MVE051/MSG810 2017 Lecture 7

Petter Mostad

Chalmers

November 20, 2017

The purpose of collecting and analyzing data

- Purpose: To build and select *models* for parts of the real world (which can be used for prediction).
- The first part of data analysis is always to summarize and visualize the data. This is called *descriptive statistics*. (Most people call it just "statistics").
- What separates *mathematical statistics* from descriptive statistics is that we use *probaility theory* to formulate, build, and select the models for parts of the real world.

Data collection and analysis is always subjective

- What one decides to study, how one decides to study it, and what data one decides to collect, is necessarily based on ones preconceptions.
- Your way to summarize and visualize data is always influenced by your preconceptions; indeed, different ways to summarize data can be used to promote different ideas.
- The choice between different statistical models (and in some settings the choice of different statistical methods) is necessarily subjective.

Summarizing data

- **Graphical summaries**: Illustrating the data (or part of the data) in an plot or figure.
- **Numerical summaries**: Computing from the data (or part of the data) one or more numbers that tells something important about the data.

There are a large number of ways to summarize; you should at least know the ones we go through below.

Numerical summaries

Let $x_1, x_2, \ldots, x_n$ be observed real values.

- Mean: $\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$.
- Median: If we sort the data so we write it $y_1, \ldots, y_n$ in order of size, then the median is $y_{(n+1)/2}$ if n is odd and the mean of $y_{n/2}$ and $y_{n+1}/2$ if n is even.
- Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- The sample standard deviation is the square root of this.
- Min and max.

Quantiles and percentiles

- *Quantile*: A number such that a certain *proportion* of the data values is smaller than the number.
- We may also talk about a *perentile*, where the proportion is specified with a percentage.
- Example: The 30th percentile is a number such that 30% of the data is smaller than the number.
- Example: The median is the same thing as the 50th percentile.
- Example: The first *quartile* is a number such that a quarter of the data is smaller than the number.
- Example: The inter-quartile range is the interval between the 25th and 75th percentile.
- We also talk about quantiles for probability densities. Example: The first quartile of a normal density Normal(μ, σ) is the number z_0 such that $\Pr(z < z_0) = 1/4$ when $z \sim \text{Normal}(\mu, \sigma)$.

Graphical summaries

- Scatterplots.
- Histograms. Note how certain parameters must be selected (and is usually selected automatically by the program making the histogram).
- Boxplots: Efficient way to illustrate and compare the spread in one or more groups of data. Plots a box with the inter-quartile range and the median, together with "whiskers" indicating the spread of the data (definitions may vary) and individual observations outside this spread.
- Exact definitions of parameter defaults in the functions above, and generally the choice of graphical functions, depends on the program you use.
- We may regard graphical summaries as a step on the way to selecting a probabilistic model for the data.
- A free and powerful tool for statistics: R (www.r-project.org).

Random variables as models for data

The second step in a statistical data analysis is to find a *probabilistic model* for your data.

- We describe a *population* of objects, or maybe possible observations, where our data represents a subset of this population.
- Example: We have measured the concentration of lead in 10 fish from a lake. The *population* may be the lead concentrations of all the fish in the lake. (Which species? Only this lake? ...). The *model* of the population could be for example a normal distribution, or a normal distribution of the logged values.
- We generally have to assume that our data is a *random sample* from the popultation, i.e., that
  - Each data value is randomly chosen from the population (so each population member has the same chance of being observed, or, given a model, the model specifies the probability (density) of each possible observation).
  - The observations are independent of each other.
- It is very important to specify the population so that the assumption that your data is a random sample is reasonable!

Finding a probabilistic model for your data

In this course, finding a probabilistic (or *stochastic*) model for your data will have two steps:

- Step 1: Find the type of model, i.e., a family of probability distributions that fit the context: The Normal family, the Binomial family, the Poisson family, etc.
  - Consider: Are the observed values real numbers or integers? (Could they be real numbers?) Is this a "sequence of trials"? Etc.
  - In our course, we may use Hypothesis Testing for selecting between possible models, but alternative methods also exist.
- Step 2: Find the parametres of the model (For example, find values for μ and σ^2 if the model is Normal, or λ if the model is Poisson).
  - In this course, we will use *estimators* which compute from the data an *estimate* for the model parameters.
  - It is also possible to use probability theory to obtain probability distributions for the parameters; this is outside this course.

Estimates and estimators

Assume we have a model with unknown parameter θ and data $x_1, \ldots, x_n$ which we assume is a random sample from the model.

- We separate between
  - An *estimator* for θ: A function or formula which from a random sample $x_1, \ldots, x_n$ computes a number which may function as a value for the parmeter θ.
  - An *estimate* for θ: The value of the estimator for specific values of $x_1, \ldots, x_n$.
- We often write $\hat{\theta}$ for the estimate, but also for the estimator for θ. (So if the parameter is called for example μ, we write $\hat{\mu}$ etc.).
- A function of a random sample is called a *statistic*. So an estimator is a statistic.
- A statistic is also a random variable, as it is a function of random variables. So we can talk about its distribution, expectation, variance, etc.

