Mathematical Statistics I

Chapter 1: Probability

Jesse Wheeler

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, Point and Interval Estimation, Uniformly most powerful tests, likelihood ratio tests, chi-squre and F tests, and nonparameteric tests.
- 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/.
- Canvas: Canvas will be used to submit assignments, view grades, and for course announcments.
- Course 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 introductory examples.
- 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 essential for building statistical theory in 4451.

Sample Spaces

Probability theory is concerned with situations in which the outcomes occur randomly. We call these situations experiments. The set of all possible outcomes is called the sample space.

Flipping a coin

Flipping a coin is an *experiment*, with possible outcomes $\{H, T\}$, which defines the *sample space* of the experiment.

An arbitrary sample space is typically denoted Ω, and an element of Ω is denoted ω.

Sample Spaces

Probability theory is concerned with situations in which the outcomes occur randomly. We call these situations experiments. The set of all possible outcomes is called the sample space.

Flipping a coin

Flipping a coin is an *experiment*, with possible outcomes $\{H, T\}$, which defines the *sample space* of the experiment.

An arbitrary sample space is typically denoted Ω, and an element of Ω is denoted ω.

- In the example above, $\Omega = \{H, T\}$.

Sample Space Examples

Example: Stoplights

Driving to work, a commuter passes through a sequence of three intersections with traffic lights. At each light, they either stop (s), or continues (c). The sample space Ω is the set of all possible outcomes:

$$\Omega = \{ccc, ccs, css, csc, sss, ssc, scc, scs\}$$

where csc denotes the outcome that the commuter continues at the first light, stops at the second, and continues through the third.

Sample Space Examples II

Example: Printing

The number of jobs in a print queue of a printing machine may be modeled as random. Here the sample space is all non-negative integers:

$$\Omega = \mathbb{N} = \{0, 1, 2, \ldots\}$$

Sample Space Examples III

Example: Earthquakes

We may want to model the *time* between successive earthquakes in a particular region. In this case, the experiment is the length of time between earthquakes, and our sample space is (uncountably) infinite:

$$\Omega = \{t \in \mathbb{R} | t \geq 0\}.$$

Looking for trends and relationships in dependent data

The first half of this course focuses on:

1. Quantifying dependence in time series data.
2. 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 evidence for regional variation? What is a good prediction interval for weather in the next year or two?

Modeling and statistical inference for dynamic systems

The second half of this course focuses on:

1. Building models for dynamic systems, which may or may not be linear and Gaussian.
2. 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 as volatility)? How do we assess our model and decide whether it is indeed an improvement?

Example: Winter in Michigan

Example: Winter in Michigan

There is a temptation to attribute a warm winter to global warming. You can then struggle to explain a subsequent cold winter. Is a trend in fact noticeable at individual locations in the presence of variability? Let's look at some data, downloaded from www.usclimatedata.com and put in ann_arbor_weather.csv.

- You can get this file from the course 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, code, homeworks and solutions for this course.

```
y <- read.table(file="ann_arbor_weather.csv",header=1)
```

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 and learning statistical methodology, since they make it easy for you to replicate and adapt analysis presented in class.

- 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 combining with Latex to produce pdf articles.
- Rmd naturally produces html.

To get a first look at our dataset, `str` summarizes its structure:

```
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,

As statisticians, we want an uncertainty estimate. We want to know how reliable our estimate is, since it is based on only a limited amount of data.

- The data are $y_1^*, \ldots, y_N^*$, which we also write as $y_{1:N}^*$.
- Basic estimates of the mean and standard deviation are

$$\hat{\mu}_1 = \frac{1}{N} \sum_{n=1}^{N} y_n^*, \qquad \hat{\sigma}_1 = \sqrt{\frac{1}{N-1} \sum_{n=1}^{N} (y_n^* - \hat{\mu}_1)^2}. \tag{1}$$

- This suggests an approximate confidence interval for μ of $\hat{\mu}_1 \pm 1.96 \, \hat{\sigma}_1 / \sqrt{N}$.

Question 1.1. What are the assumptions behind this confidence interval?

- 1955 has missing data, coded as NA, requiring a minor modification. So, we compute $\hat{\mu}_1$ and $\mathrm{SE}_1 = \hat{\sigma}_1/\sqrt{N}$ as

