Institute of Economic Studies, Faculty of Social Sciences
Charles University in Prague (*established 1348*)

ROBUST STATISTICS AND ECONOMETRICS

INSTITUTE OF ECONOMIC STUDIES
FACULTY OF SOCIAL SCIENCES
CHARLES UNIVERSITY IN PRAGUE

JAN ÁMOS VÍŠEK

Week 2

Content of lecture

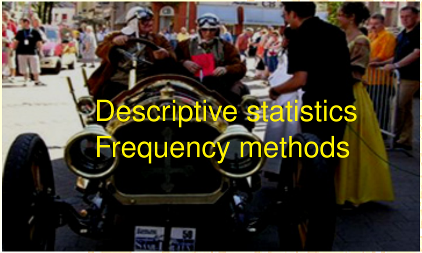
1. Repetition of findings from previous lecture
   - How did we start to study the statistics ?
2. Main goals of robust statistics and problems to be solved

How did we start to study the statistics?

How did we start to study the statistics?

Descriptive statistics
Frequency methods

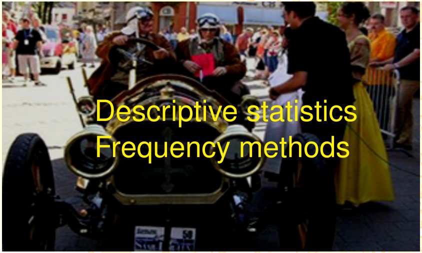
How did we start to study the statistics?



Descriptive statistics
Frequency methods

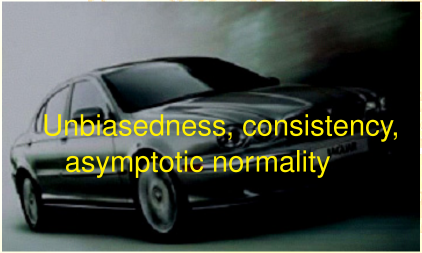
Maximum likelihood
The least squares

How did we start to study the statistics?



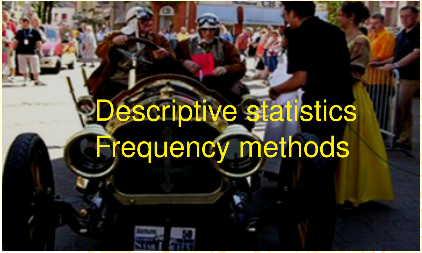
Descriptive statistics
Frequency methods

Maximum likelihood
The least squares



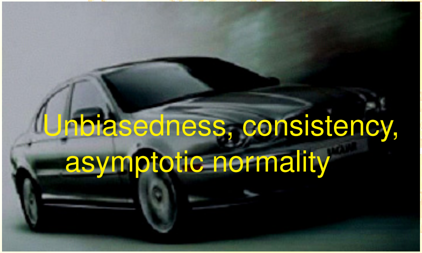
Unbiasedness, consistency,
asymptotic normality

How did we start to study the statistics?



Descriptive statistics
Frequency methods

Maximum likelihood
The least squares



Unbiasedness, consistency,
asymptotic normality

Sufficiency, efficiency,
sequential tests

A small historical chimney-corner

Could You assign some years or persons to this development ?

When did we start to study the statistics?

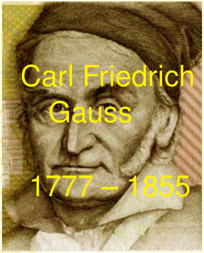
When did we start to study the statistics?


Pierre-Simon Laplace

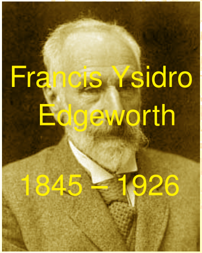
1749 –1827


Adrien-Marie Legendre

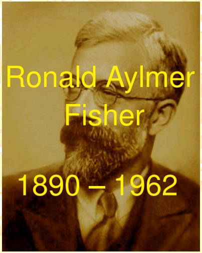
1752 – 1833


Carl Friedrich Gauss

1777 – 1855


Francis Ysidro Edgeworth

1845 – 1926


Ronald Aylmer Fisher

1890 – 1962


Andrej Nikolajevič Kolmogorov

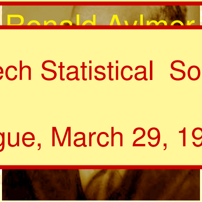
1903 – 1987

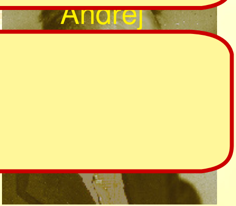
When did we start to study the statistics?

The first statistical society in the world

- STATISTICAL SOCIETY in LONDON -

founded on February 21, 1834





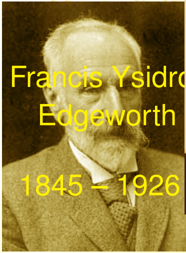
the Czech Statistical Society

- Prague, March 29, 1990

When did we start to study the statistics?

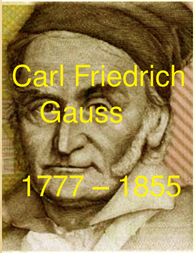
Pierre-Simon Laplace

1749 –1827



Francis Ysidr
Edgeworth

1845 – 1926



Carl Friedrich Gauss

1777 – 1855

Andrej Nikolajevič Kolmogorov

1903 – 1987

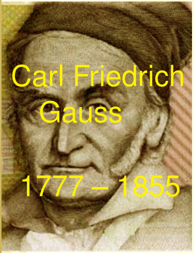
When did we start to study the statistics?

