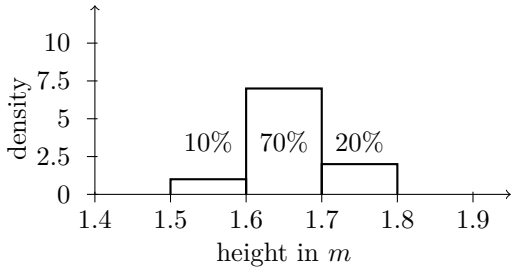


Figure 5: Histogram

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 6.

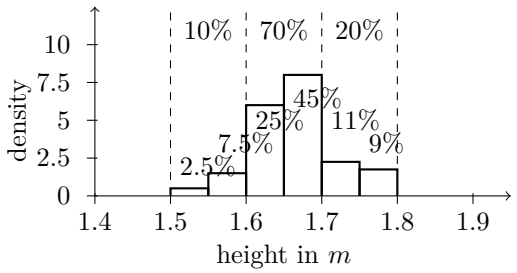


Figure 6: Histogram with higher resolution

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 7).

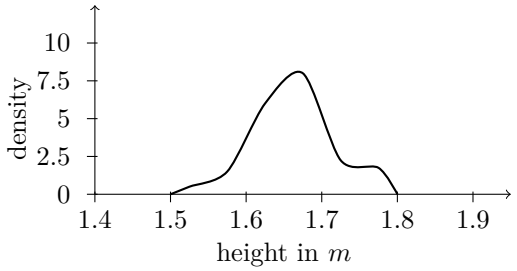


Figure 7: Probability density function

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 $[2, 3[$.

4.1.2 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 8).

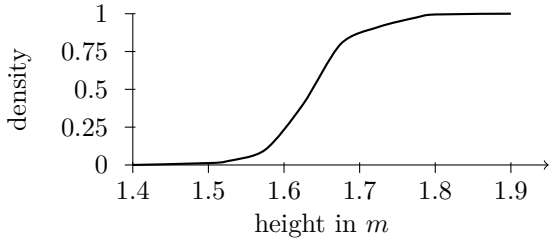


Figure 8: Cumulative density function

4.1.3 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 9) 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 9).

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

The area of the standard deviation ($\mu - \sigma \leq X \leq \mu + \sigma$) has a probability of approximately 68 % (see figure 10).

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

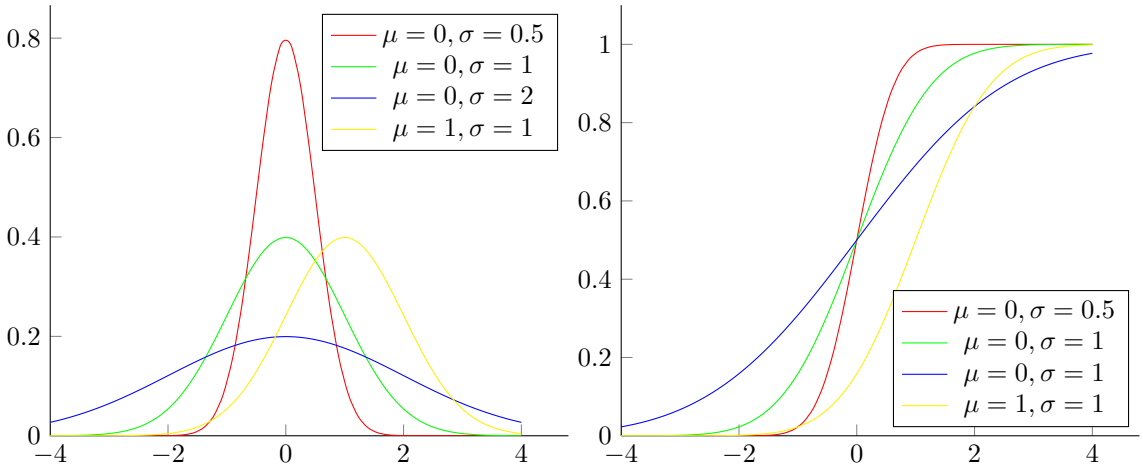


Figure 9: Left: Gaussian distributions. Right: Their corresponding integrals. Parameters: $\mu = 0$ and $\sigma = 0.5, 1, 2$, and $\mu = 1, \sigma = 1$.

