SME006 - Técnicas de Amostragem

Lista 1

Francisco Rosa Dias de Miranda - 4402962

Abril 2023

Questão 1.1.1

Qual é o grau de conhecimento e adoção de blockchain pelas empresas de tecnologia no Brasil?

- Unidade de pesquisa: empresas de tecnologia no Brasil
- População: todas as empresas de tecnologia registradas no Brasil
- Instrumento de coleta de dados: questionário online ou por telefone com perguntas fechadas e abertas sobre conhecimento e adoção de blockchain
- Unidade respondente: representante autorizado da empresa
- Possível sistema de referência: registro de empresas de tecnologia no Brasil
- Unidade amostral mais provável: empresas de tecnologia selecionadas aleatoriamente a partir do registro nacional de empresas de tecnologia
- Unidades amostrais alternativas: empresas de tecnologia selecionadas aleatoriamente a partir de listas de associações de empresas de tecnologia, diretórios de negócios, entre outros
- Tamanho da amostra: pode ser determinado utilizando a fórmula de amostragem aleatória simples para uma população finita, levando em consideração o tamanho da população, nível de confiança e margem de erro desejados.

Outras considerações relevantes podem incluir o tamanho mínimo para subgrupos relevantes, o custo da coleta de dados e a possibilidade de desistência das empresas selecionadas em participar da pesquisa.

Assumindo que a população seja grande e com variabilidade moderada, e considerando um nível de confiança de 95%, o tamanho da amostra necessário para um erro amostral de 2% pode ser calculado pela fórmula:

$$n = \frac{(Z^2 * p * (1 - p))}{\epsilon^2}$$

Onde:

- n é o tamanho da amostra necessário
- Z é o valor crítico da distribuição normal para o nível de confiança desejado (1,96 para 95%)
- p é a proporção esperada na população (se não for conhecida, assume-se $p = 0,5$ para o maior tamanho amostral possível)
- ϵ é o erro amostral desejado, em decimal (0,02)

Assumindo uma proporção esperada na população de 0,5, temos:

$$n = (1,96^2 * 0,5 * (1 - 0,5))/0,02^2 = 2401$$

Portanto, seria necessário amostrar 2401 indivíduos para obter um erro amostral de 2%, com um nível de confiança de 95%. No entanto, é importante ressaltar que outros fatores, como a representatividade da amostra e o custo da pesquisa, também devem ser considerados na definição do tamanho da amostra.

Questão 1.1.4

Para realizar a estimação da proporção de alunos de pós graducação favoráveis a mudança no exame, o plano amostral deve levar em consideração alguns pontos importantes, tais como:

- Sistema de referência: a população alvo é formada por todos os alunos de pós-graduação da universidade, logo, o sistema de referência utilizado será o cadastro completo desses alunos.
- Tamanho da amostra: é necessário calcular o tamanho da amostra para garantir que a pesquisa seja representativa da população alvo e que os resultados obtidos sejam precisos e confiáveis. Considerando que a proporção esperada dos favoráveis é de 5%, o tamanho mínimo da amostra pode ser calculado utilizando a fórmula $n = (Z_{\alpha/2})^2 * p * (1-p)/\epsilon^2$, onde $Z_{\alpha/2}$ é o valor crítico da distribuição normal padrão correspondente ao nível de confiança desejado (em nosso caso, para um nível de confiança de 95%, $Z_{\alpha/2} = 1,96$), p é a proporção esperada dos favoráveis e ϵ é o erro amostral tolerado. Supondo um erro amostral de 2%, o tamanho mínimo da amostra seria $n = (1,96)^2 * 0,05 * 0,95/0,02^2 = 377,61$. Arredondando para cima, o tamanho da amostra seria de pelo menos 378 alunos.
- Unidade primária de amostragem (UPA): a UPA pode ser definida como o curso de pós-graduação ao qual o aluno está matriculado.
- Unidade secundária de amostragem (USA): a USA pode ser definida como o próprio aluno de pós-graduação.
- Estimadores e variâncias: para estimar a proporção dos favoráveis, pode-se utilizar o estimador da proporção amostral, que é dado por $\hat{p} = X/n$, onde X é o número de alunos favoráveis na amostra e n é o tamanho da amostra. A variância do estimador pode ser calculada utilizando a fórmula $Var(\hat{p}) = p(1-p)/n$, onde p é a proporção populacional dos favoráveis.

Dessa forma, o plano amostral seria definido como:

- Sistema de referência: cadastro completo dos alunos de pós-graduação da universidade.
- Tamanho da amostra: pelo menos 378 alunos de pós-graduação.
- UPA: o curso de pós-graduação.
- USA: o próprio aluno de pós-graduação.
- Estimadores e variâncias: estimador da proporção amostral $\hat{p} = X/n$ e variância $Var(\hat{p}) = p(1-p)/n$, onde p é a proporção populacional dos favoráveis.