Pierre-Simon Laplace

1749 –1827

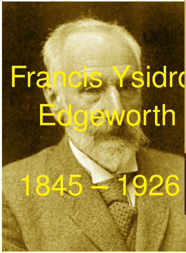
Jaroslav Hájek

1926 – 1974



Carl Friedrich Gauss

1777 – 1855



Francis Ysidr
Edgeworth

1845 – 1926

Andrej Nikolajevič Kolmogorov

1903 – 1987

Repeating from previous lecture - small deviation from exact model can cause ...

Huber, P. J. (1980): *Robust Statistics.*
New York: J.Wiley and Sons.

$$s_n = \left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \right]^{\frac{1}{2}} \qquad d_n = \frac{\pi}{2n} \sum_{i=1}^{n} |x_i - \bar{x}_n|$$

$$F(x) = (1 - \varepsilon)\Phi(x) + \varepsilon\Phi(\frac{x}{3})$$

$$ARE_F(\varepsilon) = \lim_{n \to \infty} \frac{\mathrm{var}_F s_n / \mathbb{E}_F^2 s_n}{\mathrm{var}_F d_n / \mathbb{E}_F^2 d_n}$$

Small deviation from exact model can cause ...

ε	0	0.001	0.002	0.05
$ARE(\varepsilon)$	0.876	0.948	1.016	2.035

So, 5% of contamination $\rightarrow$ d_n is two times better than s_n.

Is 5% contamination too much or too little?

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel. (1986):
Robust Statistic - The Approach Based on Influence Curve.
New York: J.Wiley and Sons.

E. g. Switzerland has 6% of errors in mortality tables.

Is the efficiency really important or a bit misleading?

Fisher, R. A. (1922): On the mathematical foundation of theoretical statistics.
Philos. Trans. Roy. Soc. London Ser. A 222, 309 - 368.

$$\lim_{n\to\infty} \frac{\mathrm{var}_{N(0,1)}(\overline{x}_n)}{\mathrm{var}_{t(\nu)}(\overline{x}_n)} = 1 - \frac{6}{\nu(\nu+1)}$$

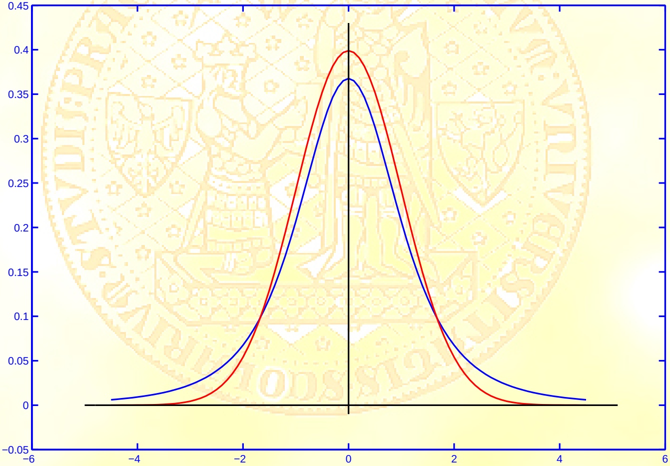
$$\lim_{n\to\infty} \frac{\mathrm{var}_{N(0,1)}(s_n^2)}{\mathrm{var}_{t(\nu)}(s_n^2)} = 1 - \frac{12}{\nu(\nu+1)}$$

Is the efficiency really important or a bit misleading?

$\lim_{n\to\infty} \frac{\mathrm{var}_{N(0,1)}(T_n)}{\mathrm{var}_{t(\nu)}(T_n)}$	t_9	t_5	t_3
$\bar{x}_n$	0.93	0.80	0.50
s_n^2	0.83	0.40	0!

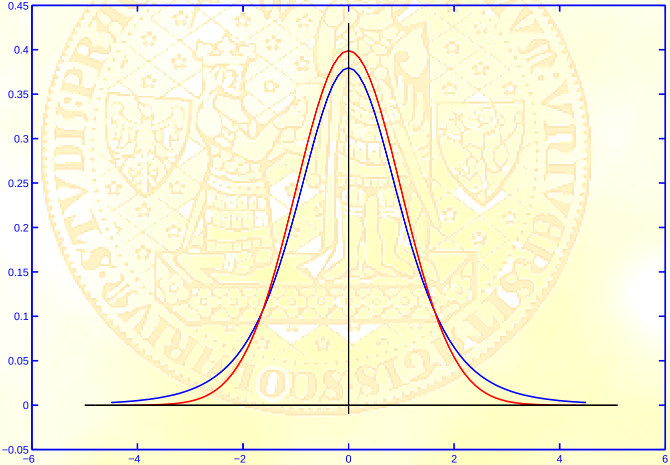
How far is Student density from the normal one ?

THE BLUE CURVE IS STANDARD NORMAL WHILE THE RED ONE IS THE STUDENT'S WITH 3 DEGREES OF FREEDOM.



