Nível de Significância (α) ou Valor-p?

SÉRGIO FRANCISCO COSTA
Prof. de Metodologia de Investigação e de Estatística
UnG/Universidade de Guarulhos, USP/Universidade de São Paulo,
UniABC/Universidade do Grande ABC
costasf@terra.com.br

Resumo
Nas últimas décadas, alguns pesquisadores têm preferido trabalhar, quando fazem *testes de hipóteses estatísticas*, com o *nível de significância efetivo*, chamado de **valor-p**, em lugar do tradicional *nível de significância teórico (α)*, fixado *a priori*, por ocasião do *planejamento do experimento*. Neste artigo, procura-se mostrar que tal prática *flexibiliza demasiadamente* a *região crítica*, deixando a questão da *rejeição* ou da *não-rejeição* da *hipótese* sob a quase inevitável *influência de fatores emocionais*. Ressaltam-se também aspectos ligados à ocorrência do chamado *Erro Tipo II* e conseqüências ligadas à *sensibilidade do experimento*, isto é, do *poder*.
Palavras-chave: Nível de significância. Valor-p. Região crítica. Erro Tipo I. Erro Tipo II. Poder.

Resumen
En las últimas décadas, algunos investigadores, al hacer **pruebas de hipótesis estadísticas**, han preferido trabajar con el **nivel de significáncia efectivo**, de nombre valor-p, en lugar del tradicional **nivel de significáncia teórico (α)**, fijado *a priori*, por ocurrencia del **planeamiento del experimento**. En este artículo, se procura mostrar que tal práctica **flexibiliza demasiado** la **región crítica**, dejando el tema de la **renuncia** o de la **no-renuncia** de la **hipótesis** sujeta a la influencia inevitable de **factores emocionales**. Se resaltan también aspectos referentes a la ocurrencia del llamado **Error Tipo II** y consecuencias vinculadas a la **sensibilidad del experimento**, o sea, del **poder**.
Palabras-clave: Nivel de significáncia, Valor-p, Región crítica, Error Tipo I, Error Tipo II, Poder.

Abstract
During the last few decades, some researchers have shown strong preference, when *testing statistical hypoteses*, for the *effective level of significance*, also called **p-value**, in place of the traditional *theoretical level of significance (α)*, established *in advance* – at the time of the *design of the experiment*. This article tries to show that such practice adds extreme *flexibility* to the *critical region*, leaving the question of *rejecting* or *not rejecting* the hypoteses under the almost inevitable *influence of emotional factors*. Some aspects in relation with the occurrence of *Type II Errors* as well as the corresponding *consequences* connected with the *sensitivity of the experiment* (*power*) are dealt with.
Keywords: Level of significance. p-value. Critical region. Type I Error. Type II Error. Power.

1. Nas últimas décadas, alguns pesquisadores, principalmente por influência norte-americana, têm adotado, ao *testarem hipóteses, critérios* que, embora *complementares*, não são rigorosamente *eqüivalentes*. Referimo-nos aos seguintes procedimentos:
   a) especificação, *a priori*, do *nível de significância* (a) ou
   b) cálculo do **valor-p** (*p-value*).

2. Para tornar claros os conceitos acima, vamos recordar, brevemente, o caminho que torna possível *testar* uma *hipótese estatística*. De início, recapitulemos que o correto é falar em *hipóteses estatísticas*, no plural, uma vez que a *lógica subjacente* ao processo exige que haja *duas* hipóteses construídas de tal forma que, não sendo possível admitir a possibilidade de uma delas ser *falsa*, a outra, *por exclusão*, seja *considerada como tal*.

3. Rigorosamente, *nunca* o pesquisador sabe se está diante de uma *hipótese verdadeira*. Entretanto, para ser *falsa* basta que a hipótese sob análise deixe de preencher um *único* quesito. Na verdade, isso remete à questão da *indução*: não é possível afirmar categoricamente que todos os cisnes sejam brancos, porque *não há número suficiente de observações* (por grande que seja esse número) *capaz de garantir a veracidade da declaração*. Por outro lado, basta que, em qualquer tempo, um **único** cisne seja de outra cor para tornar-se possível a inequívoca conclusão de que nem todos os cisnes são brancos. A *Ciência* adota essa postura: *prefere rejeitar inúmeras hipóteses verdadeiras a ter de conviver com uma única falsa*. Outra forma de dizer isso: *a Ciência só tem certeza das coisas que nega; das coisas que afirma, tem, no máximo, probabilidades*. Em síntese, **a Ciência é "negativa"**.

