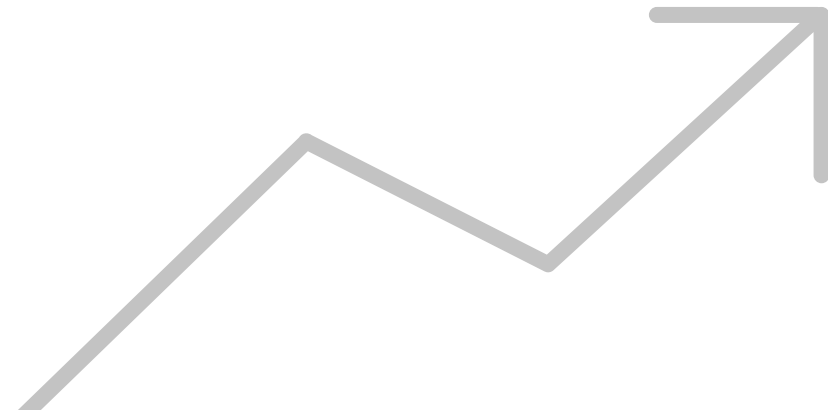


Data Science-Based Innovation in Official Statistics

German Data Science Days 2022



Official Statistics

Is not about ...

60 3. Uniformly Most Powerful Tests

yields the largest number of miles per hour. Analogously in the present problem the most valuable points x are those with the highest value of

$$r(x) = \frac{P_1(x)}{P_0(x)}$$

The points are therefore rated according to the value of this ratio and selected for S in this order, as many as one can afford under restriction (3.6). Formally this means that S is the set of all points x for which $r(x) > c$, where c is determined by the condition

$$P_0\{X \in S\} = \sum_{x:r(x)>c} P_0(x) = \alpha.$$

Here a difficulty is seen to arise. It may happen that when a certain point is included, the value α has not yet been reached but that it would be exceeded if the point were also included. The exact value α can then either not be achieved at all, or it can be attained only by breaking the preference order established by the points were also included. The exact value α can then either not be achieved at all, or it can be attained only by breaking the preference order established by the points were also included. The exact value α can then either not be achieved at all, or it can be attained only by breaking the preference order established by the points were also included.

Theorem 3.2.1 Let P_0 and P_1 be probability distributions possessing densities p_0 and p_1 respectively with respect to a measure μ .

(i) Existence. For testing $H: p_0$ against the alternative $K: p_1$ there exists a test ϕ and a constant k such that

$$E_0\phi(X) = \alpha \quad (3.7)$$

and

$$\phi(x) = \begin{cases} 1 & \text{when } p_1(x) > kp_0(x), \\ 0 & \text{when } p_1(x) < kp_0(x). \end{cases} \quad (3.8)$$

(ii) Sufficient condition for a most powerful test. If a test satisfies (3.7) and (3.8) for some k , then it is most powerful for testing p_0 against p_1 at level α .

(iii) Necessary condition for a most powerful test. If ϕ is most powerful at level α for testing p_0 against p_1 , then for some k it satisfies (3.8) a.e. μ . It also satisfies (3.7) unless there exists a test of size $< \alpha$ and with power 1.

PROOF. For $\alpha = 0$ and $\alpha = 1$ the theorem is easily seen to be true provided the value $k = +\infty$ is admitted in (3.8) and $0, \infty$ is interpreted as 0 . Throughout the proof we shall therefore assume $0 < \alpha < 1$.

¹In practice, typically either the breaking of the r -order nor randomization is considered acceptable. The common solution, instead, is to adopt a value of α that can be attained exactly and therefore does not present this problem.

²There is no loss of generality in this assumption, since one can take $\mu = P_0 + P_1$.

(i): Let $\alpha(c) = P_0\{p_1(X) > cp_0(X)\}$. Since the probability is computed under P_0 , the inequality need be considered only for the set where $p_0(x) > 0$, so that $\alpha(c)$ is the probability that the random variable $p_1(X)/p_0(X)$ exceeds c . Thus $\alpha(c)$ is a cumulative distribution function, and $\alpha(c)$ is nonincreasing and $1 - \alpha(c)$ on the right, $\alpha(c-0) - \alpha(c) = P_0\{p_1(X)/p_0(X) = c\}$, $\alpha(-\infty) = 1$, and $\alpha(\infty) = 0$. Given any $0 < \alpha < 1$, let c_0 be such that $\alpha(c_0) \leq \alpha < \alpha(c_0-0)$, and consider the test ϕ defined by

$$\phi(x) = \begin{cases} 1 & \text{when } p_1(x) > c_0 p_0(x), \\ \frac{\alpha - \alpha(c_0)}{\alpha(c_0-0) - \alpha(c_0)} & \text{when } p_1(x) = c_0 p_0(x), \\ 0 & \text{when } p_1(x) < c_0 p_0(x). \end{cases}$$

Here the middle expression is meaningful unless $\alpha(c_0) = \alpha(c_0-0)$; since then $P_0\{p_1(X) = c_0 p_0(X)\} = 0$, ϕ is defined a.e. The size of ϕ is

$$E_0\phi(X) = P_0\left\{\frac{p_1(X)}{p_0(X)} > c_0\right\} + \frac{\alpha - \alpha(c_0)}{\alpha(c_0-0) - \alpha(c_0)} P_0\left\{\frac{p_1(X)}{p_0(X)} = c_0\right\} = \alpha,$$

so that c_0 can be taken as the k of the theorem.

(ii): Suppose that ϕ is a test satisfying (3.7) and (3.8) and that ϕ^* is any other test with $E_0\phi^*(X) \leq \alpha$. Denote by S^+ and S^- the sets in the sample space where $\phi(x) - \phi^*(x) > 0$ and < 0 respectively. If x is in S^+ , $\phi(x)$ must be > 0 and $p_1(x) \geq kp_0(x)$. In the same way $p_1(x) \leq kp_0(x)$ for all x in S^- , and hence

$$\int_{S^+ \cup S^-} (\phi - \phi^*)(p_1 - kp_0) d\mu = \int_{S^+ \cup S^-} (\phi - \phi^*)(p_1 - kp_0) d\mu \geq 0.$$

The difference in power between ϕ and ϕ^* therefore satisfies

$$\int (\phi - \phi^*) p_1 d\mu \geq k \int (\phi - \phi^*) p_0 d\mu \geq 0,$$

as was to be proved.

(iii): Let ϕ^* be most powerful at level α for testing p_0 against p_1 , and let ϕ satisfy (3.7) and (3.8). Let S be the intersection of the set $S^+ \cup S^-$, on which ϕ and ϕ^* differ, with the set $\{x: p_1(x) \geq kp_0(x)\}$, and suppose that $\mu(S) > 0$. Since $(\phi - \phi^*)(p_1 - kp_0)$ is positive on S , it follows from Problem 2.4 that

$$\int_{S^+ \cup S^-} (\phi - \phi^*)(p_1 - kp_0) d\mu = \int_S (\phi - \phi^*)(p_1 - kp_0) d\mu > 0,$$

and hence that ϕ is more powerful against p_1 than ϕ^* . This is a contradiction, and therefore $\mu(S) = 0$, as was to be proved.

If ϕ^* were of size $< \alpha$ and power < 1 , it would be possible to include in the rejection region additional points or portions of points, and thereby to increase the power until either the power is 1 or the size is α . Thus either $E_0\phi^*(X) = \alpha$ or $E_1\phi^*(X) = 1$. ■

The proof of part (iii) shows that the most powerful test is uniquely determined by (3.7) and (3.8) except on the set on which $p_1(x) = kp_0(x)$. On this set, ϕ can be defined arbitrarily provided the resulting test has size α . Actually, we have shown that there is always to define ϕ to be constant over this boundary set. In the trivial case that there exists a test of power 1, the constant k of (3.8) is 0, and one will accept H for all points for which $p_1(x) = kp_0(x)$ even though the test may then have size $< \alpha$.

386

D. Neal and S. Rosen

economy at large, the model helps us understand the nature, determinants, and distribution of economic rents in labor markets. It also illustrates how selection may generate asymmetric, long-tailed densities for earnings distributions even when the distribution of potential earnings in each job is symmetric.

2.1. The Roy model

Consider an economy with n different types of jobs. Earnings are worker specific because worker productivity varies from job to job, and some workers are better at some jobs than others. Assume that supply and demand are equal in all job markets and examine the assignment of workers to jobs in market equilibrium. Let y_{ij} be the earnings capacity of worker i on job j . It is what worker i could earn in job j . Let y_i be the earnings actually observed for worker i . Maintaining the economic hypothesis that self-interested workers choose jobs that maximize their earnings, observed earnings for worker i are

$$y_i = \max\{y_{i1}, y_{i2}, \dots, y_{in}\}. \quad (2.1)$$

The model is completed by specifying a joint probability distribution $f(y_1, y_2, \dots, y_n)$ for earnings prospects in the population at large. Roy (1951) assumed that $f(y)$ is jointly normal. Mandelbrot (1962) assumed that $f(\cdot)$ follows the Pareto-Levy form. The observed earnings distribution, $g(y)$, is the transformation of $f(y)$ implied by Eq. (2.1).

This simple model yields surprisingly complicated outcomes. The observed distribution $g(y)$ is a mixture of conditional distributions resulting from maximization. Note that if n is large, Eq. (2.1) suggests the extreme value distribution (the distribution of the first order-statistic) as a possible stationary distribution. The extreme value theorem, like the central limit theorem, proves that the first order-statistic of independent and identically distributed parents tends to a unique limiting distribution, independent of how the parents are distributed. The stationary distribution is a double exponential. It is skewed and leptokurtic, characteristic of the upper tail of earnings distributions.

Though this idea is insightful, the extreme-value theorem cannot be applied here because the component distributions of earning potential y_{ij} in each job are not identically distributed (they have different means and variances) and may not be independent. It proves necessary to work through all the conditioning information to derive the observed distribution $g(y)$. Mandelbrot (1962) shows that $g(y)$ follows the Pareto-Levy distribution if $f(\cdot)$ is Pareto-Levy. In fact, this is the only case for which $g(y)$ is in the same family as its parent $f(\cdot)$. Otherwise the conditional distributions which compose it are truncated, and $g(y)$ is a complicated mixture that is not easily described in closed form.

The model is best illustrated when there are two jobs ($n = 2$). Suppose $f(y_1, y_2)$ is normal (or lognormal), with means μ_1 and μ_2 , variances σ_1^2 and σ_2^2 and covariance σ_{12} .

387

Ch. 7. Theories of the Distribution of Earnings

Fig. 2.1.

The correlation coefficient between earning capacities y_1 and y_2 is $\rho = \sigma_{12}/\sigma_1\sigma_2$. The assignment of workers to jobs is depicted in Fig. 2.1. The probability contours depicted assume $\mu_1 = \mu_2$, $\sigma_1 > \sigma_2$ and $\rho > 0$. Each person in the population is a point in the (y_1, y_2) plane. The probability mass there is proportional to the number of people in the population with those prospects. All workers along the 45° line ($y_1 = y_2$) find the two jobs equally attractive. Earnings in job 1 exceed those in job 2 for all people above the line. They choose job 1 and $y_i = y_{i1}$. Similarly, all people in the region below the 45° line maximize earnings by choosing job 2, so $y_i = y_{i2}$ for them. Therefore the observed distribution $g(y)$ is the weighted sum of two truncated normals, one above the 45° line and the other below it. Maddala (1977) and Heckman and Honore (1990) present complete details of this model.

Economic rents enter this model both in the sense of relative talents ("producer" surplus or comparative advantage) and in the sense of absolute talents ("ability" rents and absolute advantage). In Fig. 2.1, workers with prospects along the equal earnings line represent the extensive margin between job types. Relative to their alternative, they receive zero surplus from their actual job choice. All others are inframarginal, and workers furthest from the 45° line (in either direction) receive the largest surplus from their choices. The length of the (y_{i1}, y_{i2}) vector provides a measure of general ability. Holding the ratio (y_{i1}, y_{i2}) constant, earnings rise with general ability.

With a little stretch, the variances and covariance in the probability contours can be given an economic interpretation. For instance, $\sigma_1 > \sigma_2$ implies a sense in which job 1 is more difficult than job 2. In this case, the outcome in job 1 is more dependent

Official Statistics

Is about ...

Die Statistik für Bundeszwecke (Bundesstatistik) hat im föderativ gegliederten Gesamtsystem der amtlichen Statistik die Aufgabe, **laufend Daten über Massenerscheinungen zu erheben, zu sammeln, aufzubereiten, darzustellen und zu analysieren.** Für sie gelten die Grundsätze der Neutralität, Objektivität und fachlichen Unabhängigkeit. Sie gewinnt die Daten unter Verwendung wissenschaftlicher Erkenntnisse und unter Einsatz der jeweils sachgerechten Methoden und Informationstechniken. Durch die Ergebnisse der Bundesstatistik werden gesellschaftliche, wirtschaftliche und ökologische Zusammenhänge für Bund, Länder einschließlich Gemeinden und Gemeindeverbände, Gesellschaft, Wirtschaft, Wissenschaft und Forschung aufgeschlüsselt. Die Bundesstatistik ist **Voraussetzung für eine am Sozialstaatsprinzip ausgerichtete Politik.** [...]

In the federally structured overall system of official statistics, statistics for federal purposes (federal statistics) have the task of **continuously collecting, collating, processing, presenting and analysing data on mass phenomena.** It is governed by the principles of neutrality, objectivity and professional independence. It collects data using scientific knowledge and appropriate methods and information technology. The results of federal statistics provide a breakdown of social, economic and ecological relationships for the Federation, the Länder including municipalities and municipal associations, society, the economy, science and research. Federal statistics are a **prerequisite for a policy oriented towards the welfare state principle.** [...]

§ 1 BStatG and an unauthorised translation

Official Statistics

We uphold the trustworthiness and enhance the usefulness of our results.

The office continues to develop in its roles as

- a reliable statistics producer,
- digital data manager and centre of data excellence and
- digital, customer-focused information service provider.

The Federal Statistical Office is the guarantor of independent, quality-assured information and part of a sustainable data infrastructure in Germany.

