In [1]:
```
%matplotlib notebook
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import math
from scipy import stats
import scipy.stats as st
```

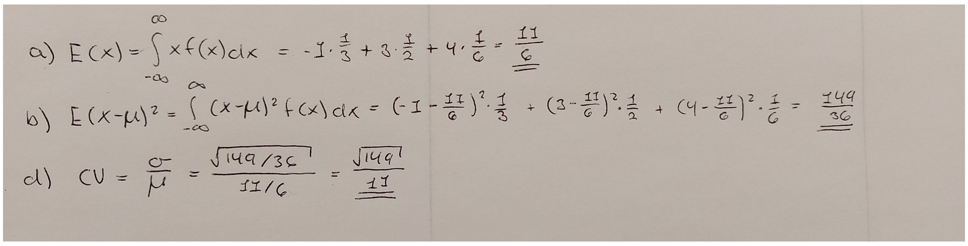
Eksamen GEO4300

candidate: 15410

Oppgave 1: Random variable parameter estimation

Answer a), b) and d) in the picture.

picture downloaded for Random variable



a) $E(x) = \int_{-\infty}^{\infty} x f(x) dx = -1 \cdot \frac{1}{3} + 3 \cdot \frac{1}{2} + 4 \cdot \frac{1}{6} = \frac{11}{6}$

b) $E(x-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = (-1 - \frac{11}{6})^2 \cdot \frac{1}{3} + (3 - \frac{11}{6})^2 \cdot \frac{1}{2} + (4 - \frac{11}{6})^2 \cdot \frac{1}{6} = \frac{149}{36}$

d) $CV = \frac{\sigma}{\mu} = \frac{\sqrt{149/36}}{11/6} = \frac{\sqrt{149}}{11}$

c) The mode is the most frequently occuring value in set of discrete data. For continous variables, the probability distribution function f(x) has a maximum for x=mode.

- Here the mode is x=3, for prob=1/2

picture downloaded for Formulas from the compendium used in the calculation:

TABLE 11.3.1
Population parameters and sample statistics

Population parameter	Sample statistic
1. *Midpoint*	
Arithmetic mean	
$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Median	
x such that $F(x) = 0.5$	50th-percentile value of data
Geometric mean	
antilog $[E(\log x)]$	$\left(\prod_{i=1}^{n} x_i\right)^{1/n}$
2. *Variability*	
Variance	
$\sigma^2 = E[(x-\mu)^2]$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard deviation	
$\sigma = \{E[(x-\mu)^2]\}^{1/2}$	$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}$
Coefficient of variation	
$CV = \frac{\sigma}{\mu}$	$CV = \frac{s}{\bar{x}}$
3. *Symmetry*	
Coefficient of skewness	
$\gamma = \frac{E[(x-\mu)^3]}{\sigma^3}$	$C_s = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s^3}$

The *variability* of data is measured by the *variance* σ^2, which is the second moment about the mean:

$$E[(x-\mu)^2] = \sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\,dx \qquad (11.3.3)$$

Oppgave 2: Frequency analysis and linear regression

In [13]:
```
# a) What is the probability to observe at least one 100-years floo
d or larger within a period of 10 years?
# Look at the changes for not 1-'not happening'
n = 10
flood = 100

p = 1 - (1-(1/flood))**n

print ('probability for at least one 100 years flood within a 10 ye
ars periode is ',np.round(p,3))
```

```
probability for at least one 100 years flood within a 10 years per
iode is  0.096
```

b) Describe which assumption of a simple linear regression is violated in this analysis, and discuss strategies that can be used to improve the analysis.

There are four assumptions associated with a linear regression model:

- Linearity: The relationship between X and the mean of Y is linear.
- Homoscedasticity: The variance of residual is the same for any value of X.
- Independence: Observations are independent of each other.
- Normality: For any fixed value of X, Y is normally distributed.

Here: the homoscedasticity is violated. For large values are the spread is larger (large disturbance), should be equal spread around the linear regression.

Strategies to improve the analysis could be to only use the data where the spread is equal around the linear regression line. Want the QQ-plot to be linear, thus the data is normal distributed.

Oppgave 3: Confidence intervals

In [12]:
```
n = 30
mean = 145
var = 20

# a) What is the 95% confidence interval on the mean assuming a nor
mal distribution if
#(i)  the true variance is unknown and estimated as 20

a = 1-0.95
t = stats.t.ppf(1-a/2,n-1)

L = mean - t*(np.sqrt(var)/np.sqrt(n))
U = mean + t*(np.sqrt(var)/np.sqrt(n))
print ('lower interval is ',np.round(L,2),' and upper interval is '
,np.round(U,2),' without known variance.')

#(ii)  the true variance is 20

z = stats.norm.ppf(1-a/2)

L = mean - z*(np.sqrt(var)/np.sqrt(n))
U = mean + z*(np.sqrt(var)/np.sqrt(n))
print ('lower interval is ',np.round(L,3),' and upper interval is '
,np.round(U,3),' with known variance.')
```

```
lower interval is  143.33  and upper interval is  146.67  without
known variance.
lower interval is  143.4  and upper interval is  146.6  with known
variance.
```

b) What is the reason for the difference of results in part (i) and part (ii)?

