¡article¿
¡presentation¿ Outline
Course Overview
Course Overview
This course is the first part of a two semester introductory course on Mathematical Statistics.
Our goal is to cover Chapters 1-10 of "Mathematical Statistics and Data Analysis", by John A. Rice.
Topics include: Probability, Random Variables, Discrete and Continuous distributions, Order Statistics, Limit Theorems
Roughly speaking, 4450 and 4451 can be broken into two parts:
Math 4450: Probability (mathematics of randomness)
Math 4451: Statistics (procedures for analyzing data)
Course Logistics
About Me (TODO)
Course Website: https://jeswheel.github.io/4450_f25/https://jeswheel.github.io/4450_f25/.
Canvas: Canvas will be used to submit assignments, view grades, and for course announcments.
https://jeswheel.github.io/4450_f25/syllabus.pdfCourse Syllabus
Probability: Chapter I Overview
Probability has been around for a long time.
Probability theory originated in the study of games of chance (i.e., dice, cards, etc.). These provide some nice introduct
More modern examples of probability in practice include:
Modeling mutations in genetics, playing a central role in bioinformatics.
Designing and analyzing computer operating systems.
Modeling atmospheric turbulence.
Probability theory is a cornerstone of the theory of finance, machine learning, and artificial intelligence.
Much more...
This semester will focus on the theory of probability as a mathematical model for chance phenomena. This will be essen
Looking for trends and relationships in dependent data
The first half of this course focuses on:
Quantifying dependence in time series data.
Finding statistical arguments for the presence or absence of associations that are valid in situations with dependence.
Example questions: Does Michigan show evidence for global warming? Does Michigan follow global trends, or is there e
Modeling and statistical inference for dynamic systems
The second half of this course focuses on:
Building models for dynamic systems, which may or may not be linear and Gaussian.
Using time series data to carry out statistical inference on these models.
Example questions: Can we develop a better model for understanding variability of financial markets (known in finance
Example: Winter in Michigan
Course files on Github
[fragile]Example: Winter in Michigan
There is a temptation to attribute a warm winter to global warming. You can then struggle to explain a subsequent col
You can get this file from the https://github.com/ionides/531w24course repository on GitHub.
Better, you can make a local clone of this git repository that will give you an up-to-date copy of all the data, notes, cod
shadecolorrgb0.969, 0.969, 0.969fgcolor y ¡- read.table(file="ann$_arbor_weather.csv$",$header$=1)
Rmarkdown and knitr
Rmarkdown and knitr
The notes combine source code with text, to generate statistical analysis that is
Reproducible
Easily modified or extended
These two properties are useful for developing your own statistical research projects. Also, they are useful for teaching a
Many of you will already know `Rmarkdown` (Rmd format) and/or Jupyter notebooks.
`knitr` (Rnw format) is similar, and is also supported by Rstudio. The notes are in Rnw, since it is superior for combinin
Rmd naturally produces html.
Some basic investigation using R
[fragile]
To get a first look at our dataset, `str` summarizes its structure:
shadecolorrgb0.969, 0.969, 0.969fgcolor str(y)

```
'data.frame': 124 obs. of  12 variables:
 $ Year     : int  1900 1901 1902 1903 1904 1905 1906 1907 1908..
 $ Low      : num  -7 -7 -4 -7 -11 -3 11 -8 -8 -1 ...
 $ High     : num  50 48 41 50 38 47 62 61 42 61 ...
 $ Hi_min   : num  36 37 27 36 31 32 53 38 32 50 ...
 $ Lo_max   : num  12 20 11 12 6 14 20 11 15 13 ...
 $ Avg_min  : num  18 17 15 15.1 8.2 10.9 25.8 17.2 17.6 20 ...
 $ Avg_max  : num  34.7 31.8 30.4 29.6 22.9 25.9 38.8 31.8 28.9..
 $ Mean     : num  26.3 24.4 22.7 22.4 15.3 18.4 32.3 24.5 23.2..
 $ Precip   : num  1.06 1.45 0.6 1.27 2.51 1.64 1.91 4.68 1.06 ..
 $ Snow     : num  4 10.1 6 7.3 11 7.9 3.6 16.1 4.3 8.7 ...
 $ Hi_Pricip: num  0.28 0.4 0.25 0.4 0.67 0.84 0.43 1.27 0.63 1..
 $ Hi_Snow  : num  1.1 3.2 2.5 3.2 2.1 2.5 2 5 1.3 7 ...
```

We focus on `Low`, which is the lowest temperature, in Fahrenheit, for January.
As statisticians, we want an uncertainty estimate. We want to know how reliable our estimate is, since it is based on on
The data are $y_1^*, \dots, y_N^*$, which we also write as $y_{1:N}^*$.
Basic estimates of the mean and standard deviation are

shadecolorrgb0.969, 0.969, 0.969fgcolor plot(Year,Low,data=y,ty="l")
shadecolorrgb0.969, 0.969, 0.969fgcolor

