Probability & Statistics Notes

Bjarke Kingo Iversen

Winter 16/17

Contents

1 Introduction

This compilation of notes are originally meant for exam uses regarding Probability and/or Statistics. It covers most stuff from the 'SS' course held on institute of mathematics. These notes can be freely used, but there might be some overlooked errors. Formulas, pictures and examples are primarily from the books "**Introduction to Probability**" by *Blitzstein & Hwang* and "**Introduktion til Statistik**" by *Ditlevsen & Sørensen*. It is highly recommended that these books are used for the intend of learn, since these notes only covers a small subset of the contents.
The content is in English, with Danish translations on some of the terms.
I wish you a pleasant time using these notes/cheat-sheet.

2 Notation

Some of the notation used.

- μ = Expected mean value
- σ = Standard deviation
- σ^2 = Variance
- X = Capital letter. Usually designated to be random/stochastic variable
- x = Lowercase letter. Usually designated to be a specific value
- DS = "Introduktion til Statistik" by Susanne Ditlevsen and Helle Sørensen
- $\hat{\mu}$ = Estimated value for the mean value etc.
- $\tilde{\sigma^2}$ = Estimated value for sigma etc. But with correction
- $\bar{y}$ = Mean value
- $\sim$ = Distributed as
- SSD_y = Sum og squared deviations: $\sum\limits_{i=1}^{n}(y_i - \bar{y})^2$
- s = Designated to be the standard-deviation (spredning) estimate
- s^2 = Designated to be the variance-estimate

3 Sandsynlighed - (Probability theory)

We focus on two major parts of probability theory: Discrete probability distributions Continous probability distributions.
Common experiments to look at in probability theory would be i.e. Throwing dice, coin tossing, experiments with deck of cards etc.

Given af sample space S and some probabilistic outcomes of a **discrete** experiment we say:

$$\forall x \in S.(f(x) \in [0,1])$$

$$\sum\nolimits_{x \in S} f(x) = 1$$

That is all possible outcomes has a probability between 0 and 1, and the sample space S is equivalent with summing all probabilistic outcomes to 1.
Also, the probability of a subset E of our sample space S would be:

$$P(E) = \sum\nolimits_{x \in E} f(x)$$

In the naive probability-definition we say that:

$$P(A) = \frac{\text{Number of favourable outcomes}}{\text{Number of possible outcomes}}$$

Two sets A and B are disjoint if $A \cap B = \emptyset$, if they are disjunct we also have that $P(A \cup B) = P(A) + P(B)$.
We also have an additions rule for sets:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

and since we have that S = 1, we can calculate the complement probability by:

$$P(\neg A) = 1 - P(A)$$

3.1 Brugbare kombinatorikker - (useful combinatorics)

When calculating the probability of an event, its often necessary to calculate the number of favourable outcomes in different scenarios. Here are some examples:

3.1.1 Additionsmetoden - (Addition Method)

a + b = number of combinations

Example: You may choose **one** fruit from basket 1 = {Apple, Banana, Tomato} **or** from basket 2 = {Kiwi, Pineapple}. How many possible combinations?
Solution: |basket 1| + |basket 2| = 3 + 2 = 5.

3.1.2 Multiplikationsmetoden - (Multiplication Method)

a * b = number of combinations

Example: You may choose **one** fruit from basket 1 = {Apple, Banana, Tomato} **and** one from basket 2 = {Kiwi, Pineapple}. How many possible combinations?
Solution: For each fruit in basket 1, you have to different choices for basket 2. |basket 1| * |basket 2| = 3 * 2 = 6.

3.1.3 Rækkefølge betyder noget, uden tilbagelægning - (Order matters, no replacement)

$\frac{n!}{(n-k)!}$ = number of combinations

Example: We have 8 car-racers competing in a race, there are 3 major rewards, for the 1st, the 2nd, and the 3rd place. How many ways can we pick out a a driver for the 1st, 2nd, and 3rd place?
Solution: Let n = 8, and k = 3, we have $\frac{8!}{(8-3)!} = \frac{8!}{5!} = 336$ ways.

3.1.4 Fødselsdagsproblemet - (Birthday problem)

The birthday problem is a problem that considers a classroom with k students each having a birthday in one of the $n = 365$ days, utilizing the 'Order matters, no replacement' formula. It asks the question, what the probability is that two students have birthday the same day.
This is the same as asking the following: What is the probability that we have a birthday-collision given k birthdays?

For the first birthday, we can choose 365 out of 365 days and have no collision. For student number 2, we can choose 364 days out of 365 days. We see that the pattern is:

$$\frac{365}{365} * \frac{364}{365} * \frac{363}{365} * \cdots * \frac{n-k+1}{365}$$

We have the bound k for the number of students.

$$\frac{365 * 364 * 363 * \cdots * (365 - n + 1)}{365^k}$$

We can rewrite this by fully completing $n!$ as the nominator, and dividing it by $(n-k)!$ in the denominator as followed:

$$\frac{365!}{365^k * (365 - k)!}$$

The answer to the problem can be obtained by subtracting it from 1 since we computed the probability that we do not have a collision.
For $k = 23$ people and $n = 365$ days. We have:

$$P(2 \text{ People have same birthday}) = 1 - \frac{365!}{365^{23} * (365 - 23)!} \approx 0.507297$$

Thus it might be surprising. By just having 23 students in the same class, the probability of 2 having the same birthday is more than 50%.

This method can also be applied to similar problems, e.g. the probability of a hash-function hashing to the same value, given a hash-table of n size.

3.1.5 Rækkefølge betyder ikke noget, uden tilbagelægning - Binomial Coefficient - (Order doesn't matter, no replacement)

$\binom{n}{k} = \frac{n!}{(n-k)! * k!}$ = number of combinations

Example: We have 8 car-racers competing in a race, only 3 drivers gets to the final race. How many ways can we pick drivers to get to the final race?
Solution: Now the order of the 3 drivers doesn't matter, let n = 8, and k = 3, we have $\frac{8!}{(8-3)! * 3!} = 56$ ways.

3.1.6 Rækkefølge betyder noget, med tilbagelægning - (Order matters, with replacement)

n^k = number of combinations

Example: A 5-digit phone number consists of numbers 0 to 9. If all numbers are valid, how many combinations of phone numbers are possible?
Solution: Since replacements are allowed we're allowed to pick the same numbers (e.g. 55555). Let $n = 10, k = 5, 10^5 = 10000$ different phone numbers.

3.1.7 Rækkefølge betyder ikke noget, med tilbagelægning - (Order doesn't matter, with replacement)

$\frac{(n-1+r)!}{(n-1)! * r!}$ = number of combinations

Example: In an iceshop they have 7 different variants of ice cream. If you'd have to pick 3 scoops of ice cream, how many combinations of ice cream scoops are possible?
Solution: Since replacement are allowed, we are allowed to pick the same variant multiple times. Let n = 7, r = 3, we have $\frac{(7-1+3)!}{(7-1)! * 3!} = \frac{9!}{6! * 3!} = 84$ combinations of 3 ice cream scoops.

3.2 Betinget sandsynlighed m. Bayes Lov - (Conditional Probability w. Bayes' theorem)

Conditional probability is the probability that some event occurs given (by assumption, presumption, statement etc.) that some other event occurs/has occurred. It is read $P(A|B)$ *"Probability of A given B"*.
If i have an event A and B that are **independent**, it obviously doesn't have any effect on A that B occurred, thus making:

$$\boldsymbol{P(A|B) = P(A)}$$

Otherwise we have that:

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

Using Bayes' Theorem we can write it as:

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

Now we can use use $P(B|A), P(A)$ and $P(B)$ to calculate the conditional probability. Sometimes we may not have all the components needed, continue reading for LOTP.

3.3 Lov om total Sandsynlighed - (LOTP: Law of total probability)

The law of total probability we can use in many cases for calculating the probability of a single event.
Lets assume that we have $P(B|A)$, and $P(A)$ but not $P(B)$ for calculating $P(A|B)$. The law of total probability states that:

$$\boldsymbol{P(B) = \sum_{i=1}^{k} P(B|A_i) * P(A_i)}$$

Example: (2.11 in BH.) Consider an election with two candidates, Candidate A and Candidate B. Every voter is invited to participate in an exit poll, where they are asked whom they voted for; some accept and some refuse. For a randomly selected voter, let A be the event that they voted for A, and W be the event that they are willing to participate in the exit poll. Suppose that $P(W|A) = 0.7$ but $P(W|\neg A) = 0.3$. In the exit poll, 60% of the respondents say they voted for A (assume that they are all honest), suggesting a comfortable victory for A. Find $P(A)$, the proportion of people who voted for A.

