Introdução à Análise de Dados em Arqueologia

Relatório de Unidade Curricular

João Cascalheira

ICArEHB, Universidade do Algarve

Relatório apresentado no âmbito das provas de habilitação para o título de Agregado no ramo de conhecimento de Arqueologia pela Universidade do Algarve, de acordo com o Art.º 8.º do Decreto-Lei n.º 239/2007, de 19 de junho, e com o Art.º 4.º do Despacho n.º 2251/2020, de 17 de fevereiro.

Conteúdos

Nota Introdutória

Em termos gerais, a Arqueologia pré-década de 1960 estava maioritariamente baseada numa descrição empírica da cultura material, o que incluía acreditar que uma grande quantidade de dados “falariam” sempre por si próprios. Os padrões emergiriam do estudo descritivo de coleções de dados, permitindo que cerâmicas, ferramentas em pedra ou estruturas arqueológicas fizessem sentido quando agrupadas segundo determinadas características, às quais depois eram atribuídos limites espaciais e temporais por meio do conceito normativo Childeano de “Cultura Arqueológica” (Johnson, 2019). Devido a este contexto teórico, os dados eram tidos como adquiridos pela maior parte dos arqueólogos e o processo de observação, registo e interpretação necessitava de pouca, se não mesmo nenhuma, justificação (Lock, 2003).

As mudanças com a Nova Arqueologia, iniciadas na década de 1960 e intensificadas na década seguinte, marcaram a introdução do método científico e a rejeição da percepção subjetiva do empirismo arqueológico (Johnson, 2019; Trigger, 1989). No centro desta nova corrente estava a crença na objetividade por meio da observação sistemática, medição e registo de dados através da adoção dos denominados métodos quantitativos. Acreditava-se que a objetividade seria alcançada ao separar a teoria da prática, permitindo que os dados fossem medidos de forma independente por qualquer sujeito observador (Lock, 2003). Assim, enquanto a ligação anterior entre dados e teoria era indutiva, *i.e.*, uma recolha imparcial de "todos" os dados iria produzir teoria, o novo paradigma da arqueologia processual tinha como ponto central um raciocínio hipotético-dedutivo em que o conhecimento é acumulado através do teste de hipóteses explícitas (muitas vezes através do uso de testes estatísticos de significância) sobre dados recolhidos na base do que seria relevante para a análise (Trigger, 1989). A grande implicação do método científico para a Arqueologia tornou-se, assim, na possibilidade de criar uma Arqueologia global, unida por métodos padronizados de análise (Clarke, 1968), aplicáveis a qualquer conjunto de dados para estabelecer generalizações interculturais ou até mesmo "leis" do conhecimento arqueológico.

Se avançarmos para os dias de hoje, estas mudanças serviram como pedra basilar para algumas das mais recentes revoluções na forma como se pensa e se pratica Arqueologia. Fenómenos como a chamada "crise da replicação" nas ciências sociais (Baker, 2016; Camerer et al., 2018; Serra-Garcia & Gneezy, 2021), e consequentes alterações na forma de recolher, gerir e partilhar dados em Arqueologia estão a transformar a disciplina em todas as suas vertentes (Karoune & Plomp, 2022; Marwick, 2016).

Como em outras disciplinas, quando vários estudos arqueológicos independentes apresentam resultados semelhantes, consideramos que esses resultados são uma aproximação razoável ao comportamento humano do passado. Esta capacidade de reproduzir os resultados de outros estudos é um princípio fundamental do método científico e, quando as reproduções são bem sucedidas, a disciplina avança. Em Arqueologia há uma longa tradição em realizar testes empíricos de reprodutibilidade quando, por exemplo, se regressa a sítios escavados por gerações anteriores de arqueólogos ou quando se reanalisam coleções de museus utilizando novos métodos. Contudo, pouco progresso foi feito no que diz respeito a testar a replicabilidade dos dados recolhidos e das suas análises quantitativas. Este problema advém de duas razões principais. Primeiro, a maior parte das publicações raramente disponibiliza informações suficientes que permitam a outro arqueólogo reproduzir os resultados (Marwick, 2016). Segundo, a formação base dos arqueólogos, particularmente em Portugal, raramente inclui o desenvolvimento de competências de organização, gestão e análise de dados. A ausência de uma componente de literacia de dados (do inglês *data literacy*, definida como a capacidade de ler, trabalhar, analisar e comunicar com dados) na formação em Arqueologia, compromete a qualidade dos dados arqueológicos, limita a sua reutilização e restringe a capacidade que os arqueólogos muitas vezes têm de se expressarem em contextos interdisciplinares. Como resultado, a contribuição que poderíamos ter em conversas contemporâneas fora do âmbito da Arqueologia também se torna bastante limitada, apesar de, teoricamente, estarmos bem posicionados para o fazer (Kintigh et al., 2014).

Neste contexto, a literacia de dados não consiste primariamente em capacitar os indivíduos a dominar uma habilidade específica ou a tornarem-se proficientes numa determinada plataforma de tecnologia. Em vez disso, trata-se de equipar os arqueólogos com as ferramentas necessárias para compreender os princípios subjacentes e os desafios associados aos dados (Bhargava et al., 2015; Kansa & Kansa, 2021).

Atualmente são abundantes as discussões e iniciativas acerca das promessas e dos perigos de alavancar dados em diversos tamanhos e formatos para enfrentar os desafios do mundo como parte da *Data Revolution*, invocada em 2015 pelas Nações Unidas na sua Agenda de Desenvolvimento Global. Um artigo, amplamente citado, publicado pelo jornal *The Economist* intitulado "The Data Deluge: Businesses, Governments and Society Are Only Starting to Tap Its Vast Potential" (Cukier, 2010) teve como um dos primeiros comentários online o seguinte: "Aqui estão os empregos do século XXI (...). Por favor entendam e eduquem a próxima geração em conformidade". A importância da literacia de dados (assim como da literacia digital) para estudantes que terminam qualquer curso superior nos dias que correm, não deve ser desvalorizada. Em Arqueologia, em particular, é urgente a mudança para perfis profissionais em

que a especialização e a profundidade de conhecimentos se conjugam com uma maior amplitude de conhecimento no uso, tratamento e análise quantitativa de dados digitais (ver Figura 1). A crescente importância dos métodos digitais de registo, gestão e comunicação dos achados arqueológicos, obrigam, necessariamente, que todos os profissionais tenham um nível básico de literacia de dados, de preferência por meio de programas educativos especialmente direcionados para as especificidades dos dados arqueológicos.

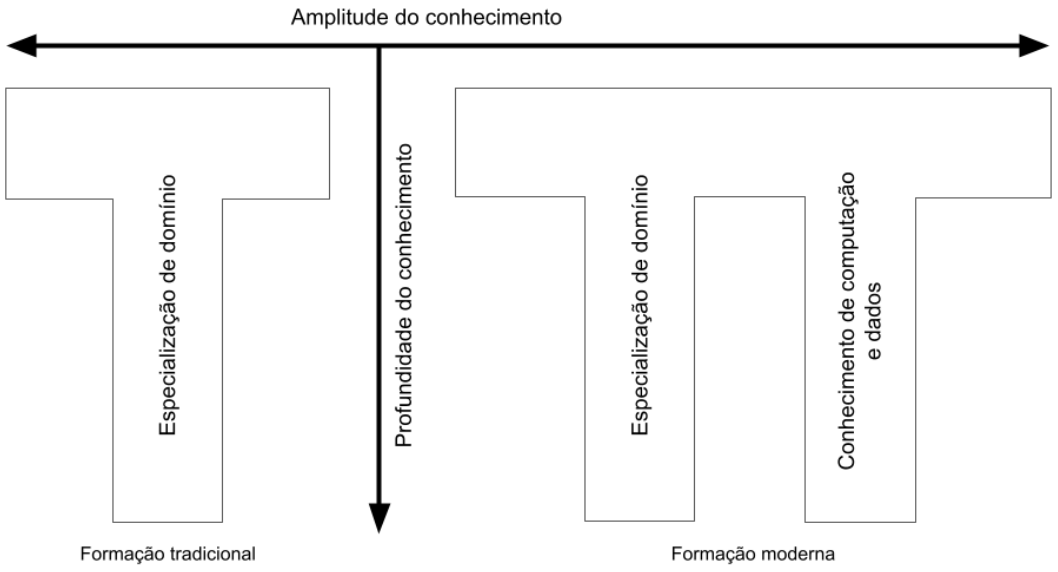


*Figura 1. Perfis profissionais em forma de "**T**" vs forma de "**Π**". Adaptado de Marwick (2016).*

Em 1998, Aldenderfer identificava esta lacuna nos cursos de 1º ciclo na área da Antropologia nos Estados Unidos da América, dizendo:

"Undergraduates, except for the most motivated, are unlikely to ever see a quantitative class, and therefore, many are poorly prepared for graduate study. Those who choose to go into cultural resources management also will be poorly trained, a perennial complaint from the managers who hire them." (Aldenderfer, 1998, p. 108)

Do lado de cá do Atlântico, e no caso de Portugal em particular, a situação não é, infelizmente, muito diferente. As páginas que se seguem, bem como a unidade curricular que representam, são um modesto contributo para a mudança deste paradigma. Serão, a seu tempo, disponibilizadas online na íntegra (em conjunto com os ficheiros necessários para a realização dos exercícios), para que sirvam de suporte a todos os estudantes que procurem material introdutório sobre análise de dados em Arqueologia.

IADA no quadro da Licenciatura em Património Cultural e Arqueologia

A Introdução à Análise de Dados em Arqueologia (abreviadamente, IADA) é uma das unidades curriculares optativas do curso de licenciatura em Património Cultural e Arqueologia da Universidade do Algarve, destinando-se, na estrutura do curso, aos alunos que frequentem o 2º e 3º ano[1].

Esta unidade curricular funcionou pela primeira vez no ano letivo de 2016/2017, repetindo-se em 2017/2018 e 2020/2021. A sua existência resulta de uma necessidade muito concreta: a de preencher uma lacuna na estrutura do curso no que diz respeito ao ensino das técnicas e metodologias computacionais de tratamento, análise e apresentação de dados arqueológicos. Tradicionalmente, no âmbito da licenciatura em Património Cultural e Arqueologia, estes aspectos fundamentais para qualquer profissional de arqueologia, eram lecionados ou no âmbito de uma disciplina de Metodologia do Trabalho Científico (em conjunto com aspectos mais relacionados com a escrita, bibliografia, etc.), das disciplinas de técnicas de análise laboratorial (*e.g.*, cerâmicas, líticos, faunas), ou no contexto da realização de trabalhos de seminário de conclusão de curso. Em qualquer dos casos, os alunos para além de terem de desenvolver conhecimentos específicos sobre cada uma das disciplinas, bem como desenvolver o trabalho analítico dos materiais em questão, têm ainda de aprender a construir e manter as bases de dados resultantes da análise de conjuntos arqueológicos, escolher as melhores abordagens para as analisar, e finalmente criar e interpretar elementos gráficos com base nos dados recolhidos. Com base na experiência anterior do candidato na leccionação da disciplina de Análise de Materiais Líticos, bem como na orientação de alunos de licenciatura e do mestrado em Arqueologia da UAlg, frequentemente ficou patente a dificuldade que os alunos demonstram neste âmbito, fruto das lacunas relacionadas com a literacia dos dados, nomeadamente a falta de um conhecimento base sobre os diferentes métodos e procedimentos de recolha, análise e apresentação de dados arqueológicos.

No panorama nacional, IADA não se trata de uma unidade curricular ímpar, sendo que a maior parte dos planos curriculares de licenciaturas em Arqueologia têm atualmente pelo menos uma disciplina dedicada às técnicas informáticas e métodos digitais aplicados à Arqueologia (Tabela

[1] A partir de 2020/2021, com a alteração na estrutura do curso de Património Cultural e Arqueologia, os alunos só escolhem o ramo de especialização no final do 2º ano e a unidade curricular de IADA é apenas destinada a alunos do 3º ano da licenciatura.

1). Ainda assim, torna-se evidente, com base na informação disponibilizada nos respectivos portais, que os conteúdos programáticos de algumas dessas disciplinas têm como foco dominante o uso de Sistemas de Informação Geográfica (SIG) ou o Desenho Assistido por Computador. Estes tópicos não fazem parte do programa de IADA, sendo tratados em unidades curriculares optativas independentes, como por exemplo a de Sistemas de Informação Geográfica Aplicados à Arqueologia. Por outro lado, é certo que os alunos de grande parte destas licenciaturas, incluindo os da UAlg, têm a opção de frequentar outras unidades curriculares optativas de cursos distintos, dedicadas, por exemplo, à ciência de dados ou à estatística introdutória. Nem sempre esta é a melhor opção, tendo em conta, principalmente, as particularidades (inclusivamente éticas) dos dados arqueológicos e as respectivas limitações e oportunidades na aplicação de determinadas abordagens.

Instituição - Curso	Unidade Curricular	Links
Universidade Nova de Lisboa - Licenciatura em Arqueologia	Métodos Digitais em Arqueologia	https://guia.unl.pt/pt/2022/fcsh/program/9006/course/01105677
Universidade de Coimbra - Licenciatura em Arqueologia	Informática Aplicada à Arqueologia	https://apps.uc.pt/courses/PT/unit/80479/22541/2023-2024?type=ram&id=5382
Universidade de Évora - Licenciatura em História e Arqueologia	Introdução às Humanidades Digitais	https://www.uevora.pt/estudar/cursos/licenciaturas?cod=8251&v=plano-estudos&uc=HIS12024L
Universidade do Minho - Licenciatura em Arqueologia	Tratamento Digital da Informação	https://www.ics.uminho.pt/pt/Ensino/Licenciaturas/Arqueologia

Tabela 1. Exemplos de unidades curriculares disponibilizadas noutras instituições com componentes relacionadas com a análise de dados em Arqueologia.

O número de alunos inscritos em IADA, nas três edições do curso em que esta foi leccionada, foi de quatro alunos em 2016/2017, e sete alunos em 2017/2018 e em 2020/2021. Tratando-se de uma unidade curricular optativa destinada apenas aos alunos que optem pelo ramo de Arqueologia no 3º ano do curso (em anos anteriores no 2º ano do curso), a consolidação numérica das duas últimas edições reflete uma relativamente boa aceitação por parte dos alunos. Por outro lado, tendo em conta a grande componente prática da unidade curricular, estes números permitiram cumprir os pressupostos do processo de Bolonha, promovendo um processo de ensino/aprendizagem mais centrado no aluno.

Do ponto de vista da avaliação, a média geral de todos alunos de todas as edições de IADA situa-se em 14.8 valores. Deve salientar-se ainda, a este respeito, que as avaliações individuais registadas acima destas médias se distribuem entre 15 e 19 valores, e abaixo entre os 14 e os 4 valores. Neste último caso, as duas reprovações registadas estão claramente relacionadas com a fraca assiduidade dos alunos, o que revela a importância da componente lectiva presencial no âmbito desta unidade curricular.

Apesar do sucesso relativamente elevado dos alunos de IADA, são de salientar dois casos específicos, por terem não só cumprido os objetivos da unidade curricular, mas também porque demonstraram, posteriormente, a importância dos conteúdos aprendidos para o sucesso do seu percurso enquanto alunos e investigadores. Estes dois casos são os de David Nora e Joana Belmiro, cujos seminários finais de licenciatura, bem como as teses de mestrado defendidas na UAlg, incluem uma ampla utilização das ferramentas e conceitos de recolha, gestão, análise e partilha de dados introduzidas e exploradas nas versões anteriores de IADA. No caso da Joana Belmiro, a publicação com revisão por pares que teve origem na sua tese de mestrado (Belmiro et al., 2021), utilizou os princípios discutidos em IADA sobre a importância da transparência e da partilha dos dados em Arqueologia.

Estrutura da unidade curricular

Nesta secção esclarecem-se algumas das opções tomadas nos campos mais relevantes para a estruturação da unidade curricular, incluindo algumas das alterações efetuadas com base na experiência tida nos anos em que esta foi lecionada.

Objetivos da aprendizagem

Esta unidade curricular tem dois objetivos principais. Por um lado, visa dotar os alunos de conhecimentos que lhes permitam identificar, classificar e compreender os diferentes tipos de dados arqueológicos. Neste sentido, os alunos devem ser capazes de: reconhecer as limitações e o potencial dos diferentes tipos e fontes de dados; compreender como integrar múltiplas formas de evidência numa interpretação coesa do registo arqueológico; e identificar e resolver problemas comuns na gestão de dados arqueológicos.

Por outro lado, procura-se formar competências no âmbito da utilização das várias ferramentas computacionais utilizadas na recolha, processamento e análise quantitativa de dados arqueológicos. Neste âmbito, os alunos devem ser capazes de: selecionar o método de análise mais adequado para um determinado conjunto de dados; realizar análises exploratórias de dados; e representar e interpretar análises simples de dados arqueológicos.

Metodologias de ensino e avaliação

A avaliação é distribuída pelos três seguintes elementos:

- Elaboração de exercícios práticos: 40%;
- Trabalho final: 50%;
- Participação e assiduidade: 10%.

Os exercícios práticos são individuais e realizam-se durante três aulas de cariz puramente prático, nas quais os alunos são acompanhados pelo docente na resolução dos enunciados. Essencialmente, estes exercícios correspondem à aplicação prática de cada um dos principais blocos práticos da unidade curricular (registo de dados, transformação de dados, e análise descritiva). Na maioria dos casos, estes exercícios iniciam-se e terminam-se em contexto de aula. Contudo, em alguns casos, quando, por vários motivos, a conclusão dos mesmos não é alcançada em aula, é pedido aos alunos que terminem até à próxima aula, onde serão reservados trinta

minutos para revisão das dúvidas e dos resultados finais. A avaliação dos exercícios foca-se sobretudo nas soluções encontradas pelos alunos para resolver e apresentar cada um dos pontos do enunciado. No caso do exercício de criação de um ficheiro E5 para recolha de dados, são tidos em consideração, por exemplo, as escolhas dos nomes das variáveis, ou as condições criadas para preenchimento de variáveis com base em valores preenchidos em variáveis anteriores. No caso do exercício de transformação de dados em folhas de cálculo, são avaliadas as soluções encontradas pelos alunos no uso de funções para criação de novas variáveis ou operações de consulta e limpeza dos dados.

O trabalho final é também iniciado em sala de aula na última aula do semestre. Ao contrário dos exercícios práticos, neste caso os alunos não contam com o suporte do docente, a não ser em questões relacionadas com a interpretação do enunciado.

Conteúdos programáticos

A unidade curricular de IADA tem um total de 39 horas de teórico-práticas e 5 horas de orientação tutorial; portanto, 140 horas totais de trabalho, equivalentes a 5 ECTS. Tendo em conta que os horários semanais-tipo estabelecem 3 horas de aulas por semana, as aulas distribuem-se ao longo de 13 semanas do respectivo semestre. As 13 aulas teórico-práticas estão organizadas de uma forma lógica com base em quatro blocos temáticos e respectivos subtemas:

- **Bloco 1**. **O registo de dados em Arqueologia (Aula 01)** - Neste bloco, apresentam-se, em primeiro lugar, as particularidades do registo em Arqueologia e a sua história, com particular relevância para a introdução dos métodos de registo totalmente digitais e as suas respectivas vantagens. Em segundo lugar, introduz-se a importância dos métodos quantitativos em Arqueologia, salientando-se as diferenças entre Matemática e Estatística, como o raciocínio quantitativo é fundamental para a Arqueologia e uma melhor compreensão das suas implicações é suscetível de melhorar o nosso trabalho como arqueólogos.
- **Bloco 2**. **Recolha e transformação de dados (Aulas 02 a 06)** - Este bloco está dividido em três partes, relativas a: 1) fundamentos da construção de uma base de dados, incluindo o reconhecimento dos diferentes tipos de dados e dos vários tipos de bases de dados mais utilizados em Arqueologia; 2) diferentes abordagens ao registo de dados arqueológicos, focando-se, em maior detalhe, na utilização do programa E5; 3) por fim, gestão e transformação de dados em folhas de cálculo, nomeadamente através da utilização dos modificadores matemáticos e de texto, bem como das funções para criação

de novas variáveis ou correção das existentes. Para este último ponto, é partilhado com os alunos um ficheiro com dados arqueológicos reais (com informação detalhada sobre uma coleção de ferramentas líticas solutrenses do sítio arqueológico de Vale Boi - Vila do Bispo), que serve de base para a reprodução de todos os exemplos explorados em aula.

- **Bloco 3**. **Métodos quantitativos e estatística descritiva (Aulas 07 a 11)** - Este bloco, que ocupa a maior parte das aulas no programa, não só engloba a maior quantidade de conteúdo, mas também requer um contexto teórico mais extenso para uma compreensão completa da aplicação prática. O bloco é dividido em dois grandes sub-blocos: análise univariada e análise bivariada, que são organizados de acordo com os dois principais tipos de variáveis, categóricas e numéricas. Em ambos os sub-blocos, são explicados conceitos estatísticos e matemáticos que permitem a construção e interpretação dos dois elementos fundamentais para a representação dos dados arqueológicos, as tabelas e os gráficos. Apesar de os estudantes poderem seguir e reproduzir os muitos exemplos fornecidos ao longo das aulas com recurso à base de dados já mencionada, a penúltima aula deste bloco (Aula 10) é dedicada a explorar um conjunto de casos de estudo (publicados sob a forma de artigos científicos) em que são analisadas não apenas as opções tomadas pelos autores para representar os dados, mas também as interpretações oferecidas por cada representação. O principal objetivo desta aula é mostrar exemplos concretos da utilização dos conceitos e ferramentas explorados nas aulas anteriores, para que haja uma melhor compreensão da utilidade e flexibilidade dos mesmos.
- **Bloco 4**. **Amostras, populações e a interpretação dos dados em Arqueologia (Aula 12)** - Este bloco, condensado em apenas uma aula, procura funcionar como uma introdução ao processo de amostragem e a sua importância para compreender e integrar os resultados da análise de um determinado conjunto de dados arqueológicos. São focadas as diferenças entre populações e amostras estatísticas e a sua relevância para a avaliação e reconhecimento de padrões em dados arqueológicos.

Bibliografia

Tratando-se de uma unidade curricular com uma componente prática muito significativa, a lista de referências bibliográficas fornecida aos alunos contempla sobretudo os manuais atualmente disponíveis sobre o uso de métodos quantitativos em Arqueologia. Muitos dos exemplos explorados em aula e mencionados nas próximas secções deste relatório são, aliás, readaptados de alguns desses manuais utilizando dados de sítios arqueológicos reais recolhidos pelo docente. Entre as obras mais relevantes incluem-se:

- Banning, E. B. (2020). *The Archaeologist's Laboratory: The Analysis of Archaeological Evidence*. Springer Nature.
- Baxter, M. J. (2003). *Statistics in archaeology*. Arnold.Shennan, S. (1997). *Quantifying archaeology*. University of Iowa Press.
- Drennan, R. D. (2010). *Statistics for archaeologists. A common sense approach.* Springer.
- Lock, G. (2003). *Using computers in archaeology: Towards virtual pasts*. Routledge.
- McCall, G. S. (2018). *Strategies for quantitative research: Archaeology by numbers*. Routledge.
- McPherron, S. P., & Dibble, H. L. (2002). *Using computers in archaeology: A practical guide*. McGraw-Hill.
- Orton, C. (2000). *Sampling in archaeology*. Cambridge University Press.
- VanPool, T. L., & Leonard, R. D. (2011). *Quantitative analysis in archaeology*. John Wiley & Sons.

Naturalmente, muitos outros recursos estão disponíveis para os alunos consultarem, incluindo todas as referências citadas ao longo deste relatório e que aparecem na secção de bibliografia no final do documento. Por outro lado, e porque se trata de uma unidade curricular com uma grande componente prática, os alunos são também incentivados a utilizar a internet como fonte valiosíssima de informação relativamente, por exemplo, à realização de operações em folhas de cálculo. Atualmente, com a proliferação das plataformas de Inteligência Artificial capazes de gerar texto semelhante a um humano com base no contexto e em conversas anteriores, como o ChatGPT, a consulta de soluções para a análise e gestão de dados torna-se ainda mais fácil. Futuras edições desta unidade curricular, não poderão deixar, por isso, de juntar essa componente como recurso de consulta online.

Organização e conteúdos das aulas

Aula 01 - Introdução

A primeira aula da unidade curricular é dedicada à importância do registo digital e da análise quantitativa em Arqueologia. Esta aula centra-se, por um lado, em como a transição para a recolha, processamento e análise de dados num ambiente totalmente digital pode agilizar o fluxo de trabalho, permitir uma melhor planificação do trabalho arqueológico e reduzir o potencial de erros e a perda de informação irrecuperável do registo arqueológico. Por outro lado, explora como as ferramentas digitais têm facilitado a realização e apresentação de análises exploratórias de dados e estatística descritiva.

O registo em Arqueologia: do físico ao digital

Embora a opção preferencial para o registo arqueológico seja a sua preservação *in situ*, a maior parte das realidades com que o arqueólogo se depara não permitem esta conservação. Muitos dos sítios arqueológicos são destruídos, tapados ou transformados após a intervenção. A escavação de um sítio arqueológico é, desta forma, uma experiência que não se pode voltar a repetir e, portanto, a importância do registo exato de todas as suas particularidades deve ser a principal preocupação do arqueólogo.

As abordagens atuais ao processo de registo durante a escavação arqueológica prevêem, sobretudo, e de uma forma o mais lógica possível, ligar todos os elementos díspares de um sítio arqueológico no âmbito do registo espacial tridimensional. Naturalmente, a escavação arqueológica não é uma prática precisa, uma vez que existe muito espaço para ambiguidade. O processo de escavar e registar um sítio arqueológico é uma mistura curiosa entre intuição, interpretação e rigor científico (Lock, 2003). O resultado desse processo é um arquivo composto por dados escritos, desenhados e fotografados dos objetos escavados e da sua relação espacial. Esta preservação por registo tem como principal vantagem o facto de que esse arquivo estará acessível para análise, interpretação e futuras reinterpretações. Coletivamente, esse registo fornece o contexto dos achados feitos, ou seja, os padrões específicos de associação física que os relacionam entre si e com os sedimentos em que foram encontrados (Mitchell, 2018).

No entanto, para garantir sucesso neste processo é essencial que os dados estejam estruturados e guardados de forma lógica e sem ambiguidades. O uso crescente de computadores e, particularmente, de software de Sistemas de Bases de Dados, tem sido fundamental para o desenvolvimento dos sistemas de registo de escavação nas últimas décadas.

Devido às particularidades dos sítios arqueológicos e às diferentes ideias e objetivos das equipas de investigação e responsáveis, conhecemos hoje um conjunto muito alargado de sistemas de registo. Ainda assim, é possível reconhecer alguns conceitos-chave e requisitos-chave que integram, atualmente, a maior parte dos trabalhos de escavação. Um destes conceitos é, por exemplo, o da Matriz de Harris (Harris, 1997) nomeadamente o da conceptualização do registo de contextos únicos, em que um contexto é qualquer evento natural ou antrópico que pode ser distinguido no registo arqueológico.

As primeiras abordagens ao registo de escavação consistiam em texto descritivo, notas curtas e esquemas anotados nos cadernos de campo, produzindo um tipo de registo que é difícil de armazenar e, principalmente, de analisar de forma objetiva, por ser pouco lógico na sua estrutura e não ter sentido explícito (Figura 2). O início da busca por cumprir com estes dois últimos requisitos foi aquando da introdução das folhas proforma de registo de campo (Figura 3). Ainda hoje, em várias escavações, é comum ter vários formulários de registo que representam várias classes diferentes de informação, *e.g.* formulários de contexto, formulários de contextos especiais, formulários individuais de achados, etc.

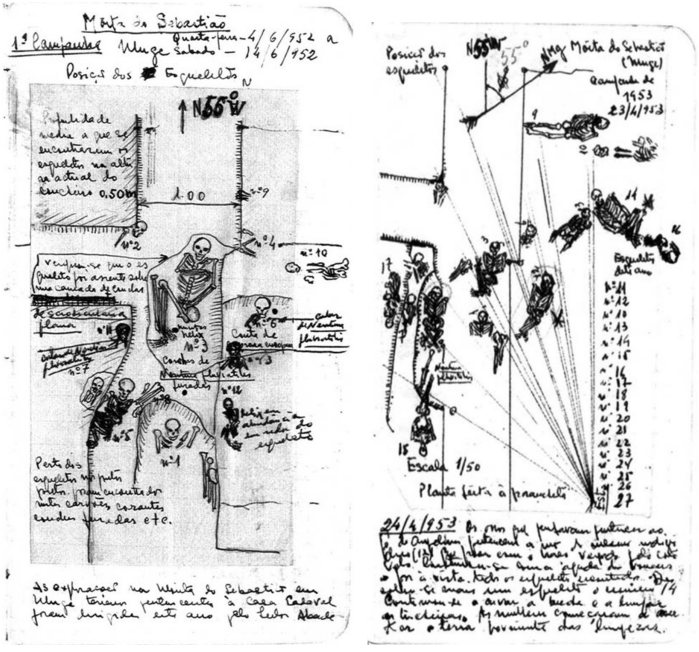
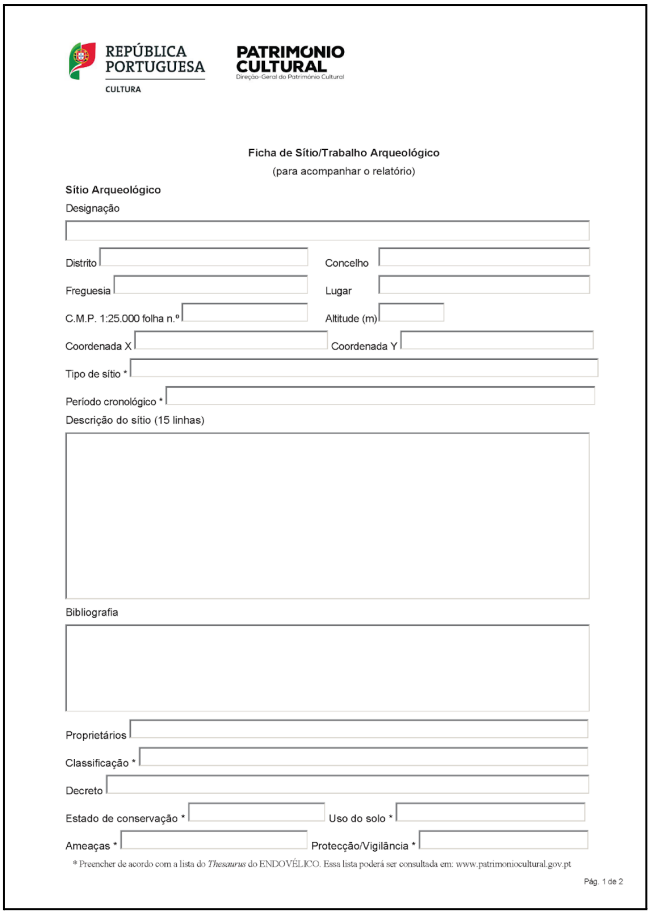


Figura 2. Páginas dos cadernos de campo de O. da Veiga Ferreira relativas às escavações do concheiro da Moita do Sebastião de 1952-53 (Cardoso & Rolão, 1999/2000).



REPÚBLICA PORTUGUESA
CULTURA

PATRIMÓNIO CULTURAL
Direção-Geral do Património Cultural

Ficha de Sítio/Trabalho Arqueológico
(para acompanhar o relatório)

Sítio Arqueológico

Designação

Distrito | Concelho

Freguesia | Lugar

C.M.P. 1:25.000 folha n.º | Altitude (m)

Coordenada X | Coordenada Y

Tipo de sítio *

Período cronológico *

Descrição do sítio (15 linhas)

Bibliografia

Proprietários

Classificação *

Decreto

Estado de conservação * | Uso do solo *

Ameaças * | Protecção/Vigilância *

* Preencher de acordo com a lista do *Thesaurus* do ENDOVÉLICO. Essa lista poderá ser consultada em: www.patrimoniocultural.gov.pt

Pág. 1 de 2

Figura 3. Ficha de sítio da Direção Geral do Património Cultural. Disponível em: https://www.patrimoniocultural.gov.pt/static/data/recursos/formularios/fichasitiodgpc_distributed.pdf

No entanto, aspectos como a ilegibilidade, folhas de dados incompletas e discrepâncias entre observadores são alguns dos inúmeros desafios que os arqueólogos enfrentam quando recolhem dados no terreno utilizando métodos estritamente analógicos. Por outro lado, frequentemente, os arqueólogos lamentam as horas despendidas a preencher formulários, que consomem um tempo valioso que poderia ser dedicado a outras tarefas.

Por estas razões, atualmente, muitos destes formulários foram substituídos por abordagens totalmente digitais ao trabalho de campo arqueológico. De facto, a completa digitalização de dados está a tornar-se um requisito prático para todo o trabalho de campo e de laboratório (Austin, 2014; Kintigh et al., 2014; Ogburn, 2010). Várias entidades de financiamento e tutelares de Património e Arqueologia, obrigam os profissionais a apresentarem uma estratégia para o arquivo e divulgação a longo prazo de todos os dados criados durante a investigação. Em Portugal, por exemplo, a Direção Geral do Património Cultural (DGPC) exige a submissão dos dados recolhidos em todos os trabalhos de campo em suporte físico e digital, para efeitos de preservação dos dados a longo prazo.

A digitalização é frequentemente vista como proporcionadora de uma maior flexibilidade através da sua separação entre função e forma, entre conteúdo e meio, na forma como pode quebrar as fronteiras entre os dados, encoraja e apoia uma utilização dinâmica e colaborativa e proporciona mais oportunidades para a recombinação de dados e a criação de novos conjuntos de dados. Kaufman & Jeandesboz (2017) sugerem uma série de possibilidades digitais, muitas das quais diretamente relacionadas com a utilização e relação com os dados. Estas incluem a maleabilidade e flexibilidade dos dispositivos digitais, as suas capacidades de armazenamento, a sua capacidade de pesquisa, a sua conetividade, a sua computabilidade, a sua natureza interactiva e a sua criação e organização de dados. A combinação de todos estes fatores, e não só, cria um ambiente indiscutivelmente atrativo para a produção, transformação, consumo e criação de conhecimento a partir dos dados.

No trabalho arqueológico, a transição para a recolha e gestão de dados de forma totalmente digital é uma inevitabilidade, dadas as características dos instrumentos de medição utilizados atualmente no campo e no laboratório. A adoção generalizada da Estação Total em trabalhos de escavação e prospecção (Bernatchez & Marean, 2011; McPherron & Dibble, 2002), por exemplo, resultou no armazenamento totalmente digital dos dados espaciais, seja na memória interna do equipamento, seja em computadores ligados por cabo ou por bluetooth ao equipamento topográfico. Este último cenário tem frequentemente levado ao desenvolvimento de software próprio para estabelecer essa ligação e cumprir com necessidades muito específicas em Arqueologia - ver, por exemplo, as soluções disponibilizadas em www.oldstoneage.com/osa/tech/index/. No laboratório, por outro lado, quase todos os equipamentos utilizados para leitura, registo e análise de materiais arqueológicos são, hoje em dia, completamente digitais. Até mesmo o paquímetro, um instrumento amplamente utilizado para obter medições lineares de objetos, tem versões digitais desde o final da década de 1980 que permitem a sua ligação a um computador, e com um simples pressionar de botão, as medidas são enviadas de forma imediata para o campo da base de dados utilizada.

Um campo particular onde a digitalização total dos dados tem estado particularmente ativa nos últimos anos é o da utilização de dispositivos móveis (*i.e.*, smartphones e tablets) para recolha de dados no campo e no laboratório (Averett et al., 2016; Cascalheira et al., 2014). A razão para tal é bastante simples: estes dispositivos integram, à partida, ferramentas que são essenciais para o trabalho arqueológico. Componentes como chip de GPS, câmara fotográfica e capacidade de comunicação via bluetooth tornam os dispositivos móveis equipamentos eficazes para registar localizações de sítios arqueológicos, registar achados fotograficamente ou comunicar com equipamentos de topografia para registo de coordenadas durante os trabalhos de escavação. Um dos melhores exemplos da aplicação destas ferramentas, em conjunto com

software personalizado, foi o desenvolvido por Cascalheira et al. (2017) com base em serviços da Google e no sistema operativo Android. Esta solução, criada para registar, gerir e partilhar dados de prospecção arqueológica no âmbito de um projeto de investigação sobre a Idade da Pedra em Moçambique, foi concebida para ser totalmente integrada e personalizável (Figura 4). O núcleo do sistema é constituído por duas aplicações personalizadas para dispositivos móveis, como smartphones ou tablets, desenvolvidas através do MIT App Inventor. Com estas aplicações, os prospetores conseguem obter coordenadas geográficas e dados relevantes de locais com vestígios arqueológicos e ainda realizar análises de artefactos *in situ*, incluindo medições precisas com paquímetros digitais conectáveis diretamente aos dispositivos móveis (Figura 5). Todos os dados recolhidos são armazenados na memória interna dos dispositivos, bem como, originalmente, numa base de dados espacial baseada na nuvem (a já extinta Google Fusion Tables). Esta base de dados permitia a partilha automática e análise dos dados, utilizando um conjunto bastante intuitivo de ferramentas de visualização para a criação imediata de mapas e gráficos exploratórios.

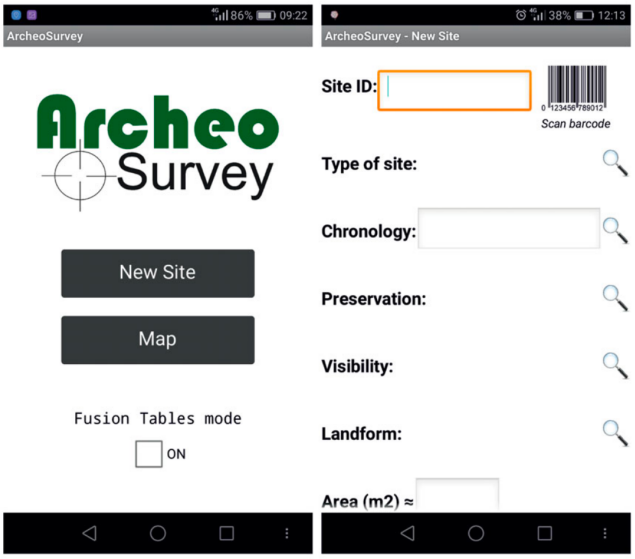


Figura 4. Ecrãs principais da aplicação ArcheoSurvey para dispositivos móveis Android.

Figura 5. Exemplo de utilização de um paquímetro digital conectado diretamente a um smartphone para recolha de dados de indústrias líticas durante o trabalho de campo.

A análise quantitativa em Arqueologia

Os avanços na tecnologia digital não se limitaram apenas ao equipamento e à rapidez e eficácia com que recolhemos dados arqueológicos. No que concerne ao processamento e análise, os desenvolvimentos informáticos das últimas duas décadas possibilitam, atualmente, que qualquer pessoa, independentemente da sua experiência, aplique facilmente métodos quantitativos ou estatísticos a um determinado conjunto de dados.

Na maior parte dos casos, quando se fala de estatística a um grupo de alunos de Arqueologia, a reação tende a ser de receio e surpresa, uma vez que, frequentemente, estes alunos referem terem escolhido as Humanidades como forma de "fugir à matemática". Contudo, os cálculos matemáticos envolvidos na análise quantitativa ou estatística, por mais simples ou complexos que sejam, são hoje considerados uma coisa do passado. Atualmente, nenhum estudante ou profissional de Arqueologia precisa de realizar manualmente cálculos como a média ou o desvio padrão de um conjunto de dados. Em vez disso, praticamente todos os cálculos são automaticamente realizados pela miríade de pacotes de software estatístico disponíveis. Naturalmente, isso torna a pesquisa estatística muito mais acessível, rápida e exacta.

Certamente, a matemática está envolvida nos cálculos estatísticos, bem como nas representações gráficas e interpretações. No entanto, a estatística é mais uma abordagem lógica

para fazer inferências sobre a natureza de toda uma população com base numa amostra dessa população, o que (de um modo geral) envolve a análise de factos numéricos. Embora nunca seja demais saber um pouco de matemática, não é necessário ser um génio da matemática para realizar uma investigação estatística sólida e coerente (McCall, 2018).

O grande avanço no pensamento estatístico ocorreu no final do século XIX e início do século XX, quando cientistas como Karl Pearson e Ronald Fisher começaram a pensar na forma como os investigadores poderiam avaliar as relações entre amostras e populações através do desenvolvimento de uma melhor compreensão matemática de questões como a distribuição, a variância e a probabilidade. Alguns destes cálculos, como o cálculo dos desvios-padrão para a avaliação da variância, são de facto bastante simples. Outros, como a modelação de distribuições estatísticas utilizando funções matemáticas, são bastante complexos. A questão é que as perguntas sobre a relação entre amostras e populações são, na verdade, questões de lógica, e a matemática só entra em jogo quando realizamos cálculos para avaliar estas questões.

Dito isto, como argumentado por McCall (2018), é possível fazer uma investigação estatística sólida utilizando o software sofisticado que está atualmente disponível, mesmo sem compreender toda a matemática envolvida nos cálculos estatísticos selecionados. Como analogia, é um pouco como conduzir um carro sem compreender completamente todos os pormenores do funcionamento do motor de combustão interna. Contudo, existem ainda muitas coisas que podem correr mal nesta situação e, tal como acontece com a condução de um automóvel, são necessárias algumas competências para obter bons resultados. Acima de tudo, os arqueólogos que utilizam métodos quantitativos precisam sempre de conhecer as circunstâncias em que devem usar certos modelos e abordagens, e isso requer um conhecimento prático dos diferentes tipos de dados que podemos recolher.

Em Arqueologia, temos duas formas principais de quantificar no nosso trabalho: 1) contando "coisas" classificadas como pertencentes a tipos específicos ou outras classes de fenómenos; e 2) medindo "coisas" de forma padronizada (medidas lineares, peso, composição, etc.). O objetivo fundamental destes procedimentos de quantificação é facilitar comparações sistemáticas entre as populações de "coisas" provenientes de diferentes contextos, como, por exemplo, as coleções de artefactos de diferentes sítios arqueológicos ou de diferentes unidades estratigráficas de um mesmo sítio arqueológico. Assim, é inevitável constatar que a Arqueologia se baseia fortemente na quantificação. Em parte, isto está relacionado com a gama limitada de “coisas” às quais os arqueólogos têm acesso para fazer observações. Ao contrário de, por exemplo, os antropólogos culturais, que são capazes de fazer observações diretas sobre o amplo e complexo meio de

interações sociais no mundo humano, a investigação arqueológica concentra-se principalmente no estudo das relações entre certas "coisas" e outras certas "coisas".

A quantificação e a análise quantitativa são instrumentos cruciais para caracterizar estas relações de forma sistemática, permitindo uma avaliação mais aprofundada do seu significado.

Aula 02 - Dados e Bases de Dados

Nesta aula, são apresentados os princípios básicos de uma base de dados, incluindo os tipos de bases de dados mais utilizados, bem como os tipos de variáveis e escalas de medida. O objetivo é, além da familiarização com os diversos conceitos, demonstrar a importância das decisões tomadas antes do início da recolha de dados, e como essas decisões influenciarão os tipos de análises que poderão ser posteriormente realizadas.

Tipos de bases de dados

Uma base de dados é um repositório de informação relacionada com um determinado assunto ou finalidade, ou seja, é uma coleção de dados ou itens de informação estruturados de maneira específica, que permite a sua consulta, atualização e outros tipos de operações processadas por meios informáticos. Serve para gerir vastos conjuntos de informação de modo a facilitar a organização, manutenção e pesquisa de dados. Existem vários tipos de bases de dados digitais, incluindo bases de dados relacionais, NoSQL, orientadas a objetos, multimodais, espaciais, entre outras. Em Arqueologia, são utilizados, sobretudo, três tipos principais de bases de dados: **folhas de cálculo**, **bases de dados relacionais** e **bases de dados espaciais**.

Folhas de cálculo

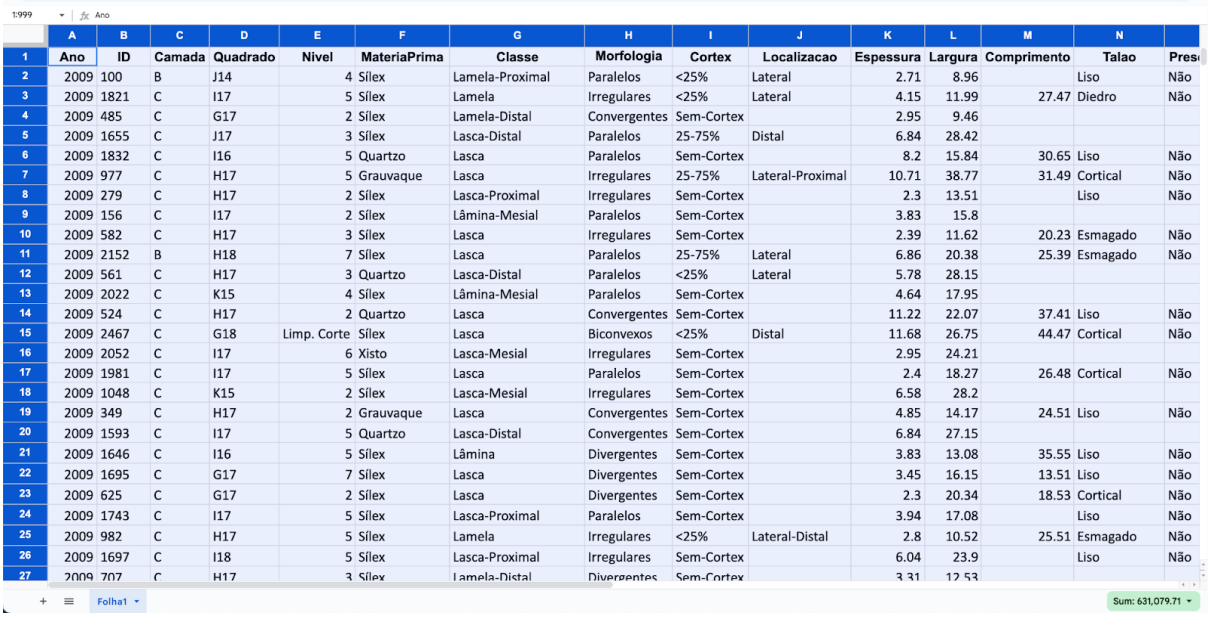
Uma folha de cálculo é um tipo de ficheiro para armazenamento de dados, onde os dados são guardados numa estrutura semelhante a uma tabela. É, em certo sentido, muito semelhante ao sistema de quadrícula arqueológica usado para identificar unidades de escavação, com **linhas** horizontais (normalmente identificadas com números) e **colunas** verticais (normalmente identificadas com letras) (McPherron & Dibble, 2002). A intersecção entre uma linha específica e uma coluna específica denomina-se **célula**. Numa folha de cálculo utilizada como base de dados, cada linha representa um registo, cada coluna representa um campo e cada célula contém os valores observados.

Em Arqueologia, as folhas de cálculo são amplamente utilizadas para a recolha e análise de dados. A sua principal vantagem reside na simplicidade e na capacidade de serem lidas pela maioria dos programas mais utilizados para análises quantitativas de dados (como MS Excel, Google Sheets, SPSS, entre outros), sistemas de gestão de bases de dados (por exemplo, MS

Access) e linguagens de programação (por exemplo, R, Python). Muitas ferramentas de transformação de dados e bibliotecas de programação oferecem suporte integrado para a leitura e escrita destes ficheiros, tornando-as uma escolha popular para a troca de dados em vários domínios (Figura 6).

Os formatos mais comuns e aconselháveis para este tipo de base de dados são os CSV (comma-separated values) e TXT (ficheiro de texto simples). Nos ficheiros CSV, os campos são sempre separados por vírgulas. Já nos ficheiros TXT, os campos podem ser separados por vírgula, ponto e vírgula ou por tabulação. Ao contrário de outros formatos, estes são completamente abertos do ponto de vista do licenciamento e, por essa razão, são facilmente partilhados entre utilizadores que usem diferentes programas ou sistemas operativos. Além disso, por não precisarem de um programa específico, podem ser abertos e modificados até com um simples editor de texto.

Contudo, à medida que os requisitos de armazenamento de dados crescem em tamanho e complexidade, as folhas de cálculo podem tornar-se menos eficientes e escaláveis comparativamente às bases de dados relacionais, que oferecem capacidades mais avançadas de consulta, indexação e gestão de dados.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Ano	ID	Camada	Quadrado	Nivel	MateriaPrima	Classe	Morfologia	Cortex	Localizacao	Espessura	Largura	Comprimento	Talao	Pres
2	2009	100	B	J14	4	Sílex	Lamela-Proximal	Paralelos	<25%	Lateral	2.71	8.96		Liso	Não
3	2009	1821	C	I17	5	Sílex	Lamela	Irregulares	<25%	Lateral	4.15	11.99	27.47	Diedro	Não
4	2009	485	C	G17	2	Sílex	Lamela-Distal	Convergentes	Sem-Cortex		2.95	9.46			
5	2009	1655	C	J17	3	Sílex	Lasca-Distal	Paralelos	25-75%	Distal	6.84	28.42			
6	2009	1832	C	I16	5	Quartzo	Lasca	Paralelos	Sem-Cortex		8.2	15.84	30.65	Liso	Não
7	2009	977	C	H17	5	Grauvaque	Lasca	Irregulares	25-75%	Lateral-Proximal	10.71	38.77	31.49	Cortical	Não
8	2009	279	C	H17	2	Sílex	Lasca-Proximal	Irregulares	Sem-Cortex		2.3	13.51		Liso	Não
9	2009	156	C	I17	2	Sílex	Lâmina-Mesial	Paralelos	Sem-Cortex		3.83	15.8			
10	2009	582	C	H17	3	Sílex	Lasca	Irregulares	Sem-Cortex		2.39	11.62	20.23	Esmagado	Não
11	2009	2152	B	H18	7	Sílex	Lasca	Paralelos	25-75%	Lateral	6.86	20.38	25.39	Esmagado	Não
12	2009	561	C	H17	3	Quartzo	Lasca-Distal	Paralelos	<25%	Lateral	5.78	28.15			
13	2009	2022	C	K15	4	Sílex	Lâmina-Mesial	Paralelos	Sem-Cortex		4.64	17.95			
14	2009	524	C	H17	2	Quartzo	Lasca	Convergentes	Sem-Cortex		11.22	22.07	37.41	Liso	Não
15	2009	2467	C	G18	Limp. Corte	Sílex	Lasca	Biconvexos	<25%	Distal	11.68	26.75	44.47	Cortical	Não
16	2009	2052	C	I17	6	Xisto	Lasca-Mesial	Irregulares	Sem-Cortex		2.95	24.21			
17	2009	1981	C	I17	5	Sílex	Lasca	Paralelos	Sem-Cortex		2.4	18.27	26.48	Cortical	Não
18	2009	1048	C	K15	2	Sílex	Lasca-Mesial	Irregulares	Sem-Cortex		6.58	28.2			
19	2009	349	C	H17	2	Grauvaque	Lasca	Convergentes	Sem-Cortex		4.85	14.17	24.51	Liso	Não
20	2009	1593	C	I17	5	Quartzo	Lasca-Distal	Convergentes	Sem-Cortex		6.84	27.15			
21	2009	1646	C	I16	5	Sílex	Lâmina	Divergentes	Sem-Cortex		3.83	13.08	35.55	Liso	Não
22	2009	1695	C	G17	7	Sílex	Lasca	Divergentes	Sem-Cortex		3.45	16.15	13.51	Liso	Não
23	2009	625	C	G17	2	Sílex	Lasca	Divergentes	Sem-Cortex		2.3	20.34	18.53	Cortical	Não
24	2009	1743	C	I17	5	Sílex	Lasca-Proximal	Paralelos	Sem-Cortex		3.94	17.08		Liso	Não
25	2009	982	C	H17	5	Sílex	Lamela	Irregulares	<25%	Lateral-Distal	2.8	10.52	25.51	Esmagado	Não
26	2009	1697	C	I18	5	Sílex	Lasca-Proximal	Irregulares	Sem-Cortex		6.04	23.9		Liso	Não
27	2009	707	C	H17	3	Sílex	Lamela-Distal	Divergentes	Sem-Cortex		3.31	12.53			

Figura 6. Exemplo de folha de cálculo com dados provenientes da análise de ferramentas em pedra do sítio arqueológico de Vale Boi. Cada linha representa um artefacto, e cada coluna, uma das variáveis analisadas. Os dados presentes nas células representam os valores para cada uma das variáveis ou campos.

Bases de dados relacionais

As bases de dados relacionais são construídas com base nos princípios do modelo relacional, que define os dados como conjuntos de tabelas. De uma forma simplificada, uma base de dados relacional agrega várias folhas de cálculo ou ficheiros planos, que contêm informação distinta sobre o mesmo conjunto de dados.

A estrutura de uma base de dados relacional baseia-se no conceito de **chaves** ou **identificadores únicos** (em inglês, *Globally Unique Identifier,* ou GUID), que desempenham um papel vital na manutenção da integridade dos dados e no estabelecimento de relações entre tabelas. A seleção de uma **chave primária** é essencial, pois influencia o desempenho e a eficiência das consultas e das junções na base de dados. Eis algumas das principais características das chaves primárias:

- Unicidade: Cada valor presente na coluna da chave primária deve ser único. Esta característica assegura que não existam dois registos na mesma tabela com o mesmo identificador.
- Não-nulidade: Uma chave primária não pode conter valores NULL (*i.e.*, nulos). É imperativo que cada registo na tabela possua um valor de chave primária válido e não nulo.
- Imutabilidade: Idealmente, os valores atribuídos às chaves primárias não devem ser alterados após a sua atribuição a um registo. Isto garante a consistência e previne potenciais problemas na referência de dados.
- Coluna única ou chave composta: Uma chave primária pode ser composta por uma única coluna ou por uma combinação de múltiplas colunas. Neste último caso, é designada por chave composta.

Um exemplo prático de uma base de dados relacional é a utilizada pelo software Newplot, um programa de código aberto (*i.e.*, *open source)*, desenvolvido para a gestão de dados provenientes de escavações arqueológicas. O Newplot funciona sobre uma base de dados relacional em MS Access, que compreende várias tabelas, sendo as mais significativas as tabelas *Context* e *XYZ*. A tabela *Context* armazena informação sobre determinados artefactos ou conjunto de artefactos e o contexto da sua descoberta (*e.g.*, unidade estratigráfica, tipo de material). Por sua vez, a tabela *XYZ* regista informações espaciais relativas à proveniência desses artefactos, incluindo as coordenadas tridimensionais x, y e z. Ambas as tabelas estão interligadas por uma chave primária composta, representada pelas variáveis *UNIT* e *ID*, que correspondem, respetivamente, ao quadrado, área de escavação ou sítio arqueológico (*e.g.*, A1, L15, Abrigo, Vale Boi), e a um

número sequencial de identificação. A combinação *UNIT* e *ID* deve ser única em toda a base de dados. Por exemplo, em toda a base de dados, apenas um objeto pode ter como chave primária "A1-100", permitindo correlacionar as suas coordenadas na tabela *XYZ* com a respetiva informação contextual na tabela *Context*. Em certos casos, torna-se necessário registar múltiplas coordenadas tridimensionais para um único objeto, com o objetivo de recolher dados relativos à sua orientação ou às suas dimensões e morfologia geral. No entanto, esta metodologia implica a criação de mais do que um registo na tabela *XYZ* para cada objeto. A base de dados do Newplot está preparada para aceitar várias entradas na tabela *XYZ* para um único objeto, mantendo o mesmo conjunto UNIT-ID, mas adicionando uma outra variável, denominada *SUFFIX*. Neste caso, a relação entre as tabelas *Context* e *XYZ* é caracterizada como um-para-múltiplos (1:M), em que várias entradas em *XYZ* correspondem a uma única entrada em *Context*.

Bases de dados espaciais

As bases de dados espaciais constituem um tipo especializado de bases de dados, concebidas para armazenar, gerir e processar dados espaciais ou geográficos. Desempenham um papel crucial em diversas aplicações que necessitam de informações baseadas na localização, como é o caso dos Sistemas de Informação Geográfica (SIG), a monitorização ambiental, o planeamento urbano, a logística, os serviços baseados na localização e, claro, em Arqueologia. Estas bases de dados estão equipadas com técnicas avançadas de indexação espacial e capacidades de consulta espacial, as quais permitem a recuperação e análise eficientes de dados espaciais. Facilitam a representação de objetos geométricos simples, tais como pontos, linhas e polígonos, e, em certos casos, são também capazes de lidar com estruturas mais complexas, como objetos tridimensionais e redes triangulares irregulares (conhecidas em inglês como *Triangulated Irregular Network* ou TIN).

