Statistics: A Primer

Lecture BigData Analytics

Julian M. Kunkel

julian.kunkel@googlemail.com

University of Hamburg / German Climate Computing Center (DKRZ)

2016-10-28

Disclaimer: Big Data software is constantly updated, code samples may be outdated.

Outline

Statistics: Overview

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data [21]

Either **describe** properties of a sample or **infer** properties of a population

Important terms [10]

- **Unit of observation**: the entity described by the data
- **Unit of analysis**: the major entity that is being analyzed
  - Example: Observe income of each person, analyse differences of countries
- **Statistical population**: Complete set of items that share at least one property that is subject of analysis
  - Subpopulation share additional properties
- **Sample**: (sub)set of data collected and/or selected from a population
  - If chosen properly, they can represent the population
  - There are many sampling methods, we can never capture ALL items
- **Independence**: one observation does not effect another
  - Example: select two people living in Germany randomly
  - Dependent: select one household and pick a married couple

Statistics: Variables

- **Dependent variable**: represents the output/effect
  - Example: word count of a Wikipedia article; income of people
- **Independent variable**: assumed input/cause/explanation
  - Example: number of sentences; age, educational level

Characterization

- **Univariate analysis**: characterize a single variable
- **Bivariate analysis**: describes/analyze relationships between two vars
- **Multivariate statistics**: analyze/observe multiple dependent variables
  - Example: chemicals in the blood stream of people or chance for cancers, Independent variables are personal information/habits

Descriptive Statistics [10]

- The discipline of quantitatively describing main features of sampled data
  - Summarize observations/selected samples
- **Exploratory data analysis** (EDA): approach for inspecting data
  - Using different chart types, e.g., box plots, histograms, scatter plot
- Methods for univariate analysis
  - Distribution of values, e.g., mean, variance, quantiles
  - Probability distribution and density
  - t-test (e.g., check if data is t-distributed)
- Methods for bivariate analysis
  - Correlation coefficient[1] describes linear relationship
  - Rank correlation[2]: extent by which one variable increases with another var
- Methods for multivariate analysis
  - Principal component analysis (PCA) converts correlated variables into linearly uncorrelated variables called principal components

[1] Pearson's product-moment coefficient

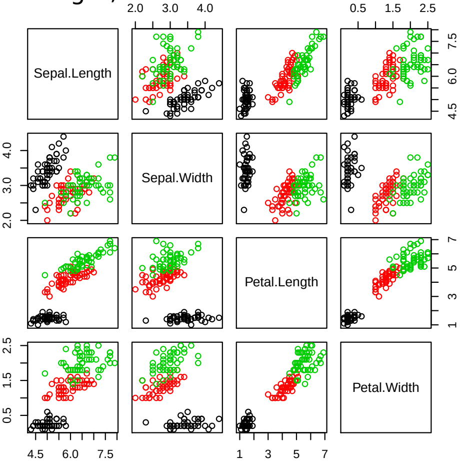
[2] By Spearman or Kendall

Example Dataset: Iris Flower Data Set

- Contains information about iris flower
- Three species: Iris Setosa, Iris Virginica, Iris Versicolor
- Data: Sepal.length, Sepal.width, Petal.length, Petal.width

R example

```
> data(iris) # load iris data
> summary(iris)
Sepal.Length    Sepal.Width     Petal.Length
Min.   :4.300   Min.   :2.000   Min.   :1.000
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600
Median :5.800   Median :3.000   Median :4.350
Mean   :5.843   Mean   :3.057   Mean   :3.758
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100
Max.   :7.900   Max.   :4.400   Max.   :6.900

Petal.Width          Species
Min.   :0.100    setosa    :50
1st Qu.:0.300    versicolor:50
Median :1.300    virginica :50
Mean   :1.199
3rd Qu.:1.800
Max.   :2.500

# Draw a matrix of all variables
> plot(iris[,1:4], col=iris$Species)
```



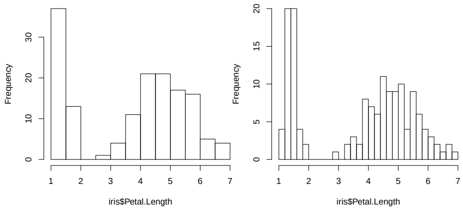
R plot of the iris data

Distribution of Values: Histograms [10]

- Distribution: frequency of outcomes (values) in a sample
  - Example: Species in the Iris data set
    - setosa: 50
    - versicolor: 50
    - virginica: 50
- Histogram: graphical representation of the distribution
  - Partition observed values into bins
  - Count number of occurrences in each bin
  - It is an estimate for the probability distribution