- The reason for the different result are that the t-test have a longer tail than in the z-test and the spread of t-test are larger than the spread of z-test. t-test is for small sample size and unknown variance and z-test if for large sample size and known variance. With unknown variance one need to be more carefull, higher uncertianty.

picture downloaded form copendium with normal distribution (solide line) and t-distribution (dottet line). The red box indicate the 'tail' of the two distributions.

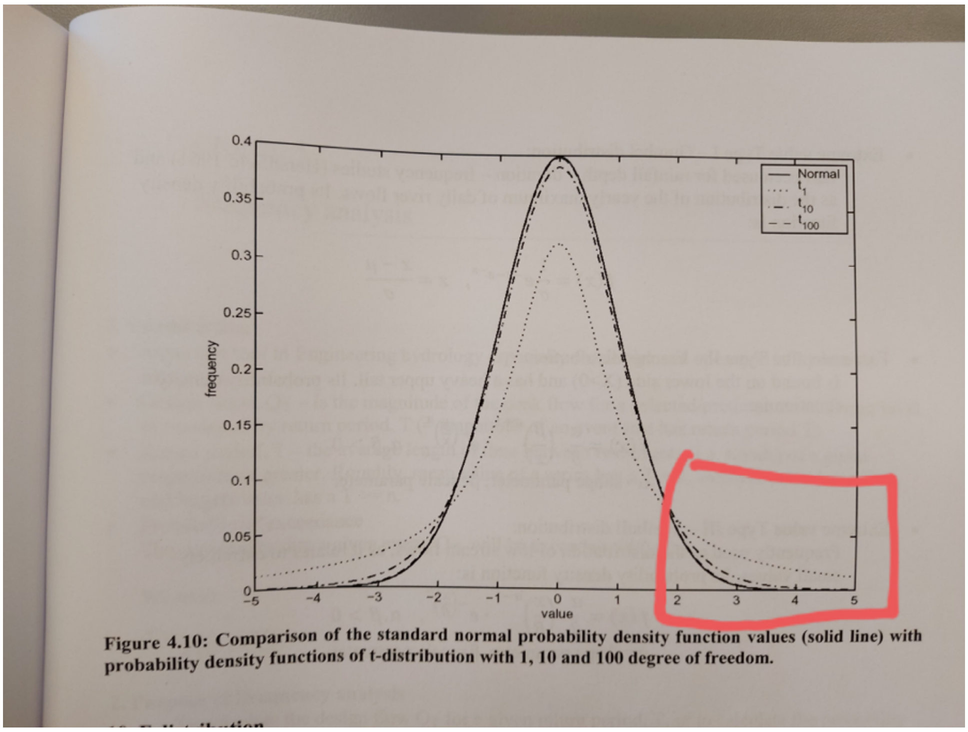


Figure 4.10: Comparison of the standard normal probability density function values (solid line) with probability density functions of t-distribution with 1, 10 and 100 degree of freedom.

In [8]:
```
#c)  What is the 95% confidence interval on the variance?
x_a = stats.chi2.ppf(a/2, n-1)
x_1_a = stats.chi2.ppf(1-a/2, n-1)

L = ((n-1)*np.sqrt(var)**2)/x_1_a
U = ((n-1)*np.sqrt(var)**2)/x_a

print ('The lower interval is ',np.round(L,2),' and upper interval is ',np.round(U,2),' for the variance.')
```

```
The lower interval is  12.69  and upper interval is  36.14  for the variance.
```

Oppgave 4: Machine learning

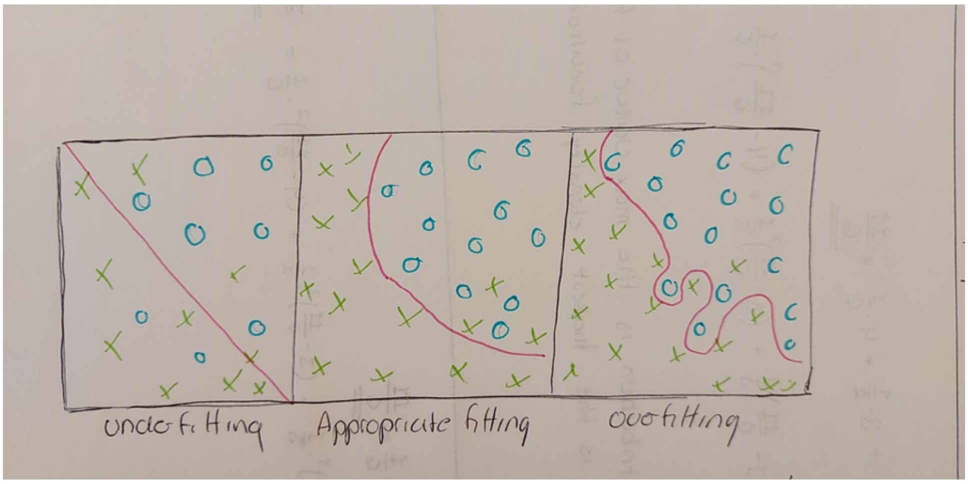
a) Why is it common to split the dataset into a training set and a test set when doing machine learning? In your answer, include in a relevant way the terms training error and test error.