```
mu1 <- mean(y$Low,na.rm=TRUE)
se1 <- sd(y$Low,na.rm=TRUE)/sqrt(sum(!is.na(y$Low)))
cat("mu1 =", mu1, ",  se1 =", se1, "\n")

mu1 = -2.829268 ,  se1 = 0.6767317
```
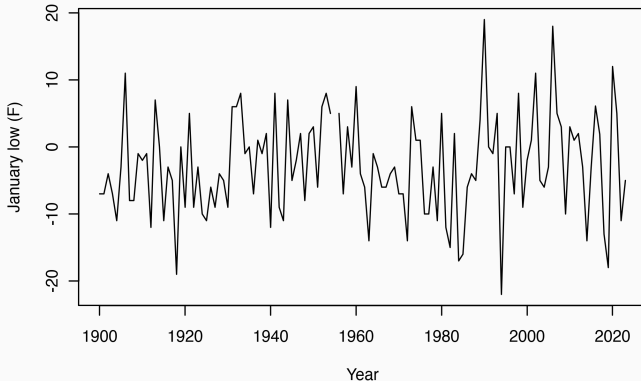
Question 1.2. If you had to give an uncertainty estimate on the mean, is it reasonable to present the confidence interval, -2.83 ± 1.33? Do you have ideas of a better alternative?

Some data analysis

- The first rule of data analysis is to plot the data in as many ways as you can think of. For time series, we usually start with a time plot.

```
plot(Year,Low,data=y,ty="l")
```

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 in much more detail later in the course. Let's fit an ARMA model given by

$$Y_n = \mu + \alpha(Y_{n-1} - \mu) + \epsilon_n + \beta\epsilon_{n-1}. \qquad (2)$$

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 an ARMA(1,1) model. These lags give the model some time dependence.

- 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 our resonant drum noise as identically distributed

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 noise processes. Thus, these symbols are used to build models.

- We use lower case Roman letters (x, y, z, ...) to denote numbers. These are not random variables. We use y^* to denote a data point.

- "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

Maximum likelihood

We can readily fit the ARMA(1,1) model by maximum likelihood,

```
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 called intercept.

```
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

Investigating R objects

Some poking around is required to extract the quantities of primary interest from the fitted ARMA model in R.

```
names(arma11)
```

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

```
mu2 <- arma11$coef["intercept"]
se2 <- sqrt(arma11$var.coef["intercept","intercept"])
cat("mu2 =", mu2, ",  se2 =", se2, "\n")
```

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

Model diagnostics

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 close.
- 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 for their distinct estimators?

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 ahead prediction of y_n^* based on $y_{1:n-1}^*$, written as y_n^P.

From the ARMA(1,1) definition,

$$Y_n = \mu + \alpha(Y_{n-1} - \mu) + \epsilon_n + \beta\epsilon_{n-1}, \tag{3}$$

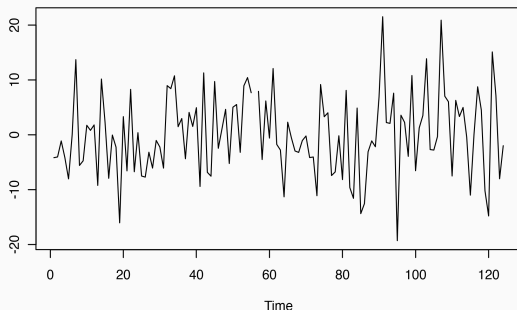a basic one-step-ahead predicted value corresponding to parameter estimates $\hat{\mu}$ and $\hat{\alpha}$ could be

$$y_n^P = \hat{\mu} + \hat{\alpha}(y_{n-1}^* - \hat{\mu}). \tag{4}$$

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

$$r_n = y_n^* - y_n^P. \tag{5}$$

In fact, R does something slightly more sophisticated.