4. Retomando o que ficou dito no parágrafo dois, as hipóteses estatísticas são: H_0, chamada universalmente de *hipótese nula* ou *hipótese probanda*, e H_a, denominada *hipótese alternativa* ou *hipótese experimental*. Sob a *hipótese nula* o pesquisador coloca, em *linguagem probabilística*, o que seria *razoável esperar* em função do que *conhece do fenômeno* ou, quando nada conhece, o que seria *possível imaginar* como conseqüência da *ação do acaso*. Sob a *hipótese alternativa*, como o próprio nome já o sugere, o pesquisador coloca, também em *termos probabilísticos*, o que *deveria ocorrer* se, além do acaso, *outra causa* pudesse ser *prioritariamente responsável pelo resultado observado*.

5. O que está dito no parágrafo quatro pode ser mais bem compreendido à luz de um exemplo simples. Suponhamos então que, numa urna, haja bolas vermelhas (V) e bolas amarelas (A), em quantidades desconhecidas. Seja também admitido que as bolas, exceto no que se refere à cor, tenham todas as *mesmas características*: mesmo peso, mesma compactação, mesmo polimento. Por alguma razão, um pesquisador deseja avaliar se as bolas vermelhas *superam* em número as amarelas ou se, contrariamente, as amarelas *superam* as vermelhas. Também existe a possibilidade de o número de bolas vermelhas ser *diferente* do número de bolas amarelas, não sendo, entretanto, identificável em que proporção. Dependendo da suspeita do pesquisador (baseada em alguma pista - expressa por conhecimento anterior ou por alguma sinalização circunstancial), *três hipóteses alternativas* (H_a) podem ser construídas, embora apenas *uma hipótese nula* (H_0) seja possível. Conseqüentemente, as hipóteses possíveis seriam:

a) H_0: P(V) = P(A)
   H_a: P(V) > P(A)

b) H_0: P(V) = P(A)
   H_a: P(V) < P(A)

c) H_0: P(V) = P(A)
   H_a: P(V) ≠ P(A)

Em síntese: no exemplo (a), o pesquisador admite ser possível que o número de bolas vermelhas supere o de amarelas; no caso (b), ele resolve testar a hipótese contrária, qual a de que haveria menos bolas vermelhas do que amarelas; finalmente, em (c), não tendo nenhuma razão para optar, ele decide verificar se seria razoável imaginar que o número de bolas vermelhas fosse diferente do de bolas amarelas.

6. Aqui é que entra a questão do *nível de significância*. Na terminologia de Neyman e Pearson (In: Rodrigues, 1970, p. 132), *rejeitar uma hipótese verdadeira* é cometer um *Erro de Tipo I* e *deixar de rejeitar uma hipótese falsa* é cometer *Erro Tipo II*. Desses dois erros, o *bom senso* tem levado a considerar *a ocorrência de um Erro Tipo I mais grave* (em função das conseqüências) do que *a ocorrência de um Erro Tipo II*. Pois bem, o nível de significância, α, representa a *probabilidade de o pesquisador rejeitar uma hipótese verdadeira*, ou seja: α = *P(Erro Tipo I);* por outro lado, *se o*

pesquisador não rejeitar uma hipótese falsa, estará cometendo um Erro Tipo II, com probabilidade β. Portanto, **β = P (Erro Tipo II).** É evidente que não sendo humanamente possível evitar erros, embora tecnicamente seja sempre viável minimizá-los, o pesquisador procura trabalhar com um nível de significância que não comprometa dramaticamente o outro erro (Tipo II), porque *a redução de um deles sempre acarreta o aumento do outro.* Para visualizar essa situação, basta colocar esses erros nos extremos de uma gangorra, com a num dos extremos e b no outro.

7. Para tornar mais "competente" a utilização dos conceitos acima (V. § 6.), Neyman e Pearson criaram também o conceito de **poder,** que pode ser expresso pela expressão: *Poder* = **(1 - β)** e com significado eqüivalente à *capacidade do experimento para perceber a falsidade de uma hipótese.* Então, o *ideal* é que a e b sejam não só *pequenos* como *os menores possíveis,* mas atenção: *nenhum* deles pode ser *nulo* (zero), sob pena de o *outro* ser *máximo*! Aqui cabe um exemplo. Se um juiz não quiser carregar na consciência o fato de ter *condenado um inocente,* deverá *absolver* todos os supostos criminosos; por outro lado, se ele não quiser arcar com o fardo de ter *deixado livre um criminoso,* deverá *condenar,* sem exceção, todos os acusados. O julgamento "mais correto", portanto, depende da *reunião de um conjunto satisfatório de evidências a favor ou contra* o acusado. Nas provas estatísticas, o procedimento é o mesmo. E o conjunto satisfatório de evidências corresponde a uma *soma de probabilidades individuais,* até o *limite* estabelecido pelo *nível de significância.*

8. Outra maneira de trabalhar com o conceito acima é a seguinte: definir *poder* como a *probabilidade de não cometer Erro Tipo II,* e designar essa probabilidade por β, atentando para o fato de que α deverá *continuar sendo* o *menor possível,* mas β, nesse caso, o *maior possível.* Não faremos uso dessa possibilidade, menos comum, a fim de não complicar a compreensão dos gráficos apresentados no § 19. Ademais, a grande maioria dos livros que tratam do assunto prefere definir β como a *probabilidade de ocorrer Erro Tipo II.*

9. Retomemos agora o exemplo do § 5 e suponhamos que o pesquisador tenha razões para suspeitar da *prevalência de bolas vermelhas.* Então, as suas hipóteses estatísticas são:

$$H_0: P(V) = P(A)$$
$$H_a: P(V) > P(A)$$

Supondo impossível uma *contagem direta*, o pesquisador deve valer-se de um *experimento* que o auxilie a *tomar uma decisão inteligente*. Esse experimento pode ser o seguinte: retirar, com reposição, (n = 12) bolas e estudar o comportamento da variável X = número de bolas vermelhas (também poderiam ser amarelas). Suponhamos realizado o experimento, com o seguinte resultado: das 12 bolas retiradas (sempre com reposição!), 9 foram de cor vermelha. E agora? Achar que só porque 9/12=0,75 deve haver 75% de bolas vermelhas na urna é um raciocínio extremamente simplista – uma vez que exclui duas importantes considerações: (1.ª) a questão da *reposição*; (2.ª) a influência do *acaso*. A análise desse resultado depende de reconhecer que a distribuição de probabilidade subjacente ao experimento, isto é, que *regula* as ocorrências do tipo "sair bola vermelha", é a *distribuição binomial*, cujos parâmetros, nos termos do problema em foco, são:

$$n = 12$$

$$P(V) = P(A) = p = 0{,}5 \text{ (Modelo de cara-coroa.)}$$

10. As hipóteses estatísticas, com esses refinamentos, transformam-se, conseqüentemente, em:

H_0: **p = 0.50**
H_a: **p > 0,50** (Hipótese alternativa **unicaudal direita**)

Como salientado anteriormente, o *nível* de significância, a, deverá ser *pequeno*. E embora a fixação desse nível não seja da alçada exclusiva do estatístico – mas, com prioridade, do *pesquisador* familiarizado com a área de estudo -, vamos fixá-lo em 0,05, ou seja, 5%, de modo que seja possível *confiar em 95 de cada grupo de 100 repetições* do *mesmo* experimento.

11. O passo seguinte é escrever o *conjunto de todos os resultados possíveis* decorrentes do experimento. Esse conjunto, designado por **R**, e denominado *espaço experimental*, é

R : 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 bolas vermelhas (V),
sendo que X = 9 representa o resultado efetivamente conseguido.

Ora, se, de fato, houver prevalência das bolas vermelhas (V), o pesquisador espera que o *número de V seja grande*. Por isso, ele fixa sua atenção na *extremidade direita do espaço experimental* (R), e determina, para *cada valor, da direita para a esquerda*, a respectiva *probabilidade*. É essa lateralização que faz com que a hipótese alternativa seja denominada *unicaudal direita*. Uma tábua binomial resolve facilmente a questão. Assim:

$$P(X = 12) = 0^{+}$$
$$P(X = 11) = 0{,}003$$
$$P(X = 10) = 0{,}016$$
$$P(X = 9) = 0{,}054$$

Recapitulemos que, no parágrafo 7, o nível de significância foi associado a um conjunto satisfatório de evidências representado pela *soma* de probabilidades individuais. Pois bem, as probabilidades acima, somadas, dão 0,073, o que *ultrapassa o limite estabelecido a priori de 0,05*. Por essa razão, como a entrada se dá pela direita, a *solução* é *parar no valor X = 10*, donde resulta a seguinte nova soma: 0,019, agora bem *inferior* a 0,05.