- It is commen to split the data set into training (2/3 of the data) and test (1/3 of the data) set when doing machine learning. In machinlearning you wanna learn your model to map from inputs to outputs based on traning data. You use the training data to learn you model how you want the out put to be of the given input, then you test if it manage to get the same conection with in- and out-put from new data, your test data, to see if the model works. Your model learn by reducing measures of the training error. The training error underestimate the test error and is the error that occures when running the model of the same data that it was trained on. Test error is the error that we incur on new data and is the error that acctual gives an indication on how well the model wil do on future data the model hasn't seen yet.

b) In many machine learning algorithms you have a parameter that controls the complexity of the model. Why do we want to control this complexity?

- Complexity in machine learning is the number of features or terms included in a given predictive model. Models with high capacity (complexity) may over fit and generalize badly even though the model have minimized the training error. These are also more likely be more computationally expensive. If the model is made to simple, it's underfitting and we get bad generalization with large errors. We want to control the complexity so we can adjust our model and find the 'sweet spot'; the compelxity that gives the smallest error.

picture downloaded for complxity in machinlearning.



Oppgave 5: Time series analysis and Fourier transformation

a) How could you test if there is a significant trend in Xt? Explain a suitable test.

- A suitable test to check if there is a significant trend in Xt is the Ordinary Least Squares methode for Simple Linear Regression. OSL gives parameters of a linear function from a set of variables by using the principle of least squares which is minimizing the sum of the squares of the differences between the observed variable in the given dataset and those predicted by the linear function. These are used to do the Smiple Linear Regression. Simple Linear Regression is a statistical model based on the idea that the relationship between two variables can be explained by the following formula: $Y_i = \alpha + \beta X_i$. The idea of Simple Linear Regression is to find the parameters α and β for which the error term is minimized so you can fit a 'straight line' which is as close as possible to your data points. The best fitted parameters to the data are given by the OSL.
- In phyton one can to use the function: sm.OLS(year,sm.add_constant(x)) that gives out the 95% confident interval from a student t-tast (small sample size and unknown variance), it the confident interval changes sign there is not a significant trend, if they do not change sign there is a trend. The function also give the Beta coffecient, the slope, if there is a trend, the trend is positive for beta positiv, and negative for beta negative.

In [11]:
```
# example on OSL sm.OLS(year,sm.add_constant(x)) output:
from statsmodels.tools.tools import add_constant
import statsmodels.api as sm

x = [0.5, 1, -0.5, 1.4, 0.8, 1.5, -0.4, 1, 0.7, -0.5, 0, 0.6] #would have to read the data points better, but just an example on how it works
year = np.arange(12)

model = sm.OLS(year,sm.add_constant(x))
results = model.fit()
results.summary()


# The alpha is the constant[coef]
# The beta is the x1[coef]
# The confident interval is x1 [   ] [   ] --> here it change sign and are not significant, hence no trend.

#print(results.t_test([1, 0]))
```

Out[11]:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.051
Model:	OLS	**Adj. R-squared:**	-0.044
Method:	Least Squares	**F-statistic:**	0.5394
Date:	Mon, 23 Nov 2020	**Prob (F-statistic):**	0.480
Time:	10:24:20	**Log-Likelihood:**	-31.580
No. Observations:	12	**AIC:**	67.16
Df Residuals:	10	**BIC:**	68.13
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	6.0859	1.329	4.578	0.001	3.124	9.048
x1	-1.1526	1.569	-0.734	0.480	-4.649	2.344

Omnibus:	0.714	**Durbin-Watson:**	0.226
Prob(Omnibus):	0.700	**Jarque-Bera (JB):**	0.600
Skew:	-0.123	**Prob(JB):**	0.741
Kurtosis:	1.933	**Cond. No.**	2.05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

b) The following three graphs show the absolute values for Fourier coefficients, defined as: (see exam) Which one of them (A, B or C) shows the Fourier transform of Xt? Explain your answer

- The right answer is A, beacuse the time series show one peak at every 5 sek, the frequency then become f=1/T=1/5=0.2, where we can find the peaks in A. We also have symmerty around the periode; T=5 in A.

In [ ]: