UNIVERSIDADE DE UBERABA

PROGRAMA DE EDUCAÇÃO A DISTÂNCIA

ESPECIALIZAÇÃO A DISTÂNCIA EM CAFEICULTURA IRRIGADA

UNIDADE DE ESTUDO V - METODOLOGIA DA PESQUISA

BLOCO TEMÁTICO II - CAMINHOS PARA A PESQUISA

AUTOR: DANIEL DURANTE PEREIRA ALVES

SUMÁRIO

1 Entendendo as várias Possibilidades do Método Científico

Uma vez que já vimos as características mais gerais da ciência, que a distinguem das outras formas de conhecimento, e que já vimos também as características mais gerais das diversas formulações do método científico, cabe agora nos aprofundarmos nas técnicas e métodos específicos ao tipo de pesquisa que nos interessa aqui. A saber, a pesquisa agropecuária em geral e, mais especificamente, a cafeicultura irrigada. Neste contexto, o aspecto metodológico mais importante para o desenvolvimento da pesquisa são os métodos e técnicas estatísticas, que tanto nos ajudam a projetar um experimento, como também a analisar e interpretar os dados obtidos deste experimento.

Mas antes disso, nos deteremos de maneira bastante rápida em uma questão fundamental que diz respeito a todas as diversas possibilidades de desenvolvimento do método. Possibilidades estas que se distinguem entre si na medida em que são diferentes os diversos tipos de problemas e questões que buscamos respostas através da atividade científica. Mesmo que concordem sobre os aspectos fundamentais do método científico, um antropólogo que deseje investigar, por exemplo, qual a auto-imagem dos menores abandonados que dormem na Praça da Sé, em São Paulo, usará procedimentos e métodos de pesquisa bastante diversos de um agrônomo interessado em compreender melhor os efeitos da adição de certos complementos minerais ao solo de uma cultura de café irrigado no município de Catalão-GO.

1.1 Pesquisa Qualitativa e Pesquisa Quantitativa

Uma distinção fundamental que se faz na metodologia da pesquisa científica diz respeito à utilização da matemática tanto para a produção quanto para a interpretação dos dados de uma pesquisa. Diz se que uma pesquisa é *quantitativa*, ou utiliza uma *metodologia quantitativa* quando a natureza dos dados desta pesquisa e as principais técnicas de análise destes dados são numéricas ou matemáticas. Quando a natureza tanto dos dados quanto das técnicas de análise destes não for numérica ou matemática, dizemos que se trata de uma *pesquisa qualitativa*.

Vejamos alguns exemplos:

- um engenheiro mecânico que pesquisa uma forma de diminuir o consumo de energia de um sistema de irrigação por pivô central lida com dados tais como velocidade, tempo, pressão, escoamento, aceleração, entre outros. Tais dados são expressados numericamente e exigem técnicas numéricas de análise e interpretação que caracterizam a sua atividade de pesquisa como *quantitativa*;
- um agrônomo interessado em comparar a eficiência de dois sistemas de irrigação de café distintos, trabalha com dados tais como: volume de água gasto por hora, taxa de crescimento, produtividade por pé de café, relação custo/benefício, entre outros. Estes, apesar de serem dados menos abstratos que os do exemplo anterior, também são em sua maioria expressos numericamente e analisados por métodos matemáticos e estatísticos, caracterizando assim a pesquisa como *quantitativa*;
- um antropólogo pesquisando a auto-imagem de menores abandonados muito dificilmente utilizará como ferramentas principais dados e métodos de análise numéricos. Como medir numericamente auto-imagem? Muito provavelmente o antropólogo utilizará em sua pesquisa categorias da psicanálise, desenvolverá histórias de vida, estudos de caso. Poderá fazer análises comparativas com a auto-imagem de crianças em outras situações. Talvez até utilize, em um argumento ou outro, dados e informações numéricas. Mas compreender a auto-imagem de um grupo social exige técnicas e métodos não matemáticos, característicos de uma pesquisa *qualitativa*.

É importante notarmos que o que distingue a pesquisa quantitativa da qualitativa não é a disciplina científica à qual a pesquisa se insere. Mesmo em ciências humanas faz-se muita pesquisa quantitativa. Por exemplo, um cientista político pode estar interessado em avaliar numericamente, em termos de porcentagem de apoio e rejeição, o impacto que a adoção pelo governo de certo programa econômico teria em uma determinada população. Uma pesquisa desse tipo exige métodos matemáticos e estatísticos, sendo portanto quantitativa. O que distingue fundamentalmente a forma de pesquisar quantitativa da qualitativa é o tipo de pergunta que o pesquisador quer responder e como ele encaminha seu trabalho para obter as respostas. Questões relacionadas ao significado dos fenômenos sociais, por exemplo, que levem em conta motivações, crenças, valores são inadequadas para serem tratadas por métodos matemáticos e estatísticos. Elas representam o universo onde se insere a pesquisa qualitativa.

Este universo é recente na história das ciências. Costumava-se considerar científico apenas o conhecimento obtido através dos métodos quantitativos de pesquisa, que possuem uma uniformidade e objetividade tal, que podemos dizer que o método quantitativo é único para todas as disciplinas científicas. Tem por base a filosofia positivista de Auguste Comte e está "apoiado na experimentação, mensuração e controle rigoroso dos dados (fatos), tanto nas ciências naturais como nas ciências humanas. [...] Sua objetividade seria garantida pelos instrumentos e técnicas de mensuração e pela *neutralidade* do próprio pesquisador frente à investigação da realidade"(**PÁDUA[1998]**, p.31).

O que importa é termos em mente que os métodos e técnicas de pesquisa quantitativa que veremos a seguir não são os métodos e técnicas de toda e qualquer atividade científica. São apenas os métodos mais adequados ao tipo de pesquisa que normalmente se faz na área agrícola. Há ciência fora dos procedimentos estatísticos padrões. O fator fundamental que determina o tipo de metodologia que deve ser adotada em uma pesquisa científica corresponde à identificação do tipo de problema ou questão que se deseja compreender melhor através da pesquisa. Além deste fator, podemos acrescentar também as convicções filosóficas do pesquisador. Este no entanto é um fator secundário, uma vez que a questão geradora e hipóteses de uma pesquisa científica já carregam embutidas as convicções filosóficas de seus proponentes que são relevantes para a pesquisa.

1.1.2 Exercício

Descreva sucintamente duas pesquisas científicas que possam ser caracterizadas uma como pesquisa qualitativa e a outra como pesquisa quantitativa.

2 Uma Metodologia Quantitativa da Investigação Adequada à Pesquisa Agropecuária

O Ponto de partida para a atividade científica é um problema, uma lacuna, uma questão para a qual não temos uma resposta. É a chamada *questão geradora* da pesquisa. Mas a questão geradora não é suficiente. É necessário mais: uma *hipótese*. Uma hipótese é uma resposta provisória, uma estimativa, uma crença do pesquisador sobre qual seria a solução para a questão geradora. Mas enquanto estimativa e crença, uma hipótese não tem valor científico. Não é um produto da ciência. Uma vez que o pesquisador possui uma hipótese, o próximo passo representa a elaboração de uma pesquisa que possa confirmar ou rejeitar a hipótese. O objetivo de toda a atividade de pesquisa é este: transformar hipóteses em teses. Em outras palavras, tornar crenças incertas, palpites, em resultados mais seguros e confiáveis que possam carregar o status de conhecimento científico, de teses científicas.

Para a grande maioria das pesquisas em agropecuária, especialmente para as ligadas aos fatores físicos e biológicos dos cultivares, é conveniente adotarmos a postura metodológica defendida por Karl Popper. Segundo a sugestão de Popper, uma pesquisa bem projetada é uma tentativa de falseamento da hipótese. Pesquisar é testar hipóteses. Se uma hipótese resiste a severos testes, severas tentativas de falseamento, então é porque ela é forte o suficiente para tornar-se uma tese e adquirir o status de conhecimento científico. Dessa forma, segundo (**REY[1998]**, p. 31), "o planejamento de uma pesquisa consiste na elaboração de um plano de observação, ou de experimentação, destinado a contestar determinada hipótese, por mais justa e sólida que possa parecer."

Considere a seguinte hipótese:

(2.1) a adição de 5% de compostos orgânicos à água de irrigação de uma cultura jovem de mudas de café aumenta em 50% a velocidade de crescimento das mudas. Chamemos de X a variável que indica a presença de 5% de compostos orgânicos na água de irrigação, e de Y a variável que mede a taxa de crescimento das mudas. Nossa hipótese então representa uma relação causal entre X e Y. Podemos rescrevê-la como: na presença de X, Y deve a aumentar 50%.

Elaborar uma pesquisa que teste a hipótese 2.1 é elaborar uma experiência que possa verificar que esta relação entre X e Y sempre se dá. Deve-se realizar experimentos onde há ausência de X e onde há a presença de X sob circunstâncias diversas. O pesquisador deve tentar imaginar alguma possibilidade em que esta relação não se daria e verificar testando-a.

Apenas como notação, chama-se as variáveis como X, que precedem a outra e que devem ser manipuladas pelo pesquisador para verificar a hipótese de *variáveis independentes*, e as variáveis como Y, que podem ter seus valores alterados conforme a alteração de X, de *variáveis dependentes*.

2.1.1 Exercício

Atribua variáveis às hipóteses abaixo, identificando as variáveis independentes e as dependentes.

a) O cruzamento da espécie (A) de café com a espécie (B) produz um café híbrido (AB) com produtividade (grãos por planta) 20% superior à produtividade do café da espécie (A) e 30% superior à do café (B).

b) O tratamento de uma lavoura de milho com o composto (C) diminui, em 2 meses de aplicação, a incidência da praga (P) em 85%.

c) Para manter a produtividade de lavouras irrigadas do café (A), o método de gotejamento necessita de 40% menos água que o método de pivô central e 70% menos que o método de asperção.

2.2 Pesquisa Observacional e Pesquisa Experimental

Podemos caracterizar uma pesquisa quanto a sua natureza sob duas categorias: *pesquisa observacional*, que é própria de fenômenos em que o pesquisador não pode influir nem produzir experimentos, mas tem que se limitar a observações, e a *pesquisa experimental*, na qual o pesquisador, para verificar sua hipótese, pode interferir nos valores da variável independente.

Por exemplo, se estivermos interessados em verificar como o regime de chuvas afeta a incidência de pragas nas culturas de milho de uma certa região, não há nada que possamos fazer a não ser observar. Não temos o poder de manipular a variável independente (quantidade de chuva). Esta será portanto uma pesquisa observacional. Já a verificação da aceleração de crescimento de mudas devido a adição de componentes orgânicos à água de irrigação pode e deve ser desenvolvida como uma pesquisa experimental, pois é possível manipular a variável independente (presença de compostos orgânicos na água).

2.3 Planejando uma Pesquisa Experimental

Como já vimos, o que caracteriza a pesquisa experimental ou *experimentação* é a possibilidade de manipulação das variáveis independentes pelo pesquisador. Assim, a primeira tarefa do pesquisador é descobrir, através de sua hipótese, quais parâmetros devem ser medidos, reescrever sua hipótese em termos de variáveis e identificar dentre essas quais serão submetidas a manipulação, representando as variáveis independentes, e quais serão as variáveis dependentes.

É fundamental que o projeto da pesquisa permita que se façam comparações, a fim de sabermos como as variáveis dependentes reagiram às manipulações impostas às variáveis independentes. Estas comparações serão cruciais para testarmos nossa hipótese.

A diversidade dos fenômenos naturais e das condições de experimentação exigem uma grande variedade de tipos de planejamento. Apenas a título de ilustração, seguem alguns dos exemplos mais comuns.

2.3.1 Pesquisas Com Um Só Grupo Experimental

Neste tipo de pesquisa há um único grupo de sujeitos pesquisados no qual se fará uma ou um conjunto de repetidas observações preliminares. Em seguida aplica-se um incremento à variável independente e faz-se novas observações no grupo para avaliar os efeitos de tal manipulação. Este tipo de pesquisa é muito comum quando se deseja avaliar a resposta de determinados indivíduos a alguma droga ou tratamento.

2.3.2 Pesquisas Com Grupos Experimentais Independentes

Aqui ocorrem dois grupos pesquisados, independentes, que devem se submeter a condições distintas de alteração da variável independente. Por exemplo, podemos projetar uma pesquisa para testar a hipótese 2.1 separando dois grupos experimentais, submetidos a exatamente as mesmas condições, com exceção que no primeiro, em um dado momento, adicionaremos componentes orgânicos à água da irrigação e no outro não. Ao final de um período comparamos as alturas das mudas dos dois grupos. A hipótese estará confirmada se as mudas do primeiro grupo forem em média 50% maiores que as mudas do segundo. Caso isso não ocorra a hipótese está refutada.

2.3.3 Pesquisas Com Tratamentos Múltiplos e uma Só Variável Independente

Suponha que queiramos verificar como varia o crescimento de mudas de café irrigado conforme a concentração de componentes orgânicos na água de irrigação. Para isso podemos separar, por exemplo, 4 grupos pesquisados sendo que a eles, aplicamos, respectivamente, concentrações de 5%, 10%, 15% e 0% de componentes orgânicos na água de irrigação destes grupos. Fazendo observações preliminares e após um período das aplicações, temos meios de avaliar como a concentração de componentes orgânicos influencia no crescimento das mudas. Este é um caso de pesquisa com múltiplos tratamentos e uma só variável independente.

2.3.4 Pesquisas com Múltiplas Variáveis Independentes

Neste caso mede-se, em um mesmo experimento, os efeitos de múltiplas variáveis independentes. É claro que quanto maior o número de variáveis independentes, maior o número de combinações possíveis para os casos a serem analisados. Considere um experimento com duas variáveis independentes A e B, sendo que nos interessam duas condições de análise (valores) para cada uma (A_1, A_2 e B_1, B_2). Temos dessa forma 4 possibilidades diferentes para o conjunto dos valores de A e B. São elas:

ocorrem simultaneamente A_1 e B_1;
ocorrem simultaneamente A_1 e B_2;
ocorrem simultaneamente A_2 e B_1;
ocorrem simultaneamente A_2 e B_2.

Este tipo de pesquisa permite a identificação de possíveis interações de fatores diversos em um dado fenômeno. Se, por exemplo, quisermos verificar as interações dos fatores *componentes orgânicos* e *quantidade de água* sobre o crescimento de mudas de café irrigado, podemos planejar uma pesquisa com as variáveis independentes *quantidade de água* e *porcentagem de componentes orgânicos na água*, e analisar os resultados que as várias combinações possíveis destes fatores têm no crescimento das mudas.

2.3.5 Exercícios

Quais, dentre os quatro tipos de modelagem experimental apresentados acima, são mais adequados para as seguintes situações:

a) Deseja-se verificar como varia a produtividade de uma lavoura de café de acordo com a quantidade de água disponível para irrigação.

b) Deseja-se investigar os efeitos que a aplicação da droga (X) tem em uma população bovina.

c) Deseja-se verificar qual dos métodos de irrigação (gotejamento ou pivô central) é mais eficiente (necessita de menos água).

d) Deseja-se verificar quais as opções mais produtivas com relação à quantidade de água e porcentagem de componentes orgânicos na água de irrigação de uma lavoura de café

2.4 Técnicas de Amostragem

Chamamos de *população* ou *universo* o total de elementos de um conjunto, como por exemplo todos os pés de café de uma lavoura. Uma *amostra* representa um subconjunto qualquer de uma população. Por exemplo, 5 pés de café desta lavoura, tomados ao acaso, representam uma amostra. Os 10 pés de café mais altos da lavoura representam outra amostra.

Quando fazemos pesquisa, em muitos casos não é possível obter dados de toda uma população. Quando isto ocorre temos que trabalhar com uma amostra. Escolher uma amostra que seja representativa de toda a população para o desenvolvimento da pesquisa é tarefa bastante importante. Para tal existem certas *técnicas de amostragem*, que dependem da natureza da população e dos dados que se quer obter. Aqui vão algumas delas.

2.4.1 Amostra Casual Simples

Composta por elementos selecionados ao acaso na população, de modo que todo elemento da população tem igual probabilidade de compor a amostra. É recomendada quando a população a ser estudada não se encontra organizada sob nenhum aspecto.

2.4.2 Amostra Sistemática

Composta por elementos selecionados de acordo com algum sistema. Por exemplo, um sistema pode ser o seguinte: em uma lavoura de milho, toma-se para a amostra todos os pés das filas pares cuja ordem da esquerda para a direita é um número múltiplo de 3. Esquematicamente teríamos:

fila 1: 01 02 03 04 05 06 07 08 09 10 11 12 13
fila 2: 01 02 03 04 05 06 07 08 09 10 11 12 13
fila 3: 01 02 03 04 05 06 07 08 09 10 11 12 13
fila 4: 01 02 03 04 05 06 07 08 09 10 11 12 13
fila 5: 01 02 03 04 05 06 07 08 09 10 11 12 13

Esta técnica de amostragem é recomendada quando a população a ser estudada encontra-se organizada de alguma forma.

2.4.3 Amostra Estratificada

Composta por elementos pertencentes a todos os estratos da população. Os estratos são subgrupos da população reunidos de acordo com certos valores de uma variável. Por exemplo, podemos estratificar uma população de pés de milho de acordo com o número de espigas que eles apresentam. Assim, podemos considerar que os pés que têm de 0 a 2 espigas estão no estrato 1, os que têm de 3 a 4 estão no estrato 2 e assim por diante. Uma amostra estratificada desta população deve conter pés de milho de todos os estratos. Esta técnica é recomendada quando a característica que define os extratos é importante em nossa pesquisa.

2.4.4 Amostra por Conveniência

Composta por elementos selecionados de acordo com critérios não objetivos de conveniência. Suponha que se queira pesquisar a população de pássaros que está atacando uma determinada lavoura. Prepara-se então algumas armadilhas ao longo da lavoura. Os pássaros capturados representam uma amostra da população que ataca a lavoura composta por conveniência — o fato de terem sido capturados. Este tipo de amostra é bastante criticado pelos estatísticos pelo risco de não ser representativa de toda a população, devendo ser evitada sempre que possível. No entanto, em casos como o do exemplo, não há escolha.

3 A Estatística como Ferramenta da Pesquisa Quantitativa

Os métodos e técnicas estatísticos ocupam importante papel na atividade científica em geral, sendo fundamentais para a pesquisa quantitativa. É, de fato, a utilização destes métodos e técnicas que caracteriza uma atividade de pesquisa como quantitativa.

Segundo (**REY[1998]**, p. 45),

> "(...) a aplicação da estatística ao tratamento dos resultados de um experimento ou de uma observação científica compreende dois aspectos. O primeiro é a apresentação dos dados de forma sistematizada, clara e sintética, a fim de que se possa entender ou visualizar o comportamento das variáveis observadas e, eventualmente, alguma relação que haja entre elas. Este é o aspecto descritivo da estatística (...). O segundo é aquele que permite analisar os resultados e fazer inferências, para a tomada de decisões quanto à aceitação ou rejeição das hipóteses formuladas: essa é a função da inferência estatística".

Ou seja, a estatística é útil tanto para apresentar os dados de uma pesquisa quanto para ajudar na decisão sobre a validade de uma hipótese. A estas duas funções podemos acrescentar também a importância da estatística para o projeto dos experimentos de uma pesquisa.

Nosso objetivo aqui é apresentarmos uma breve descrição de certas técnicas estatísticas que auxiliem a atividade do pesquisador, tendo sempre como base a pesquisa agropecuária. Não temos a pretensão de sermos completos em nossa exposição, que, como dissemos, tem apenas o caráter introdutório. Três textos foram recorrentemente utilizados por nós durante a produção deste material e devem ser consultados pelo estudante em caso de necessidade de aprofundamento do que aqui expomos. São eles: **GOMES[1984]**, **REY[1998]** e **VIEIRA[1980]**.

3.1 Distribuição Aleatória

Em uma lavoura de café onde todos os pés são da mesma espécie, nascidos de sementes idênticas, plantados em solo com as mesmas características e sujeitos às mesmas condições climáticas, poderíamos logicamente supor que todos estes pés de café devessem ter um desenvolvimento idêntico, possuindo todos, por exemplo, a mesma altura. No entanto, sabemos que isso não ocorre. As variações nas alturas entre cada um dos pés de café de uma tal lavoura são devidas a um enorme conjunto de fatores que não podemos controlar nem ao menos perceber.

Este tipo de variação, que não possui uma causa definida, é chamada pelos estatísticos de *variação aleatória* e possui propriedades específicas bastante estudadas, que são de profunda utilidade na atividade do pesquisador. Por mais que a ciência avance na compreensão dos fenômenos, suas causas e relações, sempre haverá espaço para a aleatoriedade, e praticamente todos os tipos de medidas que os cientistas realizam estão sujeitas à variação aleatória. No caso das pesquisas agropecuárias, onde estão envolvidos fatores geográficos, biológicos, físicos, químicos, e muitos outros, a variação aleatória é sempre significativa e deve ser considerada com cuidado.

Quando, para testar a hipótese 2.1, por exemplo, utilizamos dois grupos experimentais e dizemos que o valor da variável Y (altura das mudas de café) no grupo A é por exemplo 40 cm e no grupo B é 30 cm, isto significa que estes números 40 e 30 foram obtidos através de medidas em muitos pés de café do grupo. São números que representam uma tendência de cada grupo à aquela altura média. As alturas reais de cada uma das mudas do grupo variam aleatoriamente em torno destes números. Além disso, junto com os números 40 e 30 precisamos informar quão longe ou perto deles as alturas reais dos pés estão. Qual o desvio médio dos pés de café com relação a estes números, ou seja, quão homogêneas são as alturas das plantas de cada grupo.

Entender como atribuir valores a variáveis, como estimar a confiança e representatividade dos dados obtidos, quão grande deve ser uma amostra a fim de que ela seja suficientemente representativa de uma população são as principais funções da estatística descritiva sobre as quais trataremos agora.

3.2 Medidas de Tendência Central

Uma *medida de tendência central* representa, como o nome sugere, um valor em torno do qual os dados de um conjunto se distribuem. A mais comum delas é a média aritmética. Também importantes são a mediana e a moda.

Quando, em uma pesquisa, utilizamos uma amostra (ou toda a população) para atribuir valor a uma variável, tal como a altura das mudas de café de uma lavoura, o valor que a variável deve assumir corresponde a uma medida de tendência central dos valores desta amostra. Em geral utilizamos a média aritmética.

3.2.1 Média Aritmética

A *média aritmética* ou simplesmente *média* representa a soma de todos os valores de um conjunto numérico dividido pelo total de elementos do conjunto.

Considere por exemplo uma amostra casual de 6 pés de café de uma lavoura, com as seguintes alturas em centímetros: 176, 214, 267, 198, 235, 212. A média aritmética desta amostra é:

$$m = \frac{176+214+267+198+235+212}{6} = 217$$

Utilizaremos a letra maiúscula *M* para representar a média de toda uma população *N*. A letra minúscula *m*, em itálico, indicará a média de uma amostra extraída da população.

A média aritmética é então obtida através da seguinte fórmula:

$$M=\frac{\sum X}{N}$$, onde $$\sum X=X_1+X_2+\cdots+X_N$$.

3.2.2 Mediana

A *mediana* representa o valor central de uma seqüência de dados, quando esta seqüência tem uma quantidade impar de elementos. Caso a seqüência tenha uma quantidade par de elementos, a mediana representa a média aritmética dos dois elementos centrais.

Considere, por exemplo as duas seqüências seguintes,

X: 123, 214, 175, 91, 188 e *Y*: 214, 123, 214, 175, 91, 188

Colocando-as em ordem crescente temos:

X: 91, 123, 174, 188, 214 e *Y*: 91, 123, 174, 188, 214, 214

Dessa forma, a mediana da seqüência *X* é 175, que é o valor central, e a mediana da seqüência *Y* é (174+188)/2 = 181, que corresponde à média aritmética dos seus dois valores centrais.

Denotaremos a mediana por *Me*.

3.2.3 Moda

A *moda* é o valor que ocorre com maior freqüência em uma seqüência de números.

Assim, na seqüência 3, 4, 4, 6, 6, 6, 8, 9, 9, a moda é 6.

Ao passo que a seqüência *X*: 91, 123, 174, 188, 214 não tem moda, pois nenhum valor se repete.

Já a seqüência 3, 4, 4, 6, 6, 6, 8, 9, 9, 9, tem duas modas, 6 e 9.

Denotaremos a moda por *Mo*.

3.2.4 Exercício

Considere a tabela seguinte como representando uma pequena lavoura de café, onde os números representam as alturas dos pés em centímetros. Calcule a média aritmética, a mediana e a moda das duas amostras sistemáticas, uma delas constituída pelos números marcados de vermelho e a outra pelos números marcados com amarelo.

098	123	112	136	087	099	107	124	085	102	114	097
096	114	135	128	109	097	136	127	110	094	102	107
085	103	118	129	127	136	142	126	112	135	107	103
101	100	145	127	114	104	116	122	098	099	103	111
081	096	092	113	121	132	111	127	097	100	120	098
079	087	101	099	103	110	125	111	103	098	101	085

3.3 Verificação da Aleatoriedade de Uma Distribuição

Além de informarem os pontos em torno dos quais os valores de uma amostra se distribuem, as medidas de tendência central nos dão uma outra informação bastante importante. Quando tomamos uma amostra razoavelmente grande, a estatística nos diz que em uma distribuição aleatória, os valores da média aritmética *M*, mediana *Me* e moda *Mo* devem coincidir. Ou seja, devemos ter $M = Me = Mo$. Não se trata de obter uma igualdade absoluta, mas a diferença entre eles não pode ser significativa. Se estes valores diferirem significativamente, isso quer dizer que a distribuição dos valores da amostra não é aleatória, devendo haver alguma causa identificável que provoca esta distorção.

Tal fato é bastante útil para identificarmos certos fatores que podem estar afetando os dados de uma pesquisa. Identificados estes fatores temos duas opções: ou os isolamos, mantendo-os inalterados, ou, caso não seja possível, associamos uma variável independente a eles e medimos sua variação no decorrer da pesquisa.

3.4 Medidas de Dispersão

Apesar de as medidas de tendência central corresponderem a valores em torno dos quais os demais se distribuem, elas não têm poder de expressar quão agrupados ou dispersos em torno deste valor central os demais estão. Olhando apenas para a média aritmética das alturas de pés de café de uma lavoura, não temos condição de saber se a maioria dos pés de café tem altura próxima à média ou se a altura dos pés da lavoura varia muito. Este outro tipo de informação é obtida através das *medidas de dispersão*, sendo o desvio padrão, a variância e o coeficiente de variabilidade as mais utilizadas.

Assim, quando em uma pesquisa utilizamos a média para atribuir valor a uma variável, temos também que informar qual a dispersão dos valores medidos em torno desta média. Esta informação, veremos mais adiante, será útil para calcularmos a confiança e representatividade dos dados da pesquisa. A variância, o desvio padrão e o coeficiente de variabilidade são as ferramentas estatísticas de que dispomos para calcularmos esta dispersão.

3.4.1 Variância, Desvio Padrão e Coeficiente de Variabilidade

Como os dados de um conjunto se distribuem em torno da média, nada mais natural do que medir o grau de dispersão de um conjunto de dados através do desvio destes dados em relação à média, ou seja, através da diferença destes dados em relação à média. A *variância* nada mais é do que a média dos quadrados dos desvios individuais e o *desvio padrão* corresponde à raiz quadrada da variância.

