Статистически анализ на данни от измерване

С. Лекова, Д. Гочева

Statistical Analysis of Measurement Data

S. Lekova, D. Gocheva

Key Words: *Statistical analysis of data; linear and rank correlation; analysis of variance.*

Abstract. *Good laboratory practice requires laboratories to ensure the quality of analyzes and assess the extent of human exposure to different chemicals and reagents. The subject of research and analysis are data from measurement of samples between two laboratories, by two different methods, in the processing of ore in the Ellatzite-Med AD Mine Complex. Of particular importance is the reliable measurement of the products (concentrations) from the flotation process (content of copper in the ore, intermediate products, concentrate and waste) as well as the production of molybdenum (feed, concentrate and waste). As a result of the sampling and depending on the characteristic to be determined and the test method, the concentration of the identifiable element is measured differently and with different technical means. The statistical analysis of data of parallel measurements from two or three laboratories involves calculating a linear or rank correlation and analyzing how significant the influence of a laboratory factor is, as well as studying and modeling the relationship between the presence of chemical elements in the ore, waste and concentrate in the process of laboratory control.*

Въведение

Рудник „Елаците" е най-големият открит рудник в България и най-големият местен производител на меден имолибденов концентрат. Добивът на руда започва през 1982 г. В момента се изпълнява програма за експлоатация на рудника до 2032 г. В рудничен комплекс „Елаците" работи най-съвременното високопроизводително минно оборудване. Добитата и натрошена руда се транспортира от откритият рудник, разположен в землищата на гр. Етрополе, до Обогатителния комплекс в с. Мирково. Във фирмата работят високоспециализирани минни инженери, геолози, еколози, икономисти, програмисти, специалисти по здравословни и безопасни условия на труд, специалисти по околна среда. Медно-молибденовият концентрат се реализира на вътрешния и световния пазар. Дружеството притежава сертификати за акредитация съгласно стандартите ISO/IEC 17025:2006 – лаборатории за изпитване и калибриране, и ISO/IEC 17020:2012 – дейност на органи, извършващи контрол. „Елаците-мед" има и четири сертификата за производствен контрол: БДС EN 12620+A1:2008 – Добавъчни материали за бетон, БДС EN 13043+AC:2005 – Скални материали за битумни смеси и настилки за пътища, самолетни писти и други транспортни площи, БДС EN 13242+A1:2007 – Скални материали за несвързани и хидравлично свързани смеси за използване в строителни съоръжения и пътно строителство, БДС EN 13450+AC:2005 – Трошен камък за жп линии.

Прилаганата технология в обогатителния комплекс (ОК) "Елаците"е в отделни цехове и може да се разглежда като съставена от три последователни етапа: рудоподготовка; обогатяване на рудата; обезводняване и складиране на продуктите от обогатяването. Пробоподготовката на продуктите и последващото им изпитване по определени характеристики е от голямо значение за всеки един технологичен или производствен процес. Въз основа пробоподготовката и резултатите от изпитването се получава ясна представа за протичането на технологичния процес. Опробването може да бъде автоматично и ръчно. При първия вариант, то се извършва с помощта на автоматични пробовземачни устройства, които отбират проби през определен интервал от време. Ръчното опробване се извършва от лаборант с ръчни пробовземници също през точно определен период от време. И двата метода са равнопоставени и достатъчно представителни. Пробовземането е етап, който е съществен фактор за използваемостта на аналитичната информация. Пробата за анализ трябва да бъде представителна, т.е. нейният състав да бъде идентичен с този на анализирания обект. Така подготвена тази представителна пробата постъпва за анализ в Химическа лаборатория.

Лабораторията за изпитване при „Елаците-мед" АД е създадена като самостоятелно звено в структурата на „Елаците-мед" АД и е на пряко подчинение на генералния директор Производство. Оборудвана с необходимите за дейността й технически и други средства и е акредитирана за анализиране на подземни повърхности и отпадъчни води, медни концентрати, руди и стерили. Нейната функция е главно да предоставя информация и анализ на ръководството за:

- качеството на отпадъчните води от рудодобивната и обогатителната дейност, чрез които да се вземат превантивни действия;
- качеството на преработваните руди, отпадъка от обогатяването и произвежданите концентрати.

Експресната лаборатория извършва анализи посредством апаратура, работеща с радиоизотопи, с цел да установи текущите съдържания на полезните компоненти в руда, отпадък, камерни продукти. Лабораторията е категоризирана и лицензирана за работа в среда на йонизиращи лъчения поради наличните два рентгенови анализатора. Двете ведомствени лаборатории измерват едни и същи

проби по различни методи. Освен това същите проби са измерени в трета, независима и сертифицирана лаборатория. Целта е да се определи доколко съществено е влиянието на лабораториите по отношение на изходните променливи: концентрацията на Cu% в рудата, концентрацията на Cu% в отпадъка, концентрацията на Cu% в концентрата, концентрацията на Mo% в рудата, концентрацията на Mo% в отпадъка, концентрацията на Mo% в концентрата. Анализирана е взаимовръзката между наличието на химични елементи в рудата и в отпадъка в процеса на лабораторен контрол.

Материали и методи

Аналитични техники, методи и процедури. Развитието на аналитичната химия, като наука и практика, е довело до значително многообразие от аналитични техники, аналитични методи и конкретни аналитични процедури. Аналитичните измервателни процедури имат различен обхват на приложимост и различни метрологични и технически характеристики. Главния критерий за избор на аналитична измервателна процедура е пригодност за целта. Другите фундаментални условия при аналитичната дейност са осигуряване на метрологична проследимост на резултатите от измерване и оценяване на неопределеността на получените резултати. Аналитичните техники, методи и процедури могат да бъдат избирани и оценявани по голяма съвкупност от характеристики. Те могат да се разделят на две групи:

- **Основни характеристики (метрологични)** – точност, вярност, прецизност, чувствителност, селективност, изместване, неопределеност от измерването, граница на откриване, граница на определяне, линеен обхват.
- **Второстепенни характеристики** – устойчивост, неподатливост, възможност за многокомпонентен анализ, продължителност на анализа, необходима маса от пробата, квалификация на оператора, вредни и опасни отпадъци за околната среда, цена и други.

Към аналитичните лабораториите се поставят все по-високи изисквания за контрол, които включват:

- осигуряване на проследимост на измерванията;
- калибриране на техническите средства;
- валидиране на методите за изпитване;
- оценка на неопределеността на резултатите;
- вътрешен контрол на качеството;
- използване на сертифицирани референтни материали (CRM) и сравнителни материали (CM);
- участие в национални и международни изпитвания за пригодност и други.

Сравнителните материали и CRM намират широко приложение в аналитичната практика. Те са придружени от документация, издадена от оторизирана организация и осигуряващи стойности на едно или повече указани свойства, свързани с неопределеност и проследимост, с използване на валидиращи процедури. Добрите измервателни практики се съдържат в инструкции за поддържане, калибриране и използване на измервателната апаратура.

В резултат на пробоподготовката и в зависимост от характеристиката, която се определя, и метода на изпитване, концентрацията на определяемия елемент се измерва по различен начин и с различни технически средства. Двете ведомствени лаборатории измерват едни и същи проби по различни методи:

- концентрацията на Cu% в рудата чрез атомно-абсорбционна спектрометрия AAS и рентгено-флуоресцентен анализ RFA;
- концентрацията на Cu% в отпадъка чрез AAS и RFA;
- концентрацията на Cu% в концентрата чрез Йодометрия и RFA;
- концентрацията на Mo% в рудата чрез атомно-емисионна спектрометрия с индуктивно свързана плазма – ICP и RFA;
- концентрацията на Mo% в отпадъка чрез ICP и RFA;
- концентрацията на Mo% в концентрата чрез ICP и RFA.

Освен това същите проби са измерени в трета, независима и сертифицирана лаборатория. Статистическият анализ на резултатите включва изчисляването на линейна или ранг корелация [1] в Matlab на извадки от двете ведомствени лаборатории, както и изчисляването на еднофакторен дисперсионен анализ на данни от три лаборатории.