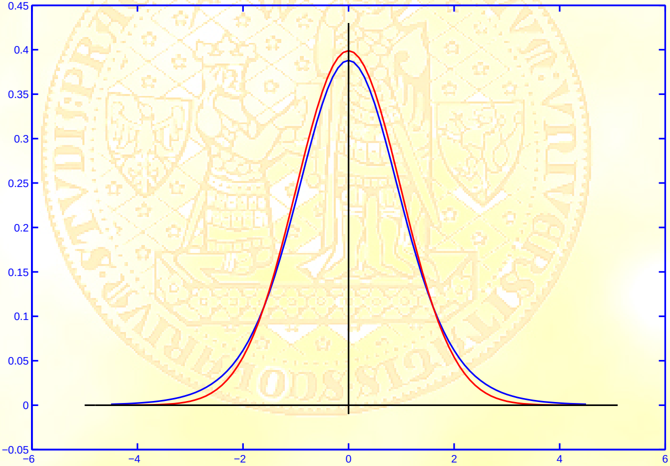
How far is Student density from the normal one ?

THE BLUE CURVE IS STANDARD NORMAL WHILE THE RED ONE IS THE STUDENT'S WITH 5 DEGREES OF FREEDOM.



How far is Student density from the normal one ?

THE BLUE CURVE IS STANDARD NORMAL WHILE THE RED ONE IS THE STUDENT'S WITH 9 DEGREES OF FREEDOM.



> *A tacit hope in ingnoring deviations from ideal models was that they would not matter; that statistical procedures which were optimal under strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians.*
>
> *John W. Tukey (1960)*

Let's study general reasons causing it - returning a few slides back.

Maximum likelihood - solving an extremal problem

$$\hat{\theta}^{(ML,n)} = \underset{\theta \in \Theta}{\arg\max} \prod_{i=1}^{n} f(x_i, \theta)$$

$$\hat{\theta}^{(ML,n)} = \underset{\theta \in \Theta}{\arg\max} \sum_{i=1}^{n} log\left(f(x_i, \theta)\right)$$

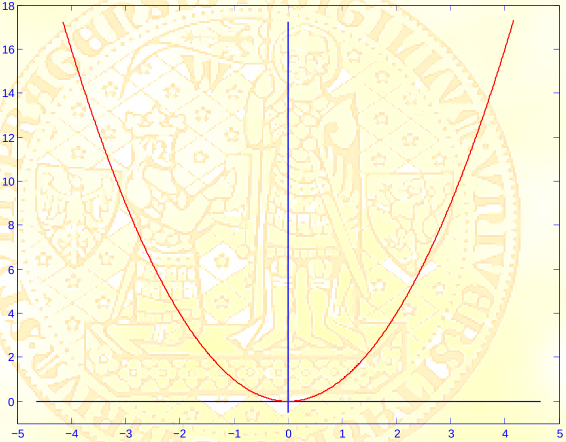
Let again $f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{\frac{(x-\mu)^2}{2\sigma^2}\right\}$ and consider only μ

$$\Rightarrow \quad \hat{\mu}^{(ML,n)} = \underset{\mu \in R}{\arg\min} \left\{\sum_{i=1}^{n} (x_i - \mu)^2\right\}$$

The observations with large $(x_i - \mu)^2$
have a large influence on solution.

Evidently, low robustness is consequence of quadratic objective function

We have such objective function.



We should depress influence of large residuals.

Let's study general reasons causing it - an alternative way.

Maximum likelihood - solving the normal equations

$$\hat{\theta}^{(ML,n)} = \underset{\theta\in\Theta}{\arg\max} \prod_{i=1}^{n} f(x_i,\theta) = \underset{\theta\in\Theta}{\arg\max} \sum_{i=1}^{n} log\left(f(x_i,\theta)\right)$$

$$\hat{\theta}^{(ML,n)} = \underset{\theta\in\Theta}{\arg} \sum_{i=1}^{n} \frac{1}{f(x_i,\theta)} \cdot \frac{\partial f(x_i,\theta)}{\partial\theta} = 0$$

Let again $f(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} exp\left\{\frac{(x-\mu)^2}{2\sigma^2}\right\}$, i. e. $\frac{\partial f(x_i,\theta)}{\partial\mu} = f(x_i,\mu,\sigma^2)\cdot\frac{(x_i-\mu)}{\sigma^2}$

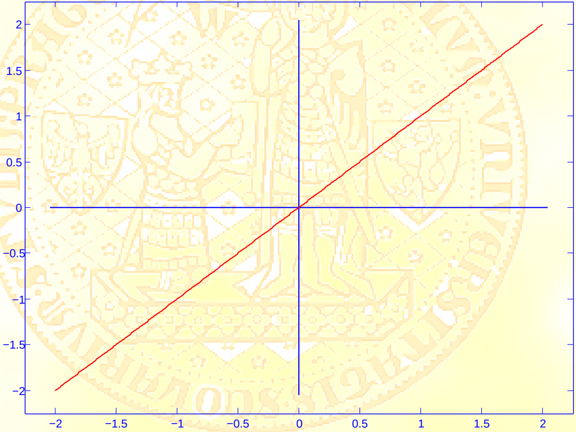
and consider only μ $\quad\Rightarrow\quad \hat{\mu}^{(ML,n)} = \underset{\mu\in R}{\arg} \left\{\sum_{i=1}^{n}(x_i-\mu) = 0\right\}$

The same conclusion:

The observations with large $|x_i - \mu|$
have a large influence on solution.