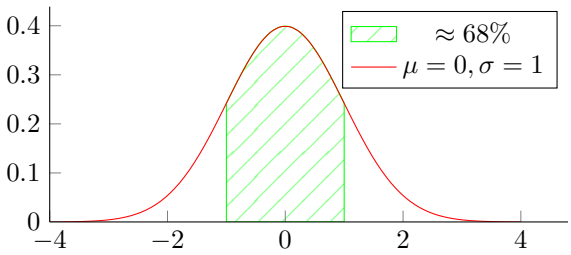


Figure 10: The area of the standard deviation yields approximately 68 %

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

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

4.2 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 17).

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.

5 Bayesian Inference Examples

This section is basically about Exercise 4 on the 4th Tutorial Sheet (see page **??**). 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?

5.1 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.

5.2 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.

6 Frequentist Inference Examples

6.1 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*.

6.2 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.*

6.3 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!*

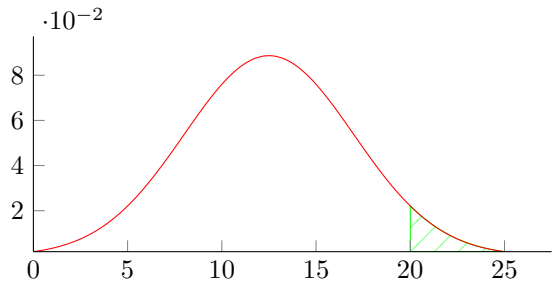
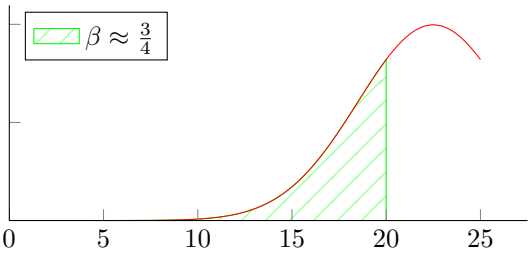


Figure 11: The corresponding Gaussian

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

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 $1 - \beta$.

7 Signal Detection Theory I

7.1 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.

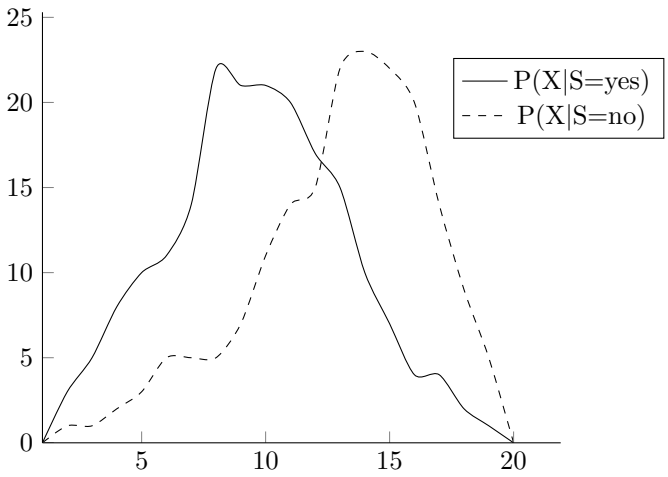
7.1.1 Examples

- 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:



7.1.2 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.

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

7.1.3 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:

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

7.1.4 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:

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

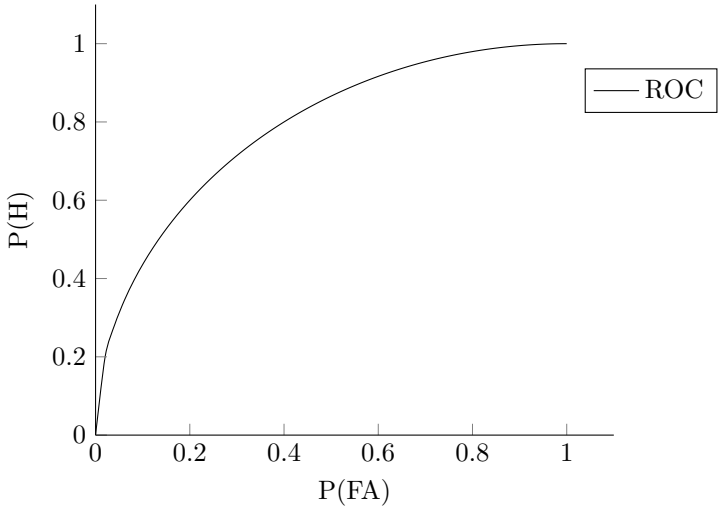
i.e. the best threshold is just in the middle of the two Gaussians, given the condition that both Gaussians have the same variance.