Solution: From the text we can pick out some information:

- A = Voting for A
- W = Participating in exit-poll
- $P(W|A) = 0.7$
- $P(W|\neg A) = 0.3$
- $P(A|W) = 0.6$
- $P(A) = ??$

Lets see what we can do with Bayes' law:

$$P(W|A) = \frac{P(A|W) * P(W)}{P(A)} \to 0.7 = \frac{0.6 * P(W)}{P(A)}$$

Lets try to use LOTP on $P(W)$:

$$0.7 = \frac{0.6 * (P(W|A) * P(A) + P(W|\neg A) * P(\neg A))}{P(A)}$$

We know that $P(\neg A) = 1 - P(A)$ (Since $P(A) + P(\neg A) = 1$ per definition). Lets use it:

$$0.7 = \frac{0.6 * (0.7 * P(A) + 0.3 * (1 - P(A)))}{P(A)} = \frac{0.6 * (0.4 * P(A) + 0.3}{P(A)} =>$$

$P(A) = \frac{12}{35} * P(A) + \frac{9}{35}$, using wolfram-alpha we get that $P(A) = \frac{9}{23} \approx 0.39$

4 Stokastisk Variabel - (Random variable)

A random variable is a variable that maps to a value represented by a specific event, an example could be a coin toss:

$$X(coin) = \begin{cases} 1, \text{if coin} = \text{Heads} \\ 0, \text{if coin} = \text{Tails} \end{cases}$$

5 Sandsynlighedsfordelinger - (Probability distributions)

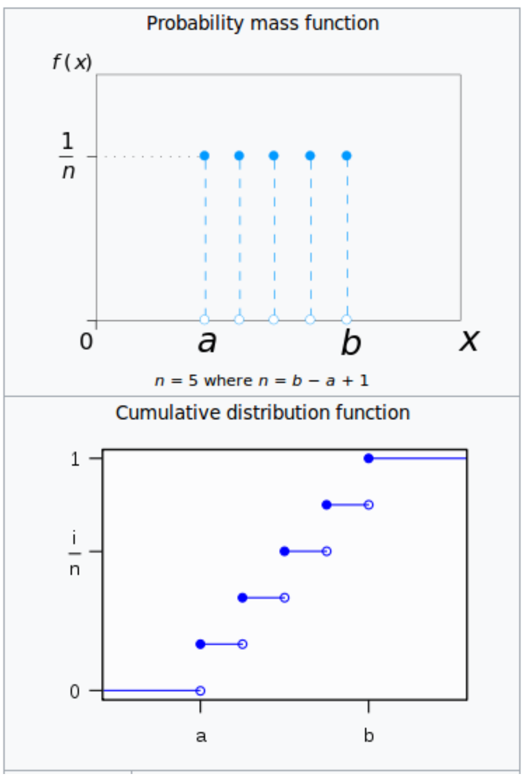
5.1 Ligefordeling - (Discrete uniform distribution)

The Discrete uniform distribution

$$X \sim DUnif(n)$$

is characterized with being one or more tests with n different outcomes, all equally likely to happen with probability $\frac{1}{n}$.

discrete uniform



Parameters	$a \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ $b \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}, b \geq a$ $n = b - a + 1$
Support	$k \in \{a, a+1, \ldots, b-1, b\}$
pmf	$\frac{1}{n}$
CDF	$\frac{\lfloor k \rfloor - a + 1}{n}$
Mean	$\frac{a+b}{2}$
Median	$\frac{a+b}{2}$
Mode	N/A
Variance	$\frac{(b-a+1)^2 - 1}{12}$

5.2 Bernouilli-fordeling - (Bernouilli distribution)

The Bernouilli-distribution

$$X \sim Bern(p)$$

is characterized with being a test with just 2 outcomes: success and failure. The random variable X of a Bernouilli-distribution maps to the value 1 with success probability p and to the value 0 with failure probability of $q = 1 - p$.
It could for instance be used to represent a coin toss where 1 and 0 would represent *head* and *tail*, in particular it could represent an unfair coin toss, thus $p \neq \frac{1}{2}$.
The Bernouilli-distribution is the special case of the binomial-distribution where just 1 experiment is conducted $(n = 1)$, that is $X \sim Bin(1, p) = Bern(p)$

Bernoulli

Parameters	$0 < p < 1, p \in \mathbb{R}$
Support	$k \in \{0, 1\}$
pmf	$\begin{cases} q = (1-p) & \text{for } k = 0 \\ p & \text{for } k = 1 \end{cases}$
CDF	$\begin{cases} 0 & \text{for } k < 0 \\ 1-p & \text{for } 0 \leq k < 1 \\ 1 & \text{for } k \geq 1 \end{cases}$
Mean	p
Median	$\begin{cases} 0 & \text{if } q > p \\ 0.5 & \text{if } q = p \\ 1 & \text{if } q < p \end{cases}$
Mode	$\begin{cases} 0 & \text{if } q > p \\ 0, 1 & \text{if } q = p \\ 1 & \text{if } q < p \end{cases}$
Variance	$p(1-p)(= pq)$

(***PMF: Probability mass function, CDF: Cumulative distribute function, Mode: The specific number that occurs the most in this distribution.***)

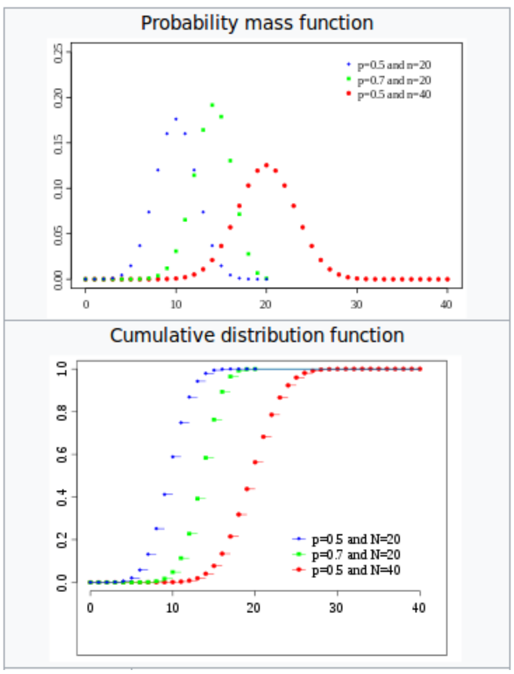
5.3 Binomial-fordeling - (Binomial distribution)

The Binomial-distribution

$$X \sim Bin(n, p)$$

is characterized with being a series of n experiments being Bernoulli-trials (yes/no answer to the experiment), with success rate p. All experiments are independent of eachother, that is, the success and failure of some experiments have no influence on others. The Binomial-distribution could for instance be used to describe an experiment with $n = 10$ cups, they have a chance of $p = 0.7$ of being filled with water and chance $1 - p$ for not being filled.
The mean-value of this experiment is then $n * p = 7$ cups that on average gets filled with water.

binomial



Notation	$B(n, p)$
Parameters	$n \in \mathbf{N}_0$ — number of trials $p \in [0,1]$ — success probability in each trial
Support	$k \in \{ 0, \ldots, n \}$ — number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$
CDF	$I_{1-p}(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$np(1-p)$

5.4 Hypergeometrisk fordeling - (Hypergeometric distribution)

The Hypergeometric distribution

$$X \sim HGeom(w, b, n)$$

is characterized with being a series of n experiments with probability rate $\frac{w}{w+b}$. Consider an urn with w white balls and b black balls. We draw n balls out of the urn at random, without replacement, such that all $\binom{w+b}{n}$ samples are equally likely. Let X be the number of white balls in the sample. Then X is hypergeometric distributed with parameters w, b, n.

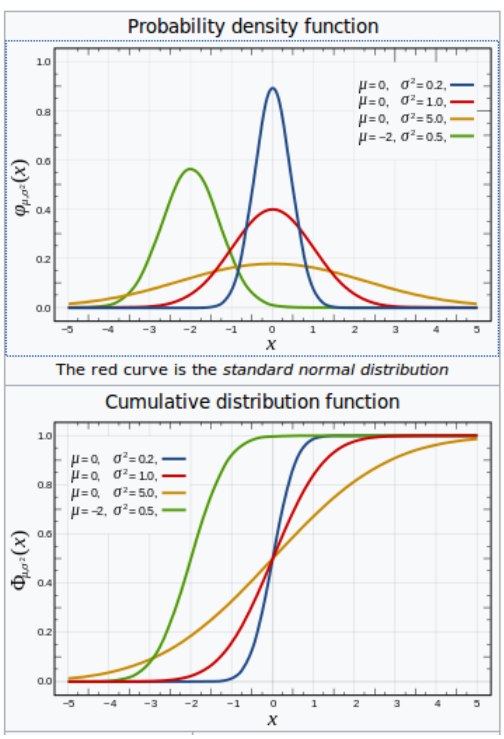
5.5 Normalfordeling - (Normal/Gaussian distribution)

The normal distribution

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

is a continuous probability distribution that are often used in real-valued experiments, whose distributions are not known. The probability density function for the standard normal distribution $\mathcal{N}(0, 1)$ is $\frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}x^2)$.