Equivalently, low robustness is consequence of identity in normal equations

We have such influence function.



We should depress influence of large residuals.

Motivation by historical experience

1. Ancient Egyptians and medieval French,
2. Sir John William Rayleigh, Nobel Prize for Physics, 1904
   (William Ramsay, Nobel Prize in chemistry, 1904)
   - 7 out of 15 atomic weight of "nitrogen" ⇒ argon,
3. J. B. Leon Foucault - 19. century,
   Albert Abraham Michelson - 1920 improved the method
   - 12 out of 16 measurements of light velocity.

(Remember Foucalt pendulum, 1851.)

The main goals of robust statistics

1. To describe the structure best fitting the bulk of data.
2. To identify deviating data points (outliers) or deviating substructures for further treatment, if desired.
3. To identify and give a warning about highly influential data points (leverage points).
4. To deal with unsuspected serial correlation, or more generally, with deviations from the assumed correlation structures.

The four main types of deviations from the strict parametric model

1. The occurence of gross errors.
2. Rounding and grouping.
3. The model may have been conceived as an approximation anyway, e.g., by virtue of CLT.
4. Apart of distributional assumptions, the assumption of independence (or of some specific correlation structure) may only be approximately fulfilled.

How have we attempted to cope with these tasks ?

Three approaches:

1. Huber's alternative to classical point estimation via neighbourhoods.
2. Huber's alternative to classical testing hypotheses via capacities.
3. Hampel's infinitesimal approach via Prokhorov metric and influence function.

Huber's proposal to robustify point estimation

1. Denote by $\mathcal{H}$ the set of all distribution functions (d. f.'s).
2. Select one fix $F_{\theta_0} \in \mathcal{F}_\Theta = \{F_\theta\}_{\theta \in \Theta} \subset \mathcal{H}$

   (called *the parent or central distribution*),

   fix also some $H \in \mathcal{H}^* \subset \mathcal{H}$ and $\delta > 0$. Then put

   $$G_{\theta_0,\delta,H}(x) = (1-\delta)F_{\theta_0}(x) + \delta H(x).$$
3. Fix $\varepsilon \in (0,1)$ and put

   $$\mathcal{G}_{\Theta,\varepsilon,\mathcal{H}^*} = \{G_{\theta,\delta,H}\}_{\theta\in\Theta,\delta\le\varepsilon,H\in\mathcal{H}^*}.$$
4. Use $\mathcal{G}_{\Theta,\varepsilon,\mathcal{H}^*}$ instead of $\mathcal{F}_\Theta$

   in the usual approach of classical statistics.

(An example on the next slide)

Huber's proposal to robustify point estimation - an example

1. Previous lecture has recalled ML estimation

$$\hat{\theta}^{(ML,n)} = \arg\max_{\theta \in \Theta} \ log \left\{ \prod_{i=1}^{n} f(x_i, \theta) \right\}, \qquad (1)$$

2. specifying example with $f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$.

3. Putting for $\Theta = (R \times R^+)$ and $\theta = (\mu, \sigma^2)$

$$\mathcal{F}_\Theta = \left\{ N(x, \mu, \sigma^2) \right\}_{(\mu,\sigma^2) \in (R \times R^+)}$$

we can write (1) as

$$\hat{\theta}^{(ML,n)} = (\hat{\mu}^{(ML,n)}, \hat{\sigma}^{(ML,n)}) = \arg\max_{f \in \mathcal{F}_\Theta} \ log \left\{ \prod_{i=1}^{n} f(x_i, \theta) \right\}.$$

(The item 3 is rewritten on the next slide.)

Huber's proposal to robustify point estimation - an example

1. Putting for $\Theta = (R \times R^{+})$ and $\theta = (\mu, \sigma^2)$

$$\mathcal{F}_{\Theta} = \left\{ N(x, \mu, \sigma^2) \right\}_{(\mu, \sigma^2) \in (R \times R^{+})}$$

we can write (1) as

For details see:

Huber, P. J. (1964): Robust estimation of a location parameter. *Ann. Math. Statist. 35, 73–101.*

$$\hat{\theta}^{(ML,n,\varepsilon)} = (\hat{\mu}^{(ML,n,\varepsilon)}, \hat{\sigma}^{(ML,n,\varepsilon)}) = \underset{g \in \mathcal{G}_{\Theta,\varepsilon}}{\arg\max} \; log \left\{ \prod_{i=1}^{n} g\left(x_i, \mu, \sigma^2, \delta, H\right) \right\}$$

where $g\left(x_i, \mu, \sigma^2, \delta, H\right)$ is the density of

$$G_{\theta,\delta,H} = G_{\mu,\sigma^2,\delta,H} = (1-\delta)\Phi(x, \mu, \sigma^2) + \delta H(x)$$

and $\mathcal{G}_{\Theta,\varepsilon,\mathcal{H}^*} = \left\{ G_{\theta,\delta,H} \right\}_{\theta \in \Theta, \delta \le \varepsilon, H \in \mathcal{H}^*}$.

Huber’s proposal to robustify testing hypotheses

For details about the capacities see:

Choquet, G. (1954): Theory of capacities.

Annales de l’institut Fourier, 5 (1954), 131-295.

$$\mathcal{G}_{\Theta,\varepsilon,\mathcal{H}^*} = \{G_{\theta,\delta,H}\}_{\theta\in\Theta,\delta\le\varepsilon,H\in\mathcal{H}^*}\,.$$

For details about the tests see:

Huber, P. J. (1965): A robust version of the probability ratio test.

Ann. Math. Statist. 36, 1753–1758.

A generalized Neyman-Pearson lemma

Hampel's approach - a bit more mathematics

The Hampel's approach is based on two basic ideas and a nice fact:

1. The first idea - any estimator can be interpreted as a function T (say) from the space of all distribution functions $\mathcal{H}$ to the parameter space Θ (say).

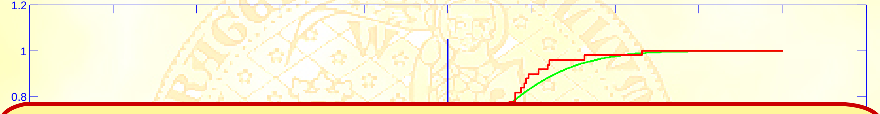
Let's start - by an illustration - with the last topic,
an exact mathematics will be delivered in some next lecture.

3. A nice fact - the Kolmogorov-Smirnov result - the empirical d.f. converge uniformly to the "true" underlying one.

In the case that the function T is defined on so rich space, as $\mathcal{H}$, we usually use for T the "name" *functional* instead of function.

Convergence of emp. d. f. and Kolmogorov-Smirnov distance

Empirical distribution function - 5010015025040007001000 observations.



Kolmogorov, A. (1933):
Sulla determinazione empirica di una legge di distribuzionc
1st. Ital. Attuari. G. 4. 1 - 11.

Smirnov, N. (1939): On the estimation of discrepancy between
empirical curves of distribution for two independent samples.
Bull. Math. Univ. Moscow 2, 3 - 14.

Víšek, J. Á (2011):
Empirical distribution function under heteroscedasticity.
Statistics 45, 497-508.

Making preparation steps for explanation of Hampel's approach

Now, let us turn to the first idea:

Any estimator can be interpreted as
a function T (say) from the space of all distribution
functions $\mathcal{H}$ to the parameter space Θ (say).

Prior to it we need to recall something about the integration of functions.

A preliminary intermezzo - the idea of integral

All of us learnt that the integral, say $\int_a^b g(y)\mathrm{d}y$, is defined as follows:

Let $a = y_0 < y_1 < y_2 < ... < y_n = b$ be an (equdistant) division of the interval $[a, b]$ and for any $i \in \{1, 2, ..., n\}$ let $\tilde{y}_i \in [y_{i-1}, y_i]$. Then put

$$\int_a^b g(y)\mathrm{d}y = \lim_{n\to\infty} \sum_{i=1}^{n} g(\tilde{y}_i)(y_i - y_{i-1}).$$

The integral represents the area under the function $g(y)$.

We say that the integral is computed

with respect to Lebesgue measure $y_i - y_{i-1}$

- it is indicated by $\mathrm{d}y$.

A preliminary intermezzo - the idea of the mean value of r. v.

Let F be a d. f., f its density (in a general sense covering continuous as well as discrete r. v.'s) and Y random variable distributed according to F. Then the mean value of Y is given as

$$\mathbb{E}_F Y = \int_{-\infty}^{\infty} y\, f(y)\, \mathrm{d}y = \lim_{n\to\infty} \sum_{i=1}^{n} \tilde{y}_i f(\tilde{y}_i)(y_i - y_{i-1}).$$

Notice the subindex in $\mathbb{E}_F$

indicating that the mean value was taken with respect to d. f. F.

But $f(\tilde{y}_i)(y_i - y_{i-1}) \approx F(y_i) - F(y_{i-1})$.

So, we can see the mean value also as

$$\mathbb{E} Y = \lim_{n\to\infty} \sum_{i=1}^{n} \tilde{y}_i \Big(F(y_i) - F(y_{i-1}) \Big) = \int_{-\infty}^{\infty} y \mathrm{d}F(y).$$

A preliminary intermezzo - the idea of the mean value of r. v.

Let F_n be an empirical d. f. and Y a random variable.

Then the mean value of Y with respect to F_n is given as

$$\mathbb{E}_{F_n} Y = \lim_{n \to \infty} \sum_{i=1}^{n} \tilde{y}_i \left(F_n(y_i) - F_n(y_{i-1}) \right) .$$

But $F_n(y_i) - F_n(y_{i-1})$ is either $\frac{1}{n}$ or 0, i. e.

$$\mathbb{E}_{F_n} Y = \frac{1}{n} \sum_{i=1}^{n} \tilde{y}_i$$

where $\tilde{y}_i$'s are points at which is jump equal to $\frac{1}{n}$.
Finally, let's recall that we have denoted the set of all d. f.'s by $\mathcal{H}$

(in what follows we'll need it).

The Hampel approach

Estimator as a function of distribution function

1. Consider e. g. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
2. Let $F_n(.) \in \mathcal{H}$ be an empirical d. f. corresponding to the observations x … $x_1, x_2, ..., x_n)$.
3. If we … , we obt … $= \mathbb{E}X$ … estimator.
4. Typica … rt so that we can write $\hat{\theta}^{(n)} = T_n(F_n)$ and $\theta = T(F)$, where F_n is the empirical d. f. corresponding to the underlying d. f..

Realize that $F_n \to F$ (Kolmogorov-Smirnov)
and immediately is clear why we adopt this idea.

If $T(.)$ is coninuous (at the point $F \in \mathcal{H}$),
we have $\hat{\theta}^{(n)} = T(F_n) \to T(F) = \theta$.

Making preparation steps for explanation of Hampel's approach

Now, let us turn to the second idea:

The function T

can be studied by an infinitesimal calculus of limits, derivaties, integrals, etc.

Prior to it we need to carry out some preliminary explanation about the uncountably dimensional vector spaces.

Countable versus uncountable - notion of cardinality

1. Two sets have the same cardinality (mohutnost),
   if they have the "same number of elements".
2. The "same number of elements" means that
   we can find a one-to-one mapping of one set on the other.
3. An example

$$\begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} \begin{matrix} \longleftrightarrow \\ \longleftrightarrow \\ \longleftrightarrow \\ \longleftrightarrow \end{matrix} \begin{bmatrix} W \\ X \\ Y \\ Z \end{bmatrix}$$

4. An example with infinite number of objects is on the next slide
   - consider the set of all positive integers
   and the set of all positive rational numbers.

Positive integers versus positive rational numbers

So, let's study cardinalities of the set of positive integers, say $\mathcal{N}$,
and of set of positive rational numbers, $\mathcal{R}$.

1. Imagine interval [0, 1].
2. Inside … the type:
   … nfinity.

Realize that although our conclusion was wrong,
we proved that $\#\mathcal{N} \leq \#\mathcal{R}$.

3. So, we have much more positive rational numbers
   than the positive integers, haven't we?

Surprising correct answer is:
The set of positive integers
has the same cardinality as the set of positive rational numbers.

Let's star to construct the mapping.

Positive integers versus positive rational numbers

The set of positive rational numbers, say $\mathcal{R}$
(some rationals are repeated in this table, e. g. $\frac{1}{2}$ and $\frac{2}{4}$, etc):

$$\begin{array}{ccccccc}
\frac{1}{1} & \frac{2}{1} & \frac{3}{1} & \frac{4}{1} & \frac{5}{1} & \frac{6}{1} & \cdots \\
\frac{1}{2} & \frac{2}{2} & \frac{3}{2} & \frac{4}{2} & \frac{5}{2} & \frac{6}{2} & \cdots \\
\frac{1}{3} & \frac{2}{3} & \frac{3}{3} & \frac{4}{3} & \frac{5}{3} & \frac{6}{3} & \cdots \\
\frac{1}{4} & \frac{2}{4} & \frac{3}{4} & \frac{4}{4} & \frac{5}{4} & \frac{6}{4} & \cdots \\
\frac{1}{5} & \frac{2}{5} & \frac{3}{5} & \frac{4}{5} & \frac{5}{5} & \frac{6}{5} & \cdots \\
\frac{1}{6} & \frac{2}{6} & \frac{3}{6} & \frac{4}{6} & \frac{5}{6} & \frac{6}{6} & \cdots \\
\frac{1}{7} & \frac{2}{7} & \frac{3}{7} & \frac{4}{7} & \frac{5}{7} & \frac{6}{7} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots
\end{array}$$

Let’s star to construct the mapping.

Positive integers versus positive rational numbers

Starting the construction of mapping $\mathcal{R}$ on $\mathcal{N}$ - the firstsecond step:

$$\begin{array}{ccccccc}
\frac{1}{1} \to 1 & \frac{2}{1} \to 3 & \frac{3}{1} & \frac{4}{1} & \frac{5}{1} & \frac{6}{1} & \cdots \\
\frac{1}{2} \to 2 & \frac{2}{2} & \frac{3}{2} & \frac{4}{2} & \frac{5}{2} & \frac{6}{2} & \cdots \\
\frac{1}{3} & \frac{2}{3} & \frac{3}{3} & \frac{4}{3} & \frac{5}{3} & \frac{6}{3} & \cdots \\
\frac{1}{4} & \frac{2}{4} & \frac{3}{4} & \frac{4}{4} & \frac{5}{4} & \frac{6}{4} & \cdots \\
\frac{1}{5} & \frac{2}{5} & \frac{3}{5} & \frac{4}{5} & \frac{5}{5} & \frac{6}{5} & \cdots \\
\frac{1}{6} & \frac{2}{6} & \frac{3}{6} & \frac{4}{6} & \frac{5}{6} & \frac{6}{6} & \cdots \\
\frac{1}{7} & \frac{2}{7} & \frac{3}{7} & \frac{4}{7} & \frac{5}{7} & \frac{6}{7} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots
\end{array}$$