7.2 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.

8 Signal Detection Theory II

8.1 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.

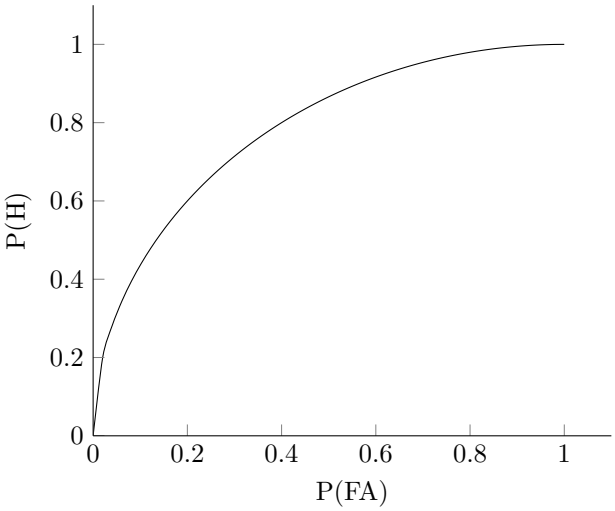


Figure 12: ROC-Curve

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:

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

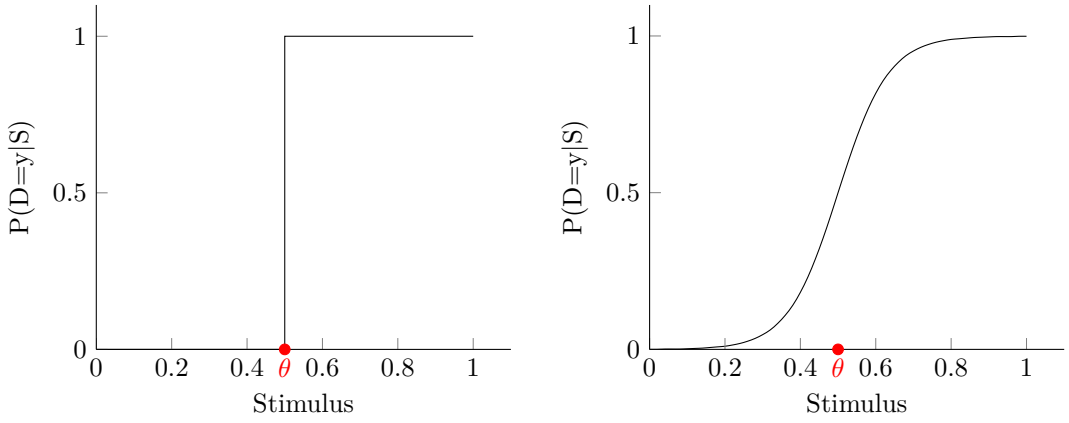
The last step is possible because the Gaussian is symmetric. Now we try to find a formula for d' as well.

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

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

8.2 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 13: 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!

8.3 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.

9 Signal Detection Theory III

9.1 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:

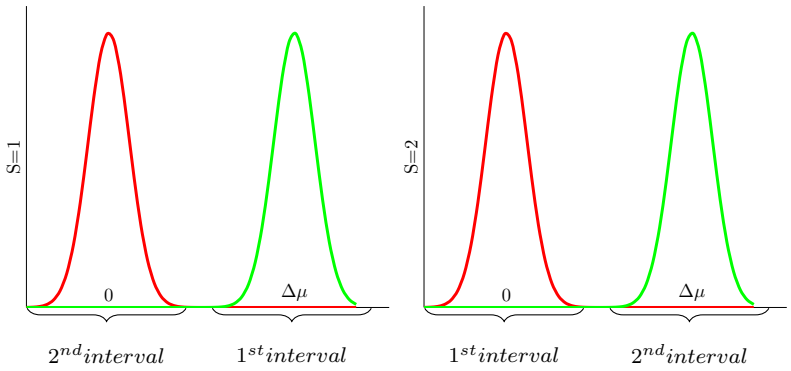


Figure 14: Mean-shifted Gaussian distributions

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:

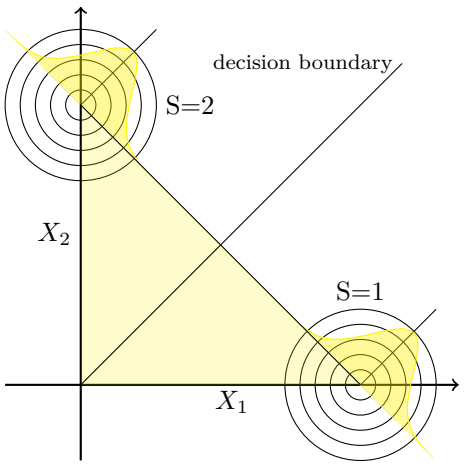


Figure 15: The distance between the two distributions is $\sqrt{2}\Delta\mu$.

