Lecture 8: Samples and point estimates

MVE055 / MSG810 Mathematical statistics and discrete mathematics)

Moritz Schauer

Last updated September 21, 2020, 2020

GU & Chalmers University of Technology

Samples and point estimators

Statistics

Definition: Sample

A sample $(x_1, \ldots, x_n)$ of size n is made of n independent observations (realisations) of a random variable X, or – the same – of random variables $X_1, \ldots, X_n$ where all X_i are independent and equally distributed (thus have the same distribution).

Example: The weight of a cat is modelled as normal random variable $X \sim \mathrm{N}(\mu, (1.2)^2)$ with unknown parameter μ. $(5.27, 4.07, 5.48, 3.38)$ is a sample of the weight of $n = 4$ independently sampled cats.

Statistics

Like in the “cat“-example we can often say what kind of distribution is appropriate for X but we do not know the right parameters.

Many statistical problems can be reduced to the following question: Given the observations $x_1, \ldots, x_n$, what can we say about the parameters in the distribution of X_i?

The sample mean as estimator

Recall: $\bar{X}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean.

Example: Let $(5.27, 4.07, 5.48, 3.38)$ our sample.
$\bar{x}^{(4)} = (5.27 + 4.07 + 5.48 + 3.38)/4 = 4.55$ is an estimate for the unknown mean parameter μ.

Standard error of the mean

Let $X_1, \ldots, X_n$ an independent sample of a $\mathrm{N}(\mu, \sigma^2)$ r.v. Then $\bar{X}^{(n)}$ is N$(\mu, \sigma^2/n)$-distributed.

$$\frac{\sigma}{\sqrt{n}} \text{ is called standard error of the mean.}$$

This means E$\bar{X}^{(n)} = \mu$ and $\mathrm{Var}(\bar{X}^{(n)}) = \sigma^2/n$. The population variance of the weight was $\sigma^2 = (1.2)^2$. Thus the standard error associated with x is $1.2/\sqrt{4} = 0.6$.

Unbiasedness of the sample mean

Point estimators

Estimation

An estimator for a parameter θ is a function $\hat{\theta}(X_1, \ldots, X_n)$ mapping the observations into the parameter space Θ.

Example: $\bar{X}^{(n)}$ is an estimator for $\mu = \mathsf{E}X$.

$\hat{\theta}$ can refer both to a random variable and to actual observed values.

- $\hat{\theta}(X_1, \ldots, X_n)$ is a random variable with a certain distribution (random in $\rightarrow$ random out).
- $\hat{\theta}(x_1, \ldots, x_n)$ is a number calculated from data. This is called the point estimate of the parameter.

Properties of estimators

Two important qualities of estimators:

- *unbiased*: $\mathsf{E}(\hat{\theta}(X_1, \ldots, X_n)) = \theta$.
- Small variance in large samples: $\mathsf{V}(\hat{\theta}(X_1, \ldots, X_n))$ small if n large.

If the expected value of the estimator is the true value (the estimator is unbiased), that means that the estimated values center on average around the true value if we make several repeated samples of size n.

- For a given sample, the value need not be close to the true value.
- The variance of an unbiased estimate gives an indication of how far it may be from the actual value.
- Often the standard error of the estimate is reported, which is the standard deviation of the estimate.

Important estimators

Sample mean and sample variance

Consider an i.i.d sample $(X_1, \ldots, X_n)$ and assume that $\mathsf{E}(X_i) = \mu$ and $\mathsf{V}(X_i) = \sigma^2$.

The **sample mean** $\hat{\mu} = \bar{X}^{(n)}$ is an unbiased estimator of μ, that is $\mathsf{E}(\hat{\mu}) = \mu$. It has standard error $\sqrt{V(\hat{\mu})} = \frac{\sigma}{\sqrt{n}}$.

An unbiased estimator for the variance σ^2 is the **sample variance**

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Sample variance can also be computed as

$$S^2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}$$

Percentiles and quantiles

The p^{th} percentile P is the value of X such that $p\%$ or less of the observations are less than P and $(100 - p)\%$ or less are greater than P. p^{th} percentiles are $p\%$-quantiles.

In particular, P_{25} is the 25^{th} percentile or the first quartile denoted also by $Q_1 . P_{50}$ is the 50^{th} percentile or the second quartile Q_2, which is also the median, and P_{75} is the 75^{th} percentile or the third quartile Q_3.