Quase a totalidade do trabalho de campo arqueológico produz grandes quantidades de informação baseada na localização, seja de sítios arqueológicos ou de objetos encontrados num determinado sítio arqueológico. A proveniência de cada artefacto e amostra é crucial para entender a sua idade, tafonomia, contexto ecológico, contexto cultural e significado comportamental. Assim, a proveniência espacial de um objeto é frequentemente tão importante quanto o próprio objeto. Esta realidade torna a utilização de bases de dados espaciais em Arqueologia particularmente vantajosa, especialmente quando os dados geográficos incluem coordenadas absolutas, referenciadas a um *datum* e a um sistema de projeção específicos.

Um exemplo do uso de uma base de dados espacial para consulta online do inventário de sítios arqueológicos de Portugal Continental é o Geoportal da Direção Geral do Património Cultural (Figura 7). Esta base de dados utiliza o formato Geodatabase da ESRI, um formato proprietário usado pelo conjunto de software ArcGIS da ESRI. Este formato oferece um modelo de dados espaciais abrangente, que inclui diversos tipos de dados, como pontos, linhas, polígonos e dados raster.

Figura 7. Geoportal da Direção Geral do Património Cultural (https://patrimoniodgpc.maps.arcgis.com)

Outro exemplo disponível para consulta é o projeto PaleoCore (Reed et al., 2015), uma plataforma online dedicada à gestão de dados arqueológicos, paleontológicos e geológicos, com foco particular nos períodos Pliocénico e Pleistocénico (Figura 8). A base de dados deste projeto é suportada por software de acesso aberto, incluindo a extensão PostGIS do popular sistema de gestão de bases de dados relacionais PostgreSQL. O PostGIS adiciona suporte para a gestão e análise de dados espaciais, convertendo o PostgreSQL num sistema de bases de dados espacial completo, que inclui tipos de dados espaciais específicos, índices espaciais e funções espaciais.

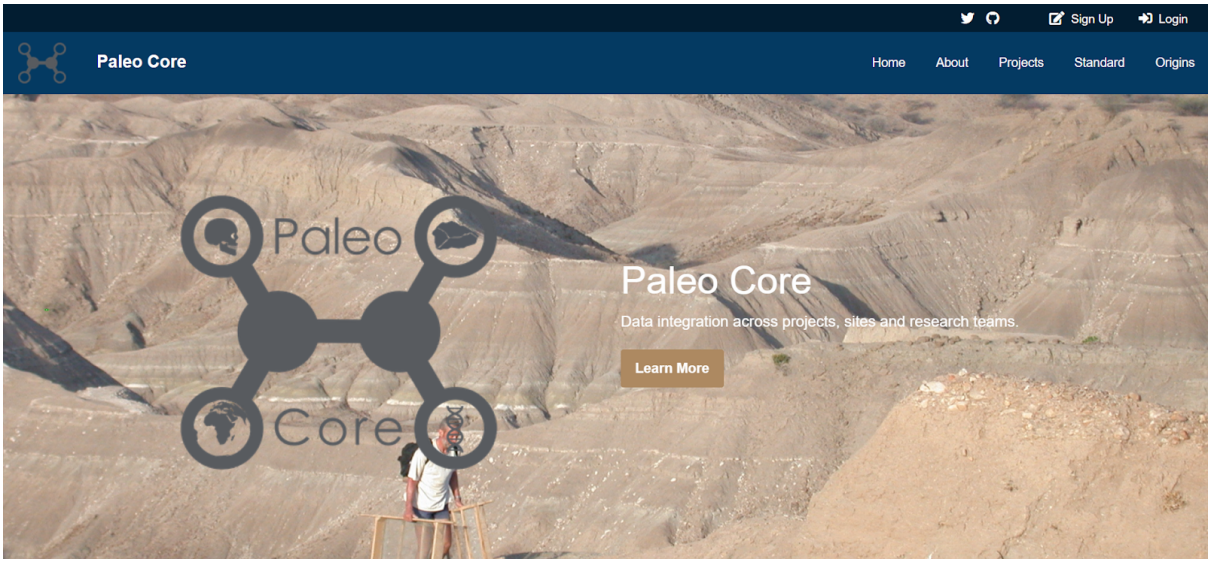


Figura 8. Plataforma Paleocore (https://paleocore.org/), dedicada à gestão de dados arqueológicos, paleontológicos e geológicos, com foco particular nos períodos Pliocénico e Pleistocénico.

Tipos de dados

Todos os formatos de bases de dados permitem a definição de diversos tipos de dados. A seleção do tipo de dados deve ser realizada conforme a natureza da informação que se pretende armazenar nesse campo específico. Devemos optar pelos tipos de dados que melhor se adaptam à informação que pretendemos armazenar/guardar. O tipo de dados escolhido determina a natureza dos dados que podem ser armazenados, o intervalo de valores que podem assumir e as operações que podem ser efectuadas sobre os mesmos. Por vezes, confundem-se os tipos de dados que podem ser selecionados numa base de dados com os tipos de variáveis estatísticas. Embora relacionados, no caso das bases de dados, são tidas em consideração principalmente as operações de armazenamento e transformação, resultando numa maior variedade de tipos em comparação com as variáveis estatísticas. Na Tabela 2, são apresentados os principais exemplos dos tipos de dados.

Tipo de dados	Descrição
Texto Texto Breve	Valores curtos e alfanuméricos, tais como um apelido ou um endereço
Número, Número Grande	Valores numéricos, como distâncias
Percentagem	Percentagens
Científico	Aceita numeração científica
Moeda	Valores monetários
Sim/Não	Também conhecido como lógico ou booleano. Os valores Sim e Não e os campos que contêm apenas um de dois valores
Data/Hora Data/Hora Prolongada	Data/Hora: valores de data e hora dos anos 100 a 9999. Data/Hora Prolongada: valores de data e hora dos anos 1 a 9999
Rich Text	Texto ou combinações de texto e números que podem ser formatados ao utilizar controlos de cor e de tipo de letra
Campo Calculado	Resultados de um cálculo. O cálculo tem de fazer referência a outros campos da mesma tabela
Anexo	Imagens, ficheiros de folha de cálculo, documentos, gráficos e outros tipos de ficheiros suportados anexados aos registos na base de dados, semelhante à forma de anexar ficheiros a mensagens de email
Hiperligação	Texto ou combinações de texto e números armazenados como texto e utilizados como um endereço de hiperligação
Memo Texto Longo	Longos blocos de texto. Uma utilização típica de um campo Memo seria uma descrição detalhada do produto
Pesquisa	Lista de valores obtida de uma tabela ou consulta ou apresenta um conjunto de valores especificados na criação do campo.

Tabela 2. Principais tipos de dados que definem os valores que uma coluna de uma base de dados pode conter.

Aula 03 - Recolha de dados

A terceira aula foca-se principalmente nos programas mais utilizados em Arqueologia para a recolha de dados, abordando as suas vantagens e desvantagens no contexto do registo de dados em laboratório. É dada especial atenção ao software E5, cujas características, desenvolvidas especificamente para a recolha de dados arqueológicos, oferecem múltiplas vantagens em comparação com outros programas. A exploração do software E5 tem também como objetivo familiarizar os alunos com, entre outros aspectos, a lógica das linguagens de programação, a nomeação e a escolha do tipo de variáveis.

Em geral, e tal como acontece em diversas disciplinas, os objetivos principais na prática arqueológica passam por recolher o máximo de dados necessários, da forma mais rápida possível e com o menor número de erros. Nesta aula, serão apresentadas e aplicadas técnicas destinadas a alcançar estes objetivos, especialmente no que diz respeito ao trabalho de laboratório e à recolha de dados a partir de artefactos arqueológicos. Algumas destas técnicas implicam o uso de hardware especializado. Outras, como a criação de menus e ficheiros de ajuda que contêm listas de respostas corretas ou válidas para as variáveis selecionadas, envolvem a utilização de software que pode ser aplicado à maioria dos programas de gestão de bases de dados.

Decidir quais variáveis incluir numa determinada base de dados para análise de materiais arqueológicos é, efetivamente, o primeiro passo antes de se iniciar a análise propriamente dita. No entanto, a escolha dessas variáveis depende em grande medida do objetivo do estudo e do tipo de objetos a analisar. Mais generalizada, contudo, é a preocupação em perceber como se podem traduzir essas observações num conjunto de variáveis numa base de dados que (1) armazene a informação de forma eficiente, (2) facilite a recuperação e análise dos dados, (3) se integre de modo harmonioso com as demais análises que eventualmente serão realizadas no contexto do projeto arqueológico em causa.

Neste contexto, uma das primeiras preocupações passa pela organização e decisão sobre a **unidade de análise**. Frequentemente, cada objeto é definido como a unidade de análise, sendo os campos da base de dados representativos das diversas observações realizadas individualmente. Num exemplo muito simples, no âmbito da análise de artefactos líticos, cada peça é considerada uma unidade de análise, com o registo de variáveis como Comprimento,

Largura, Espessura, Tipo de Talão, etc., para cada artefacto. Nestes casos, a introdução dos dados não constitui um grande desafio, tendo em conta que se pode efetuar a análise recorrendo a um único ficheiro plano ou folha de cálculo.

Noutros casos, como, por exemplo, na análise de marcas de corte em ossos de animais, a situação é muito mais desafiante. A principal razão reside no facto de existirem duas unidades principais de análise: o próprio osso (do qual se podem registar medidas, identificar a espécie, etc.), e cada uma das marcas de corte identificadas (que podem ser individualmente caracterizadas segundo as suas dimensões, perfil, etc.). Uma abordagem possível consiste em adicionar um conjunto de variáveis para cada marca de corte na base de dados. No entanto, esta estratégia levanta dois problemas: por um lado, a presença de muitas células em branco, uma vez que nem todos os ossos apresentarão marcas de corte; por outro lado, a dificuldade de prever o número de marcas de corte antes de iniciar a análise, o que implicaria adicionar novas variáveis ao longo do estudo, prática que deve ser sempre(!) evitada (McPherron & Dibble, 2003).

Assim, nestes casos, a melhor estratégia será dividir a análise em pelo menos duas tabelas distintas, correspondendo às duas unidades de análise definidas: uma para os ossos e outra para as marcas de corte. Naturalmente, estas duas tabelas devem estar interligadas através de um identificador único (*i.e.*, uma chave primária), que assegura a ligação entre os dois registos. Mais uma vez, os identificadores únicos são muito importantes, uma vez que sem essa referência seria impossível associar os dados registados ao(s) objeto(s) analisado(s). Desta forma, é também de extrema importância manter junto ao objeto analisado (por meio de uma etiqueta ou, quando possível, uma marcação direta na peça) o código do identificador único utilizado. Mesmo no caso de se tratarem de objetos pertencentes a coleções de museus, nos quais não exista um identificador único associado, deve ser criado um sistema de IDs que permita, em qualquer etapa da análise, correlacionar os dados recolhidos com o objeto em análise.

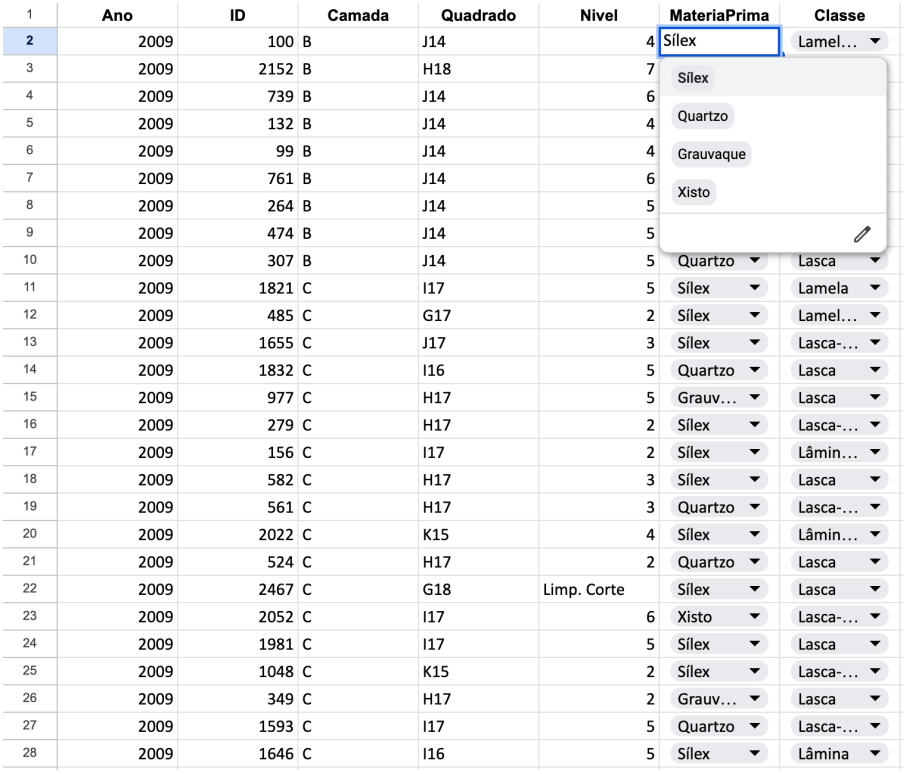
Uma vez definida(s) a(s) unidade(s) de análise, pode-se proceder à criação da base dados, atribuindo a cada variável um nome e especificando o tipo de dados a ser introduzido em cada campo. Embora possa parecer um detalhe menor, a nomeação das variáveis é de extrema importância no desenho de uma base de dados de qualquer natureza. Os nomes devem ser únicos, mas também descritivos e de fácil interpretação. Nomes genéricos como VAR1, VAR2, VAR3, etc., não facilitam a identificação da informação registrada nesse campo. Espaços em branco, acentos e símbolos especiais (*e.g.*, %), devem também ser evitados, uma vez que alguns softwares apresentam dificuldades na conversão desses caracteres. Contudo, um caractere que é

frequentemente utilizado é o *underscore* (em português, travessão cansado), especialmente para substituir espaços entre duas ou mais palavras (*e.g.*, Tipo_Talao).

Como mencionado anteriormente, um dos suportes informáticos mais utilizados para introdução de dados em Arqueologia são as folhas de cálculo. Com os princípios definidos acima, torna-se relativamente simples criar um novo documento em MS Excel, Google Sheets, LibreOffice Calc, ou outro programa de folhas de cálculo, inserir os nomes das variáveis na primeira linha e começar a análise, inserindo informação sobre cada objeto em cada uma das linhas subsequentes.

Este método é amplamente utilizado por arqueólogos em todo o mundo para analisar as suas coleções, principalmente devido à sua simplicidade. No entanto, este método apresenta problemas comuns. Um dos melhores exemplos é quando um conjunto de variáveis não é relevante para um objeto específico: o utilizador precisa de olhar para os nomes das variáveis, saltar com o cursor todas as células irrelevantes e retomar a análise na próxima variável aplicável. Este processo não só consome tempo precioso, mas também, frequentemente leva ao preenchimento incorreto de células. Por outro lado, a introdução manual de caracteres num teclado de computador numa folha de cálculo pode resultar em trocas de letras ou números, na introdução de espaços em branco antes ou após palavras, ou no uso indiferenciado de letras maiúsculas e minúsculas (*e.g.*, Lasca *vs* lasca). Todos estes problemas podem levar a que o programa de leitura de dados considere duas palavras que deveriam representar a mesma informação como sendo diferentes. A solução mais comum para estas situações, e para aumentar a rapidez com que os dados são inseridos nas folhas de cálculo, é a criação das chamadas tabelas de codificação. Por este método, em vez de o utilizador inserir uma palavra completa durante a análise (*e.g.*, Lasca), utilizam-se códigos numéricos para representar cada opção possível para uma determinada variável. Por exemplo, se as opções para uma variável como Matéria-Prima são Sílex, Quartzo, Quartzito, Outra, em vez de escrever o nome de cada uma das matérias, o utilizador poderá utilizar o algarismo 1 para Sílex, 2 para Quartzo, etc. Naturalmente, tendo em conta a quantidade e diversidade de variáveis que geralmente compõem uma base de dados arqueológica, estes códigos devem estar sempre disponíveis numa folha de apoio (digital ou, frequentemente, em papel) para que o utilizador possa sempre consultar qual o código correspondente à opção pretendida. A origem desta forma de inserir dados num computador remonta aos tempos em que a capacidade de processamento dos computadores era limitada, e a redução do espaço de memória ocupado por um código de um ou dois caracteres, em vez de uma palavra mais longa, representava uma vantagem significativa. Embora esta preocupação seja menos relevante atualmente, dada a capacidade de computação dos equipamentos modernos, a maioria dos programas de folhas de cálculo permite hoje a

criação de menus nas várias células, com opções predefinidas, para reduzir erros de escrita e acelerar a introdução dos dados (Figura 9).



	Ano	ID	Camada	Quadrado	Nivel	MateriaPrima	Classe
2	2009	100	B	J14	4	Sílex	Lamel…
3	2009	2152	B	H18	7		
4	2009	739	B	J14	6		
5	2009	132	B	J14	4		
6	2009	99	B	J14	4		
7	2009	761	B	J14	6		
8	2009	264	B	J14	5		
9	2009	474	B	J14	5		
10	2009	307	B	J14	5	Quartzo	Lasca
11	2009	1821	C	I17	5	Sílex	Lamela
12	2009	485	C	G17	2	Sílex	Lamel…
13	2009	1655	C	J17	3	Sílex	Lasca-…
14	2009	1832	C	I16	5	Quartzo	Lasca
15	2009	977	C	H17	5	Grauv…	Lasca
16	2009	279	C	H17	2	Sílex	Lasca-…
17	2009	156	C	I17	2	Sílex	Lâmin…
18	2009	582	C	H17	3	Sílex	Lasca
19	2009	561	C	H17	3	Quartzo	Lasca-…
20	2009	2022	C	K15	4	Sílex	Lâmin…
21	2009	524	C	H17	2	Quartzo	Lasca
22	2009	2467	C	G18	Limp. Corte	Sílex	Lasca
23	2009	2052	C	I17	6	Xisto	Lasca-…
24	2009	1981	C	I17	5	Sílex	Lasca
25	2009	1048	C	K15	2	Sílex	Lasca-…
26	2009	349	C	H17	2	Grauv…	Lasca
27	2009	1593	C	I17	5	Quartzo	Lasca-…
28	2009	1646	C	I16	5	Sílex	Lâmina

Figura 9. Exemplo da utilização de menus numa folha de cálculo Google Sheets, salientando as opções disponíveis para escolha das várias matérias-primas durante a análise de uma coleção de ferramentas em pedra.

No entanto, existem atualmente alternativas mais práticas que possibilitam a introdução de dados de forma mais rápida e com menor margem de erro em comparação com a tradicional folha de cálculo. Uma dessas soluções, especificamente desenvolvida para projetos arqueológicos, é o software E5, que faz parte do conjunto de programas criados por Shannon McPherron e Harold Dibble. Este software é de uso totalmente livre e está disponível para download em: https://github.com/surf3s/E5 (Figura 10).

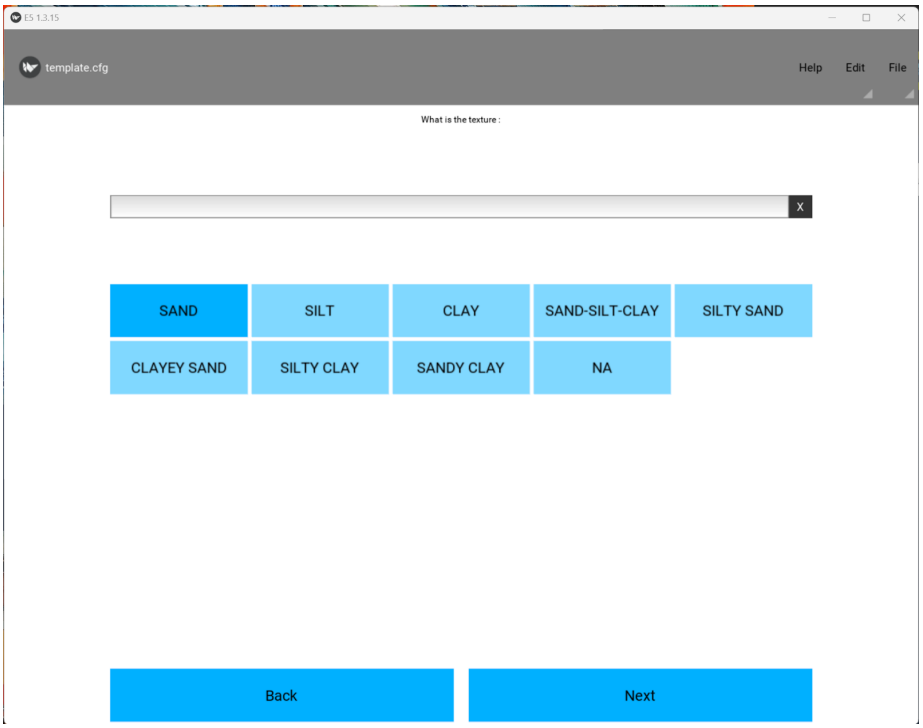


Figura 10. Página de exemplo do ambiente de seleção de atributos através de menu no programa E5.

O E5 é um programa genérico de introdução de dados, que funciona com um ficheiro de configuração no qual são definidos os campos para a introdução de dados. A sua principal vantagem reside na capacidade de condicionar o preenchimento de variáveis com base em valores já inseridos em variáveis anteriores.

O elemento-chave do E5 é o ficheiro de configuração, onde são definidos os campos para a introdução de dados. Os ficheiros de configuração, que têm a extensão CFG, podem parecer algo complexos à primeira vista e devem ser escritos num programa separado, num editor de texto como o NotePad, NotePad++, Atom ou Sublime text. Contudo, o esforço de criar um ficheiro de configuração implica uma reflexão antecipada sobre a estrutura dos dados antes de iniciar a sua recolha, diferentemente do que ocorre, por exemplo, com a folha de cálculo. Este investimento inicial tende a ser compensado mais tarde, durante a fase de análise dos dados.

Vários exemplos de ficheiros CFG estão disponíveis no website do E5. A seguir, apresenta-se um exemplo de ficheiro de configuração disponibilizado na página oficial, com o objetivo de ilustrar algumas das principais características:

```
[E5]
TABLE=lithics

[ID]
TYPE=TEXT
PROMPT=Enter the artifact ID
UNIQUE=True

[ARTIFACTTYPE]
TYPE=MENU
PROMPT=Select the artifact type
MENU=Tool,Flake,Core

[TOOLTYPE]
TYPE=MENU
PROMPT=Select the tool type
MENU=Scraper,Notch,Point,Other
CONDITION1=ArtifactType Tool

[PLATFORMTYPE]
TYPE=MENU
PROMPT=What is the platform
MENU=Plain,Cortical,Missing,Other
CONDITION1=ArtifactType Tool,Flake

[PLATFORMWIDTH]
TYPE=NUMERIC
PROMPT=Measure the platform width
CONDITION1=ArtifactType Tool,Flake
CONDITION2=PlatformType not Missing

[WEIGHT]
TYPE=NUMERIC
PROMPT=WEIGHT
```

O ficheiro de configuração está organizado em blocos, definidos pela presença, na primeira linha, dos símbolos [ ], dentro dos quais se insere o nome de cada variável. Nas linhas seguintes, são introduzidas as particularidades de cada variável. Cada ficheiro inclui também um bloco [E5] (geralmente no início) que contém definições que se aplicam a todo o ficheiro de configuração. Neste exemplo específico, existe uma opção (*TABLE=*) que indica ao E5 qual a base de dados a utilizar. Se não for especificada nenhuma tabela, o E5 recorre à tabela '_default'. Como neste caso não é especificada uma tabela, o ficheiro da base de dados (um ficheiro JSON) terá o mesmo nome do ficheiro de configuração.

De seguida, há uma série de campos de entrada de dados (mais uma vez, cada um definido com [ ]). O primeiro campo é um *ID* de artefacto. A opção *TYPE* indica ao E5 o tipo de dados a aceitar, com opções válidas incluindo texto, nota, numérico, menus, booleano (Verdadeiro/Falso) e data e hora. O *PROMPT* é especificado com uma opção e, em seguida, a opção *UNIQUE* indica ao E5 que cada registo de dados deve ter um valor único para este campo. Qualquer tentativa de duplicar um valor para este campo resultará num aviso e, caso a introdução de dados prossiga, o registo anterior com este *ID* será editado ou substituído. O campo *ArtifactType* demonstra a utilização de menus. Os itens de menu são especificados na opção *MENU* e são separados por vírgulas. Não existe um limite para o número de itens de menu, e estes são apresentados pela ordem definida (excepto se a opção *SORTED* estiver configurada como *TRUE*). O campo *ToolType*, que se segue, também é um menu, mas demonstra a utilização de condicionantes. Durante a introdução de dados, o menu *ToolType* só é apresentado quando *ArtifactType* é igual a *Tool*, caso contrário, o E5 avança para o campo seguinte, inserindo uma célula vazia ("") no campo *ToolType*.

Do mesmo modo, *PlatformType* está condicionado ao facto de o *ArtifactType* ser *Tool* ou *Flake*. O campo seguinte, *PlatformWidth*, está sujeito a duas condições que devem ser simultaneamente verdadeiras, caso contrário, o campo será ignorado e será inserido um valor vazio na tabela da base de dados. A segunda condição exemplifica a utilização da palavra *NOT* nas condições. Quando o *PlatformType* é um valor diferente de *Missing*, esta condição é verdadeira.

Tanto o campo *PlatformWidth* como o último campo, *Weight*, são campos numéricos, o que implica que apenas números são aceites como entrada. Qualquer outro tipo de entrada resultará num erro, impedindo a continuação da introdução de dados.

Aula 04 - Exercício prático de E5

Nesta aula, realiza-se o primeiro dos três exercícios práticos da unidade curricular. O principal objetivo é que os alunos construam uma base de dados para a recolha de dados utilizando o software E5, aplicando os conceitos explorados nas aulas anteriores. Dá-se especial valor às opções relacionadas com a escolha do tipo de variáveis a registar, assim como as relações de interdependência entre as várias variáveis a serem analisadas e os respectivos atributos.

Enunciado do exercício nº 1 – Criação de uma base de dados em E5

Utilize os dados fornecidos na tabela abaixo para criar uma base de dados funcional no software E5. Os ficheiros finais deverão ser entregues em formato digital, por email, no final da aula. Será avaliado não só o bom funcionamento do ficheiro, mas também as escolhas efetuadas na nomeação das variáveis e atributos, e na definição das relações e condições aplicadas.

Observações:

1. Quando o campo dos exemplos de atributos a preencher oferece uma lista de vários atributos, significa que deverá criar um menu no ficheiro de configuração.
2. Todas as informações que não são fornecidas na tabela são subjetivas e ficarão ao critério de cada aluno.

Variável	**Exemplos/Lista de atributos a utilizar**	**Condições**
Nome do sítio arqueológico	Ex.: Vale Boi	-
Nome da área de escavação	Ex.: Area 2	-
Número de ID do artefacto	Ex.: G23-3456	-
Matéria-prima do artefacto	● Sílex ● Quartzo ● Quartzito ● Outra	-
Classe do artefacto	● Lasca inteira ● Fragmento de lasca ● Produto alongado inteiro ● Fragmento de produto alongado ● Núcleo ● Fragmento de núcleo ● Elemento de preparação e manutenção do núcleo ● Bigorna ● Percutor	-

	• Fragmento inclassificável • Esquírola	
Quantidade de córtex presente no artefacto	• 0% • 1-30% • 31-60% • 61-99% • 100%	-
Localização do córtex presente no artefacto	• Total • Proximal • Mesial • Distal	1. A classe do artefacto **não** pode ser *Núcleo*, *Fragmento de núcleo*, *Bigorna*, *Percutor*, *Fragmento inclassificável*, *Esquírola.* 2. O artefacto **não** pode ter 0% de córtex
Tipo de córtex presente no artefacto	• Seixo • Nódulo • Indeterminado	1. A classe do artefacto **não** pode ser *Esquírola*. 2. O artefacto **não** pode ter *0%* de córtex
Tipo de talão do artefacto	• Liso • Diedro • Facetado • Punctiforme • Cortical • Linear • Winged • Removido • Outro	1. A classe do artefacto **não** pode ser *Fragmento de lasca*, *Fragmento de produto alongado*, *Núcleo*, *Fragmento de núcleo*, *Bigorna*, *Percutor*, *Fragmento inclassificável*, *Esquírola*
Morfologia dos bordos do artefacto	• Convergentes • Divergentes • Biconvexos • Irregulares • Circulares • Outro	1. O artefacto **tem** de ser *Lasca inteira*, *Produto alongado inteiro*, *Elemento de preparação e manutenção do núcleo*
Perfil do artefacto	• Direito • Envurvado • Torcido • Outro	1. A classe do artefacto **tem** de ser *Lasca inteira*, *Produto alongado inteiro*, *Elemento de preparação e manutenção do núcleo*
Número de levantamentos anteriores presentes no artefacto	Ex.: 5	1. A classe do artefacto **tem** de ser *Lasca inteira*, *Fragmento de lasca*, *Produto alongado inteiro*, *Fragmento de produto alongado*, *Elemento de preparação e manutenção do núcleo* 2. O artefacto **não** pode ter *100%* de córtex
Direção dos levantamentos anteriores presentes no artefacto	• Unidirecional • Bidirecional • Cruzado	1. A classe do artefacto **tem** de ser *Lasca inteira*, *Fragmento de lasca*, *Produto*

	● Centrípeto ● Outro	*alongado inteiro*, *Fragmento de produto alongado*, *Elemento de preparação e manutenção do núcleo* 2. O artefacto **não** pode ter *100%* de córtex
Espessura do artefacto	Ex.: 12,34	1. A classe do artefacto **não** pode ser *Fragmento de núcleo*, *Fragmento inclassificável*, *Esquírola*
Largura do artefacto	Ex.: 23,34	1. A classe do artefacto **não** pode ser *Fragmento de lasca, Fragmento de produto alongado*, *Núcleo*, *Fragmento de núcleo*, *Fragmento inclassificável*, *Esquírola*
Comprimento do artefacto	Ex.: 50,56	1. A classe do artefacto **não** pode ser *Fragmento de lasca*, *Fragmento de produto alongado*, *Núcleo*, *Fragmento de núcleo*, *Fragmento inclassificável*, *Esquírola*
Peso do artefacto	Ex.: 100,23	-
Largura do talão do artefacto	Ex.: 10,69	1. A classe do artefacto **não** pode ser *Fragmento de lasca*, *Fragmento de produto alongado*, *Núcleo*, *Fragmento de núcleo*, *Fragmento inclassificável*, *Esquírola* 2. O tipo de talão do artefacto **não** pode ser *Removido*
Espessura do talão do artefacto	Ex.: 15.34	1. A classe do artefacto **não** pode ser F*ragmento de lasca*, *Fragmento de produto alongado*, *Núcleo*, *Fragmento de núcleo*, *Fragmento inclassificável*, *Esquírola* 2. O tipo de talão do artefacto **não** pode ser *Removido*
Ângulo exterior do talão do artefacto	Ex.: 30	1. A classe do artefacto **não** pode ser *Fragmento de lasca, Fragmento de produto alongado*, *Núcleo*, *Fragmento de núcleo*, *Fragmento inclassificável*, *Esquírola* 2. O tipo de talão do artefacto **não** pode ser

		Removido
Tipo de núcleo	• Simples com 1 plataforma • Prismático com 1 plataforma • Piramidal • Simples com 2 plataformas • Oposto • Ortogonal • Informe • Bipolar • Centrípeto • Globular • Outro	1. A classe do artefacto **tem** de ser *Núcleo*
Secção do núcleo	• Circular • Triangular • Quadrangular • Irregular	1. A classe do artefacto **tem** de ser *Núcleo*
Número de faces de debitagem do núcleo	• Uma • Duas • Três • Quatro ou mais	1. A classe do artefacto **tem** de ser *Núcleo* 2. O tipo de núcleo **não** pode ser *Informe*, *Centrípeto*, *Globular*, *Outro*
Tipo de plataforma do núcleo	• Lisa • Diedra • Facetada • Cortical • Esmagada • Outra	1. A classe do artefacto **tem** de ser *Núcleo* 2. O tipo de núcleo **não** pode ser *Informe*, *Centrípeto*, *Globular*, *Outro*
Tratamento distal do núcleo	• Cortical • Superfície de debitagem • Plano de percussão • Crista • Esmagado • Levantamento aleatório • Outro	1. A classe do artefacto **tem** de ser *Núcleo* 2. O tipo de núcleo **não** pode ser *Informe*, *Centrípeto*, *Globular*, *Outro* 3. O número de faces de debitagem do núcleo **não** pode ser *Quatro ou mais*
Tratamento direito do núcleo	• Cortical • Superfície de debitagem • Plano de percussão • Crista • Esmagado • Levantamento aleatório • Outro	1. A classe do artefacto **tem** de ser *Núcleo* 2. O tipo de núcleo **não** pode ser *Informe*, *Centrípeto*, *Globular*, *Outro* 3. O número de faces de debitagem do núcleo **não** pode ser *Quatro ou mais*
Tratamento posterior do núcleo	• Cortical • Superfície de debitagem • Plano de percussão • Crista • Esmagado • Levantamento aleatório • Outro	1. A classe do artefacto **tem** de ser *Núcleo* 2. O tipo de núcleo **não** pode ser *Informe*, *Centrípeto*, *Globular*, *Outro*

Razão do abandono do núcleo	● Ultrapassagem ● Ressalto ● Plataforma esmagada ● Imperfeição natural ● Fratura ● Perda de ângulo ● Sem razão	1. A classe do artefacto **tem** de ser *Núcleo*
Alterações na superfície do artefacto	● Nenhuma ● Pátina ● Concreções ● Fogo ● Misto	1. A classe do artefacto **não** pode ser *Esquírola*
Tipo de alteração por fogo	● Queimado ● Rubefacto ● Tratamento térmico	1. A alteração na superfície do artefacto **tem** de ser *Fogo*
Presença de retoque no artefacto	● Sim ● Não	1. A classe do artefacto **não** pode ser *Núcleo*, *Fragmento de núcleo*, *Bigorna*, *Percutor*, *Fragmento inclassificável*, *Esquírola*
Tipologia do utensílio retocado	Ex.: 5	1. A presença de retoque no artefacto **tem** de ser *Sim*
Quantidade de esquírolas	Ex.: 345	1. A classe do artefacto **tem** de ser *Esquírola*
Observações	Ex.: O padrão dorsal é bidirecional alternante	-

Aula 05 - Transformação de dados

Esta aula tem como objetivo dotar os alunos das competências necessárias para limpar e transformar os dados brutos, tornando-os adequados para análise. Pretende-se que os alunos desenvolvam não só as competências técnicas necessárias, mas também compreendam o impacto e a relevância de uma base de dados devidamente organizada. Será dada ênfase à utilização das diferentes ferramentas disponíveis nas folhas de cálculo para a transformação dos dados, como os operadores e as funções.

Após a fase de recolha de dados, seja no campo ou em laboratório, idealmente obteremos uma tabela ou um conjunto de tabelas onde a informação está registada de forma organizada e legível, em conformidade com os princípios e boas práticas da construção de bases de dados. Assim, em princípio, os dados estarão prontos para iniciar o processo de análise estatística e, se necessário, realizar algumas operações de **limpeza**, **transformação** e **filtragem**. Este processo pode ser efetuado em diferentes programas informáticos, destacando-se dois grupos distintos: software comercial com interface de apontar e clicar (*e.g.*, MS Excel, Google Sheets, SPSS, SAS, Stata, Minitab) e linguagens de programação adequadas para computação estatística (*e.g.*, R, Python). O modo dominante de interação com as ferramentas de análise de dados para muitos investigadores é com o primeiro grupo. Apesar de serem muito potentes, a forma de interação com esses software constitui um obstáculo à reprodutibilidade, pois os gestos do rato deixam poucos vestígios duradouros e acessíveis a outros investigadores. Por outro lado, quase todos esses programas exigem a aquisição de uma licença, o que também dificulta a partilha de ficheiros com quem não possui acesso a uma licença do software em questão. Neste contexto, o Google Sheets surge com uma das melhores opções, sendo uma plataforma gratuita, com funcionalidades quase idênticas às do MS Excel e a vantagem de funcionar através de armazenamento na nuvem, facilitando a partilha de ficheiros e a colaboração em tempo real. Por estas razões, os próximos exemplos relativos à transformação de dados serão realizados com recurso ao Google Sheets. Contudo, os mesmos procedimentos podem ser aplicados no MS Excel, LibreOffice Calc, ou outro software de folhas de cálculo, visto que as ferramentas são, na maioria dos casos, semelhantes.

A transformação de dados refere-se ao processo de alterar, organizar e limpar os dados, com o objetivo de extrair informações úteis e torná-los mais adequados para diversos fins. Ainda que seja esperado que a nossa base de dados esteja o mais limpa e organizada possível, é frequente termos que trabalhar com ficheiros que não cumprem com as boas práticas e, por conseguinte,

necessitam de várias operações de formatação antes de poderem ser analisados estatisticamente. No Google Sheets, a transformação de dados pode ser realizada utilizando diversos recursos e funções, que se passam a detalhar.

Formatar, ordenar e filtrar

As três operações básicas de qualquer programa de folhas de cálculo são formatar, ordenar e filtrar. Quando se importa um ficheiro para o Google Sheets, devemos não só verificar se cada coluna corresponde a uma variável devidamente nomeada na primeira linha da tabela e se cada linha corresponde a uma entrada, mas também se o tipo de dados para cada coluna é o correto. Embora o programa identifique automaticamente os diferentes tipos de dados, é crucial verificar e selecionar o formato de dados correto para cada coluna no Google Sheets. Esta seleção é essencial para uma representação precisa dos dados e uma transformação adequada dos mesmos. Para verificar o formato, seleciona-se a coluna inteira ou o intervalo de células desejado. No menu *Format,* em *Number* pode-se selecionar o formato melhor adequado ao tipo de dados presentes naquela coluna (*e.g.*, Number, Plain text, Currency, Date).

A ordenação de dados em folhas de cálculo é fundamental para organizar as informações numa ordem específica com base em critérios selecionados. Para ordenar os dados, primeiro seleciona-se o intervalo de células ou a coluna a ser ordenada. No menu *Data*, pode-se selecionar *Sort sheet A-Z* para uma ordem ascendente ou *Sort sheet Z-A* para uma ordem decrescente. Se a variável em causa for um campo de texto, as células serão organizadas alfabeticamente. Adicionalmente, pode-se utilizar a opção *Advanced range sorting option* para ordenar com base em várias colunas, permitindo adaptar a ordenação às necessidades específicas.

A filtragem de dados é outra funcionalidade importante que ajuda a extrair informações relevantes de grandes conjuntos de dados. Para aplicar filtros, seleciona-se a coluna relevante e, no menu *Data* seleciona-se *Create a filter*. As opções de filtro surgirão no nome de cada coluna, permitindo exibir ou ocultar linhas que correspondem a critérios específicos.

Usar operadores e funções para transformação de dados

Existem duas formas básicas de transformar os dados: através da utilização de operadores ou de funções. Os operadores são sinais especiais que efectuam, normalmente, alterações simples. As funções, mais parecidas a fórmulas, frequentemente processam vários dados para obter um resultado e podem ser bastante poderosas e complexas, dependendo do software utilizado. Em seguida, apresentam-se os vários operadores e funções, consoante os formatos de dados anteriormente mencionados, uma vez que tanto operadores como funções podem variar de acordo com o formato de dados. É importante notar que, ao se trabalhar com uma base de dados, normalmente não se alteram os valores contidos numa coluna. Em vez disso, coloca-se o resultado da sua transformação numa nova coluna. A principal razão para não substituir ou alterar os valores de uma variável, pelo menos inicialmente, é que, se isso for feito incorretamente, pode ser difícil ou impossível repor os valores originais. Utilizando o exemplo dado por Mcpherron e Dibble (2002), suponhamos que temos uma variável chamada "Cor", com os valores "Azul", "Vermelho" ou "Amarelo". Pretendemos alterar todos os “Vermelho” para “Azul”, mas acidentalmente alteramos todos os “Amarelo” para “Azul”. Agora, não temos forma de saber quais os valores “Azul” que têm de ser revertidos para “Amarelo”. Por isso, é sempre preferível colocar os resultados numa nova coluna, que pode ser comparada lado a lado com a coluna original. Quando estivermos convencidos de que a nova coluna contém os resultados pretendidos, podemos eliminar a coluna original e alterar o nome da nova coluna para o da original ou, preferencialmente, manter ambas as colunas para que outros utilizadores possam ter acesso às transformações efetuadas.

Dados em formato numérico

Os operadores básicos que podem ser utilizados com uma variável numérica refletem as várias operações aritméticas com as quais estamos familiarizados: adição (+), subtração (-), multiplicação (*), divisão (/), elevação a uma potência (^), entre outras. O conceito básico é exatamente o mesmo que na álgebra. Por exemplo, se VAR1 e VAR2 são duas colunas que contêm valores numéricos, é possível realizar operações como as seguintes:

```
VAR3 = VAR1 + VAR2 - soma os dois valores

VAR3 = VAR1 - VAR2 - subtrai o valor de VAR2 do valor de VAR1

VAR3 = VAR1 * VAR2 - multiplica os dois valores

VAR3 = VAR1 / VAR2 - divide VAR1 por VAR2
```

```
VAR3 = VAR1 ^ 2 - calcula o quadrado de VAR1
```

Numa folha de cálculo este tipo de operação é muito simples de efetuar. Basta, na célula vazia onde queremos o resultado do cálculo, introduzir o símbolo de igual (=), seguido da referência espacial da célula onde se encontra o primeiro valor, o operador matemático, e a referência espacial da célula onde se encontra o segundo valor. É importante notar que a referência espacial é automaticamente dada pelo programa assim que se seleciona uma célula. Imaginemos que, com base nos valores das variáveis "Comprimento" e "Largura" de um conjunto de sítios arqueológicos, pretendemos calcular a "Área" dos mesmos. A Figura 11 ilustra como realizar este cálculo numa folha de cálculo.

	A	B	C
1	Comprimento	Largura	Area
2	0.8151457494	0.6996722407	=A2*B2
3	0.1307160952	0.9713894865	
4	0.7866764581	0.6484524091	
5	0.2738134445	0.5087544898	
6	0.6267559509	0.8686860803	
7	0.5265242372	0.9528427109	
8	0.7895443198	0.06038376018	
9	0.5864412878	0.02382212205	
10	0.303072952	0.6630649756	
11	0.701625985	0.4198544792	
12	0.1461669573	0.2293887895	
13	0.135689811	0.8059862794	
14	0.2043812627	0.5049647801	
15	0.6730313141	0.4928962306	
16	0.9587372728	0.4447629881	
17	0.5063104742	0.3706517918	
18	0.05648840621	0.7972867969	

Figura 11. Exemplo do uso de operador aritmético para calcular a Área com base nos valores de Comprimento e Largura. As referências espaciais das células, neste caso, são o A2 (valor para o Comprimento) e B2 (valor para a Largura).

Também podemos efetuar estas operações aritméticas não só com os valores contidos nas células, mas também com quaisquer parâmetros numéricos ou constantes. Por exemplo, para converter um valor de polegadas para centímetros, faríamos o seguinte:

```
Comprimento_cms = Comprimento_poleg * 2,54
```

Estas expressões algébricas (ou fórmulas ou equações) podem tornar-se muito complexas com a utilização combinada de vários operadores diferentes. Como demonstrado por Mcpherron &

Dibble (2002), é importante ter em conta que existe uma prioridade inerente a certas operações: os expoentes são calculados em primeiro lugar, seguidas da divisão e da multiplicação, e, finalmente, da adição e da subtração. Assim, na expressão:

```
VAR3 = 3+4/2^2
```

o resultado é 4. Isto ocorre porque a primeira operação efetuada é elevar 2 ao quadrado (resultando em 4), que é depois dividido por 4 (dando 1), que é depois adicionado a 3. É possível alterar esta ordem inerente através da utilização de parênteses, visto que qualquer expressão entre parênteses é calculada antes de se efetuarem outras operações. Assim, na seguinte expressão;

```
VAR3 = ((3+4)/2)^2
```

o resultado é 12,25. A primeira operação efetuada é a adição de 3 e 4 (resultando em 7); de seguida, o valor é dividido por 2 e o resultado é elevado ao quadrado. Obviamente, é fundamental ter extremo cuidado com este tipo de operações complexas, sendo sempre boa ideia verificar algumas respostas manualmente para assegurar que se estão a obter os resultados pretendidos. Note-se também que um erro comum quando se introduzem fórmulas complexas com parênteses é omitir um dos pares de parênteses. Assim, se o computador devolver um erro, deve-se sempre verificar que, para cada parêntesis que se abre, existe um parêntesis fechado correspondente.

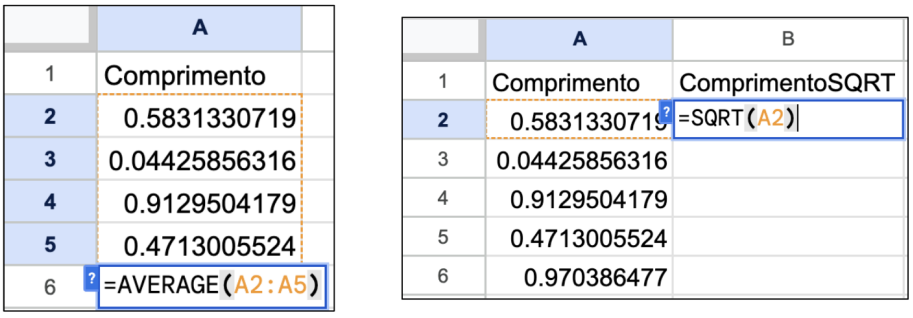
As funções são semelhantes aos operadores, mas, enquanto estes combinam valores para produzir um novo valor, as funções transformam ou atuam tipicamente sobre um determinado valor. As funções têm também um formato diferente, com o nome da função seguido de um conjunto de parêntesis que contém o argumento numérico (ou valor) sobre o qual a função atua. Se houver mais do que um argumento, estes são separados por vírgulas ou ponto e vírgula.

Nas folhas de cálculo, as funções são fórmulas pré-construídas concebidas para efetuar cálculos específicos, transformação ou tarefas de análise de dados. Estas funções são ferramentas poderosas que poupam tempo e esforço através da automatização de cálculos complexos. Seguem uma sintaxe específica, que não é muito distinta da dos operadores, iniciando-se com o símbolo de igual (=), seguido do nome da função, e, entre parêntesis, os valores que serão usados no cálculo. Algumas das funções mais comumente utilizadas com variáveis numéricas são as apresentadas na Tabela 3.

Função	Descrição
SUM	Calcula a soma de um intervalo de células.
AVERAGE	Calcula a média de um intervalo de células.
COUNT	Conta o número de células com valores numéricos num intervalo.
MIN/MAX	Encontra o valor mínimo ou máximo num intervalo.
SQRT	Calcula a raíz quadrada de um valor.
INT	Arredonda um número para o número inteiro mais próximo.
ROUND	Arredonda um número para um número especificado de casas decimais.
ABS	Devolve o valor absoluto de um número.
COS	Devolve o coseno de um ângulo em radianos.
SIN	Devolve o seno de um ângulo em radianos.
ATAN	Devolve a tangente inversa de um valor, em radianos.
TEXT	Formata um número em texto.

Tabela 3. Principais funções de uma folha de cálculo para transformação de dados em formato numérico.

No caso das funções que calculam um novo valor com base num intervalo de células, estas podem ser utilizadas para efetuar o cálculo ao longo de uma linha ou de uma coluna. A Figura 12 demonstra a utilização das funções AVERAGE e SQRT para obter a média e o desvio padrão numa folha de cálculo.



	A
1	Comprimento
2	0.5831330719
3	0.04425856316
4	0.9129504179
5	0.4713005524
6	=AVERAGE(A2:A5)

	A	B
1	Comprimento	ComprimentoSQRT
2	0.5831330719	=SQRT(A2)
3	0.04425856316	
4	0.9129504179	
5	0.4713005524	
6	0.970386477	

Figura 12. Exemplo de utilização das funções AVERAGE e SQRT numa folha de cálculo para obter a média dos valores de uma coluna, e a raíz quadrada de um valor.

Podemos também combinar funções numa expressão, tal como combinámos operadores anteriormente. Por exemplo, se quisermos calcular a raiz quadrada da diferença entre dois números. Por vezes, a diferença entre dois números é negativa, e a maioria dos sistemas informáticos não permite calcular a raiz quadrada de um número negativo. Para evitar este problema, pode-se calcular o valor absoluto (*i.e.*, o valor positivo) da diferença antes de obter a raiz quadrada:

```
VAR3 = SQRT(ABS(VAR2-VAR1))
```

Dados em formato de texto

Tipicamente, apenas um operador está disponível para trabalhar com texto, o operador de concatenação (+ ou &). A concatenação combina dois valores de caracteres juntando-os simplesmente um ao outro. Por exemplo, "Paleolítico" & "Médio" resulta em "PaleolíticoMédio". Para formatar a expressão com um espaço no meio, podemos fazer o seguinte:

```
"Paleolítico" & " " & "Médio"
```

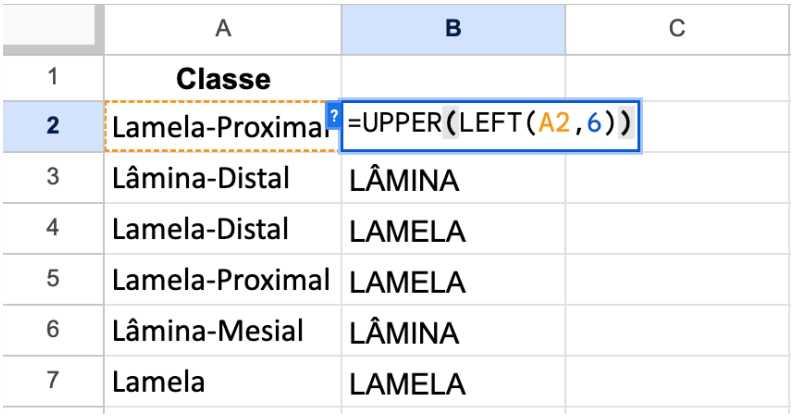
o que resulta em "Paleolítico Médio".

Em contrapartida, existem muitas funções úteis para o tratamento de texto. As mais comuns encontram-se listadas na Tabela 4.

Função	Descrição
LEN	Devolve o comprimento de uma cadeia de texto (número de caracteres).
LEFT	Extrai um número especificado de caracteres do início de uma cadeia de texto.
RIGHT	Extrai um número especificado de caracteres do fim de uma cadeia de texto.
MID	Extrai um número específico de caracteres de uma cadeia de texto, começando numa posição especificada.
UPPER	Converte todos os caracteres de uma cadeia de texto em maiúsculas.
TRIM	Remove quaisquer espaços à esquerda ou à direita de uma cadeia de texto.
PROPER	Converte a primeira letra de cada palavra numa cadeia de texto para maiúsculas.
SUBSTITUTE	Substitui as ocorrências de um texto específico numa cadeia de texto por outro texto.

Tabela 4. Principais funções de uma folha de cálculo para transformação de dados em formato de texto.

Também é possível combinar duas ou mais funções para alterar ou criar um novo valor. A Figura 13 demonstra como podemos extrair um conjunto de caracteres de uma célula e convertê-los para maiúsculas, através da combinação das funções LEFT e UPPER do Google Sheets.



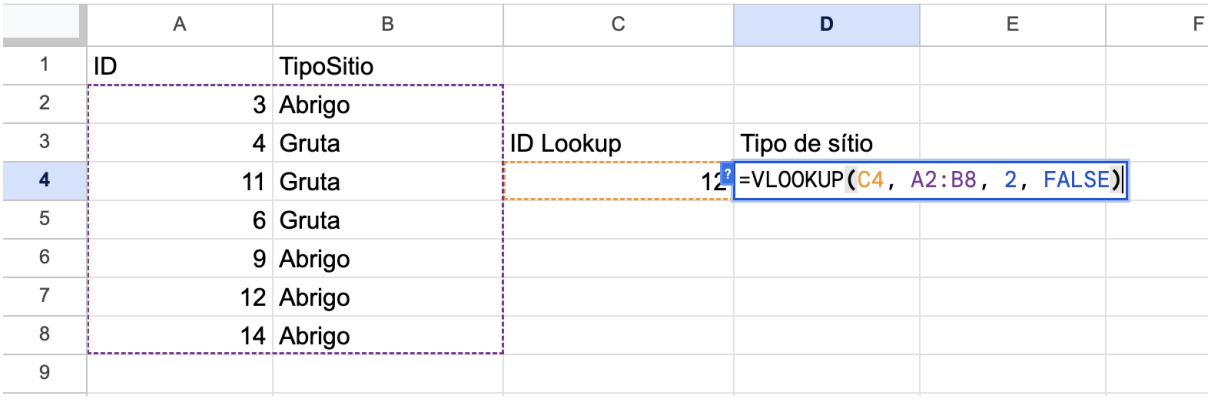
	A	B	C
1	**Classe**		
2	Lamela-Proximal	=UPPER(LEFT(A2,6))	
3	Lâmina-Distal	LÂMINA	
4	Lamela-Distal	LAMELA	
5	Lamela-Proximal	LAMELA	
6	Lâmina-Mesial	LÂMINA	
7	Lamela	LAMELA	

Figura 13. Exemplo de utilização das funções UPPER e LEFT numa folha de cálculo para obter apenas o tipo de classe de artefacto lítico (i.e., lâmina, lamela) sem indicação da porção analisada (i.e., proximal, distal, mesial).

Funções de lógica e de pesquisa

Por fim, resta-nos abordar um conjunto de funções aplicáveis à maior parte dos tipos de dados e que, apesar da sua complexidade, são de grande relevância.

As funções de pesquisa permitem obter dados com base em critérios específicos. Duas das mais utilizadas são as funções VLOOKUP e HLOOKUP. A função VLOOKUP procura um valor na coluna mais à esquerda de um intervalo especificado, retornando um valor correspondente de uma coluna diferente, na mesma linha. É normalmente usada para pesquisas verticais em grandes conjuntos de dados (Figura 14). A função HLOOKUP, por sua vez, funciona de forma semelhante à VLOOKUP, mas pesquisa horizontalmente na primeira linha de um intervalo especificado e devolve um valor correspondente de uma linha diferente. O exemplo abaixo demonstra como utilizar a função VLOOKUP para procurar o tipo de sítio arqueológico com base no ID atribuído. Esta função revela-se extremamente útil, por exemplo, quando queremos combinar dados de duas tabelas que partilham uma chave primária comum.



	A	B	C	D	E	F
1	ID	TipoSitio				
2	3	Abrigo				
3	4	Gruta	ID Lookup	Tipo de sítio		
4	11	Gruta	12	=VLOOKUP(C4, A2:B8, 2, FALSE)		
5	6	Gruta				
6	9	Abrigo				
7	12	Abrigo				
8	14	Abrigo				
9						

Figura 14. Exemplo de utilização da função VLOOKUP. Esta função requer a seguinte ordem de atributos para funcionar: search_key (o valor a procurar, que no exemplo é a célula C4), range (o intervalo a ser considerado para a pesquisa, em que a primeira coluna do intervalo é pesquisada para encontrar a search_key sendo no exemplo as células entre A2 e B8), index (a posição da coluna de onde se extrai o valor, neste caso, trata-se da segunda coluna do intervalo selecionado), is_sorted (opcional - indica se a coluna a ser pesquisada - a primeira coluna do intervalo especificado - está ordenada, caso em que será devolvida a correspondência mais próxima para a search_key).

As funções de lógica possibilitam a tomada de decisões baseadas em condições, permitindo um processamento de dados mais avançado. Uma das funções mais úteis é a função IF. Esta avalia uma condição e retorna valores diferentes conforme a condição seja verdadeira ou falsa. Possui três argumentos essenciais para o seu funcionamento correto:

```
=IF(expressão_lógica, [valor_se_verdadeiro], [valor_se_falso])
```

No exemplo a seguir, usa-se a função IF para determinar, com base nos valores de comprimento e largura de utensílios líticos, o tipo de suporte a ser preenchido numa nova coluna. A lógica é que se for verdade que o comprimento é superior ou igual a duas vezes a largura, a célula será preenchida com a opção "Alongado", se, pelo contrário a expressão lógica não se verificar, então a célula será preenchida com a opção "Lasca" (Figura 15).

	A	B	C	D	E
1	Comprimento	Largura	Tipo		
2	27.47	11.99	Alongado		
3	30.65	15.84	Lasca		
4	31.49	38.77	Lasca		
5	20.23	11.62	Lasca		
6	25.39	20.38	Lasca		
7	37.41	22.07	Lasca		
8	44.47	26.79	=IF(A8>=2*B8,"Alongado", "Lasca")		
9	26.48	18.27			
10	24.51	14.17			
11	35.55	13.08			
12	13.51	16.15			
13	18.53	20.34			
14	25.51	10.52			
15	33.19	14.99			
16	27.94	12.26			
17	32.71	16.17			

Figura 15. Exemplo da utilização da função IF numa folha de cálculo.

Aula 06 - Exercício prático de transformação de dados

Nesta aula, realiza-se o segundo dos três exercícios práticos da unidade curricular. O objetivo principal é que os alunos apliquem os diferentes operadores e funções disponíveis nas folhas de cálculo para a correção e transformação de dados, bem como para a criação de novas variáveis. Valorizam-se as escolhas efetuadas na seleção das funções e também as decisões relacionadas com a lógica das sequências de transformação dos dados.

Enunciado do exercício nº 2 – Formatação de dados através de funções de folhas de cálculo

A tabela fornecida (Exercicio2.csv) contém informação fictícia relativa a uma série de datações por radiocarbono para um conjunto de sítios atribuíveis aos tecnocomplexos do Paleolítico superior, nomeadamente o Gravetense e o Solutrense.

O objetivo deste exercício é formatar os elementos da tabela de modo a torná-la mais organizada e fácil de consultar. Para tal, deverão utilizar as funções de folhas de cálculo aprendidas na aula anterior (*e.g.*, CONCATENAR, etc.) ou outras que considerem mais adequadas para o cálculo das novas colunas.

No final, deverão obter uma série de novas colunas semelhantes a estas:

ID	Site	Date	Techno-complex	Phase	Latitud in degrees, minutes, seconds	Longitud in degrees, minutes, seconds
IADA01234	HJKI	23456±150	Gravettian	Final	23’45’’10.68	-2° 2' 30.01"
IADA00079	POIY	20567±300	Solutrean	Lower	24’34’’45.67	-4° 7' 8"

As transformações a realizar são as seguintes:

1. Acrescentar o prefixo “IADA” ao número de ID, sendo que este campo (*i.e.*, IADA+ID) terá sempre 9 caracteres.

2. Juntar os campos Date e Standard Deviation numa única célula, utilizando o símbolo ± como separador.
3. Separar em duas colunas distintas o tecnocomplexo (Solutrean ou Gravettian) e a respetiva fase (Final, Upper, etc.)
4. Converter os valores de Longitude e Latitude em coordenadas de graus, minutos e segundos. Para esta conversão, devem seguir o exemplo abaixo:

Converter 30.263888889 para graus (g),minutos (m),segundos (s):

g = integer(30.263888889°) = 30°

m = integer((graus decimais - graus) × 60) = 15'

s = (graus decimais - graus - minutos/60) × 3600 = 50"

ATENÇÃO: As colunas devem permanecer com as fórmulas inseridas e não apenas com os valores calculados. Caso contrário, não será possível confirmar que funções foram utilizadas para a transformação.

Aula 07 - Estatística descritiva univariada (parte 1)

A sétima aula marca o início do terceiro grande bloco do programa, dedicado à análise quantitativa em Arqueologia. Nesta aula, exploram-se os conceitos e ferramentas principais para a caracterização e apresentação de variáveis isoladas. O objetivo é que os alunos se familiarizem com os diferentes tipos de frequência, assim como com as várias opções disponíveis para a representação de variáveis qualitativas.

Tipos de variáveis estatísticas

Como ficou anteriormente evidente, os dados arqueológicos apresentam-se sob várias formas quando registados numa base de dados. Em linguagem de base de dados, a divisão dos vários tipos está principalmente relacionada com a forma como a informação será armazenada. No entanto, no momento de analisar essas mesmas variáveis através de métodos aritméticos e estatísticos, a sua classificação obedece a princípios ligeiramente diferentes. Assim, do ponto de vista estatístico, podemos classificar os dados como:

- **Dados qualitativos (ou não numéricos, ou categóricos)**. Captam informação categórica sem valores numéricos, expressos frequentemente através de etiquetas, códigos ou categorias. Exemplos incluem tipos de artefactos, práticas de enterramento, tipos de povoamento e estilos de cerâmica. Além disso, os dados qualitativos podem ser classificados com base nas suas características:
    - **Dados nominais**. Representam categorias distintas sem qualquer ordem ou classificação inerente. Exemplos incluem tipos de artefactos (*e.g.*, estilos de cerâmica, tipos de ferramentas líticas), classificações de sítios (*e.g.*, gruta, abrigo, cidade) e categorizações de género (*e.g.*, masculino, feminino).
    - **Dados ordinais.** Têm uma ordem inerente, mas os intervalos entre os valores não são uniformes. Exemplos incluem a classificação de bens de prestígio (*e.g.*, estatuto alto, médio, baixo), camadas estratigráficas (*e.g.*, superior, médio, inferior) e níveis de hierarquia social.
- **Dados quantitativos (ou numéricos).** Envolvem valores numéricos e podem ser medidos utilizando unidades padronizadas. Estes dados permitem análises matemáticas e comparações precisas. Exemplos incluem contagens de artefactos, medidas (comprimento, peso, etc.) e coordenadas de sítios arqueológicos. Os dados quantitativos podem ser classificados com base nas suas características:

- **Dados contínuos.** Medidas que podem assumir qualquer valor dentro de um determinado intervalo, frequentemente representados numa escala contínua. Exemplos em arqueologia incluem pesos de artefactos, profundidades de camadas arqueológicas e distâncias entre sítios.
- **Dados discretos.** Valores contados que são distintos e separados. Surgem frequentemente da contagem de ocorrências ou da enumeração de objetos. Exemplos incluem contagens de artefactos, o número de enterramentos num sítio e o número de tipos específicos de artefactos.

Por outro lado, as variáveis podem ser organizadas segundo a escala de medida. Esta refere-se à natureza e às características dos dados que ditam as operações matemáticas que podem ser efectuadas sobre eles. Existem quatro escalas de medida principais:

- **Escala Nominal.** Os dados na escala nominal são categóricos e não apresentam qualquer ordem ou classificação inerente. Exemplos incluem género, grupo étnico e tipos de artefactos. Os dados nominais permitem o agrupamento e classificação, mas não se prestam a operações matemáticas significativas, como a adição ou a multiplicação.
- **Escala ordinal.** Os dados ordinais têm uma ordem ou classificação inerente entre as categorias, mas os intervalos entre as categorias não são uniformes. Exemplos incluem respostas a inquéritos (por exemplo, concordo totalmente, concordo, neutro, discordo, discordo totalmente) e camadas estratigráficas. Estes dados permitem comparações relativas, mas não medições precisas devido aos intervalos irregulares.
- **Escala intervalar.** Os dados de escala intervalar possuem intervalos consistentes entre valores, mas não têm um verdadeiro ponto zero. Um exemplo é a temperatura medida em Celsius ou Fahrenheit. Em arqueologia, estas escalas de intervalo podem ser aplicadas para representar sequências cronológicas de estilos de artefactos ou a distribuição de percentagens na composição dos artefactos.
- **Escala de razão.** Os dados da escala de razão caracterizam-se não só pelos intervalos uniformes, mas também pela presença de um verdadeiro ponto zero. Incluem, por exemplo, contagens de artefactos, pesos de artefactos ou distâncias entre sítios arqueológicos. As escalas de razão permitem o cálculo de proporções e a realização de várias operações matemáticas, como a adição, a subtração, a multiplicação e a divisão.

O tipo de dados corresponde diretamente à escala de medida utilizada. Por exemplo, os dados qualitativos alinham-se frequentemente com escalas nominais ou ordinais, enquanto os dados quantitativos alinham-se tipicamente com escalas de intervalos ou de razão.

Análise univariada

Uma vez elaborada uma ou várias tabelas de dados, toda a informação encontra-se disponível, mas os padrões que caracterizam aquele conjunto ainda podem não estar totalmente claros para nós. Normalmente, o nosso interesse não reside nas características de cada peça individual, mas sim no conjunto do material como um todo. Quando fazemos perguntas como "Quão comuns são os diferentes tipos de matérias-primas numa coleção de artefactos líticos?" ou "As lascas dessa coleção têm um tamanho normalizado?", as respostas não estão imediatamente disponíveis na base de dados. É necessário resumir os dados (os valores das variáveis) de alguma forma, seja através de tabelas, gráficos ou de números que resumam a tendência, considerando sempre as características de medição das variáveis.

Variáveis qualitativas

Tabelas de frequência

As **tabelas de frequência** constituem um excelente ponto de partida para resumir dados. Estas são particularmente úteis para apresentar dados qualitativos, mostrando a frequência com que cada valor ocorre no conjunto.

A representação de dados qualitativos numa tabela de frequências é relativamente simples, dado que as categorias estão claramente definidas. Por exemplo, se quisermos apresentar a frequência das várias matérias-primas numa coleção de ferramentas em pedra, podemos facilmente criar uma tabela como a Tabela 5.

Matéria-prima	Frequência absoluta
Grauvaque	20
Outra	7
Quartzo	29
Sílex	138
Xisto	1
Total	195

Tabela 5. Exemplo de uma tabela de frequência absoluta (N) contabilizando as matérias-primas presentes numa coleção de ferramentas em pedra.

Este tipo de frequência denomina-se de **frequência absoluta**, uma vez que representa o número real de itens analisados para cada categoria. Consiste na contagem bruta, sem considerar a dimensão global do conjunto de dados ou o contexto.

Por outro lado, as **frequências relativas** têm em conta a proporção ou percentagem de ocorrências de um atributo específico em relação ao número total de ocorrências ou a um subconjunto específico. Proporcionam uma perspectiva sobre a importância de um atributo no contexto do conjunto completo de dados. Para calcular a frequência relativa, divide-se a frequência absoluta de um atributo específico pelo número total de ocorrências no conjunto de dados ou subconjunto. Esta abordagem permite comparar a importância de diferentes atributos em diferentes conjuntos, independentemente dos tamanhos destes. As frequências relativas podem ser expressas em frações, percentagens ou números decimais (Tabela 6).

Matéria-prima	**Frequência absoluta**	**Frequência relativa**
Grauvaque	20	10.26%
Outra	7	3.59%
Quartzo	29	14.87%
Sílex	138	70.77%
Xisto	1	0.51%
Total	195	100.00%

Tabela 6. Exemplo de uma tabela de frequências absoluta (N) relativa (%), contabilizando as matérias-primas presentes numa coleção de ferramentas em pedra.

Finalmente, existe um terceiro tipo de frequência comumente utilizado em Arqueologia: as **frequências relativas acumuladas**. Consiste na soma das frequências relativas anteriores. Para encontrar as frequências relativas acumuladas, somam-se todas as frequências relativas anteriores à frequência relativa da linha atual, como ilustrado na Tabela 7.

A criação destes e de outros tipos de tabelas num software de folhas de cálculo é relativamente simples, graças à ferramenta de **tabelas dinâmicas**, ou *Pivot Tables*. As tabelas dinâmicas oferecem uma forma eficaz de resumir os dados na folha de cálculo, agregando, ordenando, contando e calculando automaticamente os indicadores desejados, enquanto apresentam os resultados resumidos numa nova tabela. Uma tabela dinâmica funciona como uma espécie de consulta a uma base de dados.

Matéria-prima	Frequência absoluta	Frequência relativa	Frequência relativa acumulada
Grauvaque	20	10.26%	10.26%
Outra	7	3.59%	13.85%
Quartzo	29	14.87%	28.72%
Sílex	138	70.77%	99.49%
Xisto	1	0.51%	100.00%
Total	195	100.00%	

Tabela 7. Exemplo de uma tabela de frequências absoluta (N) relativa (%), e relativa acumulada, contabilizando as matérias-primas presentes numa coleção de ferramentas em pedra.

Estas tabelas são compostas por colunas, linhas, páginas e campos de dados que podem ser reorganizados, facilitando o isolamento, agrupamento, expansão e soma dos dados em tempo real.

As tabelas representam uma boa forma de organizar e apresentar dados. Contudo, os gráficos podem ser ainda mais úteis para a compreensão dos dados. Não existem regras rígidas quanto aos tipos gráficos a utilizar. Dois exemplos de gráficos frequentemente usados para apresentar dados qualitativos são os **gráficos circulares** e os **gráficos de barras**.

Gráficos ou diagramas circulares

Num gráfico circular, as categorias de dados são representadas por secções (cunhas) num círculo, sendo proporcionais em tamanho à percentagem (frequência relativa) de indivíduos em cada categoria. Utilizando o exemplo mencionado nas tabelas anteriores, a figura abaixo poderia representar os mesmos dados através de um gráfico circular (Figura 16).

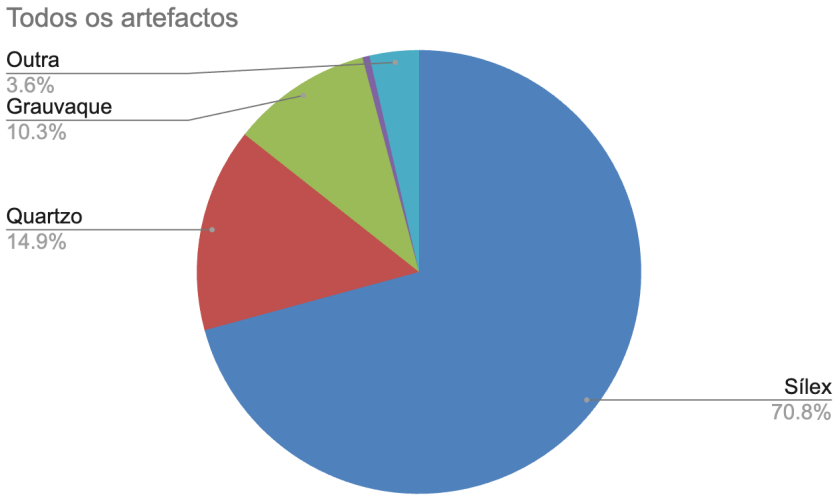


Figura 16. Diagrama circular representando frequência de matérias-primas numa coleção de ferramentas em pedra.

Gráficos ou diagramas de barras

Os gráficos de barras são compostos por barras individuais, separadas umas das outras. O comprimento ou altura de cada barra para cada categoria é proporcional ao número ou à percentagem de indivíduos nessa categoria. As barras podem ser retângulos ou caixas retangulares, utilizadas em gráficos tridimensionais, e podem ser dispostas vertical ou horizontalmente. Nos gráficos de barras apresentados nas Figuras 17 e 18, os tipos de matérias-primas estão representados no eixo dos X e as proporções no eixo dos Y.

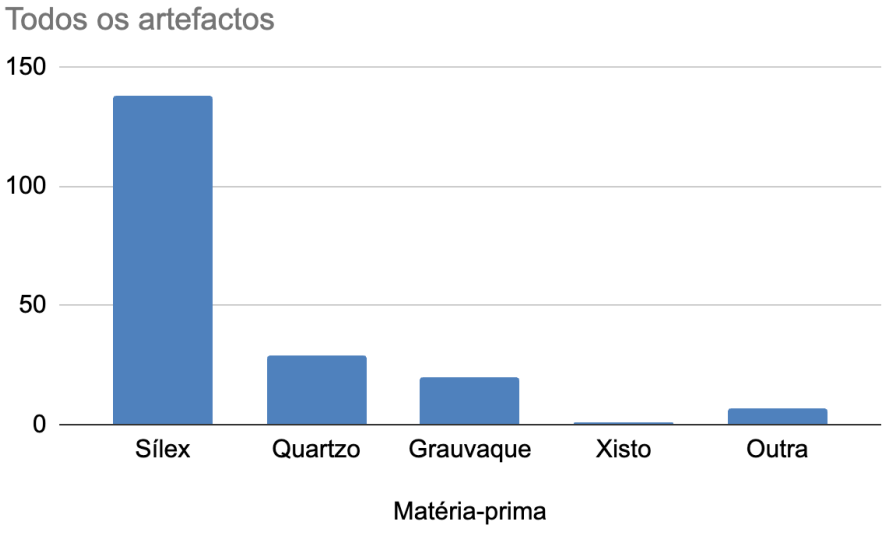


Figura 17. Gráfico de barras representando frequência absoluta de matérias-primas numa coleção de ferramentas em pedra.

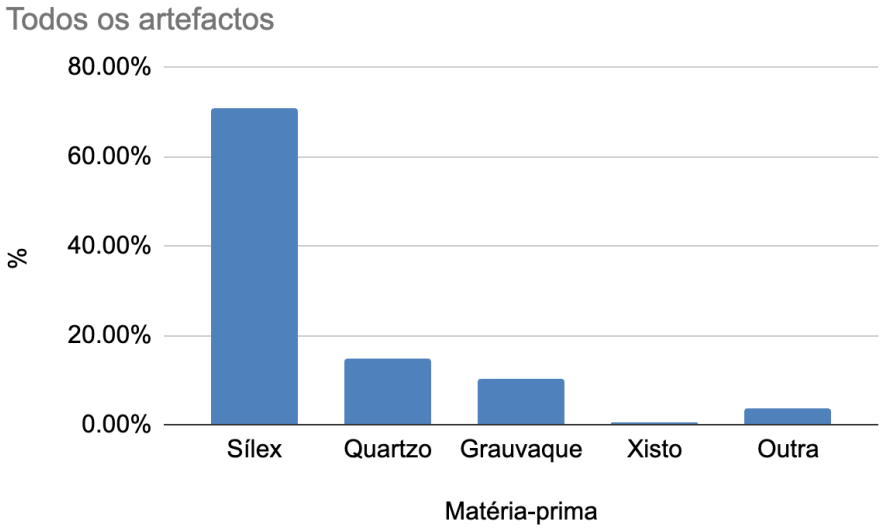


Figura 18. Gráfico de barras representando frequência relativa (em percentagem) de matérias-primas numa coleção de ferramentas em pedra.

Aula 08 - Estatística descritiva univariada (parte 2)

Nesta aula, abordam-se os principais conceitos para a caracterização e apresentação de variáveis quantitativas. Trabalham-se as medidas de tendência central e as principais medidas de dispersão, destacando-se também a importância da forma de distribuição, bem como o significado e a deteção de outliers. Os vários conceitos são sempre apresentados em conjunto com os elementos gráficos e tabulares que podem ser utilizados.

Variáveis quantitativas

As opções descritivas para dados quantitativos são significativamente mais robustas do que para dados categóricos. A principal razão para isto é que os dados quantitativos podem ser submetidos a um vasto leque de operações matemáticas e análises estatísticas, permitindo uma visão mais profunda dos padrões, relações e tendências. Por outro lado, os dados categóricos, sendo não numéricos, são mais limitados em termos das operações matemáticas que podem ser realizadas. Ao descrever uma variável quantitativa, é importante observar pelo menos quatro aspectos: as **medidas de tendência central**, as **medidas de dispersão**, a **forma da distribuição** e a presença de **valores anómalos ou *outliers***. Para cada um destes aspectos, existem elementos gráficos e tabulares úteis que facilitam a análise e interpretação da distribuição de uma variável numérica.

Medidas de tendência central

As medidas de tendência central utilizam um único valor para representar um conjunto de dados. Estas medidas fornecem um resumo conciso da distribuição e destacam o valor típico que reflete a essência dos dados. Em Arqueologia, as medidas de tendência central são fundamentais para discernir padrões, efetuar comparações e tirar conclusões acerca de um conjunto de dados.

Três medidas de tendência central são principalmente utilizadas:

- **Média**. A média é calculada somando todos os valores de um conjunto de dados e dividindo pelo número de observações. É particularmente útil quando os dados seguem uma distribuição normal (ou simétrica), sendo sensível a valores extremos. Esta estatística só tem significado para variáveis quantitativas. A fórmula para calcular a média é:

$$\overline{x} = \frac{\Sigma x}{n}$$

A média é frequentemente a medida preferida de tendência central para distribuições de dados arqueológicos, tanto por razões lógicas como matemáticas. Do ponto de vista lógico, a média de uma distribuição representa o valor mais provável se retirado aleatoriamente de uma população. Tecnicamente, a média representa o centróide ou um valor numérico que minimiza a soma das diferenças entre si e todos os pontos de dados individuais.

- **Mediana.** A mediana é o valor intermédio num conjunto de dados organizados por ordem ascendente ou decrescente. Menos afetada por valores extremos, fornece uma estimativa robusta da tendência central. Para calcular a mediana, os dados são organizados em ordem crescente ou decrescente e encontra-se o valor central, localizado no meio da distribuição, de modo que 50% das observações sejam superiores à mediana e 50% sejam inferiores à mediana. Se o conjunto de dados tiver um número par de observações, a mediana corresponde à média dos dois valores centrais. A mediana não é influenciada por valores extremos e o seu cálculo exige que as variáveis sejam medidas numa escala pelo menos ordinal.

  Mas por que que escolheríamos olhar para a mediana de uma distribuição em vez da média? A resposta a esta pergunta está relacionada com situações em que as nossas distribuições de dados são enviesadas ou assimétricas. Distribuições enviesadas criam circunstâncias em que a média pode ser fortemente influenciada por valores extremos em qualquer uma das extremidades da distribuição. Assim, para distribuições enviesadas, a média e a mediana estão frequentemente desalinhadas, e a média muitas vezes não representa adequadamente o centro da distribuição.

- **Moda.** A moda representa o valor que aparece com maior frequência num conjunto de dados. É particularmente útil para identificar o atributo mais comum num conjunto de dados. A moda pode evidenciar tipos de artefactos dominantes, práticas de enterramento e padrões recorrentes na cultura material. Um conjunto de dados pode ter uma moda (unimodal), várias modas (multimodal) ou ser amodal (sem valores repetidos). Esta estatística é relevante apenas para variáveis discretas. Quando os dados seguem uma distribuição normal, a moda coincide com a média e a mediana. Em distribuições assimétricas, a moda aparece separada das outras duas medidas de tendência central.

Todas estas medidas de tendência central podem ser obtidas de duas formas bastante simples numa folha de cálculo. A primeira é através da utilização das funções já discutidas em aulas anteriores. Neste caso, a função para cálculo da média é =AVERAGE(), para a mediana é =MEDIAN(), e para a moda é =MODE(). A segunda é através da utilização das tabelas dinâmicas. Neste caso, o procedimento é semelhante ao demonstrado para a criação de tabelas de frequência de variáveis categóricas, mas selecionando a opção de sumarizar através de uma das medidas de tendência central. A utilização de tabelas dinâmicas revela-se particularmente útil quando se pretende comparar medidas de tendência central de vários grupos de dados. A Figura 19 mostra a utilização desta ferramenta no Google Sheets para mostrar as médias da espessura de artefactos líticos de acordo com as matérias-primas utilizadas.

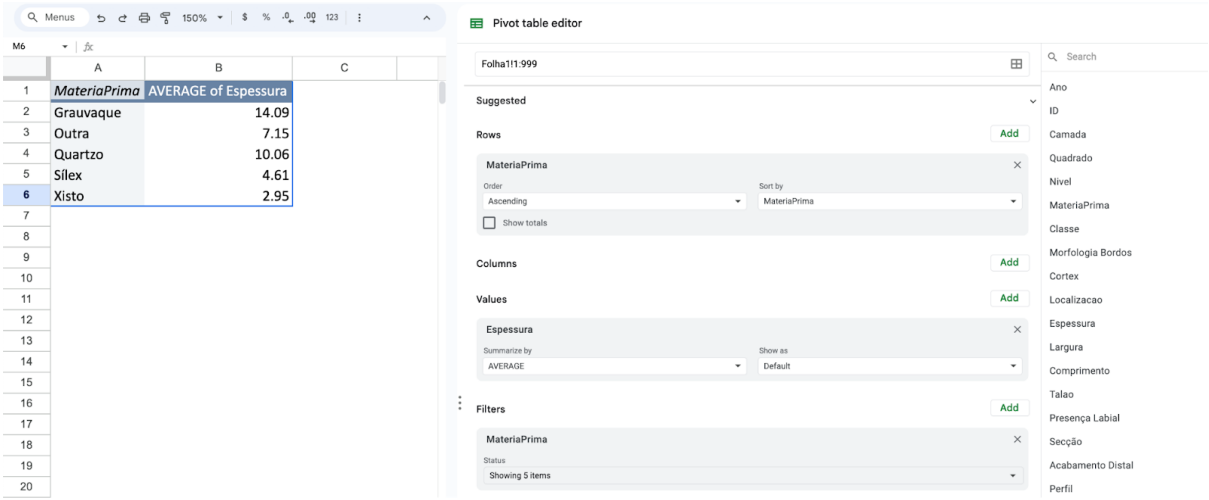


Figura 19. Exemplo da utilização de uma tabela dinâmica para a elaboração de uma tabela em que se apresentam as médias da espessura para cada uma das matérias-primas de uma coleção de ferramentas em pedra.

Medidas de dispersão

Enquanto as medidas de tendência central refletem o valor médio ou típico num conjunto de dados, as medidas de dispersão destacam a variabilidade dos dados em torno desse valor central. Infelizmente, as medidas de dispersão não são tão comuns no nosso dia a dia como as medidas de tendência central, mas são muito importantes quando pretendemos caracterizar qualquer conjunto de dados arqueológicos. Muita da sua importância está relacionada com probabilidade. Imagine-se que analisamos uma amostra de 50 pontas de seta em pedra e que o comprimento de cada uma das peças é exatamente 5 cm. Se mais tarde encontrarmos uma nova ponta de seta do mesmo conjunto, há uma probabilidade muito alta de que esta meça também exatamente 5 cm de comprimento. Isto significa que, quando não há variabilidade numa amostra, aquele conjunto de dados é muito previsível. Em contraste, imagine-se que analisamos

uma outra amostra de 50 pontas de seta em pedra, onde o valor mais pequeno para o comprimento é de 3 cm e o maior é de 8 cm, com todos os outros artefactos registando comprimentos distintos entres estes dois valores. Se encontrarmos outra peça semelhante pertencente ao conjunto, será muito mais difícil prever qual o seu comprimento, uma vez que medidas com mais variabilidade são mais difíceis de prever.

Três medidas de dispersão são comumente utilizadas em Arqueologia:

- **Intervalo.** O intervalo de variação é a medida de dispersão mais simples. Calcula a diferença entre os valores máximo e mínimo num conjunto de dados. Embora forneça uma estimativa aproximada da variabilidade, pode ser sensível a valores extremos e pode não refletir toda a distribuição. Para calcular o intervalo, subtrai-se o valor mínimo do valor máximo no conjunto de dados. Um valor de intervalo mais pequeno indica menos variabilidade, enquanto um valor mais elevado indica mais variabilidade. Numa folha de cálculo, este valor pode ser calculado utilizando a função =RANGE().

- **Variância.** A variância quantifica a diferença média ao quadrado entre cada ponto de dados e a média. Esta medida representa o quanto os valores individuais se desviam da média, fornecendo uma compreensão mais abrangente da dispersão. No entanto, a variância é apresentada em unidades quadradas, o que pode não ser tão intuitivo. Para calcular a variância, primeiramente encontra-se a média do conjunto de dados. Em seguida, calcula-se a diferença ao quadrado entre cada ponto de dados e a média. Somam-se essas diferenças ao quadrado e divide-se pelo número de observações. A fórmula de cálculo da variância é a seguinte:

$$\sigma^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$$

  Nesta fórmula, Σ representa a soma, x_i representa cada valor individual na amostra, $\bar{x}$ é a média da amostra, e $n - 1$ é o número de observações na amostra menos um. O divisor $n - 1$ é usado em vez de n para a correção de Bessel, aplicada no cálculo da variância de uma amostra para obter uma estimativa mais precisa da variância da população.

  A variância pode ser calculada a partir de uma lista de números numa folha de cálculo através da função =VARA(), ou selecionando VAR como a variável de resumo numa tabela dinâmica.

- **Desvio padrão.** O desvio padrão é a raiz quadrada da variância, expressando a dispersão nas mesmas unidades que os dados originais, o que a torna mais interpretável. Este indicador oferece uma representação equilibrada da variabilidade dos dados e é frequentemente preferido devido à sua facilidade de compreensão. Para calcular o desvio padrão, calcula-se a raiz quadrada da variância. A fórmula do desvio padrão para uma amostra é:

$$\sigma = \sqrt{\frac{\Sigma\,(x_i - \bar{x})^{\,2}}{n-1}}$$

Onde Σ indica a soma, x_i representa cada valor da amostra, $\bar{x}$ é a média da amostra, e $n - 1$ é o número total de observações na amostra menos um.

Um desvio padrão baixo indica que os valores tendem a estar próximos da média do conjunto, enquanto um desvio padrão elevado indica que os valores estão dispersos por um intervalo mais alargado. Por exemplo, as três populações {1, 1, 14, 14}, {1, 6, 8, 15} e {6, 6, 8, 10} têm uma média de 7.5. Os seus desvios padrão são 6.5, 5 e 1.7, respetivamente. A terceira população tem um desvio padrão muito mais baixo do que as outras duas porque os seus valores são todos próximos de 7. Estes desvios padrão são expressos nas mesmas unidades que os próprios pontos de dados. Por exemplo, se o conjunto de dados {1, 6, 8, 14} representa os pesos de uma amostra de quatro peças arqueológicas em gramas, o desvio padrão é 5 gramas. Como outro exemplo, a amostra {1000, 1006, 1008, 1014} pode representar as distâncias entre sítios arqueológicos e a fonte de matéria-prima mais próxima, medidas em metros. Tem uma média de 1007 metros e um desvio padrão de 5 metros.

Quando a distribuição dos dados é normal, ou seja, em que a média constitui o centro da curva (ou o "pico" do sino) com quantidades iguais de dados de ambos os lados, o desvio padrão quantifica quão larga ou estreita é a curva. Uma suposição comum sobre dados que seguem uma distribuição normal é que a área sob a curva corresponde a quantos desvios padrão estamos distantes da média, como demonstrado na Figura 20. A área entre mais e menos um desvio padrão da média contém cerca de 68% dos dados. Dois desvios padrão abrangem aproximadamente 95% dos dados, e três desvios padrão , cerca de 99.8% dos dados. Estes intervalos de probabilidade são também representados por uma unidade conhecida como sigma (cujo símbolo é σ).

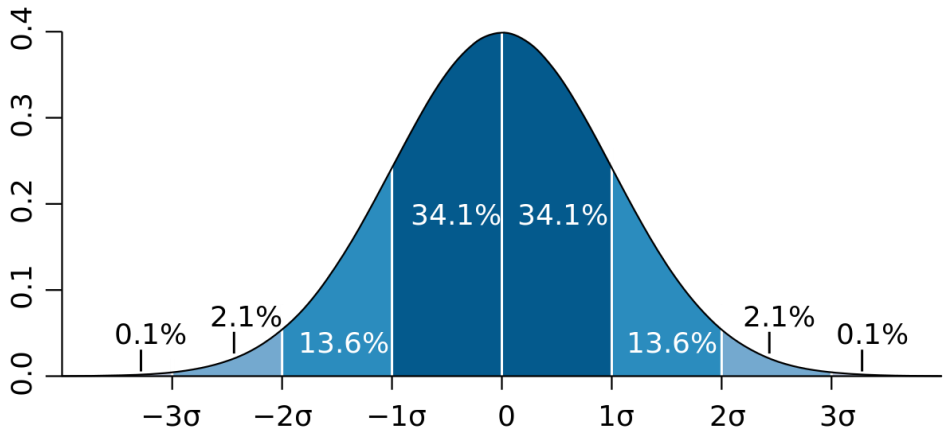


Figura 20. - Um gráfico de distribuição normal (ou curva em forma de sino) em que cada banda tem uma largura de 1 desvio-padrão. Fonte: https://en.wikipedia.org/wiki/Standard_deviation

Um dos exemplos paradigmáticos do uso do desvio padrão em Arqueologia para caracterizar um conjunto de dados ocorre na apresentação de resultados de datação por radiocarbono. As idades são habitualmente representadas por dois valores: um que representa um valor médio da idade e outro que indica o desvio padrão em torno dessa média, separados pelo símbolo "±". Assim, se temos uma idade radiocarbono de 1000±25BP, significa que existe uma probabilidade de cerca de 67% de a verdadeira idade da amostra se situar entre 1025 e 975 BP, ou uma probabilidade de cerca de 95% de se situar entre 1050 e 950 BP.

Em geral, as medidas de tendência central e de dispersão são apresentadas em conjunto sob a forma de uma tabela, onde cada coluna corresponde a um destes valores, calculados a partir de um conjunto de dados. Suponhamos que queremos mostrar os valores médios e os desvios padrão das espessuras de artefactos líticos de uma coleção. A Figura 21 mostra o resultado obtido ao utilizar uma tabela dinâmica.

A1 | AVERAGE of Espessura

	A	B	C
1	AVERAGE of Espessura	STDEV of Espessura	
2	6.50	5.45	
3			
4			

Figura 21. Exemplo da utilização de uma tabela dinâmica para representação da média e desvio padrão das espessuras de uma coleção de ferramentas em pedra.

Estes resultados são importantes por si só, pois permitem obter uma visão geral da nossa amostra. No entanto, a análise torna-se mais interessante quando, por exemplo, agrupamos os artefactos por matérias-primas (Figura 22), comparamos os diferentes valores e nos apercebemos que as peças em sílex são, em média, muito mais finas ($\bar{x}$ = 4.61) que as de quartzo ($\bar{x}$ = 10.06) ou grauvaque ($\bar{x}$ = 14.09). Além disso, verifica-se que as peças em sílex têm também uma variabilidade muito menor, revelada pelo valor baixo do desvio padrão em comparação com as restantes matérias.

	A	B	C	D
1	*MateriaPrima*	AVERAGE of Espessura	STDEV of Espessura	
2	Grauvaque	14.09	10.14	
3	Outra	7.15	5.01	
4	Quartzo	10.06	5.17	
5	Sílex	4.61	2.52	
6				
7				

Figura 22. Exemplo da utilização de uma tabela dinâmica para representação das médias e desvios padrão das espessuras de cada uma das matérias-primas de uma coleção de ferramentas em pedra.

Tal como acontece com as variáveis categóricas, outra forma de evidenciar a estrutura de um conjunto de dados é através de representações gráficas No caso das variáveis quantitativas, estes gráficos são extremamente importantes, pois permitem verificar, de forma quase instantânea, dois elementos fundamentais para a análise univariada: a forma da distribuição e a presença de valores atípicos, ou outliers.

Forma da distribuição

A forma é a principal característica que podemos determinar ao observar um gráfico. Constitui uma das primeiras análises a serem feitas, pois frequentemente dita como proceder com o resto da análise. A maioria dos métodos gráficos pode fornecer uma ideia da forma de uma distribuição, mas os mais eficazes na maioria das situações são o histograma e o gráfico de caixa e bigodes (também conhecido por diagrama de extremos e quartis, ou boxplot). Estes dois tipos de gráfico têm como objetivo mostrar, de forma visual e detalhada, os padrões sumarizados pelas medidas de tendência central e de dispersão.

O histograma é a representação gráfica de uma tabela de frequências para variáveis quantitativas. Utilizando o exemplo anterior, imaginemos que queremos construir um histograma com as espessuras dos artefactos líticos de uma coleção. O primeiro passo consiste

em agrupar todos os valores em intervalos e quantificar quantos casos existem em cada um destes. Por exemplo, se decidirmos utilizar intervalos de 5 mm (*i.e.*, 0-5; 5-10; 10-15,...), contaremos quantas peças têm espessuras medidas entre 0 e 5 mm, entre 5 e 10 mm, etc, e assim por diante. No final, teremos uma tabela como a mostrada na Tabela 8, obtida através da função =FREQUENCY().

Espessura (mm)	Frequência
0 - 5.0	99
5.1 - 10.0	63
10.1 - 15.0	19
15.1 - 20.0	6
20.1 - 25.0	3
25.1 - 30.0	1
30.1 - 35.0	0
35.1 - 40.0	0
40.1 - 45.0	0
45-1 - 50.0	1
50.1 - 55.0	0

Tabela 8. Exemplo de tabela de frequência absoluta de números de casos para intervalos específicos da variável contínua. Neste exemplo, com base nas espessuras de uma conjunto de artefactos líticos, são utilizados intervalos de 5 mm. Noutros casos, o ajuste da dimensão e número de intervalos está sempre dependente da distribuição e natureza dos dados.

O histograma é, na essência, um gráfico de barras que representa uma tabela deste tipo. Cada barra no histograma corresponde a um dos intervalos, e a altura de cada barra indica a frequência desses intervalos (Figura 23).

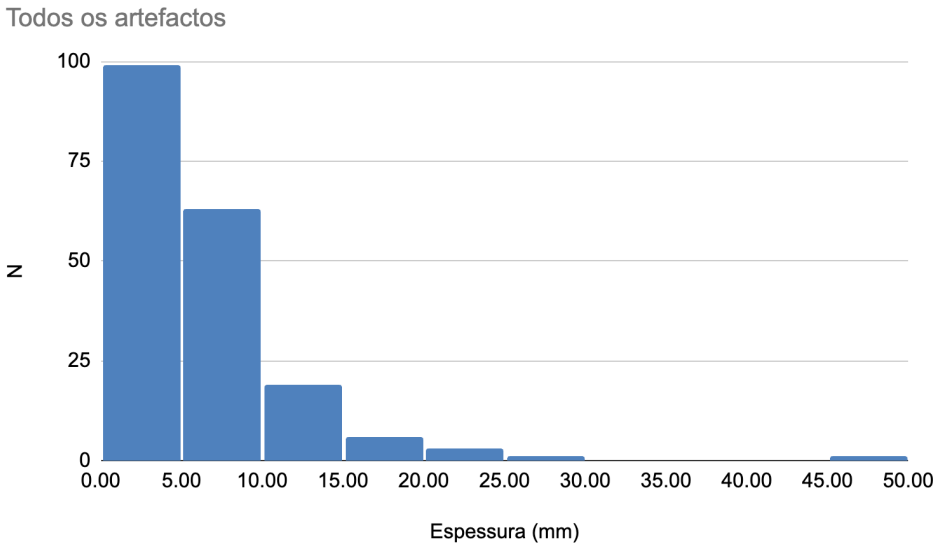


Figura 23. Histograma calculado a partir dos valores Tabela 8, representando a distribuição em intervalos de 5 mm da espessura dos artefactos em pedra de uma coleção arqueológica.

Nas folhas de cálculo, é possível simplificar o processo de construção de um histograma, uma vez que o programa calcula automaticamente os intervalos e as quantidades ao selecionar a opção de histograma. Uma das grandes vantagens dos histogramas é a sua capacidade de mostrar, de forma automática, a distribuição dos dados quantitativos contínuos. Neste contexto, alguns histogramas apresentam formas típicas que, pela frequência com que ocorrem, merecem referência especial. As distribuições mais comuns são:

Distribuições simétricas - Neste tipo de distribuição, as frequências distribuem-se de forma aproximadamente simétrica em relação a uma classe média. Um caso específico de distribuição simétrica é a chamada distribuição normal ou gaussiana, caracterizada pela sua forma de "sino". Esta distribuição de probabilidades foi descoberta no século XVIII pelo matemático e jogador Abraham de Moivre. Interessado nas probabilidades envolvidas em jogos de azar, como o lançamento de dados e moedas, de Moivre realizou experiências simples envolvendo o lançamento de moedas ao ar. Ele calculou as probabilidades de vários resultados, notando que, à medida que o número de lançamentos aumentava, a probabilidade de obter resultados extremos diminuía, enquanto a de resultados mais equilibrados aumentava. Se atirarmos uma moeda ao ar uma vez, temos 50/50 de hipóteses, ou seja, uma probabilidade de 0.5, de a moeda cair em "cara". Se lançarmos a moeda duas vezes, a complexidade aumenta. Agora, há três resultados possíveis: a moeda pode cair em "cara" duas vezes, uma vez ou nenhuma. Obter "cara" (ou "coroa") duas vezes é o resultado menos provável, com uma probabilidade de 25% ou 0.25. A probabilidade de o resultado ser dividido, com uma "cara" e uma "coroa", é a mais elevada, pois existem duas formas diferentes de alcançá-lo. Esta probabilidade é de 50%, ou seja, 0.5. À medida que o número de lançamentos aumenta, a probabilidade de obter todas "caras" ou todas "coroas" diminui. Num ensaio de apenas dez lançamentos, a probabilidade de qualquer um destes resultados é inferior a um décimo de um por cento, aproximadamente 0.00098. Em contrapartida, a probabilidade de um resultado dividido uniformemente (cinco "caras" e cinco "coroas") é de aproximadamente 25%, ou 0.25. Por que é isto importante para a análise estatística? Porque muitas características das populações estatísticas ajustam-se à distribuição normal, uma vez que a sua variância é, em grande parte, aleatória. A estatura humana é um exemplo clássico. A variância da estatura nas populações humanas é tal que a probabilidade de uma pessoa ser significativamente mais alta ou mais baixa do que a média é relativamente baixa, comparada com a probabilidade de estar mais próxima da média. Além disso, a probabilidade de uma pessoa ser mais alta ou mais baixa do que a estatura média da população pela mesma quantidade é também a mesma. Por outras palavras, a distribuição é simétrica em relação à média.

A distribuição normal também descreve muitos fenómenos quantitativos em arqueologia, especialmente aqueles relacionados com dados de medição. Os melhores exemplos de distribuições verdadeiramente normais no campo da arqueologia incluem as medições de restos de organismos. Assim, como a estatura humana, os comprimentos das tíbias das vacas, a largura dos anéis das árvores, as massas dos grãos de milho, entre outros, também seguem uma distribuição normal. Além disso, a maioria dos objetos mensuráveis no registo arqueológico exibem propriedades que seguem uma distribuição normal ou, pelo menos, quase normal. Isto ocorre porque todos os elementos mensuráveis possuem aspectos de variabilidade que são efetivamente aleatórios (McCall, 2018).

Distribuições enviesadas - A distribuição das frequências ocorre de forma acentuadamente assimétrica, apresentando valores substancialmente menores num dos lados, em comparação com o outro. Embora a maioria dos dados de medições arqueológicas se aproxima de uma distribuição normal, nenhum conjunto de dados reais é perfeitamente simétrico em relação à média. Quando as distribuições estatísticas são sistematicamente assimétricas num sentido ou noutro, denominamos isso de enviesamento. O enviesamento pode ocorrer por diversas razões, frequentemente relacionadas com a imposição de limites no intervalo possível de uma distribuição. Em Arqueologia, é comum que o zero constitua um destes limites, uma vez que as medições (ou contagens) não podem ser inferiores a zero, mas podem atingir valores infinitamente grandes. As distribuições enviesadas caracterizam-se por uma média deslocada do "pico" (o intervalo de valores com mais casos) do gráfico devido ao alongamento de uma cauda, sendo que nem o "pico" nem a média se situam no centro da distribuição. As distribuições ligeiramente enviesadas são extremamente comuns nos dados arqueológicos, devido ao facto de os valores dos dados arqueológicos tendencialmente serem limitados por zero na extremidade inferior da sua distribuição. A Figura 24 exemplifica este princípio, mostrando a distribuição dos comprimentos das lascas inteiras de uma coleção lítica Esta distribuição está positivamente enviesada, com uma média de 28.14, claramente deslocada para a direita do "pico" da distribuição, e nem a média nem o "pico" estão no centro, como seria de esperar numa distribuição gaussiana.

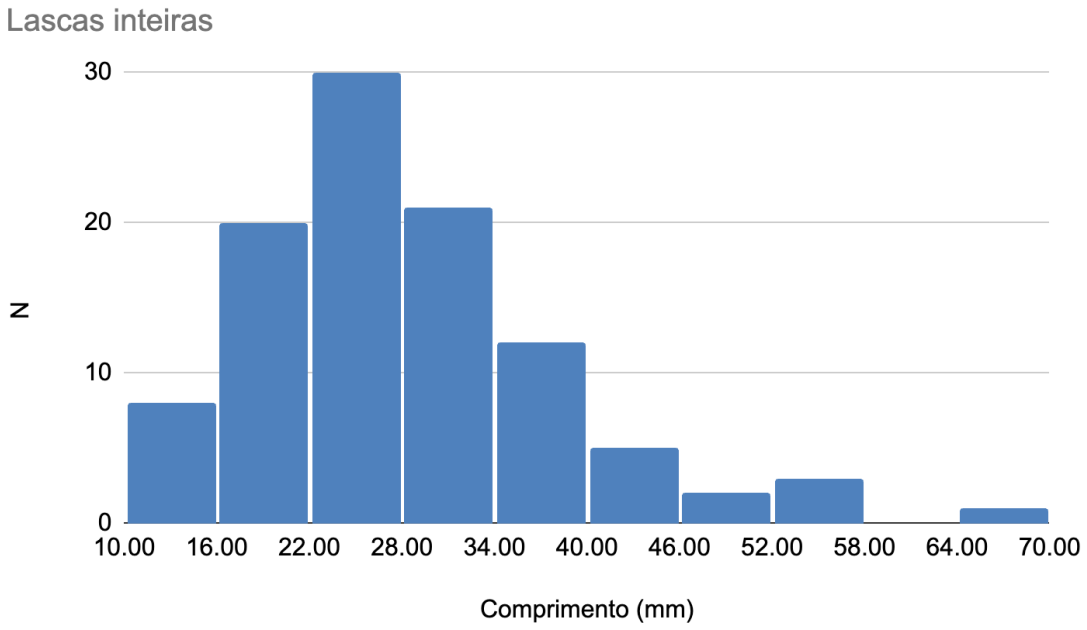


Figura 24. Histograma dos comprimentos das lascas inteiras de uma coleção de ferramentas em pedra.

Distribuições com vários "picos" ou modas - Neste tipo de distribuição, as frequências apresentam dois ou mais "picos", denominados modas (ver ponto seguinte), sugerindo que os dados são constituídos por vários grupos distintos. Em Arqueologia, a distribuição bimodal é a mais comum, sendo as multimodais muito mais raras. A diferenciação entre distribuições unimodais e bimodais tem desempenhado um papel relevante na Arqueologia, particularmente na definição de tipos de artefactos a partir da quantificação da sua forma (Ford, 1954; McPherron, 1994). Um exemplo que ilustra bem a importância de uma distribuição bimodal na análise de indústrias líticas é a separação dos chamados produtos alongados (peças cujo comprimento é igual ou superior ao dobro da sua largura) em lâminas ou lamelas. Tradicionalmente, é aceite a definição de Tixier (1963) em que as lâminas são definidas como tendo uma largura mínima de pelo menos 12 mm ou um comprimento superior a 50 mm. Tixier ilustrou estes rácios através da análise de conjuntos artefactuais datados do Epipaleolítico no Maghreb, onde a distribuição dos valores de largura e comprimento dos elementos alongados inteiros revelava dois "picos" distintos, sugerindo a intenção dos artesãos de produzir duas dimensões de peças através de cadeias de produção separadas, em vez de através de um único processo de redução contínuo. Assim, quando lidamos com dados de medição e encontramos uma distribuição multimodal, isso indica normalmente que estamos a combinar diferentes tipos na mesma categoria de observação (ou seja, vários tipos de artefactos) (ver também Spaulding, 1953).

O segundo tipo de gráfico que permite verificar a forma da distribuição é o chamado boxplot. Os princípios de construção de um boxplot diferem significativamente dos do histograma, principalmente porque não é necessário realizar o agrupamento em intervalos e a contagem do número de casos, tornando a sua construção muito mais simples. Para construir um boxplot, precisamos apenas de calcular o chamado resumo dos cinco números: mínimo, 1º quartil, mediana, 3º quartil e máximo (Figura 25). Dos valores mencionados, os quartis ainda não foram previamente explicados. O 1º quartil é o valor abaixo do qual se encontram 25% das observações, enquanto o 3º quartil é o valor abaixo do qual se encontram 75% das observações. O conceito é semelhante ao da mediana, que divide o conjunto ordenado de dados em partes iguais. Uma vez obtidos estes cinco números, a construção do gráfico é bastante simples, como representado na Figura 26, que apresenta os mesmo dados que o histograma anterior.

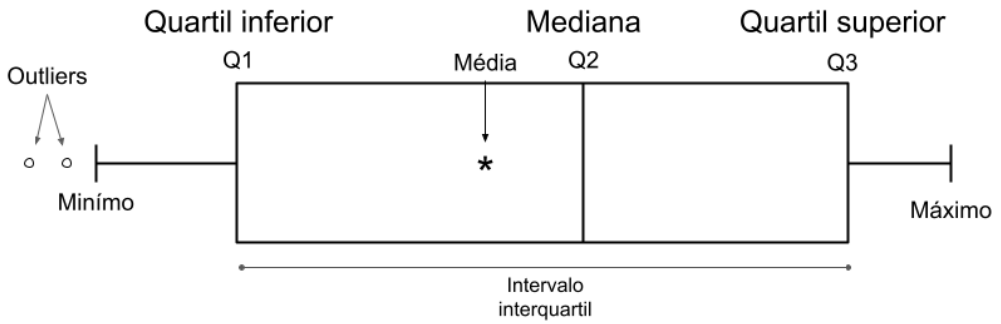


Figura 25. Principais elementos para construção e leitura de uma gráfico de caixa e bigodes (boxplot) de acordo com o denominado resumo dos cinco números: mínimo, 1º quartil, mediana, 3º quartil e máximo.

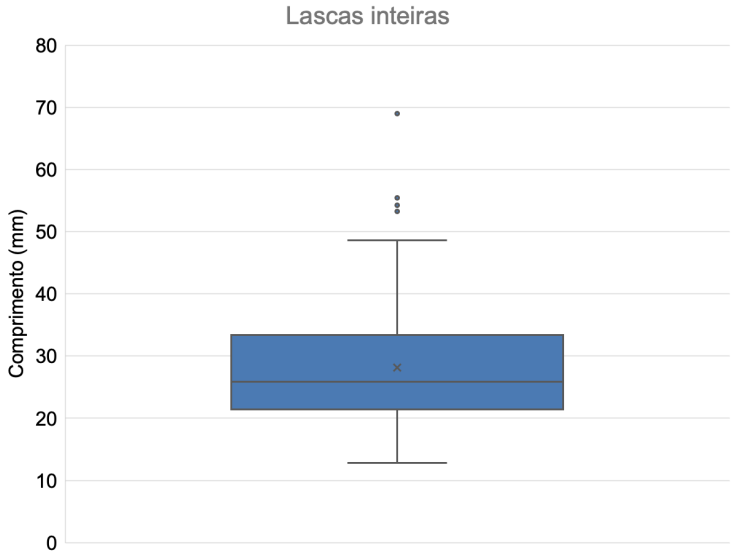


Figura 26. Boxplot dos valores de comprimento das lascas inteiras de uma coleção de ferramentas em pedra. Nota: atualmente (final de 2023) o Google Sheets não inclui uma opção para construção de boxplots, pelo que e a imagem foi elaborada com recurso ao programa MS Excel.

Também no caso dos boxplots é possível discernir padrões de simetria e enviesamento dos dados. Isto é feito através da análise da distância entre a linha da mediana e os limites superior e inferior do rectângulo, do comprimento do rectângulo, e das linhas que se estendem a partir do topo e base do rectângulo. Os boxplots são particularmente úteis para a comparação de vários conjuntos de dados. A Figura 27, por exemplo, representa os comprimentos das lascas de cada matéria-prima lado a lado, facilitando a compreensão das diferenças nas distribuições de cada grupo.

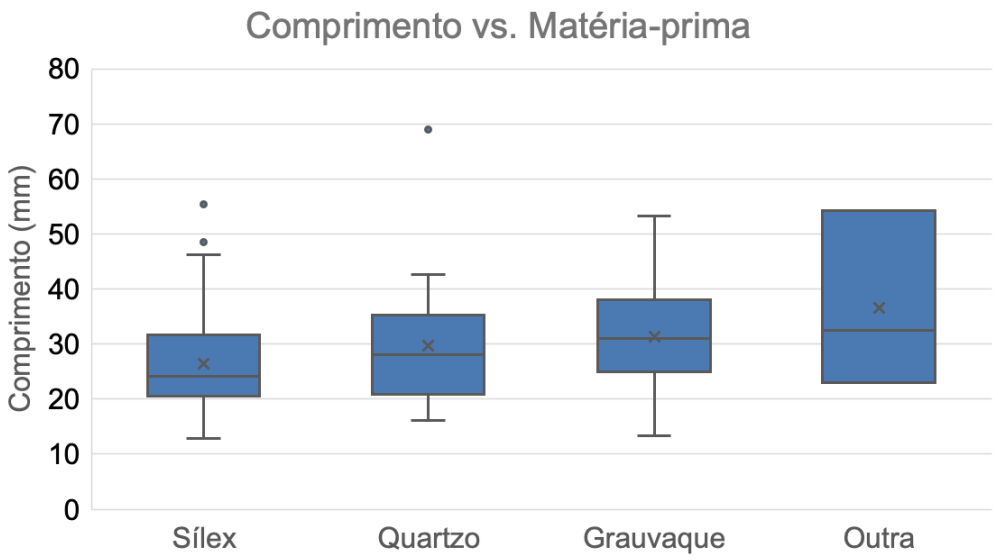


Figura 27. Boxplot dos valores de comprimento, por cada uma das matérias-primas, das lascas inteiras de uma coleção de ferramentas em pedra. Nota: atualmente (final de 2023) o Google Sheets não inclui uma opção para construção de boxplots, pelo que e a imagem foi elaborada com recurso ao programa MS Excel.

É de salientar, no entanto, que embora os boxplot forneçam informações sobre as estatísticas resumidas dos dados, eles oferecem uma visão menos detalhada dos valores reais e da sua densidade dentro da distribuição, em comparação com os histogramas. Assim, deve-se: utilizar um histograma quando se pretende compreender a forma detalhada da distribuição, identificar grupos ou modos e visualizar a densidade dos dados ao longo de todo o seu intervalo; e utilizar um boxplot quando o interesse residir mais em resumir a tendência central e a dispersão dos dados, comparar rapidamente vários conjuntos de dados e identificar potenciais valores atípicos.

Outliers

Por vezes, encontramos um ou mais pontos de dados que se destacam visualmente no conjunto. Estes valores extremos, potencialmente anómalos, podem ser, em algumas ocasiões, óbvios,

como demonstrado no histograma da Figura 24 e no boxplot da Figura 26. Estes valores são chamados de *outliers*. Os valores anómalos numa representação boxplot são normalmente identificados utilizando um método baseado no intervalo interquartil (IQR) que é, nada mais nada menos, a diferença entre o terceiro e o primeiro quartil (*i.e.*, Q3 - Q1), e que representa 50% dos dados que se encontram no meio da distribuição (Figura 28). Para chegar ao limite a partir do qual determinados dados são considerados *outliers*, a maior parte dos software efetuam os seguintes cálculos:

```
Limite inferior = Q1 - 1.5 * IQR

Limite superior = Q3 + 1.5 * IQR
```

Qualquer ponto de dados que seja inferior ao limite inferior ou superior ao limite superior é considerado um *outlier*.

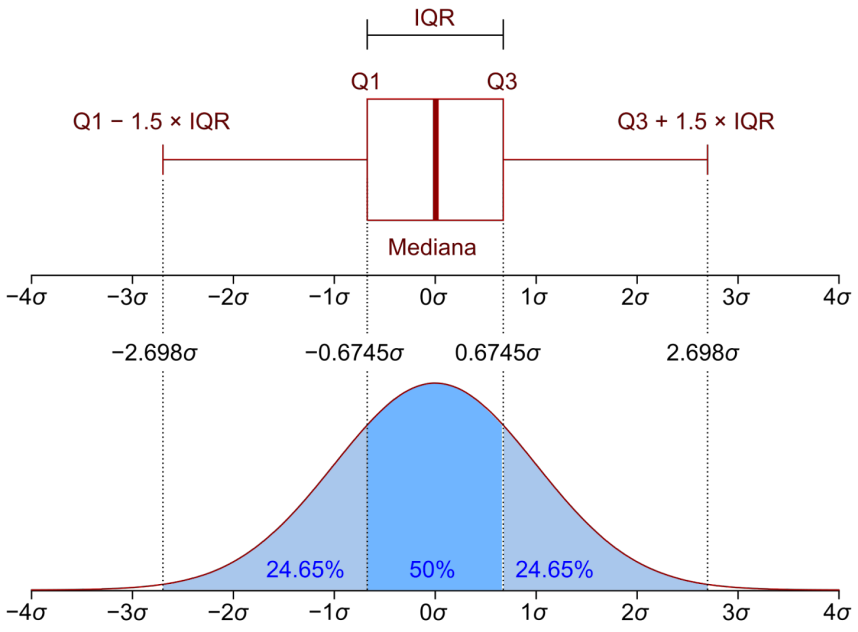


Figura 28. Boxplot com representação do intervalo interquartil e a correspondência de representação em percentagens numa curva de distribuição considerada normal. Modificado a partir de: https://pt.wikipedia.org/wiki/Amplitude_interquartil#/media/Ficheiro:Boxplot_vs_PDF_-_PT.svg

A deteção destes valores anómalos é importante por várias razões:

- **Limpeza de dados e garantia de qualidade:** As anomalias podem surgir devido a erros na recolha, introdução ou processamento de dados. A identificação destas anomalias ajuda na limpeza dos dados e garante a qualidade e a exatidão do conjunto de dados.

- **Compreender os dados**: Os valores anómalos podem fornecer informações sobre os dados. Podem indicar variabilidade na medição ou erros experimentais, mas também podem significar que os dados têm uma distribuição marcadamente assimétrica.
- **Impacto na análise estatística**: Os valores atípicos podem distorcer significativamente os resultados das análises estatísticas. Podem afetar a média, o desvio padrão e outras medidas estatísticas, levando a conclusões incorrectas. Por conseguinte, é necessário identificar e tratar os valores atípicos para uma modelação estatística exacta.
- **Valor informativo**: Em alguns casos, os valores atípicos são a parte mais interessante dos dados. Por exemplo, em Arqueologia, os valores atípicos podem representar acontecimentos invulgares e, por isso, de extrema importância para caracterizar idiossincrasias no comportamento humano.
- **Melhorar os modelos**: Na modelação preditiva e *machine learning*, os valores atípicos podem ser uma fonte de ruído ou uma fonte de informação valiosa. A sua identificação ajuda a decidir se devem ser incluídos ou excluídos do processo de formação do modelo, melhorando assim o desempenho e a precisão do modelo.
- **Pressupostos dos testes estatísticos:** Muitos testes estatísticos pressupõem que os dados são normalmente distribuídos. Os valores atípicos podem violar estes pressupostos, afetando a validade destes testes. A deteção de *outliers* é, portanto, crucial antes de aplicar estes testes.

Aula 09 - Análise bivariada

Abordados os diferentes tipos de análise univariada nas aulas anteriores, nesta aula explora-se a importância da combinação de duas variáveis para a deteção de padrões e interpretação de uma coleção arqueológica. As diversas combinações entre variáveis qualitativas e quantitativas são exploradas com recurso à representação sobre a forma de tabelas e gráficos. Na segunda parte da aula, com foco nos gráficos de dispersão, introduzem-se também os conceitos de correlação e regressão.

Duas variáveis categóricas

Apesar de a caracterização de uma única variável ser importante para a análise exploratória de dados em Arqueologia, na realidade são raros os casos em que isso é suficiente para detetar padrões ou tirar conclusões mais sólidas sobre o registo arqueológico. Frequentemente, o trabalho de arqueólogo baseia-se na comparação entre conjuntos de dados.

Um caso bastante frequente em Arqueologia é a comparação entre duas variáveis categóricas ou qualitativas. De facto, como visto anteriormente, muitas das variáveis registadas na análise de artefactos ou mesmo durante o trabalho de campo são categóricas. Por exemplo, pode ser de interesse comparar as classes de artefactos presentes em cada uma das unidades estratigráficas de um sítio arqueológico, ou a cor da pasta entre diferentes conjuntos de cerâmica. Para este fim, dispomos de duas abordagens principais: a utilização das denominadas **tabelas de contingência** ou a representação gráfica através de **gráficos de barras** ou, mais frequentemente, de **barras empilhadas**.

Uma tabela de contingência, também conhecida como tabela cruzada ou bidirecional, é uma ferramenta estatística utilizada para apresentar e analisar a relação entre duas variáveis categóricas. As tabelas de contingência são geralmente organizadas em linhas e colunas, onde cada linha corresponde a uma categoria da primeira variável e cada coluna a uma categoria da segunda variável. As células da tabela contêm as frequências ou contagens de observações que resultam da intersecção dessas categorias. Esta organização permite-nos observar facilmente a distribuição conjunta das duas variáveis e pode auxiliar na identificação de padrões ou associações entre elas.

Numa folha de cálculo, é muito fácil criar este tipo de tabelas utilizando a ferramenta de tabelas dinâmicas. Para isso, basta arrastar para as colunas e linhas duas variáveis categóricas distintas. O programa calculará automaticamente o número de casos para cada combinação possível. Na

Tabela 9, por exemplo, comparam-se as matérias-primas de uma coleção de indústria lítica com as respectivas morfologias dos bordos. Neste caso, a contagem apresentada é uma frequência absoluta, mas também é possível (e frequentemente desejável) apresentar os valores em frequências relativas, formatadas como percentagem.

	Grauvaque	Outra	Quartzo	Sílex	Xisto	TOTAL
Biconvexos	1	2		4		7
Circular	2			2		4
Convergentes	3	1	5	6		15
Divergentes	8		8	39		55
Irregulares	5	3	7	37	1	53
Outro				2		2
Paralelos		1	9	45		55
TOTAL	19	7	29	135	1	191

Tabela 9. Exemplo de tabela de contingência em que se cruzam as variáveis Matéria-prima (colunas) com a variável Morfologia dos bordos (linhas). Cada célula da tabela mostra a contagem de ocorrências para as categorias correspondentes.

Os valores apresentados numa tabela de contingência podem também ser facilmente apresentados sob a forma de gráfico. Um dos gráficos mais comuns neste âmbito é o gráfico de barras empilhadas. Este tipo de diagrama é um tipo de visualização de dados que representa dados categóricos com barras retangulares. Cada barra é dividida em segmentos ou sub-barras que correspondem a diferentes categorias dentro da mesma variável. O comprimento de cada segmento dentro de uma barra corresponde à proporção dos dados que representa. Na Figura 29, apresentam-se os mesmos dados da tabela de contingência da Tabela 9. Neste caso, foi utilizado um gráfico de barras empilhadas a 100%, significando que as frequências são apresentadas em percentagem, com todas as barras tendo o mesmo comprimento, mas variando na dimensão dos segmentos. Estes segmentos são apresentados com cores distintas para facilitar a identificação das tendências na distribuição das morfologias dos bordos pelas várias matérias-primas.

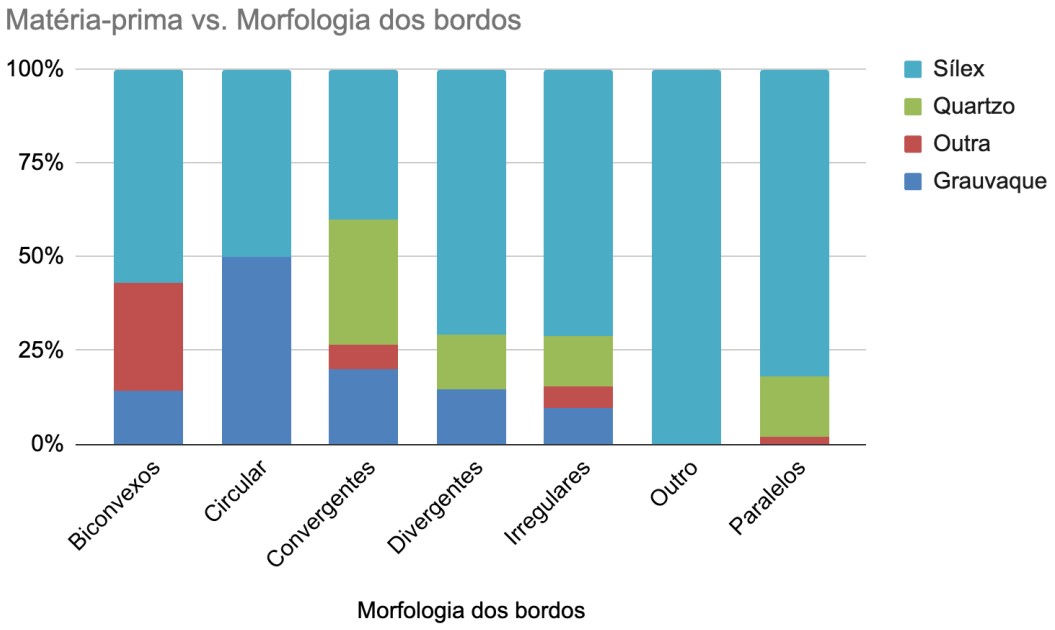


Figura 29. Gráfico de barras empilhadas a 100% com a incidência de cada uma das matérias-primas nas várias morfologias dos bordos de uma coleção de ferramentas em pedra.

Uma variável categórica e uma variável numérica

Em alguns dos exemplos já vistos no âmbito das medidas de tendência central, foram apresentadas comparações entre uma variável qualitativa e outra quantitativa. Um dos exemplos é a comparação dos diferentes comprimentos de peças líticas feitas em diferentes matérias-primas. Nesse caso, comparou-se uma variável quantitativa contínua (comprimento) com uma variável qualitativa (matéria-prima), com o objetivo de perceber como cada matéria-prima foi utilizada para (ou influenciou) a produção de lascas de diferentes tamanhos. No cruzamento entre variáveis quantitativas e qualitativas, além da utilização de tabelas com os valores centrais e de dispersão, podem ser empregues alguns métodos gráficos já mencionados. O mais comum neste contexto é o uso de um **boxplot comparativo**, em que na mesma área do gráfico aparecem representadas várias caixas correspondentes a cada grupo formado pela variável categórica (Figura 27). Este tipo de gráfico é feito diretamente na folha de dados sem a necessidade de transformação dos valores. Para criar o gráfico automaticamente, basta que a primeira coluna contenha a variável qualitativa e a segunda a variável quantitativa. Após a seleção dos dados, insere-se o tipo de gráfico boxplot.

Duas variáveis numéricas

O último tipo de análise bivariada é quando queremos cruzar informação de duas variáveis numéricas. Esta abordagem é muito comum em Arqueologia, mas também em outras ciências, como a Economia ou a Biologia, dada a sua utilidade como técnica de predição de novos valores.

Para cruzarmos duas variáveis numéricas, uma das formas gráficas mais utilizadas é o chamado **diagrama de dispersão**. Este diagrama é uma representação gráfica de pontos de dados num espaço bidimensional, geralmente num sistema de coordenadas cartesianas. Cada ponto de dados é representado como um ponto único no gráfico, com a sua posição determinada pelos valores de duas variáveis. Na Figura 30, apresenta-se um exemplo de um diagrama de dispersão que cruza informação da espessura e largura das lascas de uma coleção de ferramentas em pedra.

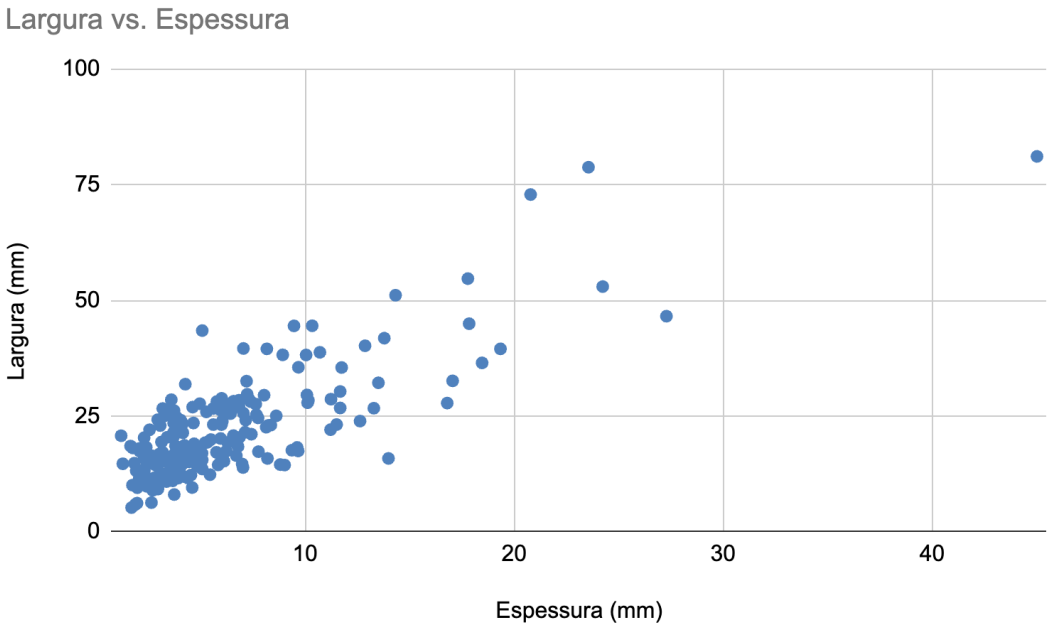


Figura 30. Gráfico de dispersão com representação da relação entre largura e espessura de artefactos em pedra de uma coleção arqueológica.

Este diagrama de dispersão é extremamente informativo. Podemos ver que a espessura das lascas aumenta à medida que a sua largura também aumenta. Além disso, podemos ver que a relação entre estas duas variáveis tem aproximadamente a forma de uma linha reta. Isto é, uma linha reta adequadamente posicionada, passaria muito perto de todos os pontos. Por outras palavras, para um dado aumento na largura, existe um correspondente aumento na espessura

das lascas, ao longo de toda a escala de largura, e todos os artefactos seguem mais ou menos esta relação.

No caso do cruzamento de variáveis contínuas, existem normalmente dois objetivos principais. O primeiro é a verificação da **correlação** entre as variáveis, e o segundo é a previsão de novos valores com base na relação dos valores existentes. Este segundo objetivo é alcançado através de uma técnica matemática conhecida como **regressão**.

Correlação

A análise de correlação constitui a pedra angular na revelação de relações entre variáveis. Procura responder a questões como: Estão estas variáveis relacionadas? Em caso afirmativo, qual é a intensidade e natureza dessa relação? Na sua essência, a análise de correlação procura quantificar o grau de associação entre variáveis, permitindo uma compreensão aprofundada das interligações nos seus conjuntos de dados. O primeiro objetivo desta análise é verificar a existência de uma relação genuína entre duas variáveis. Isto implica avaliar se as alterações numa variável coincidem com as alterações na outra. A natureza desta relação é elucidada classificando-a numa de três categorias: uma relação direta (**correlação positiva**), uma relação inversa (**correlação negativa**) ou a ausência de uma relação discernível. Uma correlação positiva indica que, à medida que uma variável aumenta, a outra também tende a aumentar, enquanto uma correlação negativa significa que, à medida que uma variável aumenta, a outra tende a diminuir. A análise da Figura 30 demonstra uma correlação positiva entre as duas métricas representadas.

Regressão

A regressão, por sua vez, é uma técnica de modelação preditiva que tem como objetivo estabelecer uma relação funcional entre variáveis. Ao contrário da correlação, a regressão procura criar um modelo matemático capaz de prever os valores de uma variável com base nos valores de uma ou mais variáveis. Ela pressupõe uma direção causal e utiliza a relação estabelecida para fazer previsões. A **regressão linear**, talvez a forma mais conhecida de regressão, procura encontrar a linha que melhor se ajusta aos pontos de dados, minimizando a soma das diferenças quadráticas entre os valores observados e os valores previstos.

Em âmbito arqueológico, as regressões envolvem normalmente a tentativa de modelar a mudança numa medida em função da mudança noutra medida. Em termos matemáticos, modelar este tipo de relação implica determinar uma função linear que inclui: (1) uma reta com um determinado declive e (2) uma constante que define a localização da reta no plano cartesiano. O declive de uma reta descreve a variação dos valores de uma variável em resposta à alteração dos valores de outra variável. Isto é, para duas variáveis x e y, o declive indica quanto o valor de y aumenta ou diminui sempre que o valor de x aumenta em uma unidade.

Porque é isto importante para a análise estatística? Ao nível mais básico, a análise de regressão informa sobre o tipo de relação existente entre as duas variáveis de interesse. Se se encontrar uma relação direta - representada por uma função com um valor positivo para o declive da linha - isso significa que uma variável aumenta à medida que a outra também aumenta. Por outro lado, uma relação inversa - uma linha com um declive negativo - indica que uma variável diminui à medida que a outra aumenta. Para além disso, a "inclinação" do declive fornece a magnitude da relação entre as duas variáveis. Valores muito baixos para o declive de uma reta podem indicar a ausência de uma relação significativa. Por estas razões, encontrar o declive da reta que melhor se ajusta aos dados bivariados é um passo fundamental no processo de identificação de padrões e na procura de relações significativas.

Frequentemente, são publicados gráficos de dispersão com base em materiais arqueológicos que incluem uma linha desenhada sobre os pontos do gráfico (Figura 31). Esta linha é denominada **reta de regressão**.

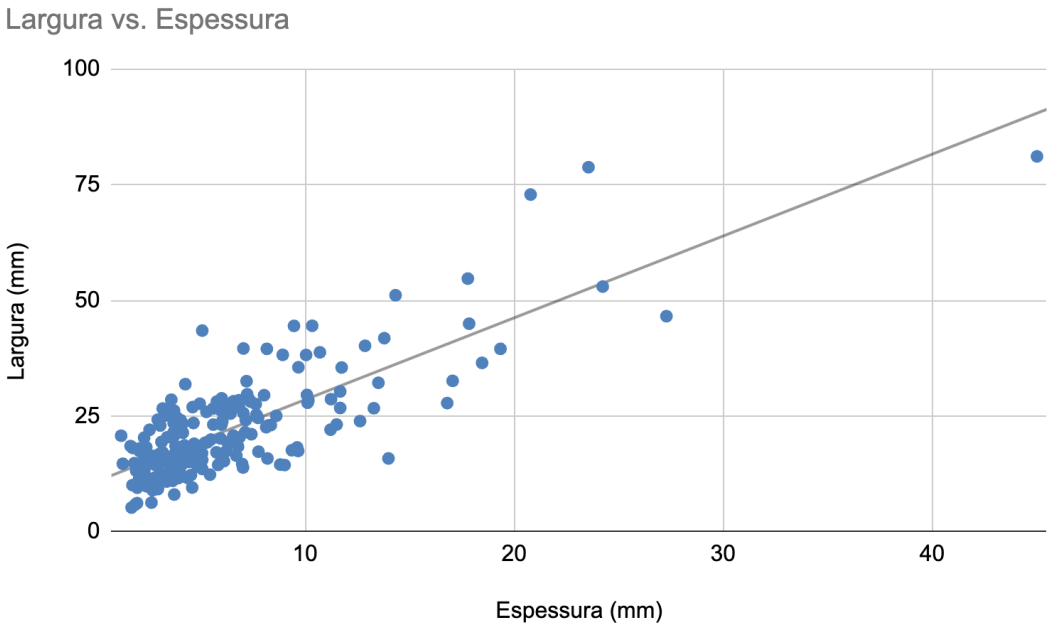


Figura 31. Gráfico de dispersão com representação da relação entre largura e espessura de artefactos em pedra de uma coleção arqueológica. A reta de regressão demonstra a relação positiva entre as duas variáveis.

O processo matemático para determinar a função que define a localização da reta é relativamente complexo, sendo necessário compreender alguns conceitos para explicar exatamente como o fazer. Em primeiro lugar, existe a prática de determinar o "melhor ajuste" de uma linha de regressão. Isto significa descobrir a equação de uma função que mais se aproxima de todos os pontos de dados do nosso gráfico. Seguidamente, encontra-se o conceito de **resíduo**. No contexto da regressão linear, o resíduo é a distância vertical (ao longo do eixo y) entre um ponto de dados específico e o valor previsto para esse ponto com base na função da reta de regressão (Figura 32). Por outras palavras, o resíduo representa a diferença entre o ponto de dados observado e as previsões baseadas na linha de melhor ajuste que explica a relação entre as duas variáveis.

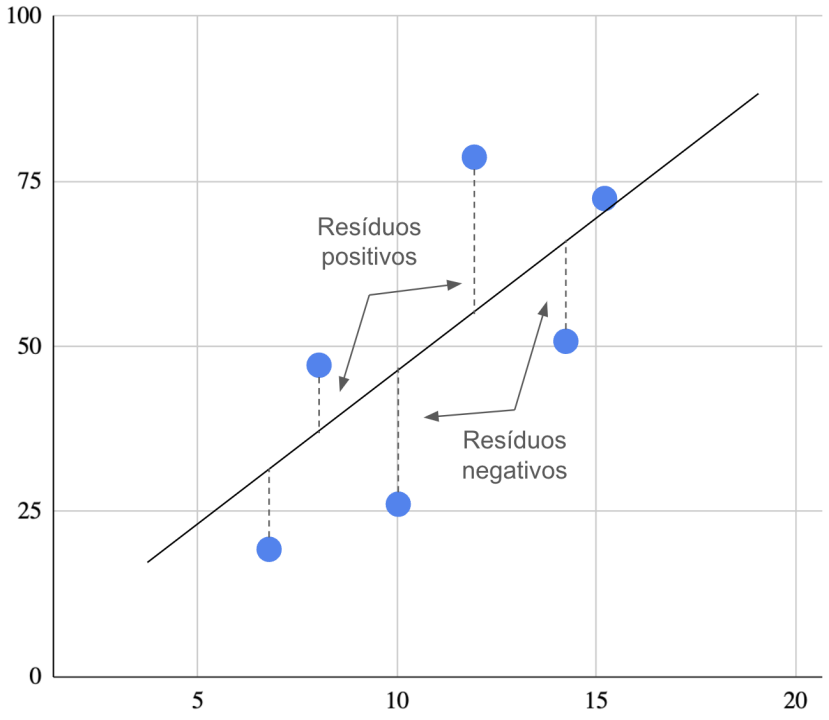


Figura 32. Quando se efectua uma regressão linear simples (ou qualquer outro tipo de análise de regressão), obtém-se uma linha de melhor ajuste. Os pontos de dados normalmente não caem exatamente nesta linha de equação de regressão; estão dispersos. Um resíduo é a distância vertical entre um ponto de dados e a linha de regressão.

Praticamente todos os programas estatísticos, incluindo os de folhas de cálculo, permitem colocar automaticamente retas de regressão linear sobre gráficos de dispersão. No entanto, nem sempre é claro para o utilizador o significado dessa reta ou como ela foi calculada. De forma

muito simples, quando o programa calcula a equação de uma reta de regressão, está a encontrar a reta que minimiza os resíduos para cada ponto de dados. Mais especificamente, pode-se imaginar o programa ajustando a reta, movendo-a para cima e para baixo e alterando o seu declive. Ao fazer isso, os valores dos resíduos para cada ponto de dados mudam à medida que a reta se aproxima ou se afasta deles. Ao determinar a linha de regressão que melhor se ajusta, o programa está a minimizar a soma total dos resíduos para todos os pontos no gráfico. A abordagem matemática mais comum para calcular a função de uma linha de regressão é o conhecida como método dos mínimos quadrados.

Em conjunto com a representação da reta no gráfico de dispersão, muitos dos programas de folhas de cálculo permitem ainda incluir no próprio gráfico dois outros elementos que são fundamentais para complementar a informação da reta: a equação da regressão e o valor de R^2. A equação é geralmente apresentada na forma:

$$y = a + bx$$

em que x é a variável independente (ou explicativa) e y é a variável dependente. O declive da reta é representado por b, e a é a interseção (*i.e.*, o valor de y quando $x = 0$) (ver Figura 33).

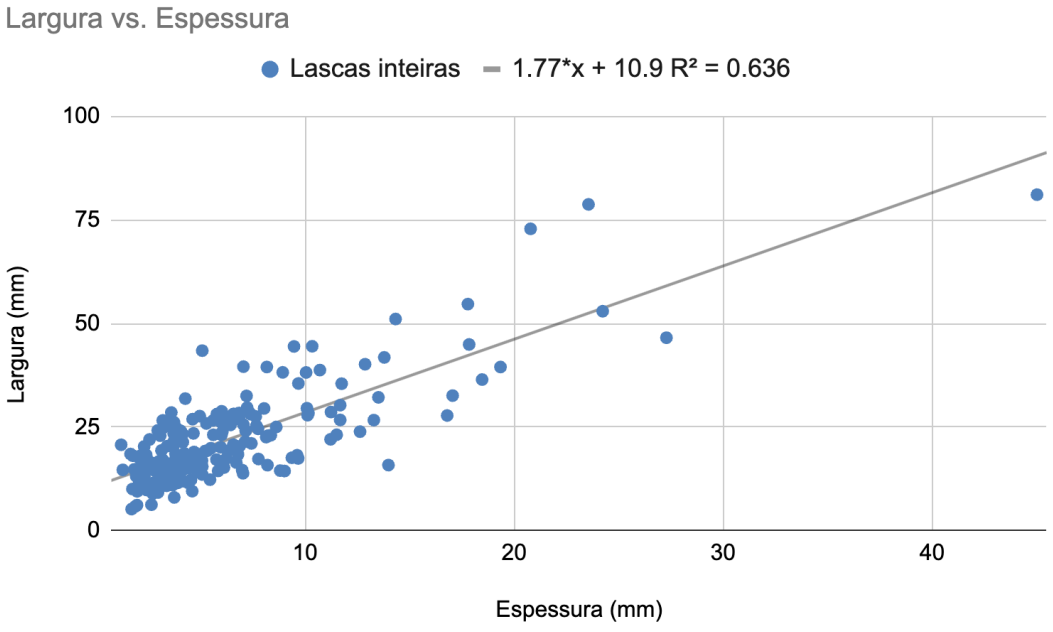


Figura 33. Gráfico de dispersão com a representação da reta calculada pela regressão linear, bem como a equação da regressão e o valor de R^2.

Esta equação é importante, pois permite prever novos valores da variável dependente, a partir dos valores da variável independente. Contudo, a diferença entre **variáveis dependentes e**

independentes é muitas das vezes confusa quando falamos de dados arqueológicos. A regressão linear é normalmente conceptualizada como envolvendo dois tipos de variáveis: a dependente e a independente. Convencionalmente, a variável independente é representada ao longo do eixo do X dos gráficos, e a variável dependente ao longo do eixo do Y. Em Arqueologia, contudo, esta distinção não é assim tão importante. Ou melhor, se fôssemos cientistas experimentais, a distinção seria muito mais clara. O cientista que administra uma certa quantidade de uma droga para a inteligência a um rato, sendo esta quantidade a variável independente, e posteriormente mede o tempo que o rato demora a percorrer um labirinto, sendo este tempo a variável dependente).

Para a Arqueologia, assim como para a maioria das outras ciências observacionais e históricas, a linha de distinção entre variáveis independentes e dependentes torna-se, na maior parte dos casos, pouco evidente. A principal razão para isso é que, com raras excepções (como na Arqueologia Experimental), não temos a capacidade de controlar nenhuma das nossas variáveis da forma como os cientistas experimentais o fazem. As frequências e medições das coisas arqueológicas são inalteráveis, e as nossas observações são feitas sem qualquer possibilidade de controle sobre elas. Além disso, frequentemente procuramos relações entre variáveis em contextos onde a causalidade é, na melhor das hipóteses, ambígua e, na pior, absurda (McCall, 2018). Por exemplo, ao cruzarmos informações sobre as larguras e espessuras das lascas do exemplo anterior, sabemos à partida que ambos os valores têm uma relação instrumental, mas um não causa efetivamente o outro.

A única base lógica real para distinguir entre variáveis independentes e dependentes em análises de dados arqueológicos está, talvez, na variável que é usada para modelar outra variável. Embora a largura da lasca não cause a espessura da mesma, pretendemos usar a largura para prever a espessura. Pensando de outra forma, estamos a introduzir um valor para a largura na nossa análise de regressão para prever um valor para a espessura da lasca. Assim, mesmo que não possamos controlar literalmente a largura como uma variável independente, ela atua como tal no sentido em que usamos valores conhecidos da largura da lasca para fazer previsões sobre valores desconhecidos da sua espessura.

Aula 10 - Casos de estudo de análise exploratória de dados

Nesta aula, exploram-se e discutem-se dois casos de estudo arqueológicos que utilizam análises exploratórias de dados. O principal objetivo é demonstrar como os conceitos abordados nas aulas anteriores são aplicados em estudos arqueológicos reais. Pretende-se também que os alunos adotem uma postura crítica em relação às opções tomadas para a apresentação dos dados e que, sempre que possível, tentem reproduzir os resultados dos estudos.

A utilização dos métodos exploratórios de dados, abordados nas aulas anteriores, constitui um dos primeiros passos na caracterização de conjuntos artefactuais arqueológicos. Através destes métodos, conseguimos explorar tendências, detetar padrões e identificar anomalias que, de outra forma, poderiam não ser evidentes. Ao apresentar tabelas e gráficos durante esta fase da análise, é importante tomar várias precauções para assegurar que a informação é transmitida de forma exacta e abrangente. Eis algumas das principais precauções:

- **Evitar escalas enganadoras:** As escalas utilizadas nos gráficos, especialmente nos gráficos de barras, devem começar em zero, caso contrário, podem criar uma representação enganadora das diferenças. Por exemplo, um gráfico de barras que inicia em 50, em vez de 0, pode exagerar diferenças pequenas.
- **Fornecer contexto:** É crucial rotular sempre os eixos de forma clara e indicar as unidades. É também sempre necessário que qualquer mapa ou visualização de dados inclua uma legenda ou chave clara.
- **Cuidado com a sobreposição de gráficos:** Ao representar grandes conjuntos de dados, os pontos de dados podem sobrepor-se, tornando difícil discernir a distribuição dos mesmos.
- **Garantir a legibilidade:** As cores e os tamanhos de letra devem ser escolhidos de forma a serem claros e legíveis. Deve-se evitar utilizar demasiadas cores ou criar desenhos excessivamente complexos que possam confundir o leitor.
- **A escolha da cor é importante:** é importante garantir que as cores são perceptualmente uniformes, para que as diferenças nos dados sejam representadas de forma uniforme nas diferenças de cor. Ter em atenção o daltonismo. Ferramentas como o ColorBrewer podem ajudar a escolher esquemas de cores adequados.

- **Representar a incerteza:** Sempre que possível, apresentar medidas de incerteza, como intervalos de confiança ou erros padrão, nos diferentes elementos.
- **Evitar a seleção de dados:** Apresentar a história completa e não realçar apenas os dados que apoiam um determinado argumento ou hipótese.
- **Cuidado com o enviesamento na seleção:** é importante certificar que os dados apresentados não são afetados por enviesamento na seleção ou que, se o forem, esse enviesamento é claramente comunicado.
- **Verificar a existência de valores anómalos:** Os valores anómalos podem influenciar indevidamente os resultados. É importante identificá-los, e as decisões sobre a sua inclusão ou exclusão devem ser tomadas de forma transparente.
- **Evitar gráficos circulares com muitas categorias:** Os gráficos circulares podem ser difíceis de interpretar com exatidão, especialmente quando contêm muitas fatias. Em alternativa, deve-se considerar a utilização de um gráfico de barras ou outro método de visualização mais eficaz.
- **Contextualizar os números absolutos com taxas ou valores por unidade:** Por exemplo, ao apresentar dados sobre duas coleções de cerâmicas, mostrar apenas o número absoluto de peças com determinada característica pode ser enganador se as duas coleções forem de tamanhos drasticamente diferentes.
- **Considerar a granularidade:** Dados agregados podem, por vezes, ocultar padrões locais importantes. Por outro lado, dados demasiado granulares podem ser demasiado ruidosos. É importante escolher o nível de granularidade adequado para a análise e apresentação.
- **Relativamente às tabelas:**
    - Evitar a desordem. Apresentar apenas o que é necessário para a mensagem principal.
    - Certificar de que as tabelas têm cabeçalhos claros e são fáceis de ler.
    - Considerar a possibilidade de complementar com visualizações para tabelas grandes ou para destacar pontos-chave.
    - Documentação e fonte: Deve-se sempre fornecer a fonte dos dados e detalhes sobre qualquer processamento ou transformação aplicada.
- **Procurar feedback:** Frequentemente, a primeira tentativa de visualização ou tabulação de dados pode não ser a mais eficaz. Por isso, é sempre uma boa ideia repetir as concepções e obter feedback de outras pessoas para garantir clareza e precisão.
- Muitas vezes, a consulta de outros artigos, relatórios, teses, etc., é fundamental para tomarmos decisões informadas no que concerne à apresentação dos resultados.

Os artigos listados abaixo são alguns dos *case studies* usados em aula ao longo das várias edições de IADA para ilustrar boas e menos boas aplicações dos conceitos e metodologias da estatística descritiva na exploração de conjuntos artefactuais. Todos os textos foram publicados em revistas científicas internacionais com revisão por pares e referem-se a trabalhos elaborados por alunos, investigadores e docentes da Universidade do Algarve.

Belmiro, J., Bicho, N., Haws, J., & Cascalheira, J. (2021). The Gravettian-Solutrean transition in westernmost Iberia: New data from the sites of Vale Boi and Lapa do Picareiro. Quaternary International, 587–588, 19–40. https://doi.org/10.1016/j.quaint.2020.08.027

Bicho, N., Cascalheira, J., Haws, J., & Gonçalves, C. (2018). Middle Stone Age Technologies in Mozambique: A Preliminary Study of the Niassa and Massingir Regions. Journal of African Archaeology, 16(1), 60–82. https://doi.org/10.1163/21915784-20180006

Coelho, E., Carlos Valera, A., & Carvalho, A. (2017). O concheiro do Meu Jardim (Nazaré) no contexto das estratégias de produção e circulação de suportes lâmino lamelares no Neolítico Médio da Estremadura Portuguesa. Journal of Lithic Studies, 4(3), 79–102. https://doi.org/10.2218/jls.v4i3.2533

Horta, P., Cascalheira, J., & Bicho, N. (2019). The Role of Lithic Bipolar Technology in Western Iberia's Upper Paleolithic: The Case of Vale Boi (Southern Portugal). Journal of Paleolithic Archaeology, 2(2), 134–159. https://doi.org/10.1007/s41982-019-0022-5

Marreiros, J., Bicho, N., Gibaja, J., Pereira, T., & Cascalheira, J. (2015). Lithic technology from the Gravettian of Vale Boi: New insights into Early Upper Paleolithic human behavior in Southern Iberian Peninsula. Quaternary International, 359–360, 479–498. http://dx.doi.org/10.1016/j.quaint.2014.06.074

Paixão, E., Marreiros, J., Pereira, T., Gibaja, J., Cascalheira, J., & Bicho, N. (2018). Technology, use-wear and raw material sourcing analysis of a c. 7500 cal BP lithic assemblage from Cabeço da Amoreira shellmidden (Muge, Portugal). Archaeological and Anthropological Sciences, 1–21. https://doi.org/10.1007/s12520-018-0621-y

Aula 11 - Exercício prático de estatística descritiva

Nesta aula, realiza-se o terceiro dos três exercícios práticos da unidade curricular. O principal objetivo é que os alunos, através da análise de uma base de dados arqueológicos fictícios, consigam produzir tabelas e gráficos que caracterizem os vários tipos de variáveis, tanto de forma isolada quanto combinada. Avalia-se o reconhecimento de alguns dos conceitos estatísticos mais relevantes (como média e moda), bem como as escolhas gráficas utilizadas na representação dos dados.

Enunciado do exercício nº 3 – Elaboração de tabelas e gráficos com dados quantitativos e qualitativos

A base de dados fornecida (Exercicio3.csv) contém informação fictícia relativa à análise de um conjunto de elementos de cerâmica provenientes de dois sítios arqueológicos distintos.
Utilizando os conhecimentos adquiridos nas aulas sobre tipos de variáveis, tabelas dinâmicas e construção de gráficos numa folha de cálculo, deverá responder a cada uma das questões abaixo apresentadas.
Todos os elementos criados (tabelas e gráficos) deverão ser inseridos num documento Word, e devem ser legendados de acordo com o que representam.

1. Construa uma tabela representativa da frequência absoluta, frequência relativa (%) e totais de cada um dos tipos de cerâmica nos vários contextos estratigráficos dos sítios El Sombrero e Playa Blanca.
2. Elabore um gráfico que ilustre claramente qual a Moda dos vários tipos de cerâmica em cada um dos dois sítios.
3. Apresente uma tabela que demonstre se existe uma relação entre o tipo de cerâmica e a cor da pasta identificada.
4. Elabore uma tabela para cada sítio com as três medidas de tendência central mais utilizadas e o desvio-padrão de todas as variáveis quantitativas contínuas presentes na base de dados.
5. Construa, para cada sítio, um gráfico que represente a dispersão dos dados das variáveis utilizadas no ponto anterior com base na mediana e nos quartis.
6. Crie dois gráficos que mostrem a relação bivariada entre as variáveis Largura_Boca/ Largura_Base e Espessura_Bordo/Espessura_Parede.
7. Utilizando a variável Tipo_Decoracao, crie uma tabela que identifique a contagem de cada um dos tipos nas várias camadas de cada um dos sítios, bem como a frequência relativa acumulada (%).

Aula 12 - Amostragem em Arqueologia

Nesta última aula teórico-prática da unidade curricular, o foco principal recai sobre o tema da amostragem e a importância de definir o que é uma população e uma amostra estatística, para melhor compreender e integrar os resultados da análise de um determinado conjunto de dados arqueológicos. O principal objetivo é assegurar que os alunos assimilem os vários conceitos teóricos relacionados com os processos de amostragem. Esta aula serve também como introdução a etapas mais avançadas na análise estatística, como a utilização de testes estatísticos para inferências.

A maioria dos arqueólogos vê a estatística apenas como uma ferramenta para analisar dados. Recolhem informação, analisam-na e depois interpretam o seu significado. No entanto, alguns arqueólogos não se apercebem de que a estatística também pode ser fundamental na decisão sobre que tipo de dados recolher e como planear um projeto de investigação (Orton, 1999). Isto pode dever-se ao facto de preferirem lidar com estatística o menos possível, mas levanta, na maior parte dos casos, problemas. Não são de todo incomuns as situações em que um arqueólogo fica frustrado quando um especialista em estatística lhe diz que recolheu o tipo errado de dados, que não recolheu dados suficientes ou que recolheu demasiados.. O tipo e quantidade de dados a recolher estão diretamente relacionados com um processo denominado de amostragem, que é de enorme importância para a Arqueologia, tal como para as demais ciências observacionais. Na maioria dos estudos arqueológicos, procura-se frequentemente desenvolver teorias e explicações que sejam generalizáveis para toda uma **população**. População não no sentido dos habitantes de uma cidade ou de um território, mas no sentido estatístico do termo (também conhecida como **população do estudo**): a totalidade daquilo em que estamos interessados, ou, em termos mais técnicos, todos os membros individuais de um determinado fenómeno de classe que pretendemos estudar.

Em Arqueologia, uma população pode ser diferentes coisas, mas incluí sempre todos os elementos que pertencem a uma determinada classe de fenómenos presentes num determinado contexto, como todos os artefactos presentes num sítio, estrato ou elemento; todos os sítios de uma região; ou todos os artefactos pertencentes a um determinado tipo, entre outros. (McCall, 2018). Contudo, embora queiramos conhecer as características de uma população inteira, é frequentemente impraticável, por vezes impossível, e metodologicamente problemático tentar recolher dados sobre todos os indivíduos de toda uma população. Por outro lado, não faz sentido dedicar mais recursos a um problema do que os estritamente necessários para o resolver

(Orton, 1999). Assim, torna-se evidente a importância de desenvolver uma estratégia de amostragem eficaz, de modo que a **amostra** selecionada (definida como o subconjunto mais pequeno de toda a população sobre o qual fazemos efetivamente inferências) seja representativa da população. É claro que os arqueólogos sempre utilizaram a amostragem, no sentido de que selecionaram sítios para escavar em regiões e definiram locais para unidades de escavação dentro de sítios, sem restringirem as suas conclusões a sítios ou unidades estratigráficas específicas. Por outro lado, o registo arqueológico por si só constitui uma amostra do que foram as diferentes populações do passado. Os tipos de amostra definidos por Orton (2000) ilustram muito bem estas diferenças:

1. **Amostras não intencionais**: Neste caso, a amostragem foi efectuada, por assim dizer, antes da chegada do arqueólogo ao local. Neste sentido, é sempre importante ter em consideração que o material recuperado numa escavação arqueológica não representa a totalidade do que foi perdido ou descartado no decurso das actividades realizadas num determinado local. Por vezes, o seu estado torna-o óbvio - por exemplo, os fragmentos de ossos encontrados representam uma fração do esqueleto que representam - mas outras vezes o seu estado pode sugerir erroneamente que se trata de uma amostra completa - por exemplo, no caso de um conjunto de moedas.
2. **Amostras informais**: A escolha pode basear-se em critérios arqueológicos ou em critérios de tempo, custo e conveniência. Existe um espectro de intencionalidade neste tipo de amostras: num extremo, encontram-se amostras intencionais, como, por exemplo, unidades de escavação cuidadosamente selecionadas com base em características topográficas ou levantamentos geofísicos. No outro extremo, situam-se amostras aleatórias ou de recolha, como por exemplo, a recolha apressada de alguns objetos encontrados à superfície de um potencial sítio arqueológico. Entre estes dois extremos, podem ser consideradas as amostras típicas, selecionadas pelo arqueólogo para representarem uma coleção específica de objetos. No entanto, o que falta a todas essas amostras é o potencial de generalização a partir delas, ou seja, a capacidade de extrapolar, de forma fiável, a descrição de uma amostra para afirmações sobre uma entidade mais vasta, normalmente designada como população.
3. **Amostras formais**: Estas representam amostras selecionadas de uma população bem definida e de acordo com procedimentos estatísticos também eles bem definidos. Quando realizada corretamente, este tipo de amostragem permite-nos realizar inferências válidas sobre uma população, incluindo estimar parâmetros específicos. Outro aspecto relevante é o facto de este tipo de amostragem possibilitar calcular margens de erro, que fornecem não só indicações sobre a utilidade provável da amostra,

mas também orientam na determinação da dimensão que a amostra deverá ter para alcançar-se mais segurança nas respostas às perguntas colocadas.

Estratégias de amostragem

O critério mais importante para uma amostra é a sua representatividade, ou seja, o subconjunto deve incorporar ou refletir de forma fiável o conjunto completo. Infelizmente, não é possível assegurar a representatividade absoluta de uma amostra sem compará-la com toda a população, o que, à partida, anularia o objetivo da amostragem. A abordagem mais eficaz consiste em tentar minimizar e quantificar a probabilidade de a amostra não ser representativa, estabelecendo inicialmente uma definição clara do que significa "representativo" em cada cenário específico. Foi este requisito que estimulou o desenvolvimento da teoria da amostragem estatística e a identificação de várias estratégias de amostragem que podem ser utilizadas. Entre estas, existem três estratégias principais de amostragem probabilística utilizadas pelos arqueólogos (Figura 34):

- **Amostragem aleatória simples:** Esta técnica é utilizada com o objetivo de compreender um sítio arqueológico ou uma coleção arqueológica como um todo, e não apenas as áreas ou componentes onde se espera encontrar algo. Normalmente, as áreas para amostragem são selecionadas aleatoriamente, recorrendo a uma tabela de números aleatórios. Atualmente, para minimizar o risco de enviesamento, são frequentemente utilizados programas informáticos para determinar as áreas de amostragem aleatória. Uma vez definida a área, determina-se a dimensão das unidades de amostragem, o número de unidades que se pretende e qual a área que se pretende amostrar. Obviamente, quanto maior a amostra, mais precisa será a previsão. Este método tem, naturalmente, algumas desvantagens. Em particular, no que diz respeito à sua aplicação a uma escavação arqueológica, é necessário definir previamente os limites do sítio arqueológico. Isto significa que tem de se ter a certeza da dimensão real do mesmo, o que nem sempre é fácil de saber antes do início da escavação. Em segundo lugar, esta atribuição aleatória de números pode fazer com que as áreas de amostragem sejam organizadas (aleatoriamente) em grupos, criando disparidades no tamanho das áreas amostradas e introduzindo um potencial enviesamento.
- **Amostragem aleatória estratificada:** Utilizando o exemplo de um trabalho de prospecção arqueológica para localização de novos sítios, este método envolve a divisão da região conforme as suas fronteiras naturais, tais como terras cultivadas, florestas,

margens de rios, etc. Os quadrados para a amostragem são selecionados utilizando o mesmo sistema de numeração aleatória empregue na Amostragem Aleatória Simples, mas com a distribuição proporcional à extensão de cada tipo de área natural. Assim, por exemplo, se 50% da região a ser amostrada estiver coberta de floresta, então 50% das zonas de amostragem aleatória serão atribuídas a esta área. Isto garante que a amostragem efectuada não favorece uma área em detrimento de outra,mantendo a proporcionalidade em relação à geografia do local.

- **Amostragem sistemática.** Neste método, tal como no exemplo de um trabalho de prospeção arqueológica, cria-se uma grelha com áreas espaçadas de forma uniforme para a amostragem. Por exemplo, pode-se estabelecer um espaçamento de dois em dois metros, ou de dois em dois quadrados. No caso da aplicação deste método, por exemplo, à recolha de amostras numa escavação arqueológica, pode-se decidir guardar um em cada cinco baldes de sedimento escavado, ou recolher amostras de 10 em 10 centímetros de um perfil estratigráfico. Este método é relativamente simples de implementar. Porém, o espaçamento regular pode levar a falhas ou acertos sistemáticos em todas as componentes de diagnóstico de um padrão igualmente regular, o que pode introduzir um enviesamento na amostragem.

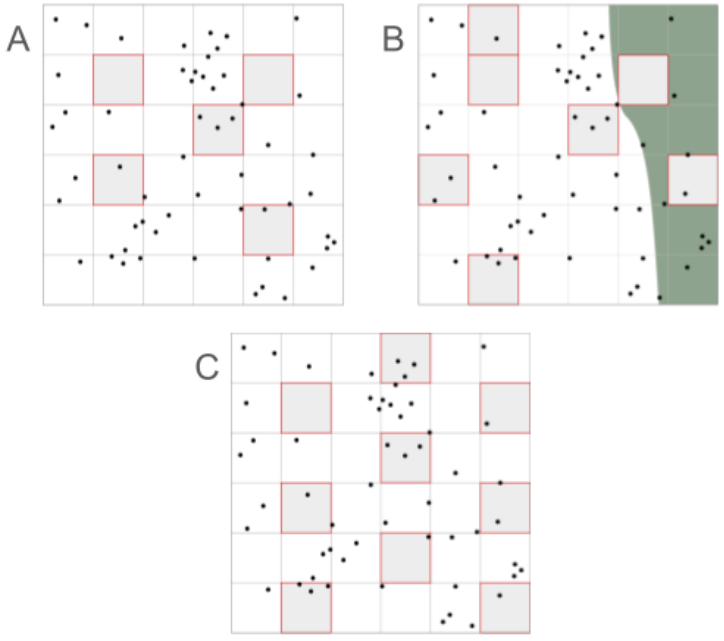


Figura 34. Principais estratégias de amostragem aplicados à seleção de unidades a inspecionar durante um trabalho de prospeção. A) Aleatória simples; B) Aleatória estratificada; C) Sistemática. Os pontos correspondem a sítios arqueológicos, e os quadrados a vermelho às unidades selecionadas para inspeção. Note-se o impacto dos vários métodos na deteção dos sítios arqueológicos.

Aula 13 - Exercício Final

Nesta aula, realiza-se o exercício final da unidade curricular. O principal objetivo é que os alunos apliquem a maior parte dos conceitos e técnicas explorados ao longo das aulas, com um foco especial na transformação de dados e na análise exploratória de diferentes tipos de variáveis.

Enunciado do exercício final

A base de dados fornecida consiste num conjunto de variáveis que dizem respeito ao registo de campo durante a escavação de um sítio arqueológico e à análise preliminar de alguns desses materiais. Utilizando esses dados, deverão responder a todos os pontos que se seguem para completar o exercício.
Os documentos finais deverão incluir a folha de cálculo devidamente transformada, bem como um ficheiro Word ou PDF com os elementos gráficos e tabelas solicitadas. Neste último caso, deverão também ser apresentadas legendas individuais que caracterizem de forma sucinta os dados representados por cada elemento.
Será avaliada não apenas a correta execução do exercício, mas também as escolhas realizadas ao nível da representação dos dados solicitados.

1. **Tópicos para transformação de dados:**
    a. Crie uma nova variável em que UNIT e ID apareçam juntos, separados por um hífen ("-"). Denomine esta variável de UNIT-ID.
    b. Todos os ID's de 2013 deverão ser apresentados sem os pontos a separar os conjuntos de caracteres.
    c. Os conteúdos da variável CODE não estão uniformizados em termos de formatação. Proceda à sua uniformização.
    d. Crie uma nova variável, denomine-a SAMPLING, e preencha as células dessa variável com os elementos YES ou NO, utilizando uma numeração aleatória.
    e. A variável Z foi registada com um erro: todos os pontos com valores superiores a 571.00 foram registados com 1.35 metros a mais. Crie uma nova variável, chame-lhe Z_MODIFIED, e corrija este erro.
    f. Uma das variáveis presentes na base de dados fornecida é denominada de SUFFIX e corresponde ao número de coordenadas registadas de uma peça com a mesma conjugação de UNIT-ID. Assim, quando o SUFFIX é 0, esta é a primeira coordenada registada do artefacto. Quando o SUFFIX é 1, significa que foi medida

uma coordenada extra no artefacto. Utilizando os valores de X, Y e Z destes artefactos, é possível calcular a ORIENTAÇÃO e INCLINAÇÃO do artefacto no momento da sua descoberta. Segundo McPherron (2005), estas duas medidas, destacadas a amarelo na tabela, podem ser calculadas utilizando as fórmulas apresentadas abaixo. Deverá criar e calcular estas duas novas variáveis na base de dados, utilizando todos os pontos que têm mais do que uma coordenada associada.

	A	B	C	D	E	F	G
1							
2							
3	ID	SUFFIX	X	Y	Z	ORIENTAÇÃO	INCLINAÇÃO
4	1	1	45.00	45.00	0.00		
5	1	2	45.05	45.05	0.03		
6	2	1	45.00	45.00	0.03	45,00	22,99
7	2	2	45.05	45.05	0.00		

ORIENTAÇÃO

=IF(E6>E7;MOD(DEGREES(ATAN2((D7-D6);(C7-C6)));360);MOD(DEGREES(ATAN2((D6-D7);(C6-C7)));360))

INCLINAÇÃO

=DEGREES(ATAN2(SQRT((C7-C6)^2+(D7-D6)^2);ABS(E7-E6)))

2. **Tópicos para exploração e apresentação de dados:**
    a. Elabore uma tabela geral que apresente as frequências absolutas e relativas (%) do total de artefactos de cada CODE encontrados em cada um dos LEVEL do sítio arqueológico.
    b. Apresente os dados acima mencionados na forma de gráfico.
    c. Represente graficamente a distribuição bivariada das coordenadas X e Y de todos os artefactos líticos do ano de 2015.
    d. Que categoria de LITHIC_TYPE está melhor representada na categoria QUARTZITE da variável LITHIC_RAW_MATERIAL? Apresente um gráfico ou tabela que evidencie este facto.
    e. Apresente, nas formas que considerar mais adequadas as medidas de tendência central mais relevantes das três dimensões (LENGHT, WIDTH e THICKNESS) para cada uma das LITHIC_TYPE.
    f. Compare, por LEVEL (A e Z), através de um boxplot, as mesmas três dimensões, mas apenas para a categoria FLAKE.

g. Que CERAMIC_TYPE são mais abundantes e como se comparam as frequências relativas acumuladas das suas distribuições pelos vários LEVEL do sítio arqueológico? Apresente um gráfico ou tabela que responda a estas perguntas.
h. Realize o mesmo procedimento da alínea anterior, mas desta vez para a variável FAUNA_SPECIES.

Referências citadas

Aldenderfer, M. (1998). Quantitative Methods in Archaeology: A Review of Recent Trends and Developments. *Journal of Archaeological Research*, *6*, 91–120.

Austin, A. (2014). Mobilizing Archaeologists: Increasing the Quantity and Quality of Data Collected in the Field with Mobile Technology. *Advances in Archaeological Practice*, *2*, 13–23.

Averett, E., Gordon, J., & Counts, D. (2016). *Mobilizing the Past for a Digital Future*. The Digital Press.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604).

Belmiro, J., Bicho, N., Haws, J., & Cascalheira, J. (2021). The Gravettian-Solutrean transition in westernmost Iberia: New data from the sites of Vale Boi and Lapa do Picareiro. *Quaternary International*, *587–588*, 19–40. https://doi.org/10.1016/j.quaint.2020.08.027

Bernatchez, J., & Marean, C. W. (2011). Total station archaeology and the use of digital photography. *SAA Archaeol. Rec*, *11*.

Bhargava, R., Deahl, E., Letouzé, E., Noonan, A., Sangokoya, D., & Shoup, N. (2015). *Beyond data literacy: Reinventing community engagement and empowerment in the age of data*.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., & Pfeiffer, T. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Cardoso, J. L., & Rolão, J. M. (1999/2000). Prospecções e escavações nos concheiros mesolíticos de Muge e de Magos (Salvaterra de Magos): Contribuição para a história dos trabalhos arqueológicos efectuados. *Estudos Arqueológicos de Oeiras*, *8*, 83–240.

Cascalheira, J., Bicho, N., & Gonçalves, C. (2017). A Google-Based Freeware Solution for Archaeological Field Survey and Onsite Artifact Analysis. *Advances in Archaeological Practice*, *5*, 328–339. https://doi.org/10.1017/aap.2017.21

Cascalheira, J., Gonçalves, C., & Bicho, N. (2014). Smartphones and the use of customized Apps in archaeological projects. *The SAA Archaeological Record*, *14*, 20–25.

Clarke, D. L. (1968). *Analytical archaeology*. Methuen.

Cukier, K. (2010). The data deluge: Businesses, governments and society are only starting to tap its vast potential. *The Economist*, *23*.

Ford, J. A. (1954). Comment on A. C. Spaulding, "Statistical Techniques for the Discovery of Artifact Types." *American Antiquity*, *19*(04), 390–391. https://doi.org/10.2307/277609

Harris, E. C. (1997). *Principles of archaeological stratigraphy* (2. ed., 3. print). Acad. Pr.

Johnson, M. (2019). *Archaeological theory: An introduction*. John Wiley & Sons.

Kansa, E., & Kansa, S. W. (2021). Digital Data and Data Literacy in Archaeology Now and in the New Decade. *Advances in Archaeological Practice*, *9*(1), 81–85. https://doi.org/10.1017/aap.2020.55

Karoune, E., & Plomp, E. (2022). Removing barriers to reproducible research in archaeology. *Zenodo, Ver. 5, Peer Reviewed and Recommended by Peer Community in Archaeology*.

Kaufmann, M., & Jeandesboz, J. (2017). Politics and 'the digital' From singularity to specificity. In *European Journal of Social Theory* (Vol. 20, Issue 3, pp. 309–328). SAGE Publications Sage UK: London, England.

Kintigh, K. W., Altschul, J. H., Beaudry, M. C., Drennan, R. D., Kinzig, A. P., Kohler, T. A., Limp, W. F., Maschner, H. D. G., Michener, W. K., Pauketat, T. R., Peregrine, P., Sabloff, J. A., Wilkinson, T. J., Wright, H. T., & Zeder, M. A. (2014). GRAND CHALLENGES FOR ARCHAEOLOGY. *AMERICAN ANTIQUITY*, *79*(1), 20.

Lock, G. (2003). *Using computers in archaeology: Towards virtual pasts*. Routledge.

Marwick, B. (2016). Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory*, *24*(2), 424–450. https://doi.org/10.1007/s10816-015-9272-9

McCall, G. S. (2018). *Strategies for quantitative research: Archaeology by numbers*. Routledge.

McPherron, S. P. (1994). *A reduction model for variability in Acheulian biface morphology* [PhD Thesis]. University of Pennsylvania.

McPherron, S. P., & Dibble, H. L. (2002). *Using computers in archaeology: A practical guide*. McGraw-Hill.

McPherron, S. P., & Dibble, H. L. (2003). *Using computers in archaeology: A practical guide*.

McGraw-Hill.

Mitchell, P. (2018). Introduction to Archaeological Methods and Sources. In *Oxford Research Encyclopedia of African History*. https://doi.org/10.1093/acrefore/9780190277734.013.367

Ogburn, J. L. (2010). The imperative for data curation. *Portal: Libraries and the Academy*, *10*(2), 241–246.

Orton, C. (1999). Plus ça change?–25 years of statistics in archaeology. *Archeology in the Age of the Internet, Oxford*.

Orton, C. (2000). *Sampling in archaeology*. Cambridge University Press.

Reed, D., Barr, W. A., McPherron, S. P., Bobe, R., Geraads, D., Wynn, J. G., & Alemseged, Z. (2015). Digital data collection in paleoanthropology. *Evol Anthropol*, *24*, 238–249. https://doi.org/10.1002/evan.21466

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), eabd1705.

Spaulding, A. C. (1953). Statistical Techniques for the Discovery of Artifact Types. *American Antiquity*, *18*(4), 305–313. https://doi.org/10.2307/277099

Tixier, J. (1963). Typologie de l'Epipaléolithique du Maghreb. Mémoires du Centre de Recherches Anthropologiques, Préhistoriques et Ethnographiques, No 2. *Paris: Arts et Métiers Graphiques*.

Trigger, B. G. (1989). *A history of archaeological thought*. Cambridge university press.