Questão 1.1.10

1. Unidade domiciliar: A UPA domiciliar pode ser uma boa escolha para estimar o consumo médio de água por domicílio em uma cidade, uma vez que permite uma amostragem simples e direta, além de fornecer informações precisas e detalhadas sobre o consumo de água de cada residência. No entanto, o uso da unidade domiciliar pode ser desvantajoso se a variação no consumo de água for muito grande entre as residências, tornando necessária uma amostra maior para garantir a representatividade.
2. Blocos de domicílios: O uso de blocos de domicílios, como casas, prédios de apartamentos e vilas, pode ser vantajoso para reduzir o tamanho da amostra necessária e aumentar a eficiência da coleta de dados. Além disso, o uso de blocos permite obter informações sobre o consumo médio em diferentes tipos de

residências, o que pode ser útil em estudos de planejamento urbano. No entanto, uma desvantagem do uso de blocos é que a heterogeneidade pode ser maior dentro de um bloco do que entre blocos, o que pode afetar a precisão da estimativa.

3. Quarteirões: O uso de quarteirões como UPA pode ser vantajoso para fornecer informações detalhadas sobre as características socioeconômicas da população em cada quarteirão, o que pode ser útil para análises posteriores. Além disso, a variação no consumo de água pode ser menor dentro de um quarteirão do que entre quarteirões, o que aumenta a precisão da estimativa. No entanto, a desvantagem do uso de quarteirões é que a amostra pode ser grande e complexa de selecionar, o que pode levar a erros de amostragem se a seleção não for feita de forma cuidadosa.

Questão 1.1.13

Cada um dos diversos métodos de coleta de dados existente possui seus próprios méritos e limitações, é importante escolher o método que melhor atenda aos objetivos e recursos disponíveis para a pesquisa. Dentre os métodos de entrevista pessoal, telefone, correio e internet podemos destacar:

- Entrevista pessoal: A entrevista pessoal é um método de coleta de dados em que o entrevistador se encontra face a face com o entrevistado. Este método é útil quando é necessário obter informações detalhadas ou sensíveis, pois o entrevistado pode fornecer respostas mais precisas e pode ser mais fácil estabelecer um relacionamento de confiança. No entanto, esse método pode ser caro e demorado, especialmente se a amostra for grande e dispersa geograficamente.
- Telefone: A coleta de dados por telefone é um método rápido e relativamente barato. É adequado para pesquisas que requerem informações simples e rápidas e é eficaz para atingir uma amostra grande e dispersa geograficamente. No entanto, pode haver uma taxa de não resposta alta e é difícil estabelecer um relacionamento de confiança com o entrevistado.
- Correio: A coleta de dados por correio é um método de baixo custo e pode ser útil para amostras pequenas. É apropriado para pesquisas em que as perguntas são claras e diretas, e não há necessidade de esclarecimentos. No entanto, a taxa de resposta pode ser baixa e há uma possibilidade de viés de resposta.
- Internet: A coleta de dados pela internet é um método de coleta de dados de baixo custo e rápido, e é adequado para amostras grandes e dispersas geograficamente. Esse método permite uma ampla cobertura demográfica, e os dados podem ser coletados de forma eficiente. No entanto, é importante estar ciente de possíveis vieses amostrais e que nem todas as pessoas têm acesso ou estão dispostas a responder pesquisas on-line.

a.

No caso do diretor de marketing de uma rede de televisão que deseja estimar a proporção de pessoas no país assistindo a determinado programa, uma opção pode ser utilizar uma combinação de métodos de coleta de dados, como entrevistas pessoais para obter informações detalhadas sobre o público, telefone para cobrir uma amostra ampla e geograficamente dispersa, e internet para alcançar uma amostra ainda maior. A escolha do método ou combinação de métodos mais adequados dependerá das especificidades, recursos e objetivos da pesquisa.

b.

Existem diversas formas de coletar opiniões dos leitores sobre os tipos de notícias de um jornal, como pesquisa por telefone, online foco e enquete por correio. É importante que a amostra de leitores represente a base de

leitores do jornal e que as perguntas sejam claras e objetivas. Com essas informações, o editor pode melhorar a qualidade do conteúdo e aumentar a satisfação dos leitores. Poderia ser interessante a adoção de *grupos de foco*: uma espécie de entrevista em que um grupo de leitores pode ser convidado a se reunir para discutir seus pensamentos e opiniões sobre as notícias do jornal. O editor ou um moderador qualificado pode liderar a discussão e registrar os comentários dos participantes.

c.

Para estimar o número de cachorros vacinados contra a raiva no ano passado, o departamento de saúde pode utilizar algumas estratégias. Uma opção é solicitar os registros de vacinação de todas as clínicas veterinárias da região, somando o número total de cachorros vacinados. Outra opção é realizar uma pesquisa por amostragem, selecionando aleatoriamente uma amostra de proprietários de cachorros e perguntando se seus animais de estimação foram vacinados contra a raiva no ano anterior. O departamento de saúde também pode solicitar relatórios de vacinação de organizações de resgate de animais e abrigos. Independentemente da abordagem escolhida, é importante que a amostra de dados seja representativa da população total de cachorros da região e que os dados sejam analisados com precisão e rigor estatístico para se obter uma estimativa confiável do número de cachorros vacinados.

Questão 2.

1. Famílias individuais: selecionar famílias individuais como unidades de amostragem pode ser útil se o objetivo for obter informações detalhadas sobre as características individuais de cada família e seu consumo de água. No entanto, pode ser difícil obter uma amostra representativa da população usando apenas famílias individuais como unidades de amostragem, especialmente se a população for muito grande e/ou dispersa. Um possível sistema de referência seria um cadastro de endereços residenciais na cidade, a partir do qual as famílias individuais poderiam ser selecionadas aleatoriamente.
2. Unidades habitacionais: escolher unidades habitacionais (como casas unifamiliares, prédios de apartamentos, vilas, etc.) como unidades de amostragem pode ser uma opção mais prática e eficiente para obter uma amostra representativa da população. Por exemplo, em vez de selecionar uma amostra aleatória de famílias individuais, poderia-se selecionar aleatoriamente um número suficiente de unidades habitacionais e, em seguida, coletar informações sobre o consumo de água de todas as famílias que residem nessas unidades. Um possível sistema de referência seria um cadastro de unidades habitacionais na cidade, a partir do qual as unidades habitacionais poderiam ser selecionadas aleatoriamente.
3. Quarteirões: selecionar quarteirões como unidades de amostragem pode ser uma opção útil se houver interesse em estudar as diferenças entre bairros ou áreas geográficas específicas da cidade. Por exemplo, poderia-se selecionar aleatoriamente alguns quarteirões em diferentes áreas da cidade e, em seguida, coletar informações sobre o consumo de água de todas as famílias que residem nesses quarteirões. Um possível sistema de referência seria um mapa da cidade, a partir do qual os quarteirões poderiam ser selecionados aleatoriamente.

Questão 4.

Existem várias unidades de amostragem possíveis para estimar o número de hectares de plantação de cana de açúcar dentro do estado de São Paulo. Algumas possibilidades são:

- Municípios: uma estratégia seria selecionar aleatoriamente um conjunto de municípios do estado e coletar informações sobre o número de hectares de plantação de cana de açúcar em cada um deles. Essa abordagem permitiria obter uma estimativa para cada município, além da estimativa para o estado como um todo. O sistema de referência seria cada município selecionado.

- Propriedades rurais: outra opção seria utilizar propriedades rurais como unidades de amostragem. Nesse caso, seria necessário ter acesso a informações cadastrais para a população de propriedades do estado. O sistema de referência seria cada propriedade rural selecionada.
- Imagens de satélite: uma alternativa seria utilizar imagens de satélite para identificar e medir as áreas de plantação de cana de açúcar em todo o estado. Nesse caso, a unidade de amostragem seria a própria imagem de satélite e o sistema de referência seria o estado de São Paulo como um todo.

Independentemente da unidade de amostragem escolhida, é importante planejar cuidadosamente o processo de amostragem para garantir que a estimativa seja precisa e confiável.

Questão 2.1

Vamos simular uma população de tamanho $N = 100$, onde a caracteristica de interesse possui distribuição Y $\mathcal{N}(50, 16)$. Após essa etapa, vamos encontrar o total populacional τ, a media populacional μ, e a variancia populacional S^2.

```
set.seed(3)
#Simulacao da amostra da v.a. Y com N = 100.
Y <- rnorm(100, 50, 16)

#Calculo do Total populacional.
T<-sum(Y)

#Calculo da media populacional mu.
mu<-mean(Y)

#////////////////////////////////////
#Calculo da varincia amostral.
S2<-var(Y)

data.frame(T, mu, S2)
```

```
##          T       mu       S2
## 1 5017.657 50.17657 187.6141
```

Questão 2.2

Considere a populacao dada na tabela 2.8 (pagina 60), onde:

- X denota o numero de apartamentos nos condominios.
- Y denota o numero de apartamentos alugados.

```
library(tidyverse, quietly = T, verbose = F)
```

```
## -- Attaching packages ------------------------------------ tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
```

```
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
dados <- read.csv("CAP_02_tab_2_8.csv",header = TRUE,
                  sep = ",", as.is = TRUE) |>
  select(-indice) |> mutate(w = ifelse(Y > 20, 1, 0))

dados |> summarise(
           across(everything(), list(
                                  media = mean,
                                  total = sum,
                                  variancia = var)))
```

```
##   Y_media Y_total Y_variancia  X_media X_total X_variancia   w_media w_total
## 1    18.6    3348    409.4369 27.37778    4928    609.4096 0.3166667      57
##   w_variancia
## 1   0.2175978
```