DIREÇÃO CONCURSOS

POLÍCIA FEDERAL

RESUMÃO DE ESTATÍSTICA

PROFESSORES **ARTHUR LIMA E HUGO LIMA**

APRESENTAÇÃO

PROFESSORES **ARTHUR LIMA E HUGO LIMA**

Futuros Policiais Federais, tudo bem?

Eu (professor Arthur Lima) e o professor Hugo Lima elaboramos este resumo de Fórmulas de Estatística para o concurso da Polícia Federal 2021. Aproveitamos para incluir alguns conceitos muito cobrados que podem ajudar a acertar várias questões nas provas!

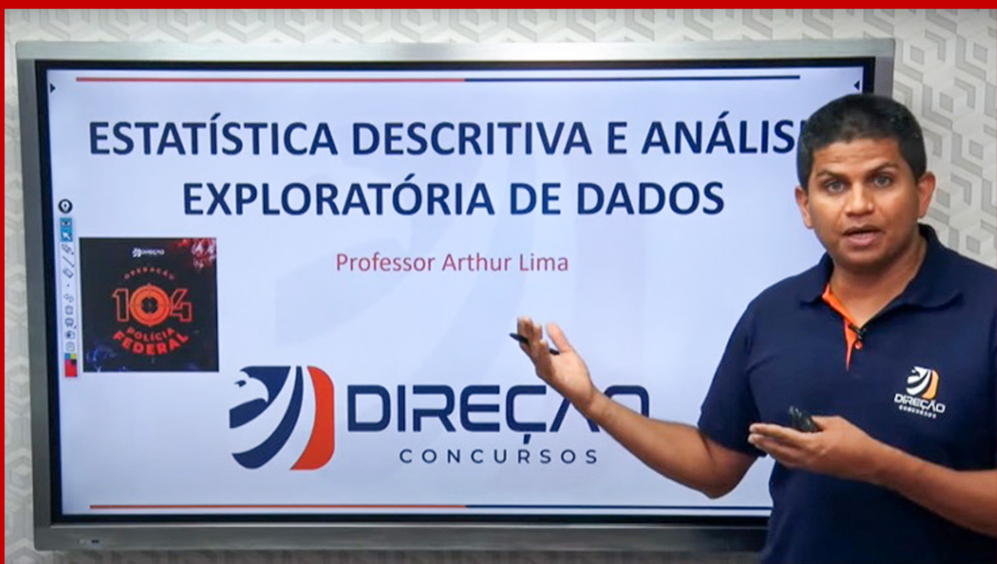
Somos Engenheiros pelo Instituto Tecnológico de Aeronáutica (ITA) e fomos aprovados nos concursos da Receita Federal para o cargo de Auditor-Fiscal. Também somos os professores responsáveis pelas disciplinas de Raciocínio Lógico, Matemática, Matemática Financeira e Estatística no Direção Concursos.

Mantenham este resumão sempre acessível para revisarem várias vezes as fórmulas de Estatística até a prova da Polícia Federal, ok?

E, por favor, compartilhem esse material com seus amigos que vão prestar o concurso da Polícia Federal.

Bons estudos! Nos vemos na ANP ainda este ano!

AULA GRATUITA DE ESTATÍSTICA – QUESTÕES CEBRASPE



RESUMÃO DE ESTATÍSTICA

PROFESSORES **ARTHUR LIMA E HUGO LIMA**

MEDIDAS DESCRITIVAS – POSIÇÃO

MÉDIA

Dados em rol (lista):

$$Média(\overline{X}) = \frac{Soma}{quantidade} = \frac{\sum X_i}{n}$$

Tabela de frequências (e média ponderada, usando os pesos no lugar de fi):

$$\overline{X} = \frac{\sum X_i \times f_i}{\sum f_i}$$

Dados agrupados em intervalos (usar ponto médio – PM):

$$\overline{X} = \frac{\sum PM_i \times f_i}{\sum f_i}$$

- Média é afetada por todos os valores da distribuição
- Se somarmos / subtrairmos ou multiplicarmos / dividirmos todos os termos de uma distribuição pelo mesmo número, a média sofrerá o mesmo efeito

MEDIANA

Passos:

1 – colocar dados em ordem crescente;

2 – calcular a posição da mediana:

Posição=(n+1)/2

3 – localizar a mediana (calcular a média entre 2 valores, se necessário).

- Mediana não é afetada pelas mudanças de valores extremos da distribuição

MODA

É o valor com maior número de frequências (repetições).

- Moda não é afetada pelas mudanças de valores extremos da distribuição

SIMETRIA

É o valor com maior número de frequências (repetições).

Simetria da Distribuição	Média, Mediana e Moda
Simétrica	Média = Mediana = Moda
Assimétrica positiva (à direita)	Média > Mediana > Moda
Assimétrica negativa (à esquerda)	Média < Mediana < Moda

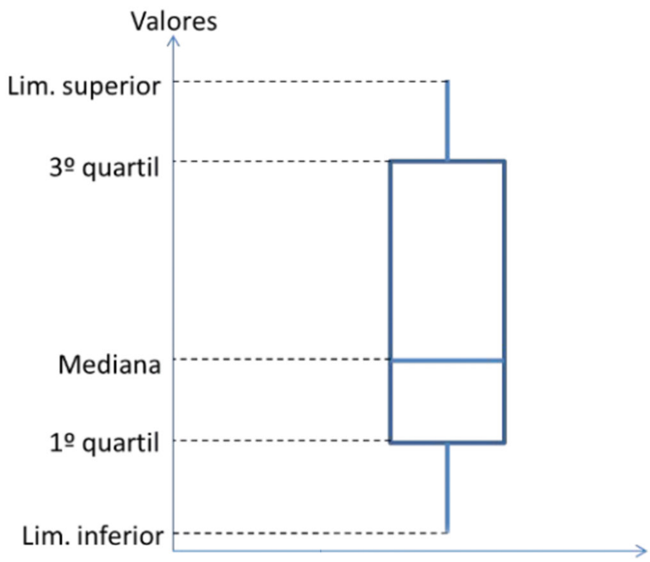
ESPERANÇA OU VALOR ESPERADO "MÉDIA"

$$E(X)=\sum x.\,p(x)$$

MEDIDAS SEPARATRIZES - QUARTIS

Quartil	Posição
1	(n+1)/4
2	2(n+1)/4
3	3(n+1)/4

DIAGRAMA BOX PLOT



Limite Superior –> MENOR valor entre o valor máximo e Q3 + 1,5.(Q3-Q1)

Limite Inferior –> MAIOR valor entre o valor mínimo e Q1 – 1,5.(Q3 – Q1)

Outliers (valores atípicos) –> acima do Limite Superior ou abaixo do Limite Inferior

TÉCNICAS DE AMOSTRAGEM PROBABILÍSTICAS (CASUAIS)

- **Aleatória Simples:** é preciso ter acesso a todos os elementos da população; todos os elementos precisam ter a mesma chance de serem escolhidos para a amostra; população única; com ou sem reposição.
- **Sistemática:** regra objetiva / lógica / sistema para seleção dos elementos da amostra.
- **Estratificada:** população é dividida em estratos (agrupamentos mais homogêneos entre si) e são selecionados ALGUNS elementos de TODOS os estratos.
- **Por conglomerados:** população é dividida em subgrupos (conglomerados, não necessariamente mais homogêneos entre si), e TODOS os elementos de ALGUNS subgrupos são analisados.

MEDIDAS DESCRITIVAS - DISPERSÃO

VARIÂNCIA

$$Variância(\sigma^2)=E(X^2)-[E(X)]^2$$

ou

$$Variância(\sigma^2)=\frac{\sum X^2-\frac{1}{n}.(\sum X)^2}{n}$$

Tabela de Frequências:

$$Variância(\sigma^2)=\frac{\sum X^2.f_i-\frac{1}{n}.(\sum X.f_i)^2}{n}$$

Intervalos de Classes:

$$Variância(\sigma^2)=\frac{\sum PM_i^2.f_i-\frac{1}{n}.(\sum PM_i.f_i)^2}{n}$$

Para calcular a variância AMOSTRAL, substitua n por "n-1" no denominador das fórmulas acima

DESVIO PADRÃO

$$Desvio\ padrão(\sigma)=\sqrt{Variância}$$

- Somas/subtrações não afetam o desvio padrão e nem a variância
- Se multiplicarmos/dividirmos todos os termos da distribuição pelo mesmo número, o desvio padrão é multiplicado/dividido pelo número, e a variância pelo quadrado do número

$$Desvio\ padrão<\frac{Amplitude}{2}$$

COEFICIENTE DE VARIAÇÃO

$$CV=\frac{Desvio\ Padrão}{Média}=\frac{\sigma}{\mu}$$

- Medida de dispersão RELATIVA (boa para comparações entre distribuições)
- Não tem unidade

MEDIDAS DE ASSIMETRIA

$$Coeficiente\ Percentílico=\frac{P90+P10-2.P50}{P90-P10}$$

$$Coeficiente\ Quartílico=\frac{Q3+Q1-2.Md}{Q3-Q1}$$

$$Primeiro\ coeficiente\ de\ Pearson=\frac{(Média-Moda)}{Desvio\ Padrão}$$

$$Segundo\ coeficiente\ de\ Pearson=3.\frac{(Média-Mediana)}{Desvio\ Padrão}$$

MEDIDAS DE CURTOSE

Medem o grau de achatamento de uma distribuição em relação à distribuição normal.

($C = m_4 / s^4$)

Classificação	Coef. Momento de Curtose
Platicúrtica (mais achatada)	C < 3
Mesocúrtica	C = 3
Leptocúrtica	C > 3

$$K=\frac{Q3-Q1}{2.(P90-P10)}$$

Classificação	Coef. Percentílico de Curtose
Platicúrtica (mais achatada)	C < 3
Mesocúrtica	C = 3
Leptocúrtica	C > 3

- Distribuição normal é mesocúrtica

PROBABILIDADE

DEFINIÇÃO

Probabilidade do evento = número de resultados favoráveis / número total de resultados

Eventos independentes:

P (A ∩ B) = P (A) X P (B)

Probabilidade da união de eventos:

P (A ∪ B) = P (A) + P (B) - P (A ∩ B)

Eventos mutuamente excludentes:

P (A ∩ B) = 0

Eventos complementares:

Probabilidade (E) = 1 - Probabilidade (E^c)

Probabilidade condicional:

P (A/B) = P(A∩B)/P (B)

PRINCIPAIS DISTRIBUIÇÕES DE PROBABILIDADE DISCRETAS

DISTRIBUIÇÃO	COMO IDENTIFICAR	VALOR ESPERADO (MÉDIA)	VARIÂNCIA
Bernoulli	1 tentativa 2 resultados possíveis	E(X) = p	Var(X) = p.(1-p)
Binomial	"n" tentativas independentes 2 resultados possíveis Probabilidade de sucesso constante	E(X) = n.p	Var(X) = n.p.(1-p)
Poisson	Fenômeno que se estende no tempo ou espaço com independência e regularidade conhecida	E(X) = λ	Var(X) = λ
Geométrica	Probabilidade de "n" tentativas até o primeiro sucesso	E(X) = 1/p	Var(X) = (1-p)/p²

FUNÇÃO DE PROBABILIDADE BINOMIAL

$$P(k,n,p) = C(n,k).p^k.(1-p)^{n-k}$$

(probabilidade de k sucessos em n tentativas com probabilidade p de sucesso a cada tentativa)

FUNÇÃO DE PROBABILIDADE POISSON

$$f(k,\lambda) = \frac{e^{-\lambda}.\lambda^k}{k!}$$

(probabilidade de k repetições do evento durante o período observado)

FUNÇÃO DE PROBABILIDADE GEOMÉTRICA

$$P(X = n) = p.(1 - p)^{n-1}$$

PRINCIPAIS DISTRIBUIÇÕES DE PROBABILIDADE CONTÍNUAS

DISTRIBUIÇÃO	VALOR ESPERADO (MÉDIA)	VARIÂNCIA
Uniforme	E(X) = (Máximo+Mínimo)/2	Var(X) = (máximo-mínimo)²/12
Exponencial	E(X) = 1/λ	Var(X) = 1/λ²
Qui-quadrado	E(X) = n	Var(X) = 2n

- Nas distribuições contínuas, a probabilidade de qualquer valor exato é ZERO
- Função de distribuição acumulada exponencial: $F(X<x)=1-e^{(-\lambda.x)}$
- Qui-quadrado é obtida pela soma dos quadrados de "n" distribuições normais padronizadas independentes.

Distribuição Normal

• Presente em fenômenos da natureza em que a maior parte dos valores está próxima da média, mas há valores mais afastados da média para ambos os lados;

• Se os dados tem distribuição normal, pode-se dizer que cerca de 68% encontram-se entre $\mu-\sigma$ e $\mu+\sigma$. Da mesma forma, 95% dos dados encontram-se entre $\mu-2\partial$ e $\mu+2\sigma$, e 99,7% entre $\mu-3\sigma$ e $\mu+3\sigma$;

• A distribuição normal padrão tem média 0 e desvio padrão 1;

• Para transformar uma distribuição normal qualquer em uma normal padrão, basta usar a transformação: $Z=(X-\mu)/\sigma$;

• A distribuição é simétrica, de modo que $P(X > a)$ é igual a $P(X < -a)$, e média = mediana = moda;

• Se X e Y forem independentes e seguirem distribuição normal, então $X+Y$ e $X-Y$ também terão distribuição normal, e:

$$E(X+Y) = E(X) + E(Y)$$

$$Var(X+Y) = Var(X) + Var(Y)$$

PROBABILIDADE

• Parâmetro: característica da população (ex.: idade média populacional);

• Estimativa: valor obtido a partir de uma amostra (ex.: média amostral);

• Estimador: função matemática usada para obter a estimativa (ex.: somar todos os valores e dividir pela quantidade).

DEFINIÇÃO

• Viés: "parcialidade" ou "tendência" do estimador. Um estimador não viesado tem como valor esperado o próprio valor do parâmetro;

• Consistência: um estimador consistente converge para o valor do parâmetro à medida que o número de observações aumenta (e a sua variância tende a zero);

• Eficiência: entre dois estimadores, o mais eficiente é o que possui MENOR variância;

• Suficiência: o estimador suficiente capta todas as informações sobre o parâmetro a ser estimado.

ESTIMADR IDEAL

Não-viesado (não tendencioso), consistente, eficiente e suficiente.

ESTIMADORES IMPORTANTES PARA A MÉDIA

• média amostral: é não-viesada e consistente

• mediana amostral: é não-viesada e consistente, mas é menos eficiente que a média amostral

• primeiro item coletado: é não-viesado, mas não é consistente

OUTROS ESTIMADORES IMPORTANTES

• Variância (com "n" no denominador): é viesado. Colocando-se n-1 no denominador, fica não-viesado (variância amostral).

• Desvio padrão (com "n" ou com "n-1" no denominador): é viesado.

ESTIMAÇÃO INTERVALAR

TEOREMA CENTRAL DO LIMITE E DISTRIBUIÇÕES AMOSTRAIS

A distribuição amostral da média é NORMAL, com:

$$Média = média\ da\ população\ (\mu)$$

$$Desvio\ padrão = \frac{\sigma}{\sqrt{n}}$$

ERRO PADRÃO NA ESTIMATIVA

$$E = \frac{\sigma}{\sqrt{n}} \rightarrow erro\ padrão\ na\ estimativa\ de\ médias$$

$$E = \sqrt{\frac{p(1-p)}{n}} \rightarrow erro\ padrão\ na\ estimativa\ de\ proporções$$

INTERVALO DE CONFIANÇA PARA A MÉDIA

$$Média\ amostral \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \quad \text{ou} \quad Média\ amostral \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

• Usar a segunda SOMENTE quando n < 30 e o desvio padrão populacional for desconhecido

• Número de graus de liberdade (t de Student) = n – 1

$$Amplitude\ do\ intervalo = 2 . Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

ou

$$Amplitude\ do\ intervalo = 2 . t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

INTERVALO DE CONFIANÇA PARA PROPORÇÕES

$$Proporção\ amostral\ (p) \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{p . \frac{1-p}{n}}$$

$$Amplitude\ do\ intervalo = 2 . Z_{\frac{\alpha}{2}} \cdot \sqrt{p . \frac{1-p}{n}}$$

$$Margem\ de\ erro = Z_{\frac{\alpha}{2}} \cdot \sqrt{p . \frac{1-p}{n}}$$

INTERVALOS DE CONFIANÇA VS. INTERVALOS DE CREDIBILIDADE

intervalos de Confiança	intervalos de Credibilidade
Estatística Frequentista	Estatística Bayesiana
Não dependem da distribuição a priori	Utilizam informações da distribuição a priori (cria intervalo de probabilidade a posterior)

TAMANHO DE AMOSTRAS

$$n = \left(\frac{Z_{\alpha/2} . \sigma}{d}\right)^2 \rightarrow para\ estimativa\ de\ médias$$

$$n = \left(\frac{Z_{\alpha/2}}{d}\right)^2 . p . q \rightarrow para\ estimativa\ de\ proporções$$

• As fórmulas acima consideram que a população é infinita (amostragem com reposição)

• Quando p e q forem desconhecidos, usar p = q = 0,5

TESTES DE HIPÓTESES

RECEITA DE BOLO – TESTES DE HIPÓTESES PARA MÉDIAS

1 – Formular a hipótese nula (H_0) e a hipótese alternativa (H_1) – mutuamente excludentes

2 – Rascunhar a curva normal, marcando as regiões de aceitação e de rejeição (região crítica) de acordo com o nível de significância α e o tipo de teste (bilateral, se H_1 for do tipo "média diferente de", ou unilateral, se H_1 for do tipo "maior que" ou "menor que")

3 – Obter $Z_{tabelado}$ com base no nível de significância e tipo de teste

4 – Obter $Z_{calculado}$ por meio da expressão:

$$Zcalculado = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

5 – Posicionar $Z_{calculado}$ no gráfico e obter a conclusão do teste – aceitação ou rejeição de H_0

Usar t de Student se n < 30 e o desvio padrão populacional for desconhecido. Neste caso,

$$t_{calculado} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Número de graus de liberdade = n – 1

Se for um teste de hipóteses para proporções, usar

$$Z_{calculado} = \frac{(p - p_0)}{\sqrt{p_0 \cdot \frac{1 - p_0}{n}}}$$

onde p é a proporção obtida na amostra e p_0 está nas hipóteses do teste.

ERROS EM TESTES DE HIPÓTESES

	H_0 é verdadeira	H_0 é Falsa
H_0 é Aceita	Certo	Erro tipo II (falso negativo, probabilidade β)
H_0 é Rejeitada	Erro tipo I (falso positivo, probabilidade α)	Certo

OUTROS CONCEITOS EM TESTES DE HIPÓTESES

• Poder (Potência) do Teste: probabilidade de rejeitar corretamente a hipótese nula (1-β)

• p-valor (nível descritivo, probabilidade de significância): p-valor ≤ α -> rejeitar H_0

ANÁLISE DE REGRESSÃO

COVARIÂNCIA

Cov(X,Y) = E(X.Y) – E(X).E(Y)

• Positiva: as duas variáveis tendem a crescer / decrescer juntas

• Negativa: quando uma variável cresce, a outra tende a decrescer

Var(X + Y) = Var(X) + Var(Y) + 2.a.b.cov(X,Y)

Var(X – Y) = Var(X) + Var(Y) – 2.a.b.cov(X,Y)

Cov(a.X + b, c.Y + d) = a.c.cov(X,Y)

CORRELAÇÃO

Correlação(X,Y)=(cov(X,Y))/(σ_X σ_Y)

• Varia de -1 (correlação fortemente negativa) a +1 (correlação fortemente positiva)

• Se variáveis são INDEPENDENTES, então a correlação é NULA (o contrário não é necessariamente verdade)

Correlação(X,X) = 1

Correlação(a.X + b, c.Y + d) = (sinal de a.c) x Correlação(X,Y)

REGRESSÃO LINEAR SIMPLES

$$Y_i = \alpha + \beta . X_i + \epsilon_i$$

- α é o coeficiente linear, β é o coeficiente angular, ε_i é o erro aleatório

- X é a variável independente, Y é a variável dependente

$$\beta = \frac{n . \sum XY - \sum X . \sum Y}{n . \sum X^2 - \left(\sum X\right)^2}$$

ou

$$\beta = \frac{cov(X, Y)}{Var(X)}$$

$$\textit{Relação entre as médias}: \overline{Y} = \alpha + \beta . \overline{X}$$

COEFICIENTE DE DETERMINAÇÃO R^2

• É o quadrado do coeficiente de correlação

• Mede o ajustamento da regressão ao modelo linear

• Varia entre 0 e 1 (quanto mais próximo de 1, melhor o ajustamento)

• Interpretação: é o percentual das variações da variável dependente que são explicadas pela variável independente.

LEIS DOS GRANDES NÚMEROS

• Afirmam que a média amostral converge para média populacional à medida que aumentamos o tamanho da amostra.

Lei FRACA dos grandes números	Lei FORTE dos grandes números
A convergência ocorre em PROBABILIDADE	A convergência é CERTA / QUASE CERTA
Se o número de elementos da amostra é suficientemente grande, a convergência é PROVÁVEL	A partir de um tamanho de amostra "n" suficientemente grande, a convergência é CERTA / QUASE CERTA

ANÁLISE DE RESÍDUOS

• Trata-se do estudo do comportamento do resíduo (erro aleatório) da regressão linear;

• Características desejáveis do resíduo:

- média igual a zero;
- variância constante (homocedasticidade, e NÃO heterocedasticidade);
- normalidade (resíduos com distribuição normal em torno do zero);

• Diagrama de dispersão de resíduos (gráfico de resíduos): permite verificar a homocedasticidade e a presença de outliers;

• Quando a suposição da homocedasticidade é violada, a distribuição de probabilidade do modelo é afetada, mas não o valor esperado.

ANÁLISE DE VARIÂNCIA

- usada para comparação de médias entre 3 ou mais grupos;

Hipótese nula (H_0): as médias não possuem diferença significativa

Hipótese alternativa (H_1): pelo menos uma média difere das demais

Fontes de variação	Soma de quadrados	Graus de Liberdade	Quadrados médios (soma/gl)	F
Entre grupos (equação)	SQE	Grupos – 1	QME = SQE / (Grupos – 1)	F = QME / QMR
Dentro dos grupos (resíduo)	SQR	N – Grupos	QMR = SQR/ (N – Grupos)	
Total (modelo)	SQT	N – 1	QMT = SQT / (N – 1)	

SQE + SQR = SQT

R^2 = SQE / SQT

$\beta = R.\sigma_Y / \sigma_X$

As variâncias são os Quadrados Médios do quadro acima