Lecture 8: Samples and point estimates

MVE055 / MSG810
Mathematical statistics and discrete mathematics

Moritz Schauer

Last updated September 20, 2021, 2021

GU & Chalmers University of Technology

Samples and point estimators

Example: $(5.27, 4.07, 5.48, 3.38)$ are measurements of the weight of $n = 4$ randomly (independent) selected cats.

Definition: Sample

A sample $(x_1, \ldots, x_n)$ of size n is made of n independent observations (realisations) of a random variable. 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: $(5.27, 4.07, 5.48, 3.38)$ are measurements of the weight of $n = 4$ randomly (independent) selected cats.

Example: $(5.27, 4.07, 5.48, 3.38)$ are measurements of the weight of $n = 4$ randomly (independent) selected cats.

The weight of a cat is modelled as normal random variable X_1, X_2, X_3, X_4 each $\mathrm{N}(\mu, (1.2)^2)$-distributed with unknown parameter μ. Here $\mathrm{N}(\mu, (1.2)^2)$ is a model for the population of *all cats.*

$(5.27, 4.07, 5.48, 3.38)$ is a sample of X_1, X_2, X_3, X_4.

Definition: Anti-Example

$(5.27, 5.27, 5.27, 5.27, 5.27)$ is perhaps not a sample

(lack of independence because some genius just weighted the same cat over and over).

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.

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 (assuming each X_i is drawn independently from the same distribution)?

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 $x_1, \dots, x_n$, what can we say about the parameters in the distribution of X_i (assuming each X_i is drawn independently from the same distribution)?

Definition: i.i.d.

We write $X_1, X_2, \dots X_n \overset{\text{i.i.d.}}{\sim} D$ if $X_1, X_2, \dots, X_n$ are independently and identically distributed with distribution D.

The sample mean as estimator

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

The sample mean as estimator

$\bar{X}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} X_i$ 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 a realisation $\bar{X}^{(n)}$.

We model $\bar{X}^{(n)}$ itself as random variable with its own expectation, variance and realization etc.

The sample mean as estimator

$\bar{X}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 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 a realisation $\bar{X}^{(n)}$.

We model $\bar{X}^{(n)}$ itself as random variable with its own expectation, variance and realization etc. Now with $\mu = \mathsf{E}(X_1) = \mathsf{E}(X_2) = \ldots$ and $\sigma^2 = \mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \ldots$

$$\mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}X_i \stackrel{(*)}{=} \mu$$

The sample mean as estimator

$\bar{X}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 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 a realisation $\bar{X}^{(n)}$.

We model $\bar{X}^{(n)}$ itself as random variable with its own expectation, variance and realization etc. Now with $\mu = \mathsf{E}(X_1) = \mathsf{E}(X_2) = \ldots$ and $\sigma^2 = \mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \ldots$

$$\mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}X_i \overset{(*)}{=} \mu$$

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i)\right) \overset{i.i.d}{=} \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n}$$

Ah! Smaller uncertainty, 4.55 is perhaps closer to μ than most the values in our sample which vary from μ by σ.

The sample mean as random variable

Expectation and variance of the sample average

$\mathsf{E}\left(\bar{X}^{(n)}\right) = \mu$ and $\mathrm{Var}\left(\bar{X}^{(n)}\right) = \sigma^2/n$.

Quiz: How fast goes uncertainty down if n increases?

The sample mean as random variable

Expectation and variance of the sample average

$\mathsf{E}\left(\bar{X}^{(n)}\right) = \mu$ and $\mathrm{Var}\left(\bar{X}^{(n)}\right) = \sigma^2/n$.

Quiz: How fast goes uncertainty down if n increases?

Standard error of the mean

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

Point estimate and standard error

Example: Take $(5.27, 4.07, 5.48, 3.38)$ our sample. Model $X_1, \dots X_n \overset{i.i.d}{\sim} \mathrm{N}(\mu, \sigma^2)$ with $n = 4$ and $\sigma = 1.2$ and μ unknown.

Point estimate and standard error

Example: Take $(5.27, 4.07, 5.48, 3.38)$ our sample. Model $X_1, \ldots X_n \overset{i.i.d}{\sim} \mathrm{N}(\mu, \sigma^2)$ with $n = 4$ and $\sigma = 1.2$ and μ unknown.

$\bar{x}^{(4)} = (5.27 + 4.07 + 5.48 + 3.38)/4 = 4.55$ is an estimate for μ

Point estimate and standard error

Example: Take $(5.27, 4.07, 5.48, 3.38)$ our sample. Model $X_1, \dots X_n \overset{i.i.d}{\sim} \mathrm{N}(\mu, \sigma^2)$ with $n = 4$ and $\sigma = 1.2$ and μ unknown.

$\bar{x}^{(4)} = (5.27 + 4.07 + 5.48 + 3.38)/4 = 4.55$ is an estimate for μ

The standard error associated with $\bar{x}^{(4)}$ is $\sigma/\sqrt{n} = 1.2/\sqrt{4} = 0.6$.

Our estimate

$$\mu \approx 4.55 \pm 0.6$$

The sample mean as random variable: Gaussian case

Average of Gaussian distributed random variables.

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

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 $\mu = \mathsf{E}X_1 = \mathsf{E}X_2 = \ldots.$

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 $\mu = \mathsf{E}X_1 = \mathsf{E}X_2 = \ldots$.