Note that $Q_1 = \frac{n+1}{4}$ th ordered observation, $Q_2 = \frac{2(n+1)}{4} = \frac{n+1}{2}$ th ordered observation, and $Q_3 = \frac{3(n+1)}{4}$ th ordered observation.

Example

Given the following set of data :

$$18, 1, 20, 15, 12, 15, 14, 7, 11, 9, 6, 4$$

Order the numbers from the lowest to the highest

$$1, 4, 6, 7, 9, 11, 12, 14, 15, 15, 18, 20$$

$\bar{x}^{(12)} = \frac{1+4+\cdots+18+20}{12} = 11.$

Median: $Me = \frac{11+12}{2} = 11.5.$

Example

Given the following set of data :

$$18, 1, 20, 15, 12, 15, 14, 7, 11, 9, 6, 4$$

Variance

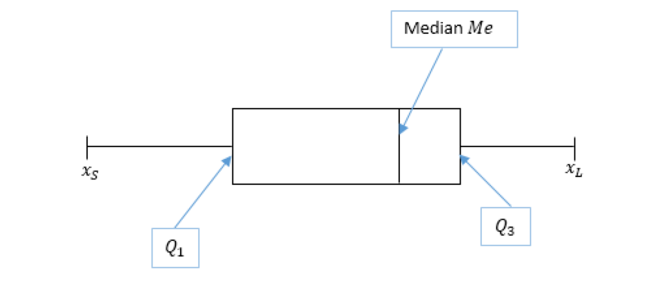
$$s^2 = \frac{(20-11)^2 + (18-11)^2 + \cdots + (-7)^2 + (-10)^2}{12-1} \approx 33.3$$

Order the numbers from the lowest to the highest

$$1, 4, 6, 7, 9, 11, 12, 14, 15, 15, 18, 20$$

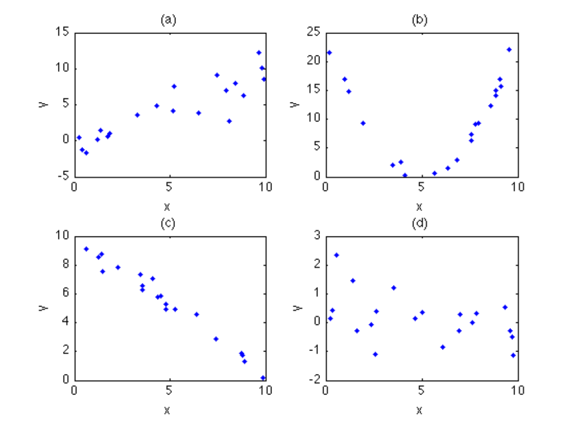
$Q_1 = 6.25$, $Q_3 = 15$.

Boxplot



Bivariate samples

Visualisation



Assume 2d measurements (x_i, y_i). A scatter plot is a two-dimensional plot in which each (x_i, y_i) measurement is represented as a point in the x-y-plane.

Descriptive statistic for bivariate data

The *sample* covariance is defined as,

$$c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

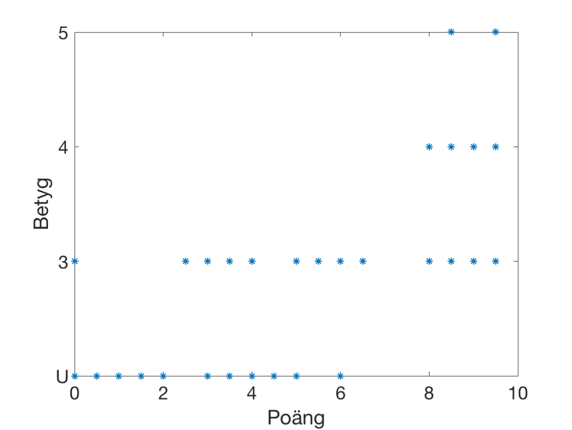
and sample correlation coefficient is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{c_{xy}}{s_x s_y}$$

The sample correlation is an empirical measure of linear dependence.

In the picture $r_{xy} = 0.8067$ i (a), $r_{xy} = 0.2912$ i (b), $r_{xy} = -0.9884$ i (c), och $r_{xy} = 0.3640$ i (d).

Example: Course results 2017



Exam grade (Y) versus points in exam question 5 (X).
Correlation: $r_{xy} = 0.7261$