Assim, considerando X uma seqüência de N números temos que a variância, que denotaremos por S^2, é definida pela seguinte fórmula:

$$\boxed{S^2 = \frac{\left(X_1 - M\right)^2 + \left(X_2 - M\right)^2 + \cdots + \left(X_N - M\right)^2}{N}}$$

Ou, mais resumidamente:

$$S^2 = \frac{\sum_{i=1}^{N} \left(X_i - M \right)^2}{N}$$

Logo, o desvio padrão é dado por:

$$S = \sqrt{S^2}$$

Vejamos um exemplo. Considere o seguinte conjunto de dados:

X: 123, 214, 175, 91, 188

Note que $M = \frac{123+214+91+175+188}{5} \quad \Rightarrow \quad M = 158{,}2$

Assim, os desvios de cada um dos valores correspondem a:

M - 123 = 158,2 - 123 = 35,2

M - 214 = 158,2 - 214 = -55,8

M - 91 = 158,2 - 91 = 67,2

M - 175 = 158,2 - 175 = -16,8

M - 188 = 158,2 - 188 = -29,8

Portanto, a variância S^2 é dada por:

$$S^2 = \frac{35{,}2^2 + (-55{,}8)^2 + 67{,}2^2 + (-16{,}8)^2 + (-29{,}8)^2}{5}$$

$$S^2 = \frac{1239{,}04 + 3113{,}64 + 4515{,}84 + 282{,}24 + 888{,}04}{5}$$

S^2 = 2007,76

O desvio padrão é portanto a raiz quadrada de 2007,76, ou seja:

$$S = \sqrt{2007{,}76} = 44{,}80$$

Note, pela fórmula que define o desvio padrão, que este representa uma grandeza expressa na mesma unidade dos valores do conjunto de dados. Para que possamos comparar o grau de dispersão de diferentes conjuntos de dados, lançaremos mão do *coeficiente de variabilidade* (*CV*), que será definido como uma grandeza absoluta, ou seja, independente da unidade.

Definimos o coeficiente de variabilidade *CV* como o valor percentual que o desvio padrão *S* representa em relação à média *M*. Ou seja:

$$\boxed{CV=\frac{100\times S}{M}}$$

Vejamos um exemplo da importância do coeficiente de variabilidade.

Considere os dois seguintes grupos de dados que representam a idade em anos de 6 indivíduos:

3, 1, 5 e 53, 51, 55

No primeiro grupo a média de idade é de 3 anos, e no segundo de 53.

Os dois grupos têm a mesma dispersão de dados, com desvio padrão *S* = 2.

Acontece que as diferenças de idade no primeiro grupo são muito mais significativas do que no segundo, pois representam nos indivíduos do primeiro grupo um percentual muito maior em relação às suas idades do que no segundo. É claro que a diferença de idade entre um bebê de 1 ano e uma criança de 5 é muito mais significativa do que a diferença de idade entre dois adultos de 51 e 55 anos. É este tipo de informação que o coeficiente de variabilidade capta. Veja:

No primeiro grupo temos: $CV=\frac{100\times 2}{3}=66,67$

Já no segundo grupo: $CV=\frac{100\times 2}{53}=3,77$

O coeficiente de variabilidade dos dados do primeiro grupo é muito maior do que no segundo, apesar dos dois grupos possuírem desvios padrão com valores idênticos e expressos na mesma unidade (anos).

3.4.2 Medidas de Dispersão sobre Amostras e Graus de Liberdade

Note que para os cálculos do desvio padrão, da variância e do coeficiente de variabilidade (respectivamente S, S^2 e CV) apresentados acima, utilizamos a média absoluta M, que como definimos em 3.2.1, é calculada a partir de todos os indivíduos de uma população.

Acontece que, na prática das pesquisas, raramente tem-se o valor da média M. Normalmente trabalha-se com a média amostral m, obtida de uma amostra que tomamos como representativa da população. Neste caso, o cálculo da variância e do desvio padrão se alteram um pouco, em função do que chamaremos de *graus de liberdade*, para refletir a diminuição da confiabilidade que o trabalho com a amostra representa. Estes novos desvio padrão e variância calculados através de amostras são denotados por s e s^2 (letras minúsculas). Assim, considerando uma amostra X com N elementos e média m, temos que as novas fórmulas para desvio padrão e variância são:

$$\boxed{s^2 = \frac{\sum (X_i - m)^2}{N-1}}$$

$$\boxed{s = \sqrt{s^2}}$$

Note que além da média amostral m no lugar de M, no denominador da expressão de s^2, diferentemente do cálculo de S^2, ocorre a subtração de uma unidade do tamanho da amostra.

Dizemos que um sistema baseado em uma amostra com N elementos possui (N-1) *graus de liberdade*. Em sistema que não é baseado em amostra, ou seja, cujos cálculos são feitos sobre todos os elementos da população, dizemos que há infinitos graus de liberdade.

3.4.3 Exercício

Considere a mesma tabela do exercício anterior, que representa uma pequena lavoura de café, onde os números correspondem às alturas dos pés em centímetros. Considerando as médias das amostras vermelhas e amarelas, calculadas anteriormente, calcule os valores do desvio padrão, da variância e do coeficiente de variabilidade para cada uma das amostras.

098	123	112	136	087	099	107	124	085	102	114	097
096	114	135	128	109	097	136	127	110	094	102	107
085	103	118	129	127	136	142	126	112	135	107	103
101	100	145	127	114	104	116	122	098	099	103	111
081	096	092	113	121	132	111	127	097	100	120	098
079	087	101	099	103	110	125	111	103	098	101	085

3.5 Interpretando as Medidas de Tendência Central e Dispersão

Veremos agora como as medidas de tendência central e de dispersão são úteis para atividade de pesquisa, informando sobre a confiança e representatividade dos dados. Novamente alertamos que não entraremos em maiores detalhes estatísticos. Para os que desejem se aprofundar no assunto, além da literatura já recomendada sugerimos **LITLE & HILLS[1976]**.

Considere uma amostra de 11 plantas escolhidas de uma lavoura de café. Seja 176cm a altura média de pés de café da amostra, com desvio padrão de 21,3 cm. Temos: m=176 cm, s= 21,3 e N=11.

Mas como, com estes dados, responder às seguintes perguntas?