A first look at an autoregressive-moving average (ARMA) model
ARMA models
Another basic thing to do is to fit an **autoregressive-moving average** (ARMA) model. We'll look at ARMA models

This has a one-lag autoregressive term, $\alpha(Y_{n-1}-\mu)$, and a one-lag moving average term, $\beta\epsilon_{n-1}$. It is therefore called a
If $\alpha=\beta=0$, we get back to the basic independent model, $Y_n=\mu+\epsilon_n$.
If $\alpha=0$ we have a moving average model with one lag, MA(1).
If $\beta=0$, we have an autoregressive model with one lag, AR(1).
We model $\epsilon_1\dots,\epsilon_N$ to be an independent, identically distributed (iid) sequence. To be concrete, let's specify a model w
A note on notation
In this course, capital Roman letters, e.g., X, Y, Z, denote random variables. We may also use ϵ, η, ξ, ζ for random noi
We use lower case Roman letters (x, y, z, ...) to denote numbers. These are not random variables. We use y^* to denote
"We must be careful not to confuse data with the abstractions we use to analyze them." (William James, 1842-1910).
Other Greek letters will usually be parameters, i.e., real numbers that form part of the model.
Fitting an ARMA model in R
fragile]Maximum likelihood
We can readily fit the ARMA(1,1) model by maximum likelihood,
shadecolorrgb0.969, 0.969, 0.969fgcolor arma11 ¡- arima(y$Low,order=c$(1,0,1))
`print(arma11)` or just `arma11` gives a summary of the fitted model, where α is called `ar1`, β is called `ma1`, and μ is call
shadecolorrgb0.969, 0.969, 0.969fgcolor

```
Coefficients:
         ar1    ma1  intercept
      -0.584  0.619     -2.823
s.e.   0.598  0.578      0.688

sigma^2 estimated as 55.8:  log likelihood = -421.84,
aic = 851.68
```

We will write the ARMA(1,1) estimate of μ as $\hat{\mu}_2$, and its standard error as SE_2.
fragile]Investigating R objects
Some poking around is required to extract the quantities of primary interest from the fitted ARMA model in R.
shadecolorrgb0.969, 0.969, 0.969fgcolor names(arma11)

```
 [1] "coef"      "sigma2"    "var.coef"  "mask"      "loglik"
 [6] "aic"       "arma"      "residuals" "call"      "series"
[11] "code"      "n.cond"    "nobs"      "model"
```

shadecolorrgb0.969, 0.969, 0.969fgcolor mu2 ¡- arma11$coef["intercept"] se2 <- sqrt(arma11$var.coef["intercept","interce

```
mu2 = -2.823284 ,  se2 = 0.6880817
```

Model diagnostics
fragile]Comparing the iid estimate with the ARMA estimate
In this case, the two estimates, $\hat{\mu}_1=-2.83$ and $\hat{\mu}_2=-2.82$, and their standard errors, $\mathrm{SE}_1=0.68$ and $\mathrm{SE}_2=0.69$, are
For data up to 2015, $\hat{\mu}_1^{2015}=-2.83$ and $\hat{\mu}_2^{2015}=-2.85$, with standard errors, $\mathrm{SE}_1^{2015}=0.68$ and $\mathrm{SE}_2^{2015}=0.83$.
In this case, the standard error for the simpler model is $100(1-\mathrm{SE}_1^{2015}/\mathrm{SE}_2^{2015})=17.5\%$ smaller.
Exactly how the ARMA(1,1) model is fitted and the standard errors computed will be covered later.
Question 1.3. When standard errors for two methods differ, which is more trustworthy? Or are they both equally valid

fragile]Model diagnostic analysis
We should do **diagnostic analysis**. The first thing to do is to look at the residuals.
For an ARMA model, the residual r_n at time t_n is defined to be the difference between the data, y_n^*, and a one-step ahe
From the ARMA(1,1) definition,

a basic one-step-ahead predicted value corresponding to parameter estimates $\hat{\mu}$ and $\hat{\alpha}$ could be

A **residual time series**, $r_{1:N}$, is then given by

In fact, R does something slightly more sophisticated.
fragile]
shadecolorrgb0.969, 0.969, 0.969fgcolor plot(arma11$resid)
shadecolorrgb0.969, 0.969, 0.969fgcolor

We see slow variation in the residuals, over a decadal time scale. However, the residuals $r_{1:N}$ are close to uncorrelated.

fragile]
shadecolorrgb0.969, 0.969, 0.969fgcolor acf(arma11$resid,na.action= na.pass)
shadecolorrgb0.969, 0.969, 0.969fgcolor