В статистиката **коефициентът на корелация на Pearson *(PCC)***, наричан също Pearson's r или двумерна корелация, е мярка за линейната корелация между две променливи X и Y. Той има стойност между +1 и -1, където 1 е строга положителна линейна корелация, 0 е липса на линейна корелация и -1 е строга отрицателна линейна корелация. Той се използва широко и е разработен от Карл Пиърсън от идея, въведена от Франсис Галтон през 1880-те. Коефициентът на корелация на Pearson е ковариацията на двете променливи, разделена на произведението от техните стандартни отклонения.

$$(1)\quad r = \frac{\sum_{i=1}^{n}(x_i \quad \bar{x})(y_i \quad \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i \quad \bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i \quad \bar{y})^2}}$$

където:

n е обемът на извадките;

x_i, y_i са отделните проби/опити, индексирани с i;

$\bar{x}$, $\bar{y}$ – средни стойности на двете извадки.

Коефициент на корелация на Кендъл. В статистиката коефициентът на корелация Kendall, обикновено наричан τ коефициент на Kendall, е статистика, използвана за измерване на ординалната връзка между две измерени количества. Tau тестът е непараметричен тест за проверка на хипотезата за статистическа зависимост въз основа на τ коефициента. Той е мярка за ранг корелация: сходството на подреждане на данните при даване на ранг на всяко от количествата. Носи името на Морис Кендъл, който го е разработил през 1938 г. Интуитивно корелацията между две

променливи в Kendall ще бъде висока, когато наблюденията имат подобен (или идентичен за корелация от 1) ранг (т.е. относителна позиция на етикета на наблюденията в променливата: 1-ва, 2-ра, 3-та и т.н.) и ниска, когато наблюденията имат различни (или напълно различни за корелация от -1) рангове между двете променливи. Коефициентът τ на Kendall се дефинира като

$$(2)\ \tau = \frac{(\text{брой съвпадащи двойки}) - (\text{брой несъответстващи двойки})}{n(n-1)/2}$$

Знаменателят е общият брой комбинации от двойки, така че коефициентът трябва да бъде в диапазона $-1 \le \tau \le 1$. Ако сходството между двете класификации е перфектно (т.е. двете класирания са еднакви), коефициентът има стойност 1. Ако няма сходство между двете класификации (т.е. една класификация е обратна на другата), коефициентът има стойност -1. Ако X и Y са независими, очакваният коефициент ще бъде приблизително нулев.

В статистиката **коефициентът на корелация на Spearman**, наречен на името на Чарлс Спирман и често обозначен с гръцката буква ρ или като r_s, е непараметрична мярка за ранг корелация: оценява колко добре може да се опише връзката между две променливи, използвайки монотонна функция. При коефициента на корелация на Spearman двете променливи, които се сравняват, са монотонно свързани, дори ако връзката им не е линейна. Корелацията Spearman между две променливи е равна на Pearson корелацията между ранговите стойности на тези две линейни отношения, корелацията на Spearman оценява монотонните връзки (линейни или не). Ако няма повтарящи се стойности на данни, перфектната Spearman корелация от +1 или -1 възниква, когато всяка от променливите е перфектна монотонна функция на другата. Коефициентът на корелация Spearman се дефинира като коефициент на корелация на Pearson между рангови променливи. Ако двете редици не съдържат повтарящи се стойности, то може r_s да се изчисли с помощта на популярната формула

$$(3)\quad r_s = 1 \quad \frac{6\sum d_i^2}{n(n^2-1)}$$

където

$$d_i = rg(X_i) \quad rg(Y_i)$$

е разликата между ранговете за всяко наблюдение.

Резултати

Изходните данни представляват матрица с 12 стълба (m=12), всеки от които съдържа извадка с 10 измервания (n=10) на съответните изходни величини. В *табл. 1* са показани последователността на измерваните изходни концентрации и съответният метод на измерване от двете заводски лаборатории

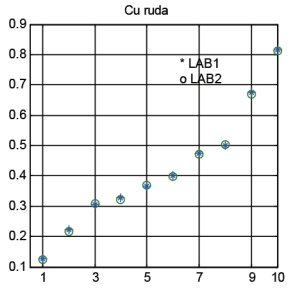
Табл. 1. (ЙМ – йодометрия)

AAS1	RFA2	AAS3	RFA4	ЙМ5	RFA6
Cu% руда	Cu% руда	Cu% отпадък	Cu% отпадък	Cu% к-т	Cu% к-т
ICP7	RFA8	ICP9	RFA10	ICP11	RFA12
Mo% руда	Mo% руда	Mo% отпадък	Mo% отпадък	Mo% к-т	Mo% к-т

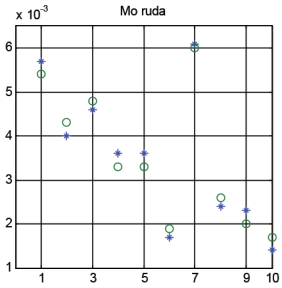
Линейна корелация на Пиърсън. RHO е матрица с размерност (mxm), съдържаща линейния корелационен коефициент между всяка двойка колони в изходната (nxm)-матрица. PVAL е матрица от p-стойностите (нива на значимост) за тестване на хипотезата за отсъствие на корелация спрямо алтернативата, че има ненулева корелация. Всеки елемент на PVAL е p-стойността за съответния елемент на RHO. Двете матрици са дадени на *фиг. 5* в края на статията поради големия им размер. Ако PVAL (i, j) е малък, да речем по-малко от 0.05, тогава корелацията RHO (i, j) е значително различна от нула. Матриците, съдържащи долни и горни граници за 95% доверителен интервал за всеки коефициент, са със същия размер като RHO, но не са приведени тук. Корелационните матрици на Kendall и Spearman също не показват други значими корелации.

От тази матрица се вижда, че коефициентите на корелация на измерванията по двата метода (в двете лаборатории) при всички разглеждани случаи са по-големи от 0.91 и са значими – p-value клони към нула (маркирано с жълто). По-нататък по подразбиране ще разглеждаме и двата метода. В матриците RHO и PVAL значимите коефициенти на корелация (p-value по-малко от 0.05) са маркирани с виолетово: коефициенти на корелация между Cu-руда/Mo-руда и коефициенти на корелация между Cu-конц./Mo-конц и са обобщени в *табл. 2* (също в края на статията). На *фиг.1* и *2* са показани данните за Cu% руда и Mo% руда, за които са получени тези резултати, а на *фиг. 3* и *4* са показани данните за Cu% концентрат и Mo% концентрат. На абцисната ос е номерът на пробата.

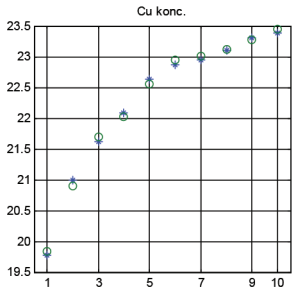
При ниво на значимост α=0.1 (маркирани със синьо) са значими коефициентите на корелация между Cu-отпадък/Cu-конц. и между Mo-отпадък/Cu-конц. Коефициентът на корелация между Cu-отпадък/Mo-отпадък е значим само по първия метод, но това по-скоро трябва да се като резултат от малкия обем на извадката.



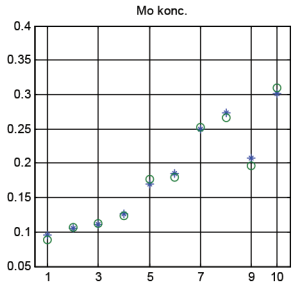
Фиг. 1. Данни за Cu% руда



Фиг. 2. Данни за Mo% руда



Фиг. 3. Данни за Cu% конц.



Фиг. 4. Данни за Mo% конц.

Обработка на данните от три лаборатории с еднофакторен дисперсионен анализ. Чрез метода на дисперсионния анализ може да се даде еднозначен отговор на въпроса дали влиянието на един или повече фактори е съществено или не по отношение на изходната променлива [1,2]. Нека означим с X фактора, чието влияние се изследва, а именно фактора Лаборатория, който при експеримента има три нива (две ведомствени лаборатории и една независима). Той може да се променя целенасочено на различни нива (m=3). За всяко от тези нива са проведени определен брой (n=50) измервания или наблюдения върху зависимата променлива У, по която се съди за поведението на изследвания обект, процес или явление. В еднофакторния дисперсионен анализ (ЕДА) се сравняват средните на няколко групи, за да се тества хипотезата, че те са равни, срещу общата алтернатива, че те са различни. Понякога тази алтернатива може да е твърде обща. Може да е необходима информация за това кои двойки средни са значително различни и кои не. Тест, който може да предостави такава информация, се нарича процедура за множествено сравнение.

Когато се извършва обикновен t-тест за средната стойност на една група спрямо друга, се задава ниво на значимост α, което определя критичната стойност на t статистиката. Обикновено в инженерни задачи се приема α = 0.05. Когато има много групови средни, има и много двойки за сравнение. Ако се приложили обикновен t-тест в тази ситуация, α ще се прилага за всяко сравнение, така че вероятността от неправилно намиране на значителна разлика би се увеличила с броя на сравненията. Често е за предпочитане да се направи тест, за да се определи кои двойки средни са значително различни и кои не са. В Matlab, multicompare прави множествен тест за сравнение, като използва информацията в структурата на статистическите данни и връща матрица на резултатите от сравнението по двойки и показва интерактивна графика на прогнозите с интервалите за сравнение около тях.

RHO =

1.0000	0.9998	0.3871	0.2905	-0.0347	-0.0270	-0.6507	-0.6494	0.1077	0.0097	-0.1360	-0.1382
0.9998	1.0000	0.3836	0.2857	-0.0359	-0.0284	-0.6489	-0.6475	0.1048	0.0075	-0.1409	-0.1432
0.3871	0.3836	1.0000	0.9790	0.5533	0.5670	-0.1669	-0.1912	0.4787	0.2267	0.3832	0.3513
0.2905	0.2857	0.9790	1.0000	0.5861	0.6027	-0.1158	-0.1490	0.5533	0.3108	0.4014	0.3740
-0.0347	-0.0359	0.5533	0.5861	1.0000	0.9982	0.4894	0.4495	0.5693	0.2184	0.8309	0.8270
-0.0270	-0.0284	0.5670	0.6027	0.9982	1.0000	0.4665	0.4218	0.5793	0.2289	0.8440	0.8395
-0.6507	-0.6489	-0.1669	-0.1158	0.4894	0.4665	1.0000	0.9862	0.3329	0.2475	0.3602	0.3906
-0.6494	-0.6475	-0.1912	-0.1490	0.4495	0.4218	0.9862	1.0000	0.2157	0.1298	0.2769	0.3025
0.1077	0.1048	0.4787	0.5533	0.5693	0.5793	0.3329	0.2157	1.0000	0.9066	0.4667	0.5110
0.0097	0.0075	0.2267	0.3108	0.2184	0.2289	0.2475	0.1298	0.9066	1.0000	0.1980	0.2595
-0.1360	-0.1409	0.3832	0.4014	0.8309	0.8440	0.3602	0.2769	0.4667	0.1980	1.0000	0.9962
-0.1382	-0.1432	0.3513	0.3740	0.8270	0.8395	0.3906	0.3025	0.5110	0.2595	0.9962	1.0000

PVAL =

0	0.0000	0.2692	0.4156	0.9242	0.9411	0.0416	0.0421	0.7672	0.9789	0.7079	0.7033
0.0000	0.0000	0.2738	0.4236	0.9215	0.9379	0.0424	0.0430	0.7733	0.9835	0.6978	0.6931
0.2692	0.2738	0	0.0000	0.0971	0.0874	0.6448	0.5966	0.1616	0.5287	0.2743	0.3195
0.4156	0.4236	0.0000	0	0.0750	0.0652	0.7501	0.6813	0.0971	0.3820	0.2503	0.2870
0.9242	0.9215	0.0971	0.0750	0.0000	0.0000	0.1511	0.1924	0.0859	0.5445	0.0029	0.0032
0.9411	0.9379	0.0874	0.0652	0.0000	0	0.1741	0.2247	0.0792	0.5247	0.0021	0.0024
0.0416	0.0424	0.6448	0.7501	0.1511	0.1741	0.0000	0.0000	0.3472	0.4906	0.3066	0.2644
0.0421	0.0430	0.5966	0.6813	0.1924	0.2247	0.0000	0	0.5495	0.7207	0.4387	0.3956
0.7672	0.7733	0.1616	0.0971	0.0859	0.0792	0.3472	0.5495	0.0000	0.0003	0.1739	0.1312
0.9789	0.9835	0.5287	0.3820	0.5445	0.5247	0.4906	0.7207	0.0003	0	0.5834	0.4691
0.7079	0.6978	0.2743	0.2503	0.0029	0.0021	0.3066	0.4387	0.1739	0.5834	0	0.0000
0.7033	0.6931	0.3195	0.2870	0.0032	0.0024	0.2644	0.3956	0.1312	0.4691	0.0000	0.0000