R example

```
# nclass specifies the number of bins
# by default, hist uses equidistant bins
hist(iris$Petal.Length, nclass=10, main="")

hist(iris$Petal.Length, nclass=25, main="")
```



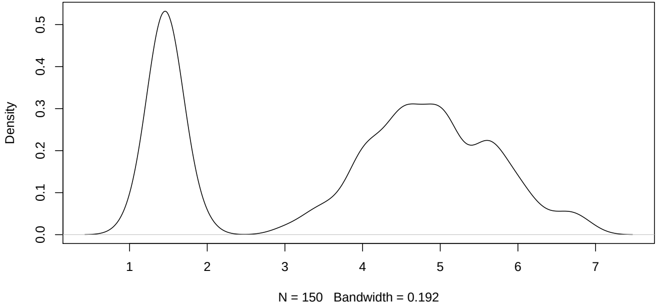
Histograms with 10 and 25 bins

Distribution of Values: Density [10]

- Probability density function (density):
  - Likelihood for a **continuous** variable to take on a given value
  - Kernel density estimation (KDE) approximates the density

R example

```r
# The kernel density estimator moves a function (kernel) in a window across samples
# With bw="SJ" or "nrd" it automatically determines the bandwidth, i.e., window size
d = density(iris$Petal.Length, bw="SJ", kernel="gaussian")
plot(d, main="")
```



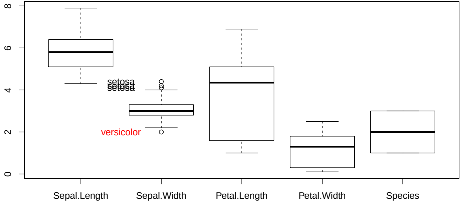
Density estimation of Petal.Length

Distribution of Values: Quantiles [10]

- Percentile: value below which a given percentage of observations fall
- q-Quantiles: values that partition a ranked set into q equal sized subsets
- **Quartiles**: 3 data points that split a ranked set into 4 equal points
  - Q1=P(25), Q2=median=P(50), Q3=P(75), **interquartile range** iqr=Q3-Q1
- **Five-number summary**: (min, Q1, Q2, Q3, max)
- Boxplot: shows quartiles (Q1,Q2,Q3) and whiskers
  - Whiskers extend to values up to 1.5 iqr from Q1 and Q3
  - Outliers are outside of whiskers

R example

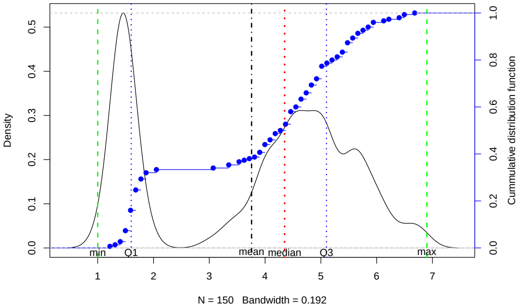
```
> boxplot(iris, range=1.5) # 1.5 interquartile range
> d = iris$Sepal.Width
> quantile(d)
  0%  25%  50%  75% 100%
 2.0  2.8  3.0  3.3  4.4
> q3 = quantile(d,0.75) # pick value below which are 75%
> q1 = quantile(d,0.25)
> irq = (q3 - q1)
# identify all outliers based on the interquartile range
> mask = d < (q1 - 1.5*irq) | d > (q3 + 1.5*irq)
# pick outlier selection from full data set
> o = iris[mask,]
# draw the species name of the outliers on the boxplot
> text(rep(1.5,nrow(o)), o$Sepal.Width, o$Species,
        ↪ col=as.numeric(o$Species))
```



Boxplot

Density Plot Including Summary

```
d = density(iris$Petal.Length, bw="SJ", kernel="gaussian")

# add space for two axes
par(mar=c(5, 4, 4, 6) + 0.1)
plot(d, main="")
# draw lines for Q1, Q2, Q3
q = quantile(iris$Petal.Length)
q = c(q, mean(iris$Petal.Length))
abline(v=q[1], lty=2, col="green", lwd=2)
abline(v=q[2], lty=3, col="blue", lwd=2)
abline(v=q[3], lty=3, col="red", lwd=3)
abline(v=q[4], lty=3, col="blue", lwd=2)
abline(v=q[5], lty=2, col="green", lwd=2)
abline(v=q[6], lty=4, col="black", lwd=2)
# Add titles
text(q, rep(-0.01, 5), c("min", "Q1", "median", "Q3", "max", "mean"))

# identify x limits
xlim = par("usr")[1:2]
par(new=TRUE)

# Empirical cummulative distribution function
e = ecdf(iris$Petal.Length)
plot(e, col="blue", axes=FALSE, xlim=xlim, ylab="", xlab="", main="")

axis(4, ylim=c(0,1.0), col="blue")
mtext("Cummulative distribution function", side=4, line=2.5)
```