Normal distribution



Parameters	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbf{R}$
PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\mu + \sigma\sqrt{2}\operatorname{erf}^{-1}(2F-1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2

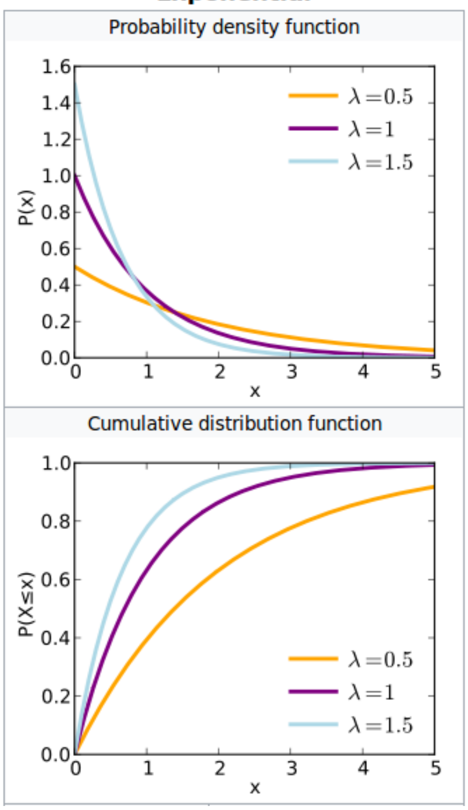
5.6 Eksponentialfordeling - Exponential distribution

The exponential distribution

$$X \sim Expo(\lambda)$$

is a continuous distribution with parameter $\lambda > 0$ and $x > 0$.

Exponential



Parameters	$\lambda > 0$ rate, or inverse scale
Support	$x \in [0, \infty)$
PDF	$\lambda e^{-\lambda x}$
CDF	$1 - e^{-\lambda x}$
Quantile	$-\ln(1 - F) / \lambda$
Mean	λ^{-1} $(= \beta)$
Median	$\lambda^{-1} \ln(2)$
Mode	0
Variance	λ^{-2} $(= \beta^2)$

6 Diskrete stokastiske variable - (Discrete random variables)

In discrete probability distributions we often work with some sort of finite dataset of Real Numbers which we can apply arithmetic calculations on.

6.1 Sandsynlighedsfunktion - (PMF: Probability Mass function)

The probability mass function tells us something about the different probabilistic events that can occur in our sample space S, and what their probabilities are. It takes a value as imput and returns a probability in an interval of $[0, 1]$. We say that:

$$\boldsymbol{f_X(x) = P(X = x)}$$

Lets consider the probability mass function for 2 fair coin tosses:
Let Q = Number of Heads, thus Q = {0,1,2}

$$f(0) = P(Q = 0) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$$

$$f(2) = P(Q = 2) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$$

$$f(1) = P(Q = 1) = 1 * \frac{1}{2} = \frac{1}{2}$$

We have now covered the possible outcomes of Heads, and of course we must have that the sum of all events with a probability > 0 in Q sums to 1:

$$\textstyle\sum_{x \in Q} f_Q(x) = 1$$

6.2 Fordelingsfunktion - (CDF: Cumulative Distribution Function)

The cumulative distribution function is defined as the sum of of our PMF, but just the values that are supported by the threshold given by our CDF, e.g.:

$$\boldsymbol{P_Q(q) = P(Q \leq q)}$$
$$P_Q(1.5) = P(Q \leq 1.5) = P(q = 0) + P(q = 1) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}$$

6.3 Middelværdi - (Mean / Weighted mean / Expectation)

The mean, often denoted as E[X] is in contradiction to the arithmetic mean, a mean of the expected value, by summing over the product of a value with its probability to occur. The mean is thus defined as:

$$\boldsymbol{E(X) = \sum_{j=1}^{\infty} x_j * P(X = x_j)}$$

Example: Lets consider the following table:

		Y		
		1	2	3
	1	0.12	0.02	0.06
X	2	0.04	0.50	0.06
	3	0.04	0.08	0.08

Let Z = max(X,Y), and lets calculate the mean of Z:
$E(X) = \max(1,1)*0.12+\max(1,2)*0.02+\max(1,3)*0.06+\max(2,1)*0.04+\max(2,2)*0.50+\max(2,3)*0.06+\max(3,1)*0.04+\max(3,2)*0.08+\max(3,3)*0.08 =$
1∗0.12+2∗0.02+3∗0.06+2∗0.04+2∗0.50+3∗0.06+3∗0.04+3∗0.08+3∗0.08 = **2.2**

6.3.1 Two variables

For two discrete random variables X and Y, the mean value can be computed using the LOTUS-principle thus:

$$E(g(X,Y)) = \sum_x \sum_y g(x,y)P(X=x, Y=y)$$

6.4 Varians & Spredning - (Variance & Standard Deviation)

The variance, often denoted as $Var(X)$, S^2 or σ^2, is a measurement on our random variable X which tells us har far X is from its mean on average. We have a little trick of squaring the value, which insures that both positive and negative deviations form the mean are considered.
The variance is defined as:

$$\boldsymbol{Var(X) = E(X - E(X))^2}$$

An easy way to calculate the variance is by using the equivalent expression:

$$\boldsymbol{Var(X) = E(X^2) - (E(X))^2}$$

The standard deviation is simply defined as the square-root of the variance:

$$\boldsymbol{\sigma = \sqrt{Var(X)}}$$

Example: Lets try to calculate the variance of the previous table. First we find the mean of squares:

$$\boldsymbol{E(X^2) = \sum_{j=1}^{\infty} x_j^2 * P(X = x_j)}$$

$$1^2 * 0.12 + 2^2 * 0.02 + 3^2 * 0.06 + 2^2 * 0.04 + 2^2 * 0.50 + 3^2 * 0.06 + 3^2 * 0.04 +$$
$$3^2 * 0.04 + 3^2 * 0.08 + 3^2 * 0.08 =$$
$$0.12 + 0.08 + 0.54 + 0.16 + 2 + 0.54 + 0.36 + 0.72 + 0.72 = \mathbf{5.24} = E(X^2)$$

Now that we've found the mean of squares we simply subtract the squared mean which we found in the previous section:

$$Var(X) = E(X^2) - (E(X))^2 =$$
$$5.24 - 2.2^2 = \mathbf{0.4}$$

6.5 Diskret stokastisk uafhængighed - (Independence of discrete random variables)

Given two discrete random variables, X and Y, we have that X and Y are independent if and only if:

$$\boldsymbol{\forall (x,y).(F_{X,Y}(x,y) = F_X(x) * F_Y(y))}$$

where $F_X(x) = P(X = x)$.

Example: Lets continue with figure 1. We have that:

$$F_X(1) = P(X = 1) = 0.12 + 0.02 + 0.06 = 0.20$$
$$F_Y(1) = P(Y = 1) = 0.12 + 0.04 + 0.04 = 0.20$$

We see that:

$$F_X(1) * F_Y(1) = 0.04$$

If we take a look at the table from figure 1 we see that:

$$F_{X,Y}(1,1) = 0.12$$

Now we definitely see that:

$$F_X(1) * F_Y(1) \neq F_{X,Y}(1,1)$$

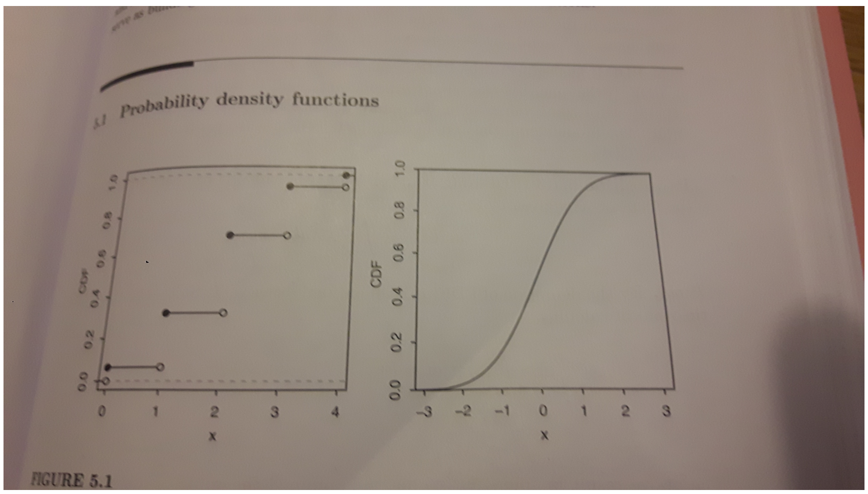
Thus we have $\exists(x,y).(F_{X,Y}(x,y) \neq F_X(x) * F_Y(y))$, and we conclude that they are **not** independent. Keep in mind, if we had that $F_X(1) * F_Y(1) = F_{X,Y}(1,1)$, we can't conclude anything, as we have to check for all possible values of (x,y).

7 Kontinuerte stokastiske variable - (Continuous random variables)

A continuous variable is a variable that can take infinitely many, uncountable values. Since we can have infinitely many points, we say for a random variable X that:

$$\mathrm{P(X = x) = 0}$$

That is, the probability that X maps to a specific x is 0.



The CDF of discrete probability distribution vs the CDF of a continuous probability distribution (from B.H.)

We are instead often interested in the area of a 2-dimensional probability room (for instance), calculated by integrating over some probability function.