Positive integers versus positive rational numbers

Constructing the mapping $\mathcal{R}$ on $\mathcal{N}$ - the third step:

$$\begin{array}{ccccccc}
\frac{1}{1} \to 1 & \frac{2}{1} \to 3 & \frac{3}{1} \to 6 & \frac{4}{1} & \frac{5}{1} & \frac{6}{1} & \cdots \\
\frac{1}{2} \to 2 & \frac{2}{2} \to 5 & \frac{3}{2} & \frac{4}{2} & \frac{5}{2} & \frac{6}{2} & \cdots \\
\frac{1}{3} \to 4 & \frac{2}{3} & \frac{3}{3} & \frac{4}{3} & \frac{5}{3} & \frac{6}{3} & \cdots \\
\frac{1}{4} & \frac{2}{4} & \frac{3}{4} & \frac{4}{4} & \frac{5}{4} & \frac{6}{4} & \cdots \\
\frac{1}{5} & \frac{2}{5} & \frac{3}{5} & \frac{4}{5} & \frac{5}{5} & \frac{6}{5} & \cdots \\
\frac{1}{6} & \frac{2}{6} & \frac{3}{6} & \frac{4}{6} & \frac{5}{6} & \frac{6}{6} & \cdots \\
\frac{1}{7} & \frac{2}{7} & \frac{3}{7} & \frac{4}{7} & \frac{5}{7} & \frac{6}{7} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots
\end{array}$$

Positive integers versus positive rational numbers

Constructing the mapping $\mathcal{R}$ on $\mathcal{N}$ - the fourth step , etc.:

$$\begin{array}{ccccccc}
\frac{1}{1} \rightarrow 1 & \frac{2}{1} \rightarrow 3 & \frac{3}{1} \rightarrow 6 & \frac{4}{1} \rightarrow 10 & \frac{5}{1} & \frac{6}{1} & \dots \\
\frac{1}{2} \rightarrow 2 & \frac{2}{2} \rightarrow 5 & \frac{3}{2} \rightarrow 9 & \frac{4}{2} & \frac{5}{2} & \frac{6}{2} & \dots \\
\frac{1}{3} \rightarrow 4 & \frac{2}{3} \rightarrow 8 & \frac{3}{3} & \frac{4}{3} & \frac{5}{3} & \frac{6}{3} & \dots \\
\frac{1}{4} \rightarrow 7 & \frac{2}{4} & \frac{3}{4} & \frac{4}{4} & \frac{5}{4} & \frac{6}{4} & \dots \\
\frac{1}{5} & \frac{2}{5} & \frac{3}{5} & \frac{4}{5} & \frac{5}{5} & \frac{6}{5} & \dots \\
\frac{1}{6} & \frac{2}{6} & \frac{3}{6} & \frac{4}{6} & \frac{5}{6} & \frac{6}{6} & \dots \\
\frac{1}{7} & \frac{2}{7} & \frac{3}{7} & \frac{4}{7} & \frac{5}{7} & \frac{6}{7} & \dots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots
\end{array}$$

Q.E.D.

Positive rational numbers versus positive irrational numbers

What about cardinalities of the set of positive rational numbers $\mathcal{R}$ and of set of positive irrational numbers, say $\mathcal{I}$.

1. Between any two positive rational numbers is at least one positive irrational number.
2. Between any two positive irrational numbers is at least one positive rational numbers.
3. So, we have the same number of positive rational numbers and of positive irrational numbers, haven't we?

Surprising correct answer is:

The set of positive irrational numbers has (much) larger cardinality than the set of positive rational numbers.

Cardinality of $\mathcal{R}$ (positive rationals), $\mathcal{I}$ (positive irrationals) and $\mathcal{RE}$ (positive real numbers)

1. We already know - $\mathcal{R}$ is countable.
2. If $\mathcal{I}$ be countable $\rightarrow$ $\mathcal{RE}$ is countable.
3. However, we'll prove that $\mathcal{RE}$ is uncountable.
4. Hence $\mathcal{I}$ is uncountable.

Famous diagonal argument by Georg Cantor (1845 - 1918)

Let's assume that the set of real numbers between 0 and 1 is countable. Assume that we have all real numbers between 0 and 1, ordered into a sequence:

(the upper index indicates position of number in question in this sequence)

$$\begin{array}{cccccccccc}
0. & c_1^{(1)} & c_2^{(1)} & c_3^{(1)} & c_4^{(1)} & c_5^{(1)} & c_6^{(1)} & c_7^{(1)} & c_8^{(1)} & \ldots \\
0. & c_1^{(2)} & c_2^{(2)} & c_3^{(2)} & c_4^{(2)} & c_5^{(2)} & c_6^{(2)} & c_7^{(2)} & c_8^{(2)} & \ldots \\
0. & c_1^{(3)} & c_2^{(3)} & c_3^{(3)} & c_4^{(3)} & c_5^{(3)} & c_6^{(3)} & c_7^{(3)} & c_8^{(3)} & \ldots \\
0. & c_1^{(4)} & c_2^{(4)} & c_3^{(4)} & c_4^{(4)} & c_5^{(4)} & c_6^{(4)} & c_7^{(4)} & c_8^{(1)} & \ldots \\
\vdots & & \vdots & & \vdots & \vdots & & \vdots & &
\end{array}$$

Let's rewrite this scheme on the next slide.