$\hat{\theta}$ 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*: $\mathsf{E}(\hat{\theta}(X_1, \ldots, X_n)) = \theta$.

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.

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 standard deviation of an unbiased estimate gives an indication of how far it may be from the actual value.

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 standard deviation 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 $\sqrt{V(\hat{\mu})} = \frac{\sigma}{\sqrt{n}}$.

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

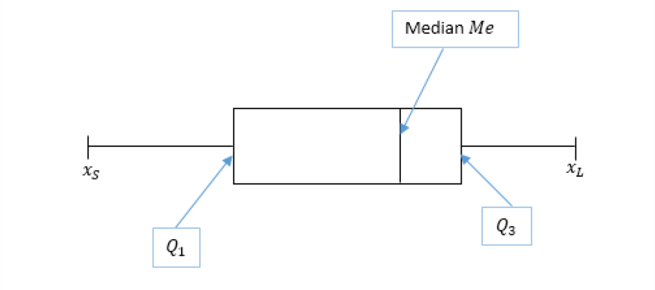
$$s^2 = \frac{(20-11)^2 + (18-11)^2 + \cdots + (-7)^2 + (-10)^2}{12-1} \approx 33.3$$

Order the numbers from the lowest to the highest

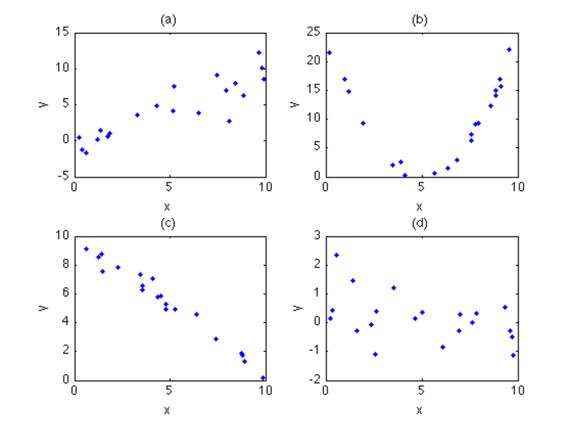
$$1, 4, 6, 7, 9, 11, 12, 14, 15, 15, 18, 20$$

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

Boxplot



Bivariate samples



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.

Statistics 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 is an unbiased estimator of the covariance $\mathrm{Cov}(X, Y)$.

Statistics 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 is an unbiased estimator of the covariance $\mathrm{Cov}(X, Y)$.

The sample correlation coefficient is defined as

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

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 is an unbiased estimator of the covariance $\mathrm{Cov}(X, Y)$.

The sample correlation coefficient is defined as

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

Statistics 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 is an unbiased estimator of the covariance $\mathrm{Cov}(X, Y)$.

The sample correlation coefficient is defined as

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

Statistics 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 is an unbiased estimator of the covariance $\mathrm{Cov}(X, Y)$.

The 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.

Statistics for bivariate data

The *sample* covariance is defined as,

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

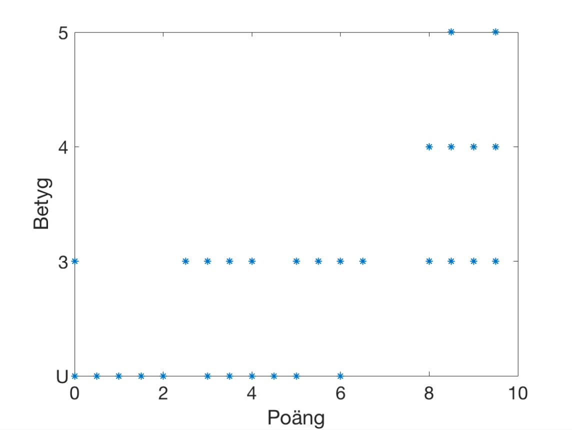
and is an unbiased estimator of the covariance $\mathrm{Cov}(X, Y)$.

The 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.

Example: Course results 2017



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

Sum of Gaussian r.v.

Let $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ with X and Y independent. Then

$$X + Y \sim \mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Note: A normal random variable with mean μ and variance σ^2 has moment generating function $m(t) = \exp(t\mu + t^2\sigma^2/2)$. So if you tell me your moment generating function, I tell you if you are normally distributed and if, what your parameters are. We can prove the theorem by computing and identifying the m.g.f of $X + Y$ (next slide)

Proof with m.g.f.

So we now $m_X(t) = \mathsf{E}\exp(tX) = \exp(t\mu_X + t^2\sigma_X^2/2)$ and $m_Y(t) = \mathsf{E}\exp(tY) = \exp(t\mu_Y + t^2\sigma_Y^2/2)$.

We compute and identify m_{X+Y}

$$\begin{aligned}
m_{X+Y}(t) &= \mathsf{E}\exp(t(X+Y)) = \mathsf{E}\left(\exp(tX)\exp(tY)\right) \\
&\overset{indep}{=} \mathsf{E}\left(\exp(tX)\right)\mathsf{E}\left(\exp(tY)\right) \\
&= m_X(t)m_Y(t) = \exp(t\mu_X + t^2\sigma_X^2/2)\exp(t\mu_Y + t^2\sigma_Y^2/2) \\
&= \exp(t(\mu_X + \mu_Y) + t^2(\sigma_X^2 + \sigma_Y^2)/2)
\end{aligned}$$

which is m.g.f of $\mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ so $X + Y$ must be $\mathrm{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ distributed.