7.1 Sandsynlighedstæthedsfunktion - (PDF: Probability Density Function)

The PDF f of a continuous random variable must satisfy the two criteria:

$$\boldsymbol{F_X f(x) \geq 0} \text{ for all } \boldsymbol{x}$$

,

$$\boldsymbol{\int_{-\infty}^{\infty} f(x)dx = 1}$$

That is, it must be positive and integrate to 1.

7.2 Simultan tæthed & Marginal tæthed - Simultaneous & Marginal PDFs

A simultaneous PDF must (like in the single variable case) satisfy the two criteria:

$$\boldsymbol{F_{XY} f(x, y) \geq 0} \text{ for all } \boldsymbol{(x, y)}$$

,

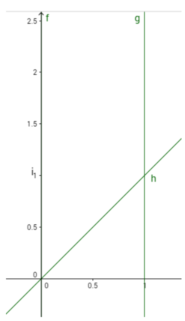
$$\boldsymbol{\int_{-\infty}^{\infty} (\int_{-\infty}^{\infty} f(x, y) dy) dx = 1}$$

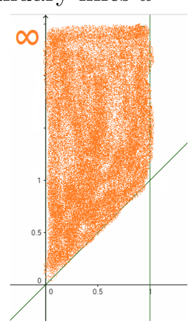
That is, $f(x, y)$ must be positive, and integrate to 1.

Example: Draw the sample space A of f and show that f is a probability density function, given:

$$f(x, y) = \begin{cases} \frac{4x^3}{y^3} & \text{if } 0 < x < 1, x < y \\ 0 & \text{otherwise} \end{cases}$$

We can start of by drawing the boundary-lines $x = 0$, $x = 1$ and $x = y$:





When now considering the boundaries $0 < x < 1$ and $x < y$, we now have A sketched in the orange area.
We see that f is defined on a positive sample space, now we see whether it integrates to 1:

$$\int_0^1 \int_x^{\infty} \frac{4x^3}{y^3} dydx = \int_0^1 4x^3 \int_x^{\infty} y^{-3} dydx = \int_0^1 4x^3 \left[-\frac{1}{2y^2} \right]_x^{\infty} dx =$$

$$\int_0^1 4x^3 (-\frac{1}{\infty} - -\frac{1}{2x^2}) dx = \int_0^1 4x^3 * \frac{1}{2x^2} dx = \int_0^1 \frac{4x^3}{2x^2} dx = \int_0^1 2xdx =$$

$$2 \int_0^1 xdx = 2 \left[\frac{x^2}{2} \right]_0^1 = 2\frac{1}{2} = \mathbf{1}$$

And thus we can conclude that f is indeed a probability density function. (It might seem weird that its 1 since y can be infinitely large, but remember $\lim_{y \to \infty} \frac{1}{y^3} = 0$, that is, the area gets infinitely small as y approaches ∞)

If X and Y have simultaneous PDF (Simultan tæthed) $F_{XY}f(x, y)$, then X has probability density function (Marginal tæthed):

$$\boldsymbol{F_X f(x) = \int_{-\infty}^{\infty} f(x, y) dy}$$

and Y has probability density function (Marginal tæthed):

$$\boldsymbol{F_Y f(y) = \int_{-\infty}^{\infty} f(x, y) dx}$$

You can think about it as 'We marginalize X out', therefore we integrate over dx if we want PDF for Y and likewise.

Example:
For the simultaneous probability density function $f(x, y) = \frac{4x^3}{y^3}$ from before, find the probability density function for x.
Solution:
Using our formula, we marginalize Y out:

$$\int_x^{\infty} \frac{4x^3}{y^3} dy = 4x^3 \int_x^{\infty} \frac{1}{y^3} dy =$$
$$4x^3 * \left[-\frac{1}{2y^2}\right]_x^{\infty} = \frac{4x^3}{2x^2} = \boldsymbol{2x}$$

7.3 Fordelingsfunktion - (CDF: Cumulative Distribution Function)

The cumulative distribution function $F_X(x)$ is the cumulative area up to x (remember X and x are not the same, so lets call $f(x)$ for $f(y)$):

$$\boldsymbol{F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy}$$

Likewise for joint CDF's ($f(x,y) = f(u,v)$):

$$\boldsymbol{F_{XY}(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u,v)dudv}$$

Example: Given $f(x)$:

$$f(x) = \begin{cases} \frac{1}{2}x^{-\frac{3}{2}} & \text{for } x > 1 \\ 0 & \text{otherwise} \end{cases}$$

Show that the CDF for x is given by:

$$F_X = \begin{cases} 0 \text{ for } x \leq 1 \\ 1 - x^{\frac{1}{2}} \text{ for } x > 1 \end{cases}$$

Solution: We see that $\int_{-\infty}^{1} = 0$ due to $x \leq 1$.
Thus we are left with the integral:

$$F_X(x) = P(X \leq x) = \int_1^x f(y)dy = \int_1^x \frac{1}{2}y^{-\frac{3}{2}}dy = \left[-y^{-\frac{1}{2}}\right]_1^x = -x^{-\frac{1}{2}} - -1 = \boldsymbol{1 - x^{-\frac{1}{2}}}$$

7.4 Middelværdi - (Mean / Weighted mean / Expectation)

The mean value for a probability destribution function (tæthed) is computed by integrating over the PDF (tæthed) multiplied by the variable of integration in the process:

> Let Y be a continuous random variable with probability density function $g(y)$ the mean of Y is:
>
> $$E(Y) = \mu = \int_{-\infty}^{\infty} y * g(y)dy$$

We also call this the '1st-moment'.

Example:
In the same PDF as before, find the mean value of X
Solution:
We use our formula, and integrate our marginal density function times the value x:

$$\int_{-\infty}^{\infty} xf(x)dx = \int_0^1 2x^2dx = 2\int_0^1 x^2dx =$$

$$2\left[\frac{x^3}{3}\right]_0^1 = 2\frac{1}{3} = \mathbf{\frac{2}{3} = \mu}$$

7.5 Varians & Spredning - (Variance & Standard Deviation)

The variance for a probability density function (tæthed) is computed by integrating over the PDF (tæthed) multiplied by the variable squared of integration in the process. We then subtract the mean value squared:

> Let Y be a continuous random variable with probability density function $g(y)$ the variance of Y is:
>
> $$Var(Y) = E(Y^2) - E(Y)^2 = \int_{-\infty}^{\infty} y^2 * g(y)dy - \mu^2$$

We also call this the '2nd-moment minus the 1st-moment'.
As usual the standard deviation is simply the square-root of the variance:

$$\boldsymbol{sd(Y) = \sigma = \sqrt{Var(Y)}}$$

Example:
Find the variance of X from the same probability function as before.
Solution:

$$\int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 = \int_0^1 2x^3 dx - \mu^2 = 2\int_0^1 x^3 dx - \mu^2 =$$

$$2\left[\frac{x^4}{4}\right]_0^1 - \mu^2 = 2\frac{1}{4} - \mu^2 = \frac{2}{4} - \left(\frac{2}{3}\right)^2 = \frac{9}{18} - \frac{8}{18} = \boldsymbol{\frac{1}{18} = Var(X)}$$

7.6 Uafhængighed - Independance of continuous random variables

When we want to check whether to continuous random variables are independent, we simply take the probability density function (marginal tæthed) of the first, and multiply it with the pdf (marginal tæthed) of the second, and check whether the product is equal to the simultaneous pdf (simultan tæthed).

Let $p(x, y)$ be a probability density function for (X, Y). Let $f(x)$ be the probability density function for X and $g(y)$ the probability density function for Y, then we have that they are independent if:

$$f(x) * g(y) = p(x, y)$$

Otherwise they're not independent.

7.7 Transformation (Ny stokastisk variabel Z) - (New random variable Z)

Sometimes we may want to introduce a new random variable Z with some relationship to another random variable Y.
Given some probability density function (tæthed) $g(y)$ for Y we can find the probability density function (tæthed) for Z.

> Let Y be a continous random variable with probability density function g and support $I \subseteq \mathbb{R}$ and let $g : I \rightarrow \mathbb{R}$ be strictly increasing/decreasing and be differentiable on I. Then $Z = h(Y)$ have the probability density function:
> $$f_Z = g(h^{-1}(z)) * |(h^{-1})'(z)|$$

To find $h^{-1}(z)$ we write $z = h(Y)$, and solve the function for Y (isolate Y).