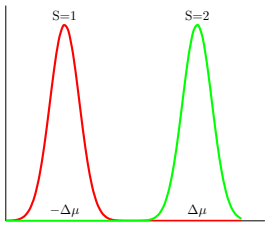
If we could discriminate perfectly our data points would lie on the x or y axis for each trial (depending on the interval). When we discussed why HT-Theory is wrong (8.3), we already stated that subjects perform better in 2AFC than in YN-Tasks. Now we see why: the distance of the two distributions is $\Delta\mu_{2AFC} = \sqrt{2}\Delta\mu$. 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.



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

9.2 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 16). Different than in the 2AFC the two distributions have the same mean – leading to the plot in figure 17. 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:

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

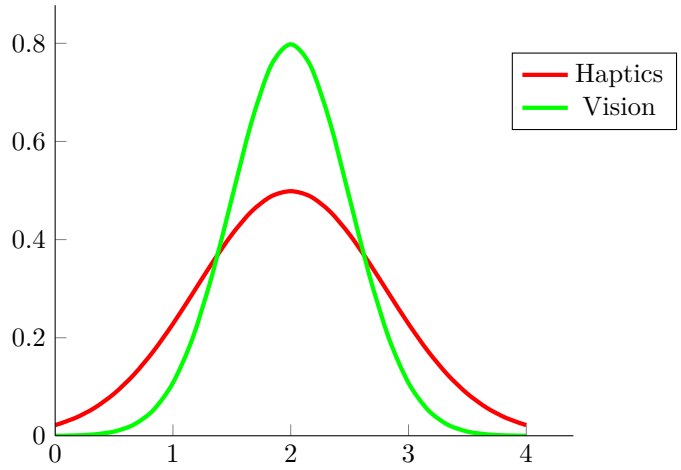


Figure 16: The visual system is more accurate than the haptic, thus the normal distribution of the visual system has a smaller variance.

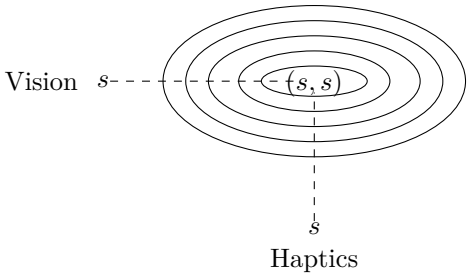


Figure 17: Vision and haptic systems' distributions have the same mean.

Use first derivative:

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

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 $\hat{s}$.

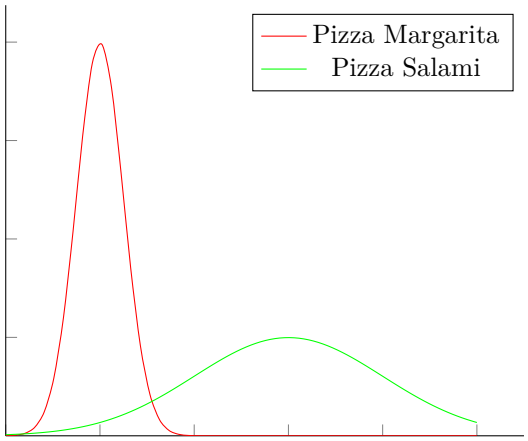


Figure 19: Red: Pizza margarita is quite popular among all subjects. Green: Pizza salami is not that popular among subjects, it has a higher variance.

10 Choice Models I

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 19).
Since the utility is dependent on the choice to made, there can only be *relative* utilities.

Figure 18: Pizzaaaa!

10.1 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 2 for an example.

Two options

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 20.

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

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 2: 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.