Фиг. 5. Корелационна матрица на Пиърсън и матрица от p-стойностите

Табл. 2

Коефициенти на корелация	Коефициенти на корелация между Cu-руда/Mo-руда и p-value в скоби			
	Cu-руда Iлаб./ Mo-руда I лаб.	Cu-руда I лаб./ Mo-руда II лаб.	Cu-руда II лаб./ Mo-руда I лаб.	Cu-руда II лаб./ Mo-руда II лаб.
Pearson	-0. 6507 (0. 0416)	-0. 6494 (0. 0421)	-0. 6489 (0. 0424)	-0. 6475 (0. 0430)
Kendall	-0. 5843 (0. 0248)	-0. 5843 (0. 0248)	-0. 5843 (0. 0248)	-0. 5843 (0. 0248)
Spearman	-0. 6687 (0. 0345)	-0. 6687 (0. 0345)	-0. 6687 (0. 0345)	-0. 6687 (0. 0345)
	Коефициенти на корелация между Cu- конц./Mo- конц.			
	Cu-конц. Iлаб./ Mo-конц. I лаб.	Cu-конц. I лаб./ Mo-конц. II лаб.	Cu-конц. II лаб./ Mo-конц. I лаб.	Cu-конц. II лаб./ Mo-конц. II лаб.
Pearson	0. 8309 (0. 0029)	0. 8270 (0. 0032)	0. 8440 (0. 0021)	0. 8395 (0. 0024)
Kendall	0. 9111 (0. 0000)	0. 9111 (0. 0000)	0. 9111 (0. 0000)	0. 9111 (0. 0000)
Spearman	0. 9636 (0. 0000)	0. 9636 (0. 0000)	0. 9636 (0. 0000)	0. 9636 (0. 0000)

Изходът съдържа резултатите от теста под формата на матрица с пет колони. Всеки ред от матрицата представлява един тест и има един ред за всяка двойка групи. Записите в реда са в следния ред: средните, които се сравняват, прогнозната разлика в средните и доверителния интервал за разликата. Ако доверителният интервал съдържа 0.0, разликата няма да бъде значителна при зададено α и обратно.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	0.0019	2	0.00095	0.1	0.9074
Error	1.43335	147	0.00975		
Total	1.43525	149			

Фиг. 6. Резултати от ЕДА на Cu% в руда

Резултатите от еднофакторен дисперсионен анализ при зависимата променлива концентрация на Cu в руда са показани на *фиг. 6*. За останалите пет целеви променливи (концентации) резултатите са аналогични – факторът Лаборатория не влияе върху резултатите от измерването.

Заключение

От направения анализ се вижда че влиянието на лабораториите е несъществено по отношение на изходната променлива (концентрацията на Cu% в рудата, концентрацията на Cu% в отпадъка, концентрацията на Cu% в концентрата и за концентрацията на Mo% в рудата, концентрацията на Mo% в отпадъка, концентрацията на Mo% в концентрата).

Благодарности

Резултатите в статията са получени в рамките на Национална научна програма „Информационни и комуникационни технологии за единенцифров пазар в науката, образованието и сигурността (ИКТвНОС)“, финансирана от МОН.

Литература

1. Akoglu, H. User's Guide to Correlation Coefficients. – *Turkish Journal of Emergency Medicine,* 18, 2018, 91– 93.

2. Вучков, И., С. Стоянов. Математическо моделиране и оптимизация на технологични обекти. София, Техника, 1980.

3. Вучков, И., С. Стоянов, В. Цочев, Н. Козарев. Ръководство за лабораторни упражнения по статистически методи. София, Нови Знания, 2002.

За контакти:
Гл. ас. д-р **Светла Лекова**
e-mail:sv_lekova@uctm. edu,
Доц. д-р **Даниела Гочева**
e-mail:dani@uctm. edu
Катедра „Автоматизация на производството“
Химикотехнологичен и металургичен
университет – София