12. Com esse procedimento, o pesquisador *dividiu* o conjunto **R** em dois *subconjuntos* $\mathbf{R_c^*}$ e $\mathbf{R_c}$, respectivamente, *região não-crítica* e *região crítica*, compostos da seguinte maneira:

$\mathbf{R_c^*}$ **: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9** ocorrências de V
$\mathbf{R_c}$ **: 10, 11, 12** ocorrências de V

Ocorre que, da maneira como foi especificada a $\mathbf{H_0}$, os valores pertencentes à *região não-crítica* decorrem naturalmente da *ação do acaso*, enquanto os valores da *região crítica*, da *ação de alguma outra variável* (além – sempre!! - da *ação do acaso*) *capaz de explicar ocorrências tão extremas*. Assim, diante desse quadro, a *conduta* do pesquisador deve ser *não-rejeitar a* H_0, uma vez que **(X = 9)** *pertence à região não-crítica*. E como a H_0 afirma que P(V) = P(A) = 0,50, *não rejeitá-la* significa *admitir a hipótese* de que, na urna, *o número de bolas vermelhas seja igual ao de bolas amarelas*.

13. Retomemos agora a questão do nível (a) fixado em 0,05 (ou 5%). Na verdade, por ter sido *estabelecido o nível antes da realização do experimento* (isto é, na fase do *planejamento experimental*), ele escapou do controle do pesquisador no sentido de que, de fato, o *ponto de corte* entre a *região*

crítica e a *região não-crítica* acabou determinando um *segundo nível de significância*, chamado de *efetivo* ou *observado*, para diferençá-lo do primeiro, obviamente *teórico*, e de *valor muito inferior a 0,05*, ou seja, *0,019* (ou 1,9%, o que dá no mesmo). A esse *segundo nível de significância (nível efetivo, observado)* os estatísticos têm, nas últimas décadas, atribuído o nome de **valor-p** (*p-value*), que corresponde à *probabilidade associada ao resultado experimental obtido*. Observemos que trabalhar com um *nível (teórico)* α ou com um **valor-p** implica utilizar *critérios de decisão distintos* – com *conseqüências* também *distintas*. O nível de significância (a) é sempre fixado *a priori*, portanto *antes da realização do experimento* (ou da pesquisa); o **valor-p**, por *não depender de um limite fixado antes, não tem restrições*, ficando sua *aceitação* ou *rejeição* por conta do *próprio pesquisador!* Então, *dois pesquisadores que testem a mesma hipótese, realizando o mesmo experimento, e obtendo o mesmo resultado (X=9), poderão tomar decisões diferentes, dependendo da disposição pessoal de transigir mais ou menos*. De fato, com p=0,073, um pesquisador poderá achar que a H_0 deva ser *não-rejeitada*, enquanto outro, mais exigente, poderá fazer exatamente o oposto. Com isso, o *tamanho da região crítica* torna-se excessivamente *elástico*, e o *consenso entre pesquisadores, menos provável*. Com o *nível de significância* (a) isso *jamais* ocorreria: *até o limite de 0,05, o que caísse na* R_c *levaria à rejeição da* H_0 *e o que fora desse limite, à não-rejeição da* H_0. O segredo por trás de tudo isso? Simples: se não declararmos, *antes* do jogo, para que time torcemos, nosso time poderá ser sempre *vencedor* – já que o *critério*, explicitado *a posteriori*, tem forte *comprometimento emocional*.

14. Autores que insistem em trabalhar com os dois critérios (**nível** e **valor-p**) costumam acenar com a seguinte *regra: se o valor-p for menor do que o nível* α, *não-rejeitar a* H_0; em caso contrário, isto é, se o valor-p for maior do que o nível α, *rejeitar a* H_0. Essa regra mostra que calcular o *nível efetivo* (rebatizado de **valor-p**) é procedimento quase *inevitável*, pois somente a partir do seu cálculo é possível *situá-lo com relação ao valor de* α. Por essa razão é que, no § 1, deste artigo, referimo-nos a critérios que, embora *não eqüivalentes*, não deixam de ser *complementares*. A discussão, portanto, não gira em torno de a **e** valor-p, ambos *necessários* (a fixado *a priori* e valor-p calculado *a posteriori*), mas em torno de a *ou* valor-p, *alternativamente*, como querem alguns.