Example: Let Y have probability density function (marginal tæthed):

$$g(y) = \begin{cases} -log(y) & \text{for } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Define new random variable $Z = -log(Y)$
Show that Z has probability density function (tæthed):

$$f(z) = \begin{cases} ze^{-z} & \text{for } z > 0 \\ 0 & \text{otherwise} \end{cases}$$

Solution:
We use the formula and determine h, g, $h^{-1}(z)$, $(h^{-1})'(z)$:

- $h = -log(Y)$ given by our definition of Z
- $g = -log(y)$ given by the probability density function for Y
- $h^{-1}(z) =$
$$z = -log(Y) \rightarrow -z = log(Y) \rightarrow e^{-z} = Y \rightarrow h^{-1}(z) = \boldsymbol{e^{-z}}$$
- $(h^{-1})'(z) = (e^{-z})' = -e^{-z}$

Now lets insert it in our formula:

$$f(z) = -log(e^{-z}) * |-e^{-z}| = -log(e^{-z}) * e^{-z} = --z * e^{-z} = \boldsymbol{ze^{-z}}$$

Example 2: Find the probability density function of X^7 for $X \sim Expo(\lambda)$
Solution:
We know that the PDF for $X \sim Expo(\lambda)$ is given by:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ for } x > 0 \\ 0 \text{ otherwise} \end{cases}$$

$g(x) = x^7$, $Y = g(x)$, and we see that $(x^7)' = 7x^6$ is strictly positive. Thus Y have the PDF:

$$h(y) = \begin{cases} \lambda e^{-\lambda g^{-1}(y)} |(g^{-1})'(y)| \text{ for } y \in J \\ 0 \text{ otherwise} \end{cases}$$

- $g^{-1}(y) = y^{\frac{1}{7}}$, by isolating the $y = x^7$.
- $(g^{-1})'(y) = \frac{1}{7} y^{\frac{1}{7}-1} = \frac{1}{7} y^{\frac{-6}{7}}$

We insert the values and obtain the PDF:

$$h(y) = \begin{cases} \lambda e^{-\lambda y^{\frac{1}{7}}} \frac{1}{7} y^{\frac{-6}{7}} \text{ for } y > 0 \\ 0 \text{ otherwise} \end{cases}$$

8 Joint functions

See BH. if you're dealing with Joint CDF's, PMF's etc.

9 Statistics

9.1 Maximum likelihood estimation

Remember the PMF for Binomial-distribution?:
If $X \sim \text{bin}(n, p)$ for a given p, then our probability-mass-function

$$f_p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ for } x = 0, 1, \cdots, n$$

describes the probabilities for the possible outcomes of X, with probability p to observe x successes.

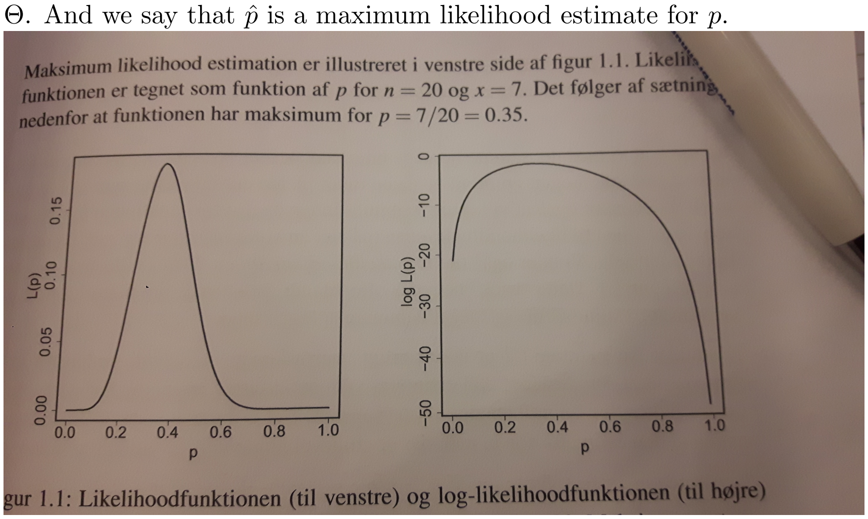
That is the way when we think in terms of probability. Now we want to do the opposite: We have an observation x, but we don't know the probabilityparameter p. Given the observation our goal is to estimate p. That is, we wish to find the value of p that fits the observation x best (Said with other words, our estimate p is the value p that maximizes the probability of observing exactly that value of X that we are given).
Thus we calculate $f_p(x) = P(X = x)$ for a given value x, but for all possible values of p, and then we pick the value that maximizes p.

The likelihood function is identical with the probability-mass-function, but is now considered a function of p for a static x rather than the opposite. If the parametersupport is Θ, then the likelihood-function for x is defined by:

$$L_x : \Theta \to [0, 1]$$

$$L_x(p) = f_p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad \text{for } p \in \Theta$$

As the estimate for p we will use the value in Θ that maximizes L_x, where x is the observed value. We also seek a value $\hat{p} \in \Theta$ thus that $L_x(\hat{p}) \geq L_x(p), \qquad$ for $p \in \Theta$. And we say that $\hat{p}$ is a maximum likelihood estimate for p.


Maksimum likelihood estimation er illustreret i venstre side af figur 1.1. Likelih
funktionen er tegnet som funktion af p for $n = 20$ og $x = 7$. Det følger af sætning
nedenfor at funktionen har maksimum for $p = 7/20 = 0.35$.

gur 1.1: Likelihoodfunktionen (til venstre) og log-likelihoodfunktionen (til højre)

Figure: The Maximum likelihood estimation is illustrated to the left. The likelihood-function is a function of p***,*** $x = 7$ ***in a binomial-distribution with*** $n = 20$***. That maximum is assumed to be*** $p = \frac{x}{n} = 0.35$***. The log-likelihood-function is illustrated to the right, and may sometimes be more useful to look at, since it doesn't sharp the edges as much (It seems like there is*** 0% ***chance that*** p ***is*** 0.8 ***in the above example.***

A good estimation for p for a binomial distribution is:

$$\hat{p}(x) = \frac{x}{n}$$

Given a model with the sample space $E = \{0, \cdots, n\}$ with the family $\wp = \{bin(n, p) : p \in \Theta\}, \Theta \subseteq [0, 1]$.

Question: We want to research the development of the stock-market. One day we pick 10 stocks and we register how many of the stocks that decreased that day.
Under which circumstances can the binomial-distribution be used to describe the number of stocks where the exchange rate decreased?
Solution: We can use the binomial-distribution to describe the number of stocks where the exchange rate decreased given that: We have n Bernoulli-trials with a success rate p and failure rate $1 - p$, and the events are independent of each-other.
Example: Under the assumption that the circumstances are satisfied, and that the exchange rate decreased on 8 stocks (that is x = 8). Construct a statistic model that can be used to describe the experiment.
Solution: $X =$ Number of stocks that decreased. $x = 8$. The model consists of the sample space $E = \{0, \cdots, 10\}$ with the family $\wp = \{bin(10, p) : p \in \Theta\}, \Theta \subseteq [0, 1]$.
Example: Give an estimate for p.
Solution:

$$\hat{p}(x) = \frac{x}{n} = \frac{x}{10}$$

$$\hat{p}(8) = \frac{8}{10} = 80\% = 0.8$$

Example: Declare the estimators distribution. Also declare the estimated mean and variance for the estimator.
Solution: The estimator:

$$\hat{p}(X)$$

can be written as

$$n\hat{p} \sim \text{ bin}(n, p)$$

in this case

$$10 * \hat{p}(X) \sim \text{ bin}(10, p)$$

(This is from sentence 1.3 in D.S. Introduktion til Statistik).
The mean of the estimator will be $E(\hat{p}(X)) = p$ since, if we use this estimator

for p we would be correct on average.
The variance of the estimator will be $Var(\hat{p}(X)) = \frac{p(1-p)}{10}$. It is also seen here that the more observations we do (in this case we have 10), the variance will faster approach 0 (Which makes totally sense if you think about it!). **Question:** Assume that the probability-parameter is 0.5. What is the probability that at least 8 stocks will decrease in exchange rate? And whats the probability that at most 8 stocks has decreased in exchange rate?
Solution:

9.1.1 When calculating likelihood on a Bernouilli-distribution

Remember that the Bernouilli-distribution is simply a Binomial distribution with $n = 1$

9.1.2 When calculating likelihood on a Normal-distribution

See sections about one/two samples with unknown/known variance.

9.2 En stikprøve med kendt varians - One Sample with known variance

See section 3 in DS.

9.3 En stikprøve med ukendt varians - (One sample with unknown variance)

"It's almost never with a known variance!" - ***Frederik***

This section is based on the normal-distribution.

The model for one sample with unknown variance consists of the sample space $\mathbb{R}^n$ and the family:

$$\wp = \{\mathcal{N}^n_{\mu,\sigma^2} : (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)\}$$

of distributions on $\mathbb{R}^n$ where $\mathcal{N}^n_{\mu,\sigma^2}$ have probability density function

For independent normaldistributed random varaibles $Y_1, \cdots, Y_n$ with mean μ and variance σ^2. The simultaneous distribution designated $\mathcal{N}^n_{\mu,\sigma^2}$ have probability density function:

$$f_{\mu,\sigma^2}(y) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right), \text{ for } y = (y_1, \cdots, y_n) \in \mathbb{R}^n$$

9.3.1 Maximum likelihood estimation

For the statistic model we just described, the maximum likelihood estimate for (μ, σ^2) is given by:

$$\boldsymbol{\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i}$$

for the mean value, and:

$$\boldsymbol{\hat{\sigma^2} = \frac{1}{n} SSD_y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

for the variance.