```
plot(arma11$resid)
```


We see slow variation in the residuals, over a decadal time scale. However, the residuals $r_{1:N}$ are close to uncorrelated. We see this by plotting their pairwise sample correlations at a range of lags. This is the **sample autocorrelation function**, or **sample ACF**, written for each lag h as

$$\hat{\rho}_h = \frac{\frac{1}{N}\sum_{n=1}^{N-h} r_n\, r_{n+h}}{\frac{1}{N}\sum_{n=1}^{N} r_n^2}. \tag{6}$$

```
acf(arma11$resid,na.action=na.pass)
```

Series arma11$resid
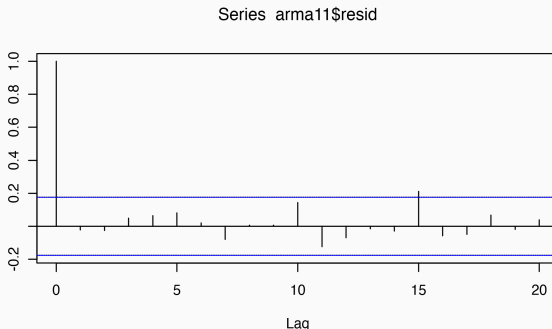
- This shows no substantial autocorrelation. An ARMA model may not be a good way to describe the slow variation present in the residuals of the ARMA(1,1) model.
- This is a simple example. However, inadequate models giving poor statistical uncertainty estimates is a general concern when working with time series data.

Model misspecification and non-reproducibility

Quantifying uncertainty for scientific reproducibility

Usually, omitted dependency in the model will give overconfident (too small) standard errors.

- This leads to scientific reproducibility problems, where chance variation is too often assigned statistical significance.
- It can also lead to improper pricing of risk in financial markets, a factor in the US financial crisis of 2007-2008.

A first look at a state-space model

Models dynamic systems: State-space models

Scientists and engineers often have equations in mind to describe a system they're interested in. Often, we have a model for how the state of a stochastic dynamic system evolves through time, and another model for how imperfect measurements on this system gives rise to a time series of observations.

This is called a **state-space model**. The **state** models the quantities that we think determine how the system changes with time. However, these idealized state variables are not usually directly and perfectly measurable.

Statistical analysis of time series data on a system should be able to

1. Help scientists choose between rival hypotheses.
2. Estimate unknown parameters in the model equations.

A finance example: fitting a model for volatility of a stock market index

- Let $\{y_n^*, n = 1, \ldots, N\}$ be the daily returns on a stock market index, such as the S&P 500.

- The **return** is the difference of the log of the index. If z_n is the index value for day n, then $y_n^* = \log(z_n) - \log(z_{n-1})$.

- Since the stock market is notoriously unpredictable, it is often unproductive to predict the mean of the returns and instead there is much emphasis on predicting the variability of the returns, known as the **volatility**.

- Volatility is critical to assessing the risk of financial investments.

Financial mathematicians have postulated the following model. We do not need to understand it in detail right now. The point is that investigators find it useful to develop models for how dynamic systems progress through time, and this gives rise to the time series analysis goals of estimating unknown parameters and assessing how successfully the fitted model describes the data.

$$Y_n = \exp\left\{\frac{H_n}{2}\right\} \epsilon_n, \quad G_n = G_{n-1} + \nu_n,$$

$$H_n = \mu_h(1 - \phi) + \phi H_{n-1}$$
$$+ Y_{n-1}\sigma_\eta\sqrt{1 - \phi^2}\tanh(G_{n-1} + \nu_n)\exp\left\{\frac{-H_{n-1}}{2}\right\} + \omega_n. \tag{7}$$

- $\{\epsilon_n\}$ is iid $N(0, 1)$, $\{\nu_n\}$ is iid $N(0, \sigma_\nu^2)$, $\{\omega_n\}$ is iid $N(0, \sigma_\omega^2)$.
- H_n is unobserved volatility at time t_n. We only observe the return, modeled by Y_n.
- H_n has auto-regressive behavior and dependence on Y_{n-1} and

Questions to be addressed later in the course

This is an example of a **mechanistic model**, where scientific or engineering considerations lead to a model of interest. Now there is data and a model of interest, it is time to recruit a statistician!

1. How can we get good estimates of the parameters, μ_h, ϕ, σ_ν, σ_ω, together with their uncertainties?
2. Does this model fit better than alternative models? So far as it does, what have we learned?
3. Does the data analysis suggest new models, or the collection of new data?

Likelihood-based inference for this partially observed stochastic dynamic system is possible, and enables addressing these questions (Bretó, 2014).

- By the end of this course, you will be able to carry out data analysis developing complex models and fitting them to time

References

Bretó C (2014). "On idiosyncratic stochasticity of financial leverage effects." *Statistics & Probability Letters*, **91**, 20–26. doi: 10.1016/j.spl.2014.04.003.

References and Acknowledgements II

- We acknowledge students and instructors for previous versions of this course.