15. Outros autores, mais prudentes, como R. S. Witte e J. S. Witte (1997), preferem tratar do assunto sob dois ângulos distintos: (a) *mérito* de um

procedimento menos estruturado; (b) *fragilidade* de um procedimento menos estruturado. No primeiro caso (a), dizem esses autores que, tendo sido *eliminada a necessidade de rejeitar ou não-rejeitar a hipótese nula* (H_0), a *decisão* pode ser *postergada* até que *mais pesquisas* sejam *realizadas* e *mais evidências* tenham sido *colhidas a favor* ou *contra* determinada hipótese. Dizem eles, complementarmente, que essa perspectiva torna-se muito atraente em casos denominados *borderline*, isto é, casos em que *o ponto de corte de* **R** (espaço experimental) *separe probabilidades muito próximas*. No segundo caso (b), e aí concordamos, dizem eles que, *dada a falta de comprometimento com algum nível de significância pré-determinado, fica difícil avaliar as conseqüências decorrentes dos erros Tipo I e II.*

16. Vamos agora retomar o experimento analisado e, resumidamente, consolidar as conclusões. Tendo sido demarcado o ponto de corte de R entre os valores X=9 e X=10, a H_0 foi não-rejeitada uma vez que $(X=9) \in R_c^*$. Isso eqüivale a dizer que, com base no resultado experimental, tudo se passa *como se*, na urna, houvesse igual quantidade de bolas vermelhas e bolas amarelas. Primeiro, observemos que essa declaração *independe* da *quantidade* de bolas de cada cor dentro da urna; segundo, *não* é preciso que, rigorosamente, o número de bolas vermelhas seja igual ao número de bolas amarelas, pois uma *pequena diferença* pode ficar razoavelmente bem *protegida* pelo nível estabelecido (5%). Além disso, o experimento, nas condições em que foi realizado, mostrou ser *impossível* cometer erro superior a 2% (arredondando), o que faz com que as *conclusões* possam ser expressas com aproximadamente *98% de confiança.*

17. A análise dos dados não termina aí. Observemos que, *não tendo sido rejeitada a hipótese nula*, certamente *não pode ter sido cometido Erro Tipo I*, mas nada impede que um *Erro Tipo II tenha sido cometido* (já que ele só ocorre quando uma hipótese falsa, que deveria ter sido rejeitada, deixou de sê-lo). Ocorre que a *hipótese alternativa* H_a: *P(V)>0,50 sinaliza um conjunto ilimitado de probabilidades*: 0,51 ou 0,60 ou 0,63 ou 0,76 ou 0,80 ou 0,95 etc. etc. para citar apenas situações com duas casas decimais, embora possam ser considerados valores do tipo 0,679 ou 0,7984 etc etc. Então, se, eventualmente, tiver ocorrido Erro Tipo II, é porque a P(V), nas condições do problema, pode ser maior do que 0,50. Para verificar as *conseqüências*, em termos de *Erro Tipo II*, vamos, simplificadamente, considerar apenas as probabilidades maiores do que 0,50 expressas por

duas casas decimais e terminadas em 0 ou 5. Novamente, recorrendo a uma tábua de probabilidades binomiais, obtemos:

p=0,55	p=0,60	p=0,65	p=0,70
P(X=12)= 0,001	P (X=12)= 0,002	P(X=12)= 0,006	P(X=12)= 0,014
P(X=11)= 0,008	P (X=11)= 0.017	P(X=11)= 0,037	P(X=11)= 0,071
P(X=19)= 0,034	P(X=10)= 0,064	P(X=10)= 0,109	P(X=10)= 0,168
Soma= **0,043**	Soma= **0,083**	Soma= **0,152**	Soma= **0, 253**
4,3%	**8,3%**	**15,2%**	**25,3%**