Official Statistics

Destatis in numbers



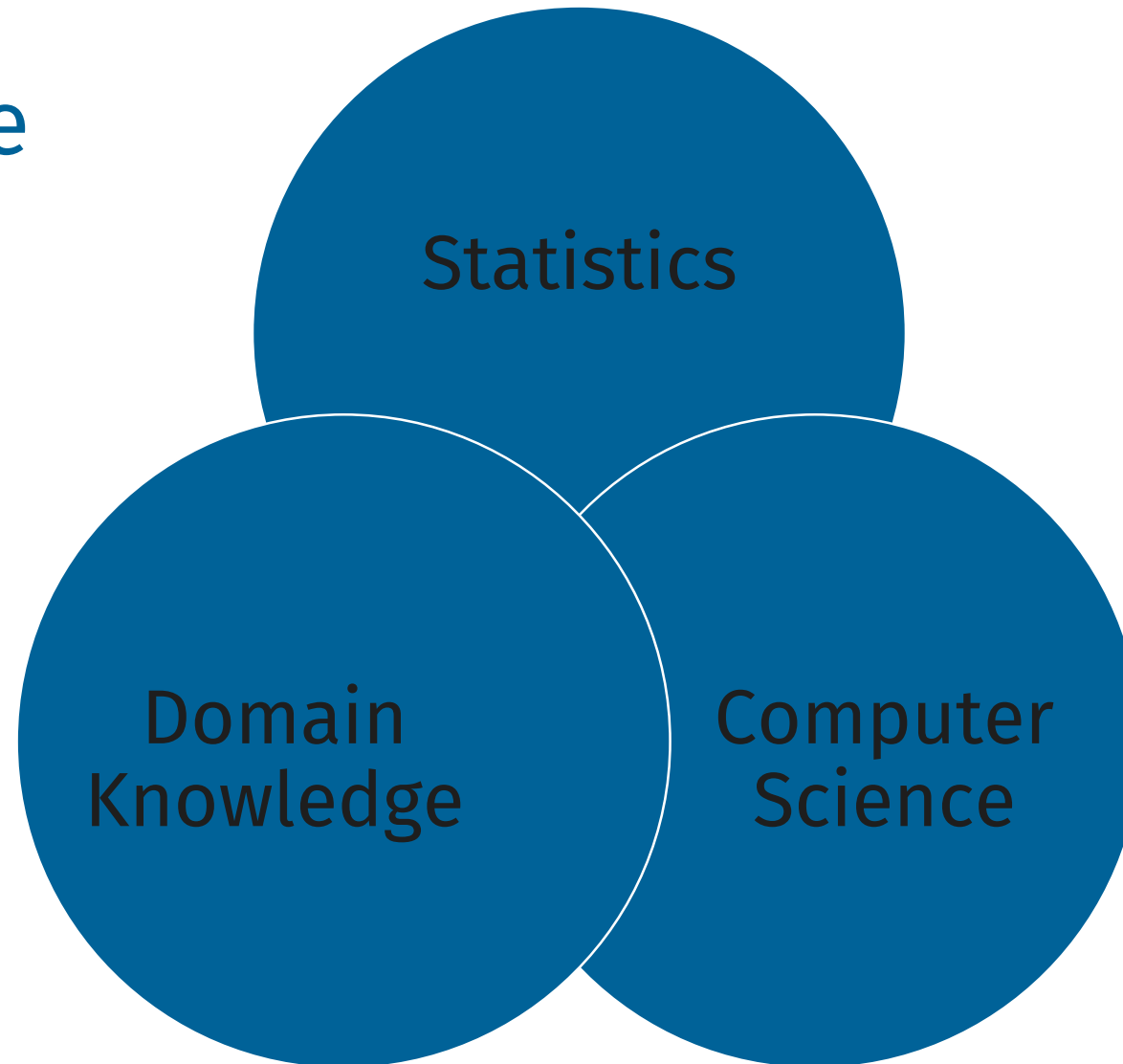
389 sets of statistics
of which **323** surveys
60 calculations
6 registers

Official Statistics

How we work (Generic Statistical Business Process Model)

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

Data Science



Data Science and Official Statistics

e.g.

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

Data Science-Based Innovation

Less effort in the collection of prices

GTIN	GTIN_description	UNIT	QUANTITY		ECOICOP
'8856900000051'	'AROMAX GEWUERZMISCHUNG'	'G'	95	→	'0119203100'
'8714100806234'	'NATUERLICH LECKER FIX SPAGHETTI CARBONARA 47g'	'G'	47	→	'0119103300'
'4102140000882'	'LANDFUERST MALZ 0,5l MW'	'L'	0,5	→	'0213300100'

Data Science-Based Innovation

Observing mobility

Pressemitteilung Nr. 284 vom 7. Juli 2022

- **Deutlich mehr Bahnreisen zwischen 30 und 300 Kilometern im Juni 2022 als im Juni 2019**
- **Moderater Rückgang im Straßenverkehr bei Reisen über 300 Kilometern**
- **Mehr Bahnreisen und weniger Mobilität im Straßenverkehr insbesondere an Wochenenden**