- Dentro de que limites variam as alturas dos pés de café da lavoura em sua maioria?
- Qual a probabilidade de haver nesta lavoura pés de café com alturas superiores a 230cm? Ou inferiores a 86 cm?

As respostas a estas perguntas são obtidas estatisticamente utilizando-se das propriedades matemáticas da famosa *curva de distribuição normal*, ou *curva de gauss*. A Curva de gauss corresponde a uma função matemática que descreve, em termos ideais, o comportamento dos fenômenos com distribuição aleatória.

A primeira pergunta acima está procurando por um intervalo de confiança. Podemos refazê-la, mais especificamente, da seguinte maneira:

- Quais seriam os limites de altura entre os quais 95% dos pés de café da lavoura se encontram?

A resposta a esta pergunta corresponde ao que chamamos de *intervalo de confiança de 95%* de uma amostra. Ela é obtida através de uma equação simples, envolvendo a média, o desvio padrão e os graus de liberdade. Chamemos de L_i o limite inferior e L_s o limite superior. Temos:

$$\boxed{L_i = m - (t_0 \times s)} \quad \text{e} \quad \boxed{L_s = m + (t_0 \times s)}$$

Onde t_0 é obtido através da consulta a uma tabela que depende do número de graus de liberdade do sistema, do grau de confiança desejado e está presente na maioria dos livros de estatística. Apresentamos a seguir uma versão resumida desta tabela retirada de (**GOMES[1987]**, p. 19).

N.° de graus de liberdade	Valores de t_0
2	4,30
4	2,78
10	2,23
15	2,13
20	2,09
30	2,04
40	2,02
60	2,00
120	1,98
∞	1,96

Note que a tabela que apresentamos vale apenas para um grau de confiança de 95%. Para outros graus de confiança deve-se consultar outras tabelas.

Como, em nosso exemplo N=11, temos 10 graus de liberdade e portanto, t_0=2,23 e os valores para L_i e L_s são:

$L_i = 176 - (2,23 \times 21,3) = 128,5$

$L_s = 176 + (2,23 \times 21,3) = 223,5$

Dessa forma, respondendo à nossa pergunta, diríamos: 95% dos pés de café da lavoura têm altura superior a 128,5 cm e inferior a 223,5 cm. Ou, dito de outra forma, a probabilidade de um pé de café desta lavoura ter altura entre 128,5 cm e 223,5 cm é de 95%.

Esta informação já nos diz algo sobre a segunda pergunta (as probabilidades de haverem pés de café de alturas superiores a 230 cm ou inferiores a 86 cm). Como estes valores estão além dos limites do intervalo de confiança de 95%, certamente há menos de 2,5% de probabilidade de haverem pés de café menores que 86 cm, e também menos de 2,5 % de probabilidade de haverem pés de café maiores que 230 cm.

3.5.1 Exercício

Considere novamente a tabela dos exercícios anteriores, que traz as alturas de uma pequena lavoura de café. Arredondando o tamanho das amostras vermelha e amarela para 10 graus de liberdade, obtenha o valor de t0 da tabela acima e calcule os limites inferior e superior do intervalo de confiança para essas duas amostras. Qual o significado deste intervalo de confiança?

3.6 Estimando o Tamanho das Amostras

Quando trabalhamos com amostras, como quase sempre ocorre na pesquisa agropecuária, uma questão importante que surge é como estimar o tamanho de uma amostra de modo que ela seja representativa da população. As técnicas de amostragem que vimos em 2.4 são úteis para nos ajudar a produzir amostras não viciadas, cujos elementos sejam realmente obtidos da população aleatoriamente. Mas, assumindo que utilizamos uma técnica de amostragem correta, quantos elementos devem compor nossa amostra de modo a que os dados dela obtidos sejam significativos?

A resposta a esta pergunta depende de quão próximo da média verdadeira queremos que a nossa média amostral esteja. Antes de obtê-la, vejamos como calcular o desvio padrão da média amostral.

Suponha uma lavoura de café da qual tomamos 4 amostras com 11 elementos cada e calculamos a média das alturas dos pés de cada uma delas. Podemos agora calcular o desvio padrão dessas médias.

Há, no entanto, uma maneira direta de calcular o *desvio padrão da média*, que chamaremos de S_m, baseada nos dados de apenas uma amostra. Seja N o número de elementos da amostra e s seu desvio padrão. O desvio padrão da média é dado por:

(3.6.1) $$\boxed{S_m = \frac{s}{\sqrt{N}}}$$

Com este desvio padrão sabemos calcular o intervalo de, por exemplo, 95% de confiabilidade para a média, ou seja, um intervalo que, com 95% de certeza, conterá a média verdadeira. Assim, de acordo com o que vimos em 3.5, em 95% dos casos:

$$\boxed{m - (t_0 \times S_m) < M < m + (t_0 \times S_m)}$$

Note que $(t_0 \times S_m)$ corresponde ao incremento/decremento que quando aplicado a m, define o intervalo de confiança da média. Este incremento/decremento é chamado de *erro padrão* e denotado por D. Temos:

$$\boxed{D = (t_0 \times S_m) = \left(t_0 \times \frac{s}{\sqrt{N}}\right)}$$

Podemos agora manipular algebricamente esta equação de modo a isolar N. Temos:

(3.6.2) $$\boxed{N = \left(t_0 \times \frac{s}{D}\right)^2}$$

Um procedimento muito utilizado para estimar o tamanho que uma amostra deve ter, consiste em trabalharmos, preliminarmente, com uma amostra piloto para que possamos obter os valores de s e t_0 da equação 3.6.2. Mas para que possamos calcular

N resta saber o valor de D. Como dissemos, o erro padrão D representa o incremento/decremento que aplicado à média amostral m produz um intervalo que com 95% de certeza conterá a média verdadeira. O pesquisador determina então que tipo de erro padrão seria aceitável em sua pesquisa, definindo assim o valor de D. Com isso ele obtém todas as informações necessárias para calcular, através da equação 3.6.2, o tamanho da amostra mais adequado à sua pesquisa.