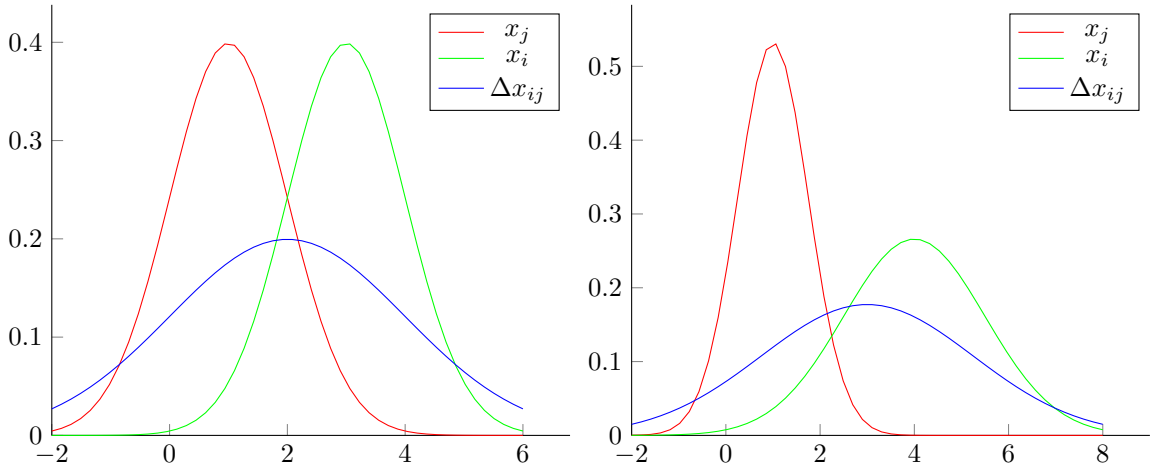


Figure 20: Relation of x_i, x_j and Δx_{ij}. Left: two options with equal variance. Right: two options with different variances. Note that the variances add up, hence Δx_{ij} gets flat and wide.

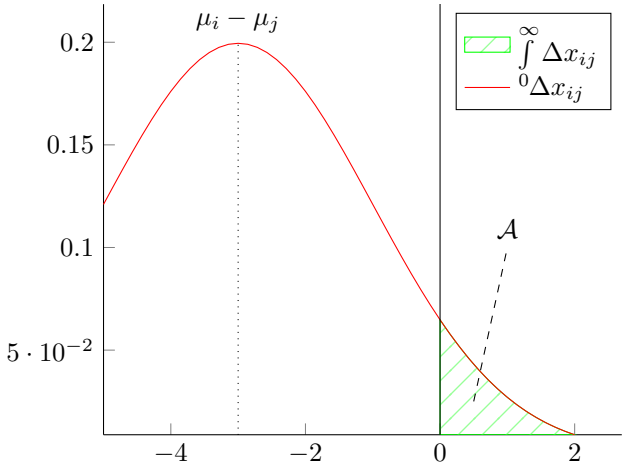


Figure 21: Δx_{ij}: If we know this area $\mathcal{A}$, we can guess $\Delta \mu_{ij}$.

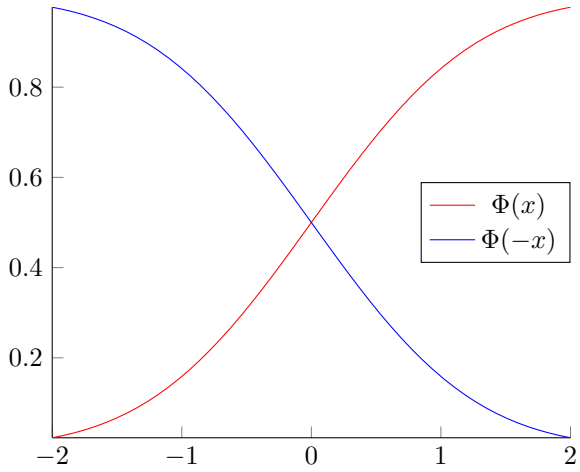


Figure 22: $\Phi(x) = 1 - \Phi(-x)$

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 22 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 23) 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 23: $\hat{d}$, 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).

10.2 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 24). For these distributions we can simply derive the matrix d and calculate the average distance between two plots.