The estimators $\hat{\mu} = \bar{Y}$ and $\hat{\sigma^2} = \frac{1}{n} SSD_y = \frac{1}{n}\sum\limits_{i=1}^{n}(y_i - \bar{y})^2$ are independent, and their marginal distributions are:

$$\boldsymbol{\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \qquad \hat{\sigma^2} \sim \frac{\sigma^2}{n}\chi^2_{n-1}}$$

that is, $\bar{Y}$ is normal distributed around mean μ and variance to be the true variance over the number of observations, and $\hat{\sigma^2}$ is χ^2-distributed (chi^2) with $n-1$ degrees of freedom, and scale-parameter $\frac{\sigma^2}{n}$.
In other words $\frac{n}{\sigma^2}\hat{\sigma^2}$ is "true" χ^2_{n-1}−distributed.
OBS According to DS, on average σ^2 will be estimated too low, using the maximum likelihood estimator.
Therefore we normalize with $n-1$ instead of n in our definition of $\hat{\sigma^2}$ and therefore we use:

$$\boldsymbol{\tilde{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1} SSD_Y}$$

as our estimator. Now we have the following:
$E(\tilde{\sigma^2}) = \sigma^2$, $\bar{Y}$ and $\tilde{\sigma^2}$ are independent, and $(n-1)\tilde{\sigma^2} \sim \sigma^2\chi^2_{n-1}$. Thus $\tilde{\sigma^2} \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$. The similar estimate, where the observations are used, is designated usually as s^2, thus:

$$\boldsymbol{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

and it will almost always be this estimate we use for the variance.

Example:
With the intention to examine whether there is a difference between the visual and the auditive reaction time on humans, someone measured both forms of reaction time on 15 basketball-players. The visual reaction time is measured as the time it takes before a test individual reacts on a light signal, while the

auditive reaction time is measured as the time it takes before before the test individual reacts on a specific sound. All measurements are in milliseconds.

Spiller	Visuel	Auditiv	Forskel
1	161	157	4
2	203	207	−4
3	235	198	37
4	176	161	15
5	201	234	−33
6	188	197	−9
7	228	180	48
8	211	165	46
9	191	202	−11
10	178	193	−15
11	159	173	−14
12	227	187	40
13	193	182	11
14	192	159	33
15	212	186	26
$\bar{y}$	197	185.4	11.6
s	23.11	20.99	25.67

Example:
Create a statistic model for data corresponding the visual reaction time, and compute the estimates for the parameters in the model.
Solution:
$n = 15$
Model:
We have the sample space: $\mathbb{R}^{15}$
corresponding to the family: $\wp = \{\mathcal{N}^{15}_{\mu,\sigma^2} : (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)\}$
The estimate for the mean value is given in the dataset by $\hat{\mu} = \bar{y} = 197$.
We know the formula to estimate the variance $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, but in this case we have the estimate for the standard-deviation given in the dataset by $s = 23.11$. Thus we must have that the estimate for the variance is $s^2 = 23.11^2$.

Example:
Specify the theoretic distribution of the belonging estimators, and the estimated standard deviation for the mean value estimate.
Solution:
$\hat{\mu} \sim \mathcal{N}(\mu, \frac{\sigma^2}{15})$ (obtained by using our definition $\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$)
$s^2 \sim \frac{\sigma^2}{n-1}\chi^2_{14}$ (obtained by using our definition $\tilde{\sigma^2} \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$)
To specify the estimator for the standard deviation for the mean value estimate, we simply take the variance given by the mean value estimate: $\frac{\sigma^2}{15}$ and take the

squareroot: $s = \sqrt{\frac{23.11^2}{15}} = 5.97$.

9.3.2 Confidence interval for the mean value

For the 1 sample with unknown variance case, we can calculate a $1 - \alpha$ % confidence interval for μ by:

$$\bar{Y} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{\tilde{\sigma}}{\sqrt{n}} = (\bar{Y} - t_{n-1,1-\frac{\alpha}{2}} \frac{\tilde{\sigma}}{\sqrt{n}}, \bar{Y} + t_{n-1,1-\frac{\alpha}{2}} \frac{\tilde{\sigma}}{\sqrt{n}})$$

where $\bar{Y}$ is the mean, $\tilde{\sigma}$ is the standard deviation estimate, n is the number of observations, and α is the error rate. For a 95% confidence interval we choose α to be 0.05, thus we get the t-fractile $t_{n-1,0.975}$.
In order to calculate the t-fractile for a 95% confidence interval, we use the r-command:

qt(0.975, n-1)

Example:
Calculate a 95% and a 90% confidence interval for the expected visual reaction time.
Solution:
We know the formula for calculating a confidence interval: $\bar{y} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{\tilde{\sigma}}{\sqrt{n}}$.
95% We put in the mean value: 197, the fractile $n-1 = 14, 0.975$, the estimated standard deviation $s = 23.11$, and $n = 15$:

$$197 \pm t_{14,0.975} \frac{23.11}{\sqrt{15}}$$

We calculate the fractile using **qt(0.975,14)**:

```
> qt(0.975,df=14)
[1] 2.144787
```

and plot it into our formula:

$$197 \pm 2.145 \frac{23.11}{\sqrt{15}}$$

Using any calculator we obtain the values

```
> 197+2.145*(23.11/sqrt(15))
[1] 209.7992
> 197-2.145*(23.11/sqrt(15))
[1] 184.2008
```

From this we can conclude that with 95% probability the true mean exist in the interval $(184.2, 209.8)$.

90% We put in the mean value: 197, the fractile $n - 1 = 14, 1 - \frac{0.10}{2}$, the estimated standard deviation $s = 23.11$, and $n = 15$:

$$197 \pm t_{14,0.95} \frac{23.11}{\sqrt{15}}$$

We calculate the fractile using **qt(0.95,14)**:

```
> qt(0.95,df=14)
[1] 1.76131
```

and plot it into our formula:

$$197 \pm 1.761 \frac{23.11}{\sqrt{15}}$$

Using any calculator we obtain the values

```
> 197-1.761*(23.11/sqrt(15))
[1] 186.4922
> 197+1.761*(23.11/sqrt(15))
[1] 207.5078
```

From this we can conclude that with 90% probability the true mean of the visual reaction time exist in the interval $(186.5, 207.5)$.

Example:
Calculate likewise a 95% confidence interval for the expected auditive reaction time.
Solution:
Using the same formula and same t-percentile:

$$\bar{x} \pm t_{14,0.975} \frac{\tilde{\sigma}}{\sqrt{n}}$$

Using the r-command **qt**, to obtain the exact same value from the t-percentile, and inserting all our values:

$$185.4 \pm 2.145 \frac{20.99}{\sqrt{15}}$$

Evaluate to obtain the confidence interval $(173.8, 197.025)$ for the auditive reaction time.

Gruppe	Årskarakter	Eksamenskarakter	Forskel
1	7.67	5.67	2.00
2	9.33	7.67	1.67
3	8.67	8.33	0.33
4	9.67	8.33	1.33
5	7.33	7.00	0.33
6	8.67	8.33	0.33
7	7.33	7.33	0.00
8	8.00	8.33	−0.33
9	8.33	7.00	1.33
10	9.00	7.67	1.33
11	7.33	7.00	0.33
12	7.67	6.33	1.33
13	8.67	9.33	0.67
14	6.33	5.33	1.00
15	8.00	7.67	0.33
16	8.00	8.00	0.00
17	9.67	9.00	0.67
$\bar{y}$	8.216	7.548	0.744
s	0.914	1.086	0.662

Figure: Data of the mean grades of different classes in a high-school with annual as well as exam grades. From DS

Example:

Create a statistic model that can be used to examine whether there is a difference between the annual grades and the exam grades.

Solution:

The model for the difference consists of the sample space $\mathbb{R}^{17}$ and the family:

$$\wp = \{\mathcal{N}^{17}_{\mu,\sigma^2} : (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)\}$$

Example:

Specify estimates for the parameters in the model. Specify also the estimators distributions and the estimated standard deviation for the mean-estimate.

Solution:

$\hat{\mu} = \bar{y} = 0.744$ given by the dataset.
$s = 0.662$ given by the dataset.
$\tilde{\sigma^2} = s^2 = 0.662^2$.
The distributions of the estimators are given by our previous formulas:
$\hat{\mu} \sim \mathcal{N}(\mu, \frac{\sigma^2}{17})$ for the mean. $\tilde{\sigma^2} \sim \frac{\sigma^2}{16}\chi^2_{16}$ for the variance. The estimated standard deviation for the mean-estimator we simply find by our formula $\sqrt{\frac{\sigma^2}{n}}$. We insert:
$\tilde{\sigma} = \sqrt{\frac{0.662^2}{17}} \approx 0.16$.

Example:

Examine whether there is niveau-difference in the 2 forms of grades.

Solution:

Lets do a confidence interval!

We simply calculate it like before:

$$\bar{y} \pm t_{16,0.975}\frac{\tilde{\sigma}}{\sqrt{17}} =$$

$$0.744 \pm 2.12\frac{0.16}{\sqrt{17}} = (0.404, 1.084)$$

We see that 0 is not a part of the interval, and that the niveau-difference is positive. With 95% probability the true mean is withing the interval.