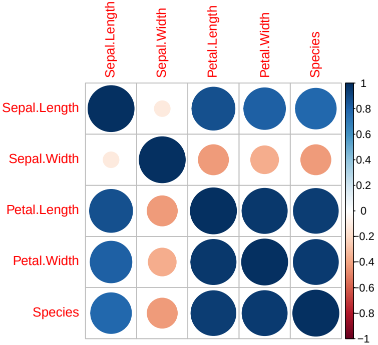
Density estimation with 5-number summary and cumulative density function

Correlation Coefficients

- Measures (linear) correlation between two variables
  - Result is between -1 and +1
    - >0.7: strong positive correlation
    - >0.2: weak positive correlation
    - 0: no correlation, < 0: negative correlation

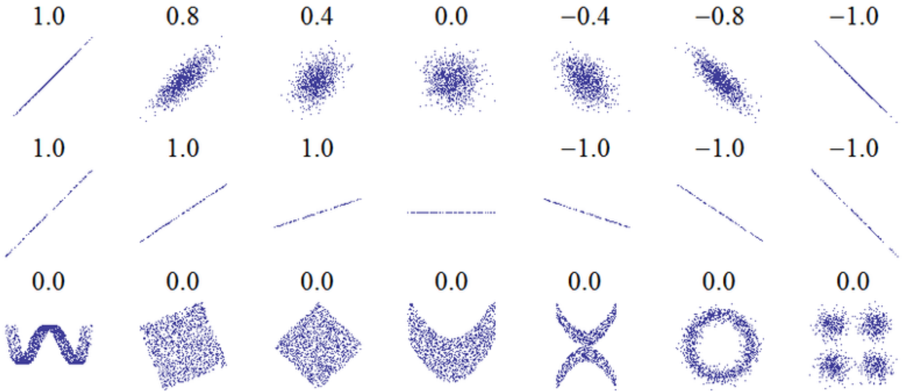
R example

```
library(corrplot)
d = iris
d$Species = as.numeric(d$Species) # It is normally not adviced to convert categorial data to numeric
corrplot(cor(d), method = "circle") # the right plot

mplot = function(x,y, name){
  pdf(name,width=5,height=5) # plot into a PDF
  p = cor(x,y, method="pearson") # compute correlation
  k = cor(x,y, method="spearman")
  plot(x,y, xlab=sprintf("x\n cor. coeff: %.2f rank coef.: %.2f", p, k))
  dev.off()
}

mplot(iris$Petal.Length, iris$Petal.Width, "iris-corr.pdf")
# cor. coeff: 0.96 rank coef.: 0.94

x = 1:10; y = c(1,3,2,5,4,7,6,9,8,10)
mplot(x,y, "linear.pdf") # cor. coeff: 0.95 rank coef.: 0.95

mplot(x, x*x*x , "x3.pdf") # cor. coeff: 0.93 rank coef.: 1
```



Corrplot

Example Correlations for Plots of Two Variables (X, Y)



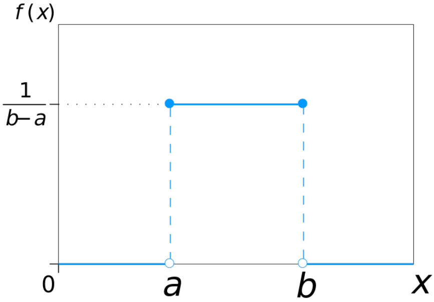
Correlations for x,y plots; Source: [22]

Distribution of Values

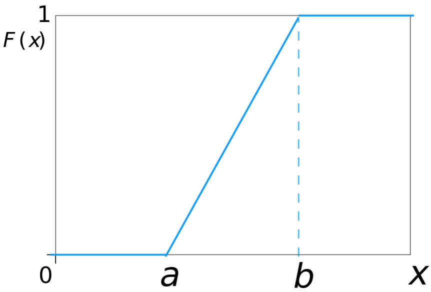
- Data exploration helps to understand distribution of values for a variable
- To illustrate data distribution, statistics uses:
  - Probability density function (PDF): value and probability; $P(x)$
  - Cumulative distribution function (CDF): sum from $-\infty$ to value; $P(X \leq x)$
- Many standard distributions exists that can be models for processes
  - Discrete probability functions are drawn from a limited set of values
  - Continuous probability functions are $\in \mathbb{R}$
- When we explore data, we may identify these common distributions
- Discrete probability functions are derived from useful processes

(Continuous) Uniform Distribution

- Notation: unif(a, b)
- Mean: $\frac{a+b}{2}$
- Variance: $\frac{(b-a)^2}{12}$



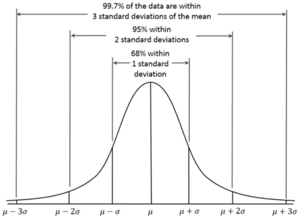
(a) Source: IkamusumeFan, PDF of the uniform probability distribution (10)



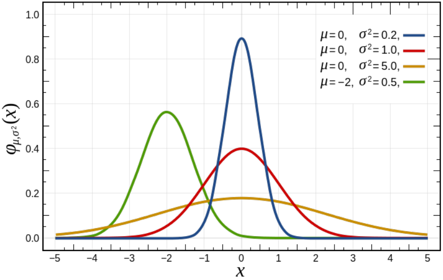
(b) Source: IkamusumeFan, CDF of the uniform probability distribution (10)

Normal Distribution

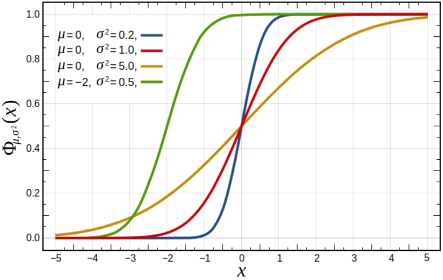
- Notation: $\mathcal{N}(\mu, \sigma^2)$
- Mean: μ
- Variance: σ^2



Source: Dan Kernler, For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%. [10]



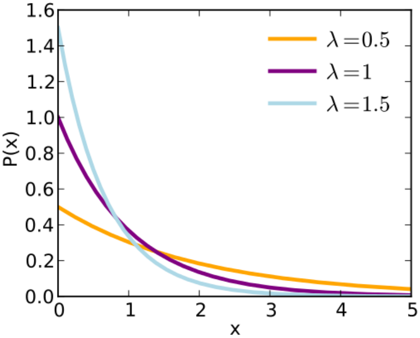
(a) Source: Inductiveload, Probability density function for the normal distribution (10)



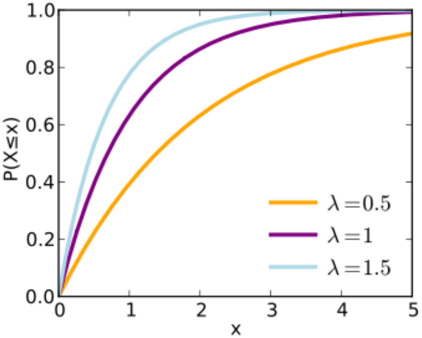
(b) Source: Inductiveload, Cumulative distribution function for the normal distribution (10)

Exponential Distribution

- Notation: Exponential(λ)
  - λ (rate) > 0
- Mean: λ^{-1}
- Variance: λ^{-2}
- Models inter-arrival time of a Poisson process
- Memoryless, useful to model failure rates



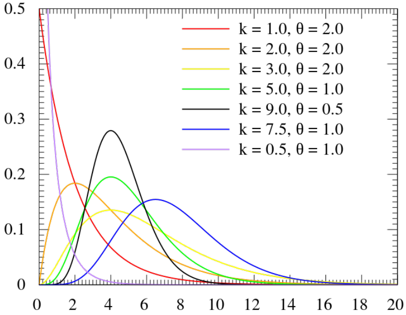
(c) Source: Skbkekas, Probability density function (10)



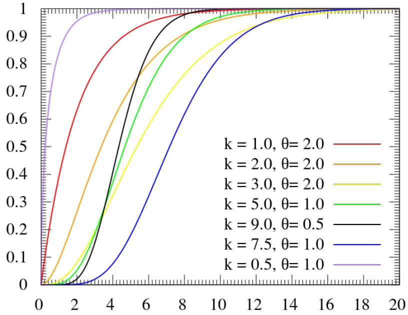
(d) Source: Skbkekas, Cumulative distribution (10)