WIESBADEN – Im ersten Monat nach Einführung des bundesweiten 9-Euro-Tickets hat sich das Reiseaufkommen im Schienenverkehr deutlich erhöht. Dies geht aus einer **Sonderauswertung von Mobilfunkdaten** des Statistischen Bundesamtes (Destatis) hervor. Im Juni 2022 lagen die bundesweiten Bewegungen im Schienenverkehr im Schnitt 42 % höher als im Juni 2019. Im Mai 2022 hatten sie noch um 3 % höher als im

https://www.destatis.de/DE/Presse/Pressemitteilungen/2022/07/PD22_284_12.html

Data Science-Based Innovation

Estimation of compliance costs

Geschätzte Fallzahl je Jahr	<input type="text" value="FZ"/>
Art der Vorgabe	bitte auswählen ArtderVorgabe <input type="text"/>
Art der anfallenden Kosten	bitte auswählen Kostenart <input type="text"/>
Themenbereich der Vorgabe	bitte auswählen Themenbereich <input type="text"/>
Art der Pflichterfüllung	bitte auswählen ArtderPflichterfuellung <input type="text"/>
Wirtschaftszweig	bitte auswählen WZ <input type="text"/>
Komplexität der Vorgabe	bitte auswählen Kompl <input type="text"/>
Initiative zur Informationsübermittlung	bitte auswählen Initiative <input type="text"/>
Endadressat der Information	bitte auswählen Endadressat <input type="text"/>

Erfüllungsaufwand (in Tsd. Euro) schätzen	<input type="text" value="JEA_neu"/>
--	--------------------------------------

JavaScript-Editor

JavaScripts erstellen und bearbeiten

```

/* 5. Berechnung der Erfüllungsaufwandsänderung */
/* wenn die Fallzahl null ist dann kann es keinen EA geben */
if (FZ <= 0.1)
  {JEA_neu = 0;
  }
else /* Regressionsmodell */
  { JEA_neu = Math.round
    (Math.pow
     (Math.E,
      (-1.79432 +
       0.62481 * Math.log (FZ+10) +
       2.03461 * ArtderVorgabe +
       0.46688 * Kostenart_2 +
       -0.64758 * Kostenart_3 +
       -0.19543 * ArtderPflichterfuellung_12 +
       -0.22409 * ArtderPflichterfuellung_13 +
       -0.31319 * ArtderPflichterfuellung_14 +
       0.22111 * ArtderPflichterfuellung_21
    )
  )
  }

```

Z. 299, Sp. 1

OK Abbrechen Gehe zu...

<https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2022/03/vereinfachtes-verfahren-erfuellungsaufwand-032022.pdf>

Data Science-Based Innovation ... and Quality

Quality aspects of machine learning

In order to bring together the already built-up empirical knowledge from the practice of official statistics with the findings of science, and thus to promote and intensify the exchange between researchers from the field of machine learning and official statisticians working on applications of modern machine learning methods, Statistics Network Bavaria intends to organize a scientific workshop **from 6th to 8th September 2022 in Munich**. The workshop will be organized for Statistics Network Bavaria by the Bavarian State Statistical Office, the Institute of Statistics of the Ludwig-Maximilians-University Munich and the ifo Institute Munich with the cooperation of the Federal Statistical Office of Germany.



Bayerisches Landesamt für
Statistik



DI STATIS
Statistisches Bundesamt

<https://www.statistiknetzwerk.bayern.de/service/aktuelles/maschinellen-lernens/>

Contact

Statistisches Bundesamt
65180 Wiesbaden
Germany

www.destatis.de

www.destatis.de/kontakt

Contact Person
Florian Dumpert
florian.dumpert@destatis.de
Phone +49 611 75-3887