p=0,75	p=0,80	p=0,85	p=0,90
P(X=12)= 0,032	P(X=12)= 0,069	P(X=12)=0,142	P(X=12)= 0,282
P(X=11)= 0,127	P(X=11)= 0,206	P(X=11)= 0,301	P(X=11)= 0,377
P(X=10)= 0,232	P(X=10)= 0,283	P(X=10)= 0,292	P(X=10)= 0,230
Soma= **0,391**	Soma= **0,558**	Soma= **0,735**	Soma= **0,889**
39,1%	**55,8%**	**73,5%**	**88,9%**

p=0,95
P(X=12)= 0,540
P(X=11)= 0,341
P(X=10)= 0,099
Soma= 0,980
99,0%

Bem, essas somas de probabilidades correspondem aos vários *poderes* que as *regiões críticas* assumem quando, para o *mesmo grupo de valores* (X=12,11,e10), fazemos *variar* o parâmetro **p=P(V)**. Ainda que redundantemente, recapitulemos que *poder é a capacidade de a região crítica perceber a falsidade da hipótese*.

Cada uma das somas de probabilidades acima, convertida em porcentagem, representa a *sensibilidade do experimento* (=*poder*) para *evitar* que o pesquisador tome, como já foi dito, "gato por lebre". E a *diferença* entre 100% e cada uma dessas porcentagens corresponde ao *erro de julgamento* (Erro Tipo II). Então, traduzindo essas porcentagens em interpretações, temos as conseqüências apresentadas abaixo.

a) Se p=0,55, o poder da Rc é da ordem de 4,3% e o risco de um julgamento errado é de (100% - 4,3%)=95,7%. *Péssimo*!
b) Se (0,55 ≤ p ≤ 0,85), o poder não é lá muito brilhante, uma vez que, na melhor das opções, o julgamento errado ocorrerá em (100 – 73,5)% = 26,5% das vezes.
c) Por igual raciocínio, verificamos que o poder começa a ficar *bom* a partir de p=0,90, quando o risco de "gato por lebre" vai de 11,1% até 1%, no caso de p=0,95.

Então, que *conclusão* podemos tirar do experimento realizado? Simplesmente a de que ele é *ótimo* para perceber a *veracidade* da H_0, no caso de ela ser, *de fato, verdadeira*; entretanto, no caso de ela ser *falsa*, ele é *péssimo* para detectar tal situação, a menos que **p** esteja no *entorno* de 0,90, quando a *"mecânica do processo"* praticamente *impedirá* que *erros de julgamento de absurdas proporções* sejam cometidos.

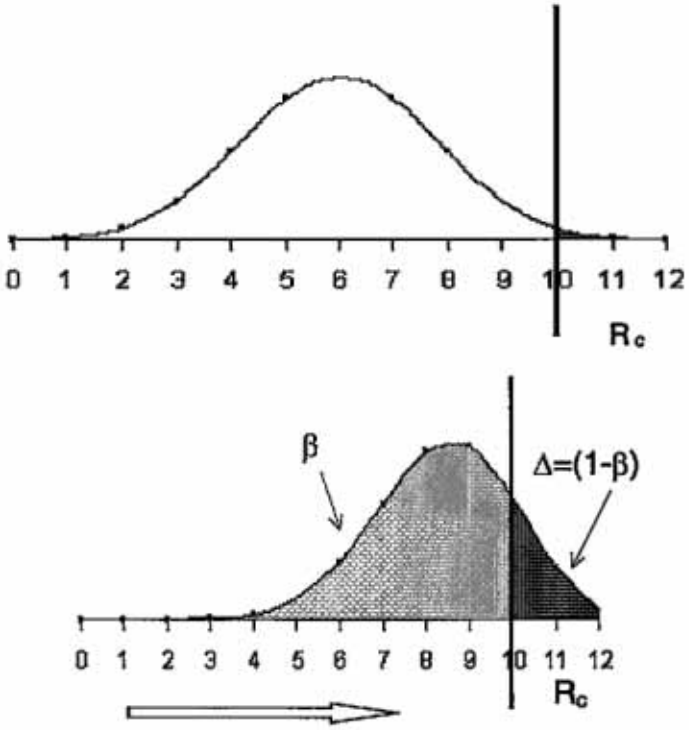
18. Observemos que, em todos os cálculos feitos no § 17, os valores de X foram sempre os mesmos: 12, 11 e 10. Por quê? Porque a *região crítica* (R_c) é *invariante* a partir do momento em que o *nível de significância* tenha sido *escolhido*. O pesquisador, obviamente, *desconhece* o fato de estar diante de uma *hipótese falsa*; por isso, comporta-se *como se a* $\mathbf{H_0}$ *fosse verdadeira*, o que determina, de forma rígida, o *limite* entre a $\mathbf{R_c}$ e a $\mathbf{R_c^*}$. A análise que acabamos de fazer mostra que o *experimento* foi *mal concebido*, tornando a *conclusão* extremamente *frágil* se, de fato, a $\mathbf{H_0}$ for *falsa*. Decorrem dessa exposição duas *importantes lições*:

1.ª) quando o *experimento* é *mal planejado*, dele resultam inevitavelmente *conclusões claudicantes* e *pouco úteis*;

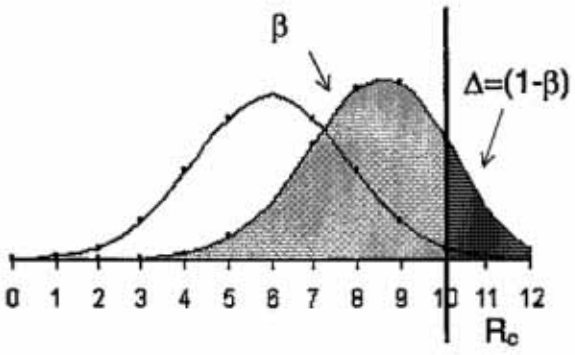
2.ª) sem a *especificação, a priori*, do *nível de significância*, os cálculos acima (§ 17) não seriam *possíveis*, uma vez que o *ponto de corte* entre a R_c e a R_c^* *não estaria identificado*.

19. Os gráficos da página seguinte mostram duas situações: a *básica*, sob a suposição de que a $\mathbf{H_0}$ seja *verdadeira*, e algumas *opções* decorrentes da *variação* de **p**, se a $\mathbf{H_0}$ for *falsa*. Chamamos a atenção do leitor para as *áreas* que vão sendo delimitadas por *sobreposição* à medida que **p** assume os valores de **0,70** e **0,90** (apenas dois, a título de ilustração).

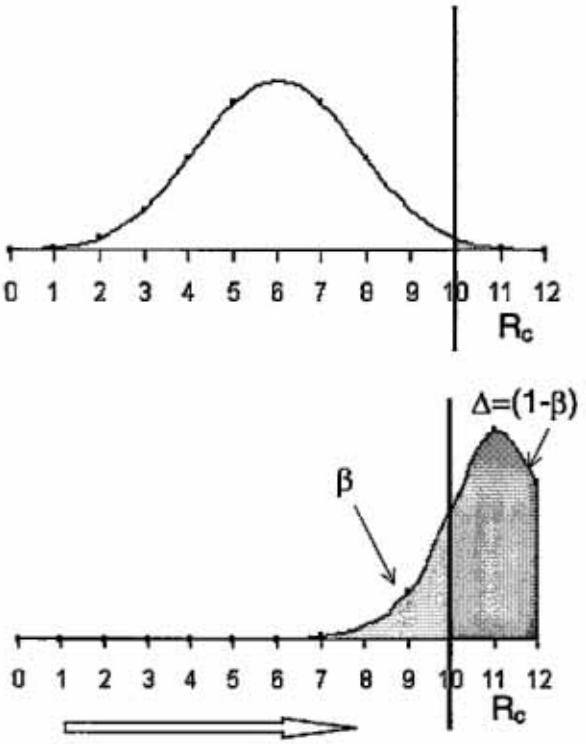
Em branco: X ç B(12; 050)
Em tonalidades de cinza: X ç B(12; 0.70)



A sobreposição dos dois gráficos mostra que, logicamente, à medida que **aumenta α diminui β** e vice-versa.