Constructing an estimator

Generally, in this course we will use "standard estimators" for each context, but here is a discussion on obtaining estimators:

- There is no general mathematical specification for how to construct an estimator. Instead one may specify some properties one believes a "good" estimator should have, and try to find estimators fulfilling these criteria.
- A good property for an estimator: To be *unbiased*: This means that the expectation of the estimator is equal to the parameter it is estimating.
- A good property for an estimator: To have as small variance as possible.
- A common way to construct an estimator (the Maximum Likelihood (ML) method): Write the probability of the observed data as a function of the model parmeters. This is the *likelihood function*. Find the parameters mazimizing this function. The formula for computing this maximum from the data becomes the estimator.

Estimator for the expectation

- Assume data is a random sample from Normal(μ, σ), so that they are represented by independent random variables

$$X_1, X_2, \ldots, X_n \sim \text{Normal}(\mu, \sigma)$$

  We want to find an estimator for μ.
- A natural estimator is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$. (This is an ML estimator).
- The estimator is unbiased, i.e., $\mathbb{E}\left[\overline{X}\right] = \mu$. (Simple proof).
- The proof works equally well for any distribution fo X_i. So the estimator $\overline{X}$ is always an unbiased estimator for the expectaiton of a distribution.
- Example: Assume the observations $x_1, x_2, \ldots, x_n$ are a random sample from a Poisson(λ) distribution, where the expectation is equal to the parameter λ. Then $\overline{X}$ is an unbiased estimator for λ.

Variance of an estimator

- The estimator $\overline{X}$ has variance σ^2/n, where σ^2 is the variance of the distribution for $X_1, \ldots, X_n$.
- The proof is good to understand. NOTE: The proof uses that, if X and Y are independent random variables, we have $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$. This is also good to understand.
- Example: The Bernoulli distribution (which is the Binomial distribution with only one trial):
  - The distribution has a parameter p and $X \sim \text{Bernoulli}(p)$ has possible values 0 and 1.
  - The expectation is p and the variance is $p(1-p)$.
  - $\hat{p} = \overline{X}$ is an unbiased estimator for p. The variance of this estimator is $p(1-p)/n$.

Estimator for variance

- If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with expectation μ and variance σ^2, then

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

  is an unbiased estimator for σ^2.

- The proof may be useful to understand:

$$\begin{aligned}\mathbb{E}\left[\hat{\sigma}^2\right] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] \\ &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}((X_i - \mu) - (\overline{X} - \mu))^2\right] = \cdots = \sigma^2\end{aligned}$$

- This is the reason why we divide with $n-1$ to compute the sample variance: It makes the estimator unbiased.

The distribution of an estimator

- To further find out how "good" an estimator is, we can study its distribution, not only its expectation and variance.
- Example: If $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma)$, then we know the estimator $\overline{X}$ has expectation μ and variance σ^2/n: Also:
  - One can show: If $X_1, \ldots, X_n$ are independent and normally distributed, then $X_1 + X_2 + \cdots + X_n$ is normally distributed.
  - We know that if Y is normally distributed then Y/n is also normally distributed.

  From this we get that $\overline{X} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$.
- If $X_1, \ldots, X_n$ has a Bernoulli-distribution with parameter p, then we get from the definitions that $X_1 + X_2 + \cdots + X_n$ has a Binomialdistribution with parmeters n och p. From this we can also get an explicit description of the distribution of $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$.
- One can show that if $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma)$ then, for the variance estimator $\hat{\sigma}^2$ we get

$$(n-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-1)$$

So: $(n-1)\hat{\sigma}^2$ has a distribution that corresponds to σ^2 multiplied with a chi-squared distribution with $n-1$ degrees of freedom.

We study estimators not estimates

- Assume we investigate a type of trials which each time result in success (1) or failure (0), and the probability of success is an unknown parameter p. Assume we make some trials and get the results

$$0, 1, 0, 0, 1, 0, 0, 1$$

  We make the estimate $3/8 = 0.375$ for p.
- How "good" is this estimate? *We cannot say anything about that before we specify the estimator*.
- ALTERNATIVE 1: The estimator consists of making 8 trials, letting x be the number of successes, and computing $\hat{p} = x/8$.
- ALTERNATIVE 2: The estimator consists of making trials until 3 successes have been observed, and letting x be the number of trials needed for this outcome. Then one computes $\hat{p} = 3/x$.
- The two estimators have different properties! One is unbiased and the other is biased.

Example, cont.

- Let us for example assume that the real value for p is 0.6. We can then study which distributions our two estimators have.
- ALTERNATIVE 1: We have $X \sim \text{Binomial}(8, 0.6)$. The possible values for $\hat{p} = X/8$ and their probabilities are found in the table below:

0/8	1/8	2/8	3/8	4/8	5/8	6/8	7/8	8/8
0.001	0.008	0.041	0.124	0.232	0.279	0.209	0.090	0.017

- The estimator has expectation 0.6; it is unbiased.
- ALTERNATIVE 2: We get $X \sim \text{Neg-Binomial}(3, 0.6)$. The possible values for $\hat{p} = 3/X$ and their probabilities are found in the table below:

3/3	3/4	3/5	3/6	3/7	3/8	3/9	3/10	3/11
0.216	0.259	0.207	0.138	0.083	0.046	0.025	0.013	0.006

3/12	3/13	3/14	3/15	3/16,3/17,...
0.003	0.001	0.001	0.000	totalt 0.000

- The estimator has expectation 0.672. It is biased.