Gamma Distribution

- Notation: Gamma(α, β)
  - α (shape), β (rate) > 0
- Mean: $\frac{\alpha}{\beta}$
- Variance: $\frac{\alpha}{\beta^2}$



(e) MarkSweep and Cburnett, Probability density plots of gamma distributions (10)

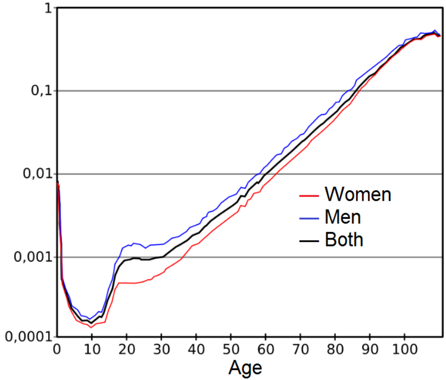


(f) MarkSweep and Cburnett, Cumulative distribution plots of gamma distributions (10)

Real Data

- Observations can be based on several (unknown) processes

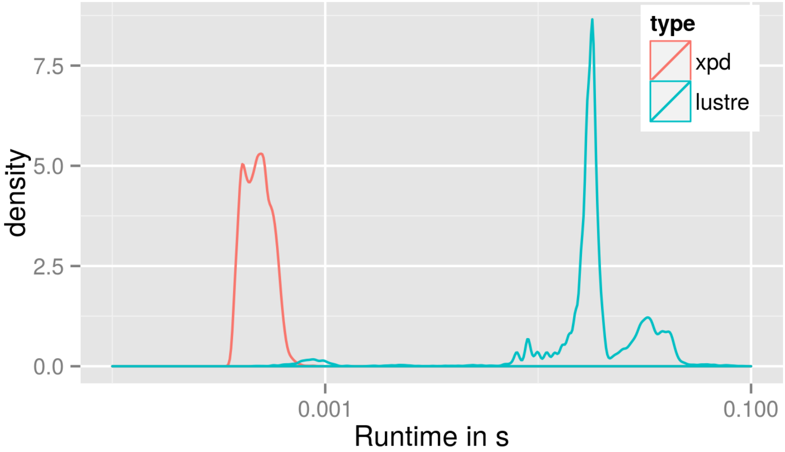
Fraction of people of a given age that died in 1999



Mark Bailin, Mortality by age [10]

Real Data (2)

Performance of reading 1 MiB of data from two different storage systems



Lustre parallel file systems vs. pooled memory of the XPD

Inductive Statistics: Some Terminology [10]

- Statistical inference is the process of **deducting properties** of a population by analyzing samples
  - Build a statistical model and test the hypothesis if it applies
  - Allows to deduct propositions (statements about data properties)
- **Statistical hypothesis**: hypothesis that is testable on a process modeled via a set of random variables
- **Statistical model**: embodies a set of assumptions concerning the generation of the observed data, and similar data from a larger population. A model represents, often in considerably idealized form, the data-generating process
- **Validation**: process to verify that a model/hypothesis is likely to represent the observation/population
- **Significance**: a significant finding is one that is determined (statistically) to be very unlikely to happen by chance
- **Residual**: difference of observation and estimated/predicted value

Statistics: Inductive Statistics [10]

Testing process

1. Formulate default (null[3]) and alternative hypothesis
2. Formulate statistical assumptions, e.g., independence of variables
3. Decide which statistical tests can be applied to disprove null hypothesis
4. Choose significance level α for wrongly rejecting null hypothesis
5. Compute test results, especially the **p-value**[4]
6. If p-value $< \alpha$, then reject null hypothesis and go for alternative
   - Be careful: (p-value $\geq \alpha$) $\not\Rightarrow$ null hypothesis is true, though it may be

Example hypotheses

- Petal.Width of each iris flowers species follow a normal distribution
- Waiting time of a supermarket checkout queue is gamma distributed

[3] We try to reject/**null**ify this hypothesis.

[4] Probability of obtaining a result equal or more extreme than observed.