Famous diagonal argument by Georg Cantor (1845 - 1918)

$$\begin{array}{l}
0.\ c_1^{(1)}\ c_2^{(1)}\ c_3^{(1)}\ c_4^{(1)}\ c_5^{(1)}\ c_6^{(1)}\ c_7^{(1)}\ c_8^{(1)}\ \ldots \\
0.\ c_1^{(2)}\ c_2^{(2)}\ c_3^{(2)}\ c_4^{(2)}\ c_5^{(2)}\ c_6^{(2)}\ c_7^{(2)}\ c_8^{(2)}\ \ldots \\
0.\ c_1^{(3)}\ c_2^{(3)}\ c_3^{(3)}\ c_4^{(3)}\ c_5^{(3)}\ c_6^{(3)}\ c_7^{(3)}\ c_8^{(3)}\ \ldots \\
0.\ c_1^{(4)}\ c_2^{(4)}\ c_3^{(4)}\ c_4^{(4)}\ c_5^{(4)}\ c_6^{(4)}\ c_7^{(4)}\ c_8^{(1)}\ \ldots \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots
\end{array}$$

Let's create a new real number (the upper index (n) indicates that it is "new" real number):

$$0.\ c_1^{(n)} \neq c_1^{(1)}\ c_2^{(n)} \neq c_2^{(2)}\ c_3^{(n)} \neq c_3^{(3)}\ c_4^{(n)} \neq c_4^{(4)}\ c_5^{(1)} \neq c_5^{(5)}\ c_6^{(1)} \neq c_6^{(6)}\ c_7^{(1)} \neq c_7^{(7)}\ c_8^{(1)} \neq c_8^{(8)}\ \ldots$$

This new number does not coincide with any number in the sequence
and it is a contradiction with the assumption that
we had all real numbers in the sequence we studied above.

More shocking facts !!

We are going to prove a much more surprising result
and what is nearly shocking - it can be done by trivial means.

Cardinality (denote by $\#$) of finite sets and sets of their subsets

Consider a finite set $A = \{1, 2, ..., n\}$, i. e. $\#A = n$.

How much subsets it has (including the whole set A and the empty set)?

The answer is hinted by the scheme :

Label of element	1	2	3	4	...	n
The element is not selected into subsample	0	0	0	0	...	0
The element is selected into subsample	1	1	1	1	...	1

The set of all subsets, say $\mathcal{A}$, has $2^{\#A}$ elements ! So, $\#A < \#\mathcal{A}$.

Does the last inequality hold also for infinite sets,
i. e. is it still true that $\#A < 2^{\#A}$?

More shocking facts !!

We are going to prove a much more surprising result
and what is nearly shocking - it can be done by trivial means.

The cardinality of a set and the set of all its subsets.

1. Consider any set A and all its subsets, say $\mathcal{A}$.
2. Assume that there is a one-to-one mapping of A and $\mathcal{A}$, say $\kappa : A \to \mathcal{A}$ - i. e.

$$\forall (s \in A) \qquad \exists (S \in \mathcal{A}) \qquad \text{so that} \quad S = \kappa(s).$$

We proved that it holds generally that $\#A < \#\mathcal{A}$.
The set of real numbers is (more or less) the same thing as the set of all subsets of rational numbers.

5. Assume that $\tilde{s} \in \tilde{S}$. But point 3 then implies that $\tilde{s} \notin \tilde{S}$.
6. So, assume that $\tilde{s} \notin \tilde{S}$. But point 3 then implies that $\tilde{s} \in \tilde{S}$.

It is a contradiction, Q.E.D.

Making preparation steps for explanation of Hampel's approach

DEFINITION

1. We say that the set is finite, if it has finite number of points.
2. We say that the set *A* is countable if its cardinality is the same as cardinality of the set of (all positive) integers.
3. Otherwise, we say that the set is uncountable.

EXAMPLES:

- The sets of (all) rational numbers is countable.
- The sets of (all) irrational numbers is uncountable.
- The sets of (all) real numbers is uncountable.

Making preparation steps for explanation of Hampel's approach

Recalling the notion of vector space

1. Consider p-dimensional vector space, say $\mathcal{U}$

   $\rightarrow$ then any vector $u \in \mathcal{U}$ has coordinates $u_i, i \in \{1, 2, ..., p\}$.

2. We can imagine that we have for any point $u \in \mathcal{U}$ one mapping from the set $\{1, 2, ..., p\}$ to the real line, i. e. for point u we have the mapping $u(.)$ such that if we plug in some i from $\{1, 2, ..., p\}$, we obtain the i-th coordinate of u.

Making preparation steps for explanation of Hampel's approach

Slightly generalize the notion of vector space

1. Consider countably-dimensional vector space, say $\mathcal{Z}$

   → then any vector $z \in \mathcal{Z}$ has coordinates z_i's, $i \in \{1, 2, ...\}$.

2. We can again imagine that we have for any point $z \in \mathcal{Z}$ one mapping from the set $\{1, 2, ...\}$ to the real line, i. e. for point z we have the mapping $z(.)$ such that if we plug in some i from $\{1, 2, ...\}$, we obtain the i-th coordinate of z.

Making preparation steps for explanation of Hampel's approach

Slightly generalize the notion of vector space

1. Finally, consider uncountably-dimensional vector space, say $\mathcal{F}$
   $\rightarrow$ then any vector $F \in \mathcal{F}$ has coordinates $F(x), x \in R$, say.
2. We can again imagine that we have for any point of $\mathcal{F}$ one mapping from the set R to the real line, i. e. for point F we have the mapping $F(.)$ such that if we plug in some x from R, we obtain the x-th coordinate of F.

An example
In the previous lecture we met with the space of all distribution function $\mathcal{H}$. It is uncountably-dimensional vector space. Every d.f. F, including the empirical ones, is one point in it - convolution.

THANKS FOR ATTENTION