9.4 To stikprøver

This section is based on the normal-distribution.

While we just looked at 1 sample with unknown variance, one may sometimes have a scenario where we want to compare to different samples. This could for example be comparing a specific treatment for shoulder pain versus no treatment, or compare 2 products on quality-reviews.

Formalia: We have two groups of observations: $x_1, \cdots, x_{n_1}$ and $y_1, \cdots, y_{n_2}$. The observations are a case of the random variables $X_1, \cdots, X_{n_1}$ with mean μ_1 and variance σ^2, and $Y_1, \cdots, Y_{n_2}$ with mean μ_2 and variance σ^2. So each group has its own mean, but we're assuming that they have the same variance. Also note that the number of observations are not necessarily the same in the two samples. Thus the total number of observations is $n = n_1 + n_2$.
It can then be graphically examined whether its a fair assuming that the variance are the same or not.

The simultaneous distribution of all X_i's and Y_j's, designated $\mathcal{N}^{n_1,n_2}_{\mu_1,\mu_2,\sigma^2}$ have density function:

$$f_{\mu_1,\mu_2,\sigma^2}(x,y) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\Bigg(- \frac{1}{2\sigma^2}\Bigg(\sum_{i=1}^{n_1}(x_i - \mu_1)^2 + \sum_{j=1}^{n_2}(y_j - \mu_2)^2 \Bigg)\Bigg)$$

Where $x = (x_1, \cdots, x_{n_1}) \in \mathbb{R}^{n_1}$ and $y = (y_1, \cdots, y_{n_2}) \in \mathbb{R}^{n_2}$

The model for 2 samples with same variance consists of the sample space $\mathbb{R}^n$ with the family

$$\boldsymbol{\wp = \{\mathcal{N}^{n_1,n_2}_{\mu_1,\mu_2,\sigma^2} : (\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty)\}}$$

of distributions on $\mathbb{R}^n$ where $\mathcal{N}^{n_1,n_2}_{\mu_1,\mu_2,\sigma^2}$ have the just defined density function.

9.4.1 Maximum likelihood estimation

For the statistic model we just described, the maximum likelihood estimate for (μ_1, μ_2, σ^2) is given by:

$$\boldsymbol{\hat{\mu_1} = \bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i}$$

for the mean of the first variable,

$$\hat{\mu_2} = \bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_i$$

for the mean of the second variable, and:

$$\hat{\sigma^2} = \frac{1}{n}(SSD_x + SSD_y) = \frac{1}{n}\left(\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2\right)$$

for the variance.
The estimators $\hat{\mu_1} = \bar{X}$, $\hat{\mu_2} = \bar{Y}$ and $\hat{\sigma^2} = \frac{1}{n}(SSD_x + SSD_y)$ are independent, and their marginal distributions are:

$$\hat{\mu_1} \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n_1}\right), \hat{\mu_2} \sim \mathcal{N}\left(\mu_2, \frac{\sigma^2}{n_2}\right), \hat{\sigma^2} \sim \frac{\sigma^2}{n}\chi^2_{n-2}$$

Like in the 1 sample case we normalize our variance-estimate in order to reach a central estimate, this time we normalize with $n-2$ thus:

$$\tilde{\sigma^2} = \frac{1}{n-2}\left(\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2\right) = \frac{1}{n-2}(SSD_X + SSD_Y)$$

as our estimator. Now we have the following:
$(n-2)\tilde{\sigma^2} \sim \sigma^2\chi^2_{n-2}$, $E(\tilde{\sigma^2}) = \sigma^2$. The similar estimate estimate, whether the observations are used, is designated as s^2, thus:

$$s^2 = \frac{1}{n-2}\left(\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2\right) = \frac{SSD_x + SSD_y}{n-2} = \frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n-2}$$

Example:
In an investigation about whether "women are more talkative than men" the number of words that 396 female and male students speaks on a day was measured. The results are shown in the following table:

	Women (n = 210)	Men (n = 186)
Mean	16215	15669
Standard Deviation	7301	8633

Make a statistic model that makes it possible to examine whether there is a difference on the number of words a woman, and a man speaks in a day.
Solution:
The model for number of words per day for women and men consists of the sample space $\mathbb{R}^{396}$ and the family:

$$\wp = \{\mathcal{N}^{n_1,n_2}_{\mu_1,\mu_2,\sigma^2} : (\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty)\}$$

with x_1 to be women and x_2 to be men.

Example:
Specify estimates for all the parameters in the model, and specify their distributions. Also compute the estimated standard deviation for the estimator of the mean.
Solution:

$$\hat{\mu_1} = 16215, \hat{\mu_2} = 15669$$

for the mean values that are pretty much given.
We know the formula for the variance $s^2 = \frac{1}{n-2}(SSD_{x_1} + SSD_{x_2})$ can be written as $s^2 = \frac{(n_1-1)s_{x_1}^2 + (n_2-1)s_{x_2}^2}{n-2}$. And we know that $n = 396$, $n_1 = 210$, $n_2 = 186$. Recall that $s_{x_1}^2$ (the marginal/emperical variance) can be computed by $s_{x_1}^2 = \frac{1}{n_1-1}SSD_{x_1}$ and standard deviation values are given in the question so we simply square them to get $s_{x_1}^2$ and $s_{x_2}^2$.
Thus we have:

$$s^2 = \frac{(209)7301^2 + (185)8633^2}{394} = 63270226$$

The distributions are then:
$\hat{\mu_1} \sim \mathcal{N}(\mu_1, \frac{\sigma^2}{210})$
$\hat{\mu_2} \sim \mathcal{N}(\mu_2, \frac{\sigma^2}{186})$
$s^2 \sim \frac{\sigma^2}{394}\chi^2_{394}$

And the estimated standard deviations for estimator of the mean value:
$\hat{\mu_1} : \sqrt{\frac{s^2}{210}} = 548.986$
$\hat{\mu_2} : \sqrt{\frac{s^2}{186}} = 583.2345$

9.4.2 Confidence intervals

In the version of one sample with unknown variance, we had that the confidence interval was:

$$\bar{Y} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{\tilde{\sigma}}{\sqrt{n}}$$

for the mean value.

Now we use both samples to estimate the joint variance.
For the statistic model we defined earlier in this section, then

$$\bar{X} \pm t_{n-2,1-\frac{\alpha}{2}} \frac{\tilde{\sigma}}{\sqrt{n_1}} = \left(\bar{X} - t_{n-2,1-\frac{\alpha}{2}} \frac{\tilde{\sigma}}{\sqrt{n_1}}, \bar{X} + t_{n-2,1-\frac{\alpha}{2}} \frac{\tilde{\sigma}}{\sqrt{n_1}}\right)$$

is a $1-\alpha$ confidence interval for μ_1, and

$$\bar{Y} \pm t_{n-2,1-\frac{\alpha}{2}} \frac{\tilde{\sigma}}{\sqrt{n_2}} = \left(\bar{Y} - t_{n-2,1-\frac{\alpha}{2}} \frac{\tilde{\sigma}}{\sqrt{n_2}}, \bar{Y} + t_{n-2,1-\frac{\alpha}{2}} \frac{\tilde{\sigma}}{\sqrt{n_2}}\right)$$

is a $1-\alpha$ confidence interval for μ_2.

When we have two samples we're often interested in the difference in their mean values. Thus it makes sense that

$$\widehat{\mu_1 - \mu_2} = \hat{\mu_1} - \hat{\mu_2} = \bar{x} - \bar{y}$$

That is, that the estimated difference in the mean value, is the estimated mean value for x minus the estimated mean value for y.
Since $\bar{X}$ and $\bar{Y}$ are independent and have the distributions we descriped before, it follows that

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

By substituting σ with the estimate s, we get the estimated standard deviation for the estimator

$$SE(\widehat{\mu_1 - \mu_2}) = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(SE means standard error)

Now

$$\bar{X} - \bar{Y} \pm t_{n-2,1-\frac{\alpha}{2}} \tilde{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \left(\bar{X} - \bar{Y} - t_{n-2,1-\frac{\alpha}{2}} \tilde{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X} - \bar{Y} + t_{n-2,1-\frac{\alpha}{2}} \tilde{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

is a $1-\alpha$ confidence interval for $\mu_1 - \mu_2$

Example:
Using the same example as before, calculate an estimate for the expected difference between the number of words the two genders say in one day. Also compute the estimated standard deviation for the belonging estimator and calculate a 95% confidence interval for the difference.
Solution:
The expected difference can be found using our formula $\widehat{\mu_1 - \mu_2} = \hat{\mu_1} - \hat{\mu_2} = 16215 - 15669 = 546$.
We have that $\widehat{\mu_1 - \mu_2} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{210} + \frac{\sigma^2}{186}\right)$.
The estimated standard deviation we can compute using our formula $\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = \sqrt{\frac{63270226}{210} + \frac{63270226}{186}} \approx 800.91$.
Finally lets find a confidence interval. Using our formula

$$\bar{X_1} - \bar{X_2} \pm t_{394,0.975} * \tilde{\sigma} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 546 \pm t_{394,0.975} * \sqrt{63270226} * \sqrt{\frac{1}{210} + \frac{1}{186}}$$

Using R, we get the confidence interval $(-1072.436, 2104.436)$. We can conclude that with 95% probability, the true mean value of difference exists in this interval. The interval consists 0. We can therefore not reject that the two gender speak equally many words, with 95% probability.