3.6.1 Exercício

Considerando D=5cm, arredondando t_0 para 2,23 (dez graus de liberdade em um intervalo de confiança de 95%) e tomando o desvio padrão s obtido da amostra amarela de nossa tabela de alturas de pés de café, calcule uma estimativa para o tamanho N de uma amostra que garanta, com 95% de certeza, que a média verdadeira M da população esteja compreendida no intervalo (m-5, m+5), onde m representa a média das alturas desta amostra.

3.7 Interpretando Resultados: Os Testes de Hipóteses

Vamos agora dar uma rápida olhada nas possibilidades de utilização da estatística como ferramenta para interpretar e tomar decisões com relação aos dados de uma pesquisa.

Vejamos um exemplo. Há 3 meses atrás verificamos que 20% dos pés de uma lavoura de café estavam infectadas com uma doença A. Hoje, suspeitamos que este número aumentou e queremos verificar.

São duas as hipóteses a verificar:

H_0: a proporção de pés de café doentes continua 20%.

H_1: a proporção de pés de café doentes é maior do que 20%.

A primeira hipótese é conhecida como *nulidade*, por razões óbvias, enquanto que a segunda é chamada de *hipótese alternativa*. Podemos reescrevê-las matematicamente como:

$$H_0: p = 0{,}2 \qquad \text{e} \qquad H_1: p > 0{,}2$$

Para verificar se o tratamento deu algum resultado basta tomarmos uma amostra da lavoura de, digamos, 100 plantas e verificarmos quantas delas apresentam a doença A. Se, por exemplo, 67 dos pés da amostra estiverem doentes, parece bastante razoável concluir que a doença se alastrou e nos decidirmos pela hipótese H_1. Mas e se encontrarmos 21, 22 ou mesmo 24 pés doentes? O que dizer? Estes valores certamente dariam um percentual acima de 20% em relação a amostra, mas e em relação a lavoura toda? Será que são suficientes para explicar uma rejeição de H_0 e escolha H_1?

Temos que estabelecer um limite a partir do qual consideraremos a hipótese H_0 rejeitada e optaremos por H_1. Mas baseado em que escolheremos este limite? O que significa, em termos estatísticos, dizer que só vamos abandonar a hipótese H_0 se encontrarmos, por exemplo, 40 ou mais pés de café doentes em nossa amostra de 100? Qual seria um bom limite para que a escolha de H_1, baseada neste limite, nos desse 95% de certeza de que é de fato H_1 que ocorre na população total?

Em geral, o que se faz é escolher um *nível de significância* para o teste. O nível de significância representa a probabilidade máxima que admitimos para o erro de rejeitar H_0 quando H_0 for verdadeira (o que neste caso é o mesmo que escolher H_1 quando H_1 é falsa).

Assim, no nosso exemplo, ao invés de atribuirmos arbitrariamente o número 40, ou qualquer outro, como limite entre a escolha das duas hipóteses, escolhemos um nível de significância para nosso teste, digamos 5%, e calculamos o limite numérico a partir deste valor. O valor mais adequado para este limite será então aquele que garantir o nível de significância de 5% que escolhemos.

Este é um exemplo simples mas bastante típico do tipo de situação em que a inferência estatística propiciada pelos testes de hipótese auxilia os pesquisadores na interpretação dos resultados de suas pesquisas. Existem inúmeros tipos de testes de hipóteses diferentes, adequados a variadas situações. A escolha e aplicação do teste de hipótese mais adequado em uma determinada circunstância nem sempre é tarefa simples e em muitos casos exige a ajuda de um estatístico profissional. Dentre os testes mais conhecidos estão o teste do qui-quadrado, o teste de variância, o teste do sinal, o teste de Walsh e muitos outros. Não vamos nos aprofundar neste tópico, mas fica aqui nossa recomendação de leitura ao estudante interessado: **VIEIRA[1980]** e **MILLER[1977]**.

3.7.1 Exercício

O que a expressão *nível de significância* quer dizer?

3.8 Algumas Considerações

Há muitos outros tópicos da estatística que são bastante importantes para a atividade do pesquisador agropecuário, que sequer mencionamos aqui. Em especial, podemos citar as noções de *correlação* e *regressão*, que certamente mereceriam ser tratadas. Nosso intuito, ao apresentar neste curso de metodologia da pesquisa alguns rudimentos das ferramentas estatísticas úteis à pesquisa agropecuária, não foi o de instrumentalizar o estudante para realizar o trabalho estatístico das pesquisas, o que seria esperado de um curso de estatística, mas apenas foi o de evidenciar o valor, a força, a importância e o caráter fundamental da estatística para o desenvolvimento da pesquisa científica agropecuária. Esperamos sinceramente tê-lo atingido e recomendamos fortemente ao estudante interessado que consulta a bibliografia indicada.

4. Bibliografia

VIIEIRA, S. *Introdução à bioestatística*. 3.ed. Rio de Janeiro: Ed. Campus, 1980.

REY, L. *Planejar e redigir trabalhos científicos*. 2.ed. São Paulo: Edgard Blücher, 1998.

LITLE, T. & HILLS, F. J. *Métodos estatísticos para la investigación en la agricultura*. México: Ed. Trillas, 1976.

MILLER, S. *Planejamento experimental e estatística*. Rio de Janeiro: Zahar, 1977.

GOMES, F. P. *A estatística moderna na pesquisa agropecuária*. 3.ed. Piracicaba: POTAFOS, 1987.

PÁDUA, E. M. M de *Metodologia da pesquisa: abordagem teórico-prática*. 3.ed. Campinas: Papirus, 1998.