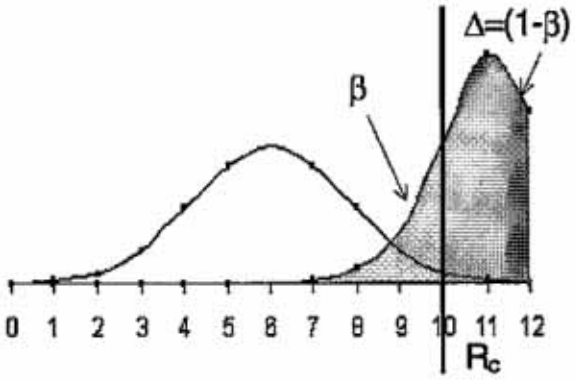
O poder, definido agora como $\Delta = (1 - \beta)$, pode ser visualizado na parte *mais escura* do gráfico (à direita).
Em branco: $X \rightarrow B(12; 0{,}50)$
Em tonalidades de cinza: $X \rightarrow B(12; 0{,}90)$



A sobreposi→ão dos gráficos mostra, mais uma vez, o que ficou dito na página anterior: o *aumento* de α implica a *diminuição* de β e vice-versa. Além disso, *mantido fixo* α (em 0,05), à medida que *varia* o valor de **p** (de 0,70 para 0,90), a curva sobreposta (a curva **assimétrica**) vai produzindo uma *redução* no valor de β e conseqüente *aumento* do *poder*, agora definido como **Δ = (1 - β)**. Resultado: mantido *inalterado* o *nível de significância*, e para o mesmo valor de **n** (no caso, 12), *diminui* progressivamente a ***probabilidade de ocorrer Erro Tipo II*** e, *a fortiori*, *aumenta* o *poder do experimento*. Isso tudo vem condensado nos cálculos apresentados no § 17.

Observação importante

A *distribuição binomial* é *discreta*. Por isso, rigorosamente, os *pontos* de seus *gráficos* não deveriam ser unidos por *linha contínua*. A regra *não* foi observada *de propósito* – para *evidenciar* as *áreas* correspondentes a **α , β e (1 - β)**.



20. Em *controle de qualidade*, a denominação *Erro Tipo I* é conhecida por *producer's risk*, ou seja, *risco do produtor*, no sentido de que, num carregamento de matéria-prima, ele poderia retirar, por *mero acaso*, uma *amostra* de *qualidade duvidosa*, embora o carregamento, em sua *quase totalidade*, estivesse em *ordem*. Isso levaria o produtor a *rejeitar* um *bom carregamento* apenas porque *casualmente* uma amostra *não correspondeu ao resultado esperado*. E por que *Erro Tipo I*? Porque *foi rejeitada uma hipótese verdadeira, quando, de fato, isso não deveria ter ocorrido*! O *Erro Tipo II*, também nesse contexto, recebe denominação especial: *consumer's risk*, isto é, *risco do consumidor*, no sentido de que ele poderia, *casualmente*, selecionar uma *boa amostra* de uma *produção ruim*, o que acarretaria uma compra indevida. Nesse caso, ele estaria levando "gato por lebre"!

21. Esperamos ter, com este artigo, mostrado que *não* se trata de *optar* pelo *nível de significância* (α) *ou* pelo **valor-p**, mas, sim, de *considerar ambos*, de forma *complementar*. Esperamos também ter deixado claro que a *fixação, a priori*, do nível de significância *implica* a existência de boas razões, por parte do pesquisador, em *aderir* a esse nível, o que traduz *compromisso* com a *rejeição* ou a *não-rejeição* da *hipótese* que seja *objeto de teste*.

22. Desejamos também salientar que esse exercício *hipotético* que acabamos de fazer (com a *distribuição binomial*) pode ser realizado com variáveis que tenham subjacentemente *outras* distribuições de probabilidade. Para facilitar os cálculos, existem hoje, disponíveis no mercado, *softwares* que dão conta de oferecer ao pesquisador, com rapidez, a maior parte dos resultados de que ele possa necessitar.

Referências Bibliográficas

COSTA, Sérgio Francisco. **Introdução Ilustrada à Estatística.** 3.ª ed. São Paulo: Harbra, 1998.

LEVIN, Jack. **Estatística Aplicada a Ciências Humanas.**2.ª ed. Tradução e adaptação de Sérgio F. Costa. São Paulo: Harbra, 1985.

RODRIGUES, Mílton da Silva. **Dicionário Brasileiro de Estatística.** 2.ª ed. Rio de Janeiro: Fundação IBGE & Instituto Brasileiro de Estatística, 1970.

ULLMAN, Neil R. **Elementary Statistics - An Applied Approach.** USA: John Wiley & Sons, Inc., 1978.

WITTE, Robert S & WITTE, John S. **Statistics.** 5.ª ed. USA: Harcourt Brace College Publishers, 1997.