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

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

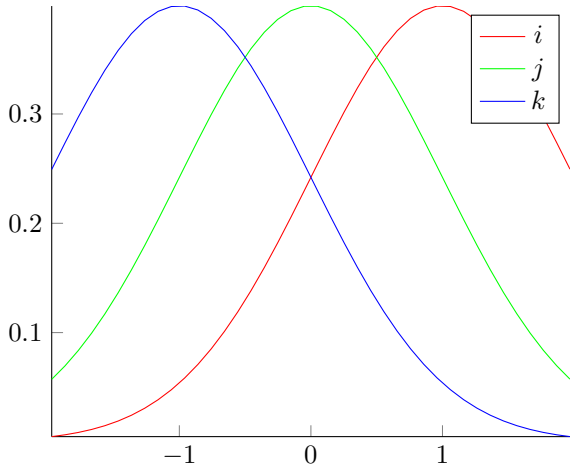


Figure 24: Three normal distributions i, j, and k.

11 Choice Models II

11.1 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?

11.2 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.

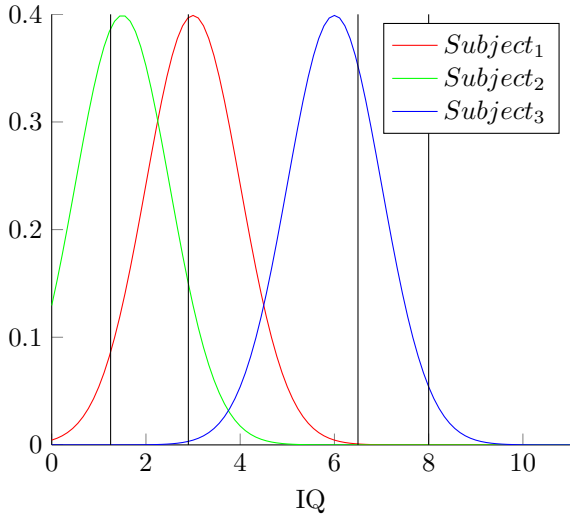


Figure 25: Performance of 3 subjects in an IQ-Test.

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:

11.2.1 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:

11.2.2 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:

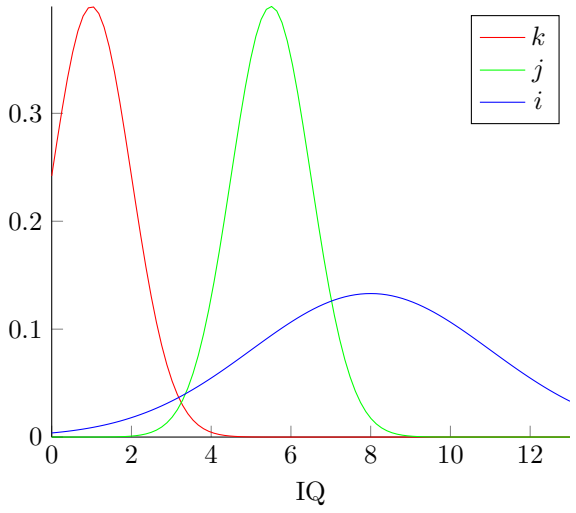


Figure 26: A problem for strong stochastic probability

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 27. 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?

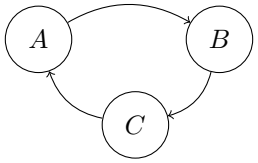


Figure 27: Three chessplayers dominate each other in a cyclic way.

11.2.3 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.

12 Everyday Predictions

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:

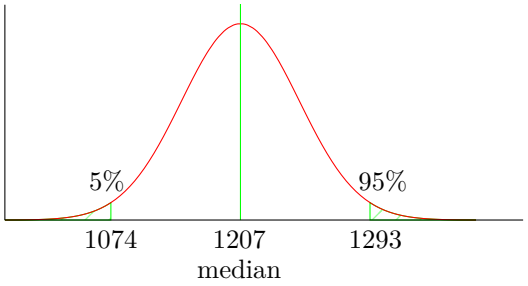


Figure 28: A popular fair game: guess the bull's weight. The *wisdom of the crowds* leads to quite a good solution.

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.*

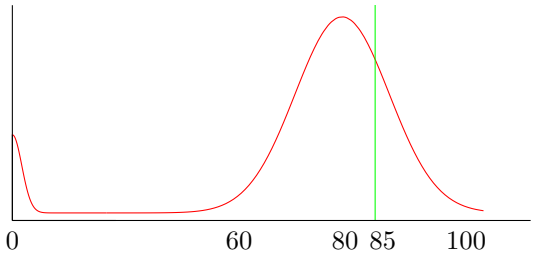


Figure 29: Example distribution of men's age of death in Germany.

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:

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

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

13 Appendix

13.1 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.