Checking if Petal.Width is Normal Distributed

R example

```
# The Shapiro-Wilk-Test allows for testing if a population represented by a sample is normal distributed
# The Null-hypothesis claims that data is normal distributed

# Let us check for the full population
> shapiro.test(iris$Petal.Width)
# W = 0.9018, p-value = 1.68e-08
# Value is almost 0, thus reject null hypothesis => in the full population, Petal.Width is not normal distributed

# Maybe the Petal.Width is normal distributed for individual species?
for (spec in levels(iris$Species)){
  print(spec)
  y = iris[iris$Species==spec,]

  # Shapiro-Wilk-test checks if data is normal distributed
  print(shapiro.test(y$Petal.Width))
}

[1] "virginica"
W = 0.9598, p-value = 0.08695
# Small p-value means a low chance this happens, here about 8.7%
# With the typical significance level of 0.05 Petal.Width is normal distributed
# For simplicitly, we may now assume Petal.Width is normal distributed for this species

[1] "setosa"
W = 0.7998, p-value = 8.659e-07 # it is not normal distributed

[1] "versicolor"
W = 0.9476, p-value = 0.02728 # still too unlikely to be normal distributed
```

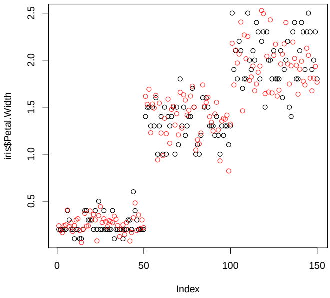
Linear Models (for Regression) [10]

- **Linear regression**: Modeling the relationship between dependent var Y and explanatory variables X_i
- Assume n samples are observed with their values in the tuples $(Y_i, X_{i1}, ..., X_{ip})$
  - Y_i is the dependent variable (label)
  - X_{ij} are independent variables
  - Assumption for linear models: normal distributed variables
- A linear regression model fits $Y_i = c_0 + c_1 \cdot f_1(X_{i1}) + ... + c_p \cdot f_p(X_{ip}) + \epsilon_i$
  - Determine coefficients c_0 to c_p to minimize the error term ϵ
  - The functions f_i can be non-linear

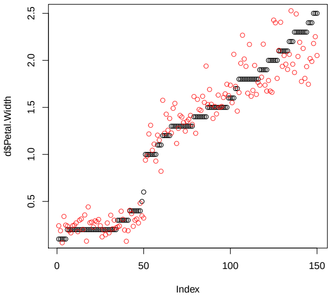
R example

```r
# R allows to define equations, here Petal.Width is our dependent var
m = lm("Petal.Width ~ Petal.Length + Sepal.Width", data=iris)

print(m) # print coefficients
# (Intercept)  Petal.Length   Sepal.Width
#     -0.7065        0.4263        0.0994
# So Petal.Width = -0.7065 + 0.4263 * Petal.Length + 0.0994 * Sepal.Width
```

Compare Prediction with Observation



Iris linear model



With sorted data

```
# Predict petal.width for a given petal.length and sepal.width
d = predict(m, iris)

# Add prediction to our data frame
iris$prediction = d

# Plot the differences
plot(iris$Petal.Width, col="black")
points(iris$prediction, col=rgb(1,0,0,alpha=0.8))

# Sort observations
d = iris[sort(iris$Petal.Width, index.return=TRUE)$ix,]
plot(d$Petal.Width, col="black")
points(d$prediction, col=rgb(1,0,0,alpha=0.8))
```

Analysing Model Accuracy [23]

- Std. error of the estimate: variability of c_i, should be lower than c_i
- t-value: measures how useful a variable is for the model
- $Pr(>|t|)$ two-sided p-value: probability that the variable is not significant
- Degrees of freedom: number of independent samples (avoid overfitting!)
- R-squared: fraction of variance explained by the model, 1 is optimal
- F-statistic: the f-test analyses the model goodness – high value is good

```
summary(m) # Provide detailed information about the model
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.53907 -0.11443 -0.01447  0.12168  0.65419
#
# Coefficients:
#              Estimate  Std.Error t value  Pr(>|t|)
# (Intercept)  -0.70648   0.15133  -4.668   6.78e-06 ***
# Petal.Length  0.42627   0.01045  40.804    < 2e-16 ***
# Sepal.Width   0.09940   0.04231   2.349     0.0201 *
# ---
# Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 " " 1
#
# Residual standard error: 0.2034 on 147 degrees of freedom
# Multiple R-squared:  0.9297,  Adjusted R-squared:  0.9288
# F-statistic: 972.7 on 2 and 147 DF,  p-value: < 2.2e-16
```

- Akaike's Information Criterion (AIC)
- Idea: prefer accurate models with smaller number of parameters
- Test various models to reduce AIC
- Improve good candidates
- AIC allows to check which models can be excluded

Time Series

- A time series is a sequence of observations
  - e.g., temperature, or stock price over time
  - Prediction of the future behavior is of high interest
- An observation may depend on any previous observation
  - Trend: tendency in the data
  - Seasonality: periodic variation