9.5 Lineær regression

See Section 6 in DS.

10 R

Here is a list of some neat functions for the R statistic program

11 Appendix

11.1 Integration

Integration is often used for finding the area underneath the graph of a function. Given a function f of a real variable x and an interval $[a, b]$ of the real line in the cartesian plane, we can compute the definite integral:

$$\boldsymbol{\int_a^b f(x)dx = F(b) - F(a)}$$

Example: Find the area in the interval $[1, 2]$ of the function $x^2 + 1$.

$$\int_1^2 x^2 + 1dx$$

First we find out what rule to use (see figure below for common rules). In this case we use the sum rule:

$$\int_1^2 x^2 + 1dx = \int_1^2 x^2 dx + \int_1^2 1dx$$

Now we find the 2 integrals. We see need to use the power rule on x^2 and the constant rule for 1:

$$\left[\frac{x^3}{3}\right]_1^2 + [x]_1^2$$

Thus we computed the integrals. Now we compute the integral by subtracting the interval up to a, from b inserting a and b on $x's$ place:

$$\left(F(2) = \frac{2^3}{3} - F(1) = \frac{1^3}{3}\right) + \left(F(2) = 2 - F(1) = 1\right) =$$

$$\left(\frac{2^3}{3} - \frac{1^3}{3}\right) + \left(2 - 1\right) = \boldsymbol{\frac{10}{3}}$$

Often we might have the interval $[0, b]$, where $F(0) = 0$, thus we simply need to calculate $F(b)$.

Common Functions	Function	Integral
Constant	∫a dx	ax + C
Variable	∫x dx	x²/2 + C
Square	∫x² dx	x³/3 + C
Reciprocal	∫(1/x) dx	ln|x| + C
Exponential	∫eˣ dx	eˣ + C
	∫aˣ dx	aˣ/ln(a) + C
	∫ln(x) dx	x ln(x) − x + C
Trigonometry (x in radians)	∫cos(x) dx	sin(x) + C
	∫sin(x) dx	-cos(x) + C
	∫sec²(x) dx	tan(x) + C
Rules	**Function**	**Integral**
Multiplication by constant	∫cf(x) dx	c∫f(x) dx
Power Rule (n≠-1)	∫xⁿ dx	xⁿ⁺¹/(n+1) + C
Sum Rule	∫(f + g) dx	∫f dx + ∫g dx
Difference Rule	∫(f - g) dx	∫f dx - ∫g dx

Figure 3: Integration rules by mathisfun.com

11.1.1 Integration by parts

Sometimes, integration can't be solved by using simple rules, and we must do a partial integration. This is often the case when we have a multiplication/division of two functions containing the same variable (e.g. $\int x^3 * 2^x dx$).
For two functions $f(x)$ and $g(x)$ we can find the integral of the multiplication/-division of these by using the rule (Remember that $\frac{f(x)}{g(x)}$ can be written as the product $f(x) * \frac{1}{g(x)}$):

$$\boldsymbol{\int_a^b f(x)g(x)dx = [F(x)g(x)]_a^b - \int_a^b F(x)g'(x)dx}$$

- Where $f(x)$ and $g(x)$ are the functions x^3 and 2^x
- Where $F(x)$ is the antiderivate (stamfunktion) of $f(x)$, $F(x) = \int_a^b x^3 dx = \left[\frac{x^4}{4}\right]_a^b = F(b) - F(a)$
- Where $g'(x)$ is the derivative of $g(x)$, $g'(x) = \frac{d}{dx} 2^x = ln(2) * 2^x$

- **I**: Inverse trigonometric functions such as $\sin^{-1}(x)$, $\cos^{-1}(x)$, $\tan^{-1}(x)$
- **L**: Logarithmic functions such as $\ln(x)$, $\log(x)$
- **A**: Algebraic functions such as x^2, x^3
- **T**: Trigonometric functions such as $\sin(x)$, $\cos(x)$, $\tan(x)$
- **E**: Exponential functions such as e^x, 3^x

Figure 4. A litte list that could be helpful. You usually want to choose the function for $f(x)$ that is highest on this list to make your integration easier (thanks to mathisfun.com).

11.1.2 Integration by substituation

Sometimes, our integration can't be solved using the above rules, nor integration by parts. This could be the case for a function applied on a function, e.g. $f(g(x)) = cos(x^2)$ where $f(x) = cos()$ and $g(x) = x^2$. Mathisfun.com made a great guide how we proceed from this:

Integration by Substitution

"Integration by Substitution" (also called "u-substitution") is a method to find an integral, but only when it can be set up in a special way.

The first and most vital step is to be able to write our integral in this form:

$$\int f(g(x))\ g'(x)\ dx$$

Note that we have **g(x)** and its derivative **g'(x)**

Like in this example:

$$\int \cos(x^2)\ 2x\ dx$$

Here **f=cos**, and we have **g=x²** and its derivative of **2x**
This integral is good to go!

When our integral is set up like that, we can do **this substitution**:

$$\int f(\underbrace{g(x)}_{u})\ \underbrace{g'(x)\ dx}_{du}$$

$$\int f(u)\ du$$

Then we can **integrate f(u)**, and finish by **putting g(x) back as u**.

Like this:

Example: $\int \cos(x^2)\ 2x\ dx$

We know (from above) that it is in the right form to do the substitution:

$$\int \cos(x^2)\ 2x\ dx$$

$$\int \cos(u)\ du$$

Now integrate:

$$\int \cos(u)\ du = \sin(u) + C$$

And finally put $\mathbf{u=x^2}$ back again:

$$\sin(x^2) + C$$

So $\boldsymbol{\int \cos(x^2)\ 2x\ dx = \sin(x^2) + C}$ worked out really nicely! (Well, I knew it would.)

This method only works on some integrals of course, and it may need rearranging:

Example: $\int \cos(x^2)\ 6x\ dx$

Oh no! It is **6x**, not **2x**. Our perfect setup is gone.

Never fear! Just rearrange the integral like this:

$$\int \cos(x^2)\ 6x\ dx = 3\int \cos(x^2)\ 2x\ dx$$

(We can pull constant multipliers outside the integration, see Rules of Integration .)

Then go ahead as before:

$$3\int \cos(u)\ du = 3\ \sin(u) + C$$

Now put $\mathbf{u=x^2}$ back again:

$$3\ \sin(x^2) + C$$

Done!

Now we are ready for a slightly harder example:

Example: $\int x/(x^2+1)\, dx$

Let me see ... the derivative of x^2+1 is $2x$ ... so how about we rearrange it like this:

$$\int x/(x^2+1)\, dx = ½ \int 2x/(x^2+1)\, dx$$

Then we have:

$$½ \int \frac{2x}{x^2+1}\, dx$$

$$½ \int 1/(u)\, du$$

Then integrate:

$$½ \int 1/u\, du = ½ \ln(u) + C$$

Now put $\mathbf{u=x^2+1}$ back again:

$$½ \ln(x^2+1) + C$$

11.2 Differentiation

No in dept guide written for differentiation, but here's a cheet-sheet that should cover your most desirable needs:

Common Functions	Function	Derivative
Constant	c	0
Line	x	1
	ax	a
Square	x^2	$2x$
Square Root	$\sqrt{x}$	$(½)x^{-½}$
Exponential	e^x	e^x
	a^x	$\ln(a)\, a^x$
Logarithms	$\ln(x)$	$1/x$
	$\log_a(x)$	$1 / (x \ln(a))$
Trigonometry (x is in radians)	$\sin(x)$	$\cos(x)$
	$\cos(x)$	$-\sin(x)$
	$\tan(x)$	$\sec^2(x)$
Inverse Trigonometry	$\sin^{-1}(x)$	$1/\sqrt{(1-x^2)}$
	$\cos^{-1}(x)$	$-1/\sqrt{(1-x^2)}$
	$\tan^{-1}(x)$	$1/(1+x^2)$

Rules	Function	Derivative
Multiplication by constant	cf	cf'
Power Rule	x^n	nx^{n-1}
Sum Rule	$f + g$	$f' + g'$
Difference Rule	$f - g$	$f' - g'$
Product Rule	fg	$f g' + f' g$
Quotient Rule	f/g	$(f' g - g' f)/g^2$
Reciprocal Rule	$1/f$	$-f'/f^2$
Chain Rule (as "Composition of Functions")	$f \circ g$	$(f' \circ g) \times g'$
Chain Rule (using ′)	$f(g(x))$	$f'(g(x))g'(x)$
Chain Rule (using $\frac{d}{dx}$)	$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$	

***Figure 5.** Also $(e^{-x})' = -e^{-x}$, Thanks to mathsisfun.com for the picture.*