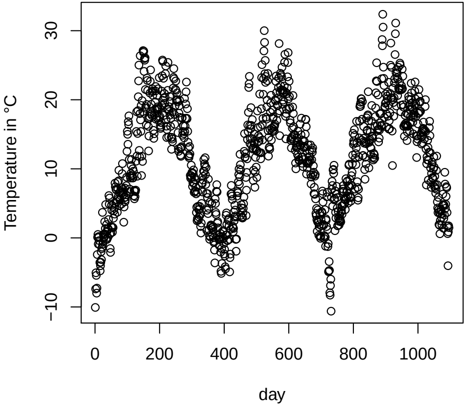
Prediction models

- Autoregressive models: AR(p)
  - Depend linearly on last p values (+ white noise)
- Moving average models: MA(q)
  - Random shocks: Depend linearly on last q white noise terms (+ white noise)
- Autoregressive moving average (ARMA) models
  - Combine AR and MA models
- Autoregressive integrated moving average: ARIMA(p, d, q)
  - Combines AR, MA and differencing (seasonal) models

Example Time Series

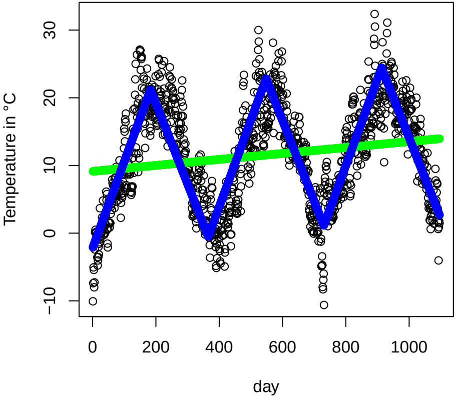
- Temperature in Hamburg every day at 12:00
- Three years of data (1980, 1996, 2014)

```r
d = read.csv("temp-hamburg.csv", header=TRUE)
d$Lat = NULL; d$Lon = NULL
colnames(d) = c("h", "t")
d$t = d$t - 273.15 # convert degree Kelvin to Celcius
plot(d$t, xlab="day", ylab="Temperature in C")

pdf("hamburg-temp-models.pdf", width=5,height=5)
plot(d$t, xlab="day", ylab="Temperature in C")

# Enumerate values
d$index=1:nrow(d)

# General trend
m = lm("t ~ index", data=d)
points(predict(m, d), col="green")

# Summer/Winter model per day of the year
d$day=c(rep(c(1:183, 182:1),3),0)
m = lm("t ~ day + index", data=d)
points(predict(m, d), col="blue");

library(forecast)
# Convert data to a time series
ts = ts(d$t, frequency = 365, start= c(1980, 1))
# Apply a model for non-seasonal data on seasonal adjusted data
tsmod = stlm(ts, modelfunction=ar)
plot(forecast(tsmod, h=365))
```

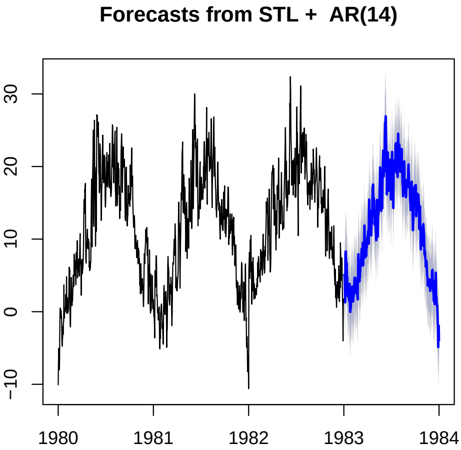


Temperature time series

Example Time Series Models



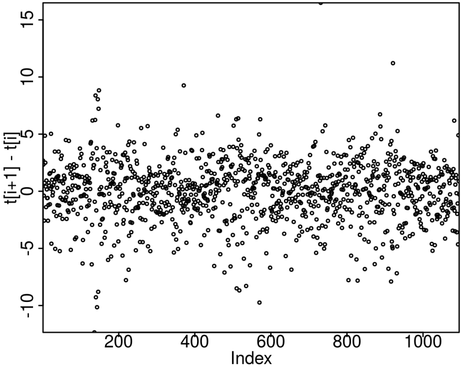
Linear models for trend & winter/summer cycle



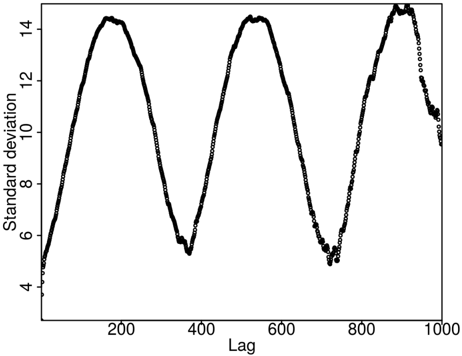
Using the forecast package/stlm()

Lag Plots

- Lag operator returns the previous element ($LX_t = X_{t-1}$) [4]



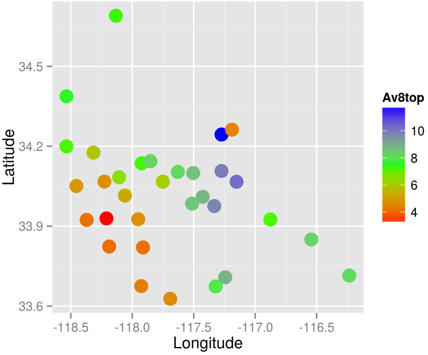
Lag plot, difference for Lag 1 for each element



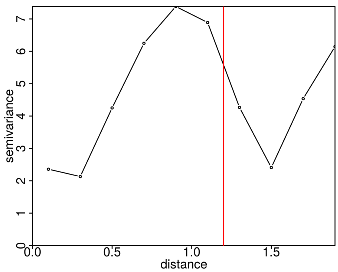
Standard deviation for all lags

Modeling Spatial Data

- Assume we measure data (z) on a surface (x,y), e.g., oil reservoirs
- **Kridging**: "*methods are statistical estimation algorithms that curve-fit known point data to produce a predictive surface*" [2, p.116]
- Variogram: variance between (x,y) pairs depending on the distance



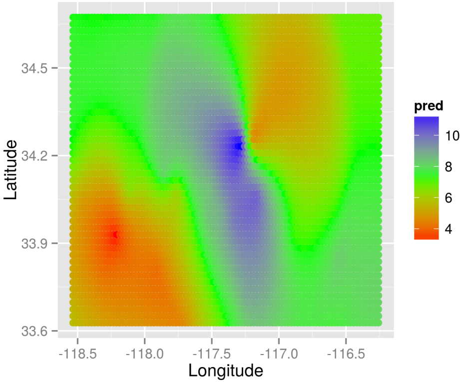
Stations and their observed ozone values



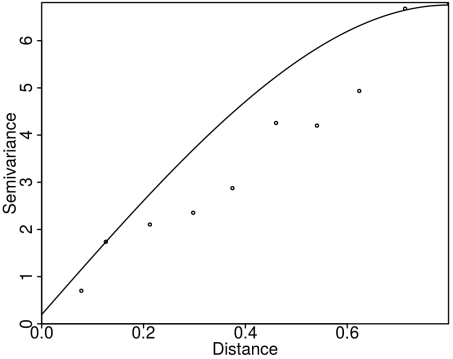
Estimated variogram with bins. A distance > red line, only green stations are compared

Kridging with Different Methods

- Models: linear, exponential, Gaussian, spherical



Spherical model



Spherical model prediction and actual variance

Summary

- Statistics allows us to describe propertiers and infer properties
  - Sample: subset of data
  - Population: complete set of items that is subject to analysis
  - Independent vs. dependent variables
- Descriptive statistics helps analyzing samples
  - Univariate: 5-number summary, quantiles, histograms, boxplots, density
  - Bivariate: correlation of two variables
- Inductive statistics provide concepts for inferring knowledge
  - Aim: reject null-hypothesis
  - Careful investigation of model accuracy is needed
  - Methods have certain requirements to data distribution
  - Regression method with linear models
  - Time series: variables depend on previous state
  - Geospatial data is treated with Kridging
- More about prediction in the machine learning lecture

Bibliography

1 Book: Statistik macchiato – Cartoon-Stochastikkurs für Schüler und Studenten. Andreas Lindenberg, Irmgard Wagner. Pearson.

2 Book: Data Science for Dummies. Lillian Pierson. A Wiley Brand.

4 Lag operator, Wikipedia. https://en.wikipedia.org/wiki/Lag_operator

10 Wikipedia

21 Book: Y. Dodge. The Oxford Dictionary of Statistical Terms. ISBN 0-19-920613-9

22 https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

23 http://blog.yhathq.com/posts/r-lm-summary.html

24 Forecasting: principles and practise https://www.otexts.org/fpp